

Un modelo *spline* para el pronóstico de la demanda de energía eléctrica

A Spline Model for Electricity Demand Forecasting

ANDRÉS FELIPE BARRIENTOS^a, JAVIER OLAYA^b,
VÍCTOR MANUEL GONZÁLEZ^c

ESCUELA DE INGENIERÍA INDUSTRIAL Y ESTADÍSTICA, FACULTAD DE INGENIERÍAS,
UNIVERSIDAD DEL VALLE, CALI, COLOMBIA

Resumen

El propósito de este trabajo es modelar, con fines de pronóstico, la demanda diaria de energía eléctrica en una región del suroccidente colombiano, mediante la implementación de modelos de regresión no paramétrica teniendo en cuenta factores de influencia tales como hora del día, día de la semana, mes y año, entre otros. Los datos empleados en el desarrollo de este proyecto provienen de una compañía local de distribución de energía eléctrica y se tomaron de Valencia (2005). La información disponible va desde enero de 2001 hasta noviembre de 2004. Estos datos muestran un comportamiento complejo, difícil de modelar con la teoría básica de los métodos paramétricos. Dado que un análisis exploratorio de la información sugiere la existencia de una curva típica diaria de demanda, se eligió estimarla utilizando modelos de regresión no paramétrica. Para efectos comparativos, se propuso la aplicación de otras metodologías que involucran modelos ARIMA y variables macroeconómicas. Todo el procesamiento estadístico se ejecutó con R.

Palabras clave: suavización, regresión no paramétrica, modelos ARIMA.

Abstract

Our goal is to model, with forecasting aims, the daily electricity demand in a southeast colombian region through a non-parametric regression model implementation. We consider some "calendar variables" such as time of the day, day of the week, month, and year, among others, on the estimation process. Data come from an electricity distribution local company and are taken from Valencia (2005). Available data go from January 2001 to November 2004. These data show such a complicated behavior that it becomes

^aProfesor auxiliar. E-mail: anfebar@pino.univalle.edu.co

^bProfesor titular. E-mail: olaya@univalle.edu.co

^cProfesor auxiliar. E-mail: vmgonzal@pino.univalle.edu.co

hard to model using classical parametric models. Since exploratory analysis suggested the existence of an electricity demand daily typical curve, we used non-parametric models instead. For comparison purposes, we made use of some other methodologies including ARIMA models and the insertion of macroeconomic variables. Statistical processing was run using R.

Key words: Smoothing, Non-parametric regression, ARIMA models.

1. Introducción

El constante desarrollo de los mercados de energía eléctrica genera escenarios cada vez más competitivos, por lo que es de crucial importancia poseer sistemas de distribución con planes de manejo bien estructurados. El objetivo es disponer de sistemas que suministren energía eléctrica con el menor número de interrupciones y al menor costo posible, mejorando así la calidad del servicio. Esta calidad depende en gran medida de la capacidad que tengan las empresas responsables de proveer y distribuir este servicio, de determinar a largo y corto plazo cuál será la demanda futura de energía. En este trabajo se utilizan datos de demanda horaria en megavatios (MW), recolectados por una entidad local de distribución que controla algunas zonas de la región objeto de estudio. Los datos provienen del trabajo de Valencia (2005).

El propósito de este trabajo es modelar y pronosticar la demanda horaria de energía eléctrica en una región del suroccidente colombiano mediante la implementación de modelos de regresión no paramétrica usando suavización *spline*, teniendo en cuenta factores de influencia tales como día de la semana, mes y año, entre otros. Dado que los registros de demanda de energía eléctrica son tomados a través del tiempo, se ajustan modelos de series de tiempo tipo ARIMA para generar pronósticos y realizar comparaciones con los obtenidos a partir del modelo propuesto. Para la aplicación y comparación de estas dos metodologías de pronóstico se tomará como ejemplo los días miércoles y para el modelo ARIMA la hora 20:00 de este mismo día.

2. Antecedentes

Varios han sido los métodos empleados para pronosticar la demanda de energía eléctrica. En general estos métodos pueden ser clasificados en tres grandes grupos: los primeros se basan en modelos estadísticos (regresión, series de tiempo y/o econométricos); los segundos en inteligencia artificial; y, finalmente, aquellos que dependen del juicio y la intuición humana. En Valencia (2005) se puede encontrar un cuadro detallado de cada uno de estos grupos, al igual que sus ventajas y desventajas.

Fernández (2006) desarrolla una aplicación de modelos de regresión. En este estudio la ecuación de demanda establece una relación no lineal entre el consumo doméstico de electricidad y el precio de ese consumo, así como un conjunto de características del hogar que se trate, personales y demográficas, de los individuos

que forman parte de la muestra. La estimación se realiza a través de un análisis paramétrico de mínimos cuadrados en dos etapas, así como de un análisis de regresión no paramétrica, mediante técnicas *kernel*, que permite fundamentar de forma más robusta la especificación funcional paramétrica.

Murillo et al. (2003) realizan un estudio usando modelos de tipo ARIMA. En este caso el pronóstico de demanda de energía eléctrica se basa en datos históricos suministrados por la Empresa de Energía de Pereira en el periodo comprendido entre el 1 de enero de 2001 y el 31 de diciembre de 2001.

Por otra parte, Blaconá & Abril (2000) realizan un estudio para el MEM (Mercado Eléctrico Mayorista) en Argentina. De acuerdo con los autores, las series relacionadas con la demanda de energía eléctrica contienen procesos estacionales muy complejos, que resultan difíciles de modelar con la teoría básica de series de tiempo. Por esta razón prefieren apoyarse en los MBEE (Modelo Básico de Espacio de Estado). La serie de tiempo de la demanda diaria promedio de energía eléctrica de Argentina contiene dos tipos de estacionalidad, una semanal y otra anual. La estacionalidad semanal resulta fácil de modelar con variables DUMMY, mientras la estacionalidad anual se modela mediante un modelo *spline* de regresión no paramétrica descrito por Poirier (1973).

Harvey & Koopman (1993) emplean también MBEE y la técnica *Time-Varying splines* para obtener un modelo de pronóstico horario de la demanda de electricidad de la compañía americana Puget Sound Power and Light. Poirier et al. (1979) realizaron un estudio para el Electric Power Research Institute, el cual involucró modelos de tipo econométrico y *splines* con el fin de estimar la demanda residencial para un pequeño conjunto de consumidores en Estados Unidos de América, de los cuales se contaba con información cada 15 minutos.

Modelos *spline* jerárquicos fueron empleados por Hendricks & Koenker (1992) para modelar la demanda de electricidad horaria de cuatrocientas residencias ubicadas en el área metropolitana de Chicago, medidas durante cuatro meses en 1985. Smith et al. (1998) modelaron 264 datos mensuales de demanda residencial tomados del trabajo realizado por Harris & Liu (1993), usando cuatro variables independientes e implementando una metodología basada en una aproximación bayesiana para regresión *spline* no paramétrica aditiva con errores autocorrelacionados.

En resumen, se han implementado distintas aproximaciones con el fin de realizar proyecciones de demanda de energía eléctrica. En todas ellas se ha considerado la inclusión de otras variables, las cuales contribuyen significativamente a la explicación de la demanda. Se deduce de esta revisión de la literatura que no se ha establecido un método estándar para solucionar este problema, debido a la complejidad que conlleva el mismo. Igualmente, se han empleado con frecuencia algunas técnicas de suavización para dar solución a este problema, algunas veces utilizadas como herramienta de apoyo a otras metodologías estadísticas de predicción. A partir de estas experiencias, este trabajo propone ajustar un modelo *spline* de regresión no paramétrica con este propósito.

3. Modelos de regresión no paramétrica

Los objetivos del análisis de regresión no paramétrica son los mismos de su contraparte paramétrica, vale decir, estimar y probar las características de la función de regresión.

(Olaya 2002)

La idea general de los modelos de regresión consiste en aproximar a partir de un conjunto de n datos (x_i, y_i) la función f en la siguiente expresión:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n$$

donde Y es continua, denominada generalmente *variable respuesta*; $X = (X_1, X_2, \dots, X_p)$ denota p variables continuas denominadas *covariables*, *variables predictoras*, *variables explicativas* o *variables independientes*; $f(x)$ es la respuesta media y referida en la literatura como *función de regresión* o *curva de regresión*; y ε se conoce comúnmente como *error* o *ruido* con $E[\varepsilon_i] = 0$ y $Var[\varepsilon_i] = \sigma^2 < \infty, \forall i$.

En el análisis de regresión paramétrica, el investigador presupone una forma de la función de regresión, de la cual solo se desconoce el valor de los parámetros asociados a la misma. Por el contrario, el análisis de regresión no paramétrica no asume un comportamiento del fenómeno a priori, sino que concibe la forma de la curva o función “a partir de lo que los datos digan”. La curva se escoge de entre un conjunto de curvas con ciertas propiedades de continuidad y diferenciabilidad.

En el caso de la regresión no paramétrica simple puede haber varias maneras de estimar $f(x)$. En este trabajo se empleó un estimador lineal de la forma $f_\lambda(x) = \sum_{i=1}^n K(x, x_i; \lambda) y_i$, donde $K(x, x_i; \lambda)$, $i = 1, 2, \dots, n$ es una colección de pesos (funciones) que depende de la covariable X y del parámetro de suavización λ . Así que se dispone de un estimador $f_\lambda(x)$ de $f(x)$ para cada λ , por lo que la elección del parámetro λ es decisiva.

El parámetro λ puede estimarse a partir de la función de *riesgo* del estimador, la cual puede expresarse como $R(\lambda) = B^2(f_\lambda) + var(f_\lambda)$, donde $B^2(f_\lambda)$ y $var(f_\lambda)$ representan el sesgo y la varianza del estimador $f_\lambda(x)$, respectivamente. El valor de λ que minimiza esta función provee una estimación de $f(\cdot)$ que puede ser empleada en una situación en general y de la cual se esperan obtener buenas predicciones. λ también podría estimarse usando el *riesgo de predicción* del estimador definido por, $p(\lambda) = n^{-1} \sum_{i=1}^n E[y_i^* - f_\lambda(x_i)]^2 = \sigma^2 + R(\lambda)$. Este indicador se emplea para evaluar la funcionalidad de un estimador como un predictor de observaciones futuras $y_1^*, y_2^*, \dots, y_n^*$.

Si se conoce σ^2 , entonces $R(\lambda)$ y $P(\lambda)$ se pueden estimar directamente; sin embargo, en la práctica este no es el caso y resulta necesario estimar σ^2 . En el marco de la regresión no paramétrica existe un buen número de estimadores de σ^2 que no dependen del parámetro de suavización y en su mayoría se apoyan en diferencias sucesivas de las respuestas y_i asociadas a una vecindad predeterminada de x_i . Un ejemplo de esta clase de estimadores es el presentado en la ecuación (1), denominado estimador de Gasser, Sroka and Jennen-Steinmetz (citado por Green & Silverman (2000)). Algunos autores denotan este estimador de la varianza como estimador GSJS; de esta manera se citará en este trabajo.

$$\hat{\sigma}_{\text{GSJS}}^2 = \sum_{i=2}^n \frac{\tilde{\varepsilon}_i^2}{n-2} = \sum_{i=2}^n \frac{\left[\frac{y_i - A_i y_{i-1} - B_i y_{i+1}}{1 + A_i^2 + B_i^2} \right]^2}{n-2} \quad (1)$$

En la ecuación (1), $A_i = \left(\frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} \right)$ y $B_i = \left(\frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} \right)$. Los $\tilde{\varepsilon}_i$ se llaman, en este contexto, pseudo-residuales. Aunque el estimador de varianza GSJS se presenta con mayor frecuencia en el contexto de suavización *kernel*, se debe tener en cuenta, como se mencionará en la sección 3.1, que un suavizador *spline* puede ser expresado como un suavizador *kernel*; por tal motivo las propiedades de este último son heredadas para la suavización *spline*. Por otra parte, es necesario considerar que existen algunos estimadores comúnmente usados en el contexto de regresión con *splines* (véase Wood 2006), tal como el presentado en la ecuación (2); no obstante Eubank (1999) menciona que existen algunos problemas debido a que este estimador está en función del parámetro de suavización, el cual puede ser calculado mediante la minimización del *riesgo de predicción*, que a su vez depende de σ^2 .

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{(y_i - f_\lambda(x_i))^2}{n - \text{tr}(S_\lambda)} \quad (2)$$

donde S_λ es una matriz equivalente a la matriz HAT en regresión paramétrica. Sin embargo, en algunos casos σ^2 es razonablemente difícil de estimar, por lo que se hace necesario el uso de métodos alternativos para estimar el parámetro de suavización (λ), tales como el *criterio de validación cruzada* $CV(\lambda)$ y el *criterio de validación cruzada generalizada* $CVG(\lambda)$, el cual no es una generalización del primero. Estos criterios están definidos por

$$CV(\lambda) = n^{-1} \sum_{i=1}^n \left[\frac{y_i - f(x_i)}{1 - s_i} \right]^2$$

y

$$CVG(\lambda) = n^{-1} \sum_{i=1}^n \left[\frac{y_i - f(x_i)}{1 - s_i} \right]^2 \left\{ \frac{1 - s_i}{n^{-1} \text{tr}[I - S_\lambda]} \right\}^2$$

En las expresiones para CV y CVG , s_i es el i -ésimo elemento de la diagonal de S_λ . El cálculo de estos criterios es computacionalmente intensivo, pero se han desarrollado rutinas muy eficientes que están disponibles en R (R Development Core Team 2006).

3.1. Suavización *spline*

Algunos autores (Wegman & Wriugh 1983, Green & Silverman 2000) describen un *spline* como una “delgada tira de madera” que se puede hacer pasar a través de un conjunto de puntos gracias a que posee una determinada energía de deflexión. En su mayoría, los *splines* han sido estudiados en el marco del análisis numérico como método de interpolación más que en el estadístico como método

de suavización. Eubank (1999, pp. 229-230) menciona que los *splines* tuvieron sus orígenes desde 1923 gracias a Whittaker, mientras que su formulación moderna fue planteada por Schönberg en 1964 y por Reinsche en 1967; finalmente su implementación en Estadística se da alrededor de 1990 por Wahba. En Estadística existen básicamente tres métodos de suavización *spline*, donde la aplicabilidad de cada uno de ellos depende de la estructura de los errores asociados a los datos.

3.2. Método *spline* - mínimos cuadrados penalizado

En regresión, una medida de bondad de ajuste para f está dada por la expresión $n^{-1} \sum_{i=1}^n (y_i - f(x_i))^2$, mientras que una medida de suavidad asociada con una función f que pertenece al espacio de Sobolev¹ W_2^m está dada por $\int_0^1 (f^{(m)}(x))^2 dx$ (penalización por aspereza). Este método usa ambas medidas para estimar una función *spline* única f_λ minimizando la siguiente combinación convexa:

$$(1 - q)n^{-1} \sum_{i=1}^n (y_i - f(x_i))^2 + q \int_0^1 (f^{(m)}(x))^2 dx \quad \text{para algún } 0 < q < 1$$

esto es equivalente a:

$$n^{-1} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx \quad (3)$$

En la ecuación (3), $\lambda = q/(q-1)$ es el parámetro de suavización y representa la tasa de intercambio entre el error residual y la variación local. El problema puede resumirse de la siguiente forma: si $f_\lambda(x)$ es un *spline* natural (definido bajo ciertas condiciones de diferenciabilidad de $f(\cdot)$ al evaluarla en el valor mínimo y máximo de la covariable), entonces el problema de minimizar (3) sobre todas las funciones en W_2^m se reduce a un espacio finito n -dimensional de los *splines* naturales. Eubank (1999) menciona que las de funciones Demmler-Reinch conforman una base del espacio de los *splines* naturales para la estimación de f , la cual se puede representar como un suavizador *kernel*; por tanto, las propiedades asintóticas establecidas para el método *kernel* son equivalentes para el método *spline*. Otra clase de base son los *B-splines* considerados como una generalización de las curvas Bezier, cuyas funciones se calculan de forma recursiva (véase Eilers & Marx 1996).

4. Base de datos

De acuerdo con Valencia (2005), el consumo de energía está distribuido principalmente en los sectores residencial (36.67%), industrial (9.98%) y comercial (6.64%). Para evitar confusiones en la base de datos, se consideró la unificación de la nomenclatura; por este motivo en las siguientes secciones de este documento, cuando se hable de la hora 0 se hará referencia al periodo comprendido entre

¹Un espacio de Sobolev provee condiciones generales suficientes de continuidad de las funciones elegibles como función de regresión.

la hora 00:00 y las 00:59, y así sucesivamente hasta la hora 23, que se refiere al periodo comprendido entre la 23:00 y las 23:59. La base de datos está conformada por los registros de la demanda horaria de energía durante los días de enero 1 de 2001 hasta noviembre 4 de 2004. Cada día registra 24 lecturas correspondientes a la hora 0, 1, 2, . . . , 23. Así, se tiene por ejemplo información de la hora 20 de cada miércoles a lo largo del periodo de análisis.

Con el fin de evaluar la calidad de los pronósticos de los modelos de regresión no paramétrica y ARIMA, se seleccionó como horizonte de pronóstico el mes de octubre de 2004. Para ajustar los modelos basados en suavización *spline* para un día de la semana en particular, la base de datos fue estructurada de tal forma que la covariable en el modelo fuese la hora del día y la variable respuesta los datos de demanda; es decir, por cada hora del día hay tantas observaciones como días de interés en el mes de análisis. Respecto a la metodología ARIMA se deben ajustar 24 modelos por día, uno por cada hora; en este caso cada serie estará conformada por los datos registrados desde enero de 2001 hasta septiembre de 2004. Por tanto, los cuatro o cinco primeros pronósticos de cada modelo estimarán la demanda horaria para determinado día durante todo el mes de octubre de 2004.

5. Resultados

El análisis exploratorio mostró que el consumo de energía se incrementa a través de los años, lo que convierte la predicción de la demanda horaria de energía eléctrica en un problema interesante. Además, la demanda horaria no pareció estar considerablemente influida por el mes, caso contrario al que ocurre con el día de la semana. En la figura 1 se puede observar que los mayores consumos se presentan los días martes, miércoles, jueves y viernes cuya distribución es similar, seguidos de los días sábado y lunes, siendo este último el día más variable. Finalmente, el domingo es el día donde el consumo de energía es menor, al igual que su variabilidad comparada con los demás.

En la figura 2 se observa que en general la demanda durante las horas de la madrugada es baja, luego se presenta un incremento a las 6 y 7 a.m. debido a la iniciación de actividades. Seguidamente se observa un crecimiento progresivo hasta las 12 del día donde presenta un pico; allí comienza un descenso que finaliza a las 6 p.m., hora en la cual el consumo se incrementa hasta las 8 p.m. cuando usualmente la demanda es máxima. A partir de las 9 p.m. la demanda desciende hasta finalizar el día. Este comportamiento seguramente obedece a los horarios establecidos para tomar las comidas, trabajar, estudiar y descansar. Nótese, además, que la distribución de la demanda en cada hora es similar, lo cual conduce a suponer que el consumo horario de energía tiene varianza constante. Esto último es un resultado importante para la validación de los supuestos de los modelos que fueron ajustados.

A partir del análisis exploratorio de la información se encontró que la demanda horaria de energía eléctrica en el Valle del Cauca posee un comportamiento característico condicionado por el día de la semana; además, fue posible concluir que los días martes, miércoles, jueves y viernes poseen comportamientos similares,

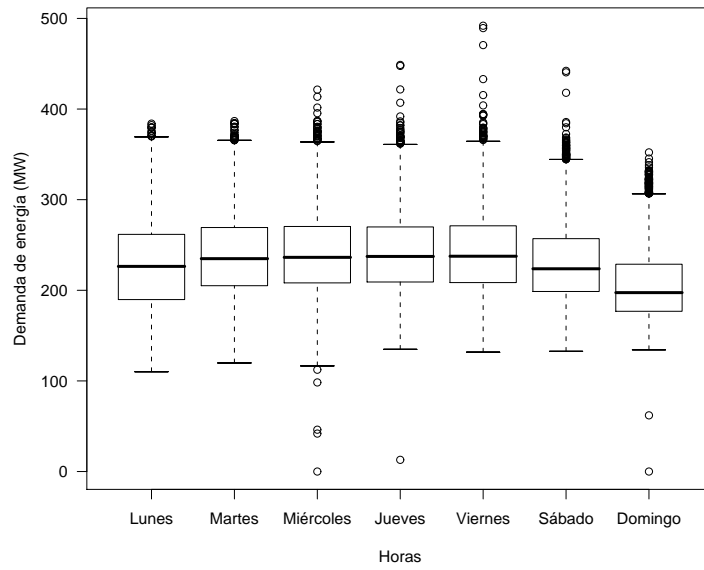


FIGURA 1: Diagrama de cajas y alambres de la demanda horaria de energía eléctrica considerando el día de la semana.

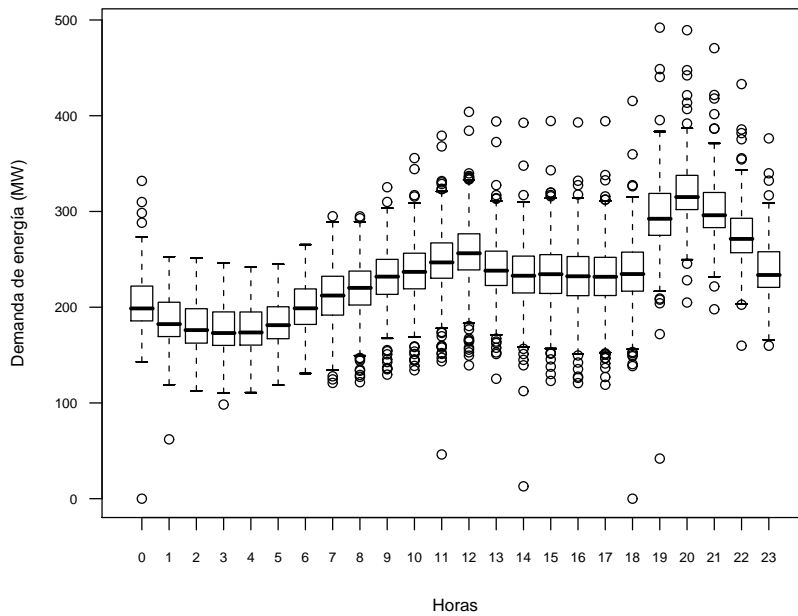


FIGURA 2: Diagrama de cajas y alambres de la demanda horaria de energía eléctrica considerando la hora del día.

razón por la cual se sugiere el agrupamiento de estos días para su posterior modelación. En la metodología planteada en este trabajo, se propuso usar la técnica de suavización *spline* para estimar la curva típica que represente la demanda de

energía de un conjunto de días en un periodo de tiempo dado, donde la variable predictora es la hora del día y la independiente es la demanda. El método *spline* seleccionado fue mínimos cuadrados penalizados calculados a partir de una base de funciones B-*splines*, donde el parámetro de suavización óptimo se obtuvo mediante el criterio de validación cruzada generalizada. El cálculo de las bandas del 95 % de confianza, asociadas a la curva estimada, se realizó asumiendo normalidad y usando el estimador de varianza $\hat{\sigma}_{\text{GSJS}}$, el cual fue comparado con $\hat{\sigma}$.

En el planteamiento y diseño de los modelos de pronóstico *spline* y ARIMA no se tuvieron en cuenta los factores regulatorios que definen las condiciones temporales reales del pronóstico. Los modelos basados en la técnica de suavización *spline* se usan para predecir a partir de la información disponible desde enero de 2001 hasta septiembre de 2004 la curva típica del día seleccionado en el horizonte de pronóstico.

Dado que la demanda de energía eléctrica es una variable que puede ser considerada de tipo macroeconómico, se espera que esta tenga algún tipo de relación posiblemente lineal con el Producto Interno Bruto (PIB). Por esta razón, el primer modelo de pronóstico planteado se basa en la tasa de crecimiento del PIB calculada para octubre de 2004. Al consultar la página web del DANE, se encontró que la tasa de crecimiento del PIB en Colombia para el último trimestre de 2004 fue del 4.415 %. Esta tasa fue hallada respecto al valor del PIB presentado en el mismo trimestre del año anterior. Obsérvese que el PIB no puede ser incluido como variable exógena explicativa para estimar la demanda, puesto que no está medido de forma horaria. Teniendo en cuenta estas consideraciones, el método de pronóstico consistió en estimar la curva típica de los días miércoles del mes de octubre de 2003 con sus respectivas bandas de confianza; luego sobre estas estimaciones se aplicó el 4.415 % y así se proyectó esta curva a octubre de 2004. Otro método de pronóstico consistió en estimar la curva típica con sus bandas de confianza del mes que precede al horizonte de pronóstico, en este caso septiembre de 2004, y luego usarla como pronóstico de octubre.

Los pronósticos de la curva de demanda horaria de energía eléctrica de los días miércoles del mes de octubre de 2004 fueron obtenidos usando la función `smooth.spline()` existente en el software estadístico R y son presentados en las figuras 3 y 4, donde se observa que al aplicar la tasa de crecimiento del PIB a la curva estimada para el mes de octubre de 2003, las bandas de confianza estimadas incluyen la curva típica y la mayoría de los registros de consumo del horizonte de pronóstico; de igual forma sucede con los pronósticos obtenidos a partir de las curvas estimadas para el mes de septiembre de 2004. Para el cálculo de la varianza fue necesario considerar que la estructura de los datos no permitía usar de forma directa el estimador GSJS, puesto que asociado a un valor de x_i (hora) hay diversos valores de y (demanda), es decir, existe un problema de medidas repetidas que hace compleja la implementación de este estimador. Por este motivo la variable respuesta fue ordenada de acuerdo con la fecha y hora en que fue tomado el registro. Según este enfoque, la variable independiente fue recodificada de la siguiente forma:

$$x = (x_1 = 1, x_2 = 2, x_3 = 3, \dots, x_n = n)$$

en este caso x_1 corresponde a la hora 00:00 del primero de los días considerados para la estimación de la curva típica, seguidamente x_2 equivale a la hora 01:00 de ese mismo día y así sucesivamente hasta x_n que representa la hora 23:00 del último día. La estimación de la varianza por medio de $\hat{\sigma}_{\text{GSJS}}$ utilizada para realizar los pronósticos presentados en las figuras 3 y 4 dio como resultado 116.5446 MW^2 y 127.9824 MW^2 , respectivamente. Estas estimaciones fueron comparadas con las obtenidas mediante el estimador $\hat{\sigma}$, dando como resultado 42.65659 MW^2 y 87.03438 MW^2 , es decir, los errores estándar obtenidos con $\hat{\sigma}_{\text{GSJS}}$ son mayores que con $\hat{\sigma}$. Para efecto de comparación con la metodología ARIMA se seleccionó el caso más crítico, esto es, las estimaciones derivadas de $\hat{\sigma}_{\text{GSJS}}$. Los estimadores $\hat{\sigma}_{\text{GSJS}}$ y $\hat{\sigma}$ fueron programados directamente en el software estadístico R.

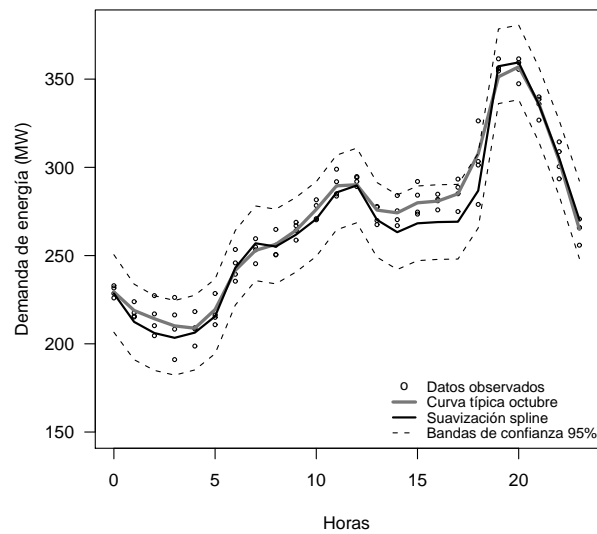


FIGURA 3: Pronóstico de la curva típica de demanda horaria de energía eléctrica de octubre de 2004 usando la tasa de crecimiento del PIB.

Respecto a la validación de los supuestos de incorrelación y valor esperado igual a cero en los errores de estos modelos según el método gráfico, se pudo observar que posiblemente estos supuestos se cumplen, dado que no se observó algún tipo de tendencia y los residuos se encuentran concentrados alrededor de cero. Al realizar la prueba de Levene y comparar el valor P con un nivel de significancia igual a 0.05, en ambos casos mostró que existe suficiente evidencia para no rechazar la hipótesis de homogeneidad de varianza de los errores en los modelos. En las figuras 5 y 6 se presentan los gráficos de los residuales asociados al modelo ajustado para el mes de septiembre de 2004 (caso similar al de octubre de 2003), en los cuales se puede observar que los residuales tienen media cero, varianza constante y no presentan un comportamiento de correlación seriada. No obstante, debido a que los datos han sido tomados a través del tiempo, se podría considerar la posibilidad de que los errores sean una secuencia estacionaria autocorrelacionada y, por tanto, modelarse mediante un proceso autorregresivo, tal como lo propone Smith et al. (1998); sin embargo la estructura de la base de datos necesaria para la implementación de

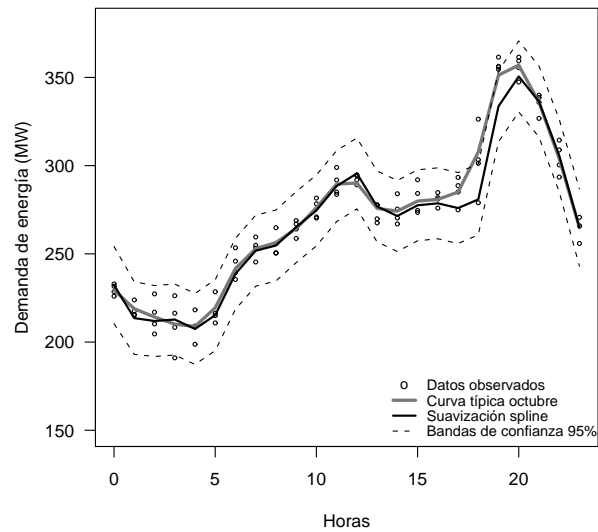


FIGURA 4: Pronóstico de la curva típica de demanda horaria de energía eléctrica de octubre de 2004 usando la curva obtenida en el mes de septiembre de 2004.

esta clase de metodologías no coincide con la usada en este problema, puesto que se requiere por cada punto de diseño un solo dato en la variable repuesta, y en este caso, por cada punto de diseño hay tantos datos como miércoles hay en el mes a suavizar. Trabajos más recientes, tales como el de Krivobokoa & Kauermann (2007) y Currie & Durban (2002), no presentan una alternativa de modelación que permita abordar el problema de regresión *spline* no paramétrica con errores autocorrelacionados y medidas repetidas por punto de diseño.

Una de las principales ventajas de usar modelos *spline* en la predicción de la demanda radica en que solo se necesita un modelo para realizar los pronósticos de un mes completo. Por su parte, el método basado en la tasa de crecimiento del PIB permite predecir la curva típica para periodos más cortos (trimestral, semestral) o más largos (superiores a un año) siempre y cuando se cuente con los pronósticos de la tasa de crecimiento para esos periodos, caso contrario al que sucede con el método que emplea el mes anterior al horizonte. Sin embargo, la exactitud de los pronósticos de la curva típica de demanda depende de la confiabilidad de la predicción de la tasa de crecimiento del PIB, lo cual se convierte en una desventaja teniendo en cuenta que determinar el PIB en el futuro es un problema que encierra mucha incertidumbre. Las predicciones obtenidas con el mes anterior al horizonte solo se pueden realizar siempre y cuando se cuente con todos los datos de demanda del mes. Esto último representa una de las principales desventajas de este método, debido a que los planes de manejo de las empresas generadoras de energía requieren las predicciones del horizonte antes de que este inicie. Una posible solución sería construir los modelos de pronóstico con un pequeño desfase, por ejemplo utilizando datos del 15 de un mes al 14 del siguiente.

El siguiente paso consistió en realizar pronósticos con modelos ARIMA. Para construir estos modelos, se debe dividir la información disponible para el día

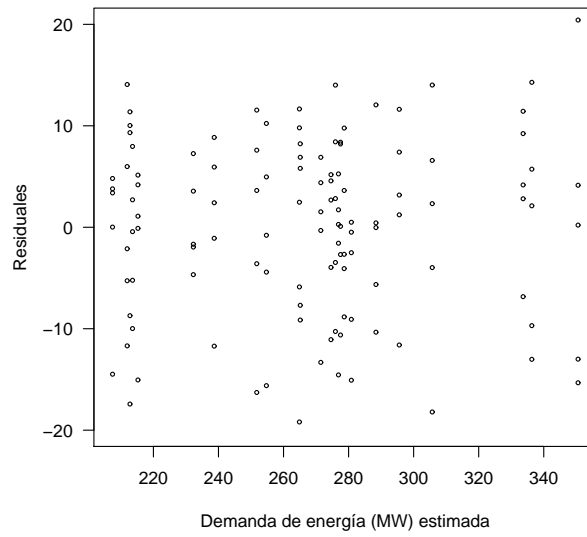


FIGURA 5: Residuales de la curva típica de demanda horaria de energía eléctrica estimada para el mes de septiembre de 2004 *vs.* demanda de energía (MW) estimada.

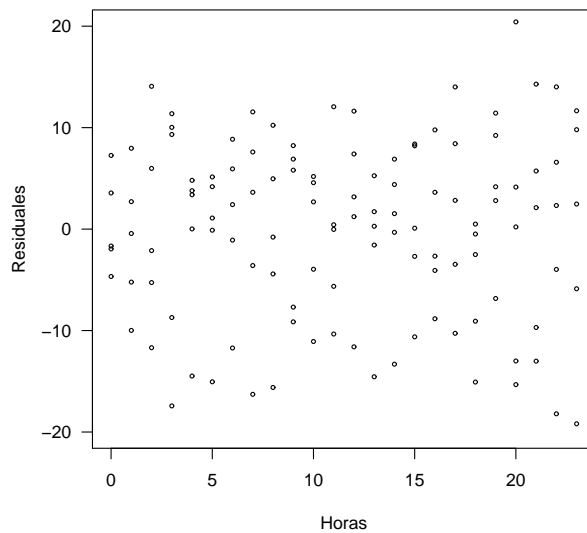


FIGURA 6: Residuales de la curva típica de demanda horaria de energía eléctrica estimada para el mes de septiembre de 2004 *vs.* hora del día.

miércoles en 24 series, de tal forma que cada una de ellas represente una hora específica. Nótese que los pronósticos no corresponderían a la curva típica asociada al horizonte de predicción establecido, ya que los modelos han sido construidos con los datos puros; por tanto, es necesario realizar tantos pronósticos como miércoles tenga este mes.

Por motivos prácticos y para realizar la comparación con los modelos *spline* solo se ajustó un ARIMA para una hora específica, en este caso la hora 20:00. Esta hora fue seleccionada porque normalmente es en ese instante del día cuando se produce el pico máximo de demanda horaria, el cual es de gran interés para las empresas que conforman el mercado eléctrico. Al aplicar el método del mínimo coeficiente de variación, se determinó que la serie no requiere ser transformada para lograr su estabilidad en varianza; además, debe diferenciarse una vez para alcanzar estacionariedad en media. Para determinar los órdenes del modelo ARIMA se implementó la metodología propuesta por BOX-JENKINS basada en el comportamiento de las funciones de autocorrelación simple y parcial. Al analizar estas funciones se estableció que no presentan periodos de estacionalidad bien definidos, puesto que la serie está conformada para una misma hora y día, eliminando con ello los efectos diarios y horarios; sin embargo, el comportamiento de estas funciones permitió establecer que los órdenes de los polinomios autorregresivos y de promedios móviles son iguales a 3. En la ecuación (4) se presenta la estimación del modelo ARIMA(3, 1, 3) asociado a la serie de demanda de energía eléctrica de los miércoles a las 20:00 conformada con los registros recolectados desde enero de 2001 hasta septiembre de 2004. Las funciones de autocorrelación simple y parcial fueron estimadas usando el software R y las rutinas `acf()` y `pacf()`, respectivamente; la función `arima()` fue implementada para estimar los coeficientes asociados al modelo ARIMA propuesto.

$$\tilde{Y}_t = -0.0207\tilde{Y}_{t-3} - 0.7732\tilde{Y}_{t-2} + 0.3055\tilde{Y}_{t-1} - 0.4402a_{t-3} + 0.6167a_{t-2} - 0.8388a_{t-1} \quad (4)$$

donde Y_t es el valor de la demanda de energía en el tiempo t , $\tilde{Y}_t = Y_t - \mu$, μ es la media de la demanda y a_t es el residual en el tiempo t . Al comparar los pronósticos generados por este modelo con datos medios en el horizonte de pronóstico se encontró que en términos generales el error porcentual de pronóstico no fue superior al 2%, siendo este último un resultado satisfactorio. Los supuestos de incorrelación y normalidad en errores fueron validados y evaluados con la prueba de Durbin-Watson y Shapiro-Wilk's, respectivamente. El método gráfico mostró que la varianza de los errores parece ser constante a través del tiempo.

La desventaja que presenta la implementación de modelos ARIMA se encuentra en la necesidad de ajustar 24 modelos para realizar los pronósticos de un día, además de realizar la validación de los supuestos en los que se basa cada uno de estos modelos. Este hecho demuestra que los ARIMA en la modelación de la demanda de energía no son viables desde un punto de vista práctico, puesto que es una tarea complicada que requiere mucho tiempo. Además no es posible realizar la predicción de la demanda para las horas en las que no se cuenta con información.

Finalmente, al comparar los resultados en la hora 20:00 los errores porcentuales de pronóstico tanto de los *splines* como los ARIMA fueron en ambos casos satisfactorios (< 2%); sin embargo, el error estándar de los pronósticos generados por el modelo *spline* basado en la tasa de crecimiento del PIB (10.3 MW) y en el mes anterior al pronóstico (10.8 MW) son menores que los de los ARIMA (> 13.3 MW). Esto último conduce a ubicar esta técnica de suavización en posición más ventajosa comparada con este modelo de series de tiempo.

Dado que se contaba con información disponible 45 meses atrás (enero de 2001 a septiembre de 2004), se propuso una metodología que permitiera estimar la curva típica de demanda de energía eléctrica del mes de octubre de 2004, en la cual se usara la técnica de suavización *spline* para encontrar la curva característica de los días miércoles en cada uno de estos 45 meses, y luego emplearlas como una serie de tiempo que pronostique la curva asociada al horizonte de pronóstico. No obstante, considerar modelos de pronóstico obtenidos a partir de la combinación de modelos paramétricos y no paramétricos implicó un complejo problema teórico relacionado con la validez del modelo y con la estimación de la varianza del estimador. Este es un problema abierto.

Se debe mencionar que en el trabajo original los modelos aplicados y propuestos en este documento fueron aplicados para generar los pronósticos de todos los días de la semana. En el caso de los ARIMA se ajustaron 24 modelos por día.

6. Conclusiones

Los resultados parecen mostrar que la variable macroeconómica Producto Interno Bruto (PIB) tiene una importante relación con la demanda de energía eléctrica del Valle del Cauca, por lo que los modelos que consideran la tasa de crecimiento de esta variable lucen adecuados para el pronóstico de la demanda horaria en este caso particular. Una generalización de este resultado requeriría un estudio diferente.

La calidad de los pronósticos obtenidos a partir de los modelos *spline* basados en el mes anterior al horizonte indican que para realizar predicciones de la demanda de energía eléctrica del Valle del Cauca en el corto plazo es suficiente contar con información reciente. La implementación de modelos ARIMA no es operacionalmente eficiente para realizar pronósticos de demanda horaria en el corto plazo, ya que para realizar predicciones se requieren tantos modelos como horas-días por estimar; es decir, desde un punto de vista práctico, la dificultad se encuentra en la validación de los supuestos a tantos modelos. Este es un problema ideal para el uso de modelos no paramétricos.

Además, no es razonable plantear un modelo multivariante de series de tiempo como VAR (Vectores Autorregresivos) para un día en particular, puesto que cada serie horaria está conformada por datos de demanda registrados en tiempos diferentes; en otras palabras, estas series no son equivalentes en el tiempo.

Finalmente, utilizando criterios tales como el error absoluto medio porcentual, el error porcentual de pronóstico y el error estándar de pronóstico se determinó que las metodologías apoyadas en la técnica de suavización *spline* en general arrojaron resultados similares y los pronósticos fueron acertados respecto a las curvas típicas asociadas al horizonte preestablecido; entre tanto, los modelos ARIMA ajustados sobre los datos puros no realizaron en la mayoría de los casos un buen pronóstico de los días del mes de octubre de 2004. Los errores estándar de pronóstico de los modelos ARIMA son mayores que los generados por los modelos *spline*.

7. Posibles extensiones de este trabajo

Con el fin de mejorar estos procesos de estimación se establece alternativamente desarrollar en el futuro algunos trabajos tales como el planteamiento de modelos de regresión no paramétrica que consideren los factores regulatorios que definen las condiciones temporales reales del pronóstico de la demanda horaria de energía eléctrica en Colombia; también podría considerarse la inclusión en el análisis de intervalos de confianza de la tasa de crecimiento del Producto Interno Bruto para realizar pronósticos de demanda horaria de energía. Igualmente, se podría intentar ajustar un modelo de pronóstico de demanda considerando el impacto de variables climáticas y macroeconómicas. Otra opción sería evaluar el uso de modelos de regresión paramétrica, tales como polinomios, usando una metodología similar a la planteada en este trabajo. En este caso se debe tener en cuenta la validez de estos modelos de acuerdo con el número de grados de libertad efectivos y la falta de ajuste. Además, es necesario determinar metodologías que permitan analizar este problema usando modelos de regresión no paramétrica con errores autocorrelacionados. Finalmente, se podrían plantear modelos de series de tiempo no paramétrico.

Recibido: marzo de 2007

Aceptado: septiembre de 2007

Referencias

- Blaconá, M. T. & Abril, J. C. (2000), Modelo estructural de espacio de estado para la demanda diaria promedio de energía eléctrica en la república Argentina, *in* 'Trabajo Presentado en la Reunión de la Asociación Argentina de Economía Política (AAEP)', Asociación Argentina de Economía Política.
- *<http://www.aaep.org.ar/espa/anales/>
- Currie, I. & Durban, M. (2002), 'Flexible Smoothing with P-splines: An Unified Approach', *Statistical Modelling* **4**, 333–349.
- Eilers, P. & Marx, B. (1996), 'Flexible smoothing with B-splines and Penalties', *Statistical Science* **11**, 89–121.
- Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*, Marcel Dekker Inc., New York.
- Fernández, L. (2006), La demanda residencial de electricidad en España: un análisis microeconómico de la demanda eléctrica residencial de corto plazo en España, Trabajo de doctorado en teoría económica y métodos cuantitativos, Universidad de Vigo, Facultad de Economía. Departamento de Teoría Económica y Econometría, Vigo (España).
- *<http://webs.uvigo.es/vijpe/pdf/FERNANDEZ.pdf>
- Green, P. J. & Silverman, B. W. (2000), *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, New York.

- Harris, J. & Liu, L. (1993), 'Dynamic Structural Analysis and Forecasting of Residential Electricity Consumption', *International Journal of Forecasting* **9**(4), 437–455.
- Harvey, A. & Koopman, S. J. (1993), 'Forecasting Hourly Electricity Demand using Time-Varying Splines', *Journal of American Statistics Association* **88**(424), 1228–1236.
- Hendricks, W. & Koenker, R. (1992), 'Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity', *Journal of the American Statistical Association* **87**(417), 58–68.
- Krivobokoa, T. & Kauermann, G. (2007), 'A Note on Penalized Smoothing with Correlated Errors', *Journal of the American Statistical Association* . (en prensa).
- Murillo, J., Trejos, . & Carvajal, P. (2003), 'Estudio del pronóstico de la demanda de energía eléctrica, utilizando modelos de series de tiempo', *Scientia et Technica* **23**, 37–40.
- Olaya, J. (2002), Suavización y regresión no paramétrica. Santiago de Cali, Universidad del Valle, Facultad de Ingenierías, Escuela de Ingeniería Industrial y Estadística.
*<http://pino.univalle.edu.co/jolaya/Seminario/Suynopar2.pdf>
- Poirier, D. L. (1973), 'Piecewise Regression using Cubic Splines', *Journal of the American Statistical Association* **68**(343), 514–524.
- Poirier, D. L., Hendricks, W. & Koenker, R. (1979), 'Residential Demand for Electricity: An Econometric Approach', *Journal of Econometrics* **9**, 33–57.
- R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Viena, Austria. ISBN 3-900051-07-0.
*<http://www.R-project.org>
- Smith, M., Wong, C. & Kohn, R. (1998), 'Additive Nonparametric Regression with Autocorrelated Errors', *Journal of the Royal Statistical Society* **60**(2), 311–331.
- Valencia, A. L. (2005), Diagnóstico del modelo de pronóstico de demanda de potencia y energía eléctrica de EPSA, Tesis de pregrado, Ingeniería Eléctrica, Universidad del Valle, Facultad de Ingenierías. Escuela de Ingeniería Eléctrica y Electrónica.
- Wegman, E. J. & Wright, I. W. (1983), 'Splines in Statistics', *Journal of American Statistical Association* **78**(382), 351–365.
- Wood, S. (2006), *Generalized Additive Models: An introduction with R*, Chapman & Hall, Florida.