Klasyfikacja lokalnych rynków pracy z wykorzystaniem przestrzenno-czasowej analizy składowych głównych

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

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

02–04.07.2024

1. Cel prezentacji
2. Opis metody
3. Źródło danych i zmienne diagnostyczne
4. Wybrane wyniki
5. Kluczowe wnioski
6. Kierunki dalszych prac badawczych
7. Literatura

Cel prezentacji

Główne cele prezentacji:

- ocena sytuacji na rynku pracy w przekroju powiatów województwa wielkopolskiego,
- omówienie zastosowanego podejścia tj. przestrzenno-czasowej analizy składowych głównych (ang. spatio-temporal principal component analysis - **STPCA**),
- przedstawienie najważniejszych wyników, ich interpretacja oraz sformułowanie wniosków z przeprowadzonej analizy.

Problem metodologiczny

- Próba składa się z $n = 35$ obserwacji opisanych $p = 12$ cechami statystycznymi w $T = 19$ momentach czasu, zatem każda z obserwacji jest reprezentowana przez macierz o wymiarach $p \times T = 12 \times 19$.
- Dokonując wektoryzacji otrzymujemy obiekty będące wektorami o $pT = 228$ wymiarach.
- **Problem** konstrukcji składowych głównych sprowadza się do estymacji macierzy kowariancji stopnia $pT = 228$. Warunkiem poprawności oszacowania jest $n > pT$, co w omawianym przypadku nie zachodzi, mamy bowiem: $35 \ll 12 \cdot 19 = 228$.
- **Rozwiązanie:** przejście z „przestrzeni pierwotnej" do „przestrzeni danych funkcjonalnych".
- Operując w konwencji FPCA (Functional Principal Component Analisys) wzbogaciliśmy model o komponent zależności przestrzennych - **STPCA (Spatio-Temporal Principal Component Analisys)**.

Funkcjonalna analiza składowych głównych (FPCA)
Dane funkcjonalne

Dla p-wymiarowego procesu stochastycznego:

$$\boldsymbol{X} = (X_1, X_2, ..., X_p)^\top$$

zakładamy, że jego i-ta składowa może być reprezentowana przez skończoną liczbę funkcji bazowych $\{\varphi_{ij}\}$:

$$X_i(t) = \sum_{j=1}^{B_i} \alpha_{ij}\varphi_{ij}(t),\ j = 1, \ldots, B_i, i = 1, \ldots, p,$$

gdzie α_{ij} są zmiennymi losowymi.

Funkcjonalna analiza składowych głównych (FPCA)
Dane funkcjonalne (cd.)

Używając notacji macierzowej, proces $\boldsymbol{X}$ ma postać:

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\,\boldsymbol{\alpha},\ E(\boldsymbol{\alpha}) = 0,\ Var(\boldsymbol{\alpha}) = \boldsymbol{\Sigma}, \tag{1}$$

gdzie

$$\boldsymbol{\alpha} = \left(\alpha_{11}, \ldots, \alpha_{1B_1}, \ldots, \alpha_{p1}, \ldots, \alpha_{pB_p}\right)^T,$$

$$\boldsymbol{\Phi}(t) = diag\left(\boldsymbol{\phi}_1^T(t), \ldots, \boldsymbol{\phi}_p^T(t)\right),$$

$$\boldsymbol{\phi}_i^T(t) = \left(\varphi_{i1}(t), \ldots, \varphi_{iB_i}(t)\right),$$

$$\boldsymbol{\alpha} \in \mathbb{R}^B, B = B_1 + \cdots + B_p, i = 1, \ldots, p.$$

Funkcjonalna analiza składowych głównych (FPCA) Rozwiązanie

Funkcjonalne składowe główne definiujemy następująco:

$$U = \langle \boldsymbol{u}, \boldsymbol{X} \rangle = \int_a^b \boldsymbol{u}^\top(t)\boldsymbol{X}(t)dt,$$

gdzie $\boldsymbol{X}$ oraz $\boldsymbol{u}$ są elementami tej samej przestrzeni $\mathcal{L}_2^p([a, b])$.
Przyjmując $\boldsymbol{X}$ postaci (1) oraz $\boldsymbol{u}$ wyrażone przez:

$$\boldsymbol{u}(t) = \boldsymbol{\Phi}(t)\gamma, \ \gamma \in \mathbb{R}^B,$$

otrzymujemy

$$U = \gamma^\top \boldsymbol{\alpha},$$

$$\mathrm{Var}(U) = \gamma^\top \boldsymbol{\Sigma}\gamma,$$

gdzie $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{\alpha})$.

Funkcjonalna analiza składowych głównych (FPCA)
Rozwiązanie (cd.)

Zakładając, że $\boldsymbol{a}$ jest estymatorem (MNK) losowego wektora $\boldsymbol{\alpha}$, oraz $\boldsymbol{A}_{n\times B} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$, mamy:

$$\boldsymbol{S} = \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{A}, \tag{2}$$

gdzie estymator wariancji FPCA ma postać:

$$\widehat{\mathrm{Var}(U)} = \boldsymbol{\gamma}^\top \boldsymbol{S}\boldsymbol{\gamma} = \frac{1}{n}\boldsymbol{\gamma}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{\gamma}.$$

W kolejnym kroku, model uzupełnimy o komponent "przestrzenny".

Zależności przestrzenne - macierz wag przestrzennych

Rozpatrujemy obserwacje prowadzone w punktach (przestrzeni) oraz relacje pomiędzy parami punktów. Formalnie mamy n punktów, w których dokonujemy obserwacji $x_1, \ldots, x_n$. Punkty te wraz z ich zależnościami tworzą sieć opisaną macierzą wag przestrzennych $\boldsymbol{W}$. Wagi w_{ij}, $(i, j = 1, 2, \ldots, n)$ wyrażają występowanie zależności pomiędzy punktami i oraz j, jak i (opcjonalnie) siłę tej zależności.
W niniejszym opracowaniu rozpatrzono 9 różnych macierzy wag $\boldsymbol{W}$. Najczęściej stosowane wagi w_{ij} są funkcjami:

- odległości między jednostkami geograficznymi (tu powiatami),
- długości granicy między jednostkami geograficznymi,
- obu wymienionych jednocześnie.

Wyniki STPCA są zależne od wybranej macierzy wag przestrzennych.

Zależności przestrzenne - zastosowane wagi

No.	Wagi przestrzenne	Parametr
1.	k-Nearest Neighbour Weights (standard form)	$k = 3, 4, 6$
2.	k-Nearest Neighbour Weights (symmetric form)	$k = 3, 4, 6$
3.	Radial Distance Weights	nd.
4.	Power Distance Weights	$\alpha = 1, 2$
5.	Exponential Distance Weights	$\alpha = 1, 2$
6.	Double-Power Distance Weights	$k = 2, 3, 4$
7.	Rook Contiguity Weights	nd.
8.	FPCA	nd.

Wszystkie macierze wag skonstruowano dla 4 wariantów odległości (S-distance, S-time, F-distance i F-time)
distance - odległość geograficzna, time - czas dojazdu,
S - shortest, F - fastest

Zależności przestrzenne - współczynnik Geary'ego

Konstruując funkcjonalne składowe główne $U = \boldsymbol{\gamma}^\top \boldsymbol{\alpha}$, szukamy wektora γ, który maksymalizuje $\widehat{\mathrm{Var}(U)}$ przy $\boldsymbol{\gamma}^\top \boldsymbol{\gamma} = 1$. Dodatkowo, uwzględniamy aspekt przestrzenny, wykorzystując współczynnik Geary'ego C.
Przyjmijmy, że w n punktach przestrzeni $s_1, \ldots, s_n$ obserwujemy $x_1 = x(s_1), \ldots x(s_n)$ oraz niech $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $y_i = x_i - \bar{x}$, $i = 1, 2, \ldots, n$. Przyjmujemy również, że wagi przestrzenne w_{ij} spełniają warunki: $w_{ij} \geq 0$ dla dowolnych $i \neq j$, $w_{ii} = 0$ dla wszystkich i.
Niech $\boldsymbol{W} = (w_{ij})_{n\times n}$ oznacza macierz wag oraz $a = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$. Wówczas współczynnik Geary'ego C zdefiniowany jest następująco (Geary, 1954):

$$C(y_1, \ldots, y_n) = \frac{n-1}{2a} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(y_i - y_j)^2}{\sum_{i=1}^n y_i^2}.$$

C jest nieujemną miarą autokorelacji przestrzennej.

Zależności przestrzenne - współczynnik Geary'ego (cd.)

Współczynniki autokorelacji przestrzennej określają, czy i w jakim stopniu obserwacje x_i oddziałują na siebie wzajemnie poprzez przyjętą strukturę. Rozważmy $\boldsymbol{B} = (B_{ij})$ - macierz stopnia n o elementach:

$$B_{ij} = \begin{cases} R_i + K_j - 2w_{ij}, & \text{if } i = j, \\ -2w_{ij}, & \text{if } i \neq j, \end{cases}$$

gdzie R_i jest sumą elementów i-tego wiersza macierzy $\boldsymbol{W}$ oraz K_j jest sumą elementów j-tej kolumny tej macierzy.
Współczynnik Geary'ego C przyjmuje wówczas postać (Jong et al., 1984):

$$C(\boldsymbol{y}) = \frac{n-1}{2a} \frac{\boldsymbol{y}^\top \boldsymbol{B}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}}, \tag{3}$$

gdzie $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$.

STPCA
Przestrzenno-czasowa analiza składowych głównych

Niech $\boldsymbol{y} = \boldsymbol{A}\gamma$ będzie rzutem macierzy obserwacji $\boldsymbol{A}$ na przestrzeń funkcjonalnych składowych głównych $U = \gamma^\top \boldsymbol{\alpha}$ określonych przez wektor kierunkowy γ.
Wówczas:

$$C(\boldsymbol{y}) = C(\boldsymbol{A}\gamma) = \frac{n-1}{2a} \frac{\gamma^\top \boldsymbol{A}^\top \boldsymbol{B}\boldsymbol{A}\gamma}{\gamma^\top \boldsymbol{A}^\top \boldsymbol{A}\gamma}.$$

Chcemy wyznaczyć wektor γ funkcjonalnej składowej głównej $U = \gamma^\top \boldsymbol{\alpha}$ tak, aby maksymalizował iloczyn wariancji tej składowej oraz współczynnik Geary'ego rzutu macierzy obserwacji na tę składową.

STPCA (cd.)

Innymi słowy, szukamy wektora γ który spełnia warunek:

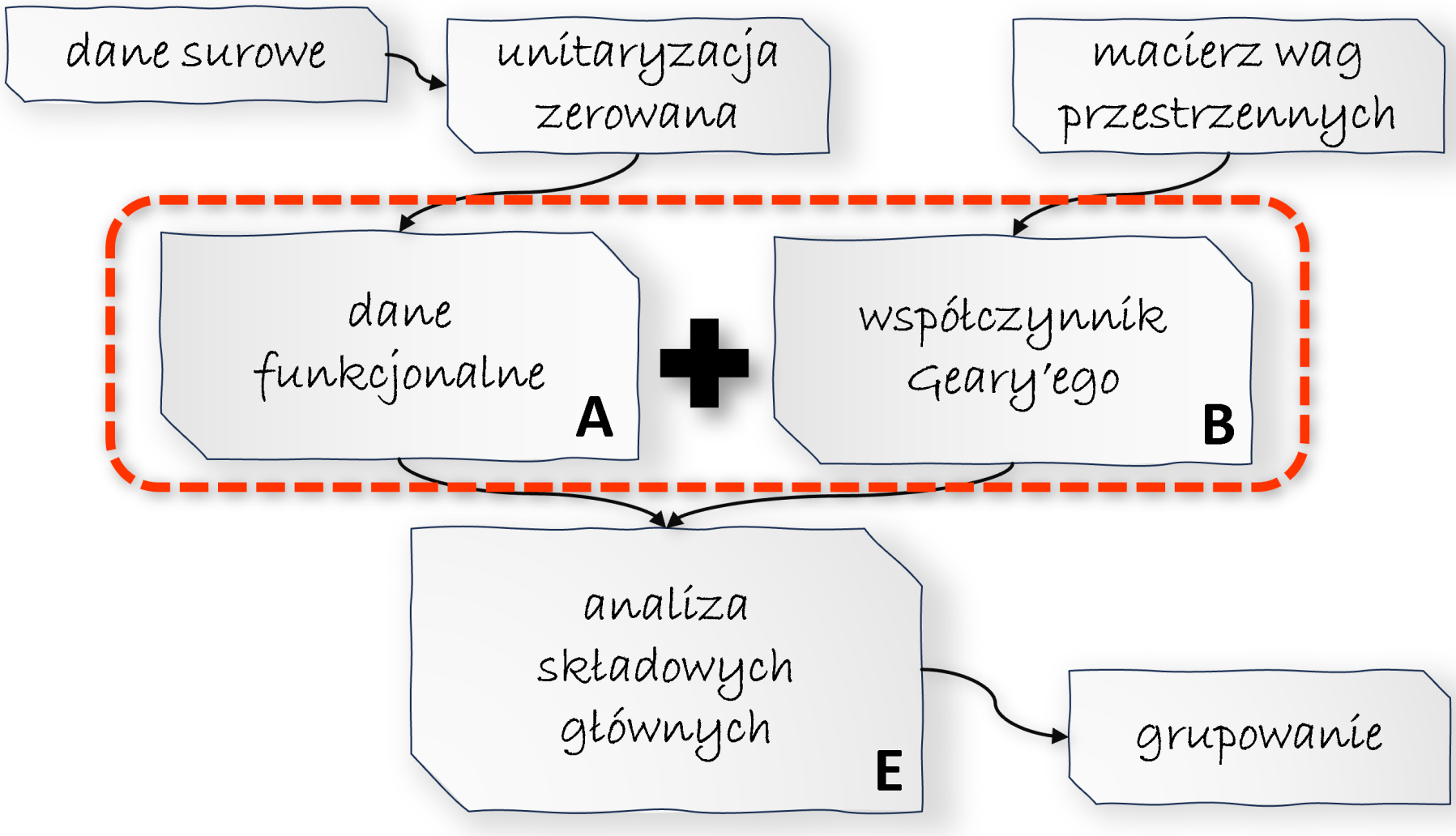
$$arg\max_{\gamma} \widehat{\mathrm{Var}(U)}C(\boldsymbol{A}\gamma) = arg\max_{\gamma} \frac{1}{n}\gamma^{\top}\boldsymbol{A}^{\top}\boldsymbol{A}\gamma\frac{n-1}{2a}\frac{\gamma^{\top}\boldsymbol{A}^{\top}\boldsymbol{B}\boldsymbol{A}\gamma}{\gamma^{\top}\boldsymbol{A}^{\top}\boldsymbol{A}\gamma}$$
$$= arg\max_{\gamma} \frac{n-1}{2na}\gamma^{\top}\boldsymbol{A}^{\top}\boldsymbol{B}\boldsymbol{A}\gamma = arg\max_{\gamma}\gamma^{\top}\boldsymbol{E}\gamma,$$

przy $\gamma^{\top}\gamma = 1$, gdzie

$$\boldsymbol{E} = \frac{n-1}{2na}\boldsymbol{A}^{\top}\boldsymbol{B}\boldsymbol{A}. \tag{4}$$

Nowy układ współrzędnych (przestrzenno-czasowe składowe główne), wyznaczony jest poprzez wektory własne macierzy $\boldsymbol{E}$ odpowiadające jej niezerowym wartościom własnym.

Schemat badania



Źródło danych

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

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

Dane zostały poddane wstępnemu przygotowaniu:

- przeliczono wielkości bezwzględne odpowiednio, w relacji do liczby mieszkańców powiatu lub do ogólnej liczby bezrobotnych w powiecie,
- zastosowano unitaryzację zerowaną (z uwzględnieniem stymulant i destymulant).

Zmienne diagnostyczne

Badaniem statystycznym objęto:

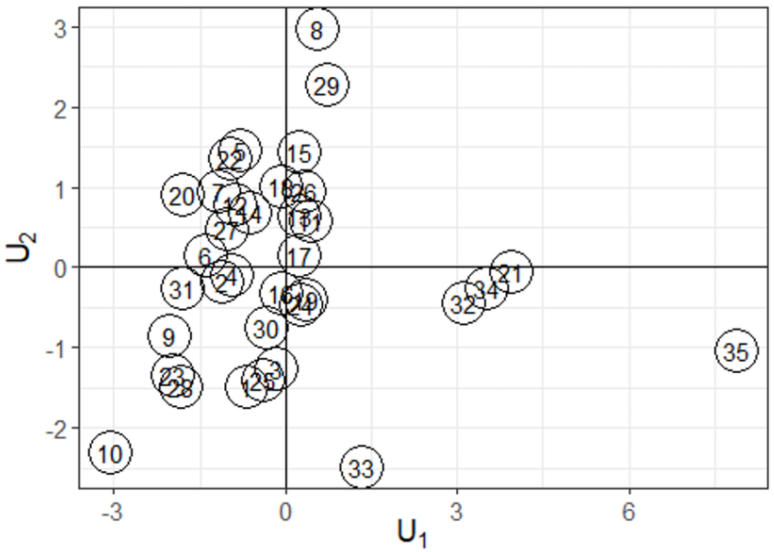
- wszystkie powiaty województwa wielkopolskiego ($n = 35$),
- dwanaście charakterystyk związanych z rynkiem pracy ($p = 12$),
- okres czasu odnoszący się do lat 2004–2022 ($T = 19$).

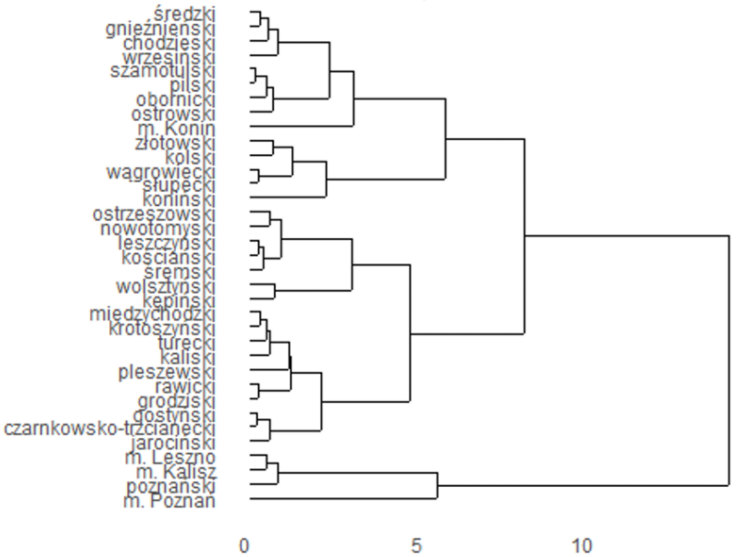
ZMN1 –	Odsetek zarejestrowanych bezrobotnych z wyższym wykształceniem (w bezrobotnych ogółem)	ZMN7 –	Przeciętne wynagrodzenie brutto
ZMN2 –	Odsetek zarejestrowanych bezrobotnych pozostających bez pracy dłużej niż 1 rok (w bezrobotnych ogółem)	ZMN8 –	Podmioty wpisane do REGON na 1000 mieszkańców
ZMN3 –	Odsetek nowo zarejestrowanych bezrobotnych (w bezrobotnych ogółem)	ZMN9 –	Podmioty nowo zarejestrowane w REGON na 1000 mieszkańców
ZMN4 –	Odsetek zarejestrowanych bezrobotnych w wieku 24 lata i mniej (w bezrobotnych ogółem)	ZMN10 –	Podmioty wyrejestrowane z REGON na 1000 mieszkańców
ZMN5 –	Odsetek zarejestrowanych bezrobotnych w wieku 55 lat i więcej (w bezrobotnych ogółem)	ZMN11 –	Stopa bezrobocia
ZMN6 –	Odsetek pracujących w ludności w wieku produkcyjnym	ZMN12 –	Oferty pracy na 1000 mieszkańców w wieku produkcyjnym
	Stymulanta		Destymulanta

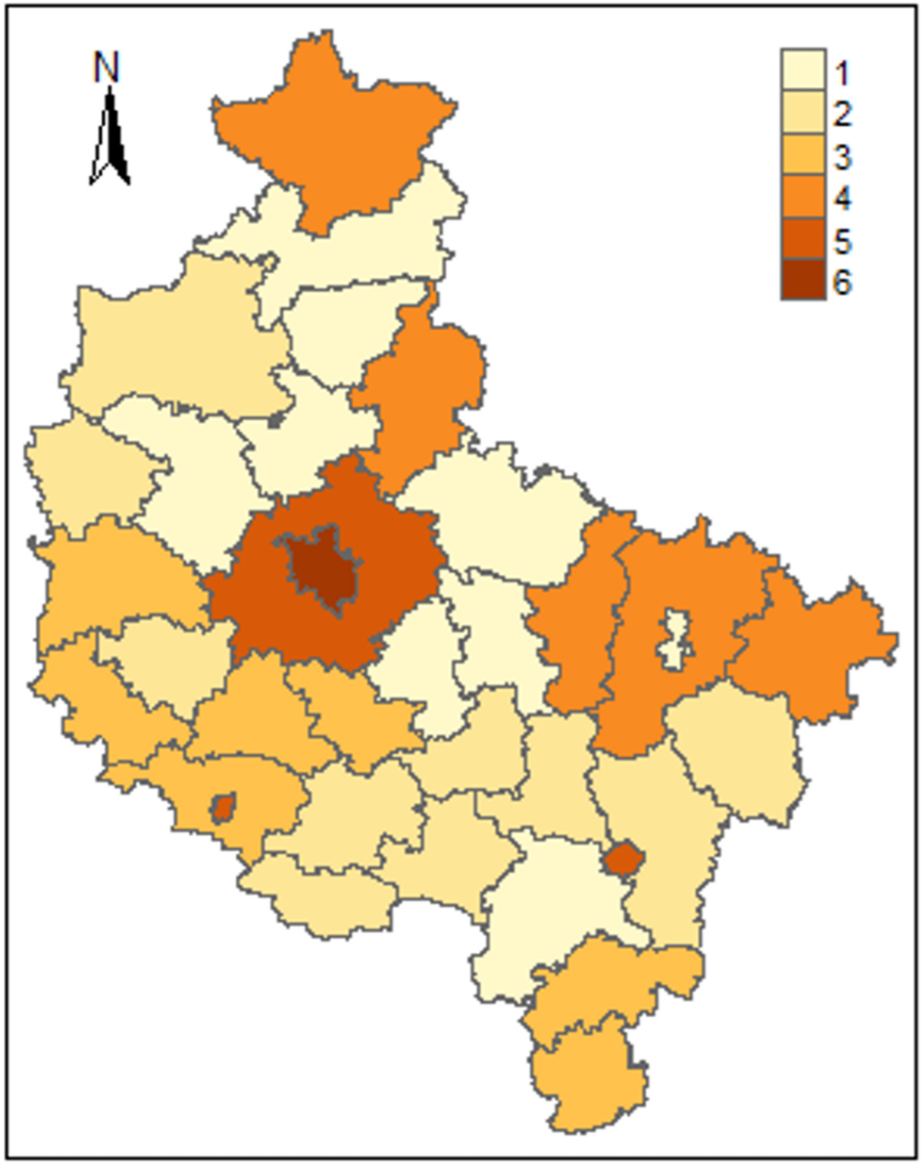
Wyniki: dla dwóch pierwszych składowych głównych

Model		S-dist	S-time	`F-dist	F-time
FPCA		60.13	60.13	60.13	60.13
***k*-Nearest Neighbour Weights**	**k = 3**	**61.39**	60.79	60.69	60.32
	k = 4	**62.54**	58.67	62.33	58.42
	k = 6	61.03	**62.42**	61.68	61.19
***k*-Nearest Neighbour Weights (symmetric form)**	**k = 3**	**62.11**	60.12	60.79	61.69
	k = 4	60.13	59.04	**61.87**	60.66
	k = 6	61.15	**63.18**	61.45	61.47
Radial Distance Weights		62.85	**63.08**	62.70	62.43
Power Distance Weights;	**α = 1**	61.97	61.98	**62.08**	61.87
	α = 2	65.10	63.31	**65.66**	62.98
Exponential Distance Weights	**α = 1**	62.11	**62.57**	62.26	62.16
	α = 2	63.25	**63.64**	63.38	63.34
Double-Power Distance Weights	**k = 2**	64.34	**65.09**	64.43	64.82
	k = 3	64.45	**65.32**	64.57	65.06
	k = 4	64.45	**65.41**	64.60	65.16
Rook Contiguity Weights		54.69	54.69	54.69	54.69

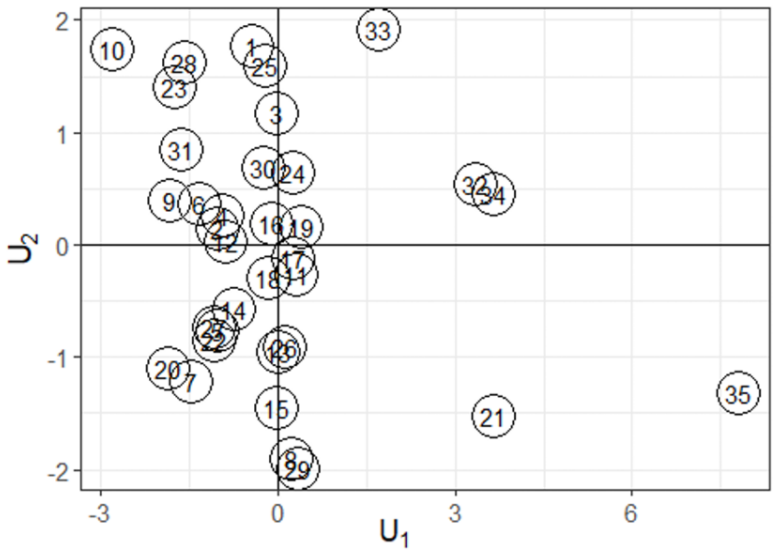
Wyniki: FPCA
(bez komponentu przestrzennego)

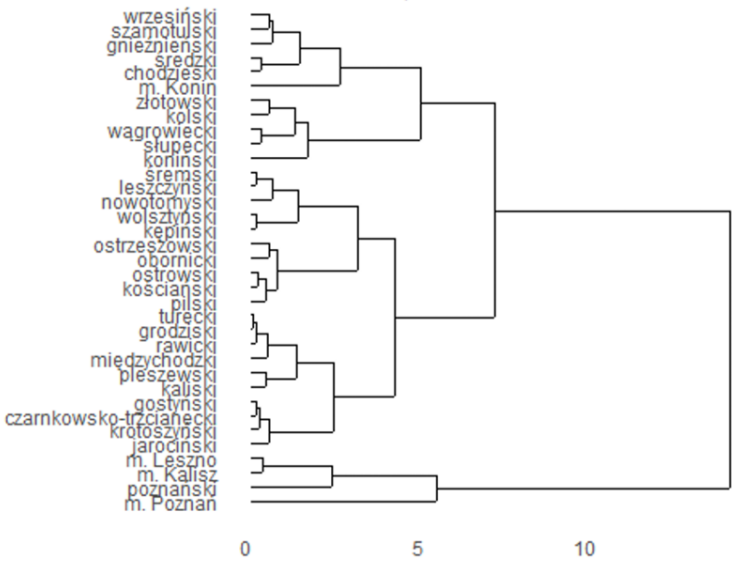


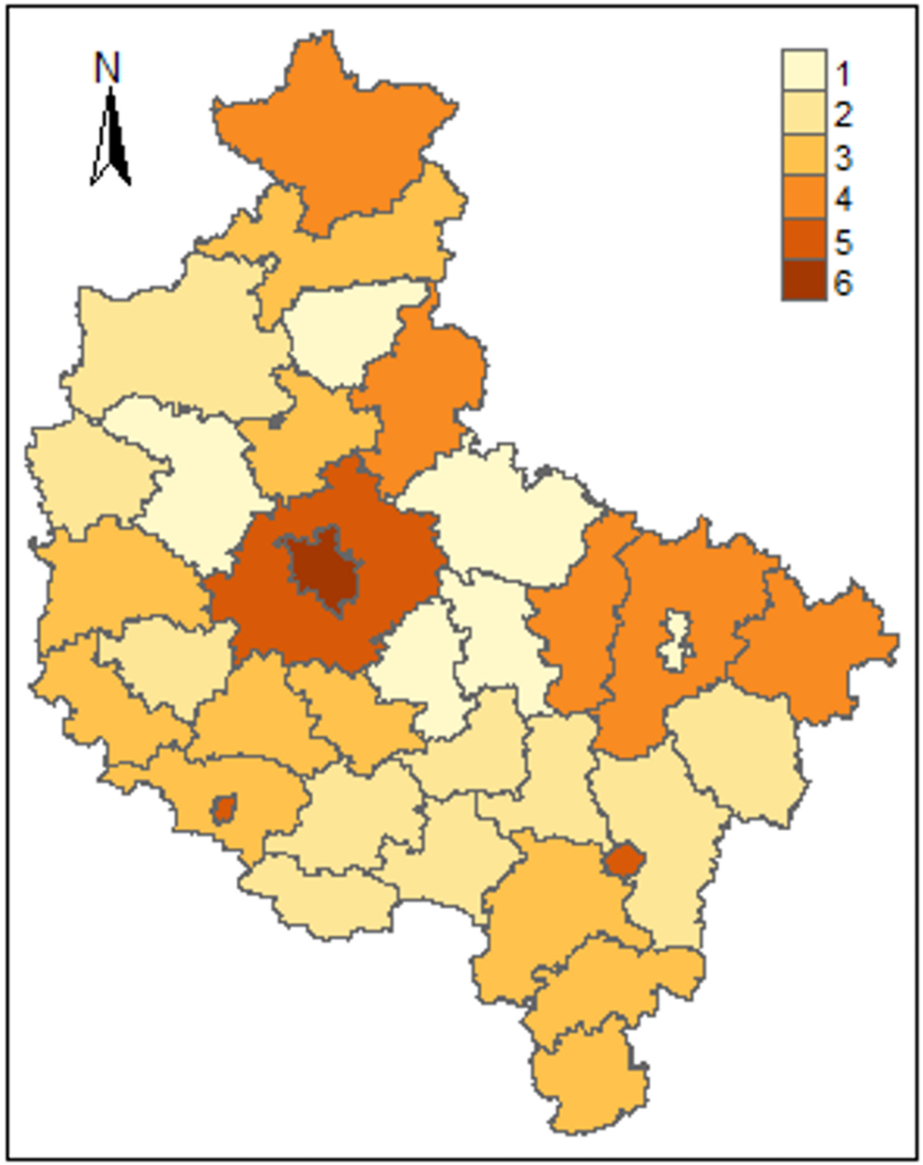




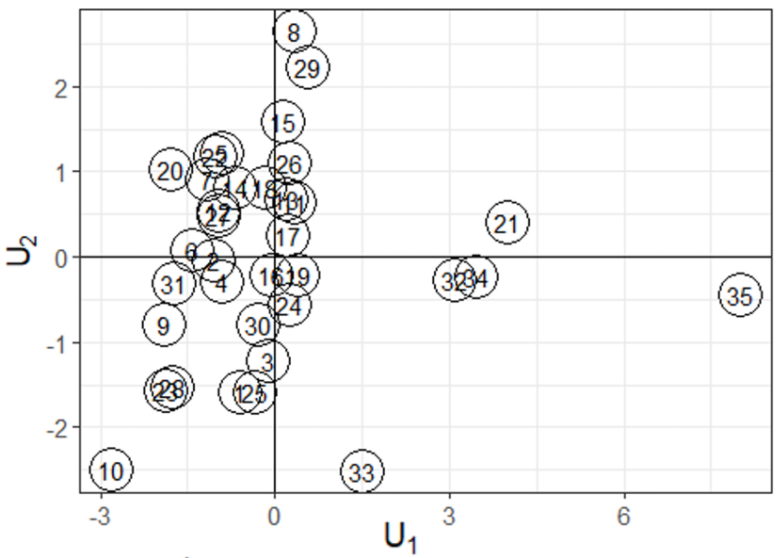
Wyniki: STPCA S-distance (Power Distance Weights, $\alpha = 2$)

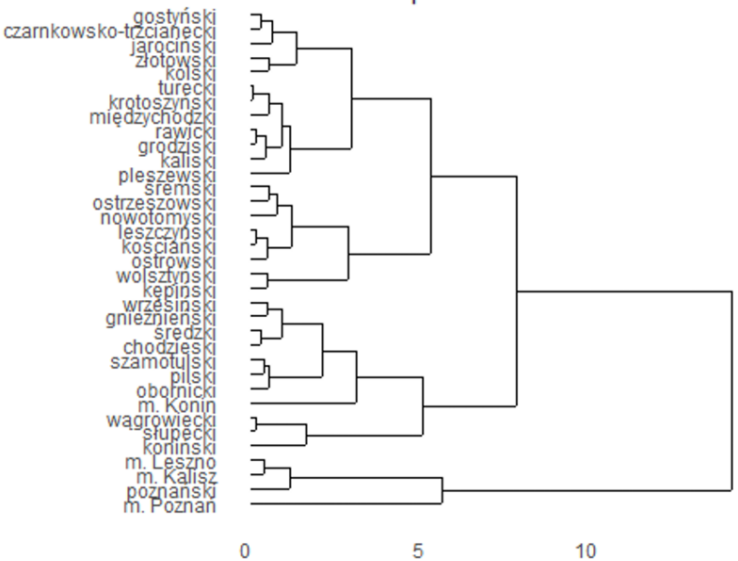


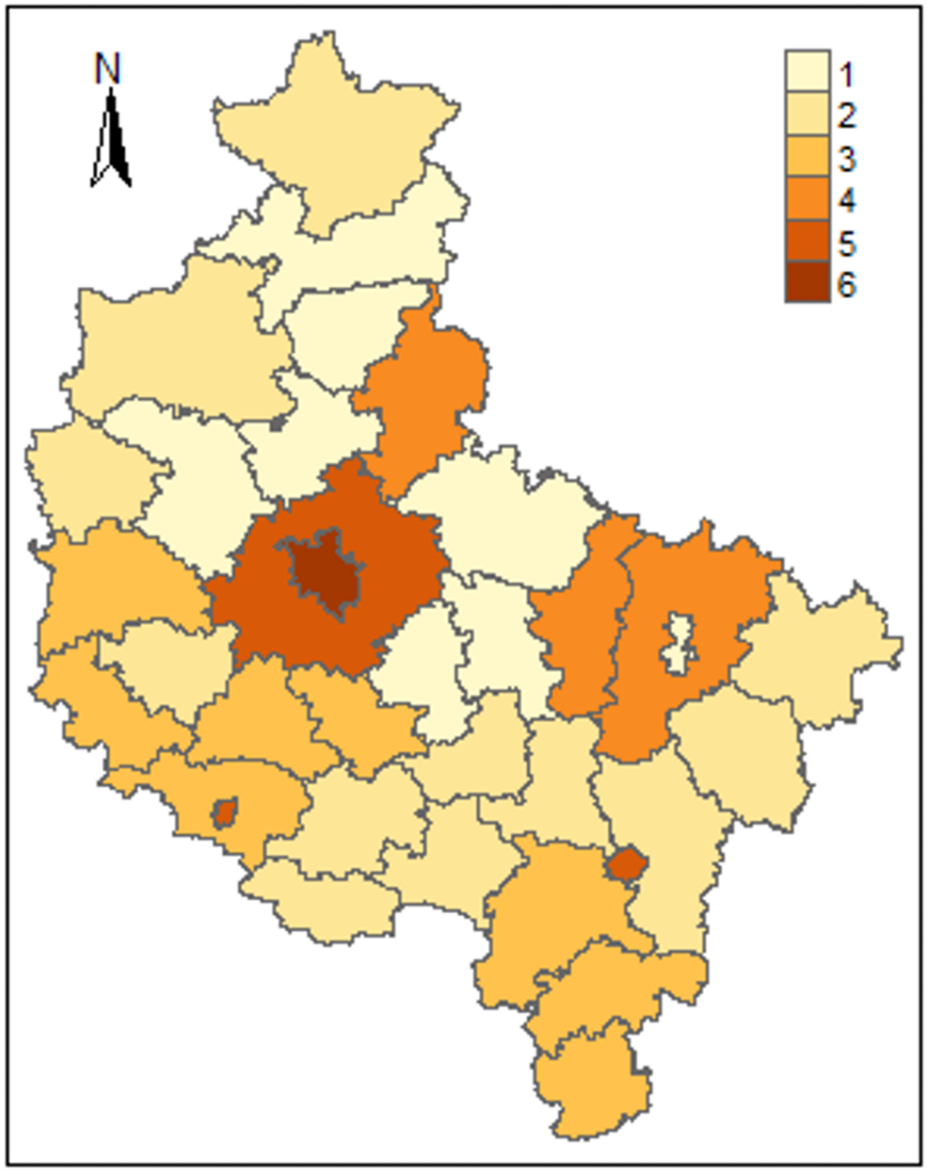




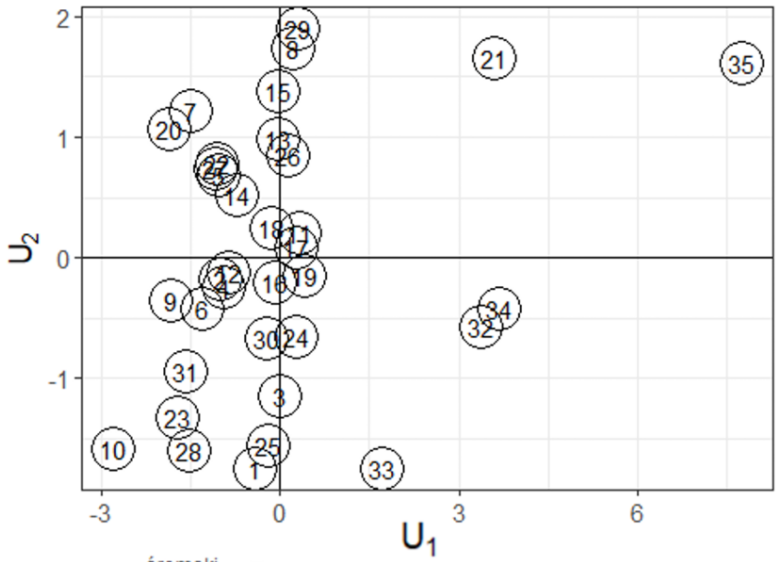
Wyniki: STPCA S-time (Double-Power Distance Weights, $k = 4$)

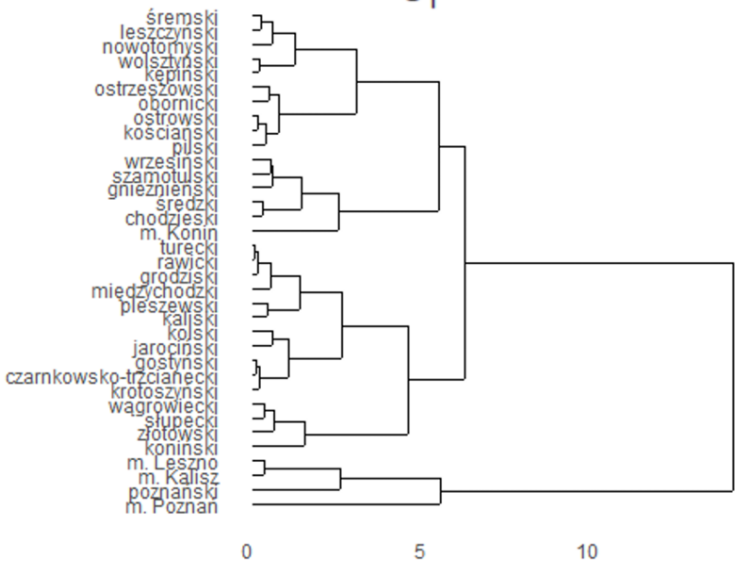


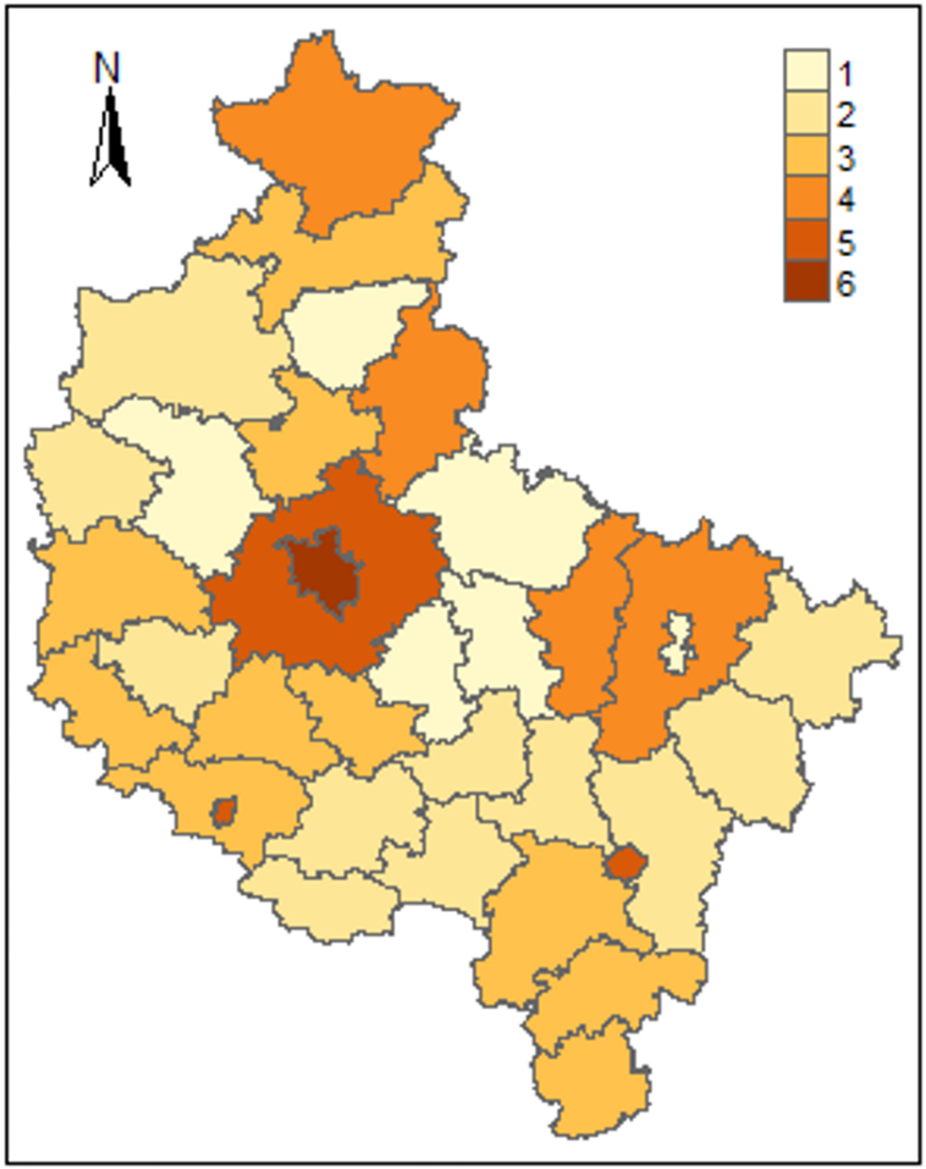




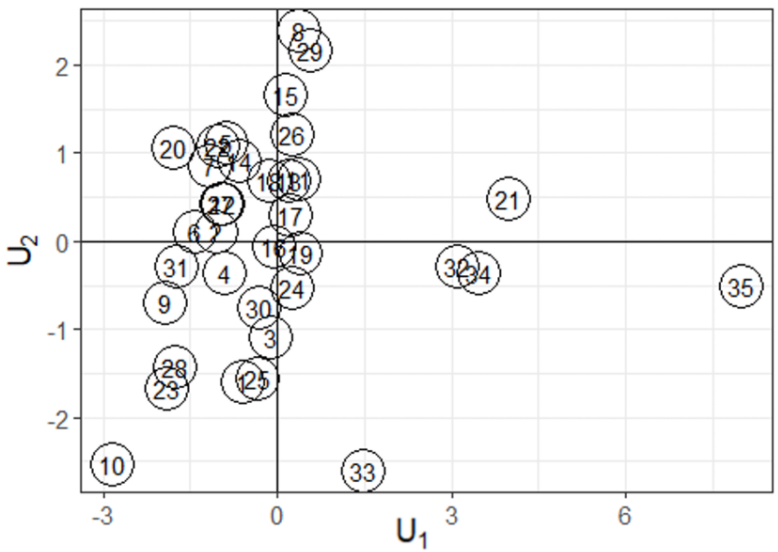
Wyniki: STPCA F-distance (Power Distance Weights, $\alpha = 2$)

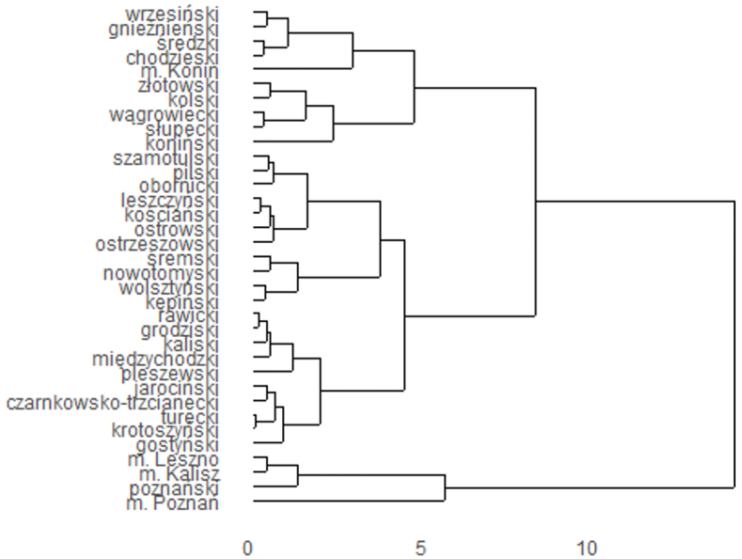


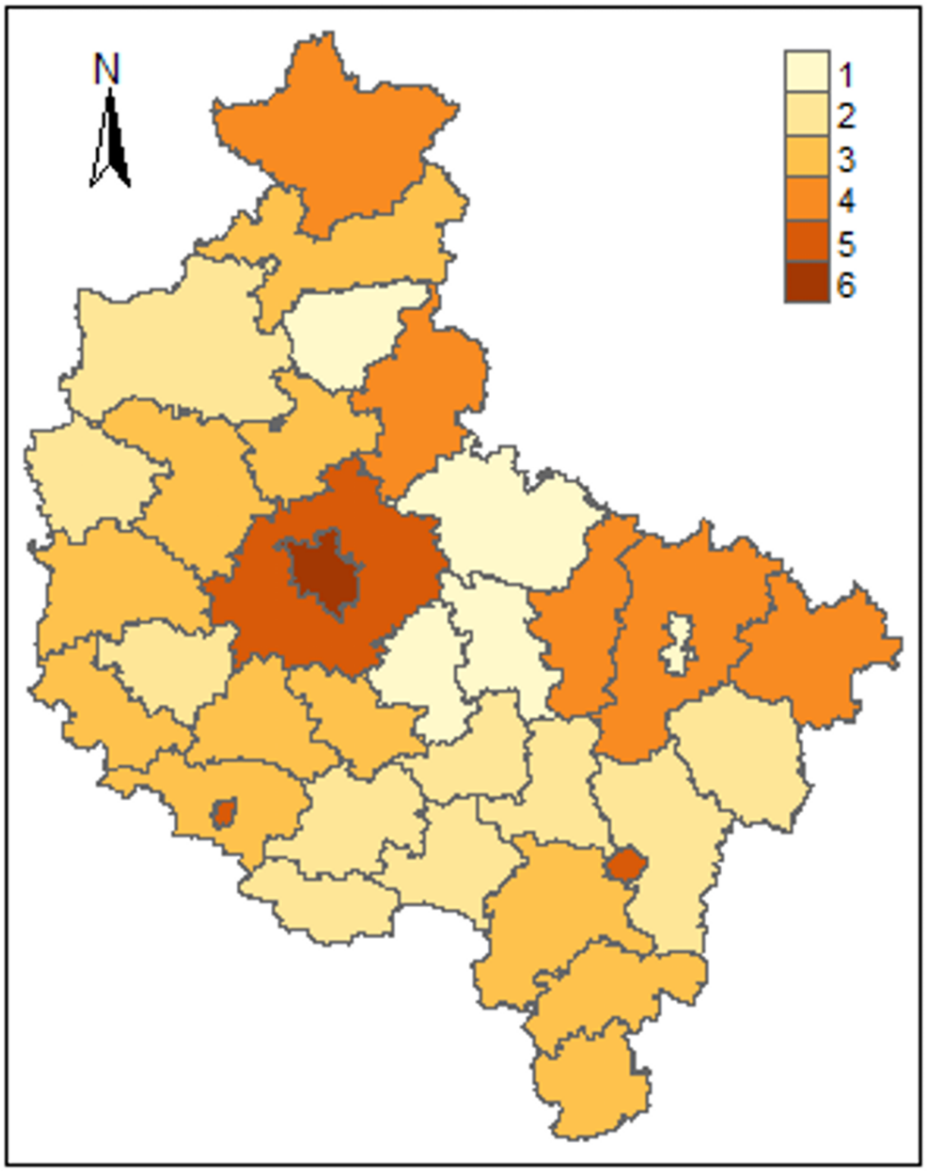




Wyniki: STPCA F-time (Power Distance Weights, $\alpha = 2$)







Kluczowe wnioski

- STPCA stanowi użyteczną metodę badawczą, w której uwaga koncentruje się na jednoczesnym uwzględnieniu czasowego komponentu cech opisujących jednostki badania wraz z ich przestrzenną lokalizacją.
- Uwzględnienie w badaniu, z wykorzystaniem funkcjonalnej analizy składowych głównych, komponentu przestrzennego w postaci odpowiednio dobranej macierzy wag przestrzennych zwiększa procent wyjaśnianej wariancji.
- Wyniki analiz w kontekście przestrzennego zróżnicowania lokalnych rynków pracy w województwie wielkopolskim są na ogół podobne, choć w niektórych przypadkach (dla pewnych macierzy wag) można jednak zaobserwować nieco odmienne wyniki grupowania powiatów.

Kierunki dalszych prac badawczych

- badanie dla wszystkich powiatów ($n = 380$),
- badanie na poziomie województw ($n = 16$),
- uwzględnienie w badaniach innych macierzy wag przestrzennych (również innych parametrów k, α),
- inne modyfikacje aparatu statystycznego (wprowadzenie w modelach elementów „wygładzania” w wymiarze czasowym).

Literatura

Geary, R. (1954), The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician. **5**, 115.

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

Hotelling, H. (1933), Analysis of app. complex of statistical variables into principal components.. Journal Of Educational Psychology. **24**, pp. 417–441.

Kossowski, T. (2020), Wybrane zagadnienia modelowania matematyczno-statystycznego struktur i procesów przestrzennych. Rozwój Regionalny i Polityka Regionalna., 159–174.

Krzyśko, M., Nijkamp, P., Ratajczak, W., Wołyński, W., Wojtyła, A. & Wenerska, B. (2023), A novel spatio-temporal principal component analysis based on Geary's contiguity ratio. Computers, Environment And Urban Systems. **103** pp. 101980.

Ramsay, J., Hooker, G. & Graves, S. (2009), Introduction to Functional Data Analysis. Functional Data Analysis With R And MATLAB. pp. 1–19.

Dziękujemy za uwagę!