PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 6 (1206) 2008
Ekonometria 21

Dariusz Biskup

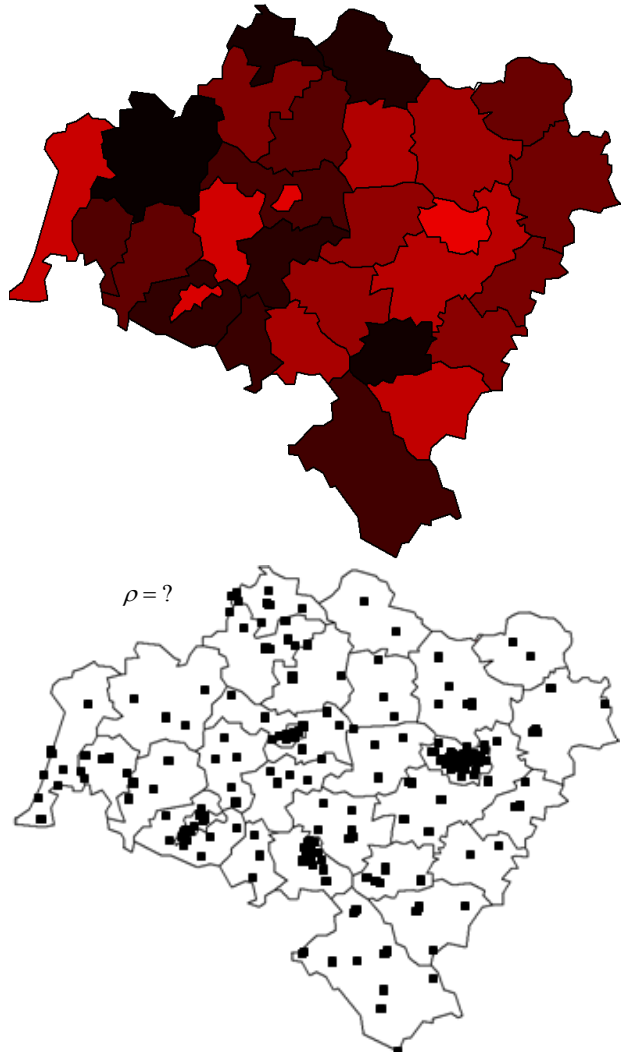
POMIAR ZALEŻNOŚCI MIĘDZY ZJAWISKAMI MIERZONYMI NA RÓŻNYCH POZIOMACH AGREGACJI PRZESTRZENNEJ

1. Wstęp

W przypadku analizy danych regionalnych często dane dotyczące pewnych zmiennych są zebrane na różnych poziomach agregacji. Jedna z tego typu sytuacji omówiona w niniejszym artykule polega na tym, iż dane są zebrane na poziomie punktowym (na przykład za pomocą stacji monitorującej poziom zanieczyszczenia środowiska w konkretnej lokalizacji), a istnieje jednak potrzeba wyznaczenia średniego poziomu rozpatrywanej cechy na pewnym obszarze geograficznym (na przykład w gminie lub w powiecie). Zatem na podstawie pewnej liczby pomiarów punktowych należy wyznaczyć poziom cechy w zadanej liczbie regionów. Celem tego rodzaju agregacji może być konieczność dopasowania obszarów, na których mierzonych jest kilka zmiennych. Na przykład jeżeli chcemy zmierzyć korelację między poziomem zachorowania na pewną chorobę, dla której dane zbierane są na poziomie regionu, a poziomem zanieczyszczenia środowiska, który mierzony jest punktowo, istnieje konieczność dopasowania obszaru pomiarowego zanieczyszczenia środowiska do obszaru, w wypadku którego mamy dane o zachorowalności.

W artykule podjęto problem predykcji poziomów zanieczyszczenia powietrza w sytuacji, gdy analizowane jest jednocześnie kilka rodzajów zanieczyszczeń. Poziom zanieczyszczenia środowiska jest zjawiskiem, które analizować można w wymiarze przestrzennym, w którym zakłada się na ogół, że wielkość korelacji poziomu zanieczyszczenia pomiędzy punktami pomiarowymi jest zależna od odległości pomiędzy nimi.

Estymacja poziomu zanieczyszczenia w sytuacji, gdy mierzone jest kilka rodzajów zanieczyszczenia, może być dokonywana na podstawie zarówno występujących korelacji przestrzennych, jak i korelacji związanych z występowaniem zależności pomiędzy kilkoma typami zanieczyszczeń. Można przy tym wyróżnić dwa typy zagadnień estymacji: sytuację, w której w danej lokalizacji nie są znane żadne pomiary, i sytuację, w której znane są poziomy zanieczyszczeń dla wybranych substancji.



Rys. 1. Dane regionalne a dane przestrzenne

Źródło: opracowanie własne.

Otrzymane wyniki prognoz posłużą do wyznaczenia średnich poziomów zanieczyszczenia w powiatach, a następnie zbadania korelacji pomiędzy nimi a zachorowalnością na nowotwory. Parametry opisywanego modelu szacowane będą metodami bayesowskimi przy użyciu algorytmów Monte Carlo.

Obliczenie współczynnika korelacji wymaga „uzgodnienia" typów danych – na podstawie punktowych danych o zanieczyszczeniu należy wyznaczyć średnie poziomy zanieczyszczeń w ramach powiatów. Opisana sytuacja została schematycznie przedstawiona na rys. 1.

Aby dopasować dane punktowe do danych regionalnych, należy wyznaczyć dla każdego obszaru A_i następującą średnią:

$$Y(A_i) = \frac{1}{|A_i|} \int_{s_0 \in A_i} Y(s_0)\, ds_0,$$

gdzie s_0 oznacza dowolną lokalizację, a $Y(s_0)$ poziom zjawiska w lokalizacji s_0.

2. Dane

Wykorzystane zostaną następujące zbiory danych:

- zachorowalność na raka na Dolnym Śląsku w 2003 r.,
- średnioroczne zanieczyszczenie dwutlenkiem siarki w 191 punktach pomiarowych zlokalizowanych na Dolnym Śląsku w 2003 r.,
- średnioroczne zanieczyszczenie dwutlenkiem azotu w 191 punktach pomiarowych zlokalizowanych na Dolnym Śląsku w 2003 r.,
- średnioroczne zanieczyszczenie pyłem zawieszonym PM10 w 67 punktach pomiarowych zlokalizowanych na Dolnym Śląsku w 2003 r.

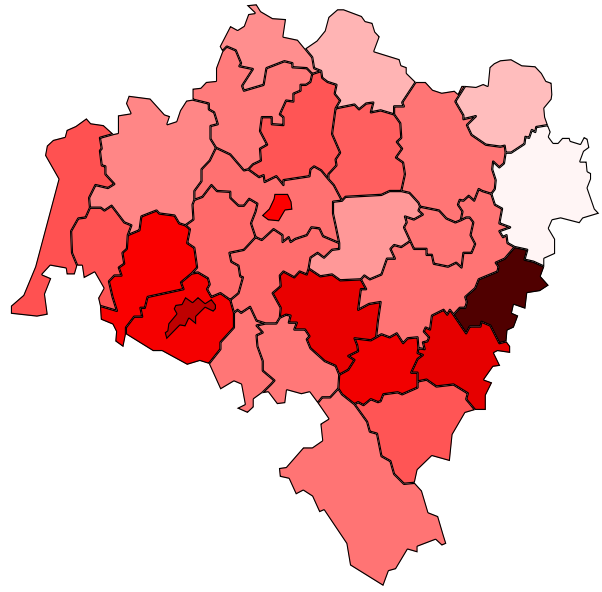
Dane pochodzą z publikacji *Raport o stanie środowiska w województwie dolnośląskim w 2003 roku* oraz *Nowotwory złośliwe w województwie dolnośląskim w roku 2003* (por. [5; 3]). Jak już wspomniano, dane dotyczące zachorowania na raka mają charakter zagregowany i rozpatrywane będą na poziomie powiatów. Dane dotyczące zanieczyszczeń mają charakter punktowy, co oznacza, że obliczenie korelacji między zachorowalnością a poziomem zanieczyszczenia wymagać będzie wyznaczenia średniego poziomu zanieczyszczenia na poziomie powiatowym. Dane dotyczące zanieczyszczeń dostępne są dla różnej liczby lokalizacji. Estymacja poziomu zanieczyszczenia w powiatach możliwa będzie dzięki występowaniu dwóch typów korelacji:

- przestrzennej, która oznacza, że korelacja pomiędzy poziomami zanieczyszczeń jest tym większa, im bliższe są punkty pomiarowe,
- pomiędzy poszczególnymi typami zanieczyszczeń.

Tabela 1. Dane o zachorowalności na raka

Powiat	Liczba	Wskaźnik	Powiat	Liczba	Wskaźnik
Bolesławiecki	250	27,77	polkowicki	181	28,95
Dzierżoniowski	390	34,92	strzeliński	167	36,23
Głogowski	248	26,97	średzki	128	25,90
Górowski	87	22,95	świdnicki	603	35,90
Jaworski	166	30,48	trzebnicki	227	29,65
Jeleniogórski	232	34,61	wałbrzyski	180	29,02
Kamiennogórski	144	29,47	wołowski	159	31,96
Kłodzki	534	29,66	wrocławski	282	29,75
Legnicki	163	30,07	ząbkowicki	240	32,98
Lubański	199	33,26	zgorzelecki	335	33,41
Lubiński	363	32,99	złotoryjski	146	30,76
Lwówecki	175	34,11	Jelenia Góra	375	40,34
Milicki	82	21,89	Legnica	368	33,72
Oleśnicki	168	**15,99**	Wałbrzych	397	29,47
Oławski	377	**52,15**	Wrocław	2590	40,86

Źródło: [3].



Rys. 2. Przestrzenne zróżnicowanie wskaźników zachorowalności na raka w województwie dolnośląskim

Źródło: opracowanie własne.

W tabeli przedstawiono liczbę przypadków zachorowań na raka w poszczególnych powiatach oraz wskaźnik opisujący liczbę przypadków na 10 000 mieszkańców. Zaznaczono (pismem płgrubym) również dwa powiaty, w których wskaźnik ten jest najmniejszy oraz największy (powiaty oleśnicki oraz oławski). Z kolei rys. 2 ilustruje graficznie przestrzenne zróżnicowanie zachorowalności na raka (kolory ciemniejsze oznaczają wyższe wartości wskaźnika).

Poziomy zanieczyszczeń zmierzono ogólnie dla 235 lokalizacji. Zanieczyszczenia SO_2 i NO_2 zmierzone zostały w 191 (tych samych) lokalizacjach. Zanieczyszczenia pyłami zmierzono w 67 lokalizacjach (z czego dla 23 lokalizacji istnieją również obserwacje SO_2 i NO_2). Dane zmierzone zostały w $\mu g/m^3$. Dla każdej lokalizacji zmierzono długość i szerokość geograficzną (przy użyciu programu *Nawigator Mapa Polski*), które zostały przeliczone na współrzędne kilometrowe.

3. Wstępna analiza danych

Zastosowany w dalszej części artykułu model wymaga, aby dane charakteryzowały się dwoma typami korelacji: przestrzenną oraz związaną z typami zanieczyszczeń. Aby stwierdzić, czy poszczególne rodzaje zanieczyszczeń są ze sobą skorelowane, obliczono macierz korelacji; przedstawiono ją w tab. 2. Ze względu na to, że poszczególne typy zanieczyszczeń mierzone są w różnych punktach pomiarowych współczynniki korelacji dla par zmiennych (SO_2, Pyły) i (NO_2, Pyły) obliczono dla 23 obserwacji, a współczynnik korelacji dla pary (SO_2, NO_2) obliczono dla 191 obserwacji. Wszystkie współczynniki korelacji są statystycznie istotne i z wyjątkiem pary SO_2 i NO_2 są stosunkowo wysokie.

Tabela 2. Macierz korelacji zanieczyszczeń

Zmienne	SO_2	NO_2	Pyły
SO_2	1	0,3	0,71
NO_2	0,3	1	0,63
Pyły	0,71	0,63	1

Źródło: opracowanie własne.

Drugi typ analizy wstępnej polegać będzie na stwierdzeniu, czy w miarę zmniejszania się odległości pomiędzy punktami pomiarowymi wzrasta korelacja pomiędzy pomiarami tego samego typu zanieczyszczenia. Podstawowym narzędziem sprawdzania, czy dane charakteryzują się zależnością przestrzenną, jest wariogram (por. [4]). Wariogram definiowany jest jako wariancja różnicy zmiennych losowych określonych w lokalizacji s i lokalizacji $s+h$:

Tabela 3. Punkty pomiarowe, ich lokalizacje oraz poziomy zanieczyszczeń

Lp.	Miasto	Ulica	SO_2	NO_2	Pyły	Długość geograficzna			Szerokość geograficzna		
						stopnie	minuty	sekundy	stopnie	minuty	sekundy
1	Białka		7,4	10,9	23,4	16	7	24,34	51	11	57,4
2	Bielawa	Bankowa	20,1	16,5		16	36	30,54	50	41	7,97
3	Bielawa	Grota-Roweckiego			19,7	16	37	7,54	50	41	0,08
4	Bogatynia	Kusocińskiego/Daszyńskiego	11	15,5		14	57	25,26	50	54	15,15
5	Bogatynia	Chopina			31,3	14	57	56,54	50	54	17,5
6	Bolesławiec	Chrobrego	10	21,7		15	34	16,54	51	15	48,84
7	Bolesławiec	Górników			42,1	15	34	6,18	51	15	27,13
8	Brzeg Dolny	Słowackiego	5,1	14,5		16	41	23,19	51	16	1,28
9	Brzeg Głogowski		6,2	13,4		15	54	56,63	51	41	50,66
10	Bystrzyca Kłodzka	Wojska Polskiego	22,8	18,8		16	38	28,07	50	18	0,02
11	Długołęka	Wiejska	6,1	20,8		17	11	26,17	51	10	37,73
12	Długopole-Zdrój	Leśna	7,7	10,2		16	38	2,76	50	14	18,5
13	Duszniki-Zdrój	Zdrojowa	3,9	9,8		16	23	19,76	50	23	55,8
14	Duszniki-Zdrój	Zielona			10,7	16	23	21,07	50	24	12,74
15	Działoszyn		10,9	9,1	19,3	14	56	44,17	50	58	44,36
16	Dzierżoniów	Mickiewicza	11,1	22,8		16	38	43,8	50	43	43,62
17	Dzierżoniów	Osiedle Błękitne	9,3	23,3		16	38	39	50	44	19,23
18	Dzierżoniów	Krasickiego			26,9	16	38	48,72	50	43	37,23
19	Gaworzyce		8	14,7		15	52	57,65	51	37	51,99
20	Głogów	Norwida	9,2	19,6	22,8	16	5	13,28	51	39	27,25
21	Głogów	Sikorskiego	10,8	14,3	26	16	3	52,8	51	39	52,36
22	Głogów	Wojska Polskiego	9,7	17,7		16	3	52,56	51	39	35,68
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
235	Żmigród	Wiejska	5,8	15		16	54	27,79	51	28	2,06

Źródło: [5].

$$Var[Y(s+h) - Y(s)] = 2\gamma(h).$$

Funkcja $2\gamma(h)$ nazywana jest wariogramem, a funkcja $\gamma(h)$ semiwariogramem. Istnieje następująca zależność między wariogramem a funkcją kowariancji:

$$\gamma(h) = Var[Y(s)] - Cov[Y(s+h), Y(s)].$$

Estymacja wariogramu odbywa się przez następujące wyrażenie:

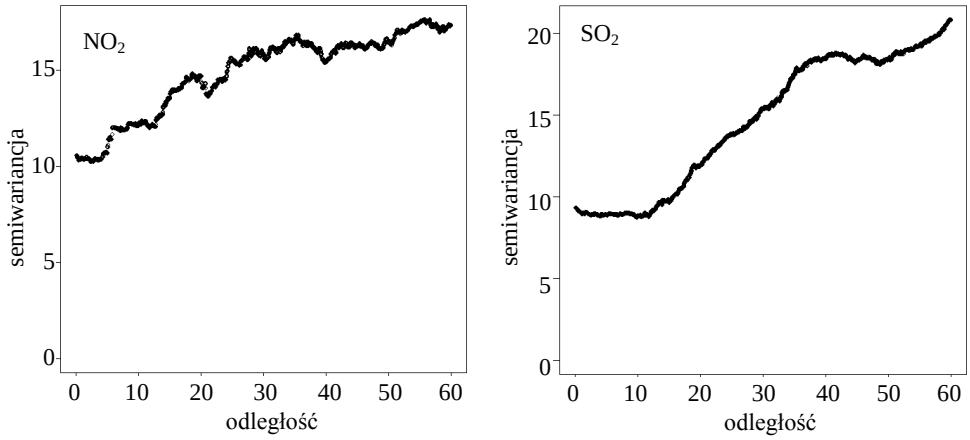
$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{s_i, s_j \in N(h)} [Y(s_i) - Y(s_j)]^2,$$

gdzie $N(h)$ oznacza zbiór par lokalizacji, które są odległe o *h*, a $|N(h)|$ oznacza liczbę takich par. Najczęściej stosowaną parametryczną formą wariogramu jest funkcja wykładnicza. Funkcja kowariancji odpowiadająca takiemu wariogramowi ma postać:

$$Cov(h) = \sigma^2 \exp(-\phi h).$$

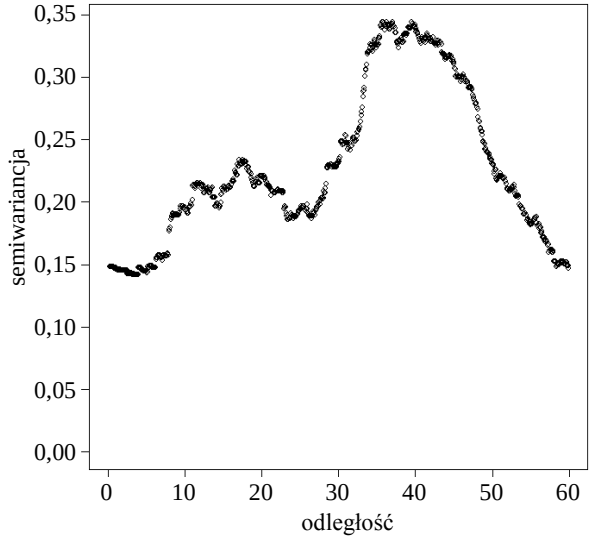
Występowanie zależności przestrzennej powoduje, że wariogram jest funkcją rosnącą (wówczas odpowiednio funkcja kowariancji maleje w miarę wzrostu odległości).

Rysunek 3 wskazuje, że semiwariancja zwiększa się w miarę wzrostu odległości, co umożliwia zastosowanie dla tych typów zanieczyszczeń modeli korelacji przestrzennej.



Rys. 3. Wariogramy dla NO_2 i SO_2

Źródło: opracowanie własne.



Rys. 4. Wariogram dla pyłów

Źródło: opracowanie własne.

Rys. 5. Mapa punktów pomiarowych

Źródło: opracowanie własne.

Wariogram dla pyłów nie pozwala na stwierdzenie występowania zależności przestrzennej będącej funkcją odległości. W związku z tym w dalszej analizie zostaną one pominięte, gdyż nie będzie istniała możliwość wiarygodnego przewidywania poziomów zanieczyszczeń w miejscach, które są pozbawione stacji pomiarowych. Oznacza to, że w analizie wykorzystanie zostanie 191 punktów pomiarowych, w których zmierzono zanieczyszczenie dwutlenkiem siarki oraz azotu.

4. Wielowymiarowe modele korelacji przestrzennej

Przedstawiony zostanie najpierw skrótowo model jednowymiarowy, który zostanie następnie uogólniony na sytuację, w której analizowane jest jednocześnie kilka rodzajów zanieczyszczeń (por.[1]).

Zmienna zależna poziomu zanieczyszczenia $Y(\mathbf{s})$ opisana jest w lokalizacji $\mathbf{s}$ następującym równaniem:

$$\mathbf{Y}(\mathbf{s}) = \mathbf{x}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \varepsilon(\mathbf{s}),$$

gdzie $\varepsilon(\mathbf{s}) \sim N(0, \tau)$, $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^{\mathrm{T}} \sim N(\mathbf{0}, \sigma^2\mathbf{R})$. Wektor $\mathbf{x}^{\mathrm{T}}(\mathbf{s})$ określa współrzędne geograficzne lokalizacji, natomiast $\boldsymbol{\beta}$ jest wektorem parametrów.

Macierz kowariancji wektora $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^{\mathrm{T}}$ ma postać:

$$Cov(\mathbf{Y}) = \sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I},$$

gdzie $\mathbf{R}$ jest to macierz korelacyjna opisana wzorem:

$$R_{ij} = \rho(\mathbf{s}_i, \mathbf{s}_j; \phi).$$

W przypadku zastosowania wykładniczej funkcji kowariancji

$$\rho(\mathbf{s}_i, \mathbf{s}_j; \phi) = \exp(-h_{ij}\phi).$$

Jeśli analizujemy jednocześnie kilka poziomów zanieczyszczeń dla każdej lokalizacji $\mathbf{s}$, określony jest wektor zmiennych zależnych

$$\mathbf{Y}(\mathbf{s}) = \left[Y_1(\mathbf{s}) \ \ldots \ Y_n(\mathbf{s}) \right]^{\mathrm{T}}$$

oraz macierz zmiennych niezależnych o wymiarach $m \times p$ postaci

$$\mathbf{X}(\mathbf{s}) = \left[\mathbf{x}_1^{\mathrm{T}}(\mathbf{s}) \ \ldots \ \mathbf{x}_m^{\mathrm{T}}(\mathbf{s}) \right]^{\mathrm{T}},$$

które powiązane są ze sobą następującym równaniem:

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + \mathbf{W}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s}),$$

gdzie $\boldsymbol{\varepsilon}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Psi})$, $\mathbf{W}(\mathbf{s}) = W_1(\mathbf{s}) \ldots W_m(\mathbf{s})^{\mathrm{T}}$ jest wielowymiarowym procesem gaussowskim o zerowym wektorze wartości oczekiwanych i funkcji kowariancyjnej postaci:

$$\mathbf{K}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = Cov\{W_i(\mathbf{s}), W_j(\mathbf{s}')\}_{i,j=1}^{m},$$

gdzie $\boldsymbol{\theta}$ oznacza wektor parametrów.

Dla zbioru n lokalizacji $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $m \times n$-elementowy wektor $\mathbf{W} = \mathbf{W}\{\mathbf{s}_i\}_{i=1}^{n}$ ma wielowymiarowy rozkład normalny o zerowej wartości oczekiwanej i macierzy kowariancji

$$\boldsymbol{\Sigma}_{\mathbf{W}}(\boldsymbol{\theta}) = \mathbf{K}\{\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta}\}_{i,j=1}^{n}.$$

Macierz $\boldsymbol{\Sigma}_{\mathbf{W}}(\boldsymbol{\theta})$ ma wymiar $mn \times mn$, a jej elementem o numerze (i, j) jest blok o wymiarze $m \times n$ równy $\mathbf{K}\{\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta}\}$. Macierz kowariancji zmiennych zależnych $\mathbf{Y} = \mathbf{Y}\{\mathbf{s}_i\}_{i=1}^{n}$ jest równa

$$\boldsymbol{\Sigma}_{\mathbf{W}}(\boldsymbol{\theta}) + \mathbf{I}_n \otimes \boldsymbol{\Psi},$$

gdzie $\mathbf{I}_n$ oznacza macierz jednostkową, a $\otimes$ iloczyn Kroneckera.

Istnieje kilka modeli pozwalających na powiązanie zależności związanej z typami zanieczyszczeń oraz zależności przestrzennej. Najprostszym z nich jest model wykorzystujący założenie o separowalności. Założenie o separowalności efektów korelacji przestrzennej oraz korelacji pomiędzy czynnikami zanieczyszczeń oznacza, że macierz

$$\boldsymbol{\Sigma}_{\mathbf{W}}(\boldsymbol{\theta}) = \mathbf{R}(\boldsymbol{\theta}) \otimes \mathbf{K}(0, \boldsymbol{\theta}),$$

gdzie $\mathbf{K}(0, \boldsymbol{\theta})$, to macierz kowariancji poziomów zanieczyszczeń w tej samej lokalizacji, a $\mathbf{R}(\boldsymbol{\theta})$ to macierz kowariancji przestrzennej analogiczna do przypadku 1-wymiarowego. Wadą tego modelu jest zastosowanie tej samej funkcji kowariancji przestrzennej dla każdego typu zanieczyszczeń.

Bardziej złożony sposób strukturyzacji macierzy kowariancji zastosowanych jest w modelach koregionalizacji. W modelach koregionalizacji zakłada się, że

$$\mathbf{W}(\mathbf{s}) = \mathbf{A}(\mathbf{s})\tilde{\mathbf{W}}(\mathbf{s}),$$

gdzie $\tilde{\mathbf{W}}(\mathbf{s}) = \{\tilde{W}_i(\mathbf{s})\}_{i=1}^{m}$, przy czym $Var\{\tilde{W}_i(\mathbf{s})\} = 1$, $Cov\{\tilde{W}_i(\mathbf{s}), \tilde{W}_i(\mathbf{s}')\} = \rho_i\{\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_i\}$, $Cov\{\tilde{W}_i(\mathbf{s}), \tilde{W}_j(\mathbf{s})\} = 0$, dla $i \neq j$.

Macierz kowarian0cji procesu $\widetilde{\mathbf{W}}(\mathbf{s})$ ma postać:

$$\widetilde{\mathbf{K}}(\mathbf{s},\mathbf{s}';\boldsymbol{\theta}) = \widetilde{\mathbf{K}}(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})_{i=1}^{n}.$$

Ponadto:

$$\boldsymbol{\Sigma}_{\widetilde{\mathbf{W}}} = \widetilde{\mathbf{K}}(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})_{i=1}^{n},$$

$$\boldsymbol{\Sigma}_{\mathbf{W}} = \left(\mathbf{I}\otimes \mathbf{A}(\mathbf{s})\right)\boldsymbol{\Sigma}_{\widetilde{\mathbf{W}}}\left(\mathbf{I}\otimes \mathbf{A}(\mathbf{s})^{\mathrm{T}}\right).$$

Zaletą modeli koregionalizacji jest możliwość niezależnej estymacji parametrów korelacji przestrzennej.

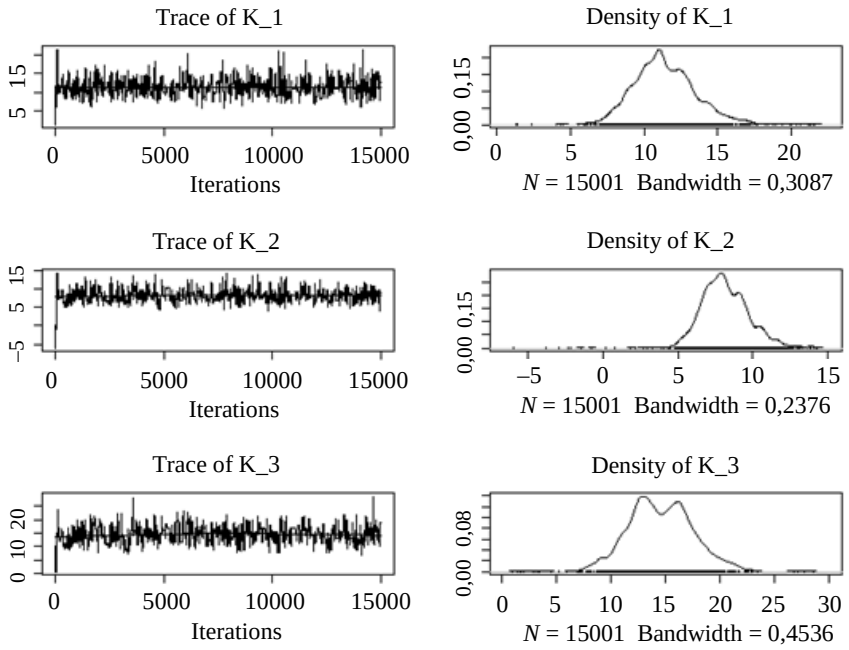
Do estymacji poziomów zanieczyszczeń dla rozpatrywanych w artykule danych zastosowany zostanie model koregionalizacji. Celem jest estymacja zbioru parametrów Ω o postaci $\Omega = (\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\theta}, \boldsymbol{\Psi})$, a następnie rozkładu *a posteriori* prognozy:

$$p(Y(\mathbf{s}^*)|\text{dane}) = \int p(Y(\mathbf{s}^*)|\text{dane}, \Omega)\, p(\Omega|\text{dane})\, d\Omega.$$

5. Estymacja modelu

Do estymacji opisanego modelu koregionalizacji zastosowane zostało podejście bayesowskie. Rozkłady *a posteriori* parametrów wyznaczone zostały za pomocą pakietu *spBayes* języka *R*. Czas obliczeń związanych z estymacją parametrów wyniósł ok. 2 godz. (Pentium IV, 2.8 GHz). Wadą algorytmu jest jego duża złożoność obliczeniowa rzędu n^3, wynikająca z konieczności odwracania macierzy rzędu n w każdej iteracji. Przeprowadzono 15 000 iteracji algorytmu MCMC. Wykorzystany został mieszany algorytm Gibbsa – dla parametrów, dla których można wyznaczyć analitycznie rozkłady warunkowe, i algorytm Metropolisa – dla pozostałych parametrów.

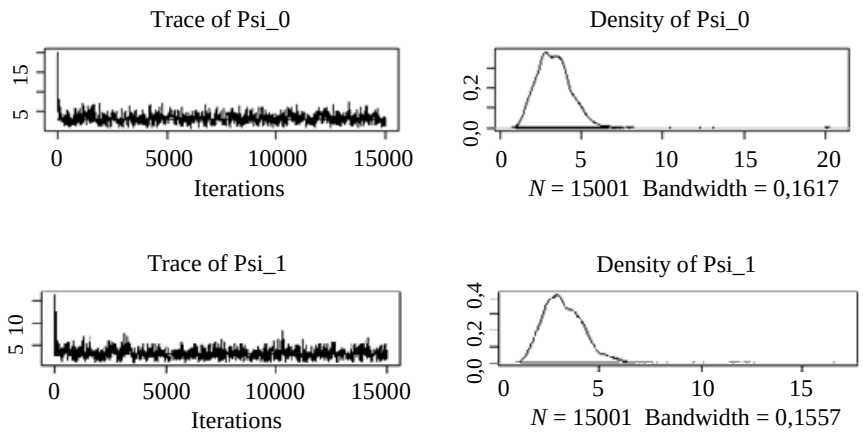
Na kolejnych rysunkach przedstawiono wyniki estymacji parametrów modelu.



Rys. 6. Rozkłady *a posteriori* parametrów – macierz **K**

Źródło: opracowanie własne.

	$E(X)$	$\sigma(X)$	2,5%	97,5%
K_{11}	11,4	2,14	7,77	16,27
K_{22}	8,17	1,62	5,42	11,59
K_{01}	14,5	2,97	8,99	20,78

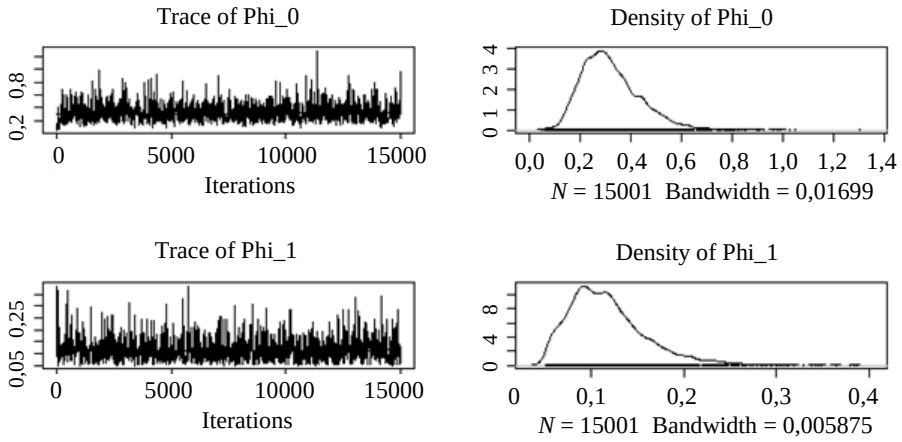


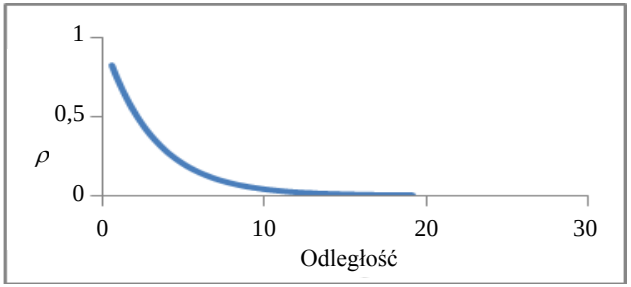
Rys. 7. Rozkłady *a posteriori* parametrów – macierz $\mathbf{\Psi}$

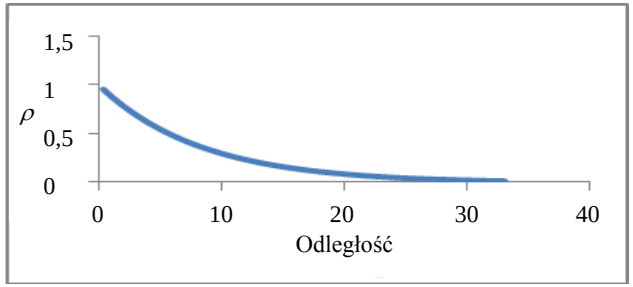
Źródło: opracowanie własne.

	$E(X)$	$\sigma(X)$	2,5%	97,5%
Ψ_{11}	3,35	1,09	1,60	5,84
Ψ_{22}	3,26	1,06	1,65	5,58

	$E(X)$	$\sigma(X)$	2,5%	97,5%
ϕ_1	0,32	0,12	0,15	0,58
ϕ_2	0,12	0,04	0,06	0,22

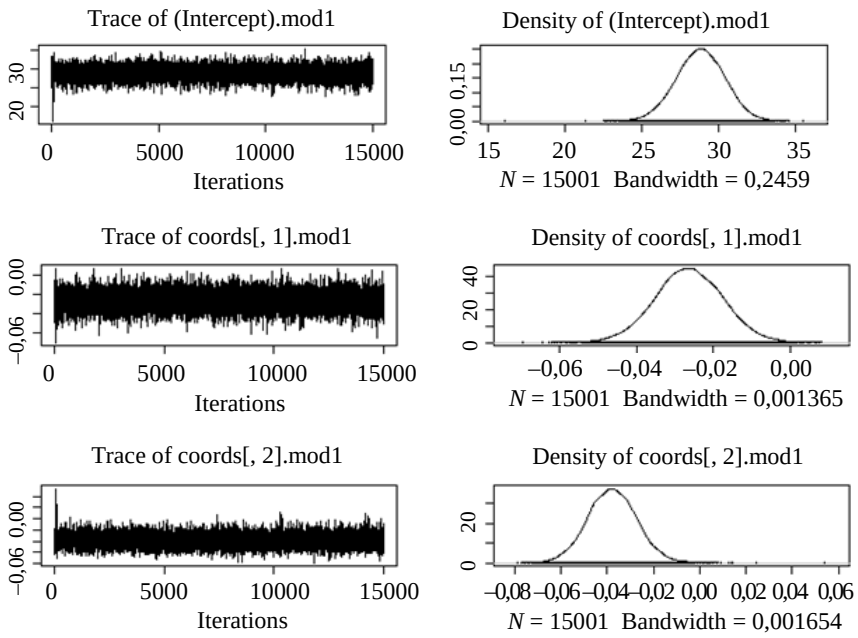






Rys. 8. Rozkłady *a posteriori* parametrów – parametry ϕ

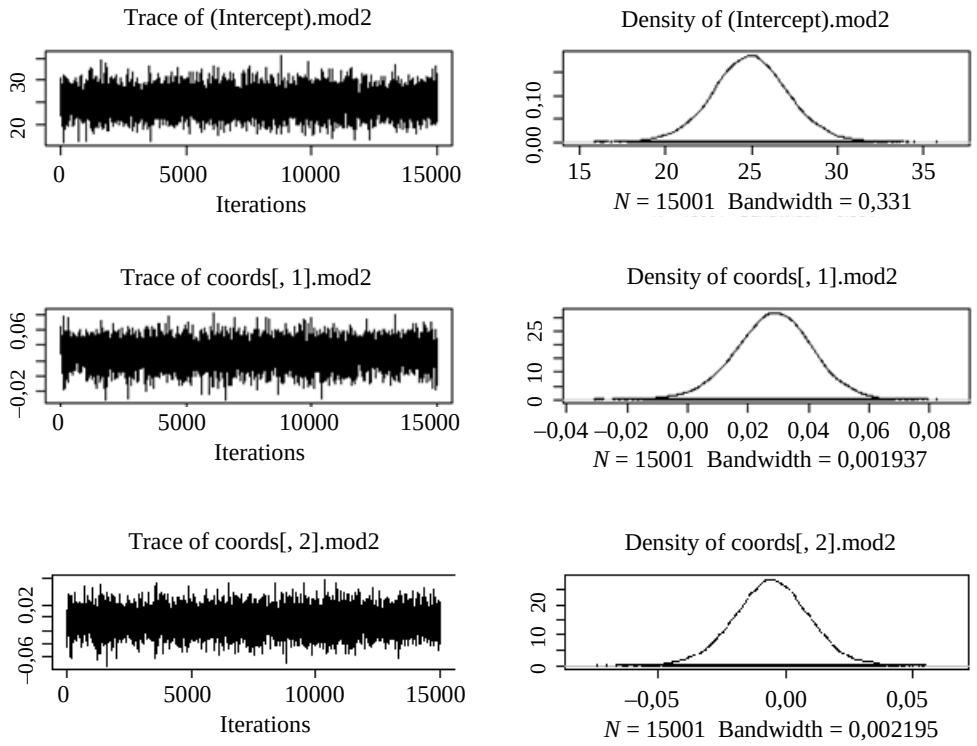
Źródło: opracowanie własne.



Rys. 9. Parametry regresyjne β dla SO_2

Źródło: opracowanie własne.

	$E(X)$	$\sigma(X)$	2,5%	97,5%
β_0	28,82	1,59	25,68	31,89
β_1	–0,03	0,01	–0,04	–0,01
β_2	–0,04	0,01	–0,06	–0,02



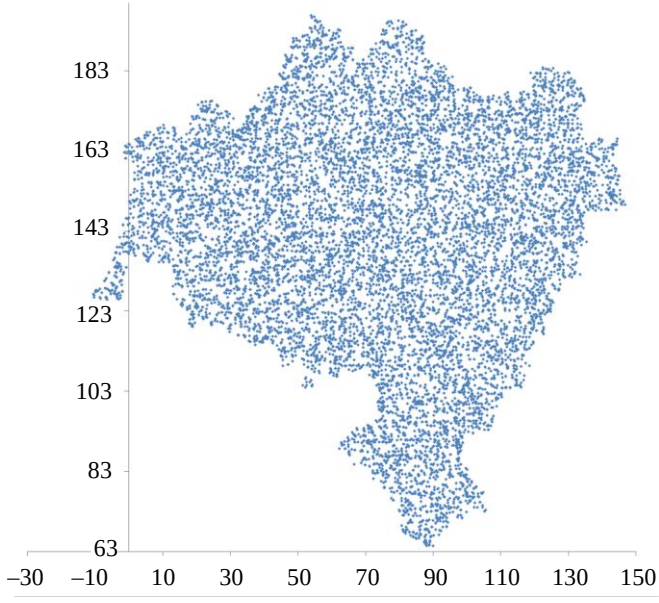
Rys. 10. Parametry regresyjne β dla NO_2

Źródło: opracowanie własne.

	$E(X)$	$\sigma(X)$	2,5%	97,5%
β_0	24,97	2,22	20,60	29,49
β_1	0,03	0,01	0,00	0,05
β_2	-0,01	0,01	-0,03	0,02

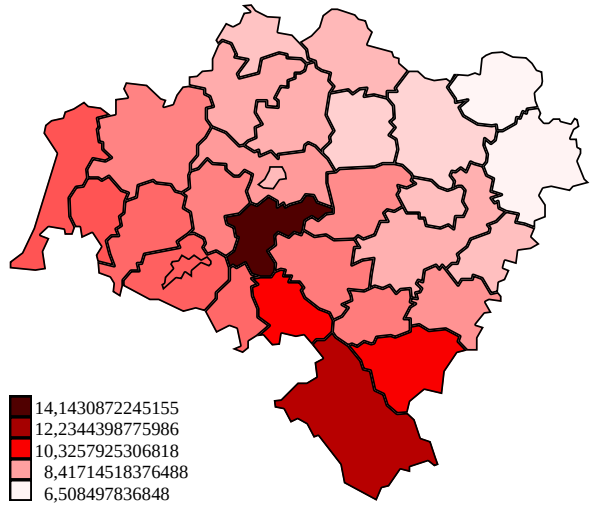
6. Predykcja

Aby dokonać oszacowania średnich poziomów zanieczyszczeń na zadanym obszarze, należy wyznaczyć wartość całki $Y(A_i) = \frac{1}{|A_i|}\int_{\mathbf{s}_0 \in A_i} Y(\mathbf{s}_0)\,d\mathbf{s}_0$. Ponieważ nie jest możliwe jej analityczne obliczenie, dokonane zostanie jej przybliżenie za pomocą wyrażenia $\frac{1}{N_i}\sum_{j=1}^{N_i} \hat{Y}(\mathbf{s}_0^{(j)})$. W związku z tym należy wylosować dostatecznie dużą liczbę punktów na obszarze poszczególnych powiatów, a następnie obliczyć średnią arytmetyczną z oszacowanych w tych punktach poziomów zanieczyszczeń. Na rysunku 11 przedstawiono 10 000 punktów wylosowanych z rozkładu jednostajnego na obszarze województwa.



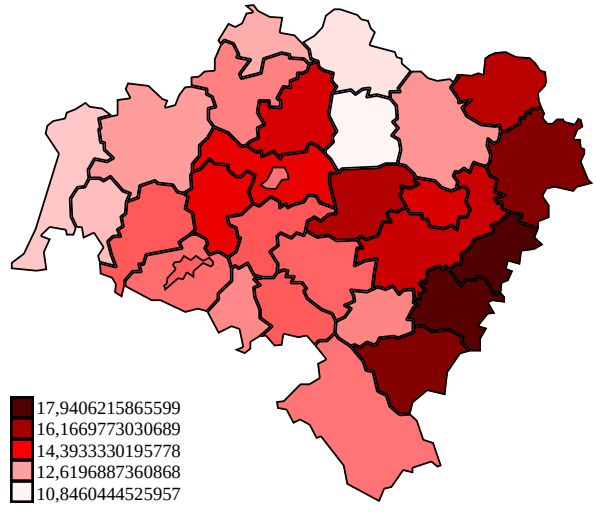
Rys. 11. Losowo wygenerowane punkty prognoz

Źródło: opracowanie własne.



Rys. 12. Średnie poziomy zanieczyszczeń SO_2

Źródło: opracowanie własne.



Rys. 13. Średnie poziomy zanieczyszczeń NO_2

Źródło: opracowanie własne.

Na podstawie opisanego przybliżenia otrzymano średnie poziomy zanieczyszczeń przedstawione na rysunkach 12 i 13. Po ich obliczeniu wyznaczono współczynniki korelacji pomiędzy zachorowalnością na raka a poziomem SO_2 ($\rho = 0{,}16$, *p-value* = 0,2) i NO_2 ($\rho = 0{,}20$, *p-value* = 0,15). W żadnym więc przypadku korelacja nie okazała się istotnie różna od zera i nie można uznać za udowodnioną tezę o występowaniu zależności pomiędzy zachorowalnością na raka a wybranymi typami zanieczyszczeń. Być może korelacja taka wystąpiłaby w przypadku, gdy analizowane byłyby tylko wybrane typy nowotworów (np. płuc). Możliwe, że wyniki zmieniłyby się również, gdyby pomiary zanieczyszczeń były prowadzone w większej liczbie punktów pomiarowych, w związku z czym oszacowania średnich poziomów zanieczyszczeń byłyby bardziej dokładne.

Literatura

[1] Banerjee S., Carlin B.P., Gelfand A.E., *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC 2003.

[2] Banerjee S., Gelfand A.E., *Prediction, Interpolation and Regression for Spatially Misaligned Data*, „The Indian Journal of Statistics" 2002, vol. 62, s. 227-245.

[3] Błaszczyk J., Pudełko M., Cisar K., *Nowotwory złośliwe w województwie dolnośląskim w roku 2003*, Dolnośląski Rejestr Nowotworów, Wrocław 2005.

[4] Cressie N., *Kriging Nonstationary Data*, „Journal of the American Statistical Association" 1986, 81, 625-634.

[5] *Raport o stanie środowiska w województwie dolnośląskim w 2003 roku*, Biblioteka Monitoringu Środowiska, Wrocław 2004.

MEASURING RELATIONSHIPS AMONG PHENOMENA MEASURED ON VARIOUS LEVELS OF SPATIAL AGGREGATION

Summary

The paper describes the problem of spatial data measured at disparate resolutions. One set of data is point-referenced and the other is areal. The problem of reconciliation of these types of data is illustrated with the measurement of the correlation coefficient between environmental pollution and cancer rates.

Dariusz Biskup – dr, adiunkt w Katedrze Statystyki Uniwersytetu Ekonomicznego we Wrocławiu.