Metody losowania prób przestrzennych w badaniach ekonomicznych

Tomasz Bąk

Uniwersytet Ekonomiczny w Katowicach, Katedra Statystyki, Ekonometrii i Matematyki

30 marca 2017

Plan prezentacji

- Wprowadzenie
- Losowanie z populacji ustalonej i skończonej
- Dobór próby na podstawie modelu nadpopulacji
- Adaptacyjne oraz pokrewne metody losowania
- Ocena wpływu lasu na redukcję emisji dwutlenku węgla

Wprowadzenie

Cele pracy

- Cel teoriopoznawczy - przekrojowe przedstawienie metod doboru próby przestrzennej wykorzystywanych w naukach ekonomicznych, ze szczególnym uwzględnieniem wyników własnych.
- Cel praktyczny (utylitarny) - dotyczy wskazania zastosowań omówionych metod w badanach ekonomicznych z uwzględnieniem wyników własnych, szczególnie skupiając się na zastosowaniu w badaniach zdolności lasu do pochłaniania dwutlenku węgla.

Powstanie teorii losowania przestrzennego

Potrzeba nadania metodzie reprezentacyjnej wymiaru przestrzennego pojawiła się w latach 50-tych XX wieku i wiąże się z pracami D. G. Krige'a ([13]) oraz S. Zubrzyckiego ([23, 24]). W literaturze polskiej pierwsze zwarte opracowanie, które ujęło metody statystyki przestrzennej w zakresie umożliwiającym aplikację przedstawił Przybycin [15].

Na przecięciu statystyki przestrzennej i ekonometrii powstała ekonometria przestrzenna. Prekursorską pracą dla tej dziedziny nauki była książka *Spatial Econometrics* napisana przez Paelincka i Klaassena [14]. Stworzyła ona teoretyczne podwaliny do dalszego rozwoju ekonometrii przestrzennej. W literaturze polskiej na uwagę za zasługuje *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych* Sucheckiego [17]. Było to pierwsze w Polsce opracowanie, które w szerokim zakresie omawia nowoczesne metody i modele ekonometrii przestrzennej.

Definicja

Zmienna zregionalizowana to pole losowe (proces stochastyczny)

$$\{Y(d), \quad d \in D\},$$

gdzie D jest ustalonym podzbiorem przestrzeni $\mathbb{R}^k$ *(por. np. [2]).*

Zmienna zregionalizowana

Zmienną zregionalizowaną wyróżnia od innych zmiennych rozważanych w statystyce to, że posiada lokalizację. Często też jej wartości układają się w struktury, co objawia się autokorelacją przestrzenną oraz przestrzenną heterogenicznością.
Zmiennej zregionalizowanej można nadać jeszcze dodatkowy wymiar, poprzez jej obserwacje w kilku różnych momentach czasu. Badania takiej zmiennej nazywa się badaniami wielookresowymi. Analizą danym pochodzących z takich badań zajmował się m.in Żądło [25].
Przykładami zmiennych zregionalizowanych są średnie dochody na mieszkańca gospodarstwa domowego, intensywność opadów deszczu, czy też ceny gruntów.

Losowanie z populacji ustalonej i skończonej

Podstawowe metody doboru próby, znane ze statystyki ‘nieprzestrzennej‘

- Próba prosta
- Losowanie systematyczne
- Losowanie warstwowe
- Losowanie dwustopniowe
- Losowanie grupowe

Uwzględnienie przestrzennej autokorelacji i heterogeniczności

Wymienione powyżej metody statystyki ’nieprzestrzennej’ można stosować w statystyce przestrzennej. Jednak autokorelacja i heterogeniczność zmiennych zregionalizowanych istotnie zmienia ocenę efektywności tych metod. W pracy omówiony zostanie wpływ charakterystyk zmiennej zregionalizowanej na efektywność wymienonych metod losowania.

Losowania przestrzenne - dobór próby z siatki

W statystyce przestrzennej często na badaną populację nakłada się siatkę wielokątów, a następnie losuje się elementy siatki (wielokąty). Dodatkowo zwiększa się odległości pomiędzy elementami w próbie, aby zwiększyć efektywność losowania. W pracy omówione zostaną w szczególności następujące metody losowania:

- Losowanie z wykorzystaniem macierzy sąsiedztwa (Wywiał [20])
- Metoda kostki (Cube method) (Deville, Tillé [4])
- Metoda GRTS (Stevens, Olsen, [16])
- Metoda lokalnych kluczy (Local pivotal method) (Grafström, Lundström, Schelin [8])
- Metoda podwójnie zrównoważonego losowania przestrzennego (Doubly balanced spatial sampling)(Grafström,Tillé [9])

Efektywność metod losowania z siatki zostanie omówiona na przykładzie badania sieci sklepów detalicznych prowincji Trentino, przeprowadzonego przez Dickson z zespołem w 2009 roku [5].

Losowanie próby uporządkowanej w oparciu o macierz sąsiedztwa

Wprowadzone przez Wywiała metody losowania wykorzystujące macierz sąsiedztwa były przeznaczone dla prób nieuporządkowanych. Pierwsza z metod preferowała elementy sąsiadujące, druga elementy nie będące sąsiadami. W pracy obie metody zostaną przeniesione na próby uporządkowane.
Dla obu metod losowania wyprowadzony zostanie plan losowania, jak również schematy losowania. Wreszcie obie metody zostaną zilustrowane przykładem.

Korygowanie podziału populacji na warstwy

Zagadnieniem warstwowania populacji na podstawie optymalnego podziału obszaru zmienności zajmowali się m.in. Dalenius [3] i Wywiał [21]. Kozak w 2004 roku wykazał, że jeżeli korelacja pomiędzy zmienną badaną a zmienną dodatkową jest wysoka (szczególnie jeżeli wartość współczynnika korelacji jest bliska 1), to warstwowanie oparte na zmiennej dodatkowej jest wystarczająco efektywne [11].

W pracy zaprezentowany zostanie algorytm optymalizacji liczby warstw w sytuacji, gdy badacz dysponuje gotowym podziałem na warstwy i planuje dobór próby w oparciu o optymalną alokację Neymana. Przedstawiony zostanie opis teoretyczny tej metody wraz z przykładem ilustrującym.

Dobór próby na podstawie modelu nadpopulacji

Optymalizacja doboru próby w oparciu o kriging

Jest to jeden z dwóch podstawowych sposobów doboru próby w podejściu modelowym. Optymalizacja doboru próby w oparciu o kriging polega na minimalizacji błędu średniokwadratowego predyktora.
Omówionym przykładem metody optymalizacji poprzez kriging będzie symulowane wyżarzanie przestrzenne(*spatially simulated annealing*). Jest ono przestrzennym wariantem metody symulowanego wyżarzania (*simulated annealing*). Nazwa algorytmu wywodzi się z metalurgii, gdzie wyżarzanie stopionego metalu prowadzi do osiągnięcia przez niego stanu krystalicznego, który to stan jest jego globalnym minimum, jeśli chodzi o energię termodynamiczną [12].
Groenigen i Stein [10] rozwinęli metodę symulowanego wyżarzania na populacje przestrzenne, przyjmując jako kryterium minimalizację wariancji krigingu i definiując w ten sposób symulowane wyżarzanie przestrzenne.

Optymalizacja doboru próby w oparciu o wariogram

Zasadnicze jest pytanie: jak należy dobrać próbę, aby uzyskać precyzyjny wariogram? Intuicyjnie wydaje się oczywiste, że próba powinna pozwalać na porównania dużych, średnich i małych (relatywnie do skali zróżnicowania przestrzennego) odległości pomiędzy elementami próby. Generalnie w metodach optymalizujących dobór próby względem wariogramu, nacisk położony jest na różnego rodzaju regularność w rozlokowaniu próby.

Funkcję celu w symulowanym wyżarzaniu przestrzennym można określić tak, aby zwiększała ona precyzję wariogramu. Zawadzki [22] rozważał użycie dwóch kryteriów zwiększających precyzję wariogramu w symulowanym wyżarzaniu przestrzennym: minimalizację średniej odległości do najbliższego sąsiada oraz kryterium Warricka-Myers'a, jako funkcji celu w symulowanym wyżarzaniu przestrzennym [22]. Obie te metody zostaną omówione w pracy.

Przykład zastosowania

Losowanie z wykorzystaniem podejścia modelowego uzupełnione zostanie przykładem badania lasów łęgowych prowadzonego przez B.N.I. Eskelson z zespołem [6]. Badanie dotyczyło mikroklimatu nadbrzeżnych lasów łęgowych w stanie Oregon (Stany Zjednoczone).

Adaptacyjne oraz pokrewne metody losowania

Adaptacyjne losowanie grupowe

Omówiony zostanie najstarszy schemat losowania adaptacyjnego - adaptacyjne losowanie grupowe. Schemat adaptacyjnego losowania grupowego został zaproponowany przez Thompsona w 1990 roku [18]. W pracy omówione na przykładach zostaną dwa kluczowe dla tej metody losowania pojęcia: sąsiedztwo oraz klaster (grupa). Dla przykładowej populacji wyznaczone zostaną również prawdopodobieństwa inkluzji pierwszego rzędu.
Ponadto przedstawiona zostanie teoria dwóch najczęściej stosowanych odmian tej metody losowania: z próbą początkową wybieraną jako próba prosta bez zwracania oraz adaptacyjne warstwowe losowanie grupowe.

Adaptacyjne losowanie sieciowe

W pracy przedstawiona zostanie również zaproponowane przez Thompsona adaptacyjne losowanie sieciowe [19]. Pod pojęciem adaptacyjnego losowania sieciowego kryje się cała klasa elastycznych planów losowania, wykorzystywanych w losowaniu z populacji przestrzennych i populacji zebranych w sieci. Dla populacji zebranych w sieci adaptacyjne losowanie sieciowe z powodzeniem stosuje się badaniach internetowych sieci społeczniościowych takich jak np. Facebook.
Przedstawienie w pracy tej metody jest istotne również z innego powodu. Dwie autorskie metody: trójkątna metoda losowania przestrzennego oraz losowanie przestrzenne wspierane krigingiem wykorzystują niektóre rozwiązania zaproponowane przez Thompsona w adaptacyjnym losowaniu sieciowym.

Trójkątna metoda losowania przestrzennego

Przedstawiona zostanie autorska metoda losowania przestrzennego - trójkątna metoda losowania przestrzennego.
Ta metoda daje możliwość modyfikowania prawdopodobieństw wyboru poszczególnych elementów populacji w kolejnych etapach losowania. Tworzenie 'na bieżąco' tych prawdopodobieństw nie jest losowe, a wynika z wiedzy o populacji zawartej w elementach już wylosowanych. Jest to więc pewnego rodzaju algorytm uczący się, który pozwala na ocenę wartości niewylosowanych elementów populacji na podstawie elementów wylosowanych. Opisywaną metodę charakteryzuje również inna cecha znana z wcześniej opisywanych metod losowania - preferuje elementy położone bliżej siebie (por. [20]).
Metoda ta została już przedstawiona w publikacji w Statistics in Transition [1], w pracy doktorskiej zostanie jednak uzupełniona o przykład ilustrujący tę metodę.

Metoda losowania przestrzennego oparta na krigingu

Przedstawiona zostanie autorska metoda losowania przestrzennego - metoda losowania przestrzennego oparta na krigingu. Jest ona w pewnym stopniu rozwinięciem trójkątnej metody losowania przestrzennego. W przeciwieństwie jednak do poprzedniego opisanej metody znajduje ona swoje zastosowania na populacjach skończonych, a nie na populacjach ciągłych.
W podrozdziale metoda opisana zostanie od strony teoretycznej jak i praktycznej. Metoda zostanie zilustrowana przykładem badania stopy bezrobocia z wykorzystaniem wykształcenia.

Przykład

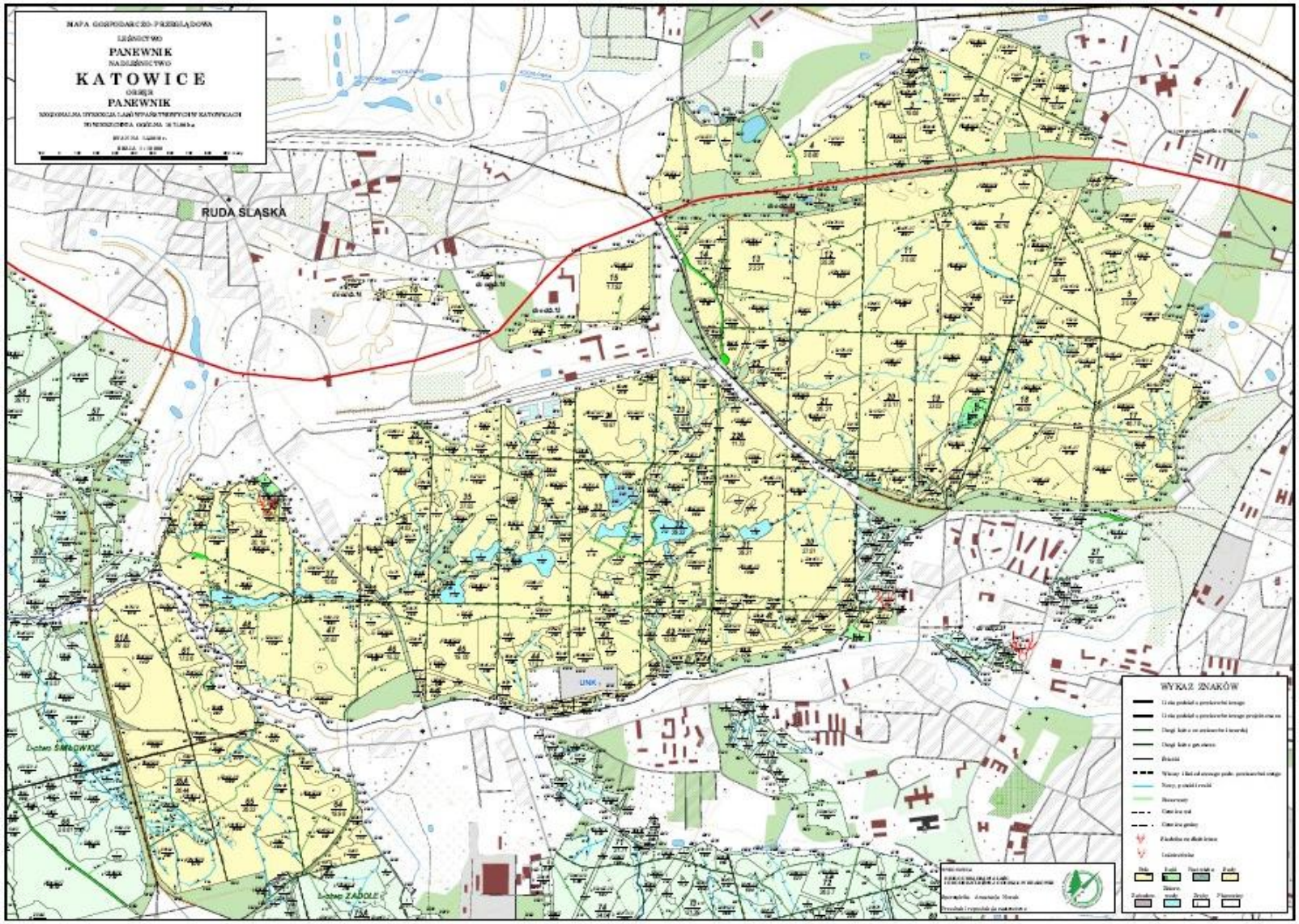
Walka z globalnym ociepleniem a potencjał lasów

Walka z globalnym ociepleniem stała się w ostatnich latach tematem dotykającym różnych sfer ludzkiego życia. Także sfery ekonomicznej. Widoczne jest to w dążeniu do ograniczenia zużycia emisjogennych paliw kopalnych (przy zachowaniu zasad zrównoważonego rozwoju) oraz w intensyfikacji zalesień. Na świecie dominują jednak wylesienia, nie zalesienia. Roczna powierzchnia wylesień na świecie wyniosła 13 mln ha w latach 2001-2010 [7]. Obowiązujący od 2005 Protokół z Kioto uwzględnił wpływ, jaki mają drzewa na pochłanianie dwutlenku węgla z atmosfery. Państwa uzyskały dzięki niemu prawo do wykorzystania w rozliczeniu swoich emisji gazów cieplarnianych pochłoniętych przez lasy, ale do pewnego limitu. Jednym ograniczeniem jest więc limit możliwego odliczenia pozytywnego wpływu lasów. Drugim jest to, że do rozliczeń można wykorzystać tylko gazy cieplarniane pochłonięte przez drzewa zasadzone po wejściu w życie Protokołu z Kioto.

Korygowanie podziału populacji na warstwy w badaniu obszaru leśnego Katowice-Panewniki

Rozważmy badanie, którego celem jest ocena zdolności obszaru leśnego do pochłaniania dwutlenku węgla. W takim badaniu obszar leśny może zostać podzielony na warstwy z wykorzystaniem mapy gospodarczo-przeglądowej lasu. Mapy te dostarczają dostarczają podziału obszaru leśnego opartego na dominujących gatunkach drzew, ich udziałom w całkowitym zalesieniu i średniego wieku gatunku dominującego. Oprócz tego mapy zawierają również informację o powierzchni każdego z podobszarów.
Do przeprowadzenia optymalizacji liczby warstw wykorzystano mapę gospodarczo-przeglądową Leśnictwa Katowice-Panewniki. Obszar leśny Katowice-Panewniki ma całkowitą powierzchnię równą 1071.96 ha. Mapa gospodarczo-przeglądowa dzieli go na 243 pododdziały zebrane w 47 działów. Jako gotowy do użycia podział populacji na warstwy wykorzystane zostaną 243 pododdziały Leśnictwa Katowice-Panewniki.

Rysunek: *Gospodarczo-przeglądowa mapa Leśnictwa Katowice-Panewniki.*



Wykorzystanie trójkątnej metody losowania przestrzennego w badaniu lasu

W poprzednim podrozdziale na przykładzie danych pochodzących z mapy gospodarczo-przeglądowej omówiono metodę optymalizującą liczbę warstw dla losowania warstwowego. W tym rozdziale przedstawiona zostanie metoda losowania, którą można wykorzystać do losowania próby wewnątrz warstwy w losowaniu warstwowym. Za przykład posłuży również badanie lasu w celu oszacowania jego zdolności do pochłaniania dwutlenku węgla

Jako zmienna pomocnicza wykorzystanie zostana pierśnica. Pierśnica jest jedną z podstawowych miar używanych do opisu drzew. W Europie, Australii i Kanadzie jest ona zdefiniowana jako średnica pnia na wysokości 130 cm nad ziemią.

Trójkątna metoda losowania przestrzennego została już opublikowana. Jednak w porównaniu do wyników z publikacji, w pracy doktorskiej zaprezentowany zostanie inny przykład, ze znacznie większą ilością powtórzeń użytą do estymacji metodą Monte Carlo.

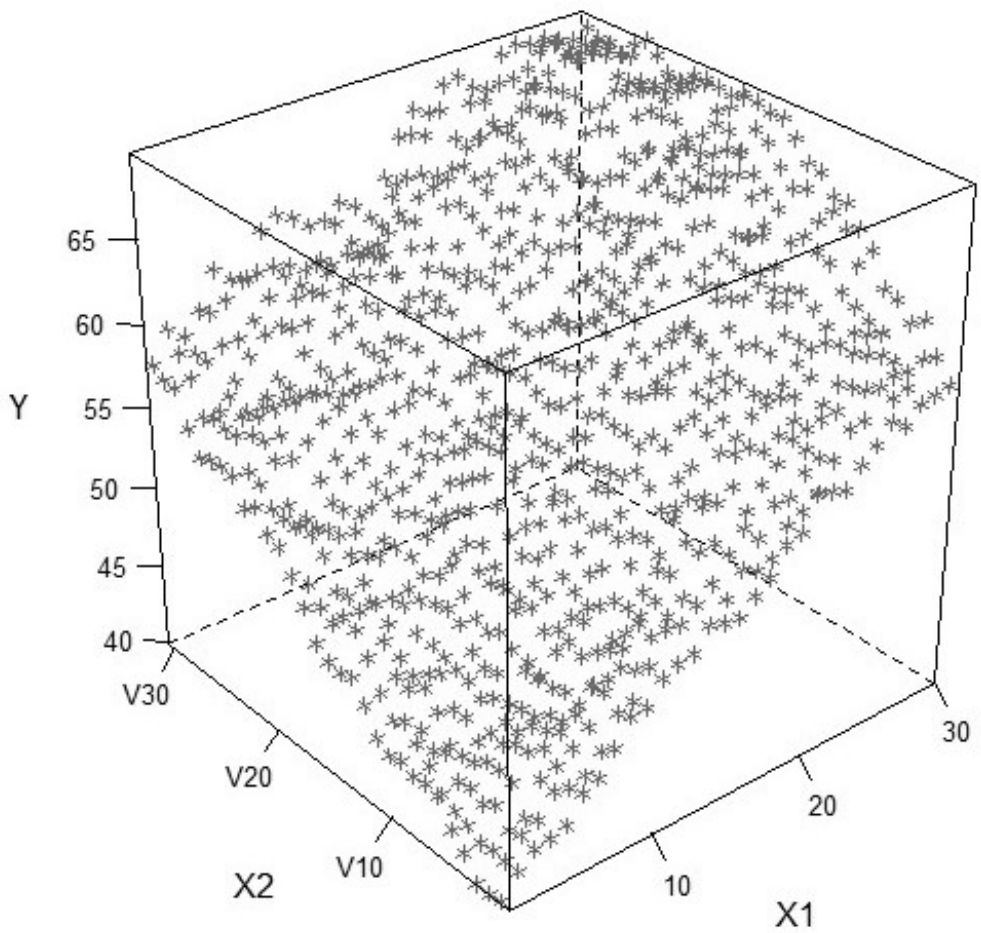
Fragment przykładu trójkątnej metody losowania przestrzennego

Wprowadzenie do przykładu

Rozważmy badanie mające na celu określenie zdolności lasu do pochłaniania dwutlenku węgla. Do oszacowania tej zdolności potrzebna jest znajomość masy drzew oraz składu gatunkowego lasu. Zmienną pomocniczą w badaniu jest pierśnica drzewa.
W przykładzie przestrzeń $[0,1] \times [0,1]$ utożsamiono z powierzchnią porośniętą przez las. Przestrzeń podzielono na 900 fragmentów. Symulacyjnie utworzono macierz 30×30 średnich wartości pierśnic, którą utożsamiono z przestrzenią $[0,1] \times [0,1]$. Wartości macierzy wygenerowano w oparciu o prosty przestrzenny model autoregresyjny:

$$[i,j] = \begin{cases} (0,4\,([i-1,j]+[i,j-1]) + 0,2[i-1,j-1])\,(1+0,025\varepsilon_{i,j})\,, \\ \text{gdy } i,\, j \in \{2,\ldots,30\}\,, \\ [i-1,j]\,(1+0,025\varepsilon_{i,j})\,, \text{ gdy } i \in \{2,\ldots,30\}\,,\ j=1, \\ [i,j-1]\,(1+0,025\varepsilon_{i,j})\,, \text{ gdy } i=1,\ j \in \{2,\ldots,30\}\,, \\ 40 \text{ cm, gdy } i=j=1, \end{cases} \tag{1}$$

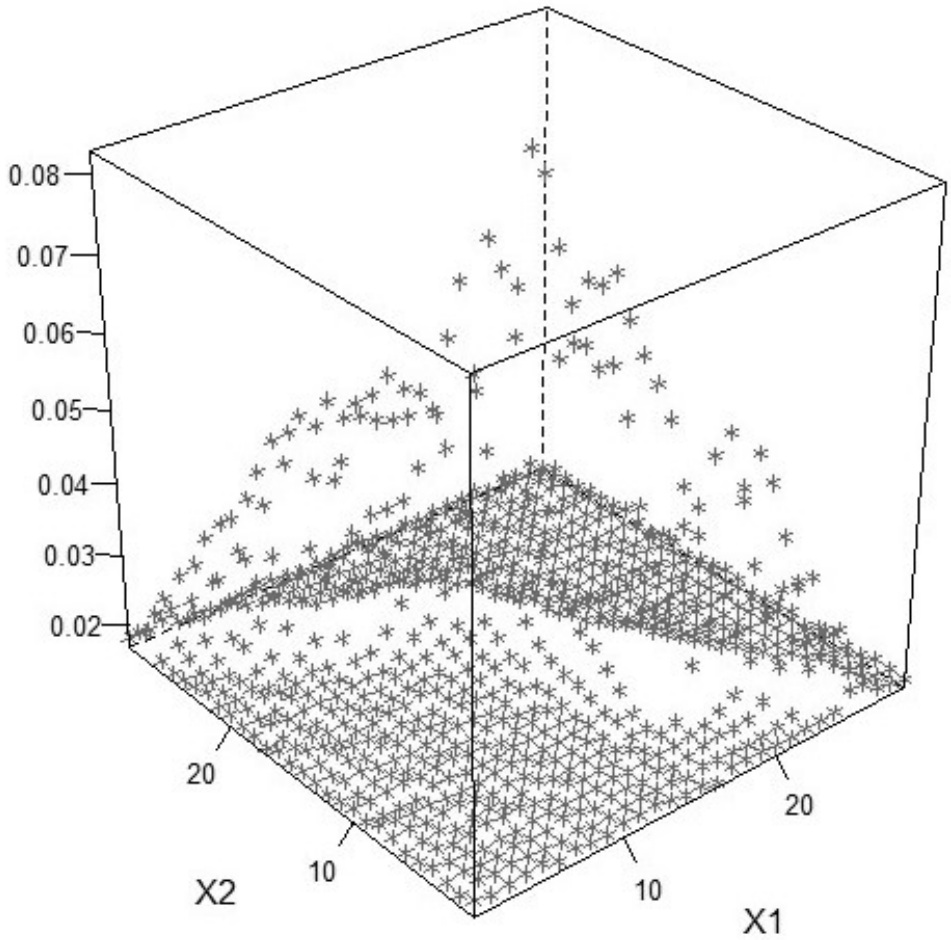
gdzie $\varepsilon_{i,j}$ $i,\, j \in \{1,\ldots,30\}$, ma rozkład jednostajny na odcinku $[-0,5;1,5]$. Wartość średnia wyniosła $57,049$ cm



Rysunek: *Wartości pierśnic uzyskane symulacyjnie*

Wyniki symulacji

- próba składała się z 20 elementów
- losowanie tej próby powtórzono 150.000 razy, a następnie metodą Monte Carlo wyznaczono prawdopodobieństwa inkluzji pierwszego rzędu.
- Aby ocenić efektywność tego schematu losowania, losowanie 20 elementowej próby powtórzono 10.000 razy. Dla każdego z losowań obliczono wartość modyfikacji estymatora Horvitza-Thompsona w postaci zaproponowanej przez Fattoriniego. Dla uzyskanych w ten sposób 10.000 wartości estymatorów wartość średnia wyniosła $57,049$ cm, zaś odchylenie standardowe z próby $1,254$ cm
- Również 10.000 razy przeprowadzono dobór próbą prostą bez zwracania 20 do estymacji. Dla każdej z prób wyznaczono wartość estymatora (średniej z próby). Średnia wartość dla estymatora z próby prostej wyniosła $57,095$, przy odchyleniu standardowym z próby równym $1,637$ cm.



Rysunek: *Funkcja gęstości prawdopodobieństwa uzyskana metodą Monte Carlo*

Bąk, T.: Triangular method of spatial sampling,

Statistics in Transition, Vol. 15, No. 1, 2014, pp.9-22.

Cressie, N. A. C.: Statistics for Spatial Data, New York:

John Wiley & Sons, Inc, 1993.

Dalenius, T.: Sampling in Sweden. Contributions to

methods and theories of sample survey practice, Almvist & Wiksells, Sztokholm, 1957.

Deville, J.-C. and Tillé, Y.: Efficient Balanced Sampling:

The Cube Method, Biometrika, 91(4), 2004,pp.893-912.

Dickson, M. M. and Benedetti, R. and Guliani, D. and

Espa, G.: The Use of Spatial Sampling Designs in Business Surveys, Open Journal of Statistics, 4, 2014,pp.345-354.

Eskelson, B.N.I. and Anderson, P.D. and Hagar, J.C. and

Temesgen, H.: Geostatistical modeling of riparian forest microclimate and its implications for sampling, Canadian Journal of Forest Research, Vol. 41(5), 2011, pp.974-985.

Gaj, K.: Pochłanianie CO_2 przez polskie ekosystemy

leśne, Leśne Prace Badawcze, Vol.73(1), 2012, pp. 17-21.

Grafström, A. and Lundström, N. L. P. and Schelin, L.:

Spatially Balanced Sampling through the Pivotal Method, Biometrics, 68, 2012, pp. 514-520.

Grafström, A. and Tillé, Y.: Doubly Spatial Sampling

with Spreading and Restitution of Auxiliary Totals, Environmetrics, 24, 2013, pp. 120-131.

van Groenigen, J. W. and Stein, A.: Constrained

optimization of spatial sampling using continuous simulated annealing, J. Environ. Qual., 27, 1998, pp. 1078-1086.

Kozak, M.: Optimal Stratification Using Random Search

Method in Agricultural Surveys,Statistics in Transition, Vol. 6, No. 5, 2004, pp.797-806.

Kirkpatrick, S. and Gelatt, C. D. and Vecchi, M. P.:

Optimization by Simulated Annealing, Science, New Series, Vol. 220, No. 4598, 1983, pp. 671-680.

Krige, D. G.: A statistical approach to some basic mine

valuation problems on the Witwatersrand, J. of the Chem., Metal. and Mining Soc. of South Africa 52 (6), 1951,pp.119-139.

Paelinck, J.H.P. and Klaassen, L.H.: Spatial

Econometrics, Saxon House Farnborough, 1979.

Przybycin, Z.: Metody i modele statystyki przestrzennej,

Wydawnictwo Akademii Ekonomicznej im Karola Adamieckiego w Katowicach, 2004.

Stevens Jr, D. L. and Olsen, A. R.: Spatially Balanced

Sampling of Natural Resources, Journal of the American Statistical Association, 99, 2004, pp. 262-278.

Suchecki, B. (red.): Ekonometria przestrzenna. Metody i

modele analizy danych przestrzennych, Wydawnictwo C. H. Beck, 2010.

Thompson, S.K.: Sampling, Wiley. New York, USA, 1992.

Thompson, S.K.: Adaptive Web Sampling, Biometrics,

62, 2006, pp. 1224-1234.

Wywiał, J. L.:On space sampling, Statistics in

Transition, Vol.2, Nr 7, 1996, pp. 1185-1191.

Wywiał, J. L.: Some contributions to multivariate

methods in survey sampling, Wydawnictwo Akademii Ekonomicznej im Karola Adamieckiego w Katowicach, 2003.

Zawadzki, J.:Symulowane wyżarzanie przestrzenne

efektywnym narzędziem planowania sieci pomiarowych, Studies & Proceedings of Polish Association for Knowledge Management, Nr 40, 2011, pp. 356-365.

Zubrzycki, S.: O szacowaniu parametrów złóż

geologicznych,Zastosowania Matematyki, 3(2), 1957, pp. 105-153.

Zubrzycki, S.:Remarks on random stratified and

systematic sampling in a plane, Colloquium Mathematicae 6, 1958, pp. 251-264.

Żądło, T.:On the Prediction of the Subpopulation Total

Based on Spatially Correlated Longitudinal Data, Mathematical Population Studies, 21, 1, 2013, pp. 30-44.

Dziękuję za uwagę