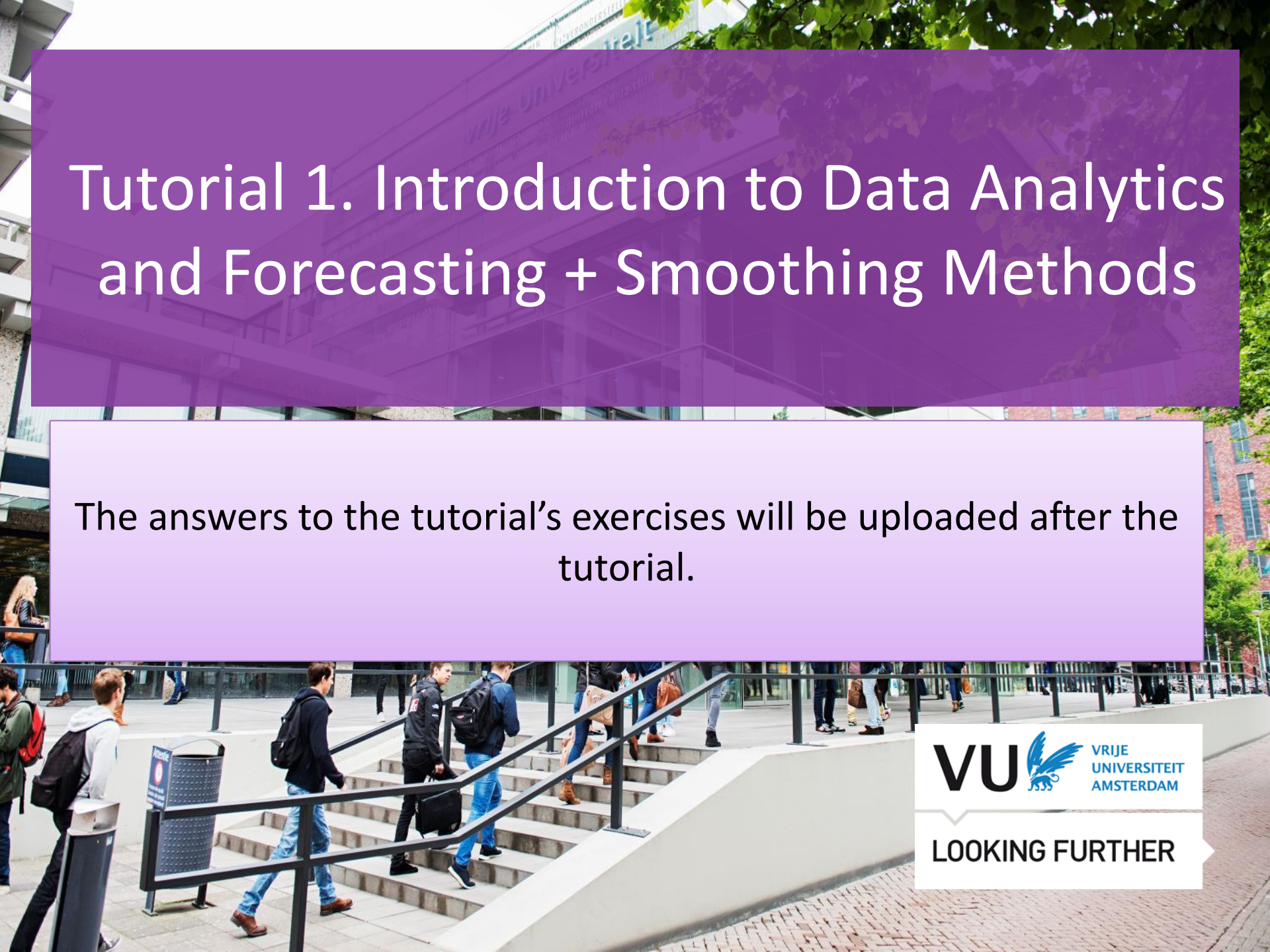
Tutorial 1. Introduction to Data Analytics and Forecasting + Smoothing Methods

The answers to the tutorial's exercises will be uploaded after the tutorial.



VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

Practical Challenges of Using Data Analytics

- When applying data analytics methods/forecasting models in practice, numerous practical challenges arise.

- In the tutorial sessions, we'll explore some of the most common challenges and learn strategies to address them effectively.

Challenges 1. Non-dominated Forecasting Models

- Assume that you have fitted various forecasting models to a time series and assessed their accuracy using multiple performance metrics (e.g., ME, MSE, RMSE).

- The evaluation reveals that no single model consistently outperforms the others across all metrics.

Forecasting Model	ME	MAE	MSE
Model 1	5	89	2500
Model 2	15	54	1200
Model 3	110	65	150

Challenges 1. Non-dominated Forecasting Models

- Assume that you have fitted various forecasting models to a time series and assessed their accuracy using multiple performance metrics (e.g., ME, MSE, RMSE).

- The evaluation reveals that no single model consistently outperforms the others across all metrics.

Forecasting Model	ME	MAE	MSE
Model 1	**5**	89	2500
Model 2	15	54	1200
Model 3	110	65	150

Challenges 1. Non-dominated Forecasting Models

- Assume that you have fitted various forecasting models to a time series and assessed their accuracy using multiple performance metrics (e.g., ME, MSE, RMSE).

- The evaluation reveals that no single model consistently outperforms the others across all metrics.

Forecasting Model	ME	MAE	MSE
Model 1	**5**	89	2500
Model 2	15	**54**	1200
Model 3	110	65	150

Challenges 1. Non-dominated Forecasting Models

- Assume that you have fitted various forecasting models to a time series and assessed their accuracy using multiple performance metrics (e.g., ME, MSE, RMSE).

- The evaluation reveals that no single model consistently outperforms the others across all metrics.

Forecasting Model	ME	MAE	MSE
Model 1	**5**	89	2500
Model 2	15	**54**	1200
Model 3	110	65	**150**

Challenges 1. Non-dominated Forecasting Models

How to select the best model for forecasting future values when you have multiple non-dominated forecasting models?

Exercise 1

We have a dataset containing time series data on the demand for two products in a supermarket over a specified period (SupermarketSales.xlsx).

1. Visualize the time series for the entire duration.
2. Prepare and clean the data as necessary to ensure accuracy for analysis.
3. Split the data into training and validation sets. What proportion of the data would be reasonable for each set based on the characteristics of the time series?
4. Implement the naïve forecast, seasonal naïve forecast, and moving average methods on both time series. Determine which method yields the best performance.
5. Use the best model and forecast demand for the upcoming year.

Exercise 2

We have a dataset containing time series data on the monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, between 1995 and 2001 (SouvenirSales.xlsx). Back in 2001, the store wanted to use the data to forecast sales for the next 12 months (year 2002). They hired an analyst to generate forecasts. The analyst first partitioned the data into training and validation periods, with the validation period containing the last 12 months of data (year 2001). She then fit a forecasting model to sales, using the training period. Partition the data into the training and validation periods as explained above.

Exercise 2

a) Why was the data partitioned?

b) Why did the analyst choice a 12-month validation period?

c) What is the naive forecast for the validation period?

d) The analyst found a forecasting model that gives satisfactory performance on the validation set. What must she do to use the forecasting model for generating forecasts for year 2002?

e) Compute the RMSE and MAPE for the naive forecasts using R.

f) Plot a histogram of the forecast errors that result from the naive forecasts (for the validation period) using R. Plot also a time plot for the naive forecasts and the actual sales numbers in the validation period. What can you say about the behaviour of the naive forecasts?

Exercise 3

Two different models were fit to the same time series. The first 100 time periods were used for the training period and the last 12 periods were treated as a validation period. Assume that both models make sense practically and fit the data reasonably well. Below are the RMSE values for each of the models:

	Training Period	Validation Period
Model A	543	690
Model B	669	675

a) Which model appears more useful for retrospectively describing the different components of this time series? Why?

b) Which model appears to be more useful for forecasting purposes? Why?

Useful Functions

Task	Function
Read data	read.csv() read_excel()
Create time series object	ts()
Partition data	window()
Plot data	plot(), lines()
Naïve and seasonal naïve forecasts	naive(), snaive()
Evaluate model's performance	accuracy()