AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ ZARZĄDZANIA

KATEDRA INFORMATYKI BIZNESOWEJ I INŻYNIERII ZARZĄDZANIA

Projekt dyplomowy

Analiza Czynników Wpływających na Frekwencję Wyborczą w Polsce

Analysis of Factors Influencing Voter Turnout in Poland

Autor: *Jakub Anczyk*
Kierunek studiów: Informatyka i Ekonometria
Opiekun pracy: *Łukasz Lach, dr hab.*

Kraków, 2024

Spis treści

1 Wprowadzenie

1.1 Cel i założenia pracy

Frekwencja wyborcza, definiowana jako "stosunek liczby oddanych głosów (ważnych kart do głosowania) do ogólnej liczby osób uprawnionych do głosowania"[1], stanowi podstawową miarę partycypacji obywateli w życiu politycznym ich kraju i społeczności. Powszechnie przyjmuje się, że wysoka frekwencja wyborcza jest niezbędna dla prawidłowego funkcjonowania procesu demokratycznego. Wynikiem zbyt niskiej frekwencji wyborczej mogą być "nierówność reprezentacji i wpływów [które] nie są losowo rozłożone, ale systematycznie stronnicze na korzyść bardziej uprzywilejowanych obywateli - tych o wyższych dochodach, większym bogactwie i lepszym wykształceniu - oraz przeciwko mniej uprzywilejowanym obywatelom" (Lijphart 1997).

Jak twierdzi International Institute for Democracy and Electoral Assistance, "wyższa frekwencja wyborcza jest w większości przypadków oznaką witalności demokracji, podczas gdy niższa frekwencja jest zwykle związana z apatią wyborców i nieufnością do procesu politycznego" (Solijonov 2016), co prowadzi do powstania rządu niereprezentatywnego, niskiej legitymizacji władzy i marginalizacji mniejszości społecznych.

Z tej przyczyny zarówno w interesie władzy jak i obywateli leży dążenie do tego, aby frekwencja wyborcza w każdych wyborach była na możliwie najwyższym poziomie. Aby umożliwić działania mające na celu zwiększenie aktywizacji politycznej obywateli, należy jednak w pierwszej kolejności zrozumieć czynniki determinujące lub wpływające na frekwencję wyborczą.

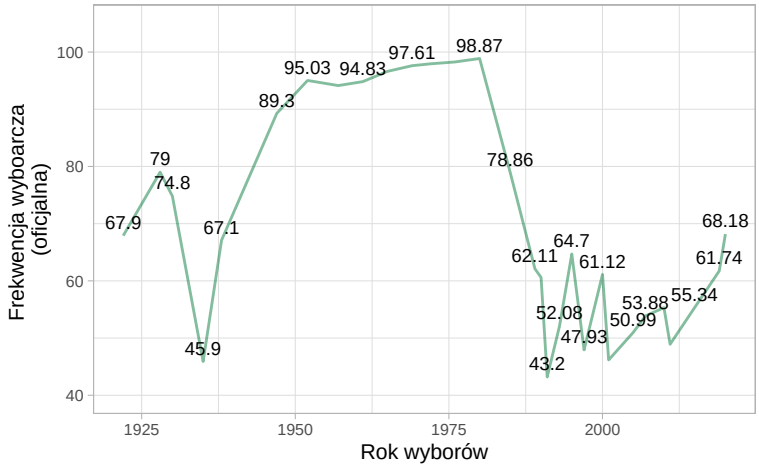
Frekwencja wyborcza jest zjawiskiem intensywnie badanym na Zachodzie, w tym przede wszystkim w Stanach Zjednoczonych, a dostępna literatura proponuje wiele metod jej modelowania. Tymczasem wydaje się, że w Polsce zjawisko to jest badane na znacznie mniejszą skalę w porównaniu do reszty świata zachodniego. Warto jednak zaznaczyć, że ta ocena opiera się głównie na dostępnych publicznie badaniach i nie jest jasne, na ile badania frekwencji są prowadzone na zlecenie partii politycznych lub innych zainteresowanych stron przez ośrodki badań opinii publicznej.

W niniejszej pracy podjęto próbę uzupełnienia tego braku oraz zaproponowano modele objaśniające frekwencję na poziomie województw, które mogą być także wykorzystane do szacowania frekwencji wyborczej na poziomie krajowym. W tym celu przeanalizowano wpływ rozmaitych czynników na frekwencję wyborczą, począwszy od wskaźników ekonomicznych, społeczno-gospodarczych i demograficznych, a skończywszy na wynikach badań opinii publicznej. W pracy przeanalizowano frekwencję wyborczą na poziomie wojewódzkim, oraz uzupełniająco na poziomie krajowym (Polska) i europejskim, począwszy od 1922 roku aż po rok 2023.

[1] Ustawa z dnia 5 stycznia 2011 r. - Kodeks wyborczy (Dz. U. z 2020 r. poz. 1319).

1.2 Wpływ historii frekwencji wyborczej na wybór przedziału czasowego

Pierwsze w Polsce pięcioprzymiotnikowe (powszechne, równe, bezpośrednie, proporcjonalne i tajne) wybory odbyły się 26 stycznia 1919 roku, w rok po odzyskaniu przez Polskę niepodległości. W okresie międzywojennym poziom frekwencji wyborczej był bardzo zróżnicowany - najwyższy oficjalny wynik na poziomie 79.00% odnotowano w 1930 roku, a najniższy na poziomie 45.00% w 1935 roku. Należy przy tym zaznaczyć, że dane te pochodzą z okresu, w którym wybory nie były w pełni demokratyczne (1930) lub zostały przez znaczną część obywateli zbojkotowane (1935), więc zestawianie ich z danymi pochodzącymi z późniejszych okresów byłoby niezasadne.



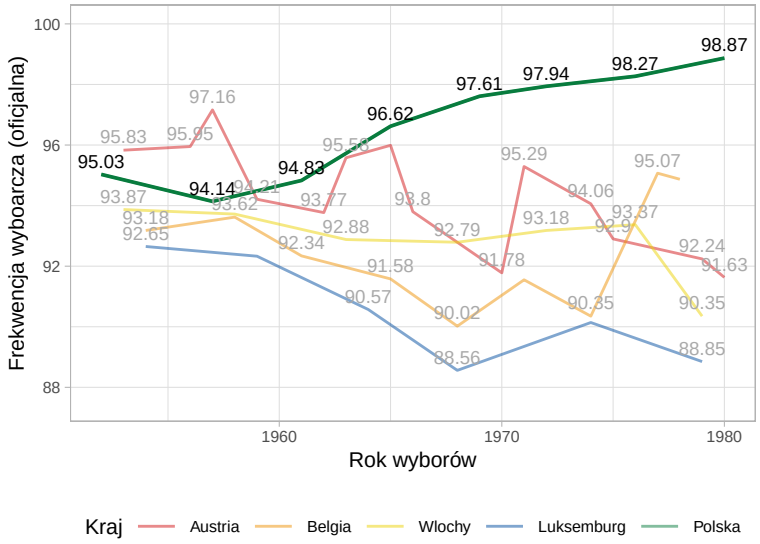
Rysunek 1. Historyczna frekwencja wyborcza w Polsce.

Ponadto, zachowane dane ekonomiczne, społeczno-gospodarcze i demograficzne z okresu między 1919 a 1952 rokiem są szczątkowe lub nie nadają się do przeprowadzenia podobnej do niniejszej analizy. Ponadto nie sposób porównywać dane z okręgów Królestwa Polskiego do aktualnego podziału terytorialnego Polski.

Z pozoru prostsze wydaje się zestawienie współczesnej frekwencji z frekwencją w Polskiej Rzeczpospolitej Ludowej. Od 1952 roku oficjalna frekwencja wyborcza oscylowała raczej w górnych granicach skali (rys. 1.), ze średnią w okolicach 96.66%, osiągając w 1980 roku rekordowy poziom 98.87%. Najniższą frekwencję w tym okresie odnotowano w 1957 roku i wyniosła ona 94.14%. Dla porównania, w tym samym okresie średnia dla państw, w których obowiązywał prawny przymus głosowania (Austria, Belgia, Włochy i Luksemburg) wynosiła jedynie 92.95%, z minimum na poziomie 88.56% i maksimum 97.16%, co ilustruje rysunek 2. Jak podają źródła historyczne (Siedziako 2016), osiągnięcie w Polsce tak wysokiego wyniku bez stosowania metod polegających na manipulacjach oraz fałszerstwach wyborczych, było technicznie niemożliwe.

Należy także zaznaczyć, że definicja głosu ważnego znacznie różniła się od współczesnej, bowiem wrzuconą do urny pustą kartę (bez tzw. “skreśleń”) traktowano jako głos ważny, a głos domyślnie przydzielano kandydatom znajdującym się na początku listy (tj. z góry ustalonym przez partię). Karty wyborcze zawierały wyłącznie jedną listę, co w praktyce uniemożliwiało oddanie głosu na kandydata niewskazanego pośrednio przez władzę, a jakiekolwiek próby rejestracji list wyborczych przez obywateli kończyły się niepowodzeniem z uwagi na rzekome “przeszkody formalne”.

Historycy oceniają, że z pewnością sfałszowano wyniki wyborów w 1952 oraz 1957 roku, co polegało przede wszystkim na sfałszowaniu frekwencji i liczby "skreśleń", przymuszaniu do głosowania, fałszowaniu liczby osób uprawnionych do głosowania i fabrykowaniu protokołów głosowania, czemu towarzyszyła ogólna atmosfera terroru i opresji. Skala fałszerstw kolejnych wyborów nie jest w pełni znana z powodu niedostatecznych badań w tym zakresie, jednak opisane mechanizmy manipulacji i fałszerstw były trwale stosowane w okresie PRL (Siedziako 2016).



Rysunek 2. Historyczna frekwencja wyborcza w Polsce i krajach z prawnym przymusem głosowania.

W związku z nierzetelnymi danymi na temat frekwencji z okresu PRL i niedemokratycznym charakterem procesu wyborczego w tym okresie, pierwszy cykl wyborczy, który może zostać poddany analizie, dotyczy wyborów parlamentarnych w 1989 roku. W praktyce jednak najwcześniejsze dostępne dane z obszarów kluczowych dla tej pracy pochodzą z lat znacznie późniejszych.

Dla przykładu, informacje o gęstości zaludnienia i strukturze ludności dostępne przez witrynę Banku Danych Lokalnych sięgają najdalej 1996 roku. O oczekiwanym trwaniu życia w zdrowiu zaczęto raportować dopiero w 2009 roku, a współczynniki skolaryzacji w szkolnictwie podstawowym i gimnazjalnym dostępne są od 2003 do 2022 roku i są w dużej mierze niekompletne. Z kolei dane o Produkcie Narodowym Brutto dostępne są tylko na poziomie krajowym oraz ponadkrajowym za pośrednictwem Eurostatu.

Większość danych niezbędnych dla przeprowadzenia analizy, w szczególności tych udostępnianych przez Główny Urząd Statystyczny za pośrednictwem Banku Danych Lokalnych (BDL) oraz Dziedzinowych Baz Wiedzy (DBW), jest dostępna począwszy od 2002 roku, w związku z czym wybór badanego przedziału czasowego padł na lata 2005 - 2023. Łącznie w tym okresie odbyło się sześć cykli wyborczych w latach 2005, 2007 (przedterminowo), 2011, 2015, 2019 oraz 2023.

2 Przegląd literatury

Obszerna literatura psefologiczna proponuje liczne modele wyjaśniające frekwencję wyborczą na poziomie krajowym i ponadkrajowym. Badacze koncentrują się jednak w głównej mierze na dużych zachodnich demokracjach o systemie dwupartyjnym - Stanach Zjednoczonych Ameryki (Gomez, Hansford, i Krause 2007), Wielkiej Brytanii (Murr i Munzert 2018) i Francji, oraz w mniejszym stopniu na pozostałych zachodnich demokracjach europejskich. Przyczyną jest względna łatwość w modelowaniu zarówno frekwencji jak i wyników wyborczych, kiedy liczebność partii na scenie politycznej jest niewielka. W przypadku tych systemów frekwencja wyborcza może mieć decydujący wpływ na ostateczny wynik wyborczy, a co za tym idzie, znajomość czynników determinujących frekwencję może dostarczyć partiom politycznym kluczowych danych niezbędnych do optymalnej aktywizacji ich elektoratu.

Literatura odnosząca się do pozostałych krajów skupia się raczej na poszukiwaniu czynników mających wpływ na frekwencją wyborczą (Blais i Dobrzynska 1998), a więc próbuje frekwencję objaśniać, a nie prognozować, lub poszukuje odpowiedzi na pytania "co jest przyczyną niskiej frekwencji wyborczej" oraz "jak podnieść poziom frekwencji wyborczej" skupiając się na pokrewnym, choć odwrotnym do frekwencji zjawisku, którym jest absencja wyborcza (Sasińska-Klas 2008; Musiał-Karg 2011).

Przykładem próby wyjaśnienia zjawiska frekwencji jest zaproponowany przez Blais i Dobrzynska (1998) model objaśniający frekwencję wyborczą w demokracjach elektoralnych, oparty o trzy kategorie czynników, tj. otoczenie socjoekonomiczne, instytucjonalne oraz system partyjny.

Do najważniejszych czynników socjoekonomicznych zaliczono średnią długość życia, gęstość zaludnienia, PNB per capita, wzrost PNB per capita względem roku poprzedniego, wskaźnik analfabetyzmu oraz wielkość populacji. Ponadto, do zmiennych socjoekonomicznych zaliczono szereg zmiennych binarnych, które odzwierciedlają wpływ innych, niemierzonych czynników, specyficznych dla danego obszaru. Zmienne te to odpowiednio Ameryka Północna, Ameryka Południowa, Afryka, Oceania i Szwajcaria, która jest jedynym krajem wśród zmiennych binarnych z uwagi na jej szczególny charakter: średnia frekwencja utrzymuje się w okolicach 46-49% od niemal 50 lat (Jackman 1987), co spowodowane jest częstymi referendami oraz złożonością systemu politycznego (Blais 2014).

W innych badaniach określa się znaczenie konkretnych czynników, takich jak wiek (Cześnik i Zagórski 2022), stabilność poparcia dla partii politycznych (Gendźwiłł, Rutkowski, i Żółtak 2014) lub cech ordynacji wyborczej i polaryzacji sceny politycznej (Najbar 2017) dla frekwencji wyborczej. W żadnym z nich nie podjęto jednak próby modelowania frekwencji na podstawie ich efektów. Pierzgalski (2012) dokonuje takiej próby, lecz na podstawie zaledwie dwóch zmiennych, jakimi są stopa bezrobocia w powiecie oraz przeciętne miesięczne wynagrodzenie brutto w powiecie w relacji do średniej krajowej. Analiza ta obejmowała wszystkie powiaty w Polsce w latach 2007 i 2011, a statystyka R-kwadrat dla obliczonego modelu wyniosła 0.4951.

Należy także zaznaczyć, że wszystkie wymienione czynniki pojawiają się w metaanalizie autorstwa Frank i Coma (2023), dotyczącej predyktorów frekwencji wyborczej, która ukazała się 20 maja 2021 roku w Political Behavior. Ich korelacja z frekwencją jest więc znana i od lat stosowana do jej szacowania.

Poza wyczerpującym modelem zdolnym do szacowania frekwencji ogólnopolskiej, brakuje w dostępnej literaturze także próby wyjaśnienia przyczyn występujących lokalnie różnic we frekwencji wyborczej, tj. między różnymi regionami tego samego kraju. Dotyczy to zarówno wspomnianej wyżej Szwajcarii (Blais 2014), ale również Polski, choć pojawiają się pojedyncze analizy dotyczące konkretnych województw a nawet powiatów (Kowalski 2016). Dostępne analizy skupiają się jednak głównie na wynikach krajowych czerpiąc również ze statystyk zebranych na poziomie krajowym. Być może jednak analiza korelatów frekwencji wyborczej na poziomie lokalnym pomoże objaśnić to zjawisko w większym stopniu także na poziomie krajowym.

3 Metodologia

3.1 Źródła danych

Dane o frekwencji mogą być analizowane na poziomie ogólnopolskim, wojewódzkim, powiatowym, a także na poziomie okręgów. Analiza na poziomie ogólnopolskim nie byłaby jednak spójna z ideą niniejszej pracy, a analiza wyłącznie na poziomie powiatów nie dawałaby pełnego obrazu sytuacji, ponieważ większość zmiennych stosowanych tradycyjnie w modelach objaśniających frekwencję wyborczą nie jest dostępna w podziale na regiony mniejsze od województw. Z kolei poszukiwanie informacji o gospodarce i innych dziedzinach na poziomie okręgów, które znacząco odbiegają od podziału administracyjnego Polski, byłoby w większości przypadków niemożliwe lub skutkowałoby niedokładnymi danymi, a w konsekwencji słabo dopasowanym modelem. Z uwagi na to zadecydowano o przeprowadzeniu niniejszej analizy wyłącznie na poziomie województw.

Dane wykorzystane w analizie pochodzą z następujących źródeł:

1. Bank Danych Lokalnych (BDL), z którego pobrano podstawowe statystyki dla poszczególnych województw,
2. Dziedzinowe Bazy Wiedzy (DBW), z których pobrano dodatkowe, nieujęte w BDL statystyki dla poszczególnych województw,
3. Portal Głównego Urzędu Statystycznego, z którego pobrano informacje o powierzchni województw,
4. Portal Państwowej Komisji Wyborczej, z którego pobrano informacje o frekwencji wyborczej w poszczególnych województwach,
5. Strona "Trendy" Centrum Badania Opinii Społecznej, z której pobrano dane o trendach wybranych wskaźników pochodzące z rozmaitych sondaży realizowanych przez CBOS od lat 90.

Pobrane i wykorzystane w pracy dane znajdują się w folderze "Data" w głównej bibliotece repozytorium GitHub niniejszej pracy[2]. Szczegółowy spis zmiennych wraz z ich źródłem, zakresem czasowym i opisem znajduje się w pliku "primary_variables.xlsx". Zmienne wtórne, które obliczono przy pomocy danych pochodzących z opisanych powyżej źródeł (np. gęstość zaludnienia może zostać obliczona przy pomocy rozmiaru terytorium oraz liczby mieszkańców zamieszkujących owo terytorium) znajdują się w pliku "secondary_variables.xlsx". Kod źródłowy, który posłużył do ekstrakcji danych z portalu Państwowej Komisji Wyborczej znajduje się w pliku "Electoral_Data_PKW.R", a kod źródłowy pozostałej części pracy, w tym obliczenia, analizy i szczegółowe ich wyniki, znajdują się w pliku "Project_Main_v8.qmd".

W celu stworzenia jednolitego zbioru danych do analizy frekwencji na poziomie wojewódzkim pobrano oficjalne dane wyborcze w rozbiciu na województwa i powiaty w latach 2005, 2007, 2011, 2015, 2019 oraz 2023 z portalu Państwowej Komisji Wyborczej. Uzyskano w ten sposób zestaw danych zawierający nazwę województwa, rok wyborów oraz frekwencję wyborczą.

Do danych o frekwencji przyłączono tabelę zawierającą powierzchnię każdego z województw, a następnie zestaw 48 zmiennych pochodzący z portali GUS. Zmienne te opóźniono o rok względem danych o frekwencji. Oznacza to, że analizowane w każdym cyklu wyborczym dane pochodzą

[2] https://github.com/kubbajb/Analysis-of-Factors-Influencing-Voter-Turnout-in-Poland

z roku poprzedzającego, ponieważ w przeciwnym razie dane obejmowałyby także wydarzenia mające miejsce bezpośrednio po wyborach, a tym samym analiza cyklu wyborczego mogłaby zostać wypaczona przez ich wynik. O ile nie powinno to mieć znaczenia dla takich statystyk jak średnia oczekiwana długość życia lub współczynnik skolaryzacji, o tyle może mieć wpływ na takie statystyki jak PKB lub PNB, które mogą zostać zaburzone na skutek reakcji inwestorów (jeżeli, na przykład, wybory wygra opcja liberalna gospodarczo, to inwestorzy mogą chcieć zwiększyć inwestycje w Polską gospodarkę, jeśli jednak wygra opcja konserwatywna, to inwestorzy zagraniczni mogą zacząć wycofywać fundusze z kraju, spodziewając się dodatkowych utrudnień takich jak cła, podatki i niekorzystne obostrzenia regulacyjne).

Do powstałej tabeli przyłączono zestaw 37 zmiennych pochodzących z sondaży realizowanych przez CBOS, który zawiera informację o średnim poziomie badanych wskaźników w 12 miesiącach poprzedzających dany cykl wyborczy. Uzyskany w ten sposób zbiór 88 zmiennych (włącznie z frekwencją, województwem oraz rokiem) poddano dalszej analizie.

Powstały zbiór podzielono na zbiór testowy i treningowy przy użyciu losowego próbkowania oryginalnego zbioru w proporcjach 7:3 (70% obserwacji posłużyło za zbiór treningowy, a 30% za zbiór testowy). Po podzieleniu zbioru na podzbiory zweryfikowano, czy każdy z nich zawiera pełen zestaw 16 województw, niezbędny do prawidłowej diagnostyki modeli. Losowe próbkowanie powtórzono aż do uzyskania zadowalającej dystrybucji województw w obu zbiorach z wartością ziarna generatora liczb losowych wynoszącą 413473.

3.2 Przetwarzanie danych

W pierwszej kolejności podjęto próbę rekonstrukcji modelu objaśniającego frekwencję wyborczą w województwach na podstawie modelu opisanego przez Blais i Dobrzynska (1998). W niektórych przypadkach niezbędne było zastosowanie pokrewnych zmiennych z uwagi na to, że nie dokonuje się ich pomiaru na poziomie wojewódzkim (np. stopień stopień alfabetyzmu lub PNB per capita). Tak stworzony model poddano analizie statystycznej, w wyniku której zredukowano liczbę zmiennych z sześciu do trzech z powodu braku dowodów na ich istotność. Zredukowany model ponownie poddano analizie uzyskując zadowalający współczynnik R-kwadrat oraz wskaźniki RMSE, MAE, a w wyniku testu Breuscha-Pagana także przesłanki do przyjęcia hipotezy o nierówności wariancji reszt. Podjęto następnie próbę stworzenia udoskonalonych modeli mogących objaśnić frekwencję wyborczą przy pomocy zebranych wcześniej danych.

Wybranie 6 zmiennych spośród 85 jest możliwe na ponad 437 milionów sposobów, a wybranie większej liczby zmiennych powoduje szybki wzrost złożoności obliczeniowej. W celu redukcji liczby kombinacji zmiennych do przetestowania zadecydowano o zastosowaniu metody klastrowania metodą k-średnich. Uzyskany zestaw danych podzielono na 6 klastrów, zawierających zmienne o podobnej wariancji. Analizę przeprowadzono również dla liczby klastrów od 7 do 10, jednak powstałe w ten sposób modele nie charakteryzowały się akceptowalnymi parametrami. Uzyskanych 6 klastrów zawierało kolejno 15, 3, 3, 14, 13 oraz 5 zmiennych, dając 122850 sposobów na ich połączenie w model składający się z sześciu zmiennych objaśniających, po jednej z każdego klastra. Zredukowano w ten sposób liczbę kombinacji z ponad 437 milionów do około 123 tysięcy.

Na każdym z 122850 potencjalnych modeli przeprowadzono zestaw testów statystycznych, mierząc je pod kątem R-kwadrat, maksymalnej istotności zmiennych, współczynnika inflacji wariancji (VIF), p-value testu Shapiro-Wilka na normalność reszt oraz p-value testu Breuscha-Pagana na homoskedastyczność. W wyniku tego procesu powstała tabela zawierająca wszystkie kombinacje zmiennych oraz wartości wymienionych statystyk dla każdej z kombinacji.

Spośród zbadanych modeli wybrano te, których statystyki były na najlepszym poziomie, tj. R-kwadrat na poziomie co najmniej 0.93, najwyższe p-value dla testu istotności zmiennych na poziomie poniżej 0.04, współczynnik inflacji wariancji na (VID) poziomie poniżej 4, p-value testu Shapiro-Wilka normalności reszt na poziomie przynajmniej 0.65 oraz p-value testu Breuscha-Pagana na homoskedastyczność na poziomie przynajmniej 0.35. W efekcie uzyskano cztery konkurencyjne modele o bardzo podobnych statystykach i podobnej selekcji zmiennych. Wśród wybranych modeli znalazły się jednak dwa, które zawierały aż po 3 zmienne pochodzące z badań opinii społecznej autorstwa CBOS. Ponieważ zmienne uzyskane z badań CBOS są stałe na poziomie całego kraju, podjęto decyzję o odrzuceniu modeli zawierających więcej niż dwie zmienne ankietowe, finalnie otrzymując dwa modele o numerach kombinacji N95425 oraz N113549, odznaczające się najlepszymi wynikami. Dalsza diagnostyka modeli obejmowała walidację krzyżową, walidację "poza próbą" oraz wykonanie testów na homoskedastyczność i współliniowość.

3.3 Ograniczenia metodologiczne

Wybrana metodologia umożliwiła realizację założeń niniejszej pracy, niemniej należy zwrócić uwagę na pewne ograniczenia i potencjalne źródła stronniczości, które mogą wpływać na interpretację wyników.

3.3.1 Potencjalnie pominięte zmienne objaśniające

Pierwszym ograniczeniem jest problem potencjalnie pominiętych zmiennych objaśniających. Praca oparta jest głównie na danych statystycznych GUS oraz trendach wskaźników z sondaży CBOS. Pomimo starań, aby wstępny zbiór zmiennych był jak najbardziej reprezentatywny, istnieje ryzyko, że kluczowe zmienne, dostępne zarówno w GUS, CBOS, jak i innych źródłach, mogły zostać pominięte w analizie.

Dodatkowo, w badaniach społecznych, takich jak analiza partycypacji wyborczej, istnieje wiele czynników wpływających na wyniki, które mogą być niedostępne z powodu trudności w ich pomiarze lub wysokich kosztów badań sondażowych. Rozważenie powtórzenia analizy z szerszym zestawem danych i uwzględnienie dodatkowych zmiennych mogłoby pozwolić na bardziej kompleksowe zrozumienie badanego zjawiska.

3.3.2 Ograniczenia modeli liniowych

Kolejne ograniczenie wynika z zastosowania regresji liniowej, której założenia obejmują liniową zależność między predyktorem a zmienną objaśnianą, homoskedastyczność, brak korelacji reszt, normalność rozkładu reszt oraz niezależność predyktorów. Ograniczenia te mogą uniemożliwić

prawidłową budowę modelu, jeśli związek między zmiennymi jest nieliniowy lub jeśli model nie spełnia któregokolwiek z założeń regresji liniowej.

Ponadto, modele liniowe są podatne na nadmierne dopasowanie i charakteryzują się podatnością na wartości odstające, co może prowadzić do zniekształcenia wyników analizy. Z założeń regresji liniowej wynika także problem ekstrapolacji, który pojawia się w momencie przenoszenia uzyskanej zależności na zakres danych, który nie był brany pod uwagę przy tworzeniu modelu. Oznacza to, że uzyskany model może nie działać prawidłowo poza zakresem czasowym, dla którego został stworzony, zwłaszcza gdy charakter którejkolwiek z wykorzystanych zależności okaże się nie być liniowy.

3.3.3 Ograniczenia badań sondażowych

Ważnym ograniczeniem są również potencjalne wady badań sondażowych CBOS. Należy zwrócić uwagę na błąd próby, który może wynikać z braku reprezentatywności próby względem populacji, na przykład z powodu wykluczenia osób bez dostępu do technologii używanej w badaniu lub osób odmawiających udziału w badaniu.

Kolejnym ograniczeniem jest błąd pomiaru, który przejawia się w niedokładności lub nieprawidłowości odpowiedzi respondentów. Należy również rozważyć ograniczenia narzędzi badawczych, które mogą nie uwzględniać pełnej złożoności opinii lub prowadzić do problemów z interpretacją pytań ankietowych. Ponadto, efekty samoświadomości i społecznej pożądliwości mogą prowadzić do udzielania niepełnych lub zniekształconych odpowiedzi przez respondentów.

3.3.4 Możliwości występowania zmian strukturalnych

Istotnym ograniczeniem jest możliwość występowania zmian strukturalnych, takich jak zmiany w strukturze badanych zjawisk, organizacji czy społeczności, które mogą wpływać na badane zjawisko lub zmienne objaśniające. Takie zmiany mogą wpłynąć na replikację analizy, zniekształcić jej wyniki lub wymagać ponownego obliczenia modeli, a w najgorszym przypadku mogą uniemożliwić zastosowanie modeli liniowych w przyszłości.

3.3.5 Pozostałe ograniczenia i uwagi

W niniejszej pracy wielokrotnie zastosowano generatory liczb losowych, w tym w kontekście klastrowania metodą k-średnich oraz podziału zbioru danych na zbiory treningowy i testowy. Z powodu ograniczeń technicznych, proces losowania liczb losowych był powtarzany tylko do momentu uzyskania pierwszej wartości ziarna, która umożliwiała spełnienie podstawowych założeń zastosowanej techniki. Istnieje jednak ryzyko, że kształt analizy oraz jej wyniki mogą być uzależnione od losowo wybranych wartości początkowych w tych procesach. W związku z tym, wnioski wyciągnięte z tak przeprowadzonej analizy mogą ulec zmianie w zależności od doboru wartości początkowych.

Dodatkowo, ze względu na niewielką próbkę danych, obejmującą jedynie 6 cykli wyborczych w okresie od 2005 do 2023 roku oraz 16 województw, analiza obejmowała łącznie 96 obserwacji.

Aby wnioski z analizy zachowały swoją trafność w przyszłości, zaleca się powtórzenie analizy oraz ponowne dopasowanie modeli wraz z uwzględnieniem kolejnych cykli wyborczych. Taki krok powinien przyczynić się do zwiększenia dokładności prognoz oraz poprawy parametrów modeli.

4 Budowa i analiza modeli regresyjnych

W niniejszym rozdziale podjęto próbę zbudowania modelu opartego na przeglądzie literatury oraz opisano jego wady. Następnie zaproponowano dwa alternatywne modele, które wykorzystują podobne typy zmiennych oraz dodatkowo zmienne ankietowe. Mają one na celu dokładniejsze wyjaśnienie frekwencji wyborczej w Polsce na poziomie województw, w porównaniu do modelu autorstwa Blais i Dobrzyńskiej.

Warto podkreślić, że model Blais i Dobrzyńskiej nie jest przeznaczony do szacowania frekwencji wyborczej na poziomie jednostek terytorialnych, takich jak województwa. Został on opracowany do badania zróżnicowania frekwencji wyborczej na świecie, jednak wciąż jest jednym z najbardziej znanych modeli stosowanych w tej dziedzinie.

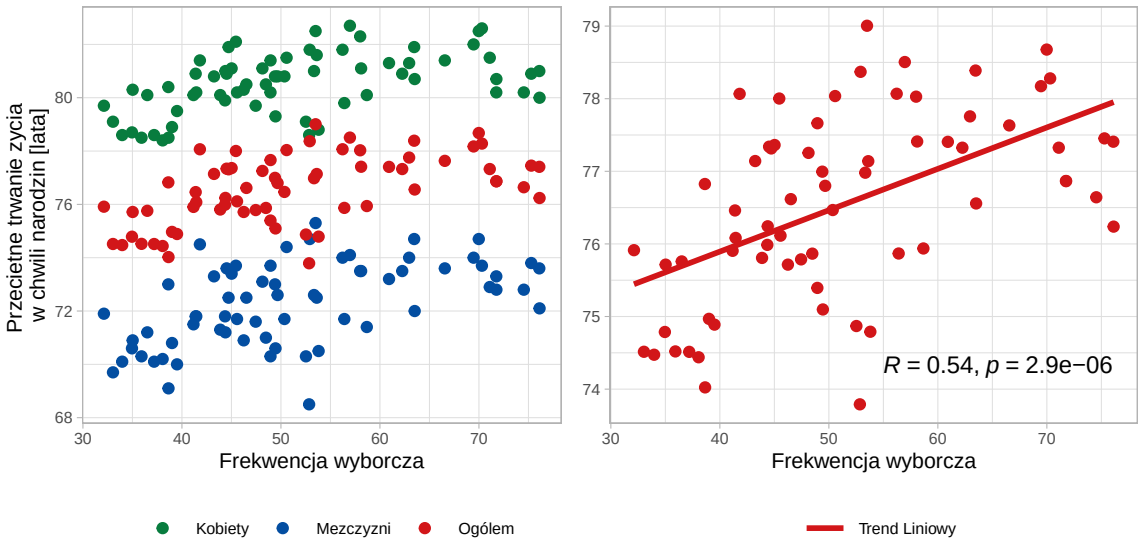
4.1 Model bazujący na modelu autorstwa Blais & Dobrzyńskiej

W modelu autorstwa Blais i Dobrzynska (1998), po usunięciu zbędnych zmiennych związanych z położeniem geograficznym oraz otoczeniem instytucjonalnym, wzięto pod uwagę szereg przytoczonych wyżej czynników socjoekonomicznych, do których należą średnia długość życia, gęstość zaludnienia, PNB per capita, wzrost PNB per capita względem roku poprzedniego, wskaźnik analfabetyzmu oraz wielkość populacji.

Znalezienie danych dla każdego z tych wskaźników na poziomie krajowym jest względnie proste, nawet w przypadku niewielkich i odległych państw. Niestety, pomiarów niektórych z tych wskaźników nie dokonuje się na poziomie regionalnym (np. produkt narodowy brutto per capita lub wskaźnik analfabetyzmu), w związku z czym przy odtwarzaniu tego modelu zastąpiono je zmiennymi pokrewnymi. Ponadto ponownie obliczono jego parametry metodą najmniejszych kwadratów, ponieważ odtwarzany model nie zawiera tej samej selekcji zmiennych, co oryginalny model, a zatem zastosowanie oryginalnych wartości parametrów nie byłoby zasadne.

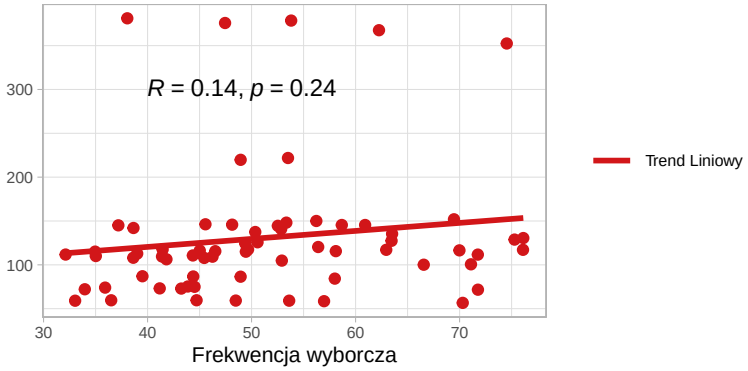
W opisanym modelu znalazły się ostatecznie następujące zmienne:

1. **Przeciętne dalsze trwanie życia w chwili narodzin**, liczone w latach, obliczone zostało na podstawie analogicznych zmiennych dla kobiet oraz mężczyzn, z wagą równą współczynnikowi feminizacji. Przeciętne dalsze trwanie życia dla kobiet oraz mężczyzn połączono z uwagi na bardzo wysoki poziom korelacji tych zmiennych wynoszący 0.8734. Powstała w ten sposób zmienna charakteryzuje się umiarkowanym poziomem korelacji z frekwencją wyborczą na poziomie 0.5360 (rys. 3.).



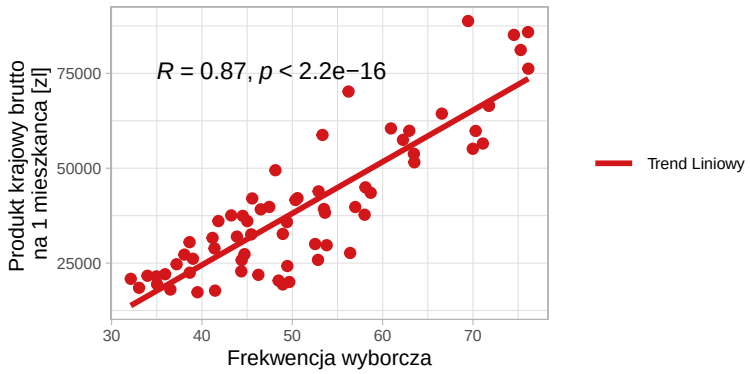
Rysunek 3. Frekwencja wyborcza a przeciętne dalsze trwanie życia [lata].

2. **Gęstość zaludnienia (liczba osób przypadająca na kilometr kwadratowy)**, obliczona jako iloraz liczby ludności zamieszkującej dane województwo i jego powierzchni. Zmienna ta bardzo słabo koreluje z frekwencją wyborczą ze współczynnikiem korelacji Pearsona wynoszącym 0.1444 (rys. 4.).



Rysunek 4. Frekwencja wyborcza a gęstość zaludnienia.

3. **Produkt krajowy brutto na 1 mieszkańca [zł]** jako niedoskonały odpowiednik zmiennej Produkt Narodowy Brutto per capita z przytoczonego wyżej modelu (Blais 2014). Różnica między tymi wskaźnikami polega na tym, ze PNB jest uzupełniony o saldo dochodów z własności za granicą. W przypadku tego modelu zastosowano PKB, ponieważ informacje o PNB nie są dostępne dla regionów i województw. Produkt Krajowy Brutto na 1 mieszkańca silnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona na poziomie 0.8705 (rys. 5.).



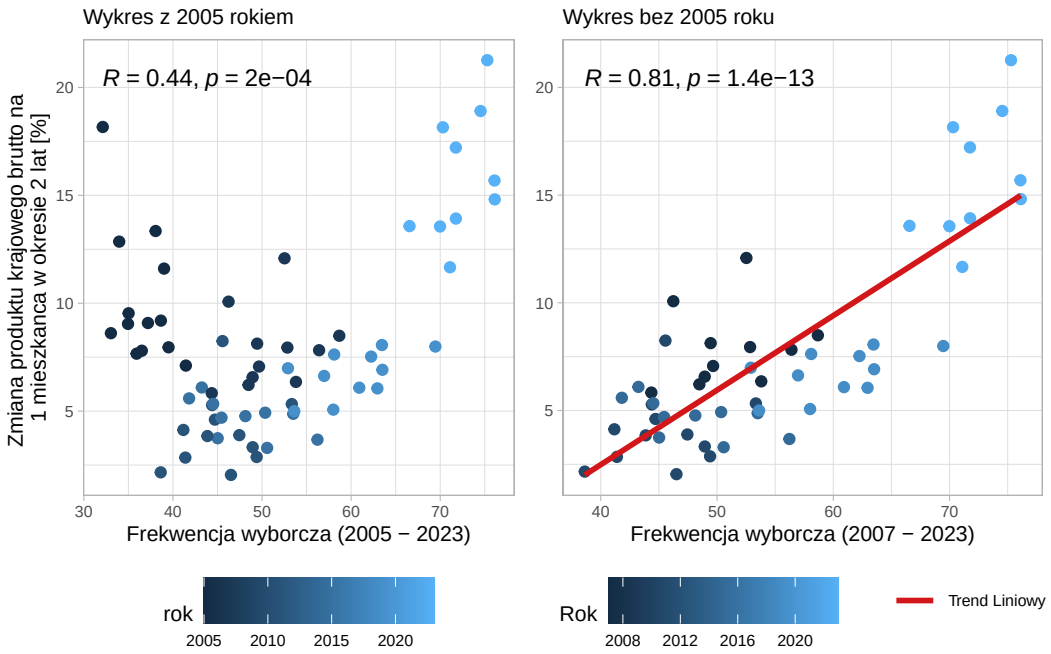
Rysunek 5. Frekwencja wyborcza a produkt krajowy brutto (PKB) na 1 mieszkańca [zł].

4. **Zmiana produktu krajowego brutto na 1 mieszkańca względem poprzedniego roku**, obliczona na jako stosunek produktu krajowego brutto danego roku względem produktu krajowego brutto w ubiegłym roku, wskazuje na dynamikę rozwoju regionu względem minionych lat, a zatem powinna korelować na przykład z jakością usług, standardem życia i wzrostem liczby nowo utworzonych miejsc pracy, co może mieć wpływ na frekwencję wyborczą.

   Podczas analizy zaobserwowano, że wykres rozrzutu zaburzony jest przez obserwacje z 2004 roku, co może wskazywać na rozbieżne trendy zachowań wyborczych w okresie od 2005 roku. Frekwencja w roku 2005 była na względnie niskim poziomie, choć zmiana PKB w okresie 2 lat była znaczna (maksymalny wzrost PKB wyniósł 18.60% w przypadku województwa opolskiego, ze średnią krajową na poziomie 10.43%).

   Od 2005 roku we wszystkich pięciu wyborach parlamentarnych i w każdym województwie zachowanie wyborców podlegało już spójnemu trendowi: im wyższy odnotowano wzrost gospodarczy, tym wyższa była frekwencja. Przytoczona zmiana może być spowodowana dołączeniem Polski do Unii Europejskiej 1 maja 2004 roku oraz nasileniem polaryzacji w Polskim społeczeństwie w następstwie nieudanej koalicji rządowej między Platformą Obywatelską a Prawem i Sprawiedliwością w 2005 roku, na co wskazuje Zagała (2023). Dokładny powód takiej zależności jest jednak niejasny i wymaga dalszych badań, aby lepiej zrozumieć jej mechanizmy.

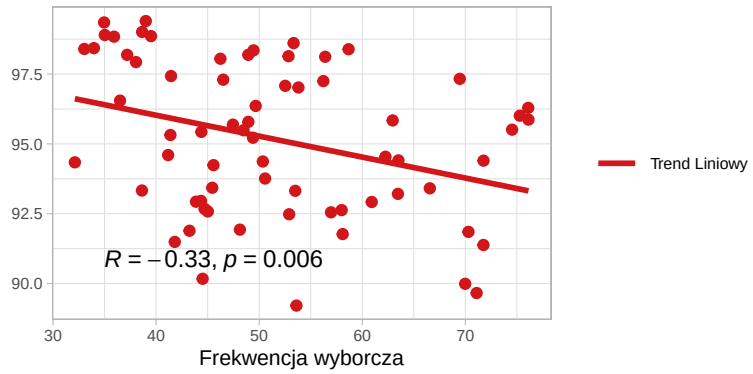
   Współczynnik korelacji Pearsona frekwencji ze zmianą PKB w porównaniu do roku poprzedniego wynosi 0.4396 przy ujęciu danych z 2005 roku i 0.8084 po ich wykluczeniu (rys. 6.). W niniejszym przykładowym modelu ujęto jednak wszystkie cykle wyborcze, włącznie z 2005 rokiem.



Rysunek 6. Frekwencja wyborcza a zmiana produktu krajowego brutto (PKB) na 1 mieszkańca [zł] z ujętym oraz nieujętym 2005 rokiem.

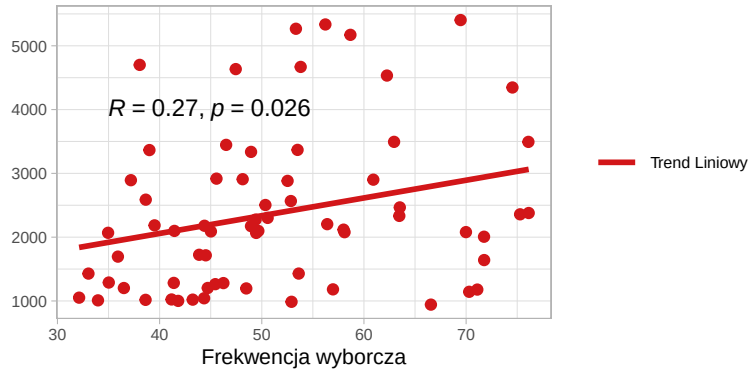
5. **Współczynnik skolaryzacji netto dla szkół podstawowych** włączono z uwagi na pokrewieństwo ze wskaźnikiem analfabetyzmu, którego pomiaru nie dokonuje się w Polsce w podziale na regiony z uwagi na wysoki stopień piśmienności Polaków. Według oficjalnych statystyk Banku Światowego, stopień alfabetyzmu Polaków utrzymuje się na poziomie powyżej 99.3 od 1988 roku[3]. Z tego powodu zarówno dokonywanie pomiarów analfabetyzmu, jak również szacowanie na jego podstawie frekwencji wyborczej może w przypadku Polski nie przynieść oczekiwanego efektu. W modelu ujęto zatem współczynnik skolaryzacji netto dla szkół podstawowych, który określa stosunek liczby wszystkich osób uczących się na danym poziomie do całej populacji osób będących w wieku nominalnie przypisanym temu poziomowi kształcenia. Współczynnik skolaryzacji netto dla szkół podstawowych koreluje z frekwencją wyborczą ujemnie na poziomie -0.3325 (rys. 7.).

[3]Wskaźnik alfabetyzacji osób dorosłych (udział osób w wieku 15 lat i więcej), UNESCO Institute for Statistics (UIS), UIS.Stat Bulk Data Download Service, dostęp 24 kwietnia 2024, https://data.worldbank.org/indicator/SE.ADT.LITR.ZS.



Rysunek 7. Frekwencja wyborcza a współczynnik skolaryzacji netto dla szkół podstawowych.

6. **Ludność w tysiącach**, oznaczająca liczbę mieszkańców danego województwa, korelująca z frekwencją wyborcą na poziomie 0.2713 (rys. 8.).



Rysunek 8. Frekwencja wyborcza a ludność w tysiącach.

Zrekonstruowany w ten sposób model charakteryzuje się współczynnikiem R-kwadrat na poziomie 0.7581. Pomimo względnie wysokiego współczynnika determinacji, związek części ujętych w modelu zmiennych ze zmienną objaśniana jest statystycznie nieistotny z bardzo wysokimi p-values. Wyniosły one kolejno 0.9854 dla przeciętnego dalszego trwania życia w chwili narodzin, 0.9608 dla gęstości zaludnienia, 0.9371 dla liczby ludności oraz 0.4885 dla współczynnika skolaryzacji. Usunięcie tych zmiennych nieznacznie poprawia współczynnik determinacji do poziomu 0.7670 i skutkuje skromniejszym modelem złożonym z zaledwie trzech zmiennych, którymi są przeciętne trwanie życia, produkt krajowy brutto na 1 mieszkańca oraz zmiana produktu krajowego brutto na 1 mieszkańca w ciągu jednego roku.

Wyniki walidacji modelu na zbiorze testowym wskazują na bardzo podobną wartość R-kwadrat na poziomie 0.7776 z pierwiastkiem błędu średniokwadratowego wynoszącym 6.0712% i średnim błędem absolutnym na poziomie 4.6615%. Niestety, na podstawie wynik testu Breuscha-Pagana (BP = 7.8307, df = 3, p-value = 0.04964) należy odrzucić hipotezę zerową o homoskedastyczności pomimo zadowalających wyników testu t-Studenta na normalność reszt (W = 0.98449, p-value = 0.5706), pozwalającego przyjąć hipotezę zerową o normalności reszt modelu.

4.2 Udoskonalone modele N95425 i N113549

Ponieważ zmodyfikowany model opierający się na modelu Blais i Dobrzynska (1998), składający się z zaledwie trzech, a właściwie z dwóch zmiennych (w tym jednej opóźnionej - produkt krajowy brutto na 1 mieszkańca), jest w stanie wyjaśnić niemal 80% zmienności we frekwencji wyborczej, podjęto próbę stworzenia udoskonalonych modeli przy pomocy znanych korelatów frekwencji wyborczej oraz szeregu zmiennych pochodzących z badań opinii publicznej ośrodka CBOS.

Metodologia obejmowała redukcję kombinacji zmiennych z 437 milionów do około 123 tysięcy wykorzystując technikę klastrowania, a następnie ocenę modeli pod kątem R-kwadrat, istotności zmiennych, współczynnika inflacji wariancji (VIF), p-value testu Shapiro-Wilka na normalność reszt oraz p-value testu Breuscha-Pagana na homoskedastyczność. W wyniku tego procesu zidentyfikowano cztery konkurencyjne modele, z których dwa (N95425 i N113549) zawierały najmniejszą ilość zmiennych stałych na poziomie krajowym. Modele N95425 i N113549 wybrano z uwagi na większe większe zróżnicowanie danych wejściowych na poziomie wojewódzkim.

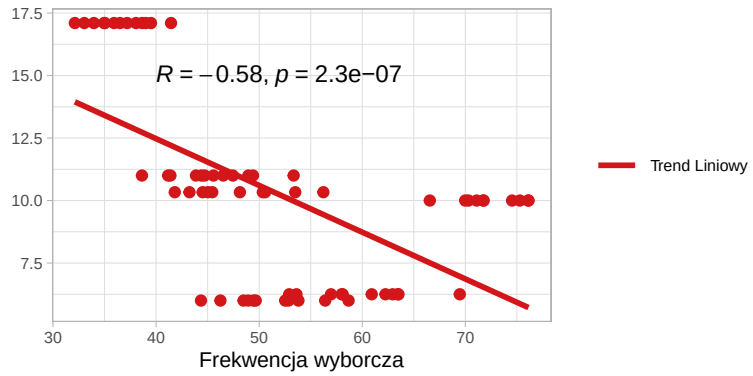
4.2.1 Opis modelu N95425

Pośród zmiennych modelu N95425 znalazły się dwie zmienne ankietowe pochodzące z badań opinii publicznej ośrodka CBOS, trzy zmienne ekonomiczne oraz jedna zmienna demograficzna:

1. Zmienne ankietowe

   1.1. **Udział osób deklarujących się jako przeciwnicy integracji z Unią Europejską** - zmienna reprezentuje średni procent osób, które w 12 miesiącach poprzedzających dane wybory parlamentarne, w ankietach CBOS pod tytułem “Stosunek do członkostwa Polski w Unii Europejskiej”, spośród dostępnych deklaracji “zwolennicy”, “przeciwnicy” oraz “niezdecydowani”, zadeklarowały się jako przeciwnicy członkostwa Polski w Unii Europejskiej.

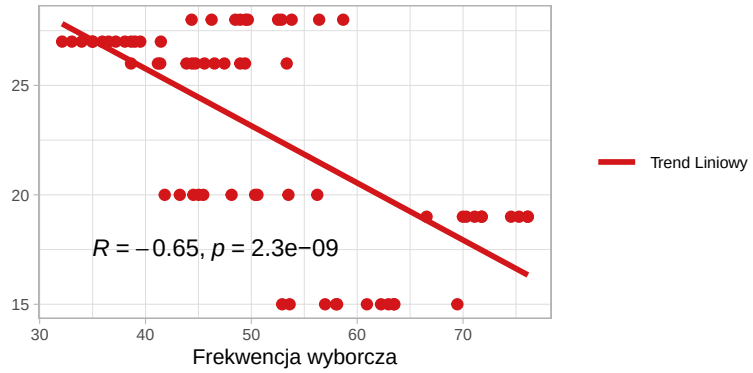
   Zmienna ta umiarkowanie i odwrotnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym -0.5826 (rys. 9.). Test istotności t-Studenta (p-value = 1.33×10^{-14}) wskazuje na istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około -1.2485, co oznacza, że wzrost odsetka przeciwników integracji z Unią Europejską o 10% powoduje spadek frekwencji wyborczej w województwie o około 12.49% (ceteris paribus). Województwa z wyższym udziałem przeciwników integracji z Unią Europejską odnotowują niższą frekwencję wyborczą w porównaniu do tych, gdzie udział ten jest mniejszy.



Rysunek 9. Frekwencja wyborcza a udział osób deklarujących się jako przeciwnicy integracji z Unią Europejską [%].

1.2. **Udział osób średnio zadowolonych z życia** - zmienna reprezentuje średni procent osób, które w 12 miesiącach poprzedzających dane wybory parlamentarne, w ankietach CBOS z pytaniem "Czy na ogół jest Pan(i) zadowolony(a) ogólnie z całego życia?", spośród odpowiedzi "zadowolony(a)", "średnio zadowolony(a)" oraz "niezadowolony(a)" wybrały drugą opcję.

Zmienna ta silnie i odwrotnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym -0.652 (rys. 10.). Test istotności t-Studenta (p-value = 0.008753) wskazuje na istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około 0.4098, co oznacza, że wzrost odsetka osób średnio zadowolonych z życia o 10% powoduje wzrost frekwencji wyborczej w województwie o około 4.0978% (ceteris paribus). Województwa z wyższym udziałem osób średnio zadowolonych z życia odnotowują niższą frekwencję wyborczą w porównaniu do tych, gdzie udział ten jest mniejszy.



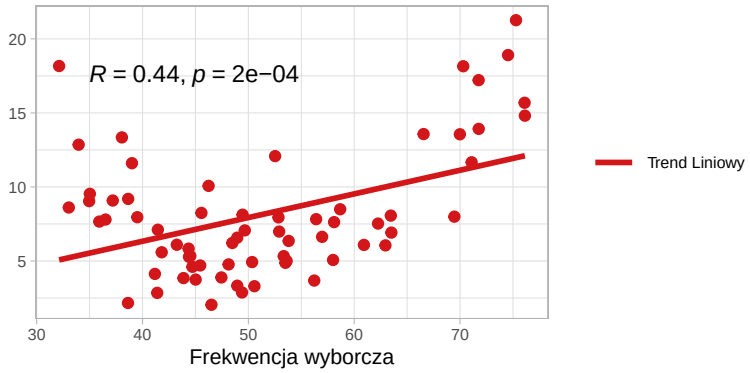
Rysunek 10. Frekwencja wyborcza a udział osób średnio zadowolonych z życia [%].

2. Zmienne ekonomiczne

2.1. **Zmiana produktu krajowego brutto na 1 mieszkańca w przeciągu 1 roku** - zmienna reprezentuje roczną zmianę produktu krajowego brutto (PKB) na jednego mieszkańca,

wyrażoną w procentach i wskazuje na dynamikę wzrostu lub spadku gospodarczego w danym roku.

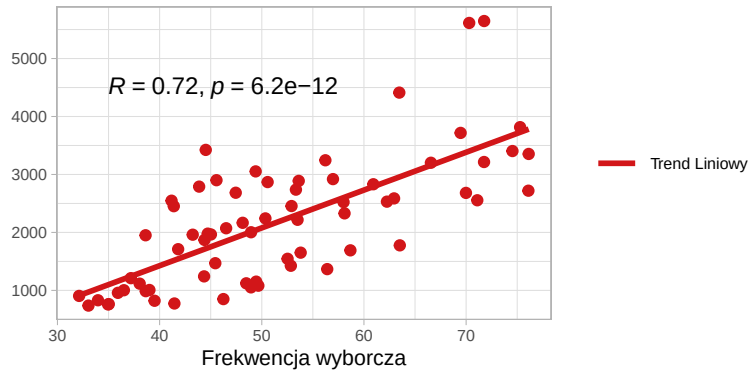
Zmienna ta umiarkowanie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym 0.4396 (rys. 11.). Test istotności t-Studenta (p-value = 0.000115) potwierdza istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około 0.5117, co oznacza, że wzrost PKB o 10% rok do roku powoduje wzrost frekwencji wyborczej o około 5.1169% (ceteris paribus). Województwa z wyższą dynamiką wzrostu PKB odnotowują wyższą frekwencję wyborczą.



Rysunek 11. Frekwencja wyborcza a zmiana produktu krajowego brutto na 1 mieszkańca w przeciągu 1 roku [%].

2.2. **Nakłady inwestycyjne na 1 mieszkańca w sektorze publicznym [zł]** - zmienna reprezentuje wartość nakładów inwestycyjnych w sektorze publicznym przypadających na jednego mieszkańca, wyrażona w złotych. Nakłady te mogą obejmować inwestycje w infrastrukturę, edukację, zdrowie i inne obszary publiczne.

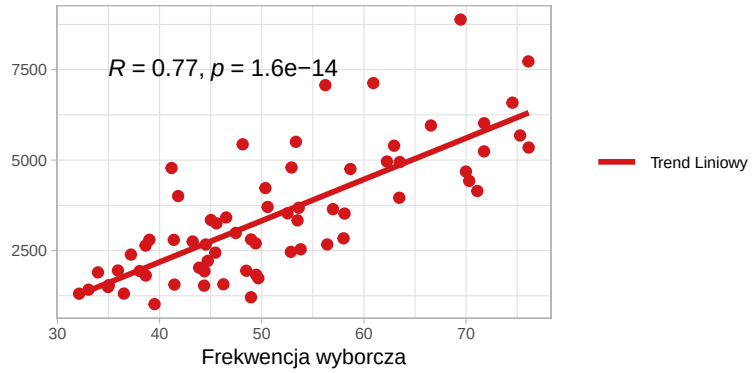
Zmienna ta silnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym 0.7206 (rys. 12.). Test istotności t-Studenta (p-value = $7.96*10^{-5}$) potwierdza istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około 0.0024, co oznacza, że wzrost nakładów inwestycyjnych o 1000zł powoduje wzrost frekwencji wyborczej o około 2.3524% (ceteris paribus). Województwa z wyższymi nakładami inwestycyjnymi na mieszkańca w sektorze publicznym odnotowują wyższą frekwencję wyborczą.



Rysunek 12. Frekwencja wyborcza a nakłady inwestycyjne na 1 mieszkańca w sektorze publicznym [zł].

2.3. **Nakłady inwestycyjne na 1 mieszkańca w sektorze prywatnym [zł]** - zmienna reprezentuje wartość nakładów inwestycyjnych w sektorze prywatnym przypadających na jednego mieszkańca, wyrażona w złotych. Nakłady te mogą obejmować inwestycje w przemysł, usługi, handel i inne obszary prywatne.

Zmienna ta barzo silnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym 0.7740 (rys. 13.). Test istotności t-Studenta (p-value = $1.30*10^{-6}$) potwierdza istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około 0.0020, co oznacza, że wzrost nakładów inwestycyjnych o 1000zł powoduje wzrost frekwencji wyborczej o około 1.9553% (ceteris paribus). Województwa z wyższymi nakładami inwestycyjnymi na mieszkańca w sektorze prywatnym odnotowują wyższą frekwencję wyborczą.

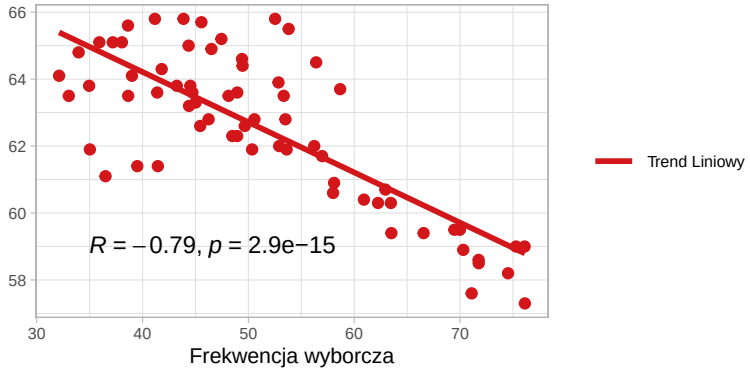


Rysunek 13. Frekwencja wyborcza a nakłady inwestycyjne na 1 mieszkańca w sektorze prywatnym [zł].

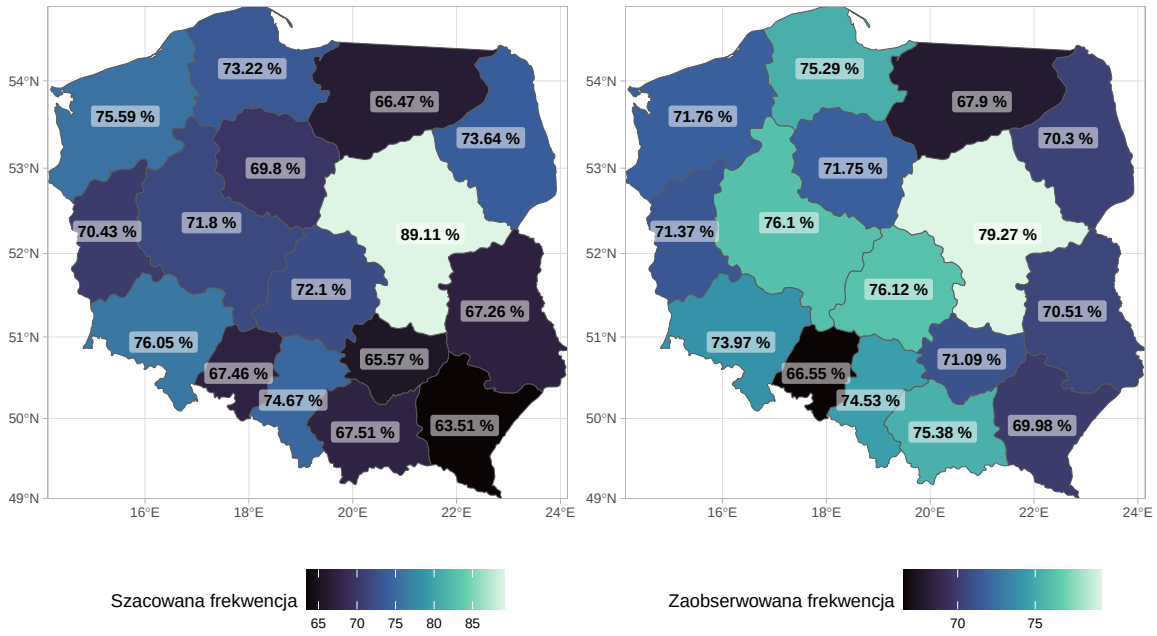
3. Zmienna demograficzna

3.1. **Udział ludności w wieku produkcyjnym** - zmienna reprezentuje procent ludności w wieku produkcyjnym (tj. mężczyźni w wieku 18-64 lat i kobiety w wieku 18-59 lat) w całkowitej populacji. Jest to kluczowy wskaźnik struktury demograficznej, wpływający na siłę roboczą i potencjał gospodarczy regionu.

Zmienna ta bardzo silnie i odwrotnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym -0.7870 (rys. 14.). Test istotności t-Studenta (p-value = 2.96×10^{-10}) potwierdza istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około -2.3044, co oznacza, że wzrost udziału osób w wieku produkcyjnym o 10% powoduje spadek frekwencji wyborczej o około 23.0442% (ceteris paribus). Województwa z wyższym udziałem ludności w wieku produkcyjnym z reguły odnotowują niższą frekwencję wyborczą.



Rysunek 14. Frekwencja wyborcza a udział ludności w wieku produkcyjnym [%].



Rysunek 15. Mapa szacowanej i faktycznej frekwencji wyborczej w wyborach parlamentarnych w 2023 roku dla modelu N95425.

Model zastosowano na danych dotyczących wyborów parlamentarnych w 2023 roku. Średnia różnica między rzeczywistą a przewidywaną frekwencją wyborczą wynosiła -1.106 punktów procentowych, z medianą -1.696 pp., co oznacza, że model ma tendencję do niedoszacowywania frekwencji.

Średnie bezwzględne odchylenie predykcji wynosiło 3.621 punktów procentowych, a jego bezwzględne odchylenie standardowe 2.688 pp. Największa odnotowana pozytywna różnica względem frekwencji zaobserwowanej wyniosła +9.839 pp. w województwie mazowieckim, z kolei największa negatywna różnica wyniosła -7.8635 w województwie Małopolskim. Najmniejsze odchylenie wyniosło +0.1322 oraz +0.9070 pp kolejno w województwie Śląskim i Opolskim. Wyniki wskazują na zróżnicowanie dokładności przewidywań modelu w zależności od regionu (rys. 15).

Model w tej postaci, po przeprowadzeniu dziesięciokrotnej walidacji krzyżowej charakteryzuje się skorygowanym współczynnikiem R-kwadrat na poziomie 0.9240. Wyniki walidacji modelu na zbiorze testowym wskazują na bardzo zbliżoną wartość R-kwadrat na poziomie 0.9309 (+0.0069 w porównaniu do skorygowanego R-kwadrat obliczonego na zbiorze treningowym), z pierwiastkiem błędu średniokwadratowego wynoszącym 3.3685 i średnim błędem absolutnym na poziomie 2.4889, wskazując na dużą wydajność modelu na różnych podzbiorach i brak nadmiernego dopasowania, co oznacza, że model daje się dobrze uogólnić na nowe dane [4].

[4]Ze względu na małą liczebność próby statystyki modelu N113549 także zostały potwierdzone za pomocą bootstrappingu (1000 iteracji). Współczynnik R-kwadrat wyniósł 0.9273 (przedział ufności 95%: 0.8735, 0.9694), pierwiastek błędu średniokwadratowego wyniósł 2.5544% (przedział ufności 95%: 1.6955, 3.2169) a średni błąd absolutny wyniósł 2.0093% (przedział ufności 95%: 1.216, 2.7376). Niewielkie odchylenia dają podstawy do uznania, że model jest

Aby zbadać wiarygodność i stabilność modelu, przetestowano go pod kątem homoskedastyczności za pomocą testu Breuscha-Pagana na poziomie istotności $\alpha = 0.05$. Wynik testu (BP = 1.7351, df = 3, p-value = 0.6292) wskazuje na brak istotnych dowodów przeciwko hipotezie zerowej o homoskedastyczności. Oznacza to, że nie ma podstaw do stwierdzenia, że wariancja reszt modelu nie jest stała.

Ponadto przeprowadzono test normalności reszt za pomocą testu Shapiro-Wilka na poziomie istotności $\alpha = 0.05$. Wynik testu (W = 0.97258, p-value = 0.6314) wskazuje na brak istotnych dowodów przeciwko hipotezie zerowej o normalności rozkładu reszt. Oznacza to, że nie ma podstaw do stwierdzenia, że rozkład reszt odbiega or rozkładu normalnego.

Dla każdej ze zmiennych w przedstawionym modelu przeprowadzono test istotności zmiennych t-Studenta na poziomie istotności $\alpha = 0.05$. Wyniki testu wskazują, że wszystkie zmienne ujęte w modelu są istotne. W przypadku każdej ze zmiennych wyniki testu dają podstawy do przyjęcia hipotezy alternatywnej o istotnym ich oddziaływaniu na na zmienną objaśnianą.

W celu zbadania występowania współliniowości pomiędzy zmiennymi modelu dokonano jego oceny za pomocą współczynnika inflacji wariancji (VIF), przyjmując za optymalny zakres od 0 do 5. Wszystkie zmienne modelu charakteryzują się wartością współczynnika VIF na poziomie odpowiadającym wyznaczonemu przedziałowi. Najwyższą wartość współczynnika VIF na poziomie 2.6686 zaobserwowano dla zmiany produktu krajowego brutto na 1 mieszkańca w przeciągu 1 roku oraz następnie na poziomie 2.6563 dla udziału ludności w wieku produkcyjnym. Oznacza to, że zmienne modelu charakteryzują się brakiem lub umiarkowaną współliniowością.

stabilny i dobrze dopasowany, a wyniki są wiarygodne nawet przy ograniczonej liczebności próby. Wartości te sugerują, że model nie jest nadmiernie dopasowany do danych treningowych i zachowuje swoją zdolność predykcyjną w odniesieniu do nowych danych.

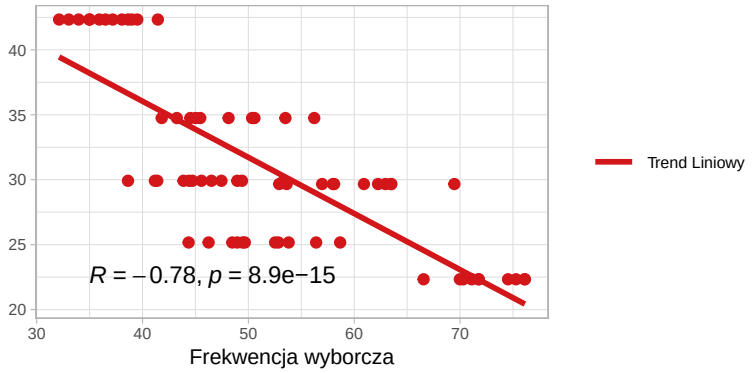
4.2.2 Opis modelu N113549

Pośród zmiennych modelu N113549 znalazły się dwie zmienne ankietowe pochodzące z badań opinii publicznej ośrodka CBOS, jedna zmienna ekonomiczna, dwie zmienne demograficzne oraz jedna zmienna związana z poziomem edukacji.

1. Zmienne ankietowe

   1.1. **Udział osób deklarujących obojętny stosunek do rządu** - zmienna ta reprezentuje średni procent osób, które w 12 miesiącach poprzedzających dane wybory parlamentarne, w ankietach CBOS pod tytułem "Stosunek do rządu", spośród dostępnych deklaracji "zwolennicy", "przeciwnicy" oraz "obojętni", zadeklarowały obojętny stosunek do rządu.

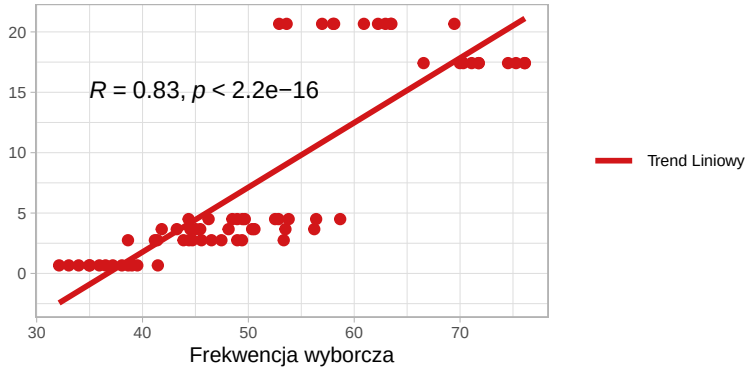
   Zmienna ta bardzo silnie i odwrotnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym -0.7784 (rys. 16.). Test istotności t-Studenta (p-value = 9.53×10^{-13}) wskazuje na istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi w przybliżeniu -0.6799, co oznacza, że wzrost odsetka osób deklarujących obojętny stosunek do rządu o 10% powoduje spadek frekwencji wyborczej w województwie o około 6.7992% (ceteris paribus). Województwa z wyższym udziałem osób deklarujących obojętny stosunek do rządu odnotowują niższą frekwencję wyborczą w porównaniu do tych, gdzie udział ten jest mniejszy. Na podstawie tej obserwacji można wywnioskować, że zarówno wyższy odsetek zwolenników jak i przeciwników rządu będzie oddziaływał motywująco na potencjalnych wyborców, przyciągając ich do urn wyborczych.



Rysunek 16. Frekwencja wyborcza a obojętny stosunek do rządu [%].

1.2. **Udział osób deklarujących brak większych problemów ze znalezieniem odpowiedniej pracy** - zmienna ta reprezentuje średni procent osób, które w 12 miesiącach poprzedzających dane wybory parlamentarne, w ankietach CBOS z pytaniem "Jak określił(a)by Pan(i) sytuację na rynku pracy w Pana(i) miejscowości lub okolicy? Czy, Pana(i) zdaniem, obecnie:", spośród dostępnych odpowiedzi "bez większych problemów można znaleźć odpowiednią pracę", "można wprawdzie znaleźć jakąś pracę, ale trudno jest o pracę odpowiednią", "trudno jest znaleźć jakąkolwiek pracę", "nie można znaleźć żadnej pracy" oraz "Trudno powiedzieć" wybrały pierwszą odpowiedź.

Zmienna ta bardzo silnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym 0.8302 (rys. 17.). Test istotności t-Studenta (p-value = 0.00203) wskazuje na istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około 0.2988, co oznacza, że wzrost odsetka osób deklarujących brak większych problemów ze znalezieniem odpowiedniej pracy o 10% powoduje wzrost frekwencji wyborczej w województwie o około 2.9881% (ceteris paribus). Województwa z wyższym udziałem osób deklarujących brak większych problemów ze znalezieniem odpowiedniej pracy odnotowują średnio wyższą frekwencję wyborczą w porównaniu do tych, gdzie udział ten jest mniejszy. Na tej podstawie można stwierdzić, że postrzeganie przez obywateli sytuacji na rynku pracy ma istotny wpływ na frekwencję wyborczą.

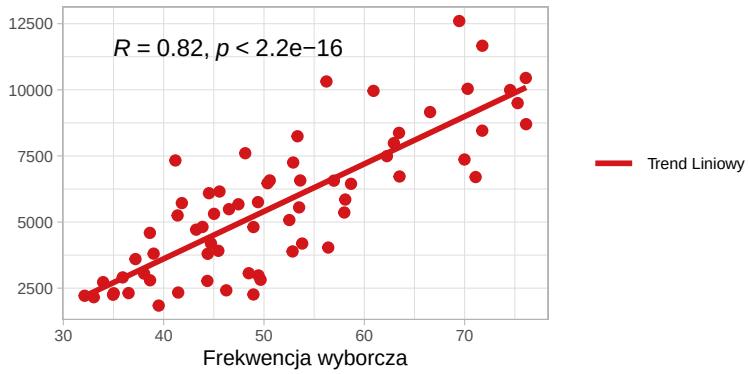


Rysunek 17. Frekwencja wyborcza a obojętny stosunek do rządu [%].

2. Zmienne ekonomiczne

2.1. **Nakłady inwestycyjne na 1 mieszkańca ogółem [zł]** - zmienna ta reprezentuje średnią wartość nakładów inwestycyjnych przypadających na jednego mieszkańca, wyrażoną w złotych. Nakłady te obejmują inwestycje zarówno w sektorze publicznym, jak i prywatnym.

Zmienna ta bardzo silnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym 0.8151 (rys. 18.). Test istotności t-Studenta (p-value = $2.17*10^{-9}$) potwierdza istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi w przybliżeniu 0.0016, co oznacza, że wzrost nakładów inwestycyjnych o 1000zł powoduje wzrost frekwencji wyborczej o około 1.5829% (ceteris paribus). Województwa z wyższymi nakładami inwestycyjnymi na mieszkańca odnotowują wyższą frekwencję wyborczą.

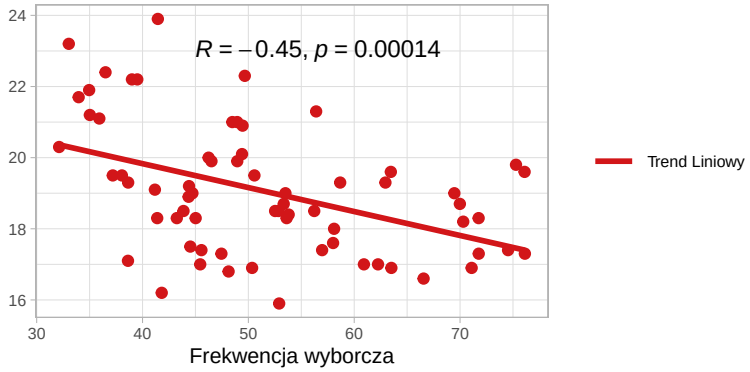


Rysunek 18. Frekwencja wyborcza a nakłady inwestycyjne na 1 mieszkańca w sektorze prywatnym [zł].

3. Zmienne demograficzne

3.1. **Udział ludności w wieku przedprodukcyjnym** - zmienna ta reprezentuje procent ludności w wieku przedprodukcyjnym (tj. dzieci i młodzież w wieku 0-17 lat) w całkowitej populacji. Jest to kluczowy wskaźnik struktury demograficznej, wpływający na obciążenie systemów edukacji i zabezpieczenia społecznego.

Zmienna ta umiarkowanie i odwrotnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym -0.4481 (rys. 19.). Test istotności t-Studenta (p-value = 2.10×10^{-5}) wskazuje na istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi w przybliżeniu 1.2374, co oznacza, że wzrost udziału ludności w wieku przedprodukcyjnym w populacji o 10% powoduje wzrost frekwencji wyborczej w województwie o około 12.3744% (ceteris paribus). Województwa z wyższym udziałem ludności w wieku przedprodukcyjnym odnotowują niższą frekwencję wyborczą w porównaniu do tych, gdzie udział ten jest mniejszy.

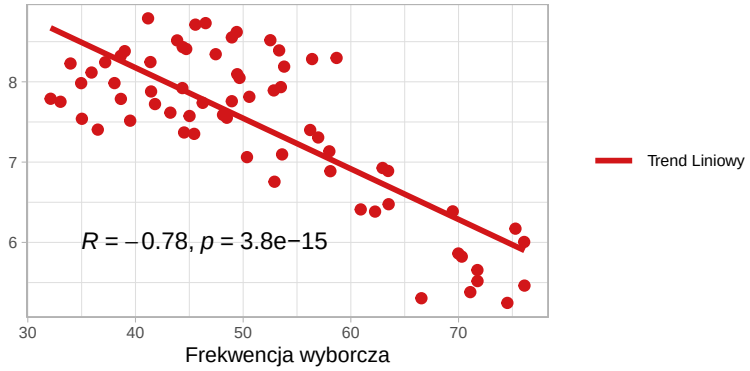


Rysunek 19. Frekwencja wyborcza a udział ludności w wieku przedprodukcyjnym [%].

3.2. **Udział ludności w wieku 25-29 lat** - zmienna ta reprezentuje całkowitą liczbę ludności w wieku 25-29 lat, bez podziału na płeć. Jest to istotny wskaźnik struktury demograficznej,

odnoszący się do młodych dorosłych, którzy zazwyczaj kończą edukację i wchodzą na rynek pracy.

Zmienna ta bardzo silnie i odwrotnie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym -0.785 (rys. 20.). Test istotności t-Studenta (p-value = 1.06×10^{-5}) wskazuje na istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi około -3.1844, co oznacza, że wzrost udzialu ludności w wieku 25-29 lat o 1% powoduje spadek frekwencji wyborczej w województwie o około 3.1844% (ceteris paribus). Województwa z wyższym udziałem ludności w wieku 25-29 lat odnotowują niższą frekwencję wyborczą w porównaniu do tych, gdzie udział ten jest mniejszy.

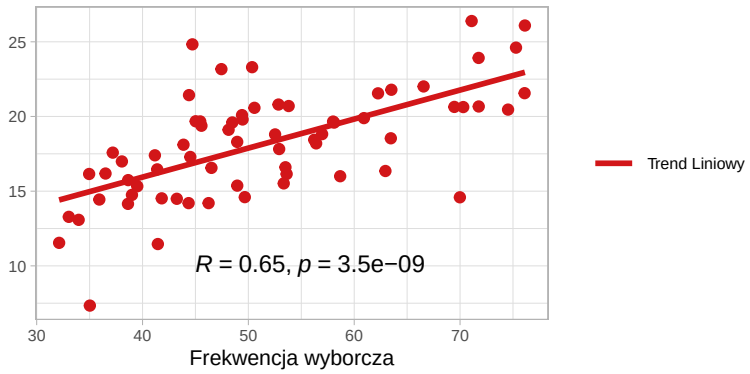


Rysunek 20. Frekwencja wyborcza a udział ludności w wieku 25-29 lat [%].

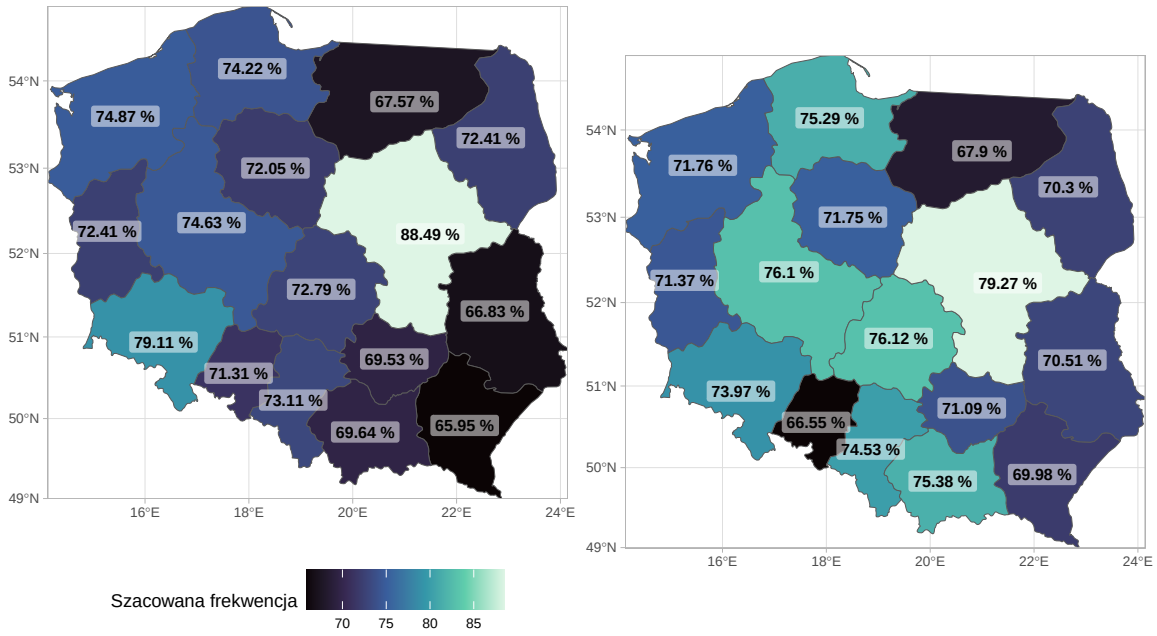
4. Zmienna edukacyjna

4.1. **Współczynnik skolaryzacji brutto w szkołach policealnych, w tym kolegiach (dla ludności w wieku 19-21 lat)** - zmienna ta reprezentuje procent młodych ludzi w wieku 19-21 lat uczęszczających do szkół policealnych lub kolegiów. Jest to wskaźnik poziomu kształcenia i zaangażowania w dalszą edukację po ukończeniu szkoły średniej.

Zmienna ta umiarkowanie koreluje z frekwencją wyborczą, ze współczynnikiem korelacji Pearsona wynoszącym 0.6462 (rys. 21.). Test istotności t-Studenta (p-value = 0.00181) wskazuje na istotność zmiennej dla modelu. Wartość parametru przy tej zmiennej wynosi w przybliżeniu 0.4512, co oznacza, że wzrost tego współczynnika skolaryzacji o 10% powoduje wzrost frekwencji wyborczej w województwie o około 4.512% (ceteris paribus). Województwa z wyższym udziałem współczynnikiem skolaryzacji brutto w szkołach policealnych, w tym kolegiach odnotowują wyższą frekwencję wyborczą w porównaniu do tych, gdzie współczynnik ten jest mniejszy.



Rysunek 21. Frekwencja wyborcza a współczynnik skolaryzacji brutto w szkołach policealnych, w tym kolegiach (dla ludności w wieku 19-21 lat) [%].



Rysunek 22. Mapa szacowanej i faktycznej frekwencji wyborczej w wyborach parlamentarnych w 2023 roku dla modelu N113549.

Model zastosowano na danych dotyczących wyborów parlamentarnych w 2023 roku. Średnia różnica między rzeczywistą a przewidywaną frekwencją wyborczą wynosiła 0.1890 punktu procentowego, z medianą -0.6964 pp., co oznacza, że model częściej zaniża frekwencję, jednocześnie zwracając w niektórych przypadkach bardzo wysokie wartości, które zawyżają średnią predykcji.

Średnie bezwzględne odchylenie predykcji wynosiło 3.0183 punktów procentowych, a jego bezwzględne odchylenie standardowe 2.3827 pp. Największa odnotowana pozytywna różnica względem frekwencji zaobserwowanej wyniosła +9.2159 pp. w województwie Mazowieckim, z kolei największa negatywna różnica wyniosła -5.7392 pp. w województwie Małopolskim. Najmniejsze odchylenie wyniosło +0.2957 oraz -0.3304 pp kolejno w województwie Kujawsko-Pomorskim oraz Warmińsko-Mazurskim. Również w tym przypadku wyniki wskazują na zróżnicowanie dokładności przewidywań modelu w zależności od regionu (rys. 22).

Model w tej postaci, po przeprowadzeniu dziesięciokrotnej walidacji krzyżowej charakteryzuje się skorygowanym współczynnikiem R-kwadrat na poziomie 0.9372. Wyniki walidacji modelu na zbiorze testowym wskazują na zbliżoną wartość R-kwadrat na poziomie 0.9429 (-0.0057 w porównaniu do skorygowanego R-kwadrat obliczonego na zbiorze treningowym), z pierwiastkiem błędu średniokwadratowego wynoszącym 4.0806% i średnim błędem absolutnym na poziomie 3.2904%, wskazując na dużą wydajność modelu na różnych podzbiorach i brak nadmiernego dopasowania, co oznacza, że model daje się dobrze uogólnić na nowe dane [5].

[5] Ze względu na małą liczebność próby statystyki modelu N113549 także zostały potwierdzone za pomocą bootstrappingu (1000 iteracji). Współczynnik R-kwadrat wyniósł 0.9273 (przedział ufności 95%: 0.8735, 0.9694), pierwiastek błędu średniokwadratowego wyniósł 2.5544% (przedział ufności 95%: 1.6955, 3.2169) a średni błąd absolutny wyniósł

Aby zbadać wiarygodność i stabilność modelu, przetestowano go pod kątem homoskedastyczności za pomocą testu Breuscha-Pagana na poziomie istotności $\alpha = 0.05$. Wynik testu (BP = 4.5045, df = 6, p-value = 0.6087) wskazuje na brak istotnych dowodów przeciwko hipotezie zerowej o homoskedastyczności. Oznacza to, że nie ma podstaw do stwierdzenia, że wariancja reszt modelu nie jest stała.

Ponadto przeprowadzono test normalności reszt za pomocą testu Shapiro-Wilka na poziomie istotności $\alpha = 0.05$. Wynik testu (W = 0.99199, p-value = 0.9465) wskazuje na brak istotnych dowodów przeciwko hipotezie zerowej o normalności rozkładu reszt. Oznacza to, że nie ma podstaw do stwierdzenia, że rozkład reszt odbiega or rozkładu normalnego.

Dla każdej ze zmiennych w przedstawionym modelu przeprowadzono test istotności zmiennych t-Studenta na poziomie istotności $\alpha = 0.05$. Wyniki testu wskazują, że wszystkie zmienne ujęte w modelu są istotne. W przypadku każdej ze zmiennych wyniki testu dają podstawy do przyjęcia hipotezy alternatywnej o istotnym ich oddziaływaniu na na zmienną objaśnianą.

W celu zbadania występowania współliniowości pomiędzy zmiennymi modelu dokonano jego oceny za pomocą współczynnika inflacji wariancji (VIF), przyjmując za optymalny zakres od 0 do 5.

Wszystkie zmienne modelu charakteryzują się wartością współczynnika VIF na poziomie odpowiadającym wyznaczonemu przedziałowi. Najwyższą wartość współczynnika VIF na poziomie 3.7493 zaobserwowano dla udziału osób deklarujących brak większych problemów ze znalezieniem odpowiedniej pracy oraz następnie na poziomie 2.9662 dla udziału ludności w wieku 25-29 lat. Oznacza to, że zmienne modelu charakteryzują się brakiem lub umiarkowaną współliniowością.

2.0093% (przedział ufności 95%: 1.216, 2.7376). Niewielkie odchylenia dają podstawy do uznania, że model jest stabilny i dobrze dopasowany, a wyniki są wiarygodne nawet przy ograniczonej liczebności próby. Wartości te sugerują, że model nie jest nadmiernie dopasowany do danych treningowych i zachowuje swoją zdolność predykcyjną w odniesieniu do nowych danych.

5 Podsumowanie i wnioski

Głównym celem niniejszej pracy było zaproponowanie modelu objaśniającego frekwencję wyborczą na poziomie wojewódzkim. W wyniku zastosowanej metodologii powstały dwa konkurencyjne modele oparte o różne rodzaje zmiennych, z których oba osiągają porównywalne wyniki w procesie diagnostycznym. Udoskonalone modele N113549 i N95425 wyraźnie przewyższają model pierwotny pod względem zarówno współczynnika determinacji, jak i statystycznej istotności zmiennych. Dzieje się tak dzięki lepszemu doborowi zmiennych, które pomagają trafniej objaśnić wariancję we frekwencji wyborczej w Polsce.

Oba zaprezentowane modele są stabilne i dobrze dopasowane, co zostało potwierdzone przez testy na homoskedastyczność i normalność reszt. W wyniku walidacji krzyżowej uzyskano porównywalne statystyki R-kwadrat wynoszące 0.9438 dla modelu N113549 i 0.9197 dla modelu N95425 (+0.0241 na korzyść modelu N113549). Wskaźniki RMSE i MAE również znajdują się na nieznacznie lepszym poziomie w przypadku modelu N113549, wynosząc kolejno 3.1784 i 2.6015 w porównaniu do 3.3787 i 2.7849 w przypadku modelu N95425 (-0.2003 i -0.1834).

Po walidacji "poza próbą" na zbiorze testowym skorygowany współczynnik R-kwadrat był wyższy w przypadku modelu N113549, wynosząc 0.9429 w porównaniu do 0.9309 dla modelu N95425 (+0.012). Wskaźnik RMSE wyniósł 4.0806 dla modelu N113549 i 3.3685 dla modelu N95425 (+0.712), a wskaźnik MAE 3.2904 dla modelu N113549 i 2.4889 dla modelu N95425 (+0.8014). Oznacza to, że pomimo lepszych wyników diagnozy modelu N113549, model N95425 osiąga nieznacznie lepsze wyniki na realnych danych. Różnice są jednak marginalne a wydajność modeli porównywalna.

Po zastosowaniu obu modeli na danych dotyczących 2023 roku zaobserwowano, że oba modele mają skłonność do zaniżania frekwencji, choć zjawisko to wyraźniej występuje w modelu N95425 niż N113549. Oba modele mają także tendencję do przeszacowywania frekwencji w województwie Mazowieckim i jednocześnie niedoszacowywania frekwencji w województwie Małopolskim. Jednocześnie, oba modele relatywnie trafnie przewidują frekwencję w takich województwach jak Kujawsko-Pomorskie, Warmińsko-Mazurskie, Lubuskie, Pomorskie i Śląskie (absolutny błąd predykcji poniżej 2%) (rys. 15 i 22.).

Cele pracy zostały w pełni osiągnięte. Opracowane modele skutecznie realizują założenia badawcze, dostarczając adekwatnych wyjaśnień zmienności frekwencji wyborczej na poziomie wojewódzkim. Modele te charakteryzują się wysoką jakością diagnostyczną i istotnością statystyczną, a także potwierdzają swoją zdolność predykcyjną na danych spoza próby. Mimo drobnych różnic, oba modele stanowią cenne narzędzie analityczne w badaniach nad zachowaniami wyborczymi, umożliwiając efektywne przewidywanie wzorców frekwencji wyborczej. Dalsze prace nad tymi modelami, bazujące na zaprezentowanej metodologii i danych z kolejnych cykli wyborczych, mogą znacząco poprawić ich precyzję.

Bibliografia

Blais, André. 2014. „Why is Turnout So Low in Switzerland? Comparing the Attitudes of Swiss and German Citizens Towards Electoral Democracy". *Swiss Political Science Review* 20 (czerwiec): 520–28. https://doi.org/10.1111/spsr.12116.

Blais, André, i Agnieszka Dobrzynska. 1998. „Turnout in Electoral Democracies". *European Journal of Political Research* 33 (styczeń): 239–61. https://doi.org/10.1111/1475-6765.00382.

Cześnik, Mikołaj, i Piotr Zagórski. 2022. „Wiek a uczestnictwo wyborcze w Polsce: cykl życia, okres, kohorta". *Studia Socjologiczne* 4: 41–66. https://doi.org/10.24425/sts.2022.143582.

Frank, Richard W., i Ferran Martínez Coma. 2023. „Correlates of Voter Turnout". *Political Behavior* 45 (2): 607–33. https://doi.org/10.1007/s11109-021-09720-y.

Gendźwiłł, Adam, Jakub Rutkowski, i Tomasz Żółtak. 2014. „O związku partycypacji wyborczej i stabilności poparcia dla partii. Lokalne efekty frekwencyjne w wyborach parlamentarnych w Polsce w latach 2005–2011". *Studia Socjologiczne* 3: 107—126. https://www.ceeol.com/search/article-detail?id=98157.

Gomez, Brad T., Thomas G. Hansford, i George A. Krause. 2007. „The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections". *The Journal of Politics* 69 (3): 649–63. https://doi.org/10.1111/j.1468-2508.2007.00565.x.

Jackman, Robert W. 1987. „Political Institutions and Voter Turnout in the Industrial Democracies". *The American Political Science Review* 81 (2): 405–23. https://doi.org/10.2307/1961959.

Kowalski, Michał. 2016. „Zróżnicowanie przestrzenne zachowań wyborczych mieszkańców gminy Szadek w wyborach prezydenckich i parlamentarnych w 2015 roku". *Biuletyn Szadkowski* 16: 243–56. https://doi.org/10.18778/1643-0700.16.12.

Lijphart, Arend. 1997. „Unequal Participation: Democracy's Unresolved Dilemma". *The American Political Science Review* 91 (1): 1–14. https://doi.org/10.2307/2952255.

Murr, Andreas, i Simon Munzert. 2018. „Forecasting National Turnout at British General Elections: Combining Polls and Structural Models". *Electoral Studies* 86 (wrzesień): 16–20. https://doi.org/10.31235/osf.io/s2j6h.

Musiał-Karg, Magdalena. 2011. „Metody zwiększania frekwencji wyborczej. Polska a doświadczenia innych państw". *Środkowoeuropejskie Studia Polityczne*, nr 2 (czerwiec): 77–100. https://doi.org/10.14746/ssp.2011.2.05.

Najbar, Marcin. 2017. „Wpływ ordynacji wyborczej i polaryzacji sceny politycznej na poziom frekwencji wyborczej". *Athenaeum. Polskie Studia Politologiczne* 55: 89–107. https://doi.org/10.15804/athena.2017.55.05.

Pierzgalski, Michał. 2012. „Modele regresyjne w badaniu zachowań wyborczych". *Athenaeum. Polskie Studia Politologiczne* 36 (styczeń): 115–42. https://doi.org/10.15804/athena.2012.36.06.

Sasińska-Klas, Teresa. 2008. „O absencji wyborczej w Polsce". W *Wybory samorządowe w kontekście mediów i polityki*, zredagowane przez Maria Magoska, 35–49. https://ruj.uj.edu.pl/xmlui/handle/item/155630.

Siedziako, Michał. 2016. „Manipulacje i fałszerstwa wyborcze w wyborach do Sejmu PRL (1952–1985)". *Pamięć i Sprawiedliwość* 1: 112–39. https://doi.org/0.12775/DN.2018.4.07.

Solijonov, Abdurashid. 2016. *Voter Turnout Trends around the World*. IDEA, International Institute for Democracy; Electoral Assistance. https://www.idea.int/sites/default/files/publications/voter-turnout-trends-around-the-world.pdf.

Zagała, Zbigniew. 2023. „Od podziału postsolidarnościowego do polaryzacji. (Krótkie) studium relacji między Prawem i Sprawiedliwością a Platformą Obywatelską i ich wyborcami". *Zeszyty Naukowe KUL*, 107–26. https://doi.org/10.31743/znkul.16429.