Project 1

Introduction

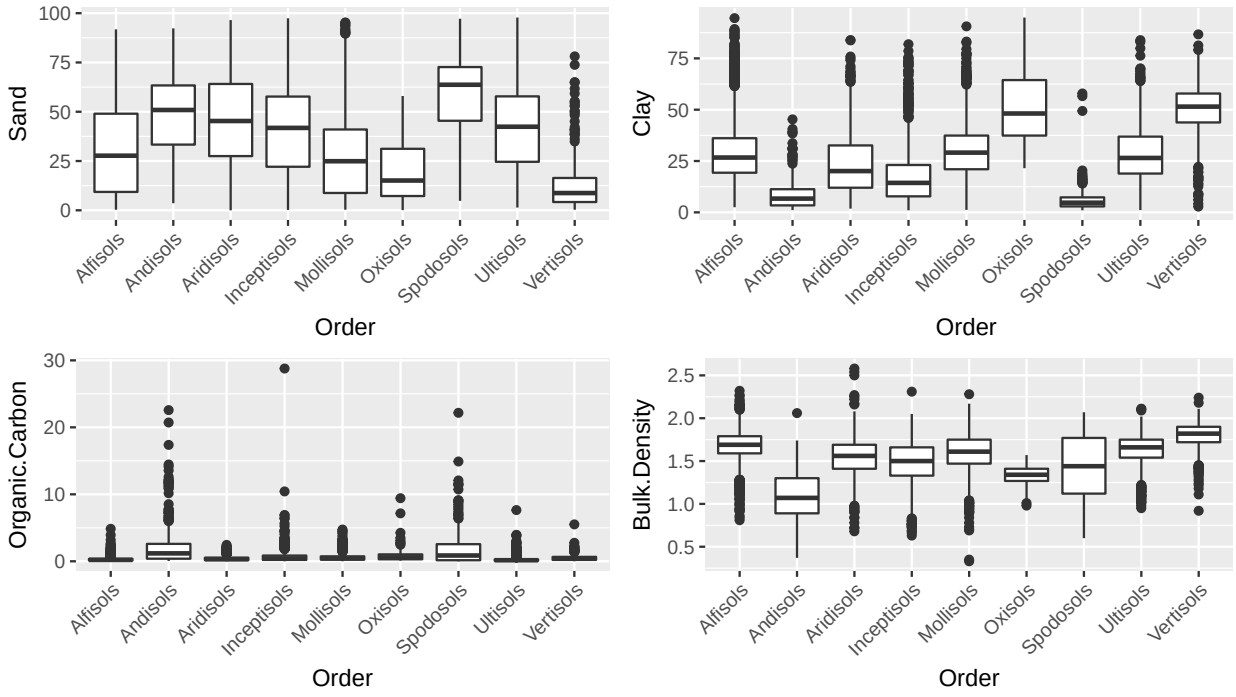
This is Project 1 for STAT 557 2018 Spring by Meridith Bartley and Fei Jiang. The aim of this project is to practice discriminant analysis and logistic regression and study basic techniques of dimension reduction. In this project we applied Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and multinomial logistic regression to soil sample data in order to classify into separate soil group (Orders).

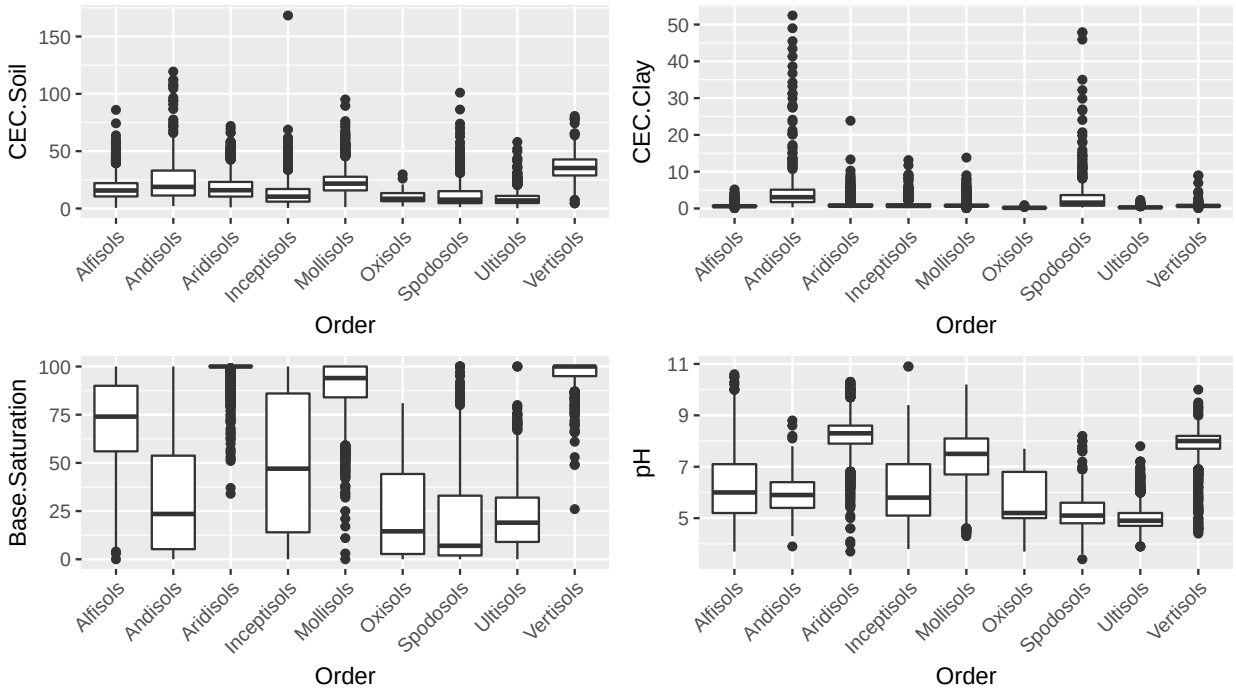
Description of Data

This dataset contains soil sample data over the US downloaded from Natural Resources Conservation Service (NRCS). After removing the incomplete data records and excluding the data records with impossible values, there are around 14,000 records left, each of which includes physical and chemical properties of soil samples (sand, silt, clay, organic carbon, bulk density, CEC soil, CEC clay, base saturation, and pH) and the corresponding soil classification group (soil order).

Boxplots for each physical and chemical property used as explanitory variables in the subsequent classification models are included below. This EDA allows for early indication of which variables may possibly be ommitted during dimention reduction. That is, what properties do not differ significantly between soil Orders.

Exploritory Data Analysis





Principle Component Analysis

In order to test whether dimension reduction will improve predictions we also conducted Principle Component Analysis on the original dataset to get a new dataset with fewer dimensions. According to our PCA results, the first four component in total can explain about 99.8% of variance of the original database. The coefficients of the relevent componets are listed in the table below. Therefore, we took the first four components and the soil order value to build a new dataset with less dimensions.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	32.63819	25.63283	11.00003	8.588096	1.538774	0.7446948	0.6472851	0.1861136
Proportion of Variance	0.55470	0.34213	0.06301	0.038410	0.001230	0.0002900	0.0002200	0.0000200
Cumulative Proportion	0.55470	0.89683	0.95984	0.998240	0.999480	0.9997600	0.9999800	1.0000000

	PC1	PC2	PC3	PC4
Sand	0.3516041	0.7770568	-0.5154614	-0.0820792
Clay	-0.2421961	-0.3976529	-0.6759072	-0.5665441
Organic.Carbon	0.0053164	-0.0050832	-0.0168647	0.0577142
Bulk.Density	-0.0019288	0.0000607	-0.0029836	-0.0104239
CEC.Soil	-0.1781833	-0.1804817	-0.5190661	0.8044793
CEC.Clay	0.0112285	0.0059963	-0.0252571	0.1423157
Base.Saturation	-0.8859172	0.4526640	0.0839830	-0.0374301
pH	-0.0309674	0.0226877	0.0059155	0.0030852

Analysis

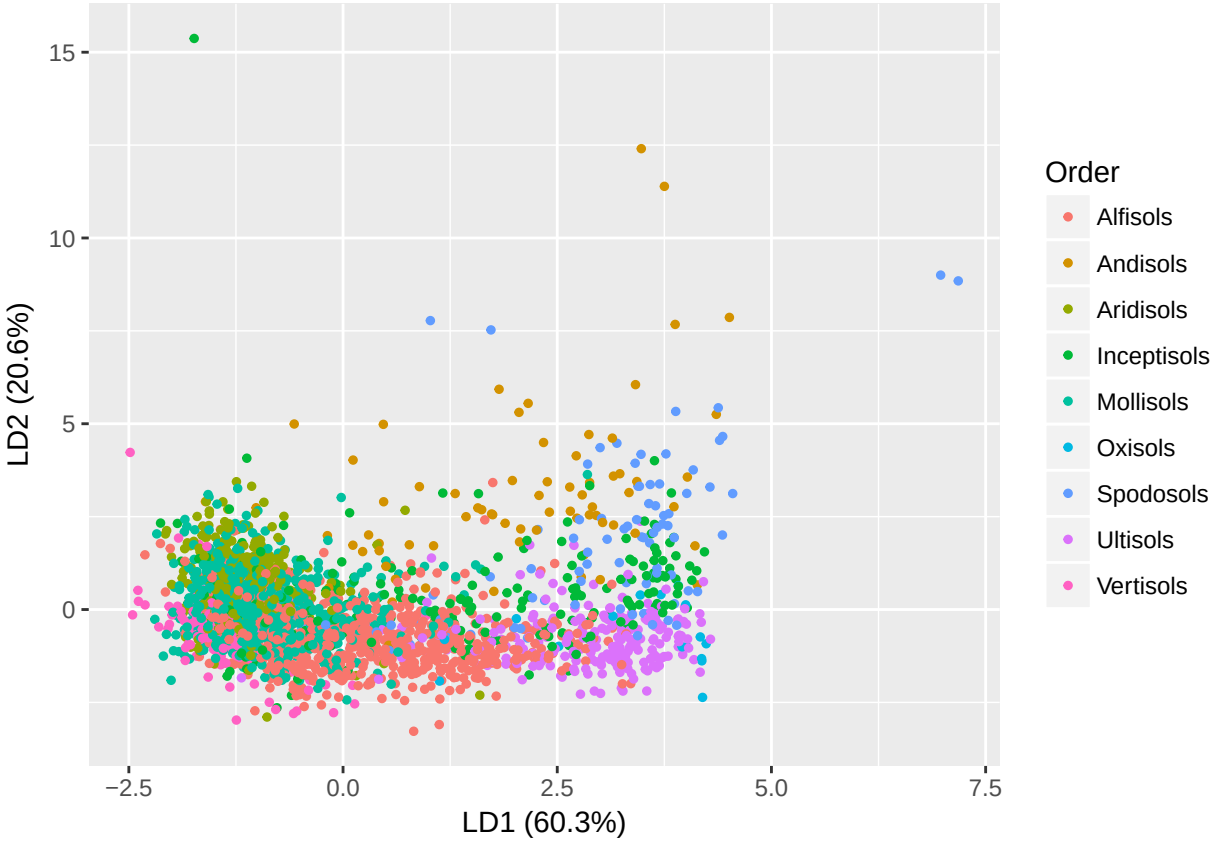
In the following analysis with three methods (LDA, QDA and logistic) and two datasets (original and dimension-reduced), we randomly selected 80% of the entire data as training data and the rest 20% as test data.

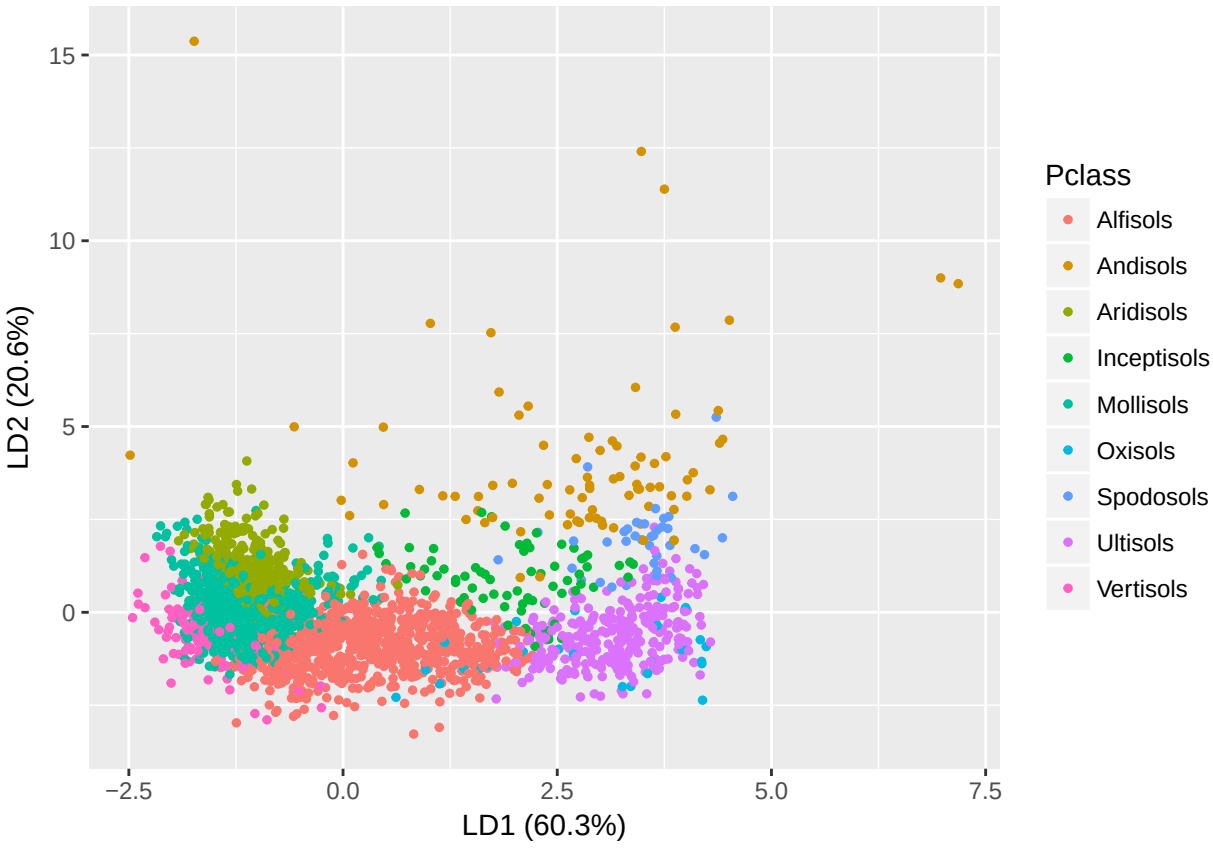
Linear Discriminant Analysis (LDA)

Original Dataset

We initially conducted the LDA on the original dataset and found that the overall prediction accuracy of our model in testing data is about 57%. Considering we have in total 9 possible classes, the accuracy rate is fairly good.

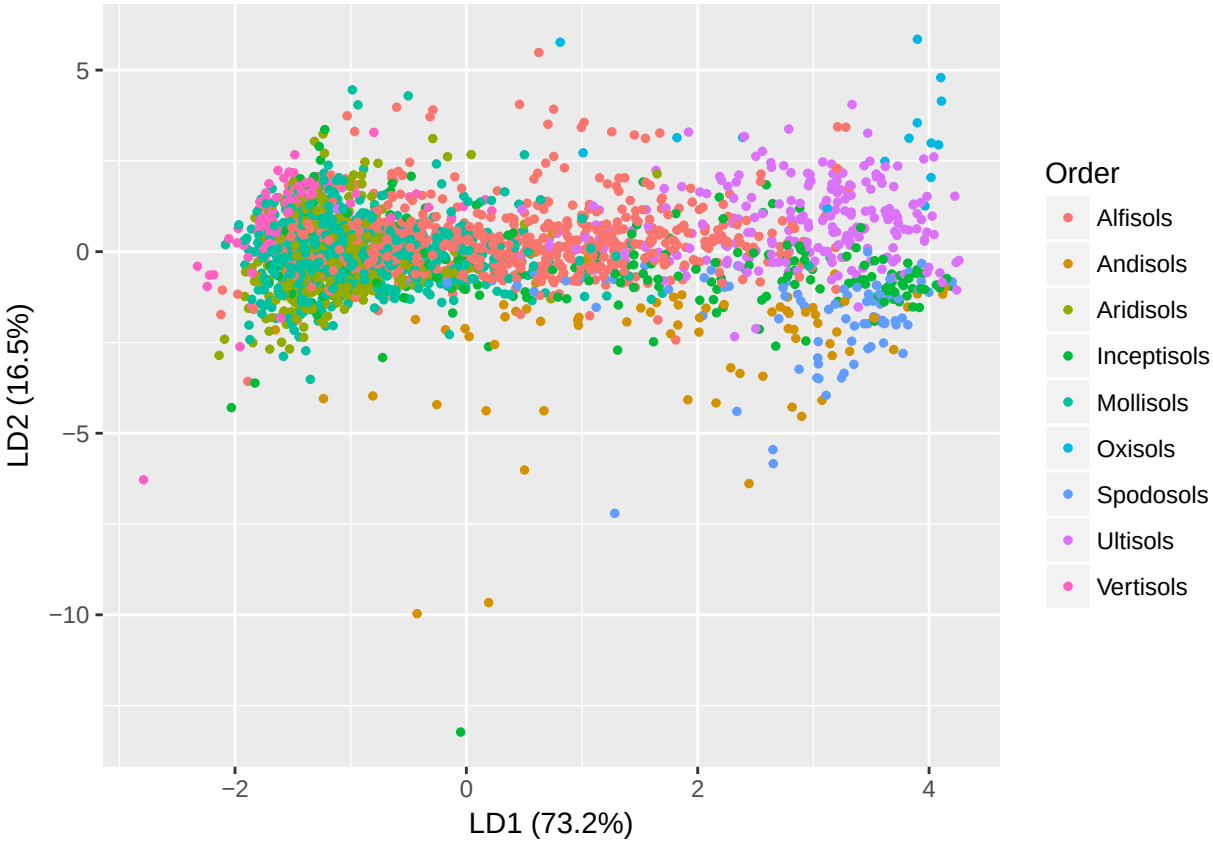
In the follow plots, we show the difference between the true and predicted classes (top and bottom plots, respectively). It can be seen that in the left part of the true class, there is a lot of overlap and in the middle part, there is some overlap. But in the prediction plot, different classes separate pretty well from each other, which indicates that our model seperate the classes more than it should be. This overlap between classes in reality also suggests us that we should consider more variables to separate them well.

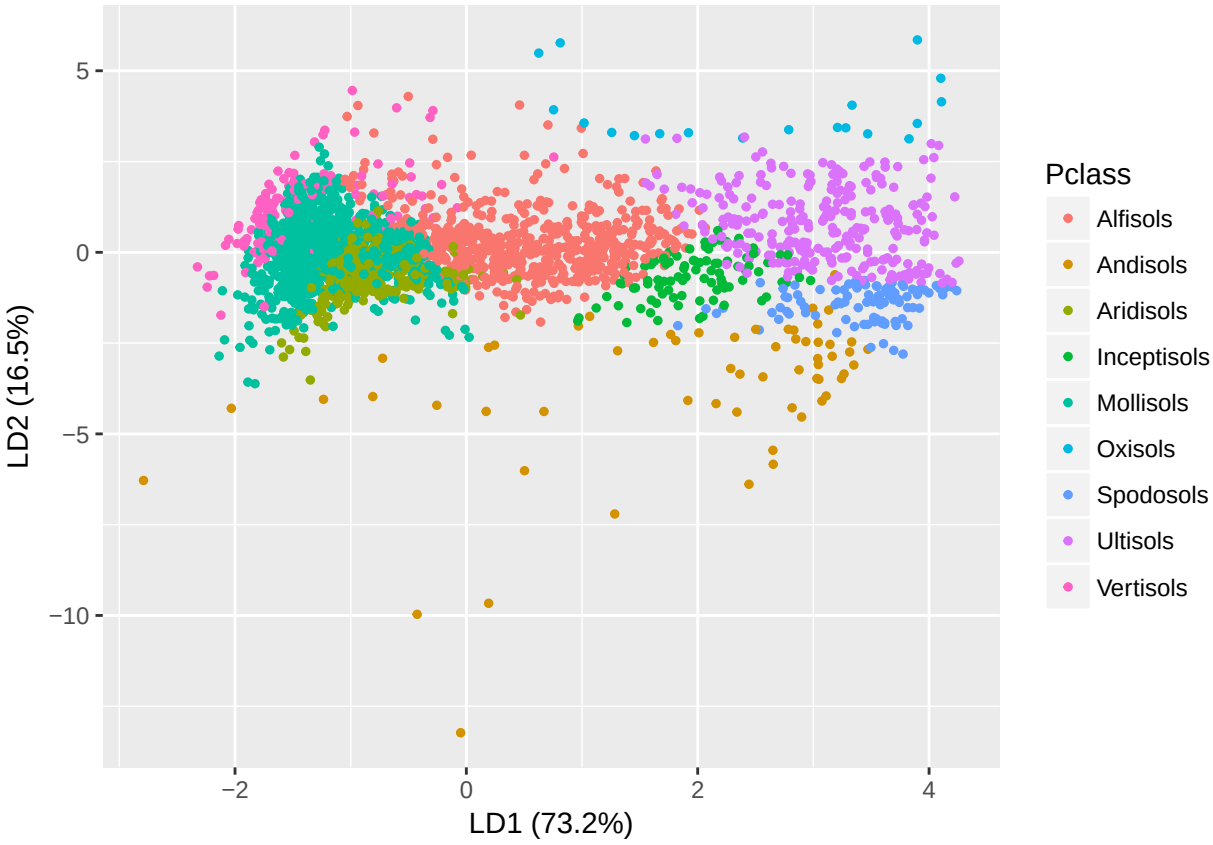




Reduced-Dimension Dataset

We also conducted the same LDA method on the dimension-reduced database. However, the result is less satisfying than using the original database. The prediction accuracy here is about 53%, less than 57% of the original database. The plot below shows the difference between the true and predicted classes. Again, the overlap in the true class plots indicate the difficulty to classify those samples.


5
0
−5
−10
LD2 (16.5%)
−2
0
2
4
LD1 (73.2%)
Order
Alfisols
Andisols
Aridisols
Inceptisols
Mollisols
Oxisols
Spodosols
Ultisols
Vertisols


5
0
−5
−10
LD2 (16.5%)
−2
0
2
4
LD1 (73.2%)
Pclass
Alfisols
Andisols
Aridisols
Inceptisols
Mollisols
Oxisols
Spodosols
Ultisols
Vertisols

In the table below we can see the prediciton accuracy for each soil Order. We can see that for most, but not all, individual soil orders the LDA method applied to a full-dimension dataset provides the highest percent prediciton accuracy.

	Alfisols	Andisols	Aridisols	Inceptisols	Mollisols	Oxisols	Spodosols	Ultisols	Vertisols
LDA	61	57	54	10	64	76	27	76	55
LDA w/ PCA	48	41	43	11	71	41	47	78	42

Quadratic Discriminant Analysis (QDA)

We used the same original and dimension-reduced dataset to apply QDA method. The prediction accuracy of QDA is very similar to LDA. For the original dataset, the overall accuracy is 56%. When applying the same method to the dimension-reduced dataset, the oveall accuracy decreases to 53%.

	Alfisols	Andisols	Aridisols	Inceptisols	Mollisols	Oxisols	Spodosols	Ultisols	Vertisols
QDA	71	39	51	22	49	65	42	78	64
QDA w/ PCA	50	35	70	8	53	35	50	80	62

Multinomial Logistic Regression

Recall that the response variable for these data is an independent 9-level categorical response. With this response variable in mind, we used the same original and dimension-reduced dataset to apply Multinomial Logistic Regression method. The prediction accuracy again is similar to LDA and QDA, albeit with a slight improvement. For the original dataset, the overall accuracy is 62%. When applying the same method to the dimension-reduced dataset, the oveall accuracy decreases to 55%.

	Alfisols	Andisols	Aridisols	Inceptisols	Mollisols	Oxisols	Spodosols	Ultisols	Vertisols
MLR	69	55	57	18	67	41	57	78	58
MLR w/ PCA	58	41	51	10	65	18	52	74	36

Results

In order to compare the reults it is important to recall the diffences between these three classification approaches. The difference between LDA and logistic regression is that linear coefficients are estimated differently. MLE for logistic models and estimated mean and variance based on Gaussian assumptions for the LDA. LDA makes more restrictive Gaussian assumptions and therefore often expected to work better than logistic models IF they are met. QDA serves as a compromise between non-parametric methods (not explored in this project) and the linear LDA and logistic regression approaches. Since QDA assumes a quadratic decision boundary, it can accurately model a wider range of problems than can the linear methods. QDA can perform better in the presence of a limited number of training observations because it does make some assumptions about the form of the decision boundary.

The results from these three approaches show that the Multinomial Logistic Regression out performed both LDA and QDA. This is likely due to not meeting the LDA's normality assumption in addition to having a very large dataset for testing/training.

	Alfisols	Andisols	Aridisols	Inceptisols	Mollisols	Oxisols	Spodosols	Ultisols	Vertisols	Overall
LDA	0.61	0.57	0.54	0.10	0.64	0.76	0.27	0.76	0.55	0.57

	Alfisols	Andisols	Aridisols	Inceptisols	Mollisols	Oxisols	Spodosols	Ultisols	Vertisols	Overall
QDA	0.71	0.39	0.51	0.22	0.49	0.65	0.42	0.78	0.64	0.58
MLR	0.69	0.55	0.57	0.18	0.67	0.41	0.57	0.78	0.58	0.62

Contributions

The different tasks required to complete this project were equally divided between Meridith and Fei. LDA and QDA analyses were completed by Fei while Meridith was responsible for MLR and model comparisons. Both members of this group contributed to the presentation slides and this report.