ASK, 2006, nr 15, strony 75–103

Marcin Kocór
Uniwersytet Jagielloński

METODY WNIOSKOWANIA EKOLOGICZNEGO W BADANIACH WYBORCZYCH

W naukach społecznych często mamy do czynienia z sytuacją, kiedy dane dotyczące poziomu jednostek są niedostępne, natomiast istnieje bogactwo danych zagregowanych. Wnioski wyciągane z analiz takich materiałów są jednak obciążone błędem ekologicznym – jeżeli dotyczą one jednostek. Z pomocą przychodzą metody wnioskowania ekologicznego, które pozwalają z dużą trafnością oszacować zachowania i charakterystyki jednostkowe oparte na danych zagregowanych. Artykuł ten przedstawia najpopularniejsze metody wnioskowania ekologicznego wykorzystujące podejście regresyjne i logistyczne. Przegląd rozpoczyna analiza regresji ekologicznej Goodmana, a kończy stosunkowo nowe rozwiązanie wykorzystujące zasadę maksymalizacji entropii. Przedstawione tu metody wnioskowania ekologicznego nie wyczerpują wszystkich spotykanych w literaturze przedmiotu, ale jedynie te najbardziej popularne i najczęściej dyskutowane. Celem mojego artykułu jest przybliżenie polskiemu czytelnikowi tych podejść.
Główne pojęcia: analiza ekologiczna, analiza danych zagregowanych, regresja ekologiczna.

Analiza danych zagregowanych

W analizach zagadnień politycznych często występują dane zagregowane – informacje uogólnione na pewnym poziomie, oparte na podziale administracyjnym bądź geograficznym. W takim przypadku informacje nie dotyczą jednostek (pojedynczych osób), ale określonych zbiorowości (agregatów) – gmin, województw, pracowników przedsiębiorstw, wyborców partii. Przykładem mogą być

Uwagi do Autora lub prośby o nadbitki prosimy kierować: Instytut Socjologii UJ, 31–044 Kraków, ul. Godzka 52.

oficjalne wyniki wyborów czy zbiorcze dane ze spisów powszechnych. Do dyspozycji jest wówczas olbrzymi materiał empiryczny, którego jedyną wadą jest agregatywny charakter, a zatem brak możliwości wnioskowania o zachowaniach na poziomie indywidualnym. Dane zagregowane są szczególnie nieocenionym źródłem informacji w przypadku prowadzenia analiz historycznych, kiedy nie można już generować potrzebnych informacji. Stąd też stosunkowo wcześnie pojawiły się próby wykorzystania tego rodzaju analiz w naukach społecznych. Już Emile Durkheim, uznawany za prekursora socjologii empirycznej, w pracy dotyczącej samobójstwa (1897) zastosował podejście badawcze polegające na analizie danych zagregowanych. Pierwsze próby wnioskowania z takich danych, w zastosowaniu do zagadnień politycznych, podjęto w Niemczech. R. Blank (1905 za: Achen i Schiveley 1995) zaproponował sposób oszacowania wpływu głosowania burżuazji na wyniki wyborcze uzyskane przez socjalistów. Dwie dekady później Ferdinand Tönnies (1924 za: Achen i Schiveley 1995) obliczył korelację ekologiczną pomiędzy rezultatami dwóch elekcji w Kolonii. Natomiast Felix Bernstein (1932 za: Achen i Schiveley 1995) opracował metodę wnioskowania z danych zagregowanych za pomocą regresji, która dwadzieścia lat później została podjęta przez Williama Robinsona (1950), a udoskonalona przez Leo Goodmana (1953). Analiza danych zagregowanych została upowszechniona dopiero w Stanach Zjednoczonych[1] w latach pięćdziesiątych XX wieku wraz z rozwojem zaawansowanych technik analizy danych i odpowiednich algorytmów matematycznych.

W socjologii podejście badawcze, które zajmuje się ustalaniem przestrzennego wpływu konfiguracji pewnych zjawisk społecznych na ich charakter, określa się mianem „ekologii społecznej". Ponieważ w przypadku danych zagregowanych mamy do czynienia właśnie z podziałem na jednostki terytorialne, tego typu analizy nazywa się analizami ekologicznymi. Można wskazać dwa rodzaje wykorzystania danych zagregowanych (Thomsen 1987).

Pierwszym jest analiza ekologiczna *sensu stricte* (*ecological analysis*). Badania tego typu polegają na analizie danych wyłącznie na poziomie zagregowanym. Konstruowane są modele matematyczne, które pozwalają uchwycić pewne wzorce wariancji na poziomie zagregowanym. Następnie na podstawie dodatkowych informacji (np. danych demograficznych opisujących skład społeczny rozpatrywanych jednostek terytorialnych) próbuje się wyjaśnić, co jest

[1] Pierwszym amerykańskim wykorzystaniem wnioskowania z danych zagregowanych była praca Ogburna i Goltry (za: Achen i Schiveley 1995) ze szkoły chicagowskiej, w której wykorzystano regresję ekologiczną.

źródłem tej zmienności. Przykładem jest geografia wyborcza, która szczególnie dobrze rozwinęła się we francuskiej tradycji badań politycznych (np. Gougel 1955). Jej celem jest poszukiwanie przestrzennych wzorców zachowań politycznych. Przykładami polskich analiz tego typu są prace Andrzeja Florczyka i Tomasza Żukowskiego (1990), Krzysztofa Ostrowskiego i Adama Przeworskiego (1996), Jacka Raciborskiego (1996) oraz Tomasza Zaryckiego i Andrzeja Nowaka (2000). Niestety wiele podobnych analiz pozostaje na poziomie prostych korelacji statystycznych, które często mogą prowadzić do błędnych wniosków.

Drugim rodzajem podejścia wykorzystującego dane zagregowane jest wnioskowanie ekologiczne (*ecological inference*). Celem jest odtworzenie z danych zagregowanych pewnych wzorów na poziomie indywidualnym. Inaczej ten sposób postępowania można nazwać deagregacją, gdyż celem postępowania jest odwrócenie porządku zebranych danych i przejście od informacji zagregowanych do indywidualnych. Wnioskowanie ekologiczne może dotyczyć wielu zagadnień: analiz ruchliwości społecznej, analiz demograficznych, studiów nad budżetami gospodarstw domowych czy marketingowej analizy lojalności konsumenckiej. Wnioskowanie ekologiczne przypomina prognozowanie, lecz nie dotyczy prognoz „wzdłużnych" – zdarzeń zachodzących w czasie, ale prognoz zdarzeń zachodzących na różnych poziomach tej samej przestrzeni historycznej. Jest to zatem prognozowanie „przekrojowe" i dlatego badania takie określa się czasami mianem „wnioskowania przekrojowego" (*cross-level inference*) [Achen i Schively 1995].

Wnioskowanie ekologiczne

Momentem zwrotnym w rozwoju metod wnioskowania ekologicznego był słynny artykułu Robinsona *Ecological Correlations and the Behavior of Individuals* (1950), w którym występował on z ostrą krytyką wnioskowania ekologicznego. Autor pokazał, że próby analizy zachowań jednostkowych przy posługiwaniu się danymi zagregowanymi prowadzą do fałszywych rezultatów. Przytoczył on przykłady wielu prac socjologicznych, w których badacze wykorzystują korelacje ekologiczne – posługując się najczęściej współczynnikiem r Pearsona – obliczając związki pomiędzy średnimi pochodzącymi z agregacji pewnych zmiennych jednostkowych. Następnie wnioski z zależności na poziomie zagregowanym przenoszone są na zachowania indywidualne. Takie postępowanie jest zdaniem Robinsona niedopuszczalne, gdyż opiera się na fałszywych przesłankach. Jako przykład podaje dwie sytuacje, w których wyniki analizy ekolo-

gicznej zestawia z danymi indywidualnymi. W pierwszym przypadku Robinson wykorzystał zależność pomiędzy pochodzeniem rasowym a poziomem analfabetyzmu w USA. Na podstawie danych ze spisu powszechnego z 1930 roku obliczył odsetki białych i czarnych w każdym z 48 stanów oraz odsetek osób potrafiących i niepotrafiących czytać i pisać (podstawą procentowania byli mieszkańcy każdego stanu w wieku 10 i więcej lat). Korelacja ekologiczna (na poziomie zagregowanym) bycia czarnym i analfabetą wyniosła 0,773, podczas gdy na poziomie jednostkowym była o wiele mniejsza: 0,203. Jeszcze dobitniejszym przykładem jest związek obliczony przez Robinsona pomiędzy narodowością i poziomem analfabetyzmu. Posługując się tymi samymi danymi, Robinson pokazał, że korelacja na poziomie jednostkowym pomiędzy byciem imigrantem a analfabetyzmem wynosiła 0,118, podczas gdy na poziomie zagregowanym: –0,526. Pokazuje to, że posługiwanie się danymi zagregowanymi prowadzi nie tylko do osiągania różnych (w kategoriach wielkości) wyników, ale wręcz do przeciwstawnych wniosków.

Odpowiednim przykładem odnoszącym się do danych dotyczących chwiejności wyborczej w Polsce są oszacowania rozmiarów przepływów poparcia wyborczego w dwóch elekcjach: 1997 i 2001 roku. Zastosowanie prostego indeksu Pedersena[2] do różnych poziomów agregacji przynosi odmienne wyniki. Dla ogółu kraju zagregowana chwiejność wyborcza wyniosła 19,07%[3]. Posłużenie

[2] Miara ta jest wyrażana równaniem:

$$V = \frac{\sum_{i=1}^{n} \left| p_{i,t} - p_{i,t+1} \right|}{2} \quad ,$$

gdzie $p_{i,t}$ – oznacza odsetek głosów uzyskanych przez *i*-tą partię w trakcie wyborów *t*, a $p_{i,t+1}$ – odsetek głosów uzyskanych przez tę partię w trakcie kolejnych wyborów *t+1*. Wartość minimalna 0 oznacza, że partia (partie) uzyskała takie samo poparcie w dwóch kolejnych wyborach, a wartość maksymalna 100 – całkowitą wymianę poparcia uzyskanego w parze wyborów.

[3] Przy obliczaniu tych miar przyjęto następujące założenia o ciągłości partii:

(i) Spadkobiercami AWS zostały: Akcja Wyborcza Solidarność Prawicy, 1/2 PO, 1/2 Ligi Polskich Rodzin oraz Prawo i Sprawiedliwość.

(ii) Spadkobiercami SLD, UP i Krajowej Partii Emerytów i Rencistów została koalicja SLD–UP.

(iii) UW i Unia Polityki Realnej w wyborach 2001 roku startowały razem, a kontynuatorami UW były UW i 1/2 PO.

(iv) Spadkobiercą ROP jest 1/2 LPR.

(v) Wszystkie ugrupowania mniejszości niemieckiej traktowano łącznie.

Do kategorii inne partie zaliczono w 1997 roku: Blok dla Polski, Krajowe Porozumienie Emerytów i Rencistów Rzeczypospolitej Polskiej oraz wszystkie lokalne komitety wyborcze, a w 2001 roku: Ruch Społeczny „Alternatywa", Polską Unię Gospodarczą oraz Polską Partię Socjalistyczną.

się danymi indywidualnymi – sondażami Polskiego Generalnego Sondażu Wyborczego 1997 i Polskiego Generalnego Studium Wyborczego 2001 – do obliczenia przesunięć poparcia na poziomie jednostkowym prowadzi do innego wniosku: wskaźnik chwiejności wyborczej wynosi 23,20%. Pokazuje to, że wyniki uzyskiwane na podstawie danych zagregowanych i indywidualnych mogą być odmienne, co nakazuje ostrożność w posługiwaniu się tego typu materiałami.

Artykuł Robinsona odegrał znaczącą rolę w dalszym rozwoju wnioskowania ekologicznego. Od jego publikacji zaczęło się powszechnie mówić o błędzie wnioskowania ekologicznego (*ecological fallacy*), chociaż stwierdzenie takie nie pojawiło się w oryginale pracy Robinsona. O upowszechnieniu się przekonania, że wnioskowanie o cechach jednostkowych na podstawie danych zagregowanych jest niemożliwe, świadczy pośrednio liczba cytowań artykułu Robinsona sięgająca ponad ośmiuset powołań się na niego (King 1997). Stwierdzenie o błędzie ekologicznym wpłynęło na dalsze badania w dziedzinie analizy danych zagregowanych w dwojaki sposób. Jednych ostatecznie zniechęciło do jakichkolwiek prób sięgania po wnioskowanie ekologiczne. Jednak dla znacznej części badaczy sformułowanie Robinsona podziałało jak wyzwanie i zaczęli oni szukać skuteczniejszych metod wnioskowania ekologicznego, co doprowadziło do opracowania kilku nowych podejść badawczych[4].

Należy również zaznaczyć, że krytyka wysunięta przez Robinsona nie jest pozbawiona błędów. Przede wszystkim autorowi można zarzucić, że w przytoczonym przykładzie zastosował on skrajnie różne poziomy analizy danych – z jednej strony obliczył korelacje dla stanów, a z drugiej dla pojedynczych jednostek. Jargowsky (2004) uważa, że główny zarzut, jaki można postawić twierdzeniu Robinsona o „błędzie wnioskowania ekologicznego", to kwestia właściwego skonstruowania modelu. Za nietrafnym wnioskiem Robinsona stoją ukryte zmienne, których wpływ autor pominął. Powodem negatywnej korelacji ekologicznej narodowości i poziomu analfabetyzmu jest fakt, że imigranci osiedlali się z reguły w stanach północnych z lepiej rozwiniętym przemysłem i o wyższych wskaźnikach rozwoju, w których równocześnie odsetek analfabetów był mniejszy (Jargowsky 2002). Przy złej specyfikacji modelu, który służy wyciąganiu wniosków o zależnościach indywidualnych z korelacji na poziomie zagregowanym, twierdzenie o „błędzie ekologicznym samo staje się błędem" (Firebaugh 1978: 570).

[4] Omówienie historii rozwoju podejść we wnioskowaniu ekologicznym znaleźć można w pracy Achena i Schively'ego (1995).

Metody wnioskowania ekologicznego

Wśród podejść stosowanych do wnioskowania ekologicznego można wskazać na dwie ogólne grupy: oparte na zaproponowanym przez Goodmana modelu regresji liniowej (1953) oraz model Thomsena (1987), wykorzystujący regresję logitową, który jednak nie jest tak popularny jak pozostałe podejścia. Metoda regresji ekologicznej Goodmana była bezpośrednią odpowiedzią na krytykę wysuniętą przez Robinsona. Inne modele wnioskowania ekologicznego oparte na schemacie regresji liniowej to: metoda wartości brzegowych (*method of bounds*) Duncana i Davisa (1953), sąsiedzki model liniowy (*linear nieighborhood model*) zaproponowany przez zespół Freedmana (1991), metoda EI opracowana przez Kinga (1997) oraz trzy modele autorstwa Grofmana i Merrilla (2002a, 2002b). Zupełnie innym podejściem jest metoda maksymalizacji entropii, będąca zastosowaniem do wnioskowania ekologicznego przez Johnstona i Pattie (2000) fizycznej trzeciej zasady termodynamiki.

Celem każdej ze wspomnianych metod wnioskowania ekologicznego, jest wyjaśnienie określonych zależności, bądź też charakterystyka pewnych zjawisk na poziomie jednostkowym na podstawie danych zagregowanych. Wyobraźmy sobie sytuację, kiedy w danym okręgu wyborczym składającym się z *i*-obwodów dwie kategorie wyborców – biali i czarni – mogą wybierać spośród dwóch kandydatów: demokraty i republikanina[5]. Demokraci uznawani są za reprezentantów interesów czarnych (i innych mniejszości rasowych), a republikanie – białych. Dysponując jedynie danymi zagregowanymi trudno uzyskać szczegółowe informacje dotyczące zachowania wyborczego pojedynczych jednostek. Jeżeli jednak istnieją dane sondażowe, należy traktować je z dużą ostrożnością. Badacz dysponując oficjalnymi danymi wyborczymi może posiłkować się informacjami ze spisów powszechnych, opisującymi skład rasowy danego obwodu. Praktycznym rozwiązaniem jest przedstawienie tych informacji w prostej tabeli krzyżowej (tabela 1).

Tabela zawiera w kolumnach, oprócz danych o frekwencji wyborczej, liczbę głosów uzyskanych przez kandydata demokratów i republikanów, natomiast w wierszach skład rasowy – liczbę czarnych i białych mieszkańców danego obszaru. Nieznane pozostają wartości poszczególnych komórek tej tabeli – liczba

[5] Najczęstszym praktycznym zastosowaniem wnioskowania ekologicznego jest wykazanie bądź zaprzeczenie przypadków głosowania rasowego (odmienne zachowanie białych i czarnych) w wyborach stanowych w USA. Dlatego ogólny model postępowania przy wnioskowaniu ekologicznym przedstawiony będzie na tym właśnie przykładzie.

białych i czarnych głosujących na kandydata demokratów i republikanów oraz tych, którzy nie wzięli udziału w wyborach.

Tabela 1. Problem wnioskowania ekologicznego

	Demokraci	Republikanie	Niegłosujący	Ogółem
Biali	?	?	?	25706
Czarni	?	?	?	55054
Ogółem	19896	10936	49928	80760

Źródło: King 1997. Dane dotyczące okręgu numer 42 z wyborów stanowych 1990 roku w Ohio.

Tabela 2. Wnioskowanie ekologiczne dla obwodu wyborczego *i*

	Głosujący na kandydata czarnych	Niegłosujący	Ogółem
Czarni	β_i^b	$1-\beta_i^b$	x_i
Biali	β_i^w	$1-\beta_i^w$	$1-x_i$
Ogółem	y_i	$1-y_i$	1

Źródło: Opracowanie własne na podstawie King (1997) oraz Grofman i Merrill (2002a).

Celem wnioskowania (tabela 2) jest obliczenie odsetka głosów oddawanych przez czarnych na kandydata demokratów (β_i^b) i odsetka głosów białych oddanych na tego samego kandydata – demokratę (β_i^w) – w obwodzie wyborczym i (i=1,2,..., p). Kategoria „niegłosujący" jest sumą odsetka czarnych, którzy wstrzymali się od udziału w wyborach oraz odsetka głosów oddanych przez czarnych na kandydata republikanów (lub odpowiednio odsetków białych niebiorących udziału w wyborach i tych białych, którzy zagłosowali na kandydata republikanów). Informacjami, które pochodzą z oficjalnych danych wyborczych i spisowych są: y_i – odsetek głosów oddanych na kandydata czarnych w obwodzie i, i x_i – odsetek głosów oddanych przez czarnych wyborców w obwodzie i.

Ponieważ liczby głosów oddanych w obwodzie sumują się do jedności, dysponując takimi danymi łatwo można obliczyć odsetek głosów oddanych przez białych wyborców ($1-x_i$) oraz pozostałych głosów ($1-y_i$).

Regresja ekologiczna Goodmana

Zaproponowane przez Goodmana podejście wnioskowania ekologicznego opiera się na zastosowaniu zwykłej procedury dwuzmiennowej regresji liniowej metodą najmniejszych kwadratów (*ordinary least squares* – OLS). Dokonywana jest regresja poparcia dla kandydata czarnych w *i*-tym obwodzie na odsetek czarnych (x_i) i białych ($1-x_i$) głosujących w tym obwodzie z pominięciem wyrazu wolnego. Równanie takiej regresji można zapisać jako:

$$y_i = \beta_i^b x_i + \beta_i^w (1-x_i) \qquad [1]$$

Zmiennymi wyjaśnianymi (szacowanymi parametrami) jest odsetek głosów oddanych w danym obwodzie *i* przez czarnych (β_i^b) oraz przez białych (β_i^w) na kandydata demokratów. Przy tak zapisanym równaniu regresji [1] znanymi informacjami są poparcie dla kandydata demokratów (y_i) oraz skład rasowy obwodu *i* – odsetki jego czarnych i białych mieszkańców (x_i oraz $1-x_i$). W takiej postaci równanie to jest niemożliwe do rozwiązania, gdyż są dwie niewiadome – szacowane parametry β_i^b i β_i^w. Goodman poradził sobie z tym problemem przyjmując założenie o *niezmienności parametrów* we wszystkich obwodach składających się na pewien okręg. Wymóg ten sprowadza się do uznania, że wyniki głosowania tak w okręgu, jak i składających się na niego obwodach nie zależą od składu rasowego. Mówiąc inaczej przyjmuje się, że w obwodach stanowiących pewien okręg dokładnie taki sam odsetek czarnych głosował na kandydata demokratów. Podobnie odsetki białych popierających tego kandydata (demokratę) były identyczne we wszystkich obwodach. W rezultacie poparcie dla demokraty wśród czarnych i białych na poziomie okręgu jest równe odsetkom głosów oddanych na tego kandydata przez obie te grupy w poszczególnych obwodach (dla wszystkich obwodów $\beta^b = \beta_i^b$ i $\beta^w = \beta_i^w$). Wówczas ze znajomości wyników głosowania w całym okręgu można oszacować wielkości poparcia dla demokraty wśród czarnych i białych w poszczególnych obwodach.

Jest to założenie o tyle nierealne, że trudno sobie wyobrazić okręg wyborczy, w którym nie ma różnic zachowań wyborczych w poszczególnych obwodach. Jeżeli różnice parametrów jednostkowych w poszczególnych obwodach wystąpią i będą zależały od składu rasowego obwodów (x_i), wyniki regresji zostaną obciążone. Przykłady radzenia sobie z zależnością od składu rasowego (x_i) – która nazywana jest problemem nieoznaczoności (*indeterminancy problem*) – można znaleźć w pracy Kinga (1997: 41–46).

Mimo swoich wad regresja ekologiczna była powszechnie stosowana przez czterdzieści lat i nadal jest wykorzystywana do wnioskowania z danych zagregowanych. Przykładem mogą być analizy przepływów elektoratów wykonane w Austrii przez Hofingera i Ogrisa (2000). Posłużyli się oni zwykłą regresją do oszacowania mobilności poparcia wyborców w dwóch elekcjach, stosując następujący model:

$$y_i = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_{NV} x_{NV} \qquad [2]$$

Poparcie dla *i*-tej partii (y_i) jest poddawane regresji na wyniki uzyskane przez *n* partii politycznych biorących udział w poprzednich wyborach (x_1, x_2,....,x_n) oraz kategorię niegłosujących i nowych wyborców (x_{NV}). Podobnie jak w oryginalnym podejściu Goodmana pominięto wyraz wolny. Wyniki stanowią oszacowania przesunięć poparcia wyborczego.

Metoda wartości brzegowych

Drugą odpowiedzią na krytykę Robinsona była propozycja Duncana i Davisa (1953), którzy do równania regresji wprowadzili dodatkowe ograniczenie. Zaproponowali wykorzystanie informacji z marginesów tabeli, co umożliwiło zawężenie przedziałów szacowanych parametrów (tabela 3).

Dla pierwszej komórki – liczby czarnoskórych głosujących na kandydata demokratów – możliwe są wartości z przedziału od 0 do 19896. Wprawdzie czarnoskórych w tym okręgu było więcej, bo 25706, ale pozostałych 5810 nie mogło oddać swych głosów na demokratę, ponieważ w całym okręgu tego kandydata poparło 19896 osób. Wartości brzegowe tabeli określają przedziały, w których mieszczą się parametry regresji. Metoda wartości brzegowych znacznie wzmacnia precyzję wnioskowania.

Tabela 3. Zastosowanie metody wartości brzegowych

	Demokraci	Republikanie	Niegłosujący	Ogółem
Czarni	[0;19896]	[0;10936]	[0;25706]	25706
Biali	[0;19896]	[0;10936]	[0;25706]	55054
Ogółem	19896	10936	49928	80760

Źródło: King 1997.

W obwodach heterogenicznych[6] można określić szerokość przedziałów szacowanych parametrów na podstawie dwóch nierówności (King 1997: 79):

$$\max\left(0, \frac{y_i - (1 - x_i)}{x_i}\right) \leq \beta_i^b \leq \min\left(\frac{y_i}{x_i}, 1\right)$$

$$\max\left(0, \frac{y_i - x_i}{1 - x_i}\right) \leq \beta_i^w \leq \min\left(\frac{y_i}{1 - x_i}, 1\right) \qquad [3]$$

Gdzie x_i to liczba czarnych w obwodzie i, a y_i liczba głosów oddanych na kandydata czarnych. Ponieważ β_i^b i β_i^w są ze sobą związane równaniem regresji, a pozostałe wyrażenia w tym równaniu są znane (y_i, x_i), liniowe relacje pomiędzy nimi można zapisać jako:

$$\beta_i^b = \frac{y_i}{x_i} - \beta_i^w \frac{(1 - x_i)}{x_i}$$

$$\beta_i^w = \frac{y_i}{1 - x_i} - \beta_i^b \frac{x_i}{1 - x_i} \qquad [4]$$

Prosta reprezentująca tę relację jest zawsze nachylona ujemnie, ze względu na ujemną wartość drugiego członu w równaniach [4]. Oznacza to, że gdy β_i^b rośnie, to β_i^w maleje i na odwrót. Jeżeli β_i^b znajduje się blisko swojej górnej granicy, to β_i^w musi znaleźć się blisko swojej dolnej granicy.

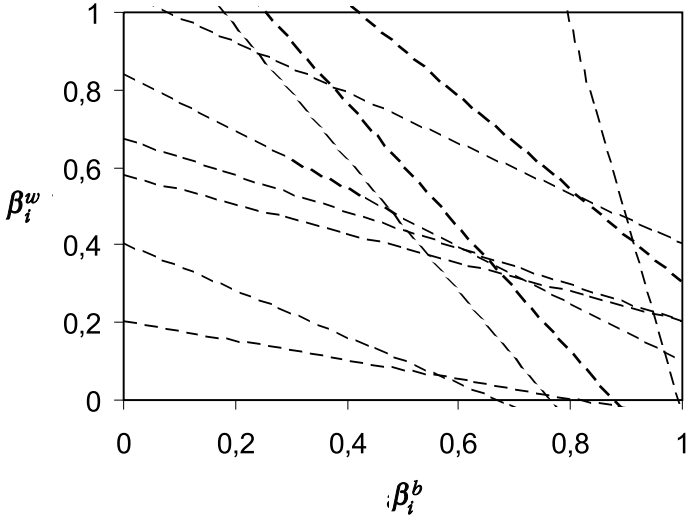
Aby wyobrazić sobie jak w praktyce wygląda zastosowanie metody wartości brzegowych, przyjmijmy, że w pewnym obwodzie wyborczym 60% uprawnionych do głosowania to czarni (x_i = 0,6), a kandydat popierany przez nich uzyskał 50% głosów (y_i = 0,5). Przy takich założeniach możemy łatwo stwierdzić, że co najwyżej 5/6 czarnych wyborców oddało głosy na swego kandydata (5/6=50%/60%). Z drugiej strony wiadomo, że biali stanowią w okręgu 40% (1-x_i =100%-60%). I jeżeli nawet oni wszyscy poparli kandydata czarnych, musiał on otrzymać co najmniej 1/6 głosów od czarnych (1/6=(50%–40%)/60%). Co się zaś

[6] Dany obwód można uznać za heterogeniczny, jeśli: 1) będzie w nim mieszkał przynajmniej jeden czarny i jeden biały uprawniony do głosowania ($0 \leqslant x_i \leqslant 1$), 2) przynajmniej jedna osoba głosuje i jedna nie głosuje i 3) kandydat czarnych (i białych) otrzymują przynajmniej po jednym głosie ($0 \leqslant y_i \leqslant 1$).

tyczy białych wyborców, to w prosty sposób można obliczyć wartości przedziału, gdyż albo wszyscy poparli kandydata czarnych (40%<50%), albo żaden biały wyborca nie oddał swego głosu na tego kandydata (Grofman i Merrill 2002a).

Dla każdego obwodu zależność pomiędzy parametramii można przedstawić na tzw. wykresie tomograficznym (*tomographic plot*) (King 1997):

Wykres 1. Wykres tomograficzny parametrów β_i^b i β_i^w



Gdy nie wykorzystuje się informacji z rozkładów brzegowych, szacowane parametry β_i^b i β_i^w mogą przyjmować wartości z przedziału (0,1), czyli być dowolnym punktem na wykresie 1. Zastosowanie metody wartości brzegowych powoduje, że możliwe wartości stają się punktem leżącym na którejś z prostych opisanych równaniami [4]. W ten sposób rozwiązanie zaproponowane przez Duncana i Davisa pozwala na uzyskanie dodatkowych deterministycznych informacji, które znacznie wzmacniają pewność oszacowania.

Sąsiedzki model liniowy

Zespół Freedmana (Klein, Sacks, Smyth i Everett) w publikacji z 1991 roku zaproponował nowy sposób poradzenia sobie z ograniczeniami metody Goodmana. Nie było to zupełnie nowe podejście, lecz jedynie modyfikacja regresji ekologicznej z wykorzystaniem metody wartości brzegowych. Freedman przyjął

dwa założenia, które miały wyeliminować problem niezmienności parametrów obwodowych: o braku głosowania rasowego we wszystkich obwodach składających się na pewien okręg oraz o liniowej zależności pomiędzy preferencjami wyborczymi i kompozycją rasową danego obwodu.

Pierwszy warunek jest w zasadzie jedynie modyfikacją założenia Goodmana o niezmienności zachowań wyborczych w obwodach. W modelu sąsiedzkim liczba głosów oddana na kandydata demokratów przez czarnych i białych może być wprawdzie różna w poszczególnych obwodach, ale autorzy przyjęli arbitralnie, że na poziomie okręgu tyle samo czarnych i białych poparło demokratę ($\beta^b = \beta^w$). W praktyce sprowadza się do założenia, że w danym okręgu połowa czarnych i połowa białych głosowała na demokratę. Natomiast poparcie dla tego kandydata w poszczególnych obwodach może się różnić wśród obu tych grup rasowych (nie musiało na niego głosować po połowie czarnych i białych).

Drugi warunek można zapisać za pomocą równania:

$$\beta_i^b = \beta_i^w = \alpha^c + \alpha^s x_i \qquad [5]$$

Gdzie α^c oznacza wyraz wolny, a α^s nachylenie. Po podstawieniu do równania regresji otrzymujemy:

$$y_i = (\alpha^c + \alpha^s) x_i + \alpha^c (1 - x_1) \qquad [6]$$

Parametry okręgu uzyskiwane z tej regresji mają odmienne znaczenie niż w modelu Goodmana (Achen i Schively 1995; King 1997). O ile współczynniki β^b i β^w oznaczały proporcje poparcia dla kandydata czarnych przez odpowiednio czarnych i białych w całym okręgu, u Freedmana wyrażają proporcję tylko czarnych ($\alpha^c + \alpha^s$) i tylko białych (α^c) głosujących we wszystkich czarnych i we wszystkich białych obwodach wyborczych.

Praktyczne zastosowanie modelu Freedmana polega na wykorzystaniu metody wartości brzegowych, która pozwala wyznaczyć liniowe zależności parametrów obwodowych (β_i^b i β_i^w). Na podstawie tych funkcji można narysować wykres tomograficzny, zawierający dane dla wszystkich p obwodów składających się na okręg. Przyjęcie w następnym kroku założenia o braku głosowania rasowego w okręgu ($\beta^b = \beta^w$), umożliwia poprowadzenie na wykresie tomograficznym przekątnej przez punkty (0,0) i (1,1). Wielkości estymowanych parametrów wyznaczone są przez punkt przecięcia przekątnej z funkcjami liniowymi β_i^b, β_i^w (Grofman i Merrill 2002a). Takie arbitralne założenie pozwala wprawdzie poradzić sobie z problemem niezmienności, ale powodować może obciążenie wyni-

ków i nie pozwala na uzyskanie oszacowania parametrów obwodowych. Zdaniem Grofmana i Merrilla przyjęcie założenia o braku różnic w zachowaniach dwóch grup jest dopuszczalne tylko w przypadkach, kiedy analizowane zjawiska rzeczywiście pozostają niezależne na poziomie zagregowanym przy kontroli innych czynników. Przykładem może tu być chęć posiadania luksusowego samochodu wśród czarnych i białych Amerykanów. Pomijając wpływ zamożności, istnieje duże prawdopodobieństwo, że w równym stopniu biali i czarni chcieliby taki samochód posiadać.

Metoda EI Kinga

Najbardziej złożonym modelem wnioskowania ekologicznego, opartym na podejściu regresyjnym jest metoda Kinga. Określenie EI pochodzi od nazwy rozbudowanego programu napisanego przez Kinga w celu szacowania parametrów jednostkowych. Opiera on swoje rozwiązanie na dwóch przedstawionych uprzednio propozycjach – regresji ekologicznej i metodzie wartości brzegowych. Oryginalność tego podejścia rozpoczyna się w momencie właściwego szacowania parametrów jednostkowych.

King proponuje dwa modele – podstawowy wymagający dwuetapowego rozwiązania oraz przebiegający w jednym etapie – model rozszerzony. Przedstawię jedynie model podstawowy, który wyjaśnia ogólny sposób postępowania przy tej metodzie. Opiera się on na równaniu:

$$T_i = \beta_i^b X_\mathrm{i} + \beta_i^w (1 - X_i) \qquad [7]$$

T_i oznacza frekwencję wyborczą w *i*-tym obwodzie[7]. X_i jest odsetkiem czarnych zamieszkałych w tym obwodzie ($1-X_\mathrm{i}$ oznacza odsetek białych w tym obwodzie), a β_i^b i β_i^w są szacowanymi parametrami jednostkowymi – odpowiednio odsetkiem czarnych i białych głosujących w tym obwodzie na kandydata czarnych (demokratę). W modelu tym przyjęte zostały następujące założenia:

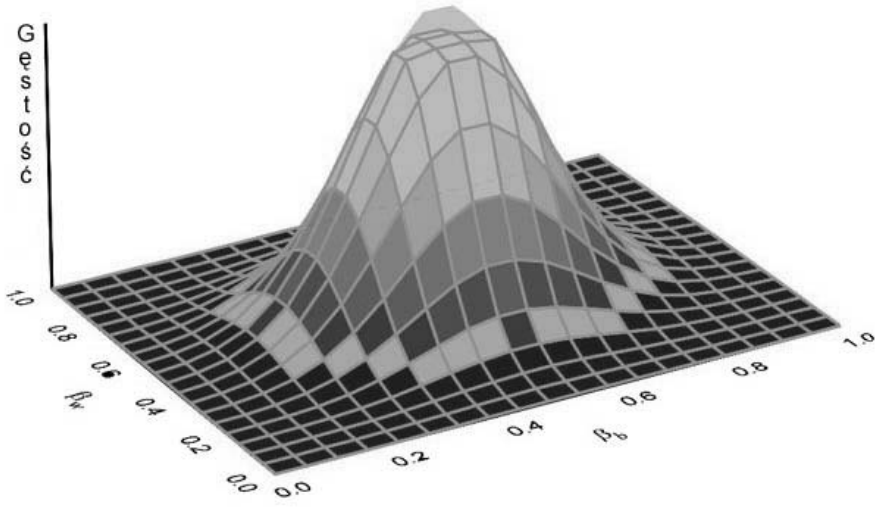
1) Dopuszczone jest, aby szacowane parametry β_i^b i β_i^w różniły się w poszczególnych obwodach. Wartości tych parametrów są modelowane na podstawie znajomości ich wzajemnego rozkładu.

[7] King (1997) w swojej metodzie stosuje inne niż poprzednio rozwiązanie – dokonuje regresji frekwencji wyborczej, a więc sumy głosów oddanych łącznie na kandydatów czarnych i białych. Wydaje się jednak, że dokonując regresji poparcia dla konkretnego kandydata wykorzystujemy bardziej szczegółowe informacje, co może mieć przełożenie na precyzję oszacowania.

2) Szacowane parametry β_i^b i β_i^w są niezależne od X_i. Co oznacza, że wymagany jest brak korelacji pomiędzy odsetkiem czarnych i białych popierających kandydata demokratów. Jest to równoznaczne z założeniem, że nie występuje błąd agregacji.
3) Wartości T_i w poszczególnych obwodach są warunkowo niezależne od X_i.

Najważniejsze jest pierwsze założenie. Wzajemne modelowanie szacowanych parametrów opiera się na założeniu normalności dwuzmiennowego rozkładu β_i^b i. Rozkład ten jest ograniczony do dopuszczalnych przedziałów wartości obu parametrów β_i^w . Ma on jedną dominantę i pewien zasób wariancji wokół jej wartości ⟨0,1⟩ (wykres 2).

Wykres 2. Przykład dwuzmiennowego rozkładu normalnego parametrów β_i^b i β_i^w



Źródło: Schuessler 1999.

W następnym kroku wykorzystywane są informacje z wartości brzegowych. Wykorzystując zwykłą metodę Duncana-Davisa można wyznaczyć funkcje liniowe parametrów β_i^b i β_i^w względem znanych wartości T_i i X_i.

Dysponując dwuzmiennowym rozkładem normalnym o znanych parametrach King dokonuje jego rzutowania na wykres tomograficzny parametrów obwodowych. Ograniczenie przestrzeni obu wykresów – tomograficznego i dwuzmiennowego – do tego samego przedziału wartości ⟨0,1⟩ na obu osiach umożliwia wyznaczenie położenia parametrów okręgu przy zastosowaniu procedury

największej wiarygodności opartej na znajomości rozkładu normalnego. Najbardziej prawdopodobne wartości parametrów okręgu można ulokować w punkcie modalnym dwuzmiennowego rozkładu normalnego – w jego wierzchołku. W ten sposób można wyodrębnić parametry okręgu.

Aby wyznaczyć wartości parametrów obwodowych konieczne jest przeprowadzenie drugiego etapu wnioskowania. King proponuje wykorzystać w tym przypadku informacje pochodzące z obu poprzednich wykresów – tomograficznego i dwuzmiennowego parametrów β_i^b i β_i^w. Każda funkcja liniowa parametrów β_i^b i β_i^w z wykresu rozkładu wartości brzegowych przecina w określonej płaszczyźnie rozkład dwuzmiennowy. W wyniku takiej operacji można wyznaczyć *i* rozkładów prawdopodobieństw *a posteriori* (otrzymanych za pomocą metody największej wiarygodności) parametrów obwodowych β_i^b lub β_i^w w zależności od tego, które z nich chcemy wyznaczyć. Na podstawie uzyskanych przekrojów, będących rozkładami prawdopodobieństw *a priori* parametrów obwodowych, można następnie uzyskać odpowiednie estymatory punktowe, stosując symulację matematyczną (metodę Monte Carlo z łańcuchami Markowa). Zastosowanie symulacji w drugim etapie powoduje, że szacowane wartości nie posiadają charakteru deterministycznego, ale są losowe. Oznacza to, że przy każdorazowym zastosowaniu modelu Kinga, nawet do tych samych danych, uzyskujemy inne wartości liczbowe. Oczywiście przedział otrzymanych rezultatów jest wąski, a co więcej metoda EI pozwala na określenie przedziałów zaufania szacowanych parametrów przez zastosowanie metody największej wiarygodności.

Modele Grofmana i Merrilla

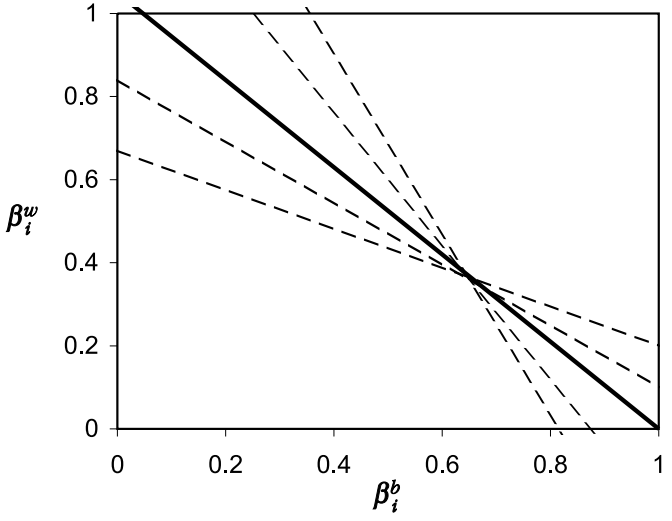
W odpowiedzi na krytykę zarówno metody Kinga, jak i liniowego modelu sąsiedzkiego Freedmana, Grofman i Merrill (2002a, 2002b) zaproponowali w sumie trzy różne sposoby wnioskowania ekologicznego. Przyjęte przez nich podejścia niewiele się różnią, jeśli chodzi o wprowadzone przez autorów założenia, ale każdy z modeli daje nieco odmienne oszacowania parametrów jednostkowych. Podobnie jak w dwóch poprzednich rozwiązaniach, również Grofman i Merrill oparli swoje podejście na klasycznej regresji ekologicznej, której trafność wzmocnili zastosowaniem metody wartości brzegowych Duncana i Davisa.

Grofman i Merrill wyszli od rozważenia przypadku szczególnego, w którym wszystkie funkcje liniowe parametrów obwodowych na wykresie tomograficznym przecinają się w jednym punkcie (wykres 3). W takiej sytuacji punktowym estymatorem parametrów okręgu β_i^b i β_i^w jest właśnie ten punkt przecięcia – co niestety zdarza się bardzo rzadko, zwłaszcza przy większej liczbie obwodów.

Model I – minimalizacja dystansów

Rozwiązanie to polega na połączeniu na jednym wykresie tomograficznym wszystkich funkcji parametrów obwodowych β_i^b i β_i^w oraz prostej reprezentującej funkcję parametrów okręgu β_i^b i β_i^w (linia ciągła na wykresie 3). Tę ostatnią można wyznaczyć w ten sam sposób, jak dla poszczególnych obwodów, lecz na podstawie danych z okręgu.

Wykres 3. Wykres tomograficzny parametrów obwodowych. Szczególny przypadek, gdy funkcje liniowe wszystkich obwodów przecinają się w jednym punkcie.



Źródło: Grofman i Merrill 2002a.
Uwaga: Linia ciągła reprezentuje funkcję parametrów okręgu β^b i β^w.

Zagregowane wartości: odsetki poparcia udzielanego przez czarnych i białych dla kandydata czarnych w całym okręgu Grofman i Merrill proponują obliczyć minimalizując sumę ważonych odległości euklidesowych pomiędzy prostą okręgu i wszystkimi p prostymi obwodów. Powstaje w ten sposób miara dystansu, którą można zapisać jako:

$$d_i^2 = \frac{\left[y_i - x_i\beta^b - (1-x_i)\beta^w\right]^2}{x_i^2 - (1-x_i)^2} \qquad [8]$$

Dystans d_i można również interpretować jako ważoną różnicę pomiędzy odsetkiem głosów oddanych przez wszystkich wyborców na kandydata czarnych w *i*-tym obwodzie, a odsetkiem głosów oddanych na tego kandydata w całym okręgu, przy założeniu, że zarówno czarni, jak i biali wyborcy głosowali tak samo we wszystkich obwodach okręgu. Wynika to z faktu, że licznik w wyrażeniu [8] można przedstawić jako różnicę $(y_i - \hat{y}_i)^2$, gdzie $\hat{y}_i$ jest odsetkiem głosów oddanych na czarnego kandydata szacowanym za pomocą regresji Goodmana (tzn. przy założeniu, że wszyscy wyborcy głosują tak samo w całym okręgu $\beta_i^b = \beta_i^w = \beta$).

Posługując się tak skalkulowanym dystansem pomiędzy prostymi okręgu i obwodów można wyznaczyć parametry okręgowe wykorzystując mnożniki Lagrange'a:

$$\beta^b = \frac{\sum_{i=1}^{p} w_i^2 (x_i - x)[(1-x)y_i - (1-x_i)y]}{\sum_{i-1}^{p} w_i^2 (x_i - x)^2}$$

$$\beta^w = \frac{\sum_{i=1}^{p} w_i^2 (x_i - x)[x_i y - x y_i]}{\sum_{i-1}^{p} w_i^2 (x_i - x)^2} \qquad [9]$$

Natomiast parametry obwodowe β_i^b i β_i^w oblicza się minimalizując wyrażenie: $(\beta_i^b - \beta^b)^2 + (\beta_i^w - \beta^w)^2$. Szacowana wartość jest miejscem przecięcia prostych prostopadłych do poszczególnych funkcji parametrów jednostkowych przechodzących przez punktowy estymator okręgu.

Przedstawiony model wymaga sformułowania dwóch zastrzeżeń. Kiedy otrzymujemy punkt leżący poza przyjętymi granicami parametrów, tzn. przedziałem ⟨0,1⟩, minimalizuje się odległości pomiędzy najbliższymi punktami leżącymi w obrębie tych granic (czy to na osi β_i^b czy β_i^w). Drugie zastrzeżenie dotyczy przypadku, kiedy wielkości obwodów są różne (tzn. obwody różnią się liczbą mieszkańców). Wówczas oprócz wag w_i^2, uwzględnia się wagi skonstruowane przy wykorzystaniu liczb uprawnionych w poszczególnych obwodach składających się na okręg n_i. W rezultacie minimalizowany jest dystans $\sum_{i=1}^{p} n_i d_i^2$.

Model opierający się na minimalizacji odległości umożliwia także obliczenie odchylenia standardowego i przedziałów zaufania szacowanych parametrów za pomocą metody *bootstrap*.

Model II – minimalizacja dystansów w przestrzeni (*m*,*b*)

Rozwiązanie to podobne jest do zaproponowanego w pierwszym modelu. Również w tym przypadku parametry szacowane są przez minimalizację odległości pomiędzy prostą okręgu i prostymi obwodów. Jednak, aby uprościć obliczenia, wszystkie działania wykonywane są nie w przestrzeni (β_i^b, β_i^w), ale w przestrzeni (*m*,*b*). Transformacja polega na przekształceniu równań regresji w następujący sposób:

$$m = \beta^b - \beta^w \qquad b = \beta^w$$

$$m = \left(b - \left(\frac{y}{1-x} \right) \left(\frac{1-x}{-x} \right) \right) - b = -\frac{b}{x} + \frac{y}{x} \qquad [10]$$

$$y = mx + b$$

W nowej przestrzeni *m* określa nachylenie prostych, a *b* – punkt przecięcia z osią β_i^b lub β_i^w. Dalej sposób postępowania jest identyczny jak w modelu I. Z tym, że zamiast w płaszczyźnie kwadratowej wyznaczonej przez pary punktów (0,0) i (1,1), teraz operacje przeprowadzane są w płaszczyźnie rombu z wierzchołkami: (0,0), (0,1), (–1,1) i (0,1). Aby w przestrzeni (*m*,*b*) obliczyć dystans d_i^2 od pewnego punktu na prostej okręgu (m_0,b_0) należy posłużyć się równaniem:

$$d_i^2 = \frac{x_i^4}{\left(1 + x_i^2\right)^3} \left(b_0 + y_i - m_0 x_i\right)^2 \qquad [11]$$

Wszystkie pozostałe założenia i rezultaty tego modelu są takie same jak w poprzednim.

Model III – minimalizacja powierzchni

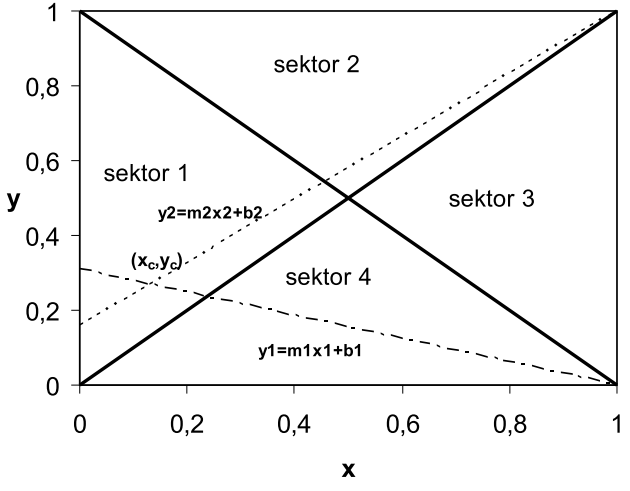
Trzeci model Grofmana i Merrilla opiera się na założeniu, że przestrzeń (*x*,*y*) wyznaczoną przez znane wartościi można podzielić dwoma przekątnymi na cztery sektory, w których może się znaleźć punkt przecięcia prostych obwodowych z prostą okręgu – punktowy estymator zagregowanych parametrów (wykres 4).

Po określeniu położenia tego punktu w którymś obszarze przestrzeni (*x*,*y*) autorzy proponują obliczyć jego dokładne parametry za pomocą minimalizacji sumy powierzchni pomiędzy *p* prostymi obwodów (pod tymi prostymi) a linia-

mi wyznaczającymi granice danego sektora tej przestrzeni. Jeżeli punkt przecięcia $x_c \in \langle 0,1 \rangle$, to powierzchnia pomiędzy każdą z prostych wynosi:

$$P = \left| \frac{\Delta b^2}{\Delta m} + \frac{\Delta m}{2} + \Delta b \right| \qquad [12]$$

Wykres 4. Sektory wyznaczające położenie punktów przecięcia (x_c,y_c) w przestrzeni (x,y)



Źródło: Grofman i Merrill 2002a.
Uwaga: Linie ciągłe wyznaczają granice obszarów, a przerywane określają przykładowe proste obwodów.

Jeśli punkt przecięcia $x_c \notin \langle 0,1 \rangle$, to powierzchnię tę można obliczyć z równania:

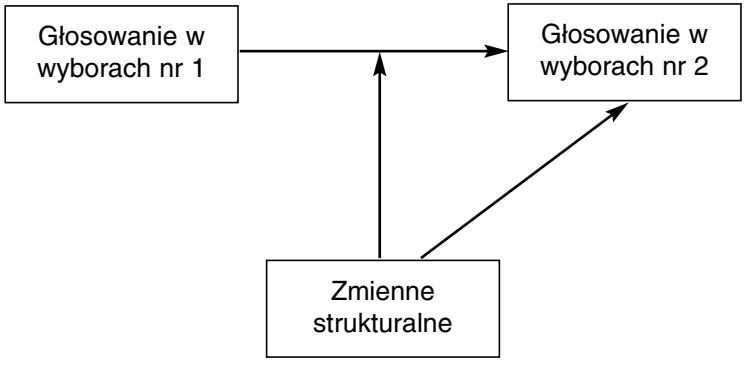
$$P = \left| \frac{\Delta m}{2} + \Delta b \right| \qquad [13]$$

W powyższych równaniach m oznacza nachylenie prostych, a b punkt przecięcia. Aby wyznaczyć zagregowany estymator punktowy należy zminimalizować sumę uzyskanej powierzchni. Dalszy sposób postępowania jest zgodny z zaproponowanym przez autorów w modelu I.

Metoda ukrytej struktury

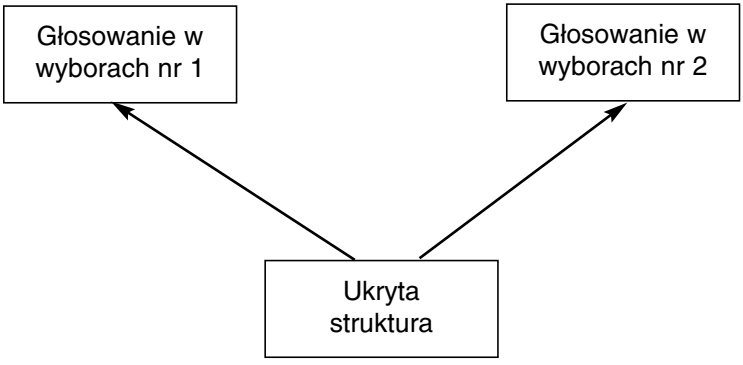
Zupełnie odmiennym rozwiązaniem problemu wnioskowania ekologicznego jest metoda zaproponowana przez Thomsena (1987). Inspiracją były prace Rascha dotyczące koncepcji pomiaru. Zdaniem Thomsena, problem „mitu wnioskowania ekologicznego" można sprowadzić do teoretycznego uzasadnienia zależności pomiędzy indywidualną wariancją w obrębie agregatów, a wariancją pomiędzy nimi. Takich teoretycznych ram dla wnioskowania ekologicznego nie daje zwykła regresja Goodmana. Podejście regresyjne podlega dwóm ograniczeniom – niewystarczającej specyfikacji zmiennych w równaniu regresji (*mis-specification*) i złożonej współzależności zmiennych (*multi-collinearity*). Odpowiednia specyfikacja modeli regresji polega na możliwie pełnym wyjaśnieniu wariancji zmiennej zależnej przez zmienne niezależne wprowadzane do równania, co nie zawsze jest możliwe. Z drugiej strony nadmierne rozbudowanie modelu wyjaśniającego prowadzi do drugiego problemu – współzależności zmiennych niezależnych. Według Thomsena wadą wszystkich podejść opartych na regresji jest przyjęcie asymetrycznego modelu wyjaśnienia kowariancji zmiennych zależnych i niezależnych – uznanie za przyczynę wariancji zmiennej zależnej, wariancji zmiennych niezależnych. W przypadku przepływów poparcia wyborczego zakłada się, że głosowanie w danych wyborach wynika z preferencji w poprzednich oraz dodatkowych zmiennych strukturalnych, na przykład cech pochodzenia społecznego (rysunek 1).

Rysunek 1. Model regresji w analizie przepływów poparcia wyborczego



Źródło: Thomsen 1987.

Rysunek 2. Model ukrytej struktury w analizie przepływów poparcia wyborczego



Źródło: Thomsen 1987.

Thomsen proponuje inną teoretyczną podstawę wnioskowania ekologicznego – koncepcję ukrytej struktury (*latent structure*), która wywodzi się od Paula Lazersfelda. Uznaje on, że za kowariancję zmiennych zależnych i niezależnych odpowiada dodatkowy, ukryty czynnik. W analizie przepływu elektoratów przyczyną określonych zachowań wyborczych w jednych i drugich wyborach jest pewna ukryta struktura – syndrom lub zespół czynników, które determinują określone zachowania wyborcze (rysunek 2). Może to być na przykład wspólnota zamieszkania, określająca podobieństwo interesów wyborców z danego terenu, które przekładają się na wyniki głosowania.

Model ukrytej struktury zakłada, że wybór pewnej opcji (w tym przypadku wybór konkretnej partii) jest funkcją ukrytej zmiennej oddziałującej na każdą jednostkę. Zagregowane prawdopodobieństwa oddania głosu na partię x w pierwszych wyborach i partię y w drugich wyborach są funkcjami tej samej ukrytej zmiennej θ oddziałującej na wyborców w obu elekcjach:

$$P(x|\theta) = f(\theta) \qquad P(y|\theta) = g(\theta) \qquad [14]$$

Najważniejszym wymogiem wprowadzonym przez Thomsena jest założenie izomorfizmu[8]. Dotyczy ono powiązania wariancji na poziomie indywidualnym

[8] Thomson wywodzi rozumienie tego pojęcia z chemii, gdzie izomorfizm oznacza „podobieństwo struktur, zwłaszcza w przypadku kryształów różniących się składem chemicznym" (Thomson 1987: 55).

(na przykład wariancji dotyczącej preferencji politycznych wyborców) wewnątrz wyróżnionych obszarów z wariancją na poziomie zagregowanym. Zgodnie z tym wymogiem, wnioskowanie ekologiczne jest możliwe, jeśli wariancja na poziomie indywidualnym ma taką samą strukturę jak wariancja na poziomie zagregowanym (jest do niej proporcjonalna).

Tabela 4. Notacja wnioskowania ekologicznego metodą ukrytej struktury. Przykład dwóch zmiennych

		Wybory nr 2		
		Partia y	Partia nie-y ($\bar{y}$)	Ogółem
Wybory nr 1	Partia x	$P(xy)_0$	$P(x\bar{y})_0$	$P(x)_0$
	Partia nie-x ($\bar{x}$)	$P(\bar{x}y)_0$	$P(\bar{x}\bar{y})_0$	$P(\bar{x})_0$
	Ogółem	$P(y)_0$	$P(\bar{y})_0$	1

Źródło: Thomsen 1987.

Jednostkową proporcję wyborców głosujących w pierwszych wyborach na partię x, którzy w drugich wyborach poparli partię y (notacja w tabeli 4) można obliczyć jako współczynnik korelacji ekologicznej. Zastosowany jest współczynnik korelacji r Pearsona pomiędzy probitami poparcia dla partii x i y w i-obwodzie wyborczym ważonymi liczebnością poszczególnych obwodów (N_i). Thomsen uważa, że ze względu na niewielką różnicę pomiędzy probitami i logitami poparcia dla poszczególnych partii można używać tych drugich, zwłaszcza że są prostsze do obliczenia.

Przedstawiony tu model dwuzmiennowy bardzo łatwo można rozszerzyć na przypadek wielu zmiennych – wówczas najważniejszą sprawą jest wybór odpowiedniej kategorii odniesienia. W modelu dwuzmiennowym była nią sztuczna kategoria „niepartia", którą stanowiła suma głosów oddanych na pozostałe partie oraz niegłosujących. W modelu wielozmiennowym najlepsze rezultaty daje przyjęcie za kategorię odniesienia odsetka niegłosujących.

Metoda maksymalizacji entropii

Metoda zaproponowana przez Johnstona i Pattiego (2000) odwołuje się bezpośrednio do trzeciej zasady termodynamiki, zgodnie z którą entropia układu odizolowanego (rozumiana jako nieuporządkowanie) jest stała. Przy wnioskowaniu ekologicznym metodą maksymalizacji entropii, punktem wyjścia jest ma-

cierz zawierająca informacje z poziomu zagregowanego – oficjalne dane wyborcze z dwóch kolejnych elekcji. W takiej macierzy $Px_1 \times Px_2$ celem wnioskowania są wartości wewnętrznych komórek. Dla uproszczenia rozpatrzmy macierz 2 × 2. Wiersze i kolumny oznaczają liczby głosów oddanych na dwie partie *P1* i *P2* w dwóch kolejnych wyborach (indeksy 1 i 2):

	$P1_1$	$P2_1$	
$P1_2$	?	?	4
$P2_2$	?	?	6
	3	7	

[15]

Macierz taką określa się *makrostanem*. Można ją zdekomponować na *mezostany* – macierze zawierające możliwe rozkłady jednostek zgodnie z wartościami brzegowymi makrostanu. Dla macierzy [15] możliwe są cztery mezostany spełniające warunki brzegowe:

1)

	$P1_1$	$P2_1$
$P1_2$	3	1
$P2_2$	0	6

2)

	$P1_1$	$P2_1$
$P1_2$	2	2
$P2_2$	1	5

3)

	$P1_1$	$P2_1$
$P1_2$	1	3
$P2_2$	2	4

4)

	$P1_1$	$P2_1$
$P1_2$	0	4
$P2_2$	3	3

[16]

Każdy mezostan posiada swój *mikrostan*, opisujący rozlokowanie konkretnych jednostek w poszczególnych wierszach i kolumnach. Dla pierwszego wiersza $P1_2$ istnieją cztery jednostki: *a*, *b*, *c*, *d*, które można uporządkować na cztery sposoby: 1) *a* w kolumnie $P1_1$, *b*, *c*, *d* w kolumnie $P2_1$, 2) *b* w kolumnie $P1_1$, *a*, *c*, *d* w kolumnie $P2_1$, 3) *c* w kolumnie $P1_1$, *a*, *b*, *d* w kolumnie $P2_1$ i 4) *d* w kolumnie $P1_1$, *a*, *b*, *c* w kolumnie $P2_1$. Jakkolwiek uporządkujemy jednostki w wierszu $P1_2$, wszystkie sześć jednostek w wierszu N_2 za każdym razem trafi do kolumny $P2_1$. Zatem liczba wszystkich uporządkowań mikrostanów dla pierwszego mezostanu wynosi cztery. Dla drugiego mezostanu można uzyskać sześć alokacji jednostek w wierszu N_1 i również sześć układów dla jednostek w wierszu $P2_2$. W re-

zultacie daje to 36 prawdopodobnych kombinacji. W przypadku trzeciego mezostanu, w pierwszym wierszu, możliwe są cztery mikrostany, a w drugim wierszu aż piętnaście mikrostanów, co daje łącznie 60 układów. Wreszcie, dla czwartego mezostanu jednostki w pierwszym wierszu dają się uporządkować tylko w jeden sposób (wszystkie trafiają do kolumny $P2_1$), natomiast jednostki z drugiego wiersza na osiemnaście sposobów, co w sumie umożliwia uzyskanie 18 mikrostanów.

Najbardziej prawdopodobny jest ten mezostan, w którym istnieje najwięcej możliwych alokacji jednostek wewnątrz macierzy. Macierz reprezentująca ten mezostan, zawiera rezultaty wnioskowania ekologicznego. Zgodnie z opisaną metodą maksymalizacji entropii, dla macierzy [15] najbardziej prawdopodobne są wartości trzeciego mezostanu.

Wnioskowanie ekologiczne za pomocą maksymalizacji entropii odbywa się z wykorzystaniem trójwymiarowych macierzy danych (kostek danych). W analizie przepływów elektoratów wiersze i kolumny dotyczą oficjalnych wyników dwóch kolejnych wyborów, a trzeci wymiar odnosi się do informacji z poszczególnych okręgów. Dzięki operowaniu na macierzach trójwymiarowych metoda maksymalizacji entropii wprowadza dodatkowe wartości brzegowe, co daje większą precyzję oszacowania. Czasami jako trzeci wymiar dekomponowanej macierzy, stosuje się informacje z badań sondażowych. Pozwala to lepiej dostosować wnioskowanie ekologiczne do rzeczywistych danych z poziomu jednostek.

Metoda maksymalizacji entropii zakłada, że alokacja jednostek w trójwymiarowej macierzy zgodnie z rozkładami brzegowymi zależy wyłącznie od relacji pomiędzy danymi. Innymi słowy przyjmuje, że na wewnętrzną strukturę tej macierzy nie wpływają inne zmienne. Do wyznaczenia ostatecznych parametrów Johnston i Pattie wykorzystali iteratywny algorytm podwójnego proporcjonalnego skalowania macierzy (*biproportional matrix scaling*). Procedura ta jest podobna do rozwiązania problemu minimalizacji kosztów transportu ludzi czy towarów pomiędzy różnymi miejscami.

Wady i zalety metod wnioskowania ekologicznego

Po omówieniu sześciu różnych podejść warto na koniec odnieść się do wymogów stawianych metodom wnioskowania ekologicznego przez Grofmana i Merrilla (2002b). Każda metoda powinna spełniać trzy główne założenia: 1) realność wyników (*feasibility*), 2) kompletność rozwiązania (*completness*) i 3)

spójność (*empirical coherence*), oraz dziewięć dodatkowych: 4) czułość (*sensitivity*), 5) trafność wyników (*substantive plausibility*), 6) oszczędność założeń (*parsimony*), 7) powtarzalność (*replicabilty*), 8) interpretowalność (*characterizability*), 9) diagnozowalność (*diagnosticability*), 10) możliwość rozszerzenia zastosowań (*expansibility*), 11) łatwość wyjaśnienia modelu (*explainability*) i 12) łatwość obliczeń (*calculatability*).

Metody wnioskowania ekologicznego powinny dawać logiczne rezultaty (realność wyników). Warunek ten nie podlega wątpliwości. Przede wszystkim szacowane parametry muszą przyjmować wartości z przedziału ⟨0,1⟩. Kompletność rozwiązania odnosi się do możliwości oszacowania zarówno parametrów dla poszczególnych obwodów, jak i dla całego okręgu. Natomiast spójność oznacza możliwość uzyskania z odpowiednio ważonych parametrów obwodowych wartości parametrów dla całego okręgu, które powinny być zbliżone do znanych wartości, pochodzących z danych oficjalnych.

Metody wnioskowania ekologicznego charakteryzuje odmienna elastyczność, która polega na dostosowaniu metod obliczania parametrów do dostępnych danych. W przypadku regresji Goodmana czy modelu sąsiedzkiego Freedmana arbitralne założenia nie umożliwiają wzmocnienia precyzji oszacowania o znajomość rozkładów brzegowych dla całego okręgu. Model Kinga i modele Grofmana i Merrilla nie wymagają żadnych arbitralnych założeń i szacują parametry na podstawie informacji na temat rozkładu cech we wszystkich obwodach składających się na dany okręg. W ten sposób uzależniają ostateczne wyniki wnioskowania od konkretnych danych.

Ważną cechą metod wnioskowania z danych zagregowanych powinna być trafność wyników – możliwość potwierdzenia uzyskanych oszacowań parametrów indywidualnych na podstawie rzeczywistych danych, dotyczących jednostek (np. wyników sondaży panelowych). Niestety nie zawsze dysponujemy takimi informacjami, dlatego Grofman i Merrill sugerują możliwość porównywania otrzymanych wartości szacowanych parametrów z rezultatami osiągniętymi przez zastosowanie innych metod wnioskowania ekologicznego, które zostały porównane z danymi rzeczywistymi i które potwierdziły ich wiarygodność. Trafność wnioskowania ekologicznego jest uzależniona od kontekstu i celu badań. W jednych zastosowaniach może dać trafne wyniki zgodne z rzeczywistymi zachowaniami indywidualnymi, a w innych nie.

Metody wnioskowania ekologicznego powinny być oszczędne co do stosowanych założeń. Nie zawsze jednak przyjęcie tylko kilku kryteriów daje dobre wyniki, zwłaszcza, jeśli są one zbyt arbitralne i rygorystyczne. W tym sensie oszczędność założeń jest silnie związana z wymogiem elastyczności przyjętej metody.

Kolejny wymóg – powtarzalność – określa rzetelność metod wnioskowania ekologicznego. Ogólnie można powiedzieć, że podejścia bazujące na metodach regresyjnych (liniowych lub logitowych) są rzetelne. Wyjątek stanowi tu procedura przyjęta przez Kinga, w której parametry jednostkowe obliczane są przy wykorzystaniu symulacji matematycznej dającej stochastyczne wyniki.

Tabela 5. Wymogi stawiane metodom wnioskowania ekologicznego według Grofmana i Merrilla

Wymogi metod wnioskowania ekologicznego	Regresja ekologiczna Goodmana	Sąsiedzki model liniowy Freedmana	Metoda EI Kinga	Modele Grofmana i Merrilla	Model ukrytej struktury Thomsena	Maksymalizacja entropii
Realność wyników	Nie	Tak	Tak	Tak	Tak	Tak
Kompletność rozwiązania	Nie	Tak	Tak	Tak	Tak	Tak
Spójność	Nie	Tak	Tak	Tak	Tak	Tak
Czułość	Nie	Nie	Tak	Tak	Tak	Tak
Trafność wyników	Tak	Nie	Tak	Tak	Tak	Tak
Oszczędność założeń	Proste	Proste	Złożone	Średnie	Średnie	Średnie
Powtarzalność	Tak	Tak	Tak[1]	Tak	Tak	Tak
Interpretowalność	Tak	Nie	Tak	Tak[2]	Tak	Tak
Diagnostyka	Tak	Nie	Tak	Nie	Tak	Nie
Wszechstronność zastosowań	Tak	Nie	Tak[3]	Nie	Tak	Tak
Łatwość wyjaśnienia	Proste	Proste	Złożone	Średnie	Złożone	Średnie
Łatwość obliczeń	Proste	Proste	Złożone	Średnie	Złożone	Złożone

[1] Pod warunkiem, że różnice poszczególnych symulacji nie są duże.
[2] Za pomocą metody *bootstrap*.
[3] Tylko w zastosowaniach dwuzmiennowych.

Źródło: Opracowanie własne na podstawie Grofman i Merril 2002b.

Interpretowalność metod wnioskowania z danych zagregowanych polega na możliwości obliczenia wiarygodnych przedziałów zaufania dla szacowanych parametrów. Jest to związane z kolejnym kryterium zaproponowanym przez Grofmana i Merrilla – diagnozowalnością, tj. możliwością oceny przyjętych rozwią-

zań. Poprawna metoda powinna umożliwiać w łatwy sposób – zarówno statystycznie, jak i graficznie – określenie poprawności założeń leżących u jej podstaw. Dobre propozycje metod wnioskowania ekologicznego powinny także dawać możliwość zastosowania w różnych analizach (autorzy zwracają szczególną uwagę na analizy wielozmiennowe).

Na koniec Grofman i Merrill dodają jeszcze dwa kryteria dotyczące praktycznej strony proponowanych rozwiązań, które powinny cechować: łatwość wyjaśniania, bez stosowania skomplikowanych przekształceń matematycznych i wizualizacji graficznych, oraz prostota obliczeń parametrów przy wykorzystaniu istniejących programów.

Podsumowanie omówionych metod zgodnie z wymogami zaproponowanymi przez Merrilla i Grofmana przedstawia tabela 5.

Literatura

Achen, Christopher H. i Philips W. Shively. 1995. *Crosslevel Inference*. Chicago and London: Chicago University Press.

Bernstein, Felix. 1932. *Über eine Methode, de Soziologische und Bevölkerungs-statistische Gliederung von Abstimmungen bei Geheimem Wahlverfahren Statistisch zu Ermitteln*. „Allgemeines Statistische Archiv" 22: 253–256.

Blank, R. 1905. *Die Soziale Zusammensetzung der Sozialdemokratischen Wählerschaft Deutschlands*. „Archiv für Sozialwissenschaft und Sozialpolitik" 20: 507–553.

Duncan, Dudley O. i Beverly Davis. 1953. *An Alternative to Ecological Correlation*. „American Sociological Review" 18: 665–666.

Durkheim, Emile. 1897. *Le Suicide*. Paris: F. Alcan.

Firebaugh, Glenn. 1978. *A Rule for Inferring Individual-Level Relationships from Aggregate Data*. „American Sociological Review" 18 (43): 557–572.

Florczyk, Andrzej i Tomasz Żukowski. 1990. *Nowa geografia polityczna Polski.* W: Lena Kolarska-Bobińska, Piotr Łukasiewicz, Zbigniew W. Reykowski. *Wyniki badań – wyniki wyborów 4 czerwca 1989*. Warszawa: PTS.

Freedman, David A., Stephen P. Klein, Jerome Sacks, C. Smyth i C. Everett. 1991. *Ecological Regression and Voting Rights*. „Evaluation Review" 15(6): 673–711.

Goodman, Leo. 1953. *Ecological Regression and Behavior of Individuals*. „American Sociology Review" 18: 663–664

Gougel, François. 1955. *Comment votent les Français*. „French Review", Febuary.

Grofman, Bernard i Samuel Merril. 2002a. *Ecological Regression and Ecological Inference*, presentation at the Ecological Inference Conference, June 17–18, 2002, Harvard University, Cambridge, MA, http://course.wilkes.edu/Merrill/.

Grofman, Bernard i Samuel Merril. 2002b. *What Does it Means to Offer a "Solution" to the Problem of Ecological Inference?*, presentation at the Ecological Inference Conference, June 17–18, 2002, Harvard University, Cambridge, MA, http://course.wilkes.edu/Merrill/

Herron, Michael C. i Kenneth W. Shotts. 2002. *Logical Inconsistency in King-based Ecological Regressions*, presentation at 2000 Annual Meeting of the American Political Science Association.

Hofinger, Christoph i Günther Ogris. 2000. *Die Analyse der Wählerströme bei der Nationalratswahl 1999*. „SWS-Rundschau (40. Jahrgang)" Heft 2/2000: 125–139.

Jargowsky, Paul A. 2004. *The Ecological Fallacy.* W: Kimberly Kempf-Leonard. *The Encyclopedia of Social Measurement*. Academic Press, s. 715–722.

Johnston, Ron i Charles Pattie. 2000. *Ecological Inference and Entropy-Maximizing: An Alternative Estimation Procedure for Split-Ticket Voting*. „Political Analysis" 8 (4): 333–345.

King, Garry. 1997. *A Solution to the Ecological Inference Problem. Reconstruction Individual Behavior From Aggregate Data*. Princeton: Princeton University Press.

King, Garry. 1999. *The Future of Ecological Inference Research: A Reply to Freedman et al.* „Journal of the American Statistical Association" 94 (445): 352–355.

Ostrowski, Krzysztof i Adam Przeworski. 1996. *The Structure of Partisan Conflict in Poland.* W: A. Jasińska-Kania, J. Raciborski (red.), *Naród–Władza–Społeczeństwo. Eseje poświęcone prof. Jerzemu J. Wiatrowi*. Warszawa: Wydawnictwo Naukowe Scholar, s. 185–206.

Ogburn, William F. i Inez Goltra. 1919. *How Women Vote: A Study of an Election in Portland, Oregon.* „Political Science Quarterly" 34: 413–438.

Raciborski, Jacek. 1996. *An Outline of the Electoral Geography of the Polish Society.* W: Jerzy J. Wiatr (red.), *Polish Sociology and Democratic Transformation in Poland.* Warszawa: Wydawnictwo Scholar.

Robinson, William S. 1950. *Ecological Correlations and the Behavior of Individuals*, „American Sociological Review" 15: 351–357.

Schuessler, Alexander A. 1999. *Ecological inference*. „Proceedings of National Academy of Science" 96 (19),), http://www.pnas.org/cgi/content/full/96/19/10578.

Thomsen, Soren R.1987. *Danish Election 1920–79. A Logit Approach to Ecological Analysis and Inference*. Aarchaus: Politica.

Thomsen, Soren R. 2001. *Logit method for ecological analysis and inference.* PART-COM MULTILEVEL Project working paper, http://www.ps.au.dk/srt/Ecology.htm.

Thomsen, Soren R. 2003. *Ecological unstandardised factor and pooled regression analysis*. PART-COM Multilevel Project working paper, http://www.ps.au.dk/srt/Ecology.htm.

Zarycki, Tomasz i Andrzej Nowak. 2000. *Hidden dimensions: the stability and structure of regional political cleavages in Poland*. „Communist and Post-Communist Studies" 33 (3): 331–354.

ECOLOGICAL ANALYSES IN ELECTORAL RESEARCH

In social sciences we often have to do with aggregated data which pose both methodological and practical problems. Erroneous inference of individual characteristics and behaviors from this kind of data, described and named by Robinson as "ecological fallacy", has been known for more than fifty years now. Seve-

ral statistical approaches were developed to deal with the problem and produce valid and reliable results. This paper provides on overview of the most popular and frequently used methods of ecological inference. First one is ecological regression as developed by Goodman in a direct answer to Robinson's article. The last one is quite new entropy maximization approach. The purpose of paper is to give Polish reader a review of useful techniques in analysis of aggregated data along with their pros and cons.

Key words: ecological analysis, agregate data analysis, ecological regression.