

Zbyszko Chojnicki, Teresa Czyż

Zastosowanie korelacji cząstkowej w analizie przyczynowej

1. Wstęp

Celem pracy jest przedstawienie modelu przyczynowej interpretacji zjawisk społeczno-ekonomicznych, opartego na koncepcji H. A. Simona (1957) i H. M. Błałocka (1964).

Nie wdając się w szerszą analizę pojęcia przyczynowości przyjmuje się w kontekście niniejszych rozważań definicję przyczynowości sformułowaną przez H. A. Simona (1957): przyczynowość jest asymetryczną relacją między pewnymi zmiennymi lub podzbiorami zmiennych w pewnej strukturze zależności. Jednokierunkowość związku przyczynowego, czyli jego asymetria wyraża jednostronną zależność skutku od przyczyny, tzn. jeśli $C \rightarrow D$ to $D \nrightarrow C$. Ujęcie przyczynowe można więc utożsamiać z uporządkowaniem zależności według takiej asymetrycznej relacji.

Interpretacja przyczynowa zależności jest zadaniem trudnym. Istnieje dość szeroko przyjęty pogląd, że najlepszą metodą sprawdzenia czy między zjawiskami zachodzą związki przyczynowe jest metoda eksperymentalna. Jak stwierdza E. Greenwood (1945, s. 28): „Eksperyment jest dowodem hipotezy, która szuka połączeń między dwoma czynnikami w związku przyczynowym, przez badanie kontrastowych sytuacji, które były kontrolowane względem wszystkich czynników, z wyjątkiem jednego, będącego przedmiotem zainteresowania, który jest hipotetyczną przyczyną albo hipotetycznym skutkiem”. Eksperyment pozwala izolować oddziaływanie poszczególnych czynników na pewne zjawiska. Sterowanie poszczególnymi czynnikami-przyczynami umożliwia w warunkach eksperymentu kontrolowanego wywołanie wystąpienia lub zmiany zjawiska-skutku oraz stwierdzenie niezależności tego skutku od innych niekontrolowanych czynników.

W naukach obserwacyjnych o wysokim stopniu kompleksowości zjawisk, takich jak geografia, napotyka się znaczne trudności przy ustalaniu związków przyczynowych. Poszczególne sytuacje muszą być badane tak jak występują w rzeczywistości, nie zaś w kontrolowanych warunkach. W warunkach rzeczywistych wielka liczba czynników oddziaływa jednocześnie i badacz musi uwzględnić nie tylko wpływ każdego z nich, ale i ich wzajemne oddziaływanie, które nie miałyby miejsca, gdyby każdy z nich oddziaływał odrębnie. Bada się system niedostatecznie izolowany i stąd zachodzi oddziaływanie dużej liczby zewnętrznych czynników.

2. Rekurencyjny model przyczynowy

Modele wielorównaniowe, przedstawiające zależności między wieloma zmiennymi, dzielimy na dwie podstawowe klasy: modele o równaniach współzależnych i modele rekurencyjne.

Model o równaniach współzależnych można zapisać w postaci:

$$\begin{cases} x_1 = a_1 + b_{12}x_2 + b_{13}x_3 + \dots + b_{1k}x_k + e_1, \\ x_2 = a_2 + b_{21}x_1 + b_{23}x_3 + \dots + b_{2k}x_k + e_2, \\ \vdots \\ x_k = a_k + b_{k1}x_1 + b_{k2}x_2 + \dots + b_{k,k-1}x_{k-1} + e_k. \end{cases} \quad (1)$$

Model ten przedstawia sytuację, w której dopuszcza się wzajemne oddziaływanie na siebie zmiennych objaśnianych w równaniach modelu. Występuje zjawisko wielostronnych współzależności między zmiennymi objaśnianymi, czego matematycznym wyrazem jest, że jeśli w pewnym równaniu modelu przyjmujemy, iż zmienna x_i jest funkcją wartości zmiennej x_j ($i \neq j$) to w innym równaniu tego modelu zmienna x_j może być przedstawiona jako (między innymi) funkcja zmiennej x_i . Model o równaniach współzależnych opiera się więc na założeniu, że wartość każdej zmiennej x_i może mieć wpływ na wszystkie pozostałe $(k-1)$ zmienne systemu. Formalnie sprowadza się to do twierdzenia, że macierz współczynników stojących przy zmiennych współzależnych w poszczególnych równaniach jest macierzą dowolną, nieosobliwą. Należy jeszcze zaznaczyć, że składnik e_i w poszczególnych równaniach modelu oznacza łączny efekt oddziaływania na zmienną zależną x_i wszystkich innych czynników, które nie zostały uwzględnione jako zmienne objaśniające.

W przeciwieństwie do modeli równań współzależnych modele rekurencyjne nakładają ograniczenia na budowę równań występujących w modelu i na postać macierzy współczynników przy zmiennych. Jeśli model jest rekurencyjny, to między zmiennymi istnieją tylko powiązania jednokierunkowe. Wymaga się, aby macierz współczynników modelu była macierzą trójkątną, tj. wszystkie elementy znajdujące się nad główną przekątną macierzy współczynników (tzn. elementy b_{ij} , dla których $i < j$) bądź wszystkie elementy pod główną przekątną były równe zero. Trójkątny charakter macierzy współczynników oznacza, że jakieś b musi być równe zero i jeśli $b_{ij} \neq 0$, to wtedy $b_{ji} = 0$.

Model rekurencyjny przybiera zatem postać:

$$\begin{cases} x_1 = e_1, \\ x_2 = b_{21}x_1 + e_2, \\ x_3 = b_{31.2}x_1 + b_{32.1}x_2 + e_3, \\ x_4 = b_{41.23}x_1 + b_{42.13}x_2 + b_{43.12}x_3 + e_4, \\ \vdots \\ x_k = b_{k1.23\dots,k-1}x_1 + b_{k2.13\dots,k-1}x_2 + \dots + b_{k,k-1.1,2\dots,k-2}x_{k-1} + e_k. \end{cases} \quad (2)$$

Model ten zawiera łańcuch kolejnych zależności między zmiennymi, z których każda następna jest określona przez ciąg poprzednich zmiennych. Zmienna x_1 jest niezależna od pozostałych zmiennych systemu. Zmienna x_2 zależy od zmiennej x_1 . Zmienna x_3 zależy od zmiennej x_1 i x_2 , a zmienna x_k zależy od wszystkich pozostałych zmiennych x_i . W modelu rekurencyjnym wykluczona jest więc taka sytuacja, by zmienna x_j objaśniała w jednym równaniu x_i , a w innym sama była (między innymi) objaśniana przez zmienną x_i .

Spróbujmy uchwycić różnicę między modelami współzależnymi i rekurencyjnymi na prostym przykładzie numerycznym opracowanym przez H. M. Białocka (1964). Spośród trzech następujących układów równań równoważnych matematycznie

$$\begin{cases} x = 5, \\ x + y = 6, \end{cases} \quad (3)$$

$$\begin{cases} x - y = 4, \\ x + y = 6, \end{cases} \quad (4)$$

$$\begin{cases} x = 5, \\ y = 1, \end{cases} \quad (5)$$

układ (3) ma charakter rekurencyjny, układ (4) jest współzależny, a układ (5) wykazuje niezależność x i y . Jeżeli w układzie (3) dokona się zmiany pierwszego równania (wprowadzając stałą równą 4), wówczas następuje zmiana zarówno x , jak i y ($x=4$, $y=2$). Jeżeli zmienimy stałą w drugim równaniu tego układu (np. na liczbę 8), to wówczas następuje zmiana wartości y ($y=3$), natomiast nie ma to wpływu na x . Ten układ ma postać rekurencyjną. Natomiast w układzie (4) zmiana stałej w każdym równaniu powoduje zarówno zmianę wartości x , jak i y . Układ ten jest współzależny i charakteryzuje się symetrią, której poprzedni układ nie posiadał. Wreszcie zmiana wartości jednej ze stałych w układzie (5) nie ma wpływu na drugą stałą.

Rozróżnienie między modelami o równaniach współzależnych a równaniami rekurencyjnymi jest bardzo istotne ze względu na ich odmienne własności matematyczno-statystyczne. Zagadnienie to szczegółowo omawia Z. Pawłowski (1971).

Parametry modeli o równaniach współzależnych nie mogą być szacowane za pomocą metody najmniejszych kwadratów. Natomiast metoda ta jest zasadniczo poprawna, gdy chodzi o estymację parametrów modeli rekurencyjnych. Parametry każdego z równań modelu rekurencyjnego można szacować oddzielnie, tzn. w oderwaniu od pozostałych równań modelu. Każde równanie modelu rekurencyjnego traktuje się przy pewnych ogólnych założeniach podobnie jak regresję.

Równania układu rekurencyjnego dają się interpretować w sposób przyczynowy w tym sensie, że zmienne objaśniające spełniają rolę przyczyn określających (z dokładnością do wielkości składnika e_i) zmienną zależną i posiadają odpowiedniki w pewnych sytuacjach empirycznych.

Budowa modelu przyczynowego na podstawie równania rekurencyjnego przebiega następująco. Badacz wychodzi z pewnej teoretycznej znajomości zależności przyczynowej zjawisk, występujących w konkretnej sytuacji będącej przedmiotem badań. Następnie opierając się na założeniu, że powiązania przyczynowe w danym zbiorze zmiennych można traktować jako równoczesne, formułuje się hipotezę wyrażoną w postaci ogólnej modelu rekurencyjnego zdefiniowanego odpowiednim układem równań.

Przykładowo w odniesieniu do 5 zmiennych model ten reprezentuje następujący układ równań:

$$\begin{cases} x_1 = e_{21}, \\ x_2 = b_{21}x_1 + e_2, \\ x_3 = b_{31.2}x_1 + b_{32.1}x_2 + e_3, \\ x_4 = b_{41.23}x_1 + b_{42.13}x_2 + b_{43.12}x_3 + e_4, \\ x_5 = b_{51.234}x_1 + b_{52.134}x_2 + b_{53.124}x_3 + e_5. \end{cases} \quad (6)$$

Każde równanie tego modelu wyraża jednostronną zależność przyczynowo-skutkową.

Z kolei nakłada się na podstawie przesłanek teoretycznych określone ograniczenia na współczynniki równań modelu. Jako ograniczenia przyjmuje się zerowanie pewnych współczynników. Przyjęcie założenia, że pewien parametr b_{ij} w równaniu rekurencyjnym (2) jest równy zeru jest równoznaczne założeniu, że w i -tym równaniu modelu zmienna x_j nie występuje jako zmienna objaśniająca, tzn. że zmienna x_j nie jest po prostu przyczyną x_i . Ograniczenia zerujące przyjmują więc formę dodatkowych równań, w których niektóre współczynniki regresji cząstkowej równań modelu rekurencyjnego przyrównuje się do zera. H. M. Blalock równania te nazywa równaniami predykcji.

Wprowadzenie do rozpatrywanego modelu rekurencyjnego 5-zmiennych ograniczeń, ujętych w następujące równania predykcji:

$$b_{32,1} = 0, \quad (7)$$

$$b_{41,23} = 0, \quad (8)$$

$$b_{51,234} = 0, \quad (9)$$

$$b_{53,124} = 0 \quad (10)$$

jest równoznaczne założeniu, że zmienna x_2 nie jest przyczyną x_3 , x_1 nie jest przyczyną x_4 , a zmienne x_1 , x_3 nie są przyczynami x_5 , i prowadzi do następującej postaci tego modelu:

$$\begin{cases} x_2 = b_{21}x_1 + e_2, \\ x_3 = b_{31,2}x_1 + e_3, \\ x_4 = b_{42,13}x_2 + b_{43,12}x_3 + e_4, \\ x_5 = b_{52,134}x_2 + b_{54,123}x_4 + e_5. \end{cases} \quad (11)$$

Procedura testowania poprawności hipotezy przyczynowej na podstawie modelu rekurencyjnego sprowadza się do sprawdzenia, czy spełnione są określone równania predykcji.

Należy dodać, że opierając się na zależności między współczynnikami regresji cząstkowej i współczynnikami korelacji cząstkowej, równania predykcji formułuje się również w postaci założeń o zerowaniu się określonych współczynników korelacji cząstkowej.

3. Współczynnik regresji cząstkowej a współczynnik korelacji cząstkowej

Współczynnik korelacji cząstkowej $r_{ij,k}$ jest miarą korelacji między x_i i x_j przy ustalonym x_k , tzn. gdy eliminuje się wpływ zmiennej x_k . Załóżmy, że badamy związki między trzema zmiennymi: x_1 , x_2 , x_3 . Współczynnik korelacji dwóch z tych cech, przy wyłączeniu oddziaływania cechy trzeciej, nazywamy współczynnikiem korelacji cząstkowej. Współczynnik ten zapisujemy jako $r_{12,3}$, $r_{13,2}$, $r_{23,1}$. Pierwsze dwie liczby jako subskrypty główne oznaczają zmienne, między którymi badamy korelację, natomiast liczby następne jako subskrypty dodatkowe oznaczają zmienne wywierające wpływ na korelację dwóch pozostałych cech, które chcemy w danym przypadku wyeliminować.

Współczynniki korelacji cząstkowej można wyrazić jako funkcję współczynników korelacji całkowitej. Współczynnik $r_{12,3}$ jest określony równaniem

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}. \quad (12)$$

Przykład zastosowania współczynnika korelacji cząstkowej podaje G. Olsson (1965), który w studium z zakresu migracji bada korelację między odległością migracji (x_1), wiekiem migranta (x_2), poziomem dochodu w miejscu pochodzenia migranta (x_3). Uzyskuje $r_{12} = 0,814$, jednak po wyłączeniu oddziaływania zmiennej x_3 , $r_{12,3} = -0,041$, co potwierdza jego hipotezę wstępną o ujemnej korelacji między odległością migracji a wiekiem migranta.

Uogólniając, współczynnik korelacji cząstkowej pierwszego rzędu (tj. z jednym dodatkowym subskryptem) określa się wzorem:

$$r_{ij,k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{1-r_{ik}^2}\sqrt{1-r_{jk}^2}}. \quad (13)$$

Jeżeli x_i i x_j nie są skorelowane ze zmienną x_k (tj. $r_{ik} = 0$), to wtedy powyższy wzór redukuje się do postaci: $r_{ij,k} = r_{ij}$. W przypadku gdy x_j jest maksymalnie skorelowane z x_k nie można obliczyć $r_{ij,k}$, ponieważ $r_{jk} = 1$ i mianownik staje się zerem. Występuje wtedy zagadnienie współliniowości zmiennych niezależnych, współczynniki regresji są wówczas nieoznaczone. Jeżeli $r_{ij,k} = 0$, to wtedy $r_{ij} = r_{ik}r_{jk}$.

Współczynnik korelacji cząstkowej rzędu p jest współczynnikiem o p dodatkowych subskryptach. Współczynnik korelacji cząstkowej drugiego rzędu przyjmuje postać:

$$r_{ij,kl} = \frac{r_{ij,k} - r_{il,k}r_{jl,k}}{\sqrt{1-r_{il,k}^2}\sqrt{1-r_{jl,k}^2}}. \quad (14)$$

H. M. Blalock (1960) stwierdza, że współczynniki regresji cząstkowej można w pewnym sensie interpretować jako współczynniki korelacji cząstkowej. Każdy współczynnik regresji cząstkowej określa wielkość zmian zmiennej zależnej, który można objaśnić zmianami jednej ze zmiennych niezależnych, podczas gdy pozostałe zmienne niezależne utrzymuje się na stałym poziomie. Wyrażne jest więc podobieństwo współczynników regresji cząstkowej do współczynników korelacji cząstkowej. Stąd wzory współczynników regresji cząstkowej okazują się bardzo podobne do wzorów na współczynniki korelacji cząstkowej.

W przypadku regresji dwóch zmiennych relacja między współczynnikiem regresji a współczynnikiem korelacji przedstawia się następująco:

$$b_{12} = r_{12} \frac{\sigma_1}{\sigma_2}, \quad (15)$$

$$b_{21} = r_{12} \frac{\sigma_2}{\sigma_1}.$$

Dla zmiennych standaryzowanych relacja ta przybiera postać

$$b_{12}^* = b_{21}^* = r_{12}, \quad (16)$$

ponieważ wariancje zmiennych standaryzowanych równe są jedności.

W przypadku regresji wielokrotnej relacja między współczynnikiem regresji cząstkowej a współczynnikiem korelacji cząstkowej według G. U. Yule'a i M. G. Kendalla (1958) wyraża się wzorem:

$$b_{ij.kl, \dots, n} = r_{ij.kl, \dots, n} \frac{\sigma_{i.kl, \dots, n}}{\sigma_{j.kl, \dots, n}}, \quad (17)$$

gdzie $\sigma_{i.kl, \dots, n}$, $\sigma_{j.kl, \dots, n}$ są odchyleniami standardowymi rzędu p . Odchylenie standardowe rzędu p daje się wyrazić za pomocą odchylenia standardowego rzędu $(p-1)$ i korelacji rzędu $(p-1)$, tj.

$$\sigma_{1.23, \dots, n} = \sigma_{1.23, \dots, (n-1)} (1 - r_{1n.23, \dots, (n-1)}^2). \quad (18)$$

Przykładowo w równaniu

$$x_1 = a_{1.23} + b_{12.3} x_2 + b_{13.2} x_3 \quad (19)$$

współczynnik regresji $b_{12.3}$ można określić wzorem:

$$b_{12.3} = r_{12.3} \frac{\sigma_{1.3}}{\sigma_{2.3}} = r_{12.3} \frac{\sigma_1 \sqrt{1-r_{13}^2}}{\sigma_2 \sqrt{1-r_{23}^2}} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1-r_{13}^2} \sqrt{1-r_{23}^2}} \frac{\sigma_1 \sqrt{1-r_{13}^2}}{\sigma_2 \sqrt{1-r_{23}^2}} = \frac{r_{12} - r_{13} r_{23}}{1-r_{23}^2} \frac{\sigma_1}{\sigma_2}. \quad (20)$$

Dla zmiennych standaryzowanych powyższy współczynnik regresji cząstkowej pierwszego rzędu można zapisać według J. O. Kima i F. J. Kohouta (1975) jako

$$b_{12.3}^* = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2}. \quad (21)$$

Współczynnik ten jest nieokreślony, jeżeli $r_{23} = 1$, tzn. zachodzi współliniowość zmiennych niezależnych.

Zerowanie się współczynnika regresji cząstkowej pierwszego rzędu, tj. $b_{ij.k}^* = 0$, zachodzi więc wtedy, gdy licznik ułamka jest równy zero, a więc $r_{ij} = r_{ik} r_{jk}$.

Jeżeli porównamy wzory na współczynnik regresji cząstkowej i współczynnik korelacji cząstkowej, tj.

$$b_{12.3}^* = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2}, \quad (22)$$

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1-r_{13}^2} \sqrt{1-r_{23}^2}}, \quad (23)$$

to stwierdzimy, że liczniki ułamków są takie same.

Zarówno zerowanie się współczynnika korelacji cząstkowej, jak i odpowiadającego mu współczynnika regresji cząstkowej zachodzi wtedy, gdy licznik ułamka jest równy zero, a więc

$$r_{12} - r_{13} r_{23} = 0. \quad (24)$$

Dla współczynników wyższego rzędu $b_{ij.kl}^*$ i $r_{ij.kl}$ zerowanie zachodzi wtedy, gdy

$$r_{ij} - r_{ik} r_{jk} - \frac{(r_{il} - r_{ik} r_{kl})(r_{jl} - r_{jk} r_{kl})}{1 - r_{kl}^2} = 0. \quad (25)$$

Analogicznie można wyprowadzić warunek zerowania dla współczynników jeszcze wyższego rzędu. Wynika z tego, że w procedurze testowania modelu przyczynowego równania predykcji mogą być przedstawione w postaci zbioru równań, w których korelacje cząstkowe między zmiennymi, nie połączonymi bezpośrednio w modelu, przyrównuje się do zera. Weryfikacja hipotezy $\rho = 0$ jest więc jednocześnie weryfikacją hipotezy $\beta = 0$ i vice versa.

Ponieważ niestandaryzowane współczynniki regresji są funkcjami jednostek pomiaru danych, w badaniach empirycznych uzasadnione jest formułowanie równań predykcji w kategoriach współczynników korelacji (N. K. Namboodiri, L. F. Carter, H. M. Blalock, 1975).

4. Testowanie istotności współczynników korelacji

Istotność statystyczną współczynnika korelacji całkowitej r sprawdza się przy zastosowaniu testu F (H. M. Blalock, 1960). Statystyka F określona jest wzorem

$$F_{1, N-2} = \frac{r^2}{1-r^2} (N-2), \quad (26)$$

gdzie

N = liczba obserwacji w próbie.

Statystyka ta ma przy założeniu prawdziwości hipotezy $H_0: \rho = 0$ rozkład F o 1 i $(N-2)$ stopniach swobody. Z tablicy rozkładu F dla ustalonego z góry poziomu istotności α odczytuje się wartość krytyczną F_α . Jeżeli z porównania obliczonej wartości F z wartością krytyczną F otrzyma się nierówność $F > F_\alpha$, to hipotezę H_0 o braku korelacji między zmiennymi można odrzucić.

Dla sprawdzenia czy kilka współczynników korelacji między zmiennymi (otrzymanych tylko z jednej próby) różni się istotnie operuje się rozkładem t . Gdy testuje się hipotezę $H_0: \rho_{xy} = \rho_{zy}$, to buduje się statystykę t w następujący sposób:

$$t = (r_{yx} - r_{zy}) \sqrt{\frac{(N-3)(1+r_{xz})}{2(1-r_{xy}^2 - r_{xz}^2 + 2r_{xy}r_{xz}r_{zy})}}. \quad (27)$$

Z tablicy rozkładu t dla określonego poziomu istotności α i $(N-3)$ stopni swobody odczytuje się wartość krytyczną t_α . Gdy $|t| < t_\alpha$, to nie ma podstaw do odrzucenia hipotezy zerowej o braku różnicy między współczynnikami korelacji.

Test istotności statystycznej współczynnika korelacji cząstkowej jest oparty na statystyce F określonej wzorem:

$$F_{1, N-k-1} = \frac{r_{ij.mn}^2}{1-r_{ij.mn}^2} (N-k-1), \quad (28)$$

gdzie

$k+1$ = liczba zmiennych.

Można zauważyć, że jeśli wartości F , N i k są znane, to przekształcenie równania (26) lub (28) ze względu na r (korelacja całkowita) albo $r_{ij.mn}$ (korelacja cząstkowa) prowadzi do określenia wartości krytycznej współczynnika przy założonym poziomie istotności α .

5. Zasady testowania adekwatności modeli przyczynowych

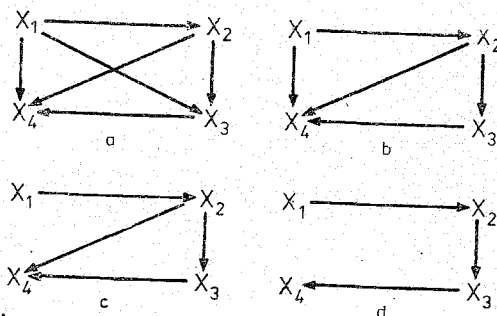
Zasady testowania modeli przyczynowych opartych na równaniach rekurencyjnych zostały opracowane przez H. M. Blalocka (1962).

Przedstawimy niektóre z nich analizując model systemu złożonego z 4 zmiennych w czterech sytuacjach.

Sytuacja 1. Model równań rekurencyjnych w odniesieniu do czterech zmiennych reprezentuje układ ¹⁾:

$$\begin{cases} x_1 = e_1, \\ x_2 = b_{21}x_1 + e_2, \\ x_3 = b_{31.2}x_1 + b_{32.1}x_2 + e_3, \\ x_4 = b_{41.23}x_1 + b_{42.13}x_2 + b_{43.12}x_3 + e_4. \end{cases} \quad (29)$$

Relacje wyrażone w równaniach przedstawia rycina 1a. Strzałki wskazują kierunek przyczynowego uwarunkowania. A więc przy założeniu, że wszystkie współczynniki regresji w równaniach rekurencyjnych są różne od zera otrzymujemy model, w którym zachodzą wszystkie możliwe połączenia przyczynowe.



Ryc. 1. Modele zależności przyczynowych w układzie 4 zmiennych. Źródło: H. M. Blalock (1964)

Sytuacja 2.* Usunięcie połączenia między x_1 i x_3 (ryc. 1b) implikuje, że $b_{31.2}=0$ i stąd równanie predykcji przybiera postać: $r_{13.2}=0$, co interpretujemy: jeśli wyłączamy oddziaływanie x_2 , to korelacja między x_1 i x_3 równa się zero. Jeśli wprowadzimy w tym równaniu predykcji jako zmienną kontrolną dodatkowo x_4 , to w konsekwencji współczynnik korelacji cząstkowej przyjmuje teraz postać

$$r_{13.24} = \frac{r_{13.2} - r_{14.2}r_{34.2}}{\sqrt{1-r_{14.2}^2}\sqrt{1-r_{34.2}^2}}. \quad (30)$$

Ponieważ $r_{13.2}=0$, to $r_{13.24}$ będzie zerować się tylko wtedy, jeśli albo $r_{14.2}=0$ albo $r_{34.2}=0$, co nie zachodzi w modelu (ryc. 1b). Stąd konkluduje się, że dodatkowa zmienna x_4 nie prowadzi w tym przypadku do zerowania się współczynnika korelacji cząstkowej.

W związku z tym należy sformułować następującą zasadę: kontrolę zależności przeprowadzamy tylko ze względu na zmienne, które są albo zmiennymi-poprzednikami w stosunku do jednej lub dwu zmiennych badanej relacji albo zmiennymi-interweniującymi między nimi.

¹⁾ W rozdziałach 2, 5, 6 pracy symbol $b_{ij.k}$... oznacza współczynnik regresji cząstkowej dla zmiennych standaryzowanych.

Sytuacja 3. Załóżmy z kolei, że usuwamy następne połączenie między x_1 i x_4 . Teraz mamy dwa równania predykcji: $r_{13.2}=0$ i $r_{14.23}=0$, które powinny być spełnione (ryc. 1c). Można udowodnić, że w równaniu dotyczącym zerowania się współczynnika korelacji cząstkowej między x_1 i x_4 nie jest konieczne wprowadzenie jako zmiennej kontrolnej – zmiennej x_3 . Analiza wzoru

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}r_{34.2}}{\sqrt{1-r_{13.2}^2}\sqrt{1-r_{34.2}^2}} \quad (31)$$

ukazuje, że para równań predykcji $r_{13.2}=0$ i $r_{14.2}=0$ automatycznie implikuje zerowanie się współczynnika $r_{14.23}$. Zauważmy, że tego rodzaju sytuacja zachodzi wówczas, gdy mamy dwa równania predykcji, zawierające współczynniki korelacji cząstkowej pierwszego rzędu stosujące tę samą zmienną kontrolną (w tym przypadku x_2). W modelu połączenie między x_2 i x_4 ma zarówno charakter bezpośredni, jak i pośredni poprzez x_3 . Numeryczna wartość r_{24} będzie więc odzwierciedlać zarówno wpływ x_3 , jak i x_2 . Nie potrzeba więc wprowadzać zmiennej kontrolnej x_3 w badaniu relacji x_1 i x_4 .

Należy zaznaczyć, że powyższa zasada symplifikacji nie musi być stosowana. Jednak operowanie tą zasadą jest dość wygodne ze względu na możliwość redukcji rzędu współczynnika.

Sytuacja 4. Z kolei usuwamy połączenie między x_2 i x_4 . W ten sposób otrzymujemy prosty łańcuch przyczynowy, w którym x_1 jest przyczyną x_2 , x_2 – przyczyną x_3 , a x_3 – przyczyną x_4 (ryc. 1d). Zgodnie z ogólną zasadą określamy trzy równania predykcji dla tego modelu: $r_{13.2}=0$, $r_{14.23}=0$, $r_{24.13}=0$. Można dowieść, że zarówno $r_{24.13}$, jak i $r_{24.3}$ zerują się. To pozwala wyprowadzić następującą zasadę praktyczną: gdy zmienna x_1 działa bezpośrednio tylko na jeden element zbioru pozostałych zmiennych (w tym przypadku x_2 , x_3 , x_4), to można pominąć jej wpływ na zerowanie się współczynników korelacji między pozostałymi zmiennymi.

Przegląd pozostałych prostych zasad testowania modeli 4-zmiennych zawarty jest w pracy H. M. Blalocka (1962).

Należy dodać, że można sformułować jeszcze dalsze zasady, które mają charakter czysto pragmatyczny związany z innymi typami sytuacji i innymi systemami zmiennych.

6. Modele empiryczne

Pierwszą próbę zastosowania modelu analizy przyczynowej Simona-Blalocka w geografii podjął K. R. Cox (1968). Uwagi krytyczne na marginesie tej pracy sformułował P. Taylor (1969).

Wzorcowy przykład zastosowania modelu analizy przyczynowej Simona-Blalocka stanowi badanie zależności między kształtowaniem się wskaźnika ruchliwości przestrzennej ludności a innymi czynnikami społeczno-ekonomicznymi, przeprowadzone przez E. G. Moore'a (1969). Dane dotyczyły 193 jednostek spisowych miasta Brisbane (Australia) w 1961 roku.

Zmienną wyrażoną jako zmienna zależna w modelu jest wskaźnik ruchliwości przestrzennej ludności (y), który określa się jako odsetek ludności zmieniającej miejsce zamieszkania.

Formułowanie modelu przyczynowego dokonuje się w dwóch etapach przez: 1) identyfikację zmiennych niezależnych, 2) określenie przyczynowych asymetrycznych zależności między zmiennymi.

Zmienne niezależne określa się na podstawie doświadczeń badawczych P. H. Rosiego (1955), który prowadził studia nad mobilnością ludności Filadelfii. Bierze się pod uwagę następujące czynniki – przyczyny migracji wewnątrzmijskich: 1) strukturę wieku, 2) sytuację rodzinną, 3) samodzielność mieszkania, 4) sytuację własnościową mieszkania, 5) ludność miejscową. Czynniki te identyfikuje się z 5 wskaźnikami zawartymi w tabeli 1. Macierz współczynników korelacji tych wskaźników przedstawia tabela 2.

TABELA 1

Wskaźniki społeczne i mieszkaniowe dla miasta Brisbane

Symbol	Nazwa wskaźnika
x_1	Wskaźnik zróżnicowania wieku ludności, tj. (ludność 0 - 14 lat) - (ludność powyżej 65 lat) w stosunku do ogółu ludności
x_2	Odsetek kobiet stanu wolnego, tj. panny powyżej 15 lat, w stosunku do ogółu kobiet
x_3	Udział domków jednorodzinnych w ogólnej liczbie mieszkań
x_4	Odsetek mieszkań własnościowych
x_5	Odsetek ludności urodzonej w Australii

Źródło: E. G. Moore (1969).

TABELA 2

Macierz korelacji

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1,00	-0,53	0,53	-0,72	-0,21	-0,67	-0,70
x_1		1,00	-0,74	0,79	-0,23	0,49	0,82
x_2			1,00	-0,65	-0,13	-0,33	-0,66
x_3				1,00	0,08	0,74	0,86
x_4					1,00	0,21	0,04
x_5						1,00	0,70
x_6							1,00

Źródło: E. G. Moore (1969).

Z kolei określa się hipotetyczne zależności przyczynowe w układzie zmiennych x_1, x_2, x_3, x_4, x_5 . Takie postępowanie prowadzi do wyodrębnienia 7 podstawowych układów powiązań (tab. 3). Postuluje się, że bezpośredni wpływ na kształtowanie się wskaźnika ruchliwości przestrzennej ludności (y) wywierają następujące zmienne: wskaźnik domków jednorodzinnych (x_3), odsetek ludności urodzonej w Australii (x_5), odsetek mieszkań własnościowych (x_4), odsetek samotnych kobiet (x_2).

Analiza relacji bezpośrednich i pośrednich prowadzi do sformułowania modelu przyczynowego ruchliwości ludności w obszarze intramiejskim, przedstawionego na rycinie 2. Model ten testuje się na podstawie równań predykcji określonych przy założeniu, że współczynniki korelacji cząstkowej między zmiennymi, dla których nie postulowano połączeń, powinny być równe zeru. Estymowane wartości współczynników korelacji porównuje się z wartościami obserwowanych korelacji (patrz tab. 4). Wielkości różnic tych wartości traktuje się jako miernik adekwatności modelu.

TABELA 3

Układy powiązań

Relacja	Interpretacja relacji
$x_3 \rightarrow x_2$	Ujemna zależność między wskaźnikiem domków jednorodzinnych (x_3) a odsetkiem kobiet samotnych (x_2) (osoby samotne zajmują z reguły mieszkania w domach wielomieszkaniowych)
$x_1 \rightarrow x_2$	Ujemna zależność między liczbą młodych osób pozostających na utrzymaniu (x_1) a odsetkiem kobiet dorosłych stanu wolnego (x_2)
$x_3 \rightarrow x_1$	Zależność dodatnia między odsetkiem ludności urodzonej w Australii (x_5) a odsetkiem ludzi młodych (x_1) (migranci, szczególnie z krajów niebrytyjskich, to przede wszystkim osoby dorosłe stanu wolnego)
$x_5 \rightarrow x_1 \rightarrow x_2$	Zależność pośrednia
$x_3 \rightarrow x_5$	Zależność między odsetkiem domków jednorodzinnych (x_3) a odsetkiem ludności urodzonej w Australii (x_5), tj. ludności o większym udziale ludzi młodych i wyższych dochodach niż mniejszości etniczne
$x_3 \rightarrow x_5 \rightarrow x_1 \rightarrow x_2$	Zależność pośrednia
$x_3 \rightarrow x_4$	Dodatnia zależność między odsetkiem samodzielnych jednostek mieszkaniowych (x_3) a odsetkiem mieszkań własnościowych (x_4) (domki jednorodzinne są najczęściej własnością osób w nich mieszkających)

TABELA 4

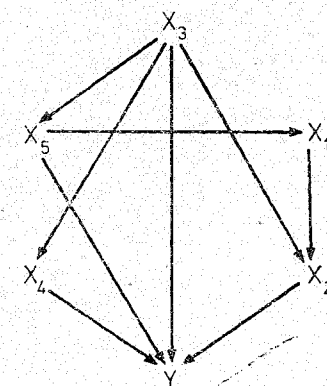
Estymowane i obserwowane wartości współczynników korelacji dla modelu I

Wartość estymowana	Wartość obserwowana	Różnica
$r_{45.3} = 0,06$	$r_{45} = 0,21$	0,15
$r_{24.153} = -0,05$	$r_{24} = -0,13$	0,08
$r_{13.5} = 0,36$	$r_{13} = 0,79$	0,43
$r_{25.13} = -0,42$	$r_{25} = -0,33$	0,09
$r_{14.35} = 0,10$	$r_{14} = -0,23$	0,33
$r_{17.25} = -0,53$	$r_{17} = -0,53$	0,00

Źródło: E. G. Moore (1969).

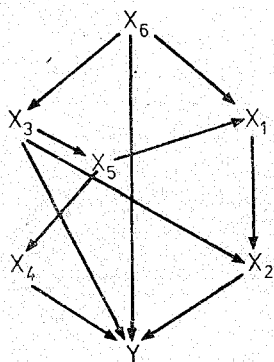
Ponieważ obserwowane korelacje nie są zgodne we wszystkich przypadkach z predykcjami określonymi przez model musi on podlegać pewnym modyfikacjom.

Do modelu wprowadza się dodatkową zmienną x_6 w postaci odległości od centrum miasta (wyrażoną czasem dojazdu publicznymi środkami transportu) sugerując, że do-



Ryc. 2. Model I. Źródło: E. G. Moore (1969)

stępnosc ma nie tylko bezpośredni hamujący wpływ na ruchliwość ludności, ale jest czynnikiem determinującym kształtowanie się innych zmiennych społeczno-ekonomicznych ujętych w modelu. Usuwa się również hipotetyczne połączenie między zmienną x_3 i x_4 . Formuluje się w ten sposób model II, który przedstawia rycina 3.



Ryc. 3. Model II. Źródło: E. G. Moore (1969)

Z kolei dokonuje się testowania tego zmodyfikowanego modelu. Wyniki testowania zawiera tabela 5. Przewidywane korelacje różnią się od obserwowanych o wartość mniejszą niż 0,10, co świadczy o wysokiej adekwatności modelu II.

TABELA 5

Estymowane i obserwowane wartości współczynników korelacji dla modelu II

Wartość estymowana	Wartość obserwowana	Różnica
$r_{56.3} = 0,63$	$r_{56} = 0,70$	0,07
$r_{26.1} = -0,61$	$r_{26} = -0,66$	0,05
$r_{24.15} = -0,17$	$r_{24} = -0,13$	0,04
$r_{46.135} = 0,00$	$r_{46} = 0,04$	0,04
$r_{25.13} = -0,42$	$r_{25} = -0,33$	0,09
$r_{13.56} = -0,39$	$r_{13} = -0,31$	0,08
$r_{34.156} = -0,02$	$r_{34} = 0,08$	0,10
$r_{17.245} = -0,52$	$r_{17} = -0,53$	0,01

Źródło: E. G. Moore (1969).

S. H. Putman (1971) przedstawia zastosowanie modelu analizy przyczynowej Simona-Blałocka do przewidywania zmian w zatrudnieniu, zachodzących pod wpływem zmian warunków transportowych, w ujęciu regionalnym. Wprowadza się 3 zmienne: Y – zmiana w zatrudnieniu w danym sektorze gospodarki, X – zmiana w potencjale rynkowym dla półproduktów, Z – zmiana w potencjale rynkowym dla produktów końcowych.

Zmienne definiuje się w następujący sposób:

$$Y_{i,g} = E_{i,g,t} - E_{i,g,t-1}, \quad (32)$$

$$X_{i,g} = D_{i,g,t} - D_{i,g,t-1}, \quad (33)$$

$$Z_{i,g} = G_{i,g,t} - G_{i,g,t-1}, \quad (34)$$

gdzie

$$D_{i,g,t} = \sum_j \pi_{i,j,g,t} \sum_h \alpha_{gh} B_{j,h,t}, \quad (35)$$

$$G_{i,g,t} = \sum_j \pi_{i,j,g,t} [I_{j,t}], \quad (36)$$

$$\pi_{i,j,g,t} = \frac{[E_{i,g,t}/C_{i,j,t}]}{\sum_k [E_{k,g,t}/C_{k,j,t}]}, \quad (37)$$

przy czym

$E_{i,g,t}$ = zatrudnienie w sektorze g w podregionie i , w czasie t ;

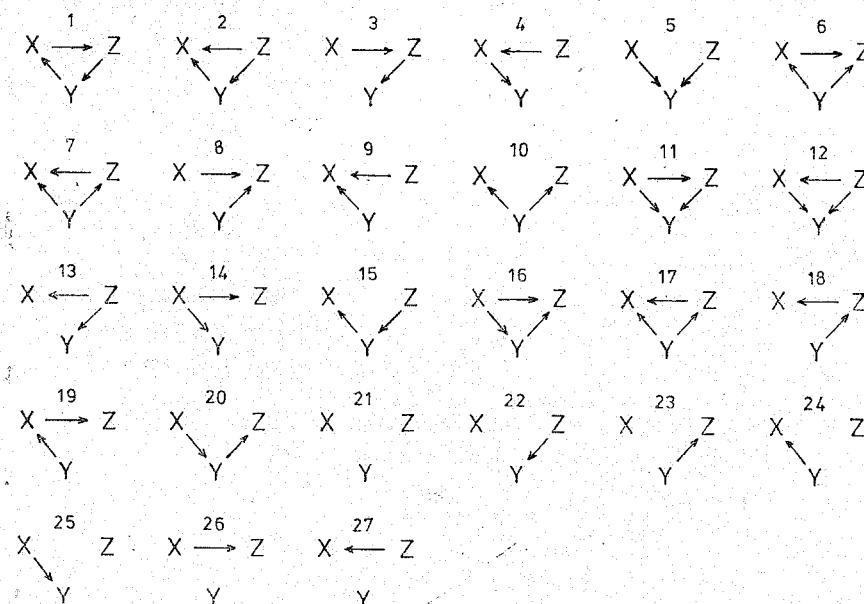
$\alpha_{g,h}$ = ilość jednostek nakładu (*input*) z sektora g potrzebna do wytworzenia jednej jednostki produktu (*output*) w sektorze h ;

$I_{j,t}$ = globalny dochód mieszkańców podregionu j w czasie t ;

$C_{i,j,t}$ = opór transportu (koszt) między podregionem i oraz podregionem j w czasie t .

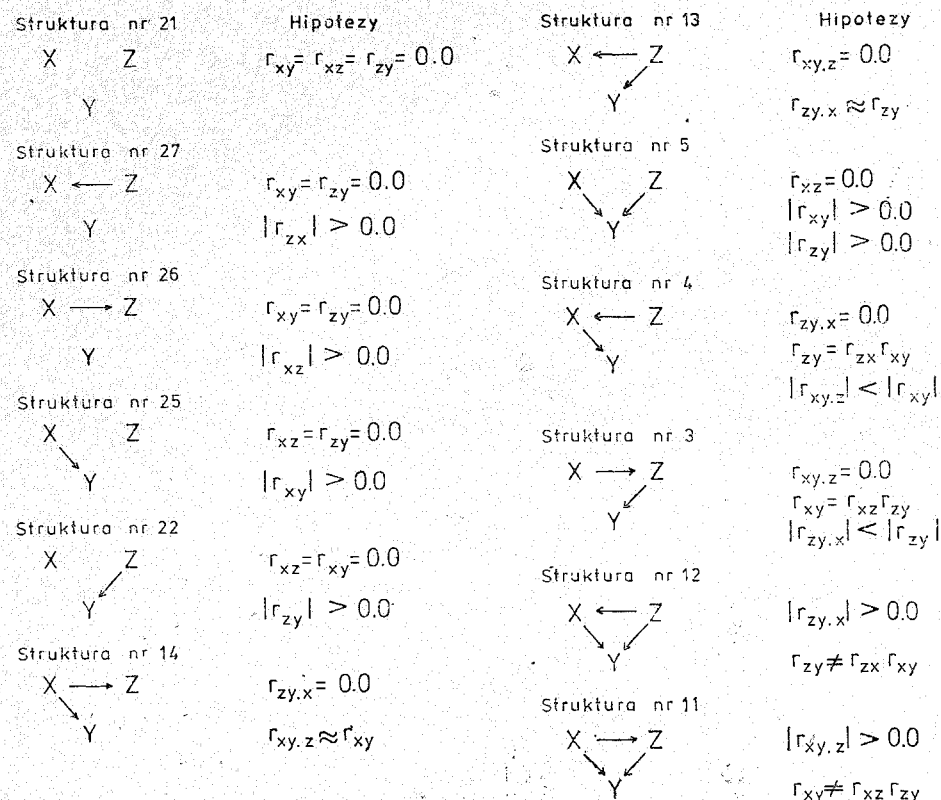
Można więc zauważyć, że zmiana oporu transportu powoduje zarówno zmianę $D_{i,g}$, jak i $G_{i,g}$.

Punktem wyjścia dla budowy modelu jest ogólne założenie, że między zmiennymi X , Y , Z zachodzą asymetryczne relacje. W tej sytuacji stwierdza się, że istnieje 27 możliwych powiązań przyczynowych między tymi zmiennymi (ryc. 4). Ponieważ jednak w badaniu przyjmuje się a priori, że zmienna Y jest zmienną zależną, to liczbę alternatywnych



Ryc. 4. Modele zależności przyczynowych w układzie 3 zmiennych. Źródło: S. H. Putman (1971)

struktur przyczynowych redukuje się w ten sposób z 27 do 12. Rozpatruje się struktury nr 3, 4, 5, 11, 12, 13, 14, 21, 22, 25, 26, 27 (ryc. 5). Postulując, że zmienne X i Z są pojedynczo lub łącznie przyczynami zmiennej zależnej Y , w dalszym postępowaniu odrzuca się jeszcze struktury nr 21, 27, 26, 14, 13, 4, 3. Tak więc przedmiotem analizy pozostają ostatecznie tylko struktury nr 25, 22, 5, 12, 11.



Ryc. 5. Modele zależności przyczynowych w układzie 3 zmiennych przy założeniu, że zmienna Y jest zmienną zależną. Źródło: S. H. Putman (1971)

Formuluje się hipotezę, że korelacje zachodzące między zmiennymi Y i X oraz między zmiennymi Y i Z nie są zależnościami pozornymi. W procedurze testowania oblicza się współczynniki korelacji cząstkowej i całkowitej zamieszczone w tabeli 6.

TABELA 6

Wartości współczynników korelacji

Sektor	r_{xy}	r_{xz}	r_{zy}	$r_{xy.z}$	$r_{yz.x}$	$r_{xz.y}$	$r_{xy.zy}$	Nr struktury
1 – Rolnictwo	0,416	0,533	0,944	-0,312	0,939	0,503	0,246	12,11
2 – Górnictwo	0,971	0,981	0,955	0,523	0,053	0,937	0,953	4,11
3 – Budownictwo	0,911	0,613	0,859	0,950	0,922	0,527	0,558	12,11
4 – Przemysł	0,976	0,165	-0,016	0,992	-0,824	-0,003	0,161	12,11
5 – Transport	0,994	0,885	0,916	0,982	0,713	0,811	0,880	12,11
6 – Użyteczność publiczna	0,989	0,848	0,767	0,996	-0,914	0,650	0,839	12,11
7 – Handel hurtowy	0,991	0,953	0,908	0,990	-0,898	0,865	0,944	12,11
8 – Handel detaliczny	0,987	0,405	0,471	0,987	0,485	0,191	0,399	12,11
9 – Finanse	0,999	0,978	0,973	0,985	-0,431	0,952	0,977	11
10 – Usługi	0,921	0,903	0,996	0,563	0,982	0,899	0,832	12,11
11 – Administracja państwa	0,434	0,687	0,811	-0,290	0,783	0,557	0,298	12,11

Źródło: S. H. Putman (1971).

Wstępna analiza wartości tych współczynników pozwala twierdzić, że nie ma podstaw do odrzucenia hipotetycznych struktur nr 11 i 12 dla poszczególnych sektorów gospodarki. Tylko w przypadku sektora 2 (górnictwo) wyjątkowo możliwa jest, oprócz akceptacji struktury nr 11, również akceptacja struktury nr 4, oznaczającej sytuację pozornej przyczynowości między X i Y.

Przeprowadza się testowanie istotności współczynników korelacji przy zastosowaniu statystyki F (tab. 7). Dla badanej próby wartości $F > 6,8$ dla $\alpha = 0,01$ i $F > 3,9$ dla $\alpha = 0,05$ implikują wartości istotne. Jak wynika z tabeli 7, z pewnymi wyjątkami, korelacje całkowite i cząstkowe są wysoce istotne. Wyjątki stanowią współczynniki $r_{zy.x}$ dla sektora 2 i r_{zy} dla sektora 4.

TABELA 7

Wartość statystyki F

Sektor	Statystyka F				
	r_{xy}	r_{xz}	r_{zy}	$r_{xy.z}$	$r_{yz.x}$
1	26,6	50,4	1039,6	13,6	934,5
2	2094,9	3247,2	1316,5	68,5	0,4
3	619,7	76,5	357,5	1176,0	718,8
4	2550,9	3,55	0,03	8159,3	267,0
5	10488,1	458,9	662,1	3321,6	130,3
6	5677,6	325,1	181,5	14317,7	641,6
7	6960,4	1256,6	596,5	6254,9	525,3
8	4789,6	24,9	36,2	4836,9	38,7
9	63400,3	2791,5	2257,0	3998,7	28,8
10	709,9	561,0	15779,7	58,5	3380,3
11	29,5	113,5	244,0	11,5	200,2

Źródło: S. H. Putman (1971).

W rezultacie takiego postępowania, zdaniem S. H. Putmana, uzasadnione jest twierdzenie, że zmiany w cechach transportu regionalnego powodują zmiany w zatrudnieniu, rozpatrywane w układzie sektorowym dla poszczególnych podregionów.

Kończąc powyższą analizę należy stwierdzić, że przedstawiona metoda Simona-Bla-locka może być zarówno stosowana do wyboru między alternatywnymi modelami przyjętymi a priori, jak i do interpretowania badanych zależności jako związków przyczynowych przez konstruowanie sukcesywnych modeli, które spełniają warunki metodologii Simona-Bla-locka.

Bibliografia

- Blalock H. M., 1960. Social statistics, New York. Tłum. polskie: H. M. Blalock, Statystyka dla socjologów. Warszawa 1975.
- Blalock H. M., 1962. Four-variable causal models and partial correlation. American Journal of Sociology, 68, s. 182 - 194.
- Blalock H. M., 1964. Causal inferences in nonexperimental research. Chapel Hill.
- Cox K. R., 1968. Suburbia and voting behaviour in the London Metropolitan Area. Annals of the Association of American Geographers, 58, 1, s. 111 - 127.
- Greenwood E., 1945. Experimental sociology. New York.

- Kim J. O., Kohout F. J., 1975. Multiple regression analysis: subprogram regression. [W:] N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, D. H. Bent, Statistical package for the social sciences. New York, s. 320 - 368.
- Moore E. G., 1969. The structure of intra-urban movement rates: an ecological model. Urban Studies, 6, s. 17 - 33.
- Namboodiri N. K., Carter L. F., Blalock H. M., 1975. Applied multivariate analysis and experimental design. New York.
- Olsson G., 1965. Distance and human interaction. A migration study. Geografiska Annaler, 47, s. 3 - 43.
- Pawłowski Z., 1971. Modele ekonometryczne równań opisowych. Warszawa.
- Putman S. H., 1971. The effect of changing transport facilities on regional economic development: a case study of causal analysis. The Regional Science Association, Papers, 27, s. 151 - 156.
- Rossi P. H., 1955. Why families move. Glencoe.
- Simon H. A., 1957. Models of man. New York.
- Taylor P., 1969. Causal models in geographic research. Annals of the Association of American Geographers, 59, s. 402 - 404.
- Yule G. U., Kendall M. G., 1958. An introduction to the theory of statistics. London. Tłum. polskie: G. U. Yule, M. G. Kendall, Wstęp do teorii statystyki. Warszawa 1966.