

Daniel Rabczenko, Bogdan Wojtyniak

METODY ANALIZY SZEREGÓW CZASOWYCH W OCENIE WPŁYWU ZANIECZYSZCZEŃ POWIETRZA ATMOSFERYCZNEGO NA UMIERALNOŚĆ LUDNOŚCI

Zakład Statystyki Medycznej Państwowego Zakładu Higieny
Kierownik: Paweł Goryński

Krótkookresowy wpływ zanieczyszczeń powietrza na umieralność jest przedmiotem badań od około 20 lat. Wzrostowi znaczenia tych badań towarzyszy znaczny rozwój metod statystycznych mających kluczowe znaczenie dla jakości i precyzji uzyskanych wyników. Praca prezentuje najważniejsze aspekty modelowania szeregów czasowych, które mogą być przydatne w badaniach innych problemów zdrowotnych tego typu.

Słowa kluczowe: zanieczyszczenia powietrza, umieralność, szeregi czasowe, uogólnione modele addytywne, meta-analiza

Key words: air pollution, mortality, time series, generalized additive models, meta-analysis

WSTĘP

Przez krótkookresowy efekt zanieczyszczeń powietrza na umieralność rozumiemy związek zmian dziennej liczby zgonów w populacji związany ze zmianami stężenia zanieczyszczenia powietrza atmosferycznego obserwowanymi w skali kilku do kilkudziesięciu dni. Impulsem do badania takiego efektu były tak zwane epizody smogowe, które w latach pomiędzy 1930 a 1952 dotknęły szereg miast w Stanach Zjednoczonych i w Europie (1,2,3). Spowodowały one kilkukrotny wzrost liczby zgonów w porównaniu z okresem, gdy stężenie zanieczyszczeń w powietrzu pozostawało na normalnym poziomie. Najsłynniejszym z nich był smog londyński, który w grudniu 1952 roku spowodował śmierć około 4000 osób (2).

Krótkookresowy efekt wzrostu dziennej liczby zgonów obserwowany w czasie epizodów smogowych związany był z wystąpieniem nienormalnie wysokich stężeń zanieczyszczeń powietrza. W obecnej chwili w krajach uprzemysłowionych nie rejestruje się już tak wysokich stężeń zanieczyszczenia powietrza atmosferycznego. Jednakże badania naukowe, epidemiologiczne i eksperymentalne dostarczają wiele dowodów na to, iż zanieczyszczenie powietrza nawet na rejestrowanym obecnie poziomie nie pozostaje obojętne dla zdrowia ludzi. Wśród badań epidemiologicznych, których przedmiotem jest ocena efektów zanieczyszczenia powietrza kluczowe miejsce zajmują badania posługujące się ana-

lizą szeregów czasowych, polegające na badaniu związku zmian w krótkich przedziałach czasowych zanieczyszczeń powietrza i zdarzeń zdrowotnych – głównie dziennej liczby zgonów oraz przyjęć do szpitali z powodu wybranych przyczyn. Spośród prowadzonych na całym świecie przy użyciu tej metodyki badań największe znaczenie miały dwa wielo-środkowe finansowane przez Komisję Europejską projekty APHEA (Air Pollution and Health Effects: An European Approach) i APHEA-2 oraz jeden amerykański NMAPS (The National Mortality, Morbidity and Air Pollution Study). Bardzo ważnym ich osiągnięciem było wypracowanie i adaptacja szeregu metod statystycznych odpowiednich do natury analizowanego problemu. Było to niezwykle istotne, gdyż we wcześniej prowadzonych badaniach stosowano różne metody statystyczne, co bardzo utrudniało porównywanie wyników. Nowe metody pozwoliły również na bardziej dokładne zbadanie różnych aspektów wpływu zanieczyszczeń powietrza na umieralność.

W pracy omówione zostały metody statystyczne stosowane obecnie na świecie w badaniu krótkookresowego wpływu zanieczyszczeń powietrza na zdrowie ludności i zastosowane przez nas w analizie wpływu trzech rodzajów zanieczyszczeń na umieralność w czterech polskich miastach. W dużej mierze podobne są one do metod stosowanych w projekcie APHEA-2, jednakże uwzględniają najnowszą wiedzę o problemach związanych z estymacją parametrów budowanych modeli statystycznych. Również sam przyjęty przez nas model umieralności różni się nieco od tego używanego w projekcie APHEA-2 większym naciskiem na pełniejsze uwzględnienie czynników atmosferycznych.

MATERIAŁ

Analizowaną zmienną zależną jest zazwyczaj dzienna liczba zdarzeń zdrowotnych mogących w jakimś stopniu zależeć od poziomu zanieczyszczeń. W naszych badaniach były to zgony z powodu ogółu przyczyn z wyłączeniem zewnętrznych oraz wybranych grup przyczyn.

Analizowanym czynnikiem ryzyka zgonu jest średniodobowy poziom zanieczyszczenia powietrza. Obliczany jest on jako średnia z dostępnych w danym dniu wyników z punktów pomiarowych w danym mieście. Do analiz włączyliśmy te punkty pomiarowe, których lokalizacja była niezmienna w czasie badania oraz w których liczba dni z brakami danych nie przekraczała 15% całego okresu badania. Dla dni, w których dostępny był przynajmniej jeden pomiar z któregośkolwiek z zaklasyfikowanych do badania punktów, brakujące dane z pozostałych punktów pomiarowych były uzupełniane jako średnia ważona z dostępnych pomiarów według następującego wzoru:

$$\hat{x}_{ijk} = \bar{x}_{i \bullet k} \cdot \frac{\bar{x}_{\bullet jk}}{\bar{x}_{\bullet \bullet k}} \quad [1]$$

gdzie:

- \hat{x}_{ijk} – estymowany poziom zanieczyszczeń w dniu i , punkcie j i roku k ,
- $\bar{x}_{i \bullet k}$ – średnia dobowa z dostępnych pomiarów w dniu i , roku k ,
- $\bar{x}_{\bullet jk}$ – średnia roczna punktu j w roku k ,
- $\bar{x}_{\bullet \bullet k}$ – średnia w roku k z wszystkich punktów.

Tak więc brakujący pomiar w danym punkcie j zastępuje się przez średnią z pomiarów dostępnych w danym dniu w innych punktach pomiarowych, pomnożoną przez stosunek średniego poziomu zanieczyszczenia w roku k w punkcie j do średniego poziomu zanieczyszczenia w danym mieście w roku k . Zauważmy, że metoda ta nie może być zastosowana do dni, w których nie dysponujemy żadnym pomiarem zanieczyszczenia powietrza. Poziom zanieczyszczeń w dniach, w których nie wykonano żadnego pomiaru estymowany był metodą interpolacji pomiarów z dni poprzedzających i następujących.

Użycie średniej z punktów pomiarowych jako wskaźnika narażenia dla całej populacji może wydawać się dyskusyjne, gdyż takie podejście zakłada jednakowe narażenie dla wszystkich ludzi mieszkających w danym mieście. Jednakże podejście to sankcjonują wyniki prac amerykańskich (4) i holenderskich (5) wykazujące istnienie dużej korelacji w czasie pomiędzy rutynowymi pomiarami a pomiarami pochodzącymi z próbników osobistych. Teoretyczne rozważania na temat błędu oszacowania związanego z błędem pomiaru zawarte w pracy Dominici i innych (6) również pozwalają na odrzucenie hipotezy, że obserwowany wzrost dziennej liczby zgonów, związany ze wzrostem stężenia zanieczyszczenia w powietrzu atmosferycznym, jest efektem różnic pomiędzy narażeniem jednostkowym a populacyjnym

Zasadniczą sprawą jest uwzględnienie w analizie czynników zakłócających związek analizowanej zmiennej zdrowotnej i zanieczyszczenia powietrza. W przypadku analiz umieralności są to: dobowe parametry warunków meteorologicznych dla każdego dnia w okresie badania (średnie dobowe temperatury, wilgotności i ciśnienia atmosferycznego a także zmian temperatury i ciśnienia pomiędzy kolejnymi dniami) a także okresy epidemii grypy.

METODY STATYSTYCZNE

W analizach szeregów czasowych mających na celu opisanie wpływu zanieczyszczeń powietrza na dzienną liczbę zgonów kluczowym zagadnieniem jest prawidłowy dobór metod statystycznych. Zmiany dziennej liczby zgonów zależą od bardzo wielu czynników o różnorodnym charakterze, a ponadto należy brać pod uwagę nieliniowość pewnych zależności. Uzyskiwane wyniki okazały się być wrażliwe na stosowane metody zarówno jeżeli chodzi o wybór modelu (7) jak i pewne problemy związane z obliczeniami matematycznymi – zbieżnością iteracyjnego procesu estymacji parametrów w uogólnionych modelach addytywnych (6,8) (zbyt mała domyślna liczba iteracji w programie SPLUS-2000) oraz uproszczeniami w oszacowaniach błędów liniowych składników modelu (9). Należy zaznaczyć, że dzięki ustaleniu ściślejszych kryteriów zbieżności oraz zastosowaniu programu likwidującego błędy oszacowania precyzji efektu (9) problemy te nie dotyczą obliczeń w naszej pracy.

B u d o w a m o d e l u . Podstawowym problemem występującym przy budowie odpowiedniego modelu statystycznego jest nieliniowość analizowanych zależności. Standardową obecnie techniką statystyczną stosowaną w analizie wyników są uogólnione modele addytywne (generalized additive models – GAM) (8) pozwalające na uwzględnienie w modelu zarówno liniowego jak i nieliniowego wpływu niektórych czynników bez zakładania a priori kształtu zależności.

Równanie uogólnionego modelu addytywnego ma następującą postać:

$$\ln(E(Y_t)) = S_0 + \underbrace{\sum_{i=1}^p S_i(X_{it}, \lambda_i)}_{(A)} + \underbrace{\sum_{i=p+1}^k \beta_i X_{it}}_{(B)} + \underbrace{\beta \cdot ZAN_t}_{(C)} \quad [2]$$

gdzie:

- $E(Y_t)$ – wartość oczekiwana zmiennej zależnej – dziennej liczby zgonów (lub innych analizowanych zdarzeń zdrowotnych) w dniu t
- X_{it} – wartość i – tego czynnika w dniu t
- S_i – funkcja opisująca zależność pomiędzy i – tym czynnikiem a dzienną liczbą zgonów
- λ_i – parametr wygładzający (określa kształt) funkcji S_i
- β_i – współczynnik obrazujący kierunek i siłę wpływu czynnika X_i związanego w sposób liniowy ze zmienną zależną (liczbą zgonów)
- ZAN_t – miara narażenia na zanieczyszczenie powietrza w dniu t

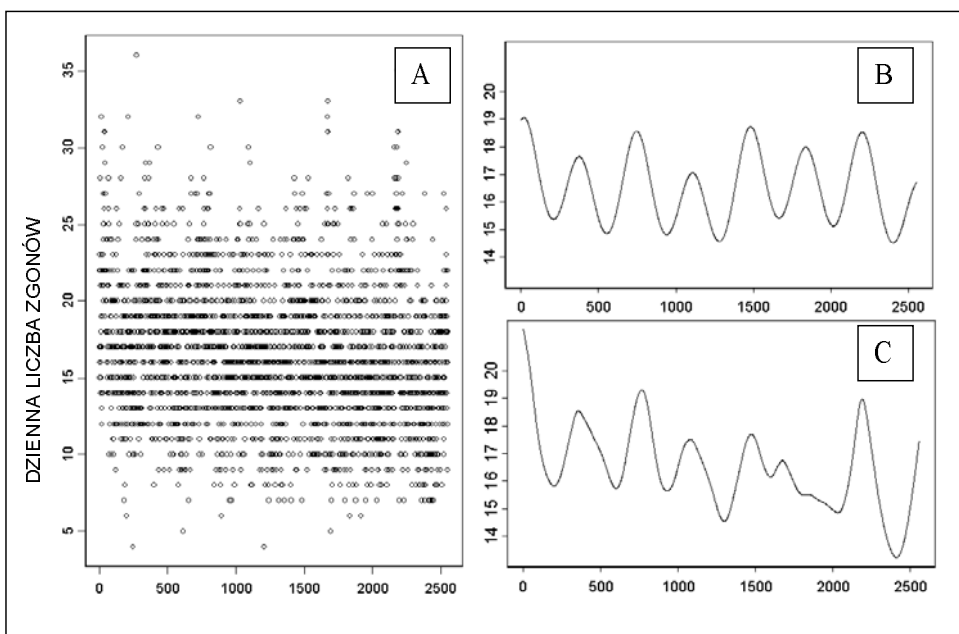
Liczba interesujących nas zdarzeń zdrowotnych Y_t w dniu t jest liczbą naturalną, liczba tych zdarzeń jest dużo mniejsza niż liczebność narażonej populacji, w związku z tym Y_t modelujemy przy użyciu rozkładu Poissona.

Równanie [2] ma 3 główne składniki. Składnik nieparametryczny (A) reprezentuje zmienne zakłócające, modelowanie których wymaga uwzględnienia nieliniowości i nieregularności zależności. Składnik (B) – parametryczny – uwzględnia zmienne modelowane w sposób klasyczny (liniowo lub w postaci zmiennych ślepych). Zanieczyszczenia powietrza stanowią trzeci składnik (C) równania modelu i w zależności od badanego problemu uwzględniane są w modelu w różny sposób – zarówno parametrycznie jak i nieparametrycznie. Warto w tym miejscu zaznaczyć, że do niedawna w modelowaniu statystycznym standardem było zastosowanie uogólnionych modeli liniowych, w których nieliniowe zależności próbowano modelować przy pomocy funkcji wielomianowych bądź też trygonometrycznych. Przeprowadzone analizy wykazały czułość wyników na rodzaj zastosowanej metody (7), a za bardziej wiarygodne uznaje się modele zbudowane w oparciu o metody nieparametryczne. Zastosowanie ich pozwala na znacznie bardziej właściwe modelowanie nieregularnych zależności (ryc. 1).

Model budowany jest w sposób krokowy, a przyjętą zasadą jest, że w pierwszej kolejności wprowadza się do niego zmienne określające długoterminowe i okresowe zmiany umieralności (analizowanego zdarzenia zdrowotnego), następnie zmienne mające działanie krótkookresowe a na samym końcu zanieczyszczenie powietrza. Wprowadzenie do modelu czynnika modelowanego w sposób nieparametryczny wiąże się z koniecznością dobrania tak zwanego parametru wygładzającego określającego kształt modelowanej zależności. Dobrany parametr musi być wynikiem „kompromisu” pomiędzy poszukiwaniem ogólnej, a więc jak najbardziej regularnej zależności z dopasowaniem krzywej do posiadanych danych. W odpowiednim dobraniu parametrów wygładzających, wspomagamy się dwoma kryteriami: funkcją autokorelacji (10) oraz kryterium informacyjnym Akaikiego (AIC) (11).

Dla każdego miasta i przyczyny zgonu konieczne było zbudowanie unikalnego bazowego modelu zmian umieralności w czasie. Uwzględniano w nim następujące czynniki:

- zmiany długookresowe: trend długookresowy i sezonowość;



Ryc. 1. Porównanie metod modelowania trendu długookresowego
 A – dzienna liczba zgonów w Poznaniu, 1990-1996
 B – trend modelowany przy użyciu funkcji trygonometrycznych
 C – trend modelowany przy użyciu funkcji nieparametrycznych (splajnów)

Fig. 1. Comparison of methods of long-term trend modeling

– warunki meteorologiczne (temperatura, ciśnienie, wilgotność, zmiany temperatury i ciśnienia pomiędzy kolejnymi dniami);

– epidemie grypy;

– dni tygodnia;

– święta państwowe i okresy wakacji letnich.

Po wprowadzeniu do modelu wszystkich zmiennych zakłócających przejść można do analizy zanieczyszczeń powietrza.

Zmianę umieralności związaną ze zmianą stężenia zanieczyszczenia powietrza o $x \mu\text{g}/\text{m}^3$ obliczamy dzieląc przez siebie oszacowaną z modelu dzienną liczbę zgonów w dniu, w którym zanieczyszczenia ma wartość $Z+x$ przez oszacowaną z modelu dzienną liczbę zgonów w dniu, w którym zanieczyszczenia ma wartość Z , natomiast wszystkie wartości zmiennych zakłócających pozostają bez zmian. Po uproszczeniu nie zmieniających się składników uzyskujemy wzór:

$$RR = \frac{e^{\beta \cdot (Z+x)}}{e^{\beta \cdot Z}} = e^{\beta \cdot x} \quad [3]$$

Obliczona wartość (RR) jest ryzykiem względnym zgonu związanym ze wzrostem stężenia zanieczyszczenia powietrza o $x \mu\text{g}/\text{m}^3$. Ryzyko względne może być przeliczone na procent zmiany dziennej liczby zgonów według wzoru.

M e t a - a n a l i z a . Metody meta-analizy pomagają przedstawić sumarycznie wyniki wielu niezależnych od siebie badań lub eksperymentów. Podstawową zasadą jest uśrednianie uzyskanych wyników przy jednoczesnym uwzględnieniu ich precyzji. Metody te znajdują zastosowanie w badaniach dotyczących krótkookresowego wpływu zanieczyszczeń powietrza na umieralność, w których analizuje się dane z większej ilości miast.

Najprostszym sposobem uogólnienia wyników jest zastosowanie tak zwanego modelu efektów stałych (*fixed effect model*) (12). Średni efekt β_{fixed} obliczany jest jako średnia ważona wyników z każdego miasta, przy czym wagi są odwrotnie proporcjonalne do wariancji oszacowania. Tak więc większą wagę uzyskują wyniki z niewielkim błędem oszacowania.

$$\beta_{fixed} = \frac{\sum_{i=1}^n \beta_i \cdot w_i}{\sum_{i=1}^n w_i} \quad [4]$$

przy czym

$$w_i = \frac{1}{\text{var } \beta_i}$$

oraz

n – liczba miast

$\beta_i, i=1...n$ – współczynnik opisujący wpływ zanieczyszczeń powietrza na umieralność w danym mieście

$se\beta_i, \text{var}\beta_i$ – błąd standardowy i wariancja współczynnika β_i

Wariancja uśrednionego oszacowania ma postać:

$$\text{var } \beta_{fixed} = \frac{1}{\sum_{i=1}^n w_i} \quad [5]$$

Model efektów stałych stosować można w przypadku jednorodności wyników. Jednorodność sprawdza się przy użyciu testu heterogeniczności (12). Jego statystyka testowa obliczana jest według wzoru:

$$Q = \sum_{i=1}^n (\beta_i - \beta_{fixed}) \cdot w_i \quad [6]$$

Ma ona rozkład χ^2 z $n-1$ stopniami swobody, gdzie n jest liczbą uśrednianych współczynników.

W przypadku, gdy test heterogeniczności odrzuca hipotezę o jednorodności wyników, to właściwym postępowaniem jest zastosowanie tak zwanego modelu efektów losowych (*random effect model*) (12). Oszacowania wielkości i wariancji tego efektu dokonuje się wg wzorów [4] i [5] z tym, że wagi użyte przy obliczaniu średniej ogólnej są skorygowane dodatkowym czynnikiem D dla uwzględnienia dużej zmienności wyników. Dany jest on wzorem:

$$D = \max \left(\frac{(Q - (S-1) \cdot \sum_{i=1}^n w_i)}{(\sum_{i=1}^n w_i)^2 - \sum_{i=1}^n (w_i)^2}, 0 \right) \quad [7]$$

S – liczba współczynników, na podstawie których przeprowadza się meta-analizę.

Średni efekt i jego wariancję oblicza się stosując odpowiednio wzory [4] i [5] zastępując w_i na w_i^* :

$$w_i^* = \frac{1}{D + \frac{1}{w_i}} \quad [6]$$

Oszacowanie efektu zanieczyszczeń uzyskane z modelu efektów losowych jest podobne do oszacowania uzyskanego z modelu efektów stałych, jednak można powiedzieć, że wyniki uzyskane tą metodą są bardziej konserwatywne, gdyż błąd standardowy oszacowania ma większą wartość.

PODSUMOWANIE

Problem krótkookresowego wpływu zanieczyszczeń powietrza na umieralność dotyczy całej populacji, zatem jego ocena ma dużą wagę. Jego modelowanie jest zadaniem złożonym, choć należy tu wspomnieć, że w porównaniu z problemem długookresowych zmian umieralności nie aż tak wiele czynników odgrywa tu rolę i wymaga uwzględnienia w modelowaniu. Do zmiennych, które niewątpliwie mają wpływ na poziom umieralności, a nie muszą być bezpośrednio uwzględnione należą rozpowszechnienie palenia, nawyki żywieniowe, styl życia itp. – ich zmiany są reprezentowane w modelu przez czynniki obrazujące długookresowe zmiany umieralności. Przeprowadzone w przeszłości analizy wykazały czułość uzyskanych wyników na metodę zastosowaną w modelowaniu (7). Metody analiz są cały czas rozwijane. Udział w europejskim projekcie APHEA pozwoliła nam na przedstawienie aktualnej metodologii badań tego ważnego z punktu widzenia zdrowia publicznego problemu. Zastosowanie przedstawionych metod do analizy danych w czterech polskich miastach: Krakowie, Łodzi, Poznaniu i Wrocławiu przedstawimy w kolejnej części niniejszej pracy.

D Rabczenko, B Wojtyniak

METHODS OF TIME SERIES ANALYSIS EFFECT OF AIR POLLUTION ON MORTALITY

SUMMARY

The paper describes main aspects of modeling of short-term effect of air pollution on mortality. Basic statistical tool - generalized additive models are presented as well as principles of inclusion of confounding variables in the model and measures of air pollution effect. In the second part of work methods of meta analysis allowing for combining results from several individual studies are presented.

PIŚMIENNICTWO

1. Firket M. Fog along the Meuse Valley; Transactions of the Faraday Society 1936; 32: 1192-1197.
2. Her Majesty's Public Health Service, Mortality and Morbidity during the London fog of December 1952; London: HMSO; 1954; Report No 95 on Public Health and Medical Subjects.
3. Shrenk H, Heiman H, Clayton G. „Air Pollution in Donora PA: Epidemiology of the Unusual Smog Episode of October 1948”, Preliminary Report, U.S. Public Health Service; Washington D.C.; Public Health Bulletin No. 306.

4. Mage D, Wilson W, Hasselblad V, Grant I. Assessment of Human Exposure to Ambient Particulate Matter; *J Air & Waste Manage Assoc*; 49: 1280-1291, Nov. 1999.
5. Janssen NAH, Hoek G, Narssema H, and Brunekreef B. Personal Exposure to Fine Particles in Children Correlates Closely with Ambient Fine Particles. *Archives of Environmental Health*, March/April 1999, Vol. 54, No 2, 95-101.
6. Dominici F. and McDermott A. and Zeger S.L. and Samet J.M On the Use of Generalized Additive Models in Time Series of Air Pollution and Health *American Journal of Epidemiology* vol 156, no 3,193-203.
7. Samoli E, Schwartz J, Wojtyniak B, Touloumi G, Spix C, Balducci F, Medina S, Rossi G, Sunyer J, Bacharova L, Anderson HR, Katsouyanni K. Investigating regional differences in short-term effects of air pollution on daily mortality in the APHEA project: a sensitivity analysis; *Env Health Perspect* 2001; 109: 349-353.
8. Hastie T, Tibshirani R. *Generalized Additive Models*, Chapman and Hall, London, 1990
9. Dominici F, Zeger SL, and Samet JM. A Measurement Error Model for Time-Series Studies of Air Pollution and Mortality. *Biostatistics*, 12, 157-175.
10. Box G, Jenkins G. *Analiza szeregów czasowych*; PWN; Warszawa 1983.
11. Akaike H. Informaton theory and an extension of the maximum likelihood principle; w: *Second International Symposium of Information Theory* (ed. Petrov BN, Czaki F.); Akademiai Kiado; Budapest; 1973.
12. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med*. 1999 8(3):321-59.

Otrzymano: 19.09.2005 r.

Adres autora:

Daniel Rabczenko
Państwowy Zakład Higieny, Zakład Statystyki Medycznej
ul. Chocimska 24, 00-791 Warszawa
tel. (22) 542 12 36
e-mail: daniel@medstat.waw.pl