Алгоритмы построения статистик для анализа и прогнозирования нестационарных временных рядов

К.П. Осминин

Аннотация. Описаны алгоритмы построения статистики горизонтного ряда и максимального горизонта прогнозирования при заданной допустимой погрешности в эмпирической выборочной плотности функции распределения нестационарного временного ряда. На основе оценивания статистических параметров горизонтного ряда разработан алгоритм прогнозирования выборочной функции распределения с помощью эмпирического уравнения эволюции. Приведен пример прогнозирования конкретного ряда, образованного движением цен на финансовом рынке.

Ключевые слова: нестационарный временной ряд, горизонтная статистика, оптимальный объем выборки, прогнозирование, кинетические уравнения

Введение

Одной из основных проблем в задачах анализа и прогнозирования временных рядов, возникающих в практической деятельности, является их нестационарность. Ряды могут быть нестационарными как в широком смысле, так и в узком [1]. Нестационарность в широком смысле означает, что, корреляционная функция ряда при фиксированном сдвиге или первый момент ряда или оба они вместе изменяются во времени. Нестационарность в узком смысле означает изменчивость во времени функции распределения.

В большинстве методов исследования, кратко описанных далее в разделе 1, ряд предполагается стационарным, а зависимость от времени считается возможным учесть различными нестатистическими методами. Например, применяется разложение ряда на три компоненты: трендовую, циклическую и случайную. Под трендом предполагается долговременно меняющаяся составляющая ряда, обусловленная влиянием фундаментальных факторов. Примером тренда может служить снижение количества соединений с использованием стационарных телефонов за счет перехода к мобильным телефонам. Циклическая компонента меняется во времени с определенным периодом и отвечает повторяющимся процессам. Примером циклической составляющей может служить сезонное снижение и увеличение теплового потребления населением. К случайной компоненте относятся все остальные составляющие, которые не могут быть отнесены к одной из первых двух групп.

Причины такого затруднения в идентификации могут быть следующими. Во-первых, несовершенство методов определения тренда и циклической компоненты. В случае, если не требуется высокая точность и ответ может быть приблизительным, можно сознательно сделать выбор в пользу не очень точных, но простых методов с невысокой вычислительной стоимостью. Во-вторых, компонента может быть устроена достаточно сложно. К примеру, ее трендовые характеристики могут быть нелинейными и неполиномиальными, как предполагается в большинстве методов выявления тренда. У циклической компоненты период может быть разным на разных промежутках времени,

а значения компоненты в пределах периода также изменчивы, хотя и подчиняются определенным законам. В-третьих, в практических задачах всегда присутствует «белый шум», наименьшая ошибка прогнозирования которого равна его дисперсии. Таким образом, ряд разбивается на три части, хотя это деление нечетко, циклическая компонента может стать трендом на промежутках времени, меньших периода, а тренд и циклическая компонента могут перейти в разряд случайных ввиду вышеуказанных причин.

Это разложение вполне эффективно при условии, что случайная компонента достаточно мала. Во многих случаях так и бывает. Однако существенная нестационарность ряда может привести к ситуации, когда циклическая компонента меняет свой вид из-за изменения периода, а долговременные тенденции уже не аппроксимируются трендовой моделью. В данной ситуации случайная компонента получается слишком большой, а анализ – неудовлетворительным. Для решения данной проблемы важно определить промежутках времени, на которых ряд меняется слабо и промежутки на которых его характеристики меняются настолько сильно, что делают неэффективными разбиение на три компоненты, и иные методы работы со стационарными рядами. Таким образом, возникает задача определения объема выборки, работая с которой можно добиться оптимальных результатов при анализе и прогнозировании. Для каждого типа исследования объем выборки «оптимален» по-своему, т.е. должен удовлетворять своим, вполне определенным, свойствам.

В задачах прогнозирования объем выборки, на основании которой делается прогноз, с одной стороны, должен быть достаточно большим для того, чтобы уловить закономерности ряда и его поведение, т.е. для его репрезентативности. Если ряд нестационарен, то его характеристики меняются со временем (вообще говоря, неизвестным образом), поэтому объем выборки не должен быть слишком большим в силу того, что делать прогноз, используя устаревшие данные, было бы неверно. С другой стороны, он не должен быть мал, иначе выборка будет нерепрезентативна. Необходимо понять, в какой момент данные «устаревают», т.е. поведение ряда меняется настолько существенно, что это мешает построению точного прогноза. Безусловно, это зависит от той точности прогноза, которую мы хотим получить. Чем большая точность необходима, тем точнее должен быть определен выборочный объем, используемый для построения прогноза.

При построении модели временного ряда первой задачей является аппроксимация имеющихся данных. В задаче аппроксимации отрезок ряда должен удовлетворять следующим свойствам. Во-первых, он должен быть достаточно велик для репрезентативности выборки. Во-вторых, на заданном промежутке времени ряд должен быть квазистационарен с некоторой точностью ε, т.е. выборочная функция распределения (далее ВФР) ряда меняется на величину меньше допустимой ошибки ε.

В этом месте требуется уточнение, объясняющее, что подразумевается под словами, «функция распределения меняется на такую-то величину» и «ряд изменяется несильно», «ряд квазистационарен с точностью ε». Иными словами, требуется ввести расстояние между двумя функциями распределения, формализовать понятие ошибки. Это составляет содержание раздела 2. В разделе 3 изложен алгоритм нахождения оптимального объема выборки для прогнозирования на заданный горизонт с точностью не выше заданной. В разделе 4 рассматривается двойственная к ней задача нахождения горизонта прогноза, оптимального для прогнозирования, на основе выборки заданного объема и с точностью не ниже заданной. В разделе 5 описан алгоритм прогнозирования ВФР с точностью не ниже заданной. В разделах 6 и 7 изложены методики прогнозирования первого момента ряда и самого ряда на основе прогноза ВФР. В заключении говорится об особенностях подхода с помощью предложенных алгоритмов и обозначаются направления дальнейших исследований. Блок-схемы алгоритмов приведены в приложениях.

1. Основные методы исследования стационарных рядов

Среди основных статистических методов исследования временных рядов [1-5] можно

выделить следующие: метод выделения тренда (временного сглаживания), регрессионный, автокорреляционный, адаптивный (скользящих средних), метод гармонического анализа, сингулярного спектрального анализа, бутстрепа (численного размножения выборок) и нейросетевой метод.

Выделение тренда предполагает выделение долговременного влияющего фактора. При описании тренда используются полиномиальные или квазиполиномиальные зависимости, дробно-рациональные и линейно-логарифмические функции и т.п. Этот метод эффективен в случаях, когда требуется невысокая точность, а также в качестве первого приближения, на основе которого применяются более точные методы.

Регрессионный метод выделяет зависимость ряда от некоторого параметра (например, времени) по существующему набору значений обоих переменных. Наиболее простой и распространенный – это метод линейной регрессии, когда зависимость предполагается линейной. Эффективность данного метода зависит от умения определить вид зависимости, а также от того, насколько сильно ряд ею определяется и насколько медленно меняются параметры зависимости.

Автокорреляционный метод использует цикличность ряда и зависимость его компонент между собой. Циклы определяются из рассмотрения автокорреляционной функции либо ее модификации – частной автокорреляционной функции. После этого строится модель, в которой элемент ряда зависит от предыдущих значений ряда, отстоящих от искомого элемента на расстояния, равные периодам выявленных циклов. Примером применения автокорреляций может служить модель Бокса-Дженкинса [3]. Метод эффективен, когда ряд имеет выраженную цикличность и периоды циклов неизменны во времени.

Адаптивный метод предполагает сглаживание ряда за значительный промежуток времени. Вместо ряда используется его скользящее среднее, возможно, взвешенное. Таким образом, из ряда удаляются все случайные колебания и циклические компоненты с периодом меньше отрезка сглаживания. После этого разумно применять автокорреляционные модели. Имеется большое количество вариаций данного метода. Например, параметрическая модель сглаживания, дополненная одновременным моделированием тренда и сглаживания, зависящих друг от друга, предложена в моделях Брауна, Хольта и Уинтерса [4, 5].

Из методов, приносящих интересные результаты в сложных ситуациях и с трудом поддающихся анализу традиционными методами, можно выделить следующие.

В методе сингулярного спектрального анализа выборки фиксированного объема T образуют T-мерное векторное пространство, в котором выбирается базис из T выборок. Далее ищется функциональная зависимость от базисной выборки значения ряда в следующий момент времени. Полученная функциональная зависимость определяет динамическую систему, у которой можно искать траектории, их функции распределения, аттрактор и его размерность и иные характеристики. Данный метод восходит к анализу нелинейных динамических систем, современный анализ хаотичности в которых проведен в монографиях [6, 7].

Спектральное разложение используется в случаях, когда объем ряда очень велик, и заключается в разложении в гармонический ряд с равномерно распределенными частотами и фазами [4, 8]. Спектральное представление корреляционной функции позволяет выявить скрытую периодичность значений ряда, но также может быть использовано и для оценки средних значений и других характеристик на основе данных выборки фиксированного объема. Знание выборочной спектральной функции может дать дополнительную информацию о корреляционной функции, что полезно для моделирования процесса.

В противоположной ситуации искусственно увеличить объем выборки помогает бутстреп - метод размножения выборки, разработанный Эфроном [9]. Метод заключается в построении ВФР на основе имеющийся выборки объемом N, извлечении из ВФР выборок с возвращением того же объема с равной вероятностью извлечения каждого значения. По получившимся выборкам производится оценка искомой величины, результаты усредняются. Возможна также модификация этого метода, предложенная в

[10], в соответствии с которой к исходному ряду добавляются малые независимые погрешности, формируя новую выборку.

Нейросетевые методы (например, [11]) используют математическую модель нейронных сетей, на основе которых функционирует мозг человека и других живых существ. Основу сети составляет нейрон – устройство, имеющее вход, преобразующее полученный сигнал и передающее его на выход, дальше по сети нейронам, соединенным с ним. Настройкой сети к применению служит обучение – сеть обучается на основе известных примеров (значений ряда), получая поощрения за правильный ответ (к примеру, прогноз) или ответ в пределах точности и наказание за ответ неправильный. Таким образом, получающийся ответ носит вероятностный характер. Особенностью нейросети является то, что создатель не должен знать закономерностей ряда при обучении сети, она обучается сама, на примерах. В этом же заключается и слабое место – сеть может выдавать очень точные ответы без указания, как они получились, представляя собой «черный ящик». Нейросетевой подход зачастую бывает очень эффективным в случаях, когда другие методы несостоятельны, является толерантным к приемлемому количеству ошибочных обучающих примеров, однако имеет и слабые стороны. К ним относятся относительность выдаваемых ответов, высокая вычислительная стоимость обучения, отсутствие гарантий приемлемости результата применения метода.

Рассмотренные методы имеют свои традиционные границы применимости, обусловленные их слабыми и сильными сторонами. Существующие методы работы с рядами как правило предполагают, что ряд стационарен либо законы его изменения во времени несложно определить. Другая особенность методов заключается в том, что при работе с характеристиками ряда традиционно рассматриваются первые два условия, функция распределения используется гораздо реже. Это связано с тем, что работать с первыми двумя условиями значительно проще, чем с функцией распределения. Также предполагается, что некоторые статистики, например, дисперсии оценок параметров эмпирического распределения, можно считать нормальными или несильно отклоняющимися от него на больших объемах выборки. В случае нестационарных рядов это предположение не выполняется.

В данной статье информация об изменчивости ряда извлекается из функции распределения. Рассматриваются промежутки квазистационарности ряда, объемы выборок, необходимые для прогнозирования на заданный горизонт, и наоборот, ищется максимальный горизонт, на который можно спрогнозировать ряд с заданного объема выборки, с заданной точностью. Указанные понятия и методы прогнозирования были введены в работах автора [12-14]. В настоящей работе описывается построенный алгоритм прогнозирования и приводится ряд примеров.

2. Основные понятия метода квазистационарных рядов

Пусть имеется временной ряд $\{x(t)\}$ с дискретным временем $t = 1,..., t_{\max}$. Обозначим $f_T(y,t)$, ВФР, зависящую от трех параметров: объема выборки T, на основе которого строится ВФР, конца выборки t (таким образом, ВФР строится на основе выборки $[t\text{-}T, t)$, правый конец не включается в силу дискретности времени) и значения y, частоту встречи которого мы ищем в $\{x(t)\}$. Введем интегральное расстояние на пространстве функций распределения:

$$\left\| f_T(y,t_1) - f_T(y,t_2) \right\| = \int \left| f_T(y,t_1) - f_T(y,t_2) \right| dy \qquad (1)$$

При прогнозировании ВФР ошибкой спрогнозированной функции $\widetilde{f}$ по отношению к фактической функции f будем называть расстояние между ними $\left\| \widetilde{f} - f \right\|$. Ошибкой прогноза временного ряда $\{x(t)\}$ на промежутке τ будем называть среднеквадратичное отклонение прогнозных значений $\widetilde{x}(t)$ от фактических $x(t)$ на промежутке горизонта прогноза:

$$\delta = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau} \left(\widetilde{x}(t) - x(t)\right)^2} \ . \qquad (2)$$

На основе нормы (1) удобно рассматривать функционал

$$V_T(t,\tau)=\int\left|f_T(y,t)-f_T(y,t+\tau)\right|dy, \qquad (3)$$

представляющий расстояние между двумя ВФР, построенными по выборке одинакового объема, но сдвинутыми одна относительно другой на промежуток τ. Если значение этого функционала не превосходит некоторого ε, распределение называется ε-стационарным [12, 13].

3. Алгоритм построения горизонтного ряда

Алгоритм CHS (Construction of Horizon Series) разработан для решения задачи об определении оптимального объема выборки, необходимого для прогнозирования на τ шагов вперед с ошибкой ВФР не выше ε, и предполагает следующие этапы.

1.Получение массива с временным рядом $\{x(t)\}$ с дискретным временем $t=1,...,t_{\max}$.

2. Предобработка ряда существующими методами. Например, для ряда, имеющего явный тренд, предобработка может включать в себя переход к разностям порядка, той же степени, что и аппроксимирующий тренд полином. Для ряда с хорошо выраженной циклической компонентой можно определить периоды, анализируя частную автокорреляционную функцию, найти циклическую составляющую путем усреднения ряда по периоду и перейти к остатку ряда после вычитания циклической компоненты. Если выраженных циклов несколько, то операцию можно повторить несколько раз, от больших периодов к малым. Предобработка заканчивается в тот момент, когда, по нашему мнению, ряд нельзя улучшить имеющимися в наличии методами. Получившийся остаток и будет объектом нашего исследования. В конце преобразуем ряд так, чтобы его значения заключались в пределах от -1 до 1. На Рис. 1 приведен фрагмент ряда дневной цены закрытия акций General Electric за 2004-2005 гг. по данным [15] до и после обработки.

3. Построение трехмерного массив ВФР $f_T(y,t)$. Область значений случайной величины *x* разбивается на *N* равных отрезков (хотя сетка может быть и неравномерной), и считается частота попадания значений *y* в эти отрезки.

Рис.1 Исходный нестационарный ряд и его остаток после вычитания кусочно-линейного тренда

Таким образом, ВФР строится с точностью до 1/2*N*:

$$f_T(i/N,t_0)=\frac{1}{T}\#\left\{x(t):\frac{2i}{N}-1\le x(t)<\frac{2(i+1)}{N}-1,t\in[t_0-T,t_0)\right\} \qquad (4)$$

Здесь *t* меняется от 2 до $t_{\max}$, *T* меняется от 1 до *t*, *i* меняется от 0 до *N*-1.

4. Составляется массив $V_T(t,\tau;x)$, представляющий набор расстояний (3) (или, если понятно, о каком ряде идет речь, просто $V_T(t,\tau)$). Этот массив в каждый момент времени *t* зависит от объема выборки *T*, на основе которого строится ВПФР в соответствии с определением (3) и величиной сдвига τ:

$$V_T(t,\tau;x)=$$

$$=\frac{1}{N}\sum_{i=0}^{N-1}\left|f_T(2i/N-1,t+\tau)-f_T(2i/N-1,t)\right|, \qquad (5)$$

где *t* и *T* изменяются в пределах, указанных выше.

5. Выбирается ε и составляется массив $h(t,\tau)=\min\{T:V_T(t,\tau)<\varepsilon\}$.

6. Для каждого момента времени *t* находится ближайший справа к $h(t,\tau)$ объем выборки T' при котором выборочная дисперсия достигает локального минимума, если он есть (Рис. 2) и имеет меньшее значение, чем дисперсия для выборки объема $h(t,\tau)$.

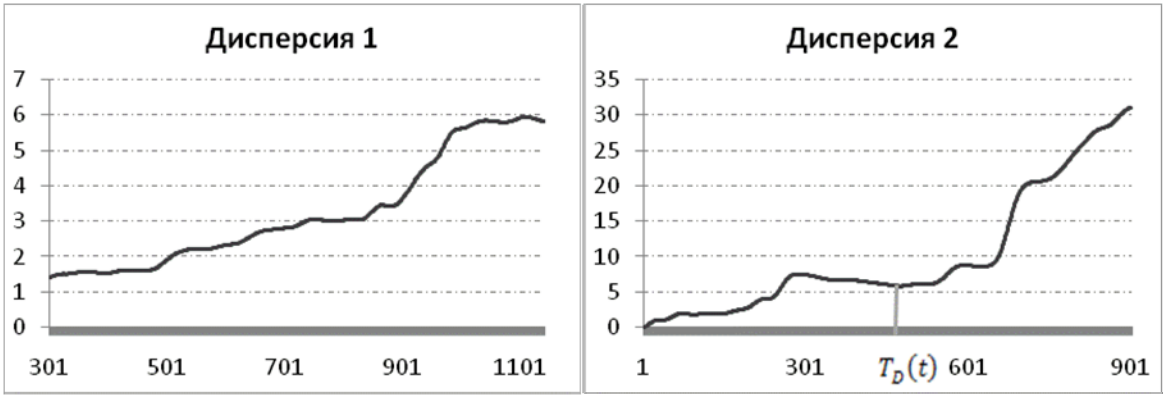

Рис.2. Выборочная дисперсия для нестационарного ряда (цена закрытия акций GE) в зависимости от объема выборки
Локальный минимум может существовать (справа) и может не существовать (слева)

7. Оптимальный объем, выборки определяется как наибольшая из величин T', $h(t,\tau)$: $h(\tau) = M\left(\max(h(t,\tau), T')\right)$. В случае, если объем исходного ряда велик, т.е. $t_{\max}$ велико, то построение всех массивов $h(\tau)$ будет довольно трудоемко. Для сокращения количества вычислений в некоторых случаях можно воспользоваться априорными соображениями о поведении ряда $\{x(t)\}$ для сужения пределов изменения параметров. При отсутствии таковых можно воспользоваться, например, следующим приемом. В вычислениях на этапах 3-6 время t изменяется от 1 до $t_{\max}$ с шагом $t_{\max}/100$, объем выборки T варьируется от 1 до t с шагом $t_{\max}/100$. После вычисления величины $h(\tau)$ алгоритм повторяется, но при этом пределы изменения времени t ограничиваются пределами от $t_{\max} - 3\sigma(h(t,\tau))$ до $t_{\max}$ с шагом $3\sigma(h(t,\tau))/100$. Левый конец диапазона изменения переменной t можно взять меньшим, если необходимо набрать большую статистику. Получив новое значения для $h(\tau)$, будем повторно применять этот алгоритм до тех пор, пока шаг изменения величин t и T не станет равным единице. Этот прием, однако, не гарантирует оптимальность полученной величины $h(\tau)$ ввиду немонотонности функционала $V_T(t,\tau;x)$. Размер массивов в 100 элементов в данной работе был подобран эмпирически и может быть заменен на другой в зависимости от конкретного ряда.

Полученная величина $h(\tau)$ является оптимальным объемом выборки, на основе которой строится прогноз на τ шагов вперед (Рис. 3). Обоснование этого факта приведено в [12, 14]. По построению, ВФР ряда на промежутке $[t_{\max} + \tau - h(\tau), t_{\max} + \tau)$ будет отличаться от ВФР, построенной на промежутке $[t_{\max} - h(\tau), t_{\max})$ не более, чем на ε. Это дает возможность прогнозировать на τ шагов вперед каким-либо методом, применяемым к стационарным рядам. Точность прогнозирования будет не ниже ε в силу того, что такую точность уже дает наивный прогноз. Важно иметь в виду, что состоятельность (т.е. асимптотическая сходимость по вероятности) этой оценки в общем случае не может быть установлена ввиду нестационарности ряда.

Распределение статистики горизонтного ряда представлена на Рис. 4

Из графика видно, что значения горизонтного ряда находятся вблизи теоретического максимума.

4. Алгоритм построения максимального горизонта прогнозирования

Опишем структуру алгоритма CHF (Construction of maximal Horizon Forecast). Пусть требуется решить двойственную к предыдущей задачу определения максимального горизонта прогноза по имеющемуся объему выборки T, при котором точность прогнозирования ВПФР будет не ниже ε. Такая задача возникает, например, при наличии априорной оценки выборочного объема, а также в ситуации ограниченного набора данных.

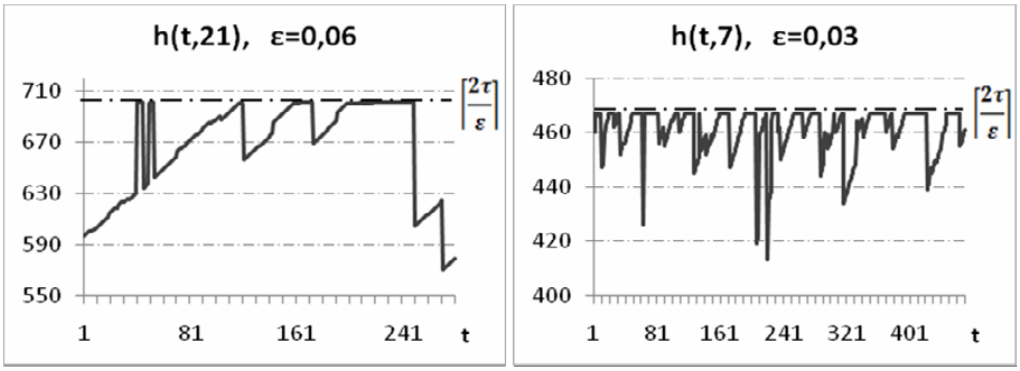

Рис.3. Горизонтный ряд нестационарного ряда (дневная цена закрытия акций General Electric за 2004-2005 гг.) для горизонта прогнозирования в 21 день и в 7 дней при точности в 0,06 и 0,03 соответственно

1-2. Этапы 1 и 2 – те же, что и в описании алгоритма CHS.

3. Задаются области изменения параметров: t меняется от 1 до $t_{\max}$, T задано условиями задачи. В первой итерации τ меняется от 1 до величины $\tau_{\max}$, определяемой априорно с таким расчетом, чтобы функционал (3) при $\tau = \tau_{\max}$ и всех значениях t, как указано выше, по модулю не превосходил ε. В принципе, значение τ может оказаться сколь угодно большим, однако в практических задачах τ оказывается на два порядка меньше T.

4. Построим массив выборочной функции распределения $f_T(y,t)$, где t и τ изменяются в пределах, указанных в пункте 3.

5. Строится массив $V_T(t,\tau;x)$, как и в пункте 4 алгоритма CHS.

6. Фиксируется величина ε и составляется массив $g(t,T) = \max\{\tau : V_T(t,\tau) < \varepsilon\}$.

7. Строятся массивы величин $g(T) = Mg(t,T)$ и $e(T) = \#\{t : g(t,T) \le g(T)\}/t_{\max}$.

В случае, если объем исходного ряда велик, т.е. $t_{\max}$ велико, можно применить следующую модификацию приема из предыдущего параграфа. В вычислениях пунктов 3-6 время t изменять от 1 до $t_{\max}$ с шагом $t_{\max}/100$, τ изменять от 1 до некоторого $\tau_{\max}$, значение которого установлено в пункте 3, с шагом $\tau_{\max}/10$. В знаменателе формулы п.7 вместо $t_{\max}$ будет стоять 100. После вычисления величины $g(T)$ алгоритм повторно выполняется, но при этом пределы изменения времени t ограничиваются пределами от $t_{\max} - 3\sigma(g(t,\tau))$ до $t_{\max}$ с шагом $3\sigma(g(t,\tau))/100$. Левый конец диапазона изменения переменной t можно взять меньшим, если необходимо набрать большую статистику. Горизонт прогноза τ будет изменяться от 1 до $g(T) + 3\sigma(g(t,\tau))$ с шагом $[g(T) + 3\sigma(g(t,\tau))]/100$. Получив новое значение для $g(T)$, будем повторно применять алгоритм до тех пор, пока шаг изменения величин t и τ не станет равным единице. Как отмечалось выше, этот прием не гарантирует оптимальность полученной величины $g(T)$ ввиду немонотонности функционала $V_T(t,\tau;x)$.

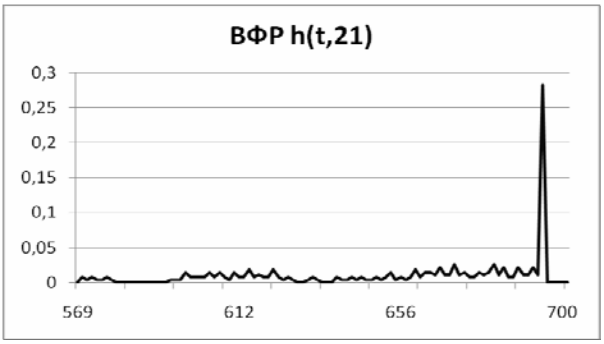

Рис. 4. Эмпирическая функция распределения статистики горизонтного ряда h(t,21) из Рис.3

Полученная величина $g(T)$ является величиной максимального горизонта, при прогнозировании наивным прогнозом на который ВПФР, построенная по выборке заданного объема T, будет отличаться от фактической ВПФР не более, чем на ε. Именно, согласно [12, 14], ВПФР данного ряда на промежутке $[t_{\max} + g(T) - T, t_{\max} + g(T))$ будет отличаться

от ВПФР, построенной на промежутке $[t_{\max}-T, t_{\max})$ не более, чем на ε. Данный факт позволяет прогнозировать ряд или ВПФР на $g(T)$ шагов вперед каким-либо методом, применяемым к стационарным рядам.

5. Прогнозирование ВФР с помощью уравнения эволюции

Имея оценку изменения ВПФР при сдвиге выборки на шаг τ, будет целесообразно прогнозировать поведение ряда с помощью метода Evolution Forecast of sample Distribution function (EFD), прогнозирующего именно ВПФР, а не сами элементы ряда. Этот метод, изложенный в предыдущей работе автора [14], позволяет построить ВПФР на промежутке $[t_{\max}, t_{\max}+\tau)$. В данной работе описывается численный алгоритм, реализующий прогноз ВПФР на τ шагов вперед с ошибкой не выше заданной.

1. Найдем величину $h(\tau)$ с точностью ε согласно описанному выше алгоритму CHS.

2. Установим границы изменения переменных:

время *t* изменяется от $t_{\min} = t_{\max} - h(\tau) - \sqrt{3D(h(t,\tau))}$ до $t_{\max}$ с шагом $(t_{\max} - t_{\min})/100$;

объем выборки $T = h(\tau)$.

3. Введем конечно-разностную производную $\dot{x}(t) = x(t) - x(t-1)$, t меняется в выбранных нами пределах.

4. Рассмотрим ВПФР по двум переменным x и $p = \dot{x}$, обозначим ее через $F_T(x,p,t)$. Аналогично определению одномерной ВФР, строим массив

$$F_T\left(\frac{2i}{N}-1, \frac{2i}{N}-2, t_0\right) =$$

$$= \frac{1}{T}\#\left\{\frac{2i}{N}-1 \le x(t) < \frac{2(i+1)}{N}-1, \frac{2j}{N}-\right.$$

$$\left. -2 \le \dot{x}(t) < \frac{2(i+1)}{N}-2, t \in [t_0-T, t_0)\right\}.$$

Здесь *i* меняется от 0 до *N*-1, *j* меняется от 0 до 2*N*-1. Таким образом, выполняется равенство $f_T(y,t) = \sum_{j=0}^{2N-1} F_T\left(y, \frac{2j}{N}-2, t\right)$.

5. Введем функцию «средней скорости» $u_T(y,t)$ согласно уравнению

$$u_T(y,t) f_T(y,t) = \int p F_T(y,p,t) dp. \qquad (6)$$

В [12] показано, что $f_T(y,t)$ удовлетворяет «эмпирическому» уравнению Лиувилля, т.е. уравнению, построенному с использованием эмпирической «средней скорости» (6):

$$\frac{\partial f}{\partial t} + \frac{\partial (uf)}{\partial x} = 0. \qquad (7)$$

6. В конечных разностях уравнение (7) имеет вид

$$f_T\left(\frac{i}{N}, t+\tau+1\right) = f_T\left(\frac{i}{N}, t+\tau\right)\left(1 - u_T\left(\frac{i}{N}, t+\tau\right)\right) +$$

$$+ u_T\left(\frac{i-1}{N}, t+\tau\right) f_T\left(\frac{i-1}{N}, t+\tau\right).$$

Данная схема позволяет рекуррентно находить ВПФР $f_T(y,t)$, основываясь на значениях этой функции и значениях произведения $u_T(y,t) f_T(y,t)$ в предыдущий момент времени.

7. Чтобы построить прогноз $f_T(y,t)$ в моменты времени $t_{\max}, ..., t_{\max}+\tau-1$, следует спрогнозировать «среднюю скорость» по предыдущим данным в объеме горизонта прогноза. Вычислим $u_T(y,t)$ при *t* и *T*, изменяющихся в указанных пределах. При *y* таких, что $f_T(y,t) = 0$, полагаем $u_T(y,t) = 0$. Поскольку функция $u_T(y,t)$ является «средней скоростью» изменения ряда, то вполне логично ожидать, что она как ряд от *t* имеет более стационарный характер, нежели *x*(*t*). В частности, она лучше поддается прогнозированию стандартным инструментарием, который, напомним, корректно применять в области ε-стационарности ВПФР. Результаты расчетов для рассматриваемого в работе примера [15] приведены на Рис. 5.

6. Прогнозирование первого момента ряда

В случае, если требуется найти не сам ряд, а, к примеру, его первый момент, изложенный метод имеет модификацию под названием First Moment Forecast (FMF) и заключается в следующем.

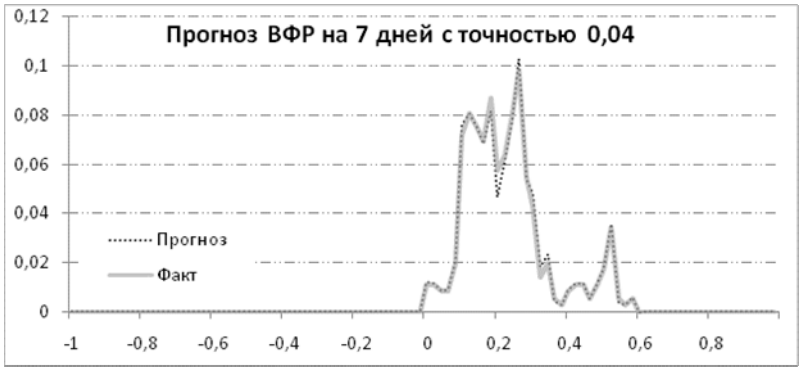

Рис. 5. Прогнозное и фактическое значение ВФР указанного выше ряда при прогнозировании на 7 дней вперед с точностью 0,04

В основном ошибка накапливается на экстремальных значениях

1. Найдем величину $h(\tau)$ с точностью ε согласно описанному выше алгоритму.

2. Установим границы изменения переменных:

время *t* изменяется от

$t_{\min} = t_{\max} - h(\tau) - \sqrt{3D(h(t,\tau))}$ до $t_{\max}$

с шагом $(t_{\max} - t_{\min})/100$;

объем выборки $T = h(\tau)$.

3. Введем конечно-разностную производную $\dot{x}(t) = x(t) - x(t-1)$, t меняется в выбранных нами пределах.

4. Вторую производную исходного ряда $x(t)$ по времени будем моделировать на основе выборочных данных: $\ddot{x} = \varphi(x,t)$.

5. Введем функцию «средней скорости» $u_T(y,t)$ согласно (6) и вычислим массив этих данных при t и T, изменяющихся в указанных пределах.

6. Введем первые два выборочных момента

$$\mu(t,T) = \sum_{i=0}^{N-1}\left(\frac{2i}{N}-1\right) f_T\left(\frac{i}{N},t\right),$$

$$\sigma^2(t,T) = \sum_{i=0}^{N-1}\left(\frac{2i}{N}-1-\mu(t,T)\right)^2 f_T\left(\frac{2i}{N}-1,t\right).$$

Тогда имеют место соотношения:

$$\frac{\partial\mu}{\partial t} = \bar{u}, \qquad (8)$$

$$\frac{\partial\sigma^2}{\partial t} = 2\sum_{i=0}^{N-1}\left(\frac{2i}{N}-1-\mu(t,T)\right)(u-\bar{u}) f_T\left(\frac{2i}{N}-1,t\right),$$

где $\bar{u} = \sum_{i=0}^{N-1} u_T\left(\frac{2i}{N}-1,t\right) f_T\left(\frac{2i}{N}-1,t\right)$.

7. Аппроксимируем *φ(x)* линейной регрессией по *x*, т.е. положим $\varphi(x,t) = ax(t) + b$. Тогда верно соотношение $\partial\mu/\partial t = a\mu + b$.

8. Зададим начальные значения $\mu_0, \bar{u}_0$ в момент времени $t_{\max}$. Тогда получаем

$$\mu(t_{\max}+t,T) = -\frac{b}{a} + \left(\mu_0 + \frac{b}{a}\right)\cos\left(\sqrt{a}t\right) + \frac{\bar{u}_0}{a}\sin\left(\sqrt{a}t\right),$$
$$t = 1,\ldots,\tau. \qquad (9)$$

Обоснования формул (8-9) приведены в [12].

7. Прогнозирование собственно временного ряда

В случае, если необходим прогноз самого ряда {*x(t)*}, то, имея прогноз ВПФР $f_T(y,t)$ методом EFD, сам ряд можно восстановить следующим образом. Рассмотрим

$$d(y,t,T) = f_T(y,t+1) - f_T(y,t) - \partial(uf)/\partial x, \quad t \in [t_{\max}, t_{\max}+\tau). \qquad (10)$$

В качестве прогноза ряда в момент *t*+1 выберем значение, при котором *d(y,t,T)* максимально. Если таких значений несколько, то из них выбирается значение с максимальной вероятностью. Если и таких несколько, то из них выбирается значение, ближе всего находящееся к медиане. Если и таких несколько (а именно, не более двух), то выбирается значение со зна-

ком, равным знаку коэффициента асимметрии ВПФР. В случае если $d(y,t,T) = const = 0$, то $x(t+1) = x(t)$.

Заметим, что перед применением данного алгоритма необходимо удостовериться, что прогнозирование функции $u_T(y,t)$ в методике EFD не осуществлялось регрессией по времени, иначе прогноз ряда данным методом будет выдавать одну и ту же величину, а именно величину, при которой коэффициент регрессии $u_T(y,t)$ максимален.

Заключение

Таким образом, в работе предложены алгоритмы, позволяющие найти объем выборки, необходимый для прогнозирования нестационарного временного ряда на заданный горизонт времени с заданной точностью (CHS). На этой основе алгоритмом EFD строится прогноз ВПФР на указанном промежутке времени, что позволяет найти прогноз первого момента (FMF) и найти прогноз самого ряда на указанный промежуток. Предложен также алгоритм CHF нахождения максимального горизонта прогноза, строящегося на основе выборки заданного объема и с заданной точностью.

Данные методы удобны тем, что позволяют априори определить верхний предел ошибки прогнозирования в смысле интегральной разности функций распределения, чего нельзя сделать непосредственно для нестационарного ряда. В этом случае предпочтительнее работать с ВПФР, чем с самим рядом или его моментами.

Каждый следующий алгоритм опирается на предыдущий, однако каждый из них обладает самостоятельной ценностью. Эти алгоритмы могут служить целям анализа динамики ряда, определения периодов его квазистационарности, исследования распределения в будущем, прогнозирования ряда на заданный промежуток времени с заданной точностью, отслеживания его изменений. Возможны модификации названных методов, позволяющие прогнозировать ВПФР вне пределов квазистационарности, но с потерей точности. Также существует алгоритм, по объему выборки T и точности ε определяющий максимальный горизонт прогнозирования τ, на который прогнозируется ВПФР с ошибкой не выше ε с использованием выборки объема T.

Отслеживание изменений означает определение новых тенденций в поведении ряда и определение моментов их появления. К примеру, горизонтный ряд $h(t,\tau)$ является своеобразным индикатором изменения закономерностей ряда $x(t)$, резкое изменение горизонтного ряда по переменной t сигнализирует о возможном появлении нового фактора, влияющего на динамику исходного ряда либо о существенном изменении имеющихся факторов. Профиль горизонтного ряда является показателем относительной неизменности совокупности определяющих ряд факторов. Данный индикатор может быть полезен, например, для анализа изменения структуры рынка ценных бумаг, изменения эластичности спроса, а период перехода горизонтного ряда к новой устойчивой структуре может характеризовать скорость протекания процессов саморегуляции в системе.

Горизонтный ряд является интересным объектом для исследований ввиду своей информативности. В практических примерах, как указывалось выше (Рис.3 и Рис. 4), значения горизонтного ряда близки к своему теоретическому максимуму [12-14]. Если указать условия, которым должна удовлетворять функция распределения исходного временного ряда и при которых горизонтный ряд ведет себя более стационарным образом, а его значения близки к теоретическому максимуму, то это будет значимым критерием при анализе нестационарных временных рядов. Это и будет направлением дальнейших исследований.

Автор благодарит к.ф.-м.н. Н.А. Митина за полезные обсуждения методики и предложений по подбору примеров для тестирования алгоритма.

Литература

1. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
2. А.И. Кобзарь Прикладная математическая статистика, Москва, 2006
3. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. (пер. с англ.) – М.: Мир, 1974.

4. Кильдишев Г.С., Френкель А.А. Анализ временных рядов и прогнозирование. М.: «Статистика», 1973.
5. Лукашин Ю.П. Адаптивные методы прогнозирования экономических показателей. М.: «Статистика», 1979.
6. Малинецкий Г.Г., Потапов А.Б. Современные проблемы нелинейной динамики. – Москва-Ижевск: Регулярная и хаотическая динамика, 2000.
7. Мандельброт Б. Фракталы, случай и финансы. (пер. с англ.) – Москва-Ижевск: Регулярная и хаотическая динамика, 2004.
8. Розанов Ю.А. Случайные процессы. – М.: Наука, 1971. – 286 с.
9. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. - М.: Финансы и статистика, 1988. - 263 с.
10. Орлов А.И. О реальных возможностях бутстрепа как статистического метода. // Заводская лаборатория. 1987. Т.53. No.10. С.82-85.
11. Хайкин С. Нейронные сети. Полный курс (пер. с англ.) – Москва, С-Петербург, Киев, «Вильямс», 2006.
12. Орлов Ю.Н., Осминин К.П. Методика определения оптимального объема выборки для прогнозирования нестационарного временного ряда, ИТВС № 4 2007
13. Орлов Ю.Н., Осминин К.П. Анализ нестационарных временных рядов, ИМП им. М.В. Келдыша РАН, препринт №36 за 2007 год.
14. Орлов Ю.Н., Осминин К.П. Построение выборочной функции распределения для прогнозирования нестационарного временного ряда. // Мат. Мод. 2008.
15. http://moneycentral.msn.com/detail/stock_quote?Symbol=GE&pkw=PI&vendor=Paid+Inclusion&OCID=iSEMPI

Осминин Константин Павлович. Аспирант Мехмата МГУ 3-го года обучения. Специализируется в областях геометрии и топологии, теории особенностей, а также прикладной математической статистики. Имеет 11 публикаций в отраслевых изданиях по анализу рынка электроэнергии и 3 публикации в научных изданиях о методах исследования нестационарных процессов. Эл. адрес: ov3159f@yandex.ru.