doi: 10.26114/sir.iung.2006.03.04

R A P O R T Y P I B

ZESZYT 3 **2006**

Krystyna Filipiak

Instytut Uprawy Nawożenia i Gleboznawstwa - Państwowy Instytut Badawczy w Puławach

METODY STATYSTYCZNE STOSOWANE DO OCENY REGIONALNEGO ZRÓŻNICOWANIA ROLNICTWA*

Wstęp

Problem regionalnych uwarunkowań rozwoju rolnictwa występuje w tematyce badawczej wielu placówek naukowych w Polsce. Dotychczasowe wyniki badań z tego zakresu wskazują na występowanie znaczących różnic między regionami w kraju, powodowanych odmiennymi warunkami przyrodniczymi, społeczno-ekonomicznymi i organizacyjnymi gospodarstw rolnych (2). Cechy charakteryzujące jakość rolniczej przestrzeni produkcyjnej, poziom kultury rolnej, agrotechniki, nakładów na produkcję rolną itp. wykazują przestrzenne zróżnicowanie. Ocena tego zróżnicowania nie jest sprawą prostą. Mamy tu bowiem do czynienia z wieloma czynnikami opisującymi rolniczą przestrzeń produkcyjną i jej wykorzystanie; występują problemy z pozyskaniem danych źródłowych o niewysokim stopniu agregacji, również wiarygodność niektórych informacji pozyskanych od respondentów może budzić pewne zastrzeżenia (3). Kolejne utrudnienie występuje na etapie opracowania zebranych danych. Duża liczba zmiennych (ilościowych i jakościowych) i obserwowanych jednostek wpływa na niejednorodność materiału liczbowego, co w następstwie stwarza problemy przy wyborze właściwych metod klasyfikacji i grupowania danych, które zależą również od celu badań.

Celem opracowania jest przegląd metod statystycznych zalecanych w analizach dotyczących zróżnicowania regionalnego rolnictwa.

Materiały źródłowe a cel badań

Dane źródłowe najczęściej pozyskuje się z rejestracji i sprawozdawczości organów administracji państwowej, w tym spisów powszechnych prowadzonych przez GUS, ankiet oraz badań naukowych, tj. obserwacji i doświadczeń polowych. Dane te przedstawiają rozmieszczenie badanej zbiorowości w przestrzeni (województwo, powiat, gmina, wieś, gospodarstwo, pole, współrzędne geograficzne), tworząc szereg

* Opracowanie wykonano w ramach zadania nr 2.1 w wieloletnim programie IUNG-PIB

terytorialny. W szeregach terytorialnych wartość badanej cechy lub liczebność jednostek posiadających określoną właściwość określa się w odniesieniu do przestrzeni, np.: plony 4 zbóż z 1 ha zasiewów, obsada zwierząt wyrażona w SD na 100 ha UR, udział GO w % powierzchni UR itp.

Dokładność wyników, ze statystycznego punktu widzenia, zależy od pochodzenia danych źródłowych. Wyniki statystyki państwowej pozyskane są najczęściej z badań pełnych, obejmujących całą populację: np. w ramach Powszechnego Spisu Rolnego wyniki zbierane są ze wszystkich gospodarstw rolnych w kraju, a następnie dane te są odpowiednio agregowane do poziomu kolejnych jednostek administracyjnych Polski (gmina, powiat, województwo). Celem badań bazujących na danych statystyki masowej może być: analiza poziomu i relacje między czynnikami produkcji rolnej, bonitacja jednostek administracyjnych pod kątem ich przydatności do produkcji rolnej, waloryzacja regionów i ich delimitacja (1). Badania pełne są jednak bardzo pracochłonne i drogie, dlatego do opracowań z pogranicza ekonomii, socjologii i rolnictwa wyniki pozyskuje się na podstawie ankiet, a informacje będące ich wynikiem stanowią cenne źródło wiedzy wspomagającej proces decyzyjny.

Badanie ankietowe jest w istocie badaniem częściowym. Z analizowanej zbiorowości jednostek, np. gospodarstw rolnych zlokalizowanych na określonym obszarze wybiera się w sposób reprezentatywny próbę, której wszystkie jednostki należy objąć badaniami ankietowymi. Operat losowania próby i jej liczebność zależą od właściwości statystycznych populacji (liczba cech, ich jednorodność, skorelowanie zmiennych) oraz założonej dokładności badań. Wyniki cech z próby wykorzystuje się do oszacowania parametrów tych cech (średnie, wariancje, korelacje, współczynniki funkcji opisujących zależności między zmiennymi) w całej populacji. Przykładowe cele badań ankietowych to: analiza i ocena technologii stosowanych w produkcji rolniczej wybranego terenu, efekty ekonomiczne lub warunki socjo-ekonomiczne gospodarstw położonych w określonym regionie, ocena wykorzystania rolniczej przestrzeni produkcyjnej wybranych regionów.

Wyniki eksperymentalne odnoszą się do konkretnych warunków siedliskowych pola i nie można ich uogólniać, jeżeli nie prowadzi się badań w różnych warunkach klimatyczno-glebowych. Ponieważ obecnie nie zakłada się wieloletnich i wielopunktowych serii doświadczeń polowych w skali całego kraju, więc na podstawie analizy wyników doświadczeń można jedynie wnioskować, że w warunkach produkcyjnych występują podobne zależności między zmiennymi (korelacje, kształt funkcji), natomiast wielkość lub zmienność badanych cech oraz wartości współczynników funkcji opisujących te zależności są najczęściej inne. Różnice te wynikają głównie z faktu, że doświadczenia prowadzone są w warunkach kontrolowanych. Podobne uwagi dotyczą doświadczeń zootechnicznych, w których odpowiednio dobiera się zwierzęta do badań tak, by zmienność osobnicza była jak najmniejsza. Doświadczenia polowe tylko w specyficznych przypadkach są wykorzystywane w tego typu badaniach, dotyczy to np. rejonizacji upraw.

Obserwacje naukowe rejestrują stan interesujących cech w punktach badań. Najczęściej wykorzystywane są do poszukiwania zależności przyczynowo-skutkowych

między badanymi cechami. Wyniki obserwacji punktowych przekształca się na dane powierzchniowe, opisujące pewne jednostki regionalne za pomocą różnych metod interpolacyjnych. W ten sposób opracowuje się np. mapy warunków pogodowych (sumy opadów, średnie temperatury, nasłonecznienie), mapy glebowo-rolnicze itp. Jeżeli jednak punkty rejestracji są znacznie od siebie oddalone, a wartości cechy nie zmieniają się liniowo między punktami siatki pomiarowej, wówczas metody interpolacji przestrzennej nie odzwierciedlają prawdziwych wartości tej cechy.

Metody opracowywania wyników

Statystyczne uporządkowanie materiału liczbowego polega na takim podziale zbiorowości niejednorodnej na możliwie homogeniczne grupy, by umożliwić ich charakterystykę, poznać strukturę tej zbiorowości oraz ustalić zależności między badanymi cechami. Wyróżnia się trzy klasy metod grupowania obiektów:

a) grupowanie strukturalne, polegające na uporządkowaniu badanej zbiorowości według wielkości, czyli wartości określonej cechy lub cech;
b) metody estymacji podobieństwa, czyli podział na klasy typologiczne poprzez wyodrębnianie z badanej zbiorowości możliwie jednorodnych grup przy uwzględnieniu kilku cech typowych do określenia badanego zjawiska;
c) grupowanie analityczne, uwzględniające korelacje między zmiennymi w celu określenia liczbowego związku między wieloma cechami.

Grupowanie strukturalne

Przystępując do wydzielenia grup – regionów na podstawie cech różnicujących daną zbiorowość należy wyznaczyć podstawowe charakterystyki statystyczne zmiennych. Najczęściej dla cech ilościowych oblicza się wartość średnią, medianę, wariancję lub odchylenie standardowe, rozstęp między minimalną i maksymalną wartością, kwartyle oraz współczynnik zmienności wyrażony w % średniej. Zmienną opisuje się średnią arytmetyczną i odchyleniem standardowym lub współczynnikiem zmienności, gdy jej rozkład jest zgodny z rozkładem normalnym, a w przypadku rozkładów skośnych za pomocą mediany i kwartyli.

Kolejnym krokiem jest konstrukcja szeregu strukturalnego, polegająca na zgrupowaniu jednostek według przyjętej przez badacza cechy różnicującej – ilościowej lub jakościowej. Klasy szeregu wyznacza się na podstawie dotychczasowej wiedzy, rozkładu zmiennej lub jej wartości bądź bonitacji. Na przykład, można skonstruować szereg strukturalny dla plonów roślin, zawartości próchnicy lub mikroskładników w glebie, poziomu nawożenia itp. w zależności od klas jakości gleby. Gleba może być opisana jako: słaba, średnia i dobra, może być oceniona za pomocą kompleksu rolniczej przydatności gleb lub klasy bonitacyjnej, a w przypadku, gdy jest wyrażona wskaźnikiem waloryzacji rolniczej przestrzeni produkcyjnej, wówczas wartości wskaźnika należy pogrupować w klasy (od – do). Do oceny istotności różnic między wartościami cech sklasyfikowanych w szereg strukturalny wykorzystuje się jednoczynnikową analizę

wariancji (zmienne o rozkładzie normalnym) lub analizę rang (rozkład zmiennej różny od rozkładu normalnego), przyjmując za czynnik wybraną cechę klasyfikującą (w omawianym przykładzie jest to jakość gleby). Istotne różnice wartości cech między klasami wskazują, że wybrana cecha klasyfikująca dobrze różnicuje analizowaną zbiorowość, więc pogrupowanie jest poprawne. W przypadku dwóch lub większej ilości zmiennych klasyfikujących wykorzystuje się wieloczynnikową analizę wariancji lub rang albo konstruuje się wielodzielcze tablice i za pomocą testu χ^2 ocenia się zależność (współdziałanie) między zmiennymi klasyfikującymi.

Metody estymacji podobieństwa

Podział zbiorowości na klasy typologiczne zależy od liczby zmiennych; w przypadku jednej lub dwóch zmiennych wykorzystuje się podstawowe charakterystyki statystyczne do podziału uporządkowanego szeregu liczbowego lub do podziału płaszczyzny. Klasyfikacji dokonuje się stosując kryteria: średniej, mediany, średniej i odchylenia standardowego, mediany i kwartyli. Na podstawie kryterium średniej i mediany uzyskuje się podział jednostek na dwie klasy: „mniejsze od…" i „większe lub równe...", albo „mniejsze lub równe..." i „większe od..." dla liniowo uporządkowanego szeregu i na cztery klasy przy podziale płaszczyzny (5). Z podziału bazującego na medianie i kwartylach mamy dla uporządkowanego szeregu liczbowego cztery klasy, a na podstawie średniej i odchylenia standardowego najczęściej wyodrębnia się sześć klas, a przypadku podziału płaszczyzny odpowiednio szesnaście i trzydzieści sześć klas.

W przypadku grupowania opartego na wielu cechach wykorzystuje się metody redukcji zmiennych, takie jak: analiza składowych głównych lub analiza czynnikowa, a następnie do nowych zmiennych (składowych lub czynników) stosuje przedstawione powyżej miary statystyczne.

Metoda składowych głównych jest uważana za szczególnie przydatną ze względu na możliwość przekształcenia uzyskanych wyników do przestrzeni trój-, dwu-, a nawet jednowymiarowej, niewielką stratę informacji przy redukcji wymiarów przestrzeni, graficzną prezentację obserwacji wielowymiarowych na płaszczyźnie wyznaczonej przez składowe oraz brak korelacji między składowymi. O ilości składowych decydują wniesiony przez nie ładunek informacji oraz przyjęta w badaniach metoda podziału jednostek na klasy typologiczne. W praktyce najczęściej uwzględnia się dwie pierwsze składowe główne, które umożliwiają ocenę podobieństwa obiektów poprzez rozkład punktów na płaszczyźnie.

Analiza czynnikowa, podobnie jak analiza składowych głównych, jest metodą analizy struktury współzależności cech pod kątem istnienia grup skorelowanych zmiennych (1). Zakłada się, że gdy cechy wewnątrz grup są silnie skorelowane, wówczas o ich wartościach decyduje jakiś wspólny czynnik wyjaśniający tę współzależność. Z praktycznego punktu widzenia analiza czynnikowa, tak jak metoda składowych głównych, prowadzi do redukcji zmiennych ze stosunkowo małą stratą informacji, a dodat-

kowo umożliwia interpretację czynników poprzez zespół cech najsilniej z nimi skorelowanych.

Do podziału obiektów wielozmiennych na grupy jednorodne można zastosować biplot rozpięty w przestrzeni dwóch lub trzech składowych głównych lub czynników, szczególnie w przypadku, gdy klasyfikuje się niezbyt dużą liczbę jednostek.

Inną metodą pozwalającą na podobną identyfikację ukrytych cech jest metoda skalowania wielowymiarowego, zwana w skrócie MDS. Algorytm metody MDS wykorzystując zasadę redukcji wielowymiarowej przestrzeni niemetrycznej (skala porządkowa) na przestrzeń metryczną (skala przedziałowa) o mniejszej liczbie wymiarów pozwala na wyznaczenie map percepcji z odpowiednią konfiguracją przestrzenną struktur obiektów. Jedną z technik skalowania wielowymiarowego (MDS) stosowaną do danych jakościowych jest analiza korespondencji. Analiza korespondencji grupuje kategorie na podstawie tablic kontyngencji, a wyniki są interpretowane na podstawie miar charakteryzujących związek pomiędzy kolumnami i wierszami. Kategorie, które są sobie bliższe są zarazem bardziej podobne pod względem strukturalnym. Jest to technika opisowa i eksploracyjna dostarczająca informacji podobnych w swej naturze do rezultatów analizy czynnikowej.

Analizę skupień wykorzystuje się najczęściej do estymacji podobieństwa jednostek scharakteryzowanych przez wiele cech i ich grupowania na podstawie odległości (1). Analiza skupień obejmuje różne metody podziału zbiorowości na podzbiory obiektów o wzajemnie podobnych elementach w przestrzeni wielowymiarowej. Podstawą podziału jest macierz odległości, przy czym definicja odległości jest dowolna i zależy od celu badań i charakteru grupowanych jednostek. Może to być odległość Euklidesa lub jej kwadrat, odległość Mahalanobisa, współczynnik determinacji albo inne miary traktowane jako odległość w przestrzeni rozpiętej na analizowanych cechach. Z metod analizy skupień najczęściej wykorzystywane są hierarchiczne metody aglomeracyjne i nie hierarchiczna metoda k-średnich. Ponieważ w każdej z metod hierarchicznych wykorzystuje się inne kryteria wyznaczania odległości między skupieniami, metody te mogą prowadzić do różnych wyników. Najczęściej stosowanymi metodami analizy skupień są:

- metoda najbliższego sąsiedztwa, zwana też taksonomią wrocławską – uzyskuje się jedno duże skupienie obiektów oraz mało liczne podzbiory lub pojedyncze jednostki o wartościach cech znacznie odstających od wartości przeciętnych dla zbiorowości i dlatego wykorzystuje się ją przy poszukiwaniu jednostek odstających;
- metoda najdalszego sąsiedztwa – konstruuje skupienia na podstawie maksymalnej odległości między obiektami, daje dobry podział na skupienia w zagadnieniach dotyczących regionalizacji;
- metoda Warda – prowadzi do skupień o zbliżonych liczebnościach, charakteryzujących się minimalną wariancją i z tego względu również często jest wykorzystywana do klasyfikacji jednostek przestrzennych;
- metoda *k*-średnich – łączy obiekty w skupienia wokół zadanych z góry jednostek,

w możliwie największym stopniu różniące się od siebie. Metodę stosuje się do dużej liczby obiektów, a za kryterium podziału na grupy przyjmuje się minimalizację sumy kwadratów odległości wewnątrzgrupowych.

Analiza skupień nie jest testem statystycznym, ale zbiorem różnych algorytmów, które prowadzą do grupowania obiektów. W odróżnieniu od wielu innych procedur statystycznych, metody analizy skupień są stosowane przeważnie wtedy, gdy nie dysponujemy żadnymi hipotezami *a priori*, natomiast jesteśmy nadal w fazie eksploracyjnej badań. Dlatego testowanie istotności statystycznej w tradycyjnym rozumieniu tego pojęcia faktycznie nie znajduje tutaj zastosowania. Należy też podkreślić, że optymalna liczba skupień nie jest znana z góry i powinna być wyliczana na podstawie danych. Zazwyczaj badamy średnie wszystkich cech dla każdego skupienia, aby oszacować na ile nasze skupienia są od siebie różne. Na podstawie testu F z analizy wariancji wykonanej dla każdej cechy określa się na ile dobrze dany wymiar dyskryminuje skupienia.

Inną metodą sprawdzania poprawności klasyfikacji obiektów jest analiza dyskryminacji, stosowana również do rozstrzygania, które zmienne pozwalają w najlepszy sposób dzielić dany zbiór jednostek na występujące w naturalny sposób grupy.

Grupowanie analityczne

Ostatnią klasą metod statystycznych wykorzystywanych do oceny regionalnego zróżnicowania rolnictwa są procedury grupowania analitycznego. Należy tu zaliczyć analizę korelacji, analizę regresji i drzewa klasyfikacyjne oraz analizę kanoniczną. Badając związek korelacyjny między zmiennymi można oceniać wielkość jednej zmiennej na podstawie wielkości drugiej zmiennej lub zbioru zmiennych, wykorzystując w tym celu modele regresji.

Równanie regresji wyznaczone na podstawie jednostek uporządkowanych przestrzennie ze względu na dwie zmienne niezależne, które są ortogonalnymi współrzędnymi geograficznymi lub jedną zmienną niezależną określającą odległość między obiektami nosi nazwę trendu powierzchniowego. Zakłada się, że funkcja trendu jest wielomianem stosunkowo niskiego stopnia. W analizie trendu wyróżnia się: zmiany zachodzące w dużej skali, zwane zmianami regionalnymi, zmiany w małej skali, czyli fluktuacje lokalne oraz zmiany losowe.

Na podstawie oceny dopasowania modelu regresji, czyli analizy punktów odstających i analizy reszt można również dokonać grupowania jednostek.

Drzewa klasyfikacyjne wykorzystuje się do wyznaczania przynależności przypadków lub obiektów do klas jakościowej zmiennej zależnej na podstawie pomiarów jednej lub więcej zmiennych objaśniających. Elastyczność analizy wykonanej za pomocą drzew klasyfikacyjnych sprawia, że jest ona bardzo atrakcyjna, szczególnie, gdy nie można korzystać z metod parametrycznych z powodu założenia dotyczącego rozkładów zmiennych.

Istnieje wiele miar korelacji wyrażających zależności między dwiema lub większą liczbą zmiennych. Można tu wymienić współczynnik korelacji Pearsona, który mierzy

stopień liniowego powiązania między dwiema zmiennymi lub nieparametryczne miary zależności, oparte na podobieństwie rang dwóch zmiennych, na przykład współczynnik korelacji rang Spearmana. Regresja wieloraka umożliwia badanie związków między zmienną zależną i zbiorem zmiennych niezależnych, natomiast wielowymiarowa analiza korespondencji pomaga wyjaśniać współzależności występujące w zbiorze zmiennych jakościowych.

Korelacja kanoniczna, to dodatkowa procedura szacowania związku między zmiennymi. W szczególności, analiza ta umożliwia badanie związku między dwoma zbiorami zmiennych. Współczynniki korelacji kanonicznych można wykorzystać do redukcji zbioru zmiennych wejściowych poprzez zastąpienie ich najsilniej skorelowanymi parami zmiennych kanonicznych.

Podsumowanie

Regionalizacja to procedura wyróżniania regionów, polegająca bądź na dzieleniu powierzchni obszaru na jednostki regionalne na podstawie przyjętych kryteriów lub na łączeniu przylegających do siebie jednostek o podobnych warunkach. Ważną cechą regionalizacji jest to, że pokrywa ona cały badany obszar, czyli każda z analizowanych jednostek musi należeć do jakiegoś regionu.

Do opracowania zagadnień związanych z klasyfikacją i typologią przestrzenną jednostek regionalnych wykorzystuje się wiele różnych metod. Omówione metody wybrano pod kątem ich przydatności w badaniach rolniczo-ekonomicznych, a nie na podstawie przeglądu literatury dotyczącej regionalnego zróżnicowania rolnictwa. Nie wszystkie z prezentowanych procedur statystycznych są powszechnie wykorzystywane, szczególnie metody wielozmienne, chociaż występują one w dwóch popularnych w Polsce programach: STATGRAPHICS i STATISTICA (6, 7). W obrębie metod wielozmiennych można wyróżnić procedury, które prowadzą do redukcji zmiennych (analiza składowych głównych, analiza czynnikowa, korelacje kanoniczne, skalowanie wielowymiarowe i analiza korespondencji) lub estymacji podobieństwa badanych obiektów (analiza skupień, drzewa decyzyjne). Należy podkreślić, że wyniki metod redukujących macierz informacji nie prowadzą bezpośrednio do wydzielenia grup jednorodnych obiektów. Na podstawie nowych, najczęściej jednej, dwóch lub trzech zmiennych (składowych, czynników lub zmiennych kanonicznych), wykorzystując metody grupowania strukturalnego lub estymacji podobieństwa dokonuje się grupowania obiektów. Dla jednej zmiennej, po liniowym uporządkowaniu obiektów, do klasyfikacji stosuje się podstawowe charakterystyki statystyczne. Podobne parametry służą do podziału płaszczyzny (dla dwóch zmiennych) lub przestrzeni (trzy wyodrębnione zmienne). W przypadku zbioru danych odwzorowanego za pomocą dwóch lub większej liczby nowych zmiennych najczęściej w dalszej analizie wyników stosuje się różne algorytmy analizy skupień i analizę dyskryminacji. Podobnie postępuje się, analizując reszty równania regresji, tu obiekty klasyfikuje się najczęściej na podstawie wartości przedziałów wyznaczonych za pomocą odchylenia standardowego od modelu.

Oczywiście, na wybór metod ma również znaczny wpływ charakter danych i sposób ich pozyskania (4). Metody stosowane dla danych wyrażonych w skali nominalnej lub porządkowej różnią się od procedur stosowanych dla zmiennych ilościowych. Wykorzystując materiał liczbowy pochodzący z różnych źródeł, należy zdawać sobie sprawę z agregacji i dokładności danych oraz możliwości zrealizowania zaplanowanego celu badań.

Przy opracowywaniu wyników badań, które można określić jako „regionalne zróżnicowanie…" najczęściej stosuje się metody grupowania typologicznego, natomiast w przypadku tematów „czynniki wpływające na regionalne zróżnicowanie…" do analizy danych wykorzystuje się również metody grupowania analitycznego.

Literatura

1. F i l i p i a k K., W i l k o s S.: Wybrane metody analizy wielozmiennej i ich zastosowanie w badaniach przestrzennych. IUNG Puławy, 1998, **R(349)**.
2. Praca zbiorowa pod red. J. D z i e c h c i a r z a: Ekonometria. Metody ilościowe. AE Wrocław, 2003, **981**.
3. Praca zbiorowa pod red. W. P o c z t y: Regionalne zróżnicowanie agrobiznesu. AR Poznań, 2002.
4. Praca zbiorowa pod red. D. S t r a h l: Metody oceny rozwoju regionalnego. AE Wrocław, 2006.
5. R a o C. R.: Modele liniowe statystyki matematycznej. PWN Warszawa, 1982.
6. STATGRAPHICS – program statystyczny.
7. STATISTICA 6 – internetowy opis programu statystycznego: www.statsoft.pl/textbook/

Adres do korespondencji:

dr Krystyna Filipiak
IUNG - PIB
ul. Czartoryskich 8
24-100 Puławy
tel. (0-81) 886-34-21
e-mail: filipiak@iung.pulawy.pl