AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Ekonometria Przestrzenna

Projekt 4

Mateusz Mulka, Jakub Le Van

Spis treści

1. Wstęp

Celem projektu jest analiza występowania zależności przestrzennych w modelu liniowym oszacowanym w ramach „Projektu 3”. Dotychczasowa analiza skupiła się na badaniu czynników wpływających na ilość wytwarzanych odpadów komunalnych w wybranych powiatach Polski, jednak nie uwzględniała możliwych zależności przestrzennych między regionami.

2. Dane do analizy

W projekcie wykorzystano ten sam zestaw danych, co w „Projekcie 3”. Dane te dotyczyły wybranych powiatów w Polsce i opisywały czynniki wpływające na poziom wytwarzanych odpadów komunalnych. Wśród zmiennych objaśniających znalazły się takie charakterystyki, jak na przykład gęstość zaludnienia. Wszystkie dane zostały wcześniej zweryfikowane pod kątem kompletności i jakości. Przygotowane dane posłużyły jako podstawa do budowy modelu regresji w „Projekcie 3”, który stanowi punkt wyjścia do dalszej analizy. W obecnym projekcie uwzględniono jednak zależności przestrzenne, co wymagało opracowania macierzy wag przestrzennych opisujących relacje między powiatami.

3. Badanie zależności przestrzennych

Aby zbadać zależności przestrzenne reszt modelu OLS z „Projektu 3”, przeprowadzono testy autokorelacji przestrzennej. Zastosowano globalne i lokalne miary autokorelacji, które umożliwiły ocenę występowania przestrzennych wzorców w analizowanych danych.

Wyniki dla reszt:

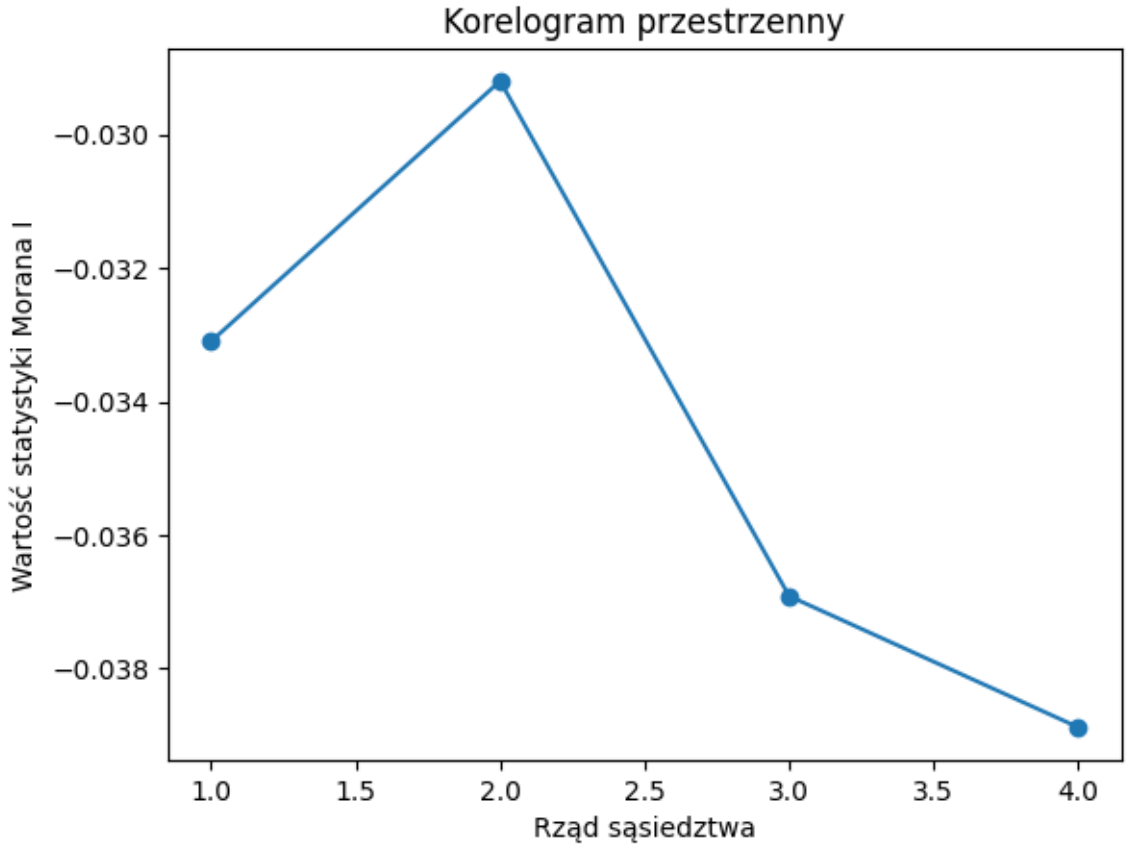
Wskaźnik Moran'a I	-0,053119854099342965
p-value	0,01
Z-score	-1,8107841838551764
Geary's C	1,2686023159804458
Geary's C p-value	0,057
Getis-Ord G	0,191680924476207
Getis-Ord G p-value	0,051

Globalna autokorelacja przestrzenna

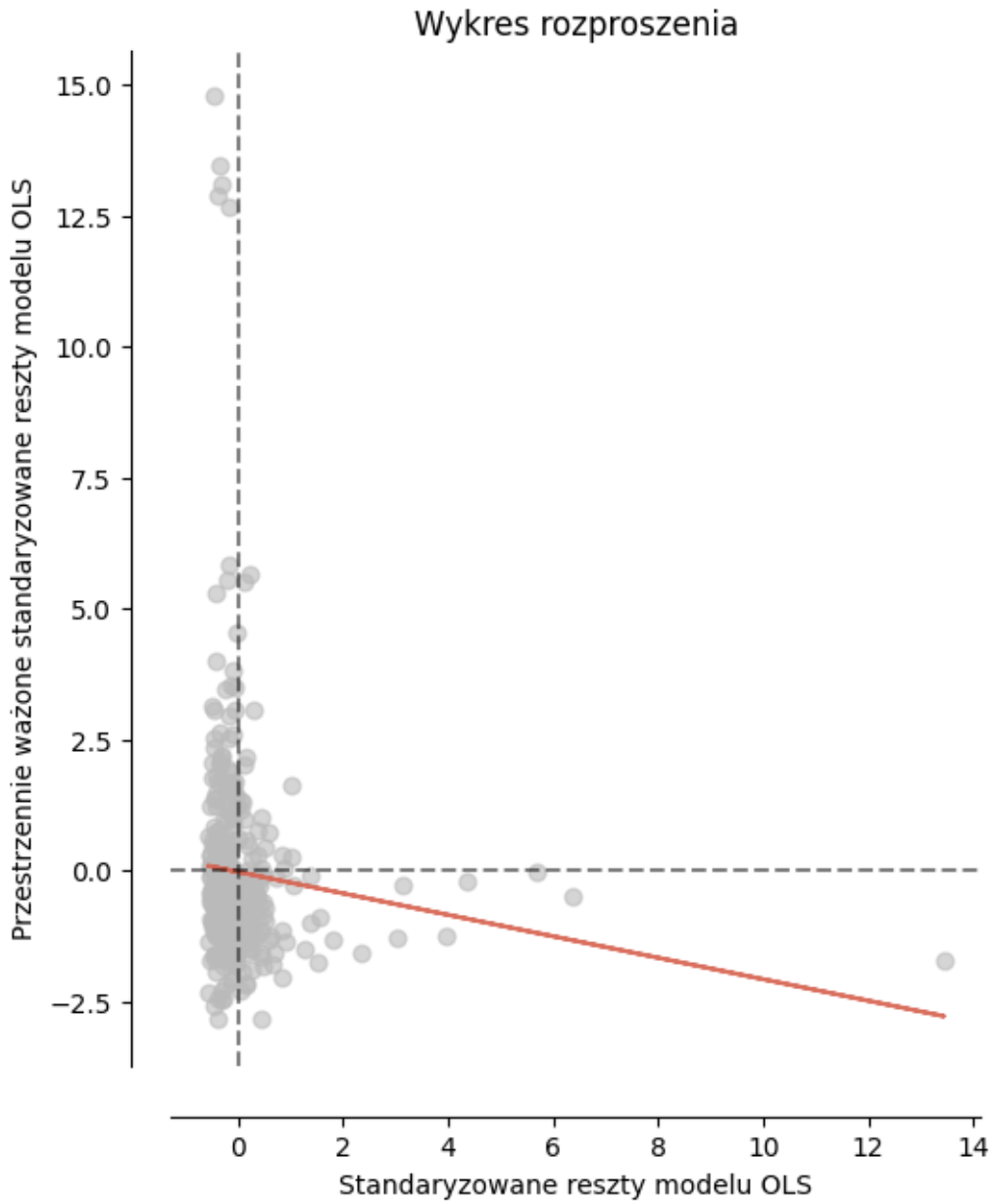
W przypadku przeprowadzania tej analizy wskaźnik Morana I jest najważniejszym wskaźnikiem na jaki powinniśmy zwracać uwagę, jednocześnie jest jedynym wskaźnikiem którego wartość wyszła mniejsza niż 0,05. Z tego też powodu skupiamy się na interpretacji Morana żeby nie zaburzyć oceny wyników. Z uwagi na wartość p-value wynoszącą 0,01

mamy podstawy do odrzucenia hipotezy zerowej, przyjmujemy zatem, że jest obecna ujemna korelacja przestenna.

Korelogram i wykres rozproszenia



Korelogram przestrzenny przedstawia wartości statystyki Morana I dla różnych rzędów sąsiedztwa, co pozwala ocenić obecnośći i siłę autokorelacji. Możemy w łatwy sposób wywnioskować, że najsilniejsza autokorelacja ujemna występuje w przypadku sąsiedztwa III i IV rzędu, a zatem w przypadku jednostek względnie odległych od siebie, niemnie nawet przy niskim rzędzie sąsiedztwa wyniki są podobne.



W przypadku wykresu rozproszenia, mamy tu przedstawione zależności między standaryzowanymi resztami modelu regresji liniowej, a przestrzennie ważonymi standaryzowanymi resztami. Na wykresie możemy zaznaczyć kilka kwestii, chociaż widoczna jest czerwona linia trendu, jej nachylenie wskazuje na słabą zależność między zmiennymi, widoczne są także wartości odstające. Większość punktów na wykresie znajduje się w okolicy punktu (0,0) – w przypadku tych wyników reszty są stosunkowo niewielkie, jednak nie możemy tego powiedzieć o całym zbiorze.

4. Budowa modeli SAR i SEM

W celu uwzględnienia zależności przestrzennych w analizie ilości odpadów komunalnych, zastosowano dwa modele regresji przestrzennej: SAR (Spatial Autoregressive Model) oraz SEM (Spatial Error Model). Modele te pozwalają na różne podejścia do uwzględnienia przestrzennych relacji między regionami:

Model SAR uwzględnia zależność przestrzenną w zmiennej objaśnianej, tj. zakłada, że wartość zmiennej w jednym regionie jest zależna od wartości w regionach sąsiednich.

Model SEM koncentruje się na korelacji przestrzennej w resztach, co oznacza, że niewyjaśnione przez model odchylenia mogą wykazywać zależność przestrzenną.

Model SAR

Model SAR uwzględnia zależności przestrzenne w zmiennej objaśnianej, co oznacza, że ilość odpadów komunalnych w jednym regionie może być powiązana z ilością w regionach sąsiednich. Wyniki estymacji modelu SAR przedstawiają się następująco:

- **Pseudo R^2 = 0,0071**, co wskazuje na bardzo niską zdolność modelu do wyjaśnienia zmienności zmiennej zależnej.
- Wartości współczynników dla zmiennych objaśniających są w większości nieistotne statystycznie:
    - **Gęstość zaludnienia**: współczynnik wyniósł **-0,00018** (p = 0,98148).
    - **Wynagrodzenie**: współczynnik wyniósł **-0,00329** (p = 0,48599).
    - Podobny brak istotności wykazano dla zmiennych takich jak liczba ludności czy turystyka.
- Kryterium Akaikego (AIC) wyniosło **4015,10**, co sugeruje, że model SAR jest nieco mniej dopasowany niż model OLS (AIC = 4013,30).

Interpretacja wyników wskazuje, że przestrzenne zależności w zmiennej objaśnianej nie mają istotnego wpływu na ilość odpadów komunalnych. Model SAR nie poprawił znacząco dopasowania w stosunku do modelu OLS, co może być wynikiem słabych przestrzennych wzorców w danych.

Model SEM

Model SEM uwzględnia przestrzenne zależności w resztach, co pozwala na wychwycenie potencjalnych przestrzennych błędów w modelu. Wyniki estymacji SEM przedstawiają się następująco:

- **Pseudo R^2 = 0,0063**, co również wskazuje na bardzo niską zdolność modelu do wyjaśnienia zmienności zmiennej zależnej.

- Kluczowy parametr **lambda**, opisujący siłę zależności przestrzennych w resztach, wyniósł **-0,01050** (p = 0,52571), co wskazuje na brak istotności statystycznej.
- Kryterium Akaikego (AIC) dla modelu SEM wyniosło **4012,91**, co jest minimalnie lepszym wynikiem w porównaniu z modelem SAR i modelem OLS.

Podobnie jak w przypadku modelu SAR, model SEM nie wprowadził istotnej poprawy w jakości dopasowania do danych. Brak istotności współczynnika lambda sugeruje, że w resztach modelu nie występują wyraźne przestrzenne zależności.

Porównanie Modeli

Wyniki estymacji modeli SAR i SEM zostały porównane z modelem OLS w oparciu o kryterium Akaikego (AIC):

Model OLS	4013,30
Model SAR	4015,10
Model SEM	4012,91

Najlepsze dopasowanie uzyskano dla modelu SEM, jednak różnice między modelami są minimalne i nie pozwalają na wyciągnięcie jednoznacznych wniosków o przewadze któregokolwiek z modeli przestrzennych. Wczytując się w szczegółowe wyniki modeli możemy zauważyć, że zarówno model SEM jak i model SAR mają lepsze R-kwadrat od standardowego modelu regresji liniowej, możemy zatem przypuszczać, że faktycznie modele wychwytują zależności przestrzenne widoczne w danych, jednak poprawa jest niewielka. W analizie zebranych danych widzieliśmy, że zmienne wykazywały wysoką lub średnią korelację ze zmienną objaśnianą, a także niską między sobą, jest to pożądana sytuacja jednak być może dobór innych zmiennych wykazałby lepsze wyniki (pomimo, iż dobrane przez nas zmienne z logicznego punktu widzenia powinny mieć istotny wpływ na produkcję odpadów komunalnych). Dodatkowo być może dodanie kolejnych korków procesowania danych – takich jak użycie logarytmów niektórych kolumn, mogłoby pomóc poprawić dane, jednak w przypadku modelu OLS widzieliśmy w dalszym stopniu słabe wyniki po zastosowania techniki usuwania wartości odstających. Widzimy, że użyte zmienne zazwyczaj nie są istotne statystycznie, być może zależności w tych danych są bardziej złożone i wymagają bardziej skomplikowanych modeli oraz doboru szerszej gammy zmiennych objaśniających.

5. Weryfikacja i ocena modeli regresji przestrzennej

W celu sprawdzenia poprawności i stabilności modeli SAR i SEM przeprowadzono szereg testów diagnostycznych, które miały na celu ocenę spełnienia założeń tych modeli oraz ich przydatności w analizie danych przestrzennych.

5.1. Stabilność parametrów (Test Chow)

Test Chow przeprowadzono w celu zbadania, czy parametry modeli SAR i SEM są stabilne w różnych podzbiorach danych. Weryfikowano następujące hipotezy:

H_0: Parametry modelu są stabilne w obu podzbiorach (brak różnic między grupami).

H_1: Parametry modelu różnią się między podzbiorami (model nie jest stabilny).

Wyniki:

	Statystyka F	P-Value
Model SAR	1,3524	0,2331
Model SEM	1,0343	0,4026

W obu przypadkach wartości **p-value** są większe niż 0,05, co oznacza brak podstaw do odrzucenia hipotezy zerowej. Parametry obu modeli można uznać za stabilne między podzbiorami danych, co wskazuje na poprawną specyfikację i brak istotnych różnic między grupami.

5.2. Homoskedastyczność reszt (Test Levene'a)

Test Levene'a został wykorzystany do oceny jednorodności wariancji reszt w modelach. Weryfikowano następujące hipotezy:

H_0: Wariancje reszt są jednorodne we wszystkich grupach (homoskedastyczność).

H_1: Wariancje reszt nie są jednorodne (heteroskedastyczność).

Wyniki:

	P-Value
Model SAR	0,6312
Model SEM	0,6219

Dla obu modeli wartości **p-value** są większe niż 0,05, co oznacza brak podstaw do odrzucenia hipotezy zerowej. Możemy zatem uznać, że wariancje reszt w modelach są jednorodne, co świadczy o braku heteroskedastyczności.

5.3. Normalność reszt

Test normalności przeprowadzono w celu sprawdzenia, czy reszty modeli SAR i SEM pochodzą z rozkładu normalnego. Weryfikowano następujące hipotezy:

H_0: Reszty modelu pochodzą z rozkładu normalnego.

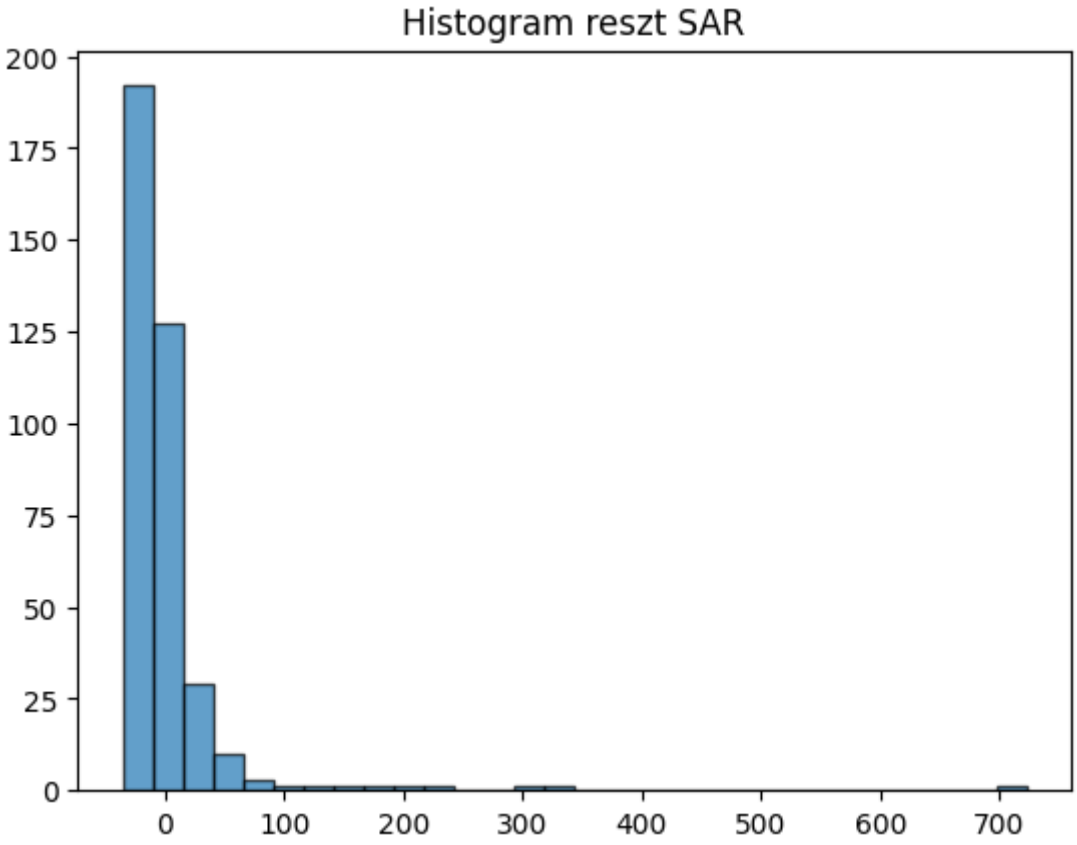
H_1: Reszty modelu nie pochodzą z rozkładu normalnego.

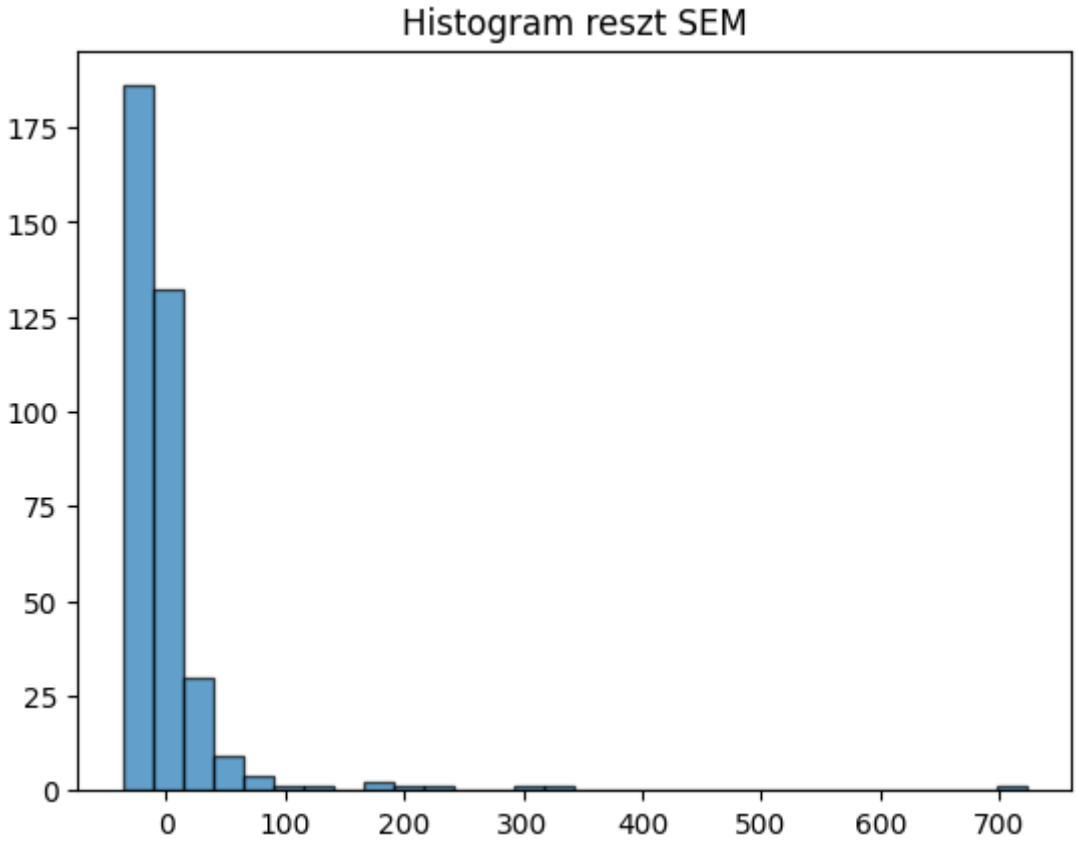
Wyniki:

	P-Value
Model SAR	$3{,}28 \times 10^{-126}$.
Model SEM	$3{,}28 \times 10^{-126}$.

Obie wartości **p-value** są niezwykle niskie, co prowadzi do odrzucenia hipotezy zerowej. Oznacza to, że reszty modeli nie spełniają założenia normalności. Może to być związane z występowaniem wartości odstających lub z nieliniowymi zależnościami, które nie zostały uchwycone przez modele – co widać na poniższych wykresach.

5.4. Wizualizacja reszt





Podsumowanie

Podsumowując, w tym projekcie użyliśmy technik przestrzennych, analizy korelacji przestrzennej oraz modeli takich jak SAR czy SEM do oceny występowania zależności przestrzennych w naszych danych. Po zastosowaniu modeli SAR i SEM widzimy nieznaczną poprawę w uzyskiwanych wartościach R-kwadrat, jednak nasze modele nie radzą sobie najlepiej. W dalszej analizie i rozwoju projektu prawdopodobnie kluczowe byłoby zastanowienie się nad rozwojem użytej bazy danych, to znaczy doborem kolejnych zmiennych objaśniających i ich dalszej manipulacji – jak opisaliśmy w projekcie. W zmiennych widzimy występującą ujemną korelację przestrzenną, wartości odstające. Parametry uzyskanych przez nas modeli są stabilne, a reszty homoskedastyczne, jednak nie pochodzą one z rozkładu normalnego. Wnioski te poparte są wynikami z użytych testów statystycznych.