RUCH PRAWNICZY, EKONOMICZNY I SOCJOLOGICZNY
Rok XLIII — zeszyt 4 — 1981

GUILLAUME WUNSCH*

WPŁYW AGREGACJI I WSPÓŁLINIOWOŚCI NA WYNIKI REGRESJI WIELOKROTNEJ W ANALIZIE PRZESTRZENI

W wielu dziedzinach życia ekonomicznego i społecznego informacje, jakimi się rozporządza, są zebrane z geograficznych jednostek administracyjnych (gmin, powiatów), podczas gdy teoria, którą usiłuje się sprawdzić, opiera się na badaniach jednostek prostych. Na przykład rozporządza się współczynnikami przeciętnego spożycia tytoniu w skali powiatu i usiłuje się odkryć na podstawie tych danych ewentualny związek między zatruciem nikotyną a umieralnością. Postępując w ten sposób narażamy się na dojście do błędnego wyniku[1]: teoria postawiona na pewnym poziomie agregacji (w naszym przypadku, jednostka) nie może być poprawnie zweryfikowana we wszystkich przypadkach przez dane ustawione na innym poziomie agregacji (np. powiat). W przypadku dwóch zmiennych (zmienna zależna *y* i niezależna *x),* łatwo jest wykazać[2] na przykład, że współczynnik korelacji *R(xy)* może być napisany w postaci *R(xy)=IR(xy)[A]+ER(xy)[B],* gdzie *IR(xy)* i *ER(xy)* są odpowiednio współczynnikami korelacji wewnątrz i między grupami, a wyrażenia w klamrach są funkcją względnych dyspersji wewnątrz i między grupami zmiennych *x* i *y*. Wysoka korelacja międzygrupowa jest więc zupełnie zgodna ze słabą korelacją wewnątrzgrupową: innymi słowy, można obserwować mocną korelację między powiatami, która nie odzwierciedla wysokiej więzi na płaszczyźnie indywidualnej.

Artykuł ten stara się sprecyzować wpływ agregacji danych, jak również związku

* Profesor Guillunme Wunsch jest kierownikiem Katedry Demografii na Université Catholique de Louvain w Belgii a także (od 1971 r.) profesorem na Uniwersytecie w Montrealu. Zajmuje się zastosowaniem metod statystyczno-matematycznych we współczesnej demografii *(Etude démographique de la nuptialité en Belgique* 1967; *Méthodes d'analyse démographique pour les pays en développement,* 1978; wspólnie z H. Gerard *Comprendre la démographie* 1973; wspólnie z M. G. Termote, *Introduction to Demographic Analysis, Principles and Methods,* 1978).

[1] Jednym z pierwszych autorów, którzy zasygnalizowali ten problem, jest W. S. Robinson, *Ecological correlations and the behavior of individuals,* American Sociological Review 1950, t. 15, s. 351 -357.

[2] Por. H. R. Alker, *Topology of Ecological Fallacies,* w: M. Dogun, S. Rokkan, *Quantitative Ecological Analysis in the Social Sciences,* Cambridge, Mass. 1969.

między zmiennymi niezależnymi lub objaśniającymi (lub współliniowości) w przypadku regresji wielokrotnej. Przypomnijmy, że analiza regresji wielokrotnej ma na celu powiązanie zmiennej zależnej y z różnymi zmiennymi „niezależnymi" x_i według modelu liniowego w rodzaju:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + e,$$

gdzie b_0, b_1, ... są stałymi oszacowanymi na podstawie danych empirycznych, a e jest błędem losowym wynikającym z faktu, że rzeczywistość przystosowuje się w sposób niedoskonały do używanego modelu liniowego. W praktyce stale są określane metodą najmniejszych kwadratów, co zaś do błędu losowego, zakłada się spełnienie różnych warunków, w tym braku autokorelacji[3]. Zaznaczmy przy tyra, że w naszym przypadku chodzi o autokorelację przestrzenną; nie stosuje się wtedy klasycznych testów ekonometrycznych, takich jak test Durbina-Watsona.

I. PIERWSZY MODEL: WSPÓŁLNIOWOŚĆ NIEDOSKONAŁA

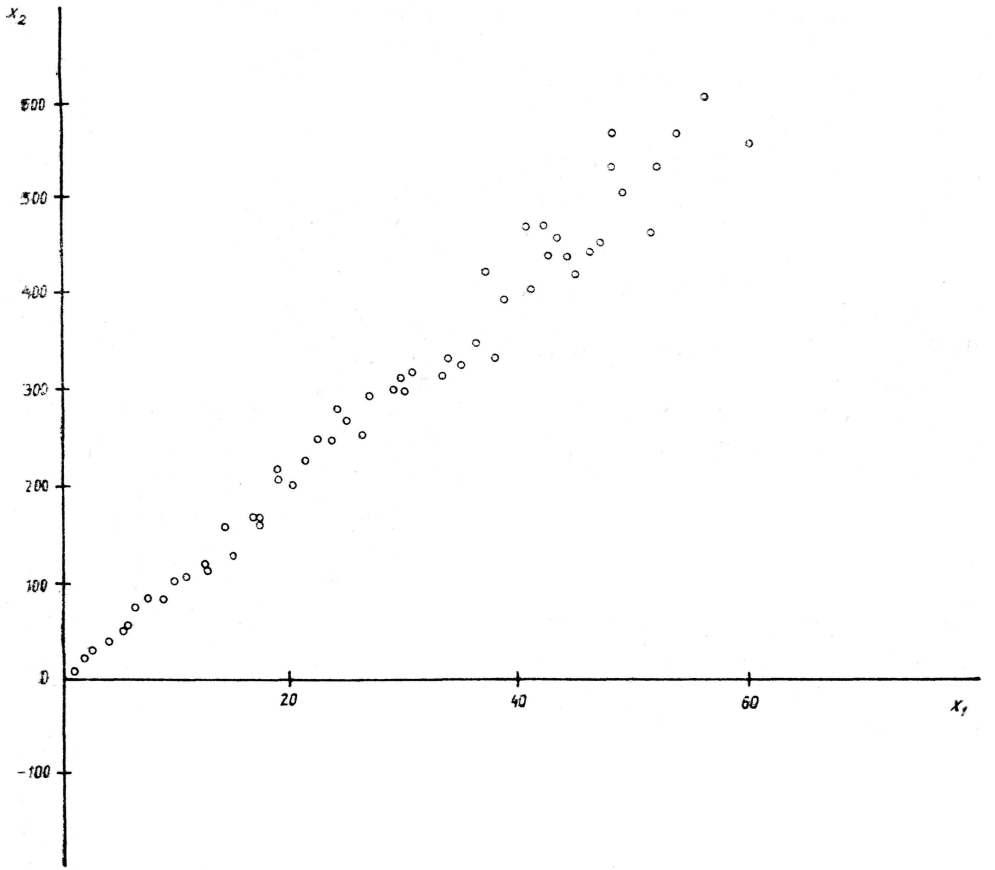
Wpływ agregacji i współliniowości był testowany w sposób następujący: pierwszy model podstawowy (model **A**) został zbudowany za pomocą trzech zmiennych „niezależnych", z których jedna była funkcją innej (patrz załącznik). Zachodzi więc współliniowość doskonała między x_i i regresja nie może być wykonana: współliniowość prowadzi do niemożliwości odwrócenia macierzy $\mathbf{X'X}$ zmiennych niezależnych[4]. Aby uniknąć współliniowości doskonałej, dane modelu **A** zostały zakłócone przez czynnik losowy (model **B**), który tym niemniej utrzymuje związek statystyczny wysoki między zmiennymi x_i (współczynnik korelacji prostej r ma wartość 0,988 między x_1 i x_2, a -0,985 między x_1 i x_3). Rycina 1 przedstawia jako przykład schemat par x_1, x_2.

Ponieważ macierz $\mathbf{X'X}$ nie jest już osobliwa, może być odwrócona i dać rozwiązanie. Podejście „krok za krokiem", przy zatrzymaniu tylko zmiennej x_1, doprowadza do wariancji „objaśnionej" (kwadrat współczynnika R korelacji wielokrotnej) równej 0,987. Zatrzymując wszystkie zmienne, współczynnik korelacji wielokrotnej niewiele się zwiększa (R^2=0,995). Można więc w tym przypadku optować bardzo rozsądnie na relację $\hat{Y}=b_0+b_1 x_1$, pomijając zmienne x_2 i x_3.

Dane zostały następnie przegrupowane, tym razem, by sprawdzić wynik agregacji. W modelu **BC** przegrupowanie jest systematyczne, w modelu **BD** przeciwnie, jest przypadkowe. Dla modelu **BC** (grupowanie systematyczne) zmienna x_1 wyjaśnia na nowo prawie całkowicie zmienność Y(R^2=0,99907 wobec 0,99955 dla trzech x jednocześnie). Przegrupowanie danych usuwa część „czynnika losowego" i to tak dobrze, że relacje między x_i i Y są prawie liniowe. Co się tyczy modelu **BD**

[3] W związku z tym patrz: J. Johnston, *Econometric Methods,* wyd. 2, Tokio 1972, rozdz. V.

[4] Przypomnijmy, że współczynniki regresji b_i otrzymuje się z równania macierzy $\mathbf{B}=(X'X)^{-1}X'Y$, gdzie Y jest wektorem zmiennych zależnych, X jest macierzą zmiennych niezależnych, a **B** jest wektorem szukanych współczynników.

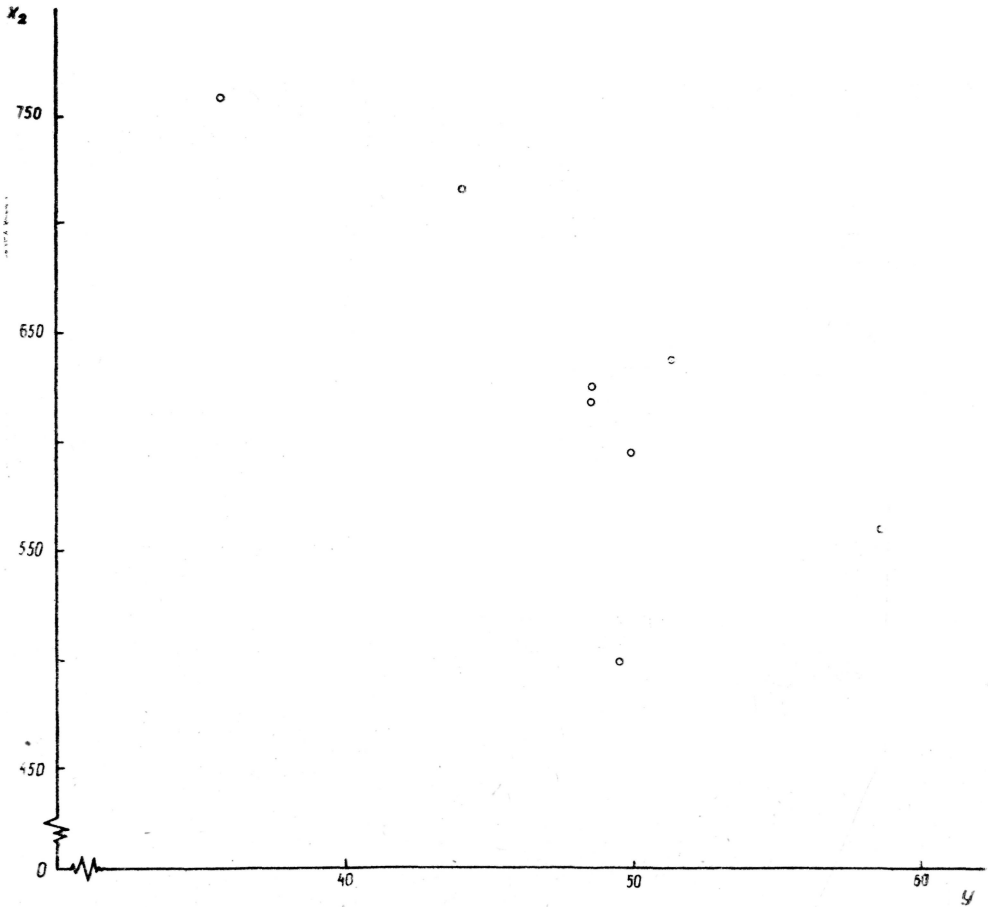


Ryc. 1. Zmienne x_1 (oś pozioma) i x_2 (oś pionowa)

(grupowanie losowe), regresja krokowa wykazuje, że Y jest wystarczająco dobrze oszacowane na podstawie x_2 (R^2 =0,997). Ukazują się więc dobrze związki, mimo przypadkowego przegrupowania danych.

II. MODEL DRUGI: DANE LOSOWE

W modelu **C** trzy serie zmiennych niezależnych x_1 są losowe. Model **CB** rozkłada te dane na 8 klas według losowania. Wychodząc od danych nie przegrupowanych otrzymuje się wariancję objaśnioną (R^2) tylko o 7%, biorąc pod uwagę tylko zmienną niezależną x_3. Dodanie dwóch innych zmiennych powiększa R^2 zaledwie o 0,5% wartości absolutnej. Ten brak korelacji szczęśliwie odzwierciedla przypadkowy charakter modelu. Z modelu **CB** (dane pogrupowane) kwadrat współczynnika korelacji wielokrotnej dochodzi do 60%, przy jednej zmiennej niezależnej (x_2) w równaniu; test istotności R (hipoteza zerowa) wykazuje jednak, że ten wynik jest nieistotny dla wydania sądu o przystosowaniu modelu do rzeczywistości. Badanie relacji graficznej między Y i x_2 (ryc. 2) wykazuje zresztą dobrze niedorzeczność uciekania się do modelu liniowego na podstawie tych danych.

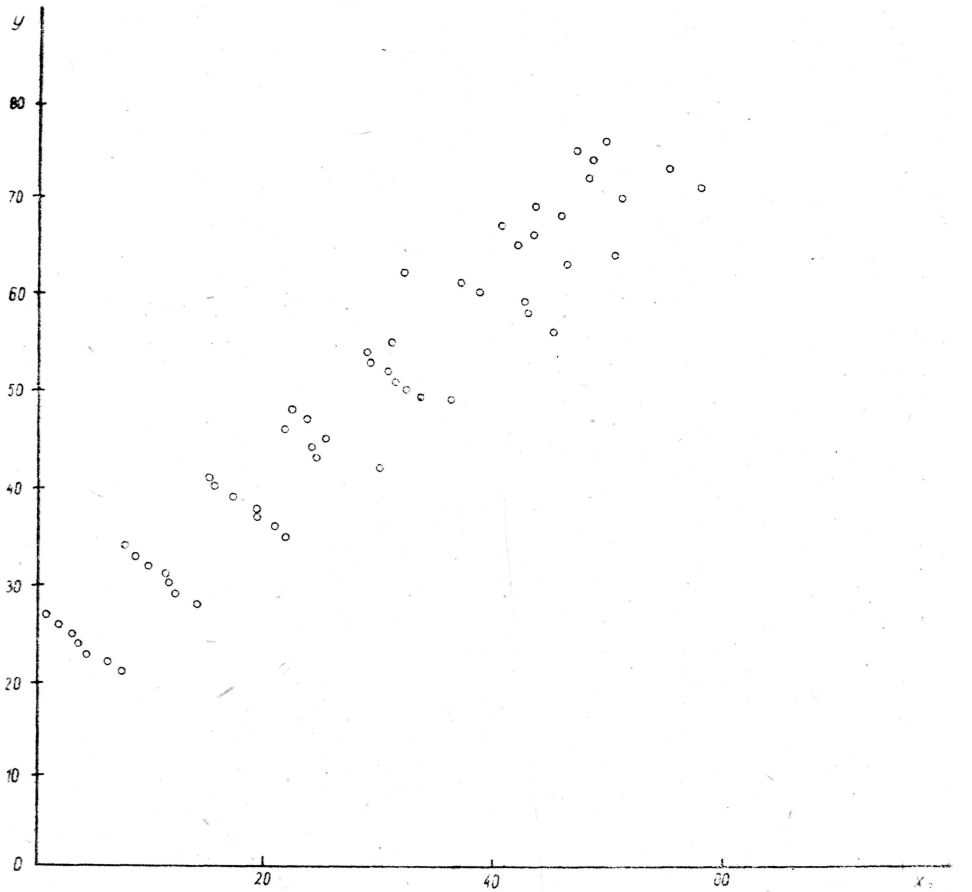


Ryc. 2. Zmienne y (oś pozioma) i x_2 (oś pionowa)

III. MODEL TRZECI: WYNIKI KONTEKSTUALNE

Model trzeci (model **DA**) jest bliski modelu **B**. Jednakże jednostki są rozdzielone na 8 grup takich, że w każdej grupie zmienna zależna jest skorelowana ujemnie z trzema zmiennymi niezależnymi x_i. Rycina 3 przedstawia, jako przykład, związek między Y i x_3. Tak jak w modelu **B**, współliniowość jest bardzo silna między zmiennymi niezależnymi. Ponadto obserwuje się tym razem efekt kontekstualny: relacja w każdej grupie jest przeciwstawna relacji całości. Problem ten traktowałby się więc klasycznie przez analizę kowariancji w braku wzajemnego oddziaływania, jednakże tutaj zwrócimy się znowu do analizy regresji klasycznej.

Regresja krokowa na zbiór danych indywidualnych zatrzymuje istotnie zmienną x_2, jako zmienną objaśniającą, doprowadzając do wariancji objaśnionej R^2= =0,9329. Wprowadzenie dwu innych zmiennych niezależnych zwiększa kwadrat współczynnika korelacji wielokrotnej tylko na trzecim miejsce po przecinku; widać więc, że dwie z vrzech zmiennych są zbyteczne, co zresztą doskonale odpowiada

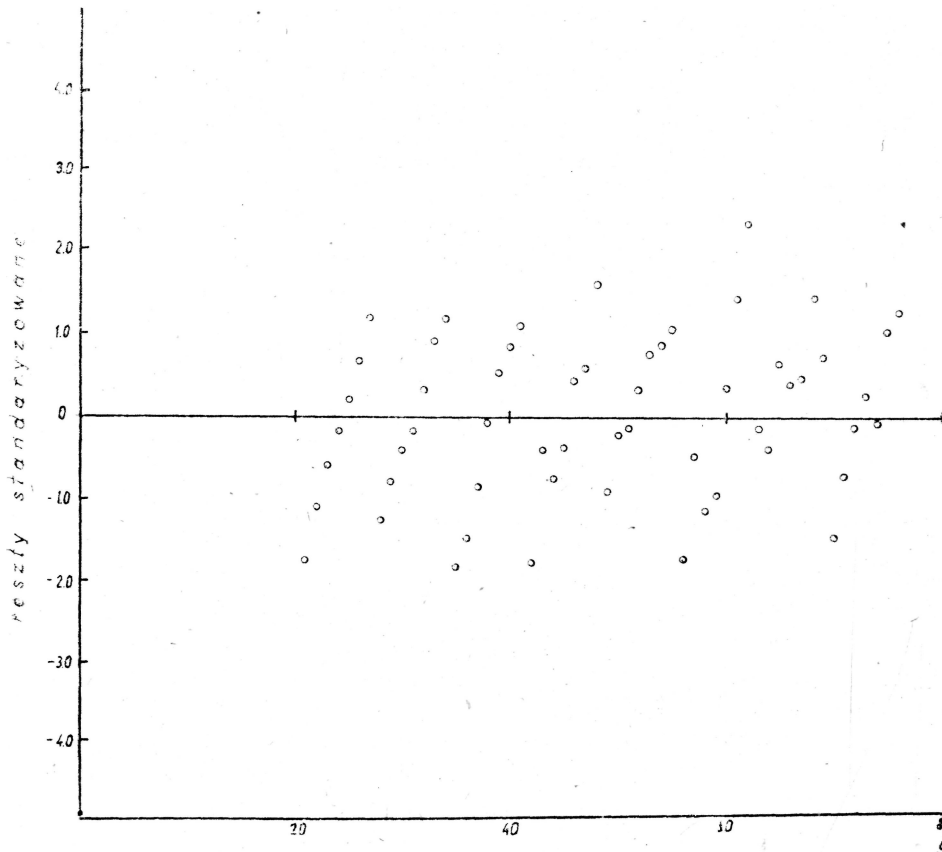


Ryc. 3. Zmienne x_3 (oś pozioma) i y (oś pionowa)

budowie danych, a mianowicie współliniowości bardzo silnej między trzema zmiennymi niezależnymi rzędu r=0,99. Należy zauważyć, że współczynniki korelacji prostej *(r)* między Y i trzema x_i są bardzo wysokie i dodatnie (rzędu +0,96). Wewnątrz grupy przeciwnie, współczynniki są ujemne[5]. Relacja całości dominuje zatem wyraźnie nad relacją rzeczywistą istniejącą wewnątrz każdej grupy. Dodajmy, że efekt kontekstualny ujawnia się również, jeżeli wykonujemy wykres reszt standaryzowanych (między wartościami rzeczywistymi a ich oszacowaniem przez regresję) w zależności od zmiennej zależnej Y, jak można to zobaczyć na rycinie 4. Trzy zmienne niezależne x_i zostały tu zatrzymane dla zapewnienia liniowości. Widać wyraźnie, może nawet lepiej niż na rycinie poprzedniej, pojawienie się istnienia ośmiu grup.

Następnie dane zostały przegrupowane (model **DB**) na 8 grup, takich samych,

[5] Przypomnijmy nawiasem, że przeciwnie do korelacji prostej między dwiema zmiennymi, znak współczynnika korelacji wielokrotnej nie ma znaczenia. Umownie bierze się go jako dodatni. Patrz: H. M. Blalock, *Social Statistics,* New York 1960, s. 347.



Ryc. 4. Reszty (oś pionowa) w stosunku do wartości zmiennej y (oś pozioma)

Wartości średnie przegrupowanych danych

Grupa	Y	X_1	X_2	X_3
1	24	3,98	39,83	4,09
2	31	11,24	105,20	11,01
3	38	18,14	185,83	18,77
4	45	24,83	255,69	24,81
5	52	32,11	313,00	31,66
6	59	39,45	395,14	40,02
7	66	45,49	451,86	44,92
8	73	52,81	551,81	51,48

jakie były podstawą budowy modelu **DA**. Dla każdej z tych grup obliczono średnie zmiennych zależnych i niezależnych, dochodząc do wyników ukazanych w tabeli.

Wykazują one, jak agregacja spowodowała zniknięcie prawdziwej relacji między Y i x_i wewnątrz grup: odtąd pojawia się tylko relacja (w znaczeniu przeciwnym) między grupami. Efekt kontekstualny, połączony z przegrupowaniem danych,

nie pozwala odtąd użytkownikowi na uznanie prawdziwego sensu relacji między zmiennymi. Byłoby tak na przykład w wypadku, gdyby w każdym powiecie obserwowano korelację ujemną między trwaniem życia jednostki a spożyciem tytoniu, podczas gdy — na skutek efektu kontekstualnego — przeciętne długości życia i spożycia tytoniu na powiat byłyby połączone dodatnio jedne z drugimi. Przegrupowanie danych jednostkowych doprowadziłoby w tym wypadku do wniosku, oczywiście błędnego, że długość życia wzrasta, gdy zwiększa się spożycie tytoniu.

W przypadku danych modelu **DB** regresja wielokrotna krokowa dokonana na podstawie danych tabeli zatrzymuje istotnie zmienną x_1 dla najlepszego objaśnienia wariancji (R^2 =0,9997). Wprowadzenie dwu innych zmiennych do modelu regresji nie zmienia prawie wartości R^2; są one więc zupełnie zbyteczne. Jeżeli chodzi o korelacje proste rzędu zero, to obserwuje się współczynnik korelacji bardzo wysoki i dodatni (r+ 0,998) między zmienną zależną Y i każdą ze zmiennych niezależnych x_i.

W stosunku do modelu całkowitego fakt zabrania się do analizy na podstawie średnich z 8 grup zwiększa wariancję objaśnianą: R^2 przechodzi z 0,9358 do 0,9998, biorąc pod uwagę trzy zmienne niezależne. Ponadto korelacja między Y i każdą zmienną x_i ulepsza się także, ponieważ osiąga teraz około 0,999 zamiast około 0,960 w modelu zupełnym (przy danych nieprzegrupowanych). Raz jeszcze więc przegrupowanie danych powoduje znaczny wzrost korelacji rzędu zero i wielokrotnych.

*

Praca miała za główny cel badanie wpływu agregacji danych na wyniki regresji wielokrotnej. Ponadto została wprowadzona do danych wysoka współliniowość, w celu zbadania ewentualnych zakłóceń wprowadzonych przez ten czynnik. Agregacja danych prowadzi do wzrostu współczynnika korelacji wielokrotnych. Kiedy jest efekt kontekstualny (zwany niekiedy „ekologicznym"), agregacja danych może ukrywać relację prawdziwą między zmienną zależną, a zmiennymi niezależnymi — gdy relacje wewnątrz- i międzygrupowe są zwrócone w kierunku przeciwnym. Jest więc rzeczą istotną przede wszystkim branie pod uwagę liczby obserwacji bardziej niż wartość współczynnika korelacji wielokrotnej, by osądzić wartość objaśniającą modelu. Widzieliśmy w przypadku danych losowych, że przegrupowanie prowadziło do wariancji objaśnionej przez model liniowy rzędu 60%. Ten absurdalny wynik odrzuciliśmy weryfikując hipotezę zerową: ze względu na niewielką ilość obserwacji, wartość R^2 — nawet wysoka — nie różni się istotnie od zera; nie można więc wyciągnąć wniosku przyjmując model liniowy w tym przypadku.

Co do efektu kontekstualnego, może on być łatwo wykryty na podstawie informacji indywidualnej kompletnej, rzucając na wykres punkty między zmienną zależną Y i każdą zmienną niezależną x_i, lub jeszcze dodatkowe reszty zależnych zmiennej zależnej Y. Jeżeli taka relacja wyłania się z danych, regresja liniowa nie jest oczywiście modelem statystycznym adekwatnym. W tym przypadku konieczna jest analiza kowariancji.

Agregacja danych może — widzieliśmy to na modelu **DB** — spowodować zni-

knięcie dowodów istnienia efektu kontekstualnego. W tym przypadku narażamy się na dojście do wniosków zawodnych, jeżeli transponuje się relacje otrzymane z danych przydzielanych na jednostki, przez jednostki geograficzne. Z braku możności posługiwania się danymi indywidualnymi, żadne środki techniczne nie pozwalają na wykrycie takiego efektu, wychodząc z danych agregowanych, chyba że dysponuje się informacjami a priori dotyczącymi relacji indywidualnych. W przypadku zatrucia nikotyną i długości życia dysponujemy licznymi badaniami indywidualnymi, wydobywającym właściwy sens związku. Gdyby dane agregowane prowadziły da wniosku przeciwnego, chodziłoby prawdopodobnie o efekt kontekstualny. W tym przypadku trzeba by zrezygnować z wyciągnięcia wniosku z tych danych. Jeżeli związki wewnątrz- i międzygrupowe zdążają w tym samym kierunku, wniosek jest mniej oczywisty. Sens relacji będzie respektowany, jeżeli uciekniemy się do danych agregowanych, ale relacja międzygrupowa nie odbije na tyle relacji wewnątrzgrupowej, ponieważ agregacja pozwoli oczywiście na przetrwanie efektu kontekstualnego[6].

Jeżeli współliniowość między zmiennymi niezależnymi istnieje, eliminacja zmiennych zbytecznych jest dokonywana przez regresję wielokrotną „krokową". Jednakże zmienna „objaśniająca" zmienia się, jak widzieliśmy, od jednego modelu do innego; nie można więc być pewnym prawdziwości wyodrębnionego związku leżącego u podstaw. Pożyteczne jest więc sprawdzenie przedtem istnienia ewentualnej współliniowości między zmiennymi. Jeżeli jest ona silna, można pominąć zmienne zbyteczne. Jeżeli nie zaleca się klasycznego wprowadzenia nowych danych pozwalających na zredukowanie współliniowości[7]; to rozwiązanie jest często niemożliwe do zrealizowania w praktyce, z braku dodatkowych informacji. Można także przegrupować zmienne skorelowane przystępując do analizy głównych składowych, które są, jak wynika z definicji, ortogonalne (niezależne) i wykorzystać ładunki czynnikowe zamiast zmiennych niezależnych początkowych. Jednakże uzyskanie niezależności między zmiennymi dokonuje się za cenę większej trudności na poziomie interpretacji wyników: ponieważ każdy ładunek czynnikowy jest kombinacją liniową zmiennych pierwotnych, interpretacja skierowuje się odtąd na kombinacje, a nie na same zmienne. Jeżeli te kombinacje łub „czynniki" mogą otrzymać właściwe znaczenie, można w konsekwencji zweryfikować teorię; często identyfikacja „czynników" jest niemożliwa. W tym wypadku żadna teoria nie może się równać z zastosowanym modelem statystycznym.

W końcu dobrze jest sprawdzić, czy reszty nie są dotknięte autokorelacją przestrzenną: reszta odnosząca się do jednostki administracyjnej nie może ulegać wpływowi reszt sąsiednich. Problem ten był niedawno przedmiotem różnych artykułów w czasopismach geografii ilościowej. Proponujemy w tym względzie badanie mapy geograficznej reszt, w celu ujawnienia ewentualnej obecności stref geograficznych.

[6] Innymi słowy, nachylenia prostych regresji wewnątrz- i międzygrupowej mogą być różne, nawet jeśli są na przykład wszystkie rosnące.

[7] J. Johnston, *Econometric* ..., s. 164-167.

w łonie których te reszty są autokorelowane[8]. Poza badaniem wizualnym można również uciec się w tym celu do metod klasyfikacji geograficznej reszt, z ograniczeniem sąsiedztwa; brak autokorelacji przestrzennej wyrazi się przez wysoką wielokrotność klas.

ZAŁĄCZNIK: BUDOWA MODELI

Model podstawowy (model A) jest zbudowany na podstawie następujących relacji:

$$Y = x_1 + 20$$

$$x_2 = 10x_1$$

$$x_3 = 57 - x_1$$

Można więc napisać równanie regresji w postaci:

$$Y(1-b_1) = (b_0 - 20b_1 + 57b_3) + (10b_2 - b_3)x_1 .$$

Relacje te zostały następnie zakłócone przez czynnik losowy: każda wartość x_1, określona przez powyższe równania, została dotknięta przez czynnik *j,* gdzie *j* jest liczbą losową zawartą między 1, 10 i 0,90. Dla każdego z modeli zostało generowanych, a później przegrupowanych 56 liczb (na 8 grup) — systematycznie lub losowo, zależnie od modelu. Wszystkie wyliczenia zostały wykonane na materiale Tektronix 4051, o wizualizacji graficznej Zakładu, za pomocą programu analizy regresji krokowej, za pomocą metody najmniejszych kwadratów (OLS) zawartej w „software" statystycznym Tektronixu.

Z języka francuskiego tłumaczyła Czesława Rutkowska

THE IMPACT OF AGGREGATION AND MULTICOLLINEARITY ON THE RESULTS OF THE MULTIPLE REGRESSION ANALYSIS OF SPACE

Summary

Simulated data are used to test the impact of aggregation and of multicollinearity on the results of the multiple regression analysis.

Aggregation increases the level of the multiple correlation coefficient. Furthermore, contextual (or ecological) effects can conceal the true relationship between dependent and independent variables. Redundant variables due to multicollinearity are usually eliminated by stepwise regression procedures. However, the variables thus selected differ from one model to another: the actual underlying relation is not dislosed. Various ways of coping with aggregation and multicollinearity are proposed as conclusions.

[8] Na wzór autokorelacji rzędu 1 w seriach chronologicznych, można sądzić, że autokorelacja — jeśli istnieje — zaznacza się bardziej między jednostkami geograficznymi sąsiednimi niż między jednostkami bardzo od siebie odległymi.