Модуль 1 · Python и анализ данных

Анализ данных метеорологических станций

Ph.D., старший преподаватель Мадина Турсумбаева

Кафедра метеорологии и гидрологии

Курс: Современные цифровые технологии в метеорологии

Цель лекции

Показать, как превратить сырой файл архива метеорологической станции в проверенный ряд и получить по нему базовый график, среднемесячные значения и суточный ход температуры.

Станционный ряд — основа климатической оценки. Лекция показывает, как превратить сырой файл архива в проверенный ряд.

Результаты обучения

После лекции студент будет:

Знать

Структура архива станционных данных: служебные строки, разделитель, кодировка

Местное время и UTC

Коды пропусков

Уметь

Загрузить CSV-файл станции и найти столбцы времени и температуры

Построить базовый график, среднемесячные и средние суточные значения

Понимать

Почему полнота наблюдений важна не меньше самого значения

Почему нельзя смешивать часы разных поясов

Почему сезонный анализ начинается с качества данных

Станционные данные: что мы получаем

Срочные наблюдения

Метеостанция фиксирует состояние атмосферы в установленные сроки, например каждые 3 часа. Правила описаны в справочнике ВМО (WMO-No. 8).

Архив RP5

Выгрузка CSV по станции за выбранный период: строка — срок наблюдения, столбцы — T, давление, влажность, ветер, явления погоды.

Качество

Пропуски, ошибки прибора, смена приборов или места станции. Ряд нужно проверить до расчётов.

- Одна строка — один срок наблюдения, а не один день.
- Время в файле может быть местным: уточните часовой пояс при выгрузке.

Структура файла архива

1

Служебные строки

Название станции, координаты, период выгрузки

2

Строка заголовка

Названия столбцов: время, температура и др.

3

Данные

По одной строке на срок наблюдения

```
# подберите параметры под свой файл
df = pd.read_csv("almaty_2024.csv", sep=";", skiprows=6, encoding="cp1251")
# skiprows — число служебных строк; cp1251 или utf-8 — если кириллица искажена
```

Разделитель, кодировка и число служебных строк проверяются по самому файлу: откройте его в текстовом редакторе.

Шаг 1–2: загрузка и определение столбцов

```
import pandas as pd
df = pd.read_csv("almaty_2024.csv", sep=";", skiprows=6, encoding="cp1251")
print(df.columns.tolist())        # названия столбцов
time_col = df.columns[0]         # проверьте: столбец времени
df[time_col] = pd.to_datetime(df[time_col], dayfirst=True)
df = df.set_index(time_col)
# температура: запятая → точка, нечисла → NaN
df["T"] = pd.to_numeric(df["T"].str.replace(",", "."), errors="coerce")
```

- Имя столбца температуры в архиве может отличаться (например, T): сверяйтесь с print().
- errors="coerce" превращает нечисловые значения в NaN, а не в ошибку — но их надо проверить.

Шаг 3: базовый график

```
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 4))
plt.plot(df.index, df["T"], lw=0.8, color="tab:red")
plt.title("Температура воздуха, Алматы, 2024 г. (срочные данные)")
plt.xlabel("Дата"); plt.ylabel("T, °C"); plt.grid(alpha=0.3)
plt.show()
```

Что на графике

Срочные значения дают много шума.

Заголовок должны содержать станцию, период и срочность.

- Подписи осей и единицы — обязательны.
- Если линия выглядит «пилообразной», это суточный ход: его разберём в шаге 5.

Шаг 4: среднемесячные значения

```
# среднее за календарный месяц
m = df["T"].resample("MS").mean()
# сколько наблюдений вошло в каждый месяц
n = df["T"].resample("MS").count()
# оставить месяцы, где есть не менее 80 % ожидаемых наблюдений
m_ok = m[n >= 0.8 * n.max()]
m_ok.plot(marker="o"); plt.ylabel("T, °C"); plt.show()
```

resample

Группирует ряд по календарному интервалу: "MS" — начало месяца, "1D" — сутки.

Месяц с 30 % наблюдений нельзя сравнивать с полным месяцем: среднее искажено выборкой.

Шаг 5: средние значения по часам (суточный ход)

```
# средняя температура по часу суток (местное время)
hourly = df["T"].groupby(df.index.hour).mean()
hourly.plot(marker="o")
plt.xlabel("Час суток"); plt.ylabel("T, °C"); plt.grid(alpha=0.3)
plt.title("Суточный ход температуры")
```

Как читать

Минимум — обычно у восхода Солнца.

Максимум — после полудня, в середине дня.

Важно

Суточный ход — это сигнал, а не шум: он показывает, как работает нагрев и выхолаживание поверхности. Сравнивайте часы только в одном часовом поясе.

Самостоятельная работа: сезонный анализ

- Постройте три вида графиков для одного и того же ряда и объясните, что показывает каждый.

1. Линия по часам

Суточный ход: средние значения по часам суток.

2. Столбцы по месяцам

Годовой ход: среднее за каждый месяц.

3. Ящики по сезонам

Разброс: boxplot по сезонам (см. лекцию 5).

```
import seaborn as sns
sns.boxplot(data=season_df, x="season", y="T")    # season — столбец сезона
```

Контроль качества станционных данных

1

Диапазон

Значения вне физического диапазона для региона — в NaN после проверки

2

Повторы

duplicated() по индексу времени: один срок — одна строка

3

Разрывы

Пропущенные сроки: сравните индекс с полной сеткой времени

4

Смена прибора

Сдвиги среднего в метаданных станции — отдельная проверка

```
bad = (df["T"] < lo) | (df["T"] > hi)    # lo, hi — границы для региона
df.loc[bad, "T"] = np.nan                  # помечаем, не удаляем
full = pd.date_range(df.index.min(), df.index.max(), freq="3h")
missing = full.difference(df.index)        # срочные сроки без данных
```

Границы lo и hi выбирайте по справочнику климата региона, а не по собственной интуиции.

Типичные ошибки и их исправление

Ошибка	Что происходит	Как исправить
Неверный разделитель	Все данные в одном столбце	Указать sep=";" или "\t"
Запятая как десятичный знак	Температура читается как текст	replace(",", ".") и to_numeric
Часы разных поясов	Суточный ход «сдвинут»	Уточнить пояс; работать в одном
Месяц с малой полнотой	Среднее смещено выборкой	Порог полноты, как в шаге 4
Пропуски заполнены нулём	Среднее уходит вниз	NaN остаётся NaN; заполнение — с пометкой

Заключение

1

Прочитать

служебные строки, sep, encoding

2

Определить

столбцы времени и температуры

3

Проверить

диапазоны, повторы, пропуски

4

Агрегировать

месяцы, часы, сутки

5

Представить

подписи, единицы, период

Главный вывод

Сырой архив станции становится пригодным рядом только после проверки: единый часовой пояс, коды пропусков, полнота наблюдений. Среднемесячные значения и суточный ход имеют смысл лишь для проверенного ряда.

Вопросы для самопроверки

1. Почему станционный ряд нужно читать в одном часовом поясе?
2. Чем resample отличается от groupby по часу суток?
3. Почему месяц с малой полнотой наблюдений нельзя сравнивать с полным?
4. Почему NaN нельзя заменять нулём при вычислении среднего?

Семинар 4 и следующая лекция

Семинар 4 · Визуализация станционных наблюдений

Лабораторная работа «Анализ температуры воздуха (Алматы, 2024)» по файлам RP5.

Шаги: загрузка, столбцы, базовый график, среднемесячные значения, суточный ход, самостоятельный сезонный анализ.

Лекция 5

Временные ряды: сезонность, нормы и аномалии.

Тот же ряд станции — уже в виде аномалий.

Литература

- WMO. Guide to Instruments and Methods of Observation (WMO-No. 8). World Meteorological Organization.
- RP5 — архив фактической погоды: https://rp5.ru
- McKinney, W. (2022). Python for Data Analysis (3rd ed.). O’Reilly Media. https://wesmckinney.com/book/
- pandas development team. Time series / date functionality. https://pandas.pydata.org/docs/
- Wilks, D. S. (2019). Statistical Methods in the Atmospheric Sciences (4th ed.). Elsevier.