KU LEUVEN

Strava data in the context of local knowledge-driven sports and physical activity policy

SPORT. VLAANDEREN

STATISTIEK VLAANDEREN

Maarten Haselaars, Tuur Deboutte and Ewout Samyn

Daily supervisor: Manon Reusens

Promotor: Professor Bart Baesens

Contact Statistiek Vlaanderen: Kasimir Putseys

Introduction

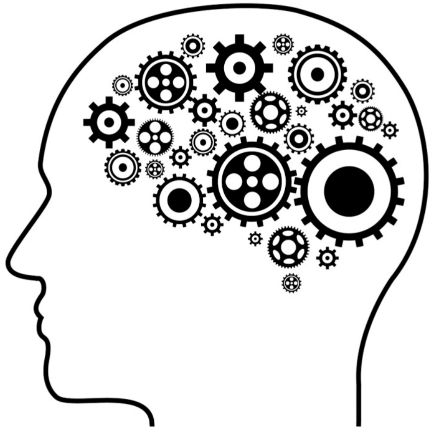
- Exercise & being outdoors is great for well-being
- Flemish government: Every city centre safely and easily reachable by bicycle
- To gain insights, bike counters were installed at various sites in Flanders

Goal of the thesis

1. Help Flemish government **gain insights** by combining data from government-issued bike counters and Strava Metro platform
2. Modest contribution to existing literature about **performance of different forecasting models for time series** (SARIMAX, Prophet & LSTM)
3. Comparison of **forecasting** performance **in practice**

Predictive Modelling

1. Statistical model
    - ARIMA
        - SARIMA(X)
2. Machine-Learning model
    - Prophet
3. Deep learning model
    - LSTM model



Models are implemented using existing Python libraries.
(e.g. Keras, statsmodels/SARIMA(X), SciKit-learn)

Data sources

- Bike counters: Agentschap Wegen en Verkeer
- Strava Metro
- Weather data: Meteostat2 API

Seasonality

- STL decomposition
  - Obvious pattern every week

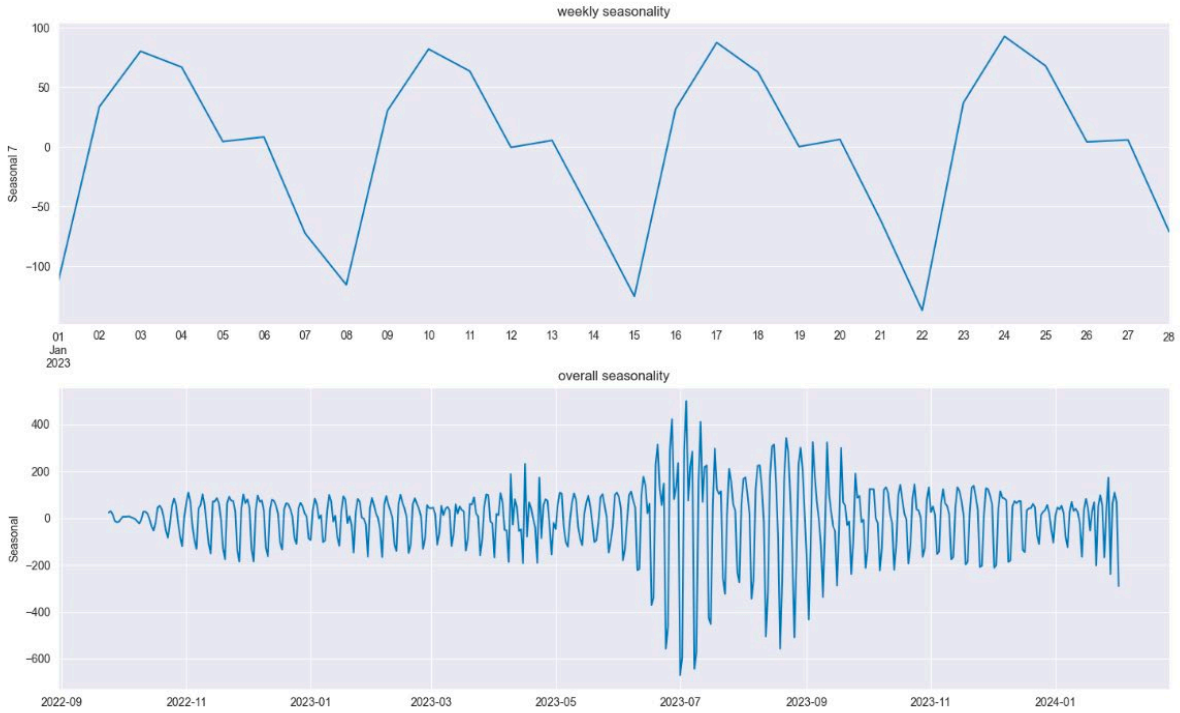
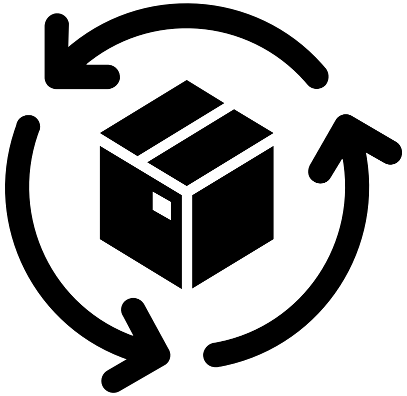


Figure 8: The y-axis shows seasonality, the x-axis shows time

Data preparation

- SARIMAX:
  - Stationarity
  - Seasonality
- LSTM:
  - Scaling
  - Sliding window (lookback = 7days)
  - Input-, LSTM- and dense layer
- Prophet:
  - Renaming columns



Data preparation

- Crossvalidation:
  - Training set (80%), validation set (10%), test set (10%)
  - Comparing predicted values to actual values of target variable in test set
  - RMSE: relatively simple and interpretable

- Hyperparameter tuning:
  - For all 3 models

Results

Table 3: Average test set RMSE over all bike counters

Variables/Model	Prophet	SARIMAX	LSTM
No Exogenous variables	290,44	189,57	156,68
Strava only	185,01	176,65	133,66
Weather only	218,73	178,60	159,87
All exogenous variables	159,76	160,71	138,30

- LSTM is the best model overall → lowest RMSE value in each row
- Adding more data sources improves model performance significantly

Results

- Prophet model benefits more from extra regressors as opposed to SARIMAX and LSTM
  - Test set RMSE for **Prophet model** <u>with all exogenous regressors</u> **decreases with 44,99%**. (as opposed to 15,11% for SARIMAX model and 11,73% for LSTM)
  - Prophet model is more robustly able to handle extra data from extra regressors, **excels with more data**

Actionable insights for government

Our models are able to generate forecasts **up to 7 days in advance** within 95% confidence interval of actual values

- Extending our research: use model to **generate forecasts** for areas that currently **do not have bicycle counter**
- Facilitate increasing number of cyclists (e.g. widen bike paths)
- More attention for areas that have a large expected number of cyclists in the near future. (e.g. increase number of cycling highways, more police/medical staff presence)

Conclusions

- **Forecasting** 7 days in advance **is possible** and to a very high degree of accuracy
- **LSTM model outperformed Prophet and SARIMAX**, based on the RMSE metric.
    - Best average RMSE value was 16,8% and 16,3% lower compared to SARIMAX and Prophet respectively