AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

FACULTY OF GEOLOGY, GEOPHYSICS AND ENVIRONMENTAL PROTECTION

Department of Geoinformatics and Applied Computer Science

Diploma Thesis

Forecasting air pollution using meteorological data
Prognozowanie zanieczyszczeń powietrza z wykorzystaniem danych meteorologicznych

Author: *Dawid Makowski*
Field of study: Geoinformatics
Supervisor: *dr. Tomasz Danek*

Kraków, 2024

Contents

Introduction

1. Preliminary information

During thesis creation, OpenAI ChatGPT was used, its use was limited to summarizing scientific articles to find conclusions or information more quickly on a given issue and to help in searching for sources.

2. Air pollutions

Air pollution is any air contamination by substances that are harmful to health or dangerous for other reasons, regardless of their physical form (Dz.U. 2004 Nr 29, poz. 255).

The most common pollutant is **dust pollution**, containing various chemical compounds. Its occurrence is related to, among others, the combustion processes of solid and liquid fuels. First, it negatively affects the respiratory system, and indirectly, dust also affects the rest of the body. This pollutant is divided into three groups: **PM_1**, **$PM_{2.5}$**, **PM_{10}**, the numbers in their names indicate the maximum particle size for a given dust in micrometers. The smaller the size of the dust, the more dangerous it is for humans because it is easier for it to penetrate the bloodstream (www.epa.gov, 2024).

Heavy metals and gases also appear in smog. These include **carbon monoxide**, the source of which is most often poorly installed or faulty gas stoves (www.malopolska.uw.gov.pl, 2024). **Sulfur oxides** also appear in the air and are emitted into the atmosphere both by natural processes (volcanic eruptions, forest fires) and as a result of human activity in urbanized areas; even short exposure to sulfur oxides can cause significant breathing difficulties (Ćwik, 2017). Smog also includes **nitrogen oxides**, which are particularly dangerous; their toxicity is much greater than carbon monoxide or sulfur dioxide. The compounds arise especially because of car exhaust fumes and toxins emitted by industrial plants entering the atmosphere (airly.org, 2024). **Tropospheric ozone** may also appear in the air, which is a secondary pollution resulting from the reaction of other chemical compounds (S. Shelton, G. Liyanage, ..., 2022). It is a strong oxidant, so when it enters the human respiratory tract, it causes irritation and breathing discomfort (WHO, 2006).

3. Air Quality Index

To inform the public about the current level of air pollution, various government agencies develop air quality indexes. When the index is high, people are encouraged to limit outdoor physical activity or even go outside altogether. Different countries have their own air quality indicators, corresponding to different national standards. In Poland, a popular one is **PIJP (Polish Air Quality Index)** developed by the GIOŚ (Chief Inspectorate of Environmental Protection) (powietrze.gios.gov.pl, 2024). The index value is determined based on the values of table *1.1.*

Tab. 1.1. Classification of pollutants according to the PIJP (powietrze.gios.gov.pl, 2024)

Quality name	Hourly pollution concentration [µg/m³]						
	PM_{10}	$PM_{2.5}$	O_3	NO_2	SO_2	C_6H_6	CO [mg/m³]
Good	0 - 20	0 - 13	0 - 70	0 - 40	0 - 50	0 - 6	0 - 3
Sufficient	20 - 50	13 - 35	70 - 120	40 - 100	50 - 100	6 - 11	3 - 7
Moderate	50 - 80	35 - 55	120 - 150	100 - 150	100 - 200	11 - 16	7 - 11
Bad	80 - 110	55 - 75	150 - 180	150 - 200	200 - 350	16 - 21	11 - 15
Highly bad	110 - 150	75 - 110	180 - 240	200 - 400	350 - 500	21 - 51	15 - 21
Extremely bad	> 150	> 110	> 240	> 400	> 500	> 51	> 21
No index	Air Quality Index is not determined due to the lack of measurement of pollution						

In Europe, the most popular index is **EAQI**, introduced in 2017, it provides information on the current situation in terms of air quality based on measurements from over 2,000 monitoring stations and is based on 5 key pollutants that harm human health and the environment: suspended particulate matter, tropospheric ozone, nitrogen dioxide and sulfur dioxide. Hourly concentrations are included in the index, and for the 24-hour average, these are data from the preceding 24 hours of measurements (airindex.eea.europa.eu, 2024).

4. Methods of measuring dust pollution

To measure air pollution in Poland, two methods of dust measurement are used. One of them is the gravimetric method, which uses the so-called dust collectors, fourteen disposable filters are placed in the collector every two weeks and the device changes them automatically every 24 hours. In laboratories, filters are weighed before and after the exposure period, and dust concentrations are calculated from the differences in mass. The method is fully accurate, but the time needed to obtain results is approximately 3 weeks. The second method is the automatic method, which uses automatic meters that have certificates confirming their equivalence with the reference method. These sensors measure dust concentrations on an ongoing basis, which enables the results of these measurements to be displayed on-line (powietrze.gios.gov.pl, 2024).

5. Causes of smog and its diversity in Poland

Map in *Fig. 1.1.* presents the average annual air pollution in Poland with PM_{10} dust. The highest concentrations were recorded in the Silesian, Lesser Poland and Łódź voivodeships, while the best air is breathed by residents of the Pomeranian and West Pomeranian voivodeships.

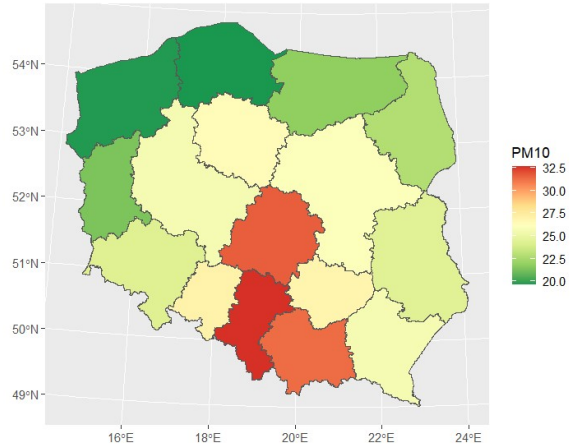

Fig. 1.1. Average PM_{10} level in every voivodeship in 2021 based on GIOŚ data (powietrze.gios.gov.pl)

Possible non-meteorological factors that have the greatest impact on increasing pollution in each location are population density, the number of individual buildings and the length of roads in the vicinity of the measuring device. These are factors related to the presence of emissions - including low emissions (single-family buildings), moreover, these factors are highly variable in space, they can be truly diverse even within the same cities. However, factors that limit air pollution may be the area of green areas and the topographic position index (TPI), i.e. the degree of location in a valley or on a mountain ridge (airly.org, 2024).

The main source responsible for the level of pollution in the Lesser Poland Voivodeship is emissions from the municipal and domestic sector, which is responsible for approximately 90% of dust emissions and 16% of NO_x emissions. Transport is responsible for approximately: 4% of dust emissions and 44% of NO_x emissions, car traffic also causes secondary dust removal from the road surface. Industrial emissions in the region generate lesser amounts of dust by approximately 2%, but they are a crucial factor in the emission of gases into the air - they are responsible for 29% of NO_x emissions. Other sources, such as agriculture (cultivation and breeding), forests and fires, are responsible for the following emissions: 11% of NO_x, 9% of PM_{10} and 1% of $PM_{2.5}$ (powietrze.malopolska.pl, 2024).

Additionally, in larger cities, pollution is influenced by the development of air corridors, which allow pollutants to be removed from the city and introduced fresh air into it (M. Dziekciarz, M. Foremniak). In the case of Krakow, factors also include inflow emissions, i.e. the movement of pollutants from neighboring municipalities not covered by the anti-smog resolution, and unfavorable geographical location, because the city is located in a river valley and is surrounded by hills on three sides, the movement of air masses around it is limited (airly.org, styczeń 2024).

6. Counteracting the problem of pollution

In order to reduce pollution, local governments can introduce various activities, such as anti-smog education, subsidizing public transport, promoting renewable energy or increasing the amount of urban greenery (airly.org, 2024).

For example, in the Lesser Poland Voivodeship there is an Air Protection Program (Uchwała Nr LXXV/1102/23), that limits the use of solid fuel boilers. In Krakow, since 2019, a resolution (UCHWAŁA nr XVIII/243/16) has been in force completely banning the use of solid fuel stoves, which has significantly reduced the average pollution level in the city (www.krakow.pl, 2024). The city also runs many preventive programs, such as the *Thermal modernization program for single-family buildings*, the *STOP SMOG program* and many others, and also undertakes many investments to help with this problem, such as the expansion of tram lines, the transport system or bicycle paths, also when the average value of pollution PM_{10} dust in the city will exceed 100 $\mu g/m^3$, free public transport is being introduced (www.krakow.pl, 2024).

7. Objective of the thesis

The thesis objective is to collect data on air pollution and meteorology in the Lesser Poland Voivodeship, using as many physical point sensors as possible in its area. Then, based on the collected data, creation and implementation of optimal air pollution forecast models for a given point, which will use meteorological data as input variables.

Methodology

1. Analysis tools

The project used the **R** language, an interpreted programming language and an environment for statistical calculations and visualization of results. It has a rich community and support, giving access to many packages, tools, and educational materials (www.r-project.org, 2024). The entire project was made in the **RStudio** environment, which allows to build scripts in R and Python, automatically import data, create reports, etc. (posit.co, 2024)

Important packages used in the thesis are **Leaflet**, which allows to create basic, interactive maps and **Shiny**, responsible for the entire visualization. The **plotly** package allows to create interactive charts that can be enlarged and check the values of selected points. The **randomForest** and **neuralnet** packages are responsible for providing forecast algorithms, i.e., neural network and random forest, based on them, later, forecast models are created. The **spatstat**, **sf**, **sp** and **gridlayout** packages were used to work with shapefile data or administrative units. The **raster** and **automap** packages were essential in creating a raster map of a given factor in the application. The rest of the packages were used for single analytical tasks or for aesthetic purposes.

The API download service of the Airly and OpenMeteo portals was used to download pollution and meteorology data. The data obtained from the Airly website was not always complete, especially the meteorology, and there was a need to supplement the information with further meteorological data. For this purpose, the OpenMeteo website was used, which provides extensive meteorological information for every point on Earth in its API. The following code fragment, in *Fig. 2.1*, contains calls that retrieve data from both services.

```
link_OpenMeteo <- paste(
        "https://api.open-meteo.com/v1/forecast?",
        "latitude=",Punkty$latitude[Punkty$id == i],
        "&longitude=",Punkty$longitude[Punkty$id == i],
        "&hourly=temperature_2m,relativehumidity_2m,dew_point_2m,rain,pressure_msl,
        surface_pressure,wind_speed_10m,wind_direction_10m,soil_temperature_0cm,
        soil_moisture_0_to_1cm&timezone=Europe%2FBerlin&past_days=1&forecast_days=2",
sep = "")

link_Airly <- paste(
        "https://airapi.airly.eu/v2/measurements/installation?installationId=", i, sep = "")
```

Fig. 2.1. Code snippet from RStudio containing calls downloading data from the Airly and OpenMeteo APIs

In the analysis there were also used data from the GIS Support Knowledge Base (gis-support.pl, 2024) based on the State Border Register (PRG), which contains files in the shp format (shapefile) containing data on the administrative boundaries of voivodeships, counties, communes and districts of Krakow. Measuring points are aggregated in the preprocessing phase, which means that each point is assigned a commune and a county.

2. Shiny web application as a visualization tool

Good visualization is crucial to understanding the problem, which is why a significant part of this project was devoted to this issue. For this purpose, a web application was created that allows for comprehensive interaction with data by manipulating various parameters and settings, as well as operating the pollution forecast according to the given settings.

The application consists of three tabs, the map service is placed in the "Interactive map" tab. It consists of a sidebar that contains options for selecting pollution or data aggregation, in addition, it is possible to choose color scaling locally to show differences within the received data or globally or scaling according to the Polish Air Quality Index. In the side menu there is a chart showing the average value of a given factor over time. In the case of wind, in addition to the speed, the azimuth of its direction is also displayed. *Fig. 3.3.* presents the tab described above.

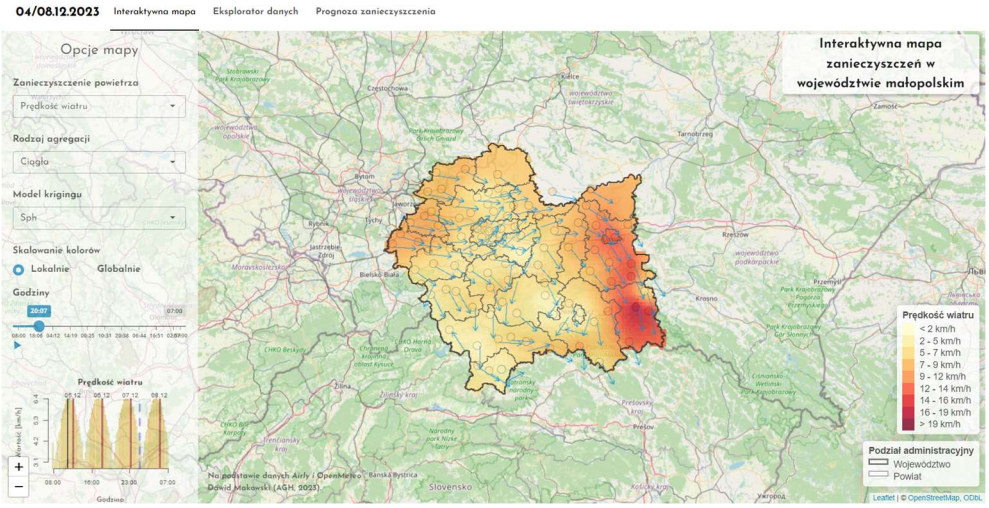

Fig. 3.3. The appearance of the "Interactive Map" tab of the Shiny application

The second tab is the "Data Explorer", shown in *Fig. 3.4.*, which compiles tabular data of points displayed on the map. When a given factor is selected in the first tab, e.g., $PM_{2.5}$, the table displays only this factor in a selected column. The table is divided into pages if there are more than 10 points, you can also filter the points according to your preferences, select the counties to which you want to limit the display of points or the minimum / maximum value of the selected factor for which the points will be displayed, it is also possible to verbally search assigning points by their address. In the "Zoom" column, by clicking on the crosshair, the user can be redirected to the first tab with a zoomed-in view of the selected point.

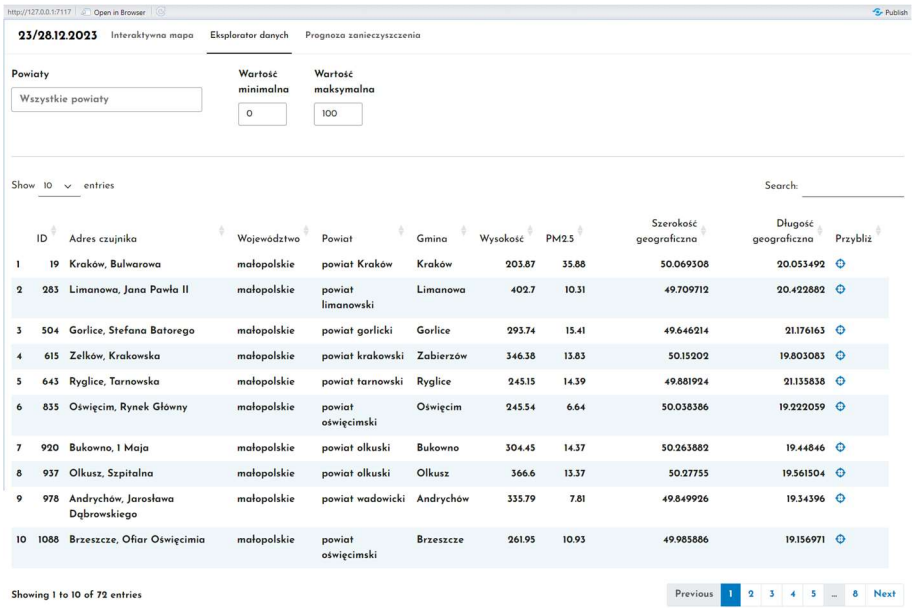

Fig. 3.4. The appearance of the "Data Explorer" tab of the Shiny application

The last tab is "Pollution Forecast", it allows you to make a live forecast based on the provided data. The sidebar allows you to adjust the parameters as you wish, you can use a random forest or neural network model, each of them has its own parameters to set. In addition, you can choose what kind of pollution will be forecast, as well as what meteorological data and what point the forecast should be based on. On the right side there is a map that shows measurement points and their distances to the point selected by the user, and the range of wind direction in the last hours is also displayed. *Fig. 3.5*. presents the tab described above. After clicking the "Run forecast" button, a forecast of the selected pollution will be made based on the selected factors. After completing the calculations, three interactive charts from the plotly package will appear, informing about the quality of the forecast. The top chart compares the data used in the forecast with the MSE and MAE quality metrics, the chart below directly compares the test data with the forecast data, displaying the Pearson correlation coefficient. The graph at the bottom informs, in the case of a neural network, about its structure, and in the case of a random forest, about the value of the mean square error depending on the number of trees.

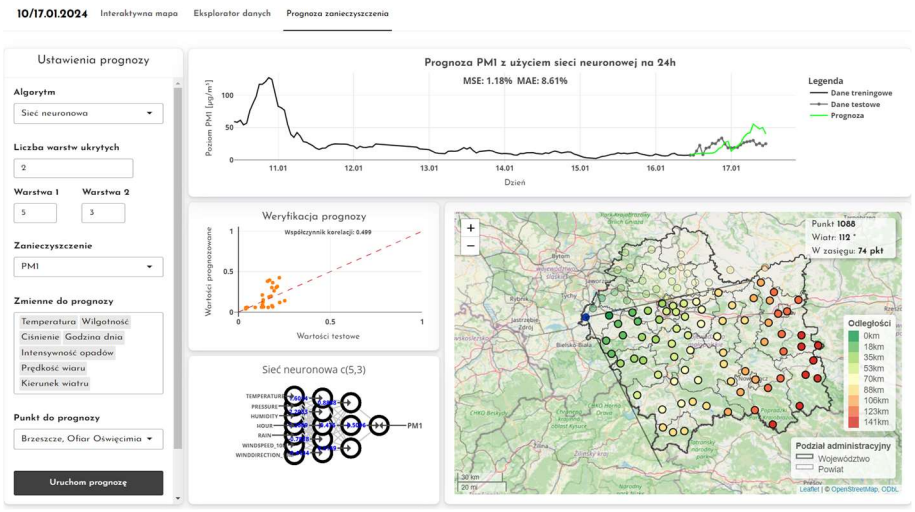

Fig. 3.5. The appearance of the "Pollution Forecast" tab of the Shiny application.

3. Machine learning models in pollution forecasting

In order to create an accurate pollution forecast, two models were used in the project.

The first of them are **artificial neural networks**, which are a mathematical model simulating the structure of biological networks. The basic element of every network is an artificial neuron, which is a mathematical function. At the input of the neuron, each input value is multiplied by the appropriate weight, then, inside the neuron, all weighted data are summed, and then at the output it passes through the activation function (A. Krenker, J. Bešter, A. Kos, 2011). There are several types of neural networks:

- **Feed-forward Artificial Neural Networks**, information must flow from input to output in only one direction with no back-loops.,
- **multi-layer perceptron**, capable of learning by back propagation, has many hidden layers,
- **recurrent neural networks**, taking input from previous sequences using feedback loops.

The second one is **random forest**, which is a machine learning model that combines the output of multiple decision trees to produce a single result. To obtain a result for a specific input object, the decision process starts from the root node and traverses the tree until it reaches the leaf containing the result. At each node, the path to follow depends on the feature value for the specific input object. The model performs well even in the case of missing data, but on the other hand, they cannot exceed the range of values of the target variable used in training (Roßbach, 2018).

To build a good prediction model for data of a specific type, you need to have an appropriate reference point to assess its quality and know whether the model with given parameters is good or whether its results are insufficiently precise. Metrics dedicated to regression problems will be used.

The most popular metric is **MSE** (mean squared error), it measures the average of squared errors, it is a risk function corresponding to the expected value of the squared loss after error, its disadvantage is that it is not very resistant to outliers, it may also lead to underestimation of the model if error values are less than 1. It is often used because its error decreases as the data set grows (Aishwarya, 2022). Its twin metric is **MSLE**, which measures the mean squared error of the logarithms of the differences between forecasts and actual data; the use of a logarithm causes larger errors to appear when the real variable is underestimated than when it is overestimated. (Saxena, 2019). Another metric is **MAE** (mean absolute error), it calculates the difference between each forecast and the actual value and then averages the absolute values of these differences, this metric treats errors in a linear way, which means it is not as sensitive to outliers as MSE (Hiregoudar, 2020). The coefficient of determination $\mathbf{R^2}$ is also used, which is a measure used to assess the degree of fit of the regression model to real data, it informs how well the regression model explains the variability of the dependent variable in relation to its mean; the closer to 1, the better the model fits the data. (Hiregoudar, 2020).

Research results

1. Selection of points

In order to conduct a complete analysis of pollution in the entire voivodeship, points were selected in such a way that they were not too close to each other, but also not too far away. Airly company has over 900 physical sensors in its database in the Lesser Poland Voivodeship, on the *Fig. 3.1.* their distribution is presented, along with the degree of proximity. However, a maximum of 100 sensors can be downloaded per day; additionally, it should be considered that not all sensors are fully operational and do not work at all in each period, so the selection had to be narrowed down to the best possible one hundred sensors.

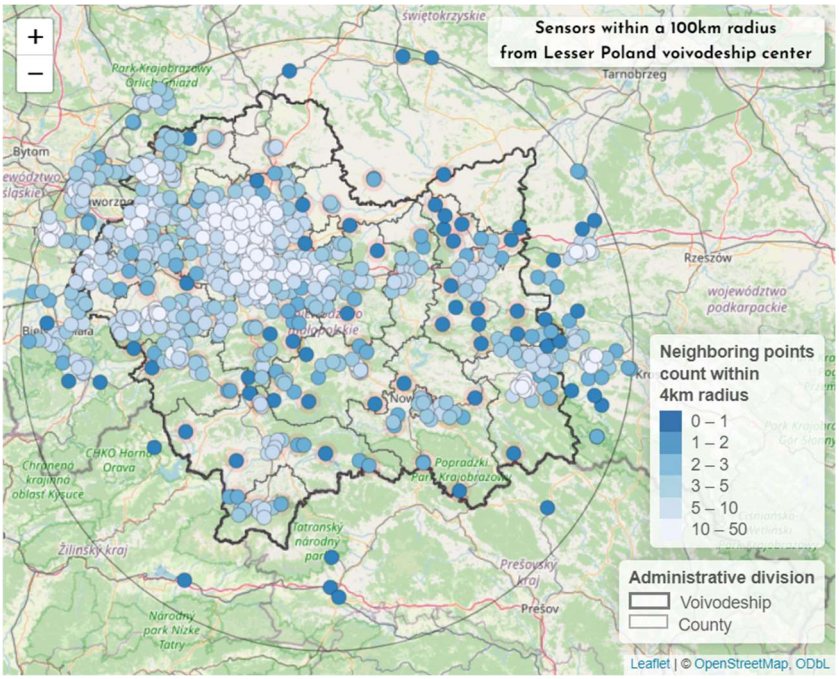

Fig. 3.1. Map of Airly sensors in the Lesser Poland Voivodeship with information about the number of neighboring points

When selecting the points for the first time, it was checked what percentage of non-empty measurements of PM2.5 or PM10 pollution the given points contained, then another set of 100 points was considered for the analysis and the procedure was repeated. Among the examined points, those that contained the most non-empty measurements were selected, while points that contained a slightly smaller but satisfactory number of observations and were in an area where, in proximity, there were no other points, were also selected.

2. Obtaining data using Airly and OpenMeteo API

The project required meteorological and air pollution data. Before the research was conducted, a set of historical data covering approximately 20 months in 2021-2022 was used, containing observations only for PM_1, $PM_{2.5}$ and PM_{10} dust along with basic meteorological data in the Krakow area. The data was used to build a visualization application, perform correlation analysis, and model tests. Below, on *Fig. 3.2.*, the location of the points covering this data is shown. The points do not cover the entire area of the voivodeship, but for early analysis and tests it is enough.

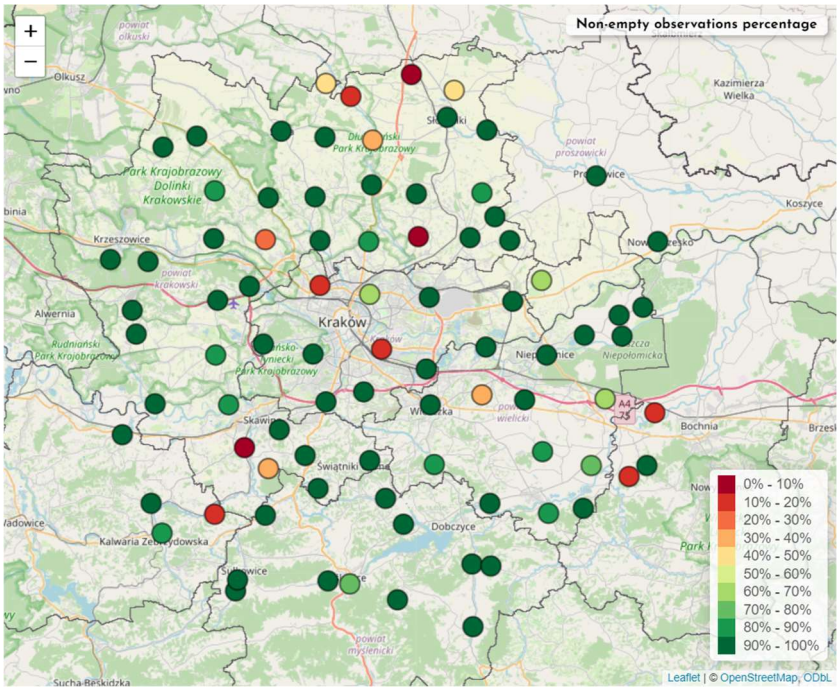

Fig. 3.2. Summary of the data used with information about the percentage of non-empty measurements (13 000 hours)

As part of the research, meteorology and air pollution data within the Lesser Poland Voivodeship were collected for over a month, using the Airly API and open-meteo.com. Airly provides data on dust, NO_2, SO_2, CO, O_3 pollution as well as temperature, humidity and air pressure reduced to sea level. The data was obtained from actual physical measurements of sensors located in the voivodeship, due to the daily limit of data downloading, it was limited to 100 points.

The selection of points prioritized measurements of $PM_{2.5}$ and PM_{10} dust, which is why they contain the most non-empty observations. It was not possible to collect a sizable number of measurements on other gases, so they will only serve as illustrative data. *Table 3.1.* presents how much information was collected in the period from December 5, 2023, to January 17, 2024. These data are very often incomplete, and the most non-empty information is available for suspended dust, meteorological data were later supplemented by OpenMeteo measurements.

Tab. 3.1. A table showing the completeness of data downloaded from Airly API.

FACTOR	NUMBER OF OBSERAVTIONS	PERCENTAGE OF NON-EMPTY OBSERVATIONS	PERCENTAGE OF TIME COVERAGE
PM25	62087	74.09%	100.00%
PM10	66810	79.73%	100.00%
NO2	4261	5.08%	88.31%
SO2	4974	5.94%	88.19%
CO	2121	2.53%	88.19%
O3	2149	2.56%	86.75%
PM1	54201	64.68%	88.31%
NO	796	0.95%	88.31%
PRESSURE AIRLY	53462	63.80%	88.31%
HUMIDITY AIRLY	52594	62.76%	88.31%
TEMPERATURE AIRLY	52594	62.76%	88.31%

3. Visualization and preliminary conclusions of research results

In the illustration Fig. 3.6. average meteorological data for the entire period, for each point, are presented. In some cases, certain regularities occur that may affect the level of air pollution. Pressure and temperature show similar relationships, they decrease towards the south, this is since in the south of the voivodeship the absolute height is higher. The wind speed is on average the highest in the north of the voivodeship, which may be due to greater windiness than in the mountain valleys, but the data set is not large enough to demonstrate this well enough. The variability of precipitation intensity in this case is the result of the fact that in a given period the precipitation was more intense in the west, only a long-term analysis would show certain constant dependencies, e.g. that precipitation is usually more intense in the mountains (Carpenter, 2018).

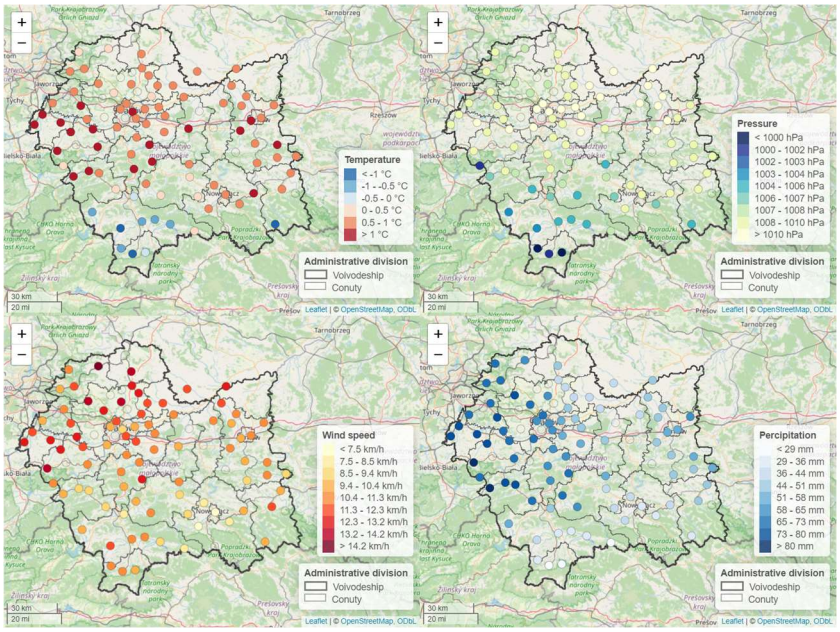

Fig. 3.6. Visualization of acquired meteorological data.

The map below (*Fig. 3.7.*) shows average dust concentration values. In the proximity of Kraków and Tarnów and the western part of the voivodeship, dust concentration is higher, while the southern part usually has lower dust concentration. A noteworthy observation is that larger towns located in deeper mountain valleys (Nowy Sącz, Nowy Targ, Sucha Beskidzka) have the highest pollution values, while neighboring sensors do not show such high values, which proves the difficulty of air flow in such an area (airly.org, 2024).

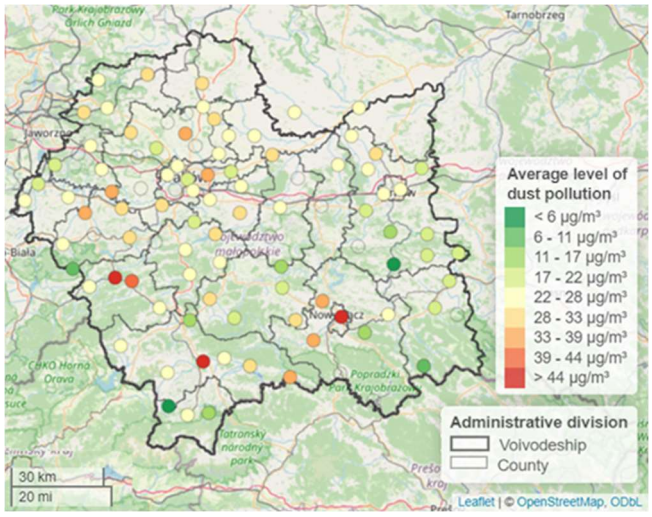

Fig. 3.7. Visualization of dust pollution

Airly also provides data on other aerosols, but they are not numerous enough to perform a significant geostatistical analysis, measurements are made at no more than five points.

4. Correlation study

In the next step, an analysis of the dependence of pollution on meteorological factors was conducted to select those that were most dependent on the level of pollution. It should be noted that some factors may have non-linear relationships, which should be remembered when comparing correlations and creating forecast models. The following relationship maps were analyzed only to find the overall relationship, i.e., which factors usually cause an increase or a decrease in the pollution value. So far, it is not possible to notice non-linear dependencies, which may be demonstrated by a more detailed analysis carried out later.

In Fig*. 3.8.* a correlation map between pollutants is presented. Dust pollutions correlate most strongly with each other because their composition is the same, they differ only in size. Other gases correlate to a lesser degree, carbon monoxide and nitrogen oxides show a strong positive correlation with dust, as does sulfur oxide, although slightly weaker. Ozone, on the other hand, correlates negatively with each pollutant, which may be since it is a secondary pollutant.

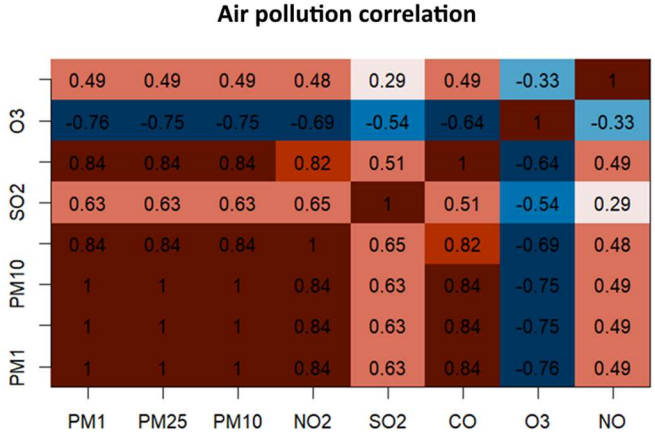

Fig. 3.8. Visualization of the relationship between air pollutants

Now that we know how pollutants correlate with each other, we can look at their correlations with meteorological variables, *Fig. 3.9.* presents a map of the relationships between them. Stronger correlations of pollutants with pressure and temperature can be seen. Although the precipitation and wind intensity variables do not correlate that well, non-linear dependencies may occur. Additionally, the analysis also included data such as dew point, soil temperature, etc., which are related to their twin factors and their correlations are similar, but weaker, because they directly concern the soil or ground level, therefore they are not as effective in correlation with pollution.

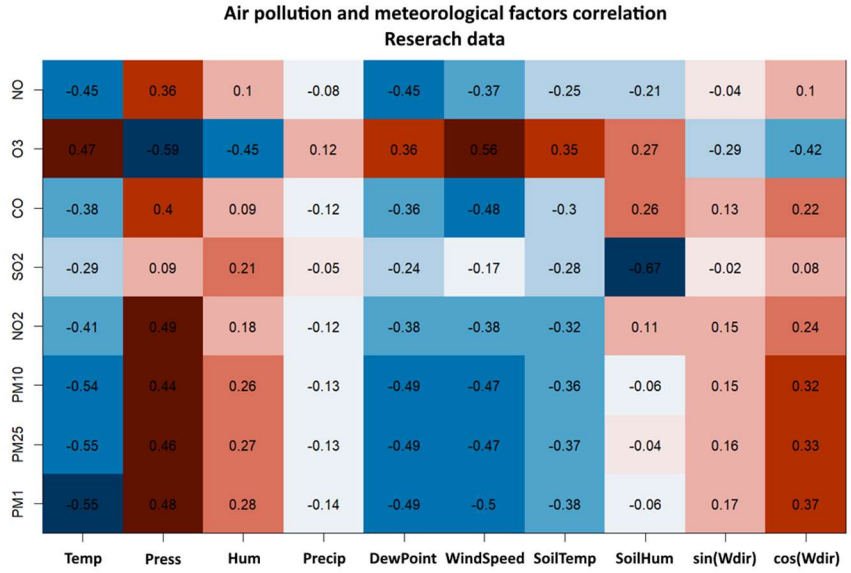

Fig. 3.9. Visualization of the relationship between air pollutants and meteorological factors

Graphs in *Fig. 3.10.* present the relationship between temperature and PM_{10} dust pollution. The negative correlation, especially in the Lesser Poland Voivodeship, is related to the heating season - the lower the temperature, the more people will burn stoves in their homes, and moreover, people are more likely to use cars instead of cycling or walking, which is the result of increased emissions of pollutants into the atmosphere. Additionally, in winter, there is a high chance of inversions occurring, which may block pollutants in the lower layers of the atmosphere (Elminir, 2005). It is also worth noting the "divergence" of points on the left and right sides of the charts, this is due to the fact that for each argument (in this case the temperature value), the arithmetic mean of the pollution values at a given temperature is taken, for the arguments extreme data, there are fewer values to average, therefore they are less representative and more exposed to deviations or anomalies.

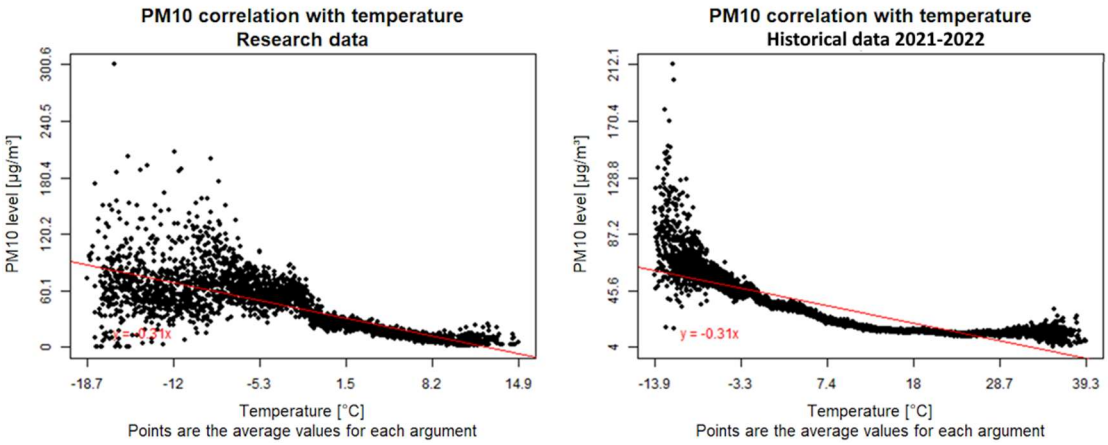

Fig. 3.10. Charts of the relationship between air temperature and PM_{10} dust pollution

As can be seen in *Fig. 3.11*, in the case of precipitation, there is also a negative correlation with pollution. This relationship is better visible in historical data, in the research data the linearity is strongly disturbed, this may be since the research period concerned December and January, where most of the precipitation is snow, which may not have the same impact on dust as rain. The negative correlation can be explained by the so-called atmospheric deposition, i.e. during rain, water collects pollutant particles and brings them to the earth's surface, thanks to which pollutants are removed from the air, the greater the rain intensity, the more effective this process is (Koch, D., J. Park, A. Del Genio, 2003).

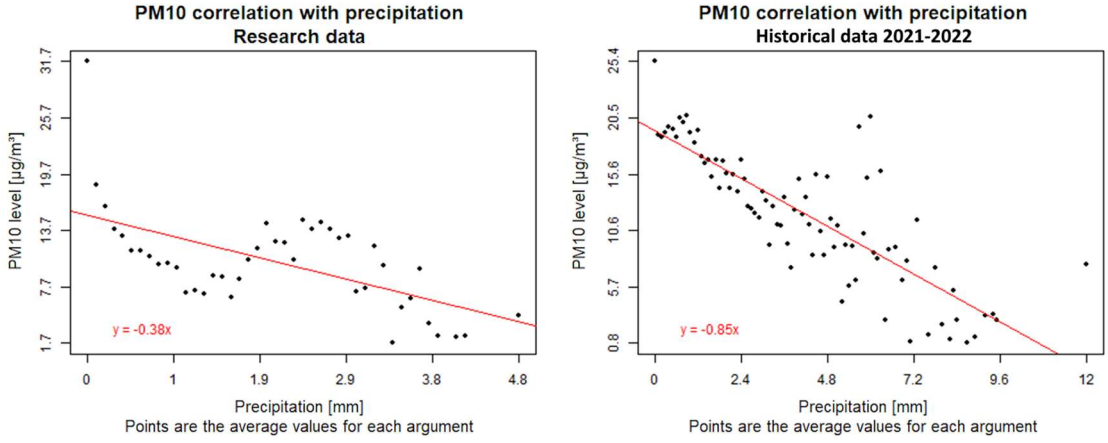

Fig. 3.11. Charts of the relationship between precipitation and PM_{10} dust pollution

Fig. 3.12. shows the correlation of PM_{10} dust with wind speed, which in both cases is strong and negative, i.e., the stronger the wind, the lower the pollution. Wind makes a significant contribution to reducing the amount of pollutants in the air by effectively dispersing pollutant particles, preventing them from accumulating in one place, while bringing fresh air from areas with lower concentrations (Elminir, 2005).

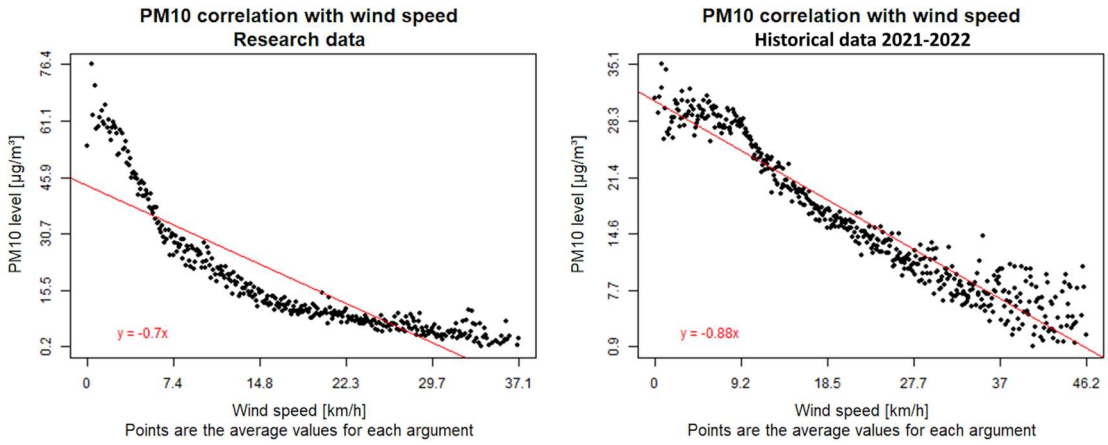

Fig. 3.12. Charts of the relationship between wind speed and PM_{10} dust pollution

In the case of pressure, there is a positive correlation with PM_{10} pollution, as shown in *Figure 3.13*. In the case of research data, the relationship is weaker than in the case of historical data and due to numerous linearity disturbances, it may even be invisible. This is due to the time interval being too short to highlight this relationship well, and other atmospheric factors may also influence the "flattening" this relationship. High pressure may favor high pollution due to the stability of atmospheric conditions, which usually occur at high pressure, this is most often associated with the lack of precipitation or weak wind movement, then the air is less susceptible to movement, which favors the accumulation of pollution in one place (Amos P.K. Tai, 2010).

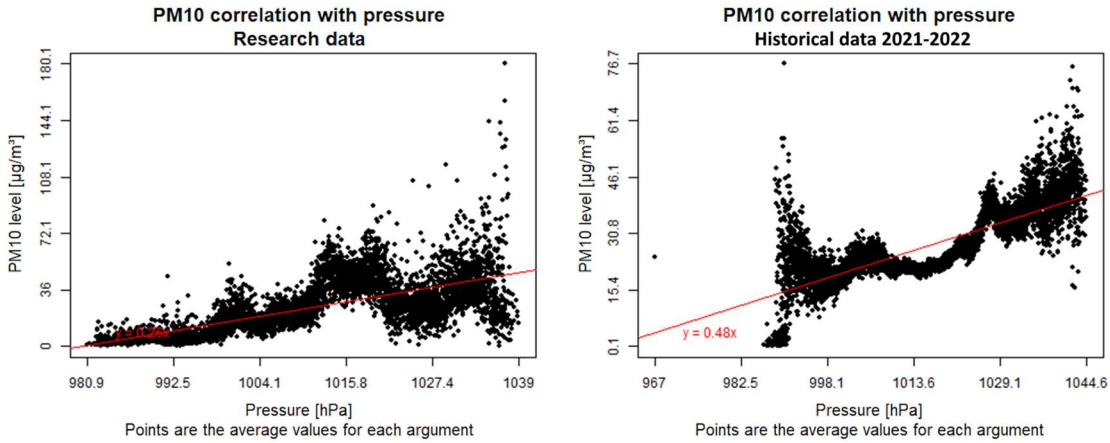

Fig. 3.13. Charts of the relationship between air pressure and PM_{10} dust pollution

In the case of humidity, which relationship is shown in *Fig. 3.14*., there is a certain positive correlation, but it is small, and for research data it is even negligible. The positive correlation may be since its elevated level facilitates the occurrence of certain chemical processes in the atmosphere, which lead to the formation of new pollutants and may also lead to more stable atmospheric conditions. On the other hand, high humidity facilitates the dissolution of some gases, such as nitrogen or sulfur oxides, but the graphs show that the effects of positive correlation predominate. (Elminir, 2005). The relationship is so small in the case of research data that it may not be profitable to use it in forecasts.

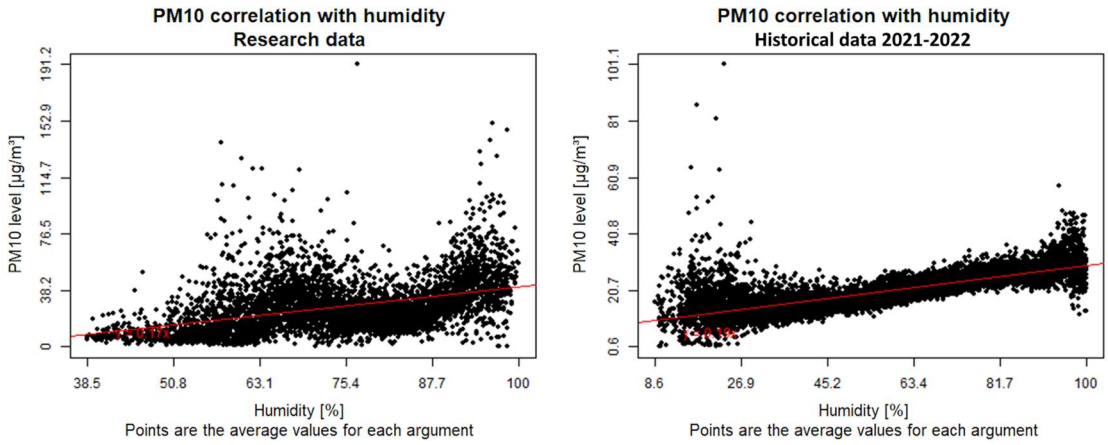

Fig. 3.14. Charts of the relationship between air humidity and PM_{10} dust pollution

Pollution is also closely correlated with the time of day, as shown in *Fig. 3.15*., although it is non-linear, this relationship is clearly visible for all pollutants. This relationship is related to the human rhythm of the day and night, it can be seen that both graphs combine high pollution values in the evening hours (5-8 p.m.), which is caused by increased street traffic when people return home from work, which particularly translates into increase in nitrogen oxide pollution, this is also the time when home furnaces are most often lit, emitting pollutants into the atmosphere. For all pollutants except ozone, the values increase in the morning, which is related to increased car traffic when people leave their homes for work or school, but the pollution values are not as high as in the evening. In turn, around 12 o'clock the lowest concentration values appear, which is related to the lack of street traffic. Ozone has minima and maxima contrary to other pollutants, which is caused by its secondary presence in the air, during the day, as a result of photochemical reactions with NO and O_3, NO_2 is formed, and at night, as a result of photolysis, it is regenerated again (S. Shelton, G. Liyanage, ..., 2022).

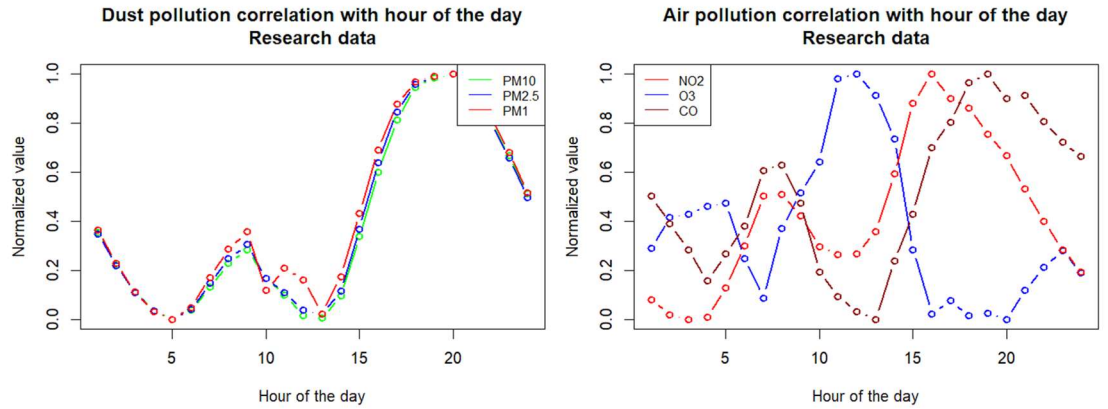

Fig. 3.15. Charts of the relationship between hour of day and PM_1, $PM_{2.5}$, PM_{10} dust

Similarly for the day of the week, from the charts in *Fig. 3.16*. it can be concluded that there are non-linear relationships. In the case of all pollutants, except ozone, the maximum occurs on Wednesday, the lowest on Friday, and a slight increase on the weekend. One of the reasons may be the differentiation of industrial activity or the intensity of services - in some plants it may be increased on days in the middle of the week and reduced on Fridays, when work may be completed faster before the

weekend. The reason may also be transport traffic, where the traffic, especially when it comes to commuting to work, will be higher, while on Friday, some people may finish work earlier, which causes the transport traffic to slow down during (He, 2023).

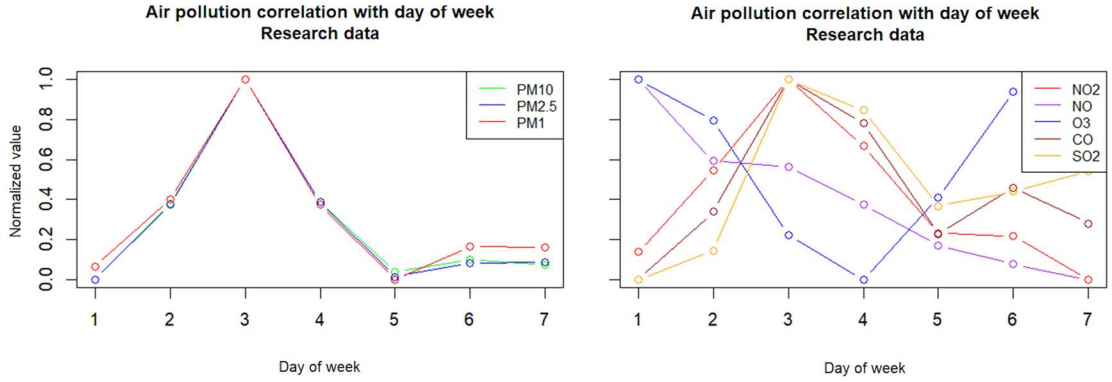

Fig. 3.16. Charts of the relationship between day of week and air pollutions (one-month long data)

It is also worth comparing these dependencies on different time intervals, which is presented in *Fig. 3.17*. For 20-month data, the relationship is very visible, however, if the forecast is to be made on the basis of data from one week, it is worth not taking this variable into account, because it may have a negative impact on the quality of the forecast, it is difficult on the graph on the left, look for some relationships.

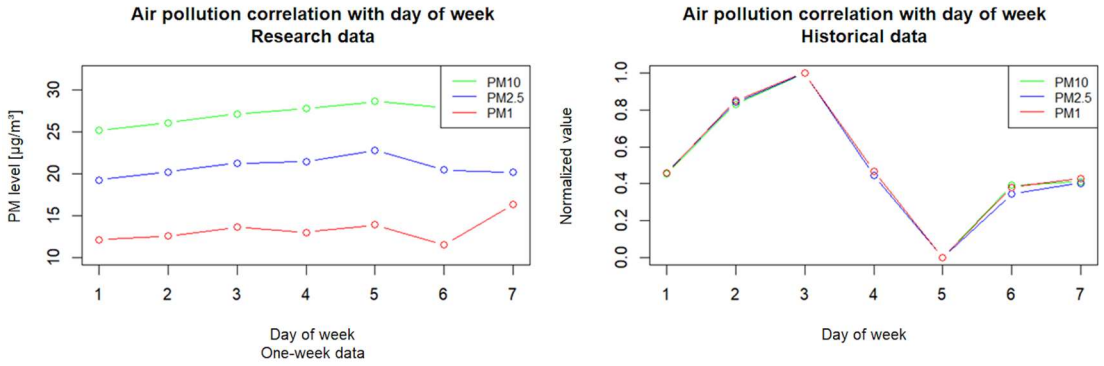

Fig. 3.17. Comparison of the relationship between the day of the week and dust pollution for one week and 20 months of data

Another important variable is the wind direction; its relationship with pollution is non-linear and takes the form of a sinusoid, which is presented in the graphs (*Fig. 3.18*). Both graphs differ because the relationship of this variable is related to the geographical location, and both sets contain data for slightly different areas. The location-related influence here is the nearby terrain, the location of industrial plants or urban clusters from which polluted air may potentially be blown.

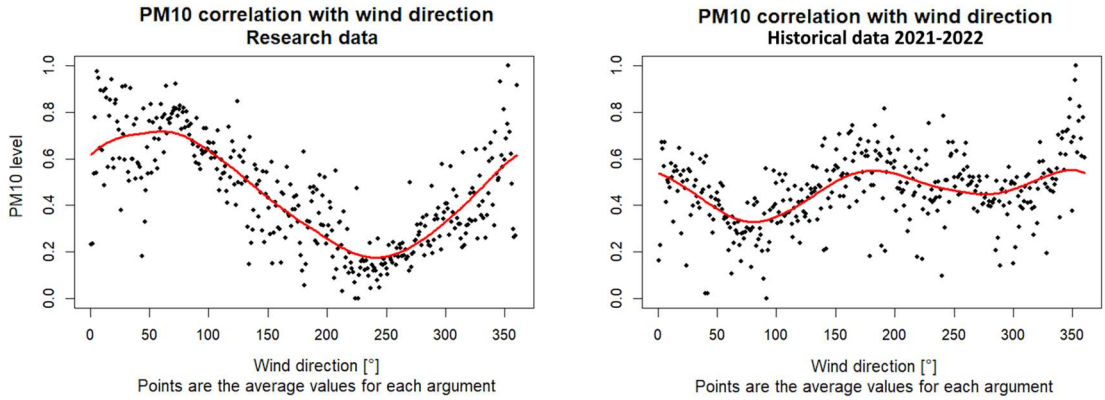

Fig. 3.18. Charts of the relationship between wind direction and PM_{10} dust pollution. Arguments represent the direction from which the wind is blowing.

The maps in *Figure 3.19* present wind directions for maximum and minimum concentration levels for the Krakow region, based on historical data, over a period of 20 months. Certain patterns can be noticed in the geographical location, with maximum values of pollution, wind directions closer to Krakow and the Sandomierska Valley are mostly from the west, which may be caused by the inflow of pollutants from industrialized Silesia, while the reverse values of wind directions may be dictated by the location in mountainous terrain where mountain ridges influence the direction of the wind. In the case of the smallest values, Kraków is, in some places, a transition from the south to the north. Although it seems that Krakow receives most of the pollution from the outskirts, and not the other way around, it is worth noting that the cause of inflow emissions to Krakow is not only the wind, but also the topography of the area in the Krakow area and the heating season. The data contains observations from over a year, in the summer period most pollution is created in Krakow by car traffic in the center, the suburbs do not generate emissions from combustion in home furnaces, then they are less polluted, which is why the wind "inward" Krakow contributes to lower pollution values in its area outside the heating season.

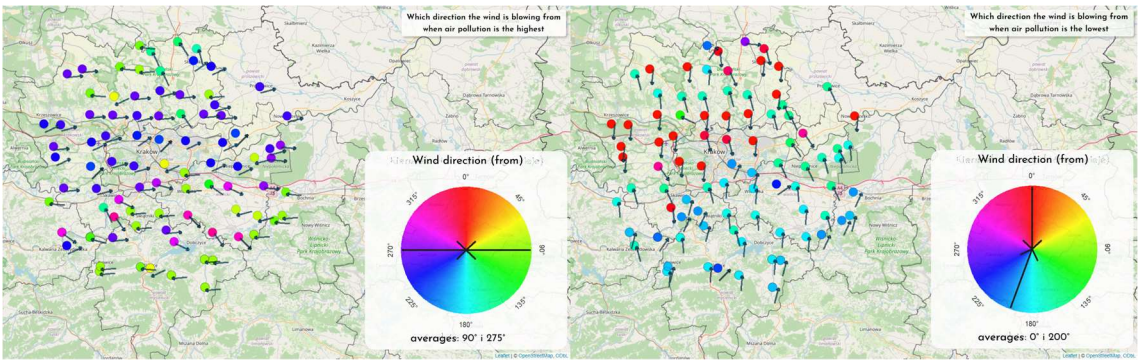

Fig. 3.19. A summary of maps showing, for each point, for which values of the wind direction, pollution is the highest and lowest, respectively. Based on historical data, covering only the Krakow region

For the analysis of the entire voivodeship, it can be seen in *Figure 3.20*, especially for the minimum pollution value, that for most points the wind direction is very similar, with different values sometimes occurring, but this is most likely due to the location in mountainous terrain, the ridges of which they block the wind from a specific direction. However, in the central and northern parts of the voivodeship, wind directions are almost identical. As for the wind direction at the highest pollution, there are not such equal values here, but in most cases, it is a direction close to the east. After

averaging the results, the values in both cases will be almost opposite. Considering that these are data from one month, it is difficult to find any permanent reason for these dependencies. The wind direction at minimum pollution values can be explained by the inflow of warm and humid air masses from the west, which may lead to lower pollution levels, while the wind directions at the highest pollution levels are probably more dictated by the location in relation to cities or industrial plants, although, given the average direction as north-eastern, it can also be considered that it was influenced by the polar - continental air flowing from Russia, which is strongly cooled in the lower layers and quite dry (B. Skowera, J. Wojkowski, 2009).

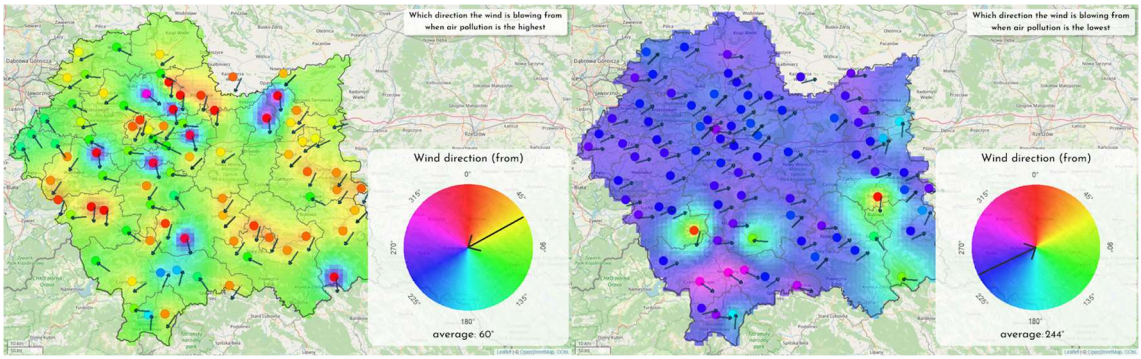

Fig. 3.20. A summary of maps showing, for each point, for which values of the wind direction, pollution is the highest and lowest, respectively. Based on data obtained from research covering the entire voivodeship.

In the spatial context, in the Lesser Poland Voivodeship, correlations with meteorological variables look almost identical throughout the entire area, only the humidity correlation seems to change in space, but these values are remarkably close to zero, so it can be assumed that these correlations are insignificant. *Fig. 3.21.* shows the variability of correlations in space.

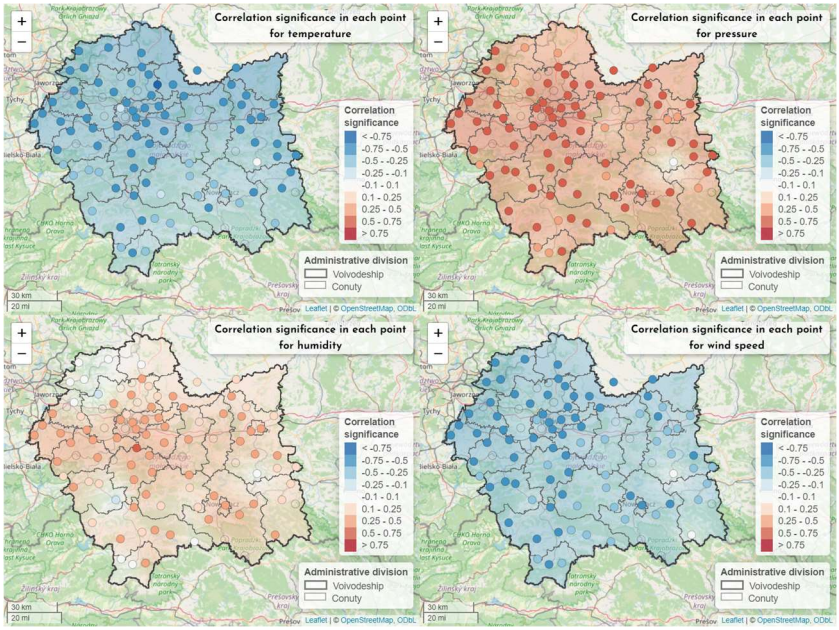

Fig. 3.21. Maps showing correlations of selected meteorological variables with PM_{10} dust for each point.

5. Testing pollution forecasts

After preprocessing, data ready for modeling is available, from which a training set (6 days long) and a test set (24 hours long) are created. Data sets prepared in this way can be used in forecasts; in the case of a neural network and a random forest, in the R language, the built-in functions "neuralnet" and "randomForest" are used. These functions take a dependency function as their first argument, where you enter the output and input variables. The next parameters are individual for each model, the random forest takes the number of trees, and the neural network takes the network structure. A forecast using a given model is made with the predict() function.

In a further step, the quality of the models was analyzed depending on various parameter settings, tests were performed on PM10 pollution. The charts below (*Fig. 3.22.*) present a comparison of test (actual) and predicted data, this is a graphical method of assessing the quality of the model, the closer the points are to the red line, the better the quality of the model, i.e. the predicted values are very close to real values. The graphs show decent quality models, an MAE error of less than 5% indicates a good fit of the data.

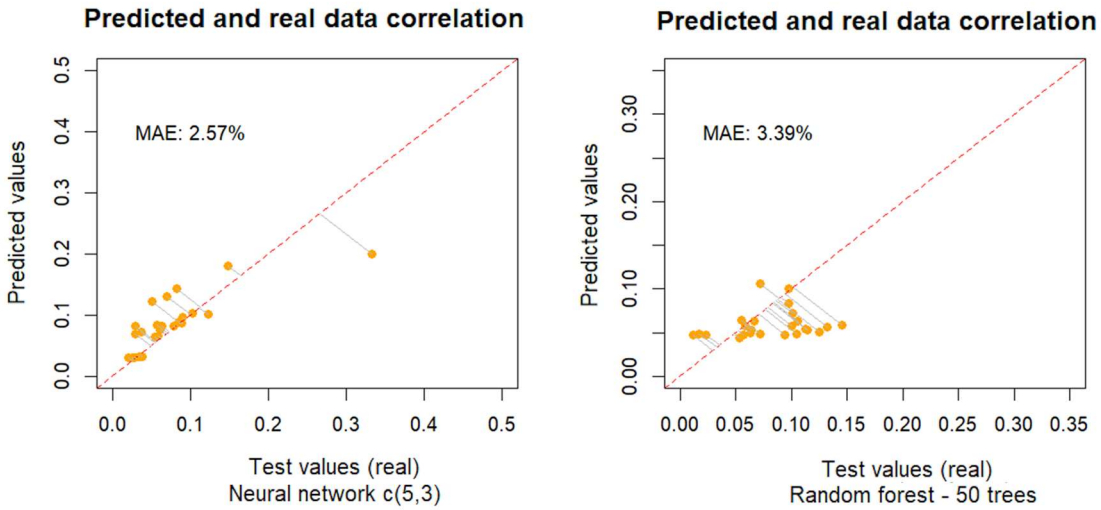

Fig. 3.22. Comparison of the relationship between predicted values and actual values on the example of a random forest and a neural network. Both models in this case are of decent quality.

It is also worth looking at the models in terms of quality variability depending on the point, which is presented in *Fig. 3.23*. In the case of both models, such variability occurs, for some points these values are on average higher, for others lower. Random forest models illustrate this well. It is also worth noting that the range of model quality assessments for points in this case is small, unlike a neural network, where the qualities differ more significantly. This may be due to the fact that the random forest is a more stable model compared to the neural network, which is a more complicated model, but more flexible (Roßbach, 2018).

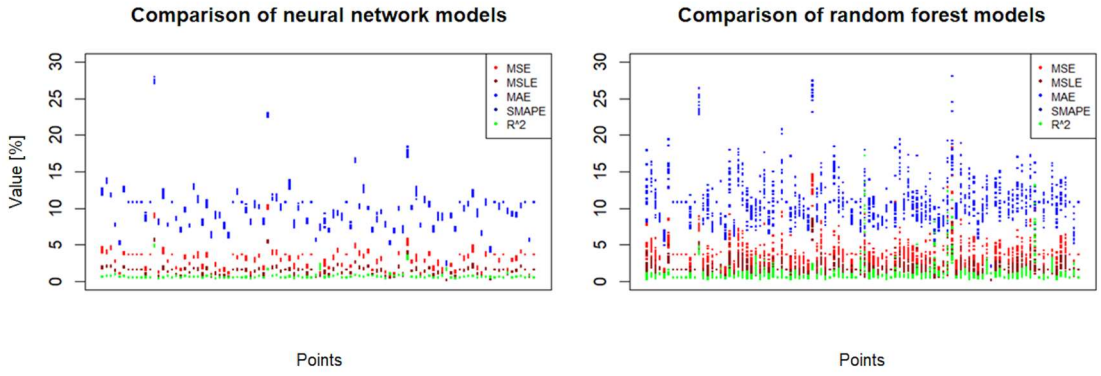

Fig. 3.23. Comparison of random forest and neural network models based on their evaluation with other measures, sorting by measurement points.

In Fig. *3.24.*, PM_{10} pollution values over time are presented for points for which the models had the best or worst quality on average. For the point on the basis of which the models perform the best on average, it can be seen that, unlike the one with poor quality, the variability is more stable, you can see clear daily changes in pollution, of course there are jumps in concentration values from time to time, but they are much more stable, than the data from the second chart. The values on the graph on the right seem to change less stably, you can notice jumps lasting several hours, as well as huge daily changes for the initial hours of observation, in addition, there are fewer observations available (they end about 100 hours faster) and an error can also be noticed measurement, which are zero values at 800 hours.

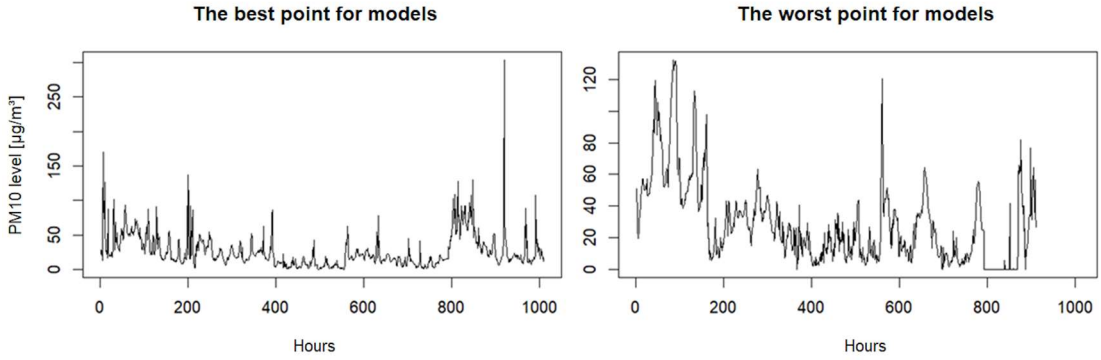

Fig. 3.24. PM_{10} dust variability charts for points, based on which the models created the best and worst forecasts, respectively.

When the neural network considers variables that correlate non-linearly, it must be informed about this in advance, then the model will "know" that it is dealing with non-linearly correlated variables. In the diagrams of *Fig. 3.25.* a comparison of models without considering nonlinearity and on the right with it is presented. Random forest does not need information about the linearity of the model because it works differently. When non-linearity is considered, the quality increases several times, so it should be considered when creating a forecast. It is also worth noting that the random forest always has better results than the neural network, this may be due to the fact that the random forest copes better with smaller data sets, but also with missing or erroneous data (Roßbach, 2018).

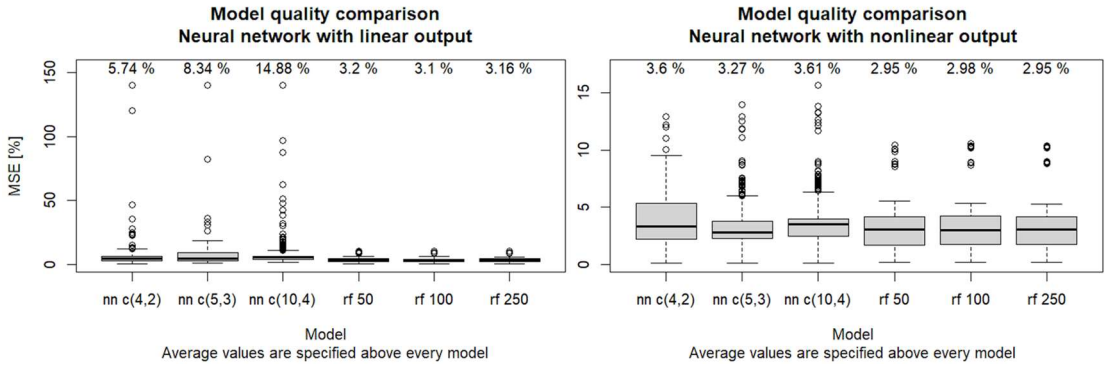

Fig. 3.25. Comparison of the quality of neural network models considering only linear dependencies on the left and considering non-linear dependencies on the right (83 repetitions). It is worth paying attention to the difference in vertical scale.

Then, models that take the day of the week as an input variable and those that do not were compared, the results are visualized in *Fig. 3.26.* A model created based on data covering more than a month, which considers the day of the week, will perform slightly better in forecasting than a model that does not take it into account. In such a period, the dependence of pollution on this variable begins to become more important and therefore its use in this case is justified, although not necessarily obligatory. Models created on the basis of data covering a week do not have as much data on the relationship with the day of the week, a forecast based on only one sample (one week) does not make sense in this case, because it leads to a flattening of the forecast data, which usually reduces the quality of the model , so it is worth building a model without taking this variable into account, it will save computational time and perhaps increase the quality of the model.

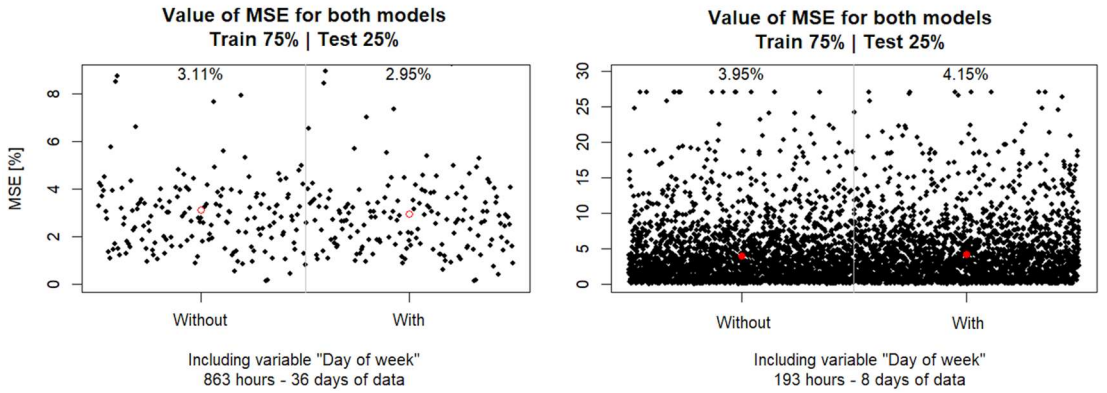

Fig. 3.26. Comparison of the quality of models considering the "Day of the week" variable with monthly and weekly periods

When selecting parameters for the model, you should also remember to choose the appropriate division between the training and test set. To check the quality of the model depending on the size of these sets, analyzes were performed at various time intervals. In the graphic *Fig. 3.27*, the graph on the left shows the quality of the models based on monthly data. As you can see, with such a large set of observations, it is worth having a slightly smaller training set, because with larger sets, overfitting may occur, which will make the model less effective and poorly it fit the data. It can be concluded that forecasting data from one month to one day is not the best idea, but it is also worth noting that less data for testing may also mean a higher value of the model quality metric, because conclusions are made on a not very representative

sample and the test set it cannot be of the order of one observation. In the case of data lasting a week, it is more reasonable to forecast for one day. Models with training sets smaller than 80% are mostly undertrained, they have an insufficient sample to learn the dependencies, therefore their quality is worse than in the case of a model with an 88% training set. In the case of shorter data, it is more difficult to overtrain the model, so you can afford larger training sets than test ones.

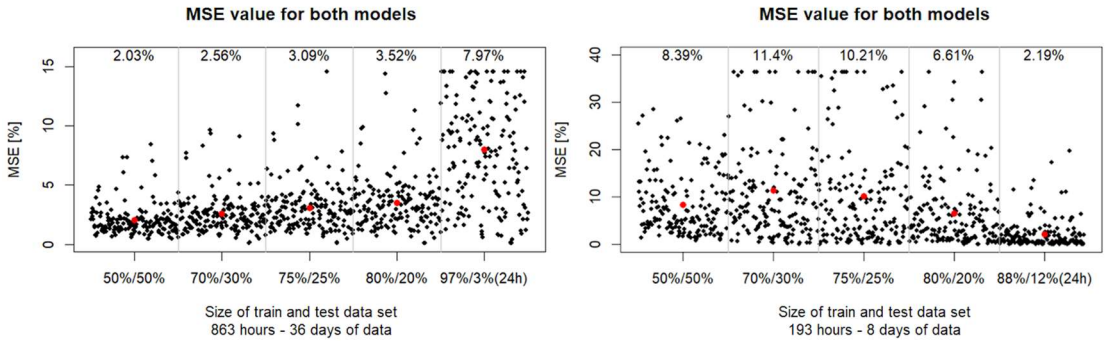

Fig. 3.27. Comparison of the quality of models with a monthly period with different ratios of training and testing set sizes.

Tests were also performed on models that included and did not include outliers; the comparison of their quality is presented in the chart *Fig. 3.28*. In many cases, in the pre-processing phase, outliers are trimmed, which may be anomalies or measurement errors that interfere with modeling. However, during model testing, no unrealistic anomalies were noticed that could be the result of measurement errors. Models that did not cut out outliers performed better.

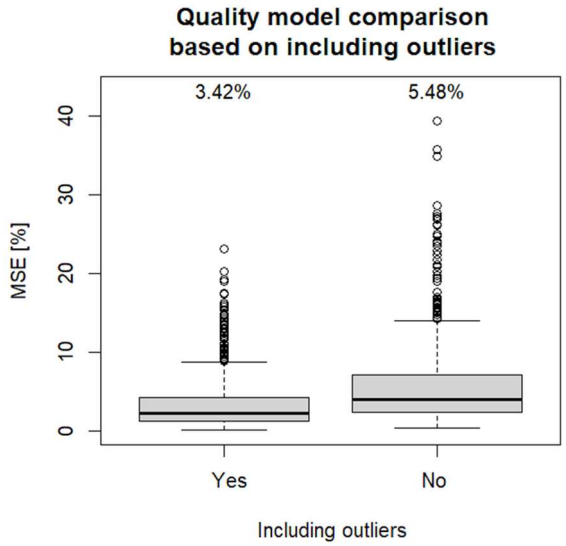

Fig. 3.28. Comparison of model quality with a weekly period based on the inclusion of outliers.

To further optimize the models, tests were performed on combinations of various input variables. The following factors were considered: sea level pressure, relative humidity, air temperature, rainfall intensity, wind speed and direction, hour of the day and day of the week. 36 combinations were created, considering 6, 7 and 8 input variables in all ways, tests were performed on PM_{10} and $PM_{2.5}$ dust, each point is 300

test models. The graphs below, in *Fig. 3.29*, show whether a given variable is included in the model or not.

From the charts, there can be seen that the variability in quality is not that large. Models with seven input variables have, on average, better quality than those with 6 or 8 variables. It cannot be clearly stated that any of the variables significantly worsens or improves the quality of the model, because each variable is in a slightly worse or better combination. It seems that the pressure and intensity of precipitation are found in most good-quality combinations, while in worse models the time of day seems to be there, but this is not a very noticeable element, so it is not possible, based on this analysis, to clearly indicate which factor performs better in the forecast.

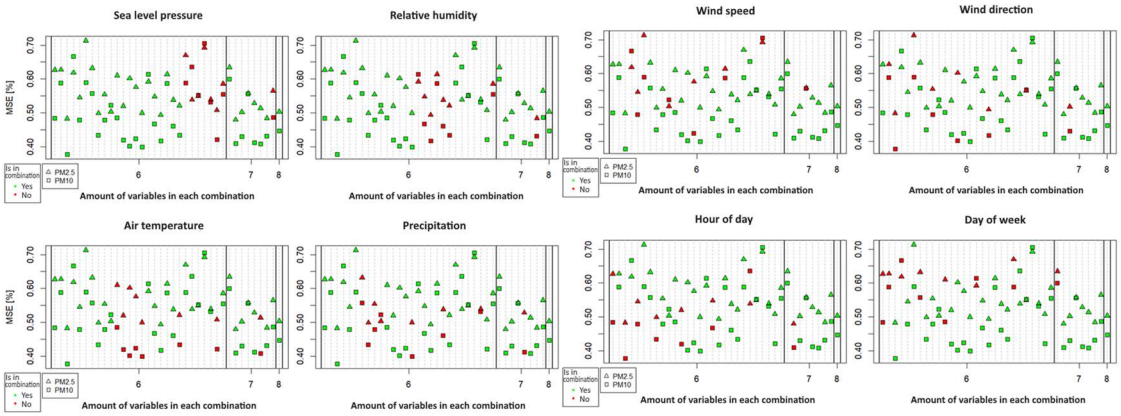

Fig. 3.29. Comparison of the quality of models with different combinations of input variables with the output variable PM_{10} and $PM_{2.5}$

6. Presentation of the dust pollution forecast in a web application.

Based on the above analyses, the following factors were considered for the forecast: temperature and air pressure, time of day, precipitation, and wind and speed, because they show the best correlation with air pollution. In the case of random forest, 50 trees were used, mainly to avoid overfitting, and it was also noticed that after this number the error did not decrease but stabilized. In the case of a neural network, the architecture consists of two hidden layers, one with five neurons and the other with three. Forecasts were made for dust pollution that contained the most non-empty observations, $PM_{2.5}$ and PM_{10}, using a training set of 6 days and a 24-hour test set.

The screenshots below, in *Fig. 3.30*, present the forecast of $PM_{2.5}$ dust pollution for the point in Gorlice and PM_{10} dust for the point in Kraków, the graphs compare the values of test data (in gray) and forecast data (in green), as well as the training data used (in black). Two metrics were used to estimate quality to better evaluate the models overall. Additionally, on the right side, there are charts directly comparing the test data with the predicted data with the given Pearson correlation coefficient.

In cases where a random forest is used, all metrics indicate that the quality of the forecast is better than for the neural network, which was also the result of the previous analysis, so it can be concluded that the forecast of the neural network in this case is incorrect. It is also worth paying attention to the unnaturally low values of the MSE meter in relation to the charts and the MAE meter, although the values are below 1% (which would indicate an exceptionally good match between the data), the prognostic and test data are noticeably inconsistent with each other, they are falling apart. In this case, the MSE is so low because, for the entire data set, the predicted values are close to zero. Bearing in mind that in the earlier days the values are much higher,

when normalizing the data, the maximum values will be close to one, while the values at the end of the data set are almost zero. Given the fact that this metric uses squared differences, errors close to zero are reduced even further, which is why the average is so small. In turn, the MAE metric is robust to such small values because it does not raise these differences to a power.

Visually speaking, it can be said that in the case of the sensor in Krakow (top graphs), the model correctly predicted the increase in air pollution values on January 17, while in the case of the sensor in Gorlice there was no such increase in concentration values, but there were minor fluctuations values, which the model also predicted to some extent.

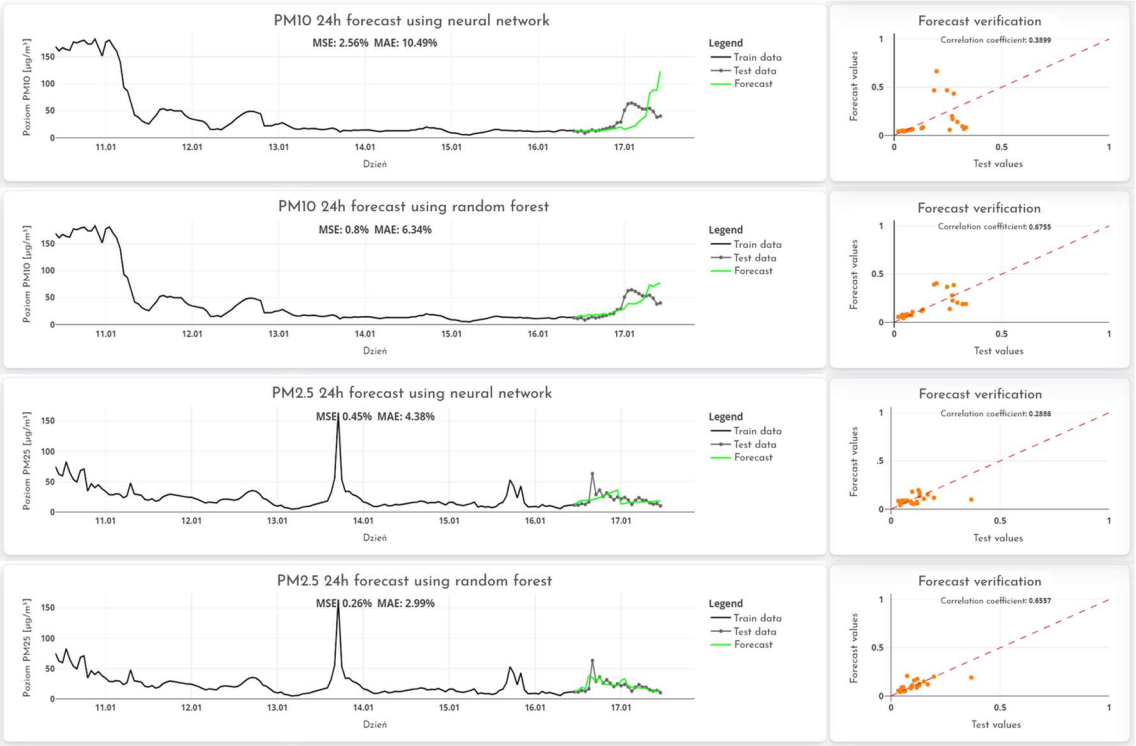

Fig. 3.30. Visualization of $PM_{2.5}$ dust pollution forecasts for a point in Gorlice and PM_{10} for a point in Krakow (Bulwarowa Street) using a random forest and a neural network

Conclusions

It can be concluded, from the forecast presentation, that models built based on one week's data, with appropriate selection of parameters and input variables, can match advanced forecast models available on popular map portals. It was noticed that in the case of prediction on a weekly dataset, random forest outperforms the neural network. It was also important to select the appropriate sizes of the test and training set to avoid both under- and over-fitting. However, this forecast is effective for a short, 24-hour period, because the model does not have information about the influence of seasons, days of the week and others. The models work well if the data provided does not contain errors or some anomalies occur for the forecast period that only continental and larger modes can detect. As with the weather in various parts of the world, pollution levels in various places will usually be higher or lower than elsewhere, but it has been shown that for the entire area, meteorological variables and air pollution correlate similarly. The greatest impact on pollution is caused by changes in the time of day and temperature; these are indirect relationships, because in most cases it is the human response to changes in these factors that directly leads to a change in the level of pollution. The strength of the wind and the intensity of rainfall also have a significant impact, which directly affect the concentration of pollutants.

Literature

1. Konwencja Nr 148 dotycząca ochrony pracowników przed zagrożeniami zawodowymi w miejscach pracy, z 20 czerwca 1977, (Dz.U. 2004 Nr 29, poz. 255).

2. A. Krenker, J. Bešter, A. Kos. (January 2011). *Artificial Neural Networks - Methodological Advances and Biomedical Applications.*

3. https://airindex.eea.europa.eu/Map/AQI/Viewer/# (January 2024), access: January 2024

4. https://airly.org/pl/gdzie-jest-najwiecej-smogu-w-polsce-przeczytaj-analize-airly/, access: January 2024

5. https://airly.org/pl/tlenek-azotu-trujace-skladniki-smogu-cz-1/, access: January 2024

6. https://airly.org/pl/jakie-dzialania-antysmogowe-moga-wdrozyc-gminy/, access: January 2024

7. Aishwarya, B. (Ferbruary 2022). *Regression Metrics - Of all metrics why MSE?* Downloaded from https://www.linkedin.com/pulse/regression-metrics-all-why-mse-aishwarya-b/

8. Amos P.K. Tai, L. J. (October 2010). Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensivity of PM2.5 to climate change. *Atmospheric Environment, 44*(32), 3976-3984.

9. B. Skowera, J. Wojkowski. (2009). Wpływ sytuacji synoptycznych na temperaturę powietrza w południowej części Wyżyny Krakowsko-Częstochowskiej. *Infrastruktura i Ekologia Terenów Wiejskich, 2009*(5), 123-135.

10. Carpenter, M. E. (March 2018). *How Do Mountains Affect Precipitation?* Downloaded from sciencing.com: https://sciencing.com/do-mountains-affect-precipitation-8691099.html

11. Ćwik, P. (April 2017). *Dwutlenek siarki. W Polsce źle, na Bałkanach gorzej*. Downloaded from https://smoglab.pl/dwutlenek-siarki-w-polsce-zle-na-balkanach-gorzej-czym-truje-nas-smog-4/

12. Elminir, H. K. (November 2015). Dependence of urban air pollutants on meteorology. *Science of The Total Environment, 350*(1-3), 225-237.

13. https://gis-support.pl/baza-wiedzy-2/dane-do-pobrania/granice-administracyjne/, access: January 2024

14. He, R.-R. (Ferbruary 2023). Quantifying the weekly cycle effect of air pollution in cities of China. *Stochastic Environmental Research and Risk Assessment, 37*, 2445-2457.

15. Hiregoudar, S. (August 2020). *Ways to Evaluate Regression Models*. Downloaded from https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70, access: January 2024

16. Koch, D., J. Park, A. Del Genio. (2003). Clouds and sulfate are anticorrelated: a new diagnostic for global sulfur models. *Journal of Geophysical Research - Atmospheres, 108*(D24), 4781.

17. M. Dziekciarz, M. Foremniak. Korytarze powietrzne a zanieczyszczenie powietrza w miastach. Downloaded: January 2024

18. https://posit.co/products/open-source/rstudio/, access: January 2024

19. https://powietrze.gios.gov.pl/pjp/content/health_informations, access: January 2024

20. https://powietrze.gios.gov.pl/pjp/content/show/1000919, access: January 2024

21. https://powietrze.malopolska.pl/program-ochrony-powietrza/, access: January 2024

22. Roßbach, P. D. (2018). Neural Networks vs. Random Forests – Does it always have to be Deep Learning?

23. S. Shelton, G. Liyanage, S. Jayasekara, B. Pushpawela, U. Rathnayake, A. Jayasundara, L. D. Jayasooriya (2022). Seasonal Variability of Air Pollutants and Their Relationships to Meteorological Parameters in an Urban Environment. *Advances in Meteorology, 2022*, 18.

24. Saxena, S. (June 2019). *What's the Difference Between RMSE and RMSLE?* Downloaded from https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a, access: January 2024

25. Uchwała nr LXXV/1102/23 Sejmiku Województwa Małopolskiego z dnia 20 Novembera 2023 r. w sprawie zmiany uchwały Nr XXV/373/20 Sejmiku Województwa Małopolskiego z dnia 28 września 2020 r. w sprawie Programu ochrony powietrza dla województwa małopolskiego.

26. UCHWAŁA NR XVIII/243/16 SEJMIKU WOJEWÓDZTWA MAŁOPOLSKIEGO z dnia 15 stycznia 2016 roku w sprawie wprowadzenia na obszarze Gminy Miejskiej Kraków ograniczeń w zakresie eksploatacji instalacji, w których następuje spalanie paliw.

27. WHO. (January 2006). Air Quality Guidelines. Global Update 2005.

28. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics, access: January 2024

29. https://www.krakow.pl/213212,1962,230530,powietrze,faq.html, access: January 2024

30. https://www.krakow.pl/aktualnosci/274334,26,komunikat,krakow_przeciwny_lagodzeniu_uchwal_antysmogowych.html, access: January 2024

31. https://www.malopolska.uw.gov.pl/default.aspx?page=tlenek_wegla, access: January 2024

32. https://www.r-project.org/, access: January 2024

33. Yun Bai, Yong Li, Xiaoxue Wang, Jingjing Xie, Chuan Li. (2016, May). Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. *Atmospheric Pollution Research, 7*(3), 557-566.