Ethanol Fuel Demand Forecasting in Brazil Using a LSTM Recurrent Neural Network Approach

J. A. Puentes, C. O. Ribeiro, E. A. Ruelas and V. Figueroa

***Abstract*—Ethanol is a biofuel widely consumed in Brazil, which functions as a substitute for gasoline since the late 1970s. Due to several fluctuations in the characteristics of the Brazilian vehicle fleet and the political-economic conditions of the country, forecasting ethanol consumption has become a difficult task to perform. Under this scenario, the aim of this paper was to forecast ethanol consumption in Brazil using an approach of Long-Short Term Memory (LSTM) Recurrent Neural Networks (RNN) and Autoregressive Integrated Moving Average (ARIMA) models. The above, taking into consideration univariate and multivariate models for each case. Likewise, single-layer and multi-layer topologies of LSTM RNN were explored in this study. The results show that LSTM models overperformed ARIMA models even working with a relatively small training dataset of just 180 instances. This, for both univariate and multivariate models. A novel approach for searching suitable LSTM Neural Network topologies is proposed in this paper.**

***Index Terms*—Brazilian ethanol, Deep Learning, LSTM, Recurrent Neural Networks, Time Series Forecasting.**

I. Introducción

La matriz energética mundial estuvo compuesta a 2017 por aproximadamente un 86% de fuentes no renovables de energía, y en sólo un 14% por fuentes renovables [1]. No obstante, para el mismo año, dicha proporción de fuentes de energía en la matriz energética de Brasil fue de un 57% para fuentes no renovables y 43% para fuentes renovables [2]. La fuente energética de Brasil que más contribuye a este gran porcentaje de renovables, con un 17% de participación en el total de la matriz, es la biomasa de caña de azúcar. Esta fuente es destinada para la producción de etanol, un biocombustible que representó a 2018 el 58% del total de combustibles consumidos por la flota brasileña de vehículos ligeros (vehículos familiares y comerciales ligeros) y un 6.4% de la energía total consumida a nivel nacional [3].

This work was supported by the Consejo Nacional de Ciencia y Tecnología (CONACyT). Likewise, it was supported in part by Project FAPESP (São Paulo Research Foundation)/BG Brasil, through the Research Centre for Gas Innovation, Grant 2014/50279-4 as well as FAPESP/RCUK-NERC Grant Proc. 2015/50684-9. The author C. O. Ribeiro received financial support from CNPq (Brazilian National Research Council), grant 307126/2018-8.

J. A. Puentes, Tecnológico Nacional de México en Celaya, Celaya, Guanajuato, México (e-mail: m1803016@itcelaya.edu.mx).

C. O. Ribeiro, Universidade de São Paulo, São Paulo, São Paulo, Brasil (e-mail: celma@usp.br).

E. A. Ruelas, Instituto Tecnológico Superior de Irapuato, Irapuato, Guanajuato, México (e-mail: edruelas@itesi.edu.mx).

V. Figueroa, Tecnológico Nacional de México en Celaya, Celaya, Guanajuato, México (e-mail: vicente.figueroa@itcelaya.edu.mx).

En Brasil, el etanol funciona como combustible substituto de la gasolina común a partir de finales de la década de 1970 [3]. Esto, debido a los cambios en las políticas energéticas nacionales que implicaron las crisis mundiales de petróleo de 1973 y 1979 [4]. Para 1989, aproximadamente 4 millones de vehículos ya eran movidos con etanol, alrededor de un tercio de la flota brasileña de vehículos a dicho año [5]. En la actualidad, el etanol se ha seguido manteniendo como uno de los principales combustibles de Brasil, siendo este ofertado en estaciones de abastecimiento de combustibles a nivel nacional juntamente con la gasolina.

La exitosa introducción del etanol como biocombustible en Brasil ha llamado la atención de instituciones brasileñas y de diversas instituciones más alrededor del mundo. Varios estudios analizan el impacto de las políticas gubernamentales brasileñas en dicho éxito [4] [6], otros, aspectos de producción, competitividad, y sustentabilidad entre el etanol producido en Brasil y Estados Unidos (principales productores de etanol en el mundo) [5] [7] [8], variables econométricas de consumo entre etanol y gasolina [9] [10], entre otros tópicos más [11] [12] [13]. No obstante, en lo referente al consumo de etanol en Brasil, hay un problema que llama especial atención.

Debido a la introducción de vehículos con tecnología *flex-fuel* en Brasil a partir de 2003, el consumo interno de etanol comenzó a presentar diversas variaciones. Esto, dado que a diferencia de un vehículo movido sólo a gasolina, un vehículo *flex-fuel* puede ser abastecido con etanol, gasolina o una mezcla entre ambos. Lo anterior, gracias a un sistema electrónico que monitoriza y procesa constantemente la información de diversas variables del vehículo con el fin de adaptar el funcionamiento del motor a la mezcla de combustibles disponible en tanque. Este suceso, acompañado de la gran recesión económica mundial de 2008, las crisis sufridas en el sector sucroenergético de Brasil en 2010-2012, y la nueva política de precios de gasolina y diésel introducida por Petrobras (principal empresa productora y comercializadora de combustible en Brasil) en 2017, ha hecho que el comportamiento de la demanda de etanol en Brasil presente diversas fluctuaciones. Esto ha ocasionado, a su vez, que las previsiones de consumo interno de etanol y de demás variables relacionadas sean difíciles de realizar [10].

Lo anterior, ha originado la necesidad de utilizar métodos de predicción que puedan abstraer con precisión el comportamiento de estas variables para minimizar así la incertidumbre asociada a sus comportamientos futuros. En este sentido, en [14] se realiza la predicción de los precios del etanol brasileño utilizando un enfoque de Redes Neuronales

Artificiales (RNA) de tipo perceptrón multicapa. Así mismo, en [15] se realiza la predicción de precios de etanol utilizando un enfoque de modelos ARIMA y delimitando el estudio al estado de São Paulo. Por otra parte, en [16] se utiliza un modelo SARIMA para predecir el consumo de etanol a nivel nacional para el periodo 2006 – 2012. Además, en [17] se realiza una predicción anual de la demanda de etanol y demás combustibles de Brasil para el periodo 2008 – 2017 utilizando un modelo integrado de tipo econométrico e insumo-producto.

Sin embargo, no se encontraron estudios publicados que atiendan la problemática descrita utilizando un enfoque de RNA recurrentes de tipo *Long-Short Term Memory* (LSTM), las cuales han demostrado recientemente ser una herramienta de alta precisión para el modelado de series de tiempo [18] [19] [20]. A nivel internacional, en [21] modelos LSTM son utilizados para predecir los precios de crudo WTI y Brent. En [22] son utilizados modelos estadísticos y computacionales, entre ellos modelos LSTM, para pronosticar datos de demanda para un producto. Por otra parte, en [18] predicciones de producción de petróleo son realizadas mediante un enfoque de RNA LSTM profundas. Los resultados de los estudios anteriormente mencionados mostraron una clara superioridad en desempeño de los modelos LSTM desarrollados sobre los demás modelos comparados.

En este orden de ideas, la presente investigación tuvo como objetivo proponer un modelo de predicción de demanda de etanol en Brasil bajo un enfoque de RNA recurrentes de tipo LSTM con el fin de disminuir la incertidumbre asociada al comportamiento de la demanda de este combustible. Las principales contribuciones del presente trabajo son el empleo de RNA LSTM utilizando enfoques profundos, univariados y multivariados para solucionar la problemática bajo estudio. Así mismo, la demostración de la eficacia de esta herramienta incluso trabajando con un subconjunto de entrenamiento de sólo 180 instancias. Por otra parte, un innovador enfoque metodológico de búsqueda de topología de red neuronal basado en *grid search* es propuesto en la presente artículo.

El resto de este artículo es organizado de la siguiente manera. En la sección II se realiza una descripción de las RNA recurrentes de tipo LSTM. En la sección III, se describen los datos empleados y el método aplicado para la generación y comparación de modelos de predicción. Luego, en la sección IV, se presentan resultados y comparaciones entre los modelos elaborados. Finalmente, en la sección V, se exponen las respectivas conclusiones.

II. Redes Neuronales Artificiales LSTM

Las RNA son algoritmos inspirados en el comportamiento de las redes neuronales biológicas existentes en el cerebro. Su objetivo es emular la manera en la cual dichas redes biológicas reciben, procesan y generan información. Estas consisten en un sistema de capas de entrada, capas ocultas, y capas de salida que se encargan de recibir, procesar y exportar información respectivamente [23]. Se denomina una topología o arquitectura de RNA a la estructura de capas que una red posee y los parámetros presentes en cada una de estas. Una topología básica es la de las RNA de tipo Perceptrón Multicapa (MLP, *MultiLayer Perceptron*), la cual consiste en un conjunto de nodos (neuronas) que se interconectan entre sí a través de enlaces. A su vez, cada enlace posee un conjunto de pesos, los cuales refieren a la intensidad de la conexión entre cada nodo. En este orden de ideas, las entradas pasan a través de cada nodo y mediante una optimización de pesos (proceso de aprendizaje) se obtienen unas salidas deseadas. Al final del dicho proceso, el aprendizaje que la RNA obtiene puede ser expresado en términos del conjunto de pesos hallados que permitió atender determinado problema utilizando determinada topología [24].

Sin embargo, a pesar de que las RNA de tipo MLP pueden ser utilizadas para modelar relaciones complejas, estas no son capaces de asimilar dependencias de largo y corto plazo presentes en datos históricos [25]. Estas dependencias refieren a la capacidad de una RNA para identificar y recordar patrones de comportamiento del pasado distante y el pasado cercano respectivamente. Esto, con fines de realizar predicciones de comportamiento de datos secuenciales [26].

Como un intento de atender dicha problemática, surgen en la década de 1980 las primeras Redes Neuronales Recurrentes (RNN, *Recurrent Neural Networks*), en donde el término "recurrente" refiere a la característica de estas redes de poseer bucles internos de retroalimentación. No obstante, estas redes presentaron una gran desventaja la cual es conocida en la literatura como el *vanishing gradient problem* o problema del gradiente de fuga. Este problema refiere a la dificultad que se tiene al entrenar estas redes con métodos basados en gradientes o algoritmos *backpropagation*. Esto, debido a que cuando este tipo de métodos calculan el ajuste de pesos con base a la regla de la cadena, se tienen comúnmente múltiplos entre 0 y 1 que se multiplican entre sí n veces. Al multiplicar n veces números menores a 1, se genera un decrecimiento exponencial en la señal del error, lo cual afectaba significativamente el entrenamiento de estas RNN [27].

Debido a lo anterior, surgen en 1997 las RNN de tipo LSTM gracias a los científicos de la computación S. Hochreiter y J. Schmidhuber [26]. En este tipo de redes se poseen bucles de retroalimentación compuestos por células de memoria que poseen un sistema de decisiones basado en compuertas, las cuales que se encargan de "recordar" patrones históricos importantes y "olvidar" patrones no relevantes. Gracias a estas células de memoria, una RNN de tipo LSTM es capaz de asimilar dependencias a corto y largo plazo y, además, atender eficientemente el problema del gradiente de fuga. La estructura de dichas células de memoria es presentada en la Fig. 1.

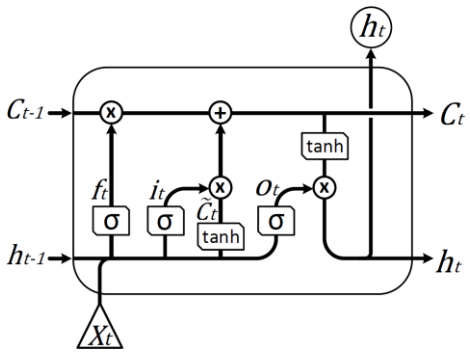


Fig. 1. Estructura de una célula de memoria de RNN de tipo LSTM.

Donde:

X_t: *entradas de la red LSTM en un tiempo t.*

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (7)$$

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (8)$$

W y b son respectivamente los pesos y bias del enlace representado en sus subíndices. Finalmente, los círculos que contienen los símbolos "+" y "x" refieren a operaciones puntuales de suma y multiplicación respectivamente. Estas operaciones, en el orden anteriormente dicho, son las responsables de hacer que cada célula de memoria de una red LSTM sea capaz de adicionar (recordar) y retirar (olvidar) información del estado de célula C_t. En el caso de f_t, esta función toma como insumo la salida de la célula de memoria inmediatamente anterior (h_{t-1}) y las entradas de datos X_t para mediante la función sigmoide (σ), generar un escalar entre 0 y 1 que multiplicará el estado de célula C_t, el cual se encarga de transportar la información relevante a través de las células de memoria. Un valor de 0 representa olvidar toda la información contenida en C_t, un valor de 1 representa lo contrario. Debido a esto, esta sección de la célula LSTM es llamada la compuerta de olvido (*forget gate*) [28].

Mediante una dinámica similar, i_t y $\tilde{C}_t$ adicionan información a C_t utilizado la operación puntual de suma, por lo cual esta sección de célula recibe el nombre de compuerta de entrada (*input gate*). Finalmente, una versión filtrada del estado de célula C_t es obtenida mediante el producto de la aplicación de la función $\tanh$ a dicho estado de célula y el escalar obtenido en o_t. Lo anterior, constituye la salida h_t de una célula de memoria y representa la sección denominada como compuerta de salida (*output gate*). Gracias a la anterior estructura, una RNN de tipo LSTM es capaz de atender problemas de procesamiento de lenguaje natural, análisis de series de tiempo, entre otras tareas [28].

III. Materiales y Métodos

A. Levantamiento y Adecuamiento de Datos

Tomando en consideración la relación dual que existe entre los consumos de etanol y gasolina en Brasil [29], las siguientes series de tiempo fueron obtenidas: consumo mensual de etanol y gasolina, flota mensual de vehículos *flex-fuel* y gasolina, y precio promedio ponderado mensual nacional de etanol y gasolina para el consumidor final. Esto, para el periodo julio de 2001 hasta junio de 2019. Levantando así, 6 series de tiempo de 18 años (216 instancias) cada una.

Las series de consumo fueron recabadas de la *Agência Nacional do Petróleo, Gás Natural e Biocombustíveis* (ANP) a través del repositorio de datos IPEADATA [30]. Los datos disponibles correspondieron al consumo aparente (producción nacional aumentada por las importaciones y disminuida por las exportaciones) promedio diario de cada mes, por lo que fue necesario multiplicar cada instancia por el número de días que tuvo cada mes en cada respectivo año para estimar así el consumo total de etanol y gasolina de cada mes. Una prueba Dickey-Fuller Aumentada (ADF) y una descomposición ETS (*Error, Trend, Seasonality*) fueron realizadas para esta serie de tiempo. En lo referente a las series de precios, estas fueron extraídas directamente del sitio web de la ANP y se encuentran expresadas en la divisa Reales Brasileros (BRL).

En el caso de las series de tiempo de flota circulante de vehículos *flex-fuel* y gasolina, fueron elaboradas dos curvas de chatarrización mediante la función Gompertz: una para vehículos comerciales leves y otra para vehículos familiares. Lo anterior, debido a que estos dos tipos de vehículos son los que consumen alternadamente gasolina y/o etanol en Brasil (vehículos de carga y buses consumen otros tipos de combustibles). Las curvas fueron calibradas utilizando una aproximación a los datos anuales de flota circulante de vehículos de la *Associação Nacional dos Fabricantes de Veículos Automotores* (ANFAVEA).

Dado que históricamente las licencias otorgadas a vehículos comerciales leves representan el 18.31% del total de las licencias otorgadas a vehículos *flex-fuel* y sólo a gasolina, se realizó una nueva curva de chatarrización teniendo en cuenta la proporción anteriormente dicha. Esto, con el fin de que esta nueva curva pudiera ser aplicada a los datos mensuales de licenciamiento de vehículos por tipo de combustible provistos por la asociación ANFAVEA. De esta manera, se obtuvieron las series mensuales de flota total circulante de vehículos *flex-fuel* y gasolina para el periodo anteriormente mencionado.

B. Construcción de Algoritmo de Búsqueda de Topología de Red LSTM

En esta etapa se construyó un algoritmo de búsqueda de tipo *grid search* con el fin de encontrar una topología de red que permitiera definir un modelo de red LSTM que brindase la mejor generalización de la serie de respuesta (consumo de etanol). La codificación del algoritmo fue realizada en el lenguaje de programación Python en su distribución Anaconda, dentro del entorno de desarrollo integrado Spyder, y utilizando como sistema operativo Windows 10 Pro. Para la construcción del bloque de código referente a las topologías de red LSTM fue utilizada la biblioteca externa Keras con TensorFlow como *backend*. Así mismo, fueron utilizadas las bibliotecas Pandas y Numpy para procesamiento de datos, y Scikit-learn y Pyplot para cálculo de errores y visualización de resultados respectivamente. Los cálculos del algoritmo de búsqueda fueron ejecutados en un ordenador con procesador Intel i7 8700K, 16 GB de RAM a 3200 MHz en canal dual, un SSD NVMe M.2 para el sistema operativo y software relacionado, y un SSD SATA dedicado a la lectura y escritura de datos por parte del algoritmo elaborado.

Dicho algoritmo recibió como insumo datos normalizados en una escala entre 0 a 1. Los cuales estuvieron divididos en subconjuntos de entrenamiento y validación. Para el entrenamiento fueron definidos 15 años (83.33%) y para la validación 3 años (16.67%). Así mismo, se elaboraron listas multidimensionales para dichos subconjuntos, en donde cada elemento de lista contiene la información de 24 o 36 pasos de tiempo (instancias) hacia atrás. Esto es un parámetro de búsqueda del algoritmo desarrollado y refiere a la cantidad de

datos históricos que la red LSTM toma en consideración para la identificación de patrones individuales de comportamiento. Un seudocódigo de dicho algoritmo es mostrado a continuación:

Inicio.
Establecer variables previsoras y normalizar datos.
Dividir subconjuntos de entrenamiento y validación.
Definir listas con parámetros de búsqueda.
Inicializar lista para almacenar resultados de pruebas.

Para ts en lista [pasos de tiempo]:
 Para hl en lista [capas ocultas a experimentar]:
  Para i_{hl} en lista [células de memoria i en capa hl]:
   Para j en lista [valores de dropout]:
    Para k en lista [funciones de activ. en neurona de salida]:
     Para l en lista [optimizadores]:
      Para m en lista [funciones de error]:
       Para n en lista [épocas]:
        Para o en lista [batch sizes]:
         Para p en rango [repeticiones]:

          tf.keras.backend.clear_session.
          Construir listas entrenamiento (ts) y validación (ts).
          Compilar red LSTM (hl, i_{hl}, j, k, l, m, n, o, p).
          Entrenar dinámicamente la red LSTM.
          Predecir a validación (ts) y entrenamiento (ts).
          Desnormalizar previsiones.
          Calcular medidas de error.
          Anexar resultados a lista previamente inicializada.
          Exportar backup de red LSTM a unidad de almacenamiento local.

Guardar resumen de resultados en archivo .csv.
Fin.

La red LSTM fue entrenada de forma dinámica mediante la función EarlyStopping de Keras, la cual detiene el entrenamiento, pasando así a la siguiente prueba del algoritmo de búsqueda, si la función de error no mejora en determinada medida durante un número determinado de épocas. Para el presente modelo, se estableció una detención en el entrenamiento si la función de error no mejora por lo menos en 0.0001 en 150 épocas. Así mismo, fue utilizada la función ReduceLROnPlateau para reducir la tasa de aprendizaje de la red LSTM en un factor de 0.1 si dicha función de error no mejora en la medida previamente especificada a lo largo de 100 épocas.

Además de lo anterior, fue implementada la instrucción tf.keras.backend.clear_session para limpiar de memoria los nodos/grafos creados por TensorFlow en la prueba inmediatamente anterior y mejorar así el uso de recursos computacionales. Adicionalmente, una exportación de cada topología de red LSTM generada es exportada en forma de archivo Hierarchical Data Format (.h5) con el fin de guardar copias de seguridad que puedan ser posteriormente cargadas. En este tipo de archivo es almacenada la estructura de la red LSTM y los pesos y bias alusivos a las interconexiones presentes entre cada capa y célula de memoria. Se definió que el nombre de estos archivos contuviese información de la estructura de red y comenzase con el valor de la medida de error RMSE alusiva al desempeño obtenido en el subconjunto de datos de validación. Esto, con el fin de catalogar fácilmente dichas topologías en el explorador de Windows de acuerdo con el valor de dicha medida de error. Lo anterior permitió que cada topología de red pudiese ser fácilmente catalogada y, posteriormente, cargada y analizada a detalle.

C. *Construcción de Modelos ARIMA*

Con fines de comparar el desempeño de los modelos LSTM elaborados con otros métodos de pronóstico, y dado el buen desempeño que han mostrado tener los modelos ARIMA y sus respectivas variaciones para predicción de series de tiempo [18], se realiza en esta etapa la construcción de 2 modelos ARIMA: un modelo SARIMA y un modelo SARIMAX. El modelo SARIMA es elaborado debido a la componente estacional que posee la serie de tiempo de consumo de etanol. Por otra parte, el modelo SARIMAX es desarrollado con fines de realizar una comparación justa entre este y el modelo de red LSTM multivariado creado. Esto, dada la capacidad que igualmente poseen los modelos SARIMAX de aceptar múltiples variables previsoras. La definición de parámetros de estos modelos fue llevada a cabo utilizando la librería statsmodels y la función auto_arima de la librería pmdarima del lenguaje de programación Python. Esta última función realiza pruebas KPSS, Augmented Dickey-Fuller, entre otras, para determinar el orden de diferenciación 'd' de los modelos. Posteriormente, ajusta diferentes modelos tomando en consideración el parámetro 'd' hallado y posibles parámetros 'p', 'q', y 'P', 'Q'. En cuanto a la definición del parámetro 'D', dicha función utilizada la prueba de estacionalidad de Canova-Hansen. Por último, es retornado el modelo con mejor índice AIC (Akaike Information Criterion) [31].

D. *Comparación de Resultados*

Finalmente, se procedió a contrastar el desempeño obtenido en cada uno de los seis modelos de predicción elaborados: un modelo SARIMA, un modelo LSTM y un modelo Deep-LSTM con única variable previsora (modelos univariados), un modelo SARIMAX, un modelo LSTM, y un modelo Deep-LSTM con múltiples variables previsoras (modelos multivariados). Dicho contraste fue llevado a cabo tomando en consideración las siguientes medidas de error: error medio absoluto (MAE) y raíz del error cuadrático medio (RMSE). Estas medidas de error fueron calculadas para los subconjuntos de validación y entrenamiento con el objetivo de analizar la calidad de cada modelo tanto en su fase de validación como de entrenamiento. Las anteriores medidas de error fueron los criterios utilizados para definir el mejor modelo y ayudaron, además, a identificar situaciones de *underfitting* y *overfitting*.

IV. Resultados

Para la serie de tiempo de consumo de etanol se realizó inicialmente una prueba ADF con el fin de conocer si dicha serie de tiempo poseía un comportamiento estacionario o no. Dicha prueba, con un p-valor igual a 0.8414, no permitió

rechazar la hipótesis nula de que existe una raíz unitaria para la serie de tiempo en cuestión. Dicho de otra manera, se infiere que la serie de tiempo no es estacionaria y que su media y varianza cambian a través del tiempo. Esto, da indicios de que se posee tendencia o estacionalidad en dicha serie. Para corroborar lo anterior, una descomposición de la serie en sus componentes de tendencia, estacionalidad, y residuos (*ETS decomposition*) fue realizada. Dicha descomposición es representada en la Fig 2.

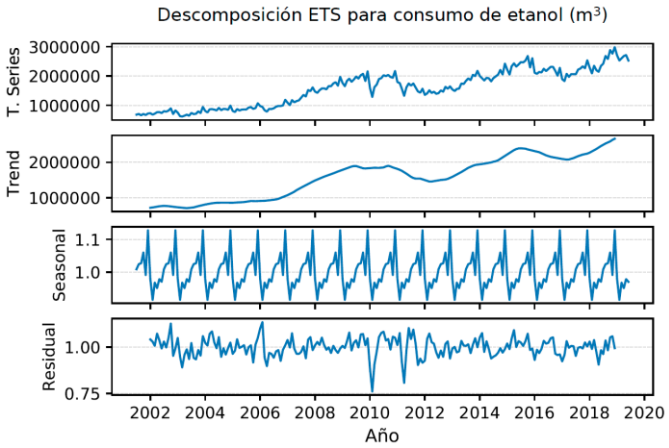


Fig. 2. Descomposición de serie de tiempo de consumo de etanol.

Como es posible apreciar en la Fig. 2, se poseen relevantes componentes de tendencia y estacionalidad para la serie de tiempo de consumo de etanol, lo cual apoya el resultado obtenido en la prueba ADF. Además, es posible observar una alta componente residual, la cual refiere a los datos que no son explicados por las componentes de tendencia y estacionalidad. Por otra parte, con relación al adecuamiento de las variables exógenas contempladas, se representa en la Fig. 3 las curvas de chatarrización empleadas para estimar los valores históricos de flota circulante de vehículos *flex-fuel* y gasolina.

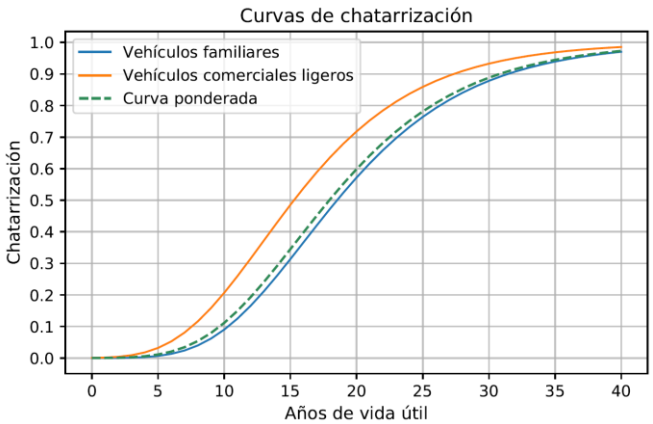


Fig. 3. Curvas de chatarrización empleadas para estimar flotas circulantes.

Las curvas de chatarrización construidas para vehículos familiares y comerciales ligeros, observables en la Fig. 3, sirvieron para estimar una nueva curva que representase a ambas. Esto fue realizado mediante una ponderación basada en la proporción histórica de licencias otorgadas a dichos tipos de vehículos en Brasil. De esta manera, fue posible obtener una nueva curva aplicable a los datos de licencias otorgadas a vehículos por tipo de combustible. Esto, debido a que las estimaciones de chatarrización se hacen sobre los vehículos y no sobre los combustibles. Por lo cual, si se desean realizar estimativas sobre flota histórica circulante por tipo de combustible, se deben hacer generalizaciones desde los tipos de vehículos hasta los tipos de combustible. Como resultado de lo anterior, y de demás actividades de levantamiento y adecuamiento de series de tiempo detalladas en la sección anterior, se obtuvieron las series tomadas en consideración en el presente estudio, la cuales son representadas en la Fig. 4.

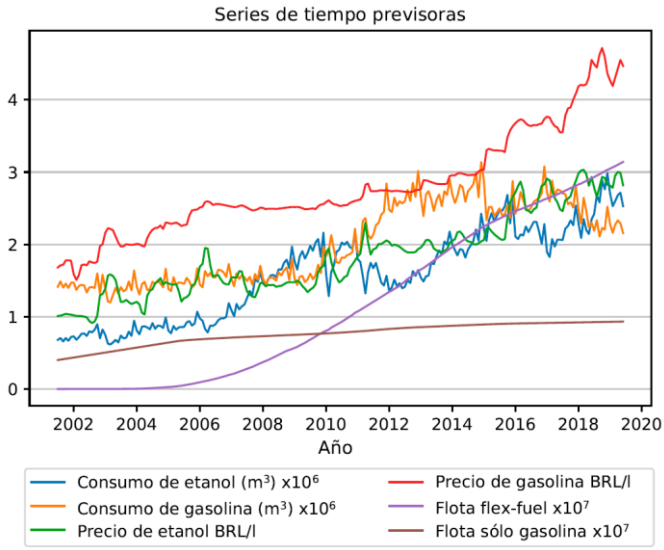


Fig. 4. Series de tiempo previsoras de consumo de etanol.

Las anteriores series de tiempo fueron utilizadas en su totalidad sólo para los modelos multivariados. Para el caso de los modelos univariados, sólo fue tomada la respectiva serie de tiempo de respuesta (consumo de etanol). La correlación presente entre variables fue medida utilizando el coeficiente de correlación Pearson. Los resultados, como se muestran en la Tabla I, muestran una alta correlación positiva entre las variables consideradas, siendo notorio que todas las variables exógenas contempladas guardan un coeficiente de correlación de más del 72% y en promedio del 83% con respecto a la serie de consumo de etanol.

TABLA I
MATRIZ DE CORRELACIÓN DE PEARSON

	Cons. Etanol	**Cons. Gasol.**	**Precio etanol**	**Precio gasol.**	**Flota F-F.**	**Flota Gasol.**
Cons. Etanol	1	0.73	0.80	0.85	0.89	0.88
Cons. Gasol.	0.73	1	0.78	0.69	0.87	0.83
Precio etanol	0.80	0.78	1	0.96	0.94	0.86
Precio gasol.	0.85	0.69	0.96	1	0.93	0.86
Flota F-F.	0.89	0.87	0.94	0.93	1	0.88
Flota Gasol.	0.88	0.83	0.86	0.86	0.88	1

Por otra parte, en términos de la aplicación del algoritmo de búsqueda de topología de red neuronal, este realizó 8064 experimentos alusivos a 2688 topologías de red LSTM. Se experimentó con 1 y 2 capas ocultas, la inclusión de 3 o más capas ocultas no representó mejorías a los resultados obtenidos. Dicha experimentación tuvo un tiempo de ejecución aproximado de 27 y 38 horas continuas de cómputo en paralelo para los modelos univariados y multivariados respectivamente. El uso total de espacio en unidad de almacenamiento SSD SATA, para todas las escrituras del algoritmo en ambas modalidades de variables previsoras consideradas, mostró ser inferior a 5 GB. Así mismo, el uso total de RAM (incluyendo el consumo del sistema operativo y servicios en segundo plano) mostró ser de, en promedio, 14 GB. Los resultados obtenidos en los modelos univariados, para el subconjunto de datos de validación, son representados en la Fig. 5 y Tabla II.

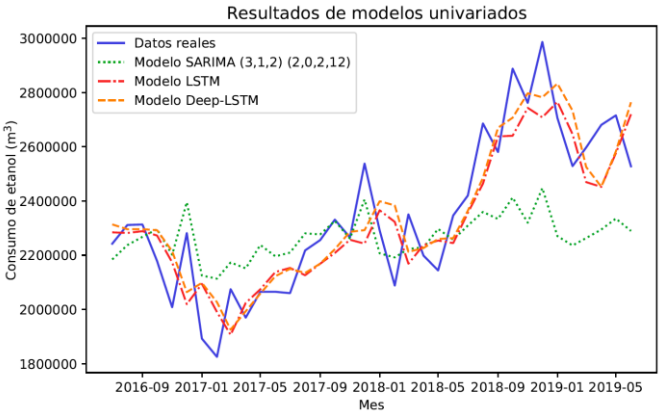


Fig. 5. Previsiones de consumo de etanol de modelos univariados.

TABLA II
MATRIZ DE ERRORES Y PARÁMETROS PARA MODELOS UNIVARIADOS

Modelo	Parámetros	MAE	RMSE
SARIMA	**Orden:** (3,1,2) (2,0,2,12)	**Test**: 191008.94	**Test**: 238122.36
LSTM	**Pasos de tiempo**: 24 **Capas ocultas**: 1 **Células de memoria**: 60 **Dropout**: 0.00 **Fn. activación**: Linear **Optimizador**: Adam **Learning rate**: 0.001 **Épocas**: 270 **Batch size**: 24	**Test**: 124580.25 **Training**: **92112.03**	**Test**: 148770.75 **Training**: **129093.22**
Deep-LSTM	**Pasos de tiempo**: 24 **Capas ocultas**: 3 **Células de memoria**: 60-30-60 **Dropout**: 0.00 **Fn. activación**: Linear **Optimizador**: Adam **Learning rate**: 0.001 **Épocas**: 150 **Batch size**: 24	**Test**: **123870.51** **Training**: 100130.77	**Test**: **146117.09** **Training**: 142072.29

Las anteriores previsiones muestran una clara superioridad de los modelos LSTM y Deep-LSTM sobre el modelo SARIMA. Además, es posible apreciar que existe una ligera diferencia entre estos dos primeros modelos en cuanto al desempeño obtenido en la etapa de validación (Test), siendo el modelo Deep-LSTM el que menor error obtuvo en ambas medidas. Sin embargo, en la fase de entrenamiento (Training) el modelo LSTM mostró un mejor desempeño con un margen de diferencia superior al observado en los resultados de validación de estos dos modelos. Debido a lo anterior, se establece hasta este punto al modelo LSTM como el modelo de mejor desempeño dentro de la clase de modelos univariados. Los resultados obtenidos en los modelos multivariados, para dicho subconjunto de validación, son mostrados en la Fig. 6 y Tabla III.

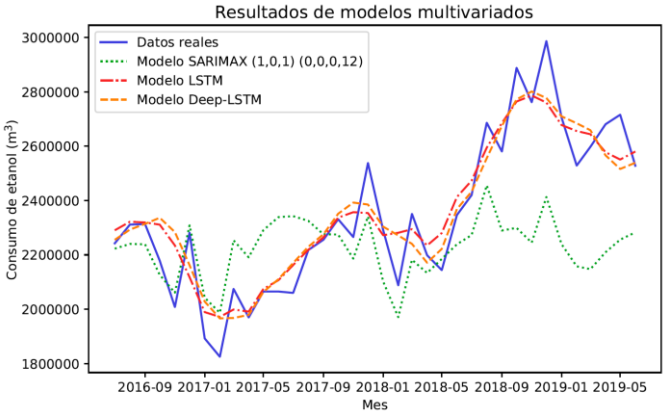


Fig. 6. Previsiones de consumo de etanol de modelos multivariados.

TABLA III
MATRIZ DE ERRORES Y PARÁMETROS PARA MODELOS MULTIVARIADOS

Modelo	Parámetros	MAE	RMSE
SARIMAX	**Orden:** (1,0,1) (0,0,0,12)	**Test**: 215255.07	**Test**: 271829.67
LSTM	**Pasos de tiempo**: 36 **Capas ocultas**: 1 **Células de memoria**: 120 **Dropout**: 0.00 **Fn. activación**: Linear **Optimizador**: Adam **Learning rate**: 0.001 **Épocas**: 200 **Batch size**: 16	**Test**: **84120.47** **Training**: **64995.82**	**Test**: **106020.78** **Training**: **89109.54**
Deep-LSTM	**Pasos de tiempo**: 24 **Capas ocultas**: 2 **Células de memoria**: 120-120 **Dropout**: 0.00 **Fn. activación**: Linear **Optimizador**: Adam **Learning rate**: 0.001 **Épocas**: 150 **Batch size**: 24	**Test**: 84754.01 **Training**: 78184.22	**Test**: 110562.99 **Training**: 105621.18

Como es posible apreciar en la Fig. 6 y Tabla III, el desempeño de los modelos LSTM y Deep-LSTM mejoró significativamente con la inclusión de variables exógenas, logrando generalizar en mayor medida la serie de tiempo de respuesta. En el caso del modelo LSTM multivariado, el desempeño de este aumentó en un 28.74% y 30.97% con respecto al modelo LSTM univariado para los subconjuntos

de validación y entrenamiento respectivamente. Esto, tomando como referencia la medida de error RMSE. Así mismo, el modelo Deep-LSTM multivariado mostró un desempeño superior en un 24.33% y 25.66% con respecto al modelo Deep-LSTM univariado para los subconjuntos de validación y entrenamiento respectivamente tomando como referencia la misma medida de error.

No obstante, la inclusión de estas variables mostró ser contraproducente en el modelo SARIMAX, el cual obtuvo un desempeño inferior al del modelo SARIMA en un -14,16%. Dada la alta correlación existente entre variables previsoras expuesta en la Tabla I, los anteriores resultados ponen en manifiesto que, para el presente caso de estudio, las redes neuronales recurrentes LSTM mostraron poseer una capacidad superior de relación y abstracción de patrones de comportamiento entre variables exógenas y variable de respuesta en comparación con los modelos ARIMA. Lo anterior, posiciona a este tipo de RNA como la mejor de las herramientas evaluadas para la predicción de consumo de etanol en Brasil. Pudiendo ser igualmente útil y aplicable en países con escenarios de incertidumbre similares.

Con fines de realizar estimaciones de consumo para el periodo de 2019-07 a 2020-12, se tomó como referencia el modelo LSTM univariado. Esto, debido a que aunque el modelo LSTM multivariado fue el que mejor desempeño obtuvo, este último modelo, así como cualquier modelo de predicción de series de tiempo basado en variables exógenas, necesita de los datos futuros de dichas variables exógenas para realizar las respectivas previsiones. Dicho de otra forma, se necesitan estimaciones de las series de tiempo de variables exógenas para el periodo 2019-07 a 2020-11 para estas ser ingresadas al modelo y así poder generar las respectivas previsiones de la variable de respuesta (consumo de etanol). Dado que la estimación de dichas variables exógenas se encontró por fuera del alcance de la presente investigación, se presentan en la Tabla. IV las predicciones realizadas por el modelo LSTM univariado para el periodo de 2019-07 a 2020-12.

TABLA IV
PREVISIONES DE MODELO LSTM UNIVARIADO

MES	PREVISIÓN (m^3)	MES	PREVISIÓN (m^3)
jul-19	2695541,14	**abr-20**	2401014,67
ago-19	2746418,03	**mai-20**	2386586,29
set-19	2797561,14	**jun-20**	2408320,57
out-19	2826678,88	**jul-20**	2452392,20
nov-19	2811336,70	**ago-20**	2509833,29
dez-19	2735678,49	**set-20**	2560539,06
jan-20	2625701,76	**out-20**	2594224,69
fev-20	2527430,80	**nov-20**	2588780,59
mar-20	2455533,91	**dez-20**	2549101,06

V. CONCLUSIONES

En el presente estudio fueron desarrollados múltiples modelos de predicción de demanda de etanol en Brasil utilizando redes neuronales artificiales recurrentes de tipo LSTM y modelos ARIMA. Esto, explorando abordajes univariados y multivariados para cada tipo de modelo. Los resultados mostraron que los modelos LSTM encontrados mediante el algoritmo de búsqueda de topología de red, tanto univariados como multivariados, tuvieron el mejor desempeño para el caso de estudio abordado. Los experimentos conducidos resaltan la alta capacidad de los modelos LSTM para abstraer patrones de comportamiento incluso en conjuntos de datos de entrenamiento no extensos (180 instancias para el presente estudio). Así mismo, se evidenció que la inclusión de capas ocultas en los modelos LSTM no significó la obtención de mejores resultados para ambos subconjuntos de entrenamiento y validación.

En lo referente a los modelos LSTM multivariados, es resaltable el alto desempeño y capacidad de generalización que estos mostraron para la problemática bajo estudio. No obstante, dada la limitación que estos modelos poseen (la cual fue explicada al final de la anterior sección), se recomienda a futuras investigaciones relacionadas involucrar este tipo de modelos cuando se posean variables exógenas fácilmente predecibles o, en su defecto, que puedan ser predichas mediante modelos univariados con gran precisión. Esto último, tomando en consideración el riesgo (ruido) que supone ingresar estimaciones a un modelo predictor para realizar más estimaciones. Este último aspecto, se propone como tópico de indagación para futuras investigaciones en donde se explore el efecto de la inclusión de previsiones de modelos LSTM univariados como insumo de modelos LSTM multivariados versus el desempeño aislado de dichos modelos LSTM univariados. Finalmente, dada la alta variabilidad de la serie de tiempo bajo estudio, se recomienda realizar labores de reentrenamiento a la red LSTM por lo menos cada 12 meses (periodicidad estacional).

REFERENCIAS

[1] International Energy Agency - IEA, «World Energy Balances 2019,» IEA Publications & Data, París, 2019.

[2] Empresa de Pesquisa Energética, «Balanço Energético Nacional 2018: ano base 2017,» Ministério de Minas e Energia, Rio de Janeiro, 2018.

[3] Empresa de Pesquisa Energética, «Balanço Energético Nacional 2019: ano base 2018,» Ministério de Minas e Energia, Rio de Janeiro, 2019.

[4] J. A. Puerto Rico y S. S. S. I. L. Mercedes, «Genesis and consolidation of the Brazilian bioethanol: A review of policies and incentive mechanisms,» *Renewable and Sustainable Energy Reviews,* vol. 14, no 7, pp. 1874-1887, 2010.

[5] A. K. de Souza y C. E. de Farias, «Bioethanol in Brazil: Status, Challenges and Perspectives to Improve the Production,» de *Bioethanol Production from Food Crops*, Academic Press, 2019, pp. 417-443.

[6] S. L. Stattman, O. Hospes y A. P. Mol, «Governing biofuels in Brazil: A comparison of ethanol and biodiesel policies,» *Energy Policy,* vol. 61, pp. 22-30, 2013.

[7] L. L. Benites-Lazaro, N. A. Mello-Théry y M. Lahsen, «Business storytelling about energy and climate change: The case of Brazil's ethanol industry,» *Energy Research and Social Science,* vol. 31, pp. 77-85, 2017.

[8] L. G. Pereira, O. Cavalett, A. Bonomi, Y. Zhang, E. Warner y H. L. Chum, «Comparison of biofuel life-cycle GHG emissions assessment tools: The case studies of ethanol produced from sugarcane, corn, and wheat,» *Renewable and Sustainable Energy Reviews,* vol. 110, pp. 1-12, 2019.

[9] L. C. Cardoso, M. V. Bittencourt, W. H. Litt y E. G. Irwin, «Biofuels policies and fuel demand elasticities in Brazil,» *Energy Policy,* vol. 128, pp. 296-305, 2019.

[10] I. G. Cesca, M. Araújo y S. Bottrel, «Análise da demanda de combustíveis veiculares no Brasil entre 2004 e 2014,» *Revista de Economia e Agronegócio,* vol. 14, pp. 167-194, 2016.

[11] S. Marssal, d. O. C. Ribeiro y M. Vieira, «Uncertainty effects on production mix and on hedging decisions: The case of Brazilian ethanol and sugar,» *Energy Economics,* vol. 70, pp. 516-524, 2018.

[12] J. A. Moncada, J. A. Verstegen, J. A. Posada, M. Junginger, Z. Lukszo, A. Faaij y M. Weijnen, «Exploring policy options to spur the expansion of ethanol production and consumption in Brazil: An agent-based modeling approach,» *Energy Policy,* vol. 123, pp. 619-641, 2018.

[13] A. L. Martins, P. Wanke, Z. Chen y N. Zhang, «Ethanol production in Brazil: An assessment of main drivers with MCMC generalized linear mixed models,» *Resources, Conservation and Recycling,* vol. 132, pp. 16-27, 2018.

[14] V. A. Sobreiro, P. H. D. S. L. Araújo y M. S. Nagano, «Precificação do etanol utilizando técnicas de redes neurais artificiais,» *Revista de Administração-RAUSP,* vol. 44, no 1, pp. 46-58, 2009.

[15] C. A. Gonçalves, «Análise da previsão do preço do etanol hidratado no estado de São Paulo: uma aplicação do modelo arima,» *Brazilian Journal of Development,* vol. 5, no 10, pp. 17763-17778, 2019.

[16] S. R. Figueira, H. L. Burnquist y M. R. P. Bacchi, «Forecasting fuel ethanol consumption in Brazil by time series models: 2006–2012,» *Applied Economics,* vol. 42, no 7, pp. 865-874, 2010.

[17] F. S. Santiago, R. S. D. Mattos y F. S. Perobelli, «Um modelo integrado econométrico + insumo-produto para previsão de longo prazo da demanda de combustíveis no Brasil,» *Nova economia,* vol. 21, no 3, pp. 423-455, 2011.

[18] A. Sagheer y M. Kotb, «Time series forecasting of petroleum production using deep LSTM recurrent networks,» *Neurocomputing,* vol. 323, pp. 203-213, 2019.

[19] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu y Y. Zhang, «Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network,» *IEEE Transactions on Smart Grid,* vol. 10, no 1, pp. 841-851, 2017.

[20] R. A. Schwalbert, T. Amado, G. Corassa, L. Pott, P. Prasad y I. A. Ciampitti, «Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil,» *Agricultural and Forest Meteorology,* vol. 284, pp. 107886, 2020.

[21] Z. Cen y J. Wang, «Crude oil price prediction model with long short term memory deep learning based on prior knowledge data transfer,» *Energy,* vol. 169, pp. 160-171, 2019.

[22] H. Abbasimehr, M. Shabani y M. Yousefi, «An optimized model using LSTM network for demand forecasting,» *Computers & Industrial Engineering,* vol. 143, no 3, pp. 106435, 2020.

[23] W. Cao, X. Wang, Z. Ming y J. Gao, «A review on neural networks with random weights,» *Neurocomputing,* vol. 275, pp. 278-287, 2018.

[24] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou y M. Ettaouil, «Multilayer Perceptron: Architecture Optimization and Training,» *IJIMAI,* vol. 4, no 1, pp. 23-30, 2016.

[25] J. Bedi y D. Toshniwal, «Deep learning framework to forecast electricity demand,» *Applied Energy,* vol. 238, pp. 1312-1326, 2019.

[26] S. Hochreiter y J. Schmidhuber, «Long Short-Term Memory,» *Neural Computation,* vol. 9, no 8, pp. 1735-1780, 1997.

[27] S. Hochreiter, «The vanishing gradient problem during learning recurrent neural nets and problem solutions,» *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, pp. 107-116, 1998.

[28] C. Olah, «Understanding lstm networks,» 2015. [En línea]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs.

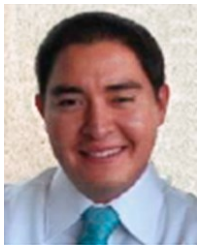
[29] C. P. Barros, L. A. Gil-Alana y P. Wanke, «Ethanol consumption in Brazil: Empirical facts based on persistence, seasonality and breaks,» *Biomass and Bioenergy,* vol. 63, pp. 313-320, 2014.

[30] Instituto de Pesquisa Econômica Aplicada, «Dados macroeconômicos de consumo e venda,» 2019. [En línea]. Available: http://www.ipeadata.gov.br/Default.aspx. [Último acceso: 27 Agosto 2019].

[31] T. G. Smith, «Docs. Pyramid: ARIMA estimators for Python,» 2018. [En línea]. Available: http://www.alkaline-ml.com/pmdarima/0.9.0/modules/generated/pyramid.arima.auto_arima.html. [Último acceso: 25 Octubre 2019].

Jorge Armando Puentes Márquez posee grado en Ingeniería Industrial por parte de la Corporación Universitaria del Caribe, Colombia (2016). Magíster en Ingeniería Industrial en formación por parte del Tecnológico Nacional de México en Celaya. Actualmente, desempeña labores de investigación en la Universidad de São Paulo y el Tecnológico Nacional de México en Celaya en tópicos relacionados a Redes Neuronales Artificiales e Industria 4.0.

Celma de Oliveira Ribeiro posee grado en Ciencias de la Computación por parte de la Universidad de São Paulo (1980), Maestría en Matemática Aplicada por parte de la Universidad de São Paulo (1987), Doctorado en Ingeniería (Engenharia de Produção) por parte de la Universidad de São Paulo (1997) y Post-doctorado por parte de la Universidad de Oporto (2002). Actualmente es profesora asociada en la Universidad de São Paulo y revisora de varias revistas nacionales e internacionales.



Edgar Augusto Ruelas Santoyo posee licenciatura y maestría en Ingeniería Industrial por parte del Tecnológico Nacional de México en Celaya, México (2008 y 2011 respectivamente). Doctorado en Ingeniería Industrial por parte del Posgrado Interinstitucional en Ciencia y Tecnología (PICyT) de CIATEC, México (2015). Actualmente es profesor titular en el Instituto Tecnológico Superior de Irapuato. Sus principales intereses de investigación son: Estadística Industrial, Procesamiento digital de imágenes, Lógica Difusa, y Redes Neuronales.

Vicente Figueroa Fernández posee Maestría en Ciencias en Ingeniería Industrial por parte del Tecnológico Nacional de México en Celaya (2000). Actualmente es profesor investigador, miembro de un cuerpo académico, y coordinador de la Maestría en Ingeniería Industrial del Tecnológico Nacional de México en Celaya. Además, se desempeña como tutor de incubadora empresarial y ha dirigido múltiples tesis de maestría. Sus áreas de investigación son Logística y operaciones, Industria 4.0, y Diseño y mejora de procesos y productos.