Michał WILKOSZ[1], Łukasz POLOCZEK[2], Piotr CZECH[3],
Mariola SATERNUS[4], Henryk KANIA[5]

Opiekunowie naukowi: Piotr CZECH[3], Mariola SATERNUS[4],
Henryk KANIA[5]

DOI: https://doi.org/10.53052/9788366249837.28

PROGNOZOWANIU ZANIECZYSZCZENIA POWIETRZA ATMOSFERYCZNEGO PRZY UŻYCIU SZEREGÓW CZASOWYCH I RÓŻNYCH TYPÓW SZTUCZNYCH SIECI NEURONOWYCH

Streszczenie: Artykuł miał na celu przedstawienie wyników badań dotyczących prognozowania wartości zanieczyszczenia powietrza atmosferycznego na podstawie zarejestrowanych na stacji pomiarowej danych. W badaniach wykorzystano różnego typu sztuczne sieci neuronowe. Przeanalizowano możliwość prognozowania różnych zanieczyszczeń. W badaniach wykorzystano szeregi czasowe uwzględniając prognozę dnia następnego. Badania zrealizowano w ramach w ramach projektu Politechnika Śląska jako Centrum Nowoczesnego Kształcenia opartego o badania i innowacje POWR-03.05.00-00-Z098/17-00.

Słowa kluczowe: sieć neuronowa, szeregi czasowe, zanieczyszczenie powietrza

FORECASTING AIR POLLUTION WITH THE USE OF TIME SERIES AND DIFFERENT TYPES OF ARTIFICIAL NEURAL NETWORKS

Summary: The aim of the article was to present the results of research on air pollution value forecasting based on the weather data recorded at the measuring station. Various types of artificial neural networks were used in the research. The possibility of forecasting various pollutants was analyzed. The research used time series taking into account the forecast of

[1] Politechnika Śląska, Wydział Automatyki, Elektroniki i Informatyki, specjalność: macrofaculty – data science
[2] Politechnika Śląska, Wydział Automatyki, Elektroniki i Informatyki, specjalność: macrofaculty – data science
[3] Prof. dr hab. inż., Politechnika Śląska, Wydział Transportu i Inżynierii Lotniczej, email: piotr.czech@polsl.pl
[4] Dr hab. inż., prof. PŚ, Politechnika Śląska, Wydział Inżynierii Materiałowej, email: mariola.saternus@polsl.pl
[5] Dr hab. inż., Politechnika Śląska, Wydział Inżynierii Materiałowej, email: henryk.kania@polsl.pl

the next day. The research was carried out as part of the project of the Silesian University of Technology as a Modern Education Center based on research and innovation POWR-03.05.00-00-Z098 / 17-00.

Keywords: neural network, time series, air pollution

1. Wprowadzenie

Celem przeprowadzonych badań było określenie możliwości wykorzystania modeli do predykcji stężenia zanieczyszczeń powietrza pochodzących z motoryzacji, hutnictwa i palenisk domowych. Wykorzystano modele bazujące na sztucznych sieciach neuronowych. W ramach badań wykorzystano zarejestrowane dane zanieczyszczeń powietrza ze stacji pomiarowych województwa dolnośląskiego. Przetestowano różne metody prognozowania, rozpoczynając od prostych metod statystycznych, a kończąc na bardziej zaawansowanych metodach korzystających z sieci neuronowych. Zostały one wykorzystane do prognozowania szeregu czasowego. Uwzględniono prognozę dnia następnego.

2. Wybór stacji pomiarowej

Przed przystąpieniem do prognozowania zanieczyszczeń będących celem badań, przeprowadzono przegląd i wybór meteorologicznej stacji pomiarowej, która posłużyła za źródło danych. Dysponując danymi pochodzącymi z wielu stacji pomiarowych z województwa dolnośląskiego, wyboru odpowiedniej jednostki dokonano na podstawie analizy następujących właściwości stacji:

- Położenie. Ze względu na różne elementy składowe zanieczyszczeń pod uwagę były brane jedynie stacje znajdujące się w miastach. Miasta będąc centrum generowania się zanieczyszczeń pochodzących między innymi z pojazdów spalinowych, przemysłu oraz pozostałych źródeł niskiej emisji pozwalają w najbardziej racjonalny sposób na dokładne przedstawienie rozpatrywanego problemu.
- Liczba danych. Większość modeli zbudowanych z sieci neuronowych celem odpowiedniego nauczenia i uzyskania dobrych wyników wymaga bardzo dużej liczby danych. Zasada generalnie jest bardzo prosta – im więcej tym lepiej. W tym celu został dokonany przegląd najdłużej działających stacji, ponieważ to właśnie te jednostki przez długi okres działania mogły zgromadzić największe liczby danych.
- Kompletność danych. Posiadanie danych dobrej jakości pod kątem ich kompletności jest kolejnym istotnym czynnikiem branym pod uwagę przy wyborze stacji. Co prawda istnieją modele radzące sobie z brakującymi danymi, jednak powszechnie przyjmuje się wybór jak najbardziej kompletnych danych. Duża liczba brakujących wartości powoduje gorsze działanie niemalże każdego modelu. Stacje pomiarowe będąc urządzeniami zasilanymi za pośrednictwem energii elektrycznej, w przypadku jej braku nie prowadziły akwizycji danych czego następstwem były duże braki w przypadku praktycznie każdej z dostępnych stacji.

Kierując się wymienionymi założeniami wybrano stację pomiarową o krajowym kodzie „DsWrocWybCon". Stacja prowadząc pomiary metodą automatyczną pozwoliła na zgromadzenie licznych pomiarów, wśród których można wymienić tlenek węgla (CO), dwutlenek azotu (NO_2), tlenki azotu (NO_X), ozon, pył zawieszony PM2,5, pył zawieszony PM10, dwutlenek siarki (SO_2), benzen (C_6H_6), oraz ciśnienie atmosferyczne, temperatura powietrza, promieniowanie RAD, opad atmosferyczny, wilgotność względna powietrza, promieniowanie UVB, kierunek wiatru, prędkość wiatru. Przedstawiając wybraną stację należy omówić jak wypadała pod względem postawionych wymagań:

- Położenie. Stacja pomiarowa znajduje się przy ulicy Wybrzeże J. Conrada Korzeniowskiego, która leży niedaleko ścisłego centrum Wrocławia. Taka lokalizacja może być intersująca pod względem zanieczyszczeń powietrza atmosferycznego, szczególnie w zakresie poszczególnych parametrów, które ze względu na położenie powinny osiągać wysokie wartości.
- Liczba danych. Pod tym względem wybrana stacja pomiarowa była jedną z kilku, które rejestrowały dane od roku 2005 do roku 2017. Stacja ze względu na dużą liczbę dostępnych lat oraz liczne pomiary parametrów była jedną z wyróżniających się stacji. Całkowity zbiór danych zgromadzonych przez stację obejmował 113952 wierszy rozmieszczonych w 21 kolumnach.
- Kompletność danych. Najbardziej istotnym elementem tego kryterium była liczba brakujących wartości względem poszczególnych parametrów. Brakujące dane zostały pokazane w tabeli 1.

Tabela 1. Zestawienie brakujących danych

Parametr	Brakująca liczba pomiarów
HG (TGM)	113952
SO_2	88907
PM10	76622
Temperatura powietrza	65369
Opad atmosferyczny	65225
C_6H_6	55431
Kierunek wiatru	29177
Prędkość wiatru	27596
CO	27136
Wilgotność względna	22824
Promieniowanie UVB	22824
O_3	18075
Promieniowanie RAD	14435
NO_x	14083
NO_2	14081
Ciśnienie atmosferyczne	11295
PM2.5	8931
Godzina	0
Dzień	0
Miesiąc	0
Rok	0

Jak można zauważyć stacja pomiarowa pomimo posiadania możliwości odczytywania wilgotności HG (TGM) nie rejestrowała tych danych. Niektóre parametry takie jak m.in. SO_2 i PM10 były rejestrowane przez bardzo krótki okres czasu, co przełożyło

się na dużą liczbę brakujących wartości. Pomimo dużych braków w przypadku większości substancji, na stacji zarejestrowano dużą liczbę danych w zakresie istotnych substancji pod względem przeprowadzonych badań.

3. Eksploracyjna analiza danych

Rok 2016 był najbardziej kompletnym z dostępnych zarejestrowanych w wybranej stacji meteorologicznej. Z tego względu z powodzeniem mogła posłużyć jako zbiór danych do przeprowadzenia eksploracyjnej analizy danych. Jak powszechnie wiadomo okresem występowania wzmożonych ilości szkodliwych pyłów w Polsce są z pewnością okresy zimowe. Za pomocą prostych analiz zarejestrowanych substancji, można sprawdzić poprawność tej tezy.

3.1. Pył PM10

Pył PM10 to mieszanina zawieszonych w powietrzu cząsteczek, których średnica nie przekracza 10 μm. Swoją szkodliwość zawdzięcza zawartości takich elementów jak benzopireny, furany i dioksyny, czyli głównie rakotwórcze metale ciężkie. Norma średniego dobowego stężenia tego pyłu wynosi według WHO 50 μg/m^3, a roczna 20 μg/m^3. W przypadku Polski normy kształtują się następująco:

- poziom dopuszczalny: 50 μg/m^3.
- poziom informowania: 100 μg/m^3.
- poziom alarmowy: 150 μg/m^3.

Zarejestrowane dane można scharakteryzować następująco:

- liczba dni pomiarowych: 357,
- wartość średnia: 24,59,
- odchylenie standardowe: 19,10,
- wartość minimalna: 4,37,
- 25%: 12,08,
- 50%: 18,01,
- 75%: 138,54,

Jak można zauważyć wartość średnia przekracza roczne dopuszczalne normy. Brak 7 dni pomiarowych wynika z problemów rejestracji wartości przez stację. Na podstawie zarejestrowanych danych można wywnioskować, że obecność pyłu PM10 w powietrzu w zasadzie nie opuszcza pobliskich mieszkańców przez cały rok i potrafi nawet w dość ciepłych miesiącach roku osiągać niebezpieczne wartości. Dane poza widocznym szumem przedstawiają również sezonowość występowania wyższych wartości w okresie zimowym. W początkowych miesiącach roku można zauważyć szczyt odczytywanych wartości, który następnie w kolejnych miesiącach spada, gdzie w połowie roku osiąga najniższe wartości. Następnie znowu zaczyna rosnąć wraz z rozpoczęciem się chłodniejszych dni by znowu osiągnąć szczyty w nadchodzącym roku.

3.2. Pył PM2,5

Pył PM2,5 to aerozole atmosferyczne, których średnica nie przekracza 2,5 μm. Tego rodzaju pył zawieszony jest uznawany za najgroźniejszy dla zdrowia człowieka. Według WHO długotrwałe narażenie na działanie pyłu zawieszonego PM2,5 skutkuje

skróceniem średniej długości życia. Nawet krótkotrwała ekspozycja na wysokie stężenia pyłu PM2,5 powoduje wzrost liczby zgonów z powodu chorób układu oddechowego i krążenia, oraz wzrost ryzyka nagłych przypadków wymagających hospitalizacji. Wszystkie powikłania spowodowane występowaniem tego pyłu następują na skutek bardzo małych wymiarów pyłu, który z łatwością może przedostać się bezpośrednio do ludzkiego krwiobiegu. Polskie normy nie określają maksymalnych stężeń dla pyłów PM2,5. Natomiast wartości stężeń pyłu zawieszonego PM2,5 zalecane przez WHO wyglądają następująco:

- stężenia 24-godzinne: 25 $\mu g/m^3$,
- stężenia średnioroczne: 10 $\mu g/m^3$.

Warto dodać również, że w przypadku Polski według WHO normy te są zwiększone do 25 $\mu g/m^3$, a od 2020 do 20 $\mu g/m^3$.

Zarejestrowane dane można scharakteryzować następująco:

- liczba dni pomiarowych: 366,
- wartość średnia: 44,48,
- odchylenie standardowe: 23,97,
- wartość minimalna: 1,16,
- 25%: 23,71,
- 50%: 45,62,
- 75%: 61,83,
- wartość maksymalna: 114,37.

Zarejestrowane dane pokazują, że mieszkańcy z okolicy stacji pomiarowej mają dość poważny problem z ilością szkodliwego pyłu w powietrzu atmosferycznym. Wartość średnioroczna w tym obszarze została przekroczona niemalże czterokrotnie. Ciekawą zależnością jest również występowanie najwyższych wartości w okolicy 150 dnia roku, czyli w maju, co jest w sprzeczności z pierwotnym założeniem dotyczącym największych ilości szkodliwych pyłów w okresie zimowym.

3.3. NO_2

Kolejną szczególnie groźną dla zdrowia człowieka substancją jest dwutlenek azotu (NO_2). To właśnie za jego sprawą smog przybiera charakterystyczną postać brunatnego zabarwienia. W miastach o dużym natężeniu ruchu samochodowego, gaz ten powoduje powstawanie zjawiska zwanego smogiem fotochemicznym, czyli występowania w słoneczne dni brunatnej mgły unoszącej się nad miastem. Dla człowieka NO_2 jest zagrożeniem przede wszystkim ze względu na oddziaływanie na układ oddechowy. Wdychanie powietrza, w którym znajdują się duże zawartości NO_2 może objawiać się powstawaniem ataków duszności, podrażnieniem śluzówek, kłuciem w klatce piersiowej czy spłyceniem oddechu. Tlenki azotu wchodzące w skład smogu powstają zwłaszcza na skutek przedostawania się do atmosfery spalin samochodowych, a także toksyn emitowanych przez zakłady przemysłowe. Dopuszczalne stężenia NO_2 w Polsce to:

- poziom dopuszczalny stężenia średniorocznego: 40 $\mu g/m^3$,
- poziom dopuszczalny stężenia średniego 1-godzinnego: 200 $\mu g/m^3$ (przekroczenie tego poziomu jest dozwolone 18 razy w ciągu roku),
- poziom alarmowy stężenia średniego 1-godzinnego: 400 $\mu g/m^3$.

Zarejestrowane dane można scharakteryzować następująco:

- liczba dni pomiarowych: 366,

- wartość średnia: 24,12,
- odchylenie standardowe: 10,41,
- wartość minimalna: 4,39,
- 25%: 15,94,
- 50%: 22,72,
- 75%: 30,81,
- wartość maksymalna: 59,42.

W przypadku NO_2 można również wskazać sezonowość oraz trendy wzrostowe, jednakże nie są one aż tak uwydatnione w porównaniu z pozostałymi substancjami. Wartość średnioroczna mieści się w wymaganym dopuszczalnym zakresie.

3.4. NO_x

Tlenki azotu to jedne z najbardziej niebezpiecznych składników smogu. Ich toksyczność jest wielokrotnie większa w porównaniu do tlenku węgla czy dwutlenku siarki. Jako NO_x można określić mieszaninę tlenku azotu (NO) i dwutlenek azotu (NO_2). Są to nieorganiczne gazy utworzone poprzez połączenie tlenu z azotem z powietrza. NO jest wytwarzany w znacznie większych ilościach niż NO_2, ale utlenia się do NO_2 w atmosferze. Szkodliwy wpływ NO_x jest taki sam jak NO_2, który jest jego głównym składnikiem. Tlenki azotu pojawiają się wszędzie tam gdzie NO_2, czyli przede wszystkim w transporcie oraz procesach przemysłowych i energetycznych. Dotychczas nie ustanowiono dopuszczalnego poziomu NO_x.

Zarejestrowane dane można scharakteryzować następująco:

- liczba dni pomiarowych: 366,
- wartość średnia: 36,71,
- odchylenie standardowe: 26,64,
- wartość minimalna: 5,92,
- 25%: 19,89,
- 50%: 29,43,
- 75%: 43,54,
- wartość maksymalna: 197,25.

Przedstawione czynniki występujące w rejonie stacji meteorologicznej pokazują dość wysokie odczyty NO_x dla niemalże całego roku. W przypadku NO_x można zauważyć sezonowość oraz trendy podobne do PM10. Widoczny jest trend wzrostowy w okresie zimowym, jednakże największe wartości zarejestrowanych wartości są pojedynczymi zdarzeniami, które są wartościami odstającymi względem innych odczytów.

3.5. SO_2

Źródłem powstawania SO_2 są paliwa kopalne, które zawierając śladowe ilości związków siarki i powodują wydzielanie SO_2 w procesie spalania. Większość SO_2 emitowana jest do powietrza atmosferycznego przy produkcji energii elektrycznej oraz są emitowane w niewielkim stopniu z transportu. Ekspozycja na SO_2 może szkodzić zdrowiu, negatywnie oddziałuje na układ oddechowy.

Kwas siarkowy wytwarzany w wyniku atmosferycznej reakcji SO_2 jest głównym składnikiem niebezpiecznych kwaśnych deszczów. Tlenki siarki reagując w powietrzu z innymi związkami powodują powstawanie groźnych drobnych pyłów PM. Normy dotyczące SO_2 obowiązujące w Polsce są następujące:

- średnie stężenie 1-godzinne: 350 µg/m^3 (dopuszczalne przekroczenie 24 razy w roku, poziom alarmowy: 500 µg/m^3),
- średnie stężenie 24-godzinne: 125 µg/m^3 (dopuszczalne przekroczenie max. 3 razy w roku).

Zarejestrowane dane można scharakteryzować następująco:

- liczba dni pomiarowych: 366,
- wartość średnia: 35,96,
- odchylenie standardowe: 19,30,
- wartość minimalna: 10,17,
- 25%: 24,06,
- 50%: 30,99,
- 75%: 41,98,
- wartość maksymalna: 144,87.

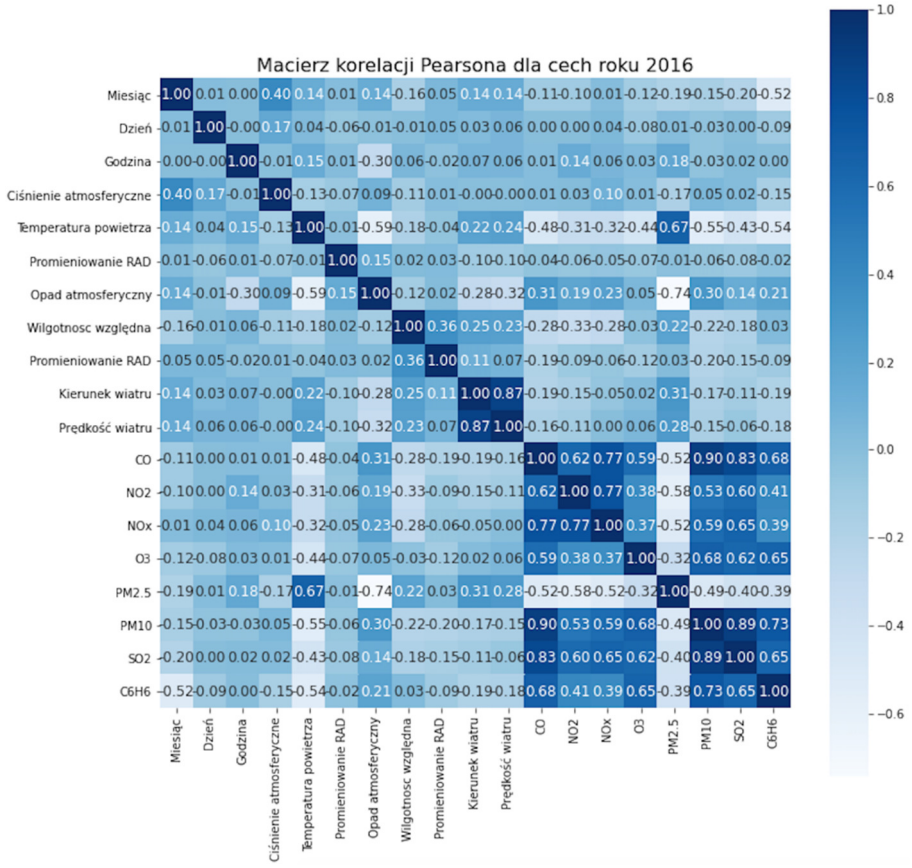
SO_2 jest kolejnym przykładem popierającym tezę dotyczącą występowania wzmożonych ilości niebezpiecznych dla ludzi substancji w okresie zimowym. Dane analogicznie jak w przypadku PM10, NO_2 czy NOx osiągają najwyższe wartości z początkiem i końcem roku. Zmierzone wartości są dalekie od przekroczenia obowiązujących norm.

3.6. Korelacje pomiędzy zarejestrowanymi substancjami

Współczynnik korelacji r Pearsona pozwala na sprawdzenie czy dwa atrybuty są ze sobą powiązane związkiem liniowym. Takie obserwacje dostarczają wielu informacji o zbiorze danych, ponieważ pozwalają w prosty sposób zrozumieć jak zachowuje się dana zmienna względem innej rozpatrywanej zmiennej. Współczynnik korelacji waha się w zakresie od -1 do +1. Gdzie -1 określa silną korelację ujemną, +1 silną korelację dodatnią, natomiast 0 oznacza brak korelacji pomiędzy dwoma atrybutami. Ważnym elementem analizy korelacji jest to, że nie bada ona związku przyczynowo-skutkowego dwóch zmiennych, a jedynie ich wzajemne współwystępowanie. Korelacja pozwala na wybranie atrybutów do budowy modeli przewidujących nieznane wartości jednych wielkości na podstawie znanych wartości innych. Analiza korelacji będąc jedną z klasycznych metod wyboru cech znalazła również zastosowanie w przypadku modeli uczenia głębokiego. Modele uczenia głębokiego będąc podzbiorem uczenia maszynowego mogą działać lepiej w przypadku mniejszej liczby atrybutów, na których następować będzie proces uczenia. W przypadku dużej liczby danych, mniej skomplikowany model ma również zaletę w postaci mniejszej ilości wymaganej pamięci, a rezultatem są szybsze obliczenia.

Na rysunku 1 pokazano mapę korelacji Pearsona dla cech roku 2016.

Analizując przedstawioną macierz korelacji można zauważyć, że w przypadku dużej części substancji zachodzi pozytywny związek między zmiennymi. Na przykład w przypadku przewidywania wartości PM10 atrybutami, które mogą okazać się pomocne są m.in. O_3, SO_2 oraz C_6H_6. Natomiast w przypadku przewidywania wartości NO_x można posłużyć się CO, NO_2 oraz SO_2.



Rysunek 1. Mapa korelacji Pearsona

Po dokonaniu eksploracyjnej analizy danych do prognozowania wybrane zostały tlenki azotu NO_x. Najbardziej istotnym czynnikiem decydującym o tym wyborze były źródła ich pochodzenia, czyli transport oraz przemysł.

3.7. Przygotowanie danych do modeli

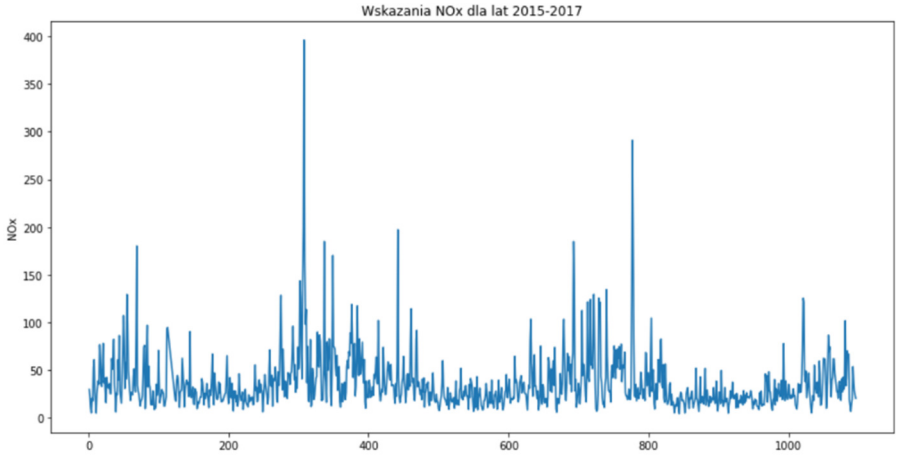
Do zbudowania modeli zostały wybrane lata 2015-2017, jednakże również w tych latach występowała pewna liczba wartości brakujących (tabela 2).

Tabela 2. Liczba brakujących danych w zarejestrowanych pomiarach

Rok	NO_x
2015	473
2016	73
2017	99

Dla pozostałych brakujących wartości NO_x została przeprowadzona metoda interpolacji liniowej wypełniająca brakujące w odczytach dla poszczególnych godzin wartości pośrednie.

Po wczytaniu danych i wybraniu z nich wartości dotyczących NO_x, dokonano uśrednienia wartości z 24 godzin i w ten sposób otrzymano 1096 wartości odczytów (rysunek 2), co jest dość małą liczbą danych jak na wymagania sieci neuronowych. Mimo wszystko pozwoliło to na przetestowanie różnych modeli, które w przyszłości mogą zostać uaktualnione o nowe dane.



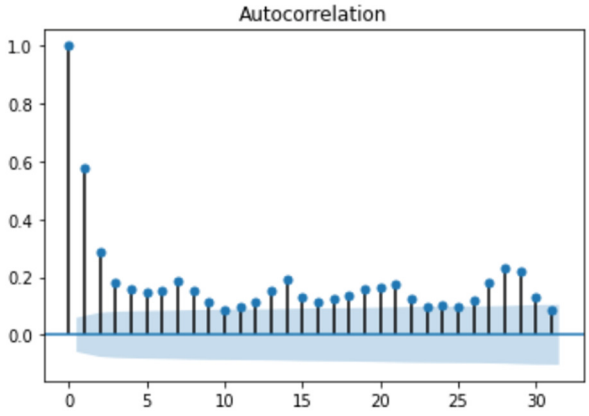
Rysunek 2. Wskazania NO_x dla lat 2015-2017

Zarejestrowane dane można scharakteryzować następująco:

- liczba dni pomiarowych: 1096,
- wartość średnia: 36,40,
- odchylenie standardowe: 29,57,
- wartość minimalna: 4,44,
- 25%: 19,11,
- 50%: 27,92,
- 75%: 43,01,
- wartość maksymalna: 395,98.

Jak można wywnioskować z przedstawionego wykresu, dla trzech ostatnich lat powtarza się schemat sezonowości zanieczyszczeń. W każdym kolejnym roku widać dominacje wysokich wartości odczytów NO_x w okresie zimowym, które następnie spadają do minimalnych wartości w okresie letnim.

Autokorelacja będąc narzędziem matematycznym stosowanym przy przetwarzaniu sygnałów do analizowania serii wartości pozwala na uzyskanie informacji, w jakim stopniu dany wyraz szeregu zależy od wyrazów poprzednich w szeregu czasowym. Autokorelacja przypisuje danemu argumentu k wartość współczynnika korelacji Pearsona pomiędzy szeregiem czasowym, a tym samym szeregiem cofniętym o k jednostek czasu. Dokonując autokorelacji na danych dotyczących NO_x uzyskano następujący wykres przedstawiony na rysunku 3.



Rysunek 3. Wykres autokorelacji dla NO_x

Z wykresu można wywnioskować, że najlepszym do prognozowania będzie kolejny nadchodzący dzień, ponieważ osiąga najwyższą wartość korelacji.

4. Zadanie prognozy dla dnia następnego

Na podstawie wykresu autokorelacji rozsądnym punktem rozpoczęcia prognoz jest próba prognozowania wartości NO_x dla dnia następnego. W dalszej części artykułu przedstawiono wyniki badań przy wykorzystaniu różnych metod prognozowania, rozpoczynając od prostych metod statystycznych, a kończąc na bardziej zaawansowanych metodach korzystających z sieci neuronowych.

Przed przystąpieniem do prognozowania należy podzielić dane na zbiór treningowy oraz zbiór testowy. Pierwszy zbiór jest wykorzystywany do trenowania modelu, natomiast drugi zbiór jest wykorzystywany do dokonywania predykcji poprzez model, a następnie otrzymane wartości są porównywane z wartościami oczekiwanymi. Celem podziału jest oszacowanie wydajności modelu uczenia maszynowego na nowych danych, czyli takich, które nie zostały wcześniej użyte do trenowania modelu. W przypadku dostępnych danych, rozsądnym podziałem jest wykorzystanie dwóch lat jako zbioru treningowego, natomiast pozostały rok został użyty jako zbiór testowy celem sprawdzenia wydajności modelu. Za pomocą takiego podziału uzyskano 731 dni, na których będą trenowane modele, oraz 365 dni, które posłużyły do oceny wydajności utworzonego modelu.

Wszystkie modele zostały ocenione dla danych dotyczących całego roku zbioru testowego za pomocą następujących metryk:

- średni błąd bezwzględny MAE (z ang. Mean Absolute Error),
- pierwiastek błędu średniokwadratowego RMSE (z ang. Root Mean Squared Error).

Średni błąd bezwzględny mierzy średnią wielkość błędów w zbiorze przewidywań, nie biorąc pod uwagę ich kierunku. Jest to średnia z badanej próby bezwzględnych różnic między prognozą a rzeczywistą wartością, gdzie wszystkie indywidualne różnice mają jednakową wagę. Wyraża się go za pomocą następującego wzoru:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y_i}\right| \quad (1)$$

Pierwiastek błędu średniokwadratowego jest wartością oczekiwaną kwadratu „błędu", czyli różnicy między wartością rzeczywistą, a wartością prognozy. Wyraża się go za pomocą następującego wzoru:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (2)$$

gdzie dla obu metryk:
- y_i – wartość rzeczywista,
- $\hat{y_i}$ – wartość prognozowana.

4.1. Prognozowanie bazowe – naiwna metoda prognozowania

Ważnym elementem przed opracowaniem zaawansowanych modeli prognostycznych jest przetestowanie najprostszych metod. Proste metody są to takie, co do których nie oczekuje się zbyt wiele odnośnie rozpatrywanego problemu, ale są szybkie do wdrożenia. Wyniki uzyskane z takich metod mogą zostać wykorzystane jako punkt odniesienia przy ocenie wydajności bardziej zaawansowanych modeli. Jeżeli rozpatrywany model jest w stanie osiągnąć lepsze wyniki niż te uzyskane w przypadku prostej metody to oznacza, że jest wydajny. Jedną z takich metod jest naiwna metoda prognozowania. Metoda ta polega na wykorzystaniu poprzedniej obserwacji bezpośrednio jako prognozy. W przypadku zastosowania tej metody nie ma potrzeby trenowania modelu. W tym przypadku wystarczy, że dla zbioru danych testowych przesunięte zostaną wartości rzeczywiste o jeden dzień do przodu. W ten sposób otrzymano pierwszą najprostszą prognozę.
Dzięki zastosowaniu tej prostej metody można uzyskać pierwsze wyniki w ramach wybranych metryk. Porównując wartości rzeczywiste oraz prognozy dla całego roku otrzymano:
- roczny RMSE: 23,34,
- roczny MAE: 14,21.

Uzyskane wyniki prognozy są wynikami bazowymi, na których podstawie były oceniane kolejne modele.

4.2. Metoda średniej ruchomej

Kolejną rozpatrywaną metodą była metoda średniej ruchomej. W przypadku danych dotyczących NO_x została zastosowana prosta średnia ruchoma SMA (z ang. Simple Moving Average), która jest zwykłą średnią arytmetyczną z wartości ostatnich n okresów. Jeżeli przez p_0 oznaczy się ostatnią wartość to:

$$SMA = \frac{p_0 + p_1 + \cdots + p_{n-1}}{n} \quad (3)$$

Prognozując wskazania NO_x na następny dzień skorzystano z siedmiu poprzednich wartości, co pozwoliło na uzyskanie następujących wyników:
- roczny RMSE: 27,08,
- roczny MAE: 15,56.

W przypadku średniej ruchomej otrzymane wyniki są gorsze niż wyniki uzyskane dla prognozy bazowej. Taka informacja pozwalała na kontynuowanie dalszych poszukiwań w kierunku lepszych wyników. Wydajniejszych metod można szukać między innymi w modelach zbudowanych ze sztucznych sieci neuronowych.

4.3. Perceptron

Dysponując wartościami siedmiu poprzednich dni, kolejną prostą metodą jaką można użyć do prognozy następnego dnia jest perceptron. Perceptron to najprostsza sieć neuronowa składająca się z jednego lub wielu niezależnych neuronów McCullocha-Pittsa.
W przypadku zastosowania perceptronu do danych treningowych otrzymano zwykły model regresyjny. Model dostrajając poszczególne wagi dla poprzednich siedmiu dni stara się estymować następny dzień minimalizując przy tym funkcje straty jaką jest błąd średniokwadratowy. Po przetrenowaniu perceptronu na dostępnych danych uzyskano poszczególne wagi odpowiadające poprzednim dniom. Rezultatem prognozowania następnych dni za pomocą danych testowych są następujące wyniki:

- roczny RMSE: 21,49,
- roczny MAE: 12,81.

Jak widać model prostego perceptronu okazał się dużo lepszy niż model średniej ruchomej, oraz udało się osiągnąć lepszy wynik niż prognoza bazowa, co pozwala określić model jako wydajny. W przypadku prostego perceptronu można również zastosować inną liczbę dni poprzednich celem dostrojenia wag, jednak lepszym sposobem kontynuacji prognoz z użyciem perceptronów było zastosowanie perceptronu wielowarstwowego.

4.4. Perceptron wielowarstwowy

Perceptron wielowarstwowy MLP (ang. Multi Layer Perceptron) jest najpopularniejszym typem sztucznej sieci neuronowej. Sieć tego typu składa się zwykle z jednej warstwy wejściowej, kilku warstw ukrytych oraz jednej warstwy wyjściowej. Warstwy ukryte składają się najczęściej z neuronów McCullocha-Pitsa. Ustalenie właściwej liczby warstw ukrytych oraz liczby neuronów znajdujących się w poszczególnych warstwach jest trudnym zadaniem, dla którego najlepszym rozwiązaniem jest przetestowanie różnych architektur sieci. Trenowanie sieci typu MLP jest możliwe dzięki zastosowaniu metody wstecznej propagacji błędów.
Stosując sieć typu MLP do rozpatrywanego problemu można przyjąć, że warstwa wejściowa do sieci miała otrzymywać tak samo jak w przypadku pojedynczego perceptronu siedem ostatnich wartości pomiarów, natomiast jej zadaniem było tak dostroić wagi, aby otrzymać prognozę na nadchodzący dzień. Proponowana sieć składała się z trzech warstw. Sieć była zbudowana na podstawie modelu sekwencyjnego, który pozwala na tworzenie modelu warstwa po warstwie. Takie rozwiązanie nie wymagało jawnej deklaracji warstwy wejściowej, ponieważ taką rolę pełni zawsze pierwsza warstwa dodana do modelu. Pierwsza warstwa przyjmująca siedem poprzednich wartości NO_x składała się z 64 neuronów oraz funkcji aktywacji typu ReLU (z ang. Rectified Linear Unit). Zadaniem funkcji aktywacji była decyzja, co zrobić z wynikiem wychodzącym z bloku sumacyjnego pojedynczego perceptronu. Nie sposób wymienić tutaj wszystkich stosowanych funkcji aktywacji, jednak są one nieodłącznym punktem uczenia głębokiego. Druga warstwa składała się natomiast

z 32 neuronów oraz kolejnej funkcji aktywacji typu ReLU. Ostatnią warstwą wyjściową był pojedynczy neuron, który był odpowiedzialny za prognozę na następny dzień.
Uzyskano następujące wyniki:
- roczny RMSE: 28,63,
- roczny MAE: 15,45.

Jak widać testowany model spisał się gorzej od prostego modelu perceptronu. Szczególnie złą prognozę można zauważyć w przypadku najwyższej wartości pomiarowej dla roku testowego. Model zamiast prognozować najwyższą wartość, złapał błąd, przez co prognozą na ten dzień jest wynik ujemny. Ponadto można zauważyć, że w przypadku dużej części dni prognozy są wyższe od wartości rzeczywistych, co może świadczyć o nadmiernym dopasowaniu modelu do danych treningowych.

4.4. LSTM

Problemy prognozy szeregów czasowych są trudnym rodzajem prognoz do zamodelowania. W przeciwieństwie do modelowania regresyjnego, szeregi czasowe zawierają złożoną zależność sekwencyjną wśród zmiennych wejściowych. Wydajnym rodzajem sieci neuronowych radzących sobie z tymi zależnościami są rekurencyjne sieci neuronowe RNN. Szczególnym typem sieci rekurencyjnych, który znalazł zastosowanie w przypadku prognozowania szeregów czasowych są sieci LSTM (z ang. Long Short-Term Memory). Najważniejszą cechą sieci LSTM w odróżnieniu od perceptronu wielowarstwowego jest możliwość przechowywania informacji przez pewien okres czasu. Cecha ta jest niezwykle przydatna w przypadku danych w postaci szeregów bądź sekwencji. Modele LSTM mając pewną pojemność pamięci, mają swobodę decydowania o tym, jakie informację będą przechowywane a jakie odrzucane. Czynności te są zrealizowane za pomocą tzw. bramek.
Stosując do uzyskanych prognoz wybrane metryki otrzymano:
- roczny RMSE: 26,99,
- roczny MAE: 17,29.

Otrzymany model jest z pewnością lepszy od perceptronu wielowarstwowego, jednakże model ten można nazwać modelem testowym do bardziej złożonych modeli. Wynika to z fakty, że nie wykorzystuje on w pełni potencjału komórek LSTM. Rozważany problem był zbyt prosty dla zbudowanej sieci, co spowodowało wyniki gorsze niż w przypadku prostszych modeli. Żeby w pełni wykorzystać potencjał komórek LSTM należy przetestować je w bardziej zaawansowanych problemach.

5. Podsumowanie

Wyniki przeprowadzonych badań jednoznacznie wykazały potencjał leżący w zastosowaniu sztucznych sieci neuronowych do prognozowania stężenia zanieczyszczenia powietrza atmosferycznego. W trakcie eksperymentów zostały przetestowane różne typy oraz architektury sieci neuronowych dostosowane do problemu prognozowania szeregu czasowego. Wnioski dotyczące uzyskanych wyników są następujące:

- Najlepszym modelem okazał się pojedynczy perceptron McCullocha-Pittsa. Powodem tak dobrego wyniku przy użyciu stosunkowo prostego modelu była prostota rozpatrywanego problemu. Problem ten nie był w gruncie rzeczy typowym problemem prognozy, a bardziej problemem regresji polegającym na dostosowaniu 7 wag poprzednich dni celem uzyskania prognozy dnia następnego.
- Bardziej zaawansowane modele takie jak perceptron wielowarstwowy oraz LSTM były zbyt rozbudowane w stosunku do rozpatrywanego problemu, co przełożyło się na osiągnięcie przez nie niezadowalających wyników.
- W przypadku tak prostego problemu prognostycznego, równie zadowalające wyniki mogłyby zostać uzyskane nawet w przypadku klasycznych metod wykorzystywanych do prognoz.

LITERATURA

1. BROWNLEE J.: Deep Learning for Time Series Forecasting. Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery 2018.
2. CHOLLET F.: Deep Learning. Praca z językiem Python i biblioteką Keras. Wydawnictwo Helion, Gliwice 2019.
3. DUCH W., KORBICZ J., RUTKOWSKI L., TADEUSIEWICZ R. (red.): Sieci neuronowe. W: NAŁĘCZ M. (red.), Biocybernetyka i inżynieria biomedyczna 2000, tom VI, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2000.
4. GERON A.: Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow. Wydawnictwo Helion, Gliwice 2018.
5. GOODFELLOW I., BENGIO Y., COURVILLE A.: Deep Learning. MIT Press, 2016.
6. KORBICZ J., OBUCHOWICZ A., UCIŃSKI D.: Sztuczne sieci neuronowe - podstawy i zastosowania. Akademicka Oficyna Wydawnicza PLJ, Warszawa 1994.
7. OSOWSKI S.: Sieci neuronowe do przetwarzania informacji. Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2020.
8. RASCHKA S.: Python machine learning. Packt Publishing Ltd. 2015.
9. SIEDLECKI J. (red.): Wykorzystanie sztucznych sieci neuronowych w modelowaniu ekonomicznym. Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2001.
10. SRA S., NOWOZIN S., WRIGHT S.J. (eds.).: Optimization for Machine Learning. MIT Press, Cambridge 2012.
11. TADEUSIEWICZ R.: Sieci neuronowe. Akademicka Oficyna Wydawnicza RM, Warszawa 1993.

Badania zrealizowano w ramach Project Base Learning sfinansowanego z Funduszu Europejskiego w ramach projektu Politechnika Śląska jako Centrum Nowoczesnego Kształcenia opartego o badania i innowacje POWR-03.05.00-00-Z098/17-00.