Analysis of Superfund Site

Authors:

Frances Chua

Alice Huynh

Christian Santizo

Professor: Sung Eun Kim

Department of Mathematics and Statistics
California State University, Long Beach
Long Beach, California, USA

May 16, 2019

Abstract

Thousands of sites are contaminated with hazardous waste being dumped, left out in the open or improperly managed, which imposed risks to human health and the environment in the U.S. The government set up a large amount of funds to clean up locations with high toxic waste which is called Superfund.

The report carries out principal component analysis (PCA) variables between age and labor force predictors. These PCA variables were then applied to discriminant analysis (DA) models to determine the classification strength in distinguishing social classes within a Superfund and non-Superfund tracts.

Unfortunately, the DA models fail to properly distinguish social classes among Superfund and non-Superfund sites.

Contents

1 Introduction

Thousands of sites are contaminated with hazardous waste being dumped, left out in the open or improperly managed, which imposed risks to human health and the environment in the U.S. The government set up a large amount of funds to clean up locations with high toxic waste which is called Superfund. The purpose for performing this statistical analysis is to identify the relationships between two data that are available through the U.S. Environmental Protection Agency (EPA) and American Community Survey (ACS). If there are relationships between the Superfund site and the demographic and social economics, researchers can propose incentive and outreach program to political officials, targeting a specific demographic and social economic group to prevent future sites from being considered as a Superfund site.

2 Data description

The Tract Level Planning Database with 2010 Census and 2009 – 2013 American Community Survey Data (also called the PDB) is a database that assembles a range of geography, demographic, socioeconomic data. The PDB variables have been extracted from census and ACS databases and summarized for all tracts in the country, including Puerto Rico. The data set contains more than 300 variables; however, the analysis of the report focuses on the age range of the total population and the age range of the labor force.

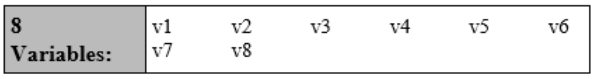
Principal component analysis (PCA) variables are created between age and labor predictors; furthermore, the PCA variables will determine the classification strength to distinguish social classes within a Superfund and non-Superfund tracts. The analysis applies discriminant analysis (DA) to investigate the classification performance.

3 Analysis

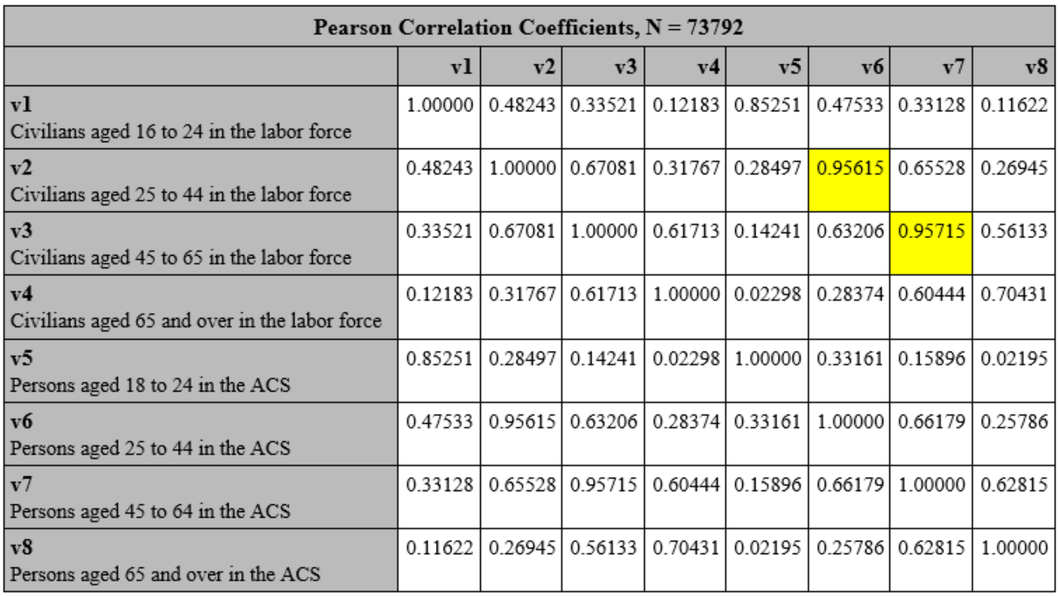
The summary statistics is shown below displays the mean, standard deviation, minimum, and maximum values of each predictor. Note that the ranges the highlighted variables: they contain similar range of values.

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
v1	Civilians aged 16 to 24 in the labor force	73792	307.80	240.87	0.00	5374.00
v2	Civilians aged 25 to 44 in the labor force	73792	925.19	559.49	0.00	10872.00
v3	Civilians aged 45 to 65 in the labor force	73792	820.37	452.79	0.00	7026.00
v4	Civilians aged 65 and over in the labor force	73792	94.04	71.16	0.00	1454.00
v5	Persons aged 18 to 24 in the ACS	73792	426.15	445.08	0.00	17801.00
v6	Persons aged 25 to 44 in the ACS	73792	1132.09	660.50	0.00	12730.00
v7	Persons aged 45 to 64 in the ACS	73792	1125.75	564.91	0.00	10187.00
v8	Persons aged 65 and over in the ACS	73792	574.80	370.47	0.00	20842.00

PCA is beneficial in data reduction conditioned on that the predictors are correlated. To ensure that the predictors are strongly correlated, a correlation matrix of the predictors is investigated. The pairwise predictors that contain the largest Pearson correlation are the labor force and population variables of ages 24 to 44 and 45 to 64. Thus, the analysis shifts towards these four particular predictors.



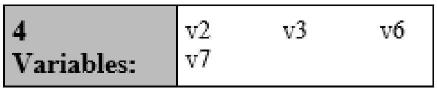
8 Variables:	v1	v2	v3	v4	v5	v6
	v7	v8				



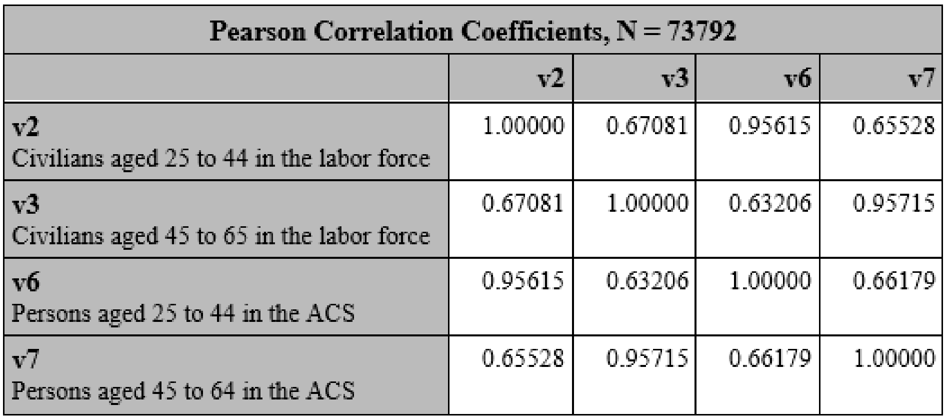
Pearson Correlation Coefficients, N = 73792

	v1	v2	v3	v4	v5	v6	v7	v8
v1 Civilians aged 16 to 24 in the labor force	1.00000	0.48243	0.33521	0.12183	0.85251	0.47533	0.33128	0.11622
v2 Civilians aged 25 to 44 in the labor force	0.48243	1.00000	0.67081	0.31767	0.28497	0.95615	0.65528	0.26945
v3 Civilians aged 45 to 65 in the labor force	0.33521	0.67081	1.00000	0.61713	0.14241	0.63206	0.95715	0.56133
v4 Civilians aged 65 and over in the labor force	0.12183	0.31767	0.61713	1.00000	0.02298	0.28374	0.60444	0.70431
v5 Persons aged 18 to 24 in the ACS	0.85251	0.28497	0.14241	0.02298	1.00000	0.33161	0.15896	0.02195
v6 Persons aged 25 to 44 in the ACS	0.47533	0.95615	0.63206	0.28374	0.33161	1.00000	0.66179	0.25786
v7 Persons aged 45 to 64 in the ACS	0.33128	0.65528	0.95715	0.60444	0.15896	0.66179	1.00000	0.62815
v8 Persons aged 65 and over in the ACS	0.11622	0.26945	0.56133	0.70431	0.02195	0.25786	0.62815	1.00000

The correlation matrix of these four predictors are show below. Notice that the correlations between the predictors range from 0.63 to 0.96.

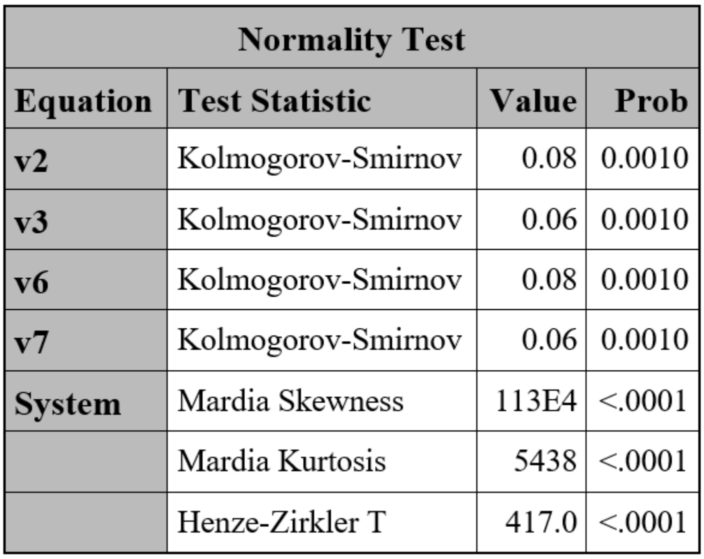


4 Variables:	v2 v3 v6 v7

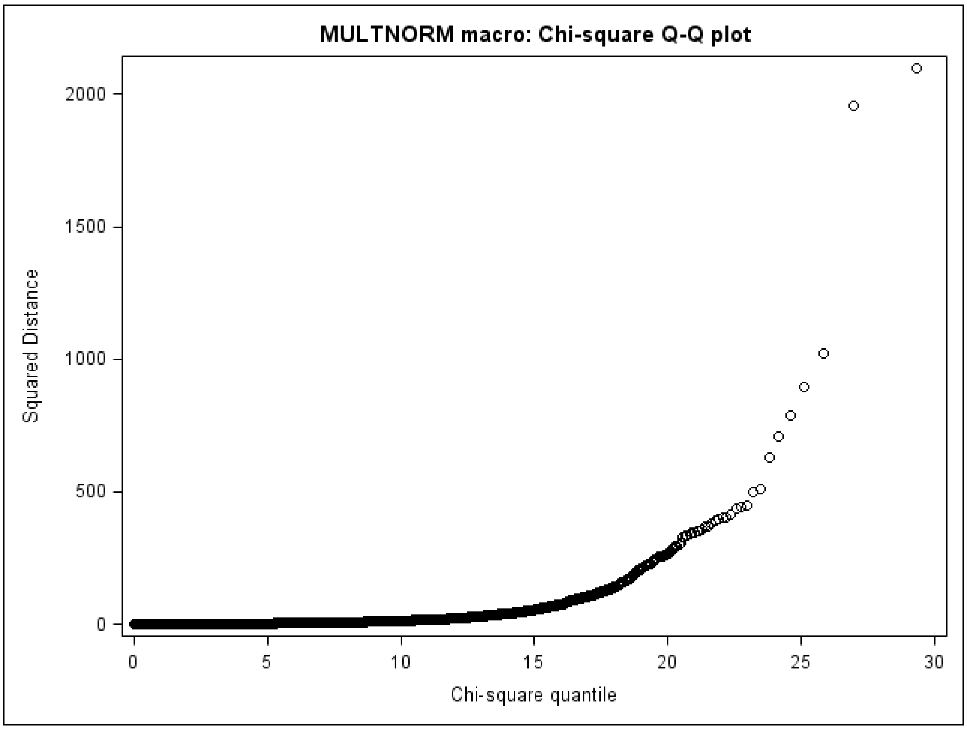


Pearson Correlation Coefficients, N = 73792				
	v2	**v3**	**v6**	**v7**
v2 Civilians aged 25 to 44 in the labor force	1.00000	0.67081	0.95615	0.65528
v3 Civilians aged 45 to 65 in the labor force	0.67081	1.00000	0.63206	0.95715
v6 Persons aged 25 to 44 in the ACS	0.95615	0.63206	1.00000	0.66179
v7 Persons aged 45 to 64 in the ACS	0.65528	0.95715	0.66179	1.00000

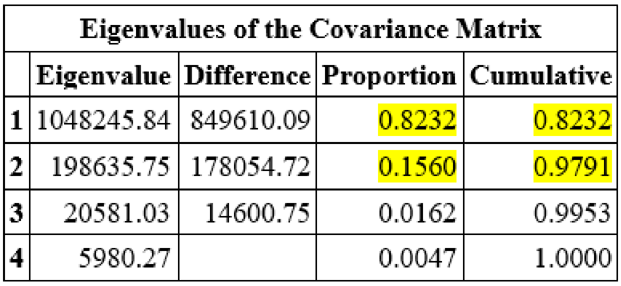
DA requires that each distinct population within the response vector to be multivariate normal. Thus, a χ^2-plot is used to validate multivariate normality. The χ^2-plot does not follow a straight line; furthermore, the univariate tests reject univariate normality, and therefore multivariate normality.



Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
v2	Kolmogorov-Smirnov	0.08	0.0010
v3	Kolmogorov-Smirnov	0.06	0.0010
v6	Kolmogorov-Smirnov	0.08	0.0010
v7	Kolmogorov-Smirnov	0.06	0.0010
System	Mardia Skewness	113E4	<.0001
	Mardia Kurtosis	5438	<.0001
	Henze-Zirkler T	417.0	<.0001

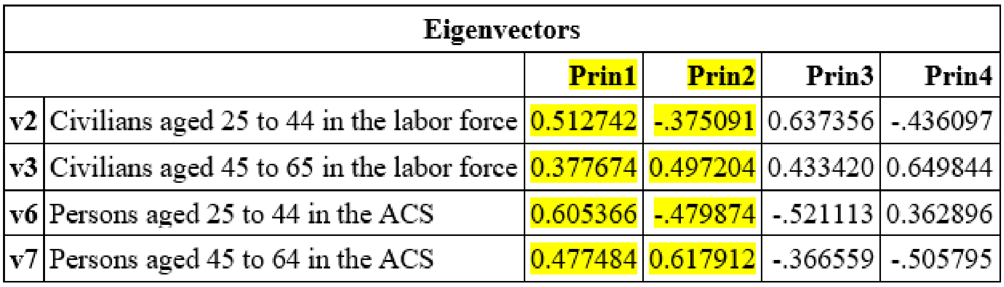


As shown above, the ranges of the predictors are similar. Therefore, PCA on covariance is carried over to retain the importance of each predictor's variability.



Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	1048245.84	849610.09	0.8232	0.8232
2	198635.75	178054.72	0.1560	0.9791
3	20581.03	14600.75	0.0162	0.9953
4	5980.27		0.0047	1.0000

The first two principal components account for 97.91% of the total variation, therefore that is all that is needed for analysis.



Eigenvectors					
		Prin1	Prin2	Prin3	Prin4
v2	Civilians aged 25 to 44 in the labor force	0.512742	-.375091	0.637356	-.436097
v3	Civilians aged 45 to 65 in the labor force	0.377674	0.497204	0.433420	0.649844
v6	Persons aged 25 to 44 in the ACS	0.605366	-.479874	-.521113	0.362896
v7	Persons aged 45 to 64 in the ACS	0.477484	0.617912	-.366559	-.505795

All classes are captured in the first PC, with Upper Class having a slightly higher proportion. Since the underlying factor of the 1st PC is thought to be an age index, it is possible that the classification plot explains that some younger people can be in Upper Class. Other classes are captured, however, showing that the income bracket for younger people is mixed.

For the second PC, most positive values are from Upper Class, with Lower and Middle Classes make up the negative values. Since the underlying factor for the 2nd PC was maturity, this plot shows that maybe older people make up the majority of Upper Class.

The classification plots using linear discrimination for nonsuperfunded and superfunded are similar. Only the Middle Class exhibits a difference between the two plots, having a smaller proportion on the plot for superfunded.

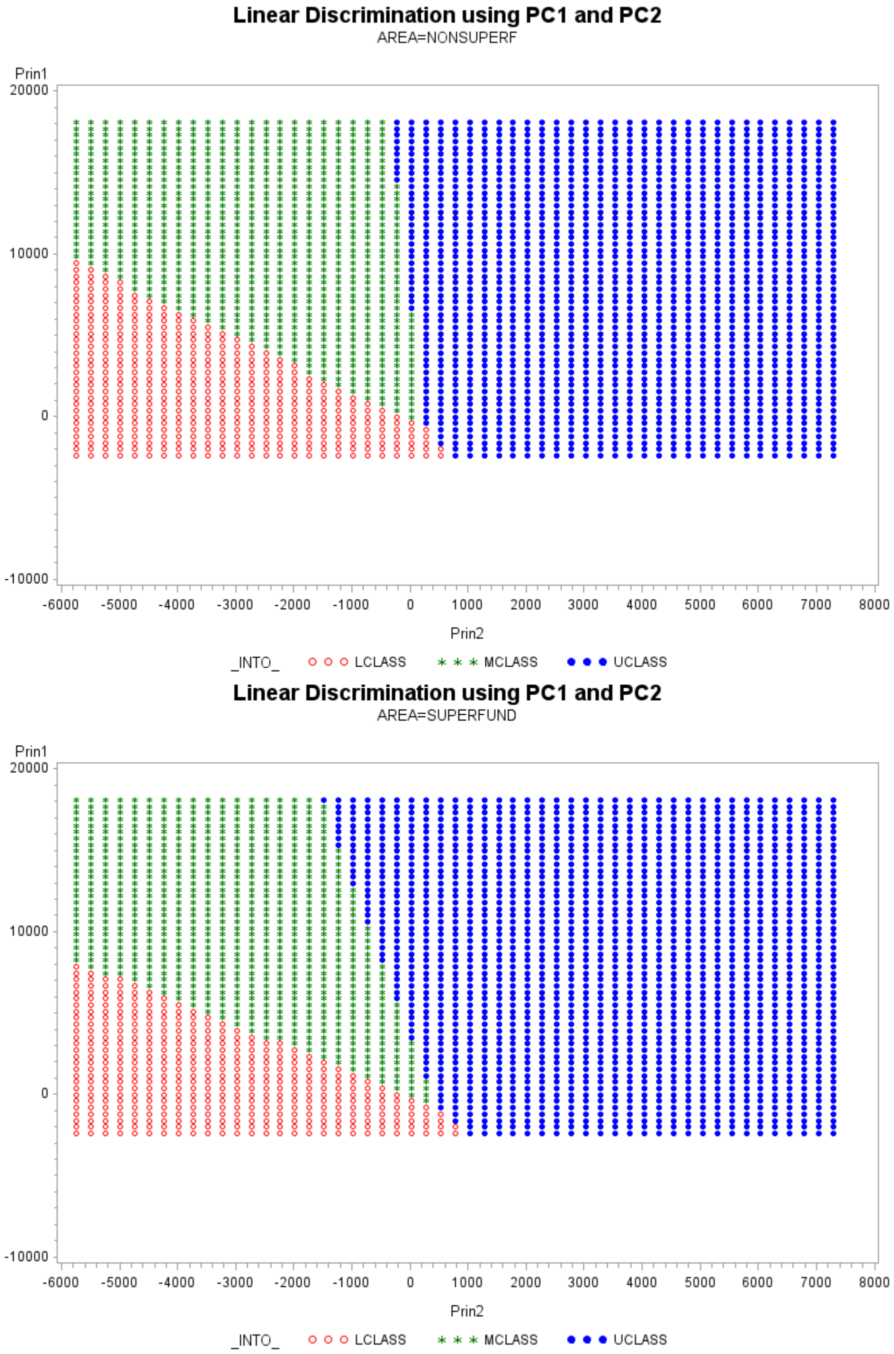


Figure 3.1: Comparison between the LDA models.

It can seen below that there is little to no difference in using quadratic discrimination.

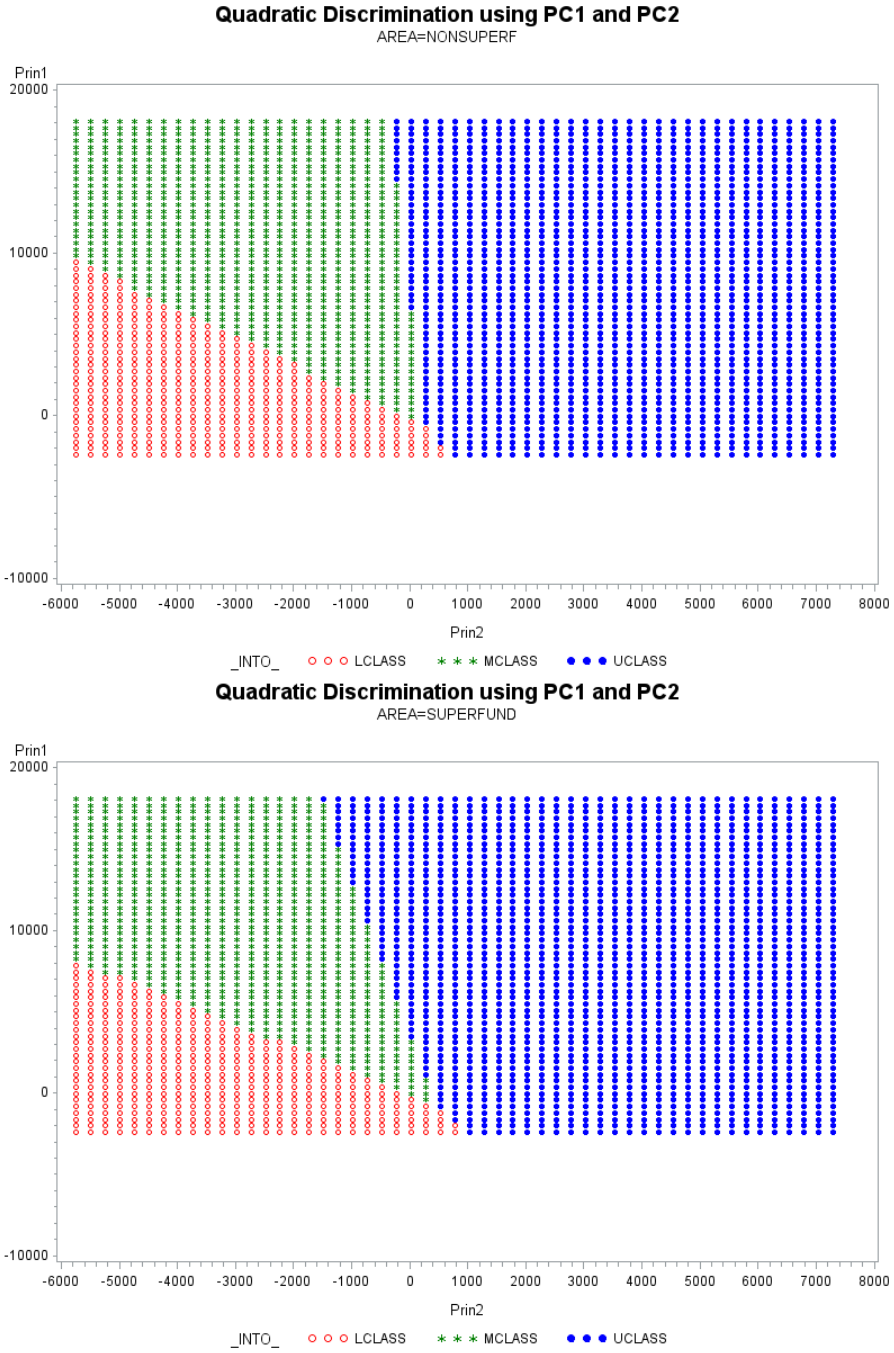
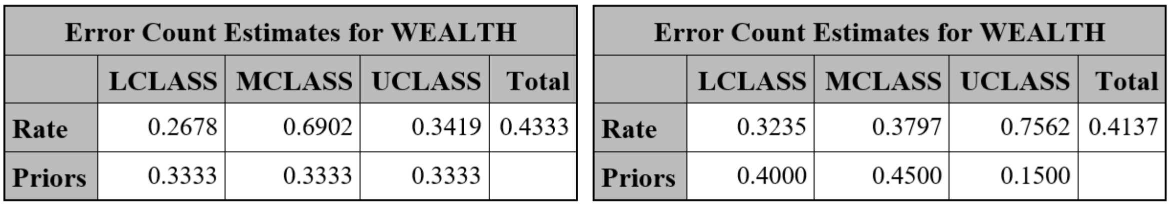


Figure 3.2: Comparison between the QDA models.

In figure 3.3, for non-Superfund, equal priors the error rate is 43 % and the unequal priors we estimated from Federal Reserves System (Board of Governors of the Federal Reserve System, 2019) is error rate is 41 % .



Error Count Estimates for WEALTH				
	LCLASS	**MCLASS**	**UCLASS**	**Total**
Rate	0.2678	0.6902	0.3419	0.4333
Priors	0.3333	0.3333	0.3333	

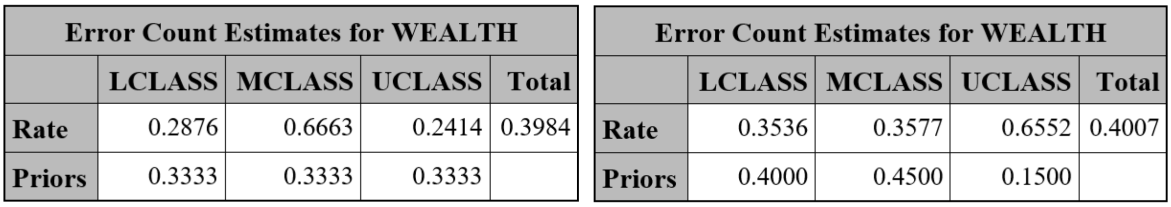
(a) Equal prior

Error Count Estimates for WEALTH				
	LCLASS	**MCLASS**	**UCLASS**	**Total**
Rate	0.3235	0.3797	0.7562	0.4137
Priors	0.4000	0.4500	0.1500	

(b) Unequal prior

Figure 3.3: Linear Discriminate Analysis
Non-Superfund

In figure 3.4, for superfund, equal priors the error rate is 40% and the unequal priors we estimated from Federal Reserves System (Board of Governors of the Federal Reserve System, 2019) is error rate is 40 % .



Error Count Estimates for WEALTH				
	LCLASS	**MCLASS**	**UCLASS**	**Total**
Rate	0.2876	0.6663	0.2414	0.3984
Priors	0.3333	0.3333	0.3333	

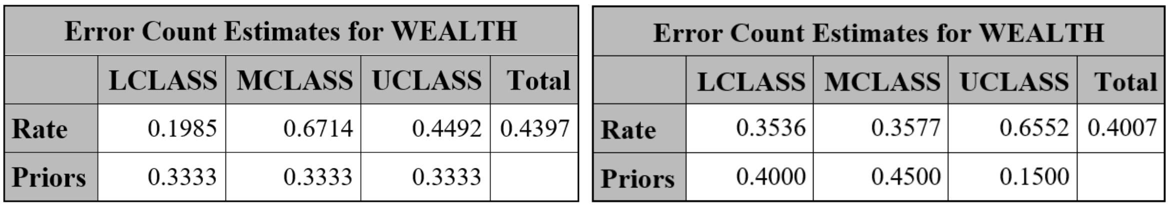
(a) Equal prior

Error Count Estimates for WEALTH				
	LCLASS	**MCLASS**	**UCLASS**	**Total**
Rate	0.3536	0.3577	0.6552	0.4007
Priors	0.4000	0.4500	0.1500	

(b) Unequal prior

Figure 3.4: Linear Discriminate Analysis
Superfund

For figure 3.5, non-Superfund, equal priors the error rate is 43 % and the unequal priors we estimated from Federal Reserves System (Board of Governors of the Federal Reserve System, 2019) is error rate is 40 % .



Error Count Estimates for WEALTH				
	LCLASS	**MCLASS**	**UCLASS**	**Total**
Rate	0.1985	0.6714	0.4492	0.4397
Priors	0.3333	0.3333	0.3333	

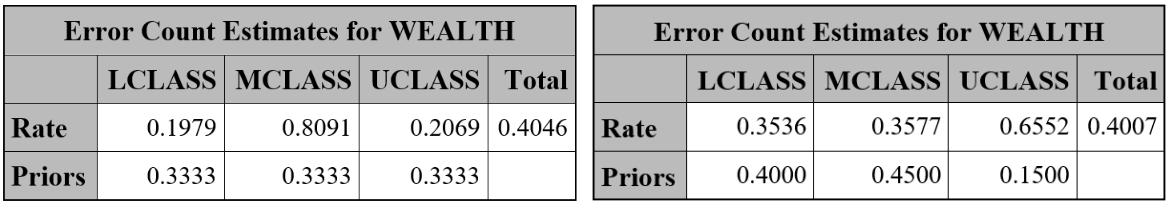
(a) Equal prior

Error Count Estimates for WEALTH				
	LCLASS	**MCLASS**	**UCLASS**	**Total**
Rate	0.3536	0.3577	0.6552	0.4007
Priors	0.4000	0.4500	0.1500	

(b) Unequal prior

Figure 3.5: Quadratic Discriminate Analysis
Non-Superfunded

For figure 3.6, Superfund, equal priors the error rate is 40% and the unequal priors we estimated from Federal Reserves System (Board of Governors of the Federal Reserve System, 2019) is error rate is 40 % .



Error Count Estimates for WEALTH

	LCLASS	MCLASS	UCLASS	Total
Rate	0.1979	0.8091	0.2069	0.4046
Priors	0.3333	0.3333	0.3333	

(a) Equal prior

Error Count Estimates for WEALTH

	LCLASS	MCLASS	UCLASS	Total
Rate	0.3536	0.3577	0.6552	0.4007
Priors	0.4000	0.4500	0.1500	

(b) Unequal prior

Figure 3.6: Quadtric Discriminate Analysis
Superfunded

The result indicate the separation between the different classes are not clear. So the DA model would not work as well as it should.

4 Conclusion

Based on the results above, the DA models fail to properly distinguish social classes among Superfund and non-Superfund sites. The failure in classification can be explained by various reasons. Mainly, the variation in the PCA variables between social classes proved to be minimal. In other words, it is almost indistinguishable to determine a person's social class solely on their age and whether they are in the labor force. It is also worth noting that DA requires the distinct social classes to contain multivariate normal distribution; yet, based on the χ^2-plot this is not true. This further contributes to the poor performance of the DA models.

5 Reference

Board of Governors of the Federal Reserve System. (2019). The Fed - Changes in U.S. Family Finances from 2013 to 2016: Evidence from the Survey of Consumer Finances. [online] Available at: https://www.federalreserve.gov/publications/2017-September-changes-in-us-family-finances-from-2013-to-2016.htm [Accessed 16 May 2019].