Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана»
(национальный исследовательский университет)
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Фундаментальные науки

КАФЕДРА Вычислительная математика и математическая физика (ФН-11)

К У Р С О В А Я Р А Б О Т А

на тему:

Применение рекуррентных нейронных сетей
для прогнозирования временных рядов.

Дисциплина:

Численные методы

Студент группы ФН11-62Б _____ **Очкин Н.В.**
(Подпись, дата) (И.О. Фамилия)

Руководитель курсовой работы _____ **Зубарев К.М.**
(Подпись, дата) (И.О. Фамилия)

Оценка: _____

Москва, 2024

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель НИР	_____ Подпись, дата	Зубарев К.М.
Исполнители: Студент группы ФН11-62Б	_____ Подпись, дата	Очкин Н.В.
Нормоконтролёр ст. преп. каф. ФН11	_____ Подпись, дата	Прозоровский А.А.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ **4**

1 Прогнозирование временных рядов **5**

1.1 Временные ряды **5**

1.1.1 Прогнозирование временного ряда [1] **6**

1.1.2 Главная особенность временных рядов **7**

1.1.3 Компоненты временных рядов **7**

1.2 Автокорреляция **8**

1.2.1 Автокорреляция, её вычисление **9**

1.2.2 Коррелограммы **10**

1.2.2.1 График задержек **10**

1.2.2.2 correlation matrix **11**

1.2.2.3 AutoCorrelation Function (ACF) **11**

1.2.2.4 Значимость автокорреляции **11**

2 Практическая часть **12**

ЗАКЛЮЧЕНИЕ **13**

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ **14**

ПРИЛОЖЕНИЕ А **15**

ПРИЛОЖЕНИЕ Б **16**

ПРИЛОЖЕНИЕ В **17**

ПРИЛОЖЕНИЕ Г **20**

ВВЕДЕНИЕ

Курсовая работа представляет собой вид учебной работы, рассчитанный на получение и закрепление необходимых навыков, развитие умения работать со сбором и анализом материала, полученного из изученных источников, а также написанию текста работы с соблюдением стандарта оформления (в данной работе используется ГОСТ 7.32).

1 Прогнозирование временных рядов

1.1 Временные ряды

Временным рядом называется последовательность значений признака y, измеряемого через постоянные временные интервалы:

$$y_1, \ldots, y_T, \ldots, \quad y_t \in \mathbb{R}.$$

В этом определении нужно обратить внимание на то, что временные интервалы между измерениями признака постоянны.

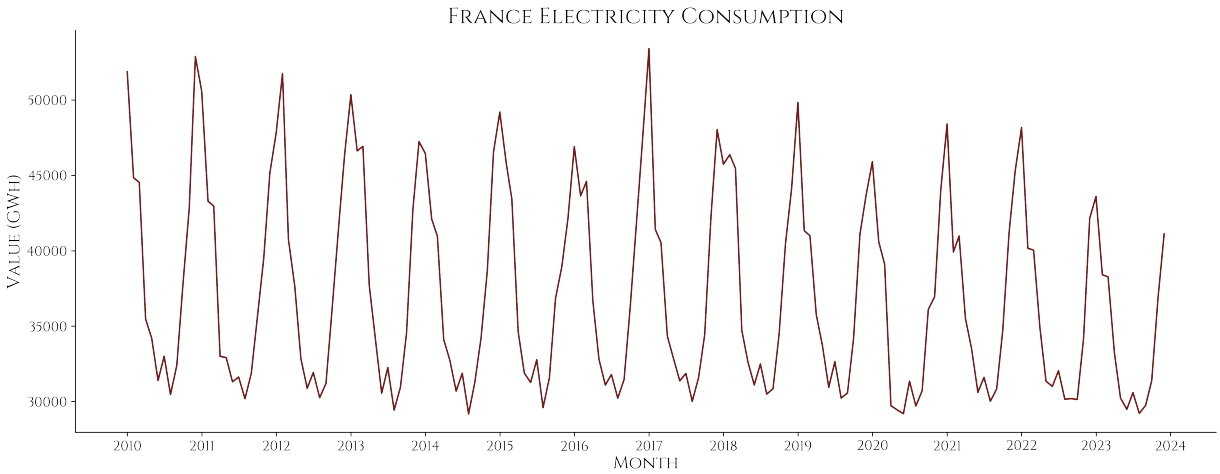

Рис. 1.1: Ежемесячное потребление электричества во Франции. Построено программой по адресу (листинг 1)

Временные ряды изобилуют в таких областях, как экономика, бизнес, инженерия, естественные науки (особенно геофизика и метеорология), а также социальные науки.

Примеры временных рядов - это ряды с ежемесячной последовательностью объемов отгруженных с завода товаров, еженедельными данными о количестве дорожно-транспортных происшествий, ежедневными объемами осадков, почасовыми наблюдениями за химическеми выбросами. Ещё один пример временного ряда (показан на рисунке 1.1) — это реальное ежемесячное потребление электричества во Франции.

1.1.1 Прогнозирование временного ряда [1]

Информация о доступных наблюдениях в момент времени t временного ряда, используемая для предсказания его значения в некотором будущем $t+l$ может предоставить фундамент для планирования в бизнесе и экономике, контроля продукции, оптимизации индустриальных процессов и так далее. Здесь l - это, так называемый, *горизонт прогнозирования*, который меняется от задачи к задаче.

Предположим, что наблюдения - *дискретные*, равноудаленные друг от друга величины. Например в задаче прогнозирования продаж, продажи z_t в текущий месяц t и продажи $z_{t-1}, z_{t-2}, z_{t-3}, ...$ в предыдущие месяцы можно использовать для прогнозирования горизонта в $l = 1, 2, 3, ..., 12$ месяцев. Обозначим за $\hat{z}_t(l)$ прогноз, составленный в *момент времени* t для продаж z_{t+l} в некотором будущем $t+l$, т.е. с *горизонтом прогнозирования* l. Функция $\hat{z}_t(l)$, прогнозирующая в момент времени t для всех будущих горизнтов прогнозирования, на основе доступной информации о текущем и предыдущих значениях $z_t, z_{t-1}, z_{t-2}, z_{t-3}, ...$ на протяжении времени t, называется *функцией прогнозирования* в момент времени t. Основная задача прогнозирования временных рядов заключается в том, чтобы найти функцию прогнозирования с наименьшим средним квадратом отклонений $z_{t+l} - \hat{z}_t(l)$ между истинными и спрогнозированными значениями для каждого из *горизонтов прогнозирования* l.

В добавок к нахождению наилучших прогнозов, также необходимо указать их точность, так чтобы, например, можно было рассчитать риски, ассоциированные с решениями, принятыми на их основе. Точность прогнозов можно выразить в качестве *доверительных интервалов* с каждой стороны прогноза. Эти интервалы могут быть рассчитаны для любого удобного набора вероятностей, например, 50 и 95%. Они указывают вероятность попадания спрогнозированной величины в данный интервал. За иллюстрацией обратимся к рис. 1.2, на котором изображены последние 20 значений временного ряда, кульминирующие в момент времени t. Также представлены прогнозы, сделанные в момент времени t с горизонтами прогнозирования $l = 1, 2, ..., 13$, вместе с 50% доверетильными интервалами.

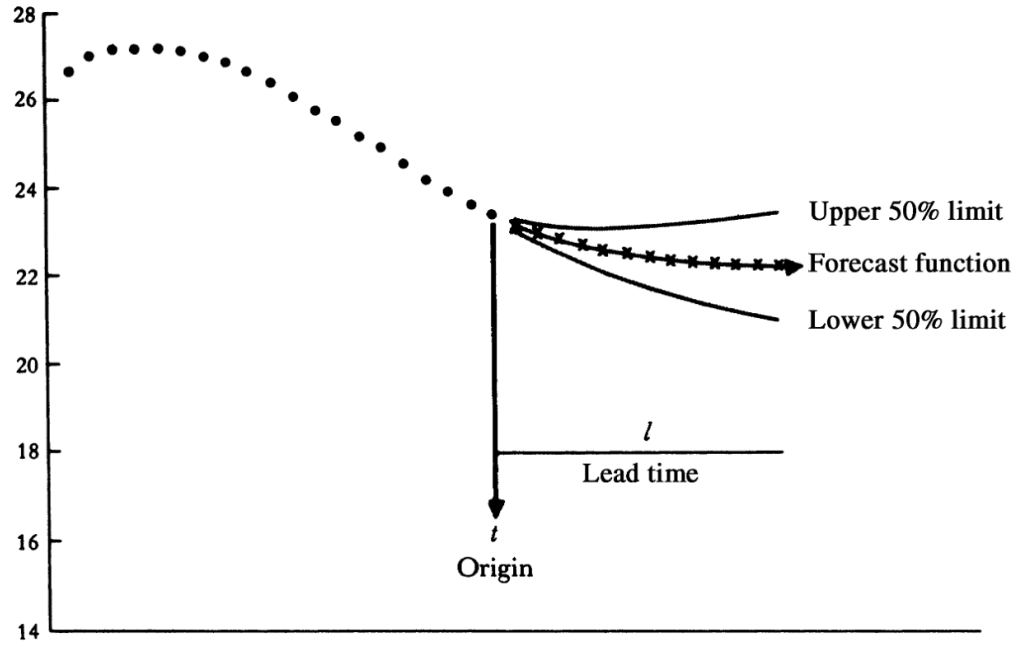

Рис. 1.2: Значения временного ряда с прогнозирующей функцией и 50% доверетильными интервалами.

1.1.2 Главная особенность временных рядов

Часто, в задачах регрессионного анализа и машинного обучения, в качестве анализируемых данных берутся простые выборки, построенные из независимо одинаково распредленных наблюдений. В задаче анализа временных рядов всё с точностью наоборот: предполагается, что данные в прошлом каким-то образом связаны с данными в будущем. Чем сильнее они связаны, тем больше имеется информации о поведении временного ряда в будущем и тем точнее можно сделать прогноз.

Полезно снова рассмотреть данные о реальном ежемесячном потреблении электричества во Франции (рис. 1.1). Видно, что на графике изображена не простая выборка (измерения не являются независимыми и одинаково распределёнными), а сложный, структурированный процесс. Выявив структуру этого процесса, можно учесть её в прогнозирующей модели и построить действительно точный прогноз.

1.1.3 Компоненты временных рядов

Полезно рассмотреть несколько понятий, которыми можно описать поведение временных рядов:

- Тренд — плавное долгосрочное изменение уровня ряда. Эту характе-

ристику можно получить, наблюдая ряд в течение достаточно долгого времени.

- Сезонность — циклические изменения уровня ряда с постоянным периодом. В данных о ежемесячном потреблении электричества во Франции (рис. 1.1) очень хорошо видны подобные сезонные колебания: признак всегда принимает максимальное значение зимой, а минимальное — летом. Это легко объяснить тем, что летом электричества необходимо меньше всего, это самый тёплый сезон во Франции. В целом профиль изменения потребления электричества внутри года остаётся более-менее постоянным.

- Цикл — изменение уровня ряда с переменным периодом. Такое поведение часто встречается в рядах, связанных с продажами, и объясняется циклическими изменениями экономической активности. В эко- номике выделяют циклы длиной 4 - 5 лет, 7 - 11 лет, 45 - 50 лет и т. д. Другой пример ряда с такой характеристикой — это солнечная активность, которая соответствует, например, количеству солнечных пятен за день. Она плавно меняется с периодом, который составляет несколько лет, причём сам период также меняется во времени.

- Ошибка — непрогнозируемая случайная компонента ряда. Сюда включены все те характеристики временного ряда, которые сложно измерить (например, слишком слабые).

1.2 Автокорреляция

Одной из важнейших характеристик временного ряда является автокорреляция. Автокорреляцией назвается математическая репрезентация степени «схожести» между исходным рядом и его версией, сдвинутой на некоторый интервал, называемый лагом. Концептуально это похоже на корреляцию Пирсона между двумя временными рядами, только автокорреляция рассматривает один и тот же ряд дважды.

Стоит напомнить, что коэффициент корреляции, а значит и автокорреляции может быть равен 0 при наличии сильной, но нелинейной

зависимости (рис. 1.3) [2].

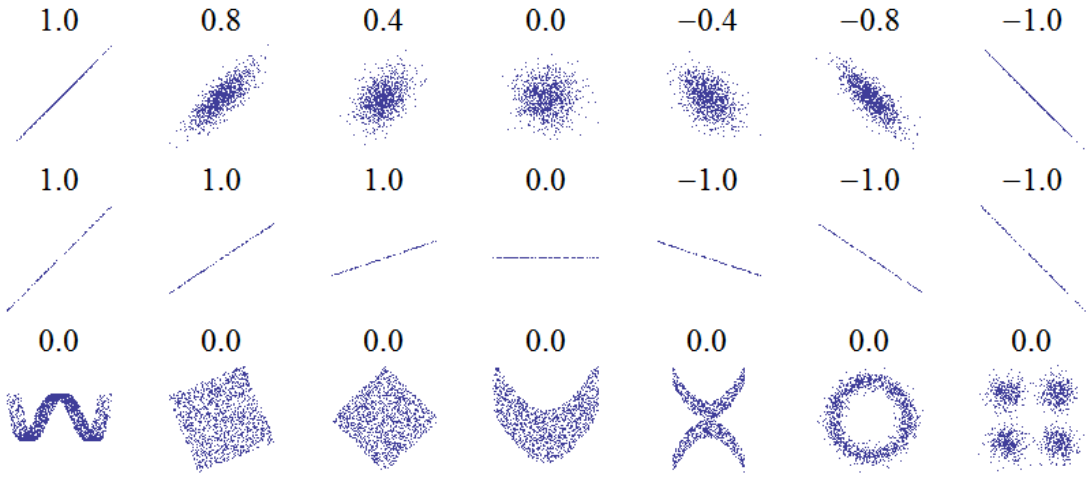

Рис. 1.3: Несколько наборов точек (x, y) и коэффициент корреляции x и y для каждого набора. Отметим, что корреляция отражает зашумленность и направление линейной зависимости (верхний ряд), но не отражает ее угловой коэффициент (средний ряд), а также многие аспекты нелинейных связей (нижний ряд). (Примечание: на рисунке в центре угловой коэффициент равен 0, но в этом случае коэффициент корреляции не определен, потому что дисперсия Y нулевая.)

1.2.1 Автокорреляция, её вычисление

Количественной характеристикой сходства между значениями ряда в соседних точках является автокорреляционная функция (или просто автокорреляция), которая задаётся следующим соотношением:

$$r_\tau \stackrel{\text{def}}{=} \frac{\mathbb{E}[(y_t - \mathbb{E}_y)(y_{t+\tau} - \mathbb{E}_y)]}{\mathbb{V}_y},$$

где количество отсчётов, на которое сдвинут ряд, называется лагом автокорреляции (τ).

Можно заметить сходство с корреляцией Пирсона:

$$\rho \stackrel{\text{def}}{=} \text{corr}[X, Y] \stackrel{\text{def}}{=} \frac{\text{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}, \qquad \text{Cov}[X, Y] \stackrel{\text{def}}{=} \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Значения, принимаемые автокорреляцией такие же, как и у коэффициента Пирсона: $r_\tau \in [-1, 1]$. Вычислить автокорреляцию по выборке можно,

заменив в формуле математическое ожидание на выборочное среднее, а дисперсию — на выборочную дисперсию.

1.2.2 Коррелограммы

Анализировать величину автокорреляции при разных значениях лагов удобно с помощью графиков. Далее различные методы графического анализа автокорреляции будут демонстрироваться на примере данных о суммарном объёме продаж вина в Австралии за месяц на протяжении 15 лет (рис. 1.4).

Рис. 1.4: Месячный объём продаж вина в Австралии, в бутылках.

Этот ряд обладает ярко выраженной годовой сезонностью: максимум продаж за год приходится на де- кабрь, а затем, в январе, происходит существенное падение.

1.2.2.1 График задержек

Графиком задержек (Lag Plot) называется особый вид диаграммы рассеяния (scatter plot), на которой по одной из ее осей откладывается значение временного ряда в момент времени t, а по другой - значение в момент времени $t + l$, где l - значение лага.

** WINE LAG PLOT (with explanation) **

График задержек помогает понять являются ли значения во временном ряду случайными. Если данные случайны, то на графике мы не увидим никакой определенной структуры. Однако если же данные не случайны, то на графике задержек можно будет увидеть ярко выраженную форму.

1.2.2.2 correlation matrix

1.2.2.3 AutoCorrelation Function (ACF)

Such a plot is also called a correlogram.

1.2.2.4 Значимость автокорреляции

2 Практическая часть

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Gregory C. Reinsel George E. P. Box, Gwilym M. Jenkins and Greta M. Ljung. Time Series Analysis: Forecasting and Control. Wiley, fifth edition edition, 2015.

[2] Kevin P. Murphy. Probabilistic Machine Learning: An introduction. MIT Press, 2022.

ПРИЛОЖЕНИЕ А

ПРИЛОЖЕНИЕ Б

ПРИЛОЖЕНИЕ В

Программный код, используемый в данной работе.

```
import pandas as pd

import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import matplotlib.font_manager as fm

font_path = 'extra/Cinzel-VariableFont_wght.ttf'

cinzel_font = fm.FontProperties(fname=font_path)
fm.fontManager.addfont(font_path)

RED = '#6F1D1B'

def decorate_plot(ax, xname, yname, loc):
    SIZE_TICKS = 12

    # Eliminate upper and right axes
    ax.spines['right'].set_color('none')
    ax.spines['top'].set_color('none')

    # Show ticks in the left and lower axes only
    ax.xaxis.set_ticks_position('bottom')
    ax.yaxis.set_ticks_position('left')

    # x axis name
    ax.set_xlabel(xname, fontsize=15)

    # y axis name
    ax.set_ylabel(yname, fontsize=15)

    # Adjust the font size of the tick labels
    ax.tick_params(axis='both', which='major', labelsize=SIZE_TICKS)

    if loc:
        plt.legend(fontsize=10, loc=loc)
```

```
    # Update font settings
    plt.rcParams.update({
        "font.family": cinzel_font.get_name(),
        "font.size": 16
    })

    # Adjust layout
    plt.tight_layout()
```

Листинг 1: Ежемесячное потребление электричества во Франции

```
df = pd.read_csv('data/global_electricity_production_data.csv')

France_df = df[(df['country_name'] == 'France') &
               (df['product'] == 'Electricity') &
               (df['parameter'] == 'Final Consumption (Calculated)')].copy()

start_date = pd.to_datetime('01/01/2010', format='%m/%d/%Y')
end_date   = pd.to_datetime('12/01/2023', format='%m/%d/%Y')

France_df['date'] = pd.to_datetime(df['date'], format='%m/%d/%Y')

France_df = France_df[(France_df['date'] >= start_date) &
                      (France_df['date'] <= end_date)]

France_df.sort_values('date', inplace=True)

# Create the plot
fig, ax = plt.subplots(figsize=(15, 6))
ax.plot(France_df['date'], France_df['value'], linestyle='-', color=RED)

ax.set_title('France Electricity Consumption')

decorate_plot(ax, 'Month', 'Value (GWh)', '')

year_locator = mdates.YearLocator()
year_formatter = mdates.DateFormatter('%Y')
ax.xaxis.set_major_locator(year_locator)
ax.xaxis.set_major_formatter(year_formatter)

plt.savefig(f'images/time_series_example_France.png',
```

```
            dpi=300,
            transparent=True)

plt.show()
```

ПРИЛОЖЕНИЕ Г