

Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie

# Geostatystyka – eksploracja danych

Zastosowanie GIS w badaniach przyrodniczych, Ćwiczenie 12

Na podstawie materiałów szkoleniowych ESRI.  
Wyłącznie do użytku wewnętrznego AGH.

2021-10-26

## Wprowadzenie

Jeśli nie wykonałeś dotąd pierwszego etapu ćwiczenia z działu Geostatystyka, przejdź do wykonania ćwiczenia [11](#).

W poprzednim ćwiczeniu zapoznaliśmy się z działaniem kreatora geostatystycznego oraz z domyślnymi ustawieniami analiz geostatystycznych, wykonaliśmy dla rejonu stanu Kalifornia, prostą, niezbyt dokładną mapę predykcji zawartości ozonu w atmosferze.

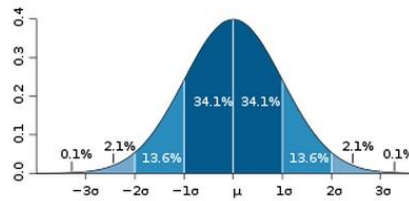
W tym etapie ćwiczenia dokonamy szczegółowej eksploracji danych. Proces analiz geostatystycznych pokazany na zakończenie etapu 1-szego ćwiczenia sugeruje, że aby umożliwić podejmowanie lepszych decyzji podczas geostatystycznej estymacji powierzchni, należy najpierw szczegółowo zbadać posiadane dane. Podczas eksploracji danych będziemy poszukiwać oczywistych błędów w wartościach danych próbkowych, które mogą wpłynąć na jakość estymowanej powierzchni geostatystycznej, przeanalizujemy rozkład danych, będziemy poszukiwać generalnych trendów oraz istnienia zależności zmienności parametru od kierunku obserwacji.

Moduł *Geostatistical Analyst* oferuje wiele narzędzi do eksploracji danych. W tym etapie ćwiczenia:

- zbadamy rozkład danych,
- zidentyfikujemy w swoich danych ewentualne trendy,
- zajmiemy się zrozumieniem czym są autokorelacja przestrzenna i wpływy kierunkowe.

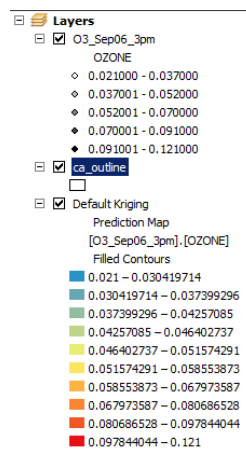
### 1. Analiza rozkładów danych z zastosowaniem narzędzia Histogram

Metody geostatystyczne wykorzystywane do interpolacji powierzchni geostatystycznych dają najlepsze wyniki, jeśli dane posiadają rozkłady normalne lub przynajmniej zbliżone do normalnego (krzywa rozkładu posiada kształt dzwonu Gaussa – [Fig. 1](#)). Jeśli dane wykazują większą asymetrię (jeden z ogonów rozkładu jest znacząco dłuższy od drugiego), przed przystąpieniem do analizy geostatystycznej należy je doprowadzić do postaci umożliwiającej przeprowadzenie wiarygodnych analiz. Wykonanie analiz geostatystycznych na danych znacząco odbiegających od rozkładu normalnego skutkuje poważnymi błędami. Dlatego ważne jest, aby przed przystąpieniem do generowania powierzchni zmienności cechy, zapoznać się z rozkładami danych. W ArcGIS istnieje narzędzie *Histogram*, które wyświetla histogramy częstości dla atrybutów zbiorów danych. Umożliwia ono zbadanie rozkładów jednej zmiennej. W kolejnym kroku ćwiczenia zapoznamy się z rozkładem ozonu w warstwie 03\_Sep06\_3pm.



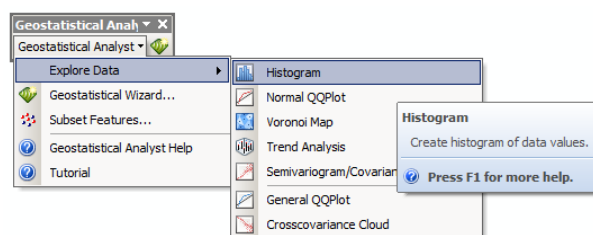
**Fig. 1. Krzywa dzwonowa rozkładu Normalnego**

- 1.1. Jeśli zamknąłeś poprzednią sesję ArcMap, uruchom program ponownie i otwórz plik projektu Ozone Prediction Map.mxd.
- 1.2. Kliknij warstwę `ca_outline` i przeciągnij ją w tabeli zawartości pod warstwę `O3_Sep06_3pm` (Fig. 2).



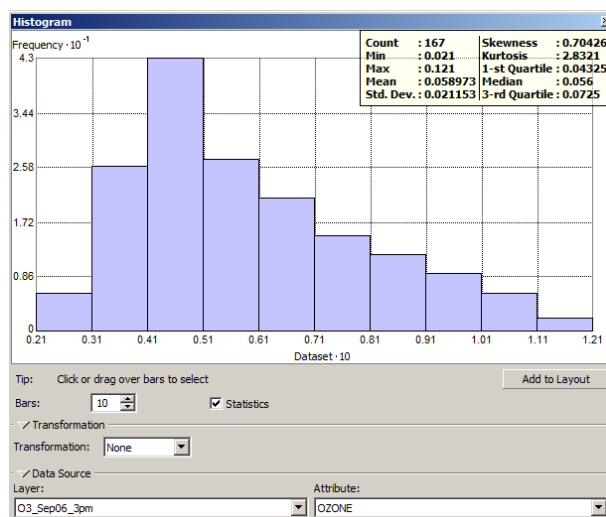
**Fig. 2. Tabela zawartości projektu**

- 1.3. Kliknij warstwę `O3_Sep06_3pm`, aby ją wybrać.
- 1.4. Na pasku narzędzi *Geostatistical Analyst* kliknij pozycję *Geostatistical Analyst > Explore Data > Histogram* (Fig. 3).



**Fig. 3. Wybór narzędzia *Histogram* na pasku narzędzi *Geostatistical Analyst***

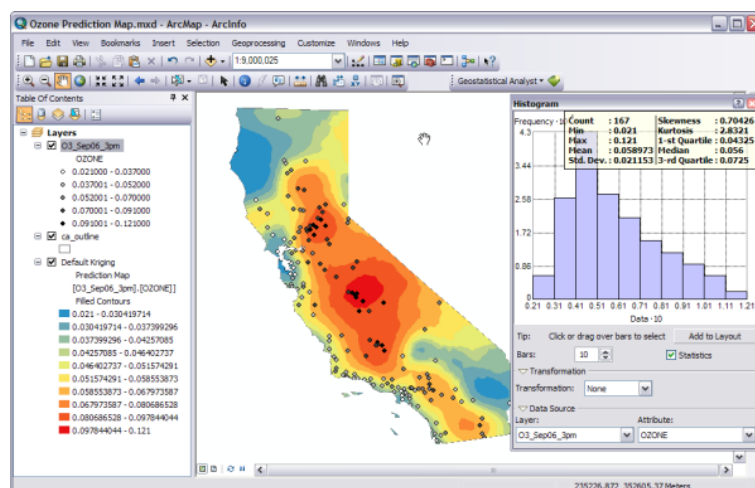
- 1.5. W oknie dialogowym *Histogram*, w polu *Atrybut (Attribute)* wybierz zmienną przechowującą wartości ozonu pomierzone przez stanowe stacje monitoringu atmosfery (`OZONE`, Fig. 4).



**Fig. 4. Histogram przedstawiający rozkład danych atrybutu OZONE**

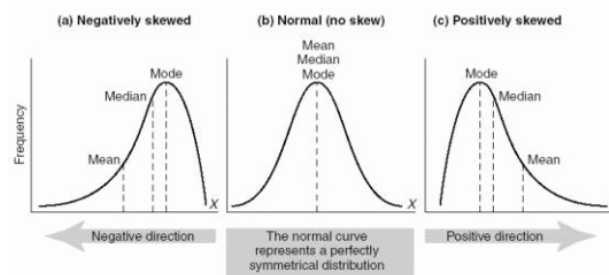
Aby ułatwić odczyt danych, wartości osi OX zostały przeskalowane o współczynnik 10.

- 1.6. Aby umożliwić jednoczesny dostęp do mapy i histogramu, zmień rozmiar okna i/lub przesunij okno dialogowe *Histogram* (Fig. 5).



**Fig. 5. Scena ArcMap umożliwiającą jednoczesną obserwację mapy i histogramu**

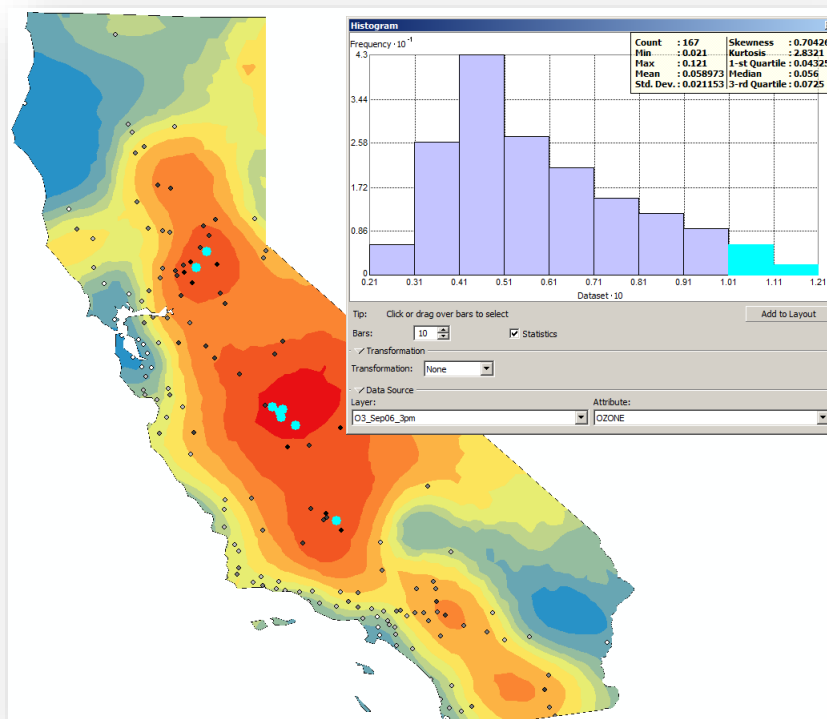
Histogram z Fig. 4 przedstawia rozkład zawartości ozonu w zbiorze danych wejściowych. Przedział zmienności parametru został podzielony na 10 klas. Częstość danych w każdej klasie jest reprezentowana przez wysokość każdego słupka. Ogólnie rzecz biorąc, ważnymi cechami rozkładów są: klasa modalna, zakres zmienności i symetria. Szybkim sposobem na sprawdzenie asymetrii rozkładu jest przeanalizowanie, czy średnia i mediana są w przybliżeniu takie same. Jeśli tak jest, mamy jeden dowód na to, że dane mogą mieć rozkład normalny (Fig. 6).



**Fig. 6. Zależność estymatorów wartości przeciętnej (średniej arytmetycznej, mediany i mody) w rozkładach lewoskośnym ( $M < Me < Mo$ ), symetrycznym ( $M \approx Me \approx Mo$ ) i prawoskośnym ( $M > Me > Mo$ )**

Histogram zawartości ozonu wskazuje, że dane mają charakter jednomodalny (jeden słupek posiada wartości maksymalne) i są prawoskośne (in. prawoasymetryczne, dłuższy prawy ogon rozkładu). Prawy ogon rozkładu wskazuje na obecność stosunkowo niewielkiej liczby elementów próby o wysokich wartościach stężenia ozonu. Wydaje się, że dane nie posiadają rozkładu zbliżonego do rozkładu normalnego.

- 1.7. Klikając i przeciągając wskaźnik nad słupkami histogramu, wybierz dwa z nich o wartościach ozonu większych niż 0,10 ppm (pamiętaj, że wartości zostały przeskalowane o współczynnik 10). Punkty poboru próbek z tego zakresu są zaznaczone na mapie. Zauważ, że większość z nich jest położona w kalifornijskiej Dolinie Centralnej (Fig. 7).



**Fig. 7. Scena ArcMap z zaznaczonymi miejscami opróbowania należącymi do dwóch ostatnich klas rozkładu zawartości ozonu w atmosferze**

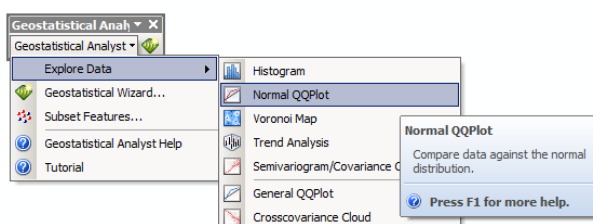
- 1.8. Aby wyczyścić wybrane punkty na mapie i histogramie, wybierz polecenie *Wyczyść wybrane zaznaczenie* (*Clear Selected Features*) znajdujące się na pasku narzędzi *Narzędzia* (*Tools*).
- 1.9. Kliknij przycisk *Zamknij* znajdujący się w górnym rogu okna dialogowego *Histogram*.

## 2. Analiza normalności rozkładu empirycznego za pomocą wykresu QQ

Wykres kwantyl-kwantyl (*quantile-quantile* – *QQ*) jest wykresem 2D, w którym względem siebie wykreślane są kwantyle<sup>1</sup> dwóch rozkładów. Służy on do porównywania rozkładu danych empirycznych z wybranym rozkładem teoretycznym (np. normalnym). Analiza wykresu QQ stanowi kolejną miarę normalności rozkładu danych próbkowych. Na dwóch osiach współrzędnych odkładane są wartości kwantyli rozkładu teoretycznego (np. normalnego) i rozkładu empirycznego (tu – rozkładu ozonu). Kiedy mamy do czynienia z pełną zgodnością obu rozkładów, poszczególne wartości kwantyli obu analizowanych rozkładów są sobie równe, dlatego wykres QQ będzie linią prostą nachyloną do osi OX pod kątem 45°. Opisywaną właściwość można wykorzystać do testowania zgodności dwóch rozkładów lub odchylenia rozkładu empirycznego od rozkładu teoretycznego.

Porównując kwantyle zawartości ozonu w atmosferze z kwantylami rozkładu Normalnego będziemy mogli określić jak bardzo rozkład zmiennej empirycznej odbiega od rozkładu Normalnego.

- 2.1. Na pasku narzędzi *Geostatistical Analyst* wybierz *Geostatistical Analyst* > *Explore Data* > *Normal QQ Plot* (Fig. 8).



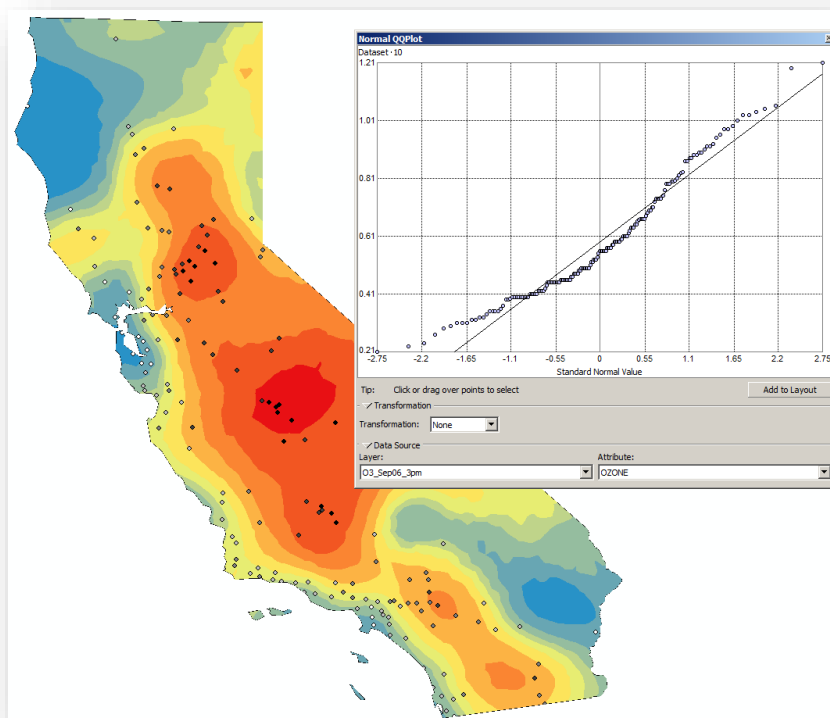
**Fig. 8. Wybór narzędzia *Normal QQ Plot* na pasku narzędzi *Geostatistical Analyst***

- 2.2. Z listy rozwijanej wybierz atrybut *OZONE*.

---

<sup>1</sup> Kwantyle to wartości, które dzielą przypadki populacji na pewną liczbę grup tej samej wielkości. Kwantyl rzędu  $q$  ( $0 < q < 1$ ) w populacji jest taką liczbą  $x_q$ , że  $q \times 100\%$  elementów tej populacji ma wartość badanej cechy nie większą od  $x_q$ . Np. kwantyl rzędu  $1/2$  ( $q = 1/2$ ) w populacji to taka liczba, że  $(1/2 \times 100\% = 50\%)$  elementów tej populacji ma wartość badanej cechy nie większą od  $x_{1/2}$ .

Na podstawie wykresu QQ (Fig. 9) można zauważyć, że ciąg punktów jest bardzo zbliżony do prostej. Największa odchyłka od tej linii następuje przy niskich wartościach stężenia ozonu. Jeśli dane nie wykazują rozkładu normalnego na histogramie ani na wykresie QQ, może okazać się konieczne przekształcenie danych w taki sposób, aby przed użyciem techniki krigingu były zgodne z rozkładem normalnym.

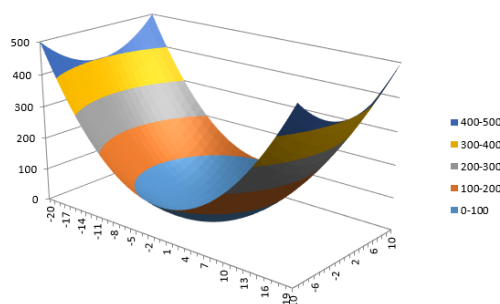


**Fig. 9. Scena ArcMap z wykresem QQ**

- 2.3. Kliknij przycisk *Zamknij* znajdujący się w górnym rogu okna dialogowego *Normal Q-Q Plot*.

### 3. Identyfikacja generalnych trendów w zmienności danych

Całkowitą zmienność parametru w populacji próby można podzielić na część zmienności, za którą odpowiada tendencja (regionalny trend) oraz część zmienności występującą w niewielkiej skali (zmienność lokalna). Czasami ta tendencja występuje, a czasami nie. Jeżeli w danych przestrzennych występuje jakaś tendencja, jest to dowód na obecność nielosowego (deterministycznego) składnika zmienności parametru. Taką zmienność powierzchniową można opisać za pomocą funkcji matematycznej, np. delikatnie opadające zbocze można przybliżyć za pomocą wzoru płaszczyzny, zaś dolinę – za pomocą nieco bardziej złożonej funkcji tworzącej powierzchnię przypominającą literę U (wielomianu drugiego rzędu; Fig. 10).



**Fig. 10. Przykładowy wykres wielomianu drugiego rzędu aproksymujący powierzchnię doliny**

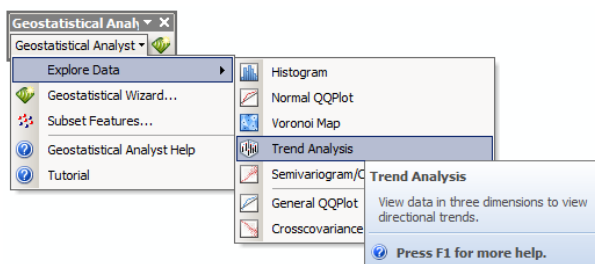
Najczęściej jest tak, że powierzchnie aproksymowane funkcjami matematycznymi, ze względu na daleko idące uproszczenia, nie są w stanie oddać skomplikowanej zmienności przestrzennej parametrów i tym samym nie nadają się do przedstawianych zjawisk przestrzennych.

W nawiązaniu do I-go prawa geografii Toblera (Tobler 1970), dla geostatystycznej estymacji wartości parametrów, największe znaczenie mają punkty o znanych wartościach tego parametru położone w pewnej niewielkiej odległości od miejsca estymacji. Im większą dokładnością będą się cechowały punkty opróbowania, tym lepsze będą rezultaty estymacji. Z tą prostą obserwacją związana jest nieodłączna część analiz geostatystycznych – analiza trendów.

Skoro na całkowitą zmienność populacji próby składają się: zmienność regionalna (trend) i poszukiwana przez nas zmienność lokalna, Dobrym pomysłem byłoby usunięcie części zmienności powodowanej istnieniem trendu. W wyniku operacji odejmowania otrzymujemy tzw. reszty (ang. *residuals*), które następnie poddawane są procedurze geostatystycznej. Podczas modelowania reszt będziemy analizowali zmienność parametru w niewielkich (lokalnych) fragmentach estymowanej powierzchni. Narzędzie analizy trendów umożliwia identyfikację obecności/braku trendów w zbiorze danych wejściowych i określenie, stopnia wielomianu, który najlepiej opisuje występujący trend. Po procesie geostatystycznej estymacji reszt, do otrzymanej powierzchni geostatystycznej należy z powrotem dołączyć powierzchnię trendu aby ostatecznie otrzymać obraz całkowitej zmienności analizowanego parametru.

- 3.1. Na pasku narzędzi *Geostatistical Analyst* wybierz *Geostatistical Analyst > Explore Data > Trend Analysis* (Fig. 11).

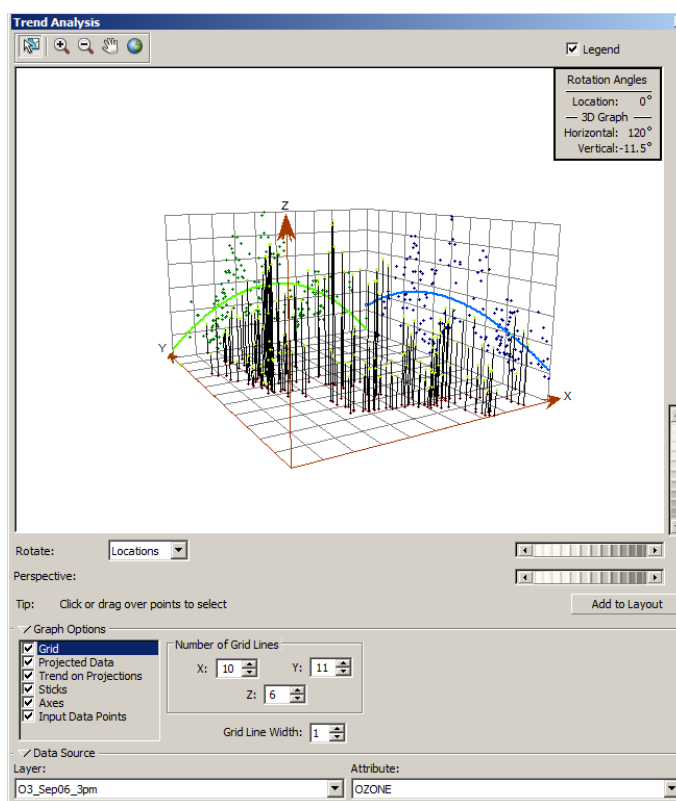




**Fig. 11. Wybór narzędzia *Trend Analysis* na pasku narzędzi *Geostatistical Analyst***

3.2. Z listy rozwijanej wybierz atrybut `OZONE`.

Każdy pionowy słupek na wykresie analizy trendów przedstawia lokalizację punktu pomiarowego i wartość (wysokość) każdego pomiaru ozonu. Punkty danych są rzutowane na prostopadłych do siebie płaszczyznach wschód – zachód i północ – południe. Poprzez rzutowane punkty poprowadzono krzywe (wielomiany), które stanowią najlepsze dopasowanie do rzutowanych punktów. Stanowią one obrazy regionalnych trendów występujących w zmienności zawartości ozonu w konkretnych kierunkach. Jeśli linia jest płaska, oznacza to brak tendencji. Jeśli spojrzymy na [Fig. 12](#) na jasnozieloną linię, widać, że zaczyna się ona niskimi wartościami, wzrasta w miarę przesuwania się ku środkowi osi X, a następnie maleje. Podobnie wygląda zmienność w płaszczyźnie prostopadłej, zobrażowana linią niebieską. Wartości ozonu rosną wzdłuż niej w kierunku północnym, a po przekroczeniu środka stanu Kalifornia zaczynają spadać. Pokazuje to, że dane wydają się wykazywać silną tendencję malejącą od środka stanu we wszystkich kierunkach.



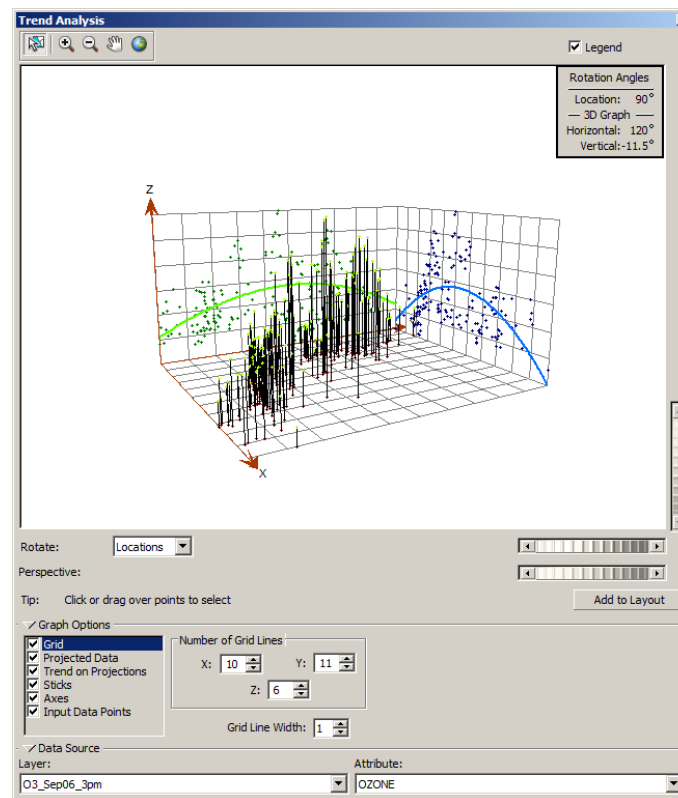
**Fig. 12. Wykres analizy trendów**

- 3.3. Kliknij pasek przewijania *Obróć Lokalizacje* (*Rotate Locations*) i przewiń w lewo, aż kąt obrotu osiągnie  $90^\circ$  (Fig. 13).



**Fig. 13. Paski modyfikacji rotacji lokalizacji i perspektywy**

Widać, że podczas obracania punktów, tendencje zawsze wykazują odwrócony kształt litery U (Fig. 14). Ponadto, dla dowolnego określonego kąta obrotu, tendencja nie wydaje się być silniejsza (bardziej wyraźny kształt U). Potwierdza to wcześniejszą obserwację, że istnieje silna tendencja od środka domeny danych we wszystkich kierunkach. Ponieważ trend posiada kształt litery U, wielomian drugiego rzędu będzie dobrym wyborem do zastosowania go jako modelu globalnego trendu. Obserwowany trend może wynikać z faktu, że w rejonie wybrzeża populacja ludzi jest stosunkowo niewielka. Najbardziej zaludnionym obszarem jest Dolina Centralna, z którą rozpoczynają się słabo zaludnione obszary gór. Usunięciem trendu zajmiemy się w części 4-tej ćwiczenia.

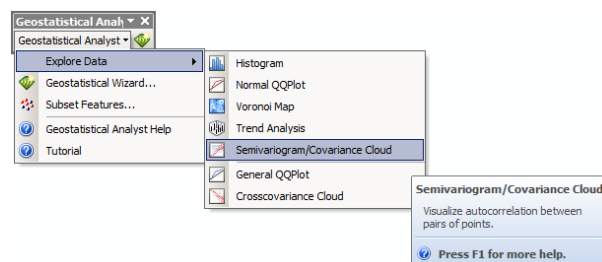


**Fig. 14. Wykres analizy trendów, lokalizacje obrócone o kąt 90°**

- 3.4. Kliknij przycisk *Zamknij* znajdujący się w górnym rogu okna dialogowego *Analiza Trendów (Trend Analysis)*.

#### 4. Poszukiwanie autokorelacji przestrzennej i wpływów kierunkowych

- 4.1. Na pasku narzędzi *Geostatistical Analyst* wybierz *Geostatistical Analyst > Explore Data > Cloud Semivariogram/Covariance Cloud* (Fig. 15).



**Fig. 15. Wybór narzędzia *Semivariogram/Covariance Cloud* na pasku narzędzi *Geostatistical Analyst***

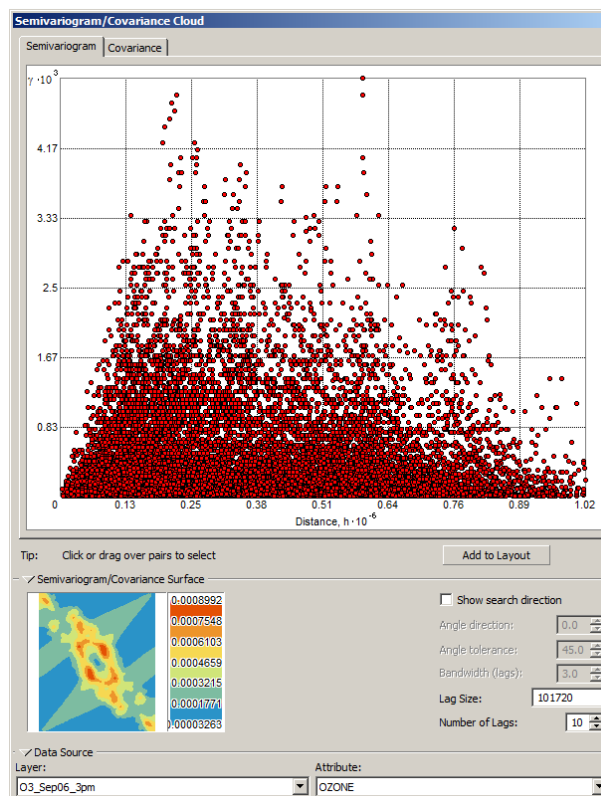
- 4.2. Z listy rozwijanej wybierz atrybut *OZONE*.

*Chmura semiwariogramu/kowariancji (Semivariogram/Covariance Cloud)* pozwala zbadać autokorelację przestrzenną pomiędzy wartościami zmiennej w punktach opróbowania. Cytowane już tutaj 1-sze prawo geografii Toblera zakłada, że cechy przestrzennie bliskie sobie są do siebie podobne. *Chmura semiwariogramu/kowariancji*

pozwała badać ten związek. Analiza wymaga skonstruowania wykresów, w których na osi OY odkładane są wartości semiwariogramów – kwadraty różnic wartości parametrów w parach punktów opróbowania (zob.: [ćwiczenie 6.14](#), [Wz. 1](#)), zaś na osi OX odległości dzielące każdą parę punktów.

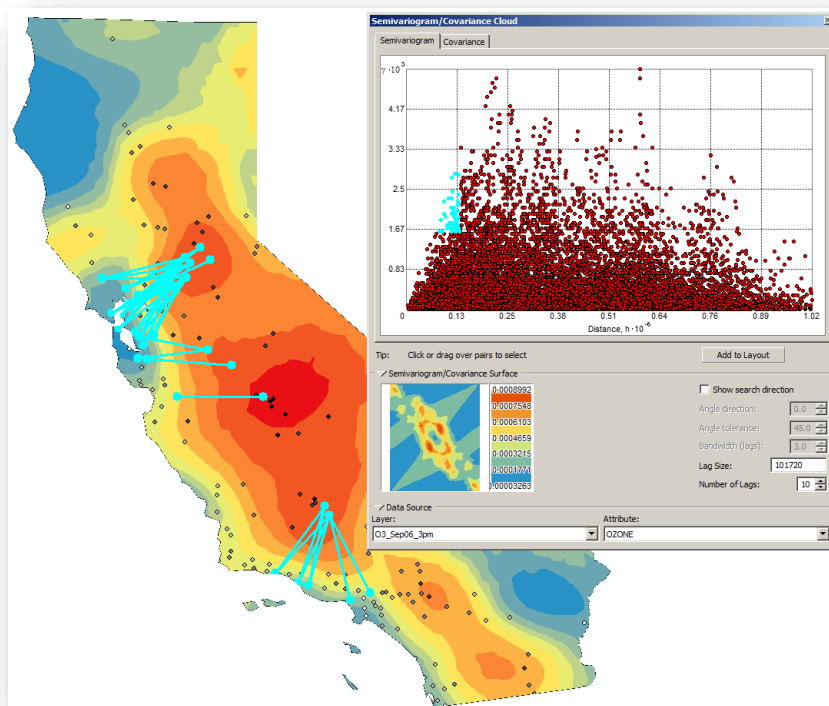
Każda czerwona kropka w *chmurze semiwariogramu/kowariancji* reprezentuje parę punktów opróbowania (tutaj – stacji monitoringu atmosfery; [Fig. 16](#)). Ponieważ miejsca położone blisko siebie powinny być do siebie bardziej podobne, lokalizacje położone na wykresie najbliżej siebie (lewy koniec osi OX) powinny mieć małe wartości semiwariogramu (niskie wartości na osi OY). W miarę wzrostu odległości pomiędzy parami punktów opróbowania (przesuwanie się w prawo na osi OX), wartości semiwariogramu powinny również wzrastać (wzdłuż osi OY). Jednakże po osiągnięciu pewnej granicznej odległości, chmura punktów spłaszczyła się, wskazując, że zawartości ozonu w parach punktów oddzielonych więcej niż tą odległość nie są już skorelowane.

Jeśli okazuje się, że niektóre lokalizacje danych są położone blisko siebie (blisko zera na osi OX) i mają względnie wyższe wartości semiwariogramu niż można byłoby się spodziewać (wysokie wartości na osi OY), należy zbadać te pary punktów, aby sprawdzić, czy istnieje możliwość, że dane zawierają jakieś błędy o typie prób huraganowych (ang. *outliers*).



**Fig. 16. Okno dialogowe *Chmura semiwariogramu/kowariancji***

- 4.3. Wybierz polecenie *Wybierz Cechy wg. Prostokąta (Select Features by Rectangle)* znajdujący się na pasku narzędzi *Narzędzia (Tools)*, a następnie w oknie dialogowym *Semivariogram/Covariance Cloud* zaznacz niektóre punkty o dużych wartościach semiwariogramu (oś OY). Nie jest ważne, aby wybrać dokładnie te same punkty, co na Fig. 17.

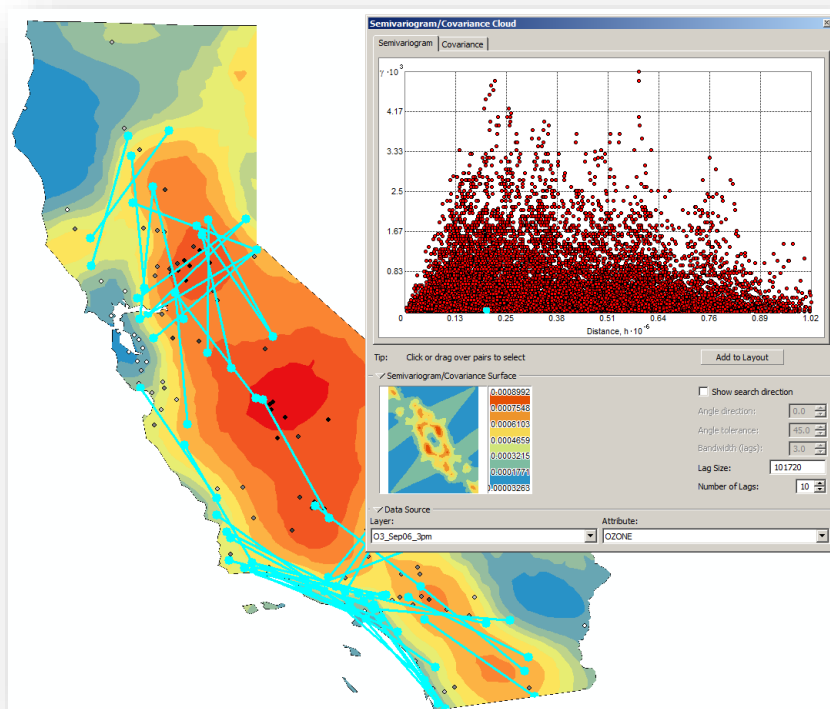


**Fig. 17. Okno dialogowe *Chmura semiwariogramu/kowariancji* z wybranymi parami punktów oddalonymi od siebie o odległość mniejszą od progu autokorelacji<sup>2</sup> i cechujących się wysokimi wartościami semiwariancji**

Lokalizacje par próbek, które są wybrane w oknie semiwariogramu, są zaznaczone na mapie. Punkty na mapie łączą linie wskazujące na parowanie. Jak można się było spodziewać na podstawie mapy predykcji zawartości ozonu wygenerowanej z ustawieniami domyślnymi, linie o wysokich wartościach semiwariogramu dla określonego zakresu odległości pomiędzy punktami opróbowania, odpowiadają największym gradientom zawartości ozonu.

Diagram Fig. 18 przedstawia pary punktów opróbowania o jednej wartości semiwariogramu oddalonych od siebie o tę samą odległość.

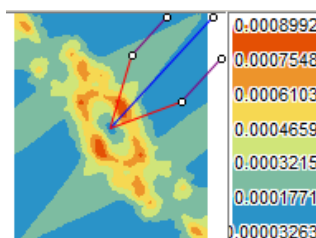
<sup>2</sup> Próg autokorelacji – odległość pomiędzy parami punktów opróbowania w obrębie której dochodzi do zjawiska autokorelacji, czyli związku korelacyjnego pomiędzy wartościami pomiarów w punktach. Po przekroczeniu progu autokorelacji zależności pomiędzy wartościami analizowanego parametru mają charakter losowy.



**Fig. 18. Okno dialogowe *Chmura semiwariogramu/kowariancji* z wybranymi parami punktów oddalonymi od siebie o tę samą odległość**

Większość linii jest mniej więcej równoległa do linii wybrzeża. Jest to dowód na istnienie wpływu kierunku obserwacji na zmienność zawartości ozonu. W następnym ćwiczeniu zajmiemy się uwzględnieniem wpływów kierunkowych podczas modelowania. *Geostatistical Analyst* dostarcza narzędzi do uwzględnienia go w procesie tworzenia powierzchni interpolacyjnych. Aby zbadać wpływ kierunku w chmurze wartości semiwariogramów, użyj narzędzi *Poszukiwanie Kierunku* (*Search Direction*).

- 4.4. Zaznacz opcję *Pokaż wyszukaj kierunek* (*Show search direction*).
- 4.5. Kliknij i przesuń wskaźnik kierunku pod dowolnym kątem (Fig. 19).

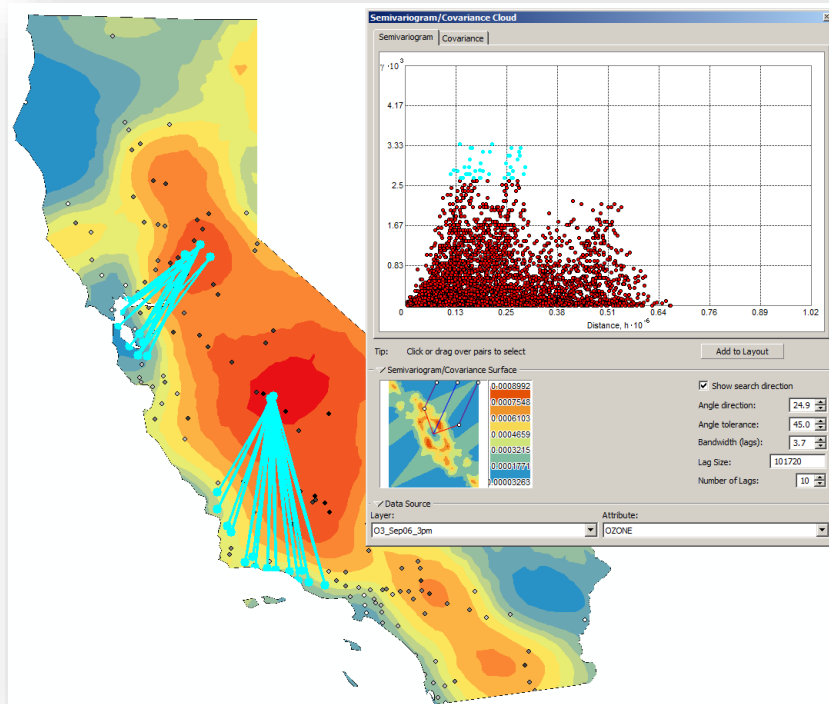


**Fig. 19. Wskaźnik kierunku na diagramie zmienności semiwariogramów**

Kierunek wskazywany przez wskaźnik określa, które pary lokalizacji danych są nanoszone na semiwariogram. Jeśli na przykład wskaźnik skierowany jest w kierunku wschód – zachód, na semiwariogramie zostaną wykreślone tylko pary lokalizacji danych, które są położone względem siebie w osi wschód – zachód. Pozwala to na

wyeliminowanie par, których zmiennością nie jesteśmy w danym momencie zainteresowani i zbadanie wpływu kierunków na dane.

- 4.6. Wybierz polecenie *Wybierz Cechy wg. Prostokąta (Select Features by Rectangle)*. Zaznacz pary punktów o najwyższych wartościach semiwariogramu (nie jest ważne, aby wybrać dokładnie te same punkty na wykresie lub używać tego samego kierunku wyszukiwania).



**Fig. 20. Wybrane pary punktów położone względem siebie wzdłuż osi NE-SW (kierunek: 24,9°, tolerancja kierunku: 45°)**

Zauważ, że większość połączeń lokalizacji (reprezentujących pary punktów na mapie) odpowiada jednemu z punktów opróbowania położonemu w kalifornijskiej Dolinie Centralnej. Dzieje się tak dlatego, że wartości ozonu w tym obszarze są najwyższe w całej Kalifornii.

- 4.7. Kliknij przycisk *Zamknij* znajdujący się w górnym rogu okna dialogowego.
- 4.8. Aby usunąć wybrane zaznaczenia, kliknij przycisk *Wyczyść Wybrane Obiekty (Clear Selected Features)* znajdujący się na pasku narzędzi *Narzędzia (Tools)*.

W tym etapie ćwiczenia dowiedzieliśmy się, że:

- ✓ Rozkład zawartości ozonu w atmosferze stanu Kalifornia z dnia 6 września 2007 r. (w godz. 3:00–16:00) ma charakter jednomodalny, ale nie jest zbliżony do rozkładu normalnego.
- ✓ Wykres QQ potwierdza, że dane nie posiadają rozkładu zbliżonego do rozkładu normalnego, ponieważ punkty na wykresie nie tworzą prostej. Może istnieć konieczność przekształcenia danych.
- ✓ Korzystając z narzędzia analizy trendów, zauważyliśmy, że dane wykazują tendencję. Powierzchnia trendu może być aproksymowana wielomianem 2-go rzędu.
- ✓ *Chmura semiwariogramu/kowariancji* pokazała, że niezwykle wysokie wartości *semiwariogramu* są w dużej mierze reprezentowane przez linie prostopadłe do wybrzeża. Analiza za pomocą tego narzędzia wskazuje, że model interpolacyjny powinien uwzględniać anizotropię (kierunkowość).
- ✓ Powierzchnia semiwariogramu wskazuje, że w danych są przestrzenne autokorelacje. W zestawie danych nie stwierdziliśmy żadnych prób huraganowych (lub wartości błędnych), wobec czego nie musimy żadnych danych usuwać.

Mając to wszystko na uwadze, będziemy w stanie stworzyć bardziej dokładną powierzchnię niż ta, którą utworzyliśmy w pierwszym etapie ćwiczenia, używając domyślnych opcji i wartości parametrów.

W 3-cim etapie ćwiczenia wykorzystamy to, czego dowiedzieliśmy się o danych, aby opracować lepszy model interpolacji niż ten, który został utworzony w pierwszym etapie.

## 5. Bibliografia

Johnston K., Ver Hoef J.M., Krivoruchko K., Lucas N., 2001. *Using ArcGIS™ Geostatistical Analyst*. ESRI, New York, 300.

Tobler, W.R., 1970. A Continuous Transformation Useful for Districting. *Annals of the New York Academy of Sciences*, 215-220.

Urbański, J., 2011. *GIS w badaniach przyrodniczych*. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk, 252.