ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 53 за 2011 г.

Орлов Ю.Н., Шагов Д.О.

Индикативные статистики для нестационарных временных рядов

Рекомендуемая форма библиографической ссылки: Орлов Ю.Н., Шагов Д.О. Индикативные статистики для нестационарных временных рядов // Препринты ИПМ им. М.В.Келдыша. 2011. № 53. 20 с. URL: http://library.keldysh.ru/preprint.asp?id=2011-53

РОССИЙСКАЯ АКАДЕМИЯ НАУК

Ордена Ленина

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ

им. М.В. Келдыша

Ю.Н. Орлов, Д.О. Шагов

ИНДИКАТИВНЫЕ СТАТИСТИКИ ДЛЯ НЕСТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ

Москва, 2011

Ю.Н. Орлов[1], Д.О. Шагов[2]

[1]ИПМ им. М.В. Келдыша РАН

[2]МФТИ (ГУ)

Индикативные статистики для нестационарных временных рядов

Строятся статистики меры хаоса в нестационарных рядах. Для построения используются горизонтные ряды и согласованные уровни нестационарности для функции распределения расстояний между выборочными распределениями. Оптимизация объема выборки для построения согласованного уровня нестационарности проводится с помощью статистической добротности ряда.

Yu.N. Orlov, D.O. Shagov

Indicative statistics for non-stationary time series

The statistics for chaos measure are constructed in the case of non-stationary time series. For this purpose the self-consistent levels of stationary is introduced for distribution function of distances between empirical distribution functions. The optimal set of data is obtained with the use of statistical good quality factor.

1. ВВЕДЕНИЕ

1.1. Цель и направление исследований

Цель работы состоит в построении индикаторов изменения по времени выборочной плотности функции распределения нестационарного временного ряда и сравнительном анализе их достоверности.

При анализе стационарного временного ряда основной задачей является нахождение его функции распределения, т.е. вероятности принятия случайной величиной значений из определенного промежутка. Согласно теореме Гливенко [1] о сходимости эмпирической вероятности к теоретическому распределению, чем больше значений будет учтено при построении эмпирического распределения, тем точнее будет приближена генеральная совокупность. Для нестационарных процессов такой асимптотической сходимости нет, и потому требуются иные методы анализа данных. В этом случае помимо количественной задачи прогнозирования значения случайной величины возникает не менее важная задача качественного анализа распределения – так называемая задача определения момента разладки.

Во многих практических примерах локальное по времени поведение нестационарной выборочной функции распределения можно представить как квазипериодическое, для которого характерны три стадии: квазистационарное состояние, переходный процесс, новое квазистационарное состояние. Продолжительность переходного процесса определяется как объективными факторами, характеризующими смену режима (собственно разладка), так и субъективными – а именно, объемом выборки, который используется для проведения статистического анализа. По прошествии переходного периода, т.е. постфактум, легко установить по классическим стационарным критериям, что распределение изменилось. Однако на практике важно понять наступление разладки в течение переходного периода, причем с минимальным запаздыванием. Следовательно, объем выборки должен быть заведомо меньше длительности переходного периода, но в то же время достаточным для того, чтобы оценки вероятностного распределения были бы достоверны. Поскольку, как уже говорилось, продолжительность переходного этапа связана в том числе и с объемом выборки, возникает самосогласованная оптимизационная задача по определению индикатора разладки с наибольшей достоверностью.

В математической статистике существует большое количество критериев, применяемых для оценки уровня стационарности или однородности двух выборок, например, критерий Колмогорова [1] о близости выборочного распределения и генеральной совокупности, критерий Колмогорова-Смирнова [1, 2] о близости двух выборочных распределений, критерий тренда в среднем значении Аббе-Линника [2, 3], критерий тренда в дисперсии Фостера-Стюарта [2, 4] и др. Каждая из таких критериальных статистик представляет собой своеобразный индикатор разладки, однако достоверность их для практических нестационарных рядов, таких, как ряды цен на финансовых рынках, низкая. Это связано с тем, что в силу нестационарности возникает большое количество ложных срабатываний

стационарного индикатора, что не позволяет эффективно применять такие статистики на практике. Фактически, уровень значимости такого индикатора является неизвестной функцией времени, что требует его постоянного изменения.

С другой стороны, существуют методы, развитые для анализа собственно нестационарных процессов и основанные на анализе эволюции семейства выборочных распределений процесса, отличающихся объемами выборок. В частности, в работах [5-7] была введена так называемая горизонтная статистика и изучены некоторые ее свойства. В настоящей работе эта статистика рассматривается с точки зрения индикации смены уровня хаоса в системе. На основе горизонтной статистики в данной работе предложены некоторые новые индикаторы разладки, имеющие малое время запаздывания, так что в определенных случаях они могут также трактоваться и как предикторы.

1.2. Актуальность темы

В настоящее время нестационарный анализ данных представляет большую практическую важность и востребованность в различных сферах человеческой деятельности: прогнозирование погоды в метеорологии, анализ функционирования организма человека в биомедицине, прогноз ценовых рядов на финансовых и сырьевых рынках, корректная обработка результатов социологических опросов. Стандартные модели стационарных процессов (регрессионные, автокорреляционные, адаптивные) не отвечают потребностям специалистов в этих областях, поскольку они дают слишком большую ошибку, превосходящую, как правило, ошибку экспертных оценок, которые также далеки от требуемой точности. Поэтому построение статистически корректной модели индикатора нестационарности необходимо для коррекции экспертной оценки. Трудность представляет математически строгое обоснование достоверности статистических оценок. В этой связи даже индикативное указание на изменение функционирования изучаемой системы (рынка акций, общества, отдельного организма) является весьма актуальной задачей. Кроме того, разработка численного метода анализа большого объема данных за конечное время, требуемое для принятия решения, также представляет практическую важность.

Построенные индикаторы разладки могут быть применены к самым разным временным рядам. Например, можно проверить письменный текст достаточной длины (более 30 тыс. знаков) на однородность с целью выявления количества возможных соавторов. Наличие численного критерия однородности позволяет указать не только величины отрывков, написанных, по-видимому, одним человеком, но и дать корректную статистическую оценку достоверности ответа. Эта задача весьма актуальна в литературоведении.

1.3. Новизна разработки

Как правило, для каждого нестационарного временного ряда индикаторы тех или иных его свойств имеют свой специфический вид, не обобщаемый на ряды другого типа. Например, индикатор линейного тренда не особенно эффективен для рядов с квазипериодическим изменением, как и индикатор нестационарности дисперсии для рядов с квазилинейным трендом. Более того, индикаторы, основанные на некоторых средних характеристиках ряда (например, несколько первых моментов), не образуют базисной системы, по которой можно определить тенденцию локального по времени изменения случайного процесса. Для этой цели требуется использовать моменты всех порядков, что не может быть реализовано на практике ввиду конечности выборки. Следовательно, работать надо не с моментами, а с самими выборочными распределениями.

В настоящей работе впервые предлагается применить кинетический подход, развитый в [7], к построению индикаторов разладки для нестационарных временных рядов. В этом подходе используются нестационарные статистики и расстояния между выборочными плотностями функций распределения в различных нормах. В работе введена статистика, названная согласованным уровнем нестационарности, а также статистическая добротность ряда. Эти статистики позволили более точно определить моменты разладки.

2. ГОРИЗОНТНЫЙ РЯД КАК ИНДИКАТОР УРОВНЯ ХАОСА

2.1. Понятие горизонтного ряда

Временной ряд $x(t)$, заданный в дискретные моменты времени, для удобства считающиеся равноотстоящими, называется стационарным в узком смысле, если его плотность распределения $f(x)$ стационарна. Ряд называется стационарным в широком смысле, если его среднее значение постоянно, а автокорреляционная функция зависит только от разности моментов времени. В противном случае при невыполнении любого из вышеприведенных условий ряд называется нестационарным.

Выборочной плотностью функции распределения (далее ВПФР) $f_T(x,t)$ будем называть ВПФР, построенную по выборке данных $x(t), x(t-1), ..., x(t-T+1)$. Расстоянием между двумя ВПФР $f_1 = f_{T_1}(x,t_1)$ и $f_2 = f_{T_2}(x,t_2)$ называется величина [5-7]

$$\rho(f_1, f_2) = \left\| f_{T_1}(x,t_1) - f_{T_2}(x,t_2) \right\| = \int \left| f_{T_1}(x,t_1) - f_{T_2}(x,t_2) \right| dx . \qquad (2.1)$$

Используя (2.1), можно определить вероятность того, что две ВПФР, сдвинутые одна относительно другой на τ шагов, интегрально отличаются не более чем на заданную величину ε. Соответствующий интегральный функционал обозначим через $V(T,\tau;t)$:

$$V(T,\tau;x,t) \equiv \rho\big(f_T(x,t+\tau),\ f_T(x,t)\big) = \int_0^1 \big| f_T(x,t+\tau) - f_T(x,t)\big| dx \le \varepsilon . \quad (2.2)$$

ВПФР $f_T(x,t)$ временного ряда $x(t)$ будем называть ε-стационарной, если выполнено условие (2.2), т.е. если $V(T,\tau;t) \le \varepsilon$.

Для функционала (2.2) близости двух ВПФР имеет место оценка

$$0 \le V(T,\tau;t) \le \min\big(2\tau/T;\ 2\big). \quad (2.3)$$

Неравенство (2.3) дает возможность сделать важный вывод о том, что при фиксированном τ функционал (2.2) равномерно ограничен по *t*. Поэтому $\forall\ \varepsilon > 0$ всегда можно подобрать такой объем выборки $T > 2\tau/\varepsilon$, что ВПФР будет ε-стационарной. Таким образом, если нельзя сравнить выборочную функцию распределения со стационарным распределением, то можно добиться близости двух нестационарных выборочных распределений. В таком подходе центральная задача состоит в эмпирическом нахождении таких параметров ε и τ, при которых достоверность выводов по той или иной конкретной проблеме будет наибольшая. Естественно, это имеет смысл только в том случае, если найденные параметры меняются во времени медленнее, чем практически требуемое время для принятия решения.

Горизонтным рядом для ряда $x(t)$ при сдвиге на промежуток τ называется такой минимальный объем выборки $h\big(t,\tau;\varepsilon\big)$, что при всех $T \ge h(t,\tau;\varepsilon)$ и для всех $k \in [1;\tau]$ выполнено условие $V(T,\tau;t) \le \varepsilon$.

В зависимости от типа случайного процесса горизонтные ряды ведут себя разным образом. Основное их свойство состоит в том, что чем более нестационарен исходный ряд, тем в более широкой полосе изменяется горизонтный ряд. Максимальные значения горизонтного ряда означают полную локальную хаотизацию процесса, а их уменьшение свидетельствует о появлении взаимосвязи между элементами исходного ряда.

2.2. Свойства горизонтной статистики

Введем плотность распределения $\psi_{\tau,\varepsilon}(T)$ значений горизонтного ряда $h\big(t,\tau;\varepsilon\big)$, т.е. определим вероятность того, что расстояние между двумя ВПФР, построенными по выборке объема T и сдвинутыми по времени на фиксированный промежуток τ, не превосходит ε для всех $T' \ge T$. Для краткости зависимость от момента времени *t* в аргументах $\psi_{\tau,\varepsilon}(T)$ опущена. В силу (2.3) достаточно рассмотреть конечное множество значений T', не превосходящих величины $2\tau/\varepsilon$.

Определим интегральную функцию распределения

$$\Psi_{\tau,\varepsilon}(T) = \sum_{n=1}^{T} \psi_{\tau,\varepsilon}(k), \quad \Psi_{\tau,\varepsilon}(2\tau/\varepsilon) = 1. \quad (2.4)$$

Смысл функции распределения (2.4) в том, что если вместо величины $[2\tau/\varepsilon]$ взять некоторое меньшее значение T^*, то с вероятностью $\Psi_{\tau,\varepsilon}(T^*)$

будет выполнено условие ε-стационарности. Выборочные квантили распределения (2.4) могут служить оценками точности в установлении соответствующих объемов выборок.

Отметим, что если ВПФР $f_T(x,t)$ ε-стационарна, то и ВПФР $\psi_{\tau,\varepsilon}(T)$ соответствующего горизонтного ряда также ε-стационарна. Область определения горизонтной статистики представляет собой целочисленный набор точек от 1 до $[2\tau/\varepsilon]$. Эта область разбивается на полосы шириной $[2/\varepsilon]$, внутри которых распределение горизонтного ряда является монотонно возрастающей функцией. Кусочно-монотонный рост распределения в последней (крайней справа) полосе, отвечающей наибольшим значениям горизонтного ряда, имеет определенную специфику. Именно, распределение горизонтного ряда в полосе $\frac{2(n-1)}{\varepsilon} < h(t, n; \varepsilon) \le \frac{2n}{\varepsilon}$ шириной $[2/\varepsilon]$ является равномерным. Доказательство указанных фактов содержится в [7].

Укажем основные типы поведения горизонтной статистики. Для белого шума особенность распределения $\psi_{\tau,\varepsilon}(T)$ состоит в том, что имеется преобладающее количество значений горизонтного ряда, которые равны своему максимальному значению. Напротив, для хаотической динамической системы последняя точка области определения горизонтного ряда не является точкой максимума, что свидетельствует о корреляциях между элементами ряда, хотя линейные корреляции в этой системе так же малы, как и в белом шуме. Для сильно нестационарного ряда существует оптимальный объема выборки, поскольку горизонтный ряд имеет абсолютный максимум во внутренней зоне области определения, заметные колебания в которой также свидетельствуют о наличии значимых корреляций между элементами исходного ряда.

Таким образом, горизонтная статистика позволяет выявить наличие скрытых зависимостей между значениями временного ряда ДС, которые не находятся средствами линейного корреляционного анализа.

2.3. Скачки значений горизонтного ряда как индикаторы разладки

Рассмотрим индикатор изменения гипотетической генеральной совокупности, из которой «как бы» берутся значения нестационарного ряда. Если, например, считать, что в каждый момент времени на бирже эта совокупность порождена ожиданиями игроков, то сигнал о том, что распределение этих ожиданий (неизвестное, заметим, в принципе) изменилось, сам по себе является важным.

Конкретное значение горизонтного ряда показывает, выборкой какого объема определяется вероятностное распределение случайной величины с заданной точностью, и изменение этого распределения. Именно, значение горизонтной статистики в момент времени t показывает, сколько значений исходного ряда надо взять к моменту $t-\tau+1$, чтобы выборочное распределение изменилось за время τ не более чем на фиксированное

значение ε. Вследствие этого большие значения горизонтного ряда свидетельствуют о высокой хаотичности процесса, а относительно малые – о включении упорядоченности.

Как следует из определения горизонтного ряда, этот индикатор имеет фиксированное запаздывание τ. Индикативный смысл этой статистики в том, что она позволяет оценить уровень консолидации участников рынка, т.е. определить возникновение корреляций между значениями временного ряда.

Периодичность промежутков между максимумами и минимумами горизонтного ряда определяет временное расстояние между консолидацией значений исходного временного ряда, а рост свидетельствует о нарастании хаоса в системе, т.е., например, об отсутствии тренда на выбранном временном промежутке. В этом смысле информация о времени, нужном для возникновения согласованного поведения или, напротив, для релаксации, содержится в распределении промежутков между локальными минимумами и максимумами горизонтного ряда и представляет несомненный практический интерес. В этой связи возникают две задачи. Первая – определить характерные промежутки между хаосом и консолидацией, т.е. между очень большими и самыми малыми значениями горизонтного ряда. Вторая задача – определить локальную периодичность в тренде поведения случайного процесса, т.е. распределение промежутков нарастания и убывания хаоса.

Чтобы построить указанные распределения, необходимо, во-первых, найти такие τ и ε, для которых поведение горизонтного ряда обладает нужными свойствами, т.е. информативно с точки зрения исследователя, и, во-вторых, провести уровень отсечки значений, претендующих на «звание» локального экстремума. Последнее требование типично при анализе поведения любой эмпирической статистики случайного процесса: она всегда изрезана, и почти каждое ее значение – точка локального экстремума. Поэтому необходимо исключить те локальные минимумы и максимумы, которые отвечают просто статистическому шуму.

Поскольку из свойств распределения стационарного горизонтного ряда следует выделенность последней полосы значений ширины τ, то на роль максимальных значений может претендовать не только последнее, равное $[2\tau/\varepsilon]$, но и все значения в промежутке $[2\tau/\varepsilon - \tau + 1; 2\tau/\varepsilon]$. К минимальным же значениям отнесем все те, которые меньше ε-квантили распределения горизонтного ряда. Итак, положим

$$H^* = \min h_{\max} = [2\tau/\varepsilon] - \tau, \quad H_1 = \max h_{\min} : \int_0^{H_1} \psi(h)dh = \varepsilon, \qquad (2.5)$$

где $\psi(h)$ есть выборочная плотность распределения горизонтного ряда. Если оказалось, что $H_1 \geq H^*$ (такое возможно, если, например, все значения горизонтного ряда равны $[2\tau/\varepsilon]$ или, наоборот, нет максимальных значений), то анализ ряда не даст интересных результатов. В первом случае ряд сильно нестационарный и требует исключения временной зависимости, а во втором –

напротив, сильно коррелированный, что также требует предобработки. Далее будем полагать, что $H_1 < H^*$.

Проведем прямые $h = H_1$ и $h = H^*$ через область значений горизонтного ряда и определим точки пересечения этих прямых с линией, соединяющей последовательные фактические значения горизонтного ряда, т.е. найдем множество точек пересечения прямых и горизонтной траектории. Эти точки разделят область определения $[1; 2\tau / \varepsilon]$ горизонтной траектории на отрезки, по каждому из которых можно найти глобальный минимум, если он относится к нижней части траектории, и глобальный максимум, если к верхней части. Части траектории, находящиеся выше и ниже определенной прямой, очевидно, перемежаются, так как возможная точка касания прямой и траектории автоматически относится к предыдущему промежутку.

При этом будем применять следующий алгоритм. Укажем первый промежуток по моменту появления экстремума горизонтной траектории, т.е. первый промежуток, содержащий значение, которое либо больше H^*, либо меньше H_1. Пусть для определенности это будет максимум. Он принадлежит некоторой части траектории, находящейся выше прямой $h = H^*$. Если следующая часть траектории, находящаяся ниже прямой $h = H^*$, не пересекает прямую минимумов $h = H_1$, то эта следующая часть объединяется с предыдущей, и к этой объединенной части присоединяется и следующая часть, находящаяся по построению выше прямой $h = H^*$. В этом объединении трех участков траектории ищется максимум. Если одинаковых максимальных значений оказывается несколько, то за единственный максимум принимаем тот, у которого наибольший аргумент. Аналогично поступаем с нахождением локального минимума. В результате получаем совокупность чередующихся максимумов и минимумов траектории.

Для изучения текущей динамики нарастания или уменьшения доли хаоса в системе достаточно одной прямой, проведенной на уровне среднего значения горизонтного ряда. Алгоритм при этом несколько меняется, поскольку требуется исключить эффект статистического шума. Для этого исключения потребуем, чтобы величина размаха между соседними максимумом и минимумом была бы больше 2τ. Если размах оказался не превосходящим этой пороговой величины, то соответствующий промежуток объединяется с предыдущим.

Рассмотрим применение описанной методики на примере горизонтного ряда, отвечающего приращениям валютного курса евро/доллар (рис. 1-2). Отметим, что широкая полоса изменения значений горизонтного ряда свидетельствует о нестационарности, а спад в крайней правой точке – по аналогии с динамическими системами, о наличии дальней нелинейной автокорреляционной связи (нелинейной – потому, что линейная часть близка к нулю).

Для этого ряда размах составил величину 127, линия максимумов проходит на уровне 390, а линия минимумов – на уровне $H_1 = 308$, чему отвечает 0,05-квантиль горизонтного распределения. Среднее значение горизонтного ряда оказалось равным $h_0 = 349$.

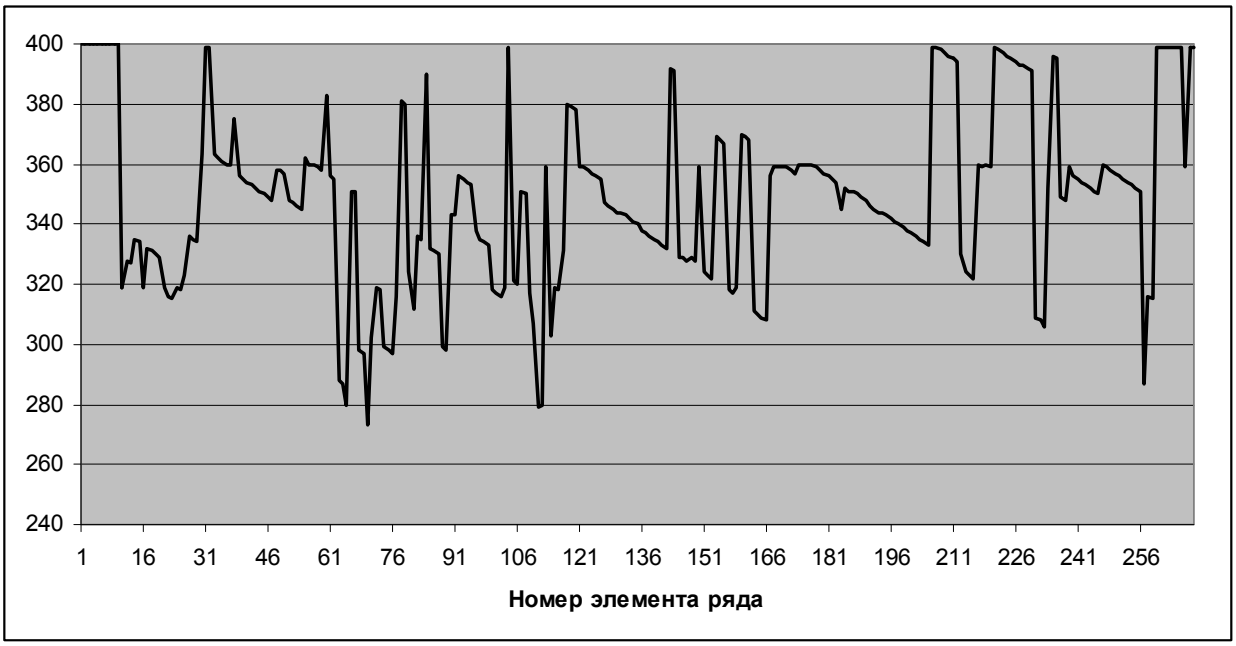

Рис. 1 – Горизонтный ряд для суточных данных курса евро/доллар, $\tau = 10$, $\varepsilon = 0{,}05$

Рис. 2 – Распределение горизонтного ряда курса евро/доллар

Выяснилось, что для рассматриваемого ряда переход от хаоса к консолидации, характеризуемый расстоянием по времени между максимумом и соседним минимумом, меньшим, чем 308, требует от 7 до 100 дней, максимум распределения расстояний приходится на промежуток 20-25 дней.

Обратный же процесс рассогласования идет от 4 до 40 дней с максимумом в промежутке 30-35 дней. Эти наблюдения относятся к долгосрочной динамике. Краткосрочное поведение описывается локальными экстремумами траектории относительно средней линии $h = h_0$. Это поведение также несимметрично. Чаще всего (с вероятностью 0,38) консолидация занимает промежуток от 3 до 5 дней, а хаотизация с вероятностью 0,44 происходит за 1-2 дня, хотя возрастание горизонтного ряда в течение 3-5 дней также имеет достаточно высокую вероятность, равную 0,33. Вероятность перехода от хаоса к порядку не более чем за 5 дней составила 0,50, а в обратную сторону за тот же период оказалась существенно выше – 0,77. При этом средний скачок горизонтного ряда в обоих направлениях оказался равным одной и той же величине 69.

Заметное изменение параметров найденного вероятностного распределения будет свидетельствовать о том, что система предпочтений участников рынка изменилась. В этом и состоит индикативная роль горизонтной статистики.

2.4. Проверка текста на однородность

Рассмотрим литературный текст как последовательность букв, исключая пробелы, знаки препинания и цифры. В книге [8] изучалось распределение расстояний (т.е. количество символов) между одинаковыми буквами в тексте. Оказалось, что это квазистационарный ряд, одинаково распределенный (по гамма-распределению) для любой буквы алфавита, и близкий к белому шуму. Возникло предположение, что вхождение соавтора изменит горизонтный ряд указанной статистики, и место этого вхождения можно будет указать с точностью до отрывка текста, содержащего τ одинаковых символов. Для гласных букв «о», «а», «и», «е» такой отрывок будет иметь минимальную длину, поскольку эти буквы наиболее часто встречаются в тексте.

Оказалось, что существует диапазон параметров $\tau = 10 \div 15$ и $\varepsilon = 0{,}04 \div 0{,}06$, при которых текст, написанный одним писателем, имеет такую последовательность расстояний между одинаковыми гласными, горизонтный ряд которой не имеет максимальных значений, что можно трактовать как реализацию хаотической динамической системы, специфической для каждого автора. В то же время в произведениях, написанных совместно, такие максимальные значения, равные $[2\tau / \varepsilon]$, встречались довольно часто, хотя в этом значении распределение горизонтного ряда по-прежнему не имело максимума.

Примеры горизонтных рядов для иллюстрации описываемой картины приведены на рис. 3-4 для значений $\tau = 10,\ \varepsilon = 0{,}05$.

Был проведен тестовый эксперимент, когда в один текст были соединены отрывки 100 произведений длиной в 15 тыс. знаков, написанные разными писателями. Места склейки отрывков предполагалось опознать по скачкам горизонтного ряда до своих максимальных значений. Результат тестирования дал нулевую эмпирическую вероятность ошибки первого рода (не срабатывание индикатора), т.е. все скачки горизонтного ряда отвечали

странице склейки отрывков, и лишь 2 отрывка из 100 не были опознаны как написанные по-разному, т.е. ошибка второго рода (отклонение правильного ответа) оценивается на уровне 0,02.

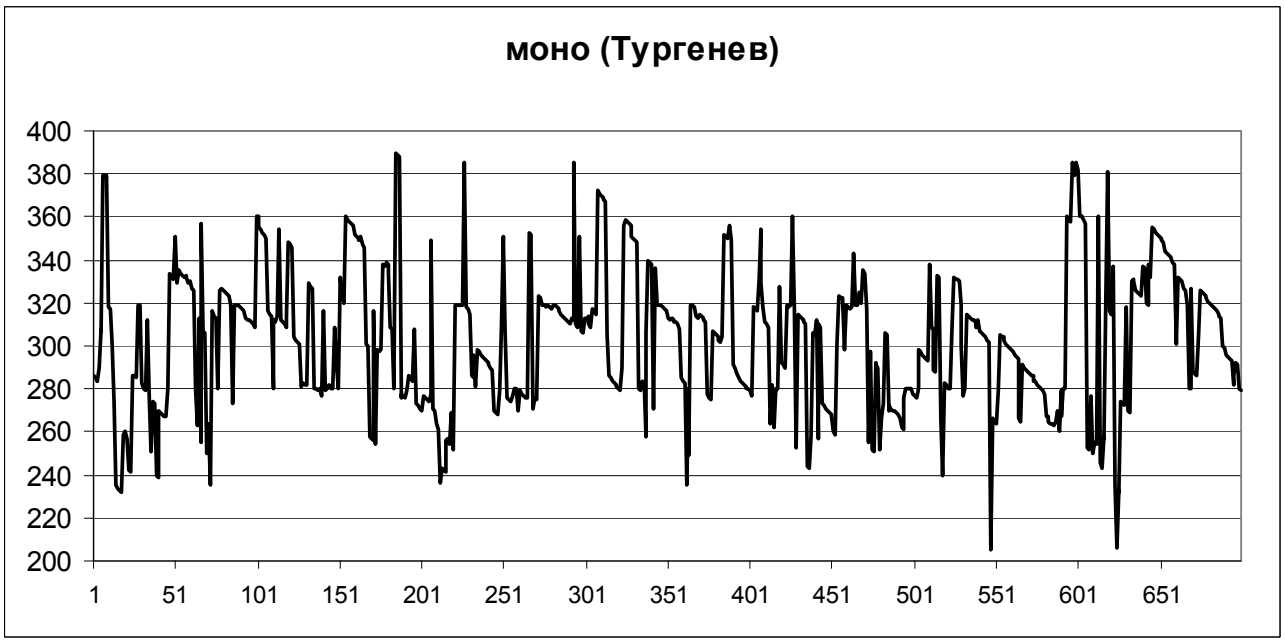

Рис. 3 –Горизонтный ряд расстояний «О-О» для одного писателя

Рис. 4 –Горизонтный ряд расстояний «О-О» для писательских тандемов

Так как тестирование показало высокую достоверность индикатора, этим методом можно попытаться определить, какие фрагменты совместного текста написаны разными людьми. В качестве примера было рассмотрено произведение «Понедельник начинается в субботу», написанное совместно А. и Б. Стругацкими. Места вхождения соавторов отмечались скачками

горизонтного ряда до максимальных значений, в данном случае до 400. После того, как текст был разбит таким образом на отрывки, написанные, предположительно, разными людьми, была сделана косвенная проверка правильности сделанного разбиения. Для этого был применен метод, описанный в [8]. Отрывок считается написанным тем из авторов, к эталонной функции распределения текста по буквам которого он ближе в норме $L1$. Поскольку каждый из соавторов имел и собственные произведения, по ним были составлены эталонные (средние) функции распределения отдельно для А. Стругацкого – «Дни кракена» (105 тыс. букв) и «Экспедиция в преисподнюю» (327 тыс. букв), а также для Б. Стругацкого – «Теорема этики» (539 тыс.) и «Бессильные мира сего» (385 тыс.).

Оказалось, что в среднем каждым из авторов были написаны отрывки объемом примерно 10 тыс. знаков (10 страниц). Косвенная проверка точности такой идентификации дала ошибку 0,06: нарушений порядка следования отрывков, авторство которых было определено сравнением с эталоном, оказалось 2 из 31 отрывков тестируемого текста.

3. СТАТИСТИЧЕСКАЯ ДОБРОТНОСТЬ ВРЕМЕННОГО РЯДА

3.1. Согласованный уровень нестационарности

Статистикой, прямо связанной с горизонтным рядом, но имеющей несколько другой акцент на способе анализа нестационарного поведения изучаемой системы, является распределение значений функционала $V(T,\tau;t)$ при заданных параметрах объема выборки T и окна сдвига τ. Особый интерес представляет анализ распределения значений расстояния между двумя выборками одинакового объема, расположенными «встык», т.е. когда окно сдвига равно объему выборки: $\tau = T$. Оказалось, что имеется достаточно хорошее соответствие между областью выраженного тренда цены и пребывания величины функционала в зоне стационарности, а также соответствие пика функционала и бокового тренда ценового ряда.

По аналогии с горизонтным рядом, малые значения V свидетельствуют о наличии взаимосвязи между значениями исходного ряда, а большие – о хаотичности процесса. Смысл этого индикатора, вычисляемого в скользящем окне, в том, что в зоне тренда он относительно мал, поскольку поведение системы (рынка) консолидировано, а на «боковике» хаос значительно возрастает.

Достаточно большая выборка значений $V(T,T;t)$ позволяет построить распределение расстояний между выборками исходного ряда и, что не менее важно, найти характерные временные расстояния между пиковыми значениями функционала и его квазистационарным уровнем. Эти расстояния покажут время, по истечении которого за ростом волатильности последует тренд. В тех случаях, когда это расстояние больше, чем объем выборки T, индикатор $V(T,T;t)$ может служить предиктором состояния рынка. Это, конечно, определенное идеалистичное представление, но в среднем

сопоставление значений функционала режимам «хаос-порядок» оправдывается с вероятностью, превышающей 0,8.

Распределения значений $V(T,T;t)$ для одного и того же инструмента качественно не особенно сильно зависят от объема выборки. При относительно небольших объемах T распределение довольно заметно изрезано, но с увеличением T оно стабилизируется. На рис. 5 показано распределение $g_T(V)$ значений функционала при разных объемах выборки.

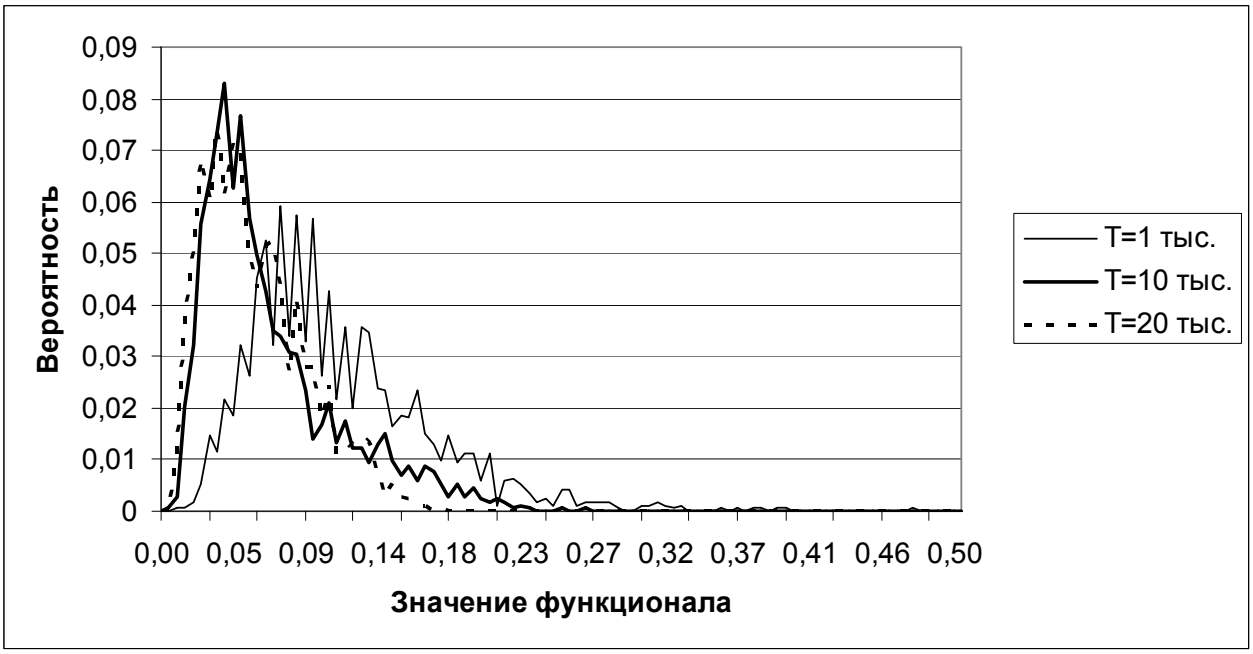

Рис. 5 – Распределение функционала для разных объемов выборок

Однако для разных временных рядов распределение $g_T(V)$ при одном и том же значении T может заметно различаться. Чем более стационарен ряд, тем левее расположен максимум распределения функционала. На рис. 6 приведены распределения функционала, построенного по объемам выборок 5 тыс. тиков для двух ценовых рядов. Первый ряд условно назван стационарным («Стац.»), поскольку для него доля стационарно объясняемых отклонений между ВПФР составила примерно 0,6, а второй – нестационарный, для него эта доля около 0,2.

Поскольку функционал представляет собой меру нестационарности ВПФР, то следует указать характерное значение $V_S(T)$, превышение которого наиболее достоверно определяет момент роста волатильности. Это значение зависит от распределения $g_T(V)$ и само по себе является индикатором нестационарности определенного инструмента. В качестве такого индикатора выберем половинный уровень нестационарности (поскольку функционал нормирован на 2), вероятность превышения которого равна ему самому. Именно, для каждого объема выборки T согласованным уровнем нестационарности будем называть такое значение $V = V_S(T)$, при котором

$$\int_0^{V_S(T)} g_T(V)dV = 1 - \frac{V_S(T)}{2}. \qquad (3.1)$$

Например, для ряда фьючерсных цен на нефть распределение значений этого функционала дано на рис. 6 тонкой линией, отвечающей объему выборки 1 тыс. тиков, согласованное значение уровня отсечки, вычисленное по формуле (3.1), оказалось равным 0,188. Все значения функционала, меньшие этого уровня, «недостаточно нестационарны» для того, чтобы служить индикатором роста волатильности.

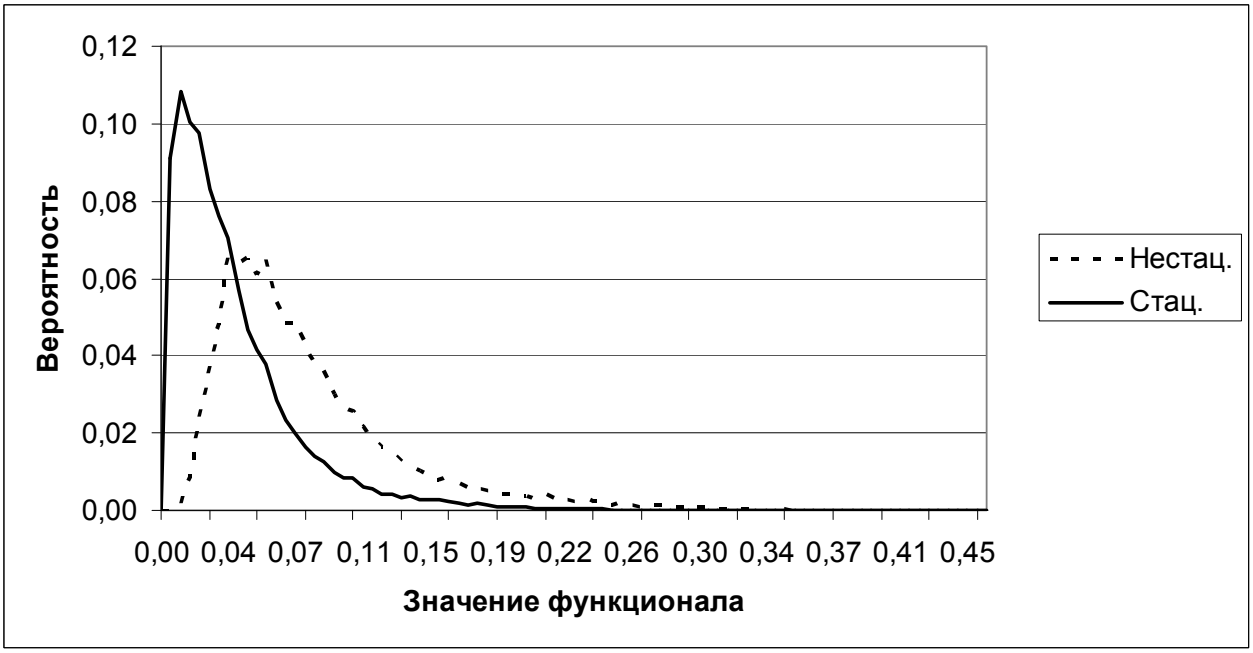

Рис. 6 – Распределение функционала для разных рядов при $T = 5$ тыс. тиков

Введенный согласованный уровень нестационарности $V_S(T)$, как и рассмотренный в предыдущем параграфе горизонтный ряд, может рассматриваться как индикатор нарастания хаоса. Отличие функциональной статистики от горизонтной в том, что горизонтный ряд принципиально строится для выборок, имеющих определенный временной нахлест, а распределение значений функционала можно строить вообще для любых выборок, в том числе имеющих и пустое пересечение по времени.

3.2. Доля стационарно объясняемых отклонений

Мера уклонения ВФР $F_T(x)$ от теоретического распределения $F(x)$ дается следующим утверждением. Если ряд стационарный, то при $T \to \infty$ разность $F_T(x) - F(x)$ распределена асимптотически нормально с параметрами $\mu = 0,\ \sigma^2 = F(x)\big(1 - F(x)\big)/T$. Асимптотическая нормальность означает, что при больших объемах выборки плотность вероятности величины $z = F_T(x) - F(x)$ является гауссовой:

$$P\big(z \in (a, a+\delta)\big) = \int_a^{a+\delta} f_G(z;\mu,\sigma)dz, \quad f_G(z;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right). \quad (3.2)$$

Однако величина уклонения неравномерна по x. Для получения равномерных оценок уклонения рассматривается статистика Колмогорова $D_T = \sup_x |F_T(x) - F(x)|$. Распределение этой статистики дается теоремой Колмогорова. Если теоретическое распределение $F(x)$ генеральной совокупности непрерывно, то ВФР статистики $\sqrt{T}D_T$ сходится при $T \to \infty$ к функции Колмогорова $K(z)$ [1]:

$$\lim_{T\to\infty} P\left\{0 < \sqrt{T}\sup_x |F_T(x) - F(x)| < z\right\} = K(z) = \sum_{k=-\infty}^{\infty}(-1)^k \exp\left(-2k^2z^2\right). \quad (3.3)$$

Обычно распределение генеральной совокупности не бывает известно. Тогда, в предположении, что оно существует, для изучения вопроса принадлежности двух ВФР одному и тому же распределению генеральной совокупности применяется статистика Смирнова $D_{m,n} = \sup_x |F_{1,m}(x) - F_{2,n}(x)|$, построенная по двум выборкам объемов m и n. Для этой статистики справедливо следующее утверждение. Пусть проводятся две независимых серии испытаний по составлению выборок объемов m и n из некоторой генеральной совокупности. Тогда ВФР статистики $\sqrt{mn/(m+n)}D_{m,n}$ сходится при $n \to \infty$ к функции Колмогорова в смысле формулы (3.3). Это означает, что если для двух ВФР $F_1(n,x)$ и $F_2(n,x)$ было найдено значение $S_n = D_{n,n}$ и вычислена величина $z = \sqrt{\frac{n}{2}}S_n$, то величина $1 - K(z)$ приближенно считается равной вероятности того, что $\sqrt{\frac{n}{2}}S_n \ge z$. Если величина $1 - K(z)$ мала, то осуществилось маловероятное событие, несовместимое с понятием случайности, и эти выборки следует считать различными. Некоторая неопределенность вывода состоит в том, что надо априори задать желаемый уровень малости критерия $1 - K(z)$. Какую вероятность считать достаточной для того, чтобы признать выборки одинаковыми? Поскольку замена статистики S на статистику V в критерии (3.3) не уменьшает достоверности оценки вероятности нестационарности, то согласованным уровнем стационарно объясняемых отклонений между ВПФР, построенных по объемам выборки n, естественно считать такое значение $z = V_0(n)$, при котором вероятность того, что расстояние между двумя выборками с объемами n больше V_0, по стационарному критерию (3.3) равна $V_0/2$, т.е.

$$1 - \frac{V_0}{2} = K\left(\sqrt{\frac{n}{2}}V_0\right). \quad (3.4)$$

Численное решения уравнения (3.4) относительно согласованного уровня стационарности V_0 показывает, что, например, при сравнении двух распределений, построенных по выборкам объемом в 1000 данных каждая, вероятность того, что расстояние между ними будет меньше 0,065, равна приблизительно 0,97. Именно эта вероятность согласована с уровнем стационарности. Для выборок объемом 60 тыс. данных уровень стационарно объясняемых отклонений равен приблизительно 0,01 с согласованной вероятностью 0,995.

3.3. Статистическая добротность ряда

Эффективность введенного в п. 3.2 согласованного уровня нестационарности $V_S(T)$ как индикатора нарастания хаоса зависит от объема выборки T. При малых объемах индикатор сам является весьма нестационарным. При больших же объемах ряд функционала становится излишне стационарным. Следовательно, требуется оптимизировать объем выборки, по которому надо вычислять индикатор. Сформулируем подходящий критерий оптимизации.

При анализе нестационарного поведения ряда превышение индикатором определенного порогового значения $V_S(T)$ является полезным сигналом. Этот сигнал проявляется на фоне квазистационарного шума. Очевидно, чем выше проведена отсечка V_S, тем меньше вероятность принять шум за полезный сигнал. С другой стороны, если провести уровень V_S слишком высоко, то будут ошибочно отвергнуты многие полезные сигналы. Таким образом, наряду с $V_S(T)$ надо рассмотреть еще две характеристики временного ряда функционалов: максимальное значение $V_m(T)$ по выборке объема T и уровень $V_0(T)$, который может быть объяснен в рамках стационарной модели. Интеграл

$$\eta(T) = \int_0^{V_0(T)} g_T(V)dV \tag{3.5}$$

дает долю значений функционала, находящихся в области V_0-стационарности.

Введем индикатор $Q(T)$, называемый статистической добротностью исходного временного ряда. Он показывает отношение доли согласованной нестационарной части распределения функционала к доле согласованной стационарной части. Именно, для каждого объема выборки T «статистической добротностью» $Q(T)$ индикатора $V(T)$ будем называть отношение

$$Q(T) = \frac{\int_{V_S(T)}^{2} g_T(V)dV}{\int_0^{V_0(T)} g_T(V)dV} = \frac{V_S(T)}{2\eta(T)}. \tag{3.6}$$

Смысл добротности (3.6) в том, что если полезным сигналом является высокий уровень нестационарности, то он должен четко выделяться на фоне стационарного шума. Заметим, что если процесс стационарный, то $V_S = V_0$, $\eta = V_0 / 2$, так что добротность равна единице: $Q_{st}(T) = 1$. Следовательно, само по себе значение добротности может служить мерой нестационарности процесса.

Недостаток индикатора (3.6) в том, что если процесс существенно нестационарный, то у него не окажется стационарно объясняемой части значений, т.е. добротность при любом объеме выборки будет велика, что не позволит провести оптимизацию этого объема. В таком случае полезно ввести второй индикатор, призванный снизить ошибку отклонения верной гипотезы. Будем называть его чувствительностью и обозначать $r(T)$.

Чувствительность статистики по величине нестационарности показывает, какую часть от максимального уровня нестационарности составляет значение уровня отсечки. Если величина индикатора слишком большая, то некоторые события будут ошибочно игнорированы, т.е. отношение

$$r(T) = \frac{V_S(T)}{V_{\max}(T)} \tag{3.7}$$

показывает, какая доля области значений функционала (безотносительно к его распределению) будет игнорирована критерием $V_S(T)$. Чем меньше величина чувствительности (3.7), тем меньше ошибочных отклонений верного сигнала допустит критерий отсечки $V_S(T)$.

Однако и у индикатора (3.7) есть недостаток, проявляющийся в области слабой нестационарности: если V_S будет приближаться к V_0, то чувствительность будет мала, но будет и большое количество ошибок, связанных с принятием стационарного шума за полезный сигнал.

Следовательно, выбор оптимального объема T_{opt} данных для построения функционала разности ВПФР основывается на следующей идее. Согласованное значение $V_S(T)$ индикатора должно быть, с одной стороны, достаточно удалено от стационарного уровня $V_0(T)$, но также должно быть и заметно меньше, чем $V_{\max}(T)$. Доля области значений полезного сигнала равна $1 - r(T)$, а относительная полезность его определяется добротностью $Q(T)$. Одним из простых условий оптимальности может быть требование максимальности величины

$$G(T) = Q(T)\big(1 - r(T)\big). \tag{3.8}$$

Тогда положим

$$T_{opt} = \arg\max Q(T)\big(1 - r(T)\big). \tag{3.9}$$

Соответствующее согласованное значение функционала будем считать оптимальным уровнем индикатора:

$$V_{opt} = V_S(T_{opt}). \tag{3.11}$$

Например, для фрагмента ряда цен на нефть WTI оптимизируемая функция $G(T)$ имеет вид, показанный на рис. 7.

Рис. 7 – Оптимизируемая функция для определения объема выборки

Из рис. 7 следует, что наилучшим объемом для использования функционала $V(T,T)$ как индикатора смены локального поведения ряда является в этом примере объем $T_{opt} = 15$ тыс. тиков. Для другого временного участка этот объем будет, вообще говоря, другим. Изменение оптимального объема само является индикатором – но не локального, а более крупномасштабного поведения временного ряда. Индикатор, вычисляемый в скользящем окне с шагом, на порядок меньшим оптимального объема, позволяет оперативно принимать решения по текущей ситуации.

ЛИТЕРАТУРА

1. Гнеденко Б.В. Курс теории вероятностей. – М.: Физматлит, 1961. – 406 с.

2. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.

3. Каган А.М., Линник Ю.В., Рао С.Р. Характеризационные задачи математической статистики. – М.: Наука, 1972.

4. 84. Foster F.G., Stuart A. A distribution-free test in time series dated on the breaking of records // JRSS, 1954. V. B16, 1, p. 1-22.

5. Орлов Ю.Н., Осминин К.П. Методика определения оптимального объема выборки для прогнозирования нестационарного временного ряда. // ИТВС, 2008, № 3, с. 3-13.

6. Орлов Ю.Н., Осминин К.П. Построение выборочной функции распределения для прогнозирования нестационарного временного ряда. // Мат. Мод., 2008, № 9, с. 23-33.

7. Орлов Ю.Н., Осминин К.П. Нестационарные временные ряды: методы прогнозирования с примерами анализа финансовых и сырьевых рынков. – М.: Editorial URSS, Книжный дом «ЛИБРОКОМ», 2011. – 384 с.

8. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Editorial URSS, Книжный дом «ЛИБРОКОМ», 2012. – 312 с.