Analysis of Census Data With Clustering Techniques

Jonah Williams¹,*

¹University of Nebraska at Omaha
*Contact author: jonahwilliams@unomaha.edu

Keywords: Clustering, Python, Demographic, KL-Divergence

Demographic information from the United States Census [1] provides a great deal of insight into the nature of the people who live in a given area. This information can be used to inform advertising campaigns, political strategies, to identify good locations for new businesses, among many others. We require a method to establish which areas are similar to each other enough to be consider part of one ‘neighborhood’ that is extendable and robust. While criteria such as ‘majority/minority’ might work for racial demographics, they result in a given amount of information loss and do not work well for age groups or language groups. Here, we consider groups of of variables such as race, age, language spoken as realizations from different multinomial distributions. This informs the use of symmetric f-divergences to measure ‘similarity’ across different areas. We then compare different clustering techniques using a novel method for updating cluster centers. This analysis opens up the possibility for considering the covariance structure of different multinomial distributions as a clustering criteria or outlier identification based on cluster membership probability. Python with IPython [5] is the primary tool for analysis, including packages scikit-learn [4], matplotlib [2], and pandas [3].

References