APROKSYMACJA POZIOMU IMISJI NA STACJACH MONITORINGU POWIETRZA ZA POMOCĄ AUTONOMICZNYCH MODELI NEURONOWYCH

Szymon HOFFMAN
Katedra Chemii, Technologii Wody i Ścieków, Politechnika Częstochowska
ul. Dąbrowskiego 69, 42-200 Częstochowa, szymon@is.pcz.czest.pl

STRESZCZENIE

W analizie wykorzystano dane zarejestrowane w okresie wieloletnim na stacjach monitoringu powietrza, działających w tym samym regionie. Badano dokładność aproksymacji stężeń chwilowych różnych zanieczyszczeń powietrza. Do predykcji stężeń użyto autonomicznych modeli neuronowych, dla których jedynym źródłem wiedzy o aproksymowanych wartościach są zależności występujące w historycznych danych rejestrowanych w sieciach monitoringu powietrza. Błąd modelowania, obliczony na podstawie porównania danych rzeczywistych i danych uzyskanych w wyniku aproksymacji, przyjęto w porównaniach jako miarę dokładności różnych metod modelowania. Wykonana analiza pozwoliła wybrać najdokładniejsze techniki modelowania, które można rekomendować do aproksymacji stężeń poszczególnych zanieczyszczeń powietrza.

1. Wprowadzenie

Dane zbierane w sposób ciągły na stacjach monitoringu powietrza nigdy nie są kompletne. Gdy kompletność wyników dla danej serii pomiarowej jest za mała, wtedy pojawia się potrzeba wprowadzenia wartości modelowanych do luk pomiarowych [1-2]. W przeciwnym wypadku seria pomiarowa zostaje uznana za niekompletną i nie powinna być wykorzystywana do oceny jakości powietrza [3]. Obowiązujące akty prawne nie rekomendują określonych metod modelowania. Najprostszą metodą uzupełniania serii pomiarowej jest interpolacja liniowa. Jednak dla dłuższych luk pomiarowych ta technika aproksymacji jest mało dokładna. Dokładniejsze mogą być metody eksplorujące wiedzę zgromadzoną w danych historycznych, zarejestrowanych na wybranej stacji monitoringu i na stacjach sąsiednich. Modele umożliwiające predykcję stężeń bez konieczności wykorzystywania danych zewnętrznych, pochodzących spoza systemu monitoringu, nazwano modelami autonomicznymi [4]. Wśród modeli autonomicznych można wyróżnić dwa podstawowe typy – modele regresyjne i modele szeregów czasowych. Modele regresyjne eksplorują zależności obserwowane między różnymi seriami pomiarowymi, natomiast modele szeregów czasowych bazują na autoregresji charakteryzującej poszczególne serie czasowe. Wybór metody modelowania powinien zapewniać możliwie najwyższą dokładność predykcji.

Zasadniczym celem przeprowadzonych badań było porównanie dokładności różnych typów modeli autonomicznych i znalezienie wśród nich optymalnych metod modelowania dla poszczególnych zanieczyszczeń powietrza. Przyjęto, że istnieje potrzeba modelowania luki pomiarowej, obejmującej pewien fragment serii czasowej wybranego zanieczyszczenia powietrza, przy założeniu, że są dostępne kompletne dane dotyczące innych zanieczyszczeń i parametrów meteorologicznych z okresu obejmującego lukę oraz dane z modelowanej serii czasowej poprzedzające taką lukę pomiarową. Dla kolejnych przypadków w tak zdefiniowanej luce pomiarowej porównywano błędy predykcji różnych metod modelowania.

W przypadku modeli regresyjnych wykorzystywano dostępne dane dla rozpatrywanego przypadku, tzn. dane zarejestrowane w tym samym dniu i o tej samej godzinie. W przypadku modeli szeregów czasowych kolejne przypadki w luce pomiarowej traktowano jako następujące po sobie kroki prognozy i przypisywano im wartości błędów otrzymane dla równoważnych horyzontów prognozy. Do budowy modeli predykcyjnych wykorzystano sieci neuronowe, ponieważ umożliwiają one optymalną aproksymację modelowanych zmiennych.

W prezentowanej pracy analizie poddano wieloletni zbiór danych zarejestrowanych na kilku stacjach monitoringu powietrza usytuowanych w centralnej Polsce. Aproksymację stężeń przeprowadzono dla stężeń zarejestrowanych na stacji monitoringu w Radomiu.

2. Metodyka badań

W analizie wykorzystano dane zarejestrowane w latach 2004-2008 na ośmiu stacjach monitoringu powietrza działających w różnych miejscowościach województw łódzkiego i mazowieckiego. Lokalizację stacji ilustruje mapa przedstawiona na rys. 1. Odpowiednim stacjom przypisano nazwy związane z ich usytuowaniem: Widzew, Gajew, Granica, Piotrków Trybunalski, Legionowo, Radom, Tłuszcz, Ursynów.

Rys. 1. Lokalizacja rozpatrywanych stacji monitoringu powietrza

Analizowano serie czasowe stężeń chwilowych podstawowych zanieczyszczeń powietrza (O_3, NO, NO_2, PM_{10}, SO_2, CO). Wykorzystano także rejestrowane na stacjach monitoringu dane meteorologiczne, w tym kierunek i prędkość wiatru, temperaturę, natężenie promieniowania słonecznego, wilgotność względną. W pracy przeprowadzono modelowanie stężeń zanieczyszczeń powietrza pochodzących ze stacji monitoringu Radom, a następnie dokonano oceny jakości modelowania poprzez porównanie modelowanych stężeń ze stężeniami rzeczywistymi.

Predykcję stężeń każdego z zanieczyszczeń powietrza wykonano za pomocą różnych metod modelowania. We wszystkich metodach wielkością modelowaną (wyjściem modelu) było stężenie wybranego zanieczyszczenia w określonym czasie. Modele różniły się liczbą i charakterem zmiennych objaśniających oraz techniką modelowania. Wygenerowano trzy podstawowe grupy modeli:

1. TS-L – Modele szeregów czasowych (Time-Series – Linear). Wejściami modeli TS-L były stężenia wybranego do modelowania zanieczyszczenia zarejestrowane w godzinach wcześniejszych. Wszystkie modele szeregów czasowych miały stałą liczbę

(24) wartości opóźnionych, stanowiących wejścia do modelu. W obrębie grupy modeli TS-L wygenerowano modele różniące się horyzontem prognozy, czyli liczbą kroków od ostatniej z wartości opóźnionych do wartości prognozowanej. Przyjęto następujące horyzonty prognozy: 1, 2, 3, 4, 5, 6, 8, 12, 24. Taki zakres horyzontów prognozy pozwolił ocenić jakość modelowania predykcyjnego w 24-godzinnym okresie. Do analizy szeregów czasowych wykorzystano liniowe sieci neuronowe. Wyjątkowo w przypadku aproksymacji za pomocą analizy szeregów czasowych ograniczono się do modeli liniowych. Dotychczasowe doświadczenia wykazują, że modele nieliniowe nie poprawiają w sposób istotny jakości modelowania [5].

2. MR-P – Modele regresji wielowymiarowej (Multiple Regression – Perceptron). W modelach tych predyktorami stężenia wybranego zanieczyszczenia były mierzone na tej samej stacji monitoringu stężenia innych zanieczyszczeń, a także dzień i godzina pomiaru oraz dane meteorologiczne. Do analizy regresji wykorzystano nieliniowe sieci neuronowe o strukturze tzw. perceptronu. W każdym modelu perceptronowym przyjęto architekturę sieci z pięcioma neuronami umieszczonymi w pojedynczej warstwie ukrytej. Taka stosunkowo prosta budowa sieci neuronowej pozwala na efektywną eksplorację wiedzy ukrytej w danych [6].
3. EMR-P – Modele regresji wielowymiarowej eksplorujące dane pochodzące z sąsiednich stacji monitoringu (Eexternal Multiple Regression - Perceptron). W modelach tych predyktorami stężenia wybranego zanieczyszczenia były stężenia tego samego zanieczyszczenia zarejestrowane w tym samym czasie na innych sąsiednich stacjach monitoringu powietrza. Do modelowania stężeń zanieczyszczeń na stacji monitoringu Widzew-Łódź wykorzystano dane pochodzące z siedmiu innych stacji usytuowanych w województwach łódzkim i mazowieckim (rys. 1). Zastosowana sieć miała architekturę perceptronu, analogiczną do sieci typu MR-P.

Obliczenia przeprowadzono korzystając z programu STATISTICA Data Mining. W przypadku każdej sieci neuronowej zbiór wszystkich przypadków został losowo podzielony na trzy podzbiory: zbiór uczący (50% przypadków), zbiór weryfikujący (25% przypadków), zbiór testujący (25% przypadków).

W porównaniach modeli wykorzystano cztery różne kategorie błędu predykcji, wynikające z porównania stężeń rzeczywistych i modelowanych:

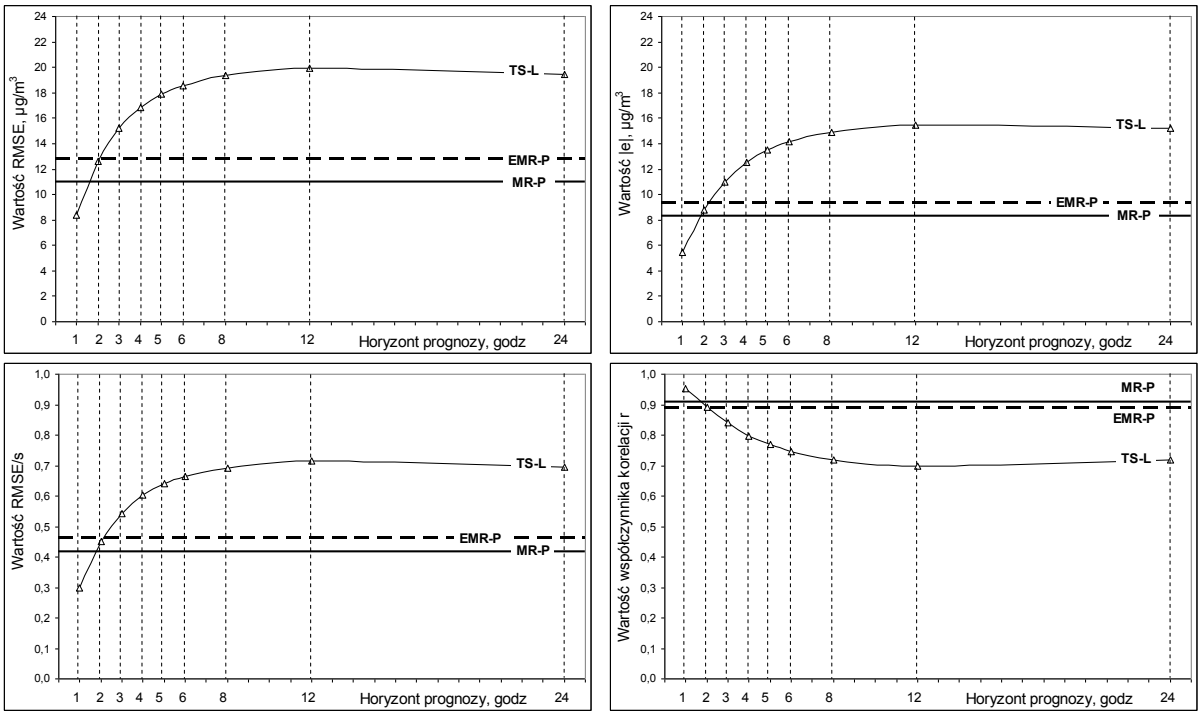
- wartość współczynnika korelacji Pearsona (r),
- wartość pierwiastka z błędu średniokwadratowego (RMSE),
- wartość średniego błędu bezwzględnego (|e|),
- stosunek RMSE do odchylenia standardowego (RMSE/s).

Wartości RMSE i |e| są użytecznym kryterium przy porównywaniu sieci modelujących stężenie wybranego zanieczyszczenia na konkretnej stacji monitoringu. Kryteriami bardziej uniwersalnymi są współczynnik korelacji i stosunek RMSE/s. Te dwie miary błędu umożliwiają porównanie dokładności predykcji stężeń różnych zanieczyszczeń powietrza.

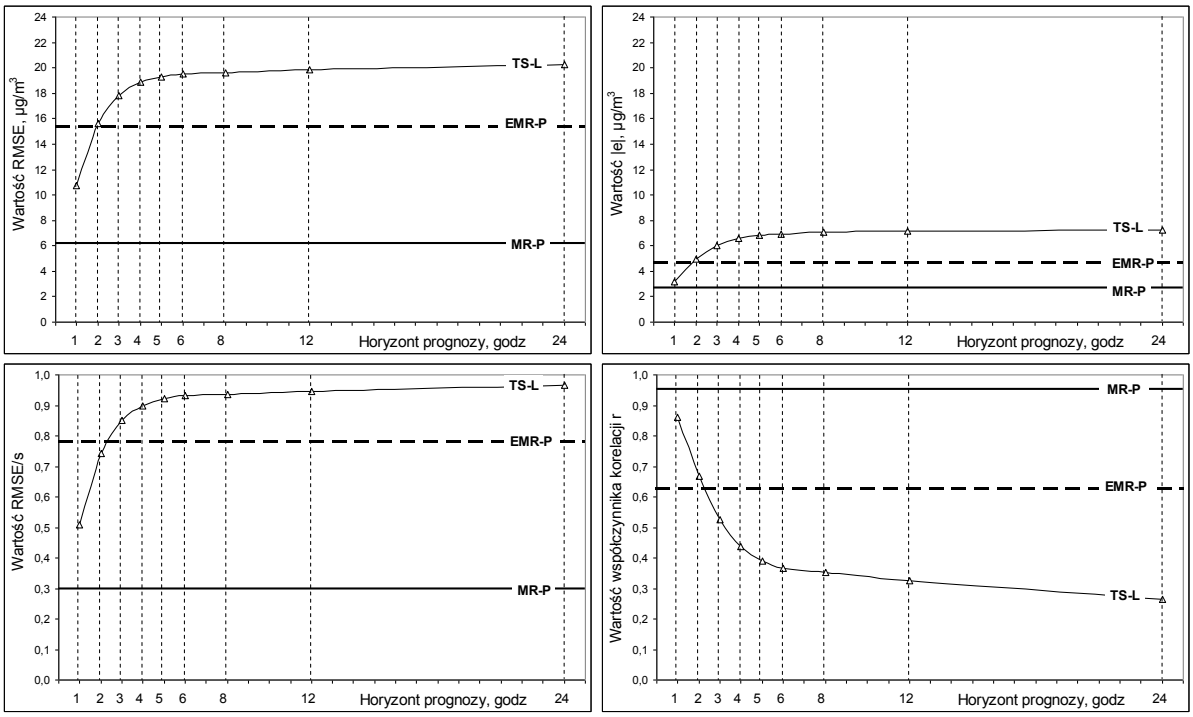
3. Wyniki

Na wykresach przedstawiono zmiany wartości błędów modelowania w zależności od horyzontu prognozy (rys. 2-7). Poszczególne rysunki ilustrują wyniki otrzymane kolejno dla O_3, NO, NO_2, CO, SO_2, PM_{10}. Dla każdego z wymienionych zanieczyszczeń zaprezentowano wykresy zmian wartości czterech różnych kategorii błędu: RMSE, |e|, RMSE/s i r, dla horyzontów prognozy w zakresie od 1 do 24 godzin. Na poszczególnych wykresach porównano wyniki dla modeli wygenerowanych 3 różnymi technikami predykcji, w tym dla modeli szeregów czasowych (TS-L), dla nieliniowych model regresji wielowymiarowej (MR-

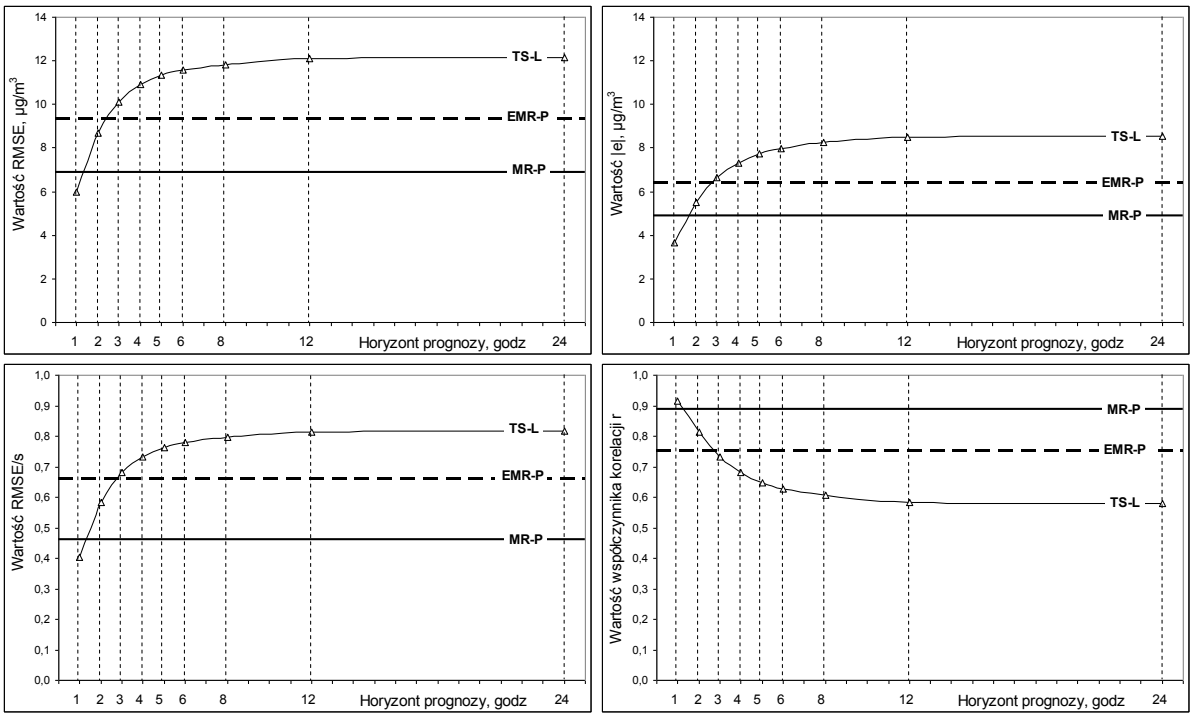
P) i dla nieliniowych modeli regresji wielowymiarowej eksplorujących dane pochodzące z sąsiednich stacji monitoringu (EMR-P). Do przygotowania wykresów wykorzystano błędy modelowania, których dokładne wartości zostały podane we wcześniejszej publikacji [7].



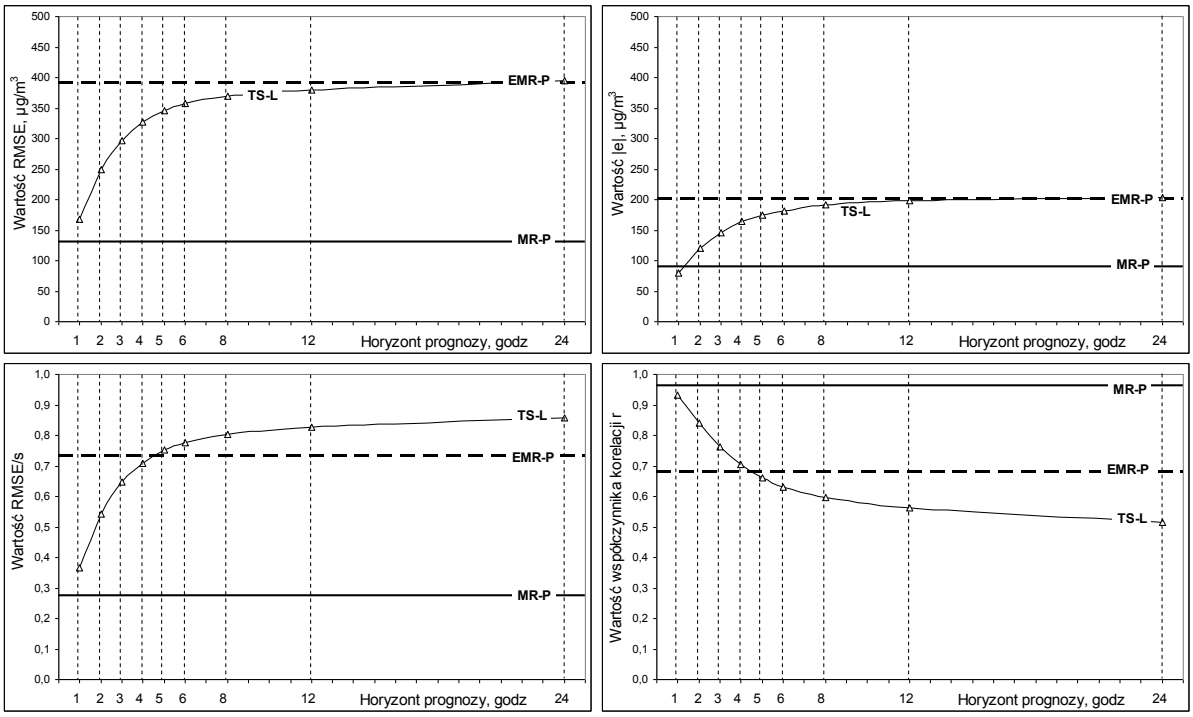
Rys. 2. Błąd modelowania stężeń chwilowych O_3 w zależności od horyzontu prognozy (Modele TS-L, MR-P, EMR-P, Radom 2004-2008)



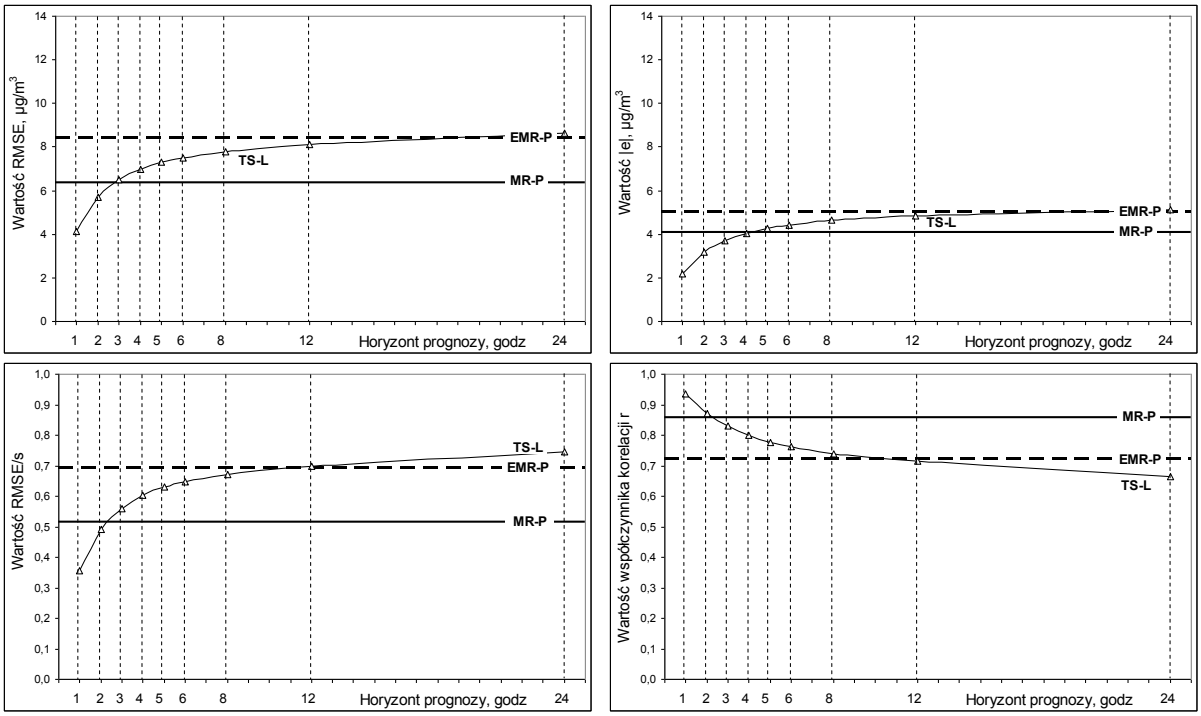
Rys. 3. Błąd modelowania stężeń chwilowych NO w zależności od horyzontu prognozy (Modele TS-L, MR-P, EMR-P, Radom 2004-2008)



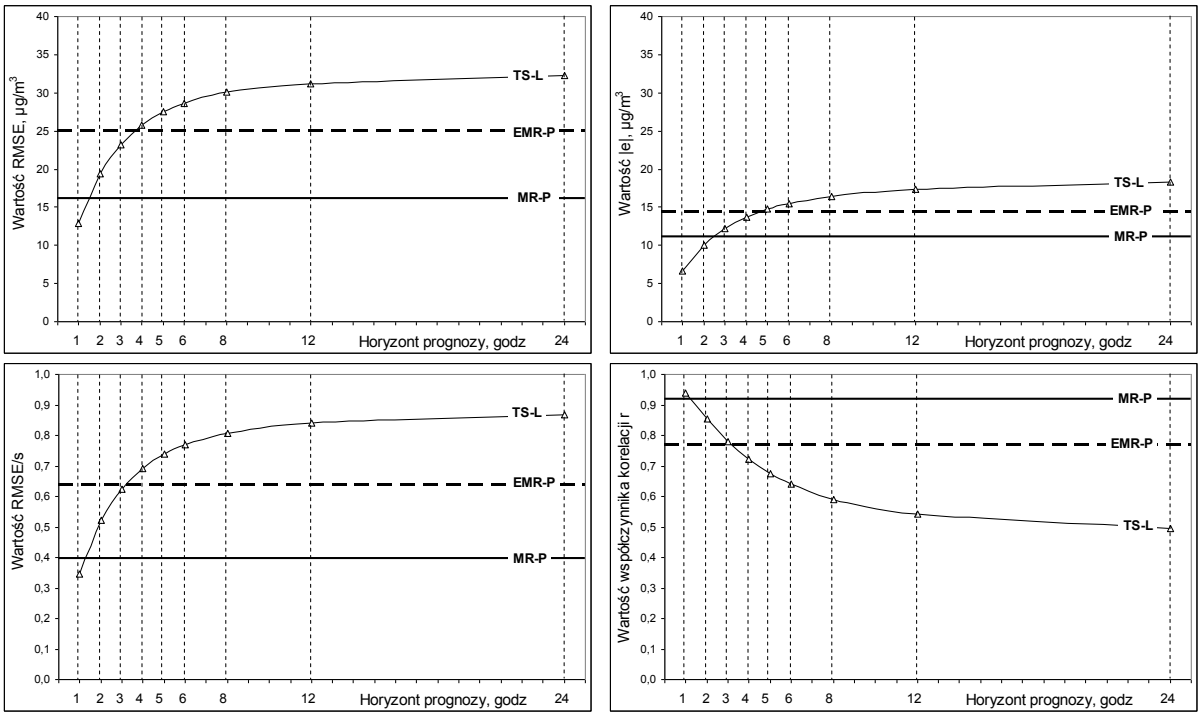
Rys. 4. Błąd modelowania stężeń chwilowych NO_2 w zależności od horyzontu prognozy (Modele TS-L, MR-P, EMR-P, Radom 2004-2008)



Rys. 5. Błąd modelowania stężeń chwilowych CO w zależności od horyzontu prognozy (Modele TS-L, MR-P, EMR-P, Radom 2004-2008)



Rys. 6. Błąd modelowania stężeń chwilowych SO_2 w zależności od horyzontu prognozy (Modele TS-L, MR-P, EMR-P, Radom 2004-2008)



Rys. 7. Błąd modelowania stężeń chwilowych PM_{10} w zależności od horyzontu prognozy (Modele TS-L, MR-P, EMR-P, Radom 2004-2008)

4. Dyskusja wyników i podsumowanie

Modele regresyjne MR-P i EMR-P mają stałą wartość błędu, niezależną od horyzontu prognozy. Tylko modele TS-L charakteryzują się zmienną wartością błędów predykcji w

miarę wydłużania horyzontu prognozy (luki pomiarowej). W przypadku tych modeli wartości błędów RMSE, |e|, RMSE/s stopniowo rosną w miarę wydłużania horyzontu prognozy. Odmiennie zachowują się wartości współczynnika korelacji, które stopniowo maleją, w miarę pogarszania się jakości prognozy. Analiza którejkolwiek z miar błędów prowadzi do wniosku, że jakość modelowania w grupie modeli TS-L silnie zależy od długości luk pomiarowych. Te prawidłowości są obserwowane dla wszystkich zanieczyszczeń powietrza. Tylko dla krótkich luk pomiarowych modele TS-L mogą być dokładniejsze od pozostałych. W tab. 1-6 zestawiono najdokładniejsze metody modelowania stężeń poszczególnych zanieczyszczeń powietrza dla kolejnych horyzontów prognozy.

W przypadku stężeń ozonu modele TS-L zapewniają najwyższą jakość predykcji w lukach pomiarowych o długościach 1 godziny (tab. 1). Rekomendacja optymalnej metody nie zależy od rozpatrywanego rodzaju błędu. Następne przypadki w lukach pomiarowych powinny być modelowane metodą regresji MR-P.

Tabela 1. Najdokładniejsze modele stężeń O_3 dla luk pomiarowych o różnej długości.

Rodzaj błędu	Długość luki pomiarowej, godz.	
	1	>1
RMSE	TS-L	MR-P
|e|	TS-L	MR-P
RMSE/s	TS-L	MR-P
r	TS-L	MR-P

W przypadku stężeń NO modele TS-L są wyjątkowo mało dokładne, nawet dla najkrótszych horyzontów prognozy. Model regresyjny typu MR-P jest natomiast wyjątkowo dokładny, nawet w porównaniu do analogicznych modeli dla innych zanieczyszczeń. Współczynnik korelacji dla tego modelu wynosi 0,954. Rozpatrując wszystkie miary błędu model MR-P należy rekomendować do prognozowania stężeń NO już od pierwszego przypadku w luce (tab. 2).

Tabela 2. Najdokładniejsze modele stężeń NO dla luk pomiarowych o różnej długości.

Rodzaj błędu	Długość luki pomiarowej, godz.	
	1	>1
RMSE	MR-P	MR-P
|e|	MR-P	MR-P
RMSE/s	MR-P	MR-P
r	MR-P	MR-P

W przypadku stężeń NO_2 modele TS-L zapewniają najwyższą jakość predykcji tylko w pierwszym kroku prognozy (tab. 3). Następne przypadki w lukach pomiarowych powinny być modelowane metodą regresji MR-P.

Tabela 3. Najdokładniejsze modele stężeń NO_2 dla luk pomiarowych o różnej długości.

Rodzaj błędu	Długość luki pomiarowej, godz.	
	1	>1
RMSE	TS-L	MR-P
|e|	TS-L	MR-P
RMSE/s	TS-L	MR-P
r	TS-L	MR-P

W przypadku stężeń CO, dla horyzontów prognozy dłuższych niż 1 godz. najdokładniejszym modelem jest model regresyjny typu MR-P (tab. 4). Dla pierwszego kroku prognozy wyższą

dokładność zapewnia model TS-L, ale tylko przy traktowaniu wartości średniego błędu bezwzględnego |e| jako kryterium jakości predykcji.

Tabela 4. Najdokładniejsze modele stężeń CO dla luk pomiarowych o różnej długości.

Rodzaj błędu	Długość luki pomiarowej, godz.	
	1	>1
RMSE	MR-P	MR-P
|e|	TS-L	MR-P
RMSE/s	MR-P	MR-P
r	MR-P	MR-P

Najlepsze rezultaty w modelowaniu stężeń SO_2 dla pierwszych przypadków luk pomiarowych daje metoda TS-L (tab. 5). Rozpatrując kategorię średniego błędu bezwzględnego |e| ta metoda okazuje się najdokładniejsza dla luk pomiarowych nieprzekraczających 4 godzin. Wartości innych błędów pozwalają rekomendować modele TS-L do uzupełniania dwóch pierwszych przypadków w lukach pomiarowych. Dla dłuższych luk pomiarowych najdokładniejszymi modelami stają się modele MR-P.

Tabela 5. Najdokładniejsze modele stężeń SO_2 dla luk pomiarowych o różnej długości.

Rodzaj błędu	Długość luki pomiarowej, godz.				
	1	2	3	4	>4
RMSE	TS-L	TS-L	MR-P	MR-P	MR-P
|e|	TS-L	TS-L	TS-L	TS-L	MR-P
RMSE/s	TS-L	TS-L	MR-P	MR-P	MR-P
r	TS-L	TS-L	MR-P	MR-P	MR-P

Dla stężeń PM_{10} najlepsze wyniki modelowania w pierwszym kroku prognozy daje zawsze metoda TS-L (tab. 6). Dla drugiego kroku prognozy o rekomendacji rozstrzyga kryterium błędu. Analizując wartości średniego błędu bezwzględnego |e| można rekomendować modele szeregów czasowych TS-L. Po porównaniu wartości innych miar błędów, w drugim kroku prognozy dokładniejsze okazują się modele regresyjne typu MR-P. Te modele zapewniają najlepszą dokładność predykcji w każdym następnym kroku prognozy.

Tabela 6. Najdokładniejsze modele stężeń PM_{10} dla luk pomiarowych o różnej długości.

Rodzaj błędu	Długość luki pomiarowej, godz.		
	1	2	>2
RMSE	TS-L	MR-P	MR-P
|e|	TS-L	TS-L	MR-P
RMSE/s	TS-L	MR-P	MR-P
r	TS-L	MR-P	MR-P

Uzyskane wyniki wskazują, że dla danych rejestrowanych na stacji monitoringu powietrza w Radomiu największą dokładność predykcji zapewniają modele TS-L i MR-P. Dokładność modeli regresji wielowymiarowej eksplorujących dane pochodzące z sąsiednich stacji monitoringu (EMR-P) jest dla wszystkich zanieczyszczeń mniejsza od dokładności wewnętrznych modeli regresyjnych (MR-P). Dlatego ta metoda nie może być rekomendowana do predykcji luk pomiarowych, o ile pozostałe z metod mogłyby być zastosowane. Niedokładność modeli EMR-P może wynikać z wyjątkowo nieregularnej emisji zanieczyszczeń pierwotnych pochodzących z lokalnych źródeł emisji znajdujących się w pobliżu stacji monitoringu powietrza w Radomiu, a także z usytuowania tej stacji na skraju obszaru monitorowanego przez pozostałe stacje pomiarowe.

NO jest jedynym zanieczyszczeniem, dla którego w całej luce pomiarowej największą dokładność aproksymacji stężeń zapewniają modele typu MR-P. Metoda szeregów czasowych może być zalecana do predykcji tylko w pierwszym kroku prognozy dla luk pomiarowych w seriach czasowych stężeń O_3, NO_2, CO. Wyjątkowo, można rozpatrywać stosowanie modeli TS-L do uzupełniania większej liczby przypadków - pierwszych 1-2 przypadków w luce dla serii czasowej stężeń PM_{10} oraz 2-4 przypadków w luce dla serii czasowej stężeń SO_2.

Wybór metody modelowania dla kolejnych miejsc luki pomiarowej jest jednoznaczny wtedy, gdy podczas wypełniania brakujących danych przyjmie się tylko jedną z miar błędów jako kryterium wyboru metody modelowania. Oceny dokładności wynikające z analizy różnych miar błędów nie są tak bardzo rozbieżne, aby można było odrzucić którąś z tych miar jako niewłaściwe kryterium oceny.

5. Wnioski

Na podstawie przeprowadzonej analizy sformułowano następujące wnioski:

1. Dla każdego z zanieczyszczeń powietrza należy rekomendować inne metody predykcji, ponieważ występują duże różnice w możliwościach modelowania stężeń poszczególnych zanieczyszczeń powietrza.
2. W celu zapewnienia optymalnej dokładności predykcji należy uwzględnić zmianę sposobu modelowania w miarę wydłużania horyzontu prognozy.
3. W przypadku NO zasadne jest stosowanie dla całej luki pomiarowej modeli regresji wielowymiarowej typu MR-P.
4. Stężenia zanieczyszczeń takich jak O_3, NO_2, CO, SO_2, PM_{10} można efektywnie modelować metodą szeregów czasowych, ale tylko dla bardzo krótkich horyzontów prognozy (1-4 godzin), po których regresyjne metody modelowania okazują się dokładniejsze.
5. Podczas wypełniania brakujących danych należy przyjąć tylko jedną z miar błędów jako kryterium wyboru metody modelowania. Wtedy wybór dla kolejnych miejsc luki pomiarowej będzie jednoznaczny.

Wnioski wynikające z przeprowadzonej analizy odnoszą się tylko do danych zarejestrowanych na stacji monitoringu powietrza w Radomiu i tylko do metod modelowania wykorzystywanych w autonomicznych modelach neuronowych. W wykonanym porównaniu nie uwzględniono innych możliwych technik modelowania, w tym interpolacji, która może być konkurencyjna dla krótkich luk w seriach czasowych.

Praca naukowa została wykonana w ramach badań własnych Politechniki Częstochowskiej BW 402-201/06. W analizie wykorzystano wyniki uzyskane w projekcie badawczym nr 1 T09D 037 30.

Literatura

1. Plaia A., Bondı A.L.: Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment, 40 (2006), is. 38, 7316–7330.
2. Gentili S., Magnaterra L., Passerini G.: An introduction to the statistical filling of environmental data time series. In Latini G., Passerini G. (Eds.): Handling Missing Data: Applications to Environmental Analysis. Wit Press, Southampton, Boston 2006.

3. Rozporządzenie Ministra Środowiska z dnia 17 grudnia 2008 r. w sprawie dokonywania oceny poziomów substancji w powietrzu (Dz. U. Nr 5, poz. 31).
4. Hoffman S.: Treating missing data at air monitoring stations. In Pawłowski L., Dudzińska M. R., Pawłowski A. (Eds.): Environmental Engineering, Taylor & Francis Group, London 2007, 349-353.
5. Hoffman S.: Short-Time forecasting of atmospheric NO_x concentration by neural networks. Environmental Engineering Science, 23 (4), 2006, 603-609.
6. Hoffman S.: Zastosowanie sieci neuronowych w modelowaniu regresyjnym stężeń zanieczyszczeń powietrza. Wydawnictwa Politechniki Częstochowskiej, Częstochowa 2004.
7. Hoffman S., Jasiński R.: Uzupełnianie brakujących danych w systemach monitoringu powietrza. Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2009.