Models of detection relationship between time series in forecasting problems*

K. R. Usmanova [1], ***V. V. Strijov*** [2]

Abstract: The problem of forecasting requires relationship between multiple time series. Engagement of related time series in a forecast model boosts the forecast quality. This paper introduces the convergent cross mapping method to establish a relationship between time series. This method estimates accuracy of reconstruction of one time series using the other series. The CCM detects relationship between series not only in full trajectory spaces, but in trajectory subspaces. The computational experiment is carried out on two sets of time series: electricity consumption and air temperature, oil transportation volume and oil production volume.

Keywords: time series; forecasting; trajectory subspace; phase trajectory; convergent cross mapping

*This research was supported by RFBR (projects 17-20-01212, 19-07-0885). This paper contains results of the project Statistical methods of machine learning, which is carried out within the framework of the Program "Center of Big Data Storage and Analysis" of the National Technology Initiative Competence Center. It is supported by the Ministry of Science and Higher Education of the Russian Federation according to the agreement between the M.V. Lomonosov Moscow State University and the Foundation of project support of the National Technology Initiative from 11.12.2018, No 13/1251/2018.

[1]Moscow Institute of Physics and Technologies, karina.usmanova@phystech.edu

[2]A. A. Dorodnicyn Computing Center, Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, Moscow Institute of Physics and Technology, strijov@ccas.ru

References

[1] Data from research about relationship between ECG indicators and pulse http://smartlab.ws/component/content/article?id=60.

[2] Granger, C. W. J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society.* 37(3):424–438.

[3] Barrett, A. B., L. Barnett, and A. K. Seth. 2010. Multivariate granger causality and generalized variance. *Physical Review* 81(4):041907.

[4] Sugihara, G., R. May, Y. Hao, H. Chih-hao, E. Deyle, M. Fogarty, and S. Munch. 2012. Detecting causality in complex ecosystems. *Science.* 338(6106): 496-500.

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

[6] Hiemstra, C., and J. D. Jones. 1994. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance.* 49(5):1639–1664.

[7] Hoffmann, R., C.-G. Lee, B. Ramasamy, and M. Yeung. 2005. FDI and pollution: a Granger causality test using panel data *Journal of International Development.* 17(3):311–317.

[8] White, H., and L. Xun. 2010. Granger causality and dynamic structural systems *Journal of Financial Econometrics.* 8(2):193–243.

[9] Katrutsa, A. M., and V. V. Strijov. 2015. Stress test procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems.* 142:172–183.

[10] Li, J., K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J.Tang, and L. Huan. 2017. Feature selection: A data perspective *ACM Computing Surveys* 50(6):1-45.

[11] Geladi P. 1988. Notes on the history and nature of partial least squares modelling. *Journal of Chemometrics.* 2(4):231–246.

[12] Hoskuldsson A. Pls regression methods. 1988. *Journal of Chemometrics.* 2(3):211–228.

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

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

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

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

Модели обнаружения зависимостей во временных рядах в задачах построения прогностических моделей*

***К. Р. Усманова*[1], *В. В. Стрижов*[2]**

Аннотация: При прогнозировании сложноорганизованных временных рядов, зависящих от экзогенных факторов и имеющих множественную периодичность, требуется решить задачу выявления связанных пар рядов. Предполагается, что добавление этих рядов в модель повышает качество прогноза. В данной работе для обнаружения связей между временными рядами предлагается использовать метод сходящегося перекрестного отображения. При таком подходе два временных ряда связаны, если существуют их траекторные подпространства, проекции на которые связаны. В свою очередь, проекции рядов на траекторные подпространства связаны, если окрестность фазовой траектории одного ряда отображается в окрестность фазовой траектории другого ряда. Ставится задача отыскания траекторных подпространств, обнаруживающих связь рядов. Решение этой задачи продемонстрировано на двух наборах рядов: потребление электроэнергии и температура воздуха, объем железнодорожных перевозок нефти и объем добычи нефти.

Ключевые слова: временные ряды; прогнозирование; траекторное подпространство; фазовая траектория; сходящееся перекрестное отображение

1 Введение

Работа посвящена обнаружению причинно-следственных связей между разнородными временным рядами. Например, прогноз железнодорожных грузоперевозок по различным группам грузов, связь сигналов ЭКГ и пульса [1].

Если прогноз временного ряда $\mathbf{x}$ строится с использованием временных рядов $\mathbf{y}_1, \ldots, \mathbf{y}_k$, то установление связей ряда $\mathbf{x}$ с $\mathbf{y}_1, \ldots, \mathbf{y}_k$ может повысить качество прогноза и при этом упростить прогностическую модель. Если установлено, что ряд $\mathbf{x}$ не зависит от ряда $\mathbf{y}_i$, то $\mathbf{y}_i$ должен быть исключен из прогностической модели. В работе анализируются два подхода обнаружения связи между рядами: тест Гренджера [2, 3] и метод сходящегося перекрестного отображения [4, 5].

*Работа выполнена при поддержке РФФИ (проекты 17-20-01212, 19-07-0885). Настоящая статья содержит результаты проекта «Статистические методы машинного обучения», выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору МГУ им. М. В.Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 11.12.2018 № 13/1251/2018.

[1]Московский физико-технический институт, karina.usmanova@phystech.edu

[2]Вычислительный центр имени А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, Московский физико-технический институт, strijov@ccas.ru

В основе теста Гренджера лежит следующий подход. Считаем, 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 статистически значимо повышает качество прогноза ряда $\mathbf{x}$ [2, 3]. Тест Гренджера устанавливает связи между рядами и основан на сравнении качества прогноза, в котором используется история только прогнозируемого ряда, либо прогноза, который дополнительно использует историю других рядов. Если улучшение качества прогноза подтверждается статистически, то говорят, что прогнозируемый ряд связан с этими рядами. Тест Гренджера используется в тех задачах, где требуется исследовать взаимосвязь между развивающимися во времени процессами [6, 7]. Тест Гренджера применим к стационарным временным рядам, поэтому в случае нестационарных рядов их необходимо продифференцировать перед проведением теста.

Недостатком теста Гренджера является то, что при используемом в нем подходе невозможно точно определить структуру зависимости рядов. Например, два ряда могут следовать из третьего, но при отсутствии информации о третьем ряде тест Грернджера установит причинно-следственную связь между первым и вторым рядом, хотя она отсутствует. Проблема точного определения структуры зависимости рядов рассмотрена в работе [8].

В случае, когда тест Гренджера неприменим или не может обнаружить связь между рядами, применяется метод сходящегося перекрестного отображения (convergent cross mapping, CCM). Этот метод основан на оценке того, насколько хорошо один ряд может быть восстановлен с использованием второго. Считается [4, 5], что ряд $\mathbf{x}$ восстанавливается по ряду $\mathbf{y}$, только если ряд $\mathbf{y}$ влияет на ряд $\mathbf{x}$. Метод CCM основан на сравнении ближайших соседей в траекторном пространстве ряда $\mathbf{x}$, полученных с помощью ряда $\mathbf{x}$ и с помощью ряда $\mathbf{y}$. Проверяется, насколько точно моменты времени, соответствующие ближайшим соседям вектора $\mathbf{y}_t$, определяют ближайших соседей вектора $\mathbf{x}_t$.

При построении линейной прогностической модели по временному ряду строится траекторная матрица, играющая роль матрицы объектов. Ответами являются значения ряда в последующие моменты времени. Когда размерность траекторного пространства избыточна, прогностическая модель становится неустойчивой. В этом случае необходимо производить отбор признаков [9, 10]. Метод проекций на латентные структуры (partial least squares, PLS) отбирает наиболее значимые признаки и строит новые признаки как их линейные комбинации [11, 12]. Таким образом, PLS находит подпространство траекторного пространства, проекция на которое наилучшим образом приближает исходный ряд. Снижение размерности применяется при изучении связей между рядами. Проекции на траекторные подпространства позволяют более детально изучить связь между главными компонентами рядов и найти подпространство, в котором наблюдается связь между рядами.

В данной работе для построения прогноза одного временного ряда по нескольким рядам используется алгоритм многомерной гусеницы (multivariate singular spectrum analysis, MSSA-L) [13]. Этот алгоритм является обобщением на многомерный случай алгоритма анализа спектральных компонент (singular spectrum analysis, SSA) [14–16]. Метод SSA основан на разложении временного ряда в сумму интерпретируемых компонент. Он подразумевает 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.

Эксперимент проводился на трех парах временных рядов: искусственные ряды, среднесуточная температура и потребление электроэнергии, объем железнодорожных перевозок нефти и объем добычи нефти. Для каждой пары исследовалось наличие связи между компонен-

тами входящих в нее рядов. В экспериментах на реальных данных строился прогноз рядов, использующий обнаруженные связанные компоненты рядов.

2 Обнаружение связей временных рядов

В данном разделе описываются метод сходящегося перекрестного отображения и тест Гренджера.

Метод сходящегося перекрестного отображения. Для заданного временного ряда $\mathbf{x} = [x_1, \ldots, x_N]^\mathsf{T}$ построим траекторную матрицу

$$\mathbf{H}_\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \ldots & x_{n-1} & x_n \\ x_2 & x_3 & \ldots & x_n & x_{n+1} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{N-n+1} & x_{N-n+2} & \ldots & x_{N-1} & x_N \end{bmatrix}, \tag{1}$$

где n – ширина окна. Обозначим t-ю строку матрицы $\mathbf{H}_\mathbf{x}$ как $\mathbf{x}_t$. Тогда матрица $\mathbf{H}_\mathbf{x}$ принимает вид

$$\mathbf{H}_\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_T \end{bmatrix}, \quad \mathbf{x}_t = [x_t, x_{t+1}, \ldots, x_{t+n-1}], \quad T = N - n + 1. \tag{2}$$

Заметим, что все векторы $\mathbf{x}_t$ принадлежат n-мерному траекторному пространству $\mathbb{H}_\mathbf{x} \subseteq \mathbb{R}^n$ ряда $\mathbf{x}$.

Обнаружение зависимости между рядами $\mathbf{x}$ и $\mathbf{y}$ осуществляется следующим образом. Выберем момент $t^* \in [1, T]$ и найдем k ближайших соседей вектора $\mathbf{x}_t$ в $\mathbb{H}_\mathbf{x}$. Обозначим их множество как $U_k(\mathbf{x}_{t^*}) = \{\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}\}$, где

$$\mathbf{x}_{t_i} = [x_{t_i}, x_{t_i+1}, \ldots, x_{t_i+n-1}], \quad i = 1, \ldots, k. \tag{3}$$

Так как оба ряда $\mathbf{x}$ и $\mathbf{y}$ определены на единой временной оси, отыщем $U_k(\mathbf{y}_{t^*})$, см. рис. 1, в пространстве $\mathbb{H}_\mathbf{y}$ с целью анализа связей между этими двумя рядами. Для этого повторим в пространстве $\mathbb{H}_\mathbf{y}$ построения (1)–(3). Аналогично строим матрицу $\mathbf{H}_\mathbf{y}$. Поставим в соответствие каждому вектору $\mathbf{x}_{t_i} \in U_k(\mathbf{x}_{t^*})$ вектор $\mathbf{y}_{t_i}$:

$$\varphi : \mathbf{x}_{t_i} \to \mathbf{y}_{t_i}, \quad i = 1, \ldots, k.$$

Найденные векторы $\mathbf{y}_{t_i}$ образуют множество $U_k(\mathbf{y}_{t^*})$. Утверждается, что ряды $\mathbf{x}$ и $\mathbf{y}$ связаны, отображение φ из пространства $\mathbb{H}_\mathbf{x}$ в пространство $\mathbb{H}_\mathbf{y}$ липшицево:

$$\rho_{\mathbb{H}_\mathbf{y}}\Big(\varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j)\Big) \leq n \cdot \rho_{\mathbb{H}_\mathbf{x}}(\mathbf{x}_i, \mathbf{x}_j), \quad \forall\, \mathbf{x}_i, \mathbf{x}_j \in \mathbb{H}_\mathbf{x},$$

где $\rho_{\mathbb{H}_\mathbf{x}}(\cdot,\cdot), \rho_{\mathbb{H}_\mathbf{y}}(\cdot,\cdot)$ – метрики в пространствах $\mathbb{H}_\mathbf{x}$ и $\mathbb{H}_\mathbf{y}$ соответственно. Проверим наличие такого отображения следующим образом. Введем меру близости векторов в окрестностях $U_k(\mathbf{x}_{t^*})$ и $U_k(\mathbf{y}_{t^*})$:

$$L(\mathbf{x}, \mathbf{y}) = \frac{R\Big(U_k(\mathbf{x}_{t^*})\Big)}{R\Big(U_k(\mathbf{y}_{t^*})\Big)}, \quad R\Big(U_k(\mathbf{x}_{t^*})\Big) = \frac{1}{k}\sum_{i=1}^{k} \rho_{\mathbb{H}_\mathbf{x}}\Big(\mathbf{x}_{t^*}, \mathbf{x}_{t_i}\Big). \tag{4}$$

Если $L(\mathbf{x}, \mathbf{y})$ больше некоторого порога $L(n)$, то ряд $\mathbf{y}$ зависит от ряда $\mathbf{x}$.

На рис. 1 показан описанный способ обнаружения связи между рядами. На верхнем рисунке ряд $\mathbf{y}$ зависит от ряда $\mathbf{x}$, на нижнем – не зависит.

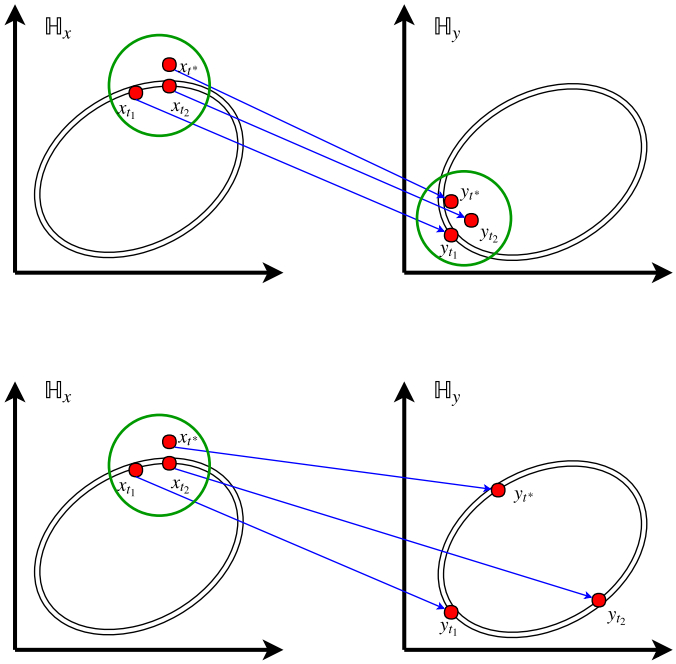


Рис. 1: Отображение из траекторного пространства ряда $\mathbf{x}$ в траекторное пространство ряда $\mathbf{y}$

Прогноз $\hat{x}_t$, первого элемента вектора $\mathbf{x}_t$, строится следующим образом:

$$\hat{x}_t = \sum_{i=1}^{k} \mathrm{w}_i x_{t_i}, \quad \text{где } t_i \text{ – индексы ближайших соседей } \mathbf{x}_t,$$
$$\mathrm{w}_i = \frac{u_i}{\sum_i u_i}, \quad u_i = \exp\left(-\frac{\rho(\mathbf{x}_t, \mathbf{x}_{t_i})}{\rho(\mathbf{x}_t, \mathbf{x}_{t_{n+1}})}\right), \tag{5}$$

где $\rho(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_1 - \mathbf{x}_2||_2$ – метрика в пространстве $\mathbb{H}_\mathbf{x}$.

Анализ связей подпространств траекторных пространств. Предлагается обнаружить связь не в траекторных пространствах $\mathbb{H}_\mathbf{x}$ и $\mathbb{H}_\mathbf{y}$, а только в некоторых их подпространствах, натянутых на главные компоненты тракторных матриц $\mathbf{H}_\mathbf{x}$ и $\mathbf{H}_\mathbf{y}$, не обязательно первые. Сингулярное разложение матрицы $\mathbf{H}_\mathbf{x}$:

$$\mathbf{H}_\mathbf{x} = \mathbf{U}_\mathbf{x} \mathbf{\Lambda}_\mathbf{x} \mathbf{V}_\mathbf{x}.$$

Обозначим $\mathcal{J}_\mathbf{x} = \{1, \ldots, n\}$ – индексы компонент ряда $\mathbf{x}$. Подпространство траекторного пространства зададим с помощью набора индексов компонент $\mathcal{A} \subseteq \mathcal{J}$, на которые оно будет натянуто. Построим проекцию ряда $\mathbf{x}$ на подпространство, натянутое на компоненты с номерами из $\mathcal{A}_\mathbf{x}$. Обозначим это подпространство $\mathbb{H}_{\mathcal{A}_\mathbf{x}} \subseteq \mathbb{H}_\mathbf{x}$. Заменим в матрице $\mathbf{\Lambda}_\mathbf{x}$ элементы, находящиеся в строках с индексами не из $\mathcal{A}_\mathbf{x}$, нулями. Обозначим полученную матрицу $\tilde{\mathbf{\Lambda}}_\mathbf{x}$.

Тогда проекция ряда $\mathbf{x}$ на подпространство, натянутое на компоненты с индексами из $\mathcal{A}_\mathbf{x}$, задается траекторной матрицей

$$\mathbf{P}_{\mathcal{A}_\mathbf{x}} = \mathbf{U}_\mathbf{x}\tilde{\mathbf{\Lambda}}_\mathbf{x}\mathbf{V}_\mathbf{x}.$$

Аналогично по некоторому набору $\mathcal{A}_\mathbf{y}$ строится подпространство $\mathbb{H}_{\mathcal{A}_\mathbf{y}} \subseteq \mathbb{H}_\mathbf{y}$ и траекторная матрица $\mathbf{P}_{\mathcal{A}_\mathbf{y}}$. Предлагается искать ближайших соседей не в полных траекторных пространствах $\mathbb{H}_\mathbf{x}$ и $\mathbb{H}_\mathbf{y}$, задающихся траекторными матрицами $\mathbf{H}_\mathbf{x}$ и $\mathbf{H}_\mathbf{y}$ соответственно, а в подпространствах $\mathbb{H}_{\mathcal{A}_\mathbf{x}}$ и $\mathbb{H}_{\mathcal{A}_\mathbf{y}}$, задающихся матрицами $\mathbf{P}_{\mathcal{A}_\mathbf{x}}$ и $\mathbf{P}_{\mathcal{A}_\mathbf{y}}$. Предполагается, что переход к траекторным подпространствам меньшей размерности повышает устойчивость прогностической модели и позволяет более подробно изучить связь между рядами.

Рассмотрев различные подпространства, выбираем то, которое будет наилучшим образом описывать исследуемый временной ряд и иметь минимальную размерность. Перебор различных подпространств позволяет установить, между какими именно компонентами рядов $\mathbf{x}$ и $\mathbf{y}$ существует зависимость.

Будем перебирать различные комбинации индексов главных компонент и соответствующие им подпространства $\mathbb{H}_{\mathcal{A}_\mathbf{x}}$ и $\mathbb{H}_{\mathcal{A}_\mathbf{y}}$. Для каждой пары $(\mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y})$ индексов главных компонент рядов $\mathbf{x}$ и $\mathbf{y}$ соответственно будем находить среднее расстояние между k ближайшими соседями для ряда $\mathbf{x}$ и между ближайшими соседями для ряда $\mathbf{y}$. Введем меру близости векторов, аналогичную (4) и нормированную на размерность траекторных подпространств:

$$L(\mathbf{x}, \mathbf{y} \,|\, \mathcal{A}_\mathbf{x}\mathcal{A}_\mathbf{y}) = \frac{R\Big(U_k\left(\mathbf{x}_{t^*}\right)\big|\mathcal{A}_\mathbf{x}\Big)}{R\Big(U_k\left(\mathbf{y}_{t^*}\right)\big|\mathcal{A}_\mathbf{y}\Big)} \cdot \frac{|\mathcal{A}_\mathbf{y}|}{|\mathcal{A}_\mathbf{x}|}, \quad R(\mathbf{x} \,|\, \mathcal{A}_\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} \rho_{\mathbb{H}_\mathbf{x}}\Big(\mathbf{x}_{t^*}, \mathbf{x}_{t_i}\Big). \tag{6}$$

Тогда задача поиска подпространств $\mathbb{H}_{\mathcal{A}_\mathbf{x}}$ и $\mathbb{H}_{\mathcal{A}_\mathbf{y}}$ эквивалентна поиску номеров главных компонент $(\mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y})$ и имеет вид

$$(\mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y}) = \arg\max_{\mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y}} L(\mathbf{x}, \mathbf{y} \,|\, \mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y}), \quad \mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y} \subseteq \{1, \ldots, n\}. \tag{7}$$

Тест Гренджера. В качестве базового метода установления связей используется статистический тест Гренджера. 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. Для этого строим прогноз $\hat{\mathbf{x}}$ ряда $\mathbf{x}$ без использования ряда $\mathbf{y}$ и находим невязку $\boldsymbol{\varepsilon}_\mathbf{x} = \hat{\mathbf{x}} - \mathbf{x}$. Строим прогноз ряда $\mathbf{x}$ с использованием ряда $\mathbf{y}$ и находим невязку $\boldsymbol{\varepsilon}_{\mathbf{xy}}$. Рассмотрим статистику

$$\Theta(\mathbf{x}, \mathbf{y}) = \frac{N - 2n}{n} \cdot \frac{\|\boldsymbol{\varepsilon}_\mathbf{x}\|_2^2 - \|\boldsymbol{\varepsilon}_{\mathbf{xy}}\|_2^2}{\|\boldsymbol{\varepsilon}_{\mathbf{xy}}\|_2^2},$$

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

Если ряд $\mathbf{x}$ не зависит от ряда $\mathbf{y}$, то значения $\boldsymbol{\varepsilon}_\mathbf{x}$ и $\boldsymbol{\varepsilon}_{\mathbf{xy}}$ будут близки, значение статистики $\Theta(\mathbf{x}, \mathbf{y})$ – мало. Значит, p-value гипотезы о независимости рядов $\mathbf{x}$ и $\mathbf{y}$ на уровне значимости α имеет вид

$$\text{p-value} = \inf\{\alpha : \Theta(\mathbf{x}, \mathbf{y}) > \theta(\alpha)\}.$$

Другими словами, для уровня значимости α утверждаем, что ряд $\mathbf{x}$ зависит от ряда $\mathbf{y}$, если $\Theta(\mathbf{x}, \mathbf{y}) > \theta(\alpha)$. Критерий зависимости ряда $\mathbf{x}$ от ряда $\mathbf{y}$ размера α выглядит следующим образом:

$$\text{Из } \Theta(\mathbf{x}, \mathbf{y}) > \theta(\alpha) \quad \text{следует, что ряд } \mathbf{x} \text{ зависит от ряда } \mathbf{y}.$$

Для приближенного, восстановленного с помощью модели MSSA-L ряда $\hat{\mathbf{y}}$ проверим зависимость от него ряда $\mathbf{x}$ Для этого используем статистику $\Theta(\mathbf{x}, \hat{\mathbf{y}})$.

Чтобы найти сдвиг по времени ряда $\mathbf{y}$ относительно ряда $\mathbf{x}$ вычислим кросскорреляционную функцию

$$\gamma_{\mathbf{xy}}(h) = \frac{\mathsf{E}\left[\left(\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x}}\right)\left(\mathbf{y}_{t+h} - \boldsymbol{\mu}_{\mathbf{y}}\right)\right]}{\sigma_{\mathbf{x}}^2 \sigma_{\mathbf{y}}^2}, \tag{8}$$

где E – математическое ожидание, $\boldsymbol{\mu}$ – выборочное среднее, σ^2 – выборочная дисперсия. $h^* = \arg\max_h \gamma_{\mathbf{xy}}(h)$ – сдвиг по времени ряда $\mathbf{y}$ относительно ряда $\mathbf{x}$. Использование ряда $\mathbf{y}$, сдвинутого на h отсчетов назад, при построении прогноза ряда $\mathbf{x}$ повышает качество прогноза.

3 Постановка задачи прогнозирования

Поставим задачу прогноза набора временных рядов. Обозначим через

$$\mathbf{X} = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)}\right]^{\mathsf{T}} \tag{9}$$

заданный набор из s временных рядов. Рассмотрим сегмент набора рядов $\mathbf{X}_{(t-n+1)\div t}$ от момента времени $(t-n+1)$ до момента t:

$$\begin{matrix} x_1^{(1)} & \ldots & x_{t-n}^{(1)} \\ \vdots & & \vdots \\ x_1^{(s)} & \ldots & x_{t-n}^{(s)} \end{matrix} \underbrace{\begin{pmatrix} x_{t-n+1}^{(1)} & \ldots & \overbrace{x_t^{(1)}}^{\boldsymbol{\chi}_t} \\ \vdots & & \vdots \\ x_{t-n+1}^{(s)} & \ldots & x_t^{(s)} \end{pmatrix}}_{\mathbf{X}_{(t-n+1)\div t}} \begin{matrix} x_t^{(1)} & \ldots \\ \vdots & \\ x_t^{(s)} & \ldots \end{matrix} \tag{10}$$

Пусть $\boldsymbol{\chi}_t = \left[x_t^{(1)}, \ldots x_t^{(s)}\right]^{\mathsf{T}}$ – столбец матрицы (10), значение ряда $\mathbf{X}$ в момент времени t. Построим прогноз $\hat{\boldsymbol{\chi}}_t$ ряда $\mathbf{X}$ в точке $\boldsymbol{\chi}_t$. Проделаем это l раз для сегментов ряда $\mathbf{X}_{(t-n+1+i)\div(t+i)}$, где $i = 1, \ldots, l$. Получим l прогнозов $\hat{\boldsymbol{\chi}}_{t\div(t+l-1)} = [\boldsymbol{\chi}_t, \boldsymbol{\chi}_{t+1}, \ldots \boldsymbol{\chi}_{t+l-1}]$ ряда $\mathbf{X}$ в точках $\boldsymbol{\chi}_t, \boldsymbol{\chi}_{t+1}, \ldots \boldsymbol{\chi}_{t+l-1}$. Построение прогнозов $\hat{\boldsymbol{\chi}}_{t\div(t+l-1)}$ изображено на рис. 2.

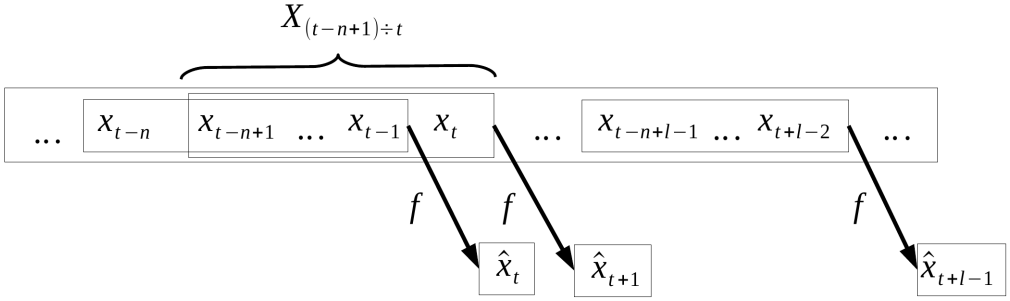


Рис. 2: Построение прогнозов для значений набора $\boldsymbol{\chi}_t, \boldsymbol{\chi}_{t+1}, \ldots \boldsymbol{\chi}_{t+l-1}$ временных рядов в моменты времени $t, t+1, \ldots, t+l-1$

.

Прогностическая модель имеет вид

$$\begin{gathered}\hat{\boldsymbol{\chi}}_t = \mathbf{f}(\hat{\mathbf{w}}, \boldsymbol{\chi}_t, \boldsymbol{\chi}_{t-1}, \ldots, \boldsymbol{\chi}_{t-n+1}), \\ \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} S(\mathbf{w}, \boldsymbol{X}, \hat{\boldsymbol{\chi}}_t, \hat{\boldsymbol{\chi}}_{t+1}, \ldots \hat{\boldsymbol{\chi}}_{t+n-1}) = S(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)}),\end{gathered} \tag{11}$$

где функция потерь

$$S(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\chi}_{t+i}, \hat{\boldsymbol{\chi}}_{t+i}) = \frac{1}{n}\sum_{i=1}^{n} \left(x_{t+i}^{(1)} - \hat{x}_{t+i}^{(1)}\right)^2. \tag{12}$$

В данной работе для построения прогноза набора из s временных рядов (9) используется алгоритм многомерной гусеницы (MSSA-L), представляющий собой обобщение 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 (1), сингулярное разложение этой матрицы (14), группировка компонент (15), полученных при сингулярном разложении, в интерпретируемые компоненты и восстановление временного ряда по каждой из интерпретируемых компонент.

По ряду $\mathbf{x}^{(i)}$ построим траекторную матрицу $\mathbf{H}^{(i)} \in \mathbb{R}^{T\times n}$, $T = N - n + 1$, согласно (1). Для набора временных рядов (10) построим объединенную траекторную матрицу

$$\mathbf{H} = [\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(s)}]. \tag{13}$$

По траекторной матрице $\mathbf{H}$ можно восстановить 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

$$\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}, \tag{14}$$

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

$$\hat{\mathbf{H}} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}. \tag{15}$$

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

Алгоритм многомерной гусеницы строит прогноз временного ряда $\mathbf{X}$ в момент t по $(n-1)$ предыдущим значениям ряда. Алгоритм находит такой вектор коэффициентов $\mathbf{p} \in \mathbb{R}^{(n-1)}$, что значения набора рядов $\mathbf{X}$ в момент t равны

$$\boldsymbol{\chi}_t = \begin{pmatrix} x_{t-n+1}^{(1)} & \ldots & x_{t-1}^{(1)} \\ x_{t-n+1}^{(2)} & \ldots & x_{t-1}^{(2)} \\ & \vdots & \\ x_{t-n+1}^{(s)} & \ldots & x_{t-1}^{(s)} \end{pmatrix} \cdot \mathbf{p}. \tag{16}$$

Заметим, что коэффициенты $\mathbf{p}$ оказываются общими для всех компонент ряда $\mathbf{X}$. Вектор p – это вектор коэффициентов в линейной комбинации первых $(n-1)$ столбцов траекторной матрицы, наилучшим образом приближающей последний столбец матрицы. Вектор p определяется методом наименьших квадратов.

Для каждого $i \in [1, r]$ обозначим через $\tilde{\mathbf{u}}_i$ первые $(n-1)$ 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

$$\mathbf{p} = \frac{1}{1-\nu^2} \sum_{i=1}^{r} \pi_i \tilde{\mathbf{u}}_i. \tag{17}$$

4 Вычислительный эксперимент

Эксперимент проводился на трех парах временных рядов: искусственные ряды, потребление электроэнергии и среднесуточная температура воздуха, объем грузоперевозок нефти и цена на нефть. Для каждой пары рядов ставилась задача нахождения связанных траекторных подпространств с помощью метода сходящегося перекрестного отображения.

4.1 Сгенерированные данные

Эксперимент проводился на двух сгенерированных рядах:

$$\mathbf{x} = \sin t + 2 \sin \frac{t}{2} + \sigma_x^2 \boldsymbol{\varepsilon}, \quad \sigma_x^2 = 0{,}3,$$

$$\mathbf{y} = \sin(2t+5) + \sigma_y^2 \boldsymbol{\varepsilon}, \quad \sigma_y^2 = 0{,}25,$$

где $\boldsymbol{\varepsilon} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$.

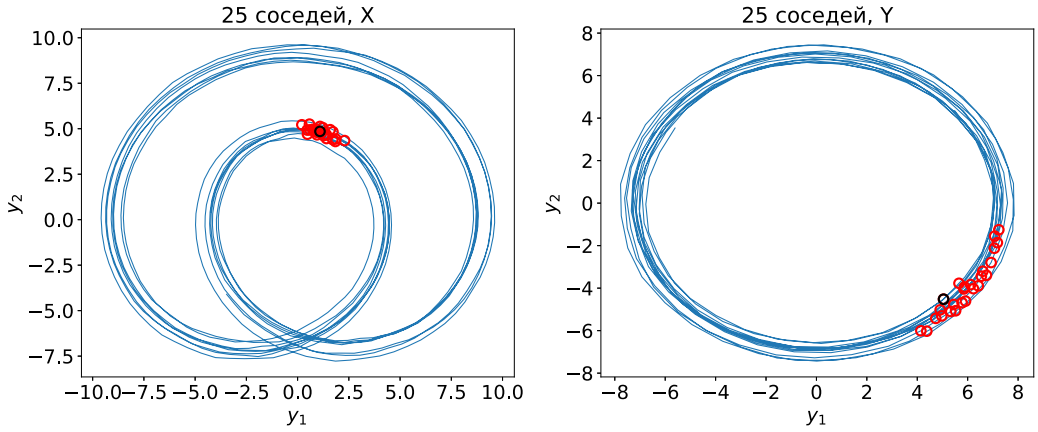
Строим траекторную матрицу $\mathbf{H_x}$ по ряду $\mathbf{x}$, взяв ширину окна $n = 250$. Для некоторого момента времени t^* рассмотрим вектор $\mathbf{x}_{t^*}$, равный t^*-й строке матрицы $\mathbf{H_x}$. Выберем k и найдем среди строк матрицы $\mathbf{H_x}$ k ближайших (в смысле евклидовой нормы) соседей вектора $\mathbf{x}_{t^*}$. Обозначим индексы найденных векторов $t_1, \ldots, t_k$, а сами найденные вектора – $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$. Обозначим i-ю строку траекторной матрицы $\mathbf{H_y}$ ряда $\mathbf{y}$ как $\mathbf{y}_i$. Тогда по найденным индексам $t_1, \ldots, t_k$ можно отобрать соответствующие $\mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$. Если ряд $\mathbf{y}$ зависит от ряда $\mathbf{x}$, то векторы $\mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$, как и $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$, будут находиться рядом в траекторном пространстве.

Аналогично для некоторого t^* находим ближайших соседей вектора $\mathbf{y}_{t^*}$. Обозначим их $\mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$. На рис. 3 изображены фазовые траектории рядов $\mathbf{x}$ и $\mathbf{y}$, а также проекции векторов $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ и $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ на фазовые траектории.

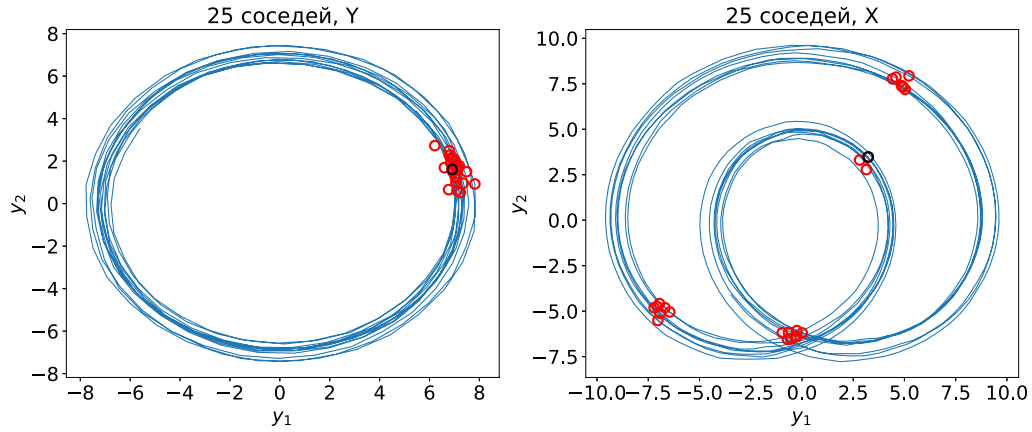
Видим, что окрестность фазовой траектории ряда $\mathbf{x}$ отображается в окрестность фазовой траектории ряда $\mathbf{y}$ и наоборот. При отображении из траекторного пространства ряда $\mathbf{y}$ в траекторное пространство ряда $\mathbf{x}$ векторы $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_{15}}$ распадаются на четыре плотные группы. Это связано с тем, что период ряда $\mathbf{y}$ в четыре раза меньше периода ряда $\mathbf{x}$.

4.2 Эксперимент на данных потребления электроэнергии и температуры

В эксперименте исследуется ряд объема потребления электроэнергии $\mathbf{x}$ и ряд значений температуры $\mathbf{y}$ в течение года. Так как эти ряды не являются стационарными, их необходимо продифференцировать и отнормировать перед тем, как исследовать зависимости между



(a) Отображение из пространства $\mathbb{H}_{\mathbf{x}}$ в пространство $\mathbb{H}_{\mathbf{y}}$



(b) Отображение из пространства $\mathbb{H}_{\mathbf{y}}$ в пространство $\mathbb{H}_{\mathbf{x}}$

Рис. 3: Проекции $U_k\left(\mathbf{x}_{t^*}\right)$ и $U_k\left(\mathbf{y}_{t^*}\right)$ на фазовые траектории

ними. Ряд температуры приведем к стационарной форме следующим образом. Рассмотрим ряд длины светового дня в течение года $\mathbf{z}$. С помощью вычисления кросскорреляционной функции (8) рядов $\mathbf{y}$ и $\mathbf{z}$ определим, насколько ряд $\mathbf{z}$ опережает ряд $\mathbf{y}$. То есть найдем такое h^*, что $\mathbf{y}(t+h^*) = \mathbf{z}(t)$. Вычтем из ряда $\mathbf{y}$ ряд $\mathbf{z}$ с учетом сдвига h^*. Полученный ряд $\mathbf{y}(t) = \mathbf{y}(t) - \mathbf{z}(t-h^*)$ будет стационарным рядом температуры.

Исходные ряды потребления электроэнергии $\mathbf{x}$, температуры $\mathbf{y}$ и длины светового дня $\mathbf{z}$ изображены на рис. 4.

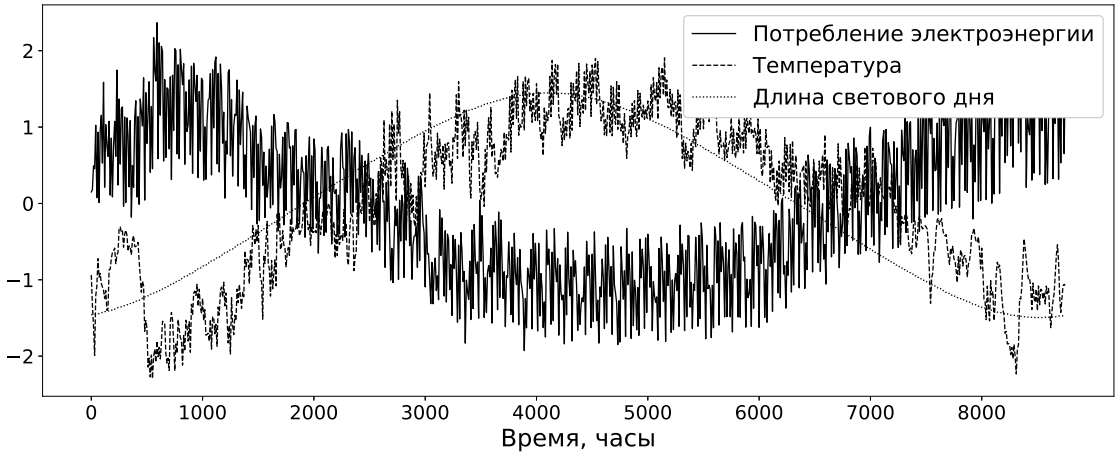


Рис. 4: Нормированные ряды потребления электроэнергии, температуры и длины светового дня

Максимум модуля кросскорреляции $\gamma_{\mathbf{yz}}(h)$ достигается при $h = 560$. Значит,

$$\mathbf{z}(t) = \mathbf{y}(t + 560).$$

И новый стационарный ряд температуры имеет вид

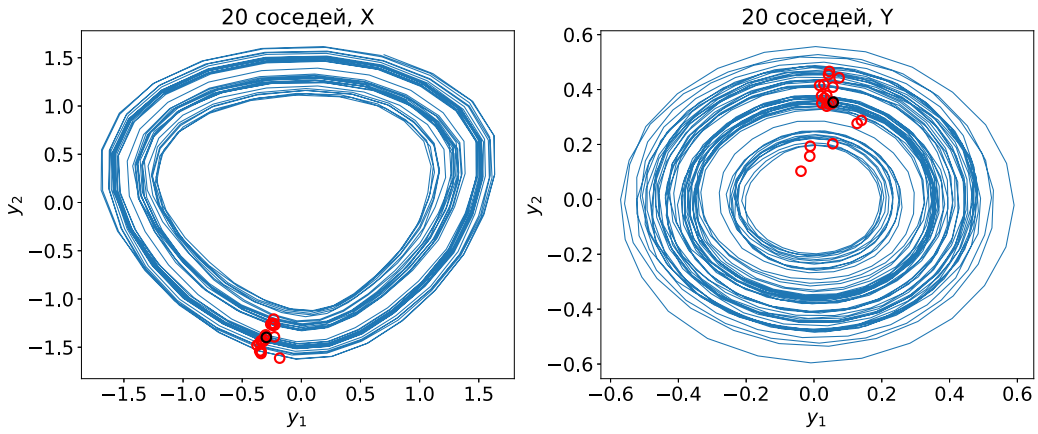
$$\mathbf{y}^*(t) = \mathbf{y}(t) - \mathbf{z}(t - 560).$$

Далее для удобства полученный ряд температуры $\mathbf{y}^*$ будем обозначать $\mathbf{y}$, как и исходный.

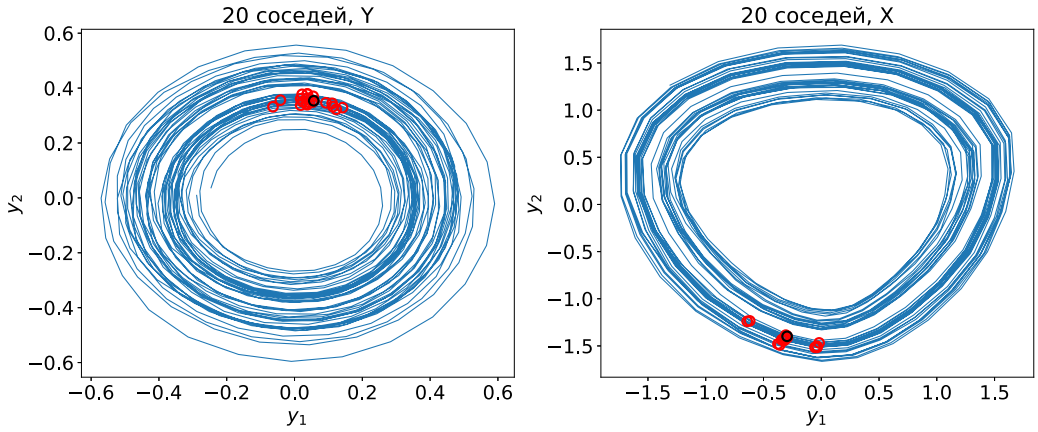
Исследуем зависимость ряда температуры $\mathbf{y}$ от ряда потребления электроэнергии $\mathbf{x}$. Делаем это аналогично эксперименту на искусственных данных. Возьмем $n = 170$, что соответствует периоду в семь дней, и $t^* = 400$. Найдем k ближайших соседей векторов $\mathbf{x}_{t^*}$ и $\mathbf{y}_{t^*}$, их расположение в траекторном пространстве показано на рис. 5.

Перебор подпространств. Переберем траекторные подпространства рядов $\mathbf{x}$ и $\mathbf{y}$ размером не больше пяти. Для этого будем перебирать пары множеств индексов главных компонент $(\mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y})$. Для каждой пары $(\mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y})$ найдем $R(\mathbf{x}, \mathbf{y} \,|\, \mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y})$, задающееся (6). Полученные значение $R(\mathbf{x}, \mathbf{y} \,|\, \mathcal{A}_\mathbf{x}, \mathcal{A}_\mathbf{y})$ представлены на рис. 6.

Построение прогноза. Построим прогноз ряда потребления электроэнергии $\mathbf{x}$, используя только его собственную историю, и сравним с прогнозом, строящимся с использованием ряда температуры $\mathbf{y}$. Рассмотрим также прогнозы, строящиеся при помощи только первых главных компонент ряда $\mathbf{y}$. На рис. 7 представлены графики зависимости среднеквадратичной ошибки прогноза от ширины окна n.



(a) Отображение из пространства $\mathbb{H}_{\mathbf{x}}$ в пространство $\mathbb{H}_{\mathbf{y}}$



(b) Отображение из пространства $\mathbb{H}_{\mathbf{y}}$ в пространство $\mathbb{H}_{\mathbf{x}}$

Рис. 5: Векторы $U_k(\mathbf{y}_{t^*})$ (ближайшие соседи вектора $\mathbf{y}_{t^*}$) и соответствующие векторы $U_k(\mathbf{x}_{t^*})$ на фазовых диаграммах с периодом 24 часа

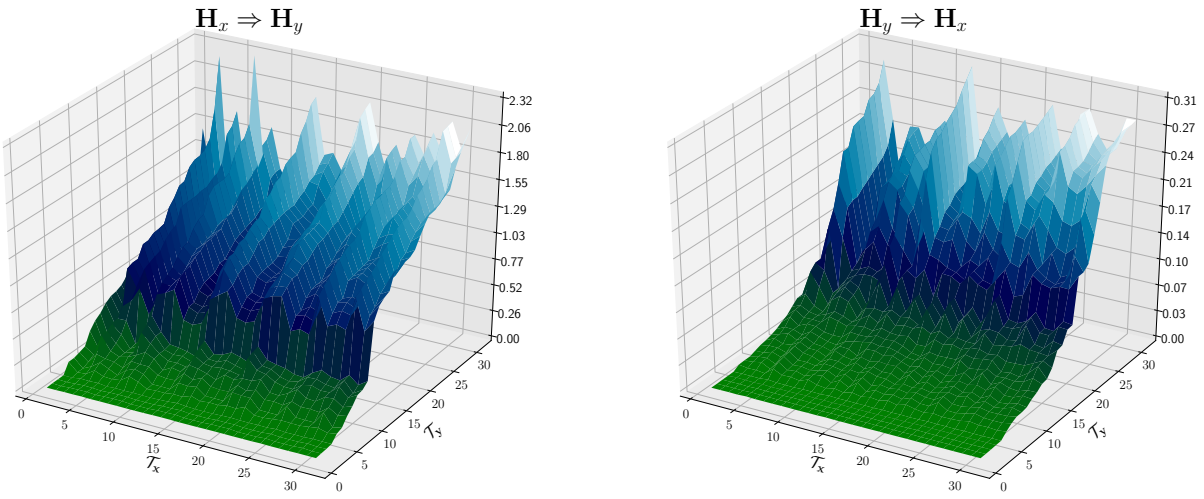


Рис. 6: Значение функционала (6) при отображении из траекторного пространства одного ряда в траекторное пространство другого

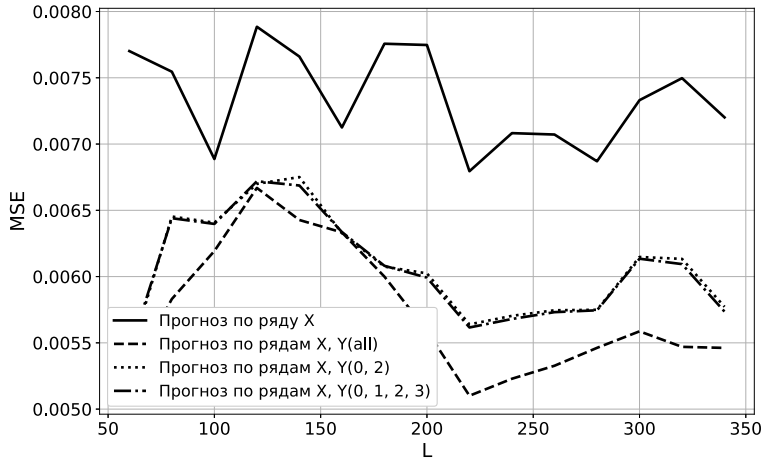


Рис. 7: Среднеквадратичная ошибка прогноза в зависимости от ширины окна

4.3 Эксперимент на данных об объеме перевозок нефти

В эксперименте проверяются связи между двумя временными рядами: объем грузоперевозок нефти $\mathbf{x}$ и объем добычи нефти $\mathbf{q}$. Временные ряды заданы за период в 21 год по месяцам. Исходные ряды представлены на рис. 8.

Исследуем связи рядов в паре $(\mathbf{x}, \mathbf{q})$. Аналогично предыдущим экспериментам, для проверки наличия связи между рядами смотрим, как точки, близкие в траекторном пространстве одного ряда, отображаются в траекторное подпространство другого ряда. Отображение между фазовыми траекториями рядов $\mathbf{x}$ и $\mathbf{q}$ показано на рис. 9.

Сравним ошибку прогноза ряда $\mathbf{x}$ с использованием ряда $\mathbf{q}$ и без него. Когда ряд $\mathbf{q}$ включен в прогностическую модель, рассмотрим два случая: для прогноза используются все компоненты ряда $\mathbf{q}$, используются только первые компоненты ряда $\mathbf{q}$. Аналогично сравним прогноз ряда $\mathbf{q}$, построенный только по его собственной истории, с прогнозом, использующим

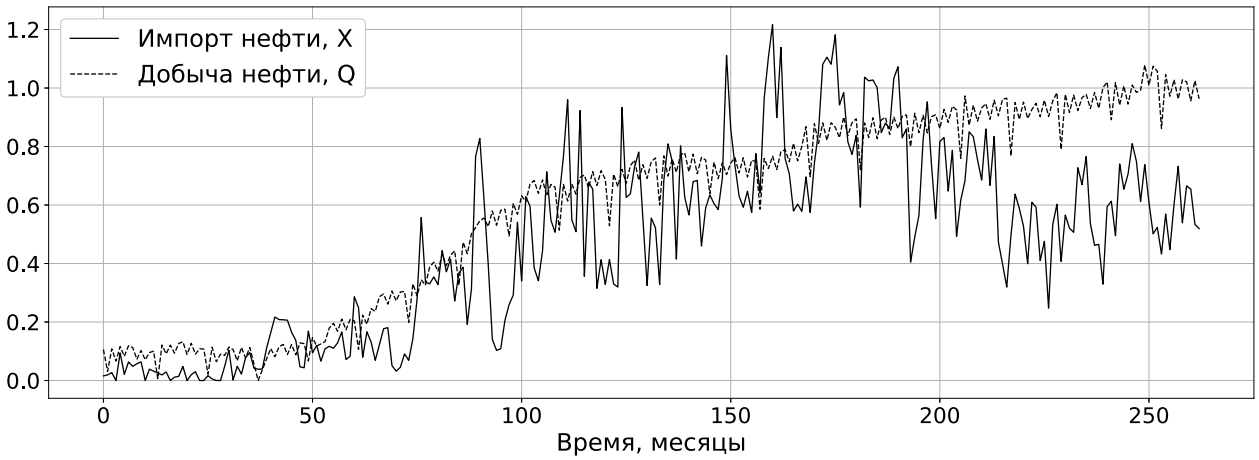
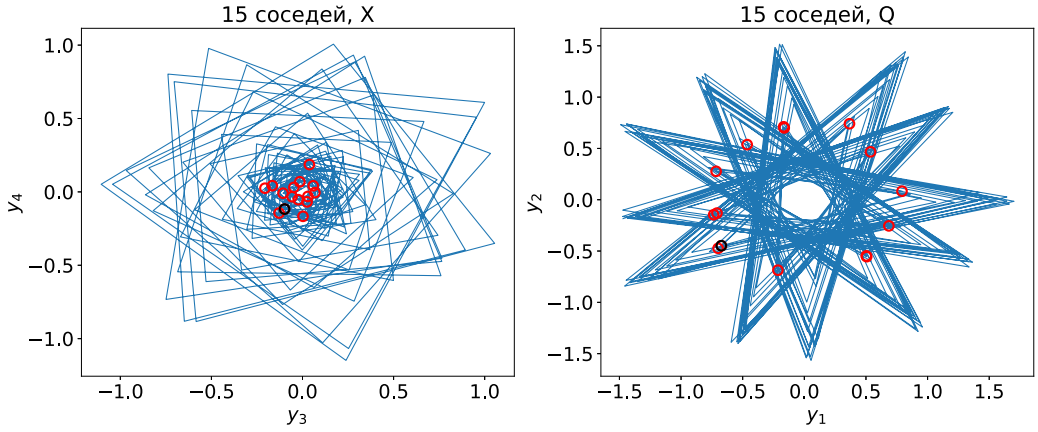
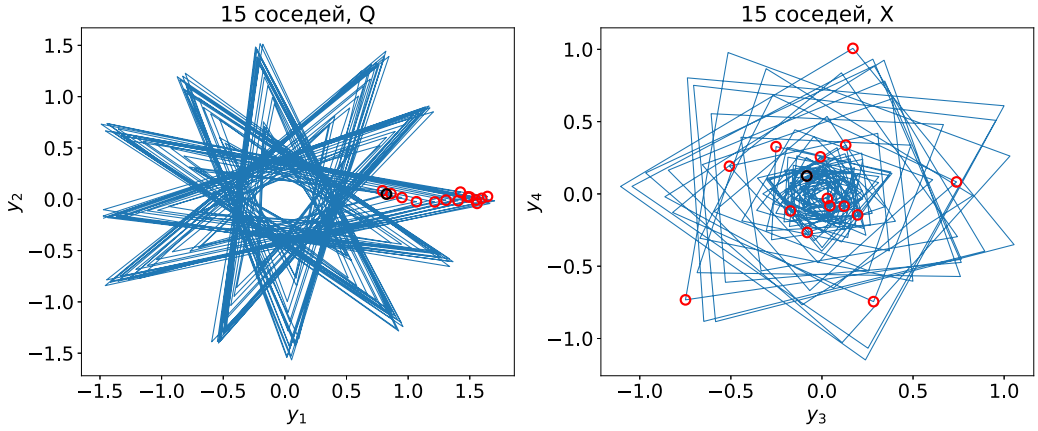


Рис. 8: Исследуемые временные ряды



(a) Отображение из пространства $\mathbb{H}_{\mathbf{x}}$ в пространство $\mathbb{H}_{\mathbf{q}}$



(b) Отображение из пространства $\mathbb{H}_{\mathbf{q}}$ в пространство $\mathbb{H}_{\mathbf{x}}$

Рис. 9: Связь между траекторными подпространствами рядов $\mathbf{x}$ и $\mathbf{q}$

историю ряда $\mathbf{x}$. Результаты эксперимента представлены на рис. 10.

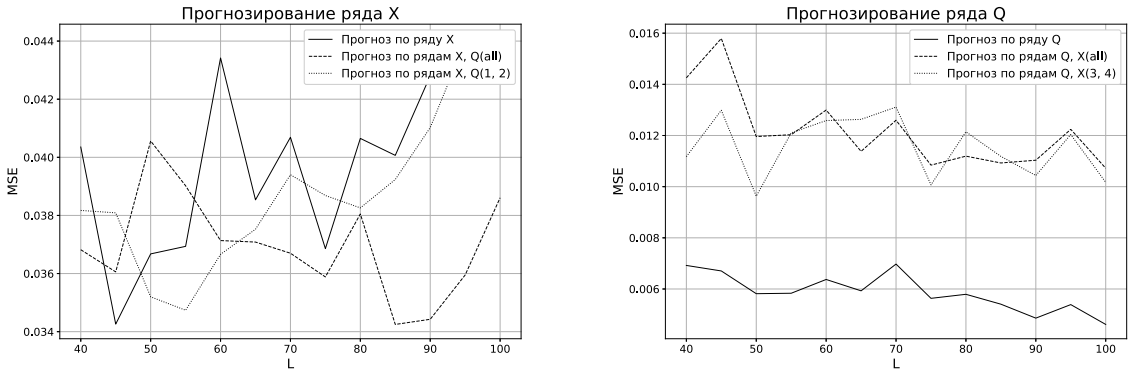


Рис. 10: Зависимость ошибки прогноза от ширины окна n для пары рядов $\mathbf{x}$ и $\mathbf{q}$

5 Заключение

В работе решалась задача обнаружения связи между временными рядами, а также между их компонентами. Связи между временными рядами устанавливались с помощью метода сходящегося перекрестного отображения. Вывод о наличии связи между парой рядов сравнивался с результатом теста Гренджера, проведенного на этих же рядах.

Эксперимент проводился на трех наборах данных: искусственные данные, данные потребления электроэнергии, данные РЖД объема грузоперевозок нефти. Для каждого набора данных сделаны выводы о наличии связей между исследуемыми рядами и их проекциями на траекторные подпространства. На данных РЖД видно, что метод CCM более чувствителен к наличию связи между рядами.

Список литературы

[1] Данные работы по исследованию связи показателей ЭКГ и пульса. http://smartlab.ws/component/content/article?id=60.

[2] *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.

[3] *Barrett A. B., Barnett L., Seth. A. K.* Multivariate Granger causality and generalized variance // Physical Review, 2010. Vol. 81. No. 4. P. 041907.

[4] *Sugihara G., May R., Hao Ye, Chih-hao H., Deyle E., Fogarty M., Munch S.* Detecting causality in complex ecosystems // Science, 2012. Vol. 338. P. 1227079.

[5] *Sugihara G., May R.* Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series // Nature, 1990. Vol. 344. No. 6268. P. 734–741.

[6] *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.

[7] *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.

[8] *White H., Xun L.* Granger causality and dynamic structural systems // Journal of Financial Econometrics, 2010. Vol. 8. No. 2. P. 193–243.

[9] *Katrutsa A. M., Strijov V. V.* Stress test procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015. Vol. 142. P. 172–183.

[10] *Li J., Cheng K., Wang S., Morstatter F., Trevino R. P., Tang J., Huan Liu.* Feature selection: A data perspective // ACM Computing Surveys, 2017. Vol. 50. No. 6. P. 1–45.

[11] *Geladi P.* Notes on the history and nature of partial least squares modelling // Journal of Chemometrics, 1988. Vol. 2. No. 4. P. 231–246.

[12] *Hoskuldsson A.* Pls regression methods // Journal of Chemometrics, 1988. Vol. 2. No. 3. P. 211–228.

[13] *Golyandina N., Stepanov D.* SSA-based approaches to analysis and forecast of multidimensional time series // Simulation 2005: Proceedings of the 5th St. Petersburg Workshop on Simulation. – Saint Petersburg: NII Chemistry Saint Petersburg University Publishers, 2005. P. 293–298.

[14] *Golyandina N., Nekrutkin V., Zhigljavsky A. A.* Analysis of time series structure: SSA and related techniques. – Chapman and Hall, 2002. 320 p.

[15] *Golyandina N., Zhigljavsky A.* Singular Spectrum Analysis for time series. – Springer Science & Business Media, 2013. 120 p.

[16] *Elsner J. B., Tsonis A. A.* Singular spectrum analysis: a new tool in time series analysis. – Springer Science & Business Media, 2013. 164 p.