METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXVII/1, 2026, s. 41 – 54

Received: 11.02.2026
Accepted: 20.03.2026

ANALIZA WPŁYWU LOKALIZACJI I CECH NIERUCHOMOŚCI NA ICH CENY W POLSCE W LATACH 2023-2024

Angelika Samson
Wydział Zastosowań Informatyki i Matematyki
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: angelika_samson@sggw.edu.pl
Monika Zielińska-Sitkiewicz https://orcid.org/0000-0003-4829-3239
Instytut Ekonomii i Finansów
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: monika_zielinska-sitkiewicz@sggw.edu.pl

Streszczenie: W artykule przeprowadzono badanie wpływu metrażu, lokalizacji i udogodnień na ceny ofertowe mieszkań w Polsce w latach 2023-2024, ze szczególnym uwzględnieniem Warszawy, Krakowa, Łodzi i Trójmiasta. Połączono metody ekonometryczne i uczenie maszynowe, aby wskazać kluczowe czynniki cenotwórcze. W analizie zastosowano modele OLS, Lasso, GWR, k-means oraz Random Forest. Badanie oparte na próbie 49 531 ofert wykazuje wyraźną przewagę modeli nieliniowych w predykcji. Potwierdzono również dominującą rolę lokalizacji i cech fizycznych w objaśnianiu przestrzennego zróżnicowania cen na rynku.

Słowa kluczowe: wycena nieruchomości, metraż, lokalizacja, udogodnienia, modele OLS, Lasso, GWR, k-means, Random Forest

JEL classification: R31, C53, C55, R12

WSTĘP

Rynek nieruchomości stanowi jeden z kluczowych sektorów gospodarki, ściśle powiązany z cyklem koniunkturalnym, polityką pieniężną oraz sytuacją dochodową społeczeństwa. W ostatnich latach, w warunkach rosnącej inflacji, zmieniających się preferencji mieszkaniowych oraz rozwoju pracy zdalnej, temat wyceny nieruchomości zyskał szczególne znaczenie. Dostęp do rzetelnych modeli wyceny staje się dziś niezwykle istotny dla rzeczoznawców majątkowych, deweloperów, instytucji finansowych oraz gospodarstw domowych.

https://doi.org/10.22630/MIBE.2026.27.1.3

W perspektywie długookresowej polski rynek mieszkaniowy charakteryzował się systematycznym wzrostem cen, przerywanym okresami spowolnienia gospodarczego (np. globalny kryzys finansowy, pandemia) oraz zmian polityki pieniężnej i warunków finansowania [NBP 2025]. W latach 2023-2024 obserwowano łagodniejsze tempo wzrostu cen, a w I-II kw. 2025 r. sygnały stabilizacji i miejscami niewielkich korekt w części lokalizacji, co NBP wiązał m.in. z wygaszeniem programu dopłat do kredytów oraz dostosowaniami po stronie popytu i podaży [NBP 2024; NBP 2025]. Jednocześnie analizy Polskiego Instytutu Ekonomicznego dla 2024 r. wskazywały na istotne zróżnicowanie dynamiki między miastami oraz różnice pomiędzy rynkiem pierwotnym a wtórnym [PIE 2024]. Ponadto istotnym czynnikiem kształtującym koniunkturę były instrumenty publiczne – funkcjonowanie programu „Bezpieczny Kredyt 2%" w 2023 roku sprzyjało wzrostowi popytu, podczas gdy jego wygaśnięcie z końcem roku osłabiło dynamikę transakcyjną [NBP 2024; UKNF 2024; Szelągowska 2023]. Równolegle uwarunkowania podażowe - koszty gruntów i materiałów, ograniczenia planistyczne oraz dostępność terenów utrwalały istotne zróżnicowanie regionalne i lokalne [NBP 2025; PIE 2024].

W literaturze współczesna analiza rynku nieruchomości opiera się na trzech fundamentalnych podejściach wyceny: porównawczym, dochodowym i hybrydowym [Pagourtzi et al., 2003; Sirmans et al., 2005] przy czym modele hedoniczne pozostają głównym narzędziem badawczym. Teoretyczne podstawy koncepcji cen ukrytych, gdzie wartość nieruchomości jest postrzegana jako suma jej indywidualnych atrybutów, zostały ustanowione przez Rosena [1974], natomiast kompleksowy przegląd ich empirycznych zastosowań i metod selekcji zmiennych przedstawili Sirmans, Macpherson i Zietz [2005]. Jednakże, skuteczne ilościowe określenie wpływu cech fizycznych i lokalizacyjnych na cenę wymaga uwzględnienia autokorelacji przestrzennej, ponieważ zdaniem Anselina [2001] pominięcie zależności sąsiedztwa w modelach SAR lub SEM prowadzi do obciążonych estymatorów i błędnych wniosków. Ponadto, pełne zrozumienie mechanizmów rynkowych wymaga umiejscowienia ich w kontekście makroekonomicznym, gdzie sektor mieszkaniowy jest uznawany za kluczowy wiodący wskaźnik cyklu koniunkturalnego [Leamer 2007] oraz włączenia perspektywy behawioralnej. Jak zauważyli Case i Shiller [2003], dynamika cen jest często napędzana przez czynniki psychologiczne i specyficzne oczekiwania kupujących, co może prowadzić do baniek spekulacyjnych odbiegających od obiektywnych fundamentów ekonomicznych.

W badaniach nad analizą cen nieruchomości szczególną wagę przywiązuje się do metod ilościowych. Tradycyjne indeksy cen często nie są w stanie w pełni oddać dynamiki rynku, gdyż nie uwzględniają zmienności cech nieruchomości [Foryś, 2015]. W celu ograniczenia tych niedoskonałości stosuje się modele oparte na regresji hedonicznej, dekomponujące cenę na wartości poszczególnych atrybutów, takich jak powierzchnia, rok budowy czy lokalizacja. Badania Foryś [2015] nad rynkiem szczecińskim potwierdziły skuteczność tej metody, wykazując dominujący

wpływ powierzchni użytkowej na ceny mieszkań. Metoda ta posiada jednak ograniczenia w przypadku gwałtownych zmian rynkowych i wymaga dużych, homogenicznych zbiorów danych.
Z drugiej strony Dąbrowski [2010] udokumentował, że podejście oparte wyłącznie na lokalnych cechach nieruchomości jest niewystarczające do opisu złożonych zjawisk rynkowych. Wprowadzenie do modeli tzw. atrybutów globalnych (wskaźników makroekonomicznych i społeczno-gospodarczych, np. inflacji, produkcji przemysłowej) znacząco poprawia trafność prognoz. Dąbrowski wykazał silne powiązania między sytuacją gospodarczą kraju a poziomem cen, postulując potrzebę stosowania wielowymiarowych analiz dynamicznych.
Współczesna literatura coraz częściej odchodzi od klasycznej ekonometrii w stronę zaawansowanych algorytmów. Naji [2024], badając rynek nieruchomości, zaprezentował podejście z wykorzystaniem uczenia maszynowego (m.in. Random Forest, Lasso, Stacking Regressor). Wyniki jego pracy dowiodły, że modele zespołowe (ensemble models) osiągają najwyższą dokładność predykcyjną dzięki efektywnej integracji danych z różnych źródeł (w tym web scrapingu i analityki przestrzennej). Uczenie maszynowe pozwala na lepsze radzenie sobie ze złożonym i nieliniowym charakterem rynków, precyzyjniej oceniając wpływ lokalizacji, metrażu i udogodnień.

W kontekście zmian rynkowych i ewolucji metodologicznej, analiza empiryczna mechanizmów wyceny na polskim rynku zyskuje szczególne znaczenie. Celem artykułu jest empiryczne zbadanie stopnia, w jakim metraż, lokalizacja oraz wybrane udogodnienia determinują ceny ofertowe mieszkań w Polsce, ze szczególnym uwzględnieniem największych rynków: Warszawy, Krakowa, Łodzi i Trójmiasta w latach 2023-2024. Badanie łączy tradycyjne podejście ekonometryczne z nowoczesnymi metodami uczenia maszynowego, aby zidentyfikować czynniki o największej sile wyjaśniającej w warunkach silnego zróżnicowania przestrzennego i makroekonomicznego polskiego rynku mieszkaniowego.

METODOLOGIA I DANE

Materiał badawczy stanowił pozyskany z platformy Kaggle publiczny zbiór 49 531 ogłoszeń z polskiego rynku nieruchomości [Jamróz 2024]. Na potrzeby realizacji celu naukowego zbiór ograniczono wyłącznie do ofert sprzedaży opublikowanych w okresie od sierpnia 2023 r. do czerwca 2024 r., wykluczając rekordy dotyczące najmu. Takie zawężenie pozwoliło na precyzyjną analizę czynników kształtujących ceny rynkowe oraz badanie ich zróżnicowania przestrzennego. Przed przystąpieniem do modelowania dane poddano procedurom oczyszczania, standaryzacji i transformacji, co umożliwiło ich efektywne wykorzystanie w analizach statystycznych.

W badaniu uwzględniono pełne spektrum zmiennych pozwalających na wielowymiarowy opis nieruchomości. Zmienną zależną była cena ofertowa mieszkania wyrażona w PLN (price). Zbiór zmiennych niezależnych obejmował:

- parametry fizyczne i strukturalne: powierzchnia użytkowa w m² (squareMeters), rok oddania budynku do użytku (buildYear), piętro, na którym znajduje się lokal (floor), oraz całkowita liczba kondygnacji w budynku (floorCount);
- atrybuty budynkowe: typ zabudowy (buildingType, np. blok, kamienica) oraz materiał konstrukcyjny (buildingMaterial, np. cegła, żelbeton);
- zmienne lokalizacyjne i przestrzenne: miasto zakodowane binarnie (city, m.in. Warszawa, Kraków, Gdańsk i Łódź), współrzędne geograficzne (latitude, longitude), odległość od ścisłego centrum miasta w kilometrach (centreDistance) oraz wskaźnik dostępności usług mierzony jako średnia odległość do najbliższych punktów użyteczności publicznej (avgPOIDistance);
- udogodnienia (zmienne binarne): posiadanie miejsca parkingowego (parking_space), balkonu (balcony), windy (elevator), ochrony (security) oraz komórki lokatorskiej (storage_room);
- czas: miesiąc (month) oraz rok (year) publikacji ogłoszenia.

Proces badawczy zrealizowano w środowisku Python, wykorzystując biblioteki pandas, numpy, scikit-learn, statsmodels, matplotlib oraz geopandas. Obejmował on następujące etapy:

1. Eksploracyjna analiza danych (EDA) - przeprowadzono winsoryzację wartości odstających w rozkładach cen i powierzchni. Obliczono podstawowe miary statystyczne oraz macierz korelacji Pearsona. Zidentyfikowano i usunięto zmienne silnie skorelowane w celu redukcji kolinearności [Ligas & Czaja 2010; Śleszyński 2020; Chaim & Łukasik 2024; Rana & Singhal 2015].
2. Modelowanie regresyjne - estymowano model klasycznej regresji liniowej (OLS) w celu identyfikacji globalnego wpływu predyktorów na zmienną zależną. Równolegle zastosowano regresję Lasso (Least Absolute Shrinkage and Selection Operator) wykorzystując regularyzację L_1 w celu jednoczesnej estymacji i selekcji zmiennych [Berry & Feldman 1985; Bun & Harrison 2019; Fatih 2024]
3. Segmentacja rynku (klasteryzacja) - wykorzystano algorytm k-means do identyfikacji homogenicznych grup nieruchomości. Optymalna liczba klastrów została określona metodą „łokcia" (elbow method), co zapewniło kompromis między liczbą grup a ich spójnością [MacQueen 1967; Arthur & Vassilvitskii 2007; Huang & Lai 2023].
4. Analityka przestrzenna (GWR) - zastosowano lokalną regresję ważoną geograficznie (Geographically Weighted Regression – GWR) z adaptacyjnym pasmem umożliwiającym estymację współczynników lokalnych dla poszczególnych lokalizacji. Dodatkowo zweryfikowano występowanie

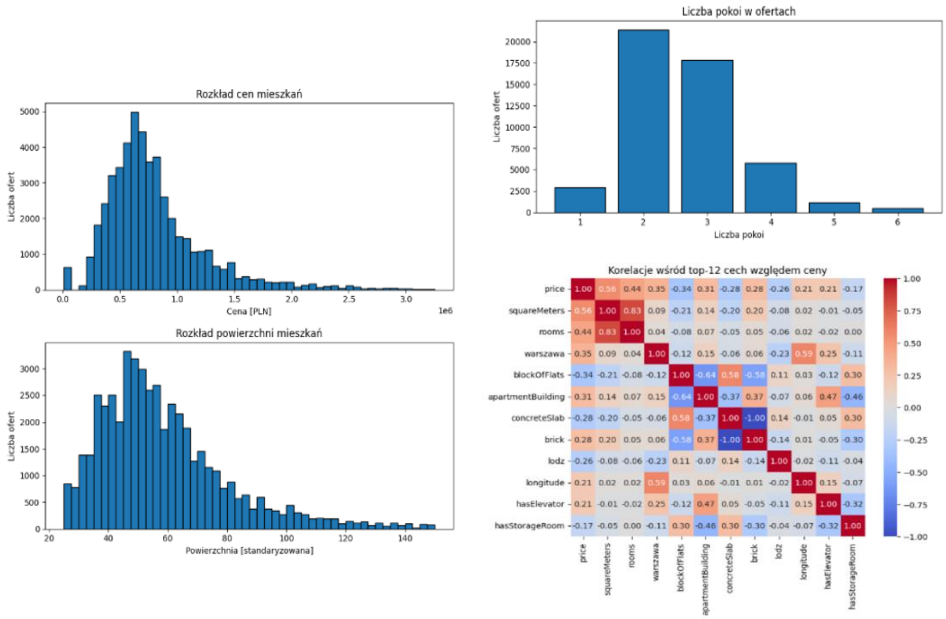
autokorelacji przestrzennej przy użyciu indeksu Morana (Moran's I) [Brunsdon, Fotheringham & Charlton 1996, 2000, 2006; Matthews & Yang 2012].

5. Modelowanie nieliniowe (Random Forest) - w celu uchwycenia złożonych i nieliniowych zależności między atrybutami a ceną, zbudowano model lasu losowego z wykorzystaniem pięciokrotnej walidacji krzyżowej. Dokonano oceny jakości predykcji na zbiorze testowym (miary R^2 i RMSE), a także przeanalizowano ważność poszczególnych cech [Breiman 2001].
6. Analiza porównawcza aglomeracji - dla czterech największych aglomeracji w Polsce opracowano odrębne modele OLS i Random Forest. Porównano wartości współczynników β, miary dopasowania R^2 oraz błędy RMSE.

WYNIKI BADAŃ

Eksploracyjną analizą danych scharakteryzowano rozkłady najważniejszych zmiennych oraz zidentyfikowano potencjalne odchylenia od typowych wartości rynkowych. Na podstawie całego zbioru obliczono medianę ceny - 525 000 PLN oraz medianę powierzchni mieszkania 48 m². Dane te potwierdzają, że analizowany zbiór obejmuje w przeważającej mierze mieszkania średniej wielkości, o przeciętnej wartości transakcyjnej.

Rysunek 1. Charakterystyka rozkładów i korelacja badanych zmiennych

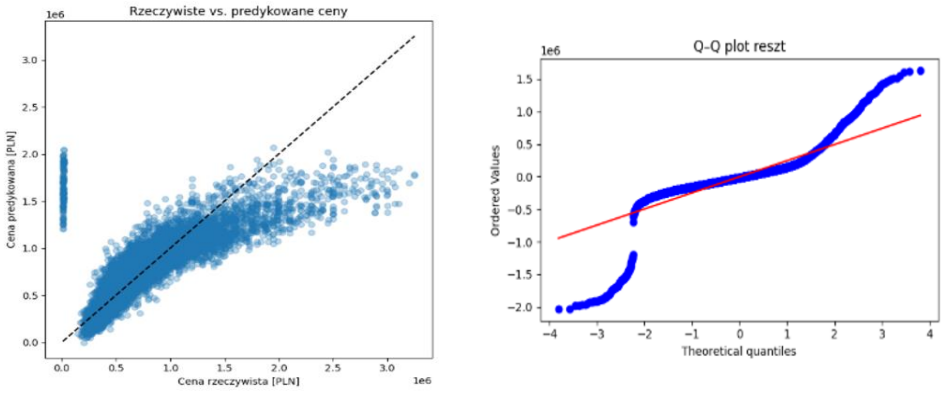


Źródło: Obliczenia i opracowanie własne

Na rysunku 1 przedstawiono rozkłady cen ofertowych oraz powierzchni charakteryzujące się wyraźną prawoskośnością. Występowanie dodatniej asymetrii świadczy o rynkowej dominacji lokali o niższych i średnich parametrach oraz obecności nielicznej grupy ofert o skrajnie wysokich wartościach cenowych i powierzchniowych, przy jednoczesnej przewadze rynkowej mieszkań dwu- i trzypokojowych. Na podstawie analizy korelacji (heatmapa na rys. 1) stwierdzono, że cena jest najsilniej powiązana z metrażem (r=0,84), liczbą pokoi (r=0,60) oraz obecnością windy (r=0,55). W celu wyeliminowania multikolinearności, wynikającej z silnej zależności między zmiennymi squareMeters i rooms oraz tożsamości cech brick i concreteSlab, przeprowadzono redukcję wymiarów. Usunięcie zmiennej rooms (na rzecz squareMeters o wyższej sile objaśniającej) oraz concreteSlab pozwoliło na zapewnienie stabilności numerycznej i poprawę interpretowalności estymowanych modeli.

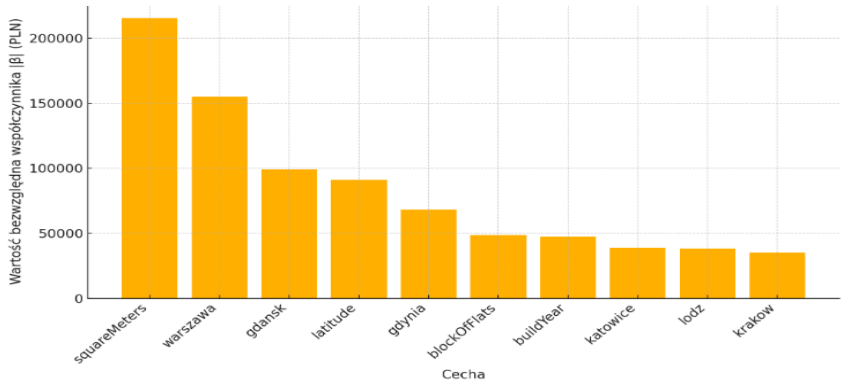
W celu oszacowania determinant cen ofertowych zastosowano klasyczny model najmniejszych kwadratów (OLS) oraz regresję Lasso z regularyzacją L_1. Model OLS uzyskał dopasowanie na poziomie $R^2 = 0{,}57$ oraz RMSE $\approx$ 286 379 PLN, poprawnie odwzorowując środek rozkładu, przy jednoczesnej tendencji do niedoszacowania wartości najdroższych nieruchomości. Analiza reszt wykazała występowanie heteroskedastyczności oraz odstępstwa od rozkładu normalnego, co jest zjawiskiem typowym dla danych rynkowych o nieliniowych zależnościach (Rysunek 2). Wyniki regresji Lasso ($\alpha \approx 0{,}015$) okazały się zbliżone do OLS: $R^2 = 0{,}565$, RMSE $\approx$ 286 718 PLN, co potwierdziło stabilność oszacowań. Zidentyfikowano, że kluczowym predyktorem ceny jest metraż mieszkania (squareMeters), a wśród zmiennych lokalizacyjnych największy wpływ wykazują aglomeracje Warszawy, Gdańska i Gdyni (Rysunek 3). Istotnymi czynnikami różnicującymi ceny okazały się również parametry geograficzne (latitude), rok budowy (buildYear) oraz typ zabudowy.

Rysunek 2. Zależność pomiędzy cenami rzeczywistymi a predykowanymi (OLS) oraz analiza reszt modelu OLS



Źródło: Obliczenia i opracowanie własne

Rysunek 3. Dziesięć cech o największych modułach współczynników (Lasso)

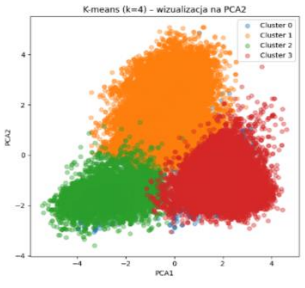


Źródło: Obliczenia i opracowanie własne

Identyfikując homogeniczne grupy ofert o zbliżonej charakterystyce rynkowej zastosowano algorytm k-means. Optymalną liczbę klastrów (k=4) wyznaczono metodą „łokcia", analizując spadek inercji wewnątrzklasowej. Wyniki segmentacji, przedstawione w tabeli 1, pozwoliły na wyodrębnienie zróżnicowanych podgrup mieszkań różniących się istotnie pod względem parametrów cenowych, metrażu oraz lokalizacji. Klaster 0 identyfikuje segment mieszkań o charakterze ekonomicznym, w którym dominują lokale o relatywnie niższych metrażach i cenach ofertowych, zlokalizowane głównie w blokach mieszkalnych w niewielkiej odległości od punktów usługowych. Klaster 1 reprezentuje segment ekonomiczny o najniższej średniej cenie (606 tys. PLN), natomiast Klaster 3 grupuje lokale o najwyższym standardzie cenowym (980 tys. PLN) i największej powierzchni średniej (63,11 m^2). Z kolei Klaster 2 charakteryzuje się najmniejszą średnią odległością od punktów usługowych (0,364 km), co sugeruje lokalizacje w ścisłych centrach miast. Separację i strukturę wyodrębnionych segmentów potwierdza wizualizacja w przestrzeni dwóch głównych składowych (PCA), wskazująca na wyraźne zróżnicowanie między rynkiem masowym a segmentem premium.

Tabela 1 Charakterystyka klastrów rynku mieszkaniowego

Klaster	Średnia cena (PLN)	Mediana ceny (PLN)	Średni metraż m^2	Średnia odległość od POI	Średnia dl. geogr. E	Średnia szer. geogr. N
0	870499,61	775000	58,85	0,638	18,62	54,37
1	605897,55	580000	52,46	0,586	19,62	51,76
2	857588,93	735000	62,14	0,364	19,25	51,82
3	979864,74	875000	63,11	0,720	19,89	51,76



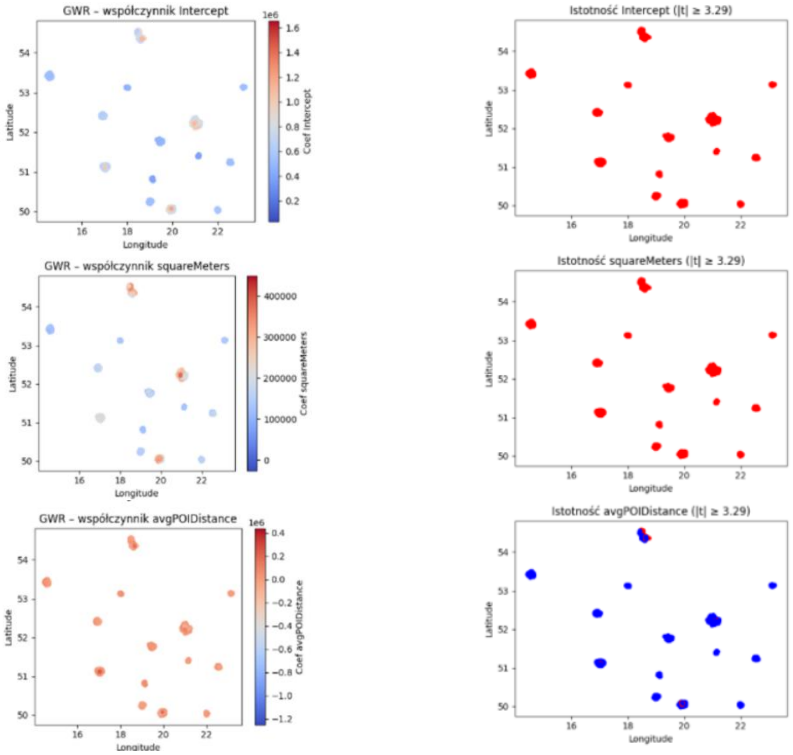
Źródło: Obliczenia własne

Uwzględnienie niestacjonarności przestrzennej mechanizmów kształtowania cen za pomocą lokalnej regresji ważonej geograficznie GWR pozwoliło na zidentyfikowanie najwyższej ceny bazowej (wyrazu wolnego) w centralnych obszarach aglomeracji oraz najsilniejszego wpływu metrażu (squareMeters) na premię cenową w strefach o najwyższym stopniu zurbanizowania. Zaobserwowano również regionalną zmienność wpływu dostępności usług (avgPOIDistance), którego oddziaływanie na cenę jest wyraźniejsze w południowo-zachodniej części badanego obszaru. Weryfikacja istotności statystycznej lokalnych współczynników regresji ($\alpha = 0{,}001$, $|t| \geq 3{,}29$) umożliwiła precyzyjne określenie przestrzennego zasięgu oddziaływania poszczególnych zmiennych. Stwierdzono, że lokalny wyraz wolny zachowuje istotność niemal w całym obszarze badań, co potwierdza powszechność lokalnego efektu cenowego, występującego niezależnie od cech fizycznych nieruchomości. Z kolei wpływ metrażu koncentruje się wokół głównych ośrodków miejskich i ich przedmieść. Istotność dostępności usług wykazuje natomiast charakter punktowy i ogranicza się głównie do południowo-zachodniej części obszaru, co sugeruje, że w tym regionie bliskość punktów usługowych stanowi krytyczny czynnik kształtujący ceny. Wyniki te dowodzą, że choć większość cech wykazuje znaczenie globalne, siła ich oddziaływania ulega istotnemu zróżnicowaniu przestrzennemu, osiągając najwyższą stabilność statystyczną w rejonach o dużej intensywności procesów rynkowych (Rysunek 4). Diagnostyka modelu przy użyciu statystyki Morana wykazała niewielką, choć istotną autokorelację reszt (Moran's $I = 0{,}04$ $p<0{,}01$), co przy tak niskiej wartości współczynnika potwierdza skuteczność modelu GWR w eliminowaniu efektów sąsiedztwa i poprawnym ujęciu struktury przestrzennej danych.

Wykorzystanie modelu lasu losowego umożliwiło uwzględnienie nieliniowych zależności oraz złożonych interakcji zachodzących między predyktorami. Algorytm wytrenowany przy użyciu 5-krotnej walidacji krzyżowej (z parametrami n_estimators = 500, max_depth = None oraz max_features = 'sqrt') wykazał wysoką zdolność predykcyjną i stabilność, osiągając współczynnik determinacji $R^2 = 0{,}742$ na zbiorze testowym oraz błąd RMSE = 220 729 PLN. Minimalna rozbieżność między wynikami walidacji ($0{,}745 \pm 0{,}01$) a zbiorem testowym, przy błędzie OOB na poziomie 0,75, potwierdza wysoką zdolność generalizacji i brak przeuczenia modelu. Analiza ważności zmiennych ujawniła dominującą rolę metrażu (squareMeters) w kształtowaniu cen ofertowych. Kluczowe znaczenie przypisano również atrybutom przestrzennym (współrzędne geograficzne, odległość od centrum, lokalizacja w Warszawie) oraz strukturalnym (buildYear) (Rysunek 5).

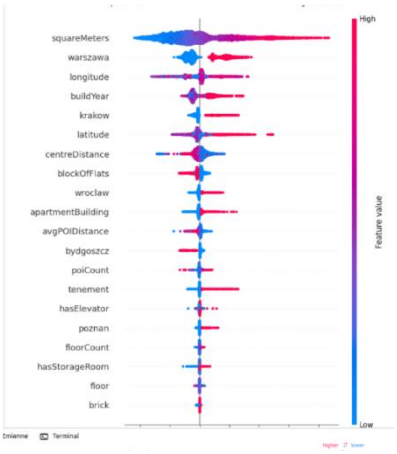
Analiza porównawcza czterech kluczowych rynków — Warszawy, Trójmiasta, Krakowa i Łodzi — pozwoliła na identyfikację istotnych różnic w strukturze cenowej oraz hierarchii determinant wartości mieszkań. Podczas gdy Warszawa i Trójmiasto charakteryzują się najwyższymi medianami cen oraz znaczną heterogenicznością powierzchniową, rynki w Krakowie i Łodzi wykazują relatywnie większą jednorodność ofert.

Rysunek 4 Rozkład lokalnych współczynników regresji ważonej geograficznie (wyraz wolny, metraż, odległość od POI) wraz z mapami istotności parametrów.



Źródło: Obliczenia własne

Rysunek 5 Ważność cech w modelu Random Forest



Źródło: obliczenia własne

W czterech badanych miastach odnotowano prawoskośność rozkładów cen i metrażu, typową dla sektora mieszkaniowego o przewadze lokali o mniejszej powierzchni. Porównanie efektywności modeli OLS i Random Forest (RF) jednoznacznie wskazuje na wyższą zdolność predykcyjną algorytmów zespołowych, które pozwoliły na redukcję błędu RMSE oraz wzrost współczynnika R^2 średnio o 0,25–0,30, przy czym największą poprawę dopasowania odnotowano dla Warszawy i Trójmiasta.

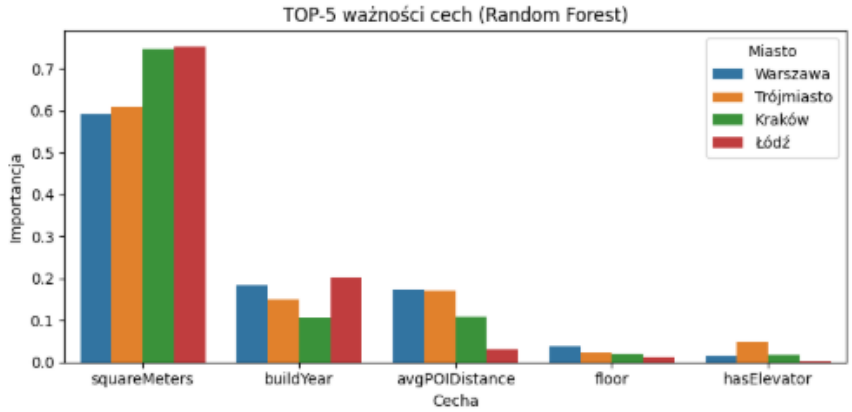
Tabela 2 Wyniki R^2 i RMSE dla modeli OLS i Random Forest

Miasto	N ofert	R^2 OLS	RMSE OLS	R^2 RF	RMSE RF
Warszawa	17481	0,285	426 389	0,608	314 922
Trójmiasto	5368	0,621	247 463	0,828	165 927
Kraków	7494	0,646	239 267	0,781	202 131
Łódź	4414	0,781	76 108	0,885	58 109

Źródło: obliczenia własne

Mimo różnic lokalnych, metraż (squareMeters) pozostaje najsilniejszym predyktorem ceny w każdym z analizowanych miast. Interesujące zróżnicowanie zaobserwowano w przypadku dostępności infrastruktury usługowej - o ile w Warszawie, Trójmieście i Krakowie bliskość usług generuje istotną premię cenową (ujemny współczynnik β), o tyle w Łodzi zależność ta przyjmuje kierunek dodatni, co sugeruje odmienną specyfikę przestrzenną tego ośrodka. Analiza ważności cech w modelach RF (Rysunek 6) potwierdza dominację metrażu, wskazując jednocześnie na istotną, choć zróżnicowaną rolę roku budowy (buildYear) oraz dostępności punktów usługowych, przy marginalnym wpływie pozostałych atrybutów fizycznych, takich jak piętro czy obecność windy.

Rysunek 6 Ważność cech w modelu Random Forest dla czterech miast



Źródło: obliczenia własne

PODSUMOWANIE

Celem pracy było empiryczne zbadanie, w jakim stopniu metraż, lokalizacja i wybrane udogodnienia determinują ceny ofertowe mieszkań w Polsce oraz w wybranych dużych miastach (Warszawa, Trójmiasto, Kraków, Łódź). Analiza została przeprowadzona na zbiorze 49 531 unikalnych ofert z lat 2023–2024, po uprzednim oczyszczeniu danych (eliminacja duplikatów, imputacja braków, winsoryzacja 1% skrajnych wartości) oraz podziale na zbiory uczący i testowy. Wykazano prawoskośność rozkładów cen i metraży, dominację mieszkań 2–3-pokojowych oraz różnice cen między typami budynków. Redukcja cech ograniczyła multikolinearność (m.in. rezygnacja z concreteSlab oraz pozostawienie squareMeters zamiast rooms).

Wyniki potwierdziły, że metraż jest najsilniejszym predyktorem ceny (korelacja Pearsona z ceną r = 0,84; dominująca ważność w Random Forest). Wskazano jednocześnie zróżnicowanie przestrzenne znaczenia determinant: w Warszawie przeważają czynniki lokalizacyjne (m.in. odległość od centrum, gęstość usług), podczas gdy w Łodzi większą rolę odgrywają wielkość i cechy budynku. Modele globalne (OLS, Lasso) wyjaśniały ok. 57% wariancji cen ($R^2 \approx 0{,}57$; RMSE ≈ 286 tys. PLN), z typową dla danych rynkowych heteroskedastycznością reszt. Uwzględnienie niestacjonarności przestrzennej w GWR podniosło dopasowanie do $R^2 = 0{,}62$ (RMSE ≈ 268 tys. PLN), co potwierdziło istotność lokalnych efektów. Najwyższą trafność predykcji uzyskano w Random Forest (R^2 test = 0,742; RMSE ≈ 220 729 PLN; CV $R^2 = 0{,}745 \pm 0{,}010$; OOB ≈ 0,75), co wskazuje, że nieliniowe zależności i interakcje są na tym rynku znaczące. Segmentacja k-means wyodrębniła cztery spójne klastry (od segmentu ekonomicznego po premium), potwierdzając heterogeniczność struktury popytowo-podażowej.

Uzyskane wyniki empirycznie potwierdzają mechanizmy formowania się cen mieszkań w Polsce w latach 2023–2024 oraz wskazują, że połączenie klasycznych metod ekonometrycznych z modelami uczenia maszynowego (zwłaszcza Random Forest) stanowi skuteczne podejście do analizy złożonych, przestrzennie zróżnicowanych danych rynkowych.

BIBLIOGRAFIA

Anselin L. (2001) Spatial Econometrics [in:] B. H. Baltagi (red.) A Companion to Theoretical Econometrics, Chapter 14, 310-330, Blackwell Publishing Ltd.

Arthur D., Vassilvitskii S. (2007) K-means++: The Advantages of Careful Seeding, SODA ’07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, New Orleans Louisiana, 1027-1035.

Berry W. D., Feldman S. (1985) Multiple Regression in Practice. Sage University Paper series on Quantitative Applications in the Social Sciences, Series No. 07-050, Sage, Newbury Park.

Breiman L. (2001) Random Forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Brunsdon C., Fotheringham A. S., Charlton M. (1996) Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 28(4), 281-298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Brunsdon C., Fotheringham A. S., Charlton M. (2000) Geographically Weighted Regression as a Statistical Model. University of Newcastle-upon-Tyne: Newcastle upon Tyne, UK.

Bun M. J. G., Harrison T. D. (2019) OLS and IV Estimation of Regression Models Including Endogenous Interaction Terms. Econometric Reviews, 38(7), 814-827. https://doi.org/10.1080/07474938.2018.1427486

Case K. E., Shiller R. J. (2003) Is There a Bubble in the Housing Market? Brookings Papers on Economic Activity, 34(2), 299-362.

Chaim A., Łukasik N. (2024) Analiza korelacji w stosunkach międzynarodowych na przykładzie wybranych aspektów stosunków polsko-niemieckich. Krakowskie Studia Małopolskie, 3, 83-110

Charlton M., Fotheringham A. S. Brunsdon C. (2006) Geographically Weighted Regression: NCRM Methods Review Papers/NCRM/006. http://eprints.ncrm.ac.uk/90/ (dostęp: 12.01.2026)

Dąbrowski J. (2010) Zastosowanie wybranych metod statystycznych do analizy rynku nieruchomości. StatSoft Polska. http://www.statsoft.pl/Portals/0/Downloads/Zast_met_stat_analizy_rynku_nieruchomosci.pdf.

Fatih C. (2024) Lasso Regression Model [Technical Report]. University Ferhat Abbas of Sétif. https://doi.org/10.13140/RG.2.2.23949.14563

Foryś I. (2015) Indeks hedoniczny na wtórnym rynku mieszkań spółdzielczych na przykładzie wybranego osiedla w Szczecinie. Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego, 42(1), 149-159.

Huang Z., Lai, G (2023) A House Price Prediction Model Based on K-means Clustering and Random Forest in Guangzhou. Frontiers in Business, Economics and Management, 10(2), 377-381. https://doi.org/10.54097/fbem.v10i2.11077

Jamróz K. (2024) Apartment prices in Poland [zbiór danych]. Kaggle. https://www.kaggle.com/datasets/krzysztofjamroz/apartment-prices-in-poland

Leamer E. E. (2007) Housing IS the Business Cycle. NBER Working Paper 13428. https://doi.org/10.3386/w13428

Ligas M., Czaja, J. (2010) Zaawansowane metody analizy statystycznej rynku nieruchomości. Studia i Materiały Towarzystwa Naukowego Nieruchomości, 18(1), 7-19.

Matthews S. A., Yang T.-C. (2012) Mapping the Results of Local Statistics: Using Geographically Weighted Regression. Demographic Research, 26, 151-166. https://doi.org/10.4054/DemRes.2012.26.6

MacQueen J. B. (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1: Statistics, University of California Press, Berkeley, 281-297. http://projecteuclid.org/euclid.bsmsp/1200512992.

Naji M. (2024) Real estate market analysis and prediction using machine learning. Master's thesis, Faculty of Science, Department of Applied Mathematics. https://www.esrilebanon.com/content/dam/esrisites/en-us/education/higher-education/Masters/Projects/2024/Naji/Thesis_Report_MohamadNaji.pdf

NBP (2024) Informacja o cenach mieszkań i sytuacji na rynku nieruchomości w II i IV kw.2024 r. https://nbp.pl/wp-content/uploads/2025/03/Informacja-o-cenach-mieszkan-w-IV-2024.pdf

NBP (2025) Informacja o cenach mieszkań i sytuacji na rynku nieruchomości mieszkaniowych i komercyjnych w Polsce w I kw. 2025 r. https://nbp.pl/wp-content/uploads/2025/06/Informacja-o-cenach-mieszkan-i-sytuacji-na-rynku-nieruchomosci-mieszkaniowych-i-komercyjnych-w-Polsce-w-I-kw.-2025-r.pdf?

Pagourtzi E., Assimakopoulos V., Hatzichristos T., French N. (2003) Real Estate Appraisal: A Review of Valuation Methods. Journal of Property Investment & Finance, 21(4), 383-401. https://doi.org/10.47772/IJRISS.2026.100300057

PIE – Polski Instytut Ekonomiczny (2024) Analiza rynku mieszkaniowego – IV kw. 2024 r. https://pie.net.pl/wp-content/uploads/2025/02/Rynek-mieszk-4_kwartal_2024.pdf

Rana R. K., Singhal R. (2015) Chi-square Test and its Application in Hypothesis Testing. Journal of the Practice of Cardiovascular Sciences, 1(1), 69-71. https://doi.org/10.4103/2395-5414.157577

Rosen S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82(1), 34-55.

Sirmans G. S., Macpherson D. A., Zietz E. N. (2005) The Composition of Hedonic Pricing Models. Journal of Real Estate Literature, 13(1), 3-43.

Szelągowska A. (2023) Bezpieczny kredyt 2% oraz konto mieszkaniowe jako nowe instrumenty polityki mieszkaniowej w Polsce. Studia BAS, 4, 55-84. https://doi.org/10.31268/StudiaBAS.2023.30

Śleszyński Z. (2020) Wyznaczanie współczynników korelacji liniowej – podstawy. Wiadomości Statystyczne, 65(6), 69-87.

UKNF (2024) Informacja na temat sytuacji sektora bankowego w 2023 r. Pobrane z:
https://www.knf.gov.pl/knf/pl/komponenty/img/Informacja_na_temat_sytuacji_sektora_bankowego_w_2023_91099.pdf

ANALYSIS OF THE IMPACT OF LOCATION AND PROPERTY CHARACTERISTICS ON REAL ESTATE PRICES IN POLAND IN 2023–2024

Abstract: This article examines the impact of floor space, location, and amenities on the asking prices of apartments in Poland between 2023 and 2024, with a particular focus on Warsaw, Krakow, Lodz, and the Tricity. The study combines econometric methods and machine learning to identify key price-forming factors. The analysis employs OLS, Lasso, GWR, k-means, and Random Forest models. Based on a sample of 49,531 listings, the results demonstrate a clear superiority of non-linear models in price prediction. The

findings also confirm the dominant role of location and physical characteristics in explaining spatial price differentiation in the real estate market.

Keywords: real estate valuation, floor space, location, amenities, OLS, Lasso, GWR, k-means, Random Forest

JEL classification: R31, C53, C55, R12

CC BY NC