III

Cel, zakres i metody pracy

III.1. Cel i zakres pracy

Głównym celem pracy była analiza prawidłowości przestrzennej i czasowej zmienności miesięcznych oraz rocznych maksymalnych dobowych sum opadów (**MSDO**).

Specyfika tych danych polega na tym, że oprócz typowej dla opadów nieciągłości przestrzennej są także niesynchroniczne. Podana wartość miesięcznej MSDO mogła wystąpić w dowolnym stanowisku w którymkolwiek dniu miesiąca. O ile można z uproszczeniem przyjmować, że pole sum miesięcznych, a zwłaszcza rocznych czy wieloletnich, ma charakter ciągły, to w tym przypadku założenie to nie ma uzasadnienia. Mierzone punktowo MSDO nie są jednakże efektem niezależnych, całkowicie losowych zjawisk. Każdy epizod opadowy trwa przez pewien czas i ma określony zasięg przestrzenny. Istnieje zatem niezerowe prawdopodobieństwo, że zostanie zarejestrowany w większej liczbie stanowisk, i w przynajmniej w części z nich będzie zaklasyfikowany jako okresowa MSDO. Założeniem metodologicznym niniejszej pracy była hipoteza (potwierdzona w dalszym toku postępowania), że prawdopodobieństwo wystąpienia MSDO charakteryzuje się autokorelacją przestrzenną, a zatem ciągłością przestrzenną. Logiczną konsekwencją takiego podejścia jest możliwość zastosowania do analizy tych danych metod geostatystycznych (Chilés, Delfiner 1999, Cressie 1993, Goovaerts 1997, Isaaks, Srivastava 1989, Namysłowska-Wilczyńska 2006, Webster, Oliver 2001, Wackernagel 2003, Zawadzki 2005).

Główny cel rozprawy był realizowany poprzez następujące cele szczegółowe:

- określenie głównych cech statystycznych analizowanych zbiorów miesięcznych i rocznych MSDO, a w tym także zmienności sezonowej i wieloletniej;
- identyfikację i typologię struktury przestrzennej miesięcznych oraz rocznych MSDO;

- określenie charakteru i przypuszczalnej genezy zjawisk generujących MSDO;
- ilościową ocenę udziału poszczególnych zjawisk w całkowitych rozmiarach MSDO;
- zbadanie potencjalnej zmienności sezonowej struktury przestrzennej MSDO;
- weryfikację hipotezy o zróżnicowaniu struktury przestrzennej MSDO w różnych klasach wysokości opadów;
- sprawdzenie, czy struktura przestrzenna MSDO wykazuje istotne tendencje w całym analizowanym okresie;
- określenie sezonowego zróżnicowania prawdopodobieństwa wystąpienia i wysokości rocznych MSDO;
- zbadanie zmienności regionalnej terminu występowania rocznych MDSO.

Ten szeroko zakrojony program wymagał wykonania wielu czasochłonnych prac związanych z utworzeniem i weryfikacją źródłowej bazy danych, selekcją i przetwarzaniem numerycznym różnorodnych podzbiorów MSDO, a później analizą oraz interpretacją uzyskanych wyników. Najważniejsze z nich to:

- analiza statystyczna zmian ilości i rozmieszczenia przestrzennego sieci punktów danych w analizowanym wieloleciu;
- globalna i lokalna statystyczna charakterystyka różnorodnych podzbiorów bazy danych MSDO;
- budowanie modeli wielomianowych potrzebnych do normalizacji każdego z 325 podzbiorów miesięcznych i rocznych MSDO;
- binarne kodowanie każdego podzbioru według 13 wartości progowych, określonych na podstawie empirycznych funkcji skumulowanego rozkładu;
- maskowanie, w każdym analizowanym podzbiorze, wartości odstających zaburzających obraz struktury przestrzennej;
- obliczanie 325 semiwariogramów empirycznych danych normalizowanych i 4225 danych kodowanych;
- interaktywne modelowanie matematyczne 4550 semiwariogramów,

- klasyfikacja uzyskanych modeli struktury przestrzennej; testowanie istotności ich zróżnicowania w zależności od sezonu, względnej i bezwzględnej wysokości opadów oraz miejsca w wieloletnim ciągu pomiarowym;
- optymalizacja algorytmu interpolacji prawdopodobieństwa wystąpienia i wysokości rocznych MSDO;
- interpolacja statystyk terminów wystąpienia rocznych MSDO.

Podstawy zastosowanej metodyki przedstawiono w kolejnym podrozdziale (III.2). Szczegółowe, specyficzne dla analizowanego zbioru danych, zagadnienia metodyczne opisano w dodatkach (rozdział X). Sposoby rozwiązania szeregu drobniejszych problemów metodycznych omawiane są także w rozdziałach analitycznych, wraz z prezentacją uzyskanych wyników.

III.2. Podstawy metodyki

III.2.1. Wprowadzenie

Podstawę metodyki niniejszej analizy zmienności przestrzennej maksymalnych sum dobowych opadów na obszarze Polski stanowi geostatystyka. Ponieważ w ostatnich latach ukazały się drukiem dwa obszerne opracowania, które po polsku prezentowały teorię i zastosowania tej dziedziny statystyki przestrzennej (Namysłowska-Wilczyńska 2006, Zawadzki 2005), autor czuje się zwolniony z obowiązku szczegółowego omówienia jej podstaw. Zostaną one przedstawione jedynie w takim zakresie, jaki jest niezbędny dla uzasadnienia zastosowania oraz wyjaśnienia teorii i algorytmów konkretnych, wykorzystanych w niniejszej pracy technik geostatystycznych. Bardziej dokładnie omówione zostaną te zagadnienia, które w cytowanych wyżej pracach były pominięte, albo potraktowane skrótowo, a mają istotne znaczenie dla zrozumienia uzyskanych przez autora wyników. Niezbędne jest przy tym używanie zwięzłej i jednoznacznej notacji matematycznej. We wszystkich takich przypadkach używano konwencji wprowadzonej w serii podręczników wydawanych przez Oxford University Press pod wspólną nazwą „ *Applied Geostatistics Series*” (Deutsch, Journel 1992, 1998, Goovaets 1997).

III.2.2. Miary ciągłości lub zmienności przestrzennej

Niech $z(\mathbf{u}_\alpha)$, α=1, 2, …, n określa zbiór n wartości pomiarów dowolnej cechy z (wielkości) dokonanych w obrębie badanego obszaru, gdzie $\mathbf{u}_\alpha$ oznacza wektor współrzędnych konkretnej obserwacji α. W większości przypadków nie są to wartości kompletnie losowe, co oznacza, że pomiary wykonane bliżej są zazwyczaj do siebie bardziej podobne od tych, które dzieli większa odległość. Podobieństwo to można określić ilościowo, porównując wyniki pomiarów dla stanowisk odległych od siebie o coraz większe odległości. Symbolicznie określane to jest jako porównanie dowolnej danej z określonej w lokalizacji $\mathbf{u}_\alpha$ czyli $z(\mathbf{u}_\alpha)$, z dowolną inną odległą o wektor $\mathbf{h}$, czyli $z(\mathbf{u}_\alpha + \mathbf{h})$. W przypadku każdego typu próbkowania poza regularnym, wartość wektora $\mathbf{h}$ oznacza w rzeczywistości pewien przedział odległości[7]. Na przykład, porównujemy pomiary odległe od siebie 0-1 km, 1-2 km, 2-3 km i tak dalej[8]. Najprościej można tego dokonać używając wykresu kartezjańskiego XY. Zazwyczaj za pomocą tego wykresu przedstawia się relacje między dwoma cechami (parametrami) zmierzonymi w tej samej lokalizacji i/lub w tym samym czasie. Tym razem służy on do porównania wartości tej samej cechy zmierzonej w dwóch różnych lokalizacjach i w związku z tym na osi X odkłada się wartości $z(\mathbf{u}_\alpha)$, a na Y – $z(\mathbf{u}_\alpha + \mathbf{h})$. Wykres taki nazywany jest „**h**-*scattergram*”, co można przetłumaczyć na „wykres rozrzutu z przesunięciem **h**” (Goovaerts 1997, Pannatier 1996, Zawadzki 2005). W geostatystyce przyjęto określenie „ogona” (ang. *tail value*) dla wartości będącej początkiem wektora $\mathbf{h}$, czyli $z(\mathbf{u}_\alpha)$, podczas gdy wartość stanowiąca jego koniec, czyli $z(\mathbf{u}_\alpha + \mathbf{h})$, nazywana jest „głową” (ang. *head value*). Jako przykład zaprezentowano wykresy rozrzutu z przesunięciem dla pierwszych sześciu odstępów, o szerokości 2,5 km każdy, zbioru danych MSDO z lipca roku 1977 (ryc. 3). Pokazują on wyraźnie, wyczuwaną intuicyjnie, właściwość spadku podobieństwa wyników pomiarów wraz z odległością. Na kolejnych wykresach chmura punktów staje się coraz „szersza’ – są one bardziej oddalone on przekątnej symbolizującej idealną zależność wprost proporcjonalną. Tradycyjnie do wyrażenie siły relacji między dwoma zmiennymi używa się współczynnika korelacji liniowej Pearsona, oznaczanego symbolem ρ. Jest on standaryzowaną (niezależną od skali pomiarowych) formą kowariancji

[7] Przy analizie kierunkowej (anizotropowej) również kierunek wektora ma charakter przedziałowy.
[8] Żeby wyraźnie zaznaczyć charakter wektora **h**, który nie określa jednej konkretnej odległości między konkretną parą lokalizacji, ale ogólną różnicę położenia w przestrzeni dowolnej pary danych, w geostatystyce nazwa się go *lag*, a nie *distance*. Najbardziej trafnym polskim odpowiednikiem tego pojęcia jest „odstęp”, ewentualnie „przesunięcie”. W niniejszej rozprawie stosowane będzie najczęściej to pierwsze określenie.

obu zmiennych. W tym przypadku można go wyrazić w sposób następujący (Goovaerts 1997):

$$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{\sqrt{\sigma^2_{\text{-}\mathbf{h}} \cdot \sigma^2_{\text{+}\mathbf{h}}}} \qquad \in [-1, +1] \qquad [\mathbf{1}]$$

gdzie:

$C(\mathbf{h})$ oznacza autokowariancję par pomiarów oddalonych od siebie o **h** (wzór [2]),

natomiast $\sigma^2_{\text{-}\mathbf{h}}$ i $\sigma^2_{\text{+}\mathbf{h}}$ – odpowiednio wariancję podzbioru danych ogona i głowy (wzory [3] i [4]):

$$C(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[z(\mathbf{u}_\alpha) \cdot z(\mathbf{u}_\alpha + \mathbf{h}) - m_{\text{-}\mathbf{h}} \cdot m_{\text{+}\mathbf{h}} \right] \qquad [\mathbf{2}]$$

$$\sigma^2_{\text{-}\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[z(\mathbf{u}_\alpha) - m_{\text{-}\mathbf{h}} \right]^2 \qquad [\mathbf{3}]$$

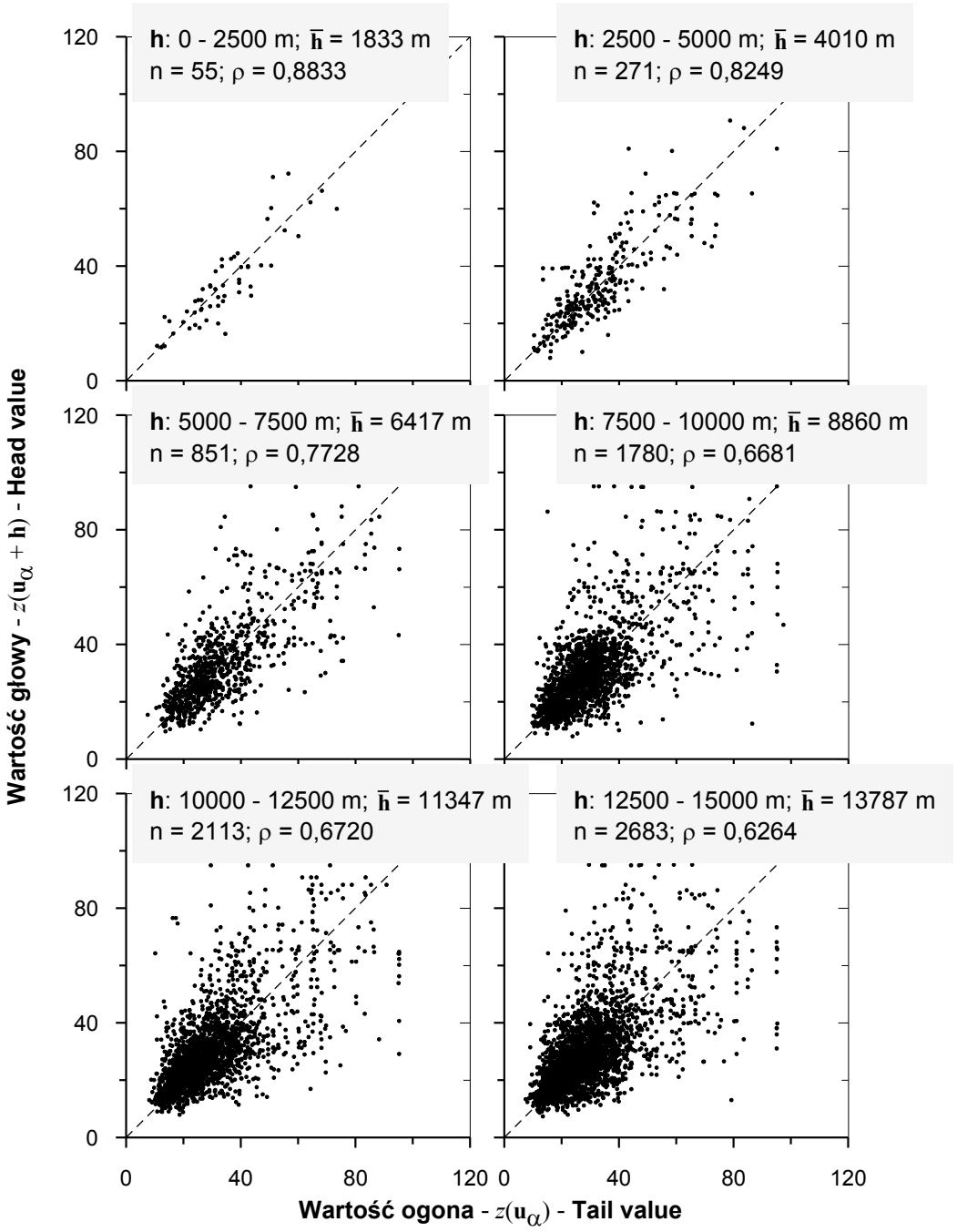
$$\sigma^2_{\text{+}\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[z(\mathbf{u}_\alpha + \mathbf{h}) - m_{\text{+}\mathbf{h}} \right]^2 \qquad [\mathbf{4}]$$

$N(\mathbf{h})$ oznacza liczbę par danych w obrębie określonej klasy odległości i kierunku, a symbolami $m_{\text{-}\mathbf{h}}$ i $m_{\text{+}\mathbf{h}}$ we wzorze autokowariancji [2] i wzorach wariancji podzbiorów ogona [3] i głowy [4] określono odpowiednie średnie arytmetyczne obu tych podzbiorów [5] i [6]:

$$m_{\text{-}\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} z(\mathbf{u}_\alpha) \qquad [\mathbf{5}]$$

$$m_{\text{+}\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} z(\mathbf{u}_\alpha + \mathbf{h}) \qquad [\mathbf{6}]$$

Na wykresie (ryc. 3) współczynnik korelacji (autokorelacji) maleje od 0,8833 dla par punktów należących do pierwszego odstępu (0-2,5 km) do 0,6264 dla szóstego odstępu (pary odległe o 12,5-15,0 km). Obliczone, dla kolejnych rosnących odstępów, współczynniki autokorelacji, przedstawione w relacji do średniej odległości par należących do danego odstępu, tworzą tak zwany korelogramem eksperymentalny. Rycina 4 jest przykładem takiego wykresu dla danych MSDO z lipca 1977 roku. Wynika z niego, że oprócz kilku niewielkich wahnięć, spadek autokorelacji następuje konsekwentnie do odległości około 150 km. Na takim dystansie nie ma już statystycznie żadnego podobieństwa wyników pomiarów – współczynnik korelacji oscyluje wokół zera.



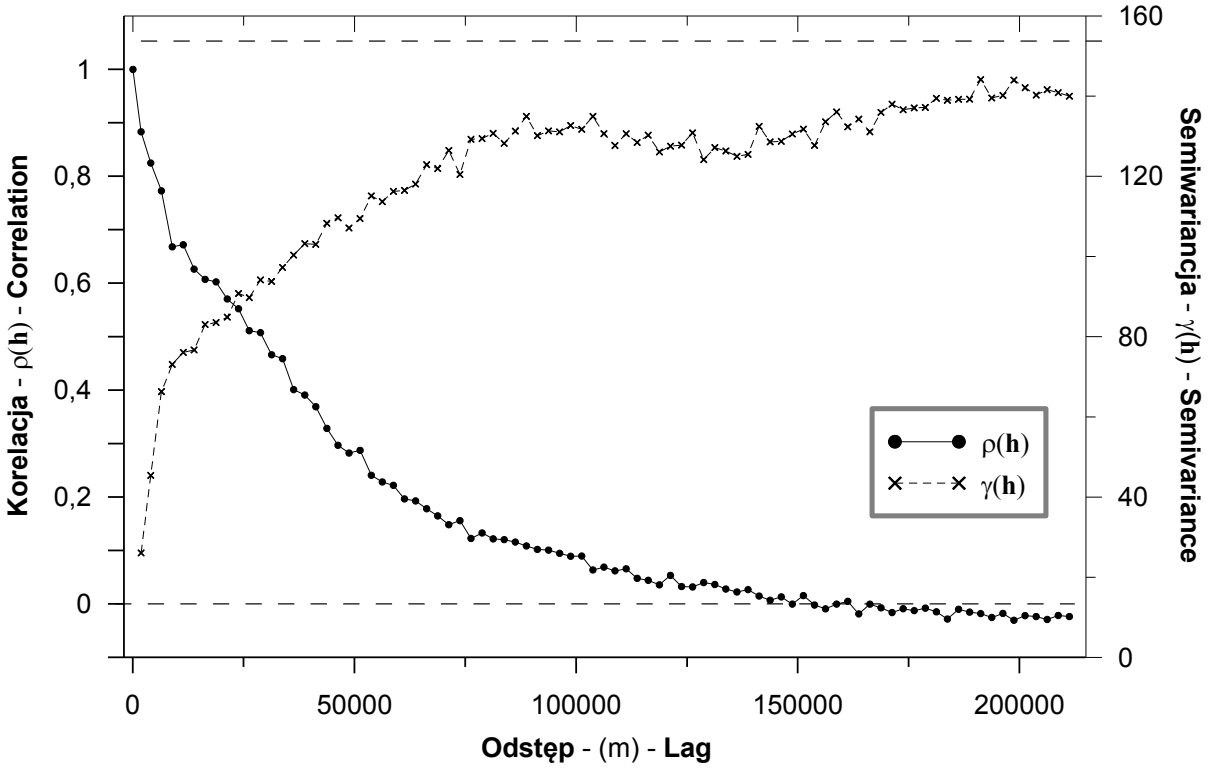
Ryc. 3. Wykresy rozrzutu z przesunięciem (ang. **h**-*scattegram*) pierwszych 6 odstępów, o szerokości 2500 m każdy, dla danych MSDO z lipca roku 1977. Przy wykresach podano zakres odległości, średni odstęp dla wszystkich par danych znajdujących się w danym przedziale, ilość par i współczynnik korelacji liniowej pomiędzy wynikami pomiarów. Należy zwrócić uwagę na konsekwentny spadek korelacji MSDO wraz ze wzrostem odstępu między stanowiskami.

Alternatywnym, i silniej ugruntowanym w literaturze geostatystycznej, sposobem przedstawiania relacji przestrzennych jest kategoria niepodobieństwa (zamiast podobieństwa) pomiędzy obserwacjami, jako funkcji dzielącej je odległości. Miarą średniego niepodobieństwa jest semiwariancja, zdefiniowana jako połowa średniej kwadratów różnic

wartości cechy w lokalizacjach odległych o wektor **h** (Gringarten, Deutsch 2001, Goovaerts 1997, Zawadzki 2005):

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h}) \right]^2 \qquad [7]$$

Geometryczną interpretację semiwariancji stanowi odległość dzieląca każdą naniesioną na wykres rozrzutu z przesunięciem parę pomiarów oddalonych o wektor **h** od przekątnej wykresu (pierwszego bisektora) symbolizującej idealną relację, wprost proporcjonalną między porównywanymi cechami (Goovaerts 1997, ryc. 5).

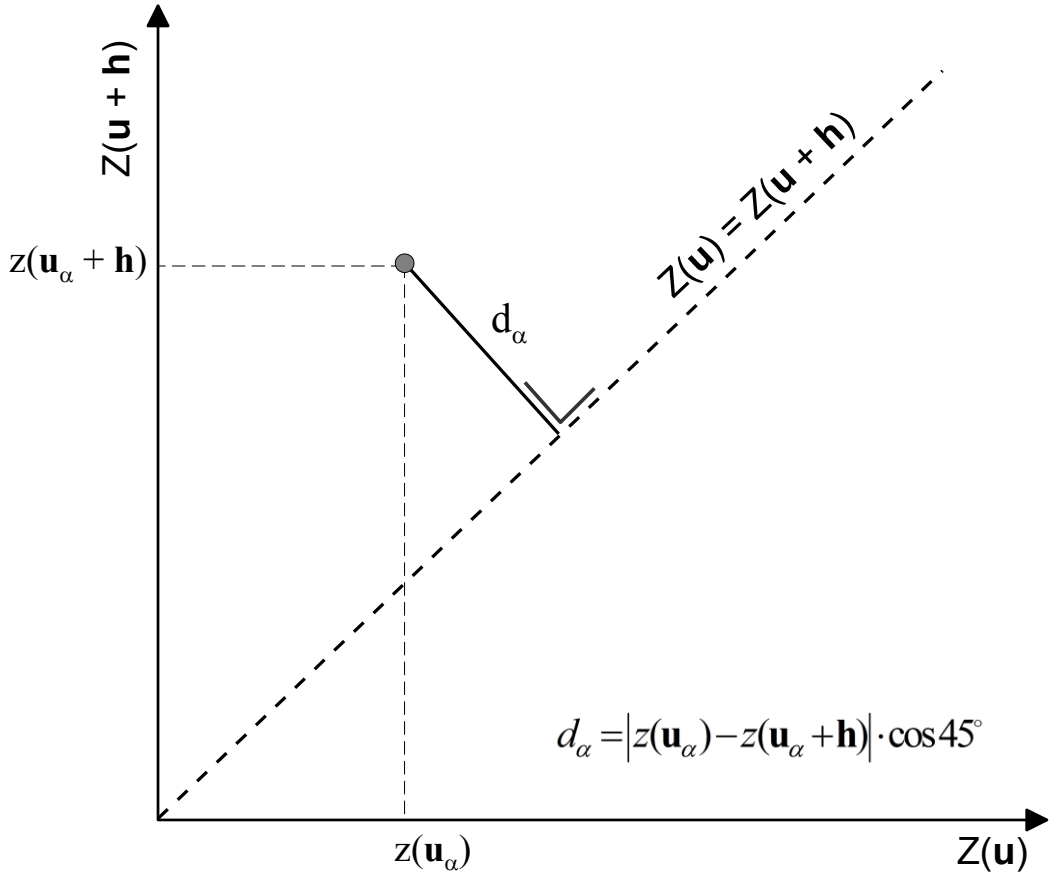


Ryc. 4. Korelogram i semiwariogram danych MSDO z lipca 1977 obliczony dla 85 przedziałów odległości po 2500 m. Na wykresie zaznaczono zerowy poziom autokorelacji i wariancję próby wynoszącą 153,72, a oznaczanej w geostatystyce jak $C(0)$ – autokowariancja dla odstępu równego 0.

W geostatystyce wynik pomiaru $z(\mathbf{u}_\alpha)$ jest interpretowany jako jedna z możliwych realizacji funkcji losowej $FL(\mathbf{u}_\alpha)$, która jest w pełni charakteryzowana przez jej rozkład prawdopodobieństwa $F(\mathbf{u}_\alpha;z) = \text{Prob}\{Z(\mathbf{u}_\alpha) \leq z\}$. Dla stacjonarnej FL[9] istnieje funkcyjna

[9] Model funkcji losowej spełnia założenie stacjonarności jeśli: (1) wartość oczekiwana $E\{Z(\mathbf{u})\}$ istnieje i nie zależy od lokalizacji w obrębie analizowanego obszaru, (2) dla każdej pary zmiennych losowych $\{Z(\mathbf{u}), Z(\mathbf{u} + \mathbf{h})\}$ istnieje ich kowariancja, która jest zależna jedynie od wektora odstępu **h**.

relacja pomiędzy semiwariancją a poprzednio zdefiniowanymi ([1] i [2]) autokorelacją i autokowariancją (Goovaerts 1997):



Ryc. 5. Interpretacja geometryczna wartości semiwariancji γ(**h**), jako średniej wszystkich podniesionych do kwadratu ortogonalnych odległości d_α od przekątnej wykresu (pierwszego bisektora) rozrzutu z przesunięciem (Goovaerts 1997).

$$2\gamma(\mathbf{h}) = \mathrm{Var}\{Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})\}$$
$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}), \quad \forall \mathbf{u} \qquad [\mathbf{8}]$$

oraz,

$$\rho(\mathbf{h}) = 1 - \frac{\gamma(\mathbf{h})}{C(0)} \qquad [\mathbf{9}]$$

gdzie $C(0)$ oznacza wariancję próby.

Na wykresie (ryc. 4) semiwariogram danych MSDO z lipca 1977 roku rośnie, a jego obraz jest do pewnego stopnia lustrzanym odbiciem przebiegu korelogramu. Są jednak wyraźne różnice. Wykres osiąga maksimum (staje się płaski) na dystansie około 80 km, a

dalsze zmiany mają charakter raczej oscylacji niż konsekwentnego wzrostu. Odległość, na której zachodzi stabilizacja semiwariancji nazywana jest zasięgiem (ang. *range*), a jej wartość – semiwariancją progową (ang. *sill*). Wzrost niepodobieństwa reprezentowany przez semiwariogram nie zachodzi jednak w stałym tempie. Wykres wykazuje co najmniej dwa załamania: jedno na dystansie około 9 km, drugie, słabiej widoczne, przy odstępie około 55 km. Określa się je jako zasięgi cząstkowe (ang. *partial ranges*), a odpowiadające im wartości semiwariancji jako progi cząstkowe (ang. *partial sills*). Świadczą one, że analizowana cecha (parametr) jest efektem działania kilku zjawisk operujących w różnych skalach przestrzennych (Goovaerts 1997).

Wraz ze wzrostem odstępu $|\mathbf{h}|$ korelacja między jakimikolwiek dwoma zmiennymi losowymi $Z(\mathbf{u})$ i $Z(\mathbf{u} + \mathbf{h})$ zazwyczaj dąży do zera:

$$C(h) \to 0 \quad dla \quad |\mathbf{h}| \to \infty \qquad [\mathbf{10}]$$

Biorąc pod uwagę zależność [8], wariancja progowa (ang. *sill*) semiwariogramu ograniczonego dąży do wariancji $C(0)$:

$$\gamma(\mathbf{h}) \to C(0) \quad dla \quad |\mathbf{h}| \to \infty \qquad [\mathbf{11}]$$

Nieciągłość na początku semiwariogramu (to jest dla odstępu równego zero) nazywana jest efektem nuggetowym[10] (Goovaerts 1997, Gringarten, Deutsch 2001, Chilés, Delfiner 1999). Jego źródłem są błędy pomiarowe i/lub zmienność przestrzenna w skali mniejszej niż najkrótszy odstęp pomiarów, uwzględniając w tym także występującą w obrębie próbki[11].

III.2.3. Struktura przestrzenna w klasach natężenia analizowanej cechy

Funkcja autokowariancji i semiwariogram to charakterystyki ciągłości przestrzennej (lub zmienności) dla całego zakresu wartości cechy. Struktura ciągłości przestrzennej może jednak różnić się, zależnie czy pod uwagę bierzemy rozkład punktów danych o niskich, średnich czy wysokich wartościach (Deutsch, Journel 1998, Goovaerts 1997). W wielu sytuacjach spotykanych w środowisku przyrodniczym lub społeczno-gospodarczym, losowo występujące wysokie wyniki pomiarów, są otoczone większymi obszarami o średnich lub niskich

[10] Źródłem tej terminologii były analizy zasobów złóż złota w RPA, dokonywane w latach pięćdziesiątych XX wieku przez D. Krige'a (1951, 1952). Głównym powodem występowania nieciągłości było tam losowe występowanie samorodków złota (ang. *nugget*). W związku z tym w polskiej literaturze geostatystycznej używa się często terminu „efekt samorodka" (Zawadzki 2005, Namysłowska-Wilczyńska 2006).
[11] W tym przypadku zmienność w obrębie powierzchni zbiorczej deszczomierza, czyli w skali około 16 cm.

wartościach, które zmieniają się w sposób ciągły i stopniowy. Czy wartości ekstremalne są w przestrzeni rozproszone czy skupione, jaki jest ich zasięg ma duże znaczenie dla wyjaśniania zjawiska, oraz jakości estymacji.

Określenie prawidłowości rozkładu przestrzennego wartości cechy z występujących powyżej lub poniżej ustalonego poziomu progowego z_k wymaga uprzedniego przekodowania każdego wyniku pomiaru $z(\mathbf{u}_\alpha)$ do formy binarnej zgodnie z poniższą regułą:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1 & jeżeli\ z(\mathbf{u}_\alpha) \le z_k \\ 0 & poza\ tym \end{cases} \qquad [\mathbf{12}]$$

Dane kodowane $i(\mathbf{u}_\alpha; z_k)$ mogą dalej być analizowane przy użyciu każdej z wymienionych poprzednio miar struktury przestrzennej. Po odpowiednim zmodyfikowaniu wzoru [2] oblicza się na jego podstawie eksperymentalną autokowariancję kodów:

$$\begin{aligned} C_I(\mathbf{h}; z_k) &= \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} i(\mathbf{u}_\alpha; z_k) \cdot i(\mathbf{u}_\alpha + \mathbf{h}; z_k) - F_{\text{-}\mathbf{h}}(z_k) \cdot F_{+\mathbf{h}}(z_k) \\ &= F(\mathbf{h}; z_k) - F_{\text{-}\mathbf{h}}(z_k) \cdot F_{+\mathbf{h}}(z_k) \end{aligned} \qquad [\mathbf{13}]$$

gdzie:

$$F_{\text{-}\mathbf{h}}(z_k) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} i(\mathbf{u}_\alpha; z_k) \qquad F_{+\mathbf{h}}(z_k) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} i(\mathbf{u}_\alpha + \mathbf{h}; z_k)$$

$F_{\text{-}\mathbf{h}}(z_k)$ i $F_{+\mathbf{h}}(z_k)$ oznaczają proporcje (ułamek) wartości ogona i głowy nie przekraczających poziomu wartości progowej z_k. Autokowariancja kodów określa, jak często dwie wartości tej samej cechy oddalone od siebie o wektor **h** są jednocześnie nie większe od wartości progowej z_k.

Autokowariancja kodów podzielona przez pierwiastek z iloczynu wariancji podzbiorów ogona i głowy przybiera postać standaryzowaną – eksperymentalną autokorelację kodów:

$$\rho_I(\mathbf{h}; z_k) = \frac{C_I(\mathbf{h}; z_k)}{\sqrt{\sigma^2_{\text{-}\mathbf{h}}(z_k) \cdot \sigma^2_{+\mathbf{h}}(z_k)}} \in [-1, +1] \qquad [\mathbf{14}]$$

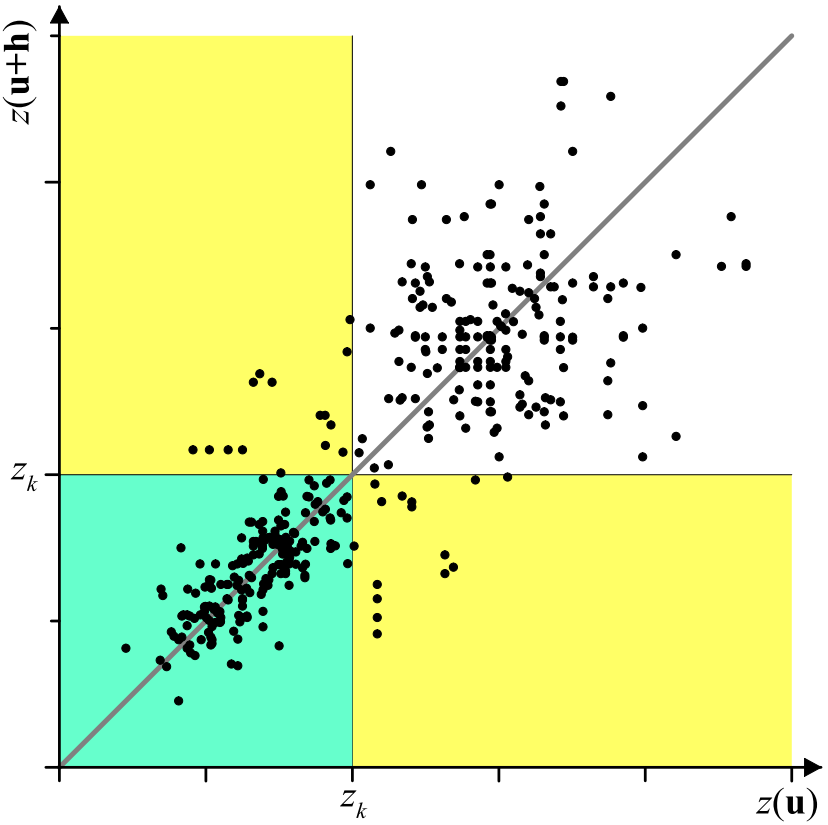
gdzie:

$\sigma^2_{\text{-}\mathbf{h}}(z_k) = F_{\text{-}\mathbf{h}}(z_k)\left[1 - F_{\text{-}\mathbf{h}}(z_k)\right]$ oznacza wariancję wartości kodów ogona, a

$\sigma^2_{+\mathbf{h}}(z_k) = F_{+\mathbf{h}}(z_k)\left[1 - F_{+\mathbf{h}}(z_k)\right]$ – wariancję wartości kodów głowy.

Analogicznie eksperymentalna semiwariancja kodów obliczana jest na podstawie wzoru [7] po podstawieniu zamiast wartości zmierzonej $z(\mathbf{u}_\alpha)$ jej binarnej transformacji $i(\mathbf{u}_\alpha; z_k)$:

$$\gamma_I(\mathbf{h}; z_k) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[i(\mathbf{u}_\alpha; z_k) - i(\mathbf{u}_\alpha + \mathbf{h}; z_k) \right]^2 \qquad [15]$$



Ryc. 6. Autokowariancję i semiwariancję danych kodowanych można interpretować jako proporcję punktów (par danych), które występują w określonych częściach wykresu rozrzutu z przesunięciem: autokowariancja – obszar zaznaczony kolorem żółtym, semiwariancja – obszar zaznaczony kolorem niebieskim (Goovaerts 1997). Do wykonania ryciny wykorzystano dane maksymalnej rocznej sumy dobowej opadu z dnia 18 lipca roku 1970. Na wykresie znajdują się wartości dla 392 par stanowisk odległych od siebie o 1,5-4,5 km. Wartość progową z_k stanowi suma dobowa równa 100 mm.

Semiwariancja kodów ($2\gamma_I(\mathbf{h}; z_k)$) określa, jak często dwie wartości analizowanej cechy oddalone o wektor **h** znajdują się po przeciwnych stronach wartości progowej z_k[12]. Innymi słowy $2\gamma_I(\mathbf{h}; z_k)$ odzwierciedla częstość przejść między dwoma klasami wartości cechy jako funkcję odległości (**h**). Im jest większy, tym mniejszą ciągłość przestrzenną wykazują niskie lub wysokie wartości.

[12] Semiwariancja wartości kodowanych (patrz wzór 15) jest tylko wtedy niezerowa kiedy $i(\mathbf{u}_\alpha; z_k)$ równa się 0, a $i(\mathbf{u}_\alpha + \mathbf{h}; z_k)$ równa się 1, lub odwrotnie.

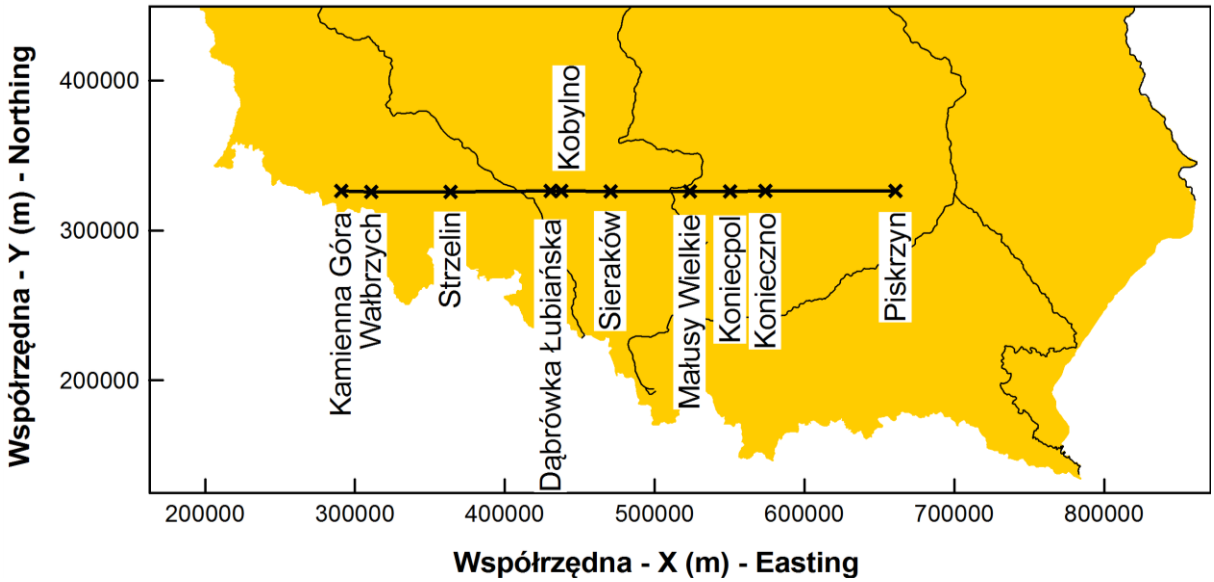
Kowariancja i semiwariancja danych kodowanych może być przedstawiona graficznie jako proporcja punktów (par danych) znajdujących się w określonych częściach wykresu rozrzutu z przesunięciem (ryc. 6, Goovaerts 1997). Autokowariancja kodów $C_I(\mathbf{h}; z_k)$ odzwierciedla udział par danych ($z(\mathbf{u}_\alpha)$, $z(\mathbf{u}_\alpha+\mathbf{h})$), które jednocześnie nie przekraczają wartości progowej z_k (ryc. 6: obszar zaznaczony kolorem żółtym). W obliczonej semiwariancji kodów $2\gamma_I(\mathbf{h}; z_k)$ udział mają jedynie te pary $z(\mathbf{u}_\alpha)$ i $z(\mathbf{u}_\alpha+\mathbf{h})$, które znajdują się po przeciwnych stronach ustalonej wartości progowej z_k. Stanowi go zatem ułamek całego zbioru par dla danego odstępu, zlokalizowany na rycinie 6 w obszarze zaznaczonym kolorem niebieskim.

III.2.4. Przykład analizy danych kodowanych

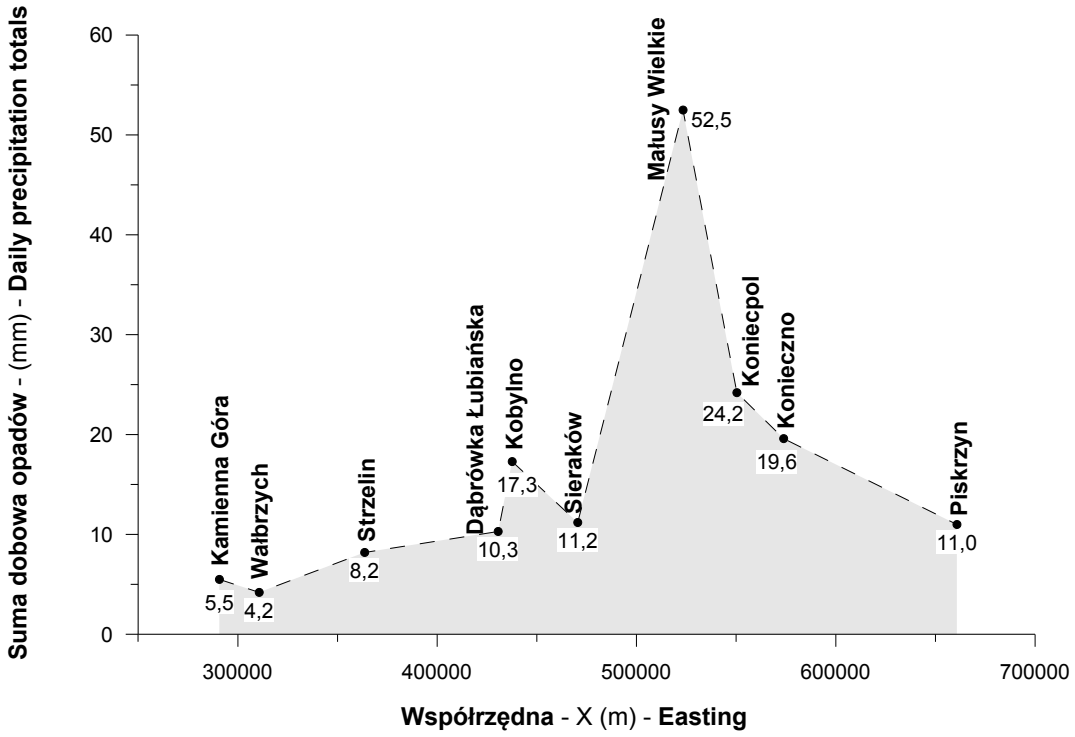
W celu prostego i obrazowego zilustrowania stosowanej w niniejszej rozprawie, raczej mało znanej i wykorzystywanej, metodyki analizy danych kodowanych posłużono się specjalnie do tego celu przygotowanym, jednowymiarowym przykładem (ryc. 7 i 8). Ze całego zbioru maksymalnych opadów dobowych zarejestrowanych na terytorium Polski w maju 1980 roku wybrano 10 stanowisk leżących w przybliżeniu na jednej rzędnej Y w układzie współrzędnych 1992 (N = 326 000 ± 500 m). Profil o długości 370 km rozpoczyna się na zachodzie stanowiskiem Kamienna Góra w Sudetach Środkowych, a kończy na wschodzie w Piskorzynie na Wyżynie Sandomierskiej. Największe wysokości terenu są na obu końcach profilu: Kamienna Góra – 420, Wałbrzych – 490 oraz Małusy Wielkie – 280 i Piskorzyn – 300 m n.p.m. Najniżej usytuowane są stanowiska na Równinie Wrocławskiej (Strzelin, 165 m n.p.m.) oraz Równinie Opolskiej (Dąbrówka Łubiańska, Kobylno, 165 i 175 m n.p.m.). Omawiane dane zostały wybrane głównie ze względu na relatywnie duże zróżnicowanie zmierzonych w maju 1980 maksymalnych opadów dobowych (ryc. 8). Najniższe opady zarejestrowano w dwóch stanowiskach leżących na zachodnim krańcu profilu (5,5 i 4,2 mm). W kolejnych punktach pomiarowych usytuowanych dalej w kierunku wschodnim zanotowano coraz wyższe sumy dobowe, z maksimum w Małusach Wielkich – 52,5 mm. Dalej ku wschodowi opady konsekwentnie malały do wartości 11,0 mm w Piskorzynie. Na wybranym profilu duże jest także zróżnicowanie odległości sąsiadujących stanowisk – od 7 km pomiędzy Kobylnem a Dąbrówką Łubiańską do 87 km pomiędzy Piskorzynem a Koniecznem.

Procedura postępowania jest następująca:

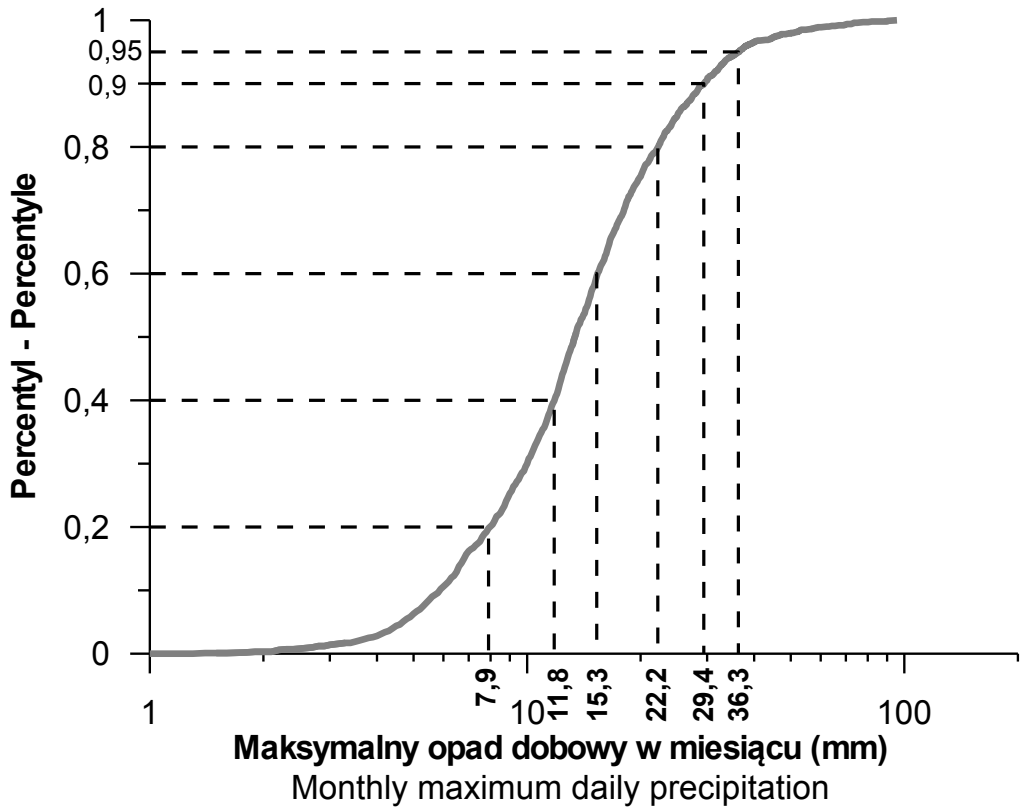
1) **konstrukcja i dyskredytyzacja globalnej skumulowanej funkcji rozkładu** (ang. *cumulative distribution function* = *cdf*, ryc. 9). Na podstawie dostępnych danych pomiarowych (próby) obliczana jest i wykreślana krzywa *cdf*. W przedstawianym poniżej przykładzie posłużono się wszystkimi wynikami pomiarów maksymalnych sum dobowych opadów na terenie Polski w maju 1980. Następnie wybiera się wartości progowe służące do dyskredytyzacji *cdf*. Powinno ich być jak najmniej (ze względu na czasochłonność analizy), ale dostatecznie dużo, żeby uchwycić najbardziej charakterystyczne cechy rozkładu. Szczególną uwagę zwracać zazwyczaj trzeba na skrajne części krzywej, obrazujące częstość występowania wartości ekstremalnych. Z drugiej strony, wybranie bardzo skrajnych wartości progowych, na przykład 0,01 lub 0,99, przy małym zbiorze danych pomiarowych pociąga za sobą ryzyko dużych problemów z określeniem wiarygodnego modelu struktury przestrzennej. W prezentowanym jednowymiarowym przykładzie dla uproszczenia wybrano jedynie 6 wartości progowych odpowiadających 20, 40, 60, 80, 90 i 95 percentylowi rozkładu MSDO na obszarze Polski w maju 1980 roku (ryc. 9);



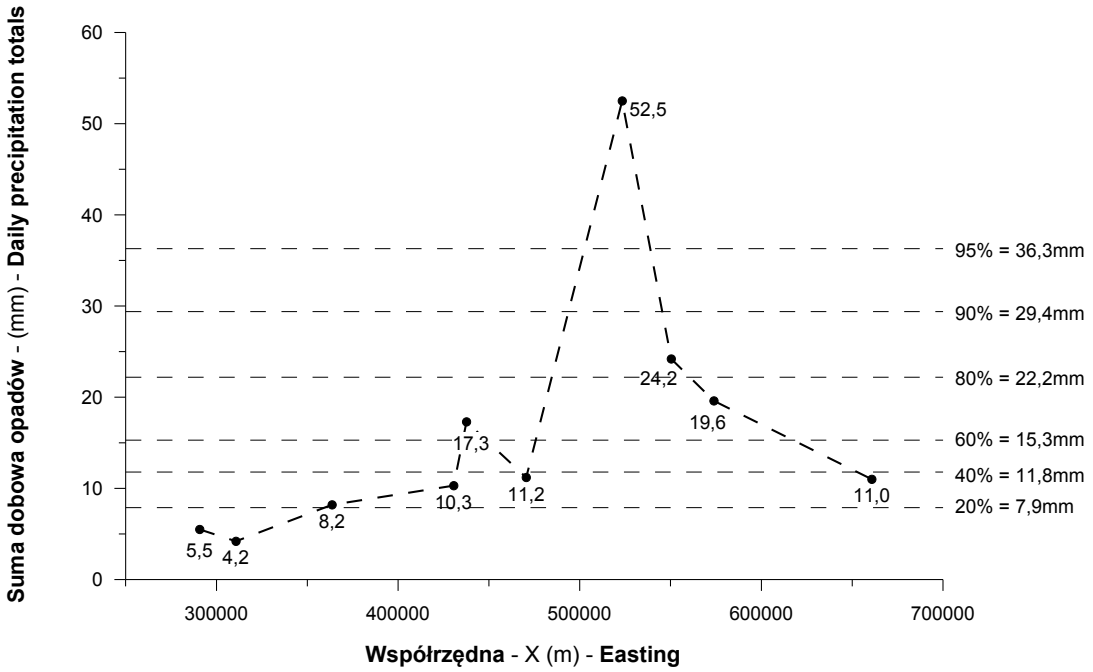
Ryc. 7. Profil posterunków opadowych (dane jednowymiarowe) funkcjonujących w maju 1980 roku wykorzystany do ilustracji metodyki geostatystycznej analizy danych kodowanych.



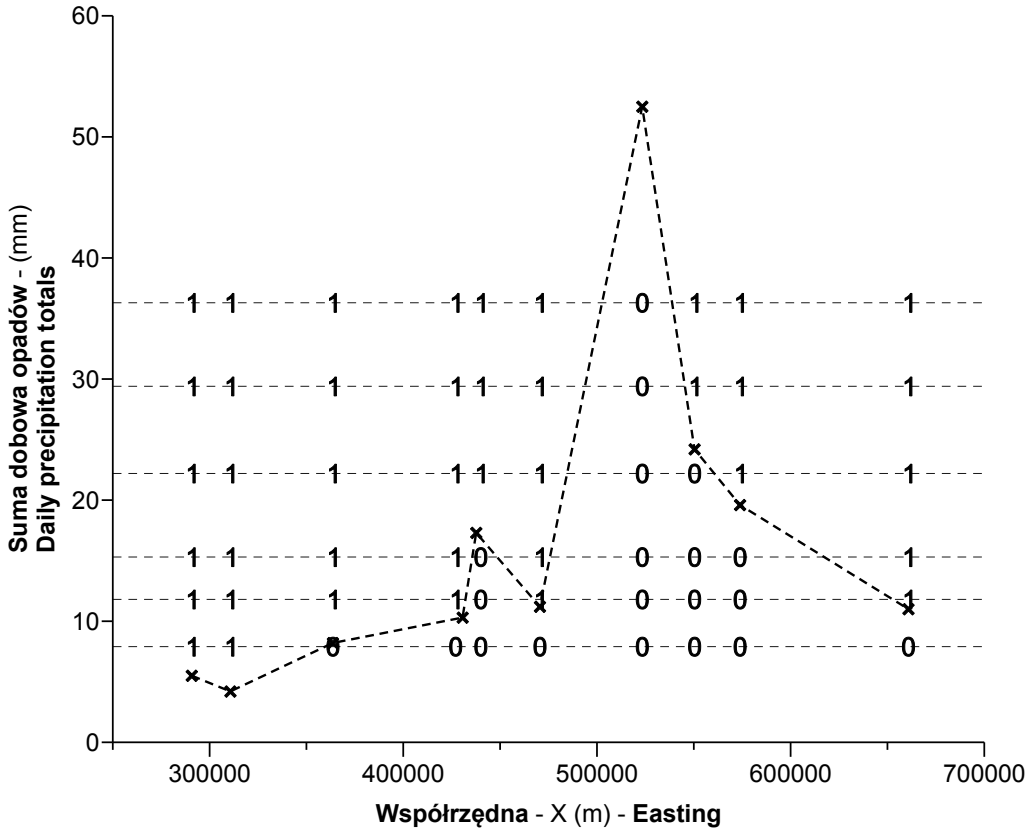
Ryc. 8. Maksymalne sumy dobowe opadów zarejestrowane w maju 1980 roku na posterunkach usytuowanych na profilu (ryc. 7).



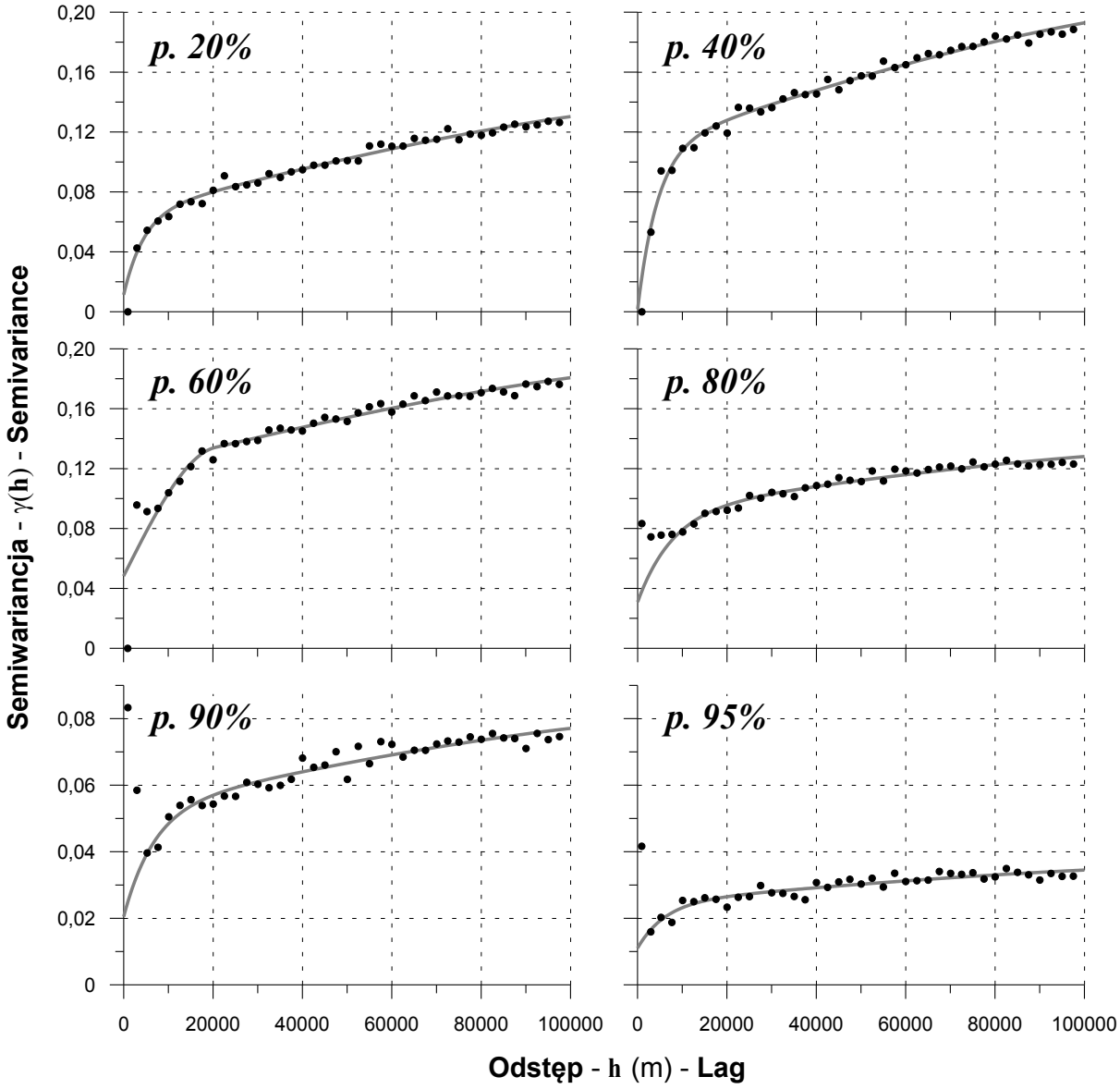
Ryc. 9. Skumulowany rozkład prawdopodobieństwa (*cdf*) maksymalnych dobowych sum opadów zarejestrowanych w posterunkach opadowych na terenie Polski w maju 1980 roku. Na wykresie zaznaczono wysokości sum dobowych o prawdopodobieństwie przewyższenia 0,2, 0,4, 0,6, 0,8, 0,9 i 0,95 (percentyle 20, 40, …., 95%).



Ryc. 10. Wartości progowe maksymalnych sum dobowych opadów w maju 1980 roku o prawdopodobieństwie 0,2, 0,4, 0,6, 0,8, 0,9 i 0,95 (ryc. 9, percentyle 20, 40, 60, 80, 90 i 95%) naniesione na dane profilowe (ryc. 7 i 8).



Ryc. 11. Maksymalne sumy dobowe opadów zarejestrowane w maju 1980 roku na analizowanym profilu (ryc. 7 i 8) przekodowane na wektory danych binarnych zgodnie ze wzorem [12] w zależności od przekroczenia wartości progowych wyznaczonych z globalnej krzywej skumulowanego rozkładu prawdopodobieństwa (precentyle 20, 40, 60, 80, 90 i 95%, ryc. 9).



Ryc. 12. Semiwariogramy empiryczne i ich modele dla wartości kodowanych (percentyle 20, 40, 60, 80, 90 i 95% - ryc. 9) maksymalnych sum dobowych opadów na terenie Polski w maju 1980 roku.

2) **utworzenie dla każdej lokalizacji danych pomiarowych wektora danych binarnych** (ryc. 10 i 11). Oryginalne, ciągłe dane pomiarowe przetwarza się do formy binarnej zgodnie ze wzorem [12]. Wartość MSDO zmierzona w Kamiennej Górze (5,5 mm) jest mniejsza od najniższego progu (20 percentyl = 7,8 mm), dlatego po przekodowaniu uzyskujemy wektor składający się tylko z jedynek (ryc 11). W Kobylnie zanotowano 17,3 mm, a zatem więcej od pierwszych trzech wartości progowych (20% = 7,9 mm, 40% = 11,8 mm i 60% = 15,3 mm), i dlatego wektor wartości binarnych jest następujący: 0,0,0,1,1,1. Dla lokalizacji stacji Małusy Wielkie wektor ten składa się z samych zer, bo zmierzona tam w

maju 1980 roku MSDO wynosząca 52,5 mm była większa od wszystkich wybranych wartości progowych.

3) **obliczanie i modelowanie miar struktury przestrzennej wartości kodowanych** (ryc. 12). Dla kolejnych progów obliczane są empiryczne miary struktury przestrzennej, najczęściej semiwariancje danych kodowanych. Do dyskretnych wartości semiwariancji dopasowywane są następnie funkcje matematyczne, zgodnie z regułami liniowego modelu regionalizacji (patrz podrozdział III.2.5.2). Jeśli modele te mają być później używane do estymacji w węzłach siatki interpolacyjnej, warunkowych funkcji skumulowanego rozkładu (ang. *conditional cumlative distribition finction – ccdf*) metodą krigingu danych kodowanych (patrz rozdział X.2), należy na etapie modelowania przestrzegać określonych zasad. Mają one na celu ograniczenie występowanie w estymowanych *ccdf* błędów relacji porządkowych[13] (ang. *order relations errors*, Deutsch, Journel 1998, Goovaerts 1997). W szczególności modele dla poszczególnych wartości progowych powinny być tworzone jako kombinacja tych samych modeli elementarnych, a ich parametry winny zmieniać się między kolejnymi progami w sposób stopniowy. W prezentowanym jednowymiarowym przykładzie (ryc. 7 i 8) semiwariancje empiryczne obliczono z kodowanych danych całego zbioru MSDO na obszarze Polski w maju 1980 roku (ryc. 12). Do ich modelowania (por. rozdział III.2.5.2) użyto za każdym razem identycznego zestawu trzech modeli elementarnych: nuggetowego, wykładniczego o zasięgu kilku do kilkunastu kilometrów i sferycznego o arbitralnie przyjętym zasięgu 150 km. Zastosowane funkcje były bardzo dobrze dopasowane do danych empirycznych oprócz pierwszych 1-2 odstępów niektórych progów(np. od 0,6 do 0,95), gdzie ze względu na małą ilość pomiarów wartości semiwariogramu były chaotyczne. Uzyskane modele (ryc. 12) wyraźnie pokazują odmienność struktury przestrzennej dla MSDO o różnej wysokości. Względny udział[14] semiwariancji nuggetowej był najmniejszy przy niskich opadach (percentyl 40 i 20 – odpowiednio 0,007 i 0,07), a osiągał maksimum przy najwyższych progach (percentyl 90 i 95 – odpowiednio 0,234 i 0,1). Zasięg pierwszej struktury był

[13] W dowolnej lokalizacji **u** każde estymowane *posteriori* prawdopodobieństwo $[F(\mathbf{u};z_k \mid (n))]^*$ musi należeć do przedziału [0,1], a seria K takich szacunków musi być niemalejącą funkcją wielkości wartości progowej z_k.
[14] Porównania dokonuje się po standaryzacji polegającej na podzieleniu każdej wartości semiwariogramu empirycznego i modelowego przez wariancję danych kodowanych analizowanego zbioru.

największy przy percentylu 80 (23,5 km), nieco mniejszy przy opadach najwyższych (20,5 oraz 18,5 km), a najmniejszy (13,5 km) przy pierwszym i drugim progu. Opady najniższe (20 i 40 percentyl) charakteryzowały się największym gradientem spadku podobieństwa wraz z odległością, najwyższe (95 percentyl) – najmniejszym.

III.2.5. Modelowanie struktury przestrzennej danych

W niniejszej pracy modelowanie struktury przestrzennej było zasadniczym czynnikiem decydującym o jakości wszystkich najważniejszych jego wyników[15]. Z drugiej strony, liczba modeli jakie trzeba było opracować, zmuszała do szukania takiego rozwiązania, które umożliwiałoby uzyskanie dobrych wyników w rozsądnym czasie. Dlatego, problematykę tę przedstawiono w niniejszym podrozdziale bardzo szeroko.

III.2.5.1. Wprowadzenie do problematyki modelowania struktury przestrzennej

Wszystkie geostatystyczne algorytmy estymacji i symulacji przestrzennej wymagają podania w matematycznej (parametrycznej) formie modelu struktury przestrzennej analizowanej cechy (Goovaerts 1997, Deutsch, Journel 1998, Isaaks, Srivastava 1989). Nie mogą to być po prostu wartości empirycznych miar struktury przestrzennej, jak kowariogram, korelogram czy semiwariogram. Przyczyn jest kilka.

W macierzach krigingowych dla każdego oczka siatki interpolacyjnej (patrz dodatek X.2) potrzebne są wartości kowariancji pomiędzy danymi pomiarowymi znajdującymi się w sąsiedztwie szukania (ang. *data covariances*) oraz kowariancji danych do lokalizacji estymowanej nieznanej (ang. *data-to-unknown covariances*). Oznacza to, że muszą one być możliwe do określenia dla dowolnej kombinacji odległości i kierunku. Wszystkie zaś wymienione wyżej miary empiryczne autokorelacji przestrzennej są dyskretne, czyli nieciągłe, obliczane są bowiem jako średnie przedziałowe (odstępy odległości i sektory kierunków). W estymacjach potrzebne są też często wartości kowariancji dla odległości i kierunków, dla których nie ma danych empirycznych. Nie bywa jednak stosowane, wydawałoby się najprostsze, rozwiązanie, to jest interpolacja i ekstrapolacja kowariancji empirycznych. Składają się na to z kolei dwie przyczyny.

[15] Wyniki modelowania zamieszczono na dołączonym dysku DVD (patrz. załącznik XII.2).

Po pierwsze, empiryczne miary struktury przestrzennej stanowią statystyki obliczane z próby i jako takie są tylko przybliżeniem relacji przestrzennych w całej populacji. Przy założeniu, że dysponuje się danymi niezawierającymi błędów, ich wiarygodność jest uzależniona z jednej strony od wielkości i reprezentatywności próby, a z drugiej – od skali zmienności analizowanej cechy (Webster, Oliver 2001). Zawsze dochodzi jednak trzeci element, jakim są nieuniknione błędy, zarówno pomiaru cechy, jak i określenia lokalizacji stanowiska, gdzie pomiar został dokonany. Zazwyczaj analizowane są próby stanowiące milionowe, miliardowe lub nawet mniejsze części całej populacji[16], a zalecenie ich losowości często nie jest w pełni spełnione. Wszystkie wymienione czynniki, a także czułość stosowanych zazwyczaj miar struktury przestrzennej na naturalnie występujące wartości anomalne (por. dodatek X.5) powoduje, że ich wykresy często są chaotyczne i nie odzwierciedlają wiarygodnie stosunków istniejących w całej zbiorowości. Matematyczny model może przynajmniej zredukować znaczenie pierwszej z wyżej wymienionych wad, wygładzając chaotyczne fluktuacje danych empirycznych.

Druga przyczyna jest natury „wewnętrznej" i wiąże się z matematyczną formą estymatora krigingowego. Jak wspominano poprzednio (por. rozdział III.2.2), estymowane lub symulowane wartości są w geostatystyce traktowane jako zmienne losowe będące liniową kombinacją innych znanych zmiennych losowych. Wariancja zaś jakiejkolwiek liniowej kombinacji Y zmiennych losowych $Z(\mathbf{u}_\alpha)$, $\mathbf{u}_\alpha \in A$, jest wówczas liniową kombinacją wartości kowariancji owych zmiennych i musi nieujemna (ang. *non-negative*):

$$\begin{aligned} \mathrm{Var}\{Y\} &= \mathrm{Var}\left\{\sum_{\alpha=1}^{n} \lambda_\alpha Z(\mathbf{u}_\alpha)\right\} \\ &= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C(\mathbf{u}_\alpha - \mathbf{u}_\beta) \;\geq\; 0 \end{aligned} \qquad [\mathbf{16}]$$

dla jakiejkolwiek z wybranych n lokalizacji $\mathbf{u}_\alpha \in A$ i dla jakiejkolwiek wagi λ_α. Spełnienie tego warunku jest możliwe tylko przy zastosowaniu takich funkcji kowariancji C(**h**), nieparametrycznych czy parametrycznych, które są pozytywnie połowicznie określone (ang. *positive semidefinite*). Stosowanie interpolowanych / ekstrapolowanych wartości empirycznych miar struktury przestrzennej nigdy nie gwarantuje, że obliczenia estymacji /

[16] W niniejszym opracowaniu (por. rozdział V) korzystano z wyników pomiarów w średnio 2485 lokalizacjach na powierzchni 319 114 km^2 (319,114 × 10^9 m^2). Ponieważ powierzchnia zbiorcza standardowego deszczomierza czy pluwiografu wynosi 200 cm^2, to średnia sumaryczna powierzchnia próby MSDO którą dysponowano wynosiła 49,6 m^2. W stosunku do całej populacji stanowi to 1,554 × 10^{-10}, a zatem dziesięciomiliardową jej część.

symulacji dadzą jakikolwiek wynik. Gwarancję taką można mieć jedynie przy zastosowaniu modelu matematycznego o takiej postaci, który jest z góry pozytywnie połowicznie określony (Deutsch, Journel 1998, Goovaerts 1997, Zawadzki 2005). Modele takie określa się jako dozwolone (ang. *permissible*).

III.2.5.2. Proste i złożone, dopuszczalne funkcje używane przy modelowaniu struktury przestrzennej

W literaturze geostatystycznej (Bleinès i in. 2007, Chilès, Delfiner 1999, Cressie 1993, Olea 1999, Webster, Oliver 2001) podawane jest w sumie kilkanaście funkcji spełniających warunek pozytywnej połowicznej określoności. Nie wszystkie jednak mogą być stosowane bez ograniczeń – na przykład dotyczących liczby wymiarów przestrzeni danych. W praktyce wykorzystywane jest jedynie kilka z nich, które zapewniają w większości przypadków bardzo dobre, lub dobre odwzorowanie struktur przestrzennych spotykanych w środowisku. Decyzja o wykonywaniu w ramach całego projektu badawczego estymacji pola prawdopodobieństwa MSDO metodą krigingu wartości kodowanych (ang. *Indicator Kriging*, IK) i jego symulacji metodą *p-pola* (ang. *p-field*), przy wykorzystaniu programu IKSIM (Ying 2000), miała również konsekwencje dotyczące modelowania struktury przestrzennej (patrz dodatek X.2). We wspomnianym bowiem programie komputerowym dopuszczalne są jedynie cztery wymienione i opisane niżej oraz przedstawione na rycinie 13A dozwolone modele struktury przestrzennej danych.

- Model nuggetowy[17] (ang. *nugget effect model*):

$$\gamma(h) = \begin{cases} 0, & jeżeli\ \ h = 0 \\ C_0, & poza\ \ tym \end{cases} \qquad [\mathbf{17}]$$

- Model sferyczny (ang. *spherical*) o zasięgu *a*:

$$\gamma(h) = C \cdot Sph\left(\frac{h}{a}\right) = \begin{cases} C \cdot \left[1{,}5 \cdot \frac{h}{a} - 0{,}5 \cdot \left(\frac{h}{a}\right)^3\right], & jeżeli\ \ h \le a \\ C, & poza\ \ tym \end{cases} \qquad [\mathbf{18}]$$

- Model wykładniczy (ang. *exponential*) o „praktycznym”[18] zasięgu *a*:

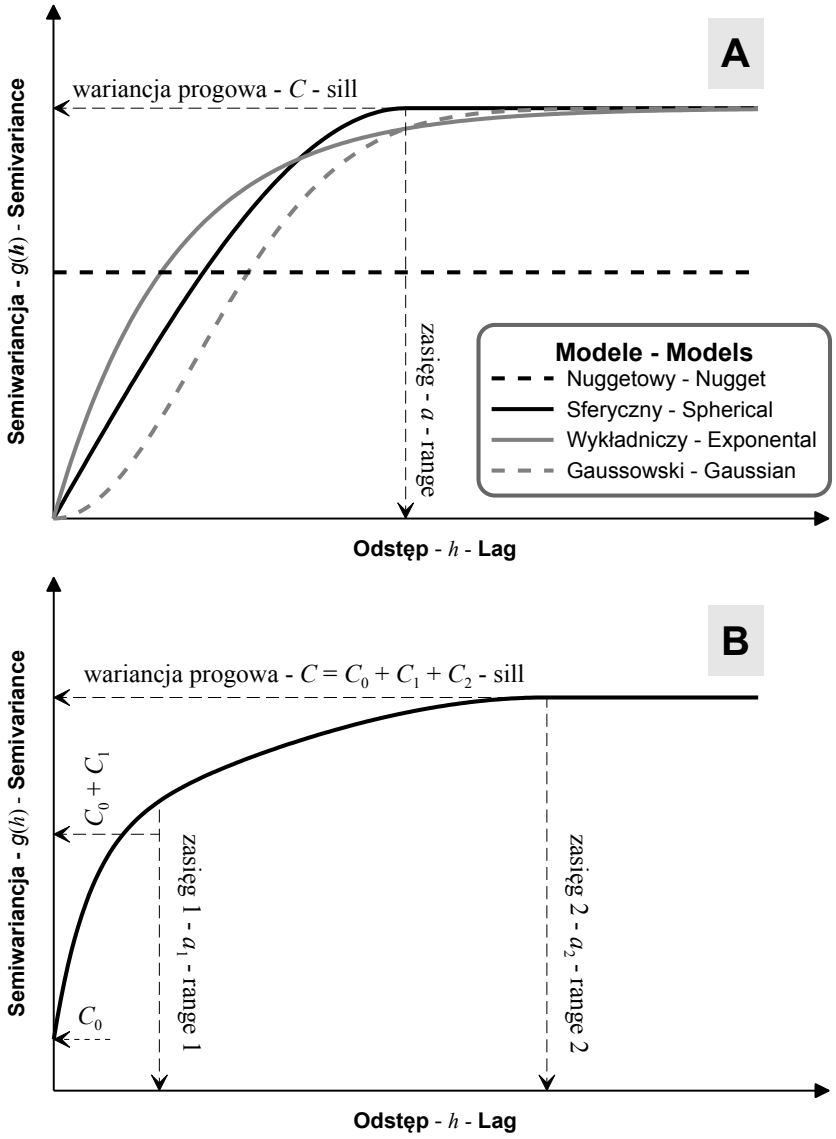
[17] Jak wspomniano, w polskiej literaturze (Zawadzki 2005, Namysłowska-Wilczyńska 2006) stosowany jest termin model „efektu samorodka”.
[18] Pojęcie zasięgu praktycznego jest wyjaśnione w dalszej części tekstu.

$$g(h) = C \cdot Exp\left(\frac{h}{a}\right) = C \cdot \left[1 - \exp\left(\frac{-3h}{a}\right)\right] \quad [\mathbf{19}]$$

- Model gaussowski (ang. *gaussian*) o „praktycznym" zasięgu *a*:

$$g(h) = C \cdot \left[1 - \exp\left(-\frac{(3h)^2}{a^2}\right)\right] \quad [\mathbf{20}]$$

gdzie C_0 oznacza wariancję nuggetową, C – wariancję progową (ang. *sill*), a – zasięg lub zasięg „praktyczny" (ang. *range, practical range*), zaś h – odległość (odstęp, ang. *lag*).



Ryc. 13. Podstawowe modele semiwariogramów wykorzystywane do opisu struktury przestrzennej maksymalnych sum dobowych opadów oraz estymacji i symulacji ich pola prawdopodobieństwa (A). Przykład złożonego (zagnieżdżonego) modelu semiwariogramu składającego się z trzech modeli podstawowych: nuggetowego, wykładniczego i sferycznego (B). Na wykresach zaznaczono podstawowe parametry modeli: wariancję progową (C, *sill*), wariancję składowych (C_x, *partial sills*), zasięg (a, *range*), zasięgi składowych (a_x, *partial ranges*).

Zastosowanie modelu nuggetowego ([17], ryc. 13A) oznacza w praktyce stwierdzenie braku autokorelacji zjawiska. Wariancja procesu jest stała dla każdej odległości większej od zera. Ponieważ większość parametrów środowiskowych jest w jakiejś skali przestrzennej ciągłe, konieczność użycia tego modelu oznacza zazwyczaj, że odstęp próbkowania był większy niż zasięg autokorelacji. Inną możliwość stwarza istnienie zjawiska w 100% losowego, nieciągłego, którego charakterystyki zmieniają gwałtownie przy przejściu z jednego punktu do drugiego. Obrazem symulacji bezwarunkowej przy użyciu tego modelu jest, zgodnie z terminologią stosowaną w analizie serii czasowych, „biały szum", podobny do tego co jest obserwowane na ekranie telewizora (lampy katodowej), do którego nie dociera żaden sygnał[19].

Model sferyczny ([18], ryc. 13A) może być stosowany w 1, 2 i 3 wymiarach, i jest jednym z najczęściej stosowanych w geostatystyce do charakterystyki struktury przestrzennej. Daje on reprezentację cech ciągłych, które mają podobną rozciągłość, a ich zmienność ma charakter przeplatających się nieregularnych płatów z wysokimi i niskimi wartościami. Średnia średnica owych płatów jest reprezentowana przez zasięg modelu. Model sferyczny ma w początkowym odcinku charakter funkcji liniowej o nachyleniu $3C/2a$.

Podobnie często wykorzystywany jest model ujemnie wykładniczy ([19], ryc. 13A). Funkcja ta osiąga wariancję progową asymptotycznie i dlatego nie ma skończonego zasięgu. Zamiast niego podawany jest tak zwany zasięg „praktyczny" lub „efektywny" zdefiniowany jako odległość na jakiej model osiąga 95% wartości wariancji progowej. Model wykładniczy ma również liniowy charakter w fazie początkowej, ale o większym nachyleniu niż w przypadku sferycznego: C/a. Funkcja ta odgrywa bardzo ważną rolę teoretyczną. Stanowi bowiem istotę losowości w ujęciu przestrzennym. Jest to semiwariogram procesów autoregresyjnych pierwszego rzędu i procesów Markowa. Można się spodziewać semiwariogramu wykładniczego wówczas, kiedy głównym źródłem zmienności cechy jest występowanie odmiennych typów systemów, a granice między typami występują losowo zgodnie z procesem Poissona. Przykładem może być sytuacja, kiedy zmienność przestrzenna pewnej cechy gleb, na przykład odczynu, jest uwarunkowana głównie różnicami pomiędzy typami gleb. Inaczej mówiąc, jest to semiwariogram cech ciągłych, których struktury mają losowy zasięg.

[19] Model nuggetowy opisuje również pewne mikro struktury regularne typu mozaiki (por. Chilès, Delfiner 1999, ryc. 2.10, s. 53).

Ostatni z wykorzystywanych, dopuszczalnych, modeli – gaussowski ([20], ryc. 13A) – również osiąga poziom wariancji progowej asymptotycznie, dlatego tak jak w poprzednim przypadku, zasięg praktyczny definiowany jest jako odległość dla której wartość semiwariancji wyliczona z modelu osiąga 95% wariancji progowej. Model gaussowski odróżnia się od dwóch wymienionych poprzednio przede wszystkim parabolicznym kształtem w początkowym odcinku. Oznacza to, że „dotyka" on osi z zerowym nachyleniem. Stanowi to granicę zmienności losowej, przy której w rzeczywistości istnieje wartość stała i podwójnie różniczkowalna (Wackernagel 1998). Model ten ze względu na swój deterministyczny charakter daje w wielu przypadkach nierealistyczne wyniki estymacji i zazwyczaj nie może być stosowany samodzielnie. Jego użycie jest czasami uzasadnione przy analizie przestrzennej i prognozowaniu parametrów o bardzo regularnej i łagodnej zmienności przestrzennej, na przykład poziomu wód gruntowych w obszarach o mało zróżnicowanej rzeźbie i jednolitej budowie geologicznej.

W wielu sytuacjach, w celu dokładnego odwzorowania kształtu semiwariogramu empirycznego konieczne jest połączenie dwóch lub większej ilości modeli podstawowych $\boldsymbol{g}(\mathbf{h})$. Problem tkwi w tym, że nie wszystkie kombinacje dopuszczalnych modeli dają w efekcie funkcję dopuszczalną, to znaczy z nieujemną wariancją. Najprostszym sposobem utworzenia modelu dopuszczalnego jest stworzenie najpierw funkcji losowej. Semiwariogram takiej funkcji jest z definicji dopuszczalny. Praktycznie rzecz biorąc, konieczne jest spełnienie dwóch warunków tak zwanego liniowego modelu regionalizacji[20] (ang. *linear regionalisation model*):

1) wszystkie użyte w modelu złożonym podstawowe funkcje $\boldsymbol{g}_l(\mathbf{h})$ muszą być dopuszczalne,
2) wariancja progowa b^l każdego podstawowego modelu semiwariogramu musi być dodatnia [21], a wówczas:

$$\gamma(\mathbf{h}) = \frac{1}{2} E\left\{\left[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right]^2\right\} = \sum_{l=0}^{L} b^l \boldsymbol{g}_l(\mathbf{h}) \quad \text{z} \quad b^l = \left(a^l\right)^2 \geq 0 \qquad [\mathbf{21}]$$

Model złożony $\gamma(\mathbf{h})$ w takiej sytuacji jest wyrażony jako pozytywna liniowa kombinacja podstawowych modeli semiwariogramów $\boldsymbol{g}_l(\mathbf{h})$. W literaturze przedmiotu (Deutsch, Journel

[20] Oczywiście pojęcie to nie ma nic wspólnego z regionalizacją w sensie nauk geograficznych.

1998, Goovaerts 1997) popularna jest jego także alternatywna nazwa: „*nested model*", czyli model zagnieżdżony. Rycina 13B prezentuje przykład takiego modelu składającego się z trzech podstawowych: nuggetowego, wykładniczego i sferycznego. Oprócz kilku wyjątków wszystkie z około 4800 opracowanych w niniejszej pracy modeli struktury przestrzennej miały charakter złożony zawierając minimum dwa elementy (w tym nuggetowy), a maksimum – pięć. **Złożona struktura przestrzenna analizowanej cechy (w tym przypadku MSDO), świadczy, że jest ona efektem działania kilku procesów operujących w różnych skalach**.

III.2.5.3. Specyfika matematycznego modelowania struktury przestrzennej

Modelowanie matematyczne struktury przestrzennej nie jest, jakby się wydawało, nawet przy aktualnym poziomie mocy obliczeniowej komputerów osobistych, zagadnieniem łatwym. Wręcz przeciwnie, automatyczne algorytmy zdają egzamin jedynie przy izotropowych, i to raczej najprostszych przypadkach. Próby stworzenie takich uniwersalnych procedur podejmowane są od ponad 30 lat bez większego powodzenia (Webster, Oliver 2001). W modułach modelowania struktury przestrzennej w większości dostępnych programów geostatystycznych dokonuje się tego na drodze manualnej lub półautomatycznej, przy czym i w tym drugim przypadku operator ma całkowitą kontrolę nad przebiegiem obliczeń, łącznie z możliwością całkowitego wyłączenia „automatyki" i / albo „ręcznej" modyfikacji wyników. Tam gdzie stosuje się wariant w 100% manualny, operator obserwując na ekranie wykres semiwariogramu empirycznego (albo innej miary struktury przestrzennej), wybiera podstawowe składowe modelu: ilość, typ i następstwo struktur elementarnych, a następnie metodą prób i błędów optymalizuje ich parametry: zasięg, wariancje cząstkowe, kierunek i stopień anizotropii. W procedurach półautomatycznych drugi z wymienionych wyżej etapów, to jest optymalizacja parametrów, dokonywana jest w mniejszym lub większym stopniu automatycznie. Jeśli operator nie jest zadowolony z końcowego wyniku, może dokonać jego modyfikacji, cały czas obserwując efekt swoich działań na ekranowym wykresie. Ten element procedury geostatystycznej jest bardzo często krytykowany ze względu na subiektywizm. Bywa jednak, że może to doprowadzić do radykalnego polepszenia wyników estymacji lub symulacji. Związane jest to bowiem z możliwością uwzględnienia wiedzy *a priori*, wiedzy eksperta, o naturze zmienności przestrzennej zjawiska, które wygenerowało obserwowany rozkład analizowanej cechy (Goovaerts 1997). Dane pomiarowe ze względu na małą próbę, czy błędy pomiarowe, mogą tych relacji nie

wykazywać. Kiedy na przykład zadaniem jest estymacja skażenia gleb wokół emitora, jakim może być komin elektrowni cieplnej, wiemy że musi ona wykazywać anizotropię o kierunku i rozmiarach uzależnionych od lokalnego reżimu wiatrów. Wiedzę tą możemy w trakcie modelowania wykorzystać, „wymuszając" uwzględnienie anizotropii.

Oprócz problemów merytorycznych wadą ręcznego i półautomatycznego modelowania struktury przestrzennej jest także jego czasochłonność, co było szczególnie istotne w kontekście niniejszej pracy. Opracowanie jednego złożonego, anizotropowego modelu może trwać wiele minut. W trakcie prac ze zbiorami danych miesięcznych i rocznych MSDO wykonano ich blisko 4800.

III.2.5.4. Kryterium jakości dopasowania modelu

Zarówno przy ręcznym, jak i automatycznym modelowaniu struktury przestrzennej potrzebne jest obiektywne kryterium jakości dopasowania modelu do danych empirycznych. W pierwszym przypadku ma ono charakter pomocniczy. Operator może, ale nie musi, opierając się na obliczanych po każdej wprowadzonej przez niego zmianie w parametrach modelu, wartościach owego kryterium, dążyć do optymalizacji wyniku. Optymalizacja automatyczna musi być dokonywana w odniesieniu do precyzyjnie zdefiniowanego jej celu, którym zazwyczaj jest minimalizacja albo maksymalizacja wartości jakiejś funkcji. Konieczne jest również podanie warunku zakończenia obliczeń – zazwyczaj dokonywanych metodą kolejnych przybliżeń (iteracyjnie). Najczęściej w geostatystyce stosowanym kryterium dopasowania modelu do danych empirycznych jest Ważona Suma Kwadratów (ang. *Weighted Sum of Squares*, WSS, Cressie 1985, 1991, Jian i in. 1996, Pardo-Igúzquiza 1999) różnic pomiędzy eksperymentalnymi $\hat{\gamma}(\mathbf{h}_k)$ a modelowanymi $\gamma(\mathbf{h}_k)$ wartościami semiwariogramu:

$$\mathrm{WSS} = \sum_{k=1}^{K} \omega(\mathbf{h}_k) \cdot \left[\hat{\gamma}(\mathbf{h}_k) - \gamma(\mathbf{h}_k) \right]^2 \qquad [\mathbf{22}]$$

Waga $\omega(\mathbf{h}_k)$ przypisana do każdego odstępu $\mathbf{h}_k$ jest zazwyczaj proporcjonalna do liczby $N(\mathbf{h}_k)$ par danych, które są uwzględniane w obliczeniu wartości semiwariogramu empirycznego $\hat{\gamma}(\mathbf{h}_k)$. Podstawą takiego rozwiązania jest założenie, że wiarygodność semiwariogramu empirycznego wzrasta wraz z wielkością próby. Nie jest to jednakże jedyny wariant wagi w kryterium WSS. Inny, równie często stosowany, przywiązuje większe

znaczenie do semiwariogramów obliczanych dla pierwszych odstępów[21], poprzez podzielenie ilości par danych przez podniesioną do kwadratu wartość modelu: $\frac{N(\mathbf{h}_k)}{\left[\gamma(\mathbf{h}_k)\right]^2}$. W innych podejściach wiarygodność statystyczną wartości semiwariogramu empirycznego ocenia się nie poprzez ilość par danych, która posłużyła do jego obliczenia, ale poprzez zróżnicowanie indywidualnych wartości różnic obliczonych dla każdej pary: małe zróżnicowanie – duża wiarygodność. Waga WSS jest wówczas odwrotnie proporcjonalna do odchylenia standardowego indywidualnych wartości różnic. Na koniec omawiania tego zagadnienia należy wspomnieć, że metoda WSS, jakkolwiek najbardziej popularna, nie jest jedynym rozwiązaniem problemu automatycznego dopasowania parametrów modeli struktury przestrzennej. Stosowane są również algorytmy tzw. maksymalnej wiarygodności (ang. *maximum likelihood, ML*) lub ograniczonej ML (ang. *restricted ML*, *REML)*, gdzie model tworzony jest bezpośrednio na podstawie surowych wartości różnic (Dietrich, Osborne 1991, Pardo-Igúzquiza 1997, 1998, Zimmerman 1989). Ponieważ jednak bazują one na założeniu rozkładu normalnego, ich oszacowania są często obciążone. Obliczenia wykonywane metodami ML i REML są także stosunkowo wolne przy dużych próbach (Olea 1999).

III.2.5.5. Optymalizacja modelu struktury przestrzennej

Z półautomatycznym i ręcznym konstruowaniem złożonego (zagnieżdżonego) modelu struktury przestrzennej wiąże się jeszcze jeden istotny problem – niepewności co do wyboru optymalnej liczby i kombinacji funkcji podstawowych (elementarnych). Z jednej strony model powinien być jak najlepiej dopasowany do danych eksperymentalnych, z drugiej zaś wiadomo, że ich niewielkie fluktuacje mogą być zupełnie przypadkowe. Zagadnienie to można rozpatrywać w dwóch kontekstach. Pierwszy z nich ma charakter optymalizacyjny drugi – praktyczny.

Z praktycznego punktu widzenia najlepszym modelem jest nie ten najlepiej dopasowany do danych obserwacyjnych, ale dający najbardziej dokładną prognozę. Budując model rzadko dysponuje się taką liczbą danych, która umożliwia dokonanie niezależnej jego walidacji – porównanie rzeczywistych wartości analizowanej cechy z prognozowanymi na jego podstawie. Dlatego, zazwyczaj w geostatystyce stosuje się uproszczoną metodę testowania

[21] Jakość estymacji zależy głównie od poprawnego określenia „kształtu" modelu u jego początku, natomiast semiwariancje empiryczne pierwszych odstępów obliczane są zazwyczaj na podstawie znacznie mniejszej liczby par punktów niż dalsze.

jakości modelu, zwaną kroswalidacją (Davis 1987, Goovaerts 1997, Webster, Oliver 2001). Polega ona na wykonywaniu sekwencyjnie *n* estymacji dla każdej lokalizacji z posiadanego zbioru danych, z wyłączeniem każdorazowo w trakcie obliczeń zmierzonej w tym miejscu wartości cechy. Wykonany szacunek jest oparty na pozostałych w puli *n* - 1danych, i model opracowany na podstawie wszystkich *n* danych. W efekcie, dla każdego z *n* punktów pomiarowych dokonuje się porównania rzeczywiście zmierzonych wartości analizowanej cechy z prognozą. Używając różnych syntetycznych miar jakości estymacji, jak na przykład średni błąd (*ME*), pierwiastek średniego błędu kwadratowego (*RMSE*), czy współczynnik korelacji (*r*) wykonuje się porównania alternatywnych modeli i wybiera najlepszy. Należy jednakże podkreślić, że kroswalidacja nie jest techniką w pełni obiektywną i dającą całkowicie wiarygodne wyniki. Jej słabość polega na używaniu na wszystkich etapach obliczeń (budowy modelu, estymacji/walidacji) tych samych danych, a oceny błędów dotyczą tylko lokalizacji pomiarowych, a nie najbardziej interesujących miejsc, w których pomiarów nie wykonano.

Optymalizacyjne podejście do zagadnienia budowy złożonego modelu semiwariancji opiera się idei, że musi istnieć równowaga pomiędzy prostotą modelu, a więc także łatwością wykonywanych na jego podstawie estymacji, a jakością jego dopasowania do danych eksperymentalnych. Polepszenie modelu poprzez minimalizację kryterium WSS może się dokonywać praktycznie w nieskończoność, jeśli zwiększać się będzie ilość funkcji elementarnych. Jednakże kolejne komplikowanie jego postaci skutkuje coraz mniejszym przyrostem jakości dopasowania, dlatego ważne byłoby ustalenie kryterium, które umożliwiłoby w obiektywny i powtarzalny sposób zachowanie proporcji między dążeniem do prostoty modelu, a wiernością odwzorowania wyników pomiarów. Webster i McBratney (1989) zaproponowali zastosowanie do tego celu kryterium informacyjnego Akaike (ang. *Akaike Information Criterion*, *AIC*, [23]):

$$AIC = \left\{ n\ln\left(\frac{2\pi}{n}\right) + n + 2 \right\} + n\ln R + 2p \qquad [\mathbf{23}]$$

gdzie: *n* jest liczbą punktów na wariogramie, *p* – ilością parametrów modelu, a *R* stanowi średnią podniesionych do kwadratu różnic pomiędzy wartościami eksperymentalnymi a modelem. Do dalszego etapu procedury geostatystycznej wybiera się ten model, dla którego AIC jest najmniejsze. Fragment wzoru [23] znajdujący się w nawiasie jest stały dla każdego konkretnego semiwariogramu, dlatego można go uprościć do postaci [24]:

$$AIC = n\ln R + 2p \qquad [\mathbf{24}]$$

Minimalizacja kwadratów odchyleń (WSS) zmniejsza wartość R, lecz jeśli dalsze jego obniżanie dokonuje się jedynie poprzez zwiększanie p, to w pewnym momencie spadek AIC zostaje zatrzymany.

Zastosowane oprogramowanie (patrz dalej), ale przede wszystkim niespotykana w typowych opracowaniach geostatystycznych liczba modeli, które trzeba było przygotować, uniemożliwiały rutynowe wykonywanie oceny ich optymalności. Przy czym, rzadko występowała taka potrzeba, szczególnie w odniesieniu do semiwariogramów danych znormalizowanych (patrz dalej w tym podrozdziale). W większości przypadków układ punktów był bowiem bardzo regularny, a odstępy pomiędzy kolejnymi załamaniami krzywej na tyle duże, że nie było wątpliwości, które z podstawowych funkcji należy użyć. Więcej problemów było z niektórymi semiwariogramami danych kodowanych, co szczegółowo opisano dalej w poniższym podrozdziale. Trzeba wyraźnie podkreślić, że jakkolwiek w ocenie autora znaczenie optymalizacji przy modelowaniu struktury przestrzennej w kontekście analizowanych w niniejszej pracy zbiorów danych nie jest duże, to jednak ten problem istnieje i nie został rozwiązany. Będzie to tematem osobnego opracowania.

Przedstawiony powyżej obraz modelowania struktury przestrzennej był paradygmatem geostatystyki przez ostatnie 40 lat. W każdym dostępnym oprogramowaniu realizującym funkcje geostatystycznej estymacji, symulacji bądź optymalizacji wymagane jest podanie przez operatora modelu struktury przestrzennej w postaci parametrycznej. Prawdopodobnie ulegnie to w najbliższym czasie zmianie. Pojawiła się bowiem idea pominięcia tradycyjnego sposobu budowy tego modelu – nieparametryczna alternatywa umożliwiająca automatyzację i obiektywizację procedury (Yao, Journel 1998). Sprowadza się to do transformacji eksperymentalnych map korelacji (lub kroskorelacji) do map gęstości spektralnej przy użyciu szybkiej transformaty Fouriera (FFT). Owe mapy gęstości spektralnej są następnie wygładzane przy zastosowaniu ograniczeń dodatniości i sumowania do jedności. Przeprowadzana następnie transformacja zwrotna przez odwrotność FFT daje w efekcie dozwolone, pozytywnie połowicznie określone, mapy korelacji. Dzięki tej metodzie „prawidłowe" i praktycznie użyteczne mapy korelacji uzyskuje się automatycznie bez konieczności analitycznego tworzenia modelu struktury. Potrzebne do obliczeń estymacji / symulacji wartości kowariancji dla dowolnej odległości i dowolnego kierunku uzyskiwane są tych map przez interpolację lub ekstrapolację. Trudność weryfikacji w przestrzeni danych

surowych warunku połowicznej pozytywnej określoności jest usunięta, ponieważ w domenie częstości jedynym wymogiem jest pozytywność uzyskanej funkcji gęstości i jej sumowanie do jedności.

III.2.5.6. Modelowanie struktury przestrzennej MSDO w programie ISATIS

Modelowanie struktury przestrzennej całych[22] analizowanych zbiorów miesięcznych i rocznych MSDO wykonywano w programie ISATIS (Bleinès i in. 2007) metodą półautomatyczną. Zdecydowano się na przeprowadzenie jedynie analizy izotropowej, co bardzo ją ułatwiło i przyspieszyło (por. dodatek X.4). Automatyczna procedura dopasowania wariancji cząstkowych w programie ISATIS zmierza do minimalizacji odległości pomiędzy wartością semiwariogramu empirycznego dla danego odstępu a odpowiadającą mu wartością modelu (za pomocą omówionego poprzednio kryterium WSS). Minimalizacja ta jest przeprowadzana przy uwzględnieniu jednej z czterech możliwych kombinacji wag:

1) każda wartość semiwariogramu niezależnie od odstępu i kierunku jest traktowana identycznie (bez wag);
2) waga dla każdego odstępu danego kierunku jest proporcjonalna do liczby par punktów wszystkich odstępów tego kierunku;
3) waga każdego odstępu danego kierunku jest wprost proporcjonalna do liczby par i odwrotnie proporcjonalna do średniej odległości tego odstępu;
4) waga każdego odstępu danego kierunku jest odwrotnie proporcjonalna do liczby odstępów na tym kierunku.

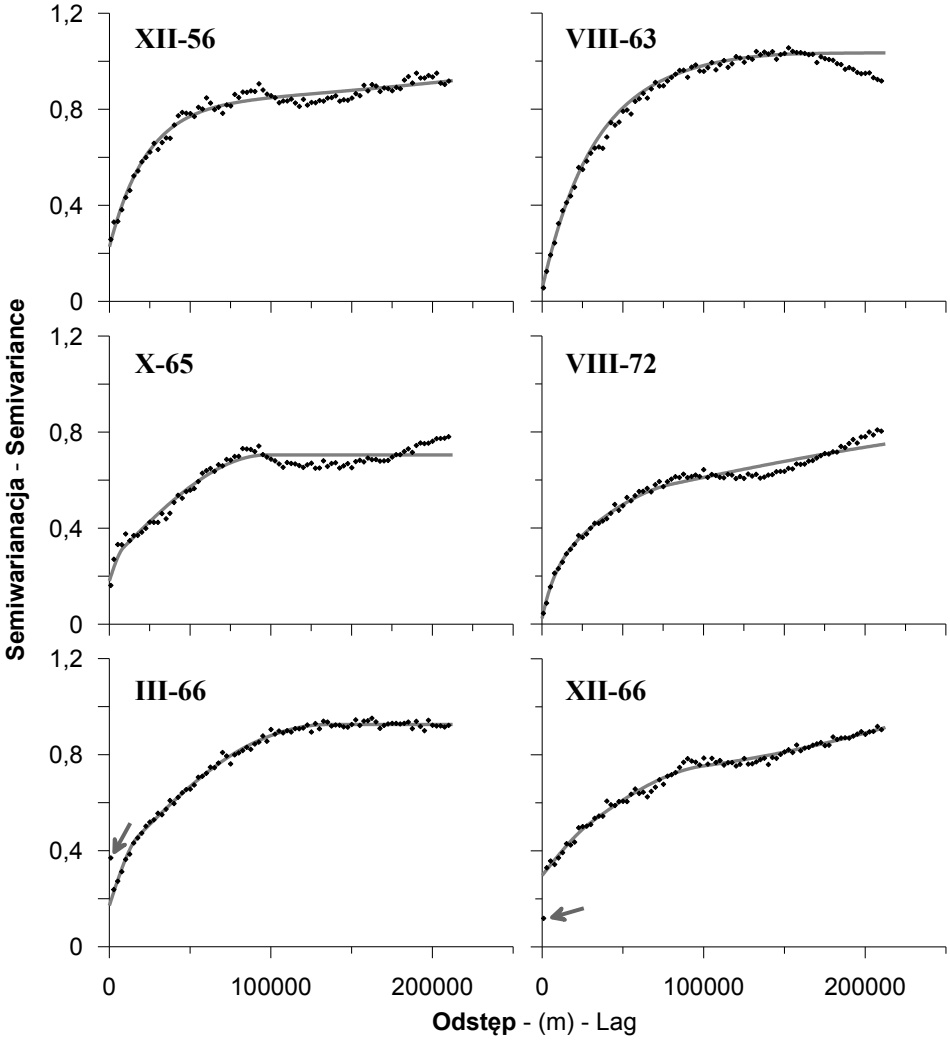
Szczegóły algorytmu optymalizacji wariancji cząstkowych poszczególnych składowych modelu przedstawione są cytowanej publikacji Bleinèsa i in. (2007). W trakcie całego przebiegu prac modelowania semiwariogramów, zarówno danych znormalizowanych jak i kodowanych, używano trzeciego wariantu ważenia.

III.2.5.7. Modelowanie struktury przestrzennej danych znormalizowanych

Tworzenie modeli danych znormalizowanych było stosunkowo proste. Zazwyczaj po kilku próbach z wyborem typu struktury i jej zasięgu uzyskiwano bardzo zadawalający efekt. Nie było żadnej potrzeby ingerencji w automatyczną część całej procedury. Czasami jednak

[22] Oprócz tego wykonano również analizę struktury przestrzennej 208 indywidualnych przypadków rocznych MSDO (roczne MSDO które miały miejsce tego samego dnia). Wyniki te zostaną opublikowane osobno.

nie można było, używając dopuszczalnych typów struktur, uzyskać optymalnego dopasowania. Rozbieżności zazwyczaj jednak nie były duże i dotyczyły wyłącznie miesięcznych danych MSDO. Najczęściej spotykane sytuacje przedstawiono na rycinie 14.



Ryc. 14. Przykłady problemów przy modelowaniu struktury przestrzennej danych znormalizowanych. Szersze omówienie w tekście.

Pierwsza z nich (na ryc. 14, przykład z grudnia 1956), kiedy wartości semiwariogramu wykazywały chaotyczne fluktuacje na krótkich dystansach, zdarzała się najrzadziej i w zasadzie ograniczyła się do kilku miesięcy w roku 1956 i na początku 1957. Był to okres z najmniejszą ilością stanowisk pluwiometrycznych, i z jednocześnie najszybszym tempem ich przyrostu. Oznacza to, że wielu obserwatorów dopiero uczyło się wykonywać pomiary, i chyba to jest przyczyną opisywanych anomalii.

Częściej, bo kilkanaście razy, zanotowano sytuacje ilustrowane przykładami z sierpnia roku 1963, października 1965 i znów sierpnia, tym razem roku 1972 (ryc. 14). Polegały one

występowaniu długodystansowych, raczej regularnych fluktuacji, czasem wyraźnie cyklicznych, zachodzących po osiągnięciu przez semiwariogram poziomu plateau, które można było interpretować jako semiwariancję progową. Świadczyłoby to o występowaniu stosunkowo mało znaczących (o niewielkiej amplitudzie) długodystansowych struktur opadów (o skali 50, 100 i więcej kilometrów), na które „nałożone" były silnie zmienne opady „lokalne". Jeśli posiadane *a priori* informacje wskazują na autentyczną cykliczność zjawiska przestrzennego, tego typu struktura empiryczna jest przedstawiana za pomocą odpowiednich modeli zwanych „*hole effect*", których elementem jest funkcja *cosinus* (Deutsch, Journel 1998, Olea 1999, Webster, Oliver 2001). W odniesieniu do analizowanych w niniejszej pracy zbiorów danych takich informacji nie było, a stosowane do estymacji i symulacji oprogramowanie nie dopuszczało użycia modeli typu „*hole effect*". Dlatego, opisane wyżej fluktuacje były w modelowaniu pomijane. Ich znaczenie związane jest tylko z pytaniem o przypuszczalną genezę, ponieważ stosowanie modeli, w których ich obecności nie uwzględniono, nie wpływa w istotny sposób na jakość uzyskiwanych estymacji i symulacji pola prawdopodobieństwa MSDO.

Dwa ostatnie przedstawione na rycinie 14 przykłady (z marca i grudnia 1966) ilustrują siedem przypadków, kiedy wartości semiwariancji dla tak zwanego „zerowego" odstępu znacznie odbiegają[23] od tendencji zmian autokorelacji wykazywanych przez semiwariancje obliczone dla kolejnych, dalszych odstępów. Odstęp „zerowy" zwany także „połówkowym", obejmuje wszystkie pary punktów pomiarowych, które są od siebie odległe nie więcej niż 1250 m. W rozdziale V.2 omówiono szeroko problem dokładności określenia położenia stanowisk pluwiometrycznych, z których pochodzą analizowane dane MSDO. Lokalizacja ich była znana z dokładnością do 1 minuty kątowej, co oznacza że w „połówkowym" przedziale znalazły się zarówno stanowiska odległe dokładnie o 1 minutę długości geograficznej, ale tylko położone w północnej Polsce oraz tak zwane „duplikaty" (których z kolei było więcej w Polsce południowej). Te drugie miały takie same współrzędne geograficzne (por. roz. V.2). Po przeliczeniu ich do układu współrzędnych płaskich GUGIK 92/19, w celu uniknięcia obecności w zbiorach danych punktów o tej samej lokalizacji, zmieniano losowo wartości współrzędnych prostokątnych tak, aby mieściły się w obszarze znajdującym się w promieniu 0,5 km od lokalizacji „wyjściowej". Par punktów w odstępie „połówkowym" było od ponad siedemdziesięciu w roku 1956 do dziewięciu w roku 1980. Oznacza to znacznie mniejszą wiarygodność statystyczną wyliczonej z nich wartości semiwariancji w porównaniu z

[23] Zaznaczone na rycinie 14, dla zwrócenia uwagi, strzałkami.

odstępami następnymi, gdzie tych par było od kilkuset do kilku tysięcy. Wystąpienie pojedynczej anomalnej wartości opadu mogło w takiej sytuacji znacząco wpłynąć na wynik obliczenia. W dodatku X.5 omówiono procedurę maskowania w trakcie obliczeń semiwariancji danych tzw. „odstających”[24]. Jest ona jednakże obarczona pewną dozą subiektywizmu, co oznacza możliwość pominięcia niektórych przypadków. Podsumowując ów długi i wielowątkowy wywód można napisać, że przedstawione dwa ostatnie przykłady (ryc. 14) związane są z niedokładnościami określenia lokalizacji stanowisk pomiarowych, które były szczególnie istotne przy ich niewielkiej od siebie odległości, oraz ze specyfiką procedury maskowania danych anomalnych. Również w tym przypadku różnice między modelem a danymi empirycznymi są nie istotne. Odstająca wartość semiwariancji odstępu „połówkowego” była ignorowana, a wówczas model idealnie „pasował” do danych kolejnych odstępów. W sumie, wszystkie omówione wyżej „trudne” przypadki stanowiły w trakcie modelowania struktury przestrzennej zbiorów znormalizowanych danych MSDO tylko około 10% całości.

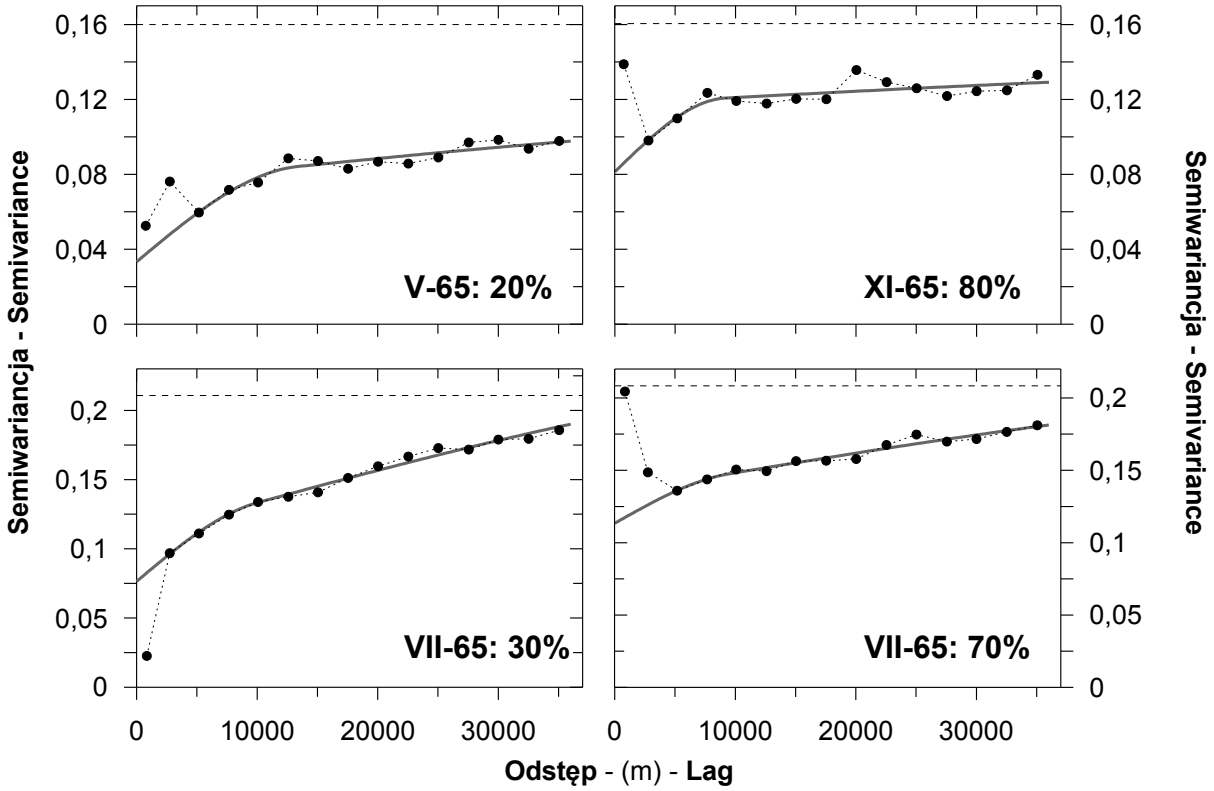
III.2.5.8. Modelowanie struktury przestrzennej danych kodowanych

Procedurę modelowania semiwariogramów danych kodowanych wykonywano analogicznie. Ze względu na 13 wartości progowych (percentyle 1, 5, 10, 20, …., 90, 95, 99) dla każdego z 325 zbiorów danych MSDO = 4225 modeli był to jeden z najbardziej pracochłonnych i czasochłonnych etapów niniejszej pracy.

Z jednej strony procedura ta była prostsza niż omówiona wyżej. Mniejsza liczba uwzględnianych odstępów i krótszy w związku z tym zasięg semiwariogramu, powodował, że zazwyczaj dobrze dopasowany model składał się jedynie z trzech struktur w tym nuggetowej. Dalej, semiwariogramy empiryczne dla „sąsiadujących” ze sobą wartości progowych, szczególnie w środkowym przedziale zakresu rozkładu, zazwyczaj różniły się od siebie niewiele, co również ułatwiało szybkie opracowanie optymalnego modelu.

W trakcie tworzenia modeli struktury przestrzennej danych kodowanych napotykano jednak pewne problemy. Można je pogrupować w trzy klasy. Przykłady takich przypadków przedstawiono na rycinach 15-17.

[24] Dane te (patrz podrozdział V.2) to albo naturalnie występujące anomalie – głównie opady orograficzne, albo błędne pomiary.



Ryc. 15. Przykłady semiwariogramów empirycznych danych kodowanych z odstającymi wartościami pierwszych odstępów i ich modele. Linią przerywaną zaznaczono poziom wariancji danych.

Pierwszy, często spotykany problem, występujący w przy wszystkich wartościach progowych oraz wyraźnie częściej w miesiącach zimowych (I – III), stanowią anomalne układy wartości semiwariancji dla pierwszego lub dwóch pierwszych odstępów (ryc. 15, tab. 1). Odbiegają one wyraźnie od konsekwentnego przebiegu zmian struktury przestrzennej widocznego dla kolejnych, dalszych odstępów. Przyczynę tego zjawiska omawiano już poprzednio w niniejszym podrozdziale. Wynika ono z jednej strony z relatywnie małej liczby par punktów, z której obliczane były wartości semiwariancji dla pierwszych odstępów, oraz z niskiej precyzji lokalizacji stanowisk. Mała liczba danych i niepewność ich klasyfikacji do poszczególnych odstępów powodowała niższą reprezentatywność obliczonej wartości. Kodowanie binarne redukuje co prawda czułość semiwariancji na występowanie ekstremalnych przypadków, ale z drugiej strony przy małej liczbie danych niewielkie jest także prawdopodobieństwo istnienia takich samych proporcji wartości powyżej i poniżej progu, jak w całej populacji. Są to zatem anomalie w pełni „uzasadnione" i stosunkowo łatwe do skorygowania. Nie mają również większego wpływu na jakość modelu. Procedura półautomatycznego modelowania w programie ISATIS umożliwia bowiem wykluczenie z

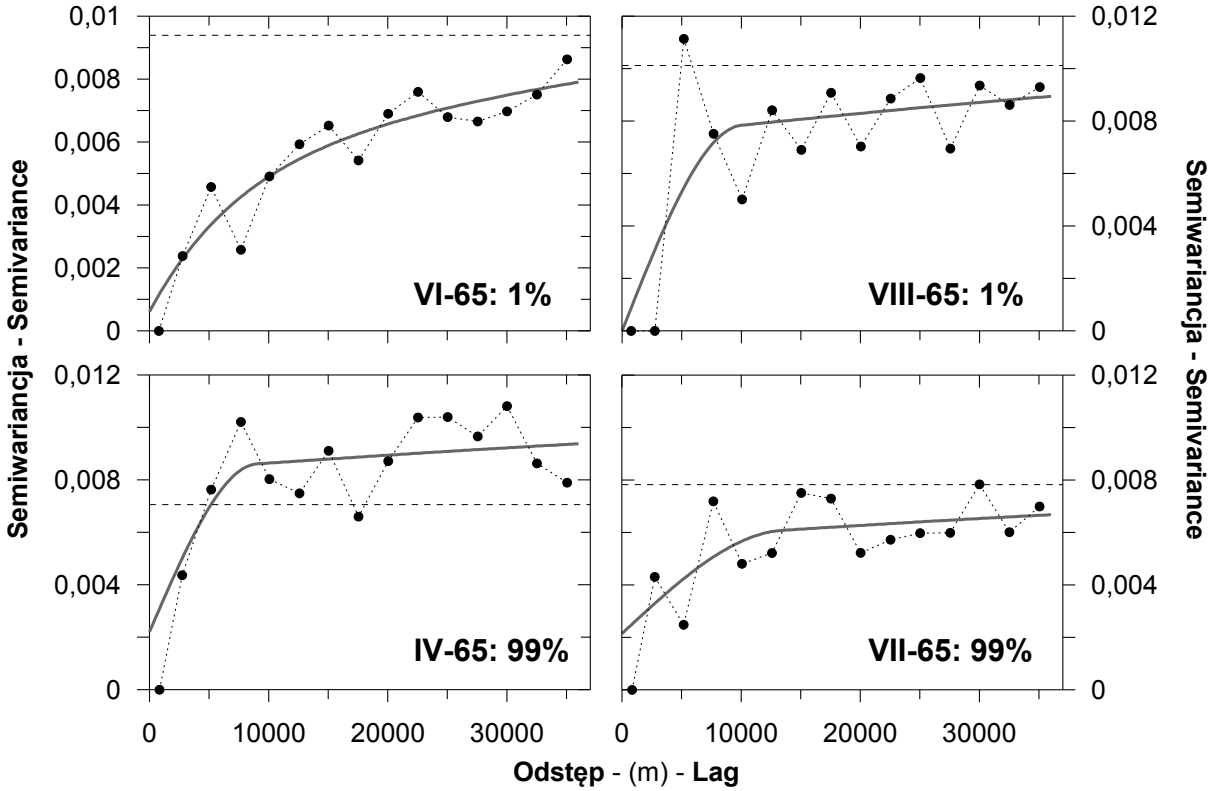
obliczeń optymalizacji wartości poszczególnych odstępów. Selekcji dokonuje się na podstawie kryterium odległości lub ilości par danych.

Tabela 1. Przybliżony procentowy udział semiwariogramów kodowanych z anomalnymi wartościami dla pierwszych odstępów w zależności od miesiąca i wartości progowej (percentyla, P). Objaśnienia: R – roczne MSDO, $\overline{x}$ - wartość średnia. Kolorem oznaczono przedziały wartości: 0 – 5, 6 – 10, a następnie co 10 od 11 do 80.

P	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	R. / Y.	$\overline{x}$
1	12	16	8	12	16	28	20	16	24	20	12	12	0	16,3
5	28	44	32	24	36	32	40	32	24	28	36	40	28	33,0
10	52	36	68	48	36	28	44	20	36	36	44	48	40	41,3
20	56	48	64	36	28	28	48	44	48	32	36	48	36	43,0
30	48	52	68	48	48	36	20	36	36	40	44	40	52	43,0
40	64	44	40	48	44	40	36	28	40	36	40	32	36	41,0
50	68	52	36	36	72	48	52	52	36	40	44	60	32	49,7
60	64	36	56	56	48	52	36	36	36	44	60	40	40	47,0
70	56	64	52	56	40	52	44	36	56	48	60	52	40	51,3
80	44	72	52	72	48	60	44	72	52	48	48	52	40	55,3
90	48	40	44	28	32	44	44	48	52	56	36	36	36	42,3
95	12	20	20	20	20	20	36	32	24	32	12	24	12	22,7
99	12	0	4	8	8	8	4	0	8	12	4	4	8	6,0
$\overline{x}$	43,4	40,3	41,8	37,8	36,6	36,6	36,0	34,8	36,3	36,3	36,6	37,5	37,5	

Tabela 2. Przybliżony procentowy udział semiwariogramów kodowanych z chaotycznymi fluktuacjami wartości dla całego analizowanego zakresu odległości w zależności od miesiąca i wartości progowej (percentyla, P). Objaśnienia: R – roczne MSDO, $\overline{x}$ - wartość średnia. Kolorem oznaczono przedziały wartości: 0 – 5, 6 – 10, a następnie co 10 od 11 do 80.

P	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	R. / Y.	$\overline{x}$
1	56	60	64	76	84	80	68	80	64	60	80	64	76	69,7
5	56	32	44	52	48	48	20	52	56	56	56	48	44	47,3
10	28	32	16	40	24	20	16	8	28	28	32	16	28	24,0
20	8	20	0	4	8	12	8	8	4	8	8	12	4	8,3
30	4	8	0	8	0	4	0	4	4	12	4	8	0	4,7
40	4	0	0	8	0	4	0	4	0	8	0	0	0	2,3
50	4	0	0	0	0	16	0	4	4	4	0	4	0	3,0
60	4	0	0	8	0	4	0	4	4	4	0	0	0	2,3
70	12	0	0	8	4	4	0	0	0	4	4	4	0	3,3
80	20	4	8	8	4	12	8	8	8	4	12	8	4	8,7
90	32	32	20	48	28	24	28	32	20	40	40	28	40	27,7
95	44	40	36	44	44	40	60	44	48	40	32	36	52	39,7
99	40	36	32	40	36	20	48	48	52	32	40	16	28	33,3
$\overline{x}$	24,0	20,3	16,9	26,5	21,5	22,2	19,7	22,8	22,5	23,1	15,1	18,8	18,8	



Ryc. 16. Przykłady chaotycznych semiwariogramów empirycznych danych kodowanych i dopasowane do nich modele. Linią przerywaną zaznaczono poziom wariancji danych.

Tabela 3. Przybliżony procentowy udział semiwariogramów kodowanych nie wykazujących wyraźnej struktury przestrzennej w zależności od miesiąca i wartości progowej (percentyla, P). Objaśnienia: R – roczne MSDO, $\bar{x}$ - wartość średnia. Kolorem oznaczono przedziały wartości: 0 – 5, 6 – 10, a następnie co 10 od 11 do 80.

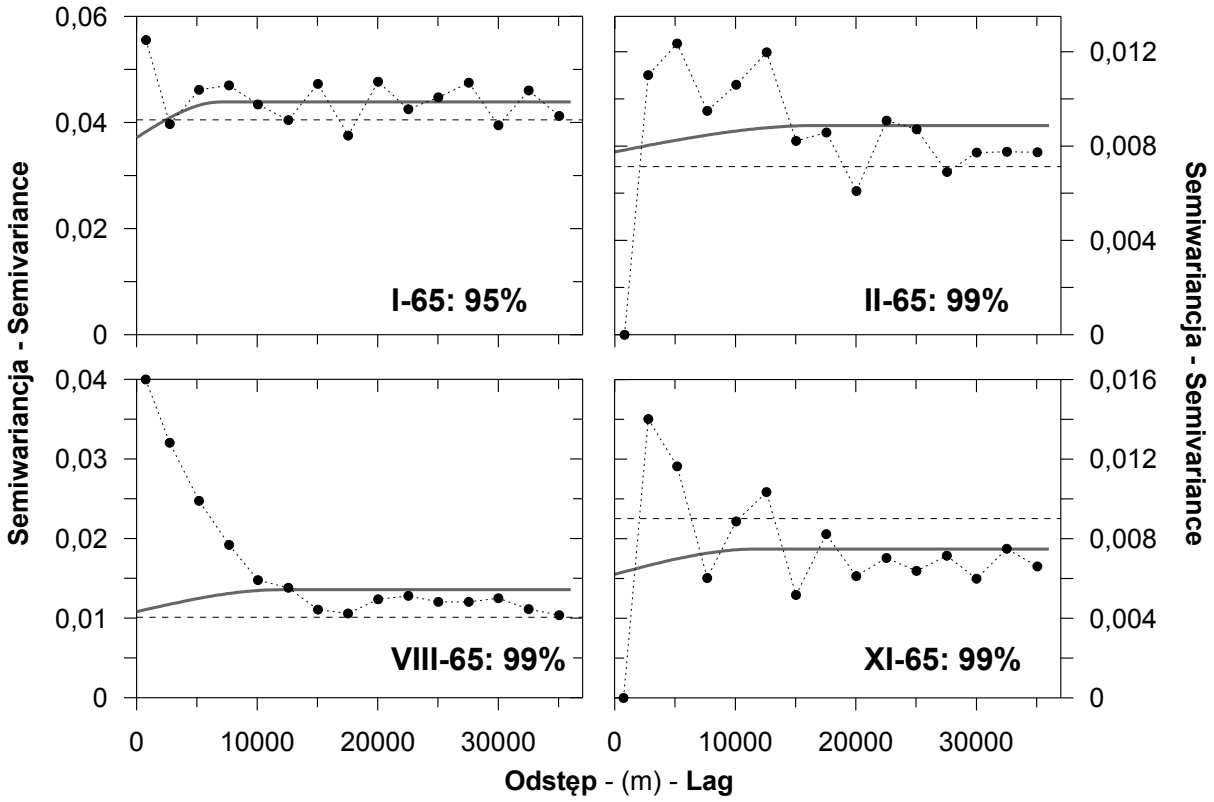
P	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	R. / Y.	$\bar{x}$
1	36	24	28	24	8	8	20	16	24	28	12	36	20	22,0
5	12	20	16	4	0	4	8	4	4	8	0	8	24	7,3
10	4	12	0	0	4	0	0	4	8	4	4	16	4	4,7
20	0	0	0	0	0	0	0	0	0	0	0	4	0	0,3
30	0	0	0	0	0	0	0	0	0	0	0	4	0	0,3
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0
80	4	0	0	4	4	0	0	0	0	0	0	4	4	1,3
90	8	4	16	12	8	0	8	8	8	4	12	32	8	10,0
95	36	32	40	48	28	24	8	20	20	40	44	44	16	32,0
99	48	60	60	56	56	64	40	52	40	60	60	80	60	56,3
$\bar{x}$	11,4	11,7	12,3	11,4	8,3	7,7	6,5	8,0	8,0	11,1	10,2	17,5	10,5	

Drugi typ problemów przy modelowaniu semiwariogramów danych kodowanych występował przy skrajnych progach, najczęściej przy 1 i 5 percentylu (ryc. 16 i tab. 2). Nie wykazywał on tym razem zmienności sezonowej, a polegał na chaotycznych fluktuacjach wartości semiwariancji w całym zakresie analizowanych odległości. Widoczne jest istnienie struktury przestrzennej, to jest generalnie semiwariancja rośnie wraz ze przyrostem odległości pomiędzy porównywanymi danymi, ale trudno zidentyfikować typ struktury i jej zasięg. Jest to oczywiście efekt braku stabilności dwupunktowej statystyki w sytuacji małej ilości danych poniżej / powyżej progu i ich raczej rozproszonego rozmieszczenia. Takie układy są zatem częstsze w miesiącach z większą ilością opadów typu konwekcyjnego o stosunkowo małym zasięgu przestrzennym. W wyborze typu struktur i ich zasięgu w trakcie modelowania takich przypadków kierowano się zarówno widocznym na wykresie kształtem semiwariogramu empirycznego, jak i charakterem modelu dla sąsiedniej wartości progowej. Wychodzono bowiem z założenia, że w większości przypadków zmiany struktury przestrzennej pomiędzy poszczególnymi progami dokonują się w sposób „płynny".

Trzeci, najbardziej kłopotliwy typ semiwariogramów danych kodowanych związany był głównie z najwyższymi wartościami progowymi – 95 i 99 percentylem (ryc. 17, tab. 3). Więcej było też takich przypadków pomiędzy październikiem a kwietniem. Wykazywały one brak struktury przestrzennej albo wręcz zmniejszanie się semiwariancji wraz ze wzrostem odległości, zwłaszcza dla kilku pierwszych odstępów (przykład z sierpnia 1965 dla 99 percentyla). Średnie wartości semiwariancji dla wszystkich odstępów w całym uwzględnianym zakresie odległości pomiędzy stanowiskami były często wyższe od wariancji próby. Tego typu semiwariogramy empiryczne przy braku jakichkolwiek informacji *a priori* o charakterze zmienności przestrzennej analizowanej cechy należałoby modelować używając pojedynczej struktury – nuggetowej. Oznaczałoby to, że najwyższe sumy dobowe dla danego zbioru danych MSDO występowały całkowicie losowo. Praktyka modelowania w opisywanych wyżej przypadkach polegała na całkowitej rezygnacji z automatycznej optymalizacji parametrów. Używane do estymacji i symulacji oprogramowanie (Ying 2000), nie dopuszcza wykorzystania modelu nuggetowego jako jedynego do stosowania samodzielnie (pojedynczo) do opisu struktury przestrzennej. Dlatego za każdym razem konieczne było „sztuczne" dodanie składowej „regularnej" – zazwyczaj sferycznej – o bardzo niewielkiej wariancji cząstkowej. Takie postępowanie umożliwiało ominięcie numerycznych ograniczeń stosowanych algorytmów, bez znaczącego zniekształcenia wyników estymacji i symulacji. Dodając „sztuczną" składową należało także podjąć decyzję o jej zasięgu. Brano

wówczas pod uwagę zarówno szczegóły kształtu modelowanego chaotycznego semiwariogramu, jak i zasięg modeli opracowanych dla bardziej regularnych semiwariogramów sąsiednich, niższych wartości progowych. Należy jednak brać pod uwagę, szczególnie przy interpretacji zmienności uwarunkowanej względnym zróżnicowaniem wysokości sumy opadów, że struktura przestrzenna MSDO interpretowana na podstawie modeli danych kodowanych jest przy ich górnej granicy obciążona dużą dozą subiektywizmu.

Ocena liczby semiwariogramów danych kodowanych, które można by zakwalifikować do trzech opisanych wyżej grup jest nieco utrudniona (tab. 1-3). Wynika to przede wszystkim z płynnej, subiektywnej oceny stopnia regularności / chaotyczności semiwariogramu (odróżnienie grupy 2). Również jednoznaczne rozdzielenie przypadków należących do grupy drugiej i trzeciej nie było w wielu przypadkach możliwe, dlatego też podane niżej odsetki należy traktować jako przybliżone. Do grupy pierwszej (anomalne wartości dla pierwszych odstępów) zaliczono 34,9% przypadków modelowanych semiwariogramów danych kodowanych, do grupy drugiej (semiwariogramy chaotyczne) – 19,5%, a do trzeciej (brak struktury) – 9,5%.



Ryc. 17. Przykłady semiwariogramów empirycznych danych kodowanych nie wykazujących wyraźnej struktury oraz ich modele. Linią przerywaną zaznaczono poziom wariancji danych.

III.2.5.9. Nieciągłość i asynchroniczność danych MSDO a ich struktura przestrzenna

Przy omawianiu szczegółów estymacji metodą krigingu danych kodowanych (dodatek X.2) wspomniano o problemie naruszania relacji porządkowych, polegającym głównie na braku konsekwentnego następstwa szacowanych prawdopodobieństw dla kolejnych wartości progowych odczytywanych ze skumulowanych funkcji rozkładu (*cdf*). W celu zmniejszenia możliwości zaistnienia takich wyników zaleca się stosowanie przy modelowaniu zawsze tej samej kombinacji modeli podstawowych, a ich parametry powinny przy kolejnych wartościach progowych zmieniać się stopniowo. Takie postępowanie ma pełne uzasadnienie w przypadku struktury przestrzennej synchronicznych zmiennych (cech) ciągłych. W kolejnych klasach wielkości / natężenia takich cech zmienia się ona niewiele i raczej w sposób „płynny". Sumy dobowe opadów, a tym bardziej ich podzbiór wykorzystywany w niniejszej pracy, takiej natury nie mają. Z jednej strony opady są nieciągłe przestrzennie, z drugiej – dane MSDO dla poszczególnych miesięcy i lat pochodzą z różnych terminów, czyli są niesynchroniczne. Zróżnicowana jest także ich geneza – powstają w efekcie działania kilku zjawisk operujących w innych skalach przestrzennych i czasowych. Dlatego, nie można było ignorować faktu, często spotykanego w trakcie niniejszej pracy, że semiwariogramy empiryczne dla kolejnych wartości progowych zmieniały się czasami dość znacznie, raczej skokowo niż „płynnie".