

Syntetyczne zbiory danych w statystyce małych obszarów

Wojciech Roszka

Katedra Statystyki
Uniwersytet Ekonomiczny w Poznaniu

11.07.2018



UNIWERSYTET
EKONOMICZNY
W POZNANIU

Plan prezentacji

1. Syntetyczne zbiory danych
2. Główne podejścia w generacji danych
3. Badanie empiryczne
4. Podsumowanie

Determinanty podjęcia tematu badawczego

- ❶ Konieczność pomiaru wielu zjawisk społeczno-ekonomicznych
 - wymogi krajowe i międzynarodowe
 - detekcja i analiza zdarzeń (np. różnych rodzajów ubóstwa) w wielowymiarowych przekrojach
- ❷ Wysoki popyt na informacje na niskich poziomach agregacji
 - instytucje publiczne, prywatne, organizacje pozarządowe
 - wielowymiarowość wielu badanych problemów
- ❸ Problemy związane z badaniami częściowymi
 - niska liczebność próby → nieakceptowalne błędy losowe
 - zwiększający się wskaźnik odmów udziału w badaniu i braków odpowiedzi → błędy nielosowe
 - stosunkowo wysokie koszty i czas pozyskania danych
- ❹ Dostępność wielu źródeł danych
 - dane statystyczne i administracyjne
 - integracja danych

Definicja

Definicja

Syntetyczny zbiór danych to zanonimizowany zbiór danych jednostkowych z odpowiednimi zmiennymi oraz brzegowymi i łącznymi rozkładami, które są przynajmniej quasi-identyczne z rzeczywistymi.

Składowe

- 1 **Zbiór danych jednostkowych z badania reprezentacyjnego**
- dostarcza wyczerpujących informacji na temat różnych charakterystyk osób i/lub gospodarstw domowych.
- 2 **Prawdziwe rozkłady dla małych domen** - zagregowane tabele spisowe (lub z innych rzetelnych źródeł) dla małych obszarów geograficznych lub innych domen.

Zasady tworzenia

- Rozkład syntetycznej populacji w ujęciu jednostek geograficznych i innych domen musi być przynajmniej *quasi*-identyczny z rozkładem rzeczywistej zbiorowości.
- Brzegowe i łączne rozkłady pomiędzy zmiennymi - struktura współzależności rzeczywistej populacji - musi być odzwierciedlona jak najdokładniej.
- Niejednorodność pomiędzy subpopulacjami, w szczególności w ujęciu przestrzennym, musi być zachowana.
- Nie powinno się tworzyć jednostek syntetycznej populacji poprzez prostą replikację w oparciu o wagi, ponieważ prowadzi to nierealistycznego spadku zmienności cech w małych domenach.
- Należy zapewnić poufność danych.

Syntetyczna rekonstrukcja

Podejście to opiera się na integracji dwóch źródeł danych:

Źródła

- 1 Zagregowane **tabele spisowe (lub z innych pełnych źródeł)** – dostarczają informacji o brzegowych i łącznych rozkładach podstawowych charakterystyk społeczno-ekonomicznych w całej badanej populacji.
- 2 **Jednostkowy zbiór danych** reprezentatywny dla badanej zbiorowości – zawierający przynajmniej te same zmienne dla jednostek próby (określany jako *plik źródłowy*).

Syntetyczna rekonstrukcja

Syntetyczny zbiór danych dla całej populacji tworzony jest za pomocą dwustopniowej procedury:

Kroki

- 1 **Estymacja** – szacowany jest łączny rozkład przy użyciu obu źródeł. Powinna zostać zachowana struktura współzależności w pliku źródłowym.
- 2 **Selekcja** – jednostki z pliku źródłowego są losowane do syntetycznej populacji w taki sposób, by zachować oszacowane w poprzednim kroku łączne rozkłady prawdopodobieństwa.

Najczęściej stosowaną techniką w syntetycznej rekonstrukcji jest **iteracyjne dopasowanie proporcjonalne** (*iterative proportional fitting, IPF*).

Iteracyjne dopasowanie proporcjonalne

Algorytm IPF jest procedurą iteracyjną, która dopasowuje n -wymiarową tabelę z częstościami z niepełnego źródła do znanych rozkładów brzegowych populacji – **wagi analityczne z badania kalibrowane są do znanych sum z całej zbiorowości.**

Algorytm IPF

- Stosowany jest nieobciążony estymator Horvitz-Thompsona $\hat{Y}_d = d_i y_i$, gdzie $d_i = \frac{1}{\pi_i}$. Jeżeli natomiast zmienna pomocnicza x jest dostępna w próbie przy założeniu, że wartość globalna $X = \sum_{i=1}^N x_i$ jest znana, zwykle $\hat{X} = \sum_{i=1}^n d_i x_i \neq X$ – stąd konieczna jest kalibracja wag.
- Posiadając wyjściowe wartości liczbowe p_{ij} w komórkach tabeli pochodzące z badania częściowego oraz liczebności brzegowe R_i i C_j z badania pełnego przeprowadza się następujący algorytm: dla każdej k -tej iteracji wykonywane są dwa kroki oszacowania wartości p_{ij} , gdzie indeks górny oznacza

$$\text{kolejny numer kroku: } p_{ij}^{2k} = \frac{p_{ij}^{2k-1}}{\sum_j p_{ij}^{2k-1}} \times R_i \rightarrow p_{ij}^{2k+1} = \frac{p_{ij}^{2k}}{\sum_i p_{ij}^{2k}} \times C_j, \text{ itd.}$$

- Gdy oczekiwana liczba jednostek zostanie oszacowana dla wszystkich domen z wielowymiarowej tabeli kontyngencji, każdej jednostce w próbie przypisywane jest prawdopodobieństwo inkluzji w oparciu o zmodyfikowane procedurą wagi.
- Jednostki są losowane z próby dopóki oczekiwana liczebność w każdej domenie zostanie osiągnięta.
- Każda jednostka włączona do syntetycznej zbiorowości opisana jest przez wszystkie charakterystyki – nie tylko te, dla których następowało rozszacowanie.

Przeważanie metodą optyimizacji kombinatorycznej

Przeważanie metodą CO polega na losowaniu **odpowiedniej kombinacji jednostek** z próby w celu odtworzenia znanych rozkładów łącznych dla małych domen w wykorzystaniem narzędzi optyimizacyjnych.

Kroki w algorytmie CO

- 1 Losowanie zestawu jednostek dla małego obszaru, który będzie pełnił funkcję początkowej kombinacji.
- 2 Porównanie rozkładu łącznego w kombinacji ze znanym rozkładem. Obliczenie miary całkowitej bezwzględnej różnicy (*total absolute distance*, TAD).
- 3 Losowy wybór jednostki, która zostanie zastąpiona nową jednostką wylosowaną z próby. Powrót do kroku 2.
- 4 Powtarzanie kroku 3 dopóki nie będzie już możliwa większa poprawa TAD.

Problemy z CO

Problemy

- Dla skończonej liczby jednostek wyznaczenie wszystkich możliwych kombinacji jest teoretycznie możliwe.
- Teoretycznie również możliwe jest znalezienie najlepszej możliwej kombinacji jednostek.
- W praktyce jednak istnieje 2.63×10^{23} (co jest przybliżoną liczbą ziaren piasku na Ziemi) możliwości dopasowań do 10-elementowej małej domeny z populacji o $N = 1000$.
- Rozwiązaniem problemu jest metoda symulowanego wyżarzania (*Simulated Annealing*, SA).

Symulowane wyżarzanie

Symulowane wyżarzanie w CO

- SA jest oparte na wyżarzaniu będącego zagadnieniem z dziedziny metalurgii.
 - 1 Topienie** – ciało stałe jest najpierw topione poprzez zwiększenie temperatury do maksymalnego poziomu. W tym stanie cząsteczki ciała charakteryzują się bardzo wysoką energią i stanem ciekłym.
 - 2 Faza ochładzania** – temperatura układu jest powoli zmniejszana.
 - 3** Jeżeli temperatura maksymalna jest wystarczająco wysoka, a ochładzanie odpowiednio powolne, to wszystkie cząsteczki osiągną w końcu stan wysokiej gęstości i minimalnej energii.
- Kombinacja jednostek w małych domenach pełni rolę stanu ciała stałego.
- TAD i parametr kontrolny (np. stopa redukcji) pełnią rolę energii (E) i temperatury (T).
- Prawdopodobieństwo zaakceptowania stanu jest wyrażone poprzez $p(\delta E) = \exp\left(-\frac{\delta E}{T}\right)$:
 - im mniejsze jest δE , tym większe jest prawdopodobieństwo dokonania wymiany jednostki w kombinacji;
 - im mniejsze jest T , tym mniejsza jest szansa na zaakceptowanie nowego układu.
- Jeżeli δE jest ujemna, nowa kombinacja jest akceptowana bezwarunkowo.
- Jeżeli $\delta E \geq 0$, prawdopodobieństwo zaakceptowania zmiany jest określone przez $\exp\left(-\frac{\delta E}{T}\right)$.

Celem badania było oszacowanie przeciętnego dochodu netto w ujęciu podregionów (wg klasyfikacji NUTS 2013).

Zbiory danych

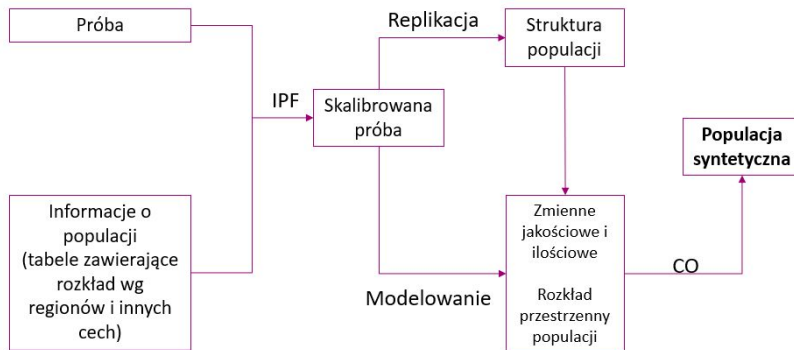
1 Jednostkowy zbiór danych EU-SILC z 2011 roku:

- województwo
- podregion
- płeć
- wiek
- status na rynku pracy
- poziom wykształcenia
- wszystkie składowe dochodu (zmienne dychotomiczne - otrzymuje/nie otrzymuje)
- wielkość GD
- klasa miejscowości
- **dochód netto** ← zmienna celu

2 Tabele spisowe (NSP 2011):

- Struktura populacji: podregion × płeć × wiek
- Struktura GD: podregion × wielkość GD × klasa miejscowości

Plan badania



Modelowanie

Modelowanie cech jakościowych

- W celu zachowania zmienności zmiennej celu, jej wartości nie powinny być replikowane w oparciu o wagi.
- Warianty zmiennych jakościowych są losowane z prawdopodobieństwa warunkowego z wykorzystaniem oszacowań wielomianowej regresji logistycznej:
 - 1 Symulowana zmienna wybierana jest z próby **S**. Zmienne niezależne muszą znajdować się zarówno w próbie **S**, jak i populacji **U**,

$$S = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,j} & x_{1,j+1} & x_{1,j+2} & \dots \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & x_{2,j+1} & x_{1,j+2} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ x_{n,1} & x_{n,2} & \dots & x_{n,j} & x_{n,j+1} & x_{n,j+2} & \dots \end{bmatrix}$$

Modelowanie

Modelowanie cech jakościowych c.d.

- 2 Model konstruowany jest w oparciu o zmienne z próby **S**. Budowany jest model wielomianowej regresji logistycznej → oszacowanie współczynników β .
- 3 Prawdopodobieństwo warunkowe wylosowania r -tego wariantu dla każdego i -tego $\hat{x}_{i,j+1}^*$ wynosi:

$$\hat{p}_{i1} = \frac{1}{1 + \sum_{r=2}^R \exp(\hat{\beta}_{0r} + \hat{\beta}_{1r}\hat{x}_{i,1} + \dots + \hat{\beta}_{jr}\hat{x}_{i,j})},$$

$$\hat{p}_{ir} = \frac{\exp(\hat{\beta}_{0r} + \hat{\beta}_{1r}\hat{x}_{i,1} + \dots + \hat{\beta}_{jr}\hat{x}_{i,j})}{1 + \sum_{r=2}^R \exp(\hat{\beta}_{0r} + \hat{\beta}_{1r}\hat{x}_{i,1} + \dots + \hat{\beta}_{jr}\hat{x}_{i,j})},$$

gdzie $r = 2, \dots, R$ a $\hat{\beta}_{0r}, \dots, \hat{\beta}_{jr}$ to oszacowania modelu regresji.

- 4 Populacja U przyjmuje postać:

$$U = \begin{bmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \dots & \hat{x}_{1,j} & \hat{x}_{1,j+1}^* \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \dots & \hat{x}_{2,j} & \hat{x}_{2,j+1}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{x}_{N,1} & \hat{x}_{N,2} & \dots & \hat{x}_{N,j} & \hat{x}_{N,j+1}^* \end{bmatrix}$$

Modelowanie

Modelowanie cech ilościowych

Modelowanie cech ilościowych przebiega w trzech krokach:

- 1 Cechę zależną x_{j+1} dyskretyzuje się w y_{j+1} w taki sposób, że tworzy się R przedziałów $c_1 < \dots < c_R$:

$$y = \begin{cases} 1 & \text{jeżeli } c_1 < x_{ij} < c_2 \\ 2 & \text{jeżeli } c_2 < x_{ij} < c_3 \\ \vdots & \vdots \\ R & \text{jeżeli } c_{R-1} < x_{ij} < c_R \end{cases}$$

- 2 Szacowany jest wielomianowy model regresji logistycznej ze zmienną zależną y_{j+1} i wektorem cech niezależnych $x_1, x_2, \dots, x_j$ dla każdej k -tej domeny oddzielnie tak jak w przypadku modelowania cech jakościowych.

Modelowanie

Modelowanie cech ilościowych c.d.

- 3 W ramach każdego r -tego przedziału szacunki $\hat{x}$ uzyskiwane są poprzez generowanie liczb losowych z rozkładu jednostajnego w ramach granic przedziału. Wyjątkiem jest przedział ostatni, gdzie ze względu na występowanie wartości odstających i skrajnych wartości losowane są z uogólnionego rozkładu Pareto:

$$\hat{x}_{ij}^* \approx \begin{cases} U(c_r, c_{r+1}) & \text{jeżeli } \hat{y}_i = r \text{ i } 1 \leq r \leq R-1 \\ GPD(\mu, \sigma, \xi, x) & \text{jeżeli } \hat{y}_i = R \end{cases}$$

Ocena jakości szacunków

Wybrane metody oceny jakości szacunków:

- 1 Walidacja wewnętrzna – porównanie rozkładów w syntetycznej populacji i próbie,

$$AD_i = |\hat{\theta}_i^* - \hat{\theta}_i|,$$

$$RD_i = \frac{\hat{\theta}_i^*}{\hat{\theta}_i} - 1.$$

- 2 Porównanie szacunków z rzeczywistymi wartościami:

$$\text{– test Z (dla cech niemierzalnych): } Z_i = \frac{\hat{p}_i^m - \hat{p}_{i0}^m}{\sqrt{\frac{\hat{p}_{i0}^m (1 - \hat{p}_{i0}^m)}{N_i}}},$$

gdzie:

$\hat{p}_i^m$ - oszacowana frakcja m -tego wariantu w domenie i ,

$\hat{p}_{i0}^m$ - znana frakcja m -tego wariantu w domenie i z populacji,

N_i - liczebność w i -tej domenie;

$$\text{– test t (dla cech mierzalnych): } t_i = \frac{\bar{x}_i - \mu_{i0}}{S_{x_i}} \sqrt{N_i - 1},$$

gdzie:

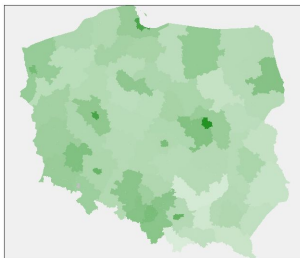
$\bar{x}_i$ - oszacowana średnia w domenie i ,

μ_{i0} - znana średnia w domenie i z populacji,

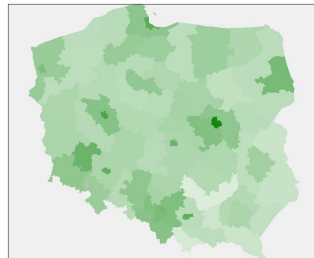
S_{x_i} - oszacowane odchylenie standardowe cechy w domenie i .

- 3 Metody symulacyjne (w opracowaniu).

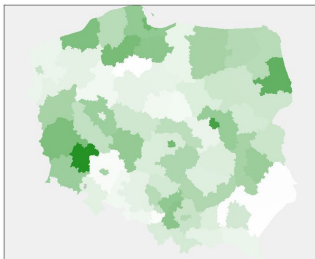
Wyniki - syntetyczny zbiór 38 mln Polaków



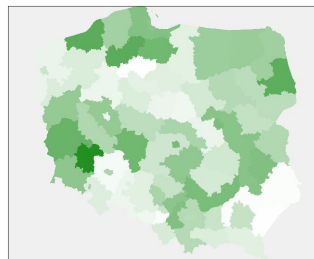
SAMPLE
8,000
7,000
6,000
5,000
4,000
3,000
2,000



SYNT
8,000
7,000
6,000
5,000
4,000
3,000
2,000



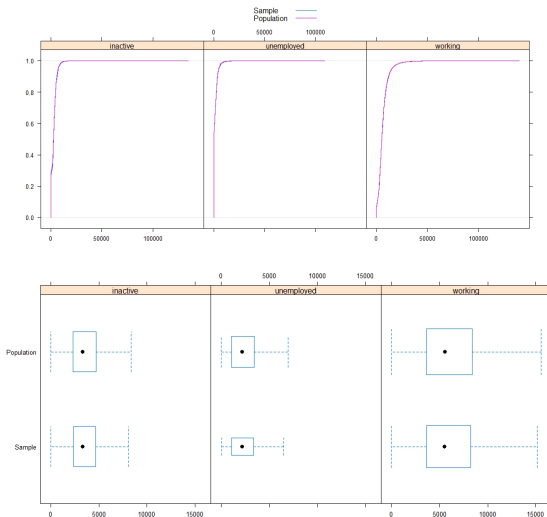
ATD
600
500
400
300
200
100
0



RTD
0.12
0.10
0.08
0.06
0.04
0.02
0.00

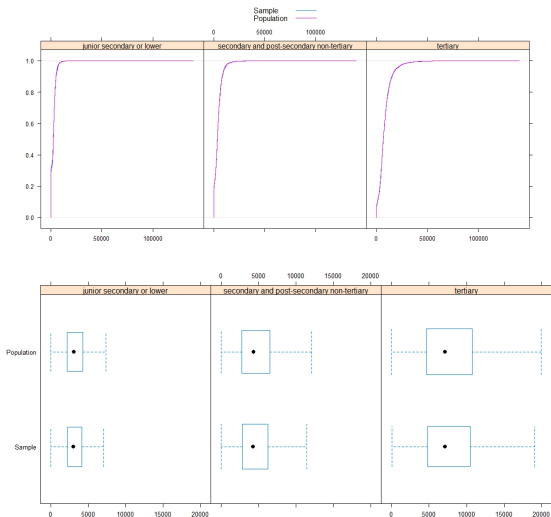
Porównanie dystrybuant dochodu netto

W ujęciu sytuacji na rynku pracy



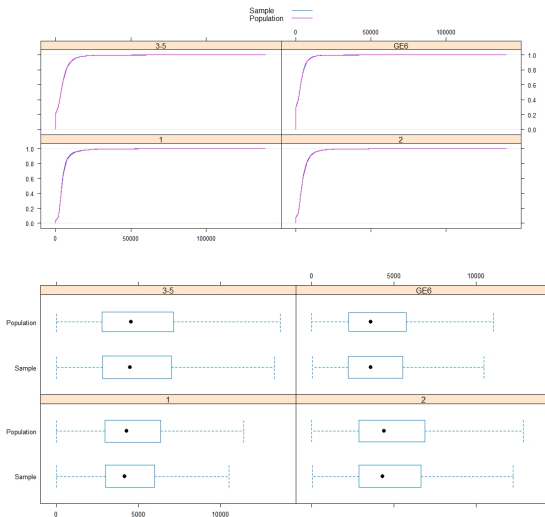
Porównanie dystrybuant dochodu netto

W ujęciu poziomu wykształcenia



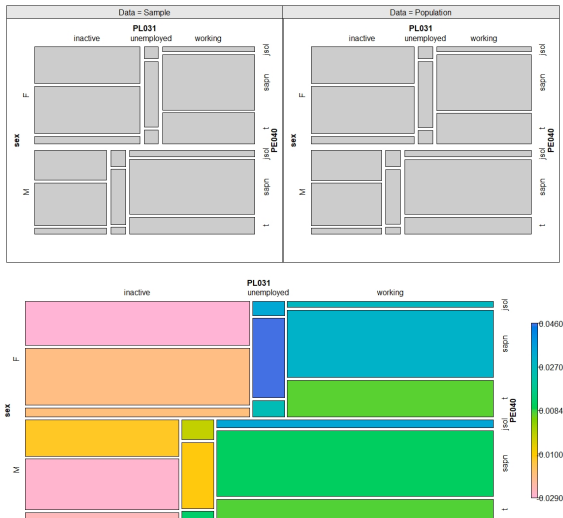
Porównanie dystrybuant dochodu netto

W ujęciu wielkości GD



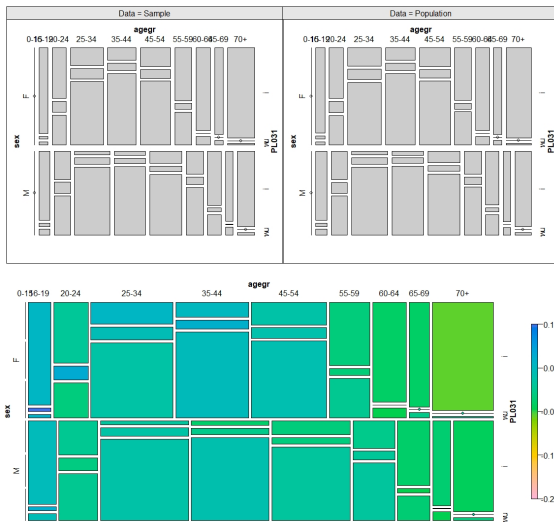
Porównanie rozkładów łącznych

Płeć × status na rynku pracy × poziom wykształcenia



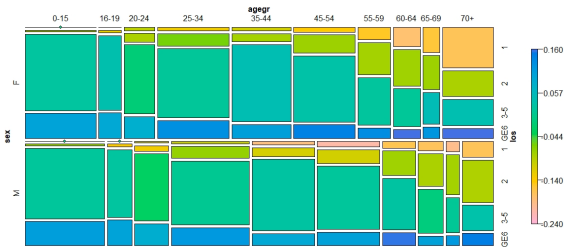
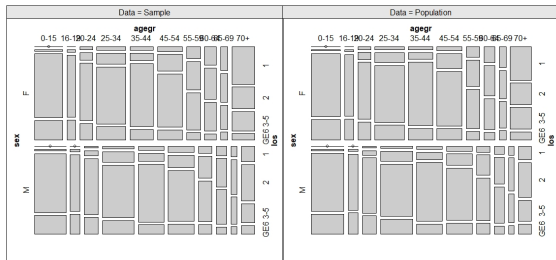
Porównanie rozkładów łącznych

Płeć × wiek × status na rynku pracy



Porównanie rozkładów łącznych

Płeć × wiek × wielkość GD



Struktura wybranych miar jakości w ujęciu podregionów

Percentyl	p-wartość testu $t^{1,2}$	AD ³	RD ⁴
MIN	≈ 0	0,2782	0,0001
5	≈ 0	6,9875	0,0017
10	≈ 0	30,6797	0,0073
25	2.36E-252	57,9740	0,0135
50	2.23E-83	114,7795	0,0293
75	2.53E-22	210,4090	0,0524
90	1.11E-07	299,0844	0,0631
95	0.1823	330,2550	0,0792
MAX	0.9610	514,6829	0,1036

¹W przypadku 5 podregionów stwierdzono p-wartość mniejszą od $\alpha = 0,05$.

²Z uwagi na brak znanych wartości średnich dochodów z populacji, za wartości „prawdziwe” przyjęto oszacowania z próby.

³Bezwzględna różnica (*absolute difference*, AD).

⁴Względna różnica (*relative difference*, RD).

Podsumowanie

- 1 Zaprezentowane metody w zadowalający sposób odzwierciedlają rozkłady brzegowe i łączne cech.
- 2 Sztuczna populacja charakteryzuje się dobrą jakością również w małych domenach.
- 3 Trwają prace nad wyznaczeniem MSE i innych charakterystyk jakości oszacowań umożliwiających porównywanie ich z innymi metodami.
- 4 Symulowanie populacji może mieć znacznie więcej zastosowań, np. analiza scenariuszy.
- 5 I dobrze być stwórcą...

Dziękuję za uwagę!

Wybrana literatura

- Alfons A., Kraft S., Templ M. (2011), "Simulation of close-to-reality data for households surveys with application to EU-SILC", Statistical Methods & Applications, Springer
- Lovelace R., Dumont M. (2016), "Spatial Microsimulation with R", CRC Press, Taylor & Francis Group, Boca Ranton, FL, USA
- Rahman A. (2009), "Small Area Estimation Through Spatial Microsimulation Models: Some Methodological Issues", NATSEM, University of Canberra
- Rahman A., Harding A. (2017), "Small Area Estimation and Microsimulation Modeling", CRC Press, Taylor & Francis Group, Boca Ranton, FL, USA
- Rahman A., Harding A., Tanton R., Liu S. (2010), "Methodological Issues in Spatial Microsimulation Modelling for Small Area Estimation", International Journal of Microsimulation
- Tanton R., Edwards K. (ed.) (2013), "Spatial Microsimulation: A Reference Guide for Users", Springer
- Templ M., Meindl B., Kowarik A., Dupriez O. (2017), "Simulation of Synthetic Complex Data: The R Package simPop", Journal of Statistical Software