Statystyczna eksploracja danych wyborczych
praca inżynierska, autor **Katarzyna Żogała**
opiekun: dr inż. Julian Sienkiewicz

W swojej pracy zajmowałam się analiza danych związanych z wyborami do Sejmu RP przeprowadzonymi we wrześniu roku 2005 oraz w październiku roku 2007. Głównym celem analizy było poszukiwanie zależności w badanym zbiorze przy zastosowaniu takich miar jak entropia Shannona oraz dywergencja Jensena-Shannona (JSD). Uzyskane wyniki opisują zachowania oraz zmiany zachowań elektoratów poszczególnych regionów.

W pierwszym rozdziale przybliżam idee analizowania danych wyborczych przez fizyków. Po krótkim wprowadzeniu, opisuję cztery przykładowe prace, których autorzy szukali uniwersalności lub bezskalowości w badanych zbiorach, bądź też posługiwali się technikami eksploracji danych, podobnymi do stosowanych przeze mnie. Kończąc rozdział uzasadniam wybór użytych narzędzi informatycznych (PostreSQL oraz Python z komponentami pomocniczymi). W drugim rozdziale podaję źródła danych. Tutaj również tłumaczę relacje pomiędzy poszczególnymi poziomami agregacji danych (województwami, powiatami i gminami oraz okręgami wyborczymi i obwodami głosowania). Ta cześć pracy jest także miejscem, w którym opisuję wykonaną relacyjną bazę danych. Ostatni z rozdziałów zawiera kwintesencję całej pracy. Po wstępie dotyczącym zastosowanych miar, czyli entropii Shannona i dywergencji Jensena-Shannona, opisuję otrzymane wyniki. Stwierdziłam bliskie podobieństwo pomiędzy siedmioma z dziesięciu utworzonych histogramów unormowanej entropii rozkładów głosów na partie na poszczególnych poziomach agregacji z roku 2005 i 2007. Badane za pomocą dywergencji Jensena-Shannona różnice w rozkładach unormowanych entropii dotyczących odpowiadających sobie poziomom w latach 2005 i 2007 okazały się niewielkie. Pomiędzy unormowana JSD liczona między wynikami powiatu z obydwu wyborów, a jego liczbą mieszkańców, czy też powierzchnią nie stwierdziłam korelacji Pearsona. Dla obydwu wyborów wykryłam i opisałam dodatnie liniowe korelacje pomiędzy frekwencją w powiecie, a jego liczbą mieszkańców oraz pomiędzy maksymalną procentową ilością głosów oddanych na jedną partie w powiecie, a liczba jego mieszkańców.

Wykonując tę pracę wykorzystałam wywodzące się z fizyki entropię Shannona i dywergencję Jensena-Shannona. Zastosowanie ich do poszukiwania prawidłowości w danych wyborczych nie jest powszechną praktyką. Stworzona baza danych w przyszłości ma posłużyć do dalszych analiz zgromadzonego zbioru.