

KLASYFIKACJA LUK Z BRAKUJĄCYMI DANymi WYSTĘPUJĄCYCH W ZBIORACH DANymi REJESTROWANYCH NA STACJACH MONITORINGU POWIETRZA

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

STRESZCZENIE

Celem badań było wytypowanie charakterystycznych struktur luk pomiarowych w zbiorach danych rejestrowanych na stacjach monitoringu powietrza. Klasyfikację przypadków z brakującymi danymi zaproponowano na podstawie przeglądu danych, pochodzących z różnych stacji pomiarowych automatycznego monitoringu powietrza. Analizę przeprowadzono wykorzystując wieloletnie zbiory danych chwilowych (średnich 1-godzinnych), zarejestrowane w systemach monitoringu powietrza działających w różnych miejscowościach w Polsce. Porównano ze sobą wyniki uzyskane dla różnych stacji.

1. Wprowadzenie

Dane zbierane w sposób ciągły na stacjach monitoringu powietrza nigdy nie są kompletne [1]. Brakujące dane mogą mieć różną postać [2]. Luki pomiarowe mogą obejmować pojedyncze dane, serie danych odnoszących się do jednego zanieczyszczenia, pojedyncze przypadki w serii czasowej i całe bloki danych (o różnej długości) itp. Rozpoznanie typowych struktur obszarów z brakującymi danymi umożliwi ich klasyfikację, a następnie rekomendację odpowiednich metod modelowania dla wyszczególnionych klas.

Celem badań było wytypowanie charakterystycznych struktur luk pomiarowych w zbiorach danych. Klasyfikację przypadków z brakującymi danymi zaproponowano na podstawie przeglądu danych, pochodzących z różnych stacji pomiarowych automatycznego monitoringu powietrza. Analizę przeprowadzono wykorzystując wieloletnie zbiory danych chwilowych (średnich 1-godzinnych), zarejestrowane w systemach monitoringu powietrza działających w różnych miejscowościach w Polsce. Porównano ze sobą wyniki uzyskane dla różnych stacji.

2. Metodyka badań

W analizie wykorzystano serie pomiarowe pochodzące z sieci monitoringu powietrza funkcjonującej w województwie łódzkim. Analizowano serie czasowe stężeń chwilowych podstawowych zanieczyszczeń powietrza (PM_{10} , SO_2 , CO, NO_2 , NO_x), zarejestrowanych w latach 2004-2006. Wstępnie, dla badanych serii czasowych wymienionych zanieczyszczeń obliczono podstawowe parametry statystyczne, takie jak: kompletność serii, wartość maksymalna, wartość minimalna, średnia arytmetyczna i odchylenie standardowe. W dalszym etapie analizy oszacowano liczbę luk pomiarowych o różnych długościach, oddzielnie dla każdego z zanieczyszczeń.

3. Wyniki i ich omówienie

Podstawowe parametry statystyczne analizowanych serii czasowych przedstawiono w tabeli 1. Kompletność wszystkich serii pomiarowych przekraczała 90%. Częstość występowania

nia luk w rocznych seriach pomiarowych wynosiła od 0,8% (NO_x, Łódź Centrum) do 9,2% (NO₂, Zgierz).

Tabela 1. Podstawowe parametry statystyczne analizowanych serii czasowych, zarejestrowanych na stacjach pomiarowych w Zgierzu, Łodzi-Widzewie i w Łodzi-Centrum.

a) Zgierz

Parametr statystyczny	Jedn.	PM ₁₀	SO ₂	CO	NO ₂	NO _x	O ₃
Kompletność	%	97,8	96,1	98,1	90,8	91,2	–
Wartość minimalna	µg/m ³	1,0	0,0	0	0,0	0,0	–
Wartość maksymalna	µg/m ³	416,0	264,1	7211	141,0	401,2	–
Średnia	µg/m ³	36,0	21,8	681	19,9	28,8	–
Odchylenie standardowe	µg/m ³	30,6	19,6	486	15,0	25,9	–

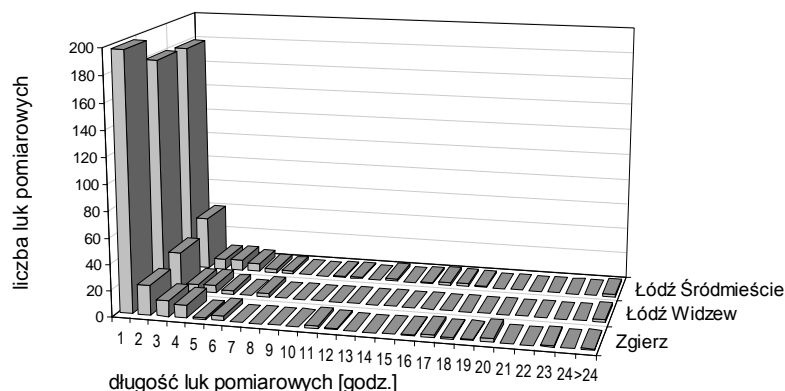
b) Łódź Widzew

Parametr statystyczny	Jedn.	PM ₁₀	SO ₂	CO	NO ₂	NO _x	O ₃
Kompletność	%	94,3	97,3	96,5	96,5	96,7	95,3
Wartość minimalna	µg/m ³	1,0	0,0	20	0,0	0,0	1,5
Wartość maksymalna	µg/m ³	264,0	471,2	5794	170,1	618,6	197,9
Średnia	µg/m ³	24,0	14,5	452	18,4	23,2	61,1
Odchylenie standardowe	µg/m ³	17,4	14,7	257	13,7	22,5	31,0

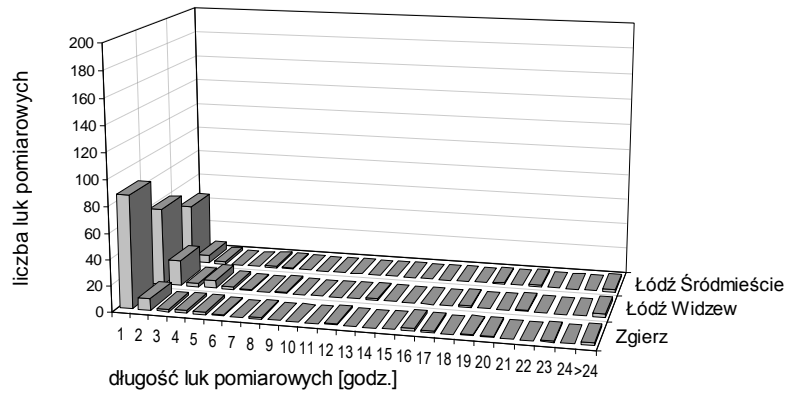
c) Łódź Centrum

Parametr statystyczny	Jedn.	PM ₁₀	SO ₂	CO	NO ₂	NO _x	O ₃
Kompletność	%	97,6	98,4	99,1	98,9	99,2	–
Wartość minimalna	µg/m ³	0,0	0,0	16	0,0	0,2	–
Wartość maksymalna	µg/m ³	382,0	208,2	6264	154,2	646,6	–
Średnia	µg/m ³	28,1	14,3	558	24,6	35,2	–
Odchylenie standardowe	µg/m ³	21,1	14,2	365	16,4	30,1	–

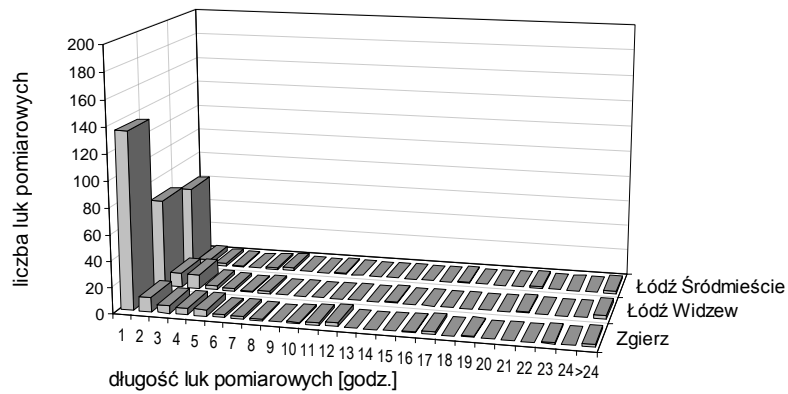
Wykresy ilustrujące liczebność luk pomiarowych w zależności od ich długości przedstawiono na rys. 1 do 6, oddzielnie dla każdego z zanieczyszczeń. Dla wszystkich zanieczyszczeń, na wszystkich stacjach najczęściej występującymi lukami pomiarowymi są pojedyncze przypadki. Częstość występowania luk bardzo szybko maleje z rosnącą ich długością (czasem trwania). Bloki brakujących danych o długościach przekraczających 5 przypadków (5 godzin) występują stosunkowo rzadko.



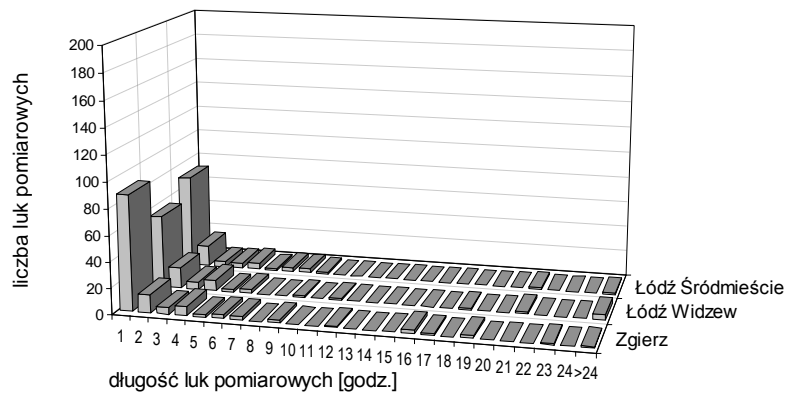
Rys. 1. Liczebność luk pomiarowych w zależności od ich długości dla stężeń PM₁₀



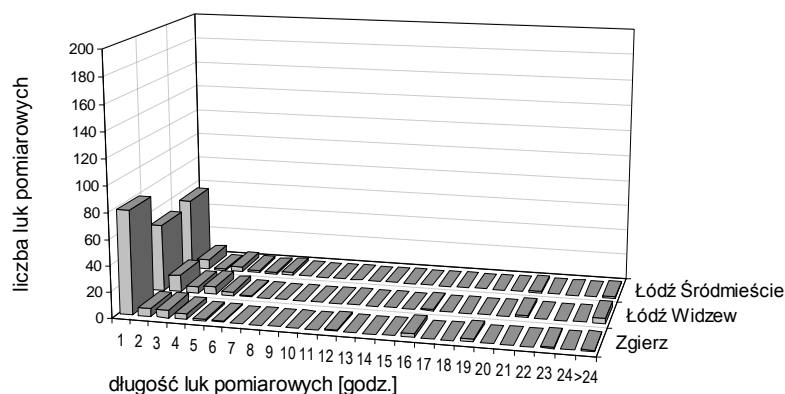
Rys. 2. Liczebność luk pomiarowych w zależności od ich długości dla stężeń SO₂



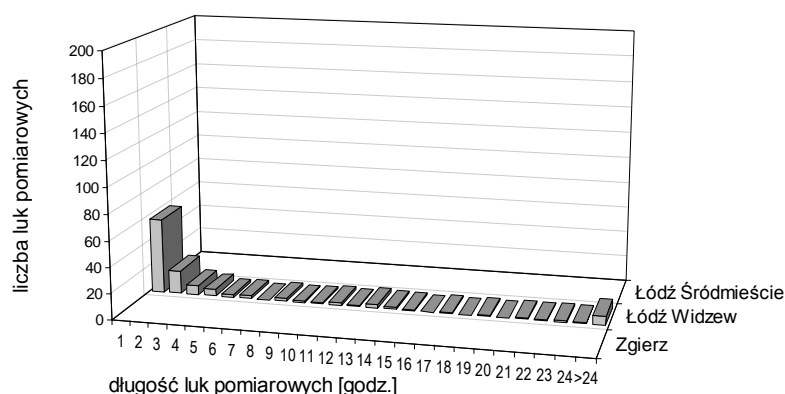
Rys. 3. Liczebność luk pomiarowych w zależności od ich długości dla stężeń CO



Rys. 4. Liczebność luk pomiarowych w zależności od ich długości dla stężeń NO₂



Rys. 5. Liczebność luk pomiarowych w zależności od ich długości dla stężeń NO_x



Rys. 6. Liczebność luk pomiarowych w zależności od ich długości dla stężeń O_3

4. Wnioski

Brakujące dane występują powszechnie w zbiorach danych pochodzących z monitoringu powietrza. Stwierdzono, że częstość ich występowania w rocznych seriach pomiarowych wynosi od 0,8 do 9,2%.

Większość luk pomiarowych jest krótka – stanowią je głównie pojedyncze przypadki. Zdecydowanie rzadziej występują bloki brakujących danych, przekraczające 5 przypadków (> 5 godzin).

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

Literatura

1. Plaia A., Bondi 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 Latini G., Passerini G., (Eds.) *Handling Missing Data: Applications to Environmental Analysis*, Wit Press, Southampton, Boston, 2006.