Методика определения оптимального объема выборки для прогнозирования нестационарного временного ряда

Ю.Н. Орлов, К.П. Осминин

Аннотация. Предложен метод статистического анализа и прогнозирования нестационарных временных рядов, основанный на введении критерия квазистационарности выборочной функции распределения. Построена прогнозная модель для конкретного временного ряда с вложенной циклической структурой данных, содержащая алгоритм определения оптимального объема выборки и основанная на решении уравнения эволюции функции распределения с дискретным шагом по времени. Проведено сравнение результатов прогнозирования данным методом с некоторыми стандартными моделями – регрессионной, автокорреляционной и скользящей средней.

Введение и постановка задачи

В работе предлагается подход к статистическому анализу и прогнозированию нестационарных временных рядов, основанный на использовании уравнения эволюции выборочной функции распределения (далее ВФР). Для этого вводятся понятие ε-стационарности выборки и функционал, играющий роль соответствующего критерия. Определяются квазистационарные временные ряды и предлагается метод, позволяющий оптимизировать объем выборки для функции распределения с точки зрения минимизации ошибки прогноза. Соответствующий оптимальный объем выборки зависит как от требуемой точности прогноза, так и от горизонта прогнозирования.

В качестве примера в работе рассматривается временной ряд, имеющий сильно выраженную циклическую составляющую: это ряд, образованный из индексов почасовых цен на электрическую энергию на оптовом рынке электроэнергии (мощности). Соответствующие данные содержатся в [1]. Прогноз цен на электроэнергию в заданном регионе или по энергосистеме имеет значение для оптимизации стратегии управления и хеджирования рисков реально функционирующих производственных систем как производителями электроэнергии, так и ее потребителями. Таким образом, интерес к этому объекту вызван и его практической важностью, и характерными свойствами данного временного ряда, требующими разработки специфических методов анализа.

Спрос на электроэнергию имеет весьма высокую неслучайную составляющую, определяемую технологическими процессами выработки, суточной и недельной периодичностью спроса. Кроме того, спрос является сильно неэластичным (нелинейно зависит от уровня цен). На него оказывают влияние такие внешние факторы, как суточная температура, сезонность, структура потребителей по секторам хозяйственной системы региона и др. На предложение влияют такие факторы, как гидрологическая ситуация, режимы работы ГЭС и ТЭЦ, графики ремонтных работ. Таким образом, для анализа такого ряда надо уметь решать задачу устранения влияния детерминированных факторов, т.к. иначе может оказаться, что вариация ряда (относительное среднеквадратичное отклонение) слишком велика с точки зрения желаемой точности прогнозирования.

В данной работе мы не будем заниматься анализом влияния вышеуказанных факторов на спрос и динамику цен, а сосредоточимся на задаче собственно прогнозирования нестационарного временного ряда, полученного из исходного после исключения его детерминированной составляющей. Предположим, что с приемлемой точностью оставшийся ряд представляет собой реализацию некоторого регулярного случайного процесса. В частности, для рассматриваемого ряда его детерминированная составляющая в значительной мере определяется так называемой назначенной периодичностью, состоящей в суточной и недельной периодичности спроса на электроэнергию и, как следствие, в периодичности изменения цен на нее. Эта периодичность в большей части может быть исключена переходом к ряду из остатков, полученных вычитанием из исходного ряда некоторого среднего графика цен. Поскольку режим потребления электроэнергии в выходные и рабочие дни существенно разный, то для повышения точности анализа эти дни было бы желательно рассматривать отдельно. Однако в этом случае могут появиться статистические артефакты, связанные с сортировкой данных. Поэтому мы в качестве характерного цикла выберем неделю и будем рассматривать так называемый «ряд нормированных остатков», получаемых после вычитания из исходного ряда среднего недельного графика, т.е. почасовых данных за среднюю неделю, и деления результата на среднее значение цены на рассматриваемом интервале времени. Как показано в [2], эти остатки уже слабо коррелируют между собой. Ряд из указанных остатков, представленный на Рис. 1, и будет являться основным объектом нашего исследования. Существенно, что практически полное исключение суточной автокорреляции между элементами исходного временного ряда путем перехода к нормированным остаткам не делает новый ряд стационарным, а всего лишь снижает размах выборки.

Существует большое количество моделей и приемов исследования временных рядов [3]. Подавляющее их число относится к стационарным в широком смысле рядам, т.е. к рядам для которых, первые два или, возможно, большее, но конечное число моментов не зависят от времени. Если ряд нестационарный, то его часто пытаются свести к стационарному путем перехода к первым (или вторым и т.д.) разностям или используют модели со скользящими параметрами. Если бы ряд из остатков или их первых (или более высоких) разностей оказался стационарным, то можно было бы, используя теорему Гливенко и критерий согласия Колмогорова [4], попытаться определить вид распределения, к которому относилась бы изучаемая выборка данных, после чего с известной доверительной вероятностью строить прогноз.

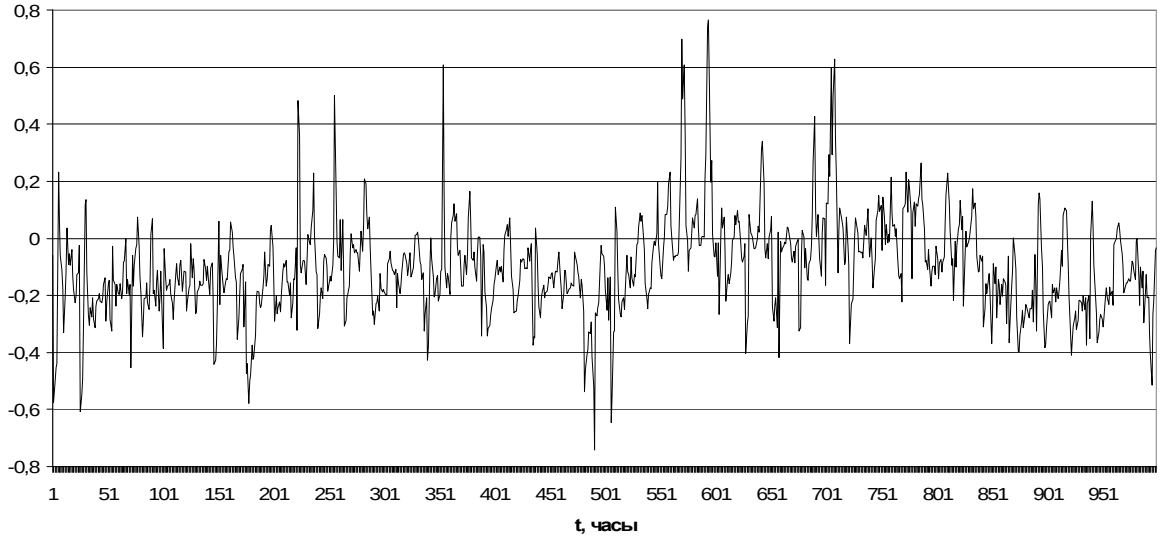

Рис. 1. Фрагмент исследуемого временного ряда, построенного по данным [1, 2]

Отличие от модельных ситуаций для рассматриваемого нами случая в том, что разностным дифференцированием ряда остатков не удается исключить тренд в дисперсии и других моментах. Существенно также, что для скользящего усреднения нужно все же выбрать некоторый объем данных, которые, по предположению, принадлежат одной генеральной совокупности. Но если тренд в средних величинах большой, это условие может не быть выполнено с требуемой точностью.

Желание иметь дело со стационарным рядом вызвано также возможностью обосновать прогнозные модели для такого ряда применением теоремы Вольда о разложении ([5], стр. 269), согласно которой всякий стационарный процесс может быть единственным образом представлен в виде суммы двух некоррелированных между собой процессов: детерминированного (сингулярного процесса), прогноз которого на любое время вперед безошибочен, и чисто случайного (регулярного белого шума). Поэтому, хотя реальные процессы рассматриваемого нами происхождения, как правило, не являются стационарными, тем не менее возникает желание в первом приближении считать их таковыми. Такой подход может дать удовлетворительный результат в задачах краткосрочного прогнозирования.

Итак, цель данного исследования состоит в разработке метода, который направлен на решение следующих задач: определение объема выборки из нестационарного ряда для формирования квазистационарной ВФР; минимизация ошибки аппроксимации временного ряда на заданном интервале времени; определение интервала времени, на котором ошибка аппроксимации не превосходит заданную величину; прогнозирование временного ряда внутри и вне границ квазистационарности ВФР.

1. Критерий ε-стационарности и согласованность объема выборки с горизонтом прогноза

Введем следующие обозначения: $x(t)$ – значение случайной величины x в дискретный момент времени t; $\Delta(t_1, t_2)$ – промежуток времени $[t_1, t_2]$, по которому набирается статистика ряда $x(t)$; $T = t_2 - t_1$ – объем соответствующей выборки; $f_T(x,t)$ – выборочная функция распределения, построенная по выборке $\Delta(t_1, t_2)$ объема T в скользящий момент времени $t_2 = t$. Среднее значение величины $g(x)$ по ВФР $f_T(x,t)$ обозначим

$$\langle g(x)\rangle_{T,t} = \int g(x) f_T(x,t)dx .$$

Определение 1. ВФР $f_T(x,t)$ временного ряда $x(t)$ будем называть ε-стационарной на промежутке Δ, если

$$\forall\, t, \tau \ge 0 \quad \int_{-\infty}^{\infty} \left| f_T(x, t+\tau) - f_T(x,t) \right| dx \le \varepsilon . \quad (1)$$

Это определение записано в виде интегрального критерия с бесконечными пределами лишь для математического удобства, чтобы иметь возможность рассматривать любые значения временного ряда $x(t)$, хотя на практике они на каждом промежутке Δ ограничены. Количество членов ряда, участвующих в формировании ВФР, равно объему T выборки, поэтому ВФР представляет собой гистограмму с некоторым шагом h, высота которой на интервале $(x, x+h)$ равна отношению числа членов ряда $x(t), t \in \Delta$, попавших в этот интервал, к объему выборки T. Функционал

$$V_T(t,\tau) = \int_{-\infty}^{\infty} \left| f_T(x, t+\tau) - f_T(x,t) \right| dx \quad (2)$$

совместно с определением 1 играет роль критерия ε-стационарности распределения случайной величины x. Он показывает, насколько ВФР с одинаковым объемом статистической базы отличаются одна от другой в разные моменты времени.

Из неотрицательности ВФР и ее нормированности на единицу в любой момент времени следует оценка

$$0 \le V_T(t,\tau) \le \min(2\tau / T; 2). \quad (3)$$

Действительно, количество новых членов временного ряда при сдвиге на τ шагов равно, очевидно, τ, и столько же старых членов оказываются исключенными из выборки для ВФР

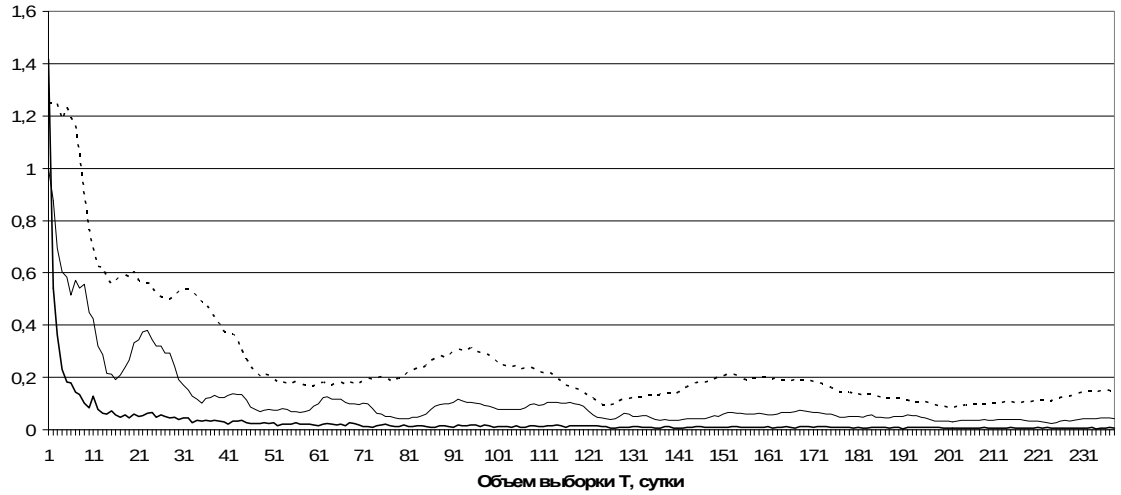

Рис. 2. Зависимость функционала $V_T(t,\tau)$ от объема выборки *T* в некоторый фиксированный момент времени *t* для трех значений параметра сдвига (снизу вверх): τ = 1, 7 и 28 дней

в момент времени $t+\tau$, поэтому суммарное изменение числа различных членов ряда при сдвиге окна выборки на τ шагов не превосходит 2τ. Если же функционал $V_T(t,\tau)$ рассматривается при $\tau \geq T$, то из (2) следует оценка $0 \leq V_T(t,\tau) \leq 2$.

Оценка (3), несмотря на свою тривиальность, дает возможность сделать важный вывод о том, что при фиксированном τ функционал (2) равномерно ограничен по *t*. Поэтому $\forall\, \varepsilon > 0$ всегда можно подобрать такой объем выборки $T > 2\tau/\varepsilon$, что ВФР будет ε-стационарной. Вид функционала $V_T(t,\tau)$ для ряда из почасовых данных [1] приведен на Рис. 2.

При увеличении точности в определении ε-стационарности, т.е. при уменьшении ε, объем выборки, при которой достигается условие $V_T(t,\tau) \leq \varepsilon$, растет. В силу равномерной ограниченности (3) функционала $V_T(t,\tau)$ для каждого момента времени *t* существуют такое минимальное значение $T_0(t,\tau;\varepsilon)$ и максимальное $\tau_0(t,T;\varepsilon)$, что при всех $T \geq T_0$ и при всех $\tau \leq \tau_0$ значения функционала $V_T(t,\tau)$ не будут превосходить ε.

Рассмотрим функции

$$\theta(T;\varepsilon) = \min_t \tau_0(t,T;\varepsilon),$$
$$T(\tau;\varepsilon) = \max_t T_0(t,\tau;\varepsilon). \qquad (4)$$

Первая из них представляет эмпирическую оценку максимального горизонта прогноза, внутри которого распределение остается ε-стационарным, а вторая определяет минимальный объем необходимой для этой цели выборки. Подчеркнем, что T и τ в (4) не любые, а удовлетворяющие условиям $T \geq T_0$ и $\tau \leq \tau_0$. Фигурирующие в этих неравенствах величины $T_0(t,\tau;\varepsilon)$ и $\tau_0(t,T;\varepsilon)$ отмечают граничные значения объемов выборок и допустимых сдвигов, т.е. являются локальными индикаторами некоторых статистических свойств ряда $x(t)$.

Из (3) следует, что в качестве максимальной оценки минимально допустимого объема выборки $T(\tau;\varepsilon)$ можно взять $2\tau/\varepsilon$. Эта оценка может быть затем уточнена (т.е. уменьшена) путем исследования статистических свойств конкретного ряда $x(t)$. Может оказаться, что распределение величин $T_0(t,\tau;\varepsilon)$, представляющих при заданном τ самостоятельный временной ряд, имеет выборочное среднее по времени *t* значение $\langle T_0(t,\tau;\varepsilon)\rangle$, значимо меньшее этой равномерной по *t* оценки. В этом случае квадратный корень из относительной выборочной дисперсии этого распределения даст оценку точности, с которой вместо (4) в качестве оптимального объема выборки можно использовать не наибольшее из наименьших значений $T_0(t,\tau;\varepsilon)$, а, например, среднее $\langle T_0(t,\tau;\varepsilon)\rangle$.

В частности, для рассматриваемого нами ряда $x(t)$ объем выборки, при котором распределение 0,01-стационарно при сдвиге на 1 сутки (24 часовых значения), составляет в среднем 175 суток. При том же сдвиге на 1 сутки минимальный объем выборки для 0,05-стационарности составляет в среднем 30 суток, а для 0,1-стационарности – 10 суток. Максимальный горизонт прогноза растет приблизительно линейно с увеличением ε и составляет: 1 сутки для $\varepsilon = 0{,}01$, 7 суток для $\varepsilon = 0{,}05$ и 15 суток для $\varepsilon = 0{,}1$.

Функция распределения величин $T_0(t, 1; 0{,}05)$ при сдвиге на 1 сутки и соблюдении требования 0,05-стационарности ВФР приведена на Рис. 3. Среднеквадратичное отклонение распределения минимальных объемов выборок составило $\sigma(\varepsilon) =$ 4 дня, т.е. вариация ряда из величин $T_0(t, 1; 0{,}05)$ составила 13%. Заметим, что наиболее вероятное значение (28 суток) не совпадает со средним значением по данному распределению (30 суток), которое, напротив, относительно маловероятно. Из (3) следует, что при $T >$ 40 суток распределение величин x заведомо 0,05-стационарно.

Введем функционалы

$$U_T(\tau) = \sup_t V_T(t,\tau), \quad E(T) = \sup_\tau U_T(\tau), \quad \eta(T) = \inf_{\tau \ge 1} U_T(\tau) \qquad (5)$$

Очевидно, полагая $\varepsilon = E(T)$, любую ВФР с объемом выборки T можно считать ε-стационарной. В силу оценки (3) ясно, что $E(T)$ уменьшается с увеличением T, но не обязательно монотонно.

Определение 2. ВФР будем называть асимптотически ε-стационарной, если

$$\exists\, T_\infty \ \forall\, \Delta : T_\Delta > T_\infty \quad V_{T_\Delta}(t, T_\Delta) \le \varepsilon. \qquad (6)$$

Если же существует предел $\lim_{T\to\infty} V_T(t,T) = 0$, *то ВФР будем называть просто асимптотически стационарной.*

Смысл определения 2 в том, что изменение асимптотически ε-стационарной ВФР при сдвиге на интервал, равный объему статистической базы, не превосходит ε.

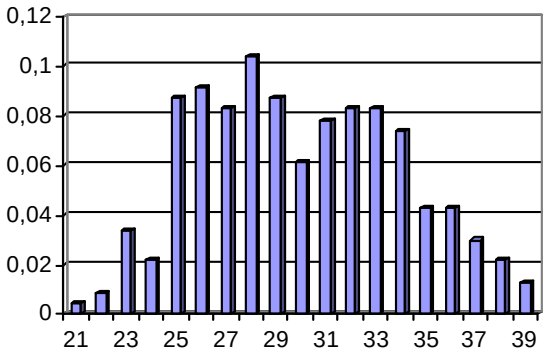

Рис. 3. Функция распределения минимального объема выборки для обеспечения 0,05-стационарности распределения значений исходного ряда при сдвиге на τ = 1 сутки

Пусть ε_0 – некоторый априори заданный желательный уровень стационарности ВФР. Тогда, если $\eta(T) \le \varepsilon_0$, существуют такие величины сдвигов τ, при которых $U_T(\tau) \le \varepsilon$, где $\eta(T) \le \varepsilon \le \varepsilon_0$. Заметим, что в силу оценки (3) $\eta(T) \le 2/T$, поэтому, увеличивая объем выборки, всегда можно добиться выполнения условия (6). Однако может оказаться, что ВФР, построенная по большей выборке, имеет и большую дисперсию, что будет препятствовать построению более точного прогноза. Таким образом, возникает проблема оптимизации объема выборки с тем, чтобы при заданном горизонте прогноза τ суммарная ошибка прогноза вследствие статистического разброса и временного тренда была минимальной. Обозначим через

$$\sigma_\Delta^2(t) = \left\langle \left(x - \langle x \rangle_{\Delta,t} \right)^2 \right\rangle_{\Delta,t} \qquad (7)$$

скользящую дисперсию распределения $f_T(x,t)$. Предположим, что рассматриваемый временной ряд $x(t)$ является стационарным и $f(x)$ есть его теоретическое распределение, для которого существуют первый μ_1 и центральные высшие моменты m_k. Известно ([5] стр. 540, формула 4.29), что тогда дисперсия смещенной выборочной дисперсии (7) равна

$$D\left(\sigma_\Delta^2\right) = \frac{m_4 - m_2^2}{T} - 2\frac{m_4 - m_2^2}{T^2} + \frac{m_4 - 3m_2^2}{T^3}. \qquad (8)$$

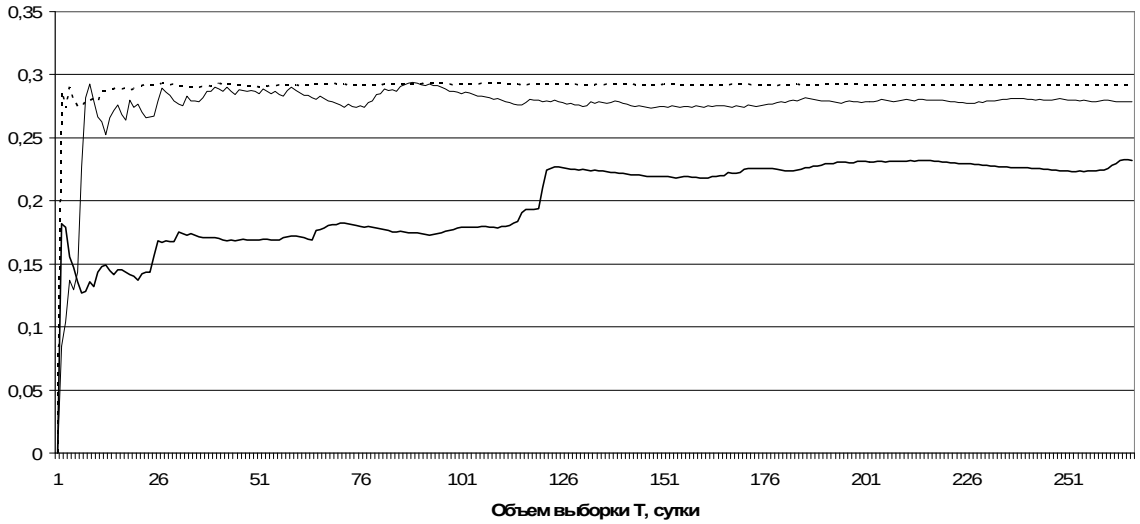

Рис. 4. Снизу вверх: зависимость среднеквадратичного отклонения от объема выборки для исследуемого нестационарного ряда (жирная линия), типичного стационарного ряда (тонкая линия) и ряда, образованного значениями $T_0(t, 1; 0{,}05)$ (точечная линия)

Известно также (теорема Гофдинга [6] стр. 215), что величина $\xi = \dfrac{\sqrt{T}\left(\sigma_\Delta^2 - m_2\right)}{\sqrt{D\left(\sigma_\Delta^2\right)}}$ имеет при $T \to \infty$ (независимо от смещенности или несмещенности выборочной дисперсии) асимптотически нормальное распределение с нулевым средним и единичной дисперсией, т.е. плотность ее распределения имеет вид $f_{norm}(\xi) = \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{\xi^2}{2}\right)$. Таким образом, вероятность того, что величина ξ не превосходит по модулю некоторое значение ξ_0, равна $\Phi(\xi_0) = 2\int_0^{\xi_0} f_{norm}(\xi)d\xi$. Этот факт лежит в основе многих тестов на нормальность генеральной совокупности, а также на стационарность или нестационарность временных рядов [3].

На практике для количественных оценок вероятности отклонения нормально распределенных величин используются таблицы α-квантилей распределения, т.е. таких величин q_α, что

$$\int_{-\infty}^{q_\alpha} f_{norm}(\xi)d\xi = \alpha = \frac{\Phi(q_\alpha)+1}{2}.$$

Из последнего равенства следует, что если задать доверительную вероятность вышеуказанного отклонения $\Phi(\xi_0) = \gamma$, то соответствующее значение ξ_0 будет равно $\dfrac{\gamma+1}{2}$-квантилю стандартного нормального распределения. В частности, для $\alpha = 0{,}95$ имеем по таблицам [5] $q_{0{,}975} \approx 1{,}97$. Это означает, что вероятность отклонения выборочной дисперсии стационарного ряда от теоретического значения m_2 на абсолютную величину большую, чем $q_{0{,}975}\sqrt{D\left(\sigma_\Delta^2\right)/T}$, не превосходит 0,05. Рассмотрим с указанных позиций выборочную дисперсию (7).

Типичная зависимость σ_Δ от объема T выборки Δ в некоторый начальный момент времени t для исследуемого нами нестационарного временного ряда показана на Рис. 4. Там же приведена аналогичная зависимость для среднеквадратичного отклонения некоторого стационарного ряда, полученного генератором псевдослучайных чисел, а также и для ряда из величин $T_0(t, 1; 0{,}05)$, определенных выше.

Как следует из формулы (8) и последующих оценок, вероятность того, что при больших T амплитуда колебаний дисперсии стационарного ряда будет больше, чем корень из дисперсии

выборочной дисперсии, весьма мала, поскольку амплитуда колебаний выборочной дисперсии убывает с ростом объема выборки как $1/\sqrt{T}$. Предположим, что выборочные дисперсии приближенно стабилизируются к соответствующим теоретическим значениям m_2, т.е. среднее значение выборочного среднеквадратичного отклонения временного ряда при $T > 200$ (Рис. 4) является приемлемой оценкой величины $\sqrt{m_2}$. Для исследуемого нами ряда $x(t)$ отклонение выборочной дисперсии от величины $q_{0,975}\sqrt{D\left(\sigma_\Delta^2\right)/T}$ при $T < 30$ дней колеблется от 10% до 25%, т.е. достаточно велико, тогда как для двух других рядов оно не превышает 0,5%. В то же время при $T > 200$ дней для всех трех рядов указанное отклонение составляет менее 0,02%.

Таким образом, по сравнению с рядом $x(t)$ можно считать, что ряд $T_0(t, 1; 0{,}05)$ является квазистационарным, по крайней мере в широком смысле. Это является важным наблюдением, поскольку оно может служить основанием для использования среднего по t значения $\langle T_0(t, 1; 0{,}05)\rangle$ в качестве оптимального объема выборки для формирования ε-стационарной ВФР.

Изменение с течением времени статистических свойств ряда $T_0(t, \tau; \varepsilon)$ может служить индикатором того, что реальная сложно функционирующая система (рынок, энергетический комплекс и т.п.), порождающая ряд $x(t)$, качественно меняется. Будем называть распределение величин $T_0(t, \tau; \varepsilon)$, представленное на Рис. 3, индикатором оптимального объема выборки.

Определение 3. Временной ряд $x(t)$, для которого индикатор оптимального объема выборки является квазистационарным временным рядом, будем называть рядом со стационарным индикатором.

Развиваемая далее теория может применяться к различным временным рядам, но корректно обоснованной она является именно для рядов со стационарным индикатором, поскольку при прогнозировании распределения на любой промежуток, как превышающий границы квазистационарности, так и находящийся внутри них, необходимо быть уверенным в том, что объем выборки сохраняет свойство своей оптимальности. Это означает, что оптимальный объем выборки $T(\tau; \varepsilon)$, определяемый по данным до начала прогнозирования, в последующем не будет иметь статистически значимого тренда в течение промежутка времени τ.

Как показывает график 1 на Рис. 4, при анализе временного ряда может возникнуть следующая ситуация: ряд, являющийся нестационарным в широком смысле на одном промежутке времени (например, при $T < 30$ дней), оказывается стационарным на другом, большем промежутке (при $T > 200$ дней), но и оценка его дисперсии на втором промежутке больше, чем на первом. При этом ВФР рассматриваемого ряда на обоих промежутках является ε-стационарной при сдвиге на некоторый фиксированный шаг τ. Следовательно, слишком большой объем выборки может не быть оптимальным для задачи прогнозирования на промежуток времени τ. Поэтому среди всех допустимых объемов выборок $T \geq T(\tau; \varepsilon)$, где $T(\tau; \varepsilon)$ мы определим как $\langle T_0(t, \tau; \varepsilon)\rangle$ (что можно сделать в силу квазистационарности ряда $T_0(t, \tau; \varepsilon)$), естественно выбрать такой объем, при котором выборочная дисперсия минимальна. Из Рис. 4 видно, что для рассматриваемого нами ряда $x(t)$ наилучшим будет вариант, при котором $T(\tau; \varepsilon) \in [10, 30]$. В частности, в этот промежуток попадает значение объема выборки, отвечающей $\varepsilon = 0{,}05$ в определении стационарности распределения при прогнозе на $\tau = 1$ сутки.

Описанная методика определения оптимального объема выборки для нестационарного временного ряда является основой для его прогнозирования на заданный промежуток времени по базе, объем которой согласован как с требованием ε-стационарности, так и с горизонтом прогноза. Поскольку выбирается тот объем выборки, при котором для данных значений ε и τ выборочная дисперсия минимальна, то прогноз, основанный на статистических свойствах этой выборки, будет иметь наименьшую ошибку по сравнению с таким же прогнозом (т.е. с прогнозом, полу-

чаемым по одинаковой методике конкретного построения будущих значений временного ряда), но использующим анализ данных по выборке другого объема.

2. Моделирование эволюции нестационарной ВФР

Определение 4. Прогнозом временного ряда на шаг τ будем называть ε-стационарную ВФР, построенную к моменту времени $t+\tau$ на основе данных выборки $\Delta(t-T,t)=[t-T,\,t]\equiv\Delta(t)$, известных к моменту времени t.

Потребуем, чтобы объем выборки и горизонт прогноза удовлетворяли описанным выше условиям оптимального согласования величин, определенных в (4). В этом случае мы прогнозируем ВФР внутри границ ее ε-стационарности.

Можно поставить вопрос и о прогнозе на промежутках времени, превышающих $\theta(T;\varepsilon)$. В этом случае на промежутках вида $[\theta, n\theta]$ ВФР не будет ε-стационарной, хотя на каждом промежутке $[(n-1)\theta, n\theta]$ она ε-стационарна. Тогда требуется алгоритм, по которому следует переходить от ε-стационарной ВФР, отнесенной к одному промежутку, к ВФР в другом промежутке. Как и для прогноза внутри промежутка ε-стационарности, такой алгоритм фактически представляет собой решение соответствующего дискретного уравнения Лиувилля, построенного с помощью текущих выборочных данных.

Обозначим $f(x,t;\tau)$ ВФР, полученную по выборке из скользящего окна $\Delta(t+\tau)=[t+\tau-T,\,t+\tau]$, т.е. в обозначениях п. 2 $f(x,t;\tau)=f_T(x,t+\tau)$. Введем также двумерную ВФР $F(x,\dot{x},t;\tau)$, случайных величин x и $\dot{x}$ исходного ряда и ряда его производной, полученного взятием разностей $x(t)-x(t-1)$ в соседние моменты времени. «Силовой член» в уравнении Лиувилля, зависящий от $\ddot{x}$, будем моделировать на основе выборочных данных, например, с помощью регрессионной зависимости $\ddot{x}$ как функции от x: $\ddot{x}=\varphi(x)$. Тогда уравнение Лиувилля эволюции введенной двумерной ВФР автоматически сохраняет ее нормировку и имеет вид

$$\frac{\partial F(x,\dot{x},t;\tau)}{\partial\tau}+\dot{x}\frac{\partial F(x,\dot{x},t;\tau)}{\partial x}+\varphi(x)\frac{\partial F(x,\dot{x},t;\tau)}{\partial\dot{x}}=0. \tag{9}$$

Формально, чтобы записать уравнение Лиувилля в дифференциальной форме (9), надо потребовать от $F(x,\dot{x},t;\tau)$ равномерной сходимости относительно τ в измеримой по Жордану области изменения аргументов x и $\dot{x}$, причем F и ее частные производные должны быть непрерывны по всей совокупности переменных $x,\dot{x},\tau$. Разумеется, для ВФР в виде ступенчатых функций последнее условие не выполняется, но мы используем здесь дифференциальную и интегральную формы записи не по существу, а из соображений удобства, подразумевая конечно-разностный аналог производных и конечную сумму вместо интеграла. Далее для краткости опускаем аргументы t и τ, если это не будет сказываться на понимании смысла уравнений.

Так как одномерное распределение получается из двумерного интегрированием (суммированием) по второму аргументу, т.е. $f(x)=\int F(x,\dot{x})d\dot{x}$, то, вводя функцию $u(x)$ по формуле

$$u(x)f(x)=\int \dot{x}F(x,\dot{x})d\dot{x}, \tag{10}$$

получаем из (9) уравнение неразрывности потока плотности вероятности:

$$\frac{\partial f}{\partial\tau}+\frac{\partial(uf)}{\partial x}=0. \tag{11}$$

Эволюция моментов одномерного распределения

$$\mu=\int xf(x)dx\ ,\ \sigma^2=\int(x-\mu)^2f(x)dx \tag{12}$$

в силу (11) имеет вид:

$$\frac{\partial\mu}{\partial\tau}=\bar{u},\quad \frac{\partial\sigma^2}{\partial\tau}=2\ (x-\mu)(u-\bar{u})f(x)dx,\qquad \bar{u}=\ u(x)f(x)dx \tag{13}$$

В частности, рассмотрим пример, в котором $\varphi(x)$ моделируется линейной регрессией на x: $\varphi(x)=ax+b$:

$$\ddot{x}=ax+b,\quad a=\langle\Delta x\Delta\ddot{x}\rangle_{\Delta(t)}/\langle(\Delta x)^2\rangle_{\Delta(t)},\qquad b=\langle\ddot{x}\rangle_{\Delta(t)}-a\langle x\rangle_{\Delta(t)} \tag{14}$$

Здесь употреблены обозначения средних по выборке $\Delta(t)$ функций времени:

$$\langle g\rangle_{\Delta(t)} = \frac{1}{T}\sum_{\tau=1}^{T} g(t-T+\tau),$$
$$\Delta g = g(t-T+\tau) - \langle g\rangle_{\Delta(t)}$$

Коэффициенты a и b в (14) являются функциями параметра t (это время начала прогноза). Для линейной модели (14) уравнение (9) представляет эволюцию ансамбля гармонических осцилляторов, частота и амплитуда колебаний которых зависит от параметра t. Используя (9), получаем из (13) уравнение для эволюции величины $\bar{u}$, которое для линейной зависимости $\varphi(x) = ax + b$ позволяет найти в явном виде $\mu(\tau)$:

$$\frac{\partial^2 \mu}{\partial \tau^2} = \frac{\partial \bar{u}}{\partial \tau} = \int \dot{x}\frac{\partial F(x,\dot{x})}{\partial \tau}dxd\dot{x} = \int \phi(x)F(x,\dot{x})dxd\dot{x} = a\mu + b \qquad (15)$$

Следовательно, задавая μ_0 и $\bar{u}_0$ в некоторый начальный момент времени, из (15) находим

$$\mu(\tau) = -\frac{b}{a} + \left(\mu_0 + \frac{b}{a}\right)\cos\left(\sqrt{a}\,\tau\right) + \frac{\bar{u}_0}{\sqrt{a}}\sin\left(\sqrt{a}\,\tau\right). \qquad (16)$$

Таким образом, частота колебаний выборочного среднего определяется в этом примере ковариацией величин x и $\ddot{x}$.

3. Построение прогноза ВФР временного ряда

Рассмотрим построение прогноза ВФР величин [1] на некотором интервале времени, меньшем горизонта прогноза $\theta(\varepsilon)$, по ε-стационарному распределению. Дискретный аналог уравнения (11) имеет вид

$$f(x,\tau+1) = f(x,\tau)\big(1-u(x,\tau)\big) + u(x-1,\tau)f(x-1,\tau) \qquad (17)$$

Это – явная разностная схема с левой разностью, которая имеет первый порядок аппроксимации по x и τ. Поскольку в действительности в нашей задаче дифференциального уравнения (11) не возникает, а исследуемым уравнением как раз и является разностное уравнение Лиувилля, определенное на фиксированной сетке с шагом по времени в 1 час и шагом по пространству, определяемым заданной точностью ε разбиения области изменения величины x, то мы не будем исследовать вопросы точности аппроксимации уравнения (11), устойчивости схемы и т.п. В этом смысле уравнение (17) для нас является точным, а (11) – его предельным аналогом.

Возможные варианты прогноза по уравнению (17) определяются выбором функции $u(x,\tau)$ по известным данным в предыдущие моменты времени, согласно (10). Для прогноза внутри промежутка ε-стационарности эту функцию естественно определять методами, применяемыми к стационарным рядам – регрессионными или автокорреляционными. Если же шаг по времени больше, чем $\theta(T;\varepsilon)$, то указанные рецепты не гарантируют желаемую точность прогноза, но, разумеется, они могут быть использованы для оценки того, как в среднем связаны одна с другой функции распределения вне промежутка своей квазистационарности. Мы ограничимся примером прогноза на сутки вперед, считая, что распределение должно быть 0,05-стационарно.

Согласно анализу, проведенному в п.2, оптимальным объемом выборки для этой задачи является $T(1;0{,}05) = 30$ суток, а максимальный горизонт прогноза равен $\theta(30;0{,}05) = 7$ суток. Внутри последнего промежутка функцию $u(x,t)$, определенную в (10), можно аппроксимировать, например, линейной регрессией на время:

$$u(x,t) = \bar{u}(x) + 6a(x)\frac{2t-p-1}{p^2-1}, \quad 1 < p \le 7;$$
$$\bar{u}(x) = \frac{1}{p}\sum_{t=1}^{p} u(x,t), \qquad (18)$$
$$a(x) = \frac{1}{2p}\sum_{t=1}^{p}\big(u(x,t)-\bar{u}(x)\big)\big(2t-p-1\big).$$

Иными словами, при построении прогноза (17) в качестве $u(x,\tau)$ можно взять ее аппроксимацию по данным за последнюю неделю, а ВФР строится по данным из скользящего окна объемом в месяц.

Так как отклонение прогнозной ВФР от фактической в предшествующий момент вре-

мени не превосходит ε и сами ВФР ε-стационарны, то отличие прогнозной ВФР от фактической в тот же момент времени не превышает 2ε по абсолютной величине:

$$\left|\tilde{f}(x,t+1)-f(x,t+1)\right|dx \le$$
$$\le \left|\tilde{f}(x,t+1)-f(x,t)\right|dx +$$
$$+ \left|f(x,t)-f(x,t+1)\right|dx \le 2\varepsilon$$

где f и $\tilde{f}$ – соответственно фактическая и прогнозная ВФР. На Рис. 5 приведено сравнение фактической ВФР и ее прогноза по описанному алгоритму. Видно, что точность прогноза достаточно высокая. Суммарное отличие прогнозной ВФР от фактической составило в этом примере 0,06.

Прогноз ВФР с шагом в 1 час позволяет также определить и сам временной ряд на рассматриваемом интервале времени. Можно, например, взять среднее значение по прогнозному распределению, либо наиболее вероятное, либо определить иной прогнозный функционал. Мы предлагаем следующий алгоритм прогнозирования собственно временного ряда, основанный на анализе эволюции его ВФР.

Поскольку выборки объемом T с интервалом в 1 шаг по времени отличаются только одним значением, то рассмотрим разность $\tilde{f}(x,t+1)-f(x,t)$. Согласно (11), (17), она равна $\partial(uf)/\partial x$. Если оказалось, что это выражение при всех x равно нулю, то полагаем $\tilde{x}(t+1)=x(t-T+1)$. В противном случае в качестве прогнозного значения $\tilde{x}(t+1)$ выберем аргумент, при котором указанная разность максимальна. Если таких аргументов несколько, то из них выбирается тот, вероятность которого выше. Если и таких аргументов несколько, то выбирается тот из них, который лежит ближе к глобальному максимуму ВФР. Если же вышеуказанные наиболее вероятные значения x находятся на равном расстоянии от максимума ВФР (таких точек может быть ровно 2), то берется тот аргумент, чей знак совпадает со знаком коэффициента асимметрии ВФР.

Построенный по описанному алгоритму суточный прогноз ряда [1] имеет ошибку 7%. В работе [2] были рассмотрены применения к ряду [1] некоторых стандартных методов прогнозирования. Оказалось, что при тех же условиях (объем выборки 1 месяц, горизонт прогноза 1 сутки) средняя ошибка регрессионного прогноза составила 13%, а автокорреляционные модели дали ошибку 8-9%. Таким образом, предложенный прогноз дает более точные результаты по сравнению с известными стандартными методами.

Представляется важным в дальнейшем провести исследование классов нестационарных временных рядов на предмет обладания ими стационарными индикаторами оптимального объема выборки. Это позволит выявить те из них, помимо примера, рассмотренного в данной работе, метод построения оптимального прогноза которых может быть корректно обоснован.

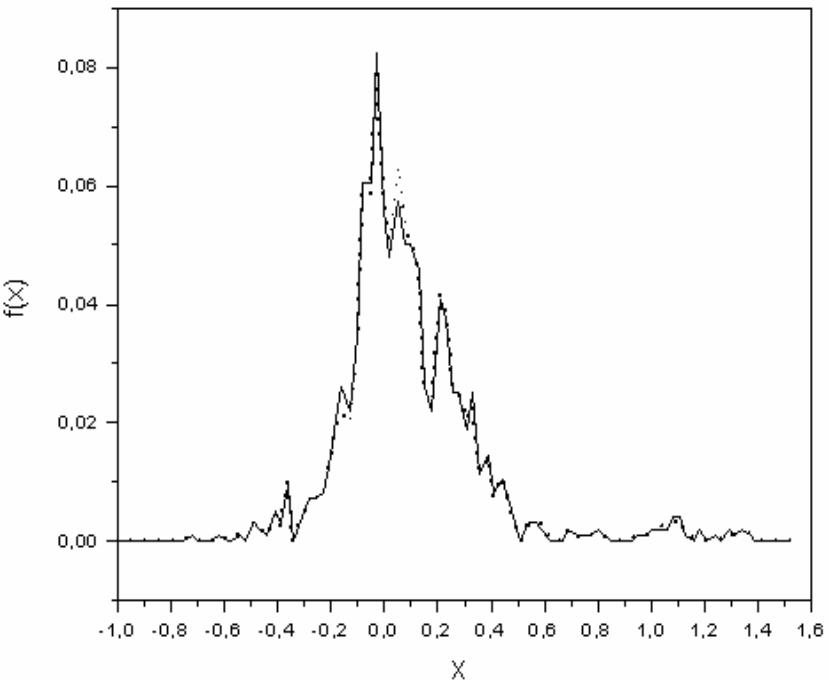

Рис. 5. Прогнозная ВФР (точечная линия) на 1 сутки вперед и фактическая ВФР (сплошная линия) по выборке в 30 суток

Литература

1. Статистика НП «АТС». http://www.np-ats.ru/index.jsp?pid=36
2. Орлов Ю.Н., Осминин К.П. Анализ нестационарных временных рядов. / Препринт ИПМ им. М.В. Келдыша РАН, № 36, 2007.
3. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006.
4. Лоэв М. Теория вероятностей. – ИЛ, 1962.
5. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985.
6. Уилкс С. Математическая статистика. – М.: Наука, 1967.

Орлов Юрий Николаевич. Зав. сектором кинетических уравнений Института прикладной математики им. М.В. Келдыша РАН, член Совета РАН по физико-техническому анализу энергетических систем, доцент МФТИ. Окончил МФТИ в 1987 году. Доктор физико-математических наук, специалист в областях классической и квантовой статистической механики, динамических систем, энергетики. Автор более 70 научных публикаций.

Осминин Константин Павлович. Аспирант механико-математического факультета МГУ. Специализируется в областях геометрии и топологии, теории особенностей, а также прикладной математической статистики. Имеет 10 публикаций в отраслевых изданиях и 3 публикации в научных изданиях.