PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Stanisław Jabłonowski
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

KLASYFIKACJA POWIATÓW WEDŁUG POZIOMU POTENCJAŁU PRODUKCYJNEGO ROLNICTWA OPARTA NA METODACH ANALIZY SKUPIEŃ STOSOWANYCH W GENETYCE

1. Wstęp

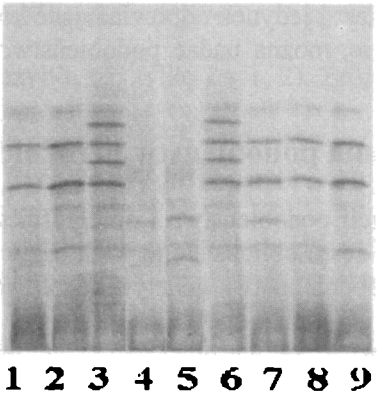
Analiza statystyczna danych genetycznych jest bardzo rozbudowaną dziedziną wiedzy. Istnieje bardzo wiele metod służących do badania podobieństwa genetycznego, analizy populacji, analizy skupień dla danych genetycznych itd. Wydaje się, że warto spróbować wykorzystać te metody i w innych dziedzinach, o ile to możliwe. Podstawowym pytaniem postawionym w artykule jest, czy istnieje taka możliwość w badaniach ekonomiczno-rolniczych. Odpowiedź nie jest oczywista ze względu na specyfikę danych genetycznych. Gdyby odpowiedź była pozytywna, otwierałoby to duże pole przed dalszym wykorzystaniem bogactwa metod badań statystycznych stosowanych w genetyce. Można byłoby stosować liczne programy komputerowe z tej dziedziny. W artykule pokazana jest próba takiego sposobu przedstawienia danych związanych z rolnictwem, by można było zastosować wspomniane gotowe metody i programy komputerowe. Sposób ten zilustrowano na danych o potencjale produkcyjnym rolnictwa w przekroju gmin województwa mazowieckiego.

2. Badanie podobieństw genetycznych

Danymi źródłowymi do badania podobieństwa genotypów organizmów żywych i ich populacji i do stosowania metod klasyfikacyjnych w genetyce mogą być elektroforegramy. Są to obrazy otrzymane w wyniku reakcji łańcuchowej polimerazy (PCR – *polymerase chain reaction*). Technika PCR umożliwia namnażanie fragmentów DNA. Na elektroforegramach przedstawione są ścieżki, na których znajdują się prążki sygnalizujące istnienie określonego (choć nie zawsze znanego) fragmentu łańcucha DNA. Ścieżki odpowiadają konkretnym próbkom materiału

genetycznego (np. igiełkom sosen), prążki zaś, w zależności od swego położenia na ścieżkach, sygnalizują istnienie takiego czy innego fragmentu DNA.

Reakcja polimerazy ma różne odmiany, zależnie od posiadanych urządzeń. Jedną z łatwiej dostępnych jest metoda RAPD, czyli metoda losowej amplifikacji. W RAPD na podstawie pewnej „próby" fragmentów DNA (starterów) otrzymuje się na ścieżkach zestawy prążków, o ile w materiale genetycznym są odpowiednie fragmenty DNA.



Rys. 1. Przykład elektroforegramu
(numery odpowiadają kolejnym ścieżkom, czyli próbkom organizmów żywych)
Źródło: [*Elektroforeza...* 2007].

Tabela 1. Zestawy zer i jedynek odpowiadające elektroforegramowi z rys. 1

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	1	1	0	0	1	1	1	1
0	0	1	0	0	1	0	0	0
1	1	1	0	0	1	1	1	1
0	0	1	0	0	1	0	0	0
1	1	1	0	1	1	1	1	1
1	1	1	0	0	1	1	1	1
1	0	1	1	1	1	1	0	0
1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Źródło: opracowanie własne.

Prążkom można przyporządkować liczby, które określają ich położenie (*locus*) na ścieżkach. Jednakowy fragment DNA może jednak dawać prążki o położeniu nie-

znacznie wahającym od ścieżki do ścieżki, co stwarza pewien problem przy porównywaniu otrzymanych liczb. Po pokonaniu tego problemu można ścieżkom przyporządkować ciągi zer i jedynek. Jedynka na określonej pozycji ciągu oznacza istnienie pewnego fragmentu DNA. W ten sposób można porównywać różne ścieżki.

W metodzie RAPD nie otrzymuje się tak wielu informacji, jak w niektórych innych metodach, nie znamy fragmentu DNA, któremu odpowiada prążek, ale podobnie położony prążek na różnych ścieżkach odpowiada temu samemu fragmentowi.

Otrzymane zestawy zer i jedynek odpowiadające kolejnym organizmom można przetwarzać statystycznie, można badać podobieństwo genetyczne, można budować skupienia.

3. Sposoby obliczania podobieństw i zróżnicowania genetycznego

Istnieje wiele definicji podobieństwa i odległości genetycznych między osobnikami i populacjami. Często oblicza się te wielkości np. według wzorów (1) i (2) (odległość Dice'a) dla podobieństwa genetycznego dwóch genotypów i odległości genetycznej między genotypami, odpowiednio:

$$GS_{ij} = 2N_{ij}/(N_i + N_j), \tag{1}$$

$$D_{ij} = 1 - GS_{ij}, \tag{2}$$

gdzie: N_{ij} – liczba wspólnych prążków dla obu genotypów,

N_i, N_j – liczba prążków w genotypach i oraz j [Nei, Li 1979; Weir 1990].

Podobieństwo genetyczne (I_N) między populacjami X i Y oblicza się np. według wzoru (3) dla określonego *locus*, w którym może wystąpić jeden albo dwa z n alleli[1]. Wzory (3)-(7) dotyczą jednego *locus*.

$$I_N = J_{XY}/(J_X J_Y)^{1/2}, \tag{3}$$

$$J_{XY} = \sum_i P_{iX} P_{iY}, \tag{4}$$

P_{iX}, P_{iY} – częstość i-tego allela ($i = 1, \ldots, n$) w populacji X i Y odpowiednio

$$J_X = \sum_i P_{iX}^2, \tag{5}$$

$$J_Y = \sum_i P_{iY}^2. \tag{6}$$

[1] RAPD może w zasadzie badać genotypy haploidalne (tylko jeden allel) dla określonego *locus* albo diploidalne (dwa allele), przy pewnych założeniach. W naszym przypadku allelem jest 0 albo 1 i mamy odpowiedniki genotypów haploidalnych, czyli w określonym *locus* jest albo 0, albo 1, a liczba n ma wartość 2.

Odległość genetyczna:

$$D_N = -\ln(I_N). \tag{7}$$

By obliczyć ogólną miarę podobieństwa (i odległości) uśrednia się wartości J_X, J_Y oraz J_{XY} po wszystkich *loci*, a dalej stosuje się wzory na I_N i D_N tak samo [Nei 1987].

Istnieje też wiele miar zróżnicowania genetycznego w populacji [Nei 1987; Weir 1996], np. h – zróżnicowanie genów Nei oblicza według wzoru: $h_L = 1 - \sum_i p_{Li}^2$ – dla poszczególnych *loci* (L = 1, ..., n), gdzie p_{Li} stanowi częstość i-tego allela w L-tym *loci* (czyli u nas tylko i = 1, 2). Zróżnicowanie genów h jest średnią z h_L po wszystkich *loci*.

4. Programy komputerowe

Z dużej liczby programów komputerowych do analizy statystycznej danych genetycznych w niniejszej pracy wykorzystano do obliczeń program POPGENE VERSION 1.32[2]. Inne programy to np.: GDA, RAPDistance, BIO-PROFIL Bio-Gene, PyPop.

Program POPGENE może m.in. rysować dendrogramy metodą UPGMA[3] na bazie odległości genetycznych, obliczać częstości alleli, obliczać wskaźniki homozygotyczności i heterozygotyczności i różne inne mierniki zróżnicowania genetycznego.

5. Zastosowanie w innych dziedzinach

Metody analizy licznych zbiorów danych stosowane w genetyce, mimo że specyficznie ukierunkowane, są w jakimś zakresie możliwe do przeniesienia do innych dziedzin i do rozwiązywania innych problemów. Wydaje się, że dane o pewnych obiektach wielocechowych związanych z rolnictwem, np. w gospodarstwach rolnych czy gminach wiejskich, można potraktować podobnie jak dane pochodzące z analizy RAPD i przetwarzać je podobnie jak tamte. Na przykład każde badane gospodarstwo z pewnej grupy można przedstawić jako taką ścieżkę, czyli „zestaw prążków". Pod pewnymi warunkami ostateczne zestawy liczbowe można przetwarzać statystycznie podobnie jak dane genetyczne.

6. Opis zastosowania

6.1. Dane

Dane dotyczą 225 gmin wiejskich województwa mazowieckiego (wybrane z 309 wszystkich gmin województwa; 5 gmin wiejskich pominięto ze względu na

[2] Microsoft Window-based Freeware for Population Genetic Analysis; autorzy: Francis C. Yeh and Rong-cai Yang, University of Alberta, Tim Boyle, Centre for International Forestry Research.

[3] Jedna z aglomeracyjnych metod hierarchicznych: metoda średnich połączeń.

niepełne dane), pochodzą ze spisu rolnego z 2002 r., dostępne są w GUS. Rozpatrywane cechy dotyczą rolnictwa. Dla każdej z gmin obliczono wartości następujących cech:

1. Liczba sztuk bydła na 100 ha powierzchni użytków rolnych.
2. Liczba sztuk trzody na 100 ha powierzchni użytków rolnych.
3. Liczba ciągników na jedno gospodarstwo.
4. Liczba samochodów ciężarowych na jedno gospodarstwo.
5. Liczba kombajnów zbożowych na 100 ha powierzchni gruntów ornych.
6. Liczba kombajnów ziemniaczanych na 100 ha powierzchni gruntów ornych.

Jako odpowiedniki organizmów żywych przyjęto poszczególne gminy; jako odpowiedniki populacji – podregiony (podział stosowany przez GUS), na które dzieli się woj. mazowieckie (bez Warszawy): ciechanowsko-płocki (A), ostrołęcko-siedlecki (B), warszawski (C), radomski (D).

W drugim podejściu badano powiaty województwa mazowieckiego, których jest 38, ale jeden, czysto miejski (powiat warszawski), pominięto. Powiatom przyporządkowane są numery od 1 do 30 i od 32 do 38 w kolejności alfabetycznej (nr 31 odpowiada powiatowi warszawskiemu).

Tabela 2. Statystyki opisowe cech opisujących gminy

Nr cechy	N	Minimum	Maksimum	Średnia	Odchylenie standardowe
1	225	0	112,8998	48,2175	20,4912
2	225	3,6581	127,7432	66,9826	23,5794
3	225	0,1708	1,4956	0,7289	0,2732
4	225	0,0056	0,5930	0,0915	0,0995
5	225	0	2,2601	0,9837	0,3855
6	225	0	6,3996	1,0331	0,9720

Źródło: [Jabłonowski, Kluza 2006].

6.2. Schemat przekształcenia danych

Dane opisujące wymienionych 6 cech dla każdej gminy przedstawiono jako ciągi zer i jedynek według następującego schematu: dla każdej cechy zakres możliwych wartości podzielono na 3 równe odcinki, a wartość cechy zastąpiono 3-elementowym ciągiem, w którym występują jedna albo dwie jedynki i zera (zero). Jedynka znajduje się na pozycji odpowiadającej odcinkowi, do którego trafia cecha, ale gdy wartość cechy jest blisko brzegu sąsiedniego odcinka (w granicach 10% długości odcinka), to jedynka jest wpisywana drugi raz na miejscu odpowiadającym sąsiedniemu odcinkowi. Pozostałe elementy tego 3-elementowego ciągu są równe zeru. Te ciągi składane są w ciąg 18-elementowy.

Tak przygotowane dane poddano obróbce programem POPGENE.

7. Wyniki

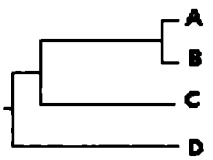
Niżej podane są wyniki obliczenia podobieństwa genetycznego (powyżej przekątnej) i odległości genetycznych (poniżej przekątnej) dla czterech populacji – podregionów woj. mazowieckiego.

Tabela 3. Podobieństwa i odległości (genetyczne) między podregionami woj. mazowieckiego

pop ID	A	B	C	D
A	****	0.9853	0.9016	0.8589
B	0.0148	****	0.8966	0.8947
C	0.1036	0.1091	****	0.8867
D	0.1521	0.1113	0.1203	****

Źródło: wyniki programu POPGENE V.1.32 („pop ID" oznacza w programie „nazwa populacji").

Niżej podano dendrogramy wygenerowane przez ten program (długości poszczególnych odcinków są proporcjonalne do liczb podanych osobno przez program, które pomijam):



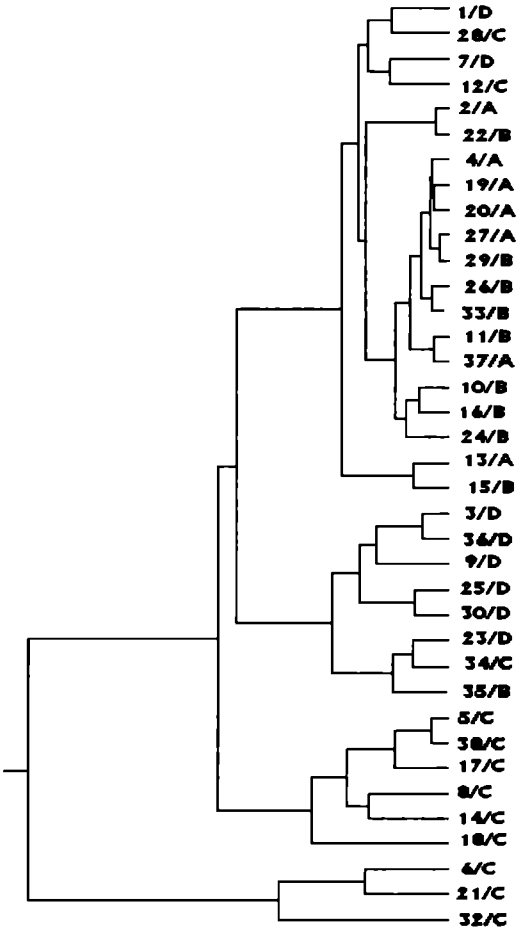
Rys. 2. Grupowanie podregionów (oznaczonych literami)
Źródło: opracowanie własne na podstawie wyników programu POPGENE (nazwy populacji zostały zmienione przez autora, gdyż POPGENE przyjmuje nazwy standardowe).

8. Ocena wyników

W pracy [Jabłonowski, Kluza 2006] zajęto się badaniem współzależności przynależności gmin i powiatów do podregionu i przedstawionymi wyżej cechami określającymi poziom potencjału produkcyjnego rolnictwa w regionie. Zbudowano tam wielomianowe modele logitowe, w których zmiennymi objaśniającymi było sześć cech podanych wyżej. Jakościowa zmienna objaśniana reprezentowała cechę przynależności gmin do podregionu. Te modele ekonometryczne miały dość dobre charakterystyki i uznano, że mogą służyć do prognozowania przynależności do podregionu[4]. Zatem uzasadniono tezę, że podział województwa mazowieckiego na podregiony jest czymś więcej niż czystym podziałem administracyjnym i do pewnego stopnia może stanowić kryterium oceny poprawności klasyfikacji. Wydaje się, że pogrupowania otrzymane w przyjętej metodzie są

[4] Przy zastosowaniu tych modeli do prognozy przynależności gmin do podregionów i wyborze takiej prognozy przynależności powiatu do podregionu, który wynika z najliczniej reprezentowanego wariantu przynależności gmin do podregionu, okazało się, że zliczeniowy R^2 wynosi dla jednego z wariantów modeli 97% (lub 92%, jeśli przy prognozowaniu właściwy wariant jest równoliczny z innym, niewłaściwym).

podobne do pogrupowań w podregiony przyjętych przez GUS (por. rys. 3). Uwaga ta może być argumentem za pozytywną oceną opisywanego w niniejszym artykule podejścia. Nie należy jednak tego traktować jako wystarczające uzasadnienie. Wymagałoby ono większej liczby zbiorów danych i porównań z podziałami i ich ocenami dokonywanymi w klasyczny sposób (por. [Kolenda 2006]).



Rys. 3. Grupowanie powiatów (oznaczonych numerami wraz z oznaczeniem podregionu)
Źródło: opracowanie własne na bazie wyników programu POPGENE (nazwy populacji zostały zmienione przez autora, gdyż POPGENE przyjmuje nazwy standardowe).

Warto też zwrócić uwagę, że przedstawione w punkcie 6.2 podejście pozwala na uwzględnienie różnych wariantów zamiany wartości cech na układy zer i jedynek, zależnie od charakterystyki danej cechy[5].

[5] W dalszych badaniach warto przeanalizować te różne warianty poprzez porównanie jakości podziałów populacji wynikających z przyjęcia różnych wariantów.

9. Podsumowanie i wnioski

Najważniejszym punktem w artykule było sprawdzenie, czy w ogóle jest możliwe zastosowanie specyficznych metod analizy statystycznej pochodzących z genetyki w badaniach ekonomiczno-rolniczych. Pokazano pewną propozycję odpowiedniego przekształcenia danych dotyczących rolnictwa, by można było te metody wykorzystać. Propozycję zilustrowano danymi określającymi poziom potencjału produkcyjnego w przekroju gmin województwa mazowieckiego. Pierwsze wyniki dają pewną nadzieję, że metody pochodzące z genetyki mogą być tu przydatne. Konieczne są dalsze badania i porównania wyników pogrupowań metodami klasycznymi i proponowanymi w pracy na szerszym materiale liczbowym.

Literatura

Elektroforeza przykłady zastosowań (2007), red. B. Walkowiak, V. Kochmańska, http://infekcja.net/katalog/Chemia/elektroforeza.pdf (stan z września 2007).

Jabłonowski S., Kluza A. (2006), *Taksonometryczna analiza poziomu rozwoju rolniczego gmin województwa mazowieckiego w oparciu o modele logitowe i ocena tej metody*, Warszawa, Materiały konferencji naukowej „Metody ilościowe w badaniach ekonomicznych", SGGW, Warszawa.

Kolenda M. (2006), *Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych*, AE, Wrocław.

Nei M. (1987), *Molecular Evolutionary Genetics*, Columbia University Press, New York.

Nei M., Li W.H. (1979), *Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases*, Proc. Natl. Acad. Sci. USA, 76, s. 5269-5273.

Weir B.S. (1990), *Genetic Data Analysis*, Sinauer Associates, Sunderland, MA.

Weir B.S. (1996), *Genetic Data Analysis II*, Sinauer Associates, Sunderland, MA.

POVIATS CLASSIFICATION BY THE PRODUCTIVE CAPABILITIES OF AGRICULTURE BASED ON CLUSTERING ANALYSIS METHODS USED IN GENETICS

Summary

Multivariate objects linked with agriculture, like farms, rural communities etc. are classified by assorted means. The purpose of the hereto study is to check if one could transfer some taxonomic methods used in genetics into other areas, especially to agroeconomical research. Source data on which the methods are applied are electroforegrams received within PCR reaction (Polymerase Chain Reaction). In effect geneticians can estimate genetic similarities between live organisms and their populations. Agroeconomical data can be presented likewise data on an electroforegram.