

03061

2



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

**UNIDAD ACADÉMICA DE LOS CICLOS PROFESIONAL
Y DE POSGRADO
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS
APLICADAS Y EN SISTEMAS**

**ANÁLISIS ESPACIAL DE EXPERIMENTOS DE
CAMPO, UNA SIMULACIÓN**

T E S I S

**QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS EN
ESTADÍSTICA E INVESTIGACIÓN DE OPERACIONES
P R E S E N T A:
JOSÉ ALEJANDRO REYES CÁMARA**



Cd. Universitaria, D. F.

282462

Septiembre 2000

La presente tesis titulada: **Análisis Espacial de Experimentos de Campo, Una Simulación**, realizada por el tesista José Alejandro Reyes Cámara, bajo la dirección de los sinodales del jurado indicados, ha sido aprobada por los mismos y aceptada como requisito para obtener el grado de MAESTRO EN CIENCIAS EN ESTADÍSTICA E INVESTIGACIÓN DE OPERACIONES:

SINODALES

Vocal: **Dr. Ignacio Méndez Ramírez**

Presidente: **Dr. Federico O'Realy Togno**

Secretario: **M. en C. Patricia Isabel Romero Mares**

Suplente: **M. en C. José Salvador Zamora Muñoz**

Suplente: **Dr. Carlos Díaz Ávalos**

Ciudad Universitaria, D. F. en septiembre del 2000

A black and white photograph of a park path. The path is paved and leads into a wooded area. On the right side of the path, there is a wooden bench. The trees are tall and leafy, creating a canopy over the path. The lighting is bright, suggesting a sunny day. The overall mood is peaceful and contemplative.

Dios:

En los momentos difíciles de la vida,
a no olvidar lo que me ha pasado,

a esperar sabiendo que nada me faltará,

a no olvidarme de mí,

a ver sin ojos,

a gritar sin ser escuchado,

a estar acompañado en la soledad,

a sonreír en la tristeza,

a dejar de amar lo que se ama,

a llorar en silencio.

AGRADECIMIENTOS

Al Dr. Ignacio Méndez Ramírez por su calidez humana y valioso tiempo aportado para la dirección de la presente tesis.

Al Dr. Federico O'Reilly Togno, M. en C. Patricia Isabel Romero Mares, M. en C. José Salvador Zamora Muñoz y Dr. Carlos Días Ávalos por sus sugerencias y observaciones para el mejoramiento del presente escrito.

Profesores y personal del IIMAS del Departamento de Probabilidad y Estadística, Personal Administrativo del IIMAS y a las personas del Proyecto Académico de la Maestría en Estadística e Investigación, por sus facilidades brindadas durante mi estancia como estudiante y académico.

A todos los miembros, presentes y ausente, de la Familia Galilei Cervantes por su invaluable apoyo.

Dra. Sonia Fernández Cantón, Dr. Luis Manuel Torres Palacios y Dr. Miguel Angel Lezana Fernández, que me brindaron su apoyo incondicional durante mi estancia laboral en la Dirección General de Estadística de la SSA, que me permitió tomar la decisión de realizar esta tesis.

Finalmente, a todos aquellos que me faltaron y que voluntaria o involuntariamente me han apoyado.

DEDICATORIA

A DIOS todopoderoso que ha sido el más grande consuelo en innumerables momentos de mi vida.

A mis PADRES, Elsy Soledad Cámara Escamilla y Diego Reyes Pérez, que me dieron la vida, amor y apoyo incondicional y hacen hasta lo imposible para que sea un hombre capaz y digno.

A mis HERMANOS, Ernesto Javier, Julissa, Héctor, William y Martha Cecilia que han hecho un gran esfuerzo por apoyarme y motivarme para seguir adelante

A mis AMIGOS Virginia del Carmen Ramírez Alcocer, Marco Antonio Rendón Teran, Lorena Ramírez Alcocer, familia Bravo Ramírez, Antonio González Hernández y Vicente Bracamonte Carrillo, todos ellos por su invaluable apoyo, amistad y paciencia brindada. También a la niña Sofía Victoria Rendón Ramírez por su sonrisa y alegría por la vida.

CONTENIDO

Índice de Cuadros.....	v
Índice de Figuras.....	vi
I. Introducción.....	1
II. Revisión Bibliográfica.....	5
2.1. Antecedentes.....	5
2.2. Supuestos del Modelo.....	16
2.3. Especificaciones.....	17
III. Objetivos.....	20
3.1. General.....	20
3.2. Específicos.....	20
IV. Marco Teórico.....	21
4.1. Antecedentes Estadísticos Para Datos Espaciales.....	21
4.1.1. Entropía y Modelo Para Diferentes Datos.....	21
4.1.2. Datos y Modelos Espaciales.....	23
4.1.2.1. El modelo general espacial.....	24
4.1.3. Principales Enfoques del Análisis Espacial.....	25
4.1.3.1. Datos geoestadísticos.....	25
4.1.3.2. Datos de látice.....	27
4.1.3.3. Datos en patrones de puntos.....	28
4.1.4. Razón de la Estadística Para Datos Espaciales.....	29
4.1.4.1. Estimación, predicción y datos para diseños experimentales.....	29
4.1.4.2. Modelos lineales con errores espacialmente dependientes.....	35

4.1.4.3. Validez del modelo espacial.....	38
4.1.4.4. Estimación de mínimos cuadrados generalizados.....	39
4.2. Geoestadística.....	40
4.2.1. Origen.....	40
4.2.2. Análisis Espacial de Datos.....	41
4.2.2.1. Estacionariedad intrínseca.....	41
4.2.2.2. Descomposición de los datos en variación de escala grande y escala pequeña.....	42
4.2.3. Procesos Estacionarios.....	43
4.2.3.1. Estacionariedad de segundo orden y ergodicidad.....	44
4.2.4. Variograma y efecto nugget.....	44
4.2.4.1. Algunos modelos de variogramas isotropicos y proceso isotropico.....	45
4.2.5. Covariograma y Correlograma.....	46
4.2.6. Estimación del Variograma.....	47
4.2.7. Representación Espectral.....	48
4.2.8. Ajuste de los Modelos de Variograma.....	49
4.3. Predicción Espacial y Kriging.....	49
4.3.1. El Mejor Predictor Lineal.....	51
4.3.2. Escala de Variación Observacional y Espacial.....	52
4.3.2.1. Suposiciones paramétricas de la función de medias.....	53
4.3.3. Kriging Ordinario.....	54
4.3.3.1. Predicción espacial óptima de el proceso Z.....	55
4.3.3.2. Kriging en términos de la función de covarianza.....	57
4.3.4. Kriging Robusto.....	58
4.3.4.1. Estimación del variograma.....	58
4.3.5. Kriging Universal.....	59
4.3.5.1. Supuestos del modelo.....	59
4.3.5.2. Predicción espacial óptima de el proceso Z para las ecuaciones de kriging.....	60
4.3.5.3. Estimación del variograma para kriging universal.....	62

4.3.5.4. Kriging para bloque.....	63
4.3.5.5. Tendencia polinomial.....	64
4.3.5.6. Estimación de la Media Parametral.....	64
4.3.6. Predicción de Superficie de Tendencia.....	65
4.3.6.1. Predicción de la superficie de tendencia y predicción del cuadrado medio del error.....	66
4.3.7. Kriging de Mediana Destacada(Polish).....	66
4.3.7.1. Análisis de medias.....	67
4.3.7.2. Algoritmo de la mediana destacada (polish).....	68
4.3.7.3. Kriging basado en residuales.....	70
4.3.8. Datos Geoestadísticos Simulados y Reales.....	70
4.3.8.1. Simulación de procesos espaciales.....	71
4.4. Diseños Espaciales.....	71
4.4.1. Diseño de Muestreo Espacial.....	72
4.4.1.1. Criterio.....	73
4.4.1.2. Planes de muestreo espacial.....	75
4.4.1.3. Diseño espacial experimental en regiones discretas.....	75
4.4.2. Diseños Experimentales Optimos en Presencia de Dependencia Espacial.....	76
4.4.2.1. La función de covarianza y el variograma en ieseños experimentales.....	78
4.4.2.2. Aleatorización para el diseño de bloques completos.....	79
4.4.2.3. Eficiencia de los diseños de bloques con errores en un proceso <i>AR</i> de segundo orden.....	82
4.4.2.4. Diseño de bloques balanceados de vecinos más cercanos de primer orden.....	85
4.4.2.5. Ensayos de campo.....	85
4.4.3. Análisis de Vecinos Más Cercanos.....	86
4.4.3.1. Estimación de efectos de tratamiento.....	86
4.4.3.2. Caso de una dimensión y caso de dos dimensiones.....	87
4.4.3.3. Estimación de los parámetros y de variación de escala pequeña.....	89
4.4.4. Análisis Basado en Modelación Espacial.....	90

V.	Materiales y Métodos.....	91
5.1.	Descripción de los Datos Experimentales Utilizados.....	91
5.2.	Metodología de Análisis Espacial Alternativo.....	94
5.2.1	Evaluación de Cinco Propuestas de Análisis Mediante Simulación.....	96
5.2.2	Descripción del Modelo Base.....	96
5.2.3	Descripción de la Generación de Datos.....	98
5.2.4	Los Tratamientos.....	98
5.3.	Métodos de Refinamiento a los Bloques.....	100
5.3.1	Bloques al Azar.....	100
5.3.1.1	Estadísticas de evaluación del modelo.....	101
5.3.2	Bloques con Covariable de Posición.....	102
5.3.3	Papadakis Dentro de Bloque.....	103
5.4	Métodos que Ignoran Bloques.....	105
5.4.1	Papadakis Clásico en Dos Dimensiones.....	105
5.4.2	Polinomio en Coordenadas de Posición.....	107
VI.	Resultados.....	109
6.1.	Estadísticas Básicas de Cada Conjunto de Datos a Analizar.....	109
6.2.	Relación entre la Varianza de los Errores Estimados Obtenido de la Simulación con la del Segmento del Ensayo de Uniformidad.....	110
6.3.	Comportamiento de la Significancia del Modelo de Cada Método.....	113
6.4.	Potencia Empírica Mediante la Diferencia de Tratamientos.....	117
6.5.	Sesgo Empírico de la Diferencia de Tratamientos.....	123
6.6.	Error Cuadrático Medio Empírico de la Diferencia de Tratamientos.....	132
VII.	Conclusiones.....	141
VIII.	Bibliografía.....	144
IX.	Anexo (CD).....	157

ÍNDICE DE CUADROS

Número	Descripción	Pág.
Cuadro 4.1.	Análisis de varianza para el diseño de bloques completos aleatorizados.....	81
Cuadro 5.1.	Características de los conjuntos de experimentos en blanco.....	93
Cuadro 5.2.	Nuevas características de los conjuntos de experimentos en blanco.....	94
Cuadro 5.3.	Representación del número total de simulaciones en los ensayos en blanco.....	100
Cuadro 6.1.	Algunas estadísticas básicas de los conjuntos de datos elegidos.....	109
Cuadro 6.2.	Algunas estadísticas básicas de los conjuntos de datos compactados.....	110
Cuadro 6.3.	Algunas estadísticas básicas de los conjuntos de datos de la muestra.....	110
Cuadro 6.4.	Cuadro de resultados, Y_{ij}^* , con dos criterios de clasificación para los datos de Baker <i>et al.</i> (1952).....	111
Cuadro 6.5.	Estadísticas del modelo y significancia del factor Método para los conjuntos de ensayos de uniformidad.....	112
Cuadro 6.6.	Prueba de Tukey para la respuesta media en el factor Método de Análisis en cinco Fuentes de Datos.....	113
Cuadro 6.7.	Representación del PPS para el Método de Bloques al Azar en el segmento del ensayo de uniformidad de Baker <i>et al.</i> (1952).....	121
Cuadro 6.8.	Sesgo Empírico y sus tres criterios de clasificación (Immer, 1932).....	126
Cuadro 6.9.	Estadísticas del modelo (6.4) y significancia de la interacción para los conjuntos de datos.....	127
Cuadro 6.10.	Prueba de Tukey para el Sesgo Empírico de los cinco Métodos por cada uno de los niveles de Δ para los datos de Baker <i>et al.</i> (1952).....	128
Cuadro 6.11.	Prueba de Tukey para el Sesgo Empírico en $\Delta_1 = 0.00mu$ con los 5 métodos para 4 fuentes de datos.....	129
Cuadro 6.12.	Estadísticas del modelo (6.5) y significancia de dos interacciones para los conjuntos de datos.....	133
Cuadro 6.13.	Prueba de Tukey para el Log($ECM_{Empírico}$) en $\Delta_1 = 0.00mu$ o $Diferencia_E=E12$ para los cinco métodos en tres fuentes de datos.....	136

ÍNDICE DE FIGURAS

Número	Descripción	Pág.
Figura 5.1.-	Ubicación general de la estructura de diseño y variables de posición en el segmento central del ensayo.....	98
Figura 5.2.-	Representación del cálculo del Índice de Papadakis por Bloque.....	104
Figura 5.3.-	Representación del cálculo del Índice de Papadakis en 2 Dimensiones a partir de residuos.....	106-107
Figura 6.1.	Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Beker <i>et al.</i> (1952). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.....	114
Figura 6.2.	Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Immer (1932). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.....	114
Figura 6.3.	Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Lander <i>et al.</i> (1938). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.....	115
Figura 6.4.	Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Mercer y Hall (1911). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.....	115
Figura 6.5.	Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Wassom y Kalton (1953). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.....	116
Figura 6.6.	Gráfica de dispersión para la media $Prob[Fo \geq F] = \hat{p}$ según Δ de los Métodos de Análisis para los datos de Baker <i>et al.</i> (1952).....	116
Figura 6.7.	Gráfica de dispersión para la media $Prob[Fo \geq F] = \hat{p}$ según Δ de los Métodos de Análisis para los datos de Immer (1932).....	118
Figura 6.8.	Gráfica de dispersión para la media de $Prob[Fo \geq F] = \hat{p}$ por Δ de los Métodos de Análisis para los datos de Lander <i>et al.</i> (1938).....	118
Figura 6.9.	Gráfica de dispersión para la media de $Prob[Fo \geq F] = \hat{p}$ por Δ de los Métodos de Análisis para los datos de Mercer y Hall (1911).....	119
Figura 6.10.	Gráfica de dispersión para la media $Prob[Fo \geq F] = \hat{p}$ por Δ de los Métodos de Análisis para los datos de Wasom y Kalton (1953).....	119

Figura 6.11. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Baker et al. (1952).....	122
Figura 6.12. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Immer (1932).....	122
Figura 6.13. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Lander <i>et al.</i> (1938).....	124
Figura 6.14. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Mercer y Hall (1911).....	124
Figura 6.15. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Wassom y Kalton (1953).....	125
Figura 6.16. Gráfica de dispersión para la media del Sesgo Empírico en la interacción <i>Metodox</i> Δ para los datos de Baker <i>et al.</i> (1952).....	128
Figura 6.17. Gráfica de dispersión para la media del Sesgo Empírico en la interacción <i>Metodox</i> Δ para los datos de Immer (1932).....	130
Figura 6.18. Gráfica de dispersión para la media del Sesgo Empírico en la interacción <i>Metodo</i> \times Δ para los datos de Lander <i>et al.</i> (1938).....	130
Figura 6.19. Gráfica de dispersión para la media del Sesgo Empírico en la interacción <i>Metodo</i> \times Δ para los datos de Mercer y Hall (1911).....	131
Figura 6.20. Gráfica de dispersión para la media del Sesgo Empírico en la interacción <i>Metodo</i> \times Δ para los datos de Wassom y Kalton (1953).....	131
Figura 6.21. Gráfica de dispersión para la media del Log(ECM Empírico) en la interacción <i>Metodo</i> \times Δ para los datos de Lander <i>et al.</i> (1938).....	135
Figura 6.22. Gráfica de dispersión para la media del Log(ECM Empírico) en la interacción <i>MetodoxDiferenciaE</i> para los datos de Lander <i>et al.</i> (1938).....	137
Figura 6.23. Gráfica de dispersión para la media del Log(ECM Empírico) en la interacción <i>Metodox</i> Δ para los datos de Mercer y Hall (1911).....	137
Figura 6.24. Gráfica de dispersión para la media del Log(ECM Empírico) en la interacción <i>Metodo</i> \times Δ para los datos de Wassom y Kalton (1953).....	138

I.

INTRODUCCIÓN

El presente trabajo es un estudio de simulación de algunas alternativas de análisis de diseños experimentales agrícolas cuyos tratamientos están sometidos a efectos de tendencia, principalmente ambientales, las cuales son no controlables debido a que su existencia no se observa a simple vista.

La metodología clásica de diseño y análisis de experimentos utilizan algunos diseños (Bloques al Azar, Cuadro Latino y Látices y otros diseños de bloques) que incluyen en el modelo de análisis parte de la variabilidad espacial, pero no se modela explícitamente la mayor parte de la variabilidad restante debido a que un modelo de análisis es por lo general sólo una aproximación útil y no una descripción matemática exacta de los datos con ubicación geográfica.

La experimentación agrícola, industrial y social inició alrededor de la década de los 20, cuando Fisher propone la técnica de análisis de varianza para estudiar los resultados de experimentos. La metodología aplicada a la experimentación, bajo condiciones no homogéneas, incluye: la aleatorización, la repetición de los tratamientos y el uso de bloques en el diseño; y las covariables para controlar las fuentes de variación que no son claramente obvias o que aparecen de modo repentino durante el desarrollo del experimento y no fueron consideradas en el diseño experimental inicial. En la experimentación agrícola de campo, las principales fuentes de variación son: la heterogeneidad del suelo, plagas, enfermedades, vientos, etcétera; dentro de la experimentación en ciencias sociales se puede mencionar a el estado de ánimo, información inadecuada por ignorancia, mentira, etc.; y en la experimentación industrial se puede mencionar a las variaciones en composición de las

materias primas, estado anímico de los operadores de un proceso, materia prima, desgaste de los equipos, etc. La presencia de la heterogeneidad de las unidades experimentales dentro de los bloques es una situación bastante común en la mayoría de los experimentos que se realizan.

En el proceso de experimentación actual, la aleatorización y el agrupamiento de las unidades experimentales en bloques se realiza con la supuesta finalidad de controlar la mayor parte de la heterogeneidad debida a condiciones variantes mencionadas anteriormente, las cuales en su mayoría, no son obvias a simple vista. Lo que se hace es estructurar cada bloque con un grupo de unidades experimentales donde se espera una respuesta medible más homogénea, con lo que las diferencias grandes en la respuesta este entre bloques y no dentro de bloques.

Cabe mencionar que los objetivos de la aleatorización son el de impartir propiedades de independencia a los errores, aproximar el modelo normal infinito a las pruebas paramétricas y producir estimaciones insesgadas de los efectos de tratamiento considerando todas las posibles aleatorizaciones en el diseño experimental. Por lo anterior, la heterogeneidad que se presenta entre las unidades experimentales dentro de bloques es, supuestamente, eliminada mediante el uso de la aleatorización, sin embargo, esta circunstancia se refiere sólo a insesgamiento de los efectos de tratamiento considerando las posibles aleatorizaciones, así, para un experimento específico se pueden afectar las estimaciones de los efectos de tratamientos, adicional a que una parte considerable de heterogeneidad quede dentro de bloques produciendo una sobre estimación de la varianza del error, debido a que la suposición de igualdad de efectos de bloques para las unidades de un mismo es falsa porque se incrementa la variabilidad de las observaciones dentro del bloque y como consecuencia se incrementa el error.

La sobreestimación de la varianza del error provoca que el diseño resulte ineficiente cuando se presenten unidades experimentales heterogéneas dentro de bloques. La consecuencia inmediata es que en las pruebas de detección de diferencias entre tratamientos

exista una falta de potencia, lo cual se puede subsanar con un mayor número de repeticiones, con su respectivo costo, para detectar diferencias que de otro modo podrían detectarse con menos repeticiones de cada tratamiento.

La forma de detectar el efecto de fuentes de variación no controlables en agricultura es por medio de ensayos (experimentos) de uniformidad, también llamados ensayos en blanco, los cuales han indicado que la heterogeneidad en la respuesta de las unidades experimentales no sigue un patrón definido o regular, donde podrían localizarse los bloques. Por lo que para una localización adecuada de los bloques se necesitaría conocer las tendencias de producción y que éstas fuesen lo suficientemente regulares para poder ubicar en forma eficiente los bloques. Esta situación, por lo general, no se presenta en la práctica, siendo cualquiera de las condiciones siguientes: forma del terreno, pendiente, etc., las que determinan la localización de los bloques. Por tal razón es muy probable que se tenga mucha variación dentro de los bloques; para intentar disminuir los inconvenientes de esta situación se puede reducir el tamaño del bloque para que sean más homogéneas las unidades experimentales dentro de los bloques. Aún así, existe la posible heterogeneidad entre unidades experimentales dentro de bloques en los casos en que no se conozca las tendencias de producción del terreno o que se observen, pero que sean muy complejas, lo cual es lo más frecuente en la práctica.

Una manera ideal de evitar estos problemas es el empleo de métodos de experimentación que no utilicen bloques, pero con eliminación en la respuesta obtenida en cada unidad experimental de la fracción correspondiente a factores ambientales sistemáticos. Para lograrlo se tiene que tener en cuenta que el rendimiento de las unidades experimentales en un experimento está compuesto de tres partes principales: i.- un efecto de tratamiento, que es el que se quiere estudiar, ii.- un efecto de productividad de cada unidad experimental en función de las características sistemáticas de su localización geográfica y iii.- un efecto aleatorio (estimable), independientemente del tratamiento aplicado a ella y de las características de su localización geográfica.

El efecto de la productividad (intrínseca) de la parcela como función de su ubicación geográfica y de otros factores sistemáticos se puede entender cuando se obtiene una respuesta de experimentos realizados en el contexto de los ensayos de uniformidad, es decir, las respuestas de las unidades experimentales están compuestas de dos partes: una parte que tenga variación sistemática continua, la cual pueda expresarse por una función matemática de la posición y además los efectos aleatorios con esperanza cero. El efecto de posición reflejará el patrón de variabilidad de la unidad experimental y de otros factores que afectan el experimento en forma sistemática; a la función que representa estos efectos, se le llama función de tendencia o función de tendencia geográfica de productividad, la cual ha sido utilizada también en Geografía y Geología.

Para lo anterior se plantea:

1. La revisión de métodos para representar la función de tendencia.
2. Simulación a partir de ensayos en blanco, para determinar las propiedades empíricas de los métodos que se utilicen y si es posible recomendar aquellos cuyas propiedades sean superiores.

II.

REVISIÓN BIBLIOGRAFICA

2.1. Antecedentes

La heterogeneidad del ambiente donde se realiza un experimento es generalmente la mayor causa de variación en los datos de respuesta de la unidad experimental y de su interpretación. Si está presente un alto grado de variabilidad en un ambiente donde se experimenta, se debe encontrar un método para controlarla. El control de la variabilidad experimental es el principal objetivo de un buen diseño experimental. Sin embargo, el control adecuado de la variabilidad ambiental puede requerir diseños experimentales complejos que pueden ser difíciles de implementar, debido a restricciones en la aleatorización o por que se requiere conocer de modo sencillo cómo realizar su análisis.

A la forma en que sistemáticamente cambian los valores esperados de la variable de respuesta, en función de su localización geográfica o por efecto ambiental se le llama *Tendencia*. Su estudio se puede hacer mediante el *Análisis Espacial*, el que se entenderá como todas aquellas formas de modelar la tendencia que no ha sido considerada con el diseño y el análisis del experimento de forma clásica. El estudio de la tendencia es un problema antiguo que data desde principios del siglo, cuando Fisher y “Student”, iniciaron su análisis (De Vos, 1991 y Jansen, 1993) y que ha tomado importancia en las últimas décadas.

Fisher (1970), desde mediados de la década de los 20 presenta el diseño de bloques al azar y la aleatorización con el propósito de controlar gradientes de fertilidad comúnmente encontrados en los experimentos de campo. Por su parte Yates(1936), plantea el uso de los

diseños de bloques incompletos; también los látices ordinarios o látices generalizados son presentados como una alternativa para eliminar el sesgo originado por la tendencia en las estimaciones, sin embargo esta tendencia puede ser en ocasiones muy variable, lo que haría necesario bloques con pocas unidades experimentales para poder eliminarla.

Papadakis (1937) propone la metodología de ajustar el rendimiento de un experimento de campo, para disminuir los efectos de tendencia local, por medio de un análisis de covarianza con respecto a los rendimientos de tratamiento corregidos a partir de parcelas adjuntas, es decir, propone el uso de la información de las parcelas vecinas como covariable para modelar la tendencia; Bartlett (1938) presenta un análisis y discusión sobre éste método y Atkinson (1969), después de un largo período de tiempo, compara la estimación de los efectos de tratamiento usando residuos de unidades experimentales adjuntas como una variable adicional para los métodos de máxima verosimilitud y Papadakis, observándose que los estimadores y sus propiedades respectivas, son muy cercanas; cabe mencionar que las unidades experimentales del diseño están alineadas en una sola dirección (una dimensión) y se supone la existencia de correlación entre rendimientos de unidades experimentales sucesivas.

Méndez (1970, 1971a y 1972) realizó un estudio con 34 ensayos de uniformidad probando tres modelos para representar la variabilidad geográfica (funciones de tendencia) de la producción de las unidades experimentales. Se encontró como mejor modelo al llamado 'método de selección de variables ortogonales' el cual está basado en el ajuste de un polinomio de grado 8 en ambas coordenadas, usando polinomios ortogonales, seguido de una selección de los términos que explican mayor cantidad de variación en la respuesta de las parcelas. Cuando se comparó las funciones de tendencia se encontró que los bloques son inadecuados en la mayoría de los casos para explicar la productividad de los suelos donde se usan parcelas medianas o grandes de diversas especies de cultivo; así mismo se encontró que si hay una marcada tendencia de producción en franjas, los bloques dan una buena representación de la productividad cuando son colocados en forma perpendicular a los gradientes de producción, de modo que coincidan con franjas de similar productividad. Así

mismo, realizó una simulación de experimentos agrícolas con 48 parcelas en 4 cultivos agregando efectos de tratamiento y errores normales a los datos de 4 ensayos de uniformidad. Cada conjunto de datos generado fue analizado por 6 métodos alternativos de diseño que no usan bloques y por 2 tipos de bloques al azar (con bloques de 8x1 o 4x2) considerando 8 tratamientos y 6 repeticiones de cada tratamiento. En la simulación hubo 2 métodos alternativos: Papadakis y Tendencias en Residuos, que resultaron con buen comportamiento. Se confirmó que el diseño de bloques al azar incorpora parte de la variación en productividad de las parcelas al error experimental, debido a que se encontró una sobreestimación de la varianza del error, con su respectivo efecto sobre las pruebas de hipótesis, al disminuir la potencia de la prueba y un error tipo II más probable. Lo último implica una menor información de los bloques al azar o diseños basados en bloques con respecto al uso del diseño completamente al azar analizado por el método de Papadakis o el de Tendencias en Residuos, por lo que con un menor número de repeticiones en estos métodos se obtiene la misma precisión comparado con los métodos basados en bloques. En México se han realizado algunos ensayos prácticos y relevantes en el campo agrícola para estimar los efectos ambientales sistemáticos mediante promedios móviles (Méndez, 1971b), polinomios y series de Fourier (Zárate y Méndez, 1972) o combinación de polinomios y series de Fourier (Méndez y Rivera, 1976) y el planteamiento de modelos para el análisis de un grupo de datos experimentales (Trujillo, 1994 y Trujillo *et al.*, 1996).

Pearce y Moore (1976) en un estudio para controlar la variación ambiental utilizaron la metodología propuesta por Papadakis y con algunas modificaciones en las cuales usaron de dos a ocho unidades experimentales vecinas con simple y doble covarianza, independientemente de la estructura de bloqueo. Encontraron que el sistema de bloqueo frecuentemente no es útil para reducir el error experimental y la doble covarianza fue superior a la simple.

Bartlett (1978) reexamina el método de Papadakis, propuesto en 1937, en una y dos dimensiones. Obtiene una ganancia de eficiencia, con respecto al análisis de bloques aleatorizados, cuando el número de tratamientos es considerablemente grande mediante un

análisis iterativo; también se reduce la discrepancia entre la aparente precisión de la suma de los cuadrados de los residuales y la verdadera precisión de comparación de tratamientos después del ajuste.

Warren y Méndez (1981) estudian la sensibilidad del diseño al tamaño y orientación de bloques. Se encontraron ensayos insensibles y sensibles, en éstos últimos se observó un aumento considerable del error experimental, cuando los bloques son largos o impropriamente orientados. Así mismo, proponen rutinas de análisis que dan lugar a estadísticas para señalar la posible existencia de bloques inadecuados y su posible análisis correctivo. Así, Warren y Méndez (1982) encontraron que el método de Dos Mejores Bloques o Bloques Dummy fue más útil que el método de Regresión Polinomial, el método de Papadakis Modificado fue prometedor. Los métodos que no fueron efectivos para considerar la variación ambiental o para llevar a cabo un análisis correctivo fueron: Hileras y Columnas, Papadakis Ordinario, Análisis de Fourier y Polinomios Inversos. En este mismo contexto Modjeska y Rawlings (1983) encontraron que el análisis de correlación espacial proporciona indicios de la naturaleza y orientación de los patrones de comportamiento micro ambiental que es útil en la determinación de la estructura óptima de bloqueo.

Martin (1982) al utilizar el método de Papadakis demuestra que tiene solución sencilla; también presenta la relación entre los estimadores teóricos de Papadakis con los estimadores de mínimos cuadrados generalizados para un diseño experimental; por su parte Wilkinson *et al.* (1983) muestran que el método no iterativo de Papadakis es conservativamente sesgado, que el método no iterativo (Bartlett, 1938) tiene un sesgo sustancialmente positivo en la razón de F y que el método es muy eficiente cuando hay efectos notables de tendencia en los datos. Además, propone un modelo de análisis de parcelas adjuntas, el cual es esencialmente un "bloques móviles", análogo a la forma de análisis clásica de bloques "fijos" o hileras y columnas; la cual es más eficiente en promedio en el análisis clásico de experimentos de bloques completos o incompletos, y también más eficiente que el análisis estándar para cuadro latino o látice, cuando hay una clara

interacción entre hileras y columnas en los datos. La validez de la propuesta lo verifican empíricamente mediante simulación.

Williams (1985 y 1986) propone un modelo del tipo de parcelas vecinas alternativo al propuesto por Wilkinson *et al.* (1983) para el análisis de una gran variedad de ensayos de campo. El modelo supone que las variedades son agrupadas en repeticiones completas y que las parcelas dentro de las repeticiones son lado por lado; así mismo, elabora criterios para la construcción de los diseños de vecinos óptimos; Ball *et al.* (1993) identifican la presencia, magnitud y patrones de la heterogeneidad espacial como elementos para asegurar la utilidad del análisis de varianza, para ello estimaron los efectos de la heterogeneidad espacial sobre la independencia de los errores y realizaron la evaluación del análisis de un método de vecinos más cercanos para los diferentes experimentos analizados. Por su parte Stroup *et al.* (1994) compararon el análisis de bloques completos al azar con dos métodos de ajuste de vecinos más cercanos y también realizaron el análisis por el método de campo aleatorio para remover la variabilidad espacial. El enfoque de vecinos más cercanos fue superior para todas las localidades del experimento, en base a un menor coeficiente de variación y gran habilidad para distinguir la diferencia entre los tratamientos.

Otra alternativa para detectar la estructura de la variabilidad espacial es el uso del variograma y Kriging, los cuales tienen su base en técnicas geoestadísticas. Tabor *et al.* (1985) presentan un análisis de la variabilidad de una variable química de suelo agrícola y su relación con otros grupos de variables de respuesta de cultivos.

Por su parte, Green (1985) propone un método de suavizamiento para la estimación de efectos de tratamiento en el contexto de los métodos espaciales. Selecciona la forma y grado de suavizamiento requerido por medio de criterios de validación cruzada. Lo anterior da lugar a una propuesta de análisis de experimentos de campo mediante mínimos cuadrados suavizados (Green *et al.*, 1985) en el cual asume un modelo con tendencia suavizada mas error independiente de cada unidad experimental; se estima efectos de tratamiento y tendencia para los casos unidimensionales y bidimensionales. Continuando con

este enfoque, Clarke *et al.* (1994) utilizan el método de análisis de media móvil y mínimos cuadrados suavizados para la estimación y ajustes de las tendencias de fertilidad del suelo donde se desarrollan experimentos en amplias áreas, encontrando que ambos métodos son útiles para remover la tendencia de fertilidad, mejora de la precisión de la comparación de tratamientos, lo cual lo confirman (Clarke y Baker, 1996) cuando comparan el análisis espacial de mínimos cuadrados suavizados y el de bloques aleatorizados y observan que ambos reducen considerablemente el error estándar de las diferencias para las respuestas medidas.

Los estimadores unidimensionales y bidimensionales del método de Papadakis son revisados y comparados con estimadores de máxima verosimilitud por Draper y Faraggi (1985), que encontraron que la solución por máxima verosimilitud es adecuada cuando existen estructuras de covarianza.

Modelar la tendencia como un proceso autoregresivo de orden bajo, es uno de los diferentes enfoques del análisis de parcelas vecinas que han sido utilizados por Besag y Kempton (1986). Gleeson y Cullis (1987) proponen usar los residuos de máxima verosimilitud para representar la tendencia de errores expresada como proceso $ARIMA(p, d, q)$. Por su parte Basu y Reinsel (1994) consideran la utilización de modelos de regresión para datos espaciales que son medidos a partir de un arreglo rectangular en dos dimensiones, junto con otras variables explicativas; los errores son correlacionados y se considera que siguen un modelo espacial unilateral $ARMA$ de orden uno. Dan detalle del cálculo de los estimadores obtenidos por mínimos cuadrados generalizados en presencia de errores espacialmente correlacionados y comparan estos estimadores con los obtenidos por mínimos cuadrados ordinarios. Consideran que los estimadores de máxima verosimilitud restringida son mejores, para el modelo de correlación espacial comparado con los estimadores de máxima verosimilitud. Por su parte Martin (1995), considera que la aleatorización de los tratamientos es adecuada cuando los errores son espacialmente dependientes y el análisis usual es mínimos cuadrados generalizados, para una estructura de dependencia conocida.

Gilmour *et al.* (1997) identificaron los tres mayores componentes de la variación espacial en los errores de las unidades experimentales de experimentos de campo y extendió los procedimientos espaciales de dos dimensiones de Cullis y Gleeson (1991). Los componentes son: la no estacionariedad, variación de gran-escala (tendencia global) a través del campo con variación estacionaria dentro del ensayo (variación natural o tendencia local) y variación extraña la cual es frecuentemente inducida por procedimientos experimentales y es predominantemente alineado con hileras y columnas. También presentan una estrategia para identificar a modelos para los errores de las unidades experimentales que usan parcelas delimitadas, gráficas en perspectiva del variograma muestral y donde fue posible, pruebas de proporción de la verosimilitud para identificar que componentes están presentes. Concluyen que no hay un modelo que ajuste todos los experimentos de campo, aunque el modelo autorregresivo separable es predominantemente mejor y hay frecuentemente variación identificable adicional. Por su parte Lill *et al.* (1988) demuestran que los estimadores de máxima verosimilitud residual es aceptable para la estimación de los errores de los parámetros en un modelo de orden bajo propuesto por Gleeson y Cullis (1987).

Martin (1990) presenta los modelos con dominio en el tiempo y considera la generalización de los modelos en una dimensión a dos dimensiones mediante el uso de látices separables; continuando esta idea, Cullis y Gleeson (1991) generalizan el modelo propuesto por Gleeson y Cullis (1987), de una a dos dimensiones, usando un proceso para la subclase de látices separables para modelar los errores, considerándolo como un proceso $ARIMA(p_1, d_1, q_1) * ARIMA(p_2, d_2, q_2)$, con $ARIMA(p_1, d_1, q_1)$ en el sentido de las columnas, $ARIMA(p_2, d_2, q_2)$ en el sentido de las hileras y los valores $p_1, q_1, p_2,$ y q_2 podrían ser pequeños, usualmente 0 o 1 y ocasionalmente 2. Por su parte, Gilmour *et al.* (1995) estiman componentes de varianza por medio de máxima verosimilitud restringida en modelos lineales mixtos por medio de la aplicación del algoritmo para el análisis de diseños experimentales para bloques incompletos y de análisis espacial por medio de procesos ARIMA separables y de orden bajo. Así, Cullis *et al.* (1998) realiza análisis espacial de variedades de cultivos diferentes etapas de acuerdo a una estructura de covarianza general o también llamado procesos autorregresivos de orden bajo, $AR(1) \times AR(1)$, para los errores de cada ensayo.

Obtienen los mejores predictores lineales e insesgados de los efectos investigados y los estimadores de máxima verosimilitud de los residuos de los parámetros espaciales y componentes de varianza. Así mismo, indican que se debe hacer más uso de los análisis espaciales y de bloques incompletos, ya que estos reconocen y remueven pérdidas potenciales en la eficiencia comparado con los análisis clásicos; además se requiere de herramientas de diagnóstico más factibles para determinar las estructuras espaciales apropiadas para los datos.

Zimmerman y Harville (1989) utilizaron varios métodos espaciales para el análisis de resultados de experimentos de campo; obtienen estimadores de contrastes de tratamientos, los cuales son por lo general funciones no lineales de las observaciones, también obtienen condiciones suficientes para que los estimadores sean insesgados bajo los postulados del verdadero modelo. Posteriormente aplican (Zimmerman y Harville, 1991) el enfoque de campo aleatorio, similar al análisis geoestadístico de kriging, en la cual las observaciones en su conjunto son consideradas como una realización parcial de un campo aleatorio y con su modelación respectiva en forma directa. Así mismo, comparan su método propuesto con el enfoque de vecinos cercanos y observan que es más adecuado para estimar contrastes de tratamiento.

Otra propuesta de análisis espacial es presentada por Grondona y Cressie (1991) en la cual se obtienen estimadores de la respuesta más eficientes, comparado con el enfoque clásico para la estimación de los contrastes de tratamiento. Se basaron en mínimos cuadrados generalizados empíricos, en la cual, los parámetros de dependencia espacial son estimados a partir de un ajuste de los datos de respuesta. En este mismo contexto, Taam y Hamada (1993) proponen métodos para detectar efectos espaciales de experimentos, a partir de respuesta binaria.

Un método que obtiene las predicciones de vecinos cercanos a partir del mejor predictor lineal insesgado es desarrollado por Stroup y Muiltze (1991), para lo cual consideran varios métodos alternativos de análisis espacial mediante simulación donde

comparan variantes del método de vecinos cercanos con el análisis convencional de vecinos cercanos ajustados y no vecinos cercanos.

Con la finalidad de determinar si el análisis de promedios móviles reduce la varianza del error Dier *et al.* (1991) utilizaron un análisis de covarianza, en la cual la covariable para cada parcela fue la media de las parcelas adjuntas y/o vecinas determinada por cuatro métodos, encuentran que el análisis de promedios móviles reduce la varianza del error comparado con las medias no ajustadas.

Baird y Mead (1991) estudiaron el fenómeno bajo cuatro diferentes modelos de respuesta mediante simulación, los modelos fueron: dos variantes de modelos de vecinos cercanos, modelo de la primera diferencia (Besag y Kempton, 1986) y el modelo extendido de error dentro de las variables; de éste último, los parámetros fueron estimados usando máxima verosimilitud restringida. La eficiencia de los modelos de vecinos cercanos fueron comparados con los análisis clásicos de bloques aleatorizados y bloques incompletos. Además encontraron que el análisis de vecinos es más eficiente que el análisis clásico donde el modelo contiene componentes de tendencia; cuando el análisis de bloques incompletos, es mejor con respecto al análisis de bloques completos, se encontró que el modelo de vecinos es mucho más eficiente.

Méndez (1992) presenta un estudio donde prueba doce modelos para representar la variación geográfica y encontró que los mejores métodos, tanto en propiedades de los errores como distribución geográfica, son los polinomios como covariables y el método de Papadakis, los cuales reducen la magnitud del error o tienen buenas propiedades. Otra aplicación de los métodos de análisis espacial (de dos dimensiones mediante proceso ARIMA separable) es realizada por Grondona *et al.* (1996), quienes utilizan varias alternativas de los modelos propuestos por Cullis y Gleeson (1991).

Otro enfoque de análisis espacial es presentado por Lefkovich (1992), quien propone que la descomposición en valores singulares de la matriz de residuos determina la posible

dependencia y propone una corrección sobre las observaciones para la eliminación de dicha dependencia.

Los métodos de ajuste de la respuesta para modelar la heterogeneidad espacial en dos dimensiones fueron utilizados por Kempton *et al.* (1994), quienes utilizaron los métodos de análisis de hileras y columnas y el de vecinos en su primera diferencia; se observó una disminución considerable de la varianza comparado con el análisis de modelos unidimensionales.

La utilización de la metodología de variables espacialmente dependientes mediante la metodología de kriging, como alternativa para modelar la variabilidad espacial en experimentación se observa en Marx y Thompson (1987), quienes publican algunos aspectos prácticos de la utilización de kriging en agricultura. Así mismo, presentan el análisis de la variabilidad espacial utilizando tanto “métodos clásicos” (bloques, vecinos cercanos, etc.) como geoestadísticos (kriging) y su combinación; encontrando que las medias de tratamiento no son correctamente estimadas al utilizar métodos clásicos, los errores estándar de la diferencia entre tratamientos es menor con el método de mínimos cuadrados generalizados y por último, encontraron que el bloqueo no fue efectivo para controlar la variabilidad espacial.

Laslett (1994) afirma que una muestra espaciada de fenómenos espaciales produce conocimiento insuficiente de una fuente de variación, en la cual, la predicción de las partes no observadas son hechas; el método estocástico probabilístico de kriging para datos espaciales resuelve el problema, aunque con dudas. Para ello se han planteado alternativas como regresión no paramétrica y splines (metodología de estimación de superficies mediante métodos numéricos). La determinación de qué método es mejor debe basarse en resultados operacionales, ya que los métodos son basados en diferentes modelos para los procesos y la precisión y predicción basado en datos reales. Para el análisis de información dividió en conjunto de datos en: el conjunto del modelo y el conjunto de predicción. En los casos analizados, kriging superó a splines por un considerable margen. Varias configuraciones de

los datos muestran que el régimen de muestreo determina cuando kriging podría superar el comportamiento de la técnica de splines.

Gotway y Stroup (1997) plantean un conjunto de herramientas para el análisis de datos no normales, a partir de la teoría de modelos lineales generalizados y casi-verosimilitud, para demostrar cómo esta teoría se puede extender para incluir el análisis de datos discretos y espacial categóricos y también para estimar parámetros y hacer pruebas para los efectos de tratamientos en diseños experimentales que incluyen respuestas categóricas o categóricas espaciales, también proporcionan un método para predicción espacial usando datos no normales; incluyen a kriging universal y kriging indicador como casos especiales.

Federer y Basford (1991) proponen diseños experimentales del tipo de vecinos más cercanos y de efectos de competencia con arreglo espacial para contabilizar la posible existencia de correlación entre unidades experimentales adjuntas, por su parte Martin (1997), considera el proceso de análisis de vecinos más cercanos de tres dimensiones y muestra que las correlaciones son difícil de obtener y que es posible que su valor disminuya ligeramente. Por su parte Azaïs y Druilhet (1997) estudian las propiedades óptimas de diseños de vecinos balanceados cuando los efectos de vecinos no son tomados en cuenta en el análisis del modelo y consideran dos tipos de diseños para los tratamientos, diseños con $t-I$ bloques de longitud t y diseños con t bloques de longitud $t-I$; cuando los diseños no son aleatorizados proponen la utilización de un nuevo criterio en el sesgo matricial y cuando hay aleatorización, el sesgo y la varianza son usados para la comparación, en ambos casos el diseño de vecinos balanceados demostró ser óptimo. Así, Azaïs *et al.* (1998) investigan la validez y eficiencia de algunos métodos de análisis vecinos y su relación con el diseño y encuentran que los métodos de vecinos son frecuentemente válidos o conservadores con alta frecuencia comparado con los métodos clásicos de análisis. Para asegurar lo anterior se utiliza el diseño de vecinos aleatorizados y corrección de grados de libertad con ciertas restricciones debido a que no siempre el método de análisis clásico es más eficiente. Por su parte Wu *et al.* (1998) realizan análisis de vecinos más cercanos y geoestadísticos para

determinar la validez, eficiencia y la extensión de la estructura de variación espacial en experimentos; encontraron que cuando dicha estructura esta presente, el análisis de vecinos puede ser superior al análisis clásico y cuando no hay presencia de estructura de variación espacial, un tipo de análisis de vecinos más cercanos no causa pérdida de eficiencia. Casler (1999) evalúa y compara los diseños de bloques completos al azar, el diseño de látice, análisis de tendencia y análisis de vecinos más cercanos en cultivos de seis especies de pasto de las cuales se cultivan 27 variedades durante quince años; encuentra que el análisis de tendencia o análisis de vecinos más cercanos es útil para contabilizar la variabilidad intrabloque en campo cuando no se conoce o no se puede predecir los patrones de bloque adecuado. Por su parte Helms *et al.* (1999), llevan a cabo un análisis de ajuste de vecinos más cercanos para residuos espacialmente correlacionados para incrementar la precisión. Usaron, alternativo a la magnitud del cuadrado medio del error para la interacción bloque x tratamiento, un método para evaluar la precisión del análisis de vecinos más cercanos y el diseño clásico de no ajuste de bloques completamente aleatorizados por el uso del variograma entre tratamientos duplicados dentro de cada bloque.

2.2. Supuestos del Modelo

Cuando se formula un modelo lineal de diseños de experimentos y para analizar la información experimental, por lo general se tiene en cuenta los supuestos siguientes:

- La forma del modelo es correcta.
- La variable de respuesta es medida sin errores humanos.
- El valor esperado de los errores es cero.
- La varianza de los errores es constante.
- Los errores son independientes.
- Los errores son normalmente distribuidos (cuando se prueba hipótesis).

Cuando el modelo satisface todos los supuestos, se obtienen los mejores estimadores lineales insesgados para los parámetros de interés en el modelo. Cuando se cumplen sucede que:

- Las diferencias de medias de tratamiento o contrastes tienden a tener distribución muestral normal, la cual requieren las pruebas de hipótesis.
- Las sumas de cuadrados se distribuyen como una distribución *Ji-cuadrada* bajo la hipótesis nula.
- Los cocientes apropiados de sumas de cuadrados tienen distribución de *F* bajo algunas hipótesis.

Por lo que cuando no se cumplen los supuestos, los resultados no son válidos debido principalmente a que no son robustas las pruebas de hipótesis y conclusiones a que se puedan llegar.

El no considerar en el análisis general, la variabilidad espacial tiene principalmente las consecuencias (que pueden disminuirse mediante el uso de las metodologías de análisis espacial) siguientes:

- Causa sesgo en las estimaciones de los parámetros de interés.
- Incrementa la variación de los errores.
- Distorsiona la distribución de los errores.
- Provoca falta de independencia de los errores, al pasar parte de la tendencia al error.

2.3. Especificaciones

Dado que en la experimentación de campo se tiene como objetivo la estimación insesgada y eficiente de contrastes de efectos tratamientos, también es de interés probar los efectos de dos o más tratamientos, por lo cual se supone que la variable de respuesta bajo estudio puede ser representada satisfactoriamente por el modelo lineal (se supone que no existe interacción entre el efecto de tendencia y el efecto de tratamiento) siguiente:

$$\text{Variable de Respuesta} = \text{Efecto de Tratamiento} + \text{Tendencia} + \text{Error} \quad (2.1)$$

$$Y_{ij} = \mu + \tau_{ij} + \theta_j + \varepsilon_{ij},$$

donde

- Y_{ij} : es la variable de respuesta observable,
- μ : es la media general,
- τ_{ij} : es el efecto de tratamiento en un sitio particular; se define como el valor esperado de la variable de respuesta menos la parte de la tendencia que modifica la respuesta. Sea Y_{ij} el valor observado en la respuesta para el tratamiento i en un sitio particular j , entonces $E(Y_{ij}) = \mu + \tau_{ij} + \theta_j$, de donde $\tau_{ij} = E(Y_{ij}) - \theta_j - \mu$,
- θ_j : es la tendencia, es un componente sistemático inherente a cada parcela u unidad experimental y depende de la posición de la parcela en relación al patrón de gradientes de fertilidad del suelo, humedad, manto freático captación de luz, cambios topográficos, manejo del experimento, plagas, enfermedades y otros factores ambientales que afectan en forma sistemática el experimento. Por el supuesto de no interacción, no depende del tratamiento aplicado al sitio j .
- ε_{ij} : es el error, es una componente aleatoria que no está relacionada con la posición de la parcela y se supone que el conjunto de errores no están correlacionados y provienen de una distribución normal con media cero y varianza constante.

La tendencia puede clasificarse en tendencia global y tendencia local. La tendencia global, se puede considerar a las fuentes de variación de escala grande, las cuales pueden ser determinada antes del establecimiento del experimento de acuerdo a la posible existencia (obvia) de canales de riego, grandes cambios topográficos, etc. Este tipo de tendencia puede ser modelada mediante el uso de diseños experimentales, bloques completos o bloques incompletos, los cuales consideran a *priori* heterogeneidad espacial, pero suponen homogeneidad dentro del bloque; si este supuesto no se cumple, los errores pueden presentar correlación, se incrementa la varianza del error y puede haber sesgo en la estimación de los efectos de tratamientos. Como la tendencia local se puede considerar a las fuentes de variación de escala pequeña de las cuales no se tiene un conocimiento claro y previo al establecimiento del experimento, como son los pequeños cambios de gradientes de fertilidad, distribución no homogénea de luz por las nubes o sombra, etc. En la metodología clásica de diseño y análisis de experimentos, la tendencia local no es modelada, con la justificación de que la aleatorización da validez a las estimaciones de efecto fijos y

aleatorios. El propósito de modelar la tendencia local (en forma más eficiente, en términos del mejor ajuste y con menor número de parámetros) es:

- Reducir o controlar la variabilidad residual.
- Remover los efectos de asociación entre unidades experimentales de los contrastes de tratamientos.
- Lograr que las propiedades distribucionales de los residuos sean más cercanos a los supuestos de normalidad, homogeneidad de varianza e independencia.

III.

OBJETIVOS

3.1. General

Perfeccionar las técnicas de experimentación y análisis que se basan en bloques y también aplicar las técnicas de compensación por las funciones de tendencias basadas en la localización de las parcelas.

3.2. Específicos

- Estudio del comportamiento de algunos métodos propuestos en experimentos de campo y su comparación con los diseños actualmente utilizados. Considerando experimentos hechos.

- Realizar estudios de simulación donde se incluyan: una muestra aleatoria de ensayos de uniformidad, mayor variación en los efectos de tratamientos incluyendo el caso de tratamientos iguales e incluir métodos de análisis de residuos.

- Revisar los métodos de análisis espacial como marco teórico de referencia, aun que no todos se usan directamente en la simulación.

IV.

MARCO TEÓRICO

4.1. Antecedentes Estadísticos Para Datos Espaciales

4.1.1. Entropía y Modelo Para Diferentes Datos

El nivel de desorden de un fenómeno se ha medido mediante una cantidad llamada *entropía*. Suponga que hay $i = 1, \dots, k$ posibles estados de la naturaleza que pueden ocurrir en forma aleatoria de acuerdo a una distribución de probabilidades (p_1, p_2, \dots, p_k) , donde $0 \leq p_i \leq 1$, $i = 1, \dots, k$, y $p_1 + p_2 + \dots + p_k = 1$. Por lo que la entropía es:

$$E \equiv -\sum_{i=1}^k p_i \cdot \log p_i, \quad (4.1.1)$$

donde $x \cdot \log x \equiv 0$, para $x = 0$. Cuando el estado de la naturaleza i tiene una interpretación numérica, la media $\mu \equiv \sum_{i=1}^k i \cdot p_i$ podría considerarse como el centro de la distribución y la varianza $\sigma^2 \equiv \sum_{i=1}^k (i - \mu)^2 p_i$ como la cantidad de variabilidad en la distribución. Es la σ^2 , la cantidad que se utiliza para cuantificar el desorden, pero no debe sorprender que E y σ^2 estén cercanamente relacionadas.

Se supone que las observaciones de un fenómeno se toman bajo idénticas condiciones y que cada observación es tomada independientemente de cualquier otra. Los datos forman entonces una muestra aleatoria (es decir, son variables independientes e idénticamente distribuidas); las técnicas estadísticas estándar pueden ser aplicadas para construir un modelo estadístico y estimar los parámetros del modelo.

La falta de homogeneidad en los datos es usualmente explicada en un modelo estadístico por la suposición de media no constante; frecuentemente se supone que la media es una combinación lineal de varias variables explicativas. Sin embargo, aun después de contabilizar las variaciones de escala grande, hay razones suficientes para sospechar de la existencia de variaciones no homogéneas a escala pequeña.

La independencia es una suposición muy conveniente que hace más manejable la teoría estadística matemática. Sin embargo, los modelos que incluyen dependencia estadística son más reales; dos clases de modelos que comúnmente se han usado incluyen estructuras de correlación intraclases y estructuras de correlación serial. Estos ofrecen pequeño alcance para datos espaciales, donde la dependencia esta presente en todas direcciones y llega a ser más débil que la localización de los datos que son más dispersos. La noción de que los datos de unidades que están cercanamente juntas, en el espacio o en el tiempo, estan probablemente correlacionados (no pueden ser modelados como estadísticamente independientes) de modo natural y son utilizados por estadísticos para modelar fenómenos fisico-químicos y probablemente sociales. También se usan los modelos puramente temporal o de series de tiempo y estan basados en observaciones idénticamente distribuidas que son dependientes y ocurren en puntos igualmente espaciados en el tiempo con un flujo unidireccional.

El análisis y modelación de datos espaciales tiene componentes básicos como *localidades espaciales* $\{s_1, \dots, s_n\}$ y *datos observados* de esas localidades $\{Z(s_1), \dots, Z(s_n)\}$. Por lo general se supone que los datos observados son aleatorios y algunas veces las localidades se suponen aleatorias. Una vez que la localización esta dada, la posibilidad de equivocación o posición imprecisa no es modelada. El análisis de datos espaciales se puede realizar por lo menos desde cuatro enfoques; *procesos espaciales indexados sobre espacio continuo* (conocido como geoestadística), *procesos espaciales reindexados como látices en el espacio* (análogo espacial de las series de tiempo), *procesos espaciales puntuales* (incluye procesos puntuales marcados y conjunto de procesos aleatorios) y *procesos de objetos*. En los datos geoestadísticos, sea D un subconjunto fijo de \mathbb{R}^d que contiene un rectángulo d -dimensional o volumen positivo y $\mathbf{Z}(\mathbf{s})$ un vector aleatorio,

definido mas adelante, en la localidad $s \in D$; para los datos de látice, D es una colección fija (regular o irregular) de muchos puntos contables de \mathbb{R}^d , una generalización es donde D es una gráfica en \mathbb{R}^d ; los patrones de puntos tienen a un D , el cual es un punto en \mathbb{R}^d o un subconjunto de \mathbb{R}^d , en su forma general esto genera un proceso puntual marcado; cuando no se especifica Z se obtiene el proceso espacial puntual (o se podría considerar el escalar $Z(s) \equiv 1$, para todo $s \in D$) y en los procesos de objetos D es un proceso puntual en \mathbb{R}^d y $Z(s)$ es un objeto aleatorio intrínseco; esto produce procesos como el modelo Boleano, el cual es un proceso relacionado con el crecimiento celular.

4.1.2. Datos y Modelos Espaciales

Los inicios estadísticos para datos espaciales y modelos espaciales se remontan desde Halley (1686), Student (1907) y Fisher (1935), este último, estaba consiente de la dependencia espacial en experimentos en campos agrícolas, ya que realizó grandes esfuerzos para removerlo, él estableció los principios de aleatorización, bloqueo y repetición, así como el control de sesgo no deseado; la aleatorización también neutraliza (pero no remueve) el efecto de la correlación espacial aunque se debe realizar para *no* neutralizar la correlación espacial a escalas mas grandes o mas pequeñas que las dimensiones de las unidades experimentales.

Los estudios de Smith y Karr (1985) relacionados con la selección de las dimensiones de las parcelas, de manera tal que cualquier incremento en el tamaño de la parcela produzca un pequeño decremento en la varianza del error; sin embargo su análisis fue empírico, por lo que la verdadera formulación del problema reconoce la presencia de correlación espacial en experimentos de campo.

El método de vecinos más cercanos para el análisis de experimentos agrícolas considera la dependencia espacial contabilizándola indirectamente por el uso de residuos de las parcelas vecinas como covariables o por diferenciación.

Dado que actualmente la estadística puede ser aplicada a cualquier disciplina

cuantitativa y que en áreas como la geología, ecología y ciencias ambientales no es posible (o no es apreciado) aleatorizar, bloquear y repetir los datos, hay una necesidad para nuevos modelos estadísticos y enfoques que presentan nuevas dudas de las antiguas y nuevas tecnologías. Muchos de los problemas resultantes, como la valoración de los recursos naturales, monitoreo ambiental (calentamiento global), imágenes médicas, etcétera, los cuales son espaciales en su naturaleza.

4.1.2.1. El modelo general espacial

En todos los aspectos del análisis exploratorio de datos, los parámetros estimados a partir de la teoría de distribución asintótica, son en mayor o menor grado parte de un modelo estocástico. Hay que tener presente a los modelos para datos espaciales, los cuales tienen una estructura muy simple que es suficientemente flexible para ser utilizado en una clase grande de problemas. Los datos pueden ser continuos o discretos, pueden ser agregaciones espaciales y observaciones de puntos en el espacio, sus localidades espaciales pueden ser regulares o irregulares y sus localidades pueden ser de un conjunto espacial continuo o discreto. Al menos el modelo estocástico es usado para el resumen de todos los datos o predecir datos no observados. Se puede o no explicar el por qué ocurre un fenómeno particular y debe distinguirse del uso más común aceptado del “modelo” genérico de datos en el espacio Euclidiano n dimensional y suponer que los datos potenciales $\mathbf{Z}(s)$ en la localidad espacial s es una cantidad aleatoria. Sea s que varia sobre el conjunto índice $D \subset \mathbb{R}^d$ de manera que genere el campo aleatorio multivariado (o proceso aleatorio o proceso estocástico)

$$\{\mathbf{Z}(s): s \in D\}; \quad (4.1.2)$$

una realización de (4.1.2) es denotada como

$$\{\mathbf{z}(s): s \in D\}. \quad (4.1.3)$$

Usualmente se supone que D es un subconjunto fijo (es decir, no aleatorio) de \mathbb{R}^d es un conjunto aleatorio, éste último es un mapeo medible de un espacio de probabilidad en un espacio medible, usualmente cerrado de subconjuntos de \mathbb{R}^d , en el cual se supone que D es igual a como \mathbf{Z} puede variar de realización a realización, dada otra fuente de aleatoriedad

del problema. Esta estructura permite estudiar el problema en los contextos de índice espacial continuo, índice de látice, patrones de puntos espaciales y otros. La flexibilidad de (4.1.2) es aparente, aunque D y $\mathbf{Z}(s)$ pueden ser cantidades aleatorias más generales y debido a que D es un subconjunto de \mathbb{R}^d , permite que sea nombrado un *proceso espacial*.

Por otro lado, se maneja por separado los procesos de *series de tiempo* de los de procesos espaciales, por medio del índice t , esto se denota como

$$\{\mathbf{Z}(t): -\infty < t < \infty\}, \quad (4.1.4)$$

debido al flujo unidimensional del tiempo, algunas veces se tiende a distinguir entre un proceso aleatorio (4.1.2) y un proceso de series de tiempo (4.1.4) en \mathbb{R}^1 . A partir de (4.1.4) se puede establecer un proceso de tiempo-espacio, el cual se denota

$$\{\mathbf{Z}(s; t): s \in D(t); t \in T\} \quad (4.1.5)$$

donde cada \mathbf{Z} , D y T son posiblemente aleatorios.

4.1.3. Principales Enfoques del Análisis Espacial

4.1.3.1. Datos geoestadísticos

La geoestadística actual tiene su origen alrededor del año de 1960 y es considerada en fechas recientes como una disciplina híbrida de ingeniería de minas, geología, matemáticas y estadística. Es el más fuerte enfoque clásico para la estimación de las reservas minerales en la que se reconoce variabilidad espacial, tanto de escala grande y escala pequeña o en términos estadísticos se puede decir que el modelo presenta tendencia espacial y correlación.

Los métodos de tendencia de superficie (Whitten, 1970) incluyen sólo variación de escala grande suponiendo errores independientes. Watson (1972) encontró que la mayoría de los datos de estudios geológicos tienen variación de escala grande y exhiben una fuerte correlación positiva entre los datos localidades espaciales cercanas.

Uno de los problemas mas importantes en geoestadística es la predicción del grado mineral en una roca observada de las muestras, a este proceso de predicción se le ha llamado *kriging* (Matheron, 1963b). El análisis de datos de las minas de grado mineral (Gomez y Hazen, 1970) es un ejemplo de datos regularmente espaciados en \mathbb{R}^2 . Los métodos geoestadísticos han sido utilizados por los científicos de suelos que buscan en los mapas propiedades del suelo de un área a partir de un número pequeño de muestras en localidades conocidas. Por ejemplo, el pH del suelo en agua, la conductividad eléctrica, potasio intercambiable, tensión e infiltración agua-suelo, son algunas de las variables que podrían ser muestreadas. La erosión de las capas fértiles del suelo por el agua es una gran preocupación por los agricultores por lo que los estudios espaciales de la infiltración suelo-agua son muy importantes. La mayor infiltración, la menor corriente de agua superficial resulta de una menor erosión del suelo y menor contaminación de las corrientes de agua (ríos, arroyos, etc.) por pesticidas y fertilizantes. Así mismo, una gran infiltración implica una mejor estructura del suelo, lo cual es más adecuado para el crecimiento del cultivo. En Cressie y Horton (1987) se presentan estudios espaciales de datos de suelo, comparan tratamientos y aplican *kriging*.

Hay una considerable cantidad de estudios donde se usa geoestadística, los datos originales usualmente no son presentados, por razones confidenciales. Las publicaciones de David (1977, 1988), Journel y Huijbregts (1978) y Clark (1979) están enfocados a aplicaciones en la minería. Algunas aplicaciones en otras áreas son en: la cantidad precipitación de lluvia (Ord y Rees, 1979), ciencias atmosféricas (Thiebaut y Pedder, 1987), mapeo de suelos (Burgess y Webster, 1980), predicción de concentración de contaminantes de agua subterránea (Istok y Cooper, 1988) y respecto a propiedades agronómicas se presentan en Viera *et al.* (1982).

Los estudios de tipo geoestadístico son distinguibles de los datos de tipo látice y patrones de puntos por la habilidad del índice espacial s para variar continuamente sobre el subconjunto de \mathbb{R}^d .

4.1.3.2. Datos de ládice

Un ládice de localizaciones evoca la idea de puntos regularmente espaciados en \mathbb{R}^d , relacionando vecinos mas cercanos, segundos vecinos más cercanos y así sucesivamente. Los ládices irregulares tiene distribución relativa que no siguen un patrón predecible y cuyas asociaciones no siempre son obvias de su geometría. De todas las posibles estructuras espaciales que un proceso estocástico puede generar, un conjunto de datos cuya localización espacial son un ládice regular en \mathbb{R}^d es el análogo más cercano a una serie de tiempo observado a puntos de tiempo igualmente espaciados. Mediante los sensores remotos de los satélites es una forma eficiente de obtener datos. Además permiten obtener información rápidamente de patrones climáticos, distribución de minerales y edad de cultivos sin tener que realizar largos e intensos muestreos tradicionales, debido a que los satélites en órbita de la tierra reciben datos en forma de reflejo de ondas electromagnéticas a un determinado número de frecuencias, incluidas aquellas de la parte visible del espectro; mediante varios métodos de muestreo e integración, la superficie de la tierra es dividida en pequeñas áreas, cada una llamada pixel. En el campo agrícola hay algunas áreas con cultivos comerciales de trigo, maíz, soya, etcétera que requieren ser estimados. Cada especie de cultivo tiene sus propias propiedades de reflejo que, junto con ruido (distorsión) son de modo remoto censados (medidos). Estos dato son recibidos como un ládice regular en \mathbb{R}^2 (considerando no significativa la curvatura de la tierra) y son identificados en el centro de su respectivo pixel; esto último es análogo a problemas temporales donde los datos son series de tiempo económicas, las cuales son una suma de un año, pero son identificadas individualmente espaciales en el eje del tiempo equis. Hay un número considerable de sobre posición entre la técnica de sensores remotos y las técnicas de imágenes médicas (consideradas de bajo nivel); aunque las escalas espaciales son completamente diferentes, la forma de los datos y las preguntas que se buscan responder son frecuentemente similares. Los modelos estadísticos para tales datos requieren expresar el hecho de que las observaciones cercanas (en el espacio o tiempo) tienden a ser iguales. Aunque algunas veces uno podría forzar a que las observaciones cercanas tiendan a ser diferentes, como es el caso (Mead, 1967) de la competencia entre plantas por luz y nutrientes del suelo puede generar que una gran cantidad de plantas sanas estén rodeadas por algunas menos fuertes.

En contraste al problema geostadístico, los datos de problemas de látice puede ser exhaustivo de los fenómenos. El muestreo puede hacerse, por ejemplo, suponiendo que sólo una pequeña parte de todos los datos son espaciales. Hay diversos estudios publicados basados en datos de látice (aunque tales investigaciones no siempre explotan el componente espacial), entre los que se encuentran están Mercer y Hall (1911), Batchelor y Reed (1918), Cochran (1936), Bartlett (1974), Cliff y Ord (1981), Ripley (1981), Symons *et al.* (1983) y Cressie y Chan (1989).

4.1.3.3. Datos en patrones de puntos

Los patrones de punto se obtienen cuando la variable importante a ser analizada es la localización de los “eventos”. La más frecuente pregunta a ser contestada es si los patrones de puntos exhiben una componente de aleatoriedad espacial, agrupación o regularidad. Por ejemplo, si se considera la localización de la longitud de las hojas de pino en un bosque; de aquí se puede investigar qué es lo biológicamente importante de la agrupación de esos árboles. Podría ser, por ejemplo, la variable diámetro a cierta altura junto con el registro de su localización. Hacer grandes (pequeños) grupos y cómo interactúan los árboles grandes y pequeños. La variable tamaño es usualmente una *variable de marca* y el proceso completo es llamado *proceso espacial puntual marcado*. La variable de marca, no necesariamente es una variable real, podría por ejemplo, ser un conjunto. El proceso consiste de $\{Z(s_i): s_i \in D\}$, con D un proceso espacial puntual y $Z(s_i)$ un proceso aleatorio localizado en $s_i \in D$. Los datos de este proceso son observados como una realización de

$$X = \cup \{Z(s_i): s_i \in D\}, \quad (4.1.6)$$

el cual es referido como modelo Booleano generalizado (Serra, 1980) o un proceso mosaico (Diggle, 1981). Los datos son por lo general imágenes y su asimilación a un modelo parecido a (4.1.6) requiere el más alto nivel de análisis de imagen. El objetivo es estimar parámetros de un conjunto aleatorio y el proceso puntual (Cressie y Laslett, 1987). Diggle (1981) analiza la incidencia de algunas enfermedades a acuerdo a un modelo Booleano. Cressie (1991) presenta el análisis para el crecimiento de tumores que itera el modelo

Booleano a través de sucesivas etapas; también se presentan dibujos de células aisladas creciendo *in vitro* y analiza la manera en que toma la forma y tamaño.

Entre las publicaciones del análisis de patrones de puntos espaciales se puede encontrar a Pielou (1959 y 1977), Matern (1960), Getis y Boots (1978), Marquiss *et al.* (1978), Ripley (1981), Diggle (1983) y Upton y Fingleton (1985). También hay métodos de análisis para patrones que ocurren en el espacio y para aquellos que ocurren en el tiempo; los métodos temporales se pueden revisar en Cox y Lewis (1966), Lewis (1972), Daley y Vere-Jones (1972), Cox y Ishan (1980) y Daley y Vere-Jones (1988).

4.1.4. Razón de la Estadística Para Datos Espaciales

Mediante modelos espaciales simples se puede mostrar el efecto de la correlación sobre la estimación, predicción y diseño experimental.

4.1.4.1. Estimación, predicción y datos para diseños experimentales

Sea $Z(1), \dots, Z(n)$ que son independientes e idénticamente distribuidos de una distribución normal con media μ y varianza σ_o^2 desconocida. El estimador, insesgado de mínima varianza, para μ es

$$\bar{Z} \equiv \sum_{i=1}^n \frac{Z(i)}{n}, \quad (4.1.7)$$

y se puede hacer la inferencia sobre μ . El estimador de \bar{Z} tiene distribución *Normal* con media μ y varianza σ_o^2 / n . Por lo que un intervalo de confianza bilateral del 95% para μ es:

$$\left(\bar{Z} - 1.96 \frac{\sigma_o}{\sqrt{n}}, \bar{Z} + 1.96 \frac{\sigma_o}{\sqrt{n}} \right). \quad (4.1.8)$$

En lugar de datos independientes, si ahora suponemos que los datos son positivamente correlacionados con una correlación que decrece cuando la separación entre los datos se incrementa, la

$$\text{Cov}(Z(i), Z(j)) = \sigma_o^2 \cdot \rho^{|i-j|}; \quad i, j = 1, \dots, n, \quad 0 < \rho < 1. \quad (4.1.9)$$

La función de autocorrelación anterior resulta de un proceso autorregresivo de primer orden $Z(i) = \rho Z(i-1) + \varepsilon(i)$; $i \geq 1$, donde $\varepsilon(i)$ es parte de una secuencia independiente e idénticamente distribuida de una variable aleatoria normal con media cero y varianza $\sigma_o^2(1 - \rho^2)$ y es independiente de $Z(i-1)$. Sin embargo, no es el único mecanismo de generación; así $Z \equiv (Z(1), \dots, Z(n))$ son datos espaciales en \mathbb{R}^1 de la cual la predicción, por lo general, de $Z(0)$ o $Z(3/2)$ es tan sólo apropiado como el de $Z(n+1)$.

Si

$$\begin{aligned} \text{Var}(\bar{Z}) &= n^{-2} \left\{ \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(Z(i), Z(j)) \right\} \\ &= \left\{ \sigma_o^2 / n \right\} \left[1 + 2\{\rho / (1 - \rho)\} \{1 - (1/n)\} - 2\{\rho / (1 - \rho)\}^2 (1 - \rho^{n-1}) / n \right]. \end{aligned} \quad (4.1.10)$$

Para $n = 10$ y $\rho = 0.26$ se tiene que

$$\text{Var}(\bar{Z}) = \left\{ \sigma_o^2 / 10 \right\} (1.608), \quad (4.1.11)$$

y el intervalo de confianza bilateral del 95% para μ es:

$$\left(\bar{Z} - 2.485 \frac{\sigma_o}{\sqrt{10}}, \bar{Z} + 2.485 \frac{\sigma_o}{\sqrt{10}} \right). \quad (4.1.12)$$

La falla para contabilizar la presencia de correlación positiva de los datos provoca que el intervalo de confianza en (4.1.8) sea angosto, con una cobertura de probabilidad del 87.8%, no del 95.0%.

El entendimiento intuitivo del efecto de la correlación espacial se obtiene de (4.1.10) al reescribirlo como

$$\text{Var}(\bar{Z}) = \frac{\sigma_o^2}{n}, \quad (4.1.13)$$

donde

$$n' \equiv \frac{n}{1 + 2\{\rho/(1-\rho)\}\{1 - (1/n)\} - 2\{\rho/(1-\rho)\}^2(1 - \rho^{n-1})/n} \quad (4.1.14)$$

se interpreta como el *número equivalente de observaciones independientes*. Si $n = 10$ y $\rho = 0.26$ entonces $n' = 6.2$; es decir, aproximadamente 6 observaciones independientes alcanzan para la misma precisión que 10 observaciones (correlacionadas de acuerdo con (4.1.9) y $\rho = 0.26$).

Para n grande, n' es:

$$n' = n / [(1 + \rho)/(1 - \rho)], \quad (4.1.15)$$

demostrando que la correlación tiene efecto sobre la inferencia, aun para grandes muestras.

Para lo referente a la predicción, supongamos que la observación $Z(n+1)$ se desea predecir de los datos $\mathbf{Z} \equiv (Z(1), \dots, Z(n))$, donde se supone que $Z(1), \dots, Z(n), Z(n+1)$ son conjuntamente normal, idénticamente distribuidos con media desconocida μ y varianza desconocida σ^2 e independientes. El predicho (o predictor o estimador) $p(\mathbf{Z}; n+1)$ que satisface la condición de insesgamiento $E(p(\mathbf{Z}; n+1)) = \mu$ y que minimiza el cuadrado medio del error predicho

$$E(Z(n+1) - p(\mathbf{Z}; n+1))^2 \quad (4.1.16)$$

es

$$p_o(\mathbf{Z}; n+1) = \bar{Z} \quad (4.1.17)$$

la media muestral. La minimización del cuadrado medio del error predicho es

$$ms_o = \sigma_o^2 \{1 + (1/n)\}. \quad (4.1.18)$$

Cuando la suposición de independencia es reemplazada con la función (4.1.9), el predictor insesgado que minimiza el cuadrado medio del error predicho es

$$p_\rho(Z; n+1) = \rho Z(n) + (1-\rho) \frac{Z(1) + (1-\rho) \sum_{i=1}^{n-1} Z(i) + Z(n)}{n - (n-2)\rho}; \quad (4.1.19)$$

si $\rho = 0$ en (4.1.19) da lugar a (4.1.17). Así mismo, si \bar{Z} es usado en lugar de (4.1.19), el cuadrado medio del error predicho es

$$E(Z(n+1) - \bar{Z})^2 = \sigma_o^2 \left\{ 1 + (1/n) \left[1 + 2\left\{ \rho/(1-\rho) \right\} \left\{ \rho^n - (1/n) \right\} - 2\left\{ \rho/(1-\rho) \right\}^2 \left(1 - \rho^{n-1} \right) / n \right] \right\}. \quad (4.1.20)$$

Para $n = 0$ y $\rho = 0.26$ la expresión (4.1.20) produce

$$E(Z(n+1) - \bar{Z})^2 = \sigma_o^2 \{ 1.09051 \}. \quad (4.1.21)$$

Comparando (4.1.21) con el resultado de (4.1.18) se produce

$$ms_o = \sigma_o^2 (1.1), \quad (4.1.22)$$

cuya diferencia no es grande. Cabe mencionar que cuando la correlación espacial decae geoméricamente con la distancia, el intervalo de predicción clásico es a menudo frecuentemente válido. Sin embargo, ello puede ser altamente ineficiente.

Considerando el modelo de dependencia en (4.1.9), el minimizado cuadrado medio del error predicho es obtenido por el cálculo de (4.1.16) con el predictor óptimo (4.1.19), es decir,

$$ms_\rho = \sigma_o^2 \left[1 - \rho^2 + \left\{ (1-\rho)^2 (1+\rho) / (n - (n-2)\rho) \right\} \right]; \quad (4.1.23)$$

si $\rho = 0$ en (4.1.23) se produce (4.1.18).

Para $n = 0$ y $\rho = 0.26$ sustituidos en (4.1.23) produce

$$ms_\rho = \sigma_o^2 (1.01952); \quad (4.1.24)$$

por lo que un intervalo de predicción del 95% para $Z(n+1)$, basado en el predictor óptimo (4.1.19) es

$$\left(p_\rho(\mathbf{Z}) - (1.96)\sigma_o \{1.01952\}^{1/2}, p_\rho(\mathbf{Z}) + (1.96)\sigma_o \{1.01952\}^{1/2} \right). \quad (4.1.25)$$

El intervalo (4.1.25) debe ser comparado con el más "amplio" intervalo de predicción del 95% para $Z(n+1)$ basado en \bar{Z} :

$$\left(\bar{Z} - (1.96)\sigma_o \{1.09051\}^{1/2}, \bar{Z} + (1.96)\sigma_o \{1.09051\}^{1/2} \right). \quad (4.1.26)$$

Para tamaños de muestra n grande, la razón al cuadrado de el intervalo de predicción amplio por el uso de $p_\rho(\mathbf{Z})$ para que el uso de \bar{Z} sea aproximadamente igual a $1 - \rho^2$ lo cual es una medida asintótica de inferencia basada en \bar{Z} contra la inferencia basada en el predictor óptimo $p_\rho(\mathbf{Z})$. Es decir, para $\rho = 0.26$ la eficiencia es del 93% y para $\rho = 0.5$ es del 75%.

En lo referente a datos de diseños experimentales, es importante la consideración de la correlación espacial en el área de los diseños de experimentos. Si las unidades experimentales son colocadas en un arreglo de $t \times b$ hecho para b bloques (columnas), cada uno con t unidades experimentales dentro de ellos. Las unidades experimentales que son mas semejantes son colocadas en el mismo bloque de modo que todo el grueso de la heterogeneidad ocurra entre los bloques. Suponiendo además que los t tratamientos son comparados a partir de la variable de respuesta, que es medida después de aplicar los tratamientos a las unidades experimentales en forma tal que cada tratamiento aparece exactamente una vez en cada bloque. Para el análisis de este diseño (bloques completos al azar), la respuesta de la i -ésima unidad experimental del j -ésimo bloque, $Y_{ij(k)}$, $i = 1, \dots, t$, $j = 1, \dots, b$, $k = 1, \dots, t$ donde k denota el tratamiento asignado (de acuerdo al diseño) a la unidad experimental. Suponiendo el modelo aditivo

$$Y_{ij(k)} = \mu + \alpha_k + \delta(i, j), \quad (4.1.27)$$

donde $\delta(\cdot, \cdot)$ es un proceso de error normal con media cero. Adicionalmente suponemos independencia de los δ 's dentro de los bloques pero la dependencia espacial de (4.1.9) disminuye en los bloques. Es decir,

$$\text{Cov}(\delta(i, j), \delta(i', j')) = \begin{cases} 0, & j \neq j', \\ \sigma^2 \rho^{|i-i'|}, & j = j', \end{cases} \quad (4.1.28)$$

donde $|\rho| < 1$.

Uno de los objetivos de un diseño experimental es encontrar la ubicación (asignación) de los tratamientos a las unidades experimentales para que sea más precisa la estimación de los efectos de los tratamientos,

$$\tau_k = -\frac{\sum_{l=1}^t \alpha_l}{t}, \quad k = 1, \dots, t. \quad (4.1.29)$$

Específicamente se puede minimizar a

$$A \equiv \sum_{1 \leq k < l \leq t} \text{Var}(\hat{\tau}_k - \hat{\tau}_l) / \binom{t}{2}, \quad (4.1.30)$$

donde $\{\hat{\tau}_k : k = 1, \dots, t\}$ son los estimadores de mínimos cuadrados generalizados de los efectos de tratamientos.

La precisión de la estimación de los efectos de tratamiento se puede mejorar por la utilización de diseños óptimos, en este sentido, Gill y Shukla (1985) encontraron que diseño de vecinos cercanos de primer orden balanceados, que satisface las condiciones adicionales de balance, es un diseño óptimo. En este diseño, el número de veces que los tratamientos aparecen adjuntos en un bloque es constante para cada par posible de tratamientos. Los diseños óptimos pueden no existir; el diseño de vecinos más cercanos de primer orden balanceados es relativamente fácil de construir y parecen ser casi tan eficientes como el diseño más óptimo (Cressie, 1991 y Grondona y Cressie, 1992).

La estrategia, por lo general, cuando se aplica el enfoque de modelación espacial es estimar y modelar la correlación espacial de los residuos y entonces utilizarlos para obtener los estimadores de los efectos de tratamientos $\{\hat{\tau}_k : k = 1, \dots, t\}$ mediante mínimos cuadrados generalizados. Un experimento artificial es construido por adición de efectos de tratamientos (de acuerdo a un diseño de bloques completos al aleatorizados) a los datos especialmente correlacionados uniformemente, el análisis espacial produce una ganancia de eficiencia de al menos 30% (Grondona y Cressie, 1991).

En muchas aplicaciones de la estadística a problemas, donde hay datos espaciales inherentes, los datos son observaciones que resultan de experimentos controlados. Por lo que la posibilidad de aleatorizar, bloquear y repetir, no es posible. En éstas circunstancias, el modelo ajustado a los datos puede suponer implícito que algunos efectos no controlados o interacciones son cero. Con la esperanza de que tales suposiciones estén basadas en conocimientos previos del fenómeno y no por conveniencia matemática (Wiens *et al.*, 1986).

4.1.4.2. Modelos lineales con errores espacialmente dependientes

Sean los datos $Z(s_1), \dots, Z(s_n)$, observados en localidades espaciales $\{s_1, \dots, s_n\}$, que son modelados como una colección de variables aleatorias, generadas por un proceso aleatorio

$$Z(s) = \sum_{l=1}^q \beta_l x_l(s) + \delta(s), \quad s \in D \subset \mathbb{R}^d \quad (4.1.31)$$

donde $\{x_l(\cdot) : l = 1, \dots, q\}$ es una colección de variables explicativas no aleatorias que pueden o no depender de una localización espacial y $\delta(\cdot)$ es un proceso de error con media cero y varianza finita que puede o no ser espacialmente correlacionada.

Así, el modelo con media constante μ y covarianza dada en (4.1.31) en \mathbb{R}^1 , con sólo una variable explicativa $x_1(\cdot) \equiv 1$, $\beta_1 \equiv \mu$ y $Cov(\delta(u), \delta(v)) = \sigma_o^2 \cdot \rho^{|u-v|}$, $u, v \in \mathbb{R}^1$. También, el modelo asociado al diseño de bloques completos (4.1.27) y (4.1.28) es un caso

especial de (4.1.31) en \mathbb{R}^2 , donde las variables explicativas son $x_1(\cdot) \equiv 1$ y $x_{k+1}(i, j) = 1$, si la (i, j) -ésima parcela recibe el k -ésimo tratamiento y cero de otro modo ($k = 1, \dots, t$) y $x_{l+t+1}(i, j) = 1$ si $j = l$ y cero de otro modo ($l = 1, \dots, b$). El proceso de error con correlación espacial anisotrópico dado por (4.1.28) tienen correlación específica dentro de los bloques, pero no entre bloques.

El modelo dado en (4.1.31) y las observaciones $\mathbf{Z} \equiv (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$ se pueden escribir en notación de matrices y vectores del modo siguiente

$$\mathbf{Z} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\delta}, \quad (4.1.32)$$

donde \mathbf{X} es una matriz $n \times q$ cuyo (i, j) -ésimo elemento es $x_j(\mathbf{s}_i)$, $\boldsymbol{\beta} \equiv (\beta_1, \dots, \beta_q)'$ y $\boldsymbol{\delta} \equiv (\delta(\mathbf{s}_1), \dots, \delta(\mathbf{s}_n))'$.

Por lo general, los parámetros de mayor interés son los parámetros, $\boldsymbol{\beta}$, de variación a escala grande y se supone que cualquier parámetro de variación a escala pequeña está asociado con el proceso de error $\delta(\cdot)$ son conocidos. Es decir, se puede obtener

$$\text{Var}(\boldsymbol{\delta}) = \boldsymbol{\Sigma} \quad (4.1.33)$$

donde $\boldsymbol{\Sigma}$ es una matriz simétrica conocida de $n \times n$, no negativa definida. Los coeficientes de regresión $\boldsymbol{\beta}$'s son estimables y $\boldsymbol{\Sigma}$ es invertible y existe su inversa.

Cuando el objetivo es una estimación eficiente de $\boldsymbol{\beta}$ (también puede ser requerida una estimación eficiente de un valor desconocido $Z(\mathbf{s}_0)$), el mejores estimadores lineales insesgados son a su vez los estimadores de mínimos cuadrados generalizados, es decir, el valor de $\boldsymbol{\beta}$ que minimiza a

$$(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}). \quad (4.1.34)$$

Por lo que

$$\hat{\beta}_{mco} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Z, \quad (4.1.35)$$

los cuales son los estimadores de máxima verosimilitud de $\delta(\cdot)$ que es un proceso normal.

La matriz de varianzas para (4.1.35) es

$$Var(\hat{\beta}_{mco}) = (X' \Sigma^{-1} X)^{-1}. \quad (4.1.36)$$

Cuando $\Sigma = \sigma^2 I$, bajo los supuestos clásicos de los errores (no espacial) independientes e idénticamente distribuidos, la expresión (4.1.34) se reduce $(Z - X\beta)'(Z - X\beta) / \sigma^2$. Por lo que la minimización apropiada produce

$$\hat{\beta}_{mco} = (X' X)^{-1} X' Z, \quad (4.1.37)$$

los cuales son los estimadores de mínimos cuadrados ordinarios para β . Cuando $\Sigma \neq \sigma^2 I$, los estimadores de mínimos cuadrados ordinarios son frecuentemente ineficientes. Específicamente,

$$Var(\hat{\beta}_{mco}) = (X' X)^{-1} (X' \Sigma X) (X' X)^{-1} \quad (4.1.38)$$

y la diferencia $Var(\hat{\beta}_{mco}) - Var(\hat{\beta}_{mco})$ es no negativa definida; por lo que cuando $\Sigma = \sigma^2 I$, la diferencia es 0.

Debido a que los estimadores de mínimos cuadrados ordinarios para (4.1.37) no requiere que se conozca Σ , es de importancia encontrar las condiciones en las que X y Σ , para los cuales mínimos cuadrados ordinarios y mínimos cuadrados generalizados son igualmente eficientes o para los cuales los estimadores de mínimos cuadrados ordinarios y mínimos cuadrados generalizados son los mismos. Cuando mínimos cuadrados ordinarios y mínimos cuadrados generalizados producen los mismos estimadores, se debe tener la precaución de no suponer que las inferencias son válidas y útiles.

El comportamiento, por lo general mal, de los procedimientos de inferencia estadística clásicos en presencia de dependencia espacial (Cliff y Ord, 1981; Dow *et al.*, 1982; Anselin y Griffith, 1988) y la ambigüedad de dicha dependencia espacial, demuestran por qué las estadísticas para datos espaciales deben ser parte del repertorio de información que se proporciona en los reportes estadísticos.

4.1.4.3. Validez del modelo espacial

Dado que en realidad, Σ es desconocido, típicamente dependerá de los parámetros en θ obtenido de un modelo de covarianza paramétrica espacial

$$C(\mathbf{u}, \mathbf{v}; \theta) \equiv \text{Cov}(\delta(\mathbf{u}), \delta(\mathbf{v})), \quad \mathbf{u}, \mathbf{v} \in D \subset \mathbb{R}^d. \quad (4.1.39)$$

Suponiendo homocedasticidad, un parámetro podría ser $\sigma^2 \equiv \text{Var}(\delta(u))$. Sea $\Sigma(\theta)$ que denota a una matriz $n \times n$ de varianzas $\text{Var}(Z)$ cuyo (i, j) -ésimo elemento es $C(\mathbf{s}_i, \mathbf{s}_j; \theta)$.

La selección de un modelo apropiado puede ser una tarea difícil debido a que se debe tener cuidado de usar modelos válidos; es decir, usar sólo aquellos θ para el cual

$$\sum_{i=1}^m \sum_{j=1}^m a_i C(\mathbf{u}_i, \mathbf{u}_j; \theta) a_j \geq 0, \quad (4.1.40)$$

para cualquier m , cualquier secuencia de números reales $\{a_i : i = 1, \dots, m\}$.

Para el modelo de covarianza válido en \mathbb{R}^1 $C(\mathbf{s}_i, \mathbf{s}_j; \theta) = C_T(\|\mathbf{s}_i - \mathbf{s}_j\|; \theta)$, y su covarianza podría ser

$$C_T(h; \theta) \equiv \begin{cases} \sigma^2 \left(1 - \frac{|h|}{\theta}\right), & 0 \leq |h| \leq \theta, \\ 0, & |h| > \theta. \end{cases} \quad (4.1.41)$$

Sin embargo, en \mathbb{R}^2 , $C(\mathbf{s}_i, \mathbf{s}_j; \theta) \equiv C_T(\|\mathbf{s}_i - \mathbf{s}_j\|; \theta)$ no es un modelo válido. Por otro

lado, si el modelo espacial para los datos $\mathbf{Z} \equiv (Z(s_1), \dots, Z(s_n))'$ tiene suficiente error medido, de manera tal que $\Sigma(\theta)$ es diagonalmente dominante, esto es, si

$$C(s_i, s_j; \theta) \geq \sum_{\substack{j=1 \\ j \neq i}}^n |C(s_i, s_j; \theta)|, \quad i = 1, \dots, n, \quad (4.1.42)$$

entonces $\Sigma(\theta)$ es siempre no negativa definida y por lo tanto un modelo válido es obtenido.

En los ejemplos para modelos anteriores (4.1.9) y (4.1.28), son válidos para cuando el parámetro $\rho \in (-1, 1)$. Es importante asegurar que el espacio parametral Θ de θ consista sólo de modelos válidos y que un estimado $\hat{\theta}$ de θ pertenezca a Θ .

4.1.4.4. Estimación de mínimos cuadrados generalizados

Suponiendo que el parámetro de dependencia espacial θ esta contenido en un espacio parametral conocido Θ , cualquier elemento de este, puede generar un modelo espacial que podría ser válido. Cualquier estimador $\hat{\theta}$ se puede asumir que pertenece a Θ , con probabilidad 1.

Los estimadores de mínimos *cuadrados generalizados estimados* son obtenidos por la sustitución de $\Sigma(\hat{\theta})$ por $\Sigma(\theta)$ en (4.1.35):

$$\hat{\beta}_{moge} = (\mathbf{X}' \Sigma(\hat{\theta})^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma(\hat{\theta})^{-1} \mathbf{Z} \quad (4.1.43)$$

donde $\hat{\theta}$ es un estimador de los parámetros θ , de variación a pequeña escala.

La predicción de un $Z(s_0)$ desconocida procede en una forma similar, frecuentemente se asume que el vector estimado de $\hat{\theta}$ es el valor verdadero. Dos enfoques diferentes son frecuentemente empleados para obtener $\hat{\theta}$. Si la función de densidad de probabilidad (o masa) de \mathbf{Z} es completamente especificada como $f(\mathbf{z}; \beta, \theta)$, entonces la verosimilitud

$$l(\beta, \theta) \equiv f(\mathbf{Z}; \beta, \theta) \quad (4.1.44)$$

puede ser maximizada con respecto a β y θ para obtener los estimadores $\hat{\beta}$ y $\hat{\theta}$. Entonces (4.1.43) es el estimador de máxima verosimilitud de β .

Una segunda posibilidad para obtener una estimación inicial de β (es decir $\hat{\beta}_{mco}$) a partir de la construcción de los residuos

$$\mathbf{R} \equiv \mathbf{Z} - \mathbf{X}\hat{\beta}_{mco} \quad (4.1.45)$$

y usando estos residuos como una aproximación para δ , la estimación de θ libre de la presencia de β . Este método es empleado frecuentemente en un contexto semiparámetro, donde la distribución conjunta de \mathbf{Z} no es especificada, pero la media $\mathbf{X}\beta$ y la varianza $\Sigma(\theta)$, sí son especificadas. Así, $\hat{\theta}$ puede ser obtenido por aproximación mediante el enfoque del método de momentos basados en \mathbf{R} . $\hat{\beta}_{moge}$ se obtiene de $\hat{\theta}$ y (4.1.43). Se tiene entonces la opción de sustituir esta estimación anterior en (4.1.45) para obtener un nuevo \mathbf{R} , un $\hat{\theta}$ y un nuevo $\hat{\beta}_{moge}$ y así sucesivamente; terminando en esta segunda etapa se genera el proceso de estimación en dos etapas.

4.2. Geoestadística

4.2.1. Origen

El prefijo “geo” es referido como parte del planeta tierra y la idea original fue usado en la minería. Hart (1954) presentó el término “geoestadística” en un contexto geográfico para denotar técnicas estadísticas que enfatizan la localidad dentro de un área de distribución.

Matheron (1962, 1963a, 1963b) usó el término en el contexto de la geología para denotar la teoría y métodos para inferir las reservas de mineral de datos espacialmente distribuidos a través de rocas minerales. Actualmente la geoestadística no está asociado

exclusivamente a estudios relacionados con el planeta tierra, teniendo actualmente un papel más universal, relacionado con la teoría estadística y aplicaciones para procesos con índice espacial continuo.

4.2.2. Análisis Espacial de Datos

La liga común es que los datos pueden ser tratados como una realización (usualmente parcial) de un proceso aleatorio (o proceso estocástico o campo aleatorio)

$$\{Z(\mathbf{s}): \mathbf{s} \in D\}, \quad (4.2.1)$$

donde D es un subconjunto fijo de \mathbb{R}^d con volumen d -dimensional positivo. Es decir, el índice espacial \mathbf{s} varía continuamente a través de la región D . También es posible considerar datos espacio temporal considerándola como una variable $Z(\mathbf{s}, t)$. El enfoque es considerar el modelo (4.2.1) de manera tal que permita la inferencia sobre los parámetros del modelo a partir de los datos $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$ en localidades espaciales conocidas. La herramienta útil es el variograma (llamado también análisis estructural).

Cuando se inicia el análisis espacial de datos, se puede empezar con un análisis de datos exploratorios, por medio de gráficos estadísticos (considerando su posición o localización geográfica) y estadísticas elementales, que proporcionan una buena aproximación para conocer la posible existencia de datos anormales (aberrantes) y si los supuestos de los datos (modelo) se cumplen o no.

4.2.2.1. Estacionariedad intrínseca

Para considerar el modelo estadístico y sus parámetros, supongamos que se mide una variable de respuesta en los puntos \mathbf{s} (en \mathbb{R}^2) que son una realización de un proceso aleatorio $\{Z(\mathbf{s}): \mathbf{s} \in D\}$ y que son observados algunos puntos $\{\mathbf{s}_i: i = 1, \dots, n\}$, en un entorno aproximadamente cuadrado, de la variable de respuesta. Entonces, *intrínsecamente estacionario*, es definido a través de las primeras diferencias

$$E(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})) = 0, \quad (4.2.2)$$

$$\text{Var}((Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})) = 2\gamma(\mathbf{h}). \tag{4.2.3}$$

La cantidad $2\gamma(\mathbf{h})$ se conoce como el *variograma* y es un parámetro crucial de la geoestadística. Un estimador clásico del variograma es

$$2\hat{\gamma}(\mathbf{h}) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(s_i) - Z(s_j))^2, \tag{4.2.4}$$

donde la suma es sobre $N(\mathbf{h}) \equiv \{(i, j): s_i - s_j = \mathbf{h}\}$ y $|N(\mathbf{h})|$ es el número de elementos diferentes de $N(\mathbf{h})$. Debido a que tiene ciertas desventajas el estimador del variograma (4.2.4), Cressie y Hawkins (1980) proponen uno mas robusto (menor sesgo y curtosis)

$$2\bar{\gamma}(\mathbf{h}) \equiv \frac{\left\{ \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} |Z(s_i) - Z(s_j)|^{1/2} \right\}^4}{\left(0.457 + \frac{0.494}{|N(\mathbf{h})|} \right)}. \tag{4.2.5}$$

Otros estimadores del variograma son propuestos por Chauvet (1982) y por Hawkins and Cressie (1984).

4.2.2.2. Descomposición de los datos en variación de escala grande y escala pequeña

La idea de analizar los datos ubicados en un contorno regular definido, viene del análisis de varianza de doble entrada, pero para las que no presentan estacionariedad sucede que las pruebas de F para columnas hileras no son válidas, porque los datos están por lo general correlacionados, sin embargo, el análisis de varianza de una tabla de doble entrada presenta una descomposición aditiva muy útil

$$dato = todos + hileras + columnas + residual, \tag{4.2.6}$$

cuando hay r hileras y c columnas, y cada localidad del cuadrículado tiene una observación Y_{ij} , su descomposición es

$$Y_{ij} = Y_{..} + (Y_{i.} - Y_{..}) + (Y_{.j} - Y_{..}) + (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..}), \tag{4.2.7}$$

donde el punto (\cdot) denota promedio sobre el subíndice faltante; se modifica ligeramente cuando hay localidades con observaciones perdidas, una forma consistente de aplicar la descomposición (4.2.6) es por medio a la aplicación de técnicas de análisis exploratorio como la *mediana destacada* (median polishing) para cada dato, la cual constituye la variación a escala pequeña.

Cabe mencionar que las técnicas para remover la tendencia (por hileras o columnas) es dependiente de la orientación del arreglo.

Por otra parte, los residuos de (4.2.6) pueden tener sesgo pequeño. Cuando el número efectivo de observaciones es pequeño, el sesgo al cuadrado es típicamente mucho mayor que la varianza, de manera que domine sobre el cuadrado medio del error.

4.2.3. Procesos Estacionarios

Sean los datos de la forma $\{z(x, y): x = 1, \dots, r; y = 1, \dots, c\}$, los cuales son números potencialmente infinitos de medidas $\{z(s): s \in D\}$ que son tomados a partir de un D . A la cantidad $z(\cdot)$ se le llama *variable regionalizada* para enfatizar la naturaleza espacial continua del conjunto indexado D .

El proceso aleatorio $\{Z(s): s \in D\}$ es usualmente definido a través de una distribución de dimensión finita

$$F_{s_1, \dots, s_m}(z_1, \dots, z_m) \equiv \Pr[Z(s_1) \leq z_1, \dots, Z(s_m) \leq z_m]; \quad m \geq 1 \quad (4.2.8)$$

la cual satisface la condición de Kolmogorov de simetría (F permanece invariante cuando z_j y s_j son subconjuntos de la misma permutación) y consistente con

$$F_{s_1, \dots, s_{k+1}}(z_1, \dots, z_k, \infty, \dots, \infty) = F_{s_1, \dots, s_k}(z_1, \dots, z_k). \quad (4.2.9)$$

Sea $\mu(s) \equiv E(Z(s))$ existente para todo $s \in D$; a $\mu(\cdot)$ se le llama *tendencia*. La existencia de la $Var(Z(s))$ para todo $s \in D$ permite la existencia de la definición de

procesos estacionarios de segundo orden e intrínsecamente lineal.

4.2.3.1. Estacionariedad de segundo orden y ergodicidad

Para que los datos, no representen un muestreo incompleto de una sola realización de Z , se supone que

$$E(Z(s)) = \mu, \text{ para todo } s \in D, \quad (4.2.10)$$

o que $F_s(z) = Pr(Z(s) \leq z)$ no dependa de s (distribución invariante para Z existente).

Con la finalidad de estimar predictores lineales óptimos, se tiene que

$$Cov(Z(s_1), Z(s_2)) = C(s_1 - s_2), \text{ para todo } s_1, s_2 \in D; \quad (4.2.11)$$

a la función $C(\cdot)$ se le llama *covariograma* o *función de covarianza estacionaria*.

Por lo tanto, una función $Z(\cdot)$ que satisface (4.2.10) y (4.2.11) es estacionario de segundo orden (o débil o amplia sensibilidad). Adicionalmente, si $C(s_1 - s_2)$ es función sólo de $\|s_1 - s_2\|$, entonces $C(\cdot)$ es llamado *isotropico*, es decir, que el proceso es estacionario e invariante en dirección (Cliff y Ord, 1981).

Así mismo, existe un subconjunto de funciones estacionarias de segundo orden que tienen una propiedad conocida como *ergodicidad*, que permite obtener esperanzas sobre el espacio Ω , los cuales son estimados por promedios espaciales. La utilidad de este concepto es en el análisis de series de tiempo y en los análisis geoestadísticos.

4.2.4. Variograma y efecto nugget

Sea

$$\text{Var}(Z(s_1) - Z(s_2)) = 2\gamma(s_1 - s_2), \text{ para todo } s_1, s_2 \in D. \quad (4.2.12)$$

La cantidad $2\gamma(\cdot)$ que es función sólo de $s_1 - s_2$, es llamada *variograma* (y a la parte $\gamma(\cdot)$ se le llama *semivariograma*), también es llamada función estructurada o cuadrado

medio de la diferencia (Matheron, 1962).

Por lo general a la función $2\gamma(\cdot)$, si existe, se le considerará como un parámetro de un proceso aleatorio $Z(\cdot)$. El parámetro puede ser de valor real o valor vectorial, de varianza no negativa, su probabilidad en el intervalo $[0, 1]$ y restringida a ser condicionada negativa definida.

Por otra parte, cuando $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$ y $\gamma(\mathbf{0}) = 0$ y si $\gamma(\mathbf{h}) \rightarrow c_0$, cuando $h \rightarrow 0$, entonces a c_0 se le llama *efecto nugget* (Matheron, 1962), el cual se cree que es la variación a nivel de microescala (pequeñas fracciones), la cual es la causa de la discontinuidad en el origen; aunque matemáticamente esta no puede suceder porque es un proceso continuo. Por lo tanto, si la continuidad de un fenómeno se espera que exista a nivel de microescala, es sólo la razón posible para que $c_0 > 0$ como medida de error la cual es la *varianza de la medida del error* c_{ME} ; además, se supone que existe la *varianza de un proceso de ruido blanco* (media cero, varianza constante y varianza cero) c_{MS} , que representa el efecto nugget del proceso a nivel de *microescala*. Es decir

$$c_0 = c_{MS} + c_{ME}, \quad (4.2.13)$$

la determinación de c_0 , en la práctica, es a partir de extrapolación en el variograma estimado del rezago más cercano a cero; en algunos procesos espaciales se supone que $c_{ME} = 0$. El comportamiento de el variograma cerca del origen es muy informativo acerca de sus propiedades de continuidad, las cuales están especificados en Cressie (1991).

4.2.4.1. Algunos modelos de variogramas isotrópicos y proceso anisotrópico

Existen tres modelos básicos isotrópicos, los cuales expresados en términos del semivariograma, son

Modelo lineal (válido en \mathbb{R}^d , $d \geq 1$):

$$\gamma(\mathbf{h}; \theta) = \begin{cases} 0, & \mathbf{h} = \mathbf{0} \\ c_0 + b_l \|\mathbf{h}\|, & \mathbf{h} \neq \mathbf{0} \end{cases} \quad (4.2.14)$$

con $\theta = (c_0, b_i)$, donde $c_0 \geq 0$ y $b_i \geq 0$.

Modelo esférico (válido en \mathbb{R}^1 , \mathbb{R}^2 y \mathbb{R}^3):

$$\gamma(\mathbf{h}; \theta) = \begin{cases} 0, & \mathbf{h} = 0 \\ c_0 + c_s \left\{ (3/2) (\|\mathbf{h}\|/a_s) - (1/2) (\|\mathbf{h}\|/a_s)^3 \right\}, & 0 < \|\mathbf{h}\| \leq a_s \\ c_0 + c_s, & \|\mathbf{h}\| \geq a_s \end{cases} \quad (4.2.15)$$

con $\theta = (c_0, c_s, a_s)$, donde $c_0 \geq 0$, $c_s \geq 0$ y $a_s \geq 0$.

Modelo exponencial (válido en \mathbb{R}^d , $d \geq 1$):

$$\gamma(\mathbf{h}; \theta) = \begin{cases} 0, & \mathbf{h} = 0 \\ c_0 + c_e \{1 - \exp(-\|\mathbf{h}\|/a_e)\}, & \mathbf{h} \neq 0 \end{cases} \quad (4.2.16)$$

con $\theta = (c_0, c_e, a_e)$, donde $c_0 \geq 0$, $c_e \geq 0$ y $a_e \geq 0$.

Otros modelos de semivariograma que normalmente se pueden utilizar son el modelo cuadrático general, modelo periódico(ondas) y el modelo de potencias, los cuales se encuentran definidos en Cressie (1991).

Por su parte, un proceso Z es *anisotrópico* cuando la dependencia entre $Z(\mathbf{s})$ y $Z(\mathbf{s} + \mathbf{h})$ es una función de tanto la magnitud y la dirección de \mathbf{h} . El variograma, de un proceso intrínsecamente estacionario, no es alargado como una función de la distancia entre localidades espaciales. La anisotropía es causada por la subordinación de procesos físicos que incluyen diferencias en el espacio; el proceso más común es en la dirección vertical (por ejemplo, donde hay diferencias de gravedad o presión) la cual es completamente diferente de la dirección horizontal. La anisotropía puede ser corregida por transformaciones lineales en el vector \mathbf{h} .

4.2.5. Covariograma y Correlograma

A la función covariograma,

$$\text{Cov}(Z(s_1), Z(s_2)) = C(s_1 - s_2), \text{ para todo } s_1, s_2 \in D, \quad (4.2.17)$$

también se le llama *función de autocovarianza* en el contexto del análisis de series de tiempo. A partir de ella, dado $C(\mathbf{0}) > 0$, se llama a

$$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{C(\mathbf{0})} \quad (4.2.18)$$

correlograma (conocido como *función autocorrelación*, la cual es usada en el contexto de series de tiempo para diagnosticar la no estacionariedad, para determinar el tipo de dependencia estacionaria y ajuste del modelo).

4.2.6. Estimación del Variograma

El variograma $2\gamma(\cdot)$ definido como

$$2\gamma(\mathbf{s}_1 - \mathbf{s}_2) = \text{Var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)), \quad (4.2.19)$$

y el proceso $Z(\cdot)$ es representado como

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}) \quad (4.2.20)$$

donde $\delta(\cdot)$ tiene media cero con un proceso estocástico intrínsecamente estacionario con variograma 2γ y la media $\mu(\cdot)$ no es constante, entonces

$$E(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2 = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2) + (\mu(\mathbf{s}_1) - \mu(\mathbf{s}_2))^2, \quad (4.2.21)$$

la cual, por lo general, no es función de $\mathbf{s}_1 - \mathbf{s}_2$. Ni tampoco satisface la condición

$$\frac{E(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2}{\|\mathbf{s}_1 - \mathbf{s}_2\|^2} \rightarrow 0, \quad \text{cuando } \|\mathbf{s}_1 - \mathbf{s}_2\| \rightarrow \infty$$

que todo variograma debe satisfacer.

Cabe mencionar que, bajo el supuesto de media constante (4.2.10), un estimador natural basado en el método de momentos (Matheron, 1962) es

$$2\hat{\gamma}(\mathbf{h}) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2, \quad \mathbf{h} \in \mathbb{R}^d, \quad (4.2.23)$$

donde

$$N(\mathbf{h}) \equiv \left\{ (s_i, s_j) : s_i - s_j = \mathbf{h}; i, j = 1, \dots, n \right\} \text{ y} \quad (4.2.24)$$

$|N(\mathbf{h})|$ es el número de pares distintos en $N(\mathbf{h})$. El estimador (4.2.23) es conocido como el *estimador clásico del variograma*.

Cuando los datos s se encuentran irregularmente espaciados en \mathbb{R}^d , el estimador del variograma es usualmente suavizado por el uso del estimador

$$2\gamma^+(\mathbf{h}(l)) \equiv \text{MedP} \left\{ (Z(s_i) - Z(s_j))^2 : (i, j) \in N(\mathbf{h}); \mathbf{h} \in T(\mathbf{h}(l)) \right\} \quad (4.2.25)$$

la región $T(\mathbf{h}(l))$ es algunas veces especificada como la región de *tolerancia* en \mathbb{R}^d alrededor de $\mathbf{h}(l)$, $l=1, \dots, k$ y $\text{MedP}\{\cdot\}$ denota un promedio ponderados sobre los elementos en $\{\cdot\}$. Se recomienda que el número de pares disjuntos $|\cup \{N(\mathbf{h}) : \mathbf{h} \in T(\mathbf{h}(l))\}|$ en $T(\mathbf{h}(l))$ sea de al menos 30.

Por otro lado, cabe mencionar, que el covariograma $C(\mathbf{h})$ o el correlograma $\rho(\mathbf{h})$ pueden ser estimados, con algunas precauciones, ya que ambos son bien definidos cuando $Z(\cdot)$ esta bien definido y es un proceso estacionario de segundo orden, pero puede no existir cuando $Z(\cdot)$ es intrinsecamente estacionario.

4.2.7. Representación Espectral

Se sabe, de los análisis de series de tiempo, que los covariogramas ajustados a los datos pueden solamente tomar formas especiales. Esas funciones $C(\cdot)$ en \mathbb{R}^d deben ser positivo definidas, es decir,

$$\sum_{i=1}^m \sum_{j=1}^m a_i a_j C(s_i - s_j) \geq 0, \quad (4.2.26)$$

para cualquier número infinito de localidades espaciales $\{s_i : i = 1, \dots, m\}$ y números reales $\{a_i : i = 1, \dots, m\}$. Esta propiedad caracteriza el covariograma de un proceso estacionario de segundo orden.

4.2.8. Ajuste de los Modelos de Variograma

De los diferentes estimadores de variogramas existentes, no pueden ser utilizados directamente para predicción espacial (kriging) debido a que no son condicionados a ser negativo definido, ya que la ausencia de esta propiedad puede resultar en predicciones impropias. La idea por medio de un variograma válido que, como un medida de la dependencia espacial, sea lo mas cercano a la dependencia espacial presente en los datos

$$\mathbf{Z} = (Z(s_1), \dots, Z(s_n))' \quad (4.2.27)$$

El espacio de todo los variogramas válidos es un conjunto grande sobre el cual buscar, de ellos, se puede seleccionar una familia paramétrica mediante diferentes métodos, entre éstos esta el de Máxima Verosimilitud, Máxima Verosimilitud Restringida, Estimación Mínimo Cuadrático Normal, Mínimos Cuadrados (Ordinarios o Generalizados o Ponderados), Estimación del Variograma Clásico y Estimación del Variograma Robusto. Los modelos de variograma ajustados por algunos de los métodos anteriores es útil en la cuantificación de la dependencia espacial de los datos de \mathbf{Z} o en el proceso $Z(\cdot)$, mas importante es para kriging, en el cálculo de ponderaciones y varianzas.

4.3. Predicción Espacial y Kriging

La necesidad de obtener predicciones adecuadas a partir de los datos observados puede ser encontrada en muchas disciplinas científicas. Las nociones de estadística que se requieren es para considerar la dependencia estadística entre los datos y la variable o variables a ser predichas.

Para ello sea $\{Z(s): s \in D \subset \mathbb{R}^d\}$ una función o proceso aleatorio de n datos $Z(s_1), \dots, Z(s_n)$ que son muestreados. Los datos son usados para llevar a cabo inferencia de el proceso para producir alguna función conocida $g(\{Z(s): s \in D\}) = g(Z(\cdot))$ de la función aleatoria $Z(\cdot)$. Así, la predicción de un punto supone que $g(Z(\cdot)) = Z(s_0)$, donde s_0 es una localidad espacial desconocida y g es de valores reales o valores de vector. Cabe mencionar que el interés no es en $Z(\cdot)$, pero si en su versión “contaminada de ruido” de él.

Así, suponga que

$$Z(\mathbf{s}) = S(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad (4.3.1)$$

donde

$\varepsilon(\mathbf{s})$ es un proceso de medida de error de ruido blanco.

En este caso, el primer interés es predecir una función conocida $g(Z(\cdot))$ de una función aleatoria de ruido $S(\cdot)$.

Al referirse a *predicción espacial* se hace énfasis a la predicción $g(Z(\cdot))$ o de $g(S(\cdot))$ a partir de los datos $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$ observados en una localidad espacial conocida $\mathbf{s}_1, \dots, \mathbf{s}_n$. Esta situación hace referencia a la noción de suavizamiento (o interpolación), filtración y predicción. Si los datos temporales están disponibles del pasado y se incluyen en el presente, el suavizamiento se refiere a la predicción de $g(S(\cdot))$ en algunos puntos en el pasado, la filtración se refiere a la predicción de $g(S(\cdot))$ en el presente y la predicción se refiere a la predicción $g(S(\cdot))$ a un determinado número de puntos en el futuro. En este contexto, la “estimación” es usada para inferencia de parámetros fijos desconocidos; la “predicción” es reservada para inferencia de cantidades aleatorias.

Kriging es un método mínimo cuadrado medio del error para predicción espacial que usualmente depende de las propiedades de segundo orden y de las propiedades de el proceso $Z(\cdot)$. El ingeniero de minas D. G. Krige desarrolló alrededor de 1950, métodos empíricos para determinar la distribución de la verdadera cantidad de mineral de rocas muestreadas distribuidos en una mina (Krige, 1951). Sin embargo, la determinación de el predictor lineal espacial óptimo se dió después de los trabajos de Krige. Las investigaciones de Wold (1938), Kolmogorov (1941), Wiener (1949), Gandin (1963) y Matheron (1963b y 1971) quienes nombraron al enfoque como *análisis objetivo* y uso la terminología de *óptima interpolación*. Estos enfoques contienen información de la predicción lineal óptima, dentro del llamado método de predicción espacial óptima, que refleja la noción de que las observaciones más cercanas se pueden utilizar para la predicción de puntos.

4.3.1. El Mejor Predictor Lineal

El mejor predictor lineal, es encontrar l_1, \dots, l_n , y k en

$$p(\mathbf{Z}; \mathbf{s}_0) = \sum_{i=1}^n l_i Z(\mathbf{s}_i) + k, \quad (4.3.2)$$

tal que $E(Z(\mathbf{s}_0) - p(\mathbf{Z}; \mathbf{s}_0))^2$ sea mínimo. Equivalentemente, minimizar (sobre l_1, \dots, l_n, k)

$$E\left(Z(\mathbf{s}_0) - \sum_{i=1}^n l_i Z(\mathbf{s}_i) - k\right)^2 = \text{Var}\left(Z(\mathbf{s}_0) - \sum_{i=1}^n l_i Z(\mathbf{s}_i)\right) + \left(\mu(\mathbf{s}_0) - \sum_{i=1}^n l_i \mu(\mathbf{s}_i) - k\right)^2, \quad (4.3.3)$$

donde $\mu(\mathbf{s}_0) = E(Z(\mathbf{s}_0))$, $\mathbf{s} \in D$. Seleccionar de $k = \mu(\mathbf{s}_0) - \sum_{i=1}^n l_i \mu(\mathbf{s}_i)$ y

$$\mathbf{l}' = \mathbf{c}' \Sigma^{-1}, \quad (4.3.4)$$

donde $\mathbf{c} \equiv (C(\mathbf{s}_0, \mathbf{s}_1), \dots, C(\mathbf{s}_0, \mathbf{s}_n))'$ y Σ es una matriz $n \times n$ cuyo (i, j) -ésimo elemento es $C(\mathbf{s}_i, \mathbf{s}_j)$, produce el predictor lineal óptimo $p^*(\mathbf{Z}; \mathbf{s}_0)$ o $Z^*(\mathbf{s}_0)$. Es decir,

$$p^*(\mathbf{Z}; \mathbf{s}_0) = \mathbf{c}' \Sigma^{-1} (\mathbf{Z} - \boldsymbol{\mu}) + \mu(\mathbf{s}_0), \quad (4.3.5)$$

donde $\boldsymbol{\mu} \equiv (\mu(\mathbf{s}_1), \dots, \mu(\mathbf{s}_n))'$. La predicción del cuadrado medio del error minimizado es

$$\sigma_{sk}^2(\mathbf{s}_0) \equiv C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}' \Sigma^{-1} \mathbf{c}. \quad (4.3.6)$$

Es predictor espacial es llamado como *kriging simple* porque permite conocer la función de la media $\mu(\cdot)$. Cabe mencionar que kriging ordinario y kriging universal producen predictores óptimos ordinarios lineales en los datos, pero donde la función de medias es lineal en un número fijo de parámetros desconocidos. Esos parámetros son estimados óptimamente, pero el precio a pagar es que la predicción de los cuadrados medios de los errores es mayor que $\sigma_{sk}^2(\cdot)$.

4.3.2. Escala de Variación Observacional y Espacial

Sean los datos $(Z(s_1), \dots, Z(s_n))$ que representan los valores de Z en los puntos $D \subset \mathbb{R}^d$ y que son modelados como realizaciones parciales de un proceso aleatorio

$$\{Z(\mathbf{s}): \mathbf{s} \in D \subset \mathbb{R}^d\}, \quad (4.3.7)$$

el cual satisface la descomposición

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \eta(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad (4.3.8)$$

donde:

$\mu(\cdot) \equiv E(Z(\cdot))$ es la estructura de media determinística la cual es llamada *variación a escala grande*.

$W(\cdot)$ es un proceso de media cero L_2 -continuo, es decir $E(W(\mathbf{s} + \mathbf{h}) - W(\mathbf{s}))^2 \rightarrow 0$ cuando $\|\mathbf{h}\| \rightarrow 0$ es un proceso intrínsecamente estacionario cuyo rango de variograma (si existe) es mayor que $\text{Min}\{\|\mathbf{s}_i - \mathbf{s}_j\|: 1 \leq i < j \leq n\}$. Es llamado *variación de escala pequeña suavizado*.

$\eta(\cdot)$ es un proceso con media cero, independiente de W , cuyo rango de variograma existe y es más pequeño que $\text{Min}\{\|\mathbf{s}_i - \mathbf{s}_j\|: 1 \leq i < j \leq n\}$. Es llamado *variación a microescala*.

$\varepsilon(\cdot)$ es un proceso de ruido blanco con media cero, es independiente de W y η . Se llama *medida del error* o *ruido blanco* y es denotado $\text{Var}(\varepsilon(\mathbf{s})) = c_{ME}$. Hay ocasiones en que $\varepsilon(\cdot)$ posee más estructura que el proceso de ruido blanco.

Entonces, en la notación

$$2\gamma_Z(\cdot) = 2\gamma_W(\cdot) + 2\gamma_\eta(\cdot) + 2c_{ME} \quad (4.3.9)$$

Las cantidades c_{ME} , $\gamma_Z(\mathbf{h})$ y $\|\mathbf{h}\|$ pertenecen a la escala observacional; las otras cantidades contienen información en escala espacial.

La descomposición (4.3.8) se puede escribir

$$Z(\mathbf{s}) = S(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad (4.3.10)$$

donde la “señal” o proceso suavizado $S(\cdot)$ es dado por $S(\cdot) = \mu(\cdot) + W(\cdot) + \eta(\cdot)$. El proceso S es referido como la versión *con menos ruido* de el proceso Z o como el *estado del proceso*.

También se escribe

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in D \quad (4.3.11)$$

donde el proceso de error correlacionado $\delta(\cdot)$ es dado por $\delta(\cdot) = W(\cdot) + \eta(\cdot) + \varepsilon(\cdot)$. Cuando la correlación de $\delta(\mathbf{h})$ con $\delta(\mathbf{s} + \mathbf{h})$ puede ser escrita como una función de $\mathbf{h}/a(\mathbf{h})$, donde $0 < a(\mathbf{h}) < \infty$, entonces $a(\mathbf{h})$ es llamado *escala de correlación espacial* de el proceso en dirección \mathbf{h} .

La descomposición (4.3.8) no es única y es de naturaleza largamente operacional. Esto provoca que diferentes analistas de información podrían llegar a diferentes conclusiones a partir de un mismo conjunto de datos, dependiendo de qué tanta variación pueda ser atribuida a los componentes de (4.3.8). Se puede tener una idea de la medida del error mediante la repetición de observaciones en la misma localidad espacial. Más observaciones en una nueva localidad espacial podría permitir la estimación de escalas debajo del $\text{Min}\{\|s_i - s_j\| : 1 \leq i < j \leq n\}$. En general, la más importante descomposición de $Z(\cdot)$ en $\mu(\cdot)$ y $\delta(\cdot)$ dado por (4.3.11) no puede ser determinada en forma única.

4.3.2.1. Suposiciones paramétricas de la función de medias

La descomposición de (4.3.11) puede ser de forma única simplemente por la especificación de un algoritmo de suavizamiento para $\mu(\cdot)$ basado en los datos $\{s_i, Z(s_i) : i = 1, \dots, n\}$. Alternativamente, se podría parametrizar $Z(\cdot)$ o $\mu(\cdot)$. En \mathbb{R}^2 , algunos ejemplos de parametrización de $\mu(\cdot)$ son

$$\mu(\mathbf{s}) = a + c(x) + r(y), \quad \mathbf{s} = (x, y)' \quad (4.3.12)$$

$$\mu(\mathbf{s}) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2, \quad \mathbf{s} = (x, y)' \quad (4.3.13)$$

$$\mu(\mathbf{s}) = \beta_0 + \beta_1 \cos(w_1 x + w_2 y), \quad \mathbf{s} = (x, y)' \quad (4.3.14)$$

la expresión (4.3.12) es la base de kriging y mediana destacada (polish), (4.3.13) es la base de kriging universal y (4.3.14) es la base del análisis de Fourier de dos dimensiones, donde el par de frecuencias (w_1, w_2) es considerado dominante siempre y cuando el proceso $Z(\cdot)$ sea oscilatorio en su naturaleza.

4.3.3. Kriging Ordinario

La palabra “kriging” se sinónimo de “predicción óptima” o como “óptimamente predicho”. Es decir, se refiere a la realización de inferencias de valores no observados de el proceso aleatorio $Z(\cdot)$ o de $S(\cdot)$, a partir de los datos

$$\mathbf{Z} \equiv (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))', \quad (4.3.15)$$

observados en localidades espaciales conocidas $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, se denota al predictor genérico de $g(Z(\cdot))$ o $g(S(\cdot))$ por $p(\mathbf{Z}; g)$.

Kriging ordinario tiene dos supuestos:

1.- Suposición del modelo. En,

$$Z(\mathbf{s}) = \mu + \delta(\mathbf{s}), \quad \mathbf{s} \in D, \quad \mu \in \mathbb{R} \text{ y } \mu \text{ desconocido.} \quad (4.3.16)$$

2.- Suposición de predicción

$$p(\mathbf{Z}; B) = \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i), \quad \sum_{i=1}^n \lambda_i = 1, \quad (4.3.17)$$

que los coeficientes de el predictor lineal sume 1, garantiza el insesgamiento uniforme: $E(p(\mathbf{Z}; B)) = \mu = E(Z(B))$, para todo $\mu \in \mathbb{R}$. Hay una versión de kriging llamado *kriging simple*, donde μ en (4.3.16) es conocido y los coeficientes no son restringidos a sumar 1.

4.3.3.1. Predicción espacial óptima de el proceso Z

Para kriging ordinario, la minimización de

$$\sigma_e^2 = E(Z(B) - p(Z; B))^2 \quad (4.3.18)$$

es llevada a cabo sobre $(\lambda_1, \dots, \lambda_n)$, sujeto a $\sum_{i=1}^n \lambda_i = 1$, donde el modelo (4.3.16) se supone que es apoyado con el variograma

$$2\gamma(\mathbf{h}) = \text{Var}(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})), \quad \mathbf{h} \in \mathbb{R}^d. \quad (4.3.19)$$

Sea $B = \{\mathbf{s}_0\}$ en

$$g(Z(\cdot)) = Z(B) = \begin{cases} \int_B Z(\mathbf{u}) d\mathbf{u} / |B|, & |B| > 0, \\ \text{Ave}\{Z(\mathbf{u}) : \mathbf{u} \in B\}, & |B| = 0. \end{cases} \quad (4.3.20)$$

Entonces, minimizar

$$E\left(Z(\mathbf{s}_0) - \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)\right)^2 - 2m\left(\sum_{i=1}^n \lambda_i - 1\right) \quad (4.3.21)$$

con respecto a $\lambda_1, \dots, \lambda_n$ y m (el parámetro m es un multiplicador de Lagrange que asegure que $\sum_{i=1}^n \lambda_i = 1$).

La condición que $\sum_{i=1}^n \lambda_i = 1$ implica que

$$E\left(Z(\mathbf{s}_0) - \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)\right)^2 = -\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 / 2 + 2\sum_{i=1}^n \lambda_i (Z(\mathbf{s}_0) - Z(\mathbf{s}_i))^2 / 2, \quad (4.3.22)$$

de tal manera, que bajo el modelo (4.3.16), (4.3.21) se convierta a

$$-\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j) + 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - 2m \left(\sum_{i=1}^n \lambda_i - 1 \right). \quad (4.3.23)$$

Después de diferenciar (4.3.23) con respecto a $\lambda_1, \dots, \lambda_n$ y m e igualando a cero, los parámetros óptimos satisfacen las dos condiciones siguientes

$$-\sum_{j=1}^n \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j) + \gamma(\mathbf{s}_0 - \mathbf{s}_i) - m = 0, \quad i = 1, \dots, n, \quad \text{y} \quad (4.3.24)$$

$$\sum_{i=1}^n \lambda_i = 1.$$

Es decir, los $\lambda_1, \dots, \lambda_n$ óptimos pueden ser obtenidos de

$$\lambda_o = \Gamma_o^{-1} \gamma_o \quad (4.3.25)$$

donde

$$\lambda_o \equiv (\lambda_1, \dots, \lambda_n, m)', \quad (4.3.26)$$

$$\gamma_o \equiv (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \dots, \gamma(\mathbf{s}_0 - \mathbf{s}_n))', \quad (4.3.27)$$

$$\Gamma_o \equiv \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j), & i = 1, \dots, n, \quad j = 1, \dots, n, \\ 1, & i = n+1, \quad j = 1, \dots, n, \\ 0, & i = n+1, \quad j = n+1, \end{cases} \quad (4.3.28)$$

y Γ_o es una matriz simétrica $(n+1) \times (n+1)$. Cabe hacer notar que (4.3.25) permanece sin cambio si $\gamma(\mathbf{h})$ es reemplazado con $\gamma(\mathbf{h}) + c$, para cualquier $c \in \mathbb{R}$. La sustitución, es algunas veces necesaria para obtener una solución numérica estable para $\lambda_o \Gamma = \gamma_o$.

De (4.3.25), los coeficientes $\lambda \equiv (\lambda_1, \dots, \lambda_n)'$ son dados por

$$\lambda' = \left(\gamma + \mathbf{1} \frac{\mathbf{1}' \Gamma^{-1} \gamma}{\mathbf{1}' \Gamma^{-1} \mathbf{1}} \right)' \Gamma^{-1}, \quad \text{y} \quad (4.3.29)$$

$$m = -(\mathbf{1} - \mathbf{1}' \Gamma^{-1} \gamma) / (\mathbf{1}' \Gamma^{-1} \mathbf{1}), \quad (4.3.30)$$

donde $\gamma \equiv (\gamma(s_0 - s_1), \dots, \gamma(s_0 - s_n))'$ y Γ es una matriz $n \times n$ cuyo (i, j) -ésimo elemento es $\gamma(s_i - s_j)$. La escritura del predictor óptimo para kriging ordinario (4.3.17) a partir de (4.3.25) se puede hacer como $\hat{p}(Z; s_0)$ o $\hat{Z}(s_0)$ y del cual se puede construir intervalos de confianza, adicional a el valor de $\sigma_k^2(s_0)$. La predicción del cuadrado medio del error minimizado (3.2.7) es algunas veces llamado *varianza de kriging* o *predicción de la varianza de kriging*, es decir,

$$\begin{aligned} \sigma_k^2(s_0) &= \lambda' \gamma \gamma' = \sum_{i=1}^n \lambda_i \gamma(s_0 - s_i) + m \\ &= \gamma' \Gamma^{-1} \gamma - (\mathbf{1}' \Gamma^{-1} \gamma - 1) / (\mathbf{1}' \Gamma^{-1} \mathbf{1}); \end{aligned} \quad (4.3.31)$$

así mismo,

$$\sigma_k^2(s_0) = 2 \sum_{i=1}^n \lambda_i \gamma(s_0 - s_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(s_i - s_j). \quad (4.3.32)$$

4.3.3.2. Kriging en términos de la función de covarianza

El error cuadrado de la predicción (4.3.22) puede ser escrito como

$$\begin{aligned} \left(Z(s_0) - \sum_{i=1}^n \lambda_i Z(s_i) \right)^2 &= (Z(s_0) - \mu)^2 \\ &+ \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j (Z(s_i) - \mu)(Z(s_j) - \mu) \\ &- 2 \sum_{i=1}^n \lambda_i (Z(s_0) - \mu)(Z(s_i) - \mu), \end{aligned} \quad (4.3.33)$$

restringido a que $\sum_{i=1}^n \lambda_i = 1$. Sea el modelo (4.3.16) con $\delta(\cdot)$ un proceso estacionario de segundo orden y media cero, que tiene covariograma $C(\mathbf{h})$, $\mathbf{h} \in \mathbb{R}^d$. Entonces (4.3.21) es

$$C(\mathbf{0}) + \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(s_i - s_j) - 2 \sum_{i=1}^n \lambda_i C(s_0 - s_i) - 2m \left(\sum_{i=1}^n \lambda_i - 1 \right), \quad (4.3.34)$$

minimizando esta última expresión con respecto a $\{\lambda_i : i = 1, \dots, n\}$ y m produce la ecuación de kriging ordinario

$$\begin{aligned} \hat{\rho}(\mathbf{Z}; \mathbf{s}_0) &= \boldsymbol{\lambda}' \mathbf{Z}, \\ \sigma_k^2(\mathbf{s}_0) &= C(\mathbf{0}) - \boldsymbol{\lambda}' \mathbf{c} + m \end{aligned} \quad (4.3.35)$$

donde

$$\boldsymbol{\lambda}' = \left(\mathbf{c} + \mathbf{1} \frac{\mathbf{1}' \boldsymbol{\Sigma}^{-1} \mathbf{c}}{\mathbf{1}' \boldsymbol{\Sigma}^{-1} \mathbf{1}} \right)' \boldsymbol{\Sigma}^{-1} \quad \text{y} \quad m = \frac{\mathbf{1}' \boldsymbol{\Sigma}^{-1} \mathbf{c}}{\mathbf{1}' \boldsymbol{\Sigma}^{-1} \mathbf{1}}, \quad (4.3.36)$$

aquí, $\mathbf{c} \equiv (C(\mathbf{s}_0 - \mathbf{s}_1), \dots, C(\mathbf{s}_0 - \mathbf{s}_1))'$ y $\boldsymbol{\Sigma}$ es una matriz $n \times n$ cuyo (i, j) -ésimo elemento es $C(\mathbf{s}_i - \mathbf{s}_j)$. Estas ecuaciones requieren que el $Z(\cdot)$ sea estacionario de segundo orden.

4.3.4. Kriging Robusto

La optimalidad de la predicción lineal espacial (o kriging simple) está supeditado a que $Z(\cdot)$ sea un proceso Gaussiano intrínsecamente estacionario.

Sea la descomposición

$$Z(\mathbf{s}) = \mu + W(\mathbf{s}) + \eta(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad (4.3.37)$$

con $W(\cdot)$ intrínsecamente estacionario y Gaussiano. Sin embargo $E(\cdot) \equiv \eta(\cdot) + \varepsilon(\cdot)$ no se supone que es Gaussiano (Normal).

Sea

$$E(\mathbf{s}) \sim \begin{cases} N(0, c_0), & \text{con probabilidad } 1 - \varepsilon, \\ H, & \text{con probabilidad } \varepsilon, \end{cases} \quad (4.3.38)$$

donde ε es pequeño (por ejemplo $\varepsilon = 0.05$) y H es alguna distribución con extremos extendidos, media cero y varianza finita. El modelo (4.3.37) es referido como la variabilidad muestra a muestra dentro de una subárea D .

4.3.4.1. Estimación del variograma

Debido a que los componentes de (4.3.37) se suponen independientes,

$$2\gamma_Z(\mathbf{h}) = 2\gamma_W(\mathbf{h}) + 2\left\{(1-\varepsilon)c_0 + \varepsilon \int z^2 dH(z)\right\}, \mathbf{h} \in \mathbb{R}^d, \quad (4.3.39)$$

donde se asume que $\int z^2 dH(z) < \infty$. Adicionalmente se asume que cualquier observación de H tiene error, el proceso de interés real es

$$Y(\mathbf{s}) = \mu + W(\mathbf{s}) + \tilde{E}(\mathbf{s}), \mathbf{s} \in D, \quad (4.3.40)$$

donde $\tilde{E}(\mathbf{s})$ es un proceso Gaussiano de ruido blanco con distribución marginal $N(0, c_0)$. El variograma a ser estimado o usado para propósitos de predicción es

$$2\gamma_Y(\mathbf{h}) = 2\gamma_W(\mathbf{h}) + 2c_0, \mathbf{h} \in \mathbb{R}^d. \quad (4.3.41)$$

4.3.5. Kriging Universal

Para el modelo de media constante

$$Z(\mathbf{s}) = \mu + \delta(\mathbf{s}), \mathbf{s} \in D, \quad (4.3.42)$$

donde $\mu \in \mathbb{R}^d$ es desconocido y $\delta(\cdot)$ es un proceso intrínsecamente estacionario con media cero y variograma $2\gamma(\cdot)$. La predicción de un $Z(B)$ desconocido de los datos \mathbf{Z} es referido como kriging ordinario. Una posible generalización de (4.3.42) es suponer un proceso de error más general $\delta(\cdot)$. Se supone que $E(Z(\mathbf{s})) [= \mu(\mathbf{s})]$ no es una constante grande pero es una combinación lineal desconocida de funciones conocidas $\{f_0(\mathbf{s}), \dots, f_p(\mathbf{s})\}$; cada $f_j(\mathbf{s})$ es una función de localización.

4.3.5.1. Supuestos del modelo y del predictor

Consideración del supuesto

$$Z(\mathbf{s}) = \sum_{j=1}^{p+1} f_{j-1}(\mathbf{s})\beta_{j-1} + \delta(\mathbf{s}), \mathbf{s} \in D \quad (4.3.43)$$

donde $\beta \equiv (\beta_0, \dots, \beta_p)' \in \mathbb{R}^{p+1}$ es un vector desconocido de parámetros y $\delta(\cdot)$ es un proceso

intrínsecamente de media cero con variograma $2\gamma(\cdot)$. Los datos de \mathbf{Z} se pueden escribir como

$$\mathbf{Z} = X\boldsymbol{\beta} + \boldsymbol{\delta} \quad (4.3.44)$$

donde X es una matriz $n \times (p+1)$ cuyo (i, j) -ésimo elemento es $f_{j-1}(\mathbf{s}_i)$.

Adicionalmente tenemos que

$$Z(B) = \mathbf{x}'\boldsymbol{\beta} + \delta(B), \quad (4.3.45)$$

donde $\mathbf{x} \equiv (f_0(B), \dots, f_p(B))'$, $f_j(B) = \int_B f_j(\mathbf{u}) d\mathbf{u} / |B|$, $j = 0, \dots, p$ y $\delta(B) = \int_B \delta(\mathbf{u}) d\mathbf{u} / |B|$.

Cuando $B = \{\mathbf{s}_0\}$, $\mathbf{x} \equiv (f_0(\mathbf{s}_0), \dots, f_p(\mathbf{s}_0))'$.

Respecto a los supuestos del predictor, es deseable predecir $Z(B)$ linealmente de los datos \mathbf{Z} usando un predictor uniformemente insesgado. Es decir, un predictor de la forma

$$p(\mathbf{Z}; B) = \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i), \text{ para } \boldsymbol{\lambda}' X = \mathbf{x}' \quad (4.3.46)$$

4.3.5.2. Predicción espacial óptima de el proceso \mathbf{Z} para las ecuaciones de kriging universal

Para kriging universal, el predictor lineal insesgado óptimo se representa como $p(\mathbf{Z}; B)$ o $\hat{\mathbf{Z}}(B)$, minimiza la predicción del cuadrado medio del error

$$\sigma_e^2 = E(Z(B) - p(\mathbf{Z}; B))^2 \quad (4.3.47)$$

sobre $\lambda_1, \dots, \lambda_n$ sujeto a $\boldsymbol{\lambda}' X = \mathbf{x}'$. Lo "universal" es usado para referir a lo insesgado de el predictor, cuando la tendencia es una combinación lineal desconocida de funciones conocidas.

De las ecuaciones de kriging universal tenemos que sobre diferenciando a

$$\begin{aligned}
& E\left(Z(\mathbf{s}_0) - \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)\right)^2 - 2 \sum_{j=1}^{p+1} m_{j-1} \left\{ \sum_{i=1}^n \lambda_i f_{j-1}(\mathbf{s}_i) - f_{j-1}(\mathbf{s}_0) \right\} \\
& \Rightarrow \\
& \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j) + 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - 2 \sum_{j=1}^{p+1} m_{j-1} \left\{ \sum_{i=1}^n \lambda_i f_{j-1}(\mathbf{s}_i) - f_{j-1}(\mathbf{s}_0) \right\}
\end{aligned} \tag{4.3.48}$$

con respecto a $\lambda_1, \dots, \lambda_n, m_0, \dots, m_p$ e igualando el resultado a cero, las ponderaciones óptimas son obtenidos de

$$\lambda_U = \Gamma_U^{-1} \gamma_U, \tag{4.3.49}$$

donde

$$\lambda_U \equiv (\lambda_1, \dots, \lambda_n, m_0, \dots, m_p)', \tag{4.3.50}$$

$$\gamma_U \equiv (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \dots, \gamma(\mathbf{s}_0 - \mathbf{s}_n), 1, f(\mathbf{s}_0), \dots, f_p(\mathbf{s}_0))', \tag{4.3.51}$$

y Γ_U es una matriz simétrica $(n+p+1) \times (n+p+1)$:

$$\Gamma_U = \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j), & i = 1, \dots, n, & j = 1, \dots, n, \\ f_{j-1-n}(\mathbf{s}_i), & i = 1, \dots, n, & j = n+1, \dots, n+p+1, \\ 0, & i = n+1, \dots, n+p+1, & j = n+1, \dots, n+p+1, \end{cases} \tag{4.3.52}$$

con $f_0(\mathbf{s}) \equiv 1$. Es decir, los coeficientes λ estan dados por

$$\lambda' = \left\{ \gamma + X(X'\Gamma^{-1}X)^{-1}(\mathbf{x} - X'\Gamma^{-1}\gamma) \right\}' \Gamma^{-1} \mathbf{y} \tag{4.3.53}$$

$$\mathbf{m}' = -(\mathbf{x} - X'\Gamma^{-1}\gamma)'(X'\Gamma^{-1}X)^{-1}, \tag{4.3.54}$$

donde $\gamma \equiv (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \dots, \gamma(\mathbf{s}_0 - \mathbf{s}_n))'$ y Γ es una matriz $n \times n$ cuyo (i, j) -ésimo elemento es $\gamma(\mathbf{s}_i - \mathbf{s}_j)$.

Cuando se supone que $Z(\cdot)$ es intrínsecamente estacionario de segundo orden, el covariograma $C(\mathbf{h}) = \text{Cov}(Z(\mathbf{s} + \mathbf{h}), Z(\mathbf{s}))$ está bien definido y la ecuación análoga a (4.3.49) es

$$(\lambda_1, \dots, \lambda_n, -m_0, \dots, m_p)' = \Sigma_U^{-1} \mathbf{c}_U \quad (4.3.55)$$

donde $\mathbf{c}_U \equiv (C(\mathbf{s}_0 - \mathbf{s}_1), \dots, C(\mathbf{s}_0 - \mathbf{s}_n), f_0(\mathbf{s}_0), f_1(\mathbf{s}_0), \dots, f_p(\mathbf{s}_0))'$ y Σ_U es una matriz simétrica $(n + p + 1) \times (n + p + 1)$

$$\Sigma_U \equiv \begin{cases} C(\mathbf{s}_i - \mathbf{s}_j), & i = 1, \dots, n, & j = 1, \dots, n, \\ f_{j-1-n}(\mathbf{s}_i), & i = 1, \dots, n, & j = n+1, \dots, n+p+1, \\ 0, & i = n+1, \dots, n+p+1, & j = n+1, \dots, n+p+1; \end{cases} \quad (4.3.56)$$

no se requiere que $f_0(\cdot) \equiv 1$.

Por lo tanto, los coeficientes λ están dados por

$$\lambda' = \left\{ \mathbf{c} + X(X' \Sigma^{-1} X)^{-1} (\mathbf{x} - X' \Sigma^{-1} \mathbf{c}) \right\} \Sigma^{-1}, \quad (4.3.57)$$

donde $\mathbf{c} \equiv (C(\mathbf{s}_0 - \mathbf{s}_1), \dots, C(\mathbf{s}_0 - \mathbf{s}_n))$ y Σ es una matriz $n \times n$ cuyo (i, j) -ésimo elemento es $C(\mathbf{s}_i - \mathbf{s}_j)$. Además, $\mathbf{m}' = (\mathbf{x} - X' \Sigma^{-1} \mathbf{c})' (X' \Sigma^{-1} X)^{-1}$.

4.3.5.3 Estimación del variograma para kriging universal

Dado que se asume que el variograma es conocido y su estimación por medio de $2\gamma(\cdot)$ u otro puede ser inapropiado ya que

$$\begin{aligned} E(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 &= \text{Var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) + \{\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)\}^2 \\ &= 2\gamma(\mathbf{s}_i - \mathbf{s}_j) + \left\{ \sum_{k=1}^{p+1} \beta_{k-1} (f_{k-1}(\mathbf{s}_i) - f_{k-1}(\mathbf{s}_j)) \right\}^2. \end{aligned} \quad (4.3.58)$$

Si β es conocido, un estimador de el variograma puede ser basado en

$$\delta(\cdot) \equiv Z(\cdot) - \sum_{k=1}^{p+1} \beta_{k-1} f_{k-1} \quad (4.3.59)$$

porque

$$E(\delta(\mathbf{s}_i) - \delta(\mathbf{s}_j))^2 = 2\gamma(\mathbf{s}_i - \mathbf{s}_j), \quad (4.3.60)$$

pero si β es desconocido, se requiere el conocimiento de $2\gamma(\cdot)$ para que sea una estimación eficiente. Pero, dado que $2\gamma(\cdot)$ es desconocido, esto lleva nuevamente las circunstancias al inicio. Este ciclo lleva a alguna ineficiencia para kriging universal.

4.3.5.4 Kriging para bloque

Cuando $|B| > 0$ en (4.3.47), las ecuaciones de kriging universal (4.3.49) y la varianza de kriging

$$\begin{aligned} \sigma_k^2(\mathbf{s}_0) &= \lambda_U' \gamma_U = \sum_{i=1}^n \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) + \sum_{j=1}^{p+1} m_{j-1} f_{j-1}(\mathbf{s}_0) \\ &= \boldsymbol{\gamma}' \Gamma \boldsymbol{\gamma} - (\mathbf{x} - X' \Gamma^{-1} \boldsymbol{\gamma})' (X' \Gamma^{-1} X)^{-1} (\mathbf{x} - X' \Gamma^{-1} \boldsymbol{\gamma}) \end{aligned} \quad (4.3.61)$$

$$= 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j); \quad (4.3.62)$$

para el covariograma $Z(\cdot)$ es bien definido la varianza de kriging $\sigma_k^2(\mathbf{s}_0)$, es

$$\begin{aligned} \sigma_k^2(\mathbf{s}_0) &= C(\mathbf{0}) - \sum_{i=1}^n \lambda_i C(\mathbf{s}_0 - \mathbf{s}_i) + \sum_{j=1}^{p+1} m_{j-1} f_{j-1}(\mathbf{s}_0) \\ &= C(\mathbf{0}) - \mathbf{c}' \Sigma^{-1} \mathbf{c} + (\mathbf{x} - X' \Sigma^{-1} \mathbf{c})' (X' \Sigma^{-1} X)^{-1} (\mathbf{x} - X' \Sigma^{-1} \mathbf{c}) \\ &= C(\mathbf{0}) - 2 \sum_{i=1}^n \lambda_i C(\mathbf{s}_0 - \mathbf{s}_i) + \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(\mathbf{s}_i - \mathbf{s}_j) \end{aligned} \quad (4.3.63)$$

los cuales dan lugar a

$$\lambda_U = \Gamma_U^{-1} \boldsymbol{\gamma}_U(B), \quad (4.3.64)$$

$$\sigma_k^2(B) = \lambda_U' \boldsymbol{\gamma}_U(B) - \gamma(B, B), \quad (4.3.65)$$

donde

$$\gamma_U(B) \equiv \left(\gamma(B, \mathbf{s}_1), \dots, \gamma(B, \mathbf{s}_n), 1, f_1(B), \dots, f_p(B) \right)', \quad (4.3.66)$$

$$\gamma(B, \mathbf{s}_i) \equiv \int_B \gamma(\mathbf{u} - \mathbf{s}_i) d\mathbf{u} / |B|, \quad i = 1, \dots, n, \quad (4.3.67)$$

$$f_j(B) \equiv \int f_j(\mathbf{u}) d\mathbf{u} / |B|, \quad j = 1, \dots, p \text{ y} \quad (4.3.68)$$

$$\gamma(B, B) \equiv \int_B \int_B \gamma(\mathbf{u} - \mathbf{v}) d\mathbf{u} d\mathbf{v} / |B|^2. \quad (4.3.69)$$

4.3.5.5 Tendencia polinomial

Supóngase a un $\mathbf{s} \in \mathbb{R}^2$ y que $E(Z(\mathbf{s})) = \mu(\mathbf{s})$ es expresada como una combinación lineal de polinomios en el espacio de coordenadas $\mathbf{s} = (x, y)'$. Una superficie de tendencia de grado r es

$$\mu(\mathbf{s}) = \sum_{0 \leq k+l \leq r} \sum \alpha_{kl} x^k y^l, \quad \mathbf{s} = (x, y)'. \quad (4.3.70)$$

Por ejemplo, una superficie de tendencia cuadrática es

$$\mu(\mathbf{s}) = \alpha_{00} + \alpha_{10}x + \alpha_{01}y + \alpha_{20}x^2 + \alpha_{11}xy + \alpha_{02}y^2. \quad (4.3.71)$$

La superficie de tendencia polinomial (4.3.70) tiene $p+1=(r+1)(r+2)/2$ términos. Un caso espacial de (4.3.43) es con

$$\begin{aligned} f_0(\mathbf{s}) &= 1, & \mathbf{s} &= (x, y)' \in D, \\ f_1(\mathbf{s}) &= x, & \mathbf{s} &= (x, y)' \in D, \\ & \vdots & & \\ f_p(\mathbf{s}) &= y^r, & \mathbf{s} &= (x, y)' \in D, \end{aligned} \quad (4.3.72)$$

donde $p = \{(r+1)(r+2)/2\} - 1$. Para $\mathbf{s} \in \mathbb{R}^d$, $p = ((r+d)/d) - 1$.

4.3.5.6 Estimación de la Media Parametral

La estimación óptima de la media parametral β puede ser calculado. De $\mathbf{Z} = \mathbf{X}\beta + \delta$, los datos de \mathbf{Z} satisfacen el modelo lineal general, donde $E(\mathbf{Z}) = \mathbf{X}\beta$ y $\text{Var}(\mathbf{Z}) = \Sigma$. El estimador de minimos cuadrados generalizados es

$$\hat{\beta}_{mcg} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Z, \quad (4.3.73)$$

el cual tiene la propiedad de que la $Var(\hat{\beta}) - Var(\hat{\beta}_{mcg})$ es no negativa definida para todos los estimadores lineales insesgados $\hat{\beta}$. Para estimar la media parametral óptimamente, se requiere conocer $\{Cov(Z(s_i), Z(s_j)): 1 \leq i \leq j \leq n\}$. Sin embargo, para predecir óptimamente un valor desconocido $Z(s_0)$ se requiere conocer $\{Var(Z(s_i) - Z(s_j)); 0 \leq i \leq j \leq n\}$. Tanto la estimación óptima y la predicción óptima, son cercanamente relacionados.

4.3.6 Predicción de Superficie de Tendencia

Cuando se utilizan series de tiempo o datos espaciales conocidos, se asume que la descomposición (4.3.11)

$$Z(\cdot) = \text{Variación_a_Escala_Grande} + \text{Variación_a_Escala_Pequeña}, \quad (4.3.74)$$

no se puede especificar de modo único. Cabe recordar, que la variación a escala grande es sinónimo de tendencia o estructura de medias, cualquier variación de escala pequeña es un proceso aleatorio de media cero.

Por lo que la predicción de una superficie de tendencia esta asociado a un proceso aleatorio subyacente

$$Z(\cdot) = \text{Variación_a_Escala_Grande} + \text{Ruido_Blanco}. \quad (4.3.75)$$

Es decir, toda la variación, a parte de las fluctuaciones pequeñas no correlacionadas es absorbida por la media. Por lo que el proceso aleatorio es representado como

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad (4.3.76)$$

donde la variación a nivel microescala $\eta(\cdot)$ es un proceso de ruido blanco. Es decir,

$$\begin{aligned} E(Z(\mathbf{s})) &= \mu(\mathbf{s}), \quad \mathbf{s} \in D, \\ (Z(\mathbf{s}), Z(\mathbf{u})) &= \begin{cases} c_0, & \mathbf{s} = \mathbf{u}, \\ 0, & \mathbf{s} \neq \mathbf{u}. \end{cases} \end{aligned} \quad (4.3.77)$$

4.3.6.1. Predicción de la superficie de tendencia y predicción del cuadrado medio del error

Si $\mu(\cdot)$ en (4.3.76) puede escribirse como

$$\mu(\mathbf{s}) = \sum_{j=1}^{p+1} \beta_{j-1} f_{j-1}(\mathbf{s}), \quad \mathbf{s} \in D; \quad (4.3.78)$$

el modelo de superficie de tendencia (4.3.76) tiene el mejor estimador lineal insesgado para el siguiente β es

$$\hat{\beta} = (X'X)^{-1} X'Z. \quad (4.3.79)$$

Así, el mejor predictor lineal insesgado de $Z(\mathbf{s}_0)$ es

$$\hat{Z}(\mathbf{s}_0) = \mathbf{x}'\hat{\beta}; \quad (4.3.80)$$

y el predictor del cuadrado medio del error predicho es

$$E(Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0))^2 = (1 + \mathbf{x}'(X'X)^{-1}\mathbf{x})c_0. \quad (4.3.81)$$

4.3.7. Kriging de Mediana Destacada(Polish)

Los datos espaciales pueden ser obtenidos a partir de muestreo parcial de una realización de un proceso aleatorio $\{Z(\mathbf{s}); \mathbf{s} \in D\}$. Dado que la descomposición

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in D, \quad (4.3.82)$$

donde $\mu(\cdot) \equiv E(Z(\cdot))$ es la estructura de media y $\delta(\cdot)$ es la estructura del error. Si $\mu(\cdot)$ es conocida, el predictor óptimo, para kriging simple es

$$Z^*(B) = \sum_{i=1}^n l_i Z(\mathbf{s}_i) + \mu(B) - \sum_{i=1}^n l_i \mu(\mathbf{s}_i), \quad (4.3.83)$$

de $Z(\mathbf{s}_0)$ es

$$Z^*(\mathbf{s}_0) = \mu(\mathbf{s}_0) + \sum_{i=1}^n l_i (Z(\mathbf{s}_i) - \mu(\mathbf{s}_i)), \quad (4.3.84)$$

donde

$$\begin{aligned} \Gamma &\equiv (l_1, \dots, l_n) = \mathbf{c}' \Sigma^{-1}, \\ \mathbf{c} &\equiv (C(\mathbf{s}_0, \mathbf{s}_1), \dots, C(\mathbf{s}_0, \mathbf{s}_n))', \\ \Sigma &\text{ tiene el } (i, j)\text{-ésimo elemento } C(\mathbf{s}_i, \mathbf{s}_j), \quad i, j = 1, \dots, n \text{ y} \\ C(\mathbf{s}, \mathbf{u}) &= \text{Cov}(Z(\mathbf{s}), Z(\mathbf{u})). \end{aligned}$$

Pero en realidad, $\mu(\cdot)$ no es conocido. En dimensiones superiores a uno, se supone que $\mu(\cdot)$ se descompone aditivamente en componentes direccionales. En \mathbb{R}^2 se asume que

$$\mu(\mathbf{s}) = a + c(x) + r(y), \quad \mathbf{s} = (x, y)' \in D. \quad (4.3.85)$$

Adicionalmente, si $\{\mathbf{s}_i: i = 1, \dots, n\}$ están en un entorno $\{(x_l, y_k): k = 1, \dots, p; l = 1, \dots, q\}$, entonces $\mathbf{s}_i = (x_l, y_k)'$ implica que

$$\mu(\mathbf{s}_i) = a + r_k + c_l. \quad (4.3.86)$$

Es decir, el efecto de la hilera r_k puede ser estimado tomando en cuenta la repetición de la otra dimensión; por lo que r_k puede ser estimado de $\{Z(\mathbf{s}_i); \text{ la segunda coordenada de } \mathbf{s}_i \text{ es } y_k; i = 1, \dots, n\}$, donde $k = 1, \dots, p$. Similar consideración permite que el efecto de columna c_l sea estimada con $l = 1, \dots, q$.

Cabe mencionar que los datos espaciales en contornos definidos en \mathbb{R}^2 pueden ser vistos como un cuadro de dos entradas (o múltiples entradas en \mathbb{R}^d). El espaciamiento en el contorno no necesariamente tiene que ser igual en la dirección horizontal o vertical.

4.3.7.1. Análisis de medias

Suponiendo sólo una observación en cada punto (nodo) del entorno, los estimadores de mínimos cuadrados ordinarios son

$$\begin{aligned} \hat{a} &= \sum_{i=1}^n Z(\mathbf{s}_i) / n, \\ \hat{r}_k &= \left\{ \sum_{N(y_k)} Z(\mathbf{s}_i) / q \right\} - \hat{a}, \quad k = 1, \dots, p, \end{aligned} \quad (4.3.87)$$

$$\hat{c}_l = \left\{ \sum_{M(x_l)} Z(s_i) / p \right\} - \hat{\alpha}, \quad l = 1, \dots, q,$$

donde

$$\begin{aligned} N(y_k) &\equiv \{i : s_i = (\cdot, y_k)'; i = 1, \dots, n\}, \\ M(x_l) &\equiv \{i : s_i = (x_l, \cdot)'; i = 1, \dots, n\} \end{aligned} \quad (4.3.88)$$

Entonces, los estimadores de mínimos cuadrados ordinarios de $\mu(s_i)$ es

$$\hat{\mu}(s_i) = \hat{\alpha} + \hat{\rho}_k + \hat{c}_l, \quad s_i = (x_l, y_k)'. \quad (4.3.89)$$

Para $s = (x, y)'$ en la región limitada por las líneas conjuntas de cuatro nodos, $(x_l, y_k)'$, $(x_{l+1}, y_k)'$, $(x_l, y_{k+1})'$ y $(x_{l+1}, y_{l+1})'$, donde $x_l < x_{l+1}$ y $y_k < y_{k+1}$ definen la interpolación planar

$$\begin{aligned} \hat{\mu}(s) &\equiv \hat{\alpha} + \hat{\rho}_k + \left(\frac{y - y_k}{y_{k+1} - y_k} \right) (\hat{\rho}_{k+1} - \hat{\rho}_k) + \hat{c}_l + \left(\frac{x - x_k}{x_{l+1} - x_l} \right) (\hat{c}_{l+1} - \hat{c}_l), \\ &k = 1, \dots, p-1, \quad l = 1, \dots, q-1. \end{aligned} \quad (4.3.90)$$

4.3.7.2. Algoritmo de la mediana destacada (polish)

Este algoritmo que produce todos los efectos de $\tilde{\alpha}$, el efecto de hileras $\{\hat{\rho}_k : k = 1, \dots, p\}$ y el efecto de columnas $\{\hat{c}_l : l = 1, \dots, q\}$ de arreglo $p \times q$ de números $\{Y_{kl} : k = 1, \dots, p; l = 1, \dots, q\}$ que son dados. En el contexto espacial, los datos del contorno $\{Z(s_i) : i = 1, \dots, n\}$ juegan el papel de las Y 's.

El proceso puede iniciar del modo siguiente

Para $i = 1, 3, 5, \dots$ definir

$$\begin{aligned} Y_{kl}^{(i)} &\equiv Y_{kl}^{(i-1)} - \text{Med}\{Y_{kl}^{(i-1)} : l = 1, \dots, q\}, & k = 1, \dots, p+1, l = 1, \dots, q, \\ Y_{k,q+1}^{(i)} &\equiv Y_{k,q+1}^{(i-1)} + \text{Med}\{Y_{kl}^{(i-1)} : l = 1, \dots, q\}, & k = 1, \dots, p+1, \end{aligned} \quad (4.3.91)$$

y para $i = 2, 4, 6, \dots$ definir

$$\begin{aligned} Y_{kl}^{(i)} &\equiv Y_{kl}^{(i-1)} - \text{Med}\{Y_{kl}^{(i-1)} : k = 1, \dots, p\}, & k = 1, \dots, p, l = 1, \dots, q+1, \\ Y_{p+1,l}^{(i)} &\equiv Y_{p+1,l}^{(i-1)} + \text{Med}\{Y_{kl}^{(i-1)} : k = 1, \dots, p\}, & l = 1, \dots, q+1, \end{aligned} \quad (4.3.92)$$

donde la $\text{Med}(y_1, \dots, y_n)$ es la mediana de y_1, \dots, y_n (definido como el número tal que la mitad de los y_1, \dots, y_n están a la izquierda y la otra mitad a la derecha). Para iniciar, suponga que

$$Y_{kl}^{(0)} = \begin{cases} Y_{kl}, & k = 1, \dots, p, \quad l = 1, \dots, q, \\ 0, & \text{de cualquier otro modo.} \end{cases} \quad (4.3.93)$$

Es decir, iniciar con los $p \cdot q$ datos y crear $p+q+1$ celdas (entradas) extras con cero en ellas. Usar (4.3.91) para remover la hilera de medianas de los datos y aumentar la cantidad removida en las p celdas de hileras extra. Hacer lo mismo con las columnas de la tabla, remover la columna de la mediana no sólo de los datos sino también de la columna de hileras acumuladas removidas. Esta última cantidad removerla de la celda extra ($p+1, q+1$). Repetir hasta convergencia. Asumiendo convergencia, los efectos estimados son

$$\begin{aligned} \bar{\alpha} &\equiv Y_{p+1,q+1}^{(\infty)}, \\ \bar{\tau}_k &\equiv Y_{k,q+1}^{(\infty)}, \quad k = 1, \dots, p \\ \bar{\tau}_l &\equiv Y_{p+1,l}^{(\infty)}, \quad l = 1, \dots, q, \end{aligned} \quad (4.3.94)$$

con la propiedad que

$$Y_{kl} = \bar{\alpha} + \bar{\tau}_k + \bar{\tau}_l + Y_{kl}^{(\infty)}, \quad k = 1, \dots, p, l = 1, \dots, q. \quad (4.3.95)$$

Así, el cuadro original $p \times q$ es reemplazado por un cuadro $p \times q$ de residuales $\{Y_{kl}^{(\infty)} : k = 1, \dots, p, l = 1, \dots, q\}$ y las $p+q+1$ celdas adicionales que contienen los efectos de hileras $\{\bar{\tau}_k\}$, los efectos de columna $\{\bar{\tau}_l\}$ y el efecto general $\bar{\alpha}$. Cuando hay diferente número de observaciones en cada nódulo (intersección) del contorno sólo se requiere modificación en la notación. El algoritmo no cambia en las sucesivas remociones de las medianas de hilera y medianas de columna de las entradas del cuadro. Si un total de hileras

(columna) de las de los nodos del contorno no tiene observaciones, entonces la hilera (columna) es ignorada. Así, el algoritmo inicia con la remoción de la hilera, pero debería iniciar con la remoción de la columna. Un criterio de paro es requerido.

4.3.7.3. Kriging basado en residuos

Considerando los residuos de la mediana destacada (polish), $\{R(s_i) : i = 1, \dots, n\}$, como un nuevo conjunto espacial de datos y se puede hacer una apropiada eliminación de la tendencia por kriging ordinario. El análisis de series de tiempo toma exactamente el mismo enfoque antes de ajustar (estimar) un modelo estacionario o de una serie de tiempo.

El comportamiento real de la variación a nivel microescala puede estar alejado de un ruido blanco, pero no hay forma de verlo sin tomar las observaciones mas cercanas.

Basado en el variograma ajustado y en los datos de los residuos $\{R(s_i) : i = 1, \dots, n\}$, las ecuaciones de kriging ordinario y $\lambda_0 = \Gamma_0^{-1} \gamma_0$ producen

$$\hat{R}(s_0) = \sum_{i=1}^n \lambda_i R(s_i), \quad s_0 \in \mathbb{R}^2. \quad (4.3.96)$$

Este kriging esta basado en los residuos que son aproximados para kriging basado en errores no conocidos $\{\delta(s_i) : i = 1, \dots, n\}$.

4.3.8. Datos Geoestadísticos Simulados y Reales

La fuente de datos es importante para el desarrollo de métodos estadísticos en una disciplina relativamente joven como la geoestadística. Debido a que mucha de la investigación estaba relacionada con la minería, la mayoría de la información generada no es proporcionada por el propietario debido a su importancia económica, actualmente hay disciplinas que generan y proporcionan información de datos geoestadísticos. Si los datos reales no son disponibles, uno puede siempre simularlos.

4.3.8.1. Simulación de procesos espaciales

Para datos geoestadísticos, el índice espacial de el proceso subyacente $\{Z(\mathbf{s}): \mathbf{s} \in D\}$ varía continuamente sobre algún subconjunto D d -dimensional de \mathbb{R}^d . La simulación de un proceso completo es usualmente llevado a cabo en contorno reticulado (cuadrículado) discreto (d -dimensional) y entonces aproximarlos a D . Alternativamente, sólo se puede simular $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$ en las n localidades, posiblemente irregularmente espaciales, $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, las cuales pueden ser hechos directamente o por seleccionar valores apropiados por un proceso de simulación completo. Adicionalmente, los métodos de simulación para datos geoestadísticos pueden frecuentemente ser usados para simular datos de látice.

En una dimensión, los datos los datos dependientes son por lo general modelados por procesos autorregresivos de promedios móviles (*ARMA*) de orden finito, el cual es factible de simular. Aunque hay restricciones en la clase de covariograma generado, su principal desventaja es que no se puede generar de modo natural a altas dimensiones.

Usualmente, los métodos de simulación espacial no son los más apropiados para procesos de distribución Gausiana, sin embargo con pequeñas modificaciones se puede dar lugar a procesos aleatorios más generales; lo más obvio proviene de transformar un proceso Gausiano $\{Z(\mathbf{s}): \mathbf{s} \in D\}$ a, por ejemplo, un proceso lognormal $\{Exp(Z(\mathbf{s})): \mathbf{s} \in D\}$ o un proceso *Ji cuadrada* $\{(Z(\mathbf{s}))^2: \mathbf{s} \in D\}$ y así sucesivamente.

Un requerimiento para simulación de cualquier proceso es que todos los parámetros del proceso sean especificados. La fuente de variación aleatorio es por lo general un generador de números pseudo aleatorios, seleccionado con mucho cuidado. Para ello se pueden utilizar procedimientos y métodos como la descomposición Cholesky, métodos espectrales, métodos de bandas cambiantes, etc. El uso de uno u otro dependerá del tamaño de muestra.

4.4. Diseños Espaciales

Para los diseños espaciales, los tratamientos son sobrepuestos en un proceso

aleatorio $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$; el objetivo es estimar los efectos de tratamientos a partir de las respuestas que son espacialmente correlacionados. Así, en el diseño de muestreo espacial, los datos potenciales deben ser observaciones de el proceso aleatorio $Z(\cdot)$; el objetivo es para determinar el tamaño de muestra n y las localidades de muestra $\{s_1, \dots, s_n\}$ de la cual los datos $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))'$ pueden ser usados para producir $g(Z(\cdot))$. Por ejemplo, el interés podría ser la predicción de

$$g(Z(\cdot)) = \int Z(\mathbf{u}) d\mathbf{u} / |B| \quad (4.4.1)$$

donde B es un bloque dado (de volumen $|B|$) en la predicción de $g(Z(\cdot)) = \text{Sup}\{Z(\mathbf{u}) : \mathbf{u} \in B\}$.

4.4.1. Diseño de Muestreo Espacial

Cuando una fuente de contaminación no se puede evitar en una investigación, es prácticamente imposible (financiera y operacionamente) obtener datos de todas las localidades espaciales en el dominio de interés D . Un buen diseño de muestreo espacial reconoce la no homogeneidad espacial de el proceso y lo toma en cuenta cuando se selecciona el tamaño de muestra n y la muestra de localidades $\{s_1, \dots, s_n\} \subset D$. Cabe mencionar que el ajuste de modelos de variograma y kriging han sido utilizados en el diseño de muestreo espacial, en la determinación de las localización de sitios en un marco de muestreo con el propósito de la predicción de puntos o bloques.

Para el modelo de campo aleatorio

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in D, \quad (4.4.2)$$

donde $E(Z(\cdot)) = \mu(\cdot)$ y se supone que sólo una observación potencial es disponible en cualquier localidad espacial. Es decir, el número de repeticiones en cualquier sitio no es considerado como un parámetro de diseño.

En el contexto del diseño espacial de muestreo, la noción de el mejor plan de muestreo es central. Es importante especificar el criterio de optimalidad a ser minimizado, el cual en el contexto estadístico es usualmente alguna medida de cercanía de el estimador (o predictor) a un parámetro desconocido (o dato). Sin embargo, la base, de el diseño estadístico óptimo es primero, la selección de qué es lo que va ha ser estimado o predicho, lo segundo es la selección de el estimador o predictor y finalmente la medida de cercanía que va ser minimizada. Una vez que la primera opción es hecha, la segunda es por medio de la utilización de procedimientos estadísticos usuales, como el mejor estimador lineal insesgado, el mejor predictor lineal insesgado, estimador de mínimos cuadrados generalizados, etc. Por lo general, el cuadrado medio del error, la varianza, la varianza generalizada o el cuadrado medio del error predicho es seleccionado como criterio a ser minimizado.

4.4.1.1. Criterio

En el modelo (4.4.2), $Z(\cdot)$ satisface las suposiciones

$$E(Z(\cdot)) = \mu, \text{ para todo } s \in D \text{ y} \quad (4.4.3)$$

$$\text{Var}(Z(s_1) - Z(s_n)) = 2\gamma(s_1 - s_n), \quad (4.4.4)$$

que lo hacen intrínsecamente estacionario; además $Z(s_0)$ va ser predicho y que las ecuaciones de kriging, $\lambda_o = \Gamma^{-1}\gamma_o$, son utilizadas para obtener el mejor predictor lineal insesgado. Entonces, el criterio del cuadrado medio del error predicho minimizado, es simplemente la varianza de kriging la cual es

$$\sigma_k^2(s_0) = 2 \sum_{i=1}^n \lambda_i \gamma(s_0 - s_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(s_i - s_j), \quad (4.4.5)$$

donde se observa que $\sigma_k^2(s_0)$ depende del número de puntos n usados para resolver las ecuaciones de kriging, en la muestra de localidades de esos puntos y de el variograma, así mismo se observa que no depende de los valores actuales de los puntos observados; esta propiedad hace que kriging sea muy útil para diseñar planes de muestreo espaciales.

En general, se supone que $\mu(\cdot)$ en (4.4.2) no es constante, pero el proceso de error $\delta(\cdot)$ permanecerá intrínsecamente estacionario con media cero. Sea $S \equiv \{s_1, \dots, s_n\}$ que denota un plan de muestreo; entonces minimizar sobre S

$$V_n(S) \equiv \int_D \sigma_k^2(B(s_0); S) \cdot w[s_0; \mu(B(s_0)), \sigma_k^2(B(s_0); S)] ds_0 \quad (4.4.6)$$

o

$$M_n(S) = \text{Sup} \left\{ \sigma_k^2(B(s_0); S) \cdot w[s_0; \mu(B(s_0)), \sigma_k^2(B(s_0); S)] : s_0 \in D \right\} \quad (4.4.7)$$

da una medida de el promedio y la máxima varianza predicha respectivamente. En las dos últimas expresiones anteriores $\sigma_k^2(B(s_0); S)$ denota la varianza de kriging para el bloque $B(s_0)$, localizado en s_0 , usando el plan S. Adicionalmente, $w[\cdot]$ es posible seleccionarlo como una función de ponderaciones no negativa, cuya integral (sobre D) es igual a 1. En la práctica, (4.4.6) y (4.4.7) es difícil de evaluar porque D es una región continua, una versión discreta de ambas es respectivamente dado por

$$V_n(S) \equiv \frac{1}{N} \sum_{j=1}^N \sigma_k^2(B(s_{0j}); S) \cdot w[s_{0j}; \mu(B(s_{0j})), \sigma_k^2(B(s_{0j}); S)] \quad (4.4.8)$$

y

$$M_n(S) = \text{Max} \left\{ \sigma_k^2(B(s_{0j}); S) \cdot w[s_{0j}; \mu(B(s_{0j})), \sigma_k^2(B(s_{0j}); S)] : j = 1, \dots, N \right\}, \quad (4.4.9)$$

donde $\{s_{01}, s_{02}, \dots, s_{0N}\}$ es una región discreta de interés (usualmente la intersección o nodo de un cuadrículado fino en D). También se tiene que minimizar las dos últimas expresiones anteriores.

La comparación óptima de los diseños óptimos de planes de muestreo es realizado suponiendo que la variación de escala grande $\mu(\cdot)$ en (4.4.2) es una combinación lineal desconocida de variables conocidas y que la variación a escala pequeña $\delta(\cdot)$ en (4.4.2) es un proceso de media cero con estructura espacial dependiente conocida. En la práctica este conocimiento por lo general no es disponible, por lo que un estudio de muestreo piloto se debe realizar.

A pesar de las propiedades óptimas de los predictores de kriging, algún predictor no óptimo (no requiere la inversión de una matriz de covarianza o variograma) se puede utilizar. El más común es la media muestral. Otros predictores pueden ser la mediana muestral, promedio móvil, etc.

4.4.1.2. Planes de muestreo espacial

Hay varios planes clásicos de muestreo que pueden ser seleccionados para seleccionar n sitios dentro de una región D : muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo aleatorio por grupos, muestreo aleatorio regular o sistemático y muestreo no aleatorio regular, los cuales son los más comúnmente usados. Estos esquemas de muestreo tienen entre sí ventajas o desventajas cuando se comparan, dependiendo de los patrones de reticulado; éstos pueden ser triangular, cuadrado o hexagonal. También los planes de muestreo regular pueden generar ahorro de cálculos computacionales sobre los planes aleatorizados. Primero, los estimadores del variograma son evaluados en un pequeño número de retrasos en lugares de muchos que el muestreo regular puede generar; en la práctica esta cantidad puede ser agrupada en algunos cuantos, de un modo no del todo satisfactorio o porque esto provoca otra fuente de imprecisión en el estimador del variograma. Segundo, un fuente de “efecto pantalla” puede reducir el número de puntos requeridos para resolver las ecuaciones de kriging; este efecto va a ser más pronunciado en patrones regulares. Los planes de muestreo regular (aleatorio y no aleatorio) son por lo general más eficientes que el muestreo aleatorio simple y estratificado, la eficiencia se mide en términos del promedio de la máxima varianza de kriging. Entre los planes de muestreo regulares, el plan triangular equilátero es más eficiente. Cuando la media es usada para predecir la integral de $Z(\mathbf{u})d\mathbf{u}/|D|$ sobre D , los planes de muestreo regular es el más eficiente.

4.4.1.3. Diseño espacial experimental en regiones discretas

El diseño experimental clásico está basado en los tres conceptos de aleatorización, bloqueo y repetición. La aleatorización neutraliza el efecto de la correlación espacial y produce un análisis de varianza válido; el bloqueo ayuda a reducir la variación de los residuales y las repeticiones la eficiencia de la estimación de el contraste de tratamiento, el

cual depende sólo de la varianza de el error. Sin embargo, en el enfoque clásico, la posición espacial de los tratamientos en el diseño no es considerado; en el contexto espacial, la eficiencia típicamente depende no sólo de la variación del error sino también de su posición. Por lo que es natural entonces, considerar los diseños experimentales espaciales y los métodos de análisis espacial.

Se la región de diseño continua $D \subset \mathbb{R}^d$, la cual es aproximada por un conjunto finito de N sitios o que los N sitios potenciales en los cuales el muestreo puede tomarse y son seleccionados desde D ; el conjunto finito de sitios son $\{s_{01}, \dots, s_{0N}\}$. Es decir, la construcción de un diseño de muestreo espacial óptimo, el tamaño de n se reduce a encontrar los “mejores” n sitios de entre todos los $\binom{N}{n}$ posibles planes de muestreo.

4.4.2. Diseños Experimentales Optimos en Presencia de Dependencia Espacial

Suponiendo que la covarianza para los observaciones de un experimento es conocida, el diseño óptimo para el experimento depende de tres factores: la estimabilidad (el (los) parámetros a ser estimados), el estimador y el criterio de estimabilidad (una medida de la cercanía de el estimador a ser estimado). Dado que en diseños experimentales uno de los principales objetivos es la estimación de los contrastes de tratamientos τ .

Tradicionalmente dos estimadores son considerados en la literatura de diseños experimentales cuando las observaciones son correlacionadas, los cuales son los estimadores de mínimos cuadrados ordinarios y mínimos cuadrados generalizados. Los estimadores de mínimos cuadrados ordinarios son por lo general seleccionados por consideraciones prácticas, es más fácil de calcular que los estimadores de mínimos cuadrados generalizados, ya que no requieren la matriz de covarianza Σ . Sin embargo, debido a que Σ se requiere para obtener la varianza de los estimadores, el mérito de usar mínimos cuadrados ordinarios, estimador menos eficiente, no es tan adecuado.

Si suponemos el modelo siguiente para las respuestas de un experimento:

$$\mathbf{Z} = X_1\alpha + X_2\zeta + \delta, \quad (4.4.10)$$

donde α contiene los t efectos de tratamiento y ζ contiene los $(p - t)$ parámetros de control local (por ejemplo, efectos de bloque, efectos de hilera y columna, etc.) incluyendo el efecto de la media general. Sea $\tau \equiv \Delta\alpha$ el vector de los t contrastes de tratamiento centrados; es decir, $\Delta = I - J/t$, donde $J = \mathbf{1}\mathbf{1}'$ es una matriz $t \times t$ cuyos elementos son todos 1. Sea $\hat{\tau}_{mco}$ el estimador de mínimos cuadrados ordinarios de τ , representado como

$$\hat{\tau}_{mco} = \Delta C(X, I)^{-1} X_1' (I - X_2 (X_2' X_2)^{-1} X_2') \mathbf{Z}, \quad (4.4.11)$$

donde

$$C(X, I) = X_1' (I - X_2 (X_2' X_2)^{-1} X_2') X_1 \quad (4.4.12)$$

y $X \equiv (X_1, X_2)$. La inversa generalizada (por ejemplo A^- de A) es utilizada porque las matrices pueden no ser de rango completo.

Los estimadores de mínimos cuadrados generalizados de τ es dado por

$$\hat{\tau}_{mco} = \Delta C(X, \Sigma)^{-1} X_1' (\Sigma^{-1} - \Sigma^{-1} X_2 (X_2' \Sigma^{-1} X_2)^{-1} X_2' \Sigma^{-1}) \mathbf{Z}, \quad (4.4.13)$$

donde

$$C(X, \Sigma) = X_1' (\Sigma^{-1} - \Sigma^{-1} X_2 (X_2' \Sigma^{-1} X_2)^{-1} X_2' \Sigma^{-1}) X_1, \quad (4.4.14)$$

cuando $\Sigma = I$, $\hat{\tau}_{mco}$ se reduce a $\hat{\tau}_{mco}$.

Sea $D(\hat{\tau}, X, \Sigma)$ la varianza de el estimador $\hat{\tau}$ de τ . La matriz $D(\hat{\tau}, X, \Sigma)$ tiene hileras y columnas suman cero porque Δ tiene hileras y columnas que suman cero. Si $C(\hat{\tau}, X, \Sigma) \equiv C(\hat{\tau}, X, \Sigma)^+$ es la matriz de información para $\hat{\tau}$, donde A^+ denota la inversa Moore-Penrose para la matriz A de números reales, que también tienen hileras y columnas que suman cero, porque $A^+ = A^+ A A^+ = A' (A^+)' A^+ = A^+ (A^+)' A'$. Cuando $\hat{\tau} = \hat{\tau}_{mco}$ sucede

que $C(\hat{\tau}_{mcg}, X, \Sigma) = C(X, \Sigma)$. Así mismo, en el modelo de errores correlacionados, la matriz de información $C(X, \Sigma)$ es la extensión de la matriz C utilizada, por ejemplo, en el modelo de errores no correlacionados.

Dado el estimador, el criterio de optimalidad natural por la cual un diseño puede ser seleccionado es la matriz de covarianza “mas pequeña” de el estimador de τ (la $Var(\hat{\tau})$ es “mas pequeña que” la $Var(\tilde{\tau})$ si la matriz $\{Var(\tilde{\tau}) - Var(\hat{\tau})\}$ sea no nula y no negativa definida). Por lo general, la minimización de alguna función de la matriz de covarianza de el estimador es todo lo que se puede hacer.

Por otra parte, una clase grande de criterios de optimalidad puede ser obtenido por minimización, para un valor fijo de λ , la función siguiente cubre la clase de diseños \mathcal{E}

$$\psi_{\lambda}(C(\hat{\tau}, X, \Sigma)) \equiv \left\{ \frac{1}{t-1} \sum_{k=1}^{t-1} v_k^{-\lambda} \right\}^{1/\lambda}, \quad 0 \leq \lambda \leq \infty, \quad (4.4.15)$$

donde para $\lambda = 0$ y $\lambda = \infty$ la función es definida en el límite y $\{v_k: k = 1, 2, \dots, t-1\}$ y los $t-1$ eigenvalores positivos son de $C(\hat{\tau}, X, \Sigma)$. Esta clase de criterio incluye los criterios de optimalidad $A(\lambda-1)$, optimalidad $D(\lambda = 0)$ y optimalidad $E(\lambda = \infty)$.

4.4.2.1. La función de covarianza y el variograma en diseños experimentales

El enfoque espacial del diseño y análisis de experimentos depende de qué tan bien la estructura de covarianza espacial pueda ser modelada. Una suposición frecuentemente hecha es la existencia de un covariograma, $C(\mathbf{s}_1 - \mathbf{s}_2) \equiv Cov(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))$, $\mathbf{s}_1, \mathbf{s}_2 \in D \subset \mathbb{R}^d$. Una forma alternativa de caracterizar la dependencia estadística entre las observaciones espaciales es a través del variograma $2\gamma(\mathbf{s}_1 - \mathbf{s}_2) \equiv Var(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))$, $\mathbf{s}_1, \mathbf{s}_2 \in D \subset \mathbb{R}^d$. En diseños experimentales espaciales, el variograma hecho de la esperanza de el cuadrado medio de los residuales de el diseño experimental aleatorizado y de las expresiones para el mejor estimador lineal insesgado de los contrastes de tratamiento.

Para el modelo (4.4.10), donde $X \equiv (X_1, X_2)$ satisface que $X1 \propto 1$; sea $\tau \equiv \lambda' \beta$ una función estimable de $\beta \equiv (\alpha', \zeta')$, tal que $\lambda' \beta = 0$. Entonces el mejor estimador lineal insesgado de τ es dado por

$$\hat{\tau} = \lambda' (X' \Gamma^{-1} X)^{-1} X' \Gamma^{-1} Z, \quad (4.4.16)$$

y su varianza es

$$\text{Var}(\hat{\tau}) = -\lambda' (X' \Gamma^{-1} X)^{-1} \lambda, \quad (4.4.17)$$

donde Γ es la matriz del semivariograma, es decir, si las localidades espaciales de los datos son enfatizados por $Z = (Z(s_1), Z(s_2), \dots, Z(s_n))'$, entonces el (i, j) -ésimo elemento de Γ es $\gamma(s_i - s_j)$.

4.4.2.2. Aleatorización para el diseño de bloques completos

Suponiendo que el modelo (4.4.10) es válido y que el experimento es llevado a cabo en $b=r$ bloques, cada uno de tamaño t (para diseños de bloques incompletos, el número de repeticiones de los tratamientos, r es más pequeño que el número b de bloques). Representando la respuesta (por ejemplo, Rendimiento) de la i -ésima unidad experimental (parcela) en el j -ésimo bloque y que recibió el k -ésimo tratamiento como

$$Z_{ij,k} = \mu_{ij} + \alpha_k + W_{ij}, \quad i = 1, \dots, t, \quad j = 1, \dots, r, \quad k = 1, \dots, t, \quad (4.4.18)$$

donde $\{\alpha_k: k = 1, \dots, t\}$ es un efecto de tratamiento aditivo, $\{\mu_{ij}\}$ constituye la tendencia espacial y $\{W_{ij}\}$ es una colección de errores espacialmente correlacionados con media cero.

Equivalentemente,

$$Z_{ij,k} = \{\mu_{..} + W_{..}\} + \{(\mu_{.j} - \mu_{..}) + (W_{.j} - W_{..})\} + \alpha_k + \{(\mu_{ij} - \mu_{.j}) + (W_{ij} - W_{.j})\}, \quad (4.4.19)$$

donde el punto indica un promedio sobre el índice omitido. Por lo que, los componentes μ , $\{\beta_j\}$, $\{\alpha_k\}$ y $\{e_{ij}\}$ son definidos a partir de (4.4.19) por la identificación de términos; es decir,

$$Z_{ij,k} = \mu + \beta_j + \alpha_k + e_{ij}. \quad (4.4.20)$$

El término de bloque β_j es algo artificial excepto cuando $E(e_{ij}) = 0$, es decir, a menos que $i = 1, \dots, r$. Un error con esperanza cero se puede asegurar por la aleatorización dentro de cada bloque. Debido que, aún con la aleatorización, cada unidad experimental puede recibir sólo un tratamiento, el rendimiento del tratamiento k del bloque j es

$$\begin{aligned} Z_{jk} &\equiv \sum_{i=1}^t \delta_{ij,k} Z_{ij,k} = \mu + \beta_j + \alpha_k + \sum_{i=1}^t \delta_{ij,k} e_{ij} \\ &\equiv \mu + \beta_j + \alpha_k + w_{jk}, \end{aligned} \quad (4.4.21)$$

donde

$$\delta_{ij,k} \equiv \begin{cases} 1, & \text{si la parcela } i \text{ en el bloque } j \text{ recibe el tratamiento } k, \\ 0, & \text{de otro modo.} \end{cases}$$

La relación anterior (4.4.21) describe la respuesta al tratamiento k en el bloque j $\{\delta_{ij,k}\}$ son variables aleatorias cuya distribución de probabilidades conjunta es inducida por aleatorización; por lo que cumple lo siguiente:

$$E(w_{ij}) = E_W E_R(w_{kj} | \{\delta_{ij} : i = 1, \dots, t\}) = 0,$$

$$\text{Cov}(w_{jk}, w_{j'k'}) = \begin{cases} \sigma_{w_j}^2, & j = j', k = k', \\ -\sigma_{w_j}^2 / (t-1), & j = j', k \neq k', \\ 0, & \text{de otro modo,} \end{cases} \quad (4.4.22)$$

$$\text{Cov}(\beta_j, w_{jk}) = 0,$$

donde $E_W(\cdot)$ denota la esperanza con respecto a la distribución de probabilidad de el error del proceso aleatorio $\{W_{ij}\}$ y $E_R(\cdot)$ denota la esperanza con respecto a la distribución de

aleatorización. También

$$\sigma_{w_j}^2 \equiv (1/t) \sum_{i=1}^t E_W(e_{ij}^2), \quad (4.4.23)$$

donde

$$E_W(e_{ij}^2) = (\mu_{ij} - \mu_{.j})^2 + E_W(\delta_{ij} - \delta_{.j})^2. \quad (4.4.24)$$

De la parte dos de (4.4.22) se desprende que la aleatorización no elimina la correlación espacial; aunque algo se neutraliza por el cambio a donde hay una pequeña correlación intrabloque negativo. Sea \mathbf{Z} que denota la respuesta en (4.4.21) y $\Sigma \equiv \text{Var}(\mathbf{Z})$; Σ no es proporcional a la matriz identidad y no es difícil verificar las condiciones de Zyskind, por lo que los estimadores de mínimos cuadrados ordinarios de cualquier contraste de tratamiento τ es también el mejor estimador lineal insesgado.

Si la medida del error esta presente, entonces la respuesta $\{Z_{jk}\}$ satisface $Z_{jk} = \mu + \beta_j + \alpha_k + w_{jk} + \varepsilon_{jk}$, donde $\{\varepsilon_{jk}\}$ es una colección de variables aleatorias independientes e idénticamente distribuidas con varianza σ_ε^2 e independientes de $\{w_{jk}\}$. Este modelo lineal para el diseño de bloques completos aleatorizados tiene una forma conveniente para el análisis de varianza, como se ve en el Cuadro 4.1.

Cuadro 4.1. Análisis de varianza para el diseño de bloques completos aleatorizados.

Fuente	Suma de Cuadrados	Esperanza de Cuadrados Medios
Bloque	$t \sum_{j=1}^r (Z_{.j} - Z_{..})^2$	$\frac{t}{r-1} \sum_{j=1}^r E_W(\beta_j^2) + \sigma_\varepsilon^2$
Tratamiento	$r \sum_{k=1}^t (Z_{k.} - Z_{..})^2$	$\frac{r}{t-1} \sum_{k=1}^t \alpha_k^2 + \frac{t}{r(t-1)} \sum_{j=1}^r \sigma_{w_j}^2 + \sigma_\varepsilon^2$
Residual	$\sum_{k=1}^t \sum_{j=1}^r (Z_{kj} - Z_{k.} - Z_{.j} + Z_{..})^2$	$\frac{t}{r(t-1)} \sum_{j=1}^r \sigma_{w_j}^2 + \sigma_\varepsilon^2$
Total	$\sum_{k=1}^t \sum_{j=1}^r (Z_{kj} - Z_{..})^2$	

Donde las esperanzas de los cuadrados medios se calculan sobre la distribución de probabilidades inducida por la aleatorización, la distribución del componente de escala pequeña $\{W_{ij}\}$ y la distribución de la medida del error $\{e_{jk}\}$; el enfoque usual de aleatorización considera los datos como fijos y calcula las esperanzas bajo la distribución de aleatorización inducida. En cualquier caso, bajo la hipótesis nula de igual efecto de tratamientos, la esperanza de cuadrados medios de tratamiento es igual a la esperanza del cuadrado medio de los residuos; es decir, la aleatorización provee pruebas válidas para la prueba de hipótesis del efecto de igualdad de tratamientos.

Debido a que la esperanza del cuadrado medio de los residuales es un promedio sobre todas las posibles permutaciones generador por el procedimiento de aleatorización, debe ser posible encontrar el conjunto de diseños sistemáticos o restringidos de diseños aleatorizados con la más pequeña esperanza del cuadrado medio de los residuales que bajo el procedimiento no restringido o con la misma esperanza del cuadrado medio de los residuales pero con menos variabilidad. Es decir, en presencia de correlación espacial, el diseño de bloques completos aleatorizados puede ser inferior a otros diseños.

4.4.2.3. Eficiencia de los diseños de bloques con errores en un proceso AR de segundo orden

Sea $\mathcal{X}_{t,b,k}$ una clase de diseños de bloques incompletos con t repeticiones de tratamientos de tratamientos y b bloques de tamaño k , de tal manera que cada tratamiento aparece una vez en cada uno de los r bloques; es decir $r = bk/t$. Considerando el modelo

$$\mathbf{Z} = \mu \mathbf{1}_n + X_a \boldsymbol{\alpha} + X_b \boldsymbol{\beta} + \boldsymbol{\delta}, \quad (4.4.25)$$

donde

$\mathbf{Z} \equiv (Z_{11}, Z_{12}, \dots, Z_{1k}, Z_{21}, \dots, Z_{bk})'$ es el vector de datos de dimensión $n=bk$ y Z_{ij} es la respuesta de la i -ésima unidad en el j -ésimo bloque.

μ representa la media general.

$\boldsymbol{\alpha}$ representa los t efectos fijos de tratamientos desconocidos.

$\boldsymbol{\beta}$ representa los b efectos fijos de bloques desconocidos.

$\boldsymbol{\delta}$ es un vector aleatorio con media cero y la estructura de dependencia es una realización parcial de un proceso autoregresivo de segundo orden $[AR(2)]$.

$\mathbf{1}_n$ es un vector de dimensión n con todos los elementos igual a 1, X_a es la matriz de

incidencia de tratamientos de dimensión $n \times t$ y $X_b \equiv I_b \otimes \mathbf{1}_k$ es la matriz de incidencia de bloques de dimensión $n \times b$ (I_b es la matriz identidad de orden b y \otimes denota el producto Kronecker); $X \equiv (\mathbf{1}, X_a, X_b)$ es definida como la matriz diseño, donde para facilitar la notación, el subíndice en $\mathbf{1}_n$ se eliminó.

Si suponemos que las observaciones de diferentes bloques son no correlacionados y que dentro de cada bloque j el proceso de error $\{\delta_{ij}; i=1, \dots, k\}$ es una realización parcial de un proceso autorregresivo de segundo orden, el cual es definido como

$$\delta_{ij} - \rho_1 \delta_{i-1,j} - \rho_2 \delta_{i-2,j} = \varepsilon_{ij}, \quad i = 0, \pm 1, \pm 2, \dots, \pm \infty \quad (4.4.26)$$

donde los ε_{ij} s son variables independientes e idénticamente distribuidos con varianza constante σ^2 y específica a $Var(\mathbf{Z}) \equiv \Sigma$.

Las ecuaciones normales reducidas de mínimos cuadrados generalizados para el vector de efectos de tratamientos α son

$$X_a' W X_a \alpha = X_a' W Z, \quad (4.4.27)$$

donde

$$W = I_b \otimes (\Sigma^{-1} - (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} \Sigma^{-1} \mathbf{1}' \Sigma^{-1}). \quad (4.4.28)$$

Considere los contrastes de tratamiento estimables, $\tau = (I - \mathbf{1}'/\mathbf{1})\alpha$; es decir $\tau_i = \alpha_i - (1/t)(\alpha_1 + \alpha_2 + \dots + \alpha_t)$. Sea $(X_a' W X_a)^+$ la inversa Moore-Penrose de $(X_a' W X_a)$.

Entonces

$$\begin{aligned} \hat{\tau}_{mcg} &= (I - \mathbf{1}'/\mathbf{1})(X_a' W X_a)^+ X_a' W Z \\ &= (X_a' W X_a)^+ X_a' W Z. \end{aligned} \quad (4.4.29)$$

Además,

$$Var(\hat{\tau}_{mcg}) = (X_a' W X_a)^+ \quad (4.4.30)$$

y de aquí, la cantidad

$$C(\hat{\tau}_{mcg}, X, \Sigma) \equiv X_a' W X_a \quad (4.4.31)$$

es la matriz de información para $\hat{\tau}_{mcg}$.

Debido a que $C(\hat{\tau}_{mcg}, X, \Sigma)$ tiene hileras y columnas que suman cero y la $Tr(C(\hat{\tau}_{mcg}, X, \Sigma))$ es constante para cada diseño en $\mathcal{E}_{t,b,k}$, una condición suficiente para que un X sea óptimamente universal sobre $\mathcal{E}_{t,b,k}$ es porque la matriz de información correspondiente sea completamente simétrica, para lo cual deben de cumplir las 6 condiciones de suficiencia (bajo los modelos (4.4.25) y (4.4.26)); por ejemplo, para $k=4$ y $l \neq m=1, 2, \dots, t$ las condiciones son

$$\lambda_{l,m} = \lambda \equiv \frac{r(k-1)}{t-1}, \quad (4.4.32)$$

$$N_{l,m}^{(1)} = \frac{2\lambda}{k}, \quad (4.4.33)$$

$$N_{l,m}^{(2)} = \frac{2b(k-2)}{t(t-1)}, \quad (4.4.34)$$

$$e_{l,m}^* = f_{l,m}^* = \frac{4(k-1)b}{t(t-1)}, \quad (4.4.35)$$

$$\sum_{j=1}^b f_{j,l} f_{j,m} = \sum_{j=1}^b e_{j,l} e_{j,m} = \frac{2b}{t(t-1)} \text{ y} \quad (4.4.36)$$

$$\sum_{j=1}^b N_{j,l,m}^{(1)} (e_{j,l} f_{j,m} + e_{j,m} f_{j,l}) = 1/2 \sum (e_{j,l} f_{j,m} + e_{j,m} f_{j,l}) = \frac{4b}{t(t-1)}, \quad (4.4.37)$$

donde

$$N_{j,l,m}^{(1)} = \begin{cases} 1, & \text{si los tratamientos } l \text{ y } m \text{ son adjuntos en el bloque } j, \\ 0, & \text{de otro modo,} \end{cases}$$

$$e_{j,l} = \begin{cases} 1, & \text{si el tratamiento } l \text{ es aplicado a un borde en el bloque } j, \\ 0, & \text{de otro modo,} \end{cases}$$

$$f_{j,l} = \begin{cases} 1, & \text{si el tratamiento } l \text{ es aplicado a la parcela proxima al borde de el bloque } j, \\ 0, & \text{de otro modo,} \end{cases}$$

$N_{l,m}^{(h)}$ es el número de bloques en la cual los tratamientos l y m ocurren como el h -ésimo vecino más cercano,

$\lambda_{l,m}$ es el número de bloques en la cual los tratamientos l y m ocurren juntos,

$e_{l,m}^*$ es el número de bloques en la cual los tratamientos l y m ocurren juntos y l o m son aplicados a una parcela en el borde y

$f_{l,m}^*$ es el número de bloques en la cual los tratamientos l y m ocurren juntos y l o m son aplicados a la parcela próxima al borde.

4.4.2.4. Diseño de bloques balanceados de vecinos más cercanos de primer orden

El diseño de bloques óptimo universal para un proceso de error $AR(2)$ puede no existir para la selección particular de t , b y k . En la quinta condición de suficiencia, $2b/t(t-1)$ debe ser un entero. Es decir, el número de bloques tiene que ser un múltiplo de $t(t-1)/2$, el cual tiene que ser grande. Sin embargo, el diseño de bloques balanceados de vecinos más cercanos de primer orden típicamente requiere de un número más pequeño de bloques. Un diseño de bloques es llamado de vecinos más cercanos balanceados de primer orden si el número de veces que dos tratamientos aparecen adjuntos en un bloque es constante para cada par de tratamientos, es decir, la condición 2, $N_{l,m}^{(1)}$, de las condiciones de suficiencia. En este contexto, cualquier alejamiento pequeño del modelo de errores no correlacionados puede causar que el diseño de bloques completos aleatorizados sea altamente ineficiente.

El diseño de bloques completos aleatorizados tiene ventajas prácticas sobre otros diseños: es muy fácil su construcción, la ineficiencia por el uso de este diseños simple puede ser mitigado por el uso del análisis espacial; para ello se tiene que proporcionar la localización de los tratamientos dentro de los bloques.

4.4.2.5. Ensayos de campo

Aunque su empleo original fue en el contexto agrícola, el propósito general de los ensayos de campo es para comparar los efectos de un determinado número de tratamientos aplicados a un conjunto de unidades experimentales (la unión de unidades experimentales espacialmente contiguas es lo que constituye un campo). La obtención de estimadores precisos de efectos de tratamiento de primordial importancia para reducir o controlar la variación residual que no es debida a los efectos de tratamiento, lo cual se puede lograr en dos formas. La primera, y el mas común enfoque, es el uso de diseños que incrementan la complejidad, como por ejemplo, diseños de bloques, diseños de bloques incompletos,

diseños de látice, etc. El segundo enfoque, el cual no es incompatible con el primero, es el uso de modelos y análisis alternativos que toman en cuenta la heterogeneidad espacial, dado que el objetivo principal es una eficiente estimación de los efectos de tratamiento.

4.4.3. Análisis de Vecinos Más Cercanos

Papadakis (1937) fue el primero en presentar el método de análisis de vecinos más cercanos, sin embargo, la investigación de las propiedades teóricas fue iniciada 30 años posteriormente (Atkinson, 1969). Las posteriores contribuciones a este método se presenta en Bartlett (1978), Wilkinson *et al.* (1983), Papadakis (1984), Green *et al.* (1985), Besag y Kempton (1986), Gleeson y Cullis (1987), Zimmerman y Harville (1989), entre otros.

4.4.3.1. Estimación de efectos de tratamiento

Sea \mathbf{Z} que denota el *vector de respuestas* de un ensayo de campo y suponiendo que el diseño experimental empleado permite escribir

$$\mathbf{Z} = X_1\alpha + \mathbf{W} + \varepsilon, \quad (4.4.38)$$

donde α es un vector de contrastes de tratamiento y $(\mathbf{W} + \varepsilon)$ es la variación espacial subyacente. \mathbf{W} es un vector de media cero (espacialmente dependiente) que representa la variación a escala pequeña suavizada (por ejemplo, variación de fertilidad) y es independiente de ε , el cual es un vector de media cero de errores idénticamente distribuidos no correlacionados. Green (1985) generaliza este modelo a

$$\mathbf{Z} = X_1\alpha + X_2\zeta + \mathbf{W} + \varepsilon, \quad (4.4.39)$$

donde $X_2\zeta$ representa no homogeneidad espacial de escala grande, los elementos de ζ son efectos de bloque.

Sea $\mathbf{W}^{(l)}$ el valor actual de \mathbf{W} . Por lo tanto un estimador de mínimos cuadrados ordinarios para α es

$$\alpha^{(l+1)} = (X_1'X_1)^{-1} X_1'\{\mathbf{Z} - \mathbf{W}^{(l)}\}; \quad (4.4.40)$$

es decir, de $\mathbf{Z} - \mathbf{X}_1 \boldsymbol{\alpha}^{(i+1)}$, un subconjunto indicador de \mathbf{W} puede ser encontrado; Papadakis propone resolver

$$\mathbf{W}^{(i+1)} = C(\gamma)(\mathbf{Z} - \mathbf{X}_1 \boldsymbol{\alpha}^{(i+1)}) \quad (4.4.41)$$

y usarlo como valor inicial, es decir, $\mathbf{W}^{(0)} = (0, \dots, 0)'$. En la matriz $C(\gamma)$ de tamaño $n \times n$ refleja similaridad en las unidades experimentales; en una dimensión $C(\gamma)$ podría promediar los residuos en las unidades vecinas más cercanas (excepto en los bordes, donde sólo un vecino más cercano es usado) y multiplicar el resultado por $\gamma \in \mathbb{R}$. Iterando (4.4.40) y (4.4.41) se genera el método de Papadakis de vecinos más cercanos iterado, el cual propone detenerse en $\boldsymbol{\alpha}^{(2)}$, por lo que la estimación final $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^{(\infty)}$, satisface

$$\mathbf{X}_1'(I - C(\gamma))\mathbf{X}_1 \hat{\boldsymbol{\alpha}} = \mathbf{X}_1'(I - C(\gamma))\mathbf{Z}. \quad (4.4.42)$$

Suponiendo que el vector $k \times 1$ de parámetros γ en $C(\gamma)$ es conocido, la relación (4.4.42) es un estimador de mínimos cuadrados generalizados para $\boldsymbol{\alpha}$ en el modelo lineal

$$E(\mathbf{Z}) = \mathbf{X}_1 \boldsymbol{\alpha}, \quad \text{Var}(\mathbf{Z}) = (I - C(\gamma))^{-1} \tau^2. \quad (4.4.43)$$

El componente de error en (4.4.38) es,

$$\boldsymbol{\delta} \equiv \mathbf{W} + \boldsymbol{\epsilon}, \quad (4.4.44)$$

puede ser identificado con media cero condicionado a un modelo espacial Gaussiano específico. Por lo que el vector γ puede ser interpretado como un conjunto espacial de parámetros de variación a escala pequeña y se asume que es conocido.

4.4.3.2. Caso de una dimensión y caso de dos dimensiones

Por que en el campo, las unidades contiguas son con frecuencia franjas de parcelas largas y angostas o son hechas de parcelas de bloques adjuntas que tienen más longitud horizontal que vertical, el caso de dependencia espacial de una dimensión recibe considerable atención. Gleeson y Culis (1987) han llevado a cabo todos los enfoques juntos

de una dimensión suponiendo que la dependencia en \mathbf{W} se comporta de acuerdo a un proceso de series de tiempo $ARIMA(p, m, q)$. Para ello se supone que $\{W(s): s \in \mathbf{Z}\}$ es un proceso aleatorio de media cero en los enteros, donde el índice s denota las localidades espaciales de una unidad experimental en una franja de una dimensión. El operador de retraso B definido por $B\delta(s) \equiv \delta(s-1)$ y por consecuencia $B^2\delta(s) \equiv \delta(s-2)$ y así sucesivamente. Sea $\{a(s): s \in \mathbf{Z}\}$ que denota a una sucesión de variables aleatorias independientes e idénticamente distribuidos con media cero y varianza σ_a^2 , constante. Entonces $W(\cdot)$ es un $ARIMA(p, m, q)$ y puede ser representar como

$$\{1 - \phi_1 B - \dots - \phi_p B^p\} (1 - B)^m W(s) = (1 - \theta_1 B - \dots - \theta_q B^q) a(s), \quad (4.4.45)$$

donde las raíces de las ecuaciones $1 - \phi_1 y - \dots - \phi_p y^p = 0$ y $1 - \theta_1 y - \dots - \theta_q y^q = 0$ están fueran del círculo unitario y de las ecuaciones no tienen raíces en común. Debido a que los elementos de ε en (4.4.44) se asumen independientes e idénticamente distribuidos, entonces $(1-B)^m(W(s)+\varepsilon(s))$ es un $ARIMA(p, 0, Q)$, donde $Q = \text{Max}(p+q, m)$; el método original propuesto por Papadakis supone $m=0$, muchos de los análisis de vecinos más cercanos supone algún orden de diferenciación, esto último se presenta en Ripley (1981), Wilkinson *et al.* (1983), Green *et al.* (1985), Besag y Kempton (1986), Williams(1986) y Gleeson y Culis (1987). Por su parte Papadakis (1937), Bartlett (1938 y 1978) y Atkinson (1969) no llevan a cabo diferenciación.

En dos dimensiones, se puede asumir un modelo espacial análogo al de una dimensión, donde no se lleva a cabo diferenciación [$ARIMA(1, 0, 0)$]:

$$E(\delta(s) | \{\delta(u): u \neq s\}) = \gamma \{\delta(s-1) + \delta(s+1)\}, \quad 2|\gamma| < 1, \quad (4.4.46)$$

la cual, junto con la suposición de un proceso Gaussiano en la innovación del proceso $a(\cdot)$ en (4.4.45) implica que

$$\delta \sim \text{Guasiano}(\mathbf{0}, (I - \gamma H)^{-1} \tau^2); \quad (4.4.47)$$

por lo que para este caso el modelo espacial es

$$E(\delta(u, v) | \{\delta(k, l) : (k, l) \neq (u, v)\}) = \gamma_1 \{\delta(u-1, v) + \delta(u+1, v)\} + \gamma_2 \{\delta(u, v-1) + \delta(u, v+1)\}, \quad 2|\gamma_1| + 2|\gamma_2| < 1, \quad (4.4.48)$$

el cual, bajo el supuesto de homocedasticidad Gaussiana de la distribución condicional de $\{\delta(u, v)\}$, implica que $\delta(\cdot)$ es un campo aleatorio Markov Gaussiano y que

$$\delta \sim \text{Gusiano}(\mathbf{0}, (I - C(\gamma))^{-1} \tau^2), \quad (4.4.49)$$

donde $C(\gamma)$ es definido por la comparación de la esperanza condicional $E(Z(s_i) | \{z(s_j) : j \neq i\}) \equiv \theta_i(\{z(s_j) : j \neq i\})$ con (4.4.48) y $Z = \text{Gusiano}(\mu, (I - C)^{-1} M)$ con (4.4.49).

4.4.3.3. Estimación de los parámetros γ y de variación de escala pequeña

En el caso de una dimensión, Papadakis (1937) propuso un análisis de covarianza sobre los datos \mathbf{Z} , con la covariable $H(\mathbf{Z} - X_1(X_1'X_1)^{-1}X_1'\mathbf{Z})$, donde H es una matriz simétrica $n \times n$ con el (i, j) -ésimo elemento h_{ij} , tiene los valores siguientes: $h_{i, i-1} = 1$ ($i=2, \dots, n$), $h_{i, i+1} = 1$ ($i=1, \dots, n-1$) y $h_{ij} = 0$ de otro modo. El parámetro de variación a escala pequeña γ representa el coeficiente de regresión de los datos sobre la covariable y de aquí

$$\hat{\gamma} = \frac{\mathbf{Z}' H (I - P_1 \mathbf{Z})}{\mathbf{Z}' (I - P_1) H' H (I - P_1) \mathbf{Z}}, \quad (4.4.50)$$

donde $P_1 = X_1(X_1'X_1)^{-1}X_1'$. Este estimador es ineficiente cuando se compara con los estimadores de máxima verosimilitud resultante de el modelo

$$\mathbf{Z} \sim \text{Gusiano}(X_1 \alpha, (I - \gamma H)^{-1} \tau^2). \quad (4.4.51)$$

4.4.4. Análisis Basado en Modelación Espacial

Todos los análisis de vecinos más cercanos pueden ser basados en un modelo lineal espacial. Considerando una generalización de el modelo (4.4.38), el cual permite no homogeneidad espacial de escala grande. Sea

$$\mathbf{Z} = X\boldsymbol{\beta} + \boldsymbol{\delta} , \quad (4.4.52)$$

donde un subconjunto de los elementos de los efectos fijos $\boldsymbol{\beta}$ son los efecto de tratamiento α y $\boldsymbol{\delta}$ es un vector de error de media cero que es espacialmente correlacionado. Suponiendo que $Var(\boldsymbol{\delta}) = \Sigma(\gamma)$.

En la expresión anterior $\Sigma(\gamma)$ es la matriz de varianzas de los datos de látice regional que resulta de integrar el modelo soportado en puntos con índice espacial continuo.

Una estimación comúnmente usada para estimar la varianza de los efectos estimados es dado por

$$\hat{Var}(\boldsymbol{\beta}_{mce}) = (X' \Sigma(\hat{\gamma}) X)^{-1} , \quad (4.4.53)$$

pero esto no toma en cuenta la estimación de γ en $\boldsymbol{\beta}_{mce}$. Est fuente de variabilidad extra da la inadecuada condición de los modelos basados en los errores estándar.

Los estimadores de máxima verosimilitud restringida para los parámetros de variación de pequeña escala γ tienen superior sesgo para pequeñas muestras y al parecer producen un valor de F para las pruebas de hipótesis sobre tratamiento.

Un análisis de datos basados en experimentos de campo y bajo un modelo espacial adopta la cantidad y calidad de la dependencia espacial mostrada por los datos y explota la dependencia para obtener estimadores más precisos de los parámetros de variación de escala grande, como son los efectos de tratamiento (Besag y Kempton, 1996; Cullis y Gleeson, 1991; Grondona y Cressie, 1991; Zimmerman y Harville, 1991;). Esto es un ejemplo del por qué, de la estadística para datos espaciales, debe tener un lugar en el repertorio de los métodos utilizados por los estadísticos.

V.

MATERIALES Y MÉTODOS

5.1. Descripción de los Datos Experimentales Utilizados

Para esta investigación se colectaron 12 conjuntos de datos provenientes de 5 experimentos en blanco o ensayos de uniformidad que presenten de modo preciso la ubicación en el espacio de las unidades experimentales; se utilizaron para el análisis de simulación solo cinco de ellos debido a la reducción de los datos, por análisis de un sólo conjunto de datos de los experimentos que presentan mas de uno o por ser redundantes. Los cultivos fueron de alfalfa, cebada, chari (gramínea para pasto), remolacha azucarera y trigo, y realizaron en diversas áreas. La mayoría de los detalles de dimensiones se presentan en las unidades de medida originales (sistema métrico inglés) y sus respectivas conversiones al sistema métrico decimal, a los cuales se les resume información de sus características y procesos en el Cuadro 5.1.

Baker *et al.* (1952) llevaron a cabo el ensayo de uniformidad de una variedad de cebada durante 12 años interrumpidos, de 1924-1935 (en los años 1927 y 1931 no se cultivo). Se utilizó un área rectangular orientada, se fraccionó en tres subdivisiones (bloques) orientados, separados por 15 pies (4.572 m) entre ellos y con 19 parcelas por cada uno, dando lugar a 57 unidades experimentales en total; cada una tiene 161 pies x 43.55 pies, (48.768 m x 13.274 m) aproximadamente 0.1 de acre (405 m²), no se indicó separación entre unidades experimentales en el sentido vertical. Se evaluó el rendimiento de cebada durante 10 años respectivamente, dando lugar a un total de 570 observaciones (Cuadro 9.1a del Anexo).

Immer (1932) realizó un ensayo de uniformidad en la que utilizó un área de 0.9 acres (3645 m²) orientada; fue plantada igual cantidad semilla de remolacha azucarera de la

variedad "Pioneer" en surcos, su manejo anterior al cultivo fue homogéneo durante varios años. Las unidades experimentales se obtuvieron de 60 surcos (hileras) de 350 pies (106.680 m) de largo, las cuales se subdividieron en 10 series de parcelas de 33 pies (10.058m) o 2 rods con 2 pies (0.610 m) de separación vertical entre cada una, la separación (horizontal) entre hileras fue de 22 pulgadas (0.559 m). Se evaluaron tres variables: el Rendimiento de remolacha azucarera (variable utilizada en la simulación) en libras de 600 parcelas de hilera simple (Cuadro 9.2a del Anexo), el Porcentaje de Azúcar en las muestras de remolachas (Cuadro 9.2c del Anexo) y Porcentaje de Pureza Aparente en las muestras de remolachas (Cuadro 9.2d del Anexo).

Lander *et al.*(1938) realizaron cuatro ensayos de uniformidad donde, en cada caso, utilizaron un área mayor que consistió de cinco lotes contiguos denominados D₄, D₅, D₆, D₇ y D₈, cada uno de 5 acres (2.025 ha) y con separación de 12 pies, sólo entre los bloques D₈-D₇ y D₆-D₅. Cada uno de estos cinco lotes anteriores fue subdividido en 3 lotes denominados A, B y C separados (horizontalmente) por dos franjas de terreno contiguas de 5 pies (1.524 m) de ancho, cada lote de las subdivisiones (A, B y C) se fraccionó en 13 unidades experimentales separadas horizontalmente por franjas de terreno de 5 pies (1.524 m) de ancho; las dimensiones de las unidades experimentales son de 207 pies 5 pulgadas (63.221 m) de largo por 19 pies 1 pulgada de ancho (5.816 m). El área se utilizó durante 4 temporadas en dos diferentes cultivos; trigo en los ciclos I (Cuadro 9.3a del Anexo), III (Cuadro 9.3c del Anexo) y IV (Cuadro 9.3d del Anexo) y chari en el ciclo II (Cuadro 9.3b del Anexo). Se evaluó el rendimiento de semilla en "maunds". Se utilizó la parte correspondiente a los ciclos III y V para la simulación.

Mercer y Hall (1911) presentan los datos de un ensayo de uniformidad que consisten de un látice (arreglo) de 20 x 25 de parcelas orientadas (20 hileras en dirección este a oeste por 25 columnas en dirección norte a sur). Las dimensiones de cada parcela fueron de 10.82 pies x 8.05 pies (3.298 m x 2.454 m) aproximadamente, se supone que con espacio de quizás un surco entre parcelas en el sentido vertical y horizontal (no se indica la distancia); la unidad experimental en último sentido indicado estaba constituida por 11

surcos. Se presentan los rendimientos de granos de trigo en libras para 500 parcelas (Cuadro 9.4 del Anexo).

Wassom y Kalton (1953) presentan tres conjuntos de datos de obtenidos, fueron de tres experimentos de uniformidad en tres diferentes áreas orientadas y homogéneas, fueron cultivadas con variantes de cultivo para forraje. En cada campo el área total fue de 126 pies x 168 pies, (38.400 m x 51.210 m) con parcelas de 4 pies x 3.5 pies (1.219 m x 1.067 m) con una separación de 3.5 pies (1.067 m) en forma horizontal y 0.667 pies (8 pulgadas) en el sentido vertical. Treinta y seis parcelas fueron cosechadas en cada dirección, siendo un total de 1296 parcelas por cada uno de los terrenos. El peso húmedo de forraje en libras se evaluó al realizar cortes de material vegetal a 2 pulgadas de la superficie. Las parcelas fueron sembradas con pasto para forraje (Cuadro 10.5a del Anexo) la cual se utilizó para la simulación, pasto para semilla (Cuadro 10.5c del Anexo) y mezcla de pasto y alfalfa en la misma proporción (Cuadro 10.5d del Anexo).

Las características de los conjuntos de datos seleccionados se observa en el Cuadro 5.1.

Cuadro 5.1.-Características de los conjuntos de experimentos en blanco.

Referencia	Conjunto de datos	Dimensiones en Pies (Largo x Ancho)	Número de Parcelas	Forma del cultivo	Suma
Baker <i>et al.</i> (1952)	1	161.00 x 43.55	570	Área	2
Immer (1932)	3	33.000 x 1.833	600	Surco	3
Lander <i>et al.</i> (1938)*	4	207.417 x 19.083	195	Área	2
Mercer y Hall (1911)	1	10.82 x 8.05	500	Surco	11
Wassom y Kalton (1953)	3	4.0 x 3.5	1296	Surco	2

* Datos ya reducidos en el artículo publicado.

La suma indicada en el Cuadro 5.1 se realizó con la finalidad de tener más variación en los datos; para los datos de Baker *et al.*(1952) la suma realizada es para el área de dos años consecutivos; en Immer (1932) se sumó la información de tres surcos contiguos por hilera (horizontalmente); para los datos de Lander *et al.* (1938) se realizó la suma de las dos últimas temporadas para el mismo cultivo; con los datos de Mercer y Hall (1911) no se realizó la suma de los once surcos debido a que los datos ya estaban sumados; en los datos

de Wassom y Kalton (1953) se sumó los resultados de dos surcos contiguos en el sentido horizontal.

Al conjunto de datos seleccionados, en forma de matrices, se le calcularan algunas estadísticas básicas mediante en modulo matricial (IML) del sistema estadístico SAS (Release. 6.12 for Windows 9X).

Al concentrar la información de las variables de respuesta, según sea el caso, se obtuvieron los datos nuevos indicado en el Cuadro 5.2 en donde se observa un total de 10 variables evaluadas, las dimensiones (largo x ancho) varían de 4.0 a 414.83 pies, el número de unidades experimentales varía de 144 a 648 y las unidades sumadas son 2 años (áreas) o surcos (3 a 11).

Cuadro 5.2. Nuevas características de los conjuntos de experimentos en blanco.

Origen	Variables Evaluadas	Dimensiones (pies)	Unidades Experimentales	Unidades Sumadas
Baker <i>et al.</i> (1952)	1	322.0 x 87.1	285	2 años
Immer (1932)	3	33.0 x 5.5	200	3 surcos
Lander <i>et al.</i> (1938)	1	414.83 x 36.17	195	2 años
Mercer y Hall (1911)	1	10.85 x 8.05	500	11 surcos
Wassom y Kalton (1953)	3	4.0 x 7.0	648	2 surcos

Los segmentos seleccionados (7 x 4) del experimento en blanco para Baker *et al.* (1952), Immer (1932), Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton (1953) estan en los Cuadros 9.1c, 9.2b, 9.3e, 9.4 y 9.5b respectivamente del Anexo en CD.

5.2. Metodología de Análisis Espacial Alternativo

Para ensayos de uniformidad se ha establecido que el rendimiento de las parcelas en campo (o alguna otra característica medible en la parcela, tal como el porcentaje de materia orgánica, número de frutos, altura de la planta, etc.) esta compuesto de dos partes: una directamente relacionada con la posición de la parcela en el campo y otra no relacionada con la posición. La primera parte de la variación es atribuida a un gran número de factores como es la cantidad de malezas, distribución de plagas y enfermedades, dirección de los vientos y de la fertilidad del suelo. La segunda parte de la variación es conocida como Error

Puro y puede ser debido a errores de medición más un gran número de pequeños efectos en escala local.

Para los experimentos de campo se acostumbra la notación de Y_{ij} o Y_{kl} , para representar variable de respuesta de una parcela individual. Cuando se requiere referirse a la parcela en términos del tratamiento y su bloque o repetición se utilizan los subíndices i y j respectivamente, pero cuando se requiere referirse a la posición de la parcela se utilizan los subíndices k y l , donde k indica la hilera o el eje Z y l indica la columna o el eje X . Cuando sea necesario referirse al tratamiento y a su posición, uno de los conjuntos de subíndices se escriben entre paréntesis, es decir $Y_{ij(k)}$ o $Y_{(ij)kl}$.

Un modelo general para representar los rendimientos de un ensayo de uniformidad puede ser escrito como:

$$Y_{kl} = \tau_{kl} + \varepsilon_{kl}, \quad (5.2.1)$$

donde

Y_{kl} = es la respuesta kl -ésima. Con $E[Y_{kl}] = \tau_{kl}$.

τ_{kl} = es el efecto de posición geográfica para la parcela con las coordenadas (x_l, z_k) . Este es el efecto de la tendencia y es una función de la posición; x_l es el valor de la coordenada de la l -ésima columna y z_k es el valor de la coordenada de la k -ésima hilera.

ε_{kl} = es el kl -ésimo error aleatorio asociado en la respuesta. Con $E[\varepsilon_{kl}] = 0$, $E[\varepsilon_{kl}^2] = \sigma^2$ y $E[\varepsilon_{kl}\varepsilon_{k'l'}] = 0$ para $k \neq k'$ y $l \neq l'$.

Por lo general la función de tendencia τ_{kl} es arbitraria, aunque se pueden establecer algunas formas matemáticas que puedan ajustar bien a algunos ensayos de uniformidad.

5.2.1 Evaluación de Cinco Propuestas de Análisis Mediante Simulación

Se realiza un estudio de simulación para la evaluación de 5 propuestas alternativas al uso de diseños en bloques para experimentos de campo. La razón de la simulación es la determinación de un *buen* método de análisis de experimentos de campo. Un buen método debe cumplir con las siguientes características:

- Bajo sesgo para la diferencia de tratamientos.
- Estimación insesgada de la varianza del error experimental.
- Alta potencia en la detección de diferencias.
- Correcta proporción del error Tipo I de acuerdo al nivel de significancia utilizado.
- Independencia de sesgos en la estimación de diferencias entre medias de tratamiento.
- De aplicación sencilla con paquetes estadísticos estándar, conocidos por agrónomos.

5.2.2 Descripción del Modelo Base

La simulación consiste de la generación de datos de acuerdo al modelo

$$Y_{ij(kl)} = \mu_i + \tau_{kl} + \varepsilon_{ij(kl)} \quad (5.2.2)$$

donde:

$$i = 1, 2, \dots, t.$$

$$j = 1, 2, \dots, n_i.$$

τ_{kl} : es la función que representa la tendencia en un ensayo de uniformidad.

$\varepsilon_{ij(kl)}$: es el error asociado con $Y_{ij(kl)}$; se asume que se distribuyen idénticamente como $N(0, \sigma^2)$.

μ_i : es la media del *i-ésimo* tratamiento.

Las μ_i en el modelo (5.2.2) son construidas de acuerdo a un diseño específico y arbitrario. Cabe mencionar que la tendencia no se debe construir artificialmente ya que su forma matemática usual no es conocida. La suposición de una forma específica de la tendencia puede ser especificada para algunas situaciones, un aspecto importante es que no siempre representa satisfactoriamente la realidad. Para representar la realidad, se utilizarán segmentos de ensayos de uniformidad como base para la generación de datos. El modelo (5.1.1) en la parte de la tendencia no tiene una forma explícita, pero una forma implícita de

la tendencia es inherente a los datos y se usará en el modelo sin una forma matemática exacta conocida.

Sean los datos del ensayo de uniformidad representados por el modelo

$$Y'_k = \tau'_k + \delta_k \quad (5.2.3)$$

donde

Y'_k : es la respuesta observada de un ensayo de uniformidad para la parcela localizada en (x_l, z_k) ,
 τ'_k : es el efecto de tendencia que contribuye a la respuesta de la parcela y
 δ_k es el error aleatorio que ocurrió dentro de la parcela.

La parte de la función de tendencia ubicada en el lado derecho en (5.2.3), τ'_k puede ser expresada como

$$\tau'_k = \mu + \tau_k \quad (5.2.4)$$

donde

μ : es la media de todas las observaciones del experimento de uniformidad y
 τ_k : es la función de tendencia definida como una desviación de la media.

Por lo que (5.2.3) se puede representar como

$$Y'_k = \mu + \tau_k + \delta_k \quad (5.2.5)$$

Se agregaron efectos de tratamiento, α_i , a los rendimientos de las parcelas del ensayo de uniformidad, de acuerdo a una asignación aleatoria en bloques al azar de los tratamientos. Así, los datos generados, a partir de (5.2.5), pueden ser expresados como

$$Y'_{ij(k)} + \alpha_i = \mu + \alpha_i + \tau_k + \delta_{ij(k)}$$

o

$$Y_{ij(k)} = \mu_i + \tau_k + \varepsilon_{ij(k)} \quad (5.2.6)$$

donde

$$Y_{ij(k)} = Y'_{ij(k)} + \alpha_i,$$

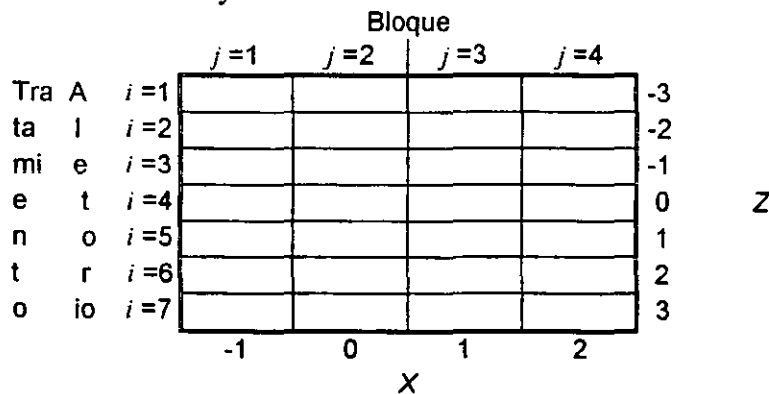
$$\varepsilon_{ij(k)} = \delta_{ij(k)} \sim NID(0, \sigma_\varepsilon^2 = \sigma_\delta^2) \text{ y}$$

μ_i : representa la media verdadera del tratamiento i ; $\mu_i = \mu + \alpha_i$.

5.2.3 Descripción de la Generación de Datos

La generación de los datos para simulación fue a partir del modelo (5.2.6) por la adición a los datos del ensayo de uniformidad de los efectos de tratamiento. Los datos del ensayo de uniformidad ($Y'_{y(kl)}$) fueron obtenidos de un arreglo de 7x4 en la parte central de cada ensayo seleccionado. De los cuales se consideraron 7 tratamientos (hileras) y 4 bloques (columnas); así mismo, se les asignó las coordenadas Z (hileras) y X (columna) de posición asignadas de izquierda a derecha y de arriba hacia abajo, son -3, -2, -1, 0, 1, 2, 3 y -1, 0, 1, 2 respectivamente, la asignación de cada uno de los siete tratamientos dentro de cada bloque fue realizado en forma aleatoria. Un esquema general de ello se observa en la Figura 5.1. La característica de los segmentos de ensayo de uniformidad para introducir la función de tendencia en el modelo, es que los conjuntos de datos de ensayos o experimentos de uniformidad especificaran de modo preciso la ubicación geográfica de las parcelas.

Figura 5.1.- Ubicación general de la estructura de diseño y variables de posición en el segmento central del ensayo.



5.2.4 Los Tratamientos

La estructura de tratamiento, un solo factor, utilizado en la simulación consistió en incrementar (decrementar) el valor observado ($Y'_{y(kl)}$) en los datos originales del

experimento en blanco, en el segmento seleccionado, con uno de los siete tratamientos ($\alpha_i; i = 1, \dots, 7$). Así mismo cada grupo de efectos de tratamiento fue variado en 9 incrementos ($\Delta_m; m = 1, \dots, 9$) de la media de los datos.

Así se corre una simulación para cada valor de Δ_m y en cada uno se definen los 7 tratamientos aplicados como:

$$\begin{aligned} \alpha_1 &= -3\Delta_m & \alpha_2 &= -2\Delta_m & \alpha_3 &= -1\Delta_m & \alpha_4 &= 0\Delta_m \\ \alpha_5 &= 1\Delta_m & \alpha_6 &= 2\Delta_m & \alpha_7 &= 3\Delta_m; & \forall m &= 1, \dots, 9. \end{aligned} \quad (5.2.7)$$

Los valores de $\Delta_m; m = 1, \dots, 9$ usados son los siguientes:

$$\begin{aligned} \Delta_1 &= 0.00\mu & \Delta_2 &= 0.03\mu & \Delta_3 &= 0.06\mu & \Delta_4 &= 0.09\mu & \Delta_5 &= 0.12\mu \\ \Delta_6 &= 0.15\mu & \Delta_7 &= 0.18\mu & \Delta_8 &= 0.21\mu & \Delta_9 &= 0.25\mu, \end{aligned} \quad (5.2.8)$$

donde μ es la media general del segmento de ensayo en blanco seleccionado.

Para cada $\Delta_m; m = 1, \dots, 9$ y dentro de las $\alpha_i; i = 1, \dots, 7$ se examinan la Potencia Empírica en la Proporción de Rechazos de $H_0: \alpha_1 = \dots = \alpha_7$, estimar las medias $\mu_i; \forall i = 1, \dots, 7$ y sus diferencias, así como la estimación de la varianza del error experimental. Para esto se generaran 1000 asignaciones de tratamiento aleatoriamente de acuerdo a un diseño de bloques completos al azar, dando lugar a 9000 (1000 x 9) grupos de datos, cada uno con 28 respuestas. Las asignaciones aleatorias y su respectiva identificación de tratamiento fue mediante un programa en lenguaje C++ (Ver. 5.0 for Windows 9X). Cada grupo de datos se analizaron con cada uno de los cinco métodos que se investiga y que se detallan más adelante. Para cada ensayo en blanco se realizaron entonces 45000 asignaciones aleatorias de tratamiento (para los 5 conjuntos de datos genera un total de 225000 aleatorizaciones) (Cuadro 5.3). Para el ajuste de modelos y otros cálculos se utilizó de modo conjunto los módulos Estadístico (Proc GLM y Proc REG) y Matricial del sistema SAS.

Cuadro 5.3. Representación del número total de simulaciones en los ensayos en blanco.

Si Bloque				No Bloque				No. de Ensayos en Blanco: 5
Δ	01	02	03	04	05			
0.00mu	1000	00	00	00	00	00	Metodologías Si Bloque: 3 Metodologías No Bloque: 2 Total de metodologías: 5 9000x5= 45000 45000x5=225000	
0.03mu	1000	03	03	03	03	03		
0.06mu	1000	06	06	06	06	06		
0.09mu	1000	09	09	09	09	09		
0.12mu	1000	12	12	12	12	12		
0.15mu	1000	15	15	15	15	15		
0.18mu	1000	18	18	18	18	18		
0.21mu	1000	21	21	21	21	21		
0.25mu	1000	25	25	25	25	25		
	9000	9000	9000	9000	9000	9000		

La forma de representar en el modelo la variación espacial no explicada por los efectos de tratamiento se ha investigado de 5 maneras, las cuales se dividirán en 2 grupos de modelos, uno de ellos se denominará *Métodos de Refinamiento a los Bloques* y el otro será denominado *Métodos que Ignoran Bloques*; dentro del primer grupo se estudian los métodos de *Bloques al Azar (BA)*, *Bloques con Covariable de Posición (BCP)* y *Papadakis Dentro de Bloque (PDB)* y para los métodos que ignoran bloques se estudia *Papadakis Clásico en Dos Dimensiones (PC2)* y *Polinomio en Coordenadas de Posición (PCP)*, estos métodos se detallan a continuación.

5.3. Métodos de Refinamiento a los Bloques

5.3.1 Bloques al Azar

Para este diseño se tiene t tratamientos que se asignan aleatoriamente en b bloques y se ajusta el modelo mediante mínimos cuadrados clásicos. Con la finalidad de obtener algunas estadísticas necesarias se utiliza el procedimiento Proc GLM de SAS.

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad \forall i = 1, \dots, 7 \text{ y } j = 1, \dots, 4 \quad (5.2.9)$$

donde

Y_{ij} : Respuesta observada de la parcela que recibió el i -ésimo tratamiento en el j -ésimo bloque.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

β_j : efecto fijo del j -ésimo bloque.

ε_{ij} : ij -ésimo error aleatorio; el cual, si el modelo es adecuado, se $\sim NID(0, \sigma_\varepsilon^2)$.

5.3.1.1. Estadísticas de evaluación del modelo

Para cada tipo de modelo una vez *ajustado*, para los 1000 grupos de datos, dentro de cada $\Delta_m; m = 1, \dots, 9$, se obtienen (resultado de proceso en SAS) y calculan a partir de los respectivos predichos \hat{Y}_{ij} las medias de tratamientos ajustadas, $\hat{\mu}_i, \forall i = 1, \dots, 7$ y las siguientes *estadísticas promedio* sobre los 1000 grupos:

1.- CM_{Error} , F_0 , $\text{Prob}[F \geq F_0]$ y la *Significancia del Modelo*; se obtienen de los resultados de proceso en el programa SAS.

2.- *Potencia Empírica en Tukey*; para ello se realiza la Prueba de Comparaciones de Medias de Tukey, con un nivel de significancia del 5 % para las siguientes hipótesis nulas $H_0: \mu_1 - \mu_{i'}, \forall i' = 2, \dots, 7$.

3.- *Sesgo Empírico* para diferencias de tratamientos; se calculan del modo siguiente

$$\text{Sesgo}E_{i',j} = (\hat{\mu}_1 - \hat{\mu}_{i'}) - (i'-1)\Delta_m; \forall i' = 2, \dots, 7 \text{ y } m = 1, \dots, 9. \quad (5.2.10)$$

4.- *Error Cuadrático Medio Empírico* para diferencia de tratamientos; se calculan del modo siguiente

$$ECME_{i',j} = [(\hat{\mu}_1 - \hat{\mu}_{i'}) - (i'-1)\Delta_m]^2; \forall i' = 2, \dots, 7 \text{ y } m = 1, \dots, 9. \quad (5.2.11)$$

En el Anexo en CD (D:\Análisis\M0?) se presentan los programas en C++ y los programas en SAS para uno de los experimentos en blanco (Baker *et al.*, 1952) analizados con cada uno de los 5 métodos de análisis y para un sólo $\Delta_m; m = 1, \dots, 9$.

5.3.2 Bloques con Covariable de Posición

Para esta parte se aplican dos etapas, en una de ellas se ajusta el modelo siguiente

$$Y_{ij(kl)} = \mu + \tau_i + \beta_j + \theta_{1j}Z_{kl} + \theta_{2j}Z_{kl}^2 + \theta_{3j}Z_{kl}^3 + \theta_{4j}Z_{kl}^4 + \varepsilon_{ij(kl)} \quad \forall i, k = 1, \dots, 7 \text{ y } j, l = 1, \dots, 4, \quad (5.2.12)$$

el cual es equivalente al siguiente

$$\begin{aligned} Y_{ij(kl)} = & \mu + \tau_i + \beta_j \\ & + \theta_{11}Z_{kl}I_1 + \theta_{21}Z_{kl}^2I_1 + \theta_{31}Z_{kl}^3I_1 + \theta_{41}Z_{kl}^4I_1 \\ & + \theta_{12}Z_{kl}I_2 + \theta_{22}Z_{kl}^2I_2 + \theta_{32}Z_{kl}^3I_2 + \theta_{42}Z_{kl}^4I_2 \\ & + \theta_{13}Z_{kl}I_3 + \theta_{23}Z_{kl}^2I_3 + \theta_{33}Z_{kl}^3I_3 + \theta_{43}Z_{kl}^4I_3 + \varepsilon_{ij(kl)} \quad \forall i, k = 1, \dots, 7 \text{ y } j, l = 1, \dots, 4 \end{aligned} \quad (5.2.13)$$

donde

I_1, I_2, I_3 : son covariables como indicadora de bloque que tienen un valor de 1 si corresponden al bloque 1, 2 y 3 respectivamente y 0 de otro modo.

$Y_{ij(kl)}$: Respuesta observada del i -ésimo tratamiento en el j -ésimo bloque en la posición kl -ésima.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

β_j : efecto fijo del j -ésimo bloque.

$\varepsilon_{ij(kl)}$: ij -ésimo error aleatorio en la posición kl -ésima, el cual, si el modelo es adecuado, se $\sim NID(0, \sigma_\varepsilon^2)$.

Los parámetros del modelo (5.2.13) se estiman como un modelo de regresión lineal múltiple (Proc REG) a la vez que se aplica el método de Selección de Variables por Etapas (Stepwise) (Montgomery y Peck, 1982) con permanencia en el modelo de los términos asociados a tratamientos y bloques.

Como una segunda etapa, con los términos que permanecen en el modelo indicado en (5.2.14) se estima el modelo restringido (indicado entre corchetes “[]” lo que significa que pueden estar los 12 términos asociados a Z , algunos o ninguno de ellos).

$$\begin{aligned}
Y_{ij(k)} = & \mu + \tau_i + \beta_j \\
& + [\theta_{11}Z_{kl}I_1 + \theta_{21}Z_{kl}^2I_1 + \theta_{31}Z_{kl}^3I_1 + \theta_{41}Z_{kl}^4I_1 \\
& + \theta_{12}Z_{kl}I_2 + \theta_{22}Z_{kl}^2I_2 + \theta_{32}Z_{kl}^3I_2 + \theta_{42}Z_{kl}^4I_2 \\
& + \theta_{13}Z_{kl}I_3 + \theta_{23}Z_{kl}^2I_3 + \theta_{33}Z_{kl}^3I_3 + \theta_{43}Z_{kl}^4I_3] + \varepsilon_{ik(k)} \quad \forall i, k = 1, \dots, 7 \text{ y } j, l = 1, \dots, 4
\end{aligned}
\tag{5.2.14}$$

donde

I_1, I_2, I_3 : son covariables que tiene un valor de 1 si corresponden al bloque 1, 2 y 3 respectivamente y 0 de otro modo.

$Y_{ij(k)}$: Respuesta observada del i -ésimo tratamiento en el j -ésimo bloque en la posición kl -ésima.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

β_j : efecto fijo del j -ésimo bloque.

$\varepsilon_{ij(k)}$: ij -ésimo error aleatorio en la posición kl -ésima; el cual, si el modelo es adecuado se $\sim NID(0, \sigma_\varepsilon^2)$.

Del ajuste (Proc GLM) del modelo anterior (5.2.14) restringido con los términos seleccionados, se obtienen las estadísticas de evaluación del modelo para cada $\Delta_m; m = 1, \dots, 9$ como se indica en (5.3.1.1).

5.3.3 Papadakis Dentro de Bloque

Se considera el análisis en una dimensión para un diseño de bloques completos al azar. Par a ellos lo que se realiza primero es el ajuste (Proc GLM) del modelo siguiente

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad \forall i = 1, \dots, 7 \text{ y } j = 1, \dots, 4 \tag{5.2.14}$$

donde

Y_{ij} : Respuesta observada del i -ésimo tratamiento en el j -ésimo bloque.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

β_j : efecto fijo del j -ésimo bloque.

ε_{ij} : ij -ésimo error aleatorio; el cual, si el modelo es adecuado se $\sim NID(0, \sigma_\varepsilon^2)$.

donde, en términos de los estimadores, se obtienen los residuos del modo siguiente

$$r_{ij} = Y_{ij(kl)} - \hat{Y}_{ij(kl)} \tag{5.2.15}$$

Se espera que el residuo dependa fundamentalmente de la posición (kl) , es este sentido, se demuestra en Méndez (1970) que si hay contaminación por los tratamientos, pero esta es pequeña.

Con estos residuos se calcula el Índice de Papadakis por Bloque (IPb) del modo siguiente (ver Figura 5.2):

Parte extremo superior del bloque

$$IPb_{kl} = r_{(k+1)l} \quad \forall k = 1 \text{ y } l = b = 1, 2, 3, 4. \tag{5.2.16}$$

Parte extremo inferior del bloque

$$IPb_{kl} = r_{(k-1)l} \quad \forall k = 7 \text{ y } l = b = 1, 2, 3, 4. \tag{5.2.17}$$

Parte intermedia del bloque

$$IPb_{kl} = \frac{r_{(k-1)l} + r_{(k+1)l}}{2} \quad \forall k = 2, \dots, 6 \text{ y } l = b = 1, 2, 3, 4. \tag{5.2.18}$$

Figura 5.2-Representación del cálculo del Índice de Papadakis por Bloque.

		<i>l</i>				⇒	<i>l</i>			
		1	2	3	4		1	2	3	4
<i>k</i>	1	r_{11}	...				$r_{(k+1)l}$...		
	2	r_{21}	...	⋮			$(r_{(k-1)l} + r_{(k+1)l})/2$...	⋮	
	3	r_{31}					.			
	4	r_{41}					.			
	5	r_{51}					.			
	6	r_{61}					$(r_{(k-1)l} + r_{(k+1)l})/2$			
	7	r_{71}				r_{74}	$r_{(k-1)l}$			$r_{(7-1)4}$
		<i>Residuos</i>					<i>IPb_{kl}</i>			

Con el índice IPb_{kl} obtenido en (5.2.16)-(5.2.18) se ajusta mediante mínimos cuadrados el modelo siguiente

$$Y_{ij(k)} = \mu + \tau_i + \beta_j + \theta IPb_{kl} + \varepsilon_{ij(k)} \quad \forall i, k = 1, \dots, 7 \text{ y } j, l = 1, \dots, 4. \quad (5.2.19)$$

donde

Y_{ij} : Respuesta observada del i -ésimo tratamiento en el j -ésimo bloque.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

β_j : efecto fijo del j -ésimo bloque.

θ : es el parámetro de la covariable.

IPb_{kl} : Índice de Papadakis por bloque en la kl -ésima posición.

$\varepsilon_{ij(k)}$: ij -ésimo error aleatorio en la posición kl -ésima; el cual, si el modelo es adecuado se $\sim NID(0, \sigma_\varepsilon^2)$.

Del ajuste (Proc GLM) del modelo anterior (5.2.19) se obtienen las estadísticas de evaluación del modelo para cada $\Delta_m; m = 1, \dots, 9$ como se indica en (5.3.1.1).

5.4 Métodos que Ignoran Bloques

5.4.1 Papadakis Clásico en Dos Dimensiones

Para este método se ajusta el modelo para un diseño completamente al azar. Para ello lo que se realiza primero es el ajuste (Proc GLM) del modelo siguiente

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \forall i = 1, \dots, 7 \text{ y } j = 1, \dots, 4 \quad (5.2.20)$$

donde

Y_{ij} : Respuesta observada del i -ésimo tratamiento en el j -ésimo bloque.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

ε_{ij} : ij -ésimo error aleatorio; el cual, si el modelo es adecuado se $\sim NID(0, \sigma_\varepsilon^2)$.

Donde, en términos de los estimadores, se obtienen los residuos del modo siguiente

$$r_{ij} = Y_{ij(k)} - \hat{Y}_{ij(k)}. \quad (5.2.21)$$

Con estos residuos se calcula el Índice de Papadakis (IP_{kl}) del modo siguiente

(Figura 5.3):

Esquinas del área

$$IP_{kl} = \frac{r_{(k+1)l} + r_{k(l+1)}}{2} \quad \forall k = 1 \text{ y } l = 1 \tag{5.2.22}$$

$$IP_{kl} = \frac{r_{(k+1)l} + r_{k(l-1)}}{2} \quad \forall k = 1 \text{ y } l = 4 \tag{5.2.23}$$

$$IP_{kl} = \frac{r_{(k-1)l} + r_{k(l+1)}}{2} \quad \forall k = 7 \text{ y } l = 1 \tag{5.2.24}$$

$$IP_{kl} = \frac{r_{(k-1)l} + r_{k(l-1)}}{2} \quad \forall k = 7 \text{ y } l = 4 \tag{5.2.25}$$

Lados del área

Izquierdo y Derecho

$$IP_{kl} = \frac{r_{(k-1)l} + r_{(k+1)l} + r_{k(l+1)}}{3} \quad \forall k = 2, \dots, 6 \text{ y } l = 1 \tag{5.2.26}$$

$$IP_{kl} = \frac{r_{(k-1)l} + r_{(k+1)l} + r_{k(l-1)}}{3} \quad \forall k = 2, \dots, 6 \text{ y } l = 4 \tag{5.2.27}$$

Arriba y Abajo

$$IP_{kl} = \frac{r_{(k+1)l} + r_{k(l-1)} + r_{k(l+1)}}{3} \quad \forall k = 1 \text{ y } l = 2, 3 \tag{5.2.28}$$

$$IP_{kl} = \frac{r_{(k-1)l} + r_{k(l-1)} + r_{k(l+1)}}{3} \quad \forall k = 7 \text{ y } l = 2, 3 \tag{5.2.29}$$

Interior del área

$$IP_{kl} = \frac{r_{(k-1)l} + r_{(k+1)l} + r_{k(l-1)} + r_{k(l+1)}}{4} \quad \forall k = 2, \dots, 6, \text{ y } l = 2, 3 \tag{5.2.30}$$

Figura 5.3-Representación del cálculo del Índice de Papadakis en 2 Dimensiones a partir de residuos.

		l			
		1	2	3	4
k	1	r_{11}	r_{12}	...	r_{14}
	2	r_{21}	r_{22}		r_{24}
	3	.			.
	4	.			.
	5	.			.
	6	r_{61}		...	r_{64}
	7	r_{71}	r_{72}	...	r_{74}
		Residuos			

		l			
		1	2	3	4
k	1	$r_{(k+1)l} + r_{k(l+1)} / 2$	$r_{(k+1)l} + r_{k(l-1)} + r_{k(l+1)} / 3$...	$r_{(k+1)l} + r_{k(l-1)} / 2$
	2	$r_{(r-1)l} + r_{(k+1)l} + r_{k(l+1)} / 3$	$r_{(k-1)l} + r_{(k+1)l} + r_{k(l-1)} + r_{k(l+1)} / 4$		$r_{(k-1)l} + r_{(k+1)l} + r_{k(l-1)} / 3$
	3	.			.
	4	.			.
	5	.			.
	6	$r_{(r-1)l} + r_{(k+1)l} + r_{k(l+1)} / 3$...	$r_{(k-1)l} + r_{(k+1)l} + r_{k(l-1)} / 3$
	7	$r_{(k-1)l} + r_{k(l+1)} / 2$	$r_{(k-1)l} + r_{k(l-1)} + r_{k(l+1)} / 3$...	$r_{(k-1)l} + r_{k(l-1)} / 2$
IP_{kl}					

Con el índice IP_{kl} obtenido en (5.2.22)-(5.2.30) se ajusta mediante mínimos cuadrados el modelo siguiente

$$Y_{ij(kl)} = \mu + \tau_i + \theta IP_{kl} + \varepsilon_{ij(kl)} \quad \forall i, k = 1, \dots, 7 \text{ y } j, l = 1, \dots, 4. \tag{5.2.31}$$

donde

Y_{ij} : Respuesta observado del *i*-ésimo tratamiento en el *j*-ésimo bloque en la posición *kl*-ésima.

μ : media general.

τ_i : efecto fijo de la *i*-ésimo tratamiento.

θ : es el coeficiente de regresión de la covariable.

IP_{kl} : Índice de Papadakis para la *kl*-ésima posición.

$\varepsilon_{ij(kl)}$: *ij*-ésimo error aleatorio en la posición *kl*-ésima; el cual, si el modelo es adecuado se $\sim NID(0, \sigma_\varepsilon^2)$.

Del ajuste (Proc GLM) del modelo anterior (5.2.31) se obtienen las estadísticas de evaluación del modelo para cada $\Delta_j; j = 1, \dots, 9$ como se indica en (5.3.1.1).

5.4.2 Polinomio en Coordenadas de Posición

Tomando las ideas de Méndez (1970), se supone una tendencia polinomial en sentido horizontal (hileras) y en el sentido vertical (columnas); para modelar esta tendencia se ajusta el modelo.

Para esta parte se aplican dos etapas, en una de ellas se ajusta el modelo siguiente

$$\begin{aligned}
Y_{ij} = & \mu + \tau_i + \theta_{10}Z + \theta_{01}X + \theta_{11}ZX \\
& + \theta_{20}Z^2 + \theta_{21}Z^2X + \theta_{22}Z^2X^2 + \theta_{23}Z^2X^3 \\
& + \theta_{30}Z^3 + \theta_{31}Z^3X + \theta_{32}Z^3X^2 + \theta_{33}Z^3X^3 \\
& + \theta_{40}Z^4 + \theta_{40}Z^4X + \theta_{42}Z^4X^2 + \theta_{43}Z^4X^3 + \varepsilon_{ij} \quad \forall i = 1, \dots, t; j = 1, \dots, b
\end{aligned}
\tag{5.2.32}$$

donde

Y_{ij} : Respuesta observada del i -ésimo tratamiento en el j -ésimo bloque.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

Z, X : Variables de posición.

θ_{nm} : es el coeficiente de regresión del término $Z^n X^m$.

ε_{ij} : ij -ésimo error aleatorio en la posición kl -ésima; el cual, si el modelo es adecuado se $\sim NID(0, \sigma_\varepsilon^2)$.

Como una segunda etapa, con los términos que permanecen en el modelo indicado en (5.2.33) se estima el modelo restringido (indicado entre corchetes “[]” lo que significa que pueden estar los 15 términos asociados a Z, X , algunos o ninguno de ellos).

$$\begin{aligned}
Y_{ij} = & \mu + \tau_i + [\theta_{10}Z + \theta_{01}X + \theta_{11}ZX \\
& + \theta_{20}Z^2 + \theta_{21}Z^2X + \theta_{22}Z^2X^2 + \theta_{23}Z^2X^3 \\
& + \theta_{30}Z^3 + \theta_{31}Z^3X + \theta_{32}Z^3X^2 + \theta_{33}Z^3X^3 \\
& + \theta_{40}Z^4 + \theta_{40}Z^4X + \theta_{42}Z^4X^2 + \theta_{43}Z^4X^3] + \varepsilon'_{ij} \quad \forall i = 1, \dots, t; j = 1, \dots, b
\end{aligned}
\tag{5.2.33}$$

donde

Y_{ij} : Respuesta observada del i -ésimo tratamiento en el j -ésimo bloque.

μ : media general.

τ_i : efecto fijo del i -ésimo tratamiento.

Z, X : Variables de posición.

θ_{nm} : es el coeficiente de regresión seleccionado del término $Z^n X^m$.

ε'_{ij} : ij -ésimo error aleatorio en la posición kl -ésima; el cual, si el modelo es adecuado se $\sim NID(0, \sigma_\varepsilon^2)$.

Del ajuste (Proc GLM) del modelo anterior (5.2.33) restringido con los términos seleccionados, se obtienen las estadísticas de evaluación del modelo para cada $\Delta_m; m = 1, \dots, 9$ como se indica en (5.3.1.1).

VI.

RESULTADOS

6.1. Estadísticas Básicas de Cada Conjunto de Datos a Analizar

Para los conjuntos de datos que tenían especificado de modo preciso su distribución espacial se le realizó una análisis de estadísticas básicas para los datos sin ninguna concentración, es decir, como están publicados en los artículos respectivos. La información para los conjuntos de datos se muestra en el Cuadro 6.1.

Cuadro 6.1.-Algunas estadísticas básicas de los conjuntos de datos elegidos.

Fuente de Datos	Número		N	Valor		Media	Varianza
	Filas	Col.		Mínimo	Máximo		
Baker <i>et al.</i> (1952)	57	10	570	720.00	4 590.00	2 551.3228	645 708.3279
Immer (1932)							
-Imm32a	60	10	600	37.10	67.70	51.1875	28.8710
-Imm32b	60	10	600	13.15	16.27	14.5232	0.3118
-Imm32c	60	10	600	77.70	90.90	84.9815	4.6222
Lander <i>et al.</i> (1938)	26	15	390	6.30	42.40	18.4359	48.0030
Mercer y Hall (1911)	20	25	500	2.73	5.16	3.9486	0.2100
Wassom y Kalton (1953)							
-WyK53a	36	36	1 296	0.70	7.80	4.0662	0.8606
-WyK53b	36	36	1 296	0.60	3.43	1.9000	0.2414
-WyK53c	36	36	1 296	2.10	7.40	5.1180	0.4490

En el Cuadro 6.1 destacan por su varianza “grande” los datos de rendimiento de Baker *et al.* (1952) con 645 708.3279 libras².

A los datos de los experimentos seleccionados y compactados por surcos o bloques, según sea el caso, se le obtuvo de igual manera, al Cuadro 6.1, algunas estadísticas básicas, la cual se presenta en el Cuadro 6.2, en este cuadro se observa que la media y varianza para todos los conjuntos de datos aumenta.

Cuadro 6.2.-Algunas estadísticas básicas de los conjuntos de datos compactados.

Fuente de Datos	Número		N	Valor		Media	Varianza
	Filas	Col.		Mínimo	Máximo		
Baker <i>et al.</i> (1952)	57	5	285	1 850.00	7 610.00	5 102.6456	1 941 201.5536
Immer (1932)							
-Imm32a	20	10	200	124.30	190.30	153.5625	141.2828
-Imm32b	20	10	200	40.33	47.64	43.5695	1.6725
-Imm32c	20	10	200	244.60	267.70	254.7830	16.2287
Lander <i>et al.</i> (1938)	16	15	195	19.80	60.30	36.8718	67.1607
Mercer y Hall (1911)	20	25	500	2.73	5.16	3.9486	0.2100
Wassom y Kalton (1953)							
-WyK53a	18	36	648	1.60	12.20	8.1324	2.4063
-WyK53b	18	36	648	1.44	6.31	3.8000	0.7420
-WyK53c	18	36	648	5.90	13.00	10.2360	0.9755

Así mismo, a los datos de la muestra (7 x 4 respuestas) en el experimento en blanco se les calculó algunas estadísticas básicas, las cuales se observan en el Cuadro 6.3. Se observa que en términos generales que la Media de la mayoría de los conjuntos de datos permanece igual, con respecto a los datos compactados (Cuadro 6.2), la Varianza tiende a disminuir.

Cuadro 6.3.-Algunas estadísticas básicas de los conjuntos de datos de la muestra.

Fuente de Datos	Número		N	Valor		Media	Varianza
	Filas	Col.		Mínimo	Máximo		
Baker <i>et al.</i> (1952)	7	4	28	4 880.00	7 610.00	<u>6 523.5714</u>	<u>610 571.9577</u>
Immer (1932)							
-Imm32a	7	4	28	139.50	173.70	<u>158.8179</u>	<u>92.2830</u>
-Imm32b	7	4	28	42.16	44.61	43.4168	0.4316
-Imm32c	7		28	249.10	259.40	254.0500	5.6648
Lander <i>et al.</i> (1938)	7	4	28	28.60	58.60	<u>40.6107</u>	<u>67.9091</u>
Mercer y Hall (1911)	7	4	28	3.04	4.99	<u>3.7975</u>	<u>0.1940</u>
Wassom y Kalton (1953)							
-WyK53a	7	4	28	2.50	9.30	<u>5.9429</u>	<u>3.0862</u>
-WyK53b	7	4	28	2.58	4.58	3.7164	0.2355
-WyK53c	7	4	28	9.00	11.50	10.0964	0.4263

6.2. Relación entre la Varianza de los Errores Estimados Obtenido de la Simulación con la del Segmento del Ensayo de Uniformidad

Para cada ensayo de uniformidad y para cada combinación de Δ (9 niveles de incremento) y Método de Análisis (5 niveles de método) se crea un cuadro de doble entrada

con 45 entradas (Cuadro 6.4) a partir de las cuales realiza un Análisis de Varianza en el sistema SAS para el modelo:

$$Y_{ijk}^* = \mu + \text{Metodo}_i + \Delta_j + \varepsilon_{ij}; \forall i = 1, \dots, 5 \text{ y } j = 1, \dots, 9 \quad (6.1)$$

donde

Y_{ij}^* : Respuesta medida en la celda correspondiente al i -ésimo método en el j -ésimo incremento.

μ : Media General.

Metodo_i : Efecto fijo del i -ésimo método.

Δ_j : Efecto del j -ésimo incremento.

ε_{ij} : ij -ésimo error aleatorio del modelo.

El valor de la respuesta Y_{ij}^* se define del modo siguiente:

$$Y_{ij}^* = \text{Log}(\text{CME}_{ij}) - \text{Log}(\sigma^2); \forall i = 1, \dots, 5 \text{ y } j = 1, \dots, 9, \quad (6.2)$$

Cabe mencionar que el modelo (6.1) y los que se utilizan más adelante se utilizan sólo para analizar los resultados por lo que sólo son una forma de detectar significancia de factores o interacciones por lo que no se puede asumir como válido que los $\varepsilon_{(i)} \sim N(0, \sigma^2)$.

Cuadro 6.4. Cuadro de resultados, Y_{ij}^* , con dos criterios de clasificación para los datos de Baker *et al.* (1952).

Δ	Metodo				
	BA	BCP	PDB	PC2	PCP
0.00mu	-0.71680	-0.90467	-0.71242	-0.27832	-0.10297
0.03mu	-0.71602	-0.89574	-0.71497	-0.27792	-0.10524
0.06mu	-0.71573	-0.89370	-0.71312	-0.27828	-0.10323
0.09mu	-0.71705	-0.90060	-0.71448	-0.27676	-0.10513
0.12mu	-0.71575	-0.89630	-0.71586	-0.28116	-0.10695
0.15mu	-0.71167	-0.89933	-0.71166	-0.27760	-0.10050
0.18mu	-0.71321	-0.90644	-0.71672	-0.27745	0.09015
0.21mu	-0.71633	-0.89864	-0.71684	-0.27973	-0.10278
0.25mu	-0.71551	-0.90068	-0.71588	-0.27844	-0.10409

donde el valor de σ^2 corresponde a la varianza de los datos del segmento seleccionado de la parte del ensayo de uniformidad. Por ejemplo, los valores de CME_{1j} , CME_{2j} , CME_{3j} , CME_{4j} y CME_{5j} para todo $j=1, \dots, 9$, corresponde a los datos del ensayo de uniformidad de Baker *et al.* (1952) resultado de los 5 métodos investigados y se encuentra en los Cuadros 9.6a-9.10c del Anexo en CD y su correspondiente valor de $\sigma^2 = 610571.9577$ (subrayado en Cuadro 6.3). Las estadísticas (especificadas en 5.3.1.1) resultantes de la simulación de los 5 métodos al que fueron sometidos los ensayos de uniformidad de Immer (1932), Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton (1953) están en los Cuadros 9.11a-9.15c, 9.16a-9.20c, 9.21a-9.25c y 9.26a-9.30c respectivamente del Anexo (en CD).

Las estadísticas resultantes para el modelo (6.1) y el factor Método de Análisis para los 5 conjuntos de datos se presentan en el Cuadro 6.5.

Cuadro 6.5.-Estadísticas del modelo y significancia del factor *Metodo* para los conjuntos de ensayos de uniformidad .

Fuente de Datos	Modelo				Metodo
	CME	Fo	Prob[Fo ≥ F]	R ²	Prob[Fo ≥ F]
Baker <i>et al.</i> (1952)	0.000856120	410.98	0.0001**	0.9936	0.0001**
Immer (1932)	0.000007700	1 915.26	0.0001**	0.9986	0.0001**
Lander <i>et al.</i> (1938)	0.000003910	31 178.49	0.0001**	0.9999	0.0001**
Mercer y Hall (1911)	0.000442850	48.33	0.0001**	0.9477	0.0001**
Wassom y Kalton (1953)	0.000011090	1 703.49	0.0001**	0.9984	0.0001**

** : Indica una alta significancia del efecto respectivo para un Error Tipo I del 1%.

En el Cuadro 6.5 se observa alta significancia ($p \leq 0.01$) del modelo y del factor Método ($p \leq 0.01$) para los cinco conjuntos de datos investigados. Esto último lleva a realizar una Prueba de Comparaciones Múltiples de Medias de Tukey, con un nivel de significancia del 5 % para las siguientes hipótesis nulas $H_0 : \mu_i^* - \mu_{i'}^* ; \forall i \neq i' \text{ y } i, i' = 1, \dots, 5$. Los resultados de la aplicación de Tukey para los cinco conjuntos de datos se pueden observar en el Cuadro 6.6 y en las Figuras 6.1-6.5 donde las gráficas están dibujadas en la misma escala.

Cuadro 6.6.-Prueba de Tukey para la respuesta media en el factor Método de Análisis en cinco Fuentes de Datos.

Baker <i>et al.</i> (1952) ¹		Immer (1932) ¹		Lander <i>et al.</i> (1938) ¹		Mercer y Hall (1911) ¹		Wassom y Kalton (1953) ¹	
Y_i^*	Método	Y_i^*	Método	Y_i^*	Método	Y_i^*	Método	Y_i^*	Método
-0.082a	PCP	0.033a	PDB	-0.261a	PC2	0.004a	PC2	-0.266a	PC2
-0.278 b	PC2	0.032a	BA	-0.281 b	PCP	-0.085 b	PDB	-0.282 b	BA
-0.715 c	PDB	-0.065 b	PC2	-0.572 c	BA	-0.090 b	BA	-0.286 b	PDB
-0.715 c	BA	-0.072 c	BCP	-0.698 d	PDB	-0.166 c	PCP	-0.414 c	BCP
-0.900 d	BCP	-0.127 d	PCP	-0.881 e	BCP	-0.214 d	BCP	-0.430 d	PCP

¹Medias con la misma letra son estadísticamente iguales con un nivel de significancia del 5% en la Prueba de Comparaciones Múltiples de Medias de Tukey.

El menor valor medio (valor absoluto) de Y_{ij}^* o de la diferencia media entre los logaritmo de varianzas se observa en el método de Polinomio en Coordenadas de Posición que corresponde al conjunto de Métodos que Ignoran Bloques, para la fuente de Datos de Baker *et al.*(1952). Es menor para el método de Papadakis Clásico en 2 Dimensiones, para el mismo conjunto de métodos, en la fuente de datos de Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton (1953) con valor mayor de la media de -0.082, -0.261, -0.004 y -0.266 respectivamente. El conjunto de datos de Immer (1932) se comporta diferente de los otro cuatro fuente de datos, el menor valor medio de Y_{ij}^* para los métodos se observa en los métodos Bloques al Azar (0.032) y Papadakis Dentro de Bloques (0.033) los cuales corresponden al conjunto de Métodos de Refinamiento a los Bloques.

6.3. Comportamiento de la Significancia del Modelo de Cada Método

Para la evaluación de este aspecto se grafica el valor de \hat{p}_j (nivel de significancia estimado) para su respectivo valor de Δ_j , $\forall j = 1, \dots, 9$ en los cinco métodos.

Para el segmento de datos de Baker *et al.* (1953) se observa un mejor comportamiento o tendencia hacia la significancia (menor \hat{p}) para los tres Métodos de Refinamiento a los Bloques (BA, BCP y PDB), los cuales tienen un comportamiento similar. Los métodos de PC2 y PCP tienen valores de $\hat{p} = Prob[F \geq F_0]$ cercano a 1 cuando $\Delta_1 = 0.00\mu$, los otros 3 métodos (BA, BCP y PDB) métodos tienden a dar valores de \hat{p} menores y en los cuales el nivel de significancia de 0.05 lo alcanzan en valores de Δ menores de 0.03μ (Fig. 6.6).

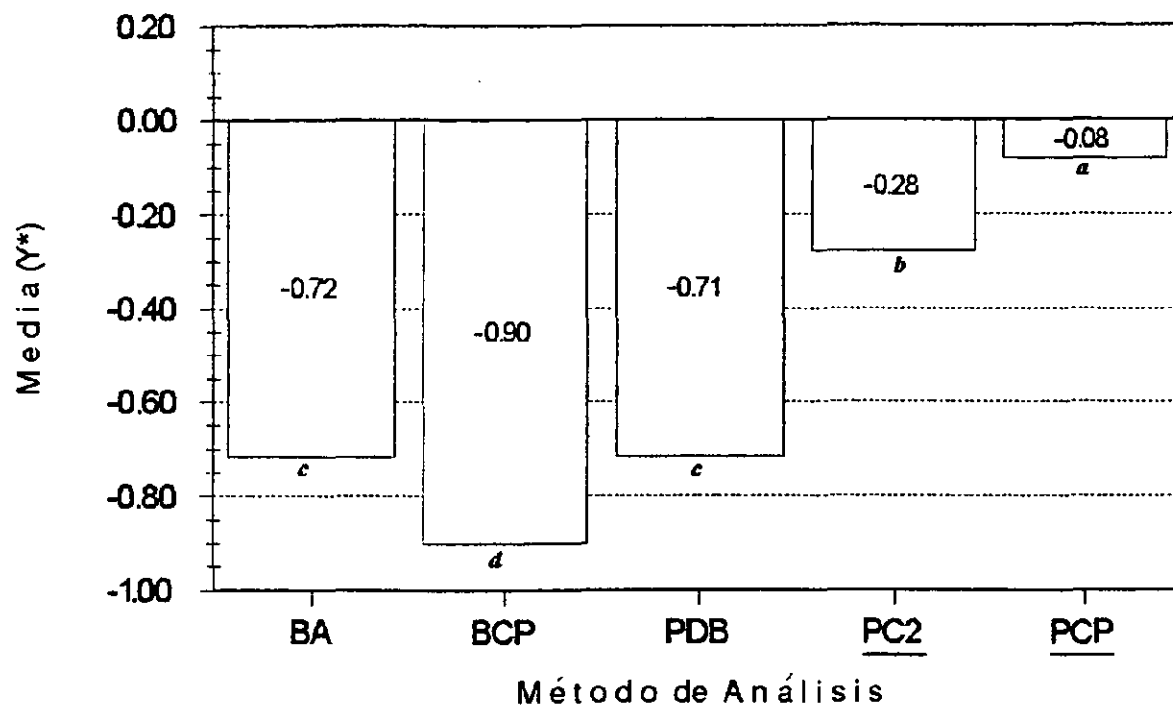


Figura 6.1. Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Beker *et al.* (1952). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.

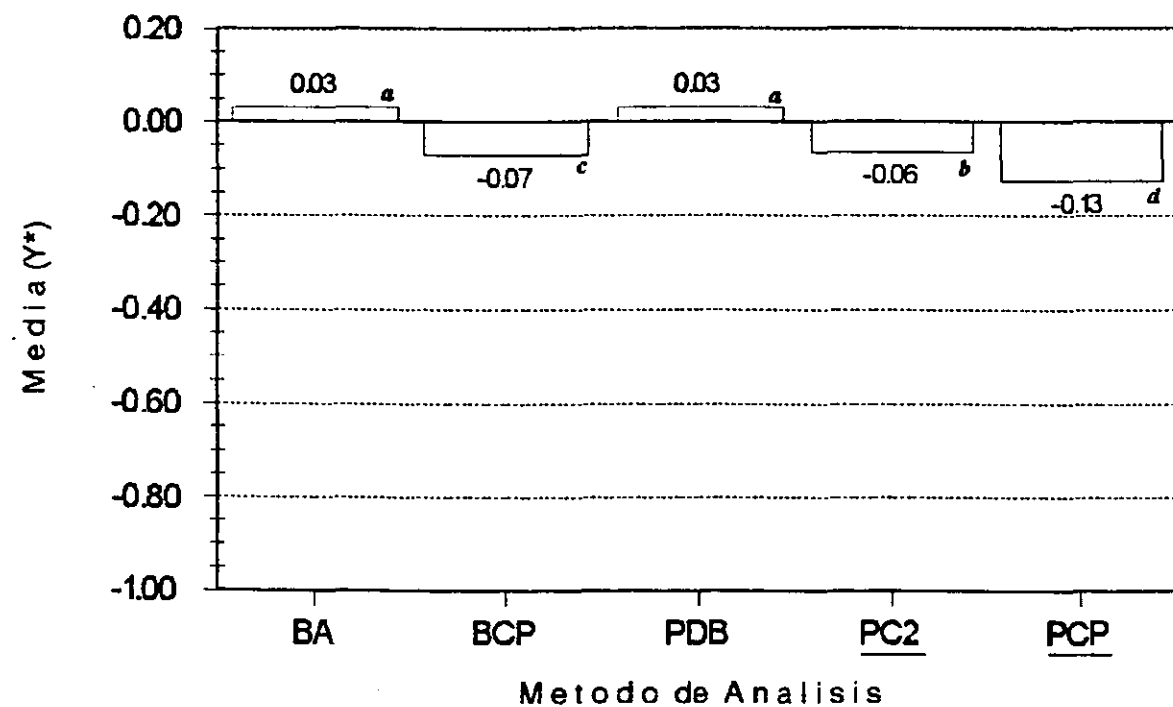


Figura 6.2. Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Immer (1932). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.

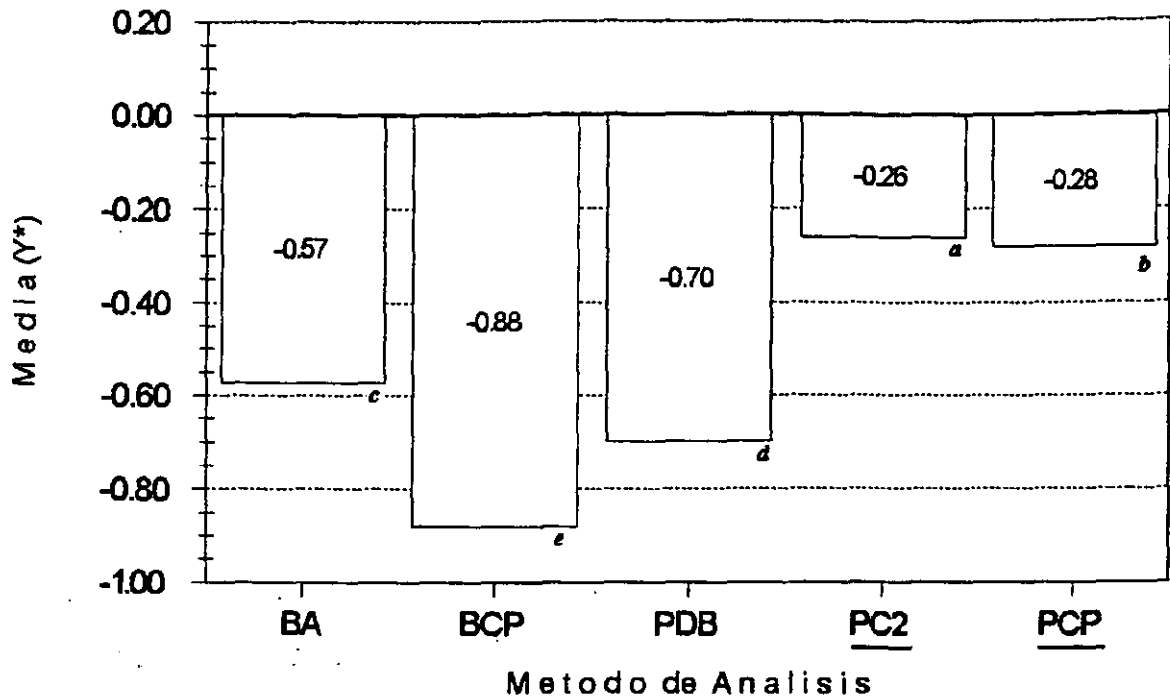


Figura 6.3. Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Lander *et al.* (1938). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.

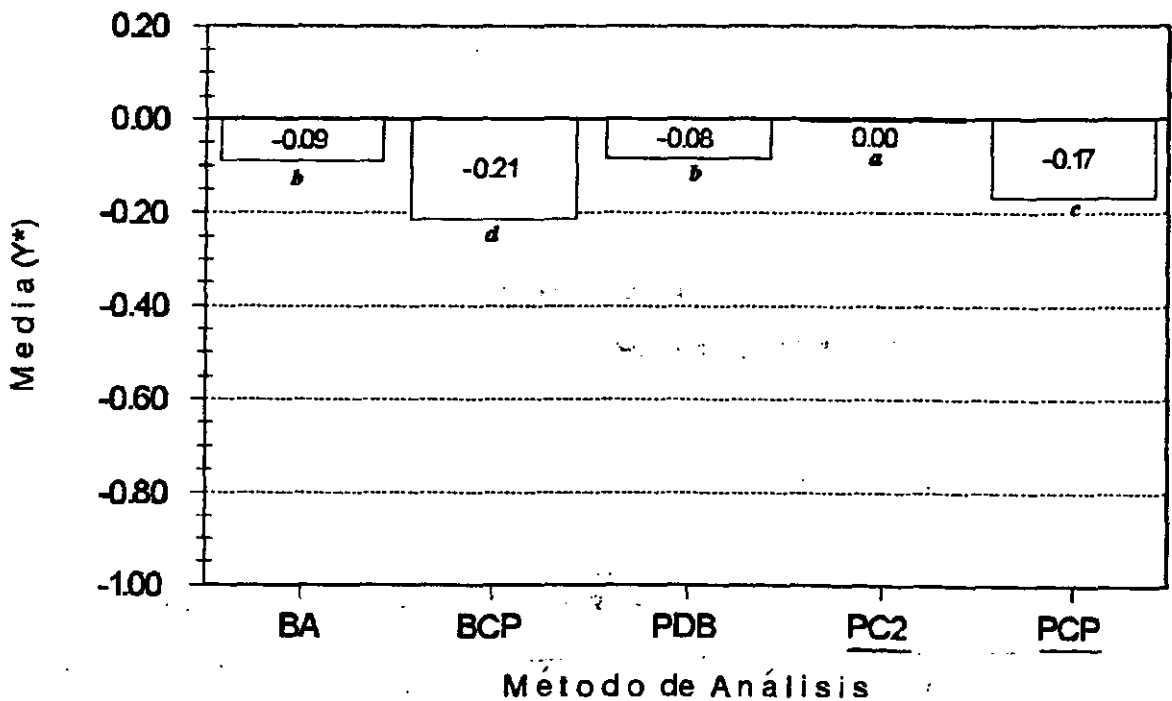


Figura 6.4. Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Mercer y Hall (1911). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.

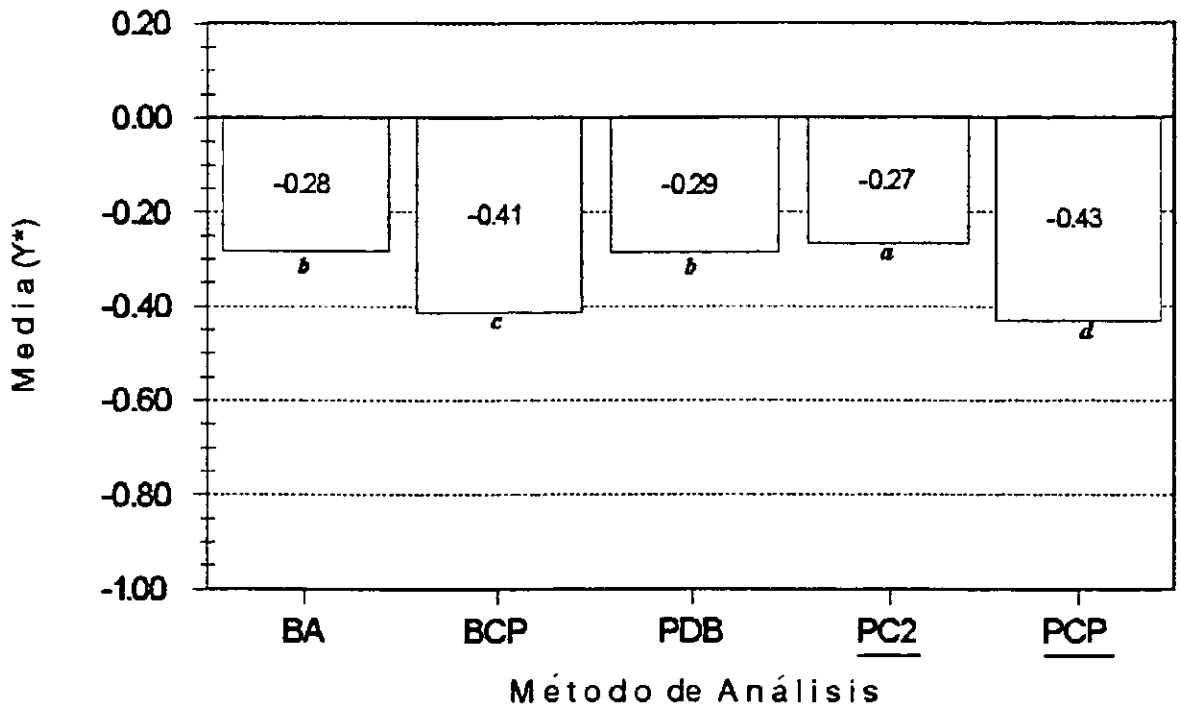


Figura 6.5. Gráfica de barras para la media de la respuesta de los Métodos de Análisis para los datos de Wassom y Kalton (1953). Medias con la misma letra son estadísticamente iguales, prueba de Tukey al 5%.

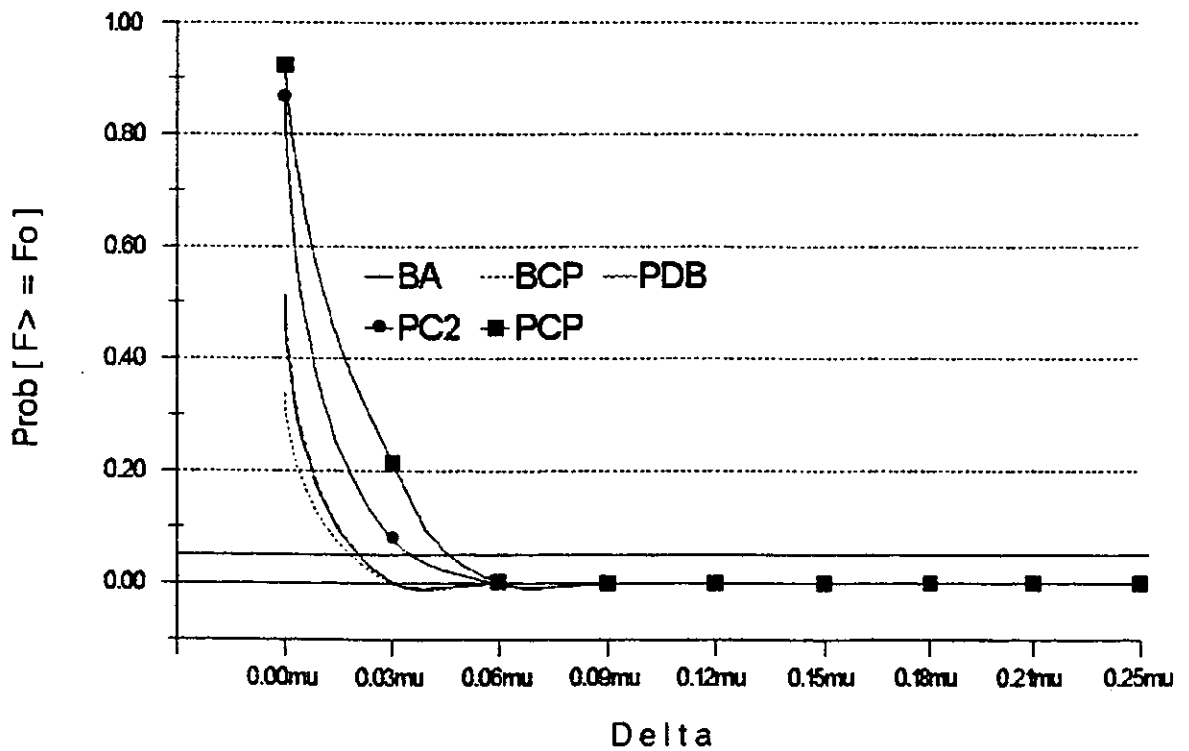


Figura 6.6. Gráfica de dispersión para la media $Prob[F_0 \geq F] = \beta$ según Δ de los Métodos de Análisis para los datos de Baker *et al.* (1952).

En los resultados de los datos de Immer (1932) los cinco métodos se comportan de modo similar, para su menor valor de Δ su valor respectivo de \hat{p} varia entre 0.35 y 0.50 y tienden a disminuir a 0 en Δ_2 . El nivel de significancia de 0.05 lo alcanza para Δ menores a 0.03μ (Fig. 6.7).

Para el segmento de datos de Lander *et al.* (1938) se observa un menor de \hat{p} para los tres Métodos de Refinamiento a los Bloques (BA, BCP y PDB), los cuales tienen un similar y mejor comportamiento. Para Δ_1 su valor de \hat{p} esta alrededor de 0.4, mientras que los otros dos Métodos que Ignoran Bloques su valor inicia en alrededor de 0.8; la significancia de 0.05 lo alcanzan con valores cercanos a Δ_2 en los Métodos de Refinamiento a los Bloques y en 0.06μ para los Métodos que Ignoran Bloques (Fig. 6.8).

En los datos de Mercer y Hall (1911) los cinco métodos se comportan de modo similar, para su menor valor de Δ su valor \hat{p} varia entre 0.45 y 0.60 y tienden a disminuir a 0 en Δ_3 . El nivel de significancia de 0.05 lo alcanzan para valores de Δ ligeramente menores a 0.06 (Fig. 6.9).

En los resultados de los datos de Wassom y Kalton (1953) los cinco métodos se comportan de modo similar, para su menor valor de Δ su valor \hat{p} varia entre 0.375 y 0.50 y tienden a disminuir a 0 en $\Delta_5 = 0.12\mu$. Así mismo, la significancia mínima adecuada (0.05) se observa alredeodr de Δ_4 (Fig. 6.10).

De las Figuras 6.6-6.10 se observa que los conjuntos de datos Immer (1932), Mercer y Hall (1911) y Wassom y Kalton (1953) se comportan de modo similar en el sentido de que no se separan claramente los dos conjuntos de métodos investigados (Fig. 6.7, 6.9 y 6.10), mientras que para los conjuntos de datos de Baker *et al.* (1952) y Lander *et al.* (1938) se observa cierta tendencia a agruparse de los métodos dentro de su respectivo conjunto de métodos (Fig. 6.6 y 6.8).

6.4. Potencia Empírica Mediante la Diferencia de Tratamientos

La capacidad de los métodos para detectar la diferencia entre tratamientos se

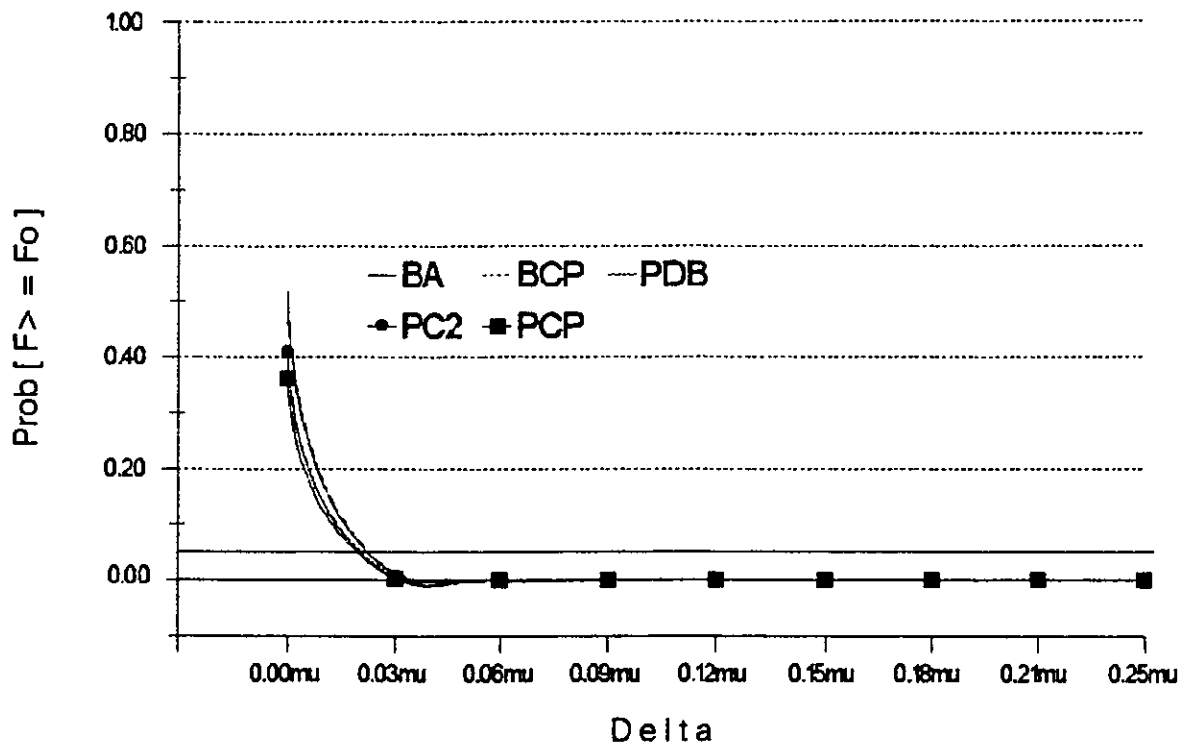


Figura 6.7. Gráfica de dispersión para la media $Prob[F_o \geq F] = \hat{p}$ según Δ de los Métodos de Análisis para los datos de Immer (1932).

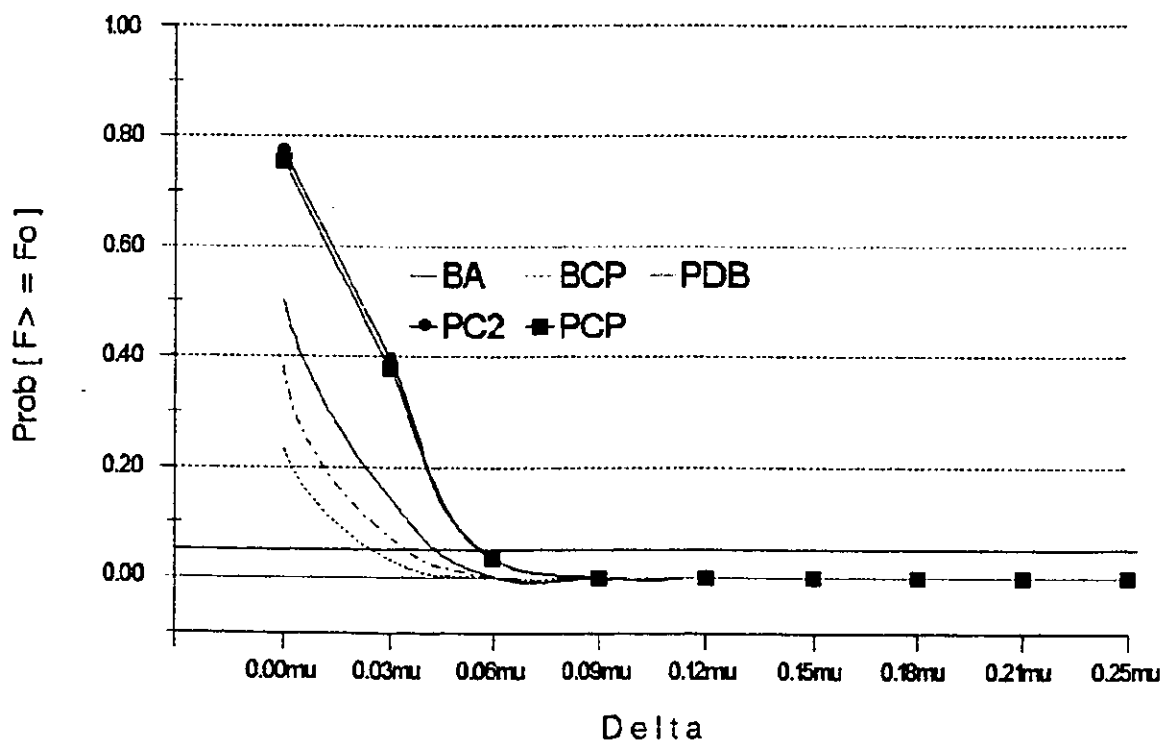


Figura 6.8. Gráfica de dispersión para la media de $Prob[F_o \geq F] = \hat{p}$ por Δ de los Métodos de Análisis para los datos de Lander et al. (1938).

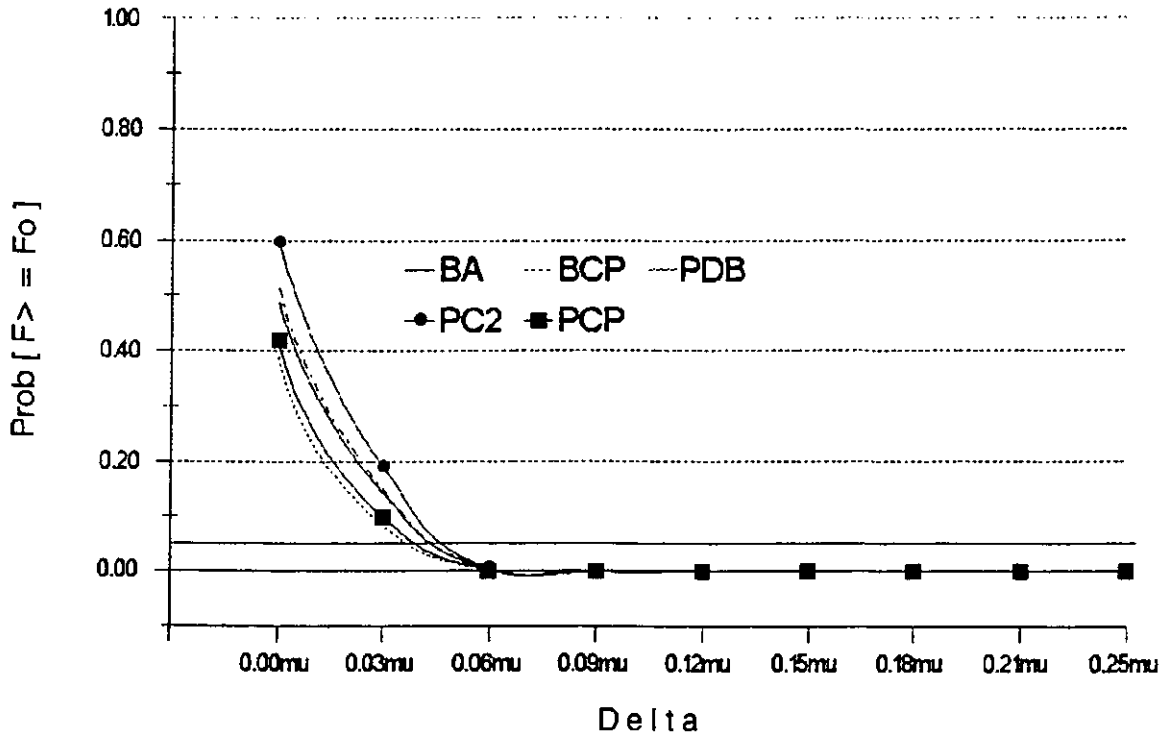


Figura 6.9. Gráfica de dispersión para la media de $Prob[F_0 \geq F] = \hat{p}$ por Δ de los Métodos de Análisis para los datos de Mercer y Hall (1911).

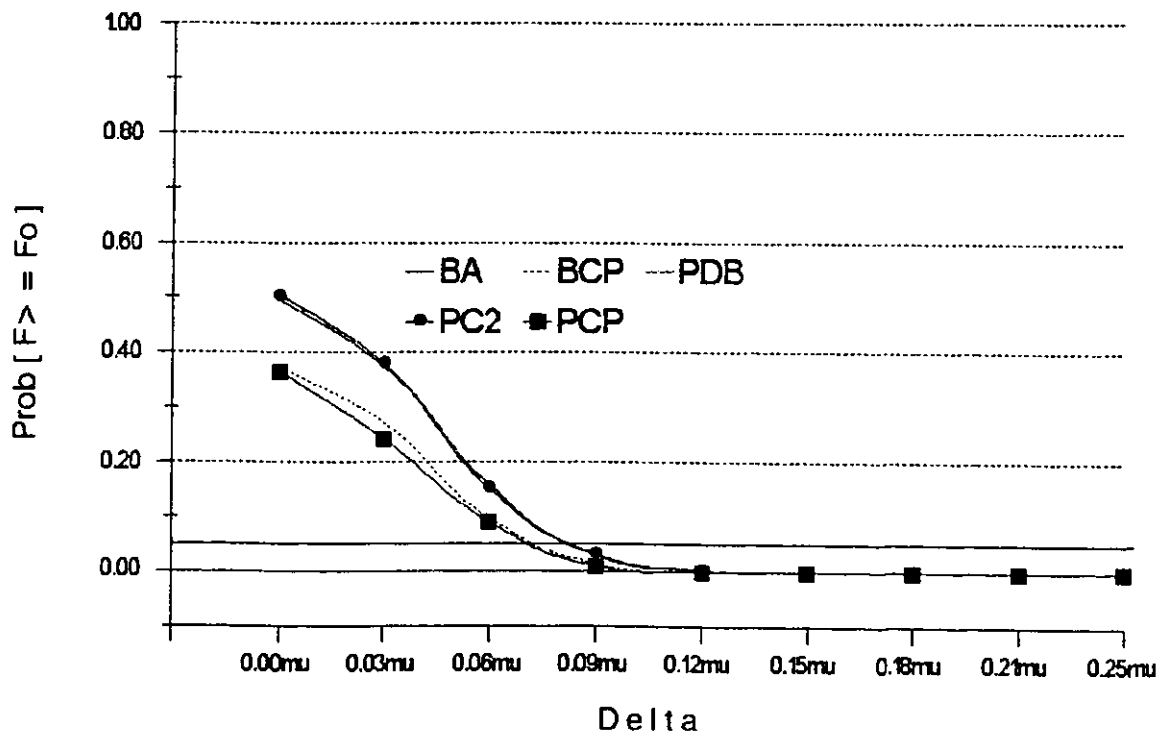


Figura 6.10. Gráfica de dispersión para la media $Prob[F_0 \geq F] = \hat{p}$ por Δ de los Métodos de Análisis para los datos de Wasom y Kalton (1953).

interpreta a partir de la proporción de diferencias significativas detectadas con la Prueba de Comparaciones Múltiples de Medias de Tukey. Los resultados se pueden observar (Anexo en CD) en los Cuadros 9.6a-9.10a para datos de Baker *et al.* (1952), Cuadros 9.11a-9.15a para los datos de Immer (1932), Cuadros 9.16a-9.20a para los datos de Lander *et al.* (1938), Cuadros 9.21a-9.25a para los datos de Mercer y Hall (1911) y en los Cuadros 9.26a-9.30a para los datos de Wassom y Kalton (1953).

Para su interpretación se grafica como fuente de datos principalmente los valores del Promedio de la Proporción de Significancia (*PPS*) de Δ por la magnitud de su incrementos $j\Delta \forall j = 1, \dots, 6$ para cuando $\Delta = 0.03\mu$ como se muestra para el Método de Bloques al Azar para los datos de Baker *et al.* (1952) (Cuadro 9.10a del anexo en CD) en el Cuadro 6.7.

El calculo de $PPS_i \forall i = 1, \dots, 7$ se realiza del modo siguiente:

$$PPS_1 = \frac{0.007 + 0.004 + 0.005 + 0.004 + 0.006 + 0.006}{6} = 0.005 ,$$

$$PPS_2 = 0.012 ,$$

$$PPS_3 = \frac{0.062 + 0.055}{2} = 0.059 ,$$

...

$$PPS_7 = \frac{0.938 + 0.947}{2} = 0.943$$

Dado que para algunos conjuntos de datos el valor de *PPS* podría no indicar su tendencia, se calcula su valor para las siguientes Distancia Entre Medias: 0.24, 0.36, 0.45, 0.50, 0.75 y 1.00, para su cálculo se considera al Cuadro 6.7 como un cuadro de doble entrada. Cabe mencionar que para el ejemplo ilustrado algunos valores son redundantes y cuando esto sucede no se grafica, dependiendo de la fuente del conjunto de datos.

$$PPS_{0.24} = \frac{PPS_{34} + PPS_{52}}{2} = \frac{1.000 + 1.000}{2} = 1.000,$$

$$PPS_{0.36} = \frac{PPS_{36} + PPS_{44} + PPS_{53} + PPS_{72}}{4},$$

$$PPS_{0.45} = \frac{PPS_{45} + PPS_{63}}{2},$$

$$PPS_{0.50} = PPS_{92}, PPS_{0.75} = PPS_{93} \text{ y } PPS_{1.00} = PPS_{94}.$$

Cuadro 6.7. Representación del PPS para el Método de Bloques al Azar en el segmento del ensayo de uniformidad de Baker et al. (1952)

Δ	(Significancia de las medias, de Tukey)						PPS
	1 Δ	2 Δ	3 Δ	4 Δ	5 Δ	6 Δ	
0.00mu	0.007	0.004	0.005	0.004	0.006	0.006	0.005
0.03mu	0.012	0.062	0.200	0.427	0.746	0.938	0.012
0.06mu	0.055	0.459	0.956	1.000	1.000	1.000	0.059
0.09mu	0.221	0.955	1.000	1.000	1.000	1.000	0.211
0.12mu	0.439	1.000	1.000	1.000	1.000	1.000	0.433
0.15mu	0.751	1.000	1.000	1.000	1.000	1.000	0.749
0.18mu	0.947	1.000	1.000	1.000	1.000	1.000	0.943
0.21mu	0.998	1.000	1.000	1.000	1.000	1.000	
0.25mu	1.000	1.000	1.000	1.000	1.000	1.000	

Al graficar el PPS x Δ (en la parte de la Distancia Entre Medias) para los cinco métodos investigados, se encontró que para los datos de Baker *et al.* (1952) el PPS tiende a valores mayores para los tres Métodos de Refinamiento a los Bloques (BA, BCP y PDB), sobresaliendo el método de BCP; mientras que para los dos Métodos que Ignoran Bloques (Papadakis Clásico en 2 Dimensiones y Polinomio en Coordenadas de Posición) tienden a tener valores menores en todo para el intervalo de Distancia Entre Medias considerado. Todos los métodos tienden a hacerse asintóticos alrededor de 0.36 de la Distancia Entre Medias. Se separan claramente los 2 conjuntos de métodos investigados (Fig. 6.11).

En la gráfica correspondiente a los datos de Immer (1932) el valor de PPS tiende a tener valores cercanos a 1.0 a la Distancia Entre Medias de 0.24 en los cinco métodos considerados, con un ligero mejor comportamiento de los Métodos que Ignoran Bloques (PC2 y PCP) con respecto a los tres Métodos de Refinamiento a los Bloques (BA, BCP y PDB) (Fig. 6.12).

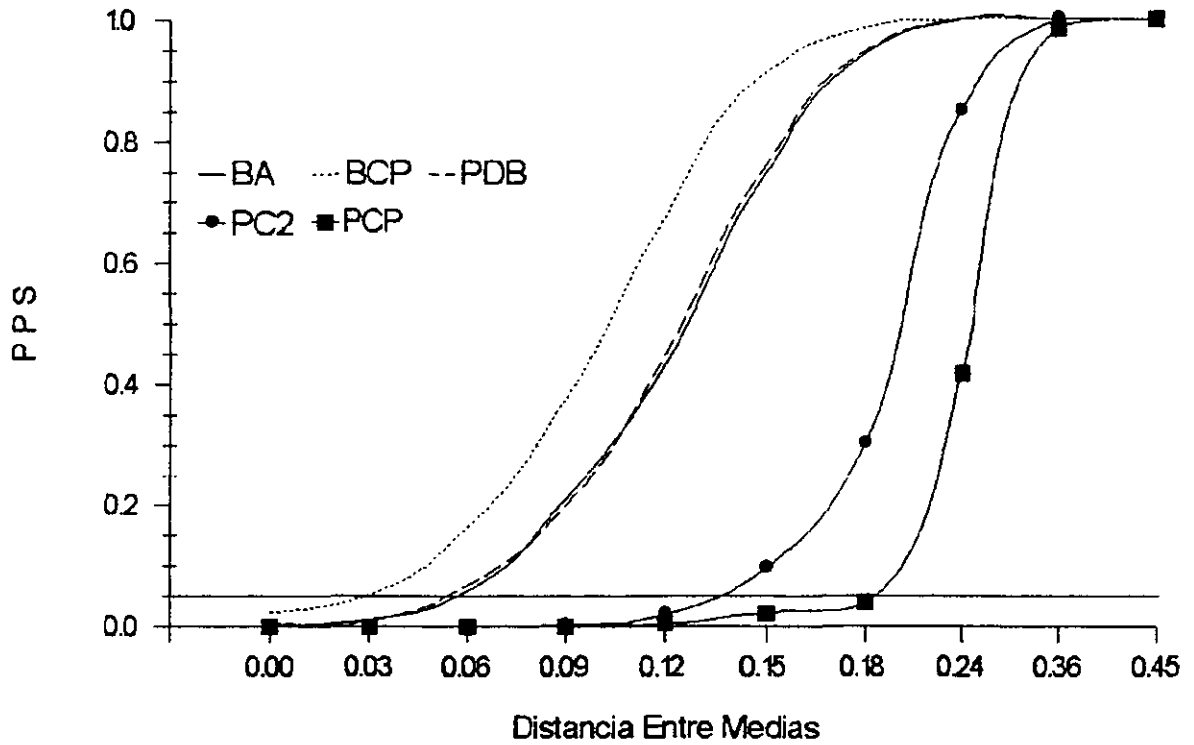


Figura 6.11. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Baker et al. (1952).

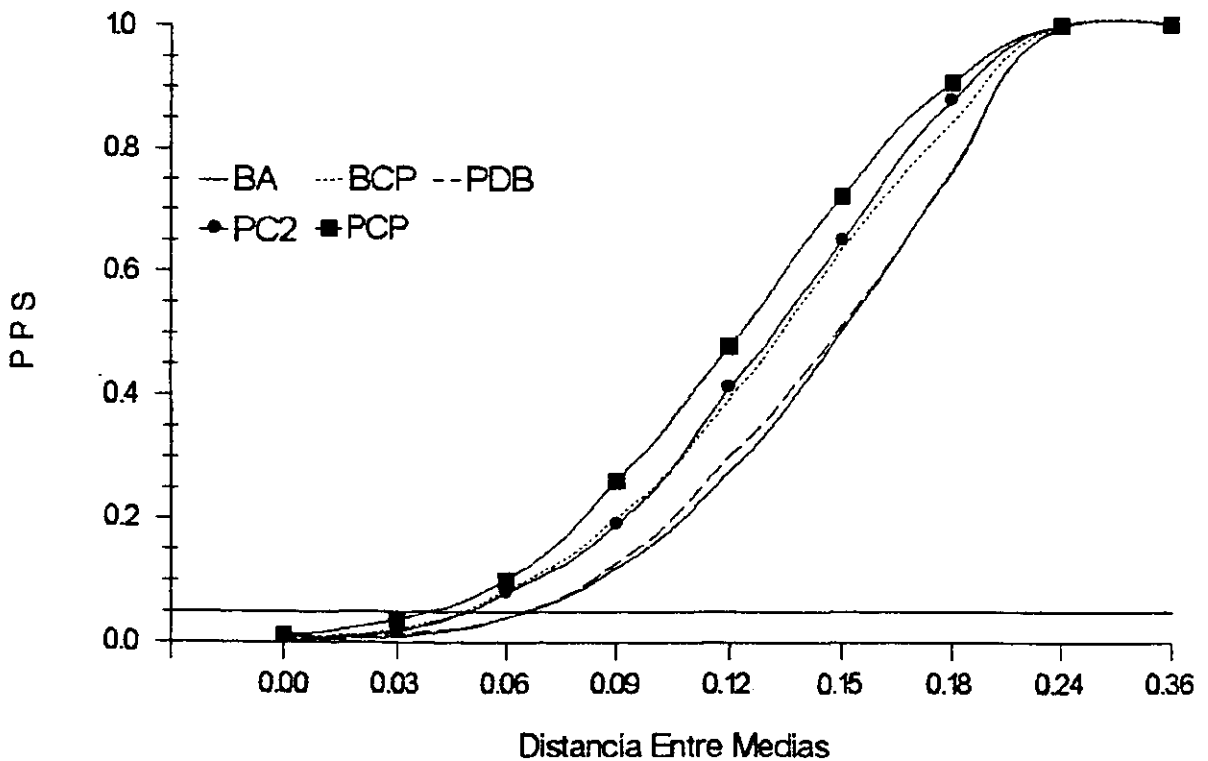


Figura 6.12. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Immer (1932).

En los datos de Lander *et al.* (1938) el PPS tiende a tener valores más mayores para los tres Métodos de Refinamiento a los Bloques (BA, BCP y PDB) los cuales se separan claramente de los Métodos que Ignoran Bloques (PC2 y PCP) que tienden a tener valores más bajos y cercanos en todo el intervalo de Distancia Entre Medias considerado. De manera notoria se observa un mejor comportamiento, con valores más alto de PPS, del método de BCP, así mismo los 5 métodos alcanzan su nivel asintótico en alrededor de 0.05 de la Distancia Entre Medias. También se observa una clara separación entre los 2 conjuntos de métodos investigados (Figura 6.13).

Los datos para Mercer y Hall(1911) (Fig. 6.14) y Wassom y Kalton (1953) (Fig. 6.15) no presentan para los cinco métodos (BA, BCP, PDB, PC2 y PCP) separación clara de los dos conjuntos de métodos (Métodos de Refinamiento a los Bloques y Métodos que Ignoran Bloques) o alguno de ellos en forma individual. Los valores de PPS máximo (que hace la dispersión asintótica) se alcanzan para la Distancia Entre Medias de 0.45 y 1.00 respectivamente.

En ninguno de los cinco conjuntos de datos investigados y ninguno de los cinco métodos el valor de PPS cercano a 0.05 (línea horizontal en las Fig. 6.11-6.15) para cuando la Distancia Entre Medias es mínima (0.00).

6.5. Sesgo Empírico de la Diferencia de Tratamientos

Para su análisis se utiliza como fuente de datos la combinación de Δ (9 niveles de incremento), Diferencia del Par Medias (6 niveles de pares de diferencias) y Métodos de Análisis (5 niveles de métodos) que dan lugar a un cuadro de triple entrada con 270 respuestas de Sesgo Empírico. Como ejemplo, en el Cuadro 6.8 se detalla la respuesta para los Datos de Immer (1932).

Los resultados de las simulaciones, para los cinco métodos, se pueden observar en el Anexo (en CD) en los Cuadros 9.6b-9.10b para datos de Baker *et al.* (1952), Cuadros 9.11b-9.15b para los datos de Immer (1932), Cuadros 9.16b-9.20b para los datos de Lander *et al.*(1938), Cuadros 9.21b-9.25b para los datos de Mercer y Hall (1911) y en los Cuadros 9.26b-9.30b para los datos de Wassom y Kalton (1953).

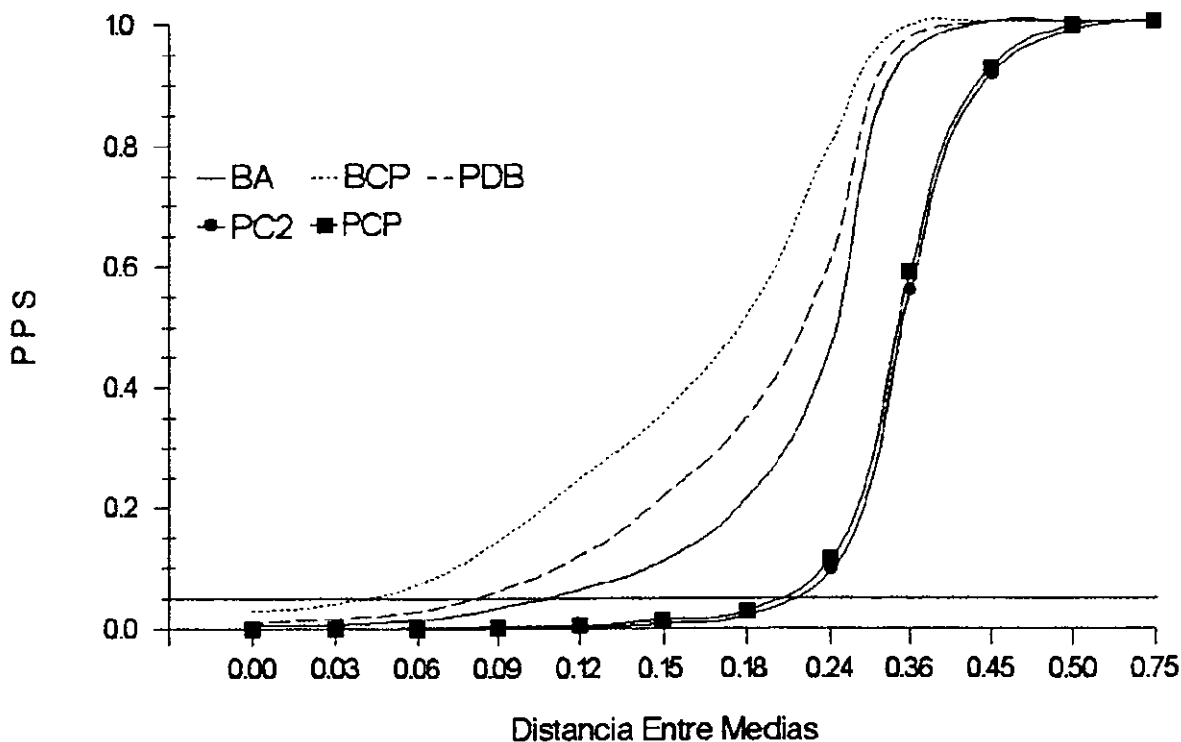


Figura 6.13. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Lander et al. (1938).

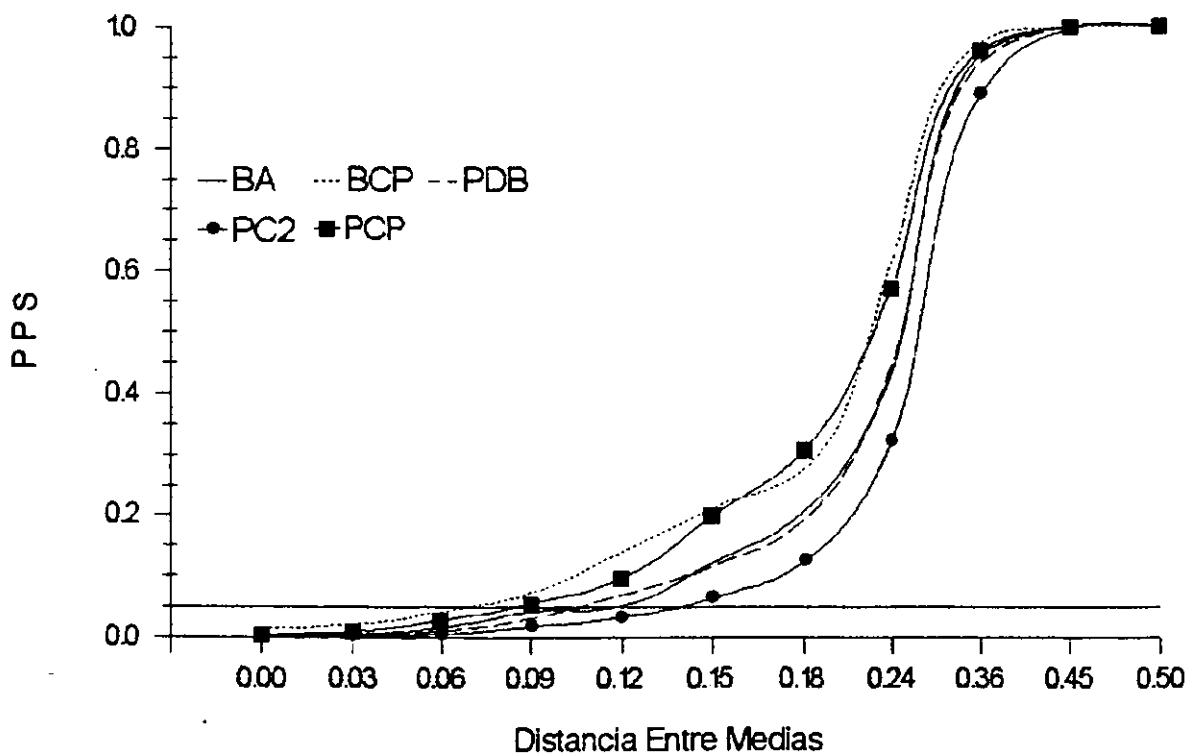


Figura 6.14. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Mercer y Hall (1911).

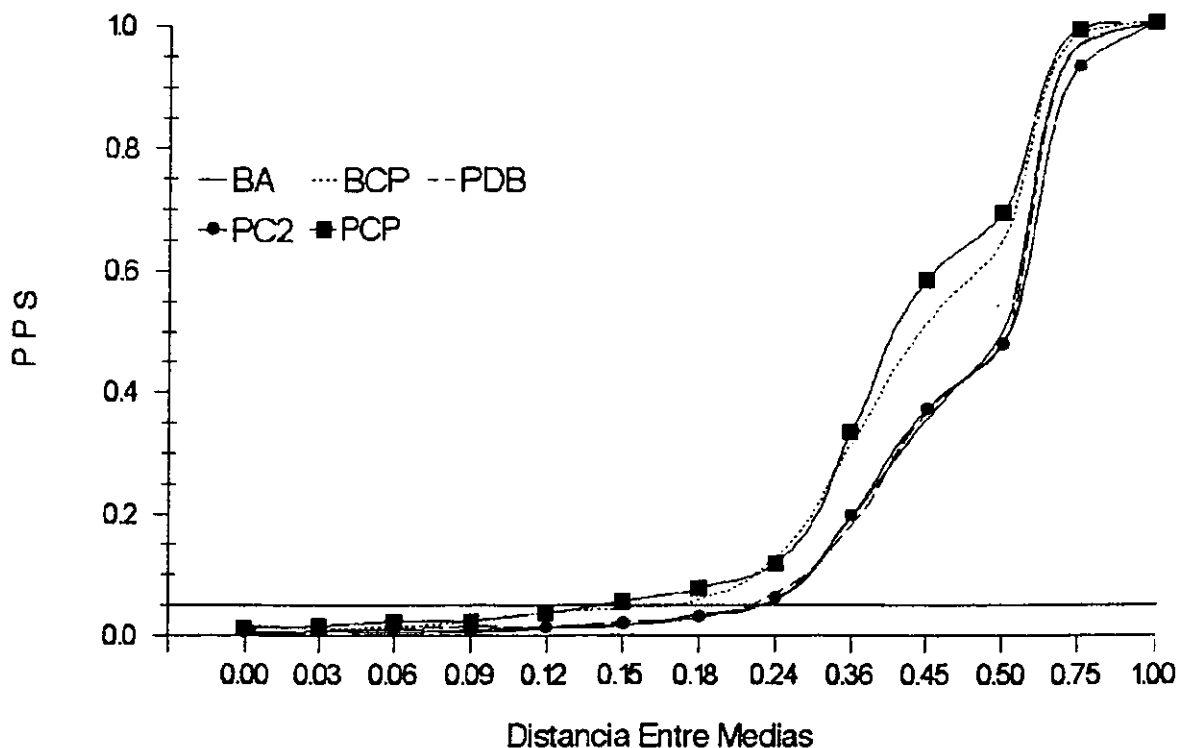


Figura 6.15. Gráfica de dispersión del Promedio de la Proporción de la Significancia por la Distancia Entre Medias para los datos de Wassom y Kalton (1953).

Con estos resultados se realiza un Análisis de Varianza para el modelo

$$Y_{ijk}^S = \mu + \text{Metodo}_i + \Delta_j + \text{Diferencia}S_k + (\text{Metodo} * \Delta)_{ij} + (\text{Metodo} * \text{Diferencia}S)_{ik} + (\Delta * \text{Diferencia}S)_{jk} + \varepsilon_{ijk} \quad \forall i = 1, \dots, 6, j = 1, \dots, 9 \text{ y } k = 1, \dots, 6 \quad (6.3)$$

donde

μ : Media General.

Y_{ijk}^S : Sesgo en el i -ésimo método, j -ésimo incremento y la k -ésima diferencia de par de medias.

Metodo_i : Efecto del i -ésimo método.

Δ_j : Efecto del j -ésimo incremento.

$\text{Diferencia}S_k$: Efecto de la k -ésima diferencia de par de medias.

$(\text{Metodo} * \Delta)_{ij}$: Efecto de la ij -ésima combinación de los niveles de los efectos de método de incremento.

$(\text{Metodo} * \text{Diferencia}S)_{ik}$: Efecto de la ik -ésima combinación de los niveles de los efectos de método y diferencia del par de medias.

$(\Delta * \text{Diferencia}S)_{jk}$: Efecto de la jk -ésima combinación de los niveles de los efectos de incremento y diferencia del par de medias.

ε_{ijk} : ijk -ésimo error aleatorio del modelo.

Cuadro 6.8. Sesgo Empírico y sus tres criterios de clasificación (Immer, 1932).

Método	Δ	Diferencia de Par de Medias					
		1-2	1-3	1-4	1-5	1-6	1-7
B A	0.00mu	- 0.1430	- 0.4616	- 0.1245	- 0.4004	- 0.3537	- 0.1634
	0.03mu	- 9.7821	- 19.2667	- 28.4195	- 38.2654	- 47.9649	- 57.6300
	0.06mu	- 19.1413	- 38.2106	- 57.2684	- 76.4260	- 95.2261	- 114.6691
	0.09mu	- 28.5453	- 57.1499	- 85.8520	- 114.5248	- 143.0870	- 171.2337
	0.12mu	- 38.0883	- 76.0798	- 114.3116	- 152.2533	- 190.4189	- 228.7624
	0.15mu	- 47.4925	- 95.1771	- 142.6909	- 190.4515	- 238.1161	- 285.6360
	0.18mu	- 57.1202	- 114.3188	- 171.8243	- 228.5642	- 286.0573	- 343.1369
	0.21mu	- 66.8331	- 133.3850	- 200.2410	- 267.0791	- 333.6453	- 400.2460
	0.25mu	- 79.3634	- 158.8084	- 238.2548	- 317.5369	- 397.3971	- 476.7067
C B P	0.00mu	- 0.2355	- 0.1476	- 0.5189	- 0.0812	- 0.4862	- 0.2193
	0.03mu	- 9.3006	- 19.0646	- 28.3286	- 37.9959	- 47.4418	- 56.8762
	0.06mu	- 19.3416	- 38.5484	- 57.4283	- 76.2914	- 95.6475	- 114.8297
	0.09mu	- 28.4733	- 56.9205	- 85.8804	- 114.1618	- 142.8094	- 171.2893
	0.12mu	- 37.8599	- 76.0779	- 114.2286	- 152.0815	- 190.3193	- 228.6521
	0.15mu	- 47.7341	- 95.4167	- 142.8062	- 190.8072	- 238.0796	- 285.7844
	0.18mu	- 57.0958	- 114.4662	- 171.4925	- 228.7495	- 286.0898	- 343.2986
	0.21mu	- 66.7022	- 133.5302	- 200.2288	- 266.7240	- 333.4630	- 400.3335
	0.25mu	- 79.5835	- 158.7802	- 238.3280	- 317.6109	- 397.1298	- 476.6020
P D B	0.00mu	- 0.1410	- 0.4242	0.2620	0.2003	- 0.1387	0.0584
	0.03mu	- 9.6506	- 19.0761	- 28.8086	- 38.4838	- 47.6857	- 57.6844
	0.06mu	- 19.0381	- 37.9416	- 57.0322	- 76.1494	- 95.1760	- 114.4524
	0.09mu	- 28.4896	- 57.3525	- 85.7055	- 114.0263	- 142.8087	- 170.9628
	0.12mu	- 38.5475	- 76.7448	- 114.3746	- 152.7662	- 191.0686	- 229.0488
	0.15mu	- 48.0955	- 95.3433	- 143.2370	- 190.7538	- 238.3593	- 286.2498
	0.18mu	- 57.4031	- 114.3386	- 171.8425	- 228.5628	- 285.8779	- 343.0045
	0.21mu	- 66.7534	- 133.3338	- 200.1935	- 266.8921	- 333.4071	- 400.2325
	0.25mu	- 79.5118	- 158.9941	- 238.2307	- 317.5668	- 397.1506	- 476.5767
P C 2	0.00mu	0.2984	0.2344	0.2055	0.1225	0.1547	- 0.0072
	0.03mu	- 9.5801	- 19.1548	- 28.8205	- 38.1519	- 47.7964	- 57.1185
	0.06mu	- 19.1029	- 38.2067	- 57.5844	- 76.0793	- 94.9696	- 114.5407
	0.09mu	- 28.4458	- 56.8467	- 85.5950	- 114.0396	- 142.5531	- 171.3904
	0.12mu	- 38.4662	- 76.3950	- 114.7509	- 152.7985	- 190.6510	- 228.8738
	0.15mu	- 47.3513	- 95.3806	- 142.7958	- 190.8453	- 238.3342	- 285.7389
	0.18mu	- 57.3742	- 114.2679	- 171.6250	- 228.7021	- 286.1590	- 343.1832
	0.21mu	- 66.4646	- 133.5935	- 200.0612	- 266.7384	- 333.5661	- 400.2684
	0.25mu	- 79.5494	- 158.8660	- 237.8815	- 317.9139	- 397.0144	- 476.1005
P C P	0.00mu	- 0.2078	0.2085	0.0863	0.1873	- 0.1261	0.1102
	0.03mu	- 9.5669	- 19.2758	- 28.5438	- 38.1832	- 47.7552	- 57.2657
	0.06mu	- 18.8206	- 38.4131	- 57.1266	- 76.1148	- 95.2931	- 114.3271
	0.09mu	- 28.7848	- 57.4181	- 85.9479	- 114.6405	- 142.9901	- 171.5977
	0.12mu	- 38.1151	- 76.3644	- 114.5374	- 152.4727	- 190.9627	- 228.9670
	0.15mu	- 47.6682	- 95.0127	- 143.0461	- 190.4016	- 238.5342	- 285.9871
	0.18mu	- 57.4641	- 114.1664	- 171.2848	- 228.7536	- 285.7091	- 343.1540
	0.21mu	- 66.7861	- 133.5320	- 200.3933	- 266.9803	- 333.4017	- 400.0978
	0.25mu	- 79.2050	- 158.7404	- 238.4540	- 317.3380	- 396.9536	- 476.1998

De los efectos del modelo (6.3) se investiga las interacciones de primer orden de *Metodo* x Δ y *Metodo* x *Diferencia*S, por ser los efectos más importante para el estudio. En el Cuadro 6.9 se muestran estadísticas del modelo y significancia de las interacciones estudiadas para los conjuntos de ensayos de uniformidad, se observa alta significancia ($p \leq 0.01$ o $Prob[F \geq F_0]$ menor de 0.0001) para el modelo en todos los conjuntos de datos, también se observa en todos los conjuntos de datos un valor alto, alrededor de 99.99%, del Coeficiente de Determinación (R^2). La significancia de la interacción *Metodo* x Δ es altamente significativa ($p \leq 0.01$) para todos los conjuntos de datos; mientras que la interacción *Metodo* x *Diferencia*S es no significativa ($p \geq 0.05$) para cinco conjuntos de datos estudiados.

Cuadro 6.9. Estadísticas del modelo (6.4) y significancia de la interacción para los conjuntos de datos.

Fuente de Datos	Modelo				Prob[F \geq Fo] = \hat{p}	
	CME	Fo	Prob[F \geq Fo]	R ²	Met x Δ	Met x Dif
Baker <i>et al.</i> (1952)	30.4600	1 909 367.52	<0.0001**	0.9998	0.0001**	0.9204 ^{NS}
Immer (1932)	0.0280	1 231 291.28	<0.0001**	0.9996	0.0001**	0.9706 ^{NS}
Lander <i>et al.</i> (1938)	0.0041	552 711.81	<0.0001**	0.9997	0.0001**	0.2574 ^{NS}
Mercer y Hall (1911)	0.0001	413 912.07	<0.0001**	0.9996	0.0001**	0.3238 ^{NS}
Wassom y Kalton (1953)	0.0004	123 442.78	<0.0001**	0.9995	0.0001**	0.3017 ^{NS}

** : Indica una alta significancia del efecto respectivo para un Error Tipo I del 1%.

NS: No Significancia del efecto respectivo para un máximo del Error Tipo I del 5%.

Para la interacción de primer orden significativa, para los respectivos conjuntos de datos, se realiza una gráfica de interacción con los niveles respectivos de *Metodo* y Δ para la media del Sesgo Empírico y además se realiza una Prueba de Comparaciones múltiples de Medias, con un nivel de significancia del 5%, para los cinco Métodos por cada uno de los nueve incrementos de Δ .

Para los datos de Baker *et al.* (1952), que tienen una amplia variación de la media del Sesgo Empírico en el intervalo de -11424.7 a 9.15 y aumenta el sesgo al incrementarse Δ (Fig. 6.16). No se observa claramente la separación de los métodos por lo amplio del intervalo de valores. La significancia de Tukey para los Métodos en cada uno de los Δ , muestran que sólo para $\Delta_1 = 0.00mu$ hay diferencia entre los métodos (Cuadro 6.10) con el mejor valor de Sesgo Empírico (cerca a cero) para los métodos del grupo *ab* en la cual está el método de Polinomios en Coordenadas de Posición(1.7) que pertenece al conjunto

de Métodos que Ignoran Bloques.

Cuadro 6.10. Prueba de Tukey para el Sesgo Empírico de los cinco Métodos por cada uno de los niveles de Δ para los datos de Baker *et al.* (1952).

Delta									
0.00mu		0.03mu		0.06mu		0.09mu		0.12mu	
Método	SE ¹	Método	SE ¹	Método	SE ¹	Método	SE ¹	Método	SE ¹
BCP	9.2a	PC2	-1364.6a	PC2	-2730.7a	BCP	-4101.9a	BA	-5475.7a
PCP	1.7ab	BA	-1368.4a	PCP	-2731.5a	PCP	-4106.8a	PCP	-5478.3a
PC2	-2.1 bc	PCP	-1370.0a	PDB	-2739.5a	PC2	-4110.7a	BCP	-5480.1a
PDB	-3.5 bc	BCP	-1370.9a	BA	-2743.7a	BA	-4112.7a	PDB	-5483.0a
BA	-7.3 c	PDB	-1371.0a	BCP	-2745.3a	PDB	-4114.7a	PC2	-5486.8a

¹Medias con la misma letra son estadísticamente iguales con un nivel de significancia del 5% en la Prueba de Comparaciones Múltiples de Medias de Tukey.

Cuadro 6.10. Continuación.

Delta							
0.15mu		0.18mu		0.21mu		0.25mu	
Método	SE ¹	Método	SE ¹	Método	SE ¹	Método	SE ¹
BA	-6840.0a	BA	-8210.9a	PCP	-9583.9a	PDB	-11396.5a
PCP	-6841.0a	PDB	-8216.1a	BCP	-9589.2a	BCP	-11408.9a
BPC	-6850.1a	BCP	-8223.3a	BA	-9592.0a	PCP	-11411.8a
PC2	-6851.4a	PCP	-8223.6a	PDB	-9594.2a	BA	-11416.0a
PDB	-6854.2a	PC2	-8225.1a	PC2	-9598.8a	PC2	-11424.7a

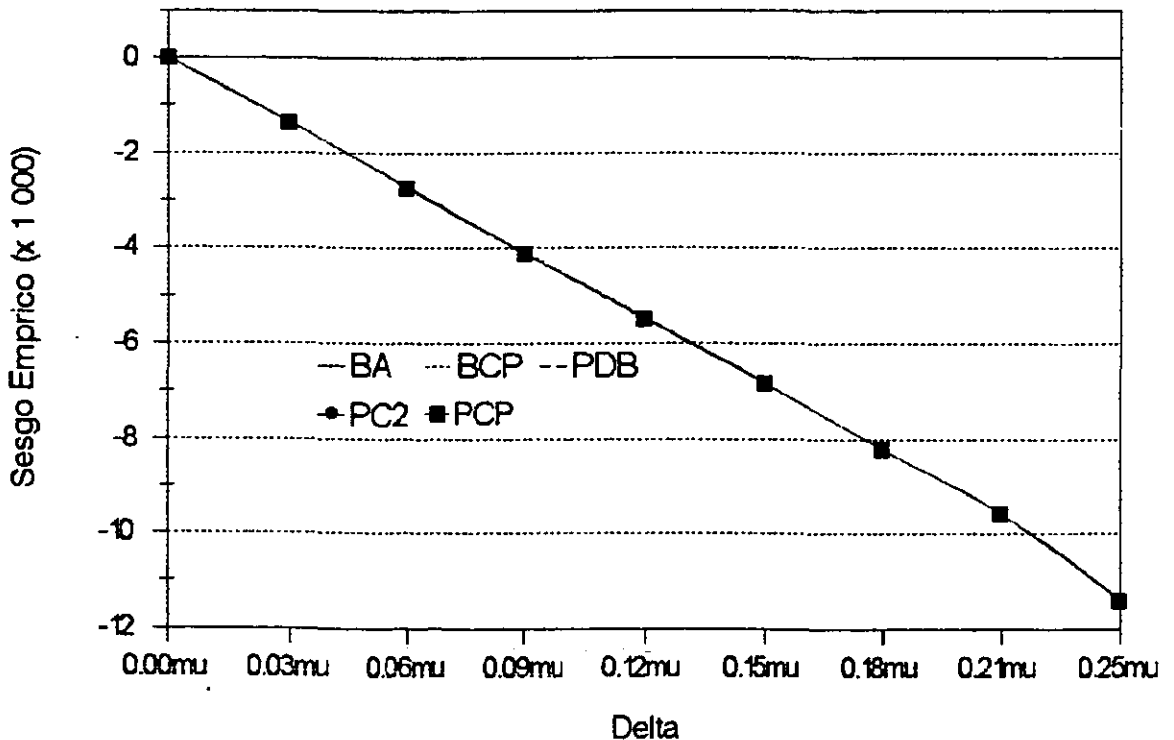


Figura 6.16. Gráfica de dispersión para la media del Sesgo Empírico en la interacción *Metodo* x Δ para los datos de Baker *et al.* (1952).

La interacción para los datos de Immer (1932), Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton (1953) (Figuras 6.17-6.20) muestran la misma tendencia de la media del Sesgo Empírico para los 5 métodos investigados en los diferentes niveles de Δ , es decir, disminuye el Sesgo Empírico a medida que aumenta Δ .

Con respecto a la Prueba de Tukey para los métodos en cada nivel de Δ , se observa significancia sólo para $\Delta_1 = 0.00mu$ en los conjuntos de datos de Immer (1932), Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton (1953) (Cuadro 6.11). El mejor Sesgo para los datos de Immer (1932) se ubica en el método de Papadakis Dentro de Bloques (-0.0305, grupo de medias *ab*), para Lander *et al.* (1938) la magnitud del Sesgo mejor es de 0.0030 en el método de Polinomio en Coordenadas de Posición (grupo *ab*), en Mercer y Hall (1938) con 0.0014 de Sesgo mejor para el método Polinomio en Coordenadas de Posición (grupo *ab*) y para Wassom y Kalton (1953) el mejor Sesgo incluye a 4 métodos, los cuales son los métodos de Polinomio en Coordenadas de Posición (0.096), Bloques al Azar (0.0136) y los dos métodos que incluye el conjunto de Métodos que Ignoran Bloques (grupo *a*).

Cuadro 6.11. Prueba de Tukey para el Sesgo Empírico en $\Delta_1 = 0.00mu$ con los 5 métodos para 4 fuentes de datos.

Immer (1932)		Lander <i>et al.</i> (1938)		Mercer y Hall (1938)		Wassom y Kalton (1953)	
Método	SE ¹	Método	SE ¹	Método	SE ¹	Método	SE ¹
PC2	0.1681a	PC2	0.0293a	BA	0.0099a	PCP	0.0196a
PCP	0.0431a	PCP	0.0030ab	PDB	0.0070a	BA	0.0136a
PDB	-0.0305ab	BA	-0.0178 b	PCP	0.0014ab	PC2	0.0098a
BA	-0.2744 b	BCP	-0.0209 b	PC2	-0.0064 b	BCP	-0.0126ab
BCP	-0.2814 b	PDB	-0.0298 b	BCP	-0.0077 b	PDB	-0.0257 b

¹Medias con la misma letra son estadísticamente iguales con un nivel de significancia del 5% en la Prueba de Comparaciones Múltiples de Medias de Tukey. SE: Sesgo Empírico.

Cabe mencionar que en la interacción *Metodos* \times *Diferencias* para las cinco fuentes de datos mantuvieron una tendencia similar a la interacción *Metodos* \times Δ , es decir, la media del Sesgo Empírico tiende a disminuir a medida que aumenta la diferencia entre las medias y la magnitud de la media es similar entre métodos para nivel de *Diferencias*. Los valores medios mayores (alredor de -9500) se encontraron en los datos de Baker *et al.* (1952) y los menores (alredor de -1) se encontraron en los datos de Mercer y Hall (1911).

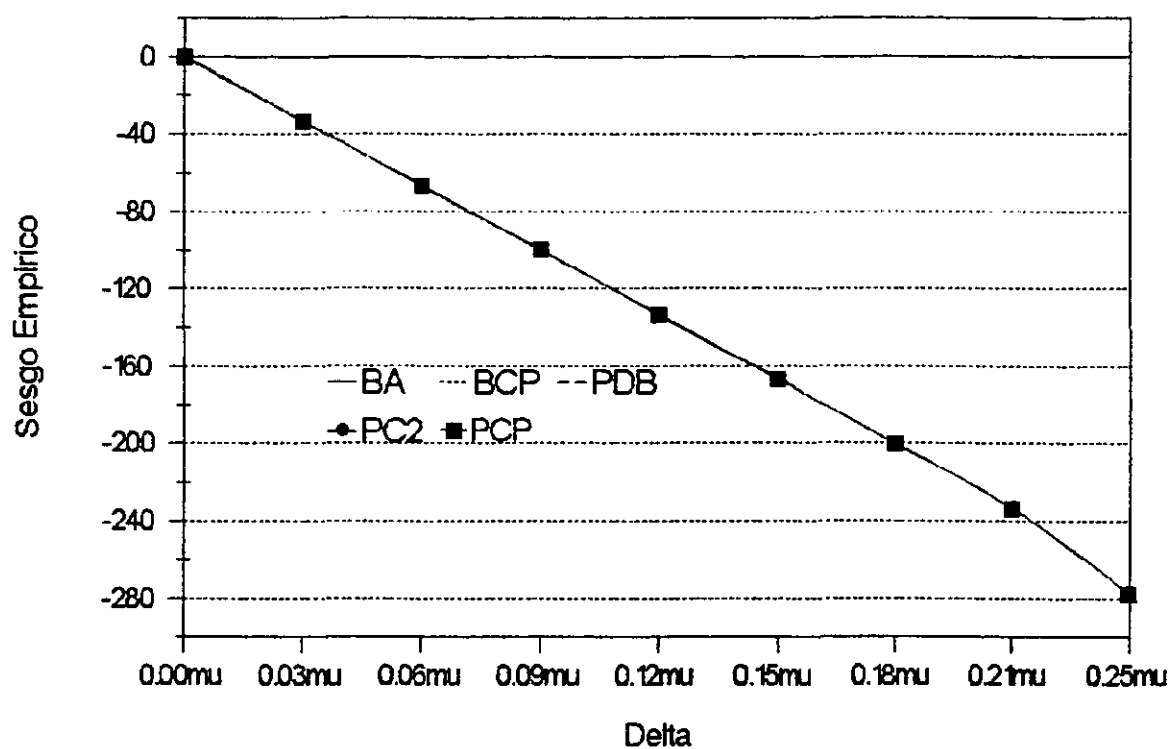


Figura 6.17. Gráfica de dispersión para la media del Sesgo Empírico en la interacción *Metodo x Δ* para los datos de Immer (1932).

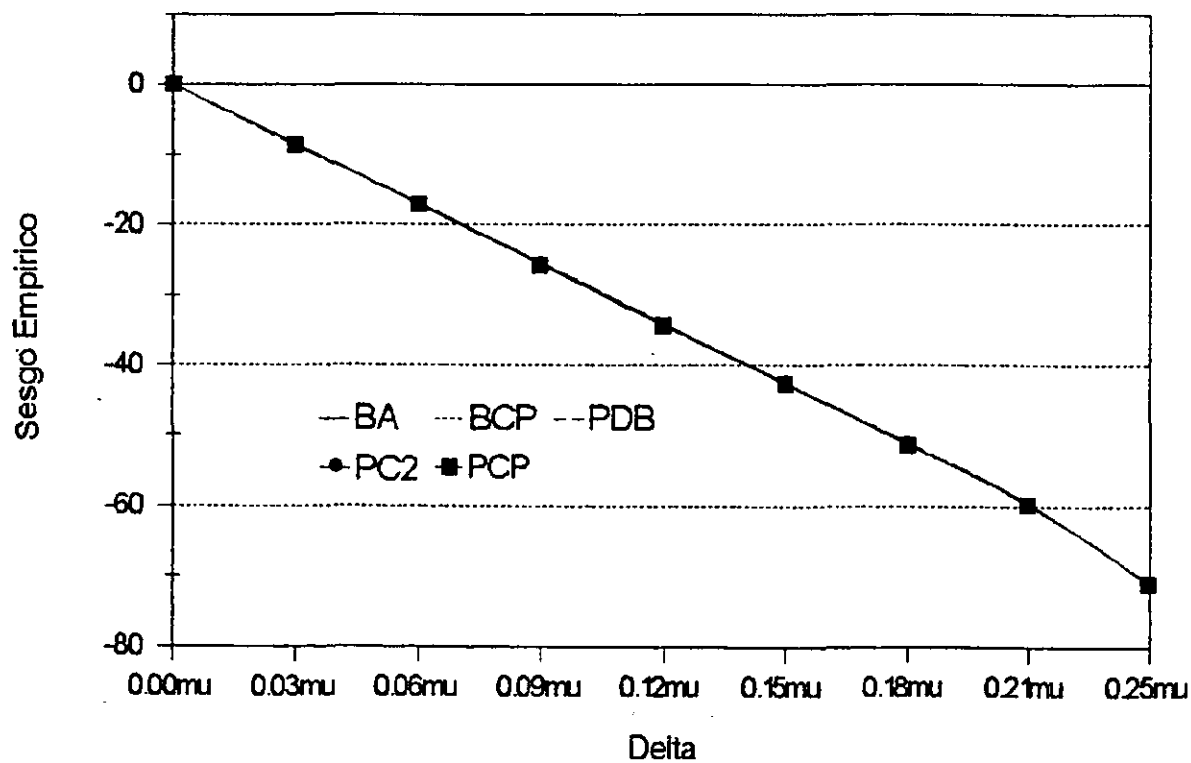


Figura 6.18. Gráfica de dispersión para la media del Sesgo Empírico en la interacción *Metodo x Δ* para los datos de Lander et al. (1938).

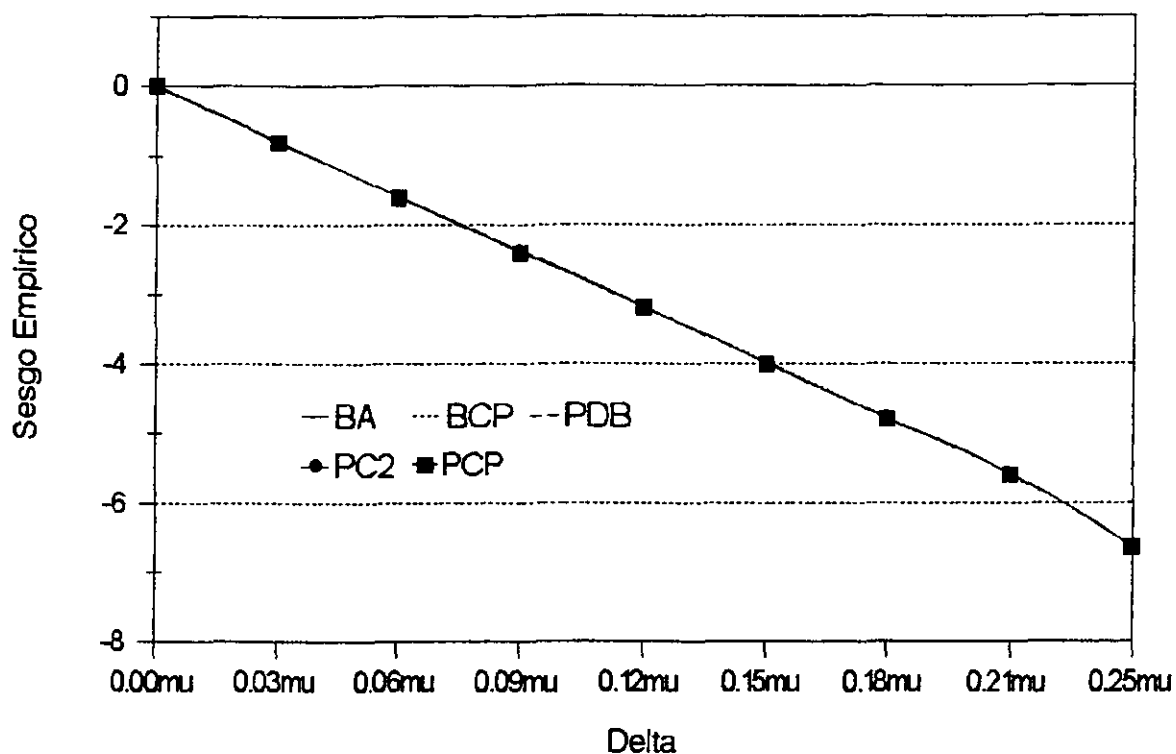


Figura 6.19. Gráfica de dispersión para la media del Sesgo Empírico en la interacción *Metodox* Δ para los datos de Mercer y Hall (1911).

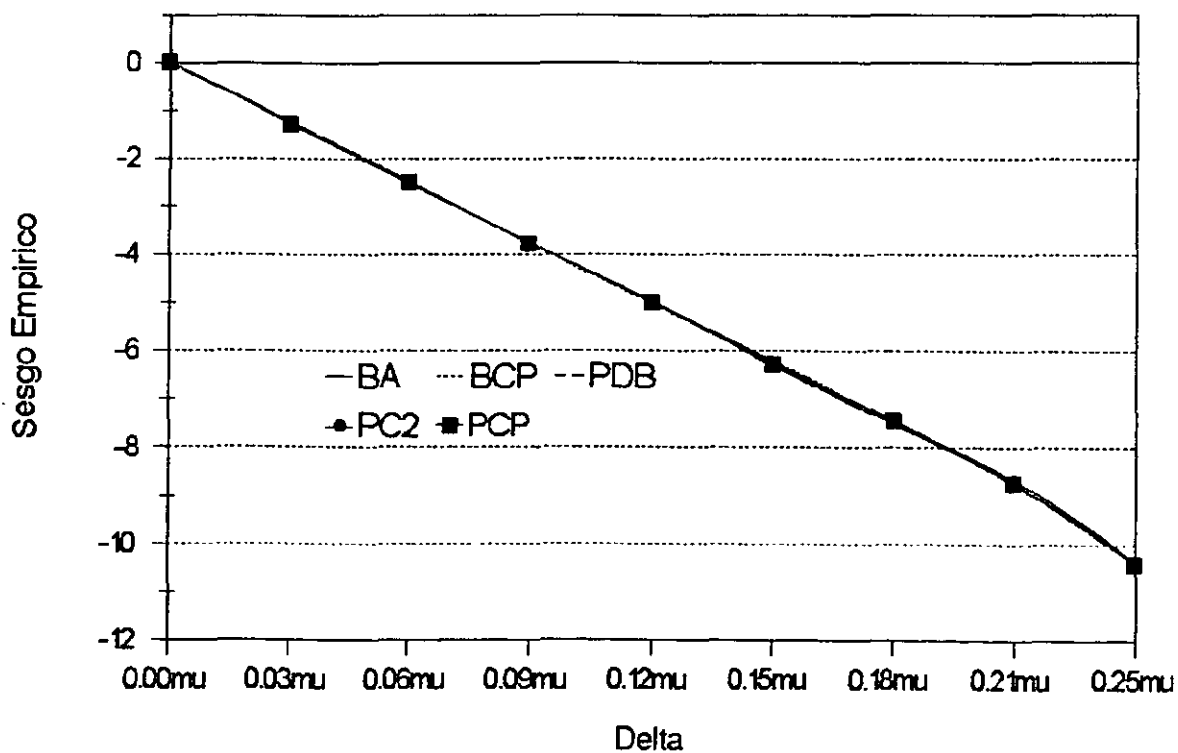


Figura 6.20. Gráfica de dispersión para la media del Sesgo Empírico en la interacción *Metodox* Δ para los datos de Wassom y Kalton (1953).

6.6. Error Cuadrático Medio Empírico de la Diferencia de Tratamientos

Para el análisis de la respuesta Error Cuadrático Medio Empírico para la diferencia del par de medias se utiliza como fuente de datos la combinación de Δ (9 niveles de incremento), *Diferencia del Par de Medias* (6 niveles de diferencias) y *Métodos de Análisis* (5 niveles de métodos) que dan lugar a un cuadro de triple entrada con 270 entradas (similar al Cuadro 6.8). Los resultados (Error Cuadrático Medio Empírico) de las simulaciones se pueden observar en el Anexo en CD de los Cuadros 9.6c-9.10c para datos de Baker *et al.* (1952), Cuadros 9.11c-9.15c para los datos de Immer (1932), Cuadros 9.16c-9.20c para los datos de Lander *et al.* (1938), Cuadros 9.21c-9.25c para los datos de Mercer y Hall (1911) y en los Cuadros 9.26c-9.30c para los datos de Wassom y Kalton (1953).

Con estos resultados se realiza un Análisis de Varianza para el modelo

$$Y_{ijk}^E = \mu + \text{Metodo}_i + \Delta_j + \text{Diferencia}E_k + (\text{Metodo} * \Delta)_{ij} + (\text{Metodo} * \text{Diferencia}E)_{ik} + (\Delta * \text{Diferencia}E)_{jk} + \varepsilon_{ijk} \quad \forall i = 1, \dots, 6, j = 1, \dots, 9 \text{ y } k = 1, \dots, 6 \quad (6.4)$$

donde

μ : Media General.

Y_{ijk}^E : Logaritmo del Error Cuadrático Medio Empírico en el *i-ésimo* método, *j-ésimo* incremento y la *k-ésima* diferencia de par de medias.

Metodo_i : Efecto del *i-ésimo* método.

Δ_j : Efecto del *j-ésimo* incremento.

$\text{Diferencia}E_k$: Efecto de la *k-ésima* diferencia de par de medias.

$(\text{Metodo} * \Delta)_{ij}$: Efecto de la *ij-ésima* combinación de los niveles de los efectos de método e incremento.

$(\text{Metodo} * \text{Diferencia}E)_{ik}$: Efecto de la *ik-ésima* combinación de los niveles de los efectos de método y diferencia de par de medias.

$(\Delta * \text{Diferencia}E)_{jk}$: Efecto de la *jk-ésima* combinación de los niveles de los efectos de incremento y diferencia de par de medias.

ε_{ijk} : *ijk-ésimo* error aleatorio del modelo.

El valor de la respuesta Y_{ijk}^E se obtiene del modo siguiente:

$$Y_{ijk}^E = \text{Log}(ECM_{Empirico})_{ijk} \quad \forall i = 1, \dots, 5, j = 1, \dots, 9 \quad y \quad k = 1, \dots, 6. \quad (6.5)$$

De los efectos del modelo (6.4) se investiga las interacciones $Metodo \times \Delta$ y $Metodo \times DiferenciaE$, por contener el efecto *Metodo*, que es uno de los más importantes para el estudio. Se encontró (Cuadro 6.12) alta significancia ($p \leq 0.01$) para el modelo en todos los conjuntos de datos considerados, también se observa en todos los conjuntos de datos un valor alto, alrededor de 99.90%, en el Coeficiente de Determinación (R^2). La significancia de la interacción $Metodo \times \Delta$ es altamente significativa ($p \leq 0.01$) sólo para los conjuntos de datos de Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton (1953); mientras que la interacción $Metodo \times DiferenciaE$ es altamente significativa ($p \leq 0.01$) sólo para el conjunto de datos de Lander *et al.* (1938).

Cuadro 6.12. Estadísticas del modelo (6.5) y significancia de dos interacciones para los conjuntos de datos.

Fuente de Datos	Modelo				Prob[F ≥ Fo] = p	
	CME	Fo	Prob[F ≥ Fo]	R ²	Met x Δ	Met x DifE
Baker et al. (1952)	5.60x10 ⁻⁵	51 033.08	0.0001**	0.9997	0.2963 ^{NS}	0.2509 ^{NS}
Immer (1932)	5.48x10 ⁻⁵	48 792.11	0.0001**	0.9997	0.1084 ^{NS}	0.4627 ^{NS}
Lander et al. (1938)	3.50x10 ⁻⁵	60 661.67	0.0001**	0.9997	0.0001**	0.0156**
Mercer y Hall (1911)	5.57x10 ⁻⁵	38 407.24	0.0001**	0.9996	0.0001**	0.4122 ^{NS}
Wassom y Kalton (1953)	9.32x10 ⁻⁵	15 676.50	0.0001**	0.9990	0.0001**	0.7269 ^{NS}

** : Indica una alta significancia del efecto respectivo para un Error Tipo I del 1%.

NS: No Significancia del efecto respectivo para un máximo del Error Tipo I del 5%.

En la fuente de datos de Baker *et al.* (1952), el análisis de varianza para la interacción $Metodo \times \Delta$ indica que la media del Logaritmo del Error Cuadrático Medio Empírico tiende a aumentar al aumentar Δ y el valor mínimo de la media es para la combinación $\Delta_1 \times PC2$ (0.00mu y Papadakis Clásico en 2 Dimensiones) con valor de 4.7628 y un máximo con un valor de 7.9809 que corresponde a la combinación de los niveles de Δ_9 (0.25mu) y PC2 (Papadakis Clásico en 2 Dimensiones) de los factores Δ y *Metodo*; el comportamiento de la media de la respuesta dentro de los diferentes métodos para cada uno de los niveles de Δ es similar en valor el valor numérico. Respecto a la interacción $DiferenciaE \times Metodo$ se encontró un comportamiento de similar a la interacción anteriormente descrita, el menor valor medio del Logaritmo del Error Cuadrático Medio

Empírico es de 6.1873 para los niveles de DE12 (Diferencia del par 1 y2 de medias) y PCP (Polinomio en Coordenadas de Posición) y el mayor valor de la media, 7.5539, se encontró en la combinación de DE17 y BCP (Bloques con Covariable de Posición).

El comportamiento, de la fuente de datos de Immer (1932), de la media del Logaritmo del Error Cuadrático Medio Empírico para la interacción *Metodo* x Δ tiene el comportamiento de aumentar la media a medida que aumenta Δ , dentro de los métodos la respuesta media tiene comportamiento numéricamente similar; el menor valor medio para la interacción es de 1.6819 que corresponde a los niveles Δ_1 x BCP y el máximo valor medio se encontró un una magnitud de 4.7534 que corresponde a los niveles de Δ_9 y BCP. En lo que respecta a la media de la interacción *DiferenciaE* x *Metodo* también tiende a aumentar a en la medida que aumenta la diferencia entre las medias (DE12 a DE17); así mismo, el valor de la media es muy similar dentro del factor *Metodo* para cada uno de los niveles de *DiferenciaE*, el valor mínimo medio encontrado en la interacción es de 2.9860 para los niveles DE12 y BCP y un máximo de 4.3436 para los niveles DE17 y BA (Bloques al Azar) de los factores involucrados.

La fuente datos de Mercer y Hall (1911) y Wassom y Kalton(1953) tienen un comportamiento similar, en la interacción *DiferenciaE* x *Metodo*, al aumentar el nivel de *DiferenciaE* aumenta la media del Logaritmo del Error Cuadrático Medio Empírico y se comportan aproximadamente constante para los diferentes métodos dentro de cada nivel de *DiferenciaE*; el mínimo valor de la media encontrado es de -0.1661 y 0.4265 para los niveles de DE12-PDB y DE12-PC2 para las dos fuentes respectivas de datos y el valor máximo encontrado es de 1.1524 y 1.6190 que corresponde con los niveles de DE17-PCP y DE17-BA respectivamente de las dos fuentes de datos mencionados anteriormente.

Para la interacción significativa, de los respectivos conjuntos de datos, se realiza una gráfica de dispersión de interacción con los niveles respectivos de *Metodo* y Δ para la media del Logaritmo del Error Cuadrático Medio Empírico, así mismo, se realiza una Prueba de Comparaciones múltiples de Medias, con un nivel de significancia del 5%, para los cinco Métodos por cada uno de los nueve incrementos de Δ . Del mismo modo se

realiza la gráfica de interacción de los niveles de *Metodo* x *DiferenciaE* y su respectiva prueba de Tukey para *Metodo* dentro de cada uno de los 6 niveles de *DiferenciaE*, en la interacción que es significativa.

La media del $\text{Log}(\text{ECM}_{\text{Empírico}})$ para la interacción *Metodo* x Δ , en los datos de Lander *et al.* (1938), tiene la tendencia de aumentar a medida que aumentan los niveles de Δ para los cinco Métodos investigados en magnitudes muy cercanas, por lo que sus respectivas líneas de dispersión se superponen (Fig. 6.21). Los menores valores medios de la media de $\text{Log}(\text{ECM}_{\text{Empírico}})$ se encontraron para $\Delta_1 = 0.00\mu$, con un valor cercano a 1.0 y un máximo de cercano a 3.5 para Δ_9 . En que respecta a la prueba de Tukey para los cinco Métodos dentro de su respectivo Δ sólo fue significativo para Δ_1 (Cuadro 6.13),

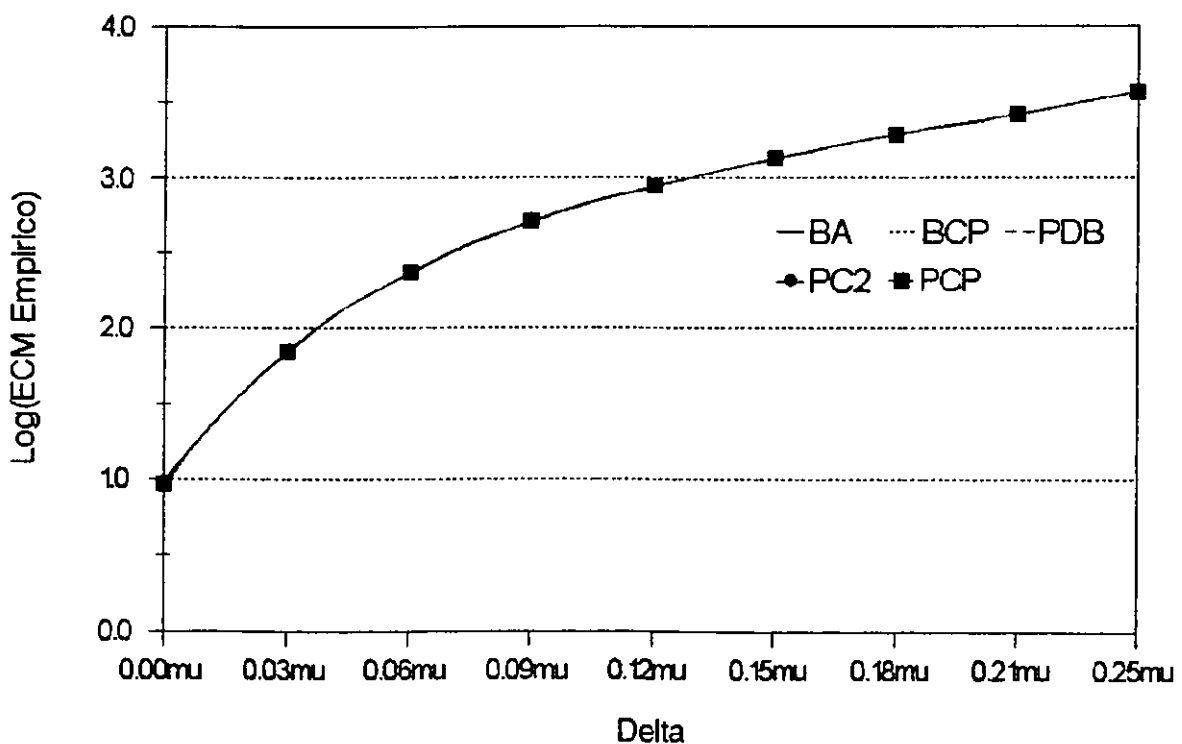


Figura 6.21. Gráfica de dispersión para la media del $\text{Log}(\text{ECM}_{\text{Empírico}})$ en la interacción *Metodo* x Δ para los datos de Lander *et al.* (1938).

(para los otros niveles de Δ no hay diferencia entre Métodos) que presenta los menores valores medios de $\text{Log}(\text{ECM}_{\text{Empírico}})$, en el grupo de medias *c* de Tukey para los 3 métodos del conjunto de Métodos de Refinamiento a los Bloques, es decir, los métodos de BA, BCP y PDB con un valor medio del logaritmo de 0.9553, 0.9422 y 0.9499 respectivamente. En

lo referente a la prueba de Tukey para métodos dentro de cada nivel de *DiferenciaE*, no se encontró significancia (solo un agrupamiento de métodos) en ningún nivel de esta fuente de variación, en el Cuadro 6.13 se presenta el tipo de agrupamiento (nivel DE12) que se encontró en todos los seis niveles (Fig. 22).

Los datos de Mercer y Hall (1911) y Wassom y Kalton (1953) tienen un comportamiento similar a los datos de Lander *et al.* (1938) anterior. El valor medio del $\text{Log}(\text{ECM}_{\text{Empírico}})$ tiende a aumentar a medida que aumenta Δ para los cinco métodos con valores cercanos por lo que la línea de dispersión se sobreponen (Fig. 6.23 y Fig. 6.24) la diferencia esta en el intervalo de la respuesta, para los datos de Mercer y Hall (1911) su valor mínimo es ligeramente inferior a de -1.0 (es el menor valor de $\text{Log}(\text{ECM}_{\text{Empírico}})$ en las interacciones significativas) y su valor máximo de 1.5, mientras que para los datos de Wassom y Kalton (1953) su menor valor de ligeramente inferior a 0.0 y su máximo valor es ligeramente inferior a 2.0.

Cuadro 6.13. Prueba de Tukey para el $\text{Log}(\text{ECM}_{\text{Empírico}})$ en $\Delta_1 = 0.00mu$ o *DiferenciaE=E12* para los cinco métodos en tres fuentes de datos.

Lander <i>et al.</i> (1938)				Mercer y Hall (1938)		Wassom y Kalton (1953)	
<i>Mx</i> Δ	$\text{Log}(\text{ECM}_E)^1$	<i>MxD</i> _E	$\text{Log}(\text{ECM}_E)^1$	<i>Mx</i> Δ	SE ¹	<i>Mx</i> Δ	$\text{Log}(\text{ECM}_E)^1$
PC2	0.9818a	PCP	1.9103a	BA	-1.0726a	BGP	-0.0802a
PCP	0.9685ab	BA	1.9048a	PCP	-1.1009ab	BA	-0.0848a
BA	0.9553 bc	PDB	1.9004a	PC2	-1.1195 b	PDB	-0.0876a
PDB	0.9499 bc	PC2	1.9003a	PDB	-1.1209 b	PC2	-0.0997a
BGP	0.9422 c	BGP	1.8964a	BGP	-1.1243 b	PCP	-0.1100a

¹Medias con la misma letra son estadísticamente iguales con un nivel de significancia del 5% en la Prueba de Comparaciones Múltiples de Medias de Tukey. *M*: Factor Metodo. *D_E*: Factor DiferenciaE.

En lo referente a la prueba de Tukey se encontró significancia, es decir, diferentes de grupos de métodos para el nivel Δ_1 en el conjunto de datos de Mercer y Hall (1911) con el menor valor medio de $\text{Log}(\text{ECM}_{\text{Empírico}})$ en los métodos de BGP, PDB, PC2 y PCP con valores de -1.1243, -1.1209, -1.1195 y -1.1009 respectivamente en el grupo b de Tukey (Cuadro 6.13), mientras que para los datos de Wassom y Kalton (1953) no se encontró significancia de *Metodo* en ninguno de los niveles de Δ ; en el Cuadro 6.13 se muestra como ejemplo el agrupamiento único observado para Δ_1 ; es similar para los demás Δ 's.

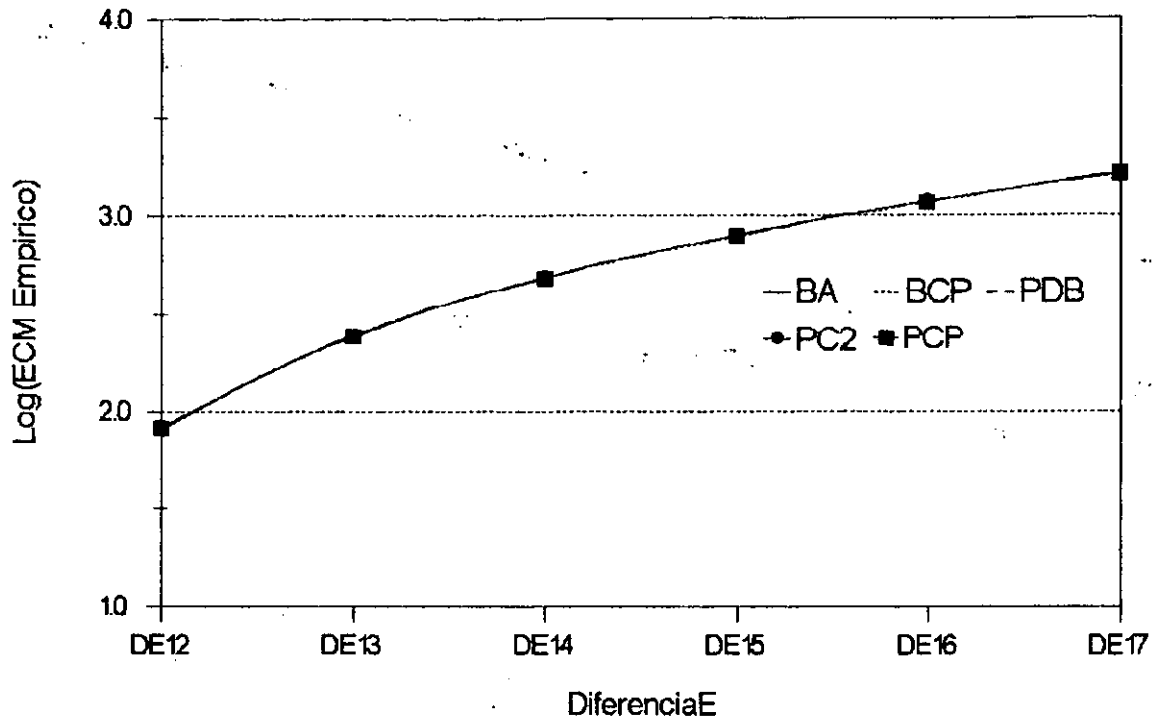


Figura 6.22. Gráfica de dispersión para la media del Log(ECM Empírico) en la interacción *Metodo x DiferenciaE* para los datos de Lander *et al.* (1938).

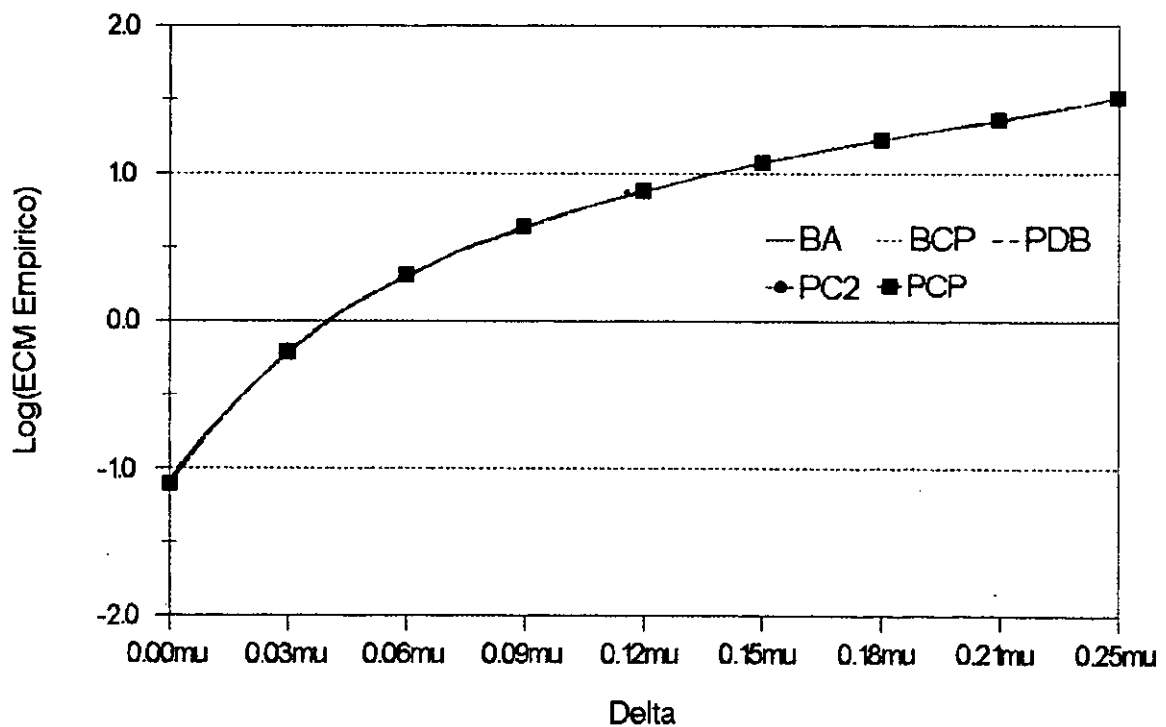


Figura 6.23. Gráfica de dispersión para la media del Log(ECM Empírico) en la interacción *Metodo x Δ* para los datos de Mercer y Hall (1911).

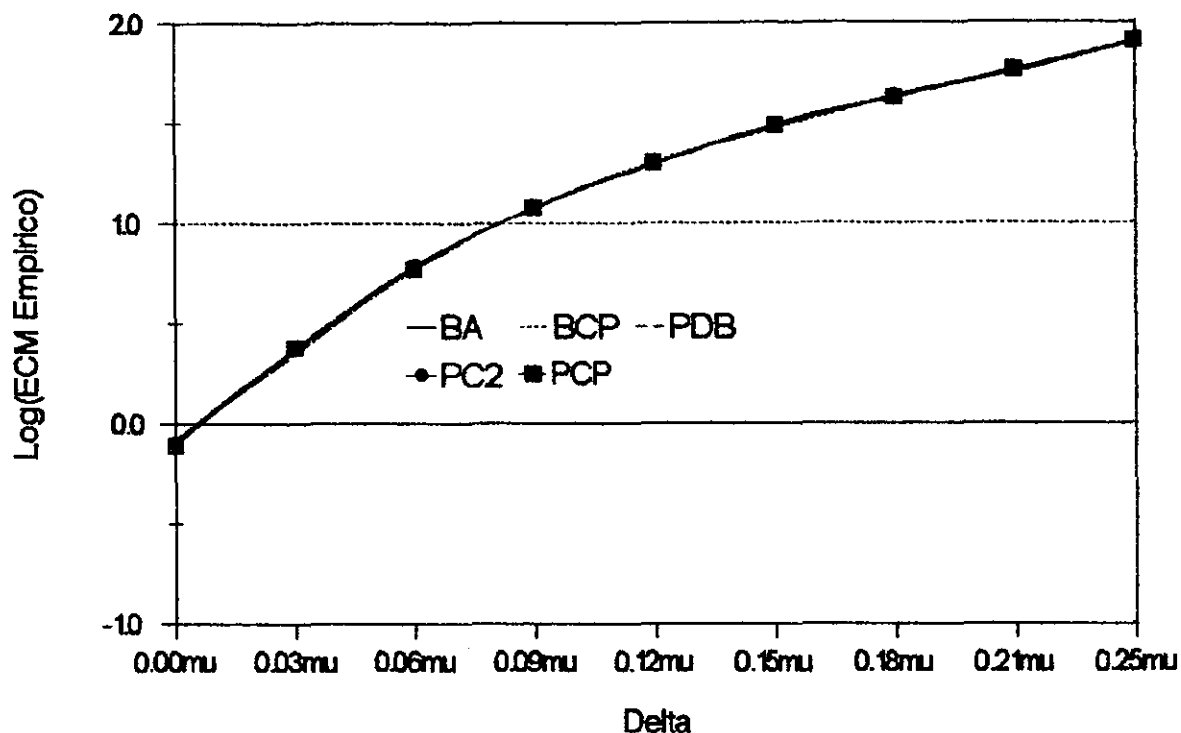


Figura 6.24. Gráfica de dispersión para la media del Log(ECM Empírico) en la interacción *Metodo x Δ* para los datos de Wassom y Kalton (1953).

Para identificar el tipo de comportamiento de los métodos investigados se les asigna un orden de rango (R) al factor *Metodo*, según los valores promedio de las variables de análisis evaluadas, las cuales son menor Y_{ij}^* (menor diferencia absoluta entre los logaritmos de las varianzas estimadas y observada), menor significancia estimada (\hat{p}) del modelo en el intervalo de Δ (de acuerdo a las Figuras 6.6-6.10), mayor promedio de PPS (Promedio de la Proporción de Significancia), menor Y_{ijk}^S y menor Y_{ijk}^E (error cuadrático medio empírico). Los resultados para cada conjunto de datos se observa en los Cuadros 6.14-6.18.

Con los resultados mencionados (Cuadros 6.14-6.18) se calcula la mediana del rango para cada tipo de respuesta por cada método sin considerar el conjunto de datos de donde proviene, se encontró (Cuadro 6.19) que el método PC2 (Papadakis Clásico en 2 Dimensiones) es el mejor cuando se evalúan mediante las respuestas de Y_{ij}^* y Y_{ijk}^S , mientras que el método BCP (Bloques con Coordenadas de Posición) resulto ser el mejor cuando se evalúa mediante \hat{p} , PPS y Y_{ijk}^E . Cada uno de ellos pertenece a uno de los dos conjuntos de

métodos investigados. Cabe mencionar que los rangos que se utilizan para el cálculo de la mediana sólo se indican para la variable de respuesta Y_{ij}^* del Cuadro 6.19.

Cuadro 6.14.-Resumen de resultados del rango por orden de magnitud de los métodos en orden de preferencia, de arriba hacia abajo, para los datos de Baker *et al.* (1952).

Tipo de Respuesta									
Y_{ij}^*	R	\hat{p}	R	PPS	R	Y_{ijk}^S	R	Y_{ijk}^E	R
(<)		(<)		(>)		(<)		(<)	
PCP	1.0	BCP	1.0	BCP	1.0	PCP	1.0	PC2	1.0
PC2	2.0	BA	2.5	PDB	2.0	BCP	2.0	PCP	2.0
PDB	3.0	PDB	2.5	BA	3.0	BA	3.0	PDB	3.0
BA	4.0	PC2	4.0	PC2	4.0	PDB	4.0	BA	4.0
BCP	5.0	PCP	5.0	PCP	5.0	PC2	5.0	BCP	5.0

Cuadro 6.15.-Resumen de resultados del rango por orden de magnitud de los métodos en orden de preferencia, de arriba hacia abajo, para los datos de Immer (1932).

Tipo de Respuesta									
Y_{ij}^*	R	\hat{p}	R	PPS	R	Y_{ijk}^S	R	Y_{ijk}^E	R
(<)		(<)		(>)		(<)		(<)	
PDB	1.0	BA	3.0	PCP	1.0	PC2	1.0	BCP	1.0
BA	2.0	BCP	3.0	PC2	2.0	BCP	2.0	PC2	2.0
PC2	3.0	PDB	3.0	BCP	3.0	PCP	3.0	PDB	3.0
BCP	4.0	PC2	3.0	PDB	4.0	BA	4.0	PCP	4.0
PCP	5.0	PCP	3.0	BA	5.0	PDB	5.0	BA	5.0

Cuadro 6.16.-Resumen de resultados del rango por orden de magnitud de los métodos en orden de preferencia, de arriba hacia abajo, para los datos de Lander *et al.* (1938).

Tipo de Respuesta									
Y_{ij}^*	R	\hat{p}	R	PPS	R	Y_{ijk}^S	R	Y_{ijk}^E	R
(<)		(<)		(>)		(<)		(<)	
PC2	1.0	BCP	1.0	BCP	1.0	PC2	1.0	BCP	1.0
PCP	2.0	PDB	2.0	PDB	2.0	PDB	2.0	PDB	2.0
BA	3.0	BA	3.0	BA	3.0	BCP	3.0	BA	3.0
PDB	4.0	PC2	4.5	PCP	4.0	PCP	4.0	PC2	4.0
BCP	5.0	PCP	4.5	PC2	5.0	BA	5.0	PCP	5.0

Cuadro 6.17.-Resumen de resultados del rango por orden de magnitud de los métodos en orden de preferencia, de arriba hacia abajo, para los datos de Mercer y Hall (1911).

Tipo de Respuesta									
Y_{ij}^*	R	\hat{p}	R	PPS	R	Y_{ijk}^S	R	Y_{ijk}^E	R
(<)		(<)		(>)		(<)		(<)	
PC2	1.0	BCP	1.0	BCP	1.0	PDB	1.0	PDB	1.0
PDB	2.0	PCP	2.0	PCP	2.0	BA	2.0	BCP	2.0
BA	3.0	BA	3.5	BA	3.0	BCP	3.0	PC2	3.0
PCP	4.0	PDB	3.5	PDB	4.0	PCP	4.0	PCP	4.0
BCP	5.0	PC2	5.0	PC2	5.0	PC2	5.0	BA	5.0

Cuadro 6.18.-Resumen de resultados del rango por orden de magnitud de los métodos en orden de preferencia, de arriba hacia abajo, para los datos de Wassom y Kalton (1953).

Tipo de Respuesta									
Y_{ij}^*	R	\hat{p}	R	PPS	R	Y_{ijk}^S	R	Y_{ijk}^E	R
(<)		(<)		(>)		(<)		(<)	
PC2	1.0	PCP	1.0	PCP	1.0	PC2	1.0	BCP	1.0
BA	2.0	BCP	2.0	BCP	2.0	BCP	2.0	PDB	2.0
PDB	3.0	BA	4.0	PDB	3.0	BA	3.0	PC2	3.0
BCP	4.0	PDB	4.0	BA	4.0	PDB	4.0	PCP	4.0
PCP	5.0	PC2	4.0	PC2	5.0	PCP	5.0	BA	5.0

Cuadro 6.19.- Evaluación general donde se reporta la Mediana del Rango en el ordenamientos de los métodos de menor a mayor.

Met.	Y_{ij}^*		\hat{p}		PPS		Y_{ijk}^S		Y_{ijk}^E	
	R	Med_R	Met	Med_R	Met	Med_R	Met	Med_R	Met	Med_R
PC2	2, 3, 1, 1, 1	1.0	BCP	1.0	BCP	1.0	PC2	1.0	BCP	1.0
PDB	3, 1, 4, 2, 3	3.0	BA	3.0	PCP	2.0	BCP	2.0	PDB	2.0
BA	4, 2, 3, 3, 2	3.0	PDB	3.0	PDB	3.0	BA	3.0	PC2	3.0
PCP	1, 5, 2, 4, 5	4.0	PCP	3.0	BA	3.0	PCP	4.0	PCP	4.0
BCP	4, 4, 5, 5, 4	5.0	PC2	4.0	PC2	5.0	PDB	4.0	BA	5.0

VII.

CONCLUSIONES

La varianza de los diferentes conjuntos de datos de muestra presenta magnitudes “pequeñas” y “grandes”. La más pequeña se encontró para los datos de Lander *et al.* (1938) y la mayor varianza esta en los datos de Baker *et al.* (1952). En terminos del Coeficiente de Variación, para los datos de Immer (1932) presentaron un 6.05%, los de Mercer y Hall (1911) un 11.60%, para Baker *et al.* (1952) 11.98%, Lander *et al.* (1938) un coeficiente de variación de 20.29% y para Wassom y Kalton (1953) un 29.56%.

La diferencia media entre los logaritmo de las varianza estimada y la varianza del segmento de datos (Y_{ij}^*), es significativamente influenciada por el *Metodo* de representar la variación geográfica y Δ la magnitud de los de tratamientos.

Los menores valores de (Y_{ij}^*) se encontraron en los métodos de PCP para los datos de Baker *et al.* (1952), PC2 para los datos de Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton (1953), dichos métodos pertenecen al conjunto de Métodos que Ignoran Bloques. Para la fuente de datos de Immer (1932) el mejor valor de (Y_{ij}^*) se encuentra en los métodos PDB y BA que son parte (2 de 3) del conjunto de Métodos de Refinamiento a los Bloques.

Para la significancia del modelo de cada uno de los cinco métodos investigados se encontró una separación clara de los dos conjuntos de métodos (Métodos de Refinamiento a los Bloques y Métodos que Ignoran Bloques) para los datos de Baker *et al.* (1938) y Lander *et al.* (1952); con menores valores de significancia para los métodos que utilizan bloques. Esto nos indica que se puede mejorar el análisis tradicional que sólo usa bloques. Para los datos de Immer (1932), Mercer y Hall (1911) y Wassom y Kalton (1953) no hay agrupación de métodos en sus respectivos conjuntos. En general se observa, como se

esperaba, que la significancia tiende a cero a medida que aumenta Δ para todos los métodos y conjuntos de datos.

La potencia empírica investigada mediante la diferencia de tratamientos al graficar el Promedio de la Proporción de Significancia (*PPS*) contra Δ encontró separación sólo para los conjuntos de métodos para los de datos de Baker *et al.* (1938), Immer (1932) y Lander *et al.* (1952). Los mejores (mayores) valores de *PPS* para los tres Métodos de Refinamiento a los Bloques en el conjunto de datos de Baker *et al.* (1938), el métodos de Bloques al Azar(BA) se ubicó en una posición intermedia; mientras que para los conjuntos de datos de Immer (1932) y Lander *et al.* (1952) presentaron mayores valores de *PPS* en los dos métodos que pertenecen al conjuntos de Métodos que ignoran Bloques (PC2 y PCP) el comportamiento del BA fue de peor a intermedio respectivamente.

El sesgo empírico evaluado en los modelos de análisis de la simulación mediante la interacciones de primer orden de los factores *Metodo*, Δ y *DiferenciaS* y considerando como factor más importante a *Metodo* encontró que la interacción *Metodo* \times Δ no fue significativa para ninguna fuente de datos mientras que la interacción *Metodo* \times *DiferenciaS* fue significativa para todos los conjuntos de datos investigados. En dicha interacción significativa el sesgo empírico tiende a disminuir a medida que aumenta Δ para todos los conjuntos de datos investigados. La diferencia entre métodos sólo se encontró para cuando Δ , en las cinco fuentes de datos; donde el menor sesgo empírico que se encontró fue para método PCP para la fuente de datos de Baker *et al.* (1952), Lander *et al.* (1938) y Mercer y Hall (1911), el método PDB para la fuente de datos Immer (1932) y los métodos de BA, BCP, PC2 y PCP para la fuente de datos de Wassom y Kalton (1953).

Por otra parte, el Error Cuadrático Medio Empírico evaluado por las interacciones de primer orden de, *Metodo*, Δ y *DiferenciaE* y donde estuviera involucrado el factor *Metodo*, es decir, *Metodo* \times Δ y *Metodo* \times *DiferenciaE*. Se encontró que el modelo fue significativo para las cinco fuentes de datos analizados; respecto a la significancia de los factores *Metodo* \times Δ , ésta fue significativa para la fuente de datos de Lander *et al.* (1938), Mercer y Hall (1911) y Wassom y Kalton(1953), mientras que la interacción

MetodoxDiferenciaE sólo fue significativo para la fuente de datos de Lander *et al.* (1938). En las interacciones significativas se observa la tendencia de aumentar el logaritmo de Error Cuadrático Medio Empírico a medida que aumentan los niveles de Δ o *DiferenciaE* para los cinco Métodos investigados en magnitudes muy cercanas, por lo que sus respectivas líneas de dispersión se sobreponen. La diferencia entre métodos sólo se encontró para Δ_1 en la fuente de datos de Lander *et al.* (1938) y Mercer y Hall (1911).

Cuando se calcula la mediana del rango para cada tipo de respuesta por cada método, se encontró que el método PC2 (Papadakis Clásico en 2 Dimensiones) es el mejor cuando se evalúan mediante las respuestas de la menor diferencia absoluta entre los logaritmos de las varianzas estimadas y observada (Y_{ij}^*) y sesgo empírico (Y_{ijk}^S), mientras que el método BCP (Bloques con Coordenadas de Posición) resultó ser el mejor cuando se evalúa mediante menor significancia estimada del modelo en el intervalo de Δ (\hat{p}), Promedio de la Proporción de Significancia (*PPS*) y error cuadrático medio empírico (Y_{ijk}^E). Por lo anterior recomendable utilizar algún método de análisis alternativo a Bloques al Azar, ya que éste fue en general de intermedio a peor comportamiento.

VIII.

BIBLIOGRAFÍA

- Anselin, L. and Griffith, D. A. 1988. Do spatial effects really matter in regression analysis?. *Papers of the Regional Science Association*, **65**:11-34.
- Atkinson, A. C. 1969. The use of residuals as a concomitant variable. *Biometrika*, **56**:33-41.
- Azaïs, J-M. and Druilhet, P. 1997. Optimality of neighbour balanced designs when neighbour effects are neglected. *J. Statist. Planning Infer.*, **64**:353-367.
- Azaïs, J-M., Monod, H. and Bailey, R. A. 1998. The influence of design on validity and efficiency neighbour methods. *Biometrics*, **54**:1374-1387.
- Baird, D. and Mead, R. 1991. The empirical efficiency and validity of two neighbour models. *Biometrics*, **47**:1473-1487.
- Ball, S. T., Mulla, D. J. and Konzak, C. F. 1993. Spatial heterogeneity affects variety trial interpretation. *Crop Science*, **33**:931-935.
- Bartlett, M. S. 1974. The statistical analysis of spatial pattern. *Advances in Applied Probability*, **6**:336-358.
- Bartlett, M. S. 1938. The proximate recovery of information from replicated field experiments with large blocks. *J. Agr. Sci.*, **28**:418-427.
- Bartlett, M. S. 1978. Nearest neighbour models in the analysis of field experiments. *J. R. Statist. Soc. B*, **40**:147-174.

- Basu, S. and Reinsel, G. C. 1994. Regression models with spatially correlated errors. *Journal of the American Statistical Association*, **89**:88-99.
- Batchelor, L. D. and Reed, H. S. 1918. Relation of the variability of yields of fruit trees to the accuracy of field trials. *Journal of Agricultural Research*, **12**:245-283.
- Baker, G. A., Huberty, M. R. and Veihmeyer, F. J. 1952. A uniformity trial on unirrigated barley of ten years' duration. *Agronomy Journal*, **44**:267-270.
- Baker, G. A. and Baker, R. E. 1953. Strawberry uniformity field trials. *Biometrics*, **9**:412-421.
- Baker, G. A. And Roessler, E. B. 1957. Implications of a uniformity trial with small plots of wheat. *Hilgardia*, **26**:183-188.
- Besag, J. and Kempton, R. 1986. Statistical analysis of field experiments using neighbouring plots. *Biometrics*, **42**:231-251.
- Casler, M. D. 1999. Spatial variation affects of perennial cool-season forage grass trials. *Agronomy Journal*, **91**:75-81.
- Clark, I. 1979. *Practical Geostatistical*. Applied Science Publishers, Essex, England.
- Clarke, F. R. and Baker, R. J. 1996. Spatial analysis improves precision of seed lot comparisons. *Crop Science*, **36**:1180-1184.
- Clarke, F. R., Baker, R. J. and DePauw, R. M. 1994. Moving mean and least squares for analysis of grain yield data. *Crop Science*, **34**:1479-1483.
- Cliff, A. D. and Ord, J. K. 1981. *Spatial Processes: Models and Applications*. Pion, London.
- Cochran, W. G. 1936. The statistical analysis of the distribution of field counts of diseased plants. *J. of the R. Stat. Soc, Supplement*, **3**:49-67.

- Cox, D. R. and Isham, V. 1980. *Point Processes*. Chapman and Hall, London.
- Cox, D. R. and Lewis, P. A. W. 1966. *The Statistical Analysis of Series of Events*. Methuen, London.
- Cressie, N and Laslett, G. M. 1987. Random set theory and problem of modeling. *SIAM Review*, **29**:557-574.
- Cressie, N. 1987. A nonparametric view of generalized covariances for kriging. *Mathematical Geology*, **19**:425-449.
- Cressie, N. 1991. Modeling growth with random sets. In *Spatial Statistics and Imaging* (Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference), A. Possolo, ed. Institute of Mathematical Statistics, Hayward, CA.
- Cressie, N. A. C. 1991. *Statistical for Spatial Data*. Wiley, New York.
- Cressie, N. and Chan, N. H. 1989. Spatial modeling of regional variables. *Journal of the American Statistical Association*, **84**:393-401.
- Cressie, N. and Horton, R. 1987. A robust/resistant spatial analysis of soil-water infiltration. *Water Resources Research*, **23**:911-917.
- Cressie, N. and Hawkins, D. M. 1980. Robust estimation of the variogram, I. *Journal of the International Association for Mathematical Geology*, **12**:115-125.
- Cullis, B. R. and Gleeson, A. C. 1991. Spatial analysis of field experiments-an extension to two dimensions. *Biometrics*, **47**:1449-1460.
- Cullis, B., Gogel, B., Verbyla, A., and Thompson, R. 1998. Spatial analysis of multi-environment early generation variety trials. *Biometrics*, **54**:1-48.
- Daley, D. J. and Vere-Jones, D. 1972. A summary of the theory of point processes. In *Stochastic Point Processes: Statistical Analysis, Theory, and Applications*, P. A. W.

- Lewis, ed. Wiley, New York, 299-383.
- Daley, D. J. and Vere-Jones, D. 1988. *Introduction to the Theory of Point Processes*. Springer, New York.
- David, M. 1977. *Geostatistical Ore Reserve Estimation*. Elsevier, Amsterdam.
- David, M. 1988. *Handbook of Applied Advanced Geostatistical Ore Reserve Estimation*. Elsevier, Amsterdam.
- De Vos, A. F. 1991. On randomization, modeling and experimental design; a new example in an old discussion. *Statistica Neerlandica*, **45**:365-377.
- Dier, B. W., Voss, B. K. and Fehr, W. R. 1991. Moving-Mean analysis of field tests for iron efficiency of soybean. *Crop Science*, **31**:54-56.
- Diggle, P. J. 1981. Binary mosaics and the spatial pattern of heather. *Biometrics*, **37**:531-539.
- Diggle, P. J. 1983. *Statistical Analysis of Spatial Point Patterns*. Academic Press, New York.
- Diggle, P. J., Tawn, J. A. and Moyeed, R. A. 1998. Model-based geostatistics. *Appl. Statist.*, **47**:299-350.
- Dow, M. M., Burton, M. L. and White, D. R. 1982. Network autocorrelation: A Simulation study of a foundational problem in regression and survey research. *Social Networks*, **4**:169-200.
- Draper, N. R. and Faraggi, D. 1985. Role of the Papadakis estimator in one- and two-dimensional field trials. *Biometrika*, **72**:223-226.
- Federer, W. T. and Basford, K. E. 1991. Competing effects designs and models for two-dimensional field arrangements. *Biometrics*, **47**:1461-1472.

- Fisher, R. A. 1935. *The Design of Experiments*. Oliver and Boyd, Edinburgh.
- Fisher, R. A. 1970. *Statistical Methodos for Research Workers*, 14th. Edition. Oliver and Boyd. Edinburg.
- Gandin, L. S. 1963. *Objetive Analysis of Meteorological Fields*. Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad (translated by Israel Program for Scientific Translations, Jerusalem, 1965).
- Getis, A. and Boots, B. 1978. *Modeles of Spatial Processes*. Cambridge University Press, Cambridge.
- Gilmour, A. R., Cullis, B. R. and Verbyla, A. P. 1997. Accounting for natural and extraneous variation in the analysis of field experiments. *J. Agric. Biol. Environm. Statist.*, **2**:269-293
- Gilmour, A. R., Thompson, R. and Cullis, B. R. 1995. Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**:1440-1450.
- Gill, P. S. and Shukla, G. K. 1985. Efficiency of nearest neighbour balanced block designs for correlated observations. *Biometrika*, **72**:539-544.
- Glesson, A. C. and Cullis, B. R. 1987. Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. *Biometrics*, **43**:277-287.
- Gomez, M. and Hazen, K. 1970. Evaluating sulfur and ash distribution in coal seams by statistical response surface regression analysis. U. S. Bureau of Mines Report RI 7377.
- Gotway, C. A. and Stroup, W. W. 1997. A generalized linear model approach to spatial data analysis and prediction. *J. Agric. Biol. Environm. Statist.*, **2**:157-178.

- Green, P. J. 1985. Linear models for field trials, smoothing and cross-validation. *Biometrika*, **72**:527-537.
- Green, P., Jenninson, C. and Seheult, A. 1985. Analysis of field experiments by least squares smoothing. *J. Royal Statist. Soc. B.*, **47**:299-315.
- Grondona, M. O. 1991. Using spatial considerations in the analysis of experiments. *Technometrics*, **33**:150-165.
- Grondona, M. O. and Cressie, N. 1991. Using spatial considerations in the analysis of experiments. *Technometrics*, **33**:381-392.
- Grondona, M. O. and Cressie, N. 1992. Efficiency of block designs under stationary second-order autoregressive error. *Sankhya A*, **54**:59-75.
- Grondona, M. O., Crossa, J., Fox, P. N. and Pfeiffer, W. H. 1996. Analysis of variety yield trials using two-dimensional separable ARIMA processes. *Biometrics*, **52**:763-770 .
- Halley, E. 1686. An historical account of the trade winds, and monsoons, observable in the seas between and near the tropicks; with and attempt to assign the physical cause of said winds. *Philosophical Transactions*, **183**:153-168.
- Hart, J. F. 1954. Central tendency in areal distributions. *Economic Geography*, **30**:4-59.
- Hawkins, D. M. and Cressie, N. 1984. Robust kriging-a proposal. *Journal of the International Association for Mathematical Geology*, **16**:3-18.
- Helms, T. C., Scott, R. A. and Hammond, J. J. 1999. Intrablock variance among duplicate treatment for nearest-neighbor analyses. *Agronomy Journal*, **91**:317-320.
- Immer, F. R. 1932. Size and shape of plot in relation to field experiments with sugar beets. *Journal of Agricultural Research*, **44**:649-668.
- Istok, J. D. and Cooper, R. M. 1988. Geostatistical applied to groundwater pollution. III:

- Global estimates. *Journal of Environmental Engineering*, **114**:915-928.
- Jansen, M. J. W. 1993. On randomization, modeling and experimental design: some remarks on paper by De Vos. *Statistica Neerlandica*, **47**:285-290.
- Journel, A. G. and Huijbregts, C. J. 1978. *Mining Geostatistics*. Academic Press, London.
- Kempton, R. A., Seraphin, J. C. y Sword, A. M. 1994. Statistical analysis of two-dimensional variation in variety yield trials. *Journal of Agricultural Science (Cambridge)*, **122**:335-342.
- Kolgomorov, A. N. 1941. Interpolation and extrapolation of stationary random sequenses. *Izvestiia Akademii Nauk SSSR, Serii Matematicheskii*, **5**:3-14.
- Krige, D. C. 1951. A statistical approach to some basic mine valuation problems on the Witwatersand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**:119-139.
- Lander, P. E., Narain, R. and Singh, A. 1938. Soil uniformity trials in the Punjab, I. *Ind. J. Agr. Sc.* **8**:271-307.
- Laslett, G. M. 1994. Kriging and splines: An empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association*, **89**:391-409.
- Lefkovitch, L. P. 1992. Fertility adjustments in rectangular field trials. *Biometrical Journal*, **34**:1-15.
- Lewis, P. A. W., ed. 1972. *Stochastic Point Processes: Statistical Analysis, Theory, and Applications*. Wiley, New York.
- Lill, W. J., Glesson, A. C. and Cullis, B. R. 1988. Relative accuracy of a neighbour method for field trials. *Journal of Agricultural Science (Cambridge)*, **111**:339-346.

- Marquiss, M., Newton, I., and Radcliffe, D. A. 1978. The decline of the raven, *Corvus corax*, in relation to afforestation in southern Scotland and northern England. *Journal of Applied Ecology*, **15**:129-144.
- Martin, R. J. 1982. Some aspects of experimental design and analysis when errors are correlated. *Biometrika*, **69**:597-612.
- Martin, R. J. 1990. The use of time-series models and methods in the analysis of agricultural field trials. *Communications in Statistics, Part A-Theory and Methods*, **19**:55-81.
- Martin, R. J. 1995. Some results on randomization when observations are spatially dependent and the analysis uses generalized least-squares. *Communications in Statistics, Part A-Theory and Methods*, **24**:2199-2210.
- Martin, R. J. 1997. A three-dimensional unilateral autoregressive lattice process. *J. Statist. Planning Infer.*, **59**:1-18.
- Marx, D. B. y Stroup, W. W. 1991. Accounting for spatial variability in designed experiments. *Trends in Soil Sc.*, **1**:357-372.
- Marx, D. y Thompson, K. 1987. Practical aspects of agricultural kriging. Bolletín 903. Division of Agriculture. *Arkansas Agric. Exp. Station*. University of Arkansas.
- Matern, B. 1960. *Spatial Variation. Lectures Notes in Statistics, No. 26*. Springer, New York.
- Matheron, G. 1962. *Traite de Geostatistique Appliquee, Tome I. Memories du Bureau de Recherches Geologiques et Minieres, No. 14*. Editions Technip, Paris.
- Matheron, G. 1963a. *Traite de Geostatistique Appliquee, Tome II. Le Krigeage. Memories du Bureau de Recherche Geologiques et Minieres, No. 24*. Editions Bureau de Recherche Geologiques et Minieres, Paris.
- Matheron, G. 1963b. Principles of geostatistics. *Economic Geology*, **58**:1246-1266.

- Matheron, G. 1971. The Theory of Regionalized Variables and Its Applications. Cahiers du Centre de Morphologie Mathématique, No. 5. Fontainebleau, France.
- Mead, R. 1967. A mathematical model for the estimation of interplant competition. *Biometrics*, 23:189-205.
- Méndez R., I. 1970. Study of uniformity trials and six proposals as alternative to blocking for design and analysis of field experiments. Thesis Ph. D. Thesis North Carolina State University.
- Méndez R., I. 1971a. Determinación de funciones de tendencia y su comparación con bloques en los ensayos de uniformidad. *Agrociencia Serie A*, 6:17-28.
- Méndez R., I. 1971b. Refinamiento a la técnica de selección masal moderna. *Agrociencia Serie A*, 6:87-91.
- Méndez R., I. 1972. Estudio de seis métodos alternativos para el uso de bloques en la experimentación de campo. *Agrociencia Serie A*, 16:3-16.
- Méndez R., I. 1992. Métodos para representar la variación geográfica en un ensayo de campo, un estudio de caso. Presentado en el encuentro México-Cuba de estadística, La Habana, Cuba. *Sin publicar*.
- Méndez R., I. y Rivera H., G. 1976. Métodos alternativos para comparación de un número grande de variedades, usando funciones de tendencia. *Agrociencia Serie A*, 26:153-165.
- Mercer, W. B. and Hall, A. D. 1911. The experimental error of field trials. *Journal of Agricultural Science (Cambridge)*, 4:107-132.
- Modjeska, J. S. and Rawlings, J. O. 1983. Spatial correlation analysis of uniformity data. *Biometrics*, 39:373-384.

- Montgomery, D. C. and Peck, E. A. 1982. Introduction to linear regression analysis. John Wiley & Sons. New York. 504 p.
- Ord, J. K. and Rees, M. 1979. Spatial Processes: Recent developments with applications to hydrology. In *The Mathematics of Hydrology and Water Resources*, E. H. Lloyd, T. O'Donnell, and J. C. Wilkinson, eds. Academic Press, London. 95-118.
- Papadakis, J. S. 1937. Méthode statistique pour des expériences sur champ. Bulletin de l'Institut d'Amélioration des Plantes a Thessalonike(Grece). Bulletin Scientifique, No. 23.
- Papadakis, J. S. 1984. Advances in the analysis of field experiments. *Proceedings of the Academy of Athens*, 59:362-342.
- Pearse, S. C. and More, C. S. 1976. Reduction of experimental error in perennial crops, using adjustment by neighbouring plot. *Expl. Agric*, 12:267-272.
- Pielou, E. C. 1959. The use of point-to-plant distances in the pattern of plant populations. *Journal of Ecology*, 47:607-613.
- Pielou, E. C. 1977. *Mathematical Ecology*. Wiley, New York.
- Ripley, B. D. 1981. *Spatial Statistics*. Wiley, New York.
- Serra, J. 1980. The Boolean model and random sets. *Computer Graphics and image Processing*, 12:99-126.
- Smith, J. A. and Karr, A. F. 1985. Parameter estimation for a model of space-time rainfall. *Water Resources Research*, 21:1251-1257.
- Stroup, W. W. and Muiltze, D. K. 1991. Nearest neighbour adjusted best linear unbiased prediction. *The American Statistician*, 45:194-200.
- Stroup, W. W., Baenzinger, S. P. and Muiltze, D. K. 1994. Removing spatial variation from

- wheat yield trials: A comparison of methods. *Crop Science*, **86**:62-66.
- Student 1907. On the error of counting with a haemocytometer. *Biometrika*, **5**:351-360.
- Symons, M. J., Grimson, R. C., and Yuan, Y. C. 1983. Clustering of rate events. *Biometrics*, **39**:193-205.
- Taam, W. and Hamada, M. 1993. Detecting spatial effects from factorial experiments: An application from integrated-circuit manufacturing. *Technometrics*, **35**:149-160.
- Tabor, J. A., Warrick, A. W., Myers, D. E. and Pennington, D. E. 1985. Spatial variability of nitrate in irrigated cotton: II. Soil nitrate and correlated variables. *The Soil Sc. Soc. of Am. J.*, **49**:390-394.
- Thiebaux, H. J. and Pedder, M. A. 1987. *Spatial Objective Analysis with Applications in Atmospheric Science*. Academic Press, London.
- Trujillo U., M. P. 1994. Comparación de métodos para modelar la variabilidad espacial en ensayos de campo. Tesis Maestría en Ciencias. Colegio de Postgraduados.
- Trujillo U., M. P., Méndez R., I. y Castillo M., A. 1996. Comparación de diseños experimentales para modelar variabilidad espacial en ensayos de campo. *Agrociencia*, **30**:533-540.
- Upton, G. J. G. and Fingleton, B. 1985. *Spatial Data Analysis by Example, Volume 1: Point Pattern and Quantitative Data*. Wiley, Chichester.
- Vieira, S. R., Hatfield, J. L., Nielsen, D. R. and Biggar, J. W. 1982. Geostatistical theory and application to variabilidad of some agronomical properties. *Hilgardia*, **51**:1-75.
- Warren, J. A. and Méndez R., I. 1981. Block size and orientation, and allowance for positional effects, in field experiments. *Expl. Agric.* **17**:17-24.
- Warren, J. A. and Méndez, I. 1982. Methods for estimating background variation in field experiments. *Agronomy Journal*, **74**:1004-1009.

- Wassom, C. E. and Kalton, R. R. 1953. Estimation of optimum plot size using data from bromegrass uniformity trials. *Iowa State College Res.*, **Bull.** 396.
- Watson, G. S. 1972. Trend surface analysis and spatial correlation. *Geological Society of America, Special Paper*, 146:39-46.
- Whitten, E. H. T. 1970. Orthogonal polynomial trend surfaces for irregularly spaced data. *Journal of the International Association for Mathematical Geology*, 2:141-152.
- Wiebe, G. A. 1935. Variation and correlation in grain yield among 1,500 wheat nursery plots. *J. Agric. Res.* 50:331-357.
- Wiener, N. 1949. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, Cambridge, MA.
- Wiens, J. A., Rotenberry, J. T., and Van Horne, B. 1986. A lesson in the limitations of field experiments: Shrubsteppe birds and habitat alteration. *Ecology*, 67:365-376.
- Wilkinson, G. N., Ecrert, S. R., Hancock, T. W. and Mayo, O. 1983. Nearest neighbour (NN) analysis of field experiments. *J. R. Statist. Soc. B*, 45:151-211.
- Williams, E. R. 1985. A criterion for the construction of optimal neighbour designs. *J. Royal Statist. Soc. B.*, 47:489-497.
- Williams, E. R. 1986. A neighbour model for field experiments. *Biometrika*, 73:279-287.
- Wold, H. 1938. *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksells, Uppsala.
- Wu, T., Mather, D. E. and Dutilleul, P. 1998. Application of geostatistical and neighbor analyses to data from plant breeding trials. *Crop Science*, 38:1545-1553.
- Yates, F. 1936. A new method of arranging variety trials involving a large number of varieties. *J. Agr. Sci.*, 26:424-455.

- Zárate de L., G. y Méndez R., I. 1972. Representación de variables aleatorias asociadas a una localidad geográfica, aplicación de la metodología en la representación de las sales en los suelos. *Agrociencia Serie A*, 10:129-142.
- Zimmerman, D. L. and Harville, D. A. 1989. On the unbiasedness of the Papadakis estimator and other nonlinear estimators of treatment contrasts in field field-plot experiments. *Biometrika*, 76:253-259.
- Zimmerman, D. L. and Harville, D. A. 1991. A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, 47:223-239.