Machine Learning for **Time Series Forecasting**

ABOUT ME

VITOR CERQUEIRA

Postdoc at Dalhousie University, Canada
Ph.D (with honors) from U. Porto

- Researcher focused on learning from time series
- 13 journal papers on top quartile venues + several conf.s
    - best paper on ECML/PKDD 2017
- Top Writer on Medium (AI and Time Series)

Find me @ https://bio.link/vcerq

Time Series **Forecasting**

Introduction

1

motivation, objectives, basic components and properties

Supervised Learning

2

classical models, pre-processing, feature engineering, modeling

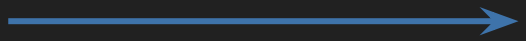
Evaluation

3

metrics and cross-validation best practices

01 Introduction

- **Motivation and objectives of time series analysis**
- The basic components of a time series
- A few useful things to know about time series



Time series are **relevant across industries**

ITS

Origin-destination flow estimation, travel time prediction

Stock Market

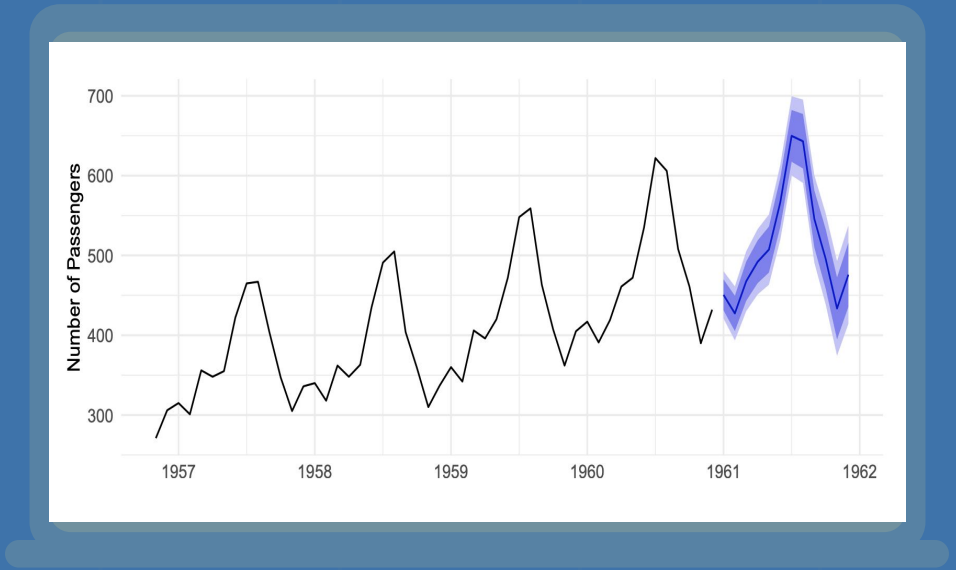
Monitoring price action for trading financial assets

Energy

Managing energy supply and demand

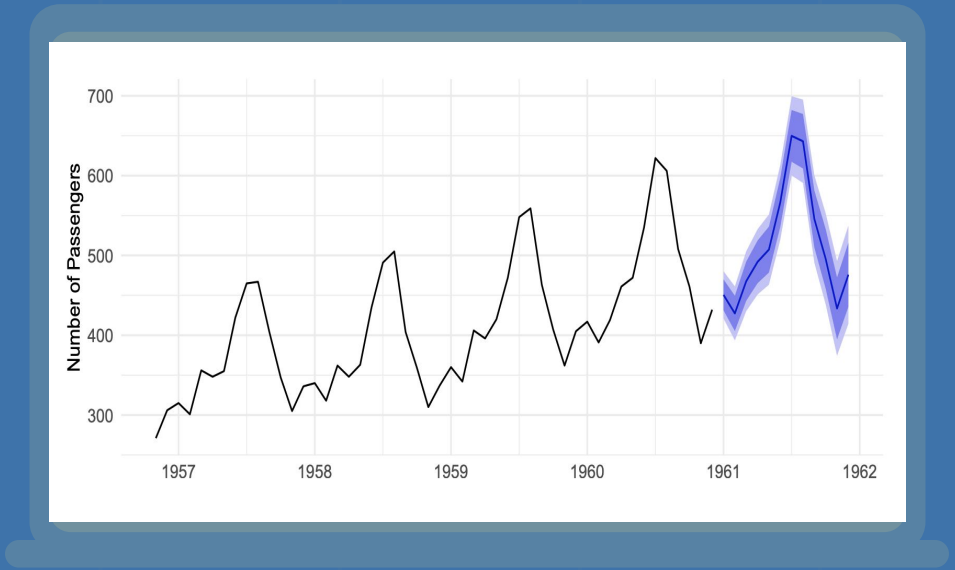
What is a **Time Series?**

- A time series is a sequence of values ordered in time.
- Denoted as $Y = y_1, y_2, \ldots, y_n$, where y_i is the value at time *i*.



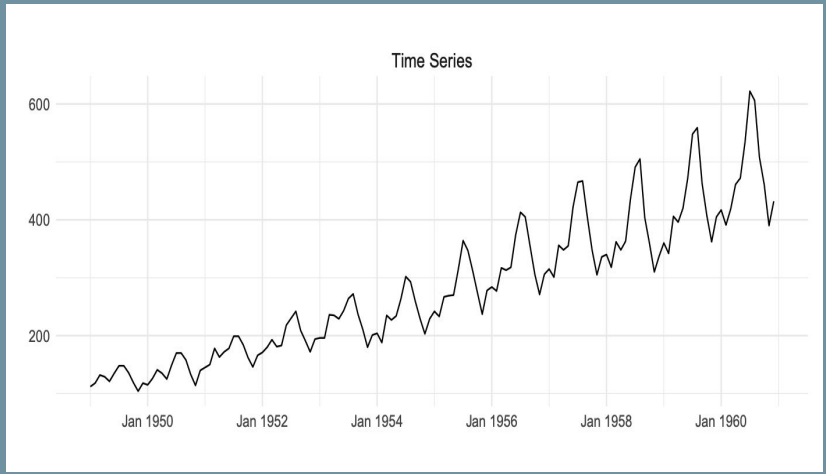
What is a **Time Series?**

- Observations are numeric
- There is a temporal dependency
- Susceptibility to change



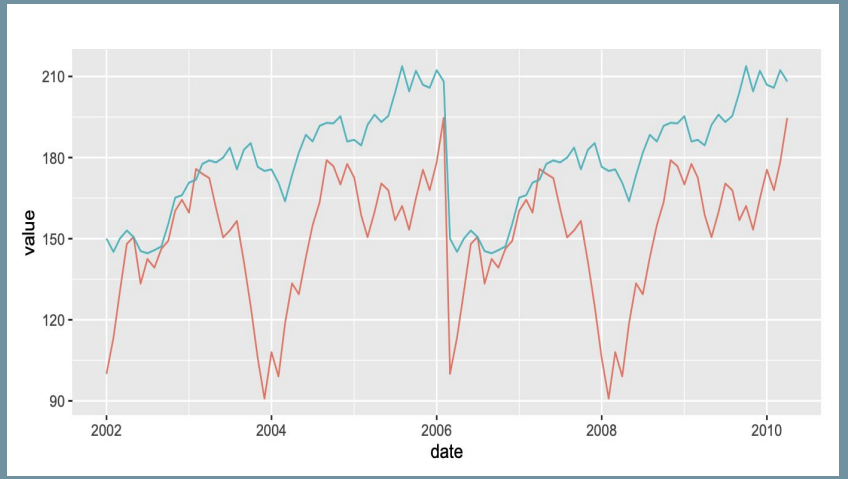
Objectives of Time Series Analysis

- Describe relevant patterns, such as trend or seasonality
- Explain how the past affects the future



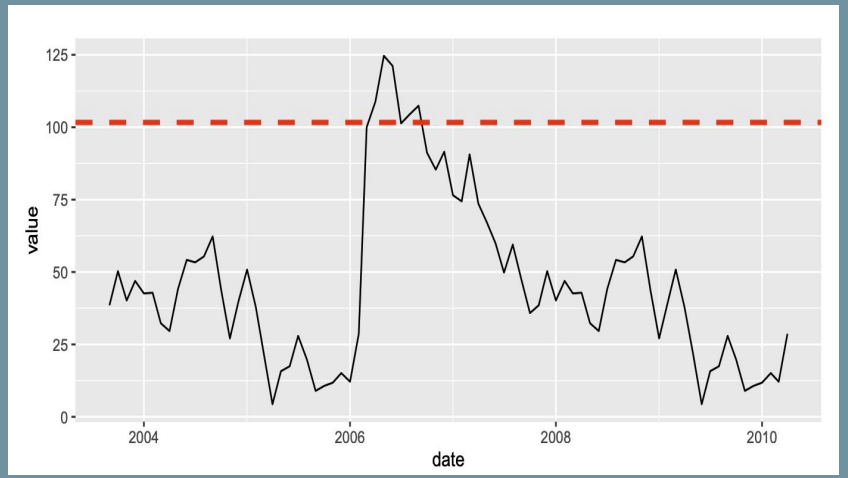
Objectives of Time Series Analysis

- Explain how two time series interact with each other



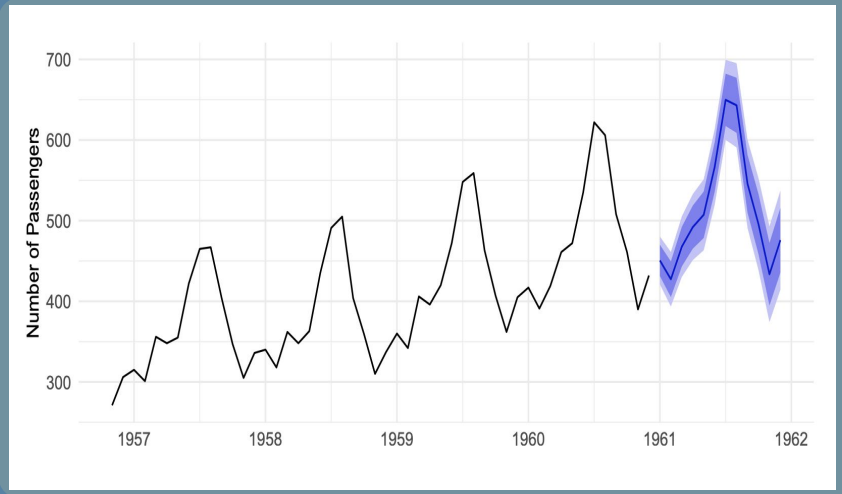
Objectives of Time Series Analysis

- Process control – analyse when time series exceed critical thresholds



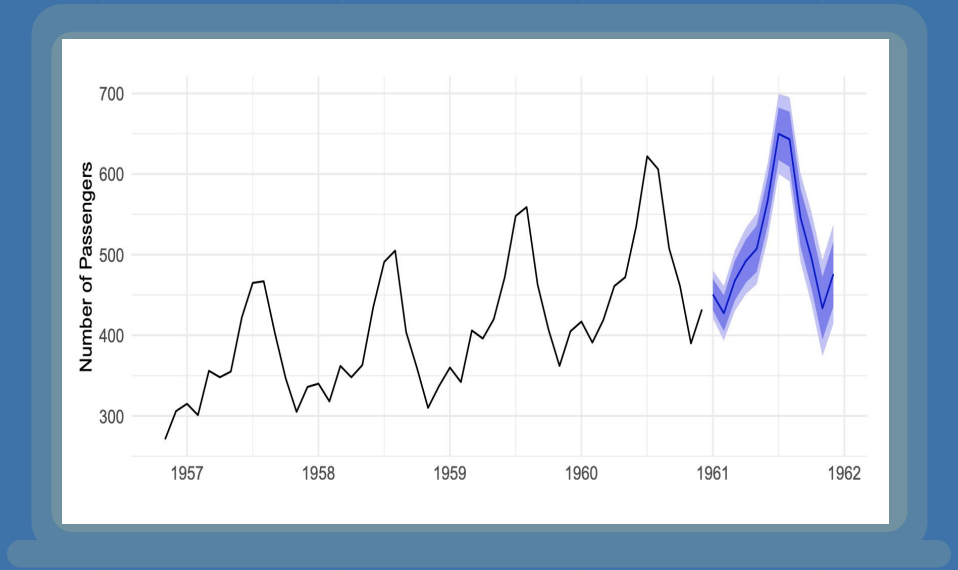
Objectives of Time Series Analysis

- Forecast future observations



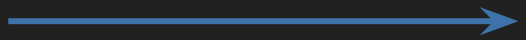
Objectives of Time Series Analysis

- **Ultimately**: Understand how the time series evolves over time to improve decision-making within organizations



01 Introduction

- Motivation and objectives of time series analysis
- **The basic components of a time series**
- A few useful things to know about time series



Components

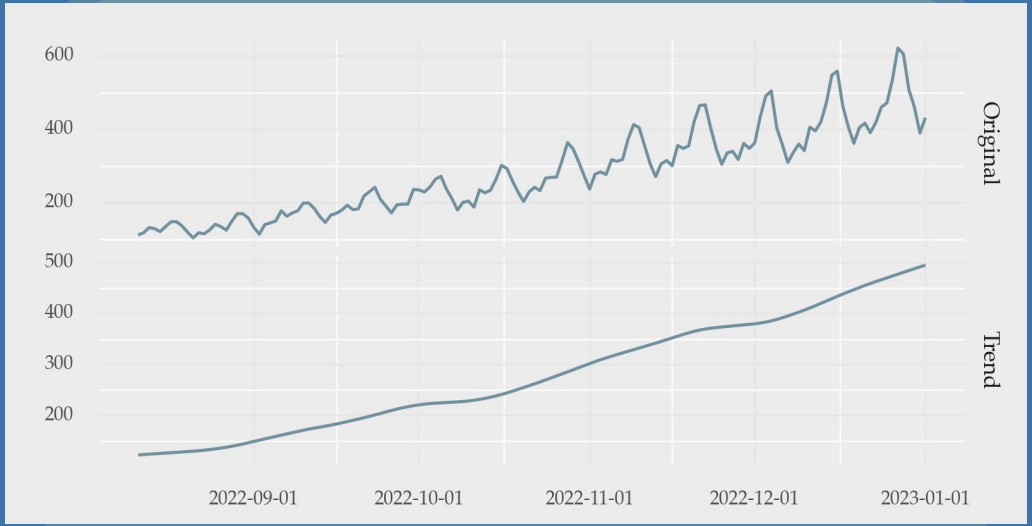
Trend

Long-term change in the level of the series

how to detect (two examples):

- KPSS test:
  - null hypothesis that the time series is stationary around a deterministic trend; alternatively, the series contains a unit root
- ADF test:
  - null hypothesis that the time series contains a unit root; alternatively, the series is stationary

time series with unit roots contain a stochastic trend



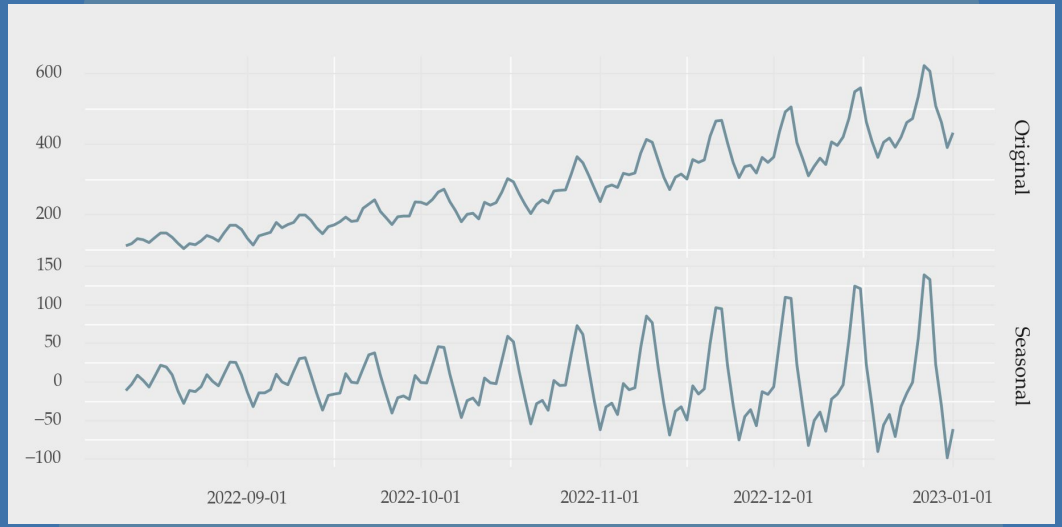
Components

Seasonality

Regular and predictable changes in fixed periods (e.g. every day).

how to detect (2 examples):

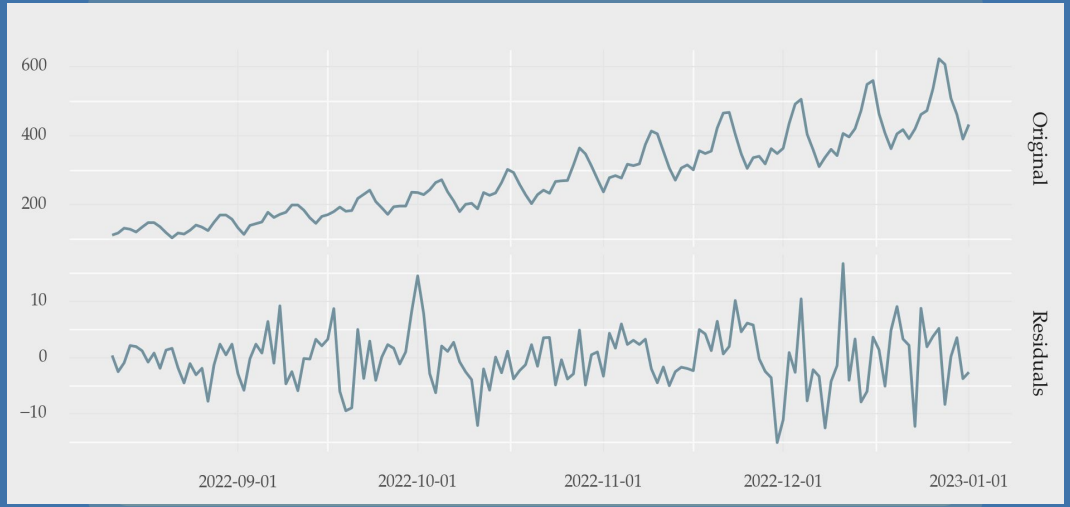
- OCSB test:
  - null hypothesis that the time series contains a seasonal unit root; alternatively, the series is stationary
- Canova-Hansen test
  - null hypothesis that the time series is stationary around a deterministic seasonality; alternatively, the series contains a unit root



Components

Remainder

The remaining fluctuations after removing trend and seasonality



Decomposition

At any given time step *t*:

$y_i = Trend_i + Seasonal_i + Remainder_i$

Decomposition can also be multiplicative

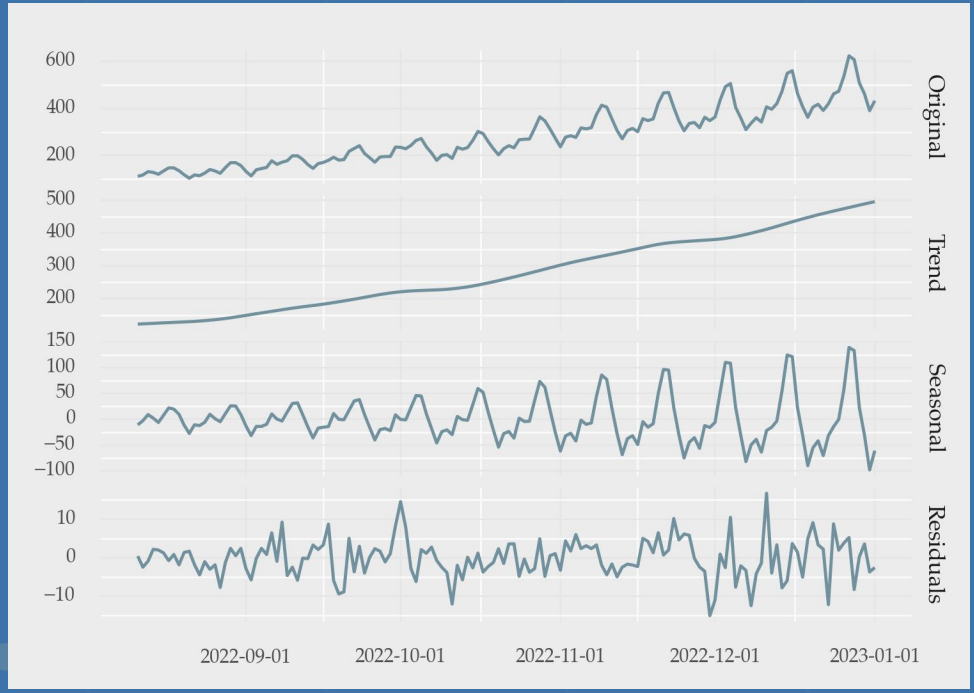
- if magnitude of fluctuations varies with the level of the series

but,

$y_i = Trend_i \times Seasonal_i \times Remainder_i$

is the same as:

$\log(y_i) = \log(Trend_i) + \log(Seasonal_i) + \log(Remainder_i)$

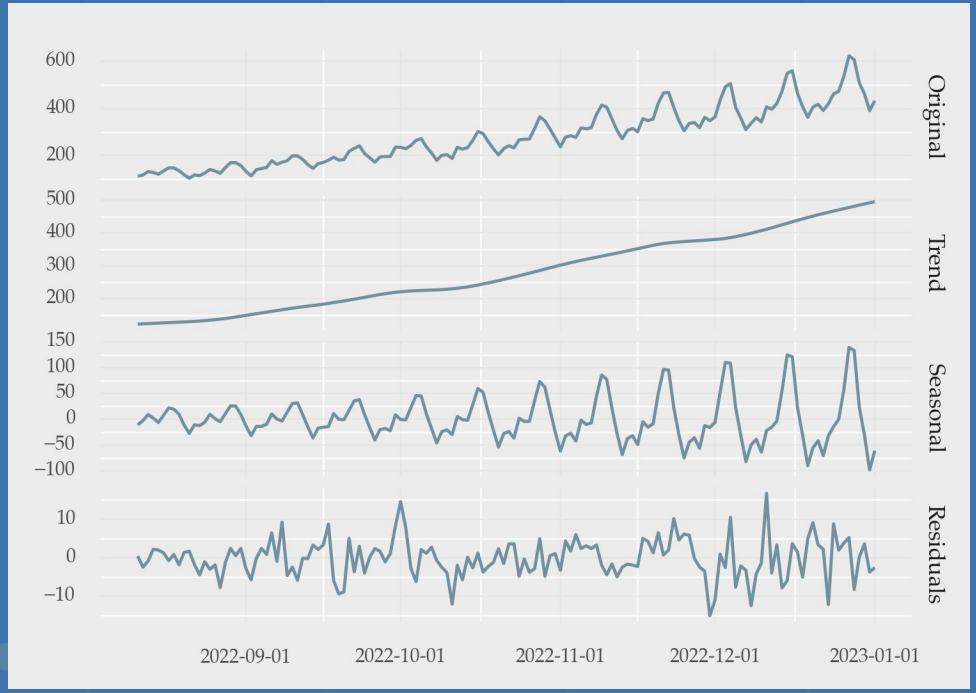


Decomposition

Two example methods:

- **Classical**
  - Trend = moving average of order m
  - Seasonality = average values by season

- **STL** (seasonal and trend decomp. with LOESS)
  - LOESS: Local regression or local polynomial regression
  - handles any type of seasonality, unlike some methods
  - handles changes in the seasonal component
  - MSTL: for multiple seasonal periods



01 Introduction

- Motivation and objectives of time series analysis
- The basic components of a time series
- **A few useful things to know about time series**

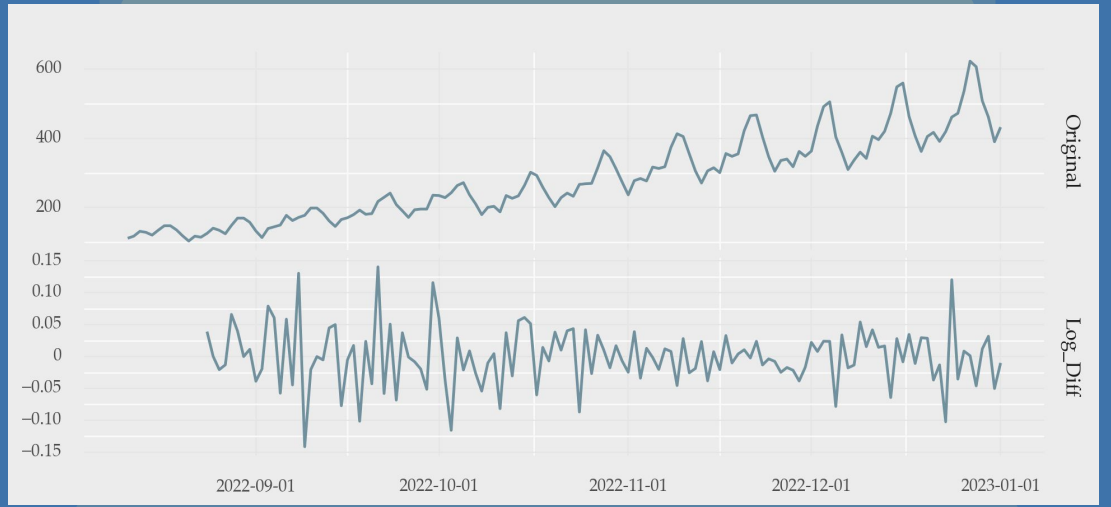
Key Properties

Stationarity

A time series is stationary

if its properties do not change over time.

in practice: constant mean and variance, and auto covariance does not depend on time

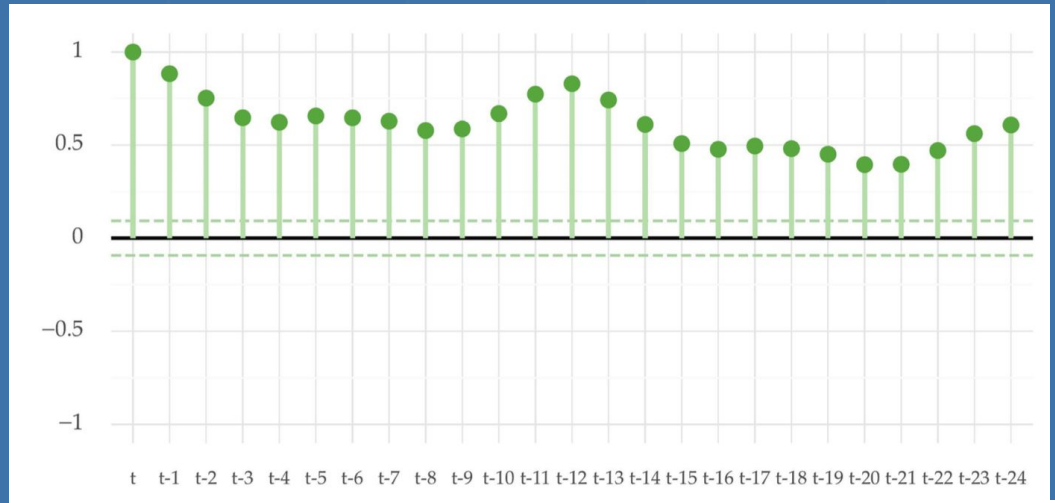


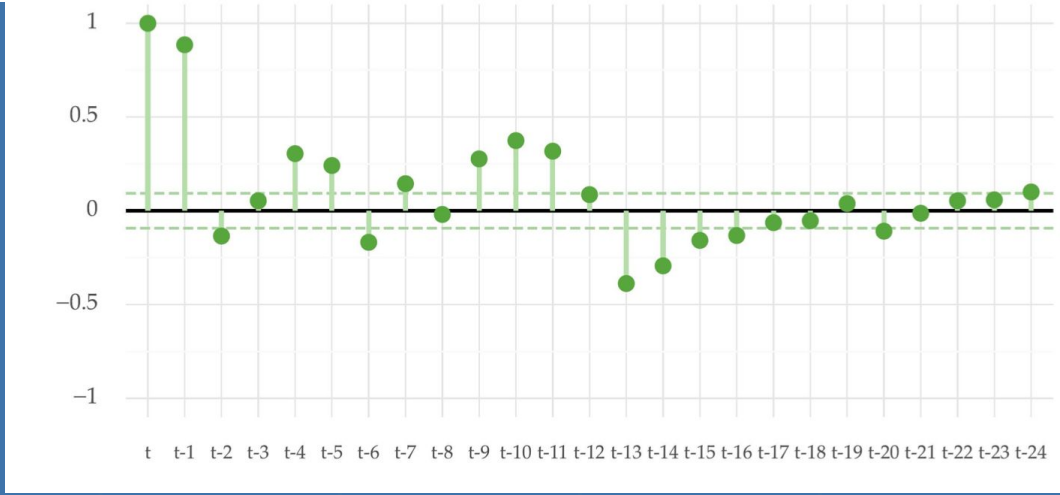
Key Properties

Auto-Correlation

Correlation of the series with its past lags

- auto-correlation
- partial auto-correlation





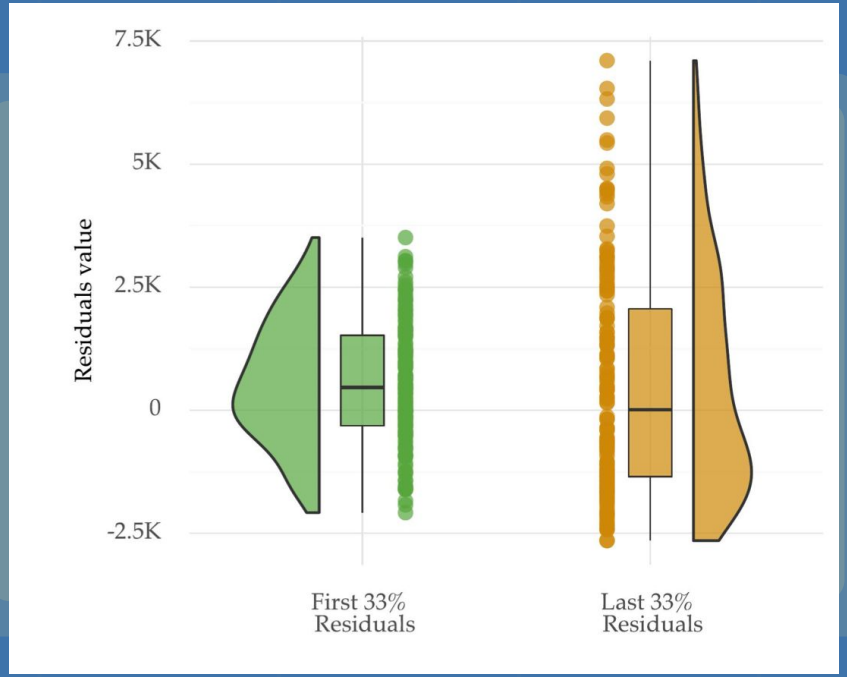
Key Properties

Heteroskedasticity

When the variance is not constant

how to detect:

- Goldfeld-Quandt test
- White test



Key Properties

Frequency

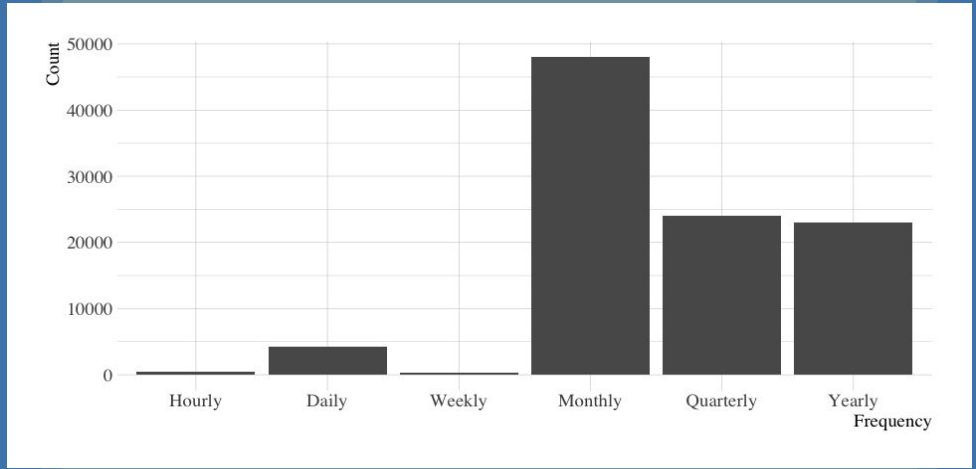
How often is the data observed

high frequency:

- lots of data
- many seasonal patterns

low frequency:

- data scarcity



Key Properties

Regularity

Most methods assume that the time series is sampled with a regular frequency

Some time series are naturally irregular.

e.g. sales of a product

2021-08-14 11:30:00

2021-08-14 15:18:00

2021-08-14 15:53:00

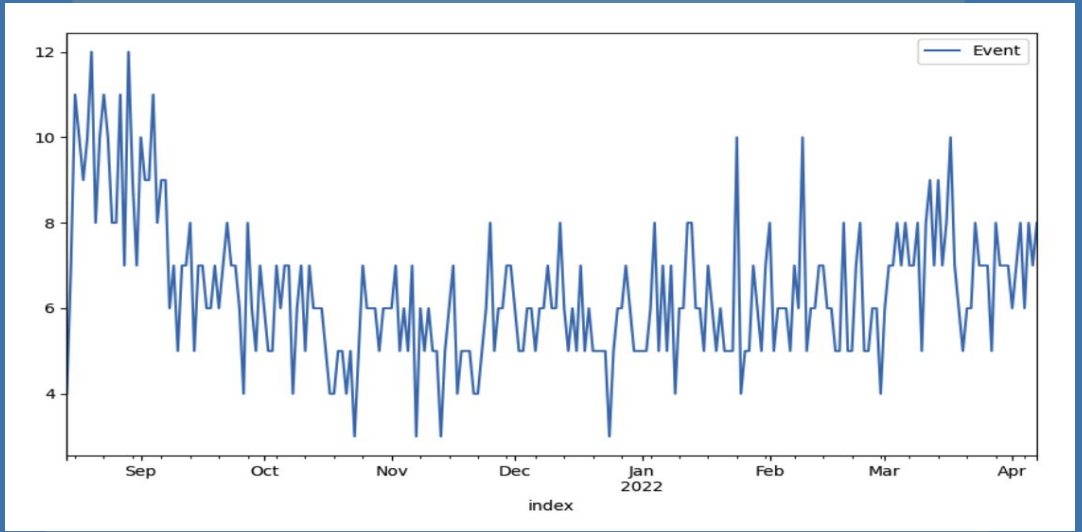
2021-08-14 18:55:00

2021-08-15 01:00:00

...

Standard solution: resampling

But, this might lead to **intermittency**



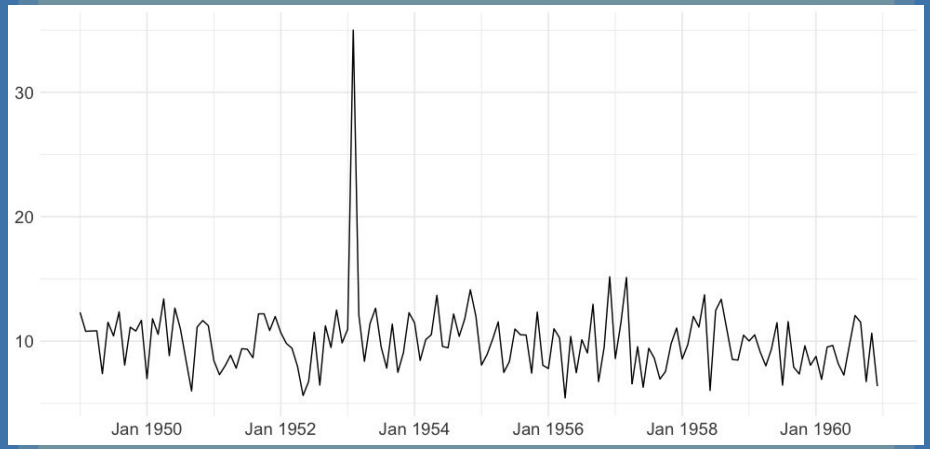
Key Properties

Outliers

Unusual observations or sequences of observations

can be:

- unwanted data, e.g. annotation errors, noise, damaged sensores
- critical instances, e.g. fraud

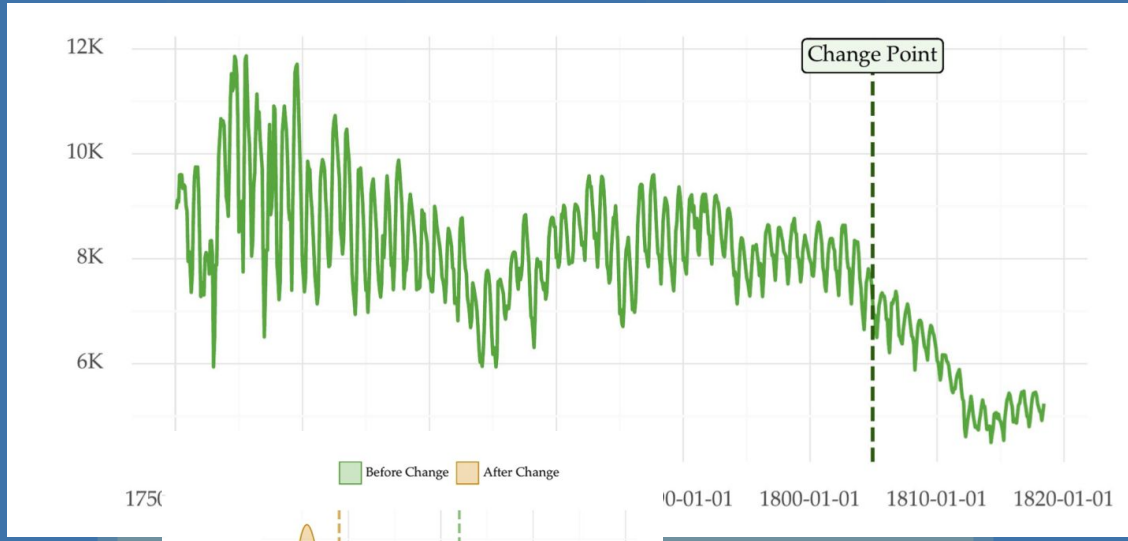


Key Properties

Regimes and Change Points

the process that generates the observations

- it can change, which leads to a change in the distribution
- changes can be permanent or recurrent
- changes can be abrupt (structural breaks) or gradual



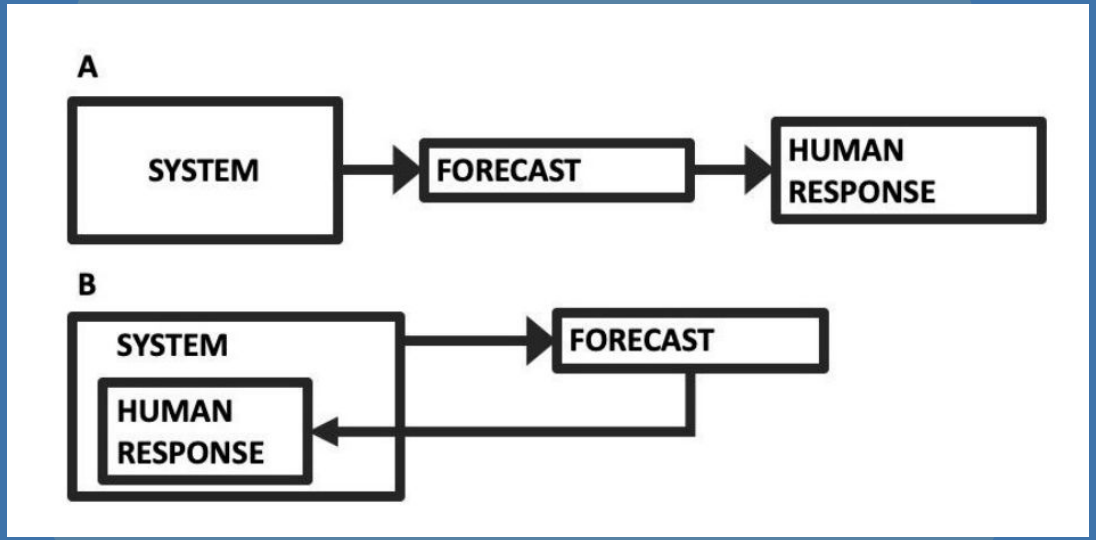
Key Properties

Reflexivity

when the forecast affects the thing being forecasted

stock market: forecasts of an increase of an assets' price attracts investors. This increases the demand for this assets, thereby increasing its price.

self-fulfilling prophecies



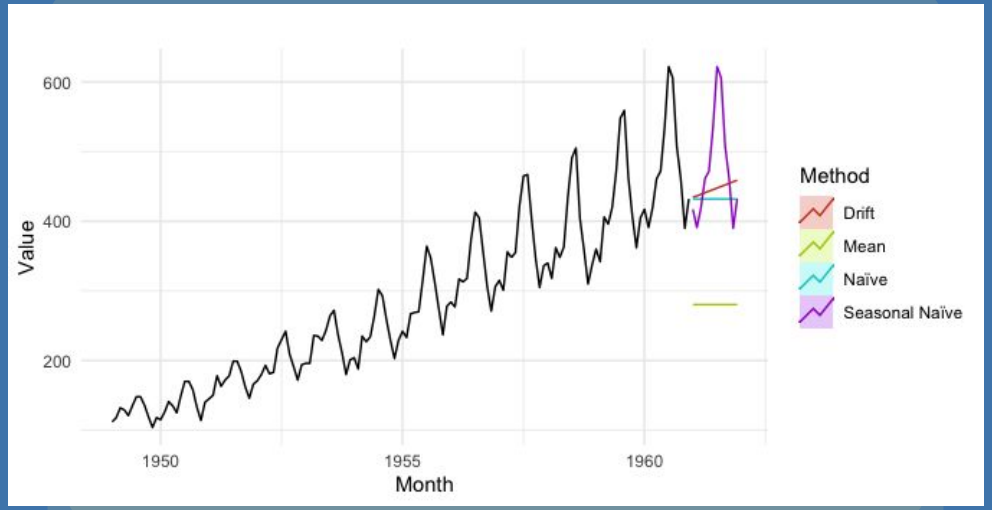
Record, Nicholas R., and Andrew J. Pershing. "Facing the Forecaster's Dilemma: Reflexivity in Ocean System Forecasting." Oceans. Vol. 2. No. 4. MDPI, 2021.

02 Supervised Learning

- **Classical models**
- Supervised learning via time delay embedding
- Transformations and feature engineering

Baselines

- ➔ **mean**: historical average
- ➔ **naive**: previous observation
- ➔ **seasonal naive**: previous observation of the same season
- ➔ **drift:** naive plus average change



Exponential smoothing

Each observation is a weighted average of past ones. Weights decay exponentially.

Several flavors:

- Simple exponential smoothing: the most basic
- Double exponential smoothing: for time series with trend
- Triple exponential smoothing (aka Holt-Winters): for time series with trend and seasonality.

Can be coupled with machine learning models:

- Winner of the M4 forecasting competition: normalize time series with ES and forecast with LSTM

Smyl, Slawek. "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting." International Journal of Forecasting 36.1 (2020): 75-85.

ARIMA

Auto-regressive Integrated Moving Average

Also known as Box-Jenkins method

AR(p):

- Past p recent values are explanatory variables
- Current value is the target

MA(q):

- Past q recent errors are explanatory variables
- Current value is the target

I(d):

- Difference d times (until stationarity)

02 Supervised Learning

- Classical models
- **Supervised learning via time delay embedding**
- Transformations and feature engineering

Supervised Learning

Two pieces of data:

- *X*: explanatory variables
- *y*: target variable
  - regression task if y is numeric, classification task otherwise

Auto-Regression (p)

- *X*: past recent observations
  - p is the lag size
- *y*: future observation(s)
  - size according to the forecasting horizon

$$Y_{[n,p]} = \left[\begin{array}{ccccc|c} y_1 & y_2 & \cdots & y_{p-1} & y_p & y_{p+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i-p+1} & y_{i-p+2} & \cdots & y_{i-1} & y_i & y_{i+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n-p+1} & y_{n-p+2} & \cdots & y_{n-1} & y_n & y_{n+1} \end{array}\right]$$

Takens, Floris. "Detecting strange attractors in turbulence." Dynamical systems and turbulence, Warwick 1980. Springer, Berlin, Heidelberg, 1981. 366-381.

Supervised Learning

Extensions to Auto-Regression (p)

Global Forecasting Models

- Training a model using multiple (related) time series
- Time series are concatenated row-wise after time delay embedding and normalization

Auto-regressive Distributed Lags (ARDL)

- Extra time series are added as explanatory variables

$$Y_{[n,p]} = \left[\begin{array}{ccccc|c} y_1 & y_2 & \cdots & y_{p-1} & y_p & y_{p+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i-p+1} & y_{i-p+2} & \cdots & y_{i-1} & y_i & y_{i+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n-p+1} & y_{n-p+2} & \cdots & y_{n-1} & y_n & y_{n+1} \end{array}\right]$$

Supervised Learning

Multi-step ahead forecasting

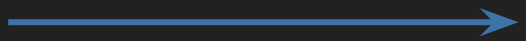
Common approaches:

- Recursive
- Direct
- Chaining (Recursive+Direct)
- Multi-output

$$Y_{[n,p]} = \left[\begin{array}{ccccc|ccc} y_1 & y_2 & \cdots & y_{p-1} & y_p & y_{p+1} & \cdots & y_{p+h} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i-p+1} & y_{i-p+2} & \cdots & y_{i-1} & y_i & y_{i+1} & \cdots & y_{i+h} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n-p+1} & y_{n-p+2} & \cdots & y_{n-1} & y_n & y_{n+1} & \cdots & y_{n+h} \end{array}\right]$$

02 Supervised Learning

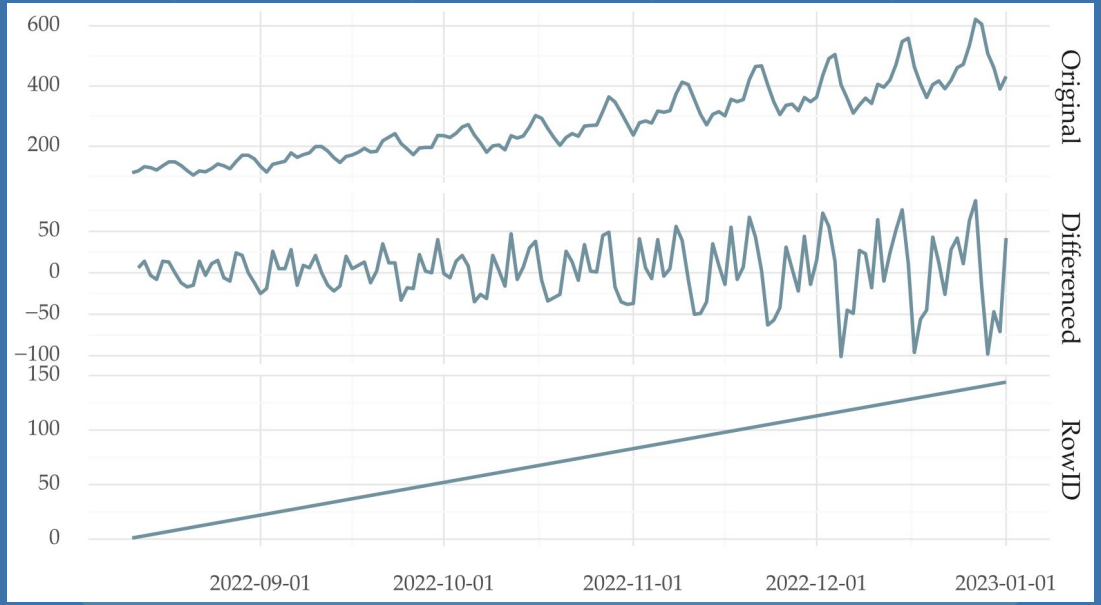
- Classical models
- Supervised learning via time delay embedding
- **Transformations and feature engineering**



Feature Engineering

How to deal with trend

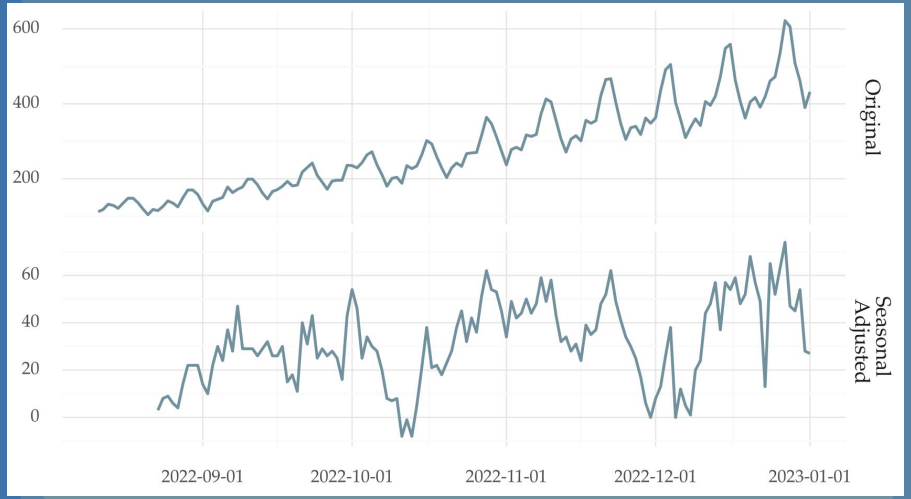
1. Differencing
2. Curve fitting
    a. modeling the residuals
3. Row id as feature



Feature Engineering

How to deal with seasonality

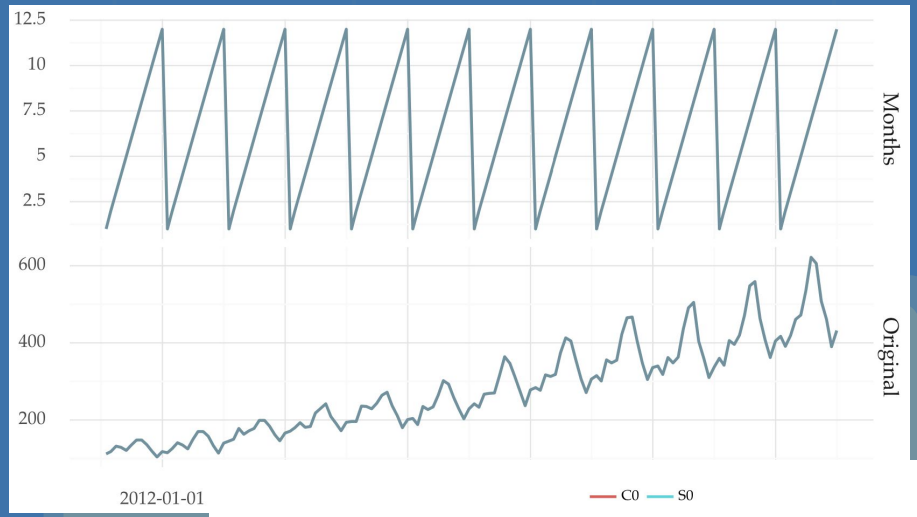
1. Seasonal differencing
2. Period information as explanatory variable
   a. blind to continuity
3. Trigonometric representations (Fourier terms)

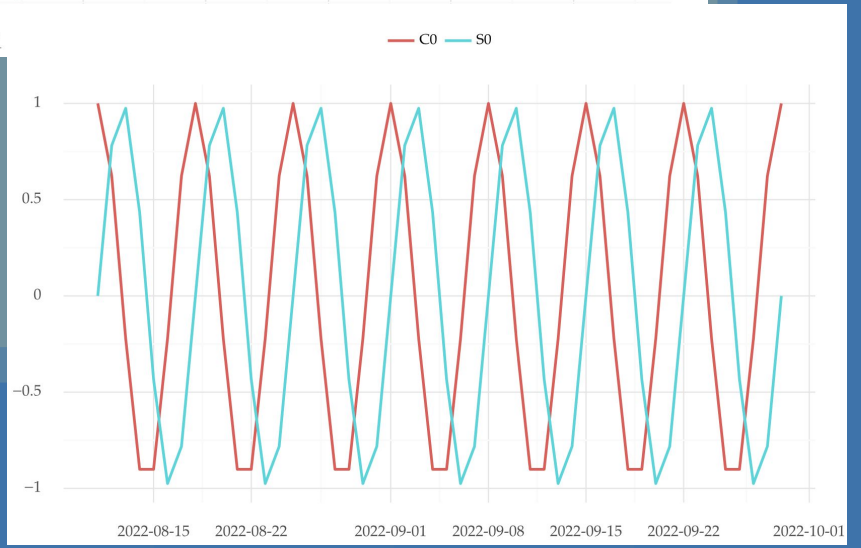


Feature Engineering

How to deal with seasonality

1. Seasonal differencing
2. **Period information as explanatory variable**
   a. **blind to continuity**
3. **Trigonometric representations (Fourier terms)**
4. Repeating basis functions

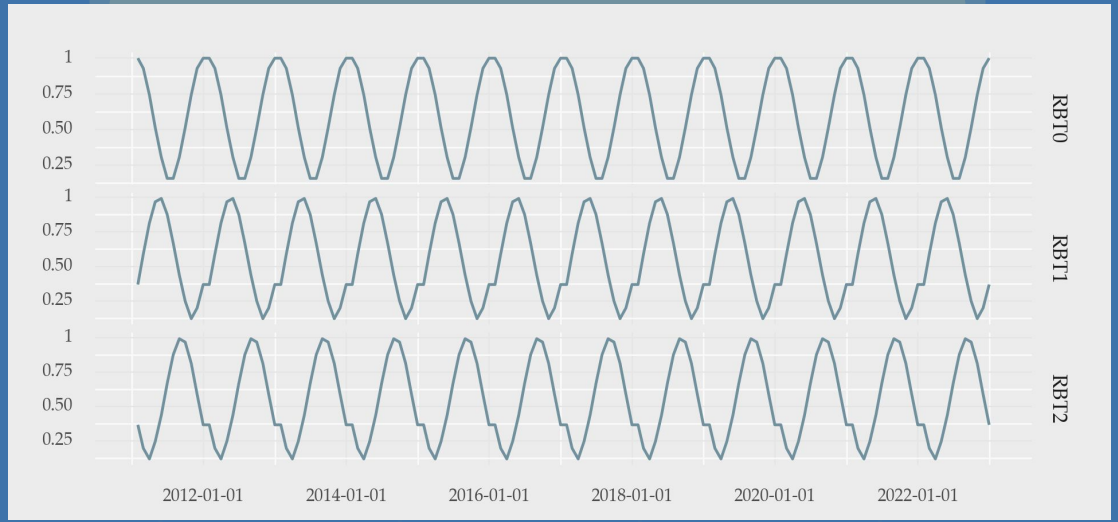




Feature Engineering

How to deal with seasonality

1. Seasonal differencing
2. Period information as explanatory variable
    a. blind to continuity
3. Trigonometric representations (Fourier terms)
4. **Repeating basis functions**



Feature Engineering

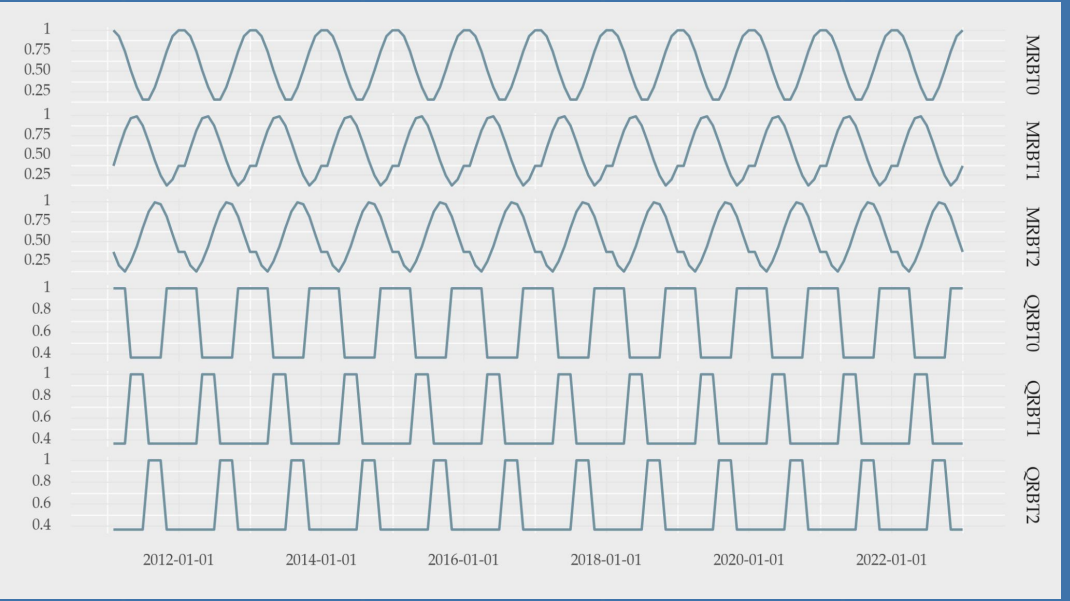
How to deal with multiple seasonal effects

for each relevant period:

1. Period information as explanatory variable
    a. blind to continuity
2. Trigonometric representations (Fourier terms)
3. Repeating basis functions

Off-the-shelf models

- Multi-Seasonal Time Series Decomposition
- TBATS
- Prophet
- Kernel-based Time-varying Regression (Orbit)



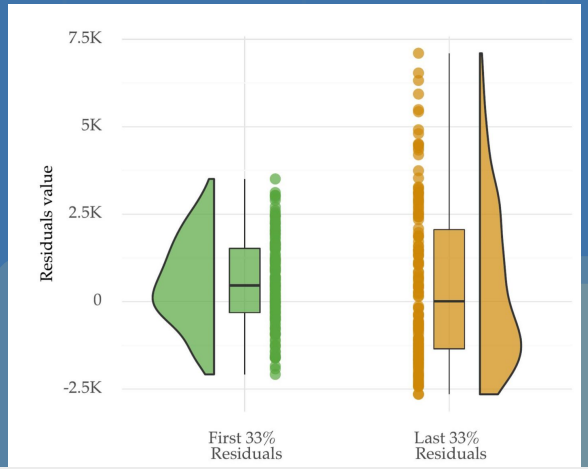
Feature Engineering

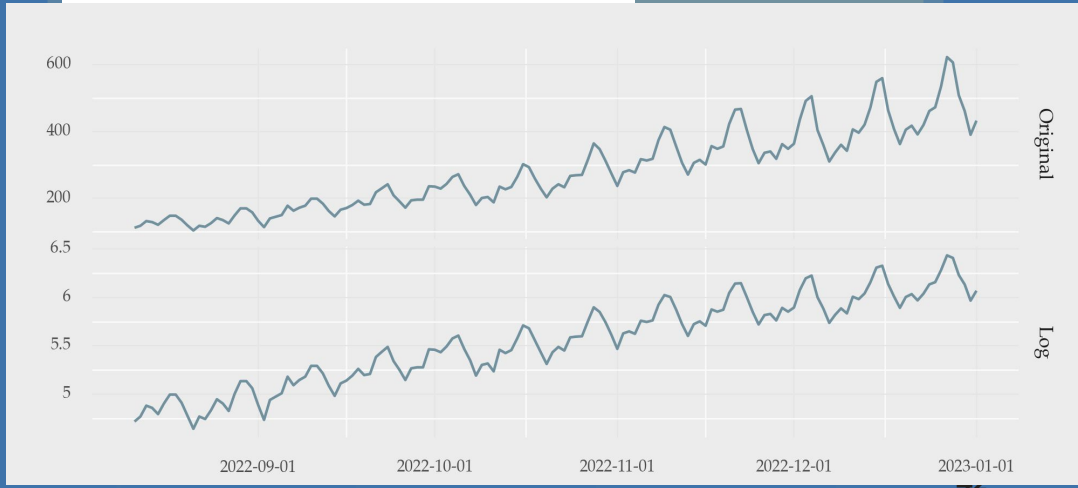
How to deal with non-constant variance

1. Log transformation
2. Box-cox transformation
3. Volatility standardization
    a. for multiple time series

If changes in variance have a structure:

- model variance with GARCH



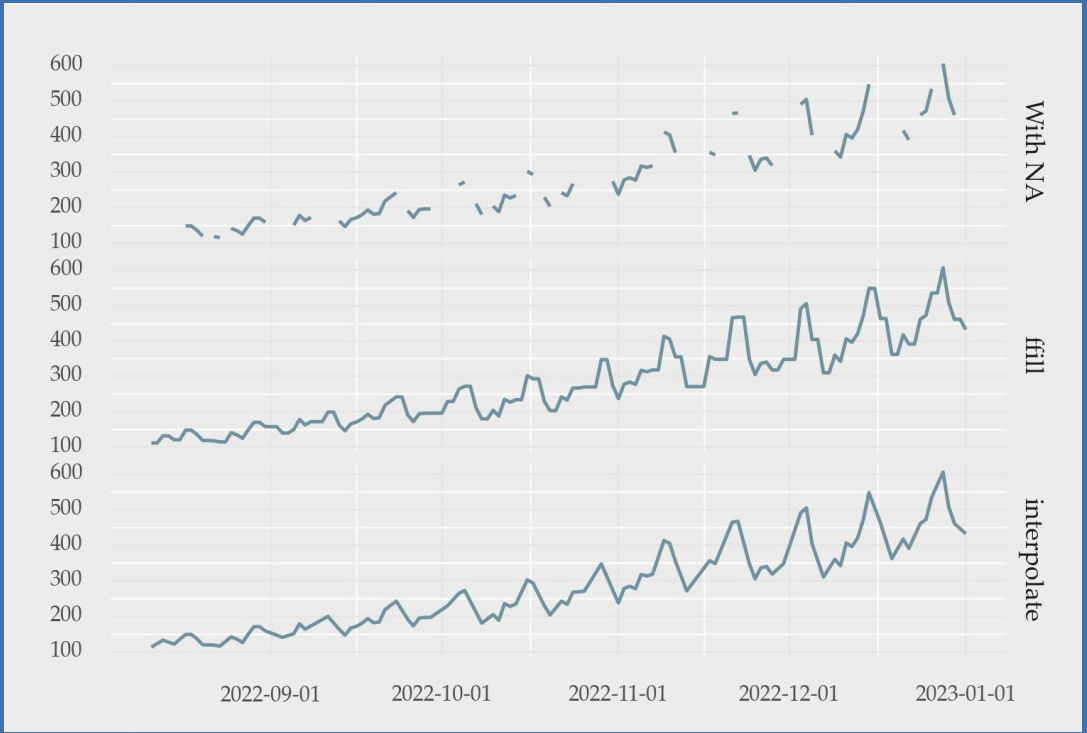


Feature Engineering

Imputation

1. Linear interpolation
2. Forward filling
3. Forward filling by period

Any other standard approach that doesn't leak future information to the past



Feature Engineering

Special Events and Holidays

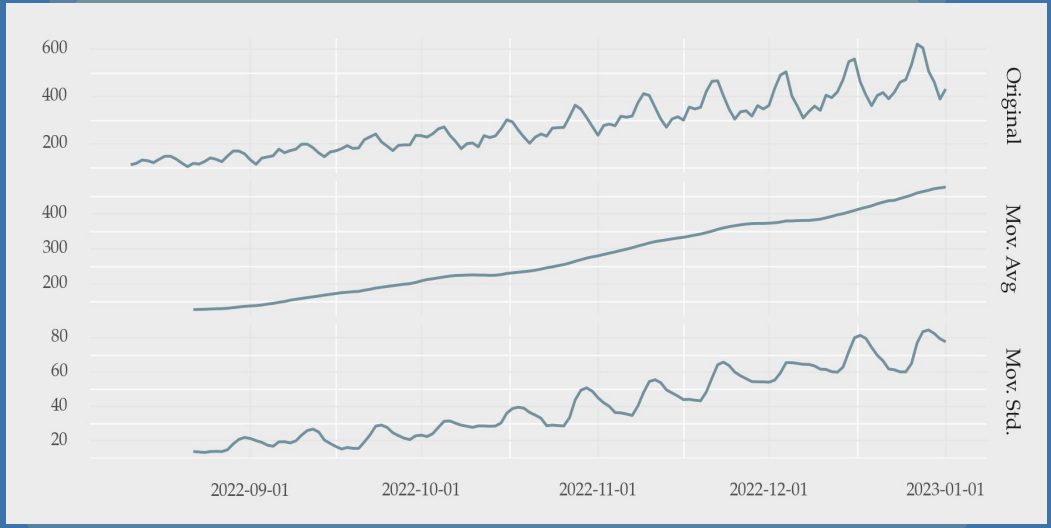
1. Information encoded as a binary variable
    a. consider also marking data before and after these events

Nice example: https://linkedin.github.io/greykite/docs/0.1.0/html/pages/model_components/0400_events.html

Feature Engineering

Summary statistics

- Summarise past recent observation using statistics:
  a. average
  b. variance
  c. quantiles
  d. plenty of other possibilities [1]
- Consider multiple rolling windows
  a. e.g. past day, week, etc.



Related:

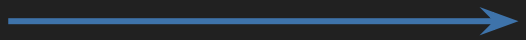
- Feature extraction to summarise a complete time series
- Example application: meta-learning the best model for an input time series [2]

[1] Cerqueira, Vitor, Nuno Moniz, and Carlos Soares. "Vest: Automatic feature engineering for forecasting." Machine Learning (2021): 1-23.

[2] Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). Fforma: Feature-based forecast model averaging. International Journal of Forecasting, 36(1), 86–92.

03 Evaluation

- **Cross-validating with time series**
- Evaluation metrics

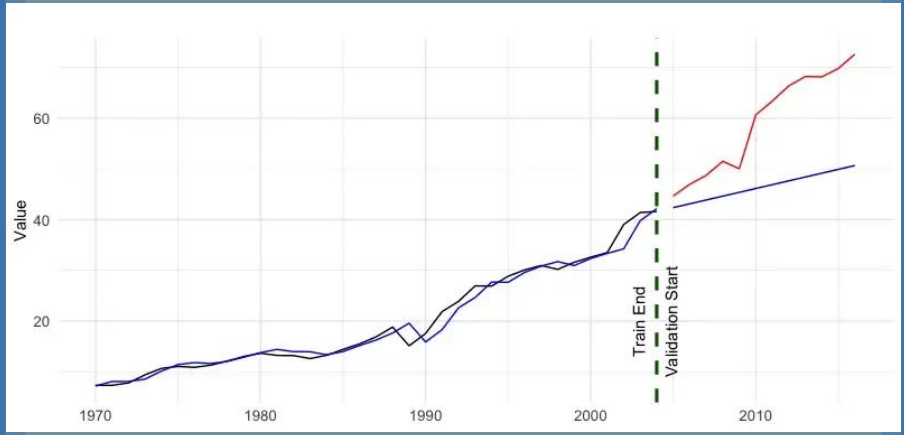


Cross-validation

- Training and testing a model on the same data leads to optimistic results
- Evaluation should be done using out-of-sample data

- Cross-validation: splitting the data many times
  - part of the data is used for training
  - remaining is used for validation

The idea is to replicate a realistic scenario



Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. J. Edic. Psychol., 22:45–55.

Cross-validation

It's important to make sure cross-validation addresses the time-dependent nature of time series

Some standard approaches (e.g. k-fold CV) fail to meet this condition

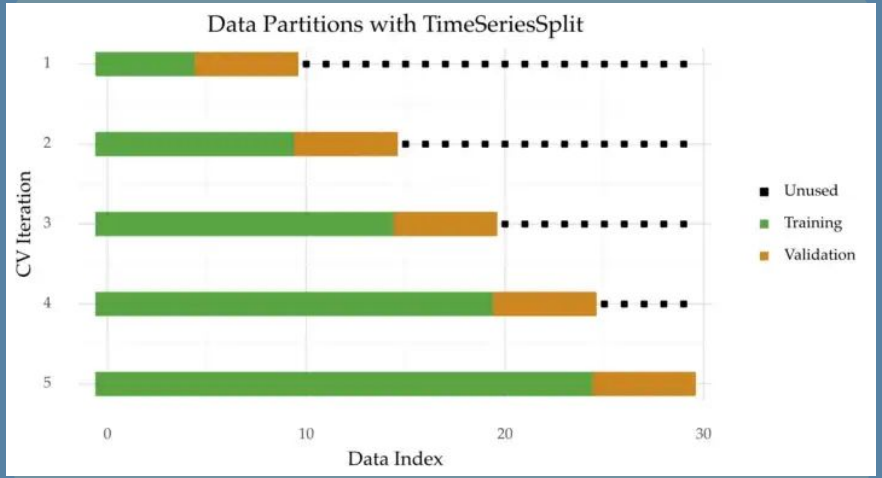
Let's look at a few aspects to consider

Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." Statistics surveys 4 (2010): 40-79.

Cross-validation

Preserve temporal order

- Otherwise the model learns nuances from the future that have not revealed themselves in the past
    - This leads to optimistic estimates

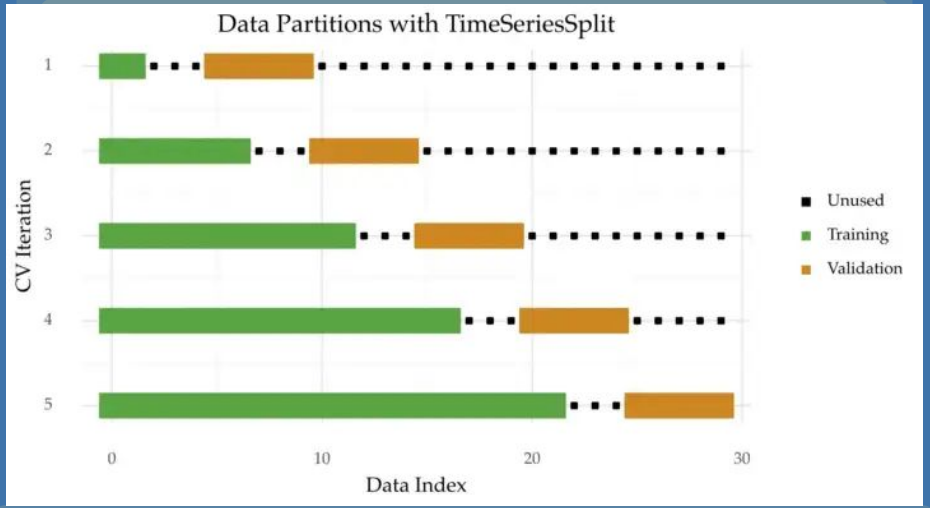


Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." Statistics surveys 4 (2010): 40-79.

Cross-validation

Purging (Gap Between Training and validation)

- Remove training observations close to the validation set
    - Increases independence



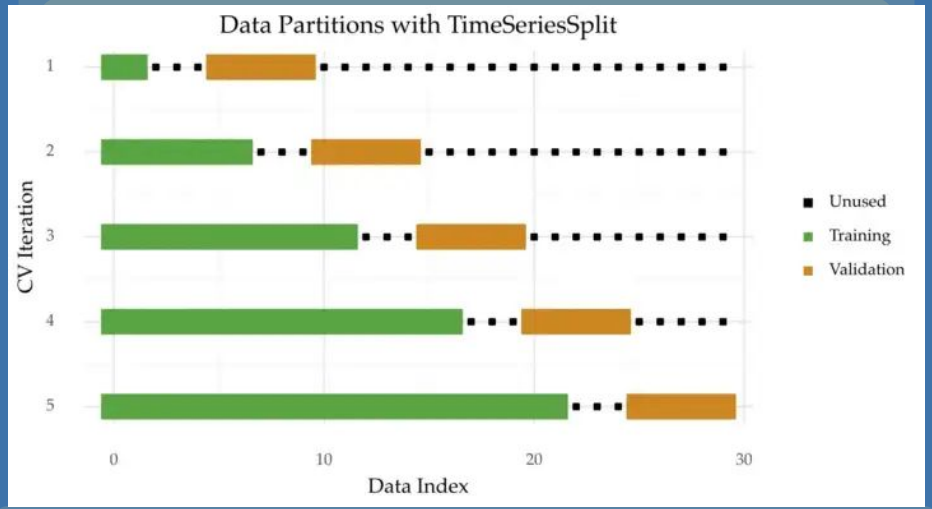
Cross-validation

Multiple Splits

- Carry out several iterations of cross-validation

- A single split may be biased due to the particularities of the selected origin
- Many splits will cover different parts of the time series.
    - For example, different trend or seasonality patterns.

Not as important if the data is large



Cross-validation

Re-training and Nesting

- Re-train the model using all data after cross-validation
- Use a three-way split for unbiased performance estimates (Nested CV)

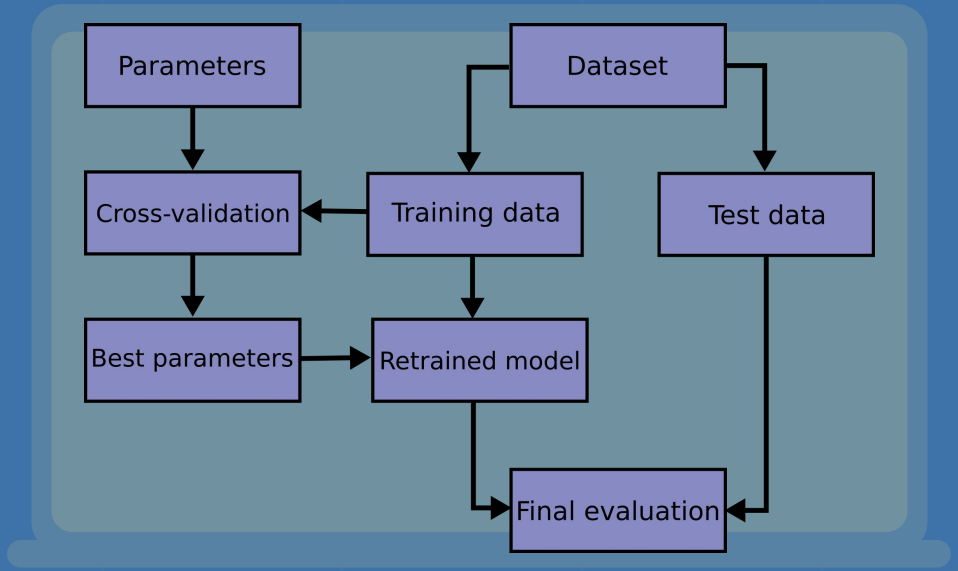
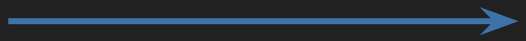


Image source: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-evaluating-estimator-performance

03 Evaluation

- Cross-validating with time series
- **Evaluation metrics**

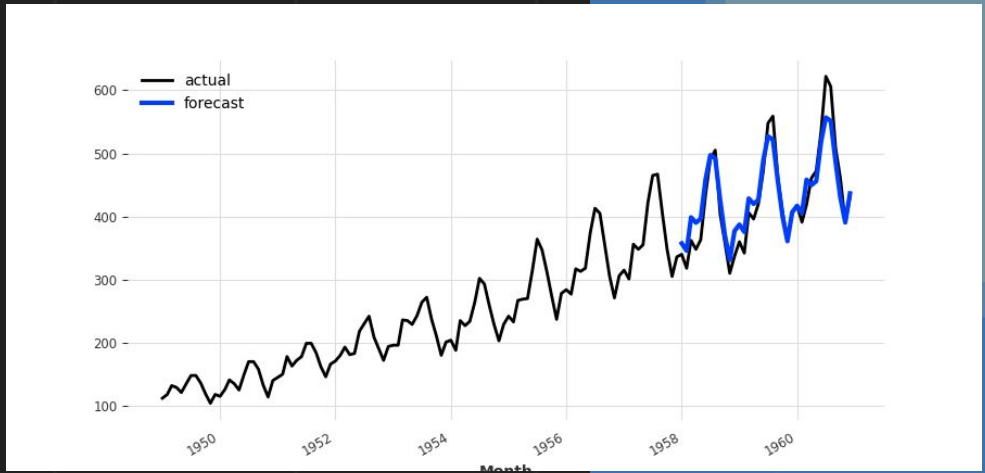


Evaluation

Forecasting Performance

A measure of how forecasts approximate actual values

$$e_i = y_i - \hat{y}_i$$



Evaluation

Regression Metrics

- scale dependent
  - difficult to compare across series

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Evaluation

Percentage or Relative Metrics

- scale independent
- MAPE runs into problems for y close to 0, unlike sMAPE or MASE
- sMAPE is actually not symmetric
  - penalizes under-estimations more than over-estimations
  - not interpretable
- MASE is MAE scaled by the MAE of a baseline

$$\text{MAPE} = 100 \times \text{mean}\left(\frac{|e|}{|y|}\right)$$

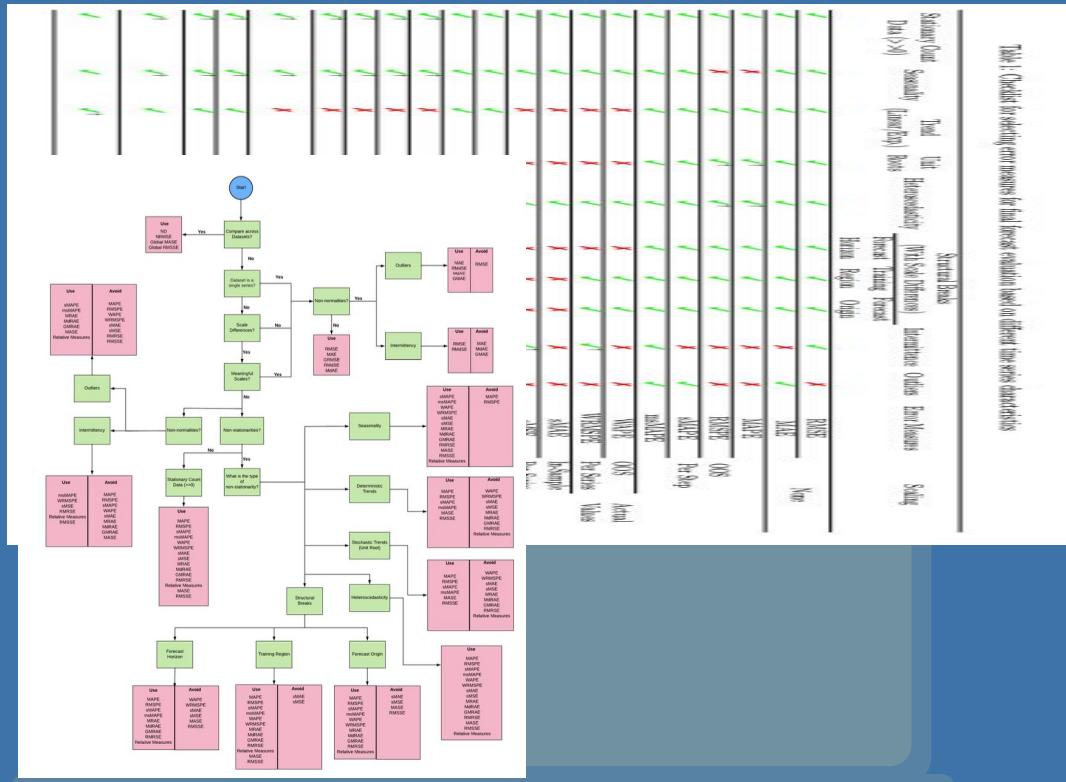
$$\text{sMAPE} = \frac{1}{n}\sum_{n}^{i=1}\left(\frac{200 * |e_i|}{|y_i| + |\hat{y}_i|}\right)$$

$$\text{MASE} = \frac{\text{MAE}}{\text{MAE}_{naive}(training)}$$

Evaluation

Many other metrics...

- different metrics for different scenarios
- some are more appropriate than others depending on the type of non-stationarities
- most metrics agree on which is the best model (provided a large enough sample) [2]
  - besides, using different metrics during optimization and evaluation does not have an effect on performance



[1] Images source: Hewamalage, Hansika, Klaus Ackermann, and Christoph Bergmeir. "Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices." arXiv preprint arXiv:2203.10716 (2022).

[2] Koutsandreas, Diamantis, et al. "On the selection of forecasting accuracy measures." Journal of the Operational Research Society 73.5 (2022): 937-954.

Beyond average performance

- Check bias and variance
  - Over-predictions or under-predictions have different effects on the business side
    - e.g. in retail, systematic under-predictions lead to stock-outs

- Two models with the same average performance may have different variability
  - Check error distribution
  - Large errors may be prohibitive

Beyond accuracy

“One does not eat [...] forecasts”

- It’s not how accurate but how valuable your model is
  - Forecasting performance is often used as proxy for value... not the same thing!
- **Utility evaluation**: how will you use the forecasts? how does a forecast impact the decision? what are the costs of a given error?
  - Difficult to decouple these questions from the domain
- Well understood in the financial domain, where accuracy is often not correlated with profits

Taleb, Nassim Nicholas. "On the statistical differences between binary forecasts and real-world payoffs." International Journal of Forecasting 36.4 (2020): 1228-1240.

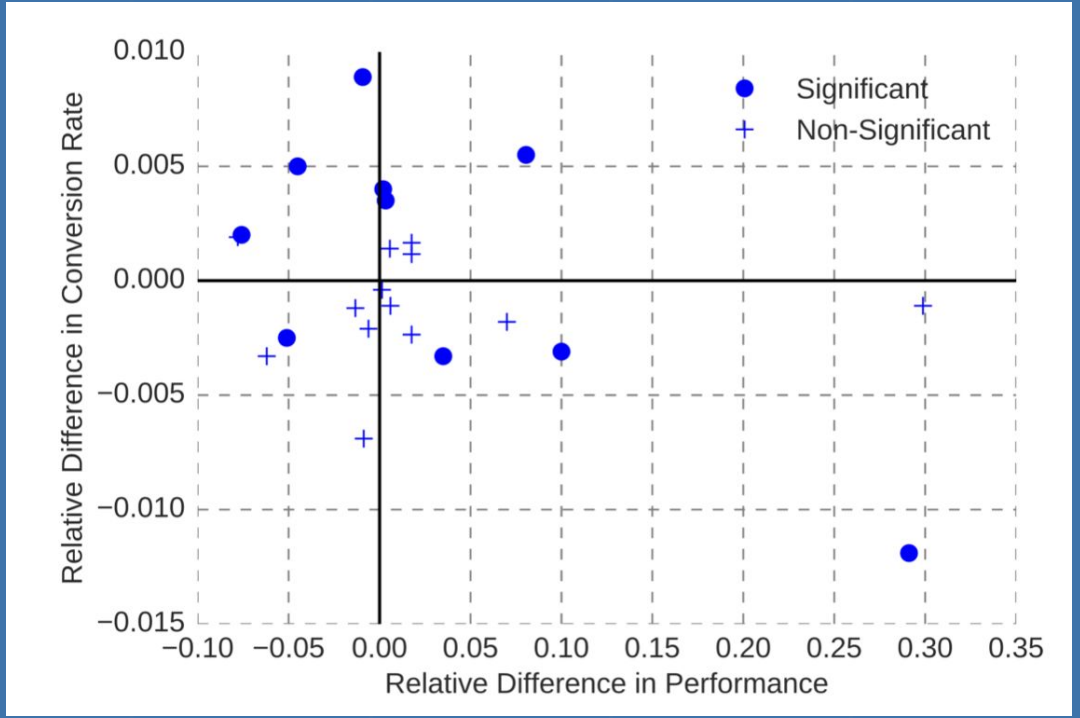
Evaluation

Booking.com example

Performance is seen as just a health check

- value is estimated through RCT and specific business metrics like conversion, customer service tickets or cancellations.

Figure: Gains in performance metrics do not necessarily translate into added value



Bernardi, Lucas, Themistoklis Mavridis, and Pablo Estevez. "150 successful machine learning models: 6 lessons learned at booking. com." Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019.

Key Takeaways

In this presentation, we explored:

- The basics of time series analysis and forecasting
- How to do supervised learning with time series
- The main transformations and pre-processing steps before training models
    - How to deal with…
        - trends
        - (multiple) seasonality
        - non-constant variance
- How to properly apply cross-validation with time series and evaluate forecasting models

Thanks!
Any questions?

You can find me at:

https://bio.link/vcerq

cerqueira.vitormanuel@gmail.com