Motrenko A.P., Rudakov K.V., Strijov V.V. Combining endogenous and exogenous variables in a special case of non-parametric time series forecasting model // Moscow University Computational Mathematics and Cybernetics, 2016, 40(2) : 71-78.

УДК 004.852

А. П. Мотренко[1], **К. В. Рудаков**[2], **В. В. Стрижов**[3]

УЧЕТ ВЛИЯНИЯ ЭКЗОГЕННЫХ ФАКТОРОВ ПРИ НЕПАРАМЕТРИЧЕСКОМ ПРОГНОЗИРОВАНИИ ВРЕМЕННЫХ РЯДОВ*

Решается задача повышения качества прогнозирования временных рядов путем учета влияния экзогенных факторов. Для учета экзогенных факторов предлагается уточнить гистограмму прогнозируемого (эндогенного) временного ряда, используя информацию о реализации значений экзогенных временных рядов. Исследуется алгоритм *hist*, основанный на методах квантильной регрессии. Алгоритм hist вычисляет прогноз на основе свертки гистограммы временного ряда с функцией потерь. Рассматриваются методы уточнения гистограммы: условные гистограммы, смеси гистограмм. Предлагаемые подходы проиллюстрированы задачей о прогнозировании объемов грузовых железнодорожных перевозок.

Ключевые слова: прогнозирование временных рядов, уточнение прогнозов, экзогенные факторы, гистограммное прогнозирование, квантильная регрессия.

1. Введение. Включение экзогенных временных рядов в прогностическую модель позволяет увеличить точность прогнозов за счет учета скрытых изменений системы [17, 22]. Способ учета экзогенных временных рядов зависит от структуры модели. В случае, когда модель линейна, учет внешнего фактора заключается в аддитивном добавлении нескольких значений, или их преобразований, экзогенного временного ряда в модель. Модель авторегрессионного скользящего среднего ARMA (autoregressive moving average) [24, 17], широко используемая при краткосрочном прогнозировании временных рядов [10], содержит три аддитивных компоненты: авторегрессионную, скользящее среднее и ошибку. Модель ARMAX (eXogenous autoregressive moving average) [14] — экзогенная модификация модели авторегрессионного скользящего среднего — включает также комбинацию экзогенных временных рядов в качестве дополнительной аддитивной компоненты.

Целью данной работы является повышение качества прогнозирования алгоритма *hist*, предложенного в работе [23]. Прогноз алгоритма *hist* равен центру столбца гистограммы прогнозируемого временного ряда, соответствующему оптимальному значению свертки гистограммы с функцией потерь. Для учета влияния экзогенных факторов предлагается уточнить гистограмму прогнозируемого временного ряда с использованием значений

[1]факультет УПМ МФТИ, аспирант, E–mail: anastasiya.motrenko@phystech.edu

[2]ВЦ РАН, профессор, чл.-корр. РАН E–mail: rudakov@ccas.ru

[3]факультет УПМ МФТИ, доцент., д.ф.-м.н. E–mail: strijov@ccas.ru

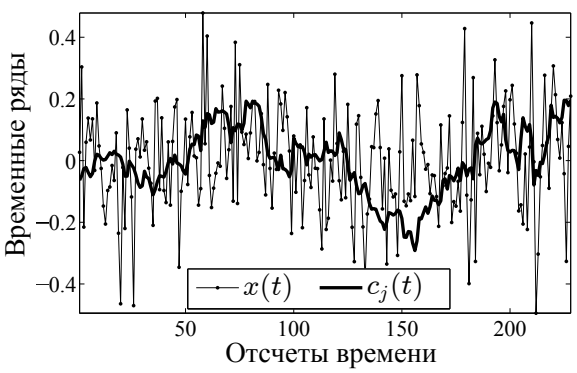
*Работа выполнена при финансовой поддержке Министерства образование и науки Российской Федерации, по соглашению RFMEFI60414X0041.

экзогенных временных рядов. Простейший способ уточнения гистограммы заключается в построении гистограммы, условной по значениям экзогенных временных рядов. Для этого оценивается многомерная гистограмма, приближающая совместное распределение прогнозируемого временного ряда и экзогенных временных рядов, а затем выбирается срез этой гистограммы на основе конкретных реализаций значений экзогенных временных рядов. При построении многомерных гистограмм ограничением является длина предыстории рассматриваемых временных рядов. Это ограничение связано с разреженностью таблиц при увеличении их размерности (количества экзогенных временных рядов). Существуют методы снижения размерности в некоторых частных случаях [7, 6]. В данной работе в качестве альтернативы многомерным гистограммам предлагается использование взвешенной суммы условных гистограмм для учета экзогенных временных рядов при гистограммном прогнозировании. Такой подход менее требователен к длине истории, но требует оценки большего числа параметров. Приближая гистограмму взвешенной суммой гистограмм, достаточно вычислить n двумерных гистограмм (где n — число экзогенных временных рядов), вместо одной n-мерной, что значительно ослабляет требования к длине предыстории рассматриваемых временных рядов: количество оцениваемых значений уменьшается от экспоненциального по n до линейного.

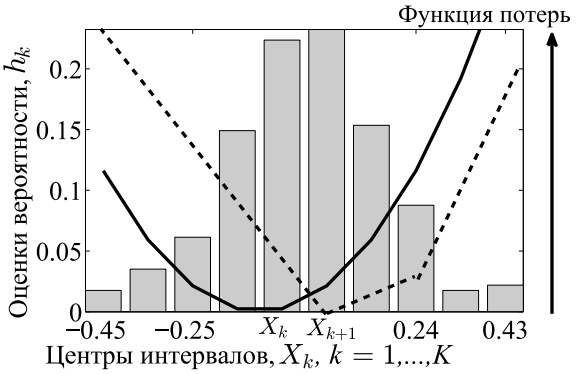
Смеси гистограмм используются как более устойчивая альтернатива смесям моделей, применяемым для оценки широкого класса плотностей распределений [15]. Плотность распределения исследуемой величины при этом приближается взвешенной суммой параметрических функций плотности. Согласно [15], любую функцию плотности можно приблизить смесью гауссиан с произвольной точностью. В работах [25, 4, 3] описаны некоторые модификации смесей моделей. Авторы работ [1, 9, **?**] применяли смеси гистограмм при решении задачи распознавании объектов на изображениях и видео. При вычислении взвешенной суммы в перечисленных работах предполагается, что интервалы разбиения гистограмм совпадают. В работах [18, 2] вводятся арифметики над гистограммами, позволяющие также рассматривать сумму гистограмм с произвольным разбиением на интервалы.

3. Постановка задачи. Обозначим через $\mathbf{x} = \{x(t)\}_{t=1}^{T-1}$ эндогенный временной ряд, $\mathbf{c}_j = \{c_j(t)\}_{t=1}^{T}$ — j-ый экзогенный временной ряд, $j = 1, \ldots, n$. На рис. 1(а) приведен пример эндогенного и экзогенного рядов. Эндогенный временной ряд $\mathbf{x}$ отложен тонкой линией, каждый отсчет $x(t)$ обозначен точкой. Пример экзогенного временного ряда $\mathbf{c}_j$ отложен жирной черной линией.

3.1. Алгоритм *hist*. Гистограмма $H = \langle \mathbf{X}, \mathbf{h} \rangle$ временного ряда $\mathbf{x}$, где $\mathbf{X}$ — вектор центров интервалов $\mathbf{X} = [X_1, \ldots, X_k, \ldots, X_K]$, а $\mathbf{h} = [h_1, \ldots, h_k, \ldots, h_K]$ — вектор связанных с



(a)



(b)

Рис. 1: (a) Примеры эндогенного $\mathbf{x}$ и экзогенного $\mathbf{c}_j$ временных рядов (b) Иллюстрация алгоритма *hist*: гистограмма прогнозируемого временного ряда сворачивается с квадратичной и асимметричной кусочно-линейной функциями потерь в различных точках X_k

X_k вероятностей, задает вероятностное распределение

$$h_k = P\big(x = X_k\big), \quad \sum_{k=1}^{K} h_k = 1.$$

Будем приближать распределение отсчетов временного ряда $x(t)$ с помощью гистограммы $H = \langle \mathbf{X}, \mathbf{h} \rangle$

$$h_k = \frac{1}{T} \sum_{t=1}^{T} [x \text{ принадлежит } k\text{-му интервалу}],$$

где $[\,\cdot\,]$ — индикаторная функция, а интервалы с центрами X_k задают равномерное разбиение множество значений временного ряда $\mathbf{x}$.

При заданных функции потерь $L(x, \hat{x}) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_+$ и некоторой оценке гистограммы H значение прогноза $hist(L, H)$ равно центру X_k столбца гистограммы, соответствующему минимальному значению свертки

$$\sum_{x \in X_1, \ldots, X_K} h_k L(X_k, x)$$

гистограммы H с заданной функцией потерь L. Значение свертки вычисляется в каждом из центров столбцов X_k, $k = 1, \ldots, K$. Затем в качестве прогноза $\hat{x}$ выбирается значение X_k, соответствующее минимуму свертки. На рис. 1(b) две функции потерь — квадратичная, график которой отложен сплошной линией, и асимметричная кусочно-линейная, отложенная пунктирной линия, — сворачиваются с гистограммой H в точках X_k и X_{k+1}, соответственно. Таким образом, прогноз временного ряда $\mathbf{x}$ с помощью алгоритма *hist*

есть решение задачи оптимизации:

$$\hat{x} = hist(H, L) = \underset{x \in X_1, \ldots, X_K}{\arg\min} \sum_{k=1}^{K} h_k L(x, X_k). \tag{1}$$

3.2. Уточнение гистограмм. Требуется на основе измеренных значений эндогенного временного ряда $\mathbf{x}$ и экзогенных временных рядов $\mathbf{c}_1, \ldots, \mathbf{c}_n$ построить уточненную гистограмму H, минимизирующую функцию потерь L. При фиксированной функции потерь L, прогноз (1) зависит от вектора $\mathbf{X}$ центров интервалов и вектора $\mathbf{h}$ высот столбцов гистограммы H. Для уточнения гистограммы H фиксируем центры интервалов X_k (считая временной ряд $\mathbf{x}$ стационарным) и будем варьировать h_k:

$$L(x(T), \hat{x}) \to \min_{\substack{\mathbf{h} \in [0,1]^K \\ \sum h_k = 1}}. \tag{2}$$

Будем искать решение задачи (2) в виде смеси условных гистограмм. Табл. 1 (a) и (b) иллюстрирует суть предлагаемого метода. Обозначим с помощью $H(t)$ уточненную гистограмму, оцененную в момент времени t. Предполагается, что

$$H(t) \equiv H(\mathbf{x}^{1:t-1}, \mathbf{c}_1^{1:t}, \ldots, \mathbf{c}_n^{1:t}),$$

то есть $H(t)$ построена с учетом первых $t-1$ значений $\mathbf{x}^{1:t-1}$ ряда $\mathbf{x}$ и первых t значений $\mathbf{c}_j^{1:t}$ экзогенных временных рядов $\mathbf{c}_j$, $1 \le j \le n$. Введем также обозначения $H_j(t) = \langle \mathbf{X}, \mathbf{h}^j \rangle$, $j = 1, \ldots, n$ для условной по $\mathbf{c}_j$ гистограммы, соответствующий значению $c_j(t)$; $H_0(t)$ для гистограммы временного ряда $\mathbf{x}$, построенной без учета внешних факторов.

Связь между маргинальной H_0 и условной H_j гистограммами проиллюстрирована табл. 1. Столбцы таблицы Табл. 1 (a) представляет двумерную гистограмму, значения p_{kg} которой приближают совместное распределение $\mathbf{x}$ и $\mathbf{c}_j$:

$$p_{kg} \approx \mathrm{P}\big(x(t) \text{ интервалу с центром } X_k,\ c_j(t) \text{ интервалу с центром } C_g^j\big),$$

и связаны со столбцами гистограмм H_0 и H_j соотношениями

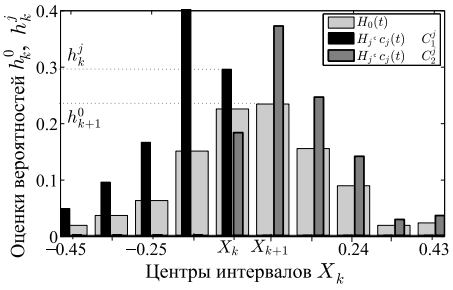
$$h_k^0 = \sum_{g=1}^{N} p_{kg} \quad \text{и} \quad h_k^j = \frac{p_{kg}}{p_g}, \quad p_g = \sum_{k=1}^{K} p_{kg}.$$

Рисунок в табл. 1 (b) иллюстрирует случай для количества $N = 2$ интервалов разбиения значений экзогенного временного ряда $\mathbf{c}_j$. Условные гистограммы $H_j(t)$, вычисленные по историческим значениям $\mathbf{x}^{1:t-1}$ при $c_j(t)$, принадлежащем первому и второму столбцу соответствующей маргинальной гистограммы, изображены с помощью более тонких столбцов. Светло-серым цветом изображена маргинальная гистограмма $H_0(t)$ эндогенного временного ряда.

Таблица 1: Иллюстрация связей между совместной, условными и маргинальными гистограммами.

	C_1^j	…	C_g^j	…	C_N^j	$\sum_g$
X_1	p_{11}	…	$p_{1g} = \boldsymbol{h}_1^j \cdot p_g$	…	p_{1N}	$\boldsymbol{h}_1^0$
X_2	p_{21}	…	$p_{2g} = \boldsymbol{h}_2^j \cdot p_g$	…	p_{2N}	$\boldsymbol{h}_2^0$
…	…	…	…	…	…	…
X_K	p_{K1}	…	$p_{K2} = \boldsymbol{h}_k^j \cdot p_g$	…	p_{KN}	$\boldsymbol{h}_K^0$
$\sum_k$	p_1^j	…	p_g^j	…	p_N^j	1

(a) Двумерная совместная гистограмма



(b) Примеры маргинальной и условных гистограмм

Используя введенные обозначения, запишем искомое решение задачи (2) в виде

$$H(T) = w_0 H_0(T) + \sum_{j=1}^{n} w_j H_j(T), \tag{3}$$

где вектор весов $\mathbf{w} = [w_0, \ldots, w_n]^\intercal$ доставляет максимальное правдоподобие модели $P(\mathbf{w}|\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_n)$.

Для оценки правдоподобия модели (3) разложим его на множители

$$P(\mathbf{w}|\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_n) = \frac{P(\mathbf{x}|\mathbf{w}, \mathbf{c}_1, \ldots, \mathbf{c}_n) P(\mathbf{w}, \mathbf{c}_1, \ldots, \mathbf{c}_n)}{P(\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_n)}.$$

Предполагается, что $\mathbf{c}$ и $\mathbf{w}$ независимы $P(\mathbf{w}, \mathbf{c}_1, \ldots, \mathbf{c}_n) = P(\mathbf{w}) P(\mathbf{c}_1, \ldots, \mathbf{c}_n)$, а веса распределены равномерно на единичной сфере в L_1. Совместные вероятности $P(\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_n)$ и $P(\mathbf{c}_1, \ldots, \mathbf{c}_n)$ не зависят от $\mathbf{w}$ и, таким образом, не влияют на результат оптимизации $P(\mathbf{w}|\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_n)$ при фиксированном наборе $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ экзогенных временных рядов. Плотность распределения $P(\mathbf{x}|\mathbf{w}, \mathbf{c}_1, \ldots, \mathbf{c}_n\})$ будем приближать с помощью гистограммы $H(t)$:

$$P(x(t)|\mathbf{w}, \mathbf{c}_1^{1:t}, \ldots, \mathbf{c}_n^{1:t}) \approx h_{k(t)},$$

где $k(t)$ — номер интервала гистограммы $H(t)$, содержащего значение $x(t)$:

$$\frac{1}{2}(X_{k(t)} + X_{k(t)-1}) \leq x(t) < \frac{1}{2}(X_{k(t)+1} + X_{k(t)}).$$

Тогда, считая значения $x(t)$ независимыми при известных $\mathbf{w}$ и $\{\mathbf{c}\}$ оценим

$$P(\mathbf{x}|\mathbf{w}, \mathbf{c}_1, \ldots, \mathbf{c}_n) = \prod_{t=t_{\min}}^{T} P(x(t)|\mathbf{w}, \mathbf{c}_1^{1:t}, \ldots, \mathbf{c}_n^{1:t}) \approx \prod_{t=t_{\min}}^{T} h_{k(t)},$$

где $t_{\min}$ — минимальная длина временного ряда $\mathbf{x}$, необходимая для оценки гистограммы. Так как количество рассматриваемых экзогенных временных рядов может быть слишком

велико, при оценке весов предлагается отобрать наиболее информативные временные ряды $\mathbf{c}_j$, $j \in \mathcal{J} \subseteq \{0, 1, \ldots, n\}$. В результате, переформулируем оптимизационую задачу (2) в виде:

$$\mathbf{w} = \underset{\substack{\mathbf{w}\in[0,1]^{|\mathcal{J}|} \\ \sum_{j\in\mathcal{J}} w_j = 1}}{\arg\max} \frac{1}{|\mathcal{J}|} \sum_{t=1}^{T} \log \left(\sum_{j\in\mathcal{J}} w_j h_k^j(t) \right),$$

учитывающем количество $|\mathcal{J}|$ отобранных компонент смеси.

4. Оценка весов смеси с выбором компонент. Пусть заданы центры $X_1, \ldots, X_K$ и $C_1^j, \ldots, C_N^j$ интервалов гистограмм $H_0(t)$ и $H_j(t)$. Для оценки весов w_j компонент смеси воспользуемс стохастической модификацией EM-алгоритма с отбором компонент (*Stochastic Expectation Maximization*, SEM), описанной в работе [8]. Алгоритм состоит из двух итеративно повторяющихся этапов: этапа сэмплирования временного ряда из распределения, задаваемого текущей оценкой $H(t)$, и EM-этапа, на котором отбираются информативные компонеты и пересчитываются веса оставшихся компонент смеси. Идея процедуры отбора компонент заключается в последовательной настройке весов с удалением компонент смеси, описывающих слишком малую часть выборки, сгенерированной на этапе сэмплирования.

Пусть заданы: максимальное число $n_{\max}$ компонент смеси, минимальна длина $t_{\min}$ временного ряда $\mathbf{x}$; минимальная доля α выборки, описываемая каждой из компонент; начальное приближение весов w_j компонент и набора информативных компонент $\mathcal{J} = 0, \ldots, n$. Процедура оценки компонент с отбором компонент состоит в следующем:

1. Сэмплируется выборка $\tilde{\mathbf{x}}^{t_{\min}+1:T} = \{\tilde{x}(t)\}_{t=t_{\min}+1}^{T}$ согласно распределению вероятностей w_{jt}:

   $$\hat{x}_t \sim \sum_{j=\in\mathcal{J}} w_{jt} H_j(t).$$

   На первой итерации w_{jt} полагаются равными значениям w_j для всех t. Пусть $\tilde{x}(t)$ принадлежит $k(t)$-му интервалу гистограммы $H(t)$. На основе сэмплированной выборки $\tilde{\mathbf{x}}$ вычисляется количество T_j элементов выборки, описываемых j-ой компонентой:

   $$T_j = \sum_{t=t_{\min}}^{T} \left[\underset{j\in\mathcal{J}}{\arg\max}\, h_{k(t)}^j = j \right], \tag{4}$$

   где $[\cdot]$ — индикаторная функция. Компоненты, сгенерировавшие $T_j < \alpha(T - t_{\min})$ элементов выборки, удаляются из модели (3):

   $$\mathcal{J} := \mathcal{J} \setminus \{j :\ T_j < \alpha(T - t_{\min})\}.$$

   Для оставшихся компонент T_j пересчитываются согласно (4) и вычисляются оценки весов w_j

   $$w_j := T_j / T.$$

2. Пересчитывается распределение w_{jt}:

$$w_{jt} := \frac{w_j h_k^j(t)}{\sum w_j h_k^j(t)}, \text{ где } x(t) \text{ принадлежит } k(t)\text{-му интервалу.}$$

Шаги алгоритма повторяются, пока число компонент $|\mathcal{J}|$ не станет меньше или равно $n_{\max}$. В результате работы алгоритма будет отобран набор наиболее информативных временных рядов $\mathbf{J}$ и получены оценки весов w_j, $j \in \mathcal{J}$ соотвествующих компонет смеси. Учитывая результаты исследований сходимости алгоритма SEM [8], для получения более стабильных оценок весов w_j, усредним оценки $\mathbf{w}^{(i)}$ весов, полученные при запуске алгоритма без выбора компонент ($\alpha = 0$) после завершения алгоритма с выбором компонет:

$$\mathbf{w} = \frac{1}{r} \sum_{i=1}^{R} \boldsymbol{w}^{(i)},$$

где R – количество дополнительных итераций, $\mathbf{w}^{(i)}$ – оценка, полученная на i-той дополнительной итерации.

5. Вычислительный эксперимент. Описанный метод метод уточнения гистограммы с выбором наиболее информативных экзогенных временных рядов был протестирован на данных о грузовых железнодорожных перевозках. Было рассмотрено 38 временных рядов, содержащих информацию о железнодорожных перевозках различных типов грузов. Каждый отсчет $x(t)$ временного ряда $\mathbf{x}$ которых соответствует одному дню и равен суммарному весу (в тоннах) определенного груза. В качестве экзогенных временных рядов $\mathbf{c}_j$ были рассмотрены цены на сахар, бензин, медь, цинк, золото, никель, пшеницу, мазут, газ, олово, нефть, серебро и свинец за рассматриваемый период времени. Данные содержат значительное число пропусков; длина временных рядов после обработки составляет 228 измерений. В экспериментах из всех рядов был удален линейный тренд. Дополнительно значения временных рядов были нормированы на отрезок [0, 1].

Выбор параметров N и K. Следуя оценке $K = \lceil 3\sqrt[3]{T} \rceil$, полученной в работе [19], было выбрано $K = 15$. Для выбора оптимального значения N интервалов разбиения величины $\mathbf{c}_j$, для каждого экзогенного временного ряда были получены оценки вероятности его включения в модель на основе нескольких запусков процедуру отбора компонент. С помощью теста Крускала-Уоллиса [21] протестирована гипотеза о независимости вероятностей от N. Наблюдаемые величины p-value составили около 0.95, то есть наблюдаемых данных недостаточно для отвержения гипотезы, и результаты отбора компонент скорее всего не зависят от N. Также не удалось обнаружить зависимость качества прогнозирования эндогенных временных рядов от N. Так как минимальная необходимая длина $t_{\min}$ временного ряда зависит от N линейно, было выбрано минимальное значение $N = 2$.

Результаты отбора компонент. При поиске информативных экзогенных временных рядов были рассмотрены также производные $\mathbf{c}_j^\circ$ от исходных временные ряды, соответствующие промежуткам возрастания и невозрастания

$$\mathbf{c}_j^\circ(t) = \begin{cases} 1, \text{ если } \bar{\mathbf{c}}_j^{t:t-t_0} > 0, \\ 0, \text{ иначе,} \end{cases} \qquad \text{где } \bar{\mathbf{c}}^{t:t-t_0,j} = \frac{1}{t_0}\sum_{\tau=0}^{t_0-1} c_j(t-\tau).$$

в среднем по последним $t_0 = 10$ отсчетам. Для учета лагирования в модели (3), набор временных рядов $\{\mathbf{c}_j, \mathring{\mathbf{c}}_j\}$ был расширен добавлением лагированных временных рядов

$$\mathbf{c}_j^{1:T-l\tau},\ \mathring{\mathbf{c}}_j^{1:T-l\tau},\ j = 1,\dots n,\ \text{и } \mathbf{x}^{1:T-l\tau-1},$$

для $l = 0,\dots,L-1$, где L — максимальный порядок лагирования. Значения параметров были выбраны экспериментально. Ниже приводятся результаты экспериментов с $\alpha = 0.07$, максимальным числом компонент $n_{\max} = 5$ и максимальным порядком лагирования $L = 3$.

Обозначим через $\hat{x}_0$ и $\hat{x}_{Ex}$ прогнозы (1) исходного и уточненного алгоритмов *hist*, соответственно. Для оценки качества уточненного алгоритма было использовано изменение потерь

$$\Delta L = L_0 - L_{Ex}, \quad L\big(x,\hat{x}\big) = (x-\hat{x})^2.$$

С помощью алгоритма SEM для каждого временного ряда $\mathbf{x}$ были выбраны экзогенные временные ряды $\mathbf{c}_j, j \in \mathcal{J}$ и настроены веса w_j соответствующих компонент смеси. Затем в 50 контрольных точках $x(t)$ была построена уточненная гистограмма $H(t)$ на основе $\mathbf{x}^{1:t-1}$ и $\mathbf{c}_j^{1:t}, j \in \mathcal{J}$ согласно (3), вычислены прогноз (1) и изменение потерь $\Delta L(t)$, связанное с уточнением гистограммы $H(t)$. Таким образом была получена выборка $\Delta L(t)$. На основе выборки $\{\Delta L(t)\}$ с помощью критерия Стьюдента тестировалась гипотеза о равенстве нулю ожидаемых потерь $\mathsf{E}(\Delta L)$ при альтернативе $\mathsf{E}(\Delta L) > 0$.

В первых двух столбцах таблицы 2 перечислены типы грузов и экзогенные факторы, выбранные для каждого из них. Кружком помечены производные временные ряды, число перед символом τ равно порядку лагирования. Величина лагирования τ была выбрана равной одной неделе $\tau = 7$. В остальных столбцах приведены количественные показатели изменения качества прогнозирования: среднее изменение потерь ΔL, среднее относительное изменение потерь $\Delta L/L_0$, минимальный уровень значимости p-value при проверке гипотезы $H_0 : \mathsf{E}(\Delta L) = 0$ при альтернативе $H_1 : \mathsf{E}(\Delta L) > 0$ с помощью одностороннего t-критерия Стьюдента. В последних двух столбцах приводится доля контрольной выборки, в которой качество прогнозирования строго улучшилось и строго ухудшилось, соответственно. Меньшее значения p-value соответствуют большему уровню значимости решения об улучшении качества прогнозирования. В таблице перечислены только те временные ряды, положительное изменение потерь для которых было подтверждено с p-value не менее

0.1. Заметим, что сумма долей отрицательного и положительного изменения потерь $\Delta L(t)$ в большинстве случаев далека от единицы: для большинства рассмотренных временных рядов примерно в половине контрольных точек качество прогноза не изменилось.

Таблица 2: Результаты проверки увеличения качества прогнозирования на статистическую достоверность.

Группа грузов	$\mathbf{c}_j$	ΔL	$\Delta L/L_0$	p-value	$\lvert\{t : L_0(t) > L_{Ex}(t)\}\rvert$	$\lvert\{t : L_0(t) < L_{Ex}(t)\}$
Нефть и нефтепродукты	$\mathring{C}$винец, 2τ	0.15311	0.27873	0.0002305	0.31373	0.039216
Черные металлы	Свинец, τ	0.11195	0.29659	0.00099359	0.17647	0
Металлические конструкции	Свинец, τ	0.066946	0.1044	0.039325	0.21569	0.058824
Метизы	Свинец, τ	0.2046	0.43858	0.00022156	0.64706	0.2549
Хмикаты и сода	$\mathring{C}$винец, 2τ	0.26036	0.41155	4.9025e-08	0.47059	0.019608
Строительные грузы	Свинец, τ	0.14043	0.34462	0.00011159	0.35294	0.098039
Шлаки гранулированные	Свинец, τ	0.062775	0.17258	0.03783	0.15686	0.039216
Огнеупоры	$\mathring{C}$винец, 2τ	0.035985	0.086756	0.099958	0.039216	0
Цемент	Свинец, τ	0.25782	0.31271	0.00010496	0.5098	0.058824
Рыба	$\mathring{C}$винец, 2τ	0.25241	0.34342	1.0663e-06	0.43137	0.019608
Зерно	Свинец, τ	0.19695	0.29581	5.3108e-05	0.39216	0.13725

6. Заключение. Описан метод учета экзогенных временных рядов в гистограммной модели прогнозирования *hist*. Предложенный метод основан на уточнении гистограммы прогнозируемого временного ряда с помощью смеси условных гистограмм. Было проведено экспериментальное сравнение расширенной и базовой версии алгоритма *hist*. Продемонстрировано уменьшение потерь для некоторых временных рядов. Дальнейшее исследования планируется проводить в следующих направлениях. Во-первых, можно рассмотреть усовершенствованные методы оценки плотностей распределения, например ядерные оценки плотности, позволяющие учесть информацию о некотором измерении при оценке плотности более чем в одной точке. Применение ядерных оценок позволит сократить разреженность снизить требования к минимальной длине временного ряда, позволив улучшить качество расширенного алгоритма *hist*. Также возможно расширение предлагаемого подхода путем введения временной зависимости для w_j и рассмотрения большего числа производных временных рядов. В данной работе эти подходы не были рассмотрены в связи с относительно небольшим объемом данных.

Альтернативой предлагаемому способу расширения алгоритма *hist* являются метода многомерной квантильной регрессии. Хотя задаче разработки методов многомерной квантильной регрессии посвящен целый ряд работ (например, [12, 20, 13]), каждое из предлагаемых решений сфокусировано на сохранении определенных характеристик одномерной

квантильной регрессии. Так как алгоритм *hist* является модификацией квантильной регрессии, создание многомерной модели предоставило бы естественный способ включения экзогенных факторов в прогностическую модель.

СПИСОК ЛИТЕРАТУРЫ

1. Aitnouri E., Wang S., Ziou D. et al. Estimation of multi-modal histogram's pdf using a mixture model // Neural, Perallel and Scientific Computations. 1999. N. 7. P. 103–118.
2. Arroyo J., Gonzalez-Rivera G., Maté C. et al. Smoothing methods for histogram-valued time series: an application to value at risk // Statistical Analysis and Data Mining. 2011. **4**. N. 2. P. 216–228.
3. Bartolucci F. and Farcomeni A. A note on the mixture transition distribution and hidden markov models // J. Time Ser. Anal. 2010. **31**. N. 2. P. 132–138.
4. Berchtold A. Mixture transition distribution (mtd) modeling of heteroscedastic time series // Computational Statistics & Data Analysis. 2003. N. 41. P. 399–411.
5. Berchtold A. and Raftery A. E. The mixture transition distribution model for high-order markov chains and non-gaussian time series // Stat. Sci. 2002. **17**. N. 3. P. 328–356.
6. Bishop Y. M. M., Fienberg S. E., and Holland P. W. Discrete Multivariate Analysis: Theory and Practice, Cambridge: M.I.T. Press, 1976.
7. Deming W. E. and Stephan F. F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known // Ann. Math. Stat. 1940. **11**. N. 4. P. 427–444.
8. Diebolt J. and Ip E. H. S. Stochastic EM: method and application // Markov Chain Monte Carlo in Practice (chapter 15), London: Chapman and Hall, 1996. P. 259–273.
9. Everingham M. and Thomas B. Supervised segmentation and tracking of nonrigid objects using a "mixture of histograms" model // IEEE Image Proc. 2001. **1**. P. 62–65.
10. De Gooijer J. G. and Hyndman R. J. 25 years of time series forecasting // Int. J. Forecasting. 2006. N. 22. P. 443–473.
11. Koenker R. and Bassett G. J. Regression quantiles // Econometrica. 1978. N. 46. P. 33–50.
12. Koltchinskii V. M-estimation, convexity and quantiles // Ann. Stat./ 1997. **25**. N. 2. P. 435–477.
13. Kong L. and Mizera I. Quantile tomography: using quantiles with multivariate data // Stat. Sinica. 2012. N. 22. P. 1589–1610.

14. Peña D. and Sánchez I. Multifold predictive validation in armax time series models // J. Am. Stat. Assoc. 2005. N. 100. P. 135–146.

15. Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition // P. IEEE. 1989. **77**. N. 2. P. 257–286.

16. Raftery A. E. A model for high-order markov chains // J. Roy. Stat. Soc. B Met. 1985. **47**. N. 3. P. 528–539.

17. Roopaei M., Zolghadri M. and Emadi A. Economical forecasting by exogenous variables // IEEE Int. Conf. Fuzzy. 2008. P. 1491–1495.

18. Schopf J. M. A practical methodology for defining histograms for predictions and scheduling // Proceedings of the International Conference ParCo99, Imperial College Press, 2000. P. 664–671

19. Scott D. W. On optimal and data-based histograms // Biometrika. 1979. **66**. N. 3. P. 605–610.

20. Serfling R. Quantile functions for multivariate analysis: Approaches and applications // Stat. Neerl. 2002. N. 56. P. 214–232.

21. Spurrier, J. D. On the null distribution of the Kruskal–Wallis statistic // J. Nonparametr. Statist. 2003. **15**. N. 6. P. 685–691.

22. Syrovátka P. and Grega L. Analysis of methodological approaches to evaluation of complementary and substitution relationship in consumer demand for food // Agr. Econ. 2002. **48**. N. 10. P. 456–462.

23. Вальков А. С., Кожанов Е. М., Медведникова М. М. и др. Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным // Машинное обучение и анализ данных. 2012. **1**. № 4. С. 448-465.

24. Whittle P. Prediction and Regulation by Linear Least-Square Methods, University of Minnesota Press, 1983.

25. Wong C. S. and Li W. K. On a mixture autoregressive conditional heteroscedastic model// J. Am. Stat. Assoc. 2001. N. 96. P. 982–995.