MECOM 99
Mendoza
6-10 de septiembre de 1999

TÉCNICAS ESTADÍSTICAS AVANZADAS EN EL ANÁLISIS DE GRANDES MODELOS COMPUTACIONALES

Jorge E. Núñez Mc Leod y Jorge H. Barón
Instituto CEDIAC, Facultad de Ingeniería
Universidad Nacional de Cuyo, Casilla de Correo 405, 5500 Mendoza, Argentina
cediac@cediac.uncu.edu.ar

RESUMEN

En el presente trabajo se explica una metodología que conjuga una serie de herramientas estadísticas avanzadas para realizar análisis de incertidumbres, sensibilidad e importancias sobre grandes modelos computacionales. El conjunto de herramientas estadísticas aplicadas son independientes del modelo y este es en sí tratado como una caja negra. Este enfoque permite aplicar estas técnicas a un amplio conjunto de modelos de diversa índole. Así por ejemplo se pueden mencionar: modelos en base a elementos finitos, elementos de contorno, modelos de estructura lógica (tales como expresiones booleanas), etc.

ABSTRACT

In the present work a methodology to perform uncertainty, sensitivity and importance analysis of large numerical models is presented. This methodology combines several advanced statistical tools. The proposed tools applied are independent of the model under analysis and the model itself is treated as a black box. This flexibility allows for the analysis of a variety of numerical models. As an example it can be mentioned the application to finite element models, boundary element models, logical models (boolean equations), etc.

INTRODUCCIÓN

Los estudios de incertidumbre, sensibilidad e importancias sobre las variables de los modelos computacionales y su/s respuesta/s se han vuelto una práctica relevante en los últimos años. Estos estudios permiten realizar un análisis del comportamiento del modelo frente a la variación de las variables de entrada. Esto permite caracterizar la incertidumbre de la respuesta, hacer un estudio de la sensibilidad del modelo a la variación de las variables de entrada y realizar una clasificación de la importancia de éstas en base a la sensibilidad. Este mejor conocimiento de la influencia de cada variable en el comportamiento del modelo permite realizar estudios más detallados sobre un subconjunto de éstas (las de mayor importancia).

Diversas tareas componen un análisis estadístico de un modelo computacional: selección de las variables bajo estudio, asignación de las distribuciones características de cada variable, generación de las muestras para el modelo, ejecución del programa (una vez por cada muestra), análisis de las relaciones de entrada/salida, clasificación de variables y determinación de la incertidumbre de los resultados del modelo. Otras tareas pueden adicionarse; pero centraremos nuestra atención en las enumeradas.

El núcleo de todo el problema se basa en el tiempo que demanda la corrida del modelo computacional y sobre este punto se han hecho los mayores esfuerzos. Desde dos puntos de vista. El primero tratando de minimizar la cantidad de ejecuciones del modelo necesarias para obtener en forma adecuada la función de distribución de los resultados (centrando la atención en los métodos de muestreo), una propuesta alternativa a la presentada en este trabajo se pude ver en O'Hagan, Kennedy and Oakley[1]. Y por otro lado reemplazando el modelo original por uno con el

mismo comportamiento estadístico; pero cuyo tiempo de ejecución no sea relevante (técnicas de la superficie de respuesta y campos aleatorios), ver Currin, Mitchell, Morris and Ylvisaker[2].

En el presente trabajo se adopta el primer punto de vista.

TAREAS DEL ESTUDIO

1. SELECCIÓN DE LAS VARIABLES A ESTUDIAR

La selección de las variables a estudiar no es una tarea trivial y depende fuertemente de la motivación del análisis y del tipo de modelo. Por esta razón y sólo a modo de ejemplo enunciaremos el caso de modelos en algebra de boole donde se pueden determinar las variables más representativas a través de medidas tales como la de Fussell-Vesely (caso de árboles de falla).

2. ASIGNACIÓN DE DISTRIBUCIONES

La asignación de una distribución a una variable puede normalmente provenir de las siguientes fuentes: datos de campo, base de datos y opinión de expertos.

Datos de campo: si se dispone de éstos, previamente seleccionados y clasificados, se puede proceder a ajustarlo a una distribución y probar el ajuste mediante un test de la chi-cuadrado. Esta es la mejor de las opciones, cuando la cantidad de datos de campo es significativo, caso contrario es contraproducente su uso.

Base de datos: Al no disponer de datos de planta o siendo éstos escasos o de baja significación, se debe recurrir al uso de bases de datos, con la información que se requiere, similar o genérica. En este punto es importante aclarar que, por ejemplo, para el caso del licenciamiento de instalaciones nucleares, estas bases de datos deben encontrarse reconocidas y por ello datos suministrados por los fabricantes no son normalmente aplicables.

Al usar bases de datos, para el caso de componentes industriales, se parte de la suposición, entre otras, que las condiciones de fabricación, operación y mantenimiento son similares con lo que se añade un cierto grado de incertidumbre al análisis. Dependerá del analista utilizar en todo el estudio la misma calidad de datos, de tal manera de no enmascarar los resultados con datos de componentes que no se ajusten a la realidad. La penalización de estos datos queda a criterio del analista.

Opinión de Experto: Este última posibilidad es la menos recomendable y si debe ser llevada a cabo en general un factor de error de 10 debe tenerse en cuenta. Una comparación de técnicas puede consultarse en Cojazzi et al.[3]. Una guía para el uso de Juicio de Expertos en análisis de incertidumbre se puede ver en Goossens and Cooke[4]

3. GENERACIÓN DE MUESTRAS

Tradicional es la técnica de Muestreo Aleatorio, una versión mejorada es el Muestreo Estratificado y más recientemente, el Muestreo por Hipercubo Latino (LHS, Latin Hypercube Sampling), McKay, Beckman and Conover[5] y el Muestreo por Hipercubo Latino Escalable (LHS-S), Barón y Núñez McLeod[6]. La primera de estas permite la generación de muestras para una variable mediante un método de muestreo aleatorio simple. En el segundo método todas las áreas del espacio muestral de las variables se encuentran representadas. El tercer método basado en un esquema de muestreo restringido, proporciona mejores resultados que los anteriores. Finalmente el último método brinda la posibilidad de ir ajustando el tamaño de la muestra si han resultado insuficientes los resultados obtenidos para la correcta caracterización de las variables (principalmente para el análisis de importancias).

A continuación veremos en más detalle estos métodos.

3.1. MUESTREO ALEATORIO

El muestreo de Montecarlo consiste en generar una muestra de una variable aleatoria, eligiendo números al azar sobre la distribución de probabilidad acumulada de dicha variable, para luego obtener los valores correspondientes de dicha variable, que constituyen la muestra en sí. Esa técnica requiere un elevado número de muestras para conseguir una adecuada representatividad de la función de distribución de la variable (típicamente miles o decenas de miles).

Una vez generada la muestra, el modelo numérico es ejecutado una vez para cada valor muestral, y la variable resultado es luego caracterizada en sus propiedades estadísticas basándose en el principio de que *la transformación de una variable aleatoria es una variable aleatoria*.

Este proceso debe repetirse para cada variable de entrada de interés, haciendo que el estudio multivariable sea extremadamente costoso para el estudio de modelos complejos.

3.2. MUESTREO ESTRATIFICADO

Para reducir el tamaño de muestra necesario se desarrollaron métodos de muestreo de Montecarlo Estratificado, que permiten obtener una buena representatividad generando muestras por sectores (o estratos) de cada variable.

3.3. MUESTREO POR HIPERCUBO LATINO

El método LHS consiste en la selección de los parámetros y variables a muestrear, la asignación de distribuciones de probabilidad a cada una (que pueden estar basados en estudios teóricos o mediciones experimentales), la división de cada distribución en un número fijado *apriori* de intervalos equiprobables, la generación de una muestra aleatoria dentro de cada intervalo y para cada variable, y el apareamiento aleatorio de muestras entre variables, de modo de obtener vectores de valores de entrada, uno por cada intervalo.

Con cada vector de valores de entrada, el modelo numérico es ejecutado una vez. Es decir, el método requiere correr el modelo tantas veces como intervalos se hayan supuesto en la división de las distribuciones de probabilidad, independientemente del número de variables muestreadas. Normalmente esta técnica permite reducir en uno o más órdenes de magnitud la cantidad de corridas necesarias para obtener una determinada representatividad, en comparación con un Montecarlo clásico.

Sin embargo, un problema que aparece, indicado por McKay, Beckman and Conover[5], es que la representatividad de los resultados solamente se puede evaluar luego de efectuar las corridas del modelo numérico, y en caso de que no sea satisfactoria, se deberían repetir todos los pasos, con un número mayor de muestras (intervalos), no pudiendo utilizar los resultados anteriores. Esta limitación es muy seria cuando se trata de correr un modelo complejo (que implica, por ejemplo, horas de CPU) y es por ello que la obtención de un método que permita escalar muestras resulta de utilidad.

3.4. MUESTREO POR HIPERCUBO LATINO ESCALABLE

El procedimiento desarrollado consiste en el uso de las funciones de distribución de probabilidad de las variables a muestrear. Sobre estas funciones de distribución se estratifica el eje de las ordenadas (para obtener estratos equiprobables), en sectores adyacentes disjuntos, y se realiza un muestreo aleatorio generando un valor por cada intervalo, obteniendo a continuación los valores en abscisa de cada variable. A continuación se aparean las muestras de cada variable de manera aleatoria, obteniendo y verificando un bajo nivel de correlación entre las mismas. Para este ciclo se han utilizado dos generaciones independientes de números aleatorios, uno para el muestreo y otro para el apareo. Hasta aquí la técnica es la de LHS.

A continuación se ejecuta el modelo numérico, una vez con cada vector de muestras (igual al número de intervalos) y se realiza la caracterización estadística de los resultados deseados. En caso que dicha caracterización no sea satisfactoria, es decir, que sus propiedades estadísticas (media, varianza, percentiles, etc.) no estén adecuadamente representados, se inicia un nuevo ciclo aumentando el tamaño de la muestra. Este es el inicio del proceso de escalado.

En el nuevo ciclo se retoman los sectores muestreados previamente, y se subdividen en sectores adyacentes disjuntos equiprobables (por ejemplo, dividiendo por dos o por tres), se reconoce en cuales de los nuevos sectores no existen muestras, y se generan nuevos valores, de manera aleatoria, en dichos sectores. Ver figura 1.

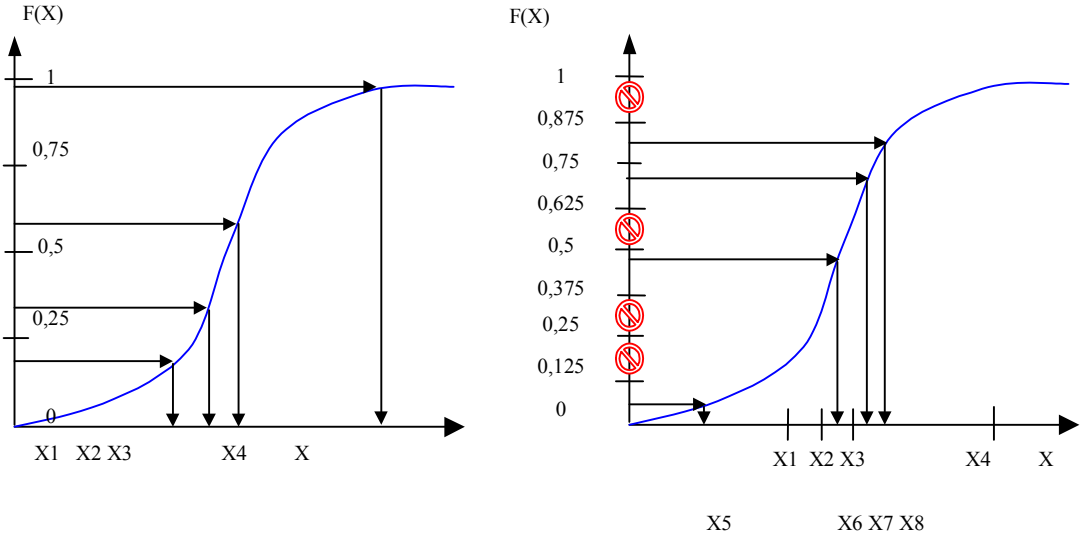

Figura 1. Esquema de muestreo y escalado

A continuación se aparean aleatoriamente los nuevos valores obtenidos, evitando las combinaciones con valores ya pertenecientes al tamaño de muestra anterior, y se obtiene un nuevo número de vectores de variables de entrada, con los que se corre el modelo numérico (Fig. 2).

Los vectores de muestra generados inicialmente y los generados en el nuevo ciclo forman el conjunto de vectores de entrada, y los resultados obtenidos inicialmente junto con los obtenidos en el nuevo ciclo forman el conjunto de resultados. Dicho conjunto de resultados es nuevamente analizado desde el punto de vista estadístico, y de ser necesario, se repite un nuevo ciclo aumentando el número de intervalos, hasta obtener la representatividad deseada en las características estadísticas de los resultados.

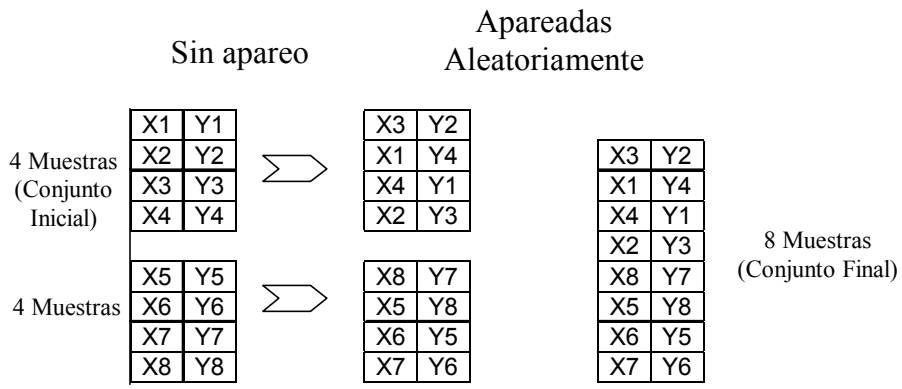

Figura 2. Apareo de las muestras

De este modo se evita tener que rehacer un número elevado de corridas del modelo, usando dentro del conjunto de resultados, aquellos obtenidos previamente.

3.5. COMPARACIÓN

Unicamente con fines ilustrativos supondremos la aplicación de los muestreos aleatorio, LHS y LHS-S a un modelo cuyo tiempo de ejecución es de 10 horas. Se analizarían 6 variables y se procederá a generar primero 20 muestras, luego 40 y finalmente 80 (Tabla 1). Esto bajo la suposición que con los dos primeros conjuntos de muestras, no se

logró una adecuada representación de las características estadísticas de los resultados o de la relación entrada/salida para el análisis de sensibilidad e importancias.

Muestras	Muestreo Aleatorio	LHS	LHS-S
20	20	20	20
40	40	40	20
80	80	80	40

Tabla 1. Cantidad de muestras generadas

Muestras	Muestreo Aleatorio	LHS	LHS-S
20	1200	200	200
40	2400	400	200
80	4800	800	400
Total [horas]	8400	1400	800
Total [días]	350	58,3	33,3

Tabla 2. Cantidad de tiempo en horas demandado por cada método y conjunto de muestras

En la tabla 2 se expresa en horas el tiempo de ejecución del modelo para cada conjunto de muestras. En esta tabla también se ha totalizado el tiempo total que se ha requerido para alcanzar resultados adecuados. El muestreo aleatorio requiere casi un año de trabajo contínuo, el LHS requiere aproximadamente 2 meses y el LHS-S alrededor de 1 mes.

Esta visión simplificada demuestra la importancia del método de generación en todo el trabajo con grandes modelos computacionales.

4. ANÁLISIS DE LAS RELACIONES DE ENTRADA/SALIDA

Una vez que se han completado las corridas del código, es necesario cuantificar la sensibilidad de la/s salida/s a cada una de las entradas. Para esto se pueden usar los Coeficientes de Regresión Estandarizados y los Coeficientes de Correlación Parcial, ver Iman, Shortencarier and Johnson[7]. Por simplicidad se harán las explicaciones sobre modelos lineales.

4.1. COEFICIENTES DE REGRESIÓN ESTANDARIZADOS

El análisis de sensibilidad en conjunción con los tipos de muestreo vistos está estrechamente relacionado con la construcción de modelos de regresión cuales aproximan el comportamiento calculado por el modelo. Donde los coeficientes de regresión, b_j, y la ordenada al origen, b_0, se calculan por los métodos tradicionales de mínimos cuadrados, obteniéndose:

$$\hat{Y} = b_0 + \sum_j b_j X_j \tag{1}$$

Estos coeficientes de regresión son las derivadas parciales del modelo de regresión con respecto a las variables de entrada. Sin embargo estos coeficientes son fácilmente influenciados por las unidades en las cuales se miden las variables. Por esto no proveen una medida confiable de la importancia relativa de cada variable.

El problema anterior puede ser eliminado estandarizando todas las variables usadas en el modelo de regresión.

$$X^* = (X - \bar{X}) / s_x \tag{2}$$

Donde $\bar{X}$ es la media muestral y s_x la desviación estándar.

$$Y^* = \sum_j b_j^* X_j^* \tag{3}$$

Los coeficientes de este nuevo modelo se denominan Coeficientes de Regresión Estandarizados, b_j^*. Estos pueden ser usados para medir directamente la importancia relativa de cada variable de entrada. Naturalmente la confiabilidad de estos resultados está condicionada a que la relación entre las variables de entrada y salida sea adecuadamente descrita por el modelo de regresión.

4.2. COEFICIENTES DE CORRELACIÓN PARCIAL

Un coeficiente de correlación de la muestra provee una medida de la relación lineal entre la variable dependiente, Y y las variables independientes X_j. Si este coeficiente de correlación lo indicamos por r_{yj}, entonces el $\max[r_{yj}]$ puede ser usado para identificar la variable que tiene la relación mas fuertemente lineal con la variable dependiente.

$$r_{yj} = \frac{\sum_i X_{ij} Y_i - \frac{\sum_i Xij \sum_i Y_i}{n}}{\sqrt{\left[\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}\right]\left[\sum_i Y_i^2 - \frac{(\sum_i Y_i)^2}{n}\right]}} \tag{4}$$

En otras palabras el coeficiente de correlación parcial da una medida de la capacidad de cada variable de ajustar el comportamiento de la variable dependiente, teniendo en cuenta la relación entre las variables de entrada.

Puede demostrarse que si las variables de entrada son independientes o casi independientes los coeficientes de correlación parcial serán aproximadamente iguales a los coeficientes de correlación de Pearson.

5. CLASIFICACIÓN DE VARIABLES

La clasificación de variables para el análisis de importancias puede surgir de los valores de los coeficientes de regresión estandarizados y de los coeficientes de correlación parcial. Esta clasificación puede en muchos casos ayudar a decidir si el tamaño de muestra utilizado es el adecuado o se requiere incrementarlo.

6. DETERMINACIÓN DE LA INCERTIDUMBRE DE LOS RESULTADOS

Con los resultados obtenidos una serie de valores estadísticos deben ser calculados. Un conjunto de estos útil es: media muestral, varianza muestral, mínimo, primer cuartil, mediana, tercer cuartil, máximo, coeficiente de asimetría y coeficiente de apuntamiento.

CONCLUSIONES

Se han mostrado diversas herramientas estadísticas para el análisis de grandes modelos computacionales, siguiendo las etapas normales de un estudio de incertidumbres, sensibilidad e importancias. Se ha hecho un fuerte énfasis en la etapa de generación de muestras, ya que de ella depende en gran medida una reducción de los tiempos de trabajo con grandes modelos.

REFERENCIAS

[1] O'Hagan, A., Kennedy, M. and Oakley, J. Uncertainty analysis and other inference tools for complex computer codes. Bayesian Statistics, 6, 1-19. 1998

[2] Currin, C., Mitchell, T., Morris, M. and Ylvisaker, D. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. Journal of the American Statistical Association, 86, 953-963. 1991

[3] Cojazzi, G., Pulkkinen, U. (JRC-ISIS), De Gelder, P., Gryffroy, D. (AVN), Bolado, R. (FGUPM), Hofer, E. (GRS), Virolainen, R. (STUK), Coe, I. (NNC), Bassanelli, A. (ENEL), Puga, J. (UNESA), Papazoglou, I. (NCSR) and Zuchuat, O. (HSK). Benchmark exercise on expert judgment techniques in PSA level 2. FISA-97 Symposium, 450-459. 1997

[4] Goossens, L. and Cooke, R. Procedures guide for the use of expert judgement in uncertainty analyses. Probabilistic Safety Assessment and Management. Pergamon, 978-985. 1996.

[5] McKay, M., Beckman, R. and Conover, W. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21, 239-245. 1979.

[6] Barón, J. y Núñez Mc Leod, J. Generación Escalable de Muestras. Simulación con Métodos Numéricos: Nuevas Tendencias y Aplicaciones. SVMNI, VA1-VA8. 1998.

[7] Iman, R., Shortencarier, M. and Johnson, J. A FORTRAN 77 Program and User´s Guide for the calculation of partial correlation and standardized regression coefficients. NUREG/CR-4122, SAND85-0044. 1985.