Klasyfikacja lokalnych rynków pracy w Polsce na podstawie analizy głównych składowych dla danych podwójnie wielowymiarowych

Wojciech Łukaszonek[1,2,3]
Marcin Szymkowiak[2,3], Waldemar Wołyński[4]

[1]Uniwersytet Kaliski
[2]Uniwersytet Ekonomiczny w Poznaniu
[3]Urząd Statystyczny w Poznaniu
[4]Uniwersytet Adama Mickiewicza

1–3 lipca 2025 r.

1. Cele prezentacji
2. Opis metod
3. Źródło danych
4. Zmienne diagnostyczne
5. Wybrane wyniki
6. Podsumowanie
7. Literatura

Cele prezentacji

Główne cele prezentacji:

- ocena sytuacji na rynku pracy w województwie wielkopolskim na poziomie powiatów,
- omówienie metod statystycznych użytych w badaniu z uwzględnieniem analizy głównych składowych dla danych podwójnie wielowymiarowych,
- przedstawienie wybranych wyników, ich interpretacja oraz sformułowanie wniosków na podstawie przeprowadzonej analizy.

Cele badania

- Celem badania jest klasyfikacja powiatów województwa wielkopolskiego z wykorzystaniem wybranych modeli statystycznych analizy głównych składowych dla danych podwójnie wielowymiarowych.
- Zastosowanie zaawansowanych modeli statystycznych analizy głównych składowych pozwala na uwzględnienie zarówno zmienności czasowej, jak i przekrojowej. Takie podejście skutecznie rozwiązuje problem niewystarczającej liczebności próby, a jednocześnie zapewnia uchwycenie ciągłości obserwowanych procesów w czasie.
- Ponadto podjęto próbę oceny zastosowanych w badaniu modeli analizy danych, wykorzystując w tym celu wyniki klasyfikacji uzyskane metodą Warda.

Problem badawczy

Próba składa się z $n = 35$ obserwacji (powiatów), z których każda opisana jest przez $p = 12$ cech statystycznych (związanych z rynkiem pracy) rejestrowanych w $T = 19$ momentach czasu (lata 2004–2022), tym samym każda obserwacja reprezentowana jest przez macierz o wymiarach $p \times T$. W wyniku wektoryzacji otrzymujemy obiekty (wektory) o pT-współrzędnych.

Problem konstrukcji głównych składowych oparty jest na estymowanej macierzy kowariancji (w omawianym przypadku o wymiarze $pT = 12 \times 19 = 228$). Specyfikacja procesu estymacji narzuca warunek $n > pT$. W praktyce jest on trudny do spełnienia (w prezentowanym przypadku mamy $35 \ll 12 \cdot 19 = 228$).

Rozwiązanie: proponujemy dwa podejścia:

- model wykorzystujący transformację poprzez funkcję jądrową - **KPCA**,
- model oparty na analizie danych funkcjonalnych - **FPCA**.

Dodatkowym atutem proponowanych rozwiązań jest uwzględnienie ciągłości czasowej obserwowanych zjawisk.

Analiza głównych składowych - PCA

- Analiza głównych składowych (PCA) jest metodą przekształcającą obserwowane zmienne pierwotne w nowe, wzajemnie ortogonalne zmienne, znane jako główne składowe. Nowych zmiennych jest tyle samo co zmiennych pierwotnych.
- Główne składowe definiuje się jako kombinację liniową zmiennych pierwotnych:

$$Z_j = a_{j1}X_1 + a_{j2}X_2 + \ldots + a_{jk}X_k,$$

gdzie $a_{j1}, a_{j2}, \ldots, a_{jk}$ są elementami wektora własnego odpowiadającego j-tej największej wartości własnej macierzy kowariancji (lub korelacji).

- Każda kolejna główna składowa wyjaśnia malejącą część całkowitej wariancji, a ich łączny wkład sumuje się do wartości odpowiadającej zmiennym pierwotnym.
- Konstrukcja głównych składowych sprowadza się do obliczenia wartości własnych macierzy kowariancji $\mathbf{\Sigma}$, a następnie wektorów własnych a_i z równań:

$$|\mathbf{\Sigma} - \lambda \boldsymbol{I}| = 0, \ (\mathbf{\Sigma} - \lambda \boldsymbol{I})\, a_i = 0.$$

Model jądrowy - KPCA

Rozważmy zbiór danych:

$$\{X_1, X_2, ..., X_n\}, X_i \in \mathbb{R}^{T \times p}, i = 1, 2, ..., n,$$

dla którego obliczamy macierz jądrową:

$$\mathbf{K} = (K_{ij}), K_{ij} = k(\mathbf{X_i}, \mathbf{X_j}),$$

gdzie $k : \mathbb{R}^{T \times p} \times \mathbb{R}^{T \times p} \to \mathbb{R}$ nazywana jest funkcją jądrową (w naszym przypadku jest to gaussowska funkcja jądrowa). Następnie obliczamy:

$$\tilde{\mathbf{K}} = \mathbf{HKH}, \text{ gdzie } \mathbf{H} = \mathbf{I_n} - \frac{1}{n} 1_n 1_n^T.$$

Wyznaczona macierz $\tilde{\mathbf{K}}$ (określana jako scentrowana macierz jądrową) stanowi podstawę konstrukcji głównych składowych.

Model danych funkcjonalnych - FPCA

Dla p-wymiarowego procesu stochastycznego:

$$\mathbf{X} = (X_1, X_2, ..., X_p)'$$

zakładamy, że jego k-ty element może być reprezentowany przez skończoną liczbę funkcji bazowych $\{\varphi_b\}$ (w tym przypadku baza Fouriera):

$$X_k(t) = \sum_{b=0}^{B_k} c_{kb}\varphi_b(t),\ k = 1, 2, ..., p,$$

gdzie c_{kb} są zmiennymi losowymi.
Wykorzystując zapis macierzowy, proces $\mathbf{X}$ może być wyrażony następująco:

$$\mathbf{X}(t) = \mathbf{\Phi}(t)\mathbf{c},\ E(\mathbf{c}) = 0,\ Var(\mathbf{c}) = \mathbf{\Sigma_c},$$

wówczas macierz $\mathbf{\Sigma_c}$ stanowi podstawę konstrukcji głównych składowych.

Źródło danych

Wykorzystano zasoby **Banku Danych Lokalnych** (`https:/bdl.stat.gov.pl`) z następujących kategorii:

- Ludność (K3),
- Rynek pracy (K4),
- Wynagrodzenia i świadczenia społeczne (K40),
- Podmioty gospodarki narodowej, przekształcenia własnościowe i strukturalne (K25).

Dane obejmują:

- wszystkie powiaty województwa wielkopolskiego ($n = 35$),
- kilkanaście zmiennych opisujących sytuację na rynku pracy ($p = 12$),
- okres od 2004 do 2022 roku ($T = 19$).

Wartości wszystkich zmiennych poddano unitaryzacji zerowanej z uwzględnieniem charakteru: stymulanta/destymulanta. Motywacją takiego podejścia był fakt, że zmienne były wyrażone w różnych jednostkach i miały różne poziomy wartości.

Dane rynku pracy

Zmienna		Opis	stymulanta ⊕	destymulanta ⊖
X1	⊖	udział zarejestrowanych bezrobotnych z wyższym wykształceniem w ogólnej liczbie zarejestrowanych bezrobotnych		
X2	⊖	udział zarejestrowanych bezrobotnych pozostających bez pracy dłużej niż 1 rok w ogólnej liczbie zarejestrowanych bezrobotnych		
X3	⊖	udział nowo zarejestrowanych bezrobotnych w ogólnej liczbie zarejestrowanych bezrobotnych		
X4	⊖	udział zarejestrowanych bezrobotnych w wieku poniżej 25 lat w ogólnej liczbie zarejestrowanych bezrobotnych		
X5	⊖	udział zarejestrowanych bezrobotnych w wieku 55 lat i więcej w ogólnej liczbie zarejestrowanych bezrobotnych		
X6	⊖	udział zarejestrowanych bezrobotnych w populacji w wieku produkcyjnym		
X7	⊕	przeciętne miesięczne wynagrodzenie brutto		
X8	⊕	podmioty zarejestrowane w REGON na 1000 mieszkańców		
X9	⊕	podmioty nowo zarejestrowane w REGON na 1000 mieszkańców		
X10	⊖	podmioty wykreślone z REGON na 1000 mieszkańców		
X11	⊖	stopa bezrobocia rejestrowanego		
X12	⊕	oferty pracy na 1000 osób w wieku produkcyjnym		

Struktura danych

Wybrane zmienne tworzą zrównoważoną strukturę, która obejmuje cztery kluczowe aspekty analizy rynku pracy:

1. **Charakterystyka podaży pracy** (Zmienne 1–5),
2. **Wskaźniki wydajności rynku pracy** (Zmienne 6, 11, 12),
3. **Aktywność gospodarcza i przedsiębiorczość** (Zmienne 8–10),
4. **Wartość i produktywność rynku pracy** (Zmienna 7).

Te zmienne skutecznie oddają wielowymiarowy charakter zmienności i dynamiki rynku pracy, umożliwiając tym samym adekwatną ocenę warunków gospodarczych (w kontekście rynku pracy) w powiatach województwa Wielkopolskiego.

Procedura badawcza

Krok 1: Przygotowanie zmiennych

Pobranie zmiennych opisujących rynek pracy z Banku Danych Lokalnych (`bdl.stat.gov.pl`) i przeliczenie ich na odpowiednie wskaźniki.

Krok 2: Zastosowanie unitaryzacji zerowanej

Z uwzględnieniem charakteru zmiennych (stymulanty i destymulanty).

Krok 3: Konstrukcja składowych głównych

- z użyciem gaussowskiej funkcji jądrowej - (Model jądrowy - KPCA),
- z użyciem transformacji do przestrzeni danych funkcjonalnych wykorzystując bazę Fouriera - (Model danych funkcjonalnych - FPCA),
- z użyciem klasycznej analizy składowych głównych (PCA).

Krok 4: Analiza skupień

Zastosowanie hierarchicznej metody Warda opartej na dwóch pierwszych głównych składowych w celu identyfikacji skupień, wykreślenie dendrogramów i kartogramów.

Step 5: Interpretacja wyników

Ocena jakości uzyskanych wyników, rozpoznanie różnic i podobieństw oraz ich interpretacja.

Wyniki - wyjaśniana wariancja

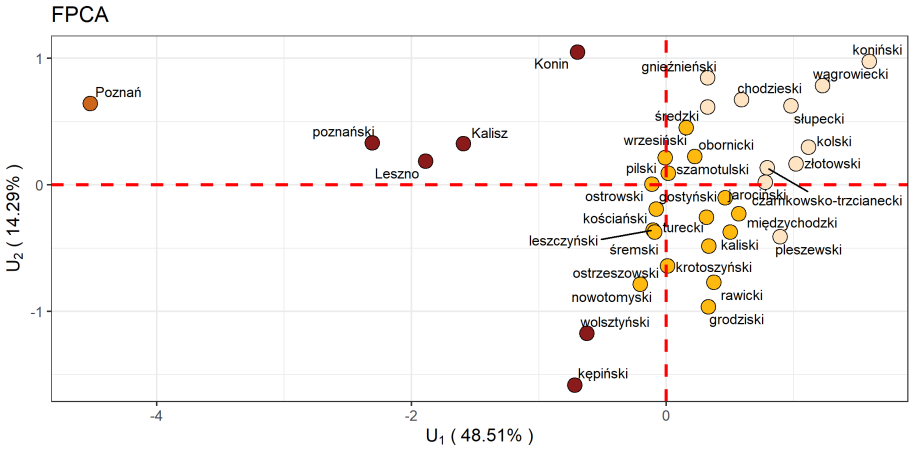
Stopień wyjaśnianej wariancji (trzy pierwsze główne składowe) dla wszystkich modeli zastosowanych w badaniu przedstawiono w poniższej tabeli.

Tabela: Wariancja danych pierwotnych wyjaśniana przez główne składowe

Model		Składowa 1	Składowa 2	Składowa 3
PCA	% wariancji	43.52%	12.97%	7.34%
	skumulowane %	43.52%	56.49%	63.83%
KPCA	% wariancji	16.17%	12.11%	8.97%
	skumulowane %	16.17%	28.28%	37.25%
FPCA	% wariancji	48.51%	14.29%	7.55%
	skumulowane %	48.51%	62.80%	70.35%

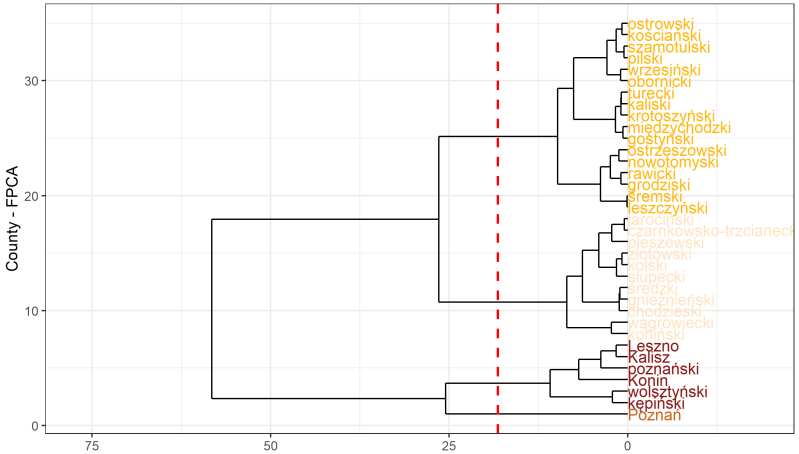
Funkcjonalna analiza głównych składowych (FPCA) wykazuje wyższą skuteczność w identyfikowaniu ukrytych wzorców, podkreślając zalety podejścia funkcjonalnego w efektywnym uchwyceniu dynamicznych aspektów (ciągłości w czasie) obecnych w danych (dotyczących rynku pracy) obejmujących lata 2004–2022.

Wyniki - powiaty (rzut na płaszczyznę) - FPCA



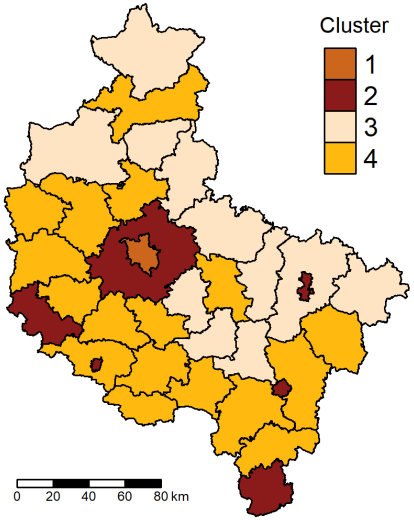
Rysunek: Powiaty na płaszczyźnie dwóch pierwszych głównych składowych - model FPCA

Wyniki - analiza skupień - FPCA



Rysunek: Dendrogram metody Warda - wyodrębnione skupienia - model FPCA

Wyniki - FPCA



Wyodrębnione skupienia powiatów na bazie metody Warda dla dwóch pierwszych głównych składowych

Skupienie	Powiat	Skupienie	Powiat
	Poznań		gostyński
	Kalisz		grodziski
	kępiński		kaliski
	Konin		kościański
	Leszno		krotoszyński
	poznański		leszczyński
	wolsztyński		międzychodzki
	koniński		nowotomyski
	gnieźnieński		obornicki
	chodzieski		ostrowski
	wągrowiecki		ostrzeszowski
	średzki		pilski
	słupecki		rawicki
	kolski		szamotulski
	złotowski		śremski
	jarociński		turecki
	cza-trz		wrzesiński
	pleszewski		

Uwaga: cza-trz = czarnkowsko-trzcianecki

Wnioski

- Funkcjonalna analiza głównych składowych (FPCA) okazała się najefektywniej wyjaśnianiać zmienność badanych zjawisk (spośród rozważanych modeli: PCA, KPCA, FPCA).
- FPCA traktuje obserwacje jako funkcje ciągłe w czasie, co umożliwia głębszą analizę dynamiki zjawisk niż w przypadku klasycznej PCA.
- Struktury skupień uzyskane na podstawie FPCA wykazują wysoką zgodność przestrzenną z wynikami uzyskanymi przy użyciu tradycyjnej PCA.
- Analiza skupień hierarchicznych pozwoliła na wyodrębnienie czterech odrębnych profili rynku pracy w województwie wielkopolskim.
- Zidentyfikowano wyraźny układ centrum–peryferie, z Poznaniem jako ośrodkiem najbardziej korzystnych warunków na rynku pracy.
- Obszary peryferyjne charakteryzują się słabszymi wskaźnikami rynku pracy, przy czym obserwowane wyjątki wynikają głównie z uwarunkowań infrastrukturalnych oraz historycznych.

Wybrana literatura

Górecki, T., Krzyśko, M., Waszak, Ł., Wołyński, W. (2014), *Methods of reducing dimension for functional data*. Statistics in transition, 15(2), 231-242.

Górecki, T., Krzyśko, M., Waszak, Ł., Wołyński, W. (2018), *Selected statistical methods of data analysis for multivariate functional data*. Stat Papers, 59, 153-182.

Gray, V., (2017). *Principal component analysis: methods, applications and technology.* Nova Science Publishers.

Jajuga, K., Walesiak, M., (2000). Standardisation of data set under different measurement scales. *Classification and information processing at the turn of the millennium*, pp. 105–112. Springer.

Lewandowski, P., Magda, I., (2023). *The labor market in Poland, 2000–2021.* IZA World of Labor.

Ramsay, J.O., Silverman, B.W., (2005). *Functional data analysis, 2nd edn.* Springer, Berlin.

Ward, J.H., (1963). *Hierarchical Grouping to Optimize an Objective Function.* Journal of the American Statistical Association, 58(301), pp. 236–244.

Woźniak, M., (2021). *Spatial matching on the urban labor market: estimates with unique micro data.* Journal for Labour Market Research, 55(1), 11.

Yenilmez, F., Esin, K. (Eds.). (2016). *Handbook of research on unemployment and labor market sustainability in the era of globalization.* IGI Global.

Dziękujemy za uwagę