Построение кросс-корреляционных зависимостей при прогнозе загруженности железнодорожного узла*

Вальков А. С.,[1] ***Кожанов Е. М.,***[2] ***Мотренко А. П.,***[3] ***Хусаинов Ф. И.***[4]
valkov@forecsys.ru, vinger4@gmail.com, anastasia.motrenko@gmail.com, f-husainov@yandex.ru
1 — Вычислительный центр РАН, 2 — Московский государственный технический университет имени Н. Э. Баумана, 3 — Московский физико-технический институт, 4 — Российская открытая академия транспорта Московского государственного университета путей сообщения (МИИТ)

Рассматривается проблема обнаружения причинно-следственных связей в разнородных временных рядах. Предлагается прогностическая модель, использующая выявленные связи. Модель предназначена для прогнозирования загруженности железнодорожного узла. Модель использует как исторические данные о загруженности, так и внешние данные: биржевые цены на основные инструменты и нормативные документы. При построении модели используются экспертные высказывания относительно вида связей. Предложен метод оценки достоверности экспертных высказываний. Метод проиллюстрирован данными грузовых перевозок РЖД.

Ключевые слова: *временные ряды, прогнозирование, загрузка железнодорожного узла, прогностическая модель, экспертное высказывание.*

Constructing a cross-correlation model to forecast the utilization of a railway junction station*

Valkov A. S.,[1] ***Kozhanov E. M.,***[2] ***Motrenko A. P.,***[3] ***Husainov F. I.***[4]
1 — Computing Center of the Russian Academy of Sciences; 2 — Bauman Moscow State Technical University; 3 — Moscow Institute of Physics and Technology; 4 — Moscow State University of Railway Engineering

The problem of detecting causal relationships between time series is studied. The authors propose a forecasting model that considers detected relationships. The model is aimed to forecast the utilization of a railway junction station. The model relies on the history of a junction station utilization as well as on the time series for the main financial instruments and regulations. Expert's assessments are used to construct the model. A method that evaluates plausibility of the expert's assessments is proposed. The method is illustrated with the Russian Railways data.

Keywords: *time series, forecasting, utilization of a railway junction station, forecasting model, expert assessment.*

Введение

При прогнозировании грузоперевозок необходимо учитывать как предысторию самих перевозок, так и различные внешние факторы. Так как множество внешних факторов, влияющих на объемы перевозимых грузов и, как следствие, на загруженность железнодорожных узлов, велико, то для выбора факторов предполагается использовать экспертные оценки. Эксперты высказывают гипотезы о влиянии внешних факторов на грузоперевоз-

Работа выполнена при финансовой поддержке РФФИ, проект № 12-07-31095.

ки. Требуется проверить эти гипотезы, обнаружить влияние внешних факторов и оценить достоверность экспертных суждений.

В данной работе экспертные высказывания о влиянии внешних событий на грузоперевозки представлены в табл. 1. В правой колонке перечислены факторы, которые с точки зрения экспертов оказывают влияние на объем грузоперевозок. В центральной колонке перечислены группы груза, соответствующие этим факторам. В левой колонке перечислены высказывания эксперта о характере влияния. В табл.2 перечислены факторы, влияние которых подтвердить или опровергнуть затруднительно из-за недостаточного количества данных.

Экспертные высказывания носят качественный характер и представлены в номинальных (оказывает данный фактор влияние на загруженность железнодорожного узла или нет) и ранговых шкалах (степень влияния — высокая, низкая, фактор не оказывает влияния). В связи с возможным несоответствием экспертного суждения измеряемым данным или с возможной несогласованностью [1] высказываний нескольких экспертов предлагается решить задачу определения достоверности экспертного высказывания и разработать алгоритм получения оценки достоверности. Методы получения экспертных оценок и высказываний и проверки их непротиворечивости описаны в [2].

Таблица 1: Факторы экономического и производственного характера, влияющие на динамику перевозок.

№	Вид фактора, влияющего на объем грузоперевозок	Группы грузов и отрасли, на которые оказывается влияние	Степень и характер влияния
1	Мировые и внутренние цены на соответствующие активы	Нефть и нефтепродукты, черные металлы, цветные металлы, удобрения, уголь и др.	На экспортные перевозки влияние сильное. Связь бывает как прямой, так и обратной.
2	Курс рубля к доллару	Грузы, отправляемые на экспорт (нефть и нефтепродукты, металлы, уголь)	Степень влияния для экспортных перевозок зачастую высокая.
3	Сезонность производства природного-климатического характера	Зерно, овощи, бахчевые культуры	Степень влияния высокая. Динамика перевозки связана со сбором урожая. Повышенный объем перевозок приходится на июль, август и сентябрь, иногда на октябрь.
4	Сезонность спроса на продукцию	Строительные грузы (щебень, кирпич, цемент, промсырьё)	Степень влияния высокая. Динамика перевозки связана с сезоном строительных работ. Пик приходится на летние месяцы.
5	Особенности технологического цикла в различных отраслях	Различные грузы. Например, в соледобыче в период “соленавигации” высокий уровень добычи, а в остальные периоды зависит от величины складских запасов	Во многих случаях являются ограничителем (как верхней, так и нижней границей) объема погрузки грузов.

Таблица 2: Нерассматриваемые факторы экономического и производственного характера, влияющие на динамику перевозок.

№	Вид фактора, влияющего на объем грузоперевозок	Группы грузов и отрасли, на которые оказывается влияние	Степень и характер влияния
1	Сезонность, связанная с навигацией	Грузы, которые перевозятся водным транспортом	Степень влияния высокая для производств расположенных в пределах досягаемости судоходных путей.
2	Производственные мощности заводов и других предприятий	Добывающие отрасли (угледобыча, добыча соли, камня гипсового, добыча нефти); нефтеперерабатывающие, нефтехимические, металлургические производства.	Во многих случаях являются верхней границей объема погрузки грузов.
3	Запасы готовой продукции на складах предприятий	—	Во многих случаях динамика этого показателя являются мерой спроса на продукцию данного предприятия, а спрос в свою очередь оказывает влияние на планируемые объемы производства.
4	Доли погрузки предприятия, приходящиеся на другие виды транспорта	Все грузы, перевозка которых возможна альтернативными видами транспорта	На сверхдальних расстояниях перевозки влияние низкое, для массовых грузов низкая; для остальных грузов, особенно на короткие и средние расстояния.
5	Мощности станций погрузки	Все грузы	Высокое в случаях увеличения производственных мощностей предприятий-грузоотправителей или низких мощностей станции или низких перерабатывающих способностей грузовых фронтов.

Экспертное утверждение о влиянии некоторого фактора на загруженность железнодорожного узла интерпретируется как утверждение о причинно-следственной связи между временными рядами. Таким образом, для оценки достоверности экспертного высказывания необходимо исследовать ряды на наличие причинно-следственной связи.

Способы обнаружения причинно-следственной связи исследованы в [?] и широко распространены [3, 4]. Подход Грейнджера, принятый в этих работах, основан на сравнении точности прогноза исследуемого временного ряда исключительно по его истории и при наличии информации о других временных рядах. В случае, если улучшение прогноза подтверждается статистически, говорят о G-зависимости [5] временных рядов. При этом никакой информации о структуре зависимости между рядами получить не удается. Например, если изменения двух исследуемых рядов вызваны третьим временным рядом, о котором исследователь не знает, ряду будут признаны G-зависимыми, хотя причинно-следственной связи между ними нет. В работе [6] описаны способы расширения подхода Грейнджера для обнаружения структуры связей между временными рядами.

Альтернативный подход, моделирование структурных уравнений (structural equation modelling), рассмотрен в работах [7, 8]. В этом случае исследователь строит модель, основываясь на своих предположениях о структуре причинно-следственных связей между измеренными временными рядами, а также рядами, значения которых по каким-либо причинам неизвестны, а затем настраивает ее в соответствии с данными. В работе [9] обсуждается возможность, сделав вывод о наличии связи между временными рядами, перенести этот результат на временные ряды того же характера, но измеренные в других условиях.

Метод сходящегося перекрестного отображения (англ. convergent cross mapping, CCM) [10, 11], предложенный в 2012 году позволяет определить, что временные ряды принадлежат одной динамической системе. Этот метод был разработан для выявления причинно-следственных связей в случаях, когда тест Грейнджера неприменим или не может обнаружить связи. Метод основан на преобразовании пространства состояний системы, и сравнении ближайших соседей одной и той же точки в преобразованных системах и заключается в проверке сходимости коэффициента корреляции между спрогнозированными и исходными значениями исследуемого ряда при увеличении объема выборки.

Постановка задачи

Задано множество временных рядов $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$, в котором временной ряд $\mathbf{s}_i \in \mathbb{Z}^n,\ i \in \mathcal{I}$ — временной ряд, состоящий из элементов, соответствующих числу вагонов, проходящих через станцию. *Положительное значение элемента вектора s_{ij} означает приходящий вагон, отрицательное — отправляющийся.* Каждому $s_{ij},\ j \in \mathcal{J} = \{1, \ldots, n$ поставлены в соответствие две метки $g(s_{ij}) \in G$ и $h(s_{ij}) \in H$ — тип вагона и вид перевозимого груза.

Множество индексов $\mathcal{I}$ соответствует множеству железнодорожных узлов (станций), а множество $\mathcal{J}$ соответствует множеству отсчетов времени.

Задано множество внешних факторов $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, в котором временной ряд $\mathbf{x}_i \in \mathbb{R}^n,\ i \in \mathcal{I}$. Задан набор экспертных высказываний μ о влиянии внешних факторов $\mathbf{x}$ на временные ряды $\mathbf{s}$,

$$\mu = \mu(\mathbf{x}, \mathbf{s}) \in \{\text{«+»}, \text{«-»}\}.$$

Требуется проверить экспертные высказывания и поставить в соответствие каждому высказыванию значение достоверности.

Выявление G-зависимости между временными рядами

Тест Грейнджера применяется при прогнозировании с помощью линейных регрессионных моделей. Пусть $\mathbf{s}$ и $\mathbf{x}$ — исследуемые временные ряды, тогда линейная регрессионная модель имеет вид

$$\hat{s}_j = \sum_{t=1}^{\tau} a_t s_{j-t} + \sum_{t=1}^{\tau} b_t x_{j-t} + \varepsilon_j, \quad j \in \mathcal{J}, \tag{1}$$

$$\hat{x}_j = \sum_{t=1}^{\tau} c_t s_{j-t} + \sum_{t=1}^{\tau} d_t x_{j-t} + \xi_j \quad j \in \mathcal{J}. \tag{2}$$

Здесь τ – количество предыдущих значений, принимаемых во внимание, ряды $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ содержат веса учитываемых объектов, а ε_j и ξ_j – ошибки прогнозирования. Будем считать, что временной ряд $\mathbf{x}$ влечет за собой временной ряд $\mathbf{s}$, если модуль ошибки $\varepsilon_j = s_j - \hat{s}_j$ прогнозирования ряда $\mathbf{s}$ уменьшается при включении в модель значений ряда $\mathbf{x}$. Для проверки значимости уменьшения ошибки при добавлении в модель ряда $\mathbf{x}$ воспользуемся

статистикой Фишера. Пусть для ряда $\mathbf{s}$ построена модель (1), учитывающая историю ряда $\mathbf{x}$. Построим также линейную регрессионную модель, основанную только на истории $\mathbf{s}$:

$$\tilde{s}_j = \sum_{t=1}^{\tau} a_t s_{j-t} + \tilde{\varepsilon}_j, \quad j \in \mathcal{J}. \tag{3}$$

Тогда применение теста Грейнджера сводится к вычислению статистики

$$F = \frac{(\mathrm{RSS}_{\tilde{\mathbf{s}}} - \mathrm{RSS}_{\hat{\mathbf{s}}})/\tau}{\mathrm{RSS}_{\hat{\mathbf{s}}}/(|J| - 2\tau)} \tag{4}$$

и сравнению ее с критическим значением при заданном уровне значимости. Здесь

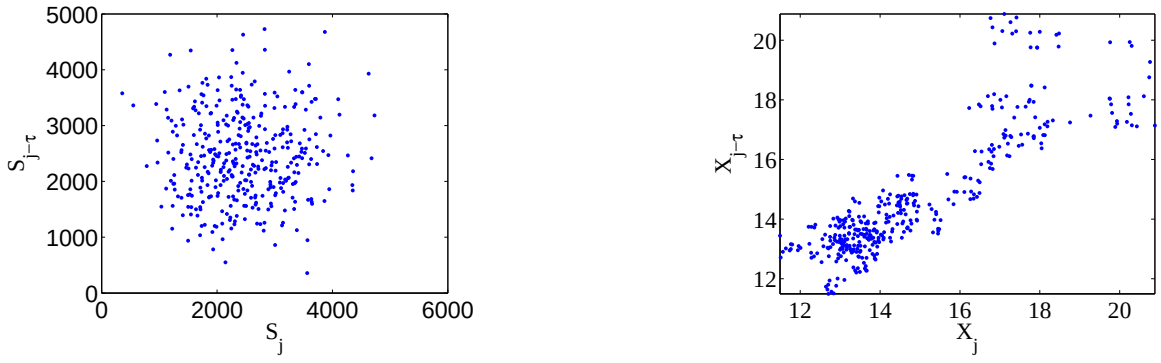
$$\mathrm{RSS}_{\hat{\mathbf{s}}} = \sum_{j\in\mathcal{J}} \varepsilon_j^2 = \sum_{j\in\mathcal{J}} (s_j - \hat{s}_j)^2,$$

$$\mathrm{RSS}_{\tilde{\mathbf{s}}} = \sum_{j\in\mathcal{J}} \tilde{\varepsilon}_j^2 = \sum_{j\in\mathcal{J}} (s_j - \tilde{s}_j)^2.$$

Нулевая гипотеза заключается в предположении, что ряд $\mathbf{x}$ не оказывает влияния на значения ряда $\mathbf{s}$. При нулевой гипотезе F принадлежит распределению Фишера со степенями свободы τ и $|J| - 2\tau$. Полученную вероятность отклонить нулевую гипотезу $1 - p(\mathbf{s}, \mathbf{x})$, где $p(\mathbf{s}, \mathbf{x})$ — критическое значение F-статистики будем интерпретировать как достоверность экспертного высказывания $\mu(\mathbf{s}, \mathbf{x})$.

Исследование данных о влиянии цен на основные инструменты на перевозки соответствующих групп грузов приведено ниже, в разделе «Вычислительный эксперимент».

Тест Грейнджера применим к рядам, обладающим не зависящими от времени матожиданием и дисперсией. Если исследуемый ряд не обладает этими свойствами, необходимо привести его к соответствующему виду. Кроме того, в самом определении G-зависимости, данном Грейнджером, заключается предположение о разделимости исследуемых временных рядов. Под разделимостью рядов здесь имеется в виду, что для ряда $\mathbf{s}$ можно построить модель (3), не учитывающую никакой информации о временном ряде $\mathbf{x}$. В случае линейной зависимости это выполняется, однако в случае сложных динамических систем разделимость, как правило, отсутствует.



a. b.

Рис. 1: a. Множество $M_{\mathbf{s}} = \{(s_j, s_{j-\tau}),\ j \in \mathcal{J}\}$. b. Множество $M_{\mathbf{x}} = \{(x_j, x_{j-\tau}),\ j \in \mathcal{J}\}$.

Использование метода сходящегося перекрестного отображения для выявления зависимости между временными рядами

Для исследования более сложных динамических систем, находящихся вне области применения теста Грейнджера, существует метод сходящегося перекрестного отображения. Этот метод существенно опирается на теорему Такенса [12, 13]. Рассмотрим множество M состояний некоторой динамической системы

$$\mathbf{x}_j = \varphi(\mathbf{x}_{j-1}), \quad \mathbf{x} \in \mathbb{R}^d. \tag{5}$$

. Согласно теореме Такенса, при соблюдении определенных условий динамическая система может быть реконструирована с помощью последовательности ее наблюдаемых состояний. А именно, утверждается, что множество $M_{\mathbf{x}} = f(M)$, где

$$f(x) = (x, \varphi(x), \ldots, \varphi^{d-1}(x)).$$

может быть использовано для описания системы (5).

Опишем процедуру исследования зависимости между временными рядами $\mathbf{s}$ и $\mathbf{x}$ с помощью метода сходящегося перекрестного отображения. Пусть ряды $\mathbf{s}$ и $\mathbf{x}$ и имеют достаточно большую историю. Рассмотрим множество $M = \{(s_j, x_j), j \in |\mathcal{J}|\}$ пар значений временных рядов $\mathbf{s}$ и $\mathbf{x}$, измеренных в j-й момент времени. В силу предположений о линейной зависимости φ между рядами,

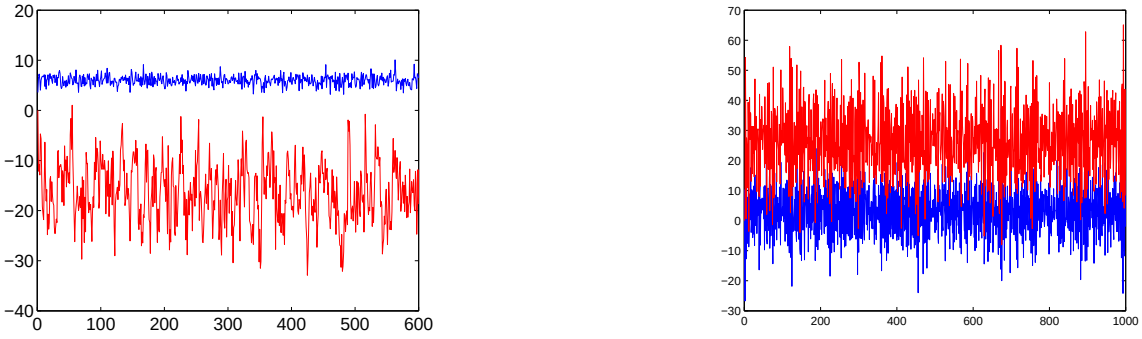
$$\varphi(x_\tau) = x_{j-\tau}, \quad \varphi(x_\tau) = x_{j-\tau},$$

где τ — выбранное в (1), (2) значение задержки. Тогда отображения $f_{\mathbf{s}} : M \to M_{\mathbf{s}}$ и $f_{\mathbf{x}} : M \to M_{\mathbf{x}}$, определенные как

$$f_{\mathbf{s}}(s_j, x_j) = (s_j, s_{j-\tau}) \equiv \boldsymbol{s}(j),$$

$$f_{\mathbf{x}}(s_j, x_j) = (x_j, x_{j-\tau}) \equiv \boldsymbol{x}(j),$$

должны сохранять свойства системы (5).

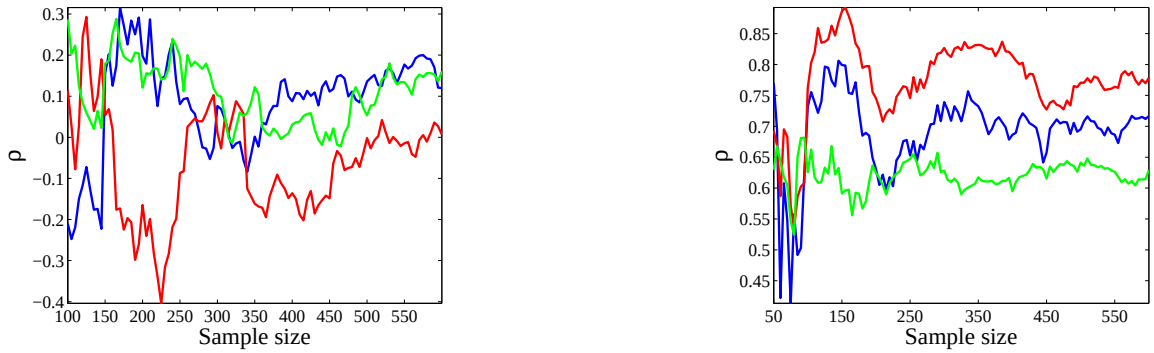


(a) (b)

Рис. 2: a. Синтетические временные ряды. Синий ряд независимый, красный зависит от своей истории и истории синего ряда. b. Временные ряды с двусторонней связью.

Рассмотрим точку $\boldsymbol{x}(j)$ множества $M_{\mathbf{x}}$. Найдем для нее $d+1$ ближайших соседей $\boldsymbol{x}(j_1), \boldsymbol{x}(j_2), \ldots, \boldsymbol{x}(j_{d+1})$. Предполагая связь между рядами $\mathbf{x}$ и $\mathbf{s}$, считаем, что индексы

$$j_1, \ldots, j_{d+1} \tag{6}$$



(a) (b)

Рис. 3: Зависимость различных корреляций от объема выборки. a — для рядов с односторонней причинно-следственной связью, b — для рядов в двусторонней связью.

ближайших к $\boldsymbol{x}(j)$ соседей в $M_{\mathbf{x}}$ являются также индексами ближайших к (j) соседей в $M_{\mathbf{s}}$. Используя значения $s_{j_1}, s_{j_2}, \ldots, s_{j_{d+1}}$ временного ряда $\mathbf{s}$, получим прогноз значения s_j:

$$\hat{s}_j = \sum_{i=1}^{d+1} w_i s_{j_i}. \tag{7}$$

Веса w_i получаются экспоненциальным взвешиванием евклидовых расстояний $r(s_{j_i}, s_j)$ от s_j до элементов $s_{j_1}, s_{j_2}, \ldots, s_{j_{d+1}}$:

$$w_i = \frac{u_i}{\sum u_k}, \quad u_i = \exp\left(-\frac{r(s_j, s_{j_i})}{r(s_j, s_{j_1})}\right), \quad i = 1, \ldots, d+1,$$

$$r(s, s') = |s - s'|.$$

При увеличении истории $|\mathcal{J}|$ измеряемых временных рядов расстояния между соседними точками множеств $M_{\mathbf{s}}$ и $M_{\mathbf{x}}$ сокращаются. В случае, если ряд $\mathbf{x}$ действительно влияет на $\mathbf{s}$, при $|\mathcal{J}| \to \infty$ прогноз (7) j-го значения ряда $\mathbf{s}$ становится точнее. Тогда коэффициент корреляции

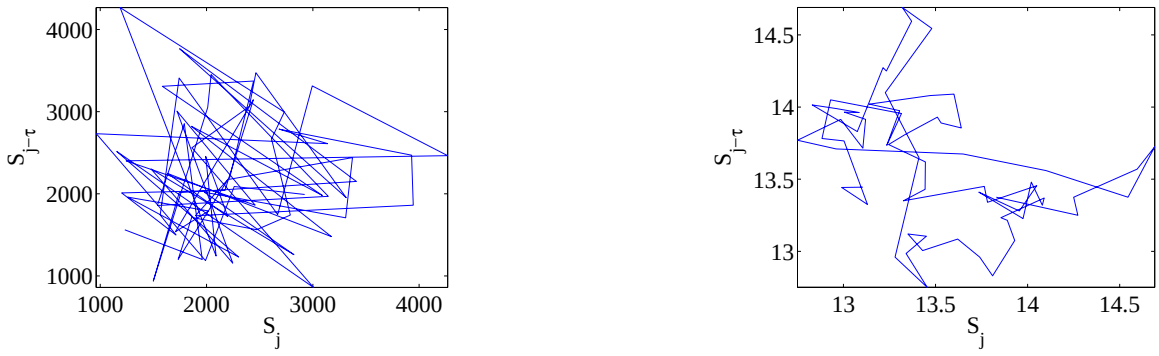
$$\rho(\hat{s}_j, s_j) = \frac{1}{\sigma_s \sigma_{\hat{s}}} \mathsf{E}(\hat{s}_j - \mathsf{E}\hat{s}_j)(s_j - \mathsf{E}s_j) \tag{8}$$

должен стремиться к некоторому ρ_0, при этом значение ρ_0 не должно равняться нулю. Здесь E — математическое ожидание случайной величины, σ — ее дисперсия. Чем сильнее ρ_0 отклоняется от нуля, тем сильнее зависимость между рядами. Наличие этой сходимости будет проверяться в вычислительном эксперименте.

В качестве примера рассмотрим ряд $\mathbf{s}$ загруженности железнодорожного узла цистернами с нефтью и ряд $\mathbf{x}$ цен на нефть. Опишем процедуру добавления элементов к выборке, предназначенную для выявления сходимости выражения (8) к ρ_0. Предположим, что ряд $\mathbf{s}$ зависит от ряда $\mathbf{x}$, тогда ближайшим соседям $\boldsymbol{x}(j)$ будут соответствовать ближайшие соседи $\boldsymbol{s}(j)$. Это позволит получать прогноз $\hat{s}_j$ значений ряда $\mathbf{s}$ по s_{j_i}, $i = 1, \ldots, d+1$, соответствующим ближайшим соседям $\boldsymbol{x}(j)$. Представим множество индексов $|\mathcal{J}|$ значений временных рядов в виде

$$\mathcal{J} = \mathcal{L} \bigsqcup \mathcal{T},$$

причем в $\mathcal{T}$ содержатся последние исторические индексы временного ряда. Исключим элементы с индексами $j \in \mathcal{T}$ из множества M, и для каждого из них вычислим $\hat{s}_j$ по



(a) (b)

Рис. 4: Первые сто точек множества $M_{\mathbf{s}}$ и $M_{\mathbf{x}}$, соединенные по возрастанию индекса j.

формуле (7). Спрогнозировав значения отрезка временного ряда, вычислим коэффициент корреляции (8) между спрогнозированными значениями и реальными значениями иссдедуемого ряда $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$. Здесь имеется в виду эмпирический коээфициент корреляции, для подсчета которого в формуле (8) матожидание E и дисперсия σ^2 заменяются на среднее значение ряда, и среднеквадратичное отклонение:

$$\mathsf{E}s \mapsto \bar{\mathbf{s}}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}} s_j,$$

$$\sigma_s^2 \mapsto \frac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}}(s_j - \bar{\mathbf{s}}_{\mathcal{T}})^2.$$

Рассмотрим зависимость коэффициентов корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ исходного ряда $\mathbf{s}_{\mathcal{T}}$ и спрогнозированного ряда $\hat{\mathbf{s}}_{\mathcal{T}}$ и коэффициента корреляции $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ рядов $\hat{\mathbf{x}}_{\mathcal{T}}$ и $\mathbf{x}_{\mathcal{T}}$ от размера выборки. При увеличении объема выборки корреляция (8) между спрогнозированными значениями зависимого временнного ряда и его измеренными значениями должна возрастать. Будем наращивать размер выборки $|\mathcal{J}|$, оставляя для прогноза последние $0.25|\mathcal{J}|$ значений временного ряда, $\mathcal{T}(|\mathcal{J}|) = \{\lfloor 0.75|\mathcal{J}|\rfloor, \ldots, |\mathcal{J}|\}$, где $\lfloor a\rfloor$ означает округление вниз. Увеличивая $|\mathcal{J}|$ и решая (7), будем вычислять $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$, $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ и корреляцию $\rho(\mathbf{s}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ между оставленными для прогноза отрезками рядов.

Рассмотрим предполагаемую связь между биржевыми ценами на нефть и загруженностью железнодорожного узла вагонами с нефтью (ожидается, что цены на нефть влияют на загруженность). Тогда модуль коэффициента корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ должен возрастать с объемом выборки и выходить на ненулевую асимптоту. На рисунке 2c изображены зависимости коэффициентов корреляции от объема выборки, по оси абсцисс отложена длина временного ряда, полученного добавлением в выборку $|\mathcal{J}|$-го элемента. На графике красного цвета по оси ординат отложены значения коэффициента корреляции спрогнозированых цен на нефть с их истинными значениями, синего — корреляция прогноза загруженности железнодорожного узла с измеренными значениями загруженности. На графике зеленого цвета отложены значения коэффициента корреляция истинных значений исследуемых рядов рядов. Хотя коэффициент $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ принимает значения в диапазоне от 0.6 до -0.8, сходимости нет. В вычислительном эксперименте в таких случаях делается вывод об отсутствии связи между временными рядами.

Рассмотрим пример с синтетическими данными, изображенными на рисунке 2a. Синий ряд, обозначим его $\mathbf{x}$, генерируется независимо от красного, $\mathbf{s}$. Ряд $\mathbf{s}$ зависит от $\mathbf{x}$. Зави-

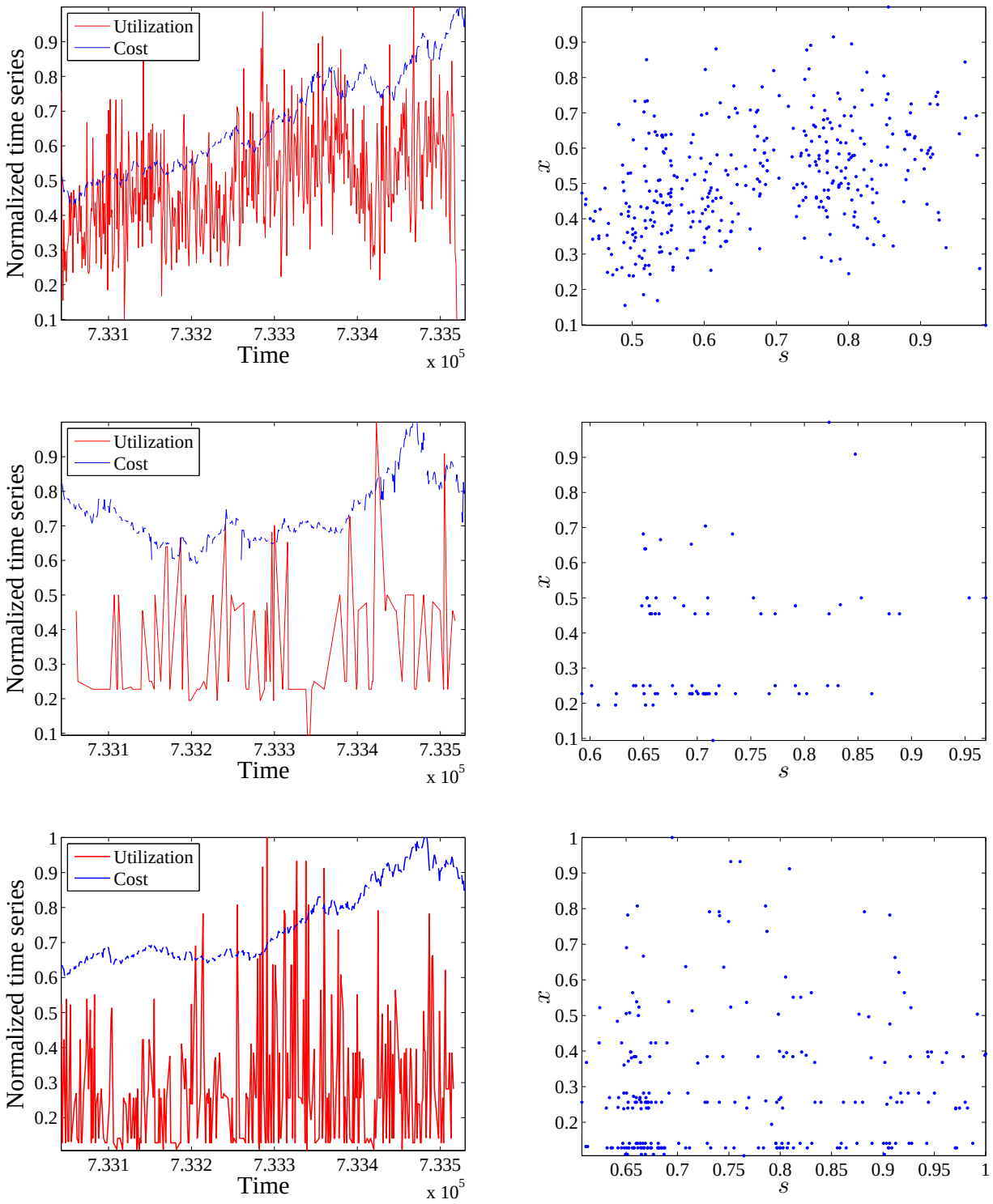


Рис. 5: Пары рядов «Нефть и нефтепродукты» —«Цены на нефть»; «Сахар» —«Цены на сахар»; «Цветные металлы, изделия из них и лом цв. металлов» —«Цены на золото».

симость коэффициентов корреляции между различными рядами от объема выборки для них изображена на рисунке 3а. Здесь синим цветом изображена зависимость $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ от

объема выборки, красным — зависимость $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ от объема выборки. Зеленый цвет соответствует зависимости $\rho(\mathbf{x}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$. Здесь $\rho_0(\hat{\mathbf{s}}, \mathbf{s}) \approx 0.1$, но наблюдается сходимость. Можно сделать вывод о наличии слабой связи между рядами.

Рассмотрим теперь синтетический пример с двусторонней связью 2b, 3b. В этом случае наблюдается сходимость к $\rho_0(\hat{\mathbf{x}}, \mathbf{x}) \approx 0.8$, $\rho_0(\hat{\mathbf{s}}, \mathbf{s}) \approx 0.7$, причем $\rho_0(\hat{\mathbf{x}}, \mathbf{x}) > \rho_0(\mathbf{x}, \mathbf{s})$, $\rho_0(\hat{\mathbf{s}}, \mathbf{s}) > \rho_0(\mathbf{x}, \mathbf{s})$, то есть для рядов с двусторонней связью метод CCM не только выявляет наличие связи, но и справляется лучше чем кросс-корреляция. Метод CCM дает здесь лучшие результаты, так как он разработан специально для выявления взаимных связей, с которыми тест Грейнджера справляется не всегда. Кроме того, для получения адекватных результатов с помощью CCM необходимо, чтобы отображения $f_{\mathbf{s}}$, $f_{\mathbf{x}}$ были взаимнооднозначными, то есть не наблюдалось самопересечений траекторий $s(j)$ и $x(j)$. Из рисунков 4 ясно, что в для исследуемых данных это не так: наблюдается значительное число самопересечений.

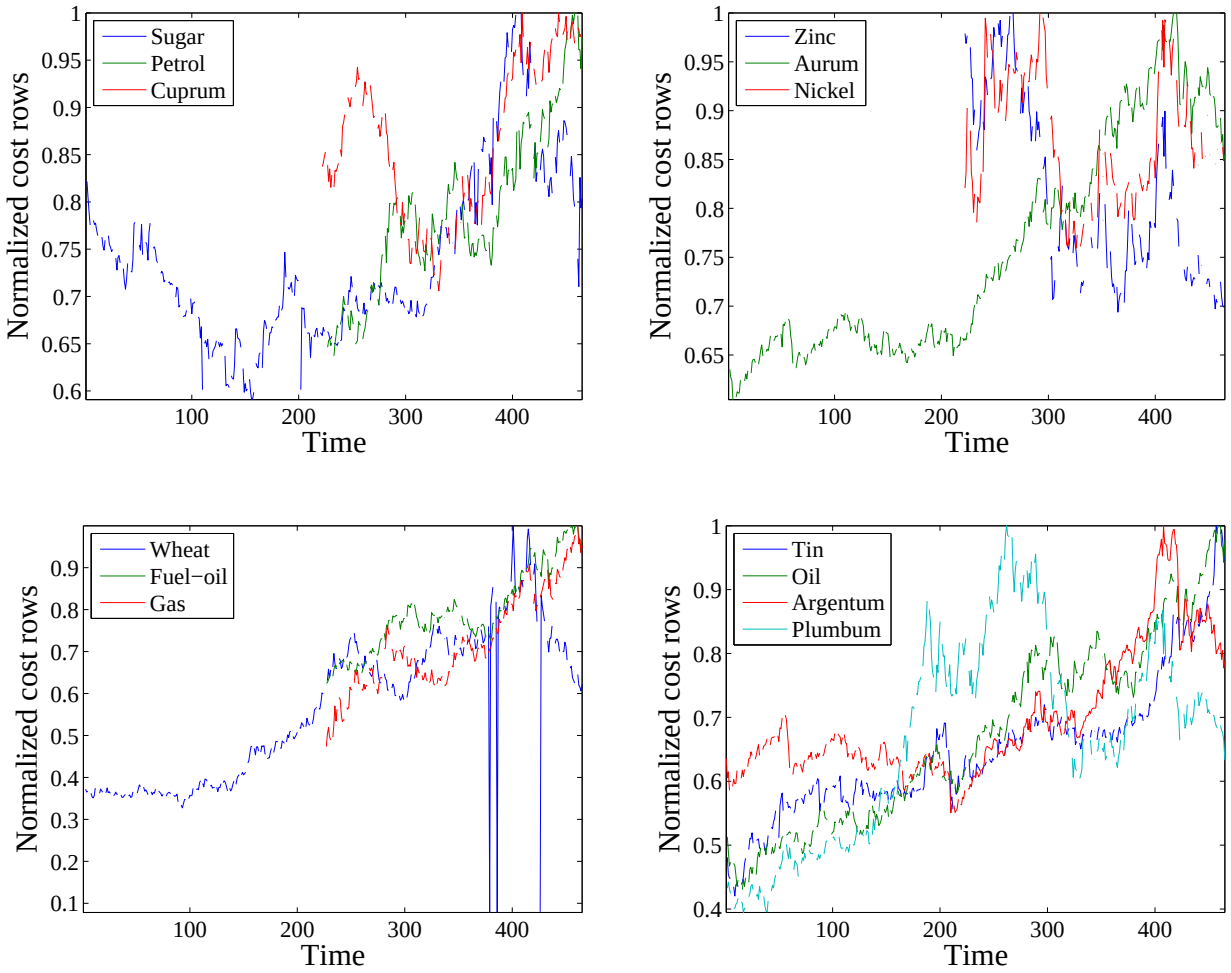


Рис. 6: Временные ряды для цен на основные инструменты.

Вычислительный эксперимент

Экспертами предоставлены временные ряды, содержащие данные о биржевых ценах на основные инструменты: сахар, бензин, медь, цинк, золото, никель, пшеницу, мазут, газ, олово, нефть, серебро, свинец за 2007 —2008 год. Эти временные ряды, нормированные на отрезок $[0, 1]$, изображены на рисунке 6. Также есть данные о загруженности некоторого железнодорожного узла грузами различных групп каждый день с 01 января 2007 года в течение 473 дней. В табл. 3 перечислены рассматриваемые группы грузов.

Таблица 3: Группы грузов.

3 — Нефть и нефтепродукты	13 — Лом черных металлов
7 — Руда железная и марганцевая	16 — Цветные металлы, изделия из них и лом цв. металлов
8 — Руда цветная и серное сырье	9 — Черные металлы
11 — Металлические конструкции	25 — Сахар
12 — Метизы	33 — Сахарная свекла и семена
34 — Зерно	35 — Продукты перемола

Опираясь на экспертные высказывания о наличии причинно-следственной связи между временным рядами, перечисленные в табл. 1, будем рассматривать различные пары $(\mathbf{s}, \mathbf{x})$ временных рядов. На рисунках 5 слева изображены пары временных рядов «Группа груза» —«Цена на некоторый инструмент». Справа на этих рисунках изображено взаимное расположение значений этих рядов. Здесь по оси абсцисс отложены значения ряда загруженности узла грузами соответствующей группы, а по оси ординат — цены на некоторый основной инструмент. Каждая точка соответствует одному дню. Полный список исследуемых зависимостей приведен в табл. 4. На первом месте стоит прогнозируемый ряд $\mathbf{s}$, на втором — ряд $\mathbf{x}$, предположительно оказывающий влияние на $\mathbf{s}$. Для каждой пары рядов в табл. 4 приведен номер соответствующего этой паре фактора из таблицы 1, а также результат теста на G-зависимость и метода сходящегося перекрестного отображения. В качестве результата теста Грейнджера в таблице приводится значение p-value для F-статистики (4). Решение о наличии G-зависимости принималось при $p < 0.1$ Также для каждой пары рядов приводится значение параметра задержки τ из модели (1). Перед применением теста Грейнждера все ряды были сглажены, после чего тестировались на стационарность с помощью теста Дики-Фуллера. Нестационарные рядам были продифференцированы. В данной работе дифференцирование проводилось не более одного раза, большее количество преобразований может затруднить интерпретацию полученных результатов. В качестве результатов метода сходящегося перекрестного отображения приведены средние значения $\bar{\rho}_{\mathbf{s}}$ и $\bar{\rho}$ коэффициентов корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ и $\rho(\mathbf{s}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$. Решение о наличии или отсутствии сходимости коэффициента (8) принималось по виду графика зависимости соответствующего коэффициента от длины выборки $|\mathcal{J}|$. Графики для некоторых пар временных рядов изображены на рисунке 2. На графике красного цвета по оси ординат отложены значения коэффициента корреляции $\rho(\hat{\mathbf{x}}, \mathbf{x})$ спрогнозированых цен на нефть с их истинными значениями, синего — корреляция $\rho(\hat{\mathbf{s}}, \mathbf{s})$ прогноза загруженности железнодорожного узла с измеренными значениями загруженности. На графике зеленого цвета отложены значения коэффициента корреляции истинных значений исследуемых рядов рядов. В графе «ССМ» встречается формулировка «Недостаточно данных». Так как метод сходящегося перекрестного отображения основан на проверке сходимости

коэффициента корреляции (8) при неограниченном увеличении объема выборки, он требует большого объема выборки. Решения принимались при $|\mathcal{J}| > 50$. Для исследования

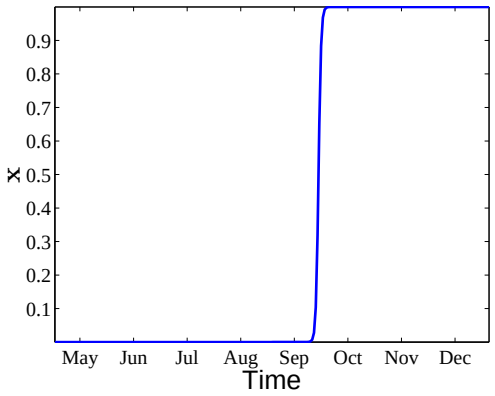


Рис. 7: Зависимость различных корреляций от объема выборки.

влияния сезонности строится вспомогательный временной ряд $\mathbf{x}$, $x_j \in [0, 1]$. Например, свекловичный сахар производится в сентябре-октябре. Для исследования влияния сезонности производства сахара на перевозки сахара будет составлен временной ряд, каждая точка которого соответствует одному дню. Значения ряда, соотвествующие месяцам с января по август включительно, приравниваются нулю, а значения ряда, соответствующие месяцам после октября — единице. Значения в сентябре-октябре получим сглаживанием имеющихся значений. Описанный ряд изображен на рисунке 7.

Результаты, приведенные в таблице следует понимать следующим образом. Каждый из рассматриваемых методов исследует временные ряды на наличие некоторой связи, причем каждый в своем смысле. Таким образом, ни для одной из пар рядов, рассмотренных в вычислительном эксперименте, не принято решение о наличии связи в смысле метода сходящегося перекрестного отображения, хотя для некоторых из них сделан вывод о наличии G-зависимости. При этом ни наличие G-зависимости, ни зависимость в смысле метода сходящегося перекрестного отображения не доказывает наличие причинно-следственной связи между рядами. Однако определение G-зависимости согласуется с целью уточнить прогноз временного ряда $\mathbf{s}$, используя значения временного ряда $\mathbf{x}$. Таким образом, если принимается решение о G-зависимости рядов, будем считать, что экспертное высказывание $\mu(\mathbf{s}, \mathbf{x})$ подтвердилось, и припишем ему оценку достоверности $1 - p(\mathbf{s}, \mathbf{x})$.

Заключение

Рассматривается задача прогнозирования загруженности железнодорожного узла. При построении модели используются экспертные высказывания о влиянии внешних факторов на прогнозируемые ряды. В связи с этим ставится задача оценки достоверности экспертного высказывания, и возникает необходимость исследования временных рядов на наличие причинно-следственной связи. В работе рассмотрены два способа выявления связи между рядами, тест Грейнджера и метод перекрестного сходящегося отображения. Описаны области применимости методов и показано, что для достижения целей, описанных в работе, следует применять тест Грейнджера. Достоверность экспертного высказывания оценивается как вероятность на основе исследуемых временных рядов принять решение о наличии G-зависимости между ними.

Таблица 4: Пары временных рядов рядов, исследуемые на зависимость в вычислительном эксперименте.

№	Пара временных рядов **s** —**x**	№ факт.	Тест Грейнджера		CCM	
1	«Нефть и нефтепродукты» — «Цены на бензин»	1	$p = 0.1131,\ \tau = 6$	−	$\bar{\rho} = -0.1296,\ \bar{\rho}_{\mathbf{s}} = 0.1630$	−
2	«Нефть и нефтепродукты» — «Цены на мазут»	1	$p = 0.0581,\ \tau = 7$	+	$\bar{\rho} = 0.1731,\ \bar{\rho}_{\mathbf{s}} = 0.1773$	−
3	«Нефть и нефтепродукты» — «Цены на нефть»	1	$p = 0.0010,\ \tau = 7$	+	$\bar{\rho} = 0.1434,\ \bar{\rho}_{\mathbf{s}} = -0.3157$	−
4	«Сахар» —«Цены на сахар»	1	$p = 0.0038,\ \tau = 7$	+	$\bar{\rho} = 0.1071,\ \bar{\rho}_{\mathbf{s}} = 0.0175$	−
5	«Сахарная свекла и семена» — «Цены на сахар»	1	$p = 0.0111,\ \tau = 8$	+	N/A	N/A
6	«Зерно» —«Цены на пшеницу»	1	$p = 0.1667,\ \tau = 1$	−	N/A	N/A
7	«Продукты перемола» — «Цены на пшеницу»	1	$p = 0.5369,\ \tau = 1$	−	N/A	N/A
8	«Сахарная свекла и семена» — «Сезонность производства сахарной свеклы»	3	$p = 0.6601,\ \tau = 8$	−	N/A	N/A

Обозначения: «+» — сделан вывод о наличии связи между временными рядами, «−» — сделан вывод об отсутствии связи, «N/A» — не достаточно данных для проведения эксперимента.

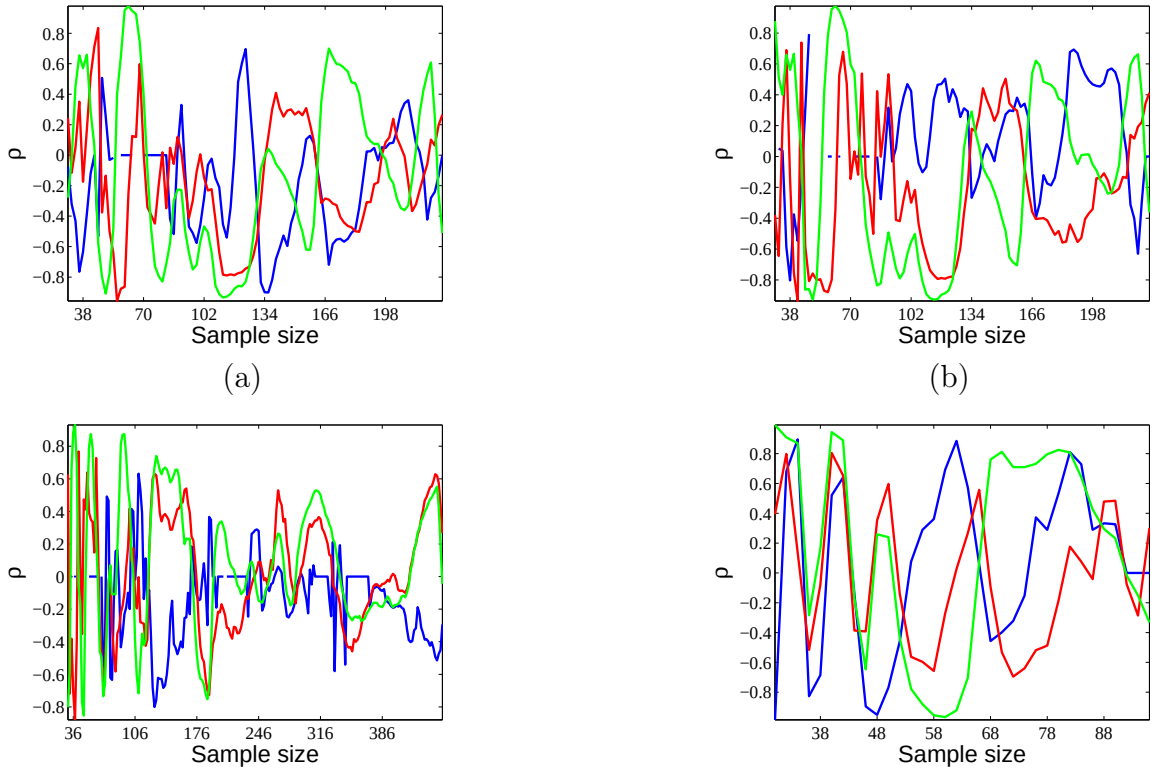


Рис. 8: Зависимости различных коэффициентов корреляции от объема выборки для пар «Нефть и нефтепродукты» —«Цены на бензин», «Нефть и нефтепродукты» —«Цены на мазут», «Нефть и нефтепродукты» —«Цены на нефть», «Сахар» —«Цены на сахар».

Литература

[1] Орлов А. И. Экспертные оценки. Заводская лаборатория, 62-1(1996). 54-60.

[2] Орлов А. И. Организационно-экономическое моделирование : учебник : в 3 ч.. / Ч. 2 : Экспертные оценки. – М.: Изд-во МГТУ им. Н.Э. Баумана. 2011.

[3] A. B. Barrett, L. Barnett, A. K. Seth. Multivariate Granger Causality and Generalized Variance //Phys. Rev. E 81-4(2010).

[4] Мотренко А. П. Использование теста Гренджера при прогнозировании временных рядов // Машинное обучение и анализ данных, 1(2011), 51-60.

[5] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica, 37(1969). 424 - 432.

[6] H. White, Xun Lu, Granger Causality and Dynamic Structural Systems // Journal of Financial Econometrics, 8-2(2012), 193–243.

[7] R. B. Kline, Principles and Practice of Structural Equation Modeling. New York: Guilford. 2005.

[8] J. Pearl, Graphs, Causality and Structural Equation Models. Sociological Methods and Research, 27-2(1998), 226-284.

[9] J. Pearl, E. Bareinboim, Transportability of Causal and Statistical Relations: A Formal Approach // Proceedings of the 25th AAAI Conference on Artificial Intelligence, August 7-11, 2011, San Francisco. 247-254.

[10] G. Sugihara et al, Detecting Causality in Complex Ecosystems // Science, 338-6106(2012), 496-500.

[11] G. Sugihara, R.M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature, 344-6268(1990). 734–741.

[12] F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L.-S. Young. Dynamical Systems and Turbulence, Lecture Notes in Mathematics, 898-1981. 366–381.

[13] E. R. Deyle, G. Sugihara. Generalized theorems for nonlinear state space reconstruction. PLoS ONE 6(3)-2011, e18295.

[14] E. Said, D. A. Dickey. Testing for Unit Roots in Autoregressive Moving Average Models of Unknown Order. Biometrika, 71-(1984). 599–607.