

STATYSTYKA MAŁYCH OBSZARÓW – TEORIA I PRAKTYKA

Anna Małasiewicz

Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu

Streszczenie

Działalność władz lokalnych, różnych instytucji i podmiotów gospodarczych uzależniona jest od posiadania kompleksowej informacji z różnych obszarów i dziedzin życia. Powoduje to, że z jednej strony niezbędna staje się informacja dostępna na niskim poziomie agregacji przestrzennej (np. gmin), czy też dla bardziej szczegółowych domen, z drugiej zaś strony, pozyskanie tego typu informacji uniemożliwiają rosnące koszty badań, ograniczenia czasowe, czy obciążenia respondentów. Te pozornie sprzeczne cele, jakimi są niskie koszty badań przy jednoczesnym zapewnieniu wysokiej jakości informacji, mogą być osiągnięte dzięki zastosowaniu najnowszych osiągnięć metodologicznych z zakresu tzw. statystyki małych obszarów. Metody te odgrywają wspólnie istotną rolę w kształtowaniu technik pozyskiwania informacji, a dzięki swoim własnościom umożliwiają uzyskanie wiarygodnych szacunków na niższych poziomach agregacji i bardziej szczegółowych domen, w sytuacji niewielkiej liczebności próby lub nawet braku obserwacji w próbie dla danego przekroju. Ich znaczenie nieustannie wzrasta, co przy postępującej informatyzacji, gromadzeniu coraz większych wolumenów danych z różnych źródeł (spisy, rejestry administracyjne czy badania reprezentacyjne) daje szerokie spektrum zastosowań tej gałęzi statystyki.

Wprowadzenie

W przypadku gdy liczebność próby w badaniu reprezentacyjnym, w odpowiednio zdefiniowanych domenach, jest duża można wykorzystać estymatory bezpośrednie. Problem pojawia się w sytuacji, gdy domena jest reprezentowana w próbie przez niewielką liczbę jednostek. W takich przypadkach stosowanie estymatora bezpośredniego jest obarczone zbyt dużym ryzykiem co związane jest zazwyczaj z jego większą wariancją. W konsekwencji, nie chcąc zwiększać kosztów badania poprzez zwiększanie próby, można rozważyć zastosowanie technik estymacji pośredniej. Ich wykorzystanie może zapewnić zmniejszenie wariancji stosowanych estymatorów poprzez wykorzystanie informacji z innych domen i danych z innych źródeł, na przykład rejestrów administracyjnych czy spisów. W zależności od przyjętego sposobu postępowania, wnioskowanie statystyczne może bazować jedynie na schemacie losowania (design-based), być wspomagane modelem (model-assisted) lub też być całkowicie oparte na modelu (model-based).

- **Wnioskowanie oparte na schemacie losowania (design-based)** – całkowicie opiera się na prawdopodobieństwach dostania się jednostek do próby zgodnie z określonym schematem losowania; do estymatorów należących do tej grupy można zaliczyć m.in. estymator bezpośredni, który stanowi punkt odniesienia dla innych estymatorów z zakresu statystyki małych obszarów,
- **Wnioskowanie wspomagane modelem (model-assisted)**– w ramach tej grupy estymatorów wyróżnia się te, które nie tylko wykorzystują informacje o prawdopodobieństwach inkluzji ale również bazują na modelu; w grupie tej znajdują się m.in. estymator regresyjny i estymator złożony,
- **Wnioskowanie oparte na modelu (model-based)** – w ramach tej grupy estymatorów wyróżnia się te, które wykorzystują na szeroką skalę odpowiednio skonstruowany model. Można zaliczyć tutaj przede wszystkim estymatory bazujące na modelu Fay’a - Herriota i modelu Poissona, empiryczne estymatory bayesowskie, hierarchiczne estymatory bayesowskie.

Estymatory statystyki małych obszarów

Poniżej znajdują się podstawowe estymatory używane w statystyce małych obszarów.

Estymator	Postać
	Brak informacji pomocniczych
Horvitz-Thompsona	$\hat{y}_i^{HT} = \sum_{j \in s} w_j y_j$, gdzie $w_j = (\pi_j)^{-1}$, a π_j to prawdopodobieństwo dostania się do próby j -tej jednostki. Estymator Horvitz-Thompsona (bezpośredni) – jest nieobciążony, – może charakteryzować się dużą wariancją.
	Wykorzystanie informacji pomocniczych
Regresyjny (syntetyczny)	$\hat{y}_i^{reg} = X_i^T \hat{B}$, gdzie \hat{B} jest rozwiązaniem równania $(\sum_{j \in s} w_j \mathbf{x}_j \mathbf{x}_j^T) \hat{B} = \sum_{j \in s} w_j \mathbf{x}_j y_j$.
Ilorazowy (syntetyczny)	$\hat{y}_i^{ilo} = X_i \frac{\hat{Y}}{\bar{X}}$, Obciążenie będzie małe, gdy iloraz $\frac{\hat{Y}}{\bar{X}}$ obliczony jedynie dla domeny, będzie bliski ilorazowi populacyjnemu.
Złożony	$\hat{y}_i^{comp} = \gamma_i \hat{y}_i^B + (1 - \gamma_i) \hat{y}_i^S$, Estymator ten stanowi kombinację liniową estymatora bezpośredniego i syntetycznego. Balansuje potencjalne obciążenie estymatora syntetycznego \hat{y}_i^S wobec dużej wariancji estymatora bezpośredniego \hat{y}_i^B . Waga γ_i może być uzależniona od zmienności cechy dodatkowej, zależna od liczebności populacji w domenach lub wielkości próby. Może również zależeć ona od średniego błędu kwadratowego estymatorów składowych.
	Podejście modelowe
model Fay’a-Herriota (I)	gdy dostępne są dane na poziomie jednostki: $y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$ wtedy estymatorem opartym na modelu będzie średnia ważona estymatora „regresji badania” (Rao (2003)) i syntetycznego regresyjnego: $\hat{\mu}_i = \hat{\gamma}_i (\bar{y}_{ia} + (\bar{X}_i - \bar{x}_{ia})^T \hat{\beta}) + (1 - \hat{\gamma}_i) \bar{X}_i^T \hat{\beta}$.
model Fay’a-Herriota (II)	gdy dostępne są dane na poziomie obszaru: $y_i = x_i^T \beta + v_i + e_i$ wtedy estymatorem opartym na modelu będzie średnia ważona estymatora bezpośredniego i syntetycznego regresyjnego: $\hat{\mu}_i = \hat{\gamma}_i \hat{y}_i + (1 - \hat{\gamma}_i) \bar{X}_i^T \hat{\beta}$.
model Poissona	Przyjęte jest założenie, że liczba osób z pewną cechą N_i podlega rozkładowi Poissona z parametrem μ_i oraz $g(\mu_i) = x_i^T \beta + v_i$, gdzie $g(\cdot)$ jest odpowiednio dobraną funkcją. Jeśli $g(\cdot) = \log(\cdot)$, wtedy estymatorem opartym na modelu będzie: $\hat{N}_i = \exp(\bar{X}_i^T \hat{\beta} + \hat{v}_i)$ dla domen, które są reprezentowane w próbie, $\hat{N}_i = \exp(\bar{X}_i^T \hat{\beta})$ dla domen, które nie są reprezentowane w próbie.

Statystyka małych obszarów zaczyna odgrywać coraz większą rolę w działalności krajowych urzędów statystycznych i wielu instytucji międzynarodowych. Przykładem może być Bank Światowy, który wykorzystuje metodę estymacji pośredniej do mapowania ubóstwa na niskich poziomach agregacji przestrzennej w wielu krajach. Stosuje w tym celu estymatory bazujące na modelu Fay’a-Herriota - w zależności od stopnia dostępności danych na poziomie jednostkowym lub obszaru.

Wyrazem coraz większego znaczenia statystyki małych obszarów są liczne konferencje, publikacje oraz projekty o charakterze międzynarodowym, w których rozwija się teorię ale również stosuje na szeroką skalę w praktyce różnego rodzaju estymatory. Poniżej opisane są najważniejsze zastosowania estymacji pośredniej oraz opis wybranych międzynarodowych projektów dedykowanych tej dziedzinie wiedzy.

Główne obszary zastosowań statystyki małych obszarów

Statystyka małych obszarów może być stosowana w wielu dziedzinach życia. Najczęściej wykorzystywana jest w Stanach Zjednoczonych i Kanadzie. W coraz większym stopniu wykorzystuje się ją również w polskiej praktyce badawczej. Ze względu na dostarczanie aktualnych i szczegółowych informacji, statystyka małych obszarów z powodzeniem stosowana jest w badaniach odnoszących się do:

- **ochrony zdrowia**– badania dotyczące m.in. niezdolności psychicznej ludności w krótkich okresach, liczby urodzeń żywych i martwych, informacji o niepełnosprawnych, zachorowalności na raka,
- **demografii i statystyki ludności**– badania dotyczące m.in. liczby ludności i liczby mieszkań w różnych przekrojach,
- **rynku pracy i bezrobocia**– badania dotyczące m.in. wskaźnika stopy bezrobocia, liczby bezrobotnych, aktywnych ekonomicznie,
- **badajń rynkowych**– badania dotyczące m.in. dochodów osobistych, zatrudnienia w pełnym i w niepełnym wymiarze, dochodów rodzin,
- **rolnictwa**– badania związane m.in. z szacowaniem powierzchni upraw (z wykorzystaniem danych ze zdjęć satelitarnych), wielkości plonów, wartości hodowli i zbiorów.

Międzynarodowe projekty dotyczące statystyki małych obszarów

Poniżej opisane zostały najważniejsze projekty, w których rozwijano teorię statystyki małych obszarów. W projektach tych kładziono również duży nacisk na warstwę aplikacyjną estymacji pośredniej w wielu obszarach życia.

- **ESSnet for Small Area Estimation** – ogólnym celem projektu jest stworzenie ram umożliwiających tworzenie szacunków dla małych obszarów w ramach Europejskich Badań Społecznych (ESS), dotyczących m.in. demografii, bezrobocia, ubóstwa, warunków życia, zdrowia i edukacji. W ramach projektu powstały publikacje (WP2, WP3, WP4, WP6 i WP7), w których znaleźć można szczegółowe opisy zastosowań estymacji małych obszarów, wykorzystywane estymatory i rodzaj wykorzystywanego oprogramowania.

- **The EURAREA project** – program badawczy finansowany przez Eurostat w ramach Programu Fifth Framework Programme FP5 Unii Europejskiej. Głównym celem projektu było wypracowanie technik estymacji oraz opracowanie odpowiedniego oprogramowania wspierającego Europejski System Statystyczny w zakresie szacowania podstawowych charakterystyk rynku pracy oraz poziomu życia ludności w skali lokalnej.

- **SAMPLE – Small Area Methods for Poverty and Living Condition Estimates** – program badawczy finansowany przez Komisję Europejską w ramach Seventh Framework Programme (FP7). Celem projektu było opracowanie nowych wskaźników zapewniających lepsze zrozumienie nierówności społecznej i ubóstwa na lokalnym poziomie oraz metod ich estymacji z wykorzystaniem statystyki małych obszarów. Dodatkowym celem było wdrożenie procedur konstrukcji i interpretacji wskaźników ubóstwa oraz określenie stopnia ich przydatności dla lokalnych samorządów.

Wykorzystanie estymatorów statystyki małych obszarów w Polsce

Zainteresowanie statystyką małych obszarów w Polsce znacząco wzrosło w ostatnich latach, głównie ze względu na zwiększające się zapotrzebowanie na informacje na niskich poziomach agregacji przestrzennej. Obszarami, w których zastosowano m.in. metody estymacji pośredniej są:

- rolnictwo,
- rynek pracy,
- ubóstwo,
- demografia,
- statystyka przedsiębiorstw.

Wyrazem coraz większego znaczenia metod estymacji pośredniej było utworzenie Ośrodka Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu. Ośrodek ten prowadzi prace badawcze w zakresie teorii i zastosowań metod statystyki małych obszarów w różnych dziedzinach (rynek pracy, statystyka przedsiębiorstw, ubóstwo itp.).

Wnioski

- Statystyka małych obszarów jest prężnie rozwijającą się dziedziną badań nie tylko na świecie ale również i w Polsce,
- Określenia „mały obszar” nie należy rozumieć dosłownie – termin ten odnosić się może zarówno do domeny lub subpopulacji, dla których wielkość próby jest na tyle mała, że estymacja bezpośrednia może być nieefektywna,
- Użycie metod estymacji pośredniej zazwyczaj pozwala uzyskać porównywalną lub znacznie lepszą precyzję szacunków w stosunku do klasycznych metod estymacji.

Literatura

- [1] Van den Brakel J.A. Models in official statistics. 2012.
- [2] Dehnel G. Statystyka Małych Obszarów jako narzędzie oceny rozwoju ekonomicznego regionów. Wyd. Akademii Ekonomicznej w Poznaniu, 2003.
- [3] Rao J.N.K. Small Area Estimation. Wiley, 2003.