Wielowymiarowe warstwowanie i alokacja próby

za pomocą metod optymalizacji numerycznej

Robert Wieczorkowski, GUS

IV Kongres Statystyki Polskiej, Warszawa, 2-4.07.2024

Plan prezentacji

- Sformułowanie matematyczne problemu
- Proponowany algorytm optymalizacji numerycznej
- Rozwiązanie początkowe
- Przykłady zastosowań praktycznych
- Wybór parametrów algorytmu i porównanie z innymi metodami

GUS

Sformułowanie matematyczne problemu

Celem klasycznego problemu z zakresu metody reprezentacyjne jest ustalenie **optymalnej liczebności próby** n do badania według schematu **losowania warstwowego**, przy ustalonej **liczbie warstw** L oraz narzuconych ograniczeniach na oczekiwane poziomy **błędów względnych** uogólnień (wartości globalnych lub średnich) dla K zmiennych dostępnych w operacie losowania (tzw. **zmiennych warstwujących**). Zakładamy, że informacje wejściowe o badanej populacji N elementarnych jednostek dla K zmiennych $X_1, \cdots, X_K$ zawarte są w macierzy:

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1K} \\ X_{21} & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{NK} \end{bmatrix}$$

Sformułowanie matematyczne problemu (2)

Poszukiwane jest rozwiązanie zadania w postaci L-elementowego wektora z liczebnościami w poszczególnych warstwach tzn. $\underline{n} = (n_1, n_2, \cdots, n_L)$, dla którego spełnione są warunki:

$$\min n = \sum_{h=1}^{L} n_h \tag{1}$$

$$CV(\widehat{T}_k) \leqslant c_k, \quad k = 1, \ldots, K \tag{2}$$

$$2 \leqslant n_h \leqslant N_h \tag{3}$$

$$n_L = N_L \tag{4}$$

gdzie n jest sumaryczną wielkością próby, N_h oznacza liczbę jednostek populacji w warstwie h. $\hat{T}_k$ oznaczają estymatory wartości globalnych dla każdej z K cech warstwujących, dla których ustalone są wymogi precyzji c_k (względne błędy standardowe).

Sformułowanie matematyczne problemu (3)

Warunek (4) określa, że stosujemy tzw. warstwę górną (*take-all stratum*), co jest uzasadnione w przypadku gdy badane zmienne mają rozkłady prawostronnie asymetryczne.
Wielkości błędów CV można obliczyć za pomocą następujących formuł wynikających z teorii klasycznego schematu losowania warstwowego:

$$CV(\hat{T}_k) = \frac{\sqrt{\sum_{h=1}^{L} N_h^2 (\frac{1}{n_h} - \frac{1}{N_h}) S_{hk}^2}}{\sum_{i=1}^{N} X_{ik}},$$

gdzie S_{hk}^2 oznacza wariancję populacyjną dla zmiennej k w warstwie h.

Sformułowanie matematyczne problemu (4)

Warstwy badanej populacji definiowane są za pomocą zestawu granic zapisanych w postaci macierzy:

$$\boldsymbol{b} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1K} \\ b_{21} & b_{22} & \cdots & b_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ b_{(L-1)1} & b_{(L-1)2} & \cdots & b_{(L-1)K} \end{bmatrix}$$

Granice dla danej cechy zapisane w ustalonej kolumnie są posortowane od wartości najmniejszych do największych.

GUS

Sformułowanie matematyczne problemu (5)

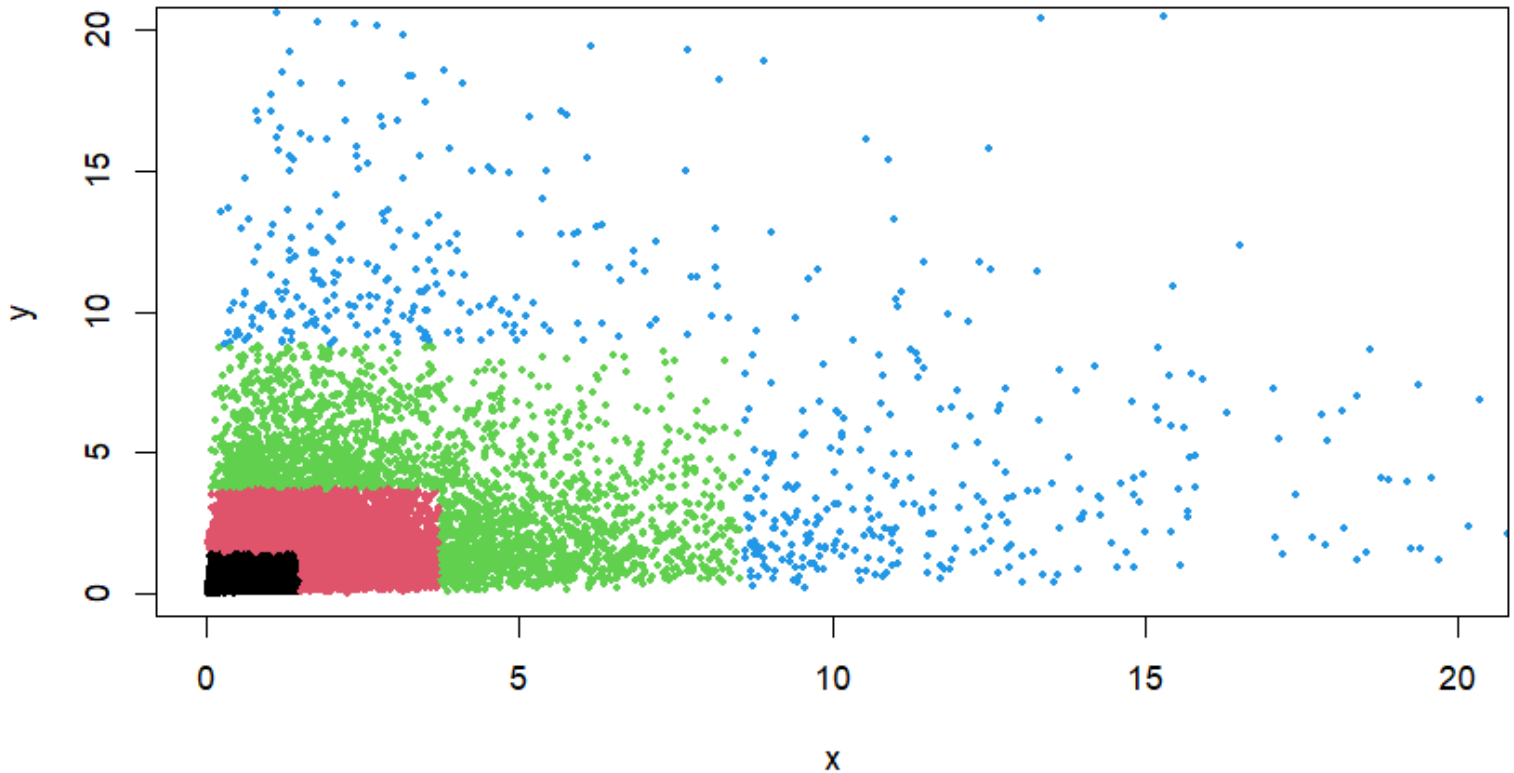
Zdefiniowanie warstw jest możliwe na wiele sposobów. W dalszych rozważaniach przyjęto metodę tzw. geometrii *L-rot-180*, która w przypadku 2 cech warstwujących ma formę odwróconej litery **L**. Na postawie macierzy $\boldsymbol{b}$ definuje się w następujący sposób przynależność każdej jednostki populacji do warstwy:

$$h_i(\boldsymbol{X}_{i*}) = \max\{h_{i1}, \ldots, h_{iK}\},\ i = 1, 2, \ldots, N$$

$$h_{ij} = \max_{t \in \{1,2,\ldots,L\}} \{t : b_{(t-1)j} \leqslant X_{ij} < b_{tj}\},\ j = 1, 2, \ldots, K,$$

gdzie h_{ij} oznacza numer warstwy dla i-tej jednostki populacji wyznaczony według granic dla j-tej cechy tzn. na podstawie wektora $(b_{1j}, b_{2j}, \ldots, b_{(L-1)j})$; dodatkowo przyjmujemy, że $b_{0j} = \min(X_{*j})$ oraz $b_{Lj} = \max(X_{*j})$.

Geometria warstw dla 2 zmiennych



Rysunek 1: L-kształtne warstwy dla 2 wymiarów

GUS

Optymalna alokacja próby

Dla ustalonego podziału populacji na L warstw należy rozwiązać problem optymalnej alokacji określony warunkami (1) - (4). Zakładając, że mamy rozwiązanie optymalnej alokacji oddzielnie dla każdej z K zmiennych, określamy alokację dla wszystkich zmiennych jako:

$$n_h = \max\{n_{hk}, k = 1, \ldots, K\}, \ h = 1, \ldots, L, \qquad (5)$$

gdzie liczebności $n_{hk}, \ h = 1, \ldots, L$ wynikają z optymalnej alokacji dla cechy k.

Optymalna alokacja dla ustalonej cechy

- Dla ustalonej zmiennej k problem alokacji jest rozwiązywany za pomocą algorytmu RNABOX (Wesołowski, Wieczorkowski, Wójciak, 2024), który jest uogólnieniem klasycznej reguły alokacji Neymana przez uwzględnienie ograniczeń dolnych i górnych na liczebności próby w warstwach.
- Algorytm RNABOX, zaimplementowany w pakiecie *stratallo* środowiska R minimalizuje oczekiwany błąd estymacji przy zadanej sumarycznej wielkości próby n. Aby go zastosować do naszego problemu, w którym minimalizujemy wielkość próby przy ustalonym względnym błędzie standardowym, stosujemy odpowiednią transformację zmiennych tzn.

GUS

Optymalna alokacja dla ustalonej cechy (2)

1. Rozwiązujemy zadanie alokacji dla wymaganej liczebności $\tilde{n} = \sum_{h=1}^{L} N_h S_{hk}^2 + c_k^2 (\sum_{i=1}^{N} X_{ik})^2$, przy ograniczeniach:

$$\frac{N_h^2 S_{hk}^2}{N_h} \leqslant \tilde{n}_{hk} \leqslant \frac{N_h^2 S_{hk}^2}{2}$$

2. Stosujemy transformację odwrotną $n_{hk} = \frac{N_h^2 S_{hk}^2}{\tilde{n}_{hk}}$
3. Zaokrąglamy uzyskane wartości do liczb całkowitych

Minimalizowana funkcja

- Opisana metoda wyznaczania alokacji dla ustalonych parametrów: L, $\boldsymbol{b}$ oraz $c_k, k = 1, \ldots, K$ może być traktowana jako numeryczny sposób obliczenia wartości funkcji wielu zmiennych $n = n(\boldsymbol{b})$ w konkretnym punkcie (dla ustalonych granic warstw).
- Rozwiązanie postawionego zadania jednoczesnej optymalnej alokacji i warstwowania można traktować jako zadanie minimalizacji funkcji n; liczba zmiennych w argumencie tej funkcji wynosi dla naszego problemu $(L-1) * K$ (są to połączone w jeden wektor kolejne kolumny macierzy $\boldsymbol{b}$).
- Ze względu na skomplikowany sposób obliczania wartości funkcji n zaproponowano przybliżone rozwiązanie problemu z użyciem metod optymalizacji numerycznej.

GUS

Algorytmy minimalizacji funkcji celu

Do minimalizacji funkcji celu wykorzystano dwie metody ze zbioru metod numerycznych opartych wyłącznie na obliczaniu wartości funkcji (w naszym przypadku wartości pochodnych są trudne do wyznaczenia):

- metoda *simplex*, zaproponowana przez Neldera i Meada (1965)
- metoda *subplex*, zaproponowana przez Rowana (1990), która stanowi uogólnienie metody *simplex*

Implementacja obu algorytmów jest dostępna w ramach pakietu *nloptr* środowiska R.

Startowe granice warstw

Algorytmy optymalizacji wymagają ustalenia wartości początkowych, czyli startowej macierzy z granicami warstw.

- Wykorzystano uogólnienie metody $cum(\sqrt{f(X)})$ Daleniusa i Hodgesa (1959), która definuje granice warstwowania za pomocą podziału na L równych części zakresu zmienności na skali skumulowanych wartości pierwiastka z estymowanej gęstości analizowanej cechy X.
- Uogólnienie polega na sparametryzowaniu powyższego sposobu w postaci metody $cum(f(X)^p)$, gdzie parametr p należy do przedziału $(0, 1)$.

GUS

Startowe granice warstw (2)

W celu wyznaczenia początkowych wartości granic warstw dla wybranych metod minimalizacji funkcji celu rozwiązywany jest podproblem optymalizacyjny:

$$\min_{(p_1,p_2,\ldots,p_K)} n(cum(f(X_1)^{p_1}),\ldots,cum(f(X_K)^{p_K})) \qquad (6)$$

Zadanie (6) rozwiązujemy algorytmem *simplex* lub *subplex*, przy czym jako wartości początkowe przyjmujemy $p_k = 0.5,\ k = 1, 2, \ldots, K$, czyli stosujemy regułę Daleniusa-Hodgesa oddzielnie do każdej z K zmiennych. W efekcie opisanego algorytmu uzyskujemy optymalny zbiór parametrów $(p_1^*, \ldots, p_K^*)$ oraz zestaw granic początkowych w postaci macierzy

$$\boldsymbol{b}_0 = (cum(f(X_1)^{p_1^*}),\ldots,cum(f(X_K)^{p_K^*}))$$

Zasadniczy algorytm minimalizacji

Zasadnicza procedura rozwiązania problemu jednoczesnego warstwowania i alokacji przebiega według następujących kroków iteracyjnych:

1. Ustalamy wartości startowe granic warstw tzn. $\boldsymbol{b}_0$ rozwiązując problem minimalizacji (6) wykonując ustaloną liczbę iteracji *maxit1* algorytmem *simplex* lub *subplex*
2. Stosujemy ustaloną liczbę iteracji *maxit2* algorytmu *simplex* lub *subplex* do minimalizacji funkcji $n(\boldsymbol{b})$, uzyskując nowe granice warstw $\boldsymbol{b}_1$
3. Powtarzamy krok 2, biorąc jako wartości startowe zestaw granic $\boldsymbol{b}_1$; ten krok wykonujemy ponownie, gdy względna poprawa uzyskanego minimum funkcji w stosunku do minimum z etapu poprzedniego jest większa niż ustalony parametr *rel_tol* (np. *rel_tol*=0.01).

Implementacja algorytmu

Algortm omawiany w referacie został zaimplementowany w języku R i jest dostępny w repozytorium na Githubie pod linkiem `https://github.com/rwieczor/mstratal`

Podstawowe parametry algorytmu

- *data* - macierz danych wejściowych (zmienne w kolumnach)
- L - liczba warstw
- $(c_1, \ldots, c_K)$ - wektor oczekiwanych błędów względnych estymacji dla poszczególnych zmiennych
- *opt_alg* - wybór algorytmu minimalizacji (*simplex* lub *subplex*)
- *maxit1* - liczba iteracji pierwszego etapu algorytmu (ustalenie przybliżenia początkowego)
- *maxit2* - liczba iteracji zasadniczego etapu algorytmu
- *rel_tol* - względna poprawa uzyskanego w 2 etapie algorytmu rozwiązania - warunek stopu dla sekwencyjnej optymalizacji

GUS

Uogólnienie dla wielu podpopulacji

- Opisane podejście wyznaczania metodami numerycznymi optymalnych granic warstw dla K cech przy ustalonej liczbie warstw L oraz poziomach precyzji $(c_1, \ldots, c_K)$ można łatwo uogólnić do istotnego w praktyce zagadnienia stałoprecyzyjnej alokacji próby między subpopulacje.
- Jeśli w konkretnym badaniu reprezentacyjnym należy dokonać alokacji n jednostek próby między J subpopulacji zapewniając wspólne poziomy precyzji c_k dla wszystkich subpopulacji i danej cechy k , to można wykorzystać opisany powyżej algorytm warstwowania oddzielnie w każdej subpopulacji, uzyskując w sumie alokację próby o wielkości $n \approx \sum_{j=1}^{J} n_j(L, c_1, c_2, \ldots, c_K)$.
- Sterując odpowiednio parametrami c_k uzyskamy w rezultacie wielkość próby odpowiednio bliską liczebności zakładanej (często przyjmuje się założenie, że $c_1 = c_2 = \ldots = c_K = c$).
- W praktyce badań rolniczych subpopulacje są najczęściej definiowane jako regiony (NUTS2), czyli $K = 17$.

GUS

Przykłady wykorzystania metody warstwowania i alokacji w GUS

- *Badanie pogłowia świń oraz produkcji żywca wieprzowego (R-ZW-S)*; co roku losowana jest próba gospodarstw rolnych; uwzględnia się jedną zmienną warstwującą - *pogłowie trzody*, stosowana jest alokacja stałoprecyzyjna dla regionów NUTS 2
- *Badanie pogłowia drobiu oraz produkcji zwierzęcej (R-ZW-B)*; co roku losowana jest próba gospodarstw rolnych; przy planowaniu głównej części próby uwzględnia się dwie zmienne warstwujące: *liczbę sztuk krów* oraz *liczbę sztuk drobiu*, stosowana jest alokacja stałoprecyzyjna dla regionów NUTS 2
- *Badanie Struktury Gospodarstw Rolnych (R-SGR)*; próba losowana do badania realizowanego co 3 lata; do alokacji głównej części próby pomiędzy regiony NUTS 2 w 2023 roku uwzględniono 4 zmienne.

Oprogramowanie przydatne do warstwowania

- Pakiet *stratification* w środowisku R: algorytmy optymalnego warstwowania i alokacji w przypadku jednowymiarowym; szczegóły opisane w artykule: Sophie Baillargeon and Louis-Paul Rivest, The construction of stratified designs in R with the package *stratification*, *Survey Methodology*, June 2011, Vol. 37, No. 1, pp. 53-65.
- w ramach pakietu *stratification* zaimplementowano m.in. algorytm Kozaka: Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 797-806,
- algorytm Kozaka jest bardzo efektywny (ale do warstwowania cechy jednowymiarowej); zastosowano w nim ideę poszukiwań losowych (*ang. random search*)

GUS

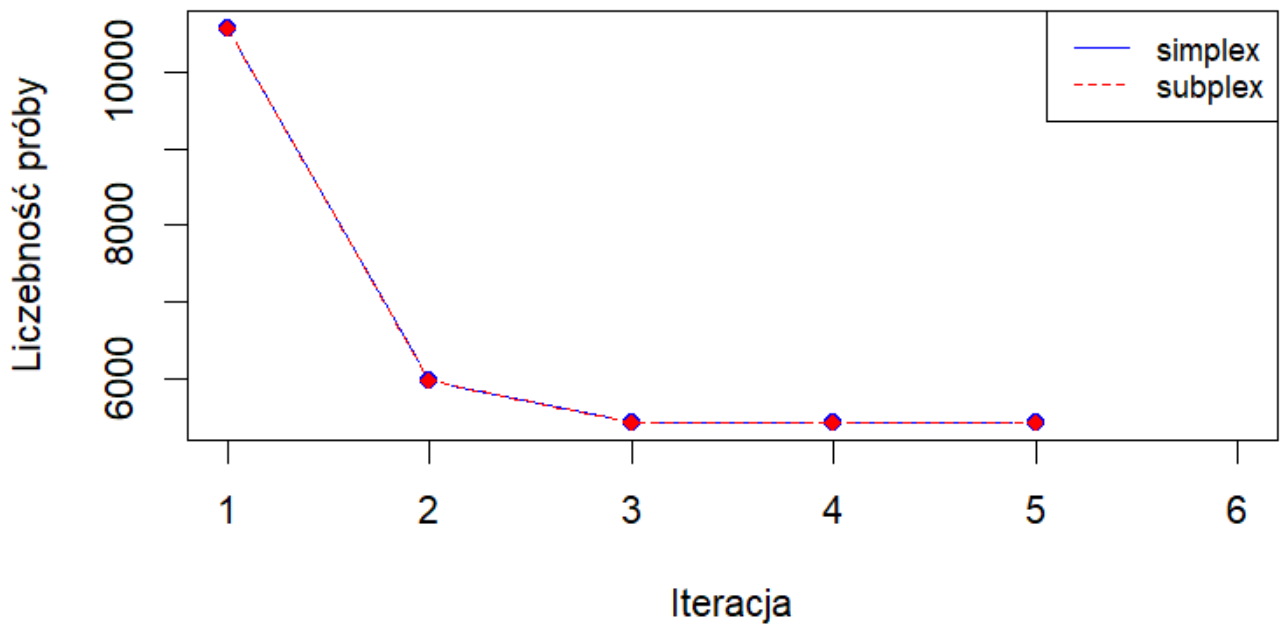
Oprogramowanie przydatne do warstwowania (2)

- Pakiet *SamplingStrata* udostępnia funkcje do numerycznego wyznaczania optymalnych granic warstw oraz alokacji próby w schemacie losowania warstwowego dla wielu zmiennych; szczegóły opisano w pracy: Barcaroli, Giulio. 2014. SamplingStrata: An R Package for the Optimization of Stratified Sampling. Journal of Statistical Software 61 (4), 1–24. `https://doi.org/10.18637/jss.v061.i04`
- Algorytmy zaimplementowane w pakiecie *SamplingStrata* rozwiązują ogólny problem omawiany w referacie; do minimalizacji funkcji celu autorzy wykorzystali metody oparte na tzw. algorytmach genetycznych.

GUS

Porównania nowego algorytmu - jedna zmienna

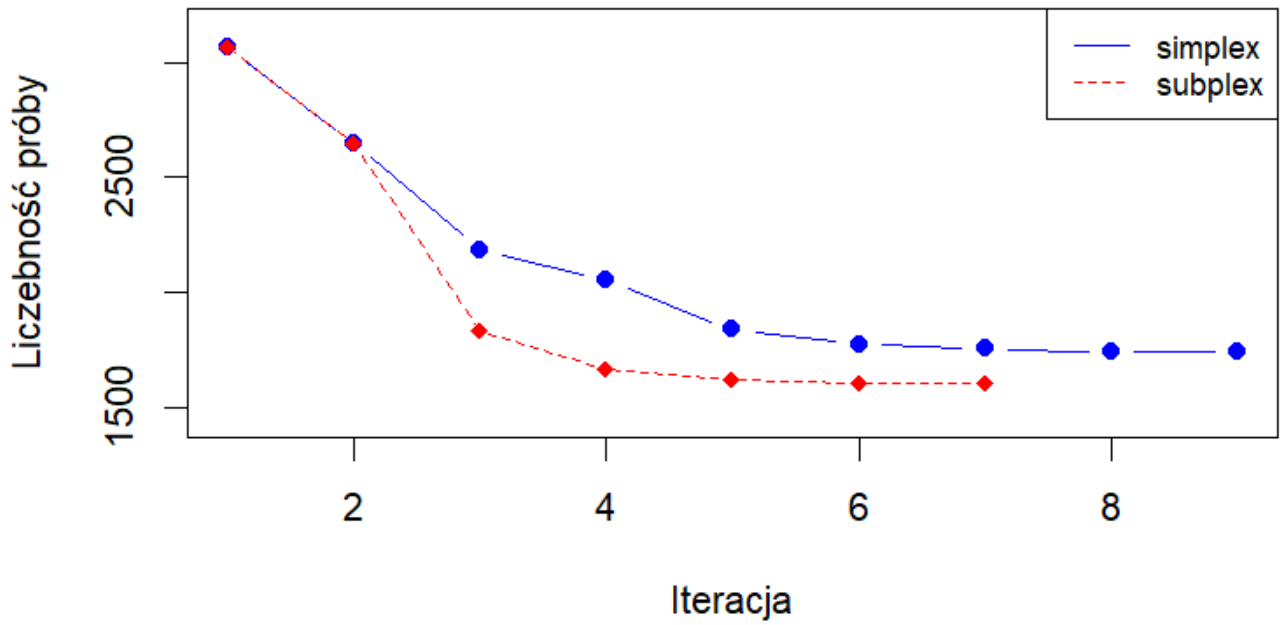
L	*c*	**algorytm**	*n*	**czas (s)**
5	0.001	*stratification: kozak*	6392	10
5	0.001	*mstratal: simplex*	5432	33
5	0.001	*mstratal: subplex*	5438	34



Rysunek 2: Przebieg optymalizacji - 1 zmienna

Porównania nowego algorytmu - dwie zmienne

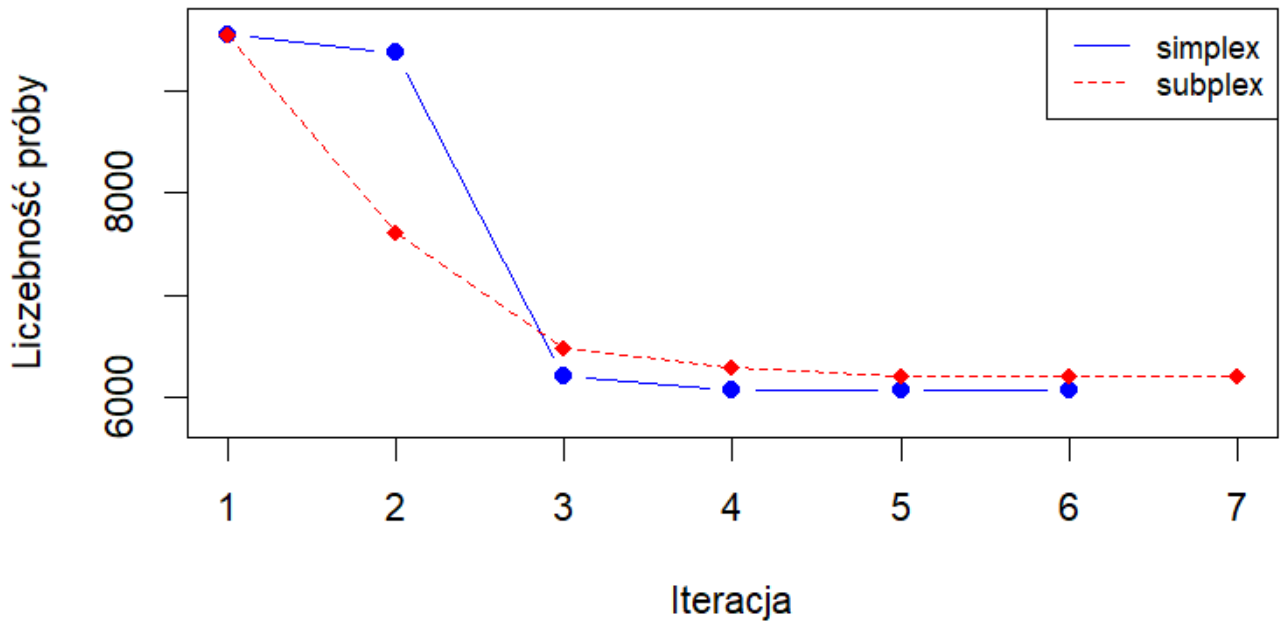
L	*c*	**algorytm**	*n*	**czas (s)**
5	0.01	*mstratal: simplex*	1740	94
5	0.01	*mstratal: subplex*	1604	69
5	0.01	*SamplingStrata*	2490	184



Rysunek 3: Przebieg optymalizacji - 2 zmienne

Porównania nowego algorytmu - 4 zmienne

L	*c*	**algorytm**	*n*	**czas (s)**
5	0.01	*mstratal: simplex*	6060	221
5	0.01	*mstratal: subplex*	6206	283
5	0.01	*SamplingStrata*	6621	924



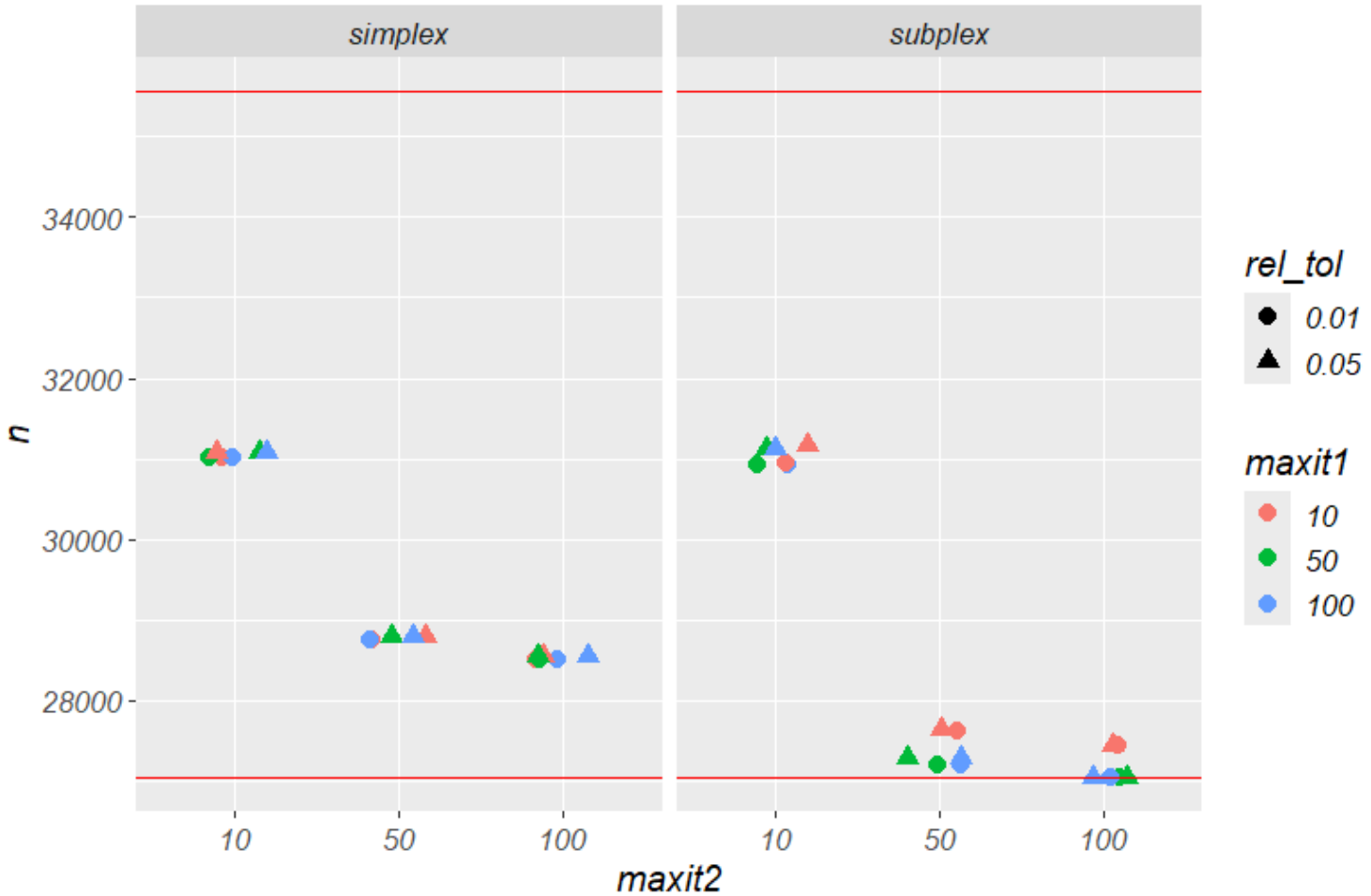
Rysunek 4: Przebieg optymalizacji - 4 zmienne

Wybór parametrów algorytmów - eksperymenty obliczeniowe

- Nowy algorytm warstwowania i alokacji zastosowano na danych jednostkowych według operatu dla 3 reprezentacyjnych badań rolniczych w 2023 roku; dla każdego z badań celem była optymalna alokacja stałoprecyzyjna w regionach NUTS 2 (ustalono parametry L i c)
- Obliczenia realizowano dla kombinacji możliwych wariantów następujących parametrów:
  - *opt_alg* - *simplex* lub *subplex*
  - *maxit1* - ze zbioru $(10, 50, 100)$
  - *maxit2* - ze zbioru $(10, 50, 100)$
  - *rel_tol* - ze zbioru $(0.05, 0.01)$
- Jako punkt odniesienia wykorzystano: algorytm Kozaka z pakietu *stratification* (dla jednej zmiennej) oraz metodę z pakietu *SamplingStrata* (dla dwóch oraz czterech zmiennych)

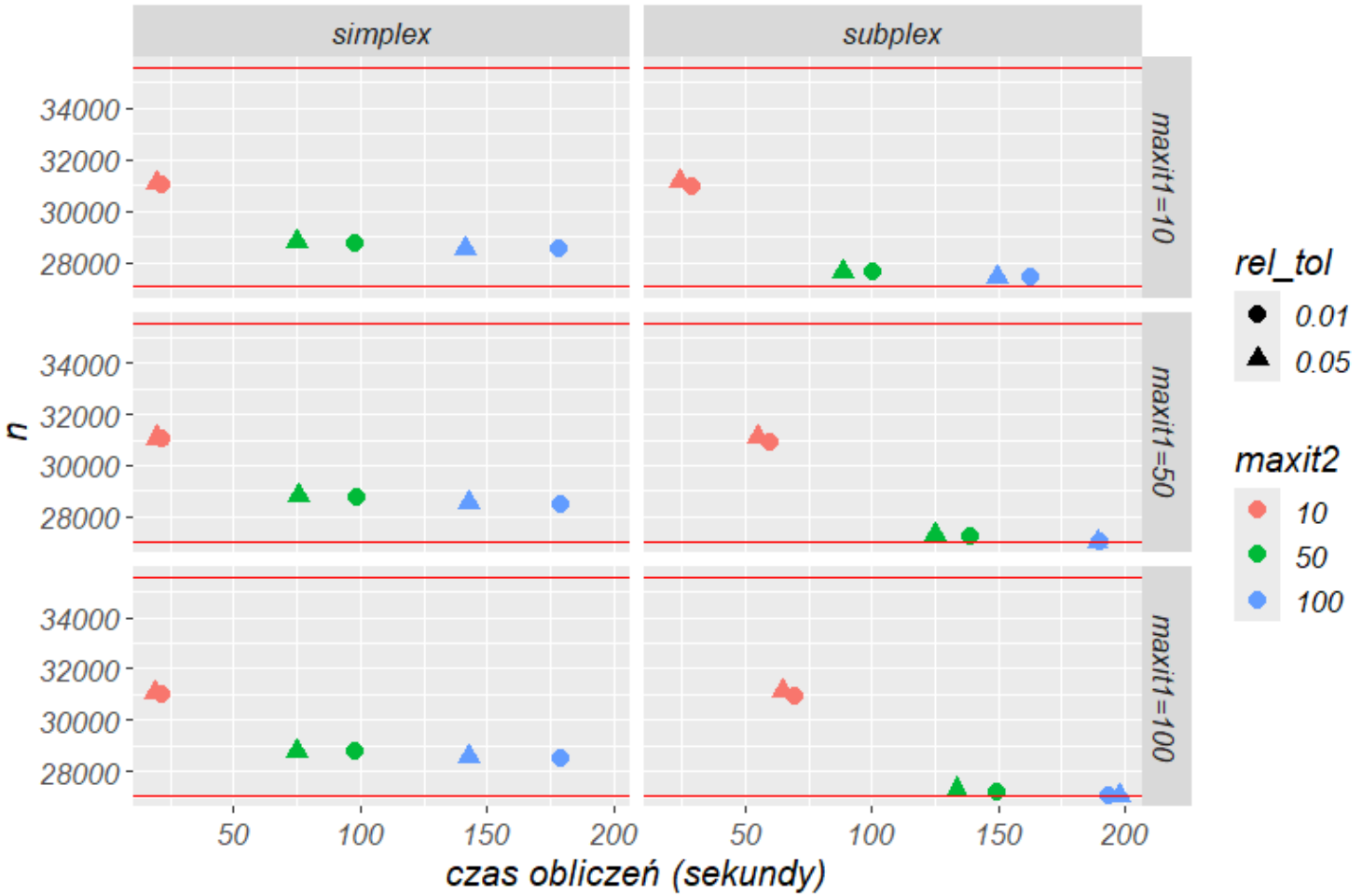
GUS

Analiza różnych parametrów algorytmu - 1 zmienna



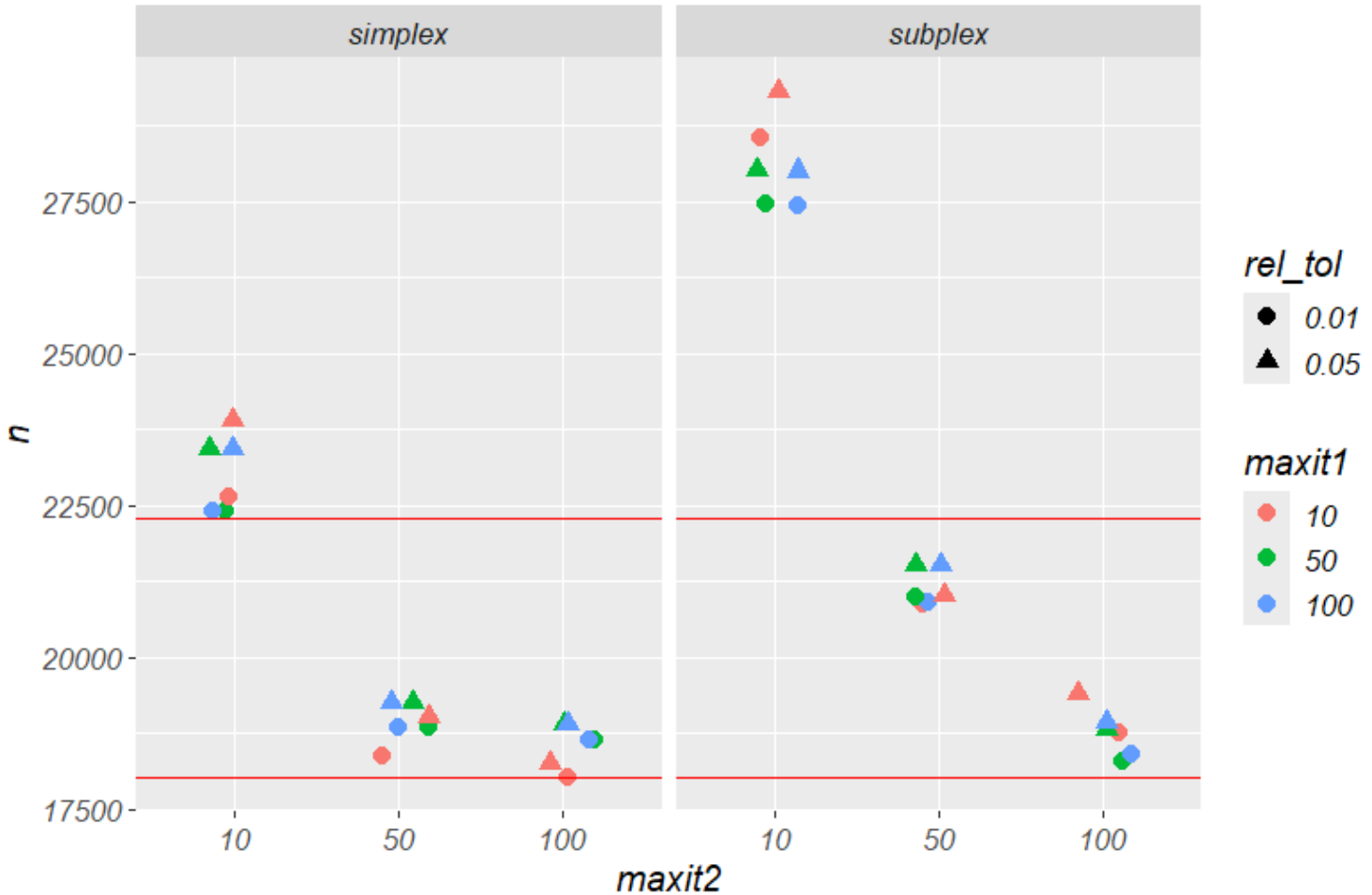
Rysunek 5: Efektywność algorytmu dla zestawu parametrów

Analiza różnych parametrów algorytmu - 1 zmienna, cd.



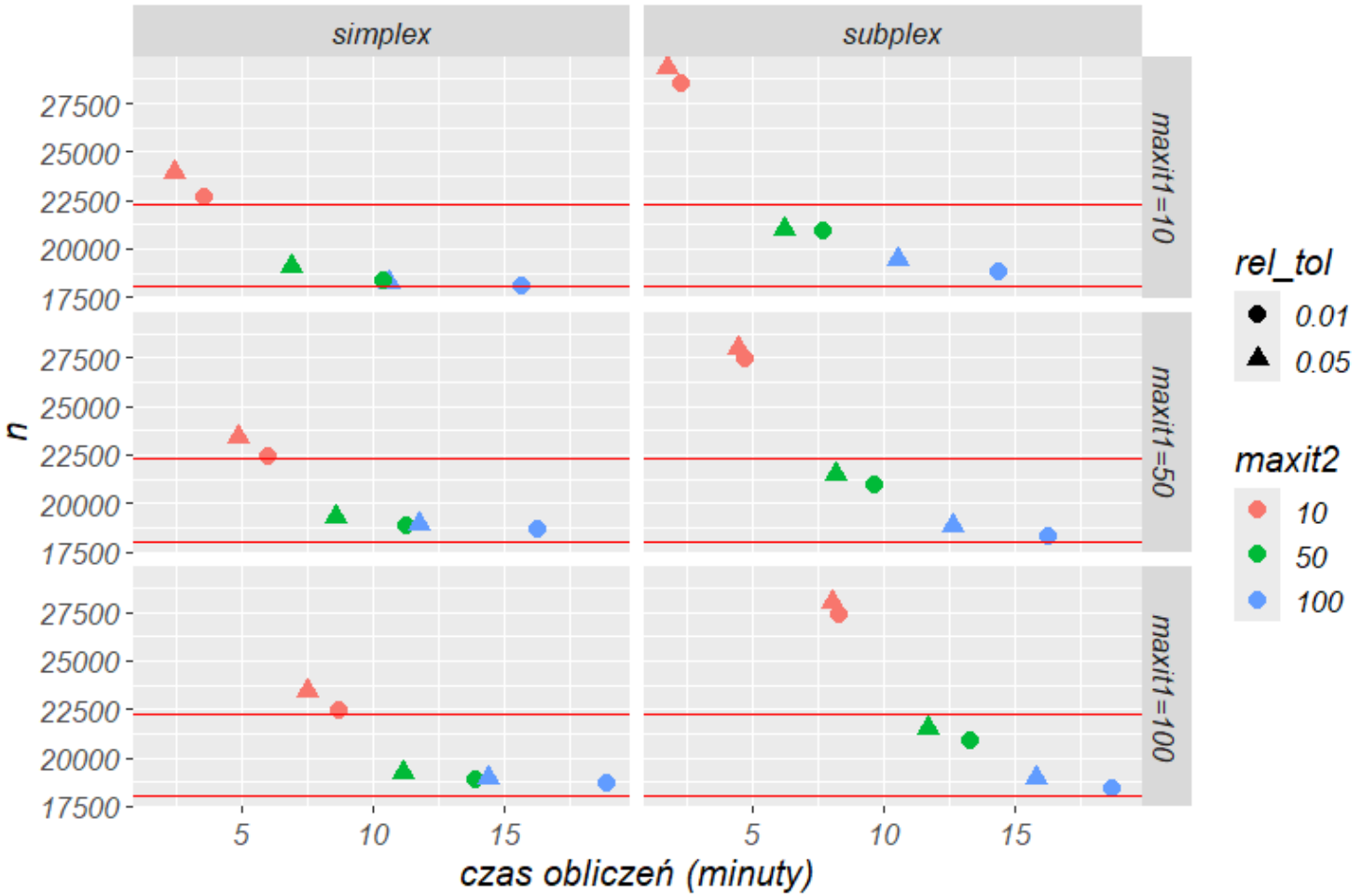
Rysunek 6: Efektywność algorytmu dla zestawu parametrów

Analiza różnych parametrów algorytmu - 2 zmienne



Rysunek 7: Efektywność algorytmu dla zestawu parametrów

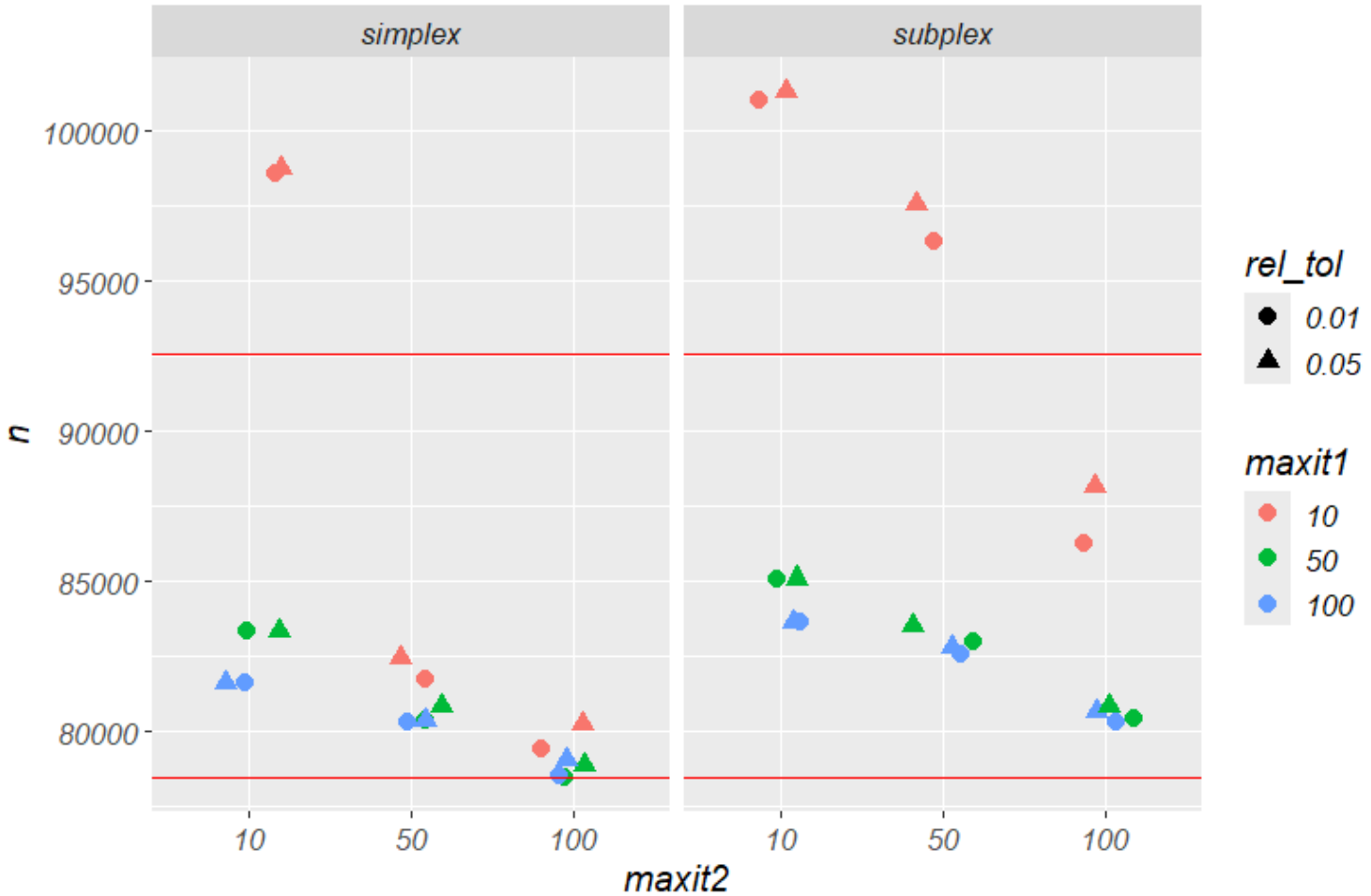
Analiza różnych parametrów algorytmu - 2 zmienne, cd.



Rysunek 8: Efektywność algorytmu dla zestawu parametrów

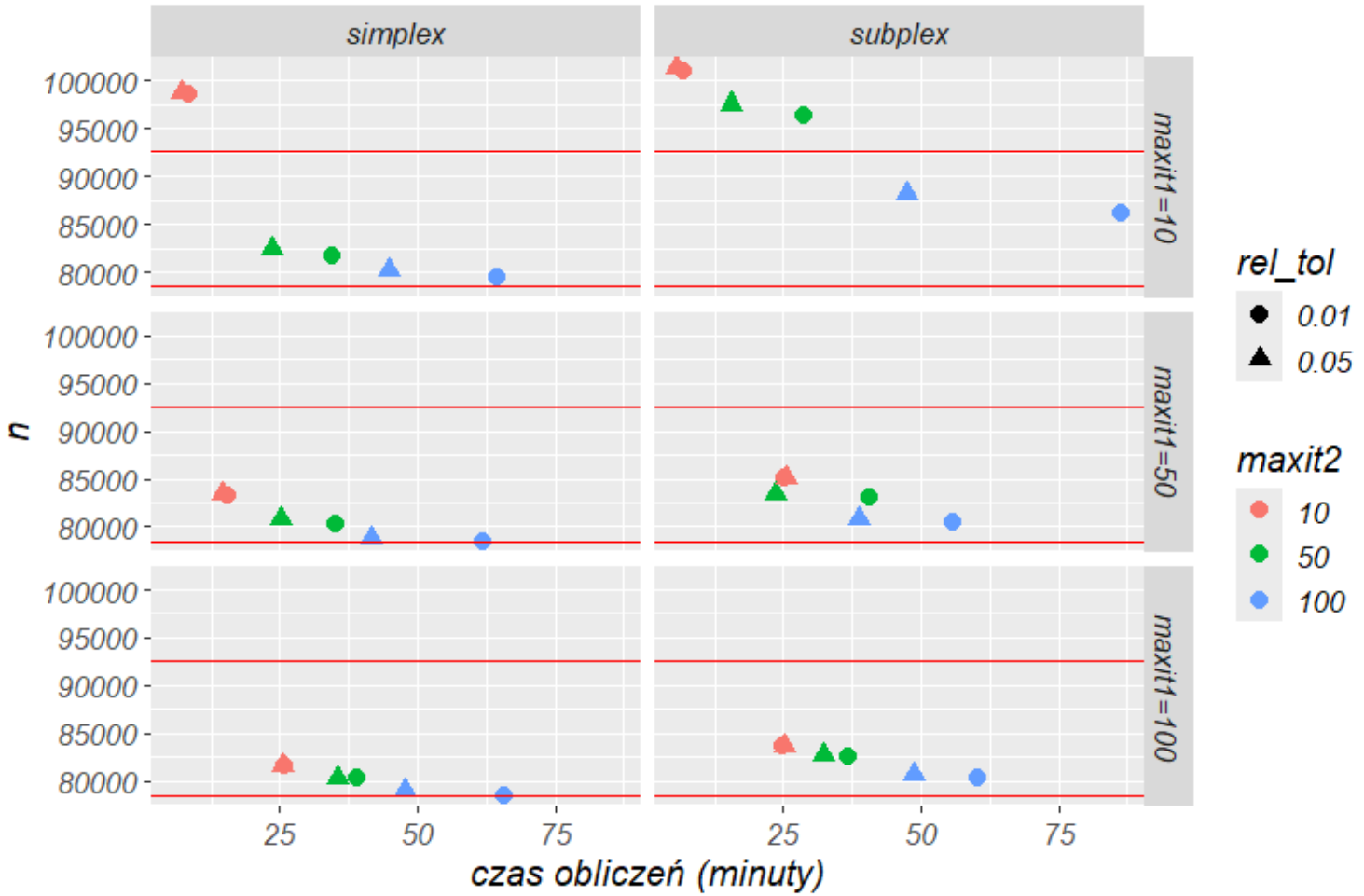
GUS

Analiza różnych parametrów algorytmu - 4 zmienne



Rysunek 9: Efektywność algorytmu dla zestawu parametrów

Analiza różnych parametrów algorytmu - 4 zmienne, cd.



Rysunek 10: Efektywność algorytmu dla zestawu parametrów

Literatura I

[1] G. Barcaroli. "SamplingStrata: An R Package for the Optimization of Stratified Sampling". In: *Journal of Statistical Software* 61.4 (2014), pp. 1–24. URL: https://doi.org/10.18637/jss.v061.i04.

[2] M.J. Box. "A new method of constrained optimization and a comparison with other methods". In: *The Computer Journal* 8.1 (1965), pp. 42–52.

[3] T. Dalenius and J.L. Hodges. "Minimum variance stratification". In: *Journal of the American Statistical Association* 54 (1959), pp. 88–101.

[4] M. Kozak. "Optimal stratification using random search method in agricultural surveys". In: *Statistics in Transition* 6.2 (2004), pp. 797–806.

GUS

Literatura II

[5] B. Lednicki and R. Wieczorkowski. "Optimal Stratification and sample allocation between subpopulations and strata". In: *Statistics in Transition* 6.2 (2003), pp. 287–305.

[6] J.A. Nelder and R. Mead. "A simplex method for function minimization". In: *The Computer Journal* 7 (1965), pp. 308–313.

[7] Rowan T. *Functional Stability Analysis of Numerical Algorithms*. Ph.D. thesis, Department of Computer Sciences: University of Texas at Austin, 1990.

[8] J. Wesołowski, R. Wieczorkowski, and W. Wójciak. "Recursive Neyman algorithm for optimum sample allocation under box constraints on sample sizes in strata". In: *Survey Methodology* 50.2 (2024).

[9] R. Wieczorkowski. *Repozytorium pakietu 'mstratal'*. URL: https://github.com/rwieczor/mstratal.

GUS

Dziękuję za uwagę !

Robert Wieczorkowski, R.Wieczorkowski@stat.gov.pl

GUS