Reducing Archives to Build Non Linear Models Using Neural Networks

Gerardo A. Colmenares L.

Universidad de Los Andes, IIES-FACES
Núcleo La Liria. Edif. G. Piso 3, Ofic. 233. Mérida. Venezuela
gcolmen@.ula.ve

Abstract

This paper describes a method to select from an original huge data set, representative data to train, test, and validate neural network models. It was called Stratified/PCA and applies stratification and principal component analysis to efficiently reduce the amount of observations (records) and original variables. The new set keeps a high amount of the original data set. The performance of neural network models built using those reduced data sets is very similar to that of neural network models built using the entire data set. In fact, it is both, significantly better and consistent than any known data selection method, including those based on random selection criterion. A kind of recognition pattern can be found within Stratified/PCA. Therefore, this novel technique can be applied as a data mining or preprocessing method to efficiently build non linear models using neural networks.

Keywords

Data mining, neural network models, stratification, principal component analysis, data reduction, and variable reduction.

1. Introduction

Frequently, when historical data is collected, a large number of observations are stored and, with each observation, a large number of variables are included. One reason for the large size of these data sets may be that once the initial cost of setting up the data collection mechanism is incurred, the additional cost of collecting more data is comparatively small and may avoid future data collection costs if unforeseen use of the data is later identified. Historical series of great volumes of data can be taken advantage of the generation of new information. It can be originated with the presence of inherent data which was indirectly collected when was created the original source of data.

Neural networks are generally recognized as an important model-building technique that can take advantage of the existence of historical data sets. There are numerous

examples of successful applications using this technique [6], [17]. However, large number of observations and variables in historical data present challenging and interesting problems for neural networks. The entire data set available for building neural network models is usually divided into two categories: training and testing. All relevant characteristics of the problem should be represented in each of these categories, otherwise poor models would be built or misleading performance results would be reported. If the data used to train the neural network is not representative of the entire data set, then the model will perform poorly on the data selected for testing the trained neural network. On the other hand, a neural network trained with representative data will perform poorly during testing if the data set selected for testing purposes is not representative of the entire data set. Another problem that neural network builders face when using data sets with large number of observations and variables is the large amount of time that they take to train the networks and the increased complexity of the resulting network architecture.

This paper deals with the design and implementation of a reliable and consistent data preprocessing method to reduce observations and variables from large data sets. Field of *exploratory data analysis* is eventually aimed to analyze those data sets using both univariate or multivariate statistical techniques and, heuristic techniques well-known like *data mining*. [17], [18]. Data exploration involves certain prior processes before building neural network models able to reach an acceptable performance to predict or recognize patterns. In that case, well combined procedures using multivariate and statistical techniques are useful as alternative tools capable of aiming a *data pre-processing*. This particular method integrates the concepts of stratification and principal component analysis to select representative observations and to eliminate redundant variables from these data sets. Neural network models trained and tested with data sets selected from an original data set using this method perform better than neural network models trained and tested with data sets selected randomly from the same original data set. In addition, neural network models built with a reduced number of variables selected by this method perform quite similar to models built using all the variables in the original data set. [4], [5]

2. Reducing the Number of Observations and the Number of Variables

Stratification reduces the number of observations using the dependent variable as screening variable and ignores the effect of independent variables. Therefore, it cannot be used to eliminate unnecessary independent variables. [2], [4], [5]. Stratified sampling technique, [3], also known as a variance reduction technique, aims to achieve a reduction in size but still maintaining a data set with similar statistical properties to the entire data. The main property of this technique is to achieve a minimum deviation between the mean value for the same variable founded with the selected sample and the entire data.

Large data sets, however, often include a large number of independent variables and observations. It is very likely that some of the variables are irrelevant for the problem at hand. Also, a group of variables may carry the same information about a particular problem. Eliminating irrelevant variables from a data set or replacing a group of variables with one variable carrying the same information can significantly reduce the size of the data. Neural network models built using this reduced data are likely to have fewer computational units and require less training time. [4], [8], [14].

Principal Component Analysis (PCA) [10], [13] is a proven statistical method that can reduce the number of variables in a data set with minimal loss of relevant information. PCA uses the covariance matrix and the correlation matrix to create a new and smaller set of variables with principal components equivalent to the principal components of the original set of variables. But in doing so, PCA keeps all of the observations in the original data set. Therefore, even if PCA is used to reduce the dimension of a data set, the problem of selecting truly representative training and test sets required to build reliable neural network models still remains. In other words, the principal components in the samples selected for these subsets should be as close to the principal components in the original data set as possible.

The proposed method is able to reduce both observations and variables from an original data set. Therefore, it can also be used to select truly representative training and test data sets to build reliable neural networks. This method integrates stratification and PCA and is referred to in this paper as Stratified/PCA [4].

Some relevant characteristics of this Stratified/PCA method are:

- The dependent and independent variables are examined as a whole.

- The new set of independent variables is based on the concept of explained variance.
- The new set of independent variables is in the same orthonormal basis.
- It relies not only on the variances provided by the first principal components, but also on the correlation between the minor components and the dependent variable.

In addition to the characteristics mentioned above, this method relies strongly on guidelines established by Kaiser H. [10], Jolliffe I. [9], Mardia K. et al. [13], to improve the reduction of the number of independent variables. These guidelines suggest that

- The eigenvalues of the correlation matrix R or covariance matrix Σ of the independent variables are the appropriate mechanism to reduce the number of independent variables.
- The correlation between the reduced independent variables and the dependent variable is important in determining the final reduced set of independent variables.
- The explained variance for the reduced independent variables guarantees that these new variables will be representative of the original independent variables.
- The correlation between the principal components and the original independent variables is recommended as a way to determine whether or not the new independent variables will be representative.

Before describing the Stratified/PCA method, some previous efforts made to address this problem are worth noting. Kramer, M. [11], [12] introduced the concept of autoassociative neural networks as a mechanism to reduce the number of independent variables. With a network architecture of five layers, Kramer's network uses the first two layers (input and mapping layers) to reduce the original independent variables to a new set of independent variables in the intermediate layer (the bottleneck layer). From the intermediate layer, the reduced set of variables is used as input variables to recover the original independent variables using the next two other layers (mapping and output layers). This reduction process uses nonlinear functions in the mapping and bottleneck layers as a mechanism to recover the non-linearity present in the original variables.

Huge data sets with a large number of variables would be difficult to train using Kramer's network because of the five-layer network and the number of nodes required in the training process. Another drawback of this proposed method is that the intermediate layer includes an arbitrary number of reduced independent variables and it is not clear how to determine their number.

Similar to Kramer's work, Tan et al. [16] introduced the concept of IT-net (Input-Training-net) as a variation of Kramer's autoassociative neural networks. They proposed a neural network with a single hidden layer to reduce data size. The single hidden layer includes the reduced independent variables. Tan et al., analyze the correlation matrix of the original independent variables, before reducing the variables using the IT-net. They create groups of independent variables highly correlated among them and select only the first principal component for every group of correlated variables. The rest of the variables remain the same since they are themselves independent. After completing this primary reduction, the new temporary set of independent variables determined by the selected principal components and the variables that do not belong to any group are used as input variables in the IT-net network. Those first principal components might belong to a different orthonormal basis since every group of correlated variables might determine different dimensional spaces. Moreover, the original variables not being grouped remain in the original dimensional space.

The stratified/PCA preserves the reduced independent variables in the same orthonormal basis. The values of the reduced independent variables in the selected samples would keep a high degree of correspondence with those of the entire data. These issues are not met by IT-net because the selected principal components do not represent the entire original variables.

Dong et al. [7] developed a model with two three-layer networks to compress and decompress data with a nonlinear behavior within the variables. They assumed that if PCA is applied to nonlinear problems, important information held in the last principal components (minor components) could be ignored due to their very small variances. They assumed that an excessive number of principal components would be present if the minor components were also being kept. If this reduction were applied to large data sets with a large number of observations where the variables are separated in one dependent variable and the rest as independent variables, the previous assumption would not consider the importance of the correlation of both the independent and dependent variables.

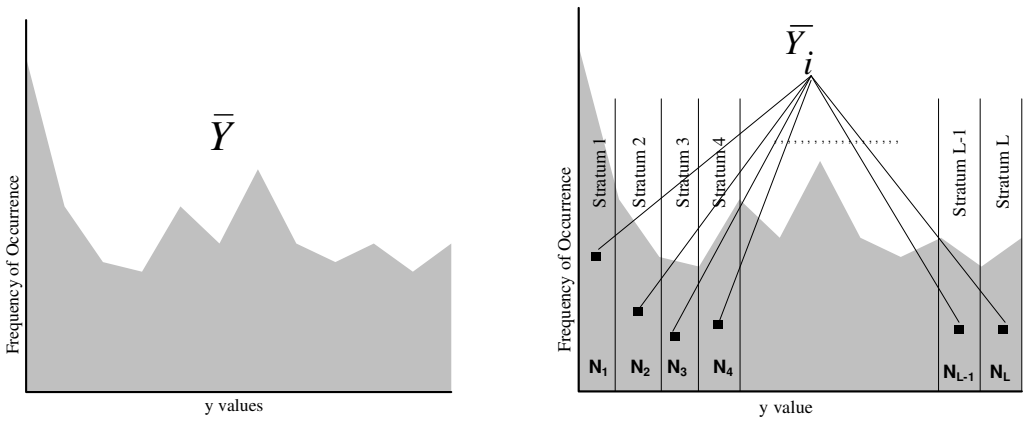
The stratified/PCA method analyzes not only the variances provided by the first principal components, but also the correlation between the minor components and the dependent variable. [9]. This method adds any minor components correlated with the dependent variable. In fact, this method showed that part of the nonlinearity of the independent variables could be captured.

Dong et al. [8] randomly separated the data sets for training. The stratified/PCA method proved that the training data sets selected from the original data set by stratification were more reliable. [4], [5]. Therefore, it can be applied as a data mining or preprocessing method to efficiently build non linear models using neural networks. [17].

2.1. The Stratified/PCA Method

The Stratified/PCA method can be described as the following sequence of steps:

Step 1. The entire data is separated in strata using the dependent variable like a variable of stratification as described in the previous section. The new set of observations of the dependent variable must be selected samples with a high confidence of representation of the entire data. Data is separated into L strata such that the summation of the sizes for every stratum i, N_i, must be the size N of the entire data, $N = \sum_L N_i$, and the mean for the stratification variable of the entire data must be $\overline{Y} = \sum_L \frac{N_i \cdot Y_i}{N}$.



Entire data for the dependent variable Stratified data of the dependent variable

Figure 1. Stratification of the main variable

Step 2. The eigenvalues of the correlation matrix are computed for every stratum of the entire data. By stratum, those values will be compared and evaluated one by one between the entire data and the entire sample. PCA is applied to every stratum, *i* in the entire data. Eigenvectors $E^{i}_{[pop]}$ and eigenvalues $\Lambda^{i}_{[pop]}$ are estimated for every stratum from the correlation matrix $R^{i}_{[pop]}$ of the independent variables $X^{i}_{[pop]}$.

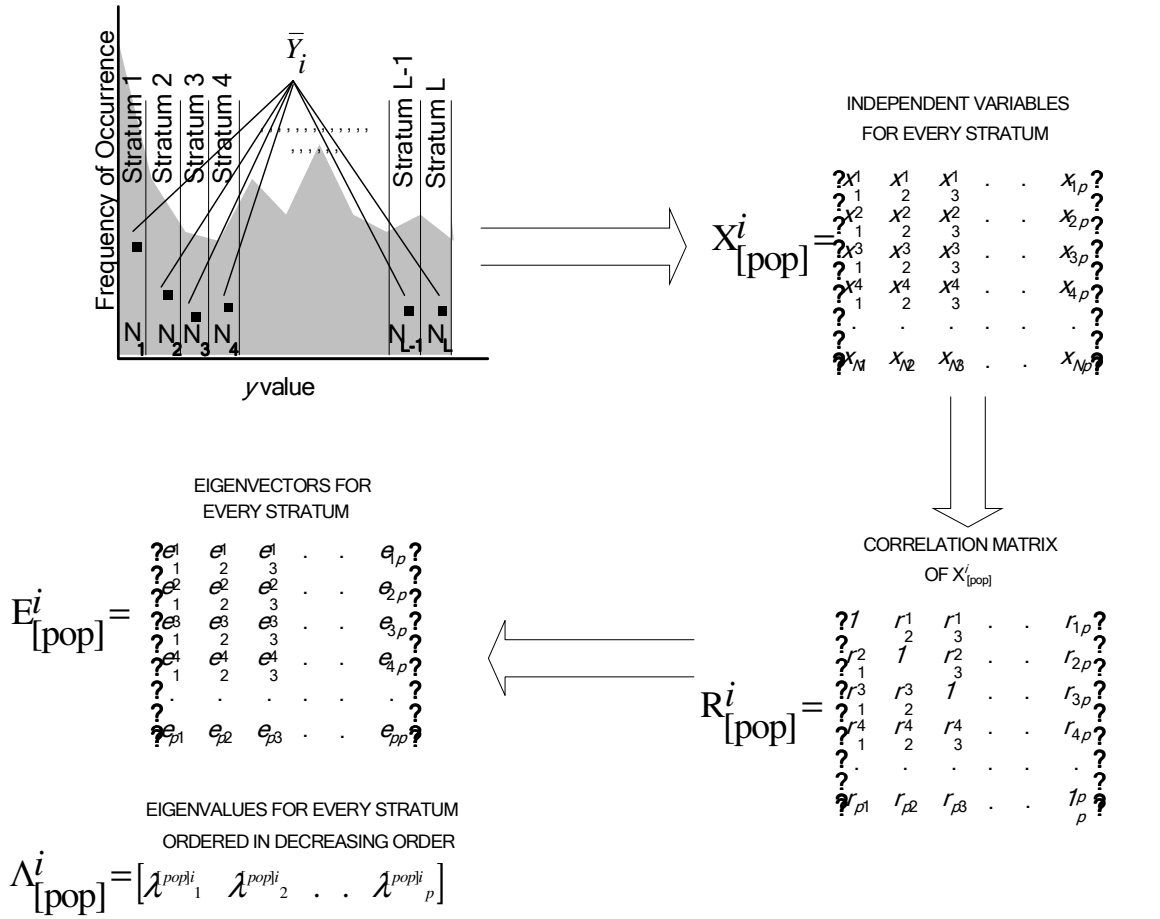
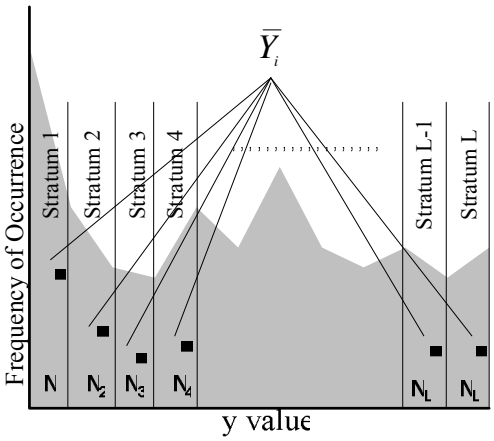
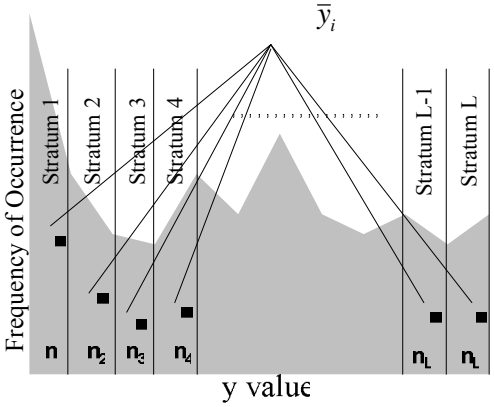


Figure 2. Principal components of the entire data in every stratum

Step 3. Stratified samples of size n_i are selected from every stratum of size N_i of the entire data.



Entire data after stratification



Samples from the entire data using stratification sampling

Properties $\overline{Y} \approx \overline{y}$ $\overline{Y}_i \approx \overline{y}_i$

Sizes $\mathbf{N}_i$ and $\mathbf{n}_i$ are proportional

Variance is minimized

Figure 3. Stratification process between the entire data and the selected samples

Step 4. PCA is applied to the independent variables $X_{[smp]}$ for the entire sample selected in step 3. Using the correlation matrix $R_{[smp]}$ of the independent variables, the eigenvectors $E_{[smp]}$ and eigenvalues $\Lambda_{[smp]}$ are estimated.

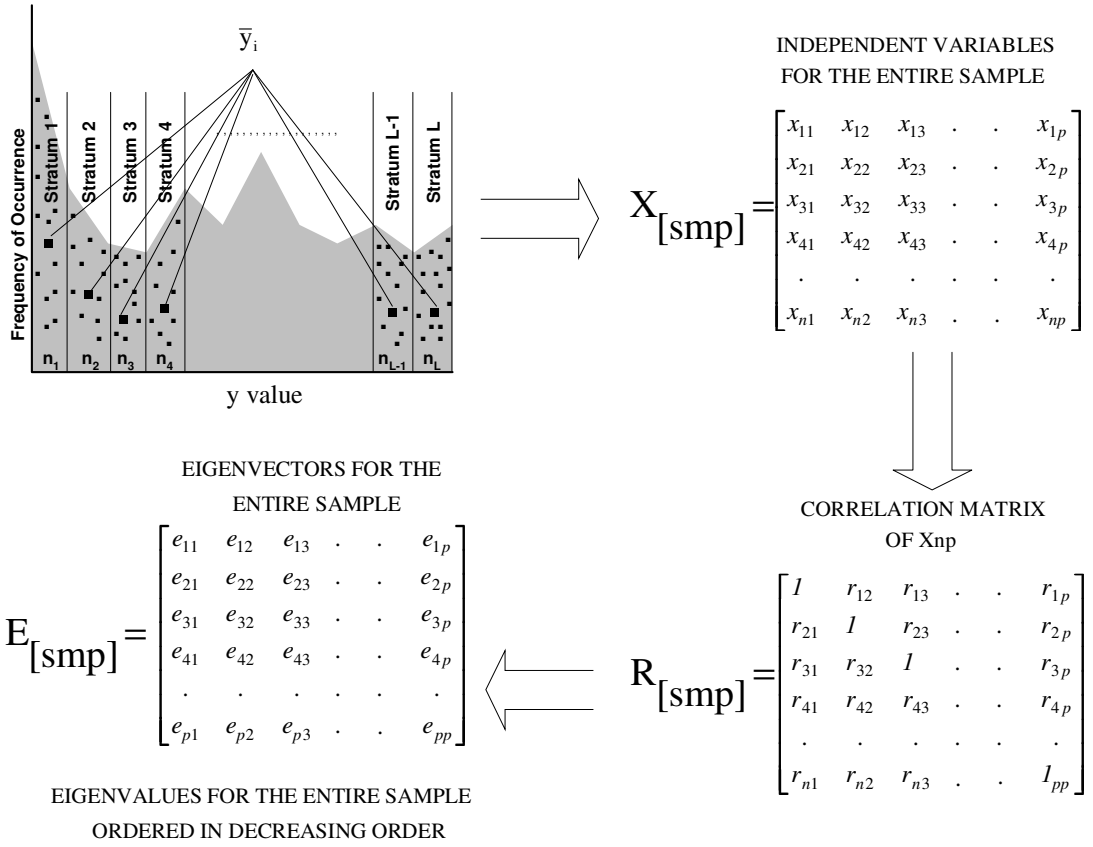


Figure 4. Principal components for the entire sample

Step 5. PCA is applied to the independent variables $X^{i}_{[smp]}$ for every stratum *i* of the selected sample $X_{[smp]}$ obtained in Step 3. Using the correlation matrix $R^{i}_{[smp]}$ of the independent variables, the eigenvectors $E^{i}_{[smp]}$ and eigenvalues $\Lambda^{i}_{[smp]}$ are estimated.

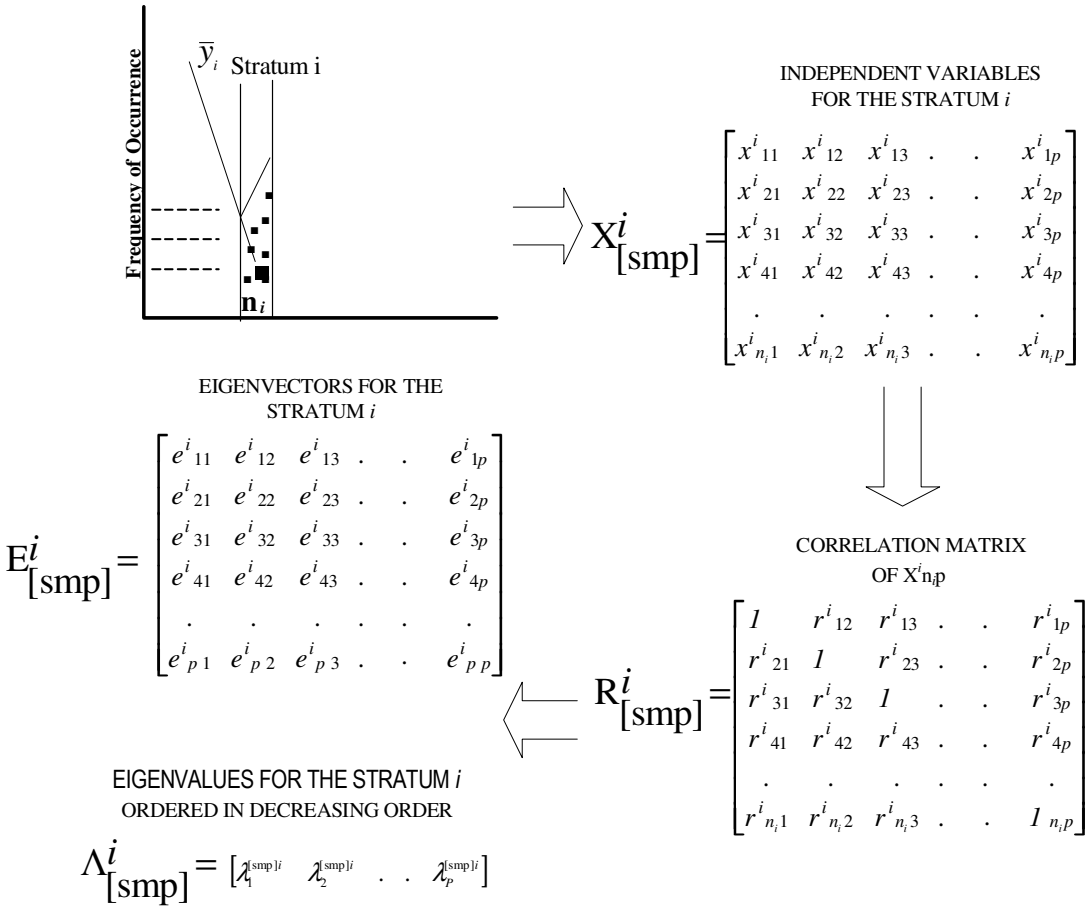


Figure 5. Principal components of the sample in every stratum

Step 6. The eigenvalues $\Lambda^{i}_{[smp]}$ and the correlation matrix $R^{i}_{[smp]}$ are evaluated stratum by stratum using the guidelines established by Kaiser H. [10], Jolliffe I. [9], Mardia K. et al. [13]. The eigenvalues for every stratum $\Lambda^{i}_{[smp]}$ in the selected sample $X_{[smp]}$ greater than a given threshold value allow the selection of the first candidates of principal components k_1^i.

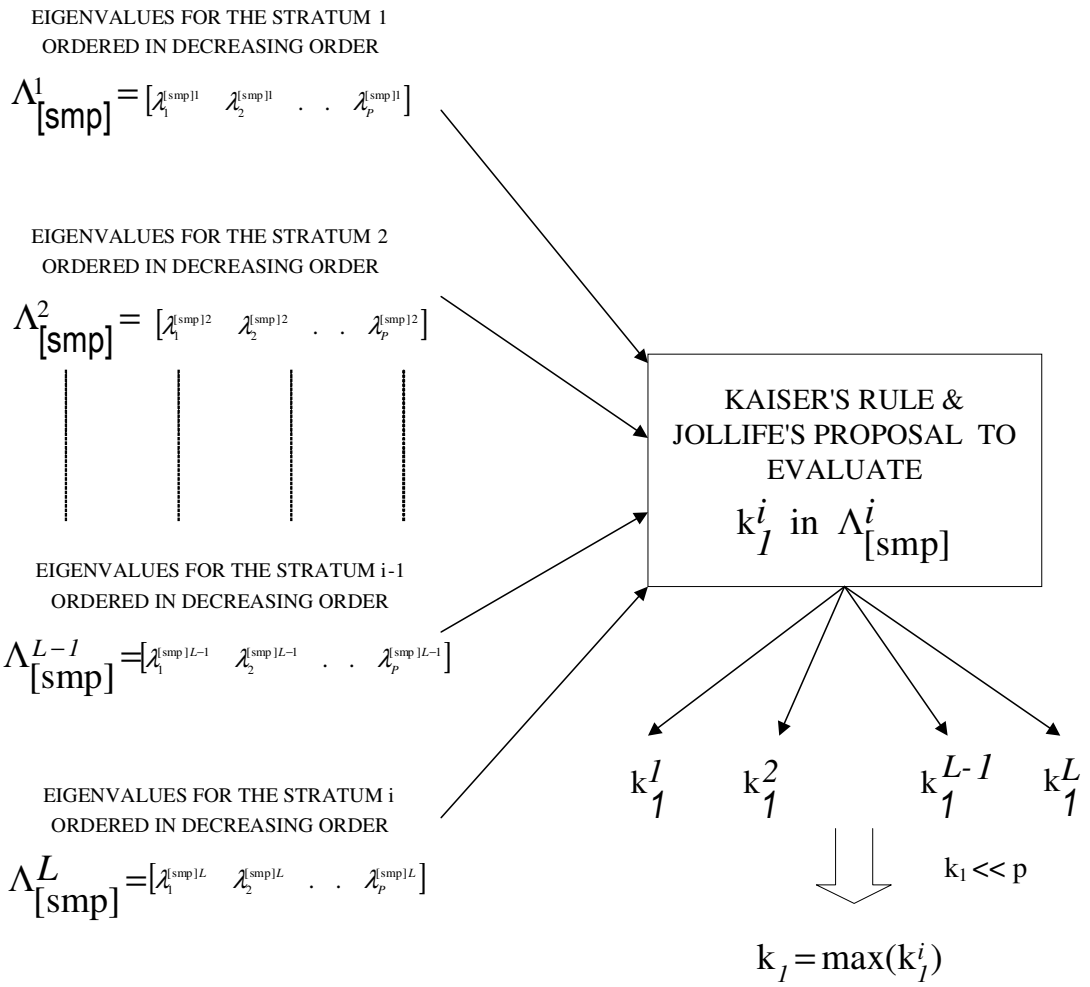


Figure 6. Selection of the k_1 principal components

Step 7. The percentage of explained variance PS_i given by the eigenvalues $\Lambda^{i}_{[smp]}$ of the k_1^i principal components for every stratum in the selected sample is compared with the percentage of explained variance PP_i given by the corresponding eigenvalues $\Lambda^{i}_{[pop]}$ of the k_1^i principal components for every stratum in the entire data. If every percentage in PS_i is greater than its corresponding percentages in PP_i, then the stratified sample is selected and the last eigenvectors should determine the new orthogonal axis for the new independent variables. Otherwise, a new stratified sample is selected from Step 3.

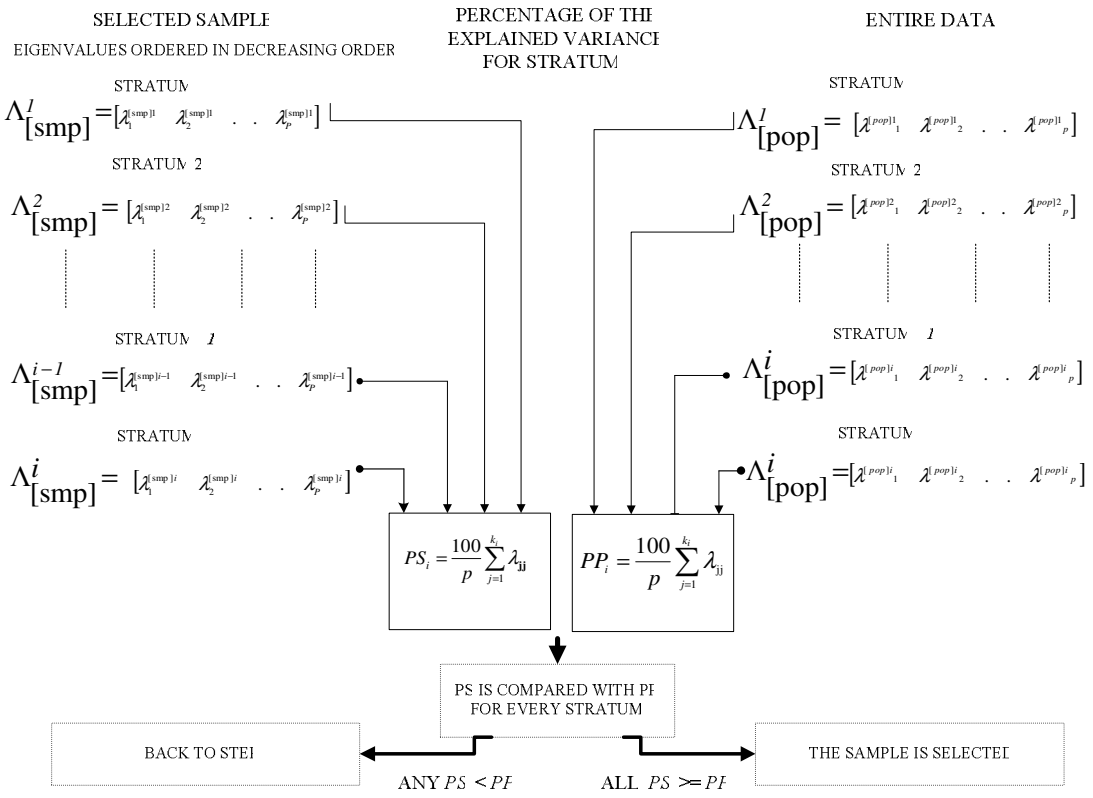


Figure 7. Confirm the selected k_1 principal components

Step 8. The principal component scores for the selected sample $Z_{[smp]}$ are estimated by projecting their original values onto the orthonormal basis. The last principal components k_2 that might be retained are determined based on the guidelines previously described. The correlation matrix between the principal component scores $Z_{[smp]}$ and the dependent variable $Y_{[smp]}$ determine whether or not new principal components will be retained from the minor components. The correlation of the selected k_2 principal components must exceed the given threshold value. If the correlation of the minor components with the dependent variable is not significant, then no minor components are added.

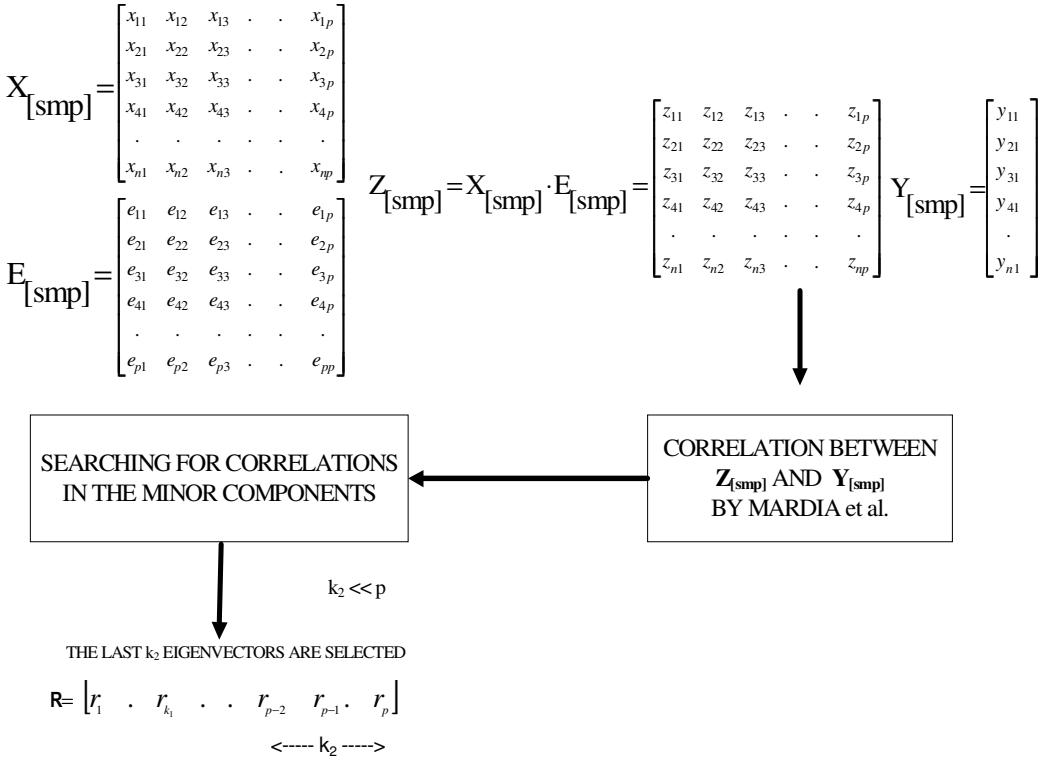


Figure 8. Selection of k_2 minor components

Step 9. The final value of the selected principal components k are the first k_1 of them, which is the maximum of the k_1^i(s) (steps 6 and 7), and the last k_2 principal components (step 8). This final selection is extracted from the entire matrix of principal component scores $Z_{[smp]}$ and they represent the new non-correlated variables explaining a high percentage of the variance of the entire set of the original independent variables.

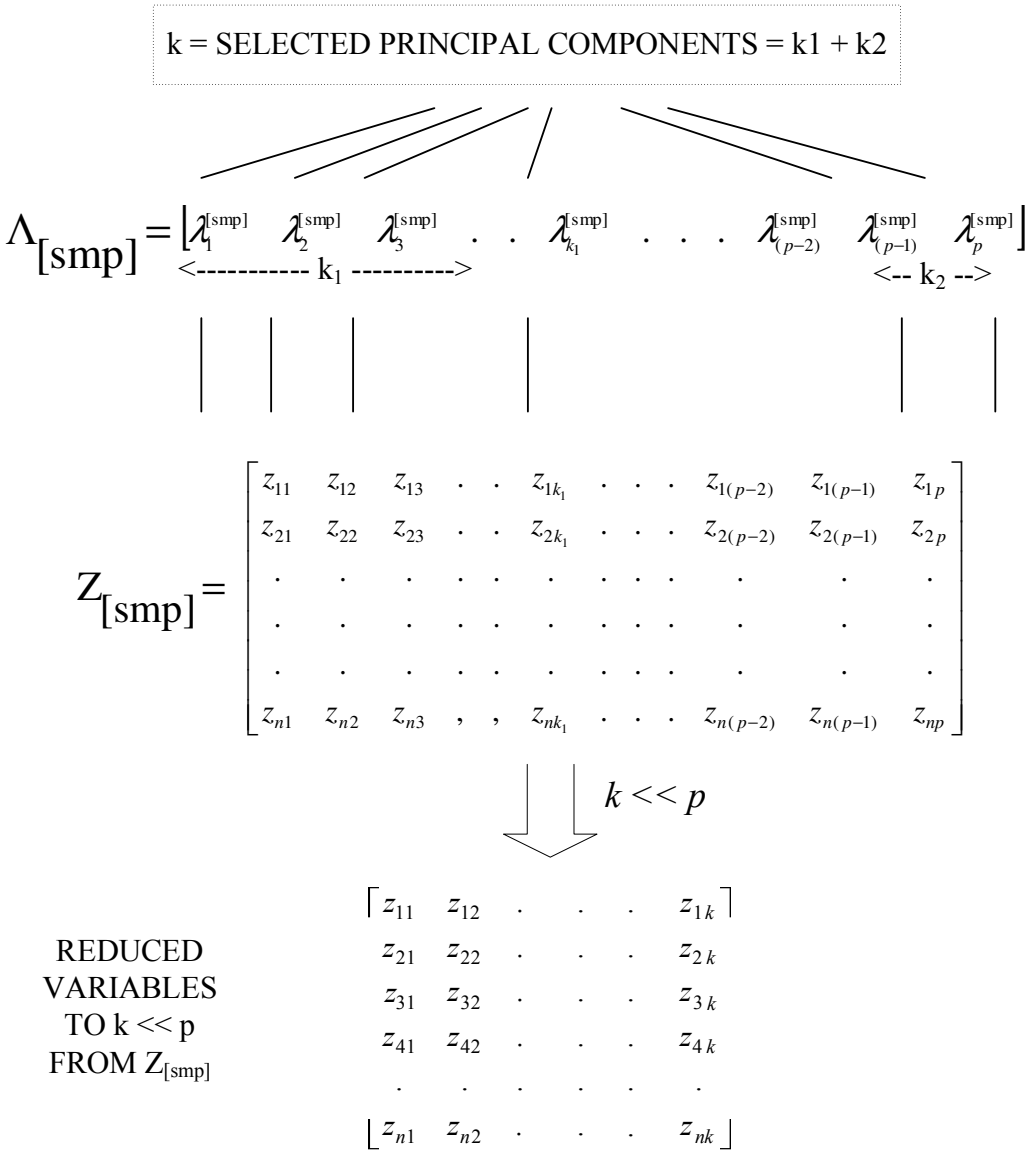


Figure 9. Final selection of the reduced variable set

3. Stratified/PCA and Neural Networks

The reliability of the Stratified/PCA method was evaluated in a manner similar to that used to evaluate the stratified method. We first created an original data set by selecting discrete values from nine independent terms, $4xe^{x/3}$, $e^{-\cos\left(\frac{3}{4}\pi x\right)}$, $\frac{x}{2}$, $\cos\left(\frac{\pi}{2}x\right)$, $\cos\left(\frac{\pi}{4}x\right)$, $x^{1/2}$, $e^{\sin(\pi x)}$, $0.00756x^3$, and $0.169x^2$ and for values of x in the interval $0 \leq x \geq 25$. Each discrete value of x is denoted by x_i where i is an integer between 1 and 7,500, where 7,500 is the number of data points created. The values of the terms were grouped as observations in the vectors $\underline{X}_1$ through $\underline{X}_9$ for every x_i. The summation of the nine terms describes the function shown on Figure 10.

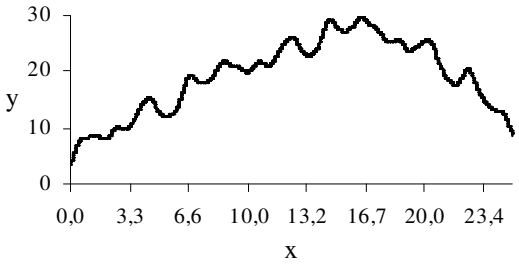


Figure 10. Experimental function

The original data set of 7,500 records was divided into two sets using the Stratified/PCA method. One subset of 85% of the data is used to train neural network models and the rest is used to test the previously trained neural networks. Additionally, two pairs of reduced new subsets, one from the set of 85% of data and the other from the set of 15% of data were created using the stratified method (without PCA) and random selection. Neural networks were trained using those two reduced 85% subset of data. Finally, the reduced sets of data from the data set of 7,500 records were created using separately the three different previously mentioned reduction methods: Stratified/PCA, the stratification, and random selection. Same numbers of set were selected by each method applying seven different sample sizes: 250, 500, 800, 1600, 2,000, 3,000 and 4,500 samples. At this instance, the presence of a reliable number of variables in the reduced data set is a consequence of applying a threshold value of lambda. The stratified/PCA method was

applied with two different values of lambda, 0.7 (suggested by Jolliffe [9]) and 1 (suggested by Kaiser [10]).

Figure 11 shows a guide to construct and generalize neural networks using examples selected by Stratified/PCA. The following steps describe a general guide when Straified/PCA and a Neural Network Software is used. [14].

- Setup the Original Archive recognizing the set of variables for each observation.
- Apply Straified/PCA to the original set using stratification at the first stage. Two sets are extracted. Almost 85% of the original data as a set of example data to apply the whole method. The rest of data (validation data) is used for neural network model validation.
- These two sets are used as input data in the Stratified/PCA process to create the new reduced data set. Two main objectives are reached at this stage: reliable stratified samples of observations and a new set of latent variables with a high percentage of the explained variance for the entire set of variables.
- At this stage the correlation matrix is prepared as the input data to use PCA. The new latent variables are identified and selected for every stratified samples. This process is applied in the same way to the different sample sizes from the examples data. Similarly, the validation data is prepared by reducing the original variables in an equivalent set of latent variables to those created from the examples data.
- These new sets of examples with less variables and observations are used as training and testing data sets in the construction and validation of neural network models. The network architecture was similar for each built model: input nodes, hidden nodes, hidden layers, activation functions, initial weights and output nodes. Additionally, the set of parameters and evaluation measurements were similar for each neural network model: leaning rate, momentum value, backpropagation algorithm, Root Mean Square Error (RMSE), etc.
- Finally, the generalization of each trained and tested model was validated with the validation data set. Outcomes of consistency and reliability were evaluated using the similar method with several replications to different samples of observations at different sizes.

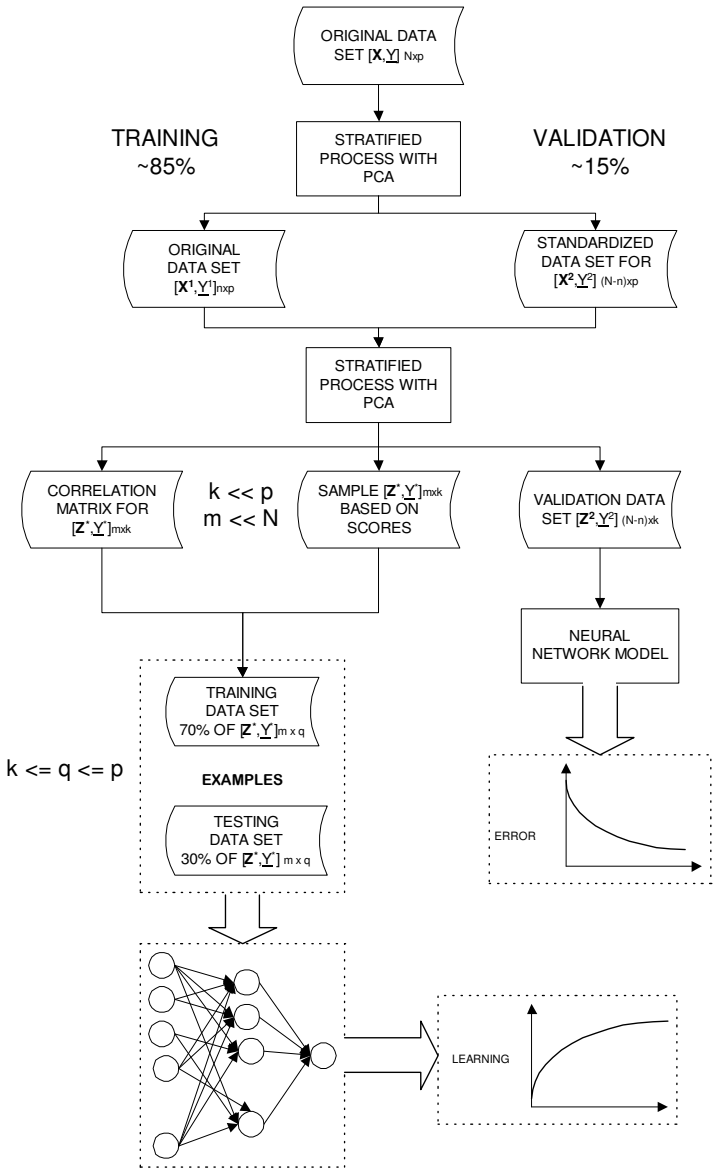


Figure 11. A block diagram to describe the use of the preprocessing method of Stratified/PCA and Neural Networks

In this experimental case, seven different neural networks were trained for each of the three selection methods: random, stratified and stratified/PCA. Seven data sets were prepared using the previous algorithm for each sample size indicated above. Each network was tested by measuring the RMSE value which is a mechanism to identify the ability of

replication of the dependent variable when the observed and computed values are compared.

Table 1. RMSE values of the performance of neural network models tested

Validation using 1500 examples

Data set size	Random	Stratified	Stratified/PCA Lambda > 1	Stratified/PCA Lambda > 0.7
Average	0.62717	0.39482	0.61886	0.56625
Standard deviation	0.28382	0.11100	0.16342	0.12659

Table 1 shows the average and the standard deviation of the RMSE values over the seven models built for each of the sampling methods discussed above. The Stratified/PCA method reduced the nine independent variables in the original data set to five (with lambda set to 1) and seven (with lambda set to 0.7). The table shows that stratification, as was previously shown [4], can create test and training data to consistently build accurate neural networks. Moreover, the size of the training data can be significantly reduced (from 6,375 samples to 250 samples in this case) with the corresponding decrease in training time. The Stratified/PCA method, with lambda set at 0.7, yielded similar results to those of stratification and significantly better results than the random method while still reducing the number of variables from nine to seven.

Figures 12, 13 and 14 show the plot of the dependent variable for the validation data set and the models built with Stratified/PCA data sets of 500 and 800 observations and using the Jolliffe and Kaiser threshold values. Figures for the other five sample sizes show similar results. To facilitate visual comparison, they are shown with the plot of the entire data set (7,500 records) and the test set (1,500 records) in the first line. The plot on the left side corresponds to the training data set using Stratified/PCA methods on both lambda threshold values (1 or 0.7). The plot on the right side shows the prediction curves using the test data set provided by stratified/PCA method for both lambda threshold values.

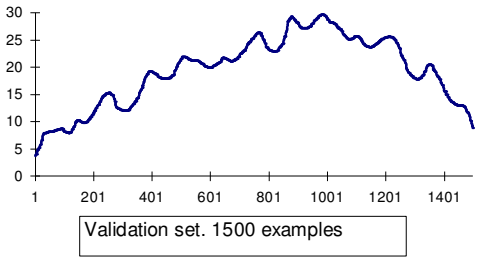


Figure 12. Validation data set of 1,500 observations.

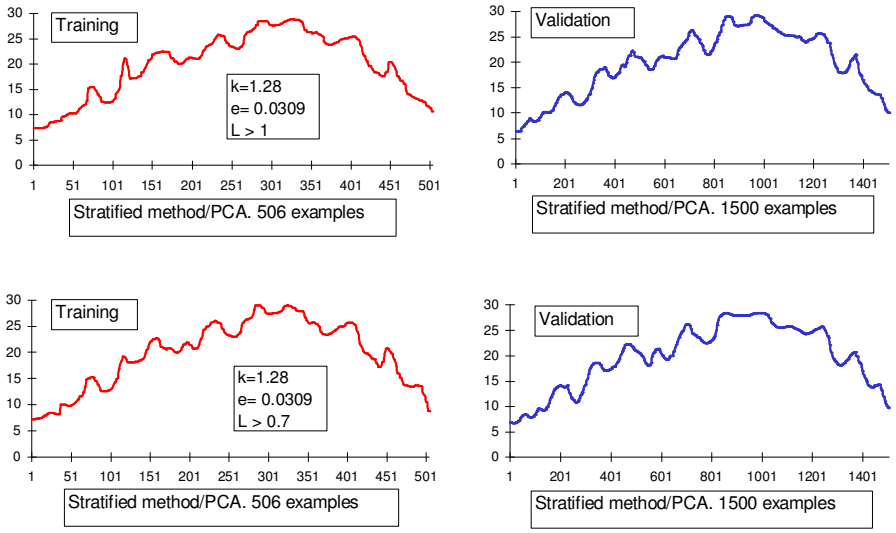


Figure 13. Training and Validation of Neural Networks models built using Straified/PCA with samples of 500 observations.

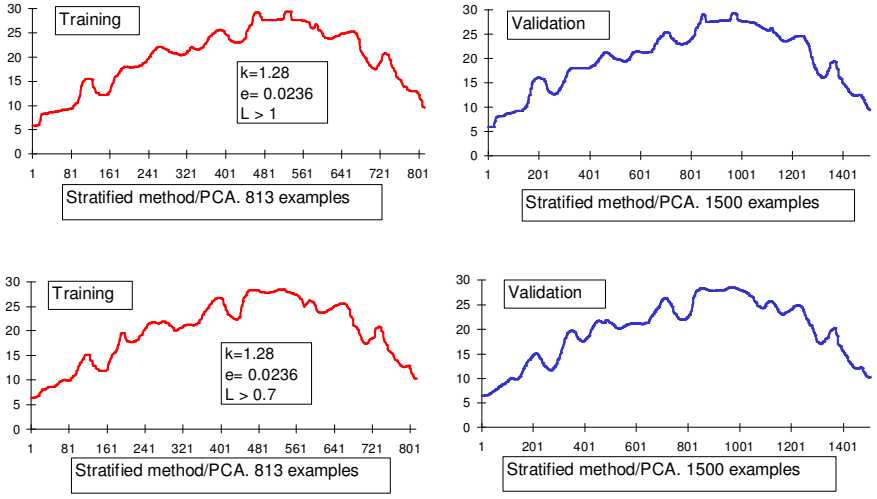


Figure 14. Training and Validation of Neural Networks models built using Straified/PCA with samples of 800 observations.

4. Conclusions

Historical data sets often include a large number of observations and variables that present difficult problems for building neural networks. Large numbers of observations lead to very long training times and large number of variables lead to large network architectures. In addition, neural networks require the selection of training and test sets that are representative of the entire data set. The random method used to either reduce the size of the original data set, or to select training and test sets from the original data can yield neural networks with widely different performances.

The Stratified/PCA method presented here can be used to consistently select samples from a data set that are representative of the entire data set and therefore maintain the original problem characteristics. The method also eliminates unnecessary variables or replaces groups of collinear variables with a smaller set of independent variables. The end result is that by reducing the number of observations and variables in a large data set, the amount of time required to train neural networks is reduced. Also, a reduction in the number of variables leads to networks with simpler architectures. Furthermore, this method can be used to select representative samples to build training and test sets needed to build and test consistent and reliable neural networks.

5. References

[1] Carpenter W. and Hoffman M. E. (1997). "Selecting the architecture of a class of back-propagation neural networks used as approximators," *Artificial Intelligence for Engineering Design. Analysis and Manufacturing*, no. 11, pp. 33-34.

[2] Cheng R. and Davenport T. (1989). "The problem of dimensionality in stratified sampling," *Management Science*, vol. 35. no. 11, pp. 1278-1296.

[3] Cochran W. (1963). *Sampling techniques*, Second edition. John Willey & Sons, Inc., New York.

[4] Colmenares G. (1999). *Reducing samples and variables to Train, Test, and Validate Neural Networks*, Doctoral Dissertation. University of South Florida.

[5] Colmenares G. and Pérez R. (1998). "A Data Reduction Method to Train, Test, and Validate Neural Networks," *Proceedings IEEE Southeastcon '98*, pp. 277-280.

[6] Don S. and Babu J. (1992). "Exploratory Data Analysis using Inductive Partitioning and Regression Trees," *Ind. Eng. Chem. Res.* 31, pp. 1989-1998.

[7] Dong D. and McAvoy T. (1996). "Nonlinear Principal Component Analysis -Based on Principal Curves and Neural Networks," *Computer Chem. Engng.*, vol 20, no. 1, pp. 65-78.

[8] Haykin S.. (1994). *Neural Networks. A comprehensive foundation*, Macmillan College Publishing Company, Inc.
[9] Jolliffe, I. T. (1986). *Principal Component Analysis*, Springer-Verlag.
[10] Kaiser, H. F. (1960). "The application of electronic computers to factor analysis," *Educ. Psychol. Meas.*, vol. 20, pp. 141- 1.
[11] Kramer M. (1991). "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks," *AIChE Journal*, vol. 37, no. 2, pp. 233-243.
[12] Kramer M. (1992). "Autoassociative Neural Networks," *Computer Chem. Engng.*, vol. 16, no. 4, pp. 313-328.
[13] Mardia K. V., Kent J. T. and Bibby J. M. (1979). *Multivariate Analysis*, Academic Press. London.
[14] Neuralware, Inc. (1995) NeuralWorks *Predict: Complete Solution for Neural Data Modeling*.
[15] Sukhatme Pandurang. (1963). *Sampling theory of surveys with applications*, Bangalore press, India.
[16] Tan S. and Mavrovouniotis M. (1995). "Reducing data dimensionality through optimizing neural network inputs," *AIChE Journal*, vol 41, no. 6, pp. 1471-1480.
[17] Witten Ian H. and Eibe Frank. (2000). *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufman Publishers. San Francisco.USA.
[18] Xue Z. Wang. (1999).*Data mining and knowledge discovery for process monitoring and control,* Springer-Verlag. Great Britain.