



AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ GEOLOGII, GEOFIZYKI I OCHRONY ŚRODOWISKA

Katedra Geoinformatyki i Informatyki Stosowanej

Projekt dyplomowy

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

Autor: *Dawid Makowski*
Kierunek studiów: Geoinformatyka
Opiekun pracy: *dr. Tomasz Danek*

Kraków, 2024

Spis treści

Wprowadzenie.....	2
Wstęp	2
Zanieczyszczenia powietrza	2
Indeksy jakości powietrza	2
Metody badań pyłu zawieszonego	3
Przyczyny smogu i jego różnicowanie w Polsce.....	3
Przeciwdziałanie problemowi zanieczyszczeń	4
Cel pracy	4
Metodologia.....	5
Narzędzia analizy	5
Aplikacja Shiny jako narzędzie wizualizacyjne.....	6
Modele uczenia maszynowego w prognozowaniu zanieczyszczeń.....	8
Wyniki badań.....	9
Wybór punktów	9
Uzyskanie danych korzystając Airly i OpenMeteo API	9
Wizualizacja i wstępne wnioski wyników badań	11
Badanie korelacji	12
Testowanie prognozy zanieczyszczeń – budowa modeli	20
Prezentacja prognozy zanieczyszczenia pyłami w aplikacji webowej.....	24
Wnioski	26
Literatura	27

Wprowadzenie

1. Wstęp

W projekcie korzystano z narzędzia ChatGPT firmy OpenAI, jego wykorzystanie ograniczało się do streszczania artykułów naukowych, żeby szybciej znaleźć wnioski albo informacje do danego zagadnienia i pomocy w szukaniu źródeł.

2. Zanieczyszczenia powietrza

Zanieczyszczenia powietrza, to wszelkie skażenie powietrza przez substancje, które są szkodliwe dla zdrowia lub niebezpieczne z innych przyczyn, bez względu na ich postać fizyczną (Dz.U. 2004 Nr 29, poz. 255).

Najczęściej występujące zanieczyszczenia to **pyły zawieszone**, zawierające różne związki chemiczne, ich występowanie związane jest m.in. z procesami spalania paliw stałych i ciekłych. Przede wszystkim wpływa on negatywnie na układ oddechowy, w sposób pośredni pyły wpływają również negatywnie na resztę organizmu. Zanieczyszczenia te podzielone są na trzy grupy: **PM₁**, **PM_{2,5}**, **PM₁₀**, liczby w ich nazwach oznaczają maksymalny rozmiar cząsteczek dla danego pyłu w mikrometrach. Im mniejszy rozmiar pyłu, tym groźniejszy jest on dla człowieka, ponieważ łatwiej jest mu przeniknąć do krwioobiegu (www.epa.gov, 2024).

W smogu pojawiają się również metale ciężkie oraz gazy. Wśród nich można wyróżnić **tlenek węgla**, którego źródłem są najczęściej źle zainstalowane lub niesprawne kuchenki gazowe, (www.malopolska.uw.gov.pl, 2024). W powietrzu pojawiają się również **tlenki siarki**, które są emitowane do atmosfery zarówno w procesach naturalnych (erupcje wulkanów, pożary lasów), jak i w wyniku działalności człowieka w rejonach zurbanizowanych, nawet krótka ekspozycja na tlenki siarki potrafi spowodować spore trudności z oddychaniem (Ćwik, 2017). Do smogu można również zaliczyć, szczególnie groźne, **tlenki azotu**, ich toksyczność jest wielokrotnie większa w porównaniu do tlenku węgla czy dwutlenku siarki. Związki powstają zwłaszcza na skutek przedostawania się do atmosfery spalin samochodowych, a także toksyn emitowanych przez zakłady przemysłowe (airly.org, 2024). W powietrzu może również pojawić się **ozon troposferyczny**, który jest zanieczyszczeniem wtórnym, powstały w wyniku reakcji innych związków chemicznych (S. Shelton, G. Liyanage, ..., 2022). Jest on silnym utleniaczem, dlatego wnikając do dróg oddechowych człowieka, powoduje ich podrażnienie i dyskomfort w oddychaniu (WHO, 2006).

3. Indeksy jakości powietrza

W celu informowania społeczeństwa o obecnym stopniu zanieczyszczenia powietrza różne agencje rządowe opracowują indeksy jakości powietrza. Kiedy indeks jest wysoki, zachęca się ludzi do ograniczenia aktywności fizycznej na świeżym powietrzu lub nawet całkowitego wychodzenia na zewnątrz. Różne kraje mają własne wskaźniki jakości powietrza, odpowiadające różnym krajowym normom. W Polsce popularnym jest, opracowany przez Główny Inspektorat Ochrony Środowiska, **PIJP (Polski Indeks Jakości Powietrza)** (powietrze.gios.gov.pl, 2024). Wartość indeksu jest ustalana w oparciu o wartości tabeli 1.1.

*Tab. 1.1. Klasyfikacja zanieczyszczeń wg Polskiego Indeksu Jakości Powietrza
(powietrze.gios.gov.pl, 2024)*

Nazwa jakości	Godzinowe stężenie zanieczyszczenia [$\mu\text{g}/\text{m}^3$]						
	PM ₁₀	PM _{2.5}	O ₃	NO ₂	SO ₂	C ₆ H ₆	CO [mg/m^3]
Dobry	0 - 20	0 - 13	0 - 70	0 - 40	0 - 50	0 - 6	0 - 3
Dostateczny	20 - 50	13 - 35	70 - 120	40 - 100	50 - 100	6 - 11	3 - 7
Umiarkowany	50 - 80	35 - 55	120 - 150	100 - 150	100 - 200	11 - 16	7 - 11
Zły	80 - 110	55 - 75	150 - 180	150 - 200	200 - 350	16 - 21	11 - 15
Bardzo zły	110 - 150	75 - 110	180 - 240	200 - 400	350 - 500	21 - 51	15 - 21
Ekstremalnie zły	> 150	> 110	> 240	> 400	> 500	> 51	> 21
Brak indeksu	Indeks jakości powietrza nie jest wyznaczony z powodu braku pomiaru zanieczyszczenia						

W Europie najpopularniejszym indeksem jest, wprowadzony w 2017 roku, **EAQI**, dostarcza on informacji o aktualnej sytuacji w zakresie jakości powietrza w oparciu o pomiary z ponad 2000 stacji monitorowania i opiera się na 5 kluczowych zanieczyszczeniach, które szkodzą zdrowiu ludzi i środowiska: pył zawieszony, ozon troposferyczny, dwutlenek azotu i dwutlenek siarki. Do indeksu wliczanie są stężenia godzinowe, a dla średniej 24-godzinnej są to dane z poprzedzających 24 godzin pomiarów (airindex.eea.europa.eu, 2024).

4. Metody badań pyłu zawieszonego

W celu pomiaru zanieczyszczenia powietrza w Polsce, stosowane są dwie metody pomiaru pyłu. Jedną z nich jest **metoda grawimetryczna**, w której używa się tzw. poborników pyłowych, co dwa tygodnie do pobornika zakłada się 14 jednorazowych filtrów, które urządzenie zmienia automatycznie co 24 godziny. W laboratoriach filtry są ważone przed i po okresie ekspozycji, a z różnic mas wyliczane są stężenia pyłów. Metoda jest bardzo dokładna, ale czas potrzebny na uzyskanie wyników wynosi ok. 3 tygodni. Drugą metodą jest **metoda automatyczna**, w której stosuje się mierniki automatyczne, które posiadają certyfikaty potwierdzające ich równoważność z metodą referencyjną. Mierniki te na bieżąco mierzą stężenia pyłu, co umożliwia pokazywanie wyników tych pomiarów w trybie „on-line” (powietrze.gios.gov.pl, 2024).

5. Przyczyny smogu i jego zróżnicowanie w Polsce

Mapa na ilustracji Fig. 1.1. prezentuje średnioroczne zanieczyszczenie powietrza w Polsce pyłem PM₁₀. Najwyższe stężenia zarejestrowano w województwach śląskim, małopolskim i łódzkim, natomiast najlepszym powietrzem oddychają mieszkańcy województw pomorskiego oraz zachodniopomorskiego.

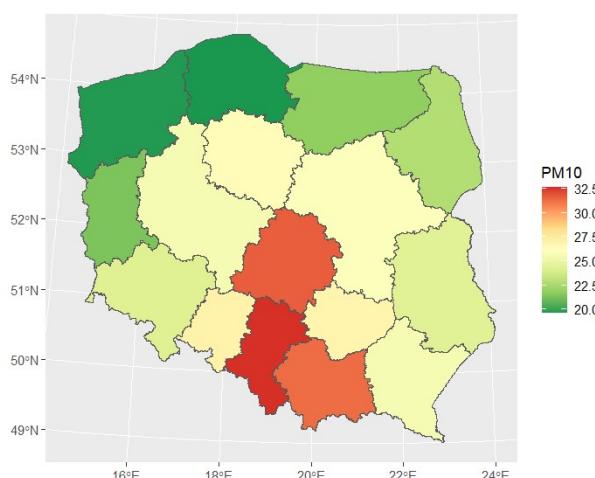


Fig. 1.1. Średni poziom PM₁₀ w danym województwie w 2021 roku na podstawie danych GIOŚ (powietrze.gios.gov.pl)

Możliwymi czynnikami niemeteorologicznymi, które w największym stopniu wpływają na zwiększenie zanieczyszczenia w danej lokalizacji są: gęstość zaludnienia, liczba budynków indywidualnych oraz długość dróg w otoczeniu urządzenia pomiarowego. Są to czynniki powiązane z obecnością emisji – w tym niskiej emisji (budynki jednorodzinne) i co więcej, czynniki te posiadają dużą zmienność w przestrzeni, potrafią być one bardzo zróżnicowane nawet w obrębie tych samych miast. Natomiast czynnikami, które ograniczają zanieczyszczenia powietrza może być powierzchnia terenów zielonych oraz topograficzny indeks pozycji (TPI), czyli stopień położenia w dolinie lub na grzbiecie górskim (airly.org, 2024).

Głównym źródłem odpowiedzialnym za poziom zanieczyszczenia w województwie małopolskim jest emisja pochodząca z sektora komunalno-bytowego, która odpowiada odpowiednio za około: 90% emisji pyłów oraz 16% emisji NO_x. Transport odpowiada za około: 4% emisji pyłów oraz 44% emisji NO_x, ruch samochodowy powoduje również wtórny unos pyłu z powierzchni jezdni. Emisja przemysłowa w skali województwa generuje niewielkie ilości pyłów o wartościach ok. 2%, jednak jest istotnym czynnikiem emisji gazów do powietrza – odpowiada za 29% emisji NO_x. Pozostałe źródła, takie jak rolnictwo (uprawa oraz hodowla), lasy oraz pożary, odpowiadają za emisję: 11% NO_x, 9% PM₁₀ oraz 1% PM_{2.5} (powietrze.malopolska.pl, 2024).

Dodatkowo, w większych miastach, wpływ na zanieczyszczenie ma zabudowa korytarzy powietrznych, które pozwalają wywierać zanieczyszczenia z miasta i wprowadzać do niego świeże powietrze (M. Dziekciarz, M. Foremniak). W przypadku Krakowa czynnikami są również, emisja napływna, czyli przemieszczanie się zanieczyszczeń z gmin ościennych, nieobjętych uchwałą antysmogową oraz niekorzystne położenie geograficzne, ponieważ miasto leży w dolinie rzecznej i jest z trzech stron otoczony wzgórzami, ruch mas powietrza jest wokół niego ograniczony (airly.org, styczeń 2024).

6. Przeciwdziałanie problemowi zanieczyszczeń

Samorządy, w celu zmniejszenia zanieczyszczeń mogą wprowadzać różne działania, takie jak edukacja antysmogowa, dofinansowanie komunikacji miejskiej, promowanie energii odnawialnej czy zwiększenie ilości zieleni miejskiej (airly.org, 2024).

Przykładowo, na terenie województwa małopolskiego, obowiązuje Program Ochrony Powietrza (Uchwała Nr LXXV/1102/23), który ogranicza eksploatację kotłów na paliwo stałe. W Krakowie od 2019 roku obowiązuje uchwała (UCHWAŁA nr XVIII/243/16) całkowicie zakazująca używania pieców na paliwo stałe, co znacznie zmniejszyło średni poziom zanieczyszczenia w mieście (www.krakow.pl, 2024). Miasto prowadzi też wiele programów zapobiegawczych, taki jak *Program termomodernizacji budynków jednorodzinnych*, *Program STOP SMOG* i wiele innych, a także podejmuje wiele inwestycji mających pomóc w tym problemie, jak rozbudowa linii tramwajowych, systemu transportu czy ścieżek rowerowych, a gdy średnia wartość zanieczyszczenia pyłem PM₁₀ w mieście przekroczy 100 µg/m³, wprowadzana jest darmowa komunikacja miejska (www.krakow.pl, 2024).

7. Cel pracy

Celem pracy jest zebranie danych o zanieczyszczeniach powietrza i meteorologii na terenie województwa małopolskiego, z wykorzystaniem jak największej ilości fizycznych czujników punktowych na jego obszarze. Następnie na podstawie zebranych danych stworzenie i zaimplementowanie optymalnych modeli prognostycznych zanieczyszczeń powietrza dla danego punktu, które będą wykorzystywały jako zmienne wejściowe dane meteorologiczne.

Metodologia

1. Narzędzia analizy

W projekcie wykorzystano język **R**., jest to interpretowany język programowania oraz środowisko do obliczeń statystycznych i wizualizacji wyników. Posiada on bogatą społeczność i wsparcie, daje on dostęp do wielu pakietów, narzędzi i materiałów edukacyjnych (www.r-project.org, 2024). Cały projekt wykonano w środowisku **RStudio**, który pozwala na budowanie skryptów w języku R i Python, automatyczny import danych, tworzenie raportów, itd. (posit.co, 2024)

Ważnymi pakietami użytymi w tej pracy jest **Leaflet**, który pozwala na tworzenie podstawowych, interaktywnych map oraz **Shiny**, odpowiadający za całą stronę wizualną. Pakiet **plotly** pozwala tworzyć interaktywne wykresy, które można powiększać i sprawdzać wartości wybranych punktów. Pakiety **randomForest** i **neuralnet** są odpowiedzialne za dostarczenie algorytmów prognozy, tj. sieci neuronowej i lasu losowego, na ich podstawie później tworzone są modele prognostyczne. Pakiety **spatstat**, **sf**, **sp** i **gridlayout** zostały użyte w pracy z danymi shapefile, czyli jednostkami administracyjnymi. Pakiety **raster** i **automap** były niezbędne w tworzeniu rastrowej mapy danego czynnika w aplikacji. Reszta pakietów była wykorzystywana do pojętycznych zadań analitycznych lub do celów estetycznych.

Do pobrania danych o zanieczyszczeniach i meteorologii wykorzystano usługę pobierania API portalu Airly i OpenMeteo. Dane uzyskane z serwisu Airly nie zawsze były kompletne, szczególnie meteorologia, dodatkowo istniała potrzeba uzupełnienia informacji o kolejne dane meteorologiczne. W tym celu skorzystano z serwisu OpenMeteo, który w swoim API dostarcza obszerne informacje meteorologiczne dla każdego punktu na Ziemi. Poniższy fragment kodu, na ilustracji Fig. 2.1., zawiera wywołania pobierające dane z obu serwisów.

```
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. Fragment kodu z RStudio zawierający wywołania pobierające dane z API Airly i OpenMeteo

Do analizy wykorzystano również dane zaczerpnięte z Bazy Wiedzy GIS Support (gis-support.pl, 2024) na podstawie Państwowego Rejestru Granic (PRG), gdzie znajdują się pliki w formacie shp (shapefile) zawierające dane o granicach administracyjnych województw, powiatów, gmin i dzielnic Krakowa. Punkty pomiarowe są agregowane w fazie preprocessingu, polega to na tym, że każdemu punktowi jest przypisywany gmina i powiat.

2. Aplikacja Shiny jako narzędzie wizualizacyjne

Dobra wizualizacja jest kluczowa dla zrozumienia problemu, dlatego też temu zagadnieniu została poświęcona znaczna część tego projektu. W tym celu stworzono aplikację webową, która pozwala na kompleksową interakcję z danymi poprzez manipulację różnymi parametrami i ustawieniami, a także operowanie prognozą zanieczyszczeń według zadanego ustawięń.

Aplikacja składa się z trzech zakładek, serwis mapowy został umieszczony w zakładce „Interaktywna mapa”. Składa się z paska bocznego, który zawiera opcje wyboru zanieczyszczenia czy agregacji danych, oprócz tego jest możliwość wyboru skalowania kolorów lokalnie, żeby ukazać różnicowanie wewnętrz otrzymanych danych lub globalnie, czyli skalowanie wg Polskiego Indeksu Jakości Powietrza. W menu bocznym znajduje się wykres obrazujący średnią wartość danego czynnika w czasie. W przypadku wiatru, oprócz prędkości wyświetlany jest też azymut jego kierunku. Grafika Fig. 3.3. prezentuje wyżej opisaną zakładkę.

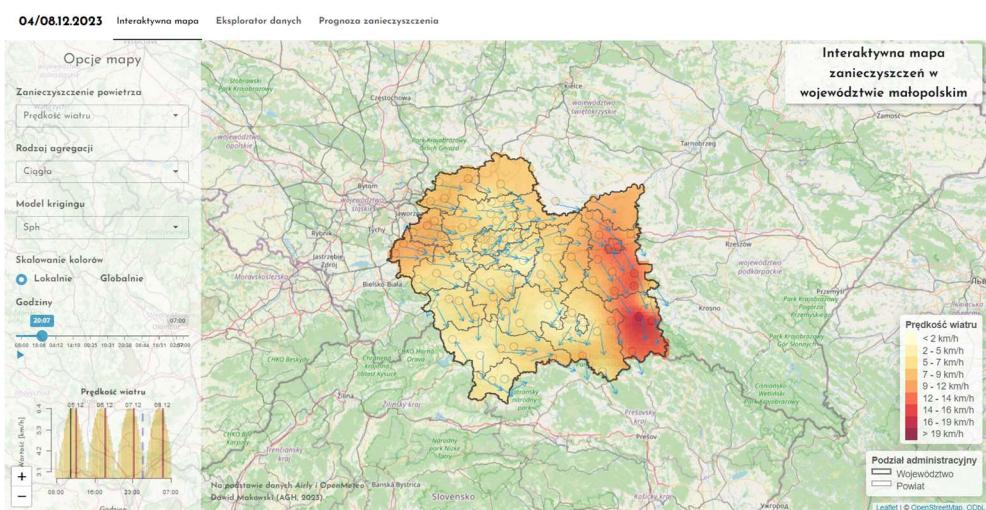


Fig. 3.3. Wygląd zakładki "Interaktywna mapa" aplikacji Shiny

Drugą zakładką jest „Eksplorator danych”, przedstawiony na grafice Fig.3.4., który zestawia dane tabelaryczne wyświetlanego na mapie punktów, gdy w pierwszej zakładce wybierany jest dany czynnik, np. PM_{2,5}, to tabela wyświetla tylko ten czynnik w specjalnie do tego wybranej kolumnie. Tabela podzielona jest na strony jeśli punktów jest więcej niż 10, można też filtrować punkty wg własnego upodobania, wybrać powiaty, do których chce się ograniczyć wyświetlanie punktów lub wartość minimalną / maksymalną wybranego czynnika, dla których punkty będą wyświetlane, istnieje też możliwość słownego wyszukiwania punktów po ich adresie. Użytkownik w kolumnie „Przybliż”, klikając celownik, może przekierować się do pierwszej zakładki ze zblizonym widokiem na wybrany punkt.

http://127.0.0.1:7117 | Open in Browser | Publish

23/28.12.2023 | Interaktywna mapa | Eksplorator danych | Prognoza zanieczyszczenia

Powiaty
Wszystkie powiaty | Wartość minimalna | Wartość maksymalna

Show 10 entries | Search:

ID	Adres czujnika	Województwo	Powiat	Gmina	Wysokość	PM2.5	Szerokość geograficzna	Długość geograficzna	Przybliż
1	19 Kraków, Bulwarowa	małopolskie	powiat Kraków	Kraków	203.87	35.88	50.069308	20.053492	
2	283 Limanowa, Jana Pawła II	małopolskie	powiat limanowski	Limanowa	409.7	10.31	49.709712	20.422682	
3	504 Gorlice, Stefana Batorego	małopolskie	powiat gorlicki	Gorlice	293.74	15.41	49.646214	21.176163	
4	615 Zelków, Krakowska	małopolskie	powiat krakowski	Zabierzów	346.38	13.83	50.15202	19.803083	
5	643 Ryglice, Tarnowska	małopolskie	powiat tarnowski	Ryglice	245.15	14.39	49.881924	21.155636	
6	835 Oświęcim, Rynek Główny	małopolskie	powiat oświęcimski	Oświęcim	245.54	6.64	50.038386	19.222059	
7	920 Bukowno, I Maja	małopolskie	powiat olkuski	Bukowno	304.45	14.37	50.263882	19.44846	
8	937 Olkusz, Szpitalna	małopolskie	powiat olkuski	Olkusz	366.6	13.37	50.27755	19.561504	
9	978 Andrychów, Jarosława Dąbrowskiego	małopolskie	powiat wadowicki	Andrychów	335.79	7.81	49.849926	19.34396	
10	1088 Brzeszcze, Ofiar Oświęcimia	małopolskie	powiat oświęcimski	Brzeszcze	261.95	10.93	49.985886	19.156971	

Showing 1 to 10 of 72 entries | Previous | 1 | 2 | 3 | 4 | 5 | ... | 8 | Next

Fig. 3.4. Wygląd zakładki „Eksplorator danych” aplikacji Shiny

Ostatnią zakładką jest „Prognoza zanieczyszczeń”, pozwala ona na wykonanie prognozy na żywo, na podstawie dostarczonych danych. Pasek boczny umożliwia dostosowanie parametrów według własnego uznania, można zastosować model lasu losowego lub sieci neuronowej, każdy z nich ma własne parametry do ustawienia. Poza tym można wybrać jakie zanieczyszczenie będzie prognozowane, a także na bazie jakich danych meteorologicznych i jakiego punktu ma zostać wykonana prognoza. Z prawej strony znajduje się mapa, która pokazuje punkty pomiarowe oraz ich odległości do wybranego przez użytkownika punktu, wyświetlany jest też zakres kierunku wiatru z ostatnich godzin. Grafika Fig. 3.5. prezentuje wyżej opisaną zakładkę. Po kliknięciu przycisku „Uruchom prognozę” zostanie wykonana prognoza wybranego zanieczyszczenia na bazie wybranych czynników. Po zakończeniu obliczeń pojawią się trzy interaktywne wykresy z pakietu plotly, informujące o jakości wykonanej prognozy. Górnny wykres porównuje wykorzystane w prognozie dane wraz z miernikami jakości MSE i MAE, wykres niżej bezpośrednio porównuje dane testowe z prognozowanymi, wyświetlając przy tym współczynnik korelacji Pearson'a. Wykres na dole informuje, w przypadku sieci neuronowej - o jej strukturze, a w przypadku lasu losowego - o wartości błędów średniokwadratowego w zależności od liczby drzew.

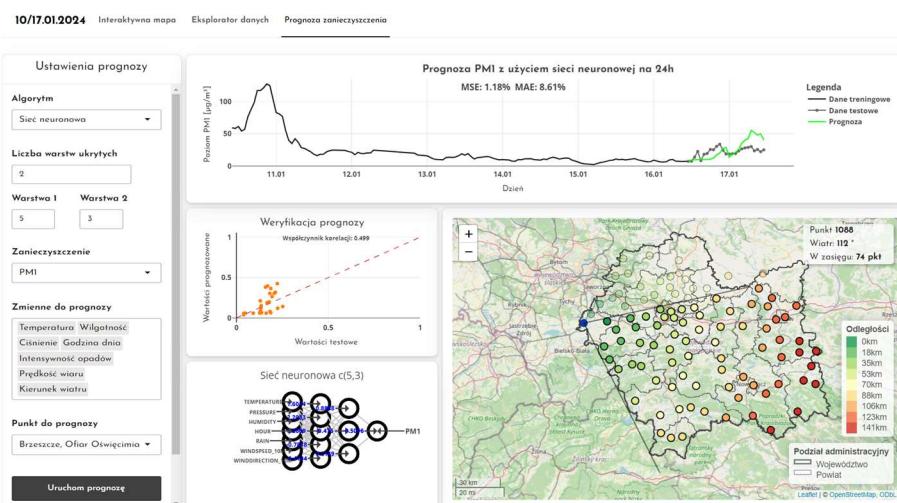


Fig. 3.5. Wygląd zakładki „Prognoza zanieczyszczenia” aplikacji Shiny

3. Modele uczenia maszynowego w prognozowaniu zanieczyszczeń

W celu stworzenia dokładnej prognozy zanieczyszczeń, w projekcie użyte zostały dwa modele.

Pierwszym z nich są **sztuczne sieci neuronowe**, które są matematycznym modelem, symulującym strukturę biologicznych sieci. Podstawowym elementem każdej sieci jest sztuczny neuron, który jest matematyczną funkcją. Na wejściu neuronu każda wartość wejściowa mnożona jest przez odpowiednią wagę, następnie, we wnętrzu neuronu wszystkie ważone dane są sumowane, po czym na wyjściu przechodzi przez funkcję aktywacji (A. Krenker, J. Bešter, A. Kos, 2011). Możemy wyróżnić kilka sieci neuronowych:

- **sieci neuronowe ze sprzężeniem zwrotnym**, złożone z warstwy wejściowej, ukrytej i wyjściowej,
- **perceptrony wielowarstwowe**, posiadające możliwość uczenia się przez propagację wsteczną, posiada wiele warstw ukrytych, stosowany jest w tym projekcie
- **rekurencyjne sieci neuronowe**, pobierające dane wejściowe z poprzednich sekwencji za pomocą pętli sprzężenia zwrotnego

Drugim jest **las losowy**, który jest modelem uczenia maszynowego, łączącym dane wyjściowe wielu drzew decyzyjnych w celu uzyskania jednego wyniku. Do uzyskania wyniku dla konkretnego obiektu wejściowego, proces decyzyjny rozpoczyna się od węzła głównego i przechodzi przez drzewo, aż dotrze do liścia zawierającego wynik. W każdym węźle ścieżka, którą należy podążać, zależy od wartości cechy dla konkretnego obiektu wejściowego. Model radzi sobie dobrze nawet w przypadku brakujących danych, z drugiej strony nie mogą one przekroczyć zakresu wartości zmiennej docelowej stosowanej w treningu (Roßbach, 2018).

Do zbudowania dobrego modelu predykcji danych konkretnego typu, należy mieć odpowiedni punkt odniesienia, żeby ocenić jego jakość i wiedzieć czy model o danych parametrach jest dobry, czy też jego wyniki są niewystarczająco precyzyjne. W projekcie wykorzystane zostaną metryki dedykowane dla problemów regresyjnych.

Najpopularniejszą metryką, jest **MSE** (błąd średniokwadratowy), mierzy on średnią kwadratów błędów, jest funkcją ryzyka odpowiadającą oczekiwanej wartości kwadratu straty po błędzie, jego wadą natomiast jest to, że jest mało odporny na wartości odstające, może też prowadzić do niedoszacowania modelu jeśli wartości błędów są mniejsze niż 1. Jest często stosowany, ponieważ jego błąd maleje wraz ze wzrostem zbioru danych (Aishwarya, 2022). Bliźniaczą do niego metryką jest **MSLE**, który mierzy średni błąd kwadratowy logarytmów różnic między prognozami a rzeczywistymi danymi, zastosowanie logarytmu powoduje, że większe błędy pojawiają się, gdy zmienna rzeczywista jest niedoszacowana niż gdy jest przeszacowana (Saxena, 2019). Kolejną metryką jest **MAE** (średni błąd bezwzględny), oblicza on różnicę między każdą prognozą a rzeczywistą wartością, a następnie uśrednia bezwzględne wartości tych różnic, miernik ten traktuje błędy w sposób liniowy, co oznacza, że nie jest tak wrażliwy na odstające wartości, jak MSE (Hiregoudar, 2020). Stosowany jest też **współczynnik determinacji R²**, który jest miarą używaną do oceny stopnia dopasowania modelu regresji do danych rzeczywistych, informuje, jak dobrze model regresyjny wyjaśnia zmienność zmiennej zależnej w odniesieniu do jej średniej, im bliżej wartości 1, tym lepiej model pasuje do danych (Hiregoudar, 2020).

Wyniki badań

1. Wybór punktów

W celu przeprowadzenia kompletnej analizy zanieczyszczeń na terenie całego województwa wybrano punkty w taki sposób, aby nie były za blisko siebie, ale też nie za daleko. Serwis Airly posiada w swojej bazie ponad 900 czujników fizycznych w rejonie województwa małopolskiego, na poniższej mapie Fig. 3.1. jest przedstawione ich rozmieszczenie, wraz ze stopniem sąsiedztwa. Jednakże, dziennie można pobrać maksymalnie dane ze 100 czujników, dodatkowo należy uwzględnić fakt, że nie wszystkie czujniki są w pełni sprawne i w danym okresie w ogóle nie działają, dlatego należało zawęzić wybór do jak najlepszych stu czujników.

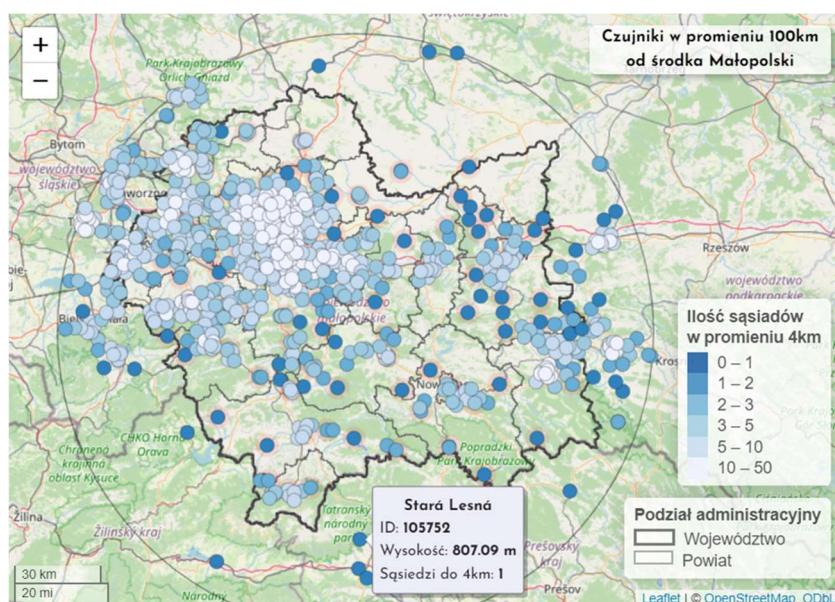


Fig. 3.1. Mapa czujników Airly w rejonie województwa małopolskiego z naniesioną informacją o liczebności sąsiednich punktów

Przy pierwszym wyborze punktów było sprawdzane, jaki procent niepustych pomiarów zanieczyszczenia PM_{2.5} lub PM₁₀ zawierają dane punkty, następnie do analizy brano pod uwagę inny zbiór 100 punktów i powtarzano procedurę. Spośród zbadań punktów wybrano te, które zawierały najwięcej pomiarów niepustych, natomiast punkty, które zawierały nieco mniejszą, ale satysfakcjonującą liczbę obserwacji, a leżały w obszarze, w którym, w bliskiej odległości, nie było innych punktów, również były wybierane.

2. Uzyskanie danych korzystając z Airly i OpenMeteo API

Do projektu wymagane były dane meteorologiczne i o zanieczyszczeniach powietrza. Zanim przeprowadzono badania, wykorzystano zbiór danych historycznych obejmujących zakres około 20 miesięcy w latach 2021-2022, zawierające obserwacje tylko dla pyłu PM₁, PM_{2.5} i PM₁₀ wraz z podstawowymi danymi meteorologicznymi w okolicy Krakowa. Dane wykorzystano, aby wcześniej zbudować aplikację do wizualizacji, przeprowadzić analizę korelacji oraz testy modelów. Poniżej, na mapie Fig. 3.2., przedstawiono położenie punktów obejmujących te dane, punkty nie obejmują całego obszaru województwa, ale dla wcześniejszych analiz i testów to wystarczy.

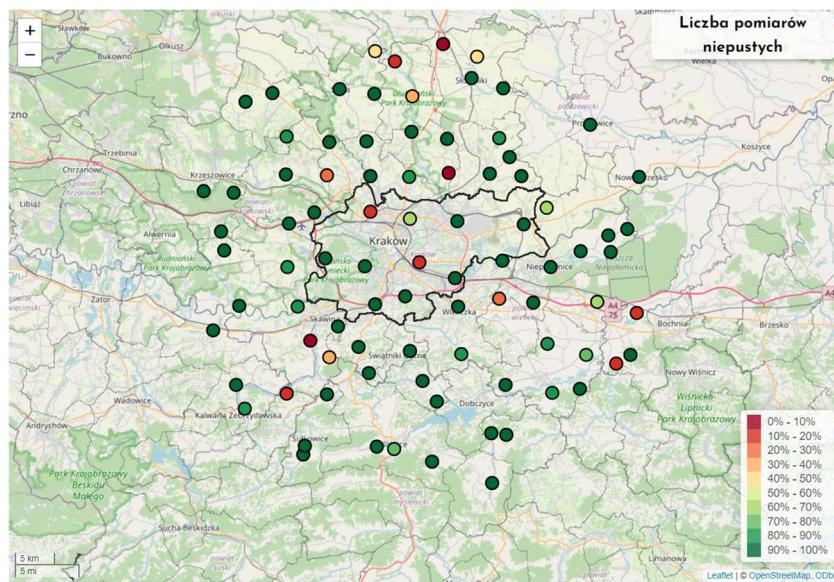


Fig. 3.2. Zestawienie wykorzystanych danych z informacją o procentowej ilości pomiarów niepustych (13 000 godzin)

W ramach badań, przez ponad miesiąc, korzystając z API Airly oraz open-meteo.com, pobierano dane o meteorologii i zanieczyszczeniach powietrza w granicach województwa małopolskiego. Airly dostarcza dane o zanieczyszczeniach pyłami, NO₂, SO₂, CO, O₃ oraz temperaturę, wilgotność i ciśnienie powietrza zredukowane do poziomu morsa. Dane uzyskiwane były z rzeczywistych pomiarów fizycznych czujników rozmieszczonych w województwie, w związku z dziennym limitem pobierania danych, ograniczono się do 100 punktów.

Wybór punktów priorytetyzował pomiary pyłów PM_{2.5} oraz PM₁₀, dlatego to one zawierają najwięcej obserwacji niepustych. Nie udało się zebrać dobrej ilości pomiarów o innych gazach, dlatego posłużą one tylko jako dane poglądowe. Tabela 3.1. prezentuje, ile informacji udało się zebrać w okresie od 5 grudnia 2023 do 17 stycznia 2024. Dane te są bardzo często niekompletne i najwięcej niepustych informacji dostępne jest dla pyłów zawieszonych, dane meteorologiczne zostały później uzupełnione przez pomiary OpenMeteo.

Tab. 3.1. Tabela przedstawiająca kompletność danych pobranych z Airly API.

CZYNNIK	ILOŚĆ OBSERWACJI	PROCENT OBSERWACJI NIEPUSTYCH	PROCENT POKRYCIA CZASOWEGO
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%
CIŚNIENIE AIRLY	53462	63.80%	88.31%
WILGOTNOŚĆ AIRLY	52594	62.76%	88.31%
TEMPERATURA AIRLY	52594	62.76%	88.31%

3. Wizualizacja i wstępne wnioski wyników badań

Na ilustracji Fig. 3.6. zaprezentowane są uśrednione dane meteorologiczne dla całego okresu, dla każdego punktu. W niektórych przypadkach zachodzą pewne prawidłowości, które mogą mieć wpływ na poziom zanieczyszczenia w powietrzu. Ciśnienie i temperatura wykazują podobne zależności, na południe maleją, jest to spowodowane tym, że na południu województwa, wysokość bezwzględna jest większa. Prędkość wiatru jest średnio najwyższa na północy województwa, co może być spowodowane większym przewiewem, niż w dolinach górskich, ale zbiór danych nie jest na tyle duży, żeby to wystarczająco dobrze wykazać. Zmienność intensywności opadów w tym przypadku jest akurat wynikiem, tego, że w danym okresie opady były intensywniejsze na zachodzie, dopiero wieloletnia analiza pozwoliłaby wykazać, pewne stałe zależności, np. że w górach opady są z reguły intensywniejsze (Carpenter, 2018).

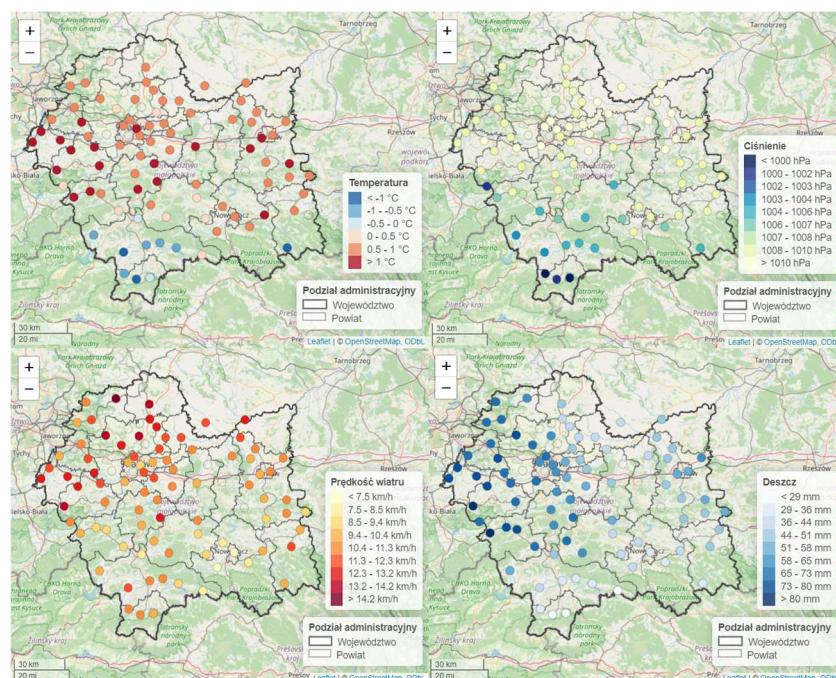


Fig. 3.6. Wizualizacja otrzymanych danych meteorologicznych

Poniższa mapa (Fig. 3.7.) przedstawia uśrednione wartości stężenia pyłów, widać że w rejonie Krakowa i Tarnowa oraz zachodniej części województwa, stężenie pyłu jest wyższe, natomiast południowa część z reguły ma mniejsze stężenie pyłów. Wartym uwagi spostrzeżeniem jest to, że punktowo, większe miejscowości położone w głębszych dolinach górskich (Nowy Sącz, Nowy Targ, Sucha Beskidzka), posiadają najwyższe wartości zanieczyszczeń, podczas kiedy sąsiednie czujniki nie wykazują tak wysokich wartości, świadczy to trudności przepływu powietrza w takim terenie (airly.org, 2024).

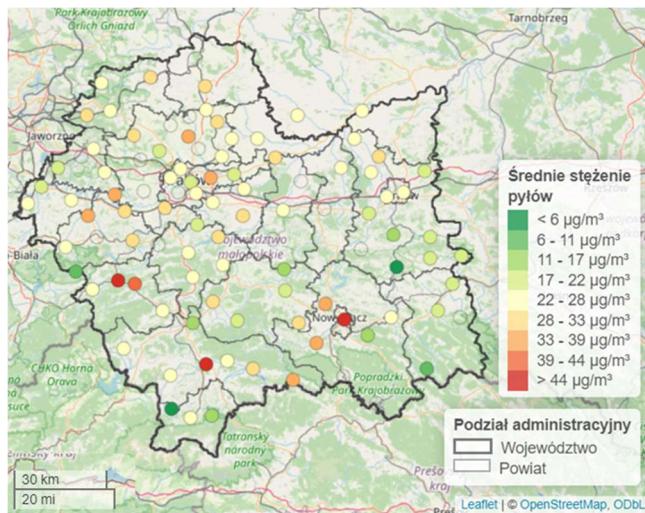


Fig. 3.7. Wizualizacja danych zanieczyszczeń pyłami

Airly dostarcza również dane o innych aerozolach, jednak są one niewystarczająco liczne, aby przeprowadzić istotną analizę geostatystyczną, pomiary odbywają się w nie więcej niż 5 punktach.

4. Badanie korelacji

W kolejnym kroku przeprowadzono analizę zależności zanieczyszczeń od czynników meteorologicznych, aby później wybrać te, które w największym stopniu zależne są poziomu zanieczyszczenia. Należy zaznaczyć, że niektóre czynniki mogą wykazywać zależności nieliniowe o czym trzeba pamiętać przy porównywaniu korelacji i tworzeniu modeli prognozy. Poniższe mapy zależności zostały przeanalizowane, tylko w celu znalezienia ogólnego związku, czyli które czynniki z reguły powodują wzrost, a które spadek wartości zanieczyszczenia. Póki co nie da się na razie zauważać zależności nieliniowych, co może wykazać, przeprowadzona później, bardziej szczegółowa analiza.

Na ilustracji Fig. 3.8. przedstawiona jest mapa korelacji między zanieczyszczeniami. Pyły zawieszone korelują ze sobą najmocniej, ponieważ ich skład jest taki sam, jedynie różnią się rozmiarem. Inne gazy korelują już w mniejszym stopniu, tlenek węgla i tlenki azotu wykazują silną korelację dodatnią z pyłami, podobnie tlenek siarki, choć nieco słabszą. Ozon natomiast koreluje ujemnie z każdym zanieczyszczeniem, co może mieć związek z tym, że jest to zanieczyszczenie wtórne.

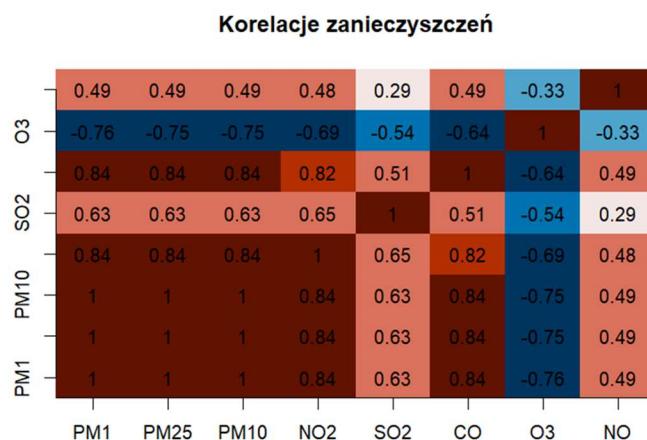


Fig. 3.8. Wizualizacja zależności między zanieczyszczeniami powietrza

Wiedząc, już jak korelują ze sobą zanieczyszczenia, można przyjrzeć się ich korelacjom ze zmiennymi meteorologicznymi, ilustracja Fig. 3.9. przedstawia mapę zależności między nimi. Można zauważać silniejsze korelacje zanieczyszczeń z ciśnieniem i temperaturą. Zmienne intensywności opadów i wiatru nie korelują co prawda tak dobrze, ale mogą tutaj wystąpić zależności nieliniowe. Dodatkowo w analizie zostały również uwzględnione dane takie jak punkt rosy, temperatura gleby, itp., które mają bezpośredni związek z bliźniaczymi im czynnikami i ich korelacje są podobne, ale słabsze, ponieważ dotyczą one bezpośrednio gleby czy też poziomu gruntu, dlatego nie są tak efektywne w korelacji z zanieczyszczeniami.

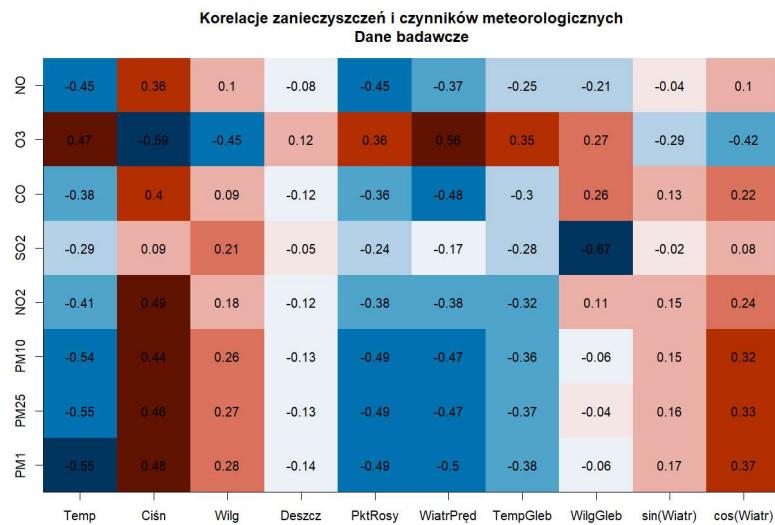


Fig. 3.9. Wizualizacja zależności między zanieczyszczeniami powietrza a czynnikami meteorologicznymi

Wykresy na Fig. 3.10. przedstawiają zależność temperatury i zanieczyszczenia pyłem PM₁₀. Ujemna korelacja ma, przede wszystkim w województwie małopolskim, związek z sezonem grzewczym, im niższa temperatura, tym więcej ludzi będzie paliło w piecu w swoich domach, poza tym częściej korzysta się wtedy z samochodów zamiast roweru czy pójścia pieszo, co jest skutkiem nasilenia się emisji zanieczyszczeń do atmosfery. Dodatkowo, w zimie, jest duża szansa pojawiania się inwersji, które mogą blokować zanieczyszczenia w dolnych warstwach atmosfery (Elminir, 2005). Warto też zaznaczyć, widoczne tu, „rozjeżdżanie się” punktów z lewej i prawej strony wykresów, jest to spowodowane tym, że dla każdego argumentu (w tym przypadku wartości temperatury) jest wyciągana średnia arytmetyczna z wartości zanieczyszczenia przypadającym danej temperaturze, dla argumentów skrajnych dane jest mniej wartości do uśrednienia, dlatego są one mniej reprezentatywne i bardziej narażone na odchylenia czy anomalie.

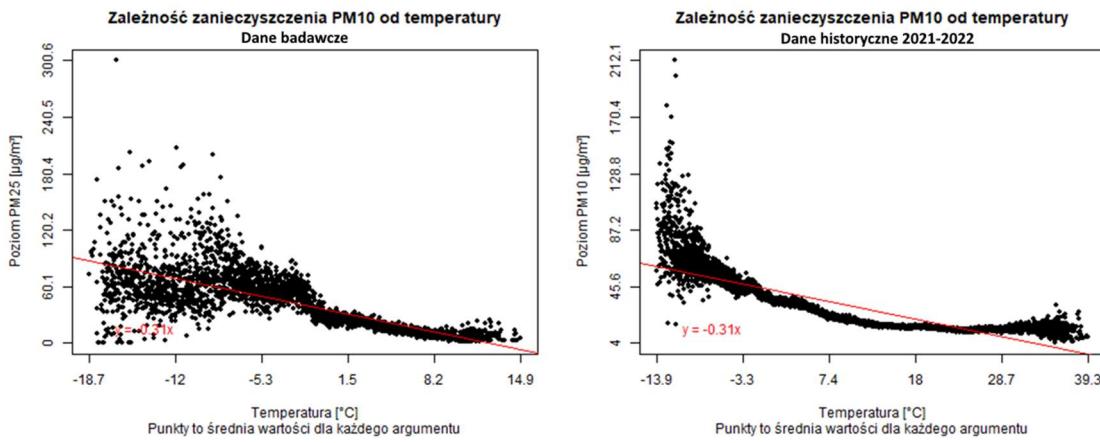


Fig. 3.10. Wykresy zależności temperatury powietrza i gleby do zanieczyszczenia pyłem PM₁₀

Jak widać na ilustracji Fig. 3.11., w przypadku opadów deszczu, również zachodzi ujemna korelacja z zanieczyszczeniami. Zależność ta jest lepiej widoczna na danych historycznych, w danych badawczych liniowość jest silnie zaburzona, może być to spowodowane tym, że okres badań dotyczył grudnia i stycznia, gdzie większość opadów to śnieg, który może nie mieć takiego wpływu na pył jak deszcz. Ujemną korelację można tłumaczyć przez tzw. depozycję atmosferyczną, tj. w trakcie opadów deszczu woda zbiera cząstki zanieczyszczeń i sprowadza je na powierzchnię ziemi, dzięki temu zanieczyszczenia są usuwane z powietrza, im większa intensywność opadów, tym ten proces jest efektywniejszy (Koch, D., J. Park, A. Del Genio, 2003).

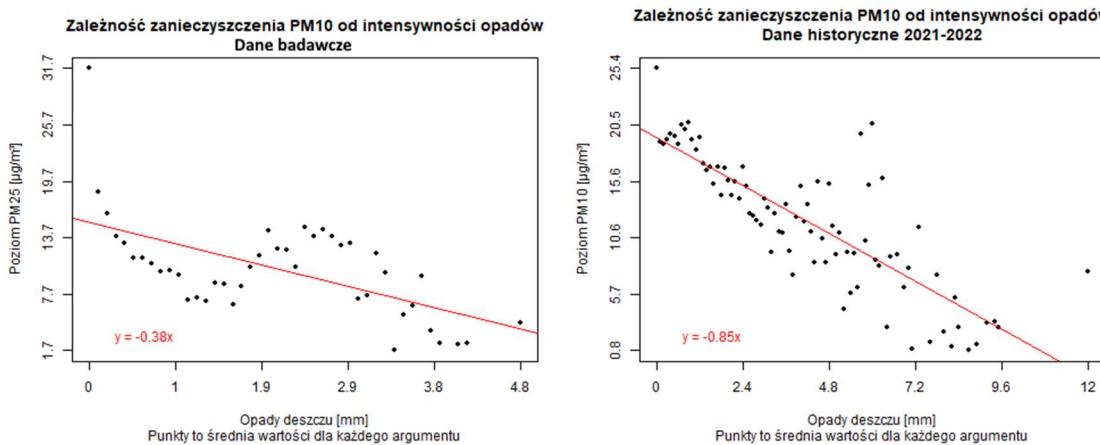


Fig. 3.11. Wykres zależności intensywności opadów deszczu do zanieczyszczenia pyłem PM₁₀

Ilustracja Fig. 3.12. przedstawia korelację pyłu PM₁₀ z prędkością wiatru, która w obu przypadkach jest silna i ujemna, czyli im silniejszy wiatr, tym niższe zanieczyszczenie. Wiatr w znacznym stopniu przyczynia się do zmniejszenia ilości zanieczyszczeń w powietrzu przez skuteczne rozpraszanie cząstek zanieczyszczeń, uniemożliwiając ich gromadzenie się w jednym miejscu, jednocześnie przynosząc świeże powietrze z obszarów o niższych stężeniach (Elminir, 2005).

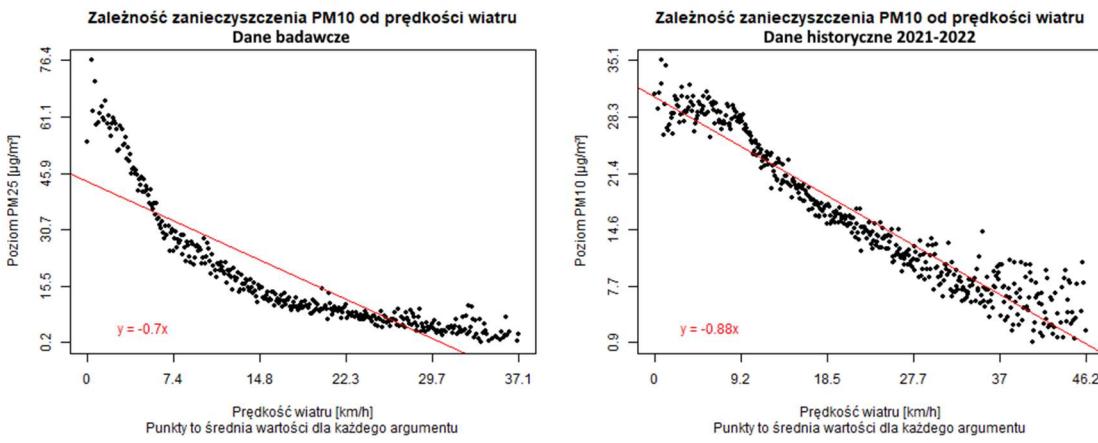


Fig. 3.12. Wykres zależności prędkości wiatru do zanieczyszczenia pyłem PM₁₀

W przypadku ciśnienia, zachodzi dodatnia korelacja z zanieczyszczeniem PM₁₀, co przedstawia ilustracja Fig. 3.13. W przypadku danych badawczych, zależność jest słabsza niż w przypadku danych historycznych i przez liczne zaburzenia liniowości można wręcz jej nie dostrzec, jest to prawdopodobnie spowodowane zbyt krótkim przedziałem czasowym, żeby uwydatnić dobrze tą zależność, również inne czynniki atmosferyczne mogą mieć wpływ na „wypłaszczenie” tej zależności. Wysokie ciśnienie może sprzyjać wysokiemu zanieczyszczeniu ze względu na występującą wtedy stabilność warunków atmosferycznych, które zazwyczaj występują właśnie przy wysokim ciśnieniu, wiąże się to najczęściej z brakiem opadów deszczu czy słabym ruchem wiatru, wówczas wtedy powietrze jest mniej podatne na ruch, co sprzyja gromadzeniu się zanieczyszczenia w jednym miejscu (Amos P.K. Tai, 2010).

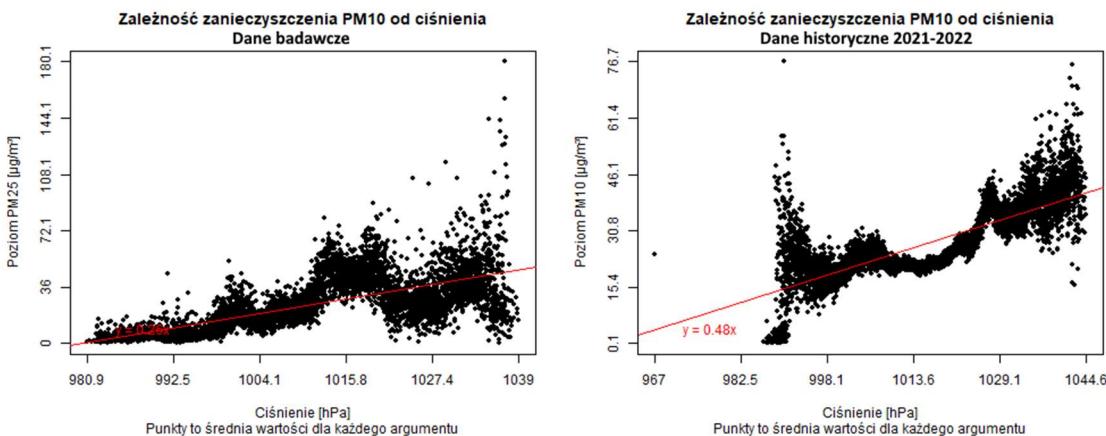


Fig. 3.13. Wykres zależności ciśnienia powietrza do zanieczyszczenia pyłem PM₁₀

W przypadku wilgotności, której zależność przedstawiona jest na ilustracji Fig. 3.14., co prawda zachodzi pewna, dodatnia korelacja, ale jest ona niewielka, dla danych badawczych jest wręcz znikoma. Za dodatnią korelacją może stać fakt, że jej wysoki poziom ułatwia zachodzenie niektórych procesów chemicznych w atmosferze, które prowadzą do powstawania nowych zanieczyszczeń, może też prowadzić do bardziej stabilnych warunków atmosferycznych. Z drugiej strony wysoka wilgotność ułatwia rozpuszczanie niektórych gazów, jak tlenki azotu czy siarki, jednak z wykresów wynika, że bardziej przeważają skutki korelacji dodatniej (Elminir, 2005). Zależność jest na tyle mała w przypadku danych badawczych, że może być nieopłacalne jej użycie w prognozach.

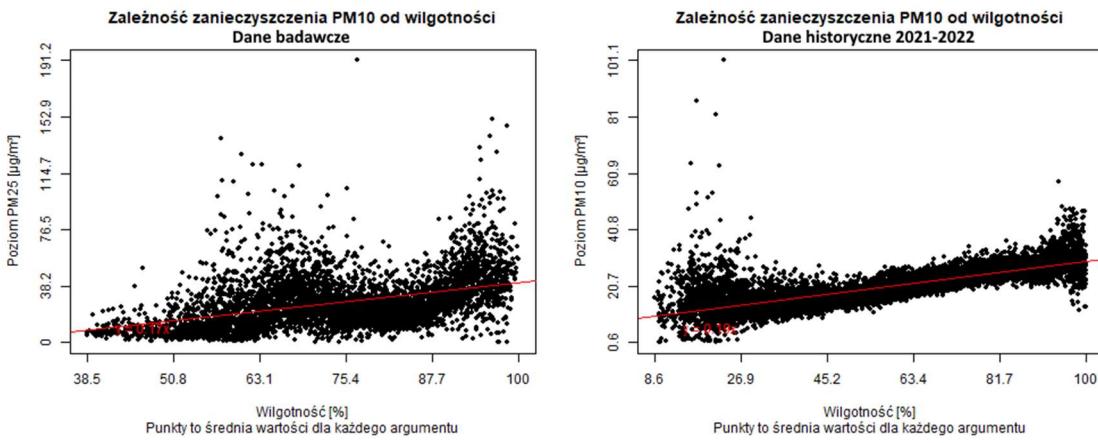


Fig. 3.14. Wykres zależności wilgotności do zanieczyszczenia pyłem PM_{10}

Zanieczyszczenie jest też ściśle skorelowane z godziną dnia, co przedstawia ilustracja Fig. 3.15., co prawda nieliniowo, ale zależność ta jest mocno widoczna, dla wszystkich zanieczyszczeń. Zależność ta jest powiązana z ludzkim rytem dnia i nocy, można zauważyć, że oba wykresy łączą wysokie wartości zanieczyszczeń w godzinach wieczornych (17-20), co spowodowane jest zwiększym ruchem ulicznym, gdy ludzie wracają z pracy do domów, co szczególnie przekłada się na wzrost zanieczyszczenia tlenkami azotu, jest to też pora, w której najczęściej rozpalane są domowe piece, które emitują zanieczyszczenia do atmosfery. Dla wszystkich zanieczyszczeń oprócz ozonu, w godzinach porannych następuje wzrost wartości, co ma związek ze zwiększym ruchem samochodowym, gdy ludzie wyjeżdżają z mieszkań do pracy, do szkoły, jednak nie widać tutaj takich wartości zanieczyszczeń jak wieczorem. Z kolei w okolicach godziny 12 pojawiają się najmniejsze wartości stężenia, co ma związek z brakiem ruchu ulicznego. Ozon posiada minima i maksima odwrotnie do innych zanieczyszczeń co jest spowodowane jego wtórnym występowaniem w powietrzu, w trakcie dnia, w wyniku reakcji fotochemicznych z NO i O_3 powstaje NO_2 , w nocy natomiast, w wyniku fotolizy, znów się regeneruje (S. Shelton, G. Liyanage, ..., 2022).

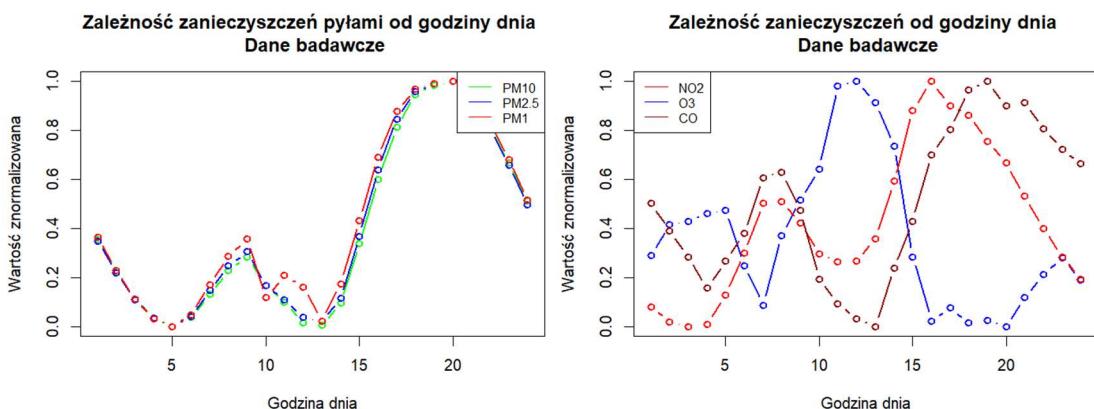


Fig. 3.15. Wykres zależności godziny dnia do zanieczyszczeń pyłami PM_1 , $PM_{2.5}$ i PM_{10}

Podobnie dla dnia tygodnia, z wykresów na ilustracji Fig. 3.16. można wywnioskować, że zachodzą zależności nieliniowe, w przypadku wszystkich zanieczyszczeń, oprócz ozonu, w środę występuje maksimum, w piątek zanieczyszczenie jest najmniejsze, a na weekend następuje niewielki wzrost. Jedną z przyczyn może być zróżnicowanie aktywności przemysłowej czy intensywności usług, w niektórych zakładach może być zwiększa w dniach pośrodku tygodnia i zmniejszona w piątki, kiedy to może być

kończona szybciej praca przed weekendem. Powodem może też być ruch komunikacyjny, gdzie ruch, zwłaszcza jeśli chodzi o dojeżdżanie do pracy, będzie większy, z kolei w piątek, niektórzy mogą kończyć pracę szybciej, co powoduje rozładowanie ruchu komunikacyjnego w czasie (He, 2023).

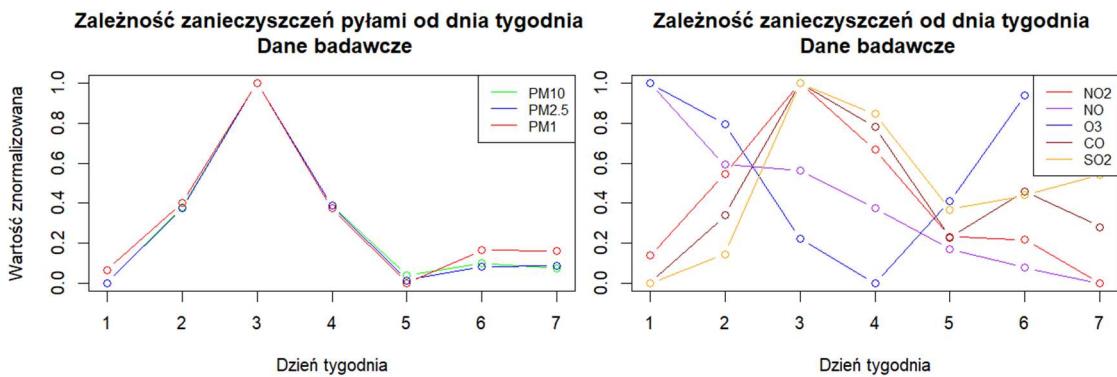


Fig. 3.16. Wykres zależności dnia tygodnia do zanieczyszczeń (dane z jednego miesiąca)

Warto też porównać te zależności na różnych przedziałach czasowych, co przedstawione jest na ilustracji Fig. 3.17. Dla danych 20-miesięcznych zależność jest ona bardzo dobrze widoczna, jednak, jeśli prognoza ma zostać wykonana na podstawie danych z jednego tygodnia, to warto zrezygnować ze wzięcia tej zmiennej pod uwagę, ponieważ może mieć ona negatywny wpływ na jakość prognozy, ciężko, na wykresie po lewej, doszukać się jakichś zależności.

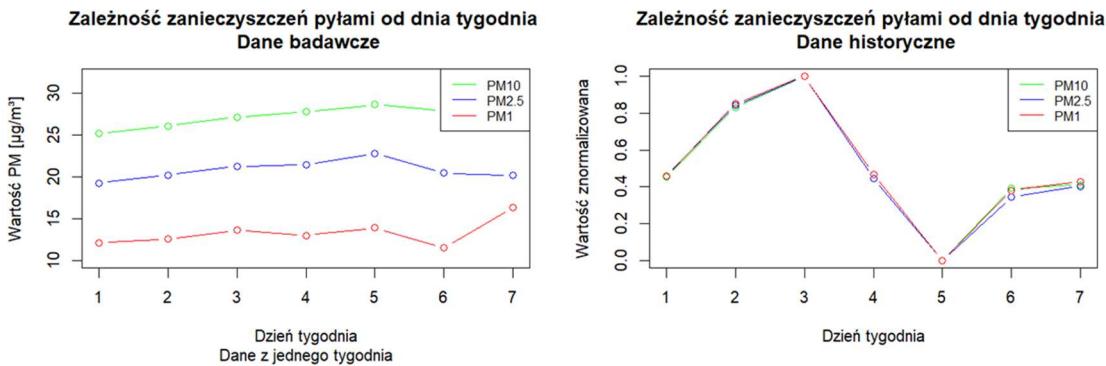


Fig. 3.17. Porównanie zależności dnia tygodnia do zanieczyszczeń pyłami dla danych z jednego tygodnia i 20-miesięcznych

Kolejną istotną zmienną jest kierunek wiatru, jej zależność z zanieczyszczeniem jest nieliniowa i przybiera charakter sinusoidy, co przedstawione jest na wykresach (Fig. 3.18.) Oba wykresy różnią się, ponieważ zależność tej zmiennej w dużej mierze jest powiązana z położeniem geograficznym, a oba zbiory zawierają dane dla nieco innych obszarów. Powiązany z lokalizacją wpływ tutaj ma pobliskie ukształtowanie terenu, położenie zakładów przemysłowych, czy skupisk miejskich, z których potencjalnie, może być nawiewane zanieczyszczone powietrze.

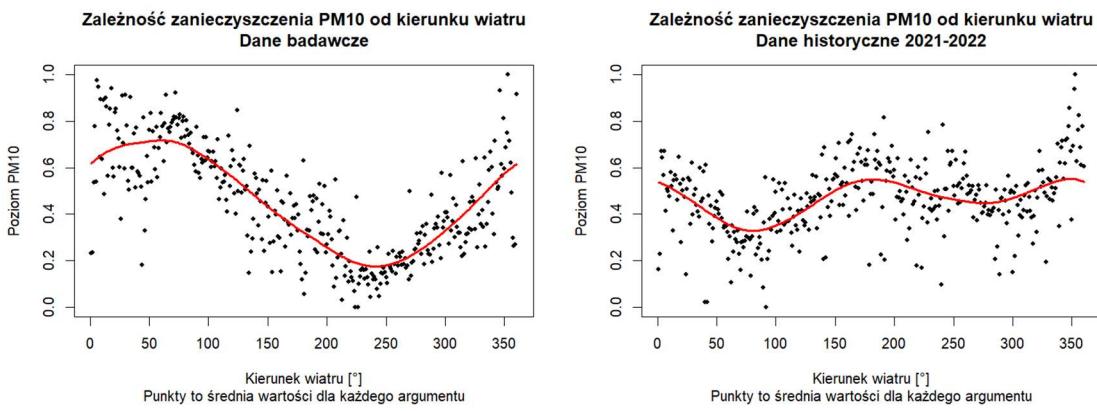


Fig. 3.18. Wykres zależności kierunku wiatru do zanieczyszczeń pyłami PM₁, PM_{2.5} i PM₁₀. Argumenty przedstawiają kierunek, z którego wieje wiatr

Mapy na ilustracji Fig. 3.19., prezentują kierunki wiatrów dla maksymalnych i minimalnych poziomów stężenia, dla rejonu Krakowa, na podstawie danych historycznych, w okresie 20 miesięcy. Można zauważać pewne wzorce w położeniu geograficznym, przy wartościach maksymalnych zanieczyszczenia, kierunki wiatru bliżej Krakowa i Kotliny Sandomierskiej, układają się przeważnie z zachodu, co może być spowodowane napływem zanieczyszczeń z uprzemysłowionego Śląska, z kolei wartości odwrotne kierunków wiatru mogą być podyktowane położeniem w terenie górkowatym, gdzie grzbiety górskie mają wpływ na kierunek wiatru. W przypadku wartości najmniejszych, Kraków stanowi, w niektórych miejscach, przejście od kierunku południowego do północnego. Mimo iż wydaje się, że Kraków otrzymuje najwięcej zanieczyszczeń z obrzeży, a nie na odwrót, to warto zaznaczyć, że przyczyną emisji napływowej do Krakowa jest nie tylko wiatr, ale również topografia terenu w rejonie Krakowa i okres sezonu grzewczego. Dane zawierają obserwacje z ponad roku, w okresie letnim najwięcej zanieczyszczeń tworzy się w Krakowie przez ruch samochodowy w centrum, przedmieścia nie generują emisji ze spalania w piecach domowych, wówczas są one mniej zanieczyszczone, przez co wiatr „do wewnętrz” Krakowa, przyczynia się do niższych wartości zanieczyszczenia w jego rejonie poza sezonem grzewczym.

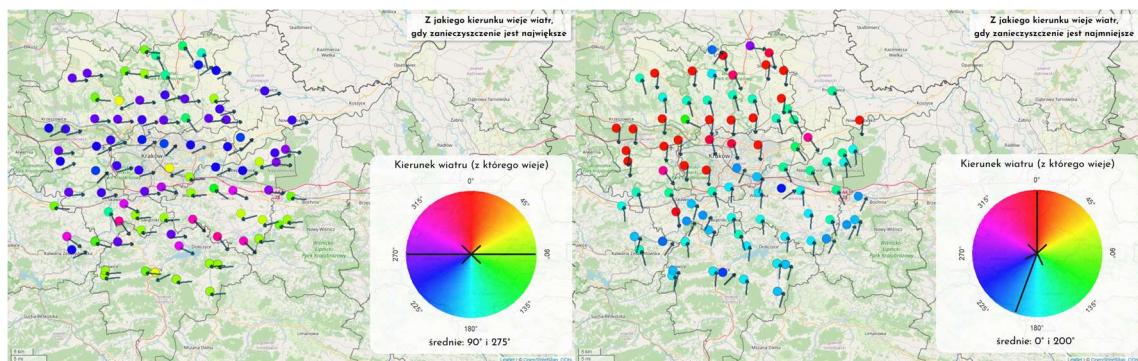


Fig. 3.19. Zestawienie map pokazujące, dla każdego punktu, dla jakich wartości kierunku wartości, zanieczyszczenie jest odpowiednio największe i najmniejsze. Na podstawie danych historycznych, obejmujących tylko rejon Krakowa

Dla analizy całego województwa widać, na ilustracji Fig. 3.20., szczególnie dla wartości minimalnego zanieczyszczenia, że dla większości punktów kierunek wiatru jest bardzo podobny, gdzie nie występują inne wartości, ale najprawdopodobniej jest to spowodowane położeniem w terenie górzystym, którego grzbiety blokują wiatr z konkretnego kierunku. Jednak w środkowej i północnej części województwa,

kierunki wiatru są niemal identyczne. Jeśli chodzi o kierunek wiatru przy najwyższym zanieczyszczeniu, to tutaj nie ma tak równych wartości, jednak w większości przypadków jest to kierunek zbliżony do wschodniego. Po uśrednieniu wyników, w obu przypadkach wyjdą wartości niemal odwrotne. Biorąc pod uwagę, że są to dane z jednego miesiąca, ciężko tutaj doszukać się jakiegoś stałej przyczyny dla tych zależności. Kierunek wiatru przy minimalnych wartościach zanieczyszczenia można解释为 napływem mas ciepłego i wilgotnego powietrza z zachodu, które może prowadzić do obniżenia poziomu zanieczyszczenia, natomiast kierunki wiatrów przy najwyższym zanieczyszczeniu, są prawdopodobnie bardziej podyktowane położeniem względem miast lub zakładów przemysłowych, chociaż, biorąc pod uwagę średni kierunek jako północno-wschodni, można też uznać, że miało na to wpływ powiaty polarne – kontynentalne napływanie z Rosji, które jest silnie ożebione w dolnych warstwach i dość suche (B. Skowera, J. Wojkowski, 2009).

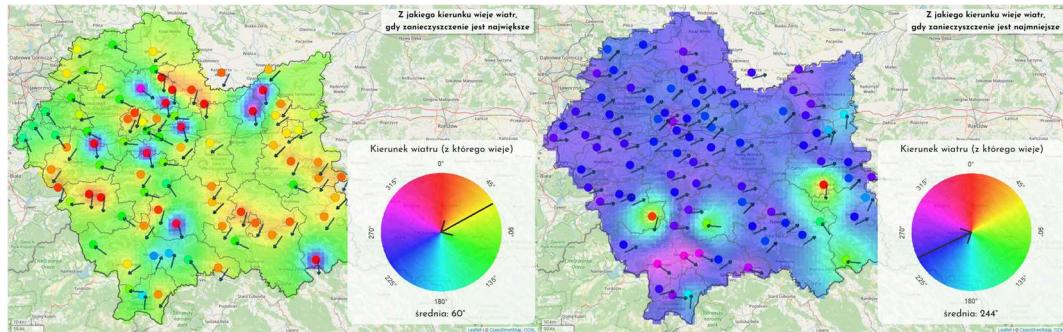


Fig. 3.20. Zestawienie map pokazujące, dla każdego punktu, dla jakich wartości kierunku wiatru, zanieczyszczenie jest odpowiednio największe i najmniejsze. Na podstawie danych otrzymanych z badań, obejmujących całe województwo.

W kontekście przestrzennym, na terenie województwa małopolskiego, korelacje ze zmiennymi meteorologicznymi wyglądają niemal identycznie na całym obszarze, jedynie korelacja wilgotności wydaje się zmieniać w przestrzeni, jednak są to wartości bardzo bliskie zeru, można więc założyć, że są to korelacje nieistotne. Grafika Fig. 3.21. ukazuje zmienność korelacji w przestrzeni.

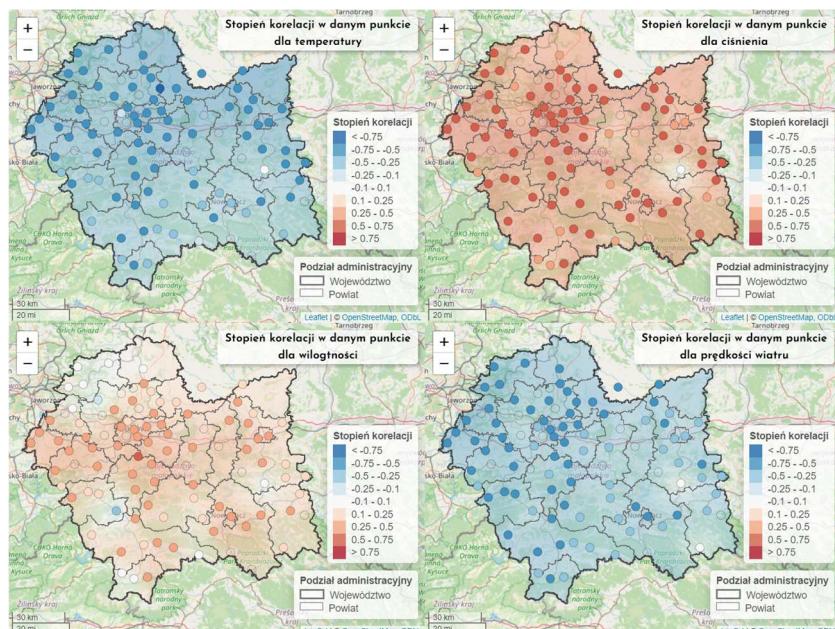


Fig. 3.21. Mapy obrazujące korelacje wybranych zmiennych meteorologicznych z pyłem zawieszonym PM₁₀ dla każdego punktu

5. Testowanie prognozy zanieczyszczeń – budowa modeli

Po przeprowadzeniu preprocessingu, do dyspozycji pozostają gotowe do modelowania dane, z których tworzony jest zbiór treningowy (długość 6 dni) i testowy (24-godzinny). Tak przygotowane zestawy danych można użyć w prognozie, w przypadku sieci neuronowej i lasu losowego, w języku R, korzysta się z wbudowanych funkcji „neuralnet” i „randomForest”. Funkcje te przyjmują jako pierwszy argument funkcję zależności, gdzie wpisuje się zmienną wyjściową i wejściowe. Kolejne parametry są indywidualne dla każdego modelu, las losowy przyjmuje liczbę drzew, a sieć neuronowa, strukturę sieci. Prognoza, z użyciem danego modelu, jest wykonywana funkcją predict().

W dalszym kroku wykonano analizę jakości modelów w zależności od różnych ustawień parametrów, testy wykonano na zanieczyszczeniu PM₁₀. Poniższe wykresy (Fig. 3.22.) prezentują porównanie danych testowych (rzeczywistych) i prognozowanych, jest to graficzna metoda oceny jakości modelu, im bliżej punkty leżą czerwonej linii, tym lepsza jakość modelu, tzn., że wartości prognozowane są bardzo zbliżone do wartości rzeczywistych. Wykresy przedstawiają modele dobre jakościowo, błąd MAE o wysokości mniejszej niż 5% świadczy o dobrym dopasowaniu danych.

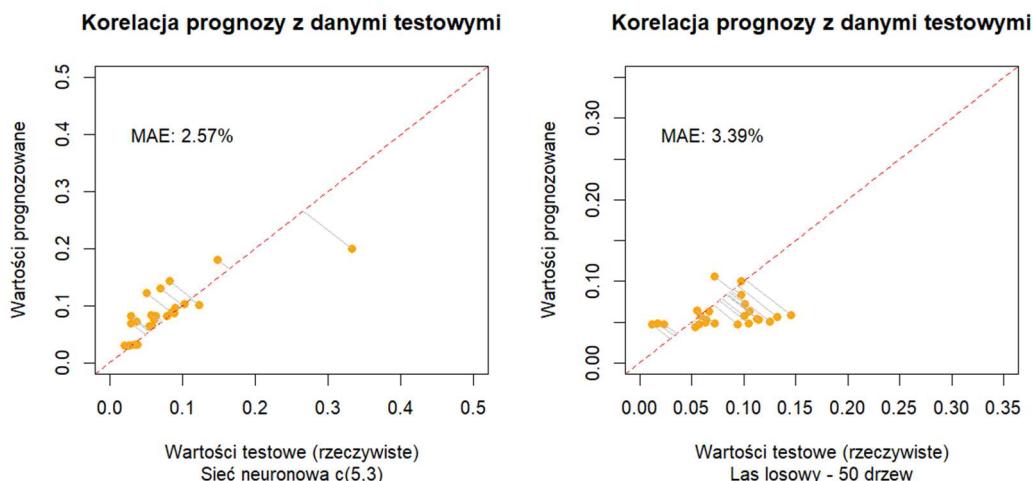


Fig. 3.22. Zestawienie zależności wartości prognozowanych z wartościami rzeczywistymi na przykładzie lasu losowego i sieci neuronowej. Oba modele w tym przypadku są dobrej jakości

Warto też przyjrzeć się modelom pod kątem zmienności jakości w zależności od punktu, co przedstawione jest na Fig. 3.23. Zarówno, w przypadku obu modeli, taka zmienność zachodzi, dla niektórych punktów wartości te są średnio wyższe, dla niektórych niższe. Szczególnie modele lasów losowych to dobrze obrazują, warto też zwrócić uwagę, że rozstęp oceny jakości modeli dla punktów, w tym przypadku jest bardzo mały, przeciwnie do sieci neuronowej, gdzie jakości różnią się znacznie. Może być to spowodowane tym, że las losowy jest bardziej stabilnym modelem w porównaniu do sieci neuronowej, która jest bardziej skomplikowanym modelem, ale za to bardziej elastycznym (Roßbach, 2018).

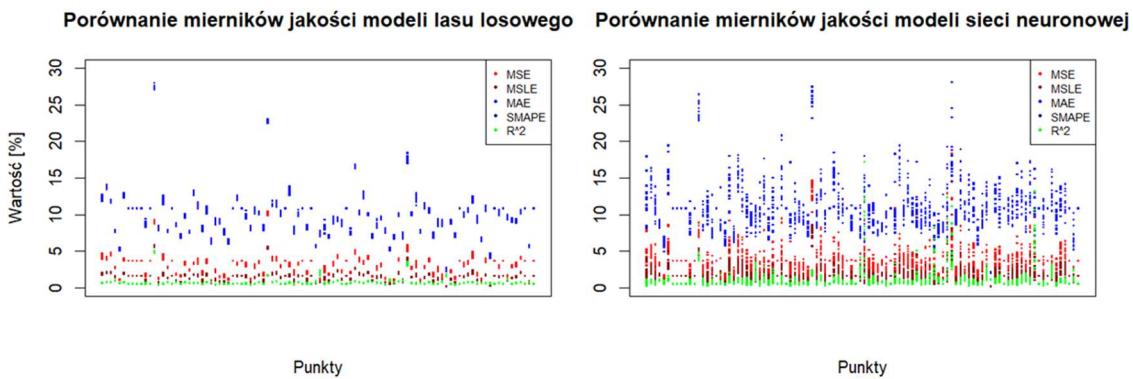


Fig. 3.23. Porównanie modeli lasu losowego i sieci neuronowej na podstawie ich oceny innymi miernikami, sortując po punktach pomiarowych.

Na grafice Fig. 3.24. zaprezentowane są wartości zanieczyszczenia PM₁₀ w czasie, dla punktów, dla których modele miały średnio najlepszą lub najgorszą jakość. Dla punktu, na podstawie którego modele radzą sobie średnio najlepiej da się zauważać, w przeciwieństwie do tego ze złą jakością, zmienność jest stabilniejsza, widać wyraźne dobowe zmiany zanieczyszczenia, oczywiście występują co jakiś czas skoki wartości stężenia, ale są one dużo stabilniejsze, niż w przypadku danych z drugiego wykresu. Wartości na wykresie po prawej wydają się zmieniać mniej stabilnie, można zauważać, kilkugodzinne skoki, a także ogromne zmiany dobowe dla początkowych godzin obserwacji, dodatkowo dostępnych jest mniej obserwacji (kończą się one ok. 100 godzin szybciej) i da się też zauważać błąd pomiarowy, jakim są wartości zerowe przy 800 godzinie.

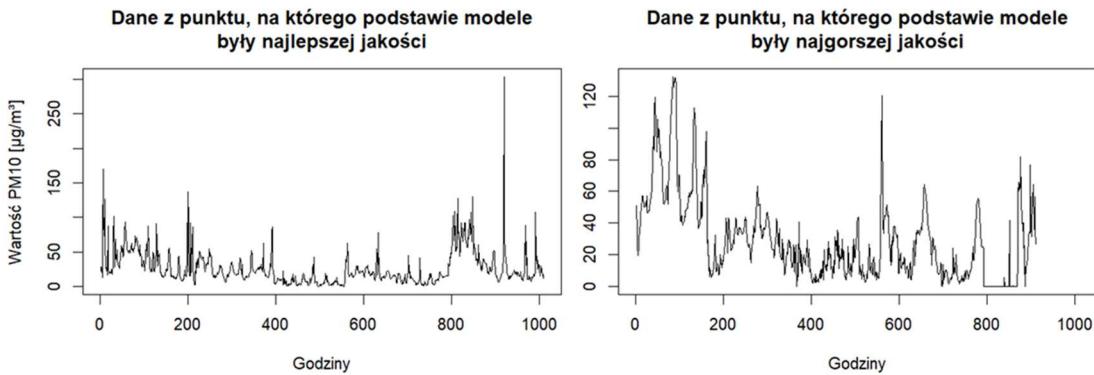


Fig. 3.24. Wykresy zmienności pyłu PM10 dla punktów, na podstawie których modele tworzyły odpowiednio najlepsze i najgorsze prognozy

Gdy sieć neuronowa uwzględnia zmienne, które korelują nieliniowo, należy o tym ją wcześniej poinformować, wówczas model będzie „wiedział”, że ma do czynienia ze zmiennymi skorelowanymi nieliniowo. Na wykresach ilustracji Fig. 3.25. prezentowane jest porównanie modeli bez uwzględnienia nieliniowości i po prawej z jej uwzględnieniem. Las losowy nie potrzebuje informacji o liniowości modelu, ponieważ działa on inaczej. Gdy uwzględni się nieliniowość jakość podnosi się kilkukrotnie, dla tego należy to uwzględnić przy tworzeniu prognozy. Warto też zwrócić uwagę, że las losowy zawsze ma lepsze wyniki niż sieć neuronowa, spowodowane może być to tym, że las losowy lepiej radzi sobie z mniejszymi zbiorami danych, ale też z brakami lub błędymi danymi (Roßbach, 2018).

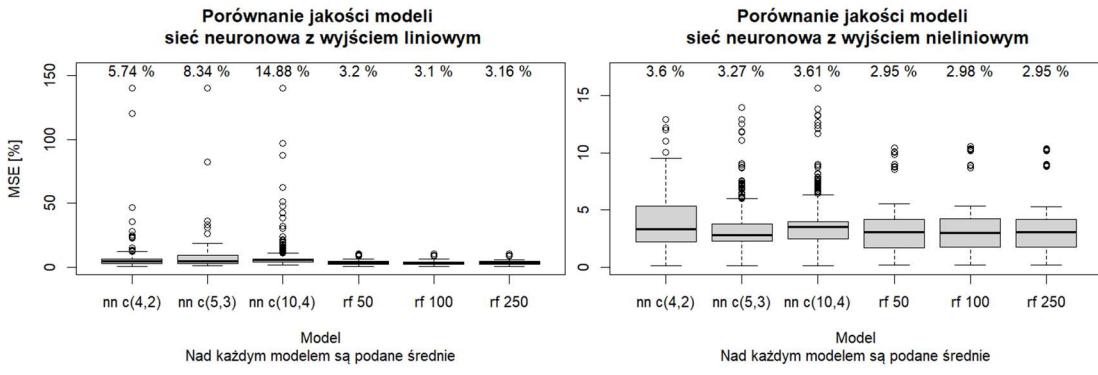


Fig. 3.25. Porównanie jakości modeli sieci neuronowej uwzględniającą tylko zależności liniowe z uwzględniającą zależności nieliniowe miernikiem MSE (83 powtórzenia). Warto zwrócić uwagę na różnicę skali pionowej

Następnie porównano modele, które biorą za zmienną wejściową dzień tygodnia i te które nie biorą, wyniki zwizualizowane są na ilustracji Fig. 3.26. Model stworzony na bazie danych obejmujących ponad miesiąc, który uwzględnia dzień tygodnia, będzie radził sobie nieco lepiej w prognozie, niż model tego nie uwzględniający. W takim okresie, zależność zanieczyszczenia od tej zmiennej zaczyna nabierać więcej istotności i dlatego użycie jej w tym przypadku jest jak najbardziej zasadne, chociaż niekoniecznie obowiązkowe. Modele tworzone na bazie danych obejmujących tydzień, już nie mają tyle danych o relacji z dniem tygodnia, prognoza na podstawie tylko jednej próby (jednego tygodnia) nie ma w tym przypadku sensu, ponieważ prowadzi ona do wypłaszczenia danych prognozowanych, przez co przeważnie jakość modelu spada, dlatego warto zbudować model bez uwzględniania tej zmiennej, zaoszczędzi to czasu obliczeniowego i być może zwiększy jakość modelu.

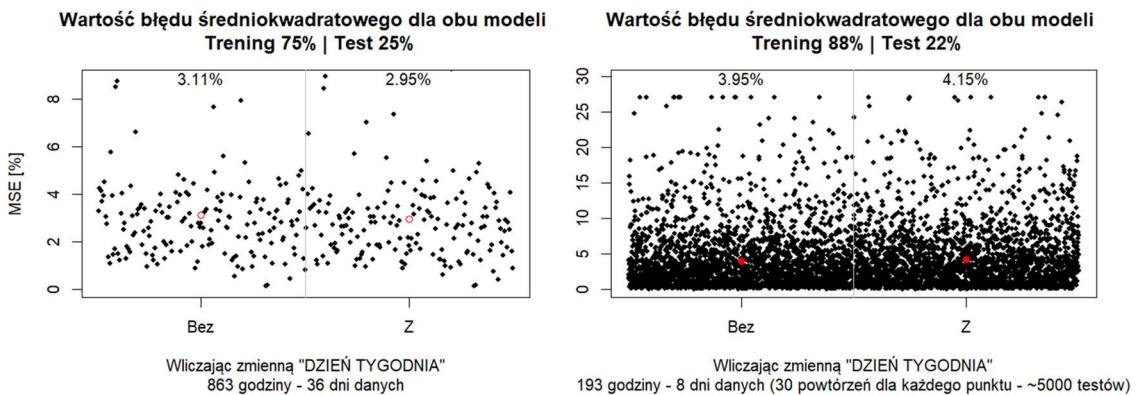


Fig. 3.26. Porównanie jakości modeli uwzględniających zmienią „Dzień tygodnia” z miesięcznym i tygodniowym przedziałem czasowym

Przy dobieraniu parametrów do modelu, należy też pamiętać o wyborze odpowiedniego podziału na zbiór treningowy i testowy. W ramach sprawdzenia jakości modelu w zależności od wielkości tych zbiorów, wykonano analizy na różnych przedziałach czasowych. Na grafice Fig. 3.27., wykres po lewej obrazuje jakości modelów na podstawie danych miesięcznych, jak widać przy tak licznych zbiorze obserwacji, warto mieć nieco mniejszy zbiór treningowy, ponieważ przy większych zbiorach może dojść do przeuczenia, przez co model będzie mniej efektywny i źle dopasowywał się do danych. Można wywnioskować, że prognoza danych z jednego miesiąca na jeden dzień nie jest najlepszym pomysłem, ale warto też zaznaczyć, że mniej danych do testowania może oznaczać również wyższą wartość miernika jakości modelu, ponieważ wnioskuje się na mało reprezentatywnej próbce, a też zbiór testowy nie może

być wielkości rzędu jednej obserwacji. W przypadku danych o długości tygodnia, zasadniejszym jest prognozowanie na jedną dobę. Modele mające zbiory treningowe mniejsze niż 80% są w większości niedouczone, mają niewystarczającą próbke do nauczenia się zależności, dlatego ich jakość jest gorsza niż w przypadku modelu z 88 % zbiorem treningowym. W przypadku krótszych danych, ciężej jest przeuczyć model, dlatego można sobie pozwolić na większe zbiory treningowe względem testowych.

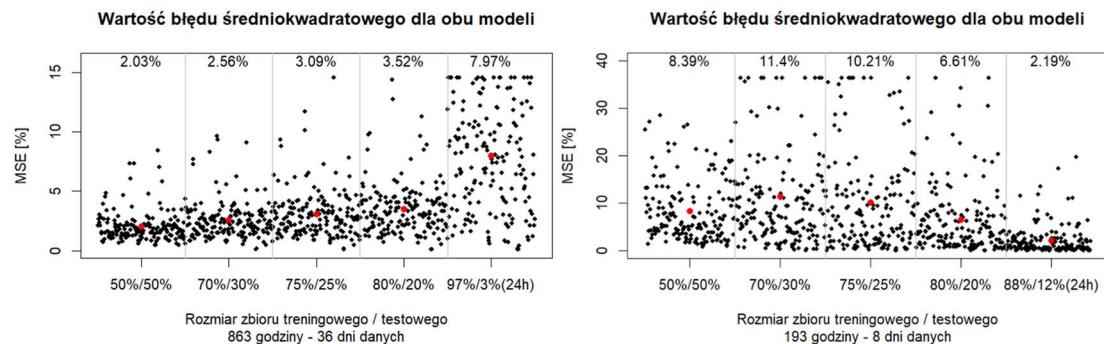


Fig. 3.27. Porównanie jakości modeli z miesięcznym przedziałem czasowym z różnym stosunkiem rozmiarów zbiorów treningowego i testowego

Wykonano też testy na modelach uwzględniających i nieuwzględniających wartości odstających, porównanie ich jakości jest przedstawione na wykresie Fig. 3.28. W wielu przypadkach, w fazie preprocessingu, obcina się wartości odstające, które mogą być przeszkadzającymi w modelowaniu anomaliami lub błędami pomiarowymi. Jednak w trakcie testowania modeli nie zauważono żadnych, nierealnych anomali, które mogłyby być wynikiem błędów pomiarowych. Końcowo, lepiej spisały się modele, które nie odcinały wartości odstających.

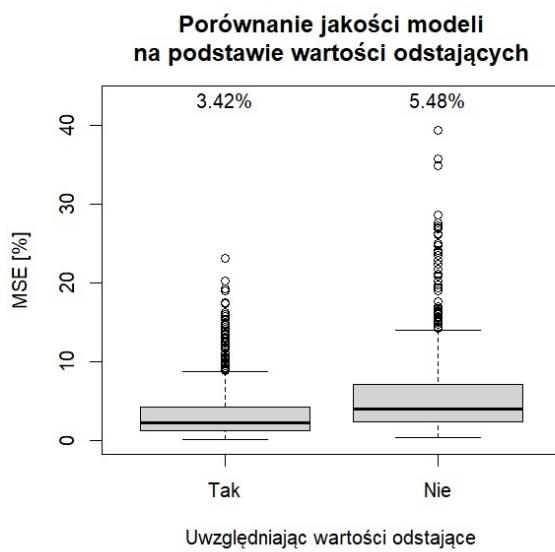


Fig. 3.28. Porównanie jakości modeli z tygodniowym przedziałem czasowym na podstawie uwzględnienia wartości odstających

W celu dalszej optymalizacji modelów, wykonano testy na kombinacjach różnych zmiennych wejściowych. Pod uwagę wzięto: ciśnienie na poziomie morza, wilgotność względna, temperaturę powietrza, intensywność opadów, prędkość i kierunek wiatru, godzinę dnia i dzień tygodnia. Stworzono 36 kombinacji, uwzględniających 6, 7 i 8

zmiennych wejściowych na wszystkie sposoby, testy wykonano na pyłach PM₁₀ i PM_{2.5}, każdy punkt to 300 modeli testowych. Poniższe wykresy, na ilustracji Fig. 3.29., przedstawiają, jeśli dana zmienna jest uwzględniona w modelu, czy też nie.

Analizując wykresy, można zauważyc, że zmienność w jakości nie jest taka duża. Modele uwzględniające 7 zmiennych wejściowych mają średnią jakość lepszą niż w przypadku tych uwzględniających 6 lub 8 zmiennych. Nie można jednoznacznie stwierdzić, że któraś zmienna w sposób znaczny pogarsza czy polepsza jakość modelu, ponieważ każda znajduje się w nieco gorszej lub lepszej kombinacji. Wydaje się, że ciśnienie i intensywność opadów, znajdują się w większości dobrych jakościowo kombinacji, z kolei w gorszych modelach zdaje się być godzina dnia, ale nie jest to mocno zauważalny element, wobec czego nie można, na podstawie tej analizy, jednoznacznie wskazać, który czynnik sprawdza się lepiej w prognozie.

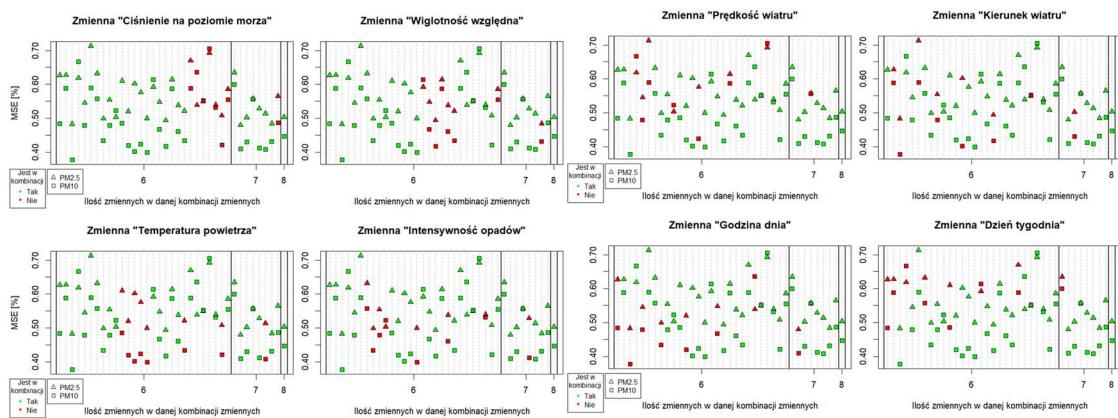


Fig. 3.29. Porównanie jakości modeli z różną kombinacją zmiennych wejściowych ze zmienną wyjściową PM₁₀ i PM_{2.5}

6. Prezentacja prognozy zanieczyszczenia pyłami w aplikacji webowej

Na podstawie powyższych analiz, do prognozy, pod uwagę wzięte zostały: temperatura i ciśnienie powietrza, godzina dnia, intensywność opadów oraz prędkość i wiatru, ponieważ to one wykazują najlepszą korelację z zanieczyszczeniem powietrza. W przypadku lasu losowego użyto 50 drzew, głównie, żeby uniknąć przeuczenia, zauważono też, że po tej liczbie błęd już się nie zmniejszał, tylko stabilizował. W przypadku sieci neuronowej architektura składa się z dwóch warstw ukrytych, jednej z pięcioma neuronami i drugiej z trzema. Prognozy wykonano dla zanieczyszczeń pyłami, które zawierały najwięcej niepustych obserwacji, PM_{2.5} oraz PM₁₀, użyto zbioru treningowego o długości 6 dni oraz 24-godzinnego zbioru testowego.

Poniższe zrzuty ekranu, na Fig. 3.30., prezentują prognozę zanieczyszczenia pyłem PM_{2.5} dla punktu w Gorlicach oraz pyłem PM₁₀ dla punktu w Krakowie, wykresy zestawiają wartości danych testowych (na szaro) i prognozowanych (na zielono), a także użytych danych treningowych (na czarno). Do szacowania jakości zostały użyte dwa mierniki, aby całkowicie lepiej ocenić dane modele. Dodatkowo, z prawej strony, znajdują się też wykresy zestawiające bezpośrednio dane testowe z prognozowanymi z podanym współczynnikiem korelacji Pearson'a.

W przypadkach, gdy wykorzystywany jest las losowy, wszystkie mierniki wskazują na to, że jakość prognozy jest lepsza niż dla sieci neuronowej, co wynikło również z wcześniejszej analizy, dlatego można uznać, że prognoza sieci neuronowej w tym przypadku jest błędna. Warto też zwrócić uwagę na nienaturalnie niskie wartości miernika

MSE w odniesieniu do wykresów i miernika MAE, mimo tego, że wartości są poniżej 1% (co wskazywałoby na bardzo dobre dopasowanie danych do siebie), to dane prognostyczne i testowe zauważalnie się ze sobą rozjeżdżają. W tym przypadku MSE jest takie niskie, dlatego że w odniesieniu do całego zbioru danych, prognozowane wartości są bliskie zeru. Mając na uwadze, że we wcześniejszych dniach wartości są znacznie wyższe, podczas normalizacji danych, te maksymalne przyjmą wartość bliską 1, natomiast wartości na końcu zbioru danych są prawie zerowe. Biorąc pod uwagę fakt, że metryka ta używa kwadratów różnic, to błędy bliskie zeru są jeszcze bardziej zmniejszane, dlatego średnia wychodzi taka mała. Z kolei metryka MAE jest odporna takie małe wartości, ponieważ nie podnosi do potęgi tych różnic.

Oceniając od strony wizualnej, można powiedzieć, że w przypadku czujnika w Krakowie (górnego wykresy), model dobrze przewidział skok wartości zanieczyszczenia powietrza 17 stycznia, z kolei w przypadku czujnika w Gorlicach nie nastąpił taki wzrost wartości stężenia, pojawiły się natomiast drobne wahania wartości, co model również w pewnym stopniu przewidział.

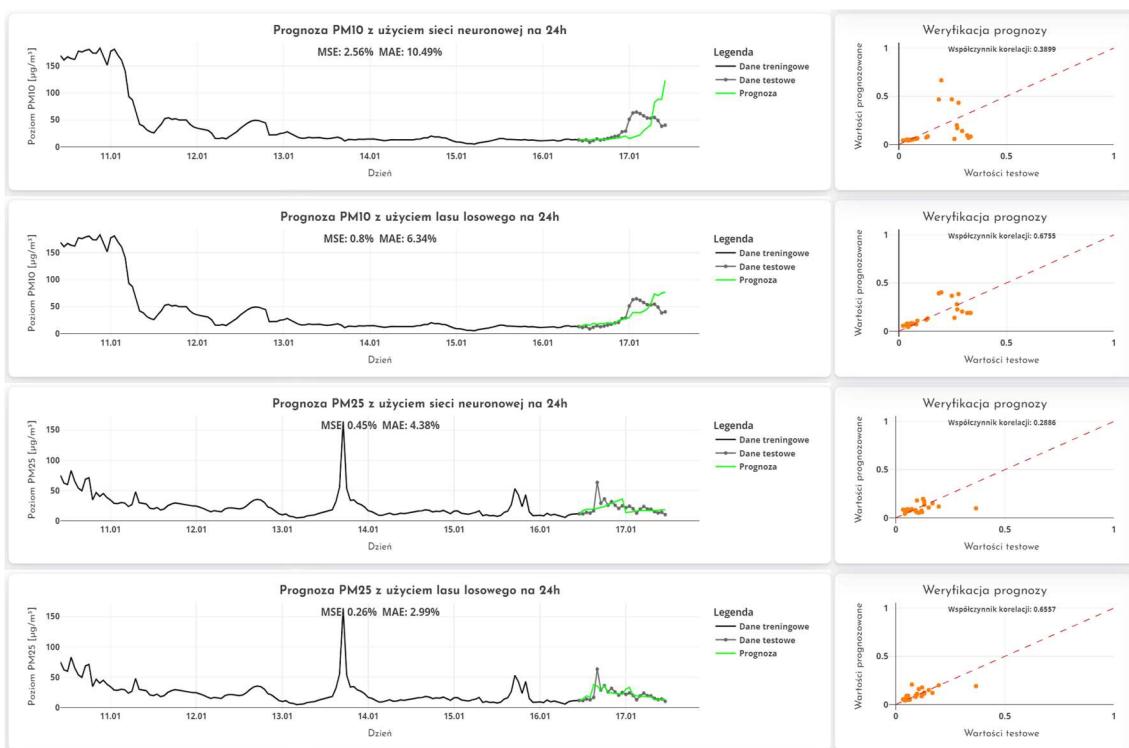


Fig. 3.30. Wizualizacja prognoz zanieczyszczenia pyłem PM_{2.5} dla punktu w Gorlicach oraz PM₁₀ dla punktu w Krakowie (ulica Bulwarowa) z użyciem lasu losowego i sieci neuronowej

Wnioski

Z prezentacji prognozy można wywnioskować, że modele zbudowane na bazie danych z jednego tygodnia, z odpowiednim dobraniem parametrów i zmiennych wejściowych potrafią dorównać zaawansowanym modelom prognostycznym, widniejącym na popularnych portalach mapowych. Zauważono, że w przypadku prognozy na tygodniowym zbiorze danych, losowy radzi sobie lepiej niż sieć neuronowa. Ważnym było również dobranie odpowiednich rozmiarów zbioru testowego i treningowego, aby uniknąć zarówno niedouczenia jak i przeuczenia. Prognoza ta jednak jest efektywna na krótkim, 24-godzinnym okresie, ponieważ model nie posiada informacji o wpływie sezonów roku, dnia tygodnia i innych. Modele działają dobrze, dopóki dostarczone dane nie zawierają błędów lub nie wystąpią jakieś anomalie na okres prognozowany, które mogą wykryć tylko modele o zasięgu kontynentalnym i większym. Podobnie jak z pogodą w różnych częściach świata, poziom zanieczyszczenia w różnych miejscach będzie przeważnie wyższy lub niższy niż gdzie indziej, jednak udało się wykazać, że dla całego obszaru zmienne meteorologiczne i zanieczyszczenia powietrza korelują podobnie. Największy wpływ na zanieczyszczenie wykazuje zmiana godziny dnia, a także temperatura, są to zależności w dużej części pośrednie, ponieważ, w większości to reakcja człowieka na zmianę tych czynników prowadzi bezpośrednio do zmiany poziomu zanieczyszczeń. Duży wpływ ma również siła wiatru oraz intensywność opadów, które już w sposób bezpośredni wpływają na stężenie zanieczyszczeń.

Literatura

1. Konwencja Nr 148 dotycząca ochrony pracowników przed zagrożeniami zawodowymi w miejscowościach pracy, z 20 czerwca 1977, (Dz.U. 2004 Nr 29, poz. 255).
2. A. Krenker, J. Bešter, A. Kos. (styczeń 2011). *Artificial Neural Networks - Methodological Advances and Biomedical Applications*.
3. <https://airindex.eea.europa.eu/Map/AQI/Viewer/#> (styczeń 2024),
dostęp: styczeń 2024
4. <https://airly.org/pl/gdzie-jest-najwiecej-smogu-w-polsce-przeczytaj-analize-airly/>,
dostęp: styczeń 2024
5. <https://airly.org/pl/tlenek-azotu-trujace-skladniki-smogu-cz-1/>, dostęp: styczeń 2024
6. <https://airly.org/pl/jakie-dzialania-antysmogowe-moga-wdrozyc-gminy/>,
dostęp: styczeń 2024
7. Aishwarya, B. (luty 2022). *Regression Metrics - Of all metrics why MSE?* Pobrano z lokalizacji <https://www.linkedin.com/pulse/regression-metrics-all-why-mse-aishwarya-b/>
8. Amos P.K. Tai, L. J. (październik 2010). Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity 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. (marzec 2018). *How Do Mountains Affect Precipitation?* Pobrano z lokalizacji sciencing.com: <https://sciencing.com/do-mountains-affect-precipitation-8691099.html>
11. Ćwik, P. (kwiecień 2017). *Dwutlenek siarki. W Polsce źle, na Bałkanach gorzej.* Pobrano z lokalizacji <https://smoglab.pl/dwutlenek-siarki-w-polsce-zle-na-balkanach-gorzej-czym-truje-nas-smog-4/>
12. Elminir, H. K. (listopad 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/>,
dostęp: styczeń 2024
14. He, R.-R. (luty 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. (sierpień 2020). *Ways to Evaluate Regression Models.* Pobrano z lokalizacji <https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70>,
dostęp: styczeń 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. Pobrano: styczeń 2024
18. <https://posit.co/products/open-source/rstudio/>, dostęp: styczeń 2024
19. https://powietrze.gios.gov.pl/pjp/content/health_informations, dostęp: styczeń 2024
20. <https://powietrze.gios.gov.pl/pjp/content/show/1000919>, dostęp: styczeń 2024
21. <https://powietrze.malopolska.pl/program-ochrony-powietrza/>, dostęp: styczeń 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. (Czerwiec 2019). *What's the Difference Between RMSE and RMSLE?* Pobrano z lokalizacji <https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmse-935c6cc1802a>, dostęp: styczeń 2024
25. Uchwała nr LXXV/1102/23 Sejmiku Województwa Małopolskiego z dnia 20 listopada 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. (styczeń 2006). Air Quality Guidelines. Global Update 2005.
28. <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics>, dostęp: styczeń 2024
29. <https://www.krakow.pl/213212,1962,230530,powietrze,faq.html>, dostęp: styczeń 2024
30. https://www.krakow.pl/aktualnosci/274334,26,komunikat,krakow_przeciwny_lagodzeniu_u_chwal_antysmogowych.html, dostęp: styczeń 2024
31. https://www.malopolska.uw.gov.pl/default.aspx?page=tlenek_wegla, dostęp: styczeń 2024
32. <https://www.r-project.org/>, dostęp: styczeń 2024
33. Yun Bai, Yong Li, Xiaoxue Wang, Jingjing Xie, Chuan Li. (2016, Maj). Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. *Atmospheric Pollution Research*, 7(3), 557-566.