Vorverarbeitung und Merkmalsextraktion

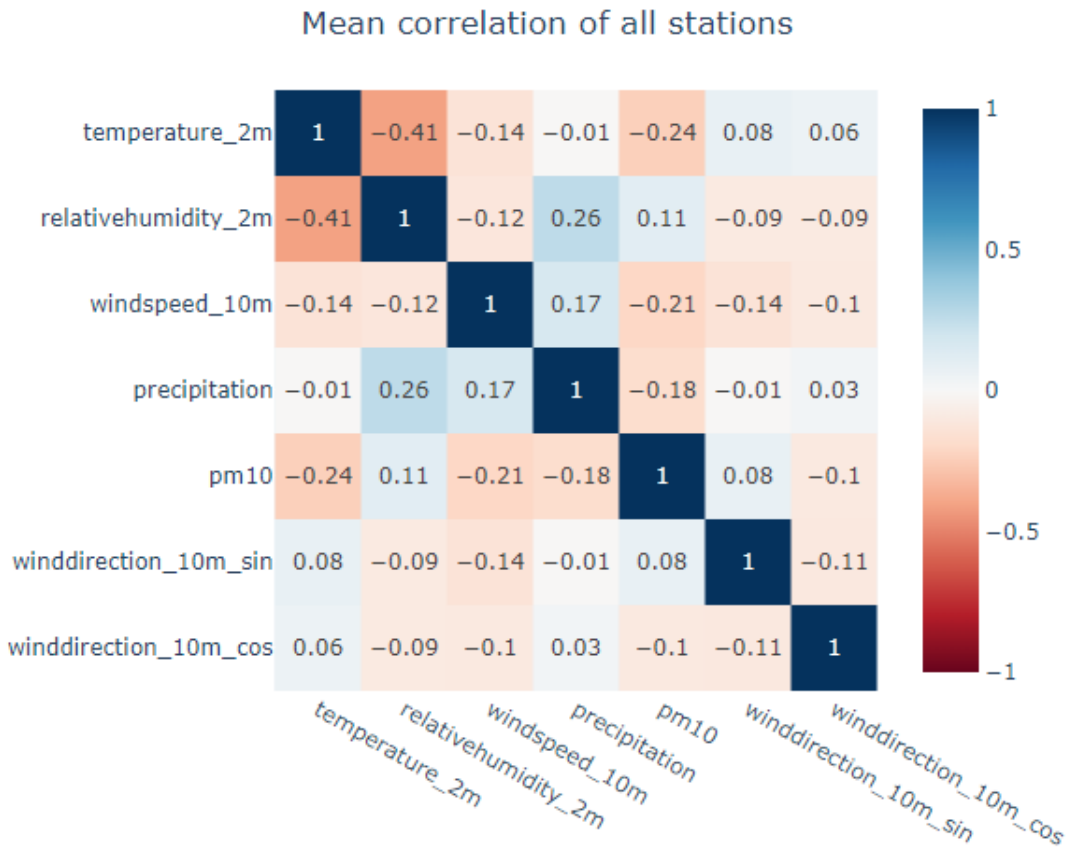
Meilenstein 2

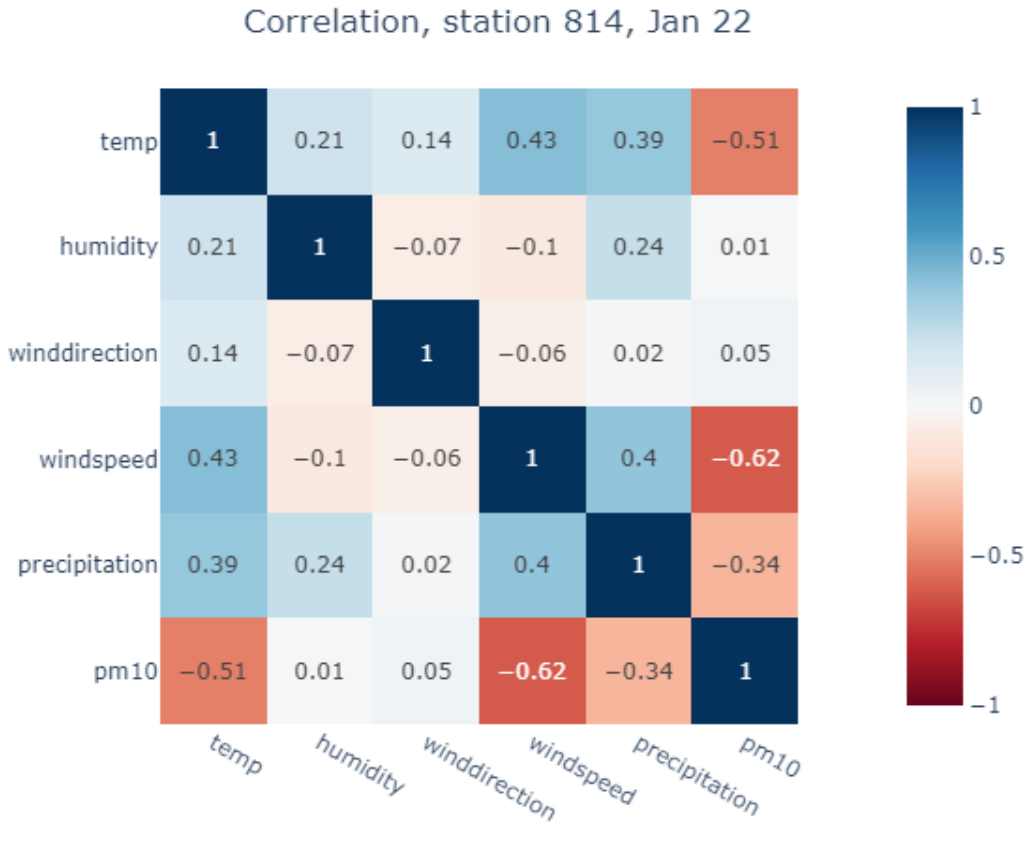
Inhalt

- Visualisierung Datenanalyse
- Vorverarbeitung
  - Interpolation PM_{10}
- Feature Engineering
  - Zeit
  - Windrichtung
  - Normalisierung
- Window Generator und ML-Vorbereitungen

Vorverarbeitung

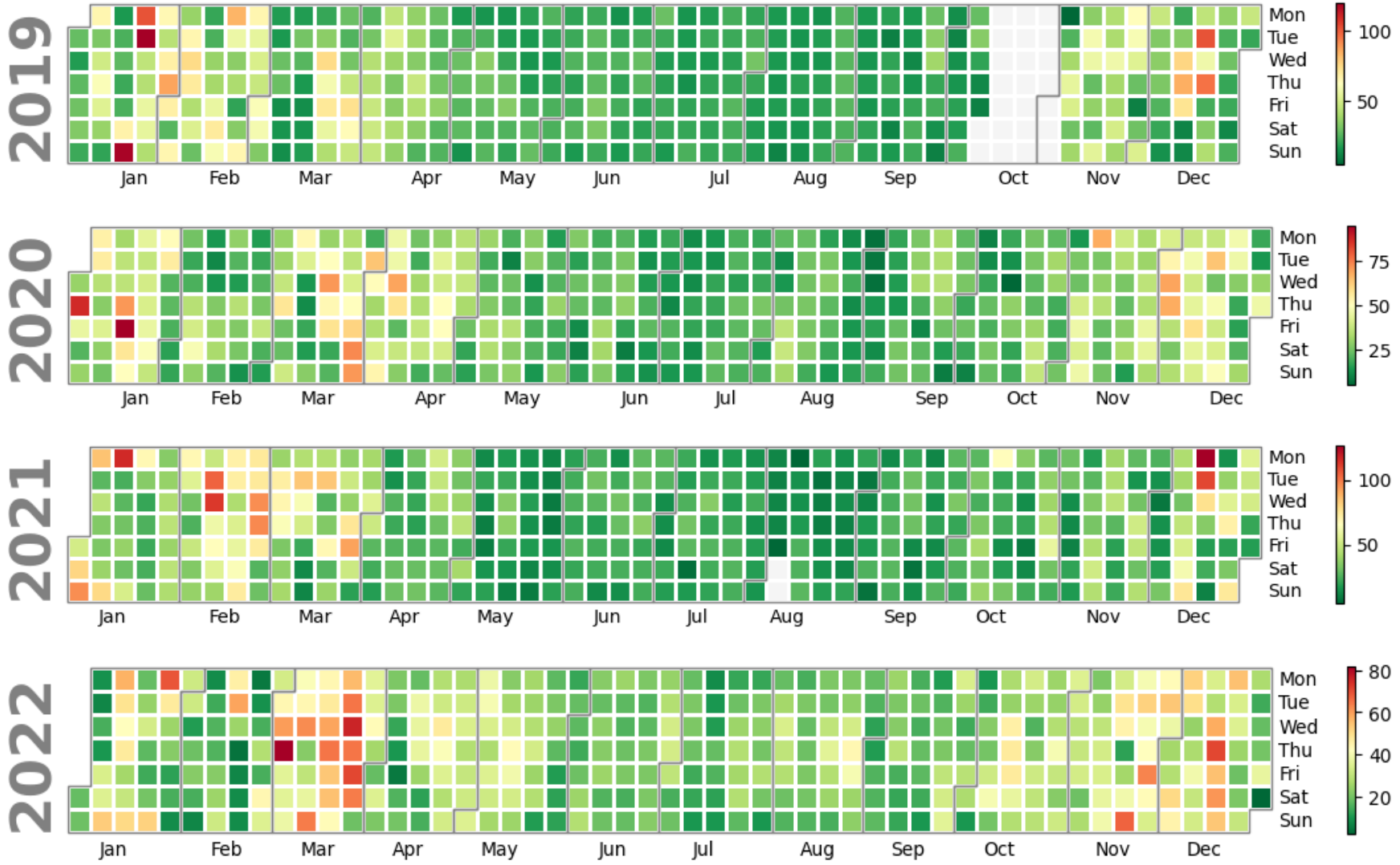
Visualisierung Datenanalyse





Vorverarbeitung

Visualisierung Datenanalyse

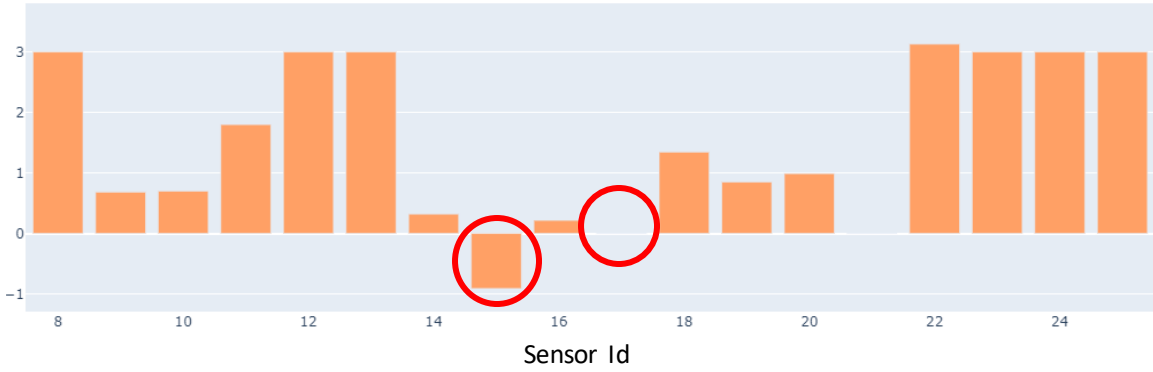


Station 814

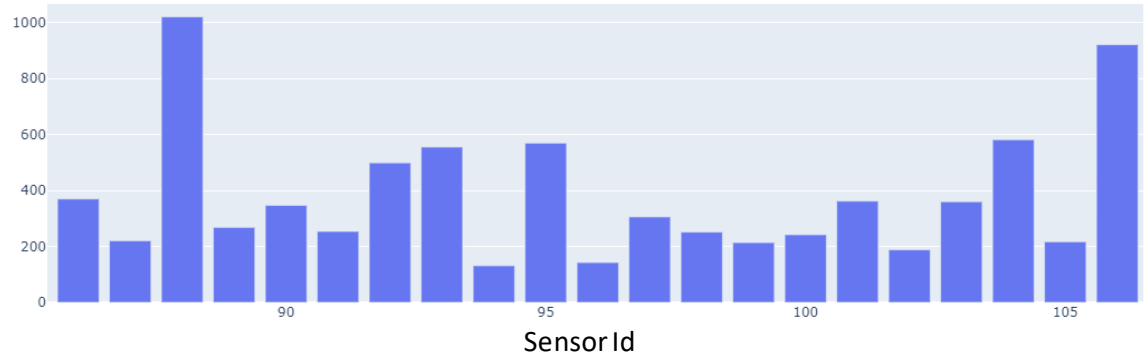
Interpolation PM_{10}

- Maximalwerte aller Stationen valide
  - Outlier Detection nicht notwendig
- Inkorrekte Werte
  - Fehlende Werte → NaN
  - Negative PM_{10} Werte
- Interpolation inkorrekter Werte für bis zu 5 aufeinander folgende Stunden
- Entfernen der restlichen Zeiträume

Minimalwerte der Datasets einiger Stationen:

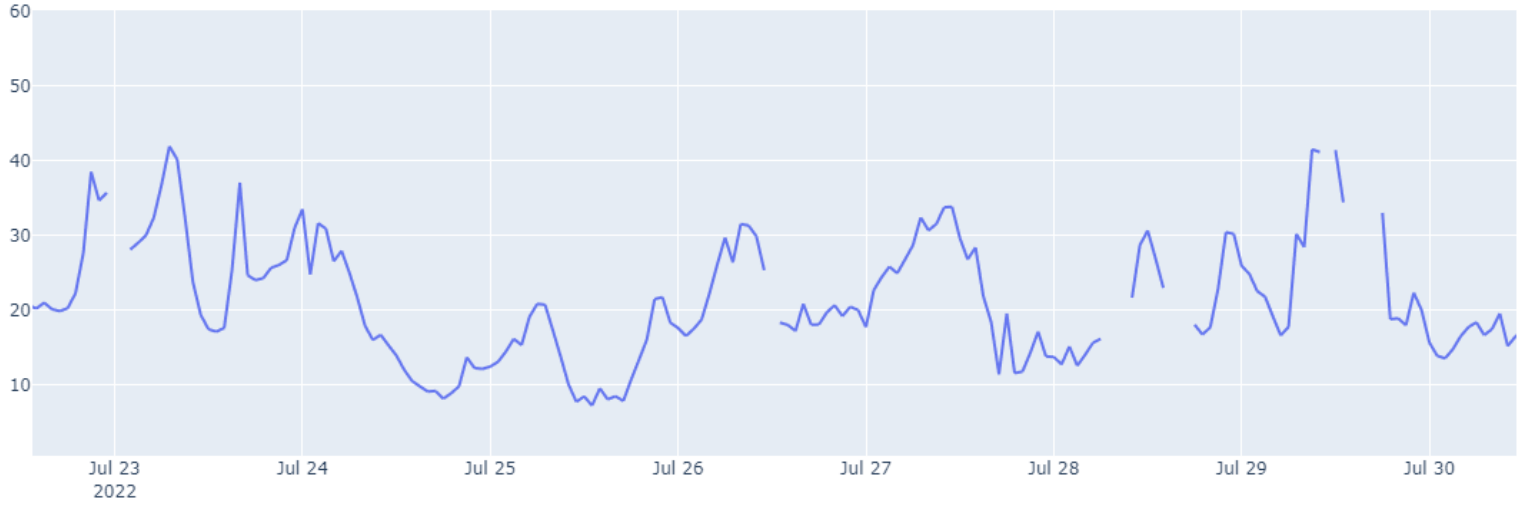


Maximalwerte der Datasets einiger Stationen:



Interpolation PM_{10}

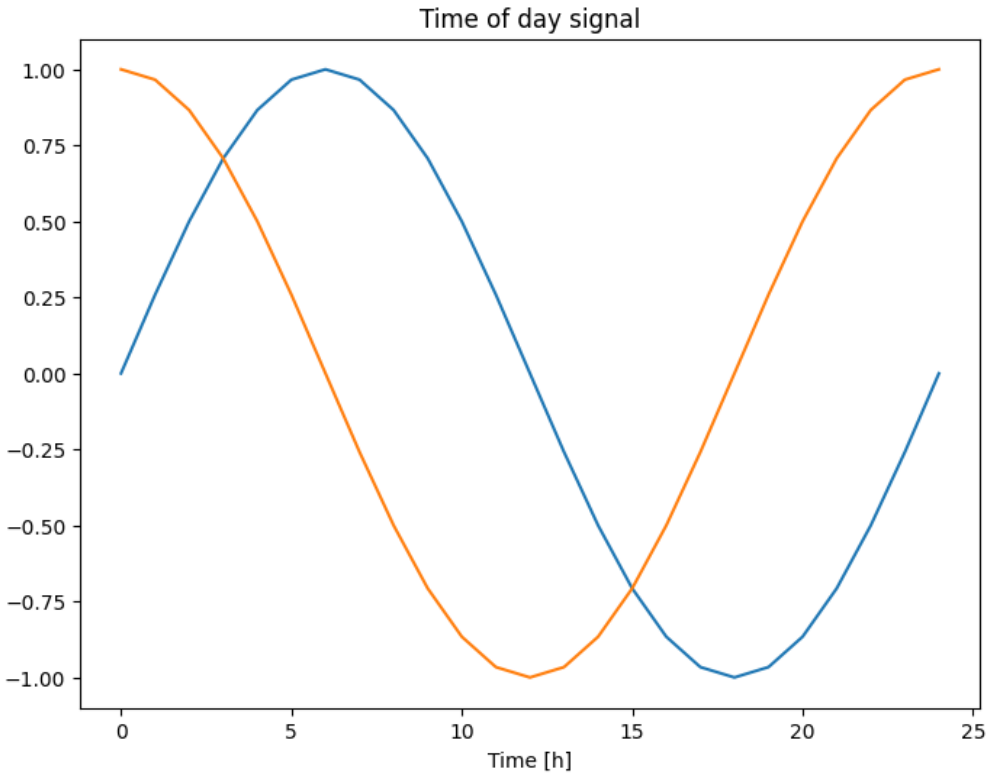
Original Daten für PM_{10}



Interpolierte Daten für PM_{10}

Zeit

- Zeit periodisch angeben statt absolut
- Perioden für Tag und Jahr
- Für Modell besser verwendbar
- Periode durch Sin und Cos



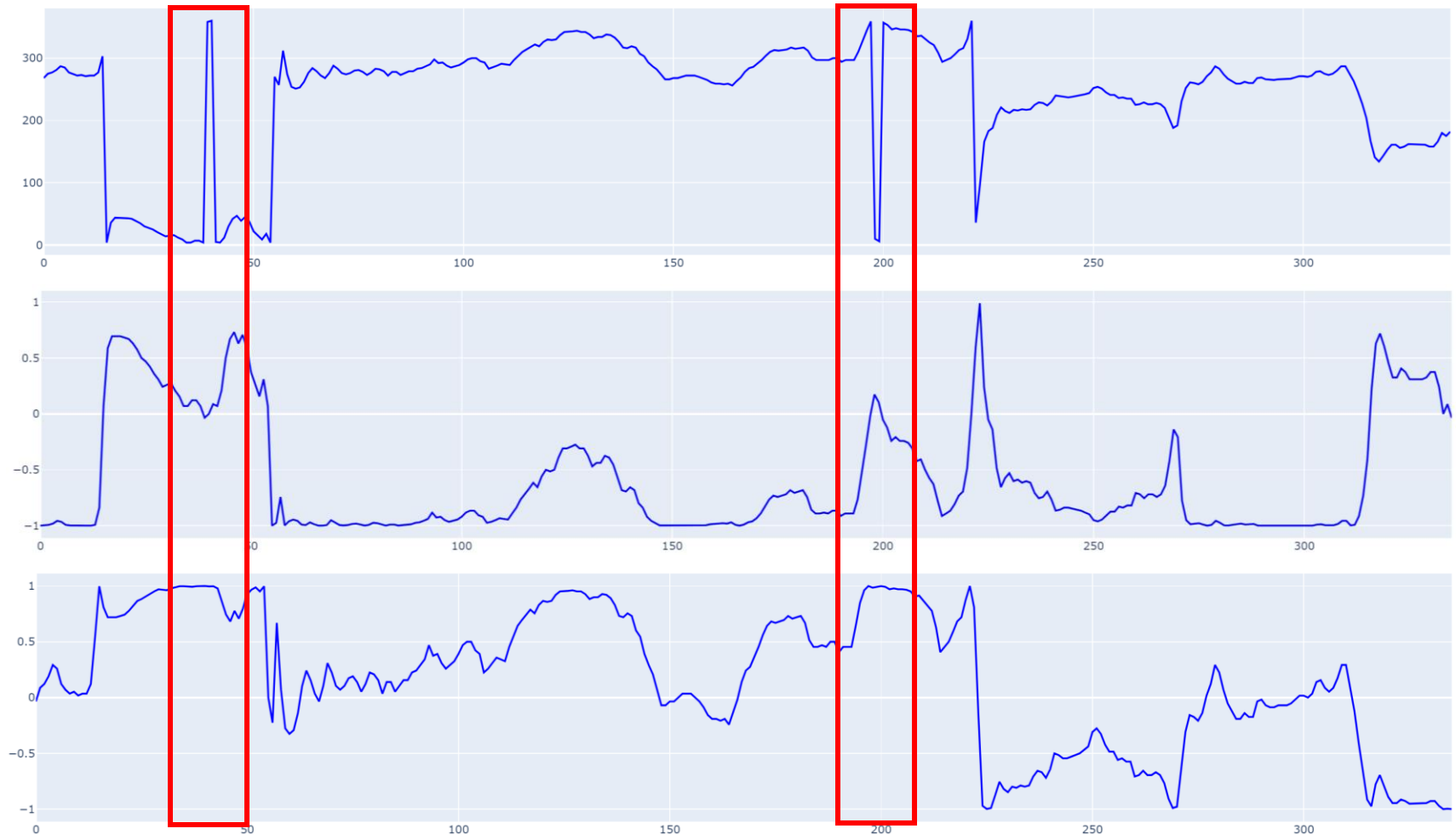
Feature Engineering

Windrichtung

Windrichtung in Grad

Windrichtung in Radiant (Sinus)

Windrichtung in Radiant (Cos)



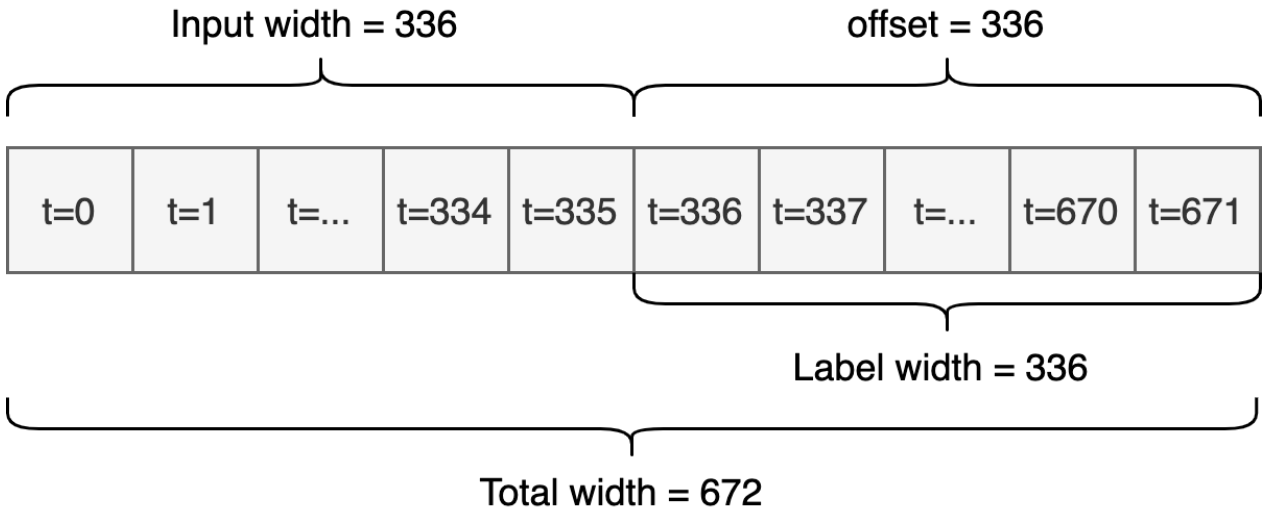
14 Tage Fenster

Normalisierung

- Normalisierung der numerischen Werte mittels Standard Skalierung
  - (value - mean) / sqrt(var)
- Skalierung durch Normalization Layer im neuronalen Netz
- Mittelwert und Varianz der Daten werden während des Trainings gelernt

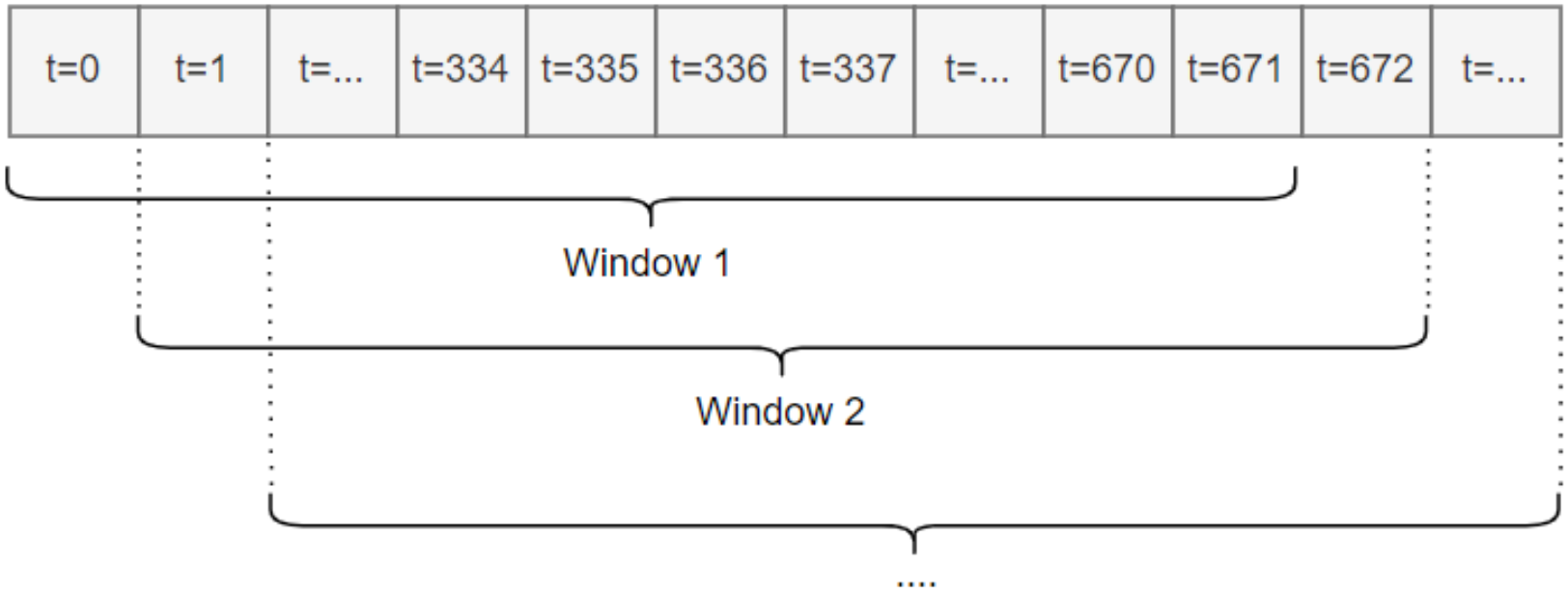
Window Generator

- Erstellen von 28-tägigen Trainingwindows aus den Daten
  - Generierung von N Windows mit vollständigen Feature Vektoren
    - ➢ Keine Zeitsprünge vorhanden
- Länge des Windows variabel einstellbar
  - Definition von Input-, Offset- und Labelbreite der Daten

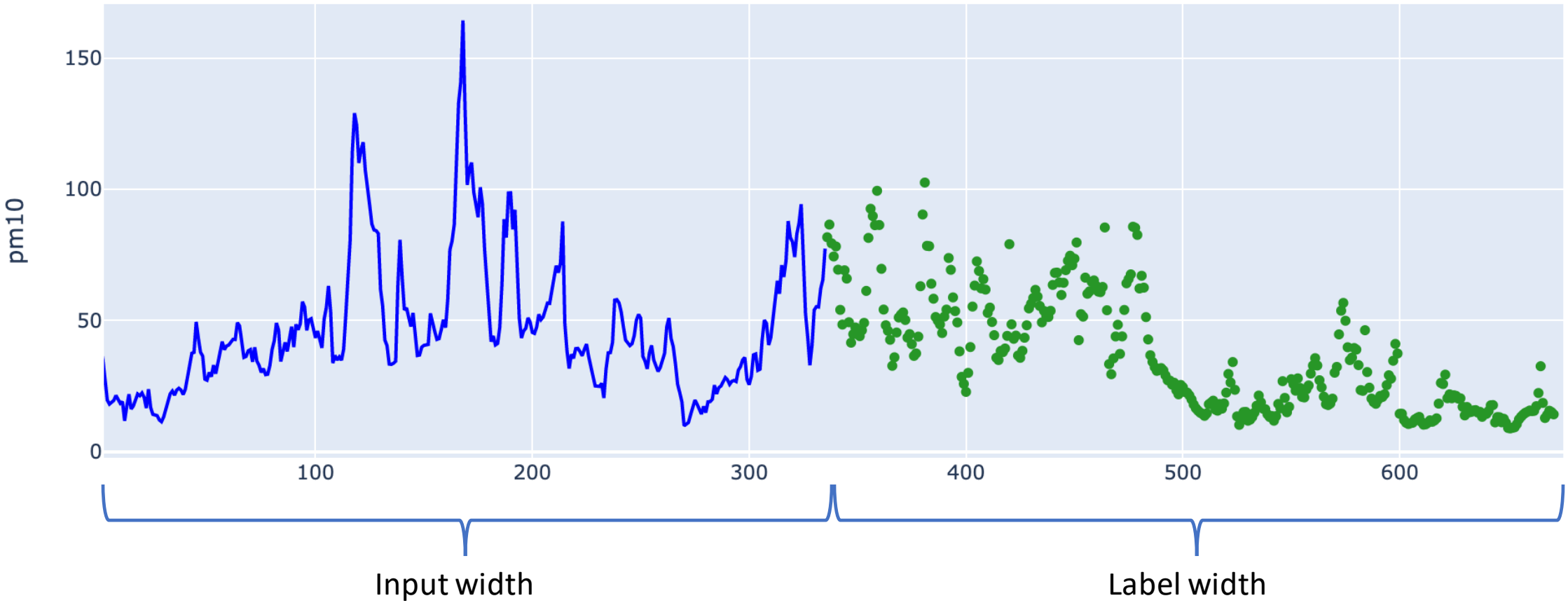


Window Generator

- Erstellung des Windows in stündlichen Schritten
  - Überprüfung auf zeitliche Lücken



Window Generator



Feature Vector

- Temperatur (°C)
- Luftfeuchtigkeit (%)
- Windgeschwindigkeit (m/s)
- Niederschlag (l/m)
- Windrichtung (sin & cos)
- Tag (sin & cos)
- Jahr (sin & cos)
- PM_{10}

➢ 11 Features

ML-Vorbereitungen

- Generierung von Tensorflow Datasets aus allen Windows
  - Trainset: 70 %
  - Validationset: 10 %
  - Testset: 20 %
  - ➢ Zufälliges Shuffeln + Batchen der Sets
- Format der Sets
  - Inputs shape (batch, time, features): (None, 336, 11)
  - Labels shape (batch, time, features): (None, 336, 1)