Analiza regresji przestrzennej w języku Python - dokumentacja końcowa

Maksim Makaranka

2024L

1 Temat

Zbadanie funkcjonalności dostępnej w środowisku R/Python dotyczącej regresji przestrzennej, w tym autoregresji. Celem zadania jest ocena dostępnych metod pod kątem ich przydatności do analizy danych przestrzennych: poprawność działania, czas wykonania w zależności od wielkości danych, łatwość użycia, dostępność materiałów pomocniczych.

2 Opis projektu

Projekt skupia się na badaniu bibliotek i narzędzi dostępnych w języku **Python**, które umożliwiają przeprowadzenie regresji przestrzennej. W ramach projektu zostały porównane 4 rodzaje regresji:

- **Ordinary Least Squares** (OLS): Jest to najprostszy model regresji, który minimalizuje sumę kwadratów różnic między obserwowanymi a przewidywanymi wartościami. Jest łatwy do zrozumienia i implementacji, ale nie uwzględnia przestrzennych zależności między danymi. Jest to podstawowy wybór dla wielu problemów regresji, ale może nie być optymalny, gdy dane mają silne zależności przestrzenne.
- **Spatial Two Stage Least Squares** ($S2SLS$): Jest to rozszerzenie modelu OLS, które uwzględnia przestrzenne zależności w danych. W pierwszym etapie estymuje się parametry modelu, a w drugim etapie uwzględnia się przestrzenne zależności. Jest to bardziej zaawansowany model, który może dawać lepsze wyniki dla danych przestrzennych.
- **Maximum Likelihood Spatial Lag Model** (SLM): Jest to model regresji przestrzennej, który uwzględnia przestrzenne zależności między obserwacjami. W modelu tym zakłada się, że wartość zmiennej zależnej w danym miejscu jest funkcją wartości tej zmiennej w sąsiednich miejscach. Jest to potężny model, który może uwzględniać skomplikowane zależności przestrzenne.
- **Maximum Likelihood Spatial Error Model** (SEM): Jest to inny model regresji przestrzennej, który zakłada, że błędy są przestrzennie skorelowane. W tym modelu, błędy w jednym miejscu mogą wpływać na błędy w sąsiednich miejscach. Jest to użyteczne, gdy modele są niedoszacowane lub przeszacowane w określonych obszarach przestrzeni.

W ramach projektu, nacisk został położony na praktyczne zastosowanie modeli regresji przestrzennej w Pythonie, z wykorzystaniem bibliotek takich jak **PySAL**, **spreg** oraz **GeoPandas**, które oferują zaawansowane funkcje do pracy z danymi przestrzennymi.

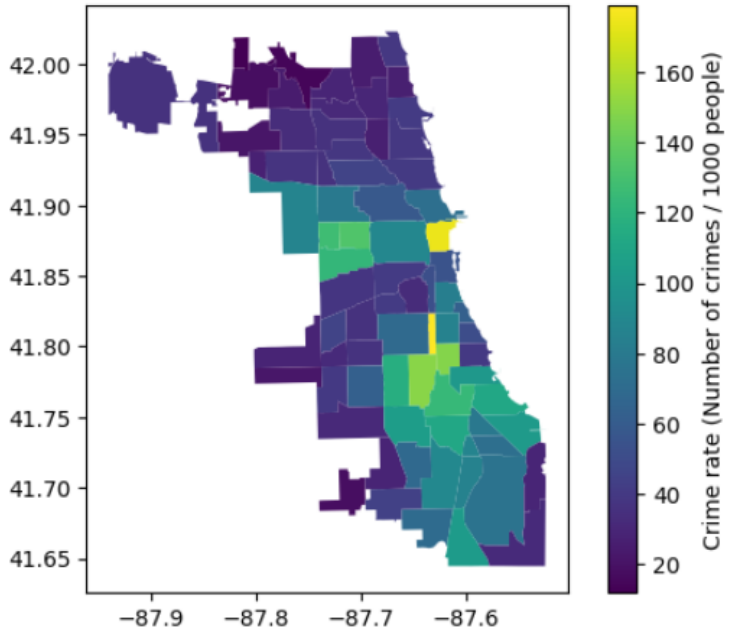
3 Opis narzędzi

W projekcie zostały wykorzystane następujące narzędzia i biblioteki:

- **Python** - język programowania wysokiego poziomu.
- **GeoPandas** - rozszerzenie **Pandas** ułatwiające pracę z danymi przestrzennymi[6]. Jest otwartoźródłowym projektem, który dodaje obsługę danych geograficznych do obiektów **Pandas**. Głównym celem **GeoPandas** jest ułatwienie pracy z danymi przestrzennymi w Pythonie. Łączy on możliwości **Pandas** i **Shapely**, dostarczając operacje przestrzenne w **Pandas** oraz wysokopoziomowy interfejs do wielu geometrii dla **Shapely**. **GeoPandas** pozwala na łatwe wczytywanie, przetwarzanie, analizę i wizualizację danych przestrzennych.
- **PySAL** - biblioteka do analizy danych przestrzennych[7]. Jest otwartoźródłowa, oferuje wiele funkcji do analizy przestrzennej, takich jak przestrzenne statystyki, analiza klastrów, interpolacja, analiza sieci i wiele innych.
- **spreg** - moduł w ramach **PySAL**, który umożliwia przeprowadzanie zaawansowanych analiz regresji przestrzennej, w tym autoregresji[8].
- **geodatasets**[9] - biblioteka zawierająca zestawy danych geograficznych ułatwiająca dostęp i manipulację danymi przestrzennymi, kompatybilna z innymi narzędziami do analizy danych przestrzennych, takimi jak **GeoPandas** i **PySAL**.

4 Zbiór danych

W ramach tego projektu skupiłem się na badaniu wpływu różnych czynników na poziom przestępczości w społecznościach (communities) w Chicago. Wykorzystałem dane udostępniane przez bibliotekę **Python geodatasets**. Poniżej przedstawiony jest wykres demonstrujący poziom przestępczości w poszczególnych społecznościach.



Rysunek 1: Wskaźnik przestępczości (Liczba przestępstw / 1000 osób) w społecznościach Chicago

Zbiory danych zawierają wiele interesujących zmiennych, takich jak procent populacji poniżej 125% progu ubóstwa, dane procentowe dotyczące wieku ludzi, dane o procentach ludzi z wyższym wykształceniem, a także dane dotyczące procentów różnych ras ludzi w społecznościach.

Dodatkowo, postanowiłem zbadać wpływ liczby sklepów alkoholowych na 1000 mieszkańców oraz gęstość zaludnienia na przestępczość. W tym celu pobrano dane dotyczące lokalizacji sklepów alkoholowych, które zostały zagregowane za pomocą operacji przestrzennej intersekcji udostępnionej przez **GeoPandas**, co pozwoliło na obliczenie wspomnianego współczynnika dla każdej społeczności.

Szczegółowy opis przetwarzania danych, wraz z wykresami demonstrującymi rozkład poszczególnych zmiennych w społecznościach, jest przedstawiony w repozytorium w pliku *1-Data_preprocessing.ipynb*[1], gdzie wyjaśniono ostateczny wybór lub odrzucenie zmiennych do analizy. W niniejszej dokumentacji przedstawiony jest jedynie sam wybór zmiennych - w modelach regresyjnych zostały użyte następujące zmienne: `POP_DENSITY` (gęstość zaludnienia), `LIQUOR_STORES_DENSITY` (gęstość sklepów alkoholowych), `POP_BELOW_125_POVERTY_PCT` (procent populacji poniżej 125% progu ubóstwa), `WHITE_POP_PCT` (procent białej populacji), `BLACK_POP_PCT` (procent czarnej populacji), `ASIAN_POP_PCT` (procent azjatyckiej populacji).

Analiza ta dostarcza cennych informacji dla organów ścigania i planistów miejskich, pomagając im lepiej zrozumieć i przewidywać przestępczość, a także w projektowaniu skuteczniejszych strategii prewencyjnych i reagowania.

5 Analiza wyników regresji

W wybranym modelu danych występuje wysoka korelacja między `WHITE_POP_PCT` a `BLACK_POP_PCT`. Stanowi to problem, który wymaga przeprowadzenia każdej regresji oddzielnie dla każdej z tych grup. Pozwoli to zbadać, czy zmienne wpływają na zmienną zależną, jednocześnie unikając współliniowości i zamieszania wyników spowodowanego istnieniem społeczności mieszanych. Na podstawie przeprowadzonych regresji, możemy zauważyć następujące wyniki:

- Model regresji Ordinal Least Squares (OLS) z `WHITE_POP_PCT` jako zmienną niezależną ma wartość R^2 równą 0.8133. Oznacza to, że około 81.33% wariancji `CRIME_RATE` można wyjaśnić za pomocą predyktorów w modelu. Wartość skorygowanego R^2 jest trochę niższa i wynosi 0.8001. Mała różnica pomiędzy tymi współczynnikami wskazuje, że w modelu prawdopodobnie nie występują predyktory, które nie przyczyniają się znacząco do wyjaśnienia `CRIME_RATE`. Te wartości R^2 i skorygowanego R^2 są podobne do tych, gdy wykorzystywany jest `BLACK_POP_PCT`, które wynoszą odpowiednio 0.8205 i 0.8079.

  Wśród wszystkich predyktorów, zmienne `POP_DENSITY`, `LIQUOR_STORES_DENSITY`, `POP_BELOW_125_POVERTY_PCT`, `WHITE_POP_PCT` i `BLACK_POP_PCT` okazały się statystycznie istotne na poziomie 5%. Dodatnie współczynniki dla

POP_DENSITY, LIQUOR_STORES_DENSITY, POP_BELOW_125_POVERTY_PCT i BLACK_POP_PCT sugerują, że wzrost gęstości sklepów z alkoholem, procentu populacji poniżej 125% progu ubóstwa, procentu populacji czarnej i gęstości zaludnienia są wszystkie związane ze wzrostem wskaźnika przestępczości.

Co ciekawe, WHITE_POP_PCT jest istotnym predyktorem o ujemnym współczynniku, co wskazuje, że wzrost procentu populacji białej wiązał się ze spadkiem wskaźnika przestępczości, a ASIAN_POP_PCT jest istotnym predyktorem ujemnym tylko w przypadku regresji z WHITE_POP_PCT.

Diagnostyka zależności przestrzennej wykazała istotną autokorelację przestrzenną w modelu. Testy mnożnika Lagrange'a dla zależności opóźnienia i błędu były istotne, co sugeruje, że CRIME_RATE w jednym miejscu może być wpływany przez wartości w sąsiednich lokalizacjach.

Biorąc pod uwagę te wyniki, korzystne jest rozważenie innego modelu regresji, który może uwzględnić zależność przestrzenną i błędu, takiego jak model Spatial Two Stage Least Squares, Spatial Lag Model lub Spatial Error Model. Pomogłoby to zapewnić prawidłowe uwzględnienie opóźnienia przestrzennego.

- Przechodząc do modelu Spatial Two Stage Least Squares ($S2SLS$), zauważamy, że w pierwszej regresji, gdzie używana jest zmienna WHITE_POP_PCT, model $S2SLS$ wykazuje wartość Pseudo R^2 równą 0.8702, co wskazuje, że model wyjaśnia około 87.02% watiancji CRIME_RATE. Jest to poprawa w porównaniu z modelem OLS. W drugiej regresji, gdzie używana jest zmienna BLACK_POP_PCT, model $S2SLS$ pokazuje wartość Pseudo R^2 równą 0.8674.

  Wśród zmiennych niezależnych, zmienne POP_DENSITY, LIQUOR_STORES_DENSITY, POP_BELOW_125_POVERTY_PCT, WHITE_POP_PCT i ASIAN_POP_PCT okazały się statystycznie istotne. Ujemne współczynniki dla POP_DENSITY, WHITE_POP_PCT i ASIAN_POP_PCT sugerują, że wzrost tych zmiennych wiąże się ze spadkiem wskaźnika przestępczości. Negatywny współczynnik regresji dla zmiennej ASIAN_POP_PCT sugeruje, że wyższy procent populacji azjatyckiej w Chinatown, jest skorelowany z niższym poziomem przestępczości w tej społeczności.

  Ciekawe jest, że regresja $S2SLS$, która wykorzystuje BLACK_POP_PCT jako zmienną niezależną, pokazuje statystycznie nieistotne wyniki dla ASIAN_POP_PCT. To samo zaobserwowano również w tej samej regresji przy użyciu OLS. Dodatkowo, podobne wyniki zostaną zaobserwowane w modelu SLM. Można to przypisać do utrzymania się zarówno białej, jak i czarnej populacji w Chinatown na podobnym poziomie, około 15-20%, co można zobaczyć na wizualizacjach danych. Podczas gdy wpływ populacji białej na wskaźnik przestępczości jest pozytywny i koreluje z wpływem populacji azjatyckiej, nie zmniejszając reprezentatywności wpływu populacji azjatyckiej, wpływ populacji czarnej na wskaźnik przestępczości jest negatywny. Ta negatywna korelacja zmniejsza wpływ populacji azjatyckiej na wskaźnik przestępczości społeczności i robi go nieistotnym. Te czynniki, wraz z relatywnie małym rozmiarem społeczności azjatyckiej, były powodami nieprzeprowadzania osobnej regresji dla populacji azjatyckiej.

  Zmienna W_CRIME_RATE jest również istotna, reprezentując opóźnienie przestrzenne (ang. spatial lag) zmiennej zależnej CRIME_RATE. Sugeruje to, że wskaźnik przestępczości w danym miejscu jest wpływany przez wskaźniki przestępczości w sąsiednich lokalizacjach. Dodatni współczynnik W_CRIME_RATE wskazuje, że miejsca o wysokim wskaźniku przestępczości zwykle są otoczone przez inne miejsca o wysokim wskaźniku przestępczości.

  Model $S2SLS$, z jego zdolnością do uwzględniania zależności przestrzennej w danych, okazuje się bardziej niezawodny i dokładny w przewidywaniu wskaźnika przestępczości. Jest to szczególnie prawdziwe, biorąc pod uwagę istotne testy mnożnika Lagrange'a w modelu OLS. Dlatego model $S2SLS$ jest lepiej dopasowany do tej analizy.

- Następnie, przechodząc do modelu Spatial Lag Model (SLM), zauważamy, że model SLM pokazuje wartość Pseudo R^2 równą 0.8707, gdzie używana jest zmienna WHITE_POP_PCT, i 0.8693, gdzie używany jest BLACK_POP_PCT. Wskazuje to, że model SLM jest porównywalny z modelem $S2SLS$, oraz zdecydowanie lepszy od modelu OLS.

  Wśród wszystkich predyktorów, zmienne LIQUOR_STORES_DENSITY, POP_BELOW_125_POVERTY_PCT, WHITE_POP_PCT, BLACK_POP_PCT i ASIAN_POP_PCT (w przypadku regresji z WHITE_POP_PCT) okazały się statystycznie istotne na poziomie 5%. Te wyniki potwierdzają obserwacje z poprzednio używanych modeli.

  Zmienna W_CRIME_RATE jest również istotna, reprezentując wpływ wskaźników przestępczości w sąsiednich lokalizacjach na zmienną zależną CRIME_RATE.

- Na koniec, przechodząc do modelu Spatial Error Model (SEM), zauważamy, że SEM pokazuje wartości Pseudo R^2 0.8058 i 0.8138, które są niższe niż w innych modelach oprócz OLS.

  Wśród wszystkich predyktorów, LIQUOR_STORES_DENSITY, POP_BELOW_125_POVERTY_PCT, WHITE_POP_PCT oraz BLACK_POP_PCT okazały się statystycznie istotne na poziomie 5%. Jest to mniej w porównaniu do innych modeli, gdzie więcej zmiennych było istotnych. Może to sugerować, że inne modele są lepsze w wykrywaniu struktury zależności przestrzennej.

  Podsumowując, chociaż SEM dostarcza cennych informacji, wydaje się, że SLM lub $S2SLS$ mogą być lepszym wyborem dla tego konkretnego zestawu danych i pytania badawczego.

Szczegółowe analizy są zawarte w repozytorium w plikach *2-Regression_OLS.ipynb*[2], *3-Regression_S2SLS.ipynb*[3], *4-Regression_SLM.ipynb*[4] oraz *5-Regression_SEM.ipynb*[5].

6 Wnioski

Projekt dotyczył badania funkcjonalności dostępnej w środowisku **Python** dotyczącej regresji przestrzennej. Celem zadania było ocenić dostępne metody pod kątem ich przydatności do analizy danych przestrzennych.

- **Poprawność działania**: Na podstawie przeprowadzonych analiz, wszystkie modele regresji - OLS, $S2SLS$, SLM i SEM - wykazały istotne wyniki dla różnych predyktorów. Modele $S2SLS$ i SLM okazały się być najbardziej efektywne, wyjaśniając odpowiednio około 87% zmienności wskaźnika przestępczości.
- **Łatwość użycia**: Wszystkie użyte narzędzia były łatwe w użyciu. Większość problemów wynikała z konieczności oceny i dopasowania danych, a nie z samego korzystania z narzędzi.
- **Dostępność materiałów pomocniczych**: Materiały pomocnicze były szeroko dostępne dla wszystkich używanych narzędzi. Jednakże, brakowało nieco dokumentacji tekstowej dla niektórych modeli w module **spreg**.

Podsumowując, przeprowadzenie regresji przestrzennej jest skomplikowanym procesem, który wymaga starannego wyboru modelu i zrozumienia danych. W tym projekcie, różne modele regresji przestrzennej zostały skutecznie zastosowane do analizy wskaźnika przestępczości. Wyniki te dostarczają cenne informacje, które mogą pomóc w podejmowaniu decyzji dotyczących bezpieczeństwa i planowania miejskiego. Jednakże, jak pokazał ten projekt, ważne jest również zrozumienie ograniczeń każdego modelu i kontekstu, w którym są stosowane. W przyszłości, warto byłoby przeprowadzić dalsze badania, aby lepiej zrozumieć wpływ różnych czynników na wskaźnik przestępczości i jak te czynniki mogą być różne w różnych kontekstach przestrzennych.

Ponadto, poza obecnym zakresem projektu, pozostają nierozwiązane kwestie takie jak analiza czasu wykonania regresji w zależności od wielkości danych oraz badanie autoregresji. Autor jest świadomy tych problemów, zwłaszcza autoregresji, i postara się dostarczyć dodatkowe informacje na ten temat, nawet jeśli nie będą one formalnie udokumentowane.

Literatura

[1] Data preprocessing - https://github.com/maksanm/Spatial-Regression/blob/main/1-Data_Preprocessing.ipynb

[2] Ordinary Least Squares (OLS) - https://github.com/maksanm/Spatial-Regression/blob/main/2-Regression_OLS.ipynb

[3] Spatial Two Stage Least Squares (S2SLS) - https://github.com/maksanm/Spatial-Regression/blob/main/3-Regression_S2SLS.ipynb

[4] Maximum Likelihood Spatial Lag Mode (SLM) - https://github.com/maksanm/Spatial-Regression/blob/main/4-Regression_SLM.ipynb

[5] Maximum Likelihood Spatial Error Model (SEM) - https://github.com/maksanm/Spatial-Regression/blob/main/5-Regression_SEM.ipynb

[6] Dokumentacja **GeoPandas** - https://geopandas.org/en/stable/docs.html

[7] Dokumentacja **PySAL** - https://pysal.org/libpysal/api.html

[8] Dokumentacja **spreg** - https://pysal.org/spreg/

[9] Dokumentacja **geodatasets** - https://geodatasets.readthedocs.io/en/latest/