
**Revista
Energética**



**Energy
Magazine**

**Año 18
número 3
sept.-dic. 1994**

**Year 18
number 3
Sept.-Dec. 1994**



**Tema: Informática y desarrollo energético
en América Latina y el Caribe**

**Topic: Informatics and energy development in
Latin America and the Caribbean**



Acondicionamiento de la Información para Construir Modelos Neuronales: Estudio de Caso del Precio Promedio del Crudo OPEP

Melio Sáenz y Carlos Galarza*
Diego Erazo y Mario Játiva**

El principal problema al utilizar modelos basados en redes neuronales artificiales es el elevado tiempo de procesamiento necesario para preparar los valores de las variables que intervienen en los escenarios de previsión. Para paliar esta dificultad hemos escogido la vía de buscar nuevas formas de representación de la información, para lo cual en esta primera aproximación hemos ensayado un pre-acondicionamiento de la información mediante la utilización del análisis por componentes principales. Los resultados logrados permanecen en los mismos intervalos de precisión que aquellos obtenidos con el modelo neuronal sin pre-acondicionamiento, mientras que el tiempo utilizado en tratar la información es notablemente menor.

INTRODUCCION

Elaborar pronósticos de precios en el sector petrolero requiere del uso de recursos humanos y tecnológicos sofisticados por parte de las empresas y gobiernos interesados en ejecutarlos. La serie de esfuerzos emprendidos en la década de los 70 condujeron a un desarrollo notable de los modelos algorítmicos. Desgraciadamente la complejidad de los escenarios petroleros no ha permitido alcanzar niveles razonables de precisión en la predicción de los precios del petróleo con la utilización de este tipo de modelos.

La disponibilidad actual de herramientas computacionales sofisticadas nos ha permitido abordar el problema de una manera no-algorítmica.

En trabajos anteriores (5, 6) pudimos constatar que los resultados obtenidos con los modelos basados en redes neuronales artificiales son de mejor calidad que aquellos calculados con los modelos algorítmicos tradicionales.

Con el presente trabajo buscamos reducir el tiempo de proce-

samiento de los datos, con la mínima pérdida de precisión en el pronóstico, para lo cual, antes de utilizar los modelos de redes neuronales, acondicionamos la información aplicando una técnica basada en el análisis de componentes principales.

1.- FORMULACION DEL PROBLEMA

En el estudio precedente (5, 6) sobre la previsión del precio promedio del crudo OPEP en el mercado internacional mediante modelos neuronales, verificamos que estos generaban resultados mas confiables que aquellos obtenidos con los modelos algorítmicos, Oil Market Simulation System (OMS) y Harvard Oil Market Simulation System (HOMS), (3, 4, 7).

En el mencionado estudio utilizamos como información la publicada en los documentos estadísticos del sector petrolero y que guardan concordancia con los mo-

* Unidad de Investigaciones y Desarrollo Tecnológico, PETROE-CUADOR

** Centro Latinoamericano de Cálculo Científico e Informática Industrial

delos algorítmicos referenciales, OMS y HOMMS. Los datos se refieren a trece variables observadas durante 31 años, desde 1960 hasta 1990.

Para crear los escenarios de previsión, mediante la evaluación de los valores de las variables que permiten pronosticar en función del tiempo, a partir de 1991 en adelante, habíamos estudiado dos casos:

CASO 1: en el cual utilizando como variable independiente el tiempo, predecíamos el comportamiento del sistema mediante la evaluación de todas las otras variables;

CASO 2: en el que los valores de las variables fueron obtenidos mediante redes neuronales específicas, entrenadas para cada variable en función del tiempo.

En los dos casos, los resultados obtenidos utilizando redes neuronales son de mayor precisión que aquellos calculados mediante los modelos algorítmicos, con el problema de excesivo consumo de tiempo de cálculo (5).

Para reducir el tiempo de procesamiento hay dos alternativas: optimizar los procesos o buscar una representación sintética de la información, sin perder la fidelidad de la representación de los escenarios reales.

Adoptando la segunda alternativa hemos pre-procesado la información para introducirla en el escenario en mejores condiciones de regularidad funcional mediante

la aplicación del análisis por componentes principales de la matriz de observaciones de los precios promedio del crudo OPEP en el mercado internacional.

2. HERRAMIENTAS COMPUTACIONALES

Para procesar la información y realizar el análisis de componentes principales, hemos utilizado un paquete desarrollado en anteriores trabajos (7). El modelo neuronal artificial está instrumentado bajo un algoritmo del tipo Backpropagation.

Análisis de componentes principales (7, 10)

Sea X una matriz ($m \times n$) de elementos x_{ij} asociada a las m observaciones de las n variables en estudio. Geométricamente esta matriz puede ser representada por una nube N de m puntos l_i en el espacio R_n en el cual el elemento x_{ij} es la coordenada en el eje j de la i -ésima observación l_i .

El análisis por componentes principales nos conduce a encontrar un subespacio R^p tal que p sea menor o igual que n y que minimice la distorsión del contenido de la información. La nube N es proyectada sobre R^p de tal manera que las p componentes principales constituyen una base ortonormal de este subespacio. Conservar las n componentes principales equivale a realizar una rotación del sistema de ejes originales. El procedimiento es el siguiente:

1) Representamos R^n mediante una recta F_p obtenida con un

ajuste por mínimos cuadrados a los puntos de N . La recta F_1 así obtenida es el primer eje factorial sobre el cual proyectamos cada uno de los l_i , elementos de N ;

2) Si el ajuste de la nube N por F_1 no es satisfactorio, buscamos el siguiente eje F_k , $k > 1$ hasta encontrar un subespacio R^p contenido en R^n que minimice la distorsión.

En síntesis, el análisis por componentes principales permite reducir la dimensionalidad del problema en estudio, mediante la identificación de un conjunto de variables no correlacionadas que describan los escenarios con precisión comparable a la producida con la descripción con variables originales.

Redes Neuronales Artificiales (1, 2, 5, 9)

Las redes neuronales artificiales (RNA) son sistemas de procesamiento paralelo y distribuido, cuyos elementos computacionales básicos o neuronas se encuentran densamente conectados entre sí; su estructura y funcionamiento se inspiraron en los del sistema nervioso biológico del cerebro humano. Esto no significa que la neurocomputación no esté abierta a la incorporación de nuevos elementos que no guarden fidelidad con la estructura biológica comprobable del cerebro, cuando éstos contribuyan favorablemente a la mejora de la productividad de los modelos neuronales.

Las redes neuronales artificiales pueden funcionar en dos mo-

dos: en el modo de aprendizaje, la red neuronal, que está definida por una arquitectura A y un conjunto de coeficientes de ponderación Ω , recibe una señal x y calcula un valor y que puede ser el resultado de una clasificación, un vector del mismo tipo del de la entrada, un valor pronosticado o la cuantificación de alguna variable. La red puede caracterizarse por una función global de transferencia $y = F_A \omega(x)$.

En el modo de aprendizaje, la red utiliza ejemplos previamente conocidos para adaptar al problema. La red puede modificar las ponderaciones de las conexiones, la arquitectura o las funciones de transición. Los algoritmos de aprendizaje pueden ser de dos tipos: de aprendizaje supervisado o no supervisado.

Características básicas

La neurona artificial se caracteriza por:

- 1) Su estado interno, s_i , valor real o entero que pertenece al espacio S de estados posibles;
- 2) Las señales de entrada, $s_1, s_2, \dots, s_n$ generadas en otras neuronas o en el entorno de la neurona en estado S_i .
- 3) La función de transición de estados, $s_i = f(s_1, \dots, s_n)$ que es la que describe la asignación de los diferentes estados por los que transita la neurona. La dependencia de f respecto a las señales de entrada puede ser representada por conexiones, la mayor parte de las cuales se

encuentran indexadas mediante coeficientes de ponderación de conexión, ω_{ij} .

Las características de una red neuronal dependen de:

- 1) Su arquitectura, es decir el número de elementos y la estructura de las conexiones;
- 2) La función de sus elementos, la misma que es automáticamente definida por los algoritmos de aprendizaje.

En el caso del algoritmo de BACKPROPAGATION, que es el utilizado en el paquete computacional BRAINMAKER con el que hemos trabajado, minimiza una función que mide la distancia entre la salida deseada d_j y la salida calculada Y_j .

$$(\omega) = \frac{1}{m} \sum \|y_j - d_j\|^2$$

expresión en la que, m es el número de neuronas de la capa de salida y el vector de los coeficientes de ponderación.

La función es minimizada mediante una técnica de gradiente descendente. Los coeficientes de ponderación son modificados por los valores obtenidos mediante la expresión

$$\Delta \omega_{ij} = k \frac{e}{\omega}$$

donde k es una constante de proporcionalidad y los valores de e miden la precisión de la aproximación.

Existen variantes de esta corrección que utilizan expresiones de la forma

$$\Delta \omega_{ij} = kg(Y_j - \omega_{ij})$$

siendo g la función de activación de la neurona.

Puesto que la corrección de los valores de los coeficientes de ponderación comienza en la última capa de neuronas y se avanza hacia la primera, mediante un mecanismo de retorno, los cálculos son más sencillos. El método tiene dos etapas:

Etapas I: En la que se calculan los niveles de actividad Y_j de cada neurona.

Etapas II: Se evalúan las derivadas, comenzando por la capa de salida.

El algoritmo de Back propagation presenta las siguientes ventajas:

1. El proceso de cálculo es lento;
2. El algoritmo guarda el conocimiento adquirido en el transcurso del proceso, por lo que necesariamente ocupa mayor cantidad de memoria del computador;
3. Cuando el proceso encuentra un valor mínimo local, difícilmente sale de él;

Un aspecto de mucho interés consiste en analizar la compatibilidad biológica del algoritmo ya que no existe evidencia de que las neuronas cerebrales funcionen

transmitiendo información en los dos sentidos.

3. APLICACION AL CASO DEL PRECIO PROMEDIO DEL CRUDO OPEP

En los trabajos anteriores hemos estudiado el problema de previsión del precio promedio del crudo producido por los países miembros de la OPEP en el mercado internacional, habiendo obtenido resultados satisfactorios. En ellos hemos utilizado los valores observados de 13 variables a partir del año 1960, los mismos que los presentamos en la tabla N°1.

Para aplicar la metodología descrita en los párrafos anteriores hemos seguido los siguientes pasos:

1. Organización de la base de datos estadísticos de 13 variables y 21 años (cf. Tabla 1).
2. Análisis por componentes principales.
3. Construcción y entrenamiento de los modelos neuronales.
4. Cálculo de previsiones con los modelos neuronales.
5. Comparación con los resultados obtenidos con otros modelos.

Existe la posibilidad de pasar directamente de la base de datos al análisis por redes neuronales, obviando el acondicionamiento previo de la información.

La base de datos estadísticos que utilizamos es la misma

que sirvió para operar con los modelos algorítmicos OMS y HOMS y los modelos neuronales NOMS1 y NOMS2 (5, 6). Suponemos, en todos los casos, que los efectos de las políticas gubernamentales, las estrategias para incrementar la productividad incorporando nuevas tecnologías y los precios relativos de los otros tipos de energía se toman en cuenta en las variables de la base de datos utilizada.

Las características intrínsecas de las redes neuronales utilizadas son las que se describen en la Tabla 3:

TABLA 3 Características de las redes neuronales utilizadas.

RED	NUMERO DE CAPAS			USO
	Entrada	Interme.	Salida	
1	1	2 de 25 neuronas	1	Generación de los valores de las variables que intervienen en los escenarios.
2	4	2 de 25 neuronas	1	Previsión de las variables de escenarios.
3	1	3 de 20-30-50 neuronas	1	Previsión de los precios

4. ANALISIS DE RESULTADOS

Analicemos los resultados obtenidos de la aplicación de las componentes principales a los datos estadísticos utilizados en nuestro estudio.

En el gráfico de la fig. 1 tenemos en ordenadas el valor de la varianza explicada por cada uno de los valores propios de la matriz de correlación de los datos originales.

Constatamos que el primer valor propio explica el 68.97% de la varianza, el segundo el 18.22%, el tercero 8.80% y el cuarto, 2.92%, es decir con los tres primeros valores propios explicamos el 96% de la varianza. Esto significa que la matriz original de datos puede ser representada por la matriz de la tabla 2 con una pérdida de información del 4%, valor que se encuentra en el intervalo de tolerancia del error de nuestro estudio.

Del cálculo de correlaciones entre componentes principales y las variables originales (fig. 2)

constatamos que las dos primeras componentes principales tienen coeficientes de correlación con valores cercanos a la unidad respecto a casi todas las variables originales. Sólo para las variables correspondientes a las importaciones, Balance de la OPEP y Producción de Refinados los coeficientes de correlación son bajos.

En la fig. 3 tenemos la ubicación de los coeficientes de co-

relación de la tercera y cuarta componentes respecto a las variables originales. Vemos que sólo para la variable original BALANCE DE LA OPEP el coeficiente es cercano a la unidad.

Este análisis nos conduce a decidir trabajar con las tres primeras componentes principales, lo que nos ha permitido disminuir el número de variables de 13 (Tab.1) a 3 (Tab. 2). De acuerdo con la metodología empleada en trabajos anteriores (cf. 5,6) luego de haber entrenado las redes neuronales pasamos a la etapa de evaluación de los valores futuros del precio promedio del crudo OPEP en el mercado internacional para lo cual necesitamos generar los valores de las variables de entrada de la red mediante redes en las cuales el valor de entrada es el año y obtenemos como valor de salida el de la variable estudiada. Estos constituyen los valores de entrada de la red neuronal de precios

En la Tab. 4 hemos tabulado el año, el precio real observado del crudo y los valores obtenidos con dos redes neuronales: CP1 en la cual el error tolerado durante el entrenamiento de la red lo hemos fijado en el 5% y CP2 con un error tolerado del 10%.

Constatamos que los pronósticos realizados para los cuatro años de comprobación son particularmente precisos, especialmente para los modelos en los que hemos utilizado el análisis por componentes principales, lo que convalida la metodología utilizada para pronósticos de este plazo.

En los gráficos de la fig. 4, tenemos en ordenadas los resultados obtenidos con los modelos neuronales y los precios reales mientras que en abscisas tenemos el tiempo. Durante el período de entrenamiento de las redes obtenemos los mejores resultados con la red de tolerancia de entrenamiento del 5% mientras que para los valores pronosticados, la reducción del error tolerado durante el entrenamiento no da lugar a una mejor aproximación.

Comparando los pronósticos obtenidos con los modelos entrenados directamente con los datos observados (DO1 y DO2), con aquellos obtenidos a partir de los modelos para los cuales hemos preparado la información utilizando el análisis por componentes principales (CP1 y CP2), constatamos que estos últimos son de mejor calidad, confirmando de esta manera la recomendación de los fabricantes de BRAINMAKER en el sentido de que es mejor utilizar valores normalizados de la información suministrada para el entrenamiento de las redes neuronales.

En la tabla N. 5 hemos sintetizado los tiempos empleados por el computador para procesar los diferentes tipos de redes. Así, para el proceso de preparación de los datos, con una red de una entrada y de una salida tenemos un tiempo promedio de procesamiento de 80 segundos, mientras que para las redes utilizadas para el pronóstico, el tiempo promedio es de 6 segundos. Se confirma, así, el mismo comportamiento encontrado en los trabajos anteriores.

CONCLUSIONES

El acondicionamiento de la información, mediante un tratamiento analítico previo, para utilizar modelos neuronales es una herramienta interesante para el conocimiento y aprehensión de los escenarios así como para optimizar los requerimientos de recursos computacionales. En efecto, al analizar la información con las componentes principales hemos comprendido de manera más clara las relaciones existentes entre las variables originales del modelo. Desde el punto de vista computacional hemos logrado discernir algunas limitaciones de las redes neuronales, sobretodo aquella relacionada con los órdenes de magnitud de los intervalos de variación de las variables empleadas como valores de entrada. El error tolerable de trabajo es un porcentaje de este intervalo, de dónde se genera el interés para trabajar con valores centrados y reducidos.

Los modelos neuronales entrenados con información pre-procesada mediante el análisis por componentes principales presenta las siguientes ventajas:

- 1) La dimensionalidad del problema se reduce. En nuestro caso hemos conseguido representar el 96% de la varianza de la matriz de 13 variables conservando solamente las tres primeras componentes principales.
- 2) El tiempo de procesamiento se ha reducido notablemente. En el caso de la aproximación con el 5% de error tolerable, el tiempo utilizado con informa-

ción preacondicionada es 5.64 veces menor que con la información original y en el caso del error del 10%, la reducción es de 635 veces.

- 3) Los órdenes de magnitud de los errores calculados en el período para el cual disponemos de datos reales observados se mantienen en valores menores al 20% en el caso de la red con error tolerable de aprendizaje del 5%, y con un valor máximo del 11.28% en el caso del error tolerable del 10%.
- 4) El intervalo de variación de los errores al final del período de previsión es mayor cuando varía el error tolerable (del 5% al 10% para CP1 y CP2 respectivamente) que cuando cambia la estrategia de procesamiento (de un solo valor de entrada en NOMS1 al procesamiento en cascada en NOMS2).

Agradecimiento: Este trabajo fue realizado en el marco del PROGRAMA DE REDES NEURONALES de la Unidad de Investigación y Desarrollo Tecnológico de PETROECUADOR. Los autores agradecen a Rafael Melgarejo Heredia por las valiosas sugerencias respecto a los modelos neuronales.

REFERENCIAS

- 1) California Scientific Software, *Introduction to Neural Networks Computer Simulation of Biological Intelligence*. Sierra Madre, California 1990.
- 2) Do Amaral Pereira, Paulo César *Inteligencia Artificial y Redes Neuronales* ESPE Quito, 1992
- 3) Hogan, William *Oil Demand and OPEC's Recovery Energy and environmental Policy Center* Harvard Energy Security Program, John F. Kennedy School of Government, Harvard University, June 1980
- 4) Hogan, William, *Oil Market risk Analysis* Energy and environmental Policy Center Harvard Energy Security Program, John F. Kennedy School of Government, Harvard University, Dec. 1985
- 5) Melgarejo, R; Calderón A., *Estudio de modelos de mercado petrolero a través de redes neuronales artificiales*, Tesis presentada a la Facultad de Ingeniería de Sistemas de la Escuela Politécnica Nacional, Quito, 1992
- 6) Sáenz, M; Izurieta, C; Melgarejo, Calderón, A; *Estudio de los precios del crudo en el mercado OPEP utilizando redes neuronales* PC-World Ecuador, marzo 1993
- 7) Sáenz, M. *Análisis por componentes principales* Unidad de Sistemas de Petroecuador, I.T. N. 13. May 1991
- 8) Smith-Perera, R. *Optimal energy pricing in oil exporting countries: second best alternatives* Energy and Environmental Policy Center, Harvard Energy Security Program. John F. Kennedy School of Government, Harvard University, Sept. 1987
- 9) Engelman-Souille, F *Neural networks and their applications for the oil industry* Revue de l'Institut Français du Pétrole, Vol 47, N. 3 mayo-junio 1992
- 10) Pla, L.E. *Análisis multivariado: método de componentes principales* Serie de Publicaciones Científicas de la Organización de los Estados Americanos, OEA, Washington 1986

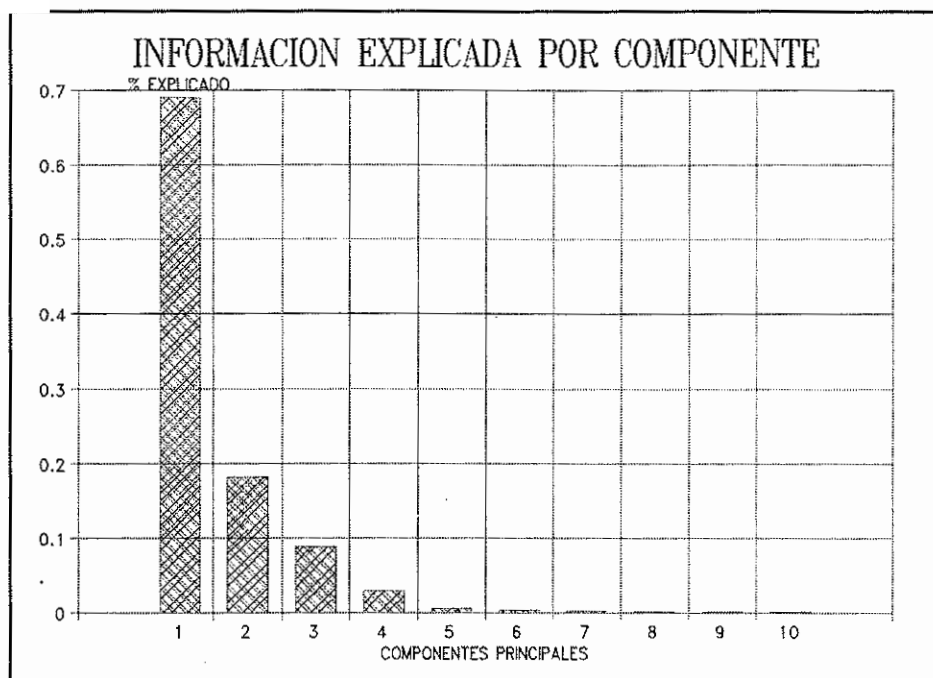


FIG. 1 Varianza explicada por cada una de las componentes principales calculadas en el pre-acondicionamiento de la información.

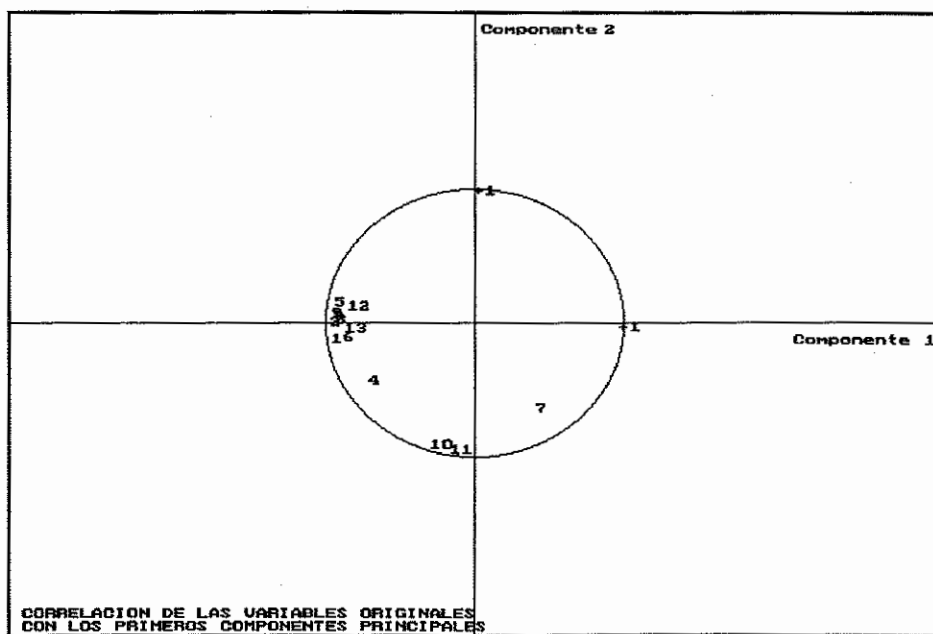


FIG 2 Correlación de las variables originales. Proyección en el plano de las dos primeras componentes principales.

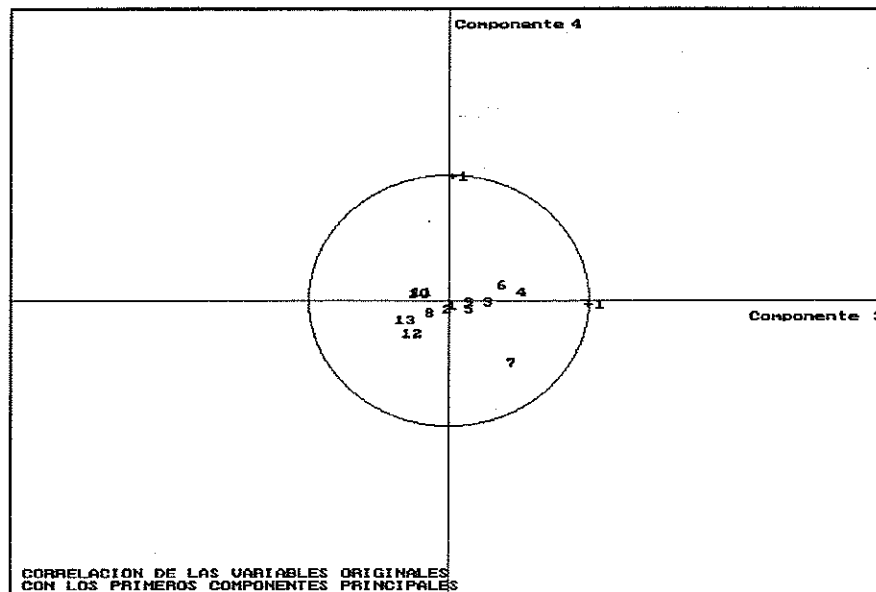


FIG. 3 Correlación de las variables originales. Proyección en el plano de las componentes principales 3 y 4

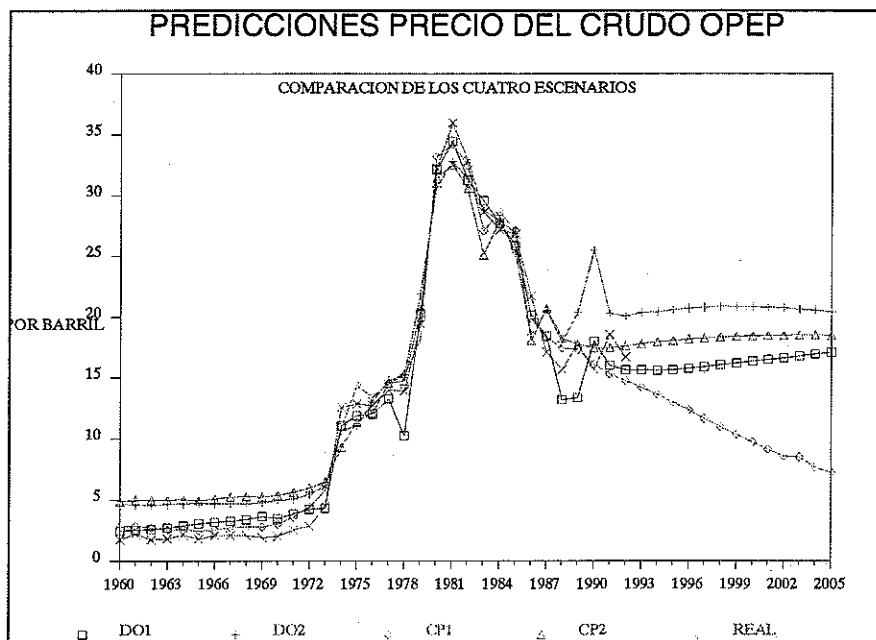


FIG 4.- Resultados obtenidos con las redes neuronales:

- DO1 red entrenada con valores reales observados y tolerancia del 5%
- DO2 red entrenada con valores reales observados y tolerancia del 10%
- CP1 red entrenada con información preprocesada con componentes principales y tolerancia del 5%
- CP2 red entrenada con información preprocesada con componentes principales y tolerancia del 10%

BASE DE DATOS UTILIZADA PARA EL ENTRENAMIENTO.

AÑO	CPRF	PBLC	PIB	EXPT	EXNP	IMPR	BLNC	PRRF	CSRF	EXPTBR	PROD	EXRF	RSVA
1960	2,154.0	205851	27421	5149	2157	5454	-493	2,020.7	330.8	6,747.0	8,696.9	1,236.3	218,180.0
1961	2,181.6	210796	28025	5468	1814	5547	-735	2,072.6	349.9	7,407.6	9,363.8	1,222.0	223,834.0
1962	2,317.7	215380	29573	5931	1691	4822	356	2,334.0	358.9	8,344.2	10,508.1	1,316.4	231,980.0
1963	2,423.7	220545	32078	6752	1853	4994	991	2,221.1	379.3	9,250.9	11,514.2	1,477.8	248,075.0
1964	2,487.8	225973	33617	7751	2010	5727	589	2,178.8	426.7	10,617.2	12,986.9	1,517.8	254,950.0
1965	2,651.1	231324	36168	8447	1851	6602	247	2,385.6	488.6	12,768.3	14,739.2	1,637.9	260,345.0
1966	2,767.2	237360	37279	9030	1820	6581	616	2,451.7	512.7	13,507.7	15,770.2	1,626.8	280,068.0
1967	2,905.4	243191	40767	10005	2425	7101	671	2,500.8	542.8	14,239.5	16,849.3	1,667.5	310,425.0
1968	3,107.7	249522	44691	11620	2239	7885	256	2,757.9	603.5	16,076.7	18,785.6	1,717.4	332,030.0
1969	3,150.4	256043	51464	12678	2788	8915	-220	2,775.4	647.8	18,010.0	20,906.3	1,682.7	402,216.0
1970	3,565.4	262809	56884	14543	2742	9923	783	3,217.0	681.8	20,223.4	23,413.0	1,973.0	402,216.0
1971	3,663.6	270143	66497	19662	2607	11557	3311	3,252.9	744.7	22,031.7	25,326.3	1,814.5	430,820.0
1972	3,948.4	277853	78046	23542	3112	14217	4104	3,131.9	858.2	24,078.5	27,194.4	1,792.3	428,492.0
1973	4,128.2	285791	107642	37145	4660	20364	7025	3,465.1	1002.8	27,547.2	30,988.5	1,974.4	420,887.7
1974	4,290.6	294348	210653	120436	6398	33015	66653	3,386.6	1107.9	27,258.9	30,729.2	1,888.8	487,439.1
1975	4,488.1	303672	250285	107773	5994	52173	31457	3,041.9	1281.3	24,063.9	27,155.0	1,559.7	448,947.6
1976	4,637.7	312117	313522	130199	7470	64262	36854	3,416.9	1521.9	27,462.6	30,737.8	1,870.9	440,414.8
1977	4,972.3	321568	375258	141899	9568	85655	20173	3,613.8	1761.2	27,641.1	31,253.5	1,752.0	438,247.4
1978	5,189.3	331628	395527	134982	11322	97181	-3846	3,792.6	1922.1	26,088.7	29,805.3	1,905.5	443,793.9
1979	5,214.4	341472	492681	203193	16011	101154	60136	3,992.4	2177.8	26,764.9	30,928.8	2,141.6	436,179.0
1980	5,767.5	352650	641919	284491	22113	127266	105039	3,853.3	2429.8	22,864.1	26,880.0	2,008.6	435,955.9
1981	5,995.5	363083	709857	263337	21840	149865	52904	3,943.4	2819.4	18,568.8	22,598.6	1,979.3	438,371.5
1982	5,136.2	373843	734894	202789	21149	148957	2422	4,350.8	3043.6	14,296.1	18,992.2	2,225.7	466,252.8
1983	5,893.9	384861	718861	156945	22265	139742	-16595	4,493.2	3275	12,506.5	16,992.3	2,023.1	470,951.8
1984	6,179.5	396114	715836	143569	23586	122359	-3374	4,648.9	3156.6	12,043.9	16,347.2	2,281.3	510,411.2
1985	6,429.5	407987	709728	127182	23601	102112	7656	5,003.6	3199.2	10,926.4	15,447.7	2,368.2	536,976.6
1986	6,394.2	419602	686233	76741	25167	90482	-19889	5,334.4	3163.8	12,939.8	18,333.0	2,642.0	644,791.3
1987	6,530.0	432464	713210	92602	27002	89334	-4123	5,371.3	3270.9	12,060.4	17,324.9	3,017.2	676,488.7
1988	6,472.0	444403	621211	84731	31945	103961	-17115	5,914.7	3423.6	13,367.0	19,618.4	3,395.3	762,928.9

TABLA 2.- Datos calculados con las componentes principales.

Año	CP1	CP2	CP3
1960	1.324	1.129	0.911
1961	1.303	1.059	0.842
1962	1.229	0.956	0.657
1963	1.168	0.867	0.476
1964	1.126	0.716	0.351
1965	1.036	0.521	0.106
1966	0.982	0.421	-0.032
1967	0.915	0.330	-0.163
1968	0.816	0.133	-0.405
1969	0.724	-0.085	-0.698
1970	0.556	-0.309	-1.076
1971	0.516	-0.554	-1.133
1972	0.459	-0.786	-1.174
1973	0.277	-1.169	-1.402
1974	0.086	-1.478	-0.631
1975	0.144	-1.049	-0.091
1976	-0.104	-1.418	-0.265
1977	-0.270	-1.432	-0.024
1978	-0.423	-1.084	-0.160
1979	-0.651	-1.463	0.492
1980	-0.890	-1.299	1.842
1981	-1.004	-0.637	2.046
1982	-1.028	0.148	1.546
1983	-1.173	1.223	0.354
1984	-1.129	0.657	0.945
1985	-1.160	0.744	0.761
1986	-1.441	1.444	-1.395
1987	-1.491	0.890	-0.518
1988	-1.899	1.523	-2.163

TABLA 4.- Valores obtenidos con los modelos de redes neuronales

AÑO	VALOR OBSERVADO	PORCENTAJES DE ERROR ASIGNADOS PARA ENTRENAMIENTO							
		5% DO		10% DO		5% CP		10% CP	
		VALOR	ERROR %	VALOR	ERROR %	VALOR	ERROR %	VALOR	ERROR %
1989	17.75	13.4	15.94	20.38	14.82	17.32	2.42	17.75	0.00
1990	15.69	18.01	24.51	25.45	62.21	16.11	2.68	17.46	11.28
1991	18.57	16.02	14.79	20.28	9.21	15.32	17.50	17.48	5.87
1992	16.71	15.67	13.73	20.09	20.23	14.79	11.49	17.63	5.51

TABLA 5.- Tiempo de procesamiento empleado en el entrenamiento.

Tipo de entrenamiento	Tipo de red neuronal (# entradas/# salidas)	
	1/1	13/1
Máximo	150'	6'
Mínimo	5'	
Promedio	80'	
Total	17 x 80'	6'

Pre-processing Information to Build Neural Models: Case Study of the Average Price of OPEC Crude Oil

Melio Sáenz and Carlos Galarza*
Diego Erazo and Mario Játiva**

The major problem in using models based on artificial neural networks is the large amount of processing time needed to prepare the values of the variables involved in the forecasting scenarios. In order to mitigate this difficulty, we have decided to look for new ways of representing the information. To do this, in this first attempt, we have tried a preliminary treatment of the information by using the analysis by principal components. The results that have been achieved remain at the same precision intervals as those obtained with the neural model without pre-processing, while the time used in processing the information is considerably lower

INTRODUCTION

In order to prepare forecasts for oil sector prices, companies and governments interested in these forecasts require sophisticated human and technological resources. Efforts undertaken in the seventies have fostered the widespread development of algorithmic models. Unfortunately, the complexity of oil scenarios has not permitted reasonable levels of accuracy in forecasting oil prices using this type of model.

The current availability of sophisticated computer tools, however, enables us to tackle the problem non-algorithmically.

In previous work (5, 6), we were able to observe that the results obtained with models based on artificial neural network were better than those obtained from traditional algorithmic models.

The present effort is aimed at reducing data-processing time with the least loss of accuracy in the forecasting. To do this, before applying the neural network models, we pre-process the information using a technique based on the analysis of principal components.

1. FORMULATION OF THE PROBLEM

In the previous study (5, 6) on forecasting the average price of OPEC crude oil on the international market by means of neural models, we verified that these models yielded results that were more reliable than the ones obtained with the algorithmic models: Oil Market Simulation System (OMS) and Harvard Oil Market Simulation System (HOMS) (3, 4, 7).

In the above-mentioned study, we used the information published in the statistical documents of the oil sector and in keeping with the referential algorithmic models OMS and HOMS. The data refer to 13 variables observed over 31 years, from 1960 to 1990.

In order to create the forecasting scenarios by evaluating the values of the variables that permit forecasting on the basis of time, from 1991 to the present, we had studied two cases:

* Research and Development Unit, PETROECUADOR, Quito, Ecuador

** Latin American Center of Scientific Calculation and Industrial Informatics, Quito, Ecuador

CASE 1: In which, using time as an independent variable, we predicted the system's behavior by evaluating all the other variables.

CASE 2: In which the values of the variables were obtained by means of specific neural networks trained for each variable on the basis of time.

In the two cases, the results that were obtained using neural networks were more accurate than the ones calculated using algorithmic models; the latter method, in addition, involved the problem of an excessive amount of time spent for calculation (5).

To reduce processing time, there are two alternatives: to optimize the processes or seek a synthetic representation of the information, without losing the authenticity of the representation of real scenarios.

We adopted the second alternative and pre-processed the information to introduce it into the scenario under better conditions of functional regularity by applying the analysis by principal components of the matrix of observations of the average prices of OPEC crude oil on the international market.

2. COMPUTER TOOLS

In order to process the information and conduct the analysis of principal components, we used a package developed in previous work (7). The artificial neural model is implemented using a backpropagation algorithm.

Analysis of Principal Components (7, 10)

Let X be a matrix ($m \times n$) of elements x_{ij} associated to m observations of n variables under study. Geometrically, this matrix can be represented by an N cloud of m points I_i in space R_n , in which the x_{ij} element is the coordinate in the j -axis of the i th observation I_i .

The analysis by principal components leads us to find a subspace R^p , such that p is less than or equal to n and minimizes the distortion of the contents of the information. The N cloud is projected onto R^p so that the major p components constitute an orthonormal base of this subspace. Keeping the principal n components is equivalent to rotating the original axis system. The procedure is as follows:

1. We represent R^n by means of a straight line F_p obtained with an adjustment for mean squared to the points of N . The straight line F_1 thus obtained is the first factorial axis on which we project each one of the I_i elements of N .
2. If the adjustment of N cloud for F_1 is unsatisfactory, we seek the following F_k , $k > 1$ until we find a subspace R^p contained in R^n that minimizes the distortion.

In short, the analysis by principal components enables us to reduce the dimensionality of the problem being studied, by identifying a set of non-correlated variables that describe the scenarios with a precision that is comparable

to the one produced with the description of the original variables.

Artificial Neural Networks (1, 2, 5, 9)

Artificial neural networks (ANN) are parallel and distributed processing systems, whose basic computer elements or neurons are densely interconnected; their structure and operation are based on those of the human brain's nervous system. This does not mean that neurocomputer technology is not open to the incorporation of new elements that do not correspond to the human brain's verifiable biological structure if and when these elements contribute to improving the productivity of neural models.

Artificial neural networks can function in two modes: in the consulting mode, the neural network, which is defined by an A architecture and a set of W weighting coefficients, receives an x signal and calculates a y value that can be the outcome of a classification, a vector of the same type as that of the input, a forecast value, or the quantification of some variable. The network can be characterized by a global transfer function $y = F_A W(X)$.

In the consulting mode, the network uses previously known examples to adapt to the problem. The network can modify the weightings of the connections, the architecture or the transition functions. The consulting algorithms can be of two kinds: supervised or nonsupervised consulting.

Basic characteristics

The artificial neuron is characterized by:

1. Its internal state, s_i , real or integer value that belongs to space S of possible states.
2. The input signals, $s_1, s_2, \dots, s_n$ generated in other neurons or in the environment of the neuron in state S_i .
3. The transition function of states $s_i = f(s_1, \dots, s_n)$ which describes the assignation of the different states that the neuron passes through. The dependence of f with respect to the input signals can be represented by connections, most of which are indexed by means of connection weighting coefficients, W_{ij} .

The characteristics of the neural network depend on:

1. Its architecture, that is, the number of elements and the structure of the connections.
2. The function of its elements, which is automatically defined by the consulting algorithms.

In the case of the backpropagation algorithm, which is used in the brainmaker computer package with which we have worked, it minimizes a function that measures the distance between the desired output d_j and the calculated output Y_j :

$$(\omega) = \frac{1}{m} \sum \|y_j - d_j\|^2$$

an equation in which m is the number of neurons of the output layer and W the vector of the weighting coefficients.

The function is minimized by means of a downward slope technique. The weighting coefficients are modified by the values obtained by means of the following equation:

$$\Delta\omega_{ij} = k \frac{e}{\omega}$$

where k is a constant of proportionality and the values of e measure the accuracy of the approximation.

There are variants of this correction that use equations in the following form:

$$\Delta\omega_{ij} = kg(Y_j - \omega_{ij})$$

g being the function of the activation of the neuron.

Since correction of the values of the weighting coefficients begins in the last layer of neurons and moves backwards toward the first, by means of a return mechanism, the calculations are simpler. The method involves two phases:

Phase I: In which the activity levels Y_j of each neuron are calculated.

Phase II: The derivatives are evaluated beginning with the output layer.

The backpropagation algorithm has the following drawbacks:

1. The calculation process is slow.
2. The algorithm keeps the knowledge acquired during the process and because of that it uses up a larger amount of computer memory.
3. When the process finds a minimum local value, it has trouble leaving it.

One highly interesting aspect consists of analyzing the biological compatibility of the algorithm since there is no evidence that the cerebral neurons operate transmitting information in both directions.

3. APPLICATION TO THE CASE OF THE AVERAGE PRICE OF OPEC CRUDE OIL

In the previous work, we studied the problem of forecasting the average price of crude oil produced by the member countries of OPEC on the international market and obtained satisfactory results. We used the observed values of 13 variables as of the year 1960, as indicated in Table 1.

In order to apply the methodology described in the preceding paragraphs, we took the following steps:

1. Organization of the statistical data base of 13 variables and 21 years (see Table 1).
2. Analysis by principal components.
3. Construction and training of the neural models.

4. Calculation of the forecasts with the neural models.
5. Comparison with the results obtained with other models.

There is the possibility of going directly from the data base to the analysis by means of neural networks, obviating the prior processing of the information.

The statistical data base that we used is the same as the one used to operate with the OMS and HOMS algorithmic models and the NOMS1 and NOMS2 neural models (5, 6). Let us assume, in any case, that the effects of government policies, the strategies to increase productivity by incorporating new technologies, and the relative prices of other types of energy are taken into account in the variables of the data base used.

The intrinsic characteristics of the neural networks used are described in the following table 3.

4. ANALYSIS OF THE RESULTS

Let us analyze the results obtained from the application of the principal components to the statistical data used in our study.

In the chart of Figure 1, we have the value of the variance explained for each one of the values of the correlation matrix of the original data on the ordinates. We observe that the first value explains 68.97% of the variance, the second 18.22%, the third 8.80%, and the fourth 2.92%; in other words, with the first three values we can explain 96% of the variance. This means that the original data matrix can be represented by the matrix of Table 2 with a 4% loss of information, a value that can be found in the error tolerance interval of our study.

On the basis of the calculations of the correlations between principal components and original variables (see Figure 2), we

observe that the first two principal components have correlation coefficients with values close to one with respect to almost all the original variables. Only for the variables corresponding to imports, OPEC balance, and production of oil products are the correlation coefficients low.

Figure 3 shows the location of the correlation coefficients of the third and fourth components regarding the original variables. We observe that only for the original variable OPEC balance is the coefficient close to one.

This analysis leads us to decide to work with the first three principal components, which has enabled us to reduce the number of variables from 13 (Table 1) to 3 (Table 2). According to the methodology used in previous work (5, 6), after having trained the neural networks, we shift to the evaluation phase of future values of the average price of OPEC crude oil on the international market. For this we need to generate the values of the network input variables in which the input value is the year and we obtain as an output value that of the variable studied. These are the input values of the neural price network.

Table 4 tabulates the year, real price observed of crude oil, and values obtained with two neural networks: CP1 in which the tolerated error during the training of the network has been set at 5% and CP2 with a tolerated error of 10%.

Table 3: Characteristic of the neural networks used

NET- WORK	NUMBER OF LAYERS			USE
	Input	Intermediate	Output	
1	1	2-25 neurons	1	Forecasting the variables of the scenarios
2	4	2-25 neurons	1	Price forecasting
3	1	3-20-30-50 neurons	1	Price forecasting

We observe that the forecasts conducted for the four years of testing are especially accurate, above all for the models we have used in analyzing the principal components, which validates the methodology used for forecasting this time period.

In the charts of Figure 4, the results obtained with the neural models and the real prices are in the ordinates, while time is on the abscissas. During the training period of the networks, we have better results with the training tolerance network of 5%, whereas for the forecast values, reduction of the tolerated error during the training does not lead to a closer approximation.

Comparing the forecasts obtained with the models trained with the observed data (DO1 and DO2) with those obtained on the basis of the models for which we pre-processed the information using the analysis by principal components (CP1 and CP2), we observe that the latter are higher quality, thus confirming the recommendation of the manufacturers of BRAINMAKER to the effect that it is better to use standardized values of the information supplied for training of the neural networks.

In Table 5 we have synthesized the time used by the computer to process the different types of networks. Thus, for the data-preparing process with a network of one input and one output we have an average processing time of 80 seconds, whereas for the networks used for the forecasting, the average time is 6 seconds. Thus,

the same behavior found in the previous work is confirmed.

CONCLUSIONS

Information pre-processing, applying a prior analytical treatment, to use neural models is an interesting tool for learning about and apprehending the scenarios as well as optimizing computer resource requirements. Indeed, upon analyzing the information with the principal components, we have understood more clearly the existing relations between the model's original variables. From the standpoint of computer technology, we have managed to discern some of the limitations of neural networks, especially with respect to the orders of magnitude of the variation intervals of the variables used as input values. The tolerable error of the work is a percentage of this interval, from which comes the interest to work with centered and reduced values.

The neural models trained with pre-processed information by means of the analysis by principal components provide the following advantages:

1. The dimensionality of the problem is reduced. In our case, we have managed to represent 96% of the variance of the matrix of 13 variables, keeping only the first three principal components.
2. The processing time has dropped considerably. In the case of the approximation with the 5% of tolerable error, the time spent with pre-

processed information is 5.64 times less than with the original information and, in the case of the 10% error, the reduction is 635 times.

3. The orders of magnitude of the errors calculated in the period for which we have real observed data are kept at values under 20% in the case of the network with the tolerable learning error of 5% and with a maximum value of 11.28% in the case of the tolerable error of 10%.
4. The variation interval of the errors at the end of the forecasting period is greater when the tolerable error varies (from 5% to 10% for CP1 and CP2, respectively) than when the processing strategy changes (from one single input value in NOMS1 to cascade processing in NOMS2).

Acknowledgments: The present work was conducted within the framework of the Neural Network Program of the Research and Development Unit of PETROECUADOR. The authors wish to thank Rafel Melgarejo-Heredia for his valuable suggestions regarding neural models.

REFERENCES

1. California Scientific Software, *Introduction to Neural Networks Computer Simulation of Biological Intelligence*, Sierra Madre, California, 1990.
2. Do Amaral Pereira, Paulo César, *Inteligencia Artificial y*

- Redes Neuronales*, ESPE, Quito, 1992.
3. Hogan, William, *Oil Demand and OPEC's Recovery*, Energy and Environmental Policy Center, Harvard Energy Security Program, John F. Kennedy School of Government, Harvard University, June 1980.
 4. Hogan, William, *Oil Market Risk Analysis*, Energy and Environmental Policy Center, Harvard Energy Security Program, John F. Kennedy School of Government, Harvard University, December 1985.
 5. Melgarejo, R., Calderón, A., *Estudio de Modelos de Mercado Petrolero a través de Redes Neuronales Artificiales*, thesis submitted to the School of Systems Engineering of the National Polytechnic School, Quito, 1982.
 6. Sáenz, M., Izurieta, C., Melgarejo, R., Calderón A., *Estudio de los Precios del Crudo en el Mercado OPEP Utilizando Redes Neuronales*, PC-World Ecuador, March 1993.
 7. Sáenz, M., *Análisis por Componentes Principales*, Systems Unit, PETROE-CUADOR, ITN, No. 13, May 1991.
 8. Smith-Perera, R., *Optimal Energy Pricing in Oil Exporting Countries: Second Best Alternatives*, Energy and Environmental Policy Center, Harvard Energy Security Program, John F. Kennedy School of Government, Harvard University, September 1987.
 9. Engelman-Souille, F., *Neural Networks and their Applications for the Oil Industry*, Revue de l'Institut français du pétrole, Vol. 47, No. 3, May-June 1992.
 10. Pla, L.E., *Análisis Multivariado: Método de Componentes Principales*, Scientific Publications Series of the Organization of American States, Washington, D.C., 1986.

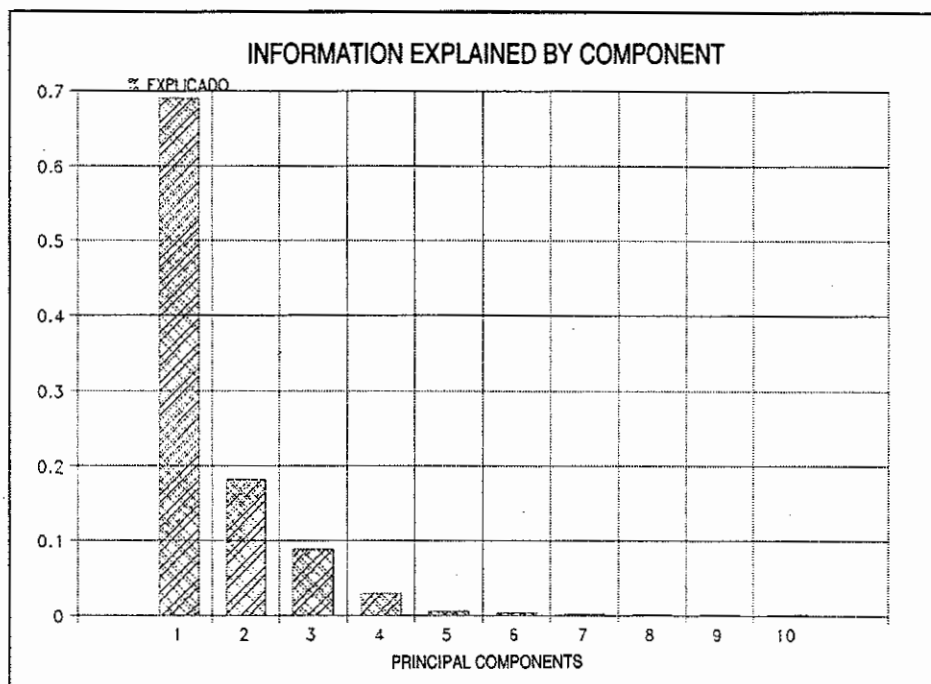


Figure 1: Variance explained for each one of the principal components calculated in the preliminary processing of the information

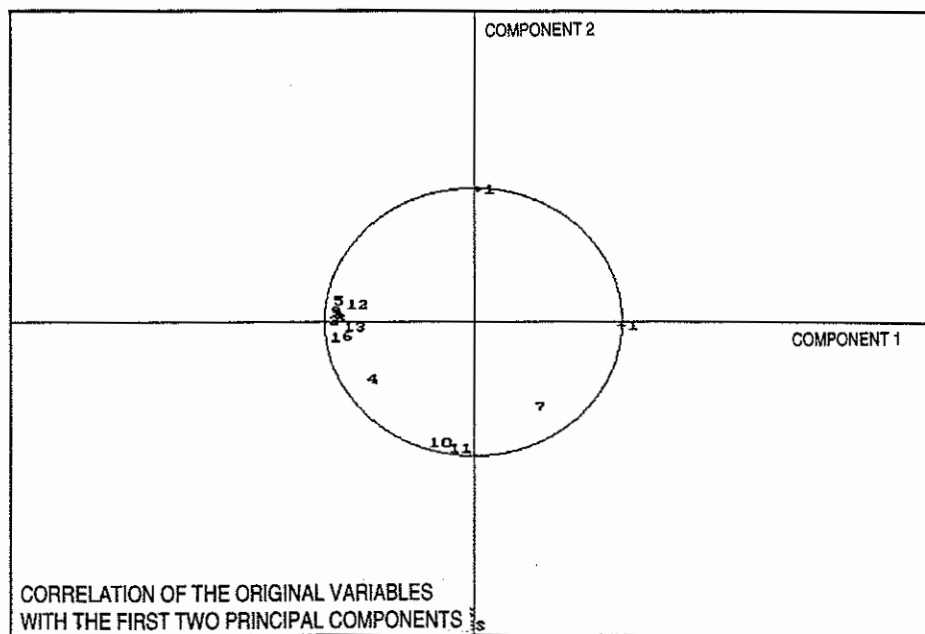


Figure 2: Correlation of the original variables: Flat projection of the first two principal components

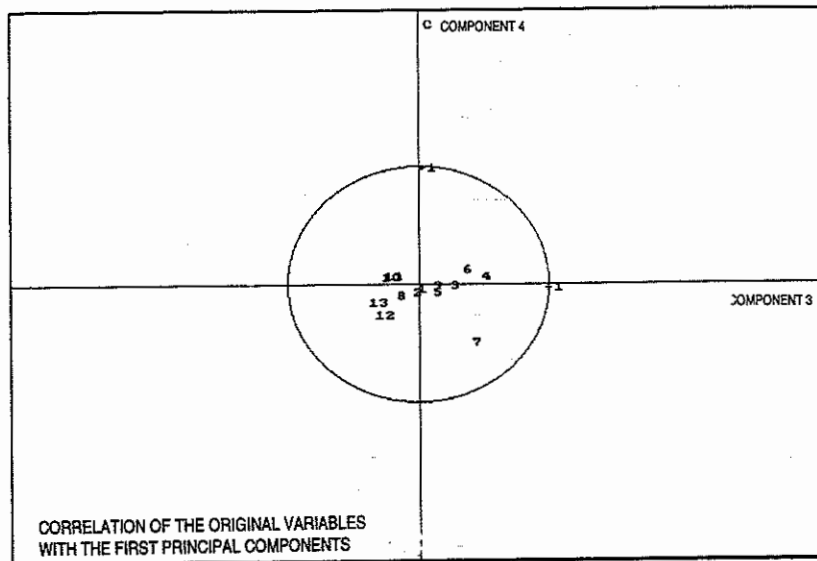


Figure 3: Correlation of the original variables: Flat projection of the principal components 3 and 4

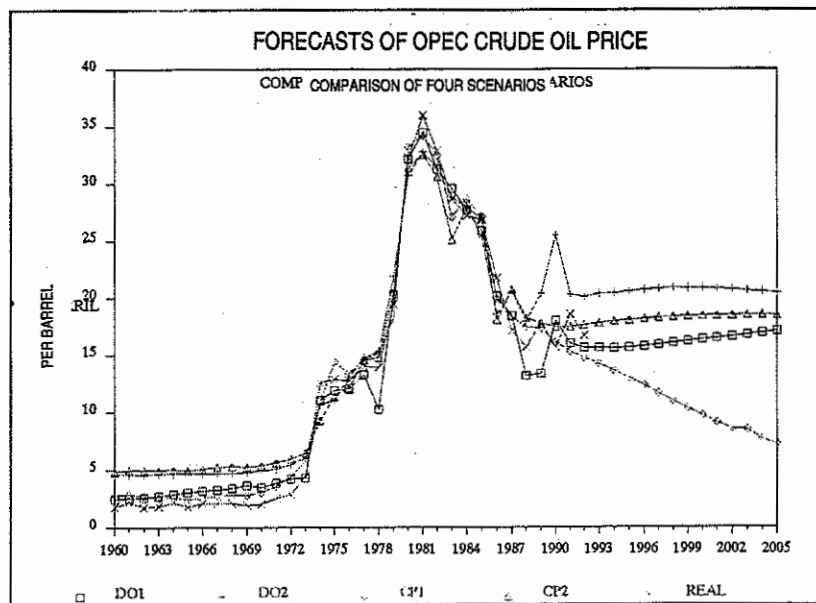


Figure 4: Results obtained with neural networks

- DO1 Trained network with observed real values and a tolerance of 5%
- DO2 Trained network with observed real values and a tolerance of 10%
- CP1 Trained network with preprocessed information with principal components and a tolerance of 5%
- CP2 Trained network with preprocessed information with principal components and a tolerance of 10%

DATA BASE USED FOR THE TRAINING

ANO	CPRF	PBLC	PIB	EXPT	EXNP	IMPR	BLNC	PRRF	CSRF	EXPTBR	PROD	EXRF	RSVA
1960	2,154.0	205851	27421	5149	2157	5454	-493	2,020.7	330.8	6,747.0	8,696.9	1,236.3	218,180.0
1961	2,181.6	210796	28025	5468	1814	5547	-735	2,072.6	349.9	7,407.6	9,363.8	1,222.0	223,834.0
1962	2,317.7	215380	29573	5931	1691	4822	356	2,334.0	358.9	8,344.2	10,508.1	1,316.4	231,980.0
1963	2,423.7	220545	32078	6752	1853	4994	991	2,221.1	379.3	9,250.9	11,514.2	1,477.8	248,075.0
1964	2,487.8	225973	33617	7751	2010	5727	589	2,178.8	426.7	10,617.2	12,986.9	1,517.8	254,950.0
1965	2,651.1	231324	36168	8447	1851	6802	247	2,385.6	488.6	12,768.3	14,739.2	1,637.9	260,345.0
1966	2,767.2	237360	37279	9030	1820	6581	616	2,451.7	512.7	13,507.7	15,770.2	1,626.8	290,068.0
1967	2,905.4	243191	40767	10005	2425	7101	671	2,500.8	542.8	14,239.5	16,849.3	1,667.5	310,425.0
1968	3,107.7	249522	44691	11620	2239	7885	256	2,757.9	603.5	16,076.7	18,785.6	1,717.4	332,030.0
1969	3,150.4	256043	51464	12678	2788	8915	-220	2,775.4	647.8	18,010.0	20,906.3	1,682.7	402,216.0
1970	3,565.4	262809	56884	14543	2742	9923	783	3,217.0	681.8	20,223.4	23,413.0	1,973.0	402,216.0
1971	3,663.6	270143	66497	19662	2607	11557	3311	3,252.9	744.7	22,031.7	25,326.3	1,814.5	430,820.0
1972	3,948.4	277853	78046	23542	3112	14217	4104	3,131.9	858.2	24,078.5	27,194.4	1,792.3	428,492.0
1973	4,128.2	285791	107642	37145	4660	20364	7025	3,465.1	1002.8	27,547.2	30,988.5	1,974.4	420,887.7
1974	4,290.6	294348	210653	120436	6398	33015	6653	3,386.6	1107.9	27,258.9	30,729.2	1,888.8	487,439.1
1975	4,488.1	303672	250285	107773	5994	52173	31457	3,041.9	1281.3	24,063.9	27,155.0	1,559.7	448,947.6
1976	4,637.7	312117	313522	130199	7470	64262	36854	3,416.9	1521.9	27,462.6	30,737.8	1,870.9	440,414.8
1977	4,972.3	321568	375258	141899	9568	85855	20173	3,613.8	1761.2	27,641.1	31,253.5	1,752.0	438,247.4
1978	5,189.3	331628	395527	134982	11322	97181	-3846	3,792.6	1922.1	26,088.7	29,805.3	1,905.5	443,793.9
1979	5,214.4	341472	492681	203193	16011	101154	60136	3,992.4	2177.8	26,764.9	30,928.8	2,141.6	436,179.0
1980	5,767.5	352650	641919	284491	22113	127266	105039	3,853.3	2429.8	22,864.1	26,880.0	2,008.6	435,955.9
1981	5,995.5	363083	709857	263337	21840	149865	52904	3,943.4	2819.4	18,568.8	22,598.6	1,979.3	438,371.5
1982	5,136.2	373843	734894	202789	21149	148957	2422	4,350.8	3043.6	14,296.1	18,992.2	2,225.7	466,252.8
1983	5,893.9	384861	718861	156945	22265	139742	-16595	4,493.2	3275	12,506.5	16,992.3	2,023.1	470,951.8
1984	6,179.5	396114	715836	143569	23586	122359	-3374	4,648.9	3156.6	12,043.9	16,347.2	2,281.3	510,411.2
1985	6,429.5	407987	709728	127182	23601	102112	7656	5,003.6	3199.2	10,926.4	15,447.7	2,368.2	536,976.6
1986	6,394.2	419602	686233	76741	25167	90482	-19889	5,334.4	3163.8	12,939.8	18,333.0	2,642.0	644,791.3
1987	6,530.0	432464	713210	92602	27002	89334	-4123	5,371.3	3270.9	12,060.4	17,324.9	3,017.2	676,488.7
1988	6,472.0	444403	621211	84731	31945	103961	-17115	5,914.7	3423.6	13,367.0	19,618.4	3,395.3	762,928.9

Table 2: Data calculated with the principal components

Año	CP1	CP2	CP3
1960	1.324	1.129	0.911
1961	1.303	1.059	0.842
1962	1.229	0.956	0.657
1963	1.168	0.867	0.476
1964	1.126	0.716	0.351
1965	1.036	0.521	0.106
1966	0.982	0.421	-0.032
1967	0.915	0.330	-0.163
1968	0.816	0.133	-0.405
1969	0.724	-0.085	-0.698
1970	0.556	-0.309	-1.076
1971	0.516	-0.554	-1.133
1972	0.459	-0.786	-1.174
1973	0.277	-1.169	-1.402
1974	0.086	-1.478	-0.631
1975	0.144	-1.049	-0.091
1976	-0.104	-1.418	-0.265
1977	-0.270	-1.432	-0.024
1978	-0.423	-1.084	-0.160
1979	-0.651	-1.463	0.492
1980	-0.890	-1.299	1.842
1981	-1.004	-0.637	2.046
1982	-1.028	0.148	1.546
1983	-1.173	1.223	0.354
1984	-1.129	0.657	0.945
1985	-1.160	0.744	0.761
1986	-1.441	1.444	-1.395
1987	-1.491	0.890	-0.518
1988	-1.899	1.523	-2.163

TABLE 4.- Values obtained with the neural network models

Year	Observed Value	Error percentages assigned for training							
		5% DO		10% DO		5% CP		10% CP	
		Value	Error %	Value	Error %	Value	Error %	Value	Error %
1989	17.75	13.4	15.94	20.38	14.82	17.32	2.42	17.75	0.00
1990	15.69	18.01	24.51	25.45	62.21	16.11	2.68	17.46	11.28
1991	18.57	16.02	14.79	20.28	9.21	15.32	17.50	17.48	5.87
1992	16.71	15.67	13.73	20.09	20.23	14.79	11.49	17.63	5.51

Table 5: Processing time used in the training

Type of training
Type of neural network
(# inputs/# outputs)

	1/1	13/1
Maximum	150'	6'
Minimum	5'	
Average	80'	
Total	17X80'	6'