Información Tecnológica
Vol. 26(1), 13-22 (2015)
doi: 10.4067/S0718-07642015000100003

Análisis de la Entrada en Simulación Estocástica

David F. Muñoz[1,,2] y Diana Villafuerte[1]
(1) Departamento de Ingeniería Industrial y Operaciones, Instituto Tecnológico Autónomo de México, Río Hondo # 1, Colonia Tizapan San Angel, 01080 México, Distrito Federal, México.
(2) Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX 78712, USA.
(e-mail: davidm@itam.mx y villafuerteolvera@hotmail.com)

Recibido Jun. 13, 2013; Aceptado Ago. 20, 2014; Versión final recibida Oct. 19, 2014

Resumen

En este artículo se presenta una revisión de las principales técnicas que se han propuesto para el análisis de los componentes aleatorios de una simulación estocástica. Se discuten dos temas principales: el modelado clásico utilizando familias univariadas de distribuciones y el modelado avanzado de la entrada. En el primer tema se introduce el uso de una librería, llamada *Analizador Simple* (*Simple Analyzer)*, que puede ser descargada libremente desde la red, mientras que en el segundo tema se discuten las técnicas para el modelado de la entrada que actualmente no están incluidas en el software disponible para el análisis de la entrada de experimentos por simulación. Se analiza también el ajuste de datos de entradas multimodales, correlacionadas y dependientes del tiempo, y se incorpora la incertidumbre paramétrica en el modelo de entrada.

Palabras clave: simulación estocástica, análisis de la entrada, ajuste de distribuciones, componentes aleatorios

Input Analysis for Stochastic Simulations

Abstract

In this article, a review of the main techniques proposed for the analysis of random components in stochastic simulations is presented. Two main topics are discussed: classical modeling using univariate distributions and advanced modeling techniques. The use of a library is introduced in the first topic; this library is named *Simple Analyzer* and can be downloaded for free from the web. In the second topic, techniques for input analysis that are currently not included in available commercial software of simulation experiments are discussed. Also the fitting of data of multimodal, correlated and time-dependent input components, and the incorporation of parameter uncertainty in the input model are analyzed

Keywords: stochastic simulation, input analysis, distribution fitting, random components

INTRODUCCIÓN

La simulación está ampliamente reconocida como una técnica efectiva para construir pronósticos, evaluar riesgos, animar e ilustrar la evolución de un sistema en muchas áreas (Kelton et al., 2012). Cuando existe incertidumbre en el comportamiento de algunos de los componentes del modelo de simulación, estos componentes aleatorios deben modelarse utilizando distribuciones de probabilidad y/o procesos estocásticos que son generados durante la simulación. Con la finalidad de ilustrar cómo aparece naturalmente el concepto de componente aleatorio en un experimento por simulación, supongamos que se desea simular la congestión en un cajero automático (ATM) durante cierto lapso de tiempo, digamos, en el intervalo de tiempo [*0, t*]. Con este fin, podemos asumir que el ATM está desocupado al tiempo *0*.

Sea A_1 el tiempo de llegada del primer cliente, y A_i el lapso de tiempo entre la llegada del cliente *i-1* y el cliente *i*, *i = 2, 3,* …; además, sea S_i el tiempo que demora la atención del cliente *i* (para *i = 2, 3,* …). Entonces, el tiempo de espera en la fila de espera del primer cliente es $W_q(1) = 0$, y para los clientes *i = 2, 3,* … los tiempos de espera pueden obtenerse utilizando la relación de recurrencia de Lindley (1952):

$$W_q(i) = \max\{W_q(i-1) + S_{i-1} - A_i, 0\}. \quad (1)$$

La congestión en el ATM puede simularse (e.g., en una hoja de cálculo) utilizando la ecuación (1) para producir los tiempos de espera (en la cola) de los clientes, hasta que el tiempo de salida del último cliente exceda por primera vez el tiempo *t*. Sin embargo, para simular las salidas definidas en (1), deben producirse dos cadenas de entradas aleatorias: $A_1, A_2, \ldots$, y $S_1, S_2, \ldots$. Por lo general (aunque no siempre), estas cadenas se generan bajo la suposición de que los tiempos entre llegadas sucesivas son variables aleatorias independientes e idénticamente distribuidas (i.i.d.) con función de densidad $f_A(x, \theta_A)$, y los tiempos de atención son variables aleatorias i.i.d. con función de densidad $f_S(x, \theta_S)$, donde θ_A y θ_S son parámetros cuyo valor es conocido. Bajo estas suposiciones, se pueden utilizar los métodos clásicos para la generación de variables aleatorias (Law, 2014) para obtener las cadenas requeridas $A_1, A_2, \ldots$, y $S_1, S_2, \ldots$.

Un componente aleatorio (también llamado entrada aleatoria) de una simulación estocástica es una secuencia $U_1, U_2, \ldots$ de valores aleatorios (a veces multivariados) que se requieren como entrada de la simulación. Cuando las U_i son i.i.d., el componente aleatorio se identifica por la correspondiente distribución de probabilidades, que generalmente es un miembro de una familia paramétrica. El análisis de la entrada de una simulación estocástica consiste en modelar adecuadamente los componentes aleatorios que están presentes en el modelo de simulación, y es particularmente relevante cuando se dispone de datos provenientes del sistema en estudio. En el caso de la simulación del ATM, sería conveniente recopilar observaciones del sistema real, por ejemplo, a través de una muestra $x_1, x_2, \ldots, x_n$ de observaciones de los tiempos entre llegadas de los clientes y, a partir de ella, encontrar la distribución de entrada $f_A(x, \theta_A)$ que se ajuste mejor a los datos observados $x_1, x_2, \ldots, x_n$. En este punto, introducimos la convención de que una "familia de distribuciones $f(x, \theta)$" hace referencia a la correspondiente función de probabilidades (f.p.) cuando la distribución es discreta, o a la correspondiente función de densidad (f.d.) si la distribución es continua. Es conveniente mencionar que, si bien en muchas aplicaciones de la simulación se remarca la importancia de la recolección de datos para ajustar los parámetros de un componente aleatorio (Callejas-Cuervo et al., 2014; Martins et al., 2012, Vargas y Giraldo, 2014), en otras los parámetros de los componentes aleatorios se establecen por opiniones subjetivas de expertos (Vanalle et al., 2012; Muñoz y Muñoz, 2010).

En este artículo discutimos las principales técnicas que se han propuesto para el análisis de la entrada de una simulación estocástica. En la siguiente sección mencionamos las principales familias de distribuciones que son útiles para modelar un componente aleatorio univariado, así como ilustramos el uso de la librería que se distribuye libremente con este artículo. En las siguientes secciones revisaremos las técnicas para el modelado de la entrada que no están disponibles en el software comercial para el análisis de la entrada, incluyendo el modelado de entradas multimodales y correlacionadas, y de procesos estocásticos con estructura dependiente del tiempo; así como la incorporación de la incertidumbre paramétrica que es inducida por el proceso de estimación de los parámetros del modelo de entrada. En la última sección presentamos nuestras conclusiones.

ANÁLISIS DE LA ENTRADA CLÁSICO

El enfoque clásico para el análisis de la entrada de un componente aleatorio (univariado) consiste en identificar una familia de distribuciones que se ajuste a nuestras necesidades, bajo la suposición de que el componente aleatorio consiste de observaciones i.i.d. de cierta distribución de probabilidades (que pertenece a una familia paramétrica). La buena selección de la familia puede depender de que las propiedades particulares de la familia son las adecuadas para los experimentos por simulación y/o de que la

familia de distribuciones es la que se ajusta mejor a una muestra $X_1, X_2,\ldots, X_n$ de observaciones (reales) del componente aleatorio. Cuando no existen observaciones del componente aleatorio, se acostumbra utilizar la distribución triangular (Law, 2014), estableciendo los parámetros con base en la opinión de expertos. Cuando sí se dispone de una muestra de observaciones $X_1, X_2,\ldots, X_n$, se identifica la familia de distribuciones $f(x,\theta)$ aplicando los siguientes pasos a cada una de las familias candidatas: (i) encontrar un buen estimador para el parámetro θ, ii) agrupar las observaciones y comparar (visualmente) la gráfica de las frecuencias relativas versus la gráfica de las probabilidades correspondientes a $f(x,\theta)$, y iii) calcular los estadísticos de bondad de ajuste. La librería *Simple Analyzer* permite ejecutar estos pasos, desde una hoja de cálculo, para las familias de distribuciones más conocidas.

Las familias de distribuciones continuas que se consideran en el *Simple Analyzer* son: uniforme, triangular, exponencial, Weibull, gamma, normal, lognormal, beta, Johnson SB y Johnson SU. Las primeras dos (uniforme y triangular) son apropiadas como modelo de entrada cuando no se dispone de observaciones del componente aleatorio, y las otras distribuciones pueden utilizarse cuando existen observaciones $x_1, x_2,\ldots, x_n$ del componente aleatorio. Por lo general, los parámetros que identifican al miembro específico de la familia de distribuciones pueden clasificarse dentro de alguno de los siguientes tipos: de localización, de escala,de dispersión o de forma (ver detalles en Law, 2014). Las familias de distribuciones discretas que hasta el momento están consideradas en el *Simple Analyzer* son: binomial, binomial negativa y Poisson. Las características y propiedades más importantes de estas distribuciones, incluyendo el cálculo del estimador máximo verosímil (EMV) están resumidas en la Tabla 6.3 (para distribuciones continuas) y en la Tabla 6.4 (para distribuciones discretas) de Law (2014).

Una distribución discreta que puede ser útil como modelo de entrada es la llamada distribución empírica, que a continuación discutimos. Dadas observaciones (univariadas) $x_1, x_2,\ldots, x_n$ de una variable aleatoria (VA) X, la distribución empírica correspondiente a este conjunto de observaciones se define por:

$$F_n(x) = \frac{\sum_{i=1}^{n} I[x_i \le x]}{n}, \tag{2}$$

para $-\infty < x < \infty$, donde $I[y \le x]$ es 1 cuando $y \le x$ y 0 de otra forma. La distribución empírica definida en (2) es la distribución de probabilidades acumuladas correspondiente a una VA discreta que asigna (a cada valor x_i) una probabilidad que es proporcional al número de veces que el valor (x_i) se repite en la muestra. Esta alternativa puede ser útil como modelo de entrada, cuando se desea que los datos observados determinen directamente al componente aleatorio, y es particularmente relevante cuando el componente aleatorio se identifica como una variable discreta que toma valores en un conjunto finito $E = \{x_1,\ldots, x_k\}$. En este caso particular, las probabilidades de la distribución empírica son estimadores consistentes de las correspondientes probabilidades $P[X = x_j]$, $j = 1,\ldots,k$, respectivamente. Sin embargo, cuando el componente aleatorio debe modelarse como una VA continua, el uso de la distribución empírica como modelo de entrada tiene las siguientes desventajas:

i) El número de valores diferentes en una muestra es finito y, en una simulación típica (larga), es muy probable que el muestreo de la distribución empírica produzca valores repetidos en algún momento, lo que es inaceptable para una variable continua.

ii) Los valores observados representan sólo el escenario observado cuando se registraron los datos, y pudieran no ser representativos de lo que podría pasar en el futuro; por ejemplo, debido a redondeos o a un número pequeño de observaciones.

iii) En muchas situaciones, el investigador puede estar interesado en simular diferentes escenarios para la entrada. Por ejemplo, si el componente aleatorio representa la demanda de un producto, podría ser de interés la simulación de escenarios correspondientes a diferentes demandas esperadas; en este caso, modificar la distribución empírica para producir diferentes demandas esperadas puede ser complicado y, por el contrario, adaptar una distribución $f(x,\theta)$ para producir diferentes escenarios puede ser tan simple como cambiar el valor del parámetro θ.

USO DEL SIMPLE ANALYZER

Desde la página web http://ciep.itam.mx/~davidm/sofdop.htm se puede descargar el archivo Excel de nombre *SimpleAnalyzer.xls*, que sirve de interfaz para facilitar el análisis de la entrada. Para cargar la librería de procedimientos que se llaman desde el archivo Excel, debe instalarse previamente el software *Random Number Generators*, que está disponible en la misma página web. En cualquiera de las hojas de

cálculo de nombre *Continuous* o *Discrete* se pueden introducir las observaciones del componente aleatorio que se desea modelar, y los botones disponibles en cada hoja permiten ajustar cada una de las distribuciones mencionadas; la hoja *Continuous* corresponde a las distribuciones continuas, y la hoja *Discrete* corresponde a las distribuciones discretas.

Cuando se usa el *Simple Analyzer* para ajustar una familia de distribuciones $f(x,\theta)$ a un conjunto de datos $x_1,\ldots, x_n$, se calcula el EMV $\hat{\theta}$. Algunas familias (exponencial, Weibull, gamma, lognormal, beta y Johnson SB) pueden incluir un traslado (además de los parámetros de escala y/o forma); en este caso, el software sugiere un valor estimado para el traslado, y el usuario puede aceptar este valor o proponer otro valor (antes de obtener el EMV para los otros parámetros).

Para los casos en que el EMV no tiene una expresión analítica, el *Simple Analyzer* resuelve numéricamente las ecuaciones descritas en la Tabla 6.3 de Law (2007), excepto para el caso de la distribución Weibull, en el que se usa el algoritmo propuesto en Qiao y Tsokos (1994). Luego de calcular el EMV $\hat{\theta}$, se construye un histograma con el número de intervalos (k) especificado por el usuario. El valor sugerido corresponde a la conocida fórmula de Sturges $k \approx 1+\log_2(n)$. Para cada intervalo de clase $i = 1,\ldots,n$, el *Simple Analyzer* proporciona un límite inferior L_i, un límite superior U_i, la frecuencia absoluta $Af_i = \sum_{j=1}^{n} I[L_i < x_j \le U_i]$, la probabilidad observada $\hat{p}_i = Af_i / n$, y la probabilidad esperada $p_i = P[L_i < X \le U_i]$, donde X tiene f.d. (f.p.) $f(x,\hat{\theta})$. Finalmente, el *Simple Analyzer* calcula los estadísticos de bondad de ajuste de Kolmogorov-Smirnov (K-S), chi- cuadrado y el error cuadrático promedio (ver detalles en Law, 2014).

Nótese que los botones al lado izquierdo de las hojas *Continuous* y *Discrete* producen una muestra de observaciones simuladas de la correspondiente distribución; para tal fin, el usuario debe indicar el tamaño de la muestra (en la celda B2) y el número de intervalos del histograma (en la celda B4). Luego de presionar un botón del lado izquierdo (e.g., el botón "Simulate Uniform"), los valores generados se muestran en la columna D, se construye el histograma y se grafican las probabilidades observadas (corresponden a las barras) y las probabilidades esperadas (los puntos unidos por líneas). El botón "Initialize Seed" permite al usuario inicializar las semillas del generador de números aleatorios con la finalidad de reproducir alguna salida deseada.

En la hoja de nombre *Example1* se presentan (en la columna A) datos de 990 observaciones del tiempo (en horas) entre las llamadas que llegan a un centro de atención, y en la Tabla 1 se presentan las valores del estadístico K-S obtenidos luego de ajustar cada una de las distribuciones disponibles en la hoja *Continuous*. En todas las corridas se usaron los valores sugeridos para el traslado, máximo y multiplicador. Se sugiere al lector verificar estos resultados copiando los datos en la hoja *Continuous*. Como puede observarse de la tabla, el mejor ajuste corresponde a la distribución exponencial, siendo el traslado sugerido muy cercano a cero. Se puede verificar que si se propone un traslado de cero, se obtendrá un valor K-S muy parecido (el mismo hasta 5 decimales). Nótese que, aunque la distribución exponencial es un caso particular de las familias Weibull y gamma, las estadísticas K-S obtenidas con cada una de estas familias son ligeramente mayores a la de la distribución exponencial.

Tabla 1. Estadísticas K-S para los datos de la hoja Example1

Distribución	Uniforme	Triangular	Exponencial	Weibull	Gamma	Normal	Lognormal	Beta	Johnson SB
K-S	0.6045	0.4529	0.0157	0.0160	0.0170	0.1524	0.0856	0.0234	0.0675

MODELADO AVANZADO DE LA ENTRADA

Existen varios software especializados en el análisis de la entrada de las simulaciones estocásticas, incluyendo el Arena Input Analyzer, Stat::Fit, @RISK y el ExpertFit. Además, el software comercial de propósito general (e.g., SAS o SPSS) generalmente incluye procedimientos que son útiles para el análisis de la entrada. El software comercial disponible permite el ajuste de familias (univariadas) de distribuciones a un conjunto de datos (como el *Simple Analyzer*). Sin embargo, en algunas situaciones el análisis de la entrada de una simulación puede requerir de metodologías que todavía no están disponibles en el software comercial y, en particular, por alguna de las siguientes razones: 1) El ajuste de una familia estándar a un conjunto de datos $x_1,\ldots, x_n$ es pobre, debido a las limitaciones de forma de las familias estándar; 2) El componente aleatorio que se requiere generar es una secuencia de entradas multivariadas $U_1, U_2,\ldots$, cuyas componentes univariadas no son independientes. Nótese que si las componentes univariadas son independientes, se pueden analizar por separado utilizando familias univariadas; 3) La estructura del proceso de entrada cambia en el tiempo; y 4) El número de observaciones (n) del componente aleatorio es

pequeño, lo que ocasiona que la incertidumbre en el valor del parámetro θ (inducida por el proceso de estimación) sea significativa, y deba ser incorporada en el análisis de la salida de la simulación.

Limitaciones de Forma de las Familias Estándar

Aunque algunas familias con tres parámetros (e.g., beta o gamma) pueden ajustarse a una variedad de formas, existen familias con cuatro parámetros, como las familias de Pearson y de Johnson (Law 2014) que tienen formas más flexibles, aunque estas distribuciones siguen siendo inadecuadas para ajustar datos con más de una moda. Notar que el EMV para ciertas distribuciones con traslado puede no existir en algunos casos, donde deben utilizarse otros métodos de estimación (Nagatsuka, 2013).

Wagner y Wilson (1996) proponen un método para ajustar datos con formas arbitrarias que consiste en incorporar un número suficiente de parámetros y polinomios de varias formas, con la finalidad de obtener una distribución (llamada distribución de Bézier) que exhibe una forma apropiada para los datos. Se han propuesto también métodos no-paramétricos, basados en estimación de densidades (Scott, 2012; Moreira y Van Keilegom, 2013; Wang y Wetelecki, 2013; Gu et al., 2013; Papp y Alizadeh, 2014), que también son útiles para ajustar datos con formas arbitrarias. Sin embargo, la adaptación de estos métodos para producir diferentes escenarios puede ser tan complicada como en el caso de usar la distribución empírica.

En el caso particular en que los datos exhiben más de una moda, la mayoría de las familias de distribuciones no son capaces de proporcionar un buen ajuste. Como ilustramos con el siguiente ejemplo, el modelado de datos multimodales puede abordarse utilizando mezclas de distribuciones. Supongamos que $f_1(x), f_2(x), \ldots, f_k(x)$ son las f.d. (f.p.) correspondientes a k diferentes distribuciones, respectivamente; una mezcla de estas distribuciones es la f.d. (f.p.) que tiene la forma:

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \ldots + \alpha_k f_k(x), \tag{3}$$

para $-\infty < x < \infty$, donde $\alpha_i \geq 0$, $i = 1, \ldots, k$, y $\sum_{i=1}^{k} \alpha_i = 1$.

En la columna A de la hoja de cálculo *Example2* se proporcionan observaciones del tiempo (en minutos) que toma el registro de los clientes en una oficina de renta de autos. En la columna B se indica si el cliente correspondiente está viajando solo o con su familia. Si ignoramos la información de la columna B y ajustamos los datos de la columna A con cada una de las familias disponibles en la hoja *Continuous*, encontraremos que el mejor estadístico K-S corresponde a la distribución lognormal, $1.01+LoN(-0.67,1.15^2)$ (ver la segunda fila de la Tabla 2) y, de acuerdo con la Fig. 1 y el valor K-S de *0.1189*, podemos concluir que el ajuste no es satisfactorio. Una rápida inspección a la gráfica de la Fig. 1 sugiere que este pobre ajuste se debe a que los datos son bimodales.

Tabla 2. Estadísticas K-S para los tres conjuntos de datos de la hoja *Example2*

Distribución	Uniforme	Triangular	Exponen.	Weibull	Gamma	Normal	Lognormal	Beta
K-S todos	0.5779	0.3916	0.1425	0.1466	0.1476	0.2724	0.1189	0.2401
K-S solo	0.1913	0.0302	0.2175	0.0404	0.0754	0.0800	0.1409	0.0487
K-S familia	0.2315	0.0567	0.3287	0.0905	0.1317	0.0757	0.1687	0.0563

Si consideramos la información de la columna B, podríamos ajustar por separado los datos de los clientes que viajan solos y luego los datos de los clientes que viajan con familia. En la columna C se presentan las $n_1=847$ observaciones de los clientes que viajan solos, y en la columna D se presentan los datos de los $n_2=131$ clientes que viajan con familia. El mejor ajuste para los datos de la columna C corresponde a una distribución triangular con *K-S=0.0302* (ver la fila 3 de la Tabla 2) y, para los datos de la columna D, el mejor ajuste corresponde a una beta con *K-S=0.0563* (ver la fila 4 de la Tabla 2). Como se ilustra en la Fig. 2, ambos ajustes son aceptables.

Finalmente, nótese que la proporción de clientes que viajan solos es $\alpha_1 = 847/978 \approx 0.8661$, por lo que una buena propuesta para modelar la entrada correspondiente a los datos de la columna A es una mezcla cuya f.d. tiene la forma de la ecuación (2), con $k=2$, $\alpha_1=0.8661$, $\alpha_2=1-\alpha_1$, $f_1(x)$ es la f.d. de una VA distribuida triangularmente y $f_2(x)$ es la f.d. de una VA distribuida como beta (con traslado de *2.9* y multiplicador de *1.96*). En la hoja *Example2* se ha implementado un procedimiento que ajusta los datos de la columna A a esta mezcla; y se ha obtenido la gráfica de la Fig. 3 con *K-S=0.0261* (mucho mejor que el ajuste lognormal de la Fig. 1). Por supuesto que no estamos sugiriendo seguir siempre este mismo procedimiento para ajustar mezclas; ya que existe software especializado para ajustar mezclas y, en particular, SS tiene el procedimiento FMM (ver Kesler y McDowell, 2012).

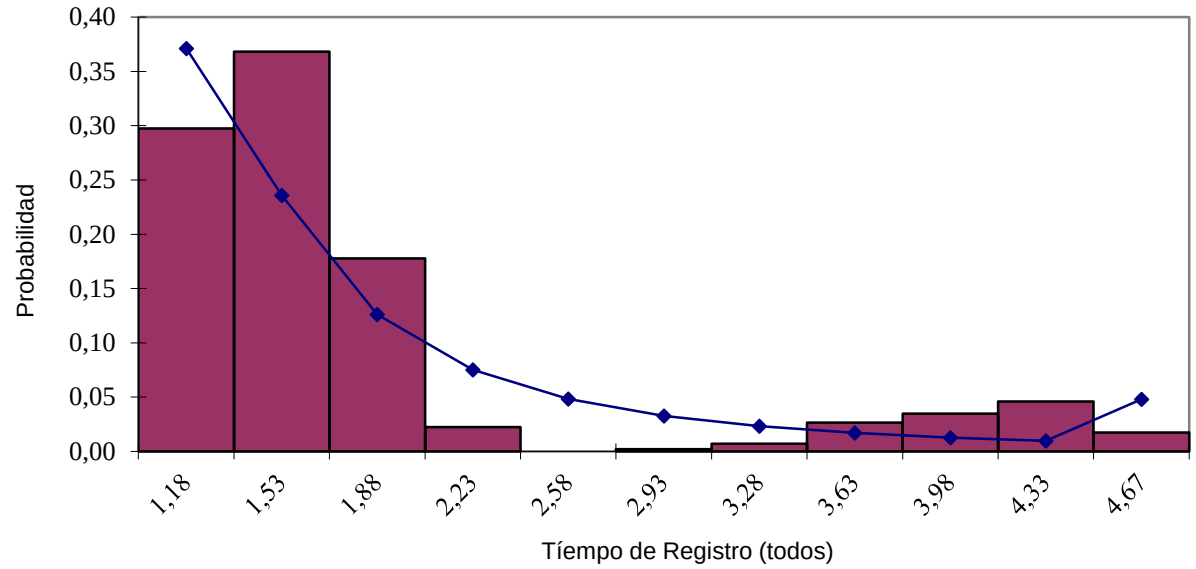
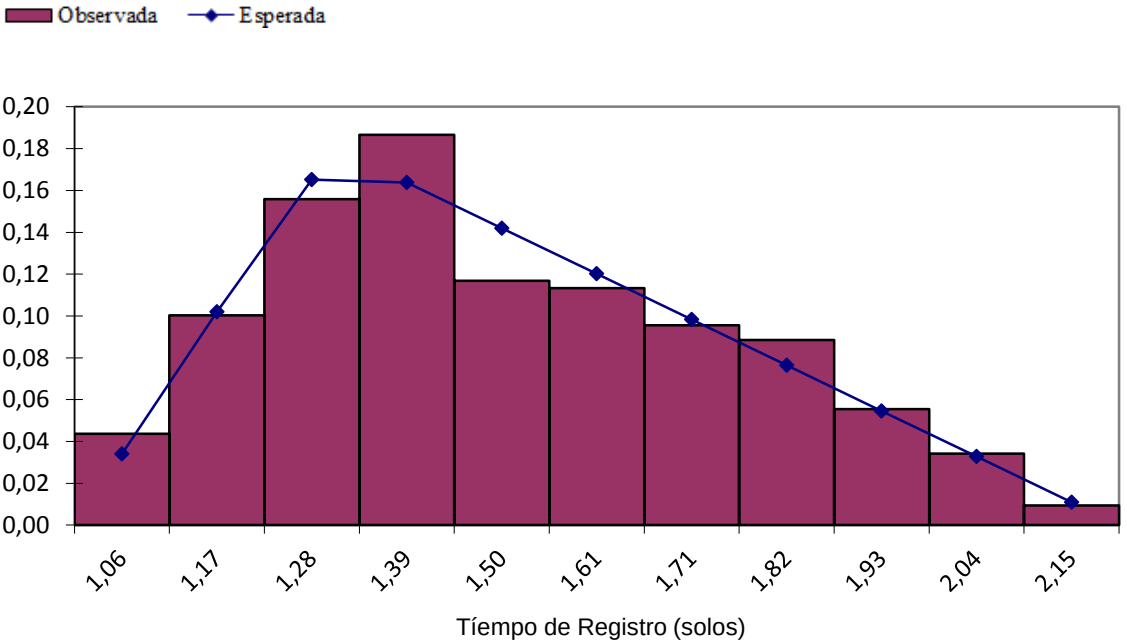


Fig. 1. Probabilidades esperadas y observadas para el ajuste lognormal de la hoja *Example2*.



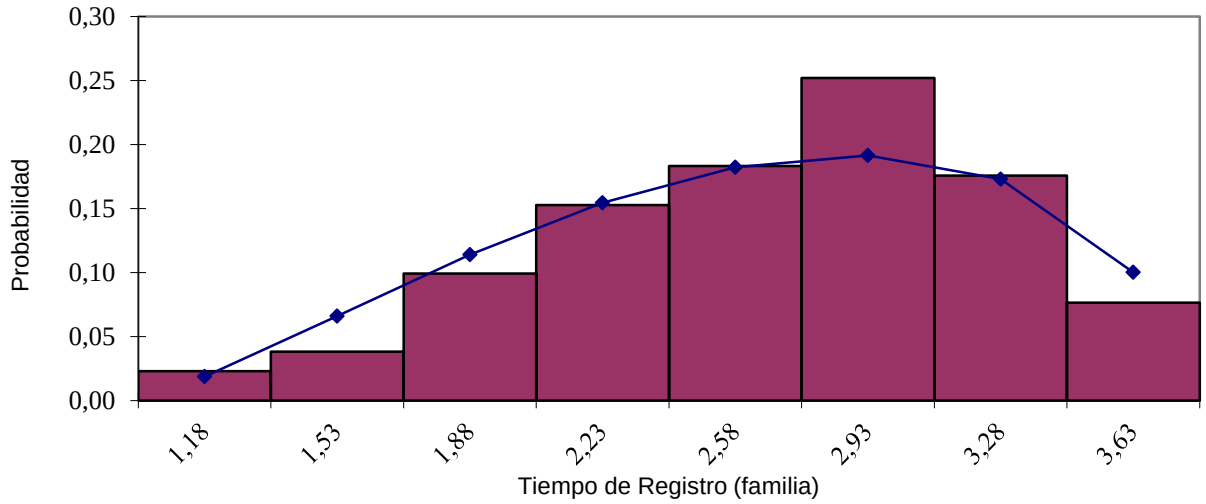


Fig. 2. Probabilidades esperadas y observadas para los ajustes triangular y beta de la hoja *Example2*.

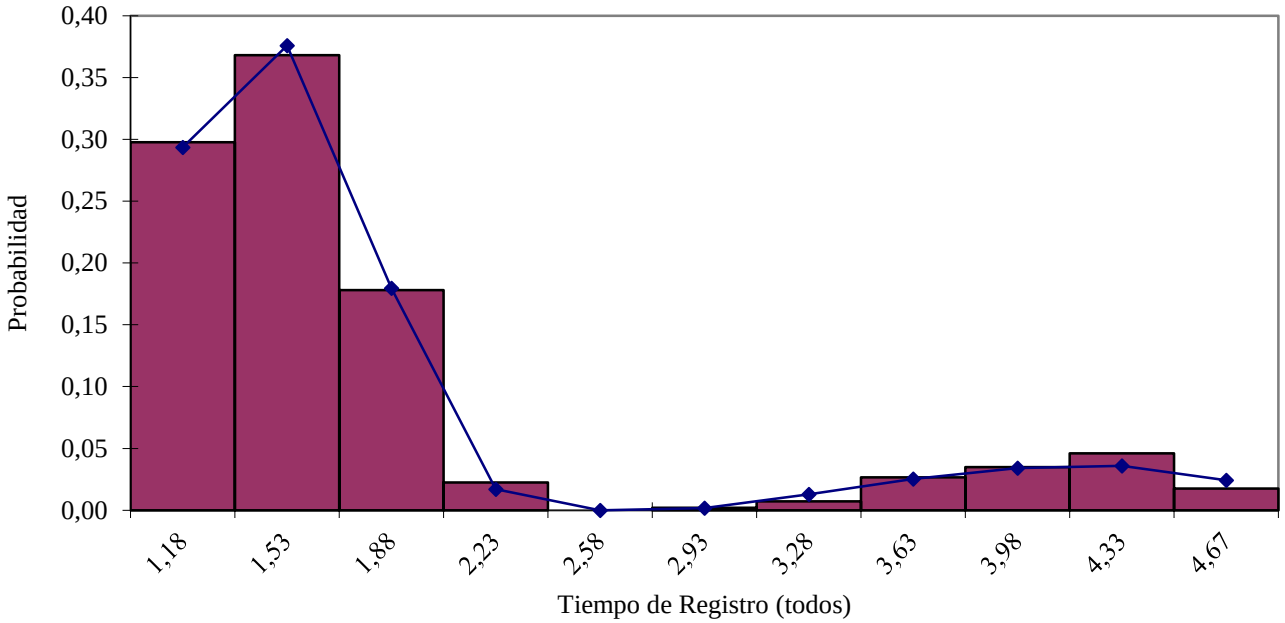


Fig. 3. Probabilidades esperadas y observadas para el ajuste de la mezcla de la hoja *Example2*.

Entradas Multivariadas y Dependientes del Tiempo

En muchas situaciones se puede requerir el modelado de un componente aleatorio cuya estructura varía en el tiempo; por ejemplo, se puede modelar un proceso de llegadas de clientes asumiendo que los tiempos entre llegadas son i.i.d., exponenciales con esperanza $\beta>0$, en cuyo caso el número de llegadas (N_t) en el intervalo [*0,t*] sigue la distribución de Poisson con esperanza λt; en este caso, el proceso de llegadas es un proceso de Poisson con tasa $\lambda=1/\beta$. Sin embargo, en un proceso de Poisson la tasa de llegadas es constante en el tiempo, y esta situación puede ser problemática; ya que en muchas situaciones de la vida real la tasa varía en el tiempo (imaginar la diferencia entre la tasa de autos que ingresan a una vía rápida en hora pico y a la medianoche). Para modelar un proceso de llegadas con tasa variable, se utiliza el proceso de Poisson no-estacionario, donde la tasa de llegadas es una función no-negativa $\lambda(t)$, y en este caso el número de llegadas en el intervalo [*0,t*] sigue una distribución de Poisson con esperanza $\Lambda(t)=\int_0^t \lambda(t)dt$. La manera más sencilla de especificar una tasa de llegadas que varía en el tiempo consiste en asumir que es constante por tramos de cierta duración (por ejemplo, 15 minutos), y que puede saltar hacia arriba o hacia abajo al final de cada tramo. Para estimar la tasa en cada tramo, basta con calcular el correspondiente promedio de llegadas con base en conteos de observaciones múltiples de cada periodo. Leemis (1991) propone una metodología no-paramétrica para estimar una función de llegadas $\lambda(t)$ constante por tramos, con base en observaciones del proceso de llegadas. Además, se han propuesto métodos para estimar funciones de llegadas $\lambda(t)$ que tienen comportamientos especiales (Kuhl, 2006) y que, además del tiempo, dependen de otros parámetros como la localización (Alizadeh y Papp, 2013).

Los modelos de series de tiempo autorregresivos (AR) o autorregresivos de promedios móviles (ARMA) pueden también ser útiles para modelar una entrada, ya que estos modelos son muy usados para el pronóstico (e.g., de la demanda). Cario y Nelson (1998) desarrollaron procesos llamados ARTA para modelar series de tiempo con distribuciones marginales y estructura de correlación dadas, y Biller y Nelson (2005) proponen una metodología para ajustar un proceso ARTA con distribuciones marginales de la familia de Johnson.

En algunas situaciones podríamos requerir que el componente aleatorio $U_1, U_2,\ldots$ de la simulación sea multivariado (i.e., U_i es un vector aleatorio, $i=1,2,\ldots$), donde las componentes (univariadas) de cada vector están correlacionadas (aunque U_i y U_j son independientes para $i \neq j$). Si las distribuciones marginales de las componentes de las U_i son normales, la distribución conjunta está determinada por la estructura de correlación, por lo que el ajuste de un conjunto de datos (multivariados) a una distribución normal se reduce a estimar la estructura de correlación. Si se requieren distribuciones marginales que no son normales, una metodología que parece funcionar bien consiste en transformar la distribución normal multivariada a una distribución uniforme multivariada. Luego se aplica transformación inversa a los componentes de la

distribución multivariada uniforme para obtener las distribuciones marginales y la estructura de correlación deseada (ver detalles, en Clemen y Reilly, 1999; Biller y Corlu, 2012; Kim et al., 2013).

Incorporación de la Incertidumbre Paramétrica

Las metodologías discutidas, hasta el momento, sugieren la propuesta de un modelo de entrada y la correspondiente estimación de los parámetros a partir de datos disponibles. Estamos asumiendo implícitamente que las simulaciones se correrán fijando el modelo y el valor de los correspondientes parámetros, método que tradicionalmente se ha venido usando en los experimentos por simulación estocástica. Este enfoque es apropiado cuando el número de observaciones (*n*) del conjunto de datos es grande, y la incertidumbre paramétrica es despreciable, debido a la consistencia de los estimadores. Sin embargo, cuando el tamaño de las muestras es pequeño, puede existir una incertidumbre significativa sobre el verdadero valor de los parámetros, y es importante incorporar esta incertidumbre en el análisis de los experimentos por simulación.

Se han propuesto varios métodos para incorporar la incertidumbre paramétrica en experimentos por simulación; algunos de los cuales son consistentes con métodos de la estadística "frecuentista". Por ejemplo, Cheng y Holland (2004) proponen utilizar el bootstrap paramétrico, que consiste en re-muestrear (por simulación) los parámetros, a partir de una función de verosimilitud donde se ha reemplazado el valor del parámetro por el EMV calculado a partir de los datos reales. Por otro lado, el enfoque Bayesiano que sugerimos en este artículo permite incorporar la incertidumbre paramétrica por medio de una distribución de probabilidades, y es consistente con el enfoque descrito, en Chick (2001), Geweke y Whiteman (2006) y Muñoz et al. (2013), donde los autores discuten la incorporación de la incertidumbre paramétrica en modelos de pronóstico que utilizan la simulación.

Este enfoque Bayesiano se resume en dos pasos. Primero se mide la incertidumbre paramétrica construyendo la densidad posterior $p(\theta|x)$ a partir de los datos disponibles (x) y de la distribución a priori $p(\theta)$. A continuación se generan las observaciones simuladas $W = g(Y(s), 0 \le s \le T; \Theta)$ muestreando el valor del parámetro Θ de la distribución posterior $p(\theta|x)$ y luego se corre la simulación con el valor muestreado de Θ para generar los componentes aleatorios. En corto, la principal diferencia es que, bajo el enfoque Bayesiano, el valor del parámetro no se fija, sino que se simula a partir de la distribución posterior $p(\theta|x)$, antes de correr una simulación. Algunas aplicaciones reales de este enfoque se reportan en Muñoz et al. (2010) y Muñoz y Muñoz (2011).

CONCLUSIONES

En este artículo se han discutido los principales problemas y técnicas relacionados con el modelado de los componentes aleatorios de una simulación estocástica. Luego de un ejemplo introductorio, hemos discutido el enfoque clásico para ajustar una familia estándar de distribuciones a un conjunto de datos $x_1, x_2, \ldots, x_n$ disponibles del componente aleatorio. Posteriormente se ilustra el uso de la librería (de distribución libre) *Simple Analyzer*, y finalmente se discuten problemas relacionados con el análisis de la entrada que no son abordados por el software comercial disponible, incluyendo el modelado de entradas multivariadas y de procesos estocásticos no estacionarios, y la incorporación de la incertidumbre paramétrica. Deseamos remarcar que, a pesar de los avances recientes en velocidad de cómputo y visualización, existen importantes problemas relacionados con el análisis de le entrada de los experimentos por simulación que los usuarios tienen que abordar sin la ayuda del software comercial especializado.

AGRADECIMIENTOS

Este trabajo se ha desarrollado con el apoyo de la Asociación Mexicana de Cultura A.C., y los autores desean expresar su agradecimiento por el apoyo proporcionado por el CONACYT a través del convenio no. 206107. Asimismo, las sugerencias y comentarios del Editor y de un Árbitro anónimo han sido muy valiosas para mejorar la presentación, en forma y fondo, de este artículo. Esta investigación se realizó durante la estancia, como Visiting Scholar, del primer autor en la Universidad de Texas en Austin.

REFERENCIAS

Alizadeh, F. y D. Papp; *Estimating arrival rate of nonhomogeneous Poisson processes with semidefinite programming*, Annals of Operations Research, 208(1), 291-308 (2013).

Biller, B., y B. L. Nelson; *Fitting time-series input processes for simulation*, Operations Research, 53(3), 549–559 (2005).

Biller, B. y C. G. Corlu; *Copula-based multivariate input modeling*, Surveys in Operations Research and Management Science, 17(2), 69-84 (2012).

Callejas-Cuervo, M.; H. A. Valero-Bustos y A. C. Alarcón-Aldana; *Agentes de software como herramienta para medir la calidad de servicio prestado en un sistema de transporte público colectivo urbano*, Información. Tecnológica, 25(5), 147-154 (2014).

Cario, M. C. y B. L. Nelson; *Numerical methods for fitting and simulating autoregressive-to-anything processes*, INFORMS Journal on Computing, 10(1), 72–81 (1998).

Cheng, R. C. H. y W. Holland; *Calculation of confidence intervals for simulation output,* ACM Transactions on Modeling and Computer Simulation, 14 (4), 344-362 (2004).

Chick, S. E.; *Input distribution selection for simulation experiments: accounting for input uncertainty*, Operations Research, 49 (5), 744-758 (2001).

Clemen, R. T. y T. Reilly; *Correlations and copulas for decision and risk analysis*, Management Science 45(2), 208–224 (1999).

Geweke, J. H., y C. H. Whiteman; *Bayesian forecasting*, en G. Elliot, C. Granger, & A. Timmermann (Eds.), Handbook of Economic Forecasting, 3-80, North-Holland (Holanda) (2006).

Gu C., Y. Jeon y Y. Lin; *Nonparametric density estimation in high dimensions*, Statistica Sinica, 23(3), 1131-1153 (2013).

Kelton W. D., J. S. Smith, D. T. Sturrock y D. F. Muñoz; *Simio y Simulación, Modelado, Análisis, Aplicaciones* (2a Ed.), Simio LLC, Sewickley (USA) (2012).

Kesler, D., y A. McDowell; *Introducing the FMM procedure for finite mixture models*, Proceedings of the SAS Global Forum 2012 (paper 328-2012). Orlando, Florida: SAS Institute Inc. (2012).

Kim, D., J. M. Kim, S. M. Liao y Y. S. Jung; *Mixture of D-vine copulas for modeling dependence,* Computational Statistics and Data Analysis, 64, 1-19 (2013).

Kuhl, M., S. G. Sumant y J. R. Wilson; *An automated multiresolution procedure for modeling complex arrival processes*, INFORMS Journal on Computing, 18(3), 277-280 (2006).

Law, A. M.; *Simulation Modeling & Analysis* (5a Ed.), McGraw-Hill (USA) (2014).

Leemis, L. M.; *Nonparametric estimation of the intensity function for a nonhomogeneous Poisson process*, Management Science, 37(7), 886–900 (1991).

Lindley, D. V.; *The theory of queues with a single server*, Proceedings of the Cambridge Philosophical Society, 48(2), 277-280 (1952).

Martins, J. L. F., M. L. R. Ferreira, J. M. Pardal, y C. A. R. Morano; *Comparación de la estimación de la productividad del proceso de soldadura eléctrica por los métodos de simulación de Monte Carlo e Hipercubo Latino*, Información. Tecnológica, 23(4), 21-32 (2012).

Moreira, C. y I. Van Keilegom; *Bandwidth selection for kernel density estimation with doubly truncated data*, Computational Statistics and Data Analysis, 61, 107-123 (2013).

Muñoz, D. F., O. Romero-Hernández, J. E. Detta Silveira y D. G. Muñoz; *Forecasting demand for educational material for adult learners in Mexico*, International Transactions in Operational Research, 17(1), 71-84 (2010).

Muñoz, D. F. y D. F. Muñoz; *Planeación y control de proyectos con diferentes tipos de precedencias utilizando simulación estocástica*, Información Tecnológica, 21(4), 25-33 (2010).

Muñoz, D. F. y D. F. Muñoz; *Bayesian forecasting of spare parts using simulation*, en N. Altay, & L. A. Litteral (eds.), Service Parts Management: Demand Forecasting and Inventory Control (pp. 105-124). Springer (USA) (2011).

Muñoz, D. F., D. G. Muñoz y A. Ramírez-López; *On the incorporation of parameter uncertainty for inventory management using simulation*, International Transactions in Operational Research, 20(4), 493-513 (2013).

Nagatsuka, H., T. Kamakura y N. Balakrishnan; *A consistent method of estimation for the three-parameter Weibull distribution*, Computational Statistics and Data Analysis, 58(2), 210-226 (2013).

Papp, D. y F. Alizadeh; *Shape-Constrained Estimation Using Nonnegative Splines, Journal of Computational and Graphical Statistics*, 23(1), 211-231 (2014).

Qiao, H., y C. P. Tsokos; *Parameter estimation of the Weibull probability distribution*, Mathematics and Computers in Simulation, 37(1), 47-55 (1994).

Scott, D. W.; *Multivariate density estimation and visualization*, en J. E. Gentle, W. A. Härdle, & Y. Mori (eds.), Handbook of Computational Statistics (pp.549-569). Springer (USA) (2012).

Vanalle, R. M.; W. C. Lucato, M. Vieira Junior y I. D. Sato; *Uso de la simulación Monte Carlo para la toma de decisiones en una línea de montaje de una fábrica*, Información. Tecnológica, 23(4), 33-44 (2012).

Vargas, J. M. y J. A. Giraldo; *Modelo de predicción de costos en servicios de salud soportado en simulación discreta*, Información Tecnológica, 25(4), 175-184 (2014).

Wagner, M. A. F. y J. R. Wilson; *Using univariate Bézier distributions to model simulation input processes*, IIE Transactions, 28(9), 699-711 (1996).

Wang, B. y W. Wetelecki; *Density estimation for data with rounding errors*, Computational Statistics and Data Analysis, 65, 4-13 (2013).