SPDB

dokumentacja projektu

Mateusz Zawiślak

semestr 14Z

1. Temat projektu.

Porównanie jakości i szybkości grupowania danych przestrzennych przy użyciu wybranego algorytmu grupowania (np. K-means, Clara, PAM, O-cluster) w dwóch wersjach:

- Dane przestrzenne traktowane są tak samo jak inne dane – wyliczana jest jedna odległość (podobieństwo) na podstawie wszystkich dostępnych atrybutów.
- Odległość (podobieństwo) między obiektami jest liczona na podstawie dwóch składowych: odległości przestrzennej i odległości (podobieństwa) obliczonego na podstawie atrybutów opisowych.

Do realizacji projektu można wykorzystać implementacje algorytmów dostępne w Internecie.

2. Badana hipoteza.

Celem poszczególnych eksperymentów będzie zbadanie jakości otrzymanego grupowania obiektów z testowanego zbioru danych w zależności od sposobu traktowania poszczególnych atrybutów opisujących dane przestrzenne. **Hipotezą** stawianą w tym zadaniu jest stwierdzenie, że wyliczanie odległości przestrzennej dla atrybutów opisujących cechy **przestrzenne** może **poprawić jakość** otrzymywanych grupowań. Jednocześnie istnieje podejrzenie o tym, że miary stosowane dla danych przestrzennych mogą spowodować **spadek szybkości** wykonywanych obliczeń.

3. Opis rozwiązania.

a. Opis algorytmu

Poniżej został przedstawiony **pseudokod głównego algorytmu** badającego wpływ zastosowanych miar podobieństwa obiektów dla cech przestrzennych na jakość oraz szybkość grupowania.

1) Wczytaj dane z pliku do pamięci programu.
2) Kolejno dla coraz większy części danych *dataPart*:
    a. Wybierz losowo *dataPart* obiektów z pełnego zbioru danych.
    b. Zbuduj macierze odległości między poszczególnymi parami obiektów w dwóch wersjach:
        i. Dane przestrzenne traktuj jak inne dane.
        ii. Dla danych przestrzennych wylicz adekwatną odległość przestrzenną.

c. Wykonaj wybrany algorytm grupowania (np. POM) obiektów na podstawie zbudowanej macierzy odległości.
d. Zbadaj jakość otrzymanego grupowania.

3) Zbuduj wykresy:
   a. Wykres zależności jakości grupowania od rozmiaru grupowanych danych (jeżeli znany jest idealne, oczekiwane grupowanie danych).
   b. Wykres zależności czasu wykonania grupowania od rozmiaru grupowanych danych.

b. Macierz podobieństwa

Najważniejszą częścią powyższego algorytmu jest zbudowanie **macierzy odległości/podobieństwa** między obiektami. Poniższa tabela przedstawia przykładową macierz odległości:

	Obiekt1	**Obiekt2**	**Obiekt3**	**Obiekt4**	**Obiekt5**
Obiekt1	0				
Obiekt2	0.737318159	0			
Obiekt3	1.891328136	1.448405890	0		
Obiekt4	2.260788513	1.814128250	0.623819206	0	
Obiekt5	2.177160445	1.799533741	0.523588356	0.370073633	0

Budowana macierz będzie oczywiście symetryczna, dlatego tak jak pokazano powyżej wystarczy uzupełnić jedynie połowę macierzy.

c. Opis sposobów obliczania podobieństwa obiektów.

Odległość *d* między dwoma obiektami (zawierającymi *N* atrybutów) dla przypadku traktowania danych przestrzennych tak samo jak innych danych opisowych będzie równa sumie przeskalowanych podobieństw pomiędzy poszczególnymi cechami tych dwóch obiektów (1).

$$d(obiekt1, obiekt2) = \sum_{j}^{N} d(atrybut_j) \quad (1)$$

W drugim przypadku odległość między dwoma obiektami będzie równa sumie dwóch składowych (dla atrybutów przestrzennych oraz atrybutów opisowych) (2):

$$d(obiekt1, obiekt2) = w_1 \cdot d(atr.nie\ przestrzenne) + w_2 \cdot d(atr.przestrzenne) \quad (2)$$

Wektor wag $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ składa się z właściwych wag w_1 oraz w_2 z jakimi są sumowane składowe dla atrybutów przestrzennych oraz atrybutów nie przestrzennych (opisowych). Pozwala to na określenie tego, jak duży nacisk powinien być kładziony na poszczególne składowe w trakcie grupowania.

d. Opis sposobów obliczania podobieństwa poszczególnych typów atrybutów.

Poniżej przedstawiono sposoby wyliczania odległości pomiędzy wartościami poszczególnych typów atrybutów. Każda z zaproponowanych miar odległości daje przeskalowane wartości odległości między wartościami atrybutów. Oznacza to, że wyliczona odległość jest zazwyczaj dzielona przez maksymalną (zaobserwowaną w badanym zbiorze obiektów) odległość między wartościami dla danego atrybutu. Tak zdefiniowane odległość są z przedziału (0,1). Dla dwóch identycznych wartości miara odległości wynosi 0.

i. Atrybuty numeryczne

Dla atrybutów **typu numerycznego**, których wartości to val_1 oraz val_2 odległość wynosi (3):

$$d(val_1, val_2) = \frac{|val_1 - val_2|}{|val_{max} - val_{\min}|} \tag{3}$$

gdzie:
val_{max} – maksymalna wartość analizowanego atrybutu w grupowanym zbiorze obiektów
val_{min} – minimalna wartość analizowanego atrybutu w grupowanym zbiorze obiektów

ii. Atrybuty reprezentujące ciągi znaków (tj. napisy)

Dla atrybutów, których wartości są **ciągami znaków** s_1 oraz s_2 odległość wynosi (4):

$$d(s_1, s_2) = \frac{odległość\ Levenstheina\ między\ s_1\ oraz\ s_2}{max(długość\ ciągu\ s_1, długość\ ciągu\ s_2)} \tag{4}$$

iii. Atrybuty reprezentujące obiekty na płaszczyźnie dwuwymiarowej

W sytuacji gdy atrybut reprezentuje dowolne obiekty na płaszczyźnie dwuwymiarowej konieczne będzie znalezienie najmniejszego prostokąta *mbb* (ang. *minimum bounding box*) zawierającego dany obiekt. Dla znalezionego *mbb* określamy 4 wartości x_{min}, x_{max}, y_{min}, y_{max} jednoznacznie identyfikujące ten *mbb*.

Odległością przestrzenną między wartościami atrybutu opisującego obiekty na płaszczyźnie będzie długość najkrótszego odcinka łączącego te dwa obiekty podzieloną przez maksymalną możliwą odległość w analizowanym zbiorze obiektów (5).

$$d(obj_1, obj_2) = \frac{długość(najkrótszy\ odcinek\ łączący\ obj_1\ i\ obj_2)}{\sqrt{|X_{max} - X_{min}|^2 + |Y_{max} - Y_{min}|^2}} \tag{5}$$

gdzie:
X_{max} – to największa wartość x_{max} rozważanego atrybutu w analizowanym zbiorze obiektów
X_{min} – to najmniejsza wartość x_{min} rozważanego atrybutu w analizowanym zbiorze obiektów
Y_{max} – to największa wartość y_{max} rozważanego atrybutu w analizowanym zbiorze obiektów
Y_{max} – to najmniejsza wartość y_{max} rozważanego atrybutu w analizowanym zbiorze obiektów

Gdy **dane przestrzenne traktowane są tak samo jak inne dane** obiekty przestrzenne będą traktowane jak 4 niezależne, składowe atrybuty: x_{min}, x_{max}, y_{min}, y_{max}. Odległość między dwoma obiektami będzie średnią czterech odległości pomiędzy wartościami „nowych" atrybutów dwóch analizowanych obiektów. Sposób wyliczania odległości dla atrybutów numerycznych (jakimi są atrybuty x_{min}, x_{max}, y_{min}, y_{max} pokazano w jednym z poprzednich punktów) (6).

$$d(obj_1, obj_2) = \frac{d(x_{max}) + d(x_{min}) + d(y_{max}) + d(y_{min})}{4} \tag{6}$$

iv. Atrybuty reprezentujące punkt we współrzędnych geometrycznych

Odległością przestrzenną między punktami we współrzędnych geograficznych będzie odległość Haversine między dwoma analizowanymi obiektami podzielona przez maksymalną możliwą do zaobserwowania w analizowanym zbiorze (7).

$$d(obj_1, obj_2) = \frac{haversine(obj_1, obj_2)}{\max(\sum_i \sum_{j \neq i} haversine(obj_i, obj_j))} \tag{7}$$

Gdy **dane przestrzenne traktowane są tak samo jak inne dane** punkty we współrzędnych geograficznych będą rozbijane na dwie składowe: szerokość geograficzną (*lat*) oraz długość geograficzną (*long*). Odległością finalną będzie średnia odległości policzonych niezależnie dla każdego z atrybutów numerycznych *lat* oraz *long* (8).

$$d(obj_1, obj_2) = \frac{d(lat) + d(long)}{2} \tag{8}$$

v. Atrybuty reprezentujące punkt na płaszczyźnie dwuwymiarowej

Odległością przestrzenną między punktami będzie stosunek ich odległości euklidesowej do maksymalnej odległości możliwej do zaobserwowania w analizowanym zbiorze (9).

$$d(point_1, point_2) = \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{\sqrt{(x_{max} - x_{min})^2 + (y_{max} - y_{min})^2}} \tag{9}$$

Gdy **dane przestrzenne traktowane są tak samo jak inne dane** punkty *(x,y)* będą rozbijane na dwie składowe *x* oraz *y*. Odległością finalną będzie średnia odległości policzonych niezależnie dla każdej składowej numerycznej *x* oraz *y* (10).

$$d(point_1, point_2) = \frac{d(x) + d(y)}{2} \tag{10}$$

e. Miara jakości grupowania.

Jeżeli dla badanego zbioru danych **nie znane jest idealne grupowanie** ocena otrzymanego grupowania będzie dokonana **empirycznie** przez użytkownika.

Jeżeli dla badanego zbioru danych znamy **idealne grupowanie** to możemy sprawdzić jakość otrzymanego grupowania za pomocą miary znanej jako ***Rand index.*** Zatem wyniki grupowania zostaną porównane z oczekiwanym, idealnym grupowaniem.

Rand index

Dany jest zbiór n obiektów $S = \{o_1, \ldots, o_n\}$, które mają zostać pogrupowane w r klastrach oraz dwa grupowania zbioru S: $X = \{X_1, \ldots, X_r\}$ oraz $Y = \{Y_1, \ldots, Y_s\}$. Należy zdefiniować następujące zmienne:

- *a* – liczba par obiektów ze zbioru S które są w tym samym klastrze w grupowaniu X, a także są w tym samym klastrze w grupowaniu Y,
- *b* – liczba par obiektów ze zbioru S które są w innym klastrze w grupowaniu X, a także są w innym klastrze w grupowaniu Y.

Dla tak zdefiniowanych zmiennych wartość Rand index wynosi:

$$R = \frac{a+b}{\binom{n}{2}}$$

Porównując otrzymane grupowanie z oczekiwanym klastrowaniem wartość Rand index wynosi:

- 0 – dla bardzo złego grupowanie
- 1 – dla idealnego grupowania.

4. Opis implementacji.

Przedstawione w poprzednim punkcie rozwiązanie zostało zaimplementowane z wykorzystaniem **języka R**. W projekcie zostały wykorzystane gotowe implementacje algorytmów grupowania dostępne w **pakiecie *cluster***. Z tego też powodu **napisany program NIE jest uzależniony od konkretnego algorytmu grupowania**. Zmiana algorytmu na inny wymaga jedynie wywołania tegoż algorytmu dostępnego w podanym pakiecie. W trakcie implementacji programu autor niniejszego projektu analizował zachowanie algorytmu **PAM** (ang. *Partitioning Around Medoids*).

Dane są wczytywane z plików typu *CSV*. Metoda badająca zachowanie algorytmu grupowania w zależności od różnego traktowania atrybutów przestrzennych wymaga podania struktury opisującej dane. W ramach struktury opisującej testowany zbiór danych można podać następujące dane:

- *class.column* – nazwa kolumny/atrybutu zawierającej oczekiwane grupowanie,
- *numerical.columns* – lista kolumn/atrybutów zawierających atrybuty numeryczne,
- *string.columns* – lista kolumn/atrybutów, których wartościami są napisy (np. kod pocztowy miasta),
- *points* – lista kolumn/atrybutów, których wartości to punkt na płaszczyźnie – trzeba podać kolumnę zawierającą współrzędną X oraz kolumnę zawierającą współrzędną Y,
- *geographic.coordinates* – lista kolumn/atrybutów, których wartości to współrzędne geometryczne – trzeba podać kolumnę zawierającą długość geometryczną (ang. *longitude*) oraz kolumnę zawierającą szerokość geometryczną (ang. *latitude*),
- *polygons* – lista atrybutów, których wartości reprezentują obiekty na płaszczyźnie dwuwymiarowej.

Przykładowy opis danych:

```
example.data.description <- list(
  "class.column" = "region",
  "ignore.columns" = c("cityName"),
  "string.columns" = c("postalCode"),
  "geographic.coordinates" = list(list("long" = " longitude ", "lat" = "latitude")),
  "polygons" = c("geo.parcel")
 )
```

5. Opis testów.

W trakcie realizacji projektu przeprowadzono testy, których rezultat pozwala określić jakość otrzymywanych grupowań oraz czas wykonania w stosunku do liczby grupowanych obiektów. Do testów wykorzystano następujące zbiory danych:

Baza danych stref mieszkań Bostonu [2]

Liczba przykładów	Liczba wszystkich atrybutów
506	17

Opis: W tym zbiorze danych przechowywane są informacje o wybranych strefach mieszkań Bostonu oraz jego okolic. Poza atrybutami opisującymi położenie stref są tam również informacje m.in. o: średniej liczbie pokoi w mieszkaniu, poziomie dostępności do autostrad, średniej wysokości podatku od mieszkania itp. W tym zbiorze znany jest idealny podział, znane są bowiem nazwy miast lub dzielnic, w których położone są poszczególne strefy mieszkań.

Policyjne raporty przestępczości stanu Missisipi [3]

Liczba przykładów	Liczba wszystkich atrybutów
82	11

Opis: W tym zbiorze danych przechowywane są policyjne, zagregowane informacje o przestępstwach popełnionych w stanie Missisipi. Poza atrybutem przestrzennym opisującym geometrię danej jednostki administracyjnej w tej bazie znajdziemy także informacje o m.in.: wydatkach policji, przestępczości oraz innych społeczno-ekonomicznych cechach tej jednostki.

Z każdego ze zbiorów danych, w kolejnych iteracjach algorytmu wybierano kolejno od 10% do 100% losowych przykładów.

Ze względu na heurystykę występującą w badanych algorytmach selekcji atrybutów każdy z testów został powtórzony **10 razy** a wyniki uśrednione.

6. Wyniki testów.

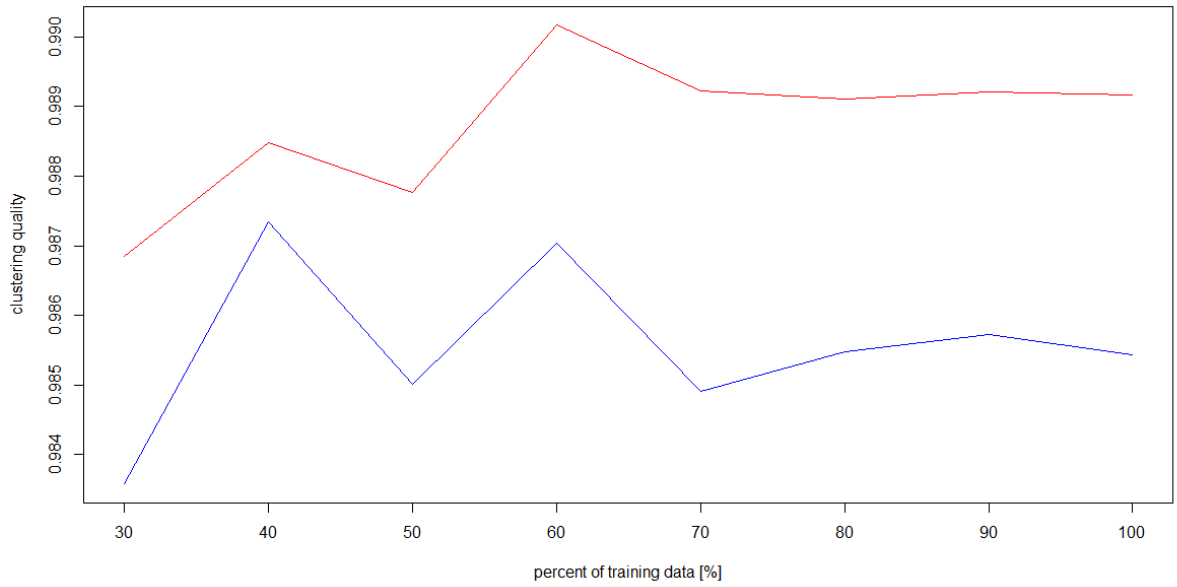
Prezentowane wykresy prezentujące przeprowadzone testy zawierają dwie linie:

- **Czerwona** – wyniki dla danych przestrzennych,
- **Niebieska** – wyniki dla danych przestrzennych traktowanych jak inne dane.

a. Baza danych mieszkań Bostonu [2]

W tej bazie danych jest zdefiniowany idealny podział, w którym każdemu z mieszkań zostało przyporządkowane miasto bądź dzielnica w którym się znajduje.

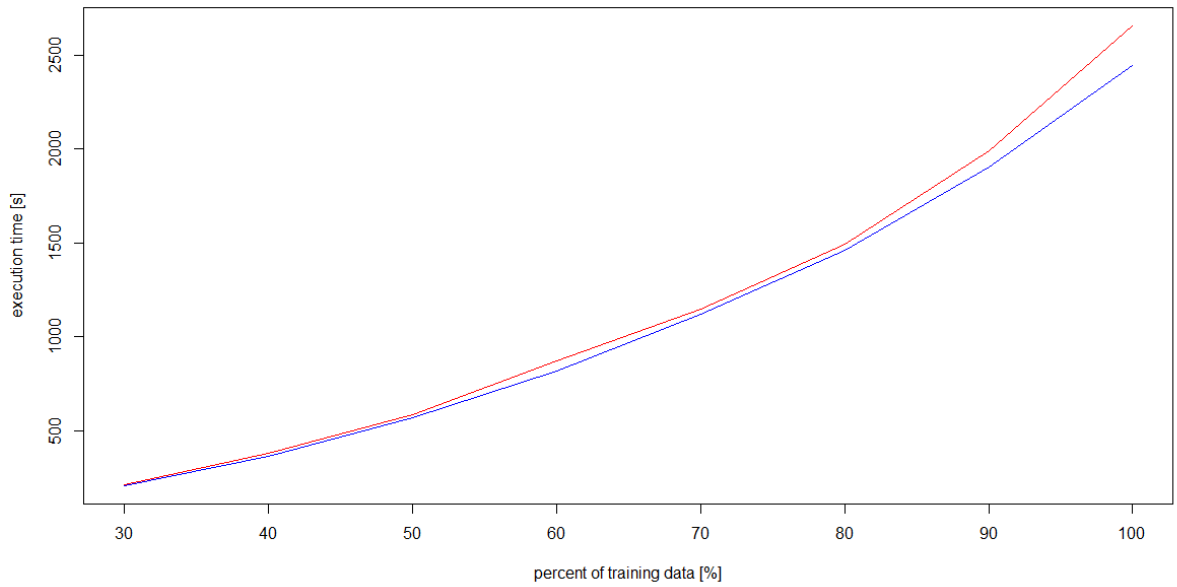
Poniższy wykres (Rys. 1) przedstawia jakość otrzymywanych grupowań w zależności od liczby grupowanych obiektów:



Rysunek 1 Jakość grupowania mieszkań Bostonu.

Jak widać jakość grupowania, niezależnie od wybranej liczby elementów (mieszkań) zawsze grupowanie traktujące dane przestrzenne jak inne dane było gorsze.

Poniższy wykres (Rys. 2) przedstawia szybkość grupowania w zależności od liczby grupowanych obiektów:



Rysunek 2 Szybkość grupowania mieszkań Bostonu.

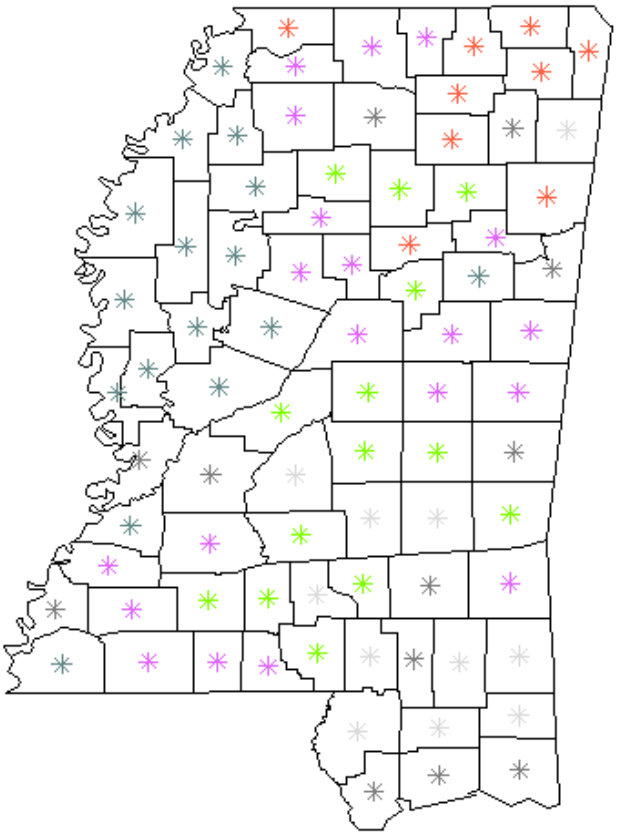
Zgodnie z oczekiwaniami w przypadku traktowania danych jak inne dane czas grupowania był niższy niż w przypadku konieczności liczenia odległości przestrzennych.

b. Baza policyjnych raportów stanu Missisipi [3]

Jednym z przykładowych zadań grupowania jakie można postawić przed omawianym zbiorem danych to: podział jednostek administracyjnych stanu Missisipi na 6 regionów, którymi będą mogły być skutecznie zarządzane przez policję stanową.

W pierwszej próbie wektor wag to $\begin{bmatrix}1\\1\end{bmatrix}$ czyli atrybuty przestrzenne mają taką samą wagę jak atrybuty nie przestrzenne.

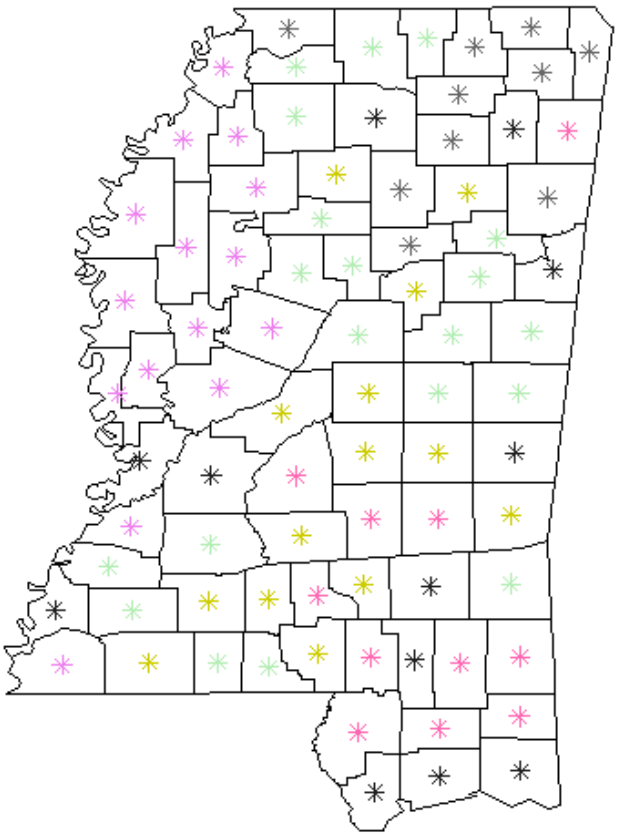
Podział w przypadku traktowania atrybutów przestrzennych jak inne dane.



Rysunek 3 Podział dla wektora 1:1 oraz traktowania atrybutów przestrzennych jak inne dane.

Jak widać na Rys. 3 zaproponowany podział zawiera jednostki, które są niespójne powierzchniowo. Wynika to z tego, że dane przestrzenne były taktowane jak inne dane, więc algorytm nie przykładał dużej uwagi do położenia geometrycznego jednostek.

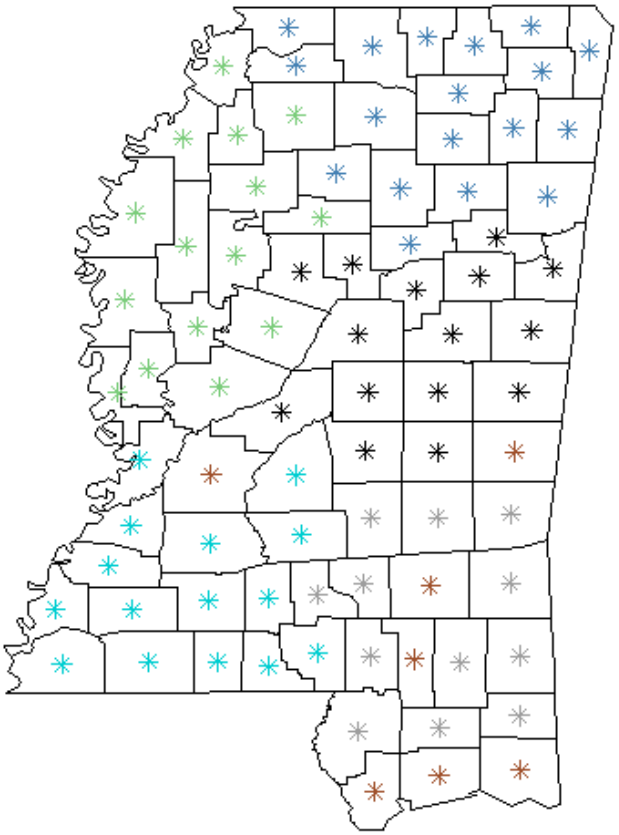
Podział w przypadku liczenia odległości przestrzennych dla atrybutów przestrzennych.



Rysunek 4 Podział dla wektora 1:1 oraz wyróżnionych danych przestrzennych.

Jak widać na Rys. 4 zaproponowany podział zawiera jednostki, które są nadal są niespójne powierzchniowo. Wynika to z tego, że waga dla atrybutów przestrzennych była taka sama jak w przypadku atrybutów opisowych. Z racji tego, że atrybutów opisowych było znacznie więcej atrybuty geometryczne nie miały dużego znaczenia. W poszukiwaniu podziału mocniej uwzględniającego położenie geograficzne jednostek zwiększamy wektor wag do $\begin{bmatrix}1\\7\end{bmatrix}$.

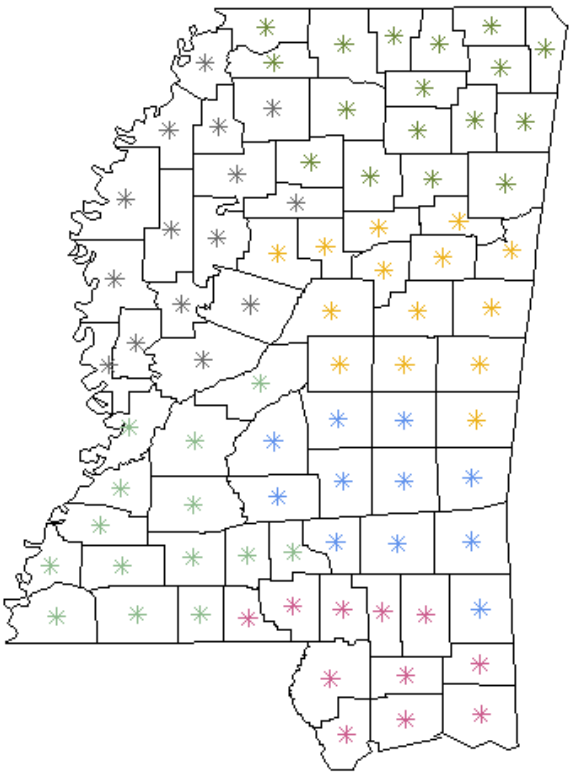
Podział w przypadku liczenia odległości przestrzennych dla atrybutów przestrzennych.



Rysunek 5 Podział dla wektora 1:7 oraz wyróżnionych danych przestrzennych.

Jak widać na Rys. 5 zaproponowany podział wygląda już dobrze, ponieważ zaproponowane jednostki policji są mocno zależne położenia. Granice między jednostkami zależą oczywiście od rodzaju przestępstw oraz pozostałych danych opisowych. Niestety „brązowa" jednostka nadal jest mocno rozstrzelona po całym stanie Missisipi. W ostatniej próbie zwiększymy zatem wektor wag do $\begin{bmatrix} 1 \\ 14 \end{bmatrix}$.

Podział w przypadku liczenia odległości przestrzennych dla atrybutów przestrzennych.



Rysunek 6 Podział dla wektora 1:14 oraz wyróżnionych danych przestrzennych.

Jak widać na Rys. 6 jednostki są już spójne. W takim problemie władze policji musiałby podjąć decyzję czy zależy im bardziej na spójności geograficznej jednostek czy na podziale opartym na specyfice popełnianych przestępstw.

7. Wnioski.

Otrzymane rezultaty potwierdzają słuszność postawionej na wstępnie hipotezy. Traktowanie danych przestrzennych jak pozostałe dane może prowadzić do pogorszenia jakości otrzymanych grupowań. Danym przestrzennym powinny zatem towarzyszyć metody poprawnego liczenia ich podobieństwa. Ponadto testy potwierdziły również przypuszczenia o tym, że prawdopodobieństwa adekwatne do danych przestrzennych skutkują dłuższym czasem wykonania algorytmów grupowania.

Dla każdego problemu z osobna należy zatem podjąć decyzję o tym, czy dla danych przestrzennych warto stosować odległości/podobieństwa dostosowane do danych przestrzennych. Istnieją bowiem problemy, w których najważniejsza jest jakość otrzymanego grupowania, zaś w innych przypadkach nacisk kładziony jest na szybkość grupowania.

8. Literatura.

[1] Miara jakości grupowania Rand index, http://en.wikipedia.org/wiki/Rand_index

[2] Baza danych mieszkań Bostonu, http://lib.stat.cmu.edu/datasets/boston_corrected.txt

[3] Mississippi Police Crime Reports, http://www.dm.unibo.it/~simoncin/police.html