Raport Deep Learning- Wiktoria Stęczna

Cel projektu

Celem projektu jest analiza rozmieszczenia mieszkań na sprzedaż w Warszawie oraz identyfikacja naturalnych, homogenicznych grup (klastrów) przy użyciu metod głębokiego uczenia Deep Learning. Dane zostały wcześniej poddane analizie z wykorzystaniem klasycznych metod ekonometrycznych (KMNK) oraz technik uwzględniających zależności przestrzenne, takich jak GWR i MGWR. W niniejszym projekcie celem jest ocena skuteczności podejścia opartego na głębokich sieciach neuronowych, ze szczególnym uwzględnieniem modelu GCN (Graph Convolutional Network).

Przygotowanie danych

Zbiór danych obejmuje oferty mieszkań wystawionych na sprzedaż w serwisie Otodom.pl w dniu 3 stycznia 2025 r., znajdujących się w granicach administracyjnych Warszawy. Dane zawierają szereg cech opisujących nieruchomości, w tym m.in.: cenę, powierzchnię (m^2), liczbę pokoi, typ rynku, rodzaj budynku, piętro, liczbę pięter w budynku, typ okien, rodzaj ogrzewania, rok budowy, konstrukcję budynku, wysokość czynszu, formę własności, materiał budowlany, a także odległości od infrastruktury (metro, tramwaj/autobus, kolej, park) oraz czas dojazdu do centrum. Zawierają również informacje o lokalizacji geograficznej w postaci współrzędnych (szerokość i długość geograficzna). Pierwotny zbiór danych zawiera ponad 14 tys. nieruchomości, jednakże ze względu obliczeniowych do analizy użyto losowo wybranej próbki niezduplikowanych 3000 mieszkań.

	OBJECTID	price	m2	rooms_number	market	building_type	floor_no	building_floor	windows_type	heating	...	longitude	building_material	metro_distance	bust_tram_
0	1929	1098000.0	57.00	4	1	1	3	7	1	5	...	20.988997	9	831.40	
1	1519	880000.0	45.43	2	1	0	0	4	1	5	...	20.909935	8	188.06	
2	2417	1999000.0	75.70	3	1	6	2	3	1	5	...	21.059330	1	1625.64	
3	2313	1255000.0	52.44	2	1	6	3	4	1	5	...	21.034454	1	1395.94	
4	6259	560000.0	20.97	1	1	1	2	10	1	5	...	20.956830	9	904.63	

5 rows × 25 columns

Tworzenie grafu powiązań między nieruchomościami

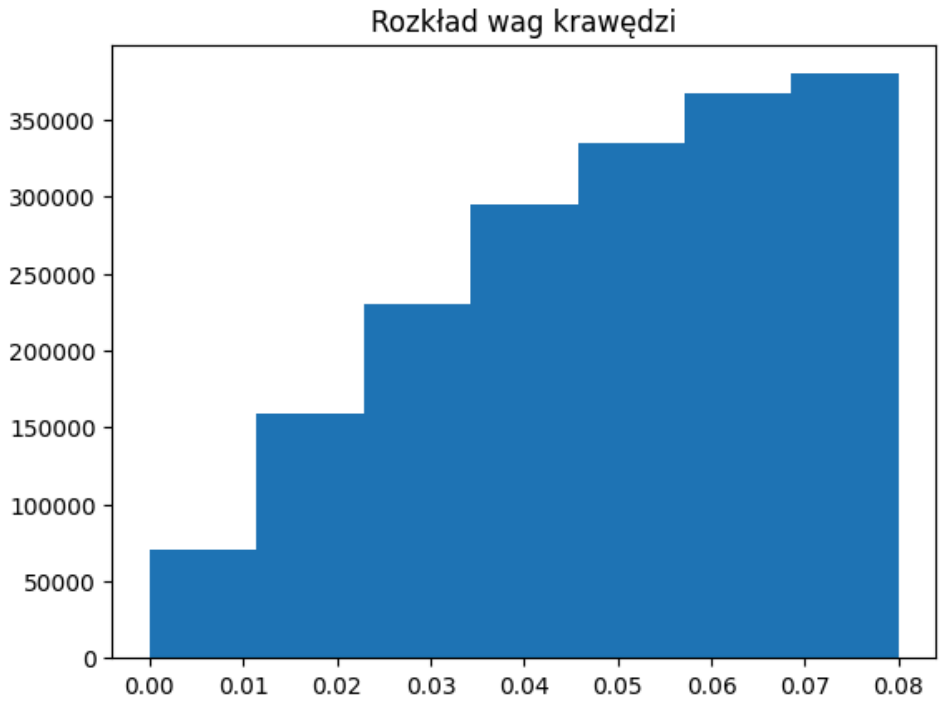
W celu zastosowania metody Graph Convolutional Network (GCN), dane przestrzenne muszą zostać odwzorowane w postaci grafu. W tym kroku tworzony jest graf nieruchomości, w którym każdy wierzchołek odpowiada jednej obserwacji - mieszkaniu, a krawędzie reprezentują relacje przestrzenne między nimi.

Graf budowany jest na podstawie współrzędnych geograficznych (latitude i longitude). Dla każdej pary mieszkań obliczana jest euklidesowa odległość pomiędzy nimi. Jeżeli odległość ta jest mniejsza niż zdefiniowany próg, pomiędzy wierzchołkami dodawana jest krawędź. Wartość tej odległości zapisywana jest jako waga krawędzi.

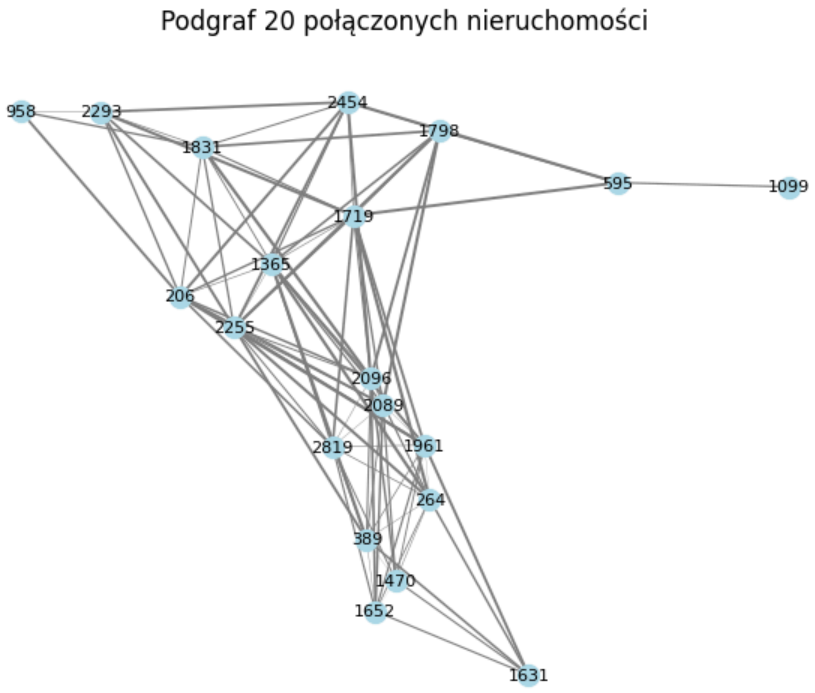
Tak skonstruowany graf umożliwia modelowi GCN uwzględnienie lokalnego kontekstu przestrzennego przy przetwarzaniu informacji o nieruchomościach, co pozwala na lepsze wykrywanie wzorców i struktur w danych.

Jednym z głównych wyzwań podczas konstruowania grafu było dobranie odpowiedniego progu odległości, poniżej którego między punktami dodawana jest krawędź. Zbyt mały próg prowadził do powstania grafu o wielu brakujących połączeniach, co utrudniało uchwycenie struktury przestrzennej. Z kolei zbyt duży próg powodował utratę lokalnego kontekstu – relacje pomiędzy bliskimi geograficznie nieruchomościami stawały się mniej istotne.

Po przeprowadzeniu testów za optymalną uznano wartość 0,08 stopnia różnicy we współrzędnych geograficznych, co odpowiada około 8 km w rzeczywistości. Dzięki temu uzyskano sensowną gęstość połączeń oraz dobrze zróżnicowane wagi krawędzi, co pozytywnie wpływa na jakość reprezentacji grafowej.



W celu lepszego zobrazowania, na czym polega reprezentacja danych w postaci grafu, poniżej przedstawiono podgraf zawierający 20 wybranych nieruchomości. Każdy wierzchołek reprezentuje pojedyncze mieszkanie, a krawędzie łączące wierzchołki wskazują na bliskość geograficzną – im bliżej siebie znajdują się dwie nieruchomości, tym większe prawdopodobieństwo utworzenia między nimi połączenia. Grubość krawędzi odpowiada odwrotności odległości między punktami – im większa waga, tym bliższe są sobie przestrzennie połączone lokalizacje.



Przygotowanie modelu Graph Convolutional Network

W dalszym etapie opracowano model GCN, którego zadaniem było przetwarzanie danych przestrzennych nieruchomości z uwzględnieniem ich wzajemnych powiązań w grafie. Dane wejściowe zostały uprzednio wystandaryzowane za pomocą StandardScaler, a następnie przekształcone do formatu zgodnego z biblioteką PyTorch Geometric. Każdy wierzchołek grafu odpowiada jednemu mieszkaniu i opisany jest przez 19 cech numerycznych. Krawędzie między wierzchołkami oraz przypisane im wagi odwzorowują zależności przestrzenne między nieruchomościami.

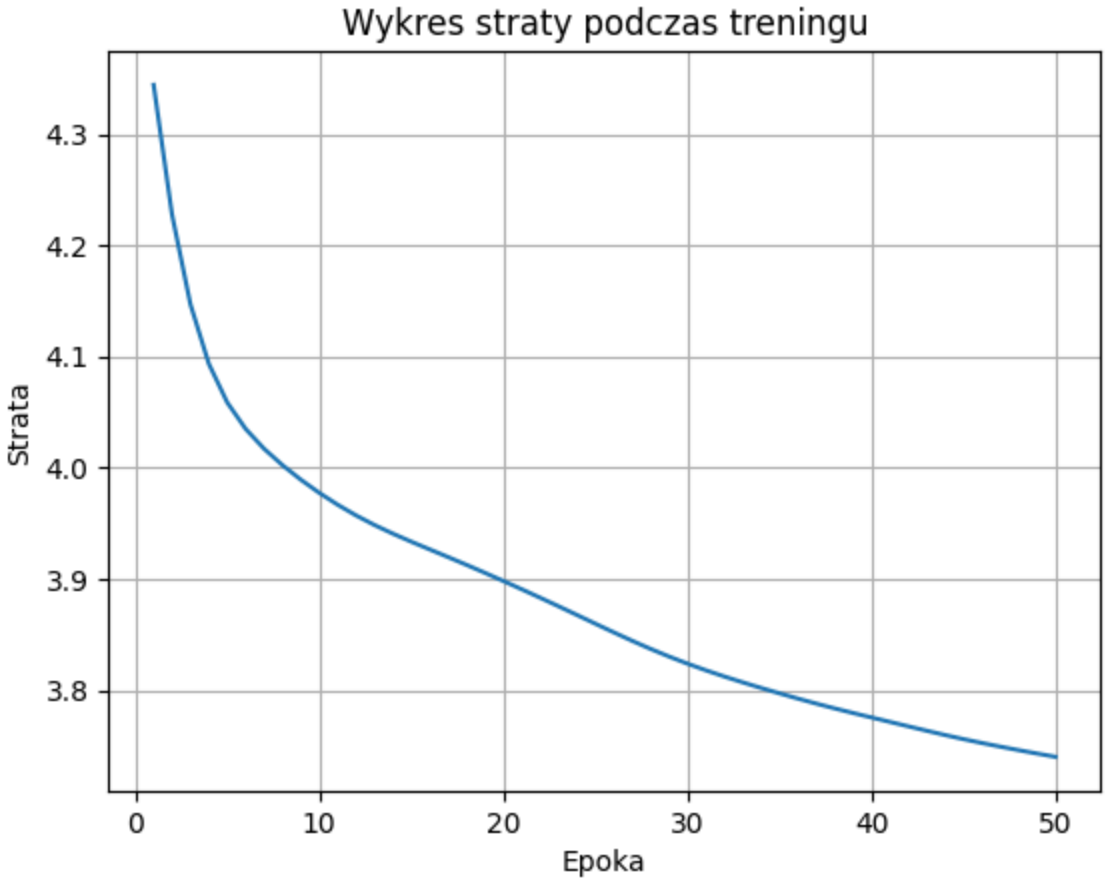
Model GCN zbudowano jako dwuwarstwową sieć grafową: pierwsza warstwa przekształca dane wejściowe do przestrzeni ukrytej z 16 ukrytymi channels, a druga przywraca wymiar wyjściowy do poziomu oryginalnego, czyli 19 channels.

Użyto Adamoptimizer z learning rate = 0,01, co pozwoliło na efektywne dopasowanie wag do struktury grafu i cech mieszkań.

Trening modelu

Trening przeprowadzano przez 50 epok, a funkcję straty liczono jako średnią długości wektora różnicy przekształconych cech i oryginalnych cech.

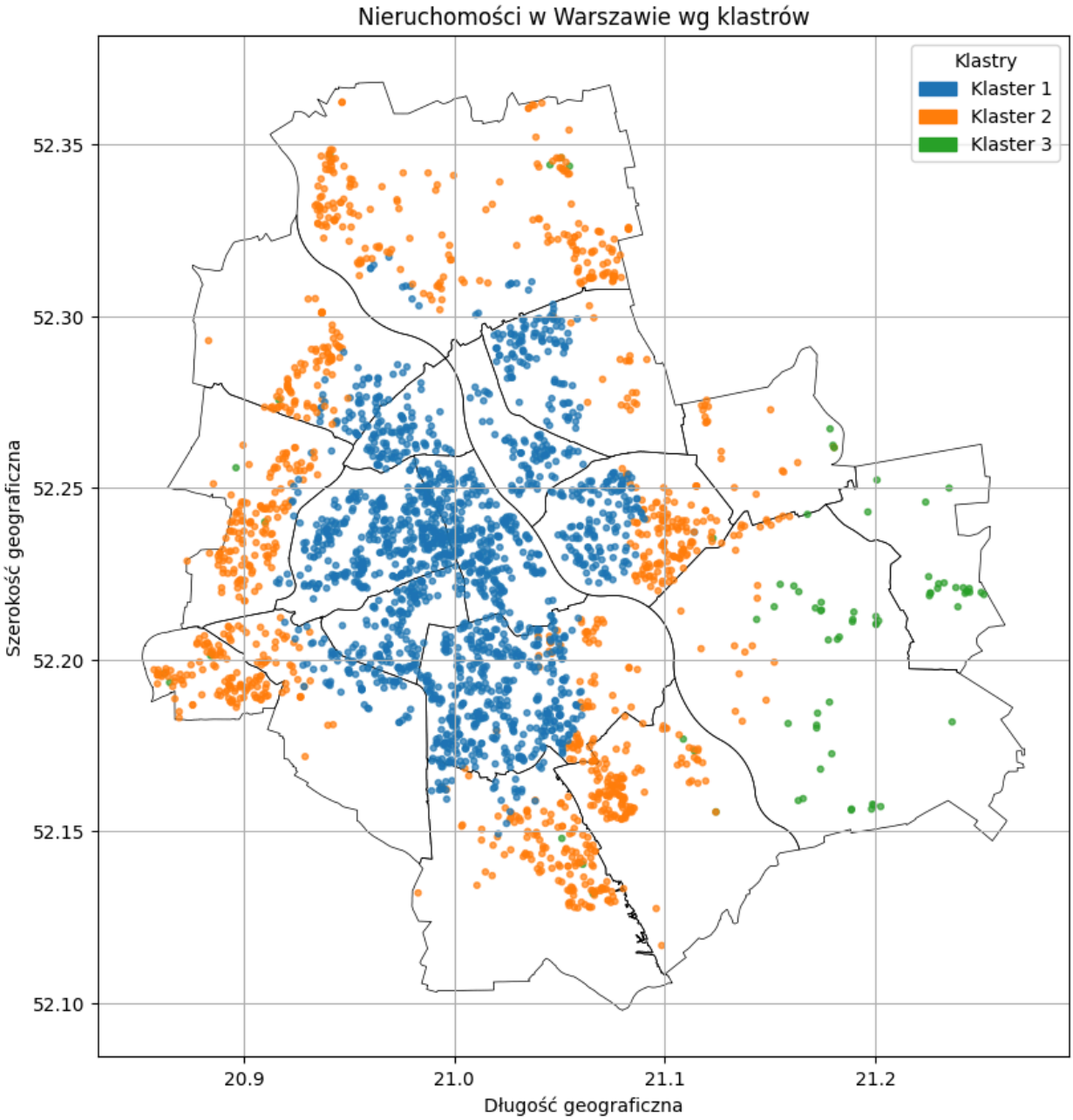
```
loss = torch.mean(torch.norm(out - data.x, dim=1))
```



Grupowanie metodą Kmeans

Po zakończeniu treningu modelu GCN, uzyskane embeddingi dla każdego wierzchołka zostały wykorzystane do dalszej analizy grupowania. W celu wybrania optymalnej liczby klastrów użyto metody łokcia.

W dalszej analizie przyjęto podział na **3 klastry**, a jako miarę jakości grupowania obliczono Silhouette Score, który wyniósł = 0.41539165. Uzyskano zatem klastry umiarkowanie dobrej jakości, również biorąc pod uwagę mnogość danych. Niemniej nie jest to też szczególnie dobry wynik - na pewno jest miejsce do poprawy.



Wnioski

Na mapie przedstawiającej wyniki klasteryzacji można zauważyć trzy wyraźne klastry: centrum miasta, obrzeża oraz dalsze dzielnice peryferyjne, takie jak Wesoła i Rembertów. Taki podział odzwierciedla intuicyjny układ przestrzenny Warszawy — różnice między centrum a obrzeżami są naturalne, zwłaszcza w kontekście cen nieruchomości.

Uzyskany podział może jednak wynikać z prostoty zastosowanego modelu GCN. W porównaniu do metod ekonometrycznych uwzględniających zależności przestrzenne, takich jak GWR i MGWR, model GCN w tej wersji może mniej precyzyjnie odwzorowywać subtelne różnice w strukturze przestrzennej. Dla porównania — modele GWR i MGWR osiągają współczynnik determinacji R^2 na poziomie ok. 0,85, co świadczy o ich wysokiej skuteczności.

W celu poprawy jakości modelowania możliwe jest zastosowanie bardziej złożonej architektury GCN, wypróbowanie innych metod grafowych, takich jak GraphSAGE, lub wykorzystanie bardziej zaawansowanych algorytmów klasteryzacji.