# Modele statystyczne wykorzystywane przy prognozowaniu stężeń zarodników alergennych

## Statistical models used for predicting concentrations of allergenic spores

**dr Agnieszka Grinn-Gofroń**
Katedra Taksonomii Roślin i Fitogeografii Uniwersytetu Szczecińskiego



**Streszczenie:** Ważnym kierunkiem badań aerobiologicznych jest poszukiwanie zależności między charakterystykami sezonu zarodnikowego a zmiennymi pogodowymi. Dotychczas powstało zaledwie kilka modeli prognostycznych dla wybranych rodzajów zarodników wywołujących alergie. Większość z nich charakteryzuje się stosunkowo niską sprawdzalnością (około 30%) i oparta jest raczej na prostych statystykach opisowych. Modelowanie koncentracji zarodników grzybów w powietrzu jest relatywnie trudnym zagadnieniem. Ze względu na stopień skomplikowania obiektu badań (duża liczba analizowanych parametrów, bardzo nieregularne zmiany koncentracji zarodników przy znaczącej rozmaitości gatunków, nieliniowe zależności między parametrami) preferowane są wielowymiarowe techniki eksploracji danych oraz inne zaawansowane metody statystyczne.

**Abstract:** The important direction in aerobiological studies is to search for relationships between the characteristics of spore season and the weather variables. Since now only a few forecasting models for selected allergenic types of spores have been created. Most of them are characterized by relatively low verifiability (about 30%) and based on simple descriptive statistics. Modelling of the concentration of fungal spores in the air is relatively difficult. Due to the complexity of the test object (a large number of parameters analyzed, very irregular changes in the concentration of spores in a significant variety of species, non-linear relationship between the parameters) the techniques for multidimensional data mining and other advanced statistical methods are preferred.

**Słowa kluczowe:** modele statystyczne, aeroalergeny, zarodniki, grzyby, czynniki meteorologiczne
**Key words:** statistical models, aeroallergens, spores, mould, meteorological variables

Badania nad zmianami koncentracji zarodników grzybów wywołujących alergie prowadzone są na całym świecie. Konieczność wykonywania takich badań wynika z obserwowanego w ostatnich dziesięcioleciach lawinowego wzrostu liczby uczuleń na alergeny spor grzybów mikroskopowych. Zjawisko to jest szczególnie intensywne na obszarach uprzemysłowionych i w dużych miastach. Zarodniki grzybów w niektórych krajach (m.in. w Kanadzie) zostały uznane za biozanieczyszczenia i opracowywane są specjalne normy ich dopuszczalnej zawartości w powietrzu. W Polsce takie normy obowiązują rodzaje wywołujące alergie i określają wartości progowe stężeń tych taksonów [1]. Podstawowe rodzaje alergogenne są podobne na całym świecie i występują w prawie wszystkich typach klimatów. Znajomość stężenia zarodników grzybów o właściwościach alergizujących w powietrzu na określonym obszarze jest problemem uniwersalnym i ważnym zarówno dla osób uczulonych, jak i dla lekarzy alergologów w profilaktyce i leczeniu tego typu alergii. Występowanie objawów alergii ma charakter sezonowy. Wysokość stężeń zarodników i osią-


Alergoprofil 2012, Vol. 8, Nr 2, 10-12
Pracę otrzymano: 2012-04-26
Zaakceptowano do druku: 2012-05-14

gane przez nie wartości progowe (wywołujące objawy alergii) zależą w bardzo dużym stopniu od czynników pogodowych dominujących w badanych sezonach. Precyzyjne określenie wartości czynników meteorologicznych, przy których stężenia zarodników osiągają wartości progowe i utrzymują się na bardzo wysokim poziomie, jest niezwykle istotne w projektowaniu długoterminowych i krótkoterminowych prognoz dla osób wrażliwych. Dodatkowo dobry model prognostyczny powinien ustalić hierarchię ważności dla najbardziej i najmniej wpływowych czynników meteorologicznych oraz zanieczyszczeń powietrza dominujących w danym środowisku.

Rozwój prognozowania aerobiologicznego zmierza w kierunku uwzględnienia w modelach nie tylko warunków pogodowych, lecz także czynników topograficznych oraz nader ważnych parametrów zanieczyszczenia atmosfery. Ważnym kierunkiem badań aerobiologicznych jest poszukiwanie zależności między charakterystyką sezonu zarodnikowego a zmiennymi pogodowymi. Dotychczas powstało zaledwie kilka modeli prognostycznych dla wybranych rodzajów zarodników wywołujących alergie. Większość z nich charakteryzuje się stosunkowo niską sprawdzalnością (około 30%) i jest oparta raczej na prostych statystykach opisowych: m.in. regresji wielokrotnej, współczynnikach korelacji Pearsona, Spearmana czy teście rang Duncana [3, 6–9]. Nie podają one także dokładnych wartości tych czynników pogodowych, które są odpowiedzialne za wywoływanie stężeń progowych alergogennych zarodników grzybów, ani nie precyzują, jaki jest najistotniejszy.

Modelowanie koncentracji zarodników grzybów w powietrzu jest relatywnie trudnym zagadnieniem. Ze względu na stopień skomplikowania obiektu badań (duża liczba analizowanych parametrów, bardzo nieregularne zmiany koncentracji zarodników przy znaczącej rozmaitości gatunków, nieliniowe zależności między parametrami) preferowane są wielowymiarowe techniki eksploracji danych oraz inne zaawansowane metody statystyczne.

Ze względu na obecność wartości zerowych w seriach czasowych koncentracji zarodników grzybów w powietrzu oraz skokowe zmiany ich wartości bardzo przydatną metodą statystyczną są wielowymiarowe drzewa regresyjne (MRT, *Multivariate Regression Trees*). Pozwalają one na określenie wartości progowych parametrów środowiskowych, po których przekroczeniu znacząco zmienia się koncentracja zarodników grzybów w powietrzu. Generalnym założeniem MRT jest tworzenie skupień składu jakościowo-ilościowego zarodników poprzez powtarzający się podział danych wzdłuż osi zmiennych objaśniających (środowiskowych), przy czym każde rozgałęzienie dobierane jest tak, aby minimalizować różnice między skupieniami, wyrażone jako suma kwadratów odległości euklidesowych (SSD, *sum of squared Euclidean distances*) [2, 4, 5]. Skupienia i ich zależność od parametrów środowiskowych prezentowane są graficznie jako drzewo. Każde skupienie (gałąź) reprezentuje pewien charakterystyczny skład gatunkowy, dodatkowo określony przez zakres parametrów środowiskowych. Jakość drzewa MR&CT określa się za pomocą błędu względnego (RE, *relative error*) (suma kwadratów odległości euklidesowych SSD w gałęziach [skupieniach] podzielona przez SSD danych pierwotnych/niepodzielonych), a jego zdolności predykcyjne za pomocą RE kroswalidacyjnego (CVRE) [2, 4]. Gatunki charakterystyczne dla danego skupienia (określonych warunków środowiskowych) są identyfikowane za pomocą indeksu gatunków wskaźnikowych (*indicator species index, indval*), obliczanego jako iloczyn względnego zagęszczenia i względnej częstotliwości pojawiania się w danej gałęzi drzewa [10]. Gatunki, dla których indeks wskaźnikowy jest > 0,25, definiowane są jako taksony wskaźnikowe.

Bardzo duża liczba wzajemnie skorelowanych parametrów wpływających często w sposób nieliniowy na koncentrację zarodników grzybów w powietrzu wskazuje na możliwość zastosowania w analizie tego typu danych jednej z najnowocześniejszych, zaawansowanych metod: sztucznych sieci neuronowych (ANN, *Artificial Neural Networks*).

Sieć neuronowa jest zespołem powiązanych ze sobą neuronów, które zwykle tworzą trzy warstwy. Pierwsza nosi nazwę wejściowej, a ostatnia wyjściowej. Wszystkie warstwy znajdujące się między nimi są nazywane warstwami ukrytymi. Dane są wprowadzane na wejścia neuronów warstwy pierwszej, a następnie, przez istniejące połączenia, wartości wyjściowe warstwy poprzedniej są przekazywane na wejścia warstw następnych. Rezultaty otrzymane na wyjściu ostatniej warstwy są wynikiem obliczeń [11].

Wyznaczanie poprawnych wartości parametrów sieci o określonej architekturze (wag połączeń między sztucznymi neuronami) nazywa się uczeniem sieci neuronowej. Jest to niezbędny etap konstrukcji modelu neuronowego. Nauczanie sieci przeprowadza się przy zastosowaniu odpowiednich algorytmów na podstawie zebranych przez użytkownika danych, opisujących przebieg badanego zjawiska. Podczas obliczeń dostępny zbiór danych jest dzielony na trzy części: zbiór uczący Tr, podawany na sieć w procesie jej uczenia, zbiór walidacyjny We, pozwalający na monitorowanie

procesu uczenia sieci, oraz zbiór testujący Te, służący do przeprowadzania ostatecznej oceny uzyskiwanego modelu [11].

Sieci neuronowe umożliwiają budowę modeli nieliniowych, opisujących złożone procesy, bezpośrednio na podstawie danych doświadczalnych, a nie przyjmowanych hipotez wymaganych przy budowie modeli strukturalnych. Stanowią nowoczesne narzędzie informatyczne do rozwiązywania problemów rzeczywistych. Mogą być stosowane w przypadku występowania zależności między zmiennymi określającymi (wejścia) i zmiennymi określanymi (wyjścia). Są one szczególnie przydatne w poszukiwaniu zależności między zmiennymi, gdy relacje te są bardzo złożone, trudne do wyrażenia metodami statystyk klasycznych [11].

Skomplikowany charakter zjawisk wielowymiarowych jest związany z koniecznością zastosowania różnych technik do opisu różnych aspektów danego zjawiska. Ostatecznie utworzone, nowatorskie dobowe i roczne modele prognostyczne mogą składać się z szeregu podmodeli statystycznych, opartych na tych z omówionych powyżej metod analizy danych doświadczalnych, które zapewnią jak najlepszą jakość prognozy. Horyzont czasowy modeli jest dobierany tak, aby zapewnić optimum pomiędzy dokładnością predykcji a praktycznym aspektem ostrzegania mieszkańców o stopniu narażenia na alergogenny aeroplankton.

## Piśmiennictwo:

1. *Rapiejko P., Lipiec A., Wojdas A., Jurkiewicz D.: Threshold pollen concentration necessary to evoke allergic symptoms. Int. Rev. Allergol. Clin. Immunol. 2004, 10(3): 91-94.*
2. *Breiman L., Friedman J.H., Olshen R.A., Stone C.G.: Classification and regression trees. Wadsworth International Group, Belmont, California 1984.*
3. *Damialis A., Gioulekas D.: Airborne allergenic fungal spores and meteorological factors in Greece: Forecasting possibilities. Grana 2006, 45: 122-129.*
4. *De'ath G.: Multivariate regression trees: A new technique for modelling species-environment relationships. Ecology 2002, 83: 1105-17.*
5. *De'ath G., Fabricus K.E.: Classification and regression trees: A powerful and simple technique for ecological data analysis. Ecology 2002, 81: 3178-92.*
6. *Herrero B., Fombella-Blanco M.A., Fernández-González D., Valencia-Barrera R.M.: The role of meteorological factors in determining the annual variation of Alternaria and Cladosporium spores in the atmosphere of Palencia, 1990-1992. Int. J. Biometeorol. 1996, 39: 139-142.*
7. *Hjelmroos M.: Relationships between airborne fungal spore presence and weather variables. Cladosporium and Alternaria. Grana 1993, 32: 40-47.*
8. *Katial R.K., Zhang Y.M., Jones R.H., Dyer P.D.: Atmospheric mold spore counts in relation to meteorological parameters. Int. J. Biometeorol. 1997, 41: 17-22.*
9. *Lyon F.L., Kramer C.L., Eversmeyer M.G.: Vertical variation of airspora concentrations in the atmosphere. Grana 1984, 23: 123-126.*
10. *Dufrene M., Legendre P.: Species assemblages and indicator species: The need for a flexible assymetrical approach. Ecol. Mon. 1997, 67: 345-66.*
11. *Żurada J., Barski M., Jędruch W.: Sztuczne sieci neuronowe. PWN, Warszawa 1996.*

*Adres do korespondencji:*
**dr Agnieszka Grinn-Gofroń**
Katedra Taksonomii Roślin i Fitogeografii
Wydział Biologii Uniwersytetu Szczecińskiego
71-415 Szczecin, ul. Wąska 13
e-mail: agofr@univ.szczecin.pl