

---

# **HIS PROJECT - TSA**

---

## **Task 03**

**Author**

Alberto Mejia Diez

# Contents

<b>1</b>	<b>Chapter 01 Summary</b>	<b>3</b>
1.1	Introduction . . . . .	3
1.2	What is a Time Series? . . . . .	3
1.3	Types of Time Series . . . . .	3
1.4	Applications of Time Series Analysis . . . . .	3
1.5	Data-Generating Process (DGP) . . . . .	3
1.6	Synthetic Time Series Generation . . . . .	3
1.7	Stationary vs. Non-Stationary Time Series . . . . .	4
1.8	Forecastability of Time Series . . . . .	4
1.9	Forecasting Terminology . . . . .	4
<b>2</b>	<b>Predicatability</b>	<b>4</b>
<b>3</b>	<b>Forecasting</b>	<b>5</b>
<b>4</b>	<b>Chapter 01 Code Analysis</b>	<b>5</b>
<b>5</b>	<b>Chapter 02 Summary</b>	<b>5</b>
<b>6</b>	<b>Data Preprocessing</b>	<b>6</b>
<b>7</b>	<b>Missing Data</b>	<b>6</b>
<b>8</b>	<b>Chapter 02 Code Analysis</b>	<b>6</b>
<b>9</b>	<b>Chapter 03 Summary</b>	<b>7</b>
9.1	Components of a Time Series . . . . .	7
9.2	Visualizing Time Series Data . . . . .	7
9.3	Decomposing a Time Series . . . . .	7
9.3.1	Detrending . . . . .	7
9.3.2	Deseasonalizing . . . . .	8
9.4	Detecting and Treating Outliers . . . . .	8
<b>10</b>	<b>Decompositon</b>	<b>8</b>
<b>11</b>	<b>Outliers</b>	<b>8</b>

# **1 Chapter 01 Summary**

## **1.1 Introduction**

This book serves as an advanced guide for data scientists and ML engineers to deepen their skills in time series analysis, a crucial and often overlooked aspect of ML. The book aims to progress beyond traditional methods such as ARIMA to the latest ML techniques, addressing the growing complexity and volume of business-related time series data.

## **1.2 What is a Time Series?**

A time series is a sequence of observations recorded over time. Examples include the number of chocolate bars consumed over a month or monthly weight measurements. Time series can reveal relationships and trends, such as the correlation between chocolate consumption and weight changes, and can extend to other domains like stock prices and climate measurements.

## **1.3 Types of Time Series**

The book distinguishes between regular time series with fixed time intervals and irregular ones without such consistency. Regular time series examples include hourly or monthly data points, while irregular ones could be patient lab test readings that occur sporadically.

## **1.4 Applications of Time Series Analysis**

Three main applications are discussed: forecasting future values, classifying time series data (e.g., normal vs. abnormal EKG readings), and deriving causal inferences from time series to understand dynamics and relationships.

## **1.5 Data-Generating Process (DGP)**

The DGP refers to the underlying mechanism that produces a time series, which is often stochastic and complex. Accurate forecasting hinges on closely approximating this DGP with a mathematical model.

## **1.6 Synthetic Time Series Generation**

The book delves into generating synthetic time series by combining fundamental components such as noise and signals to simulate real-world data complexity. Techniques for creating white and red noise processes, cyclical or seasonal signals, and autoregressive signals are explained.

## 1.7 Stationary vs. Non-Stationary Time Series

Stationary time series have constant mean and variance over time, while non-stationary series display changes in these statistical properties due to trends or varying variance (heteroscedasticity), which complicates their analysis.

## 1.8 Forecastability of Time Series

Forecastability is dependent on understanding the DGP, the amount of historical data, and the presence of repeating patterns. The book categorizes time series predictability and discusses factors that influence it, such as the high predictability of tides compared to the randomness of lottery numbers or the volatility of stock prices.

## 1.9 Forecasting Terminology

The terminology section introduces concepts like multivariate forecasting, explanatory forecasting, backtesting, and forecast combination. It explains the role of in-sample and out-sample data, as well as exogenous and endogenous variables in the context of time series forecasting.

# 2 Predictability

The predictability of a time series is influenced by several factors:

- **Understanding the Data-Generating Process (DGP):** A thorough understanding of the mechanisms that generate the time series data leads to better forecasting because the models can approximate the actual process more closely.
- **Amount of Data:** Access to a large volume of historical data improves predictability by revealing underlying trends, cycles, and variabilities.
- **Repeating Patterns:** Predictability is higher when the time series exhibits clear and consistent patterns that repeat over time.
- **Type of Data:** The inherent nature of the time series affects its predictability. Series influenced by well-understood phenomena (like tides) are more predictable than those with high randomness (like lottery numbers).
- **Statistical Characteristics:** The stationarity of a time series—constant mean, variance, and covariance—facilitates predictability. Non-stationary series are more challenging due to trends, seasonality, and other instabilities.

Each of these factors contributes to the overall predictability of a time series, with some series being inherently more predictable than others based on these characteristics.

### 3 Forecasting

### 4 Chapter 01 Code Analysis

```
ar_value = [self.previous_value[i] * self.ar_param[i] for i in
             range(len(self.ar_param))]
noise = np.random.normal(loc=0.0, scale=self.sigma, size=1)
ar_value = np.sum(ar_value) + noise
self.previous_value = self.previous_value[1:]+[ar_value[0]]
return ar_value
```

### 5 Chapter 02 Summary

Steps for Acquiring and Processing Time Series Data:

1. Understanding the time series dataset
2. Preparing the data model
3. Handling missing data: Missing data can sometimes be informative, therefore it is important to take this into account in the data generation process
  - (a) Last Observation Carried Forward or Forward Fill
  - (b) Next Observation Carried Backward or Backward Fill
  - (c) Mean Value Fill
  - (d) Linear Interpolation
  - (e) Nearest Interpolation
  - (f) Spline, Polynomial, and Other Interpolations
4. Converting data into time series data
  - (a) Time Series Identifiers
  - (b) Metadata or Static Features
  - (c) Time-Varying Features
  - (d) Find the Global End Date
  - (e) Basic Preprocessing
  - (f) Mapping additional information
5. Compact, expanded, and wide forms of data
  - (a) Wide (useless): We have the date as an index or as one of the columns and the different time series as different columns of the DataFrame. As the number of time series increases, they become wider and wider, hence the name.

- (b) Expanded: is when the time series is expanded along the rows of a DataFrame. If there are n steps in the time series, it occupies n rows in the DataFrame.
  - (c) Compact: is when any particular time series occupies only a single row in the pandas DataFrame – that is, the time dimension is managed as an array within a DataFrame row.
6. Enforcing regular intervals in time series
  - 7.

## 6 Data Preprocessing

## 7 Missing Data

## 8 Chapter 02 Code Analysis

1. Week of Year: `df.date.dt.isocalendar().week.iloc[0]` //error in reference to weekofyear
2. Why add 4 day to the date range in line 22?
3. Resampling is useful for aggregations and deaggregations
4. Difference between `y1=df.ISE` and `y2=df['ISE']`?
- 5.

This line should be skipped for the code to work

```
df = df.loc["2022-07-07 7:00":"2022-07-07 09:00", "pm2_5_1_hr"]
fig = px.line(df, x=df.index, y="pm2_5_1_hr", title="Missing_
Values_in_PM2.5")
fig = format_plot(fig, ["Original"])
fig.write_image("imgs/chapter_2/missing_values.png")
fig.show()
```

There is a code error referencing the parent directory

```
os.makedirs("imgs/chapter_2", exist_ok=True)
source_data = Path("scripts/data/london_smart_meters/") #
    scripts is added here
block_data_path = source_data/"hbblock_dataset/hbblock_dataset"
fig.show()
```

## 9 Chapter 03 Summary

Important topics discussed in chapter

1. Components of a time series
2. Visualizing time series data
3. Decomposing a time series
4. Detecting and treating outliers

### 9.1 Components of a Time Series

The following terms can be mixed in different ways, but two very commonly assumed ways are additive and multiplicative

1. Trend: is a long-term change in the mean of a time series. It is the smooth and steady movement of a time series in a particular direction.
2. Seasonal: When a time series exhibits regular, repetitive, up-and-down fluctuations.
3. Cyclical: is often confused with seasonality, but it stands apart due to a very subtle difference. Like seasonality, the cyclical component also exhibits a similar up-and-down pattern around the trend line, but instead of repeating the pattern every period, the cyclical component is irregular.
4. Irregular: is left after removing the trends, seasonality, and cyclicity from a time series. Traditionally, this component is considered unpredictable and is also called the residual or error term. (Not completely useless. Can sometimes be explained by an exogenous variable to a certain extent)

### 9.2 Visualizing Time Series Data

1. Line Charts
2. Seasonal (box)Plots
3. Calendar Heatmaps
4. Autocorrelation plot

### 9.3 Decomposing a Time Series

#### 9.3.1 Detrending

Here we estimate the trend component (which is the smooth change in the time series) and remove it from the time series, giving us a detrended time series.

Types:

1. Moving averages: It can be seen as a window that is moved along the time series in steps, and at each step, the average of all the values in the window is recorded.
2. Locally Estimated Scatterplot Smoothing (LOESS) Regression: We use an ordinal variable that moves between the time series as the independent variable and the time series signal as the dependent variable. For each value in the ordinal variable, the algorithm uses a fraction of the closest points and estimates a smoothed trend using only those points in a weighted regression. The weights in the weighted regression are the closest points to the point in question. This is given the highest weight and it decays as we move farther away from it. This gives us a very effective tool for modeling the smooth changes in the time series (trend).

### 9.3.2 Deseasonalizing

Here, we estimate the seasonality component from the detrended time series. After removing the seasonal component, what is left is the residual.

1. Moving averages: It can be seen as a window that is moved along the time series in steps, and at each step, the average of all the values in the window is recorded.
2. Fourier series: Any periodic function, no matter the shape, curve, or absence of it, or how wildly it oscillates around the axis, can be broken down into a series of sine and cosine waves.

## 9.4 Detecting and Treating Outliers

An outlier, as its name suggests, is an observation that lies at an abnormal distance from the rest of the observations. This can be for many reasons, including faulty measurement equipment, incorrect data entry, and black-swan events, to name a few. Being able to detect such outliers and treat them may help your forecasting model understand the data better.

Techniques:

1. Standard Deviation
2. Interquartile range
3. Isolation Forest
4. Extreme studentized deviate (ESD) and seasonal ESD (S-ESD)

## 10 Decomposition

## 11 Outliers