Metody losowania prób przestrzennych w badaniach ekonomicznych

Tomasz Bąk

Uniwersytet Ekonomiczny w Katowicach, Katedra Statystyki, Ekonometrii i Matematyki

17 października 2018

Cele pracy

- Cel główny: przekrojowe przedstawienie metod doboru próby przestrzennej wykorzystywanych w naukach ekonomicznych, ze szczególnym uwzględnieniem wyników własnych.
- Cel dodatkowy: omówienie dwóch autorskich metod doboru próby z przestrzeni, które w trakcie losowania wykorzystują informacje zawarte już w próbie: trójkątnej metody losowania przestrzennego oraz metody losowania przestrzennego opartej na krigingu.
- Cel dodatkowy: praktyczne wykorzystanie dwóch autorskich metod: optymalizacji gotowego podziału na warstwy w przypadku alokacji Neymana oraz wspomnianej już wcześniej trójkątnej metody losowania przestrzennego.

Plan prezentacji

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

Wprowadzenie

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 ([14]) oraz S. Zubrzyckiego ([26, 27]). W literaturze polskiej pierwsze zwarte opracowanie, które ujęło metody statystyki przestrzennej w zakresie umożliwiającym aplikację przedstawił Przybycin [16].
Na przecięciu statystyki przestrzennej i ekonometrii powstała ekonometria przestrzenna. Prekursorską pracą dla tej dziedziny nauki była książka *Spatial Econometrics* napisana przez Paelincka i Klaassena [15]. 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 [18]. Było to pierwsze w Polsce opracowanie, które w szerokim zakresie omawia nowoczesne metody i modele ekonometrii przestrzennej.

Definicja

Proces przestrzenny to pole losowe (proces stochastyczny)

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

gdzie D jest ustalonym podzbiorem przestrzeni $\mathbb{R}^k$ (por. np. [3]).

Charakterystyka

Proces przestrzenny wyróżnia od innych zmiennych rozważanych w statystyce to, że posiada lokalizację. Często też jego wartości układają się w struktury, co objawia się autokorelacją przestrzenną oraz przestrzenną heterogenicznością.
Procesowi przestrzennemu można nadać jeszcze dodatkowy wymiar, poprzez jego obserwacje w kilku różnych momentach czasu. Badania takiej zmiennej nazywa się badaniami wielookresowymi. Analizą danym pochodzących z takich badań zajmował się m.in Żądło [28].
Przykładami procesów przestrzennych są średnie dochody na mieszkańca gospodarstwa domowego, intensywność opadów, czy też ceny gruntów.

Podejście randomizacyjne

Podstawowe metody doboru próby, znane ze statystyki ‘nieprzestrzennej‘

- Dobór losowy: próba prosta, losowanie systematyczne, losowanie warstwowe, losowanie dwustopniowe, losowanie grupowe.
- Dobór celowy

Uwzględnienie przestrzennej autokorelacji i heterogeniczności

Wymienione powyżej metody statystyki ’nieprzestrzennej’ można stosować w statystyce przestrzennej. Jednak autokorelacja i heterogeniczność procesu przestrzennego istotnie zmienia ocenę efektywności tych metod.

Losowania przestrzenne - dobór próby z siatki

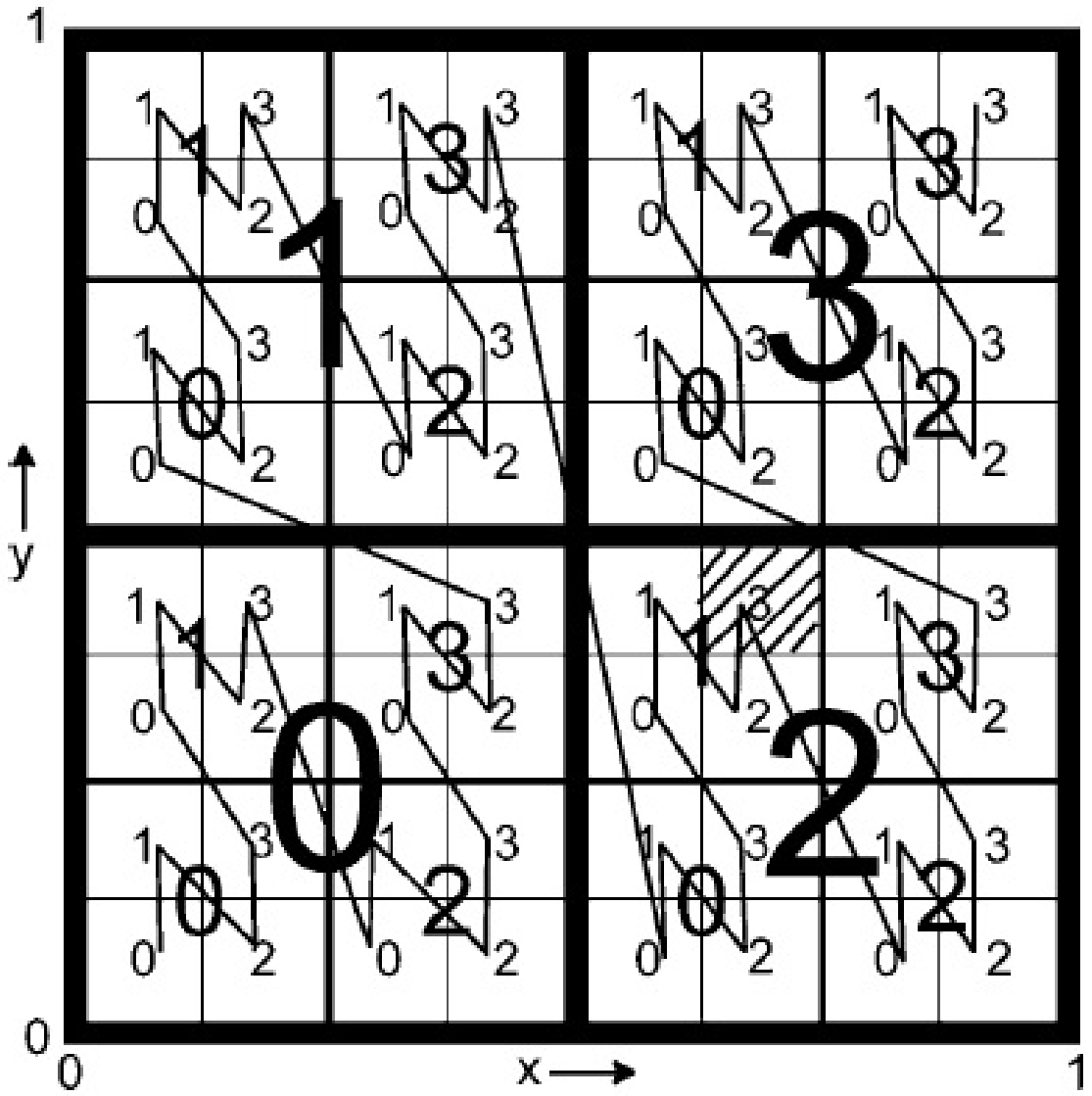
W statystyce przestrzennej często na badaną populację nakłada się siatkę wielokątów, a następnie losuje się elementy siatki (wielokąty). Dodatkowo zwiększa się odległości pomiędzy elementami w próbie, aby zwiększyć efektywność losowania. Powstaje wtedy próba zrównoważona przestrzennie (*Spatially balanced sample*). W szczególności warto wyróżnić:

- Losowanie z wykorzystaniem macierzy sąsiedztwa (Wywiał [23])
- Metoda kostki (Cube method) (Deville, Tillé [5])
- Metoda GRTS (Stevens, Olsen, [17])
- Metoda lokalnych kluczy (Local pivotal method) (Grafström, Lundström, Schelin [9])
- Metoda podwójnie zrównoważonego losowania przestrzennego (Doubly balanced spatial sampling)(Grafström,Tillé [10])

Metoda GRTS

Metoda GRTS (*generalized random-tessellation stratified*) została zaproponowana przez autorów w związku z zapotrzebowaniem na efektywny schemat doboru próby w badaniach takich zjawisk jak globalne ocieplenie, transport długodystansowy czy też zanieczyszczenia atmosfery. Ideą metody GRTS jest transformacja przestrzeni 2-wymiarowej na 1-wymiarową (odcinek), a następnie wylosowanie próby systematycznej na nowej przestrzeni. W tym celu obszar objęty badaniem wpisuje się w kwadrat o boku długości 1. Kwadrat ten dzieli się na 4 równe kwadraty, nadając każdemu z nich numer porządkowy od 0 do 3. W kolejnych krokach dokonuje się dalszego podziału kwadratów, numerując je zgodnie z ustalonym na początku porządkiem.
Po zakończonym podziale do każdego z najmniejszych kwadratów dopasowuje się unikalny adres (numer), którego k-ta cyfra odpowiada k-temu pod względem wielkości kwadratowi zawierającemu kwadrat, któremu nadawany jest adres. Transformacja przestrzeni 2-wymiarowej na 1-wymiarową odbywa się z uwzględnieniem porządku na zbiorze adresów.

Rysunek: *Przykład doboru próby metodą GRTS.*

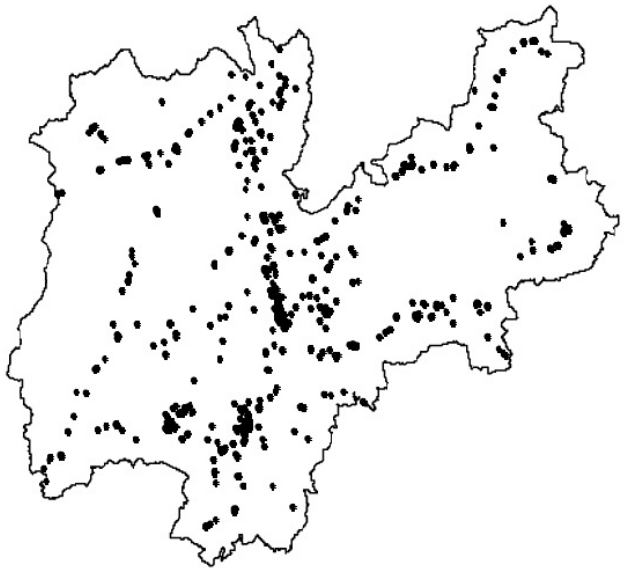


Przykład zastosowania

Coraz więcej krajowych rejestrów firm zawiera dokładne informacje o długości i szerokości geograficznej na której znajduje się firma. Tego rodzaju dane zawierają już rejestry firm w USA (LBD), Szwajcarii (STATENT) oraz Włoszech (ASIA). W Polsce Centralna Ewidencja i Informacja o Działalności Gospodarczej nie zawiera jeszcze tego typu danych [a].
Dickson z zespołem [6] objęli badaniem populację 822 niewyspecjalizowanych sklepów detalicznych działających w prowincji Trentino w 2009 roku. Celem losowania była estymacja globalnej wartości sprzedaży. Do wylosowania próby użyto metod lokalnych kluczy, metody kostki oraz metody podwójnie zrównoważonego losowania przestrzennego.

[a] Stan na 2018.09.22.

Rysunek: *Rozkład przestrzenny populacji sklepów.*



Podejście modelowe

Wariogram

Załóżmy, że

$$D^2\left(Y(x_1) - Y(x_2)\right) = 2\gamma(|x_1 - x_2|), \quad x_1, x_2 \in D, \tag{1}$$

gdzie Y jest procesem przestrzennym określonym na przestrzeni D. Funkcję $2\gamma(\cdot)$ nazywa się wariogramem, zaś $\gamma(\cdot)$ semiwariogramem.

Kriging

Kriging, podobnie jak regresja jest pojęciem generycznym. Możemy wyróżnić wiele wariantów krigingu, jak np. kriging zwyczajny (*oridinary kriging*) lub kriging uniwersalny (*universal kriging*). Matheron nazwał tę metodę optymalnej, liniowej, przestrzennej predykcji od nazwiska D. G. Krige'a, który w latach 50-tych ubiegłego wieku rozwinął metodę określania rozkładu minerałów w oparciu o próbę. Kriging jest metodą predykcji wartości procesu przestrzennego o minimalnym błędzie średniokwadratowym predyktora, która najczęściej zakłada stacjonarność drugiego rzędu modelowanego procesu przestrzennego [3, 21].

Optymalizacja w oparciu o kriging

Rozważmy wariancję estymatora krigingu zwyczajnego:

$$\sigma^2_{KZ}(x_0) = \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_0) - m, \tag{2}$$

gdzie γ jest funkcją semiwariogramu (1), a m jest mnożnikiem Lagrange'a. Na wariancję krigingu zwyczajnego (2) wpływa wyłącznie postać semiwariogramu oraz odległości punktów $x_1, \ldots, x_n$ od punktu x_0. Jeżeli więc postać wariogramu może zostać określona przed losowaniem (na podstawie wcześniejszych badań lub próby wstępnej), to wariancję (2) można wyliczyć przed przeprowadzeniem losowania. Na badany obszar D nakłada się siatkę i definiuje się 'średnią' wariancję krigingu na elementach siatki jako:

$$\phi(E) = \frac{1}{n_e} \sum_{i=1}^{n_e} \sigma^2_{KZ}(x_{e,i}|E), \tag{3}$$

gdzie E jest schematem losowania, zaś $x_{e,i}$ oznacza i-ty element n_e-elementowej siatki.

Optymalizacja 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.

Zawadzki [25] rozważał użycie dwóch kryteriów iteracyjnie zwiększających precyzję wariogramu: minimalizację średniej odległości od najbliższego sąsiada (MMSD) oraz kryterium Warricka-Myers'a.

Metody MMSD i WMSD

Generalnie funkcja celu w kryterium MMSD przyjmuje następującą postać

$$\phi(E) = \frac{1}{n_e} \sum_{i=1}^{n_e} \|x_e^i - V(E, x_e^i)\|, \tag{4}$$

gdzie x_e^i oznacza i-ty spośród n_e punktów kontrolnych, a $V(E, x_e^i)$ jego najbliższego sąsiada wśród elementów wybranych schematem losowania E. Zauważmy, że jeżeli liczba arbitralnie wybranych elementów byłaby mniejsza niż wielkość próby, to kryterium będzie dążyło do ulokowania próby w arbitralnych punktach (lub ich najbliższym sąsiedztwie). Doprowadzi to do niepożądanego zgrupowania elementów próby. Stąd oczekiwane jest, aby $n_e >> n$. Rozwinięciem metody MMSD jest kryterium ważonych średnich odległości do najbliższego sąsiada (WMSD):

$$\phi(E) = \frac{1}{n_e} \sum_{i=1}^{n_e} w(x_e^i) \|x_e^i - V(E, x_e^i)\|, \tag{5}$$

gdzie $w(x_e^i)$ jest wagą przypisaną do i-tego elementu kontrolnego.

Przykład wykorzystania metody MMSD i WMSD

Rozważmy badanie hurtowni pod kątem roli kapitału własnego w kreowaniu wartości firmy. Dla hurtowni farmaceutycznych takie badanie przeprowadziła Witczak [22]. Próbę dla tak określonego badania można dobrać metodą (MMSD). Arbitralnie wybranymi punktami kontrolnymi będą lokalizacje sklepów, które pozyskują towary z hurtowni. Minimalizowana będzie średnia odległość hurtowni wybranych do próby od tychże sklepów. Zauważmy, że liczba arbitralnie wybranych sklepów będzie znacząco większa niż liczba hurtowni w próbie, więc kryterium powinno dostarczyć zadowalających wyników ($n_e >> n$).
Dobór próby można przeprowadzić również z wykorzystaniem kryterium (WMSD). Elementami, względem których minimalizowana będzie średnia ważona odległość elementów z próby będzie ponownie sieć odbiorców towarów (sklepów). Wagi wykorzystywane w kryterium można zdefiniować jako wartości zakupionych towarów w hurtowniach objętych badaniem. Minimalizacja odległości od dużych sklepów będzie istotniejsza niż minimalizacja odległości od sklepów małych. Większe szanse na znalezienie się w próbie badawczej będą miały hurtownie położone blisko dużych odbiorców towarów.

Adaptacyjne oraz pokrewne metody losowania

Wprowadzenie

Pojęcie losowania adaptacyjnego wywodzi się z prac Thompsona [19, 20]. Początkowo dedykowane było badaniom cech rzadkich. Odnosi się ono do schematu losowania, w którym prawdopodobieństwa, z jakimi wybierane są kolejne elementy zależą od wartości obserwowanej cechy u elementów już wylosowanych. Pod pojęciem losowania adaptacyjnego kryje się wiele różnych schematów losowania, z których najbardziej pierwotnym jest adaptacyjne losowanie grupowe (*adaptive cluster sampling*). Kluczowym pojęciem dla tego losowania jest klaster. Klaster i-tego elementu posiadającego cechę rzadką powstaje poprzez dołączenie do tego elementu wszystkich jego sąsiadów, a następnie stopniowe dodawanie sąsiadów dla elementów mających poszukiwaną cechę i znajdujących się już w klastrze. Dla pozostałych elementów klaster jest równoznaczny danemu elementowi.

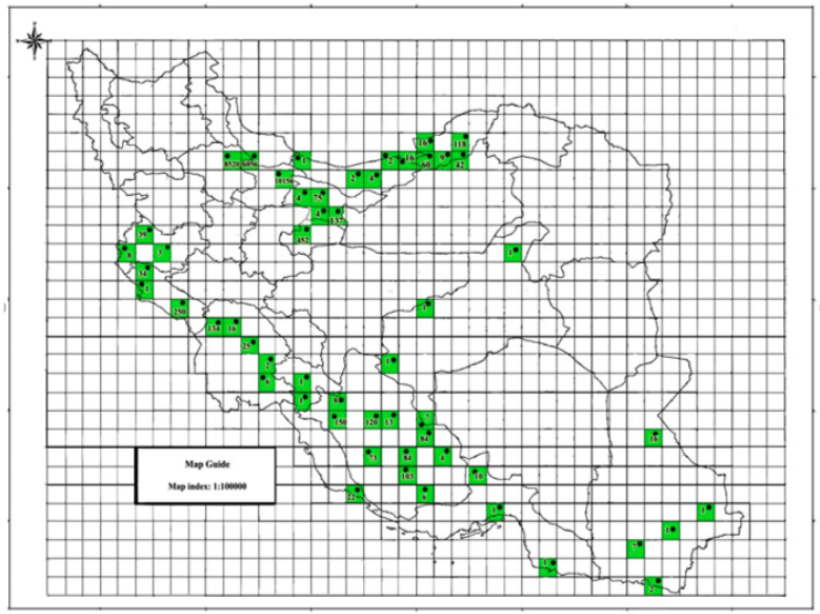
Adaptacyjne losowanie grupowe

Adaptacyjne losowanie grupowe składa się z dwóch etapów. W pierwszym etapie wybrana zostaje próba początkowa składająca się z n_0 elementów. Generalnie próbę początkową wybiera się losowaniem prostym bez zwracania lub ze zwracaniem [19]. W drugim etapie losowania do próby dodaje się wszystkie elementy klastrów, które zawierają elementy próby początkowej. Uzyskaną w ten sposób próbę nazywa się próbą adaptacyjną. W przypadku adaptacyjnego losowania grupowego prawdopodobieństwa inkluzji pierwszego rzędu są różne dla różnych elementów populacji.

Przykład zastosowania

Afshar i Navvabpour [1] wykorzystali adaptacyjne losowanie sieciowe do estymacji produkcji oliwek na terenie Iranu. Miasta zajmujące się produkcją oliwek stanowią stosunkowo mały odsetek wszystkich miast Iranu, stąd wybór tej metody losowania. Obszar kraju podzielono siatką kwadratów na 684 jednostki, wśród których 54 zawierały miasta produkujące oliwki.
Z tak utworzonej populacji wybrano po 1000 prób z wykorzystaniem adaptacyjnego losowania sieciowego oraz innych metod doboru prób przestrzennych. Adaptacyjne losowanie sieciowe dostarczyło najbardziej precyzyjnych wyników (pod kątem RMSE).

Rysunek: *Podział Iranu siatką kwadratów na 684 jednostki.*



Trójkątna metoda losowania przestrzennego

Adaptacyjne metody losowania dają badaczowi pewną kontrolę nad elementami wchodzącymi w skład próby. Taką możliwość daje modyfikowanie 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. Cecha ta charakteryzuje również trójkątną metodę losowania przestrzennego [2].
W przypadku tej metody na prawdopodobieństwa doboru próby wpływają wartości zmiennej dodatkowej a badana populacja jest populacją ciągłą. Tak więc metody adaptacyjne można również przenieść na populacje ciągłe (nieskończone).

Badanie wyceny wpływu lasu na redukcję emisji dwutlenku węgla

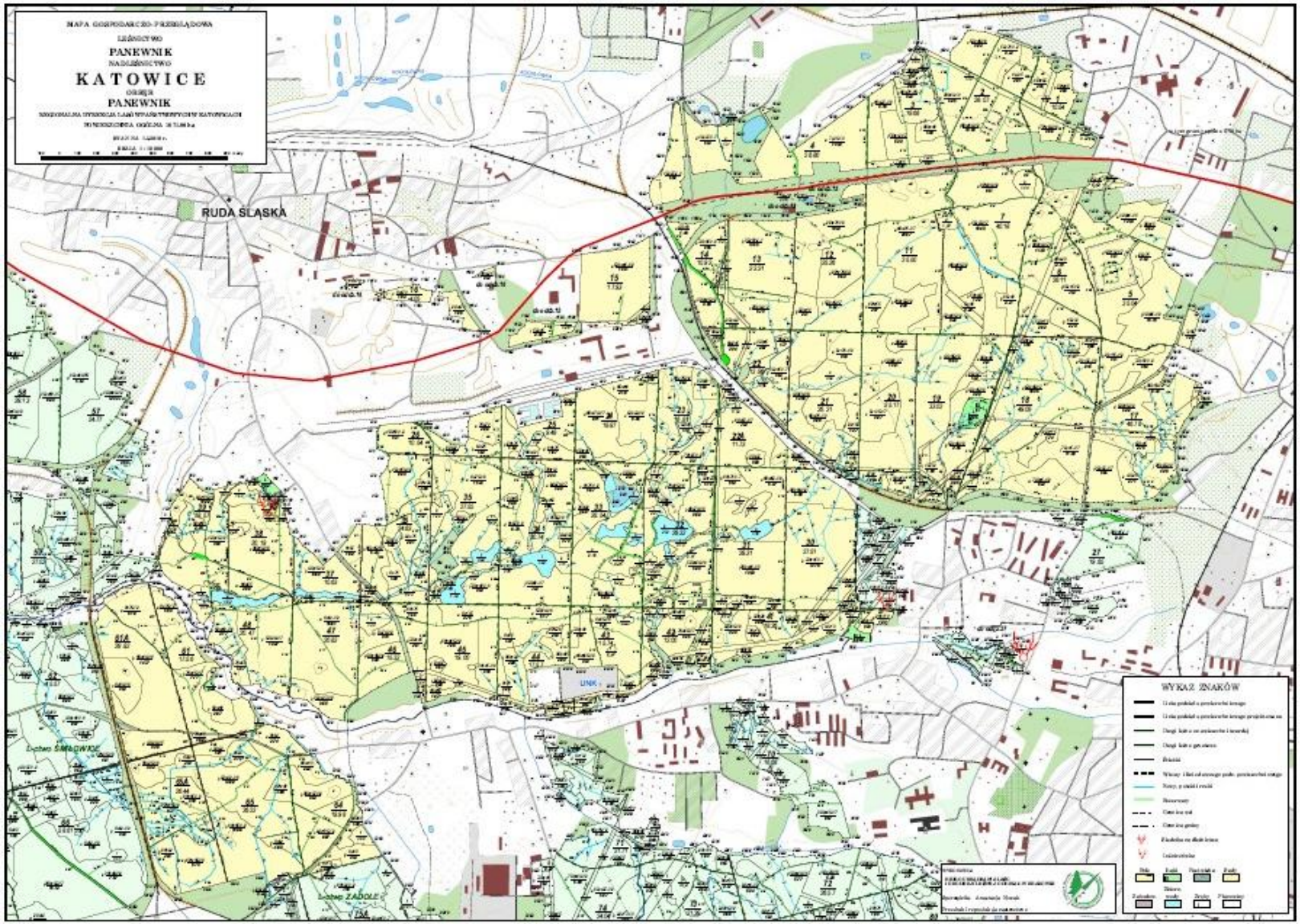
Wprowadzenie

Walka z globalnym ociepleniem stała się w ostatnich latach tematem dotykającym różnych sfer ludzkiego życia. Także sfery ekonomicznej. Obowiązujący od 2005 Protokół z Kioto uwzględnił wpływ, jaki mają drzewa na pochłanianie dwutlenku węgla z atmosfery. Państwa uzyskały dzięki niemu prawo do uwzględniania gazów cieplarnianych pochłoniętych przez lasy w rozliczaniu całkowitej ich emisji.
W grudniu 2015 roku w Paryżu podczas konferencji COP 21 podpisano globalne porozumienie klimatyczne. Jednym z rezultatów tego porozumienia jest włączenie programu REDD+, przyjętego w Warszawie na konferencji COP 19, w ramy globalnego porozumienia klimatycznego. Celem programu REDD+ jest powstrzymanie wycinki lasów tropikalnych.

Wyniki empiryczne

W kontekście przedstawionej problematyki zaproponowano zastosowanie 2 autorskich metod: optymalizację liczby warstw dla optymalnej alokacji Neymana oraz trójkątną metodę losowania przestrzennego. Optymalizację przeprowadzono na danych rzeczywistych zawartych na mapie gospodarczo-przeglądowej Leśnictwa Katowice-Panewniki. Jej zastosowanie pozwoliło zmniejszyć liczbę warstw z 243 to 201 bez straty na efektywności losowania. Trójkątną metodę losowania przestrzennego przetestowano na danych symulujących obszar leśny. W symulacjach skupiono się na rozkładzie pierśnic. Wyniki przeprowadzonego losowania pozwoliły ocenić trójkątną metodę losowania przestrzennego jako obiecujące rozwiązanie problemu doboru próby przestrzennej.
Analizę zdolności lasów do pochłaniania $C0_2$ należy rozszerzyć. Potrzebne są tutaj wieloetapowe badania, zaczynające się od odpowiedniego doboru próby, poprzez pomiar drzew, oszacowanie ich zdolności do pochłaniania dwutlenku węgla, na wycenie tej zdolności kończąc. Będzie to wymagać stworzenia interdyscyplinarnych zespołów, w których łączyć się będą kompetencje z zakresu m.in. ekonomii, biologii, ekologii i statystyki.

Rysunek: *Mapa gospodarczo-przeglądowa Leśnictwa Katowice-Panewniki.*



Afshar, A. and Navvabpour, H.: A Comparative Study of Performance of Adaptive Web Sampling and General Inverse Adaptive Sampling in Estimating Olive Production in Iran, Journal of Statistical Research of Iran 11 (1), 2014, pp.9-22.

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.

Wang, J.F. and Stein, A. and Gao, B.B. and Ge, Y.: A review of spatial sampling. Spatial Statistics 2, 2012,pp. 1-14.

Witczak, I.:Rola kapitału własnego w kreowaniu wartości firmy na przykładzie hurtowni farmaceutycznych. Studia Ekonomiczne. Zeszyty Naukowe, nr 245, 2015, pp.215-224.

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