Модели обнаружения зависимостей во временных рядах в задачах построения прогностических моделей

Карина Равилевна Усманова

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Научный руководитель: д.ф.-м.н. В.В.Стрижов

Выпускная квалификационная работа бакалавра

Москва 2018

Цель исследования

Цель

Установить связи между заданными временными рядами с помощью метода сходящегося перекрестного отображения (CCM).

Мотивация

Повысить качество прогноза временного ряда путем использования истории временных рядов, коррелирующих с ним.

Литература

- Sugihara G., May R., Ye H., Hsieh C., Deyle E., Fogarty M., and Munch S // Detecting causality in complex ecosystems. 2012.
- Sugihara G., May R. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series // Nature. 1990.
- Golyandina N. and Stepanov D. SSA-based approaches to analysis and forecast of multidimensional time series // 5th St. Petersburg workshop on simulation. 2005.

Траекторное пространство

Задан временной ряд $\mathbf{x} = [x_1, \ldots, x_N]$. По нему строится траекторная матрица $\mathbf{H_x}$:

$$\mathbf{H_x} = \begin{pmatrix} x_1 & \cdots & x_{L-1} & x_L \\ x_2 & \cdots & x_L & x_{L+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N-L+1} & \cdots & x_{N-1} & x_N \end{pmatrix} = \begin{pmatrix} \mathbf{x}_L \\ \mathbf{x}_{L+1} \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$

где L – ширина окна.

$$\mathbf{x}_i = (x_{i-L+1}, \ldots, x_{i-1}, x_i) \in \mathbf{M_x}, \quad i = L, \ldots, N,$$

где $\mathbf{M_x}$ – траекторное пространство ряда $\mathbf{x}$.

CCM, постановка задачи

Задача

Для заданных временных рядов $\mathbf{x} = [x_1, \ldots, x_N]$ и $\mathbf{y} = [y_1, \ldots, y_N]$ установить наличие или отсутствие связи между ними.

Решение

Считаем, что ряд $\mathbf{y}$ зависит от ряда $\mathbf{x}$, если существует липшицево отображение $\varphi : \mathbf{M_x} \to \mathbf{M_x}$

$$\rho_{\mathbf{M_y}}(\varphi(\mathbf{x}_i, \varphi(\mathbf{x}_j)) \leq L \cdot \rho_{\mathbf{M_x}}(\mathbf{x}_i, \mathbf{x}_j), \quad \forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{M_x}$$

CCM, постановка задачи

- Выбираем $\mathbf{x}_{t^*} = (x_{t^*-L+1}, \ldots, x_{t^*-1}, x_{t^*})$.
- Пусть $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}$ – k ближайших соседей вектора $\mathbf{x}_{t^*}$ в пространстве $\mathbf{M}_\mathbf{x}$. Тогда $\mathbf{y}_{t^*}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_k}$ – строки матрицы $\mathbf{H}_\mathbf{y}$, соответствующие индексам $t_1, \ldots, t_k$.

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

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

Построение проекции

- Сингулярное разложение траекторной матрицы:

$$\mathbf{H_x} = \mathbf{U_x \Lambda_x V_x}$$

- Выберем $\mathcal{T}_\mathbf{x}$ – некоторый набор индексов компонент ряда $\mathbf{x}$
- $\mathbf{M}_{\mathcal{T}_\mathbf{x}} \subset \mathbf{M_x}$– траекторное подпространство
- Проекция ряда $\mathbf{x}$ в подпространство $\mathbf{M}_{\mathcal{T}_\mathbf{x}}$, описывается траекторной матрицей

$$\mathbf{P}_{\mathcal{T}_\mathbf{x}} = \mathbf{U_x \tilde{\Lambda}_x V_x}$$

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

CCM на подпространствах

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

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

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

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

Эксперимент, часть 1

$$\mathbf{x} = \sin t + 2\sin\frac{t}{2} + \sigma_x^2\varepsilon, \quad \sigma_x^2 = 0.3, \quad \varepsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathbf{y} = \sin(2t + 5) + \sigma_y^2\varepsilon, \quad \sigma_y^2 = 0.25, \quad \varepsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Эксперимент, часть 1

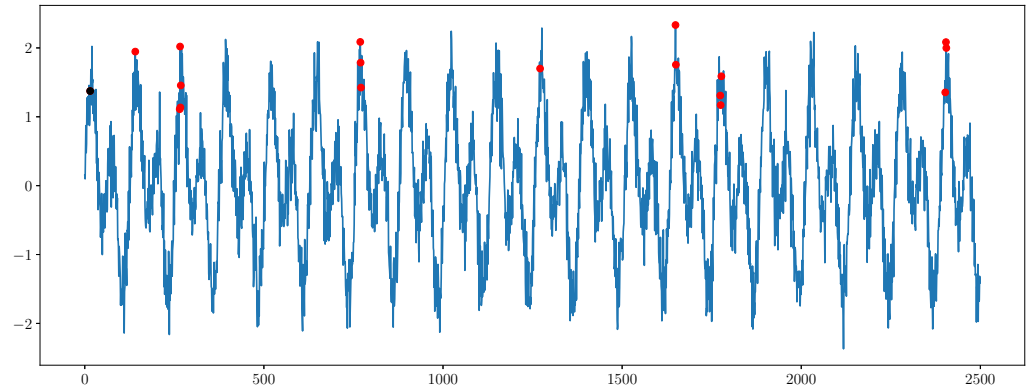

Рис.: Ближайшие соседи точки $\mathbf{x}_{15}$

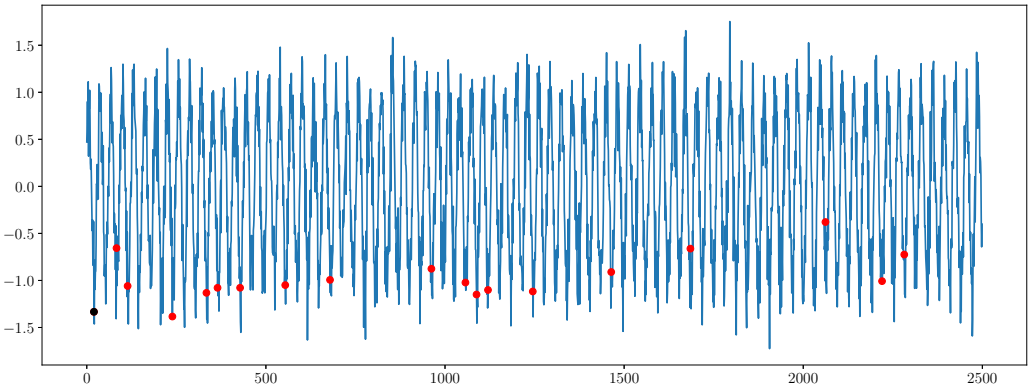

Рис.: Ближайшие соседи точки $\mathbf{y}_{15}$

Ближайшие соседи на фазовых диаграммах

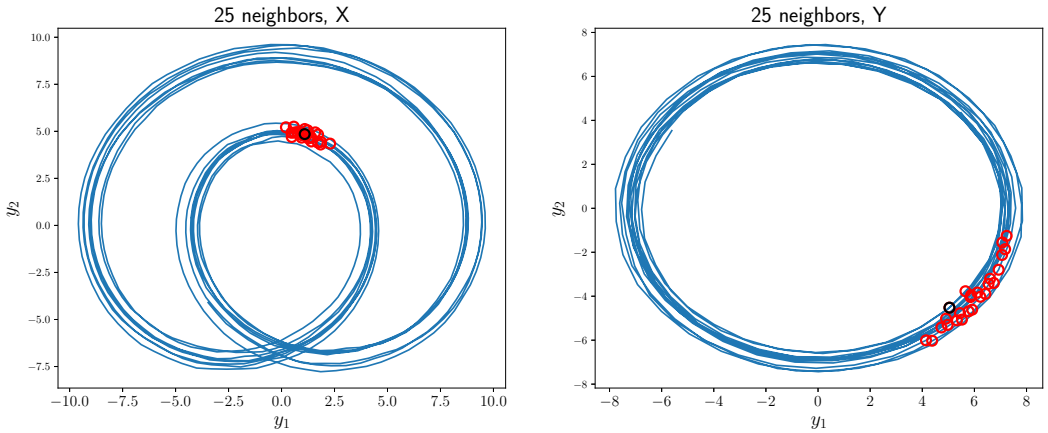

Рис.: Определение ближайших соседей по ряду **x**

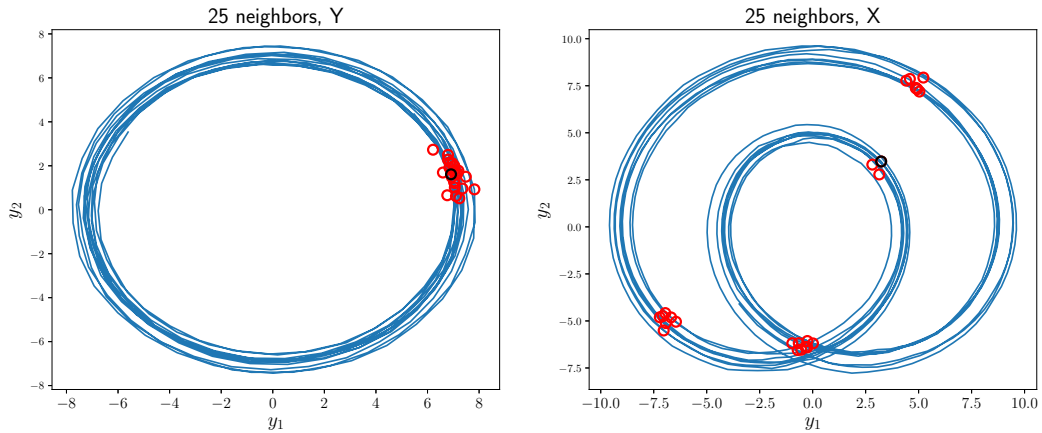

Рис.: Определение ближайших соседей по ряду **y**

Эксперимент, часть 2

Эксперимент проводился на данных потребления электроэнергии и температуры в течение года.

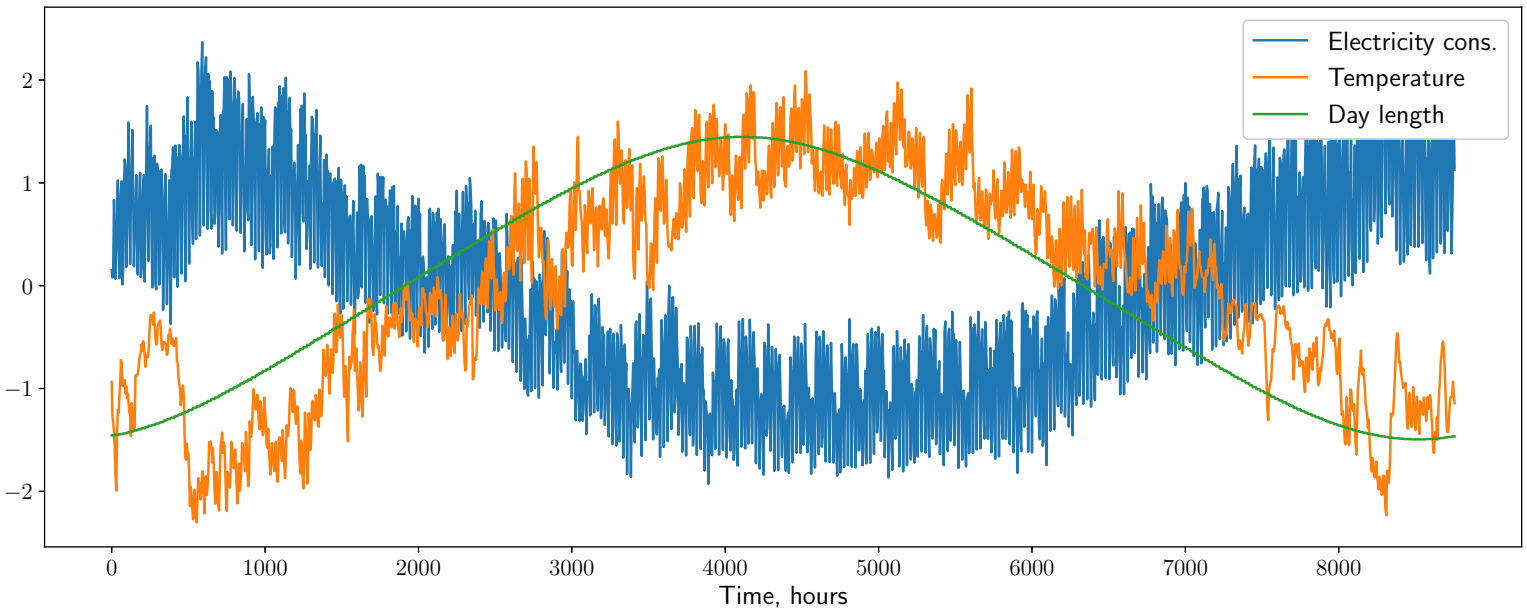

Рис.: Нормированные ряды потребления электроэнергии, температуры и длины светового дня

Эксперимент, часть 2

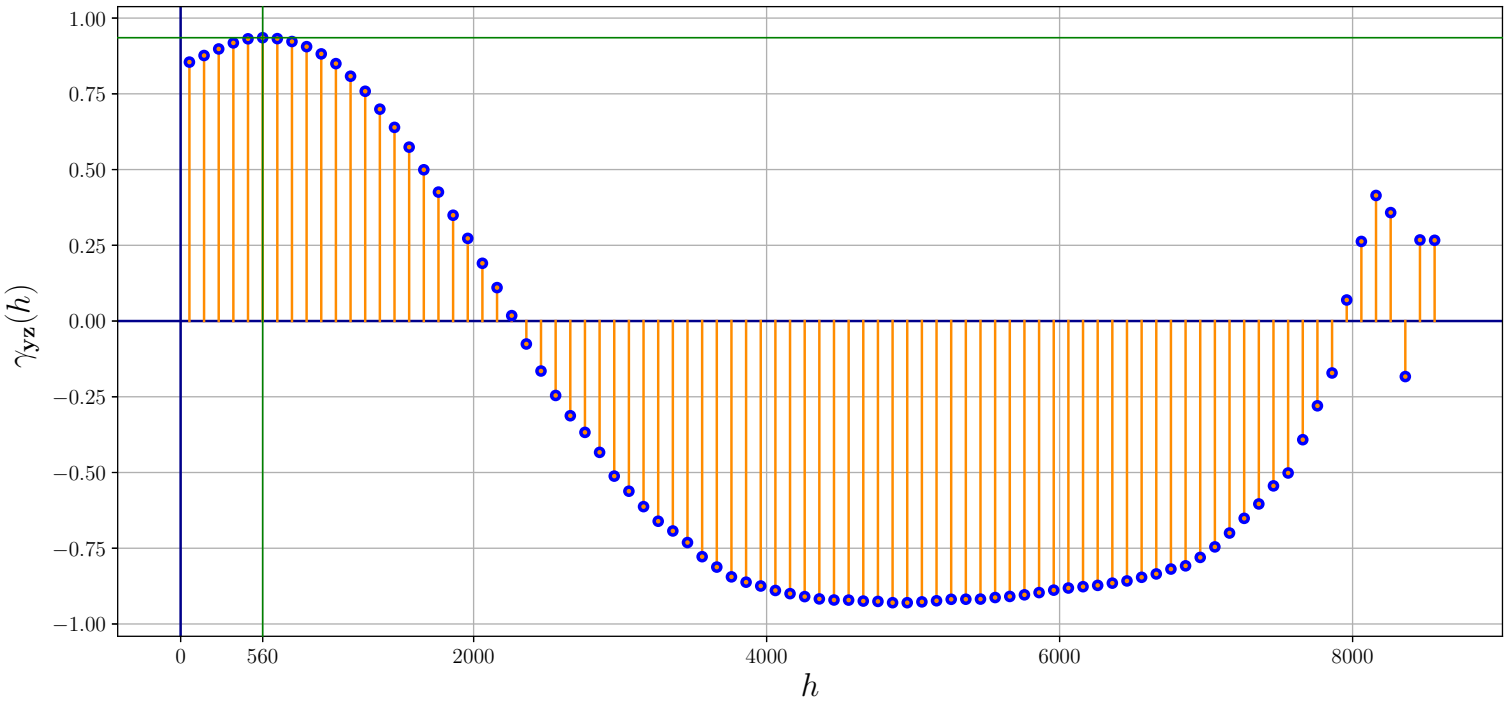

Рис.: Кросс-корреляционная диаграмма для ряда температуры и длины светового дня

Эксперимент, часть 2

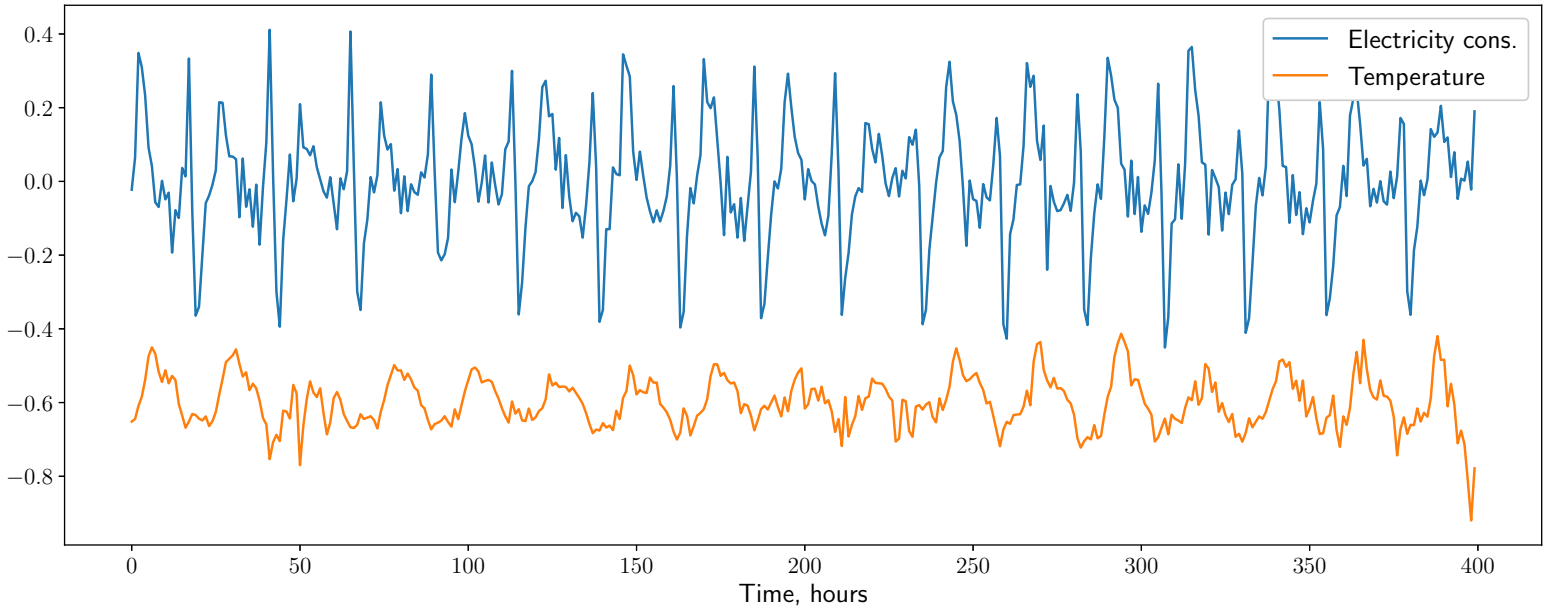

Рис.: Продифференцированные и нормированные ряды потребления электроэнергии и температуры

Ближайшие соседи на фазовых траекториях

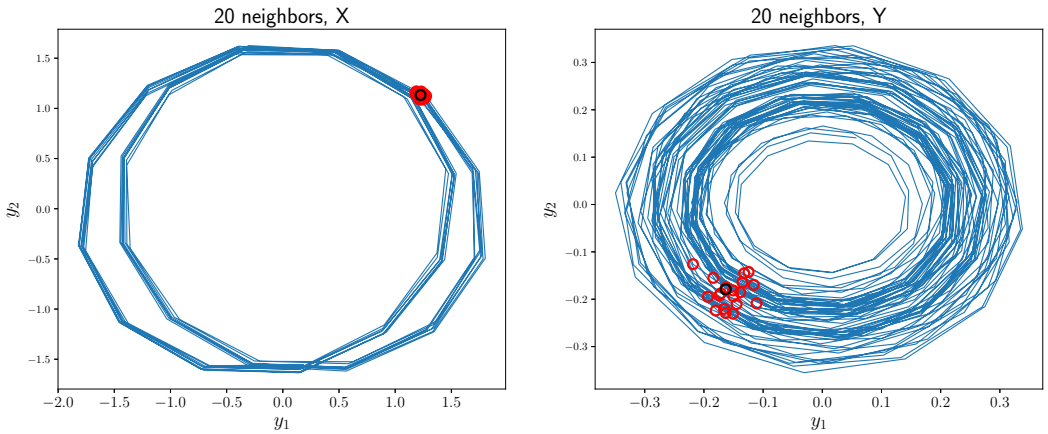

Рис.: Определение ближайших соседей по ряду **x**

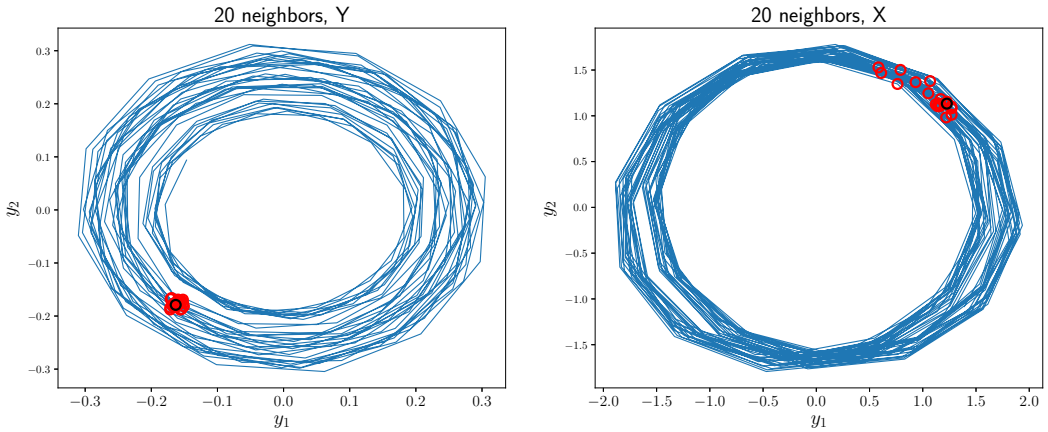

Рис.: Определение ближайших соседей по ряду **y**

Выбор подпространств

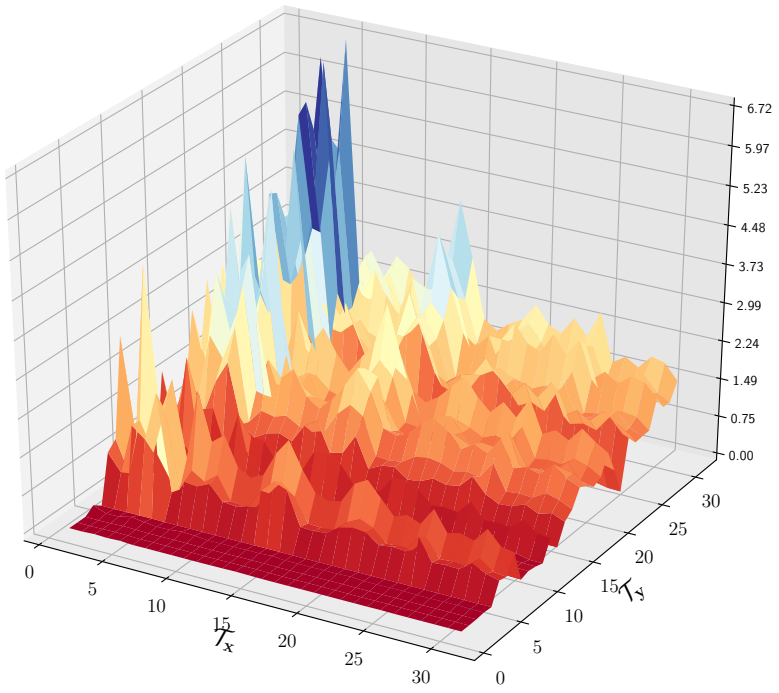

Рис.: Значения функционала $S(\mathbf{x}, \mathbf{y}, \mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y})$ для различных наборов компонент $(\mathcal{T}_\mathbf{x}, \mathcal{T}_\mathbf{y})$

Результаты

- С помощью метода CCM исследованы связи между рядами потребления электроэнергии и температуры.
- Исследована связь между проекциями этих рядов в различные подпространства.

К. Р. Усманова, С. П. Кудияров, Р. В. Мартышкин, А. А. Замковой, В. В. Стрижов // Анализ зависимостей между показателями при прогнозировании объема грузоперевозок // Системы и средства информатики, 2018.