A RELIABLE METHOD TO REDUCE OBSERVATIONS AND VARIABLES WHEN BUILDING NEURAL NETWORK MODELS

Gerardo Colmenares L[1] and Rafael Pérez[2]
[1]*Instituto de Investigaciones Económicas y Sociales, Universidad de Los Andes.*
La Hechicera, Mérida 5101. Venezuela
gcolmen@faces.ula.ve
colmenar@sunblast.eng.usf.edu
[2]*Department of Computer Science and Engineering, University of South Florida*
4202 East Fowler Avenue. Tampa, Florida 33620-5400.
perez@csee.usf.edu

Abstract: This paper describes a method to reduce the number of observations and variables of large data sets so that reliable neural network models can be built using this data and the time to build these models can be reduced. This method can also be used to select, from an original data set, representative data to train, test, and validate models. This method applies stratification and principal component analysis to select representative observations and to eliminate redundant variables. The performance of neural network models built using reduced data sets provided by this method is very similar to that of neural network models built using the entire data set. The performance is also significantly better and more consistent than that of neural networks built using data sets reduced in a random fashion. A comparison using the stratification method alone and using stratification plus principal component analysis to reduce the data set is also included.

Keywords: Neural network models, stratification, principal component analysis, data reduction, and variable reduction.

1. Introduction

When historical data is collected, a large number of observations are stored. In addition, for each observation, a large number of variables are included. Once the initial cost of setting up the collection mechanism of historical data is incurred, the cost of enlarging the number of observations is relatively small. Moreover, the collection of large number of observations and variables often avoid additional costs if unforeseen use of the data is later identified.

Neural networks is an important technique that can take advantage of the existence of historical data and numerous examples exist of successful applications using this technique [1], [6], [11], [17]. Prediction problems using historical data can be solved with neural networks. Large number of observations and variables in historical data present difficult and interesting problems for neural networks.

To ensure an acceptable neural network model performance, all relevant characteristics of the problem should be represented in the data selected. If the training data is not representative, then the model will perform poorly on testing and/or validation data. On the other hand, a model built with representative training data will perform poorly on testing and/or validation data that are not representative.

Historical data then often leads to two major problems when used for building neural network models. One is that the large number of observation and variables lead to large amount of time when building models. The other problem is how to select representative data sets for training, testing, and validating neural network models.

This paper deals with the design and implementation of a method to reduce observations and variables from large data sets. This method applies stratification and principal component analysis to select representative observations and to eliminate redundant variables Neural network models built using reduced data sets created by this method perform quite similar to models built using all the variables in the original data set and perform better than models built using reduced data sets obtained through random selection.

2. Reducing the Number of Observations

Previous work by the authors has shown that stratification can create reduced data sets that are more reliable than random methods for building neural network models [5]. However, stratification reduces the number of observations based on the dependent variable and ignores the independent variables.

A data set with N observations and M variables might be reduced to one with n observations $(n<<N)$ and M variables. The stratified sampling technique [4], [13], [15], [18], also known as variance reduction technique, aims to achieve this reduction in size but still maintain a data set with similar statistical properties as the original one.

The goal of stratified sampling is to achieve a minimum deviation between the mean value, Y_k, of a principal variable in the original data set, and the mean value, y_k, of the same principal variable in the reduced data set [4]. This variable is called the variable of stratification and in this paper corresponds to the variable to be predicted.

The random method is traditionally the sampling method used to select samples for training, testing and validating neural network models [7], [8]. However, neural network models built using this method show wide variability in their root mean squared error (RMSE) values. The next deals with an evaluation of the results provided by several neural network models built using random training and testing data.

2.1. Evaluation of the Random Method Used to Select Samples

Two experimental data sets were used to evaluate the consistency of the random method. The data sets used in these experiments were built using a function from [1], [2], who used it to analyze the performance of neural networks as a method to approximate the function below.

$$y = 2x + \sin(\pi x) + \sin(2\pi x),$$

where both, x and y, are real numbers. This particular function was selected because it is a non-linear that makes it non-trivial to build models from data derived from this function and it allows identical values in y from different x values. The function was built by selecting discrete values of x in the interval $0 \le x \le 1$. The values of the terms $2x$, $\sin(px)$, and $\sin(2px)$ were grouped as observations in the vectors $\underline{X}_1$, $\underline{X}_2$, and $\underline{X}_3$ respectively, for every x_i. Likewise, the y_i values were grouped in $\underline{Y}$, for every y_i. In general, when the discrete values for $\underline{X}_1$, $\underline{X}_2$, and $\underline{X}_3$ and $\underline{Y}$ are selected, they are referred as observations. After these observations are stored in a data set or file they are referred as records. Figure 1 shows all the observations for y in that interval.

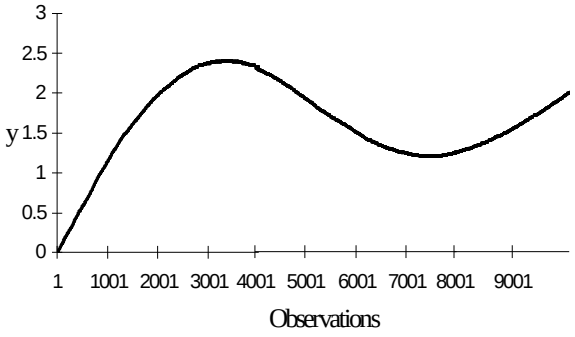


Figure 1. $y = 2x + \sin(px) + \sin(2px)$ in an interval $0 \le x \le 1$

The entire data set of 9996 records corresponding to the above example was separated at random into two new data sets. The first data set of 8505 records (approximately 85% of the entire data set) was used to provide data sets for training. The second data set of 1491 records (15% of the entire data) was used as operational data to validate the models built with the training data mentioned.

After that, ten neural network models were built using different reduced data sets of 303 records. These 303 records were selected at random from the data set of 8505 records. This number of 303 records was the number required by the stratified method to determine the sample size.

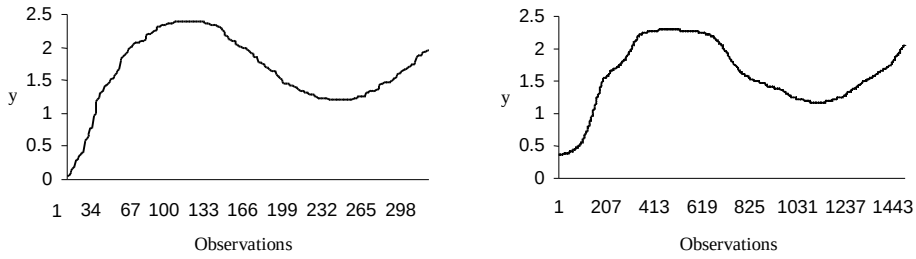
The methodology used to build every neural network model consisted of using 70% of the 303 records for training and the rest for testing. The same computational units (sigmoidal function) as well as the same tuning parameters were used for every model built. These parameters are described in Neuralware Inc. [14]. This methodology is the preferred procedure employed by PREDICT, a neural network package to train, test and validate neural network models developed by Neuralware Inc. This neural network software was used to build all the neural networks models in this research.

After the neural network models were built, they were validated using the data set of 1491 records mentioned above. Table 1 shows the RMSE values for each model using the validation set.

Figure 2 and 3 show plots of the worst and best selection among the ten random samples of 303 records used to build the neural network model. They also show plots of the values predicted by the corresponding models on the validation set. The validation of the two models shows a large difference between them. During the validation of the models, the model built with the best training and testing data of 303 records had an RMSE value of 0.01084 while the model built with the worst training and testing data had an RMSE value of 0.07877, or nearly eight times as much error. The graphs shown in Figure 2, when compared to the original data in Figure 1, give a visual indication of this large difference in performance.

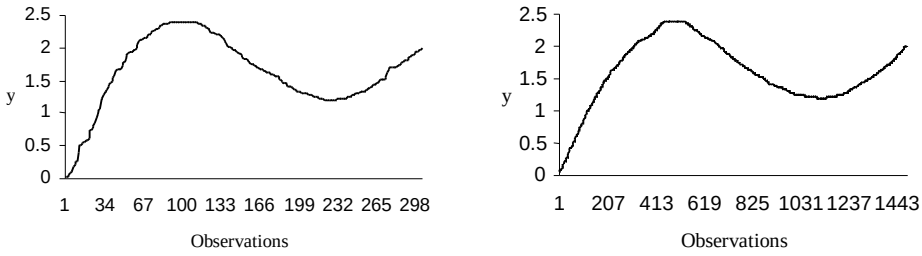
Table 1. RMSE values of the validation of ten neural network models using random samples

Model	RMSE
1	0.04607
2	0.04859
3	0.04794
4	0.07499
5	0.07877
6	0.06124
7	0.01084
8	0.02877
9	0.05637
10	0.03628
Mean	0.04899
Standard deviation	0.02056



Training data Neural network Validation

Figure 2. The worst selection using random samples



Training data Neural network Validation

Figure 3. The best selection using random samples

In summary, the above example showed that when the random method is used to reduce an original data set, the neural network models built from this reduced data set show large variation in performance.

2.2. Stratified Method

Using stratified sampling, sub-samples are selected by a random process carried out independently within each stratum. The number of stratum for a given population is determined by a procedure that continuously adjusts the number of strata until an acceptance criterion is met. This criteria is based on how close we want the variable of stratification mean and variance in the reduced data set to be to those of the original data set [3], [4], and [16]. This technique is described in more detail in Colmenares et al. [5].

The stratification method, in general, consists of two fundamental steps. The first is to determine the number of strata and the second is the selection of observations in every stratum. The next two sections describe both steps.

2.2.1. Selection of the Number of Strata

When the stratified sampling method is applied to data sets consisting of values of independent variables and one dependent variable, the variable of stratification is the dependent variable $\underline{Y}$ [4]. The original data set of $\underline{Y}$ is separated in strata. The number of strata is built using the frequency distribution of the $\underline{Y}$ vector. Figure 4 shows the dependent variable before and after the stratification.

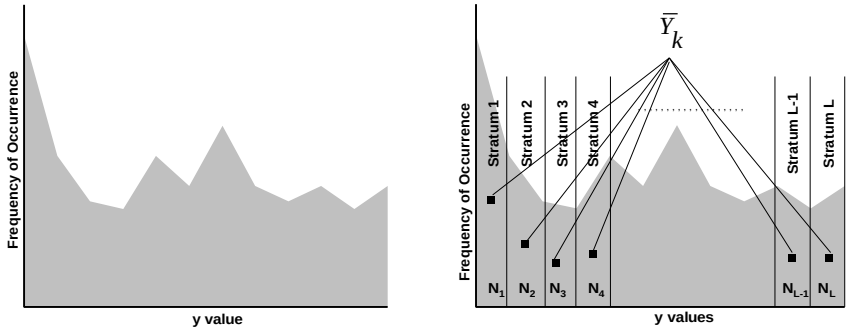


Figure 4. Stratification of the original data

The range of the variable of stratification defined by $\underline{Y}$ determines an initial number of strata. For a given original data set, the first estimation of number of strata L will be determined [18] $L = 1 + 3.3 \cdot \log_{10} N$. This expression leads to indicate a maximum value for the number of strata. For the typical size of data sets between 50,000 and 150,000 records the number of strata would be between 10 and 14.

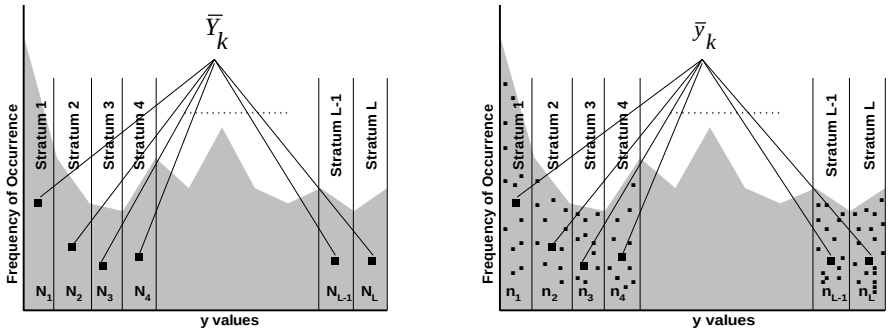
After determining the initial number of strata, the strata are equally spaced along the range of the variable of stratification. Every stratum in $\underline{Y}$ includes a set of observations that determines the size for that stratum. The observations are grouped into the initial number of strata. If any stratum has zero observations, then the number of strata is reduced and a new set is created with corresponding observations in each stratum. This process is repeated until no stratum is empty.

2.2.2. Selection of the Number of Observations

Once the observations in $\underline{Y}$ have been grouped into the final number of strata L as described above, then the reduced data $\underline{Y}'$ is selected by stratification. This method is described in more detail in Colmenares et al. [5].

Stratified sampling separates the variable of stratification $\underline{Y}$ in a set of strata and builds a reduced data set $\underline{Y}'$. The new data set $\underline{Y}'$ corresponds to the stratified sample of n observations where $n << N$. For each value of $\underline{Y}$ selected in the stratified sample, the corresponding values of the independent variables are also selected. Using stratified sampling, observations are independently selected from each stratum by a random process without reposition. If $n_1, n_2, n_3, \ldots n_L$ represent the number of selections made within each stratum, the following condition must be satisfied: sample size is n and it is

given by $n = \sum_{k=1}^{L} n_k$ and strata size in the entire data and the stratified sample are proportional. $\frac{n_k}{n} = \frac{N_k}{N}$.



a) Original data b) Sample

Figure 5. Sample selection of Y and Y' for the entire and reduced data sets

2.3. Application to Neural Networks with Experimental Data

This example consists of the function shown in Figure 1 and the same data set of 9,996 records. The entire data set was separated into two new sets by stratified selection. The first data set with 8,505 records was used to provide reduced data sets for building neural network models. The second data set with 1,491 records was used to validate the models built. Since the comparison was made with reduced data obtained at random, the entire data set was also randomly separated in two data sets with the same previous specifications: 85% for a training data set and 15% for a validation data.

The stratified data set of 8,505 records was reduced to 303 records using also the stratified method. A standard error of 0.0045 and a confidence coefficient of 0.95 were used during the stratified selection process. Ten different reduced data sets of 303 records were selected in this manner. From the randomly selected data set of 8,505 records, 303 records were selected randomly. Again, ten different reduced data sets were selected.

Twenty neural network models were built using the sets of 303 records. The results of training and testing of all these neural network models are shown in Table 2. The RMSE of the models built during training and using the stratification method show a small improvement over the random method. The models built using stratification had an RMSE value of 0.03373 while models built using randomization had an RMSE value of 0.04565.

However, the RMSE values were more stable using stratified sampling than random sampling. Table 2 shows that the standard deviation is significantly smaller (approximately 50% smaller) for the stratified method than the random method. This indicates that stratified samples are a more consistent selection than random samples.

Table 2. Summary of training models using the random and stratified methods

TRAINING OF THE SAMPLES (*)

	Stratified	Random
RMSE	0.03373	0.04565
RMSE (Std. Dev.)	0.00879	0.01998

(*) values are averaged over ten samples.

The two data sets (stratified and random) with 1491 records were used to validate the twenty neural networks models previously built using the reduced data sets. Results of both validation sets were compared with the model built using 8505 records.

Figures 6 (b) and 6 (c) show that the plots of the selected samples using stratified and random method were similar to the plot of the entire data set of 9,996 records (Figure 6(a)). However, the effects of inconsistency were observed on the validation of the neural network models built. As is shown in Table 3, the average RMSE value of 0.03847 of the ten neural network models built using reduced data by stratification (stratified models) was lower than the average RMSE value of 0.04899 of the ten neural network models built using reduced data selected at random (random models). Figures 6(d) and 6(e) show the plots of the estimated values using the validation data sets created by the stratified and random methods. In this case, Figure 6 (e) is more irregular than Figure 6(d) when compared to the plot of the entire data set of 9,996 records in Figure 6(a). Stratified models show to be more accurate than random models with a lower RMSE value.

Table 3. Performance on validation set

	VALIDATION TEST			
	RMSE	RMSE St. Dev.	Max.	Min.
Stratified models	0.03847	0.00009	0.05174	0.02188
Random models	0.04899	0.00042	0.07877	0.01084
All Data model	0.04741			

On the other hand, the standard deviation of the ten stratified models is about 70% less than the standard deviation of the ten random models. Again, stratified models are more consistent than random models. In fact, the variation range is closer in stratified models than random models.

This comparison confirms that the RMSE values on neural network models built with stratified samples were reduced. Consequently, the plot of the predicted values using the stratified method resembles closer the original plot than the plot using the random method.

Stratified Method Random Method

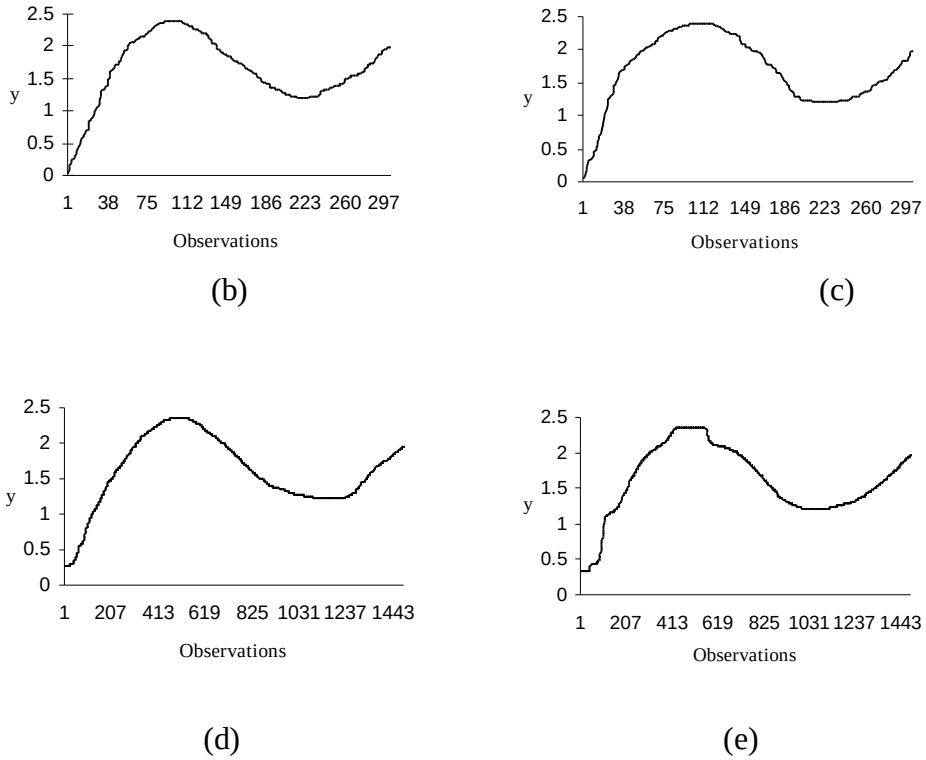


Figure 6. Plot of the dependent variable Y. a) Entire data set (9,996 records). (b and c) Reduced data set (303 records). (d and e) Validation (1,491 records)

3. Reducing the Number of Variables

A large number of independent variables in a large data set can present two major problems. Firstly, too many variables result in long training times when the model is built. Secondly, large number of observations and variables tend to retain redundant information through multicollinearity leading to unreliable models.

Some of the variables present in historical data are needed for some problems and some variables for others. Often, different variables may carry the same information. Collinear independent variables might affect the performance of neural network models. A new variable replacing subsets of variables carrying the same information can lead to more effective and reliable neural network models. Multivariate analysis methods can lead to a reduction of the number of original variables to a new set that is more representative than the original set of variables.

Thus, variable reduction is the second major issue in this research. Methods that deal with variable reduction include the following objectives:

- The redundant information from variables highly correlated can be eliminated without losing information.
- Measuring the variance contribution of the set of variables can drop variables in excess.
- The fewer the number of independent variables, the easier it is to analyze the results.

- Observation coming from one of several well-defined groups can be classified within such groups to minimize misclassification.

Principal Component Analysis and Singular Value Decomposition are the theoretical foundations used to implement an algorithm that leads to a variable reduction method.

By Principal Component Analysis, the entire set of *p* independent variables could be reduced to a new set of fully independent of variables. Some characteristics attributable to PCA are:

- It is used as a dimension reduction technique for linearly mapping high-dimensional data onto a lower dimension with minimal loss of information.
- The new set of variables is much less than the set of the original variables.
- The main input to PCA is the covariance matrix. However, the original variables could be standardized, with zero mean and unit variance, to avoid the effect of the relative variance of the variables. In this case the correlation matrix can be used instead of the covariance matrix. This substitution by the correlation matrix is useful to avoid large variances generated by the values of the original variables, which are not in the same unit [9]. Thus, the correlation matrix is also an input matrix for PCA.
- Eigenvectors of the covariance or correlation matrix are the new axes for the original variables and they are orthonormal.
- The amount of the computed principal components is equal to the amount of original variables.
- Principal component scores are the values of the new variables and they are non-correlated.
- The first principal components explain most of the variance for the entire original variables.
- PCA captures linearity among the variables.
- Correlation between the principal components and the original variables can be explained through the coefficient of determination.

4. Reducing the Number of Observations and Variables

The proposed method of reducing samples and variables analyzes independent and dependent variables as a whole using the stratification method to select samples from the dependent variable and multivariate analysis by principal component analysis (PCA) to reduce the independent variables. For this reason, the name chosen for the proposed method is Stratified/PCA.

Some relevant characteristics of this Stratified/PCA method are:

- Reduce the original independent variables determining how many reduced variables represent the original variables based on the concept of explained variance by the new set of independent variables. Through principal component analysis, the new variables would be able to explain most of the variance of the original independent variables without losing information.
- The reduced independent variables are 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.
- Analyze not only the variances provided by the first principal components, but also the correlation between the minor components and the dependent variable. 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.

- Work simultaneously with the dependent and independent variables based on: (a) statistical data analysis to obtain a reliable reduced data set, and (b) multivariate analysis to reduce the number of the independent variables without losing significant information.

Kaiser H. [10], Jolliffe I. [9], Mardia K. et al. [12], defined useful guidelines to improve the reduction of the number of independent variables. Those guidelines allow the formulation of four criteria to reduce the independent variables based on the principal components.

First, Kaiser H. [10] and Jolliffe I. [9] agree 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.

Second, Jolliffe I. [9] and Mardia K. et al. [12] agree that correlation between the reduced independent variables and the dependent variable is important to determine the final reduced set of independent variables.

Third, as was suggested by Jolliffe I. [9], the explained variance for the reduced independent variables guarantees that these new variables will be representative of the original independent variables.

Finally, Mardia K. et al. [12] reinforces the previous criterion suggesting that the correlation between the principal components and the original independent variables is a recommended way to determine whether or not the new independent variables will be representative.

5. The Stratified/PCA Method

The Stratified/PCA method sequentially examines the selected data in two ways. The first is related to the selection of a stratified sample of the dependent variable. The second is related to the reduction of the original independent variables into a new set of independent variables. The previous four criteria apply to this design.

Step 1. The entire data is separated in strata using the dependent variable as variable of stratification.

Step 2. PCA is applied to every stratum in the entire data. Eigenvectors and eigenvalues are estimated for every stratum from the correlation matrix of the independent variables.

Step 3. After specifying the initial sampling parameters, stratified samples made up of independent and dependent variables are selected from the entire data.

Step 4. After selecting the entire sample, PCA is applied to the entire sample of the independent variables. Using the correlation matrix of the independent variables, the eigenvectors and eigenvalues are estimated.

Step 5. PCA is applied to the independent variables for every stratum of the selected sample obtained in Step 3. Using the correlation matrix of the independent variables, the eigenvectors and eigenvalues are estimated.

Step 6. Based on the first criterion described in Section 4, the eigenvalues and the correlation matrix of the selected sample are evaluated stratum by stratum. The eigenvalues for every stratum greater than a given threshold value allow the selection of the first candidates of principal components.

Step 7. Using the third criterion described in Section 4, the percentage of explained variance given by the eigenvalues of the principal components for every stratum in the selected sample is compared with the percentage of explained variance given by the corresponding eigenvalues of the principal components for every stratum in the entire data. If every percentage in the sample is greater than its corresponding percentages in the entire data, 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.
Step 8. The principal component scores for the selected sample are estimated by projecting their original values onto the orthonormal basis. The last principal components that might be retained are determined based on the second criterion described in Section 4. The correlation matrix between the principal component scores and the dependent variable determine whether or not new principal components will be retained from the minor components. The correlation of the minor principal components must exceed a given threshold value. If the correlation of the minor components with the dependent variable is not significant, then no minor components are selected.
Step 9. The final value of the selected principal components corresponds to the first selection made in step 7 and the last principal components selected in step 8. This final selection is extracted from the entire matrix of principal component scores and they represent the new non-correlated variables explaining a high percentage of the variance of the entire set of the original independent variables.

6. Application of the Stratified/PCA method to Neural Networks with Experimental Data

In this section an experimental case shows the performance of neural network models using data sets selected by the Stratified/PCA method. Their performance is compared to the performance of models built using data sets selected by random and stratified methods without PCA. The sampling parameters and the neural network architecture were kept the same in all of the experiments. The data sets used in these experiments were built selecting discrete values from nine independent terms and for values of x in the interval $0 \leq x \Omega 25$ from the expressions $4xe^{x/3}$, $e^{-\cos\left(\frac{3}{4}\pi x\right)}$, $\frac{x}{2}$, $\cos\left(\frac{\pi}{2}x\right)$, $\cos\left(\frac{\pi}{4}x\right)$, $x^{1/2}$, $e^{\sin(\pi x)}$, $0.00756x^3$, and $0.169x^2$. 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 7.

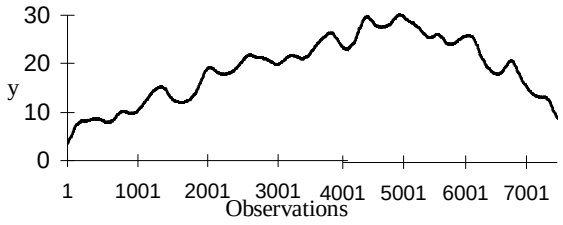


Figure 7. Experimental function

As is observed in Table 4, the average RMSE values for seven trained models are organized in two scenarios: scenario I using stratified method and scenario II using Stratified/PCA method with two values for lambda. Scenario I uses all of the independent variables and scenario II uses five new independent variables created from five principal components for lambda greater than one (Kaiser's threshold value), and seven new independent variables created from seven principal components for lambda greater than 0.7 (Jolliffe's threshold value).

Table 4. Comparison of the RMSE values of the two scenarios for models built

Data set	I	II	
Size	Stratified	Lambda > 1	Lambda > 0.7
Average	0.37642	0.45835	0.38769
Standard deviation	0.08782	0.05405	0.06835

The average RMSE value show that models built with a reduced number of independent variables using the Stratified/PCA method were similar to the models built with all of the independent variables using the stratified method. Both threshold values of lambda in scenario II showed a good performance.

However in scenario II, the Jolliffe's suggestion (lambda > 0.7) was more conservative than the Kaiser's rule (lambda > 1) retaining in the new independent variables more variance of the original independent variables.

Likewise, Jolliffe’ suggestion was more precise than the Kaiser's rule. The average RMSE value of 0.38769 for lambda > 0.7 was smaller than the average RMSE value of 0.45835 for lambda > 1.

Comparing scenario I with scenario II for lambda > 0.7, the average RMSE value of 0.37642 achieved in scenario I and the average RMSE value of 0.38769 achieved in scenario II with lambda greater than 0.7 were very similar. Thus, models built using scenario II with lambda > 0.7 showed that two variables had certain degree of collinearity or redundancy and they were not required for model building.

Three data sets of 1,505 records for the random, stratified and stratified/PCA methods were selected from the entire data set of 7,500 records. The data sets of 1500 records were used as operational data to validate the neural networks models previously built using reduced data sets provided by the three mentioned methods. The performance of the neural network models built for the three methods were compared using their RMSE values.

In Table 5, the average RMSE values of the performance of neural network models built using data sets from the stratified and stratified/PCA methods were better than the average RMSE values of the performance of neural network models built using data sets from the random method. Moreover, stratified and stratified/PCA methods were more consistent than the random method. On average, they had about 50% less standard deviation than random method.

Table 5. RMSE values of the performance of neural network models validated

Validation using 1500 examples				
Data set size			Stratified/PCA	
	Random	Stratified	Lambda > 1	Lambda > 0.7
Average	0.62717	0.39482	0.61886	0.56625
Standard deviation	0.28382	0.11100	0.16342	0.12659

Table 6 includes the correlation between the desired and estimated values of the dependent variable of the models built by the three methods: random, stratified, stratified/PCA based on the Kaiser's rule (Lambda>1), and stratified/PCA based on the Jolliffe's threshold value (Lambda>0.7). As can be seen, the performance of models built using stratified and stratified/PCA methods showed that they were more accurate models than those built using random method. The standard deviation of the correlation values for models built using stratified and stratified/PCA methods is smaller than the standard deviation of the correlation values for models built using random method. The models built using stratified and stratified/PCA methods were more consistent.

Table 6. Correlation between desired and estimated values

Validation using 1500 examples				
Data set size			Stratified/PCA	
	Random	Stratified	Lambda > 1	Lambda > 0.7
Average	0.99437	0.99789	0.99524	0.99623
Standard deviation	0.00551	0.00131	0.00228	0.00186

In summary, in similar training conditions, the Stratified/PCA method with seven independent variables (a reduction of 20% of the entire number of independent variables) and the stratified method produced models with similar RMSE values. Likewise, Stratified/PCA and stratified method were better at generalizing than the random method.

Figures 8 and 9 show the plot of the dependent variable for models built with random, stratified and Stratified/PCA data sets of 250 and 2,000 observations. To facilitate visual comparison, they are shown with the plot of the entire data set (7,500 records) and the validation set (1,500 records) in the first line. The following lines are organized with two plots of the dependent variable. Plot on the left side corresponds to the training data set using random, stratified and stratified/PCA methods. The plot on the right side shows the prediction curves using the validation data set provided by random, stratified and stratified/PCA methods.

7. Conclusions

The Stratified/PCA method yields reduced data sets that can be reliably used to build and test neural network models. From the two cases of variable reduction provided by the stratified/PCA method (lambda > 1 and lambda > 0.7), the selected data sets with fewer variables achieved similar performance as the selected data sets with all the variables provided by the stratified method. It has been previously shown [4] that the stratified method yields reduced data sets that build models as reliable as models built using the entire data set.

- The performance of the models built was similar. The average RMSE value of 0.37642 for the stratified method, 0.38769 for the stratified/PCA method with lambda > 0.7, and 0.45835 for the stratified/PCA method with lambda > 1.

- The standard deviation of the RMSE values was small. 0.08782 for the stratified method, 0.05405 for the stratified/PCA method with lambda > 1, and 0.06835 for the stratified/PCA method with lambda > 0.7.

The stratified/PCA method (lambda > 1 and lambda > 0.7) can provide data sets more consistent and reliable than data sets selected by the random method.

- The correlation between the estimated and desired values of the models built was more stable. The standard deviation was 0.00186 for the stratified/PCA method with lambda > 0.7, 0.00228 for the stratified/PCA method with lambda > 1, and 0.00551 for the random method.
- The performance of the models built was more consistent with a low standard deviation (about 50%) of the RMSE values. 0.06835 for the stratified/PCA method with lambda > 0.7, 0.05405 for the stratified/PCA method with lambda > 1, and 0.08782 for the stratified method.

In addition to creating reliable data sets, the stratified/PCA method obtained a substantial time reduction. The stratified/PCA method was able to effectively reduce observations and variables from large data sets. The large data set with nine independent variables and 7,500 observations (68,500 data points) could be reduced as few as 250 observations and 5 independent variables (1250 data points).

Therefore, the neural network architecture of models built using reduced data sets provided by the stratified/PCA method were evidently reduced since the number of observations and the number of variables used to build those models were reduced. Thus, the Stratified/PCA method is a viable alternative to select data to train and test neural network models when using large data sets.

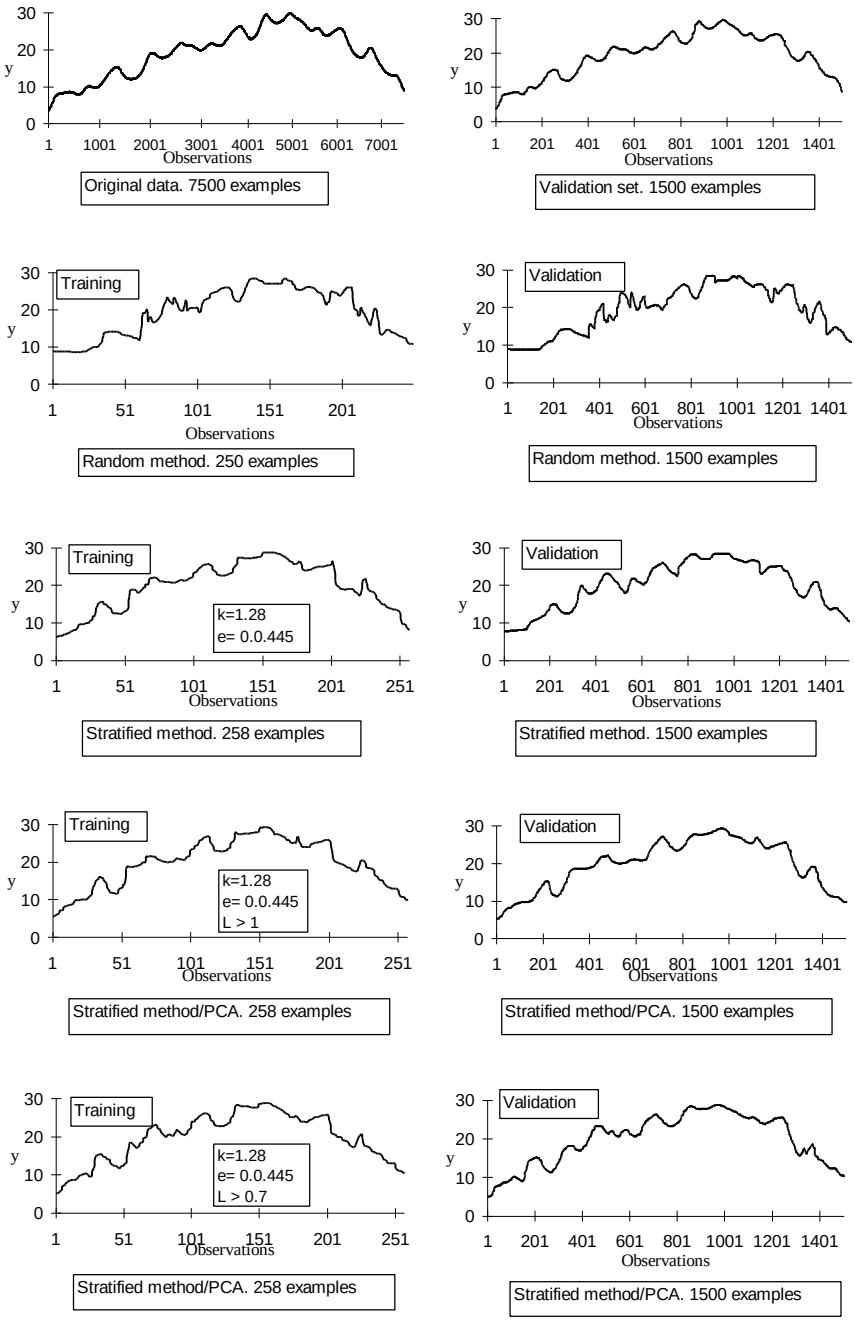


Figure 8. Training and validation for samples of 250 examples

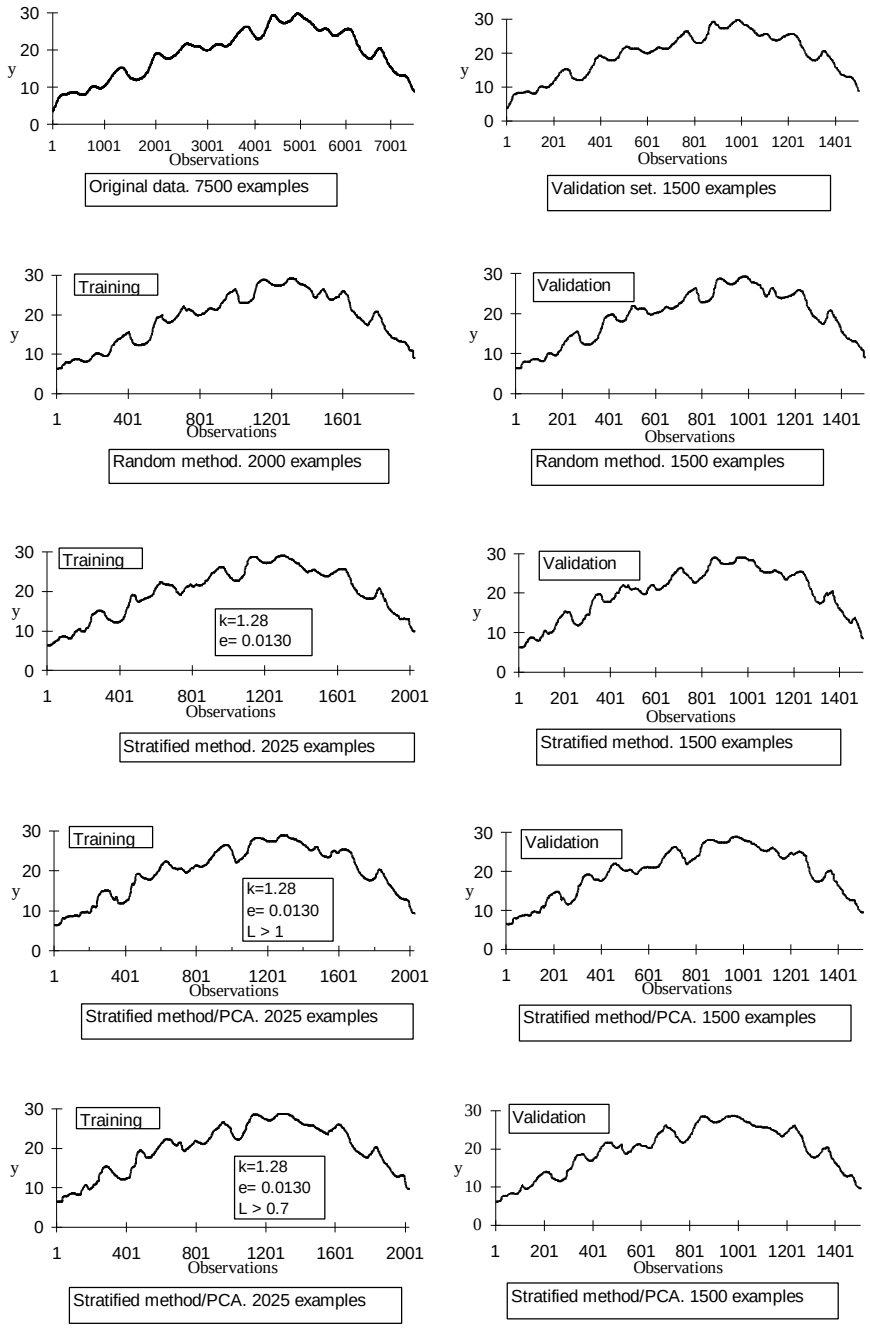


Figure 9. Training and validation for samples of 2,000 examples

8. 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] 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.

[3] Cheng Russel and Davenport Teresa. (1989). The problem of dimensionality in stratified sampling. Management Science, vol. 35. no. 11, pp. 1278-1296.
[4] Cochran William. (1963). Sampling techniques. Second edition. John Willey & Sons, Inc., New York.
[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] Haykin Simon. (1994). Neural Networks. A comprehensive foundation. Macmillan College Publishing Company, Inc.
[8] Hecht-Nielsen Robert. (1990). Neurocomputing. Addison-Wesley 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. 141-151.
[11] Majcen Nineta, Rager-Kanduc Karmen, Novic Marjana, and Zupan Jure. (1995). Modeling of Property Prediction from Multicomponent Analytical Data Using Different Neural Networks.. Anal. Chem., vol. 67, pp. 2154-2161.
[12] Mardia K. V., Kent J. T. and Bibby J. M. (1979). Multivariate Analysis. Academic Press. London.
[13] Milton Sande. (1986). A sample size formula for multiple regression studies. Public American Association for Public Opinion Researc. Opinion Quarterly, vol. 50, pp. 112-118. University of Chicago press.
[14] Neuralware, Inc. (1995) NeuralWorks Predict: Complete Solution for Neural Data Modeling.
[15] Pelletier Bertrand and Chanut Jean-Pierre. (1991). Strate: A microcomputer program for designing optimal stratified sampling in the marine environment by dynamic programming-II. Program and example. Computers & Geosciences, vol. 17, no. 2, pp. 179-196
[16] Smith Murray. (1993). Neural Networks for statistical modeling. Van Nostrand Reinhold., New York.
[17] Sovan Lek, Delacoste Marc, Baran Philippe, Dimopoulos Ioannis, Lauga Jacques, and Aulagnier Stéphane. (1996). Application of neural networks to modeling nonlinear relationships in ecology. Elsevier, Ecological Modeling, pp. 39-54.
[18] Sukhatme Pandurang. (1963). Sampling theory of surveys with applications. Bangalore press, India.