UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

DISEÑO Y ANÁLISIS ESTADÍSTICO DE UN EXPERIMENTO ATMOSFÉRICO POR COMPUTADORA

T E S I S
QUE PARA OBTENER EL TÍTULO DE
A C T U A R I A
P R E S E N T A
PATRICIA BAUTISTA OTERO

DIRECTOR DE TESIS
DR. CARLOS DÍAZ ÁVALOS

2004

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito: Diseño y Análisis Estadístico de un Experimento Atmosférico por Computadora,

realizado por Patricia Bautista Otero

con número de cuenta 94248423 , quien cubrió los créditos de la carrera de: Actuaría.

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dr. Carlos Díaz Avalos
Propietario	M. en C. Maria del Pilar Alonso Reyes
Propietario	M. en C. José Antonio Flores Díaz
Suplente	Dr. Luis Antonio Rincón Solis
Suplente	Dr. Jesús López Estrada

Consejo Departamental de Matemáticas.

Act. Jaime Vázquez Alamilla.

A mamá, Lucia y hermanos ...

Resumen

Hoy en día muchos procesos son estudiados a través de modelos por computadora. Los calculos hechos con estos modelos son referidos como experimentos por computadora o simulaciones. Un problema común en el uso de estas herramientas es encontrar la configuración más apropiada para un caso de estudio particular, lo que requiere de un número considerable de ensayos. Si un experimento no resulta demasiado costoso en cuanto a los recursos de cómputo que demanda y si el número de parámetros que se quiere ajustar no es grande, entonces es posible explorar casi por completo el espacio parametral; sin embargo lo que ocurre con mayor frecuencia es lo contrario. Una alternativa para explorar dicho espacio es ajustar un predictor a la respuesta del modelo. En este trabajo se construye un indicador que concentra la respuesta multidimensional, el cual es modelado como una realización de una función aleatoria, utilizando el método de geoestadística conocido como *Kriging*. Una de las bondades de este método es que permite estimar la incertidumbre de las predicciones. Este procedimiento es aplicado al *Sistema Regional de Simulación de la Atmósfera* (RAMS, por sus siglas en inglés). La exploración del espacio de parámetros, es decir, el diseño del experimento se realiza con el método de muestreo *cuadrado latino*.

Agradecimientos

Agradezco a CONACyT el apoyo económico, al Instituto Mexicano del Petróleo por permitirme el uso de la supercomputadora Cray Origin 2000, al Programa de Investigación en Medio Ambiente y Seguridad del mismo Intituto por permitirme la experimentación con el modelo numérico RAMS. Gracias también al Dr. Arturo I. Quintanar por el entrenamiento en el uso del modelo así como por su valiosa aportación en el área de Meteorología y gracias al Dr. Carlos Díaz por su colaboración en el área de Estadística y por ser guía de este trabajo.

Índice general

Índice de figuras

Índice de cuadros

Introducción

La simulación por computadora hoy en día tiene un impacto importante en la investigación científica. Muchos procesos son tan complejos que la experimentación física lleva demasiado tiempo o es muy costosa o incluso imposible, tal como ocurre con el modelado del estado del tiempo. En estos casos resultan de mucha utilidad los modelos numéricos que son llevados a las computadoras en forma de programas, convirtiéndose así en *modelos por computadora.* Los cálculos hechos con dichas herramientas son referidos como *experimentos por computadora* o *simulaciones.*

Como ejemplo se puede citar el Modelo de Circulación General de Océanos, Bryan-Cox (Cox y Bryan, 1984) que como su nombre lo dice, es usado en la simulación de la circulación de agua que puede haber en un océano. Otro ejemplo interesante es CALMET, (Modelo Meteorológico de California) (Scire y Robe, 2000), cuya aplicación principal está en la simulación de la dispersión de contaminantes en el aire en un área limitada. Estos sistemas son construidos con base a teorías físicas y/o matemáticas, y generalmente debido a que pueden ser aplicados a una variedad de propósitos es que presentan un número considerable de parámetros que el usuario debe establecer. Dichos parámetros pueden representar procesos físicos no incluidos explícitamente en el modelo, como por ejemplo: ondas de gravedad atmosférica o la escala de mezclado en una submalla. También podrían ser condiciones iniciales o de frontera, como la constante solar o la temperatura de la superficie del mar.

Frecuentemente el usuario de un modelo por computadora, a fin de calibrarlo, debe realizar un número considerable de simulaciones. Considérese el caso de otro modelo meteorológico, MM5 (Simulador de Mesoescala 5) (Dudhia et al., 2003) en el que el interés está únicamente en cinco parámetros de entre todos los que incluye que configurados apropiadamente permitirían la integración del modelo en una región to-

pográficamente complicada; si se hacen pruebas únicamente con los extremos y puntos medios de los intervalos en que éstos varían, entonces se deberían realizar 3^5=256 experimentos que luego deberían ser analizados para encontrar la configuración adecuada. El número de simulaciones que se puede realizar depende de los recursos de cómputo de que se disponga así como del tiempo que tome cada simulación (en algunos casos el tiempo puede ser de semanas, incluso meses, dependiendo de si se realizan en una estación de trabajo, como una PC, o en una supercómputadora). Realizar y analizar los resultados de 256 experimentos es poco práctico y pocas veces posible.

En este trabajo se presenta una alternativa que consiste en integrar el modelo bajo diferentes configuraciones de prueba y usar los resultados para predecir el efecto que causarían otras configuraciones en la respuesta del modelo. Con esto se consigue tener una descripción cualitativa de la dependencia entre la respuesta y los parámetros de estudio, con lo cual sería sencillo para el usuario llegar a la calibración.

El modelo que es usado para ejemplificar este procedimiento se llama, *Sistema Regional de Simulación de la Atmósfera*, (RAMS por sus siglas en inglés) (Walko et al., 1999). Este sistema fue creado para simular y pronosticar fenómenos meteorológicos. Los parámetros seleccionados son dos: TNUDLAT y TNUDCENT, dos magnitudes de tiempo que representan las variables de ajuste de un procedimiento de optimización llamado *Asimilación de Datos en 4 Dimensiones* (4DDA por sus siglas en inglés) (Smedstad y Fox, 1994), (M. Ghil, 1989). En el capítulo 1 se encuentra una explicación más amplia sobre los parámetros y el modelo.

Los valores para los parámetros seleccionados con los que se llevan a cabo las simulaciones son obtenidos por el muestreo cuadrado latino, que se abreviará como MCL, el cual es una variación del muestreo por cuotas (Arber, 1995). El MCL garantiza que todas las porciones del espacio estén representadas. En la sección 1 del capítulo 2 se explica la metodología de este procedimiento. La técnica para predecir el comportamiento del modelo en diferentes configuraciones esta basada en un método de interpolación exacta (exacta por que reproduce los valores observados), extraído de la geoestadística, llamado *Kriging* (Armstrong, 1998), que como es explicado en el capítulo 3, resulta más apropiado que los métodos estándar de interpolación. Hay otros enfoques potencialmente útiles diferentes al *Kriging* que se han aplicado en problemas similares. Por ejemplo, en (Bowman et al., 1993), (Sacks, 1989) y (Chapman et al., 1994) se construyen funciones estadísticas de aproximación para cada variable modelada considerando como variables independientes a los parámetros de estudio.

Capítulo 1

El modelo y los parámetros

En este capítulo se incluye información sobre el modelo por computadora, los datos con los cuales es inicializado así como la respuesta o pronóstico del mismo, los parámetros de estudio, y el lugar y días de simulación.

1.1. El modelo por computadora

RAMS, *Sistema Regional de Simulación de la Atmósfera*, es el nombre del modelo que se usa en este trabajo; la descripción completa puede ser encontrada en (Pielke et al., 1992), aunque las características más importantes son presentadas aquí. RAMS es un programa por computadora desarrollado para simular y pronosticar fenómenos meteorológicos. Las aplicaciones más comunes del modelo incluyen la simulación y modelado de: turbulencia, tornados, campos de cúmulos, sistemas convectivos de mediana escala, tormentas y dispersión atmosférica. A manera de ilustración en las figuras 1.1 y 1.2 se muestran los resultados de la simulación del huracán Georges llegando a Puerto Rico, a las 20:30 GMT del 21 de Septiembre de 1998.

El código del programa está construido a partir de un conjunto de ecuaciones que describen la dinámica y la termodinámica de la atmósfera (es decir cómo cambian en el tiempo los campos de vientos y la densidad del aire), a éstas se les añade una gran selección de parametrizaciones para los procesos de difusión turbulenta, radiación solar

y terrestre, y procesos de cambio de fase que incluyen la formación e interacción de nubes. También se consideran los efectos cinemáticos de la orografía, la convección de nubes cumulus (convección profunda) y los intercambios de calor entre la atmósfera y la superficie, considerando varias capas que incluyen vegetación y cubiertas de nieve y agua. En (Walko et al., 1999) se encuentra una descripción completa de las ecuaciones y parametrizaciones usadas en el modelo.

RAMS incluye un rango amplio de opciones que el usuario puede seleccionar en la configuración de una simulación particular. Las razones principales de ello son básicamente dos: la primera es que algunas de las funciones son necesarias para algunas aplicaciones, aunque no para todas, y la segunda es tener un modelo que permita la experimentación de diferentes esquemas de parametrización como herramienta en la investigación del modelado de la atmósfera.

El sistema operativo bajo el que se ejecuta RAMS puede ser Unix, Linux, o sistemas NT. La mayor parte del código está escrito en FORTRAN, el resto en C para facilitar los procesos de lectura y escritura.

1.1.1. Origen del modelo

RAMS fue desarrollado en conjunto por investigadores de la Universidad del Estado de Colorado y de ASTeR, Inc. (Simulación Atmosférica, Experimentación e Investigación). La finalidad fue combinar dos modelos atmosféricos, uno de física de nubes (Cotton et al., 1982) y otro de mesoescala (Pielke, 1984). En 1988 la primera versión del sistema de simulación fue utilizada en la investigación. El desarrollo de la versión en paralelo comenzó en 1991 con el uso de PVM (Máquina Virtual Paralela) y en 1996 se terminó la versión con soporte para MPI (Interfaz de Envío de Mensajes). La versión más reciente de RAMS es la Versión 4.4., (2001).

1.2. Inicialización

El modelo es inicializado con dos bases de datos: En la primera se incluyen las características topográficas del terreno y en la segunda las observaciones de las condiciones atmosféricas.

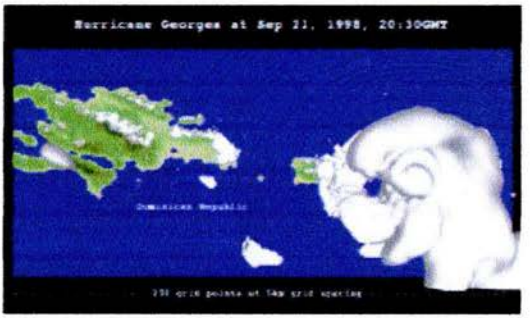

Figura 1.1: RAMS: Simulación del huracán Georges llegando a Puerto Rico a las 20:30 GMT del 21 de Septiembre de 1998. Las líneas indican la dirección del viento en la superficie. El modelo examina los efectos de la temperatura en la superficie del mar en la intensidad del huracán. A las 8:50 horas, dentro de las 48 horas de integración, el ojo del huracán está bien definido. La gráfica fue tomada del sitio http://www.npaci.edu/ envision/v17.3/hurricanes.html.

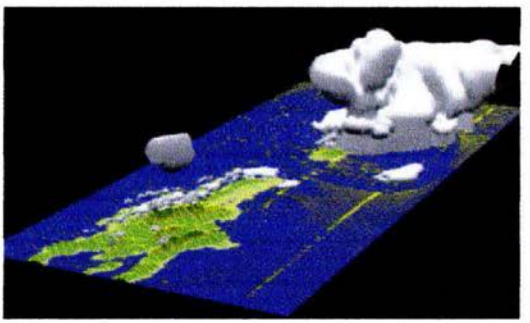

Figura 1.2: Una perspectiva de la vista del huracán Georges. Se muestra el tamaño del huracán comparado con las nubes que están sobre República Dominicana. Es interesante observar el desarrollo del ojo del huracán y la altura de las nubes que señalan la zona de convección más intensa. La gráfica fue tomada del sitio http://www.npaci.edu/envision/v17.3/hurricanes.html.

La mayoría de las observaciones son hechas con radiosondas, que son instrumentos con la capacidad de transmitir por radio, las cuales son colocadas en un globo especial y lanzadas para que en su recorrido hacia la tropósfera realicen las mediciones de la **temperatura**, **humedad**, **dirección** y **velocidad del viento** a diferentes alturas. A estas mediciones se les conoce como radiosondeos y en general se realizan en el sitio y hora de interés para el investigador.

1.2.1. Observaciones

Las observaciones con que se cuenta para esta investigación son los datos históricos de una serie de radiosondeos efectuados los días 2, 3 y 4 de febrero de 1999 a las 00:00, 6:00, 12:00 y 18:00 horas, tiempo local. Las coordenadas de los cuatro sitios donde se realizaron están señaladas en la Sección 1.4. Las mediciones fueron tomadas a 29 diferentes alturas sin rebasar los 20 kilómetros.

1.3. Predicción

Una vez que el modelo ha sido provisto de la base de datos de inicialización puede ser integrado. La respuesta que dará entonces será la predicción de las condiciones atmosféricas (**temperatura**, **humedad** y **componentes del viento**) en cada punto de una malla en 3D (latitud, longitud y altura) cada hora, o cada intervalo de tiempo seleccionado. A la malla antes citada se le llama *malla del modelo*.

A las variables atmosféricas: **temperatura**, **humedad** y **componentes del viento (horizontal y vertical)** se les abreviará en lo sucesivo como *tempc*, *relhum*, u y v respectivamente.

1.4. Dominio de interés

Para fines de este trabajo el dominio del modelo se encuentra ubicado en México, en los alrededores de Ciudad del Carmen, donde PEMEX tiene la mayoría de sus plataformas

de extracción, fuente de más del 75 % de la producción nacional de petróleo. Con mayor precisión la región se localiza entre los 17.7° y 20.7° de latitud norte y los 90.5° y 94.5° de longitud oeste y hasta una altura de 20 kilómetros. La Figura 1.3 muestra la base del dominio.

La *malla del modelo* tiene su centro en 19.25° latitud norte y 92.5° longitud oeste. El número de nodos en latitud son 45, en longitud 40 y en altura 30. La distancia entre nodos es de 10 kilómetros para una altura fija, mientras que en la vertical los niveles son equiespaciados pero con respecto a la presión barométrica.

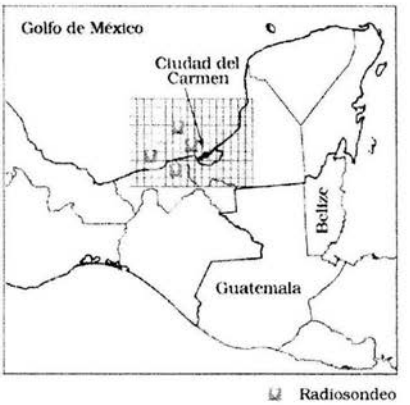

Figura 1.3: Base del dominio del modelo.

1.5. Parámetros seleccionados

Los parámetros que serán el objeto de estudio son aquellos que controlan la opción de ***Asimilación de Datos en Cuatro Dimensiones*** (4DDA por sus siglas en inglés), que se refiere a datos continuos en espacio y tiempo. A continuación se explica en qué consiste esta opción.

1.5.1. 4DDA, Asimilación de Datos en Cuatro Dimensiones

El concepto de 4DDA surge en meteorología para atacar un problema común que es generar predicciones a partir de muy pocas observaciones. En este esquema un modelo de pronóstico de campos atmosféricos[1] es actualizado secuencialmente con observaciones. A partir de que una actualización ocurre, el modelo llevará la información de un número finito de mediciones pasadas, sujeta a la dinámica apropiada, para ser combinada con las últimas observaciones. El proceso es ilustrado en la Figura 1.4, donde todos los datos dentro del intervalo de tiempo (-6h, +6h) son usados para actualizar el estado del sistema. En tiempos en los que se recibe información (0h,12h,24h,...) el pronóstico numérico es comparado contra las mediciones y luego son combinados, es decir, los datos son *asimilados* por el modelo y finalmente una nueva predicción es producida a partir del último estado estimado de la atmósfera.

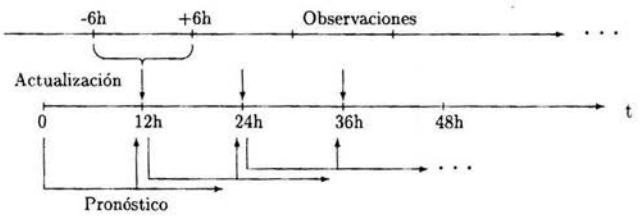

Figura 1.4: Ciclo de operación de un modelo de pronóstico en el que se combina la predicción y el proceso de asimilación de datos.

1.5.2. Parámetros

Existen varios procedimientos para implementar la asimilación de datos en cuatro dimensiones, en RAMS se usa el método llamado *Relajación Newtoniana* o también cono-

[1] En la actualidad el esquema 4DDA es utilizado en otras áreas como la oceanografía.

cido como *Nudging*, el cual consiste en *relajar* el pronóstico hacia las observaciones[2] mientras el modelo está siendo integrado. La fuerza del *nudging* está dada por

$$\frac{(\phi_{obs} - \phi_{mod})}{\tau}$$

donde ϕ representa las variables de pronóstico *u,v, tempc* y *relhum*. ϕ_{obs} es un valor observado en algún sitio particular, ϕ_{mod} es el pronóstico del modelo correspondiente y τ es el tiempo de relajación especificado por el usuario (Figura 1.5). Dicho tiempo puede ser establecido al definir sus distribuciones estándar, lo cual se logra a través de las variables **TNUDLAT**, **TNUDTOP**, y **TNUDCENT**. Estos parámetros definen tiempos de relajación en las zonas laterales, del tope y del centro del dominio de integración, respectivamente. La influencia de TNUDLAT se extiende hacia adentro a partir de las paredes del dominio hasta un cierto número de puntos de la malla especificado por el parámetro NUDLAT. La función de influencia (inversa del tiempo de *nudging*) se incrementa hacia afuera parabólicamente comenzando en el vértice de la parábola ubicado NUDLAT puntos frente al límite. El tiempo de relajación en el vértice y más adentro en el interior de la malla está definido por TNUDCENT. Por lo tanto, TNUDCENT puede ser usado para especificar un límite inferior para la fuerza del *nudging* a lo largo del dominio. La influencia de TNUDTOP se extiende hacia abajo a partir del tope del dominio hasta una altura especificada por el parámetro ZNUDTOP. La función de influencia se incrementa linealmente entre estas dos alturas, alcanzando el valor mínimo definido por TNUDCENT en y hacia abajo de ZNUDTOP.

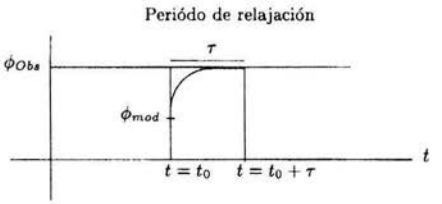

Figura 1.5: Relajación del pronóstico del modelo ϕ_{mod} hacia la observación ϕ_{Obs} en un tiempo τ. Si τ es un periódo de tiempo grande entonces la relajación es *suave*

[2]Esto quiere decir aproximar el pronóstico del modelo hacia las observaciones.

Los usuarios de RAMS que deseen hacer uso de la opción de 4DDA deben experimentar con las tres magnitudes de tiempo para determinar los valores que mejorarán el desempeño del modelo para una aplicación particular. Las únicas restricciones para la estabilidad numérica son: primero, que ninguno de los parámetros podrá ser menor que el *timestep*[3] del modelo (el cual está fijo en 12 segundos); segundo, TNUDLAT debe ser menor que TNUDCENT (considerando que la fuerza del *nudging* es la inversa del tiempo de relajación esto resulta natural). Según algunos experimentos hechos por los propios autores de RAMS (Walko et al., 1999), TNUDLAT y TNUDCENT deberían estar en rangos de 900 a 1,800 y de 7,200 a 14,000 segundos, respectivamente; TNUDCENT=3600 corresponde a un *nudging* muy fuerte. TNUDTOP usualmente no necesita ser usado y sólo debería ser activado en los niveles más altos de la estratósfera. Como en este trabajo el dominio del modelo llega hasta los 20 kilómetros de altura (atmósfera) no será necesaria la experimentación sobre TNUDTOP.

En lo sucesivo se hará referencia a un experimento como la integración del modelo bajo una configuración de prueba particular de TNUDLAT y TNUDCENT dejando al resto de los parámetros del modelo fijos en valores por omisión. El resultado del experimento será entonces el pronóstico de las varibles *u,v, tempc* y *relhum*.

[3]El timestep es una magnitud de tiempo que resulta de integrar el modelo en diferencias finitas con algún esquema de tiempo adecuado.

Capítulo 2

Diseño del experimento y concentración de los resultados

Este capítulo está dividido en dos secciones principales. La primera ha sido llamada *Diseño del experimento* y en ella se describe la forma como fueron seleccionados los valores de prueba para los parámetros. La segunda sección fue titulada *Concentración de los resultados* y en ella se muestra la construcción de un indicador que permite medir el efecto de una configuración particular de los parámetros en la respuesta del modelo.

2.1. Diseño del experimento

Considerando que cada experimento con RAMS (ejecutándose en una supercomputadora Origin 2000) toma alrededor de 7 horas, se decidió tener un diseño inicial de tamaño 13. En estos ensayos se consideró el rango de variación de TNUDLAT y TNUDCENT como (900, 1800) y (7000, 14000) respectivamente, así entonces las configuraciones posibles o espacio de parámetros fue el cuadrado formado por el producto cruz de los rangos de variación. De dicho espacio se extrajo una muestra mediante el **muestreo cuadrado latino** [1] (McKay et al., 1979) (cuadrado ya que el espacio tiene esta forma) que es

[1] En (McKay et al., 1979) se compara al MCL con el muestreo Estratificado y con el muestreo Aleatorio en la solución de un problema similar al que ahora nos ocupa y se encuentra a los primeros dos metodos mejores que el muestreo aleatorio con respecto a la varianza para una clase de estimadores que incluye la media de la muestra y la función de distribución empírica.

Cuadro 2.1: Diseño del experimento, primera parte.

Experimento	(TNUDLAT,	TNUDCENT)
1	(1326,	13833)
2	(1108,	9947)
3	(1464,	11850)
4	(1457,	8610)
5	(1125,	9278)
6	(1777,	9675)
7	(1294,	12287)
8	(950,	7895)
9	(1279,	10358)
10	(1051,	12818)
11	(1197,	8351)
13	(1361,	13146)

una extensión del *muestreo por cuotas* (Arber, 95). La metodología de este muestreo aplicado a este caso es como sigue: se dividió cada rango de variación en 13 intervalos de igual longitud y se obtuvo un valor aleatorio de cada uno de ellos. Así se generó un vector con 13 coordenadas para cada parámetro. Después en cada vector se hizo una permutación de sus componentes. La primera configuración se compone con la primera coordenada de cada vector resultante. La segunda configuración con la segunda coordenada y así sucesivamente. El resultado es un diseño Aleatorio Cuadrado Latino con dos tratamientos y 13 niveles, combinado aleatoriamente de manera que todos los tratamientos ocurren una vez (Cuadro 2.1). El objetivo es obtener una cobertura uniforme (pero no necesariamente esparcida[2]) del espacio de parámetros. Un indicador del éxito en este esquema de muestreo es la baja correlación entre los valores seleccionados de los parámetros. Sin embargo no siempre ocurre esto cuando el procedimiento es aleatorio. En (Iman y Conover, 1982) se describe un método para transformar un MCL en uno con mejores propiedades de correlación.

Una vez obtenidos los resultados de los 13 ensayos, se decidió realizar algunos otros experimentos, pero éstos deberían considerar configuraciones de los parámetros fuera de los rangos sugeridos, así el rango de TNUDLAT fue ampliado a $(100, 4000)$ y el de TNUDCENT a $(200, 20000)$, (Cuadro 2.1). El objetivo de realizar los nuevos experimentos fue saber si el modelo dejaba de ser numéricamente estable. En caso de que esto no ocurriera la cuestión sería saber en qué medida cambiaba la predicción dada

[2]La muestra puede ser o no esparcida, esto depende en gran medida de la forma como se dividan los rangos de variación de los parámetros. Cuando fueron divididos en intervalos de igual longitud se estaba asumiendo que cada parámetro se distribuye uniformemente sobre su rango de variación, pero si en base a la experiencia o a algún hecho se puede asignar alguna otra distribución entonces los intervalos en que se dividan los rangos será de igual probabilidad y esto dará lugar a una muestra uniforme con respecto a la distribución propuesta y no al espacio.

una configuración extrema de los parámetros de interés (extrema ya que sobrepasaba por completo las recomendaciones).

Cuadro 2.2: Diseño del experimento, segunda parte.

Experimento	(TNUDLAT,	TNUDCENT)
14	(300,	900)
15	(4000,	20000)
16	(300,	20000)
17	(1800,	1850)
18	(4000,	10000)
19	(300,	500)
20	(1000,	6000)
21	(100,	200)
22	(400,	12000)
23	(2300,	16000)

En total se realizaron 23 experimentos. En la Figura 2.1 se muestra la distribución de la muestra en el espacio. Los puntos dentro del rectángulo fueron obtenidos por el MCL, mientras que los que están fuera fueron seleccionados secuencialmente.

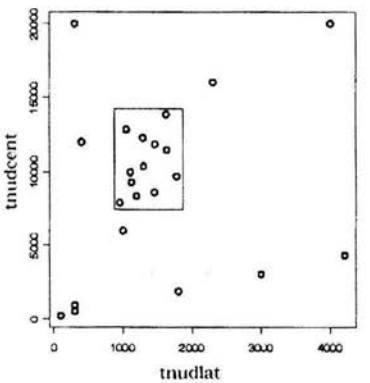

Figura 2.1: Distribución de la muestra.

2.2. Concentración de los resultados

Para evaluar el efecto que causa una configuración particular $s_i = (TNUDLAT_i, TNUDCENT_i)$ en la respuesta del modelo se ha construido un indicador llamado *índice de correlación general*, $I(s_i)$. El sustento de esta medida es la relación que existe entre el pronóstico generado por el modelo bajo una configuración de prueba particular y las observaciones.

2.2.1. Índice de correlación general

En principio se calcula el índice de correlación definido por (2.1) entre las observaciones y la predicción para cada variable (*u, v, tempc, relhum*) y bajo cada una de las 23 configuraciones de prueba.

$$\rho_x^{(s_i)} = \frac{M_{x_o x_p^{(s_i)}}}{\sqrt{M_{x_o x_o} M_{x_p^{(s_i)} x_p^{(s_i)}}}} \tag{2.1}$$

donde

$$M_{x_o x_p^{(s_i)}} = \sum_{j,k,l} (x_{oj,k,l} - \overline{x}_o)(x_p^{(s_i)}{}_{j,k,l} - \overline{x}_p^{(s_i)}),$$

$$M_{x_o x_o} = \sum_{j,k,l} (x_{oj,k,l} - \overline{x}_o)^2,$$

$$M_{x_p^{(s_i)} x_p^{(s_i)}} = \sum_{j,k,l} (x_p^{(s_i)}{}_{j,k,l} - \overline{x}_p^{(s_i)})^2,$$

Cuadro 2.3: Resultados del cálculo del índice de correlación, ρ, entre el pronóstico del modelo y las observaciones para cada una de las variables de estudio (u, v, *tempc* y *relhum*) bajo cada una de las configuraciones de prueba.

Experimento	(TNUDLAT,	TNUDCENT)	$\rho_u X10^2$	$\rho_v X10^2$	$\rho_{tempc} X10^2$	$\rho_{relhum} X10^2$
1	(1626,	13833)	96.0	81.3	99.95	96.8
2	(1108,	9947)	96.2	81.8	99.96	97.1
3	(1464,	11850)	96.1	81.6	99.95	96.9
4	(1457,	8610)	96.2	82.5	99.96	97.2
5	(1125,	9278)	96.2	82.0	99.96	97.2
6	(1777,	9675)	96.1	82.3	99.96	97.0
7	(1294,	12287)	96.1	81.4	99.95	96.9
8	(1636,	11468)	96.1	81.8	99.95	96.9
9	(950,	7895)	96.3	82.3	99.96	97.3
10	(1297,	10358)	96.1	81.8	99.96	97.1
11	(1051,	12818)	96.1	81.1	99.95	97.0
12	(1197,	8351)	96.2	82.4	99.96	97.2
13	(300,	900)	96.2	86.9	**99.98**	**97.8**
14	(4000,	20000)	95.2	80.5	99.93	95.4
15	(300,	20000)	96.3	78.7	99.96	96.9
16	(1800,	1850)	96.5	86.4	**99.98**	**97.8**
17	(301,	500)	96.1	87.3	**99.98**	**97.8**
18	(1000,	6000)	**96.7**	83.1	99.97	97.5
19	(100,	200)	95.9	**87.4**	**99.98**	**97.8**
20	(2300,	16000)	95.8	81.1	99.94	96.2
21	(400,	12000)	96.6	80.2	99.96	97.2
22	(4200,	4300)	96.6	85.0	99.97	97.5
23	(3000,	3020)	96.6	85.7	99.97	97.7

x_o representa la observación y $x_p^{(s_i)}$ la predicción de la variable x dada la configuración de prueba s_i. Los subíndices j y k representan el día y la hora de pronóstico y el subíndice l el sitio (latitud, longitud y altura) donde hay pronóstico y obsevaciones. En el Cuadro 2.3 se muestran los resultados del cálculo del índice de correlación. Los números en negrita son los máximos de la columna en que se localizan, así por ejemplo en el experimento 19 donde TNUDLAT y TNUDCENT están fijos en 100 y 200 segundos respectivamente, se obtiene la correlación más alta para las variables *tempc*, *relhum* y v. Otras observaciones interesantes sobre el cuadro son:

I. Existen cuatro configuraciones bajo las cuales ρ_{tempc} y ρ_{relhum} son mayores, y éstas son las mismas para ambas ((300,900), (1800,1850), (301,500) y (100,200)).

II. Los coeficientes ρ_u y ρ_v son inversamente proporcionales.

III. ρ_v no es mayor que 0.874, a diferencia de ρ_u, ρ_{tempc} y ρ_{relhum} que están por encima de 0.90 para todas las configuraciones de prueba.

Regresando al ensayo 19, éste es propuesto como el de mejores resultados. Es cierto que el valor de ρ_u obtenido no es el máximo, sin embargo, sí es muy alto (0.959). De (II) y (III) es claro que para mejorar el valor de ρ_u, se afectaría ρ_v, hecho que

no resulta factible, de aquí que el experimento 19 es considerado como el de mejores resultados y que la cuarta columna del cuadro no sea considerada, excepto para definir un empate. Ahora se muestra como evaluar al resto de los experimentos siguiendo el mismo razonamiento.

Sea $\overline{\rho_v}$ un vector cuya componente i-ésima es $\rho_v^{(\boldsymbol{s}_i)}$ y sea $\overline{o\rho_v}$ la ordenación descendente de $\overline{\rho_v}$, esto es

$$\overline{\rho_v} = (\rho_v^{(\boldsymbol{s}_1)}, ..., \rho_v^{(\boldsymbol{s}_{23})}),$$

$$\overline{o\rho_v} = (\rho_v^{((\boldsymbol{s}_1))}, ..., \rho_v^{((\boldsymbol{s}_{23}))}).$$

La evaluación que recibirá la configuración $\boldsymbol{s}_i$ para la variable v, denotada como $I(\boldsymbol{s}_i, v)$, será la posición j que ocupe $\rho_v^{(\boldsymbol{s}_i)}$ en $\overline{o\rho_v}$. De manera similar se definen $I(\boldsymbol{s}_i, tempc)$ y $I(\boldsymbol{s}_i, relhum)$. Finalmente

$$I(\boldsymbol{s}_i) = I(\boldsymbol{s}_i, v) + I(\boldsymbol{s}_i, tempc) + I(\boldsymbol{s}_i, relhum).$$

Es importante hacer notar que $I(\boldsymbol{s}_i)$ no es una medida de correlación, sino simplemente una forma de comparar los experimentos. El Cuadro 2.4 muestra los resultados del índice. Esta tabla ha sido ordenada de acuerdo a la última columna. Valores menores para I se asocian a mejores resultados de los experimentos y valores mayores a ensayos cuyos resultados fueron menos satisfactorios.

Con la información del Cuadro 2.4 se ha construido la Figura 2.2. El tamaño de la burbuja depende del valor de I. Algunas observaciones sobre la figura son:

- Las burbujas de mayor tamaño estan en la región de la esquina superior derecha, mientras que en la esquina inferior izquiera se localizan las de menor tamaño.
- El modelo es más sensible ante cambios en el segundo parámetro (TNUDCENT). Esto se debe a que ante variaciones pequeñas con respecto al rango de variación del parámetro el valor de I oscila en un rango mayor. No así en el caso de TNUDLAT, esto resulta claro en particular cuando los valores que toma varían entre 100 y 5000 segundos.

Cuadro 2.4: Evaluación de los experimentos a través del índice de correlación general. Valores pequeños de I se asocian a mejores resultados.

Experimento	(TNUDLAT,	TNUDCENT)	I
19	(100,	200)	3
17	(301,	500)	4
13	(300,	900)	5
16	(1800,	1850)	6
23	(3000,	3020)	9
22	(4200,	4300)	11
18	(1000,	6000)	12
4	(1457,	8610)	16
9	(950,	7895)	17
12	(1197,	8351)	18
5	(1125,	9278)	19
6	(1777,	9675)	20
2	(1108,	9947)	21
10	(1297,	10358)	22
8	(1636,	11468)	24
3	(1464,	11850)	25
21	(400,	12000)	26
11	(1051,	12818)	27
7	(1294,	12287)	27
1	(1626,	13833)	28
15	(300,	20000)	30
20	(2300,	16000)	31
14	(4000,	20000)	34

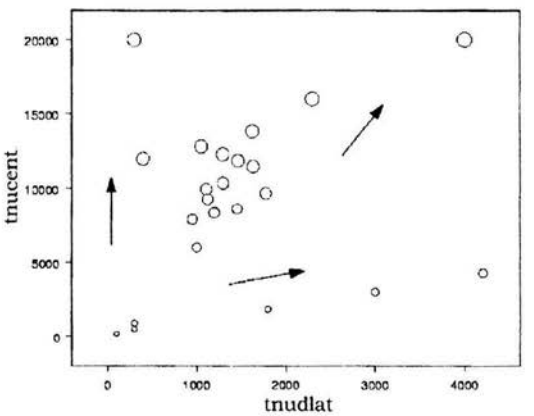

Figura 2.2: Evaluación de los experimentos. El tamaño de la burbuja depende del valor de I.

En la siguiente sección se verá como caracterizar al espacio parametral con la información que se tiene de la muestra. El procedimiento que se ocupa es un método de interpolación espacial que permitirá estimar el valor de I para configuraciones de los parámetros que no fueron probadas. Una vez que se cuente con las estimaciones de un número considerable de puntos en el espacio parametral se podrá inferir con mayor confianza sobre la totalidad del espacio.

Capítulo 3

Análisis estadístico

En este capítulo se muestra la aplicación del método de interpolación llamado *Kriging*, en la estimación de valores de $I(\boldsymbol{s})$ para configuraciones no probadas. La finalidad es tener una retícula de estimaciones que permita la caracterización del comportamiento del modelo en el espacio parametral.

En la primera sección del capítulo se da el contexto teórico del método de *Kriging*, con la intención de facilitar la comprensión de su metodología, la cual es explicada en la segunda sección. En el último apartado del capítulo se muestran los resultados del análisis de *Kriging* para el problema de estudio.

3.1. *Kriging*, un método geoestadístico de interpolación

Kriging es un método de interpolación extraído de la geoestadística. En este método la variabilidad en el espacio del atributo que se quiere estimar es modelada con una superficie estocástica. El atributo es entonces llamado una *variable regionalizada*.

Bajo la *teoría de variables regionalizadas* el valor observado del atributo de estudio en un sitio $\boldsymbol{s}$ es considerado como la realización, $z(\boldsymbol{s})$, de una variable aleatoria $Z(\boldsymbol{s})$. En sitios donde no hay mediciones los valores de $z(\boldsymbol{s})$ están bien definidos aunque no son conocidos, y también se consideran como la realización de la variable aleatoria $Z(\boldsymbol{s})$ correspondiente.

En términos matemáticos a la familia de estas variables aleatorias se le llama una *función aleatoria* y a la realización, $\{z(\boldsymbol{s}) : \boldsymbol{s}\epsilon D\}$, de una función aleatoria, $\{Z(\boldsymbol{s}) : \boldsymbol{s}\epsilon D\}$, se le conoce como una *variable regionalizada*, (con $D \subset R^n$) (Armstrong, 1998). La variable regionalizada posee dos componentes básicos:

- Un componente m local y determinístico que representa la tendencia global de la variable.
- Y un componente aleatorio R que representa la dispersión natural alrededor de m.

En este trabajo el efecto que una configuración $\boldsymbol{s}$ tiene en la respuesta del modelo, medido a través de $I(\boldsymbol{s})$, es tratado como una realización de una función aleatoria $\{I(\boldsymbol{s}); \boldsymbol{s}\epsilon D, D = R_1 X R_2\}$ (con R_1 y R_2 los rangos de variación de $TNUDLAT$ y $TNUDCENT$ respectivamente).

Para poder hacer inferencias estadísticas sobre una única realización de la función es preciso establecer algunos supuestos de estacionariedad en D.

3.1.1. Hipótesis de estacionariedad

La estacionariedad tiene diferentes grados. En la práctica se utilizan supuestos de *segundo orden* y más comúnmente la llamada *hipótesis intrínseca*, términos que son explicados a continuación .

Estacionariedad de segundo orden. Una función aleatoria $Z(\boldsymbol{s})$ es *estacionaria de segundo orden* si satisface:

- su esperanza no depende de $\boldsymbol{s}$. Esto es, para todo $\boldsymbol{s}$ en D,

$$E[Z(\boldsymbol{s})] = m,$$

- la función de covarianza entre dos puntos $\boldsymbol{s}$ y $\boldsymbol{s}+\boldsymbol{h}$ depende únicamente del vector $\boldsymbol{h}$ y no de $\boldsymbol{s}$. Es decir,

$$E[Z(\boldsymbol{s})Z(\boldsymbol{s}+\boldsymbol{h})] - m^2 = C(\boldsymbol{h}).$$

Hipótesis intrínseca. Una función aleatoria $Z(\boldsymbol{s})$ se dice intrínseca cuando los primeros dos momentos de los incrementos $Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h})$ existen y son independientes del punto $\boldsymbol{s}$. Es decir

$$E(Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h})) = 0,$$

$$Var(Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h})) = 2\gamma(\boldsymbol{h}). \tag{3.1}$$

La función $\gamma(\boldsymbol{h})$ es llamada el *variograma*, la herramienta básica para la interpretación estructural del fenómeno.

3.1.2. El variograma

El variograma $\gamma(\boldsymbol{h})$ es una medida de la continuidad espacial de una función aleatoria Z. La figura 3.1 muestra un variograma típico. Usualmente γ se incrementa con $\boldsymbol{h}$, mientras $\boldsymbol{h}$ sea menor que una cierta distancia llamada *rango*, en donde γ toma su valor máximo conocido como *meseta*. Entonces el rango es la distancia a la cual la desviación en valores de Z no depende de la separación entre ellos y de aquí que ya no están correlacionados.

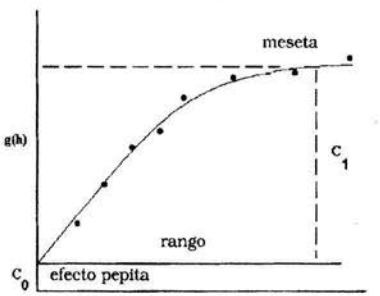

Figura 3.1: Un variograma típico que alcanza un límite, al que se le llama *meseta*, en una cierta distancia conocida como *rango*.

Es claro que $\gamma(0) = 0$, sin embargo algunas veces la presencia de variación a pequeña escala trae como consecuencia que $\gamma(\boldsymbol{h}) \to C_0$ conforme $\boldsymbol{h} \to 0$. A C_0 se le conoce como el *efecto pepita.*

3.1.3. El variograma experimental y el variograma teórico.

Para calcular el variograma experimental a partir de un grupo de observaciones $z(\boldsymbol{s}_1)$, $z(\boldsymbol{s}_2)$,..., $z(\boldsymbol{s}_n)$ de Z, los datos son divididos dentro de intervalos de distancia llamados rezagos (ver Figura 3.2). El número y tamaño de los mismos son especificados por el usuario. Una estimación de γ es hecha en cada rezago. Cada par de puntos separados por un vector $\boldsymbol{h}$, de ángulo $h_1 + \Delta h_1$ y magnitud $h_2 + \Delta h_2$ son usados para estimar el valor de $\gamma(\boldsymbol{h})$ definido por

$$\gamma(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} [Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i + \boldsymbol{h})]^2, \tag{3.2}$$

donde $N(\boldsymbol{h})$ es el número de pares que satisfacen el criterio de selección.

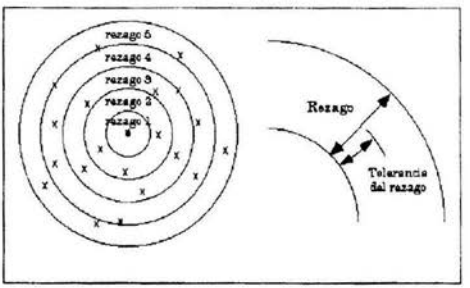

Figura 3.2: División de los datos en intervalos de distancia, llamados rezagos, en el cálculo del variograma experimental omnidireccional.

En la práctica una vez que se tiene el variograma experimental se ajusta a un modelo matemático al que se le llama *variograma teórico.* La razón de esto el poder garantizar que no se llegará a calcular varianzas negativas. El modelo que se ajusta es de la forma $g(\boldsymbol{h}) = C_0 + C_1 f(\boldsymbol{h})$, en donde C_0 es el efecto pepita, $C_0 + C_1$ es la meseta y $f(\boldsymbol{h})$ es alguna función monótona no decreciente en $\boldsymbol{h}$.

Existen varios modelos disponibles que pueden ser ajustados al variograma experimental. Éstos pueden ser categorizados por la presencia o ausencia de meseta y por el comportamiento en el origen. En el Apéndice A. se incluyen las expresiones matemáticas de los modelos más usados así como una descripción breve de ellos, además hay una sección dedicada al análisis del comportamiento del variograma cerca del origen.

3.2. Predicción espacial: *Kriging*

Kriging es un método de interpolación exacta (ya que reproduce los valores observados) en el que el estimador se construye como una combinación lineal de los datos muestrales. Esto es, la predicción de una función aleatoria Z en un punto no muestreado $\boldsymbol{s}_0$, $z^*(\boldsymbol{s}_0)$, dado el conjunto de datos muestrales $\{z(\boldsymbol{s}_1), z(\boldsymbol{s}_2), ..., z(\boldsymbol{s}_N)\}$ es:

$$z^*(\boldsymbol{s}_0) = \sum_{i=1}^{N} \lambda_i z(\boldsymbol{s}_i), \tag{3.3}$$

donde la selección de los pesos o ponderadores λ_i se hace en base a la distancia entre $\boldsymbol{s}_i$ y $\boldsymbol{s}_0$, y a la continuidad espacial presente en los datos.

La ventaja de *Kriging* sobre otros métodos de interpolación es que bajo este procedimiento se obtine un estimador lineal insesgado para el cual la varianza del error de predicción es mínima.

Del método de *Kriging* existen actualmente muchas variantes, por ejemplo simple, ordinario, universal, bayesiano, etc. En todas ellas lo que se busca es minimizar la varianza del error $Z^*(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)$ sujeto a la condición de insesgamiento.

Considérese el caso en el que $Z(\boldsymbol{s})$ es estacionaria con media m desconocida. Si el estimador es de la forma (3.3) la esperanza del error de estimación es

$$E[\sum_{i=1}^{N} \lambda_i Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_0)] = \sum_{i=1}^{N} \lambda_i m - m$$
$$= m[\sum_{i=1}^{N} \lambda_i - 1].$$

Para que el estimador sea insesgado el error esperado debe ser cero, de modo que $m = 0$ o los pesos de *Kriging* deben sumar 1. En el primer caso la media sería conocida (kringing simple), si m es desconocida entonces los pesos deben sumar 1, es decir

$$\sum_{i=1}^{N} \lambda_i = 1. \tag{3.4}$$

Lo que sigue es examinar lo relativo a la condición de varianza mínima. La varianza del error $Z^*(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)$ puede ser expresada en términos del variograma o de la covarianza como

$$\sigma^2 = 2\sum_i \lambda_i \gamma(\boldsymbol{s}_i - \boldsymbol{s}_0) - \sum_i \sum_j \lambda_i \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \tag{3.5}$$
$$= \sum_i \sum_j \lambda_i \lambda_j C(\boldsymbol{s}_i - \boldsymbol{s}_j) - 2\sum_i \lambda_i C(\boldsymbol{s}_i - \boldsymbol{s}_0).$$

El estimador óptimo será el que minimiza la varianza (3.5) sujeta a la condición de insesgamiento (3.4). Esto es

$$\phi = Var[Z^*(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)] - 2\mu(\sum \lambda_i - 1),$$

donde μ es un multiplicador de Lagrange. Después de igualar las derivadas parciales de ϕ a cero se obtiene un conjunto de $N + 1$ ecuaciones lineales conocidas como las

ecuaciones de Kriging ordinario.

$$\sum_j \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) + \mu = \gamma(\boldsymbol{s}_i - \boldsymbol{s}_0); \quad i = 1, 2, \ldots, N$$
$$\sum_i \lambda_i = 1.$$

La *varianza de Kriging ordinario* se define como

$$\sigma_k{}^2 = \sum \lambda_i \gamma(\boldsymbol{s}_i - \boldsymbol{s}_0) + \mu,$$

o equivalentemente usando notación matricial $\boldsymbol{AX} = \boldsymbol{B}$, con

$$\mathbf{A} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1n} & 1 \\ \vdots & \ddots & \vdots & \\ \gamma_{n1} & \cdots & \gamma_{nn} & 1 \\ 1 & \ldots & 1 & 0 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \gamma_{10} \\ \vdots \\ \gamma_{n0} \\ 1 \end{pmatrix}$$

y $\gamma_{ij} = \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j)$.

Si γ es un modelo admisible la matriz $\boldsymbol{A}$ será no singular y en consecuencia su inversa $\boldsymbol{A}^{-1}$ existe. Así que la solución existe y puede probarse que es única. La varianza de *Kriging* se expresa como

$$\sigma_k{}^2 = \boldsymbol{X}^T \boldsymbol{B}.$$

En el Apéndice B se dan los detalles de la estimación del valor de la media, la cual se obtiene de manera similar al procedimiento anterior. En el Apéndice C se estudia el caso del *Kriging* simple, en el que se supone a m conocida.

En la siguiente sección se muestra la aplicación del método de *Kriging* en la caracterización del espacio muestral.

3.3. Resultados de *Kriging* en los experimentos

Para explorar las relaciones espaciales entre los datos se inicia un análisis de variograma. En el Cuadro 3.1 se presentan los resultados del cálculo del variograma experimental. Los datos fueron divididos en 19 rezagos, para cada uno de los cuales una estimación de γ fue hecha. Una observación interesante es que a medida que la distancia se incrementa las estimaciones cambian de manera *suave*, lo cual indica alta continuidad en el espacio de $I(\boldsymbol{s})$. En la columna `np` se han listado el número de pares de puntos que pertenecen al rezago que aparece en la columna `Distancia`, así por ejemplo del primer renglón del cuadro se lee que se encontraron 12 pares de puntos cuya distancia de punto a punto es menor o igual que 536.35. Donde se encontraron más pares fue en el rezago 4, cuando la distancia es menor o igual que 1975.05, (np = 17). Es importante aclarar que en los calculos mencionados antes la dirección no ha sido tomada en consideración, esto se debe a que se tiene una muestra pequeña que proporciona poca información cuando se discrimina con base a la dirección.

Cuadro 3.1: Cálculo del variograma experimental.

	Distancia	γ	np
1	536.35	1.29	12
2	995.08	2.03	13
3	1500.52	4.15	10
4	1975.05	8.00	17
5	2479.28	13.31	11
6	3025.61	27.07	7
7	3554.10	21.33	15
8	4036.01	30.37	12
9	4456.56	27.87	8
10	5067.00	31.35	7
11	5519.84	41.64	7
12	5936.21	52.30	10
13	6471.75	53.18	8
14	6943.46	63.35	7
15	7555.05	89.68	8
16	8008.52	67.96	14
17	8534.35	90.25	8
18	8950.48	102.37	8
19	9498.92	132.75	8

Para el ajuste del variograma teórico se han eliminado las estimaciones de γ en las que $\mathtt{np} < 11$, ya que se asume que no hay suficiente evidencia y se omite el dato (ver Figura 3.3). El modelo ajustado es un Gausiano, y tiene la siguiente forma:

$$\gamma(\boldsymbol{h}) = C_0 + C_1(1 - e^{-(\frac{h}{r})^2}),$$

con r =6100, C_1 =82 y C_0 =0.6,

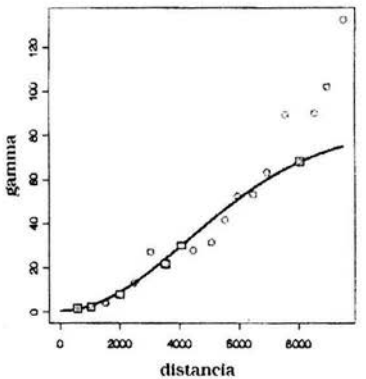

Figura 3.3: Ajuste del modelo gaussiano $\gamma(\boldsymbol{h}) = C_0 + C_1(1 - e^{-(\frac{h}{r})^2})$ con $r = 6100$, $C_1 = 82$ y $C_0 = 0{,}6$, al variograma experimental. Las estimaciones de γ con $np < 11$ fueron descartadas.

donde el parámetro r es el rango, C_1 la meseta y C_0 el efecto pepita. Esto quiere decir que la distancia mínima a la que dos observaciones ya no están correlacionadas es de 6100 unidades, o en otras palabras que el radio de influencia de una observación sobre las estimaciones es de 6100 unidades.

Si se considera que la distancia mayor a la que dos observaciones se encuentran es de 20,396 unidades entonces un rango de 6100 unidades indica continuidad a mediana escala. Esta idea es reforzada al tener que $C_1 = 82$, esto es: a medida que la distancia crece, $\boldsymbol{h} \to \infty$, la diferencia entre los valores tiende a 82, $\gamma(\boldsymbol{h}) \to 82$, que no es un valor grande si se considera que $\gamma(rango)$ =52.6.

Finalmente dado que $C_0 = 0.6$ (efecto pepita[1]) se asume la presencia de variación a pequeña escala de magnitud mínima la cual es casi imperceptible en el gráfico; si la magnitud fuera mayor se notaría al ver que el inicio de la curva no está en el origen sino C_0 unidades desplazada, hacia arriba si $C_0 > 0$ o hacia abajo si $C_0 < 0$.

Una vez definido el variograma teórico, es posible realizar con ayuda de algún paquete estadístico la estimación de $I(\boldsymbol{s})$ en una serie de sitios que cubran uniformemente el espacio parametral. En este caso se realizó la estimación en 900 puntos. Parte de los resultados se muestran en el Cuadro 3.2 en donde se puede observar que en general la

[1]La experiencia ha mostrado que el utilizar un variograma gaussiano considerando el efecto pepita nulo resulta en inestabilidad numiérica.

Cuadro 3.2: Predicción de $I(\boldsymbol{s})$ en una malla regular de 900 puntos sobre el espacio parametral. La tabla es la salida del paquete S-PLUS.

	(TNUDLAT,	TNUDCENT)	I^*	$Var(I^*)$
1	100.00	200	3.00	0
2	241.37	200	3.23	1.29
3	382.75	200	3.43	1.72
4	524.13	200	3.59	2.08
5	665.51	200	3.74	2.40
6	806.89	200	3.86	2.67
7	948.27	200	3.98	2.91
8	1089.65	200	4.09	3.12
9	1231.03	200	4.19	3.30
10	1372.41	200	4.29	3.47
11	1513.79	200	4.39	3.61
12	1655.17	200	4.50	3.75
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
893	3210.34	20000	33.25	3.02
894	3351.72	20000	33.38	2.80
895	3493.10	20000	33.52	2.53
896	3634.48	20000	33.65	2.20
897	3775.86	20000	33.79	1.76
898	3917.24	20000	33.92	1.09
899	4058.62	20000	33.96	0.93
900	4200.00	20000	33.88	1.71

varianza de una estimación es mayor en la medida que ésta dista de un dato muestral. Por ejemplo para el punto $\boldsymbol{s}_{898'}$ =(3917.24, 20 000) (se utiliza la prima para distinguir de los sitios muestrales) la varianza de la estimación $Var(I^*(\boldsymbol{s}_{898'}))$ es de 1.09, mientras que para el punto $\boldsymbol{s}_{899'}$ =(4058.62, 20 000) la varianza es de 0.93. Esto ocurre ya que el segundo está más cerca del sitio muestral $\boldsymbol{s}_{14} = (4000, 20000)$. Estas observaciones son claras en la Figura 3.4(b) que es el mapa de las curvas de nivel de $Var(I^*(\boldsymbol{s}))$. En esta figura la curva de nivel 1, C_1, se localiza donde hay más de dos observaciones (ver Figura 2.1). Es importante señalar que el valor promedio de la varianza de las estimaciones es mínimo.

El hecho de que *Kriging* sea un interpolador exacto puede ser verificado con la estimación y con la varianza de la estimación que se obtuvo en el sitio $\boldsymbol{s}_{1'} = (100, 200)$ que coincide con un sitio muestral (ver Cuadro 2.3).

A continuación se revisa lo que corresponde a las estimaciones. En la Figura 3.4(a) se muestran 6 curvas de nivel de $I^*(\boldsymbol{s})$. Por la forma de éstas es posible observar que cuando $TNUDCENT$ varía entre 5000 y 17,000 segundos. el efecto que tiene en I^* es fuerte y se ve incrementado cuando $TNUDLAT$ es grande. Mientras que el efecto de este último por si sólo es mínimo. Los valores menores de I^* se obtienen cuando $TNUDCENT$ es pequeño, menor que 5000, o en otras palabras, la relación entre

el pronóstico del modelo y las observaciones se ve mejorada a medida que la fuerza del *nudging* (inversa del tiempo de relajación) se incrementa. Es importante aclarar que este resultado debe ser usado considerando también la representatividad de las observaciones[2]. Es quizá debido al hecho de que las observaciones son típicamente muy esparcidas que los autores de RAMS sugieren rangos de variación conservadores para los parámetros.

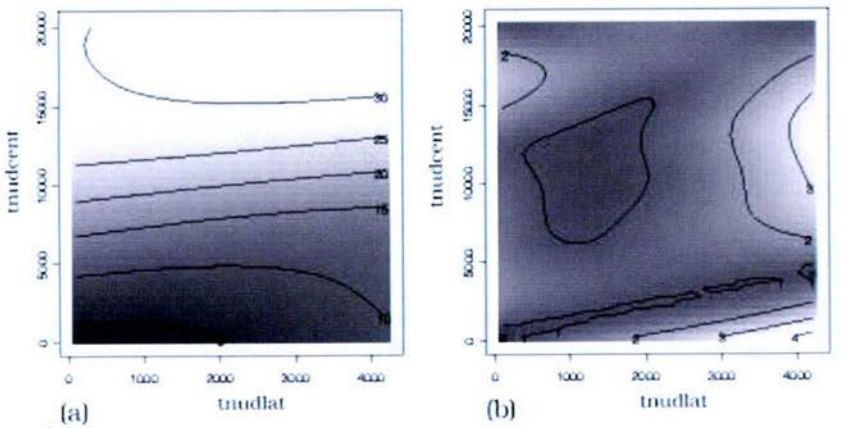

Figura 3.4: (a) Curvas de nivel de la estimación de $I(\boldsymbol{s})$. (b) Curvas de nivel de la varianza de las estimaciones de $I(\boldsymbol{s})$.

[2]Comunicación personal con Craig Tremback.

Conclusiones

I. El efecto del parámetro $TNUDCENT$ en la respuesta del modelo (medida a través de $I(\boldsymbol{s})$) fue más significativo que el que tuvo el parámetro $TNUDLAT$. Lo anterior resulta comprensible si se considera que $TNUDCENT$ es un límite inferior para la fuerza del nudging. En general cuando ambos parámetros tomaban valores menores, la relación entre el pronóstico del modelo y las observaciones era mejor; sin embargo es preciso hacer notar que este resultado debe ser usado en conjunción con consideraciones sobre la representatividad de las observaciones.

II. Se encuentra al muestreo cuadrado latino como un método sencillo de usar, flexible, ya que permite la incorporación de información adicional que se tenga para que la muestra resultante se concentre más en ciertas áreas, y particularmente valioso cuando la dimensión del espacio muestral es mayor que 2 y se tiene un tamaño de muestra pequeño.

III. Sobre el método de interpolación usado, *Kriging*, se considera que tiene dos ventajas fundamentales sobre otros métodos. La primera es que hace consideraciones sobre: la distribución de la muestra, el tamaño de la misma y la continuidad espacial de la variable de estudio. La segunda es que permite estimar la incertidumbre de las estimaciones.

Apéndice A

El variograma

A.1. Definición de variograma

La definición del variograma esta dada por la ecuación 3.1 (página 19), donde los vectores $\boldsymbol{s}$ y $\boldsymbol{s}+\boldsymbol{h}$ son puntos en un espacio de dimensión n; $\boldsymbol{h}$ es un vector, de manera que γ es función de dos componentes h_1 y h_2, con h_1 la magnitud y h_2 el sentido de $\boldsymbol{h}$. Para un ángulo fijo el variograma índica qué tan diferentes se vuelven los valores a medida que la distancia se incrementa.

La Figura 3.1 (página 20) muestra un variograma típico en el que se presentan las siguientes características:

- $\gamma(\boldsymbol{h})$ se incrementa con $\boldsymbol{h}$.
- $\gamma(\boldsymbol{h})$ se incrementa hasta un cierto límite llamado *meseta* y después se estabiliza (aunque podría seguir creciendo).

Las propiedades se ven en detalle a continuación.

A.2. Rango y zona de influencia

La medida en la que el variograma cambia con respecto a la distancia indica qué tan rápido la influencia de la muestra desaparece con la distancia. Una vez que se haya alcanzado la meseta, la correlación entre la muestra habrá desaparecido. Esta distancia crítica, llamada *rango*, proporciona una definición más precisa a la noción de *zona de influencia*.

A.2.1. Comportamiento cerca del origen

Para estudiar la continuidad y regularidad espacial de la variable es importante analizar el comportamiento del variograma en distancias pequeñas. En la Figura A.1 se muestran los siguientes cuatro tipos de comportamiento en el origen.

a) Cuadrático. Éste indica que la variable regionalizada es altamente continua.

b) Lineal. La variable es continua pero no diferenciable por lo tanto es menos regular que en el caso anterior.

c) Discontinuo. Es decir que $\gamma(\boldsymbol{h})$ no tiende a 0 cuando $\boldsymbol{h}$ se aproxima a 0. Lo cual indica que la variable es altamente irregular en distancias pequeñas.

d) Plano. Este es el caso límite en el que hay una carencia de estructura. La correlación entre las variables $z(\boldsymbol{s})$ y $z(\boldsymbol{s}+\boldsymbol{h})$ es nula sin importar qué tan cercanas estén.

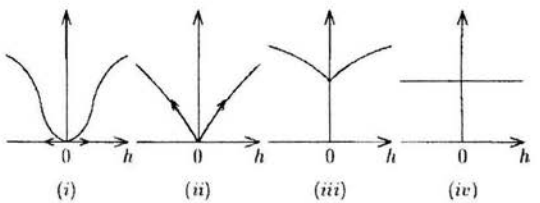

Figura A.1: Comportamiento del variograma cerca del origen. (i) Cuadrático; (ii) Lineal; (iii) Discontinuo; (iv) Plano.

A.3. Modelos teóricos de variograma

Una vez que se tiene un variograma experimental se ajusta a él un modelo matemático. La razón de esto es el poder garantizar que no se llegará a calcular varianzas negativas.

A.3.1. Varianza de combinaciones lineales admisibles

Debido a que los estimadores más comunes son combinaciones lineales de las observaciones es que es necesario poder calcular su varianza. Primero se considera a la variable estacionaria $Z(\boldsymbol{s})$ con covarianza $C(\boldsymbol{h})$. Sea Z^* la combinación lineal

$$Z^* = \sum \lambda_i Z(\boldsymbol{s}_i),$$

donde λ_i son los ponderadores y $\boldsymbol{s}_i$ las ubicaciones de la muestra. Por definición,

$$Var(Z^*) = E(Z^* - E(Z^*))^2.$$

Si m es la media de $Z(\boldsymbol{s})$ entonces $E(Z^*) = m\sum \lambda_i$. De modo que

$$\begin{aligned} Var(Z^*) &= E(\sum \lambda_i (Z(\boldsymbol{s}_i) - m))^2 \\ &= \lambda_1^2 C(\boldsymbol{s}_1 - \boldsymbol{s}_1) + \lambda_2^2 C(\boldsymbol{s}_2 - \boldsymbol{s}_2) + ... + \lambda_n^2 C(\boldsymbol{s}_n - \boldsymbol{s}_n) \\ &+2\lambda_1\lambda_2 C(\boldsymbol{s}_1 - \boldsymbol{s}_2) + ... + 2\lambda_{n-1}\lambda_n C(\boldsymbol{s}_{n-1} - \boldsymbol{s}_n), \end{aligned}$$

por lo tanto

$$Var(Z^*) = \sum_i \sum_j \lambda_i \lambda_j C(\boldsymbol{s}_i - \boldsymbol{s}_j).$$

Para cualesquiera puntos y ponderadores, esta cantidad debe ser no negativa. A una función $C(\boldsymbol{h})$ con esta propiedad se le llama *definida positiva.*

La estimación es un poco diferente cuando la variable es intrínseca en lugar de estacionaria. En este caso la varianza de una combinación lineal podría no existir, pero al menos se asegura que existe para combinaciones lineales de los incrementos. Las combinaciones serán *admisibles* si la suma de los ponderadores es cero, es decir,

$$\sum \lambda_i = 0. \tag{A.1}$$

Ya que la covarianza podría no existir la expresión para la varianza del estimador se da en términos del variograma:

$$Var(\sum \lambda_i Z(\boldsymbol{s}_i)) = -\sum \sum \lambda_i \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j).$$

Como se ha dicho, para que la varianza sea no negativa, los modelos de variograma deben satisfacer ciertas condiciones. Para cualquier conjunto de puntos $\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_k$ y cualquier grupo de ponderadores $\lambda_1, \lambda_2, ..., \lambda_k$, que satisfacen (A.1), es preciso que

$$-\sum\sum \lambda_i \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \geqslant 0,$$

por lo tanto γ es definida positiva condicionalmente. Esta condición es más débil que la encontrada sobre las funciones de covarianza por lo tanto la clase de modelos de variograma admisibles es mayor que la de covarianzas.

A.3.2. Modelos más frecuentes

Debido a que es difícil reconocer a las funciones que satisfacen las propiedades señaladas en la sección anterior es que existe una serie de modelos de variograma que pueden ser usados, éstos serán explicados a continuación.

a) Efecto pepita puro

$$\gamma(\boldsymbol{h}) = \begin{cases} 0 & \boldsymbol{h} = 0 \\ \mathrm{C} & |\boldsymbol{h}| > 0. \end{cases}$$

Este modelo se asocia a fenómenos puramente aleatorios.

b) Esférico

$$\gamma(\boldsymbol{h}) = \begin{cases} \mathrm{C}(\frac{3|\boldsymbol{h}|}{2a} - \frac{1}{2}(\frac{|\boldsymbol{h}|^3}{a^3})) & |\boldsymbol{h}| < a \\ \mathrm{C} & |\boldsymbol{h}| \geqslant a. \end{cases}$$

Este modelo es de los más usados. En el inicio crece casi linealmente y en una cierta distancia a se estabiliza.

c) Exponencial

$$\gamma(\boldsymbol{h}) = \mathrm{C}\ (1 - \exp(-\tfrac{|\boldsymbol{h}|}{a})).$$

El rango del modelo es prácticamente $3a$, ya que es ésta la distancia a la que alcanza el 95 % de su valor límite.

d) Potencias

$$\gamma(\boldsymbol{h}) = \mathrm{C}\ |\boldsymbol{h}|^{\alpha}, \text{ con } 0 < \alpha \leqslant 2.$$

El modelo líneal, $\gamma(\boldsymbol{h}) = |\boldsymbol{h}|$, es un caso particular de éste.

e) Gaussiano

$$\gamma(\boldsymbol{h}) = \mathrm{C}\ (1 - \exp(-\tfrac{|\boldsymbol{h}|^2}{a^2})).$$

El modelo gaussiano representa un fenómeno altamente continuo. Según algunos autores la experiencia ha mostrado que si se le usa con un efecto pepita nulo muy probablemente se tendrán problemas de inestabilidad numérica.

f) Cúbico

$$\gamma(\boldsymbol{h}) = \begin{cases} \mathrm{C}(7r^2 - 8{,}75r^3 + 3{,}5r^5 - 0{,}75r^7) & r < 1 \\ \mathrm{C} & r \geqslant 1 \end{cases}$$

donde $r = \frac{\boldsymbol{h}}{a}$. Éste es parabólico en el origen y usualmente similar al modelo gaussiano.

A.4. Anisotropía

Cuando se calcula el variograma en distintas direcciones a veces ocurre que se comporta diferente en algunas de ellas, es decir la variable de estudio es *anisotrópica*. Si esto no ocurre el variograma depende únicamente de la magnitud de la distancia entre los dos puntos y se dice entonces que la variable es *isotrópica*. La anisotropía debe ser representada por una sola función variograma, es decir, debe ser llevada al caso isotrópico, lo que implica hacer una transformación lineal de coordenadas rectangulares o ser expresadas de forma individual para cada dirección según sea el caso.

Apéndice B

Kriging ordinario; estimación del valor de la media

En la Sección 3.1 se mostró cómo estimar el valor de una variable regionalizada $Z(\boldsymbol{s})$ estacionaria con media m desconocida. En este apartado se explica cómo estimar m. Como antes el estimador m^* se construye como una combinación lineal de los datos muestrales, esto es

$$m^* = \sum \lambda_{m_i} Z(\boldsymbol{s}_i),$$

además m^* debe ser insesgado y de varianza mínima. Esto es,

$$E(m^* - m) = E(\sum \lambda_{m_i} Z(\boldsymbol{s}_i) - m) = 0,$$

ya que $E(Z(\boldsymbol{s})) = m$, entonces

$$\sum \lambda_{m_i} = 1.$$

La varianza del error del estimador esta dada por

$$\begin{aligned} Var(m^* - m) &= Var(\sum \lambda_{m_i} Z(\boldsymbol{s}_i) - m) \\ &= \sum\sum \lambda_{m_i}\lambda_{m_j} C(\boldsymbol{s}_i - \boldsymbol{s}_j). \end{aligned}$$

Como se hizo anteriormente la varianza se minimiza con la restricción de los ponderadores con un multiplicador de Lagrange. De modo que las ecuaciones resultantes son

$$\begin{aligned} \sum_j \lambda_{m_j} C(\boldsymbol{s}_i - \boldsymbol{s}_j) &= \mu_m; \quad 1, 2, \ldots, N \\ \sum \lambda_{m_j} &= 1. \end{aligned}$$

La varianza de *Kriging* se calcula como

$$\sigma_m^2 = Var(m^*) = \mu_m.$$

Apéndice C

Kriging simple

A diferencia del *Kriging* ordinario en el *Kriging* simple se supone que la media m de la variable regionalizada $Z(\boldsymbol{s})$ es conocida. En principio se considera a la variable regionalizada $Y(\boldsymbol{s})$ con media cero, de forma que $Z(\boldsymbol{s}) = Y(\boldsymbol{s}) + m$. El estimador de $Y(\boldsymbol{s})$ esta dado por

$$Y^*(\boldsymbol{s}_0) = \sum_i \lambda_i' Y(\boldsymbol{s}).$$

Se usan primas para distinguir estos pesos de los de *Kriging* ordinario y de los usados en la estimación de la media.

Como antes el estimador debe ser insesgado y de varianza mínima. Sobre la condición de insesgamiento hay que decir que automáticamente se satisface ya que la media de $Y(\boldsymbol{s})$ es cero. Esto es,

$$E(Y^*(\boldsymbol{s}_0) - Y(\boldsymbol{s}_0)) = E(\sum_i \lambda_i' Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_0)) = 0,$$

así que no hay restricción sobre la suma de los pesos. La varianza de la estimación del error se expresa como

$$\begin{aligned} Var(Y^*(\boldsymbol{s}_0) - Y(\boldsymbol{s}_0)) &= E(\sum_i \lambda_i' Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_0))^2 \\ &= \sum_j \sum_i \lambda_j' \lambda_i' C(\boldsymbol{s}_i - \boldsymbol{s}_j) - 2 \sum_i \lambda_i' C(\boldsymbol{s}_i - \boldsymbol{s}_0). \end{aligned}$$

Por lo anterior el sistema de *Kriging* es

$$\sum_j \lambda_j' C(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(\boldsymbol{s}_i - \boldsymbol{s}_0); \quad 1, 2, \ldots, N, \tag{C.1}$$

y la varianza correspondiente esta dada por

$$\sigma_{sk}^2 = - \sum \lambda_i' C(\boldsymbol{s}_i - \boldsymbol{s}_0).$$

Una vez que se resuelve el sistema C.1 se encuentran los ponderadores y el estimador de $Z(\boldsymbol{s}_0)$ al reemplazar $Y(\boldsymbol{s})$ por $Z(\boldsymbol{s}_0) - m$. Es decir

$$\begin{aligned} Z^*(\boldsymbol{s}_0) &= Z^*(\boldsymbol{s}_0) + m \\ &= \sum \lambda_i' (Z(\boldsymbol{s}_i) - m) + m \\ &= \sum \lambda_i' Z(\boldsymbol{s}_i) + m(1 - \sum \lambda_i') \\ &= \sum \lambda_i' Z(\boldsymbol{s}_i) + m\lambda_M. \end{aligned}$$

Bibliografía

Armstrong, Margaret, 1998. *Basic linear geostatistics.* Ed. Springer.

Arber, Sara. 1995. *Designing samples.* Researching Social Life. Editado por N. Gilbert. London.

Bowman, R., J. Sacks y Y.,-F. Chang, 1993. *Design and analysis of numerical experiments.* J. Atmos. Sci., 50, 1267-1278.

Chapman, W., W.J. Welch, R.P. Bowman, J. Sacks y J.E. Walsh, 1994. *Artic sea ice variability model sensitivities and multidecadal simulation.* J.Geohys. Research, 99, C1, 919-935.

Cotton, W.R. et al., 1982. *The Colorado State University three-dimensional cloud mesoscale model* . J. Rech. Atmos., 16, 295-320.

Cox, M. y R. Bryan, 1984. *A numerical model of the ventilated thermocline.* J. Phys. Oceanogr., 14, 674-687.

R.Daley y Kamal Puri, 1980. *Four-dimensional data assimilation and the slow manifold.* Monthly Weather Review, 108, 85-99.

Dudhia J., D. Gill et al., 2003. *PSU/NCAR Mesoscale modeling system. Tutorial Class Notes and User's Guide: MM5 Modeling System Version 3.* Mesoscale and Microscale Meteorology Division, National Center for Atmos. Research, E.U.A.

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

M. Ghil, 1989. *Metereological data assimilation for oceanographers, Part I: Description and Theorical Framework.* Dynamics of Atmospheres and Oceans, 13, 171-218.

Pielke, R.A., 1984. *Mesoscale meteorological modeling.* New York, N.Y.: Academic Press.

Pielke, R.A. et al., 1992. *A comprenhesive metereological modeling system - RAMS.* Meteor. Atmos. Phys., 49, 69-91.

Sacks, J., S.B. Schiller, y W.J. Welch, 1989. *Design and analysis of computer experiments (with discussion).* Stat. Sci., 4, 409-435.

Scire, J.S., F.R. Robe, M.E. Fernau y R.J. Yamartino, 2000. *A user's guide for the CALMET meteorological model (Version 5).* Earth Tech, Inc., Concord, MA 01742.

Steinberg, H.A. 1963. *Generalized quota sampling.* Nuc. Sci. and Engr., 15, 142-145.

Smedstad, Fox, 1994. *Assimilation of altimeter data in a two-layer primitive equation model of the gulf stream.* J. Phys. Oceanogr., 24, 305-325.

Walko, R.L., C.J. Tremback, y R.F.A. Hertenstein, 1995. *The regional atmospheric modeling system. Version 3b User's Guide.* ASTER, Inc. P.O. Box 466, Ft Collins, CO, 117 pp.

Walko et al., 1999. *RAMS, regional atmospheric modeling system.* Technical Description.