Inżynieria i Ochrona Środowiska **2009, t. 12, nr 4, s. 307-325**

Szymon HOFFMAN, Rafał JASIŃSKI
Politechnika Częstochowska, Katedra Chemii, Technologii Wody i Ścieków
ul. Dąbrowskiego 69, 42-200 Częstochowa

Porównanie dokładności różnych metod predykcji stężeń zanieczyszczeń powietrza

W analizie wykorzystano dane zarejestrowane w latach 2004-2008 na ośmiu stacjach monitoringu powietrza działających w różnych miejscowościach województw łódzkiego i mazowieckiego. W pracy badano możliwości aproksymacji stężeń zanieczyszczeń mierzonych na stacjach monitoringu powietrza. Ocenę jakości modelowania wykonano poprzez porównanie modelowanych stężeń ze stężeniami rzeczywistymi. Do predykcji stężeń wykorzystano sieci neuronowe. Porównywano dokładność pięciu różnych grup modeli: modeli szeregów czasowych, liniowych modeli regresji wielowymiarowej, nieliniowych modeli regresji wielowymiarowej, liniowych modeli regresji wielowymiarowej eksplorujących dane pochodzące z sąsiednich stacji monitoringu i nieliniowych modeli regresji wielowymiarowej eksplorujących dane pochodzące z sąsiednich stacji monitoringu. Celem praktycznym była rekomendacja optymalnych technik modelowania luki pomiarowej obejmującej pewien dłuższy fragment serii czasowej tylko jednego z zanieczyszczeń powietrza przy założeniu, że są dostępne wszystkie pozostałe dane, w tym dane pochodzące z sąsiednich stacji monitoringu powietrza.

Wykonana analiza wykazała, że dla każdego z zanieczyszczeń powietrza należy rekomendować inne metody predykcji, ponieważ występują duże różnice w możliwościach modelowania poszczególnych zanieczyszczeń powietrza. Stężenia takich zanieczyszczeń, jak O_3, SO_2, PM_{10} można efektywnie modelować metodą szeregów czasowych, ale tylko do pewnego horyzontu prognozy, po którym regresyjne metody modelowania okazują się dokładniejsze. W modelowaniu stężeń O_3 i PM_{10} efektywne może się okazać wykorzystanie stężeń tych zanieczyszczeń zarejestrowanych na innych stacjach monitoringu powietrza. W przypadku pozostałych zanieczyszczeń NO, NO_2 i CO zasadne jest stosowanie tylko jednej metody modelowania - analizy regresji. Liniowe modele regresyjne są mniej dokładne od ich nieliniowych odpowiedników. Różnice dokładności obu typów modeli nie zawsze są duże. Dlatego modele liniowe mogą stanowić praktyczną alternatywę dla nieliniowych odpowiedników.

Słowa kluczowe: **zanieczyszczenia powietrza, monitoring powietrza, stężenia chwilowe, dane monitoringu, brakujące dane, luki pomiarowe, aproksymacja, modele szeregów czasowych, modele regresyjne, sieci neuronowe**

Wprowadzenie

Dane rejestrowane na stacjach monitoringu powietrza służą do oceny jakości powietrza w sąsiedztwie stacji. Ocena taka wymaga wysokiej kompletności zarejestrowanych danych. Zgodnie z obowiązującymi aktami prawnymi, serie pomiarowe stężeń chwilowych poszczególnych zanieczyszczeń powietrza powinny mieć określoną kompletność, na ogół powyżej 90% [1]. W przypadku niższej kompletności ocena nie ma mocy prawnej. Obowiązujące przepisy prawne dopuszczają możliwość wykorzystania modelowania w celu uzupełnienia brakujących danych,

gdy kompletność zbioru jest zbyt mała. Przepisy te nie rozstrzygają jednak, jakie metody modelowania powinny być stosowane.

Najlepszym rozwiązaniem jest stosowanie modeli autonomicznych, tj. modeli wykorzystujących do predykcji stężeń wiedzę zawartą wyłącznie w danych rejestrowanych na stacjach monitoringu powietrza [2]. Wielką zaletą takich modeli jest możliwość ich stosowania w instytucjach zarządzających sieciami monitoringu, bez konieczności sprowadzania danych spoza systemu. Wartości stężeń zanieczyszczeń powietrza charakteryzują się stosunkowo silną autoregresją, co umożliwia efektywne modelowanie za pomocą analizy szeregów czasowych [3]. Ponadto stężenia różnych zanieczyszczeń są silnie skorelowane między sobą. Są również zależne od parametrów meteorologicznych. To z kolei sprawia, że możliwa jest aproksymacja stężeń metodami regresyjnymi [4]. Stężenia tych samych zanieczyszczeń mierzonych na różnych stacjach monitoringu powietrza usytuowanych w tym samym regionie mogą również silnie ze sobą korelować, ponieważ warunki meteorologiczne na takich stacjach są na ogół podobne. Tła zanieczyszczeń w regionie także wykazują duże podobieństwo. Wyniki uzyskane za pomocą modeli regresyjnych wykorzystujących dane pochodzące z innych stacji monitoringu powietrza są obiecujące [5]. Dotychczas nie porównano możliwości predykcyjnych wymienionych metod modelowania. Porównanie ich dokładności umożliwiłoby rekomendację określonych metod do rozwiązywania problemu brakujących danych na stacjach monitoringu powietrza.

Zasadniczym celem przeprowadzonych badań było znalezienie optymalnych metod modelowania dla poszczególnych zanieczyszczeń powietrza. Analizę odniesiono do sytuacji, w której istnieje potrzeba modelowania luki pomiarowej, obejmującej dłuższy fragment serii czasowej tylko jednego z zanieczyszczeń powietrza, przy założeniu, że są dostępne kompletne dane dotyczące innych zanieczyszczeń, parametrów meteorologicznych oraz dane poprzedzające taką lukę pomiarową w zakresie stężeń modelowanego zanieczyszczenia. 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 kolejne wartości horyzontu prognozy.

W prezentowanej pracy analizie poddano wieloletni zbiór danych zarejestrowanych na 8 stacjach monitoringu powietrza usytuowanych w centralnej Polsce. Aproksymację stężeń przeprowadzono, stosując sztuczne sieci neuronowe.

1. Opis danych i metodyka obliczeń

1.1. Opis analizowanych danych

W analizie wykorzystano dane zarejestrowane w latach 2004-2008 na ośmiu stacjach monitoringu powietrza działających w różnych miejscowościach woje-

województw łódzkiego i mazowieckiego. Odpowiednim stacjom przypisano nazwy związane z ich usytuowaniem: Widzew, Gajew, Granica, Piotrków Trybunalski, Legionowo, Radom, Tłuszcz, Ursynów (rys. 1).

Rys. 1. Lokalizacja rozpatrywanych stacji monitoringu powietrza

Badania przeprowadzono, wykorzystując dane pomiarowe tzw. stężeń chwilowych (średnich stężeń 1-godzinnych) sześciu podstawowych zanieczyszczeń powietrza: O_3, NO_2, NO, 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 Widzew, a następnie dokonano oceny jakości modelowania poprzez porównanie modelowanych stężeń ze stężeniami rzeczywistymi. Podstawowe parametry statystyczne rzeczywistych serii czasowych stężeń zanieczyszczeń zarejestrowanych na stacji monitoringu powietrza Widzew przedstawiono w tabeli 1.

Tabela 1

Statystyka opisowa analizowanych serii czasowych 1-godzinnych stężeń zanieczyszczeń powietrza, Widzew 2004-2008

Parametr	Jednostka	O_3	NO	NO_2	CO	SO_2	PM_{10}
Kompletność serii	%	97,1	72,1	91,4	97,5	98,1	93,7
Średnia arytmetyczna	$\mu g/m^3$	58,1	3,3	18,1	446,8	14,8	22,7
Odchylenie standardowe s	$\mu g/m^3$	31,2	8,4	13,2	240,9	14,2	16,3
Wartość minimalna	$\mu g/m^3$	0,9	0,0	0,0	20,7	0,0	0,0
Wartość maksymalna	$\mu g/m^3$	197,9	323,3	170,1	5794,8	471,2	264,0

1.2. Opis metod predykcji stężeń

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 pięć podstawowych grup modeli:

1. TS-L - modele szeregów czasowych, oznaczone akronimem pochodzącym od nazwy angielskiej szeregów czasowych (time-series) i użytej liniowej (linear) techniki modelowania. 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. 24 wartości opóźnione obejmują całodobowy okres wcześniejszych pomiarów. Jak wykazano we wcześniejszych pracach, liczba taka gwarantuje wykorzystanie wiedzy o autoregresji, zawartej w analizowanych szeregach czasowych [6]. 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, 48. Taki zakres horyzontów prognozy pozwolił ocenić jakość modelowania predykcyjnego w 48-godzinnym okresie. Do analizy szeregów czasowych wykorzystano liniowe sieci neuronowe. Trening sieci liniowych odbywał się z użyciem algorytmu pseudoinwersji [7]. 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 [6].
2. MR-L - liniowe modele regresji wielowymiarowej (multiple regression - linear). 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: kierunek i prędkość wiatru, temperatura, natężenia promieniowania słonecznego, wilgotność względna. Do analizy regresji wykorzystano liniowe sieci neuronowe. Trening sieci liniowych odbywał się za pomocą algorytmu pseudoinwersji [7].
3. MR-P - nieliniowe modele regresji wielowymiarowej (multiple regression - perceptron). W tej grupie modeli predyktorami były identyczne zmienne, jak w grupie modeli MR-L. Zastosowano jednak inny typ sieci neuronowej, tzw. perceptron, umożliwiający tworzenie modeli nieliniowych. 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 [4]. Proces uczenia sieci podzielono na dwa etapy. W pierwszym etapie stosowano algorytm wstecznej propagacji błędów, a w drugim algorytm Levenberga-Marquardta.
4. EMR-L - liniowe modele regresji wielowymiarowej eksplorujące dane pochodzące z sąsiednich stacji monitoringu, oznaczone akronimem pochodzącym od

angielskich słów external, multiple regression, linear. 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). Do budowy modeli regresyjnych wykorzystano liniowe sieci neuronowe. Do treningu sieci użyto algorytmu pseudoinwersji.

5. EMR-P - nieliniowe modele regresji wielowymiarowej eksplorujące dane pochodzące z sąsiednich stacji monitoringu, oznaczone skrótem pochodzącym od angielskich słów external, multiple regression, perceptron. W tej grupie modeli predyktorami były identyczne zmienne jak w grupie modeli EMR-L. 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).

1.3. Ocena jakości modelowania

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 pierwiastka z błędu średniokwadratowego 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.

2. Wyniki badań

W tabelach 2-7 zamieszczono wyniki pozwalające określić jakość modelowania stężeń zanieczyszczeń powietrza za pomocą analizy szeregów czasowych. We wszystkich modelach szeregów czasowych (TS-L) przyjęto liczbę wartości opóźnionych równą 24, co oznacza, że w predykcji wykorzystano stężenia z pełnej doby pomiarowej. W tabelach znajdują się wartości czterech różnych miar błędu modelowania w zależności od horyzontu prognozy, dzięki czemu można obserwować zmiany dokładności predykcji przy wydłużaniu horyzontu prognozy od 1 godziny do 48 godzin.

Tabela 2

Ocena dokładności modelowania chwilowego stężenia O_3 w zależności od horyzontu prognozy (modele TS-L; Widzew 2004-2008; liczba wartości opóźnionych = 24)

Horyzont prognozy godz.	Błąd modelowania			
	RMSE µg/m^3	|e| µg/m^3	RMSE/s	r
1	7,96	5,26	0,255	0,967
2	12,15	8,59	0,390	0,921
3	14,79	10,82	0,474	0,880
4	16,52	12,32	0,530	0,848
5	17,69	13,36	0,567	0,823
6	18,50	14,10	0,593	0,805
8	19,45	14,96	0,624	0,782
12	20,24	15,71	0,649	0,761
24	20,29	15,84	0,651	0,759
48	22,26	17,59	0,714	0,700

Tabela 3

Ocena dokładności modelowania chwilowego stężenia NO w zależności od horyzontu prognozy (modele TS-L; Widzew 2004-2008; liczba wartości opóźnionych = 24)

Horyzont prognozy godz.	Błąd modelowania			
	RMSE µg/m^3	|e| µg/m^3	RMSE/s	r
1	5,59	1,56	0,668	0,745
2	7,45	2,20	0,890	0,458
3	8,01	2,54	0,957	0,290
4	8,17	2,69	0,977	0,214
5	8,24	2,76	0,984	0,176
6	8,26	2,79	0,987	0,160
8	8,28	2,83	0,990	0,144
12	8,30	2,86	0,993	0,123
24	8,32	2,87	0,994	0,110
48	8,34	2,88	0,997	0,073

Tabela 4

Ocena dokładności modelowania chwilowego stężenia NO_2 w zależności od horyzontu prognozy (modele TS-L; Widzew 2004-2008; liczba wartości opóźnionych = 24)

Horyzont prognozy godz.	Błąd modelowania			
	RMSE µg/m^3	|e| µg/m^3	RMSE/s	r
1	5,84	3,73	0,444	0,896
2	8,39	5,50	0,637	0,770
3	9,79	6,56	0,744	0,668
4	10,55	7,17	0,801	0,598
5	10,97	7,51	0,833	0,553
6	11,21	7,72	0,852	0,524
8	11,46	7,94	0,870	0,492
12	11,72	8,18	0,890	0,455
24	11,93	8,38	0,906	0,423
48	12,47	8,87	0,948	0,319

Tabela 5

Ocena dokładności modelowania chwilowego stężenia CO w zależności od horyzontu prognozy (modele TS-L; Widzew 2004-2008; liczba wartości opóźnionych = 24)

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	92,3	45,5	0,383	0,924
2	135,8	70,7	0,564	0,826
3	156,6	85,2	0,650	0,760
4	167,4	93,5	0,695	0,719
5	173,7	99,3	0,721	0,693
6	177,6	102,6	0,737	0,676
8	182,3	107,5	0,757	0,654
12	188,5	113,5	0,783	0,623
24	198,1	121,6	0,822	0,569
48	210,8	133,5	0,875	0,484

Tabela 6

Ocena dokładności modelowania chwilowego stężenia SO_2 w zależności od horyzontu prognozy (modele TS-L; Widzew 2004-2008; liczba wartości opóźnionych = 24)

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	7,70	3,70	0,541	0,841
2	9,73	5,03	0,683	0,730
3	10,51	5,62	0,738	0,675
4	10,89	5,93	0,765	0,644
5	11,11	6,14	0,781	0,625
6	11,27	6,29	0,792	0,611
8	11,44	6,45	0,804	0,595
12	11,70	6,66	0,822	0,570
24	12,08	6,98	0,849	0,529
48	12,68	7,38	0,891	0,455

Tabela 7

Ocena dokładności modelowania chwilowego stężenia PM_{10} w zależności od horyzontu prognozy (modele TS-L; Widzew 2004-2008; liczba wartości opóźnionych = 24)

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	7,70	4,72	0,474	0,881
2	9,95	6,40	0,612	0,791
3	11,09	7,33	0,682	0,731
4	11,78	7,92	0,725	0,689
5	12,26	8,34	0,754	0,657
6	12,63	8,64	0,777	0,629
8	13,10	9,01	0,806	0,592
12	13,70	9,50	0,843	0,538
24	14,66	10,22	0,902	0,433
48	15,63	11,02	0,961	0,276

W kolejnych tabelach zamieszczono wartości błędów statystycznych pozwalające określić jakość modelowania stężeń poszczególnych zanieczyszczeń powietrza za pomocą modeli regresyjnych liniowych typu MR-L (tab. 8) i analogicznych modeli nieliniowych MR-P (tab. 9). Tabele zawierają wartości czterech różnych miar błędu modelowania i pozwalają porównać jakość tych modeli z modelami innego typu.

Tabela 8

Ocena dokładności modelowania stężeń chwilowych podstawowych zanieczyszczeń powietrza metodą liniowej regresji wielowymiarowej (modele MR-L; Widzew 2004-2008)

Rodzaj błędu	O_3	NO	NO_2	CO	SO_2	PM_{10}
RMSE, $\mu g/m^3$	16,22	5,80	7,02	98,5	12,03	10,13
$\|e\|$, $\mu g/m^3$	12,86	2,88	5,11	70,9	7,19	7,21
RMSE/s	0,509	0,714	0,522	0,419	0,830	0,613
r	0,861	0,703	0,853	0,908	0,558	0,790

Tabela 9

Ocena dokładności modelowania stężeń chwilowych podstawowych zanieczyszczeń powietrza metodą nieliniowej regresji wielowymiarowej (modele MR-P; Widzew 2004-2008).

Rodzaj błędu	O_3	NO	NO_2	CO	SO_2	PM_{10}
RMSE, $\mu g/m^3$	12,92	2,52	5,60	85,2	10,16	9,27
$\|e\|$, $\mu g/m^3$	9,85	1,03	4,06	61,3	6,22	6,59
RMSE/s	0,405	0,310	0,416	0,363	0,701	0,561
r	0,914	0,951	0,909	0,932	0,714	0,827

W tabelach 10 i 11 podano wartości błędów modelowania stężeń poszczególnych zanieczyszczeń powietrza za pomocą modeli regresyjnych typu EMR-L i EMR-P, wykorzystujących dane pochodzące z sąsiednich stacji monitoringu.

Tabela 10

Ocena dokładności modelowania stężeń chwilowych podstawowych zanieczyszczeń powietrza metodą liniowej regresji wielowymiarowej eksplorującej dane pochodzące z sąsiednich stacji monitoringu (modele EMR-L; Widzew 2004-2008)

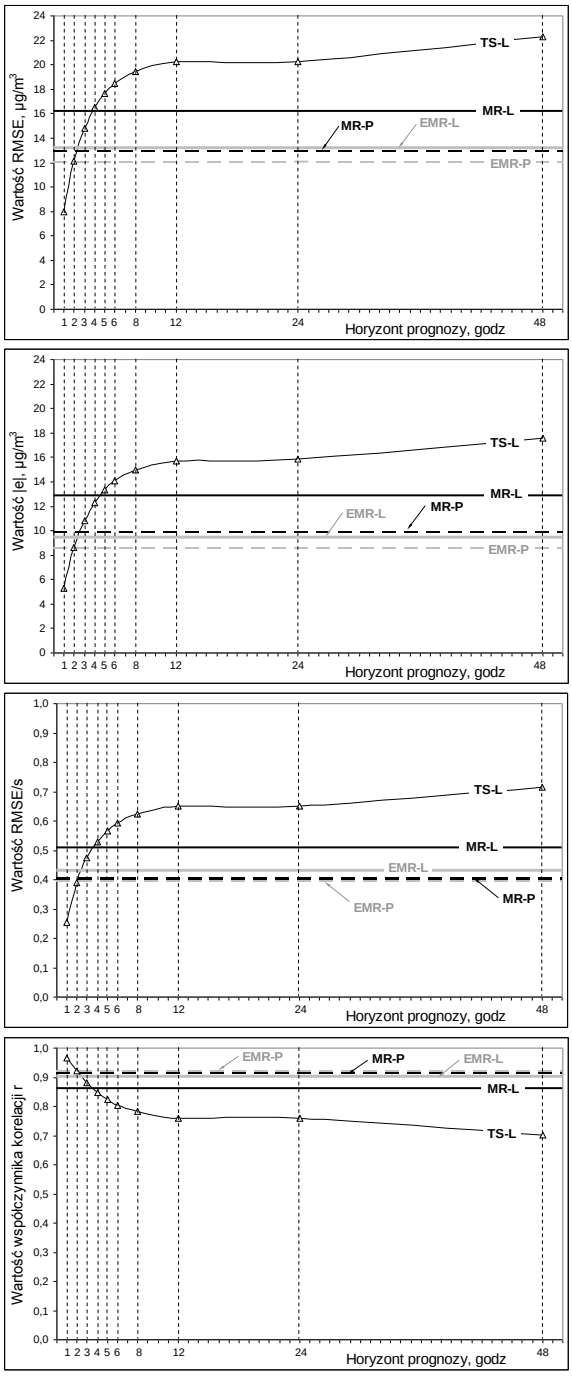
Rodzaj błędu	O_3	NO	NO_2	CO	SO_2	PM_{10}
RMSE, $\mu g/m^3$	13,19	6,81	8,49	199,2	12,30	10,15
$\|e\|$, $\mu g/m^3$	9,47	2,07	5,92	127,8	7,34	6,56
RMSE/s	0,431	0,871	0,673	0,774	0,859	0,612
r	0,903	0,491	0,740	0,635	0,511	0,791

Tabela 11

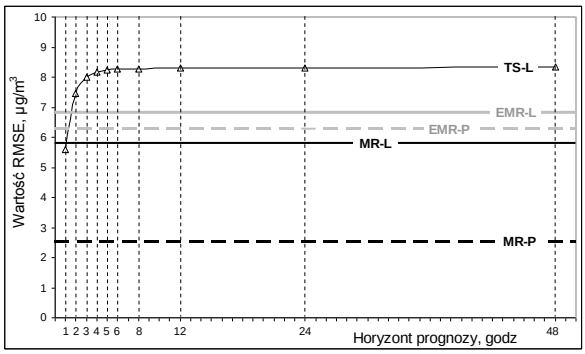
Ocena dokładności modelowania stężeń chwilowych podstawowych zanieczyszczeń powietrza metodą nieliniowej regresji wielowymiarowej eksplorującej dane pochodzące z sąsiednich stacji monitoringu (modele EMR-P; Widzew 2004-2008)

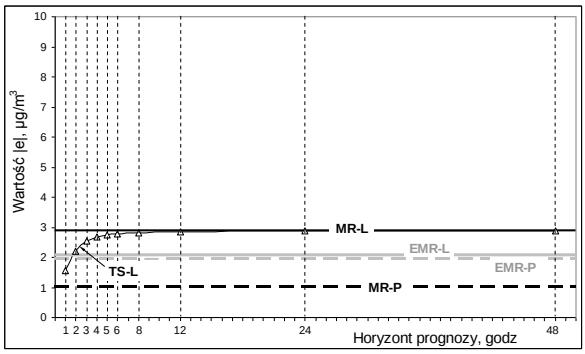
Rodzaj błędu	O_3	NO	NO_2	CO	SO_2	PM_{10}
RMSE, $\mu g/m^3$	12,08	6,27	8,22	187,8	12,15	9,46
$\|e\|$, $\mu g/m^3$	8,55	1,93	5,73	119,7	7,21	6,18
RMSE/s	0,394	0,803	0,651	0,729	0,849	0,571
r	0,919	0,597	0,759	0,684	0,529	0,821

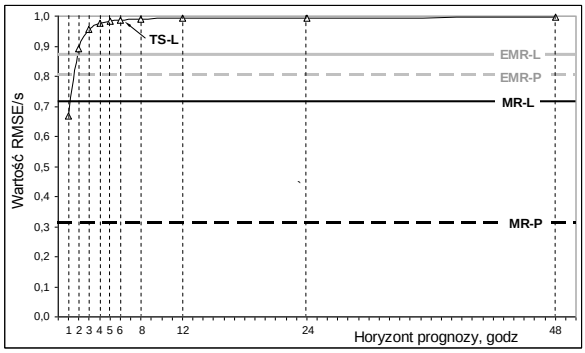
Wyniki zamieszczone w tabelach 2-11 zestawiono na rysunkach 2-7 w celu graficznego porównania jakości predykcji.

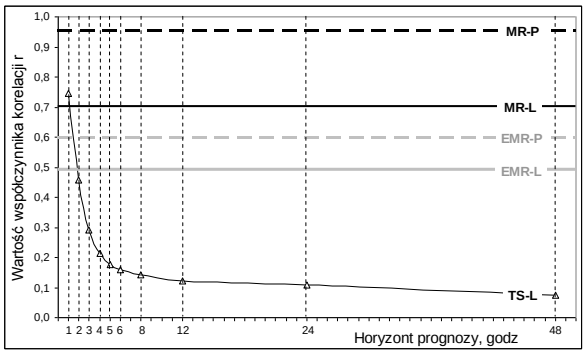


Rys. 2. Błąd modelowania stężeń O_3 w zależności od horyzontu prognozy





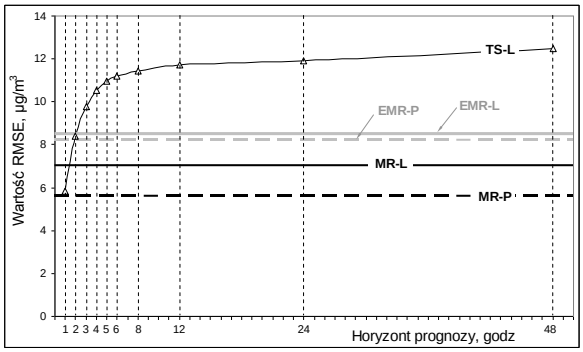


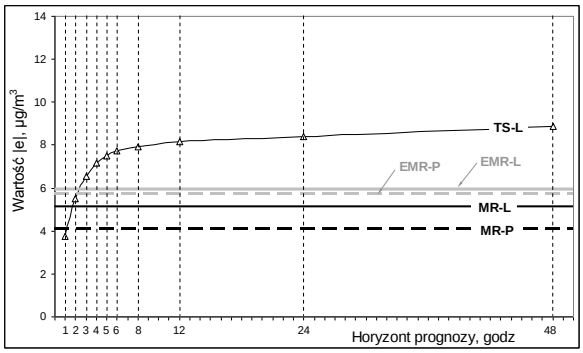


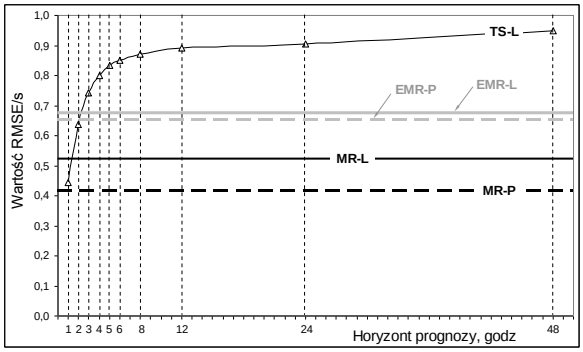
Legenda:

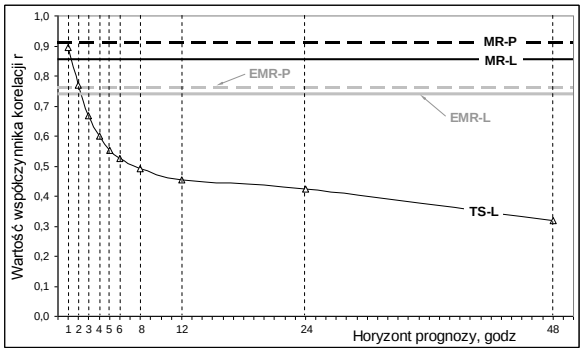
- **TS-L** – liniowy model szeregów czasowych;
- **MR-L** – liniowy model regresji wielowymiarowej
- **MR-P** – nieliniowy model regresji wielowymiarowej;
- **EMR-L** – liniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych;
- **EMR-P** – nieliniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych.

Rys. 3. Błąd modelowania stężeń NO w zależności od horyzontu prognozy





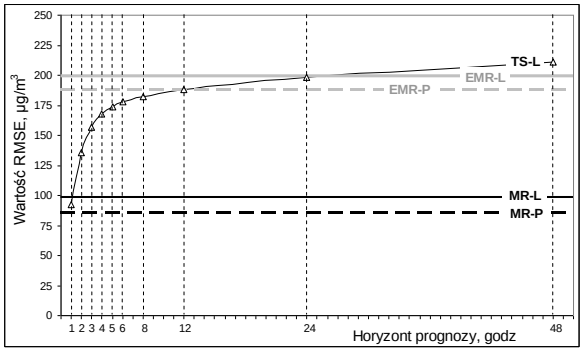


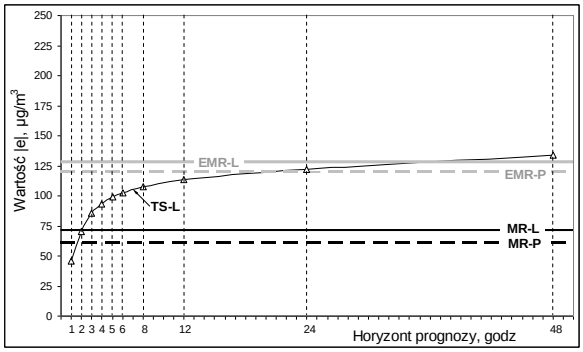


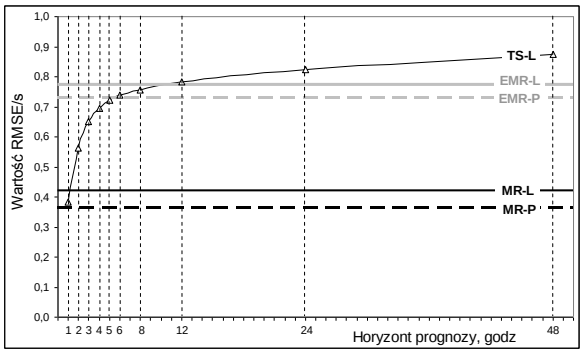
Legenda:

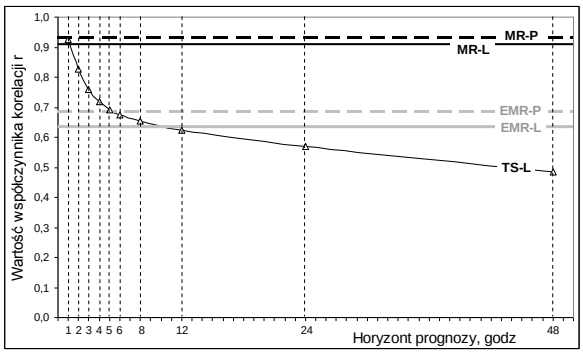
- **TS-L** – liniowy model szeregów czasowych;
- **MR-L** – liniowy model regresji wielowymiarowej
- **MR-P** – nieliniowy model regresji wielowymiarowej;
- **EMR-L** – liniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych;
- **EMR-P** – nieliniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych.

Rys. 4. Błąd modelowania stężeń NO_2 w zależności od horyzontu prognozy





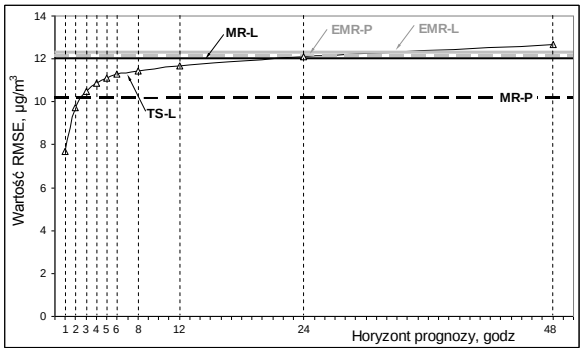


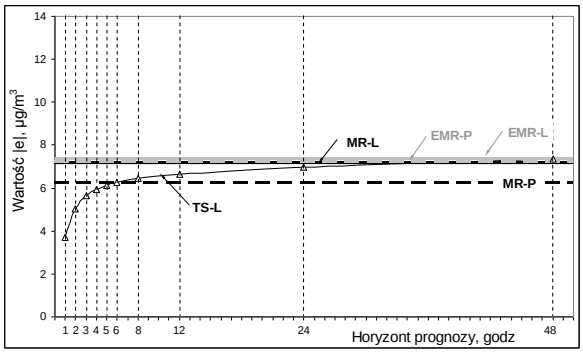


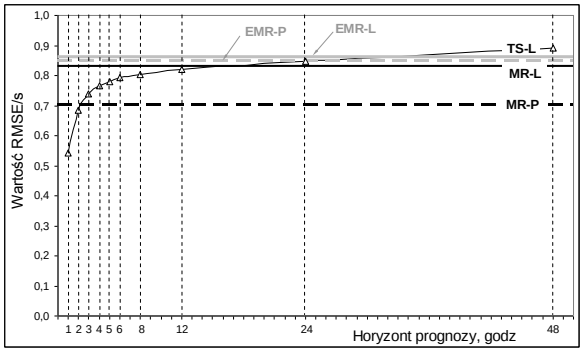
Legenda:

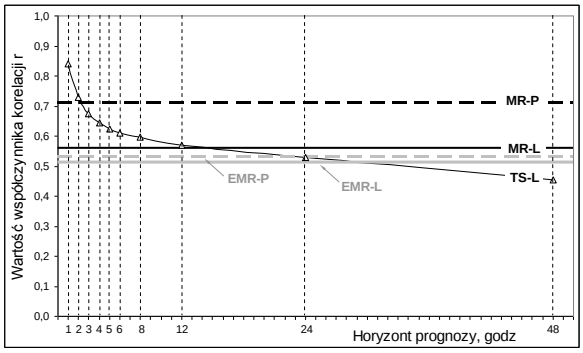
- **TS-L** – liniowy model szeregów czasowych;
- **MR-L** – liniowy model regresji wielowymiarowej
- **MR-P** – nieliniowy model regresji wielowymiarowej;
- **EMR-L** – liniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych;
- **EMR-P** – nieliniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych.

Rys. 5. Błąd modelowania stężeń CO w zależności od horyzontu prognozy





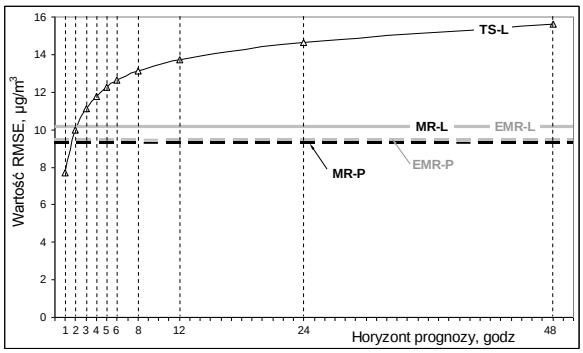


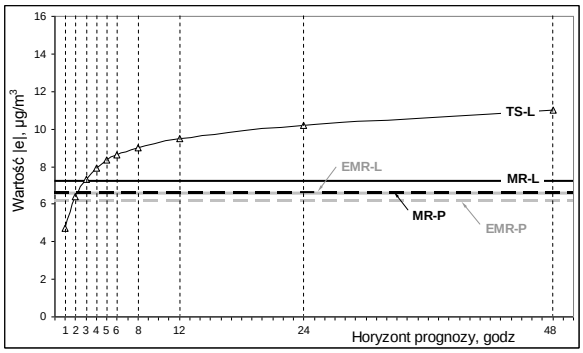


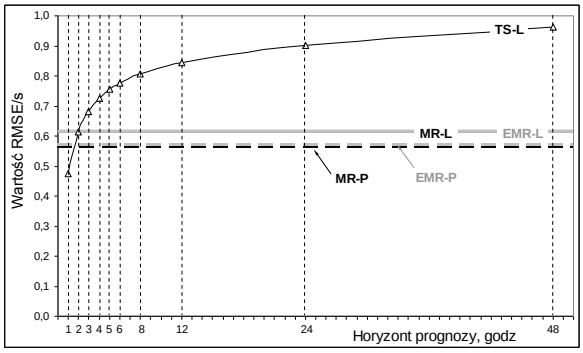
Legenda:

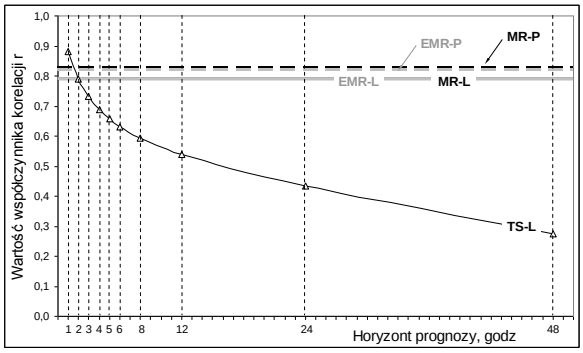
- **TS-L** – liniowy model szeregów czasowych;
- **MR-L** – liniowy model regresji wielowymiarowej
- **MR-P** – nieliniowy model regresji wielowymiarowej;
- **EMR-L** – liniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych;
- **EMR-P** – nieliniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych.

Rys. 6. Błąd modelowania stężeń SO_2 w zależności od horyzontu prognozy









Legenda:

- **TS-L** – liniowy model szeregów czasowych;
- **MR-L** – liniowy model regresji wielowymiarowej
- **MR-P** – nieliniowy model regresji wielowymiarowej;
- **EMR-L** – liniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych;
- **EMR-P** – nieliniowy model regresji wielowymiarowej eksplorujący dane pochodzące z sąsiednich stacji monitoringowych.

Rys. 7. Błąd modelowania stężeń PM_{10} w zależności od horyzontu prognozy

Na rysunkach przedstawiono wykresy zmian wartości błędów modelowania stężeń zanieczyszczeń powietrza w zależności od horyzontu prognozy. Na oddzielnych rysunkach zamieszczono wyniki otrzymane 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 48 godzin. Na poszczególnych wykresach porównano wyniki dla modeli wygenerowanych 5 różnymi technikami predykcji, w tym dla modeli szeregów czasowych (TS-L), dla liniowych modeli regresji wielowymiarowej (MR-L), dla nieliniowych modeli regresji wielowymiarowej (MR-P), dla liniowych modeli regresji wielowymiarowej eksplorujących dane pochodzące z sąsiednich stacji monitoringu (EMR-L) i dla nieliniowych modeli regresji wielowymiarowej eksplorujących dane pochodzące z sąsiednich stacji monitoringu (EMR-P).

3. Dyskusja wyników i podsumowanie

Wśród rozpatrywanych grup modeli tylko pierwsza z nich, tj. grupa modeli szeregów czasowych (TS-L), charakteryzuje się zmiennością w miarę wydłużania horyzontu prognozy. Pozostałe cztery grupy modeli regresyjnych zawsze mają stałą wartość błędu, niezależną od horyzontu prognozy. W przypadku modeli TS-L 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 tej grupie modeli silnie zależy od horyzontu prognozy.

Porównanie dokładności modelowania pozwala wybrać najdokładniejszą metodę predykcji, a w szczególności ocenić jakość prognozy uzyskanej za pomocą modeli szeregów czasowych na tle alternatywnych metod aproksymacji. Dokładność modeli szeregów czasowych dość szybko maleje w miarę wydłużania horyzontu prognozy. Tylko dla najkrótszych horyzontów prognozy metoda ta może być dokładniejsza od pozostałych; tak jest w przypadku stężeń O_3, SO_2, PM_{10}. Analiza wszystkich rozpatrywanych kategorii błędu jednoznacznie wskazuje, że dla tych trzech zanieczyszczeń modele szeregów czasowych są najdokładniejszymi modelami predykcyjnymi dla najkrótszych horyzontów prognozy. W tabelach 12-17 zestawiono najdokładniejsze metody modelowania stężeń poszczególnych zanieczyszczeń powietrza dla kolejnych horyzontów prognozy.

W przypadku stężeń ozonu modele TS-L zapewniają najwyższą jakość predykcji dla pierwszego kroku prognozy (rys. 2). Dla drugiego kroku dokładność tych modeli staje się porównywalna z dokładnością modeli typu EMR-P (tab. 12). Dla horyzontu prognozy 3 i wyższych to właśnie model typu EMR-P staje się modelem optymalnym.

Tabela 12

Najdokładniejsze metody modelowania stężeń O_3 dla różnych horyzontów prognozy. Rekomendacja na podstawie wyników analizy błędów predykcji

Rodzaj błędu	Horyzont prognozy, godz.			
	1	2	3	>3
RMSE	TS-L	EMR-P, TS-L	EMR-P	EMR-P
|e|	TS-L	EMR-P, TS-L	EMR-P	EMR-P
RMSE/s	TS-L	TS-L, EMR-P	EMR-P	EMR-P
r	TS-L	TS-L, EMR-P	EMR-P	EMR-P

W przypadku stężeń NO modele TS-L są wyjątkowo mało dokładne, nawet dla najkrótszych horyzontów prognozy (tab. 3, rys. 3). Model regresyjny typu MR-P jest natomiast wyjątkowo dokładny, nawet w porównaniu do analogicznych modeli dla innych zanieczyszczeń (tab. 9). Współczynnik korelacji dla tego modelu wynosi 0,951. Dlatego model ten należy rekomendować do prognozowania stężeń NO już od pierwszego kroku prognozy (tab. 13).

Tabela 13

Najdokładniejsze metody modelowania stężeń NO dla różnych horyzontów prognozy. Rekomendacja na podstawie wyników analizy błędów predykcji

Rodzaj błędu	Horyzont prognozy godz.	
	1	>1
RMSE	MR-P	MR-P
|e|	MR-P	MR-P
RMSE/s	MR-P	MR-P
r	MR-P	MR-P

Dla stężeń NO_2 zarówno model typu TS-L dla horyzontu prognozy wynoszącego 1, jak i model regresyjny MR-P mają porównywalną dokładność (rys. 4). W zależności od rozpatrywanej kategorii błędu jeden z nich można uznać za optymalny. Dla dłuższych horyzontów prognozy najdokładniejszym modelem jest model MR-P (tab. 14).

Tabela 14

Najdokładniejsze metody modelowania stężeń NO_2 dla różnych horyzontów prognozy. Rekomendacja na podstawie wyników analizy błędów predykcji

Rodzaj błędu	Horyzont prognozy, godz.		
	1	2	>2
RMSE	MR-P	MR-P	MR-P
|e|	TS-L	MR-P	MR-P
RMSE/s	MR-P	MR-P	MR-P
r	MR-P, TS-L	MR-P	MR-P

W przypadku stężeń CO można rekomendować podobne typy modeli jak dla stężeń NO_2. Model typu TS-L dla pierwszego kroku prognozy oraz model regresyjny MR-P mają porównywalną dokładność (rys. 5). W zależności od rozpatrywanej ka-

tegorii błędu jeden z nich można uznać za optymalny. Dla dłuższych horyzontów prognozy najdokładniejszym modelem jest model regresyjny typu MR-P (tab. 15).

Tabela 15

Najdokładniejsze metody modelowania stężeń CO dla różnych horyzontów prognozy. Rekomendacja na podstawie wyników analizy błędów predykcji

Rodzaj błędu	Horyzont prognozy, godz.		
	1	2	>2
RMSE	MR-P	MR-P	MR-P
|e|	TS-L	MR-P	MR-P
RMSE/s	MR-P	MR-P	MR-P
r	MR-P, TS-L	MR-P	MR-P

Najlepsze rezultaty w modelowaniu stężeń SO_2 dla dwóch pierwszych kroków prognozy można uzyskać, wykorzystując modele szeregów czasowych TS-L (rys. 6). Rozpatrując kategorię średniego błędu bezwzględnego |e|, ta metoda okazuje się najdokładniejsza aż do horyzontów prognozy równych 5-6 (tab. 16). Metodą najbardziej konkurencyjną w stosunku do analizy szeregów czasowych jest analiza regresji. Modele MR-P stają się najdokładniejszymi modelami dla dłuższych horyzontów prognozy.

Tabela 16

Najdokładniejsze metody modelowania stężeń SO_2 dla różnych horyzontów prognozy. Rekomendacja na podstawie wyników analizy błędów predykcji

Rodzaj błędu	Horyzont prognozy, godz.						
	1	2	3	4	5	6	>6
RMSE	TS-L	TS-L	MR-P	MR-P	MR-P	MR-P	MR-P
|e|	TS-L	TS-L	TS-L	TS-L	TS-L	MR-P, TS-L	MR-P
RMSE/s	TS-L	TS-L	MR-P	MR-P	MR-P	MR-P	MR-P
r	TS-L	TS-L	MR-P	MR-P	MR-P	MR-P	MR-P

Dla stężeń PM_{10} najlepsze wyniki modelowania w pierwszym kroku prognozy daje model szeregów czasowych TS-L (rys. 7). Dla dłuższych horyzontów prognozy najdokładniejsze i konkurencyjne względem siebie okazały się modele regresyjne MR-P i EMR-P (tab. 17). Dokładności obu modeli są porównywalne. W zależności od rozpatrywanej kategorii błędu jeden z nich można uznać za optymalny.

Tabela 17

Najdokładniejsze metody modelowania stężeń PM_{10} dla różnych horyzontów prognozy. Rekomendacja na podstawie wyników analizy błędów predykcji

Rodzaj błędu	Horyzont prognozy, godz.		
	1	2	>2
RMSE	TS-L	MR-P	MR-P
|e|	TS-L	EMR-P	EMR-P
RMSE/s	TS-L	MR-P, EMR-P	MR-P, EMR-P
r	TS-L	EMP-P, MR-P	EMP-P, MR-P

Wynikiem przeprowadzonej analizy było porównanie dokładności modelowania stężeń podstawowych zanieczyszczeń powietrza, rejestrowanych na stacjach monitoringu powietrza. Porównywano błędy predykcji pięciu różnych grup modeli. Praktycznym rezultatem tak przeprowadzonej analizy była rekomendacja optymalnych technik modelowania luki pomiarowej, obejmującej pewien fragment serii czasowej tylko jednego z zanieczyszczeń powietrza, przy założeniu, że są dostępne wszystkie pozostałe dane, w tym dane pochodzące z sąsiednich stacji monitoringu powietrza. Wcześniejsze badania pozwoliły sformułować tezę, że analiza szeregów czasowych może być rekomendowaną metodą modelowania, ale tylko dla najkrótszych horyzontów prognozy i nie dla wszystkich zanieczyszczeń powietrza [6, 8].

W wyniku dokonanej analizy stwierdzono, że występują duże różnice w możliwościach modelowania poszczególnych zanieczyszczeń powietrza. Stężenia zanieczyszczeń, takich jak O_3, SO_2, PM_{10}, można efektywnie modelować metodą szeregów czasowych, ale tylko do pewnego horyzontu prognozy, po którym regresyjne metody modelowania okazują się dokładniejsze. W przypadku tych zanieczyszczeń dla każdej dłuższej luki pomiarowej należy rekomendować różne metody modelowania - modele szeregów czasowych na początku i końcu luki, a modele regresyjne w środku tej luki. W modelowaniu stężeń O_3 i PM_{10} efektywne może się okazać wykorzystanie stężeń tych zanieczyszczeń zarejestrowanych na innych stacjach monitoringu powietrza.

W przypadku pozostałych trzech zanieczyszczeń NO, NO_2 i CO zasadne jest stosowanie tylko jednej metody modelowania - analizy regresji. Chociaż dla stężeń CO i NO_2 w pierwszym kroku prognozy porównywalną dokładność wykazują modele szeregów czasowych, to w następnych krokach prognozy ich dokładność jest zdecydowanie mniejsza.

Zgodnie z przewidywaniami liniowe modele regresyjne są mniej dokładne od nieliniowych odpowiedników. Różnice dokładności obu typów modeli nie zawsze są duże. Zwłaszcza dla modeli typu EMR większość modeli liniowych ma zbliżoną dokładność do modeli nieliniowych. Ponieważ regresję liniową można analizować za pomocą stosunkowo prostych i powszechnie dostępnych programów statystycznych, to modele liniowe mogą stanowić praktyczną alternatywę dla ich nieliniowych odpowiedników.

Wnioski

Na podstawie przeprowadzonych badań można sformułować następujące wnioski ogólne:

1. Dla każdego z zanieczyszczeń powietrza należy rekomendować inne metody predykcji, ponieważ występują duże różnice w możliwościach modelowania poszczególnych zanieczyszczeń powietrza.
2. Stężenia zanieczyszczeń, takich jak O_3, SO_2, PM_{10}, można efektywnie modelować metodą szeregów czasowych, ale tylko do pewnego horyzontu prognozy, po którym regresyjne metody modelowania okazują się dokładniejsze.

3. W modelowaniu regresyjnym stężeń O_3 i PM_{10} efektywne może się okazać wykorzystanie stężeń tych zanieczyszczeń zarejestrowanych na innych stacjach monitoringu powietrza.
4. W przypadku NO, NO_2 i CO zasadne jest stosowanie tylko jednej metody modelowania - analizy regresji.
5. Liniowe modele regresyjne są mniej dokładne od nieliniowych odpowiedników. Różnice dokładności obu typów modeli nie zawsze są duże. Dlatego modele liniowe mogą stanowić praktyczną alternatywę dla ich nieliniowych odpowiedników.

Uzyskane wyniki badań odnoszą się do możliwości implementacji danych w dużej luce pomiarowej, obejmującej serię czasową wybranego zanieczyszczenia powietrza. Wnioski wynikające z tej analizy mogą być niezasadne w odniesieniu do krótkich czasowo luk pomiarowych, dla których metody interpolacyjne mogą być dokładniejsze od rozpatrywanych w tej pracy.

Praca naukowa finansowana ze środków na naukę w latach 2006-2008 jako projekt badawczy nr 1 T09D 037 30.

Literatura

[1] Rozporządzenie Ministra Środowiska z dnia 17 grudnia 2008 r. w sprawie dokonywania oceny poziomów substancji w powietrzu, DzU Nr 5, poz. 31.

[2] Hoffman S., Treating missing data at air monitoring stations, [w:] Environmental Engineering, eds. L. Pawłowski, M.R. Dudzińska, A. Pawłowski, Taylor & Francis Group, London 2007, 349-353.

[3] Brockwell P.J., Davis R.A., Introduction to time series and forecasting, Springer-Verlag 2002.

[4] Hoffman S., Zastosowanie sieci neuronowych w modelowaniu regresyjnym stężeń zanieczyszczeń powietrza, Wydawnictwa Politechniki Częstochowskiej, Częstochowa 2004.

[5] Hoffman S., Ozone concentration modelling at air monitoring stations using external data, Polish Journal of Environmental Studies 2009, 18, 2B, 182-186.

[6] Hoffman S., Short-Time forecasting of atmospheric NO_x concentration by neural networks, Environmental Engineering Science 2006, 23(4), 603-609.

[7] Statistica Neural Networks, StatSoft 1998.

[8] Hoffman S.: Missing data completing in the air monitoring systems by forward and backward prognosis methods, Environmental Protection Engineering 2006, 32(4), 25-29.

A Comparison of Accuracies of Different Air Pollutants Concentration Prediction Methods

Air monitoring data collected over a 5-year period at 8 different measure sites in Central Poland were used as the database for analysis purposes. Approximation of concentrations of monitored air pollutants were done by means of several prediction methods: time series analysis, regression analysis with predictors from a single monitoring station, and regression analysis with external predictors. Separate models were created for O_3, NO_2, NO, PM_{10}, SO_2, CO. Modelled and measured concentrations were compared. As a result prediction errors were calculated for each model. The main objective of analysis was a comparison of prediction results, and recommendation the most accurate modelling methods, dedicated to specified pollutants. The examination was made by means of artificial neural networks, which were employed to create all types of models.

Keywords: air pollution, air monitoring, hourly concentrations, monitoring data, missing data, measure gaps, approximation, time series models, regression models, neural networks.