UZUPEŁNIANIE BRAKUJĄCYCH DANYCH METODĄ K-NAJBLIŻSZYCH SĄSIADÓW

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

STRESZCZENIE

W pracy przetestowano możliwość zastosowania statystycznej metody k-najbliższych sąsiadów do modelowania imisji. Analizę przeprowadzono wykorzystując 3-letnie zbiory chwilowych danych pomiarowych, zarejestrowane na stacji monitoringu powietrza działającej w Łodzi-Widzewie. Modelowano stężenia ozonu i NO, wykorzystując pozostałe zmienne pomiarowe jako predyktory. Oceniono błąd modelowania dla różnych wartości k, w zakresie od 1 do 20.

Wykaz oznaczeń:

- $\bar{x}$ – wartość średnia (obserwowanych $\bar{x}_{obs}$, przewidywanych $\bar{x}_{pred}$), w $\mu g/m^3$;
- s – odchylenie standardowe (obserwowanych s_{obs}, przewidywanych s_{pred}), w $\mu g/m^3$;
- RMSE – pierwiastek z błędu średniokwadratowego, w $\mu g/m^3$;
- r – współczynnik korelacji;
- |e| – średni błąd bezwzględny, w $\mu g/m^3$;

Symbole zmiennych:

- D – dzień pomiaru w postaci numerycznej;
- G – godzina pomiaru w postaci numerycznej w przedziale <0 , 1>;
- PW – prędkość wiatru, w m/s;
- T – temperatura, w °C;
- R – natężenie promieniowania słonecznego, w W/m^2;
- W_{wzg} – wilgotność względna, w %;
- O_3 – stężenie ozonu $\mu g/m^3$;
- NO_2 – stężenie NO_2, w $\mu g/m^3$;
- NO – stężenie NO, w $\mu g/m^3$;
- SO_2 – stężenie SO_2, w $\mu g/m^3$;
- CO – stężenie CO, w $\mu g/ m^3$;
- PM – stężenie pyłu zawieszonego PM_{10}, w $\mu g/m^3$;

1. Wprowadzenie

Dane zbierane na automatycznych stacjach monitoringu powietrza nie są kompletne. W skali roku ilość braków wynosi zwykle kilka procent wszystkich możliwych do zarejestrowania danych, ale dla niektórych mierzonych wielkości może osiągnąć kilkanaście, a nawet kilkadziesiąt procent. Brakujące dane imisyjne utrudniają ocenę czystości powietrza wymaganą przez regulacje prawne. Do uzupełniania danych monitoringu wykorzystuje się techniki modelowania [1, 2]. Wciąż poszukiwane są nowe metody umożliwiające predykcję stężenia wybranego zanieczyszczenia. Najbardziej obiecujące są metody regresyjne, wykorzystujące wiedzę o zależnościach między zmiennym, uzyskaną poprzez analizę zgromadzonych wieloletnich danych. W pracy oceniono jakość modelowania metodą k-najbliższych sąsiadów, jedną z technik tzw. uczenia maszyn.

Celem przeprowadzonej analizy było określenie możliwości zastosowania metody k-najbliższych sąsiadów do modelowania imisji. Analizę przeprowadzono wykorzystując

3-letni zbiór chwilowych danych pomiarowych, zarejestrowany na stacji monitoringu powietrza w Łodzi-Widzewie. Modelowano stężenia ozonu i NO.

2. Metodyka badań

Dane monitoringowe pochodziły z lat 2004-2006, ze stacji pomiarowej zlokalizowanej w Łodzi-Widzewie. Analizowany zbiór danych zawierał średnie 1-godzinne wartości stężeń podstawowych zanieczyszczeń powietrza oraz wartości parametrów meteorologicznych.

W statystycznej metodzie k-najbliższych sąsiadów jest stosowana idea prototypów. Zakłada się, że przypadki podobne w przestrzeni wielowymiarowej znajdą się w tej samej klasie. W danych treningowych są szukane przypadki, które w przestrzeni zmiennych objaśniających są najbardziej podobne do nowego przypadku [3, 4]. Procedura wstępna (klasyfikacja) polega na znalezieniu określonej liczby (k) najbliższych sąsiadów. Następnie, wartości zmiennej modelowanej dla wytypowanych przypadków są uśredniane, a otrzymana średnia jest traktowana jako wynik predykcji.

Do wykonania obliczeń wykorzystano program komputerowy Statistica Data Miner. Jako miarę odległości w przestrzeni wielowymiarowej przyjęto odległość euklidesową. Wszystkie zmienne zostały poddane wstępnej normalizacji. Zastosowano uśrednianie jednorodne (identyczne wagi dla wszystkich sąsiadów). 75% losowo wybranych przypadków zakwalifikowano do próby uczącej, a 25% do próby testowej. Jako błąd modelowania przyjęto błąd wyznaczony dla próby testowej. Dokładność modelowania oszacowano dla różnych wartości k w zakresie od 1 do 20, przy tej samej reprezentacji zmiennych objaśniających (predykatorów).

Ponieważ nie ma uniwersalnego kryterium jakości modelowania, do oceny dokładności modelowania wykorzystano cztery różne miary błędu: pierwiastek z błędu średniokwadratowego RMSE, średni błąd bezwzględny |e|, stosunek RMSE do odchylenia standardowego $RMSE/s_{obs}$ i współczynnik korelacji r.

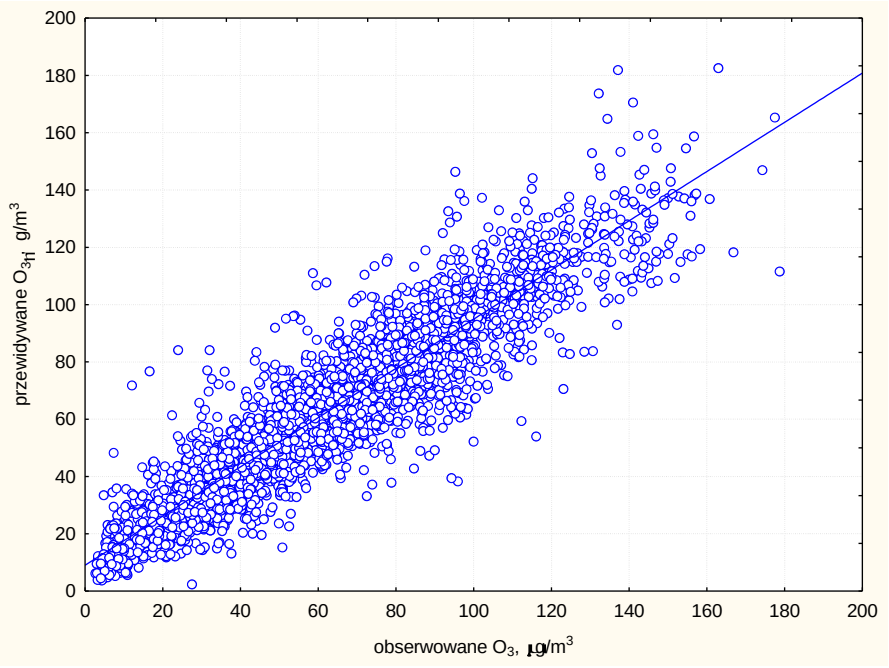
2. Wyniki i ich omówienie

W tabeli 1 przedstawiono wyniki modelowania stężenia ozonu. Najlepszą dokładność modelowania uzyskano dla k równego 2 (kryterium oceny: |e|) lub 3 (kryterium oceny: RMSE, $RMSE/s_{obs}$, r). Dla k = 1 i dla k = 20 jakość predykcji była wyraźnie gorsza. Na podstawie przedstawionych wyników można sformułować wniosek, że najlepsze wyniki modelowania stężenia O_3 uzyskuje się dla stosunkowo niedużej liczby najbliższych sąsiadów.

Tabela 1. Wyniki modelowania stężenia ozonu metodą k-najbliższych sąsiadów; predyktory: D, G, CO, NO, NO_2, SO_2, PM_{10}, PW, T, R, W_{wzg}

Statystyki opisowe i miary błędu modelowania	Liczba najbliższych sąsiadów										
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=20
$\bar{x}_{obs}$	62,6	62,6	62,6	62,6	62,6	62,6	62,6	62,6	62,6	62,6	62,6
$\bar{x}_{pred}$	62,8	62,7	62,8	62,7	62,7	62,6	62,6	62,6	62,5	62,5	62,6
s_{obs}	32,4	32,4	32,4	32,4	32,4	32,4	32,4	32,4	32,4	32,4	32,4
s_{pred}	32,0	30,7	30,0	29,5	29,1	28,8	28,6	28,4	28,2	28,0	27,2
RMSE	13,89	12,33	12,11	12,19	12,27	12,40	12,53	12,69	12,80	12,94	13,70
|e|	9,33	8,55	8,67	8,92	9,07	9,25	9,39	9,52	9,64	9,76	10,52
$RMSE/s_{obs}$	0,429	0,380	0,374	0,376	0,379	0,382	0,387	0,391	0,395	0,399	0,423
R	0,907	0,925	0,928	0,927	0,926	0,925	0,923	0,921	0,920	0,919	0,909

Na rys. 1 przedstawiono wykres rozrzutu obserwowanych i przewidywanych wartości stężeń O_3 dla modelu otrzymanego dla k = 3. Z analizy wykresu wynika, że model ma podobne właściwości predykcyjne w całym zakresie obserwowanych stężeń.



Rys.1. Wykres rozrzutu obserwowanych i przewidywanych wartości stężeń O_3 (model: k = 3)

W tabeli 2 przedstawiono wyniki modelowania stężenia NO. Najlepszą dokładność modelowania uzyskano dla k = 3 (kryterium oceny: |e|, RMSE, RMSE/s_{obs}) i dla k = 5 (kryterium oceny: r). Dla k = 1 i dla k = 20 jakość predykcji była najgorsza. Na podstawie przedstawionych wyników można sformułować wniosek, że również w przypadku stężenia NO najlepsze wyniki modelowania uzyskuje się dla stosunkowo niedużej liczby najbliższych sąsiadów.

Tabela 2. Wyniki modelowania stężenia NO metodą k-najbliższych sąsiadów; predyktory: D, G, CO, O_3, NO_2, SO_2, PM_{10}, PW, T, R, W_{wzg}

Statystyki opisowe i miary błędu modelowania	Liczba najbliższych sąsiadów										
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=20
$\bar{x}_{obs}$	3,13	3,13	3,13	3,13	3,13	3,13	3,13	3,13	3,13	3,13	3,13
$\bar{x}_{pred}$	2,99	2,92	2,88	2,82	2,82	2,81	2,81	2,80	2,78	2,76	2,70
s_{obs}	6,96	6,96	6,96	6,96	6,96	6,96	6,96	6,96	6,96	6,96	6,96
s_{pred}	6,88	6,33	5,80	5,19	4,97	4,76	4,63	4,38	4,16	4,04	3,35
RMSE	4,22	3,45	3,18	3,30	3,35	3,47	3,68	3,88	4,04	4,06	4,53
|e|	1,22	1,08	1,02	1,03	1,05	1,07	1,10	1,12	1,14	1,15	1,22
RMSE/s_{obs}	0,606	0,496	0,457	0,474	0,482	0,499	0,530	0,558	0,580	0,583	0,652
R	0,814	0,869	0,891	0,893	0,894	0,891	0,873	0,861	0,853	0,858	0,837

Do porównania dokładności modelowania różnych zanieczyszczeń można wykorzystywać tylko te miary błędu, które nie zależą od zakresu pomiarowego stężeń danego zanieczyszczenia. Do takich miar można zaliczyć współczynnik korelacji i stosunek RMSE/s_{obs}. Uzyskane wyniki wskazują, że stężenia ozonu można przewidywać z większą dokładnością.

Stosując klasyfikację jakości modeli opisaną w pracy [5], modele imisji ozonu otrzymane metodą k-najbliższych sąsiadów można uznać za bardzo dokładne ($r > 0,90$), natomiast modele NO można zaliczyć do modeli dokładnych ($0,80 < r < 0,90$). NO należy do pierwotnych zanieczyszczeń powietrza. Modelowanie imisji takich zanieczyszczeń jest trudniejsze niż zanieczyszczeń wtórnych, takich jak ozon [6, 7]. Otrzymane modele charakteryzują się podobną dokładnością jak modele uzyskane dla tych samych zanieczyszczeń za pomocą innych metod modelowania.

3. Wnioski

Na podstawie przeprowadzonej analizy sformułowano następujące wnioski:

1. Modele imisyjne otrzymane metodą k-najbliższych sąsiadów charakteryzują się podobną dokładnością jak modele uzyskane za pomocą innych metod modelowania.
2. Najlepsze wyniki modelowania stężenia O_3 uzyskuje się dla kilku, czyli stosunkowo niedużej liczby najbliższych sąsiadów. Uwzględnianie w modelu większej liczby najbliższych sąsiadów nie poprawia, lecz pogarsza jakość predykcji.
3. Modele imisyjne otrzymane dla ozonu są dokładniejsze od modeli uzyskanych dla NO.

Praca naukowa finansowana ze środków na naukę w latach 2006-2008 jako projekt badawczy 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, 2006 (40), 38, 7316–7330.
2. Gentili S,, Magnaterra L, Passerini G,: An introduction to the statistical filling of environmental data time series, in Handling Missing Data: Applications to Environmental Analysis, Latini G,, Passerini G, (eds,), Wit Press, Southampton, Boston, 2006.
3. Larose D.T.: Odkrywanie wiedzy z danych, Wydawnictwo Naukowe PWN, Warszawa, 2006.
4. Hand D., Mannila H., Smyth P.: Eksploracja danych, Wydawnictwa Naukowo-Techniczne, Warszawa, 2005.
5. Hoffman S.: Treating missing data at air monitoring stations, in Environmental engineering, L. Pawłowski, M. Dudzińska, A. Pawłowski (eds.), Taylor & Francis Group, London 2007, 349-353.
6. Hoffman S.: Zastosowanie sieci neuronowych w modelowaniu regresyjnym stężeń zanieczyszczeń powietrza, Wyd. Politechniki Częstochowskiej, Częstochowa 2004.
7. Hoffman S.: Missing data completing in the air monitoring systems by forward and backward prognosis methods, Environmental Protection Engineering, 2006, 32 (4), 25-29.