## 05.2019 r.

**Przestrzenne bazy danych**

## Projekt

**Temat nr 6**

**Eksploracja przestrzenna danych z rejestrów pokładowych statków oceanicznych z okresu 1750 - 1850**

###### Autorzy: Jakub Gwiazda Miłosz Pluta

**Warszawa 2019**

**1. Charakterystyka danych źródłowych**

Dane źródłowe pochodzą z dzienników pokładowych statków oceanicznych, z lat 1750 – 1850. Wśród wielu motywów ponownego opracowania i analizy danych z dzienników można wyróżnić dwa najważniejsze. Pierwszym była chęć analizy danych historycznych przy pomocy nowoczesnych narzędzi informatycznych, co pozwoliłoby na lepsze zobrazowanie zmian klimatu na morzach i oceanach, jakie zachodziły ponad dwa wieki temu. Natomiast drugim było dostarczenie unikalnych i bezcennych danych dla społeczności naukowej, w celu umożliwienia przeprowadzenia dalszych badań.

Dane pochodzą z dzienników pokładowych statków należących do następujących państw:

* Anglia
* Holandia
* Francja
* Hiszpania
* Argentyna

Podstawowymi parametrami, mierzonymi przez nawigatorów były:

* Prędkość / kierunek wiatru
* Długość / szerokość geograficzna
* Czas
* Temperatura powietrza / wody

Oprócz powyższych były także odnotowywane, jeżeli zachodziły, inne zjawiska występujące na morzu np. liczba błyskawic lub grzmotów.

Ze względu na brak standaryzacji pomiarów w badanym okresie, należało dane z dzienników odpowiednio przygotować przed wgraniem ich do bazy. W wielu przypadkach zdecydowano się na liczne uproszczenia, aby umożliwić także analizę niedokładnych pomiarów.

Baza danych **CLIWOC**, dostarcza wspomnianych danych w sposób uporządkowany, co ułatwia przeprowadzenie dalszej analizy. Rekordy są opisane szeregiem atrybutów. W tabeli przedstawiono dziesięć wybranych atrybutów wraz z krótkim opisem.

|  |  |
| --- | --- |
| YR | Rok(UTC) |
| MO | Miesiąc(UTC) |
| DY | Dzień(UTC) |
| HR | Godzina(UTC) |
| LAT | Szerokość geograficzna |
| LON | Długość geograficzna |
| C1 | Kod kraju macierzystego statku |
| W | Szybkość wiatru |
| AT | Temperatura powietrza |
| D | Kierunek wiatru |

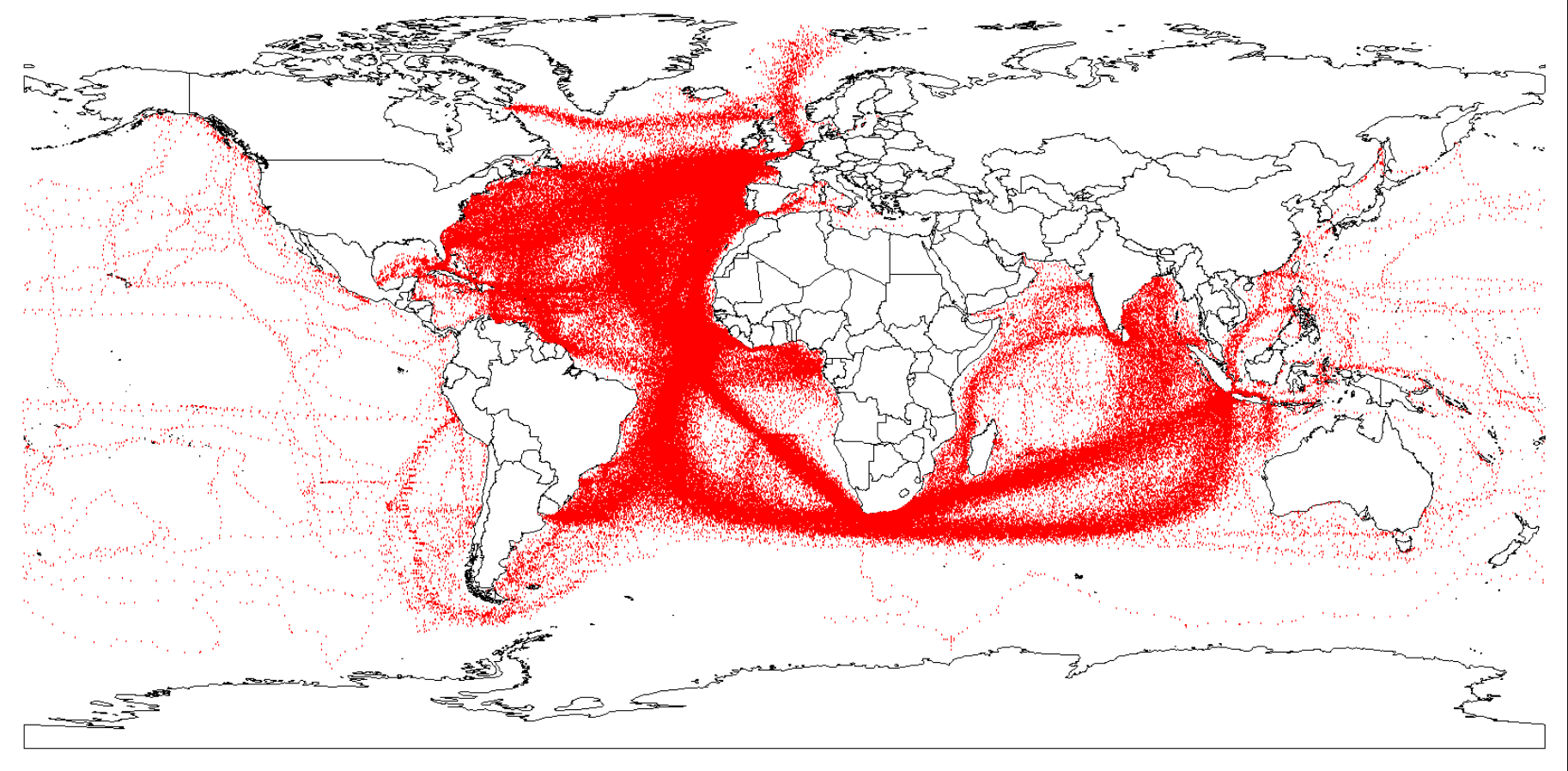
**Prezentacja danych źródłowych**

Pierwszym etapem wykonania projektu było wstępne przygotowanie danych źródłowych. W ramach tego etapu zaimplementowano:

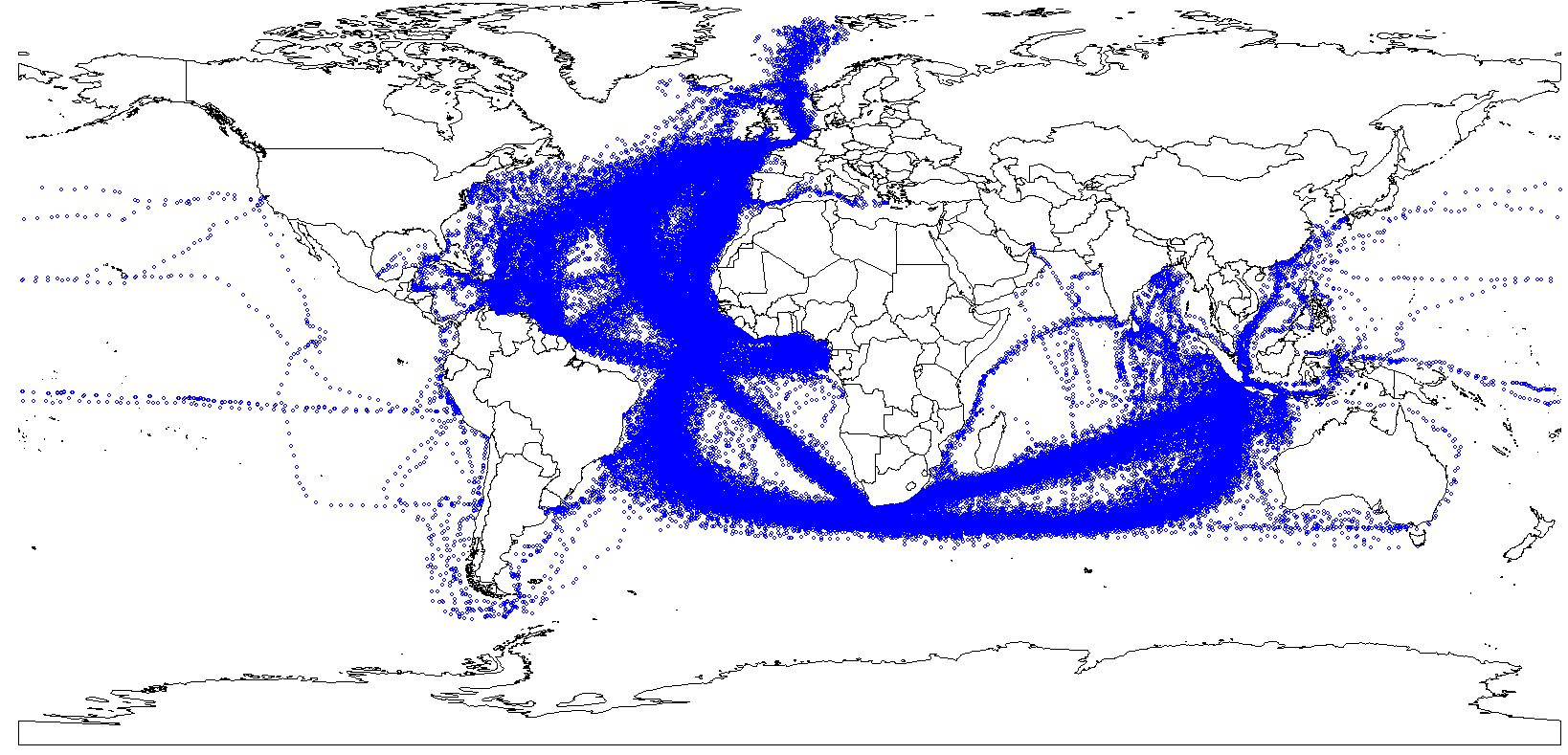
* Wczytywanie danych
* Normalizacje współrzędnych geograficznych(długość / szerokość)
* Wizualizacja znormalizowanych danych

Współrzędne geograficzne w kolumnach **LAT** oraz **LON** wymagały dostosowania do zakresów współrzędnych obsługiwanych przez biblioteki języka R. Po wykonaniu normalizacji zdecydowano się na wizualizację danych na mapie świata. Ze względu na bardzo dużą ilość zapisanych w bazie danych tras, poniżej przedstawiono poglądowo trasy wszystkich statków oraz tras statków pochodzących z Holandii. Podobnie można zobrazować trasy dla wybranego statku wyodrębniając go z bazy po numerze identyfikacyjnym, trasy statków z danego zakresu czasowego itp.

Przedstawienie wszystkich tras statków oceanicznych w latach 1750 – 1850



Przedstawienie tras oceanicznych dla statków z państwem macierzystym Holandia



**2. Cel badania**

Celem badania jest przeprowadzenie analizy danych źródłowych przy użyciu wybranych funkcji z języka R, służących do analizy danych przestrzennych. Poniżej przedstawiono eksperymenty, których wyniki zostaną przedstawione w kolejnym rozdziale:

* Analiza wariogramów danych źródłowych
* Predykcja wartości siły wiatru – Kriging
* Reprezentacja danych za pomocą funkcji pochodzących z pakietu hexbin
* Reprezentacja danych za pomocą funkcji pochodzących z pakietu fMultivar
* Zobrazowanie danych za pomocą pakietu Point Pattern

Wynikiem wymienionych eksperymentów mogą być ciekawe wnioski na temat danych klimatycznych(np. prędkość wiatru), ale także na temat danych dotyczących ruchu statków po oceanach.

Każdy z eksperymentów będzie składał się z czterech części:

* Cel eksperymentu
* Opis wykorzystywanej metody analizy oraz funkcji z języka R
* Wizualizacja wyniku
* Analiza wyniku

**3. Analiza przestrzenna danych**

W niniejszym rozdziale opisano zastosowanie wybranych funkcji do analizy przestrzennej.

**3.1 Analiza wariogramów danych źródłowych**

**Cel eksperymentu**

Celem eksperymentu jest analiza wariogramów dla przestrzennych danych źródłowych. Zdecydowano się na zawężenie analizy wariogramów dla statków pływających pod banderą Niemiec oraz USA. Taka decyzja była spowodowana zbyt długim czasem obliczania wariogramu dla większej ilości punktów(np. dla statków Hiszpańskich) oraz niemożliwością obliczenia wariogramu dla wszystkich punktów(zbyt mało pamięci obliczeniowej).

**Opis metody**

Dla danych przestrzennych wariogram opisuje stopień zależności punktów w przestrzeni.

Semiwariancja jest miarą przestrzennej zależności między dwiema obserwacjami(punktami) jako funkcja odległości między nimi.

Wyróżnia się wariogramy teoretyczne oraz empiryczne. W tym przypadku zostaną obliczone wariogramy empiryczne(na podstawie danych źródłowych). Wariogramy teoretyczne oraz ich dopasowanie będą tematem kolejnego eksperymentu.

Język R udostępnia pakiet: **geoR**, który może posłużyć do wyliczenia wariogramu. Przed przystąpieniem do liczenia należy przygotować dane, a mianowicie przekonwertować je do klasy: **geodata**. Po wykonaniu tej operacji można obliczyć wariogram poniższą funkcją:

variog(geodata = gerandusGeoData, option = "cloud", max.dist = 1)

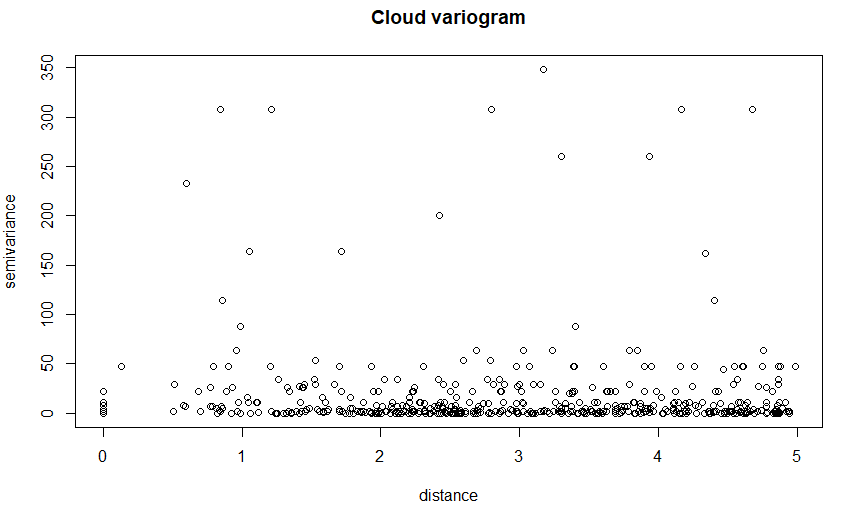
gdzie:

* geodata – obiekt klasy **geodata** zawierający koordynaty oraz dane
* option – typ wariogramu jaki zostanie wygenerowany
* max.dist – maksymalna odległość między kolejnymi punktami(zależne od wartości option)

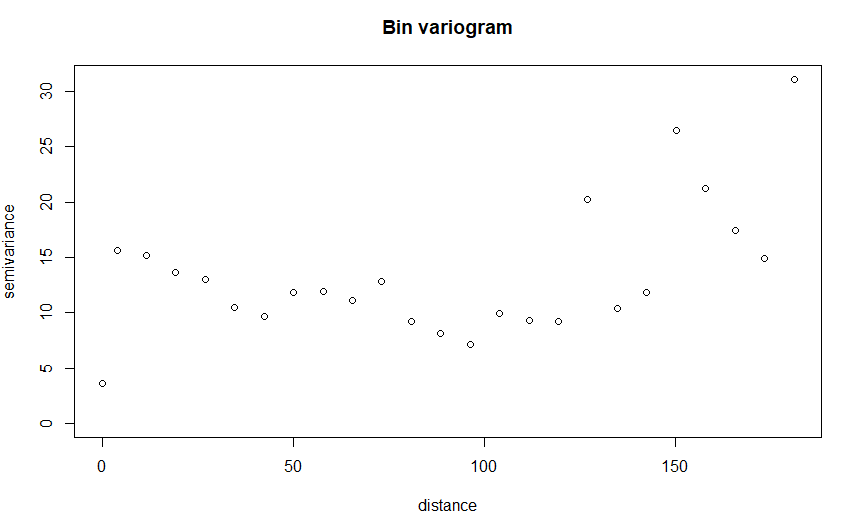
**Wizualizacja wyniku**

Poniżej przedstawiono wyniki otrzymane z dwóch eksperymentów.

1. Obliczanie wariogramu typu **cloud**



1. Obliczanie wariogramu typu **bin**



**Wnioski**

W obu eksperymentach obliczono wariogram dla tras statków niemieckich oraz amerykańskich.

W eksperymencie pierwszym użyto wariogramu typu ‘cloud’. Wynik pokazuje że niezależnie od odległości, punkty mają bardzo małą zależność przestrzenną między sobą. Co zdaje się potwierdzać fakt, iż statki niemieckie pływały z reguły w kierunku południowym(Afryka), a amerykańskie głównie do Europy.

W eksperymencie drugim użyto wariogramu typu ‘bin’ ze zwiększonymi dystansami. Ten przypadek pokazał że zmiana odległości znacząco wpłynęła na wyniki. Powiązanie między punktami utrzymuje się na stałym średnim poziomie wraz ze wzrostem odległości. Jedynie przy odległości 150 i więcej następuje wzrost powiązania punktów.

**3.2 Kriging**

**Cel eksperymentu**

Celem eksperymentu jest predykcja wartości siły wiatru na podstawie pomiarów wykonanych przez nawigatorów, na wybranych trasach.

**Opis metody**

W celu poznania rozkładu wybranej cechy na danym obszarze, należałoby poznać jej wartość w każdym punkcie tej przestrzeni – jest to niemożliwe. Do tego celu stosuje się metody interpolacji przestrzennej.

Kriging jest jedną z metod interpolacji przestrzennej, która wyróżnia się od innych metod tym, że traktuje obserwowaną cechę jak zmienną losową, co oznacza że zmienna w każdym punkcie badanej przestrzeni ma wartość oczekiwaną i wariancję.

Jednym z etapów stosowania metody Kriging’u jest obliczenie semiwariogramu empirycznego – narzędzie służące do estymacji i badania struktury zmienności badanych zjawisk w geostatyce - tj. przybliżenie semiwariogramu teoretycznego na podstawie obserwacji punktów bazowych. Następnie do tak obliczonego semiwariogramu jest dopasowywany model semiwariogramu teoretycznego. Wyróżnia się cztery podstawowe typy semiwariogramów teoretycznych:

* Wykładniczy – EXP
* Sferyczny - SPH
* Gaussa - GAU
* Liniowy – MAT

Język R udostępnia pakiet **gstat**, który oferuje szereg funkcji przestrzennych w tym kriging oraz generowanie semiwariogramów. Poniżej przedstawiono dwa kluczowe elementy wymienionego pakietu.

1. Dopasowanie wariogramu teoretycznego do empirycznego

fit.variogram(lzn.vgmbob, model = vgm(psill = max(lzn.vgmbob$gamma)\*0.5,

model = "Sph", range = max(lzn.vgmbob$dist)/2,

nugget = mean(lzn.vgmbob$gamma)/4))

gdzie:

* psill – wartość progowa, do której dąży semiwariogram
* model – model teoretyczny opisany wyżej
* range – odległość między punktami bazowymi przy której semiwariogram osiąga 95% wartości progowej
* nugget – kwadrat różnicy między obserwacjami leżącymi najbliżej siebie

1. Obliczanie wyniku metody kriging

krige(log(W)~1, bayOfBiscayTUni, grid, model = lzn.fitbob)

gdzie:

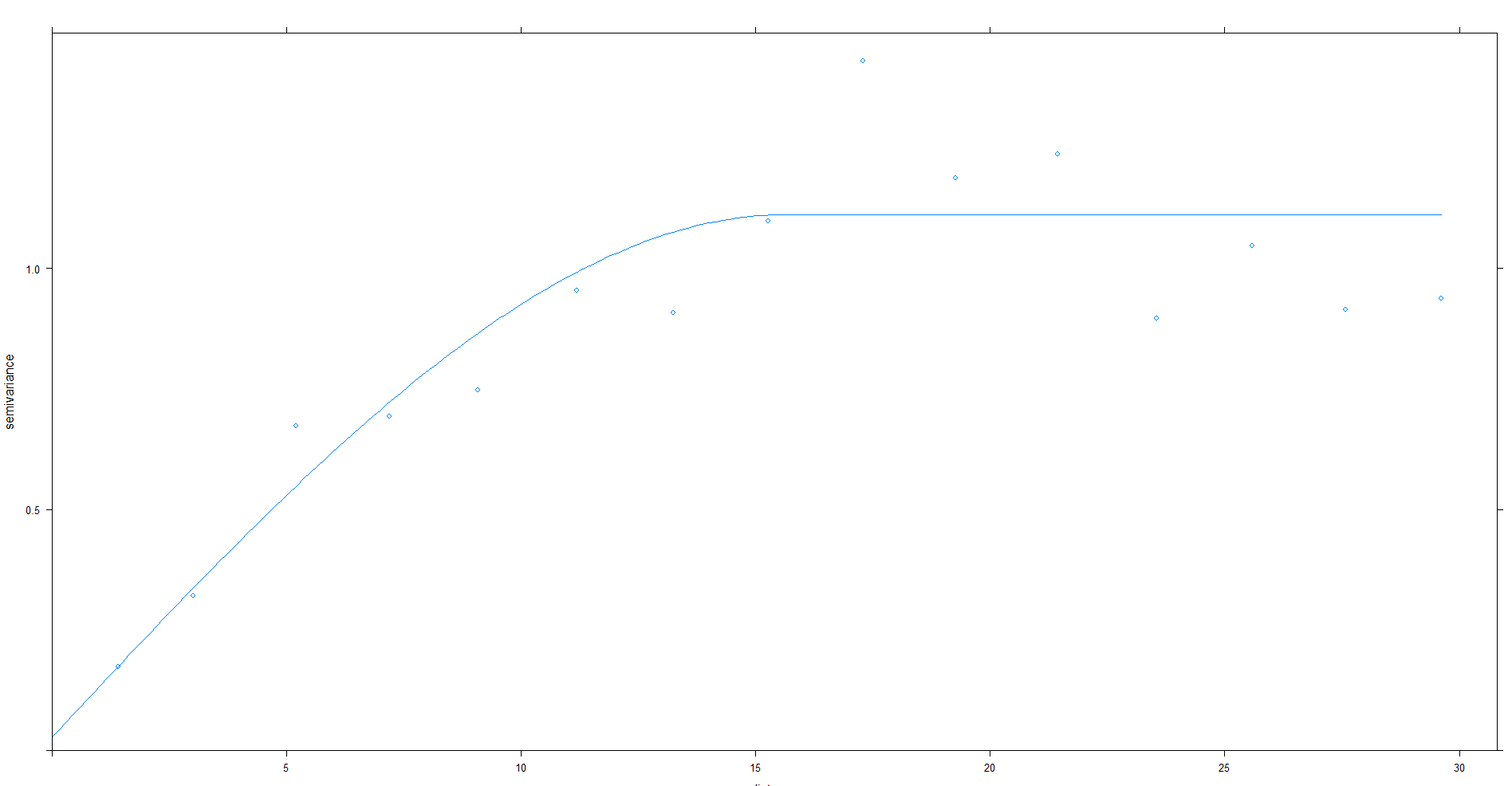
* formula(log(W)~1) – definicja zmiennej zależnej(wiatru) jako modelu liniowego zmiennych niezależnych
* locations(bayOfBiscayTUni) – dane przestrzenne
* newdata(grid) – siatka która zostanie zamodelowana
* model – wariogram dla zmiennej zależnej

**Wizualizacja wyniku**

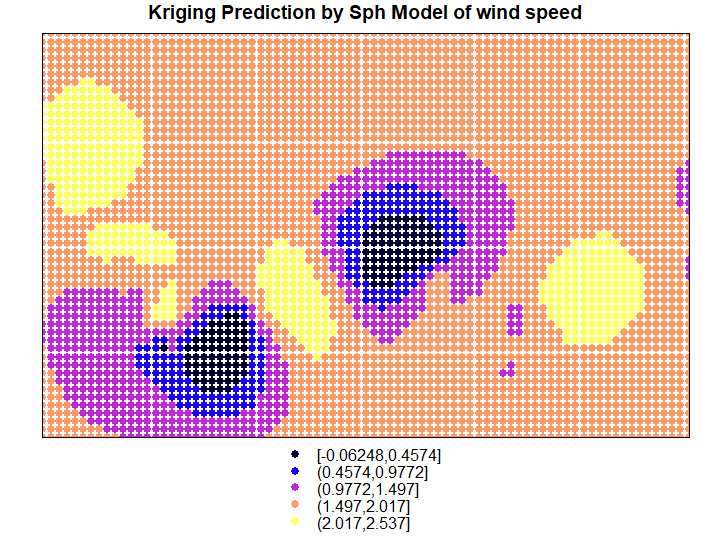
Poniżej przedstawiono wyniki otrzymane z dwóch eksperymentów.

1. Przewidywanie siły wiatru na trasach statków niemieckich

Dopasowanie semiwariogramu teoretycznego

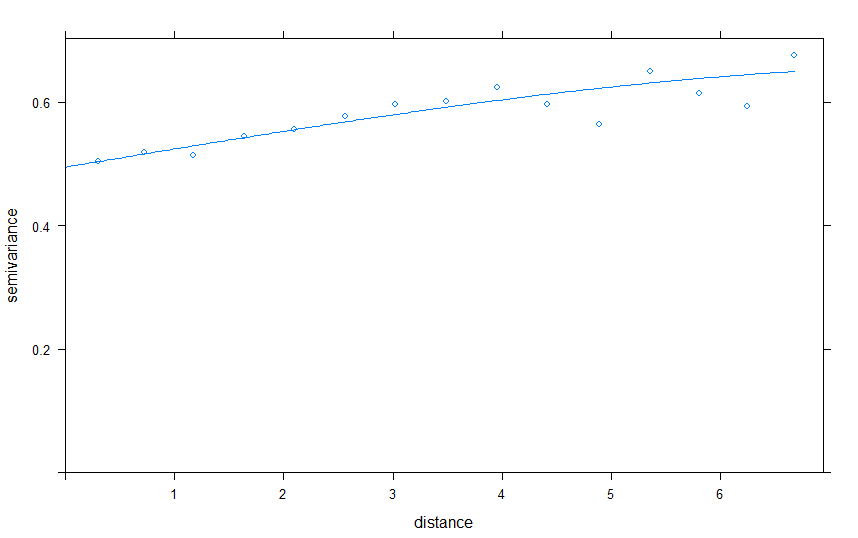


Wynik krigingu

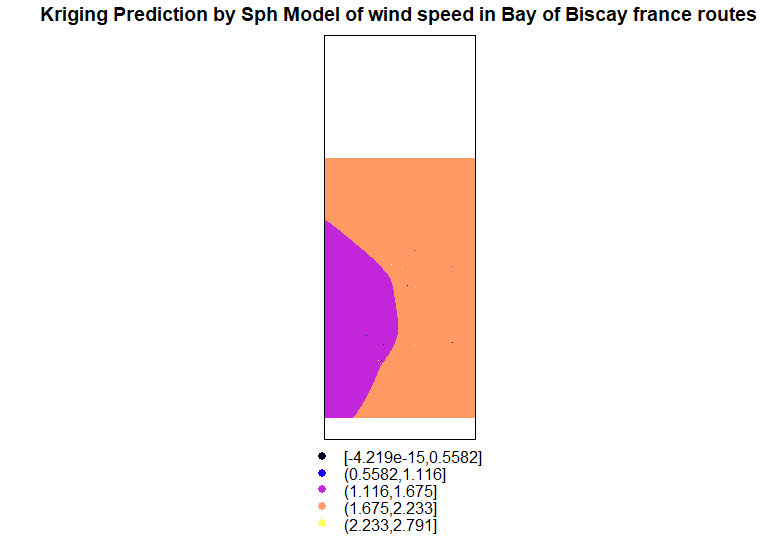


1. Przewidywanie siły wiatru na trasach statków francuskich w Zatoce Biskajskiej

Dopasowanie semiwariogramu teoretycznego



Wynik krigingu



**Wnioski**

Wynik pierwszego eksperymentu, dla tras statków niemieckich, obrazuje zmienność wiatru na zadanym obszarze. Jakość wyniku pozwala stwierdzić dobre dopasowanie semiwariogramu oraz poprawny wybór wartości początkowych parametrów. Interesujący jest otrzymany wykres, ponieważ pokazuje miejscowe zmiany wiatru.

Drugi eksperyment zdecydowano przeprowadzić się na zawężonym obszarze do Zatoki Biskajskiej(wzięto pod uwagę wszystkie statki które tamtędy przepłynęły). Zawężenie obszaru spowodowało jednak duże problemy z dopasowaniem semiwariogramów empirycznego oraz teoretycznego. Po wielu próbach i korekcie wartości początkowych uzyskano dopasowanie modelu teoretycznego do empirycznego. Jednak wynik samej operacji krigingu został bardzo uogólniony – można wnioskować że od strony oceanu Atlantyckiego(po lewej) wiatr jest słabszy niż od strony kontynentu.

**3.3 Reprezentacja danych za pomocą pakietu hexbin**

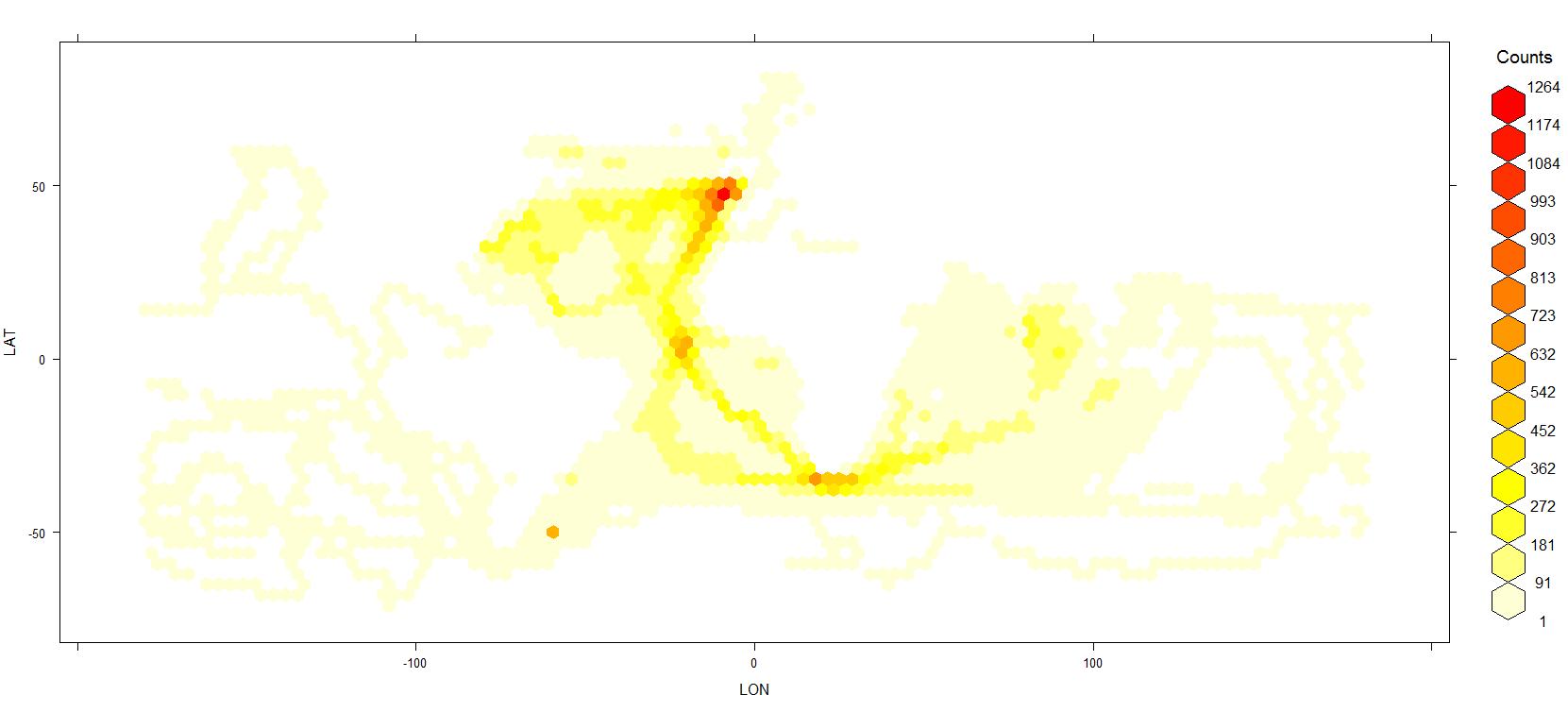
**Cel eksperymentu**

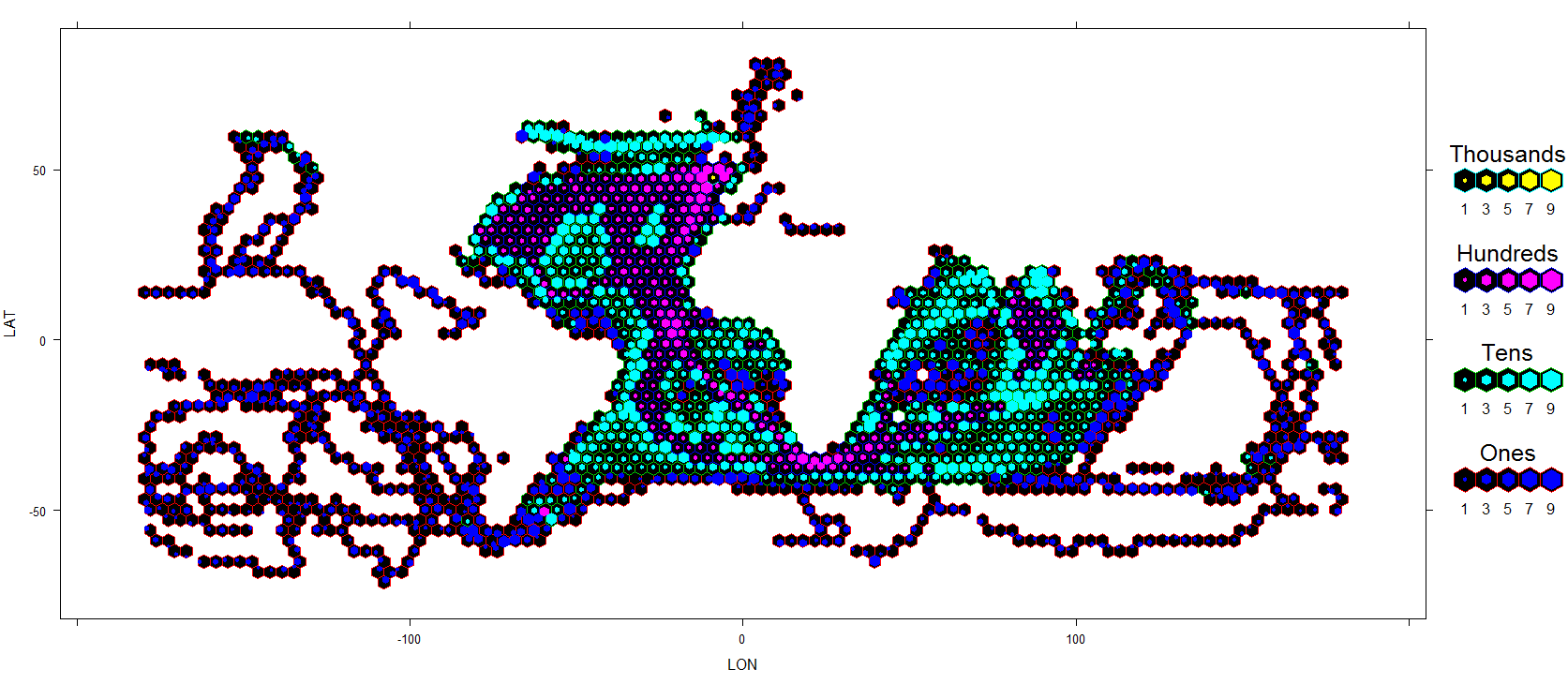
Celem badania danych za pomocą funkcji pochodzących z tego pakietu było graficzne przedstawienie tras na mapie oraz zbadanie tendencji w jakie obszary szlaki były najbardziej uczęszczane. Do badania wybrano statki pochodzące z Wielkiej Brytanii.

**Opis metody**

Funkcje z wybranego pakietu pozwalają podzielenie badanego obszaru na hexagony i zaprezentowanie w tej postaci badanych danych. Rozróżnienie i analizę zebranych danych realizujemy dzięki temu, że po podziale obszaru na hexagony funkcje wyliczają ilość wystąpień punktów na każdy hexagon z czego potem wynika dalsza analiza. W naszym przypadku skupiliśmy się na analizie gęstości wystąpień punktów w zadanym obszarze dzięki czemu udało się pokazać tendencję skąd i dokąd statki pokonywały na przestrzeni lat najczęściej trasy.

**Wizualizacja wyniku**



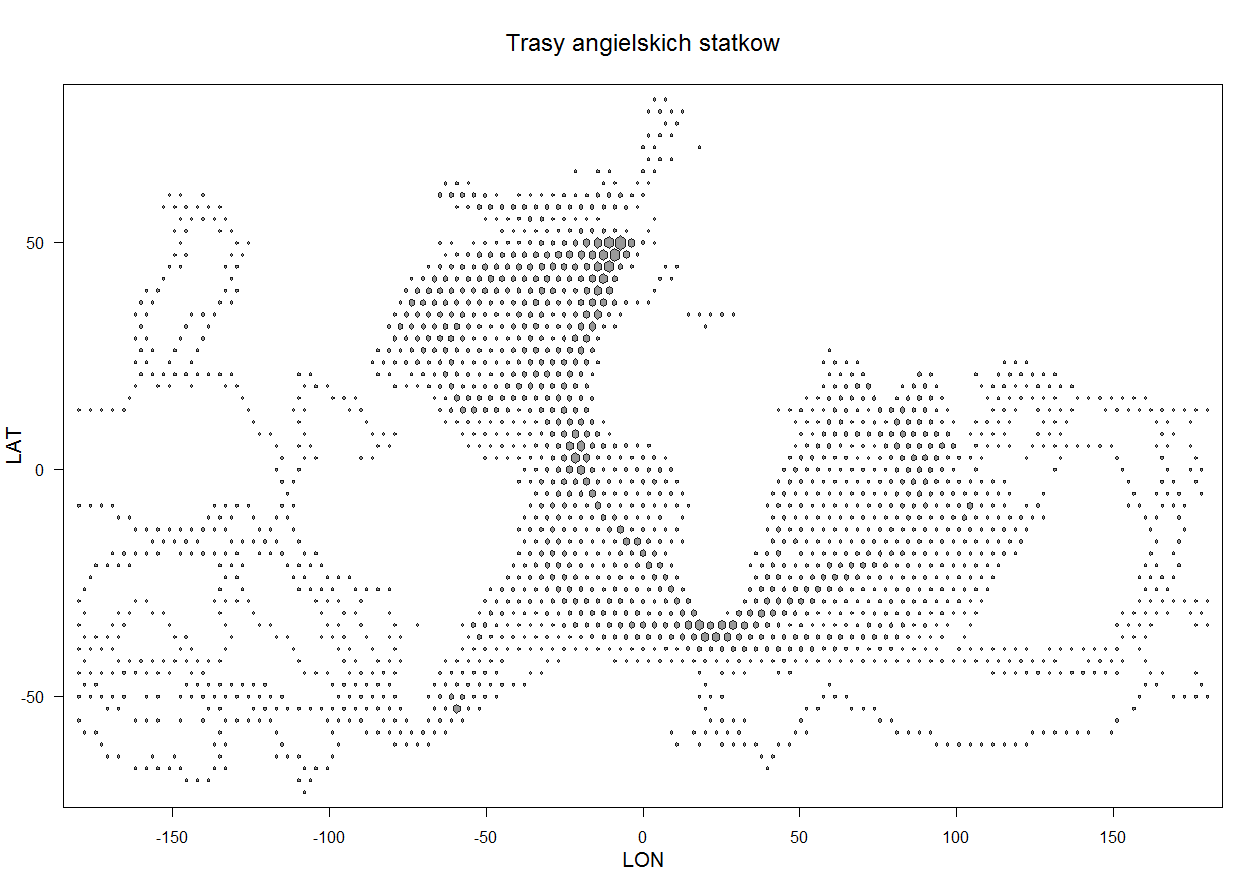


**Wnioski**

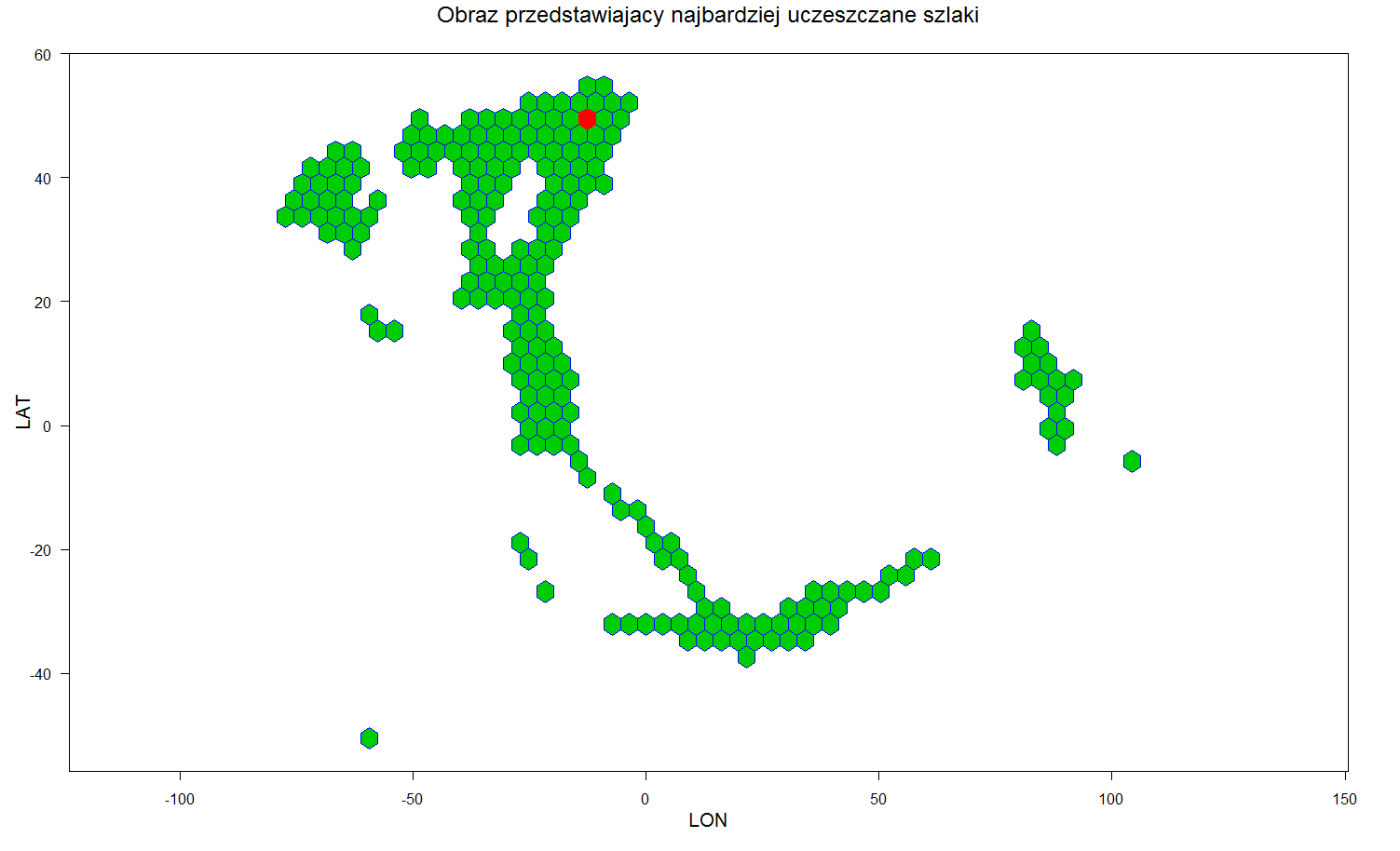
Powyżej możemy zaobserwować dwa wykresy. Pierwszy z nich prezentuje ilość trafień w dany hexagon co odzwierciedla wynik „counts” należy tutaj zauważyć że wielkość dobranego parametru wielkości hexagonu ma znaczące wpływ na otrzymany wynik. Dlatego też podczas porównywania w ten sposób kilku serii wyników należy zwracać baczną uwagę czy dane są przetwarzane w ten sam sposób.

Drugi wykres prezentuje również częstość zliczeń punktów w danym obszarze jednak do analizy użyto filtru tzw. „nestes.centroids”. Dane są prezentowane w dwojaki sposób. Na wykresie w zależności od ilości wystąpień trafień w dany hexagon rysowany jest kolorowy sześcian wewnątrz większego stanowiącego ramę, wraz ze zwiększaniem się częstości trafień wielkość rysowanego sześcianu ulega powiększeniu, aż do osiągniecia maksymalnej wielkości.

Dane za pomocą metod z pakietu hexbin można przedstawić również w postaci hexagonów, których wielkości w zależności od zagęszczenia punktów w każdym z nich ulegają zmianie, przedstawione jest to na obrazie poniżej.



W celu analizy danych wykorzystano również algorytm erozji. Służy on do wyodrębnienia pewnej frakcji danych zawierających komórki obejmujące największą liczbę punktów. Pozwala to na ukazanie wyników najbardziej istotnych. Przykładowy wynik działania funkcji pokazuje obraz poniżej.



Jak łatwo zauważyć widać znaczącą różnicę w porównaniu do wcześniej ukazanych wyników działania funkcji. Wiąże się to z tym, że mechanizm erozji działa w zadanych cyklach po każdym przebiegu eliminując pewną ilość danych i zostawiając resztę znajdujących się nad progiem decyzyjnym. Czerwony punkt przedstawia komórkę zawierającą największą liczbę zliczeń.

**3.4 Funkcje z pakietu fMultivar**

**Cel eksperymentu**

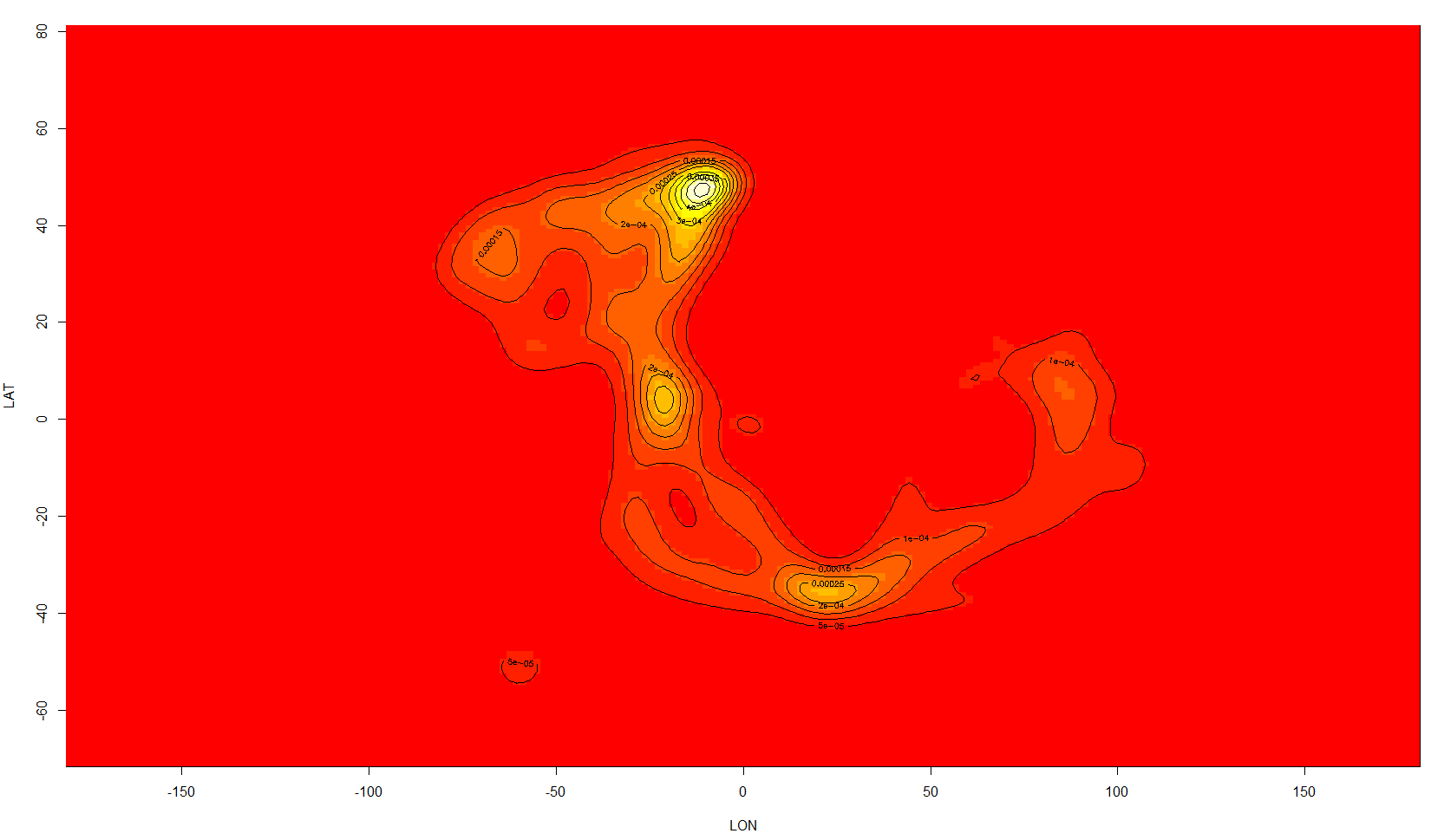
Celem badania danych za pomocą funkcji pochodzących z tego pakietu było graficzne odniesienie do obrazów otrzymanych za pomocą funkcji z pakietu hexbin. Dlatego też badanymi trasami są te pochodzące ze statków Wielkiej Brytanii.

**Opis metody**

Dzięki funkcjom z pakietu fMultivar mamy możliwość wyliczenia 3 rozkładów Cauchy’ego,

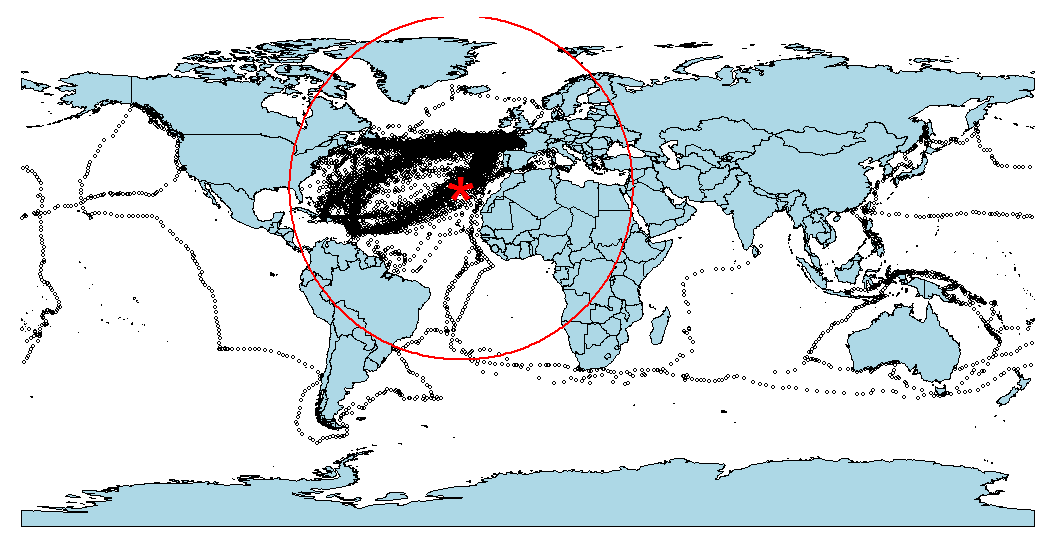
Normalnego oraz rozkładu t-Studenta. Otrzymywany obraz jest zwykle w tych przypadkach histogramem, który można przedstawić w postaci 3-d. Jednak dla naszego przykładu skupiliśmy się na zbadaniu gęstości i próbie jej przedstawienia w 2-d.

**Wizualizacja wyniku**



Zastosowanie funkcji analizującej gęstość występowania punktów pozwoliło na otrzymanie podzielenie całego zakresu danych na obszary o podobnej gęstości występowania punktów.

**3.5 Point pattern analysis**



Powyżej można zaobserwować wynik analizy za pomocą funkcji z pakietu Point Pattern. Do zobrazowania wykorzystano dane pochodzące z zapisów odnośnie statków pływających pod banderą Francji. Dzięki funkcjom pochodzącym z wcześniej wspomnianego pakietu byliśmy w stanie obrazowo zaznaczyć na średnią odległość na jaką poruszały się statki. Dodatkowo funkcje ze wspomnianego pakietu pozwalają podobnie jak poprzednie na zbadanie gęstości punktów lub inne obliczenia związane z kwestiami związanymi z odległościami punktów, które nie zostały zobrazowane w tym przypadku ze względu na skupienie się na ukazaniu funkcji udostępnionych przez ten pakiet. Należy jednak zauważyć, że aby możliwa była analiza danych w naszym przypadku wcześniej pobrane i znormalizowane dane długości i szerokości geograficznej, powiązać ze sobą oraz utworzyć obiekt przestrzenny.

**4. Porównanie wyników**

W przypadku funkcji z pakietu hexbin oraz fMultivar otrzymywane wyniki mogą się uzupełniać. W obu przypadkach badany obszar dzielony jest na odpowiedniej wielkości hexagony w których funkcje zliczają ilość wystąpień trafień. Możemy to obrazować w postaci zliczeń(pakiet hexbin), lub po przez funkcje rozkładów (pakiet fMultivar). Jednak należy jasno zauważyć że w przypadku próby przewidywania lub określenia prawdopodobieństwa należy korzystać z możliwości oferowanych przez pakiet fMultivar.

Porównane mogą być także dwie metody generowania wariogramów, które są używane do obliczania funkcji Kriging. Funkcje z pakietu **geoR** umożliwiają w stopniu podstawowym określenie jak powinien wyglądać generowany wariogram. Natomiast funkcje z pakietu **gstat** umożliwiają zaawansowane modelowanie wariogramu empirycznego, tak aby dopasować go poprawnie do wariogramu teoretycznego.

Końcowa wizualizacja za pomocą pakietu Point Pattern pokazała nam inny sposób podejścia do analizy przestrzennej. Każda para punktów może stanowić obiekt przestrzenny na rzecz którego będziemy wywoływać metody co pozwala na przeprowadzenie analizy. Znaczącą różnicą w stosunku do np. wcześniej wspomnianych funkcji z pakietów fMultivar oraz hexbin jest to, że zdecydowanie łatwiej przychodzi połączenie różnych typów wykresów np. mapy świata oraz danych tras. Kwestii tej nie udało się osiągnąć w przypadku wizualizacji przypadków pakietów hexbin i fMultivar.

**5. Wnioski**

Analiza przestrzenna bazy danych CLIWOC okazała się być zadaniem trudnym, ponieważ dane nie były znormalizowane do postaci odpowiedniej do analizy przestrzennej. Większość funkcji do analizy przestrzennej, z dostępnych pakietów, operuje na specjalnych formatach danych(np. SpatialDataframe), które nie są dostosowane do danych geograficznych oceanicznych(np. znacznie utrudnione formowanie wielokątów, wykorzystywanych między innymi do regresji przestrzennej lub innych analiz przestrzennych).

Dodatkową przeszkodą był brak normalizacji danych źródłowych. Dane były podane w specyficznych formatach, stosowanych przez nawigatorów notujących pomiary podczas rejsu, co uniemożliwiło efektywne zastosowanie większości funkcji służących do analizy przestrzennej.

Mimo opisanych przeszkód udało się wykonać normalizację danych oraz przeprowadzić analizę przestrzenną – Krigin lub analizę gęstości danych. Ponadto autorzy wzbogacili swoją wiedzę na temat danych przestrzennych oraz ich analizy.

W zależności od wymagań oraz oczekiwanych rezultatów należy zastanowić się nad doborem odpowiednich narzędzi, tak jak w przypadku chęci zobrazowania jak w naszym przypadku mapy świata wraz z zaznaczonymi punktami obrazującymi trasy statków lepszym wyborem było wybranie funkcji z pakietu Point Pattern, gdyż pakiety hexbin oraz fMultivar ze względu na specyficzny typ obiektu nie pozwoliły nam na dołączenie makiety świata.

Do sprawnych i trafnych obserwacji danych przestrzennych przekonaliśmy się, że należy wykorzystać kilka dostępnych narzędzi, a następnie zebrać wyniki uzyskane za ich pomocą i poddać dopiero analizie.

**6. Literatura**

1) <http://webcache.googleusercontent.com/search?q=cache:pE0k1gxwchcJ:www.au.poznan.pl/~rwal/ochrona_gis_pliki/wyklad_9.ppsx+&cd=2&hl=pl&ct=clnk&gl=pl>

2) <https://en.wikipedia.org/wiki/Kriging>

3) <https://wiki.landscapetoolbox.org/doku.php/spatial_analysis_methods:semivariogram_analysis>

4) <https://rspatial.org/analysis/8-pointpat.html>

5) <https://rpubs.com/nabilabd/118172>

6) <https://rpubs.com/facehappywy/159648>