

# Обзор методов статистического анализа временных рядов и проблемы, возникающие при анализе нестационарных временных рядов

## Газизов Д. И.

*Газизов Данияр Ильдарович / Gazizov Daniyar Ildarovich - студент,  
кафедра прикладной информатики,  
факультет прикладной математики и информационных технологий,  
Финансовый университет при Правительстве Российской Федерации, г. Москва*

**Аннотация:** в статье рассматриваются основные методы анализа и прогнозирования временных рядов, а также проблемы и недостатки этих методов, которые возникают при их применении к нестационарным временным рядам.

**Ключевые слова:** стационарность, нестационарные временные ряды, методы статистического анализа временных рядов.

### Основные методы анализа и прогнозирования временных рядов.

В этом разделе рассмотрены основные методы анализа временных рядов, часто применяемые на практике. Эти методы в силу своей общепотребительности служат базисом для сравнения с ними вновь разрабатываемых статистических моделей.

Основными статистическими методами исследования временных рядов являются: метод выделения тренда (временного сглаживания), регрессионный, автокорреляционный, адаптивный (скользящих средних), метод гармонического анализа, сингулярного спектрального анализа, бутстрепа (численного размножения выборок) и нейросетевой. Ниже кратко описывается идеология этих методов, даются основные определения из математической статистики и приводятся базовые уравнения соответствующих моделей.

Напомним [1, 2], что случайным процессом на некотором вероятностном пространстве называется семейство случайных величин  $x(t)$ , принимающих значения из множества, называемого областью определения процесса. Если параметр  $t$  принимает дискретные значения, то процесс называется временным рядом.

Временной ряд называется стационарным в широком смысле, если его математическое ожидание не зависит от времени  $t$ , а корреляционная функция, являющаяся математическим ожиданием произведения отклонений значений ряда от среднего в различные моменты времени  $t_1$  и  $t_2$  зависит только от разности  $t_1 - t_2$ . Более общее определение предполагает независимость от времени центральных моментов ряда вплоть до некоторого конечного порядка.

Временной ряд  $x(t)$  называется стационарным в узком смысле [3], если при любых  $t$  и  $\tau$  случайная величина  $x(t)$  распределена одинаково с величиной  $x(t + \tau)$ .

В настоящей работе используется определение стационарности в широком смысле, если речь идет о моментах ряда, и в узком смысле, если о его распределении.

Рассмотрение существующих подходов к анализу временных рядов начнем с метода временного сглаживания или выделения тренда. При исследовании временных рядов принято выделять несколько составляющих:

$$x(t) = x_{\text{тренд}}(t) + x_{\text{цикл}}(t) + \xi(t), \quad (1)$$

где  $x_{\text{тренд}}(t)$  - плавно меняющаяся компонента, определяемая долговременной тенденцией (трендом) изменения ряда признаков,  $x_{\text{цикл}}(t)$  - циклическая или сезонная компонента, которая отражает повторяемость процессов на определенных промежутках времени, а  $\xi(t)$  - случайная компонента, содержащая влияние прочих факторов, механизм которого (влияния) скрыт от наблюдателя. Первые две составляющих (тренд и цикл) в идеале должны быть описаны точно, т.к., это закономерные факторы, изучаемые в рамках детерминистских моделей. Однако следует заметить, что сами детерминистские модели представляют определенную идеализацию описываемых закономерностей, поэтому им также присуща некоторая неточность. В этом смысле представление (1) несколько условно, но оно бывает полезно на практике для интерпретации результатов статистического анализа данных.

Трендовая компонента временных рядов обычно не известна точно, а, как и ряд в целом, является случайной величиной, но ее изменение из некоторых априорных суждений часто может быть качественно описано аналитически. Для описания тренда используются так называемые кривые роста, которые позволяют моделировать процессы трех основных качественных типов: без предела роста, с пределом роста без точки перегиба, а также с пределом роста и точкой перегиба.

Процессы развития без предела роста характерны в основном для абсолютных объемных показателей. Процессы с пределом роста характерны для относительных показателей, таких, как душевое потребление продуктов питания, внесение удобрений на единицу площади, затраты на единицу произведенной продукции и т.п. Процесс с пределом роста и точкой перегиба характерен, например, для описания

изменения спроса на новые товары.

Для моделирования этих процессов используются полиномиальные или квазиполиномиальные (с экспоненциальными множителями и т.п.) зависимости, дробно-рациональные и линейно-логарифмические функции, кривые Гомперца и иные функциональные зависимости. В рамках многопараметрических моделей, часто бывает возможно провести аппроксимацию данных с требуемой точностью, однако этот подход не всегда удовлетворителен при прогнозировании, поскольку подбираемые функции не обязательно отражают реально обусловленную зависимость наблюдаемой величины от времени.

Таким образом, часто используемым методом моделирования нестационарных временных рядов является параметрическое оценивание. В этом случае подбираются параметры той или иной функциональной зависимости для трендовой составляющей, после исключения которой, остается стационарный ряд. Оставшийся ряд может и не быть стационарным в смысле математического определения этого понятия, но на практике его удобно считать таковым с доверительной вероятностью, достаточной для исследователя. Для этой цели используются различные тесты на стационарность [1], которые, как правило, разработаны для применения к известным функциональным зависимостям (например, нормальному, экспоненциальному или равномерному распределению).

В частности, для нормально распределенных случайных величин тест на наличие тренда проводится в основном по критериям Стьюдента и Фишера [1], но также существуют и другие тесты, использующие нормальность распределения. Например, статистики критерия Фостера-Стюарта [1] для обнаружения тренда в среднем значении и дисперсии используют производные случайные ряды из нулей и единиц, определяющие наличие тренда в максимумах или минимумах исходного ряда.

Если нет оснований предполагать нетривиальную функциональную зависимость трендовой составляющей ряда, ее часто считают полиномиальной. В этом случае такой тренд может быть исключен путем перехода к первым, вторым и т.д. разностям в значениях ряда, т.е. вместо ряда  $x(t)$  можно рассмотреть ряд  $x(t) - x(t-1)$  или ряд из разностей более высокого порядка, называемый производным рядом. Такой метод достаточно эффективен, если функциональный тип тренда сохраняется во времени.

Целью сведения временного ряда к стационарному является появляющаяся тогда возможность использования теоремы Гливенко о сходимости эмпирической вероятности к распределению генеральной совокупности и критерия согласия Колмогорова о близости выборочной функции распределения и распределения генеральной совокупности [4, 5] для того, чтобы попытаться определить вид распределения, к которому относилась бы изучаемая выборка данных, после чего с известной доверительной вероятностью строить прогноз.

Желание иметь дело со стационарным рядом вызвано также возможностью обосновать прогнозные модели для такого ряда применением теоремы Вальда о разложении, 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, т.е. стационарного процесса, фурье-разложение которого является константой). Поэтому, хотя реальные процессы, 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.

Ряды, которые после надлежащих подготовительных операций можно считать стационарными, далее изучаются методами регрессионного, корреляционного и гармонического анализов. Каждый из этих методов используется для создания некоторой прогнозной модели для изучаемых рядов. В зависимости от конкретной специфики ряда используются различные из перечисленных методов. Ниже кратко описаны их содержательные части.

Линейная регрессионная модель позволяет связать две величины  $Y$  и  $X$  линейной зависимостью вида  $Y = aX + b$  по имеющимся  $N$  парам значений  $(x_k, y_k)$  методом наименьших квадратов (МНК).

Обобщение линейной регрессионной модели на случай зависимости, возможно, нелинейной, от нескольких объясняющих переменных приводит к задаче выбора наиболее адекватной модели по числу переменных и виду регрессионных функций.

Регрессионные модели применяются в основном тогда, когда объясняющая переменная (в данном случае это величина  $X$ ) не является случайной, а автокорреляция между значениями другой (объясняемой величины  $Y$ ) мала.

Максимумы модуля автокорреляционной функции показывают наличие лагов, т.е. промежутков времени, на которых проявляется скрытая зависимость случайных величин. К примеру, в рядах с существенным влиянием циклической компоненты лаги выражены на графике выборочной автокорреляционной функции особенно сильно. Модели, использующие лаговую автокорреляцию, называются автокорреляционными (АМ) или авторегрессионными.

Для применения автокорреляционных моделей (АМ) желательно иметь временной ряд,

автокорреляционная функция которого имеет небольшое число максимумов и достаточно быстро спадает с ростом шага автокорреляции. Если имеется цикличность данных, которая меняется со временем, то полностью исключить ее различными методами сглаживания, как правило, не удастся. В этом случае автокорреляционные модели применяются на этапе качественного анализа, точность которого должна быть улучшена с использованием других подходов.

Последовательное усложнение автокорреляционных моделей привело к появлению т.н., адаптивных моделей прогнозирования, которые базируются на объединении двух схем - скользящего среднего (СС) и авторегрессии. Модель СС состоит в том, что для определения свойств временного ряда с целью краткосрочного прогноза берется выборка последних данных за некоторый промежуток времени  $T$ .

Модификацией модели СС является взвешенная схема СС, когда оценкой текущего уровня является взвешенное среднее всех предшествующих уровней, причем веса при наблюдениях убывают по мере удаления от последнего уровня, т.е., информационная ценность наблюдений признается тем большей, чем ближе находятся они к концу интервала наблюдений. Возможны также варианты, когда наибольшую ценность имеют наблюдения с определенным лагом.

Другим типом адаптивных прогнозных моделей является параметрическая модель сглаживания, когда значение в момент  $t+1$  определяется не только статистиками самого ряда, но и прогнозом этого ряда за несколько предшествующих моментов времени, а также оценками текущих трендов (модели Брауна, Хольта и Уинтерса). Как правило, адаптивные модели применяются для прогнозирования ряда на один шаг вперед.

Одним из методов выделения т.н. главной части или главных компонентов временного ряда является метод сингулярного спектрального анализа, разработанного в теории нелинейных динамических систем и представляющего оригинальный подход к исследованию автокорреляционной зависимости.

Из других - «нетрадиционных» методов анализа временных рядов - следует отметить нейросетевой метод и метод размножения выборок. Нейросетевой метод анализа случайных процессов является одним из активно разрабатываемых в последнее время [6]. Он, как и большинство описанных выше методов, направлен на отыскание корреляционной связи между элементами временного ряда. Целью метода является отыскание «предвестника» наступления того или иного события и определение вероятности последнего. Главной задачей является «обучение» искусственной нейронной сети на достаточно большом количестве выборок некоторого объема, после чего принятие решения о наличии «предвестника» принимается сетью на основе эмпирической вероятности наступления похожих случаев в период «обучения». Нейросетевые методы 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.

Метод размножения выборки (бутстреп), предложенный в 1977 году Б. Эфроном [7], является одним из методов случайной обработки данных стационарных процессов. Его сущность состоит в том, что по имеющимся  $N$  наблюдениям за случайной величиной, образующим по предположению выборку из генеральной совокупности, строится выборочная функция распределения, из которой извлекаются выборки с возвращением того же объема  $N$  с равными вероятностями извлечения каждого значения. По каждой выборке строится оценка интересующего параметра исходной случайной величины, а затем полученные оценки усредняются.

Из проведенного рассмотрения следует, что ни одно из перечисленных основных направлений статистического анализа временных рядов не является универсальным: конкретный случайный процесс лучше всего моделируется методом одного из направлений. Нет утверждения о том, что некоторый метод при своем практическом применении дает наименьшую ошибку прогнозирования для любых временных рядов. Напротив, каждый из методов имеет определенные ограничения, препятствующие их эффективному применению к задачам прогнозирования. Проблемы, возникающие при анализе встречающихся на практике временных рядов, очерчены в следующем параграфе.

Здесь не были упомянуты методы, развитые в теории случайных процессов определенных классов: марковских процессов, эргодических процессов, процессов с независимыми приращениями, ветвящихся

процессов и некоторых других. Эти методы предполагают определенную функциональную принадлежность временного ряда, для чего на практике обычно не бывает достаточных оснований, поскольку физическая природа изучаемых рядов, как правило, не известна.

#### **Проблемы анализа нестационарных временных рядов.**

Методы, описанные выше, корректно применимы в основном к стационарным рядам. Если ряд нестационарный, то теоремы об эффективности, состоятельности и асимптотической нормальности выборочных оценок и их дисперсий в общем случае не выполняются. Тем не менее, перечисленные методы применяются ко всем рядам, статистический анализ которых необходимо проводить для оптимизации той или иной практической деятельности. Новыми проблемами, возникающими при таком не вполне обоснованном применении, являются: задача минимизации ошибки прогнозирования для выбранного метода статистического анализа и задача выбора наиболее адекватной модели временного ряда. Последняя задача существует и при исследовании стационарных рядов, но в этом случае выбор модели может быть проведен по известным алгоритмам спецификации моделей [2], позволяющим отобрать оптимальное число параметров в рамках дисперсионного и корреляционного анализов.

При определении ошибки прогноза нестационарного временного ряда надо учесть два фактора: конечность выборки и различие распределений для разных выборок вследствие нестационарности процесса. Разные методы имеют неодинаковую чувствительность точности аппроксимации данных к действию указанных факторов.

Таким образом, модели и методы прогнозирования стационарных рядов, такие, как регрессионные и корреляционные, нуждаются в адаптации при использовании их в нестационарном случае, поскольку тогда ошибка прогноза, получаемая этими методами, может не убывать с увеличением статистической базы. Рассмотрим в этой связи ограничения, присущие методам анализа и моделирования временных рядов, перечисленным в предыдущем разделе.

В моделях регрессионного анализа средние величины (математическое ожидание, дисперсия, ковариация) постоянны. Уточнение этой модели в случае зависимости указанных величин от времени, т.е. от текущего значения  $t$ , может быть сделано посредством аналитического моделирования такой зависимости, либо переходом к первым, вторым и т.д. разностям в нестационарных временных рядах, выражающих зависимость средних величин от времени, либо тем же регрессионным анализом - но уже изучаемых величин на время. Окно усреднения становится при этом скользящим. Однако остается невыясненным, какой ширины должно быть это окно.

Те же проблемы возникают и при использовании АМ или их обобщений. При этом возникают дополнительные трудности с анализом корреллограмм: например, необходимо отличать эффекты назначенной периодичности, связанной с суточным, недельным или иным циклом, и внутренне обусловленной зависимости между членами ряда. Увеличение промежутка усреднения в этом случае не приводит к успеху, поскольку зависимость, наблюдавшаяся в одной выборке, может исчезнуть в другой того же объема, но отнесенной к иному моменту времени. Усреднение корреллограммы по некоторому промежутку времени и переход к средней корреллограмме за период наблюдений увеличивает неточность прогноза на короткий промежуток времени, а на большом интервале такую задачу ставить вообще не очень осмысленно. Как и в случае с регрессионными моделями, наилучший период усреднения не известен.

Адаптивные модели, использующие весовые коэффициенты в обобщениях АМ, требуют весьма тонкой настройки сглаживающих функций в нестационарном случае, поскольку даже для стационарных процессов оптимальный выбор этих функций является отдельной достаточно сложной задачей. При этом в стороне остаются вопросы количественной оценки точности аппроксимации, которая меняется с течением времени, т.к. длина оптимальной выборки является локально переменной величиной с неизвестной статистикой.

Что касается гармонического анализа временных рядов, то он вообще имеет ценность только для стационарных в широком смысле процессов второго порядка, когда корреляционная функция зависит от разности моментов времени. Для нестационарных рядов большое число учитываемых членов ряда приводит к достаточно высокой погрешности в оценке статистических характеристик процесса в ближайшем будущем.

Метод сингулярного спектрального анализа представляется в этом контексте наиболее устойчивым к временному тренду, поскольку его задачей и является выделение соответствующих главных компонент ряда. Изменение с течением времени размерности пространства базисных векторов матрицы задержек, представляется маловероятным событием: размерность является своеобразным индикатором данного процесса, обусловленного определенными физическими явлениями, и ее изменение будет свидетельствовать о том, что процесс изменился по своему качеству. Тем не менее, вопрос о размерности самой матрицы и количественной зависимости от этой размерности числа базисных векторов остается в этом методе открытым.

Бутстреп и искусственные нейронные сети не имеют четкой схемы обобщения на нестационарные процессы. Размножать некоторую выборку имеет смысл, если только она принадлежит генеральной совокупности, в противном случае такой метод не дает информации о будущей выборке. Размножение выборок предполагает неизменность функции распределения, в противном случае размноженные выборки, полученные с помощью выборочной функции распределения на каком-либо начальном отрезке, будут сильно отличаться от выборок, полученных на основе оригинальной функции распределения, меняющейся во времени, и это различие будет тем больше, чем сильнее меняется функция распределения ряда.

Нейронные сети имеют ограниченную функцию применимости, их обучение довольно трудоемко, а с учетом нестационарности ряда может быть сложным вдвойне.

Таким образом, задача об определении оптимального объема выборки является весьма важной для прогнозирования, и, как показывает анализ имеющейся литературы, не имеющей в настоящее время даже четкой постановки.

### *Литература*

1. *Кобзарь А. И.* Прикладная математическая статистика. М.: Физматлит, 2006. 816 с.
2. *Кремер Н. Ш., Путко Б. А.* Эконометрика. М.: ЮНИТИ-ДАНА, 2002. 311 с.
3. *Уилкс С.* Математическая статистика, (пер. с англ.) М.: Наука, 1967. 632 с.
4. *Боровков А. А.* Математическая статистика. М.: Физматлит, 2007. 704 с.
5. *Гнеденко Б. В.* Курс теории вероятностей. М.: Физматлит, 1961. 406 с.
6. *Хайкин С.* Нейронные сети. Полный курс (пер. с англ.) М.: Вильямс, 2006. 1103 с.
7. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика, 1988. 263 с.