My document

Table of contents

1 Hourly Air Quality and Weather dataset

This dataset combines hourly data from the Umweltbundesamt on Air Quality readings and from the Deutsche Wetterdienst on weather data.

Table 1: Air Quality components with unit of measure

Component	Description	Unit
PM10	Particulate Matter smaller than 10 µm	µg/m
PM25	Particulate Matter smaller than 2.5 µm	µg/m
CO	Carbon monoxide	µg/m
SO2	Sulfur dioxide	µg/m
NO	Nitrogen oxide	µg/m
NO2	Nitrogen dioxide	µg/m
O3	Ozone	µg/m

Table 2: Weather variables description and unit of measure

Short Name	Subcategory	Beschreibung	Description	unit
TT_TU	air_temperature	Lufttemperatur in 2m Höhe	temperature at 2m height	°C

Short Name	Subcategory	Beschreibung	Description	unit
RF_TU	air_temperature	relative Feuchte	relative humidity	%
V_N	cloudiness	Bedeckungsgrad aller Wolken	degree of coverage of all clouds	1/8
TT	dew_point	Lufttemperatur	temperature	°C
TD	dew_point	Taupunkttemperatur	dew point temperature	°C
F	wind	mittlere Windgeschwindigkeit	wind velocity	m/s
D	wind	mittlere Windrichtung	wind direction	°
FX_911	extreme_wind	maximale Windgeschwindigkeit	hourly maximum wind velocity	m/s
R1	precipitation	stündliche Niederschlagshöhe	hourly precipitation	mm
WRTR	precipitation	WR-Niederschlagsform	type of precipitation	
P	pressure	Luftdruck	pressure	hPa
SD_SO	sun	stündliche Sonnenscheindauer	hourly sunshine duration	min
VP_STD	moisture	berechnete Stundenwerte des Dampfdruckes	average vapor pressure	hPa
P_STD	moisture	mittlerer Druck	average pressure	hPa
RF_STD	moisture	mittlere relative Feuchte	average relative humidity	%
ABSF_STD	moisture	berechnete Stundenwerte der absoluten Feuchte	Calculated absolute humidity	g/m³
TF_STD	moisture	berechnete Stundenwerte der Feuchttemperatur	Calculated humid temperature	°C
TT_STD	moisture	Lufttemperatur in 2m Höhe	temperature at 2m height	°C

1.1 Matching Process

1.1.1 Available Data:

Air Quality data:

From Umweltbundesamt on request. The received data ranges from 2000-2022 and is made up of hourly readings from 831 stations all over Germany.

1.1.1.1 Weather Data:

Downloaded from Deutscher Wetterdienst. The data is made up of hourly measurements, partly calculated from measured variables, from 1306 stations all over Germany.

1.1.1.2 External

- Socio-economic data and data on buldings and area from Regionalatlas
- Daily Covid-policy index on a Landkreis-Level from http://corona-daten-deutschland.de
- Downloading geographical information on Air Quality surroundings from Open Streetmap (OSM) by using the package "osmdata"

1.2 Structure

1. Merge all DWD data for all stations and for all years from 2000-2022 in scirpt "MergeAQ_hourly_1.R"
2. Download and merge all UBA data for all stations and for all years from 2000-2022 in Script "Download_DWD_hourly.R".
3. Matching of closest weather station by geographical distance in script "MergeAQWD_hourly_nearestElement.R"
    - process is seperately conducted for each weather component (temperature, precipitation etc.) since not all DWD stations measure all components.
    - Finding closest stations using the function nn2 from the R package RANN and adding a constraint of a radius of 0.4 decimal degrees (native to the function) which in Germany corresponds to approx. 44km.
4. Merge Air Quality data and weather data using the matching from previous step. The result is a dataset containing 178 Air Quality stations.

5. Data cleaning

   In the script "Data_cleaning.R" some cleaning is conducted. The result is the cleaned merged dataset AQWD_merge_2000-2023_p.rds. In the script "Dataprep_AddFeatures.R" this dataset is then further processed to prepare it for modelling use.

6. Adding more features

   More data from the external sources is added by merging it with the geographical location of the Air Quality station.

Table 3: Additional variables added

Variable name	Explanation	Source
forest	Forest density in "Gemeinde" (county) in 2021	Regionatlas
Area_resi_traffic	Density of residential and traffic area in "Gemeinde" (county) in 2021	Regionatlas
area_resi	Density of residential area in "Gemeinde" (county) in 2021	Regionatlas
GDPpercap_lk	GDP per capita [EUR] in "Landkreis" (district) in 2020	Regionatlas
Avg_age	Average age in "Gemeinde" (county) in 2021	Regionatlas
Earnings	Total amount of income per taxpayer [thousand EUR] in "Gemeinde" (county) in 2018	Regionatlas
householdsize_lk	Average householdsize in "Landkreis" (district) in 2011	Regionatlas
cars_lk	Number of passenger cars per 1,000 inhabitants on 01.01.2022 in "Landkreis" (district)	Regionatlas
disp_income_lk	Average disposable income per inhbitant in "Landkreis" (district) in 2020	Regionatlas
flatsperhouse_lk	Number of flats per house in "Landkreis" (district) in 2011	Regionatlas
areaperflat_lk	Average area per flat in "Landkreis" (district) in 2011	Regionatlas

Also, some variables are computed based on existent variables, like the Heating Degree Hour or time variables like Day of the Week.

7. The script "Data_Manipulation.R prepares the dataset for modelling.
    1. Outlier detection and removal (threshold of 500 $\mu g/m^3$ for PM10 and PM2.5).
    2. Imputing of missing values (linearly imputing per station, maximum of imputation of missing values is 24 hours) of each component (only Air Quality)
    3. Using linear regression models for each weather variable missing values are imputed. For this the function regressionImp from the package VIM is used.

The result is the dataset AQWD_hourly_model_final_all.rds