

Paweł Matuszewski

Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie

CZY MOŻNA WIERZYĆ SONDAŻOM PRZEDWYBORCZYM? WYKORZYSTANIE PODEJŚCIA BAYESOWSKIEGO DO ANALIZY ROZBIEŻNOŚCI MIĘDZY WYNIKAMI WYBORÓW PARLAMENTARNYCH W POLSCE A DANymi Z BADAŃ SONDAŻOWYCH

Celem artykułu jest określenie tego, jak dobrze sondaże przedwyborcze potrafią przewidywać wyniki wyborów parlamentarnych oraz od czego zależy trafność tych prognoz. Zmiennymi wyjaśnianymi jest poprawne wskazanie zwycięskiego komitetu oraz wyniku wyborczego poszczególnych opcji politycznych biorących udział w wyborach. Pierwszą zmienną wyjaśniającą jest czas między badaniem a datą wyborów. Drugą zmienną wyjaśniającą jest różnica wskazań między dwoma komitetami o największym poparciu. Dane empiryczne obejmują wyniki sondaży w okresie 12 miesięcy przed wyborami parlamentarnymi w Polsce od 1993 do 2015 roku. W analizie wykorzystano bayesowski model hierarchiczny i symulację Monte Carlo. Częściowo potwierdziły się hipotezy, które wskazywały, że zdolność sondaży do przewidywania największego poparcia i do wskazania wyników wyborczych poszczególnych komitetów jest tym większa, im bliżej do wyborów. W pełni potwierdziła się hipoteza, że sondaże tym lepiej wskazują zwycięski komitet, im większa jest różnica między sondażowym poparciem dwóch głównych oponentów.

Główne pojęcia: sondaże przedwyborcze; prognozy polityczne; Bayes; MCMC.

Wprowadzenie

Sondaże przedwyborcze stanowią nieodłączny element współczesnych kampanii politycznych. Im bliżej daty wyborów, tym więcej badań, które są na bieżąco komentowane przez dziennikarzy, polityków, ekspertów, naukowców, a także zwykłych obywateli w codziennych rozmowach (por. Jackman 2005; Strömbäck 2012). Ich główną funkcją jest przedstawienie aktualnego poparcia poszczególnych kandydatów lub opcji politycznych. Jednakże taka informacja sama w sobie może mieć bardzo duże znaczenie i z tego powodu wokół sondaży powstaje szereg kontrowersji. Po pierwsze, dla wyborcy przekonanie o tym,

jak liczne grono zwolenników ma popierana przez niego opcja polityczna lub kandydat, może mieć istotny wpływ na decyzję wyborczą. Między innymi od obserwowanego układu sił zależy, czy wyborca przeniesie głos na partię, która ma większe szanse wygrania, mimo że nie jest to jego pierwotny wybór, albo czy zostanie w domu uspokojony znaczną przewagą swojego kandydata. Z tego względu wśród osób śledzących lub zaangażowanych w kampanię można obserwować różne podejście do sondaży – od całkowitego wyparcia wyników jako niewiarygodnych po pełną akceptację i podkreślanie ich wagi – w zależności od tego, jaki wynik pokazują (por. Rothschild i Malhotra 2014). Po drugie, przekonanie, że sondaże mogą być istotnym powodem kształtującym zachowania wyborcze (jako np. społeczny dowód słuszności), może prowadzić do sytuacji, kiedy konkretne grupy interesu przeprowadzają własne badania, które dają korzystne z ich punktu widzenia wyniki (*advocacy polling*). Osiągają to między innymi poprzez manipulację metodami zbierania danych, przeprowadzanie badań na zbyt małych i niereprezentatywnych próbach, tendencyjne zadawanie pytań oraz selektywne raportowanie wyników, interpretowanie danych w korzystny dla siebie sposób (por. Converse i Traugott 1986). Po trzecie, dla obserwatorów wyniki sondażowe służą też jako przesłanka do formułowania prognoz na temat wyniku wyścigu politycznego. W związku z powyższym powstaje pytanie, czy i jak dalece można sondażom wierzyć?

Współcześnie dominują dwa główne podejścia do przewidywania wyników wyborczych. Pierwszy z nich opiera się na tworzeniu modeli statystycznych, gdzie predyktorami są zmienne makroekonomiczne (np. wzrost PKB) i polityczne (np. czy kandydat ubiega się o reelekcję). Ich zaletą jest fakt, że prognozy faktycznie są dostępne na kilka miesięcy przed wyborami oraz charakteryzują się dość wysoką sprawdzalnością (por. Foucault i Nadeau 2012; Lebo i Norpoth 2011; Lewis-Beck i Stegmaier 2010; Lewis-Beck, Nadeau i Bélanger 2004; Norpoth i Gschwend 2010). Minusem takich modeli jest dość słabe uzasadnienie teoretyczne. Preferencje partyjne zmieniają się dużo szybciej niż warunki makroekonomiczne i niekoniecznie te zmiany są ze sobą skorelowane (por. Campbell 2014; Matuszewski 2013). W konsekwencji nie są one w stanie wyjaśnić nagłych zwrotów w sondażach (por. Gelman i King 1993: 422), a *implicite* zakładają, że kampania wyborcza nie ma żadnego znaczenia (por. Bélanger i Soroka 2012). Kolejna problematyczna kwestia jest związana z tym, że opierają się na dość nielicznych obserwacjach (kilka–kilkanaście przypadków wyborów), co też rzutuje na rzetelność tego typu modeli¹. Drugie podejście to prognozowanie wyników na podstawie sondaży. Do jego zalet należy oparcie przewidywań na danych, które są zrozumiałe. Dość intuicyjnie postrzegany jest

¹ W analizie regresji przy małej liczbie obserwacji na ogólny wynik mogą bardzo silnie rzutować pojedyncze przypadki.

związek przyczynowo-skutkowy między intencją zagłosowania na X a rzeczywistą decyzją (por. Campbell 2014). Natomiast wadą tego podejścia jest fakt, że wiarygodne prognozy są dostępne zazwyczaj na krótko przed wyborami (por. Jackman 2005).

Główny cel tego artykułu polega na określeniu na podstawie danych historycznych dotyczących wyborów parlamentarnych w Polsce, jak bardzo precyzyjne są sondaże w przewidywaniu wyniku wyborczego oraz od jakich czynników zależy ta trafność. Temat ten jest często poruszany w mediach przy okazji każdego wyborów, a szczególnie wtedy, kiedy prognozy oparte na sondażach dość wyraźnie odbiegają od rzeczywistości (Domański i Kublik 2014, „Sondażowa nerwica” 2004, „Dlaczego sondaże pobiły?” 2010, „Sondaż nasz codzienny” 2004, „Sondażokracja” 2004, „Wielka kompromitacja sondażowni. Dlaczego sondaże nie przewidziały wyniku wyborów?” 2015, „Wybory prezydenckie 2015” 2015, „Markowski: Ratujmy sondaże wyborcze” 2015, „Za błędne sondaże wyborcze odpowiedzialne są głównie media” 2005). Przyczyny tych różnic są wyjaśniane przez ekspertów i naukowców. Swoje stanowisko w sprawie nierzetelnych badań zajmowało też np. Polskie Towarzystwo Socjologiczne („Socjologia w kampanii wyborczej 2005. Stanowisko Polskiego Towarzystwa Socjologicznego” 2005).

Rozminięcie się wyników sondażowych z wynikami wyborczymi, a także rozbieżności we wskazaniach różnych ośrodków badawczych mogą mieć kilka przyczyn. Należą do nich: forma pytania (otwarte/zamknięte), sformułowanie pytania, kolejność zadawania pytań, stosowanie różnych kafeterii, różne próby badawcze oraz różne metody zbierania danych (por. Gelman i King 1993; Groves i in. 2009; Linzer 2013). Należy także pamiętać, że w badaniach opartych na próbach statystycznych ma miejsce estymacja nieznanego parametru i należy uwzględnić margines błędu. Dla sondaży przedwyborczych, które najczęściej opierają się na próbie 900–1100 osób, wynosi on przy założeniu 95% przedziału ufności około 3%. Z tym że szacowanie poparcia nie jest oparte na całej próbie, a jedynie jej fragmencie, tj. tych ankietowanych, którzy zadeklarowali, że będą uczestniczyć w wyborach. Margines błędu jest w związku z tym w rzeczywistości większy (por. Jackman 2005).

Ze względu na to, że badania sondażowe prowadzone są najczęściej wśród prawdopodobnych wyborców (*likely voters*), pojawia się kolejne utrudnienie w traktowaniu sondaży jako prognoz wyniku wyborczego. Oszacowanie tego, kim jest wyborca, jest niezwykle trudne, ponieważ do dnia wyborów ta kategoria społeczna po prostu nie istnieje. Dopiero po wyborach wiemy, ile osób zdecydowało się głosować i kim te osoby były (por. Traugott 2015). Sondaże wyborcze muszą się zatem liczyć z dodatkowymi dwoma zjawiskami wynikającymi z efektu oczekiwań społecznych i wpływającymi na trafność ich wskazań (zob. Belli, Traugott, Young i McGonagle 1999; Holbrook i Krosnick 2010;

Matuszewski 2015). Pierwsze polega na tym, że w związku z postrzeganiem uczestnictwa w wyborach jako działania społecznie oczekiwanego, ankietowani mają tendencję do deklarowania udziału, choć w rzeczywistości wcale nie zamierzają tego robić. W konsekwencji odnotowana frekwencja jest zazwyczaj zawyżona i to w sposób nielosowy, co oznacza, że wyniki niektórych opcji politycznych mogą być przeszacowane, a innych niedoszacowane. Drugi mechanizm, który wpływa na końcowe wskazania, polega na tym, że ankietowani, między innymi z tego powodu, że ich kandydat jest wyśmiewany, mówią, że nie pójdą głosować, nie chcą podać, na kogo zagłosują (wybierają opcję „trudno powiedzieć”) lub odmawiają uczestniczenia w badaniu, chociaż w rzeczywistości oddają ważny głos.

W artykule skupiam się na tym, jak – w związku z powyższymi trudnościami – sondaże radzą sobie z przewidywaniem wyników wyborów parlamentarnych. Efektem dodanym płynącym z badań przeprowadzonych jest wskazanie:

- prawdopodobieństwa, że sondaż poprawnie przewidzi zwycięzcę wyborów;
- prawdopodobieństwa, że sondaż poprawnie przewidzi poziom poparcia dla danej opcji politycznej startującej w wyborach;
- czynników, które mają związek z mocą predykcyjną sondaży. Wśród nich wyróżniam: czas do wyborów oraz różnice między notowaniami poszczególnych partii.

W związku z powyższym artykuł nawiązuje do problematyki metod prognozowania opartych na danych sondażowych. Większość tego typu analiz w literaturze światowej dotyczy wyborów prezydenckich w Stanach Zjednoczonych. Stosunkowo niewiele jest natomiast opracowań, które dotyczyłyby wyborów parlamentarnych i systemu wielopartyjnego. Celem artykułu jest uzupełnienie tej luki.

Dane i metodologia

Analiza obejmuje wyniki sondażowe pojawiające się w ciągu roku przed wyborami parlamentarnymi w Polsce w latach 1993–2015. W zestawieniu nie uwzględniono wyborów z 1989 roku i 1991 roku ze względu na kilka przyczyn. Po pierwsze, wybory w 1989 roku były tylko częściowo wolne, a więc badanie realnych preferencji wyborczych miałoby w tym przypadku dość ograniczone zastosowanie. Po drugie, w pierwszych dwóch latach po roku 1989 badania przedwyborcze były prowadzone dość rzadko. W tym okresie udało mi się odnaleźć dane jedynie z września 1991 roku (dwa sondaże) i ze względu na tak małą liczebność zdecydowałem się ich nie wprowadzać do analizy.

Do analiz wybrałem badania realizowane przez siedem pracowni: CBOS, Estymator, GfK Polonia, IBRIS, MillwardBrown, PBS i TNS (nazwy obowiązujące

obecnie²). Kierowałem się jednym zasadniczym kryterium – uwzględniałem tylko te pracownie, które regularnie prowadziły badania preferencji politycznych. Poza pracowniami Estymator i IBRIS pozostałe ośrodki są członkami Organizacji Firm Badania Opinii i Rynku zrzeszającej podmioty dbające o wysokie standardy badań. Zdecydowałem się jednak uwzględnić te dwie pracownie, ponieważ podobnie jak pozostałe regularnie prowadzą badania. Tym samym również odpowiadają za reputację sondaży, a więc wiążą się z tytułowym pytaniem tego artykułu.

Jednym z większych wyzwań było skompletowanie bazy danych sondażowych, które sięgałyby początku lat dziewięćdziesiątych XX wieku, ponieważ nie wszystkie pracownie prowadzą w tym zakresie publicznie dostępne archiwa. W związku z tym zastosowałem dwie metody zbierania danych. Pierwsza to gromadzenie archiwalnych raportów zamieszczonych na firmowych stronach internetowych. Takie zbiory posiada jedynie CBOS, TNS i MillwardBrown. W pozostałych przypadkach dane historyczne były wyszukiwane w archiwach internetowych „Gazety Wyborczej” i „Rzeczpospolitej”. Wybrałem te dwie gazety, ponieważ jako jedyne ogólnopolskie i wielonakładowe dzienniki posiadają cyfrowe archiwa sięgające do początku lat dziewięćdziesiątych.

Liczba danych jest nieproporcjonalna w dwóch wymiarach. Dużo więcej danych pochodzi z ostatnich wyborów. Przyczyny tego stanu rzeczy mają źródło w obiektywnej sytuacji dotyczącej rynku badań społecznych, który w Polsce zaczął się rozwijać dopiero po 1989 roku. Z biegiem czasu powstawało coraz więcej pracowni i zaczęto coraz częściej prowadzić badania sondażowe. Drugi wymiar wiąże się z częstością przeprowadzania sondaży, która wyraźnie wzrasta w miarę zbliżania się do daty wyborów. Z tego też względu poświęcam część miejsca w artykule, aby omówić osobno sytuację dotyczącą poszczególnych lat i porównać wyizolowane w ten sposób trendy.

Łącznie zebrano 361 sondaży, w tym przed wyborami z 1993 roku – 29, z 1997 roku – 28, z 2001 roku – 40, z 2005 roku – 31, z 2007 roku – 37, z 2011 roku – 67 i z 2015 roku – 129. Część pracowni podaje wynik uwzględniający odpowiedź „Trudno powiedzieć”, a część nie, więc wszystkie dane w bazie zostały obliczone tak, aby ukazywały rozkład preferencji jedynie wśród zdecydowanych wyborców.

W analizie wykorzystałem podejście bayesowskie. Po pierwsze, dostarcza ono teoretycznych podstaw uzasadniających agregację danych. Po drugie, nie opieram się na reprezentatywnej próbie badawczej wyników sondaży, ale na wszystkich wynikach, które udało mi się znaleźć. W związku z tym stosowanie klasycznych przedziałów ufności określających granice, w jakich znajduje

² Wcześniej MillwardBrown pojawiał się pod nazwą SMG/KRC, a IBRIS jako Homo Homini.

się nieznanym parametrem, jest w tym przypadku nieuzasadnione. Dużo bardziej adekwatne jest podejście bayesowskie (symulacja Monte Carlo), na podstawie którego obliczana jest wartość parametru oparta na aktualnie dostępnych danych (zob. Jackman 2009: XXVII–XXXIV). Dodatkową korzyścią tego podejścia jest redukcja w wyniku agregacji błędów wynikającego z faworyzowania przez pracowni badawcze wybranych opcji politycznych oraz zmniejszenie zakresu niepewności wynikającego z błędów standardowych (por. Graefe, Armstrong, Jones Jr. i Cuzán 2014; Jackman 2005). Ma to szczególne znaczenie w sytuacji, kiedy można pozyskać znaczną liczbę różnych sondaży, stosujących różną metodologię i nie ma pewności, które z nich są najbardziej dokładne (por. Armstrong 2001).

Specyfikacja modeli

Podejście bayesowskie jest z perspektywy stojących za nim założeń teoretycznych metodą wręcz intuicyjną. Ogólna koncepcja polega na tym, że dotychczasowa wiedza dotycząca wybranego parametru (a czasem fakt, że nie wiemy o nim nic) jest aktualizowana na podstawie nowych danych. Rezultatem obliczeń jest oszacowanie wraz z zakresem błędów parametru opartego na posiadanym materiale dowodowym. Wynik podawany jest jako prawdopodobieństwo, a więc do jego zrozumienia wystarczy wiedza osiągnięta na poziomie matury. Dodatkowo jest on także intuicyjnie interpretowalny, np. można stwierdzić, że z prawdopodobieństwem 95% wynik mieści się w wartościach od *a* do *b*. Wyróżnia się tym samym na tle dość skomplikowanego w interpretacji wyniku testu istotności wskazującego jedynie prawdopodobieństwo popełnienia błędów I rodzaju (por. Jackman 2009: XXVII–XXX; Sobczyk 2002: 149).

Podejście bayesowskie posiada jednakże dość poważną wadę, która w znacznym stopniu tłumaczy, dlaczego do tej pory nie stało się dominującym nurtem w statystyce. Wymaga dość dużej wiedzy matematycznej, a także umiejętności programowania. Po pierwsze, każdy model może być sformułowany w sposób całkowicie dowolny i w dużym stopniu odwzorowujący rzeczywiste relacje między zmiennymi. Jest to jednak niemożliwe bez zrozumienia matematycznych podstaw modelowania bayesowskiego. Po drugie, według mojej wiedzy nie istnieją programy komputerowe, które kompleksowo umożliwiałyby w miarę intuicyjne tworzenie dowolnych modeli statystycznych (jak np. modele regresji logistycznej oparte na statystyce Fisherowskiej, które w ciągu kilku sekund można zbudować przy użyciu między innymi IBM SPSS). Taka sytuacja wymusza opanowanie przez badaczy przynajmniej podstaw programowania (najczęściej w języku R lub Python).

Odpowiedzi na pytania badawcze zostały udzielone na podstawie trzech skonstruowanych przeze mnie modeli hierarchicznych (zob. Kruschke 2014: 221). Każdy opis rozpoczynam od określenia zmiennej niezależnej i zmiennych zależnych, tak aby osoba nieznająca metody także mogła zorientować się,

co i jak jest wnioskowane. Ponadto każdy model opisuję matematycznie. Mam świadomość, że jest to język dość hermetyczny i prawdopodobnie mało zrozumiały dla kogoś, kto po raz pierwszy się spotyka z metodą bayesowską. Jednakże określenie parametrów modeli jest jednym z najważniejszych elementów całej procedury i zdecydowałem się zamieścić taki opis, aby każdy zainteresowany mógł go zweryfikować.

Model 1

Pierwszy model wyjaśnia zależności między poprawnymi prognozami sondażowymi dotyczącymi zwycięzcy wyborów (0 – niepoprawne wskazanie; 1 – poprawne wskazanie) a czasem do wyborów. Prawdopodobieństwo poprawnego wskazania wyniku wyborów przez sondaże w czasie t zostało oznaczone jako θ_t . Całkowita trafność sondaży zależy od tego, jak dobrze prognozują one w różnych odstępach czasu przed kampanią. Ze względu na niewielką liczebność danych zdecydowałem się utworzyć zmienną kategoryjną o wartościach 1 (najwyżej miesiąc do wyborów), 2 (od miesiąca do 3 miesięcy), 3 (od 3 miesięcy do pół roku), 4 (od 7 do 12 miesięcy). Wartości zmiennej zależnej mają postać binarną, co można zapisać jako $y_{it} \sim \text{dbern}(\theta_t)$, gdzie y_{it} oznacza dane z sondażu i w czasie t , θ_t prawdopodobieństwo wskazania zwycięzcy wyborów przez sondaże w czasie t , a $\sim \text{dbern}$ jest zapisem rozkładu Bernoulli'ego. Ze względów poznawczych najłatwiej przedstawić najbardziej prawdopodobne wartości θ_t jako rozkład normalny. Aby je wywnioskować za pomocą teorematu Bayesa trzeba jednak przyjąć założenia dotyczące wstępnego rozkładu θ_t . Opierając się na zgodności rozkładów przyjmuję, że jest to rozkład beta (zob. Jackman 2009: 15–16). Zamiast jednak używać stałych parametrów kształtu rozkładu przeformułuję równanie tak, aby rozkład był opisany za pomocą dominanty ω i koncentracji κ : $\theta_t \sim \text{dbeta}(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1)$. Zapis ten ilustruje relacje między wynikami poszczególnych sondaży. Przyjmuję, że wyniki skupiają się wokół wartości modalnej, natomiast oczywiście część z nich jest poniżej tej wartości, a część powyżej. O tym, jak bardzo wyniki od niej odstają, decyduje parametr κ w taki sposób, że im większe κ , tym bliżej θ_t jest wartości ω . Zabieg ten ma na celu oszacowanie dominanty dla wszystkich sondaży. Podobnie jak w przypadku parametru θ_t wnioskowanie na temat końcowej (posterior) wartości ω wymaga przyjęcia założeń dotyczących jej wstępnego rozkładu. Ponieważ nie dysponuję danymi dotyczącymi tego, jak dobrze prognozowały sondaże w Polsce, to zakładam, że możliwa jest każda wartość (rozkład jednostajny). Dla rozkładu beta, będzie to: $\omega \sim \text{beta}(1, 1)$. Dodatkowym parametrem w modelu jest κ , którego również nie znam, i którego najbardziej prawdopodobne wartości zostaną oszacowane w symulacji. Przyjmuję, że wstępne wartości κ mają rozkład gamma o parametrze kształtu 0,0001 i parametrze skali 100. Zdecydowałem się na ten rozkład, ponieważ: 1) κ ma zawsze wartości dodatnie, a więc rozkład

musi odzwierciedlać te właściwości; 2) przyjęte parametry rozkładu sprawiają, że gęstość prawdopodobieństwa dla dowolnej wartości jest jednakowa, dzięki czemu rozkład nie rzutuje bardzo silnie na końcowe wartości κ .

Hipoteza robocza, którą weryfikuje powyższy model brzmi następująco:

H1. Trafność przewidywań sondaży jest najwyższa w miesiącu, w którym odbywają się wybory i stopniowo maleje, im więcej czasu dzieli datę sondażu od daty wyborów. Uzasadnienie tej hipotezy dostarczają klasyczne badania Fishbeina i Coombsa (1974) na temat postaw wyborczych. Im krótszy czas między pomiarem a badanym zachowaniem, tym większe prawdopodobieństwo, że pomiar odzwierciedla rzeczywistą wartość szacowanego parametru.

Model II

Konstrukcja drugiego modelu jest analogiczna do pierwszego. Bez zmian pozostaje zmienna wyjaśniana. Zmienną niezależną jest różnica punktów procentowych między dwoma głównymi oponentami. Przyjmuje ona trzy wartości: 1. Do 6 punktów procentowych, 2. Od 6 do 12 punktów procentowych, 3. Powyżej 12 punktów procentowych. Powyższe kategorie uzasadniam względami praktycznymi zorientowanymi na wartość badawczą hipotez, nie zaś dokładnymi wyliczeniami statystycznymi³. Kieruję się utrwalonymi przekonaniem społecznymi, według których margines błędu wynosi około 3 punktów procentowych⁴. W związku z tym różnica 6 punktów oznacza sytuację, kiedy liderzy sondaży mogą mieć w rzeczywistości bardzo zbliżone poparcie. Różnica między 6 a 12 punktów procentowych informuje, że jedna z partii ma wyraźnie większe poparcie, choć nie musi to być jeszcze przewaga bardzo znacząca. Różnice powyżej 12 wskazują już na dość dużą przewagę jednej z partii. Za liderów wyścigów uznane zostały partie, które w wyborach uzyskały pierwsze i drugie miejsce, czyli: w 1993 roku SLD i PSL; w 1997 roku AWS i SLD; w 2001 roku SLD i PO; w 2005 roku PiS i PO; w 2007, 2011, 2015 roku PO i PiS.

Badany parametr, czyli prawdopodobieństwo poprawnego wskazania przez sondaże zwycięzcy wyborów przy różnicy d oznaczam jako θ_d . Funkcja wiarygodności ma rozkład Bernoulliego: $y_{i|d} \sim dbern(\theta_d)$. Z kolei prawdopodobieństwo wstępne przyjmuje postać rozkładu beta wyrażonego jako $\theta_d \sim dbeta(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1)$, w którym założenia dotyczące ω i κ pozostają bez zmian w stosunku do modelu I.

³ Dokładne wyliczenia nie są możliwe, ponieważ bardzo często brakuje informacji o liczbie respondentów, którzy odpowiedzieli na pytanie.

⁴ W rzeczywistości błąd najprawdopodobniej waha się od 1,9% (95% przedział ufności, szacowana wielkość frakcji 0,05 – próg wyborczy, wielkość próby – 500 osób deklarujących udział w wyborach) do 4,4% (szacowana wielkość frakcji 0,5, pozostałe argumenty równania bez zmian).

Zadaniem modelu jest przetestowanie hipotezy (H2), według której trafność sondaży jest tym większa, im większa jest przewaga komitetu prowadzącego w sondażach nad komitetem, który zajmuje drugie miejsce. Uzasadnienie jest następujące: niewielkie różnice poparcia wskazują na wyrównane szanse, a co za tym idzie – ostateczny wynik zależy od trudnych do uchwycenia czynników, a prognoza jest mniej prawdopodobna. Im przewaga prowadzącego komitetu jest wyraźniejsza, tym większe prawdopodobieństwo, że w dniu wyborów nie zostanie on prześcignięty.

Model III

Trzeci model porusza nieco inną kwestię niż dwa poprzednie. Wiarygodność sondaży przedwyborczych może być oceniana na podstawie tego, czy trafnie wskazuje zwycięzcę wyborów, ale też od tego, jak dokładnie przewiduje poparcie dla każdego komitetu. W związku z tym zmienną zależną tym razem jest skala błędu między prognozowanym poparciem a wynikiem wyborczym. Badanie dotyczy jedynie wyników ugrupowań, które w danym roku dostały się do parlamentu. Zmienna przyjmuje dwie wartości. Uznaję, że prognoza była dokładna, jeśli różnica między wynikiem sondażowym a wyborczym nie przekracza 3 punktów procentowych. W przeciwnym razie prognoza zostaje uznana za nietrafioną. Zmienną niezależną jest czas do wyborów, która przyjmuje cztery już wcześniej przedstawione kategorie. Podobnie jak w poprzednich modelach zmienna zależna jest dychotomiczna i oznaczam ją jako θ_e , gdzie e jest skalą błędu. Funkcja wiarygodności ma rozkład Bernoulliego: $y_{ile} \sim dbern(\theta_e)$ a prawdopodobieństwo wstępne przyjmuje rozkład beta: $\theta_d \sim dbeta(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1)$.

Hipoteza (H3), którą testuję za pomocą powyższego modelu, jest bardzo zbliżona do pierwszej i opiera się na tych samych przesłankach: w miesiącu wyborów jest największe prawdopodobieństwo, że sondaże określą prawidłowy wynik poszczególnych komitetów wyborczych. Zmniejsza się ono stopniowo, im dłuższy jest okres między prognozą a datą wyborów.

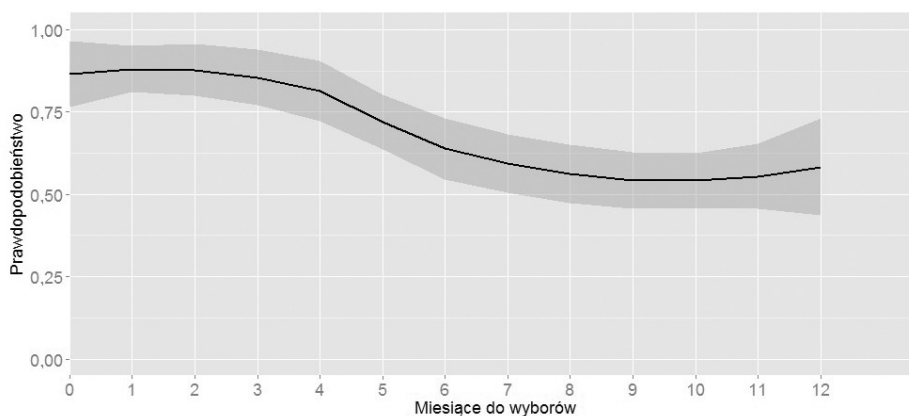
Analiza i interpretacja wyników

Model I

Przed implementacją modeli warto dokonać analizy eksploracyjnej danych (patrz wykres 1.). Wynika z niej, że dotychczas w okresie jednego roku przed wyborami prawdopodobieństwo poprawnego wskazania przez sondaż zwycięzcy wyborów było przez większość czasu powyżej 0,5. Można to zinterpretować w taki sposób, że lepiej polegać na sondażach niż prognozować, kto wygra na podstawie rzutu monety. Jedyne zastrzeżenie dotyczy sondaży, które były przeprowadzane wcześniej niż siedem miesięcy przed wyborami. Granice błędu

standardowego średniej (ciemnoszary obszar na wykresie) w tym przypadku przekraczają dość wyraźnie poziom 0,5, co wskazuje, że być może równie dobrym sposobem na wskazanie, kto wygra w wyborach parlamentarnych, byłyby wspomniani wcześniej rzut. Druga ważna konkluzja, którą można wysnuć na podstawie wykresu dotyczy zauważalnej zależności między prawdopodobieństwem przewidzenia zwycięzcy wyborów a czasem, w jakim został przeprowadzony sondaż. Przedstawiona na wykresie linia średnich lokalnych wskazuje, że moc predykcyjna sondaży rośnie, choć nie jest to funkcja liniowa.

Wykres 1. Prawdopodobieństwo poprawnego wskazania przez sondaż zwycięzcy wyborów parlamentarnych



Źródło: *Opracowanie własne*

Wyniki symulacji Monte Carlo potwierdzają i precyzują wnioski z analizy eksploracyjnej. Ogólne prawdopodobieństwo poprawnego wskazania przez sondaż zwycięzcy wyborów w okresie 12 miesięcy przed wyborami wyniosło 0,8. Jest to wartość bardzo wysoka, ale należy pamiętać, że w dużej mierze wpłynęły na nią wyniki sondaży robionych najbliższej wyborów, które są najtrafniejsze, a samych badań jest w tym czasie najwięcej. Zróżnicowanie to potwierdza bardzo szeroki 95% przedział najwyższej gęstości (HDI): 0,68–0,97. Warto zwrócić uwagę, że nie zahacza on o wartość 0,5, co oznacza, że biorąc pod uwagę dane historyczne z polskich wyborów parlamentarnych większość sondaży w okresie roku przed wyborami już poprawnie wskazywała zwycięzcę.

Zebrane dane nie pozwalają odrzucić hipotezy H1. Faktycznie sondaże najlepiej prognozują w miesiącu, w którym odbywają się wybory, ale bardzo podobne wyniki osiągają w okresie do 7 miesięcy przed wyborami (różnice między nimi nie są istotne statystycznie). Moc predykcyjna za to wyraźnie spada, gdy do wyborów jest więcej niż 6 miesięcy. Dominanta prawdopodobieństwa wynosi 0,59,

ale przedział najwyższej gęstości spada poniżej wartości 0,5. W związku z tym można powiedzieć, że w tym okresie sondaże bardzo słabo radzą sobie z prognozowaniem, a ich wyniki nie są lepsze, niż gdyby losować zwycięzcę za pomocą rzutu monety (ograniczając losowanie do dwóch faworytów).

Tabela 1. Prawdopodobieństwo poprawnego wskazania w sondażu zwycięzcy wyborów w zależności od czasu¹

Parametr	Dominanta	95% HDI	
ω	0,80	0,68	0,97
κ	9,23	2,0002	56,35
θ_{1m}	0,87	0,78	0,94
θ_{2-3m}	0,84	0,74	0,91
θ_{4-6m}	0,81	0,72	0,88
θ_{7-12m}	0,59	0,49	0,68
$\theta_{1m} - \theta_{2-3m}$	0,04	-0,07	0,15
$\theta_{1m} - \theta_{4-6m}$	0,07	-0,04	0,18
$\theta_{1m} - \theta_{7-12m}^*$	0,29	0,15	0,40
$\theta_{2-3m} - \theta_{4-6m}$	0,03	-0,08	0,14
$\theta_{2-3m} - \theta_{7-12m}^*$	0,24	0,12	0,36
$\theta_{4-6m} - \theta_{7-12m}^*$	0,21	0,09	0,32

* Istotne różnice

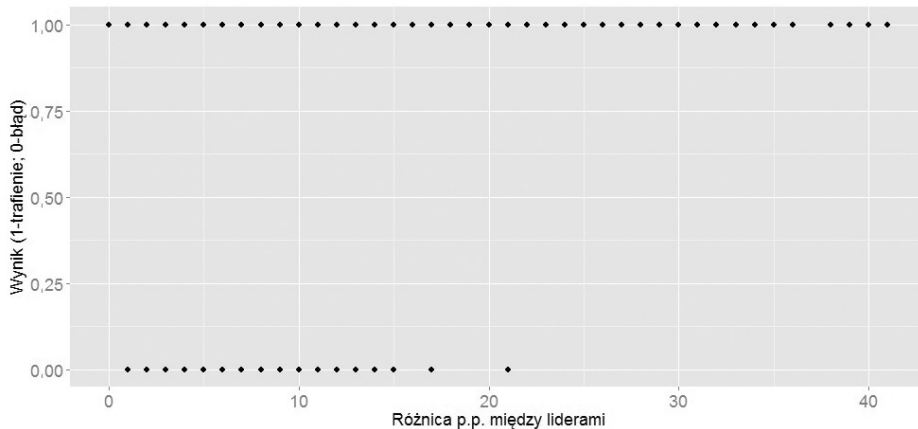
¹ W symulacji uruchomiono trzy łańcuchy z losowym punktem rozpoczęcia; na jeden łańcuch przypadło 66650 iteracji; ESS dla wszystkich parametrów powyżej 10000; statystyka Gelmana-Rubina dla wszystkich parametrów równa 1.

Model II

Drugi model opisuje zależność między prawdopodobieństwem trafnego wskazania przez sondaż zwycięzcy wyborów a różnicą punktów procentowych, która według sondażu dzieliła liderów. Należy zaznaczyć, że wykres 2. pokazuje tylko część danych. Poparcie w sondażach było wyrażane liczbami całkowitymi, co skutkuje tym, że poszczególne punkty na wykresie nakładają się na siebie (np. 14 sondaży mogło odnotować różnicę między liderami rzędu 22 punktów procentowych, ale i tak tej wartości odpowiada jeden punkt). Niemniej taka prezentacja graficzna umożliwia wyciągnięcie wstępnych wniosków. W sytuacji, kiedy różnice między liderami są poniżej 10 punktów procentowych, sondaże notują zarówno prawidłowe wskazania, jak i pomyłki. Im większa różnica, tym większe prawdopodobieństwo, że sondaż poprawnie przewidzi zwycięzcę. Dane historyczne wskazują, że żaden sondaż się nie pomylił, kiedy różnica między liderami wynosiła powyżej 21 punktów procentowych. Z punktu widzenia realiów badawczych jest to jednak poziom bardzo wysoki i rzadko osiągany

w historii polskich wyborów parlamentarnych (dotychczas tylko raz w 2001 roku, kiedy SLD prowadził bardzo wyraźnie przez cały rok przed wyborami). Warto w związku z tym przyjrzeć się wynikom symulacji Monte Carlo.

Wykres 2. Różnica poparcia między liderami wyścigu wyborczego a poprawne wskazania zwycięzcy wyborów parlamentarnych



Źródło: *Opracowanie własne*

Tabela 2. Prawdopodobieństwo poprawnego wskazania w sondażu zwycięzcy wyborów w zależności od różnicy poparcia między komitetami prowadzącymi w wyborach¹

Parametr	Dominanta	95% HDI	
ω	0,78	0,55	0,99
κ	5,12	2,0001	26,88
θ_{pon6pp}	0,55	0,45	0,63
$\theta_{\text{6-12pp}}$	0,74	0,65	0,82
θ_{pow12pp}	0,91	0,86	0,95
$\theta_{\text{pon6pp}} - \theta_{\text{6-12pp}}^*$	-0,20	-0,32	-0,07
$\theta_{\text{pon6pp}} - \theta_{\text{pow12pp}}^*$	-0,37	-0,47	-0,26
$\theta_{\text{6-12pp}} - \theta_{\text{pow12pp}}^*$	-0,17	-0,28	-0,07

* Istotne różnice

¹ W symulacji uruchomiono trzy łańcuchy z losowym punktem rozpoczęcia; na jeden łańcuch przypadło 66650 iteracji; ESS dla wszystkich parametrów powyżej 10000; statystyka Gelmana-Rubina dla wszystkich parametrów równa 1.

Źródło: *Opracowanie własne*

Symulacja potwierdza hipotezę H2, zgodnie z którą prawdopodobieństwo poprawnego wskazania zwycięzcy wyborów wzrasta wraz z rosnącą przewagą lidera sondaży. Różnice poniżej 6 punktów procentowych sprawiają, że sondaż z bardzo podobnym prawdopodobieństwem może wskazać poprawnie wynik, co się pomylić 95% przedział największej gęstości obejmuje zakres prawdopodobieństwa od 0,45 do 0,63, a dominanta dla θ_{pon6pp} jest równa 0,55. Zwiększenie się różnicy do poziomu od 6 do 12 punktów procentowych skutkuje znacznym wzrostem wartości predykcyjnej sondaży. Prawdopodobieństwo, że sondaż wskaże w takiej sytuacji poprawny wynik wyniosło 0,74. Największą moc predykcyjną sondaż posiada, kiedy lider notowań ma ponad 12 punktów procentowych przewagi nad kolejnym ugrupowaniem – dominanta dla parametru θ_{pow12pp} wynosi wówczas 0,91.

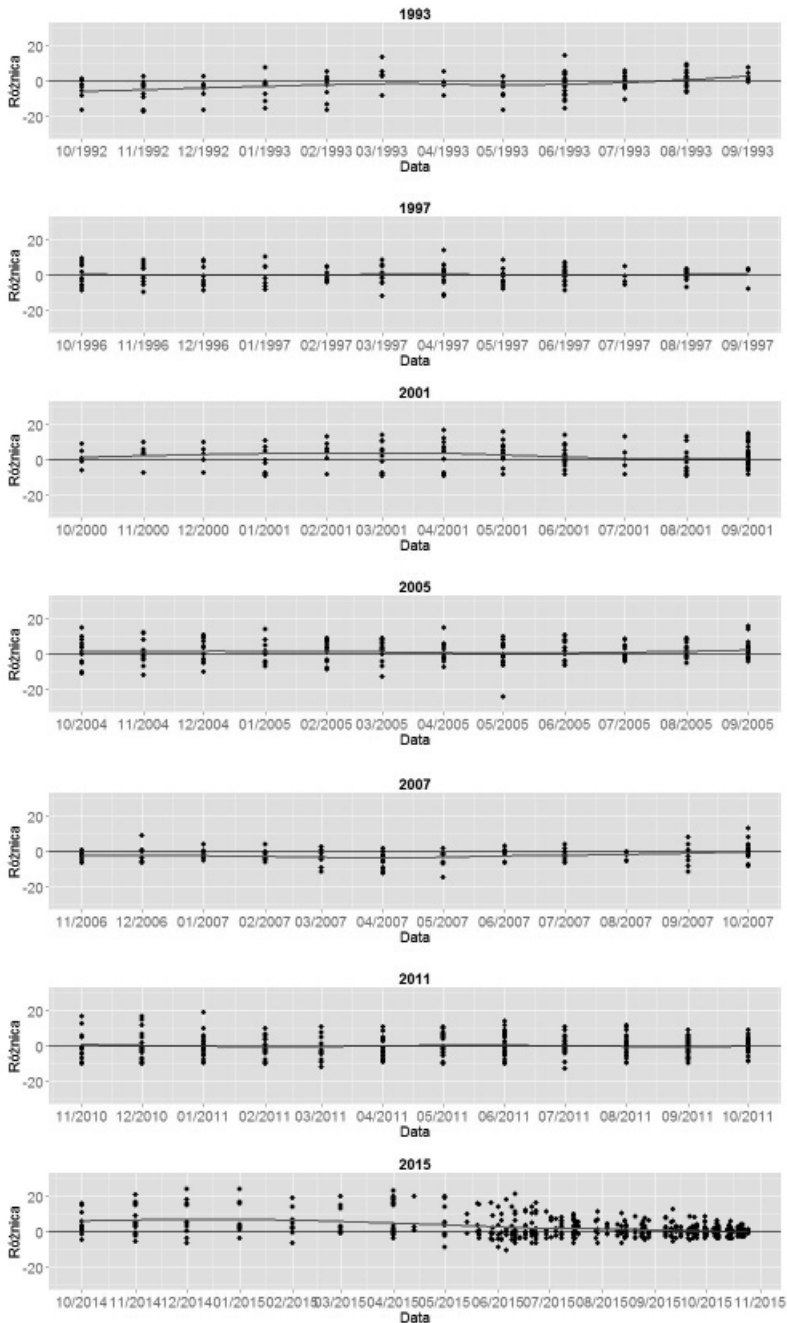
Model III

Ostatni model dotyczy tego, jak dokładnie sondaże przewidują wynik wyborczy partii politycznych. Na podstawie analizy eksploracyjnej można powiedzieć, że sondaże tym dokładniej wskazują wynik wyborczy, im bliżej daty wyborów. Zarówno wykres (wykres 3.), jak i statystyki opisowe (tabela 3.) wskazują na wyraźną prawidłowość. Sondaże bardzo dobrze przewidują wyniki wyborów, ale pod warunkiem, że ich wskazania zostaną uśrednione. Zakres błędu⁵ dla lat 1993–2015 waha się od -0,17 punktu procentowego w ostatnim miesiącu przed wyborami (niedoszacowanie wyniku) do 1,02 punktu procentowego w okresie od 7 do 12 miesięcy przed wyborami. Na te bardzo dobre wyniki dla sondaży warto spojrzeć uwzględniając odchylenie standardowe. Okazuje się, że jest ono relatywnie wysokie: od 3,88 w ostatnim miesiącu przed wyborami do 7,49 w okresie 7–12 miesięcy przed wyborami. Oznacza to, że uśrednione wyniki sondaży są bardzo trafne, ale wyniki pojedynczych sondaży są z kolei bardzo zróżnicowane: część z nich jest wyraźnie przeszacowana, a część niedoszacowana. W ostatnim miesiącu przed wyborami 95% wyników mieści się w zakresie błędu od -7,43 do 7,78 punktów procentowych. Tak duże różnice znacząco utrudniają prognozowanie.

Model hierarchiczny, który został opisany w poprzedniej części artykułu, potwierdza wnioski płynące z analizy eksploracyjnej. Biorąc pod uwagę okres 12 miesięcy przed wyborami, prawdopodobieństwo, że sondaż przewidzi wynik dowolnego komitetu w granicach błędu rzędu 3 punktów procentowych, wynosi 0,47 (dominanta), a 95% przedział najwyższej gęstości obejmuje wartości od 0,23 do 0,69. Oczywiście na tak słaby wynik wpływa fakt szerokiego zakresu czasowego – raczej nie oczekuje się od sondażu robionego na 12 miesięcy przed wyborami, że przewidzi dokładny rozkład poparcia dla wszystkich komitetów.

⁵ Mierzony jako średnia różnica między wynikiem wyborów a wynikiem sondażowym.

Wykres 3. Różnice procentowe między poparciem partii w sondażach a wynikiem wyborczym w zależności od czasu



Źródło: Opracowanie własne

Tabela 3. Miary statystyczne opisujące różnice między poparciem partii w sondażach a wynikiem wyborczym (w punktach procentowych)

Czas do wyborów (w miesiącach)	Różnice między wynikiem sondaży a wynikiem wyborów	Data wyborów							
		1993	1997	2001	2005	2007	2011	2015	1993–2015
0–1	Średnia	0,94	0,29	0,37	1,70	-0,56	-0,39	0,10	0,17
	Odchylenie stand.	3,64	3,61	5,31	4,96	5,43	3,96	2,33	3,88
2–3	Średnia	-1,30	-0,26	1,17	0,97	-2,17	-0,15	0,80	0,20
	Odchylenie stand.	5,75	4,51	6,56	5,43	2,83	6,10	3,84	4,98
4–6	Średnia	-1,39	0,15	3,50	0,64	-3,79	0,20	2,20	0,94
	Odchylenie stand.	6,48	5,63	7,80	7,04	4,56	6,66	6,49	6,72
7–12	Średnia	-4,58	-0,02	2,73	1,35	-2,41	-0,25	6,25	1,02
	Odchylenie stand.	6,50	5,57	6,32	6,65	3,53	7,17	8,20	7,49

Źródło: *Opracowanie własne*

Wyniki symulacji Monte Carlo wskazują, że jedynie w okresie miesiąca przed wyborami prawdopodobieństwo poprawnego wskazania wyniku wyborczego z dokładnością do 3 punktów procentowych jest istotnie większe niż 0,5 i wynosi 0,69 (95% HDI od 0,64 do 0,73). Oznacza to, że wartość sondaży jest wówczas umiarkowanie dobra, a opieranie się na nich wyraźnie poprawia trafność prognozy wyborczej. W okresie powyżej miesiąca do wyborów wartość ta zdecydowanie spada i właściwie jest większe prawdopodobieństwo, że sondaż wskaże poparcie, które będzie się różniło o ponad 3 punkty procentowe z wynikiem wyborczym, niż że poda odsetek mieszczący się w tych granicach.

Tabela 4. Prawdopodobieństwo podania przez sondaż poprawnego wyniku wyborczego w zależności od czasu¹

Parametr	Dominanta	95% HDI	
ω	0,47	0,23	0,69
κ	7,85	2,001	37,98
θ_{1m}	0,69	0,64	0,73
θ_{2-3m}	0,47	0,42	0,54
θ_{4-6m}	0,37	0,32	0,42
θ_{7-12m}	0,35	0,31	0,39
$\theta_{1m} - \theta_{2-3m}^*$	0,21	0,14	0,28
$\theta_{1m} - \theta_{4-6m}^*$	0,32	0,25	0,38
$\theta_{1m} - \theta_{7-12m}^*$	0,34	0,28	0,40
$\theta_{2-3m} - \theta_{4-6m}^*$	0,11	0,03	0,18
$\theta_{2-3m} - \theta_{7-12m}^*$	0,13	0,06	0,20
$\theta_{4-6m} - \theta_{7-12m}$	0,02	-0,04	0,09

* Istotne różnice

¹ W symulacji uruchomiono trzy łańcuchy z losowym punktem rozpoczęcia; na jeden łańcuch przypadło 33300 iteracji; ESS dla wszystkich parametrów powyżej 10000; statystyka Gelmana-Rubina dla wszystkich parametrów równa 1.

Dyskusja i wnioski

Biorąc pod uwagę okres jednego roku przed wyborami należy stwierdzić, że sondaże dość dobrze wskazywały, który komitet uzyska największe poparcie. Prawdopodobieństwo to wynosiło około 0,8, co oznacza, że mniej więcej 80% prognoz było trafnych. Dość szeroki przedział największej gęstości wskazuje, że wartość ta może być uzależniona od innych czynników, choć z 95% prawdopodobieństwem nigdy nie spada poniżej 0,5. Dużo gorsze wyniki osiągnęły sondaże w zakresie przewidywania odsetka oddanych głosów na poszczególne komitety. W badanym okresie prawdopodobieństwo poprawnej prognozy tego typu wyniosło 0,47, co oznacza, że jest najbardziej prawdopodobne, że sondaż nie wskaże poprawnie wyniku nawet przy uwzględnieniu marginesu błędu 3 punktów procentowych. Oczywiście, okres jednego roku przed wyborami to dość długi czas i należy do tych wyników podchodzić dość ostrożnie. Więcej informacji dostarczają obliczenia związane ze sformułowanymi hipotezami.

Na podstawie przeprowadzonych analiz należy częściowo potwierdzić hipotezę 1 i 3. Symulacje wskazują, że w miesiącu wyborów jest największe

prawdopodobieństwo, że sondaże poprawnie wskażą zwycięski komitet oraz w miesiącu wyborów jest największe prawdopodobieństwo, że sondaże określą prawidłowy wynik poszczególnych komitetów wyborczych (z marginesem błędu 3 p.p.). Nie potwierdza się jednak zależność liniowa. W pierwszym przypadku sondaże prognozowały na statystycznie takim samym poziomie w okresie miesiąca, 2–4 miesięcy i 4–6 miesięcy przed wyborami. Inaczej mówiąc, już na sześć miesięcy przed wyborami można ostrożnie wskazywać zwycięzcę wyborów i jest większe prawdopodobieństwo, że prognoza się spełni, niż że okaże się błędna. W drugim przypadku sondaże przewidywały poprawny wynik komitetu jedynie w miesiącu wyborów. Prognozy robione wcześniej najczęściej się nie sprawdzały. Moc predykcyjna w tym zakresie spada do poziomu 0,32 na 4–6 miesiące przed wyborami i pozostaje w okolicach tej wartości w okresie od 7 do 12 miesięcy przed wyborami.

Powyższe wyniki można uzasadnić w ramach obowiązującej literatury przedmiotu. Na wczesnych etapach rywalizacji wyborczej postawy elektoratu są słabsze niż pod koniec kampanii. Jednocześnie wówczas obywatele mają dostęp do mniejszej liczby informacji, a także mają mniejszą motywację, aby te informacje zdobywać (por. Gelman i King 1993; Visser, Holbrook, i Krosnick 2008).

Potwierdziła się również hipoteza H2, według której sondaże lepiej wskazują zwycięzcę wyborów, im większa jest różnica poparcia między komitetem prowadzącym w badaniach a drugim w kolejności. Małe różnice ze względu na margines błędu przy szacowaniu na podstawie próby statystycznej mogą oznaczać brak faktycznych różnic lub prowadzenie innego komitetu. W takich sytuacjach rośnie znaczenie kampanii, wizerunku, obietnic i debat (por. Gelman i King 1993), a o końcowym wyniku mogą zdecydować niuanse. Prognozy wówczas są bardzo niepewne.

Zestawienie czasu do wyborów oraz różnicy poparcia między głównymi oponentami oraz poszczególnych wyborów parlamentarnych (patrz wykres 4. i tabela 5.) umożliwia bardziej pogłębioną analizę zaobserwowanych zależności. Wynika z niej, że każde dotychczasowe wybory parlamentarne tworzyły dość unikatowy przypadek:

- Wybory w 1993 roku (liderzy sondaży: SLD i PSL): Poprawność wskazań sondaży poprawiała się wraz ze zbliżającym się terminem wyborów, jednak ich moc predykcyjna przez większość czasu była dość słaba. Powodem tego stanu rzeczy było długie prowadzenie w sondażach PSL, które ostatecznie osiągnęło wynik gorszy niż SLD. Dopiero dwa miesiące przed wyborami większość sondaży zaczęła poprawnie wskazywać zwycięstwo SLD, a różnice między ugrupowaniami rosły aż do daty wyborów.
- Wybory w 1997 roku (liderzy sondaży: AWS i SLD): Do ostatniego miesiąca przed wyborami AWS zachowywał niewielką (poniżej 5 punktów procentowych) przewagę nad SLD. W ostatnim miesiącu tendencja ta

się odwróciła. Uśredniając AWS pozostawał wówczas niedoszacowany, a wszystkie sondaże wskazywały zwycięstwo partii lewicowej. W konsekwencji prawdopodobieństwo, że zostanie poprawnie przewidziany wynik wyborów, spadło do 0 w miesiącu wyborów.

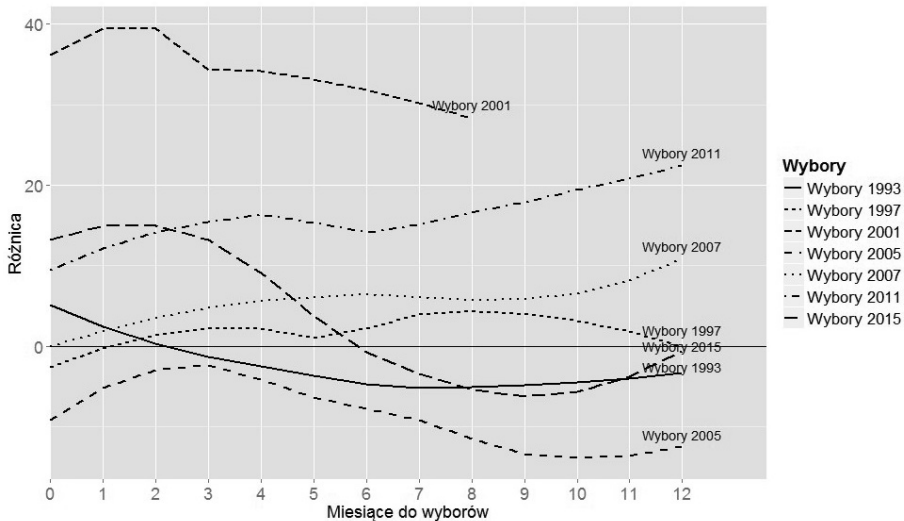
- Wybory w 2001 roku (liderzy sondaży: SLD i PO): Przez cały badany okres utrzymywała się zdecydowana, sięgająca 20 punktów procentowych przewaga SLD nad PO (patrz wykres 2.). W związku z tym żaden z sondaży się w tym czasie nie pomylił, a więc jednocześnie nie wzrosła w tym czasie moc predykcyjna badań surveyowych.
- Wybory w 2005 roku (liderzy sondaży: PiS i PO): Przez cały badany okres zdecydowana większość sondaży wskazywała na sukces PO. Średnie różnice między liderami w okresie 2–4 miesięcy przed wyborami spadły do wartości poniżej 6 punktów procentowych, ale w ostatnim miesiącu znowu wzrosły na korzyść Platformy Obywatelskiej. W konsekwencji wszystkie sondaże na miesiąc przed wyborami nieprawidłowo wskazały zwycięzcę.
- Wybory w 2007 roku (liderzy sondaży: PO i PiS): Wraz ze zbliżającą się datą wyborów malały różnice między PO i PiS, aż w ostatnim miesiącu kampanii wyborczej średnio wynosiły one poniżej 3 punktów procentowych. Z tego względu przez całą kampanię wartość predykcyjna sondaży spadała.
- Wybory w 2011 roku (liderzy sondaży: PO i PiS): przez cały badany okres PO utrzymywała wyraźną przewagę nad PiS, a sondaże trafnie przewidywały jej zwycięstwo.
- Wybory w 2015 roku (liderzy sondaży: PiS i PO): trafność przewidywań rosła w czasie. PiS, po zwycięstwie Andrzeja Dudy w wyborach prezydenckich, stał się liderem sondaży i z czasem powiększał swoją przewagę nad PO.

Rozpatrując zależność między czasem do wyborów a poprawnymi wskazaniami przez sondaże zwycięzcy, trzeba uwzględnić, że w dwóch przypadkach moc predykcyjna sondaży się nie zmieniała (nie mogła urosnąć, bo od samego początku wytyczonego okresu sondaże poprawnie wskazywały zwycięzcę), w trzech przypadkach malała (w 2005 roku trend miał charakter paraboliczny) i w dwóch przypadkach rosła. Agregując te wyniki należy zauważyć, że tendencje te nakładają się na siebie i wzajemnie znoszą. W konsekwencji, biorąc pod uwagę wszystkie dane, zauważalne jest polepszenie się mocy predykcyjnej sondaży wraz z malejącym czasem do daty wyborów, ale nie jest to trend charakterystyczny dla wszystkich przypadków. Oznacza to, że mają znaczenie czynniki inne niż sam czas.

Warto zwrócić uwagę, że błędy sondaży miały różne źródła. Między innymi wynikały z niewielkich różnic między prowadzącymi komitetami, ale np. w 1997 roku i 2005 roku okazało się, że większość sondaży w ogóle błędnie wskazywało zwycięzcę (w 1997 roku przy niewielkiej różnicy punktów procentowych w stosunku do komitetu notowanego jako drugi, a w 2005 roku przy

wyraźnej średniej różnicy bliskiej 10 p.p. na miesiąc przed wyborami). Problematyczne w kontekście tematu artykułu jest to, że informacja o tym, czy jakaś opcja polityczna była niedoszacowana albo przeszacowana, dostępna jest dopiero po wyborach. Do tego czasu nie wiadomo, czy konieczna jest korekta przewidywań, a jeśli tak, to w stosunku do których komitetów. W konsekwencji obniża to dość znacząco wartość predykcyjną sondaży rozumianą jako ich zdolność do wskazywania zwycięzcy przed faktycznymi wyborami.

Wykres 4. Różnica poparcia między dwoma komitetami o najwyższych notowaniach¹ a rok wyborów i czas do wyborów



Źródło: Opracowanie własne

¹ Liczona jako różnica między aktualnym poparciem komitetu, który zwyciężył w wyborach i komitetu, który zajął drugie miejsce.

Tabela 5. Prawdopodobieństwo poprawnego wskazania przez sondaż zwycięzcy wyborów parlamentarnych a czas do wyborów

Czas do wyborów (w miesiącach)	Rok						
	1993	1997	2001	2005	2007	2011	2015
0–1	1	0	1	0	0,67	1	1
2–3	0,57	0,33	1	0,4	0,5	1	1
4–6	0,29	0,6	1	0,29	0,83	1	0,94
7–12	0	0,71	1	0	0,94	1	0,32

Źródło: Opracowanie własne

Na zakończenie warto zwrócić uwagę, że choć badania sondażowe stanowią jeden z najbardziej rozpowszechnionych sposobów prognozowania wyborów, to nie są jedynym. Część badaczy wskazuje, że dobrym predyktorem jest pytanie respondentów o to, kto prawdopodobnie wygra wybory (zob. Graefe 2014; Graefe i Armstrong 2008; Lewis-Beck i Stegmaier 2011; Miller i in. 2012; Rothschild i Wolfers 2011). Niektórzy wskazują na duże znaczenie brania pod uwagę wskazań różnych pracowni badawczych, dzięki któremu redukują się błędy wynikające z używanych przez nich różnych metodologii (zob. Jackman 2005; Traugott 2015). Popularne są modele regresji oparte na danych historycznych, które obejmują między innymi czynniki polityczne i ekonomiczne, takie jak poparcie rządu w sondażach, poziom PKB czy stopa bezrobocia (zob. Aichholzer i Willmann 2014; Campbell 2012; Foucault i Nadeau 2012; Lebo i Norporth 2011; Norporth i Gschwend 2010). Wskazówkami pozwalającymi wnioskować o wyniku wyborów mogą też być treści prezentowane w mediach w trakcie kampanii (zob. Bélanger i Soroka 2012; Hopmann, Vliegenthart, Vreese i Albæk 2010; Soroka, Bodet, Young i Andrew 2009), a w związku z rozwojem nowych technologii informacyjnych wyniki wyszukiwań w Wikipedii (zob. Yasseri i Bright 2015) i analiza komentarzy krążących w mediach społecznościowych (zob. Schoen i in. 2013). Część badaczy z kolei jest zwolennikami bardziej jakościowego podejścia oraz badania kampanii i kandydatów w określonym, niepowtarzalnym kontekście (por. Rothenberg 2014). Istnieje zatem bardzo duże spektrum metod prognostycznych. Coraz częściej pojawiają się także opracowania wskazujące na korzyści wynikające z ich łączenia (por. Graefe i in. 2014).

Literatura

- Aichholzer, Julian i Johanna Willmann. 2014. *Forecasting Austrian National Elections: The Grand Coalition model*. „International Journal of Forecasting” 30(1): 55–64.
- Armstrong, J. Scott. 2001. *Combining Forecasts*. W: J. S. Armstrong (red.). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell: Kluwer, s. 417–439.
- Bélanger, Éric i Stuart Soroka. 2012. *Campaigns and the Prediction of Election Outcomes: Can historical and campaign-period prediction models be combined?* „Electoral Studies” 31(4): 702–714.
- Belli, Robert F., Michael W. Traugott, Margaret Young i Katherine A. McGonagle. 1999. *Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring*. „Public Opinion Quarterly” 63(1): 90–108.
- Campbell, James E. 2012. *Forecasting the Presidential and Congressional Elections of 2012: The Trial-Heat and the Seats-in-Trouble Models*. „PS: Political Science & Politics” 45(04): 630–634.

- Campbell, James E. 2014. *Issues in Presidential Election Forecasting: Election Margins, Incumbency, and Model Credibility*. „PS: Political Science & Politics” 47(02): 301–303.
- Converse, Philip E. i Michael W. Traugott. 1986. *Assessing the Accuracy of Polls and Surveys*. „Science” 234(4780): 1094–1098.
- Dlaczego sondaże pobiły. 2010. (22.06.2010) http://www.archiwum.wyborcza.pl/Archiwum/1,0,7238148,20100622RP-DGW,Dlaczego_sondaże_pobiły,zwykly.html, dostęp 18 luty 2015.
- Domański, Henryk i Agnieszka Kublik. 2014. *Są też dobre sondaże*. „Gazeta Wyborcza” 4 (20.02.2014).
- Fishbein, Martin i Fred S. Coombs. 1974. *Basis for Decision: An Attitudinal Analysis of Voting Behavior*. „Journal of Applied Social Psychology” 4(2): 95–124.
- Foucault, Martial i Richard Nadeau. 2012. *Forecasting the 2012 French Presidential Election*. „PS: Political Science & Politics” 45(02): 218–222.
- Gelman, Andrew i Gary King. 1993. *Why are American Presidential Election Campaign Polls so Variable When Votes are so Predictable?* „British Journal of Political Science” 23(4): 409–451.
- Graefe, Andreas. 2014. *Accuracy of Vote Expectation Surveys in Forecasting Elections*. „Public Opinion Quarterly” 78(S1): 204–232.
- Graefe, Andreas i J. Scott Armstrong. 2008. *Forecasting Elections from Voters' Perceptions of Candidates' Ability to Handle Issues* [MPRA Paper]. <http://mpra.ub.uni-muenchen.de/13079> (dostęp 27.05.2015).
- Graefe, Andreas, J. Scott Armstrong, Randall J. Jones Jr. i Alfred G. Cuzán. 2014. *Combining Forecasts: An application to elections*. „International Journal of Forecasting” 30(1): 43–54.
- Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James. M. Lepkowski, Eleanor Singer i Roger Tourangeau. 2009. *Survey Methodology*. Hoboken, N.J.: Wiley.
- Holbrook, Allyson L. i Jon A. Krosnick. 2010. *Social Desirability Bias in Voter Turnout Reports Tests Using the Item Count Technique*. „Public Opinion Quarterly” 74(1): 37–67.
- Hopmann, David N., Rens Vliegenthart, Claes D. Vreese i Erik Albæk. 2010. *Effects of Election News Coverage: How Visibility and Tone Influence Party Choice*. „Political Communication” 27(4): 389–405.
- Jackman, Simon. 2005. *Pooling the Polls Over an Election Campaign*. „Australian Journal of Political Science” 40(4): 499–517.
- Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Chichester: John Wiley & Sons, Inc.
- Kruschke, John. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. San Diego, CA: Academic Press.
- Lebo, Matthew i Helmut Norpoth. 2011. *Yes, Prime Minister: The Key to Forecasting British Elections*. „Electoral Studies” 30(2): 258–263.
- Lewis-Beck, Michael S., Richard Nadeau i Éric Bélanger. 2004. *General Election Forecasts in the United Kingdom: A Political Economy Model*. „Electoral Studies” 23(2): 279–290.

- Lewis-Beck, Michael S. i Mary Stegmaier. 2010. *Modele głosowania ekonomicznego*. W: R. J. Dalton i H.-D. Klingemann (red.), *Zachowania polityczne*. T. 2. Warszawa: Oxford, WN PWN, s. 103–129.
- Lewis-Beck, Michael S. i Mary Stegmaier. 2011. *Citizen forecasting: Can UK voters see the future?* „Electoral Studies” 30(2): 264–268.
- Linzer, Drew A. 2013. *Dynamic Bayesian Forecasting of Presidential Elections in the States*. „Journal of the American Statistical Association” 108(501): 124–134.
- Markowski: *Ratujmy sondaże wyborcze*. 2015. (23.03.2015). <http://wyborcza.pl/1,75478,17641464.html> (dostęp 27.05.2015).
- Matuszewski, Paweł. 2013. *Stan gospodarki a stosunek do rządu. Artefakty wynikające z modelu głosowania retrospektywnego*. „Uniwersyteckie Czasopismo Socjologiczne” (8): 72–82.
- Matuszewski, Paweł. 2015. *Techniki redukcji efektu oczekiwań społecznych w pytaniach o frekwencję wyborczą*. „Studia Politologiczne” 37: 312–333.
- Miller, Michael K., Guanchun Wang, Sanjeev R. Kulkarni, H. Vincent Poor, Daniel N. Osherson, Jeff. R. Dewitt, Jonathan Knuckey i David Moon. 2012. *Citizen Forecasts of the 2008 U.S. Presidential Election*. „Politics & Policy” 40(6): 1019–1052.
- Norpoth, Helmut i Thomas Gschwend. 2010. *The Chancellor Model: Forecasting German elections*. „International Journal of Forecasting” 26(1): 42–53.
- Prawo i Sprawiedliwość przeciw sondażom*. 2006. (28.08.2006). http://www.archiwum.wyborcza.pl/Archiwum/1,0,4703439,20060828RP-DGW,PRAWO_I_SPRAWIEDLIWOSC_PRZECIW_SONDAZOM,.html (dostęp 18.02.2015).
- Rothenberg, Stuart. 2014. *Election Forecasting and Public Opinion Polls*. „PS: Political Science & Politics” 47(02): 336–338.
- Rothschild, David i Neil Malhotra. 2014. *Are public opinion polls self-fulfilling prophecies?* „Research & Politics” 1(2) (<http://rap.sagepub.com/content/1/2/2053168014547667>, dostęp 27.05.2015).
- Rothschild, David i Justin Wolfers. 2011. *Forecasting Elections: Voter Intentions Versus Expectations* (SSRN Scholarly Paper No. ID 1884644). Rochester, NY: Social Science Research Network. <http://papers.ssrn.com/abstract=1884644>, dostęp 27.05.2015.
- Schoen, Harald, Daniel Gayo-Avello, Panagiot Takis Metaxas, Eni Mustafaraj, Markus Strohmaier i Peter Gloor. 2013. *The Power of Prediction with Social Media*. „Internet Research” 23(5): 528–543.
- Sobczyk, Mieczysław. 2002. *Statystyka*. Warszawa: WN PWN.
- Socjologia w kampanii wyborczej 2005. Stanowisko Polskiego Towarzystwa Socjologicznego*. 2005. <https://polacywybieraja.wordpress.com/2010/06/23/socjologia-w-kampanii-wyborczej-2005-stanowisko-polskiego-towarzystwa-socjologicznego/> (dostęp 27.05.2015).
- Sondaż nasz codzienny*. 2004. (23.04.2004). http://www.archiwum.wyborcza.pl/Archiwum/1,0,4040366,20040423RP-CTR,Sondaz_nasz_codzienny,.html (dostęp 18.02.2015).
- Sondażokracja*. 2004. (04.09.2004). <http://www.archiwum.wyborcza.pl/Archiwum/1,0,4022064,20040409RP-DGW,Sondazokracja,.html> (dostęp 18.02.2015).

- Sondażowa nerwica*. 2004. (09.04.2004). http://www.archiwum.wyborcza.pl/Archiwum/1,0,4022078,20040409RP-DGW,Sondazowa_nerwica.html (dostęp 18.02.2015).
- Soroka, Stuart, Marc A. Bodet, Lori Young i Blake Andrew. 2009. *Campaign News and Vote Intentions*. „Journal of Elections, Public Opinion and Parties” 19(4): 359–376.
- Strömbäck, Jesper. 2012. *The Media and their Use of Opinion Polls: Reflecting and Shaping Public Opinion*. W: C. Holtz-Bacha i J. Strömbäck (red.). *Opinion Polls and the Media: Reflecting and Shaping Public Opinion*. New York: Palgrave Macmillan, s. 1–22.
- Traugott, Michael W. 2015. *Alternative Methods for Estimating Election Outcomes*. „Estudios Públicos” 137: 7–42.
- Visser, Penny, Allyson L. Holbrook i Jon A. Krosnick. 2008. *Knowledge and Attitudes*. W: W. Donsbach i M. W. Traugott (red.). *The SAGE Handbook of Public Opinion Research*. London: SAGE Publications Ltd, s. 127–140.
- Wielka kompromitacja sondażowni. Dlaczego sondaże nie przewidziały wyniku wyborów?* 2015. (13.05.2015). <http://opinie.newsweek.pl/wyniki-wyborow-prezydenckich-a-sondaze-blamaz-sondazowni-newsweek-pl,artykuly,363070,1.html> (dostęp 14.05.2015).
- Wybory prezydenckie 2015: Dlaczego sondaże przed I turą okazały się dalekie od prawdy?* 2015. (12.05.2015). <http://polska.newsweek.pl/wyniki-wyborow-prezydenckich-sondaze-wyniki-i-tury,artykuly,363029,1.html> (dostęp 14.05.2015).
- Yasseri, Taha i Jonathan Bright. 2015. *Predicting elections from online information flows: towards theoretically informed models* (<http://arxiv.org/abs/1505.01818>, dostęp 10.09.2015).
- Za błędne sondaże wyborcze odpowiedzialne są głównie media*. 2015. (21.11.2005). <http://wiadomosci.wp.pl/kat,1342,title,Za-bledne-sondaze-wyborcze-odpowiedzia-lne-sa-glownie-media,wid,8097525,wiadomosc.html> (dostęp 17.06.2015).

Can Pre-Election Polls Be Believed? Use of a Bayesian Approach to Analyse the Disparity Between Parliamentary Election Results in Poland and Survey Data

Summary

The aim of this article is to determine how well pre-election polls can predict the results of parliamentary elections, and what determines the accuracy of these predictions. The dependent variables are 1) the correct indication of the winning party and 2) the accuracy of election surveys in forecasting voters' support. The first independent variable is the time between the poll and the date of the election. The second explanatory variable is the difference in results between the two parties with the greatest support. This study uses data from all publicly available polls that took place in the 12 months before every parliamentary election in Poland from 1993 to 2015. The analysis uses Bayesian hierarchical modeling and Markov Chain Monte Carlo simulation. The results show that the average probability that a pre-election

poll will correctly predict the winning party is around 80%, whereas the probability that it will correctly predict the distribution of voters' support (with 3% error margin) is around 50%. The evidence partially proved that the forecasting accuracy of an election poll is the better the closer the poll is taken to the date of the election. It was also proved that the ability of a poll to predict the winner is better the greater the gap between the survey results of the two leading parties.

Key words: political forecasts; pre-election polls accuracy; election polling; Bayes; MCMC.