Модели обнаружения зависимостей во временных рядах в задачах построения прогностических моделей

Аннотация

При прогнозировании сложноорганизованных временных рядов, зависящих от экзогенных факторов и имеющих множественную периодичность, требуется решить задачу выявления связанных рядов. Предполагается, что добавление этих рядов в модель повышает качество прогноза. Статистическая значимость повышения качества прогноза выявляется в с помощью теста Гренджера. В данной работе для обнаружения связей между временными рядами предлагается использовать метод сходящегося перекрестного отображения. При таком подходе два временных ряда считаются зависимыми, если существует отображение окрестности фазовой траектории из одного лагированного пространства в другое. Также при обнаружении причинно-следственных связей между временными рядами ставится задача обнаружения оптимального лагированного подпространства. Решение этой задачи продемонстрировано на двух парах рядов: потребления электроэнергии и температура, объема железнодорожных перевозок нефти и цена на нефть.

Ключевые слова: *временные ряды; прогнозирование; тест гренджера; сходящиеся перекрестные отображения; оптимальное лагированное подпространство*

1 Введение

Работа посвящена обнаружению причинно-следственных связей между разнородными временным рядами. Примеры зависимых разнородных временных рядов: связь эконометрических временных рядов, связь показателей ЭКГ и пульса (`http://smartlab.ws/component/content/article?id=60`)

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 упростить прогностическую модель. Если установлено, что ряд $\mathbf{x}$ не зависит от ряда $\mathbf{y}_i$, то $\mathbf{y}_i$ можно исключить из прогностической модели. В данной работе для обнаружения зависимостей между рядами в работе анализируются два подхода: тест Гренджера [1, 2] и метод сходящегося перекрестного отображения (convergent cross mapping, CCM) [3, 4].

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 $\mathbf{x}$, [1, 2]. Тест Гренджера позволяет установить 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 связан с использовавшихся во втором прогнозе рядов. 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 [5, 6].

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 [7].

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 и использованием второго. Считается, что ряд $\mathbf{x}$ точно 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 $\mathbf{y}$. Другими словами, проверяется, насколько точно моменты времени, соответствующие ближайшим соседям вектора $\mathbf{y}_t$, определяют ближайших соседей вектора $\mathbf{x}_t$. [3, 4].

При построении линейной прогностической модели по временному ряду строится траекторная матрица, играющая роль матрицы объектов. Ответами являются значения ряда в последующие моменты времени. Иногда размерность траекторного пространства очень велика, и, как следствие, прогностическая модель становится неустойчивой. В этом случае необходимо производить отбор признаков [8, 9]. Метод частных наименьших квадратов (partial least squares, PLS) отбирает наиболее значимые признаки и строит новые признаки как их линейные комбинации [10, 11]. Таким образом, PLS находит подпространство траекторного пространства, проекция в которое наилучшим образом приближает исходный ряд. Снижение размерности может применяться при изучении связей между рядами. Проекции в траекторное подпространство позволяет 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 используется алгоритм многомерной гусеницы (multivariate singular spectrum analysis, MSSA-L) [12]. Этот алгоритм является обобщением на многомерный случай алгоритма анализа спектральных компонен (singular spectrum analysis, SSA) [13, 14, 15]. 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 во временных рядах [16], подавления шума во временных рядах [17], прогнозирование временных рядов [18, 19].

2 Постановка задачи прогнозирования

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 $\hat{\mathbf{X}} = \left(\hat{\mathbf{X}}_n, \hat{\mathbf{X}}_{n+1}, \ldots \hat{\mathbf{X}}_{n+k-1}\right)$ 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

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

3 Обнаружение связей временных рядов

Тест Гренджера В качестве базового метода установления связей предлагается использовать статистический тест Гренджера. Ниже приведен алгоритм теста Гренджера для проверки наличия связи двух временных рядов. 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

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

   где n – длина тестовой выборки.

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

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

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

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 $\gamma_{\mathbf{xy}}(h)$

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

где E – математическое ожидание, $\boldsymbol{\mu}$ – выборочное среднее, σ^2 – 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.

Метод сходящегося перекрестного отображения, CCM Опишем, как строится прогноз ряда $\mathbf{y} = (y_1, \ldots, y_N)$ с помощью ряда $\mathbf{x} = (x_1, \ldots, x_N)$. Построим матрицу Ганкеля ряда $\mathbf{x}$.

$$\mathbf{H}_{\mathbf{x}} = \begin{pmatrix} x_1 & x_2 & \ldots & x_{L-1} & x_L \\ x_2 & x_3 & \ldots & x_L & x_{L+1} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{N-L+1} & x_{N-L+2} & \ldots & x_{N-1} & x_N \end{pmatrix},$$

где L – ширина окна, длина истории ряда, используемая при нахождении главных компонент. Аналогично строим матрицу $\mathbf{H}_{\mathbf{y}}$. Обозначим t-ю строку матрицы $\mathbf{H}_{\mathbf{x}}$ через $\mathbf{x}_{t+L-1}$ соответственно. Тогда матрица $\mathbf{H}_{\mathbf{x}}$ принимает вид

$$\mathbf{H}_{\mathbf{x}} = \begin{pmatrix} \mathbf{x}_L \\ \vdots \\ \mathbf{x}_N \end{pmatrix}, \quad \mathbf{x}_i = (x_{i-L+1}, \ldots, x_{i-1}, x_i), \quad i = L, \ldots, N.$$

Заметим, что все вектора $\mathbf{x}_L, \ldots, \mathbf{x}_N$ принадлежат L-мерному траекторному пространству $\mathbf{M}_{\mathbf{x}}$ ряда $\mathbf{x}$. Аналогично вводим $\mathbf{y}_t, t = L, \ldots, N$, лежащие в траекторном пространстве $\mathbf{M}_{\mathbf{y}}$ ряда $\mathbf{y}$. Выберем

момент $t \in [L, N]$ и найдем k ближайших соседей вектора $\mathbf{x}_t$ в $\mathbf{M}_\mathbf{x}$. Обозначим их индексы через $t_1, \ldots, t_k$. Тогда ближайшие соседи $\mathbf{x}_t$ это строки матрицы $\mathbf{H}_\mathbf{x}$ с номерами $(t_1 - L + 1), \ldots, (t_k - L + 1)$:

$$\mathbf{x}_{t_i} = (x_{t_i - L + 1}, \ldots, x_{t_i - 1}, x_{t_i}), \quad i = 1, \ldots, k$$

Прогноз $\hat{y}_t$ строится следующим образом:

$$\hat{y}_t = \sum_{i=1}^{k} \mathrm{w}_i y_{t_i}, \quad \text{где } t_i\text{– индексы ближайших соседей } \mathbf{x}_t$$

$$\mathrm{w}_i = \frac{u_i}{\sum_i u_i}, \quad u_i = \exp - \left(\frac{||\mathbf{x}_t - \mathbf{x}_{t_i}||_2}{||\mathbf{x}_t - \mathbf{x}_{t_{L+1}}||_2} \right)$$

Аналогично строится прогноз ряда $\mathbf{X}$ с использованием ряда $\mathbf{Y}$.

Покажем, как описанный подход можно применяется для обнаружения зависимости между рядами. Пусть выбран момент времени t^* и вектор $\mathbf{x}_{t^*} = (x_{t^* - L + 1}, \ldots, x_{t^* - 1}, x_{t^*})$. И пусть $\mathbf{x}_{t_1} \ldots, \mathbf{x}_{t_k}$ – ближайшие соседи вектора $\mathbf{x}_{t^*}$. Тогда вектора $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ – строки матрицы $\mathbf{H}_\mathbf{y}$, соответствующие индексам $t_1, \ldots, t_k$. Тогда, если вектора $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ расположены в $\mathbf{M}_\mathbf{y}$ достаточно близко, то утверждается, что ряд $\mathbf{x}$ зависит от ряда $\mathbf{x}$.

Введем меру близости $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ следующим образом:

$$S(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{dist}(\mathbf{x})}{\mathrm{dist}(\mathbf{y})}, \quad \mathrm{dist}(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} ||\mathbf{x}_{t^*} - \mathbf{x}_{t_i}||_2$$

Если $S(\mathbf{x}, \mathbf{y})$ меньше некоторого порога s, то ряд $\mathbf{y}$ зависит от ряда $\mathbf{x}$.

Заметим, что можно рассматривать ближайших соседей не во всем траекторном пространстве $\mathbf{M}_\mathbf{x}$ и $\mathbf{M}_\mathbf{y}$, а только в некотором его подпространстве, натянутом на первые главные компоненты. Пусть сингулярное разложение матрицы $\mathbf{H}_\mathbf{x}$ имеет вид

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

Пусть $\mathcal{T}_\mathbf{x}$ – некоторый набор индексов компонент ряда $\mathbf{x}$. Построим проекцию ряда $\mathbf{x}$ на подпространство, натянутое на компоненты с номерами из $\mathcal{T}_\mathbf{x}$. Обозначим это подпространство $\mathbf{M}_{\mathcal{T}_\mathbf{x}}$. Заменим в матрице $\mathbf{\Lambda}_\mathbf{x}$ элементы, находящиеся в строках с индексами, не из $\mathcal{T}_\mathbf{x}$, нулями. Обозначим полученную матрицу $\tilde{\mathbf{\Lambda}}_\mathbf{x}$. Тогда проекция ряда $\mathbf{x}$ в подпространство, натянутое на компоненты с индексами из $\mathcal{T}_\mathbf{x}$ задается траекторной матрицей

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

Аналогично по некоторому набору $\mathcal{T}_\mathbf{y}$ строится подпространство $\mathbf{M}_{\mathcal{T}_\mathbf{y}}$ и траекторная матрица $\mathbf{P}_{\mathcal{T}_\mathbf{y}}$. Далее предлагается искать ближайших соседей не в полных траекторных пространствах $\mathbf{M}_\mathbf{x}$ и $\mathbf{M}_\mathbf{y}$, задающихся траекторными матрицами $\mathbf{H}_x$ и $\mathbf{H}_\mathbf{y}$ соответственно, а в подпространствах $\mathbf{M}_{\mathcal{T}_\mathbf{x}}$ и $\mathbf{M}_{\mathcal{T}_\mathbf{y}}$, задающихся матрицами $\mathbf{P}_{\mathcal{T}_\mathbf{x}}$ и $\mathbf{P}_{\mathcal{T}_\mathbf{y}}$.

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

Зависимость рядов в выбранных подпространствах устанавливается аналогично зависимости в полных пространствах $\mathbf{M_x}$ и $\mathbf{M_y}$. Пусть $\mathcal{T}_\mathbf{x}$ и $\mathcal{T}_\mathbf{y}$ – наборы индексов главных компонент рядов $\mathbf{x}$ и $\mathbf{y}$ соответственно. Тогда $\mathbf{P}_{\mathcal{T}_\mathbf{x}}$ и $\mathbf{P}_{\mathcal{T}_\mathbf{y}}$ – траекторные матрицы проекций рядов $\mathbf{x}$ и $\mathbf{y}$ в подпространства, натянутые на $\mathcal{T}_\mathbf{x}$ и $\mathcal{T}_\mathbf{y}$ соответственно. Для фиксированного t^* находим k ближайших соседей $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ и соответствующие им $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$. Здесь $\mathbf{x}_t$ и $\mathbf{y}_t$ – строки матриц $\mathbf{P}_{\mathcal{T}_\mathbf{x}}$ и $\mathbf{P}_{\mathcal{T}_\mathbf{y}}$ соответственно.

Будем перебирать различные комбинации индексов главных компонент и соответствующие им подпространства $\mathbf{M}_{\mathcal{T}_\mathbf{x}}$ и $\mathbf{M}_{\mathcal{T}_\mathbf{y}}$. Для каждой пары $(\mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y})$ 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

$$S(\mathbf{x}, \mathbf{y}, \mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y}) = \frac{\mathrm{dist}(\mathbf{x}, \mathcal{T}_\mathbf{x})}{\mathrm{dist}(\mathbf{y}, \mathcal{T}_\mathbf{y})}, \quad \mathrm{dist}(\mathbf{x}, \mathcal{T}_\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} ||\mathbf{x}_{t^*} - \mathbf{x}_{t_i}||_2 \tag{2}$$

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

$$(\mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y}) = \arg\max_{\mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{x}} S(\mathbf{x}, \mathbf{y}, \mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y}), \tag{3}$$

$$|\mathcal{T}_\mathbf{x}| \to \min,$$

$$|\mathcal{T}_\mathbf{y}| \to \min$$

4 Алгоритм многомерной гусеницы (MSSA-L)

В данной работе в качестве прогностической модели $\mathbf{f}$ s-мерного ряда $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)})^\mathsf{T}$ 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.

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 $\mathbf{x}^{(i)}$,

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

По матрице Ганкеля $\mathbf{H}$ 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

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

где $\lambda_1, \dots, \lambda_d > 0$ – 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:

$$\mathbf{x}_N = \begin{pmatrix} x_{N-L+1}^{(1)} & \dots & x_{N-1}^{(1)} \\ x_{N-L+1}^{(2)} & \dots & x_{N-1}^{(2)} \\ & \vdots & \\ x_{N-L+1}^{(s)} & \dots & x_{N-1}^{(s)} \end{pmatrix} \cdot \mathbf{p} = \mathbf{Y} \cdot \mathbf{p} \tag{4}$$

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:

$$\mathbf{p} = \frac{1}{1-\nu^2} \sum_{i=1}^{r} \pi_i \tilde{\mathbf{u}}_i \tag{5}$$

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

5 Метод частных наименьших квадратов

Пусть поставлена задача построения прогноза временного ряда $\mathbf{x}$ на несколько моментов вперед. Пусть $\mathbf{X} \in \mathbb{R}^{m\times n}$ – траекторная матрица ряда $\mathbf{x}$, $\mathbf{Y} \in \mathbb{R}^{m\times r}$ – значения ряда в последующие моменты времени. Предполагается, что между строками матриц $\mathbf{X}$ и $\mathbf{Y}$ существует линейная зависимость:

$$\mathbf{Y}_i = \mathbf{X}_i \cdot \mathbf{\Theta} + \boldsymbol{\varepsilon}, \quad \mathbf{X}_i \in \mathbb{R}^n,\ \mathbf{Y}_i \in \mathbb{R}^r,\ i = 1, \dots m,$$

где $\mathbf{\Theta}$ – матрица параметров модели, $\boldsymbol{\varepsilon}$ – вектор ошибок прогноза.

Ошибка прогноза вычисляется следующим образом:

$$S(\Theta, \mathbf{X}, \mathbf{Y}) = ||\mathbf{Y} - \mathbf{X} \cdot \mathbf{\Theta}||_2^2 = \sum_{i=} ||\mathbf{Y}_i - \mathbf{X}_i \cdot \mathbf{\Theta}||_2^2$$

Для нахождения параметров модели $\mathbf{\Theta}$ используется метод частных наименьших квадратов PLS. Алгоритм PLS находит в латентном пространстве матрицу $\mathbf{T} \in \mathbb{R}^{m \times l}$, наилучшим образом описывающую матрицы $\mathbf{X}$ и $\mathbf{Y}$. Матрицы $\mathbf{X}$ и $\mathbf{Y}$ проецируются в латентное пространство следующим образом:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^T + \mathbf{F} = \sum_{k=1}^{l} \mathbf{t}_k \cdot \mathbf{p}_k^T + \mathbf{F}$$

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{Q}^T + \mathbf{E} = \sum_{k=1}^{l} \mathbf{t}_k \cdot \mathbf{q}_k^T + \mathbf{E}$$

где $\mathbf{T}$ — матрица совместного описания объектов и ответов в латентном пространстве, причём столбцы матрицы $\mathbf{T}$ ортогональны; $\mathbf{P},\ \mathbf{Q}$ — матрицы перехода из латентного пространства в исходные пространства; $\mathbf{E},\ \mathbf{F}$ — матрицы невязок. Алгоритм PLS находит матрицы $\mathbf{T}, \mathbf{P}, \mathbf{Q}$, а также такую матрицу $\mathbf{W}$, что параметры модели можно вычислить по формуле

$$\Theta = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T$$

6 CCM, эксперимент

6.1 Сгенерированные данные

Эксперимент проводился на двух сгенерированных рядах $\mathbf{x}$ и $\mathbf{x}$:

$$\mathbf{x} = \sin t + 2\sin\frac{t}{2} + \sigma_x^2\boldsymbol{\varepsilon}, \quad \sigma_x^2 = 0.3$$

$$\mathbf{x} = \sin(2t + 5) + \sigma_y^2\boldsymbol{\varepsilon}, \quad \sigma_y^2 = 0.25,$$

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

Строим матрицу Ганкеля $\mathbf{H_x}$ по ряду $\mathbf{x}$, взяв ширину окна $L = 250$. 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 – $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$.

На рисунке изображен ряд $\mathbf{x}$ и $k = 25$ ближайших соседей для момента $t^* = 15$. Моменты времени $t_1, \ldots, t_k$ выделены красным, момент t^* – черным.

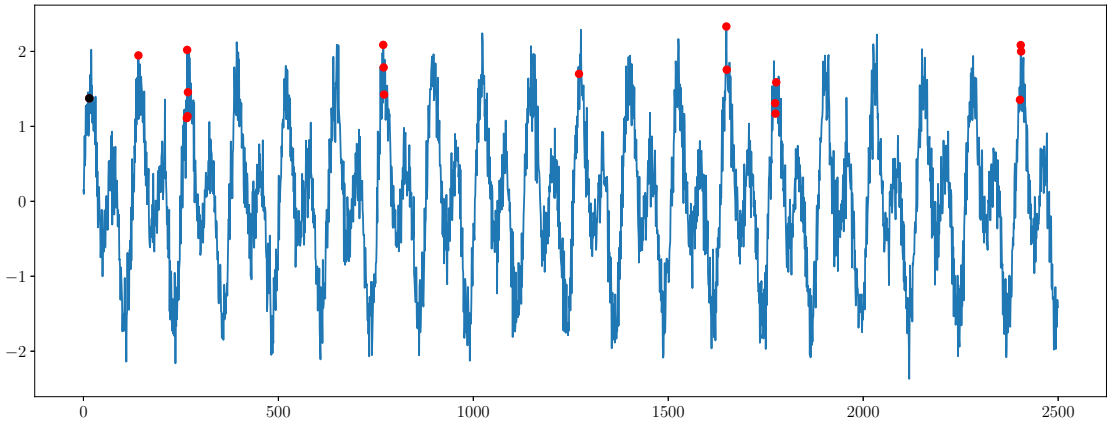

Рис. 1: Ближайшие соседи точки $\mathbf{x}_{15}$

Строим матрицу Ганкеля $\mathbf{H_y}$ по ряду $\mathbf{x}$. Обозначим i-ю строку $\mathbf{H_y}$ через $\mathbf{y}_i$. Тогда по найденным индексам $t_1, \ldots, t_k$ можно отобрать соответствующие $\mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$. Если ряд $\mathbf{x}$ зависит от ряда $\mathbf{x}$, то вектора $\mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$, как и $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$, будут находиться рядом в траекторном пространстве. Изобразим $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{y}_{t_k}$ и $\mathbf{y}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{y}_{t_k}$ на фазовых траекториях.

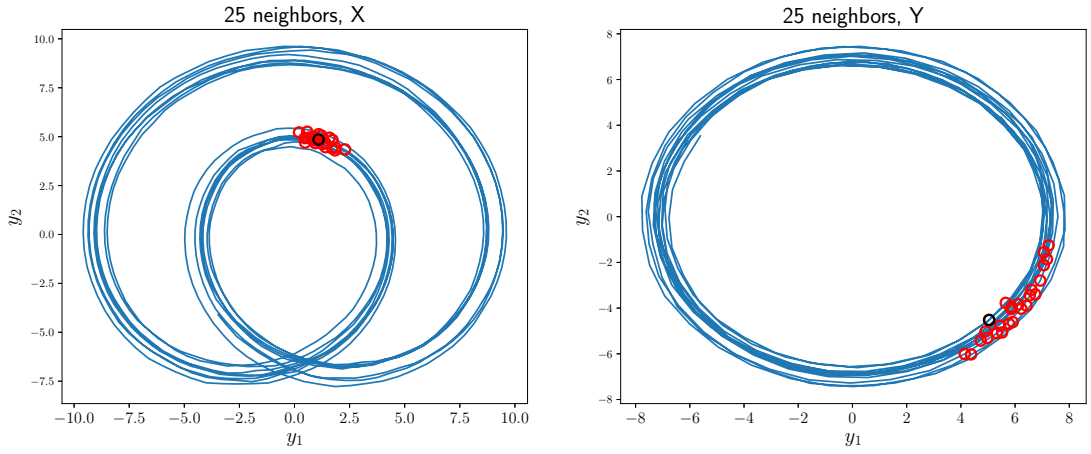

Рис. 2: Точки на фазовых траекториях рядов $\mathbf{x}$ и $\mathbf{x}$, соответствующие 15-ти ближайшим соседям $\mathbf{x}_{15}$

Видно, что точки обоих фазовых траекториях расположены близко друг другу. Значит, ряд $\mathbf{x}$ зависит от ряда $\mathbf{x}$.

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 изображен ряд $\mathbf{x}$ и $k = 25$ ближайших соседей вектора $\mathbf{y}_{20}$.

Изобразим $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ и соответствующие им $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ на фазовых траекториях (рис. 4).

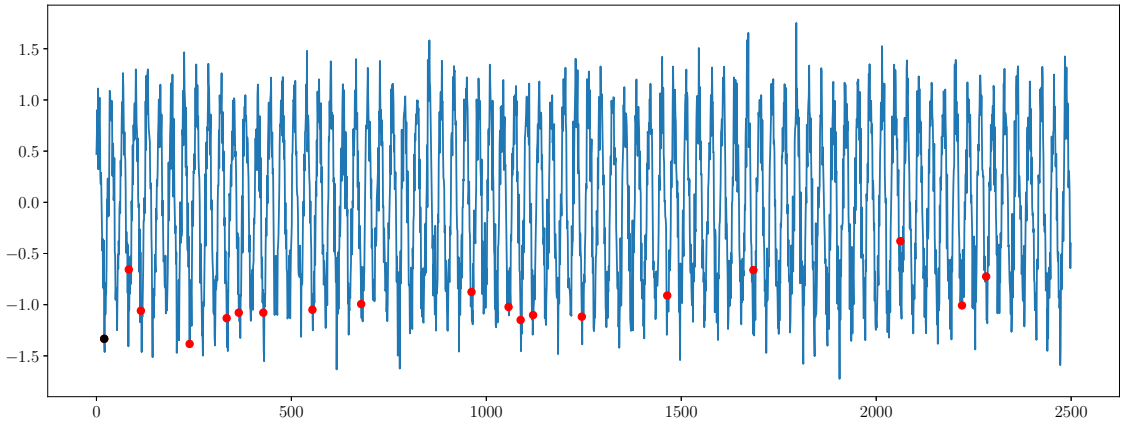

Рис. 3: Ближайшие соседи точки $\mathbf{y}_{20}$

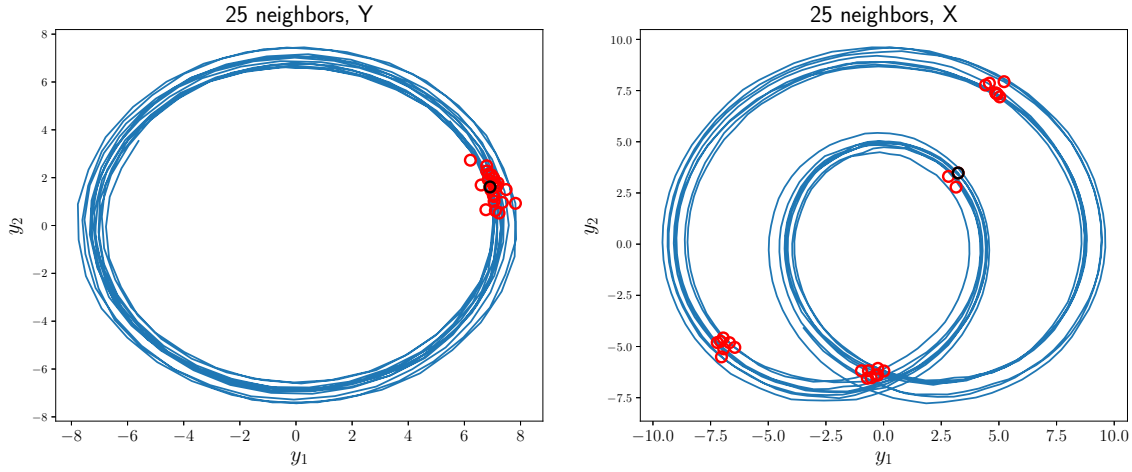

Рис. 4: Точки на фазовых траекториях рядов $\mathbf{x}$ и $\mathbf{x}$, соответствующие 15-ти ближайшим соседям $\mathbf{y}_{20}$

Видим, что точки $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_{15}}$ расположены на фазовых траекториях близко друг к другу. При этом они распадаются на четыре плотные группы. Это связано с тем, что период ряда $\mathbf{x}$ в четыре раза меньше периода ряда $\mathbf{x}$.

6.2 Эксперимент на данных потребления электроэнергии и температуры

В эксперименте исследуются ряд объема потребления электроэнергии $\mathbf{x}$ и ряд значений температуры $\mathbf{x}$ в течение года. Так как эти ряды не являются стационарными, их необходимо продифференцировать и отнормировать перед тем, как исследовать зависимости между ними. Ряд температуры будем приводить к стационарной форме следующим образом. Рассмотрим ряд длины светового дня в течение года $\mathbf{Z}$. С помощью вычисления кросс-корреляционной функции $\gamma_{\mathbf{xy}}(h)$ рядов $\mathbf{x}$ и $\mathbf{Z}$. Определим, насколько ряд $\mathbf{Z}$ опережает ряд $\mathbf{x}$. То есть найдем такое h^*, что $\mathbf{x}(t + h^*) = \mathbf{Z}(t)$. Вычтем

из ряда $\mathbf{x}$ ряд $\mathbf{Z}$ с учетом сдвига h^*. Полученный ряд $\mathbf{x}(t) = \mathbf{x}(t) - \mathbf{Z}(t - h^*)$ будет стационарным рядом температуры.

Исходные ряды потребления электроэнергии $\mathbf{x}$, температуры $\mathbf{x}$ и длины светового дня $\mathbf{Z}$ изображены на рис. 5.

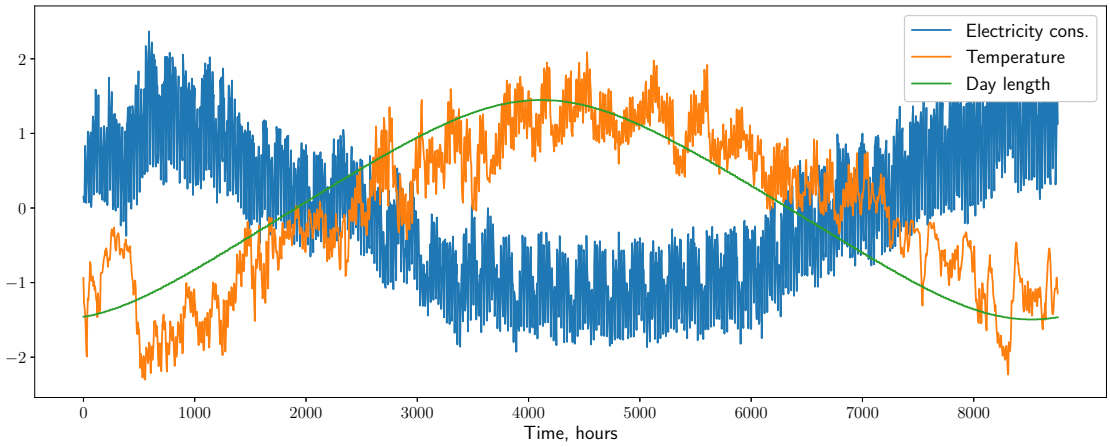

Рис. 5: Нормированные ряды потребления электроэнергии, температуры и длины светового дня

Построим кросс-корреляционную диаграмму рядов $\mathbf{x}$ и $\mathbf{Z}$

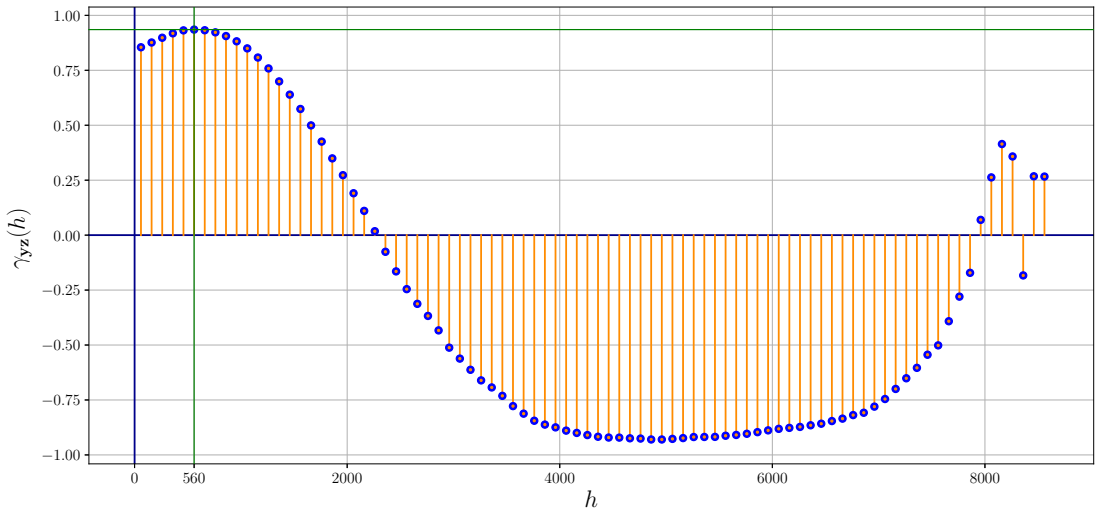

Рис. 6: Кросс-корреляционная диаграмма для ряда температуры $\mathbf{x}$ и длины светового дня $\mathbf{Z}$

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

$$\mathbf{Z}(t) = \mathbf{x}(t + 560).$$

И новый стационарный ряд температуры имеет вид

$$\mathbf{x}^*(t) = \mathbf{x}(t) - \mathbf{Z}(t - 560).$$

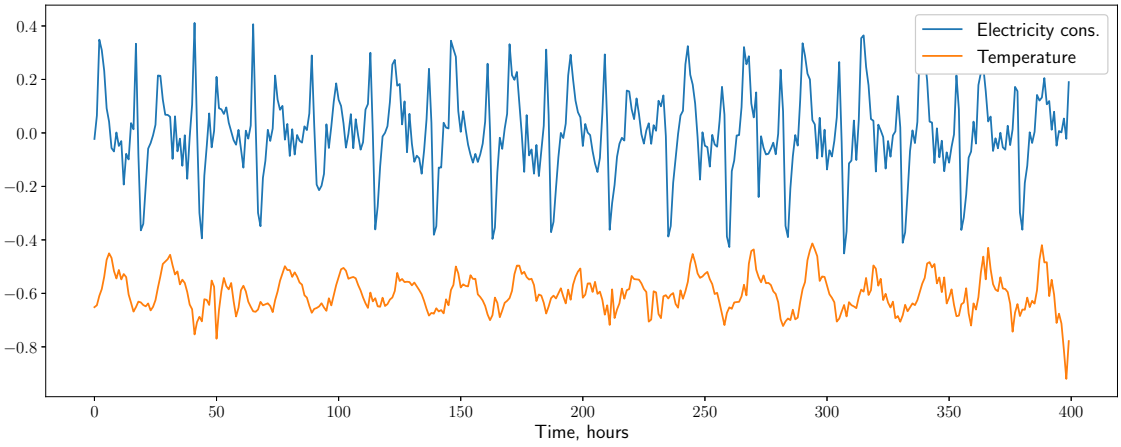

Рис. 7: Продифференцированные и нормированные ряды потребления электроэнергии и температуры

Далее, для удобства, полученный ряд температуры $\mathbf{x}^*$ будем обозначать $\mathbf{x}$. Продифференцированные и нормированные ряды потребления электроэнергии и температуры изображены на рис. 7.

Исследуем зависимость ряда температуры $\mathbf{y}$ от ряда потребления электроэнергии $\mathbf{x}$. Делаем это аналогично эксперименту на искусственных данных. Выбираем ширину окна L и некоторый момент времени t^*. Находим k ближайших соседей векторов $\mathbf{x}_{t^*}$ и $\mathbf{y}_{t^*}$ и их расположение в траекторном пространстве.

Возьмем $L = 170$, что соответствует периоду в семь дней. Возьмем $t^* = 400$. На рис. 8 красным показаны ближайшие соседи вектора $\mathbf{x}_{t^*}$

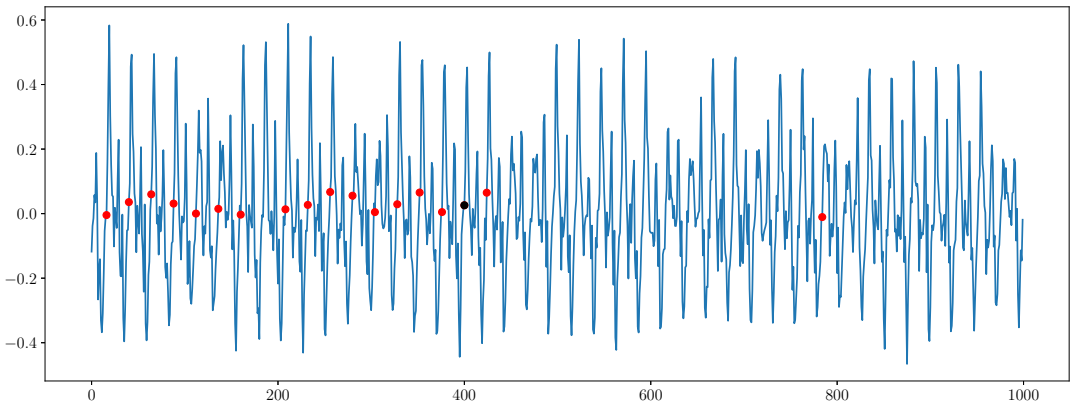

Рис. 8: Ближайшие соседи вектора $\mathbf{x}_{t^*}$, ширина окна $L = 170$

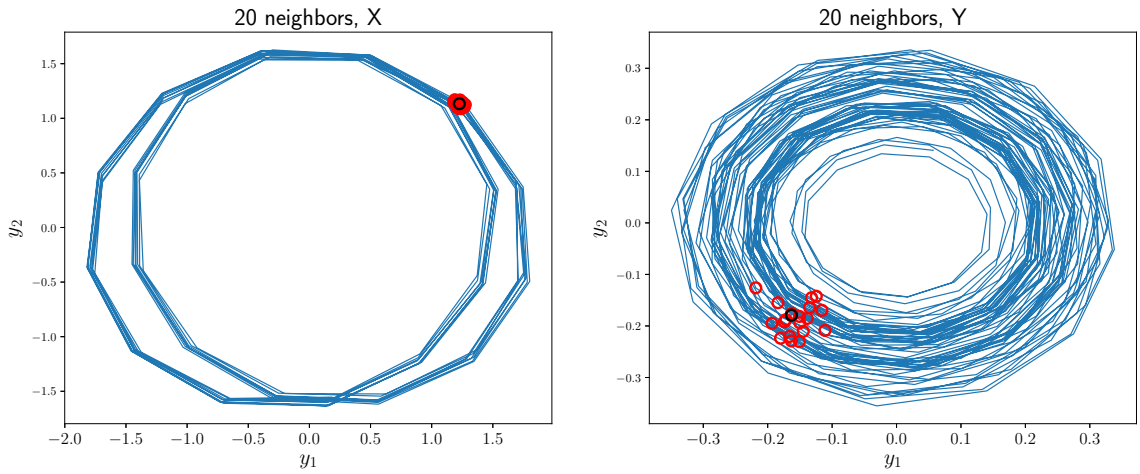

Рис. 9: Вектора $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ (ближайшие соседи вектора $\mathbf{x}_{t^*}$) и соответствующие вектора $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ на фазовых диаграммах с периодом 12 часов

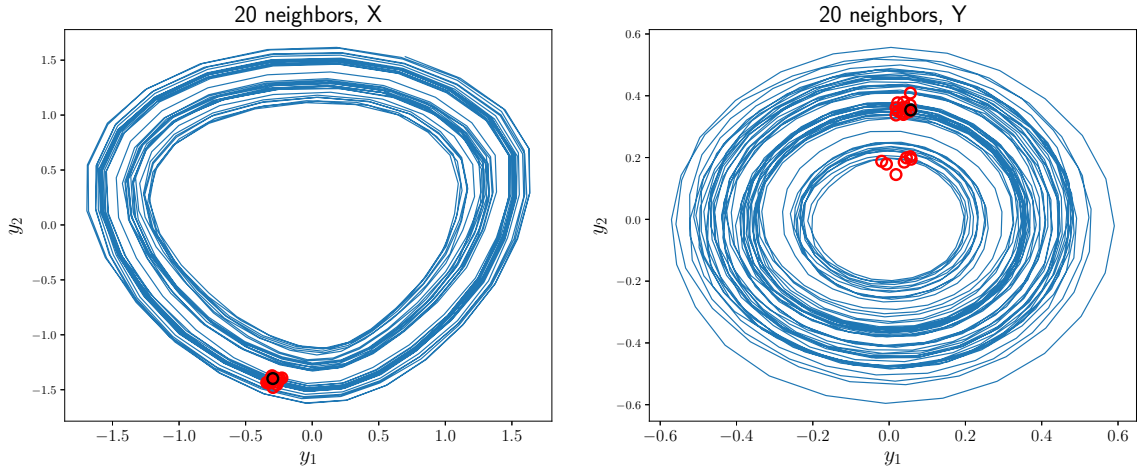

Рис. 10: Вектора $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ (ближайшие соседи вектора $\mathbf{x}_{t^*}$) и соответствующие вектора $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ на фазовых диаграммах с периодом 24 часа

Исследуем зависимость ряда $\mathbf{x}$ от ряда $\mathbf{y}$. Возьмем $t^* = 400$, $L = 170$. На рис. 11 красным показаны ближайшие соседи вектора $\mathbf{y}_{t^*}$

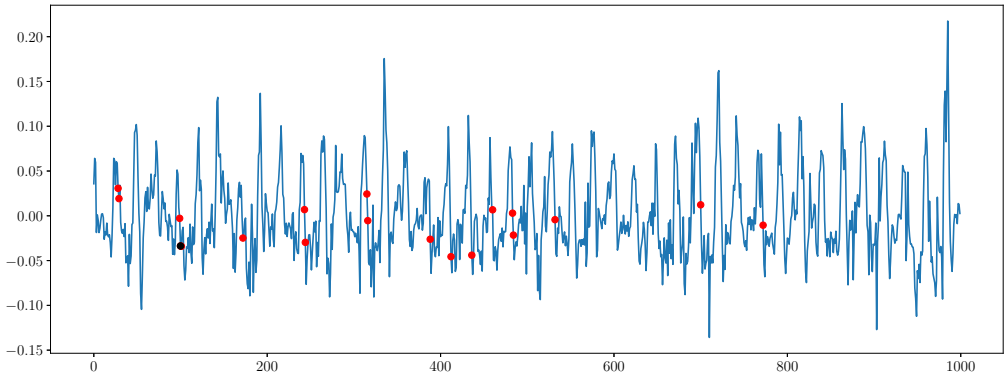

Рис. 11: Ближайшие соседи вектора $\mathbf{y}_{t^*}$, ширина окна $L = 100$

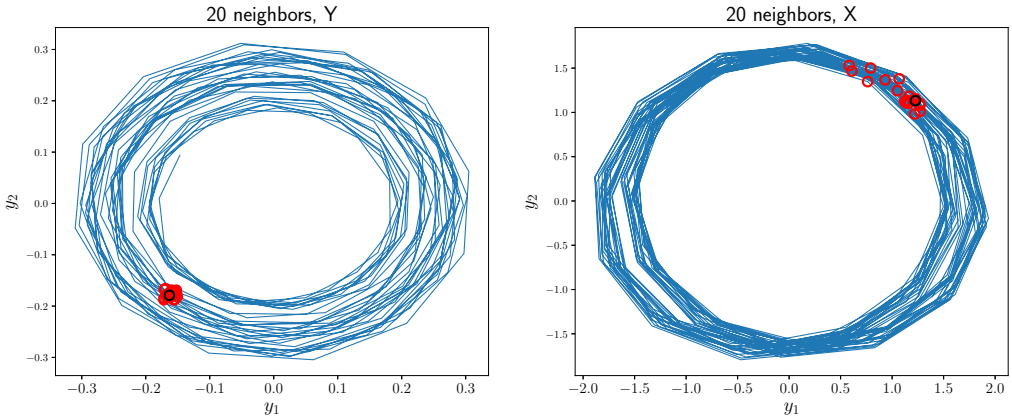

Рис. 12: Вектора $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ (ближайшие соседи вектора $\mathbf{y}_{t^*}$) и соответствующие вектора $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ на фазовых диаграммах с периодом 12 часов

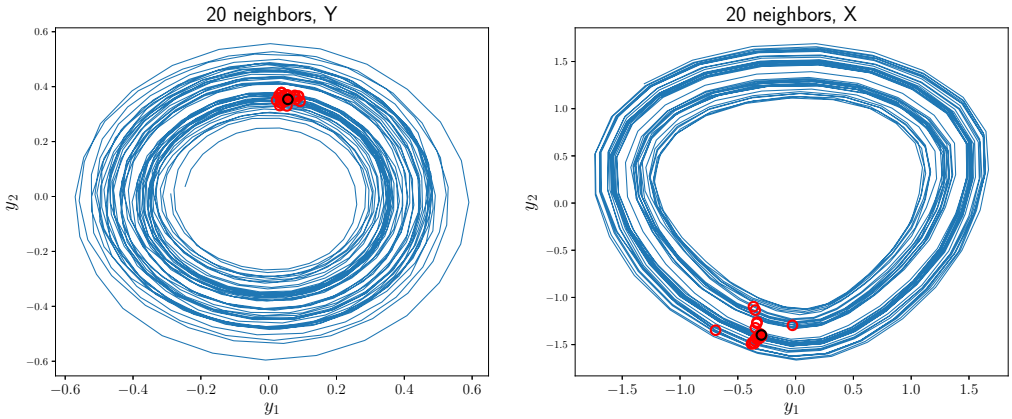

Рис. 13: Вектора $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ (ближайшие соседи вектора $\mathbf{y}_{t^*}$) и соответствующие вектора $\mathbf{x}_{t^*}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ на фазовых диаграммах с периодом 24 часа

6.3 Перебор подпространств

Переберем траекторные подпространства рядов $\mathbf{x}$ и $\mathbf{y}$ размера не больше пяти. Для этого будет перебирать пары множеств индексов главных компонент $(\mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y})$. Для каждой пары $(\mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y})$ найдем $S(\mathbf{x}, \mathbf{y}, \mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y})$, задающееся (2)

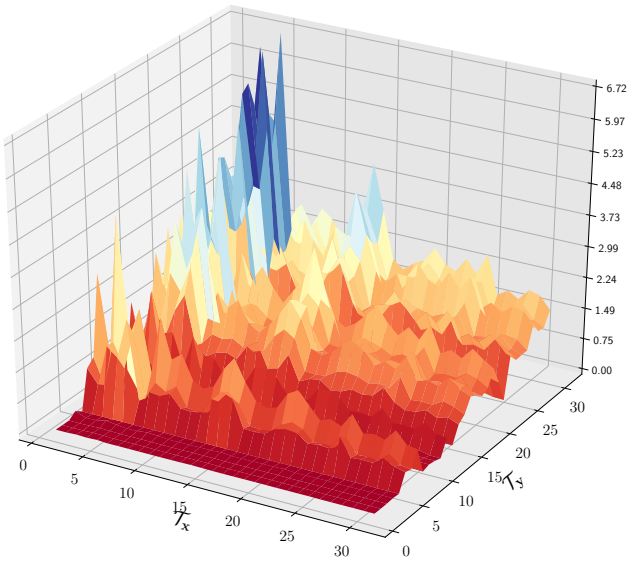

Рис. 14: Отношение расстояния между ближайшими соседями ряда $\mathbf{x}$ к расстоянию между соседями ряда $\mathbf{x}$. Ближайшие соседи определяются по ряду $\mathbf{x}$

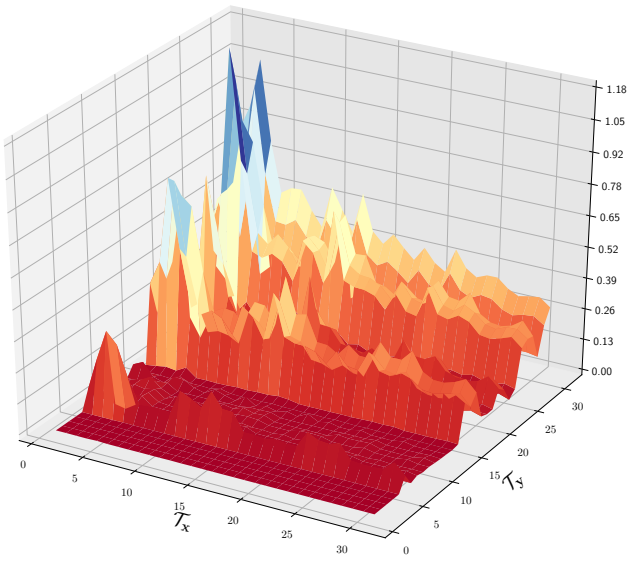

Рис. 15: Отношение расстояния между ближайшими соседями ряда **x** к расстоянию между соседями ряда **x**. Ближайшие соседи определяются по ряду **x**

6.4 Эксперимент на данных объема грузоперевозок и температуры

В эксперименте исследуется ряд объема грузоперевозок нефти в омской области **x** и ряд температуры **y**. Ряды заданы в течение года с частотой в один день. Так как ряд температуры не является стационарным, то его необходимо привести к необходимой форме аналогично пункту 6.2. На рис. 16 изображены исходные временные ряды **x** и **y**, а также ряд длины светового дня **z**.

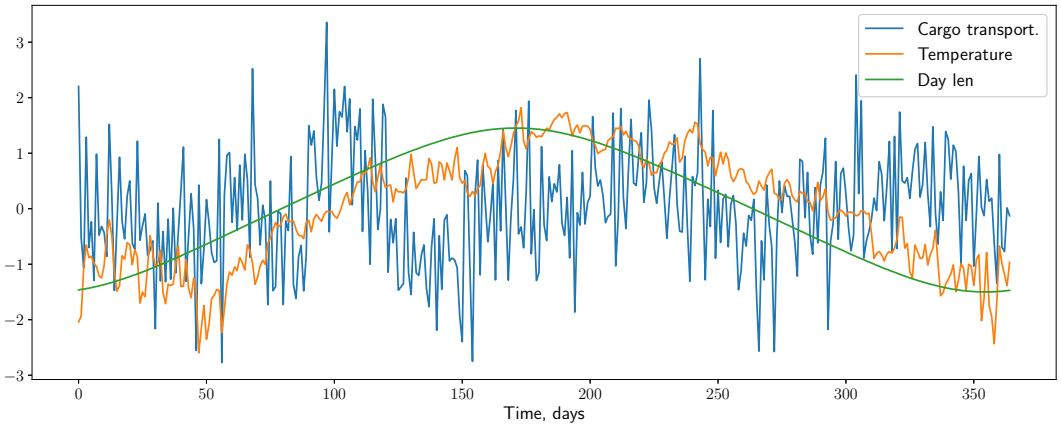

Рис. 16: Нормированные ряды объема грузоперевозок, температуры и длины светового дня

С помощью вычисления кросс-корреляционной функции $\gamma_{\mathbf{yz}}$ рядов **y** и **z** найдем, насколько ряд температуры **y** отстает от ряда длины дня **z**, и вычтем ряд **z** из ряда **y** с учетом найленного сдвига. На рис. 17 изображена кросс-корреляционная диаграмма рядов **x** и **z**

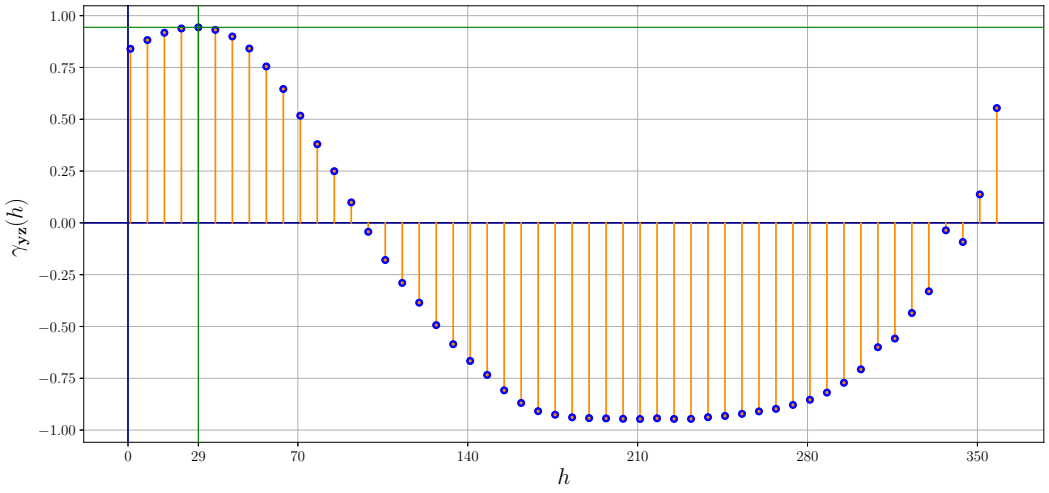

Рис. 17: Кросс-корреляционная диаграмма для ряда температуры **x** и длины светового дня **z**

Продифференцированные и нормированные ряды **x** и **y** изображены на рис. 18.

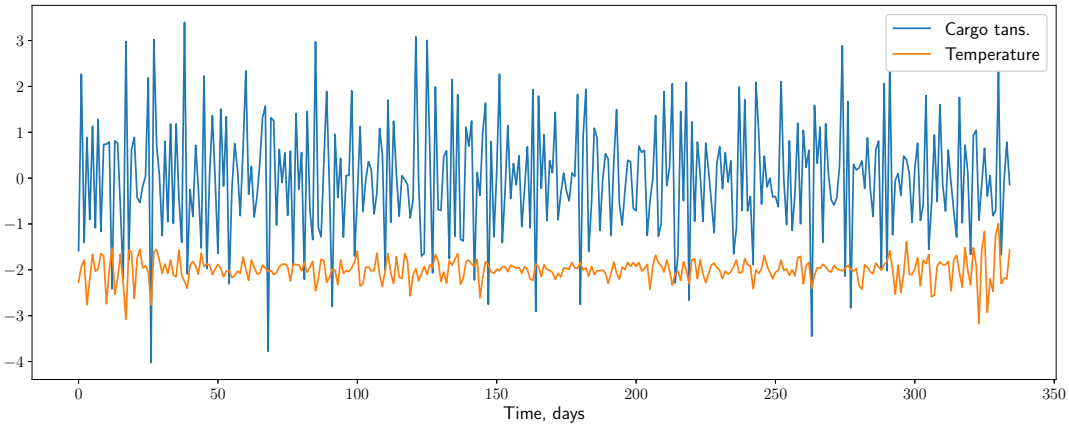

Рис. 18: Продифференцированные и нормированные ряды объема грузоперевозок $\mathbf{x}$ и температуры $\mathbf{y}$

Исследуем зависимость между рядами $\mathbf{x}$ и $\mathbf{y}$. Для этого зафиксируем ширину окна $L = 15$, что соответствует периоду в две недели. Построим траекторные матрицы $\mathbf{H_x}$ и $\mathbf{H_y}$. Их сингулярные числа изображены на рисунке:

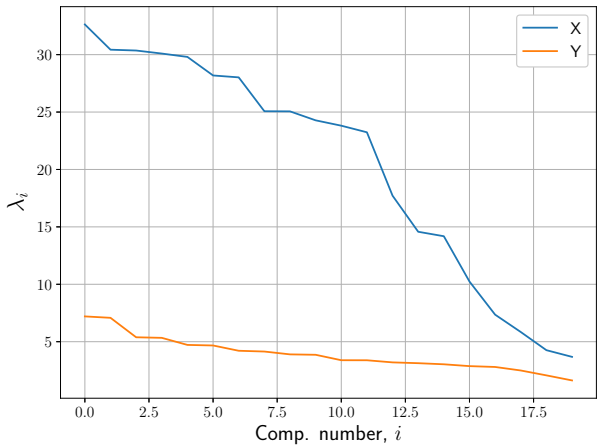

Рис. 19: Сингулярные числа матриц $\mathbf{H_x}$ и $\mathbf{H_y}$

Из графика 20 видно, что нет резкого скачка значений сингулярных чисел. Это означает, что скорее всего для рядов $\mathbf{x}$ и $\mathbf{y}$ не существует траекторных подпространств, проекции в которые точно аппроксимируют эти ряды.

Рассмотрим фазовые траектории рядов $\mathbf{x}$ и $\mathbf{y}$ в фазовые подпространства, натянутые на первые две главные компоненты. Возьмем $t^* = 10$ и изобразим ближайших соседей $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ вектора $\mathbf{x}_{t^*}$, а также соответствующие вектора $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$.

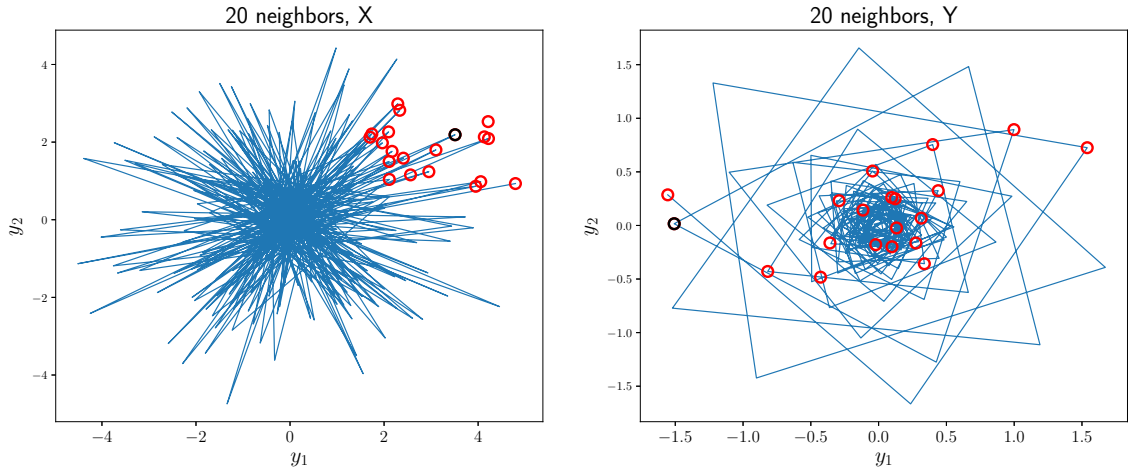

Рис. 20: Ближайшие соседи векторов $\mathbf{x}_{t^*}$ и $\mathbf{y}_{t^*}$, определенные по ряду $\mathbf{x}$

Список литературы

[1] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[2] Adam B Barrett, Lionel Barnett, and Anil K Seth. Multivariate granger causality and generalized variance. *Physical Review E*, 81(4):041907, 2010.

[3] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, page 1227079, 2012.

[4] George Sugihara and Robert M May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268):734, 1990.

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

[6] Robert Hoffmann, Chew-Ging Lee, Bala Ramasamy, and Matthew Yeung. Fdi and pollution: a granger causality test using panel data. *Journal of international development*, 17(3):311–317, 2005.

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

[8] AM Katrutsa and VV Strijov. Stress test procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems*, 142:172–183, 2015.

[9] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2017.

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

[11] Agnar Höskuldsson. Pls regression methods. *Journal of chemometrics*, 2(3):211–228, 1988.

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

[13] Nina Golyandina, Vladimir Nekrutkin, and Anatoly A Zhigljavsky. Analysis of time series structure: Ssa and related techniques (chapman & hall crc monographs on statistics & applied probability). 2001.

[14] Nina Golyandina and Anatoly Zhigljavsky. *Singular Spectrum Analysis for time series*. Springer Science & Business Media, 2013.

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

[16] Theodore Alexandrov. A method of trend extraction using singular spectrum analysis. *arXiv preprint arXiv:0804.3367*, 2008.

[17] Myles R Allen and Leonard A Smith. Monte carlo ssa: Detecting irregular oscillations in the presence of colored noise. *Journal of climate*, 9(12):3373–3404, 1996.

[18] Hossein Hassani, Saeed Heravi, and Anatoly Zhigljavsky. Forecasting uk industrial production with multivariate singular spectrum analysis. *Journal of Forecasting*, 32(5):395–408, 2013.

[19] CAF Marques, JA Ferreira, A Rocha, JM Castanheira, P Melo-Gonçalves, N Vaz, and JM Dias. Singular spectrum analysis and forecasting of hydrological time series. *Physics and Chemistry of the Earth, Parts A/B/C*, 31(18):1172–1179, 2006.