Roczniki Kolegium Analiz Ekonomicznych
Zeszyt 31/2013

Ewa Dylewska
María Purificación Galindo Villardón

Wnioski z analizy głównych składowych dla modelu Lee-Cartera – populacja Polski i Hiszpanii

Streszczenie

Model Lee-Cartera jest obecnie szeroko stosowany w prognozowaniu śmiertelności oraz tworzeniu dynamicznych tablic trwania życia. Podstawową postać modelu Lee-Cartera można rozumieć jako model analizy głównych składowych (ang. Principal Component Analysis) ograniczony do pierwszej składowej. Stosowanie modelu uproszczonego wymaga jednak sprawdzenia, czy model w wystarczającym stopniu opisuje zmienność danych dla danej populacji, a także jaką część zmienności wyjaśniłyby kolejne składowe modelu. Artykuł przedstawia wnioski dotyczące zastosowania metody głównych składowych do analizy centralnego natężenia umieralności $m(x,t)$ populacji kobiet i mężczyzn w Polsce, a także w celach porównawczych – w Hiszpanii.

1. Wstęp

W procesie wyceny długoterminowych produktów ubezpieczeniowych, w tym ubezpieczeń emerytalno-rentowych, a także w testach zyskowności, istotne jest uwzględnienie zmiany śmiertelności w trakcie trwania umowy ubezpieczenia. Jednym z modeli, które pozwalają na projekcję przyszłej śmiertelności, jest model Lee-Cartera (1992), rekomendowany przez Continuous Mortality Investigation Bureau (2007) do prognozowania śmiertelności. Model ten, bazując na informacji o wartości współczynnika natężenia zgonów $m(x,t)$ osób w wieku x w roku t, pozwala na konstrukcję modelu wykładniczego postaci $m(x,t) = e^{a(x)+b(x)k(t)+\varepsilon(x,t)}$ oraz projekcję wartości $m(x,t)$ dla kolejnych lat t.

Model Lee-Cartera w swojej podstawowej postaci można również rozumieć jako model głównych składowych (*Principal Component Analysis* – PCA), ograniczony do jednej – pierwszej – składowej: $\ln[m\,(x,t)] = a\,(x)+b_1\,(x)\,k_1\,(t)+(x,t)$.

Dla różnych populacji konieczne może się okazać stosowanie alternatywnych postaci modelu. Zastosowanie metody analizy głównych składowych pozwala sprawdzić liczbę składowych modelu potrzebnych do właściwego opisania zmienności śmiertelności w danej populacji.

Szczegółową metodologię modelu Lee-Cartera oraz wyniki prognozy dla populacji Polski przedstawiono m.in. w pracach Bijaka, Więckowskiej (2008) oraz Rossy (2009), a dla Hiszpanii m.in. w badaniu Debón Aucejo, Montes Suay, Sala Garrido (2009). Celem tego badania jest jedynie weryfikacja odpowiedniej liczby

składowych modelu za pomocą analizy głównych składowych. Jednocześnie, jako punkt odniesienia, prezentowane są analogiczne wyniki dla populacji Hiszpanii.

2. Metoda analizy głównych składowych

Metoda analizy głównych składowych (PCA) jest techniką statystyki wielowymiarowej, która umożliwia redukcję wymiaru zbioru danych przy jednoczesnym zachowaniu jak największej części wariancji systemu (Jolliffe, 2002). Redukcję liczby wymiarów uzyskuje się poprzez transformację zmiennych do nieskorelowanych ze sobą głównych składowych, które są uporządkowane malejąco wielkością objaśnianej wariancji systemu:

$$\begin{gathered}\operatorname{Cov}(Y_I, Y_J) = 0,\\ \operatorname{Var}(Y_1) \geqslant \operatorname{Var}(Y_2) \geqslant \ldots \geqslant \operatorname{Var}(Y_1).\end{gathered}$$

Składowe modelu są kombinacjami liniowymi zmiennych wejściowych, z których każda składowa określa część opisanej wariancji; tym samym możliwe jest określenie tego, jaka część informacji została utracona poprzez redukcję liczby wymiarów danych.

Zbiór danych złożony z n obserwacji p zmiennych można opisać za pomocą układu równań:

$$\begin{aligned}Y_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,\\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p,\\ &\vdots\\ Y_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p,\end{aligned}$$

gdzie X_i oznacza zmienną wejściową (oryginalną), a Y_i jest nową zmienną lub składową modelu, która jest związana ze zmiennymi wejściowymi poprzez relacje liniowe. Współczynniki modelu (a_{ij}) spełniają równanie: $\sum_{i=1}^{p} a_{ij}^2 = 1$.

W celu uzyskania głównych składowych należy znaleźć rozwiązanie następującego problemy optymalizacyjnego (Vicente Villardón, 2010):

$$(\operatorname{Var}(Y) \rightarrow \max) \text{ oraz } \mathbf{A}^T\mathbf{A} = \mathbf{I}.$$

Współczynniki dla i-składowej wyznacza się metodą mnożników Lagrange'a, rozwiązując następujące równania:

$$\begin{aligned}L &= \operatorname{Var}(Y_i) + \lambda_i(1 - a_i^T a_i),\\ \frac{\partial L}{\partial a_i} &= 2Sa_i - 2\lambda_i a_i = 0.\end{aligned}$$

Tworząc model matematyczny, należy uprzednio zweryfikować to, czy zastosowanie metody analizy głównych składowych jest w danym przypadku właściwe.

W tym celu można skorzystać z dwóch testów stosowanych częściej w analizie czynnikowej: testu sferyczności Bartletta oraz próby Kaisera-Mayera-Olkina (KMO).

Zastosowanie testu sferyczności Bartletta pozwala określić to, czy istnieje przynajmniej jedna istotna korelacja pomiędzy analizowanymi zmiennymi. Test opiera się na sprawdzeniu, czy macierz korelacji jest macierzą jednostkową, czyli $|R_p| = 1$. Weryfikuje się hipotezę:

$$H_0 : |R_p| = 1, \text{ przeciwko } H_1 : |R_p| \neq 1.$$

Sprawdzian hipotezy opiera się na wartościach macierzy współczynników korelacji reszt R dla p zmiennych i ma postać:

$$-\left[n - 1 - \tfrac{1}{6}(2p + 5)\right] \cdot \ln |R| \approx \chi_{0,5(p^2-p)}.$$

W przypadku przyjęcia hipotezy zerowej o braku korelacji pomiędzy zmiennymi należy zrezygnować z zastosowania techniki analizy głównych składowych. Brak korelacji pomiędzy zmiennymi oznacza, że nie istnieją takie ich kombinacje liniowe, które umożliwiałyby redukcję liczby wymiarów danych.

Próba Keisera-Mayera-Olkina (KMO) umożliwia określenie, w jakim stopniu próbka danych jest odpowiednia do analizy głównych składowych. Miarę KMO definiuje się jako:

$$\text{KMO} = \frac{\sum \sum_{h \neq j} r_{jh}^2}{\sum \sum_{h \neq j} r_{jh}^2 + \sum \sum_{h \neq j} a_{jh}^2},$$

gdzie r_{ij} to współczynniki korelacji pomiędzy zmiennymi wejściowymi (oryginalnymi) oraz a_{ij} to współczynniki korelacji cząstkowej pomiędzy zmiennymi. Według Keisera, wartość statystyki KMO bliska 0,9 oznacza bardzo dobrą adekwatność danych do zastosowania techniki, podczas gdy wartości KMO niższe od 0,5 oznaczają, że metoda głównych składowych nie powinna być stosowana (niejednoznaczność rozwiązania).

Analiza głównych składowych (PCA) bywa traktowana jako specjalny przypadek analizy czynnikowej (ang. *Factorial Analysis*); podejście to jest kontynuowane w wielu pakietach statystycznych (w tym również przez wykorzystywany w tej analizie SPSS). Obie metody jednak istotnie się różnią (Jolliffe, 2002). Analiza głównych składowych poszukuje takiego ortogonalnego przekształcenia p zmiennych wejściowych w p nieskorelowanych składowych, które opisze całkowitą zmienność systemu. PCA dostarcza także informacji o utracie informacji związanej z redukcją układu do $i < p$ wymiarów. Celem analizy głównych składowych jest zatem wyjaśnienie zmienności systemu, a nie budowa modelu. Składowe są uporządkowane wielkością wyjaśnionej wariancji i dodanie kolejnej składowej nie wpływa na dotychczasowe rozwiązanie. Analiza czynnikowa, podobnie jak

analiza głównych składowych, poszukuje takich kombinacji liniowych – czynników wspólnych – które pozwolą zredukować wymiar zbioru danych. W odróżnieniu jednak od PCA, celem analizy czynnikowej jest wyjaśnienie nie wariancji systemu, ale korelacji zmiennych. Analiza czynnikowa skupia się zatem na tych elementach macierzy kowariancji, które znajdują się poza przekątną macierzy. Głównym celem analizy czynnikowej jest budowa modelu. Ze względu na rotację osi dodanie kolejnego czynnika *i+1* wpływa na dotychczasowe rozwiązanie dla i czynników.

3. Model Lee-Cartera

Model Lee-Cartera (Girosi, King, 2007) w jego podstawowej postaci można rozumieć jako model analizy głównych składowych, ograniczony do pierwszej składowej $b_1(x)k_1(t)$, która wystarczy, aby opisać wystarczającą część zmienności systemu.

Rozważamy macierz $P \times T$ elementów, składającą się z logarytmów naturalnych centralnego natężenia umieralności $\ln[m(x,t)]$ dla wieku $x \in P$ oraz w momencie $t \in T$. Zakładamy, że przestrzeń P-wymiarowa może zostać zredukowana do zaledwie jednego wymiaru bez znacznej utraty informacji. Model składający się z P składowych miałby postać:

$$\ln[m\,(x,t)] = a(x) + b_1(x)k_1(t) + b_2(x)k_2(t) + \cdots + b_P(x)k_P(t).$$

Jeśli model zredukujemy do jednej składowej, otrzymamy model zaprezentowany przez Lee i Cartera. Składowa jest funkcją liniową czasu:

$$\underbrace{\ln[m\,(x,t)] = a(x) + b_1(x)k_1(t)}_{\text{model Lee-Cartera (bez części losowej)}} + \underbrace{b_2(x)k_2(t) + \cdots + b_P(x)k_P(t)}_{\varepsilon(x,t)}.$$

W modelu Lee-Cartera parametry $a(x)$ i $b(x)$ są zależne tylko od wieku x, natomiast $k(t)$ zależy wyłącznie od czasu t. Parametr $a(x)$ reprezentuje wektor średnich wartości $\ln[m(x,t)]$ obserwowanych w okresie T dla wieku x. Parametr $b(x)$ pozwala stwierdzić, dla jakich x wartości natężenia umieralności rosną lub maleją szybciej w kolejnych latach obserwacji t. Innymi słowy, $b(x)$ reprezentuje siłę, z jaką zmienia się $\ln[m(x,t)]$ w danym wieku x, gdy zmienia się parametr $k(t)$. Zasadniczo wartości parametru $b(x)$ są dodatnie. W krótkim okresie mogą być jednak również ujemne (Girosi, King, 2007), co oznacza wzrost umieralności w czasie. Błąd modelu $\varepsilon\,(x,t)$ ma rozkład normalny $N(0,\sigma^2)$ i reprezentuje zmiany śmiertelności, które nie są wyjaśnione przez model. Parametr $k(t)$ wyznacza się za pomocą odpowiedniego modelu *ARIMA*. Lee i Carter (Girosi, King, 2007), budując model dla populacji Stanów Zjednoczonych, sprawdzili, że dla

analizowanych danych wartości parametru $k(t)$ mogą być wyznaczone za pomocą modelu błądzenia losowego z dryfem. Jednocześnie dla innych populacji mogą być odpowiednie inne modele szeregów czasowych.

Jeśli model nie może być zredukowany do jednej składowej, szacuje się model dla większej liczby składowych. W procesie szacowania modelu można wyróżnić następujące etapy (Booth, Maindonald, Smith, 2002):

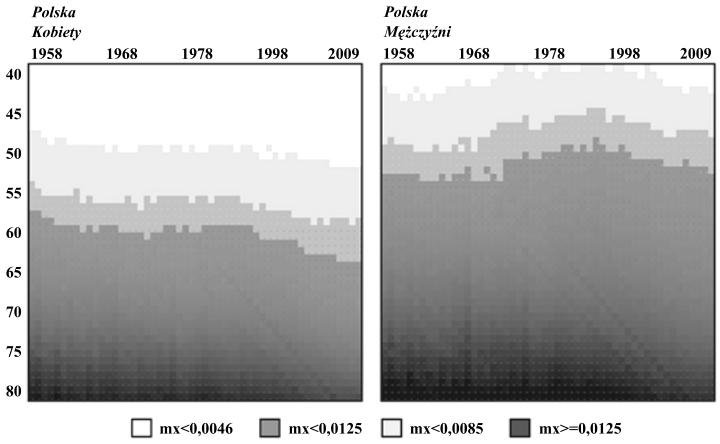
1. Oszacowanie średniej wartości logarytmów naturalnych centralnego natężenia umieralności $m(x,t)$: parametr $a(x)$.
2. Oszacowanie parametrów modelu $b_1(x)$, $k_1(t)$ dla pierwszej składowej poprzez zastosowanie metody dekompozycji macierzy danych na wektory i wartości osobliwe (SVD):
$$\ln[m\,(x,t)] = a(x) + b_1(x)k_1(t).$$
3. Obliczenie wartości reszt modelu:
$$-\ln\left[m\,(x,t)\right] - \ln\left[m\,(x,t)^*\right].$$
4. Oszacowanie parametrów modelu $b_2(x)$, $k_2(t)$ dla drugiej składowej na podstawie macierzy reszt oraz poprzez zastosowanie metody dekompozycji macierzy danych na wektory i wartości osobliwe (SVD).
5. Jeśli szacuje się więcej niż dwie składowe, proces należy powtórzyć w celu oszacowania kolejnych parametrów modelu: $b_3(x)$, $k_3(t)$, $b_4(x)$, $k_4(t)$ itd.

4. Natężenie umieralności w populacji Polski i Hiszpanii

Do analizy zmian natężenia umieralności $m(x,t)$ populacji Polski i Hiszpanii wybrano lata 1958–2009 oraz grupę wiekową $x \in \langle 40; 85\rangle$. Dane o centralnym natężeniu umieralności $m(x,t)$ populacji Polski i Hiszpanii pochodzą z „Human Mortality Database". Zawężenie grupy wiekowej jest konieczne ze względu na krótki okres obserwacji ($T = 52$ lata), dla którego są dostępne pełne dane. Wybrany przedział wieku odpowiada latom, w których spodziewana jest największa poprawa śmiertelności, a także reprezentuje potencjalnych klientów instytucji finansowych, które oferując produkty obarczone ryzykiem długowieczności, są zainteresowane dobrym dopasowaniem modelu prognostycznego szczególnie dla tego przedziału wiekowego.

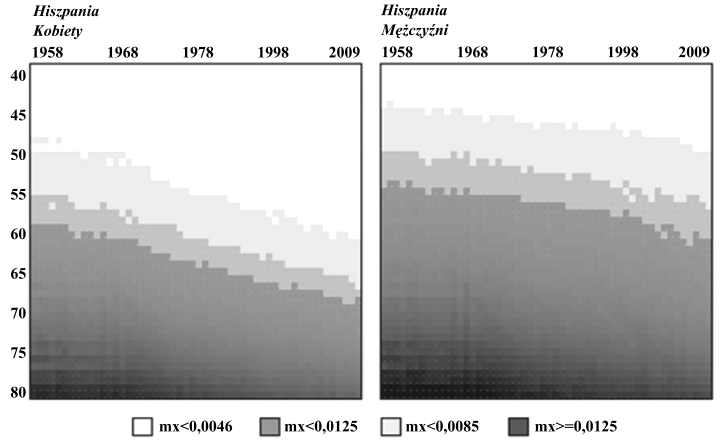
Obserwacja trendu zmian współczynnika natężania umieralności populacji Polski pozwala stwierdzić, że dla kobiet w grupie wiekowej 40–85 lat obserwuje się obniżanie natężenia umieralności. Tendencja ta była jednak spowolniona w latach 70., w latach 80. zaobserwowano nawet niewielki wzrost umieralności, począwszy jednak od lat 90., natężenia umieralności kobiet systematycznie się obniża.

Rysunek 1. Natężenie umieralności $m(x,t)$ w grupie wiekowej 40–85 w Polsce w latach 1958–2009



Źródło: opracowanie własne na podstawie danych HMD.

Rysunek 2. Natężenie umieralności $m(x,t)$ w grupie wiekowej 40–85 w Hiszpanii w latach 1958–2009



Źródło: opracowanie własne na podstawie danych HMD.

W przypadku populacji mężczyzn w Polsce w grupie wiekowej 40–85 w latach 50. i 60. obserwowano trend obniżenia śmiertelności. W latach 70. trend ten jednak zmienił kierunek i do początku lat 90. obserwuje się wzrost natężenia umieralności. Efekt ten związany był ze zwiększeniem liczby zgonów z powodu chorób układu krążenia, nowotworów złośliwych oraz urazów i zatruć (Szadkowska-Stańczyk i in., 1991). Dotyczył on całej populacji Polski w wieku aktywności zawodowej; najbardziej jednak był widoczny w populacji mężczyzn w Polsce

w wieku 40–59 lat żyjących na terenach wiejskich (Szadkowska-Stańczyk i in., 1991). Podobny wzrost umieralności był obserwowany również w innych krajach Europy Środkowo-Wschodniej, takich jak Słowacja i Węgry (Denkowska, Papież, 2009). Począwszy od lat 90., natężenie umieralności mężczyzn w Polsce obniża się.

Natężenie umieralności w grupie wiekowej 40–85 w populacji Hiszpanii systematycznie się obniża, w przypadku zarówno kobiet, jak i mężczyzn. Tempo zmiany natężenia było jednak w Hiszpanii większe w przypadku populacji kobiet. W przeciwieństwie do populacji Polski, dla populacji Hiszpanii nie obserwuje się zmian trendu ani nagłych zmian natężenia umieralności $m(x,t)$.

Zastosowanie metody głównych składowych

Obliczenia *Aplicación del modelo de Lee-Carter para la construcción de tablas de mortalidad dinámicas para Polonia y España* (Dylewska, Galindo Villardón, 2011) przeprowadzono za pomocą pakietu SPSS. Podsumowanie wyników dla populacji Polski (PL) i Hiszpanii (ES) przedstawia poniższa tabela.

Tabela 1. Wyniki analizy głównych składowych

		Populacja			
		PL M	**PL K**	**ES M**	**ES K**
Miara Kaisera-Meyera-Olkina (KMO)		0,75	0,86	0,87	0,91
Test Bartletta	przybliżone chi-kwadrat	5074,66	5547,46	6541,76	7792,62
	df	1035	1035	1035	1035
	p-value	0,00	0,00	0,00	0,00
Procent zmienności opisanej przez 1. składową		54,8	90,9	95,0	96,8
Procent zmienności opisanej przez 2. składową		34,9	1,8	1,3	1,4
Procent zmienności opisanej przez 3. składową		3,3	1,4	0,4	0,3

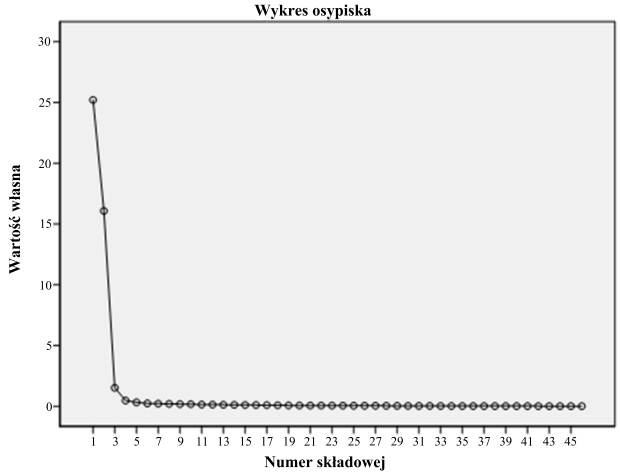
Źródło: obliczenia własne w pakiecie SPSS.

Wartości miary KMO obliczone dla populacji mężczyzn w Polsce są wysokie (0,75) lub bardzo wysokie (dla pozostałych populacji), przekraczając znacznie minimalną rekomendowaną wartość 0,5. Tym samym dane można uważać za odpowiednie do zastosowania metody głównych składowych.

Test sferyczności Bartletta ma za zadanie określić, czy istnieje przynajmniej jedna istotna relacja między zmiennymi. Testowana jest hipoteza zerowa o macierzy korelacji będącej macierzą jednostkową. Statystyka testowa obliczona dla $p = 46$ zmiennych jest mniejsza od wartości krytycznej rozkładu chi-kwadrat dla $0,5 \cdot (p^2 - p) = 1035$ stopni swobody, na poziomie istotności $\alpha \cong 0$. Hipotezę zerową o braku korelacji pomiędzy zmiennymi należy zatem odrzucić dla każdej z analizowanych populacji. Redukcja wymiarów macierzy danych jest możliwa i uzasadnione jest stosowanie metody głównych składowych.

Zmienność wyjaśniona przez pierwszą składową w przypadku populacji mężczyzn w Polsce wynosi 54,8%. Dla każdej z pozostałych populacji, zarówno kobiet w Polsce, jak i kobiet oraz mężczyzn w Hiszpanii, część zmienności systemu objaśniona przez pierwszą składową przekracza 90%. Oznacza to, że model Lee-Cartera z jedną składową jest odpowiedni do prognozowania śmiertelności w przypadku kobiet w Polsce oraz kobiet i mężczyzn w Hiszpanii. W przypadku populacji mężczyzn w Polsce do opisania przynajmniej 89,7% zmienności potrzebne są dwie składowe. Jeśli stosujemy kryterium Kaisera (wartość własna składowej większa od 1), w modelu potrzebne są aż trzy składowe. Wartości własne poszczególnych składowych modelu dla populacji mężczyzn w Polsce przedstawia poniższy wykres osypiska.

Rysunek 3. Wykres osypiska dla modelu opisującego śmiertelność populacji mężczyzn w Polsce

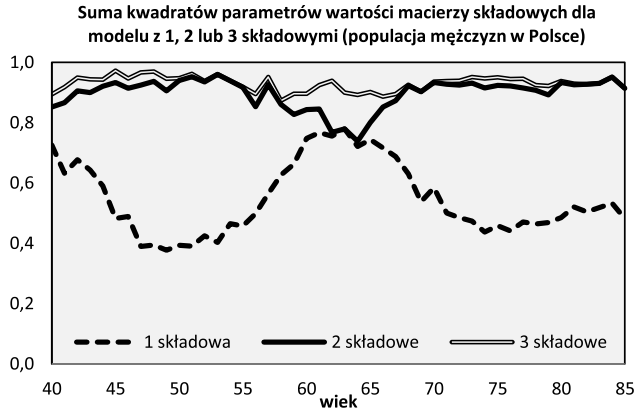


Źródło: opracowanie własne przy użyciu pakietu SPSS.

W celu oceny jakości reprezentacji zmienności danych przez jedną, dwie lub trzy składowe można wyznaczyć wskaźniki będące sumą kwadratów wartości macierzy składowych dla danej zmiennej (wiek x). Dla modelu z jedną składową wskaźniki te przyjmują wysokie wartości zarówno dla populacji mężczyzn w Hiszpanii ($\min = 0,786$), kobiet w Hiszpanii ($\min = 0,883$), jak i populacji kobiet w Polsce ($\min = 0,767$). W przypadku populacji mężczyzn w Polsce wartości tego wskaźnika dla pierwszej składowej są znacznie niższe i należą do przedziału $[0,377; 0,780]$, co w przypadku niższych wartości oznacza słabą reprezentację zmiennych przez model. Przekształcenie danych do dwóch składowych pozwala na znacznie lepsze opisanie zmienność natężenia umieralności, czego dowodem są wyższe wartości wskaźnika ($\min = 0,739$). Obserwacja tego wskaźnika dla modeli jednej, dwóch lub trzech składowych pozwala stwierdzić, że druga składowa po-

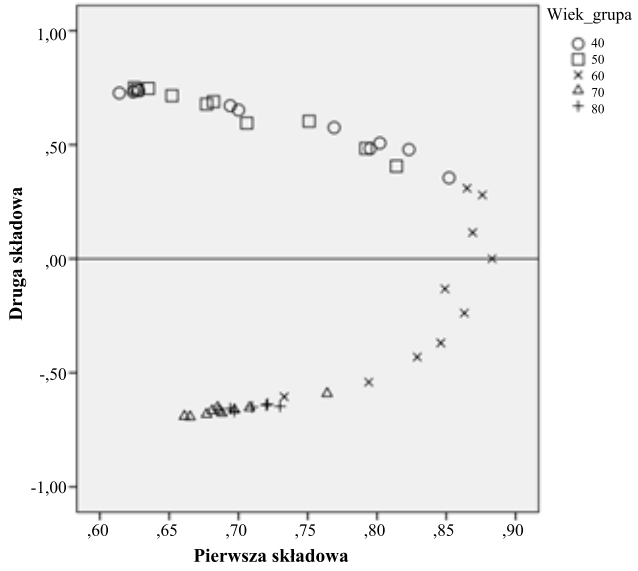
prawia stopień reprezentacji zmiennych szczególnie w grupach wiekowych: 40–59 lat oraz 71–85 lat. Wpływ trzeciej składowej jest natomiast najbardziej widoczny dla grupy wiekowej 60–70 lat.

Rysunek 4. Sumy kwadratów wartości macierzy składowych dla modelu z jedną, dwoma lub trzema składowymi oraz populacji mężczyzn w Polsce



Źródło: opracowanie własne na podstawie obliczeń w pakiecie SPSS.

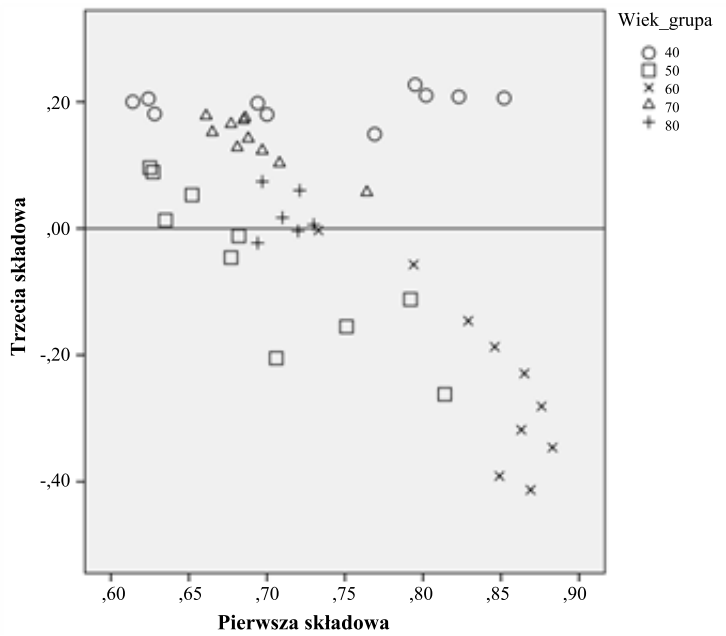
Rysunek 5. Wartości macierzy składowych pierwszej i drugiej składowej dla populacji mężczyzn w Polsce



Źródło: opracowanie własne na podstawie obliczeń w pakiecie SPSS.

Powyższe obserwacje potwierdza analiza wykresu rozrzutu wartości macierzy składowych pierwszej, drugiej i trzeciej dla populacji mężczyzn w Polsce. W celu łatwiejszej interpretacji dokonano zgrupowania zmiennej wiek (x). Pierwsza składowa najlepiej oddaje zmienność natężenia umieralności w grupie 60-latków, natomiast druga składowa jest potrzebna najbardziej w celu opisania zmienności natężenia umieralności w grupie 70-, 80- i 50-latków. Obserwacja wykresu rozrzutu wartości macierzy dla pierwszej i trzeciej składowej potwierdza udział trzeciej składowej w wyjaśnieniu natężenia umieralności głównie w grupie 60-latków i słabą relację drugiej składowej z pozostałymi grupami wiekowymi.

Rysunek 6. Wartości macierzy składowych pierwszej i trzeciej składowej dla populacji mężczyzn w Polsce



Źródło: opracowanie własne na podstawie obliczeń w pakiecie SPSS.

Badanie odpowiedniości modelu Lee-Cartera poprzez zastosowanie metody głównych składowych zrealizowali również Girosi i King (2007). Analizowane były populacje Stanów Zjednoczonych, Japonii oraz wybranych krajów europejskich. Autorzy badali dane dotyczące śmiertelności w zależności od wieku, czasu i przyczyny zgonu. Zaprezentowane wyniki oznaczają, że w przypadku populacji Stanów Zjednoczonych i śmiertelności ogólnej model składający się tylko z jednej składowej opisuje 93% zmienności danych. Analizę głównych składowych przeprowadza się również dla populacji Hiszpanii, otrzymując pierwszą składową, która wyjaśnia 89% zmienności systemu. Populacja Polski nie jest uwzględniona w badaniu Girosiego i Kinga.

5. Wnioski z badania

Zastosowanie metody analizy głównych składowych pozwoliło stwierdzić, że model Lee-Cartera w swojej podstawowej postaci oszacowany dla populacji mężczyzn w Polsce i określonych w badaniu danych wejściowych opisuje jedynie ok. 54,8% zmienności systemu. Prawdopodobną przyczyną słabego dopasowania modelu z jedną składową w przypadku populacji mężczyzn w Polsce w wieku 40–85 lat jest niemonotoniczny charakter zmiany natężenia umieralności w wieku x w okresie obserwacji. Stosowanie dla tej populacji modelu zawierającego tylko jedną składową może wpłynąć na adekwatność prognozy śmiertelności. Nawet jeśli różnica pomiędzy prognozowanymi wartościami współczynnika natężenia zgonów dla modelu podstawowego i rozszerzonego w krótkim okresie jest niewielka, efekt zakumulowany w dłuższej perspektywie może być bardziej znaczący. Wpływ drugiej składowej jest szczególnie istotny dla adekwatności oszacowań w grupach wiekowych 40–59 lat oraz 71–85 lat. Wpływ trzeciej składowej ogranicza się do oszacowań w grupie wiekowej 60–70 lat. Model złożony z dwóch lub z trzech składowych opisuje ok. 90% i 93% zmienności systemu (dla badanego okresu). W przypadku populacji kobiet w Polsce dla określonego w badaniu wieku i horyzontu szacowanego procent zmienności systemu objaśniony przez pierwszą składową wyniósł 90,9%, przez co stosowanie modelu Lee-Cartera w formie podstawowej jest uzasadnione. Procent zmienności systemu objaśnionej przez pierwszą składową dla populacji mężczyzn w Hiszpanii to 95,0%, a w przypadku kobiet 96,8%; model Lee-Cartera może być z powodzeniem stosowany w prognozowaniu śmiertelności.

Bibliografia

[1] Bijak J., Więckowska B. (2008), *Analiza ubezpieczeniowych implikacji wyników prognozy przeciętnego dalszego trwania życia uzyskanej metodą Lee i Cartera*, wydanie specjalne „Wiadomości Ubezpieczeniowych", PIU.

[2] Booth H., Maindonald J., Smith L. (2002), *Age-time interactions in mortality projection: Applying Lee-Carter to Australia*, Working Papers in Demography, no. 85.

[3] CMI Comitte (2007), *Stochastic projection methodologies: Lee-Carter model features, example results and implications*, Continuous Mortality Investigation Reports, no. 25.

[4] Debón Aucejo A., Montes Suay F., Sala Garrido R. (2009), *Tablas de mortalidad dinámicas para España. Una aplicación a la hipoteca inversa,* Universitat Valencia.

[5] Denkowska S., Papież M. (2009), *The Analysis of Mortality Changes in Poland and Selected European Countries in the Period 1960–2000*, XXVI IUSSP International Population Conference.

[6] Dylewska E., Galindo Villardón M.P. (2011), *Aplicación del modelo de Lee-Carter para la construcción de tablas de mortalidad dinámicas para Polonia y España*, Universidad de Salamanca
[7] Girosi F., King G. (2007), *Understanding the Lee-Carter Mortality Forecasting Method,* working paper, Harvard University.
[8] Jolliffe I.T. (2002), *Principal Component Analysis*, 2^{nd} ed., Springer, New York.
[9] Lee R.D., Carter L.R. (1992), *Modeling and Forecasting U.S. Mortality*, „Journal of the American Statistical Association", vol. 87 (419).
[10] Pearson K. (1901), *On Lines and Planes of Closest Fit to Systems of Points in Space*, „Philosophical Magazine", vol. 2 (6).
[11] Rossa A. (2009), *Dynamiczne tablice trwania życia oparte na metodologii Lee-Cartera i ich zastosowanie do obliczania wysokości świadczeń emerytalnych*, „Acta Universitatis Lodziensis", Folia Oeconomica, 231.
[12] Szadkowska-Stańczyk I., Hanke W., Gdulewicz T. (1991), *Analiza umieralności populacji w wieku produkcyjnym w Polsce*, cz. 2, *Udział głównych przyczyn zgonów w rosnącej umieralności mężczyzn i kobiet,* „Medycyna Pracy", nr 42 (1), s. 43–49.
[13] Vicente Villardón J.L. (2010), *Metódos clásicos*, Máster en Análisis Avanzado de Datos Multivariantes, Universidad de Salamanca.

***

Implications of Principal Component Analysis for Lee-Carter model – population of Poland and Spain

Abstract

Lee-Carter model is currently widely used in mortality forecasting and construction of dynamic life tables. Basic Lee-Carter model can be understood as Principal Component Analysis model with only one – first component. Using basic model formula requires verification what part of total variability is captured and also defining what part of variability would be explained by additional components. The article presents conclusions regarding application of Principal Component Analysis to data about central death rate $m(x,t)$ of population of males and females in Poland and additionally in Spain, for the purpose of comparison.

Autorzy:

Ewa Dylewska, Metlife Amplico, ul. Przemysłowa 26, 00-450 Warszawa,
e-mail: ewa.dylewska@metlifeamplico.pl

María Purificación Galindo Villardón, Departamento de Estadística, Facultad de Medicina, Universidad de Salamanca, c/ Alfonso X El Sabio s/n, Campus Miguel de Unamuno, 37007 Salamanca, Hiszpania,

e-mail: pgalindo@usal.es