



Studium Magisterskie

Kierunek: Metody ilościowe w ekonomii i systemy informacyjne
Specjalność: Ekonometria

Imię i nazwisko autora
Yauheni Semianiuik
Nr albumu 82591

Porównanie modeli ekonometrii przestrzennej i uczenia maszynowego do prognozowania cen mieszkań w Warszawie

Praca magisterska
pod kierunkiem naukowym
dr. hab. Andrzeja Torója, prof. SGH
Instytut
Ekonometrii

Warszawa 2023

Spis treści

Wstęp	3
I. Opis problemu prognozowania cen mieszkań	6
1.1. Prognozowanie cen mieszkań. Istotność i zastosowania we współczesnej ekonomii	6
1.2. Zjawiska specyficzne dla danych przestrzennych	7
1.3. Przegląd literatury	9
1.3.1. Przegląd modeli	9
1.3.2. Przegląd zmiennych	13
II. Metodyka badania	16
2.1. Ekonometryczne modele przestrzenne	16
2.1.1. Autoregresyjne modele przestrzenne	17
2.1.1.1. Model czystej autoregresji przestrzennej	17
2.1.1.2. Model autoregresji przestrzennej z dodatkowymi regresorami	18
2.1.1.3. Model błędu przestrzennego	18
2.1.1.4. Model opóźnienia przestrzennego regressorów	19
2.1.1.5. Rozszerzenia podstawowych autoregresyjnych modeli przestrzennych	19
2.1.2. Bayesowskie przestrzenne modele autoregresyjne	20
2.1.3. Geograficznie ważona regresja	22
2.2. Modele uczenia maszynowego	24
2.2.1. Drzewo decyzyjne. Modele uczenia zespołowego	24
2.2.1.1. Lasy losowe	25
2.2.1.2. Wyjątkowo losowe drzewa	26
2.2.1.3. Metody wzmacniania	26
2.2.2. Sztuczne sieci neuronowe	30
2.2.2.1 Perceptron wielowarstwowy	31
2.2.2.2. Przestrzenne sieci neuronowe. Geograficznie ważona sztuczna sieć neuronowa	33
2.3. Strojenie hiperparametrów. Algorytm genetyczny	36
2.4. Kryterium porównawcze	37
III. Zbiór danych	39
3.1. Gromadzenie i źródła danych	39
3.2. Inżynieria cech	42
3.3. Eksploracyjna analiza danych	42
3.3.1. Zmienne numeryczne	42

3.3.2. Zmienne kategorialne	54
IV. Analiza empiryczna.....	62
4.1. Testowanie autokorelacji przestrzennej.....	62
4.2. Optymalne hiperparametry. Parametry rozkładów a priori	65
4.3. Porównanie użytych modeli	67
4.3.1. Wyliczone kryterium porównawcze	67
4.3.2. Wykresy reszt	68
4.3.3. Istotność zmiennych. Współczynniki w modelu GWR.....	76
Zakończenie	79
Bibliografia.....	81
Spis rysunków	87
Spis tabeli	89
Streszczenie	90

Wstęp

Rynek nieruchomości jest często uważany za kamień węgielny społecznego dobrobytu, będąc jednym z 20 europejskich filarów praw socjalnych wyznaczonych przez Komisję Europejską i organizacje pokrewne.¹² W XXI stuleciu sektor mieszkaniowy jest bastionem o dużym znaczeniu ekonomicznym, odzwierciedlającym nie tylko puls systemów finansowych, ale także pragnienia i źródła utrzymania milionów ludzi. Aktualna dynamika cen mieszkań i wielopłaszczyznowe aspekty ich dostępności w największych miastach świata stanowią nieodparte wyzwanie zarówno dla ekonomistów, jak i końcowych decydentów kupna i sprzedaży.

Powstanie dużych zbiorów danych, obliczeń o wysokiej wydajności i zaawansowanych metod uczenia maszynowego daje niespotykane dotąd możliwości modelowania wartości niematerialnych mieszkań, co może usprawnić szacowanie ich ceny. Większe ilości, prędkości, odmiany i prawdziwości danych georeferencyjnych tworzonych aktywnie i pasywnie przez użytkowników zapewniają pełniejszy wgląd w przedstawianie środowisk społeczno-ekonomicznych w erze dobrowolnych informacji geograficznych (*volunteered geographic informations*, VGI)³ i dużych danych geograficznych.⁴ Na przykład, uwzględnienie zdjęć⁵ umożliwia lepsze scharakteryzowanie sąsiedztwa mieszkania z perspektywy człowieka. Ponadto szerokie rozpowszechnienie urządzeń ze wbudowanym GPS (głównie telefonów komórkowych i pojazdów) umożliwia śledzenie ruchu drogowego i pieszego. Te dynamiczne obserwacje ruchu można traktować jako uzupełnienie udogodnień lokalizacyjnych, które charakteryzują tylko statyczne aspekty mieszkań. Lepsze zrozumienie związku między wszystkimi wymiarami charakterystyk mieszkania a jego ceną może dostarczyć cennych informacji dla pojedynczych decydentów oraz dla kształtowania polityki urzędów dzielnic i stymulowania równowagi społecznej i gospodarczej w obszarach miejskich.

W związku z rozwojem inżynierii danych i wykorzystaniem technik automatycznego uczenia maszynowego, rozpoczęła się nowa debata na temat tego, czy te modele, które nie są oparte na socjoekonomicznych modelach behawioralnych i solidnym podłożu teoretycznym, mogą zapewnić lepsze prognozy niż modele ekonometrii przestrzennej. Aplikacje oparte na sieciach neuronowych do

¹ Private Real Estate Common Position on a European Pillar of Social rights,
<https://ec.europa.eu/social/BlobServlet?docId=17437> (dostęp 12.08.2023).

² <https://ec.europa.eu/social/main.jsp?catId=1567> (dostęp 12.08.2023).

³ Goodchild M.F, Citizens as sensors: the world of volunteered geography, GeoJournal Wyd. 69, 2007, s. 211–221.

⁴ Gao S., Li L., Li W., Janowicz K., Zhang Y., *Constructing gazetteers from volunteered Big Geo-Data based on Hadoop*, Computers, Environment and Urban Systems Wyd. 61, 2017, s. 172-186.

⁵ Gebru T., Krause J., Wang Y., Fei-Fei L., *Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States*, PNAS Wyd. 114 Nr. 50, 2017, s. 13108-13113.

szacowania cen mieszkań w czasie rzeczywistym są dostępne od lat; jednak wielu naukowcy odrzucili te metody, ponieważ uznali je za „czarne skrzynki”, których wyników nie da się zinterpretować.⁶⁷ Niemniej jednak ostatnie postępy w interpretowalności modeli uczenia maszynowego⁸ czynią je szczególnie atrakcyjnymi narzędziami, ponieważ ich przewidywania zapewniają większą dokładność i bardziej dogłębne wyjaśnienia zjawisk. Uczenie maszynowe dąży w stronę stworzenia hybrydowych rozwiązań, umożliwiających wysoką precyzyję prognoz wraz ze zrozumieniem zachodzących wyrafinowanych relacji.⁹¹⁰ Na obecnym etapie rozwoju, precyza osiągnięta za pomocą modeli uczenia maszynowego stopniowo skłania branżę i instytucje finansowe do przyjęcia tych technik w codziennych wycenach fiskalnych.

Przypadek Warszawy ma ogromne znaczenie w dziedzinie prognozowania rynku mieszkaniowego ze względu na żywą dynamikę miejską i unikalny kontekst geopolityczny. Jako miasto stołeczne Rzeczypospolitej Polskiej, Warszawa pulsuje nowoczesną zabudową miejską, szybką urbanizacją, wzrostem liczby ludności i wzorcami różnorodnych migracji wewnętrznych. Status miasta jako ważnego ośrodka gospodarczego i kulturalnego nieustannie przyciąga mieszkańców z całej Polski, poszukujących możliwości zatrudnienia i wysokiej jakości życia. Wraz z tym, geograficzna, historyczna, kulturowa oraz językowa bliskość z Ukrainą i Białorusią dodają istotny wymiar geopolityczny. Trwający konflikt w Ukraine wpływa na nastroje inwestorów, popyt na nieruchomości oraz ceny mieszkań w Warszawie.¹¹ Kryzys polityczny w Białorusi również oddziałuje na warszawski rynek mieszkaniowy ze względu na licznych uchodźców poszukujących nowego mieszkania. Wzajemne oddziaływanie tych lokalnych i globalnych czynników sprawia, że Warszawa jest fascynującym studium przypadku do prognozowania rynku mieszkaniowego. Zrozumienie i przewidywanie cen mieszkań w tak złożonym i dynamicznym środowisku miejskim wymaga zaawansowanych narzędzi analitycznych, takich jak przestrzenne modele ekonometryczne i algorytmy uczenia maszynowego, umożliwiające decydentom, inwestorom i badaczom podejmowanie świadomych decyzji w szybko zmieniającym się krajobrazie. Badanie rynku mieszkaniowego w Warszawie oferuje wyjątkową okazję do głębszego uświadomienia szerszych trendów na rynku mieszkaniowym w regionie Europy Wschodniej. Wyniki takiego badania mogą nie

⁶ Zhao Q, Hastie T., *Causal Interpretations Of Black-Box Models*, J Bus Econ Stat, 2019, s. 1080.

⁷ Kauko T., *The Importance of the Context and the Level of Analysis*, Druid Working Papers Wyd. 20, 2003, s. 134-136.

⁸ Lundberg S., Lee S., *A Unified Approach to Interpreting Model Predictions*, NIPS, 2017.

⁹ Cranmer M. i in., *Discovering Symbolic Models from Deep Learning with Inductive Bayes*, NeurIPS, Vancouver 2020.

¹⁰ Triebe O. i in, *NeuralProphet: Explainable Forecasting at Scale*, <https://arxiv.org/abs/2111.15397>, 2021 (dostęp 07.08.2023).

¹¹ Trojanek R., Gluszak M., *Short-run impact of the Ukrainian refugee crisis on the housing market in Poland*, Finance Research Letters Wyd. 50, 2022.

tylko wspierać decyzje dotyczące polityki mieszkaniowej w Warszawie, ale także stanowić cenny przykład dla innych miast borykających się z podobnymi wyzwaniami.

Celem niniejszej pracy jest porównanie podejścia ekonometrycznego z podejściem uczenia maszynowego do prognozowania cen mieszkań na przykładzie Warszawy i zdefiniowanie najlepszej istniejącej na moment obecny metody prognozowania. Praca ma następującą strukturę: w Rozdziale I przedstawiono zasady teoretyczne i dokonano przeglądu literatury; w Rozdziale II szczegółowo określono modele i kryterium oceny; dobór zmiennych, inżynieria cech i eksploracyjna analiza danych została opisana w Rozdziale III; dyskusja i dogłębiańska analiza wyników zawarte w Rozdziale IV, a na zakończenie podsumowano główne wnioski, wspomniano o ograniczeniach niniejszej pracy i przedstawiono kierunki dalszego rozwoju.

I. Opis problemu prognozowania cen mieszkań

Celem tego rozdziału niniejszej pracy jest zaprezentowanie jak ważne jest prognozowanie cen mieszkań we współczesnej ekonomii. Omówione również zjawiska specyficzne dla danych przestrzennych oraz dokonano przeglądu dostępnej literatury naukowej na temat używanych modeli i charakterystyk mieszkań. Przegląd literatury pozwoli zrozumieć aktualny stan badań oraz wytyczyć kierunki, na których będzie się skupiała niniejsza praca.

1.1. Prognozowanie cen mieszkań. Istotność i zastosowania we współczesnej ekonomii

Prognozowanie cen mieszkań to proces wykorzystujący dane historyczne oraz różnorodne techniki analityczne w celu przewidzenia przyszłej wartości określonego aktywa, jakim jest konkretne mieszkanie. Termin „cena mieszkania” zazwyczaj odnosi się do ceny podstawowej, wyłączając wszelkie dodatkowe koszty i opłaty. Typowe podejścia do prognozowania ceny mieszkania obejmują:

- Podejście ekonometryczne: w tej kategorii znajduje się analiza szeregów czasowych, która dotyczy badania zmieniających się trendów w czasie; analiza przestrzenna, która koncentruje się na zrozumieniu i modelowaniu przestrzennych relacji; połączenie obu analiz.
- Podejście eksperckie: obejmuje wykorzystanie wiedzy i spostrzeżeń ekspertów z rynku nieruchomości. Metody te często łączą informacje jakościowe i ilościowe.
- Zastosowanie uczenia maszynowego: techniki uczenia maszynowego znajdują szerokie wykorzystanie w przewidywaniu cen mieszkań ze względu na ich zdolność do analizy dużych zbiorów danych. Typowe algorytmy uczenia maszynowego do przewidywania cen obejmują lasy losowe oraz wzmacnianie gradientem.

Przewidywanie cen mieszkań ma ogromne znaczenie we współczesnej ekonomii ze względu na daleko idące implikacje w różnych sektorach gospodarki:

- Podejmowanie decyzji inwestycyjnych: inwestorzy i deweloperzy wykorzystują prognozy cen mieszkań do oceny potencjalnego zwrotu z inwestycji na rynku nieruchomości. Precyzyjne prognozy pozwalają im na efektywną alokację kapitału oraz zarządzanie ryzykiem w portfelach inwestycyjnych.
- Urbanistyka: przewidywanie cen mieszkań jest niezbędne dla urzędników do formułowania skutecznej polityki mieszkaniowej. Dokładne prognozy pomagają wspierać zrównoważony rozwój obszarów miejskich.

- Zachowania konsumentów i decyzje zakupowe: potencjalni nabywcy dostosowują swoje preferencje i terminy zakupu nieruchomości w oparciu o przewidywane zmiany cen, wpływając na dynamikę popytu.
- Prognozowanie ceny mieszkania jest niezbędne dla instytucji finansowych do oceny ryzyka związanego z kredytowaniem zakupu mieszkania. Trafne prognozy pomagają w zarządzaniu ryzykiem kredytowym.

W ten sposób, prognozowanie cen mieszkań odgrywa kluczową rolę na rynku nieruchomości oraz szerzej w gospodarce. Dostarcza doniosłe informacje różnym interesariuszom, umożliwiając podejmowanie świadomych decyzji, wydajną alokację zasobów, zarządzanie ryzykiem oraz formułowanie skutecznych polityk na dynamicznym i złożonym rynku mieszkaniowym.

1.2. Zjawiska specyficzne dla danych przestrzennych

Dane przestrzenne odnoszą się do wszelkich danych powiązanych ze współrzędnymi przestrzennymi (zazwyczaj geograficznymi). Te dane mogą reprezentować różne cechy występujące w określonych lokalizacjach, takie jak gęstość zaludnienia, temperatura lub liczba zachorowań. Specyfika danych przestrzennych wynika z tego, w jaki sposób przechwytyują informacje o środowiskach przestrzennych. Koncentrując się szczególnie na danych geograficznych, istnieje kilka kluczowych aspektów, które je wyróżniają:

- **Odniesienie geograficzne** (ang. *georeferencing*): oznacza to, że każda obserwacja w analizowanym zbiorze danych jest powiązana z określonym miejscem na powierzchni Ziemi za pomocą współrzędnych szerokości i długości geograficznej. Umożliwia to łatwą integrację i nakładanie wielu zbiorów danych przestrzennych z różnych źródeł.
- **Pierwsze prawo Toblera:** jest to tzw. Pierwsze prawo geografii, zdefiniowane przez Waldo Toblera w 1970 roku.¹² Dokładnie brzmi: „Wszystko jest powiązane ze wszystkim, ale pobliskie obiekty są ze sobą powiązane bardziej niż obiekty odległe.” Skoro dane przestrzenne są bardzo zależne od skali, w której są mierzone, poziom szczegółowości różni się w zależności od źródła danych i celu analizy. Prawo Toblera jest podstawowym założeniem stosowanym we wszystkich analizach przestrzennych i pomaga zrozumieć, jak odległość przestrzenna wpływa na zachodzące zależności.
- **Autokorelacja przestrzenna** (ang. *spatial autocorrelation*): jest to istotna koncepcja w analizie przestrzennej, która wynika z pierwszego prawa Toblera. Autokorelacja

¹² Tobler W., *A computer movie simulating urban growth in the Detroit region*, Economic Geography Wyd. 46, 1970, s. 234–240.

przestrzenna mierzy stopień podobieństwa w zbiorze danych przestrzennych. Oznacza, że przestrzennie bliskie obserwacje mają zwykle podobne wartości cech, podczas gdy oddalone od siebie wykazują mniejsze podobieństwo. Formalnie autokorelację przestrzenną można zdefiniować matematycznie za pomocą globalnego testu Morana I,¹³ jednej z najczęściej używanych miar do ilościowego określania autokorelacji przestrzennej:

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (1)$$

gdzie:

- N – liczba obserwacji w rozważanym zbiorze danych,
- x – zmienną będącą przedmiotem zainteresowania,
- \bar{x} – średnia wartość zmiennej x ,
- w_{ij} – element macierzy wag przestrzennych oznaczający wagę połączenia między obszarem i oraz j . Więcej na temat macierzy wag przestrzennych opowiedziano w części metodologicznej.

Statystyka I w teście Morana może przyjmować wartości w zakresie od -1 do $+1$, gdzie:

$I > 0$: Dodatnia autokorelacja przestrzenna wskazująca, że podobne wartości mają tendencję do skupiania się w przestrzeni.

$I = 0$: Sugeruje brak autokorelacji przestrzennej.

$I < 0$: Ujemna autokorelacja przestrzenna oznacza, że wartość zmiennej będącej przedmiotem zainteresowania dla obiektu i negatywnie wpływa na wartość zmiennej dla obiektu j oraz vise versa.

- **Heterogeniczność przestrzenna** (ang. *spatial heterogeneity*): oznacza, że związek między zmiennymi może się różnić w zależności od kontekstu geograficznego. Heterogeniczność przestrzenna jest powszechnym zjawiskiem w danych przestrzennych; w rzeczywistości relacje mogą różnić się lokalnie ze względu na takie czynniki, jak warunki społeczno-ekonomiczne, wzorce użytkowania gruntów, czynniki środowiskowe i inne czynniki zmieniające się przestrzennie. Zrozumienie i uwzględnienie heterogeniczności przestrzennej ma kluczowe znaczenie dla dokładnej analizy i modelowania danych przestrzennych. Pomaga to uchwycić lokalne różnice i zależności przestrzenne, które mogą zostać przeoczone przy użyciu tradycyjnych metod statystycznych.

¹³ Moran P. A., *Notes on Continuous Stochastic Phenomena*, Biometrika. Wyd. 37 Nr. 1, 1950, s. 17–23.

- **Różne typy danych:** dane przestrzenne mogą przybierać różne formy, takie jak dane punktowe (np. współrzędne GPS), dane liniowe (np. drogi, rzeki) i dane poligonalne (np. granice administracyjne dzielnicy).
- Dane przestrzenne umożliwiają zastosowanie **specjalistycznych metod analitycznych** uwzględniające kontekst przestrzenny. Techniki analizy przestrzennej pozwalają ujawnić wzorce, trendy i klastry przestrzenne, które mogą nie być widoczne w tradycyjnych aprzestrzennych zbiorach danych.

Dzięki tym specyficznym cechom dane przestrzenne zyskały liczne zastosowania w różnych dyscyplinach. Odgrywają kluczową rolę w planowaniu urbanistycznym, monitorowaniu środowiska, zarządzaniu zasobami naturalnymi, reagowaniu na katastrofy, planowaniu transportu, epidemiologii, analizie rynku nieruchomości i wielu innych dziedzinach, w których zależności i wzorce geograficzne są niezbędne do podejmowania decyzji. Wraz z postępem technologii i metod gromadzenia danych bogactwo i różnorodność danych przestrzennych stale rosną, co prowadzi do nowych spostrzeżeń i rozwiązań złożonych problemów przestrzennych.

1.3. Przegląd literatury

Ceny mieszkań są szeroko omawiane w literaturze naukowej. Ze względu na obszerną bazę wiedzy, przegląd literatury zostanie podzielony na dwie części: przegląd wykorzystywanych modeli oraz przegląd używanych zmiennych. W pierwszej części przeglądu zostaną omówione różnorodne modele używane do prognozowania cen mieszkań, takie jak modele ekonometryczne oraz modele uczenia maszynowego. Drugą część przeglądu będzie stanowiła zestawienie używanych zmiennych w analizach cen mieszkań.

1.3.1. Przegląd modeli

Historycznie, prace naukowe w dziedzinie predykcji cen mieszkań opierały się głównie na pracy Rosena,¹⁴ w której zakładano, że złożony produkt jest kompozycją jego cech, które wpływają na ostateczną cenę. Ten model jest najczęściej określany jako model hedoniczny, często wykorzystuje regresję liniową oszacowaną za pomocą klasycznej metody najmniejszych kwadratów (KMNK). Niemniej jednak wielu badaczy¹⁵¹⁶ zwracali uwagę na problemy statystyczne związane z tym modelem, ponieważ jest on co najmniej nieefektywny, a w najgorszym przypadku dostarcza

¹⁴ Rosen S., *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition*, Journal of Political Economy, Wyd. 82 Nr. 1, 1974, s. 34–55.

¹⁵ Anselin L., *Spatial Heterogeneity In Spatial Econometrics: Methods and Models*, Springer, Dordrecht 1988.

¹⁶ Cliff A. D., Ord K., *Spatial Autocorrelation: A Review of Existing and New Measures with Applications*, Economic Geography Wyd. 46, 1970, s. 269-292.

obciążonych oszacowań parametrów. Głównym powodem tego jest występowanie autokorelacji przestrzennej w danych przestrzennych. W związku z tym, w miarę rozwoju analiz przestrzennych, zaproponowano wiele innych modeli.

Modele statystyczne, uwzględniające aspekt przestrzenny, często używano do analizy cen mieszkań w różnych badaniach. Na przykład, Chrostek i Kopczewska porównywali skuteczność modeli SAR (model autokorelacji przestrzennej, ang. *Spatial Autoregressive Model*), SEM (przestrzenny model Durbina, ang. *Spatial Durbin Model*) i GWR (geograficznie ważona regresja, ang. *Geographically Weighted Regression*) w prognozowaniu cen na rynku nieruchomości we Wrocławiu.¹⁷ Podobnie, przy użyciu modelu GWR, Widłak, Waszczuk i Olszewski analizowali dynamikę cen mieszkań na wtórnym rynku mieszkaniowym w Warszawie.¹⁸ Modele statystyczne, uwzględniające aspekt przestrzenny, znalazły szerokie zastosowanie również w innych krajach Europy¹⁹²⁰²¹, Ameryki²², Azji²³ i Oceanii.²⁴

W odpowiedzi na problem wnioskowania częstościowego, wielu badaczy podjęło próby zastosowania podejścia Bayesowskiego w celu wzbogacenia tradycyjnych modeli przestrzennych, np. Larm oraz Ahelegbey.²⁵ Podejście Bayesowskie do analizy danych przestrzennych często wykazuje się lepszą dokładnością niż podejście częstościowe, szczególnie gdy uwzględnione zostaną wcześniejsze informacje. Ze względu na obfitość dostępnej wcześniejszej wiedzy, te modele powinny znaleźć szerokie zastosowanie i osiągnąć większą dokładność prognoz w porównaniu z klasycznymi statystycznymi modelami przestrzennymi. Niemniej jednak w swoim przeglądzie podejścia Bayesowskiego, Louzada, Nascimento i Egbon stwierdzili, że literatura dotycząca Bayesowskich

¹⁷ Chrostek K., Kopczewska K., *Spatial Prediction Models for Real Estate Market Analysis*, Ekonomia Wyd. 34, 2014.

¹⁸ Widłak M., Waszczuk, J., Olszewski, K., *Spatial and Hedonic Analysis of House Price Dynamics in Warsaw*, National Bank of Poland Working Paper Nr. 197, 2015.

¹⁹ Taruttis L., Weber C., *Estimating the impact of energy efficiency on housing prices in Germany: Does regional disparity matter?*, Energy Economics Wyd. 105, 2022.

²⁰ Votsis A., *Planning for green infrastructure: The spatial effects of parks, forests, and fields on Helsinki's apartment prices*, Ecological Economics Wyd. 132, 2017.

²¹ Efthymiou D., Antoniou C., *How do transport infrastructure and policies affect house prices and rents? Evidence from Athens, Greece*, Transportation Research Part A: Policy and Practice Wyd. 52, 2013.

²² Hong I., Yoo C., *Analyzing Spatial Variance of Airbnb Pricing Determinants Using Multiscale GWR Approach*, Sustainability Wyd. 12 Nr. 11, 2020.

²³ Zhao C., *Multiscale Effects of Hedonic Attributes on Airbnb Listing Prices Based on MGWR: A Case Study of Beijing, China*, Sustainability Wyd. 15 Nr. 2, 2023.

²⁴ Bottero M. i in., *Urban parks, value uplift and green gentrification: An application of the spatial hedonic model in the city of Brisbane*, Urban Forestry & Urban Greening Wyd. 74, 2022.

²⁵ Larm A., Ahelegbey, D. F, *Detecting Spatial and Temporal House Price Diffusion in the Netherlands: A Bayesian Network Approach*, Regional Science and Urban Economics Wyd. 65 Nr. 10, 2017 s. 1016

modeli przestrzennych nie zawiera wystarczających informacji na temat rozkładów a priori dla parametrów.²⁶

Jednak zarówno większość Bayesowskich, jak i klasycznych przestrzennych modeli statystycznych modeli nie liczy się z nieliniowymi relacjami zachodzącymi pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi. W odpowiedzi na to wyzwanie, od lat 2010-tych aktywnie rozwija się zastosowanie modeli uczenia maszynowego. Dzięki wielu istotnym osiągnięciom badawczym wycena nieruchomości w erze big data dokonała znacznego postępu. Wsparcie technologii umożliwiło skutecną dywersyfikację i ulepszenie zasobów danych do prac związanych z wyceną, a także zwiększenie efektywności pozyskiwania danych. Metody wyceny konsekwentnie kierują się w stronę bardziej zaawansowanych podejść. Chociaż konwencjonalne metody i tradycyjne dane wejściowe nadal dominują dziedzinę ewaluacji cen na rynku nieruchomości,²⁷ badania te stopniowo przyjmują bardziej zaawansowane techniki i innowacyjne źródła danych. Jak wspomniano w przeglądzie literatury autorstwa Hamizaha i in.,²⁸ zdolność dzisiejszej eksploracji danych do wydobywania odpowiedniej wiedzy sprawia, że przewidywanie cen domów i kluczowych atrybutów mieszkań jest bardzo efektywne. Zastosowanie modeli uczenia maszynowego w przypadku prognozowania cen mieszkań zwykle jest wykonane razem z porównaniem tych modeli z modelem hedonicznym, najczęściej z modelem regresji liniowej. W ten sposób artykuł naukowy autorstwa Gulikera, Folmerta i Sinderena porównuje różne podejścia do uczenia maszynowego do wyceny nieruchomości w holenderskich miastach.²⁹ Wyniki pokazały, że xgboost wydawał się najlepszą opcją w porównaniu do GWR i modelu liniowego, wyjaśniając 83% wariancji wartości ceny. Z kolei badacze Lorenz i in.³⁰ podkreślają znaczenie modeli uczenia maszynowego w przewidywaniu i ujawnianiu ukrytych relacji w danych w przeciwieństwie do modeli regresji liniowej i SAR. Artykuł Kalliola, Kapočiūtė-Dzikienė, Damaševičius poświęcony jest kalibracji modelu MLP (perceptron wielowarstwowy, ang. *multilayer perceptron*) w celu prognozowania cen nieruchomości

²⁶ Louzada F., Nascimento D., Egbon O. A., *Spatial Statistical Models: An Overview under the Bayesian Approach*, Axioms Wyd. 10 Nr. 4, 2021, s. 307.

²⁷ Geerts M., Broucke S., Weerdt J., *A Survey of Methods and Input Data Types for House Price Prediction*, Geo-Inf Wyd. 12 Nr. 5, 2023, s.200.

²⁸ Hamizah Z., Shuzlina A. R., Hasbiah U., *House Price Prediction using a Machine Learning Model: A Survey of Literature*, Modern Education and Computer Science Wyd. 6, 2020, s. 46-54.

²⁹ Guliker E., Folmer E., Sinderen M., *Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach*, International Journal of Geo-Information Wyd. 11, 2022, s. 125 – 149.

³⁰ Lorenz F., Willwersch J., Cajias M., Fuerst F., *Interpretable machine learning for real estate market analysis*, Real Estate Economics, 2022, str. 1–31.

w Helsinkach.³¹ Optymalizacja hiperparametrów tego modelu znacznie poprawiła jego wydajność w przeciwstawieniu do modelu bazowego (liniowego). Jednak w badaniu Przekopa, ilość wykorzystywanych danych nie pozwoliła sztucznym sieciom neuronowym konkurować z modelami przestrzennymi w odzwierciedlaniu przestrzennej struktury zależności.³² Inni autorzy podjęły próbę porównania jednocześnie wielu modeli. W pracy Rico-Juana i Taltavull de La Paza³³ przeprowadzono porównanie modeli xgboost, adaboost (wzmacnianie adaptacyjne, ang. *Adaptive Boosting*), catboost, lasów losowych, drzewa decyzyjnego, KNN (k najbliższych sąsiadów, ang. *K Nearest Neighbors*) oraz MLP do prognozowania cen mieszkań w mieście Alicante, Hiszpania, gdzie model lasów losowych osiągnął najlepsze wyniki. W badaniu autorstwa Ziyue i Lu³⁴ przewidywanie cen mieszkań w Pekinie zostało poddane analizie przy użyciu regresji hedonicznej oraz algorytmów uczenia maszynowego. Ich wyniki wskazują, że metody uczenia maszynowego przewyższają tradycyjne metody ekonometrii przestrzennej pod względem dokładności, a model xgboost jest najbardziej dokładny. W publikacji Truong i in.³⁵ zaraportowano, że modele xgboost i LGBM zdobywają przyzwoitą dokładność w prognozowaniu, ale pod względem czasowej złożoności wydajniejszy jest model LGBM. Ale w szczególności model, który łączy oba te modele, okazał się najlepszym wyborem, gdy najwyższym priorytetem jest osiągnięcie jak największej dokładności prognoz.

Jednak niewiele wysiłku włożono w modyfikację tych modeli lub dostosowanie ich wartości hiperparametrów do przewidywania cen w rzeczywistych warunkach, uwzględniając przez to specyfikę danych przestrzennych. Badacze Bin i in.³⁶ skoncentrowali się na łączaniu cech domów i cech obrazów map ulic przy użyciu sztucznej sieci neuronowej, podczas gdy sama estymacja cen została przeprowadzona przy użyciu zwykłego uczenia ze wzmacnieniem. Co więcej, Ho, Tang i Wong³⁷ wykorzystali typowe algorytmy uczenia maszynowego, takie jak maszyny wektorów

³¹ Kalliola J., Kapočiūtė-Dzikienė J., Damaševičius R., *Neural network hyperparameter optimization for prediction of real estate prices in Helsinki*, PeerJ Comput. Sc, 2021.

³² Przekop D., *Artificial Neural Networks vs Spatial Regression Approach in Property Valuation*, CEJEME Wyd. 14, 2022, s. 199-223.

³³ Rico-Juan J. R., Taltavull de La Paz P., *Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain*, Expert Systems with Applications Wyd. 171, 2021.

³⁴ Ziyue Y., Lu Z., *Spatial Prediction of Housing Prices in Beijing Using Machine Learning Algorithms*, HPCCT & BDAI '20, 2020, s. 64–71.

³⁵ Truong Q., Nguyen M., Dang H., Mei B., *Housing Price Prediction via Improved Machine Learning Techniques*, IJKI2019, 2019, s. 1-5.

³⁶ Bin J., Gardiner B., Liu Z. i in, *Attention-based multi-modal fusion for improved real estate appraisal: a case study in Los Angeles*, Multimed Tools Appl Wyd. 78, 2019, s. 31163–31184.

³⁷ Ho W., Tang B., Wong S. W., *Predicting property prices with machine learning algorithms*, Journal of Property Research Wyd. 38 Nr. 1, 2021, s. 48-57.

nośnych (ang. *Support Vector Machines*, SVM), las losowy i wzmacnianie gradientem do estymacji cen nieruchomości, ale nie przeprowadzili optymalizacji hiperparametrów tych modeli.

Najbardziej skomplikowane i najnowsze artykuły naukowe skupiały uwagę na przetestowaniu hybrydowych rozwiązań, łączących świat ekonometrii przestrzennej i uczenia maszynowego. W ten sposób geograficznie ważona sztuczna sieć neuronowa (ang. *Geographically Weighted Artificial Neural Network*, GWANN) opracowana przez Hagenauera i Helbicha³⁸ pozwala łączyć zarówno nieliniowość, jak i przestrzenną heterogeniczność cen mieszkań. Jednak w ich badaniu nie uwzględniono szczegółowych zmiennych dotyczących lokalizacji i sąsiedztwa, ani nie porównano tej metody z innymi modelami uczenia głębokiego. W odpowiedzi na to Shen i in.³⁹ połączyli sztuczną sieć neuronową z wyprzedzeniem i model GWR do przetestowania tego modelu na 4 chińskich rynkach nieruchomości w Wuhanie, Nanjingie, Pekinie oraz Xi'anie. W rezultacie zaproponowana przestrzenna sieć neuronowa przewyższyła zarówno model GWR, jak i zwykłą sztuczną sieć neuronową pod względem jakości prognoz.

W odpowiedzi na powyższe wyzwania, niniejsze badanie skupi się na głównych problematycznych kwestiach poprzez:

- Standaryzację i przejrzyste opisanie stosowanych metod,
- Szersze zastosowanie podejścia Bayesowskiego w modelach ekonometrycznych,
- Przedstawienie procesu optymalizacji hiperparametrów i inżynierii cech,
- Zaspokojenie potrzeby podejścia hybrydowego, łączącego uczenie maszynowe i ekonometrię przestrzenną.

1.3.2. Przegląd zmiennych

Jak wspomniano wcześniej, teoria stojąca za hedonicznymi modelami zakłada, że mieszkanie składa się z heterogenicznych charakterystyk, a końcowa cena jest równa sumie bezpośrednich cen za każdą z tych charakterystyk. Dość powszechną heterogeniczną charakterystyką, wykorzystywaną w modelowaniu cen mieszkań, jest dostęp do zieleni, takich jak parki, obszary przyrodnicze i krajobrazy. Badanie przeprowadzone przez Jima i Chena skupiło się na efektach sąsiedztwa parków rekreacyjnych w obszarze miejskim Hongkongu.⁴⁰ Wyniki tego badania sugerują, że osiedlowe parki

³⁸ Hagenauer J., Helbich M., *A geographically weighted artificial neural network*, International Journal of Geographical Information Science Wyd. 36 Nr. 2, 2022, s. 215–235.

³⁹ Shen H., Li L., Zhu H., Liu Y., Luo Z., *Exploring a Pricing Model for Urban Rental Houses from a Geographical Perspective*, Land Wyd. 11 Nr. 1, 2022.

⁴⁰ Jim C. Y., Chen W. Y., *External effects of neighbourhood parks and landscape elements on high-rise residential value*, Land Use Policy Wyd. 27, 2010, s. 662–670.

wynoszą pewną wartość dodatnią do cen sprzedaży mieszkań. Ponadto, model wskazuje, że nabywcy domów są skłonni zapłacić prawie 15% więcej, aby mieszkać w sąsiedztwie parku. Z kolei badanie przeprowadzone przez Bottero i in.⁴¹ wskazuje, że przekształcenie pola golfowego w park publiczny (Victoria Park) spowodowało wzrost cen nieruchomości średnio o 3% w przypadku nieruchomości położonych w promieniu 750 m od parku w Brisbane, Australia. Jednak analiza Votsisa⁴² sugeruje, że zieleń miejska koreluje z odległością od centrum miasta. Wielu autorów skupia się również na innych cechach przestrzennych, takich jak bliskość do kawiarni,⁴³ uniwersytetu,⁴⁴ sklepów spożywczych⁴⁵ czy morza.⁴⁶ Jednakże, uwzględnienie jednocześnie wielu zmiennych objaśniających zwykle nie staje się wyzwaniem w prognozowaniu cen mieszkań. W ten sposób w badaniu Rico-Juan oraz Taltavull de La Paz⁴⁷ stwierdzono, że jakość dróg, bliskość do obiektów zdrowotnych i sportowych oraz gęstość zaludnienia mają pozytywny związek z ceną nieruchomości, podczas gdy rok budowy, wiek dzielnicy oraz bliskość do obiektów religijnych wręcz przeciwnie obniżają cenę. W badaniu czynników wpływających na ceny mieszkań w Warszawie, Widlak, Waszcuk, Olszewski⁴⁸ wskazują, że bliskość do terenów zielonych, centrum miasta i stacji metra przyczyniają się do wzrostu cen mieszkań. Inni badacze skupiali swoją uwagę na analizie demograficznych czynników zewnętrznych w jednostkach geograficznych (administracyjnych), w których znajdują się mieszkania. Wyniki analiz przeprowadzonych przez Tomala sugerują, że stopa bezrobocia ma statystycznie istotny i negatywny wpływ na cenę mieszkania.⁴⁹ Z kolei Park i Yun,⁵⁰ analizując rynek mieszkaniowy w Korei Południowej, stwierdzili, że wraz ze średnim wzrostem wieku obywateli w sąsiedztwie mieszkania, cena mieszkania maleje. Należy również zaznaczyć, że każde z wymienionych wyżej badań uwzględnia co najmniej jeden z podstawowych aspektów mieszkania: piętro, na którym się znajduje, powierzchnia, liczba pokoi, rok budowy. Jednak niektórzy autorzy uwzględnili również wpływ dodatkowych atutów na ostateczną cenę mieszkania. Bondemark

⁴¹ Bottero M. i in., op. cit.

⁴² Votsis A., op. cit.

⁴³ Guliker E., Folmer E., Sinderen M., op. cit.

⁴⁴ Ho W., Tang B., Wong S. W., op. cit.

⁴⁵ Lorenz F., Willwersch J., Cajias M., Fuerst F., op. cit.

⁴⁶ Efthymiou D., Antoniou C., op. cit.

⁴⁷ Rico-Juan J. R., Taltavull de La Paz P., op. cit.

⁴⁸ Widlak M., Waszcuk, J., Olszewski, K., op. cit.

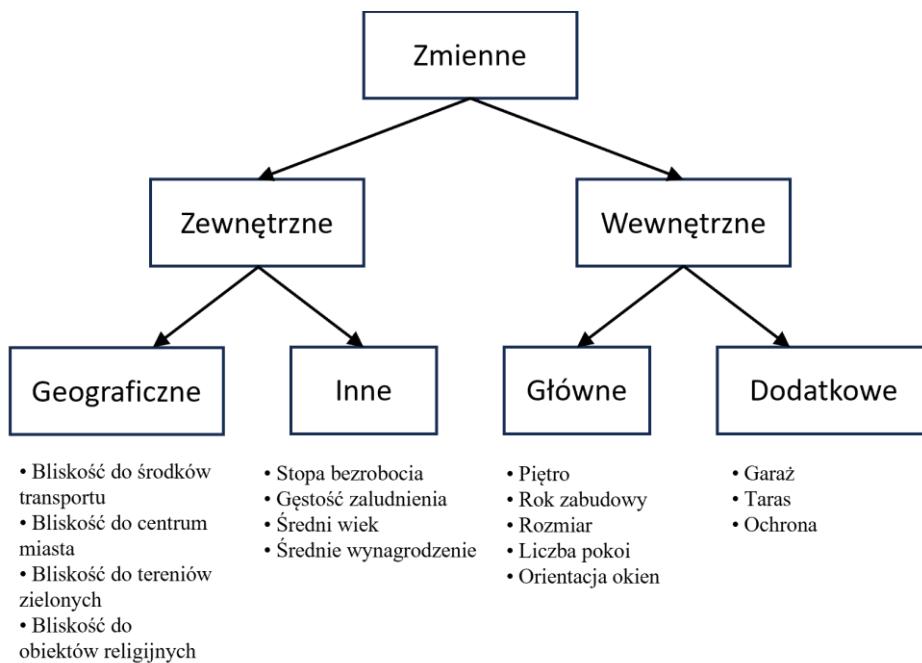
⁴⁹ Tomal M., *The Impact of Macro Factors on Apartment Prices in Polish Counties: A Two-Stage Quantile Spatial Regression Approach*, Real Estate Management and Valuation Wyd. 27 Nr. 4, 2019, s. 1-14.

⁵⁰ Park J., Yun S., *Social determinants of residential electricity consumption in Korea: Findings from a spatial panel model*, Energy Wyd. 239, 2022.

i Merkel⁵¹ doszły do kontrowersyjnego wniosku, że płatny parking miał pozytywny wpływ na ceny mieszkań w Sztokholmie, a Hong i Yoo⁵² sugerują, że obecność tarasu podwyższają cenę wynajmu. Nowoczesne zaawansowane modele uczenia głębokiego pozwalają również wyodrębnić cechy mieszkań, wynikających ze zdjęć rzeczywistych otoczenia.⁵³

Ze względu na dużą liczbę zmiennych zwykle używanych do predykcji cen mieszkań, wielu autorów skupia się na wymienieniu kategorii zmiennych, które wpływają na cenę mieszkania. Podsumowanie często używanych kategorii zmiennych wraz z przykładami zostały przedstawione na poniższym rysunku.

Rysunek 1. Taksonomia używanych zmiennych objaśniających wraz z przykładami



Źródło: opracowanie własne

W kolejnej części pracy zostanie przedstawiony aspekt metodologiczny, w którym zaprezentowano zestaw użytych modeli, dobranych w oparciu o przegląd literatury. Biorąc pod uwagę różnorodność czynników wpływających na ceny nieruchomości, zastosowanie zarówno modeli uczenia maszynowego, jak i ekonometrycznych, pozwoli na kompleksową analizę i lepsze zrozumienie zależności między zmiennymi, kształtującymi ceny mieszkań w Warszawie.

⁵¹ Bondemark A., Merkel A., *Parking not included: The effect of paid residential parking on housing prices and its relationship with public transport proximity*, Regional Science and Urban Economics Wyd. 99, 2023.

⁵² Hong I., Yoo C., op cit.

⁵³ Bin J., op cit.

II. Metodyka badania

W tym rozdziale dokonano opisu metodologii, tzn. modeli, miar dopasowania oraz algorytmu strojenia hiperparametrów, wykorzystanych podczas dalszej analizy. Uczenie maszynowe, które rozkwitło w ostatniej dekadzie, znalazło zastosowanie w różnych dziedzinach, w tym również do prognozowania cen mieszkań. Niemniej jednak, tradycyjne ekonometryczne podejście wciąż pozostaje mocnym narzędziem statystycznym, zwłaszcza w przypadku stosunkowo ograniczonych informacji na temat kształtowania cen mieszkań. Niniejsza część pracy pozwoli zrozumieć przebieg rozwoju metod używanych do analizy danych przestrzennych oraz wytlumaczyć matematyczny fundament użytych dalej modeli.

2.1. Ekonometryczne modele przestrzenne

Przestrzenne modele ekonometryczne stanowią istotne narzędzia w analizie danych przestrzennych. W przeciwieństwie do klasycznych (aprzestrzennych) modeli, uwzględniają one fakt, że dane w sąsiednich obszarach mogą być ze sobą powiązane i wzajemnie na siebie oddziaływać. Historia modeli przestrzennych sięga lat 70-ch XX wieku, kiedy zaczęła się rozwijać odpowiednia dziedzina ekonometrii.⁵⁴ Początkowo badacze skupiali się głównie na zagadnieniach ekonomicznych i społecznych. Prace te odegrały kluczową rolę w rozwoju teoretycznych podstaw przyszłych modeli. Inspiracją do powstania modeli przestrzennych były również osiągnięcia w dziedzinie ekonometrii szeregow czasowych. Od tamtego czasu modele przestrzenne zaczęły aktywnie się rozwijać, a ich zastosowanie rozszerzyło się na różne dziedziny, takie jak geologia, budownictwo, epidemiologia i ekologia. Tradycyjnie, dane przestrzenne dotyczyły informacji związanych z geograficznym umiejscowieniem obiektów. Jednak w dobie rosnącej cyfryzacji i dostępności danych z różnych dziedzin, pojęcie danych przestrzennych znacznie się rozszerzyło. Współcześnie, wiele badań koncentruje się na analizie powiązań między sąsiednimi jednostkami poza przestrzenią geograficzną. Na przykład, w analizach mediów społecznościowych dane przestrzenne mogą odnosić się do użytkowników sieci społecznych, takich jak Twitter.⁵⁵ Rozwój i udostępnianie technologii, wzrost mocy obliczeniowej oraz łatwiejszy dostęp do danych przestrzennych przyczyniły się do dalszego rozwoju ekonometrycznych modeli przestrzennych, umożliwiając bardziej zaawansowane analizy i precyzyjne wnioskowanie. Przez wiele lat modele statystyczne ignorowały aspekt przestrzenny, co prowadziło do niedokładnych i niekompletnych analiz. Wraz z udowodnieniem przewagi modeli

⁵⁴ Paelinck J., *Spatial econometrics*, Economics Letters Wyd. 1 Nr. 1, Amsterdam 1978, s. 59-63.

⁵⁵ François B., Gallic E., *Recovering the French Party Space from Twitter Data*, Science Po Quant, Paris 2015.

przestrzennych nad modelami klasycznymi, naukowcy zaczęli opracowywać nowe rozwiązania. Dzięki temu modele te mogą dostarczać bardziej trafnych informacji na temat wzorców przestrzennych oraz prognozowania zmiennych.

2.1.1. Autoregresyjne modele przestrzenne

Modele autoregresyjne stanęły u początków ekonometrii przestrzennej. Główne bodźce ich rozwoju pochodziły zarówno z osiągnięć w dziedzinie statystycznych modeli przestrzennych, jak i z ekonometrii szeregów czasowych. Podstawą dla tych modeli były metody autoregresji w czasie, w których wartości zmiennej zależnej są wyjaśniane przez wcześniejsze wartości tej zmiennej. Analogicznie, w autoregresyjnych modelach przestrzennych wprowadza się zależność między sąsiednimi obszarami, co pozwala na uwzględnienie współzależności przestrzennej. Modele autoregresyjne są ważnym narzędziem w analizie danych przestrzennych, pozwalając na bardziej realistyczne modelowanie zjawisk, które wykazują wzorce geograficzne.

2.1.1.1. Model czystej autoregresji przestrzennej

Model czystej autoregresji przestrzennej (ang. *pure Spatial Autoregressive Model*, pure SAR) jest jednym z pierwszych znanych modeli statystycznych, który uwzględnia zależności przestrzenne występujące pomiędzy jednostkami. Został opracowany niezależnie przez kilku badaczy w dziedzinie ekonometrii przestrzennej. Model pure SAR opiera się na założeniu, że wartość zmiennej objaśnianej w danym obszarze jest zależna od wartości tej zmiennej w innych obszarach. Formalnie, równanie modelu czystej autoregresji przestrzennej jest przedstawione jako:

$$y_i = \rho \cdot \sum_{j=1}^J y_j \cdot W_{ij} + \varepsilon_i \quad (2)$$

gdzie:

- y_j – wartość zmiennej objaśnianej w sąsiednim obszarze j ,
- ρ – parametr autokorelacji przestrzennej, który określa siłę zależności przestrzennej.
- W_{ij} – waga przypisana do połączenia między obszarami i oraz j , która określa, jak bardzo obserwacja z sąsiedniego obszaru j wpływa na wartość zmiennej objaśnianej w obszarze i ,
- ε_i – składnik losowy.

W formie skróconej, zapis modelu pure SAR wygląda następująco:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \boldsymbol{\varepsilon} \quad (3)$$

Jest to podstawowa forma modelu pure SAR, gdzie \mathbf{W} odpowiada za macierz wag przestrzennych.

Macierz wag przestrzennych (ang. *spatial weights matrix*) jest kluczowym elementem w przestrzennym modelowaniu statystycznym. Jej rola polega na przypisaniu odpowiednich wag

połączeniom między obszarami. Macierz ta jest reprezentowana jako macierz kwadratowa o wymiarach $N \times N$, gdzie N oznacza liczbę obserwacji w analizowanym zbiorze danych. Wagi przestrzenne są definiowane dla każdej pary obszarów i są używane do określenia wpływu jednego obszaru na drugi. Istnieje wiele różnych sposobów definiowania wag w zależności od kontekstu badania. Jednym z powszechnie stosowanych podejść jest wykorzystanie wag binarnych, gdzie wartość wagi W_{ij} wynosi 1, jeśli obszar i sąsiaduje z obszarem j , a 0 w przeciwnym przypadku. W ten sposób, jeśli dwa obszary sąsiadują ze sobą, mają one pozytywną wagę, co wskazuje na istnienie między nimi zależności przestrzennej. Istnieją również inne metody definiowania wag, takie jak wagi odwrotnie proporcjonalne do odległości między obszarami i in. Niezależnie od wybranej metody definiowania wag, macierz \mathbf{W} ma zerową diagonalę, co oznacza, że obszar nie wpływa na siebie bezpośrednio. Ważne jest, aby macierz wag przestrzennych była starannie zdefiniowana, biorąc pod uwagę istotność sąsiedztwa i potencjalne interakcje między obszarami. Wybór odpowiedniej macierzy wag przestrzennych jest kluczowy dla uzyskania trafnych wniosków. Powinno się to uwzględnić w czasie definiowania celów badawczych, aby zagwarantować solidność i wiarygodność wyników.

2.1.1.2. Model autoregresji przestrzennej z dodatkowymi regresorami

Od lat 70-ch pojawiło się kilka rozszerzeń modelu czystej autoregresji przestrzennej. Jednym z takich rozszerzeń jest model autoregresji przestrzennej z dodatkowymi regresorami (ang. *Spatial Autoregressive Model with Exogenous Variables*, SAR). Model ten uwzględnia zarówno zależności przestrzenne, jak i wpływ zmiennych niezależnych na zmienną objaśnianą. Formalnie, równanie modelu SAR można zapisać jako:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4)$$

gdzie $\boldsymbol{\beta}$ jest wektorem współczynników dla zmiennych niezależnych \mathbf{X} . Jest to ważne rozwinięcie czystej autoregresji przestrzennej, które umożliwia uwzględnienie innych czynników, które mogą wywierać wpływ na zmienną objaśnianą w przestrzeni geograficznej.

2.1.1.3. Model błędu przestrzennego.

Model błędu przestrzennego (ang. *Spatial Error Model*, SEM) został opisany przez Luca Anselina w 1988 roku.⁵⁶ Model ten jest kolejnym rozszerzeniem modelu autoregresji przestrzennej, umożliwiającym analizę zależności przestrzennej przy uwzględnieniu heteroskedastyczności danych. Równanie modelu SEM wygląda następująco:

⁵⁶ Anselin, L., op cit.

$$\begin{aligned} \mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &= \lambda \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{u} \end{aligned} \quad (5)$$

Różnice w porównaniu do modelu czystej autoregresji przestrzennej stanowi drugie równanie. Reprezentuje ono błąd przestrzenny $\boldsymbol{\varepsilon}$, który jest zależny od wartości błędu w sąsiednich obszarach, oznaczanych jako $\mathbf{W}\boldsymbol{\varepsilon}$. Parametr λ opisuje siłę przekazywania błędów między sąsiednimi obszarami, co pozwala uwzględnić wpływ przestrzennej heteroskedastyczności na dane. Z kolei \mathbf{u} odpowiada za losową, niezależną od sąsiednich obszarów część $\boldsymbol{\varepsilon}$. W ten sposób model SEM pozwala analizę zarówno zależności przestrzennych, jak i wpływu niejednorodności składnika losowego na wyniki oszacowania modelu.

2.1.1.4. Model opóźnienia przestrzennego regressorów.

Model opóźnienia przestrzennego regressorów (ang. *Spatial Lag Model*, SLX) opracowany przez Cliffa i Orda w 1973 roku⁵⁷ liczy się z zarówno bezpośrednim, jak i przestrzennie opóźnionym oddziaływaniem zmiennych niezależnych na zmienną objaśnianą. W zapisie matematycznym, równanie modelu SLX można przedstawić jako:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{WX}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (6)$$

gdzie \mathbf{WX} – przestrzenne opóźnienie zmiennych objaśniających, a $\boldsymbol{\theta}$ – wektor współczynników. W modelu SLX bieży się po uwagę wpływ zmiennych niezależnych X oraz wpływ przestrzennej autokorelacji, reprezentowanej przez opóźnienie przestrzenne $\mathbf{WX}\boldsymbol{\theta}$. Dzięki temu modelowi możliwe jest przeegzaminowanie zarówno lokalnych efektów zmiennych niezależnych, jak i efektów sąsiedztwa między obszarami.

2.1.1.5. Rozszerzenia podstawowych autoregresyjnych modeli przestrzennych.

W kontekście analizy danych przestrzennych istnieją różne kombinacje modeli SAR, SEM i SLX. W niniejszej pracy rozważono takie kombinacje, jak:

- **Model autoregresyjny drugiego rzędu** (ang. *Spatial Autoregressive*, SARAR) – łączy dwa autoregresyjne składniki przestrzenne. W modelu SARAR rozpatrywane są zarówno autoregresja przestrzenna zmiennej objaśnianej, jak i autoregresja przestrzenna błędu losowego.
- **Przestrzenny model Durbina** (ang. *Spatial Durbin Model*, SDM) – połączenie modelu SAR z dodatkowymi regresorami z modelem SLX. Model Durbina uwzględnia autoregresję przestrzenną, wpływ zmiennych niezależnych oraz oddziaływanie opóźnienia przestrzennego zmiennych niezależnych.

⁵⁷ Cliff A. D., Ord J. K., *Spatial autocorrelation*, Pion Limited, London 1973.

- **Przestrzenny model Durbina z błędem przestrzennym** (*Spatial Durbin Error Model*, SDEM) – połączenie modelu SEM z modelem SLX. Model SDEM zawiera opóźnienia składnika losowego oraz wpływ zmiennych niezależnych (bezpośrednio oraz w postaci opóźnienia przestrzennego).
- **Uogólniony model przestrzenny** (ang. *Generalized Spatial Model*, GNS) – jest kombinacją trzech możliwych komponentów zależności przestrzennej.

Poniżej przedstawiona taksonomia wszystkich opisanych modeli wraz z formułami.

Tabela 1. Taksonomia statystycznych modeli przestrzennych

Nazwa modelu	Formuła
Pure SAR	$\rho \mathbf{W} \mathbf{y} + \boldsymbol{\varepsilon}$
SAR	$\rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$
SEM	$\rho \mathbf{W} \mathbf{y} + \boldsymbol{\varepsilon}; \boldsymbol{\varepsilon} = \lambda \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{u}$
SLX	$\mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\varepsilon}$
SARAR	$\rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}; \boldsymbol{\varepsilon} = \lambda \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{u}$
SDM	$\rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\varepsilon}$
SDEM	$\mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\varepsilon}; \boldsymbol{\varepsilon} = \lambda \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{u}$
GNS	$\rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\varepsilon}; \boldsymbol{\varepsilon} = \lambda \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{u}$

Źródło: opracowanie własne

2.1.2. Bayesowskie przestrzenne modele autoregresyjne

Odrębną kategorię ekonometrycznych modeli przestrzennych stanowią Bayesowskie przestrzenne modele autoregresyjne (ang. *Bayesian Spatial Autoregressive Models*). Są to statystyczne modele, które łączą autoregresję przestrzenną z podejściem Bayesowskim. Równanie podstawowego Bayesowskiego przestrzennego modelu autoregresyjnego opartego na modelu SAR można zapisać jako:

$$\begin{aligned}
 \mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
 \boldsymbol{\varepsilon} &\sim N(0, \sigma^2 \mathbf{I}_n) \\
 \pi(\boldsymbol{\beta}) &\sim N(c, T) \\
 \pi\left(\frac{1}{\sigma^2}\right) &\sim \Gamma(d, v) \\
 \pi(\rho) &\sim U[0, 1]^{58}
 \end{aligned} \tag{7}$$

gdzie:

⁵⁸ Lesage J. P., *Bayesian Estimation of Spatial Autoregressive Models*, International Regional Science Review, Wyd. 20 Nr. 1–2, 1997, s. 113–129.

- π oznacza rozkłady a priori dla parametrów β, σ, ρ .
- N – funkcja gęstości rozkładu Gaussowskiego (normalnego),
- Γ – funkcja gęstości rozkładu Gammy,
- U – funkcja gęstości ciągłego rozkładu jednostajnego.

W przypadku dużych prób obejmujących ponad 3000 obserwacji, rozkłady a priori dla parametrów β, σ powinny wywierać stosunkowo niewielki wpływ na wyniki oszacowań.⁵⁹ Ustawienie $c = 0$ i T na bardzo dużą liczbę skutkuje nieinformacyjnym rozkładem a priori dla parametrów, co faktycznie oznacza zredukowanie wpływu wiedzy a priori dla współczynników stojących przy zmiennych wyjaśniających. Nieinformacyjne ustawienia dla rozkładu a priori dla σ obejmują ustawienie $d = 0, \nu = 0$. W przeciwnieństwie do parametrów β oraz σ , rozkład a priori współczynnika autoregresji przestrzennej ρ ma istotny wpływ na oszacowanie wyników nawet w dużych próbach. Wynika to z kluczowej roli, jaką w tych modelach odgrywa autokorelacja przestrzenna. Dlatego w typowych zastosowaniach, w których znalezienie wielkości ρ jest przedmiotem zainteresowania, używano rozkładu a priori dla parametru ρ z wielką ostrożnością.

Należy podkreślić, że powyższy model jest szczególnym przypadkiem dla autoregresyjnego modelu przestrzennego SAR. Jednak w podobny sposób podejście Bayesowskie możliwe jest do zaaplikowania do wszystkich powyżej opisanych autoregresyjnych modeli przestrzennych. W przypadku modeli z autoregresyjnym błędem przestrzennym, rozkład a priori będzie dotyczył tylko części niezależnej (komponent \mathbf{u}). Istnieje możliwość rozważania innych funkcji gęstości dla parametrów β, σ, ρ . W ten sposób na przykład, często używa się rozkład $B(a, b)$ dla parametru autoregresji przestrzennej, gdy zakładano występowanie przestrzennej kanibalizacji (ujemnej wartości współczynnika autoregresji). W ten sposób rozkład beta w naturalny sposób pozwoli ograniczyć obszar możliwych wartości parametru ρ do przedziału $[-1; 1]$. W przypadku rozkładu a priori parametru β , jest głównie zdefiniowany w zależności od problemu badawczego. Jednak na potrzeby niniejszego badania kwestie innych funkcji gęstości nie będą uwzględnione.

Aby wdrożyć Bayesowską metodę szacowania, konieczne jest określenie rozkładów warunkowych a posteriori dla parametrów β, σ, ρ . W przypadku modelu SAR rozkłady te wyglądają następująco:

- Dla parametru β :

$$\circ \quad P(\beta | \rho, \sigma) \sim N(\bar{b}, \sigma^2 \mathbf{B})$$

⁵⁹ Ibidem.

- $\bar{b} = \mathbf{A}(\mathbf{X}'\mathbf{S}\mathbf{y} + \sigma^2 T^{-1}c)$
- $\mathbf{B} = \sigma^2 \mathbf{A}$
- $\mathbf{A} = (\mathbf{X}'\mathbf{X} + \sigma^2 T^{-1})^{-1}$
- $\mathbf{S} = (\mathbf{I}_n - \rho \mathbf{W})$
- Dla parametru σ :
 - $P(\sigma^2 | \boldsymbol{\beta}, \rho) \propto (\sigma^2)^{-\left(\frac{n}{2}+d+1\right)} \cdot \exp\left(-e'e + \frac{2v}{2\sigma^2}\right)$
 - $e = (\mathbf{I}_n - \rho \mathbf{W})\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- Dla parametru ρ :
 - $P(\rho | \boldsymbol{\beta}, \sigma) \propto |\mathbf{S}| \cdot (s^2(\rho))^{\frac{n-k}{2}} \cdot \pi(\rho)$
 - $s^2(\rho) = \frac{(\mathbf{S}\mathbf{y} - \mathbf{X}b(\rho))'(\mathbf{S}\mathbf{y} - \mathbf{X}b(\rho))}{n-k}$
 - $\mathbf{S} = (\mathbf{I}_n - \rho \mathbf{W})$

Jednak w sytuacji z rozkładem a posteriori dla parametru ρ pojawia się problem polegający na tym, że nie istnieją dla niego ustalone algorytmy do generowania losowań. W odpowiedzi na to, są dwa sposoby próbkowania z wymienionego rozkładu warunkowego: za pomocą algorytmu Metropolisa-Hastingsa albo próbnika Gibbsa.⁶⁰ Zwykle oprogramowanie Bayesowskich przestrzennych modeli autoregresyjnych polega na użyciu algorytmu Metropolisa-Hastingsa.

2.1.3. Geograficznie ważona regresja

Geograficznie ważona regresja (ang. *Geographically Weighted Regression*, GWR) jest rozszerzeniem zwykłej regresji lokalnej i najczęstszym stosowanym ekonometrycznym modelem przestrzennym wśród zarówno amatorów, jak i ekspertów modelowania przestrzennego. W odróżnieniu od autoregresyjnych modeli przestrzennych, pozwala na oszacowanie współczynników regresji na poziomie lokalnym. Model został stworzony przez Brunsdona, Fotheringhama i Charltona w 1998 roku.⁶¹ Nieparametryczny model GWR opiera się na szeregu lokalnych regresji linowych, aby wyprodukować wyniki dla każdego punktu w przestrzeni. W tym celu wykorzystuje pod próbę danych z sąsiadujących obserwacji. Wzór na model wygląda następująco:

$$\mathbf{y} = \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (8)$$

⁶⁰ Ibidem.

⁶¹ Fotheringham A., Charlton M., Brunsdon C., *Geographically Weighted Regression: A Natural Evolution Of The Expansion Method for Spatial Data Analysis*, Environment and Planning Wyd. 30 Nr. 11, 1998, s. 1905-1927.

W przeciwieństwie do autoregresyjnych modeli przestrzennych, gdzie zwykle wykorzystują się zero-jedynkowe wagi, każdy wiersz macierzy wag W w przypadku modelu GWR jest konstruowany na podstawie funkcji ważenia. Jedna z najpopularniejszych funkcji ważenia wygląda odpowiednio:

$$W_i^2 = \exp\left(-\frac{d_i}{\theta}\right)$$

$$d_i = \sqrt{(Zx_i - Zx_j)^2 + (Zy_i + Zy_j)^2} \quad (9)$$

gdzie:

- θ jest parametrem zaniku, inaczej szerokością pasma (ang. decay parameter, bandwidth),
- d_i – wektor odległości (Euklidesowej) między obserwacją i a wszystkimi innymi $j = 1, \dots, N$ obserwacjami w próbie,
- Zx_j, Zy_j oznaczają współrzędne szerokości i długości (geograficznej lub innej w zależności od kontekstu zagadnienia).

Innymi słowy, dla każdej obserwacji i model GWR szacuje współczynniki na podstawie obserwacji i oraz ważonych $j = 1, \dots, N$ obserwacji, przy czym największe wagi otrzymują najbliższe (względem odległości) obserwacje. Wpływ odległości jest korygowany przez sterowanie szerokością pasma. Zmiana szerokości pasma w powyższym równaniu skutkuje innym profilem zaniku wykładniczego, co z kolei daje bardziej lub mniej lokalnie wygładzone oceny parametrów.

Chociaż GWR wyraźnie modeluje zmienność przestrzenną, wiąże się z pewnymi problemami. Jednym z problemów jest to, że struktura modelu nie pozwala na wyciąganie prawidłowych wniosków dla parametrów regresji. Aby to zobaczyć, należy wziąć pod uwagę, że lokalne liniowe oszacowania wykorzystują te same obserwacje (z różnymi wagami), aby utworzyć sekwencję oszacowań dla wszystkich punktów w przestrzeni.⁶² Z uwagi na brak niezależności między szacunkami dla każdej lokalizacji, konwencjonalne miary rozproszenia dla oszacowań parametrów będą nieprawidłowe. Innym problemem jest to, że obserwacje odstające mogą wywierać nadmierny wpływ na lokalne liniowe oszacowania. Wszystkie pobliskie obserwacje w pod próbie mogą być zanieczyszczone wartością odstającą w pojedynczym punkcie.

Na tym kończy się metodologiczny aspekt niniejszej pracy poświęcony ekonometrycznym modelom przestrzennym. Kolejna część modeli dotyczy modeli aprzestrzennych, powstałych z dziedziny uczenia maszynowego.

⁶² Ibidem.

2.2. Modele uczenia maszynowego

Modele uczenia maszynowego to zbiór matematycznych metod i algorytmów, które pozwalają na ustalenie wzorców oraz zależności na podstawie zbioru danych. Obejmują one różne dziedziny, takie jak uczenie nadzorowane (ang. *Supervised Learning*), nienadzorowane (*Unsupervised*), półnadzorowane (*Semi-Supervised*), oraz uczenie ze wzmocnieniem (*Reinforcement Learning*). Głównym celem modeli uczenia maszynowego jest optymalizacja wyników dla konkretnego zadania. Modele uczenia maszynowego często są aplikowane na dużych zbiorach danych, wykorzystując efektywne technologie obliczeniowe. Są używane w różnych dziedzinach nauki, takich jak biologia, medycyna czy ekonomia. Wprowadzenie modeli uczenia maszynowego w analizie danych pozwoliło na znaczny postęp w opracowywaniu zaawansowanych systemów rekomendacji, rozpoznawania wzorców czy predykcji przyszłych zdarzeń.

W przeciwieństwie do modeli ekonometrycznych, które często zakładają określone rozkłady danych i tworzenie modeli zgodnie z założeniami, modele uczenia maszynowego stawiają na identyfikację ukrytych wzorców i reprezentacji danych bez dodatkowych ograniczeń teoretycznych. Skupiają się na zdolnościach predykcyjnych, niekoniecznie angażując interpretowalność jako główny cel. Przy odpowiedniej kalibracji – inżynierii cech (ang. *feature engineering*) oraz strojeniu hiperparametrów (*hyperparameter tuning*) – modele uczenia maszynowego mogą uzyskać wyższą dokładność prognoz niż modele statystyczne. Skoro celem niniejszej pracy jest porównanie modeli uczenia maszynowego z modelami ekonometrycznymi w kontekście danych przestrzennych, dobrano możliwie duży zbiór danych, dokonano gruntownej inżynierii zmiennych oraz przeprowadzono dokładnej kalibracji modeli uczenia maszynowego.

2.2.1. Drzewo decyzyjne. Modele uczenia zespołowego

Modele uczenia zespołowego to popularna klasa algorytmów uczenia maszynowego. W ich strukturze najczęściej wykorzystuje się drzewo decyzyjne (czasami zwane również drzewem losowym, regresyjnym lub binarnym), gdzie węzły reprezentują testy na zmiennych objaśniających zawartych w zbiorze danych, a krawędzie prowadzą do kolejnych węzłów lub liści, w których podejmowane są ostateczne decyzje. Podczas uczenia drzewo jest konstruowane w taki sposób, aby minimalizować błąd predykcji. Ważną zaletą drzew decyzyjnych jest ich interpretowalność: każdy węzeł drzewa reprezentuje konkretne warunki na podstawie zmiennych objaśniających, co ułatwia zrozumienie procesu podejmowania akcji. Istotnym aspektem drzewa decyzyjnego jest także zdolność do pracy z danymi zarówno numerycznymi, jak i kategorycznymi, bez konieczności wprowadzania specjalnych przekształceń. Jednak pojedyncze drzewo decyzyjne jest wrażliwe na małe zmiany

w danych, co może prowadzić do zmiany całej struktury drzewa i destabilizować predykcję. Co więcej, skłonność do przeuczenia prowadzi do zbyt skomplikowanych struktur, co pogarsza zdolność do generalizacji na nowych danych. Ograniczenie głębokości drzewa zawsze powinno być stosowane w celu ograniczenia przeuczenia. Mimo tego, model drzewa decyzyjnego jest szeroko stosowany i cieszy się popularnością ze względu na swoją prostotę, interpretowalność i zdolność do rozwiązywania różnorodnych problemów. Aby uniknąć wad pojedynczego drzewa, istnieje kilka rozwiązań opartych na uśrednieniu wyniku estymacji z mnóstwa pojedynczych „słabych” drzew. Pomaga to osiągnąć wspólnego „mocnego” modelu (stąd pochodzi nazwa uczenia zespołowego, ang. *ensemble learning*).

2.2.1.1. Lasy losowe

Lasy losowe (ang. *Random Forests*) łączą wiele (B) pojedynczych drzew decyzyjnych, poprawiając w ten sposób ogólną precyzję. Pierwszy algorytm losowych lasów decyzyjnych został stworzony w 1998 roku przez Tin Kam Ho przy użyciu metody losowej podprzestrzeni.⁶³ Klasyczny algorytm uczący dla lasów losowych stosuje technikę agregacji znaną pod nazwą bagging (ang. *bootstrap aggregating*) oraz technikę feature bagging. Bagging polega na wygenerowaniu ze zwracaniem B losowych prób składających się z N obserwacji pierwotnego zbioru danych ($\mathbf{X}_b, \mathbf{y}_b$). Feature bagging z kolei polega na losowaniu k (zwykle \sqrt{p}) zmiennych ze zbioru p zmiennych ($\mathbf{X}_{bk}, \mathbf{y}_b$). Powodem tego jest korelacja drzew w zwykłej metodzie bagging: jeśli jedna lub kilka zmiennych objaśniających jest bardzo silnymi predykatorem zmiennej objaśnianej, zmienne te zostaną wybrane w wielu drzewach, powodując ich skorelowanie. Następnym etapem jest dopasowanie drzewa decyzyjnego do wygenerowanych prób $F_b(\mathbf{X}_{bk}, \mathbf{y}_b)$. Po dopasowaniu drzew do każdej z B prób, ostateczne prognozy są osiągane za pomocą uśredniania prognoz z pojedynczych drzew.

$$F(\mathbf{X}) = \frac{1}{B} \sum_{b=1}^B F_b(\mathbf{X}_{bk}, \mathbf{y}_b) \quad (10)$$

Powyższa procedura prowadzi do lepszego dopasowania modelu, ponieważ zmniejsza wariancję bez zwiększenia obciążenia. Podczas gdy prognozy pojedynczego drzewa są bardzo wrażliwe na małe zmiany wartości zmiennych, to udowodniono, że średnia wielu drzew nie jest, o ile drzewa nie są skorelowane.⁶⁴ Uczenie wielu drzew na jednym zbiorze treningowym dałoby silnie skorelowane drzewa (lub nawet to samo drzewo B razy).

⁶³ Ho T. K., *The random subspace method for constructing decision forests*, IEEE Transactions on Pattern Analysis and Machine Intelligence Wyd. 20 Nr. 8, 1998, s. 832-844.

⁶⁴ Breiman L., *Random Forests*, Machine Learning Wyd. 45, 2001, s. 5–32.

Liczba próbek/drzew B jest ważniejszym parametrem modelu. Zwykle wykorzystuje się od kilkuset do kilku tysięcy drzew, w zależności od wielkości i charakteru zbioru uczącego. Optymalną liczbę drzew B można znaleźć za pomocą walidacji krzyżowej i strojenia hiperparametrów modelu, np. za pomocą algorytmów ewolucyjnych. Oprócz parametru B i k , kontrolowaniu ulegają parametry odpowiadające za głębokość drzewa (maksymalna głębokość drzewa, minimalna liczba próbek wymagana do podziału węzła wewnętrznego, minimalna liczba próbek, które muszą znajdować się w liście końcowym i in.). Służą oni do redukcji przeuczenia.

Lasy losowe, jak i kolejne modele uczenia zespołowego, mogą służyć do uszeregowania zmiennych względem istotności. Jednak metoda określania istotności zmiennych w modelu lasów losowych ma pewne wady. W przypadku danych obejmujących zmienne kategorialne o różnej liczbie poziomów, lasy losowe są faworyzowane na korzyść tych zmiennych, które mają więcej poziomów.⁶⁵

2.2.1.2. Wyjątkowo losowe drzewa

Dodanie kolejnego etapu randomizacji daje inny model znany jako wyjątkowo losowe drzewa (*ExtraTrees*), stworzony przez Pierre Geurtsa, Damiena Ernsta i Louisa Wehenkela w roku 2006.⁶⁶ Głównym celem powstania tego modelu była redukcja obciążenia (przeuczenia) modelu lasów losowych. Chociaż jest podobny do zwykłego modelu lasów losowych, istnieją dwie poważne różnice. Po pierwsze, każde drzewo jest uczone przy użyciu całego zbioru danych. Po drugie, zamiast obliczania lokalnie optymalnego punktu odcięcia dla każdej rozważanej zmiennej (na podstawie np. zanieczyszczenia Giniego), wybierany jest losowy punkt odcięcia z jednolitego rozkładu, granicy którego są granicami zakresu wartości zmiennej wyjaśniającej. W wyniku tego, m.in. model ExtraTrees jest znacznie szybszy niż model lasów losowych. Następnie spośród wszystkich losowo wygenerowanych podziałów, do podziału węzła wybierany jest ten, który daje najwyższy wynik (względem wybranej metryki dopasowania). Podobnie jak w przypadku zwykłych lasów losowych, w każdym węźle można określić liczbę losowo выбрanych zmiennych, które należy uwzględnić.

2.2.1.3. Metody wzmacniania

W metodzie bagging każdy model uczy się niezależnie. Wzmacnianie (ang. *boosting*) zamiast tego działa na zasadzie poprawiania błędów poprzedniego modelu przez kolejny model. Koncentruje się na sekwencyjnym dodawaniu słabych modelu (estymatorów bazowych, *weak learners*), którymi

⁶⁵ Deng H., Runger G., Tuv E., *Bias of Importance Measures for Multi-valued Attributes and Solutions*, 21st International Conference on Artificial Neural Networks, Espoo 2011.

⁶⁶ Geurts P., Ernst D., Wehenkel L., *Extremely randomized trees*, Machine Learning Wyd. 63, 2006, s. 3-42.

najczęściej występują drzewa decyzyjne o niewielkiej głębokości. Obecnie dwoma najpopularniejszymi metodami wzmacniania są:

- **Wzmacnianie adaptacyjne** (*Adaboost, Adaptive Boosting*). Był jednym z pierwszych opracowanych algorytmów wzmacniania zaproponowany przez Freunda i Schapira w 1996 roku.⁶⁷ Główną ideą jest zbudowanie kombinacji wielu słabych modeli, każdy z których uczyony jest na podstawie ważonej próbki danych, gdzie wagi przydzielane są do każdej z obserwacji. Wyższa waga jest przypisana do obserwacji, które zostały źle przewidziane przez poprzednie modele, umożliwiając następnemu estymatorowi bazowemu skupienie się na trudniejszych przypadkach. Ostateczna prognoza jest uzyskiwana poprzez ważoną sumę predykcji wszystkich estymatorów bazowych. Bardziej matematycznie algorytm można opisać jako:

1. Przypisz równe wagi wszystkim $i = 1, \dots, N$ obserwacjom w zbiorze: $w_i = \frac{1}{N}$.
 2. W każdej iteracji $t = 1, \dots, T$:
 - 2.1. Trenuj estymator bazowy h_t na zbiorze danych (\mathbf{X}_i, y_i) z wagami w_i .
 - 2.2. Oblicz ważony bezwzględny błąd estymatora: $\varepsilon_t = \sum_{i=1}^N (w_i \cdot |y_i - h_t(\mathbf{X}_i)|)$.
 - 2.3. Oblicz wagę predykcji estymatora: $\alpha_t = \theta \cdot \ln\left(\frac{(1-\varepsilon_t)}{\varepsilon_t}\right)$, gdzie θ oznacza stopę uczenia (learning rate), $\theta \in (0, 1]$.
 - 2.4. Zaktualizuj wagi dla każdej obserwacji i : $w_i = w_i \cdot \exp(\alpha_t \cdot |y_i - \hat{h}_t(\mathbf{X}_i)|)$.
 - 2.5. Normalizuj wagi: $w_i = \frac{w_i}{\sum_{i=1}^N w_i}$.
 3. Ostateczna prognoza jest ważoną sumą poszczególnych estymatorów: $F(\mathbf{X}) = \sum_{t=1}^T \alpha_t \cdot h_t(\mathbf{X})$
- **Wzmacnianie gradientem** (*gradient boosting*). Gradient boosting został zaprezentowany przez Jeroma Friedmana w 2001 roku.⁶⁸ Jest to metoda uczenia zespołowego, w której słabe modele są trenowane iteracyjnie, aby poprawiać błędy poprzednich. W przeciwieństwie do Adaboost, zamiast sterowaniem wagami obserwacji, w Gradient Boosting kolejne modele są trenowane na resztach (negatywnych gradientach) pomiędzy aktualnymi predykcjami a rzeczywistymi wartościami, co pozwala na poprawę dokładności prognoz. Postać algorytmu wzmacniania gradientem wygląda następująco:
 1. Zainicjuj predykcję $F_0(x)$ na stałą wartość, zwykle na średnią zmiennej objaśnianej: $F(x) = \bar{y}$

⁶⁷ Freund Y., Schapire R. E., *Experiments with a New Boosting Algorithm*, Machine Learning: Proceedings of the Thirteenth International Conference, Nowy Jork 1996.

⁶⁸ Friedman J. H., *Greedy Function Approximation: A Gradient Boosting Machine*, The Annals of Statistics Wyd. 29 Nr. 5, 2001, s. 1189-1232.

2. W każdej iteracji $t = 1, \dots, T$:

2.1. Oblicz reszty: $r_i = -\left[\frac{\partial L(y_i, F(\mathbf{X}_i))}{\partial F(\mathbf{X}_i)}\right]_{F(\mathbf{X})=F_{t-1}(\mathbf{X})}$, gdzie L oznacza funkcję straty (loss function), na przykład błąd średniokwadratowy (MSE).

2.2. Trenuj estymator bazowy h_t na zbiorze (\mathbf{X}_i, r_i) .

2.3. Oblicz optymalną stopę uczenia: $\alpha_t = \min_{\alpha} \sum_{i=1}^N (L(y_i, F(\mathbf{X}_i) + \theta \cdot h_t(\mathbf{X}_i)))$.

2.4. Zaktualizuj predykcję: $F_t(\mathbf{X}) = F_{t-1}(\mathbf{X}) + \alpha_t \cdot h_t(\mathbf{X})$.

3. Ostateczna prognoza jest zaktualizowaną predykcją z ostatniej obserwacji procedury wzmacniania: $F(\mathbf{X}) = F_T(\mathbf{X})$, czyli skumulowaną sumą poszczególnych estymatorów bazowych skorygowanych o optymalną stopę uczenia.

Chociaż istnieje kilka różnic między tymi dwoma metodami wzmacniania, oba algorytmy podążają tą samą ścieżką i mają podobne korzenie historyczne. Oba algorytmy działają w celu zwiększenia wydajności prostego uczenia poprzez iteracyjne przesuwanie uwagi na problematyczne obserwacje, które są trudne do przewidzenia. W przypadku AdaBoost przesunięcie odbywa się poprzez zwiększanie wagi obserwacji, które zostały wcześniej błędnie zaprognozowane, podczas gdy Gradient Boosting identyfikuje trudne obserwacje za pomocą dużych reszt obliczonych w poprzednich iteracjach.

W ostatnim dziesięcioleciu podstawowa wersja wzmacniania gradientem została znacznie rozwinięta. W niniejszym badaniu rozważone będą trzy najbardziej znane nowoczesne rozszerzenia:

- **xgboost.** Początkowo będąc projektem badawczym Tianqi Chena w roku 2014,⁶⁹ model xgboost (skrajne wzmacnianie gradientem, *Extreme Gradient Boosting*) zyskał na ogromną popularność po wygraniu znanych konkursów uczenia maszynowego. Xgboost to jedna z najszybszych implementacji wzmacniania gradientem. Ten model przygląda się rozkładowi funkcji we wszystkich punktach danych w liściu i wykorzystuje te informacje do zmniejszenia przestrzeni wyszukiwania możliwych podziałów funkcji. Co więcej, xgboost implementuje kilka technik restrykcyjnych, umożliwiając szybkie przeszukiwanie hiperparametrów. Algorytm modelu wygląda odpowiednio:

1. Zainicjuj predykcję $F_0(x)$ na stałą wartość, zwykle na średnią zmiennej objaśnianej: $F(x) = \bar{y}$.

2. W każdej iteracji $t = 1, \dots, T$:

2.1. Oblicz ujemny gradient: $g_i = -\left[\frac{\partial L(y_i, F(\mathbf{X}_i))}{\partial F(\mathbf{X}_i)}\right]_{F(\mathbf{X})=F_{t-1}(\mathbf{X})}$.

⁶⁹ Chen T., *XGBoost: A Scalable Tree Boosting System*, University of Washington, Washington 2014.

$$2.1. \text{ Oblicz Hesjan: } H_i = - \left[\frac{\partial^2 L(y_i, F(\mathbf{X}_i))}{\partial F(\mathbf{X}_i)^2} \right]_{F(\mathbf{X})=F_{t-1}(\mathbf{X})}.$$

2.2. Trenuj estymator bazowy h_t na zbiorze $(\mathbf{X}_i, \frac{g_i}{H_i})$.

2.3. Oblicz optymalną stopę uczenia: $\alpha_t = \min_{\alpha} \sum_{i=1}^N \left(\frac{1}{2} H_i \left(h_t(\mathbf{X}_i) - \frac{g_i}{H_i} \right)^2 \right)$.

2.4. Zaktualizuj predykcję: $F_t(\mathbf{X}) = F_{t-1}(\mathbf{X}) + \theta \cdot \alpha_t \cdot h_t(\mathbf{X})$.

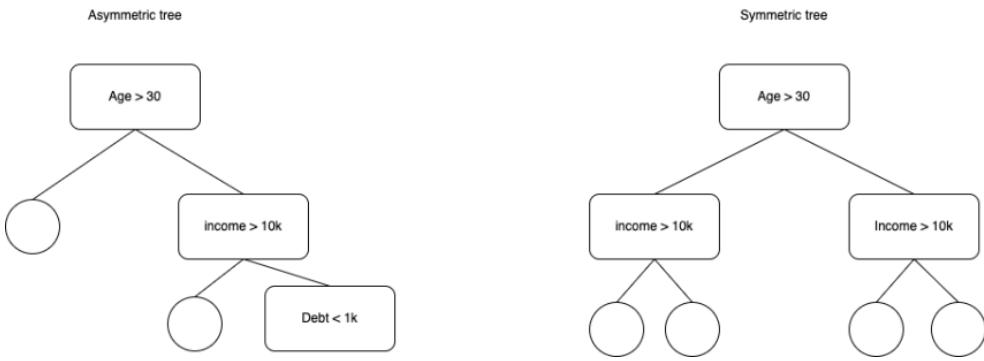
3. Ostateczna prognoza jest zaktualizowaną predykcją z ostatniej obserwacji procedury wzmacniania: $F(\mathbf{X}) = F_T(\mathbf{X})$. Xgboost działa jak metoda Newtona-Raphsona w przestrzeni funkcjnej, w przeciwieństwie do wzmacniania gradientem.

- **LightGBM.** Przedstawiony w roku 2017 przez Microsoft,⁷⁰ model ten ma wiele zalet xgboost'a, w tym możliwość równoległych obliczeń, regulowanie obciążenia i wczesne zatrzymywanie. Co do algorytmu modelu, nie różni się on od algorytmu zaprezentowanego dla modelu xgboost. Główna różnica między dwoma modelami kroi się w estymatorze bazowym. LightGBM nie rozwija drzewa poziomo, jak robi to xgboost. Zamiast tego, buduje drzewo pionowo, zaczynając od liścia, który przyniesie największy spadek strat. Zwykle oznacza to również przyspieszenie obliczeń, ponieważ taka struktura podziału procesów okazała się bardziej efektywna względem czasu obliczeń. Poza tym LightGBM nie korzysta z szeroko stosowanego algorytmu uczenia drzew decyzyjnych opartego na sortowaniu, który wyszukuje najlepszy punkt podziału na posortowanych wartościach zmiennych, jak to robi się w klasycznym drzewie decyzyjnym. Zamiast tego w LightGBM zaimplementowany jest zoptymalizowany algorytm uczenia drzewa decyzyjnego oparty na histogramie, który zapewnia korzyści zarówno pod względem wydajności, jak i zużycia pamięci komputera.
- **CatBoost** jest najbardziej nowoczesną metodą wśród metod wzmacniania, zaprezentowana w roku 2019 przez Yandex.⁷¹ CatBoost zawsze buduje symetryczne (zrównoważone) drzewa, w przeciwieństwie do xgboost i LightGBM. Oznacza to, że każda gałąź na bieżącej głębokości drzewa będzie zawierała tyle samo liści.

Rysunek 2. Drzewo asymetryczne i symetryczne

⁷⁰ Ke G. i inn., *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach 2017.

⁷¹ Prokhorenko L. i inn., *CatBoost: unbiased boosting with categorical features*, Moscow Institute of Physics and Technology, Dolgoprudny 2019.



Źródło: opracowanie własne

Ta zrównoważona architektura umożliwia kontrolowanie nadmiernego dopasowania, ponieważ służy jako regularyzacja. Co więcej, inne algorytmy wzmacniania gradientem są podatne na przeuczenie, ponieważ podczas obliczania oszacowania gradientu wykorzystują te same obserwacje ze zbioru danych, z których zbudowano model, dzięki czemu uczą się na tych samych danych kilkakrotnie. CatBoost wykorzystuje koncepcję wzmacniania uporządkowanego: jest to podejście oparte na permutacji, czyli na dopasowaniu modelu i obliczaniu reszt na różnych pod próbach. W ten sposób model zapobiega nadmierнемu dopasowaniu.

2.2.2. Sztuczne sieci neuronowe

Sztuczne sieci neuronowe (*artificial neural networks*, ANNs), są zaawansowanymi modelami uczenia maszynowego (dokładniej uczenie głębokiego, które jest częścią uczenia maszynowego), inspirowanymi działaniem ludzkiego mózgu. Powstały z próby naśladowania zdolności mózgu do przetwarzania informacji poprzez połączenia między neuronami. Początki koncepcji sztucznych sieci neuronowych sięgają 1943 roku, kiedy Warren McCulloch i Walter Pitts zaproponowali modele matematyczne, które symulowały działanie biologicznych neuronów.⁷² Jednakże prawdziwy przełom nastąpił w latach 1980-1990, kiedy David Rumelhart⁷³, Wei Zhang⁷⁴, Yann LeCun⁷⁵ i inni naukowcy rozwijali nowe algorytmy uczenia i architektury sieci neuronowych. Sieci neuronowe składają się z warstw sztucznych neuronów, które przetwarzają i przekształcają dane. Główne typy warstw to:

- Warstwa wejściowa: przyjmuje dane i przekazuje je do kolejnych warstw.

⁷² McCulloch W., Pitts W., *A Logical Calculus of Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics Wyd. 5 Nr. 4, 1943, s. 115–133.

⁷³ Rumelhart D. E., Hinton G. E., Williams R. J., *Learning representations by back-propagating errors*, Nature Wyd. 323, 1986, s. 533–536.

⁷⁴ Zhang W. i inn, *Parallel distributed processing model with local space-invariant interconnections and its optical architecture*, Applied Optics Wyd. 29 Nr. 32, 1990, s. 4790-4797.

⁷⁵ LeCun Y. i inn., *Backpropagation Applied to Handwritten Zip Code Recognition*, Neural Computation, Wyd. 1, 1989, s. 541–551.

- Warstwy ukryte: stanowią główną część modelu, w której odbywa się przetwarzanie i wyodrębnianie zmiennych.
- Warstwa wyjściowa: generuje końcowe wartości wyjściowe.

Proces dopasowania sieci neuronowych polega na dostosowywaniu wag połączeń między neuronami w trakcie „uczenia” na zbiorze danych, gdyż model stara się minimalizować różnicę między prognozami a rzeczywistymi wartościami. Proces ten wykorzystuje różne algorytmy optymalizacji, takie jak np. spadek gradientowy, aby dostosować wagi połączeń. Sieci neuronowe znalazły zastosowanie w różnych dziedzinach, takich jak rozpoznawanie obrazów, przetwarzanie języka naturalnego, klasyfikacja danych, prognozowanie, gry komputerowe itp. Ich zdolność do wyodrębniania złożonych wzorców z danych sprawia, że są niezwykle użytecznym narzędziem w analizie i przetwarzaniu informacji. Dlatego ciekawe będzie zaobserwowanie działania sieci neuronowych na przykładzie danych przestrzennych.

2.2.2.1 Perceptron wielowarstwowy

Perceptron wielowarstwowy (ang. *multilayer perceptron*, MLP) to klasa w pełni połączonych sztucznych sieci neuronowych z wyprzedzeniem (*feedforward fully connected ANN*). Termin MLP jest używany niejednoznacznie, czasami w odniesieniu do dowolnej ANN z wyprzedzeniem, czasami ściśle w odniesieniu do sieci składających się z wielu warstw perceptronów. Wielowarstwowe perceptrony są też czasami potocznie określane jako „podstawowe” sieci neuronowe, zwłaszcza gdy mają jedną ukrytą warstwę.

W MLP, z wyjątkiem warstwy wejściowej, neurony wykorzystują nieliniową funkcję aktywacji. Funkcja aktywacji określa, czy neuron powinien być aktywowany, czy nie, na podstawie ważonej sumy danych wejściowych. Kiedy neuron jest aktywowany, oznacza to, że wytwarza sygnał wyjściowy w odpowiedzi na dane wejściowe. Nie wszystkie neurony będą aktywowane w danej warstwie ukrytej, bo inaczej zgodnie z algebrą liniową, dowolną liczbę warstw można byłoby zredukować do dwuwarstwowego modelu wejścia-wyjścia. Dwie najbardziej znane funkcje aktywacji są sigmoidami i można je przedstawić jako:

$$y(v_i) = \frac{e^{v_i} - e^{-v_i}}{e^{v_i} + e^{-v_i}} \text{ (tangens hiperboliczny)}, \quad (11)$$

$$y(v_i) = (1 + e^{v_i})^{-1} \text{ (funkcja logistyczna)}, \quad (12)$$

gdzie:

- y_i jest wartością wyjściową i -tego neuronu,
- v_i jest ważoną sumą połączeń wejściowych.

Zaproponowano również alternatywne funkcje aktywacji, w tym najczęściej używana w ostatnich opracowaniach tzw. poprawiona funkcja liniowa (*Rectified Linear Units*, ReLU):⁷⁶ $\max(0, a + v_i b)$.

Ponieważ warstwy MLP są w pełni połączone, każdy neuron w jednej warstwie łączy się z określoną wagą w_{ij} z każdym neuronem w następnej warstwie. Liczba warstw ukrytych jest zadana przez użytkownika. Z kolei dopasowanie modelu odbywa się poprzez zmianę wag po przetworzeniu każdej frakcji danych, na podstawie otrzymanego błędu dopasowania. Jest to przeprowadzane poprzez propagację wsteczną (ang. *backpropagation*), algorytmie opartym na minimalizacji sumy kwadratów błędów. Matematycznie algorytm propagacji wstecznej można opisać jako:

1. Losowo zainicjuj wagi i obciążenia (*biases*).
2. Powtóż następujące kroki dla każdej obserwacji ze zbioru zmiennych objaśniających \mathbf{X} i zmiennej objaśnianej, czyli dla (\mathbf{X}, \mathbf{y}) :

2.1. Wykonaj propagację do przodu:

Dla każdej warstwy l oblicz ważoną sumę wartości wejściowych z_j^l i wartości wyjściowych a_j^l dla każdego neuronu j w warstwie l , korzystając z poniższych wzorów:

$$\begin{aligned} z_j^l &= \sum_{k=1}^D w_{jk}^l \cdot a_k^{l-1} + b_j^l \\ a_j^l &= \sigma(z_j^l) \end{aligned} \quad (13)$$

gdzie:

w_{jk}^l – waga między neuronem k w warstwie $(l - 1)$ a neuronem j w warstwie l ,

a_k^{l-1} – wartość wyjściowa neuronu k w warstwie $(l - 1)$,

b_j^l – obciążenie neuronu j w warstwie l ,

σ – funkcja aktywacji zastosowana dla każdego elementu z_j^l .

2.2. Oblicz błąd dla warstwy wyjściowej (ostatniej warstwy, L):

Oblicz błąd δ_j^L dla każdego neuronu j w warstwie wyjściowej L , korzystając z następującego wzoru:

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} \quad (14)$$

gdzie C to funkcja kosztu (np. błąd średniokwadratowy), która mierzy różnicę między przewidywaną wartością wejściową a rzeczywistą wielkością zmiennej objaśnianej.

2.3. Dokonaj propagacji wstecznej błędu:

Dla każdej warstwy l , zaczynając od przedostatniej warstwy $L - 1$ i wracając do pierwszej warstwy

⁷⁶ Fukushima K., *Visual feature extraction by a multilayered network of analog threshold elements*, IEEE Transactions on Systems Science and Cybernetics Wyd. 5 Nr. 4, 1969, s. 322–333.

ukrytej $l = 1$, oblicz błąd δ_j^l dla każdego neuronu j w warstwie l za pomocą następującego wzoru:

$$\delta_j^l = (\sum_{k=1}^{D_{l+1}} w_{kj}^{l+1} \cdot \delta_k^{l+1}) \cdot \sigma'(z_j^l) \quad (15)$$

gdzie:

D_{l+1} – liczba neuronów w warstwie $l + 1$,

w_{kj}^{l+1} – waga między neuronem j w warstwie l a neuronem k w warstwie $l + 1$,

δ_k^{l+1} – błąd neuronu k w warstwie $(l + 1)$,

σ' – pochodną funkcji aktywacji zastosowanej dla każdego elementu z_j^l .

2.4. Oblicz gradienty:

Dla każdej warstwy l oblicz gradient funkcji kosztu w odniesieniu do obciążen $\frac{\partial C}{\partial b_j^l}$ oraz gradient

funkcji kosztu w odniesieniu do wag $\frac{\partial C}{\partial w_{jk}^l}$ za pomocą następujących wzorów:

$$\begin{aligned} \frac{\partial C}{\partial b_j^l} &= \delta_j^l \\ \frac{\partial C}{\partial w_{jk}^l} &= a_k^{l-1} \cdot \delta_j^l \end{aligned} \quad (16)$$

3. Zaktualizuj wagi i obciążenia:

Dla każdej warstwy l zaktualizuj wagi i odchylenia za pomocą algorytmu optymalizacji opartego na spadku gradientowym z obliczonymi gradientami z kroku 2.4. W przypadku stochastycznego spadku gradientowego, dla każdej wagi w_{ij} i obciążenia b_j aktualizacja jest wykonywana w następujący sposób:

$$\begin{aligned} b_j &= b_j - \eta \cdot \frac{\partial C}{\partial b_j^l} \\ w_{ij} &= w_{ij} - \eta \cdot \frac{\partial C}{\partial w_{jk}^l} \end{aligned} \quad (17)$$

gdzie η jest stopą uczenia, $\eta \in (0, 1]$.

4. Powtarzaj kroki 2-3 przez określoną liczbę iteracji lub do osiągnięcia konwergencji (tzn., kiedy wydajność modelu przestanie się poprawiać).

Powtarzając iteracyjnie algorytm propagacji wstecznej dla całego zbioru danych i dostosowując parametry modelu, MLP uczy się dokonywać coraz lepszych prognoz i przybliżać zależności zachodzące pomiędzy zmennymi.

2.2.2.2. Przestrzenne sieci neuronowe. Geograficznie ważona sztuczna sieć neuronowa

Przestrzenne sieci neuronowe (*Spatial Neural Networks*, SNN) stanowią odrębną kategorię sieci neuronowych do reprezentowania i przewidywania zjawisk przestrzennych. Generalnie

zakładano, że poprawiają one dokładność przestrzennych (klasycznych) sieci neuronowych. Historia zastosowania sieci neuronowych do przestrzennych zbiorów danych sięga początku 1990-ch, kiedy Openshaw⁷⁷ oraz Hewitson⁷⁸ rozpoczęli prace nad zastosowaniem przestrzennych sieci neuronowych do badania zjawisk geograficznych. Jednak przestrzenne sieci neuronowe przezywają teraz tylko początek swojego rozwoju, gdyż pierwsze prace na temat przestrzennych sieci neuronowych powstały w latach 2021-2022. Na moment napisania niniejszej pracy, jedyną praktycznie zaimplementowaną przestrzenną siecią neuronową jest geograficznie ważona sztuczna sieć neuronowa⁷⁹, którą i będzie reprezentowała klasę przestrzennych sieci neuronowych. Perspektywy rozwoju przestrzennych sieci neuronowych z kolei będą przedyskutowane na zakończeniu niniejszej pracy.

Geograficznie ważona sztuczna sieć neuronowa (ang. *Geographically Weighted Artificial Neural Network, GWANN*) łączy ważenie geograficzne ze sztucznymi sieciami neuronowymi, które są w stanie uczyć się złożonych nieliniowych relacji bez żadnych dodatkowych założeń. Powstała w odpowiedzi na model GWR w którym przyjmuje się, że relacje między zmiennymi zależnymi i niezależnymi są liniowe, co często nie jest spełnione w praktyce. Podobnie jak GWR, GWANN wykorzystuje jądrową funkcję oporu przestrzeni (ang. *distance-decay kernel function*) i parametr szerokości pasma do przypisywania wag obserwacjom geograficznym.

Wagi połączeń w modelu perceptronu wielowarstwowego z warstwy ukrytej do warstwy wyjściowej można interpretować jako współczynniki liniowego modelu zmiennych nieliniowo przekształconych. Tak więc, gdy wagi połączeń między warstwą ukrytą a wyjściową są szacowane przy użyciu funkcji błędu ważonego geograficznie, wagi te można interpretować jako model GWR.⁸⁰ Architektura GWANN jest identyczna do architektury MLP, z tą różnicą, że każdy wyjściowy neuron w modelu GWANN jest przypisany do lokalizacji w przestrzeni geograficznej.

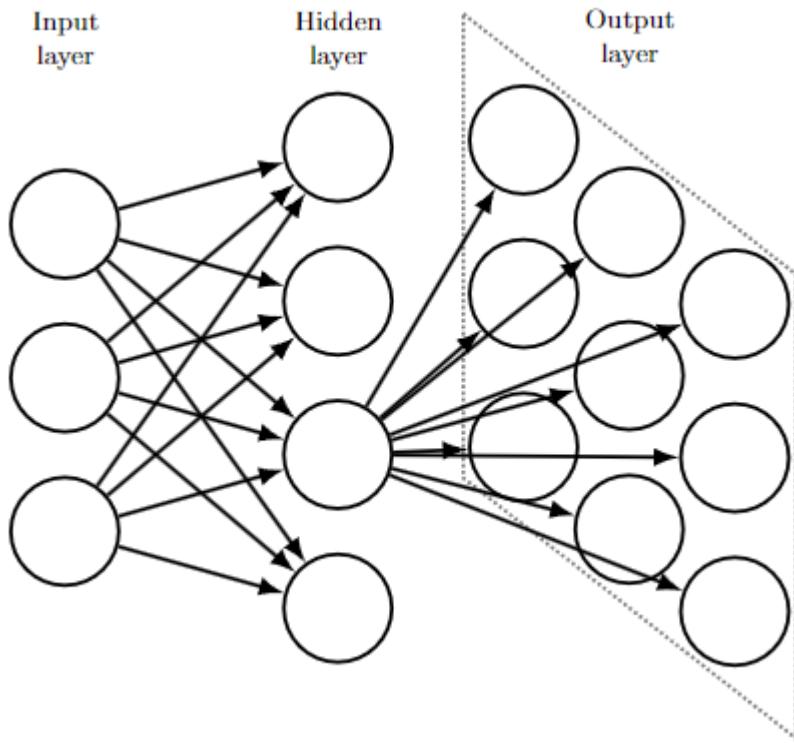
Rysunek 3. Architektura geograficznie ważonej sztucznej sieci neuronowej

⁷⁷ Openshaw S., *Modelling spatial interaction using a neural net*, Geographic Information Systems, Spatial Modelling and Policy Evaluation, 1993, s. 147–164.

⁷⁸ Hewitson B., Crane R., *Neural nets: applications in geography*, The GeoJournal Library Wyd. 29, 1994, s. 196.

⁷⁹ Hagenauer J., Helbich M., op cit.

⁸⁰ Ibidem.



Źródło: Hagenauer J., Helbich M., *A geographically weighted artificial neural network*, *International Journal of Geographical Information Science* Wyd. 36 Nr. 2, 2022, s. 215-235.

Poza architekturą, kluczowa różnica między GWANN a MLP polega na tym, że GWANN używa geograficznie ważonej funkcji błędu zamiast podstawowej kwadratowej funkcji. Geograficznie ważona funkcja błędu jest definiowana jako:

$$e = \frac{1}{2} \sum_{i=1}^N v_i (t_i - z_i)^2, \quad (18)$$

gdzie:

t_i jest wartością zmiennej objaśnianej,

z_i – wartością wyjściową neuronu wyjściowego i ,

v_i – ważoną geograficznie odległośćą między obserwacją a lokalizacją neuronu wyjściowego i .

Zgodnie z taką definicją, jeżeli odległość pomiędzy neuronem wyjściowym a obserwacją w przestrzeni jest małą, różnicom przypisuje się większa waga niż wtedy, gdy są one dalej od siebie. Należy zauważyć, że przy takiej konstrukcji błędu liczba neuronów wyjściowych musi być identyczna liczbowi obserwacji w zbiorze. W tym celu np. w przypadku, gdy chce się obliczyć wartość funkcji błędu geograficznego dla pojedynczej wartości docelowej na już dopasowanym modelu, konieczne jest powtórzenie wartości zmiennej objaśnianej tyle razy, ile jest neuronów w warstwie wyjściowej.

Po zdefiniowaniu geograficznie ważonej funkcji błędu, obliczenie błędu sygnału podczas propagacji wstecznej modyfikuje się w następujący sposób:

$$\delta_j = \begin{cases} \sigma'(z_j^l) v_j (a_k^l - t_j), & l = L \\ \sigma'(z_j^l) \sum_{k=1}^D \delta_k^l w_{jk}^l, & w \text{ p.p.} \end{cases} \quad (19)$$

gdzie:

- σ – funkcja aktywacji zastosowana dla każdego elementu z_j^l
- z_j^l – wartości wejściowe dla neuronu j w warstwie l ,
- v_j – geograficznie ważona odległość między obserwacją a lokalizacją neuronu wyjściowego j .
- a_k^l – wartość wyjściowa neuronu k w warstwie l ,
- t_j – wartość docelowa neuronu j ,
- w_{jk}^l – waga między neuronem k w warstwie $(l-1)$ a neuronem j w warstwie l ,

Ważenie geograficzne jest używane tylko do obliczania sygnału błędu neuronów wyjściowych.

Podobnie jak w MLP, wagi połączeń GWANN są dostosowywane przy użyciu spadku gradientowego dokładnie jak pokazano w modelu MLP.

W ten sposób przedstawiono wszystkie model które będą użyte do estymacji cen mieszkań w Warszawie. Kolejny podrozdział zawiera opis algorytmu strojenia hiperparametrów oraz kryterium porównawczego względem których dokonano oceny wydajności modelu.

2.3. Strojenie hiperparametrów. Algorytm genetyczny

Jak wspomniano wcześniej, strojenie parametrów odgrywa istotną rolę w kontrolowaniu wydajności modeli uczenia maszynowego. Strojenie hiperparametrów obejmuje optymalizację parametrów modelu, które nie są bezpośrednio dopasowywane do danych, ale znaczco wpływają na jego wydajność. Na przykład, w przypadku lasów losowych takim parametrem występuje liczba drzew decyzyjnych B . W niniejszej pracy zaproponowano algorytm genetyczny (ang. *genetic algorithm*, GA) przeszukiwania przestrzeni hiperparametrów. Jest on często stosowany do rozwiązywania problemów optymalizacyjnych w różnych dziedzinach ze względu na efektywność w przeszukiwaniu nieliniowych i wielowymiarowych przestrzeni.

Algorytm genetyczny GA to metaheurystyczna technika optymalizacji inspirowana naturalnym procesem ewolucji. W 1958 roku Barker opublikował serię artykułów modelujących sztuczną selekcję wśród organizmów.⁸¹ Stworzony fundament umożliwił wdrożenie modeli procesów ewolucyjnych, gdyż podejście Barkera zawierało wszystkie najważniejsze elementy współczesnych algorytmów genetycznych. Ogólna idea algorytmu polega na utrzymywaniu populacji rozwiązań

⁸¹ Barker J. S., *Simulation of Genetic Systems by Automatic Digital Computers*, Australian Journal of Biological Sciences Wyd. 11 Nr. 4, 1958, s. 603-612.

kandydujących (kandydatów) i rozwijanie je w kierunku optimum przy użyciu operatorów genetycznych, takich jak selekcja, krzyżowanie i mutacja. Funkcja dopasowania ocenia jakość każdego kandydata, kierując algorytm w stronę lepszych rozwiązań w miarę upływu czasu (pokoleń). Matematycznie ogólną postać algorytmu można przedstawić następująco:

1. Losowo inicjalizuj populację pierwotną składającą z N kandydatów: $\mathbf{P} = \{p_1, p_2, \dots, p_n\}$.
2. Oceń jakość kandydata p_i przez $F(p_i)$, gdzie $F(x)$ oznacza funkcję dopasowania (np. błąd średniokwadratowy dopasowania modelu przy użyciu hiperparametrów kandydata p_i).
3. Wybierz kandydatów z populacji \mathbf{P} z prawdopodobieństwem $P(p_i) = \frac{F(p_i)}{\sum F(p_j)}$.
4. Stwórz potomków o_1 i o_2 z rodziców p_i i p_j stosując skrzyżowanie w losowej pozycji k :

$$o_1 = p_i[1:k] + p_j[k+1:L]$$

$$o_2 = p_j[1:k] + p_i[k+1:L].$$

5. Zmutuj potomków o_i z prawdopodobieństwem P_m w losowej pozycji m :

$$o_i[m] = o_i[m] + \Delta.$$

6. Zastąp starą populację \mathbf{P} nową populacją $\mathbf{P}_{new} = \{o_1, o_2, \dots, o_n\}$.
7. Powtarzaj kroki 2-6 dla określonej liczby pokoleń lub do momentu spełnienia warunku zakończenia.

2.4. Kryterium porównawcze

Skoro celem niniejszej pracy jest ocena jakości predykcji cen mieszkań, głównym kryterium oceny występują funkcje mierzące błąd wynikający z różnicy pomiędzy wartościami rzeczywistymi zmiennej objaśnianej y a wartościami prognozowanymi przez model \hat{y} .

Tabela 2. Funkcję błędu

Skrót	Nazwa	Formuła
MSE	Błąd średniokwadratowy	$\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$
RMSE	Pierwiastek błędu średniokwadratowego	$\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$
MPE	Średni błąd procentowy	$\frac{100\%}{N} \sum_{i=1}^N \frac{ y_i - \hat{y}_i }{y_i}$

MAPE	Średni procentowy błąd bezwzględny	$\frac{100\%}{N} \sum_{i=1}^N \left \frac{y_i - \hat{y}_i}{y_i} \right $
sMAPE	Średni symetryczny procentowy błąd bezwzględny	$\frac{100\%}{N} \sum_{i=1}^N \frac{2 \cdot y_i - \hat{y}_i }{ y_i + \hat{y}_i }$

Źródło: opracowanie własne

Oprócz wymienionych błędów, ocena zostanie wzbogacona wykresami reszt dla poszczególnych modeli, interpretacją współczynników modelu (jeżeli jest to możliwe) lub oceną istotności zmiennych wyjaśniających (w przypadku metod uczenia zespołowego i modeli autoregresyjnych) oraz reprezentacją czasu obliczeniowego dla każdego ze wspomnianych modeli, co może być szczególnie istotnym dla praktycznego zastosowania podczas podejmowania rzeczywistych procesów decyzyjnych.

Powyższy gruntowny opis użytych modeli, metody strojenia hiperparametrów i kryterium porównawczych pozwoli na klarowne inicjowanie i przeprowadzenie analitycznej części niniejszego badania.

III. Zbiór danych

3.1. Gromadzenie i źródła danych

Zestaw danych analizowany w niniejszym badaniu pochodzi z trzech źródeł. Pierwsza część została pozyskana z serwisu internetowego otodom.pl⁸² – platformy zawierającej oferty sprzedaży i wynajmu tysięcy mieszkań oraz domów na obszarze całej Rzeczypospolitej Polskiej, będąc jednej z najbardziej popularnych platform tego typu w Polsce. Dane zostały zebrane przy użyciu metody web scrapingu, za pomocą interfejsu programistycznego (ang. *API*), którego implementację dostarcza pakiet BeautifulSoup dedykowany dla języka programowania Python. Procedura dotyczyła wyłącznie określonych ofert sprzedaży nieruchomości, a konkretnie lokali mieszkalnych w obszarach miasta stołecznego Warszawy dostępnych w dniu 1 czerwca 2023. W praktyce proces zbierania danych polegał na pobraniu struktury strony internetowej dla każdej oferty, a następnie jej zapisywaniu w formacie JSON. Ostatecznie uzyskano zbiór danych zawierający łącznie 10124 oferty. Warto podkreślić, że rozmiar zestawu stanowi poważną reprezentację rocznej dynamiki sprzedaży mieszkań, biorąc pod uwagę, że w roku 2022 w Warszawie zawarto łącznie około 10750 transakcji sprzedaży.⁸³ Niemniej ważne jest to, że pomimo potencjalnego braku zrealizowania transakcji sprzedaży dla wszystkich zagregowanych mieszkań, to jednakże zbiór ten zawiera informacje, stanowiące punkt odniesienia dla typowych decydentów zakupu mieszkania.

Drugą komponentą danych były cechy o charakterze geograficznym. Uzyskano je za pośrednictwem projektu społeczności internetowej OpenStreetMap.⁸⁴ W rezultacie powstał zbiór danych obejmujący konkretne lokalizacje geograficzne (na stan 1 czerwca 2023) oraz odpowiadające im współrzędne geograficzne. Poniżej przedstawiono zestawienie lokalizacji zgodnie z przypisanymi im zmiennymi:

Tabela 3. Wykorzystane lokalizacje geograficzne

Kategoria	Zmienne w ostatecznym zbiorze	Lokalizacje geograficzne
Sklepy (n = 7634)	food_shop, beauty_shop, pharmacy, jeweller, shopping_mall, shop	sklepy spożywcze od małych kiosków do supermarketów, sklepy kosmetyczne, apteki, jubilerzy, centrum handlowe, wszystkie inne sklepy (np. meblowe, sportowe, odzieżowe)

⁸² <https://www.otodom.pl/pl/wyniki/sprzedaz/mieszkanie/mazowieckie/warszawa/warszawa/warszawa> (dostęp 15.08.2023).

⁸³ <https://rednetconsulting.pl/aktualnosci.tresc.id,10504,tytul,sytuacja-na-rynk-mieszkaniowym-iv-kwartal-2022> (dostęp 15.08.2023).

⁸⁴ <https://www.openstreetmap.org/relation/336075> (dostęp 15.08.2023).

Gastronomia (n = 3880)	fast_food, bar, restaurant	restauracje fast food, puby, bary, ogródki piwne, kawiarnie, restauracje i korty gastronomiczne
Instytucje publiczne (n = 2912)	health_institution, public_institution, public_service, educational, educational_children, dormitory, temple_catholic, temple_other	przychodnie, szpitale i domy opieki, biblioteki, ambasady, sądy, ratusz, urzędy pocztowe, remizy strażackie, komisariaty policji, kolegia, uniwersytety, szkoły, przedszkola, akademiki, świątynie katolickie, inne świątynie (chrześcijańskie prawosławne, chrześcijańskie luterańskie, żydowskie, buddyjskie, muzułmańskie)
Obiekty rekreacyjne (n = 5545)	sport_object, cultural, attraction, nightclub	pływalnie, centrum sportowe, pola golfowe, stadiony, centrum sztuki, muzea, teatry, kina, ogrody zoologiczne, domy kultury, parki rozrywkowe, pomniki, ruiny, forty, fontanny, punkty widokowe, wieże, zamki, kluby nocne
Obiekty przyrodnicze (n = 2158)	swamp, water_object, river, park	Bagna, rezerwuarze wodne, rzeka Wisła, parki, miejsca piknikowe, wybiegi dla psów, kempingi, rezerwaty przyrody, lasy
Samochody i rowery (n = 29165)	car service, garage, gas station, parking, bike parking, bike rent	sprzedawcy, wypożyczalnie, myjnie samochodów, garaże, stacje paliw, miejsca parkingowe, miejsca parkingowe dla rowerów, wypożyczalnie rowerów
Komunikacja miejska i drogi (n = 211753)	subway_entrance, bus_stop, tram_stop, train_stop, roads_*	przystanki autobusowe, tramwajowe, kolejowe, stacje metra, lotniska, drogi główne, drugo- i trzeciorzędne, ścieżki, chodniki, ścieżki rowerowe, drogi ciężarówkowe, autostrady.
Inne (n = 9807)	industrial, construction, service, renthouse, office, bank, playground, graveyard, prison	zakłady produkcyjne, oczyszczalnie ścieków, fabryki i inne, budowy i renowacje, naprawy komputerów, pralnie, fryzjerzy, naprawy wodociągów, hostele, hotele, pensjonaty, motyle, biura, banki, place zabaw, cmentarze, więzienia

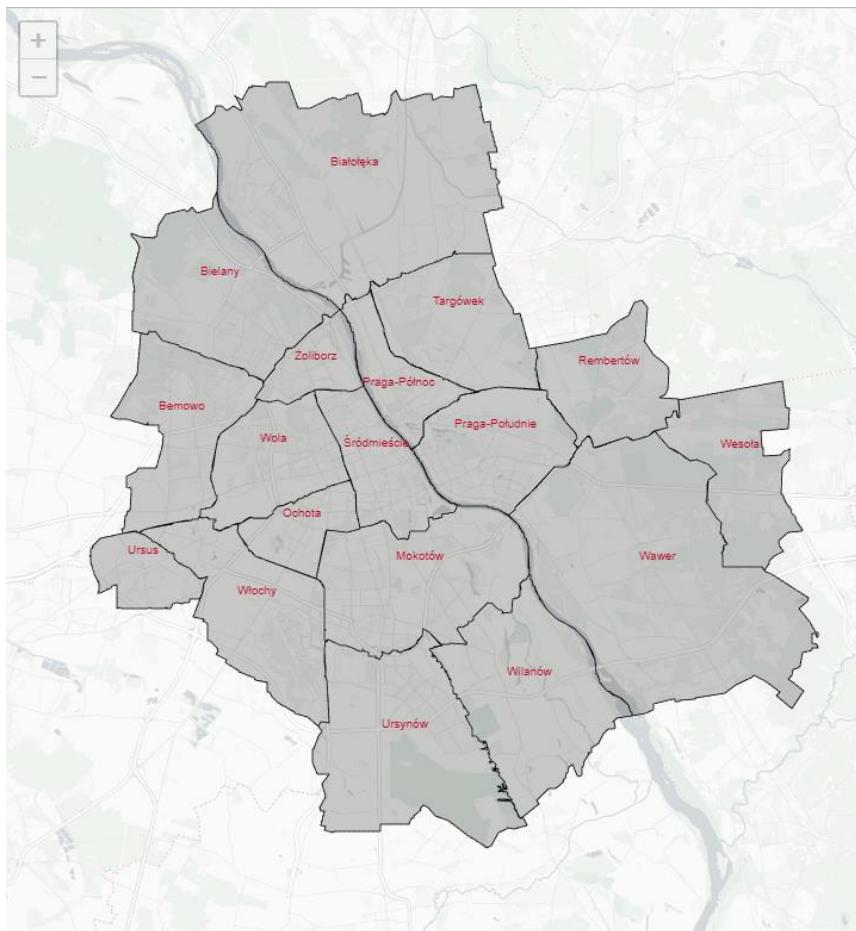
Zródło: opracowanie własne

W konsekwencji zbiór danych zawierał łącznie 272854 obiekty. Przeważającą liczbę (211753) stanowiły drogi i infrastruktura komunikacyjna miejska. W tym momencie warto poruszyć kwestię jakości danych dostarczanych przez OpenStreetMap. Aktualnie brakuje badań lub raportów dotyczących stopnia pokrycia obszaru Warszawy. Jednakże w analizie jakości danych OpenStreetMap dla wybranych powiatów przeprowadzonej przez Borkowską i Pokoniecznego, stwierdzono, że stopień pokrycia dla sąsiadującego z Warszawą powiatu piaseczyńskiego wynosił 82%.⁸⁵ Ze względu na większe znaczenie gospodarcze oraz wyższą liczbę ludności Warszawy, przyjęto założenie, że stopień pokrycia dla Warszawy wynosi co najmniej 82%.

⁸⁵ Borkowska S., Pokonieczny K, *Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development*, Sustainability Wyd. 14 Nr. 7, 2022, s.3728.

Ostatnia część danych pochodziła z zasobów bazy globalnych obszarów administracyjnych (ang. *Database of Global Administrative Areas*, GADM).⁸⁶ Pozyskano stąd granice administracyjne Warszawy wraz z określeniem poszczególnych dzielnic, umożliwiając dalszą wizualizację zmiennych na mapie geograficznej.

Rysunek 4. Umiejscowienie geograficzne dzielnic Warszawy



Źródło: opracowanie własne

Co więcej, wysoka rozdzielcość uzyskanych danych pozwoliła na zawężenie zbioru ofert sprzedaży mieszkań do 10091 obserwacji, eliminując jednostki znajdujące poza faktycznymi obszarami administracyjnymi. Z posiadanymi danymi przystąpiono do obróbki zmiennych opisujących współrzędne geograficzne obiektów na mapie, przekształcając je w cechy charakteryzujące konkretne mieszkania.

⁸⁶ https://gadm.org/download_country36.html (dostęp 15.08.2023).

3.2. Inżynieria cech

Wykorzystując dane pozyskane za pośrednictwem platformy OpenStreetMap, stworzono szereg zmiennych opartych na odległości pomiędzy mieszkaniami a każdym z obiektów geograficznych. Procedura obliczeniowa, mająca na celu wyznaczenie elipsoidalnych odległości między dwoma punktami, przyjęła formę opisaną przez Tadeusza Vincenty⁸⁷. Za pomocą tego algorytmu, wyliczone zostały odległości między każdym mieszkaniem i każdym obiektem, co daje sumarycznie ponad 2,75 miliardów pomiarów. Ze względu czasu obliczeniowego, operacja przygotowania zmiennych trwała około 24 godzin przy wykorzystaniu 12 rdzeni procesora. Następnie, na podstawie obliczonych odległości stworzone zostały dwie kategorie zmiennych: zmienne bazujące na minimalnej odległości oraz zmienne, które odzwierciedlają liczbę obiektów (każdego typu) występujących w obrębie 800 metrów (co średnio równie się 10 minutom spaceru). W przypadku rzadkich obiektów, takich jak np. aeroporty (n=2) i więzienia (n=4), oraz dróg, obliczenia dotyczyły wyłącznie minimalnych odległości. W połączeniu z danymi pochodzącyimi z otodom.pl oraz zmiennymi wyodrębnionymi w procesie inżynierii cech, utworzono zestaw danych obejmujący 131 zmienną i 10091 obserwacji.

3.3. Eksploracyjna analiza danych

Proces eksploracyjnej analizy zgromadzonego zbioru danych został rozdzielony na dwie części. Pierwsza z nich odnosi się do atrybutów numerycznych, podczas gdy druga dotyczy zmiennych kategorialnych.

3.3.1. Zmienne numeryczne

Prezentowany zbiór danych obejmuje 82 zmienne numeryczne. W ramach początkowej fazy poznano charakterystyki każdej ze zmiennych za pomocą wykresów pudełkowych oraz najważniejszych statystyk opisowych (średnia, mediana, skośność, itp.). W tym kontekście szczególną uwagę poświęcono zmiennym *area* (powierzchni mieszkania) oraz *build_year*. Dla zmiennej *area* usunięto obserwację odstającą o wartości 17000 m^2 , natomiast w przypadku roku budowy dla błędnie przypisanych lat mniejszych od 1560 roku dodano odpowiednią wartość korekcyjną równą 1900. Co więcej, w przypadku większości zmiennych numerycznych (w tym dla zmiennej objaśnianej *price_per_m*) dokonano transformacji logarytmicznej, aby zbliżyć rozkłady do postaci rozkładu normalnego. Po usunięciu wartości odstających analizie uległy gęstości jedno-

⁸⁷ Vincenty T., *Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations*, Geodetic Survey Squadron, London 1975.

i dwuwymiarowe każdej ze zmiennych oraz korelacje, mierzone za pomocą współczynnika Spearmana, definiowanym jako:

$$\rho = \frac{\text{cov}(Rx_1, Rx_2)}{\sqrt{\text{var}(Rx_1) \cdot \text{var}(Rx_2)}} \quad (20)$$

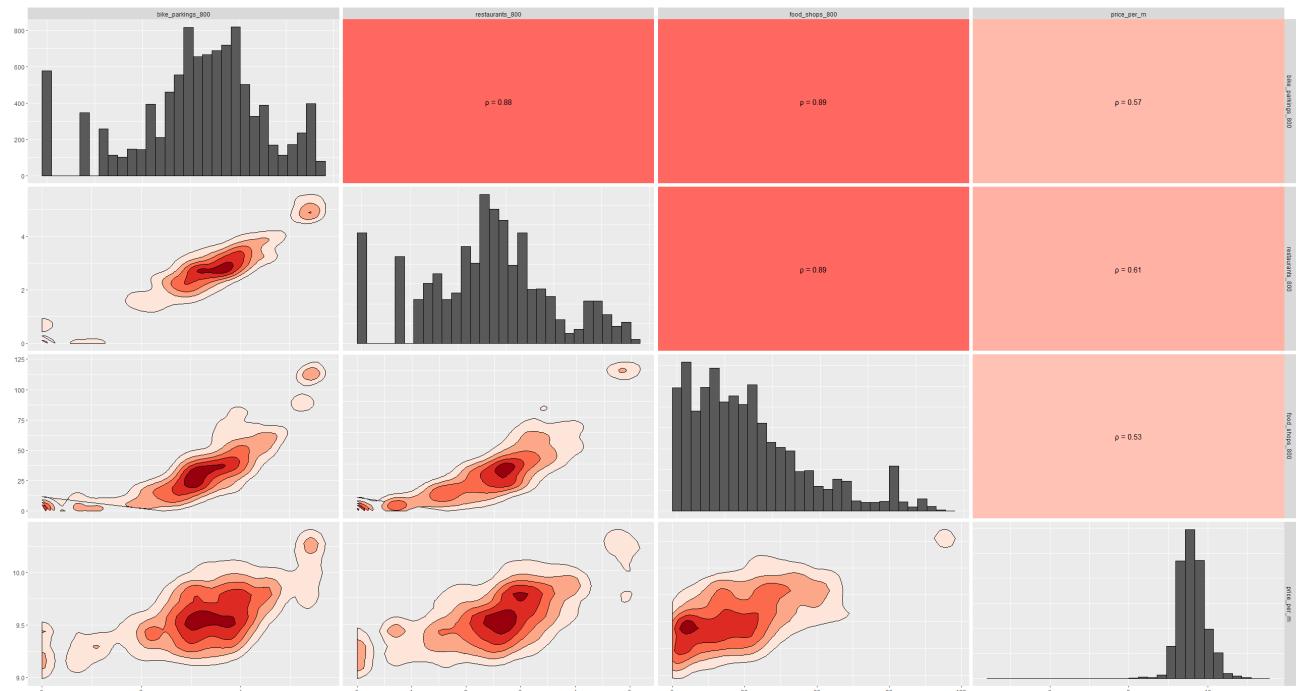
gdzie:

- Rx_1, Rx_2 – rangi (pozycje posortowanych wartości) zmiennej x_1 oraz x_2 ,
- cov, var – kowariancja i wariancje zmiennych odpowiednio.

Współczynnik korelacji rang Spearmana mieści się w granicach od -1 do 1 , gdyż wartość powyżej 0.8 zwykle przyjmuje się jako próg oznaczający mocną korelację zmiennych w przypadku, gdy są one liniowo powiązane.

W odniesieniu do zmiennych *bike_parkings_800*, *restaurants_800* oraz *food_shops_800*, współczynnik korelacji Spearmana jednoznacznie wskazuje na je silne powiązanie. Średnia wartość ρ , wynosząca 0.886 , stanowi przekonujący argument na rzecz tej asocjacji. Równocześnie, analiza dwuwymiarowych wykresów gęstości, zobrazowanych za pomocą wykresów konturowych, dostarcza widocznych dowodów na (prawie) liniową zależność między analizowanymi zmiennymi.

Rysunek 5. Rozkłady zmiennych *bike_parkings_800*, *restaurants_800*, *food_shops_800*.

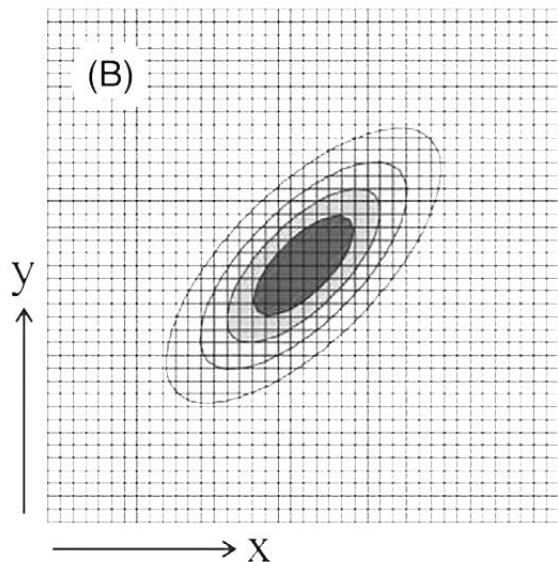


Źródło: opracowanie własne

Warto w tym miejscu wprowadzić interpretację wykresów konturowych. W sytuacji dokładnej zależności liniowej wykres konturowy powinien przedstawać postać dwuwymiarowego wykresu normalnego, nachylonego pod kątem 45 stopni i rozciągniętego od lewego dolnego do prawego

górnego rogu (w przypadku korelacji dodatniej) lub od lewego górnego do prawego dolnego rogu (w przypadku korelacji ujemnej). Wizualnie, normalność na takim wykresie wyraża się poprzez istnienie jednego centralnego punktu (piku) o największej gęstości (czyli najwyższym wzniesieniu na wykresie), oraz równomierny spadek gęstości w kierunku ogonów. W takim przypadku wykres konturowy przybrałby poniższą formę:

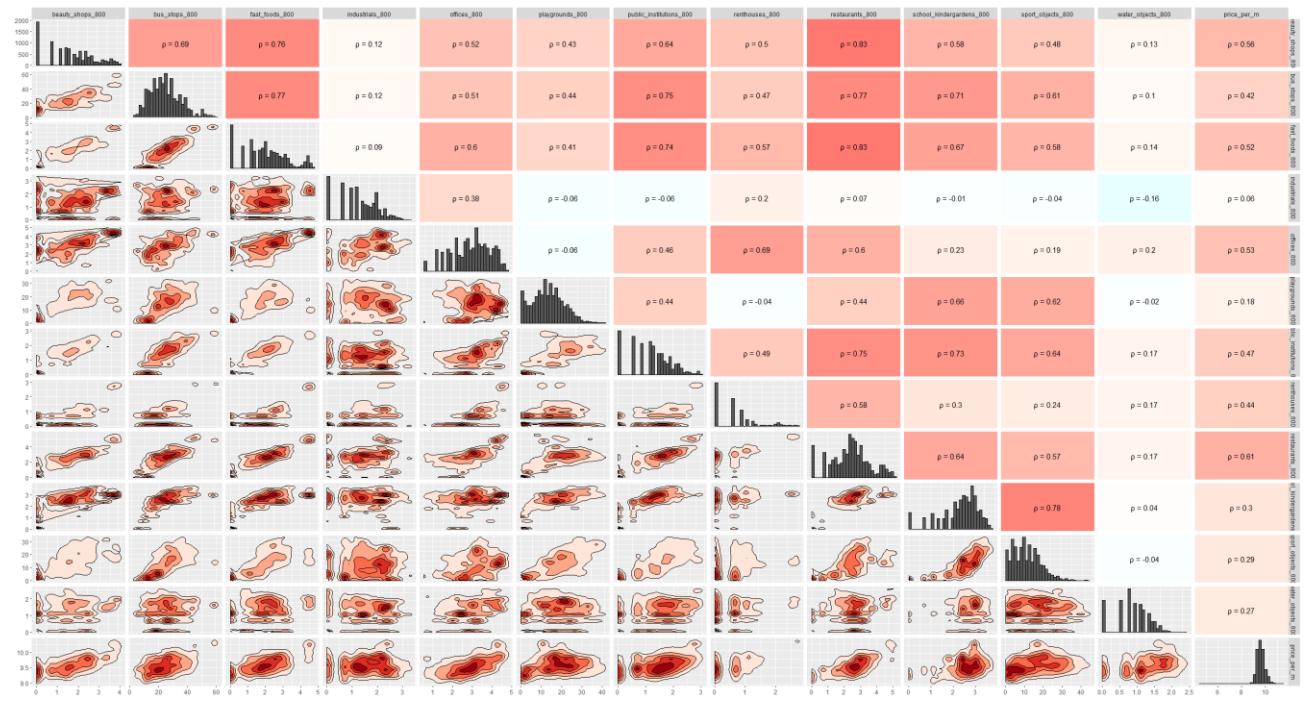
Rysunek 6. Wykres konturowy w przypadku zależności liniowej



Źródło: Paolino P.J., Teaching linear correlation using contour plots, *Teaching Statistics* Wyd. 43 Nr. 1, 2021, s. 16.

Analizując Rysunek 5, można dotrzeć, że dwuwymiarowe wykresy gęstości wykazują pewne podobieństwo do przedstawionego na Rysunku 6 wykresu konturowego, poza kilkoma obszarami gęstości „odstojącej”. Na podstawie tej obserwacji, podjęto decyzję o wyłączeniu z analizy zmiennych *bike_parkings_800* oraz *food_shops_800*. Taka sama zasada stosowana również w kontekście zmiennych *services_800* oraz *attractions_800*. Wykazane relacje między tymi zmiennymi sugerują, że liczba miejsc parkingowych, sklepów spożywczych, usług, atrakcji oraz restauracji łącznie opisują te same obszary geograficzne w obrębie Warszawy – prawdopodobnie centrum lub inne silnie rozwinięte regiony. Analogicznie postąpiono również w przypadku zmiennych *pharmacies_800* i *shops_800*. Końcowy wykres, prezentujący gęstości jedno- i dwuwymiarowe oraz współczynniki korelacji rang Spearmana prezentuje się w następujący sposób:

Rysunek 7. Rozkłady zmiennych liczące obiekty w pobliżu 800 metrów

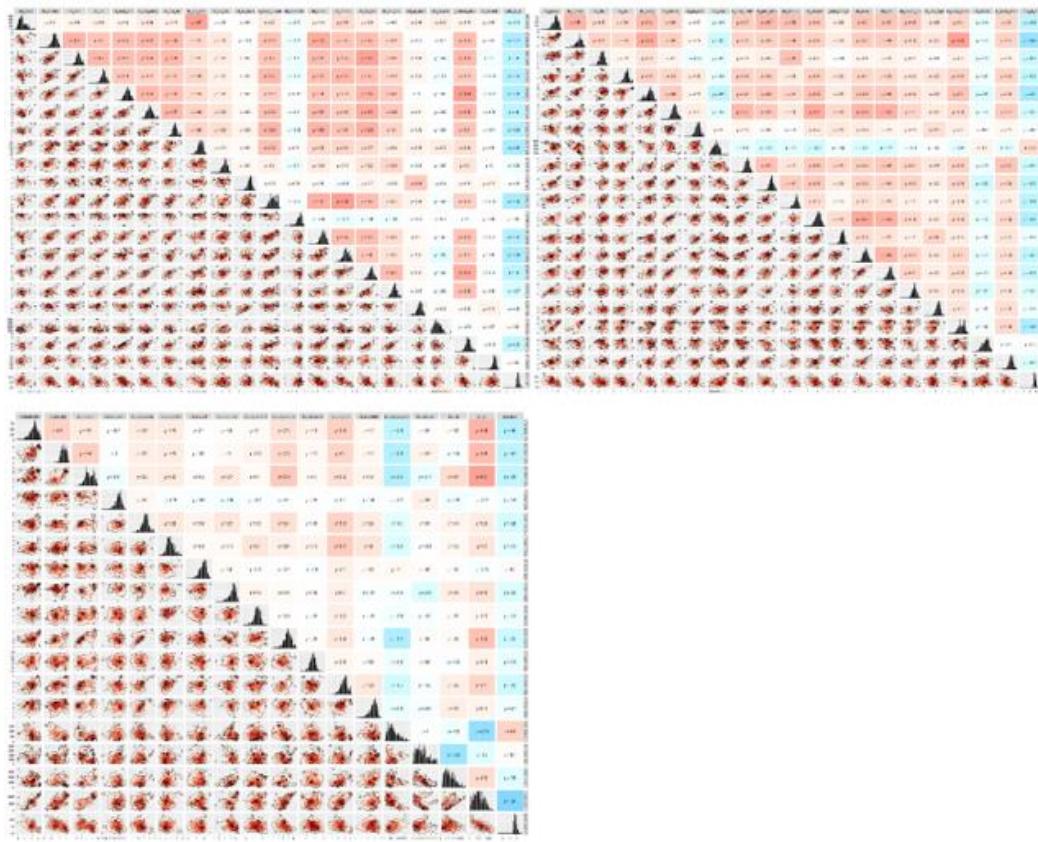


Źródło: opracowanie własne

W przypadku zmiennych *beauty_shops_800* oraz *fast_food_800*, pomimo wysokiego współczynnika korelacji Spearmana ze zmienną *restaurants_800*, nie można jednak stwierdzić, że istnieje między nimi relacja liniowa. Wynika to z nienormalnego rozkładu dwuwymiarowego. Dlatego zdecydowano się zachować te zmienne w analizowanym zbiorze. Co więcej, po dokładnym przyjrzeniu się gęstościom dwuwymiarowym, można zauważyc, że jedynie zmienne *bus_stops_800* i *offices_800* wykazują liniowe powiązania ze zmienną objaśnianą *price_per_m*. Najbardziej złożony wzorzec relacji obserwuje się w przypadku zmiennej *industrial_800*, gdzie występują cztery piki na wykresie gęstości dwuwymiarowej, co przekłada się na niską wartość współczynnika korelacji rang Spearmana.

Kolejna grupa zmiennych koncentruje się na odległościach minimalnych pomiędzy kategoriami obiektów ma mapie a mieszkaniami. W analizie tej grupy nie zaobserwowano wyraźnej współzależności między zmiennymi objaśniającymi, z wyjątkiem zmiennych mierzących odległość od najbliższego klubu nocnego oraz odległość od centrum biznesowego miasta (Pałacu Kultury i Nauki, *dist_cbd*). Z racji na większą korelację zmiennej *dist_cbd* ze zmienną objaśnianą *price_per_m*, zdecydowano się na usunięcie zmiennej *dist_nightclub*.

Rysunek 8. Rozkłady zmiennych liczące odległości minimalne



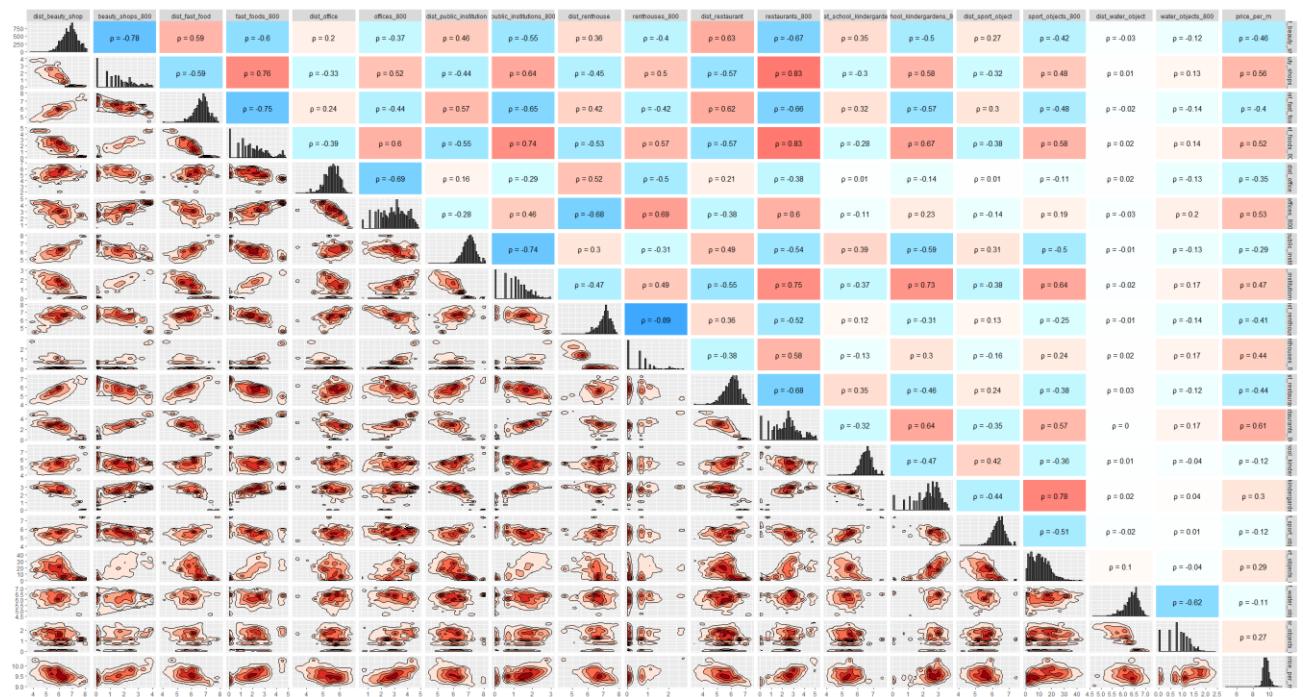
Źródło: opracowanie własne

W kontekście zmiennych pomiaru minimalnych odległości zaobserwowano liniową zależność między nimi a zmienną wyjaśnianą. Szczególnie wyraźnie widoczna była ta zależność w przypadku zmiennych: *dist_attraction*, *dist_bar*, *dist_bike_parking*, *dist_bike_rent*, *dist_fast_food*, *dist_temple_catholic*, *dist_shop*, *dist_restaurant*, *dist_office*, *dist_parking*, *dist_park*, *dist_renthouse*, *dist_public_institution*, *dist_cbd*. Natomiast brak wyraźnej zależności miał miejsce w przypadku charakterystyk: *dist_industrial*, *dist_graveyard*, *dist_gas_station*, *dist_swamps*, *dist_road_path*, *dist_bus_stop*, *dist_playgrounds*. Na wykresie konturowym brak zależności był reprezentowany jako kształt koła. Brak wpływu odległości od obiektów industrialnych, cmentarzy i bagien na cenę mieszkania może być związany z dość małą liczbą tych obiektów w przedziałach miasta Warszawy lub małą liczbą mieszkań w okolicach tych obiektów (przynajmniej w reprezentowanym zbiorze). Z kolei brak korelacji ceny mieszkania z odlegością do przystanków autobusowych, placów zabaw oraz ścieżek może wychodzić z równomiernego dostępu do tych obiektów w przedziałach Warszawy. Natomiast najbardziej złożone struktury zależności występowały w przypadku zmiennych *dist_car_service* oraz *dist_cultural*. Dlatego nawet przy niskich wartościach współczynnika korelacji rang Spearmana, zdecydowano się na pozostawienie tych zmiennych w dalszej w analizie. Dla wszystkich zmiennych w kategorii odległości minimalnych, z wyjątkiem *dist_prison* oraz

dist_road_track_grade, zauważono, że im większa odległość, tym niższa jest cena. To sugeruje, że tylko zmienne *dist_prison* i *dist_road_track_grade* opisują obszary o mniejszej atrakcyjności cenowej.

W kolejnym kroku przeprowadzono analizę współliniowości zmiennych, zarówno tych mierzących odległości minimalne, jak i liczbę obiektów. Zaobserwowano, że istnieje silny współczynnik korelacji rang Spearmana jedynie w przypadku obiektów postojowych (*renthouses*), gdzie ρ wynosi $-0,89$. Niemniej jednak, ze względu na brak związku liniowego między zmiennymi, obie charakterystyki zostały zachowane.

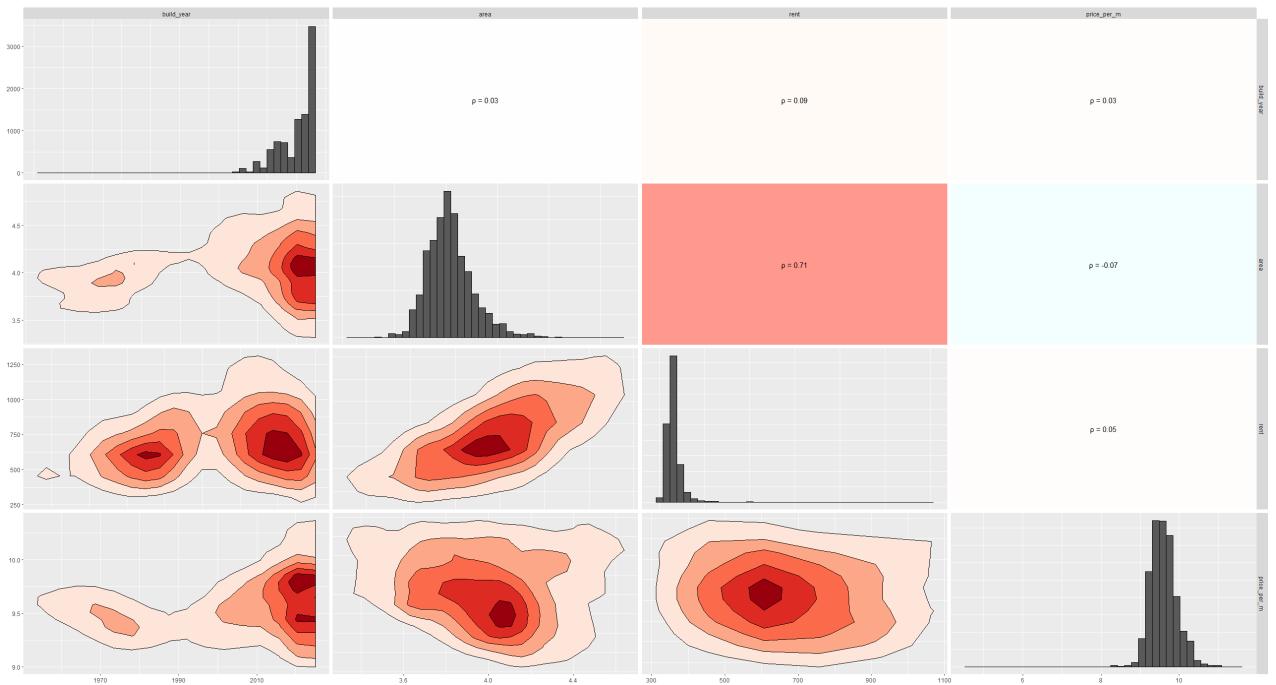
Rysunek 9. Rozkłady zmiennych odległości minimalnych oraz obiektów w pobliżu 800 metrów



Źródło: opracowanie własne

Ostatnie numeryczne zmienne zawarte w zestawie dotyczyły cech aprzestrzennych mieszkań, takich jak dodatkowa opłata za zakup (*rent*), powierzchnia (*area*) oraz rok budowy (*build_year*). Niemniej jednak, ze względu na niski współczynnik korelacji Spearmana (poniżej ustalonego progu odcięcia $\rho = |0.1|$) oraz obserwowanej normalności rozkładów dwuwymiarowych, zdecydowano się na wykluczenie tych zmiennych ze zbioru danych.

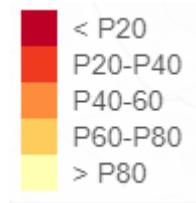
Rysunek 10. Rozkłady zmiennych area, rent, build_year



Źródło: opracowanie własne

Ostatni etap analizy zmiennych numerycznych dotyczył rozkładu przestrzennych zmiennych. Zwizualizowano je za pomocą pięciu percentylów o wartościach 0, 20, 40, 60 i 80 procent. Im wartość zbliża się do maksymalnej, tym bardziej czerwona jest obserwacja (punkt) na mapie Warszawy.

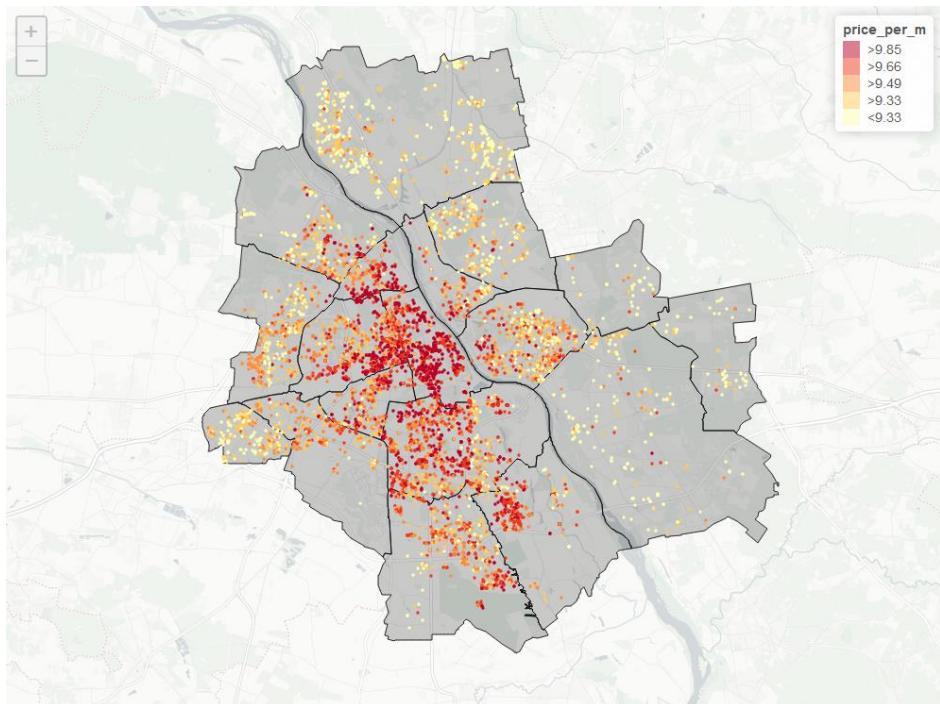
Rysunek 11. Kolory odpowiednich percentylów



Źródło: opracowanie własne

Dzięki temu największe liczby obiektów i największe odległości są oznaczone intensywnym kolorem czerwonym.

Rysunek 12. Rozkład ceny metra kwadratowego mieszkania po transformacji logarytmicznej

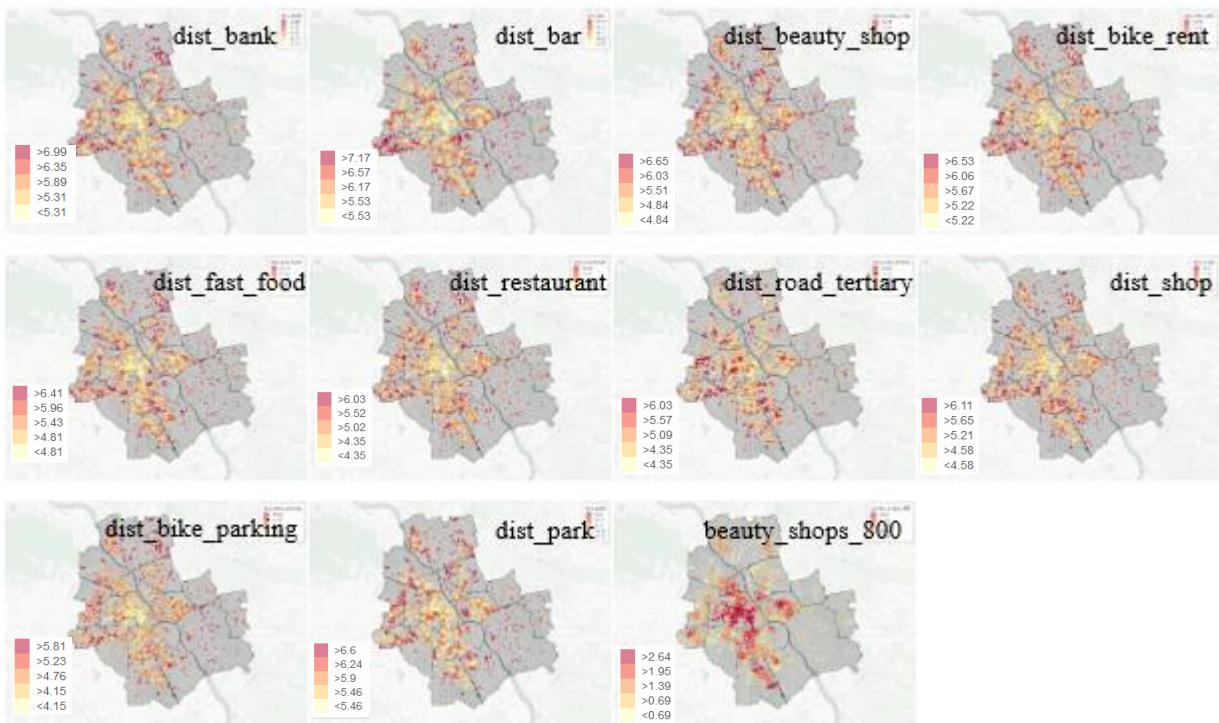


Źródło: opracowanie własne

Analiza zaczyna się od ceny za metr kwadratowy mieszkań. Widać wyraźnie, że lewy brzeg Wisły ma wyższe ceny za metr kwadratowy w porównaniu do brzegu prawego. Dodatkowo najwyższe ceny dotyczą dzielnic takich jak Śródmieście, Wola, Żoliborz oraz zachodnia część Mokotowa. Natomiast najniższe wartości obserwuje się głównie na terenach Białołeki (północ) oraz wschodnich dzielnicach Warszawy (Wawer, Wesoła, Rembertów).

Ponadto mniejsze odległości od banków, barów, restauracji, dróg rowerowych i wypożyczalni rowerów, dróg trzeciorzędnych oraz większa ilość sklepów kosmetycznych, charakteryzują dzielnice centralne takie jak Śródmieście, Wola, Żoliborz oraz Ochota. Podobne cechy można również zaobserwować w rozwiniętych częściach Pragi Południowej (zwłaszcza w okolicach ul. Francuskiej), Mokotowa (głównie wzdłuż alei Niepodległości) oraz Ursynowa (szczególnie wzdłuż alei Komisji Edukacji Narodowej).

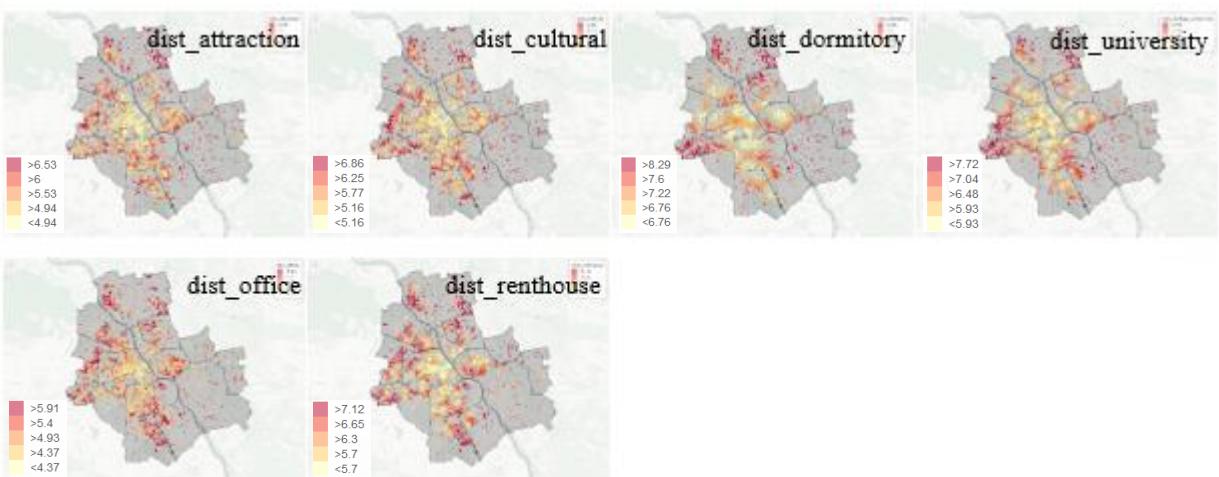
Rysunek 13. Rozkład zmiennych charakteryzujących dzielnice centralne



Źródło: opracowanie własne

Odległość od atrakcji, uniwersytetów, obiektów kulturowych, biur, hoteli oraz akademików była mniejsza przede wszystkim w centrum miasta, obejmując dzielnice takie jak Śródmieście, Ochota, Wola oraz pobliskie rejony sąsiadujących dzielnic. Niemniej jednak, z perspektywy przestrzennej cechy te odnoszą się do różnych obszarów geograficznych. Co więcej, w porównaniu ze zmiennymi przedstawionymi na Rysunku 12, koncentracja w Śródmieściu była znacznie większa.

Rysunek 14. Rozkład zmiennych charakteryzujących Śródmieście

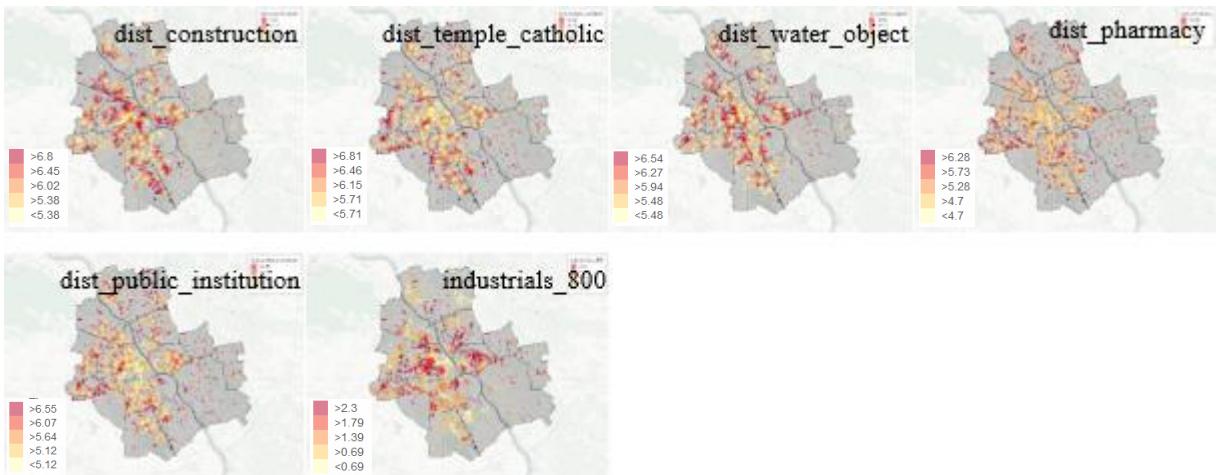


Źródło: opracowanie własne

Odległość od budów i rekonstrukcji, świątyń katolickich, aptek, instytucji publicznych, obiektów wodnych oraz gęstość obiektów industrialnych opisują mniejsze klastry przestrzenne,

zauważane na poziomie lokalnym w każdej (lub większości) dzielnic, chociaż największa ilość takich klastrów zwykle przypada na centrum Warszawy.

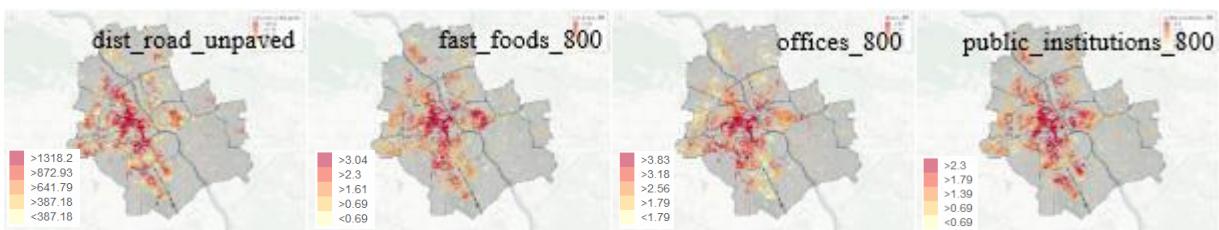
Rysunek 15. Rozkład zmiennych wyróżniających lokalne klastry przestrzenne



Źródło: opracowanie własne

W odróżnieniu od większości analizowanych zmiennych, bliskość nieutwardzonych dróg, niższa gęstość instytucji publicznych, biur oraz restauracji typu fast food zwykle idą w parze z obszarami mniej rozwiniętymi, mniej zaludnionymi lub oddalonymi od centrum i głównych obszarów działalności.

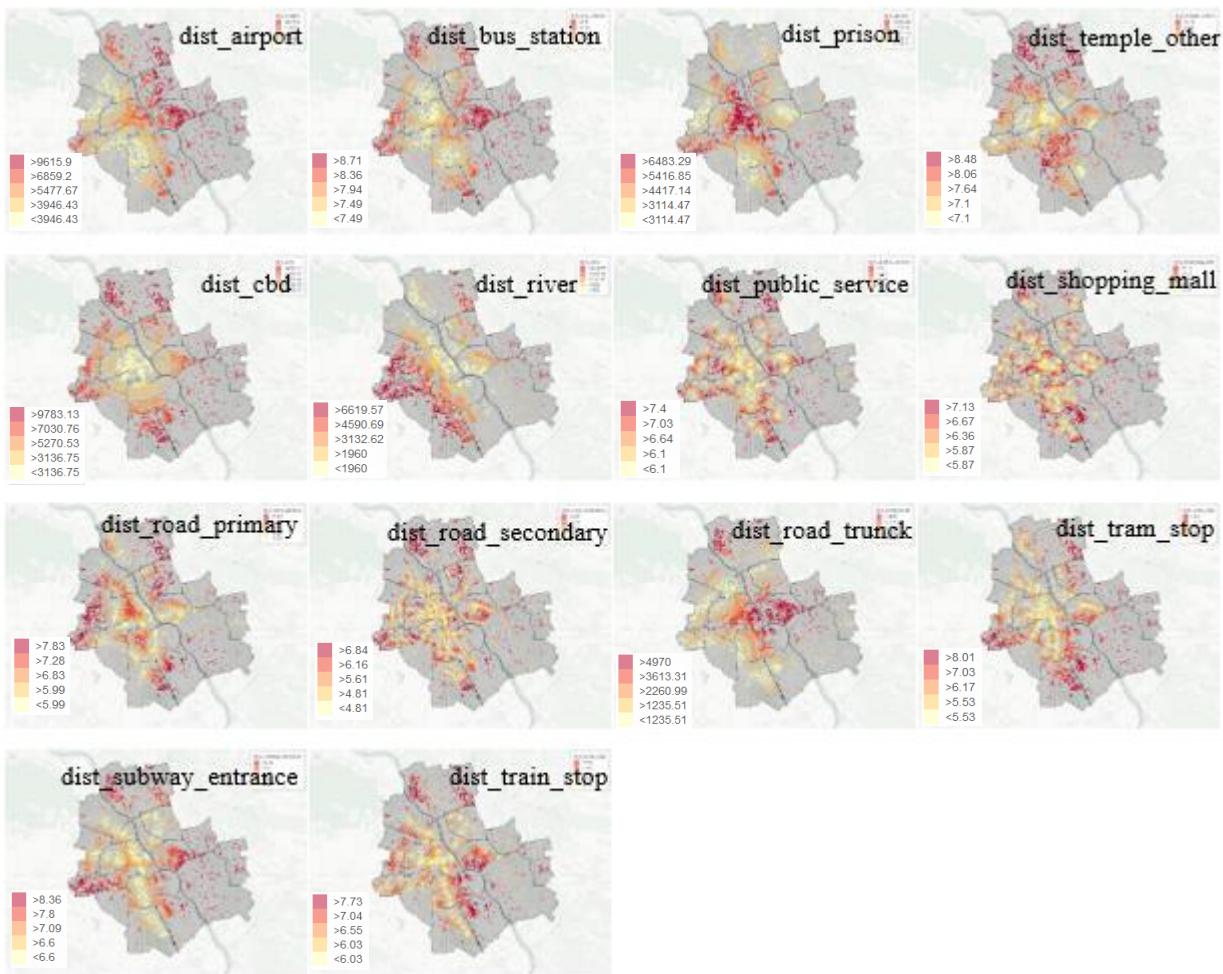
Rysunek 16. Rozkłady zmiennych charakteryzujących obszary mniej rozwinięte



Źródło: opracowanie własne

Mocny wzorzec ulokowania przestrzennego jest szczególnie widoczny w przypadku rzadko występujących obiektów, które są opisane przez zmienne mierzące minimalne odległości od lotnisk, dworców autobusowych, więzień, departamentów straży pożarnej oraz komisariatów policji, centrów handlowych, świątyń niekatolickich oraz PKiN (Pałacu Kultury i Nauki). Do tej samej grupy zaliczają się zmienne mierzące odległość od dróg pierwszo- i drugorzędnych, autostrad, przystanków metra, pociągów i tramwajów oraz rzeki Wisły.

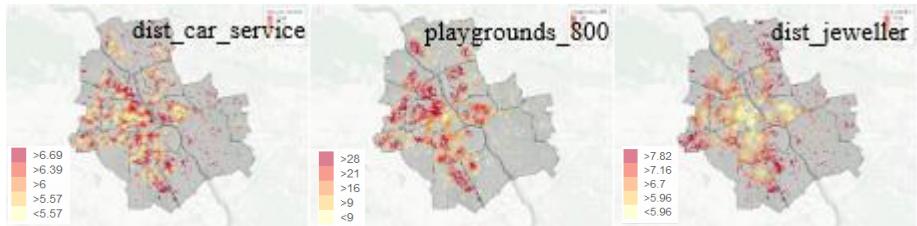
Rysunek 17. Rozkłady zmiennych z najbardziej klarownym wzorcem przestrzennym



Źródło: opracowanie własne

Rozkłady przestrzenne zmiennych *dist_car_service* (umieszczone poza głównymi drogami miasta, ale w niedużej odległości od nich), *playgrounds_800* (rzadziej występujące w centrum oraz w odległych dzielnicach) oraz *dist_jeweller* (oprócz Białołęki, ta zmienna opisuje najbardziej prestiżowe obszary Warszawy) wyodrębniają unikalne wzory przestrzenne nie zauważone w trakcie analizy innych cech.

Rysunek 18. Rozkłady zmiennych z unikalnymi wzorcami przestrzennymi

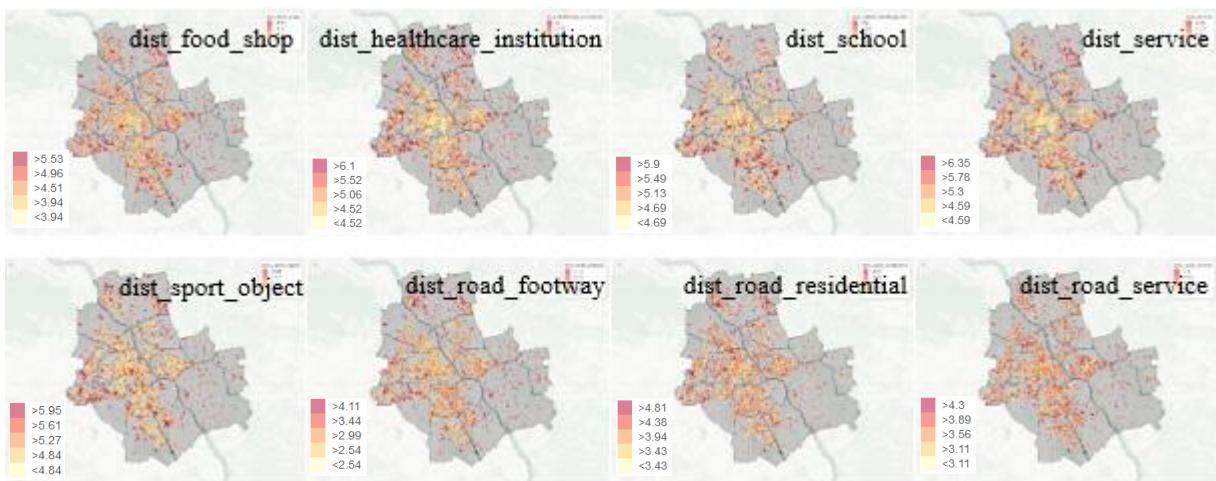


Źródło: opracowanie własne

Najbardziej równomierne rozmieszczenie przestrzenne jest obserwowane w przypadku odległości od sklepów spożywczych, instytucji zdrowotnych, szkół i przedszkoli, obiektów

sportowych, dróg serwisowych oraz chodników. Choć to nie oznacza, że te zmienne są statystycznie nieistotne, jednak zauważona jednostajność geograficzna może oznaczać, że korelacja ze zmienną objaśnianą może być wynikiem innych czynników (zmienne ukryte) lub prostych zależności liniowych (zero jedynkowych).

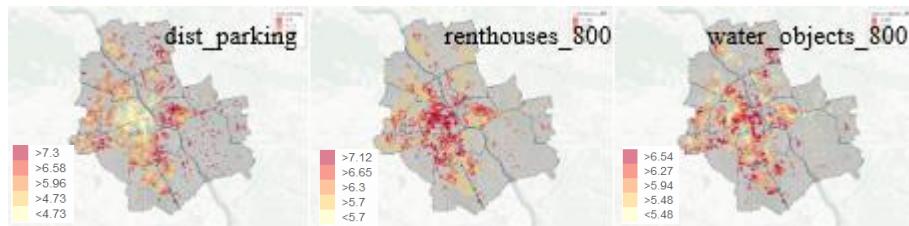
Rysunek 19. Rozkłady zmiennych bez zależności geograficznych



Źródło: opracowanie własne

Ostatnia grupa zmiennych odnosi się do charakterystyk, które zostały w końcu wykluczone z analizowanego zbioru. Zmienna *dist_parking* zdaje się być obarczona błędem pomiaru, ponieważ niska odległość od najbliższego miejsca parkingowego występowała prawie jedynie w centrum miasta. Zmienna *renthouses_800* wydaje się mało informacyjną ze względu na niewielką rozbieżność pomiędzy 20 a 40 percentylem. Z kolei w przypadku *water_objects_800*, jest ona silnie skorelowana ze zmienną *dist_water_object*; ta zależność nie została dostrzeżona podczas analizy gęstości dwuwymiarowych, ponieważ rzeka Wisła była liczona jako jeden obiekt wodny.

Rysunek 20. Rozkłady zmiennych wyeliminowanych ze zbioru



Źródło: opracowanie własne

Na tym kończy się etap analizy eksploracyjnej zmiennych numerycznych. Liczba pozostawionych w zbiorze zmiennych numerycznych wynosi 58. Jest to produkt skrupulatnej selekcji, którego rezultatem jest zestaw najbardziej potencjalnie istotnych (z punktu widzenia statystycznego) oraz niezależnych jeden od jednego numerycznych charakterystyk przestrzennych eksplorowanych mieszkań.

3.3.2. Zmienne kategorialne

W odróżnieniu od zmiennych nominalnych, siła statystycznej asocjacji pomiędzy regresorami kategorialnymi będzie oceniana za pomocą statystyki tau Kendall'a. Aby obliczyć τ Kendall'a, należy zestawić obserwacje ze zbioru we wszystkie możliwe pary (dwóch zmiennych kategorialnych), a następnie podzielić te pary na grupy P i Q . Wtedy równanie jest podane wzorem:

$$\tau = 2 \frac{P-Q}{N(N-1)}, \quad (21)$$

gdzie:

- P – pary zgodne, tzn. porównywane zmienne kategorialne w obrębie tych obserwacji zmieniają się w tę samą stronę.
- Q – pary niezgodne, tzn. pary obserwacji, dla których porównywane charakterystyki zmieniają się w przeciwną stronę.
- N – liczba obserwacji w zbiorze.

W niniejszej pracy przyjęto, że w przypadku, gdy $\tau > 0,8$, zmienne są uznane za mocno skorelowane i należy je usunąć. Natomiast siła asocjacji pomiędzy zmienną objaśnianą a kategorialnymi zmiennymi regresyjnymi będzie ustalona za pomocą testu statystycznej równości wartości oczekiwanych w dwóch populacjach, znanej jako test t Welcha. W odróżnieniu od klasycznego testu t-Studenta test Welcha uwzględnia zróżnicowane wariancje w każdej z podprób. Matematycznie jest opisany w następujący sposób:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \quad (22)$$

gdzie:

- \bar{x}_1, \bar{x}_2 – średnie zmiennej objaśnianej w obu podpróbach,
- s_1^2, s_2^2 – wariancje zmiennej objaśnianej w obu podpróbach,
- N_1, N_2 – liczba obserwacji w każdej podpróbie.

Statystyka testu t Welcha będzie porównywana z wartością krytyczną przy określonych stopniach swobody. W rezultacie istotność testu będzie wyrażona za pomocą wartości p. W niniejszej pracy przyjęto, że jeśli wartość p jest mniejsza niż 0,05, to hipoteza zerowa, która zakłada równość wariancji w obu podpróbach, zostanie odrzucona. W przypadku, gdy liczba kategorii dla zmiennej dyskretnej przekracza dwie, zastosowano analizę wariancji jednoczynnikową (ang. *one-way analysis of variance*, *one-way ANOVA*). W tym teście hipoteza zerowa zakłada równość średnich wartości dla każdej z podprób $i = 1, \dots, r$. Statystyka testowa jest obliczana przy użyciu wzoru:

$$F = \frac{\frac{1}{r-1} \sum_{i=1}^r N_i (\bar{x}_i - \hat{x})^2}{\frac{1}{N-r} \sum_{i=1}^r \sum_{j=1}^N (x_{ij} - \bar{x}_i)^2}, \quad (23)$$

gdzie:

- \bar{x}_i – średnia arytmetyczna z i-tej podpróby,
- \hat{x} – średnia arytmetyczna ze wszystkich N obserwacji w zbiorze,
- N_i – liczba obserwacji w i-tej podpróbce.

Wyznaczana statystyka F ma rozkład F-Snedecora ze znanyymi wartościami krytycznymi, dlatego podobnie jak w teście t Welcha, możliwe jest zdefiniowanie statystycznej istotności wniosków z analizy wariancji za pomocą wartości p.

Po określaniu używanych testów możliwe jest przedstawienie wyników eksploracyjnej analizy wykorzystanych zmiennych kategorialnych. Ze względu na dużą liczbę badanych zmiennych, zostały one początkowo podzielone na dwie grupy.

Rysunek 21. Rozkłady i testy statystyczne dla pierwszej grupy zmiennych kategorialnych

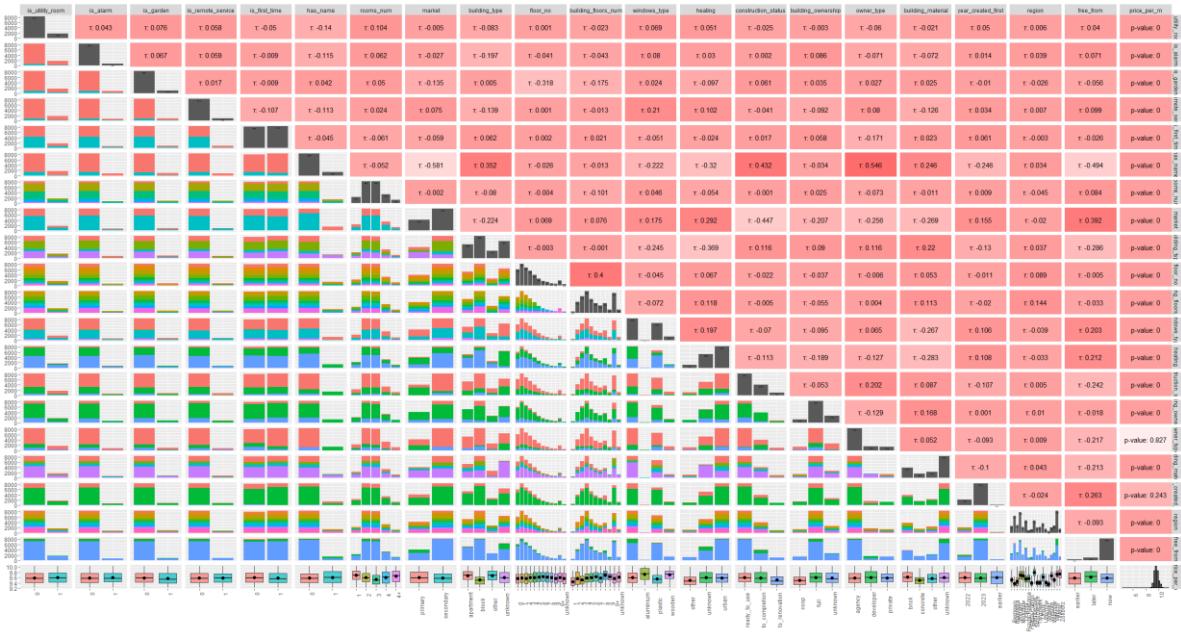


Źródło: opracowanie własne

Wartości p uzyskane z serii testów t Welcha wskazują, że wszystkie zmienne przedstawione w pierwszej grupie stanowią istotny statystycznie czynnik różnicujący średnie wartości zmiennej objaśnianej. Jednakże wartość τ Kendalla przekroczyła ustalony próg dla zmiennych *is_freezer*, *is_oven*, *is_dishwasher* oraz zmiennych *is_kable_tv* i *is_internet*. W związku z tym utworzono dwie nowe zmienne zamiast cech powiązanych: *is_kitchen_furnished*, przyjmującą wartość 1 jeśli

mieszkanie zawiera jednocześnie piekarnik, zmywarkę oraz lodówkę, oraz zmienną *is_media*, która równi się 1 w przypadku, gdy zarówno telewizja kablowa, jak i Internet zostały podłączone.

Rysunek 22. Rozkłady i testy statystyczne dla drugiej grupy zmiennych kategorialnych



Źródło: opracowanie własne

W odniesieniu do drugiej kategorii nie wykryto żadnych statystycznych asocjacji pomiędzy regresorami. Niemniej jednak stwierdzono brak istotnych różnic w średnich wartościach mieszkań dla różnych typów właścicieli (deweloper, agencja nieruchomości lub właściciel prywatny, wartość $p = 0,827$). To samo dotyczyło zmiennej *year_created_first* ($p = 0,243$). Z uwagi na powyższe, wyeliminowano te zmienne z ostatecznego zbioru. Dodatkowo ze względu na to, że analiza wariancji nie pozwala na identyfikację konkretnych podprób, w których nie zaobserwowano statystycznie istotnej zmienności średnich wartości zmiennej objaśnianej, zastosowano grupowanie kategorii w następujących regresorach na podstawie wykresów pudełkowych:

- *building_type*: poziomy „unknown” + „other”,
- *heating*: „unknown” + „urban”,
- *construction_status*: „ready_to_use” + „to_completion”,
- *building_ownership*: „full” + „unknown”,
- *building_material*: „other” + „unknown” + „brick”.

Pozwoli to skupić się na najbardziej charakterystycznych cechach mieszkań. Poniżej przedstawiono wykresy pudełkowe, wyniki testów statystycznych oraz histogramy zmiennych, dla których dokonano redukcji kategorii.

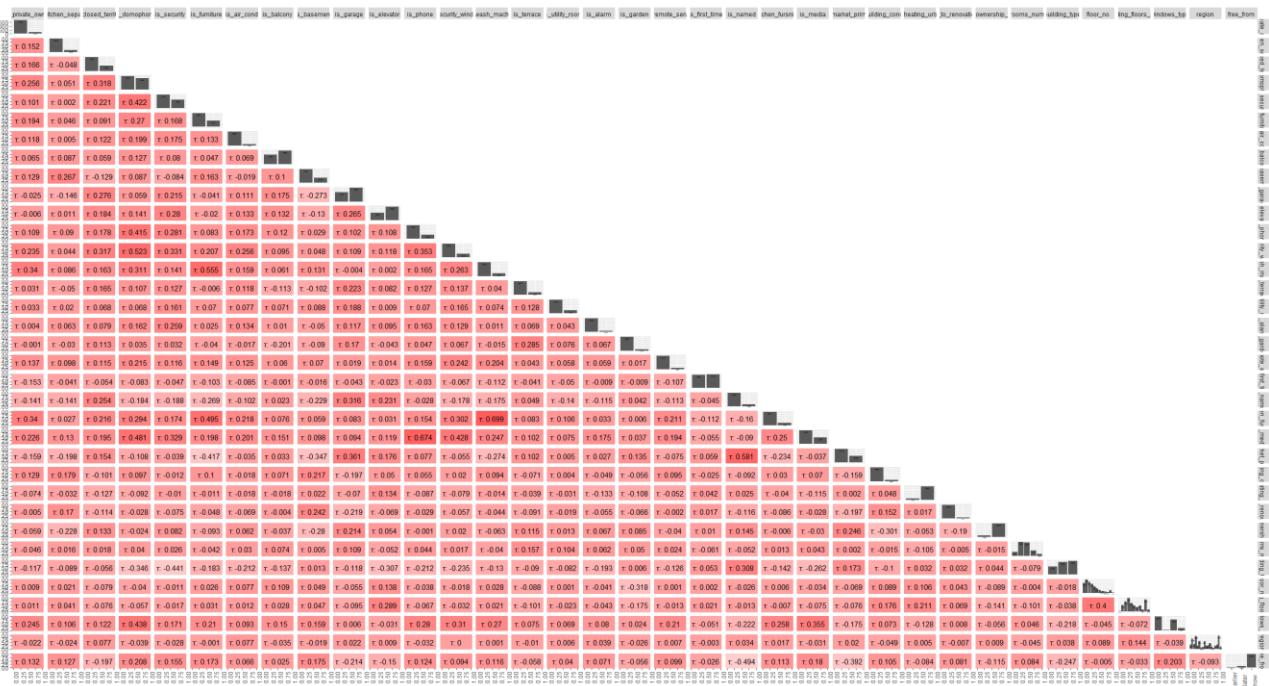
Rysunek 23. Rozkłady i testy statystyczne dla zredukowanych zmiennych



źródło: opracowanie własne

Największa korelacja została zidentyfikowana między budynkami z betonu a typem właścielstwa kooperacyjnego, choć wartość tau pozostawała poniżej ustalonego progu ($\tau = -0,3$).

Rysunek 24. Korelacje wszystkich kategorialnych regresorów

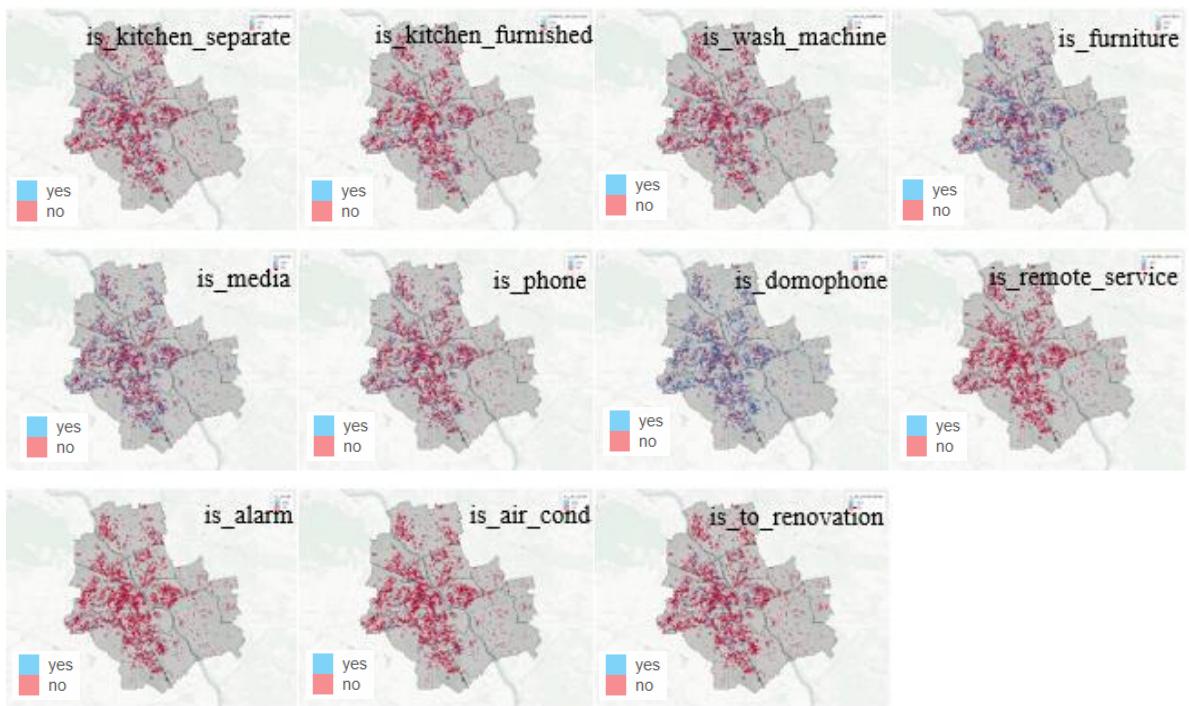


źródło: opracowanie własne

Kolejny wykres przedstawiał analizę korelacji wszystkich zmiennych kategorialnych. Największą statystyczną asocjację zaobserwowano między zmiennymi *is_phone* i *is_media*, co wynika z częstego współwystępowania tych atrybutów jako elementów komplementarnych. Analogicznie, znacząca korelacja wystąpiła również między zmiennymi *is_washing_machine*, *is_kitchen_furnished* oraz *is_furniture*, co wydaje się typową sytuację w przypadku mieszkań na rynku wtórnym. Na dodatek zmienna *is_first_time* wykazywała związek z *is_market_primary*, co jest oczekiwane, ponieważ mieszkania z rynku wtórnego prawdopodobnie nigdy wcześniej nie były umieszczane na stronie internetowej otodom.pl.

Ostatni etap eksploracji zmiennych kategorialnych polegał na przedstawieniu zmiennych na mapie Warszawy. Dla zmiennych binarnych, wartość 1 („tak”) została zaznaczona kolorem niebieskim, podczas gdy wartość 0 („nie”) została zaznaczona kolorem czerwonym. Pierwsza grupa zmiennych binarnych dotyczyła opisu wnętrza analizowanych mieszkań.

Rysunek 25. Zmienne zero-jedynkowe opisujące wnętrze mieszkań

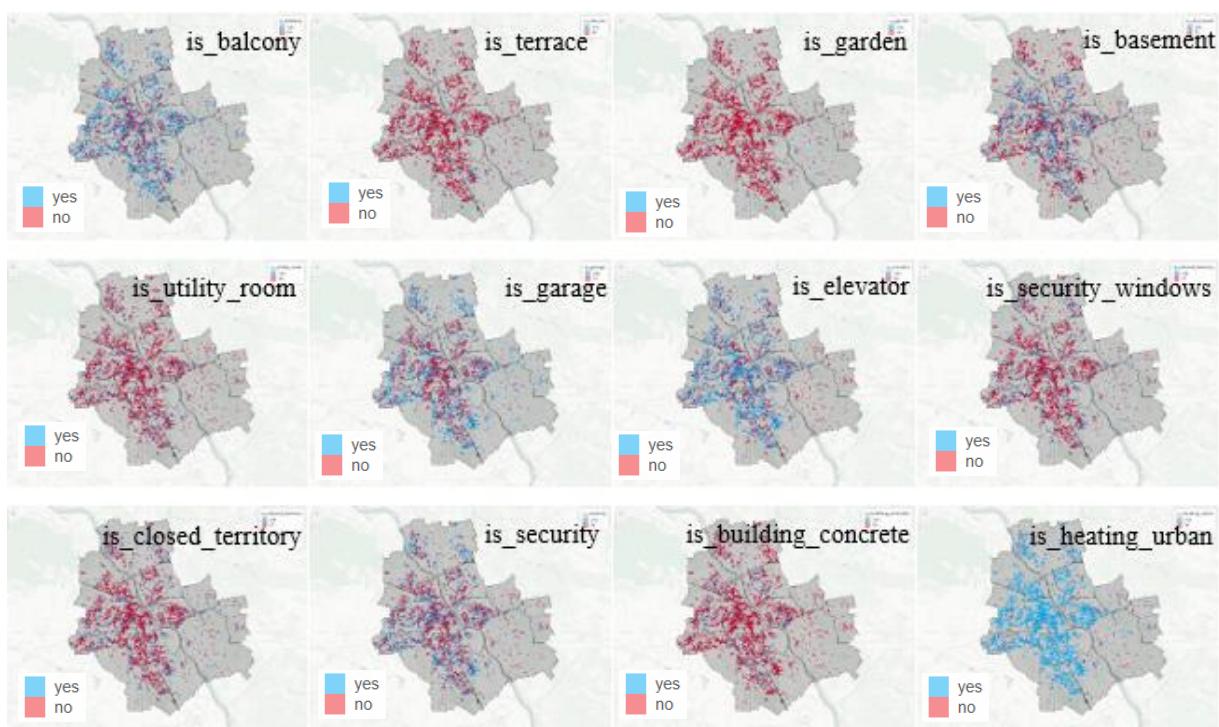


Źródło: opracowanie własne

W większości spostrzeganych mieszkańach kuchnia nie była oddzielona i umeblowana. Nie zaobserwowano również wyraźnych wzorców ulokowania przestrzennego. To samo dotyczyło pralek; wizualnie można uznać, że te zmienne są ściśle powiązane ze sobą. Biorąc pod uwagę również stosunkowo wysoką wartość tau Kendalla, pozbyto się tej zmiennej w ostatecznym zbiorze. Rozkład umeblowania innych pokoi był równomiernie umieszczany na terenie Warszawy; prawdopodobnie

wynika to z faktu, że umeblowanie dotyczyło głównie lokali z rynku wtórnego. Rozkład zmiennej *is_media* był znacznie rozproszony, przypominający bardziej rozkład *is_domophone*. Zdalna obsługa okazała się rzadką zmienną dla wartości 1, co może charakteryzować bardziej prestiżowe mieszkania. Podobnie rzadką zmienną okazały się *is_alarm* oraz *is_air_cond* – większość mieszkań nie była wyposażona w systemy alarmowe i klimatyzacje. W przypadku zmiennej *is_to_renovation* większość mieszkań z takim atrybutem znajdowała się na Żoliborzu, Bielanach oraz w Śródmieściu. To jest jedyna zmienna opisująca wnętrze mieszkań, która wykazuje wyraźny wzorzec przestrzenny.

Rysunek 26. Zmienne zero-jedynkowe opisujące zewnętrze mieszkań

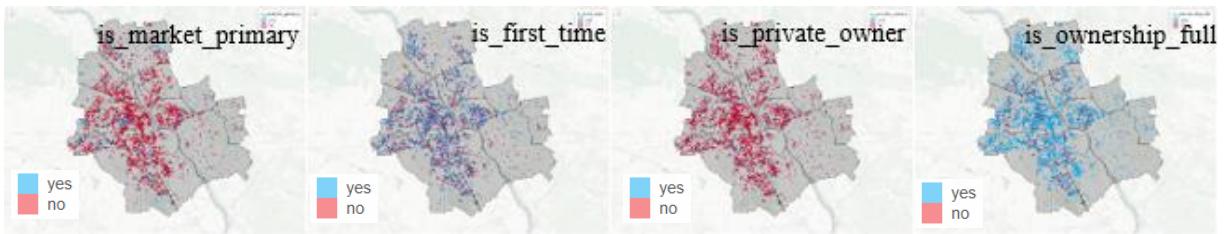


Źródło: opracowanie własne

Zmienna *is_balcony* wykazuje rozkład dość równomierny w większości obszarów, z wyjątkiem Śródmieścia, gdzie liczba balkonów jest stosunkowo niska. Natomiast zmienna *is_terrace* pojawia się głównie w odległych od centrum rejonach i stanowi zmienną relatywnie rzadką. Ogródki prywatne obserwuje się przeważnie na terenie Wawera, jak również i ogrzewanie inne od centralnego miejskiego. Z kolei istnienie piwnicy i budynków betonowych przeważnie dotyczy starych mieszkań oddanych do użytkowania jeszcze w czasach istnienia PRL (Polskiej Republiki Ludowej), kiedy piwnice i beton były standardowym elementem zabudowy. Więcej mieszkań z garażami jest zauważano poza centrum miasta, w regionach bardziej oddalonych. Zmienna *is_security* często pojawia się w obszarach aktywnie rozwijających się, w których ochrona prywatna

staje się obowiązkową częścią infrastruktury mieszkaniowej. Pozostałe zmienne wykazują się równomiernym rozkładem przestrzennym.

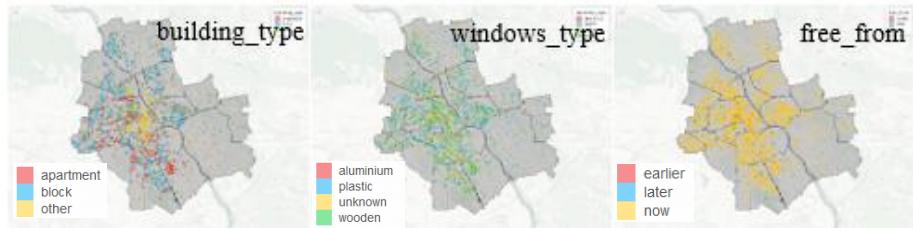
Rysunek 27. Zmienne zero-jedynkowe opisujące czynniki globalne



Źródło: opracowanie własne

W przypadku binarnych zmiennych kategorialnych, opisujących globalną sytuację na rynku mieszkaniowym, zauważono, że zazwyczaj mieszkania są sprzedawane poprzez agencję lub bezpośrednio od developerów, co potwierdzało się niską wartością p dla testów ANOVA. Osoby fizycznie bardzo rzadko występują sprzedawcami mieszkań bezpośrednio. Dodatkowo wykryto, że ofert, które zostały ogłoszone na stornie internetowej otodom.pl po raz drugi jest znacznie więcej niż ofert na rynku wtórnym. Oznacza to, że stosunkowo dużo ofert z rynku pierwotnego też zostały rozmieszczone po raz drugi. Mówiąc o strukturze rynku, większość ofert z rynku pierwotnego znajduje się na peryferii Warszawy, za wyjątkiem aktywnie rozrastającej Woli. Na koniec zauważono, że punkty skupienia kooperatyw przypadają głównie na obszary miejskie powstałe w czasach PRL.

Rysunek 28. Zmienne kategorialne opisujące faktory globalne

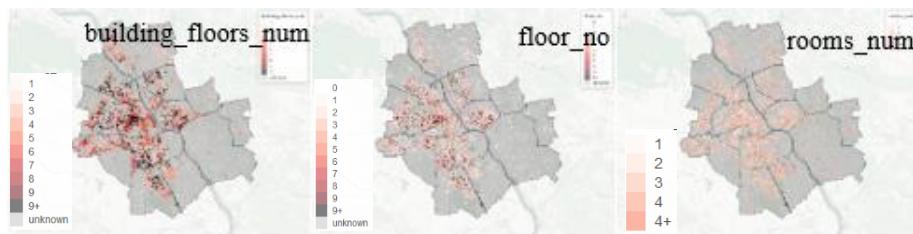


Źródło: opracowanie własne

Można zauważyć, że kategoria zmiennej *building_type* „inne” występuje głównie w centrum miasta, co jest oznaczone kolorem żółtym na mapie (Rysunek 28). Natomiast oferty apartamentów (oznaczonych czerwonym kolorem) występują na Mokotowie, w granicach Śródmieścia, na Żoliborzu oraz w rejonie Wilanowa między Aleją Rzeczypospolitej a ul. Adama Bronickiego. Na dodatek zmienna *windows_type* ilustruje, że okna aluminiowe (czerwone kropki) są mało spotykane, natomiast okna drewniane występują głównie w starych mieszkaniach, ale także w obszarach szybko rozwijających się, takich jak Wola. Z tego powodu, interpretacja tej zmiennej jest niejednoznaczna. Zmienna *free_from*, dotycząca dostępności mieszkań, ma ograniczoną wartość informacyjną.

W większości przypadków mieszkania są dostępne w chwili obecnej (żółte kropki). Dlatego wyłączono tą zmienną z analizy.

Rysunek 29. Zmienne uporządkowane opisujące fakty globalne na mapie Warszawy



Źródło: opracowanie własne

Należy zaznaczyć, że w przypadku zmiennej *building_floors_num* najwyższe budynki (oznaczone czerwonymi punktami) znajdują się głównie w regionach z zabudową pochodzącą z okresu PRL i w centrum miasta. Dla zmiennej *floor_no* dane prezentowane na mapie sugerują, że najczęściej sprzedawanych mieszkań mieści się na parterze. Natomiast w przypadku zmiennej *rooms_num*, można dostrzec, że najwięcej dużych (względem liczby pokoi) mieszkań występuje głównie w Wilanowie, Wawerze oraz zachodniej części Mokotowa.

Na bazie wnikliwej analizy zmiennych objaśniających zidentyfikowano 91 zmiennych o istotności statystycznej (w tym 58 numerycznych, 31 kategorialnych oraz 2 związane ze współrzędnymi geograficznymi). Największe współzależności zmiennych objaśnianych ujawniły się w przypadku cech numerycznych, takich jak liczba restauracji i sklepów kosmetycznych w promieniu 800 metrów oraz odległość od centrum miasta, wyznaczona jako dystans od Pałacu Kultury i Nauki. W kontekście cech kategorialnych analiza wykresów pudełkowych oraz przeprowadzone testy równości średnich dostarczyły przekonujących dowodów na istotność zmiennych takich jak obecność tarasu, potrzeba renowacji oraz rodzaj okien. Ogólnie stwierdzono, że badany zbiór danych niemal w całości pokrywał się z taksonomią zmiennych przedstawioną w przeglądzie literatury. Jedyne odstępstwa pojawiły się w kategorii zmiennych zewnętrznych innych od zmiennych geograficznych, ponieważ ta kategoria występowała rzadko. Niemniej jednak, ze względu na szczegółowy charakter zagadnienia badawczego, zbieranie takich charakterystyk bez konieczności wykorzystania kosztownych metod badawczych wydaje się wyzwaniem.

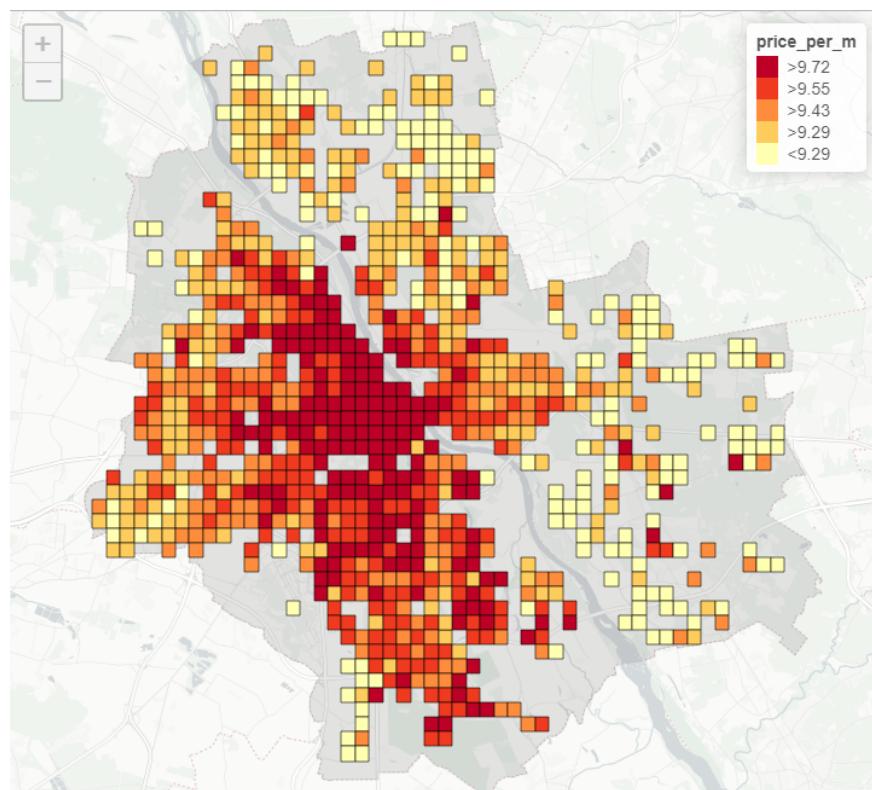
Ostatnia część pracy dotyczy opisu przebiegu testowania autokorelacji przestrzennej oraz przedstawieniu i interpretacji dopasowania użytych modeli.

IV. Analiza empiryczna

4.1. Testowanie autokorelacji przestrzennej

Część analityczna rozpoczyna się od przetestowania występowania autokorelacji przestrzennej, która jest fundamentalnym założeniem stosowania autoregresyjnych modeli przestrzennych. W tym celu skonstruowano (hedoniczny) model liniowy oraz macierz wag. W przypadku modelu regresji liniowej dokonano sekwencyjną eliminację zmiennych, zaczynając od cech z najwyższą wartością p. W wyniku tego model liniowy opierał się na 29 zmiennych objaśniających. Z kolei macierz wag przestrzennych została stworzona na bazie odwrotnych odległości pomiędzy obserwacjami. Jednakże ze względu na dużą liczbę obserwacji i towarzyszące ograniczenia obliczeniowe postanowiono zagregować mieszkania wzdłuż siatki kwadratowej, dopasowanej do mapy Warszawy, gdyż w przypadku oryginalnego zbioru zawierającego 10078 obserwacji, już sama macierz wag przestrzennych zajmowała 1GB pamięci operacyjnej. Skumulowanie danych w poszczególnych komórkach siatki dokonano poprzez użycie średniej (dla zmiennych numerycznych) oraz dominanty (dla zmiennych kategorialnych). W wyniku tych działań uzyskano zbiór obejmujący 983 obserwacje:

Rysunek 30. Rozkład zmiennej objaśnianej po zastosowaniu agregacji



Źródło: opracowanie własne

Przechodząc do samej procedury testowania, poniższa tabela przedstawia opis wszystkich dokonanych testów wykonanych na resztach z modelu regresji liniowej. Model opierał się na 29 zmiennych objaśniających zaklasyfikowanych jako istotne, mierzących odległości od centrum miasta, sklepów spożywczym, stacji autobusowych, serwisów samochodowych i innych, budów, obiektów kulturalnych, obiektów religijnych innych niż katolickie, więzienia, instytucji zdrowotnych, departamentów policji i służby pożarnej, wejścia do metra, autostrad, dróg rowerowych i serwisowych. W skład zmiennych objaśniających należały również zmienne *restaurants_800*, *sport_objects_800*, *is_kitchen_separate*, *is_security*, *is_balcony*, *is_basement*, *is_elevator*, *is_terrace*, *is_garden*, *is_kitchen_furnished*, *is_heating_urban*, *rooms_num*, *building_type*.

Tabela 4. Wyniki testów autokorelacji przestrzennej dla reszt z modelu liniowego

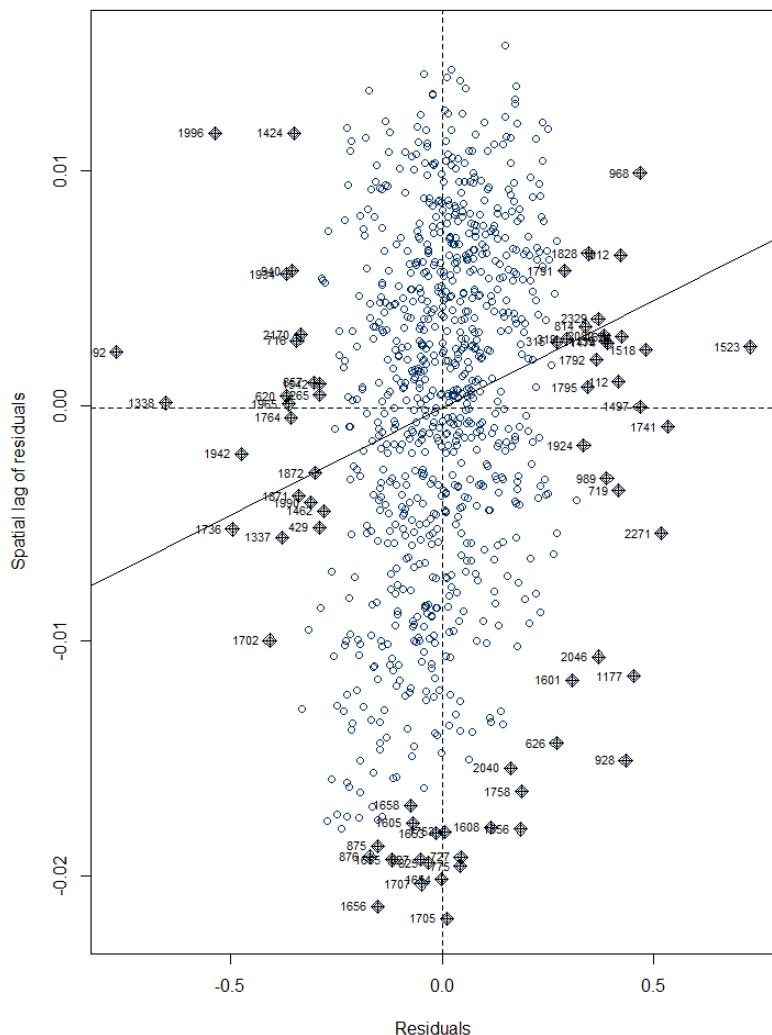
Nazwa testu	Statystyka testowa	Wartość krytyczna	Wartość p
Global Moran I	0,009	-0,003	<0,0001
C Geary'ego	0,976	1,000	<0,0001
LMerr	14,601		0,0001
LMlag	38,077		<0,0001
RLMerr	0,045		0,8303
RLMlag	23,522		<0,0001

Źródło: opracowanie własne

Pierwszym zastosowanym testem był globalny test Morana. Hipoteza zerowa w tym teście zakłada, że dane nie są w jakiś sposób zgrupowane przestrzennie. Uzyskana statystyka testowa wynosiła 0,009, co przy obliczonej wartości krytycznej (-0,003) daje możliwość stwierdzić, że w danych występują pewne klastry geograficzne (wartość p < 0,001). Jednakże otrzymana statystyka I Morana jest bardzo niska, co oznacza, że prawdopodobnie dane charakteryzują się losowymi skupiskami w losowych obszarach. Niemniej jednak niska wartość statystyki I Morana może być również konsekwencją dobrego wyjaśnienia przestrzennych zależności przez zmienne objaśniające, mierzące odległości minimalne. Wnioski na temat autokorelacji przestrzennej są wzmacniane za pomocą statystyki C Geary'ego, która jest miarą autokorelacji przestrzennej mająca na celu określenie, czy obserwacje tej samej zmiennej są globalnie przestrzennie skorelowane (a nie na poziomie lokalnym). C Geary'ego jest odwrotnie powiązana ze statystyką Morana, ale nie jest jej identyczne. Chociaż I Morana i C Geary'ego są miarami globalnej autokorelacji przestrzennej, C Geary'ego wykorzystuje sumę kwadratów odległości, podczas gdy I Morana wykorzystuje standaryzowaną kowariancję przestrzenną. Używając odległości kwadratowych, C Geary'ego jest

mniej wrażliwe na powiązania liniowe i może wykryć autokorelację tam, gdzie I Morana wskazywałaby na brak autokorelacji. Z wartością p mniejszą od 0,001 stwierdzono, że dane charakteryzują się występowaniem autokorelacji przestrzennej. Jednakże statystyka testowa była równa 0,9766, co nie wiele różni się od wartości 1, która oznacza brak autokorelacji. Lokalne testy Morana wskazują, że autokorelacja przestrzenna jest obserwowana w 126 z 893 przypadków. Testy mnożników Lagrange'a (z opóźnieniem i bez) też dostarczają wniosków na odrzucenie hipotezy zerowej o niewystępowaniu autokorelacji przestrzennej. Co więcej, wskazują one również na to, że najlepszym modelem okaże się model błędu przestrzennego, z uwagi na większą statystykę testową. Wykres punktowy Morana jest ostatnim narzędziem analitycznym, który ilustruje zależności przestrzenne.

Rysunek 31. Punktowy wykres Morana



Źródło: opracowanie własne

W przypadku mocnej autokorelacji przestrzennej, punkty oznaczające pojedyncze obserwacje muszą kumulować się wzdłuż linii nachylonej pod kątem 45 stopni. Jednakże wyraźnie widać, że wartości są skumulowane wzdłuż osi y, co oznacza brak autokorelacji przestrzennej, a statystyki testowe mogą być zanieczyszczone dwoma klastrami obserwacji odstających (w lewym dolnym i prawym górnym kwadracie), które i odzwierciadlają klastry przestrzenne. Dlatego na tym etapie stwierdzono, że nie ma jednoznacznych przyczyn stosowania autoregresyjnych modeli przestrzennych, skoro prawdopodobniej będą one doświadczali co najmniej nieefektywnych oszacowań parametrów z uwagi na małą autokorelację przestrzenną. Jednakże zostawiono te modele w następnych etapach analizy w celu sprawdzenia możliwości generalizacji tych modeli na nowe obserwacje i porównania wydajności względem jakości predykcji w przeciwnieństwie do modeli uczenia maszynowego.

4.2. Optymalne hiperparametry. Parametry rozkładów a priori

Kolejny etap analizy obejmował optymalizację hiperparametrów modeli uczenia maszynowego. W tym celu wykorzystano algorytm genetyczny z populacją równą 20, składający się z 20 generacji, wielkością turnieju wynoszącą 10, prawdopodobieństwem mutacji 0,1 oraz prawdopodobieństwem krzyżowania 0,8. Jako miary dopasowania użyto pierwiastka błędu średniokwadratowego. Proces optymalizacji przeprowadzono poprzez sprawdzian krzyżowy (ang. *cross validation*) obejmujący 10 podzbiorów. Optymalne wartości hiperparametrów wraz z czasem obliczeń zostały przedstawione w Tabeli 5:

Tabela 5. Wyniki strojenia hiperparametrów

Model	Optymalne hiperparametry	Czas obliczeń, min
Lasy losowe (RF)	n_estimators=720,max_depth=40,min_samples_split=4, min_samples_leaf=2,max_features=0,3492	118,1
Wyjątkowo losowe drzewa (ET)	n_estimators=299,max_depth=50,min_samples_split=8, min_samples_leaf=2,max_features=0,3549	19,5
Wzmacnianie adaptacyjne (AB)	n_estimators=650, learning_rate= 0,0743	84,61
Wzmacnianie gradientem (GB)	n_estimators=948, learning_rate= 0,1001	87,71
Xgboost (XGB)	n_estimators=592, learning_rate=0,1228	63,00

LightGBM (LGBM)	n_estimators=286, learning_rate=0,0758	5,96
Catboost (CB)	n_estimators=978, learning_rate=0,1239	19,46
Perceptron wielowarstwowy (MLP)	hidden_layers=938,max_iter=703,learning_rate=adaptive, optimizer=adam	28,31
Przestrzenna sieć neuronowa (GWANN)	hidden_layers=40,learning_rate=0,1,optimizer=adam, adaptive_learning_rate=FALSE	–

Źródło: opracowanie własne

W ten sposób najbardziej wydajnymi modelami pod względem czasu strojenia okazały się LGBM, ExtraTrees oraz CatBoost. W przypadku drzew wyjątkowo losowych osiągnięto to poprzez losowy wybór kryterium podziału drzewa w każdym węźle decyzyjnym. Z kolei szybki czas obliczeń LGBM i CatBoost wynikał z efektywnego algorytmu budowy drzew. W odniesieniu do metod wzmacniania estymatory bazowe nie uległy optymalizacji, ponieważ zakładano dostatecznie dużą liczbę danych, aby efektywnie korzystać z podstawowej koncepcji tych modeli, polegającej na łączeniu słabych estymatorów. Ponadto pominięto proces strojenia hiperparametrów przestrzennej sieci neuronowej GWANN ze względu na obecnie ograniczone możliwości optymalizacji, wynikające z praktycznej implementacji tego modelu. Zamiast tego, wartości hiperparametrów zostały ustalone na podstawie oryginalnej pracy Hagenauera i Helbicha.⁸⁸

W tym momencie konieczne jest również dokładne określenie parametrów rozkładów a priori oraz parametrów algorytmu Metropolisa-Hastingsa, istotne dla konstrukcji modeli Bayesowskich o charakterze przestrzennym. Wobec parametru ρ zastosowany został rozkład a priori $\pi(\rho) \sim B(1,1)$, natomiast w przypadku parametru λ w modelu SEM, SDEM i GNS przyjęto rozkład a priori w postaci $\pi(\lambda) \sim \Gamma(1,1)$. Ze względu na ograniczoną wiedzę na temat innych rozkładów parametrów, użyto nieinformacyjnych rozkładów a priori. W wybranym procesie estymacji skorzystano z czterech łańcuchów, z których każdy składał się z 6000 iteracji; pierwsze 1000 iteracji traktowano jako okres rozgrzewania (ang. *burn-in*).

⁸⁸ Hagenauer J., Helbich M., op. cit.

4.3. Porównanie użytych modeli

4.3.1. Wyliczone kryterium porównawcze

W niniejszej pracy głównymi kryteriami porównawczymi przyjęto błędy predykcji oraz czas obliczeń. Obliczenia zostały przeprowadzone na podstawie 10-krotnego sprawdzianu krzyżowego. Czas obliczeń podany w tabeli odnosi się do wygenerowania prognoz dla wszystkich 10 podzbiorów sumarycznie. Ponadto zostały również obliczone kryterium porównawcze dla modeli opartych na średniej, medianie oraz regresji liniowej – modele te stanowią punkt wyjścia do dalszych porównań.

Tabela 6. Kryterium porównawcze

Model	MSE	RMSE	MPE	MAPE	sMAPE	Czas (min)
adaboost	0,065	0,254	-0,011	1,916	1,911	7,450
bGNS	0,023	0,152	0,064	1,187	1,189	3,217
bSAR	0,0240	0,155	0,0311	1,210	1,21	1,200
bSAR (pure)	0,053	0,231	0,058	1,838	1,842	0,070
bSARAR	0,026	0,160	-0,013	1,284	1,284	2,357
bSDEM	0,023	0,152	-0,008	1,185	1,185	3,150
bSDM	0,035	0,188	0,107	1,500	1,500	2,633
bSEM	0,054	0,231	-0,063	1,835	1,836	0,650
bSLX	0,023	0,153	-0,053	1,196	1,196	0,033
catboost	0,030	0,174	-0,034	1,139	1,132	0,917
ExtraTrees	0,031	0,175	-0,042	1,149	1,142	1,467
GNS	0,022	0,148	0,023	1,149	1,150	3,200
gradient boosting	0,035	0,186	-0,033	1,222	1,216	17,217
GWANN	0,206	0,454	-0,170	3,657	3,645	834,233
GWR	0,016	0,126	-0,018	1,032	1,032	2,698
LGBM	0,030	0,174	-0,023	1,148	1,141	0,100
LM	0,026	0,161	-0,028	1,266	1,266	0,004
Średnia	0,073	0,270	-0,081	2,201	2,199	0,001
Mediania	0,073	0,271	0,016	2,210	2,210	0,001
MLP	168,945	12,998	38,787	87,367	78,930	2,433
Random Forests	0,032	0,179	-0,043	1,159	1,152	11,917
SAR	0,187	0,433	0,049	3,762	3,763	1,733
SAR (pure)	0,253	0,502	0,063	3,830	3,840	1,734
SARAR	0,023	0,151	-0,023	1,175	1,175	3,317
SDEM	0,022	0,149	-0,030	1,163	1,162	1,861
SDM	0,022	0,147	-0,024	1,143	1,143	2,033
SEM	0,027	0,164	0,019	1,294	1,295	1,688
SLX	0,101	0,318	-0,096	2,694	2,693	1,291
xgboost	0,032	0,179	-0,033	1,157	1,150	1,133

Źródło: opracowanie własne

W odniesieniu do większości wybranych metryk (oprócz czasu obliczeń i średniego błędu procentowego), najlepszym modelem okazał się model regresji geograficznie ważonej (GWR). Biorąc pod uwagę wartości RMSE i MSE, najlepsze wyniki uzyskały również większość modeli ekonometrycznych: SDM, GNS, SDEM, SARAR, bGNS, bSDEM, bSLX, bSARAR, gdzie „b” oznacza modele Bayesowskie. Pozostałe modele wykazywały większe błędy niż model liniowy. Jeśli chodzi o czas obliczeń, to modele ekonometryczne ponownie zajeły czołowe pozycje: po prostych modelach opartych na średniej, medianie i regresji liniowej następuje model bSLX. W przypadku miary MPE najdokładniejsza wartość zbliżona do zera ($-0,008$) dotyczyła Bayesowskiego przestrzennego modelu Durbina (bSDEM). Analizując powyższe wyniki, można stwierdzić, że w tym badaniu modele przestrzenne oparte na metodach ekonometrycznych przewyższyły modele uczenia maszynowego oraz model hedoniczny.

Modele perceptronu wielowarstwowego oraz czystej autoregresji przestrzennej okazały się najmniej skutecznymi pod względem wszystkich miar jakości prognozy. Jeśli chodzi o model czystej autoregresji przestrzennej, jego struktura nie pozwala na uwzględnienie złożonych zależności przestrzennych i zazwyczaj stanowi punkt odniesienia dla innych modeli autoregresyjnych. Co do słabej wydajności perceptronu wielowarstwowego, można przypuszczać, że wynika ona z niewystarczającej ilości danych w rozważanym zbiorze. Ponadto ten wniosek jest zgodny z wynikami analizy przeprowadzonej przez D. Przekopa.⁸⁹ W przypadku modelu przestrzennej sieci neuronowej GWANN, jego wydajność okazała się gorsza od modelu liniowego. Taki wynik podważa praktyczną przydatność tego modelu, przynajmniej w aspekcie prognozowania cen sprzedaży mieszkań w Warszawie. Co więcej, model ten charakteryzował się najdłuższym czasem obliczeniowym, co było konsekwencją problemu implementacyjnego opisanego w rozdziale metodologicznym: przy prognozowaniu konieczne było wielokrotne powtarzanie nowej obserwacji tyle razy, ile było neuronów wyjściowych. Przykładowo, jeśli model miał 8000 neuronów wyjściowych, a zbiór testowy obejmował 2000 obserwacji, wymagało to przewidzenia 16 milionów jednostek. W rezultacie koncepcja modelu GWANN może być ciekawa z teoretycznego punktu widzenia, jednak ma ograniczone praktyczne zastosowania na obecnym etapie rozwoju.

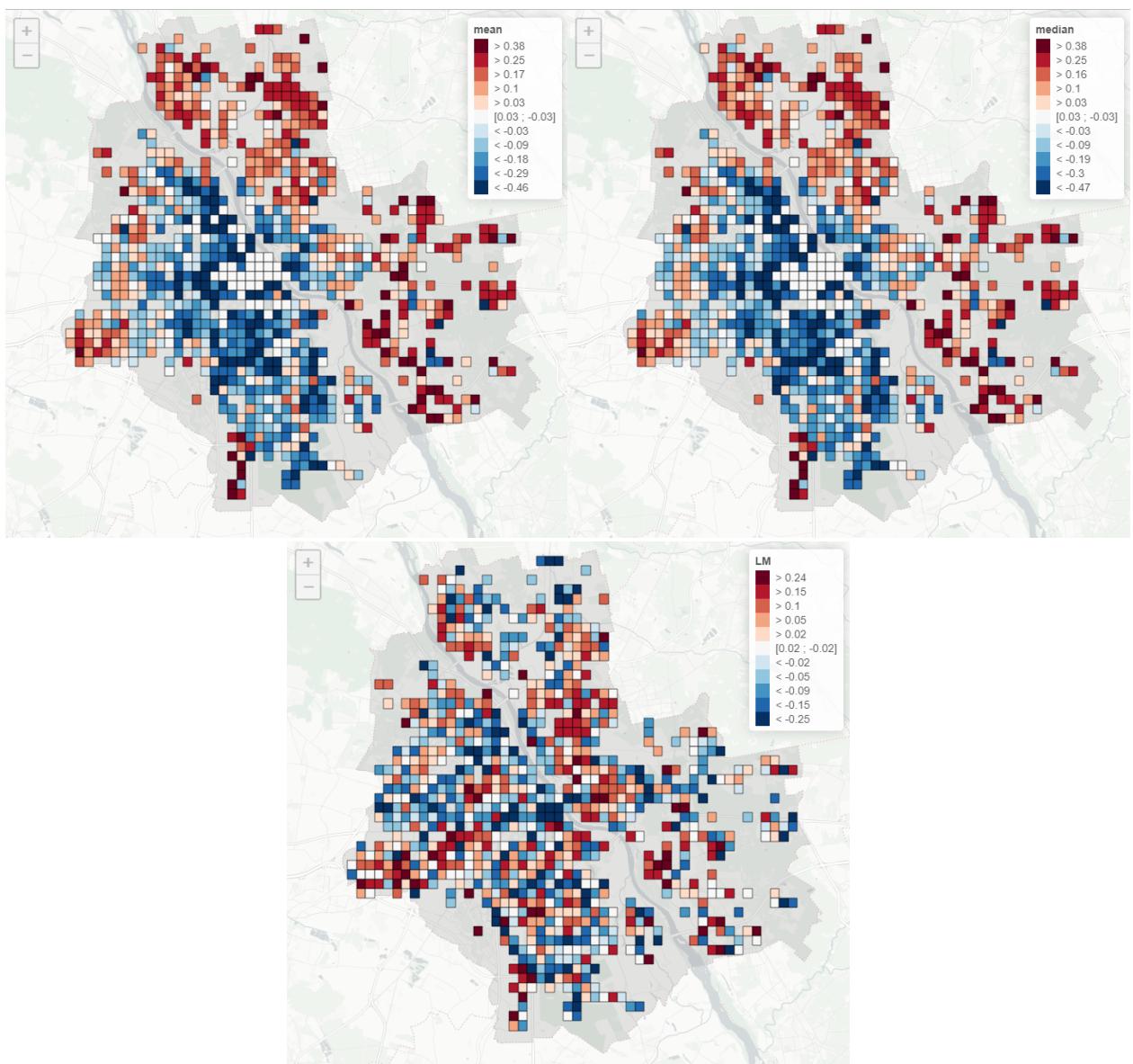
4.3.2. Wykresy reszt

Kolejnym krokiem analizy była dokładna eksploracja graficzna reszt pochodzących z każdego modelu, rozpoczynając od modelu hedonicznego, mediany i średniej. W przypadku średniej oraz

⁸⁹ Przekop D., op.cit.

mediany okazały się one adekwatnymi wskaźnikami cen mieszkań jedynie w sytuacji ofert z samego centrum Warszawy, ponieważ obszary sąsiadujące z centrum były zazwyczaj niedoszacowane, natomiast obszary na peryferiach znacznie przeszacowane. W przypadku reszt z modelu liniowego (LM), można zauważać nielosowy rozkład reszt tylko w obszarze Powiśla, gdzie wystąpiło znaczące niedoszacowanie cen sprzedaży. Niemniej jednak losowość przestrzennego rozkładu reszt sugeruje, że model hedoniczny skutecznie wykrywa zależności przestrzenne.

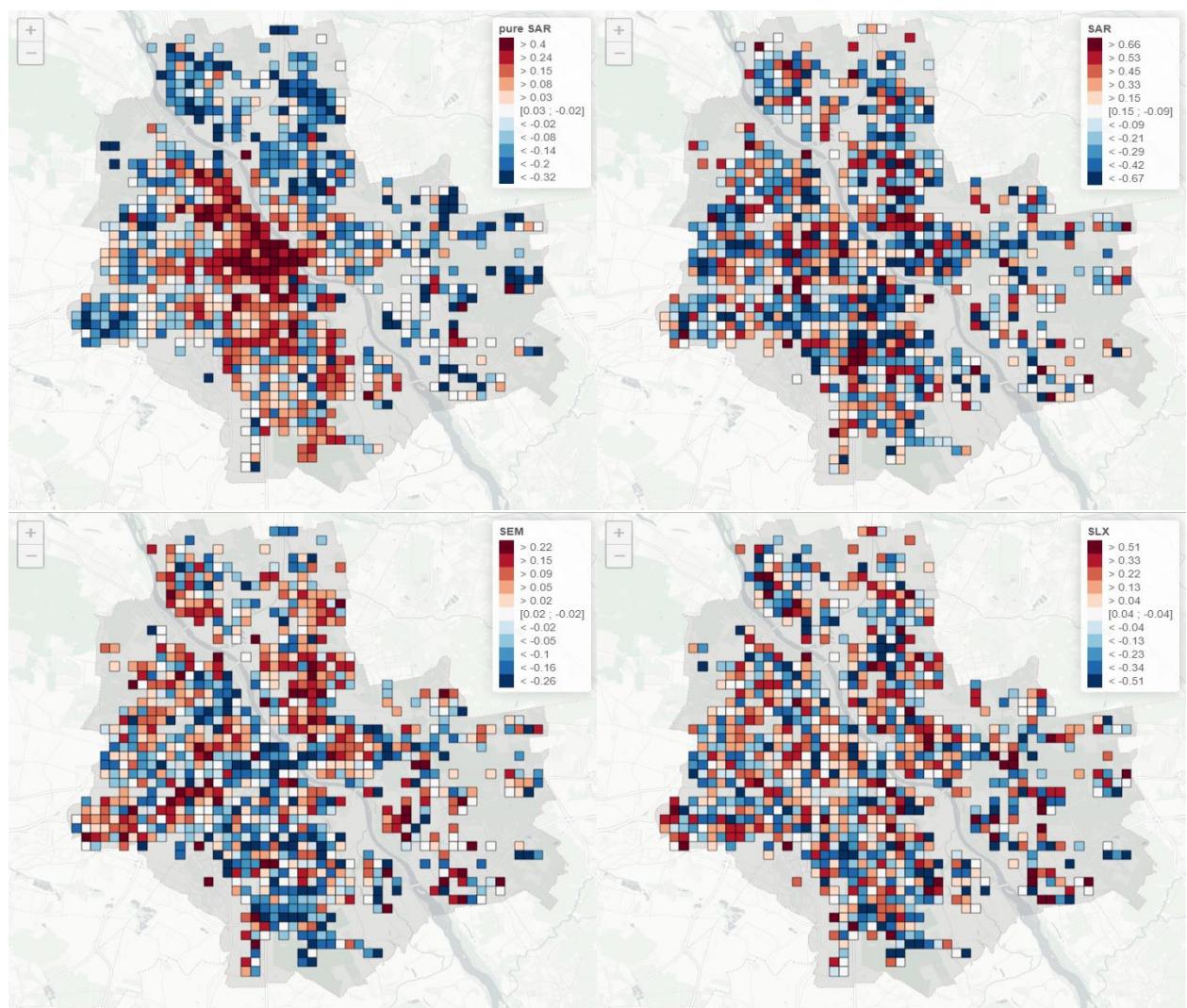
Rysunek 32. Reszty z modelu regresji liniowej oraz reszty z prognozowania za pomocą medialnej i średniej wartości

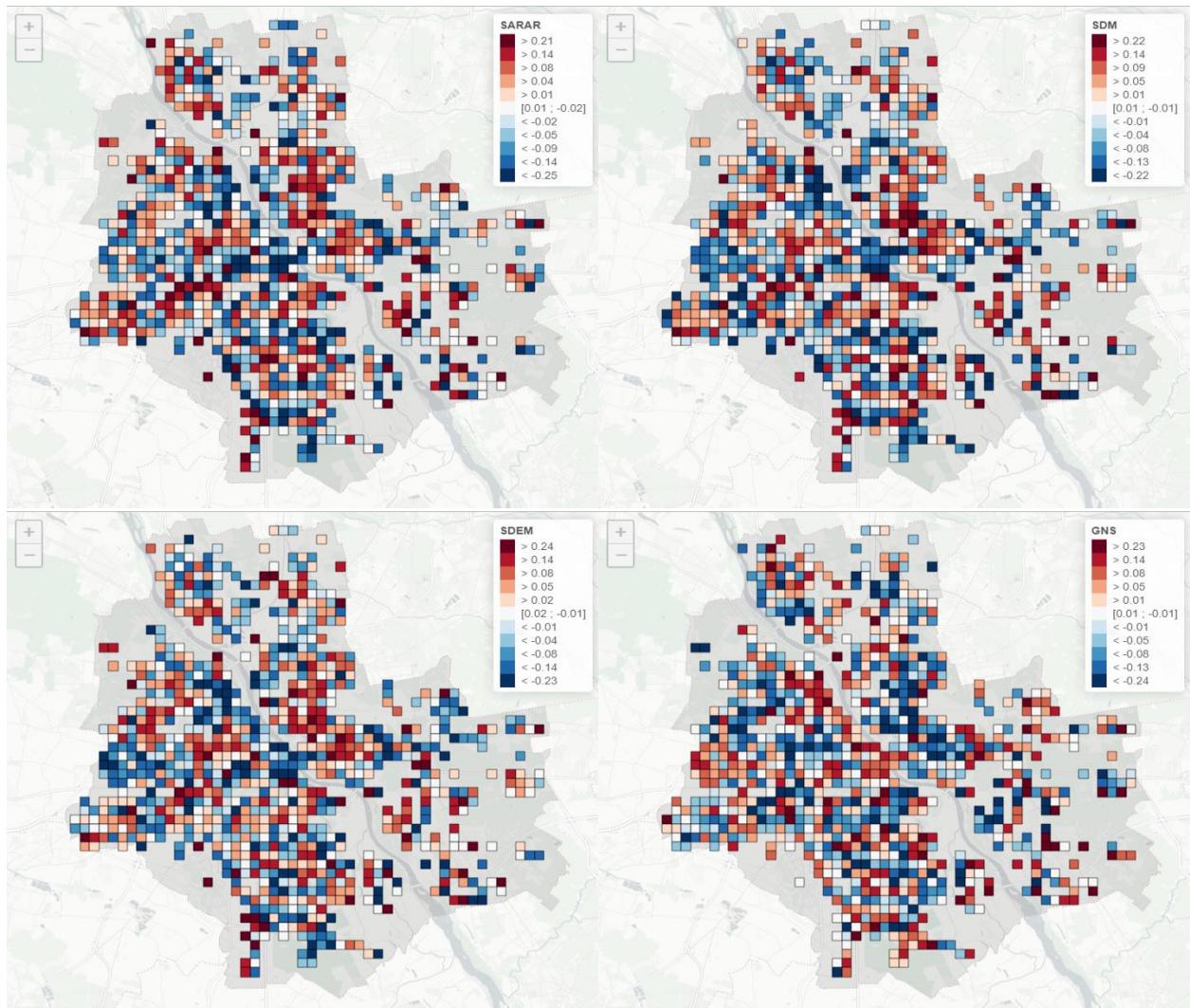


Źródło: opracowanie własne

Przechodząc do analizy modeli autoregresyjnych, warto zwrócić uwagę, że w przypadku modelu SAR zaobserwowano występowanie reszt dodatnich prawie jedynie w centrum miasta oraz reszt ujemnych na terenie obszarów peryferyjnych stolicy. W odniesieniu do pozostałych modeli, reszty przyjmowały charakter bardziej przypadkowy, chociaż należy zaznaczyć występowanie skupisk przestrzennych reszt dodatnich w obszarze Targówka (model SEM) oraz reszt ujemnych w okolicach Powiśla (SEM, SDM, SDEM, SARAR), Pragi-Północy (GNS) i Górnego Ursynowa (SEM). Najbardziej losowe reszty zaobserwowano w przypadku modelu opóźnienia przestrzennego regresorów (SLX).

Rysunek 33. Reszty z autoregresyjnych modeli przestrzennych

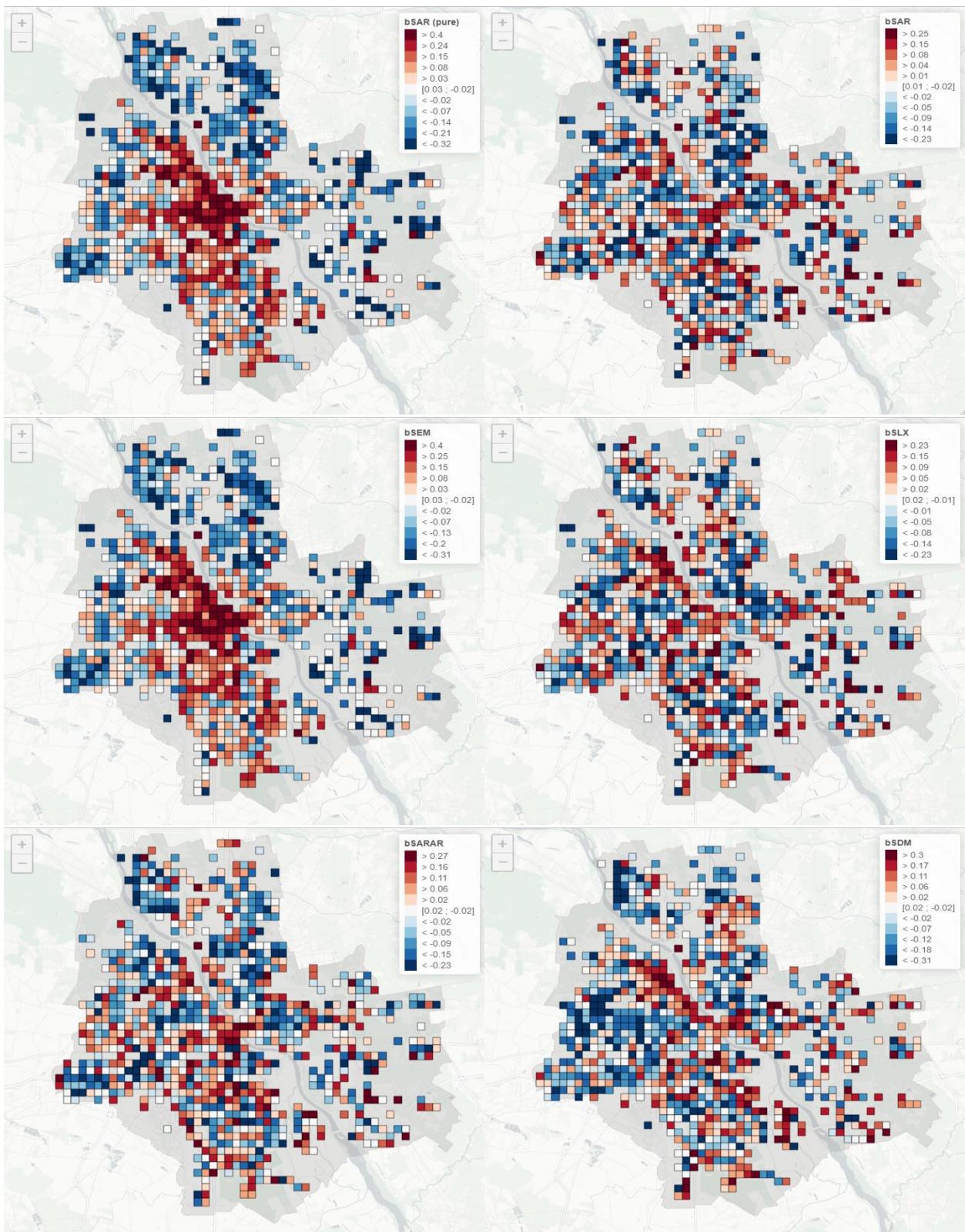


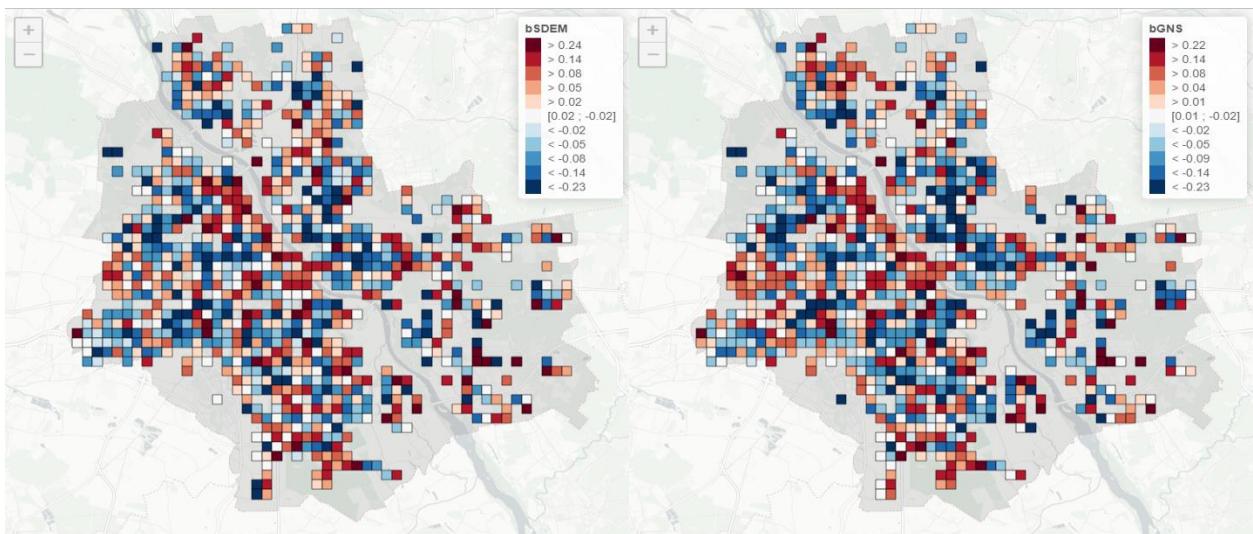


Źródło: opracowanie własne

W odniesieniu do modeli Bayesowskich, takich jak czysty Bayesowski model autoregresyjny (pure bSAR), bSDM oraz bSEM można zaobserwować wyraźne przestrzenne klastry dodatnich reszt w centrum Warszawy, oraz reszt ujemnych na Ursynowie oraz na prawym wybrzeżu Wisły. Analizując pozostałe modele, zaobserwowano bardziej nieregularny rozkład przestrzenny reszt; jedyne skupiska (ujemnych) reszt dotyczyły Pragi Południe. Ogólnie można stwierdzić, że jakość dopasowania względem rozkładu reszt w przypadku modeli Bayesowskich jest bardzo bliska do modeli klasycznych.

Rysunek 34. Reszty z Bayesowskich autoregresyjnych modeli przestrzennych

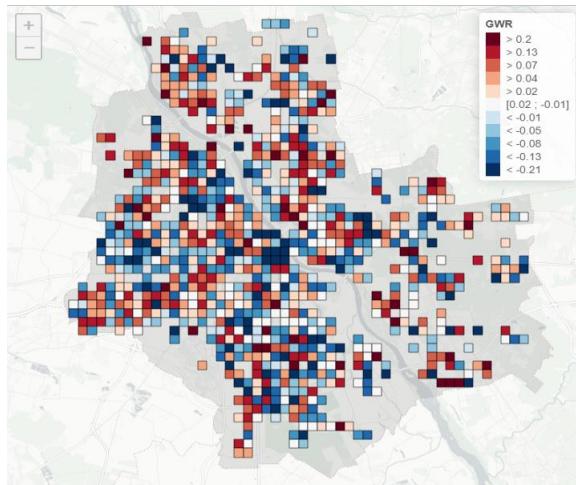




Źródło: opracowanie własne

W odniesieniu do modelu GWR nie znaleziono żadnych wyraźnych przestrzennych skupisk reszt, z wyjątkiem obszaru Powiśla, który stanowił wyzwanie również dla większości innych ekonometrycznych modeli autoregresyjnych. Wnioski z analizy graficznej dotyczącej modelu GWR w połączeniu z obliczonymi kryterium porównawczym skłania do uznania tego modelu za najlepszy model w prognozowaniu cen mieszkań metra kwadratowego w Warszawie.

Rysunek 35. Reszty z modelu GWR

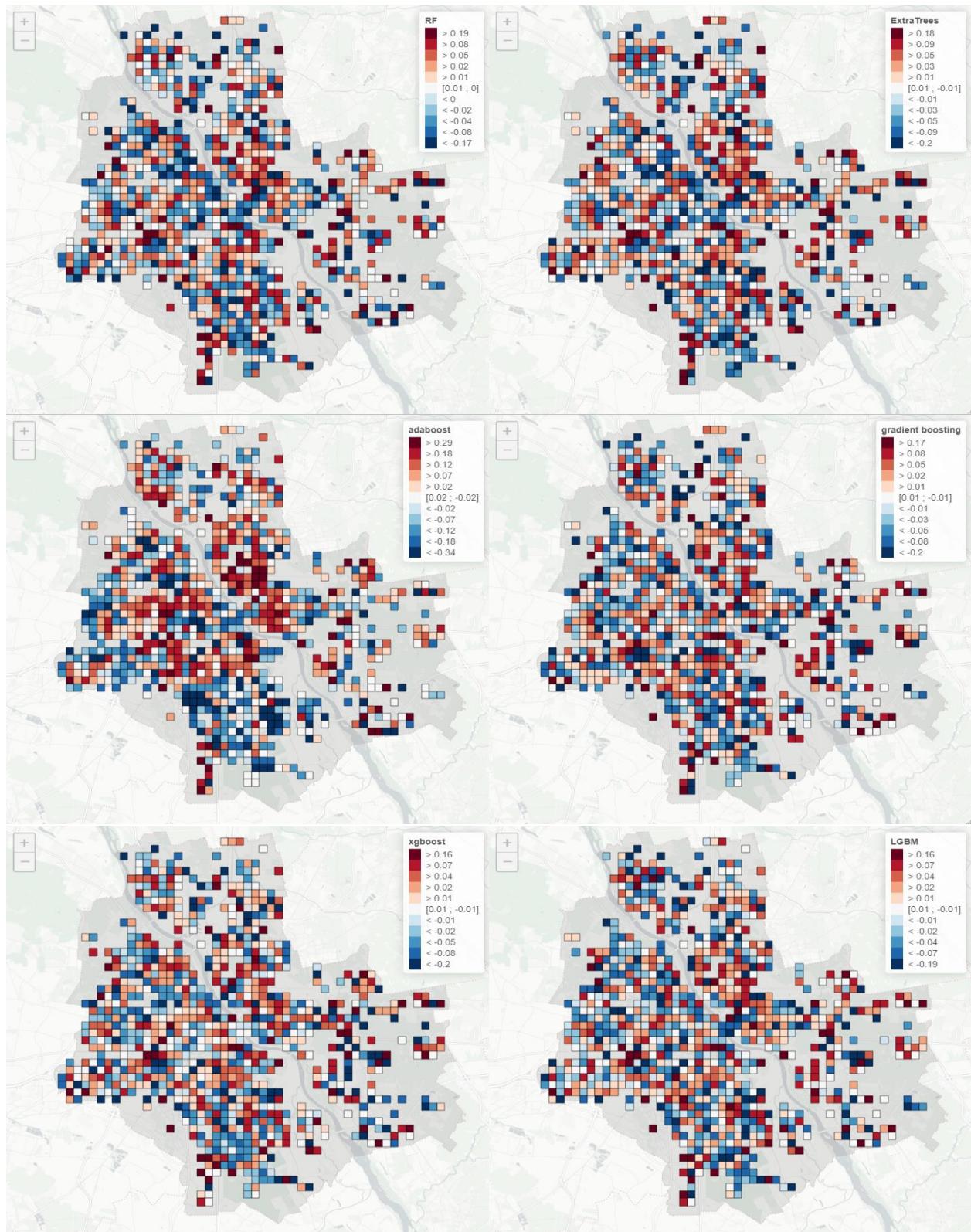


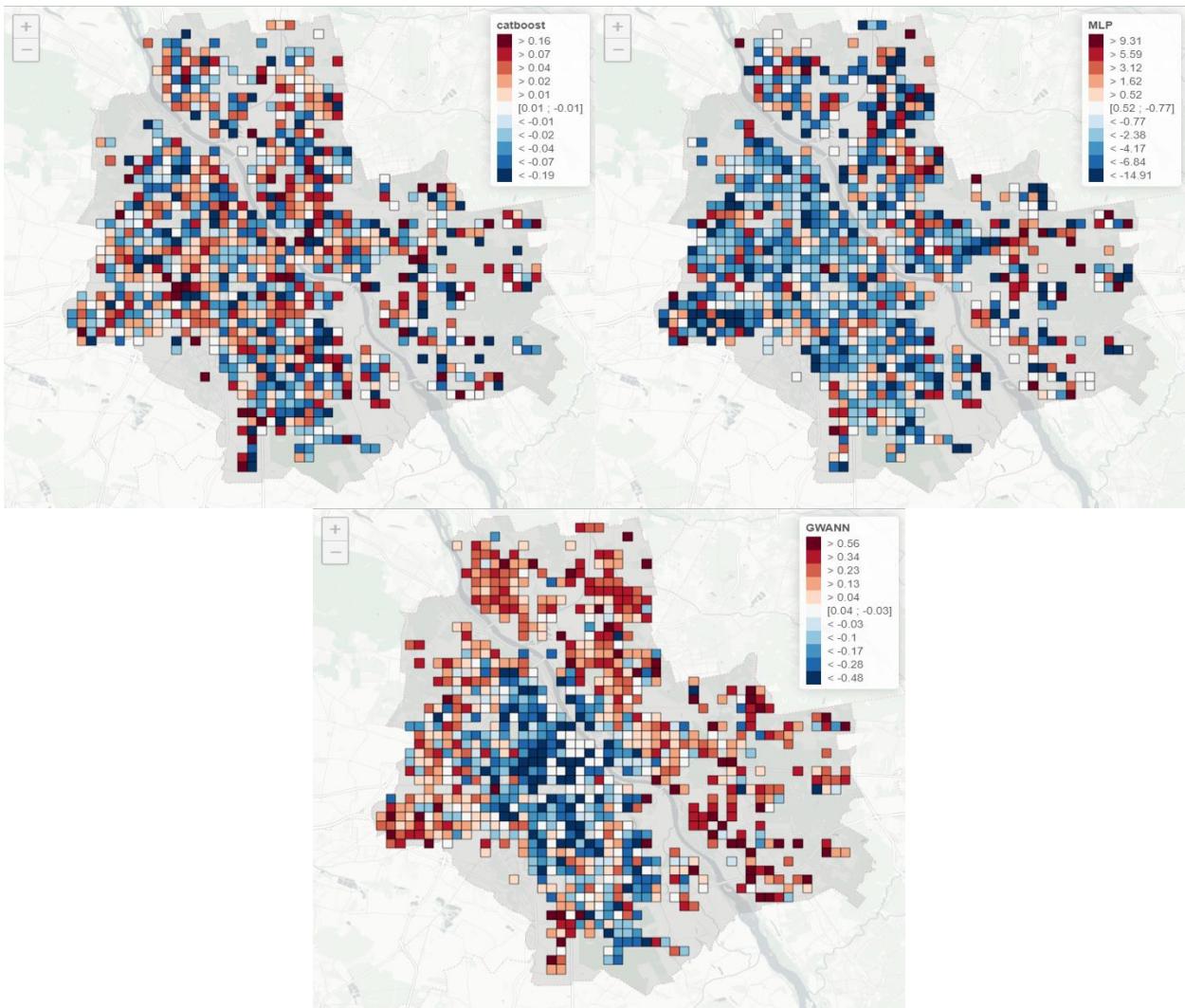
Źródło: opracowanie własne

W stosunku do ostatniej grupy modeli, czyli modeli uczenia maszynowego, warto podkreślić, że mimo że modele te cechowały się większymi błędami predykcji w porównaniu do większości modeli ekonometrycznych, to jednak przeważnie obserwowało losowy rozkład reszt w większości z nich. Należałoby wspomnieć, że przestrzenne klastry w resztach pojawiły się w przypadku modelu adaboost (niedoszacowanie cen w centrum miasta oraz na obszarze Pragi-Południe), modelu MLP (obejmującego niemal całą Wolę i Żoliborz) oraz modelu GWANN. Zastanawiające jest, że wykres

rozkładu reszt modelu GWANN przypominał wykres średniej i mediany, co budzi dodatkowe wątpliwości co do praktycznej przydatności tego modelu.

Rysunek 36. Reszty z modeli uczenia maszynowego.





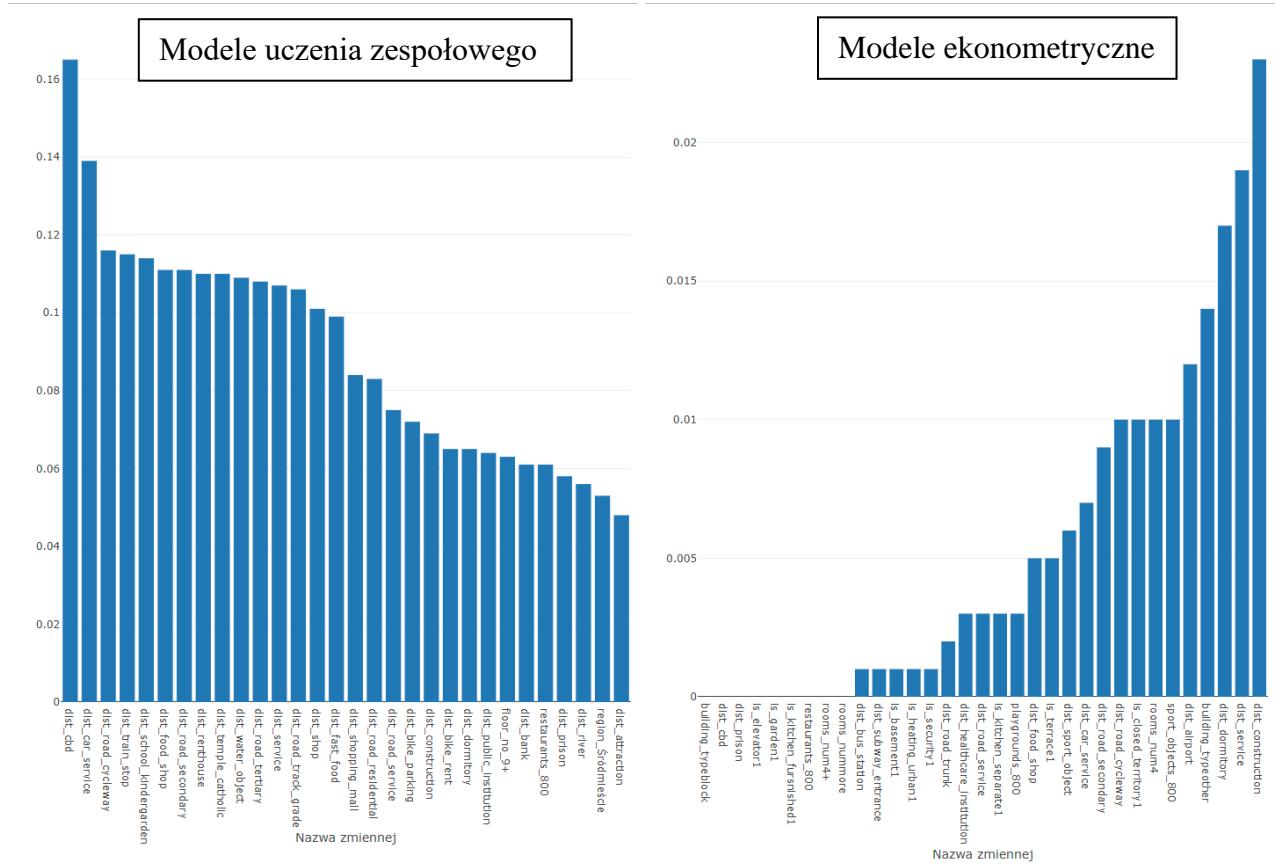
Źródło: opracowanie własne

Podsumowując, można wyciągnąć kilka fundamentalnych wniosków dotyczących różnych modeli prognozujących ceny mieszkań w Warszawie. Wykazano, że model hedoniczny, w którego jakości występuje regresja liniowa, mimo swojej prostoty, może być adekwatnym wskaźnikiem cen. Model opóźnienia przestrzennego regresorów SLX wyróżnił się najlepszą zdolnością wykrywania przestrzennych zależności w danych wśród modeli autoregresyjnych według losowości rozkładu reszt, a model geograficznie ważonej regresji wydaje się być najbardziej precyzyjnym w predykcji narzędziem statystycznym. Jednakże w przypadku prognozowania z poza próby w modelu GWR, polegającego na przypisaniu nowym obserwacjom wartości współczynników z lokalizacji sąsiadujących z pewnym wygładzaniem (w tej pracy równym 0,160), obserwacje leżące stosunkowo daleko od obiektów w próbie uczącej mogą odznaczać się większą wariancją prognozy. Modele uczenia maszynowego, mimo że cechowały się zazwyczaj losowym rozkładem reszt, miały większe błędy predykcji.

4.3.3. Istotność zmiennych. Współczynniki w modelu GWR

Na zakończenie analizy skoncentrowano się na ocenie istotności zmiennych objaśniających. W przypadku modeli wykorzystujących uczenie maszynowe, możliwość takiej oceny występuje jedynie w modelach opartych na uczeniu zespołowym. Niemniej jednak nawet w tym przypadku istotność zmiennych dotyczy jedynie częstotliwości wykorzystania poszczególnych regresorów przez estymatory bazowe mierzona w ujęciu procentowym. W celu lepszej wizualizacji skupiono się tylko na 30 najczęściej używanych zmiennych spośród siedmiu porównywanych modeli. Z kolei w kwestii modeli ekonometrycznych możliwa jest ocena istotowości zmiennych pod względem statystycznym za pomocą testów t-Studenta. Z tego powodu wybrano zmienne, które wykazały się wartością p powyżej przyjętej wartości błędu I rodzaju równej 0,05.

Rysunek 37. Średnia częstość występowania zmiennych objaśniających w estymatorach bazowych modeli zespołowych oraz średnia wartość p w modelach ekonometrycznych



Źródło: opracowanie własne

Najczęściej wykorzystywaną zmienną w modelach opartych na uczeniu zespołowym okazała się zmienna mierząca odległość od centrum Kultury i Nauki ($dist_cbd$), co jest zgodne z wynikami eksploracyjnej analizy danych, gdyż mianowicie ten regresor charakteryzował się najwyższym

współczynnikiem korelacji rang Spearmana. Ogólnie rzecz biorąc, większość zmiennych wybieranych najczęściej cechowała się wysokim współczynnikiem rang Spearmana, z wyjątkiem zmiennych, które opisywały się nieliniowymi rozkładami dwuwymiarowymi. Podobne wyniki obserwowano także w stosunku do modeli ekonometrycznych, gdzie drugą najważniejszą zmienną ponownie okazała się *dist_cbd*. Warto jednak zaznaczyć, że w przeciwieństwie do modeli opartych na uczeniu zespołowym, zmienne kategorialne częściej uzyskiwały status istotnych, co można tłumaczyć tym, że modele oparte na uczeniu zespołowym wykazują większą skłonność do uwzględniania zmiennych kategorialnych z większą liczbą kategorii, o czym także wspomniano w metodyce badania.

Na samy koniec przedstawiono uśrednione wartości współczynników β stojących przy każdej ze zmiennej w równaniu modelu regresji geograficznie ważonej.

Tabela 7. Wartości współczynników β w modelu GWR

Zmienna	Współczynnik
Stala	10.1714
dist_bus_station	-0.0743
dist_car_service	0.0248
dist_construction	0.0188
dist_cultural	0.0115
dist_dormitory	-0.0429
dist_food_shop	0.0246
dist_healthcare_institution	-0.0304
dist_prison	0.0001
dist_public_service	-0.0156
dist_service	0.0196
dist_subway_entrance	-0.0369
dist_road_cycleway	-0.0203
dist_road_service	0.0246
dist_road_trunk	0.0001
playgrounds_800	-0.0029
restaurants_800	0.0972
sport_objects_800	-0.0018
is_kitchen_separate1	-0.0621
is_closed_territory1	0.0453
is_security1	0.0359
is_basement1	-0.0607
is_elevator1	0.0526
is_terrace1	0.0669
is_garden1	-0.0913
is_kitchen_furnished1	0.0893
is_heating_urban1	0.0713
rooms_num4	-0.0576
rooms_num4+	-0.1477
building_typeblock	-0.1184
building_typeother	-0.0435

Źródło: opracowanie własne

W odniesieniu do zdecydowanej większości zmiennych kierunek oddziaływania był zgodny z oczekiwaniami, wynikającymi z analizy literatury badawczej oraz przeprowadzonej analizy eksploracyjnej w ramach niniejszej pracy. Jedynie zmienne, które wyróżniały się w tym kontekście, to istnienie ogródka, piwnicy oraz większej liczby pokoi. W przypadku tych zmiennych najczęściej występującymi wartościami była wartość 1, co jest charakterystyczne dla obszarów peryferyjnych Warszawy, gdzie ceny mieszkań są niższe. Co dotyczy siły oddziaływania każdej ze zmiennych osobno, struktura modelu nie pozwala na wyciągnięcie jednoznacznych wniosków dotyczących statystycznej istotności pojedynczych zmiennych. W związku z tym, siła oddziaływania każdego z regresorów w końcowym modelu pozostaje kwestią otwartą.

Zakończenie

Na podstawie analizy empirycznej skonstatowano, że optymalnym modelem do przewidywania cen mieszkań w Warszawie jest model regresji geograficznie ważonej. W gruncie rzeczy, modele oparte na uczeniu maszynowym odznaczały się mniejszą efektywnością, czasami nawet niż prostsze modele jak model regresji liniowej (model hedoniczny), a najwyższa jakość dopasowania wystąpiła w przypadku perceptronu wielowarstwowego. Potencjalną przyczyną tego zjawiska może być nietypowo słaba struktura autokorelacji przestrzennej cen mieszkań w Warszawie oraz (prawie) liniowy charakter zależności zmiennej objaśnianej i rozważanych regresorów. Należy zaznaczyć jednak, że osiągnięte w tej pracy wyniki opierały się na ograniczonym zbiorze danych i mogą różnić się od wyników badań wykorzystujących do analizy znacznie większe zbiorów. Pomimo to, zwiększenie rozmiaru zbioru danych stanowi wyzwanie w przypadku Warszawy, z racji ograniczonej dostępności danych dotyczących rzeczywistych transakcji cenowych. Na dzień dzisiejszy jedyne dostępne źródła danych o cenach transakcyjnych zasługujące na zaufanie to portale internetowe takie jak otodom.pl i OLX.pl⁹⁰, oraz Biuro Geodezji i Katastru.⁹¹ Jednakże w przypadku Biura Geodezji i Katastru, pozyskanie zbioru danych zawierającego informacje o sprzedaży 10000 mieszkań szacowane jest na około 5000 złotych na moment napisania pracy.

Poza tym nie tylko modele bazujące na uczeniu maszynowym napotykały na trudności podczas przebadania. Modele autoregresywne również miały pewne ograniczenia obliczeniowe, co skłoniło do zastosowania metody agregacji cen mieszkań na podstawie siatki kwadratowej w celu ich estymacji. Dodatkowo, zarówno klasyczne modele, jak i modele oparte na metodach Bayesowskich borykały się z brakiem zaawansowanych narzędzi analitycznych dostępnych w źródłach darmowych, takich jak pakiet statystyczny R czy język programowania Python. Na przykład, pierwsza próba zaspokojenia potrzeb związanych z implementacją modeli Bayesowskich w języku R miała miejsce dopiero w 2021 roku.⁹² W następstwie tego kluczowe staje się dalsze rozwijanie specjalistycznego oprogramowania, które w dzisiejszych czasach odgrywa wiodącą rolę w promowaniu zainteresowania dziedziną ekonometrii przestrzennej. Warto podkreślić jednak, że wyzwania związane z modelami ekonometrii przestrzennej przynoszą pewne korzyści w postaci rozwijania nowych metod i narzędzi, które w efekcie przyczyniają się do postępu w dziedzinie analizy danych przestrzennych.

⁹⁰ <https://www.olx.pl/nieruchomosci/mieszkania/> (dostęp 15.08.2023).

⁹¹ <https://architektura.um.warszawa.pl/geodezja> (dostęp 15.08.2023).

⁹² Kuschnig N., *Bayesian Spatial Econometrics and the Need for Software*, Department of Economics Working Paper Series Wyd. 318, Wiedeń 2021.

Odrębnie warto również skupić się na przyszłości hybrydowych modeli, takich jak model GWANN. W ramach niniejszej analizy model ten wykazał stosunkowo niską wydajność, zarówno pod względem błędu predykcji, jak i czasu obliczeń. Jednakże należy mieć na uwadze, że zastosowanie modeli hybrydowych w kontekście ekonometrii przestrzennej jest zjawiskiem nowatorskim, co oznacza, że nowe modele są nadal w fazie opracowywania. Ciekawym przykładem innego modelu hybrydowego może okazać się autoregresyjna sieć przestrzenna (ang. *Spatial Autoregressive Neural Network*, SARNN), która pojawiła się na scenie naukowej dopiero w czerwcu 2023 roku.⁹³ Ten model jest analogiczny do struktury modelu AR-Net,⁹⁴ w którym długość opóźnienia jest estymowana za pomocą sieci neuronowej. Dlatego kierunek zastosowania modeli hybrydowych łączących świat ekonometrii i uczenia maszynowego, jawi się jako obiecująca i wciąż rozwijająca się koncepcja, która otwiera nowe perspektywy badawcze w dziedzinie ekonometrii przestrzennej.

Niniejsza praca stanowi cenną wskazówkę dla podejmowania różnorodnych decyzji inwestycyjnych oraz innych działań ekonomicznych związanych z nabywaniem nieruchomości. Wynika to z faktu, że średni błąd procentowy w przypadku najlepszego modelu wyniósł zaledwie 1,03% na dziesięciu zestawach testowych, które były dobrane niezależne od zbioru wykorzystanego do estymacji parametrów modelu. Ponadto praca może stanowić punkt odniesienia do przyszłych porównań między hybrydowymi modelami a różnymi typami modeli statystycznych i opartych na uczeniu maszynowym. Dodatkowo praca może stanowić inspirację do tworzenia bardziej elastycznych narzędzi i pakietów dla języków programowania R i Python. Przykładowo, taki rozwój mógłby obejmować integrację z językiem programowania Stan, co pozwoliłoby na wykorzystanie bardziej zaawansowanych metod Bayesowskich w analizie danych przestrzennych, stwarzając tym samym możliwości dla dalszych badań i praktycznych zastosowań.

⁹³ Triebe O., Laptev N., Rajagopal R., *AR-Net: A simple Auto-Regressive Neural Network for time-series*, 2019, <https://arxiv.org/abs/1911.12436> (dostęp 28.08.2023).

⁹⁴ Wang Z., Song Y., *Deep learning for the spatial additive autoregressive model with nonparametric endogenous effect*, Spatial Statistics Wyd. 55, 2023.

Bibliografia

1. Anselin L., *Spatial Heterogeneity In Spatial Econometrics: Methods and Models*, Springer, Dordrecht 1988.
2. Barker J. S., *Simulation of Genetic Systems by Automatic Digital Computers*, Australian Journal of Biological Sciences Wyd. 11 Nr. 4, 1958, s. 603-612.
3. Bin J., Gardiner B., Liu Z. i in, *Attention-based multi-modal fusion for improved real estate appraisal: a case study in Los Angeles*, Multimed Tools Appl Wyd. 78, 2019, s. 31163–31184.
4. Breiman L., *Random Forests*, Machine Learning Wyd. 45, 2001, s. 5–32.
5. Bondemark A., Merkel A., *Parking not included: The effect of paid residential parking on housing prices and its relationship with public transport proximity*, Regional Science and Urban Economics Wyd. 99, 2023.
6. Borkowska S., Pokonieczny K, *Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development*, Sustainability Wyd. 14 Nr. 7, 2022, s.3728.
7. Bottero M. i in., *Urban parks, value uplift and green gentrification: An application of the spatial hedonic model in the city of Brisbane*, Urban Forestry & Urban Greening Wyd. 74, 2022.
8. Cankun W. i inn., *The Research Development of Hedonic Price Model-Based Real Estate Appraisal in the Era of Big Data*, Land Wyd. 11 Nr. 3, 2022.
9. Casali Y., Aydin N. Y., Comes T., *Sustainable Cities and Society, Machine learning for spatial analyses in urban areas: a scoping review*, Sustainable Cities and Society Wyd. 85, 2022.
10. Cranmer M. i in., *Discovering Symbolic Models from Deep Learning with Inductive Bayes*, NeurIPS, Vancouver 2020.
11. Chen T., *XGBoost: A Scalable Tree Boosting System*, University of Washington, Washington 2014.
12. Chipman H. A., George E. I., McCulloch R. E., *BART: Bayesian Additive Regression Trees*, The Annals of Applied Statistics Wyd. 4 Nr. 1, 2010, s. 266–298.
13. Chrostek K., Kopczewska K., *Spatial Prediction Models for Real Estate Market Analysis*, Ekonomia Wyd. 34, 2014.
14. Cliff A. D., Ord J. K., *Spatial Autocorrelation: A Review of Existing and New Measures with Applications*, Economic Geography Wyd. 46, 1970, s. 269-292.
15. Cliff A. D., Ord J. K., *Spatial autocorrelation*, Pion Limited, London 1973.
16. Congdon P., *Applied Bayesian modeling*, Wiley & Sons ltd., Hoboken 2003.

17. Deng H., Runger G., Tuv E., *Bias of Importance Measures for Multi-valued Attributes and Solutions*, 21st International Conference on Artificial Neural Networks, Espoo 2011.
18. Dreo J., Petrowski A., Siarry P., Taillard E., *Metaheuristics for Hard Optimization*, Springer, 2006.
19. Efthymiou D., Antoniou C., *How do transport infrastructure and policies affect house prices and rents? Evidence from Athens, Greece*, Transportation Research Part A: Policy and Practice Wyd. 52, 2013.
20. Fran ois B., Gallic E., *Recovering the French Party Space from Twitter Data*, Science Po Quanti, Paris 2015.
21. Freund Y., Schapire R. E., *Experiments with a New Boosting Algorithm*, Machine Learning: Proceedings of the Thirteenth International Conference, Nowy Jork 1996.
22. Friedman J. H., *Greedy Function Approximation: A Gradient Boosting Machine*, The Annals of Statistics Wyd. 29 Nr. 5, 2001, s. 1189-1232.
23. Fotheringham A., Charlton M., Brunsdon C., *Geographically Weighted Regression: A Natural Evolution Of The Expansion Method for Spatial Data Analysis*, Environment and Planning Wyd. 30 Nr. 11, 1998, s. 1905-1927.
24. Fukushima K., *Visual feature extraction by a multilayered network of analog threshold elements*, IEEE Transactions on Systems Science and Cybernetics Wyd. 5 Nr. 4, 1969, s. 322–333.
25. Gao S., Li L., Li W., Janowicz K., Zhang Y., *Constructing gazetteers from volunteered Big Geo-Data based on Hadoop*, Computers, Environment and Urban Systems Wyd. 61, 2017, s. 172-186.
26. Gebru T., Krause J., Wang Y., Fei-Fei L., *Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States*, PNAS Wyd. 114 Nr. 50, 2017, s. 13108-13113.
27. Geerts M., Broucke S., Weerdt J., *A Survey of Methods and Input Data Types for House Price Prediction*, ISPRS Int. J. Geo-Inf Wyd. 12 Nr. 5, 2023, s.200.
28. G eron A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, Sebastopol 2019.
29. Geurts P., Ernst D., Wehenkel L., *Extremely randomized trees*, Machine Learning Wyd. 63, 2006, s. 3-42.

30. Goodchild M.F, *Citizens as sensors: the world of volunteered geography*, GeoJournal Wyd. 69, 2007, s. 211–221.
31. Guliker E., Folmer E., Sinderen M., *Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach*, International Journal of Geo-Information Wyd. 11, 2022, s. 125 – 149.
32. Hagenauer J., Helbich M., *A geographically weighted artificial neural network*, International Journal of Geographical Information Science Wyd. 36 Nr. 2, 2022, s. 215-235.
33. Hamizah Z., Shuzlina A. R., Hasbiah U., *House Price Prediction using a Machine Learning Model: A Survey of Literature*, Modern Education and Computer Science Wyd. 6, 2020, s. 46-54.
34. Hewitson B., Crane R., *Neural nets: applications in geography*, The GeoJournal Library Wyd. 29, 1994, s. 196.
35. Ho T. K., *The random subspace method for constructing decision forests*, IEEE Transactions on Pattern Analysis and Machine Intelligence Wyd. 20 Nr. 8, 1998, s. 832-844.
36. Ho W., Tang B., Wong S. W., *Predicting property prices with machine learning algorithms*, Journal of Property Research Wyd. 38 Nr. 1, 2021, s. 48-57.
37. Hong I., Yoo C., *Analyzing Spatial Variance of Airbnb Pricing Determinants Using Multiscale GWR Approach*, Sustainability Wyd. 12 Nr. 11, 2020.
38. Jim C. Y., Chen W. Y., *External effects of neighbourhood parks and landscape elements on high-rise residential value*, Land Use Policy Wyd. 27, 2010, s. 662–670.
39. Kalliola J., Kapočiūtė-Dzikienė J., Damaševičius R., *Neural network hyperparameter optimization for prediction of real estate prices in Helsinki*, PeerJ Comput. Sc, 2021.
40. Kauko T., *The Importance of the Context and the Level of Analysis*, Druid Working Papers Wyd. 20, 2003, s. 134-136.
41. Ke G. i inn., *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach 2017.
42. Kuschnig N., *Bayesian Spatial Econometrics and the Need for Software*, Department of Economics Working Paper Series Wyd. 318, Wiedeń 2021.
43. Larm A., Ahelegbey, D. F, *Detecting Spatial and Temporal House Price Diffusion in the Netherlands: A Bayesian Network Approach*, Regional Science and Urban Economics Wyd. 65 Nr. 10, 2017 s. 1016

44. LeCun Y. i inn., *Backpropagation Applied to Handwritten Zip Code Recognition*, Neural Computation, Wyd. 1, 1989, s. 541–551.
45. Lesage J. P., *Bayesian Estimation of Spatial Autoregressive Models*, International Regional Science Review, Wyd. 20 Nr. 1–2, 1997, s. 113–129.
46. LeSage J. P., *Spatial Econometrics*, University of Toledo, Toledo 1998.
47. Lorenz F., Willwersch J., Cajias M., Fuerst F., *Interpretable machine learning for real estate market analysis*, Real Estate Economics, 2022, str. 1–31.
48. Louzada F., Nascimento D., Egbon O. A., *Spatial Statistical Models: An Overview under the Bayesian Approach*, Axioms Wyd. 10 Nr. 4, 2021, s. 307.
49. Lundberg S., Lee S., *A Unified Approach to Interpreting Model Predictions*, NIPS, 2017.
50. McCulloch W., Pitts W., *A Logical Calculus of Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics Wyd. 5 Nr. 4, 1943, s. 115–133.
51. Moran P. A., *Notes on Continuous Stochastic Phenomena*, Biometrika. Wyd. 37 Nr. 1, 1950, s. 17–23.
52. Openshaw S., *Modelling spatial interaction using a neural net*, Geographic Information Systems, Spatial Modelling and Policy Evaluation, 1993, s. 147–164.
53. Paelinck J., *Spatial econometrics*, Economics Letters Wyd. 1 Nr. 1, Amsterdam 1978, s. 59–63.
54. Park J., Yun S., *Social determinants of residential electricity consumption in Korea: Findings from a spatial panel model*, Energy Wyd. 239, 2022.
55. Paolino P.J., *Teaching linear correlation using contour plots*, Teaching Statistics Wyd. 43 Nr. 1, 2021, s. 13–20.
56. Prokhorenkova L. i inn., *CatBoost: unbiased boosting with categorical features*, Moscow Institute of Physics and Technology, Dolgoprudny 2019.
57. Przekop D., *Artificial Neural Networks vs Spatial Regression Approach in Property Valuation*, CEJEME Wyd. 14, 2022, s. 199–223.
58. Rico-Juan J. R., Taltavull de La Paz P., *Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain*, Expert Systems with Applications Wyd. 171, 2021.
59. Rosen S., *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition*, Journal of Political Economy, Wyd. 82 Nr. 1, 1974, s. 34–55.

60. Rumelhart D. E., Hinton G. E., Williams R. J., *Learning representations by back-propagating errors*, Nature Wyd. 323, 1986, s. 533–536.
61. Shahirah J., Junainah M., Suriatini I., *Machine learning for property price prediction and price valuation: a systematic literature review*, Journal of the Malaysian Institute of Planners Wyd. 19 Nr. 3, 2021, s. 411–422.
62. Shen H., Li L., Zhu H., Liu Y., Luo Z., *Exploring a Pricing Model for Urban Rental Houses from a Geographical Perspective*, Land Wyd. 11 Nr. 1, 2022.
63. Taruttis L., Weber C., *Estimating the impact of energy efficiency on housing prices in Germany: Does regional disparity matter?*, Energy Economics Wyd. 105, 2022.
64. Triebel O. i inn, *NeuralProphet: Explainable Forecasting at Scale*, 2021, <https://arxiv.org/abs/2111.15397> (dostęp 07.08.2023).
65. Triebel O., Laptev N., Rajagopal R., *AR-Net: A simple Auto-Regressive Neural Network for time-series*, 2019, <https://arxiv.org/abs/1911.12436> (dostęp 28.08.2023).
66. Trojanek R., Gluszak M., *Short-run impact of the Ukrainian refugee crisis on the housing market in Poland*, Finance Research Letters Wyd. 50, 2022.
67. Tobler W., *A computer movie simulating urban growth in the Detroit region*, Economic Geography Wyd. 46, 1970, s. 234–240.
68. Tomal M., *The Impact of Macro Factors on Apartment Prices in Polish Counties: A Two-Stage Quantile Spatial Regression Approach*, Real Estate Management and Valuation Wyd. 27 Nr. 4, 2019, s. 1-14.
69. Truong Q., Nguyen M., Dang H., Mei B., *Housing Price Prediction via Improved Machine Learning Techniques*, IIKI2019, 2019, s. 1-5.
70. Vincenty T., *Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations*, Geodetic Survey Squadron, London 1975.
71. Votsis A., *Planning for green infrastructure: The spatial effects of parks, forests, and fields on Helsinki's apartment prices*, Ecological Economics Wyd. 132, 2017.
72. Wang D., Li V. J., *Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review*, Sustainability Wyd. 11 Nr. 24, 2019.
73. Wang Z., Song Y., *Deep learning for the spatial additive autoregressive model with nonparametric endogenous effect*, Spatial Statistics Wyd. 55, 2023.
74. Widlak M., Waszczuk, J., Olszewski, K., *Spatial and Hedonic Analysis of House Price Dynamics in Warsaw*, National Bank of Poland Working Paper Nr. 197, 2015.

75. Zhang W. i inn, *Parallel distributed processing model with local space-invariant interconnections and its optical architecture*, Applied Optics Wyd. 29 Nr. 32, 1990, s. 4790-4797.
76. Zhao C., *Multiscale Effects of Hedonic Attributes on Airbnb Listing Prices Based on MGWR: A Case Study of Beijing, China*, Sustainability Wyd. 15 Nr. 2, 2023.
77. Zhao Q, Hastie T., *Causal Interpretations Of Black-Box Models*, J Bus Econ Stat, 2019, s. 1080.
78. Ziyue Y., Lu Z., *Spatial Prediction of Housing Prices in Beijing Using Machine Learning Algorithms*, HPCCT & BDAI '20, 2020, s. 64–71.
79. Private Real Estate Common Position on a European Pillar of Social rights, <https://ec.europa.eu/social/BlobServlet?docId=17437> (dostęp 12.08.2023).
80. <https://architektura.um.warszawa.pl/geodezja> (dostęp 15.08.2023).
81. <https://ec.europa.eu/social/main.jsp?catId=1567> (dostęp 12.08.2023).
82. https://gadm.org/download_country36.html (dostęp 15.08.2023).
83. <https://www.olx.pl/nieruchomosci/mieszkania/> (dostęp 15.08.2023).
84. <https://www.otodom.pl/pl/wyniki/sprzedaz/mieszkanie/mazowieckie/warszawa/warszawa/warszawa> (dostęp 15.08.2023).
85. <https://www.openstreetmap.org/relation/336075> (dostęp 15.08.2023).
86. <https://rednetconsulting.pl/aktualnosci,tresc,id,10504,tytul,sytuacja-na-rynkumieszkaniowym-iv-kwartal-2022> (dostęp 15.08.2023).

Spis rysunków

Rysunek 1. Taksonomia używanych zmiennych objaśniających wraz z przykładami	15
Rysunek 2. Drzewo asymetryczne i symetryczne	29
Rysunek 3. Architektura geograficznie ważonej sztucznej sieci neuronowej	34
Rysunek 4. Umiejscowienie geograficzne dzielnic Warszawy	41
Rysunek 5. Rozkłady zmiennych <i>bike_parkings_800, restaurants_800, food_shops_800</i>	43
Rysunek 6. Wykres konturowy w przypadku zależności liniowej	44
Rysunek 7. Rozkłady zmiennych liczące obiekty w pobliżu 800 metrów	44
Rysunek 8. Rozkłady zmiennych liczące odległości minimalne	45
Rysunek 9. Rozkłady zmiennych odległości minimalnych oraz obiektów w pobliżu 800 metrów ...	47
Rysunek 10. Rozkłady zmiennych area, rent, build_year	47
Rysunek 11. Kolory odpowiednich percentyle	48
Rysunek 12. Rozkład ceny metra kwadratowego mieszkania po transformacji logarytmicznej	48
Rysunek 13. Rozkład zmiennych charakteryzujących dzielnice centralne	49
Rysunek 14. Rozkład zmiennych charakteryzujących Śródmieście	50
Rysunek 15. Rozkład zmiennych wyróżniających lokalne klastry przestrzenne	51
Rysunek 16. Rozkłady zmiennych charakteryzujących obszary mniej rozwinięte	51
Rysunek 17. Rozkłady zmiennych z najbardziej klarownym wzorcem przestrzennym	51
Rysunek 18. Rozkłady zmiennych z unikalnymi wzorcami przestrzennymi	52
Rysunek 19. Rozkłady zmiennych bez zależności geograficznych	53
Rysunek 20. Rozkłady zmiennych wyeliminowanych ze zbioru	53
Rysunek 21. Rozkłady i testy statystyczne dla pierwszej grupy zmiennych kategorialnych	55
Rysunek 22. Rozkłady i testy statystyczne dla drugiej grupy zmiennych kategorialnych	56
Rysunek 23. Rozkłady i testy statystyczne dla zredukowanych zmiennych	57
Rysunek 24. Korelacje wszystkich kategorialnych regresorów	57
Rysunek 25. Zmienne zero-jedynkowe opisujące wnętrze mieszkań	58
Rysunek 26. Zmienne zero-jedynkowe opisujące zewnętrze mieszkań	59
Rysunek 27. Zmienne zero-jedynkowe opisujące czynniki globalne	60
Rysunek 28. Zmienne kategorialne opisujące faktory globalne	60
Rysunek 29. Zmienne uporządkowane opisujące faktory globalne na mapie Warszawy	61
Rysunek 30. Rozkład zmiennej objaśnianej po zastosowaniu agregacji	62
Rysunek 31. Punktowy wykres Morana	64

Rysunek 32. Reszty z modelu regresji liniowej oraz reszty z prognozowania za pomocą medialnej i średniej wartości	69
Rysunek 33. Reszty z autoregresyjnych modeli przestrzennych.....	70
Rysunek 34. Reszty z Bayesowskich autoregresyjnych modeli przestrzennych.....	71
Rysunek 35. Reszty z modelu GWR	73
Rysunek 36. Reszty z modeli uczenia maszynowego.	74
Rysunek 37. Średnia częstość występowania zmiennych objaśniających w estymatorach bazowych modeli zespołowych oraz średnia wartość p w modelach ekonometrycznych	76

Spis tabeli

Tabela 1. Taksonomia statystycznych modeli przestrzennych.....	20
Tabela 2. Funkcję błędu.....	37
Tabela 3. Wykorzystane lokalizacje geograficzne	39
Tabela 4. Wyniki testów autokorelacji przestrzennej dla reszt z modelu liniowego	63
Tabela 5. Wyniki strojenia hiperparametrów	65
Tabela 6. Kryterium porównawcze	67
Tabela 7. Wartości współczynników β w modelu GWR.....	77

Streszczenie

Przedmiotem niniejszej pracy jest dokonanie porównania podejścia ekonometrycznego a podejścia opartego na uczeniu maszynowym w kontekście prognozowania cen mieszkań na przykładzie miasta Warszawa oraz ustalenie obecnie najlepszej metody prognozowania. Struktura pracy obejmuje następujące elementy: w Rozdziale I zaprezentowano podstawy teoretyczne i dokonano przeglądu literatury; Rozdział II zawiera szczegółowe opisanie modeli oraz kryterium oceny; w Rozdziale III omówiono proces doboru zmiennych, inżynierię cech oraz analizę eksploracyjną danych; w Rozdziale IV dokonano dyskusji i empirycznej analizy wyników, a na zakończenie przedstawiono główne wnioski, omówiono ograniczenia pracy oraz wyznaczono kierunki dalszych badań. Na podstawie analizy stwierdzono, że optymalnym modelem do prognozowania cen mieszkań w Warszawie jest model geograficznie ważonej regresji GWR. Modele oparte na uczeniu maszynowym wykazywały się mniejszą skutecznością, a czasem nawet gorszymi wynikami od prostych modeli, takich jak model regresji liniowej (model hedoniczny). Najsłabsze dopasowanie obserwowano w przypadku perceptronu wielowarstwowego oraz przestrzennej sieci neuronowej.