

ENERCAST - PREVISIÓN DE DEMANDA DE ENERGÍA ELÉCTRICA

Jairo Cugliari¹; Andrés Castrillejo; Fernando Massa²; Ignacio Ramirez³

RESUMEN

Palabras clave: *energía eléctrica; predicción de demanda; estadística aplicada; R*

1 Introducción

ENERCAST es una herramienta desarrollada en el software estadístico R (R Core Team, 2014) que implementa algunas de las aproximaciones clásicas y otras más recientes para la previsión de demanda de energía eléctrica de corto plazo.

A continuación se presenta el problema de predicción y algunas de sus características, para luego en la sección 2 revisar algunos de los abordajes presentes en la literatura.

En la sección 3 se describe esta primera versión del paquete de R donde se implementan cuatro familias de modelos y en la sección 4 se presentan algunos resultados de la aplicación con de Uruguay. Finalmente en la sección 5 se comentan algunas conclusiones y perspectivas.

1.1 Predicción de demanda de corto plazo

En la predicción de corto plazo de la curva de carga (*Short-Term Load Forecasting - STLF*) la variable de interés a predecir es la carga horaria total del sistema. También son de interés los picos diarios o valores a ciertas horas del día.

El comportamiento del sistema está determinado principalmente por cuatro tipos de factores: económicos (demográficos, niveles de actividad industrial, cambios en determinados sectores, tendencias económicas), temporales (efectos estacionales, ciclos semanales y diarios, días feriados y festividades religiosas), climáticos (calefacción, acondicionamiento de aire, riego agrícola), efectos aleatorios (Gross and Galiana, 1987).

La importancia del problema tiene componentes económico-financieros, de planificación de corto y mediano plazo y operativas por la necesidad de equilibrio del sistema de carga. Adicionalmente, frente al advenimiento de las fuentes de generación autodespachadas, cobra especial importancia en la toma de decisiones de producción y distribución de energía eléctrica.

1.2 Algunas características de la curva de carga

Recordemos brevemente algunas de las características sobresalientes de la curva de carga eléctrica.

- la demanda de electricidad es altamente dependiente de las condiciones meteorológicas, como la temperatura, la cobertura nubosa o la intensidad y dirección del viento. Estos fenómenos inducen su ciclo estacional al consumo eléctrico.

¹Laboratoire ERIC, ICOM, Université Lumière Lyon 2.

²Instituto de Estadística, FCEA, UdelAR.

³Instituto de Ingeniería Eléctrica, FING, UdelAR.

- la demanda de electricidad es afectada por fenómenos sociales y económicos como por ejemplo los periodos vacacionales, la dicotomía entre días hábiles y fines de semana, o el crecimiento económico.
- la estructura semanal del calendario produce un ciclo semanal donde los perfiles de la carga eléctrica varían no sólo en el nivel medio sino también en la forma de las curvas diarias. Además, es posible distinguir diferencias al interior de la semana, especialmente en los días próximos a los fines de semana. A nivel diario, el perfil de la carga esta guiado por una clara diferencia en los niveles de carga nocturnos y diurnos.

Estacionalidad. La demanda de electricidad tiene fluctuaciones estacionales, esencialmente debidas a condiciones del clima (calefacción y acondicionamiento de aire), cambia con las condiciones del clima (temperatura y numero de horas de luz diarias). A su vez el clima afecta la oferta de electricidad, en algunos países como el nuestro, la producción hidráulica depende fuertemente de las precipitaciones. Esto puede visualizarse en los datos a través de

1. comportamiento anual ('sinusoidal')
2. Variabilidad intra-semanal, vinculada a la estructura de días laborables y fin de semana
3. Variabilidad intra-día, con pico matutino y en la tardecita.

Independientemente de que aproximación o modelización se elija, antes que nada, debe testearse cual es el comportamiento estacional de los datos. Esto se puede hacer de dos maneras complementarias, midiendo la correlación serial en el dominio del tiempo a través de la función de autocovariancia y a través de la descomposición de Fourier en el dominio espectral.

Algunos factores que afectan la curva de carga. Dos factores importantes vinculados a la predicción de la curva de carga son: la calidad de los datos y la posibilidad de incorporar factores exógenos en los modelos (Weron, 2006).

La influencia de los factores temporales como el momento del año, el día de la semana o la hora del día, se debe a los diferentes patrones de consumo en cada uno de ellos, así los fines de semana difieren de los días laborables y entre estos los lunes y viernes suelen tener diferencias con el resto, las vacaciones, feriados u otros tipos de días también tienen sus particularidades.

Las climáticas son, además de los factores temporales, las variable exógenas mas influyente en el caso de la STLF. Dentro de estas la temperatura y la humedad son las mas utilizadas, aunque otras como la nubosidad o la velocidad del viento pueden tener un rol especifico cuando se trata de hacer predicción en zonas particulares o en algunos momentos del tiempo. Algunas veces se utilizan funciones o índices de las variables climatológicas en vez de estas directamente.

Otros factores, como los precios o tarifas diferenciales y la actividad industrial también tienen influencia en el comportamiento de la curva de carga. En nuestro país, por ejemplo, a nivel de zonas hay efectos locales cuando se comienza a regar el arroz, cuando algunos de los grandes consumidores industriales tiene un cambio de comportamiento o de acuerdo a los movimientos turísticos.

2 Antecedentes y Metodología

En general, todos los abordajes para la modelización propuestos en la literatura para este tipo de datos intentan capturar las características mencionadas anteriormente en 1.2.

La bibliografía científica disponible sobre la predicción de la curva de carga eléctrica es el resultado de contribuciones tanto académicas como industriales e interdisciplinar, convergiendo aportes desde la economía, ciencias de la computación, ingeniería eléctrica, estadística, etc.

De acuerdo a la disciplina de la que provienen y las técnicas utilizadas los métodos pueden clasificarse en dos grandes áreas u orígenes, métodos estadísticos y métodos de inteligencia artificial (Weron, 2006) (Hong, 2014), aunque la distinción es algo arbitraria. Entre los llamados métodos basados en Inteligencia Artificial se encuentran las redes neuronales artificiales (ANN), la lógica difusa (fuzzy), los sistemas expertos y máquinas de vectores soportes (SVM). Dentro de la aproximación estadística, existen *grosso modo* cuatro grupos de aproximaciones metodológicas a la predicción del consumo eléctrico basados en:

1. el análisis de series temporales;
2. la búsqueda de patrones similares;
3. el análisis de regresión; y
4. algoritmos de aprendizaje automático.

Las diferencias en los abordajes radican en los supuestos necesarios para asegurar la validez de cada modelo. Los eventuales distanciamientos de los datos respecto a éstos supuestos impactan en la calidad de las predicciones que se obtienen en cada uno de ellos. Una dificultad adicional aparece con el advenimiento de nuevas tecnologías de producción, donde la distribución y transmisión son concentradas en un responsable de equilibrio del sistema pero la generación eléctrica pasa a ser descentralizada.

2.1 Antecedentes nacionales

A nivel nacional existieron algunos trabajos, a partir de convenios entre la UdelaR y UTE, que se remontan a principios de la década del 1990, en donde se estudió la demanda de energía eléctrica.

Recientemente el equipo liderado por B. Lanzilotta desde CINVE con el apoyo de ANII ha elaborado modelos más modernos con para la predicción de corto plazo de la curva de carga con un abordaje de series temporales (Lanzilotta and Rodriguez, 2014).

2.2 Antecedentes internacionales

Una referencia general con una aproximación estadística es el libro de R. Weron (Weron, 2006) que sintetiza y refiere los problemas asociados a la predicción tanto del consumo como del precio de la energía eléctrica.

Si se observan algunas revisiones de métodos (Gross and Galiana, 1987; Alfares and Nazeeruddin, 2002; Taylor, de Menezes and McSharry, 2006; Singh, Ibraheem, Khatoon, Muazzam and Chaturvedi, 2012; Weron, 2006; Hong, 2014) de las últimas dos décadas puede observarse la evolución en las técnicas utilizadas y algunos cambios en aspectos vinculados al problema de STLF.

Si antes de los 90 ((Gross and Galiana, 1987)) los métodos utilizados estaban centrados en

1. regresion lineal multiple,
2. series de tiempo (familia de modelos ARIMA)
3. alisado exponencial,
4. modelos de espacio mediante estados,
5. Knowledge-Based Approach,

durante esa década se proponen fundamentalmente métodos basados en inteligencia artificial, principalmente asociados a la creciente potencia computacional disponible (Alfares and Nazeeruddin, 2002). Se proponen aproximaciones que (Singh et al., 2012) denomina *Soft Computing Techniques*:

1. mínimos cuadrados iterativos y reponderados,
2. métodos adaptativos,
3. modelos ARMAX basados en algoritmos genéticos,
4. lógica difusa (fuzzy logic),
5. redes neuronales.

2.3 Algunos modelos utilizados

Algunos de los métodos mencionados se describen muy brevemente a continuación

- Regresión lineal múltiple. La demanda Y_t en el momento t se escribe mediante la relación lineal

$$Y_t = \beta X_t + \varepsilon_t,$$

donde X_t es el vector de variables explicativas (e.g. tiempo, temperatura, otras variables meteorológicas, día laborable, etc.), β es el vector de coeficientes y ε_t es el error del modelo en el momento t . Es una modelización popular gracias a su simpleza y a su prestacion en terminos predictivos. Los problemas que se plantean en algunas referencias están particularmente en las estaciones de transición (otoño y primavera).

- Análisis de Series Temporales. Esta aproximación combina los valores pasados de la serie utilizando en general una forma paramétrica y lineal para obtener la predicción. Los modelos utilizados son variantes de ARMA (*Autoregressive Moving Average*), ARIMA, SARIMA, ARIMA doble estacionales, etc. El principal inconveniente de esta estrategia es la incorporación de un efecto mecánico inducido por la forma paramétrica, lo que conduce a mayores errores de predicción especialmente durante los periodos de alta volatilidad del consumo eléctrico. Los a veces llamados modelos ARMAX, permiten incorporar covariables explicativas (X) en la especificación del modelo ARMA o sus variantes.

La aproximación de (Cancelo, Espasa and Grafe, 2008) es manejar 24 modelos (por cada hora del día) y hacer la predicción del día siguiente a un paso para cada modelo.

Una alternativa es utilizar métodos de *alisado exponencial* (Weron, 2006). En su forma más simple la previsión de la demanda se obtiene mediante una formula de recurrencia a partir de la previsiones pasadas. En series horarias se suelen terminos adicionales para tener en cuenta la doble estacionalidad (diaria y semanal) (Taylor et al., 2006; Taylor, 2003).

2.4 Otras propuestas metodológicas

A continuación enumeramos una serie de referencias que presentan aspectos innovadores en el modelado de la demanda eléctrica.

- GEFCom2012 La *Global Energy Forecasting Competition 2012* organizada por el grupo sobre predicción de energía de IEEE, muestra que hay mas técnicas que pueden ser exitosas en la resolución del problema de predicción de corto plazo de la curva de carga. Por detalles del problema puede verse (Hong, Pinson and Fan, 2014).

Particularmente los cuatro mejores resultados, en términos de menor error predictivo, fueron publicados en un número especial del *International Journal of Forecasting* y contienen ideas de *Machine Learning* como *Gradient Boosting Machines* y regresión con Procesos Gaussianos (Lloyd, 2014), refinamientos sobre combinaciones de modelos paramétricos de regresión (Charlton and Singleton, 2014), modelos aditivos generalizados (Nedellec, Cugliari and Goude, 2014) y boosting sobre modelos de regresión (Ben Taieb and Hyndman, 2014).

- (Hoverstad, Tidemann, Langseth and Öztürk, 2015) proponen una aproximación en tres etapas al STL. El objetivo es simplificar y automatizar las estimaciones. Para ello recurren a un pre procesamiento de la series temporales con el fin de eliminar patrones diarios y semanales utilizando datos históricos. Luego, se estima el modelo deseado con los datos procesados, se predice el consumo y se re incorpora el patrón estimado para lograr la predicción final.

En la aplicación se utilizan tres conjuntos de datos a tres niveles distintos, datos de distribución (Noruega), datos de transmisión (Canadá) y los datos a nivel del sistema de carga de la competencia GEFCom2012. Y se prueban varios modelos : *ARIMAX*, *Double Seasonal Holt Winter*, *Echo State Network*, *Wavelets*, *Hong's Vanilla Benchmark* y Modelos vectoriales (24 modelos diarios estimados con algunos de los métodos anteriores).

En (Hoverstad, Tidemann and Langseth, 2013) también ponen énfasis en la etapa de pre-procesamiento de los datos y sus resultados en la precisión de la predicción.

- (Singh and Srivastava, 2014) propone un abordaje de Redes Neuronales Artificiales (ANN) y destacan sus ventajas, fundamentalmente: adaptabilidad y operación en tiempo real. La red que plantean (no muestran ningún resultado experimental) tiene una sola capa oculta. El vector de inputs tiene 27 nodos: la carga de electricidad de 24 horas y temperatura máxima, temperatura mínima y otra variable climática.
- Los Modelos Aditivos Generalizado (GAM) (Wood, 2006) han sido aplicados con cierto suceso, por ejemplo en (Wood, Goude and Shaw, 2015; Pierrot and Goude, 2011) a datos de Francia y como se mencionó en la GEFCom2012 (Nedellec et al., 2014). Las ventajas de este tipo de modelos semi-parametricos con respecto al modelo lineal multiple es su flexibilidad para estimar el vinculo entre la demanda y las variables explicativas. Además su sencillez y claridad los hacen preferibles frente a los métodos de tipo redes neuronales. (Goude, Nedellec and Kong, 2014) Mas recientemente los han aplicado a nivel de subestaciones construyendo automáticamente para cada una un modelo de este tipo. También han sido utilizados para extraer tendencia en la construcción de métodos híbridos (Cho, Goude, Brossat and Yao, 2013).

- Las aproximaciones híbridas consisten en tomar mas de un tipo de modelo para niveles diferentes la curva de carga. Por ejemplo (Cho et al., 2013) extraen la tendencia de largo plazo con un GAM sobre los promedios semanales de demanda y luego utilizan regresión lineal funcional.
- Una serie de trabajos han puesto en evidencia la utilidad de representar la curva de carga mediante técnicas de datos funcionales ((Cugliari, 2011; Vilar, Cao and Aneiros, 2012; Antoniadis, Brosat, Cugliari, J and Poggi, 2012; Antoniadis, Brossat, Cugliari and Poggi, 2013)). La previsión se puede obtener en estos casos mediante técnicas semi o no paramétricas.
- De manera alternativa a la construcción de un solo modelo de predicción (Goude, 2008) explora la agregación de varios modelos (o expertos) mediante la combinación en línea de predictores. Trabajos en esta dirección han mostrado los beneficios teóricos de la agregación (no hay pérdida de performance en las previsión así obtenidas) así como mejoras significativas desde el punto de vista practico (Devaine, Gaillard, Goude and Stoltz, 2013).
- Finalmente, mencionamos un abordaje reciente es la utilización de Bosques Aleatorios para la predicción de la demanda (Dudek, 2015a).

3 El paquete ENERCAST

El paquete enerblast en su versión inicial (disponible en la plataforma especializada Github) propone la implementación de por lo menos un modelo de cada grupo ya mencionado dentro de la aproximación estadística al principio de la sección 2. Esto es: regresión (3.1), de series temporales (3.2, 3.3), de aprendizaje automático (3.4) y de reconocimiento de patrones (3.5).

3.1 Modelo de regresión - HVB

Modelo. T. Hong (Hong, Wang and Willis, 2011) ha propuesto como referencia o *benchmark* un sencillo modelo de regresión lineal múltiple, conocido como *Hong Vanilla Benchmark*.

El modelo incluye un cierto número de interacciones aditivas, donde la variable central es la Temperatura.

$$\begin{aligned} E(Load) = & \beta_0 + \beta_1 Trend + \beta_2 Day \times Hour + \beta_3 Month + \beta_4 Month \times Temp \\ & + \beta_5 Month \times Temp^2 + \beta_6 Month \times Temp^3 + \beta_7 Hour \times Temp \\ & + \beta_8 Hour \times Temp^2 + \beta_9 Hour \times Temp^3 \end{aligned}$$

donde *Hour*, *Day* y *Month* son las variables vinculadas al tiempo que actúan como categorías; *Trend* es una secuencia numérica que describe una tendencia lineal y *Temp* es la temperatura horaria; y sus interacciones.

Construcción. La implementación informática (*hong* y *predict.hong*) recurre de forma directa a la función `lm` del R.

Input/Output. A la función se le proporciona un `data.frame` con las variables `'year'`, `'month'`, `'wday'`, `'hour'`, `'temp'`, `'load'`.

3.2 Modelo de series de tiempo - TSB

Modelo Time Series Benchmark. Se considera como *benchmark* la siguiente generalización del modelo *ARIMA*.

$$\Phi_{P_2}(L^{s_2})\Phi_{P_1}(L^{s_1})\phi_p(L)\nabla_{s_2}^{D_2}\nabla_{s_1}^{D_1}\nabla^d(Y_t - X_t\beta) = \Theta_{Q_2}(L^{s_2})\Theta_{Q_1}(L^{s_1})\theta_q(L)\varepsilon_t$$

donde Y_t es el valor de la serie de demanda eléctrica en el momento t , X_t es el valor de las variables exógenas en t , β es un vector de coeficientes de regresión y ε_t es un ruido blanco. En cuanto a los polinomios autoregresivos, $\Phi_{P_2}(L^{s_2})$ es el correspondiente a los movimientos vinculados con la estacionalidad de período s_2 (que en este caso, es la estacionalidad semanal, por ende $s_2=168$), $\Phi_{P_1}(L^{s_1})$ es el correspondiente a los movimientos vinculados con la estacionalidad de período s_1 (que en este caso, es la estacionalidad diaria, por ende $s_1=24$) y $\phi_p(L)$ corresponde a la dinámica inmediata de corto plazo. Análogamente se definen los polinomios de medias móviles $\Theta_{Q_2}(L^{s_2})$, $\Theta_{Q_1}(L^{s_1})$ y $\theta_q(L)$.

Al variar los valores de los índices $p, d, q, P_1, D_1, Q_1, P_2, D_2$ y Q_2 se pueden representar diversas dinámicas. Típicamente los valores de los operadores de diferenciación d , D_1 y D_2 adoptan los valores 0 o 1, mientras que los valores de los índices autorregresivos y de medias móviles pueden adoptar valores un poco mayores.

Esta alternativa difiere de la comúnmente utilizada para series de esta frecuencia donde el análisis se sub-divide en 24 series (una para cada hora de día) independientes y la predicción de

un día resulta de la predicción a un paso de 24 modelos (esta metodología es la propuesta por en(Cancelo et al., 2008))

Construcción. La estimación de los parámetros se realiza mediante un método basado en el *método de los momentos* que parte de la idea de que las regularidades de la serie (adecuadamente transformada para ser estacionaria) se representan en su función de autocorrelación ρ_k . Por lo cual se trata de estimar el conjunto de parámetros que reproduzcan lo más fielmente posible la función de autocorrelación muestral $\hat{\rho}_k$. Siendo η el vector de todos los parámetros del modelo, se intenta minimizar la función:
$$\sum_{k=1}^{k=K} (\hat{\rho}_k - \rho_k(\eta))^2.$$

Ésta alternativa representa un gran incremento en cuanto a velocidad y ha probado ser una buena aproximación a los estimadores máximo verosímiles.

La variante implementada es un modelo sin variables exógenas.

Input/Output. A la función τ_{sb} se le proporciona un vector con la serie de demanda observada, los parámetros para el modelo $p, d, q, P_1, D_1, Q_1, P_2, D_2, Q_2$ y las frecuencias estacionales s_1 y s_2 .

3.3 Modelo de series de tiempo - SSM

Modelo de Espacio de Estados. Estos modelos (*SSM*) componen una familia de modelos que, entre otros, incluyen a los modelos clásicos de series temporales y los modelos de regresión, de esta manera, se puede decir que por su flexibilidad, son capaces de modelar diversas situaciones. Diversos *SSM* se pueden formular manipulando adecuadamente las siguientes 2 ecuaciones:

$$\begin{aligned} \alpha_t &= c_t + T_t \alpha_{t-1} + v_t \\ Y_t &= d_t + Z_t \alpha_t + w_t \end{aligned} \quad (1)$$

La primera de las dos ecuaciones expresadas en (1) es la llamada “ecuación de estados”. En la misma se describe la evolución temporal del vector no observable α_t a través de un vector de “control” c_t que afecta su media incondicional, una matriz T_t que afecta su media condicional y un vector de “ruido” v_t con matriz de covarianza H_t .

La segunda ecuación expresada en (1) es la llamada “ecuación de observación”. Es en esta ecuación que se modela la evolución de los datos efectivamente observados Y_t , d_t contiene elementos que afectan la media de las observaciones sin tener en cuenta de observaciones pasadas, Z_t vincula el vector de estados con el vector de observaciones y w_t es un vector aleatorio con matriz de covarianza G_t .

Dordonnat et al (Dordonnat, Koopman, Ooms, Dessertaine and Collet, 2008) proponen, entre otras alternativas, construir un modelo para la serie de demanda de energía eléctrica de cada hora del día. De esta manera la predicción de un día completo se logra al combinar la predicción a “un paso” de los 24 modelos involucrados. El modelo consta de una regresión lineal con parámetros que evolucionan temporalmente según un *random walk*. Las ecuaciones correspondientes serían las siguientes:

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + v_t \\ Y_t &= X_t^f \beta + X_t^a \alpha_t + w_t \end{aligned} \quad (2)$$

Donde X_t^f contiene variables de regresión cuyos coeficientes (β) se asume son constantes a lo largo del período de estudio, el supradice f hace referencia a que el vector de coeficientes β es de carácter *fiijo*. Por otro lado, X_t^a contiene las variables cuyos coeficientes (α) evolucionan

temporalmente según se describe en la ecuación de estados, el superíndice a hace referencia al carácter *aleatorio* de estos coeficientes.

Las variables contenidas en X_t^f son indicatrices que “marcan” la presencia de días especiales (feriados fijos, feriados móviles, carnaval, semana de turismo, etc). Las variables contenidas en X_t^a son de distinta índole:

- una constante
- una tendencia
- funciones sinusoidales en frecuencias múltiplos de 24 para captar variaciones de la demanda asociadas al momento del día.
- funciones sinusoidales en frecuencias múltiplos de 168 (24×7) para captar variaciones de la demanda asociadas al momento del semana.
- funciones sinusoidales en frecuencias múltiplos de 8760 (24×365) para captar variaciones de la demanda asociadas al momento del año.
- funciones asociadas a un “spline cúbico restringido” (Durrleman and Simon, 1989) para captar la posible relación no lineal entre la temperatura y la demanda.

Las funciones sinusoidales mencionadas anteriormente son de la siguiente forma:

$$\sum_{j=1}^{j=J} \beta_j \sin\left(\frac{2\pi jt}{s}\right) + \gamma_j \cos\left(\frac{2\pi jt}{s}\right) \quad (3)$$

Donde s representa la frecuencia que se pretende captar con el grupo de funciones sinusoidales. Según lo descrito anteriormente $s \in 24, 168, 8760$ mientras que J fue fijado en 3.

En cuanto al término spline, se lo especificó de modo cúbico y restringido para evitar posibles comportamientos “anómalos” cuando las temperaturas son muy bajas o muy altas. De esta manera, al incluir $K > 3$ o más nodos t_1, t_2, \dots, t_K para definir el spline, solo se necesitan $k - 2$ funciones (además de una constante y a la variable involucrada).

$$S_{K,k} = (X - t_k)_+^3 + \frac{t_K - t_k}{t_K - t_{K-1}} (X - t_{K-1})_+^3 + \frac{t_{K-1} - t_k}{t_K - t_{K-1}} (X - t_K)_+^3 \quad (4)$$

Donde $(z)_+ = z$ cuando $z > 0$ y 0 en otro caso.

Construcción. La implementación requiere los paquetes FKF y doParallel

Input/Output. A la función `ssm` se le pasan como argumentos la(s) fórmula(s) y el conjunto de datos que puede ser extendido previamente utilizando las funciones `rds` y `trigon`

3.4 Modelo de aprendizaje automático - RFS

Modelo Bosques Aleatorios. *Random Forests* es una técnica de agregación desarrollada por Leo Breiman (Breiman, 2001), que mejora la precisión en la predicción (clasificación o regresión) mediante la incorporación de aleatoriedad en la construcción de cada predictor individual.

El predictor consiste en una colección de árboles de regresión aleatorizados $\{r_n(X, \Theta_m, \mathcal{D}_n)\}_{m \geq 1}$, donde $\Theta_1, \Theta_2, \dots$ son variables aleatorias i.i.d.. Estos árboles se combinan formando un estimador de regresión agregado $\bar{r}_n(X, \mathcal{D}_n) = \mathbb{E}_{\Theta} [r_n(X, \Theta, \mathcal{D}_n)]$, donde \mathbb{E}_{Θ} es la esperanza respecto al parámetro aleatorio condicional en X y en la muestra de entrenamiento \mathcal{D}_n .

Los árboles utilizados como predictores individuales son variantes de *Classification and Regression Trees* (CART) de Breiman (Breiman, Friedman, Olshen and Stone, 1984), cada uno construido sobre remuestras del conjunto de entrenamiento, creciendo sin poda y en cada nodo seleccionando aleatoriamente las variables que intervienen y las particiones.

La utilización de Random Forests en la previsión de demanda de corto plazo no ha sido muy frecuente (Cheng, Qiu and P.K.K., 2012) (Dudek, 2015b) (Lahouar and Ben Hadj Slama, 2015) (Papadopoulos and Karakatsanis, 2015), a pesar de la sencillez de su implementación y sus buenos resultados en términos de performance en otras aplicaciones.

En este caso se utilizaron como variables las de calendario, rezagos de la demanda y la temperatura, construyendo un bosque para cada hora del día.

Construcción. Esta implementación inicial recurre al paquete `randomForests` del R paralelizando su uso.

Input/Output. Se requiere un set de datos, que puede ser extendido utilizando la función `newvars.rf` que genera rezagos de la temperatura y la carga y otras variables.

3.5 Modelo de reconocimiento de patrones - SPM

Modelo SPM. Sea $\vec{x}_j \in \mathbb{R}^m$, $m = 24$, el vector de demanda horaria de un día j cualquiera. La premisa básica es que las curvas de demanda de días consecutivos, \mathbf{x}_j y \mathbf{x}_{j+1} , están fuertemente vinculadas.

La hipótesis de trabajo de partida es que dicho vínculo se traduce en que el rango del espacio generado por la concatenación de pares de dichas curvas

$$\mathbf{z}_j = [\mathbf{x}_j^\top | \mathbf{x}_{j+1}^\top]^\top \in \mathbb{R}^{2m}$$

es mucho menor al espacio \mathbb{R}^{2m} que los contiene.

En los modelos esparsos (Bruckstein, Donoho and Elad, 2009) (Ramirez, 2011), se asume que existe un *diccionario* $\mathbf{D} \in \mathbb{R}^{m \times p}$ tal que todo vector \mathbf{z}_j puede aproximarse muy bien como combinación de unas pocas columnas de \mathbf{D} . Dicho matemáticamente,

$$\|\mathbf{z}_j - \mathbf{D}\mathbf{a}_j\|_2 \ll \|\mathbf{z}_j\|, \quad \|\mathbf{a}_j\|_0 \leq k \ll p \quad (5)$$

donde $\|\cdot\|_0$ es la pseudo norma que cuenta la cantidad de elementos no nulos de su argumento. Lo anterior extiende la hipótesis básica en que ahora se asume que los datos pertenecen a una unión de subespacios de bajo rango. El modelo (5) suele generalizarse sustituyendo $\|\cdot\|_2$ por una función de ajuste a datos $f(\cdot)$ y a $\|\cdot\|_1$ por una función de regularización $g(\cdot)$ que promueva soluciones esparsas.

Asumimos ahora como nueva hipótesis que alcanza con la primera mitad de \mathbf{z}_j , es decir el día actual \mathbf{x}_j , para determinar la representación de \mathbf{z}_j en la unión de subespacios. Más aún, asumimos que podemos hallar dicho \mathbf{a}_j mediante la resolución del siguiente problema:

$$\mathbf{a}_j = \arg \min_{\mathbf{a}} \|\mathbf{D}_0 \mathbf{a} - \mathbf{x}_{j+1}\|,$$

donde $\mathbf{D}_0 \in \mathbb{R}^{m \times p}$ son las primeras 24 filas de \mathbf{D} , que se corresponde con la parte que representa al “día anterior” \mathbf{x}_{j-1} en \mathbf{z}_j .

Asumiendo entonces que \mathbf{a}_j puede ser recuperado a partir de \mathbf{x}_{j1} , que es el dato que conocemos, entonces podemos reconstruir la información faltante, \vec{x}_{j+1} como $\mathbf{D}_1 \vec{a}_j$, donde \mathbf{D}_1 es ahora la mitad inferior de \mathbf{D} , “encargada” de representar el día que se quiere predecir. En resumen, la aplicación de la idea anterior consta de tres etapas.

1. Aprender diccionario. Resolver el siguiente problema, donde \mathbf{x}_j provienen de datos de entrenamiento:

$$(\mathbf{D}, \mathbf{A}) = \arg \min_{(\tilde{\mathbf{D}}, \tilde{\mathbf{A}})} \sum_{j=1}^{n-1} f(\mathbf{z}_j - \tilde{\mathbf{D}} \cdot \tilde{\mathbf{a}}_j) + g(\tilde{\mathbf{a}}_j),$$

donde

$$\mathbf{z}_j = [\mathbf{x}_j^\top | \mathbf{x}_{j+1}^\top]^\top$$

y $f()$, $g()$, son las funciones de ajuste a datos y regularización y

$$\mathbf{D} = [\mathbf{D}_0^\top | \mathbf{D}_1^\top]^\top$$

2. Ajustar a datos conocidos. Para cada día j en los datos de testing, resolvemos el siguiente problema:

$$\mathbf{a}_j = \arg \min_{\mathbf{a}} f(\mathbf{x}_j - \mathbf{D}_0 \cdot \mathbf{a}) + g(\mathbf{a})$$

3. Extrapolar.

$$\hat{\mathbf{x}}_{j+1} = \mathbf{D}_1 \mathbf{a}_j.$$

Construcción. Esta implementación inicial recurre al paquete `spams` del R.

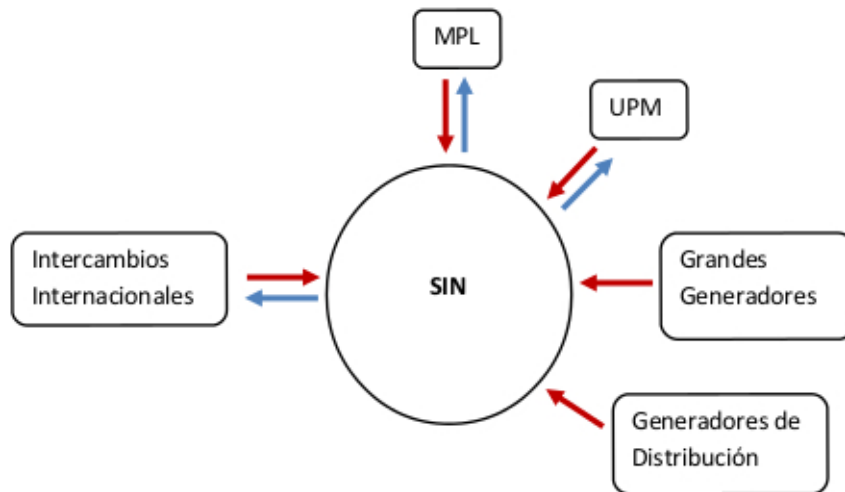
Input/Output. Solamente requiere el vector de carga introducido en forma matricial.

4 Experimentos

4.1 Los datos

Los datos horarios de demanda de Energía Eléctrica de Uruguay para el período 2003-2014, obtenidos a través de del Despacho Nacional de Carga de UTE (DNC).

Definición de Demanda del Sistema Interconectado Nacional



La definición de Demanda acordada con el DNC es la agregación de:

- los grandes generadores de transmisión de UTE (represas, parques eólicos, parques térmicos y parques solares),
- los generadores de distribución (generadores que están conectados a la red de distribución). Esta distinción es necesaria porque los medidores de distribución no se administran en el DNC, solo se poseen las lecturas.
- las importaciones menos las exportaciones (interesa considerar como demanda la importación para satisfacer el consumo del país pero no la que se importa para luego ser exportada, ni la energía que se genera para ser exportada y por ende no va a satisfacer necesidades de consumo nacional). La forma de cálculo propuesta tiene un problema cuando hay cantidades importante de energía en Tránsito en el SIN, pues en ese caso las pérdidas por transmisión quedan registradas como demanda del SIN.
- la diferencia entre la energía entrante (EE) y saliente (ES) de los Cogeneradores. El consumo de los cogeneradores para sus procesos internos no se considera en la demanda ya que no se puede predecir las decisiones del agente sobre sus procesos internos (cuándo van a apagar las calderas). La energía demandada por las plantas de celulosa (UPM, MPL) cuando están consumiendo del SIN, la brinda algún generador al SIN como respaldo al cogenerador y por tanto se resta.

Se cuenta con datos de demanda de energía medida en kilowatts por hora (kWh) desde la hora 1 del 1/1/2003, hasta la hora 24 del 31/12/2014. El conjunto de datos de 105.192 observaciones con cinco variables: cuatro de ellas especifican el momento del tiempo en que la demanda es registrada, *anio* va del 2003 al 2014, *mes* de recorrido $\{1, 2, \dots, 12\}$, la variable *día*, de recorrido $\{1, 2, \dots, 31\}$ ésta varia su alcance dependiendo del mes en el que esta situado el dato y la

variable *hora* de recorrido $\{1, 2, \dots, 24\}$, y *demanda* es la que refiere a la cantidad demandada de energía registrada en ese momento del tiempo, y registra valores entre $\{0, 2129\}$, se agrega una variable que identifica el día de la semana del cual proviene el dato.

	anio	mes	dia	hora	demanda	diasem
1	2003	1	1	1	828,220	miércoles
2	2003	1	1	2	766,573	miércoles
3	2003	1	1	3	686,091	miércoles
4	2003	1	1	4	624,981	miércoles
5	2003	1	1	5	591,963	miércoles
6	2003	1	1	6	552,376	miércoles

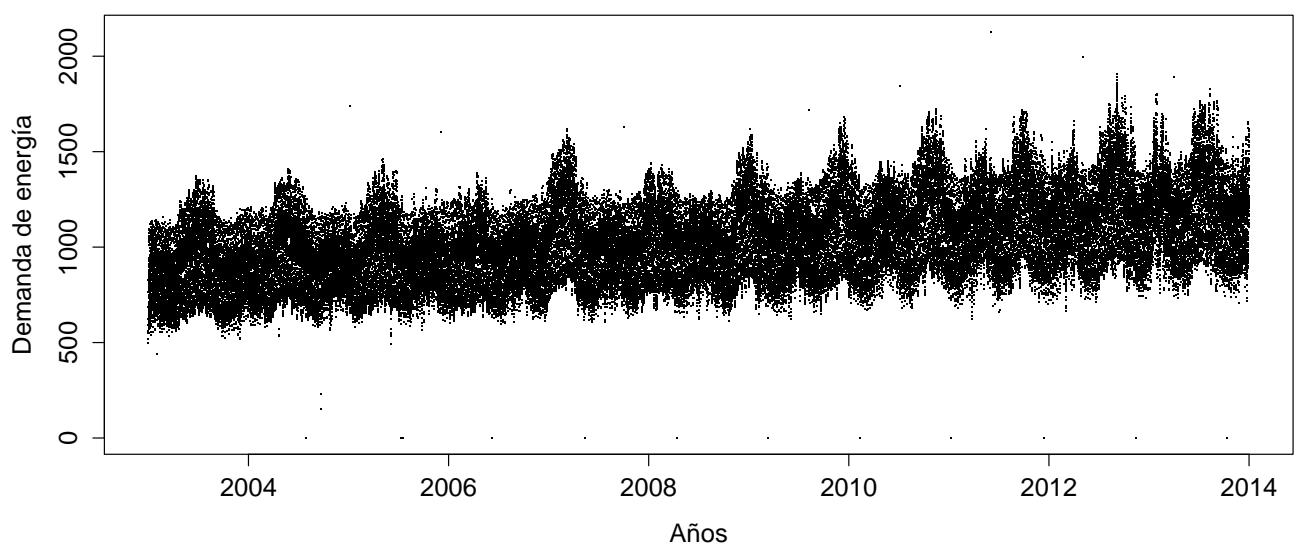
Cuadro 1: Estructura del conjunto de datos.

Dos aspectos a considerar en los datos son:

- Cambio de hora - Horario de Verano. Antes del año 2006 se dictaba un decreto puntual en donde indicaba las fechas en que se produciría el adelanto de la hora legal, a partir del 2006 se dispuso que la hora legal se adelantara a partir de la hora 2:00 AM del primer Domingo del mes de octubre de cada año y que a la hora 2:00 AM del segundo Domingo del mes de marzo de cada año, se retrasar.
- Grandes perturbaciones. Desde el DNC, se identifican los días en los que se registró alguna interrupción del servicio (en Transmisión o distribución) mayor a 10 MW. Estas perturbaciones tienen la capacidad potencial de afectar significativamente la demanda registrada, ya que el tiempo de interrupción puede ser muy pequeño (de unos pocos minutos) y no afectar mayormente la demanda horaria.

Pueden observarse (4.1) la tendencia creciente y la estacionalidad anual de la serie.

Demanda de energía de Uruguay del 2003 al 2014



4.2 Protocolo de comparación

Simultáneamente a la proposición y construcción de los distintos modelos para predecir la demanda, se requiere la definición de un protocolo de comparación adecuado. Esto es, un marco único que permita comparar la performance de cada modelo cuantificarla y resumirla según un criterio que mida la discrepancia entre las predicciones y la demanda efectiva.

Datos. Los datos disponibles provienen del Instituto Uruguayo de Meteorología (INUMET) y del Despacho Nacional de Carga de UTE (DNC).

- Los meteorológicos son datos horarios de temperatura del aire en °C, relevados en la Estación Meteorológica Carrasco (Aeropuerto de Carrasco). elevación 32.8m snmm.
- La demanda de energía eléctrica es una serie de observaciones con frecuencia horaria.
- El período que abarcan va desde la hora 00:00 del 1 de enero de 2007 hasta la hora 23:00 del 31 de diciembre de 2014.

Entrenamiento/Test. El esquema de división de los datos es el siguiente:

- *Training Set*: desde la hora 00:00 del 01/01/2007 hasta la hora 23:00 del 31/12/2013.
- *Test Set*: desde la hora 00:00 del 01/01/2014 hasta la hora 23:00 del 31/12/2014.

Criterio de error. *Mean Average Percentage (MAPE)*

$$\text{MAPE}(y, \hat{y}) = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

donde $y = (y_1, \dots, y_n)$ es la demanda efectiva, $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$ es la prevision y n el numero de observaciones (aquí $n = 24$ lo que corresponde a un día de datos).

4.3 Condiciones Experimentales

- No se utilizan predicciones de temperatura (*Temp*) sino la medida real.
- El experimento se desarrolla condiciones *aproximadamente operacionales* en donde el modelo se estima diariamente sobre el *Training Set* y se predice el día siguiente.
- El experimento se desarrolla en la versión 3.1.2 del R, en una maquina con 64Gb de RAM y 8 CPUs(2100Mhz) bajo Ubuntu 12.04-precise.

Observación. En situación operacional algunos de los modelos deberán reestimarse periódicamente, no necesariamente, de forma diaria.

4.4 Síntesis de Resultados

4.4.1. Modelo de regresión - HVB

El modelo de regresión implementado en enercast es Hong's vanilla benchmark (HVB) tal cual se describe en la sección 3.1. La predicción de cada día demora aproximadamente 5 segundos.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2,12	4,26	5,49	5,972	6,87	30,85

Cuadro 2: Resumen de MAPE para el año 2014

	regular	feriado
Min.	2,120	3,19
1st Qu.	4,238	5,56
Median	5,470	9,08
Mean	5,730	13,76
3rd Qu.	6,828	22,08
Max.	20,960	30,85

Cuadro 3: MAPE por tipo de día para el año 2014

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Min.	2,290	2,770	2,500	2,180	2,820	2,120	3,190
1st Qu.	4,128	4,198	4,270	3,982	4,258	4,625	5,002
Median	5,170	5,210	5,100	5,085	5,505	5,905	5,915
Mean	5,901	5,485	5,653	6,261	6,176	5,955	6,378
3rd Qu.	6,358	6,910	6,050	6,695	7,185	7,200	6,950
Max.	26,990	11,010	17,180	30,850	20,960	13,720	12,060

Cuadro 4: MAPE por día para el año 2014

4.4.2. Modelos de series de tiempo - TSB

Como se detalla en 3.2 el modelo de series temporales Time Series Benchmark (TSB) implementado es un ARIMA que no tiene otras covariables intervinientes mas que la propia serie, es decir que, a diferencia de los ARMAX que en etapas iniciales del proyecto se testearon no se incorporan datos de temperatura o calendario. La predicción de cada día demora aproximadamente 3 minutos.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0,65	2,23	3,14	4,127	5,11	22,03

Cuadro 5: Resumen de MAPE para el año 2014

	regular	feriado
Min.	0,650	2,970
1st Qu.	2,205	4,020
Median	3,105	6,430
Mean	4,032	7,174
3rd Qu.	5,047	9,935
Max.	22,030	14,960

Cuadro 6: MAPE por tipo de día para el año 2014

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Min.	0,750	0,860	0,650	0,770	1,020	1,040	1,170
1st Qu.	2,013	2,298	2,390	2,018	2,215	2,530	2,102
Median	2,840	3,245	3,310	3,245	3,180	3,275	3,070
Mean	3,416	4,375	4,084	4,035	4,489	4,568	3,924
3rd Qu.	4,368	5,265	4,940	5,055	6,122	5,523	5,100
Max.	11,380	17,130	12,870	14,960	11,860	22,030	13,420

Cuadro 7: MAPE por día para el año 2014

4.4.3. Modelos de series de tiempo - SSM

Uno de los Modelo de Espacio de Estados (SSM) implementado conduce al siguiente resultado general:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0,58	2,29	3,42	4,12	4,98	25,84

Cuadro 8: Resumen de MAPE para el año 2014

4.4.4. Modelo de aprendizaje automático - RFS

La implementación de Random Forests es simple, con parámetros utilizados por defecto en la función y utiliza variables de calendario, temperatura y carga rezagadas.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0,71	2,06	2,97	3,643	4,32	20,26

Cuadro 9: Resumen de MAPE para el año 2014

4.5 Modelo de reconocimiento de patrones - SPM

La implementación de Modelos Esparsos se realiza con parámetros por defecto en la función. En principio solo se ha testeado esta aproximación con la propia variable de carga, sin covariables adicionales.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1,21	2,95	3,75	4,01	4,79	9,30

Cuadro 10: Resumen de MAPE para el año 2014

5 Comentarios finales

Aún estamos en el proceso de ajustar los modelos y generar nuevos experimentos, pero los primeros resultados obtenidos en el uso de ENERCAST muestran mejoras en términos predictivos respecto a situación actual.

Este trabajo se ha desarrollado en el marco del proyecto “**Modelos de previsión de demanda corto plazo**”(FSE_1_2013_1_10886) financiado por la Agencia Nacional de Investigación e Innovación (ANII) a través del Fondo Sectorial de Energía. El objetivo del proyecto era desarrollar una modelización para la previsión de la demanda de carga en el sistema eléctrico uruguayo a corto y muy corto plazo.

Los datos utilizados en los experimentos han sido proporcionados por el Despacho Nacional de Carga(DNC) de UTE y por el Instituto Uruguayo de Meteorología(INUMET).

El vinculo productivo, generado en el trabajo conjunto con el DNC, permite que ENERCAST se comience a utilizar, de forma paulatina, en condiciones operacionales. Además de buenas predicciones se obtiene el beneficio de ser generadas internamente y generando una “trazabilidad” registrada del proceso.

Bibliografía

- Alfares, H. K. and Nazeeruddin, M. (2002), ‘Electric load forecasting: Literature survey and classification of methods’, *International Journal of Systems Science* **33**(1), 23–34.
- Antoniadis, A., Brosat, X., Cugliari, J and Poggi, J.-M. (2012), ‘Prevision d’un processus a valeurs fonctionnelles en presence de non stationnarites. Application a la consommation d’electricite’, *Journal de la Societe Française de Statistique* **153**(2), 52–78.
- Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2013), ‘Clustering functional data using wavelets’, *International Journal of Wavelets, Multiresolution and Information Processing* **11**(01), 1350003.
- Ben Taieb, S. and Hyndman, R. J. (2014), ‘A gradient boosting approach to the kaggle load forecasting competition’, *International Journal of Forecasting* **30**(2), 382–394.
- Box, G. E. P. and Jenkins, G. (1990), *Time Series Analysis, Forecasting and Control*, Holden-Day Inc.
- Breiman, L. (2001), ‘Random forests’, *Machine Learning* **45**(1), 5–32.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Chapman & Hall, New York.
- Brockwell, P. and Davis, R. (1991), *Time Series: theory and methods*, Statistics, 2nd edn, Springer.

- Bruckstein, A. M., Donoho, D. and Elad, M. (2009), ‘From sparse solutions of systems of equations to sparse modeling of signals and images’, *SIAM Rev.* **51**(1), 34–81.
- Cancelo, J., Espasa, A. and Grafe, R. (2008), ‘Forecasting the electricity load from one day to one week ahead for the spanish system operator’, *International Journal of Forecasting* **24**, 588–602.
- Charlton, N. and Singleton, C. (2014), ‘A refined parametric model for short term load forecasting’, *International Journal of Forecasting* **30**(2), 364–368.
- Cheng, Y.-Y., Qiu, Z. and P.K.K., C. (2012), Random forest based ensemble system for short term load forecasting, in ‘Machine Learning and Cybernetics (ICMLC), 2012 International Conference on’, Vol. 1, IEEE, pp. 52–56.
- Cho, H., Goude, Y., Brossat, X. and Yao, Q. (2013), ‘Modeling and Forecasting Daily Electricity Load Curves: A Hybrid Approach’, *Journal of the American Statistical Association* **108**(501), 7–21.
- Cugliari, J. (2011), Pr evision non param etrique de processus   valeurs fonctionnelles. Application   la consommation d’ lectricit ., PhD thesis, Universit  Paris Sud.
- Devaine, M., Gaillard, P., Goude, Y. and Stoltz, G. (2013), ‘Forecasting electricity consumption by aggregating specialized experts’, *Machine Learning* **90**(2), 231–260.
- Dordonnat, V., Koopman, S., Ooms, M., Dessertaine, A. and Collet, J. (2008), ‘An hourly periodic state space model for modelling French national electricity load’, *International Journal of Forecasting* **24**(4), 566–587.
- Dudek, G. (2015a), Short-term load forecasting using random forests, in D. Filev, J. Jablkowski, J. Kacprzyk, M. Krawczak, I. Popchev, L. Rutkowski, V. Sgurev, E. Sotirova, P. Szykarczyk and S. Zadrozny, eds, ‘Intelligent Systems’2014’, Vol. 323 of *Advances in Intelligent Systems and Computing*, Springer International Publishing, pp. 821–828.
- Dudek, G. (2015b), Short-Term load forecasting using random forests, in ‘Intelligent Systems’2014’, Springer, pp. 821–828.
- Durrleman, S. and Simon, R. (1989), ‘Flexible regression models with cubic splines.’, *Statistics in medicine* **8**(5), 551–561.
- Goude, Y. (2008), M lange de pr dicteurs et application a la pr vision de consommation  lectrique, PhD thesis, Th se de doctorat, Universit  Paris-Sud.
- Goude, Y., Nedellec, R. and Kong, N. (2014), ‘Local Short and Middle Term Electricity Load Forecasting With Semi-Parametric Additive Models’, *IEEE Transactions on Smart Grid* **5**(1), 440–446.
- Gross, G. and Galiana, F. (1987), ‘Short-term load forecasting’, *Proceedings of the IEEE* **75**(12), 1558–1573.
- Hong, T. (2014), ‘Energy forecasting: Past, present, and future’, *Foresight: The International Journal of Applied Forecasting* **32**, 43–48.
- Hong, T., Pinson, P. and Fan, S. (2014), ‘Global energy forecasting competition 2012’, *International Journal of Forecasting* **30**(2), 357–363.

- Hong, T., Wang, P. and Willis, H. L. (2011), A naïve multiple linear regression benchmark for short term load forecasting, in 'Power and Energy Society General Meeting, 2011 IEEE', IEEE, pp. 1–6.
- Hoverstad, B. A., Tidemann, A., Langseth, H. and Öztürk, P. (2015), 'Short-term load forecasting with seasonal decomposition using evolution for parameter tuning', *Smart Grid, IEEE Transactions on* **PP**(99), 1–1.
- Hoverstad, B., Tidemann, A. and Langseth, H. (2013), Effects of data cleansing on load prediction algorithms, in 'Computational Intelligence Applications In Smart Grid (CIASG), 2013 IEEE Symposium on', pp. 93–100.
- Lahouar, A. and Ben Hadj Slama, J. (2015), Random forests model for one day ahead load forecasting, in 'Renewable Energy Congress (IREC), 2015 6th International', IEEE, pp. 1–6.
- Lanzilotta, B. and Rodriguez, S. (2014), Modelos de predicción de energía eléctrica con datos horarios para Uruguay, Informe de investigación, Fondo Sectorial de Energía. ANII.
- Lloyd, J. R. (2014), 'GEFCom2012 hierarchical load forecasting: Gradient boosting machines and gaussian processes', *International Journal of Forecasting* **30**(2), 369–374.
- Nedellec, R., Cugliari, J. and Goude, Y. (2014), 'GEFCom2012: electric load forecasting and backcasting with semi-parametric models', *International Journal of Forecasting* **30**(2), 375–381.
- Papadopoulos, S. and Karakatsanis, I. (2015), Short-term electricity load forecasting using time series and ensemble learning methods, in 'Power and Energy Conference at Illinois (PECI), 2015 IEEE', IEEE, pp. 1–6.
- Pierrot, A. and Goude, Y. (2011), Short-term electricity load forecasting with generalized additive models., in '16th International Conference on Intelligent System Applications to Power Systems'. to appear.
- R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Ramirez, I. (2011), Second Generation Sparse Models, PhD thesis, University of Minnesota.
- Singh, A. K., Ibraheem, I., Khatoon, S., Muazzam, M. and Chaturvedi, D. K. (2012), Load forecasting techniques and methodologies: A review, in 'Power, Control and Embedded Systems (ICPCES), 2012 2nd International Conference on', IEEE, pp. 1–10.
- Singh, V. V. and Srivastava, A. (2014), 'An introduction to load forecasting: Conventional and modern technologies', *IRACST - Engineering Science and Technology: An International Journal* **4**(2), 62–66.
- Taylor, J. W. (2003), 'Short-term electricity demand forecasting using double seasonal exponential smoothing', *Journal of the Operational Research Society* **54**(8), 799–805.
- Taylor, J. W., de Menezes, L. M. and McSharry, P. E. (2006), 'A comparison of univariate methods for forecasting electricity demand up to a day ahead', *International Journal of Forecasting* **22**(1), 1–16.

- Vilar, J. M., Cao, R. and Aneiros, G. (2012), 'Forecasting next-day electricity demand and price using nonparametric functional methods', *International Journal of Electrical Power & Energy Systems* **39**(1), 48–55.
- Weron, R. (2006), *Modeling and forecasting electricity loads and prices: a statistical approach*, Vol. 396 of *Wiley finance series*, John Wiley and Sons.
- Wood, S. (2006), *Generalized additive models: an introduction with R*, CRC press.
- Wood, S. N., Goude, Y. and Shaw, S. (2015), 'Generalized additive models for large data sets', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**(1), 139–155.