3. ANÁLISIS Y MODELOS ESTADÍSTICOS

Actualmente se reconoce la importancia de la estadística aplicada en el desarrollo de investigaciones en muy diversos campos; cada vez son más los profesionales de diferentes disciplinas que requieren de métodos estadísticos como muestreo, simulación, diseño de experimentos, modelamiento estadístico e inferencia, para llevar a cabo recolección, compendio y análisis de datos y para su posterior interpretación. En sismología, los métodos estadísticos son de amplio uso; por ejemplo en la estimación del riesgo sísmico, predicción sísmica, localización de sismos, determinación de magnitudes y cuantificación de incertidumbres. Los principales modelos estocásticos usados para describir los procesos relacionados con los sismos se basan en series de tiempo y procesos puntuales. Los modelos de series de tiempo se usan generalmente para describir procesos que son muestreados en puntos de tiempo discretos, mientras que procesos puntuales se usan para modelar fenómenos que se presentan de manera irregular, sin un patrón temporal, y que pueden ocurrir en cualquier momento o espacio.

Por otro lado, y análogo al proceso de experimentación llevado a cabo en laboratorios con el objetivo de aumentar la comprensión de alguna teoría para su validación y empleo posterior, la simulación, considerada como un método de experimentación controlada, es el proceso de imitación de aspectos importantes del comportamiento de un sistema, mediante la construcción de un modelo implementado en un computador de tal forma que permita generar observaciones dadas ciertas entradas. Con el análisis estadístico de tales observaciones se estiman medidas del comportamiento del sistema de interés. Sin embargo, de esta manera no es posible encontrar resultados óptimos, sino mas bien, resultados satisfactorios a problemas de difícil, costosa o imposible resolución mediante otros métodos.

3.1. Inferencia estadística

3.1.1. Sistemas y modelos

3.1.1.1. *Sistemas*

Un sistema es una fuente de datos del comportamiento de alguna parte del mundo real. Está formado por elementos que interactuan para lograr un objetivo, los cuales poseen características o atributos, parámetros y variables, que toman valores numéricos o lógicos. Un sistema puede ser natural o artificial, dinámico o estático, estable o inestable, adaptativo o no adaptativo, lineal o no lineal; puede tener variables independientes o dependientes, no controlables o controlables, continuas, discretas o mixtas, no observables u observables. Las reglas que especifican la interacción entre los elementos de un sistema, determinan la forma

en que las variables descriptivas cambian con el tiempo. Las variables que describen las entidades, los atributos y las actividades de un sistema en un instante particular de tiempo, que permiten predecir su comportamiento futuro, se denominan variables de estado y sus valores proporcionan el estado del sistema en ese instante, además relacionan el futuro del sistema con el pasado a través del presente. Si el comportamiento de los elementos del sistema puede predecirse con seguridad, el sistema es determinístico, de lo contrario es estocástico. Si la probabilidad de encontrarse en alguno de los estados no cambia con el tiempo el sistema es estático, de lo contrario es un sistema dinámico. Si el estado de un sistema cambia sólo en ciertos instantes de tiempo se trata de un suceso discreto, de lo contrario de un suceso continuo.

Sea un sistema físico como por ejemplo la Tierra; o una estructura definida, como por ejemplo la corteza; o un estado de ella, por ejemplo la liberación de energía y reacomodamiento de esfuerzos producidos en un punto determinado (sismo). El interés puede ser la composición química en el interior de la Tierra; o un modelo de velocidades asociado a la corteza; o la identificación del punto en el interior de la Tierra a partir del cual se produjo la liberación de energía, es decir la identificación del foco sísmico, planteado como el problema de localización hipocentral. Para el último caso, el cual es el tema de interés de este proyecto, las observaciones de las cuales se dispone son mediciones indirectas logradas a través de sismogramas obtenidos en la superficie de la Tierra, que permitirán analizar lo sucedido en algún punto en el interior de la Tierra.

3.1.1.2. *Modelos*

Un modelo es una representación formal de un sistema real, con el que se pretende aumentar su comprensión, hacer predicciones y ayudar a su control. Los modelos pueden ser físicos (descritos por variables medibles), análogos (diagrama de flujo) y simbólicos (matemáticos, lingüísticos, esquemáticos). Los modelos matemáticos o cuantitativos son descritos por un conjunto de símbolos y relaciones lógico–matemáticas.

Para la construcción de un buen modelo es necesario contar con leyes (por ejemplo, físicas) que describan el comportamiento del sistema. También es importante la experiencia, la intuición, la imaginación, la simplicidad y la habilidad para seleccionar el subconjunto mas pequeño de variables. El primer paso es establecer el problema en forma clara y lógica delimitando sus fronteras; luego viene la recogida y depuración de datos; el diseño del experimento; las pruebas de contrastes; la verificación del modelo y la validación de las hipótesis. Por ejemplo, un análisis de sensibilidad determinará el grado de influencia en la solución del modelo debida a variaciones en los parámetros (robustez de un modelo). Un modelo debe ser una buena aproximación al sistema real, debe incorporar los aspectos importantes del sistema y debe resultar fácil de comprender y manejar. Un factor muy importante es que haya una alta correlación entre lo que predice el modelo y lo que actualmente ocurre en el sistema real.

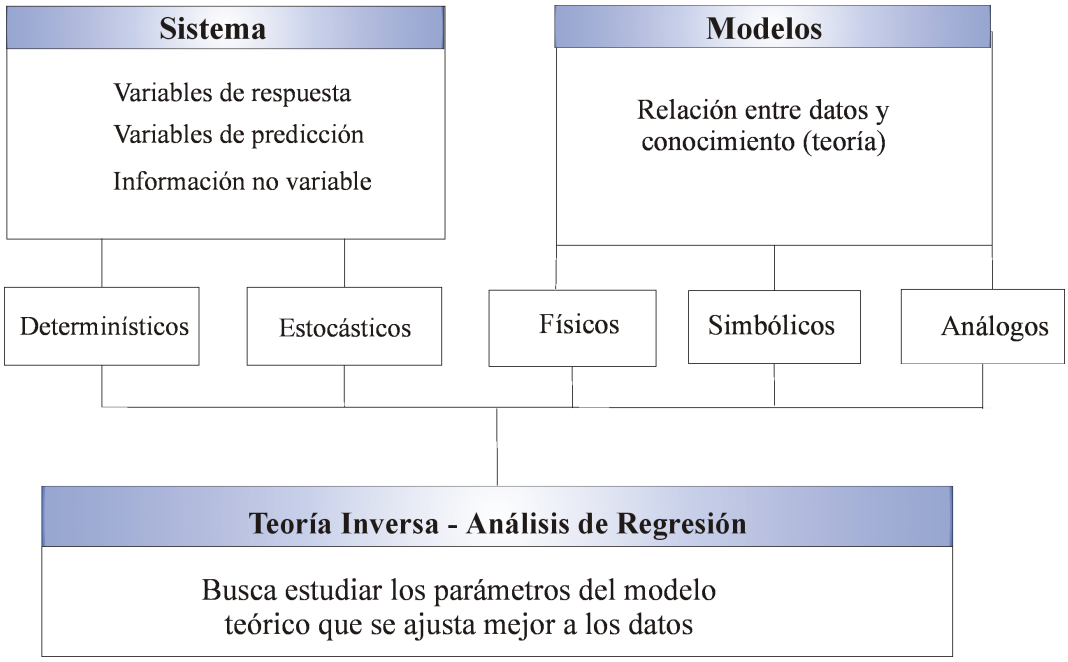

Figura 3.1: Relación entre sistemas y modelos

3.1.1.3. *Relación entre sistemas y modelos*

Un sistema puede ser descrito mediante una función que relaciona un conjunto de datos u observaciones (d, variables de respuesta) con un grupo de parámetros $P(m_1, m_2..)$; cada grupo de valores específicos de este grupo de parámetros proporciona un modelo (m) diferente. Si se dispone de un modelo físico (G) que obtenido a partir de la teoría relaciona los datos observados con los parámetros conocidos –variables–, se tiene entonces una relación funcional $G \Rightarrow F(d, m)$ que describe el fenómeno; si esta relación es lineal, se define entonces como $Gm = d$. Aquí se pueden tener dos situaciones diferentes: se conocen los parámetros del modelo pero es necesario conocer la respuesta de ese sistema, esta situación es conocida como el problema directo; o de lo contrario, se dispone de observaciones de las variables de predicción y de respuesta, pero se desconocen los parámetros del modelo que expliquen mejor la relación entre esas variables, aquí se habla del problema inverso, el cual se resuelve mediante regresión estadística. Resolver el problema inverso consiste en estimar los valores del modelo (m) que expliquen mejor las observaciones. (Menke, 1984; Tarantola and Valette, 1982b).

En el problema específico de estimar las coordenadas hipocentrales de un sismo, d son los tiempos de arribo de ondas sísmicas registradas en las diferentes estaciones sismológicas; G está conformado por una estimación inicial de las coordenadas hipocentrales, las coordenadas que definen las posiciones de las diferentes estaciones y un estructura de velocidades entre la superficie y el hipocentro; m son los parámetros del modelo planteado a partir de la

teoría física, los cuales al ser usados para calcular los tiempos de recorrido teóricos de las ondas desde diferentes estaciones producirían la menor diferencia entre éstos y los tiempos observados en el sismograma.

Tarantola and Valette (1982b) proponen que antes de formular la solución a un problema inverso es necesario que sea;

- válida tanto para problemas lineales como para problemas no lineales, tanto para problemas bien determinados (suficientes datos para la estimación, matrices invertibles) como para problemas mal determinados (información insuficiente o inconsistente)
- consistente con respecto a un cambio de variables –el cual no es el caso con aproximaciones ordinarias–
- suficientemente general para permitir diferentes distribuciones para el error en los datos (gaussiana, no gaussiana, simétrica, asimétrica, etc.), para permitir la incorporación formal de cada supuesto y para incorporar errores teóricos en una forma natural.

Afirman también, que estas restricciones pueden cumplirse si se formula el problema usando teoría de probabilidades y toda la información disponible (Inferencia bayesiana), estudiando sistemas que puedan ser descritos con un finito grupo de parámetros donde las características cuantitativas del sistema sean definidas como funciones de probabilidad (para datos y parámetros) mas que como parámetros discretos (p.e. medias). Sin embargo, existen diversos métodos estadísticos de estimación que, aunque no cumplen todas las restricciones anteriores, de igual manera y para problemas específicos proporcionan estimadores con propiedades deseables. Además debido a la complejidad de algunos problemas y al exhaustivo requerimiento computacional de los métodos bayesianos, estimadores de esta clase son poco comunes en la práctica.

3.1.2. Estimaciones y estimadores

La estimación involucra el uso de datos muestrales en conjunción con alguna técnica estadística, y se puede llevar a cabo mediante estimación puntual o por intervalo: estimación puntual es la asignación de un valor al parámetro desconocido, estimación por intervalo es una fórmula que dice cómo utilizar los datos de una muestra para calcular un intervalo en el cual con cierto nivel de confianza se encuentra el valor del parámetro. La técnica para estimar los parámetros que definan un modelo teórico que no está disponible, pretendiendo una asociación entre variables de respuesta y variables de predicción no causal, debe proporcionar estimadores con cierta propiedades.

3.1.2.1. *Estimadores*

Sea δ un grupo de características del sistema que se desean conocer a partir de la observación de la variables $x_1, \ldots, x_n$, con función de densidad de probabilidad $f(x;\delta)$ y $y_1, \ldots, y_n$ con función de densidad de probabilidad $f(y;\delta)$ observadas en una muestra aleatoria de tamaño n; la estimación de δ a partir de esta muestra se denotará como $\hat{\delta}$.

Las propiedades mas deseables de un estimador son que la distribución de muestreo esté concentrada alrededor del valor del parámetro y que la varianza del estimador sea lo menor posible. El error cuadrático medio resume estas propiedades y es definido como

$$ECM(\delta) = Var(\hat{\delta}) + [\delta - E(\hat{\delta})]^2 \tag{3.1}$$

La calidad de las estimaciones puede medirse en función de exactitud, precisión y consistencia. La exactitud es el grado en que un valor promedio coincide con el valor verdadero de la cantidad de interés; una estimación es exacta cuando no tiene desviaciones –positivas o negativas– del valor verdadero, es decir se han reducido los errores sistemáticos. La precisión está relacionada con la variabilidad de los datos, mientras exista mayor variabilidad, habrá menor precisión; buena precisión significa que se han eliminado los errores aleatorios en el procedimiento de medición. Un estimador es consistente si converge probabilísticamente al valor del parámetro que está estimando, es decir, la estimación se aproxima mas al valor del parámetro cuando el número de observaciones tiende a infinito (ley de los grandes números). Además se habla de estimadores insesgados cuando el valor promedio toma valores muy cercanos al valor real, es decir el sesgo que puede evaluarse mediante el error cuadrático medio es cero. El estimador más eficiente es aquel estimador insesgado con varianza mínima y con la propiedad de que no existe otro estimador insesgado con menor varianza. Así el ECM que es la suma de la varianza y el sesgo describe las propiedades deseables de una estimación.

Los siguientes métodos de estimación puntual (a excepción del (3) y (6)) han sido utilizados, entre otros, para resolver el problema de estimación de los parámetros hipocentrales. Una descripción mas completa de métodos de estimación se encuentra en Draper and Smith (1981).

3.1.2.2. *Métodos de estimación puntual*

1. Mínimos cuadrados, obtiene la estimación del parámetro que minimiza el error cuadrático medio (Norma L2). El objetivo es minimizar $\sum(d_i - \hat{d}_i)^2$, donde $\hat{d}_i$ son los valores estimados a partir del modelo y d_i son los datos observados, los cuales se asume que tienen asociado un error de medida que es independiente y distribuido normalmente con media μ y varianza σ^2. La varianza del error es constante entre observaciones, de lo contrario es necesario usar mínimos cuadrados con factores de peso. Se supone

que la variabilidad en los datos que no pueda explicarse mediante la ecuación de regresión se debe al error aleatorio, por tanto si la selección de la ecuación es correcta, esta última debe ser mínima. El sistema lineal a resolver es de la forma $Gm + e = d$; para lo cual es necesario plantear el sistema de ecuaciones normales que toman la forma,

$$(G^T G)m = G^T d \tag{3.2}$$

si $G^T G$ tiene inversa, la solución para m es

$$m = (G^T G)^{-1} G^T d \tag{3.3}$$

Existen diversos métodos para resolver el anterior sistema de ecuaciones, entre ellos métodos de descomposición de la matriz G en una suma de vectores, basándose en el hecho que un objeto de gran dimensión puede ser representado como la suma de productos de bajas dimensiones, lo cual hace mas fácil su análisis. Algunos métodos de descomposición son el método QR que consiste en la descomposición de la matriz G en una matriz triangular superior R y una matriz ortogonal Q, ($Q^T Q = 1$), se resuelve el sistema $RG = Q^T m$, este método involucra reducir la matriz simétrica a una matriz tridiagonal haciendo $n-2$ transformaciones ortogonales. Otros métodos son descomposición $Cholesky$ -descompone $G = R^T R$ donde R es una matriz triangular superior, el sistema a resolver es $R^T Rm = d$-, descomposición LU - descompone $G = LU$, donde L es una matriz triangular superior y U es una matriz triangular inferior- y descomposición del valor singular SVD. Algunos de estos métodos requieren mas operaciones que otros, y otros son especialmente útiles cuando el sistema está mal condicionado, o en caso de matrices simétricas, y se encuentran explicados en Press et˜al. (1997). Además, los parámetros pueden ser estimados utilizando diversas técnicas de optimización (sección 3.3). En la sección 4.1.3.1 (pág. 48) se retoma este método, como una aproximación lineal iterativa, para resolver el problema no lineal de determinación de parámetros hipocentrales y tiempo de origen de un sismo. Otro punto de vista sobre el uso del método de mínimos cuadrados para resolver problemas no lineales es expuesto por Tarantola and Valette (1982a).

2. Máxima verosimilitud, selecciona como estimador al valor del parámetro que tiene la propiedad de maximizar el valor de la probabilidad de la muestra aleatoria observada, es decir encuentra el valor de los parámetros que maximizan la función de verosimilitud. La verosimilitud es la función de densidad de probabilidad conjunta de las variables independientes. Una formulación teórica de este método para estimación de parámetros hipocentrales es dada en Rodi and Toksoz (2001).

3. Método de Momentos, consiste en igualar los momentos apropiados de la distribución de la población con los correspondientes momentos muestrales, para estimar un parámetro desconocido de la población. El r–ésimo momento se define como

$$\delta_r = \frac{1}{n} \sum_{i=1}^{n} x^r \tag{3.4}$$

así, los cuatro primeros momentos de una variable con función de densidad de probabilidad normal son la media, la varianza, la curtosis y el sesgo, en ese orden.

4. Estimación bayesiana, es la conjunción de tres estados de información: información a priori, una función de máxima verosimilitud e información a posteriori. En la sección 4.1.3.2 (pág. 51) se describe el uso de este método aplicado a la determinación de los parámetros hipocentrales y tiempo de origen. Una visión mas general en solución a problemas inversos está dada en Scales and Tenorio (2001).

5. Estimadores Robustos, funcionan muy bien para una amplia gama de distribuciones de probabilidad. Obtienen la estimación del parámetro que minimiza el error absoluto (Norma L1), es apropiado cuando la dispersion es grande o cuando hay presentes valores extremos.

6. Estimador *Jackknife* o de punto exacto, especialmente útil en la estimación de medias y varianzas, se basa en remover datos y volver a calcular la estimación. Se obtiene una estimación $\hat{\delta}_i$ omitiendo la i–ésima observación y calculando la estimación con las $n-1$ observaciones restantes. Este cálculo se realiza para cada observación del conjunto de datos, por lo tanto se producen n estimaciones $\hat{\delta}_1, \ldots, \hat{\delta}_n$; de la muestra completa se obtiene δ. Un pseudo–valor o estadístico de influencia $J_i(\hat{\delta})$ se determina de

$$J_i(\hat{\delta}) = n\hat{\delta} - (n-1)\hat{\delta}_i \tag{3.5}$$

El estimador *Jackknife* $J(\hat{\delta})$ es alguna combinación lineal de todas las estimaciones, por ejemplo el promedio de los pseudo–valores,

$$J(\hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} J_i(\hat{\delta}) \tag{3.6}$$

con varianza,

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n}(J_i(\hat{\delta}) - J(\hat{\delta}))^2}{n-1} \tag{3.7}$$

y sesgo,

$$b_i = J(\hat{\delta}) - \hat{\delta} \tag{3.8}$$

3.1.2.3. *Métodos de estimación por intervalo*

Una ventaja de estimación por intervalo es que muestra la exactitud con que estima el parámetro, a menor longitud del intervalo mayor exactitud en la estimación.Un intervalo de confianza es un rango de valores, centrado en una media muestral $\bar{X}$, dentro del cual se espera que con un nivel de confianza $(1-\alpha)$ se encuentre el valor del parámetro en cuestión. Los métodos de estimación por intervalo son el método pivotal y el método general (ver Canavos (1988); Mendenhall and Sincich (1997)).

3.1.3. Validación de modelos y evaluación de resultados

Una vez conocidos los valores de m, es necesario realizar una validación de este modelo. Se debe cuantificar qué tan adecuadamente el modelo describe los datos (observaciones o simulaciones) para los cuales fue aplicado y cómo es el ajuste. Antes de proceder a evaluar el modelo obtenido es necesario re–examinar la formulación del problema para detectar posibles errores y determinar la consistencia de las expresiones matemáticas. La siguiente etapa consiste en evaluar algunos estadísticos de prueba como el coeficiente de correlación y los resultados de una prueba F; un análisis de varianza y un análisis de residuos también son de gran utilidad en esta etapa. En la evaluación de resultados se pueden variar los parámetros de entrada y verificar el comportamiento de los resultados, y si es posible, utilizar datos históricos para reconstruir el pasado y comparar éstos con los resultados del modelo. Finalmente es necesario verificar si las condiciones o supuestos iniciales coinciden con los resultados obtenidos, para esto es necesario el uso de pruebas de bondad de ajuste.

Hay dos factores importantes que se debe tener en cuenta en esta etapa;

- Los resultados obtenidos generalmente son el resultado de la conjunción de varios factores como tiempo, dinero y trabajo en grupo; por tanto es importante obtener la mayor cantidad de información y dar a conocer los resultados para que sean de utilidad.
- los valores obtenidos son el resultado de un trabajo consciente, por lo tanto se merecen un análisis real y objetivo.

3.1.3.1. *Pruebas de validación del modelo*

Al establecer un modelo se tienen dos diferentes fuentes de variación, una fuente de variación debida a la regresión (SCR) y una fuente de variación debida al error (SCE), la variación total (SCT) es la suma de estas dos. La variación se determina de la siguiente manera,

$$SCR = m'G'd - \frac{(\sum d_i)^2}{n} \tag{3.9}$$

$$SCE = d'd - m'G'd \tag{3.10}$$

$$SCT = d'd - \frac{(\sum d_i)^2}{n} \tag{3.11}$$

El coeficiente de correlación o coeficiente de determinación R^2 mide la proporción de variación total de las observaciones con respecto a su media que puede ser atribuida a la recta de regresión estimada y es definido como,

$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{SCT} \tag{3.12}$$

donde STC representa la variación total con respecto a la media y SCR la porción de variación que es atribuible a un efecto lineal de las variables predictoras sobre las variables de respuesta. Si $R^2 = 1$ puede afirmarse que toda la variación presente en las observaciones es explicada por la presencia de las variables predictoras G en la ecuación de regresión.

Una hipótesis estadística es una afirmación con respecto a alguna característica desconocida del sistema que interesa analizar. Pruebas estadísticas como la prueba F se realizan para probar la hipótesis nula sobre los parámetros $Ho : m_j = 0$ para todo j.

$$F = \frac{SCR/(l-1)}{SCE/(n-l)} \tag{3.13}$$

es la estadística $F_{(l-1,n-l)}$ con $l-1$ grados de libertad en el numerador y $n-l$ grados de libertad en el denominador, n es el número de observaciones y l es el número de parámetros. Si F es grande la mayor proporción de variación en los datos es debida a la regresión, la región de rechazo es $F > F_\alpha$, para un α determinado. La prueba F es una prueba de idoneidad general del modelo, que dice si los datos proporcionan o no pruebas suficientes que indiquen que el modelo global contribuye con información a la predicción de d.

3.1.3.2. *Pruebas de bondad de ajuste*

La validación de modelos busca detectar si una distribución de probabilidades supuesta es congruente con un conjunto de datos dado. Para esto se utilizan pruebas de bondad de ajuste tales como la prueba chi–cuadrado o la prueba de Kolmogorov–Smirnov. Sea $X_1, \ldots, X_n$ los resultados obtenidos a partir de una muestra aleatoria de la cual se ha asumido que su distribución de probabilidades está determinada por la función de probabilidad $P_o(X)$ o la función de densidad de probabilidad $F_o(X)$, se plantea la hipótesis nula $Ho : F(X) = F_o(X)$, especificada de manera completa con respecto a todos los parámetros.

La evaluación de este supuesto se hace sobre los datos obtenidos a partir de una muestra de tamaño n los cuales se clasifican en k categorías (en caso discreto, en caso continuo los datos deben ser discretizados), el número de datos que caen en la i–ésima categoría es f_i, $\sum f_i = n$; la frecuencia observada en la i–ésima categoría se compara con e_i, el número de datos que se espera observar en la i–ésima categoría si la hipótesis nula es correcta. Ho es rechazada si existe una diferencia significativa entre lo observado y lo esperado. La prueba de bondad de ajuste Chi–cuadrado (χ^2) es calculada de,

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{3.14}$$

y tiene $k-1$ grados de libertad. χ^2 tiende a cero si Ho es cierta. La potencia de la prueba aumenta si el número (k) de categorías aumenta.

La prueba de bondad de ajuste Jarque–Bera (JB) usada para determinar si una muestra tiene

distribución normal con media y varianza no especificadas, corresponde a una prueba Chi–cuadrado con dos grados de libertad y es definida como,

$$JB = \frac{n}{6}[S^2 + \frac{(C-3)^2}{4}] \tag{3.15}$$

donde S es el sesgo, C es la curtosis y n es el tamaño de la muestra.

3.2. Simulación Estadística

La simulación es una técnica de muestreo estadístico controlado (experimentación) que se emplea conjuntamente con un modelo, para obtener respuestas aproximadas a problemas probabilísticos complejos; debe seguir las normas del diseño de experimentos para que los resultados obtenidos puedan conducir a interpretaciones significativas de las relaciones de interés. La construcción y operación de un modelo de simulación permite la observación del comportamiento dinámico de un sistema en condiciones controladas, pudiéndose efectuar experimentos para comprobar alguna hipótesis acerca del sistema bajo estudio. Básicamente se relacionan tres elementos: sistema (parte del mundo real que interesa estudiar), modelo (representación simplificada de un sistema) y computador. Generalidades sobre el tema puede consultarse en Rios et˜al. (2000); Calderón (1980); Luthe et˜al. (1982); Ross (1999).

La simulación como cualquier otra técnica, tiene algunas desventajas,

- Los modelos de simulación para computador son costosos y difíciles de construir y validar. En general debe construirse un programa para cada sistema o problema.
- La ejecución del programa de simulación, una vez construido, puede necesitar una gran cantidad de recursos.
- La gente tiende a usar la simulación cuando no es el mejor método de análisis; una vez las personas se familiarizan con la metodología, intentan emplearla en situaciones en que otras técnicas analíticas son más apropiadas.

En general, un estudio de simulación puede dividirse en las siguientes etapas,

1. Definición del problema y planificación del estudio, incluye una definición precisa del problema a resolver y de los objetivos del estudio.
2. Toma de datos.
3. Establecimiento del modelo de simulación, incluye establecer las suposiciones que se deben hacer, y definir el modelo a emplear.

4. Ejecución de la simulación. Consideración de las técnicas y metodologías simples y avanzadas requeridas para obtener los resultados de programación. Esto incluye aspectos tales como generación de variables aleatorias, métodos para manejo de variables, precisión de los estimadores, etc.

5. Ejecuciones de prueba del modelo.

6. Validación del modelo.

7. Diseño del experimento de simulación.

8. Ejecución del experimento.

9. Análisis de los resultados. Uso de técnicas estadísticas y gráficas para interpretar los resultados.

La validación del modelo de simulación se puede llevar a cabo siguiendo las técnicas estándares de los depuradores; una de ellas puede ser depuración por módulos o subrutinas, es decir, descomponer el programa en partes pequeñas y controlables, donde cada una tenga una secuencia lógica, verificando por separado cada parte. Otra técnica útil es el seguimiento o rastreo, en la cual las salidas se imprimen después de la ocurrencia de cada evento, esto permite un seguimiento para determinar si la simulación está funcionando como se esperaba (Ross, 1999).

Los problemas que pueden resolverse mediante simulación se clasifican en probabilísticos y determinísticos. Algunas aplicaciones de la simulación estadística son,

- Técnicas de remuestreo como el Bootstrap y Jackknife permiten comparar las estimaciones usando diferentes tamaños de muestras y poco análisis, pero requieren un gran esfuerzo computacional, por lo cual son mas eficientes si se llevan a cabo mediante simulación.

- En Inferencia bayesiana se requieren métodos eficientes de integración, y cuando los problemas son de alta dimensión, los métodos más eficientes de integración son los basados en simulación.

- Un algoritmo probabilístico es un algoritmo que recibe entre los datos de entrada números aleatorios; así puede producir diferentes resultados con distintos tiempos de ejecución. La simulación permite, por una lado la implementación de algoritmos probabilísticos, y por otro lado el análisis probabilístico de algoritmos.

- Muchos de los problemas tratados con inteligencia artificial son de inferencia y decisión. Por ejemplo la simulación en sistemas expertos probabilísticos –en el cual el conocimiento se representa sobre un dominio en el que hay incertidumbre, mediante

una red probabilística cuyos nodos se asocian a variables aleatorias y cuyos arcos sugieren influencia, asociando a cada nodo una tabla de probabilidades condicionadas– (Rich and Knight, 1994), o la inferencia y predicción en redes neuronales, mediante métodos monte carlo basados en cadenas de Markov –modelos de caja negra que permiten modelar rasgos no lineales en problemas de aproximación, regresión, suavizado, predicción y clasificación– (Hilera and Martinez, 1995).

Aunque los orígenes de la simulación científica se remontan a los trabajos de Student para determinar la distribución de la variable t que lleva su nombre, esta disciplina apareció mas tarde como una técnica numérica llamada métodos de Monte Carlo (Rios et˜al., 2000). A continuación se presentaran algunas características, propiedades y utilidades de este método.

3.2.1. Simulación Monte Carlo

El nombre y desarrollo del método se remontan al año 1944; su primera aplicación como herramienta de investigación se dio en el desarrollo de la bomba atómica (estudios de reacción nuclear) durante la II guerra mundial. Sin embargo, el desarrollo sistemático del método como herramienta tuvo que esperar a los trabajos de Harris y H. Kahn, en el año 1948 y de Fermi, N. Metropolis y S. Ulam en ese mismo año. En 1953 N. Metrópolis introdujo el algoritmo Metrópolis, que consiste en una caminata aleatoria sesgada cuyas iteracciones individuales se basan en reglas probabilísticas. Los primeros usos de los métodos Monte Carlo para obtener modelos de la Tierra fueron realizados por Keilis–Borok y Yanovskaya en 1967 y por Press en 1968 (Mosegaard and Tarantola, 2000). Sobre el método Monte Carlo puede consultarse Drakos (1995),Rubinstein (1981), y sobre aplicaciones en geofísica los trabajos de Mosegaard and Tarantola (1995) y Mosegaard et˜al. (1997).

Los métodos Monte Carlo son cálculos numéricos que utilizan una secuencia de números aleatorios para llevar a cabo una simulación estadística, con el fin de conocer algunas propiedades estadísticas del sistema. Estos métodos de simulación están en contraste con los métodos numéricos de discretización aplicados para resolver ecuaciones diferenciales parciales que describen el comportamiento de algún sistema físico o matemático. La característica esencial de Monte Carlo es el uso de técnicas de toma de muestras aleatorias para llegar a una solución del problema físico, mientras una solución numérica convencional inicia con un modelo matemático del sistema físico, discretizando las ecuaciones diferenciales para luego resolver un grupo de ecuaciones algebraicas.

Con estos métodos sólo se requiere que el sistema físico o matemático pueda ser descrito mediante una función de densidad de probabilidad, la cual una vez sea postulada o conocida, se requiere una forma rápida y efectiva para generar numeros aleatorios con esa distribución, y así se inicia la simulación haciendo muestreos aleatorios de la misma. Después de múltiples simulaciones, el resultado deseado se toma como el valor promedio de los resultados obtenidos en cada simulación (Fig. 3.2). En muchas aplicaciones prácticas se puede predecir

un error estadístico (varianza) para este promedio y por tanto una estimación del número de simulaciones necesarias para conseguir un error dado. De entre todos los métodos numéricos basados en evaluación de n organismos en espacios de dimensión r, los métodos de Monte Carlo tienen asociado un error absoluto de estimación que decrece como $\sqrt{n}$ mientras que para el resto de métodos tal error decrece en el mejor de los casos como $\sqrt[r]{n}$ (Drakos, 1995). Los métodos Monte Carlo se usan para simular procesos aleatorios o estocásticos, dado que ellos pueden ser descritos como funciones de densidad de probabilidad, aunque con algunas restricciones, ya que muchas aplicaciones no tienen aparente contenido estocástico, tal como la inversión de un sistema de ecuaciones lineales.

Rubinstein (1981) resalta las siguientes diferencias entre simulación y simulación mediante métodos Monte Carlo,

- En el método Monte Carlo el tiempo no es importante, como si lo es es una simulación estocástica.
- Las observaciones en el método Monte Carlo son independientes. En simulación los experimentos dependen del tiempo, de tal manera que las observaciones son correlacionadas.
- En el método Monte Carlo la respuesta se puede expresar como una simple función de las variables aleatorias de entrada, mientras que en simulación la respuesta solo puede ser expresada explícitamente por el propio programa.

En la sección 3.1.1.1 (pág. 19), se describió ampliamente lo que es un sistema. Ahora bien, la respuesta de un sistema puede estar determinada por la sucesión o combinación de estados, los estados pueden cambiar en ciertos instantes de tiempos (de lo contrario es necesario discretizar el tiempo, para llevar a cabo la simulación), los cuales pueden ser síncronos –el paso de un estado a otro depende de un tiempo fijo–, o asíncronos –el cambio de estado depende de la obtención de un suceso, el tiempo es insignificante–. Los métodos de simulación Monte Carlo son aplicados a estos últimos, y dependiendo de la manera como pasan de un estado a otro se clasifican en:

1. Métodos basados en cadenas de Markov (el algoritmo Hastings–Metrópolis, el muestreador de Gibbs, temple simulado, muestreo con remuestreo de importancia y Monte Carlo híbrido)
2. Métodos independientes (muestreo de importancia y muestreo de rechazo)

Una cadena de Markov es una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediatamente anterior, es decir son cadenas con memoria lo cual condiciona las probabilidades de los eventos futuros (probabilidad de transición).

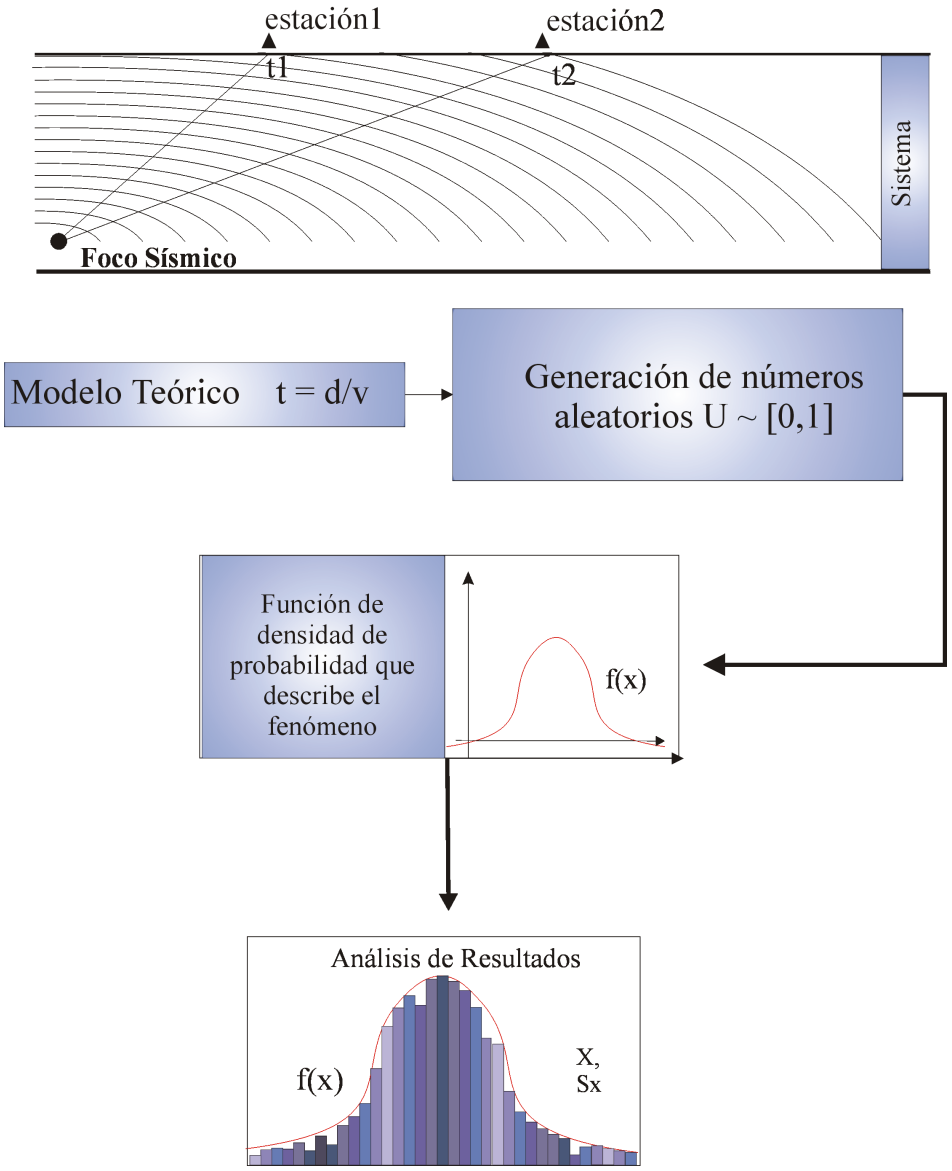

Figura 3.2: Esquematización de una simulación mediante el método Monte Carlo

3.2.2. Determinación del número de muestras

La determinación del tamaño muestral del experimento de simulación (n), es decir el número de veces que se observa el proceso, influye esencialmente en la precisión de la estimación (ver sec. 3.1.2.1), y dado que la precisión de las estimaciones aumenta en proporción directa a $\sqrt{n}$, es necesario tomar un tamaño de muestra suficientemente grande si se quiere obtener cierta precisión.

En principio, si en un proceso la variable x_i se puede observar n veces ($n > 30$), los estimadores muestrales de la media $\bar{X}$ y la varianza S^2 se pueden obtener de

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.16}$$

y

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.17}$$

respectivamente; la precisión de la estimación de $\bar{X}$ puede ser estimada a partir de $\frac{S}{\sqrt{n}}$.

Por el teorema del límite central, se tiene que la distribución muestral de una variable aleatoria tiende a una distribución normal (Z) para una muestra suficientemente grande y la desigualdad de Tchebychev proporciona una cota para la probabilidad de que una variable aleatoria X asuma un valor dentro de k desviaciones estándar alrededor de la media ($k > 1$). Las k desviaciones son entonces definidas por la distribución de probabilidades normal para un α determinado.

Por lo tanto, en el intervalo de confianza

$$[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}] \tag{3.18}$$

de amplitud $2z_{\alpha/2}\frac{S}{\sqrt{n}}$ con un nivel de confianza de $(1-\alpha)$ se encuentra el verdadero valor de X.

Ahora bien, fijando un valor aceptable como nivel de confianza $(1-\alpha)$, y determinando un valor máximo para ϵ, la amplitud del intervalo, se tiene

$$2z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \epsilon \tag{3.19}$$

Despejando n de la ecuación anterior se puede obtener una estimación del mínimo tamaño de muestra necesario para obtener estimaciones con un error aceptable d y un nivel de confianza $1-\alpha$,

$$(2z_{\alpha/2}\frac{S}{\sqrt{\varepsilon}})^2 \leq n \tag{3.20}$$

De esta manera, a partir de una muestra piloto suficientemente grande ($n > 30$) se obtiene una estimación de la varianza de los datos, y de esta forma se puede obtener el tamaño de muestra necesaria para obtener estimadores precisos, dados un nivel de confianza y un error permisible para $\bar{X}$.

3.2.3. Generación de números aleatorios

Sobre el tema existe múltiple bibliografía y en general se puede consultar cualquier libro de métodos numéricos como por ejemplo Gerald (1991); Chapra and Canale (1999); Press et˜al. (1997). Algunos métodos de generación de números aleatorios se basan en el uso de mecanismos físicos, por ejemplo el ruido blanco producido por circuitos electrónicos, el recuento de partículas emitidas, el lanzamiento de monedas, etc. El uso de estos mecanismos es poco conveniente, ya que puede presentar sesgo y dependencias; además una fuente de números aleatorios debe ser reproducible de manera que puedan hacerse réplicas de los experimentos en las mismas condiciones, lo cual implicaría el almacenamiento de los números, que conlleva al posible problema de límite de memoria y lentitud de acceso a los datos.

Otro método de generar números aleatorios es mediante el uso de algoritmos de computador, a pesar de que en principio los computadores son máquinas determinísticas incapaces por sí solas de un comportamiento aleatorio. Existen diversos algoritmos de generación de números aleatorios o pseudo–aleatorios, donde la idea es producir números que parezcan aleatorios, empleando las operaciones aritméticas del computador, partiendo de una semilla inicial. Se busca que la serie generada sea independiente, su generación sea rápida, consuma poca memoria, sea portable, sencilla de implementar, reproducible y suficientemente larga.

La generación de números aleatorios exige contrastar ciertas propiedades estadísticas de la salida. Para ello se realizan pruebas de contraste, como las pruebas de bondad de ajuste χ–cuadrado, Kolmogorov–smirnov, Cramer–von Mises, prueba de rachas y repetición de contrastes. (Ver Rubinstein (1981); Mendenhall and Sincich (1997)).

Los algoritmos de generación se clasifican en:

- generadores congruenciales, siguen la formula recursiva $x_{n+1} = (ax_n + b)$ mod m, donde a es el multiplicador, b el sesgo, m el módulo y x_o la semilla; a y b son constantes en el intervalo $(0, 1, \ldots, m-1)$, módulo m se refiere al residuo de la división entera por m; el período de la serie es $m-1$. Una adecuada selección de los parámetros a, b y m generan una sucesión de números suficientemente larga y aleatoria. Estos generadores tienen dos ciclos y la longitud del ciclo depende de los parámetros. Un generador congruencial estándar debe ser de período máximo y su implementación eficiente debe poder ser realizada en aritmética de 32 bits.

- de registro de desplazamiento, son recursivos múltiples o lineales de orden mayor, $x_n = (a_1 x_{n-1} + \ldots + a_k x_{n-k})$ mod m.

- de Fibonacci retardados, parten de la semilla inicial $x_1, x_2, x_3, \ldots$ y usan la recursión $x_{i1} = x_{i-r} \triangle x_{i-s}$, donde r y s son retardos enteros que satisfacen $r \geq s$ y $\triangle$ es una operación binaria que puede ser suma, resta, multiplicación, etc.

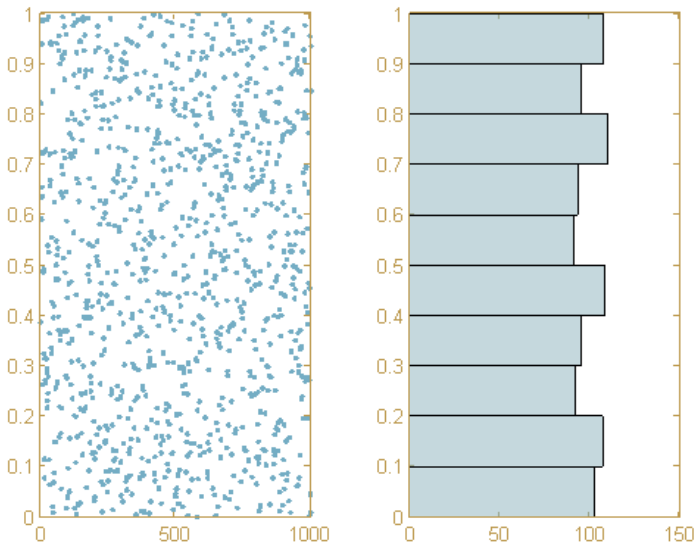

Figura 3.3: Números aleatorios distribuidos uniformemente en (0,1), generados con la función $rand$ de $Matlab$.

- no lineales, para introducir no linealidad se usa un generador con función de transición lineal, produciendo la salida mediante una transformación no lineal del estado, o usando un generador con función de transición no lineal. No producen estructura reticular, sino una estructura altamente no lineal.
- combinación de generadores (empíricos); por ejemplo, si se tienen dos sucesiones aleatorias se puede generar una nueva sucesión combinando los elementos correspondientes de ambas mediante una operación binaria.

Los números aleatorios así generados tienen la forma $u_n = x_n/m$, con distribución uniforme en el intervalo $(0, m)$. Estas series pueden ser escaladas dividiendo cada término entre m para obtener números uniformemente distribuidos en el intervalo (0,1) (Fig. 3.3). A partir de esta distribución se pueden generar series de números aleatorios con la distribución deseada.

3.2.3.1. *Generación de números aleatorios con distribución normal*

Sean U_1 y U_2 dos series de números aleatorios con distribución uniforme en (0,1). Los siguientes métodos permiten su transformación para obtener series de números aleatorios Z_1 y Z_2 distribuidas normalmente (Fig. 3.4).

- Inversión aproximada: $Z_1 = \frac{U_1^{0,135} - (1-U_1)^{0,135}}{0,1975}$

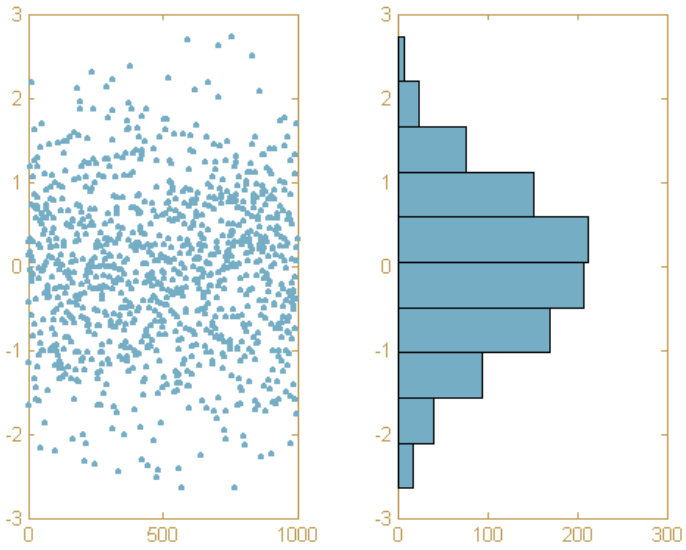

Figura 3.4: Números aleatorios con distribución normal con media 0 y desviación estándar 0.1, generados con la función $normrnd$ de $Matlab$.

- Box Muller, haciendo $r = \sqrt{-2lnU_1}$, y $\theta = 2\pi U_2$ se generan las variables con distribución normal estándar $Z_1 = r\,cos\theta$ y $Z_2 = rsen\theta$.
- Variante de Marsaglia. Es una variante del método Box Muller, que evita las operaciones de senos y cosenos. Se tiene $v_1 = 2U_1 - 1$, $v_2 = 2U_2 - 1$, $w = v_1^2 + v_2^2$, se hace mientras $w \leq 1$, $c = (-2lnw)/w$, y se obtienen las variables con distribución normal estándar $Z_1 = cv_1$ y $Z_2 = cv_2$

3.2.3.2. *generación de números aleatorios con Matlab*

La función para generación de números aleatorios con distribución uniforme en Matlab es *rand*, un generador congruencial multiplicativo con parámetros $m = 2^{31} - 1 = 2147483647$ por la facilidad de implementación y por que es un número primo, $a = 7^5 = 16807$ –ya que 7 es una raíz primitiva de $2^{31} - 1$, se obtiene el máximo periodo– y $b = 0$, valores recomendados por S. K. Park y K.W. Miller en 1988 en *Random number generators: good ones are hard to find* (Cleve, 1995; Rios et˜al., 2000). La función *rand* genera todos los números reales de la forma n/m para $n = 1, \ldots, m - 1$, la serie se repite después de generar $m - 1$ valores, lo cual es un poco mas de dos billones de números. En un computador Pentium a 75 MHz se puede agotar el período en poco más de 4 horas.

A partir de una serie de números generados aleatoriamente con distribución uniforme en $(0, 1)$ se pueden generar series de números aleatorios con distribución normal. La función de

$Matlab$ para su generación es $normrnd$ (sección 5.1.4, pág. 61)

3.2.4. Análisis estadístico de datos simulados

Como ya se ha mencionado antes, la simulación tiene como objetivo determinar el valor de una cantidad relacionada con un modelo estocástico particular. Una simulación produce datos de salida (X), cuyo valor esperado es la cantidad de interés. Un número n de repeticiones de la simulación produce $X_1, \ldots, X_n$ resultados, los cuales conforman la muestra; el promedio o media muestral (ecuación 3.16) de todos estos resultados proporciona una estimación del valor de interés. La media es un estimador insesgado, ya que su valor esperado es igual al valor del parámetro. Para determinar la bondad de la media como estimación del parámetro, se calcula la varianza de la media muestral (ecuación 3.17), si ésta es pequeña, se dice que la media es un buen estimador del parámetro. Esto es justificado por la desigualdad de *Tchebychev*, ya que para una muestra suficientemente grande, la probabilidad que una variable aleatoria quede a muchas desviaciones estándar de la media es muy pequeña.

3.3. Optimización global y local

La optimización involucra la búsqueda del mínimo o del máximo de una función, y está relacionada con la determinación del mejor resultado u óptima solución al problema. Óptimo es el punto donde la curva es plana, es decir el valor de x donde la primera derivada $f'(x)$ es cero; para determinar si el óptimo es máximo o mínimo se evalúa la segunda derivada, si $f"(x) < 0$ el punto es un máximo, de lo contrario es un mínimo. En un problema de optimización se requiere

- la definición de una función objetivo
- identificación de las variables de diseño o de predicción
- restricciones o limitaciones reales, es decir bajo las cuales se trabaja (optimización restringida).

Los métodos de optimización pueden ser unidimensionales (una sola variable) o multidimensionales (mas de una variable); restringidos o no restringidos; lineales, no lineales o cuadráticos. A continuación se resumen algunos de los métodos mas comúnmente usados. Estos métodos y otros adicionales pueden ser consultados en Press et˜al. (1997); Gerald (1991); Chapra and Canale (1999).

3.3.1. Optimización no restringida

La optimización multidimensional sin restricciones usa métodos directos, los cuales no requieren evaluación de la derivada (o gradiente en caso multidimensional) y métodos indirectos o gradiente (de descenso o ascenso) que requieren la evaluación de la derivada.

3.3.1.1. *Métodos directos*

Algunos métodos directos son aplicaciones de los métodos de Simulación Monte Carlo al problema de optimización,

1. Búsqueda aleatoria pura o fuerza bruta. Evalúa en forma repetida la función mediante la selección aleatoria de valores de la variable independiente. Si un número suficiente de muestras es evaluado, el óptimo será eventualmente localizado. Trabaja en discontinuidades y funciones no diferenciables. Siempre encuentra el óptimo global, pero no es eficiente ya que requiere mucho esfuerzo de implementación dado que no toma en cuenta el comportamiento de la función, ni los resultados de las iteraciones previas para mejorar la velocidad de convergencia. Un ejemplo es la búsqueda por malla (*Grid Search*), donde las dimensiones de x y y se dividen en pequeños incrementos para crear una malla, la función se evalúa en cada nodo: entre mas densa es la malla la probabilidad de localizar el punto óptimo es mayor (Lomax et˜al., 2000).

2. Multicomienzo, Es una mejora de la búsqueda aleatoria pura. Se genera un número de puntos desde los que se inicia una optimización local, el óptimo local obtenido es propuesto como solución inicial para la optimización global. Un ejemplo de este método es el *Neighborhood Algorithm* (Sambridge and Kennett, 2001) utilizado para resolver el problema de la determinación de los parámetros de un sismo.

3. Univariabilidad y búsquedas patrón. Cambia una variable a la vez para mejorar la aproximación, mientras las otras variables se mantienen constantes, así el problema se reduce a una búsqueda unidimensional que se puede resolver por medio de diversos métodos (Newton, de Powell, interpolación cuadrática).

4. Otras búsquedas aleatorias son de origen heurístico. Las dos primeras buscan evitar que el algoritmo quede atrapado en un óptimo local,

    - Recocido simulado. Al principio se aceptan todas las transiciones entre soluciones, lo que permite explorar todo el conjunto factible; gradualmente la aceptación de movimientos se hace mas selectiva; finalmente solo se aceptan los movimientos que mejoran la solución actual. Este método permite empeoramiento en la solución mediante reglas probabilísticas. Aplicación de este método al problema de localización hipocentral se encuentra en Billings (1994).

- Búsqueda tabú. Permite el paso de una solución a otra, aún cuando ésta empeore la solución. Para evitar el ciclado se genera una lista en la cual se guarda durante cierto tiempo un atributo que permite identificar la solución o el movimiento realizado, para de esta forma no permitir su realización. Dado que la lista está constituida por atributos y por soluciones o movimientos, el paso a soluciones óptimas podría verse restringido, para evitar esto se utilizan niveles de aspiración, si una solución o movimiento de la lista supera los niveles de aspiración, se permite el movimiento a ella.
- Algoritmos genéticos. Es una variante del método de multicomienzo. Por medio de operaciones que se derivan de los principios de la evolución natural se identifican soluciones prometedoras, a partir de las cuales se puede realizar una optimización local. Son buenos sólo para identificar regiones óptimas, aunque combinados con métodos de búsqueda local pueden encontrar la solución óptima. Billings et˜al. (1994a) presenta una aplicación de algoritmos genéticos en la estimación hipocentral.
- Redes neuronales artificiales.

3.3.1.2. *Métodos indirectos*

Algunos métodos indirectos son

1. Método de pasos ascendente (maximización) o descendente (minimización), determina la mejor dirección de búsqueda (mediante el gradiente) y establece el mejor valor a lo largo de esa dirección.

2. Gradiente avanzado, tales como gradiente conjugado, Newton, Marquardt, quasi–Newton o variable métrica. El método de Newton por ejemplo usa cálculo del gradiente a partir de la expansión de la serie de Taylor e inversión de la matriz; converge si el punto inicial está cerca del óptimo. El método Marquardt usa pasos ascendentes mientras está lejos del óptimo, y Newton cuando está cerca del óptimo. El método de localización hipocentral implementado por Geiger en 1910 (Lee and Lahr, 1975) expuesto en la sección 4.1.3.1 utiliza el método indirecto de gradiente conjugado de Newton.

A continuación se muestra un ejemplo que trae Matlab en su *toolbox* de optimización (Fig. 3.5). Consiste en la minimización de la función de *Rosenbrock* $f(x) = 100 * (x_2 - x_1^2)^2 + (1 - x_1)^2$. Es usada como ejemplo por la lenta convergencia que presenta con algunos métodos. La función tiene un único mínimo en el punto $x = [1, 1]$ donde $f(x) = 0$. El ejemplo inicia en el punto $x = [-1{,}9, 2]$. El número de evaluaciones de la función y de iteraciones son las siguientes: a. Número de iteraciones 13, evaluaciones de la función 55, b. Número de iteraciones 21, evaluaciones de la función 92, c. Número de iteraciones 68, evaluaciones de la función 302, d. Número de iteraciones 109, evaluaciones de la función 201.

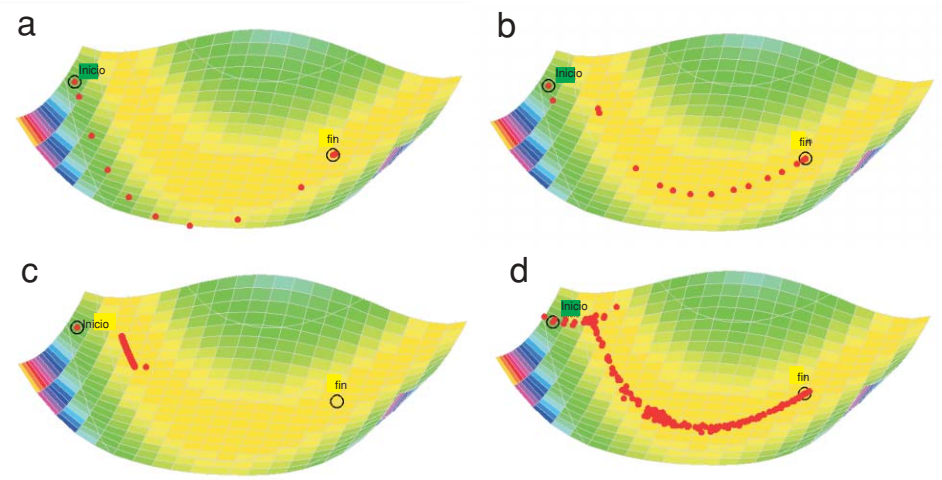

Figura 3.5: Minimización de la función *Rosenbrock*. a. Mínimos cuadrados mediante Gauss–Newton, b. Levenberg–Marquardt, c. Pasos Descendentes, d. Método Simplex

3.3.2. Optimización restringida

La optimización multidimensional con restricciones puede ser lineal (la función objetivo y las restricciones son lineales) o no lineal (la función objetivo es no lineal).

1. Optimización lineal o programación lineal. El objetivo es, dadas n variables independientes y ciertas restricciones, maximizar la función objetivo $z = a_1x_1, + \ldots + a_nx_n$. Los resultados que se pueden obtener son: a. solución única, b. soluciones alternas, c. solución no factible, d. problemas sin límite.

    - Solución gráfica. En un problema de dos dimensiones, la solución espacial se define como un plano, las restricciones se trazan sobre este plano como líneas rectas, las cuales delinean el espacio de solución factible; la función objetivo evaluada en un punto se traza como otra línea recta sobrepuesta en este espacio. El valor de la función se ajusta hasta que tome el máximo valor, mientras se conserva dentro del espacio factible. Este valor representa la solución óptima.
    - Método simplex. Se basa en la suposición que la solución óptima está en un punto extremo, la dificultad está en definir numéricamente el espacio de soluciones factibles. Las ecuaciones con restricciones se formulan como igualdades introduciéndolas como variables de holgura. Rabinowitz (2000) implementó un método de localización mediante el método de optimización simplex usando restricciones no lineales.

2. Optimización no lineal. Para problemas indirectos se utilizan funciones de penalización, las cuales involucran usar expresiones adicionales para hacer la función objetivo menos óptima en tanto la solución se aproxima a la restricción, así la solución no

será aceptada por violar las restricciones. Para problemas directos se utiliza el método del gradiente reducido generalizado, el cual reduce el problema a uno de optimización no restringida, para ser resuelto con uno de los métodos descritos anteriormente.

3.4. Clasificación de errores

Cualquier medición es una observación indirecta, por lo tanto siempre es necesario llevar a cabo un análisis de errores. El error absoluto de una medida se puede definir como la diferencia entre el valor resultante de una medición y el valor real de la variable que se está observando. Valor verdadero es un concepto ideal y en general no puede ser conocido exactamente, lo mas cercano a éste será una medida tomada con un instrumento estándar de alta calidad o exactitud que presente un error muy reducido, que reduzca el error instrumental a un nivel insignificante con respecto a otras fuentes de error. La variación restante puede ser considerada como una combinación de muchas variables independientes (Baird, 1991).

En sismología instrumental, el proceso de medición suele estar dividido en varias etapas secuenciales, en donde los resultados parciales de una investigación (modelo) se convierten en los datos de entrada (mediciones) en una nueva etapa de análisis. En efecto, partiendo de la observación directa del movimiento del suelo en varias estaciones sismológicas se pueden leer tiempos de llegada de una onda sísmica, los cuales son usados para localizar la fuente de la perturbación (hipocentro). Un conjunto de hipocentros puede ser posteriormente usado para modelar el comportamiento de una falla o, conjuntamente con los tiempos de llegada, para modelar la estructura interna de corteza y manto superior. En este proceso continuo los errores son inevitables y es por eso que se hace indispensable considerarlos explícitamente para reducirlos y compensar sus efectos. Freedman (1968) presenta un completo análisis sobre los diferentes tipos de errores encontrados en mediciones de datos sismológicos.

Si se considera que los errores presentan variabilidad que depende del fenómeno estudiado, que los errores positivos y negativos son igualmente frecuentes (su valor promedio debe estar cercano a cero), que el error total no puede exceder una cantidad razonablemente pequeña, que la probabilidad de un error pequeño es mayor que la de un error grande y que la componente del error en el modelo es un efecto compuesto que representa muchas perturbaciones pequeñas pero aleatorias (las cuales son independientes de la variable de predicción y se deben a factores que no se encuentran incluidos en el modelo), puede entonces suponerse que los errores tienen una distribución de probabilidades normal.

Parte de la incertidumbre puede ser estimada usando métodos estadísticos, mientras que otra parte sólo puede ser determinada de manera subjetiva, usando sentido común o juicio científico sobre la calidad de instrumentos o sobre efectos no tenidos en cuenta explícitamente. De acuerdo a su origen, los errores que se presentan en los procesos directos o indirectos de medición en el problema de localización hipocentral, se clasifican en aleatorios y sistemáticos.

3.4.1. Errores aleatorios

Los errores aleatorios son pequeñas variaciones positivas o negativas producidas por causas desconocidas. Pueden ser cuantificados mediante análisis estadístico, por lo tanto, sus efectos pueden ser determinados. Por lo general los errores aleatorios se asocian con la participación humana en el proceso de medición y pueden deberse a;

- la identificación errónea del patrón que se está observando; como por ejemplo, la no identificación de una fase sísmica por exceso de ruido ambiental cuando ésta llega a la estación. Este error y el error de lectura son independientes.
- error instrumental; por ejemplo, los errores de cronometraje o las variaciones en la respuesta de los instrumentos debido a su deterioro
- error de lectura de tiempos de arribo, es el error residual, el cual puede permanecer aunque los demás errores sean eliminados.

3.4.2. Errores sistemáticos

Los errores sistemáticos se deben a causas identificables. Estos pueden ser corregidos, permanecer constantes o variar en forma previsible y pueden ser clasificados como;

- Instrumentales. debidos a problemas de calibración, daño en los equipos, pérdida de señales durante su transmisión, etc.
- Observacionales, como la mala selección del instrumento, baja calidad de la respuesta instrumental, ajuste incorrecto del cero, etc.
- Naturales. Efecto de campos eléctricos y magnéticos.
- Teóricos. Simplificación de los modelos o aproximaciones en las ecuaciones

Los errores en la determinación de los tiempos de viaje de las ondas sísmicas debidos a un inadecuado modelo de velocidades, por ejemplo, incluyen componentes sistemáticas y aleatorias. Los errores sistemáticos aparecen generalmente cuando la estructura de velocidades no corresponde con las velocidades medias de la estructura real, en una escala comparable a la del modelo; mientras que pequeñas variaciones locales de la estructura, alrededor de los valores del modelo, generan diferencias que pueden ser tratadas como errores aleatorios en la mayoría de los casos.

3.4.3. Errores en la determinación hipocentral

Muchas estimaciones convencionales del error se basan en la aproximación lineal de un grupo de ecuaciones no lineales, aunque la no linealidad pueda ser vista como un diferente tipo de sesgo sistemático cuyo efecto es inseparable del efecto debido a los errores del modelo. Pavlis (1986) propone, sin embargo, que la influencia de la no linealidad probablemente es pequeña para la mayoría de las localizaciones y puede ser controlada en gran parte por sesgo en errores del modelo.

Las fuentes de error durante el proceso de localización son diversas y están presentes en los datos -errores en la medición y en la identificación de fases, diferencias entre el modelo de velocidades y la estructura real-, en los métodos de estimación -estimaciones mediante mínimos cuadrados, por ejemplo, son adecuadas cuando no se violan los supuestos sobre el error (aleatoriedad e independencia)-, y en componentes externos como el número y la distribución de las estaciones con respecto a las fuentes.