## 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. liczbę 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 |

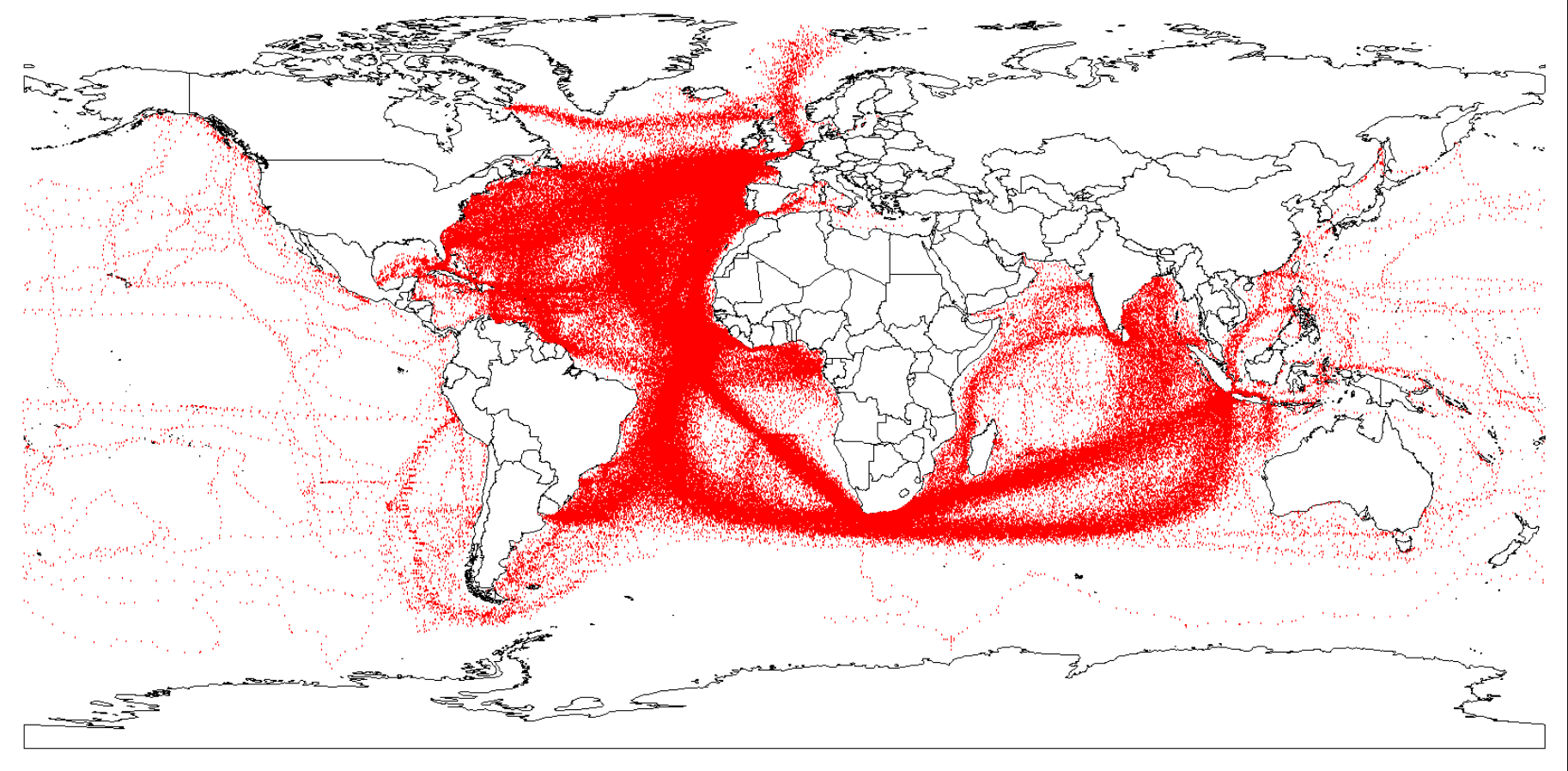
**Prezentacja danych źródłowych**

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

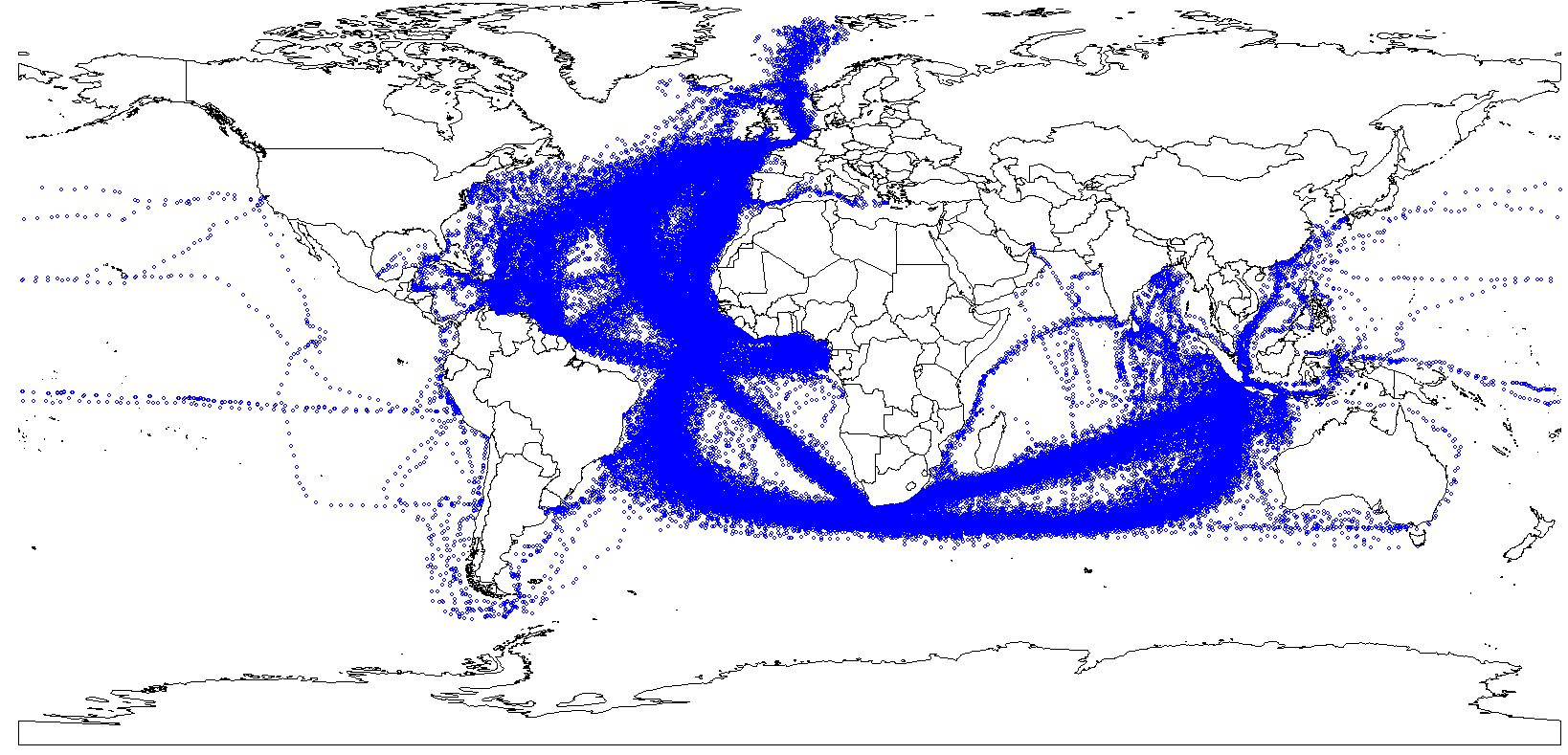
* Wczytywanie danych
* Normalizacje współrzędnych geograficznych(długość / szerokość)
* Wizualizacja znormalizowanych danych

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 przedstawiliśmy poglądowo trasy wszystkich statków oraz tras statków pochodzących z Holandii. Podobnie można zobrazować trasy dla poszczególnego 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:

* Predykcja wartości siły wiatru - Kriging
* Reprezentacja danych za pomocą funkcji pochodzących z pakietu hexbin
* Reprezentacja danych za pomocą funkcji pochodzących z pakietu fMultivar
* Ex 4 do uzupełnienia gdy będziemy wiedzieć jakie są exp

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 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 sruktury 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:

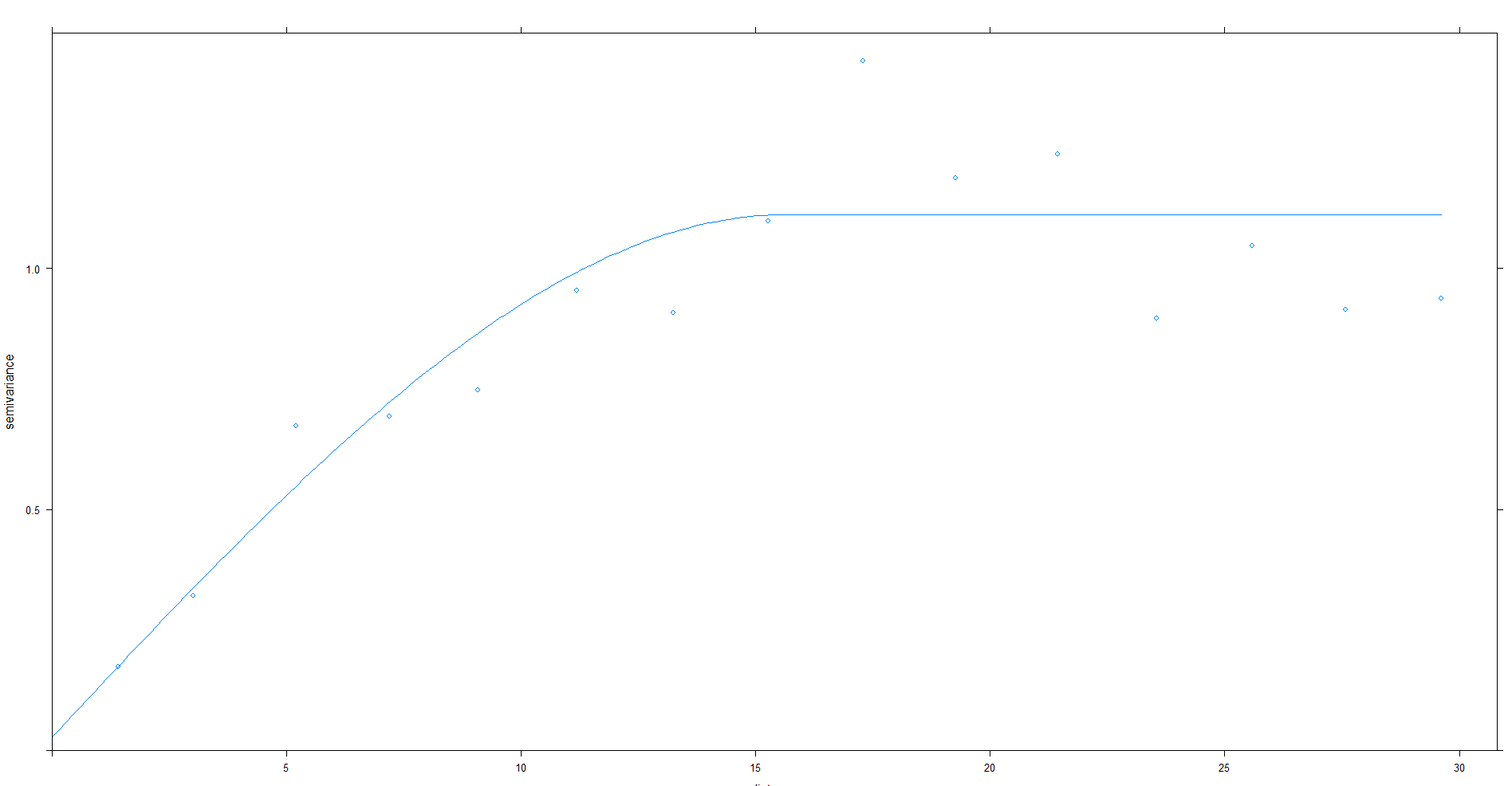
* formula(log(W)~1) – definicja zmiennej zależnej(wiatru) jako modelu liniowego zmiennych niezależnych
* locations(bayOfBiscayTUni) – dane przestrzenne
* newdata(grid) – siatka która zostanie zamodelowana
* model – wariogram dla zmiennej zależnej

**Wizualizacja wyniku**

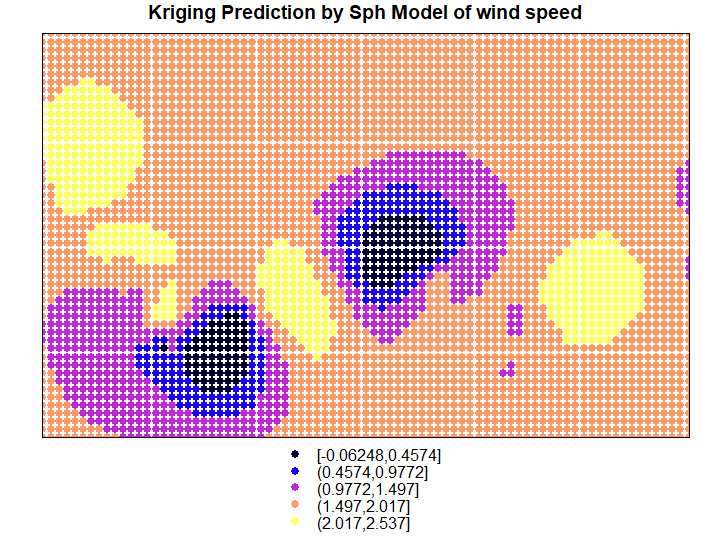
Poniżej przedstawiono wyniki otrzymane z dwóch eksperymentów.

1. Przewidywanie siły wiatru na trasach statków niemieckich

Dopasowanie semiwariogramu teoretycznego

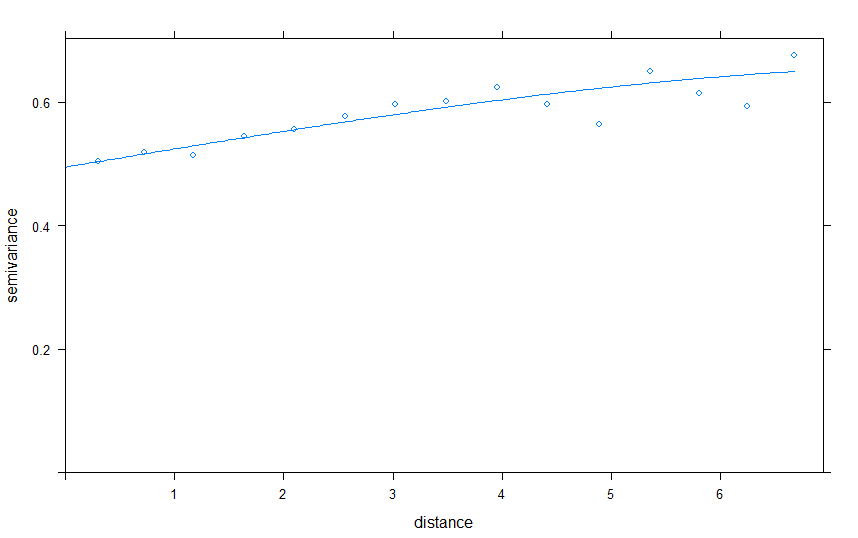


Wynik krigingu

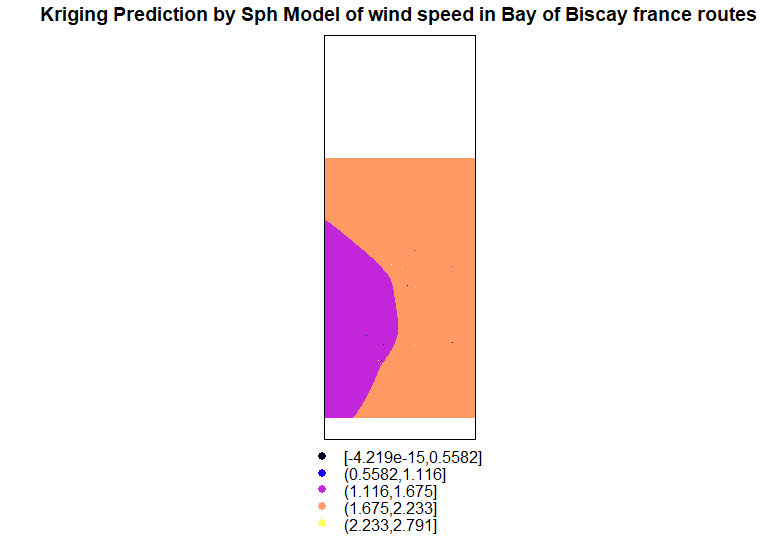


1. Przewidywanie siły wiatru na trasach statków francuskich w Zatoce Biskajskiej

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.

**3.2 Regresja przestrzenna**

**Cel eksperymentu**

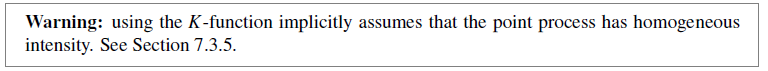
####################################################

**Uwagi – do rozkminy**

**Funkcje K,L,G z pakietu spatstat**

Powyższe funkcje służą do stwierdzenia czy wprowadzone dane(w postaci punktów o współrzędnych x,y) w jakiś sposób się grupują(klastrowanie) lub takie związki nie zachodzą.

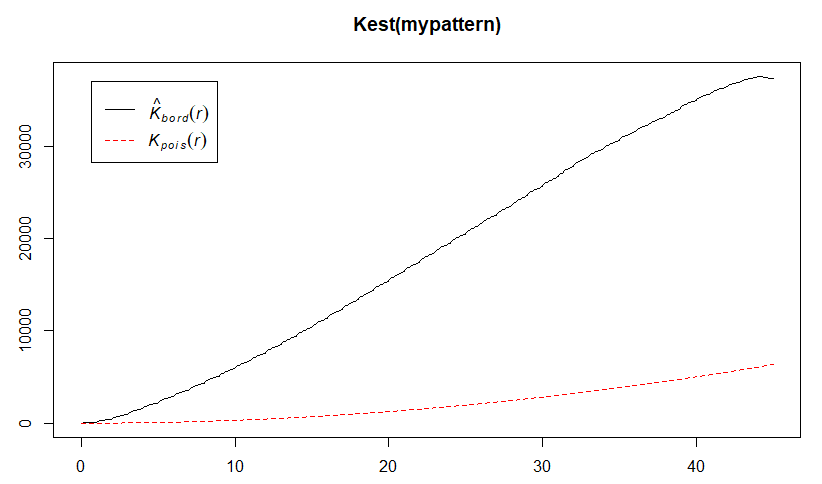
Str. 92 <https://training.fws.gov/courses/references/tutorials/geospatial/CSP7304/documents/PointPatterTutorial.pdf>  
  
U nas nie widzę dla tego zastosowania, ponieważ jedynymi współrzędnymi jakimi można określić położenie jest długość / szerokość. 1) jest podana w stopniach co już utrudnia przełożenie na liczenie odległości 2) dla nas zbędna jest informacja czy punkty gdzie znajdował się statek tworzą / nie tworzą klastrów

Kolejną przesłanką że powyższe funkcje nie mają zastosowania do naszych danych jest wpis z książki: ‘Spatial Points Patterns: Methodology and Applications with R’:   
  


Nasz zbiór punktów nie jest homogeniczny(lądy są puste od punktów) dlatego użycie tych funkcji byłoby zakłamane

Str 7 – zbiór homogeniczny wyjaśnienie  
<https://training.fws.gov/courses/references/tutorials/geospatial/CSP7304/documents/PointPatterTutorial.pdf>

Jeżeli chodzi o samą bibliotekę spatstat to można poszukać ewentualnie innych funkcji, które pozwalają na inne operacje(najpierw jednak bym analizował pozostałe wymienione z pdfa)

Próba rysowania wykresu nawet ‘na pałe’ wygląda tak jak poniżej, ale jest on przekłamany   


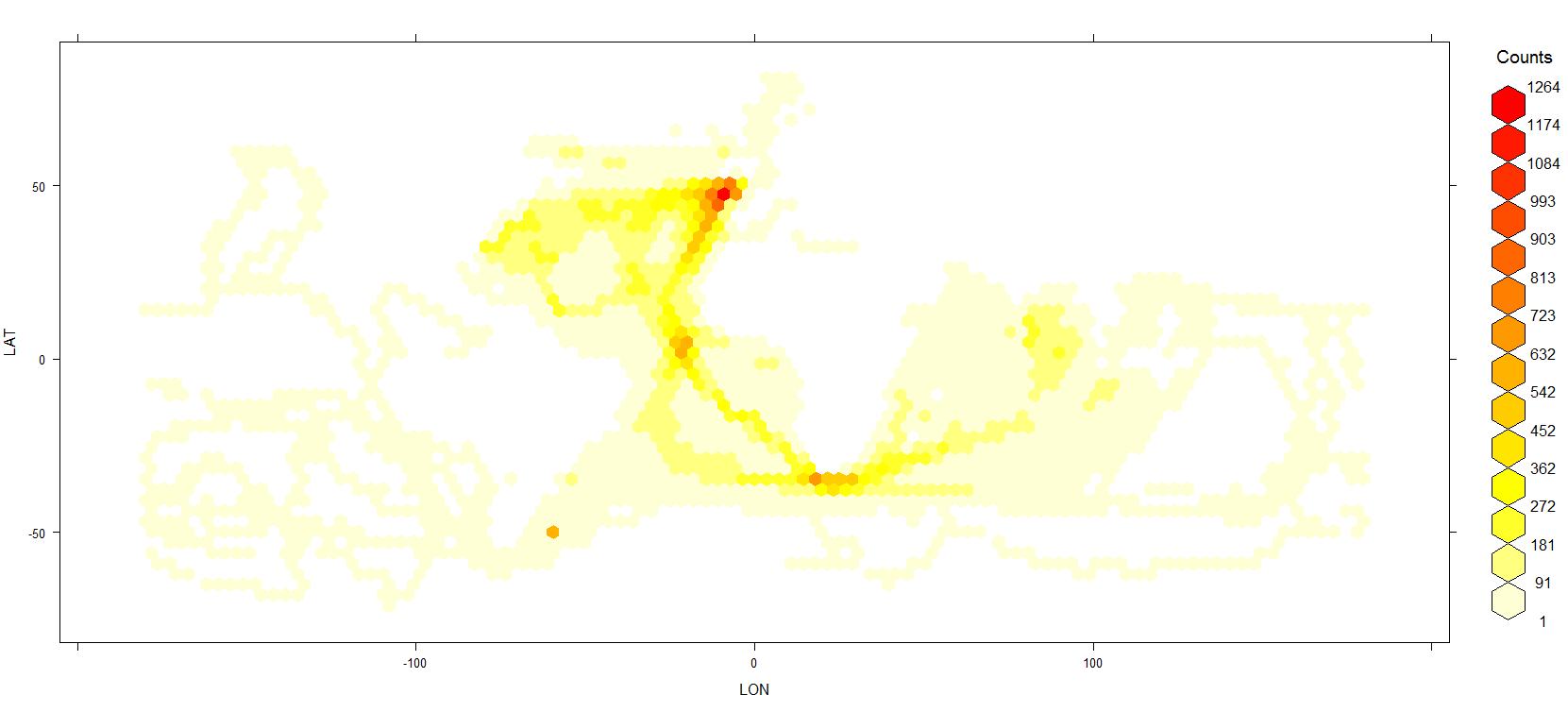
####################################################

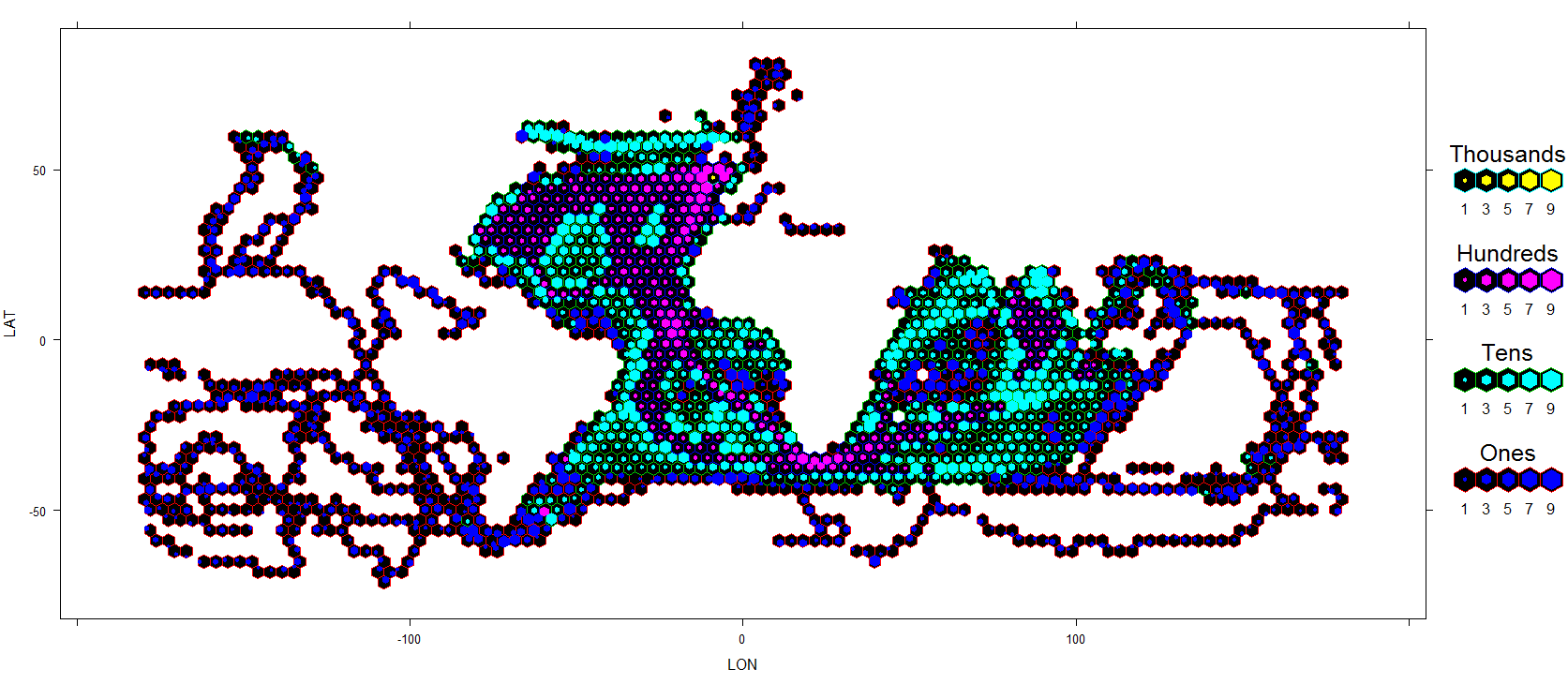
3.3 Reprezentacja danych za pomocą pakietu hexbin

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

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.

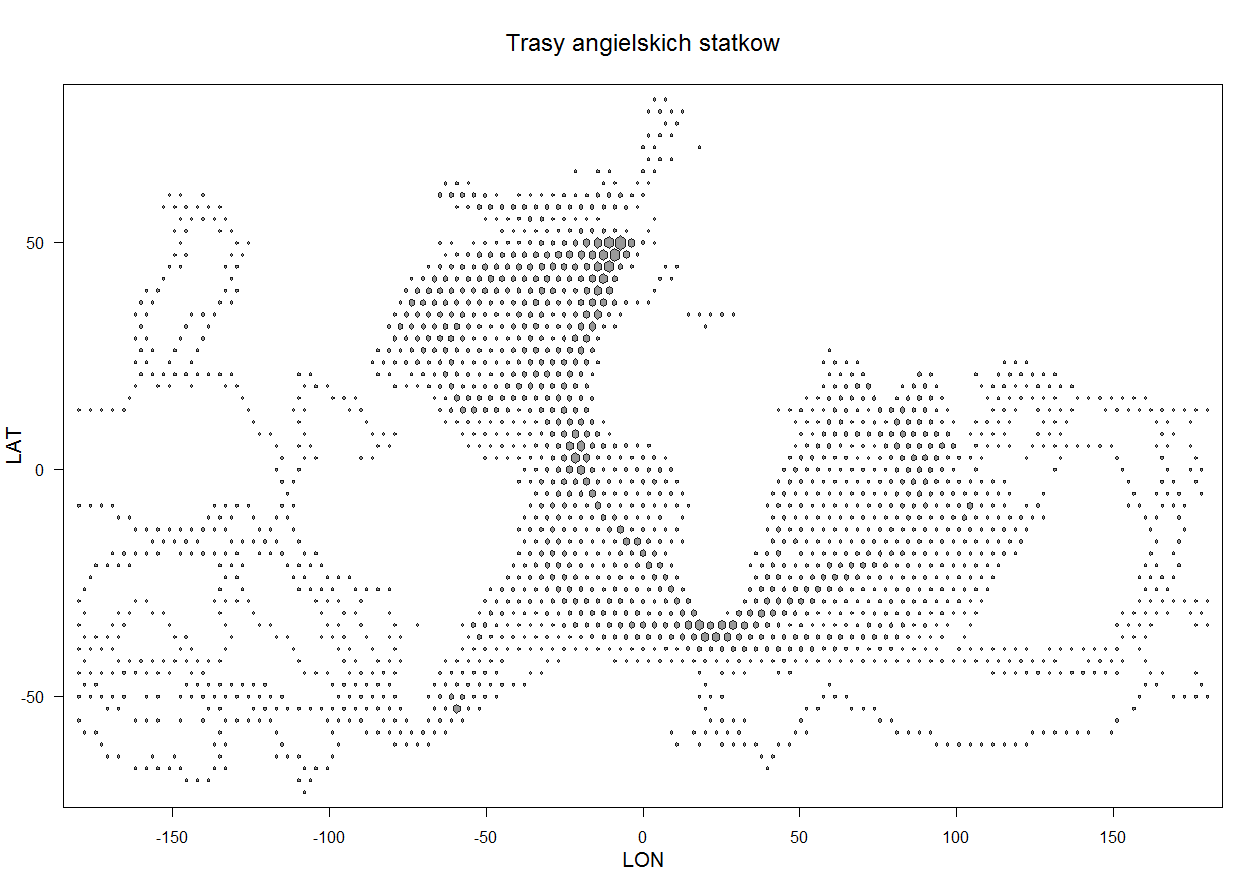




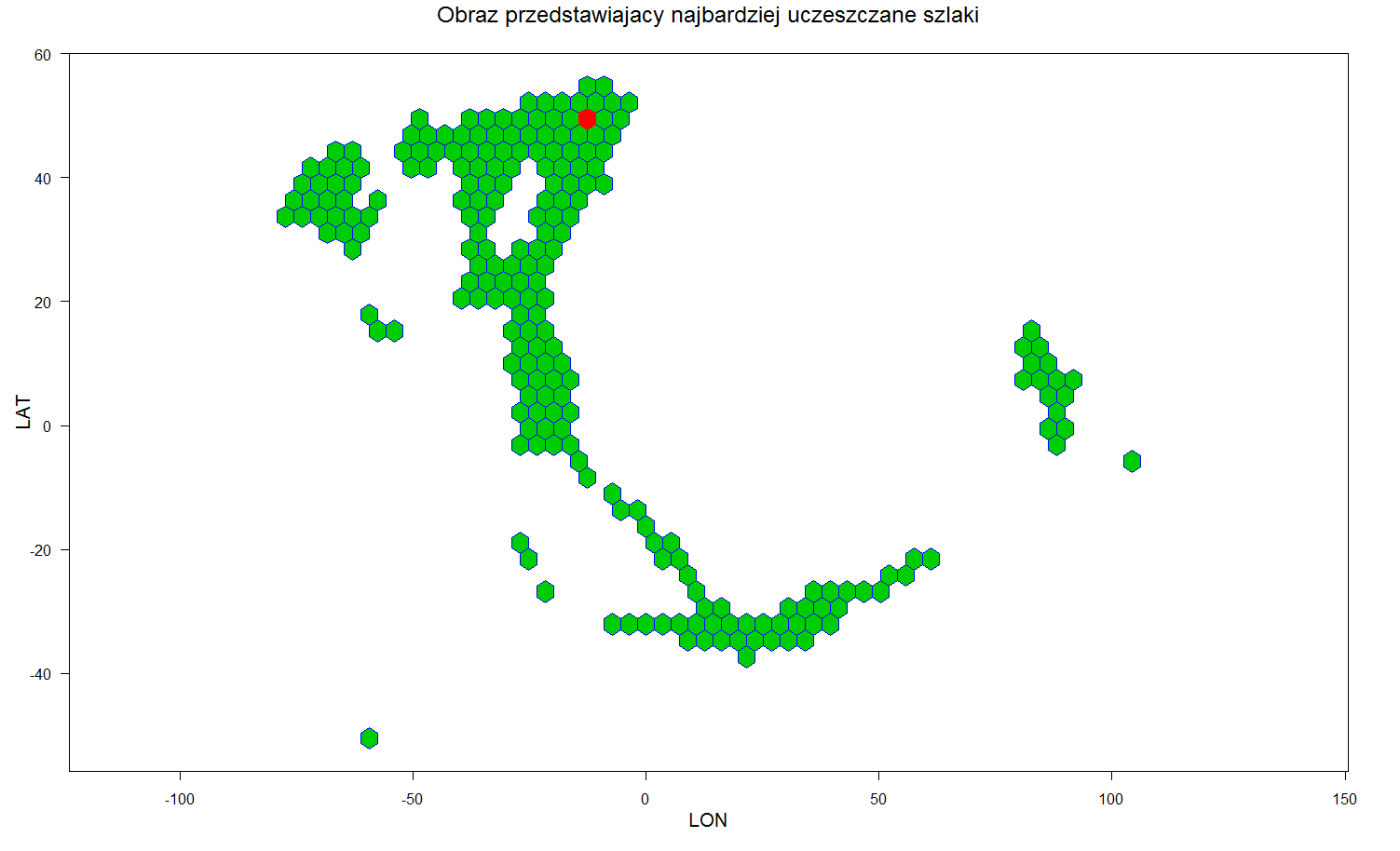
Powyżej możemy zaobserwować dwa wykresy. Pierwszy z nich prezentuje ilość trafień w dany hexagon co odzwierciedla wynik „counts” należy tutaj zauważyć że wielkość dobranego parametru wielkości hexagonu ma znaczące wpływ na otrzymany wynik. Dlatego też podczas porównywania w ten sposób kilku serii wyników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

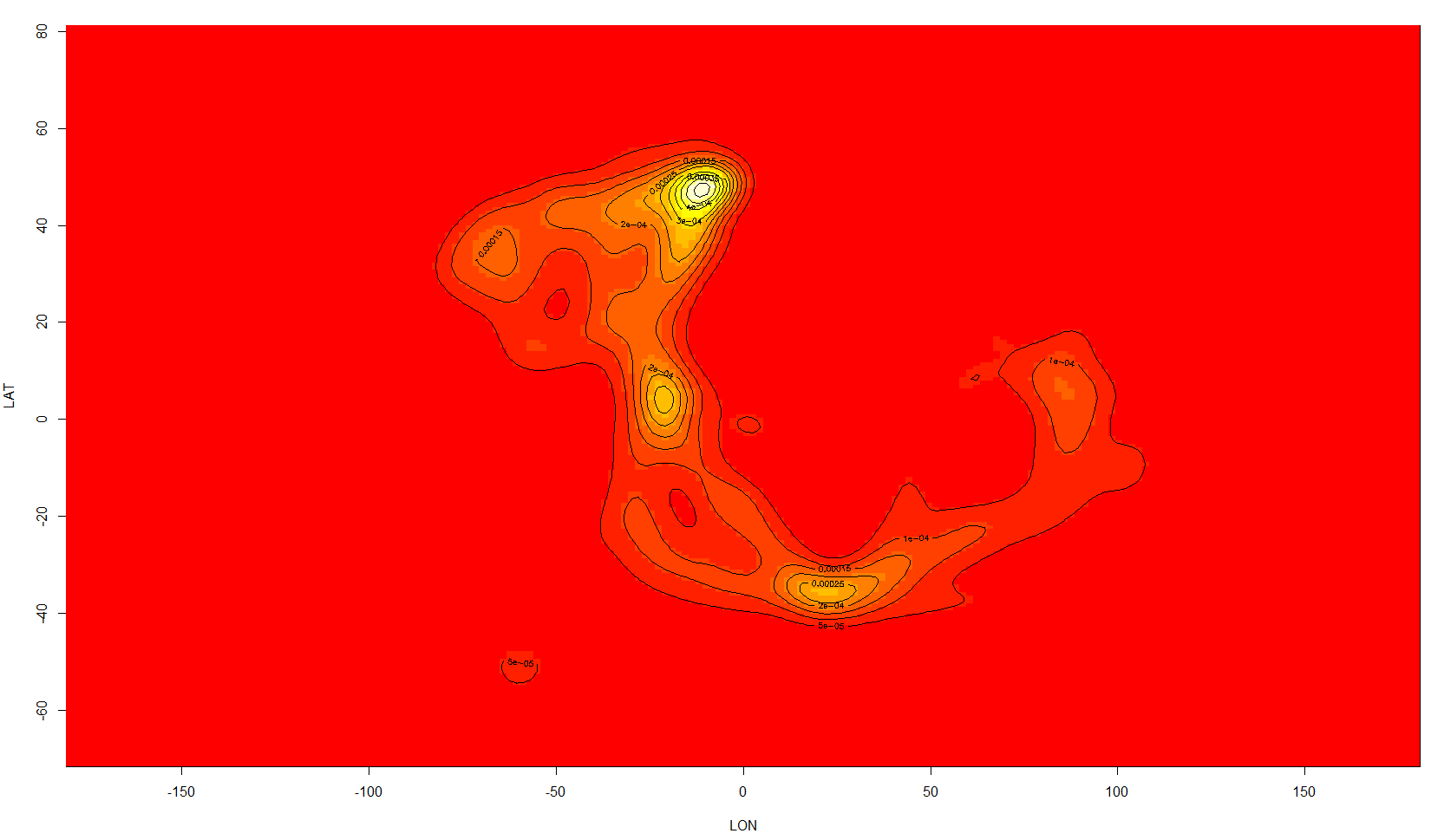
3.4 Funkcje z pakietu fMultivar

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

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.



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

**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.

**5. Wnioski**

**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>