Projekt 1

Klasteryzacja przestrzenna danych punktowych

Celem ćwiczenia jest wyznaczenie obszarów o zwiększonej intensywności zarejestrowanych wykroczeń na terenie Krakowa. Analizę wydzielenia klastrów wykonaj używając pakietu R.

Kod R:

#wczytanie bibliotek

library("spatstat")

library("rgdal")

library("dbscan")

library("ggplot2")

library("broom")

#import danych

osiedla = readOGR(dsn = "pliki", layer = "osiedla", verbose = FALSE)

point = readOGR(dsn = "pliki", layer = "punkty", verbose = FALSE)

#przygotowanie brzegowych wartosci wykresu

minmax=data.frame(osiedla@bbox)

x\_cor = minmax[1,]

y\_cor = minmax[2,]

# przygotowanie danych punktowych do analizy

point\_xy = point@coords

# DBSCAN

db = dbscan(point\_xy, eps = 500, minPts = 30)

# wyswietlenie klastra miejsc wykroczen na mapie krakowa

plot(osiedla)

par(new=TRUE)

hullplot(point\_xy, db,

ylim = c(y\_cor$min, y\_cor$max),

xlim = c(x\_cor$min, x\_cor$max),

xlab = "x [meters]", ylab = "y[meters]",

asp = 1, main = "DBSCAN")

#HDBSCAN

hdb = hdbscan(x = point\_xy, minPts = 15)

#wyswietlenie klastra miejsc wykroczen na mapie krakowa

plot(osiedla)

par(new = TRUE)

hullplot(point\_xy, hdb,

ylim = c(y\_cor$min, y\_cor$max),

xlim = c(x\_cor$min, x\_cor$max),

xlab = "x [meters]", ylab = "y[meters]",

asp = 1, main = "HDBSCAN")

#OPTICS

opt = optics(x=point\_xy, eps = 1500, minPts = 30)

#Prog do identyfikacji klastrow

db\_opt = extractDBSCAN(opt, eps\_cl = .1)

#Prog stromosci do hierarchicznej identyfikacji klastrow przy uzyciu metody Xi.

db\_opt = extractXi(opt, xi = 0.01)

# wyswietlenie klastra miejsc wykroczen na mapie krakowa

plot(osiedla)

par(new = TRUE)

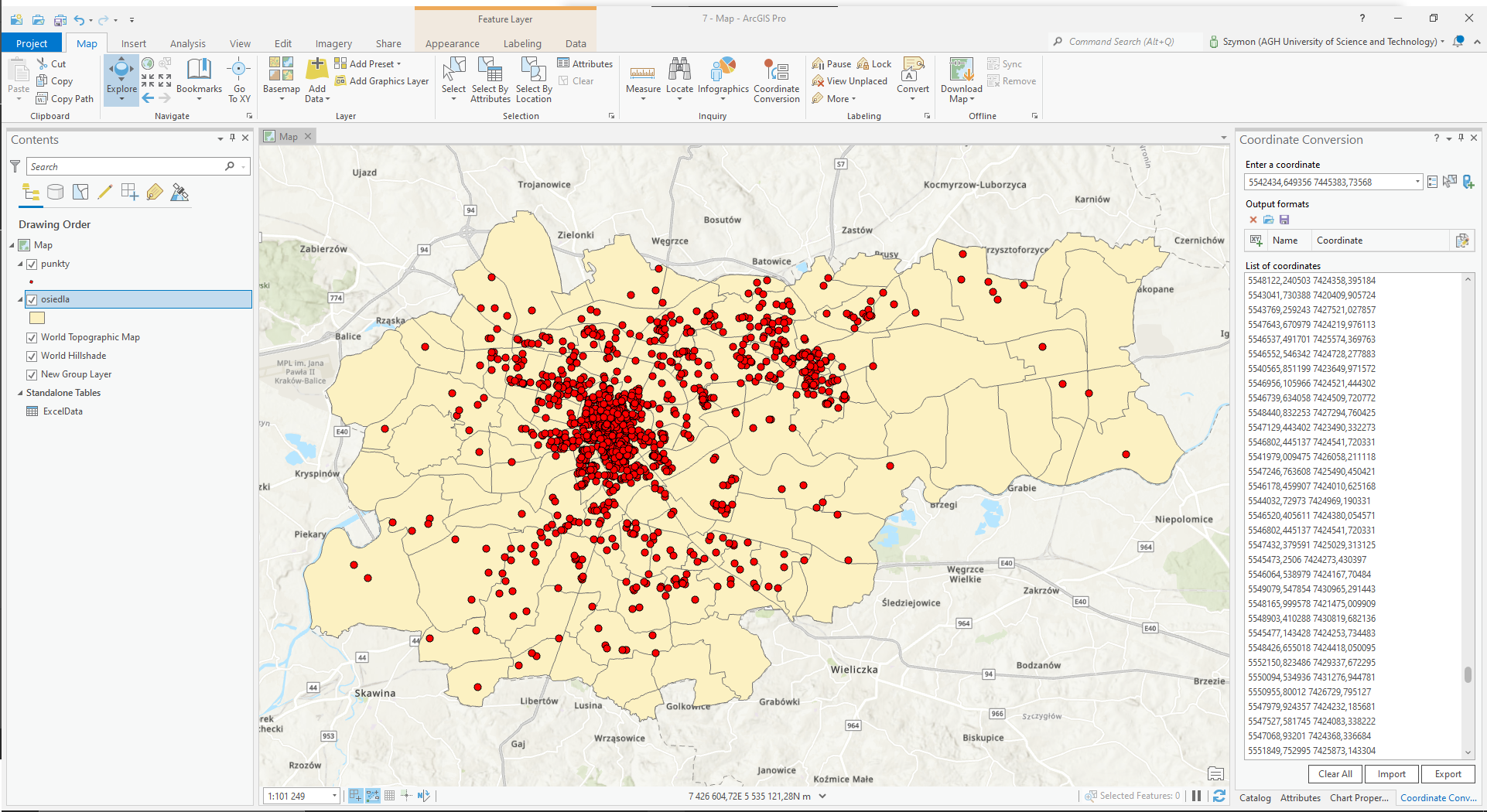
hullplot(point\_xy, db\_opt,

ylim = c(y\_cor$min, y\_cor$max),

xlim = c(x\_cor$min, x\_cor$max),

xlab = "x [meters]", ylab = "y[meters]",

asp = 1, main = "OPTICS")



Screen 1. „*wizualizacja punktów z zmienionym układem współrzędnych*”

Opis funkcji:

DBSCAN

Ma dwa argumenty „eps”, czyli minimalna odległość łącząca 2 punkty by mógł powstać klaster, oraz „minPts”, czyli minimalna ilość obserwacji by mógł powstać klaster

Zalety:

-Odporny na wpływ obserwacji odstających.

-Znakomicie radzi sobie z grupami o niewypukłym kształcie.

-Daje dobre rezultaty.

-Daje możliwość definiowania wielu miar.

Wady:

-Nie daje możliwości definiowania a priori liczby segmentów .

-Dobór odpowiednich parametrów bywa dosyć problematyczny.

HDBSCAN

Ma jeden argument „minPts” działający tak jak w DBSCAN.

HDBSCAN zasadniczo oblicza hierarchię wszystkich klastrów DBSCAN, a następnie wykorzystuje metodę ekstrakcji opartą na stabilności, aby znaleźć optymalne cięcia w hierarchii, tworząc w ten sposób płaskie rozwiązanie.

Zalety:

-Posiada wszystkie zalety DBSCAN

-Dodatkowo ma mniejszą wrażliwość niż DBSCAN

-Ma dobrą stabilność w przypadku doboru parametrów

Wady:

-Największą wadą jest dobór odpowiednich wartości domyślnych

OPTICS

Podobnie jak DBSCAN ma 2 argumenty „eps” oraz „minPts”

Zalety:

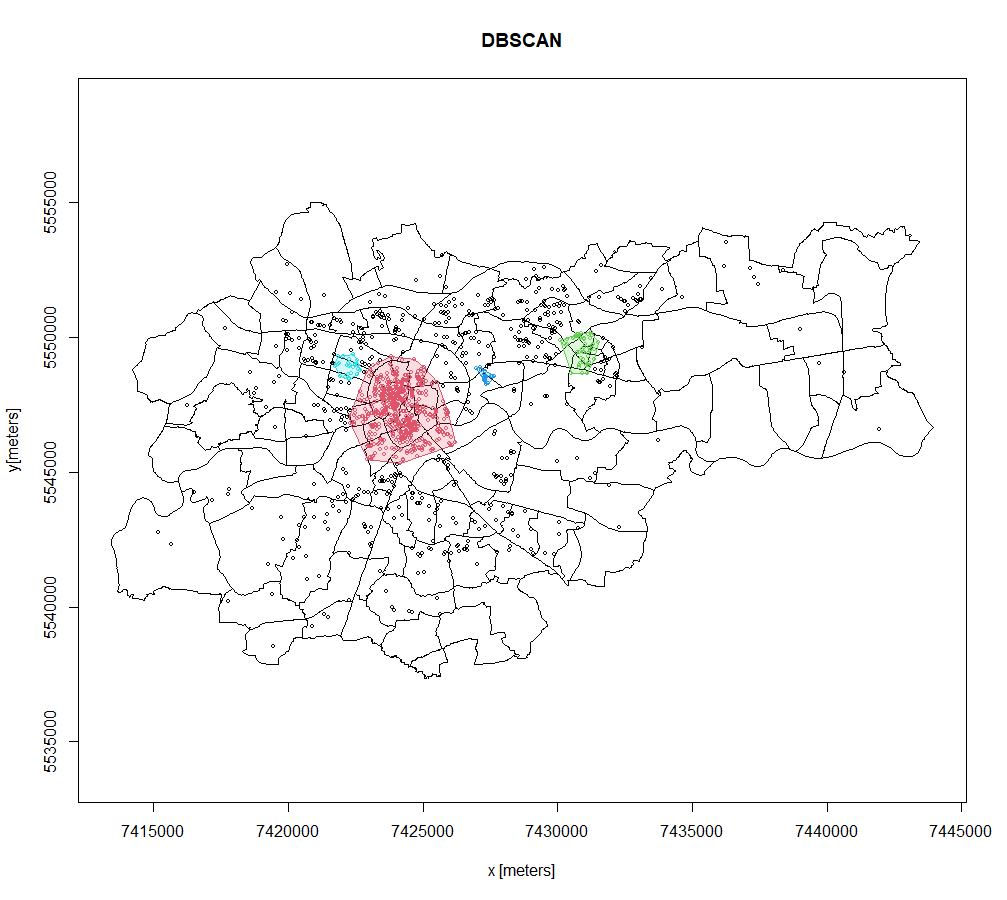
-Nie ma problemu z rozpoznawaniem klastrów o bardzo małej gęstości

-Zmniejszona wrażliwość na parametry wejściowe

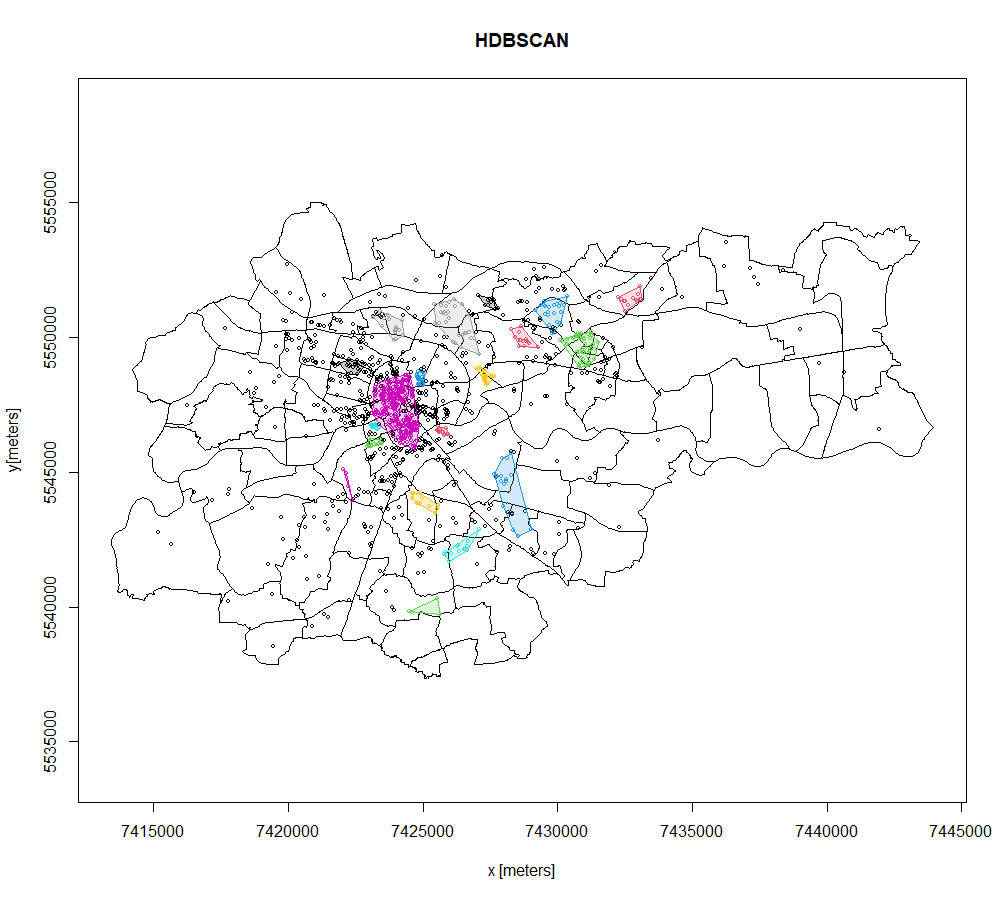
Wady:

-Trudniejszy do zrozumienia w porównaniu z metodą DBSCAN

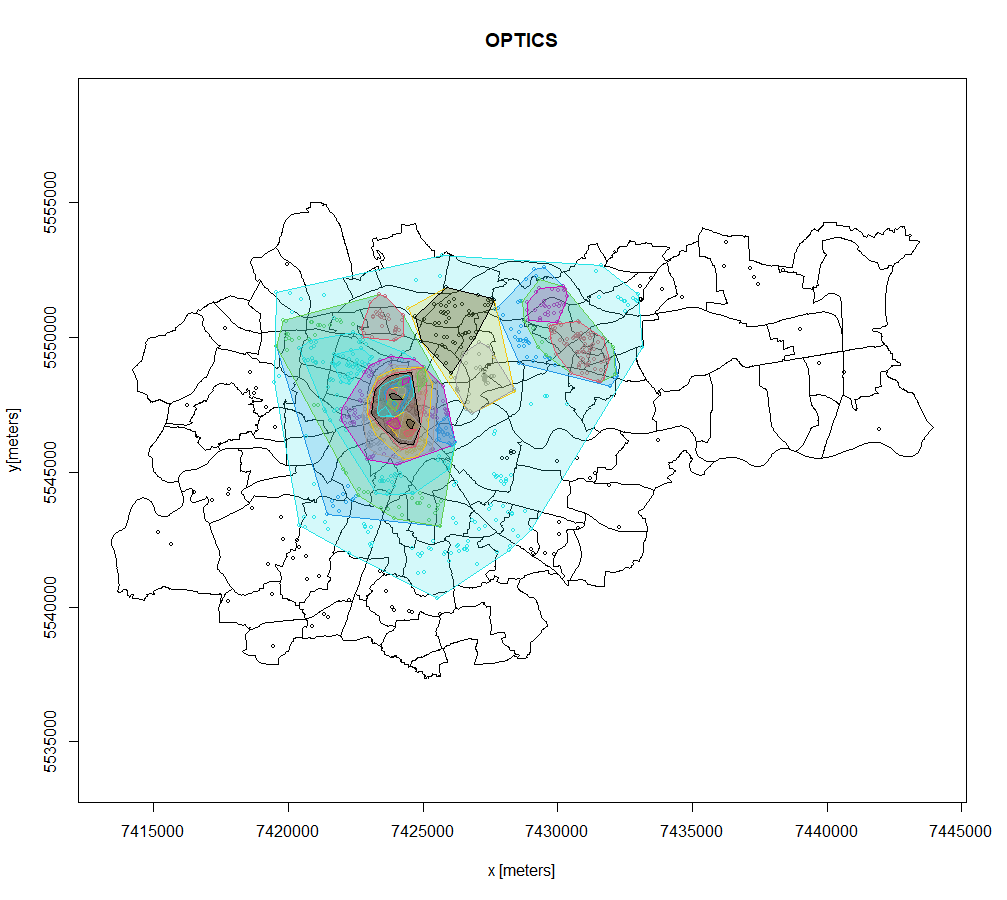
Wyniki klasteryzacji wraz z podaniem użytych parametrów:



*dbscan(point\_xy, eps = 500, minPts = 30)*



*hdbscan(x = point\_xy, minPts = 15)*



*opt = optics(x=point\_xy, eps = 1500, minPts = 30)*

*db\_opt = extractDBSCAN(opt, eps\_cl = .1)*

*db\_opt = extractXi(opt, xi = 0.01)*

Porównanie wyników i wnioski :

Na przedstawionych wykresach możemy zauważyć klasteryzację w zależności od warunków początkowych. Lecz na wszystkich możemy zauważyć że największa intensywność wykroczeń jest w okolicach rynku głównego, oraz w okolicach Bieńczyc Nowych oraz osiedla Teatralne. Dziwić zaś może bardzo niska ilość wykroczeń w okolicach dzielnicy Nowa Huta która w przeszłości miała miano bardzo niebezpiecznej, a obecnie według naszych danych jest jedną z najbezpieczniejszych.

Na wykresach klastrów możemy zauważyć że metoda DBSCAN różni się od metody HDBSCAN, DBSCAN tworzy klastry w oparciu o ilość punktów oraz maksymalny promień od punktu podany w definicji metody, tworząc grupy o różnej gęstości występowania, zaś metoda HDBSCAN tworzy wszystkie możliwe klastry a następnie zostawia tylko te które mają podobną gęstość występowania punktów na swoim obszarze (metoda ta tworzy różne klastry o różnej gęstości punktów, jedynym warunkiem jest ilość punktów jaka musi występować w nie większej odległości niż eps wyznaczony w danym momencie przez program)

Klaster OPTIC tworzy zaś wiele obszarów które mogą się nachodzić na siebie, zlicza on odpowiednie punkty podobnie jak w DBSCAN, uzupełniając to progiem do identyfikacji klastrów oraz progiem stromości do hierarchicznej identyfikacji klastrów przy użyciu metody XI.

Bibliografia:

<https://mateuszgrzyb.pl/grupowanie-gestosciowe-dbscan-teoria/>

<https://en.wikipedia.org/wiki/OPTICS_algorithm>

<https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html>