Projekt z porządkowania liniowego i analizy skupień

Jakub Zięba

# Cel pracy

W trakcie wykonywania projektu postaram się przyjrzeć różnym województwom w Polsce pod względem zmiennych związanych z szeroko pojętym dostępem do medycyny oraz jej poziomem. Celem projektu jest wyłonienie najlepszych województw pod względem poziomu i dostępu do medycyny na podstawie danych zebranych przez Główny Urząd Statystyczny.

# Opis i analiza wybranych zmiennych

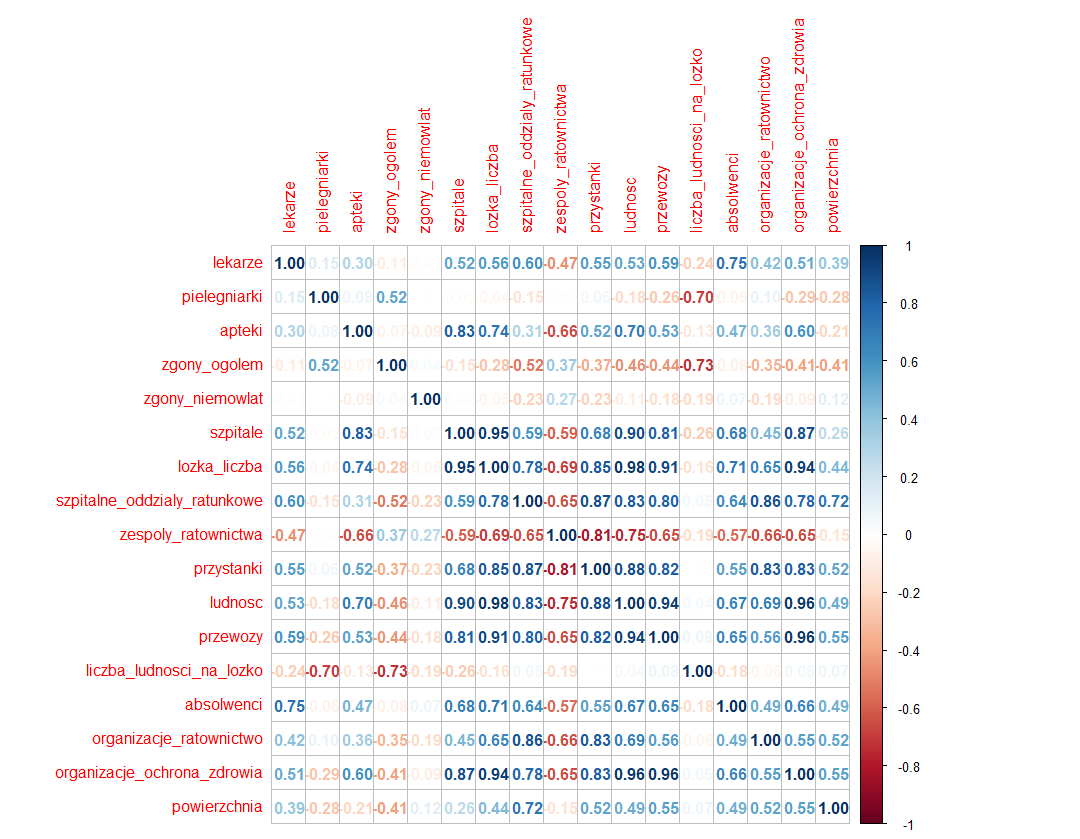
Badanie nad danymi rozpoczynam od wybrania dość dużej próbki danych. W tym przypadku jest to 17 zmiennych:

1. Lekarze, wyrażona w liczbie osób;
2. Pielęgniarki, wyrażona w liczbie osób;
3. Apteki, wyrażona w liczbie takich obiektów;
4. Zgony ogółem, wyrażona w liczbie zgonów na tysiąc urodzeń żywych;
5. Zgony niemowląt, wyrażona w liczbie zgonów niemowląt na tysiąc urodzeń żywych;
6. Szpitale, wyrażona w liczbie obiektów;
7. Łóżka szpitalne, wyrażona w liczbie łóżek;
8. Szpitalne oddziały ratunkowe, wyrażona w liczbie oddziałów;
9. Zespoły ratownictwa medycznego, wyrażona w liczbie takich oddziałów;
10. Przystanki, wyrażona w liczbie przystanków komunikacji miejskiej;
11. Ludność, wyrażona w tys. osób;
12. Przewozy pasażerskie, wyrażona w mln osób;
13. Liczba ludności na łóżko szpitalne, wyrażona w liczbie osób na jedno łóżko szpitalne w szpitalach ogólnych;
14. Absolwenci uczelni medycznych, wyrażona w liczbie osób;
15. Organizacje non profit ratownictwa, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
16. Organizacje non profit ochrona zdrowia, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
17. Powierzchnia, wyrażona w km2;

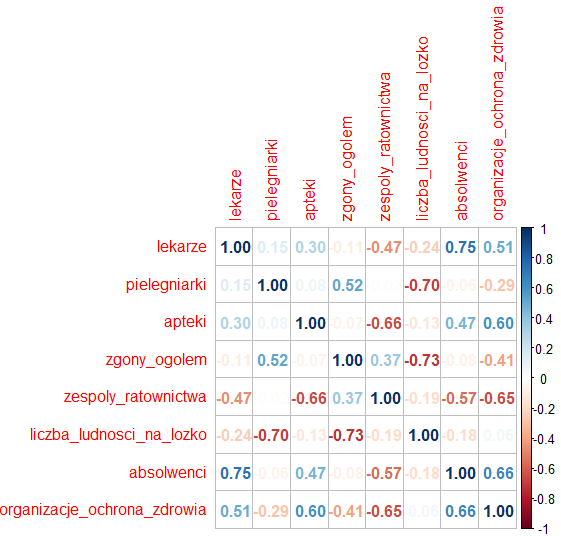
Po przejrzeniu danych można zauważyć duże rozbieżności jeśli chodzi o rzędy wielkości. Z tego powodu, oraz z uwagi na chęć wykluczenia sytuacji, w której najlepiej wypadają województwa zwyczajnie największe, dane poddałem obróbce. Obrobione zmienne prezentują się następująco:

1. Lekarze, wyrażona w liczbie osób na tys. ludności;
2. Pielęgniarki, wyrażona w liczbie osób na tys. ludności;
3. Apteki, wyrażona w liczbie obiektów na km2;
4. Zgony ogółem, wyrażona w liczbie zgonów na tysiąc urodzeń żywych;
5. Zgony niemowląt, wyrażona w liczbie zgonów niemowląt na tysiąc urodzeń żywych;
6. Szpitale, wyrażona w liczbie obiektów;
7. Łóżka szpitalne, wyrażona w liczbie łóżek;
8. Szpitalne oddziały ratunkowe, wyrażona w liczbie oddziałów;
9. Zespoły ratownictwa medycznego, wyrażona w liczbie zespołów na tys. ludności;
10. Przystanki, wyrażona w liczbie przystanków komunikacji miejskiej;
11. Ludność, wyrażona w tys. osób;
12. Przewozy pasażerskie, wyrażona w mln osób;
13. Liczba ludności na łóżko szpitalne, wyrażona w liczbie osób na jedno łóżko szpitalne w szpitalach ogólnych;
14. Absolwenci uczelni medycznych, wyrażona w liczbie osób;
15. Organizacje non profit ratownictwa, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
16. Organizacje non profit ochrona zdrowia, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
17. Powierzchnia, wyrażona w km2;

Następnie zajmuję się wyborem odpowiednich zmiennych. Dokonuję tego najpierw przez zbadanie współczynników korelacji między zmiennymi.



Aby wyeliminować zmienne bardzo mocno skorelowane, zdecydowałem się nie uwzględniać w zbiorze zmiennych odpowiedzialnych za liczbę szpitali, liczbę łóżek, liczbę szpitalnych oddziałów ratunkowych, liczbę przewozów pasażerskich, odsetek organizacji non profit związanych z ratownictwem, liczbę przystanków oraz liczbę zgonów niemowląt. Nie uwzględniam również ludności ani powierzchni, jako że te zmienne zostały wcześniej wykorzystane do wyliczenia innych zmiennych.



Po zmianie zbioru zmiennych macierz korelacji prezentuje się dużo lepiej, nie występują już wartośći korelacji na poziomie wyższym niż 0,9. Dla wybranych zmiennych weryfikuję również współczynniki zmienności.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Każdy ze współczynników zmienności jest na poziomie co najmniej 0.1, zatem możemy przyjąć taki zbiór danych do dalszej pracy.

Następnie skupimy się na podstawowych statystykach opisowych wybranego zbioru zmiennych, w stosunku do całości kraju.

Obraz zawierający tekst, zrzut ekranu, czarne, Czcionka

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

W przypadku zmiennej lekarze, średnia w całym kraju wynosi ok. 3,9 lekarza na tysiąc ludności, przy odchyleniu standardowym na poziomie ok. 0,77. Jest to dość wysokie odchylenie standardowe względem średniej. Mediana natomiast jest na poziomie lekko wyższym od średniej, co może być dobrym znakiem jeśli chodzi o dostęp do lekarzy w kraju. Wartość minimalną zmienna przyjmuje dla województwa Lubuskiego, a maksymalną dla Mazowieckiego.

Jeśli chodzi o pielęgniarki, to tutaj średnia kształtuje się na poziomie 8.17, natomiast odchylenie standardowe na poziomie 1.12, nie jest to więc aż tak duża wartość względem średniej. Podobnie jak w przypadku zmiennej powyżej, średnia i mediana są do siebie mocno zbliżone. Wartość minimalną zmienna przyjmuje w przypadku województwa Wielkopolskiego, a maksymalną dla województwa Świętokrzyskiego.

W przypadku aptek, średnia i mediana również są do siebie zbliżone. Odchylenie standardowe na poziomie 0.02 jest też bardzo bliskie średniej, która wynosi około 0,04. Najmniej aptek na kilometr kwadratowy możemy znaleźć w województwie Warmińsko-Mazurskim, najwięcej natomiast w województwie Śląskim.

Dla zgonów ogółem na tysiąc urodzeń żywych wartość minimalną znajdujemy w województwie Pomorskim, od którego bardzo blisko plasuje się województwo Małopolskie. Wartość maksymalną natomiast znajdujemy przy województwie Świętokrzyskim.

Jeśli chodzi o zespoły ratownictwa medycznego oraz ludność na łóżko szpitalne, to ich współczynniki zmienności są na podobnym poziomie, odpowiednio 15% i 12%, co nie różni się znacząco od 13% współczynnika zmienności dla pielęgniarek.

Prawdopodobnie najciekawszą zmienną będą tu absolwenci uniwersytetów medycznych. Jako że nie w każdym województwie mamy uczelnie medyczną, zarówno odchylenie standardowe jak i minimum zachowują się dość ekstremalnie. Odchylenie standardowe minimalnie przekracza średnią, a minimum jest równe zero i obejmuje wiele województw. Wartość maksymalna w wysokości prawie 2300 absolwentów znalazła się w województwie Śląskim.

W przypadku ostatniej zmiennej, odsetka organizacji non profit wśród organizacji ochrony zdrowia współczynnik zmienności również ma wysoką wartość. Rozpiętość danych również jest bardzo wysoka, jako że wartość minimalna wynosi jedynie 1.6%, natomiast maksymalna aż 18%.

Obraz zawierający linia, diagram, Wykres, tekst

Opis wygenerowany automatycznie

Obraz zawierający linia, tekst, Wykres, diagram

Opis wygenerowany automatycznie

Wykresy gęstości przedstawione powyżej pozwalają nam zobaczyć podstawową strukturę rozkładu danych. W większości przypadków wartości są raczej dość mocno skupione wokół średniej, jednak w żadnym z nich nie można mówić o rozkładzie stricte normalnym. Zdarzają się również rozkłady o grubych ogonach, szczególnie jeśli chodzi o prawe ogony. W przypadku lewych ogonów taka sytuacja nie występuję.

Obraz zawierający tekst, diagram, Plan, Równolegle

Opis wygenerowany automatycznie

Jeśli chodzi o wartości odstające, to sprawdzając pojedyncze zmienne, znajdujemy takie wartości w przypadku zmiennych dotyczących organizacji non profit, zespołów ratownictwa, ludności na łóżko szpitalne oraz aptek. Nie jesteśmy natomiast w stanie stwierdzić na tej podstawie, czy któryś z obiektów jest outlierem w przypadku analizy wielu zmiennych jednocześnie.

# Wybrane techniki porządkowania liniowego

Pierwszym krokiem w porządkowaniu liniowym jest zamiana zmiennych na stymulanty. Zmienne określiłem jako:

1. Lekarze – stymulanta;
2. Pielęgniarki – stymulanta;
3. Apteki – stymulanta;
4. Zgony ogółem – destymulanta;
5. Zespoły ratownictwa medycznego – stymulanta;
6. Ludność na łóżko szpitalne – destymulanta;
7. Absolwenci uczelni medycznych – stymulanta;
8. Organizacje non profit dot. ochrony zdrowia – nominanta na poziomie 9%.

W takim wypadku musimy zamienić zmienne 5, 6 i 8 na stymulanty. Robimy to mnożąc destymulanty przez -1 oraz stosując następujący wzór dla nominanty:

Obraz zawierający tekst, Czcionka, pismo odręczne, biały

Opis wygenerowany automatycznie

Następnym krokiem jest standaryzacja danych. Również wykonuje się ją dla większości metod, jednak dla metody standaryzowanych sum i Hellwiga stosuje się inny wzór niż dla metody TOPSIS, dla metody sumy rang standaryzację pomijamy.

## Metoda standaryzowanych sum

Wykonujemy standaryzację według wzoru: Obraz zawierający Czcionka, biały, diagram, symbol

Opis wygenerowany automatycznie

Dane po standaryzacji wyglądają następująco:

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Następnie wyliczamy i standaryzujemy syntetyczną miarę z uwzględnieniem wag, która pozwoli nam wybrać najlepsze z województw. Robimy to według wzoru:

Obraz zawierający Czcionka, biały, tekst, typografia

Opis wygenerowany automatycznieObraz zawierający tekst, Czcionka, biały, linia

Opis wygenerowany automatycznie

Wynikiem tych działań jest następująca tablica danych:

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatyczniePozostało nam jedynie posortować województwa według miary wyliczonej w kolumnie S2:

## Metoda sumy rang

W metodzie sumy rang, po zamianie zmiennych na stymulanty, musimy nadać obiektom rangi w obrębie poszczególnych zmiennych. Rangi te są rosnące, tj. najmniejsza ranga oznacza najlepszą zmienną.

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Następnie uwzględniając wagi sumujemy rangi dla poszczególnych województw. Tak otrzymaną tabelę sortujemy według ww. sum. Najmniejsza suma oznacza obiekt najlepszy.

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

## Metoda Hellwiga

W metodzie Hellwiga, po zamianie zmiennych na stymulanty, dokonujemy standaryzacji:

Obraz zawierający Czcionka, biały, diagram, symbol

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Następnie tworzymy wzorzec, tzn. najlepszy obiekt, posiadający maksymalną wartość w przypadku każdej zmiennej:

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Teraz obliczamy odległości od poszczególnych badanych obiektów do wyżej wyliczonego wzorca oraz „możliwie daleką” odległość, której użyjemy do wyliczenia miar dla obiektów.

Obraz zawierający pismo odręczne, Czcionka, kaligrafia, biały

Opis wygenerowany automatycznieObraz zawierający Czcionka, biały, typografia, kaligrafia

Opis wygenerowany automatycznie

Kiedy już wyliczymy odległości oraz odległość d0 na ich podstawie, wykorzystujemy następujący wzór aby wyliczyć kolejną miarę dla obiektów:

Obraz zawierający Czcionka, biały, numer, design

Opis wygenerowany automatycznie

Na podstawie wyliczonej miary sortujemy nasze województwa:

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

## Metoda TOPSIS

W przypadku metody TOPSIS, również standaryzujemy dane zamienione na symulanty, jednak robimy to według innego wzoru:

Obraz zawierający Czcionka, biały, diagram, pismo odręczne

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, menu, Czcionka

Opis wygenerowany automatycznie

Następnie, podobnie jak w metodzie Hellwiga, wyznaczamy wzorzec, czyli obiekt składający się z wartości najlepszych. Jednak dodatkowo tworzymy też antywzorzec, czyli obiekt, który charakteryzują wartości najgorsze.

Obraz zawierający tekst, Czcionka, biały, design

Opis wygenerowany automatycznie

Następnie na podstawie wzorca i antywzorca, obliczamy odległości obiektów od ww. tworów.

Obraz zawierający Czcionka, pismo odręczne, tekst, biały

Opis wygenerowany automatycznie

Wykorzystując wyżej wyliczone odległości, wyliczamy miarę współczynnika rankingowego, według którego sortujemy nasze województwa w poszukiwaniu najlepszego:

Obraz zawierający Czcionka, diagram, biały, design

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

# Analiza wyników porządkowania liniowego