Maciej Potyra*

Korzyści i ograniczenia związane z wykorzystaniem podejścia probabilistycznego do prognozowania ludności Polski

Wstęp

Podejście stochastyczne w prognozach demograficznych powstało w odpowiedzi na krytykę tradycyjnych prognoz deterministycznych, wskazującą na konieczność dostarczania użytkownikom bardziej precyzyjnej informacji odnośnie do niepewności co do rezultatów prognoz ludności. Najczęściej stosowaną metodą sygnalizowania niepewności w klasycznych prognozach jest przygotowywanie kilku alternatywnych wariantów prognozy. Celem prognoz probabilistycznych jest oszacowanie przedziałów predykcji dla zakładanych współczynników demograficznych oraz wyników prognozy. Prognozy stochastyczne pozwalają również na oszacowanie prawdopodobieństwa praktycznie każdego scenariusza rozwoju sytuacji demograficznej.

Za prekursorską z zakresu wykorzystania podejścia probabilistycznego do prognoz ludności uważa się często pracę Leo Torquista z 1949 r. zatytułowaną *Om de synspunkter, som bestämt valet av de primäre prognosantagendena (O poglądach, które zdecydowały o wyborze głównych założeń prognozy).* Praca ta jednak ciągle dostępna jest tylko w języku szwedzkim.

Metodologia stosowana obecnie powszechnie przy tworzeniu prognoz probabilistycznych zaczęła rozwijać się od początku lat 90. XX w., począwszy od prac Ronalda Lee i Shripada Tuljapurkara [1994]. Istotny wpływ na rozwój metodologii miały następnie prace oraz prognozy dla kolejnych krajów wykonywane m.in. przez Juhę Alho [2002], Nico Keilmana [1997, 2002], Joopa de Beera [1999, 2000] oraz Wolfganga Lutza [1998, 2005]. W Polsce najbardziej kompleksowa prognoza tego typu została przygotowana przez Annę Matysiak i Beatę Nowok [2007]. Należy nadmienić, że podejście stochastyczne jest traktowane ze znaczną rezerwą przez większość krajowych urzędów statystycznych. Jednakże częściową probabilistyczną prognozę (z migracją modelowaną deterministycznie) przygotował ONZ [2013]. Nieoficjalną metodologię tego typu prognozy przedstawił również brytyjski Narodowy Urząd Statystyczny (ONS) [2010].

* Mgr, Szkoła Główna Handlowa, Instytut Statystyki i Demografii, ul. Madalińskiego 6/8, 02-513 Warszawa; Główny Urząd Statystyczny, Departament Badań Demograficznych, mp75131@doktorant.sgh.waw.pl

Pomimo różnic w metodologiach pomiędzy poszczególnymi prognozami stochastycznymi można wskazać ogólne podstawy budowy modelu, które w przeważającej większości z nich są takie same.

Celem artykułu jest przedstawienie podstawowych zasad budowy modelu stochastycznego do prognozowania ludności oraz wskazanie jego mocnych i słabych stron w porównaniu do modeli deterministycznych. Artykuł wskazuje również na problemy z tworzeniem prognoz probabilistycznych związane z dostępnymi w Polsce danymi, w szczególności dotyczącymi migracji. W ostatniej części zaprezentowano wyniki prognozy probabilistycznej dla Polski z 2014 r. oraz ich porównanie z danymi rzeczywistymi oraz prognozą deterministyczną Głównego Urzędu Statystycznego.

1. Podstawy budowy modelu

Modele stochastyczne wychodzą od jednego deterministycznego wariantu prognozy. Zatem odpowiednie uzasadnienie przyjętych założeń jest równie ważne co w przypadku klasycznych prognoz. W związku z tym często jako podstawę do przygotowania prognozy stochastycznej wykorzystuje się główny wariant zaczerpnięty z oficjalnych prognoz wykonanych przez krajowe urzędy statystyczne, Eurostat itp. Zakładane wartości współczynnika dzietności (TFR), oczekiwanego trwania życia dla mężczyzn i kobiet w momencie urodzenia (e0) oraz salda migracji (strumieni imigracji/emigracji) będą stanowić medianę (przeważnie również średnią) wszystkich wariantów obliczonych w ramach prognozy.

W prognozach deterministycznych przeważnie zakłada się brak korelacji pomiędzy zmianami poszczególnych składowych zmiany demograficznej. Założenie to może budzić pewne kontrowersje, gdyż w teorii na przykład korzystne zmiany gospodarcze mogą prowadzić zarówno do wzrostu dzietności, jak i korzystniejszego salda migracji. Jednakże jak wskazują [Lee, Tuljapurkar, 1994] oraz [Keilman i inni, 2002], w krajach zachodnich od wielu lat dane empiryczne nie potwierdzają tego rodzaju zależności. W przypadku Polski analiza danych empirycznych również wskazuje na słuszność tego założenia. Oczywiście, jak słusznie wskazuje N. Keilman, w przypadku prognoz robionych dla krajów słabiej rozwiniętych należałoby przeprowadzić dalsze analizy potwierdzające słuszność tego założenia [Keilman i inni, 2002].

Średni wiek rodzenia, powiązana z nim struktura TFR (*fertility schedule* – tj. procentowy udział poszczególnych roczników kobiet w ogólnej dzietności), oraz struktury migracji według płci i wieku są z reguły prognozowane deterministycznie. Z kolei w przypadku współczynników zgonów (bądź przeżycia) według wieku zakłada się przeważnie całkowite skorelowanie

z przeciętnym oczekiwanym dalszym trwaniem życia dla danej płci, tj. jednej jego wartości odpowiada jeden rozkład współczynników według wieku.

Do tworzenia „ścieżek zmian" poszczególnych współczynników wykorzystuje się modele losowego błądzenia (Random Walk) bądź modele autoregresywne (AR). Wybór modelu jest jednym z czynników wpływających na szerokość przedziałów predykcji. Modele AR o parametrach o wartości absolutnej mniejszej od 1 są stacjonarne, w związku z tym ich zastosowanie sprawia, że szerokość interwałów predykcji stabilizuje się w dłuższej perspektywie. W przypadku modeli Random Walk przedziały te zwiększają się z każdym kolejnym rokiem prognozy.

Celem przygotowania prognozy stochastycznej generuje się od 300 do nawet 10 000 „ścieżek zmian" dla poszczególnych współczynników. Uzyskujemy na tej podstawie znaczną liczbę scenariuszy zmian demograficznych, na podstawie których możemy oszacować przedziały predykcji również dla wartości wynikowych prognozy (tj. ludności według wieku, obciążeń demograficznych itp.). Większość prognoz probabilistycznych powstała w oparciu o 1000 takich scenariuszy.

Kluczowym zagadnieniem przy generowaniu „ścieżek zmian" jest oszacowanie niepewności co do przewidywanych zmian poszczególnych współczynników i tym samym przewidywanie oczekiwanej wielkości pomyłki w kolejnych latach prognozy. Do tego celu wykorzystuje się trzy główne rodzaje metod: eksperckie, obliczenia na podstawie błędów wcześniejszych prognoz oraz oparte na analizie serii czasowych. Metody te są w znacznym stopniu wobec siebie komplementarne i można je na różne sposoby ze sobą łączyć.

W przypadku metod eksperckich powierza się wyznaczenie niepewności odnośnie do zmian czynników demograficznych (tj. dzietności, umieralności i migracji) wybranym specjalistom. W najczęściej spotykanej wersji tej metody wybrani eksperci proszeni są o przedstawienie swoich oszacowań odnośnie do wartości czynników na określony rok wraz z przedziałem, w którym ich zdaniem na określony procent (najczęściej 80 lub 95) zmieści się dana wartość. Szacunki ekspertów następnie uśrednia się. Otrzymujemy na tej podstawie prognozę na określony rok wraz z przedziałem predykcji. Przedziały dla pozostałych lat wyznacza się drogą interpolacji. Niewątpliwym mankamentem tych metod jest w zasadzie całkowita arbitralność wyliczonych przedziałów predykcji. Wybranie innej grupy ekspertów mogłoby całkowicie zmienić wyniki prognozy. Rozliczne wątpliwości może również budzić wnioskowanie o skali niepewności co do przewidywanych współczynników na bazie uśrednionej opinii ekspertów. Przykłady prognoz stochastycznych opartych na takich metodach można znaleźć w pracach W. Lutza [1998, 2005].

Metody oparte na wykorzystaniu informacji o błędach poprzednich prognoz bazują na porównaniu prognoz wykonywanych w przeszłości z danymi empirycznymi. Błędy *ex post* prognoz służą jako informacja o niepewności odnośnie do przewidywań. W założeniu należy przeanalizować jak największą liczbę prognoz i wyników dla poszczególnych lat, aby otrzymać przeciętne błędy przy prognozowaniu na określoną liczbę lat. Istotnym problemem z wykorzystaniem tych metod jest fakt, że wiele prognoz z przeszłości znacznie pomyliło się co do trendów zmian demograficznych. Stworzone w oparciu o takie prognozy przedziały predykcji byłyby tak szerokie, że traciłyby jakąkolwiek wartość informacyjną. Nie można oczywiście wykluczyć, że główny wariant prognozy okaże się całkowicie błędny. Przyjęcie założenia o tak wielkiej niepewności co do przewidywań prognoz stawiałoby jednak pod znakiem zapytania sam sens jej tworzenia.

Metody oparte na analizie serii czasowych są najszersze z tych trzech kategorii. Uwzględniają w praktyce wszelkie metody, które próbują oszacować niepewność na podstawie danych o zmianach danego czynnika w przeszłości. Niewątpliwą zaletą tych metod jest łatwa dostępność danych potrzebnych do ich aplikacji (znacznie łatwiej jest znaleźć dane historyczne dotyczące dzietności, umieralności czy migracji, niż założenia wykonywanych w przeszłości prognoz). Prognoza prezentowana w tym artykule również była oparta na prostej analizie szeregów czasowych.

2. Prognoza probabilistyczna na lata 2014–2050

2.1. Założenia

Przygotowana przeze mnie na początku 2015 r. prognoza stochastyczna powstała w oparciu o *Prognozę ludności na lata 2014–2050*, przygotowaną przez Główny Urząd Statystyczny. Założenia i wyniki tej prognozy posłużyły jako główny wariant prognozy, dla którego oszacowana została niepewność. Przy tworzeniu prognozy nie korzystano z żadnych informacji dotyczących danych za rok 2014. Nie dysponowano zatem żadną dodatkową wiedzą w porównaniu do *Prognozy na lata 2014–2050*. Dlatego też jej wyniki pozwalają wyraźnie wykazać korzyści uzyskane z wykorzystania podejścia probabilistycznego.

Prognoza ta miała w znacznej mierze charakter wstępny i bez wątpienia wymaga dalszych prac metodologicznych. Jednakże, ze względu na datę powstania, daje bezcenną możliwość porównania wyników z danymi empirycznymi dla ostatnich trzech lat.

Punktem wyjścia dla prognozy było obliczenie błędu standardowego prognozy na jeden rok do przodu (SD_1). Wartość ta jest szacunkiem przeciętnego błędu prognozy na jeden rok do przodu dla wszystkich wykorzystywanych współczynników (TFR, e0 dla mężczyzn i kobiet, strumienie

imigracji i emigracji). Wartość SD_1 wyliczona została na podstawie szeregów czasowych z 16 lat (1998–2013) jako średni moduł różnicy pomiędzy wartościami rzeczywistymi zmiennej a wartościami estymowanymi przez trend liniowy, oszacowany dla podanych powyżej lat.

Ścieżki zmian poszczególnych współczynników zostały utworzone przy wykorzystaniu modelu *random walk with drift* (RWD):

$$W_t = W_{t-1} + \varepsilon_t + D_t \tag{1}$$

gdzie:
W_t – współczynnik w danym roku prognozy,
D_t – dryft ($pW_{t-1} - pW_t$ w wariancie będącym medianą prognozy),
ε_t – wartość losowa o rozkładzie N(0, SD1).

W modelu RWD wartość standardowego błędu prognozy jest równa wartości SD_1 pomnożonej przez pierwiastek z liczby lat. Dla przykładu: zakładany, standardowy błąd prognozy na cztery lata do przodu jest dwa razy większy niż na rok do przodu. W modelu jest przyjęte również, że błędy predykcji mają rozkład normalny, zatem przedziały predykcji dla poszczególnych lat prognozy są obliczane według wzoru:

$$pW_t \pm U_\alpha \times SD_1 \times \sqrt{t} \tag{2}$$

gdzie:
pW_t – wartość współczynnika w głównym wariancie prognozy.

Jednakże w przypadku prognozy strumieni migracji (w szczególności emigracji) szerokość przedziału predykcji odbiega od tego wzoru. Wynika to z faktu, że ich wartość nie może być mniejsza od zera, a takie wyniki daje ten wzór w dalszych latach prognozy. Problem ten zostanie omówiony bardziej szczegółowo w dalszej części artykułu.

Wartości standardowe błędu standardowego prognozy na jeden rok do przodu (SD_1) przyjęte w prognozie dla poszczególnych komponentów były następujące:

- strumień imigracji 1792,
- strumień emigracji 5635,
- oczekiwane dalsze trwanie życia w momencie urodzenia dla mężczyzn 0,219,
- oczekiwane dalsze trwanie życia w momencie urodzenia dla kobiet 0,131,
- współczynnik dzietności 0,0554.

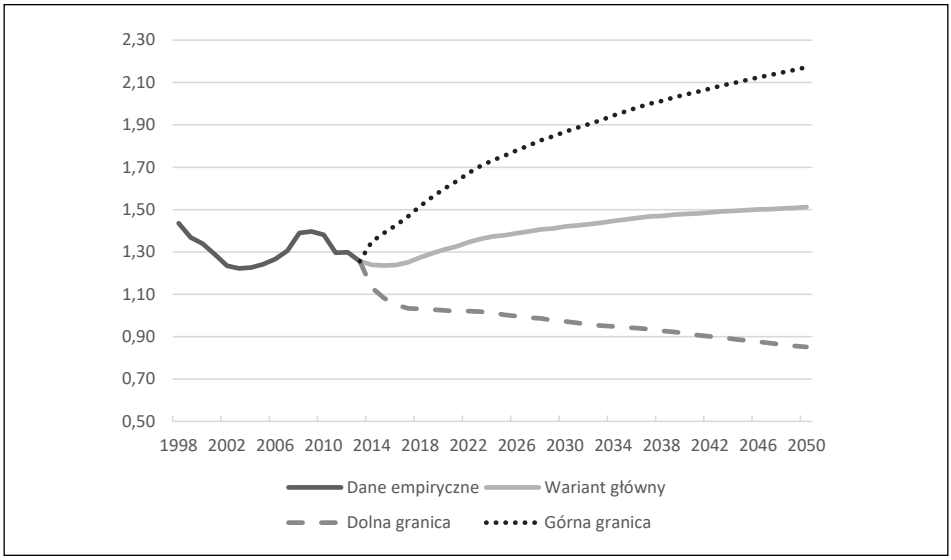
2.2. Wyniki

W przypadku współczynnika dzietności główny wariant oficjalnej prognozy ludności przygotowanej przez GUS zakładał niewielki jego spadek w najbliższej przyszłości, po czym systematyczny wzrost jego wartości do

poziomu około 1,56 w 2050 r. Prognoza probabilistyczna określa dodatkowo, że na 95% współczynnik na końcu horyzontu prognozy zmieści się w przedziale 0,85–2,17.

W 2030 r. 95% przedział predykcji obejmuje wartości od 0,97 do 1,86 (rys. 1). Należy odnotować, że 95% to bardzo wysoki poziom pewności, stąd też szerokość przedziałów jest duża. Metodologia prognoz probabilistycznych pozwala oczywiście na wyliczenie szerokości przedziałów dla dowolnego poziomu predykcji.

Rysunek 1. Prognoza z 95% przedziałem predykcji – współczynnik dzietności



Źródło: Opracowanie własne.

Prognozy probabilistyczne dają również możliwość oszacowania prawdopodobieństwa, że wartość danego współczynnika znajdzie się w określonym przedziale. Jest to niewątpliwie ogromna wartość dodana w porównaniu do klasycznych prognoz deterministycznych. Przykładowo w 2050 r. prawdopodobieństwo, że współczynnik dzietności będzie wyższy od poziomu prostej zastępowalności pokoleń (2,1) szacowane jest na 0,064, natomiast prawdopodobieństwo, że będzie niższy od 1 – na 0,058. Prawdopodobieństwo, że wartość współczynnika dzietności na końcu horyzontu prognozy będzie wyższa niż była w 2013 r. (1,255) szacowane jest z kolei na 0,792.

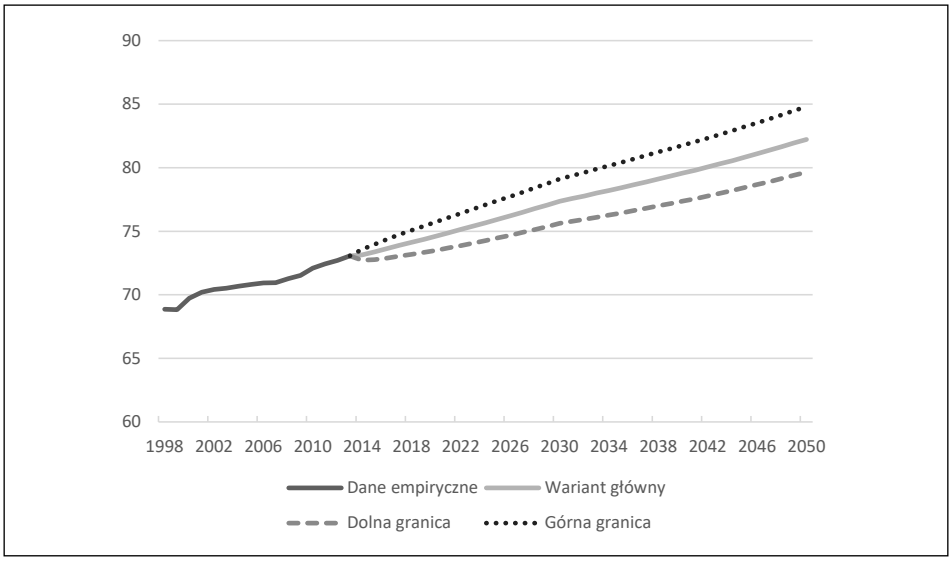
Oczekiwana długość życia w Polsce od wielu lat systematycznie wzrasta. Jest to również przyrost niemal idealnie liniowy. W związku z tym należy spodziewać się znacznie mniejszych błędów przy prognozowaniu wartości tego współczynnika niż w przypadku współczynnika dzietności.

Wykorzystanie metod opartych na analizie serii czasowych zakłada implicite, że współczynnik, który charakteryzował się w przeszłości dużą przewidywalnością, będzie się nią charakteryzował również w przyszłości. W przypadku oczekiwanej długości życia oznacza to przyjęcie założenia, że dalej będzie ona wrastała w tempie liniowym i nie nastąpi znaczne wyhamowanie bądź przyspieszenie tego procesu. Wydaje się, że na chwilę obecną można przyjąć takie założenie i próby uwzględnienia możliwości zmiany trendu prowadziłyby do niepotrzebnego poszerzenia przedziałów predykcji i utraty wartości informacyjnej prognozy.

Przewidywane oczekiwane dalsze trwanie życia w momencie urodzenia w roku 2050 wynosi 82,2 dla mężczyzn (rys. 2) i na 95% zmieści się w przedziale (79,6–84,8) oraz 87,7 dla kobiet (95% przedział predykcji od 85,1 do 90,3) (rys. 3).

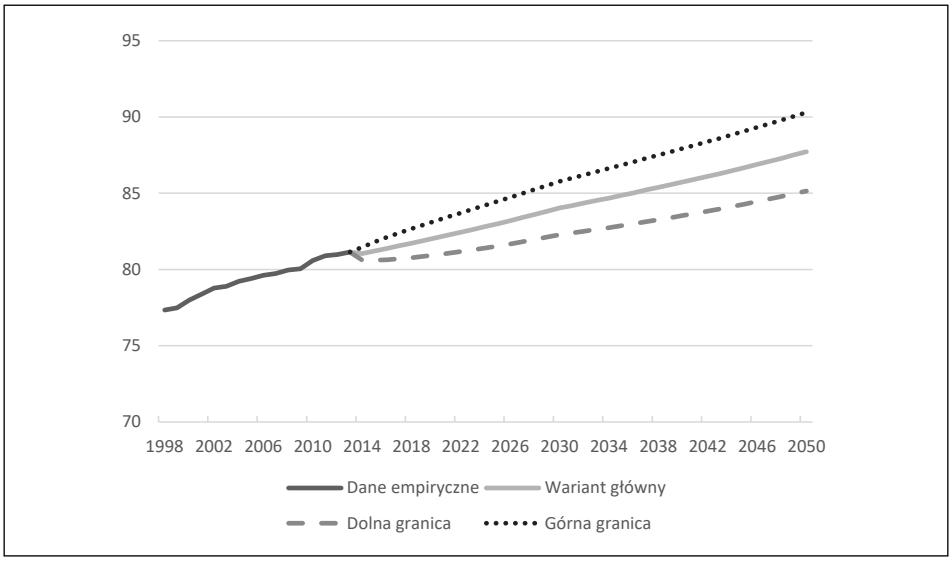
Migracje są komponentem, którego prognozowanie wiąże się z największymi trudnościami. Największym problemem, podobnie jak w przypadku prognoz deterministycznych, jest jakość danych dotyczących ruchu wędrówkowego. *Prognoza dla Polski na lata 2014–2050* brała pod uwagę tylko oficjalnie zarejestrowaną migrację na pobyt stały. W rezultacie dane dla migracji są dalece zaniżone i w praktyce marginalizują znaczenie tego komponentu dla wyników prognozy (rys. 4).

Rysunek 2. Prognoza z 95% przedziałem predykcji – e0 (mężczyźni)



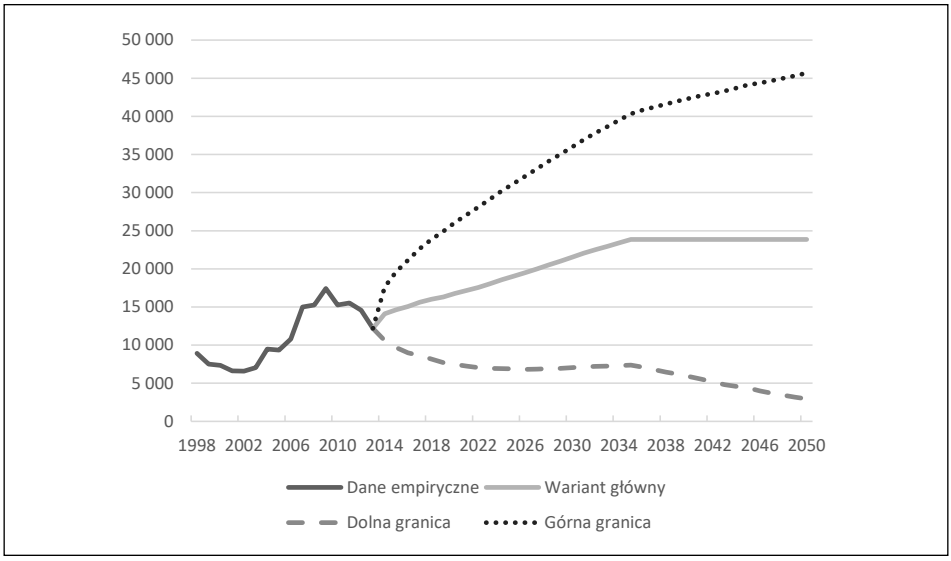
Źródło: Opracowanie własne.

Rysunek 3. Prognoza z 95% przedziałem predykcji – e0 (kobiety)



Źródło: Opracowanie własne.

Rysunek 4. Prognoza z 95% przedziałem predykcji – imigracja



Źródło: Opracowanie własne.

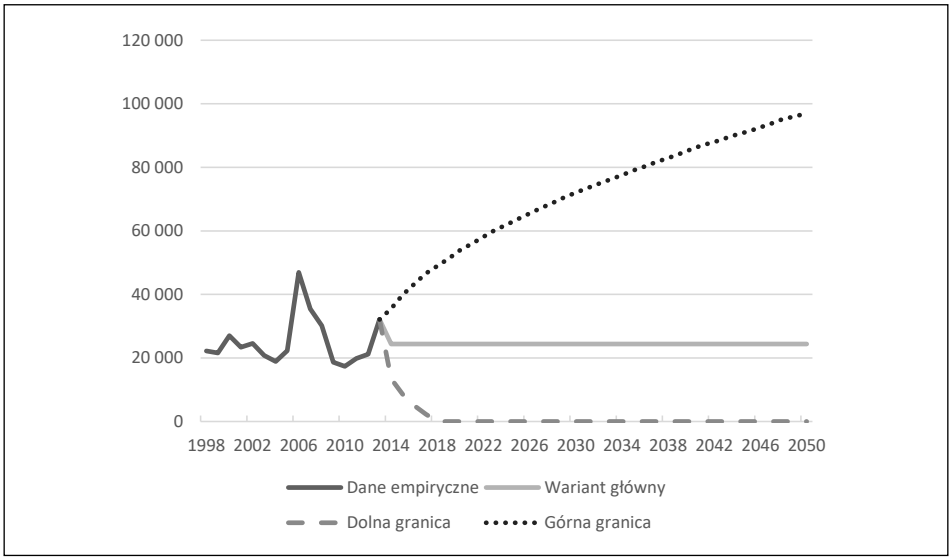
W przypadku prognoz probabilistycznych, w szczególności opartych na analizie szeregów czasowych, pojawia się dodatkowy problem. Strumienie migracji są niskie, jednak cechują się (zwłaszcza strumień emigracji) dużymi wahaniami. W rezultacie wykorzystanie klasycznych, opisanych

powyżej metod tworzenia ścieżek zmian prowadzi w wielu przypadkach do otrzymywania wartości ujemnych w dalszych latach prognozy.

W przypadku tej prognozy wszystkie wartości ujemne dla migracji zostały zamienione na zera. W rezultacie główny wariant prognozy przestał być średnią dla prognozy, choć pozostał jej medianą. Dodatkowo w przypadku prognozy emigracji dolna granica 95% przedziału predykcji już od 2018 r. wynosi zero.

Innymi możliwymi rozwiązaniami tego problemu byłoby wykorzystanie salda migracji zamiast strumieni imigracji i emigracji, użycie asymetrycznego rozkładu wartości losowej (ε_t we wzorze 1) bądź wykorzystanie innych dostępnych szacunków migracji.

Rysunek 5. Prognoza z 95% przedziałem predykcji – emigracja



Źródło: Opracowanie własne.

W wielu krajach do prognoz wykorzystuje się salda migracji a nie oddzielne strumienie. Są to jednak w większości kraje, w których imigracja zdecydowanie góruje nad emigracją i można przyjąć założenie, że imigracja będzie się utrzymywać na stałym poziomie, bądź będzie opóźnioną funkcją imigracji (jeżeli większość emigracji to migracje powrotne). W przypadku Polski wydaje się, że żadne z tych założeń nie byłoby zasadne i problemem byłoby rozdzielenie otrzymanej wartości salda pomiędzy imigrację i emigrację. Nierozdzielanie jej prowadziłoby z kolei do utraty wartości informacyjnej prognozy.

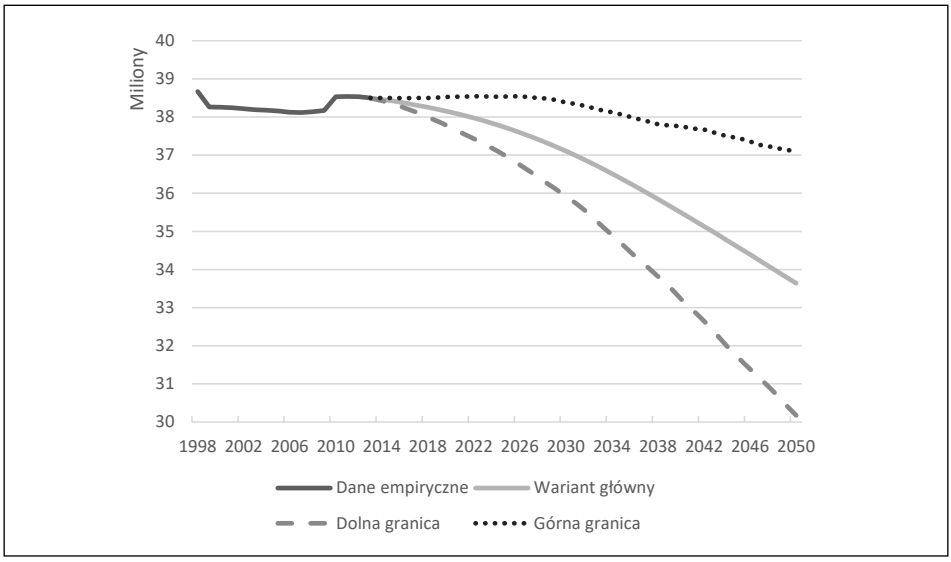
Wykorzystanie asymetrycznego rozkładu wartości losowej powoduje z kolei, że główny wariant przestaje być uznawany za najbardziej

prawdopodobny, co dodatkowo może powodować trudności w interpretacji wyników prognozy [De Beer, Alders, 1999].

Od 2008 r. dostępne są szacunki migracji oparte na tzw. statystykach lustrzanych, tj. opartych na porównaniu danych dotyczących migracji z różnych krajów. Szacowana wartość imigracji i emigracji dla Polski jest wielokrotnie wyższa od danych opartych na meldunku. Nie ulega jednak wątpliwości, że szereg czasowy oparty na tych danych jest za krótki do wykonania analiz potrzebnych do prognozy stochastycznej. Jednak być może warto podjąć próbę oszacowania wcześniejszych wartości, porównując te dane z oficjalnymi statystykami migracji.

Wyniki prognozy probabilistycznej ludności wskazują, że spadek ludności w najbliższych latach jest w zasadzie pewny. Prognoza szacuje prawdopodobieństwo, że ludność w 2030 r. będzie wyższa niż na koniec 2013 r. na 0,018, a w 2050 r. na zaledwie 0,006. Według przewidywań w 2050 r. liczba ludności na 95% będzie wahać się w granicach między 30,2 a 37,1 mln. Na 80% znajdzie się ona w przedziale pomiędzy 31,5 a 35,8 mln (rys. 6).

Rysunek 6. Prognoza ludności (95% przedział predykcji)



Źródło: Opracowanie własne.

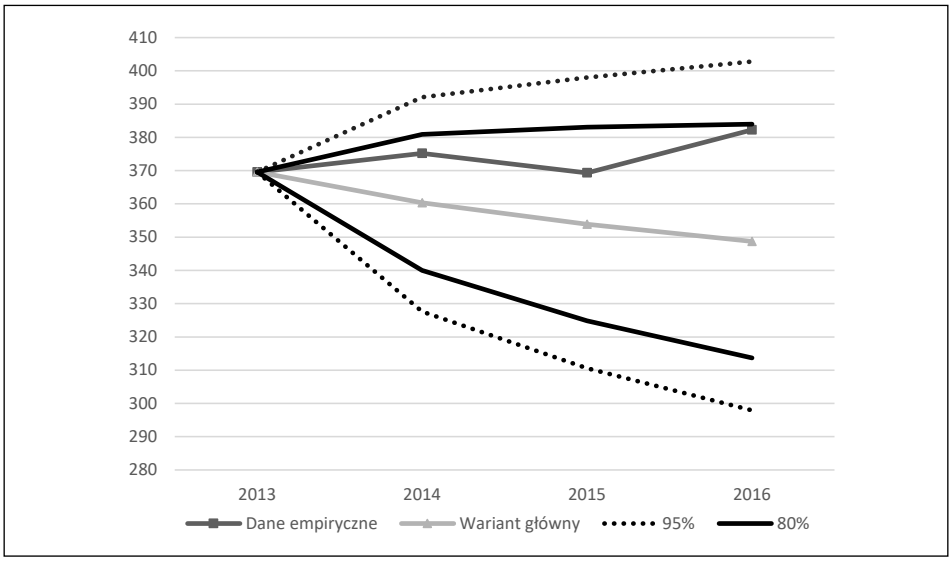
3. Porównanie wyników prognozy z danymi empirycznymi

Prognozy probabilistyczne pozwalają oszacować prawdopodobieństwo rozmaitych zdarzeń demograficznych w przyszłości. Pojawia się jednak pytanie, na ile te szacunki są trafne i w jakim stopniu są one lepsze od klasycznych prognoz wykonanych metodą deterministyczną. W jakimś stopniu odpowiedzi na to pytanie udzielić może porównanie wyników

oficjalnej prognozy GUS z 2014 r., prognozy probabilistycznej opartej na jej założeniach i danych empirycznych dla trzech kolejnych lat. Trzy lata to oczywiście bardzo krótki okres, jednak wystarczający, by korzyści z wykorzystania podejścia probabilistycznego były wyraźnie widoczne. Porównywanie wyników prognoz probabilistycznych z danymi empirycznymi może być również pomocne przy szacowaniu szerokości przedziałów predykcji w przyszłych prognozach.

W przypadku prognozy współczynnika dzietności oraz, co za tym idzie, prognozy liczby urodzeń oficjalny wariant prognozy GUS był istotnie niższy od obserwowanych wartości. W szczególności w 2016 r. było o 33,5 tys. urodzeń więcej niż zakładano (współczynnik dzietności wyniósł w tym roku 1,36 wobec zakładanego 1,24). Jeżeli weźmiemy jednak pod uwagę wyniki prognozy probabilistycznej, widać wyraźnie, że wszystkie te wyniki mieszczą się nie tylko w 95%, ale również w 80% przedziale predykcji (rys. 7). Można zatem powiedzieć, że wyniki w ostatnim okresie mieściły się w zakresie zmienności, którego należało się spodziewać w 2013 r. I właśnie uwzględnianie tej możliwej do przewidzenia zmienności wydaje się największą zaletą podejścia probabilistycznego w porównaniu do klasycznych prognoz deterministycznych.

Rysunek 7. Urodzenia (w tys.) 2013–2016



Źródło: Opracowanie własne.

W przypadku umieralności w 2014 r. zaobserwowano gwałtowny wzrost oczekiwanej długości życia zarówno dla mężczyzn (o około 0,7), jak i kobiet (o około 0,5). Tak znacznego wzrostu wartości e0 w ciągu jednego roku nie

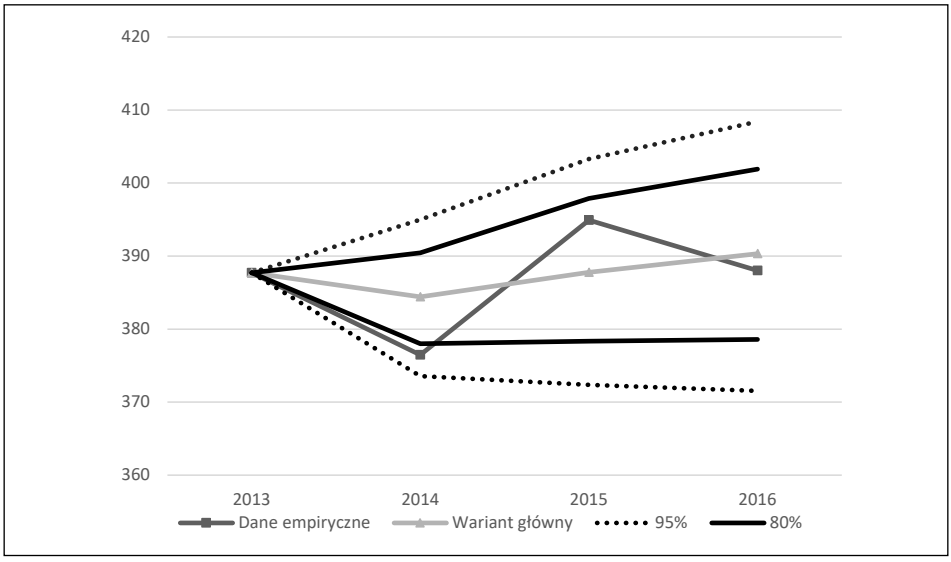
przewidział oficjalny wariant prognozy. Wartości e0 w 2014 r. nie zmieściły się nawet w 95% przedziale predykcji prognozy probabilistycznej. Należy jednak zauważyć, że rok później miał z kolei miejsce bezprecedensowy w ostatnich latach spadek oczekiwanej długości życia.

Nasuwa się w związku z tym pytanie, czy przedział predykcji dla prognozy był za wąski. Wydaje się, że jednak nie. Zmiana zaobserwowana w roku 2014 była unikatowa i nie miała precedensu w wartościach obserwowanych we wcześniejszych latach. W związku z tym można uznać ją za wydarzenie, którego prawdopodobieństwo było bardzo niskie (poniżej 0,025) i skonstatować, że nie ma konieczności poszerzenia przedziału predykcji. Trzeba zaznaczyć, że poszerzanie przedziałów predykcji do takiej szerokości, by były w stanie uwzględnić nawet najbardziej ekstremalny „szok", prowadzi nieuchronnie do utraty wartości informacyjnej prognozy.

W przypadku tego rodzaju danych nasuwa się również pytanie, czy rzeczywiście mieliśmy do czynienia z tak istotnym zjawiskiem demograficznych, czy jest to tylko „artefakt statystyczny" wynikający z innych czynników, który tym bardziej nie powinien być uwzględniany przy analizie szerokości przedziałów predykcji.

Ostatecznie prognozowana liczba zgonów na rok 2014 zmieściła się w 95% przedziale predykcji, ale w 80% już nie. Wartości w następnych latach zmieściły się w obu przedziałach (rys. 8).

Rysunek 8. Zgony (w tys.)

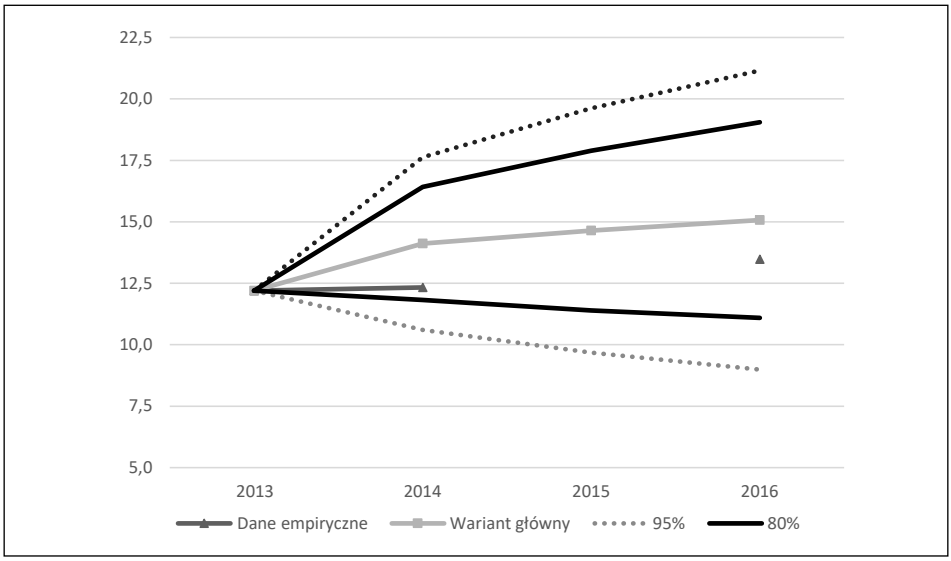


Źródło: Opracowanie własne.

W przypadku imigracji obie znane wartości empiryczne (dane za 2015 rok nie są dostępne) zmieściły się w 80% przedziale predykcji. Podobnie było również w przypadku emigracji, choć należy tu zauważyć, że w jej przypadku przedziały predykcji były znacznie szersze. 95% przedział predykcji w przypadku imigracji obejmował wartości od 9 do 21,5 tys., w przypadku emigracji od 5 do 43,2 tys. (rys. 9 i 10). Wyniki wydają się zatem potwierdzać, że wielkość przedziałów predykcji została w obu przypadkach dobrze dobrana oraz że wartości strumienia imigracji (przynajmniej oficjalnie rejestrowanej) charakteryzują się większą stabilnością niż w przypadku emigracji.

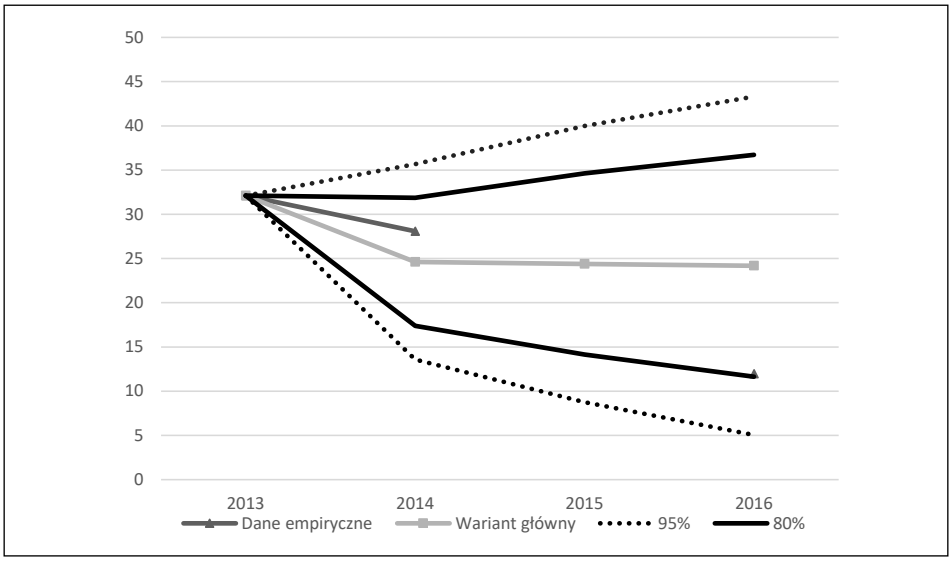
Należy zatem zauważyć, że szczególnie w przypadku prognozy urodzeń oraz emigracji prognoza probabilistyczna pozwoliła w znacznie większym stopniu przewidzieć ich zmienność i zawarła w 80% przedziałach wartości rzeczywiste, które w znacznym stopniu rozminęły się z głównym wariantem ostatniej *Prognozy ludności Polski na lata 2014–2050* wykonanej przez Główny Urząd Statystyczny.

Rysunek 9. Imigracja (w tys.)



Źródło: Opracowanie własne.

Rysunek 10. Emigracja (w tys.)



Źródło: Opracowanie własne.

Zakończenie

W niniejszym artykule starałem się przede wszystkim pokazać wartość dodaną płynącą z podejścia probabilistycznego, wskazując również na pewne ograniczenia wynikające z dostępnych w Polsce danych, w szczególności dotyczących migracji. Warto jednak odnotować, że z podejściem probabilistycznym wiążą się również problemy natury bardziej ogólnej. Kluczową kwestią jest szerokość przedziałów predykcji. Jest ona w ogromnym stopniu arbitralna i prognozy probabilistyczne wykonywane dla tego samego kraju w tym samym momencie potrafią się znaczne różnić zakładaną dokładnością przewidywań.

Jak wskazują zaprezentowane przykłady, licznych problemów dostarcza również analiza *ex post* szerokości przedziałów predykcji. Optymalna sytuacja miałaby miejsce wtedy, gdyby w 80% przedziale predykcji znajdowałoby się około 80% wyników. Analogicznie, jeżeli prawie wszystkie wyniki znajdą się na przykład w 30% przedziale predykcji, można przyjąć, że był on za szeroki. Można jednak argumentować, że przedziały predykcji powinny być szersze, by uwzględniać możliwe bardziej ekstremalne zjawiska demograficzne, bądź węższe, zapewniając większą wartość informacyjną.

Należy zauważyć również, że znacznie trudniejsza niż w przypadku prognoz deterministycznych jest kwestia błędu prognozy. Technicznie rzecz biorąc, prognozy te nigdy się nie mylą (nie wylicza się 100% przedziału predykcji). Nawet jeżeli prawdopodobieństwo zdarzenia, które nastąpiło, szacowano na 0,01, nie dowodzi to, że było to złe oszacowanie.

Niezależnie od tych problemów nie ulega wątpliwości, że prognozy probabilistyczne w znacznie większym stopniu pozwalają przewidzieć niepewność odnośnie współczynników demograficznych i dają użytkownikom pełniejszy obraz odnośnie do ich zmienności niż tradycyjne prognozy deterministyczne. Jak pokazała np. powyższa prognoza urodzeń, skala zmienności współczynnika dzietności była dobrze znana w momencie wyjściowym prognozy i jej uwzględnienie pozwoliłoby uniknąć rozbieżności między danymi empirycznymi a wynikami prognozy. Podejście probabilistyczne daje nam również bezcenną możliwość szacowania prawdopodobieństwa wystąpienia określonych zjawisk demograficznych w przyszłości. Nie ulega więc wątpliwości, że warto prowadzić dalsze analizy nad wykorzystaniem tego rodzaju prognoz.

Literatura

Alho J. (2002), *The Population of Finland in 2050 and Beyond*, The Research institute of Finnish Economy, Helsinki.

De Beer J. (2000), *Dealing with Uncertainty in Population Forecasting*, Statistics Netherlands.

De Beer J., Alders M. (1999), *Probabilistic Population and Household Forecast for the Netherlands.*

GUS (2014), *Prognoza ludności na lata 2014–2050*.

Keilman N. (1997), *Ex Post Errors in Official Population Forecast in Industrialized Countries*, „Journal of National Statistics", No. 3.

Keilman N., Pham D.G., Hetland A. (2002), *Why Population Forecasts Should Be Probabilistic? – Illustrated Case of Norway*, „Demographic Research", Vol. 6, No. 15.

Lee R., Tuljapurkar S. (1994), *Stochastic Population Forecasts for the United States: Beyond High, Medium and Low*, „Journal of American Statistical Association", Vol. 89, No. 428.

Lutz, W., Sanderson W., Scherbov S. (1998), *Expert-Based Probabilistic Population Projections*, „Population and Development Review", Vol. 24.

Matysiak A., Nowok B. (2007), *Stochastic Forecast of the Population of Poland 2005–2050*, „Demographic Research", Vol. 17, No. 11.

ONZ, Department of Economic and Social Affairs, Population Division (2013), *World Population Prospects: The 2012 Revision, Highlights and Advance Tables*.

Rowam S., Wright E. (2010), *Developing Stochastic Population Forecasts for the United Kingdom: Progress Report and Plans for Future Work*, Office for National Statistics.

Scherbov S., Mamolo M., Lutz W. (2005), *Probabilistic Population Projections for the 27 EU member States Based on Eurostat Assumptions.*

Streszczenie

Podejście probabilistyczne do budowy prognoz demograficznych powstało w odpowiedzi na zapotrzebowanie na bardziej precyzyjną informację dotyczącą niepewności przewidywań dotyczących procesów demograficznych. Projekcje

probabilistyczne budzą jednak szereg kontrowersji i obecnie większość oficjalnych prognoz ludności ma charakter deterministyczny.

W dyskusji o korzyściach i ograniczeniach projekcji probabilistycznych do przewidywania stanu i struktury ludności Polski przedstawione zostaną wyniki prac prowadzonych w GUS. Wstępna prognoza probabilistyczna, oparta na *Prognozie ludności na lata 2014–2050*, którą przygotowałem w 2015 r., daje możliwość porównania przedziałów predykcji dla poszczególnych współczynników demograficznych oraz liczby ludności z danymi empirycznymi z ostatnich trzech lat. Porównanie to jednoznacznie wskazuje, że podejście probabilistyczne pozwoliło zawrzeć zmienność procesów demograficznych w ramach przedziałów predykcji, podczas gdy wyniki prognozy deterministycznej, zaledwie po trzech latach od powstania, w znacznym stopniu odbiegają od stanu rzeczywistego.

Na przykładzie tej prognozy można wskazać również poważne ograniczenia, jakie napotyka tego typu podejście przy zastosowaniu go do danych dla Polski. Największym z nich są bez wątpienia oficjalne dane (oparte na meldunku) dotyczące migracji, które są dalece zaniżone i marginalizują znaczenie tego komponentu w prognozowaniu ludności. W przypadku prognoz probabilistycznych niskie strumienie migracji, połączone ze znacznymi ich wahaniami, powodują dodatkowe komplikacje – przy klasycznym podejściu przedziały predykcji obejmowałyby wartości ujemne. Bardziej wiarygodne szacunki migracji oparte o tzw. statystyki lustrzane dostępne są dopiero od 2008 r.

Podejście probabilistyczne bez wątpienia może posłużyć również do przygotowania znacznie bardziej złożonych prognoz na szczeblu niższym niż ogólnokrajowy. Daje na przykład możliwość oszacowania prawdopodobieństwa skrajnej depopulacji w poszczególnych regionach Polski.

Słowa kluczowe
prognozy demograficzne, ludność Polski, modele stochastyczne

Benefits and limitations of using probabilistic approach to forecast the population of Poland (Summary)

A probabilistic approach in demographic forecasting was developed in response to the need for more precise information about the uncertainty of demographic projections. Probabilistic projections methodology is however source of controversies, and most of the official population projections are still deterministic.

To discuss the benefits and limitations of probabilistic projections when applied to predicting number and structure of population of Poland, the results of the work carried out in the Central Statistical Office will be presented. The basic probabilistic projection, based on the CSO projection for 2014 -2050 which I prepared in 2015, gives the opportunity to compare confidence intervals for individual coefficients and population numbers with empirical data from the last three years. This comparison clearly indicates that the probabilistic approach has managed to include the variability of demographic processes within confidence intervals, whereas the results of the deterministic forecast, just three years after its creation, deviate significantly from the real data.

The example of this forecast also demonstrates serious constraints that this type of approach faces when applied to data for Poland. The largest of these is undoubtedly the official data for migration, which is far below understated and marginalizes the importance of this component in population projections. In the case of probabilistic forecasts, low migration flows, coupled with significant fluctuations in them, cause additional complications – with classical approaches to the confidence intervals computation giving negative values. More reliable estimates of migration based on so-called. „mirror statistics" are only available since 2008.

The probabilistic approach can undoubtedly also be used to prepare far more complex forecasts at lower levels than nationwide. It gives, for example, an opportunity to estimate the likelihood of extreme depopulation in particular regions of Poland.

Keywords
demographic forecasts, Polish population, stochastic models