Stratified/PCA: un método de preprocesamiento de datos y variables para la construcción de modelos de redes neuronales

Gerardo A. Colmenares L.[*]
Instituto de Investigaciones
Económicas y Sociales

Resumen. Este artículo describe un método para reducir observaciones y variables contenidas en grandes archivos de datos a nuevos conjuntos de datos cuyo tamaño es mucho más reducido. Estos nuevos conjuntos son confiables para mejorar tanto en calidad como en tiempo, la construcción de los modelos de redes neuronales. El método, Stratified/PCA utiliza la técnica de muestreo estratificado y análisis de componentes principales para reducir eficientemente la cantidad original de datos y variables, manteniendo a su vez, la mayor cantidad de información presente en los datos originales. El desempeño de las redes neuronales construidas con estos conjuntos reducidos de datos ha sido bastante similar al obtenido con el conjunto original de los datos. Por otro lado, la técnica de estratificación usada en Stratified/PCA mostró ser más consistente en la selección de los datos que la técnica aleatoria (convencionalmente más utilizada). Se realizaron comparaciones de reducción de datos usando solamente estratificación y el método Stratified/PCA.

Palabras claves: Modelos de redes neuronales, estratificación, análisis de componentes principales, reducción de datos, reducción de variables, preprocesamiento de datos.

[*] *Ph. D. en Ingeniería, University of South Florida, USA. gcolmen@ula.ve*

0 Introducción

En la actualidad se suele disponer de grandes volúmenes de datos almacenados en archivos históricos que incluyen en sus registros u observaciones grandes cantidades de variables. Por citar algunos ejemplos de grandes volúmenes de datos, se tienen los censos poblacionales, exploraciones petrolíferas, adquisición de datos reales en procesos automatizados de control, estudios urbanos, índices de precios, encuestas de hogares, entre otros. El costo de inferir sobre estos datos históricos y actualizarlos es relativamente bajo cuando es comparado con los costos de recopilación. Más aún, el uso posterior de estos datos históricos para actualizar o dar origen a nueva información (representada por nuevas variables u observaciones) podría eliminar costos adicionales de nuevas recopilaciones de datos.

Las Redes Neuronales Artificiales (RNA) emergen como técnica alternativa para el aprovechamiento de esas fuentes históricas de datos. Sin embargo, dado al gran tamaño de estas últimas, la selección de los datos requeridos (observaciones y variables) para la construcción de modelos confiables de RNA, requieren un alto costo de tiempo de computadora. Estos problemas se han mitigado mediante el uso individual o combinado de Métodos Exploratorios de Datos (MED) y RNA. A este proceso de preparación previa de los datos para ser usados en la construcción de modelos RNA para pronóstico o clasificación, también se le conoce como métodos de preprocesamiento de datos. Existen numerosos ejemplos que han mostrado resultados exitosos[1]. Se ha trabajado con MED basado en métodos estadísticos que se aplican a la reducción de datos (observaciones y variables) como un mecanismo de preprocesamiento antes de ser usado en la construcción de los modelos RNA. Asimismo, se ha determinado que manteniendo cierta rigurosidad científica en la selección de los datos, los

[1] Para más información, el lector puede remitirse a Carpenter (1997), Collantes y otros (2002), Don y Babu (1992), Majcen y otros (1995), Sovan y otros (1996).

modelos RNA podrían tener un mejor desempeño en sus pronósticos (Colmenares y Pérez, 1998; Don y Babu, 1992). En ese sentido, las observaciones y variables presentes en esos archivos históricos representan una fuente interesante y compleja de real investigación y aplicación para el desarrollo de técnicas de recuperación de datos mediante los modelos RNA y MED. Por otro lado, existen técnicas ad-hoc conocidas como técnicas de minería de datos, que permiten que a un conjunto de datos existente se le pueda extraer las variables de interés, transformar las variables originales a un conjunto de nuevas variables, o clasificarlas de acuerdo a un comportamiento o patrón previamente identificado. Algunas de estas técnicas no son más que métodos de exploración de datos, algoritmos de redes neuronales o una combinación de ambos (Xue, 1999). Para asegurar un pronóstico aceptable mediante RNA, todas las características que componen los datos históricos, deben estar representadas en los datos seleccionados para la construcción de la red neuronal. Adicionalmente, las observaciones utilizadas para hacer inferencia deben guardar relación con los datos históricos (Collantes y otros, 2002; Colmenares, 1999). En otras palabras, se deben incluir observaciones de todo el espacio que conforma el dominio del problema. Si los datos de entrenamiento de una RNA no son representativos para la construcción del modelo, entonces las inferencias serán de pronóstico débil. Por otro lado, un modelo construido con datos representativos no garantiza ser de pronóstico confiable si los datos de prueba, de igual modo, no son representativos.

Este artículo presenta un algoritmo de un método para reducir observaciones y variables de grandes conjuntos de datos. Es un diseño que aplica sincronizadamente las técnicas de muestreo estratificado y análisis de componentes principales para seleccionar observaciones representativas y eliminar las variables que aporten poca o ninguna información. Al método se le ha denominado Stratified/PCA (Colmenares y Pérez, 1998; Colmenares, 1999). Los ensayos realizados con datos experimentales para construir modelos RNA utilizando este tipo de preprocesamiento, han arrojado resultados similares a

los modelos RNA construidos con todos los datos originales. Incluso, los resultados han sido más halagadores que los obtenidos en modelos construidos por la vía convencional de realizar selecciones aleatorias de muestras de datos.

1 Método de muestreo estratificado

Este método consiste de una muestra de observaciones agrupadas en estratos, extraídas en porciones mediante un proceso aleatorio sin reposición de un conjunto de observaciones originales que igualmente están organizadas por estratos (tal como se hizo en la muestra). El criterio de selección se basa fundamentalmente en qué tan cerca se desean los valores de la media y la varianza de la variable estratificada en la muestra seleccionada, con respecto a los valores de la media y la varianza para la misma variable en el conjunto original de datos (Cheng y Daveport, 1989; Cochran, 1963). El número de estratos es seleccionado de acuerdo a un criterio de aceptación que depende de la media y varianza, tanto de la muestra como del conjunto de todos los datos. Esta técnica está descrita en Colmenares y Pérez (1998).

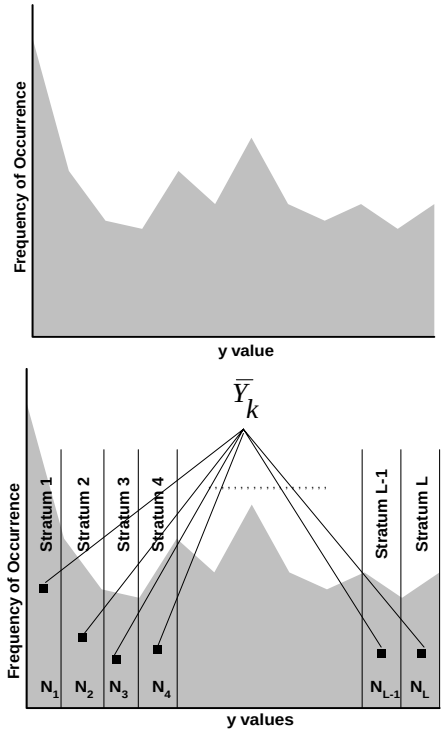
En general, la estratificación consta de dos pasos: primero, determinar el número de estratos y segundo, seleccionar de la muestra observaciones en cada estrato. Las próximas dos secciones describirán con más detalle ambos pasos.

Selección del número de estratos

La aplicación del método estratificado sobre un conjunto de observaciones ameritan de, por lo menos, una variable de estratificación o principal. En el caso de los archivos históricos sometidos a un proceso de reducción, se requiere la presencia de una variable dependiente y que el resto de ellas sean independientes. Los valores de la variable dependiente o de estratificación y pueden ser separados en estratos

construidos a partir de la distribución de frecuencia de la misma. El gráfico 1 ejemplifica la distribución de frecuencia de una variable dependiente y antes y después de la estratificación.

Gráfico 1
Proceso de estratificación de los datos originales



La primera estimación de la cantidad de estratos ***L*** en que será separado el conjunto original de ***N*** observaciones (Sukhatme, 1963) está dado por: $L = 1 + 3.3 \cdot \log_{10} N$. Para un tamaño entre 50.000 y 150.000 observaciones, el número de

estratos sería de 10 a 14. Este procedimiento inicial equivale a una distribución de frecuencia de ***N*** observaciones y ***L*** clases.

Después de determinar la cantidad inicial de ***L*** estratos, las ***N*** observaciones quedarían agrupadas en los estratos que representan las clases igualmente espaciadas a lo largo del rango de distribución de frecuencia de la variable de estratificación. Cada estrato en ***y*** incluye un conjunto de observaciones que definen el tamaño de ese estrato, equivalente al tamaño de la clase. Si alguno de los estratos no incluye observaciones, entonces se reduce la cantidad de estratos originalmente establecida y se reagrupan las ***N*** observaciones en un nuevo conjunto de estratos. Este proceso se repetiría hasta que ningún estrato permanezca vacío (Cochran, 1963; Pelletier y Chanut, 1991; Sukhatme, 1963).

1.2 Selección de la muestra de observaciones en los estratos

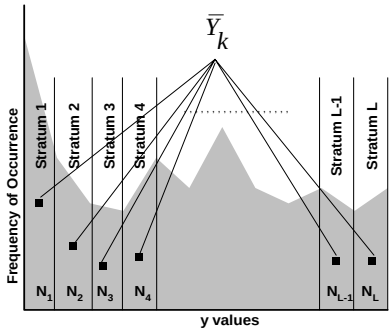
Una vez que las ***N*** observaciones de la variable ***y*** hayan sido agrupadas en ***L*** estratos, entonces el conjunto reducido o muestra de *n* observaciones ***y**** es seleccionado[2] a partir de ***y***. El nuevo conjunto de datos ***y**** se corresponde con una muestra estratificada de *n* observaciones donde $n << N$.

Cada valor de ***y**** es seleccionado mediante un proceso aleatorio y sin reposición. A cada valor seleccionado en ***y**** se le asocia sus correspondientes valores de las variables independientes ***x**.* En ese sentido, se forman matrices de observaciones [**y*** **x***] por cada estrato. De este modo, existirán *L* submuestras [**y*** **x***] con tamaños n_1, n_2, n_3, ... n_L que representan el número de selecciones realizadas en cada estrato y que deben cumplir las siguientes dos condiciones: **a)** el tamaño **n** de la muestra estratificada debe estar dada por

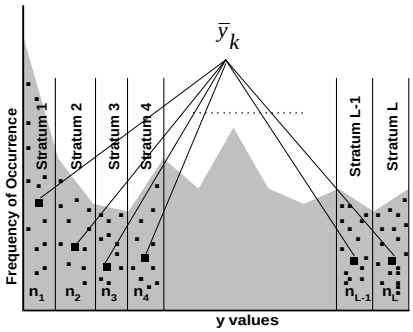
[2] Este algoritmo de selección está descrito en Colmenares (1998).

$n = \sum_{k=1}^{L} n_k$; **b)** el tamaño de los estratos en el conjunto original de datos debe ser proporcional al tamaño de los estratos en la muestra estratificada seleccionada: $\frac{n_k}{n} = \frac{N_k}{N}$. En el gráfico 2, se puede observar la estratificación tanto del conjunto original como de la muestra seleccionada.

Gráfico 2
Estratificación en el conjunto original de observaciones y muestra estratificada seleccionada



a) Conjunto original de datos



b) Muestra estratificada seleccionada

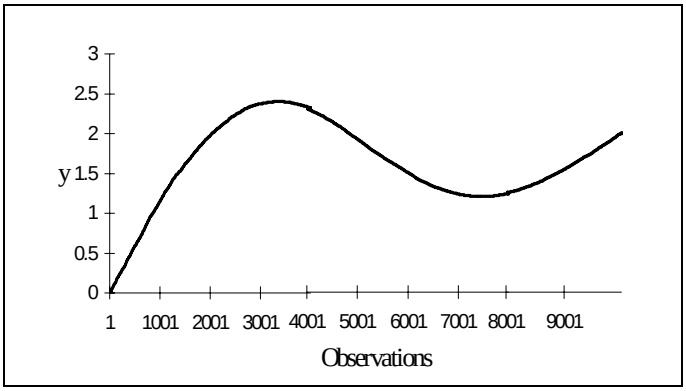
2 Aplicación del método estratificado a las RNA con datos experimentales

El conjunto de datos empleado en este experimento fue obtenido de la función $y = 2x + \text{sen}(\pi x) + \text{sen}(2\pi x)$ ya previamente utilizada para analizar el desempeño de una RNA como método de aproximación de funciones (Carpenter y Hoffman, 1997).

Esta función tiene la característica de ser no lineal y no trivial para construir modelos y permite valores idénticos de la variable ***y*** para diferentes valores de la variable ***x***, donde, ambos son variables con valores continuos. La función se construyó seleccionando valores de ***x*** ordenados crecientemente en el intervalo [0,1], hasta un máximo de 9.996 observaciones. La figura 3 muestra las observaciones obtenidas de ***y*** a partir de valores de ***x*** en el intervalo y función indicados.

A los fines de configurar un conjunto de datos con múltiples variables independientes ***x*** y una variable dependiente **y,** se originaron tres vectores de variables independientes $\mathbf{x_1}$, $\mathbf{x_2}$ y $\mathbf{x_3}$ para que contengan los valores obtenidos al sustituir cada valor dado a ***x*** entre 0 y 1 en *2x*, $\text{sen}(\pi x)$ y $\text{sen}(2\pi x)$ respectivamente (véase el gráfico 3). Los tres vectores se agruparon en $\mathbf{y = x_1 + x_2 + x_3}$, obteniéndose el vector de valores para la variable dependiente **y**. Así, se crea la matriz [$\mathbf{y\ x_1\ x_2\ x_3}$] de 9.996 observaciones, donde cada fila [$\mathbf{y\ x_1\ x_2\ x_3}$] representa una observación, **y** es la variable dependiente y las **x(s)** son las variables independientes.

Gráfico 3
$y = 2x + \text{sen}(\pi x) + \text{sen}(2\pi x)$ **en el intervalo** $0 \leq x \leq 1$



Para la preparación de los conjuntos de datos empleados en la construcción de los modelos RNA se fijaron un conjunto de premisas que fueron válidas para los dos criterios de selección, estratificado y aleatorio, a saber:

- El conjunto total de 9.996 observaciones fue separado en dos conjuntos nuevos mediante ambos métodos de selección (aleatorio y estratificado): un conjunto de aproximadamente 85% de observaciones que sirvieron de fuente de datos para la construcción de las RNA(s); y el otro 15% para inferir basado en los modelos construidos con las muestras seleccionadas del conjunto anterior.
- Del conjunto de observaciones correspondientes al 85%, se generaron un grupo de muestras aleatorias y estratificadas de igual tamaño (303 observaciones). Estas muestras fueron usadas para construir los modelos RNA que permitieron evaluar la eficiencia de la reducción de datos aplicando ambas técnicas.
- Se construyeron modelos RNA usando los conjuntos de 303 observaciones. Estas observaciones estaban preparadas por tres entradas (las variables independientes $\mathbf{x_1}$, $\mathbf{x_2}$ y $\mathbf{x_3}$) y

por una salida (la variable dependiente **y**). Se estableció un diseño fijo de arquitectura de red para las RNA(s) con igual topología para cada caso (igual cantidad de nodos de entrada y de salida, una sola capa de nodos ocultos, función de activación sigmoidal) e iguales parámetros iniciales. Estos parámetros están descritos en Neuralware Inc. (1995). Todos los modelos fueron procesados usando el programa producto PREDICT, suministrado por Neuralware. De acuerdo a PREDICT, 70% de las observaciones permitieron el entrenamiento de la red neuronal y el restante 30% la prueba de cada una de ellas.

- Se aplicó el mismo algoritmo de entrenamiento: supervisado usando backpropagation (Haykin,1994). Para evaluar el nivel de entrenamiento alcanzado en cada modelo RNA, se empleó el mismo criterio de convergencia medido a través del RMSE (error cuadrático medio).
- Para determinar la capacidad predictiva a través del nivel de generalización alcanzado por los modelos construidos, se utilizaron los dos conjuntos de observaciones provistos por el 15% del conjunto original que previamente fue elegido mediante ambos procesos de selección. Estas observaciones no fueron utilizadas en la construcción de los modelos RNA para así medir la capacidad de generalización. Se evaluó la consistencia de ambos métodos en la selección de muestras para entrenamiento y pruebas mediante la dispersión de los valores RMSE(s) obtenidos (aportados por cada modelo entrenado y verificado).

La tabla 1 muestra los resultados obtenidos durante los entrenamientos efectuados con los conjuntos de observaciones seleccionadas aleatoriamente. Se observa la alta dispersión en los resultados a través de los valores RMSE(s). El mejor conjunto de observaciones para entrenamiento fue el modelo 7 y el peor de ellos fue el modelo 5, que obtuvo un valor de RMSE casi ocho veces más alto que el obtenido por el modelo 7. Estos resultados muestran una alta dispersión y en consecuencia son inconsistentes, sin garantía de obtener pronósticos confiables.

Tabla 1
RMSE (s) de diez modelos RNA entrenados usando muestras aleatorias

1	2	3	4	5	6
0,04607	0,04859	0,04794	0,07499	0,07877	0,06124

7	8	9	10	Media	Desv. Est.
0,01084	0,02877	0,05637	0,03628	0,04899	0,01950

Fuente: Cálculos propios.

Del mismo modo diez nuevos conjuntos de 303 observaciones fueron seleccionados mediante el muestreo estratificado. Se construyeron los diez modelos RNA y los resultados promediados del entrenamiento y pruebas se muestran en la tabla 2.

Los valores RMSE(s) de los modelos construidos usando muestras estratificadas, mostraron ser mejores que los construidos usando muestras aleatorias. Casi 1.5 veces mejor fue el resultado de los valores RMSE(s) promediado sobre diez modelos escrutados en ambos casos.

Por otro lado, en lo que respecta a consistencia del método, la dispersión de los RMSE(s) calculados usando las muestras estratificadas, resultó ser más pequeña y estable. Aproximadamente 50% inferior a la obtenida en el escrutinio de los modelos RNA usando muestras aleatorias. Esto verifica que el método de selección de conjuntos de entrenamiento para los modelos basado en estratificación es más consistente en su construcción y capacidad de pronóstico que el método aleatorio el cual ha sido empleado regularmente en la selección de observaciones.

Tabla 2
Resumen de la construcción de modelos RNA usando los métodos estratificado y aleatorio

	Estratificado	Aleatorio
RMSE (*)	0,03373	0,04565
RMSE(Desv.Estandar)	0,00879	0,01998

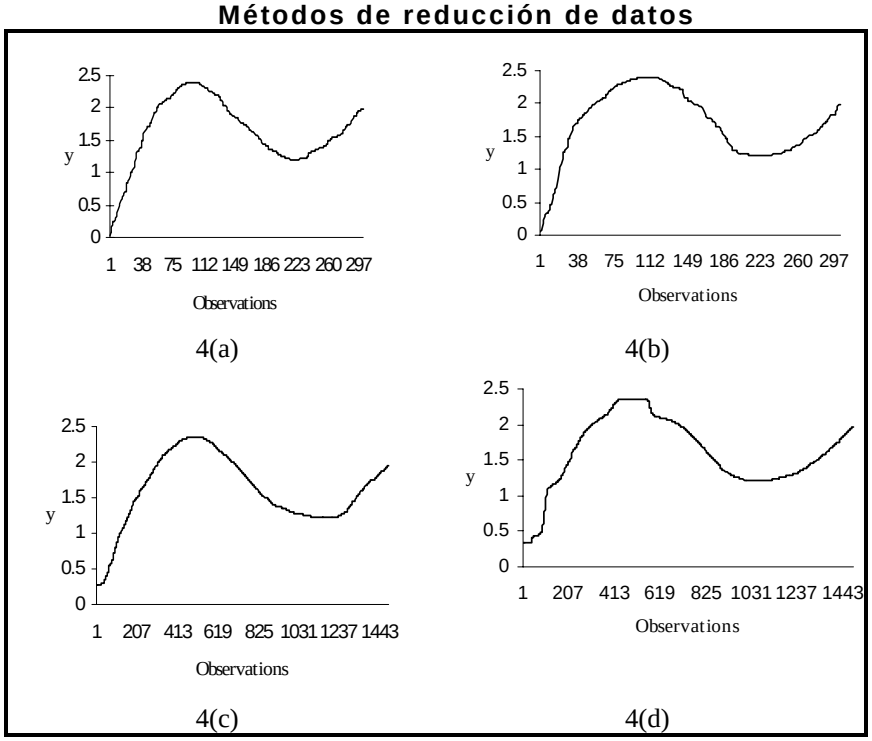
Fuente: Cálculos propios.
(*) Valores promediados en diez muestras.

El gráfico 4 muestra los resultados del mejor de los experimentos de las pruebas realizadas, usando la selecciones estratificada y la aleatoria. Adicionalmente, muestra la mejor generalización de los modelos RNA, al usar las observaciones de prueba correspondientes al 15% del conjunto original.

Los gráficos 4(a) y 4(b) muestran que las selecciones de las muestras de 303 observaciones realizadas por ambos métodos fue bastante similar cuando es comparado con el gráfico de la figura 3 que se corresponde con las 9.996 observaciones. Sin embargo, los efectos de la inconsistencia de la selección aleatoria se notaron cuando se construyeron los modelos RNA. Si se observa la tabla 3, al comparar los resultados de los valores promedio RMSE(s) obtenidos durante la verificación de los modelos RNA construidos con los diez conjuntos de datos experimentales, se nota que el RMSE haciendo estratificación (0,03847) resultó mejor que haciendo selección aleatoria (0,04899) e inclusive, que con todo el conjunto de datos (0,04741). Los gráficos 4(c) y 4(d) muestran los gráficos originados por la series de 1.491 observaciones escogidas mediante ambos métodos cuando fueron usados para inferir sobre los modelos construidos previamente. El resultado que muestra la figura 4(c) es más pobre que el mostrado por la figura 4(d) cuando son comparados con la masa total de observaciones representados en la figura 3. Más aún, de acuerdo al valor del RMSE entre los modelos estratificados y todos los datos, la estratificación resultó ser en un 18% más predictiva. En definitiva, los modelos RNA construidos con conjuntos de observaciones seleccionadas mediante

estratificación (modelos estratificados) mostraron mejor generalización que los modelos construidos usando conjunto de datos extraídos aleatoriamente (modelos aleatorios).

Gráfico 4
Métodos de reducción de datos



Nota: (a y b): selección de los conjuntos reducidos (303) y **(c y d)**: pruebas con 1.491 observaciones nuevas realizadas a los modelos RNA construidos con 303 observaciones

Tabla 3
Valores RMSE promedio y su dispersión medidos durante la verificación de los modelos RNA con 1.491 nuevas observaciones

	RMSE	RMSE Desv. estándar	Máximo	Mínimo
Modelos estratificados	0,03847	0,00009	0,05174	0,02188
Modelos aleatorios	0,04899	0,00042	0,07877	0,01084
Todos los datos	0,4741			

Fuente: Cálculos propios.

Por otro lado, la desviación estándar evaluada sobre los diez modelos estratificados, resultó ser alrededor de 70% inferior que la obtenida en los modelos aleatorios. Se prueba, como en efecto se observa, que la estratificación como técnica de reducción de datos, es más consistente que selección aleatoria (que ha sido la comúnmente utilizada).

3 Reducción del número de variables

La presencia de gran cantidad de variables independientes en la construcción y generalización de modelos RNA tiene dos grandes desventajas: por un lado, el tiempo empleado en entrenamiento de las redes puede ser bastante largo y por el otro, los grandes conjuntos de datos y variables pueden tender a retener información redundante y así conducir a modelos no confiables.

A menudo las variables contienen la misma información o puede estar relacionada entre ellas. Este tipo de variables colineales pueden afectar el buen desempeño de las redes neuronales. Una nueva variable que las reemplace, hace que la red neuronal sea más efectiva y más confiable tanto en tiempo como en resultados. Los métodos estadísticos multivariantes contribuyen a resolver este tipo de problemas mediante técnicas lineales de transformación de las variables originales a nuevos

conjuntos reducidos de variables, destacándose en ellas dos cualidades: total independencia entre ellas y más representatividad.

Los métodos de reducción de variables persiguen los siguientes objetivos:

- La información redundante entre variables altamente correlacionadas pueden ser eliminadas sin arriesgar perder valiosa información.
- La contribución de la varianza debe ser alta en el conjunto de las nuevas variables.
- Menos variables independientes hacen más fácil el análisis de los resultados y simplifica la construcción de los modelos RNA.

Los métodos lineales de Análisis de Componentes Principales (ACP) y el método de Descomposición del Valor Singular (DVS) permiten mediante técnicas de ortogonalización, la reducción del conjunto original de variables, eliminando en lo posible aquellas que presenten cierta colinealidad y creando un conjunto nuevo de variables completamente independientes (Jolliffe, 1986; Mardia y Bibby, 1979). Algunas características particulares a ACP son las siguientes:

- Es una técnica de reducción de variables desde un espacio hiperdimensional a otro de mucho más baja dimensión con muy poca pérdida de información: el nuevo conjunto de variables independientes es mucho menor que su conjunto original.
- La entrada principal al cálculo de los componentes principales es la matriz de covarianza. Sin embargo en procesos de estandarización, se puede emplear la matriz de correlación como entrada principal. De este modo, se eliminan las varianzas extremas que pueden conducir a problemas de singularidad (Jolliffe, 1986).
- Los vectores propios de la matriz de entrada representan los nuevos ejes ortonormales de las variables originales.

- El número de componentes principales es equivalente al número de variables originales.
- La transformación sufrida por los valores de las variables originales después de haber sido rotados a los nuevos ejes ortonormales en el nuevo espacio, se corresponden con los valores de las nuevas variables y son totalmente no correlacionados.
- Los primeros componentes principales explican la mayor cantidad de la varianza de las variables originales y por tanto, concentran la mayor cantidad de información. Estos componentes altamente informados permiten la reducción del conjunto original de variables.
- ACP es una técnica exclusivamente para capturar linealidad entre las variables.
- La correlación entre los componentes principales y las variables originales pueden explicar el coeficiente de determinación y la contribución de cada variable original.

4 Reducción sincronizada de observaciones y de variables

El método propuesto para reducir observaciones y variables consiste en analizar las variables dependientes e independientes como un todo. El método de estratificación reduce las observaciones usando la variable dependiente como variable de estratificación, mediante muestras que sean suficientemente confiables. Este conjunto reducido de observaciones es explorado para reducir las variables independientes y originar un nuevo conjunto de variables tal que dicho conjunto sea lo más representativo posible del espacio original, tanto en observaciones como en variables. Estos métodos fueron combinados en un algoritmo de reducción que se le denominó Stratified/PCA (Colmenares, 1999). [CAIP'99].

Algunas características de esta técnica exploratoria son las siguientes:

- Determina cuantas variables reducidas representan a las variables originales basado sobre el concepto de la varianza explicada por el nuevo conjunto de variables.
- Las variables independientes están en el misma base ortonormal, garantizando de este modo la independencia plena de todas las variables seleccionadas y la correspondencia con las variables originales por los mecanismos de rotación utilizados.
- A los fines de garantizar la mejor selección de la muestra de observaciones a través de la variable dependiente y a su vez, el mejor conjunto reducido de variables, este método analiza no sólo las varianzas de los primeros componentes principales sino también, la correlación entre los últimos componentes (menores componentes o componentes principales insignificantes) y la variable dependiente. Este método añade algunos menores componentes interpretados generalmente como ruido al considerarlos aisladamente, pero que al analizarlo como un todo, variables dependiente e independientes, contribuyen a capturar mayor información que mejora notablemente el desempeño de las RNA(s). En efecto, este método mostró que parte de la no linealidad de las variables independientes podría estar oculta y que al ser capturada pudo mejorar el desempeño de los modelos RNA (Colmanares, 1999) *[CAIP99]*.

Kaiser (1960), Jolliffe (1986), Mardia, Kent y Bibby (1979), han establecido guías que en el contexto de reducción de variables son de bastante utilidad. Se puede distinguir la formulación de cuatro criterios para reducir la presencia de variables en el espacio hiperdimensional que ellas pertenecen:

1. Kaiser y Jolliffe estiman que los valores propios de la matriz de correlación R o la matriz de covarianza Σ estimada a partir de las variables independientes, son el mecanismo apropiado para reducir la cantidad de variables presentes.
2. Jolliffe, Mardia, Kent y Bibby, coinciden en que la medida de correlación entre las variables independientes (sean éstas originales o transformadas por los componentes principales)

y la variable dependiente es importante para determinar el conjunto final de variables.

3. Jolliffe sugiere que debido a la varianza explicada presente en las variables nuevas, se garantizaría que estas nuevas variables tendrían un alto nivel de representatividad de las variables originales.
4. Mardia, Kent y Bibby refuerzan los criterios anteriores manifestando que la correlación entre los componentes principales y la variable dependiente ayuda a determinar si son o no representativas las nuevas variables del conjunto reducido.

5 El método Stratified/PCA

Stratified/PCA examina los datos y los va seleccionando en dos grandes pasos (uno a la vez): a) selección de la muestra estratificada de la variable dependiente; b) reducción de las variables originales independientes a un nuevo conjunto de variables. Si esa muestra seleccionada de observaciones representa lo más parecido posible a su conjunto original y además, si la reducción practicada al conjunto de las variables originales muestra que haya poca pérdida de la información original, entonces esta reducción practicada a los datos originales debería garantizar una construcción de modelos RNA aceptables.

La elaboración del algoritmo Stratified/PCA consta de nueve pasos que aplican los cuatro criterios señalados previamente[3]:

1. El conjunto original de datos se separa en estratos utilizando una variable dependiente como variable de estratificación.
2. ACP es aplicado a cada uno de los estratos definidos en el conjunto original de datos. Los valores y vectores propios

[3] Mayores detalles se pueden encontrar en Colmenares (1999).

son calculados en cada estrato utilizando la matriz de correlación de las variables independientes.

3. Con una definición previa de los parámetros de muestreo, se escoge por estratificación al conjunto reducido de datos donde se incluyen en cada estrato, muestras de los valores tanto de la variable dependiente como de las independientes.
4. Al conjunto reducido de datos que resulte seleccionado se le aplica ACP sobre las variables independientes. Se calculan de la matriz de correlación sus valores y vectores propios usando DVS .
5. ACP es aplicado sobre la variable independiente en cada estrato del conjunto reducido de datos conseguido en el paso 3. De su propia matriz de correlación, se calculan los valores y vectores propios usando DVS.
6. De acuerdo al primer criterio señalado en la sección 4, los valores propios de la matriz de correlación son evaluados estrato por estrato. Aquellos valores propios mayores que un valor umbral establecido, garantizan al componente principal asociado a ese valor propio su incorporación como candidato del conjunto reducido.
7. Mediante el tercer criterio (igualmente formulado en la sección 4) los porcentajes de varianza explicada dados por cada uno de los valores propios en cada uno de los estratos del conjunto reducido de datos, es comparado con el porcentaje de varianza explicada de cada uno de sus estratos equivalentes en el conjunto original de datos. Si cada porcentaje en el conjunto reducido es mayor que en el conjunto original de datos, entonces la muestra es seleccionada. De lo contrario, se estima una nueva muestra como se indica en el paso 3.
8. Se calculan las nuevas variables proyectando sus valores originales en los nuevos ejes ortonormales definidos por los componentes principales asociados a dichas variables. El resultado es una matriz transformada en nuevas variables de la misma dimensión que tendría la matriz de variables originales. Los últimos componentes que podrían ser retenidos se determinan basados en el segundo criterio referido en la sección 4. La matriz de correlación entre la

variable dependiente y los componentes principales del conjunto reducido de datos determinaría cuáles componentes de menor varianza quedarían retenidos de acuerdo a un valor umbral referencial de correlación. Aquellos componentes que estén en la vecindad de este valor serán seleccionados para completar el conjunto total de componentes que definirán las nuevas variables.

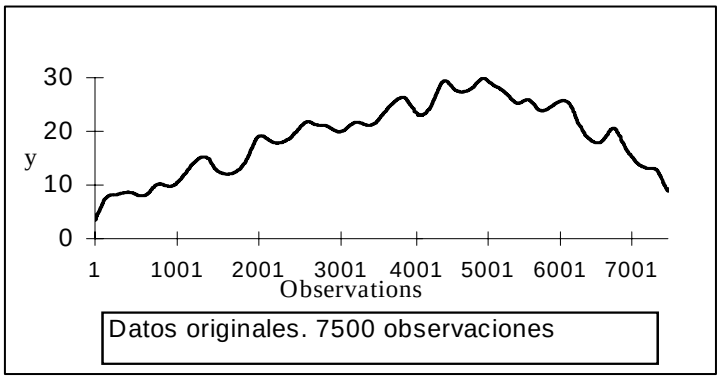
9. La selección final de componentes principales corresponde a los elegidos en el paso 7 y a los que pudieron ser capturados de acuerdo al paso previo. A cada componente seleccionado se le asocia la nueva variable que le corresponde de acuerdo al orden que indique la matriz transformada calculada en el paso 8. Los valores transformados de este conjunto de variables totalmente independientes, explican un alto porcentaje de la varianza de las variables originales. Asimismo, conjuntamente con los valores de la variable dependiente, definen al nuevo conjunto reducido de datos que puede ser aplicado a los modelos RNA.

5 Aplicación mediante experimentación del Stratified/PCA a las Redes Neuronales

En esta sección se muestra la efectividad del método en el desempeño (construcción y generalización) confiable y consistente de los modelos RNA mediante un caso experimental. Los resultados son comparados con los obtenidos en las mismas condiciones experimentales usando datos que fueron seleccionados aleatoriamente o solamente usando ACP. Los parámetros de muestreo y la arquitectura de las redes neuronales se mantienen constantes en todos los experimentos. El conjunto de datos construido, calcula un conjunto de valores (valores de la variable dependiente) en función del un conjunto de nueve términos, a los que se le denominaron variables independientes. Los nueve términos empleados en la función experimental están dados por las

siguientes expresiones: $4xe^{x/3}$, $e^{-\cos\left(\frac{3}{4}\pi x\right)}$, $\frac{x}{2}$, $\cos\left(\frac{\pi}{2}x\right)$, $\cos\left(\frac{\pi}{4}x\right)$, $x^{1/2}$, $e^{\sin(\pi x)}$, $0.00756x^3$, y $0.169x^2$. En el intervalo $0 \leq x$ 25 se le dieron valores discretos a ***x*** a cada una de las expresiones para luego acumularlas y estimar el valor de la variable dependiente ***y***. Cada valor de ***x*** es denotado por x_i, donde ***i*** es un entero entre 1 y 7.500, siendo 7.500 la cantidad creada de puntos de datos. Los valores obtenidos por cada uno de los términos fueron agrupados como observaciones en el conjunto de vectores $\underline{\mathbf{X}}_1$ hasta $\underline{\mathbf{X}}_9$ para cada x_i. La representación de los valores calculados de ***y*** se describen en el gráfico 5.

Gráfico 5
Función experimental



Tal como se puede observar en la tabla 4, se descartó experimentar con el método aleatorio debido a las inconsistencias en el desempeño de los modelos RNA evidenciadas previamente y las cuales pueden verse reseñadas en la sección 2.3 y tabla 2.

Sin embargo, en la misma tabla 4 se muestran los valores promedios de RMSE para los siete modelos RNA entrenados organizados en dos escenarios. En el escenario I se usa el método estratificado para seleccionar el número de observaciones y como no se utiliza ACP, se mantiene el conjunto de todas las variables independientes. En el escenario II, se usa Stratified/PCA para aplicar ambas reducciones: observaciones y variables. Para la reducción de variables se aplicó dos valores umbrales de lambda (valores propios): 1 y 0,7. En el caso de lambda>1, (valor umbral de Kaiser) la reducción generó cinco nuevas variables. Cuando lambda > 0,7 (valor umbral de Jolliffe), la reducción generó siete nuevas variables.

Tabla 4
Comparación mediante valores RMSE promedio obtenidos en los modelos RNA construidos, en los escenarios establecidos

Escenario	**I**	**II**	
Método	**Estratificado**	**Lambda > 1**	**Lambda > 0,7**
Promedio	0,37642	0,45835	0,38769
Desviación estándar	0,08782	0,05405	0,06835

Fuente: Cálculo propios.

El promedio de los valores RMSE, como se puede observar, es bastante similar en los tres casos. Sin embargo, también se puede notar que para los siete modelos RNA construidos, hay una mayor consistencia en los resultados de RMSE (menor dispersión) cuando se aplica el escenario II, indistintamente de cuál sea el criterio, Kaiser o Jolliffe, seleccionado. Más aún, considerando tan solo el escenario II, al aplicar el criterio de Jolliffe, los modelos entrenados fueron más precisos. Esto se debe, fundamentalmente, a que la construcción de estos modelos incluye siete variables independientes en lugar de cinco, tal como recomiendan los resultados usando el criterio de Kaiser. En consecuencia, estas nuevas variables incrementan el porcentaje de varianza retenido que explica las

variables originales. Por último, se puede verificar que al comparar los resultados obtenidos en ambos escenarios, la redundancia de información representa un mayor consumo de tiempo en el cómputo requerido en la construcción de modelos RNA y un innecesario uso de variables que mantienen colinealidad.

La generalización de los modelos RNA construidos mediante los métodos aleatorio, estratificado sin ACP y Straified/PCA, se realiza verificando sus resultados con conjuntos diferentes de 1.505 observaciones que fueron elegidas mediante el uso de los tres métodos y que a su vez no estuvieron comprometidas en las construcción de los modelos que se desean probar. El desempeño de los modelos RNA construidos son evaluados mediante la comparación de los valores de RMSE promedios alcanzados.

En la tabla 5 se observa que el valor promedio de RMSE, calculado a partir de los modelos RNA cuando son verificados con los conjuntos de 1.500 observaciones nuevas para los métodos estratificado sin ACP y Stratified/PCA, fue mucho mejor que el alcanzado por el método aleatorio. Asimismo, en promedio, muestran ser más confiables hasta en un 20% con valores RMSE bajos y en un 50% más consistente con valores bajos de desviación estándar.

Tabla 5
Valores RMSE de la verificación de los modelos RNA con 1.500 observaciones nuevas

Método y/o Escenario			**Stratified/PCA**	
	Aleatorio	**Estratificado**	**Lambda >1**	**Lmbda > 0,7**
Promedio	0,62717	0,03948	0,61886	0,56625
Desviación estándar	0,28382	0,11100	0,16342	0,12659

Fuente: Cálculos propios.

En la tabla 6 se incluye la correlación entre los valores observados y los estimados mediante los modelos RNA construidos con los tres métodos: aleatorio, estratificado sin ACP y Stratified/PCA bajo los dos criterios. Observando la tabla se reafirma que el desempeño logrado por los modelos RNA construidos con observaciones seleccionadas aleatoriamente, resultaron con una mayor dispersión (alta desviación estándar) en su capacidad predictiva, mostrando inconsistencias que se reflejan en predicciones más débiles. De acuerdo al método de estratificación sin ACP los modelos resultaron ser aproximadamente 3 veces más consistentes que los producidos por el método aleatorio. De igual forma, usando Stratified/PCA, los modelos resultaron casi dos veces más consistentes. En definitiva, Stratified/PCA reduce eficientemente y permite obtener consistentemente niveles aceptables de predicción.

Tabla 6
Correlación entre los valores observados y estimados usando conjunto de prueba de 1.500 observaciones

Método y/o Escenario			**Stratified/PCA**	
	Aleatorio	**Estratificado**	**Lambda >1**	**Lmbda > 0,7**
Promedio	0,99437	0,99789	0,99524	0,99623
Desviación estándar	0,00551	0,00131	0,00228	0,0186

Fuente: Cálculos propios.

En resumen, bajo similares condiciones de entrenamiento, Stratified/PCA con siete variables independientes (una reducción de 20% aplicada sobre las variables originales) y la selección estratificada sin ACP mostraron valores similares de RMSE. Del mismo modo, ambos métodos fueron mejores en la generalización de los modelos RNA con los grupos de 1.503 observaciones.

Los gráficos 6 y 7 son un mapa de comparación de resultados obtenidos en modelos RNA construidos y verificados

al aplicar conjuntos reducidos de datos de 250 y 2.000 observaciones mediante los tres métodos: aleatorio, estratificado sin ACP y Stratified/PCA bajos los dos criterios. Para tener una idea en conjunto, en este mapa se muestra el gráfico con las 7.500 observaciones y los gráficos con los resultados obtenidos tanto en la construcción de los modelos RNA (a la izquierda) como sus generalizaciones al ser verificados con las 1.500 observaciones seleccionadas para cada caso (a la derecha).

7 Conclusiones

A partir de los dos criterios de reducción de variables provistos por Stratified/PCA (lambda>1 o lambda>0,7), los conjuntos de datos que pudieron ser seleccionados alcanzaron un desempeño equivalente al obtenido por aquellos conjuntos reducidos de datos formados con todas las variables usando solamente estratificación. Más aún, se demostró en un estudio previo (Cochran, 1963) que el método estratificado por si sólo produce conjuntos reducidos de datos que construyen y generalizan modelos RNA tan confiables como los construidos y generalizados cuando se considera el conjunto original de todos los datos.

La construcción de los modelos RNA fue similar. Mediante el método estratificado y Stratified/PCA se alcanza valores promedios de RMSE que se mueven entre 2 y 20% para un lambda superior a 0,7 o superior a 1. Adicionalmente, la reducción de datos por si solo mejora sustancialmente, casi 45%, la calidad de los modelo construidos.

Los resultados alcanzados por la desviación estándar de los valores RMSE muestra una alta consistencia reflejada en los valores bajos para cada caso: 0,08782 para el método estratificado, 0,05405 para Stratified/PCA con lambda>1, y 0,06835 con lambda>0,7.

Stratified/PCA (lambda>1 y lambda>0,7) puede proveer modelos con habilidad predictiva y a su vez, sus resultados ser más robustos. La correlación entre los valores estimados y los observados fue estable para cada modelo construido y probado, en otras palabras, mostró robustez en la construcción de los modelos RNA. Las desviaciones estándar para cada caso fueron inferiores a la producida por el método aleatorio.

Stratified/PCA obtuvo un sustancial ahorro en el tiempo de cálculo. De acuerdo al conjunto experimental de datos empleado, se redujo el espacio de 68.500 puntos de datos a 10.000. Es decir, casi 14% de la masa de datos fue suficiente para alcanzar un aceptable nivel de efectividad. Al ser comparados todos los métodos ensayados (estratificado si ACP, Stratified/PCA con lambda >0,7 y Stratified/PCA con lambda>1) con el método aleatorio, mostraron casi 67% de mayor consistencia en los resultados de verificación de los modelos medidos a partir de la correlación y casi 35% de mejor construcción en los modelos RNA. En definitiva, Stratified/PCA es una alternativa viable de preprocesamiento de grandes volúmenes de datos históricos para producir conjunto reducidos de datos que permiten construir, probar y generalizar modelos RNA confiables.

Gráfico 6
Construcción de modelos RNA con 250 observaciones y verificación de los modelos con 1.500 observaciones nuevas

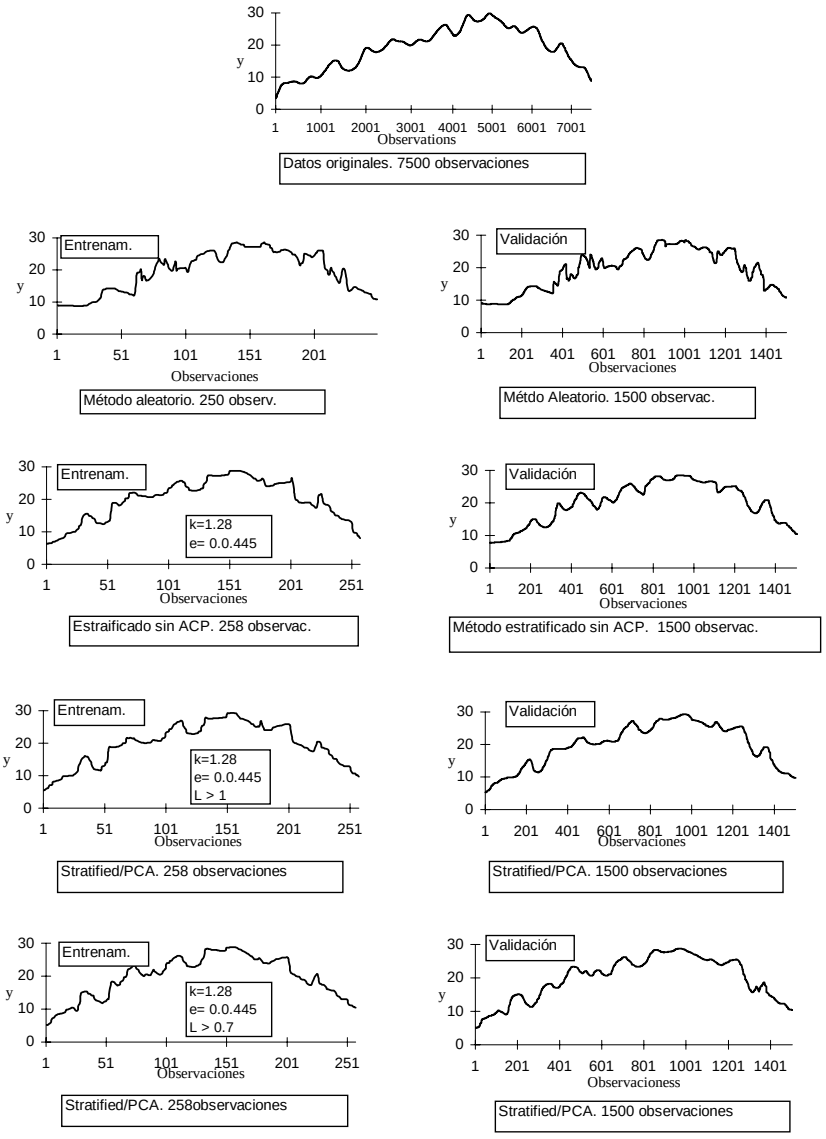
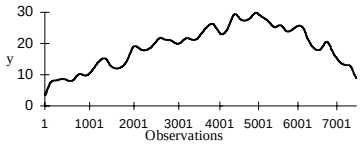
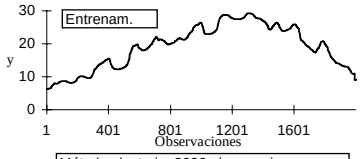


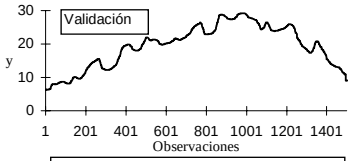
Gráfico 7
Construcción de modelos RNA con 2.000 observaciones y verificación de los modelos con 1.500 observaciones nuevas



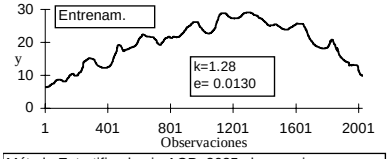
Datos originales. 7500 observaciones



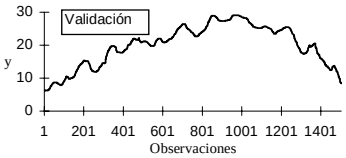
Método aleatorio. 2000 observaciones



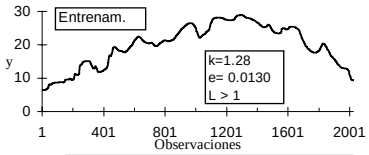
Método Aleatorio. 1500 observaciones



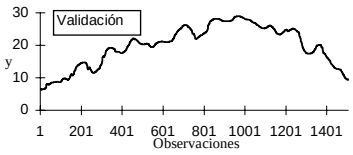
Método Estratificado sin ACP. 2025 observaciones



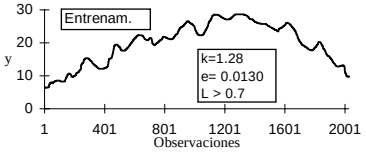
Método estratificado sin ACP. 1500 observaciones



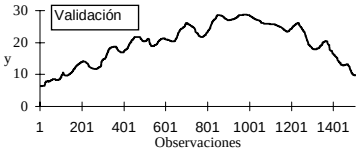
Stratified/PCA. 2025 observaciones



Stratified/PCA. 1500 observciones



Stratified/PCA. 2025 observaciones



Stratified/PCA. 1500 observaciones

8 Bibliografía

Carpenter W. and M. E. Hoffman (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.

Cheng Russel and Teresa Davenport (1989): "The problem of dimensionality in stratified sampling". **Management Science**. Vol. 35. No. 11, Pp. 1278-1296.

Cochran, William (1963): "Sampling techniques". **Second edition. John Willey & Sons, Inc**. New York.

Collantes, Joanna; G. Orlandoni; G. Colmenares y F. Rivas (2002): "Predicción con Redes Neuronales Artificiales: Comparación con las Metodologías de Box y Jenkins". **Mimeografía**. Mérida, Venezuela.

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

Colmenares G. (1999): **"Reducing Samples and Variables to Train, Test, and Validate Neural Networks. Doctoral Dissertation**. University of South Florida. Tampa.

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

Haykin, Simon (1994): **Neural Networks. A comprehensive foundation.** Macmillan College Publishing Company, Inc.

Hecht-Nielsen, Robert (1990): **Neurocomputing**. Addison-Wesley Publishing Company, Inc.

Jolliffe, I. T. (1986): **Principal Component Analysis**. Springer-Verlag.

Kaiser, H. F. (1960): The application of electronic computers to factor analysis. Educ. Psychol. Meas. Vol. 20. Pp. 141-151.

Majcen, Nineta; Karmen Rager-Kanduc; Marjana Novic and Jure Zupan (1995): **Modeling of Property Prediction from Multicomponent Analytical Data Using Different Neural Networks**. Anal. Chem. Vol. 67. Pp. 2154-2161.

Mardia, K. V.; J. T. Kent and J. M. Bibby (1979): **Multivariate Analysis**. Academic Press. London.

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.

Neuralware, Inc. (1995): **NeuralWorks Predict: Complete Solution for Neural Data Modeling**.

Pelletier, Bertrand and Jean-Pierre Chanut (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.

Smith, Murray (1993): **Neural Networks for statistical modeling**. Van Nostrand Reinhold. New York.

Sovan, Lek; Marc Delacoste; Philippe Baran; Ioannis Dimopoulos; Jacques Lauga and Stéphane Aulagnier (1996): "Application of neural networks to modeling nonlinear relationships in ecology". **Elsevier, Ecological Modeling**. Pp. 39-54.

Sukhatme, Pandurang (1963): **Sampling theory of surveys with applications**. Bangalore press, India.

Xue Z., Wang (1999): "Data mining and knowledge discovery for process monitoring and control". **Springer-Verlag.** Great Britain.