STAT 450/550 Multivariate Statistical Analysis Term Project
Dr. Sung Kim, Spring 2015

Multivariate Statistical Methods on Ozone Data in Upland City in San Bernardino County

By Elka-Anne Lacno, Hien D. Nguyen, & Denis B. Nyongesa

Abstract

The San Bernardino County ranked as the most ozone-polluted county in the U.S. according to the American Lung Association. The purpose of this study is to use multivariate statistical methods to determine the air quality data for the State of California, specifically for Upland City (in San Bernardino County) from January 2001 to December 2004. The data set from 2001-2003 consists of 1095 observations and 8 variables and the data set from 2004 consists of 353 observations and 8 variables. Preliminary Analysis such as simple statistics, normality, and correlation among variables will be checked just as a standard procedure. Then Principal Component Analysis (PCA) will be used to for data reduction and interpretation on the quantitative variables. As a result, PCA reduces the quantitative variables in our data to 3 useful linear combinations. From there, Logistic Regression Analysis on the 3 principal components as well as adding the categorical variables derives classification values from the observations in our data. We use the data from 2001-2003 as the training set to fit our model. From there, the data from 2004 will be the validation set to test the model.

1 Introduction

Formation of ozone is dependent on meteorological conditions. The meteorological data being used for analysis are from the University of California Statewide Integrated Pest Management Program (UC IPM). The meteorological conditions found in this data are shown below in table 1. One data set is recorded from the time period 2001-2003 and another data set is recorded on 2004. Also included in the data is the ozone concentration associated with the meteorological conditions. The ozone concentration will help to determine the Air Quality Index (AQI), which rates air quality from good to hazardous depending on the level of health concern shows in table 2 below. Various forms of Discriminant Analysis are used to classify the AQI based on the meteorological conditions. Prior to Discriminant analysis, principal components will help reduce the data for interpretation. Note that these data have also the variables day, month, and weekend. However, these will not be included in the principal component analysis since they are categorical variables and are more meaningful with classification. Thus once principal components have been finalized, month and weekend will be added back in for logistic regression analysis.

Table1: Description of the variables .

Variable	Description of Variable
temp	Daily maximum air temperature (F)
soil	Daily maximum soil temperature (F)
solar	Daily global radiation (Watts/m2)
eto	Daily reference evapotranspiration (inch)
r.hum	Daily maximum relative humidity (%)
prec	Daily total precipitation (inch)
wind.u	Daily west-east wind component (m/s)
wind.v	Daily south-north wind component (m/s)

Table2: Summary table of AQI state by US EPA.

Ozone Range (ppb)	Level of Health Concern	State
0-59	Good	1
60-75	Moderate	2
76-95	Unhealthy for Sensitive Group	3
96-115	Unhealthy	4
116-374	Very Unhealthy	5
375+	Hazardous	6

2 Preliminary Analysis

2.1 Simple Statistics

The variances of these variables from table 3 (shown below) have a very wide range with eto (σ^2 = 0.0056) having the smallest variance and solar (σ^2 = 30,031.57) having the largest variance. Thus, there is an indication that these variables are measured on different scales. This is a much anticipated problem which will be addressed in the start of the principal component analysis.

2.2 Normality

Any standard statistical procedure requires normality check for all variables. By using the proc univariate procedure and the construction of Q-Q plots for each variable, it can be shown that all variables in this data set violates normality. However, principal component analysis does not have the assumption of variable normality, so it is safe to go through with the analysis.

2.2 Correlation Among Variables

The numbers on the top left side of table 4 are bolded to show the variables with the highest correlations. Here the variables temp, soil, solar, and eto are highly correlated with each other with a value of more than |0.5|. Since they are all highly correlated, it is possible to remove a few variables to avoid multicollinearity. However, we will continue with the analysis using all variables because the principle component analysis will use correlation and thus apportion the weights depending on how much variation a variable contributes to the overall model.

Table3: Simple statistics table.

	Mean	St.Dev	Variance
temp	77.17717	12.72561	161.9411
soil	66.44384	9.683015	93.76078
solar	436.4703	173.2962	30031.57
eto	0.15284	0.074765	0.00559
r.hum	78.71863	15.45145	238.7473
prec	0.10411	0.305542	0.093356
wind.u	-0.73261	3.027062	9.163104
xwind.v	-1.39844	2.200575	4.84253

Table4: Correlation matrix

	temp	soil	solar	eto	r.hum	prec	wind.u	wind.v
temp	1	**0.808**	**0.688**	**0.805**	-0.242	-0.358	-0.18	-0.169
soil	**0.808**	1	**0.72**	**0.746**	0.158	-0.231	-0.484	-0.245
solar	**0.688**	**0.719**	1	**0.947**	0.002	-0.379	-0.436	-0.157
eto	**0.805**	**0.746**	**0.947**	1	-0.182	-0.38	-0.373	-0.044
r.hum	-0.242	0.158	0.002	-0.182	1	0.177	-0.318	-0.212
prec	-0.358	-0.231	-0.379	-0.38	0.177	1	0.111	-0.02
wind.u	-0.18	-0.484	-0.436	-0.373	-0.318	0.111	1	-0.113
wind.v	-0.169	-0.245	-0.157	-0.044	-0.212	-0.02	-0.113	1

3 Principal Component Analysis (PCA)

Principal Components Analysis will be used in order to interpret the air quality data given. Variables will be reduced to a few linear combinations (eigenvectors) that would explain majority of the variation. The variables, maximum temperature, maximum soil temperature, solar, evapotranspiration, maximum relative humidity, precipitation, west-east wind, and north-south wind will be included in this analysis. Based on preliminary analysis, principal components will be performed on the correlation matrix instead of the covariance matrix since more weight will be emphasized toward solar if covariance matrix was to be used. The correlation matrix will standardize the data so that all variables will be weighted equally.

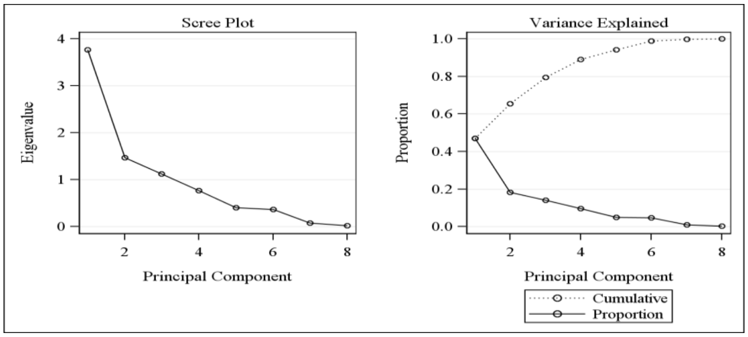
3.1 Number of Principal Components (PCs)

For a principal component to be considered important in explaining any data, it must be able to explain at least the average proportion per principal component (100/8 = 12.5%) and its associate eigenvalue should be at least 1. By the tables shown below, the first 3 principal components satisfy these rules. From graph 3, the scree plots clearly shows the first 3 PCs are above the eigenvalue of 1. All three explain 79.49% of the variation, which is satisfactory. Thus, the first three principal components can achieve an efficient description of the data without losing much of the information.

Table5: Eigenvalues of the principal components.

PC	Eigenvalue	Proportion	Cumulative
1	3.76383	0.4705	0.4705
2	1.469968	0.1837	0.6542
3	1.125384	0.1407	0.7949
4	0.76512	0.0956	0.8905
5	0.402327	0.0503	0.9408
6	0.371488	0.0464	0.9873
7	0.082245	0.0103	0.9975
8	0.019638	0.0025	1

Graph1: Scree plot and variance explained.



3.2 Component Labeling

First component can be labeled as the heat component since a high PC$_1$ would mean a warmer weather. The first four variable place high positive weights on PC$_1$ (ranging from 0.448 to 0.486). Notice that the last four variables have negative loadings. Thus, this indicates that high air and soil temperature as well as high solar and evapotranspiration result in a hotter weather. A lower negative PC$_1$ score would primarily be associated with negative values of prec and wind.u. Thus higher precipitation and strong west-east winds would mean a cooler weather. **Second component** can be named as the humidity component since a high PC$_2$ would mean a high relative humidity. Variable r.hum places the highest positive weight on PC$_2$ while wind.u and wind.v contrasts with a negative weights. Thus if PC$_2$ were to have a negative score, there would be a strong winds and low relative humidity, that is, when it's windy a person would feel the humidity less. **Third component** can be called the wind direction/speed component where a high PC$_3$ would indicate a high wind speed in the north/south direction and decreasing wind speed in the east/west direction. Variable wind.v places a high positive weight on PC$_3$ while wind.u contrasts with a high negative weight. Thus, a negative PC$_3$ score would mean stronger west-east winds and low north-south winds.

Table6: Principal Component Loadings.

	PC$_1$	**PC$_2$**	**PC$_3$**	**PC$_4$**	**PC$_5$**	**PC$_6$**	**PC$_7$**	**PC$_8$**
temp	**0.448089**	-0.18523	-0.18642	0.190308	0.312904	-.373110	-.656158	0.162098
soil	**0.456286**	0.202007	-0.08163	0.152117	0.276248	-.405348	0.692418	0.033635
solar	**0.474469**	0.02567	0.022602	0.013962	-.139687	0.603434	0.058231	0.621674
eto	**0.486147**	-0.12381	0.049861	0.149040	-.039957	0.389044	-.030745	-.754809
r.hum	-0.02436	**0.74292**	0.029256	-.261469	0.523549	0.226459	-.211747	-.089616
prec	-0.23918	0.313451	-0.09695	0.902211	-.065244	0.115202	-.051514	0.031207
wind.u	-0.25281	**-0.42257**	**-0.53497**	0.030803	0.566817	0.335943	0.189957	0.012638
wind.v	-0.08858	-0.28311	**0.811875**	0.186909	0.452799	0.059844	0.045867	0.084935

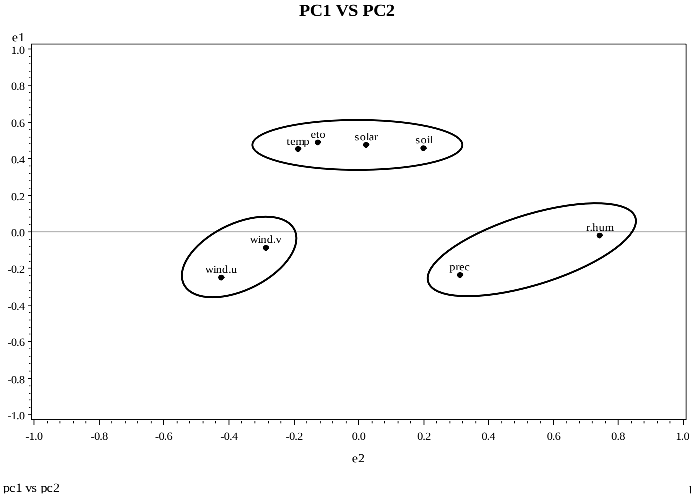
3.3 Variable Grouping

In order to show the various grouping of the variables, each of the three principal components are graphed against each other to show a clearer view of the important weights in each principal component.

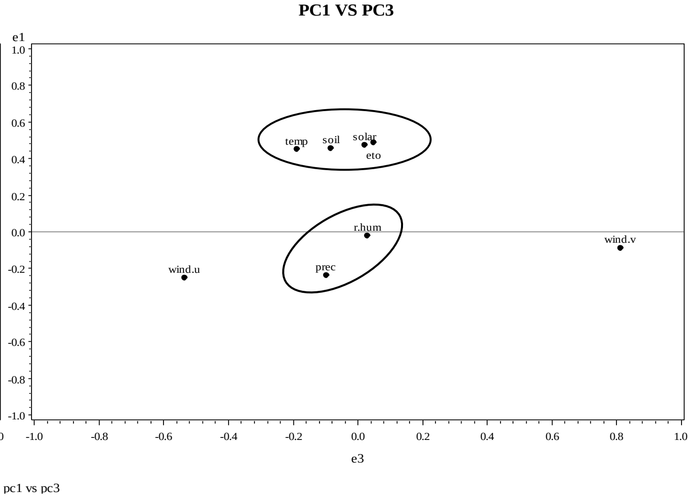
PC$_1$ grouping: In graphs 2 & 3, you can see that temp, eto, solar, and soil are grouped together positively along the vertical axis when PC$_1$ is plotted against PC$_2$ and PC$_3$. There are no specific outlying variable as it can be seen that the other four variables are roughly grouped together semi-negatively along the vertical axis. **PC$_2$ grouping**: In graphs 2 & 4, variable maximum relative humidity is the outlier lying furthest away from all the other variables, while wind.u and wind.v lie on the opposite end of the axis. Thus as relative humidity increases, wind speed decreases.**PC$_3$ grouping**: In graphs 3 & 4, both wind.u and wind.v are outlying variables opposite of each other, which all other variables are grouped together horizontally. Variable wind.v is outlying in the positive direction of the horizontal axis, which wind.u is outlying in the opposite direction. Hence, the direction of wind speed matters and cannot be grouped together in this PC$_3$.

Interpretation of variable grouping is consistent with the labeling of PC$_1$ as the heat component, PC$_2$ as the humidity component, and PC$_3$ as the wind direction/speed component.

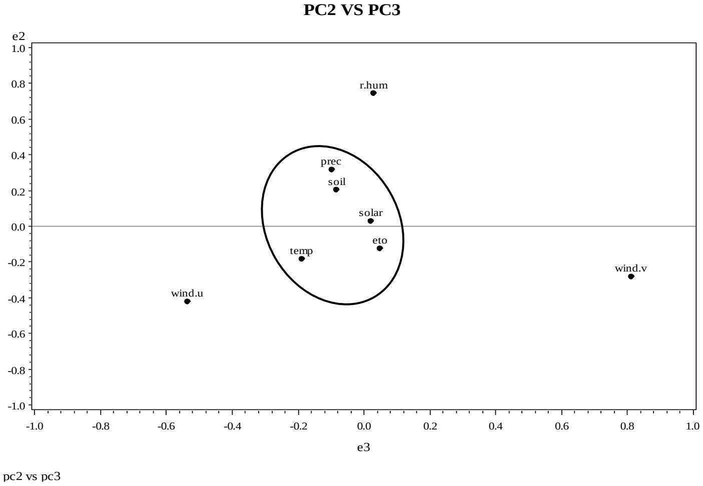
Graph2: PC$_1$ vs. PC$_2$



Graph3: PC$_1$ vs. PC$_3$



Graph4: PC$_2$ vs. PC$_3$

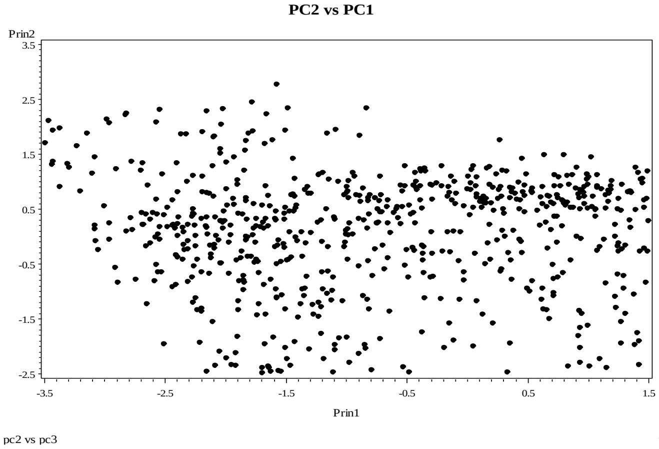


3.4 Sample Grouping

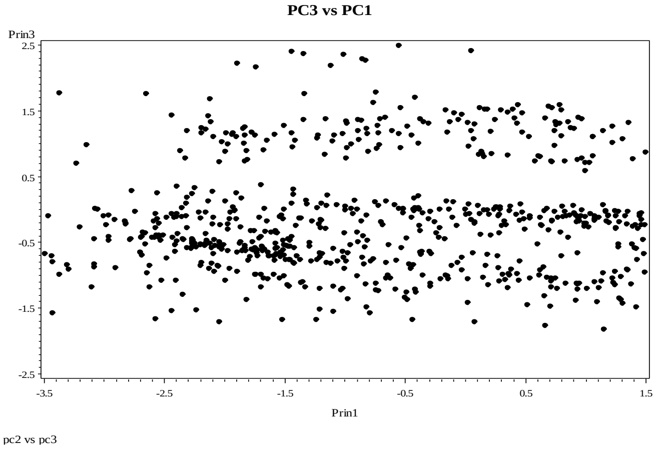
PC$_2$ vs. PC$_1$: looking at graph 5, it seems the observations spread throughout the graph. However, there is more density towards the middle of the vertical axis and across the horizontal axis. There doesn't appear to have a concrete sign on any clusterings. The cluster shows PC2 score ranging from -0.5 to 1 and a PC1 score from -3 to 2.5. Thus, from these observations, it is safe to say that a balance of low humidity and low winds is constant from low to high heat weather.

PC3 vs. PC1: graph 6, shows 2 possible clusterings on the vertical axis. The bottom cluster has a wider spread throughout the horizontal axis. However, there more dense observations towards the lower left side and upper right side of the bottom cluster. The lower left side has a PC3 score of about -0.5 and a PC1 score ranging from -2.5 to -0.5. Thus, stronger west-east winds are associated with colder weather. The upper right side has a PC3 score of around 0 and a PC1 score ranging from 0.5 and 1.5 which can indicate a warmer weather can be related to a balance of winds coming from both directions. **PC3 vs. PC2:** graph 7 shows the clearest sign of three clusterings. The top cluster has a PC3 score around 1 to 1.5 and a PC2 score around 0.5, which indicates that observations with higher north-south winds have low relative humidity and low west-east winds. The lower cluster fans out to the left, which does not come as a surprise since variable wind.u places a negative heavy weight on both PC2 and PC3. Thus a negative PC2 and PC3 indicate that strong west-east winds were observed.

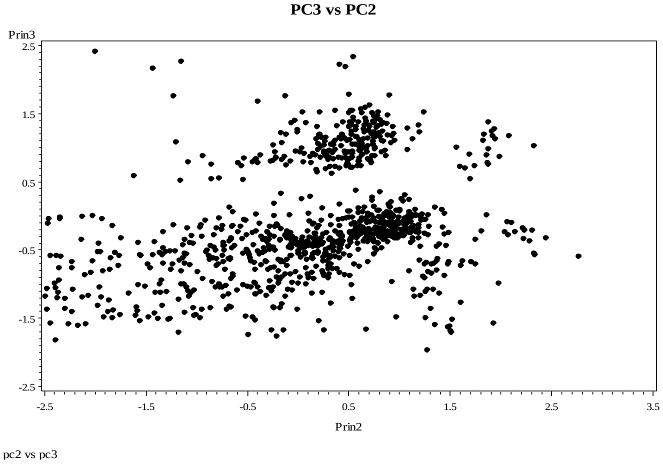
Graph5: Observation plot PC2 vs. PC1



Graph6: Observation plot PC3 vs. PC1



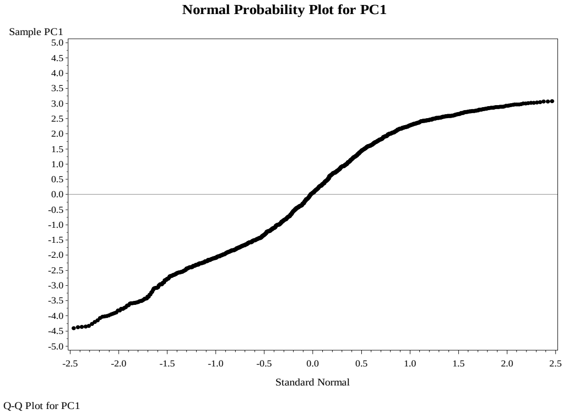
Graph7: Observation plot PC3 vs. PC2



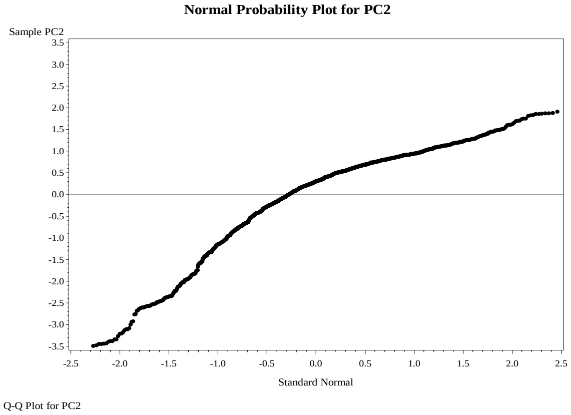
3.5 Normality of Principal Components

Normality should be checked on these principal components in order to proceed onto Discriminant analysis where normality of the data is an assumption. In order to look at the normality of the principal components, Q-Q plots on each principal components were constructed. Graphs 8-10 all show that PC1, PC2, and PC3 are individually non-normal. PC1 and PC2 are both left skewed. PC3 has a jump around the standard normal value of 0.8 and values do not fall in a straight line.

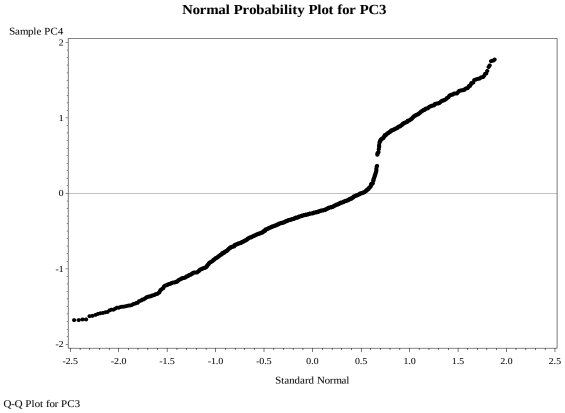
Graph8: Q-Q plot for PC1.



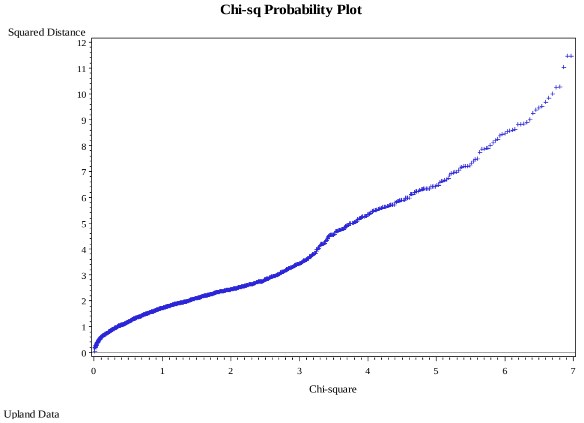
Graph9: Q-Q plot for PC2.



Graph10: Q-Q plot for PC3.



Graph11: Chi-Square Probability Plot



Since these components are not individually normal, multivariate normality must be assessed by constructing the Chi-Square probability plot (Graph 11). Notice that most of the values in the plot fall roughly in a straight line. Thus we can assume that these principal components are multivariate normal. This concludes the last step of Principal Component Analysis.

4 Logistic Regression

In Principal Component, we are able to narrow down the data to 3 variables (principal components). Since this gives us a quantitative interpretation of the data, our next step is to interpret our data set qualitatively. This will be done using Logistic Regression Analysis, a form of Discriminant Analysis, where we will be looking at a multi-way classification. Thus an ordinal logistic model is fit to a multinomial response variable. Various factors will also be considered for a more accurate classification in this method. Logistic Regression will be used first on the training data set from 2001-2003 to fit the model. Then, the data set from 2004 will be used as a validation for that model.

4.1 Creating New Variables

Looking back on table 2, the AQI levels will be further classified into 3 groups: healthy (0), unhealthy (1), and hazardous (3) refer to table 8. Variables month and weekend will also be brought back into the data set for more information in this analysis. However, the month variables will be created into dummy variables according to the season that each month falls under (shown in table9).

Table8: Classifying AQI.

Ozone Range (ppb)	Level of Health Concern	AQI
0-75	Healthy	0
76-115	Unhealthy	1
116+	Hazardous	2

Table9: Dummy Variable Assignments.

Month	Season	Dummy
12-2	Winter	1
3-6	Spring	2
7-9	Summer	3
9-11	Fall	Base level

4.2 Checking Frequency

Table10: Response Profile.

AQI	Total Frequency
0	678
1	287
2	130

Prior to logistic regression, checking the frequency of each AQI in the data set will give an idea of where a specific amount of observations fall under. Table 10 shows out of 1095 observations, 678 (61.9%) are classified as healthy, 287 (26.2%) observations as unhealthy, and 130 (11.8%) observations as hazardous.

4.3 Maximum Likelihood Estimates

Table11: Maximum Likelihood Estimates.

Effect	Wald Chi-Square	Pr> Chi-Square
PC1	199.7702	<0.0001
PC2	2.6356	0.2677
PC3	3.6571	0.1606
Season1	0.0051	0.9975
Season2	3.6456	0.1616
Season3	7.2278	0.0269
Weekend	102.327	<0.0001

In order to look at the classification, variables PC1, PC2, PC3, season1, season2, season3, and weekend will be included. First, in order to estimate the ß’s in the logistic regression model, a method of maximum likelihood is used. The proc logistics procedure gives the following results shown in table 11. Based on the p-value of the chi-square statistics of the parameter estimates, only PC1, Season3, and Weekend are significant with p-values < 0.0001 and 0.029, which is less than $\alpha = 0.05$. It would be advisable to take out all the insignificant variables and keep PC1, Season 3, and weekend. The next few sections will show deleting the insignificant variables will make a slight difference.

4.4 Confusion Matrix for AQI Including All Variables

Table12: Confusion Matrix on Training Set (2001-2003) w/ all variables

		Predicted			
	AQI	**0**	**1**	**2**	**Total**
True	**0**	625	53	0	**678**
	1	43	218	26	**287**
	2	2	65	63	**130**
	Total	**670**	**336**	**89**	**1095**

Table 12 gives us the Confusion Matrix of the training data set. It can be calculated that 92% are classified correctly as healthy, 75.9% are classified correctly as unhealthy, and 48.4% are classified correctly as hazardous. Thus a total of 82.19% of the data were correctly classified, which gives a 17.8% overall misclassification rate.

Table13: Confusion Matrix on Validation Set (2004) w/ all variables

		Predicted			
	AQI	**0**	**1**	**2**	**Total**
True	**0**	195	30	0	**225**
	1	9	70	20	**99**
	2	1	14	14	**29**
	Total	**205**	**114**	**34**	**353**

Table 13 gives us the Confusion Matrix of the validating data set. It can be calculated that 86.7% are classified correctly as healthy, 70.7% are classified correctly as unhealthy, and 48.3% are classified correctly as hazardous. Thus a total of 79.04% of the data were correctly classified, which gives a 20.96% overall misclassification rate.

4.5 Confusion Matrix for AQI With Only Significant Variables (PC1, Season3, & Weekend)

Table14: Confusion Matrix on Training Set (2001-2003) w/ all variables

		Predicted			
	AQI	**0**	**1**	**2**	**Total**
True	**0**	620	58	0	**678**
	1	46	215	26	**287**
	2	2	68	60	**130**
	Total	**668**	**341**	**86**	**1095**

Table 14 gives us the Confusion Matrix of the training data set. It can be calculated that 91.44% are classified correctly as healthy, 74.9% are classified correctly as unhealthy, and 46.15% are classified correctly as hazardous. Thus a total of 81.74% of the data were correctly classified, which gives a 18.26 % misclassification rate.

Table15: Confusion Matrix on Validation Set (2004) w/ all variables

		Predicted			
	AQI	**0**	**1**	**2**	**Total**
True	**0**	193	32	0	**225**
	1	9	71	19	**99**
	2	0	16	13	**29**
	Total	**202**	**119**	**32**	**353**

Table 15 gives us the Confusion Matrix of the validating data set. It can be calculated that 85.78% are classified correctly as healthy, 71.72% are classified correctly as unhealthy, and 44.83% are classified correctly as hazardous. Thus a total of 78.47% of the data were correctly classified, which gives a 21.53% misclassification rate.

By looking at the results from section 4.4 and 4.5, the Confusion matrices for both the training and validation set that includes all variables give a slightly higher accuracy and lower misclassification rate compared to the matrices with only the significant variables. However, the differences are negligible. Thus, looking at the model with variable PC1, Season1, and Weekend is enough to determine misclassification rates.

5. Linear Discriminant Analysis (LDA)

The Principal Component Analysis were able to narrow down the data to 3 variables, that is, the first three principle components. Since this gives us a quantitative interpretation of the data for the environmental variables, our next step is to interpret our data set qualitatively using Linear Discriminant Analysis, where we will be looking at a multi-way classification, to augment and compare with logistic regression. The discriminant analysis, unlike logistic regression, has the room to adjust priors. This will help us apportion heavier weight on the unhealthy category to minimize it’s misclassification at the expense of the healthy and moderate group. The priors suggested and adjusted were 0.15, 0.35 and 0.50 for healthy, moderate and unhealthy groups respectively. Discriminant Analysis will be used first on the training data set from 2001-2003 to fit the model. Then, the data set from 2004 will be used as a validation for that model.

5.1 Creating New Variables and Group Frequency

Looking back on table 2, the AQI levels will be further classified into 3 groups: healthy (0), unhealthy (1), and hazardous (2) refer to table 8. Variables month and weekend will also be brought back into the data set for more information in this analysis. However, the month variables will be created into dummy variables according to the season that each month falls under (shown in table9) and the distribution of the observations in the three groups as shown in table 10.

5.2 Confusion Matrix for AQI Including All Variables

Table16: Confusion Matrix on Training Set (2001-2003) w/ all variables

		Predicted			
	AQI	**0**	**1**	**2**	**Total**
True	**0**	521	129	28	**678**
	1	8	146	133	**287**
	2	0	15	115	**130**
	Total	**529**	**290**	**276**	**1095**

Table 16 alongside shows the Confusion Matrix of the training data set from discriminant analysis. It can be calculated that 76.84% are classified correctly as healthy, 50.87% are classified correctly as unhealthy, and 88.46% are classified correctly as hazardous. Thus a total of 71.42% of the data were correctly classified, which gives a 28.58% overall misclassification rate. Of more importance is the high accuracy in the unhealthy group.

Table17: Confusion Matrix on Validation Set (2004) w/ all variables

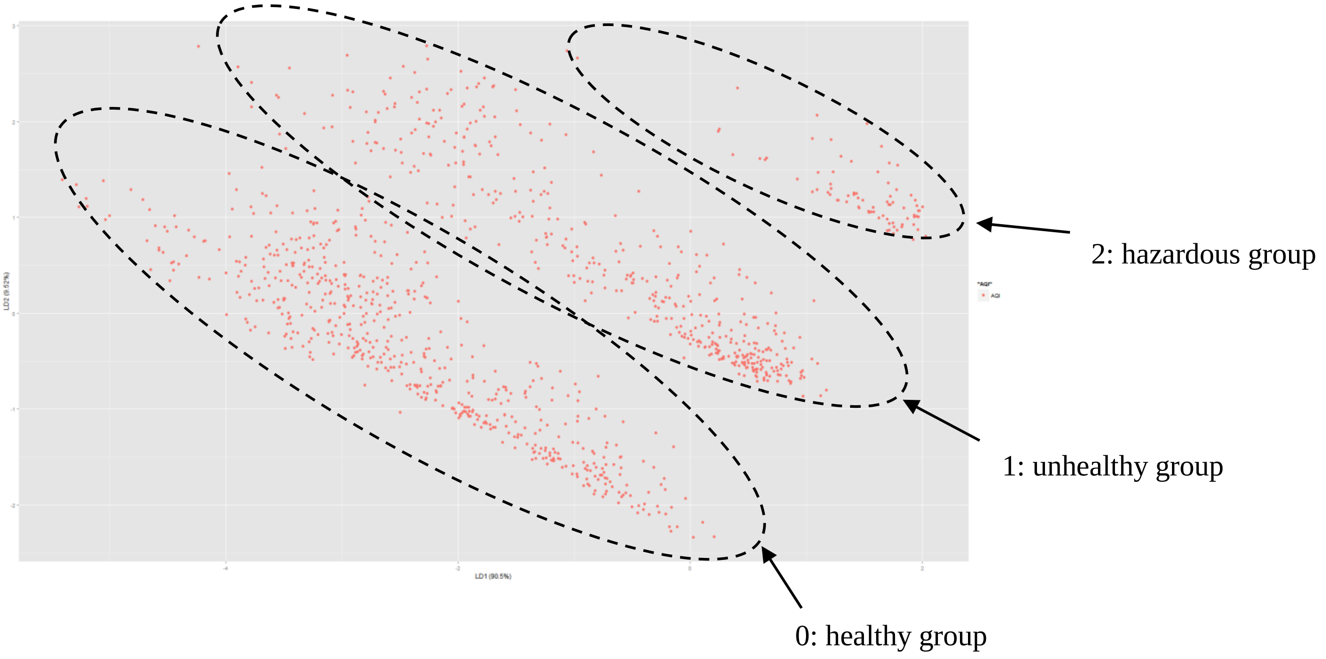
Predicted

	AQI	0	1	2	Total
True	**0**	155	66	4	**225**
	1	2	38	59	**99**
	2	0	5	24	**29**
	Total	**157**	**109**	**87**	**353**

Table 17 alongside shows the Confusion Matrix of the validation data set from discriminant analysis. It can be calculated that 68.89% are classified correctly as healthy, 38.38% are classified correctly as unhealthy, and 82.76% are classified correctly as hazardous. Thus a total of 61.47% of the data were correctly classified, which gives a 38.53% overall misclassification rate.

More emphasis was to achieve the high accuracy in the unhealthy group. It can be noted that most of the misclassified observations in the unhealthy group were classified into hazardous group, which is actually better than if they were classified into healthy group. This also applies to the 66 observations out of 225 that were classified into unhealthy instead of healthy group. Although the overall accuracy and misclassification rates in Discriminant Analysis were lower than in Logistic regression, more accuracy for the hazardous group was achieved in the Discriminant Analysis than Logistic regression. This accomplishment is particularly due to the room to adjust the priors to give more weight to the sensitive group (or the hazardous group) because misclassifying the hazardous group is extremely costly than the other two groups.

Graph 12: Discrimination graph for the Ozone data



6.0 Conclusion

PCA is an unsupervised learning technique (don't use class information) while LDA and Logistic regression are supervised techniques (uses class information), but both PCA and LDA provide the possibility of dimensionality reduction, PCA for variable reduction and LDA for observation reduction, which is very useful for visualization. LDA provides better data separation and or classification when compared to Logistic regression for respective groups of interest due to its ability to afford priors adjustment, and this is exactly what we see in tables 12-17 when both are applied to the Upland ozone dataset. The choice on which approach to use, amongst Logistic regression and LDA depends upon the data and significance of the results, whether the analyst is interested with overall accuracy and misclassification rates or individual particular sensitive group of interest.

7.0 References

[1] Venables, W. N. and Ripley, B. D. (2002). Modern applied statistics with S. Springer.
[2] Kuhn, M. and Johnson, K. (2013). Applied Predictive Modeling. Springer.
[3] Johnson, A. R. and Wichern, W. D. (2007). Applied Multivariate Analysis, 6 th. Ed. Pearson Education Inc. New Jersey, USA.
[4] Dr. Kim's Lecture Notes (Spring 2015)