Анализ зависимостей между показателями при прогнозировании объема грузоперевозок *

К. Р. Усманова [1]***, С. П. Кудияров***[2]***, Р. В. Мартышкин***[3]***,***
А. А. Замковой[4]***, В. В. Стрижов***[5]

Аннотация

В работе анализируется взаимосвязь и согласованность показателей в системе управления, мониторинга состояния и отчетности железнодорожных грузоперевозок. Рассматриваются макроэкономические временные ряды, содержащие управляющие воздействия, состояние, и целевые показатели. Предполагается, что управление, состояние и целеполагание статистически связаны. Для установления связи используется тест Гренджера. Считается, что два временных ряда связаны, если использование истории одного из рядов улучшает качество прогноза другого. Цель анализа состоит в повышении качества прогноза объема грузоперевозок. Вычислительный эксперимент выполнен на данных об объеме грузоперевозок, управляющих воздействиях и установленных целевых критериях.

Ключевые слова: временные ряды; прогнозирование; тест гренджера; сиситема управления; целевые критерии

1 Введение

Прогнозирование грузовых железнодорожных перевозок имеет большое значение как для операторов железнодорожного транспорта, в том числе ОАО «РЖД», так и для органов государственной власти. Понимание перспективного спроса на грузовые железнодорожные перевозки необходимо для управления инвестиционной деятельностью, формирования перспективной топографии железнодорожной сети, рационального обновления и распределении по сети парка подвижного состава. Отсюда появляется необходимость построения максимально достоверного прогноза на основе доступных данных.

Задача прогнозирования железнодорожных грузоперевозок усложняется тем, что существует множество факторов, которые могут влиять на объем грузоперевозок. Поэтому требуется проведение экспертного анализа на предмет отбора факторов по критериям их воздействия на объемы грузовых железнодорожных перевозок и характера этого воздействия. В данной работе исследуемые факторы делятся на экзогенные и управляемые. Экспертами высказываются гипотезы о влиянии тех или иных факторов на объем грузоперевозок, а также влиянии грузоперевозок на целевые критерии. Требуется проверить выдвинутые гипотезы и исследовать зависимости между объемом грузоперевозок, экзогенными факторами, управлением и целевыми критериями.

В таблице 1 представлены рассматриваемые в работе временные ряды.

*Работа выполнена при финансовой поддержке РФФИ (проекты 17-20-01212, 17-20-01184)

[1]Московский физико-технический институт, karina.usmanova@gmail.com

[2]АО "ИЭРТ" , s.kudiyarov@gmail.com

[3]АО "ИЭРТ" , martyshkinrv@mail.ru

[4]АО "ИЭРТ" , rtr@iert.com.ru

[5]Вычислительный центр имени А.А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, strijov@gmail.com

Таблица 1 Анализируемые факторы и соответствующие им временные ряды

Фактор	Временные ряды
Объем грузоперевозок ($\mathbb{X}$)	Импорт нефти
Экзогенные факторы ($\mathbb{H}$)	Цена на нефть Курс нац. валюты
Факторы управления ($\mathbb{U}$)	Средняя скорость дост. грузов Средний вес брутто Средняя участ. скорость дост. грузов
Целевые критерии ($\mathbb{Q}$)	Добыча нефти Темп роста ВВП

Нахождение зависимостей объема грузоперевозок от экзогенных факторов и факторов управления может повысить качество прогноза, а также упростить прогностическую модель. Если удается установить, что ряд грузоперевозок $\mathbf{x}$ не зависит от фактора управления $\mathbf{u}$ (или экзогенного фактора $\mathbf{h}$), то ряд $\mathbf{y}$ можно не учитывать при прогнозе, не уменьшая качество прогноза.

Будем говорить, что ряд $\mathbf{x}$ зависит от ряда $\mathbf{y}$ (или следует из ряда $\mathbf{y}$), если использование истории ряда $\mathbf{y}$ при построении прогностической модели улучшает прогноз ряда $\mathbf{x}$. Такой подход лежит в основе теста Гренджера [1, 2]. Тест Гренджера позволяет установить причинно-следственные связи между рядами и основан на сравнении качества прогноза, в котором используется история только прогнозируемого ряда, и прогноза, который дополнительно использует историю других рядов. Если улучшение качества прогноза подтверждается статистически, то говорят, что прогнозируемый ряд следует из использовавшихся во втором прогнозе рядов. Более формально используемый в этой работе тест Гренджера описан в разделе 4. Тест Гренджера применим к стационарным временным рядам, поэтому в случае нестационарных рядов их необходимо продифференцировать перед проведением теста Гренджера. Тест Гренджера используется в различных задачах, в котрых необходимо исследовать взаимосвязь между развивающимися во времени процессами [3, 4].

В данной работе для построения прогноза одного временного ряда по нескольким используется алгоритм многомерной гусеницы (MSSA-L) [5]. Этот алгоритм является обобщением на многомерный случай алгоритма анализа спектральных компонент SSA [6, 7, 8].

Метод SSA основан на разложении временного ряда в сумму интерпретируемых компонент. Он делится на четыре основных шага: запись ряда в виде траекторной матрицы, ее сингулярное разложение, группировка компонент полученных при сингулярном разложении, по каждой сгруппированной матрице восстанавливается временной ряд. Таким образом исходный временной ряд представляется в виде суммы временных рядов. Метод SSA применяется в таких задачах, как выявления трендов [9] во временных рядах, подавления шума во временных рядах [10], прогнозирование временных рядов [11, 12].

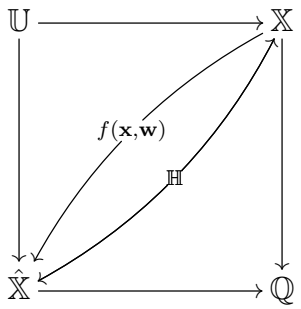

Рис. 1: Схема модели обнаружения зависимостей

На рисунке 1 изображена схема модели обнаружения зависимости. Она состоит из четырех частей: множество управляющих воздействий, состояние объекта, прогноз состояния объекта и множество кри-

териев качества состояния объекта. Принятие решения заключается в выборе воздействия $\mathbf{u} \in \mathbb{U}$, которое изменяет состояние объекта $\mathbf{x}$. На состояние объекта, помимо управляемых воздействий, могут влиять неуправляемые воздействия, называемые экзогенными факторами $\mathbf{h} \in \mathbb{H}$. Модель наблюдения объекта заключается в измерениях $\mathbf{y}(\mathbf{x}, \mathbf{h})$, которые мы можем провести, чтобы описать состояния объекта $\mathbf{x}$. Модель оценки состояния определяет качество текущего состояния объекта $\mathbf{q}(\mathbf{x}) \in \mathbb{Q}$, на основе которого принимается решение о выборе управляющего воздействия.

В данной работе модель обнаружения зависимостей рассматривается применительно к данным РДЖ. Объектом $\mathbf{x} \in \mathbb{X}$ являются показатели объема грузоперевозок. Экспертами назначаются целевые критерии $\mathbf{q}_i \in \mathbb{Q}$, управляющие факторы $\mathbf{u}_i \in \mathbb{U}$ и экзогенные факторы $\mathbf{h}_i \in \mathbb{H}$. Целевые критерии выступают в роли оценки состояния объекта. Ставится задача проверки зависимости состояния объекта $\mathbf{x}$ от управляемых ($\mathbf{u}_i$) и экзогенных ($\mathbf{h}_i$) факторов, а также зависимости целевых критериев $\mathbf{q}_i$ от состояния $\mathbf{x}$.

2 Постановка задачи прогнозирования

В данной работе строится прогноз объема грузоперевозок с использованием истории рядов управляющих факторов и целевых критериев. Поставим задачу прогноза многомерного временного ряда. Обозначим $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)})^{\mathsf{T}}$ – заданный s-мерный временной ряд. Построим матрицу плана из сегментов ряда:

$$\begin{pmatrix} x_0^{(1)} & \ldots & x_{n-1}^{(1)} \\ \vdots & & \\ x_0^{(s)} & \ldots & x_{n-1}^{(s)} \end{pmatrix} = \mathbf{X}_{0 \div (n-1)}. \tag{1}$$

Пусть $\mathbf{x}_n = \left(x_n^{(1)}, \ldots x_n^{(s)}\right)^{\mathsf{T}}$ – значение ряда $\mathbf{X}$ в момент времени n. Построим прогноз $\hat{\mathbf{x}}$ ряда $\mathbf{X}$ в точке $\mathbf{x}_n$. Проделаем это k раз для различных обучающих выборок $\mathbf{X}^i_{\text{train}} = \mathbf{X}_{i \div (n+i-1)}$, $i = 0, \ldots, (k-1)$. Получим k прогнозов $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_n, \hat{\mathbf{x}}_{n+1}, \ldots \hat{\mathbf{x}}_{n+k-1})$ ряда $\mathbf{X}$ в точках $\mathbf{x}_n, \mathbf{x}_{n+1}, \ldots \mathbf{x}_{n+k-1}$.

Прогностическая модель имеет вид

$$\hat{\mathbf{x}}_{t+1} = \mathbf{f}(\hat{\mathbf{w}}, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-L+2}),$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} S(\mathbf{w}, \mathbf{X}, \hat{\mathbf{x}}_n, \hat{\mathbf{x}}_{n+1}, \ldots \hat{\mathbf{x}}_{n+k-1}) = S(\mathbf{w}, \mathbf{X}, \hat{\mathbf{X}}),$$

где функция потерь

$$S(\mathbf{w}, \mathbf{X}, \hat{\mathbf{X}}) = \sum_{i=0}^{k-1} \mathcal{L}(\mathbf{x}^{(1)}_{n+i}, \hat{\mathbf{x}}^{(1)}_{n+i}).$$

В данной работе в качестве прогностической модели $\mathbf{f}$ используется алгоритм многомерной гусеницы (MSSA-L). Функция $\mathbf{f}$ имеет вид:

$$\mathbf{f}(\hat{\mathbf{w}}, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-L+2}) = \begin{pmatrix} x^{(1)}_{t-L+2} & \ldots & x^{(1)}_t \\ x^{(2)}_{t-L+2} & \ldots & x^{(2)}_t \\ & \vdots & \\ x^{(s)}_{t-L+2} & \ldots & x^{(s)}_t \end{pmatrix} \cdot \mathbf{p}.$$

вектор коэффициентов $\mathbf{p}$ определяется алгоритмом многомерной гусеницы MSSA-L. Алгоритм MSSA-L подробнее описан в следующем разделе.

3 Алгоритм многомерной гусеницы (MSSA-L)

Алгоритм MMSA-L является обобщением на многомерный случай алгоритма гусеницы (SSA). Задача алгоритма MSSA-L состоит в представлении временного ряда в виде суммы интерпретируемых компонент.

Это осуществляется в четыре шага: запись ряда в виде траекторной матрицы, сингулярное разложение этой матрицы, группировка компонент, полученных при сингулярном разложении, в интерпретируемые компоненты и восстановление временного ряда по каждой из интерпретируемых компонент.

По ряду (1) построим матрицу Ганкеля $\mathbf{H} \in \mathbb{R}^{L \times sK}$, $K = N - L + 1$:

$$\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_s],$$

где L – ширина окна, $\mathbf{H}_i \in \mathbb{R}^{L \times K}$ – матрица Ганкеля для ряда $\mathbf{x}^{(i)}$,

$$\mathbf{H}^{(i)} = \begin{pmatrix} x_0^{(i)} & x_1^{(i)} & \ldots & x_{N-L}^{(i)} \\ x_1^{(i)} & x_2^{(i)} & \ldots & x_{N-L+1}^{(i)} \\ & & \vdots & \\ x_{L-1}^{(i)} & x_L^{(i)} & \ldots & x_{N-1}^{(i)} \end{pmatrix}.$$

По матрице Ганкеля $\mathbf{H}$ восстановим временной ряд $\mathbf{X}$. Метод многомерной гусеницы строит приближение $\hat{\mathbf{H}}$ матрицы $\mathbf{H}$ меньшего ранга с помощью сингулярного разложения этой матрицы и восстанавливает ряд по матрице $\hat{\mathbf{H}}$. Сингулярное разложение матрицы $\mathbf{H}$ имеет вид

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$$

где $\lambda_1, \ldots, \lambda_d > 0$ – сингулярные числа матрицы $\mathbf{H}$, $\mathbf{u}_i$ и $\mathbf{v}_i$ – столбцы матриц $\mathbf{U}$ и $\mathbf{V}$. Тогда наилучшее приближение матрицы $\mathbf{H}$ матрицей ранга $r < d$ имеет вид :

$$\hat{\mathbf{H}} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$$

По матрице $\hat{\mathbf{H}}$ восстанавливается временной ряд $\mathbf{X}$ путем усреднения элементов, стоящих на антидиагоналях.

Алгоритм многомерной гусеницы также позволяет построить прогноз временного ряда в момент N по $(L-1)$ предыдущим значениям ряда. Алгоритм находит такой вектор коэффициентов $\mathbf{p} \in \mathbb{R}^{(L-1)}$, что значения ряда $\mathbf{X}$ в момент N:

$$\mathbf{x}_N = \begin{pmatrix} x_{N-L+1}^{(1)} & \ldots & x_{N-1}^{(1)} \\ x_{N-L+1}^{(2)} & \ldots & x_{N-1}^{(2)} \\ & \vdots & \\ x_{N-L+1}^{(s)} & \ldots & x_{N-1}^{(s)} \end{pmatrix} \cdot \mathbf{p} = \mathbf{Y} \cdot \mathbf{p} \tag{2}$$

Заметим, что коэффициенты $\mathbf{p}$ оказываются общими для всех компонент ряда $\mathbf{X}$.

Для каждого $i \in [1, r]$ обозначим $\tilde{\mathbf{u}}_i$ первые $(L-1)$ компонент столбца $\mathbf{u}_i$, π_i – последнюю компоненту столбца $\mathbf{u}_i$ и $\nu = \sum_{i=1}^{r} \pi_i^2$. Тогда вектор коэффициентов $\mathbf{p}$ вычисляется по формуле:

$$\mathbf{p} = \frac{1}{1 - \nu^2} \sum_{i=1}^{r} \pi_i \tilde{\mathbf{u}}_i \tag{3}$$

Заметим, что для одномерного временного ряда справедливы все приведенные соотношения при $s = 1$.

4 Анализ взаимосвязи показателей грузоперевозок

В работе для установления причинно-следственных связей предлагается использовать статистический тест Гренджера. Ниже приведен алгоритм теста Гренджера для проверки наличия зависимости одного временного ряда от другого. Пусть требуется проверить, зависит ли ряд $\mathbf{x}$ от ряда $\mathbf{y}$. Выдвинем гипотезу о независимости ряда $\mathbf{x}$ от ряда $\mathbf{y}$ и проверим ее. Делаем это следующим образом.

1. Строим прогноз ряда $\mathbf{x}$ без использования ряда $\mathbf{y}$ и находим значение функции потерь

$$S_{\mathbf{x}} = \sum_{i=1}^{n} \mathcal{L}\left(\mathbf{x}_i, \hat{\mathbf{x}}_i\right),$$

где n – длина тестовой выборки.

Функцию $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}})$ выбираем в зависимости от распределения ошибок прогноза на тестовой выборке (4).

2. Строим прогноз ряда $\mathbf{x}$ с использованием ряда $\mathbf{y}$. Вычисляем для него значение функции потерь

$$S_{\mathbf{xy}} = \sum_{i=1}^{n} \mathcal{L}\left(\mathbf{x}_i, \hat{\mathbf{x}}_i\right).$$

3. Рассмотрим статистику

$$T(\mathbf{x}, \mathbf{y}) = \frac{N - 2k}{k} \cdot \frac{S_{\mathbf{x}} - S_{\mathbf{xy}}}{S_{\mathbf{xy}}},$$

где N – длина обучающей выборки, k – размерность регрессионной модели. Статистика T имеет распределение $F(k, N-2k)$ (распределение Фишера с параметрами $(k, N-2k)$).

4. Если ряд $\mathbf{x}$ не зависит от ряда $\mathbf{y}$, то значения $S_{\mathbf{x}}$ и $S_{\mathbf{xy}}$ будут близки, а статистика $T(\mathbf{x}, \mathbf{y})$ – незначима. Поэтому в случае больших значений статистики $T(\mathbf{x}, \mathbf{y})$ отвергаем гипотезу о независимости ряда $\mathbf{x}$ от $\mathbf{y}$. Выберем некоторое критические значение t статистики $T(\mathbf{x}, \mathbf{y})$. Тогда критерий зависимости ряда $\mathbf{x}$ от ряда $\mathbf{y}$ выглядит следующим образом:

$$\text{Из } T(\mathbf{x}, \mathbf{y}) > t \quad \text{следует, что ряд } \mathbf{x} \text{ зависит от ряда } \mathbf{y}$$

5. Аналогично проверим зависимость ряда $\mathbf{x}$ от восстановленного (с помощью алгоритма MSSA-L) ряда $\hat{\mathbf{y}}$. Для этого используем статистику

$$T(\mathbf{x}, \hat{\mathbf{y}}) = \frac{N - 2k}{k} \cdot \frac{S_{\mathbf{x}} - S_{\mathbf{x}\hat{\mathbf{y}}}}{S_{\mathbf{x}\hat{\mathbf{y}}}}.$$

Для более подробного изучения связи между временными рядами $\mathbf{x}$ и $\mathbf{y}$ вычисляем кросс–корреляционную функцию $\gamma_{\mathbf{xy}}(h)$

$$\gamma_{\mathbf{xy}}(h) = \frac{\mathsf{E}\left[\left(\mathbf{x}_t - \mu_{\mathbf{x}}\right)\left(\mathbf{y}_{t+h} - \mu_{\mathbf{y}}\right)\right]}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}},$$

где E – математическое ожидание, μ – выборочное среднее, σ – выборочная дисперсия.

Если h^* соответствует максимальному значению кросс-корреляции, то говорят, что ряд $\mathbf{y}$ сдвинут на h^* относительно $\mathbf{x}$. Заметим, что если ряд $\mathbf{x}$ сдвинут на h_1 относительно ряда $\mathbf{y}$, а ряд $\mathbf{y}$ сдвинут на h_2 относительно ряда $\mathbf{z}$. То ряд $\mathbf{x}$ сдвинут на $h_3 = h_1 + h_2$ относительно ряда $\mathbf{z}$.

Пусть прогноз ряда $\mathbf{x}$ строится с использованием истории ряда $\mathbf{y}$ и пусть с помощью вычисления кросс-корреляции рядов $\mathbf{x}$ и $\mathbf{y}$ получено, что ряд $\mathbf{x}$ отстает от ряда $\mathbf{y}$ на h отсчетов времени. Тогда использование при прогнозе ряда $\mathbf{y}$, сдвинутого на h отсчетов назад, может повысить качество прогноза.

5 Вычислительный эксперимент по выявлению зависимостей между показателями грузоперевозок

Эксперимент проводится на реальных данных РЖД об объеме грузоперевозок, экзогенных факторах, факторах управления и целевых критериях [13]. Ставится задача выявления зависимости объема грузоперевозок от экзогенных факторов и управления, а также зависимости целевых критериев от объема грузоперевозок.

Данные представляют собой восемь временных рядов: импорт нефти $\mathbf{x}$, цена на нефть $\mathbf{h}_1$, курс доллара $\mathbf{h}_2$, средняя скорость доставки грузов $\mathbf{u}_1$, средний вес брутто $\mathbf{u}_2$, средняя участковая скорость доставки грузов $\mathbf{u}_3$, добыча нефти $\mathbf{q}_1$ и темп роста ВВП $\mathbf{q}_2$. Описание заданных временных рядов представлены в таблице 2.

Таблица 2 Временные ряды объема импорта нефти, целевых критериев, управляемых и экзогенных факторов

Временной ряд	**Период**	**Частота**
Импорт нефти ($\mathbf{x}$)	1996 – 2017	По месяцам
Цена на нефть ($\mathbf{h}_1$)	1996 – 2017	По месяцам
Курс доллара ($\mathbf{h}_2$),	1996 – 2017	По месяцам
Средняя скорость дост. грузов ($\mathbf{u}_1$)	1996 – 2017	По годам
Средний вес брутто ($\mathbf{u}_2$)	1996 – 2017	По годам
Средняя участ. скорость дост. грузов ($\mathbf{u}_3$)	2007 – 2017	По Годам
Добыча нефти ($\mathbf{q}_1$)	1996 – 2017	По месяцам
Темп роста ВВП ($\mathbf{q}_2$)	2006 – 2016	По годам

Проверяются следующие гипотезы:

1. Ряд $\mathbf{x}$ следует из рядов $\mathbf{h}_1$, $\mathbf{h}_2$.
2. Ряд $\mathbf{x}$ следует из рядов $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$.
3. Ряд $\mathbf{q}_1$ следует из ряда $\mathbf{x}$.
4. Ряд $\mathbf{q}_2$ следует из ряда $\mathbf{x}$.

Значения временных рядов, заданных с частотой раз в год, будем линейно аппроксимировать для каждого месяца. Временные ряды, заданные с 1996 года по 2017 изображены на рис. 2, временные, заданные с 2007 по 2016 год, изображены на рис. 3.

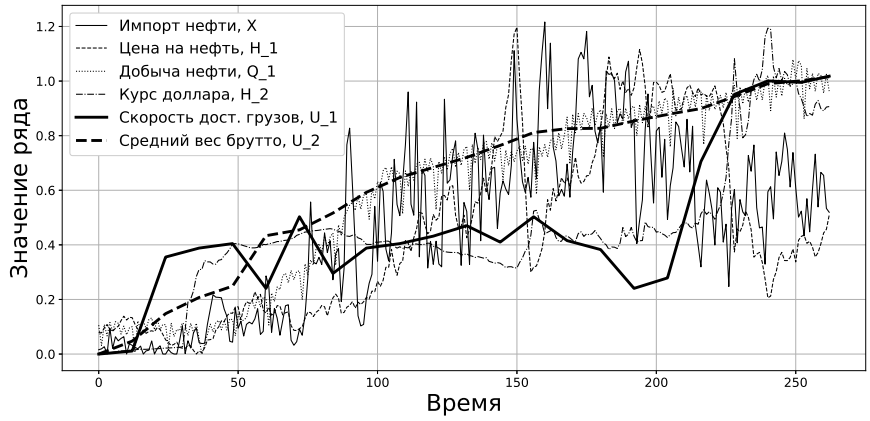

Рис. 2: Импорт нефти, цена на нефть, курс доллара, добыча нефти, скорость доставки грузов, средний вес брутто

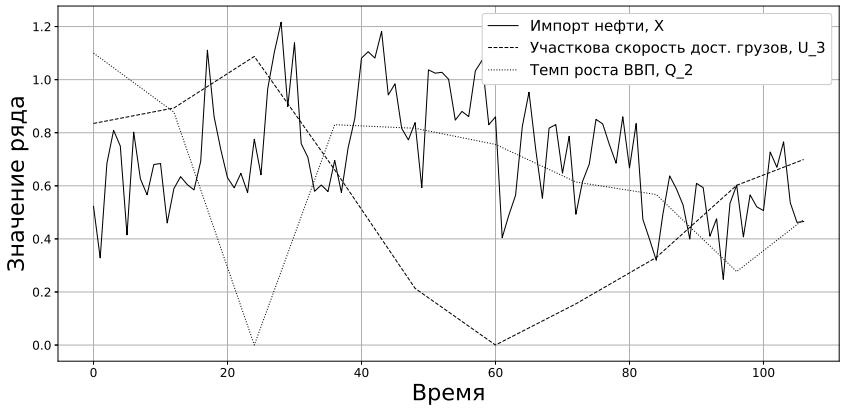

Рис. 3: Импорт нефти, средний рост ВВП, участковая скорость доставки грузов

Из графиков на рис. 2 и 3 видно, что рассматриваемые временные ряды не стационарны. Поэтому продифференцируем их. На рис. 4 представлены продифференцированные ряды.

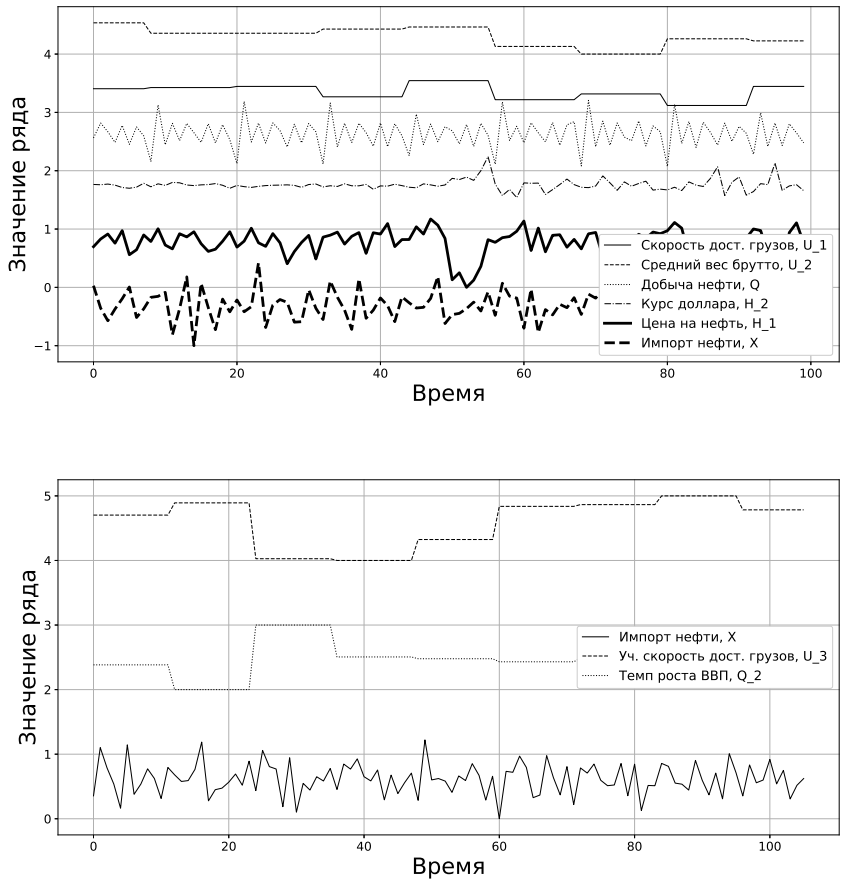

Рис. 4: Продифференцированные временные ряды

5.1 Кросс-корреляция рядов

Вычислим кросс-корреляцию для всех исследуемых пар рядов. Для примера приведем кросс-корреляционные диаграммы для пар рядов $(\mathbf{x}, \mathbf{h}_1)$ и $(\mathbf{x}, \mathbf{h}_2)$.

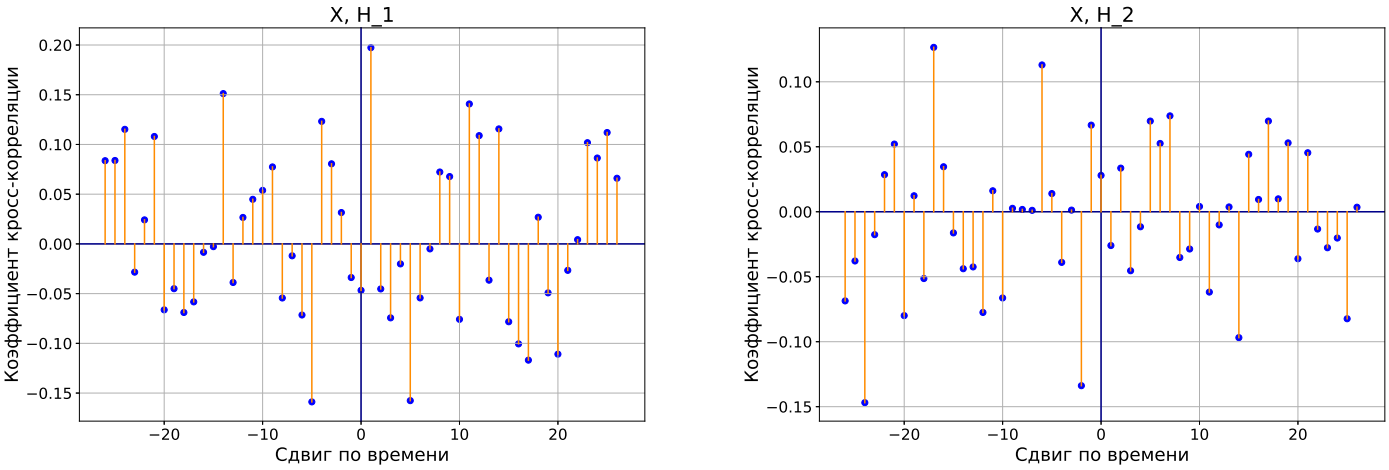

Рис. 5: Кросс-корреляционные диаграммы для пар рядов $(\mathbf{x}, \mathbf{h}_1)$ и $(\mathbf{x}, \mathbf{h}_2)$

В таблице 3 приведены найденные с помощью вычисления кросс-корреляции значения сдвигов рядов друг относительно. Для каждой пары рядов $(\mathbf{x}, \mathbf{y})$ приведено значение h такое, что значение $\gamma_{\mathbf{xy}}(h)$ максимально.

Таблица 3 Относительные сдвиги для различных пар рядов

Рассм. пары рядов	$(\mathbf{x}, \mathbf{h}_1)$	$(\mathbf{x}, \mathbf{h}_2)$	$(\mathbf{x}, \mathbf{u}_1)$	$(\mathbf{x}, \mathbf{u}_2)$	$(\mathbf{x}, \mathbf{u}_3)$	$(\mathbf{q}_1, \mathbf{x})$	$(\mathbf{q}_2, \mathbf{x})$
Сдвиг h	1	-24	-2	16	-19	24	19

5.2 Выбор функции потерь

Для оценки качества прогноза необходимо выбрать функцию потерь. Это делается, исходя из распределения ошибок прогноза. Найдем распределение ошибок прогноза рядов. Для примера на графике представлено распределение ошибок при прогнозе ряда $\mathbf{x}$.

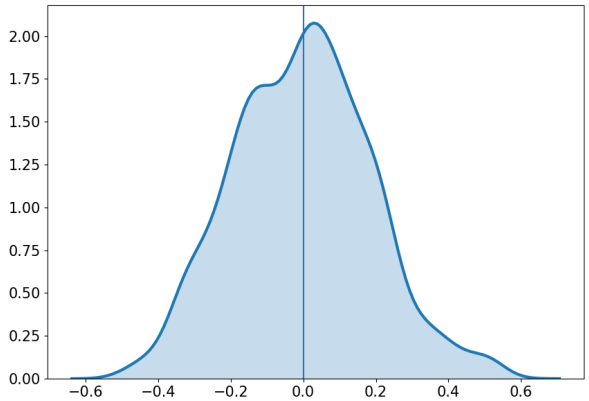

Рис. 6: Распределение ошибок прогноза ряда $\mathbf{x}$

Видно, что распределение ошибок несимметрично. Поэтому надо подбирать функционал ошибки как суперпозицию функционалов (MSE, MAE, MAPE). В данной работе эта проблема не исследуется и в качестве функционала ошибки используется MSE.

5.3 Выбор оптимальных параметров модели

Для проведения теста Гренджера построим прогноз временных рядов. В качестве прогностической модели выбрана MSSA-L. Подберем ее оптимальные параметры модели для каждой исследуемой пары рядов. Для этого найдем среднеквадратичную ошибку прогноза при различных значениях ширины окна L и ранга восстановления r.

Ширина окна перебирается по сетке $40, 45, \ldots, 175$. Ранг восстановления перебирается по сетке $3, 4, \ldots, 7$. При построении прогноза по двум рядам будем сдвигать один из них на количество отсчетов h, найденное в пункте 5.1 при помощи кросс-корреляции.

Приведем график зависимости MSE от ширины окна L для различных значений r при прогнозировании ряда $\mathbf{x}$ с использованием ряда $\mathbf{h}_1$, а также при прогнозировании ряда $\mathbf{q}_1$ с использованием ряда $\mathbf{x}$.

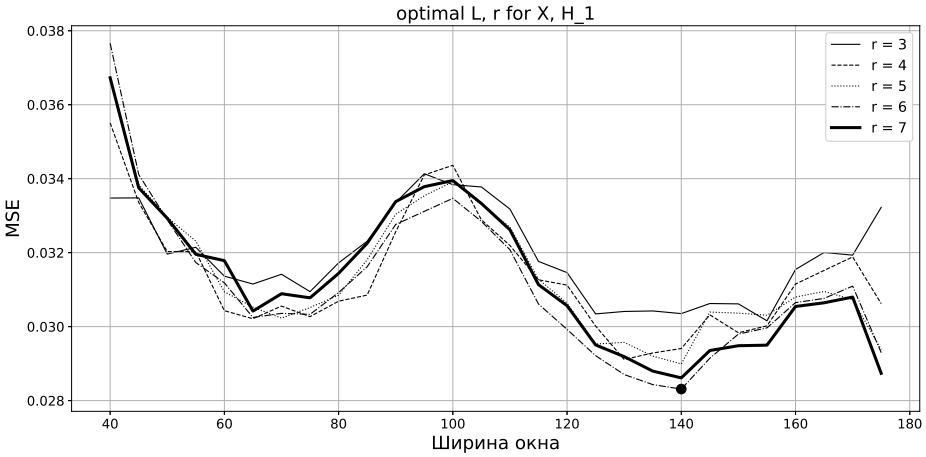

Рис. 7: MSE прогноза импорта нефти $\mathbf{x}$ с использованием ряда цен на нефть $\mathbf{h}_1$

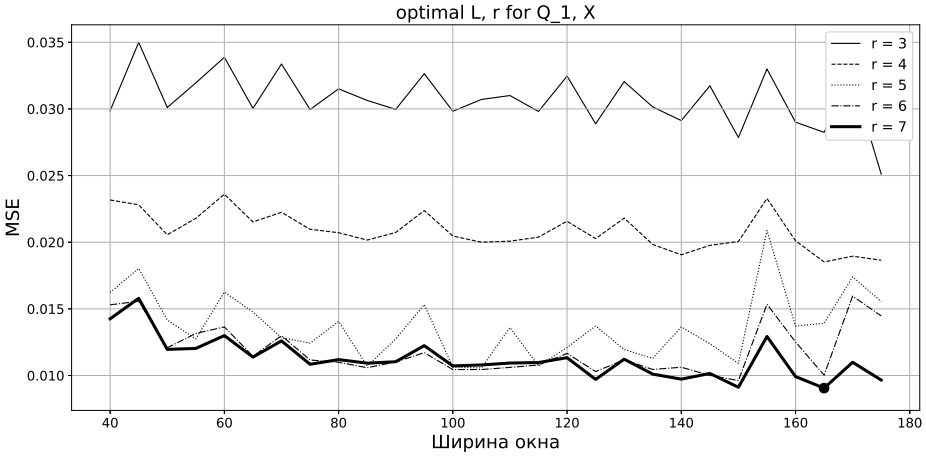

Рис. 8: MSE прогноза добычи нефти $\mathbf{q}_1$ с использованием ряда импорта нефти $\mathbf{x}$

5.4 Тест Гренджера

Исследуем зависимость объема грузоперевозок от экзогенных факторов и факторов управления, а также зависимость целевых критериев от объема грузоперевозок с помощью теста Гренджера.

На графиках представлена среднеквадратичная ошибка прогноза от длины окна L при отобранных в предыдущем пункте оптимальных значениях ранга восстановления r.

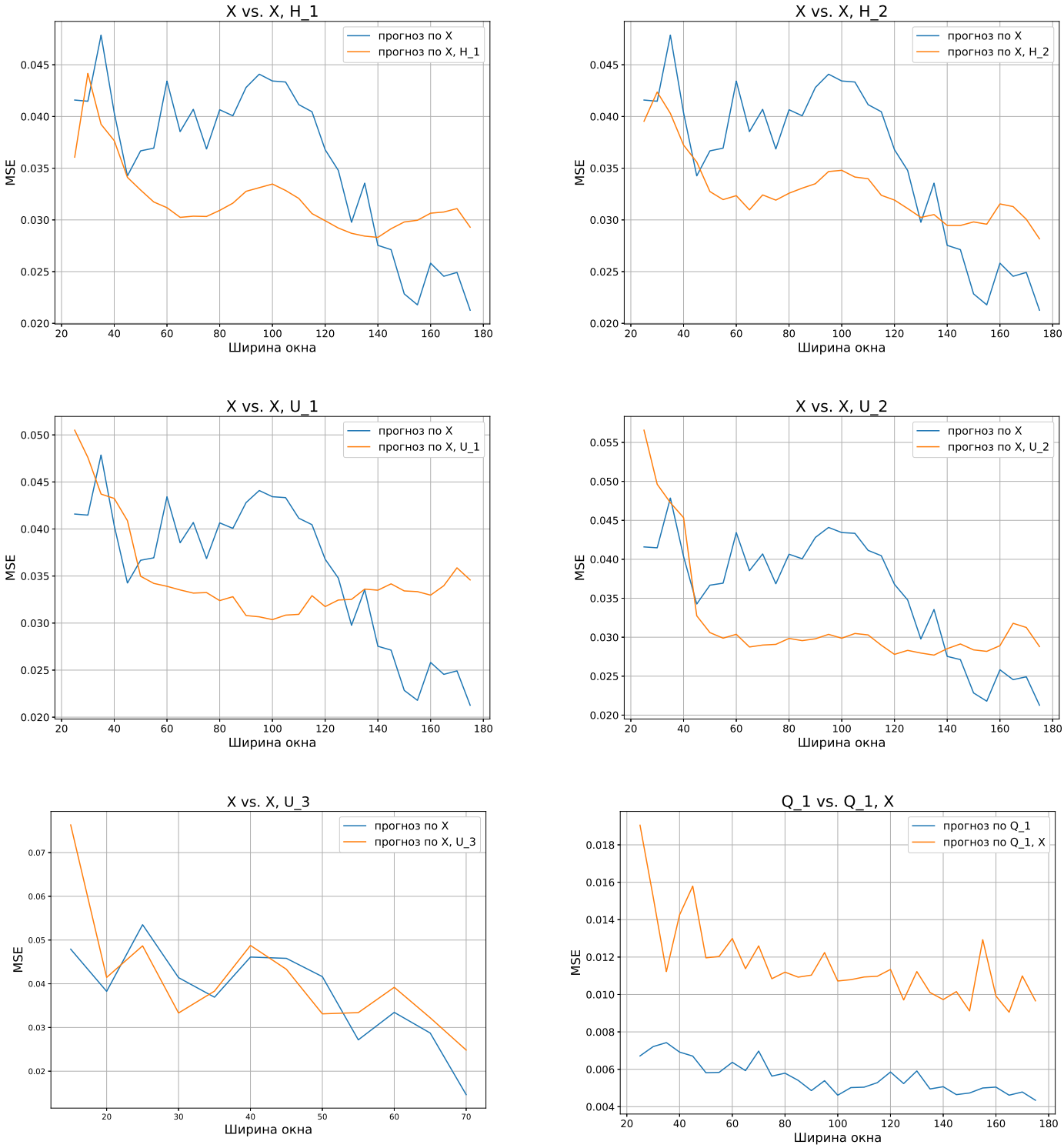

Рис. 9: Среднеквадратичная ошибка прогноза

Для каждой рассматриваемой пары рядов проверим гипотезу независимости. Для этого найдем для каждой пары значение статистики T:

$$T(\mathbf{x}, \mathbf{y}) = \frac{N - 2k}{k} \cdot \frac{\text{MSE}_{\mathbf{x}} - \text{MSE}_{\mathbf{xy}}}{\text{MSE}_{\mathbf{xy}}}, \quad \text{MSE}_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 \tag{4}$$

где k – размерность регрессионной модели, N – длина обучающей выборки.

На рисунках представлено распределение статистики T и ее значения, полученные в эксперименте для различных пар рядов. Красным выделена область, соответствующая p-value $= 0.05$.

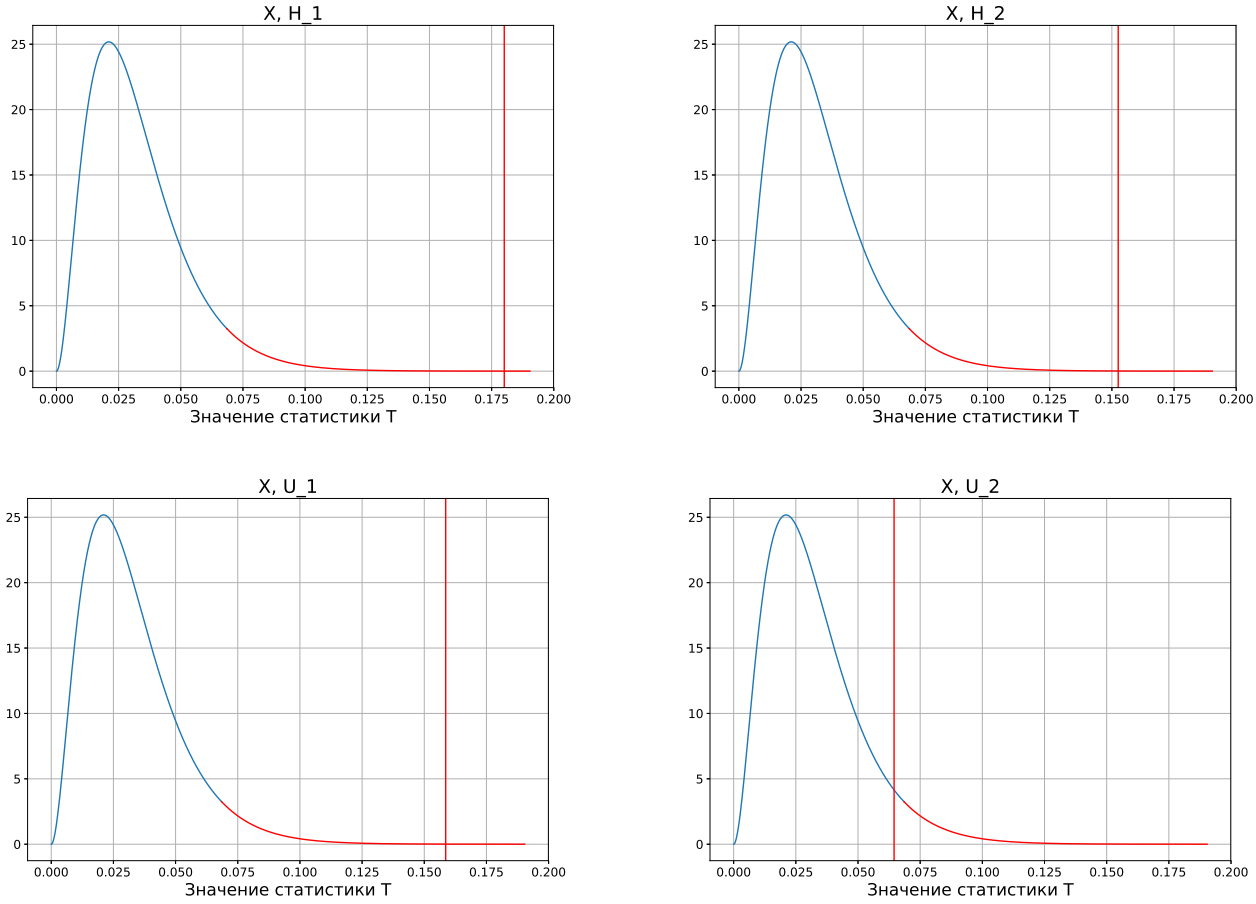

Рис. 10: Значения статистики T для пар рядов $(\mathbf{x}, \mathbf{h}_1)$, $(\mathbf{x}, \mathbf{h}_2)$, $(\mathbf{x}, \mathbf{u}_1)$, $(\mathbf{x}, \mathbf{u}_2)$

Для остальных рассматриваемых пар рядов значения статистики T получились отрицательными. Это означает, что в этих парах ряды независимы. Из графиков видно, что на уровне значимости $\alpha = 0.05$ можно утверждать, что импорт нефти $(\mathbf{x})$ зависит от курса доллара $(\mathbf{h}_1)$, цены на нефть $(\mathbf{h}_2)$ и скорости доставки грузов $(\mathbf{u}_1)$, а от среднего веса брутто $(\mathbf{u}_2)$ – не зависит.

В таблице 4 приведены результаты эксперимента. Для каждой тройки рядов (экзогенный фактор, фактор управления и целевой критерий) приведено найденное в эксперименте значение статистики T и вывод о наличии связи между состоянием и экзогенным фактором, состоянием и фактором управления, целевым критерием и состоянием.

Таблица 4 Значения статистики T для различных пар рядов

Рассм. рядыИссл. зависимости	$(\mathbf{x}, \mathbf{h})$	$(\mathbf{x}, \mathbf{u})$	$(\mathbf{q}, \mathbf{x})$
Цена на нефть ($\mathbf{h}_1$) Скорость дост. грузов ($\mathbf{u}_1$) Добыча нефти ($\mathbf{q}_1$)	$T = 0.180$ $\mathbf{x}$ зависит от $\mathbf{h}_1$	$T = 0.159$ $\mathbf{x}$ зависит от $\mathbf{u}_1$	$T < 0$ $\mathbf{q}_1$ не зависит от $\mathbf{x}$
Курс доллара ($\mathbf{h}_2$) Скорость дост. грузов ($\mathbf{u}_1$) Добыча нефти ($\mathbf{q}_1$)	$T = 0.153$ $\mathbf{x}$ зависит от $\mathbf{h}_2$	$T = 0.159$ $\mathbf{x}$ зависит от $\mathbf{u}_1$	$T < 0$ $\mathbf{q}_1$ не зависит от $\mathbf{x}$
Цена на нефть ($\mathbf{h}_1$) Средний вес брутто ($\mathbf{u}_2$) Добыча нефти ($\mathbf{q}_1$)	$T = 0.180$ $\mathbf{x}$ зависит от $\mathbf{h}_1$	$T = 0.065$ $\mathbf{x}$ не зависит от $\mathbf{u}_2$	$T < 0$ $\mathbf{q}_1$ не зависит от $\mathbf{x}$
Курс доллара ($\mathbf{h}_2$) Средний вес брутто ($\mathbf{u}_2$) Добыча нефти ($\mathbf{q}_1$)	$T = 0.153$ $\mathbf{x}$ зависит от $\mathbf{h}_2$	$T = 0.065$ $\mathbf{x}$ не зависит от $\mathbf{u}_2$	$T < 0$ $\mathbf{q}_1$ не зависит от $\mathbf{x}$
Цена на нефть ($\mathbf{h}_1$) Средняя участковая скорость ($\mathbf{u}_3$) Добыча нефти ($\mathbf{q}_1$)	$T = 0.180$ $\mathbf{x}$ зависит от $\mathbf{h}_1$	$T < 0$ $\mathbf{x}$ не зависит от $\mathbf{u}_3$	$T < 0$ $\mathbf{q}_1$ не зависит от $\mathbf{x}$
Курс доллара ($\mathbf{h}_2$) Средняя участковая скорость ($\mathbf{u}_3$) Добыча нефти ($\mathbf{q}_1$)	$T = 0.153$ $\mathbf{x}$ зависит от $\mathbf{h}_2$	$T < 0$ $\mathbf{x}$ не зависит от $\mathbf{u}_3$	$T < 0$ $\mathbf{q}_1$ не зависит от $\mathbf{x}$
Цена на нефть ($\mathbf{h}_1$) Скорость дост. грузов ($\mathbf{u}_1$) Темп роста ВВП ($\mathbf{q}_2$)	$T = 0.180$ $\mathbf{x}$ зависит от $\mathbf{h}_1$	$T = 0.159$ $\mathbf{x}$ зависит от $\mathbf{u}_1$	$T < 0$ $\mathbf{q}_2$ не зависит от $\mathbf{x}$
Курс доллара ($\mathbf{h}_2$) Скорость дост. грузов ($\mathbf{u}_1$) Темп роста ВВП ($\mathbf{q}_2$)	$T = 0.153$ $\mathbf{x}$ зависит от $\mathbf{h}_2$	$T = 0.159$ $\mathbf{x}$ зависит от $\mathbf{u}_1$	$T < 0$ $\mathbf{q}_2$ не зависит от $\mathbf{x}$
Цена на нефть ($\mathbf{h}_1$) Средний вес брутто ($\mathbf{u}_2$) Темп роста ВВП ($\mathbf{q}_2$)	$T = 0.180$ $\mathbf{x}$ зависит от $\mathbf{h}_1$	$T = 0.065$ $\mathbf{x}$ не зависит от $\mathbf{u}_2$	$T < 0$ $\mathbf{q}_2$ не зависит от $\mathbf{x}$
Курс доллара ($\mathbf{h}_2$) Средний вес брутто ($\mathbf{u}_2$) Темп роста ВВП ($\mathbf{q}_2$)	$T = 0.153$ $\mathbf{x}$ зависит от $\mathbf{h}_2$	$T = 0.065$ $\mathbf{x}$ не зависит от $\mathbf{u}_2$	$T < 0$ $\mathbf{q}_2$ не зависит от $\mathbf{x}$
Цена на нефть ($\mathbf{h}_1$) Средняя участковая скорость ($\mathbf{u}_3$) Темп роста ВВП ($\mathbf{q}_2$)	$T = 0.180$ $\mathbf{x}$ зависит от $\mathbf{h}_1$	$T < 0$ $\mathbf{x}$ не зависит от $\mathbf{u}_3$	$T < 0$ $\mathbf{q}_2$ не зависит от $\mathbf{x}$
Курс доллара ($\mathbf{h}_2$) Средняя участковая скорость ($\mathbf{u}_3$) Темп роста ВВП ($\mathbf{q}_2$)	$T = 0.153$ $\mathbf{x}$ зависит от $\mathbf{h}_2$	$T < 0$ $\mathbf{x}$ не зависит от $\mathbf{u}_3$	$T < 0$ $\mathbf{q}_2$ не зависит от $\mathbf{x}$

6 Заключение

Решалась задача проверки зависимости объема грузоперевозок от экзогенных и управляемых факторов, а также зависимости целевых критериев от объема грузоперевозок. Исследуемые ряды факторов управления и целевых критериев отбирались согласно экспертным высказываниям о наличие причинно-следственной связи между этими рядами и объемом грузоперевозок. Ставилась задача проверки достоверности экспертных оценок зависимостей рядов. В работе проведен тест Гренджера для установления взаимосвязи между рядами показателей грузоперевозок. Была исследована зависимость импорта нефти от цены на нефть, курса доллара, средней скорости доставки грузов, средней участковой скорости доставки грузов, среднего веса брутто, а также зависимость добычи нефти и темпа роста ВВП от импорта

нефти. В эксперименте было получено, что импорт нефти зависит от цены на нефть, курса доллара, скорости доставки грузов, а от средней участковой скорости и среднего веса брутто – не зависит. Также было установлено, что добыча нефти и темп роста ВВП не зависят от импорта нефти. Таким образом, результаты эксперимента показывают, что объем грузоперевозок зависит не от всех факторов управления и что рассмотренные целевые критерии не зависят от объема грузоперевозок. В дальнейшем предполагается обсуждение полученных результатов с экспертами, а также возможное изменение подхода к исследованию зависимостей между рядами.

Список литературы

[1] *Granger C. W. J.* Investigating causal relations by econometric models and cross-spectral methods // Econometrica: Journal of the Econometric Society, 1969. Vol. 37. No. 3. P. 424–438.

[2] *Barrett A. B., Barnett L., Seth. A. K.* Multivariate Granger causality and generalized variance // Physical Review E, 2010. Vol. 81. No. 4.

[3] *Hiemstra C., Jones J. D.* Testing for linear and nonlinear Granger causality in the stock price-volume relation // The Journal of Finance, 1994. Vol. 49. No. 5. P. 1639–1664.

[4] *Hoffmann R., Lee C.-G., Ramasamy B., Yeung M.* FDI and pollution: a Granger causality test using panel data // Journal of international development, 2005. Vol. 17. No. 3. P. 311–317.

[5] *Golyandina N., Stepanov D.* SSA-based approaches to analysis and forecast of multidimensional time series // In proceedings of the 5th St. Petersburg workshop on simulation, 2005. Vol. 293. P. 298.

[6] *Golyandina N., Nekrutkin V., Zhigljavsky A. A.* Analysis of time series structure: SSA and related techniques. – Chapman and Hall, 2002.

[7] *Golyandina N., Zhigljavsky A.* Singular Spectrum Analysis for time series. – Springer Science & Business Media, 2013.

[8] *Elsner J. B., Tsonis A. A.* Singular spectrum analysis: a new tool in time series analysis. – Springer Science & Business Media, 2013.

[9] *Alexandrov T.* A method of trend extraction using singular spectrum analysis // REVSTAT – Statistical Journal, 2009. Vol. 7. No. 1. P. 1-22.

[10] *Allen M. R., Smith L. A.* Monte carlo SSA: Detecting irregular oscillations in the presence of colored noise // Journal of climate, 1996. Vol. 9. No. 12. P. 3373–3404.

[11] *Hassani H., Heravi S., Zhigljavsky A.* Forecasting uk industrial production with multivariate singular spectrum analysis // Journal of Forecasting, 2013. Vol. 32. No. 5. P. 395–408.

[12] *Marques C. A. F., Ferreira J. A., Rocha A., Castanheira J. M., Melo-Gonçalves P., Vaz N., Dias J. M.* Singular spectrum analysis and forecasting of hydrological time series // Physics and Chemistry of the Earth, 2006. Vol. 31. No. 18. P. 1172-1179.

[13] Исходный код эксперимента к работе "Анализ зависимостей между показателями при прогнозировании объема грузоперевозок"(модуль на Python и Jupyter-ноутбук), `http://svn.code.sf.net/p/mlalgorithms/code/Group474/Usmanova2017TransportationQueryForecasting`