DOI: 10.37317/biul-2008-0012

ANNA RAJFURA [1]
WIESŁAW MĄDRY [1]
TADEUSZ DRZAZGA [2]
MARZENA IWAŃSKA [1]
[1] Katedra Doświadczalnictwa i Bioinformatyki, Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
[2] Przedsiębiorstwo Hodowli Roślin Rolniczych „Nasiona Kobierzyc" w Kobierzycach

Wydzielanie grup miejscowości na podstawie serii doświadczeń wielokrotnych ze zmiennym składem odmian w latach przy użyciu pakietu SEQRET
Część II. Przykład dla plonu ziarna z doświadczeń przedrejestrowych z pszenicą ozimą

The clustering of locations based on multi-environment trials with different cultivars across years using the SEQRET package
Part II. An example for grain yield from winter wheat pre-registration trials

W pracy przedstawiono zastosowanie metod analizy wzorca do wydzielenia grup miejscowości dla niekompletnych danych z doświadczeń przedrejestrowych z pszenicą ozimą. Przy użyciu pakietu SEQRET wydzielono grupy miejscowości o podobnie różnicującym wpływie na odmiany oraz wyznaczono współczynniki opisujące dopasowanie modelu. Praca prezentuje praktyczne zastosowanie metodyki, której teoretyczny opis zamieszczono w Części I.

Słowa kluczowe: pakiet SEQRET, retrospektywna sekwencyjna analiza wzorca, niekompletne historyczne bazy danych, uśrednione zredukowane macierze odległości

This work presents the example of using pattern analysis methods, which are appropriate to cluster locations for unbalanced historical data sets from multienvironmental series of experiments with winter wheat carried out for many years. The SEQRET package was used for clustering locations in the way by which they discriminate among genotypes, and for calculating determination coefficients for years. The paper presents the use of procedures described in a theoretical Part I in practice.

Key words: SEQRET package, retrospective sequential pattern analysis, unbalanced historical data sets, averaged reduced proximity matrices

WSTĘP

Na podstawie danych pochodzących z wieloletnich serii doświadczeń hodowlano-odmianowych prowadzonych w wielu miejscowościach można wykonać ocenę tych miejscowości ze względu na sposób różnicowania odmian pod kątem plonowania. Do wykonania analiz tego rodzaju służą metody pattern analysis, które wymagają danych kompletnych. Propozycję modyfikacji klasycznych metod wyznaczania odległości między miejscowościami, odpowiedniej do analizowania danych niekompletnych przedstawili DeLacy i wsp. (1996), a zaproponowaną metodykę zastosowali w pakiecie komputerowym SEQRET (nazwa od ang. SEQuential RETrospective; DeLacy i in., 1998). Prezentację idei metody oraz wzory stosowane w programach pakietu SEQRET zawarto w Części I niniejszej pracy.

Celem tej pracy, stanowiącej drugą część opracowania, jest zastosowanie pakietu SEQRET i prezentacja wyników analiz na przykładzie danych dla plonu ziarna pszenicy ozimej z doświadczeń przedrejestrowych.

MATERIAŁ I METODY

Opis programów pakietu SEQRET

Pakiet SEQRET jest zestawem siedmiu programów napisanych w języku Fortran i działających w oknie wiersza poleceń środowiska Windows lub Wine w systemie operacyjnym Linux. Programy tego pakietu wykonują retrospektywną i sekwencyjną analizę za pomocą metody typu pattern analysis, opisaną w pracy DeLacy i wsp. (1996).

Tabela 1

Zestawienie nazw programów pakietu SEQRET z nazwami rozszerzeń plików wejściowych i wyjściowych
The set of programs of the SEQRET package with the names of extensions for the input and output files

Programy Programs	Pliki wejściowe Input files	Pliki wyjściowe — Output files		
		pośrednie[1] intermediate[1]	Wynikowe — Results	
			do interpretacji[2] interpretation[2]	do grafiki[3] plotting[3]
PRESEQ (*PRE-SEQuence*)	*.NAQ; *.TXT	*.SEQ		
SEQANL (*SEQuential ANaLysis*)	*.NAQ; *.SEQ	*.PRX	*.OCC	
SEQELM *(SEQuential ELiMination)*	*.NAQ; *.PRX	*.EMA; *.MAE	*.ELM	
SEQCLU *(SEQuential CLUstering)*	*.NAQ; *.EMA	*.CLS	*.SCL	*.PCL
SEQORD *(SEQuential ORDination)*	*.NAQ; *.EMA	*.ORS	*.SOR	*.POR; *.ORP
SEQCOR *(SEQuential CORrelation)*	*.NAQ; *.SEQ; *.CLS	*.COS(n); *.ALC(n)	*.SCO(n)	
SEQSUM *(SEQuential SUMmary)*	*.NAQ; *.CLS; *.ORS; *.AL1; *.ALC		*.SUM	*.DER

[1] Pliki wejściowe dla innych programów, [2] Pliki zawierające podsumowania analiz,[3] Pliki zawierające dane do wykonania wykresów w arkuszach kalkulacyjnych

[1] Files required as input for further programs, [2] Files containing summaries of analyses,[3] Plotting files containing summary output to be imported into the worksheets for producing dendrograms

Programy wczytują dane z plików tekstowych ASCII, które można tworzyć w bazach danych lub arkuszach kalkulacyjnych oraz przeprowadzają pełną analizę dla każdego roku z wybranego przez użytkownika ciągu lat, z których pochodzą dane. Umożliwiają wybór miary odległości i metody aglomeracji, a wyniki zapisywane są w plikach wyjściowych. Pliki te zawierają opis wydzielonych grup, poziom przecięcia dendrogramu, ocenę dopasowania modelu oraz alokację miejscowości wyeliminowanych metodami opisanymi w Części I prezentowanego opracowania. Pliki wynikowe trzeba importować do pakietów rysujących wykresy, ponieważ same programy pakietu nie wykonują elementów graficznych. Zestawienie nazw programów pakietu SEQRET wraz z nazwami rozszerzeń plików wejściowych i wyjściowych w kolejności, w jakiej są wykorzystywane zawiera tabela 1. Pakiet udostępniany jest przez autorów za darmo.

Dane i analiza

Dane z prezentowanego przykładu pochodziły z doświadczeń przedrejestrowych z pszenicą ozimą przeprowadzonych w latach 1992–2007, łącznie w 16 miejscowościach z 1035 odmianami w całej serii doświadczeń. W poszczególnych latach zmieniała się zarówno liczba odmian w doświadczeniach, jak i zbiór miejscowości. Strukturę analizowanych danych przedstawia tabela 2.

Tabela 2

Liczebność odmian uprawianych w serii doświadczeń w wielu miejscowościach w latach 1992–2007
Numbers of varieties grown up in series of multienvironment experiments carried out in the years 1992–2007

Lata	Miejscowości — Locations															
Years	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16
1992	49	49	49	49	49		49	49	49	49	49	49	49	49		49
1993	53	53	53	53	53		53	53	53	53	53	53	53	53		53
1994	61	61	61	61	61		61	61	61	61	61	61	61	61		61
1996	64	64	64	64	64		64	64	64	64	64	64	64	64		64
1997	62	62	62	62	62		62	62	62	62	62	62	62	62		62
1998		59	59	59	59		59	59	59	59	59	59	59	59	59	59
1999		67	67	67			67	67	67	67	67	67	67		67	
2000			71	71				71			71	71		71	71	
2001		77	77	77				77			77	77		77	77	
2002			78	78		78		78			78	78		78	78	
2003			86	86		86		86			86	86		86	86	
2004s1			56	56				56			56	56		56	56	
2004s2			56	56				56			56	56		56	56	
2005s1			57	57				57			57	57		57	57	
2005s2			57	57				57			57	57		57	57	
2006s1			59	59				59			59	59		59	59	
2006s2			59	59				59			59	59		59	59	
2007s1			64	64				64			64	64		64	64	
2007s2			64	64				64			64	64		64	64	

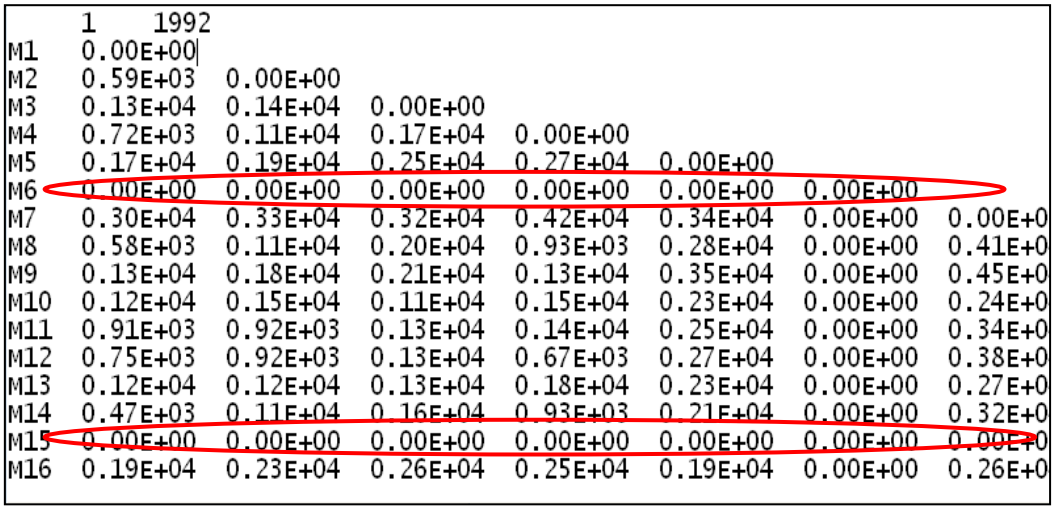
s1, s2 - oznaczenia serii doświadczeń w roku
s1, s2 - signs for series of experiments in a year

Dla tych danych przy użyciu pakietu SEQRET wykonano wydzielenie grup miejscowości pod kątem ich różnicującego wpływu na odmiany. W opcjach analizy wybrano transformację danych w sposób opisany wzorem (12) w Części I, kwadrat odległości euklidesowej jako miarę odległości między miejscowościami, wyliczanie średnich ważonych dla

odległości ze względu na różne liczby odmian w poszczególnych latach oraz metodę aglomeracji Warda dla całego zakresu lat.

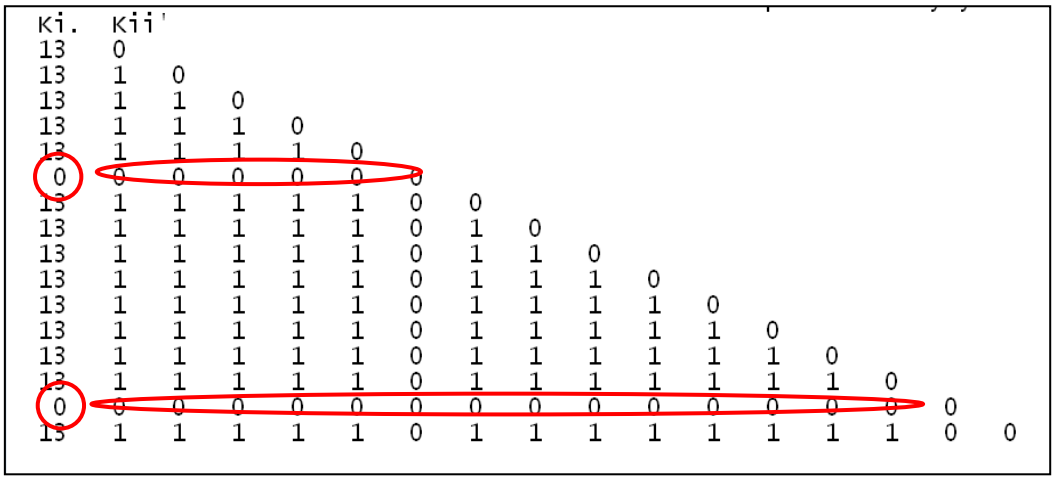
WYNIKI I DYSKUSJA

Przy użyciu programu SEQANL.EXE wyznaczono macierze incydencji oraz macierze odległości między miejscowościami dla każdego roku z analizowanego ciągu lat. Na rysunku 1 przedstawiono macierz odległości **P** dla 1992 roku.



	1	1992					
M1	0.00E+00						
M2	0.59E+03	0.00E+00					
M3	0.13E+04	0.14E+04	0.00E+00				
M4	0.72E+03	0.11E+04	0.17E+04	0.00E+00			
M5	0.17E+04	0.19E+04	0.25E+04	0.27E+04	0.00E+00		
M6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	
M7	0.30E+04	0.33E+04	0.32E+04	0.42E+04	0.34E+04	0.00E+00	0.00E+0
M8	0.58E+03	0.11E+04	0.20E+04	0.93E+03	0.28E+04	0.00E+00	0.41E+0
M9	0.13E+04	0.18E+04	0.21E+04	0.13E+04	0.35E+04	0.00E+00	0.45E+0
M10	0.12E+04	0.15E+04	0.11E+04	0.15E+04	0.23E+04	0.00E+00	0.24E+0
M11	0.91E+03	0.92E+03	0.13E+04	0.14E+04	0.25E+04	0.00E+00	0.34E+0
M12	0.75E+03	0.92E+03	0.13E+04	0.67E+03	0.27E+04	0.00E+00	0.38E+0
M13	0.12E+04	0.12E+04	0.13E+04	0.18E+04	0.23E+04	0.00E+00	0.27E+0
M14	0.47E+03	0.11E+04	0.16E+04	0.93E+03	0.21E+04	0.00E+00	0.32E+0
M15	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+0
M16	0.19E+04	0.23E+04	0.26E+04	0.25E+04	0.19E+04	0.00E+00	0.26E+0

Rys. 1. Macierz odległości P dla 1992 roku
Fig. 1. The proximity matrix P for the year 1992

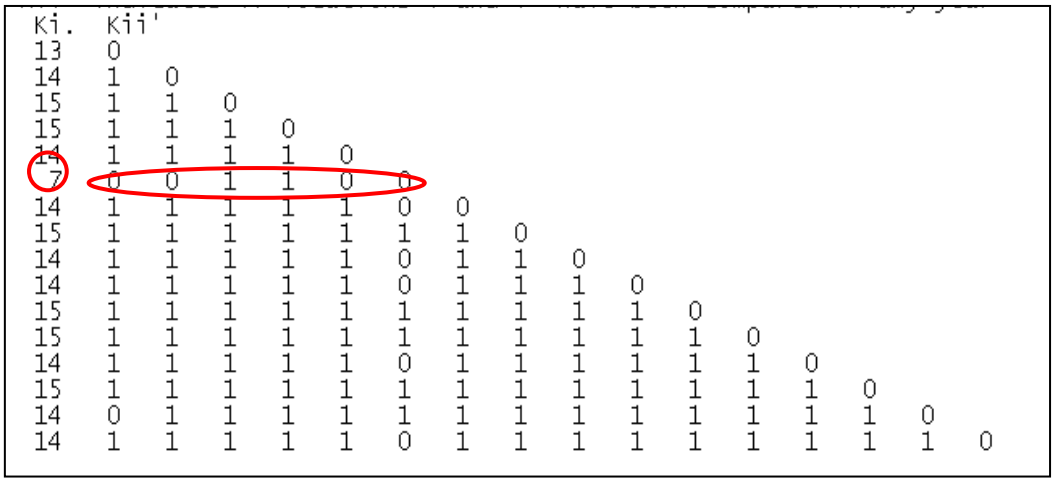


Ki.	Kii'														
13	0														
13	1	0													
13	1	1	0												
13	1	1	1	0											
13	1	1	1	1	0										
0	0	0	0	0	0	0									
13	1	1	1	1	1	0	0								
13	1	1	1	1	1	0	1	0							
13	1	1	1	1	1	0	1	1	0						
13	1	1	1	1	1	0	1	1	1	0					
13	1	1	1	1	1	0	1	1	1	1	0				
13	1	1	1	1	1	0	1	1	1	1	1	0			
13	1	1	1	1	1	0	1	1	1	1	1	1	0		
13	1	1	1	1	1	0	1	1	1	1	1	1	1	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0

Rys. 2. Macierz incydencji K dla 1992 roku
Fig. 2. The incidence matrix K for the year 1992

Zakreślono w niej wiersze odpowiadające miejscowościom M6 i M15 z wartościami zerowymi. Tylko w połączeniu z macierzą incydencji **K** dla tego samego roku przedstawioną na rysunku 2 można poprawnie zinterpretować te wartości jako brak możliwości wyznaczenia odległości między M6 (oraz M15), a jakąkolwiek inną miejscowością. Na rysunku 2 liczba *i*-ta z pierwszej kolumny przedstawia liczbę miejscowości, z którymi porównywana była *i*-ta miejscowość w 1992 roku, a pozostałe kolumny tworzą macierz incydencji **K**. Zera w *i*-tym wierszu tej macierzy oznaczają brak wspólnych odmian, a jedynki ich istnienie dla par miejscowości *i, i'*.

Program wyznacza pary macierzy — odległości i incydencji dla każdej sekwencji lat od 1992 roku. Efekt uśredniania odległości między miejscowościami poprzez lata widoczny jest na podstawie macierzy incydencji dla lat 1992–2007 zamieszczonej na rysunku 3.



Ki.	Kii'														
13	0														
14	1	0													
15	1	1	0												
15	1	1	1	0											
14	1	1	1	1	0										
7	0	0	1	1	0	0									
14	1	1	1	1	1	0	0								
15	1	1	1	1	1	1	1	0							
14	1	1	1	1	1	0	1	1	0						
14	1	1	1	1	1	0	1	1	1	0					
15	1	1	1	1	1	1	1	1	1	1	0				
15	1	1	1	1	1	1	1	1	1	1	1	0			
14	1	1	1	1	1	0	1	1	1	1	1	1	0		
15	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
14	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
14	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0

Rys. 3. Macierz incydencji K do 2007 roku
Fig. 3. The incidence matrix K until the year 2007

Zakreślono wiersze dla miejscowości M6, dla której wyznaczono 7 odległości między nią a innymi miejscowościami oraz M15, dla której wyznaczono 14 odległości. W zakreślonym szóstym wierszu dla miejscowości M6 (a także szóstej kolumnie ze względu na sposób przedstawiania wartości w tablicy trójkątnej) jest jednak osiem miejscowości (oznaczonych zerami), które nie miały wspólnych odmian z M6 w całym badanym okresie. Dla każdej miejscowości liczbę brakujących odległości można obliczyć odejmując od 15 (to maksymalna liczba odległości między *i*-tą miejscowością a każdą inną w przykładzie) wartość z pierwszej kolumny (są w niej wartości od 7 do 15). Zauważmy, że najwięcej braków — osiem zanotowano dla miejscowości M6, dwa braki dla M1 i po jednym braku dla ośmiu innych miejscowości. Dla sześciu miejscowości wyznaczono maksymalną liczbę odległości. Zatem w macierzy odległości uśrednionych poprzez lata 1992–2007 wystąpiły puste komórki (ze względów technicznych program wyświetla zera), które uniemożliwiają zastosowanie klasycznej analizy skupień.

Kolejny program SEQELM.EXE pozwala przeprowadzić eliminację miejscowości generujących puste komórki w macierzy odległości uśrednionych. Na rysunku 4 zaznaczono dwie miejscowości wyeliminowane według reguły opisanej w Części I. Są to M6 i M1. Poniżej, na tym samym rysunku w zestawieniu zamieszczono stare nazwy miejscowości (w kolumnie OLN-old location nane) oraz nowe (w kolumnie NLN-new location nane), przemianowane po eliminacji. A jeszcze niżej zamieszczono zredukowaną macierz incydencji do roku 2007. Łatwo zauważyć, że nie zawiera ona zer w wierszach dla miejscowości i dla każdej z pozostawionych czternastu miejscowości wyznaczono maksymalną liczbę trzynastu odległości. Dla tej zrównoważonej macierzy odległości przeprowadzone zostało grupowanie miejscowości przy użyciu programu SEQCLU.EXE.

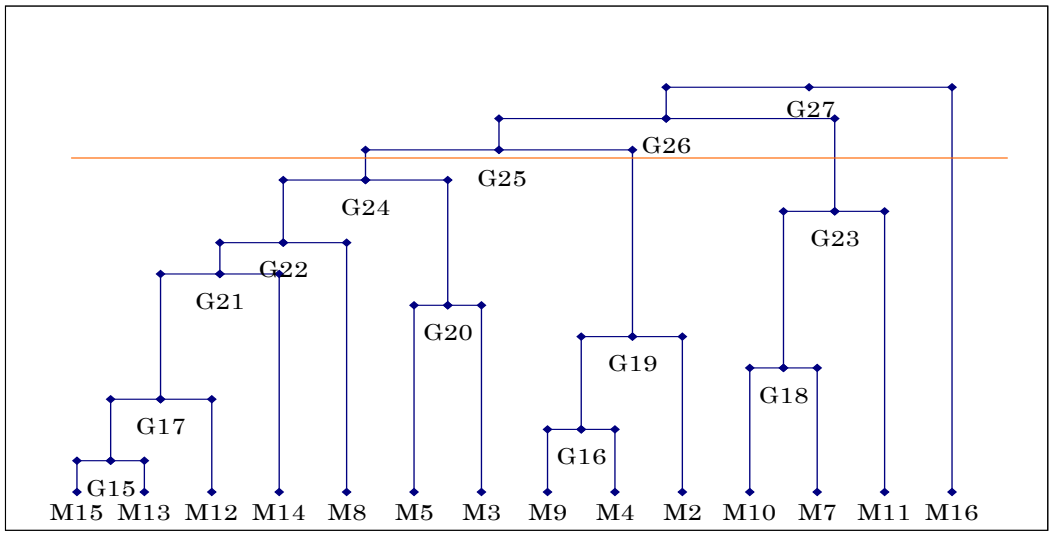
```
  For year 2007s2

Record locations with no comparisions
Locations eliminated because they have no comparisions
  Loc      name      y_com
Eliminate locations with empty cells
Locations eliminated because they have empty cells
  Loc      name      g_com      l_com      y_com
    6        M6       1148         14          7
    1        M1       3757         65         13
Recalculate marginal frequencies
Locations present and their marginal frequencies
 NLN   OLN LocName   G-Com      L_Com     Y_Com
   1     2 M2          5444         90         13
   2     3 M3          9686        156         13
   3     4 M4          9686        156         13
   4     5 M5          4235         73         13
   5     7 M7          4905         83         13
   6     8 M8          9686        156         13
   7     9 M9          4905         83         13
   8    10 M10         4905         83         13
   9    11 M11         9686        156         13
  10    12 M12         9686        156         13
  11    13 M13         4905         83         13
  12    14 M14         9016        146         13
  13    15 M15         6218         96         13
  14    16 M16         4235         73         13

Reduced Year Comparisions Matrix
   YearComp  KII'((I1=1,NLC),I2=1,I1))
  1      13      1
  2      13      1      1
  3      13      1      1      1
  4      13      1      1      1      1
  5      13      1      1      1      1      1
  6      13      1      1      1      1      1      1
  7      13      1      1      1      1      1      1      1
  8      13      1      1      1      1      1      1      1      1
  9      13      1      1      1      1      1      1      1      1      1
 10      13      1      1      1      1      1      1      1      1      1      1
 11      13      1      1      1      1      1      1      1      1      1      1      1
 12      13      1      1      1      1      1      1      1      1      1      1      1      1
 13      13      1      1      1      1      1      1      1      1      1      1      1      1      1
 14      13      1      1      1      1      1      1      1      1      1      1      1      1      1
```

Rys. 4. Miejscowości wyeliminowane i zredukowana macierz odległości do 2007 roku
Fig. 4. Locations eliminated and reduced proximity matrix until the year 2007

Jak wcześniej wspomniano, program nie wykonuje dendrogramu, a jedynie wyznacza dane na podstawie których można wykonać rysunek. W prezentowanym przykładzie dendrogram przedstawiony na rysunku 5 wykonano w arkuszu Excel. Stosowane na rysunku oznaczenia z literą M odnoszą się do miejscowości, a z literą G do grup miejscowości. Linia ciągła oznacza poziom przecięcia dendrogramu. Wydzielono cztery grupy: G24 zawierającą 7 miejscowości, G19 i G23 po 3 miejscowości i G14 z jedną miejscowością.

Po wydzieleniu przy użyciu programu SEQCOR.EXE grup miejscowości pozostawionych wykonano przyporządkowanie miejscowości wyeliminowanych do utworzonych grup. Na rysunku 6 zakreślono zestawienie z oznaczeniami miejscowości w pierwszej kolumnie (1, 6) i oznaczeniami grup w pierwszym wierszu (14, 19, 23, 24). Zawiera ono odległości miejscowości od centroidu grupy wyliczone ze wzoru (15) zamieszczonego w Części I.

W przypadku obu miejscowości najmniejszą wartość przyjęła odległość od centroidu grupy G24, zatem do tej grupy przydzielono miejscowości M1 i M6.



Rys. 5. Dendrogram wykonany na podstawie zredukowanej macierzy odległości między miejscowościami dla okresu 1992–2007
Fig. 5. A dendrogram made on the basis of reduced proximity matrix among locations for seasons 1992–2007

```
  For year 2007s2

  Location centre the years data
  Location standardise the years data

  Allocate eliminated locations
  Number of eliminated locations is       2
Number of squared distances
ElmLoc  Location group
            14         19         23         24
     1     289        289        289        289
     6       0        164        164        164
squared Euclidean distances
ElmLoc  Location group
            14         19         23         24
     1  0.98       0.80       0.85       0.61
     6   0.0        1.2        1.0       0.69
Group      14
  Number of old members      1
    They are M16
Group      19
  Number of old members      3
    They are M9       M4       M2
Group      23
  Number of old members      3
    They are M10      M7       M11
Group      24
  Number of old members      7
    They are M15      M13      M12      M14      M8       M5       M3
  Number of new members      2
    They are M1       M6
```

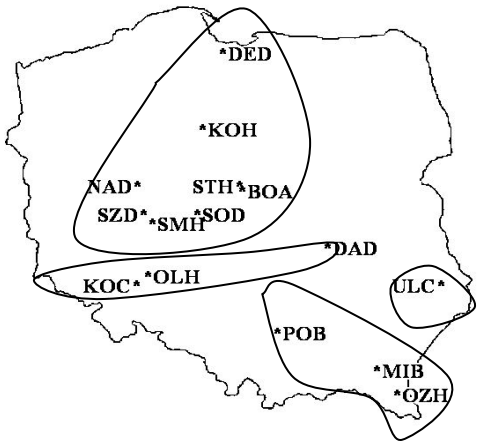
Rys. 6. Przyporządkowanie wyeliminowanych miejscowości do utworzonych grup
Fig. 6. Allocation of eliminated locations to formed clusters

Ten sam program wyznacza współczynnik dopasowania modelu według wzoru (21) zamieszczonego w Części I dla każdego z 19 lat w serii doświadczeń. Odpowiednie wartości przedstawione są w tabeli 3. Współczynniki determinacji zostały wyznaczone dwukrotnie: dla grup wydzielonych z miejscowości pozostawionych oraz dla grup po przyporządkowaniu miejscowości wyeliminowanych. Łatwo zauważyć, że wyliczone wartości nie wykazują znaczącego pogorszenia oceny dopasowania.

Tabela 3

Efektywność modelu liniowego skalibrowanego w niekompletnej serii GLY przy szacowaniu brakujących plonów ziarna pszenicy ozimej
Effectiveness of the model calibrated in incomplete GLY data for predicting grain yield of winter wheat

Rok *j* Year *j*	Współczynniki determinacji dla lat — Determination coefficients for years	
	bez wyeliminowanych miejscowości without eliminated locations	z wyeliminowanymi miejscowościami with eliminated locations
1992	0,49	0,49
1993	0,49	0,49
1994	0,59	0,58
1996	0,49	0,48
1997	0,6	0,6
1998	0,53	0,53
1999	0,56	0,56
2000	0,63	0,63
2001	0,63	0,63
2002	0,59	0,55
2003	0,64	0,63
2004s1	0,61	0,61
2004s2	0,66	0,66
2005s1	0,58	0,58
2005s2	0,61	0,61
2006s1	0,52	0,52
2006s2	0,55	0,55
2007s1	0,68	0,68
2007s2	0,64	0,64
Całkowity — Total	0,58	0,57



Rys. 7. Grupy miejscowości podobnie różnicujące plon ziarna odmian pszenicy ozimej
Fig. 7. Clusters of locations in which winter wheat cultivars differed similarly in grain yield

Wydzielone grupy miejscowości przedstawiono w rejonach geograficznych Polski na rysunku 7.

PODSUMOWANIE

Przy użyciu pakietu SEQRET wydzielono cztery grupy miejscowości biorących udział w serii doświadczeń metodą wykorzystującą uśrednianie macierzy odległości poprzez lata. Analiza skupień została przeprowadzona pomimo znacznej niekompletności wyników. W tabeli 2 łatwo można zauważyć, że miejscowość M6 pojawiła się w doświadczeniach tylko w latach 2002–2003 i nie można było porównać jej z ośmioma miejscowościami, które w tych latach nie wystąpiły. Podobnie miejscowość M1 pojawiła się w doświadczeniach tylko w latach 1992–1997 i nie można było porównać jej z dwiema miejscowościami: M6 i M15. Na tym przykładzie widać, że reguła eliminacji usuwa możliwie najmniej miejscowości z analizy (po usunięciu M1 nie jest już potrzebne usuwanie M15). Natomiast ocenę dopasowania uzyskano także dla miejscowości wyeliminowanych poprzez przydzielenie ich do najbliższych grup. Należy tu podkreślić, że opis danych uzupełniają metody ordynacyjne analizujące zależności między badanymi miejscowościami w wybranym ciągu lat, pominięte w tym opracowaniu, chociaż zawarte w pakiecie SEQRET.

LITERATURA

DeLacy I. H., Basford K. E., Cooper M., Fox P. N. 1996. Retrospective analysis of historical data sets from multi-environment trials-Theoretical development. In: Cooper M., Hammer G. L. (eds), Plant Adaptation and Crop Improvement. CAB International: 243 — 267.

DeLacy I. H., Basford K. E., Cooper M., Fox P.N. 1998. The SEQRET Package: Computer Programs for Retrospective Pattern Analysis, Version 1.1. The University of Queensland, Brisbane 4072, Australia.