9/25/2013

# Automated Detection of Pneumoconiosis Using Predictive Analytics

## Introduction:

A leading hospital has requested GE to develop an automated Pneumoconiosis early detection system for coal miners. Pneumoconiosis is an occupational lung disease typically caused by the inhalation of dust. Individuals working in mines will tend to develop this disease but it can also be brought on by the inhalation of Asbestosis, Silicosis, Siderosis and several other fine dust occupations [1] [2]. While there is no cure for Pneumoconiosis early detection can allow doctors to treat the symptoms and allow patients to continue a fairly comfortable life. The hospital has requested an automated system due to the lack of trained doctors who can examine the chest x-rays for anomalies. Success of the automated system will be measured by patient model accuracy, precision and recall.

The data that was provided consisted of 473 patient chest x-rays that have been analyzed by a team of image analysts and categorized as normal or abnormal (0 = Normal, 1 = Abnormal). Each lung x-ray has been broken down into three sections (right upper, right middle, right lower, left upper, left middle, left lower) where a set of algorithms have been developed to extract a set of features to be used for predictive modeling.

The analysis presented will show that a reasonable predictive model can be built on the features in the dataset, as well as short falls in the model and potential next steps.

## Methods:

### *Data Collection*

The data set provided was downloaded from GE shared folders [3] on August 28, 2013. The file consisted of 473 patient records where each lung was broken down into three sections each consisting of a set of extracted features. The extracted features provided for each section consisted of the mean, standard deviation, skewness, kurtosis, energy and entropy on the histogram of intensity values. Along with these 6 features at various orientations (0, 30, 35, 60, 90, 120, 135, 150 and 180 degrees) 5 additional features were provided based on the gray level co-occurrence matrix computed for the ROI (energy, entropy, local homogeneity, correlation and inertia), for a total of 39 extracted features.

While 39 extracted features were provided in the dataset the original dataset contained 222 intensity based features and 25 co-occurrence matrix based features. Without analyzing all 247 features it is unknown if the additional features would add any additional explanatory value to the model.[4] In addition to all 247 extracted features it would be beneficial to know each patients occupation, if worked in a dusty environment the number of years in such an environment as well as if the patient worked continuously in the environment or periodically each day.

### *Exploratory Analysis*

Prior to building predictive models, an exploratory analysis using JMP [5], was conducted for each factor provided. A distribution analysis was conducted to determine if the data was normal, skewed right or left, or contained outliers. Figure 1 provides a sample distribution analysis where most factors appear to be normally distributed with a few factors skewed and potentially an outlier or two.

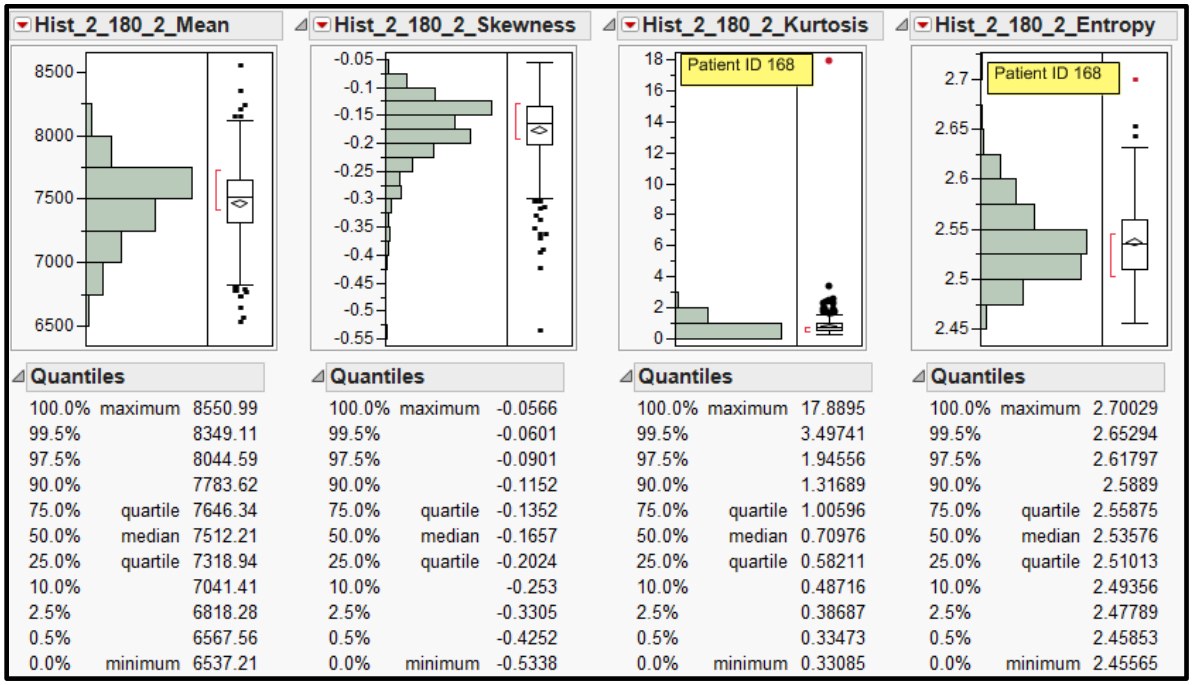


**Figure 1 Feature distribution analysis (sample), most features normally distributed and indicating potential outliers.**

Looking into the potential outliers it appears that patient number 168 and 375 have several factors where the value lie well beyond 3 standard deviations of the mean. Both of these patients should be evaluated during the modeling phase to determine how influential these values are in the model. The factors that appear to be skewed may benefit by transforming the values using a logistic transformation and should be evaluated during the modeling phase as well.

Along with the distribution analysis, each factor was plotted on a scatter plot matrix to determine if any factors are confounded. Figure 2 shows a sample scatter plot matrix where several factors appear to be correlated and should be evaluated during the modeling phase for inconsistencies or adverse influences to the model. When comparing each independent factor against the dependent variable there was no apparent correlation between the two.

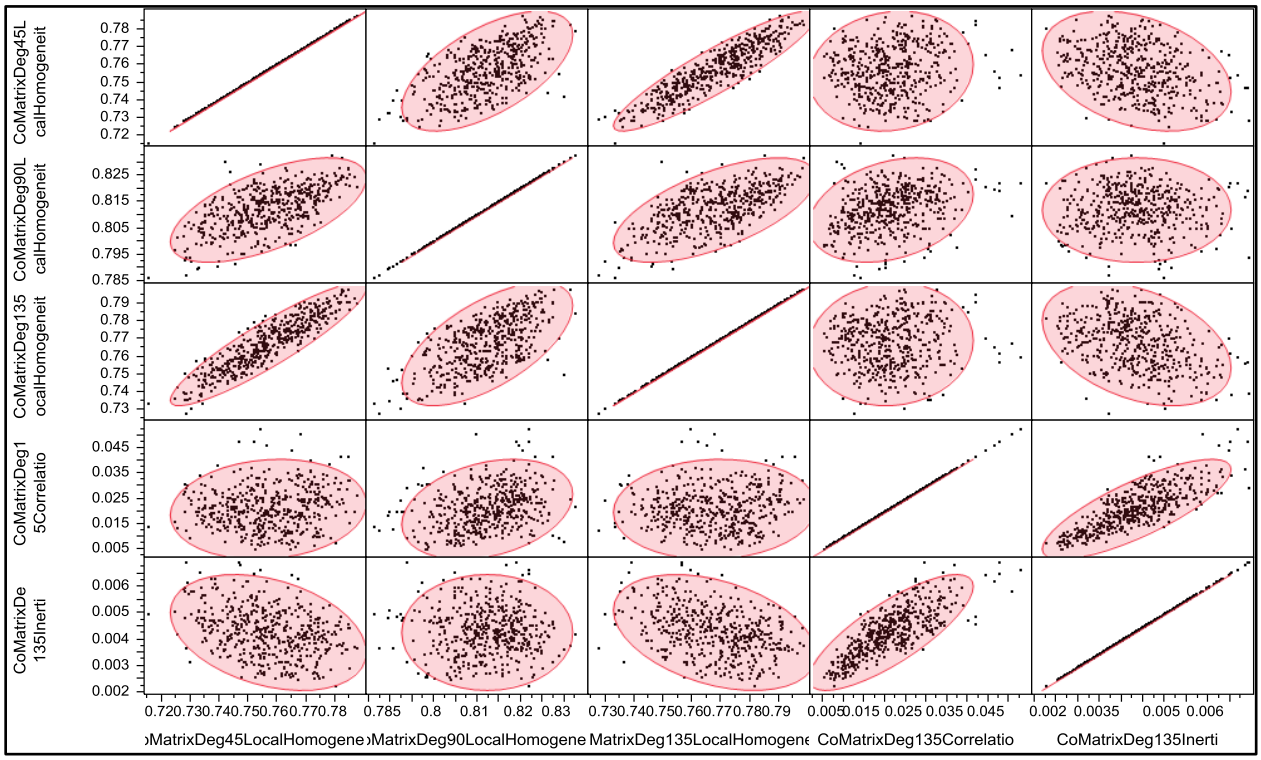


**Figure 2 Feature correlation matrix (sample), with a few factors indicating correlation and potential confounded features.**

It became evident while evaluating each lung section of data that not all patients have data for all six sections of the right and left lung. This will influence the modeling aspect and the ability to combine all sections of data into a single large model. Each lung section may have to be modeled separately or combined in a way where missing values are either ignored or imputed. Since imputing the values could result in a false image of data and a larger misclassification rate this approach is not recommended. Rather dealing with the missing data by having various modeling techniques based on the situation is a better approach and one that has been employed for this analysis.

### *Statistical Modeling*

After reviewing the exploratory analysis several modeling techniques were employed via SAS [6] as a baseline to determine which technique should be explored further. The first models generated was a Logistic Regression [7] and Random Forest [8] to set a baseline for

predictive capability. Both the Logistic Regression and Random Forest were selected because they are quick to setup and provide two separate techniques of modeling the data. Logistic Regression identifies if a linear model is appropriate and Random Forest helps to decide if multiple random iterations benefit the model.

Random Forest predicted very well on a training data set for each lung section but was only marginally better than the Logistic Regression when implemented on a validation data set. Both showed patient accuracy around 85%. As the database grows Random Forest methodology should be re-evaluated to determine if the predictive capability improves, this is because Random Forests typically do well on larger data sets and not small data sets such as the Pneumoconiosis.

The Case Study instructions indicated that each patient should be evaluated with a separate model that employs a leave-one-out cross-validation methodology [9], neither the logistic regression nor the Random Forest were setup with this methodology. Random Forest is similar to a random cross validation and due to the small number of data points would not benefit largely by writing specific code to implement a leave-one-out cross-validation. Instead of writing specific code to implement a leave-one-out cross-validation with Logistic Regression a Partial Least Squares (PLS) methodology [10] was used where leave-one-out cross-validation is an option when building a model.

The benefit of using PLS over logistic regression is that PLS procedure tries to minimize predictor variation as well as the response variation where Logistic Regression only tries to minimize the response variation. "All of the techniques implemented in the PLS procedure work by extracting successive linear combinations of the predictors, called factors (also called components or latent vectors), which optimally address one or both of these two goals (response variation and or predictor variation)". [11]

When building the PLS model several modeling approaches were evaluated. The first being individual lung section models where the final results were later aggregated to a single patient response. The second approach, and the one that provided the best response, modeled the left and right lung aggregated section data where data was available for each patient. Any patient that did not have all three sections of data for a given lung was evaluated using the individual section models. This also provided the best flexibility with new data as it would be unknown if all sections of each lung would be present.

In order to build separate models for each patient and still employ leave-one-out cross-validation a macro loop was implemented where each patient was pulled out of the dataset individually; a model was generated using leave-one-out cross-validation and scored the patient that was left out based on the model generated. This was then repeated for each subsequent patient and lung section to build out the separate models for each patient. An

excerpt of SAS code depicting the macro loop and scoring of the patient data can be found in Appendix A.

***Reproducibility***

The analysis conducted can be reproduced with the HealthCareProject.egp (SAS Enterprise Guide File) and is available upon request.

## Analysis:

Evaluating the model, potential outliers were examined and determined if they should be removed or not. Both patient ID 168 and 375 were identified as potential outliers in the exploratory analysis. The PLS model was built with both patient ID's excluded and then re-ran with both included. For patient ID 168 the results from lung section to lung section did not change even though several of the PLS factors showed this patient to potentially be an outlier when plotting response factor against the predictor, see Figure 3.

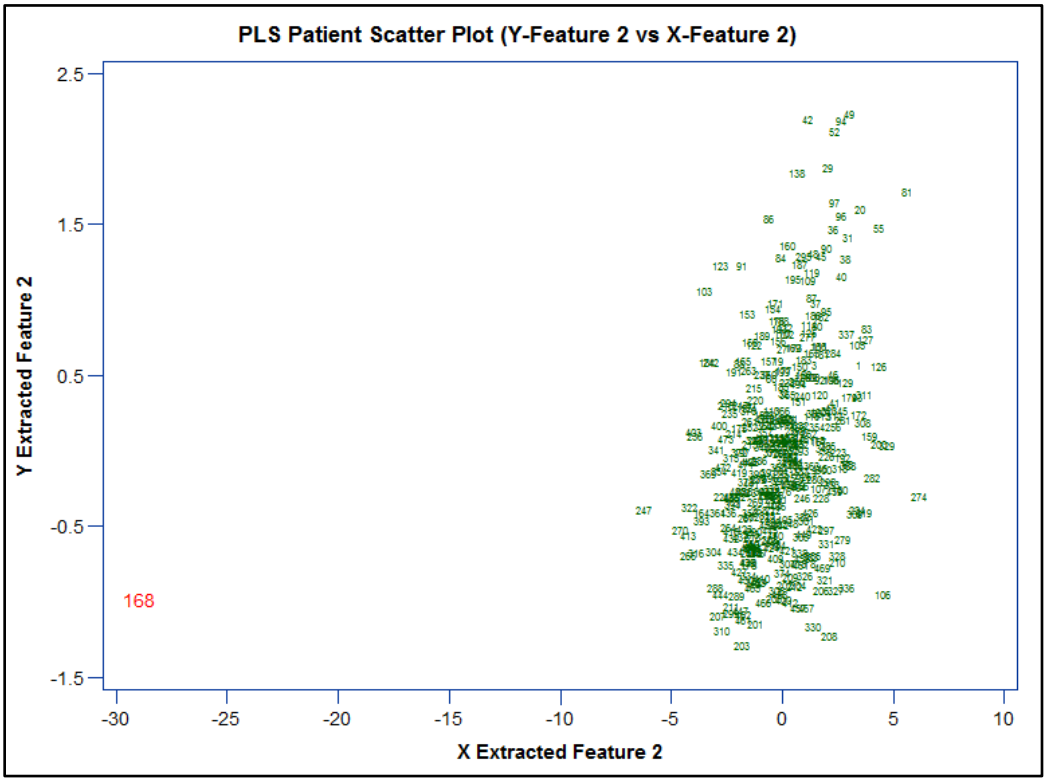


**Figure 3** PLS extracted predictor and response feature plot highlighting potential patient outliers (patient no 168)

This can also be seen when plotting one predictor against another predictor, see Figure 4. Because the results did not change, patient 168 was left in the model. However, as the database grows outliers like patient 168 should be evaluated from time to time to ensure no adverse influence is generated in the model.

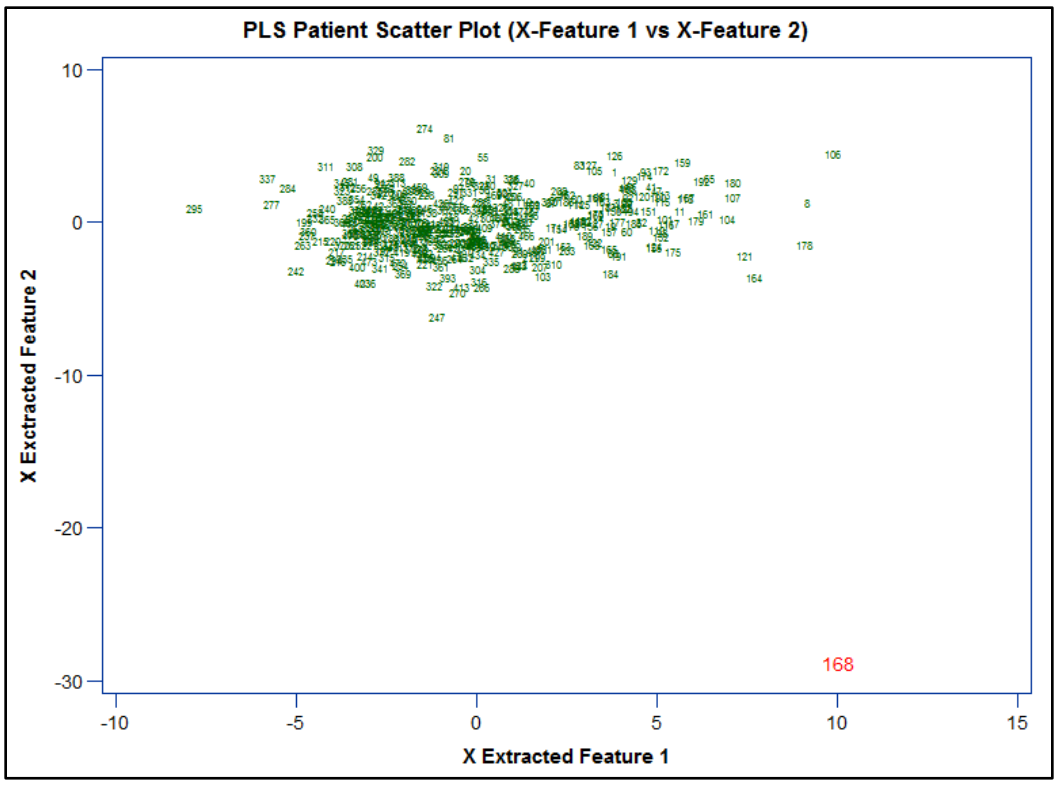


**Figure 4** PLS extracted predictor feature plot highlighting potential patient outliers (patient no 168)

With patient 375 the PLS model failed to compute the response or predictor factors. Due to the PLS model failing, patient ID 375 was omitted from the overall analysis.

During the modeling phase it was determined that transforming the features that was skewed left or right (primarily kurtosis column of data) did not provide any benefit to the model. This was primarily due to the "center" and "scale" option within the PLS procedure. This option centers each factor to have a mean of zero and a standard deviation equal to one. This is beneficial and recommended as it places all factors and the dependent variable on equal footing relative to their variation in the data.

The results of running separate lung section models rolling up to a patient having or not having Pneumoconiosis can be seen in table 1, where accuracy resulted 89.2%.

| | Actual | | | Result Measurements | |
|---|---|---|---|---|---|
| Predicted | 1 | 0 | | **Accuracy** | 89.19% |
| 1 | 184 | 40 | | **Precision** | 82.14% |
| 0 | 11 | 237 | | **Recall** | 94.36% |

**Table 1** Patient Actual vs Predicted Results, (*PLS Model for individual sections*).

This was improved upon with building separate right and left lung models where data was available for each section of a patient. The remaining patients who did not have three sections of data for each lung were diagnosed using the individual lung section models. The results can be seen in table 2, where accuracy was improved to 92.6% or a 3.4% improvement while also improving the number of false positives with precision at 94.4%.

| | Actual | | | Result Measurements | |
|---|---|---|---|---|---|
| Predicted | 1 | 0 | | **Accuracy** | 92.58% |
| 1 | 170 | 10 | | **Precision** | 94.44% |
| 0 | 25 | 267 | | **Recall** | 87.18% |

**Table 2** Patient Actual vs Predicted Results, (*PLS Model for 3 section data and individual section data*).

By implementing the second approach false positives was reduced as was true positives while false negatives and true negatives increased. Because part of the goal was to minimize the number of false positives the second modeling approach is a better approach over the first and would be recommended as the automated detection model for Pneumoconiosis.

## Conclusions:

It has been shown that an automated detection system for Pneumoconiosis can be created using Partial Least Squares methodology. PLS with leave-one-out cross-validation outperforms other modeling methods with an accuracy of 92.6% while keeping false positives low. Because the data set provided only included 39 of the 247 feature list it is unknown if the additional features would provide any improvement to the model capability. Rerunning the PLS model with the full feature list should be explored and evaluated for model improvements.

In addition to evaluating the full feature list the misclassified patients should be evaluated individually to determine if the data was misinterpreted or if the x-rays show an image quality problem and not representative of the physical state of the patient. It has also been shown that the model performs better when all three lung sections are available for analysis versus individual section models. Patients without all six sections of data should be understood since the x-rays are broken down into the different sections. Determining why the data is not present may provide insight into potential issues with the sections that were provided.

As more data becomes available additional methodologies, such as Random Forests, should be evaluated since they are well suited for larger data sets and may outperform a PLS model.

## Appendix:

Excerpt of SAS macro loop implementing PLS with leave-one-out cross-validation.

```
%Macro Loop(dataset1, results1);

proc sql noprint;
select distinct PatientNumMasked
       into: ID seperated by ","
From Outlier_Remove;
Quit;
%put &ID;
%let inc = 1;
%do %while (%scan(%bquote(&ID), &inc, %str(,)) ne);
%let PatientID = "%scan(%bquote(&ID), &inc, %str(,))";
%put &PatientID;

Data patient_subset patient_score (drop=Pred_Formula_Label round);
set Outlier_Remove;
if PatientNumMasked = &PatientID then output patient_score;
else output patient_subset;
rename label=Target ;
Run;

data patient_score (Drop=Target);
set Patient_score;
Run;

Proc append base=patient_subset data=patient_score;
run;

proc PLS Data=patient_subset CV=one ;
model Target = Hist_0_0_0_Mean Hist_0_0_0_Skewness Hist_0_0_0_Kurtosis
   Hist_0_0_0_Entropy Hist_2_45_1_Entropy Hist_2_60_1_Skewness Hist_2_90_1_Skewness
   Hist_2_90_1_Kurtosis Hist_2_135_1_Entropy Hist_1_150_1_Skewness
   Hist_2_180_1_Skewness Hist_1_30_2_Mean Hist_2_30_2_Mean Hist_2_30_2_Entropy
   Hist_2_60_2_Skewness Hist_2_60_2_Kurtosis Hist_1_90_2_Skewness Hist_2_90_2_Mean
   Hist_2_90_2_Skewness Hist_2_90_2_Kurtosis Hist_1_120_2_Mean Hist_1_135_2_Mean
   Hist_1_135_2_Entropy Hist_2_150_2_Mean Hist_2_150_2_Skewness Hist_2_150_2_Kurtosis
   Hist_2_150_2_Entropy Hist_1_180_2_Mean Hist_1_180_2_StdDev Hist_1_180_2_Skewness
   Hist_2_180_2_Mean Hist_2_180_2_Skewness Hist_2_180_2_Kurtosis Hist_2_180_2_Entropy
   CoMatrixDeg45LocalHomogeneity CoMatrixDeg90LocalHomogeneity
   CoMatrixDeg135LocalHomogeneity CoMatrixDeg135Correlation CoMatrixDeg135Inertia;
Output out=pred p=p_target;
run;

Data _pred;
set pred;
if PatientNumMasked = &PatientID then output _pred;
Run;

proc append base=&results1 data=_Pred;
Run;

%let inc = %eval(&inc+1);
%end;

%Mend Loop;
Run;
```

## References:


1. Wikipedia "Pneumoconiosis"; URL: http://en.wikipedia.org/wiki/Pneumoconiosis, accessed September 22, 2013
2. American Lung Association, "Pneumoconiosis"; URL: http://www.lung.org/lung-disease/pneumoconiosis/, accessed September 22, 2013
3. Case Study Data File, "CollatedPneumoconiosisData-GE Internal.xlsx" URL: http://libraries.ge.com/download?fileid=424754832101&entity_id=36970369101&sid=101 , accessed August 21, 2013
4. Case Study Instructions/ Field Attributes, "Case_Study_-_Detecting_Pneumoconiosis.docx" URL: http://libraries.ge.com/download?fileid=424754759101&entity_id=36970369101&sid=101 accessed August 21, 2013
5. JMP. "JMP Support and Documentation" URL: http://www.jmp.com
6. SAS. "SAS 9.3 Product Documentation" URL: http://www.sas.com
7. Shmueli G., Patel N., Bruce P., (2010). *Data Mining for Business Intelligence*, 2nd Edition, New Jersey Wiley and Sons.
8. Berkeley University - Random Forests; URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, accessed March 5, 2013.
9. Wikipedia "Cross-validation"; URL: http://en.wikipedia.org/wiki/Cross-validation_(statistics), accessed September 22, 2013
10. Witten, Frank, Hall., (2011). *Data Mining Practical Machine Learning Tools and Techniques*, 3rd Edition, Morgan Kaufmann Publishers.
11. SAS Institute Inc. 2011. *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute Inc.