



ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística Informática

**“La Econometría Espacial en la Agricultura de Precisión,
Aplicación en un ingenio azucarero”**

TESIS DE GRADO

Previa a la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentada por:

Xavier Alberto Vítores Sánchez

GUAYAQUIL – ECUADOR

AÑO 2005

AGRADECIMIENTO

Agradezco a Dios por darme fuerza y ser mi guía en todo momento. A mis padres y hermanos, quienes siempre me apoyan incondicionalmente. A los maestros que supieron incentivar me no sólo con su vasto conocimiento sino también con su calidad humana y ejemplo. A mis amigos, quienes depositaron su confianza en mi y me alentaron en todo momento. A todas las personas que de una u otra forma colaboraron en la realización de este trabajo, especialmente, al Mat. Fernando Sandoya, Mat. César Guerrero, Ec. Manuel González, Ing. Leonidas Estrada, Ing. Oscar Núñez y Phd. Egbert Spaans.

DEDICATORIA

A Dios

A mi familia

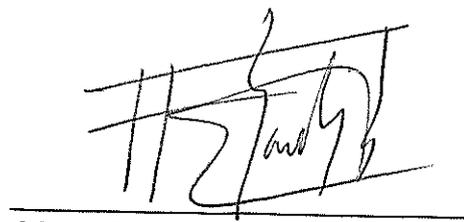
A mis maestros

A mis amigos

TRIBUNAL DE GRADUACIÓN



MAT. JORGE MEDINA
PRESIDENTE



MAT. FERNANDO SANDOZA
DIRECTOR DE TESIS

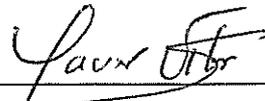


MAT. CÉSAR GUERRERO
VOCAL

DECLARACIÓN EXPRESA

“La responsabilidad del contenido de esta Tesis de Grado, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL”

(Reglamento de graduación de la ESPOL)



Xavier Alberto Vitores Sánchez

INTRODUCCIÓN

La Agricultura de Precisión o AP implica una filosofía de manejo sitio-específico del cultivo mediante la utilización de tecnologías de información geográfica que permiten al productor adecuar el manejo de suelos y cultivos a la variabilidad presente dentro de las unidades de manejo agrícola; y, consecuentemente, tener el potencial para: 1) obtener mayores rendimientos con el mismo nivel de insumos, simplemente redistribuidos; 2) obtener los mismos rendimientos con menores insumos; o, 3) obtener mayores rendimientos y reducción de insumos.

Sin embargo, la adopción de la AP conlleva algunos problemas que deben ser resueltos. Específicamente, existe una brecha de información entre el análisis de datos con referencia espacial y las recomendaciones sitio-específicas del cultivo que implican un manejo diferencial de insumos y prácticas agrícolas de acuerdo a las verdaderas necesidades del cultivo, un mayor conocimiento del sistema natural del cultivo y, consecuentemente, la maximización de la rentabilidad del negocio y la minimización de un posible impacto ambiental negativo.

La principal causa de la brecha de información es la utilización de técnicas estadísticas tradicionales en el análisis de datos espaciales, los cuales, poseen características muy diferentes a los datos con una referencia temporal. Los aspectos más sobresalientes a tratarse cuando se investiga sobre las características de las observaciones en el espacio son la dependencia espacial y la heterogeneidad espacial.

El presente trabajo trata sobre la problemática de adoptar los principios de la AP en un cultivo de caña de azúcar en el Ingenio San Carlos (ISC) durante el ciclo 2003, específicamente, en lo referente al análisis estadístico de la información recopilada por unidades de manejo agrícola de precisión sobre la producción y sus factores determinantes: características edáficas, climáticas y fisiológicas del cultivo; para luego, establecer recomendaciones de manejo agrícola.

Debido a que las técnicas tradicionales no son las más adecuadas en el contexto espacial, se plantea la hipótesis de que un modelo estadístico que incluya de manera explícita componentes espaciales en su especificación permitirá obtener estimaciones más robustas de la respuesta sitio-específica del cultivo a los factores de producción en comparación a las de un modelo tradicional. La importancia de utilizar como herramientas de decisión modelos

estadísticos que permitan realizar inferencias válidas se traduce en recomendaciones más acertadas.

El objetivo general de esta investigación es disminuir la brecha de información antes mencionada mediante técnicas estadísticas que permitan incorporar los efectos espaciales de los datos en un modelo estocástico que represente una relación causa-efecto entre la producción de un cultivo de caña de azúcar en el ISC y sus factores determinantes durante el ciclo 2003, con la finalidad de obtener un diagnóstico de la situación del cultivo en el periodo bajo estudio y establecer recomendaciones de manejo agrícola.

Los objetivos específicos son:

1. Estimar dos funciones de la respuesta sitio-específica del cultivo de caña de azúcar a sus factores de producción: la primera corresponde a un modelo econométrico espacial estimado mediante el método de Máxima Verosimilitud, y la segunda corresponde a un modelo econométrico tradicional estimado con el método de Mínimos Cuadrados Ordinarios.

2. Determinar cual es la “mejor” especificación y correspondiente método de estimación mediante pruebas de diagnóstico y medidas de ajuste para obtener estimaciones robustas de la respuesta sitio-específica del cultivo a sus factores determinantes.
3. Establecer recomendaciones de manejo agrícola a partir de un diagnóstico de la situación del cultivo sustentado por a especificación en el objetivo 3.
4. Contrastar supuestos de la teoría agronómica para tener un mejor entendimiento del sistema natural bajo estudio.

Desde el punto de vista del análisis estadístico, se busca encontrar los mejores estimadores para los parámetros del modelo econométrico de la producción de caña de azúcar. Dentro del contexto agronómico, la importancia de la investigación radica en establecer una metodología que permita realizar inferencias sobre el sistema productivo utilizando información empírica y de mejor calidad contenida en los datos de AP.

RESUMEN

En el presente trabajo se trata de resolver uno de los problemas fundamentales que conlleva la adopción de la Agricultura de Precisión (AP): la brecha de información existente entre el análisis de la información con referencia espacial y las recomendaciones sitio-específicas. Para cumplir con este propósito se utilizan técnicas estadísticas que tratan específicamente con los efectos espaciales de los datos georreferenciados reconocidos como dependencia y heterogeneidad espacial.

En el primer capítulo se expone el significado de un manejo sitio-específico del cultivo mediante la utilización de modernas herramientas de información geográfica, los procedimientos y problemas principales dentro de la adopción de la AP, resaltando lo referente al análisis de información, y describe cómo algunos productores de países en vías de desarrollo, incluido el Ecuador, pueden aprovechar los principios fundamentales de la AP para su beneficio sin necesariamente adquirir tecnologías costosas.

En el segundo capítulo se expone el caso de la adopción de la AP en el Ingenio San Carlos (ISC), uno de los ingenios azucareros más importantes del Ecuador.

El diseño de unidades de manejo sitio-específicas o lotes se constituye como el primer paso fundamental dentro del proceso. Se realiza un análisis estadístico descriptivo y exploratorio de la producción del ciclo de cultivo 2003 y los principales factores que inciden sobre la misma, expresada en toneladas métricas de caña por hectárea (TMC/ha.), cuya información se encuentra a un nivel agregado de datos de precisión (es decir, por lotes). El objetivo del capítulo es responder a la pregunta: ¿Cuánto varían las características medidas del suelo y del cultivo?

En el tercer capítulo se expone el marco teórico del análisis de regresión lineal múltiple o análisis econométrico desde dos perspectivas: el análisis tradicional y el análisis espacial. Se describen las principales especificaciones de cada rama y los procedimientos de estimación de Mínimos Cuadrados Ordinarios (MCO) y Máxima Verosimilitud (MV). Estas especificaciones son estocásticas y nos permiten explicar la variabilidad de la producción con respecto a sus factores determinantes (factores naturales y de manejo). Se describe a la producción de un cultivo como un sistema natural que conlleva una variabilidad intrínseca que puede describirse de manera adecuada mediante un modelo econométrico.

El objetivo principal del capítulo es ilustrar cómo los efectos espaciales de los datos en una sección cruzada pueden ser vistos como casos especiales en la especificación y estimación de modelos econométricos, demostrar las limitaciones de las técnicas estándar en el contexto espacial, y sugerir los métodos alternativos requeridos.

En el cuarto capítulo se evalúan dos especificaciones de modelos econométricos para la producción del cultivo de caña de azúcar en el ISC: un modelo espacial SAC estimado por el método de MV y otro modelo tradicional no espacial estimado por el método de MCO. De acuerdo a pruebas de diagnóstico y medidas de ajuste expuestas en el Capítulo 3 se escoge el modelo de mejores propiedades estadísticas y se interpretan cada uno de los coeficientes estimados. El objetivo del capítulo es que a partir de la mejor especificación se establezca una relación causa-efecto entre la producción y sus factores determinantes con la finalidad de establecer un diagnóstico de la situación del cultivo que nos conduzca a una mejor comprensión del sistema productivo y, consecuentemente, a correctas recomendaciones de manejo agrícola dentro del contexto de la AP.

En el quinto y último capítulo se exponen las conclusiones y recomendaciones que se derivan de la investigación

ÍNDICE GENERAL

	Pág.
RESUMEN	II
ÍNDICE GENERAL	III
ABREVIATURAS	IV
SIMBOLOGÍA	V
ÍNDICE DE FIGURAS	VI
ÍNDICE DE CUADROS	VII
INTRODUCCIÓN	1
CAPÍTULO 1	
1. AGRICULTURA DE PRECISIÓN	6
1.1. Manejo Sitio-específico del cultivo.....	7
El incremento de la producción por el manejo a gran escala.....	9
1.1.2. Tecnologías agrícolas dirigidas hacia el manejo uniforme.....	10
1.1.3. Beneficios potenciales de una manejo más eficiente.....	12
1.2. Origen, tecnologías y definición.....	14
1.3. Condición fundamental para la adopción.....	22

1.4. Problemas fundamentales para la adopción.....	28
1.5. Problemas fundamentales en el análisis de información:	
Perspectiva de un análisis econométrico.....	32
1.5.1. Efectos espaciales.....	34
1.5.1.1. Dependencia espacial.....	34
1.5.1.2. Heterogeneidad espacial.....	39
1.5.2. El problema de las unidades de área modificables.....	43
1.6. Secuencia general para la adopción.....	47
1.7. Adopción de los principios de la agricultura de precisión en América Latina.....	50

CAPÍTULO 2

2. ANÁLISIS ESTADÍSTICO EXPLORATORIO Y DESCRIPTIVO DE LA PRODUCCIÓN DE CAÑA DE AZÚCAR Y SUS FACTORES EN EL INGENIO SAN CARLOS	56
Características de la implementación de la agricultura de precisión	58

2.1.1. Lotización.....	61
2.1.2. Estructura y monitoreo de la producción.....	63
2.1.3. Descripción y monitoreo de los factores de producción.....	67
2.1.3.1. Factores Naturales.....	70
2.1.3.1.1. Edad del cultivo.....	70
2.1.3.1.2. Edad a la cosecha.....	70
2.1.3.1.3. Tercio de cultivo.....	71
2.1.3.1.4. Requerimientos climáticos.....	72
2.1.3.1.5. Requerimientos edáficos.....	78
2.1.3.1.6. Análisis de suelos.....	95
2.1.3.2. Factores de manejo.....	97
2.1.3.2.1. Sector.....	97
2.1.3.2.2. Variedad de la caña.....	98
2.1.3.2.3. Número de riegos.....	101
2.2. Análisis estadístico exploratorio.....	103
2.2.1. Población, marco muestral y muestra.....	104
2.2.2. Definición de variables.....	108
2.2.3. Análisis univariado.....	110

CAPÍTULO 3

3. MÉTODOS ESTADÍSTICOS PARA DATOS DE AGRICULTURA DE PRECISIÓN: ANÁLISIS ECONÓMICO DE PROCESOS

ESPACIALES	170
3.1. El modelo general de regresión lineal múltiple.....	172
3.1.1. Función de regresión poblacional (FRP).....	177
3.1.2. Notación Matricial.....	183
3.1.3. Función de regresión muestral (FRM).....	185
3.1.4. Supuestos clásicos del modelo.....	188
3.1.5. Estimación de Mínimos Cuadrados ordinarios (MCO)....	191
3.1.5.1. Propiedades del estimador MCO.....	194
3.2.5.2. Suma de cuadrados y su descomposición.....	198
3.2.5.3. Estimación de la varianza del error.....	202
3.2.5.4. Inferencia en el modelo lineal.....	205
3.1.6. Heteroscedasticidad y Autocorrelación.....	214
3.1.7. Estimación de mínimos Cuadrados Generalizados (MCG).....	221

3.1.8. Contraste de heteroscedasticidad de White.....	225
3.1.9. Multicolinealidad.....	227
3.2. El modelo espacial de regresión lineal múltiple.....	233
3.2.1. Expresión formal de los efectos espaciales:	
conectividad en el espacio.....	235
3.2.1.1. Matriz de conectividad espacial.....	238
3.2.2. Taxonomía de los modelos espaciales de regresión lineal múltiple.....	241
3.2.3. Limitaciones de la estimación MCO en los modelos de procesos espaciales.....	245
3.2.3.1. Estimadores MCO en presencia de un rezago espacial de la variable dependiente.....	246
3.2.3.2. Estimadores MCO en presencia de un rezago espacial en el error estocástico.....	250
3.2.4. Estimación de Máxima Verosimilitud (MV).....	251
3.2.4.1. Propiedades de los estimadores MV.....	254
3.2.5. La función de verosimilitud y el Jacobiano del modelo espacial general.....	258
3.2.6. Condiciones de primer orden para los estimadores MV en el modelo general.....	264

3.2.7. La matriz asintótica de varianzas del modelo general...	265
3.2.8. Pruebas de hipótesis basadas en el principio de MV...	267
3.2.8.1. Principios generales.....	268
3.2.8.2. La prueba de Wald para modelos de procesos espaciales.....	274
3.2.8.3. La prueba LR para modelos de procesos espaciales.....	276
3.2.8.4. La prueba LM para modelos de procesos espaciales.....	278
3.2.9. Dependencia espacial en el término de error.....	280
3.2.9.1. Pruebas para la autocorrelación espacial del error basadas en el principio de MV.....	281
3.2.9.2. Prueba para la autocorrelación espacial del error en un modelo SAR.....	286
3.2.10. Selección de modelos.....	290
3.2.10.1. Medidas de ajuste en los modelos espaciales.....	290
3.2.10.2. Criterios de información para la discriminación de modelos.....	293

CAPÍTULO 4

4. ESPECIFICACIÓN DE UN MODELO ECONOMETRICO PARA LA PRODUCCIÓN DE CAÑA DE AZÚCAR EN EL INGENIO SAN

CARLOS	296
4.1. Definición de variables.....	297
4.2. Especificación espacial MV.....	300
4.2.1. Análisis de los residuos.....	305
4.3. Especificación no espacial MCO.....	308
4.4. Evaluación de las especificaciones.....	311
4.5. Análisis de sensibilidad.....	314
4.6. Interpretación de resultados.....	321
4.6.1. Condiciones climáticas.....	321
4.6.2. Tercio de cultivo.....	323
4.6.3. Número de riegos.....	328
4.6.4. Variedades de caña.....	330
4.6.5. Textura del suelo.....	332
4.6.6. Sectores.....	333
4.6.7. Edad del cultivo.....	334
4.6.8. Edad a la cosecha.....	335
4.6.9. Parámetros del análisis de suelos.....	335

. 4.6.10. Interacciones.....	340
4.6.10.1. Textura de suelos y variedad.....	340
4.6.10.2. Tercio de cultivo y variedad.....	343
4.6.10.3. Tercio de cultivo y textura de suelo.....	343

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES	345
5.1. Conclusiones.....	346
5.1.1. Análisis de información.....	346
5.1.2. Manejo sitio-específico del cultivo.....	351
5.2. Recomendaciones.....	354
5.2.1. Análisis de información.....	354
5.2.2. Manejo sitio-específico del cultivo.....	356

ANEXOS

BIBLIOGRAFÍA Y REFERENCIAS

CAPÍTULO 1

1. AGRICULTURA DE PRECISIÓN

En este capítulo se tratarán las definiciones, principios elementales, principales tecnologías y problemas fundamentales que conlleva la Agricultura de Precisión. Partiendo desde el concepto general de un manejo sitio-específico del cultivo hasta la adaptación al medio local de esta nueva tendencia de por parte de los países agrícolas en vías de desarrollo, se hará un esfuerzo por no dejar de lado detalles importantes que faciliten la comprensión de estas nuevas tecnologías y métodos que están haciendo que la agricultura se dirija cada vez más hacia una *producción por especificación*, característica principal de los procesos de manufactura. El análisis de la información con referencia espacial sobre el rendimiento del cultivo y los factores que inciden sobre éste, se constituye en el determinante fundamental para demostrar la factibilidad de un proyecto de Agricultura de Precisión, por lo

que, se introducen técnicas estadísticas que modelan los efectos espaciales de los datos y cuyo objetivo es convertirse en una herramienta de gran utilidad para tomar decisiones correctas.

MANEJO SITIO-ESPECÍFICO DEL CULTIVO

¿Qué tipo de factores y en que medida influyen en el rendimiento de un cultivo?. ¿Cómo se podrían mejorar los rendimientos dentro de cada parcela o unidad mínima de manejo?. ¿Cómo deberían distribuirse los insumos agrícolas como los fertilizantes para obtener una relación costo-beneficio más eficiente?. Estas son algunas de las preguntas que, gracias a su experiencia, un productor agrícola podría contestarnos intuitivamente, aunque sin mucha precisión, y en su respuesta se resaltaría la importancia de reconocer que tanto los requerimientos de insumos (fertilizantes, pesticidas, cantidad de semilla, entre otros) y labores agrícolas como las condiciones naturales presentes (temperatura, humedad, textura del suelo, fertilidad, entre otras) que inciden sobre el rendimiento de un cultivo, no son uniformes dentro de una parcela; es decir, éstas cambian

geográficamente a medida que nos movemos de un extremo al otro dentro de la unidad de manejo. Entonces, conociendo este hecho, ¿qué sería lo ideal?. Nuestro amigo productor decidiría, en primera instancia, identificar las variaciones en las características del suelo y del cultivo así como el microclima específico del cultivo. Luego, por cada pequeña superficie dentro de la parcela que se pueda considerar “sin variaciones”, compararía los niveles medidos de las características más importantes de cada zona con los requerimientos del tipo de cultivo en el que se trabaja con el objetivo de establecer un diagnóstico de la situación en la que se encuentra cada una de estas pequeñas áreas “homogéneas”, para luego, sobre la base de un criterio sitio-específico, efectuar las recomendaciones más adecuadas de tratamientos agrícolas que permitan:

- Obtener mayores rendimientos con el mismo nivel de insumos, simplemente redistribuidos;
- Los mismos rendimientos con menores insumos; o,
- Mayores rendimientos y reducción de insumos.

Es decir, hacer de las prácticas agrícolas labores mucho más eficientes. Sin embargo, a pesar de que nos encontramos ante el hecho de que ya desde hace muchos años los productores tienen la noción de lo que debe hacerse para optimizar la aplicación de insumos y mejorar los rendimientos del cultivo, ¿por qué existe aún un marcado predominio del manejo uniforme?; y, ¿por qué no se ha dado ya la adopción definitiva de una filosofía de manejo sitio-específico entre todos los productores agrícolas?. La respuesta a estas pregunta involucra principalmente dos aspectos que deben ser analizados dentro de un contexto económico global y de costo-beneficio para la empresa: 1) el considerable incremento de la producción a partir del manejo a gran escala del cultivo; y, 2) la orientación del desarrollo de las tecnologías agrícolas hacia un manejo uniforme.

El incremento de la producción por el manejo a gran escala.

El manejo uniforme del cultivo o la aplicación homogénea de insumos sobre la unidad de manejo debe gran parte de su actual predominio al crecimiento de los campos (producciones extensivas) y a la

mecanización motorizada de la agricultura. A inicios del siglo XX los altos costos en mano de obra para trabajar sobre grandes extensiones de cultivo exigían algún tipo de automatización. La explotación en la utilización de tractores hizo que la producción fuera rentable trabajando grandes áreas con dosis uniformes de insumos debido a las ventajas de capacidad y velocidad que permiten cubrir muchas más hectáreas en un solo día. Los beneficios del manejo de unidades menores a las tradicionales, muy intensivos en mano de obra, quedaban opacados frente a las ventajas del incremento de la producción.

Tecnologías agrícolas dirigidas hacia el manejo uniforme

En el último siglo se han dado grandes avances técnicos y científicos en la agricultura que han permitido satisfacer la demanda mundial de alimentos y fibras. La mecanización, el mejoramiento genético de variedades (fitomejoramiento), la biotecnología, el manejo integrado de plagas y la fertilización se constituyen tan sólo en algunos de los

ejemplos donde se ha contribuido a obtener mayores producciones o rendimientos.

Para facilitar el entendimiento y la aplicación de estos avances, se establecieron los denominados “paquetes tecnológicos” entre los agricultores. Estos se constituyen en una “receta” para cada cultivo que indica cómo, cuándo y qué debe hacerse para asegurar cosechas y producciones satisfactorias [1]. Los paquetes tecnológicos han sido muy útiles y han dado buen resultado, particularmente, como una guía para los agricultores cuando decidían sembrar un determinado cultivo sin tener mayor experiencia en él. Sin embargo, los paquetes estaban dirigidos hacia tratamientos uniformes de las unidades de manejo agrícola y los productores notaron que la respuesta del cultivo al paquete tecnológico no era uniforme en todo el campo; es decir, en algunas áreas el cultivo germinaba y producía mejor que en otras, estaba más verde, era más vigoroso o simplemente no crecía. Era notorio que existía variabilidad en las condiciones de crecimiento, razón por la cual el manejo uniforme no era lo más apropiado [2].

Por otra parte, además de lo que sucede con los rendimientos de un cultivo, también se puede encontrar variabilidad en las características que determinan la fertilidad del suelo. Los niveles de acidez y los niveles de potasio y fósforo asimilables que se presentan en un análisis de suelos son fundamentales para establecer las recomendaciones de aplicación de fertilizantes. Las aplicaciones uniformes están basadas generalmente sobre muestras compuestas del suelo que se recolectan para representar los niveles promedio de una parcela de gran extensión. Con semejante enfoque, hay una gran posibilidad de sobre y sub aplicación de insumos.

Beneficios potenciales de un manejo más eficiente

Cada año los productores de todo el mundo aplican de manera uniforme más de 100 millones de toneladas de Nitrógeno (N), Fósforo (P) y Potasio (K). En los Estados Unidos el 97% de la superficie sembrada con maíz recibe una aplicación de fertilizante nitrogenado. Para un típico productor de maíz, los fertilizantes significan más de un cuarto del total de gastos en efectivo de una cosecha [3]. La

posibilidad de manejar estos insumos de tan alto costo puede tener un impacto significativo sobre la rentabilidad de un cultivo. Por otra parte, también debe considerarse el potencial de causar un impacto ambiental positivo debido a un manejo adecuado. Por ejemplo, los nitratos que se lixivian a través del suelo generalmente terminan en las napas, el mismo recurso que provee agua potable a más del 80% de los pobladores rurales [3]. El fósforo llega a lagos y cursos de agua como un contaminante adherido a las partículas de suelo arrastradas por la erosión. Niveles excesivos de N y P en los lagos pueden dañar varias formas de vida acuática teniendo así un impacto negativo significativo sobre algunos ecosistemas. Si se pudiesen combinar las aplicaciones de fertilizantes y herbicidas con las verdaderas necesidades de los cultivos, tendríamos como resultado, además de labores mucho más rentables y eficientes, una menor contribución de la agricultura a la contaminación ambiental y, por lo tanto, se daría un gran paso hacia la consecución de una **agricultura sostenible**.

ORÍGEN, TECNOLOGÍAS Y DEFINICIÓN

Los grandes avances técnicos y científicos en la agricultura se han visto reflejados en los ya mencionados “paquetes tecnológicos”. Estos paquetes han sido de mucha utilidad y han dado resultados satisfactorios. Sin embargo, la eficiencia del manejo actual de los recursos del suelo y de los recursos económicos en beneficio de una agricultura sostenible a largo plazo está siendo cuestionada. Los recursos naturales son cada vez más escasos y costosos, la competencia entre productores agrícolas es cada vez más reñida y las tecnologías más utilizadas no explotan completamente el potencial de los cultivos. Por esto, se hace cada vez más necesario acercarse a una nueva filosofía de manejo, dejar atrás todas aquellas prácticas tradicionales y dar inicio, lo más pronto posible, al desarrollo y adopción de nuevas formas de tecnología que estén al alcance de cualquier productor y que al mismo tiempo permitan un manejo más adecuado del cultivo.

Así, como resultado de la constante búsqueda de prácticas agrícolas más eficientes, a principios de la década de los 90 los adelantos

tecnológicos en la agricultura adquirieron un nivel que permitió retomar el viejo concepto de un manejo sitio-específico del cultivo dando lugar al nacimiento de uno de los más prometedores avances tecnológicos de finales del siglo XX: **La Agricultura de Precisión**.

La Agricultura de Precisión (AP) se originó en Estados Unidos y Europa para ser aplicada en siembras extensivas de cereales y granos, con la finalidad de facilitar la administración de una finca (toma de decisiones), después de un proceso de adquisición e interpretación de información sobre el sistema productivo [4]. Esta tendencia se basa en un manejo sitio-específico de áreas de cultivos o unidades de manejo (SSCM, del inglés: Site-Specific Crop Management), propuesto como una medida correctiva al ineficiente manejo de los recursos ambientales y financieros de las empresas agrícolas. En términos económicos, las decisiones a tomar benefician la rentabilidad de la producción del cultivo, siempre y cuándo, el cálculo y la localización precisa de los ingresos del sistema (fertilizantes, plaguicidas, riego, drenaje, entre otros) suponga y sugiera un uso más eficiente y rentable de los recursos de la empresa [2, 5 y 6].

Las tecnologías de información que se constituyen como el pilar de este avance tecnológico tienen como premisa **la recolección de datos georreferenciados**, es decir, datos con una referencia espacial o geográfica. Las principales herramientas [7] que representan a estas nuevas tecnologías de información geográfica son:

- **Sistemas de posicionamiento global por satélite (GPS).** Proporcionan la respuesta a la pregunta ¿dónde estoy? Es simplemente un “ubicador de posición” para personas, cosechadoras, pulverizadoras, etc. Ni más ni menos. La precisión de la señal gratuita GPS (del inglés: Global Positioning System) puede ser suficiente para algunas operaciones, pero o no para otras. Para lograr mayor precisión se puede pagar una corrección diferencial al GPS, lo que en conjunto se denomina DGPS (del inglés: Differential Positioning System).
- **Monitor de rendimiento y mapeo.** Mide y graba el rendimiento de pequeñas áreas o “sitios” dentro de la unidad de manejo en forma continua, a medida que se cosecha. Cada sitio tiene un ancho específico (el ancho de corte), un largo específico (la

distancia recorrida por la cosechadora en el intervalo de grabación: 1 a 5 segundos), y una ubicación única (coordenadas x, y). El monitor de rendimiento también estima y graba el contenido de humedad y la cantidad cosechada en cada sitio. El rendimiento se calcula como la cantidad que se cosecha en cada sitio dividida por su área. Con ayuda de un programa (GIS) se pueden hacer mapas que muestran la ubicación de estos sitios de cosecha, con sus respectivos rendimientos y humedades.

- **Sistema de información geográfica (GIS).** Es un programa de computación que puede recolectar, clasificar, mapear, graficar, analizar, y mostrar datos de producción con una referencia espacial (coordenadas: latitud y longitud). Un GIS (del inglés: Geographical Information System) es un sistema de “información” porque permite organizar los datos para que sea posible analizarlos, evaluarlos y tomar decisiones. Es el medio para transformar los datos en información útil para la planificación y la administración de los recursos.

- **Tecnologías para dosis variable de insumos (VRT).** Las VRT (del inglés: Variable Rate Technologies) permiten ajustar las dosis de insumos de acuerdo al mapa de aplicación realizado en un GIS. Requiere el uso de un GPS para conocer la ubicación del equipo en la unidad de manejo. Una computadora integra la información del mapa de aplicación y del GPS, enviando la información al controlador del equipo para variar la dosis recomendada sobre la marcha. Si no se dispone de un sistema de dosis variable automático, una alternativa es la dosis variable manual.
- **Percepción remota.** Es la ciencia y el arte de obtener información sobre un objeto, área o fenómeno a través del análisis de datos obtenidos con un aparato (sensor remoto) que no está en contacto físico con ese objeto, área o fenómeno de estudio. El sensor remoto puede estar a varios centímetros o a varios kilómetros, dependiendo del sistema usado y de la información deseada. Ejemplo: el sensor de nitrógeno, fotografías aéreas, imágenes satelitales, etc.

Todas estas tecnologías en conjunto nos permiten describir a la AP como *la aplicación práctica del manejo sitio-específico en la agricultura comercial mediante la automatización que nos da la informática*; o, poniendo mayor énfasis en la utilización de datos con referencia espacial:

La Agricultura de Precisión es la utilización de modernas herramientas tecnológicas que permiten la obtención y análisis de datos georreferenciados sobre las características de un cultivo, mejorando el diagnóstico, la toma de decisiones y la eficiencia del uso de los mismos [3].

Este concepto combina técnicas de cultivo con tecnologías de información permitiendo que el productor agrícola pueda medir, analizar y manejar la variabilidad del rendimiento y de sus factores con el potencial de reducir los costos de producción, aumentar la productividad y hacer un uso más eficiente de los insumos. En un sentido más específico, la AP permite:

1. Administrar los insumos en el tiempo y en el espacio (reducir desperdicios);
2. Optimizar la logística de las operaciones en el campo;
3. Manejar los riesgos de la producción;
4. Vender productos diferenciados;
5. Proveer trazabilidad de los productos para consumo humano; y,
6. Documentar los insumos aplicados para cumplir con reglas e protección ambiental.

Todo este potencial de desarrollo y el contexto tecnológico que envuelve el concepto de la AP hace que la opción de involucrarse en esta filosofía de manejo sea, sin lugar a dudas, muy atractiva. Sin embargo, las características inherentes a las principales herramientas tecnológicas de precisión y el actual contexto económico sugieren que la dinámica de cambio de un manejo uniforme a las prácticas de AP no corresponda a la de una rápida curva de adopción [8]. Esto se debe principalmente a que:

- 1) La tecnología está inmadura; es decir, los productores todavía no disponen de sistemas completos de AP en un solo paquete.

Sistemas parciales pueden ser rentables para algunos productores, pero no para todos.

- 2) La tecnología tiende en si misma a generar modificaciones. La AP no es una elección de “si o no”. Los productores pueden y van a modificar la tecnología substancialmente.
- 3) Las instituciones todavía no están listas para manejar los datos de AP. Para sacar el mayor provecho de esta tecnología se requiere de un banco de datos, pero se carece actualmente de una estructura organizacional que permita crear este banco de datos efectivamente.
- 4) La agricultura moderna es una actividad cada vez más riesgosa. Con precios inestables de los productos agrícolas, la adopción de la AP puede estar sujeta a muchos altibajos.

Cualquier cambio tecnológico es siempre confuso y desorganizado y la AP no es una excepción. La tecnología no aparece repentinamente totalmente formada y perfectamente operacional. Usualmente requiere de un período de adaptación, con productores innovadores, fabricantes y científicos, cada uno, haciendo lo que le corresponde

para hacer que la tecnología sea tanto rentable como práctica [8]. Esta en manos de la gente de negocios y los productores innovadores jugar un papel preponderante y tomar decisiones tecnológicas. No se puede esperar a que el polvo se asiente y a que la tecnología madure pues la mayor parte de la rentabilidad de cualquier nueva tecnología es aprovechada por los primeros en adoptarla correctamente. Es imperativo que cada productor aprenda todo lo esencial de estas revolucionarias prácticas agrícolas para superar cualquier barrera y lograr desarrollar estas nuevas tecnologías a un nivel que ofrezca los mayores beneficios posibles.

CONDICION FUNDAMENTAL PARA LA ADOPCIÓN

La condición fundamental que lleva a la adopción de la AP, y a la aplicación variable de insumos en particular, es la **variabilidad espacial** que consiste en la diferencias tanto de los rendimientos como de las características medidas en el cultivo y en el suelo (factores del rendimiento) en distancia y profundidad. Generalmente se encuentra variabilidad en todas las parcelas. La variabilidad se

observa en la fertilidad del suelo, contenido de humedad, textura del suelo, topografía, vigor del cultivo, incidencia de insectos, enfermedades o competencia de malezas. Sin embargo, el hecho de que esta variabilidad este presente no significa que siempre tenga sentido hacer algo al respecto. Son tres las preguntas críticas que un productor debe poder responder antes de poder adoptar la AP [3]. Las respuestas a éstas van a ayudar a determinar el potencial para que la AP sea implementada con éxito y rentabilidad:

- 1. ¿Cuánto varían las características medidas del suelo y cultivo?**
- 2. ¿Cuánto afectan estas variaciones al rendimiento del cultivo?**
- 3. ¿Puede el productor obtener suficiente información y la tecnología apropiada para manejar esta variabilidad con provecho económico?**

En el inicio, el productor debe determinar cuánta variabilidad existe en las características del suelo y del cultivo; luego, el próximo y más

importante paso es encontrar las relaciones **causa-efecto** entre las variables medidas y obtener el resultado en términos de rendimiento y calidad. Esto significa saber cuánto puede producir cada área del campo y qué es lo que está evitando que un área este rindiendo su máximo potencial. Si el productor puede averiguar qué está causando la diferencia en rendimiento o calidad, debe determinar si las herramientas y recursos que posee pueden resolver el problema.

Rendimientos altamente variables dentro de la unidad de manejo indican que las prácticas tradicionales o uniformes pueden no estar proveyendo las mejores condiciones de crecimiento en todas las partes de la parcela. En este caso la AP resultaría beneficiosa. Por el contrario, si el rendimiento de una parcela no varía significativamente y su nivel es satisfactorio, entonces hay poco incentivo para invertir en tecnologías adicionales de AP.

La variabilidad que existe dentro de una unidad de manejo se puede resumir en tres clases [7]: 1) **natural**, tal como el tipo de suelo o topografía; 2) **aleatoria**, como las precipitaciones; y, 3) **de manejo**, como la aplicación de fertilizantes y la densidad de siembra.

La variabilidad natural incluye (a) variabilidad de suelos, (b) variabilidad biológica, y (c) variabilidad en los procesos dinámicos del suelo [9].

1.a) El suelo puede variar espacialmente en la capacidad y retención de agua, materia orgánica, y en otras características físicas y químicas por tipos de topografía, como así también por una serie de elementos interactuantes. Uno de los desafíos es cuantificar la variabilidad de los suelos.

1.b) La variabilidad biológica dentro de las unidades de manejo es tan grande como la variabilidad de suelos, incluyendo la variabilidad en la población de microbios del suelo, población de malezas, población de insectos, ocurrencia de enfermedades, crecimiento del cultivo y por último y no menos importante: **el rendimiento**, que es la variable diagnóstico que permite a los productores darse cuenta del resultado de toda a variabilidad biológica ocurrida durante todo el ciclo del cultivo.

1.c) La variabilidad en los procesos dinámicos del suelo se puede entender mirando, por ejemplo, la dinámica del Nitrógeno (N): el N en el suelo representa el balance existente entre un número de procesos tales como la mineralización, nitrificación, adsorción, absorción por los cultivos, y percolación.

Los investigadores, asesores y productores han documentado numerosas fuentes naturales e inducidas (de manejo) de variación del rendimiento. El Cuadro 1.1 muestra las fuentes de variación que pueden y son expresadas en los mapas temáticos de un GIS.

Cuadro 1.1. Fuentes de variación del rendimiento de un cultivo

FUENTES DE VARIACION	EJEMPLOS
VARIABILIDAD NATURAL	
Clima	<ul style="list-style-type: none"> • Cantidad, frecuencia e intensidad de precipitaciones • Radiación solar • Vientos • Temperatura máxima, mínima, amplitud diurna y nocturna
Relación suelo/clima	<ul style="list-style-type: none"> • Drenaje • Profundidad del suelo • Capacidad de retención del agua
Propiedades físicas y químicas del suelo	<ul style="list-style-type: none"> • Textura (arcilla, limo, arena, fragmentos gruesos) • Estructura y densidad de agregados • Profundidad de capas restrictoras • Disponibilidad de nutrientes • Materia orgánica, pH, salinidad • Capacidad de intercambio calórico
Pendiente y posición del sitio en el paisaje	<ul style="list-style-type: none"> • Efecto sobre la erodabilidad, temperatura del suelo, calidad del suelo.
Infestación de plagas	<ul style="list-style-type: none"> • Malezas, insectos, enfermedades, microfauna.
VARIABILIDAD INDUCIDA POR MANEJO	
Condiciones Del cultivo e insumos	<ul style="list-style-type: none"> • Selección del híbrido o variedad • Densidad de siembra y uniformidad • Distanciamiento entre hileras • Profundidad de siembra • Control de malezas • Control de enfermedades y plagas • Fertilizante aplicado • Tipo, dosis, momento y posicionamiento del fertilizante y agroquímico • Eficiencia de aplicación de cada insumo
Historia del lote	<ul style="list-style-type: none"> • Rotaciones • Labores y compactación • Años de directa y cobertura del suelo con residuos • Estructuras y prácticas anteriores: caminos, aguadas, alambrados, etc.
Prácticas culturales y/o errores	<ul style="list-style-type: none"> • Mala aplicación de agua de riego, fertilizantes y/o agroquímicos • Problemas de sembradoras o cosechadoras • Conservación de la humedad del suelo

Fuente: Proyecto de Agricultura de Precisión INTA Manfredi [3]

Elaboración: Xavier Vítores, 2004

Es difícil determinar un orden de importancia de estos factores basado en un impacto relativo sobre el rendimiento porque éstos cambian de año en año y de parcela en parcela; sin embargo, se los menciona para mostrar que el rendimiento de una parte de la unidad de manejo puede estar indicado y ponderado por uno o varios de los factores.

PROBLEMAS FUNDAMENTALES PARA LA ADOPCIÓN

La rentabilidad a la inversión en un proyecto es el objetivo principal que todo productor agrícola busca y es el marco dentro del cual se tomarán las decisiones definitivas sobre la adopción de las prácticas de AP. Aquí es donde el ***análisis de la información georeferenciada*** se constituye en el determinante más importante. Aunque representa una gran ventaja contar con una estructura organizacional que posea un banco de datos completo, es el análisis de la información con referencia espacial el que permite obtener información agronómica y económica de utilidad práctica. Sin embargo, ¿existe un camino definido entre el análisis de datos y la adopción justificada de un manejo sitio-específico del cultivo?.

En un futuro muy cercano se marcará una gran diferencia entre los productores que sepan manejar estas nuevas tecnologías de información agronómica georeferenciada proveniente de ensayos en sus propios campos; hoy, sin embargo, existen dos barreras de información clave que deben ser superadas:

1. Determinar el **potencial de aplicación** de un manejo sitio-específico del cultivo [10]; y,
2. La existencia de una **brecha de información** entre el análisis de datos espaciales y las recomendaciones sitio-específicas que hace difícil determinar como deberían ajustarse las prácticas de manejo y las dosis variables de pesticidas, fertilizantes y otros insumos para maximizar la rentabilidad, minimizar el impacto ambiental y alcanzar otros objetivos [11].

Desde la perspectiva de un productor, decidir si existe la suficiente variabilidad en los rendimientos tanto en magnitud como en estructura espacial, es un aspecto clave para justificar el costo del cambio de un manejo tradicional (uniforme) a un manejo sitio-específico. Los mapas de rendimientos proporcionados por los GIS proveen una pista clave

hacia el **potencial de aplicación** de un manejo sitio-específico. Sin embargo, a pesar de que se han elaborado algunos métodos para modelar la variación de los rendimientos, queda aún mucho trabajo por hacer para cuantificar en una sola medida o índice la oportunidad de adopción de un manejo sitio-específico del cultivo que incorpore tanto sus posibles beneficios económicos y ambientales como la variabilidad espacial de los rendimientos.

Adicionalmente, determinar las prácticas de manejo diferenciadas y las dosis óptimas de insumos a aplicar como fertilizantes, pesticidas o semilla a partir de la información georeferenciada se ha convertido en una meta y al mismo tiempo en una barrera al momento de determinar la rentabilidad de un proyecto de AP. Esta barrera se traduce en una **brecha de información** que se ha intentado superar con una gran cantidad de métodos estadísticos que tienen como propósito describir y modelar datos. La Figura 1.1 muestra los pasos involucrados en el ciclo de la AP y la brecha de información (en amarillo).

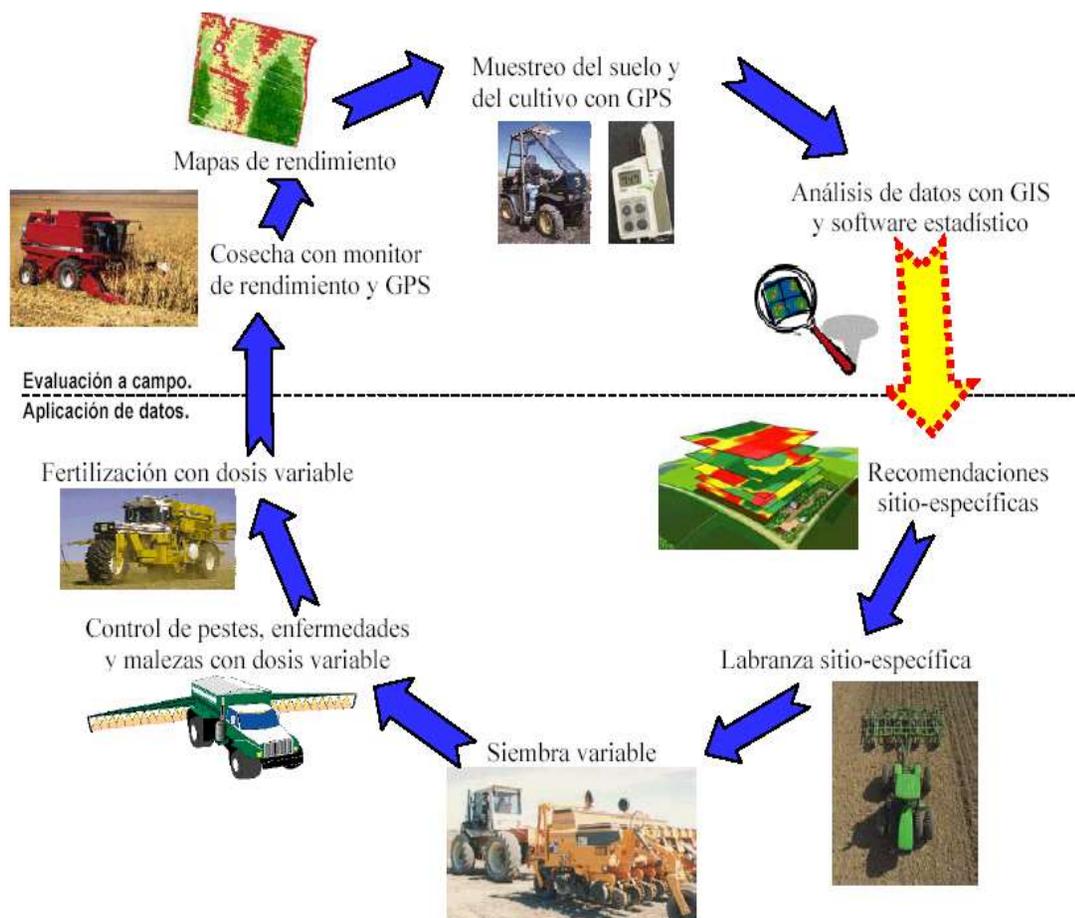


Figura 1.1. Ciclo de la AP y la brecha de información

Fuente: R. Bongiovanni [12].

PROBLEMAS FUNDAMENTALES EN EL ANÁLISIS DE INFORMACIÓN: PERSPECTIVA DE UN ANÁLISIS ECONOMÉTRICO

Dentro de las técnicas estadísticas más comúnmente aplicadas en la agricultura para la evaluación de tratamientos de insumos y prácticas agrícolas están: el diseño experimental de bloques completos aleatorizados y los modelos econométricos para la respuesta sitio-específica del cultivo con respecto a sus factores determinantes utilizando estimadores de Mínimos Cuadrados Ordinarios (MCO). Sin embargo, una condición fundamental para la aplicación de estas técnicas es que las observaciones dentro de la muestra a analizar deben ser independientes entre sí, mientras que, es muy notorio que las respuestas de las parcelas tienden a ser similares cuando son adyacentes; es decir, existe cierta dependencia o relación de tipo espacial. Partiendo de este hecho, una de las posibles explicaciones a la brecha de información antes mencionada es que la estructura espacial de los datos no se tiene debidamente en cuenta, lo que afecta la selección de tratamientos óptimos, la estimación de funciones de respuesta sitio-específica del cultivo y,

consecuentemente, a los análisis de factibilidad de proyectos de AP [12].

Afortunadamente, el gran avance que se ha dado en la actualidad en investigaciones de fenómenos físicos tanto en biología como en geología y en el estudio de modelos relacionados con economía regional, epidemiología y criminología (cuyo principal propósito es el análisis de datos espaciales), ha permitido desarrollar técnicas estadísticas específicas para detectar y modelar ciertos patrones que presentan los datos en el espacio. Entre los problemas más estudiados dentro del análisis de datos con referencia espacial se encuentran:

- Aspectos espaciales específicos de los datos o **efectos espaciales**; y,
- El problema de las **unidades de área modificables**.

1.5.1. Efectos espaciales

Los efectos espaciales son la razón esencial para la existencia de nuevas técnicas estadísticas que incorporan un componente espacial dentro de sus modelos. Se distinguen dos tipos generales de efectos: ***dependencia espacial y heterogeneidad espacial.***

1.5.1.1. Dependencia Espacial

En muchos ejercicios prácticos en las ciencias regionales aplicadas, la información es obtenida de observaciones que se encuentran ordenadas en el espacio y en el tiempo. Un ejemplo de esta situación puede encontrarse en la utilización de datos de población, empleo, y otras actividades económicas, recolectados (agregados) por unidades administrativas tales como estados, provincias, o unidades censales. Otros casos se presentan cuando las observaciones se organizan en una red de celdas en un sistema de coordenadas construido artificialmente, tal como en un mapa digitalizado de un GIS: esta aplicación de vital importancia para el análisis de información

agronómica en proyectos de AP donde por lo general los datos están almacenados por unidades de manejo en una base de datos y presentados en los mapas temáticos de rendimientos, texturas de suelos, condiciones de fertilidad, etc.

Analizar datos con un componente espacial se asemeja al problema de analizar datos que se relacionan entre sí en el tiempo (dependencia temporal) para lo cual dentro de la econometría se han desarrollado las técnicas agrupadas en el análisis de series de tiempo. Sin embargo, esta semejanza es solo parcial, pues si se examina detenidamente el caso, los resultados de un análisis de series temporales no se pueden adaptar a la existencia de **dependencia espacial** en una muestra obtenida en el espacio o sección cruzada. Esto se debe principalmente a la naturaleza multidireccional de la dependencia en el espacio, la cual, opuesta a una clara situación unidireccional en el tiempo, hace que la aplicación de muchos resultados simplificados sea inapropiada y se necesite de una metodología diferente. Por esto, dentro del análisis estadístico espacial se han desarrollado diversos métodos para analizar datos tomando en cuenta explícitamente su ubicación. Así, el concepto

estadístico de *autocorrelación temporal* tiene su equivalente en los **modelos espaciales de regresión lineal múltiple** de la **econometría espacial** en la denominada **autocorrelación espacial**. En ella, la dependencia espacial entre las observaciones de una sección cruzada de tamaño n significa que cada observación sobre una variable y en la unidad espacial $i \in S$ (donde S es el conjunto que contiene todas las unidades espaciales de observación) está relacionada formalmente a través de una función f con las magnitudes para la variable en otras unidades espaciales del sistema (Ec. 1.1); es decir, lo que se observa en un punto en el espacio es determinado (en parte) por lo que sucede en otros puntos, por lo que, la posición relativa de unas observaciones con respecto a otras se considera importante y es incluida explícitamente en el análisis.

$$y_i = f(y_1, y_2, y_3, \dots, y_n) \quad (1.1)$$

Dos tipos de condiciones llevan a la existencia de dependencia espacial [13]. La primera es el producto de los errores de medida para las observaciones en unidades espaciales vecinas. La segunda es más fundamental y está vinculada directamente al proceso bajo

investigación que implica la existencia de una variedad de fenómenos espaciales que interactúan entre sí.

En la primera situación, debido a que los datos son recolectados únicamente a un cierto nivel de agregación (escala de agregación de datos) basándose en criterios ad hoc, ***habrá poca correspondencia entre el alcance del fenómeno espacial bajo estudio y la delineación de las unidades espaciales de observación.*** Como consecuencia, los errores de medida son casi inevitables. Además, éstos tienden a “esparcirse” a través de las fronteras de las unidades espaciales. Como resultado, los errores en una observación, por ejemplo en una unidad i , muy probablemente están relacionados a los errores en una unidad vecina j . Esta situación en los errores de medición es una causa inmediata de la presencia de dependencia espacial y, como se tratará posteriormente, de heteroscedasticidad.

En el contexto de un análisis econométrico, esta situación resulta en una matriz no escalar de varianzas-covarianzas del error estocástico (como se verá en el Capítulo 3) y, además, a errores en la medición de las variables analizadas. Este hecho se ilustra en la situación

hipotética de la Figura 1.2 donde la verdadera escala espacial de las variables en estudio son las áreas A, B y C, mientras que las observaciones están agregadas en los niveles 1 y 2. Como resultado, la variable observada Y_1 será un agregado de Y_A y parte de Y_B , y la variable observada Y_2 será un agregado de Y_C y lo que queda de Y_B :

$$Y_1 = Y_A + \lambda Y_B$$

$$Y_2 = Y_C + (1 - \lambda)Y_B$$

Esta escala de agregación presentará muy probablemente errores en la asignación de un valor al parámetro λ que está presente tanto en Y_1 como en Y_2 . Como resultado, estos errores de medida generarán un patrón que exhibe dependencia espacial.

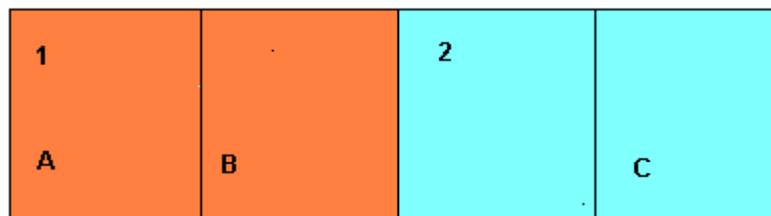


Figura 1.2. Dependencia Espacial y Agregación

Fuente: L. Anselin [13]

El segundo factor que produce dependencia espacial nace de la importancia de considerar al espacio como un elemento crucial en la especificación de modelos que tratan de explicar el comportamiento humano o de algunos sistemas naturales. La esencia de las ciencias regionales y geografía es que los componentes de distancia y ubicación son importantes y resultan en una variedad de interdependencias tanto en el espacio como en el tiempo: tal como se expresa en la **primera ley de la geografía** de Tobler [14]: *“Todo en el espacio está relacionado, pero las cosas cercanas están más relacionadas que las cosas distantes”*.

1.5.1.2. Heterogeneidad espacial

El segundo tipo de efecto espacial, la **heterogeneidad espacial**, está relacionado con la falta de estabilidad en el espacio de las relaciones bajo estudio. Siendo más precisos, esto implica que las formas funcionales y los parámetros de una especificación pueden variar con la ubicación de las unidades espaciales y no son homogéneos a través del conjunto de datos. Por ejemplo, es probable encontrar esta

situación en modelos econométricos estimados sobre secciones cruzadas compuestas por conjuntos de unidades disímiles, tales como regiones con diferentes tipos de suelos o regiones con diferentes variedades en un determinado cultivo.

En términos generales, existen dos aspectos distintos de heterogeneidad espacial: 1) **inestabilidad estructural** expresada mediante el cambio en las formas funcionales o variación de parámetros (variación sistemática o aleatoria); y, 2) **heteroscedasticidad** resultado de variables omitidas y errores de especificación que conducen a términos del error con varianza no constante. Ignorar estos efectos tiene consecuencias negativas para la validez estadística del modelo econométrico estimado: estimadores sesgados (excepto cuando sólo se presenta heteroscedasticidad), niveles de significancia incorrectos y predicciones que no resultan óptimas.

En la medida en que esta heterogeneidad este relacionada directamente con la ubicación en el espacio de las observaciones (tal como en el ejemplo de diferentes tipos de suelo en una finca), se la

considera como heterogeneidad espacial. En el trabajo econométrico, la tarea de modelar la heterogeneidad en la muestra puede llevarse a cabo explícitamente mediante una variación de parámetros, coeficientes aleatorios o varias formas de cambio estructural.

Los aspectos de heterogeneidad espacial pueden ilustrarse fácilmente en el contexto de un análisis de regresión donde los datos de una sección cruzada están combinados con una serie temporal (Ec. 1.2).

$$y_{it} = f_{it}(x_{it}, \beta_{it}, e_{it}) \quad (1.2)$$

Aquí, el índice i está vinculado a una unidad espacial, y el índice t a un periodo en el tiempo. La relación funcional f_{it} es una especificación espacio-temporal que explica el valor de la variable dependiente y_{it} (o un vector de variables dependientes) en términos de un vector de variables independientes x_{it} (regresores), un vector de parámetros β_{it} , y el término de error e_{it} . Como puede observarse, cada observación tiene su propio conjunto de parámetros, variables explicativas, especificación funcional y error. Esta formulación es claramente no operacional debido a que existen más parámetros que

observaciones; por lo que, para llevar a cabo procedimientos de estimación e inferencia y asegurar la identificación de un modelo, algunas restricciones deben ser impuestas sobre la expresión general. Es necesario no ser tan específicos en las características inherentes a cada unidad espacio-temporal y permitir la identificación de parámetros y formas funcionales que estén acorde con las restricciones impuestas por la disponibilidad de datos. En la práctica la falta de información es severa, por lo que es imperativo que la simplificación del modelo se realice bajo un buen criterio. Este problema es aún más complejo cuando la heterogeneidad espacial ocurre en combinación con dependencia espacial.

La contribución particular de la econometría espacial a este problema consiste en aplicar conocimientos de una teoría subyacente sobre la estructura espacial e interacción espacial de las observaciones que sirvan como base para la utilización de ciertas restricciones y parametrizaciones. Como en el caso de dependencia espacial, estos conocimientos teóricos se constituyen en el punto de partida para la selección e implementación de ciertas especificaciones para el

modelo, sin las cuales, se estaría tratando con un problema empírico de un alto grado de complejidad y de difícil resolución.

1.5.2. El problema de las unidades de área modificables

Tal como se indicó anteriormente, muchas aplicaciones requieren trabajar con datos espaciales agrupados por unidades de área definidas como unidades administrativas o legales (municipios, estados, distritos electorales, etc.) o como zonas naturales (zonas clasificadas por tipo de vegetación o de suelo), ya sea porque resulte natural hacerlo o porque sea necesario.

La forma en que se diseñan estas “zonas”, por lo general, es independiente de la información misma, utilizándose criterios administrativos u otros criterios que facilitan la recolección de datos. Esta situación hace que dichas zonas sean consideradas arbitrarias pues representan tan solo una regionalización de muchas posibles. Estos criterios, además, tienden a basarse en una sola variable (por ejemplo, número de habitantes o tipo de suelo) lo que implica que no

sean realmente adecuadas para analizar otros factores. Así, aún cuando una regionalización se realice siguiendo un criterio de homogeneidad, es prácticamente imposible que el mismo sea válido para todas las demás variables.

Los resultados del estudio de datos agregados arbitrariamente por área están en función de la escala, la orientación y el origen de la regionalización empleada así como del número de zonas. A este fenómeno se lo conoce como el ***problema de las unidades de área modificables*** y se puede enunciar como sigue: los resultados obtenidos del análisis de datos agregados dependen de la forma en que hayan sido agrupados dichos datos; es decir, depende tanto del número de zonas como de la forma de éstas. Así, el problema de las unidades de área modificables tiene dos partes, el llamado “efecto de escala” y el “efecto de la división en zonas”. El efecto de escala provoca que, al calcular una estadística (como varianza o correlación) a diferentes escalas, se obtengan resultados distintos. Por su parte, el efecto de la división en zonas provoca que, al reagrupar los datos en sistemas de zonas diferentes aunque a la misma escala, se obtengan también distintos valores para un mismo estadístico.

En la econometría, el problema de las unidades de área modificables está considerado como una combinación de dos problemas familiares [13]: **agregación e identificación.**

El primer aspecto del problema está relacionado con la agregación de datos en las unidades espaciales. Si el fenómeno bajo estudio es homogéneo a través de las unidades de observación, el problema no es significativo. Si este no es el caso, la heterogeneidad inherente al fenómeno y su inestabilidad estructural debe ser tomada en cuenta en el momento de realizar cualquier tipo de agregación para evitar problemas en la especificación del modelo relacionados a la forma de la heterogeneidad espacial presente.

El segundo aspecto tiene que ver a la identificación de la estructura de la dependencia espacial de los datos. En la econometría espacial, el análisis de una posible asociación espacial entre los datos es llevado a cabo relacionando una variable con su contraparte espacial o **rezago espacial.** Este último se construye como una combinación lineal de las observaciones del sistema. Comúnmente, la asociación está indicada por un coeficiente de regresión. Por ejemplo, la variable

dependiente y puede relacionarse con ρWy (Ec. 1.3), donde ρ es un coeficiente autoregresivo espacial y W es una matriz de conectividad espacial (que se tratará en más detalle en el Capítulo 3). Una implicación del problema de las unidades de área modificables es que diferentes especificaciones de W resultan en diferentes valores para ρ , y por ende, en una medida de asociación espacial indeterminada.

$$y = \rho Wy + e \quad (1.3)$$

Esta aparente indeterminación del parámetro ρ se considera un problema significativo en los análisis exploratorios donde no existe suficiente información en los datos para derivar la correcta estructura de la dependencia espacial. En la econometría espacial, razones teóricas *a priori* dictan la forma particular de las restricciones a tener en cuenta en el momento de especificar la estructura de la dependencia espacial de los datos.

1.6. SECUENCIA GENERAL PARA LA ADOPCIÓN

Una vez expuestas las principales definiciones, tecnologías involucradas y los problemas fundamentales que deben resolverse al adoptar un manejo sitio-específico del cultivo mediante tecnologías de información geográfica, es importante resumir y mostrar una secuencia general a seguir en la implementación de la AP. Cada paso es de vital importancia y debe ser analizado detenidamente tanto en forma individual como en conjunto. La secuencia implica:

1. Conocimiento e interpretación de las prestaciones de las herramientas de AP.
2. Obtención de abundante información georreferenciada de los rendimientos y sus factores.
3. Elaboración de los primeros mapas de rendimientos, fertilidad, pH, conductividad, etc., apoyados con buenas plantillas de seguimiento del cultivo;
4. Análisis de la variabilidad inducida, natural y aleatoria de las parcelas a través de toda la información georeferenciada utilizando paquetes de software específicos (GIS y software estadístico) y un

- grupo de trabajo interdisciplinario capaz de elaborar un diagnóstico de aplicación de insumos basados en conocimientos agronómicos;
5. Planificación de la aplicación variable de insumos para cada sitio específico dentro del cultivo (recomendaciones de manejo).
 6. Evaluación de las estrategias agronómicas planificadas para el cultivo a través de nuevos mapas de rendimientos.
 7. Realizar nuevos ajustes del diagnóstico.

Una vez ubicada la variabilidad de los rendimientos y sus causas, el diagnóstico preciso para cada sitio específico se convierte en la etapa crítica de la AP. La calidad de las observaciones, el seguimiento realizado por personas con conocimiento agronómico e interdisciplinario y el buen criterio permitirán profundizar en el diagnóstico.

Los mapas de fertilidad, pH, conductividad, textura de suelos, entre otros, superpuestos a los de rendimiento pueden correlacionarse orientando el diagnóstico según la necesidad de conocer la influencia de una u otra variable sobre la producción del cultivo. Luego, con toda la información previa se confeccionan mapas de prescripción con

directivas precisas de aplicación de insumos para cada sitio dentro de las unidades de manejo. Utilizando un monitor específico y un navegador satelital se podría, a través de un dosificador, variar la dosis del insumo al llegar a ese sector con posicionamiento satelital, técnica conocida como Manejo Diferencial de Insumos (MDI). Finalmente, un nuevo mapa de rendimientos del cultivo permitirá evaluar los resultados obtenidos permitiendo un ajuste del diagnóstico inicial.

Toda esta nueva tecnología de aplicación de insumos con dosis variable (VRT), podrá manifestarse como un incremento del rendimiento o reducción en el costo de aplicación, ya que lo que se pretende es evitar sub o sobre aplicaciones que pueden perjudicar el negocio o generar problemas de contaminación ambiental.

1.7. ADOPCIÓN DE LOS PRINCIPIOS DE LA AGRICULTURA DE PRECISION EN AMÉRICA LATINA

Los sistemas de producción agrícola tropicales de América Latina tienen diferencias importantes en el aspecto operativo comparado con los sistemas de Estados Unidos y Europa donde se originó la AP. Esta circunstancia demanda modificaciones y adaptaciones en las aplicaciones de la AP si realmente se aspira a implementarla ampliamente en Latinoamérica [1].

La agricultura tropical es diversa, incluye muchos cultivos anuales y perennes y la mano de obra es barata. Sin embargo, muchas labores son realizadas aún de forma manual (por ejemplo la cosecha), lo que desde un principio ha limitado la aplicación de la AP, tal como se la conoce, a algunos productores con una agricultura comercial a gran escala y mecanizada (Ej: Brasil y Argentina). Las altas tasas de interés de créditos para producción, los altos costos de algunas tecnologías y el elevado nivel de riesgo, desalientan la inversión en toda nueva tecnología incluyéndose la AP.

El monitoreo de la producción para elaborar mapas de cosecha de alta resolución y el manejo diferencial de insumos en tiempo real o “sobre la marcha” utilizando navegación satelital (GPS) son consideradas como las tecnologías más comunes de la AP; sin embargo, su aplicación práctica en América Latina involucra costos relativamente altos que comprenden:

1. Compra e instalación de equipamiento electrónico desconocido:
La mayoría de la tecnología de AP debe ser importada lo que ocasiona que sus costos sean mucho mayores que en Estados Unidos y Europa.
2. Capacitación: Entrenamiento de operarios para recolectar datos de precisión;
3. Asistencia técnica;
4. Análisis de muestras de suelos: Son la base de las decisiones en la aplicación variable de insumos; y,
5. Interpretación de datos: Es un desafío en cualquier lugar donde se aplique AP.

Estos costos hacen que la adopción de modelos de tecnologías norteamericanas o europeas resulte poco rentable para la mayoría de países de América Latina. Sin embargo, debido a que estos países dependen en gran medida de la producción de cultivos primarios, sus utilidades son reducidas y se encuentran en condiciones de fuerte competencia, los principios de AP, que implican la adquisición de información detallada e interpretación del sistema productivo para tomar mejores decisiones de manejo, resultan muy relevantes y no deben ser descartados. Por esto, es necesario buscar nuevas alternativas innovadoras, adaptar estos principios a las condiciones locales y recordar que la agronomía y economía de la AP son también sitio-específicas.

Algunas aplicaciones del concepto de AP ya han sido implementadas en algunos países de América Latina con éxito (Ej: Costa Rica, Argentina y Ecuador) y se ha demostrado que la optimización de beneficios y la adopción de mejores prácticas de manejo no se ven limitadas por el nivel de resolución espacial en la adquisición de los datos [15], sino mas bien, por la interpretación de dicha información al momento de decidir cual sería la mejor acción a tomar en el campo.

La implementación de AP se sustenta principalmente en dos aspectos [1]:

1. Determinar las condiciones de variabilidad de las características más importantes del cultivo y suelo (textura del suelo, topografía, clima, entre otras) con la finalidad de establecer unidades de manejo de precisión que permitan monitorear los rendimientos; y,
2. Determinar cuándo y en qué medida es práctico y rentable aplicar el concepto de AP.

El primer aspecto implica, en un primer paso, establecer **unidades de manejo de precisión**: Es decir, divisiones o subdivisiones de la finca utilizando como criterio (el más recomendable -teóricamente-) el recurso suelo. El proceso comienza con la recolección de información sobre tipos de suelo (textura, estructura, topografía, pedregosidad, entre otros aspectos) con el objetivo de diseñar “unidades de manejo” que se agrupen en suelos homogéneos; es decir, de forma tal que una unidad no tenga dos tipos de suelo diferentes. Luego, mediante la utilización de unidades GPS se procede a referenciar espacialmente los límites de cada una de las unidades con la finalidad de calcular su

superficie y así obtener la información de producción en unidades de producto por área (producto por hectárea). De esta manera, se diseñan unidades de manejo por tipos de suelo que, hipotéticamente, se consideran uniformes en otras características como fertilidad, temperatura y humedad.

El siguiente paso consiste en monitorear la producción (cosecha) y ciertas características importantes del suelo (fertilidad) por cada una de estas unidades de manejo para detectar áreas problemáticas. El producto cosechado se identifica con un código para su unidad y en el lugar donde se lo recibe se determina su peso o volumen con la fecha y el código de la unidad de donde proviene. Finalmente, se diseñan los mapas de producción en un GIS, se investigan y se analizan posibles causas para bajas respuestas, se toman decisiones y se aplican los tratamientos correctivos. La aplicación variable de insumos (fertilizantes, pesticidas, entre otros) se realiza entre las unidades de manejo en lugar de entre una lectura del GPS a otra. Este sistema de posicionamiento geográfico es mucho menos preciso pero es más fácil de implementar y menos costoso.

El segundo aspecto a tratar implica analizar las realidades locales, como por ejemplo, la escala del proyecto (productores pequeños, medianos o grandes), solidez económica, nivel educativo del personal, nivel de tecnología utilizada, entre otras cosas, para decidir cómo y hasta cuándo es rentable la aplicación de la AP.

CAPÍTULO 2

2. ANÁLISIS ESTADÍSTICO EXPLORATORIO Y DESCRIPTIVO DE LA PRODUCCIÓN DE CAÑA DE AZÚCAR Y SUS FACTORES EN EL INGENIO SAN CARLOS

Tal como se expuso en el primer capítulo, ciertas tecnologías de AP no están al alcance de todos los productores que quieran adoptarlas y esto se debe principalmente a los altos costos que implica su adquisición e implementación. Sin embargo, los principios básicos del manejo sitio-específico de un cultivo no deben ser dejados a un lado debido a los grandes beneficios que conllevan. Existe la posibilidad de adoptar esta nueva filosofía de manejo simplemente adaptando sus principios elementales a la situación particular de cada productor y de cada cultivo.

Para el caso de un productor de caña de azúcar, quien con frecuencia se enfrenta a precios deprimidos de su producto y a una fuerte competencia, es de vital importancia optimizar la utilización de sus recursos económicos y naturales con el fin de aumentar la rentabilidad de su negocio. Por esta razón, en el Ingenio San Carlos (ISC), uno de los ingenios azucareros más importantes del Ecuador, se estudiaron y evaluaron algunas alternativas encaminadas a alcanzar estas prioridades decidiéndose al final por la implementación de la AP utilizando una combinación de tecnologías de información geográfica con criterios agronómicos básicos.

En primera instancia este capítulo expone las características más importantes de la implementación de la AP en el ISC, incluyéndose: 1) el diseño de unidades de manejo de precisión; 2) la adquisición de información más precisa sobre el cultivo; y, 3) la manera en la cual se determinó el potencial de aplicación de la AP. Luego, se realiza un análisis estadístico descriptivo de cada una de las variables que representan a la producción y a sus factores con el objetivo de determinar cuánto varían las características del suelo y cultivo en el ingenio.

2.1. CARACTERÍSTICAS DE LA IMPLEMENTACIÓN DE LA AGRICULTURA DE PRECISIÓN

El ISC se encuentra ubicado en la costa ecuatoriana en el cantón Coronel Marcelino Maridueña, provincia del Guayas, Ecuador (80.00° O y 01.81° S). Tiene como principal actividad económica el cultivo e industrialización de caña de azúcar, para lo cual, se manejan aproximadamente 16.500 ha. de esta gramínea como monocultivo produciéndose aproximadamente 1'500.000 toneladas de azúcar por año. El cultivo tiene un ciclo de 12 meses y la cosecha o zafra comienza a mediados del mes de junio para finalizar en los últimos días de diciembre.

El ingenio esta dividido con fines de manejo administrativo y agrícola en 6 sectores de aproximadamente 3.000 ha. cada uno, 14 secciones y 347 canteros, tal como se muestra en las figuras de los Anexos 1 y 2. La división por sectores es la siguiente:

- SECTOR 1: Parte de la sección 1 y 6, y la sección 5 completa
- SECTOR 2: Comprende la sección 2 y 20

- SECTOR 3: Sección 3, 4 y 17
- SECTOR 4: Sección 10 y 9
- SECTOR 5: Sección 8, 11, 18, 19
- SECTOR 6: Parte de la sección 1 y 6, y la sección 7.

Hasta finales del año 2002 el cantero fue considerado como la mínima unidad de manejo, es decir, un área dentro de la cual la aplicación de tratamientos agrícolas se realiza de manera uniforme bajo el supuesto de obtener una respuesta uniforme del cultivo. Sin embargo, se conoce y se puede observar que dentro de un cantero existe variabilidad en la producción debido, muy probablemente, a su gran extensión (se encuentran canteros entre 3 y 138 ha. con un promedio de 48 ha.) y al efecto de las condiciones variables en cuanto a textura de suelos, niveles de fertilidad, drenaje, variedad de caña, entre otros factores. Esta variabilidad se cuantificaba hasta cierto grado pero no se aprovechaba todo el potencial de esta información debido a que ya se encontraba establecida una mínima unidad de manejo: el cantero. Se utilizaban “totales” que no permitían diferenciar entre zonas de alta y baja producción y “promedios” para representar los niveles de ciertas características cuantificables de importancia (Ej: fertilidad) para

establecer diferentes tipos de tratamientos relacionados a dosis de fertilización, aplicación de herbicidas, cantidad y frecuencia de riegos, entre otros.

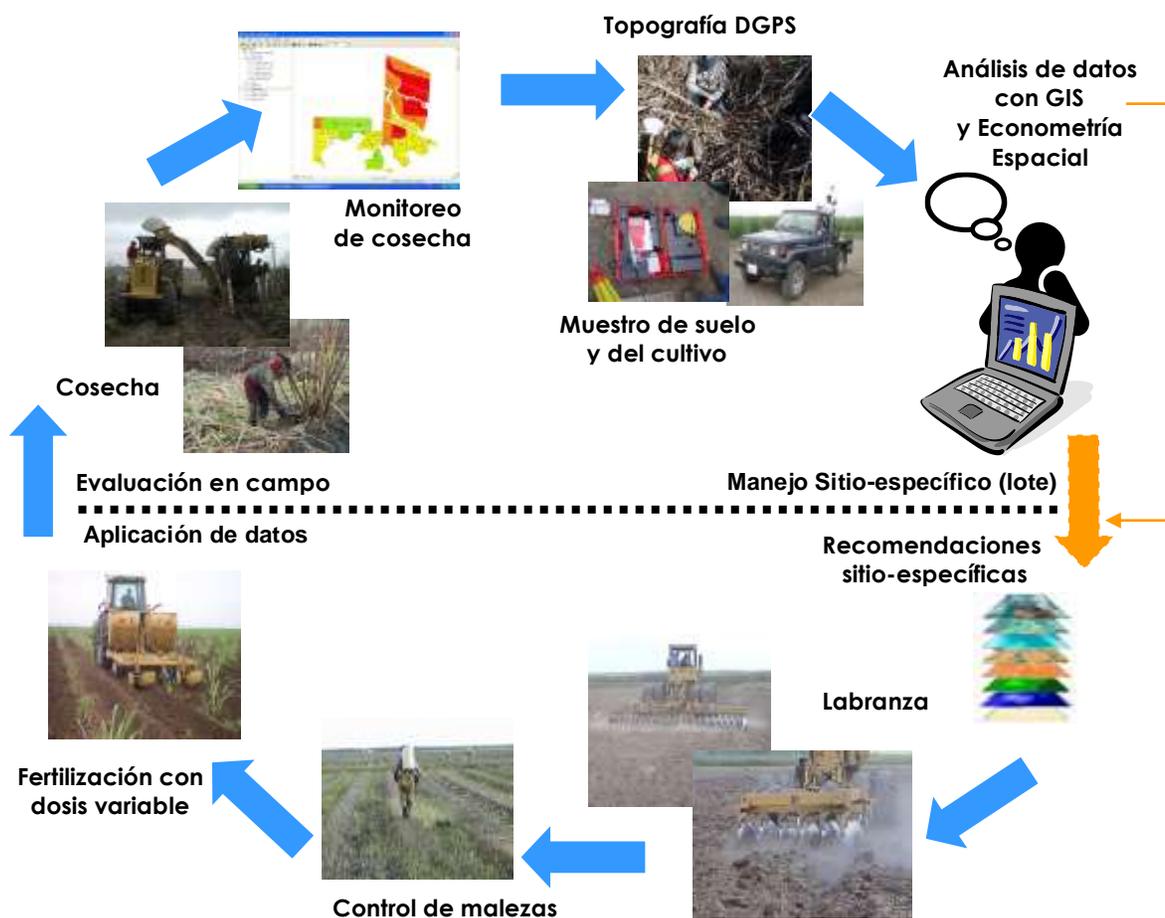


Figura 2.1. Adopción de los principios de la AP en el ISC

Elaboración: Xavier Vítors y Leonidas Estrada, 2004

Fuente: Dep. Campo, ISC

2.1.1. Lotización

Con la implementación de la AP en el ISC, se identificó la necesidad de disminuir el área de la mínima unidad de manejo existente; es decir, dividir los canteros en lotes. Lotizar, dentro del contexto de la AP, tiene como objetivo establecer unidades de manejo de menor extensión lo más homogéneas posible en cuanto a las condiciones presentes de los factores que inciden sobre la producción, para luego, recabar la mayor cantidad de información sobre el cultivo por cada una de estas nuevas unidades de manejo o lotes, de forma que, cuando se tome una decisión de manejo, esta tenga una mayor probabilidad de ser la correcta en comparación a la de un cantero.

El principal criterio utilizado para el establecimiento de los lotes fue dividir los canteros por tipos de suelo, para lo cual, se procedió a verificar la variabilidad de la textura del suelo de cada cantero apoyándose en un mapa de suelos. También se utilizaron los accidentes naturales (esteros y ríos) y la infraestructura existente (caminos, drenajes, canales de riego), los cuales generalmente son aprovechados para las diferentes operaciones de cosecha, riego y

otras labores. El objetivo fue crear unidades de manejo con un solo tipo de suelo.

Como resultado, se establecieron 1309 lotes con un promedio de 13 ha. (en un intervalo de 2 a 20 ha.). Muy pocos canteros quedaron con lotes mayores a 20 ha., los cuales en su siguiente renovación serán rediseñados para que sus divisiones tengan áreas menores. Los lotes se identifican con un código de seis dígitos y toda la información de cosecha (producción, camiones, zafreros, etc.) se recolecta y registra al nivel de lotes y no de canteros. La división por lotes del ISC se muestra en el Anexo 3.

Otro resultado importante fue que también se pudieron mostrar y cuantificar las condiciones de variabilidad en textura de suelos por cantero [16]. Se consideró homogéneo en tipo de suelo a un cantero si todos sus lotes presentaban la misma textura de suelo y, heterogéneo, si tenía por lo menos un lote de textura diferente. Se encontró que del total de los 347 canteros, el 45% (157) poseen suelos de textura heterogénea y el 55% (190) poseen suelos de textura homogénea. Los resultados se muestran en el Cuadro 2.1.

Cuadro 2.1. Canteros homogéneos y heterogéneos en textura de suelos

Canteros	Porcentaje	Número
Total	100%	347
Heterogéneos	45%	157
Homogéneos	55%	190

Elaboración: X. Vítóres, 2004

Fuente: L. Estrada [16]

Evidentemente, la variabilidad observada en textura de suelos sugiere una reingeniería de las unidades de manejo debido, principalmente, a que se esperan valores diferentes de la producción por cada textura específica.

2.1.2. Estructura y monitoreo de la producción

Para el análisis y comparación de la producción de caña de azúcar entre unidades de manejo, ésta debe ser expresada como producto por unidad de área para cada lote; por lo que, en primer lugar, se hizo necesario medir de forma precisa la superficie de cada unidad de manejo. Para llevar a cabo esta tarea se estudiaron varias

alternativas (incluida la utilización de instrumental óptico) y al final se optó por la utilización de la navegación satelital (DGPS) y un GIS. Con estas tecnologías se requirió de pocas personas para las operaciones, el posicionamiento geográfico de cada punto es independiente de los demás (lo que evita la acumulación de errores) y se pudo obtener información sobre altitud, longitud y latitud con un error aproximado de 5 cm.

Una vez reconocidos los límites de cada unidad de manejo de acuerdo a los criterios anteriormente establecidos (textura, infraestructura y accidentes naturales), se procedió a referenciarlos espacialmente mediante unidades GPS. Las tareas diarias se programaron de tal manera que la obtención de información se realizó en unidades recientemente cosechadas y mientras existía cobertura satelital. Al final de cada día de trabajo, se ingresaba la información en una computadora y con el software adecuado se procesaban datos sobre altitud, latitud y longitud de los límites de cada lote que luego eran importados al GIS para representarlos gráficamente mediante puntos en un mapa digitalizado. Estos puntos se unieron para crear las nuevas unidades de manejo y calcular su área exacta. Como

resultado, en aproximadamente 8 meses, se diseñó el mapa de las 16,500 hectáreas del ingenio divididas en lotes midiendo cerca de 70,000 puntos en campo.

Luego de cosechar un lote en el campo, los camiones con el producto llegan a la fábrica del ingenio donde se determina la cantidad de toneladas métricas de caña que transportan y se identifica a la unidad de manejo de donde provienen (código del lote). Esta información se relaciona con el área ya calculada y se cuantifica el producto en términos de toneladas métricas de caña por hectárea (TMC/ha) para cada lote. Luego, los datos sobre producción se representan gráficamente en un GIS mediante mapas digitalizados que permiten ubicar espacialmente las respuestas altas y bajas del cultivo así como los lotes de donde provienen.

Por medio estos mapas pudo visualizarse la variabilidad de la producción dentro de cada cantero [16] y, además, con la producción cuantificada por lote, se calculó la diferencia entre el valor máximo y el mínimo de la producción dentro de cada cantero; luego, se determinó la distribución de frecuencias de estos rangos y pudo observarse que

en una muestra de 231 canteros (1227 lotes) el 80% presentaban diferencias mayores a 10 TMC/ha y el otro 20% presentaba diferencias menores o iguales a 10 TMC/ha. Los resultados se muestran tanto en el Cuadro 2.2 como en la Figura 2.2.

Cuadro 2.2. Diferencia entre el máximo y mínimo de la producción por porcentaje de canteros

Rango en TMC/ha	% Canteros
5	6
10	14
20	36
30	24
40	13
50	7

Elaboración: X. Vítores, 2004

Fuente: L. Estrada [16]

Se considera a un cantero con una diferencia de 10 TMC/ha entre el valor máximo y el valor mínimo de producción como homogéneo en producción. Partiendo de este supuesto y del análisis realizado, se encontró que solo el 20% de los canteros presenta esta condición. Este hecho, junto con la variabilidad encontrada en textura de suelos por cantero, es una clara evidencia que justifica la división de los

canteros a lotes; y, lo más importante, expone una forma para determinar el potencial de aplicación de la AP en el ISC.

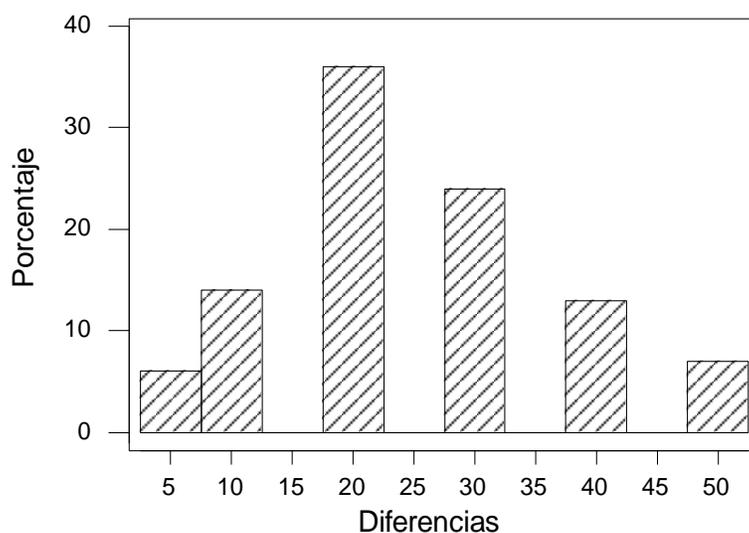


Figura 2.2. Histograma de las diferencias en producción en TMC/ha

2.1.3. Descripción y monitoreo de los factores de producción

La capacidad de producir caña y sacarosa está determinada en gran medida por las características genéticas de la variedad cultivada; sin embargo, también es necesario comprender que estos dos aspectos

son el resultado final de la interacción entre la planta, el clima, el suelo y el manejo que se le proporcione al cultivo. Por lo tanto, es normal que existan variaciones en la productividad entre zonas, de acuerdo con las condiciones climáticas prevalecientes (temperatura, humedad, radiación solar, precipitaciones, viento, entre otras), así como variaciones en las características edáficas (por ejemplo: fertilidad, textura, estructura, pH).

En esta sección, se describen y definen los principales factores naturales y de manejo que inciden sobre la producción de caña de azúcar en el ISC. El monitoreo de cada uno de ellos y su cuantificación por unidad de manejo es ya desde hace bastante tiempo una política de información dentro del ingenio. Algunos datos recopilados se encuentran a un nivel muy alto de agregación, como en el caso de los niveles de temperatura y luminosidad, y otros a nivel moderado, como el número de riegos y fertilización. En el primer caso, no es posible obtener información precisa al nivel de lotes debido a que esta proviene únicamente de una estación meteorológica central llamada "Casa Blanca" y otras 14 estaciones distribuidas a lo largo y ancho del ingenio. En el segundo caso, la información se encuentra

agregada a nivel de canteros, situación que cambia a partir del año 2004 con el manejo por lotes dentro de la implementación de la AP. La información de cosecha o zafra 2003 que se puede encontrar a distintos niveles de agregación, además de la producción, se muestra en el Cuadro 2.3.

Cuadro 2.3. Agregación de información sobre los factores de producción

Nivel de agregación de datos	Unidad de agregación	Factores de producción (naturales y de manejo)	Zafra (cosecha)
Alto	E.M.*	Temperatura	2003
Alto	E.M.	Luminosidad	2003
Alto	E.M	Precipitación	2003
Precisión	Lote	Características edáficas	2003
Precisión	Lote	Edad cultivo	2003
Precisión	Lote	Edad a la cosecha	2003
Moderado	Cantero	Fertilización (N,P y K)	2003 **
Moderado	Cantero	Control malezas	2003 **
Moderado	Cantero	Número y frecuencia de riegos	2003 **
Moderado	Cantero	Control de plagas	2003 **
Precisión	Lote	Variedad	2003

(*) Estación meteorológica

(**) Se dispondrá de información por lotes para la zafra 2004

Elaboración: Xavier Vítóres, 2004

2.1.3.1. Factores naturales

2.1.3.1.1. Edad del cultivo

El ciclo del cultivo de caña de azúcar en el ISC es de 12 meses. A partir de la plantación de la semilla o cepa hasta la primera cosecha se denomina al cultivo “caña planta” y corresponde al primer año de edad. Luego, de la cepa emergen retoños o “socas” de los cuales crecerán nuevos tallos (macollamiento) que al momento de ser cosechados serán denominados “caña soca”. La caña de dos a seis años de edad es llamada “soca nueva” y la caña de siete años en adelante “soca vieja”. Se esperan mayores producciones de la caña planta y un decrecimiento a medida que la edad aumenta.

2.1.3.1.2. Edad a la cosecha

La edad de cosecha corresponde al mes en el cual se cosecha la caña de azúcar. Se espera que a medida que se cosecha después de

los 12 meses (rezagar caña) la producción aumenta; sin embargo, las condiciones climáticas inciden significativamente sobre el resultado.

2.1.3.1.3. Tercio de cultivo

El año agrícola en el ISC corresponde a un periodo de 18 meses. El inicio del crecimiento del cultivo coincide con la cosecha o “zafra” del cultivo anterior y se da a mediados del mes de junio para terminar con la llegada del invierno a finales de diciembre o a mediados de enero si las condiciones climáticas así lo permiten. Esto se debe a la facilidad que ofrece el verano para la utilización de maquinaria en el campo. Al llegar el invierno, se deja al cultivo desarrollarse por aproximadamente seis meses para luego iniciar la cosecha (junio-julio) que culmina en otros seis meses más (diciembre-enero).

El periodo durante el cual cultivo el cultivo se desarrolla esta dividido en tercios. El primer tercio corresponde a los meses de julio y agosto (considerándose algunas veces la última semana de junio), el segundo tercio corresponde a los meses de septiembre y octubre y el

tercero a los meses de noviembre y diciembre (algunas veces las dos primeras semanas de enero). Se ha observado a través de los años que las mejores producciones se obtienen en el primero y segundo tercios.

2.1.3.1.4. Requerimientos climáticos

Para que la caña de azúcar complete satisfactoriamente sus distintas fases de crecimiento: germinación (1 mes), macollamiento y cierre (dos meses), rápido crecimiento (7 meses) y maduración (dos meses), debe contar con condiciones climáticas favorables. Los principales factores que interfieren son:

a) Temperatura

La temperatura junto con la humedad son dos de los factores que más relevancia tienen en el proceso de germinación y desarrollo. La temperatura óptima para la germinación de las yemas y el desarrollo

del cultivo se ubica ente los 27°C y 33°C. A valores de 20°C el crecimiento disminuye notoriamente; y, si la temperatura disminuye más, el crecimiento prácticamente se paraliza. Cuando la temperatura es mayor a los 35°C, aumenta la respiración y disminuye la tasa fotosintética, lo que ocasiona una reducción en el crecimiento, y por lo tanto, una menor acumulación de materia seca [17].

El ISC cuenta con una estación metereológica central llamada “Casa Blanca” para cuantificar niveles de temperatura y luminosidad. Las condiciones generales que se presentaron para la zafra 2003 fueron una temperatura media de 25,28°C con un máximo de 28,98°C y un mínimo de 21,54°C. Las condiciones específicas de temperatura se muestran en la Figura 2.3.

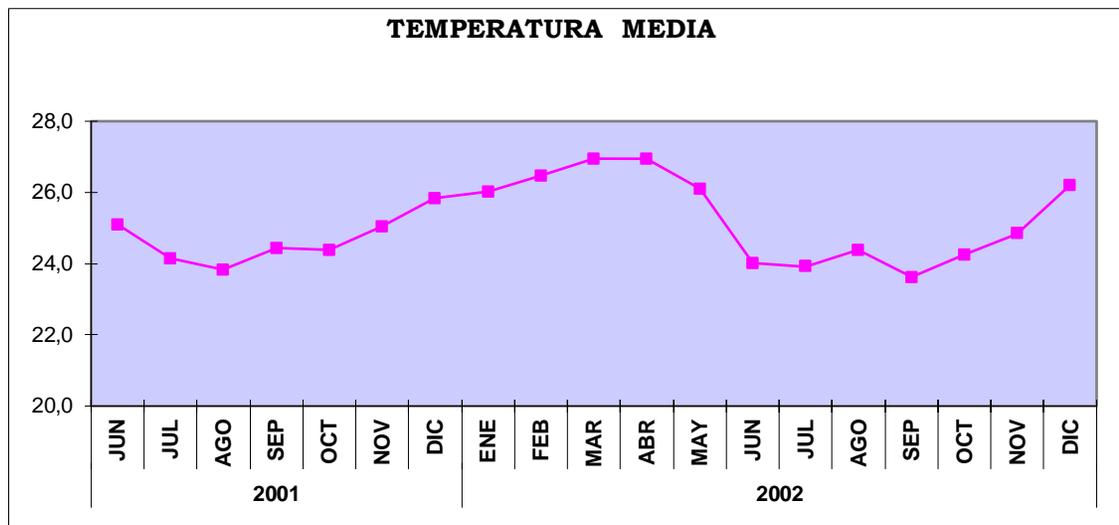


Figura 2.3. Condiciones específicas de temperatura media para la zafra 2003

b) Precipitaciones

La caña de azúcar puede prosperar en una amplia diversidad de zonas con diferentes regímenes de lluvias. El suministro de agua, ya sea mediante forma natural o mediante riego, es necesario durante todo el período de crecimiento. En promedio, se requieren de 1.200 a 1.500 mm anuales, distribuidos de la mejor manera posible durante el periodo vegetativo, de tal forma que durante esta etapa siempre se

disponga de suficiente agua. La demanda aumenta en relación con el crecimiento de la planta, debido a que la transpiración se incrementa [17].

Tanto la sequía o falta de humedad, como los excesos son perjudiciales para el cultivo. Las plantaciones que se manejan en condiciones de secano (sin riego), por lo general tienen un rendimiento pobre, el cual es aún más, si durante el ciclo, en la época de invierno, ocurren períodos de sequía.

El ISC cuenta con 14 estaciones meteorológicas para medir precipitaciones. La precipitación anual promedio varía entre 1.300 a 1.600 mm en condiciones normales y entre 800 a 900 mm para años secos. La humedad anual promedio es del 80%. Para el período correspondiente a la zafra 2003, las precipitaciones fueron de 892 mm, es decir, fue un período seco donde se requería mayor suministro de agua por riego. Las condiciones específicas de precipitaciones para la zafra 2003 se muestran en la Figura 2.4.

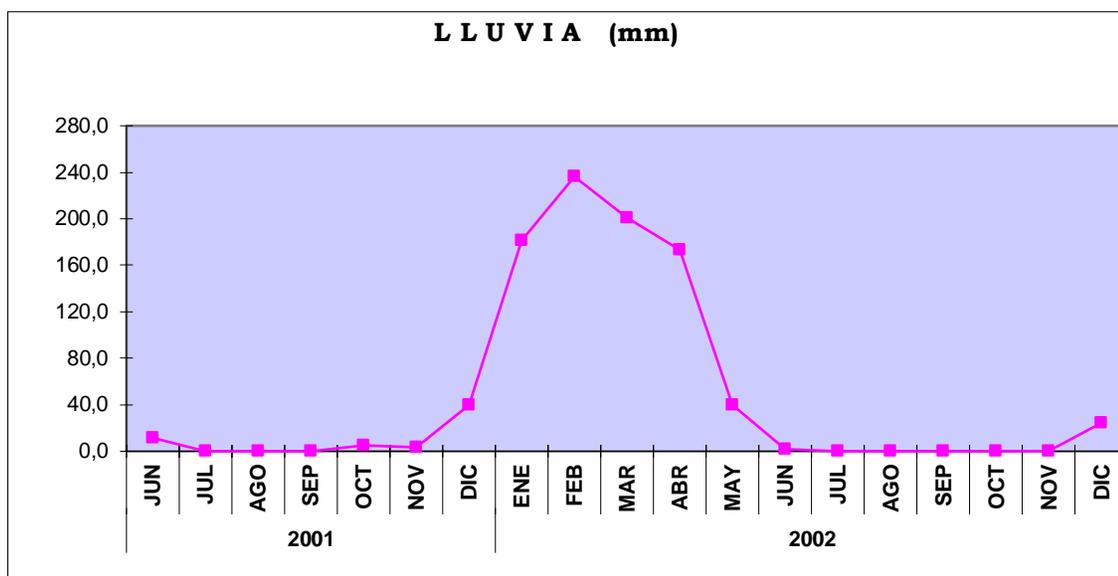


Figura 2.4. Condiciones específicas de precipitaciones para la zafra 2003

c) Luminosidad

La radiación solar es la principal fuente de energía de las plantas. Estas utilizan determinadas longitudes de onda (entre 400 y 700 nm¹, que corresponden al ámbito de radiación fotosintéticamente activa); de esta manera se lleva a cabo el proceso de la fotosíntesis y otras reacciones metabólicas. La caña de azúcar pertenece al grupo de plantas que posee un sistema fotosintético C₄, capaz de fijar de

¹ 1 nm = 1x10⁻⁹m

manera más eficiente. Existen, sin embargo, variedades con mayor capacidad que otras en su eficiencia fotosintética [17].

Cuanta mayor radiación exista, mayor será la eficiencia fotosintética, aspecto muy relacionado con la producción y acumulación de carbohidratos. Lo ideal es que, durante todo el ciclo, la planta disponga de buena luminosidad.

Por lo general, las plantas expuestas a pleno sol producen tallos cortos y gruesos, hojas de mayor longitud y de coloración verde más intensa; por el contrario, bajo condiciones de poca luminosidad, los tallos son largos y delgados, las hojas son cortas y angostas y de una tonalidad amarillenta.

En el ISC, el promedio anual de luminosidad es de 734 horas luz (heliofanía). Durante el período correspondiente a la zafra 2003 el total de horas de sol fue de aproximadamente 629 horas, es decir, por debajo del promedio. Las condiciones específicas de luminosidad para la zafra 2003 se muestran en la Figura 2.5.

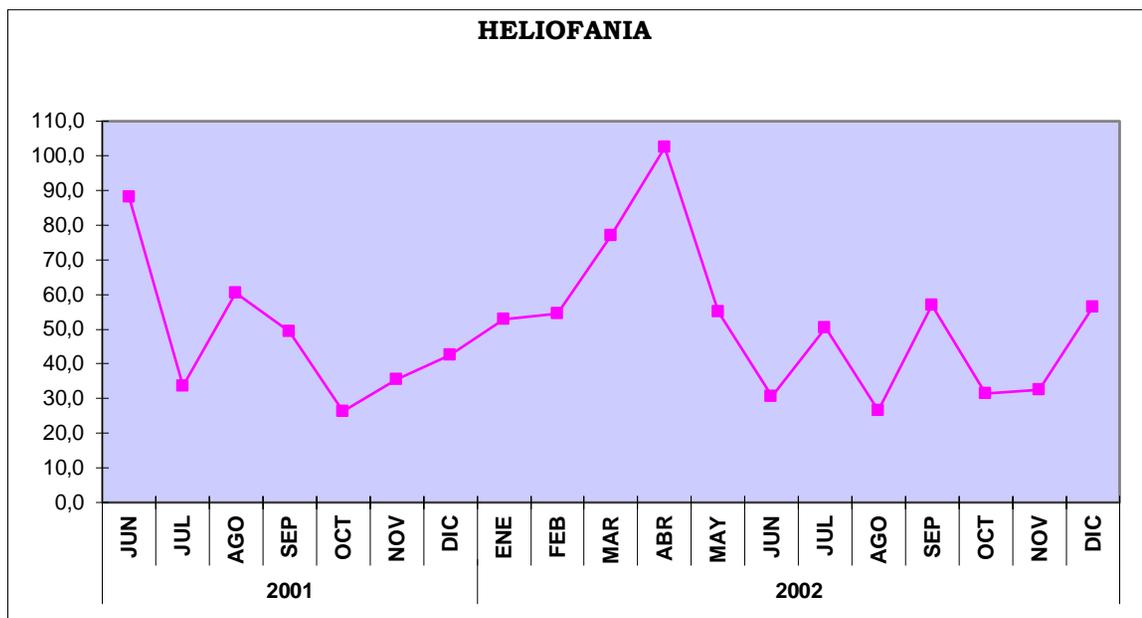


Figura 2.5. Condiciones específicas de luminosidad para la zafra 2003

2.1.3.1.5. Requerimientos edáficos

El cultivo de la caña de azúcar puede desarrollarse en una amplia diversidad de suelos: desde los arcillosos pesados hasta los completamente orgánicos; sin embargo, unos presentan mejores características que otros para que se desarrolle favorablemente la caña.

Es aconsejable que la textura del suelo sea franco arcillosa, franco arenosa o limosa, con buena estructura y capacidad de retención de humedad, pero a la vez friable, con un horizonte profundo, sin problemas de drenaje y salinidad, características que deben permanecer en al menos los primeros 50 cm del suelo, donde se localiza la mayor cantidad de raíces. El nivel freático debe estar ubicado preferentemente en profundidades mayores de 75 cm [17].

Los suelos arcillosos pesados son una barrera para la penetración de raíces y en ocasiones presentan problemas de aeración. Son más serias las limitaciones físicas que las nutricionales.

Es preferible que los suelos sean fértiles, que no muestren problemas serios de deficiencias o desbalances minerales; sin embargo, si presentan niveles de fertilidad bajos, pero las características físicas son buenas, los inconvenientes pueden solucionarse adicionando los nutrimentos requeridos. El pH debe oscilar entre 5,5 y 8,0.

a) Textura de suelos

No existe una definición universal para la textura del suelo. Inicialmente fue utilizada para describir el grado de dificultad que presentaba el suelo para ser laborado; de allí surgieron las apreciaciones de suelos livianos y pesados. Actualmente, el concepto de textura es usado como una medida de la proporción de partículas de arena, limo y arcilla que conforman un suelo [18].

Algunas de las características más importantes de los suelos se basan en su textura, particularmente, en lo relacionado a su comportamiento con respecto al suministro de agua a las cosechas y su posibilidad de cultivo. Los suelos con un predominio de las partículas más finas o arcillas son difíciles de trabajar debido a su resistencia al movimiento de agua y aire. Su gran capacidad para absorber agua hace que se vuelvan plásticos o moldeables cuando están húmedos y duros cuando se secan. Los suelos con un predominio de las partículas más gruesas o arena no retienen el agua y están secos persistentemente de forma que sólo en la estación húmeda mantienen a la cosecha abastecida con suficiente de agua

mientras que en el resto de la temporada se requiere de condiciones especiales de riego que por lo general resultan costosas. Un suelo fértil debe poseer una textura uniforme y adaptada a satisfacer las necesidades óptimas de la planta brindando un equilibrio entre los requerimientos de aire y agua.

El análisis mecánico permite conocer la proporción porcentual de partículas de arena, limo y arcilla presentes en una muestra de suelo. El nombre textural se puede obtener del triángulo de texturas desarrollado por el Departamento de Agricultura de los Estados Unidos mostrado en la Figura 2.6. Las líneas de separación entre las diferentes clases texturales fueron seleccionadas de tal manera que coincidieran con el grado de dificultad para el laboreo del suelo.

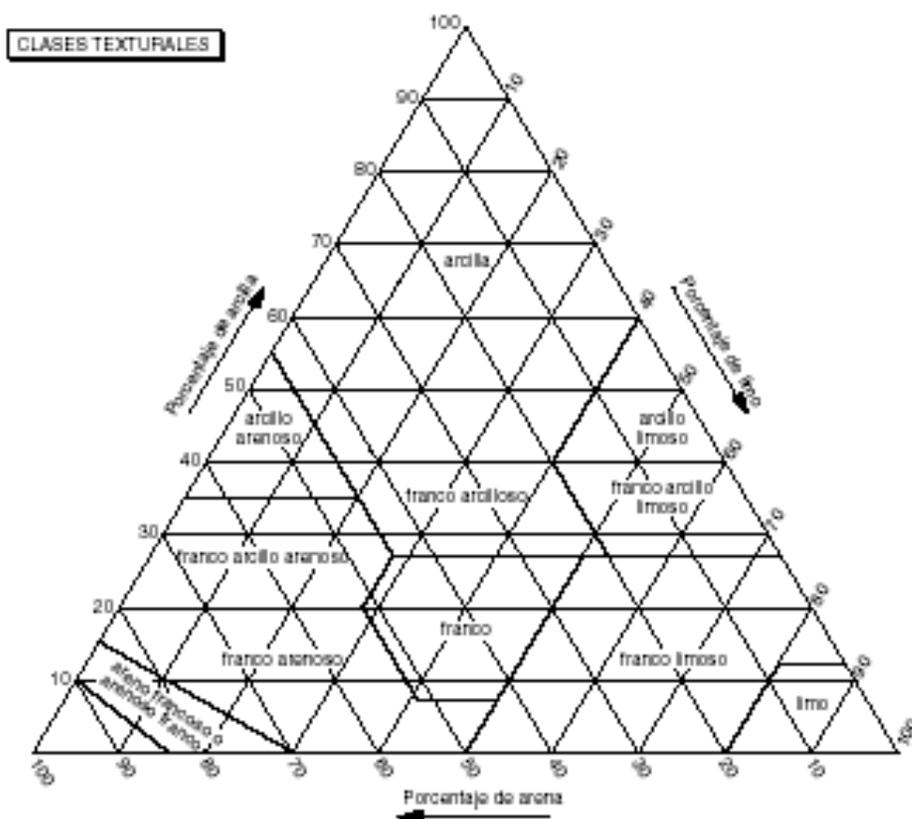


Figura 2.6. Diagrama de la determinación del nombre textural a partir del análisis mecánico del suelo

b) Niveles de Fertilidad: Nutrientos esenciales

La caña de azúcar, al igual que el resto de las plantas, requiere 16 elementos esenciales para su adecuada nutrición. Un elemento esencial es aquel cuya ausencia causa la alteración en el desarrollo

normal de la planta, según el papel que cumpla a nivel fisiológico, y altera en mayor o menor grado el rendimiento (industrial o agrícola).

Una parte de los elementos son generalmente adquiridos por la planta de la atmósfera y del agua; estos son el carbono (C), el hidrógeno (H), y el oxígeno (O). Los otros nutrientes, clasificados en macronutrientes y micronutrientes [17], los aporta el suelo si están disponibles; de lo contrario, deben agregarse como fertilizantes para que lo absorban las raíces o por vía foliar.

Macronutrientes

- **Nitrógeno (N)**

Es uno de los elementos más importantes en el desarrollo de la planta, debido a que forma parte de la estructura molecular de las proteínas, de la clorofila y de los ácidos nucleicos (ADN y ARN).

El N se encuentra en la fracción orgánica del suelo; es absorbido principalmente como nitrato (NO_3^-) y en menor proporción como

amonio (NO_4^+). Una vez que las raíces lo absorben, este se desplaza hacia las hojas y puntos de crecimiento. En esos lugares, mediante la acción de enzimas y carbohidratos, producto de la fotosíntesis, son reducidos a amoníaco, transformados en aminoácidos y en sustancias proteicas, que se utilizan en la formación de protoplasma.

La insuficiencia de N causa la mala germinación de los esquejes durante la siembra; la cantidad de brotes son menores, los tallos se tornan delgados, generalmente se reduce el crecimiento de la planta y se produce la maduración prematura del tallo (acumulación temprana de azúcares). Aplicaciones excesivas de N, sobre todo cuando se hace en forma tardía, afectan negativamente la calidad del jugo, reducen su pureza y aumentan los azúcares reductores (glucosa y fructuosa), lo que produce reducciones significativas en el contenido de sacarosa de la caña.

El N y el potasio (K) están muy correlacionados; de tal manera que cuando existe deficiencia de uno, la respuesta del otro se ve limitada, por lo que es importante el equilibrio de ambos. La absorción de ellos es lenta durante las primeras etapas del crecimiento, aumenta entre

los 5 y 6 meses de edad y posteriormente disminuye. El exceso de N provoca un desbalance con el K; este último en ocasiones debe agregarse para corregir el inconveniente.

- **Fósforo (P)**

El P es componente de los ácidos nucleicos, fosfolípidos, así como de las membranas celulares. Cumple una función importante en el metabolismo energético, debido a que es parte constituyente de la molécula ATP (adenosín trifosfato), la cual desempeña un papel muy relevante en las reacciones de transferencia de energía en la célula.

La deficiencia de P se manifiesta en el poco desarrollo de la planta: los entrenudos se acortan, la longitud y diámetro de los tallos se reduce; disminuye el número de tallos de las cepas porque el macollamiento es escaso, lo que origina menor producción de tallos secundarios y terciarios. En situaciones severas, los extremos y los márgenes de las hojas más viejas se secan. El sistema radical se desarrolla de manera raquítica: la formación de raíces secundarias es limitada, de aspecto anormal y de un color marrón, situación que impide la absorción de agua y sales minerales.

El P se absorbe de la fracción inorgánica del suelo en forma de ión fosfato monovalente o divalente (H_2PO_2^- y HPO_4^{2-}), con mayor intensidad en los primeros meses de crecimiento, en especial durante las fases de formación de raíces y macollamiento. También contribuye acelerando la maduración de la caña en la última fase del ciclo.

La disponibilidad de P puede quedar restringida en suelos con reacción alcalina o ácida. En el primer caso por la formación de fosfatos cálcicos insolubles y, en el segundo caso, por la presencia de hierro (Fe) y aluminio (Al) solubles. Estos conducen a la formación de fosfatos de hierro y aluminio insolubles. Por otro lado, niveles excesivos de P en el suelo pueden provocar deficiencias de otros elementos, tales como Fe y zinc (Zn).

- **Potasio (K)**

El K está relacionado con procesos muy importantes como la fotosíntesis, respiración, formación de clorofila, metabolismo de carbohidratos, formación de almidón y la activación de enzimas necesarias en la síntesis de proteínas. Además, tiene una participación activa en la permeabilidad de las membranas celulares,

el transporte de azúcares por el floema, en la absorción de iones y en la entrada de agua y transpiración de la planta.

La deficiencia de este elemento se manifiesta por el crecimiento lento de los tallos; estos se tornan delgados y débiles. Además, se ha observado mayor susceptibilidad al ataque de patógenos cuando existe deficiencia de este elemento, así como una reducción en el contenido de azúcares.

Es importante mencionar que el K, el magnesio (Mg) y el calcio (Ca) son las bases de cambio en el complejo coloidal y a la vez antagónicos; por tal razón, si existen desbalances grandes en el suelo, la absorción de uno o dos de ellos puede verse afectada.

- **Calcio (Ca)**

El calcio forma parte de las láminas medias constituidas de pectatos de calcio, que unen entre sí las paredes celulares. Además, desempeña un papel importante en la formación de membranas celulares y en la división celular. La carencia de Ca provoca el deterioro, e inclusive la muerte de la planta.

- **Magnesio (Mg)**

Forma parte de la molécula de clorofila en la que ocupa la posición central; por eso tiene una relación directa con el proceso de fotosíntesis.

Se supone que junto al N interviene en la conversión de azúcares reductores (glucosa y fructuosa) a sacarosa. Su deficiencia provoca tallos de menor diámetro y los entrenudos se acortan. Estudios realizados en Florida (Estados Unidos), indican que la deficiencia de Mg podría tener un papel importante en la incidencia de la roya, una de las principales enfermedades de la caña de azúcar.

Este elemento se caracteriza por ser móvil dentro de la planta. En el suelo puede perderse por lixiviación con alguna facilidad. Se ha determinado que el Mg contribuye a la absorción de P.

- **Azufre (S)**

Forma parte estructural de las proteínas y constituye los aminoácidos cistina, cisteína y metionina. Además, forma parte de algunas

vitaminas como la tiamina y biotina, de la coenzima A y de varias enzimas.

El azufre se absorbe como ión sulfato ($\text{SO}_4^{=}$) y es relativamente inmóvil en la planta. La carencia de azufre se manifiesta cuando las plantas muestran poco vigor y el grosor de los tallos se reduce.

Micronutrientes

- **Zinc (Zn)**

El zinc participa en la síntesis del ácido indol acético (AIA); es activador de una serie de enzimas y actúa en la síntesis de proteínas.

La deficiencia de este elemento produce el acortamiento de los entrenudos, reducción en el grosor de la lámina y mal macollamiento. Este nutriente presenta cierta movilidad en la planta. Es deficiente en suelos con pH alto y en los de textura arenosa.

- **Cobre (Cu)**

El cobre se encuentra en concentraciones bajas en la caña; sin embargo, desempeña funciones muy importantes. Forma parte de varias enzimas y es activador de otras reacciones. Este micronutriente se absorbe como Cu^{+2} y es inmóvil dentro de la planta.

La deficiencia de este elemento produce marchitamiento en plantas desarrolladas y, en general, el desarrollo de la cepa es raquítrico.

- **Manganeso (Mn)**

La principal función del Mn es ser agente activador enzimático en varios procesos: respiración, metabolismo de N, activador de las reductasas; interviene en el ciclo de Krebs (respiración en la mitocondria) y ciclo de Calvin (fotosíntesis en el cloroplasto), en la formación de vitamina C (ácido ascórbico). Se absorbe como Mn^{+2} y presenta cierta movilidad en la planta. Algunos suelos cuyo pH es superior a 7,5 pueden ser deficientes en este micronutriente.

- **Boro (B)**

Cumple una función muy importante en el transporte de azúcares. Se piensa que actúa en la diferenciación celular y desarrollo celular, en el metabolismo del N, y del P y en la fotosíntesis; no obstante, no se ha esclarecido completamente su papel. El B es adquirido por la planta en forma iónica de distintas maneras (BO_3^{-3} , HBO_3^{-2} , entre otras) y no es móvil dentro de la planta. Su deficiencia retarda el crecimiento de la planta y los tallos muestran un diámetro reducido.

- **Molibdeno (Mo)**

Está muy relacionado con el metabolismo del N; también con el P. Forma parte del sistema de asimilación de nitratos. Su deficiencia ocasiona que las hojas viejas se sequen prematuramente y mueran de la mitad hacia el extremo distal. Es un elemento móvil dentro de la planta y se absorbe como ión MoO_4^{-2} y HMoO_4^{-} .

- **Hierro (Fe)**

El hierro interviene en la síntesis de proteínas del cloroplasto, a la vez que juega un papel importante en la síntesis de la clorofila.

La deficiencia de Fe se asocia con altos contenidos de Mn. Se absorbe como Fe^{+3} y es inmóvil dentro de la planta.

- **Cloro (Cl)**

Su principal función es estimular la fase lumínica de la fotosíntesis. Su deficiencia produce el marchitamiento de las hojas jóvenes recién formadas. Es un elemento móvil dentro de la planta; se absorbe como ión cloruro Cl^{-1} .

c) Potencial hidrógeno (pH)

El factor pH se determina mediante el logaritmo cambiado de signo de la concentración de iones hidrógeno en los suelos. La caña de azúcar tolera una gran variabilidad en cuanto a la reacción del suelo o pH se refiere; sin embargo, se considera óptimo el valor de pH de 6,5 que equivale a un grado muy ligero de acidez [19].

En general, se considera a los suelos con valores pH inferiores a 5,5 como fuertemente ácidos; en ellos se puede restringir el desarrollo

normal de la caña de azúcar, principalmente debido a la presencia de Aluminio (Al) intercambiable, Fe y Mn en niveles tan altos que pueden causar toxicidades en el cultivo e impedir la absorción del P del suelo. En estos suelos muy ácidos se restringen las disponibilidades del Ca, el Mg y Mo.

Suelos con valores de pH entre 6,0 y 6,5 son considerados como ligeramente ácidos. Esta reacción del suelo es bastante adecuada para la caña de azúcar.

Cuando el pH varía entre 6,6 y 7,3 se considera que el suelo es casi neutro. Aunque es una reacción adecuada para la caña de azúcar, las disponibilidades del P y de los elementos menores o micronutrientes, excepto el Mo, se restringen un poco, especialmente la disponibilidad de estos últimos.

Suelos con valores de pH entre 7,3 y 8,0 se consideran alcalinos. Las disponibilidades de P y micronutrientes, excepto el Mo, son bajas.

Valores de pH superiores a 8,0 corresponden a suelos muy alcalinos, los cuales se caracterizan por la presencia de altos contenidos de sodio intercambiable y condiciones físicas inadecuadas. Estos suelos requieren prácticas de recuperación que incluyen la aplicación de enmiendas, el mejoramiento del drenaje y el lavado.

d) Materia orgánica

La materia orgánica del suelo es fuente importante de N, P, S y de algunos elementos menores o micronutrientes. Tiene marcados efectos positivos en las propiedades físicas del suelo.

Las experimentaciones sobre fertilización con N en caña de azúcar, realizadas por el Centro de Investigación de la Caña de Azúcar de Colombia (CENICAÑA) en colaboración con la mayoría de los ingenios de Colombia, han permitido establecer tres categorías para el contenido de materia orgánica del suelo. Estas se presentan en el Cuadro 2.4.

Cuadro 2.4. Categorías para la Materia orgánica

Categoría	Materia orgánica (%)
Baja	Menor que 2
Media	2 – 4
Alta	Mayor que 4

Elaboración: X. Víttores, 2004
Fuente: CENICAÑA [19]

Las categorías baja, mediana y alta están estrechamente relacionadas con las probabilidades de obtener respuestas altas, media o baja a las aplicaciones de un determinado nutrimento.

e) Conductividad Eléctrica

La medida de la conductividad eléctrica de los extractos obtenidos del suelo permite establecer una estimación aproximadamente cuantitativa de las cantidad de sales que obtiene. A valores altos de conductividad eléctrica más salinidad.

2.1.3.1.6. Análisis de suelos

Cuando no existe ninguna información relacionada con la nutrición del cultivo en una zona, lo aconsejable es tomar muestras de suelo y enviarlas al laboratorio para su análisis. Los resultados permitirán tener idea de la capacidad que tiene el suelo para suplir los nutrimentos. Los elementos se extraen mediante sustancias extractoras que tratan de simular la manera como las raíces llevan a cabo la toma de los nutrimentos. La determinación de los niveles se lleva a cabo utilizando distintos métodos, tales como espectrofotometría, colrimetría, titulación y turbidimetría.

Los valores que se obtienen pueden compararse con tablas de niveles críticos que se han diseñado mediante correlaciones entre los niveles presentes en los suelos y la respuesta de las plantas. El Ministerio de Agricultura y Ganadería de Costa Rica (MAG) ha confeccionado una guía de interpretación de análisis de suelos que se muestra a continuación en el Cuadro 2.5.

Cuadro 2.5. Guía general para la interpretación del análisis de suelos

Categoría Característica	Baja	Media	Alta
pH H ₂ O 1:2.5	<5,5	5,5-6,5	>6,5
cmol (+)/L			
Ca	<4,0	4,0-12,0	>12,0
Mg	<1,0	1,0-5,0	>5,0
K	<0,20	0,20-0,60	>0,60
ppm			
P	<10	10-20	>20
Relaciones entre elementos	Desbalance	Balance	Desbalance
Ca/Mg	<2	2-5	>5
Mg/K	<2,5	2,5-15	>15
Ca/K	<5	5-25	>25

Elaboración: Xavier Vítóres, 2004

Fuente: MAG [17]

En el ISC, tanto en los suelos nuevos como en los lotes que se han cosechado, se realiza un muestreo de suelo en forma de zigzag. Este muestreo se lo realiza con barrenos a profundidades de 30 y 60 cm por punto. Las muestras se mezclan para al final obtener sólo dos, las cuales, en el laboratorio del ingenio, pasan por un proceso de secamiento para realizar el análisis mecánico de la textura del suelo (que servirá para realizar un mapeo de suelos), pH, materia orgánica y macronutrientes como N, P, K, Ca y Mg, que servirá para las recomendaciones de fertilización en cada lote.

2.1.3.2. Factores de manejo

2.1.3.2.1. Sector

Es muy frecuente observar en una finca que para ciertas zonas geográficas dentro del cultivo creadas con fines administrativos se presentan rendimientos diferenciados. Esto puede deberse a ciertas condiciones naturales, de manejo u otro tipo de factores conocidos o desconocidos que inciden directa o indirectamente sobre la producción y que dan origen cierta heterogeneidad en el cultivo relacionada directamente con una ubicación en el espacio.

En los últimos 8 años se ha notado que, entre los seis sectores en los que está dividido el ISC, el sector 3 presenta valores altos en producción. Dentro de las posibles causas pueden considerarse ciertos factores climáticos locales (como por ejemplo una mayor luminosidad en el sector) y un manejo diferenciado de este sector con respecto a los demás. Esta última situación hace que los sectores deban ser considerados como un factor de manejo que incide sobre la producción.

2.1.3.2.2. Variedad de la caña

Las variedades existentes de caña de azúcar son el resultado de un continuo mejoramiento genético cuyo objetivo consiste en obtener materiales de mayor rendimiento en caña y en sacarosa por unidad de área. Este procedimiento es, sin duda, uno de los principales aspectos que deben considerarse para aumentar el rendimiento, debido a que poco puede lograrse, aunque se disponga de condiciones agroecológicas favorables, cuando el potencial productivo de una determinada variedad es bajo [17].

Entre las principales características deseables de una variedad se pueden citar:

- a) Elevado potencial agronómico (alta producción de caña, buena producción de tallos molederos, altos y gruesos).
- b) Buenas características industriales (alta polarización y pureza del jugo y bajo contenido de fibra)
- c) Buena adaptación a las condiciones agroecológicas de un lugar determinado.

- d) Resistencia a plagas y enfermedades.
- e) Las “socas” o retoños deben producir una buena población de tallos (macollamiento), para evitar que se produzcan pérdidas de las cepas en el surco a través de las cosechas.
- f) El tiempo entre la germinación (si es caña nueva) o retoñamiento (si es caña soca), debe ser el menor posible.
- g) El tiempo requerido entre la siembra o cosecha y el “cierre” de la plantación debe ser corto, de tal manera que el periodo crítico de competencia con malezas sea también corto y haya mayor respuesta a las labores tecnológicas.
- h) El sistema radical debe ser desarrollado con capacidad de profundizar para que proporcione un buen anclaje.
- i) El tallo no debe florear y no formar corcho.
- j) Hábito de crecimiento erecto.
- k) Tolerancia a falta de agua y a la salinidad.
- l) Ausencia de raíces adventicias, rajaduras y desintegración interna de los tejidos.
- m) Resistencia al uso de los herbicidas convencionales que se emplean en el cultivo.

Los programas de mejoramiento genético pueden trabajar sobre cualquier otra característica que resulte de interés particular.

En el ISC se determinó la variedad por lote de acuerdo al siguiente criterio: si el 75% o más del área cultivada dentro de un lote correspondía a una variedad específica, se asignaba dicha variedad al lote; caso contrario, se le asignaba una variedad ficticia denominada MISCELÁNEA que puede interpretarse como la agrupación de algunas variedades dentro del lote. Las variedades sembradas en el ISC son: B-74132, BJ-7046, CC-8592, CR-74250, NCO-310, PR671070 y RAGNAR.

2.1.3.2.3. Número de riegos

El riego tiene como objetivo suplir el agua que la planta requiere cuando no es suministrada de manera natural y así completar de manera satisfactoria su desarrollo. La cantidad debida concuerda con el desarrollo fisiológico del cultivo para tratar de proporcionar la mínima cantidad posible sin provocar efectos negativos en el

rendimiento. La determinación del momento óptimo de aplicación del riego es de suma importancia desde el punto de vista agrícola, industrial y económico.

En el ISC la programación del riego se realiza de la siguiente manera: el primer riego en caña planta debe ser realizado inmediatamente después de la siembra y la instalación de los canales de riego. El primer riego de la caña soca debe ser aplicado, una vez realizadas las labores agrícolas programadas, lo más pronto posible después de la reparación de los canales de riego. Las necesidades de agua para la caña están determinadas por la evapotranspiración del suelo y la planta y las aplicaciones de riego se basan en el programa del **balance hídrico**.

La determinación de los riegos está basada en que el 50% (como mínimo) de la capacidad de almacenamiento de los 30 cm de profundidad esté disponible para ser utilizado por la planta, en el caso de los suelos francos y arenosos; mientras que en los suelos pesados, se considera 60% (mínimo) como disponible al momento de riego.

2.2. ANÁLISIS ESTADÍSTICO EXPLORATORIO

El análisis estadístico exploratorio tiene como objetivo realizar una descripción individual y detallada de cada una de las variables que representan tanto a la producción de caña de azúcar como a sus factores. La información recopilada corresponde al período 2002-2003 o zafra 2003 y es importante mencionar que no fue posible recolectar toda la información sobre producción sino hasta finales del año 2003 cuando culminó la zafra o cosecha.

Los factores de producción a considerar para el análisis se muestran en el Cuadro 2.6 y son aquellos que presentan un nivel de agregación de datos de precisión; es decir, aquellos para los que se puede obtener la información requerida por lote, la unidad mínima de manejo resultado de la implementación de AP en el ISC.

Cuadro 2.6. Factores de producción con un nivel de agregación de precisión

Factores Naturales	Tercio del cultivo
	Textura del suelo
	Edad del cultivo
	Edad a la cosecha
	pH
	Materia orgánica (MO)
	Fósforo (P)
	Potasio (K)
	Calcio (Ca)
	Magnesio (Mg)
Relaciones entre Ca, Mg y K	
Conductividad Eléctrica (CE)	
Factores de Manejo	Sector
	Variedad de la caña
	Número de riegos

Elaboración: Xavier Vítóres, 2004

2.2.1. Población, marco y muestra

Una tarea importante para el investigador es definir cuidadosa y completamente la población antes de recolectar la muestra. Una población es una colección de elementos acerca de los cuales deseamos hacer alguna inferencia. Esta población inicial que se desea investigar se denomina **población objetivo**. Los grupos no solapados (de intersección vacía) de elementos de la población que

cubran la población objetivo se denominan **unidades de muestreo** [20].

Al momento de seleccionar el conjunto de unidades de muestreo que componen la muestra, será necesario disponer de un listado material de unidades de muestreo. Esta colección de unidades de muestreo, de la que se selecciona la muestra se denomina **marco muestral**. Lo ideal sería disponer de un marco tal que la lista de unidades muestrales que lo componen coincida con la población objetivo. Pero en la práctica el marco contiene impurezas debidas a desactualizaciones, errores, omisiones y otras causas que hacen que el marco no coincida con la población objetivo. De todas formas, la distancia entre el marco y la población objetivo ha de ser lo suficientemente pequeña como para permitir que se hagan inferencias acerca de la población basándose en una muestra obtenida del marco.

En el análisis que se presenta en esta sección, se define a la población objetivo como el conjunto de todos los lotes o unidades más pequeñas de manejo resultado de la implementación de la AP en el

ISC. La descripción de la población objetivo, periodo de recolección de datos, unidades de muestreo y marco muestral se describen a continuación en el Cuadro 2.7.

Cuadro 2.7. Población objetivo, depuración del marco muestral y muestra

Población objetivo:	Conjunto de lotes resultado de la implementación de AP en el ISC, año 2003
Periodo de recolección de datos:	2002-2003 o zafra 2003
Unidad de muestreo:	Lote
Marco muestral :	Base de datos correspondiente al periodo de cosecha 2003 o zafra 2003 (Fuente: ISC)
Número de lotes en la población objetivo:	1309
Depuración del marco muestral	
Omisiones	3
Duplicaciones	11
Rezagados	91
Barbechos	72
SP*	24
SAS**	237
Muestra (lotes)	871
Total hectáreas	11065,68
(*) Lotes sin información sobre producción	
(**) Lotes sin información completa sobre análisis de suelos	

Elaboración: Xavier Vítors, 2004

En el proceso de depuración del marco no se tomaron en cuenta ciertas unidades de muestreo o lotes debido a errores en la base de datos, omisiones o debido a que no fue posible obtener todos los valores para las variables objeto de análisis. Las unidades descartadas son:

1. Omisiones: Lotes que no se pudieron encontrar en la base de datos
2. Duplicaciones: Lotes con el mismo número de código en la base pero con información diferente o errónea para alguna de las variables medidas.
3. Rezagados: Lotes con fecha de inicio de cultivo en el año 2001. La información de los lotes no corresponde al período de análisis 2002-2003.
4. Barbechos: Lotes en donde se dejó de cultivar para mejorar la fertilidad del suelo. Se suele barbechar (práctica agrícola) al octavo, noveno o décimo año de cultivo continuo. No es posible obtener información sobre producción de caña.
5. SP: Lotes sin información de producción

6. SAS: Lotes sin información en el análisis de suelos sobre los niveles de pH, fósforo, potasio, calcio, magnesio o materia orgánica.

De esta depuración se obtiene, al final, una muestra de 871 lotes de los cuales se obtendrá toda la información necesaria para llevar a cabo el análisis estadístico exploratorio en este capítulo y el análisis econométrico en el siguiente capítulo.

2.2.2. Definición de variables

En esta sección se definen las variables que representan y cuantifican a la producción de caña de azúcar y algunos de los factores de producción naturales y de manejo anteriormente descritos.

Para el análisis resultó conveniente definir las variables sobre la base de los factores de producción. Esto significa que se contará con 14 variables que representen a los factores naturales y 3 variables que

representen a los factores de manejo. Por lo tanto, resulta conveniente definir las y codificarlas de la siguiente manera:

Variable $X_{j,k}$.- Donde el subíndice j indica si la variable corresponde a un factor natural o de manejo y el subíndice k indica el factor correspondiente. Para j el valor 1 representa un factor natural y 2 un factor de manejo. Las variables se muestran en el Cuadro 2.8.

Cuadro 2.8. Definición de variables

Valores de j	Valores de k	Variable $X_{j,k}$
1	1	Tercio de cultivo
	2	Edad del cultivo
	3	Edad a la cosecha
	4	Textura del suelo
	5	pH
	6	Materia orgánica (MO)
	7	Fósforo (P)
	8	Potasio (K)
	9	Calcio (Ca)
	10	Magnesio (Mg)
	11	Relación Ca-Mg
	12	Relación Mg-K
	13	Relación Ca-K
	14	Conductividad Eléctrica (CE)
2	1	Sector
	2	Variedad de caña
	3	Número de riegos

Elaboración: Xavier Vítores, 2004.

Por ejemplo, la variable $X_{2,2}$ representa a la variedad de caña de azúcar plantada en un determinado lote como práctica de manejo agrícola; y, la variable $X_{1,4}$ representa a la textura de suelo presente en un determinado lote como factor natural.

2.2.3. Análisis univariado

Una vez definidas las variables, se procederá a realizar un análisis univariado descriptivo para cada una de ellas, tomando en cuenta el tamaño de la muestra de 871 lotes.

En este análisis se muestran los principales gráficos y estadísticos descriptivos de cada variable. De esta manera se podrá tener información sobre la dispersión de los datos y algunos aspectos de relevancia agronómica.

**Producción: Toneladas métricas de caña de azúcar por hectárea
(TMC/ha)**

Cuadro 2.9. Estadística descriptiva de la producción

N		871
Media		75,2013
Mediana		76,5900
Desviación Estándar		18,12659
Varianza		328,573
Coef. Asimetría		-,021
Kurtosis		-,253
Rango		119,35
Mínimo		20,89
Máximo		140,24
Percentiles	25 (primer cuartil)	62,6067
	50 (segundo cuartil)	76,5900
	75 (tercer cuartil)	87,9654

Los estadísticos más importantes sobre la producción se muestran en el Cuadro 2.9. La media es igual a 75,2013 TMC/ha. El promedio de las desviaciones con respecto a la media es de 18,126 TMC/ha. El rango o diferencia entre el valor mínimo de la producción (20,89 TMC/ha.) y el máximo (140,24 TMC/ha) es de 119,35 TMC/ha. La distribución empírica para los valores de la producción se muestra en la Figura 2.7 y se puede apreciar su similitud a una distribución normal.

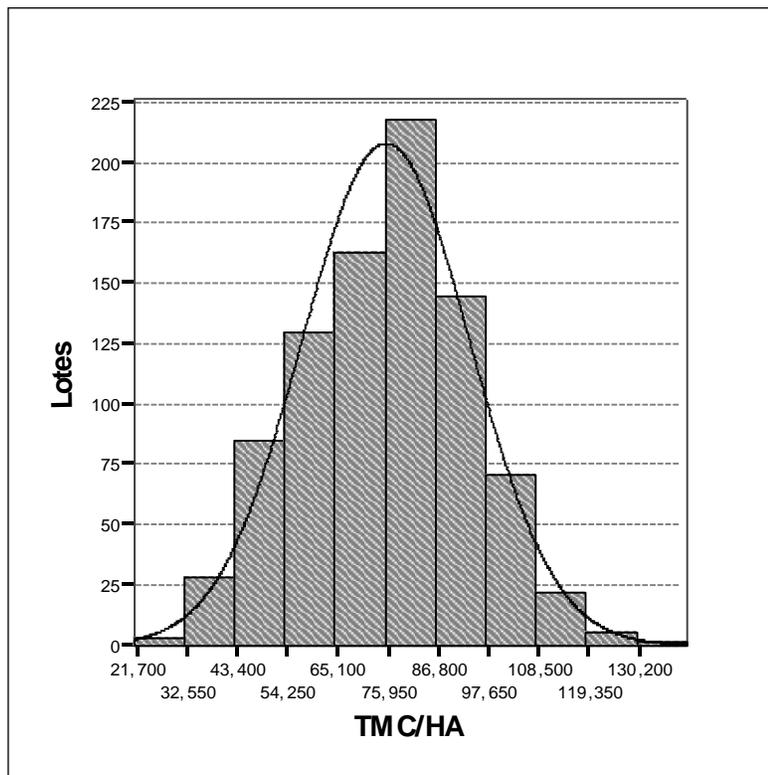


Figura 2.7. Histograma de la producción

El diagrama de caja de la Figura 2.8 muestra el primer, segundo y tercer cuartil, así como los valores extremos. De todos los lotes en la muestra, el 50% presentan una producción entre 62,6 TMC/ha (primer cuartil) y 87,96 TMC/ha (tercer cuartil). El segundo cuartil corresponde a la mediana que es una medida de tendencia central

robusta ante los valores extremos de 20,887 TMC/ha, 140,240 TMC/ha y 127,570 TMC/ha.

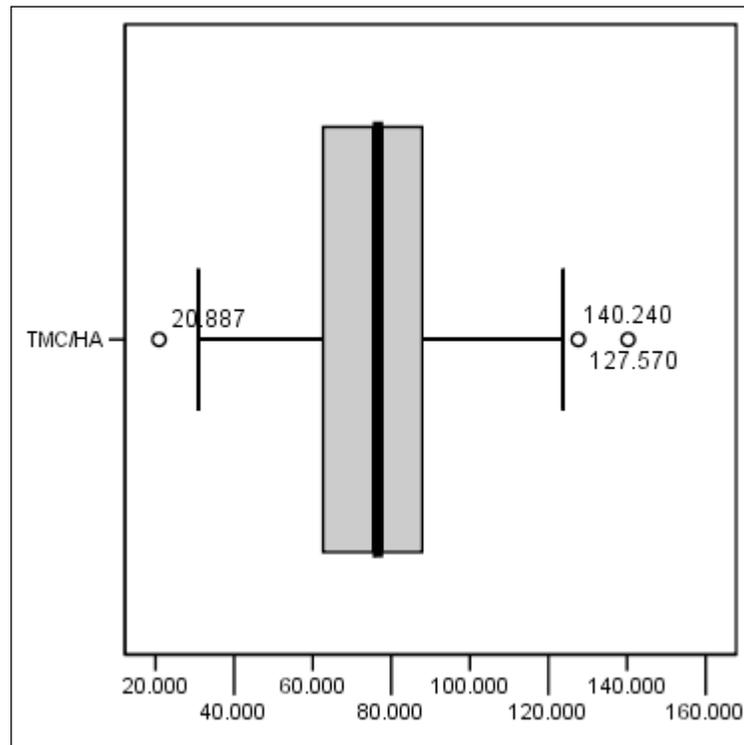


Figura 2.8. Diagrama de caja de la producción

El coeficiente de kurtosis indica el grado en que los datos se agrupan en torno a la media y que tan largas o cortas son las colas en comparación a la de una distribución normal (mostrada en la curva

acampanada de la Figura 2.7). El valor de este coeficiente para la distribución de las observaciones de TMC/ha es de -0,253 e indica que los datos se encuentran un poco menos agrupados en torno a la media en comparación a los de una distribución normal y que las colas son menos extensas.

El coeficiente de asimetría indica el grado de simetría o asimetría de la distribución. La distribución normal es simétrica y tiene un coeficiente de asimetría igual a cero. El valor para este coeficiente para las observaciones de TMC/ha. es -0,021 e indica que la distribución de éstas es casi totalmente simétrica debido a que posee un ligero sesgo negativo.

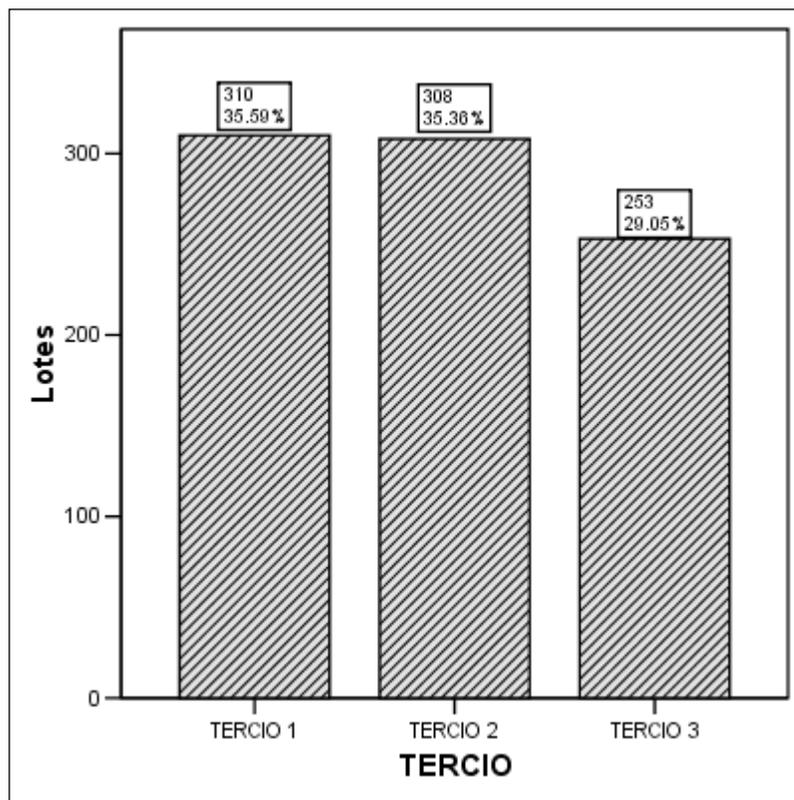
X_{1,1}: Tercio de cultivo

Figura 2.9. Diagrama de barras de Tercio de cultivo.

La estadística descriptiva del Cuadro 2.10 muestra que la mayoría de los lotes, es decir, un 35,59 % del total de la muestra (310 lotes) corresponden al primer tercio; un 35,36% (308 lotes) corresponden al segundo tercio; y, un 29,05% (253 lotes) corresponden al tercer tercio.

Cuadro 2.10. Estadística descriptiva de Tercio de cultivo

TERCIO DE CULTIVO	Frecuencia de lotes	Porcentaje de lotes
TERCIO 1	310	35,59
TERCIO 2	308	35,36
TERCIO3	253	29,05
Total	871	100,0

X_{1,2}: Edad del cultivo (años)

Cuadro 2.11. Estadística descriptiva de Edad del cultivo

N		871
Media		3,8634
Mediana		3,4961
Moda		4,00
Desviación Estándar		2,19161
Varianza		4,803
Rango		12,00
Minimo		1,00
Maximo		13,00
Percentiles	25	2,2508
	50	3,4961
	75	4,8855

La estadística descriptiva del Cuadro 2.11 muestra que la edad promedio del cultivo de la muestra es de 3,8634 años y el valor más frecuente representado por la moda es 4 años. La edad mínima es 1

año y la máxima 13 años. El 50% de los lotes tienen una edad entre 2,25 y 4,885 años. El diagrama de caja de la Figura 2.10 muestra la mediana, los percentiles 25 y 75 de la muestra y los valores extremos correspondientes a lotes con caña de 12 y 13 años. La Figura 2.11 muestra la distribución de la edad del cultivo en un diagrama de barras.

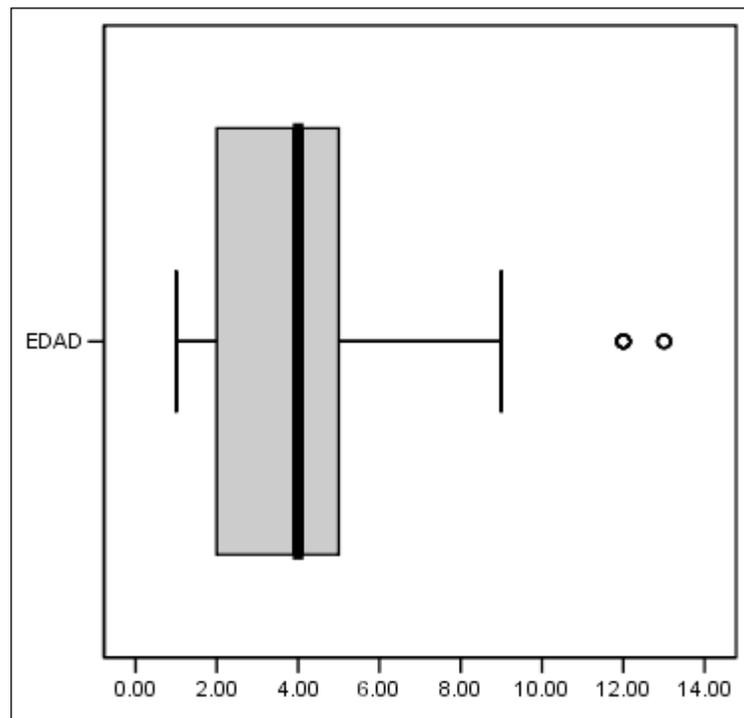


Figura 2.10. Diagrama de caja de Edad del cultivo

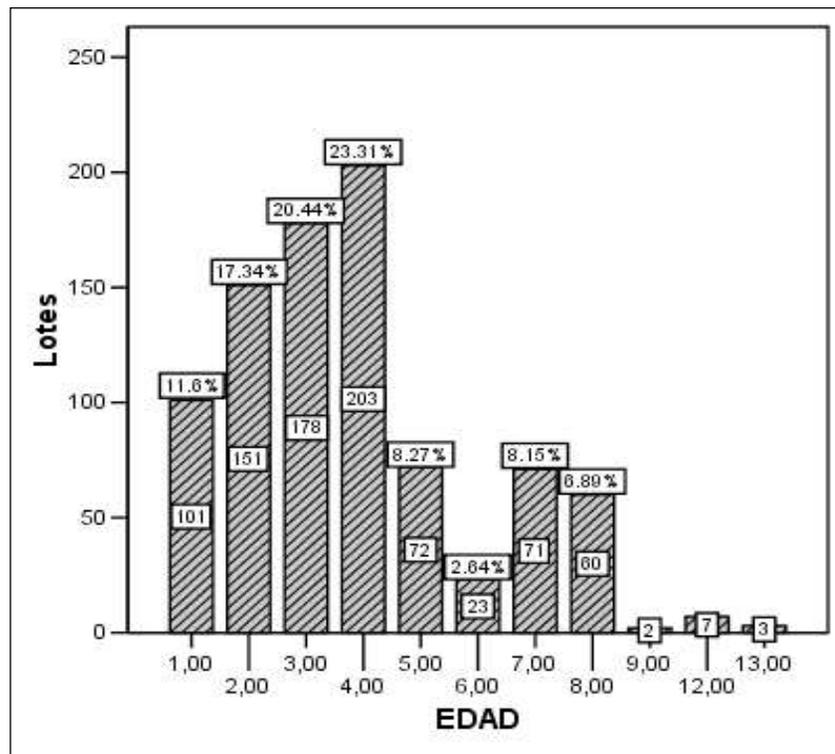


Figura 2.11. Diagrama de barras de Edad del cultivo

X_{1,3}: Edad a la cosecha (meses)**Cuadro 2.12.** Estadística descriptiva de Edad a la cosecha

N		871
Media		12,0623
Mediana		12,1200
Desviación Estándar		,65321
Varianza		,427
Rango		5,98
Mínimo		10,13
Máximo		16,11
Percentiles	25	12,0100
	50	12,1200
	75	12,2400

La estadística descriptiva del Cuadro 2.12 muestra que la edad promedio a la cosecha del cultivo es de 12,0623 meses y la mediana es de 12,12 meses. La edad mínima de cosecha es 10,13 meses y la máxima 16,11 meses. El 50% de los lotes tienen una edad a la cosecha entre 12,01 meses y 12,24 meses.

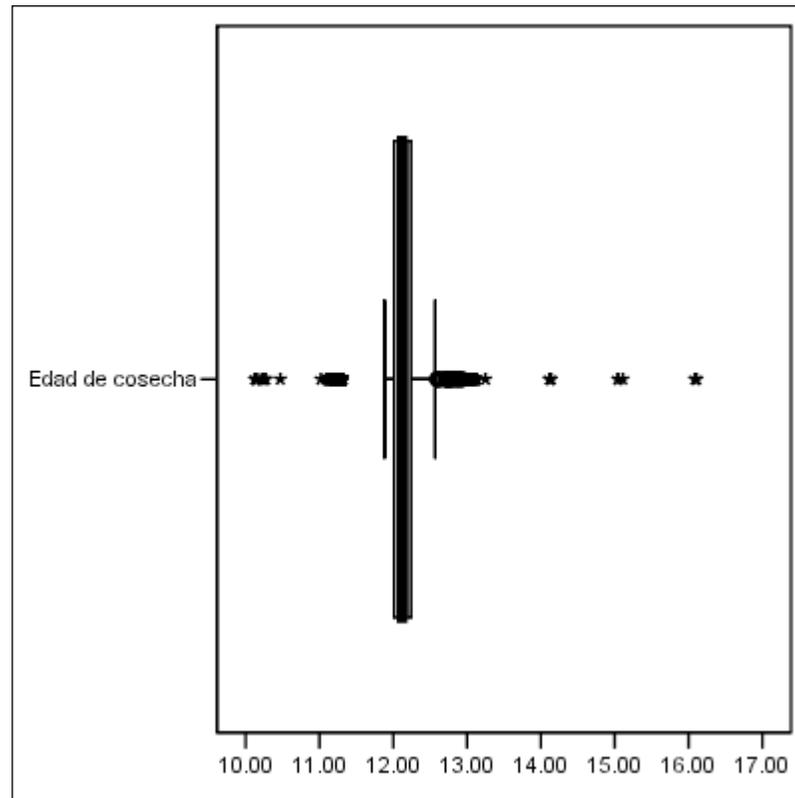


Figura 2.12. Diagrama de caja de Edad a la cosecha

El diagrama de caja de la Figura 2.12 muestra la mediana, los percentiles 25 y 75 de la muestra. Puede observarse que algunos lotes presentan caña cosechada a una edad mayor a los 14 meses y otros con caña cosechada a una edad menor a los 11 meses. La Figura 2.13 muestra la distribución de la edad a la cosecha.

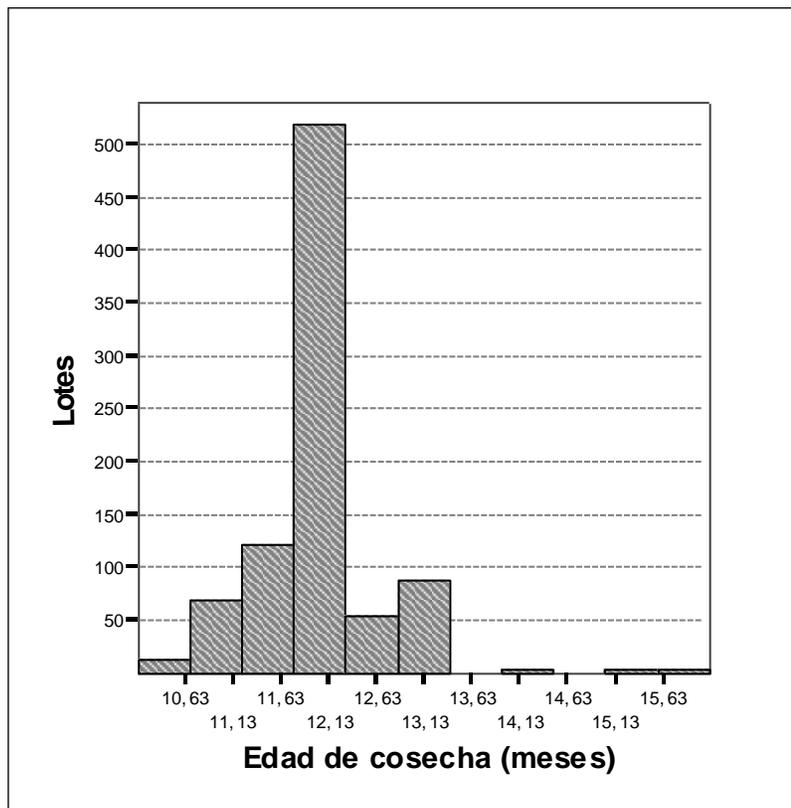


Figura 2.13. Histograma de Edad a la cosecha

X_{1,4} : Textura del suelo

Cuadro 2.13. Frecuencia de lotes por Textura del suelo

TEXTURA DEL SUELO		Frecuencia de lotes	Porcentaje de lotes
ARE.:	ARENOSO	20	2,3
F.:	FRANCO	155	17,8
ARC.:	ARCILLOSO	198	22,7
ARE. L.:	ARENO LIMOSO	15	1,7
F. ARE. :	FRANCO ARENOSO	138	15,8
F. L.:	FRANCO LIMOSO	22	2,5
F. ARC. ARE.:	FRANCO ARCILLOSO ARENOSO	28	3,2
F. ARC. :	FRANCO ARCILLOSO	272	31,2
F. ARC. L. :	FRANCO ARCILLOSO LIMOSO	14	1,6
ARC. L. :	ARCILLO LIMOSO	9	1,0
Total		871	100,0

La estadística descriptiva del Cuadro 2.13 y el diagrama de barras de la Figura 2.14 muestran que la mayoría de los lotes de la muestra, es decir, el 31,23% (272 lotes) presentan una textura de suelos franco arcillosa, seguida por la arcillosa con el 22,7% (198 lotes), franco con el 17,8% (155 lotes) y franco arenosa con el 15,8% (138 lotes). Los 108 lotes restantes (12,47%) son aquellos con texturas menos predominantes como la arenosa, areno limosa, franco limosa, franco arcillosa arenosa, franco arcillo limosa y arcillo limosa.

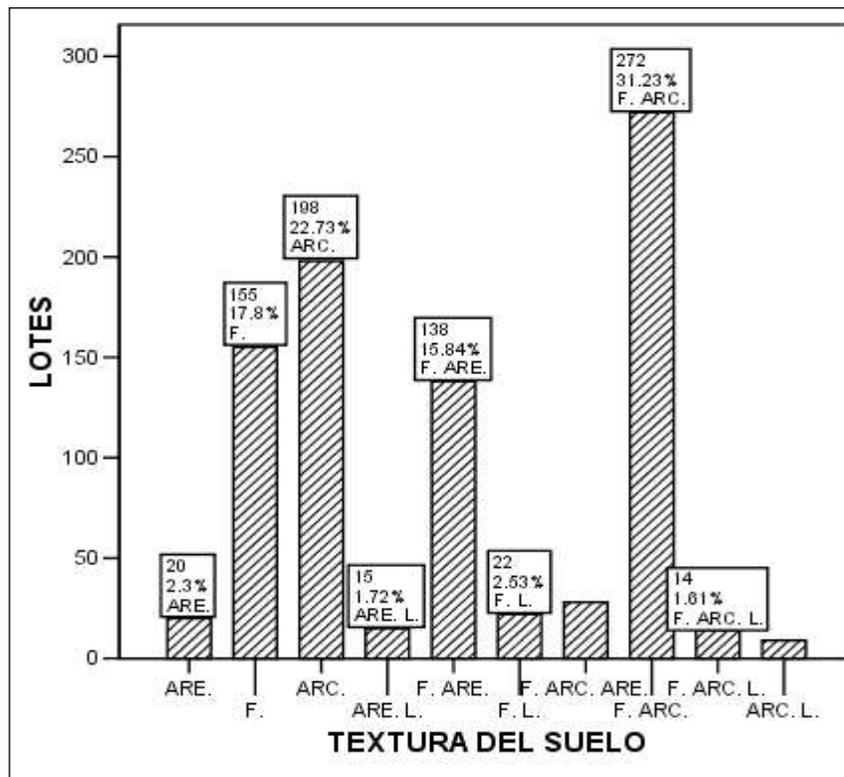


Figura 2.14. Diagrama de barras de Textura de suelo.

X_{1,5}: pH**Cuadro 2.14.** Estadística descriptiva de pH

N		871
Media		6,2721
Mediana		6,2600
Moda		6,29
Desviación Estándar		,42354
Varianza		,179
Rango		3,40
Mínimo		4,77
Máximo		8,17
Percentiles	25	5,9829
	50	6,2600
	75	6,5311

La estadística descriptiva del Cuadro 2.14 indica una media para el pH de 6,27 con un mínimo de 4,77 que representa una condición de acidez y un máximo de 8,17 que representa una condición de alcalinidad. La media, mediana y moda son similares y el promedio de las desviaciones con respecto a la media es de 0,42354.

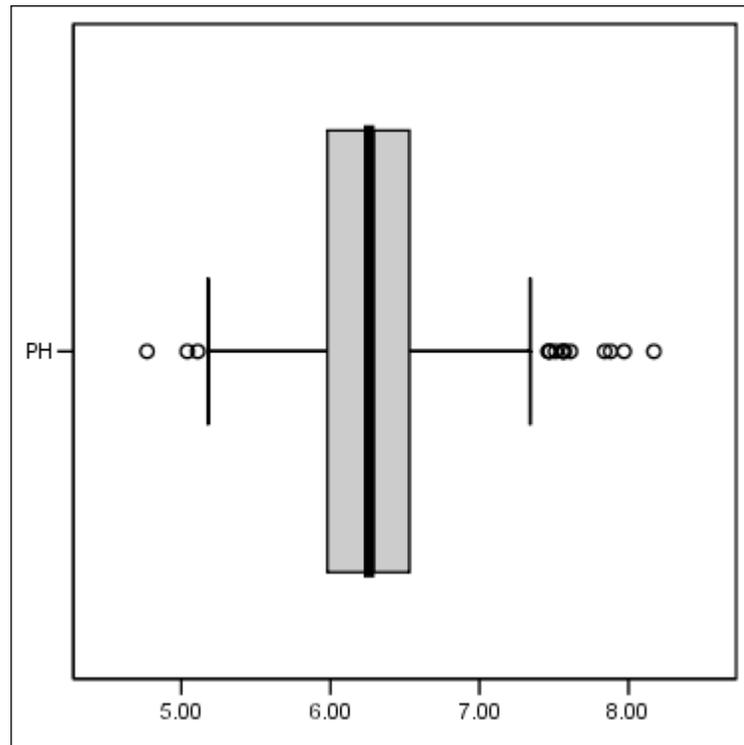


Figura 2.15. Diagrama de caja de pH

El diagrama de caja de la Figura 2.15 muestra que el 50% de los lotes de la muestra tiene valores de pH entre 5,9829 y 6,5311. Por otra parte en este diagrama puede observarse que existen más lotes con condiciones de alcalinidad ($\text{pH} > 6,5$) que condiciones de acidez ($\text{pH} < 5,5$). Esta situación también puede distinguirse en el histograma de la Figura 2.16.

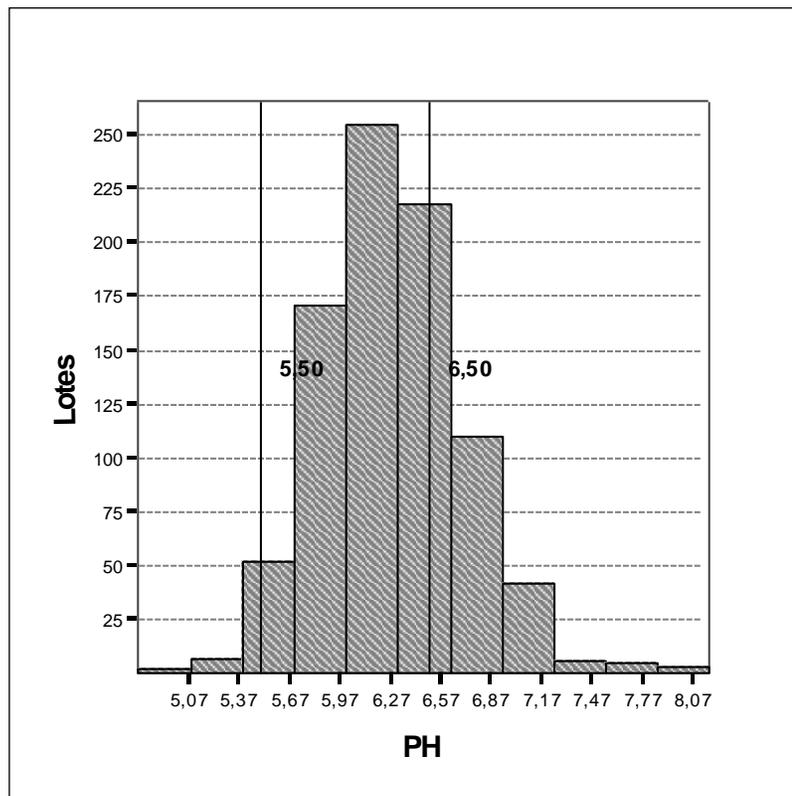


Figura 2.16. Histograma de pH

De acuerdo a la guía para el análisis de suelos mostrada en el Cuadro 2.5, los niveles de pH pueden clasificarse en las siguientes categorías por intervalos:

- Nivel alto: $>6,5$
- Nivel medio: $(5,5-6,5)$
- Nivel bajo: $<5,5$

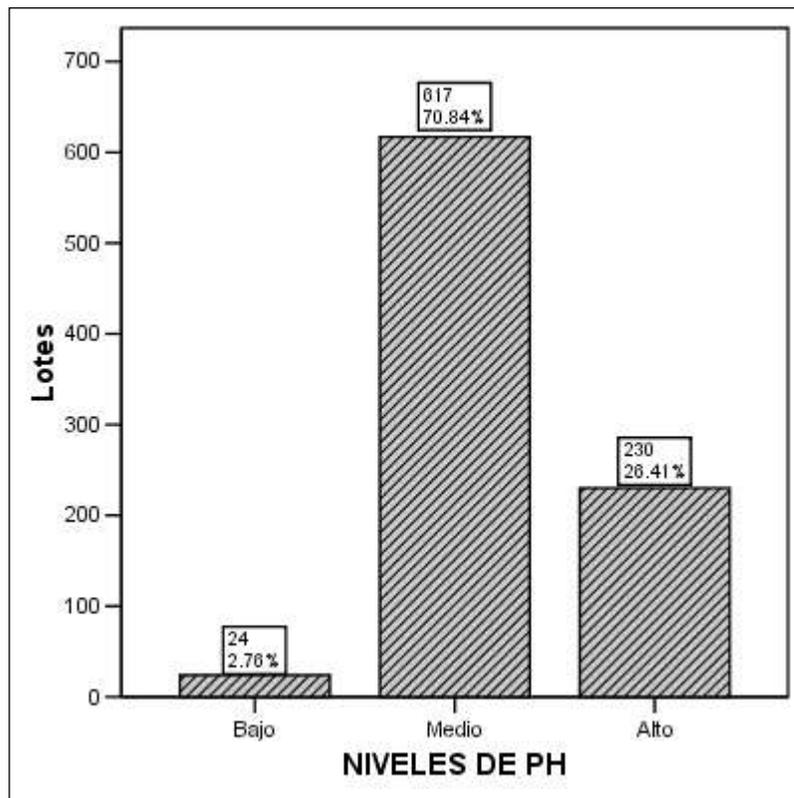


Figura 2.17. Diagrama de barras de pH

En el histograma de la Figura 2.16 y en el diagrama de barras de la Figura 2.17 se puede observar que la mayoría de los lotes, es decir, el 70,84% (617 lotes) presentan un nivel medio de pH; el 26,41% (230 lotes) un nivel alto y el 2,76% (24 lotes) niveles bajos.

X_{1,6} : Materia orgánica (%)**Cuadro 2.15.** Estadística descriptiva de Materia orgánica

N		871
Media		1,8237
Mediana		1,8220
Moda		1,82
Desviación Estándar		,45458
Varianza		,207
Rango		3,20
Mínimo		,35
Máximo		3,55
Percentiles	25	1,5163
	50	1,8220
	75	2,1437

La estadística descriptiva del Cuadro 2.15 muestra que la media para el porcentaje de materia orgánica en los lotes es de 1,8237% con un mínimo de 0,35% y un máximo de 3,55%. El promedio de las desviaciones con respecto a la media es de 0,45458%.

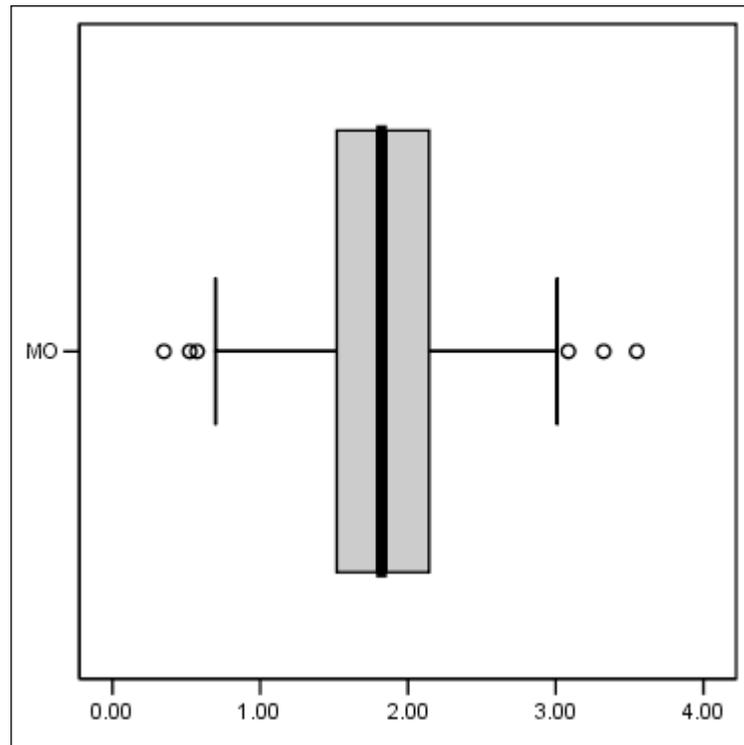


Figura 2.18. Diagrama de caja de Materia orgánica

El diagrama de caja de la Figura 2.18 muestra que el 50% de los lotes de la muestra tiene porcentajes de materia orgánica entre 1,5163% y 2,1437%.

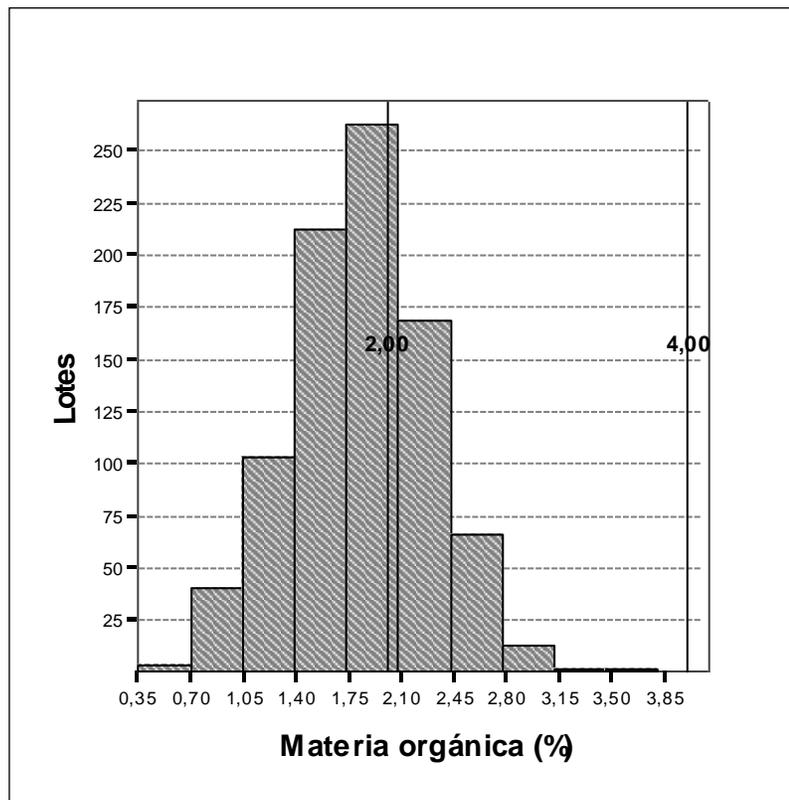


Figura 2.19. Histograma de Materia orgánica

El histograma de la Figura 2.19 muestra los valores críticos para los niveles de materia orgánica según la interpretación del análisis de suelos de CENICAÑA del Cuadro 2.4.

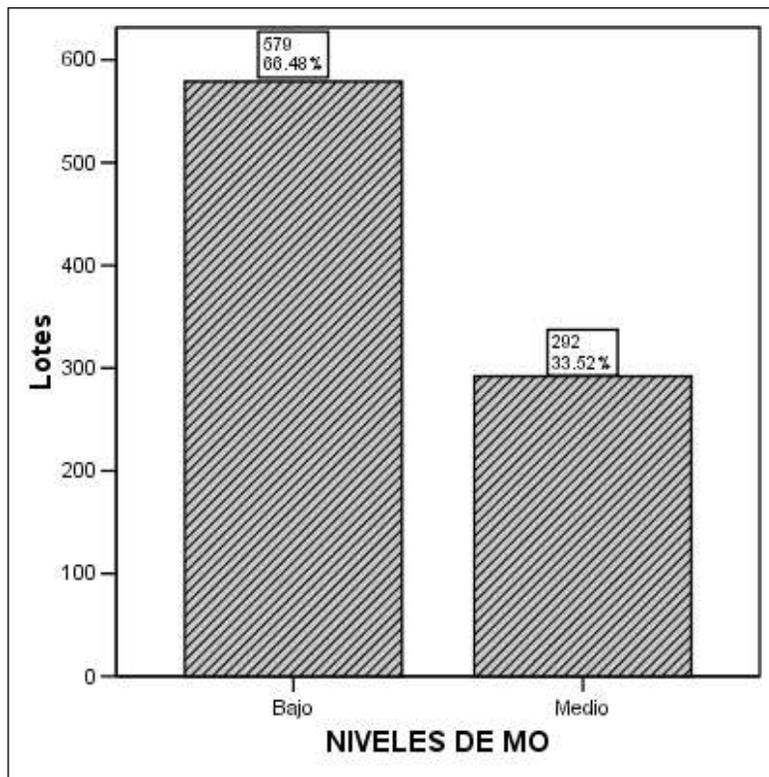


Figura 2.20. Diagrama de barras de Materia orgánica

Los niveles de materia orgánica pueden clasificarse en las siguientes categorías por intervalos:

- Alto: > 4%
- Medio: (2% - 4%)
- Bajo: <2%

El número de lotes por cada categoría puede observarse en el diagrama de barras de la Figura 2.20. La mayoría de los lotes, es decir, el 66,48% (579 lotes) tienen un nivel bajo de materia orgánica, el restante 33,52% (292 lotes) tienen un nivel medio.

X_{1,7}: Fósforo (ppm)

Cuadro 2.16. Estadística descriptiva de Fósforo

N		871
Media		15,5029
Mediana		11,6986
Moda		6,00
Desviación Estándar		12,71228
Varianza		161,602
Rango		107,00
Mínimo		1,00
Máximo		108,00
Percentiles	25	7,1903
	50	11,6986
	75	19,5595

La estadística descriptiva del Cuadro 2.16 muestra una media para los valores de P de 15,5 ppm con un mínimo de 1 ppm y un máximo de 108 ppm. El valor más frecuente o moda es 6 ppm. El promedio de las desviaciones con respecto a la media es de 12,71 ppm. El

diagrama de caja de la figura 2.15 muestra que el 50% de los lotes de la muestra tiene niveles de P entre 7,1903 ppm y 19,5595 ppm.

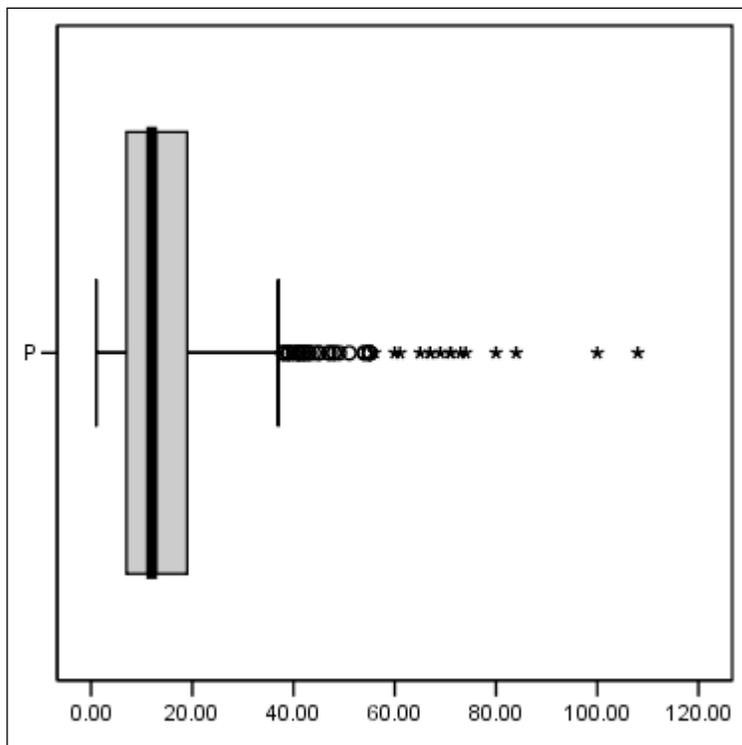


Figura 2.21. Diagrama de caja de Fósforo

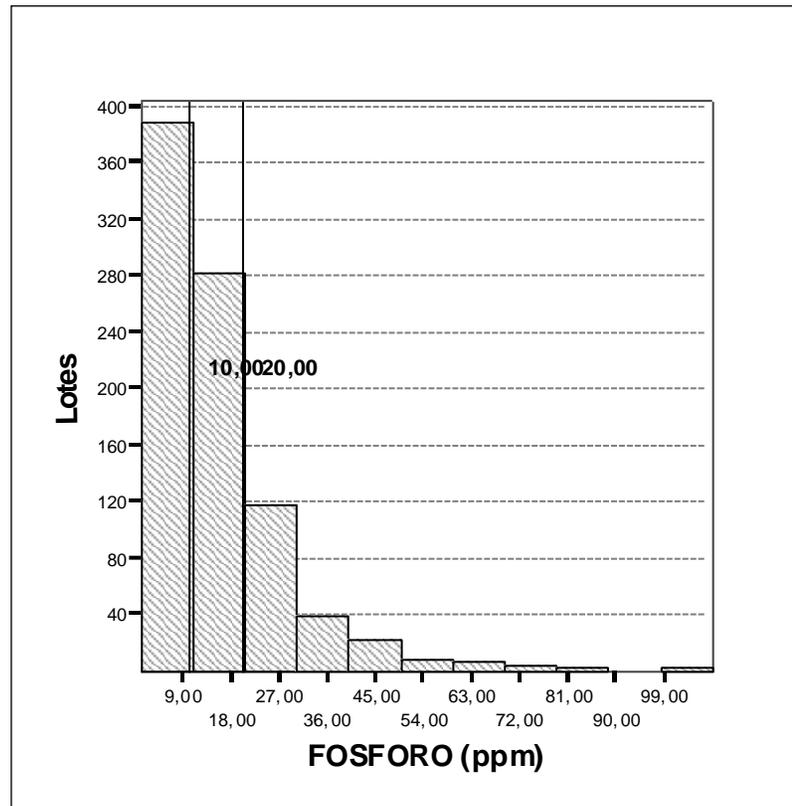


Figura 2.22. Histograma de Fósforo

De acuerdo a la guía para el análisis de suelos mostrado en el Cuadro 2.5, los niveles de P pueden clasificarse de acuerdo a las siguientes categorías por intervalos:

- Nivel alto: >20
- Nivel medio: (10 -20)
- Nivel bajo: <10

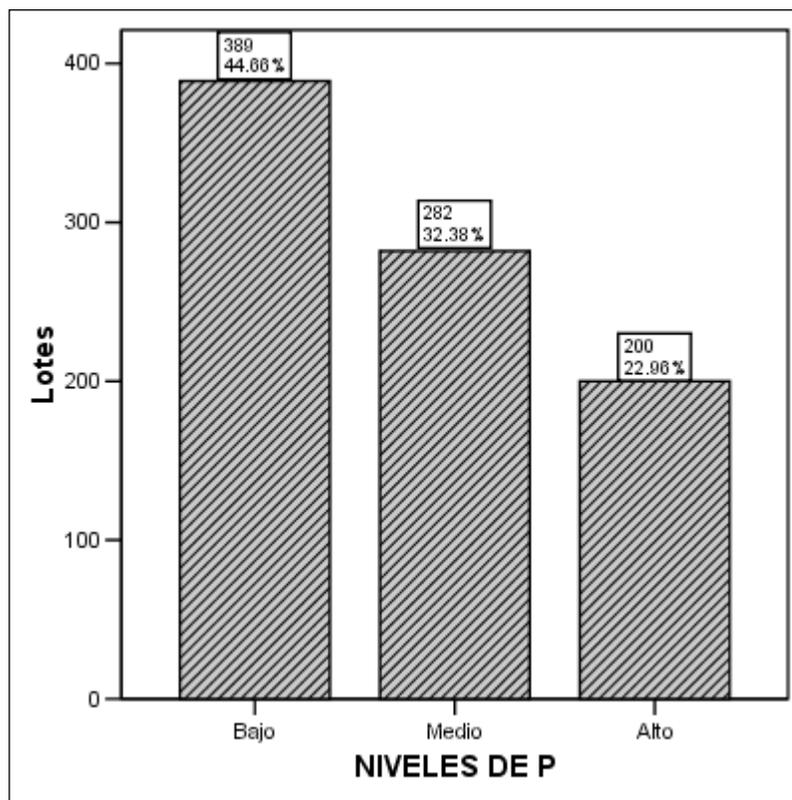


Figura 2.23. Diagrama de barras de Fósforo

En el diagrama de barras de la Figura 2.23 se puede observar que la mayoría de los lotes, es decir, el 44,66% (389 lotes) presentan un nivel bajo de P; el 32,38% (282 lotes) un nivel medio y el 22,96% (200 lotes) un nivel alto.

X_{1,8} : Potasio (cmol(+)/L)**Cuadro 2.17.** Estadística descriptiva de Potasio

N		871
Media		,4011
Mediana		,3253
Desviación Estándar		,25134
Varianza		,063
Rango		1,48
Mínimo		,06
Máximo		1,54
Percentiles	25	,2319
	50	,3253
	75	,4863

La estadística descriptiva del Cuadro 2.17 muestra una media para los valores de K de 0,4011 cmol(+)/L con un mínimo de 0,06 cmol(+)/L y un máximo de 1,54 cmol(+)/L. El promedio de las desviaciones con respecto a la media es de 0,2513 cmol(+)/L. El diagrama de caja de la Figura 2.24 muestra que un 50% de los lotes presentan valores entre 0,2319 cmol(+)/L y 0,4863 cmol(+)/L.

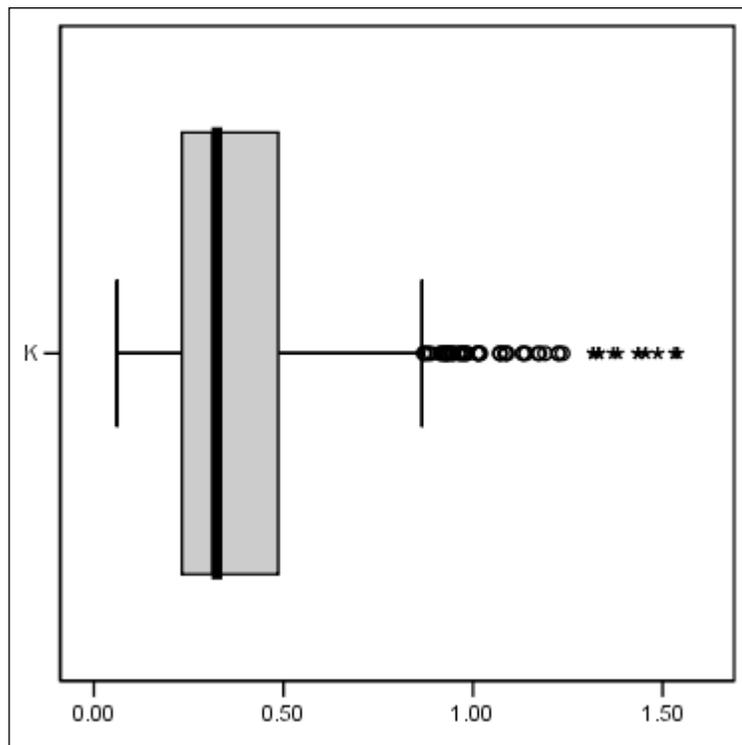


Figura 2.24. Diagrama de caja de Potasio

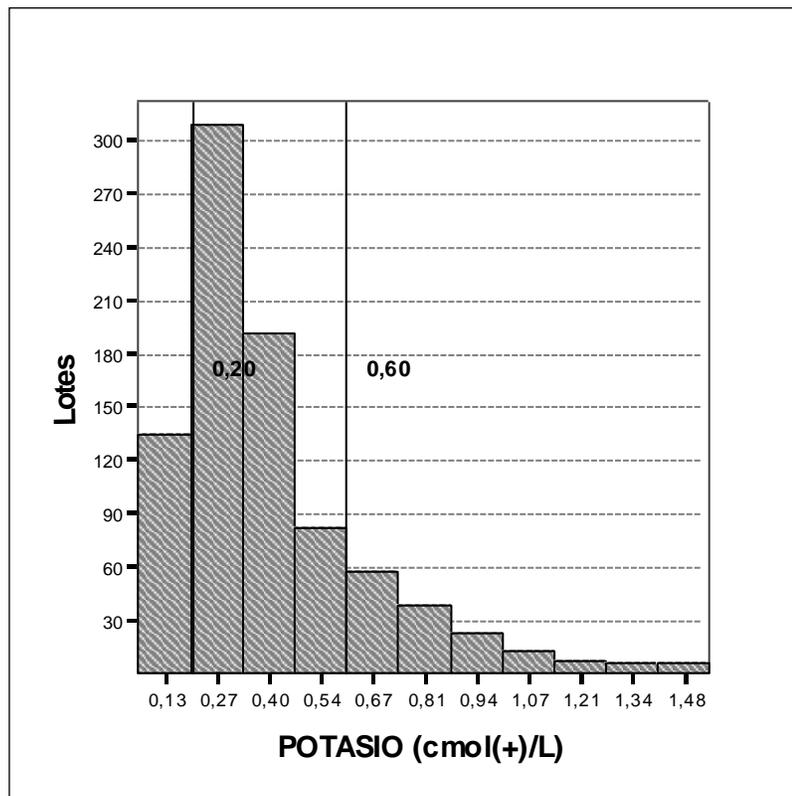


Figura 2.25. Histograma de Potasio

De acuerdo a la guía para el análisis de suelos mostrado en el Cuadro 2.5, los niveles de K pueden clasificarse de acuerdo a las siguientes categorías por intervalos:

- Nivel alto: $>0,60$
- Nivel medio: $(0,20 - 0,60)$
- Nivel bajo: $<0,20$

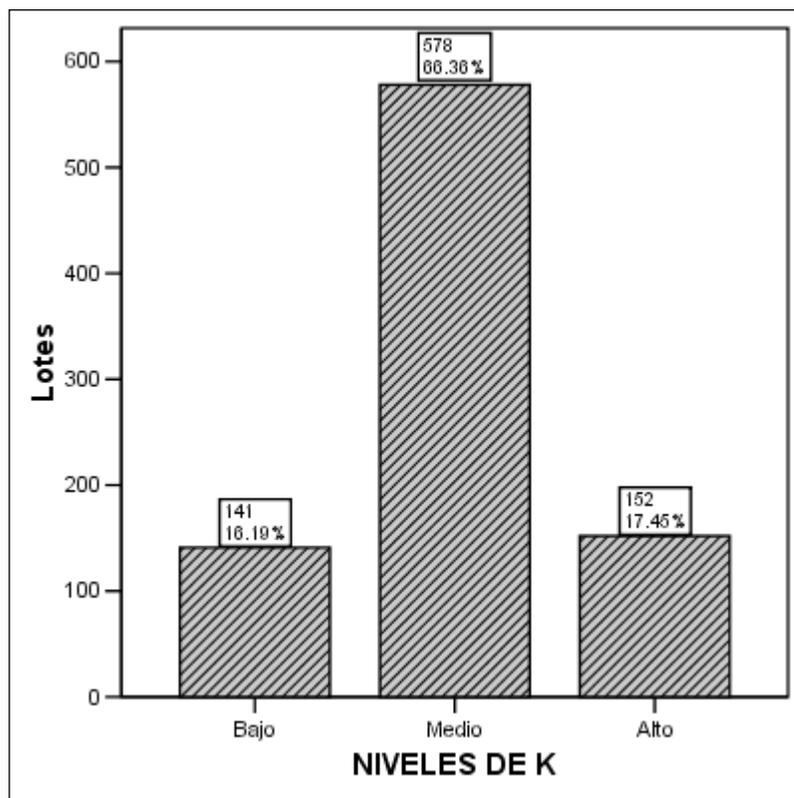


Figura 2.26. Diagrama de barras de Potasio

El diagrama de barras de la Figura 2.26 muestra que la mayoría de los lotes, es decir, el 66,36% (578 lotes) presentan un nivel medio de K; el 16,19% (141 lotes) niveles bajos y el 17,45% (152 lotes) niveles altos.

X_{1,9}: Calcio (cmol(+)/L)**Cuadro 2.18.** Estadística descriptiva de Calcio

N		871
Media		16,8093
Mediana		16,1600
Moda		15,76
Desviación Estándar		5,20702
Varianza		27,113
Rango		25,40
Mínimo		6,76
Máximo		32,16
Percentiles	25	12,6000
	50	16,1600
	75	20,5200

La estadística descriptiva del Cuadro 2.18 muestra una media para los valores de Ca de 16,8093 cmol(+)/L con un mínimo de 6,76 cmol(+)/L y un máximo de 32,16 cmol(+)/L. El promedio de las desviaciones con respecto a la media es de 5,207 cmol(+)/L. El diagrama de caja de la Figura 2.27 muestra que el 50% de los lotes presenta valores entre 12,6 cmol(+)/L y 20,52 cmol(+)/L.

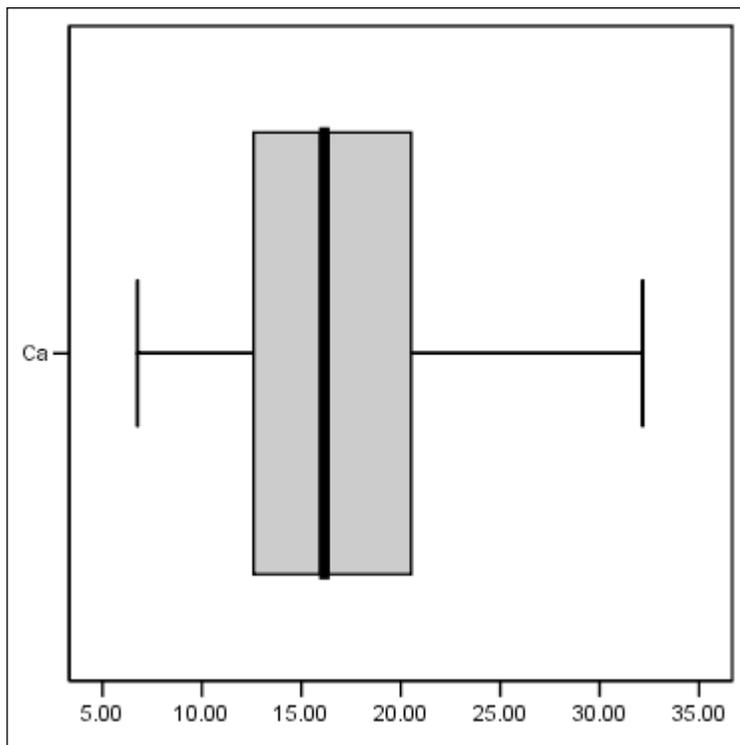


Figura 2.27. Diagrama de caja de Calcio

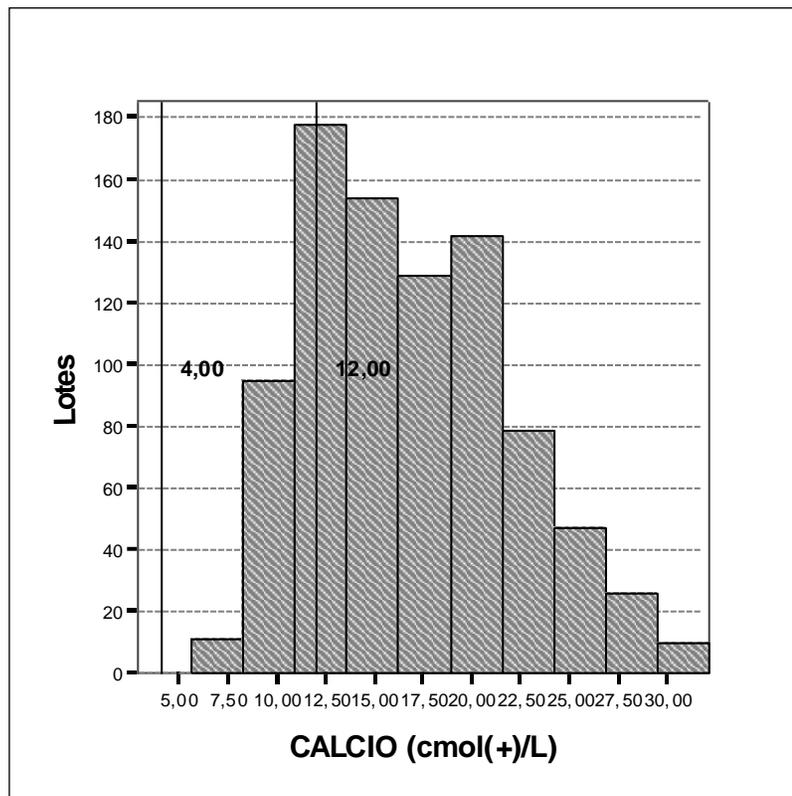


Figura 2.28. Histograma de Calcio

De acuerdo a la guía para el análisis de suelos mostrada en el Cuadro 2.5, los niveles de Ca pueden clasificarse de acuerdo a las siguientes categorías por intervalos:

- Nivel alto: $>12,0$
- Nivel medio: $(4,0 - 12,0)$
- Nivel bajo: $<4,0$

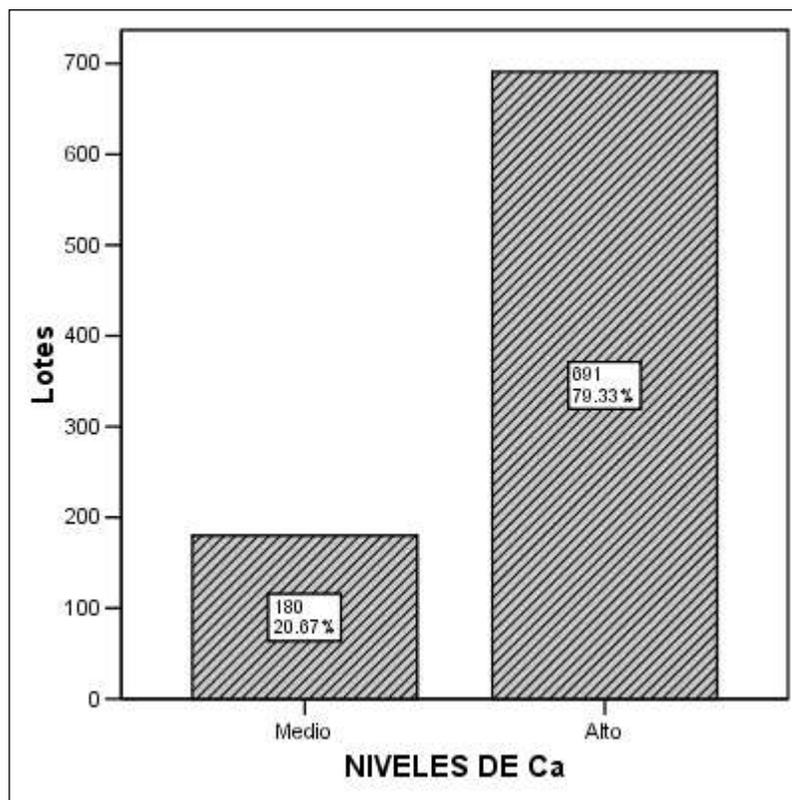


Figura 2.29. Diagrama de barras de Calcio

En el diagrama de barras de la Figura 2.29 se puede observar que la mayoría de los lotes, es decir, el 79,33% (691 lotes) presentan un nivel alto de Ca y el 20,67% (180 lotes) presentan niveles medios.

X_{1,10} : Magnesio (cmol(+)/L)**Cuadro 2.19.** Estadística descriptiva de Magnesio

N		871
Media		8,2450
Mediana		7,0000
Desviación Estándar		4,45470
Varianza		19,844
Rango		26,76
Mínimo		1,08
Máximo		27,84
Percentiles	25	4,9600
	50	7,0000
	75	10,6400

La estadística descriptiva del Cuadro 2.19 muestra una media para los valores de Mg de 8,2450 cmol(+)/L con un mínimo de 1,08 cmol(+)/L y un máximo de 27,84 cmol(+)/L. El promedio de las desviaciones con respecto a la media es de 4,4547 cmol(+)/L. El diagrama de caja de la Figura 2.30 muestra que el 50% de los lotes presentan valores entre 4,96 cmol(+)/L y 10,64 cmol(+)/L.

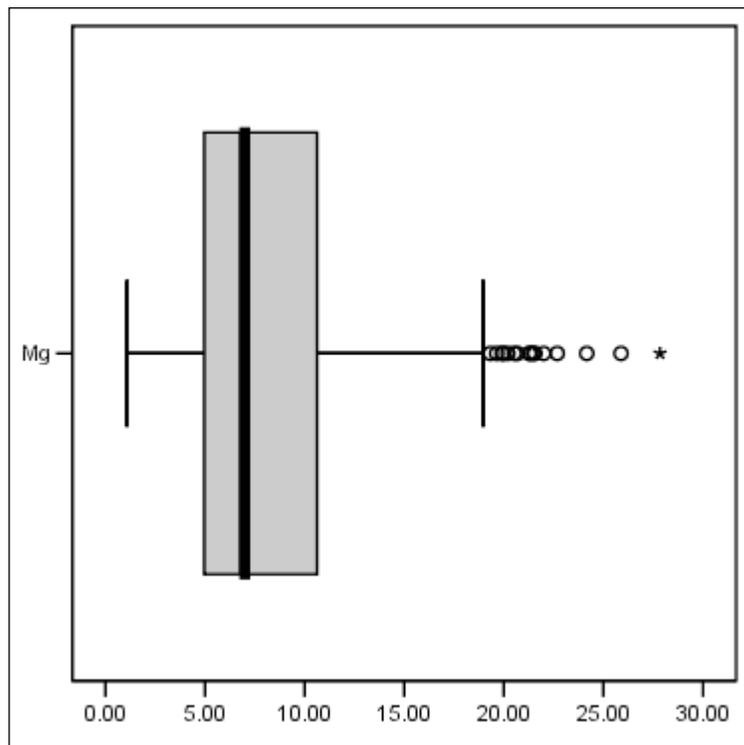


Figura 2.30. Diagrama de caja de Magnesio

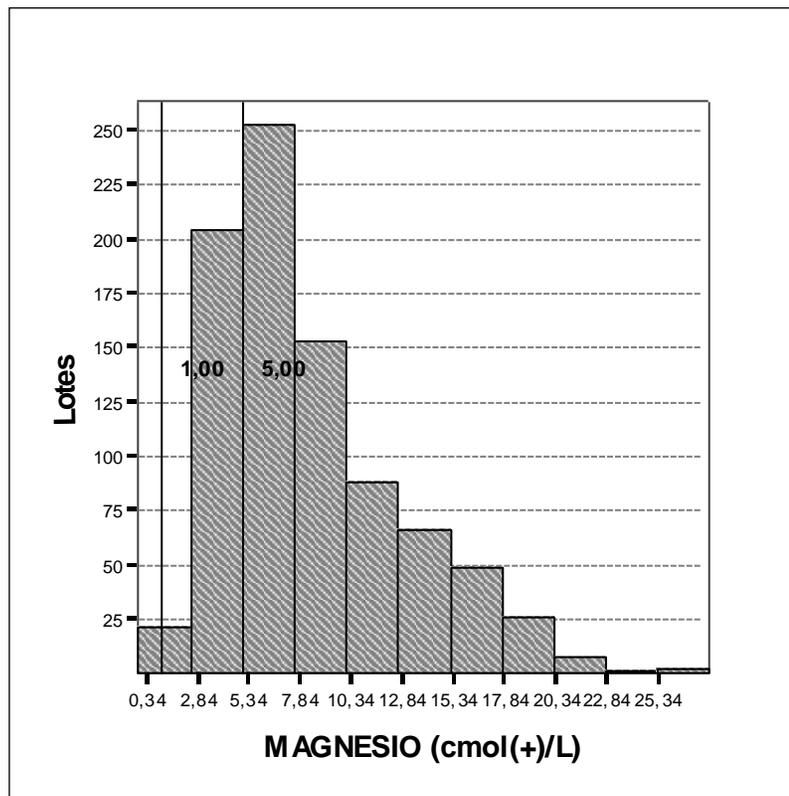


Figura 2.31. Histograma de Magnesio

De acuerdo a la guía para el análisis de suelos mostrada en el Cuadro 2.5, los niveles de Mg pueden clasificarse de acuerdo a las siguientes categorías por intervalos:

- Nivel alto: $>5,0$
- Nivel medio: $(1,0 - 5,0)$
- Nivel bajo: $<1,0$

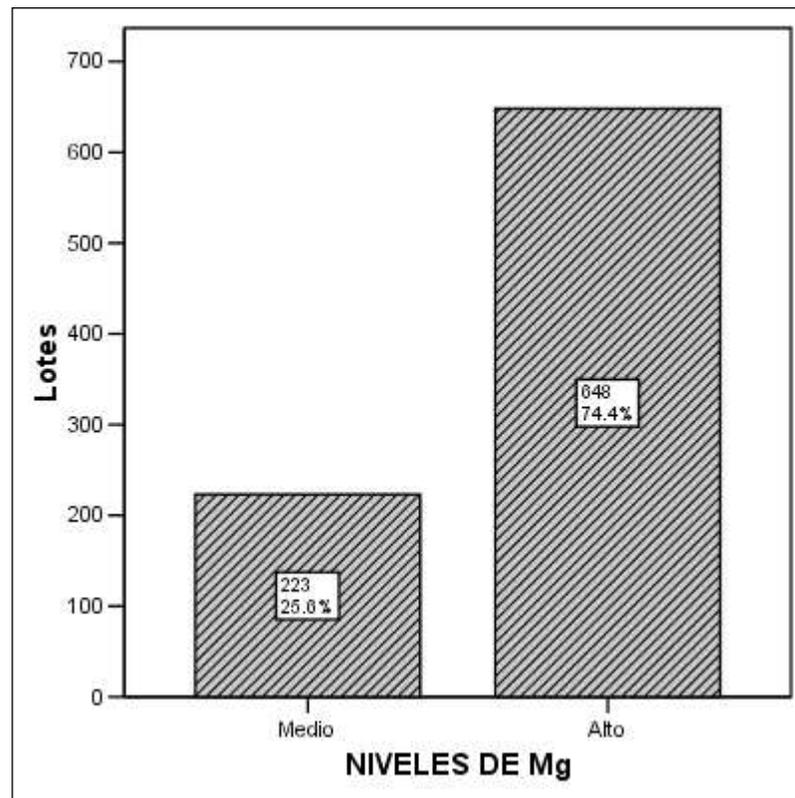


Figura 2.32. Diagrama de barras de Magnesio

En el diagrama de barras de la Figura 2.32 se puede observar que la mayoría de los lotes, es decir, el 74,4% (648 lotes) presentan un nivel alto de Mg y el 25,6% (223 lotes) presentan un nivel medio.

X_{1,11}: Relación Ca-Mg (Ca/Mg)**Cuadro 2.20** Estadística descriptiva de la relación Ca-Mg

N		871
Media		2,3926
Mediana		2,2229
Desviación Estándar		,98589
Varianza		,972
Rango		13,11
Mínimo		,78
Máximo		13,89
Percentiles	25	1,7440
	50	2,2229
	75	2,8022

La estadística descriptiva del Cuadro 2.20 muestra una media para la relación de 2,3926 con un mínimo de 0,78 y un máximo de 13,89. El promedio de las desviaciones con respecto a la media es de 0,98589. El diagrama de caja de la Figura 2.33 muestra que el 50% de los lotes presentan valores entre 1,744 y 2,8022. Los valores del extremo derecho corresponden a lotes con un desbalance ocasionado por exceso de Ca (relación >5).

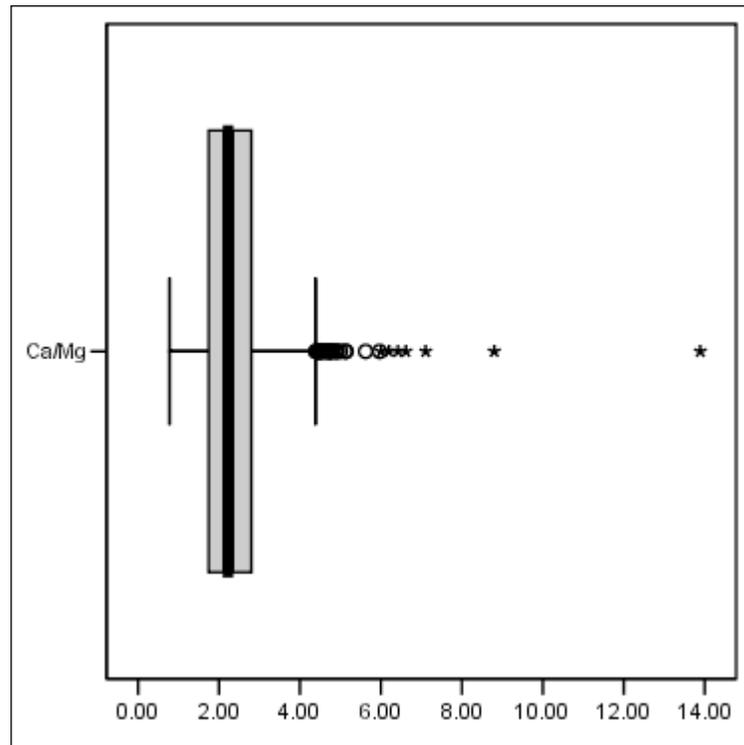


Figura 2.33. Diagrama de caja de la relación Ca-Mg

El histograma de la Figura 2.34 muestra la distribución empírica de los valores de la relación así como los valores críticos establecidos anteriormente en el Cuadro 2.5.

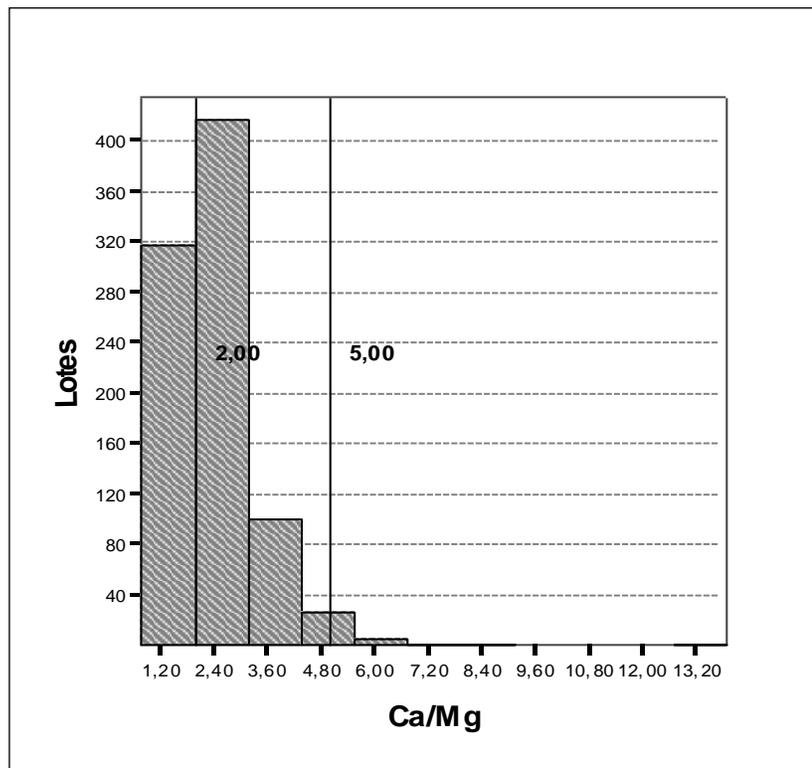


Figura 2.34. Histograma de la relación Ca-Mg

De acuerdo a la guía para el análisis de suelos mostrada en el Cuadro 2.5, los niveles de la relación pueden clasificarse de acuerdo a las siguientes categorías por intervalos:

- Desbalance por Ca: $>5,0$
- Balance: $(2,0 - 5,0)$
- Desbalance por Mg: $<2,0$

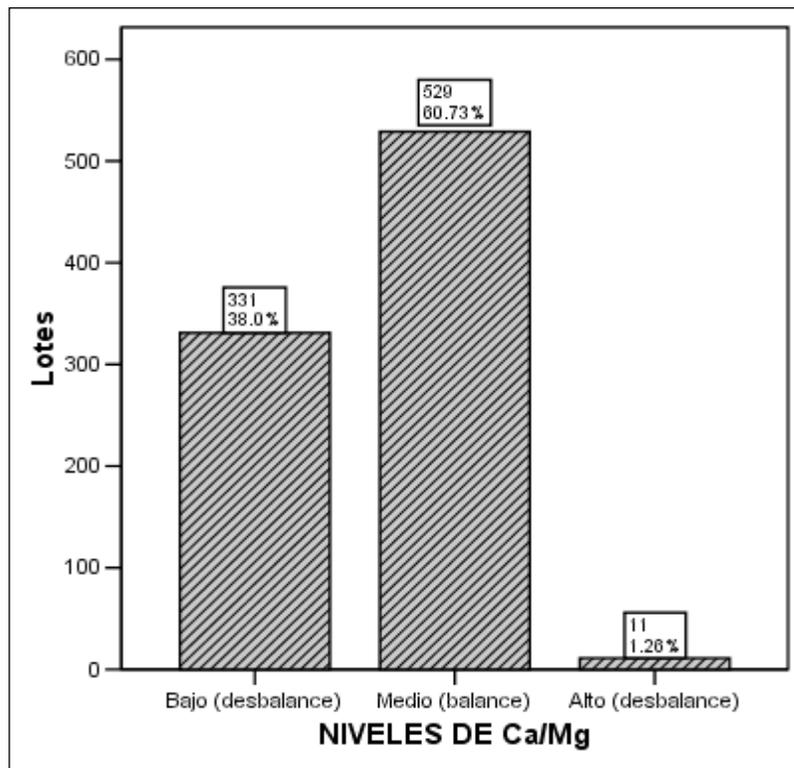


Figura 2.35. Diagrama de barras de la relación Ca-Mg

En el diagrama de barras de la Figura 2.35 se puede observar que la mayoría de los lotes, es decir, el 60,73% (529 lotes) están en balance; el 38% (331 lotes) están en desbalance con respecto al Mg; y, el 1,26% (11 lotes) presentan desbalance con respecto al Ca.

X_{1,12}: Relación Mg-K (Mg/K)**Cuadro 2.21.** Estadística descriptiva de la relación Mg-K

N		871
Media		29,0404
Mediana		23,6964
Desviación Estándar		22,42210
Varianza		502,750
Rango		140,24
Mínimo		,71
Máximo		140,95
Percentiles	25	11,8569
	50	23,6964
	75	39,4619

La estadística descriptiva del Cuadro 2.21 muestra una media para la relación de 29,0404 con un mínimo de 0,71 y un máximo de 140,95. El promedio de las desviaciones con respecto a la media es de 22,4221. El diagrama de caja de la Figura 2.36 muestra que el 50% de los lotes presentan valores entre 11,8569 y 39,4619. Los valores del extremo derecho corresponden a lotes con un desbalance ocasionado por exceso de Mg (relación >15).

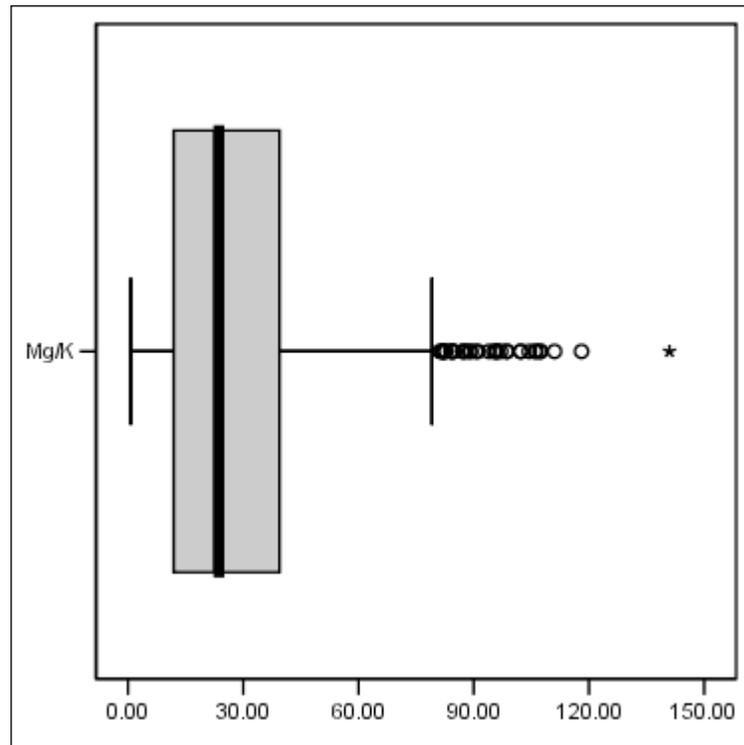


Figura 2.36. Diagrama de caja de la relación Mg-K

El histograma de la Figura 2.37 muestra la distribución empírica de los valores de la relación así como los valores críticos establecidos anteriormente en el Cuadro 2.5.

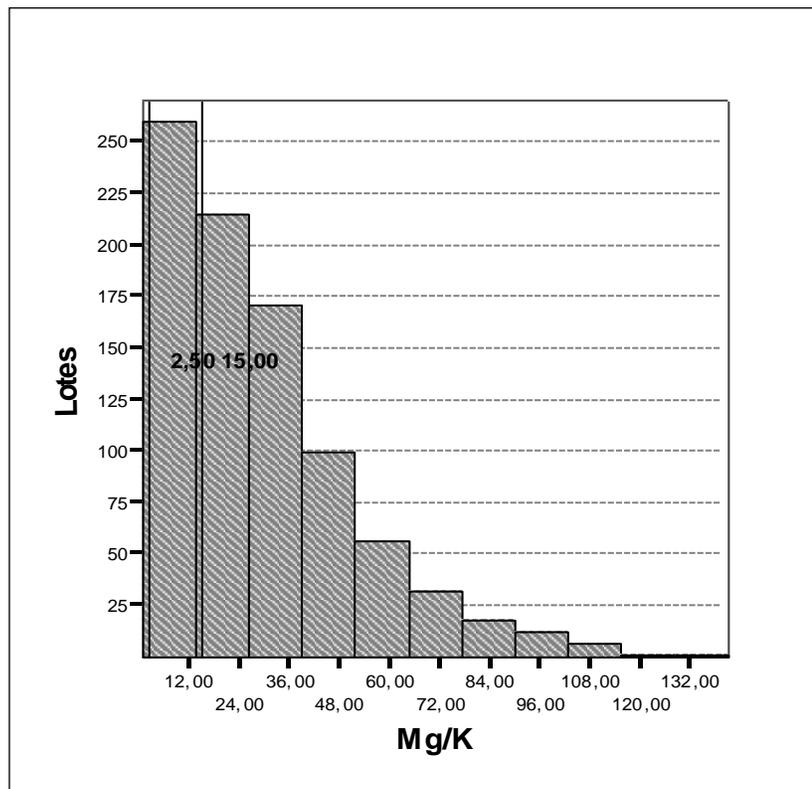


Figura 2.37. Histograma de la relación Mg-K

De acuerdo a la guía para el análisis de suelos mostrada en el Cuadro 2.5, los niveles de la relación pueden clasificarse de acuerdo a las siguientes categorías por intervalos:

- Desbalance por Mg: >15
- Balance: $(2,5 - 15,0)$
- Desbalance por K: $<2,5$

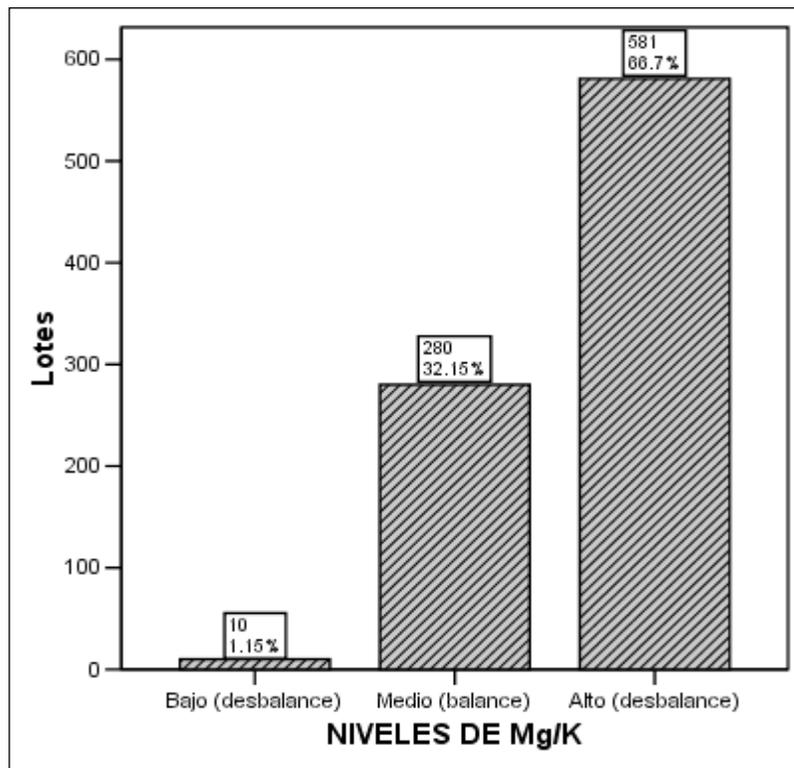


Figura 2.38. Diagrama de barras de la relación Mg-K

En el diagrama de barras de la Figura 2.38 se puede observar que la mayoría de los lotes, es decir, el 66,7% (581 lotes) están en desbalance con respecto al Mg; el 32,15% (280 lotes) están en balance; y, el 1,15% (10 lotes) presentan desbalance con respecto al K.

X_{1,13}: Relación Ca-K (Ca/K)**Cuadro 2.22.** Estadística descriptiva de la relación Ca-K

N		871
Media		57,3152
Mediana		53,7409
Moda		68,94
Desviación Estándar		33,34341
Varianza		1111,783
Rango		190,44
Mínimo		6,43
Máximo		196,88
Percentiles	25	29,5673
	50	53,7409
	75	79,2239

La estadística descriptiva del Cuadro 2.22 muestra una media para la relación de 57,3152 con un mínimo de 6,43 y un máximo de 196,88. El promedio de las desviaciones con respecto a la media es de 33,34341. El diagrama de caja de la Figura 2.39 muestra que el 50% de los lotes presentan valores entre 29,5673 y 79,2239. Los valores del extremo derecho corresponden a lotes con un desbalance ocasionado por exceso de Ca (relación >25).

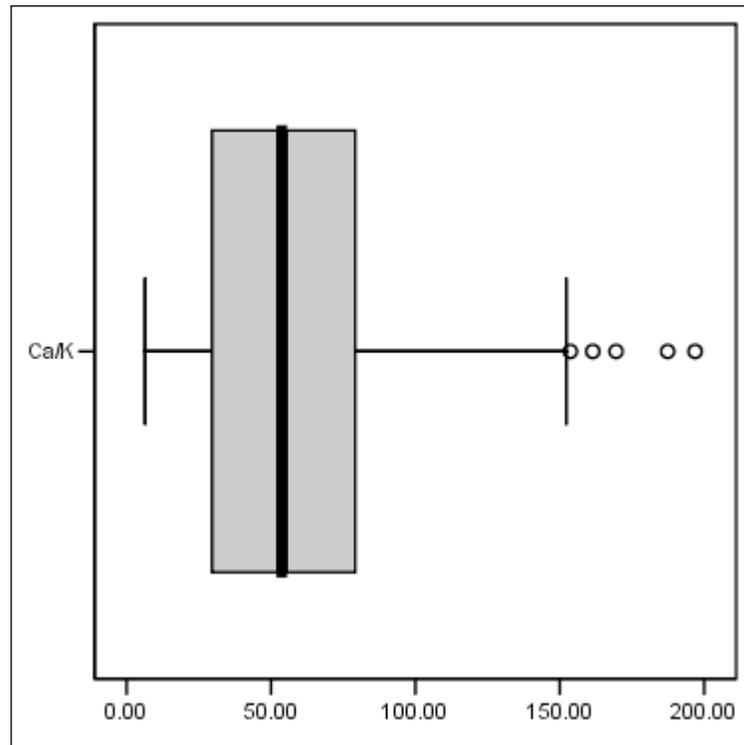


Figura 2.39. Diagrama de caja de la relación Ca-K

El histograma de la Figura 2.40 muestra la distribución empírica de los valores de la relación así como los valores críticos establecidos anteriormente en el Cuadro 2.5.

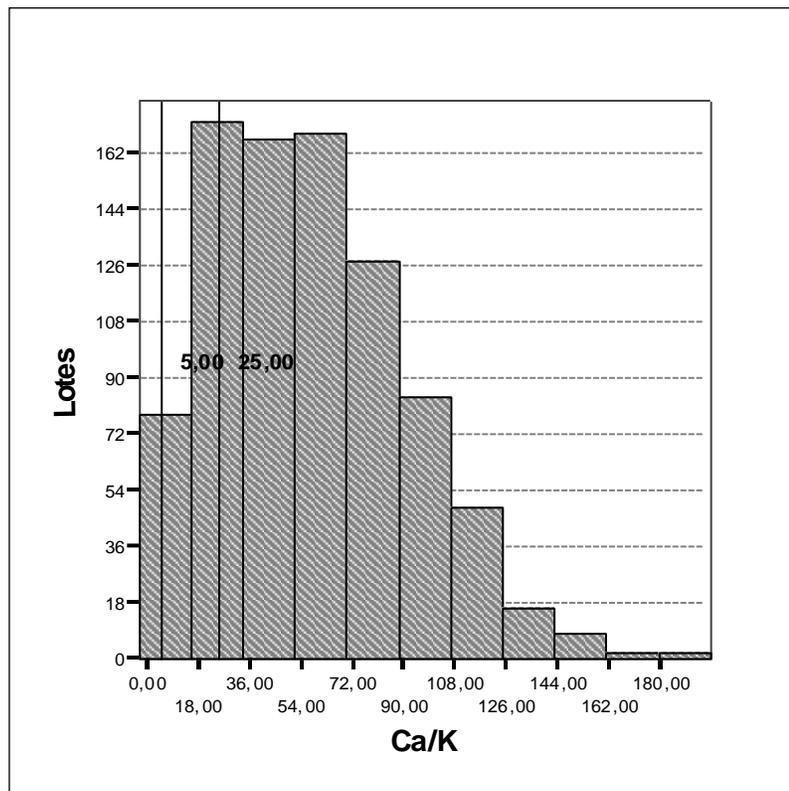


Figura 2.40. Histograma de la relación Ca-K

De acuerdo a la guía para el análisis de suelos mostrada en el Cuadro 2.5, los niveles de la relación pueden clasificarse de acuerdo a las siguientes categorías por intervalos:

- Desbalance por Ca: $>25,0$
- Balance: $(5,0 - 25,0)$
- Desbalance por K: $<5,0$

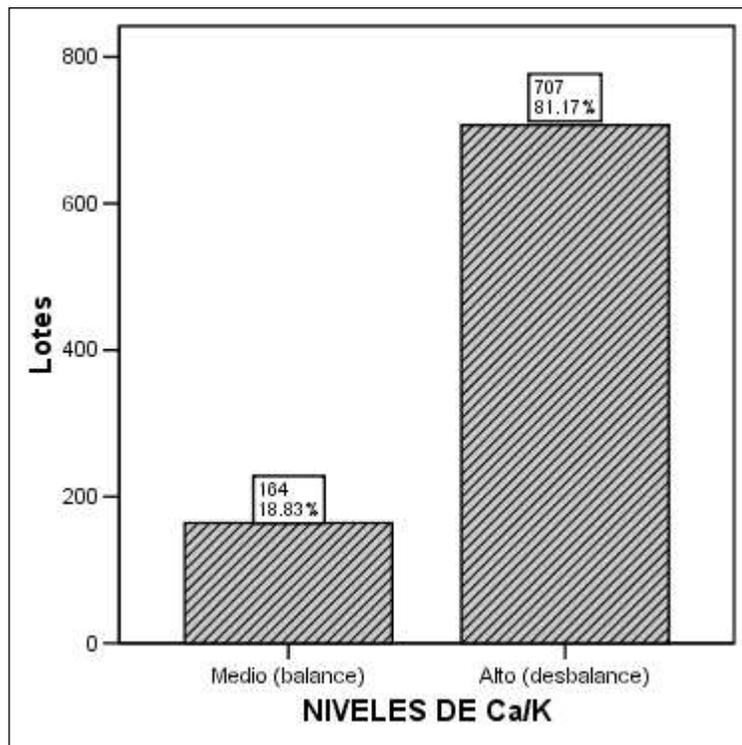


Figura 2.41. Diagrama de barras de la relación Ca-K

En el diagrama de barras de la Figura 2.41 se puede observar que la mayoría de los lotes, es decir, el 81,17% (707 lotes) están en desbalance con respecto al Ca; y, el 18,83% (164 lotes) están en balance.

X_{1,14}: Conductividad eléctrica (mmhos/cm)**Cuadro 2.23.** Estadística descriptiva de Conductividad eléctrica

N	Valid	871
Media		72,0103
Mediana		63,0000
Desviación Estándar		36,40454
Varianza		1325,291
Rango		309,00
Mínimo		25,00
Máximo		334,00
Percentiles	25	50,0000
	50	63,0000
	75	83,0000

La estadística descriptiva del Cuadro 2.23 muestra una media para la Conductividad Eléctrica (CE) de 72,0103 mmhos/cm con un mínimo de 25,00 mmhos/cm y un máximo de 334,00 mmhos/cm. El promedio de las desviaciones con respecto a la media es de 36,40454 mmhos/cm. El diagrama de caja de la Figura 2.42 muestra que el 50% de los lotes presentan valores entre 50,00 y 83,00 mmhos/cm. Los valores del extremo derecho corresponden a lotes con valores altos de CE.

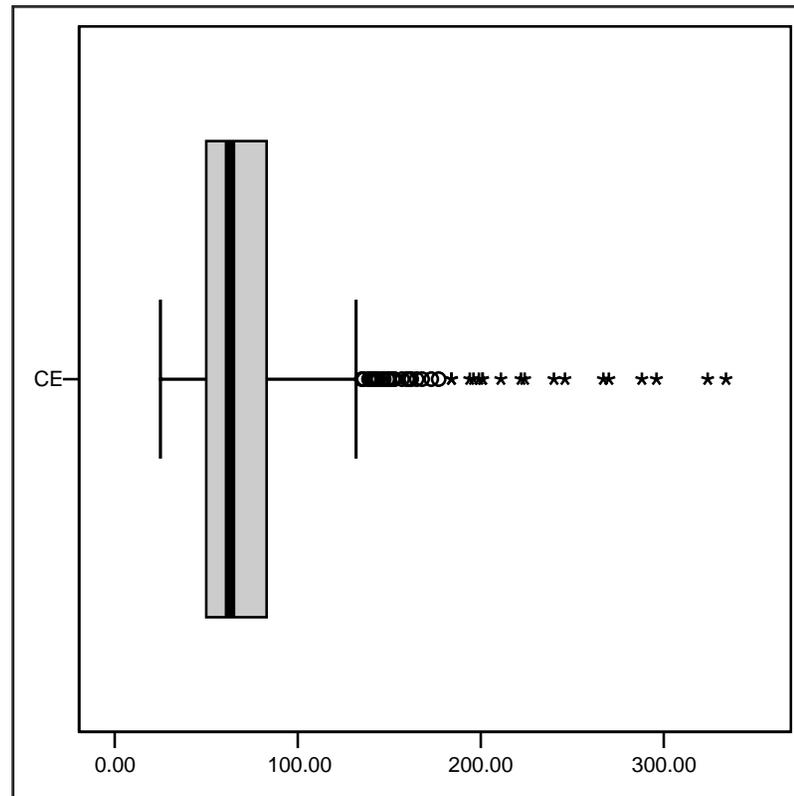


Figura 2.42. Diagrama de caja de Conductividad eléctrica

El histograma de la Figura 2.43 muestra la distribución de los valores de CE en la muestra de lotes. Los valores de la cola derecha de la distribución muestra que un pequeño porcentaje de la muestra de lotes presentan valores altos de CE.

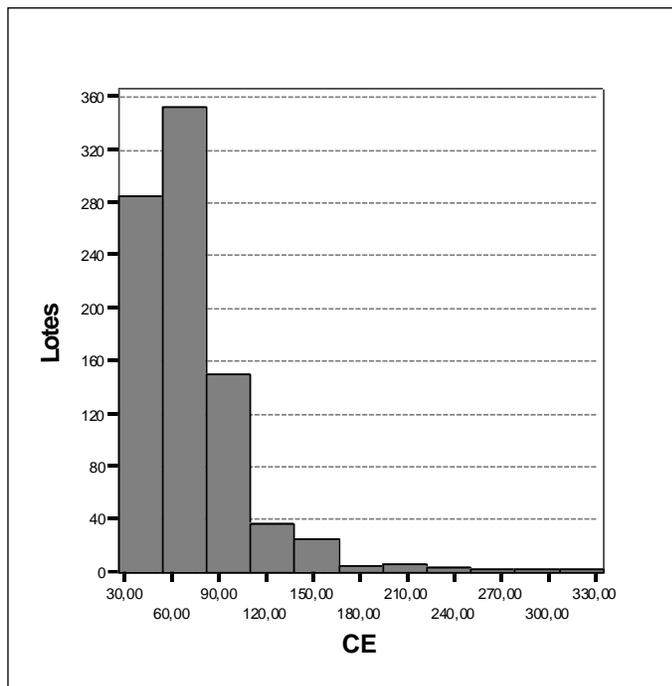


Figura 2.43. Histograma de Conductividad Eléctrica

X_{2,1} : Sector**Cuadro 2.24.** Estadística descriptiva de Sector

SECTOR	Frecuencia de lotes	Porcentaje de lotes
sector 1	124	14,2
sector 2	134	15,4
sector 3	140	16,1
sector 4	197	22,6
sector 5	164	18,8
sector 6	112	12,9
Total	871	100,0

La estadística descriptiva del Cuadro 2.24 y el diagrama de barras de la Figura 2.44 muestran la distribución de los sectores en la muestra. El sector con mayor número de lotes es el sector 4 con el 22,6% de lotes de la muestra; le sigue el sector 5 con el 18,8%, el sector 3 con el 16,1%, el sector 2 con el 15,4%, el sector 1 con el 14,4% y, finalmente, el sector 6 con el 12,9% de lotes de la muestra.

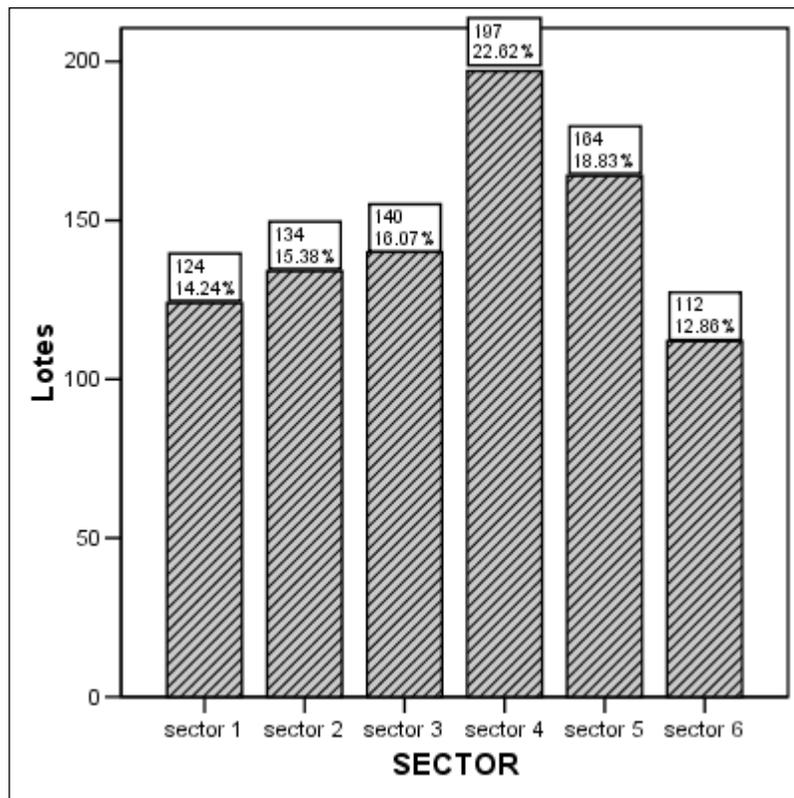


Figura 2.44. Diagrama de barras de Sector

X_{2,2} : Variedad de caña**Cuadro 2.25.** Estadística descriptiva de Variedad de caña

VARIEDAD	Frecuencia de lotes	Porcentaje de lotes
RAGNAR	444	51,0
MISCELANEA	27	3,1
MCO-310	6	,7
BJ-7046	54	6,2
PR-671070	86	9,9
CC-8592	64	7,3
B-74132	14	1,6
CR-74250	176	20,2
Total	871	100,0

La estadística descriptiva del Cuadro 2.25 y el diagrama de barras de la Figura 2.45 muestran la distribución de las variedades de caña en la muestra. La variedad con mayor número de lotes es la RAGNAR con el 51% de lotes de la muestra; le sigue la variedad CR-74250 con el 20,2%, la PR-671070 con el 9,9%, la CC-8592 con el 7,3%, la BJ-7046 con el 6,2% y, finalmente, la MISCELANEA, MCO-310 y B-74132 con el restante 5,4% de lotes de la muestra.

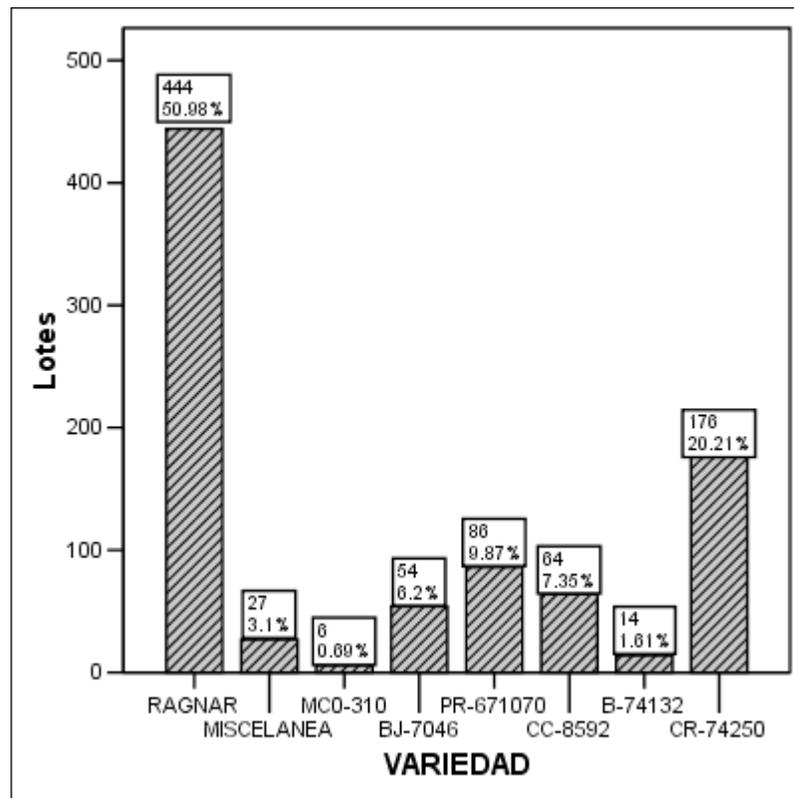


Figura 2.45. Diagrama de barras de Variedad de caña

X_{2,3} : Número de riegos**Cuadro 2.26.** Estadística descriptiva de Número de riegos

N		871
Media		7,2457
Mediana		7,0000
Desviación Estándar		2,70066
Varianza		7,294
Rango		14,00
Mínimo		2,00
Máximo		16,00
Percentiles	25	5,0000
	50	7,0000
	75	9,0000

La estadística descriptiva del Cuadro 2.26 muestra una media para el número de riegos de 7,2457 riegos con un mínimo de 2 riegos y un máximo de 16 riegos. El promedio de las desviaciones con respecto a la media es de 2,7 riegos. El diagrama de caja de la Figura 2.46 muestra que el 50% de los lotes presentan valores entre 5 y 9 riegos. Los valores del extremo derecho corresponden a lotes con valores extremos de número de riegos.

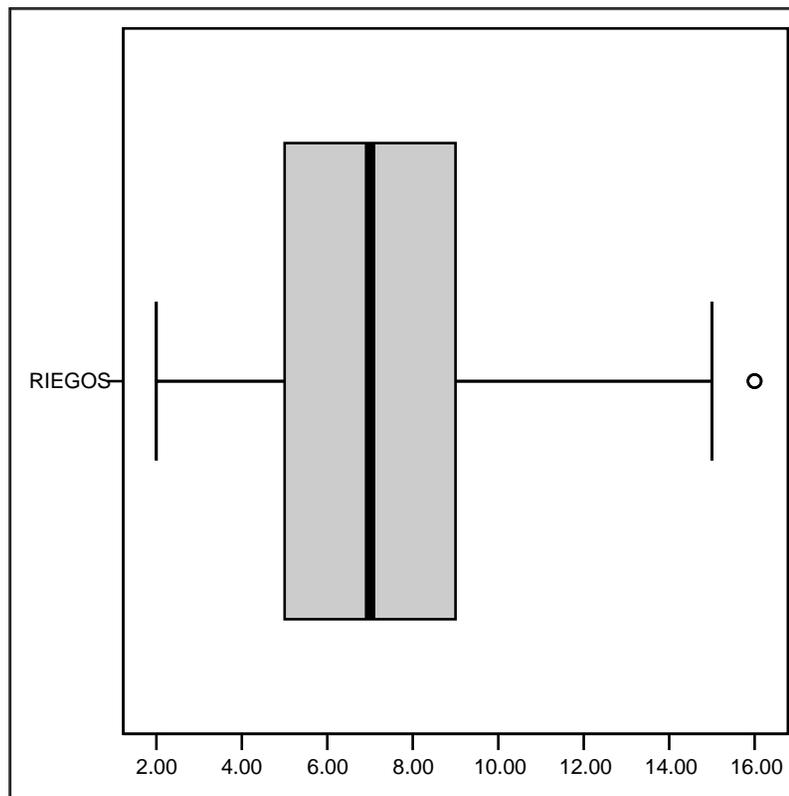


Figura 2.46. Diagrama de caja de Número de riegos

El diagrama de barras de la Figura 2.47 describe la distribución muestral del número de riegos. La mayor frecuencia de riegos es 5, la menor es 15 y 16 en dos lotes para cada frecuencia.

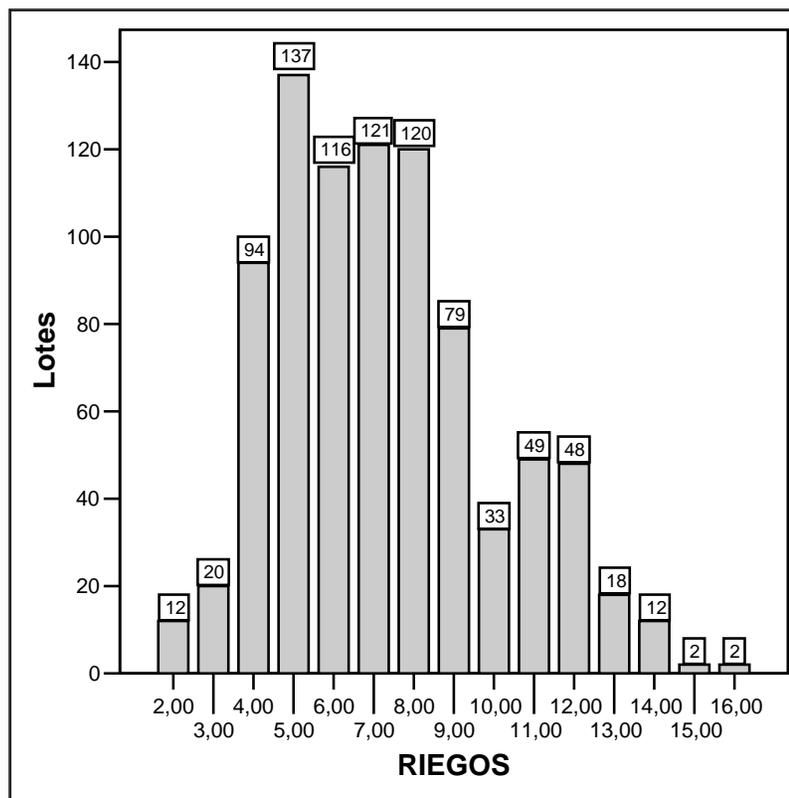


Figura 2.47. Diagrama de barras de Número de riegos

CAPÍTULO 3

3. MÉTODOS ESTADÍSTICOS PARA DATOS DE AGRICULTURA DE PRECISIÓN: ANÁLISIS ECONOMETRICO DE PROCESOS ESPACIALES

En el presente capítulo se expone el marco teórico de una de las técnicas estadísticas de mayor aplicación en muchas áreas científicas y de negocios y cuyo objetivo principal es el de especificar modelos matemáticos que permitan analizar características o propiedades de una variable objetivo utilizando como causas explicativas otras variables. A esta técnica se la conoce como análisis de regresión lineal múltiple o **análisis econométrico** y es introducida en esta investigación, principalmente, como una técnica que permite describir y evaluar la respuesta sitio-específica de un cultivo a los diferentes métodos y tecnologías de AP.

Tal como se expuso en el Capítulo 1, la mayoría de la literatura e investigaciones realizadas que muestran y utilizan análisis econométricos corresponden a teoría y aplicaciones de series temporales de datos; por lo que, en el trabajo aplicado, no se ha prestado mucha atención a ciertas características particulares de las muestras que involucran datos con una referencia espacial: por ejemplo, toda aquella información sobre la producción de un cultivo que se encuentra agregada por unidades de manejo agrícola (lotes, canteros, parcelas, entre otros) dentro de una finca.

La información agregada por unidades de observación en el espacio está caracterizada fundamentalmente por dos aspectos: *dependencia espacial* (autocorrelación espacial) entre las observaciones y *heterogeneidad espacial* (estructura espacial). Los ya mencionados *efectos espaciales*, son importantes en el análisis econométrico aplicado en la medida en que hacen que ciertas metodologías y resultados estándar resulten inválidos, demandando la adaptación de otros procedimientos, y, en algunos contextos, haciendo necesario el desarrollo de un conjunto especializado de técnicas. Estas circunstancias son frecuentemente ignoradas por la econometría tradicional y han sido enmarcadas en una rama específica de la econometría denominada ***econometría espacial***.

El análisis de datos de AP ha encontrado un sólido pilar de apoyo en esta rama de la econometría. Muchos resultados que han aparecido en una variedad de fuentes se irán juntando a medida que se avance en este capítulo y en el siguiente. A continuación, se comenzará por una explicación del modelo general de regresión lineal [21, 22 y 23], sus supuestos básicos, procedimientos de estimación, inferencia y algunos problemas elementales que deben ser considerados. Luego, se expondrá el modelo espacial de regresión lineal múltiple [13], los fundamentos para el análisis econométrico de procesos espaciales, estimación, pruebas de hipótesis y procedimientos para la validación de modelos.

3.1. EL MODELO GENERAL DE REGRESIÓN LINEAL MÚLTIPLE

En la práctica de muchas ciencias se realizan numerosos intentos para resolver una gran cantidad de problemas sobre la base de una riqueza en formulaciones teóricas que tratan de explicar el comportamiento humano o de ciertos de sistemas naturales. Estas teorías, para ser consideradas operacionales, necesitan ser traducidas de formulaciones abstractas a modelos implementables.

Esto implica que conceptos y relaciones necesitan ser expresados en especificaciones matemáticas formales; que se debe dar sentido a cada una de las variables investigadas dentro del contexto de los datos disponibles; y, que deben llevarse a cabo procedimientos de estimación, pruebas de hipótesis y predicción. Esto se realiza, comúnmente, sobre una metodología estadística o análisis econométrico.

La econometría conlleva, en gran parte, la utilización de modelos de **regresión lineal múltiple** para el estudio de la relación entre una variable llamada **dependiente**, en una o más variables denominadas **explicativas**, con la finalidad de:

- a) Especificar un modelo de relación entre variables,
- b) Utilizar información muestral acerca de los valores tomados por dichas variables con el objetivo de cuantificar la magnitud de dependencia entre ellas,
- c) Evaluar críticamente la validez de hipótesis propuestas por una teoría subyacente acerca de las relaciones estimadas; y, en algunos casos,

- d) Efectuar un ejercicio de seguimiento coyuntural y de previsión de las variables analizadas.

Así, al utilizar la econometría para datos de AP en esta investigación tratamos de responder preguntas fundamentales como:

1. Habiendo analizado de antemano cuánto varían las características del suelo y del cultivo, ¿cuánto afectan estas variaciones al rendimiento?
2. Sobre la base de información muestral, ¿cuál es la importancia cuantitativa de cada uno de los factores del rendimiento?
3. ¿Podemos contrastar algunas de las implicaciones del conocimiento agronómico acerca del efecto que variables como la textura del suelo tienen sobre la producción de caña de azúcar?
4. ¿Qué sugiere el modelo estimado acerca del comportamiento del rendimiento para una edad determinada del cultivo?

El investigador debe comenzar especificando muy claramente cuál es el centro de atención del trabajo empírico; luego, se deben tratar de identificar cuáles son los factores que inciden sobre el

comportamiento del rendimiento de un cultivo; se debe escoger cuidadosamente la información estadística relevante para cuantificar tal relación, y se debe proceder, finalmente, a su cuantificación. Por último, se utilizará el modelo de relación estimado, ya sea a efectos de contrastación de algún tipo de hipótesis o supuesto teórico, o como elemento de análisis y seguimiento de las variables que se escogieron en el estudio

Un aspecto importante a tener en cuenta es que a pesar de que el análisis de regresión tiene que ver con la dependencia de una variable de otras, esto no implica causalidad necesariamente. Para el caso del producto de una cosecha, no hay una *razón estadística* para suponer que la lluvia no depende del producto de la cosecha. El hecho de que se trata el producto de la cosecha como dependiente de la lluvia (entre otros factores) es debido a consideraciones no estadísticas: en el sentido de que la relación no puede revertirse, ya que no podemos controlar la lluvia modificando la producción de la cosecha. De aquí que, cuando se especifique un modelo econométrico para el producto de una cosecha, debe entenderse que una teoría agronómica subyacente aporta suficientes elementos como para sugerir que las

variables explicativas (factores de producción o rendimiento) influyen sobre la producción, y no al revés.

Por otra parte, el análisis econométrico se centra específicamente en lo que se conoce como **dependencia estadística** entre variables, y no en aquella dependencia funcional o determinística propia de la física clásica. En las relaciones estadísticas entre variables tratamos esencialmente con variables **aleatorias o estocásticas**, esto es, variables que tienen distribuciones de probabilidad. En la dependencia funcional o determinística también se manejan variables, pero estas no son aleatorias o estocásticas.

La dependencia del producto de una cosecha de la temperatura ambiente, la lluvia, el sol y los fertilizantes, por ejemplo, es de naturaleza estadística en el sentido que las variables explicativas, si bien son importantes, no permitirán al agrónomo predecir en forma exacta el producto de la cosecha debido a los errores involucrados en la medición de estas variables y en razón de otra serie de factores (variables), que afectan colectivamente la producción pero pueden ser difíciles de identificar individualmente. De esta manera existirá alguna

variabilidad “intrínseca” o aleatoria en la variable dependiente, producto de la cosecha, que no puede ser explicada en su totalidad sin importar cuántas otras variables explicativas consideremos. De aquí que un modelo estocástico resulta ideal para explicar el comportamiento de procesos naturales como el producto de una cosecha, sistema sobre el cual no se tiene un control o conocimiento absoluto.

3.1.1. Función de Regresión Poblacional (FRP)

Especificamos la relación de causalidad entre el producto promedio de la cosecha de un lote dado uno de sus factores en la siguiente ecuación:

$$E(y/x_i) = f(x_i) \quad (3.1)$$

donde $E(y/x_i)$ o **expectativa condicional** representa el promedio poblacional de la variable dependiente dada una variable explicativa x y el subíndice i denota el número de la observación. A la expresión

(3.1) se la denomina **función de regresión poblacional (FRP)** y su forma funcional está dispuesta por una teoría subyacente. Como una primera aproximación o hipótesis de trabajo, se mantiene el supuesto que la FRP es lineal en sus parámetros, es decir:

$$E(y/x_i) = \beta_1 + \beta_2 x_i \quad (3.2)$$

donde β_1 y β_2 son parámetros no conocidos pero fijos que se denominan **coeficientes de regresión** o también **intercepto** y **coeficiente de pendiente**, respectivamente. De esta manera se postula la existencia de relaciones simultáneas entre variables con el objetivo de estimar y/o predecir la media o valor promedio poblacional de la variable dependiente en términos de los valores conocidos o fijos (en muestras repetidas) de las variables explicativas.

Para introducir la variabilidad intrínseca del fenómeno en la especificación de un modelo de regresión expresamos a la variable dependiente como:

$$y_i = E(y/x_i) + u_i \quad (3.3)$$

donde ahora se expresa la producción de un lote particular en términos de dos componentes: (1) $E(y/x_i)$ o **componente sistemático** (determinístico) que es el valor promedio del producto de la cosecha de todos los lotes correspondientes a un nivel del factor x_i ; y, (2) u_i que es el componente aleatorio, o **no sistemático**.

La desviación u_i es una variable aleatoria no observable que toma valores negativos o positivos. Técnicamente, u_i es conocida como **perturbación estocástica** o **término de error estocástico** y es un sustituto para todas las variables que son omitidas del modelo pero que, colectivamente, afectan a y . Las razones más importantes para la presencia del error estocástico son:

1. **Vaguedad de la teoría:** De existir una teoría que determine el comportamiento de la variable dependiente, ésta puede estar incompleta y frecuentemente lo está. Podríamos saber con seguridad que las lluvias afectan al producto de una cosecha, pero se podría ignorar o no estar seguros sobre algunas otras variables que afectan a la producción.

2. **Falta de disponibilidad de información:** Aún si se sabe cuáles son algunas de las variables excluidas y se considera por consiguiente una regresión múltiple, es posible que no se pueda obtener información cuantitativa sobre algunas de éstas. Por ejemplo, en un principio se podría introducir la población de cierta plaga como una variable explicativa adicional a la cantidad de lluvias, fertilizantes, entre otras, para explicar la producción de un cultivo. Sin embargo, desafortunadamente, no se dispone de la información requerida al nivel de agregación deseado (Ej: por lotes). Por consiguiente, forzosamente se omite la variable en discusión a pesar de su importancia teórica en la explicación de la producción.
3. **Aleatoriedad intrínseca del proceso de producción:** Aún si se tiene éxito en la introducción en el modelo de todas las variables relevantes, es posible que exista alguna aleatoriedad “intrínseca” en la producción que no puede ser explicada a pesar de todos los esfuerzos que se hagan. Las perturbaciones u_i pueden reflejar muy bien esta aleatoriedad intrínseca derivada por el efecto global de un conjunto de variables no incluidas en el modelo.

4. **Errores de medición y variables próximas o “proxy”**

inadecuadas: A pesar de que el modelo de regresión supone que las variables son medidas de forma precisa, en la práctica, los datos pueden estar plagados de errores de medición. En muchas ocasiones, al no obtener información sobre una variable requerida, acudimos a las llamadas variables “proxy” en un intento de aproximar los valores de la verdadera variable de interés. En estos casos, el término de perturbación u puede también representar estos errores de medición; sin embargo, la utilización de aproximaciones puede dar origen a graves implicaciones en la estimación de los coeficientes de regresión.

5. **Nivel de agregación de los datos:** En muchas situaciones encontradas en la práctica, los datos son recolectados sólo a una escala agregada, es decir, por ejemplo, por unidades censales, distritos, unidades de manejo agrícola (parcelas, lotes), etc. De aquí que, es muy probable que exista poca correspondencia entre el comportamiento espacial del fenómeno bajo estudio y la delimitación ad hoc de las unidades espaciales bajo observación. Como consecuencia, los errores de medida son muy probables y

como se expondrá posteriormente, esta situación conduce a la autocorrelación espacial de los errores.

6. **Principio de parsimonia:** Si podemos explicar “sustancialmente” el comportamiento de nuestra variable dependiente con dos o tres variables explicativas y si nuestra teoría no es lo suficientemente fuerte para sugerir otras variables que pudieran ser incluidas, tal vez lo más aconsejable es no introducir más variables. Se puede permitir que u_i represente todas las demás variables.
7. **Forma funcional incorrecta:** Aún si se tienen variables teóricamente correctas para explicar un fenómeno y se puede obtener información sobre ellas, frecuentemente no se conoce la forma de la relación funcional entre la variable dependiente y las variables explicativas. En los modelos de regresión de dos variables la forma funcional de la relación puede ser juzgada a partir de un diagrama de dispersión. Pero en un modelo de regresión múltiple no es fácil determinar la forma funcional apropiada, puesto que no se pueden visualizar gráficamente los diagramas de dispersión en dimensiones múltiples.

Todas estas razones hacen que las perturbaciones estocásticas u_i asuman un papel extremadamente crítico en el análisis de regresión, específicamente, en el momento de efectuar algún tipo de inferencia. Esta situación se tratará con más profundidad a medida que se progrese en la exposición del capítulo.

3.1.2. Notación matricial

Esta investigación se enfoca, principalmente, en la especificación de un modelo de relación entre varias variables agronómicas, representado como:

$$y = f(x_1, x_2, x_3, \dots, x_k, \mu / \beta) \quad (3.4)$$

que trata de explicar el comportamiento del rendimiento o producción de un cultivo representado en el vector y utilizando la información proporcionada por un conjunto de k variables explicativas o factores de la producción, así como por una variable aleatoria, no observable sin un claro significado agronómico, que denotaremos por u .

El hecho de existir más de dos variables dentro de la especificación de un modelo, hace que la notación matricial resulte la más adecuada y práctica al momento de simplificar muchos cálculos y resultados. De aquí que, tomando como base la expresión en (3.4), para una muestra de tamaño n obtenemos la siguiente especificación utilizando notación matricial:

$$\begin{bmatrix} \cdot \\ \cdot \\ y \\ \cdot \\ \cdot \end{bmatrix} = \beta_1 \begin{bmatrix} \cdot \\ \cdot \\ x_1 \\ \cdot \\ \cdot \end{bmatrix} + \beta_2 \begin{bmatrix} \cdot \\ \cdot \\ x_2 \\ \cdot \\ \cdot \end{bmatrix} + \dots + \beta_k \begin{bmatrix} \cdot \\ \cdot \\ x_k \\ \cdot \\ \cdot \end{bmatrix} + \begin{bmatrix} \cdot \\ \cdot \\ \mu \\ \cdot \\ \cdot \end{bmatrix} \quad (3.5)$$

donde el vector y está expresado como una combinación lineal de los vectores x más el vector de errores estocásticos u . El vector x_1 es una columna de unos para permitir la inclusión del término de intercepto. Matricialmente podemos expresar el modelo como:

$$y = X\beta + \mu$$

(3.6)

donde:

$$X = \begin{bmatrix} 1 & X_{21} & \cdot & \cdot & X_{k1} \\ 1 & X_{22} & \cdot & \cdot & X_{k2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{2n} & \cdot & \cdot & X_{kn} \end{bmatrix} \quad y \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}$$

3.1.3. Función de Regresión Muestral (FRM)

En el análisis de regresión la tarea principal es estimar una FRP como aquella mostrada en la expresión (3.2); es decir, estimar los valores de β no conocidos utilizando las observaciones de y y X contenidas en una muestra. La contraparte muestral de la FRP se denomina **función de regresión muestral (FRM)** y puede escribirse como:

$$\hat{E}(y/X) = \hat{y} = \hat{\beta}X \quad (3.7)$$

donde:

\hat{y} = vector de estimaciones de $E(y/X)$; y ,

$\hat{\beta}$ = vector de estimadores de β

El procedimiento de estimación consiste en utilizar información muestral para asignar valores numéricos a los parámetros desconocidos del modelo de la FRP. Debe notarse que un estimador es, por tanto, una función del espacio muestral (el conjunto de todas las observaciones posibles que pudieron haberse tenido de las variables dependiente y explicativas) sobre el espacio paramétrico (el conjunto de todos los valores admisibles de los parámetros). De aquí que, un **estimador**, conocido también como **estadístico** (muestral), es simplemente una regla, fórmula o método que dice cómo estimar el parámetro poblacional a partir de la información suministrada por la muestra disponible. Un valor numérico particular obtenido por el estimador en una población es conocido como un **estimado**.

Tal como se expresa para la FRP en la expresión (3.3), también se puede expresar la FRM en forma estocástica de la siguiente manera:

$$y = \hat{\beta}X + e \quad (3.8)$$

donde, adicionalmente a los símbolos ya definidos, e es un vector que contiene a los **residuales**. Conceptualmente e es análogo a u y puede ser considerado como un estimado de u , el cual, se introduce en la FRM por las mismas razones que u fue introducida en la FRP.

La pregunta crítica es ahora: dado que la FRM es apenas una aproximación de la FRP, ¿se puede diseñar una regla o método que haga que esta aproximación sea lo más “ajustada” posible?. En otras palabras, ¿cómo se debe construir la FRM para que los valores del vector $\hat{\beta}$ estén lo más “cerca” de los verdaderos valores de β como sea posible aún cuando no se lleguen a conocer los verdaderos valores de β ?

La respuesta a esta pregunta ocupará gran parte de éste y el siguiente capítulo. Se mostrará que es posible desarrollar procedimientos que dicen cómo construir a FRM para reflejar la FRP tan fielmente como sea posible.

3.1.4. Supuestos clásicos del modelo

Los supuestos en los cuales se basa el modelo clásico de regresión lineal son los siguientes:

- I. $E(u) = 0$ donde u y 0 son vectores columna $n \times 1$, siendo 0 un vector nulo (Ec. 3.9).

$$E \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \cdot \\ \cdot \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \quad (3.9)$$

II. $E(uu') = \sigma^2 I$ donde I es una matriz identidad $n \times n$ (Ec. 3.10). Esta es la matriz simétrica definida positiva de varianzas-covarianzas del vector u . En notación escalar esto significa que $E(u_i u_j) = 0$ si $i \neq j$ y es igual a σ^2 si $i = j$. Este supuesto acoge simultáneamente dos hipótesis:

- a) Se supone que la matriz es escalar, es decir, la diagonal principal es constante. Los elementos de dicha diagonal son las varianzas de los términos de error de las distintas observaciones; por lo que, en consecuencia, estamos suponiendo **homoscedasticidad**: la varianza del término del error es constante, ya sea a lo largo del tiempo o a través de la sección cruzada.
- b) Se supone que los elementos de fuera de la diagonal de la matriz son iguales a cero. Estos elementos son las covarianzas entre los términos del error correspondientes a observaciones distintas. Al suponer que sus covarianzas son iguales a cero, estamos afirmando que dichos términos de error no están correlacionados entre sí, es decir, no existe **autocorrelación** entre las

perturbaciones, lo que implica que los errores están distribuidos independientemente unos de los otros.

$$\begin{aligned}
 E(uu') &= \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \dots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \dots & E(u_2u_n) \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ E(u_nu_1) & E(u_nu_2) & \dots & E(u_n^2) \end{bmatrix} = \begin{bmatrix} \text{Var}(u_1) & \text{Cov}(u_1u_2) & \dots & \text{Cov}(u_1u_n) \\ \text{Cov}(u_2u_1) & \text{Var}(u_2) & \dots & \text{Cov}(u_2u_n) \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \text{Cov}(u_nu_1) & \text{Cov}(u_nu_2) & \dots & \text{Var}(u_n) \end{bmatrix} \\
 &= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I_n \quad (3.10)
 \end{aligned}$$

- III. La matriz X , $n \times k$, es no estocástica; es decir, consiste en un conjunto de números fijos.
- IV. El rango de la matriz X es $p(X) = k$, donde k es el número de columnas de X y k es menor que el número de observaciones. Este supuesto implica que no hay relación lineal exacta entre las variables x , es decir, no existe **multicolinealidad**. Esta es una de las características menos estrictas de las que aquí se describen, pues excluye tan sólo la posibilidad de que alguna de las variables explicativas del modelo econométrico pueda escribirse como

combinación lineal exacta de las demás. En realidad, todas las variables agronómicas muestran algún grado de correlación entre sí, y ello no produce excesivas dificultades, excepto cuando se llega a una situación de dependencia total, lo que se excluye al afirmar que las variables explicativas no son linealmente dependientes entre sí.

V. El vector u tiene una distribución normal multivariada, es decir,

$$u \sim N_n(0_n, \sigma^2 I_n).$$

VI. Los coeficientes del modelo son constantes a través de la serie temporal o sección cruzada.

3.1.5. Estimación de Mínimos Cuadrados Ordinarios (MCO)

Si el vector desconocido β en la expresión (3.6) es reemplazado por un estimado $\hat{\beta}$, este definirá un vector de residuales:

$$e = y - X\hat{\beta} \tag{3.11}$$

El principio de mínimos cuadrados es el de elegir un vector $\hat{\beta}$ que minimice la suma de residuales al cuadrado, $e'e$, o **norma euclídea** del vector de residuales; es decir:

$$\begin{aligned}
 SR(\hat{\beta}) &= e'e \\
 &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\
 &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\
 &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}
 \end{aligned} \tag{3.12}$$

Luego, la teoría de optimización matemática permite resolver el problema:

$$\min_{\hat{\beta}} SR(\hat{\beta}) = \min_{\hat{\beta}} (y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}) \tag{3.13}$$

Utilizando los resultados acerca de derivadas matriciales se tiene, derivando respecto al vector $\hat{\beta}$ en la expresión (3.13), que:

$$\frac{\partial SR(\hat{\beta})}{\partial \hat{\beta}} = \frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} \tag{3.14}$$

La solución al problema de minimización de $SR(\hat{\beta})$ requiere, en primer lugar, que el vector gradiente en (3.14) sea igual a cero, es decir, que:

$$(X'X)\hat{\beta} = X'y \quad (3.15)$$

Además, debe cumplirse que la matriz de segundas derivadas o matriz hessiana de $SR(\hat{\beta})$ sea definida positiva. Pero dicha matriz es:

$$\frac{\partial^2 SR(\hat{\beta})}{\partial \hat{\beta} \partial \hat{\beta}'} = X'X \quad (3.16)$$

que es siempre semidefinida positiva.

Puesto que $X'X$ es una matriz $k \times k$ y $X'y$ un vector $k \times 1$ la ecuación matricial (3.15) es, en realidad, un sistema de k ecuaciones lineales en los k coeficientes desconocidos $\beta_1, \beta_2, \dots, \beta_k$. Este sistema se denomina **sistema de ecuaciones normales** y tiene, generalmente, una única solución. Dicha solución es el **estimador de mínimos cuadrados ordinarios** del vector β , que denotaremos en lo sucesivo por **estimador MCO** y que se expresa como:

$$\hat{\beta} = (X'X)^{-1} X'y \quad (3.17)$$

3.1.5.1. Propiedades del estimador MCO

Las propiedades del estimador MCO son las siguientes:

1. El estimador MCO de β es un vector aleatorio, pues, dependiendo del vector de observaciones de la variable dependiente y , depende también del término de error u .

$$\hat{\beta} = (X'X)^{-1} X'y = (X'X)^{-1} X'(X\beta + u) = \beta + (X'X)^{-1} X'u \quad (3.18)$$

2. Si $E(u) = 0_n$ (supuesto I), entonces el estimador MCO es insesgado, es decir, $E(\hat{\beta}) = \beta$. En efecto, tomando esperanzas en (3.15) se tiene:

$$E(\hat{\beta}) = E[\beta + (X'X)^{-1} X'u] = \beta + (X'X)^{-1} X'E(u) = \beta \quad (3.19)$$

3. Si $Var(u) = \sigma^2 I_n$ (supuesto II), la matriz de varianzas-covarianzas del estimador MCO es igual a $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$. Nótese que la expresión (3.18) puede utilizarse para definir el error de estimación:

$$\hat{\beta} - \beta = (X'X)^{-1} X'u$$

que utilizamos para demostrar que:

$$\begin{aligned} Var(\hat{\beta}) &= E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \\ &= E[(X'X)^{-1} X'uu'X(X'X)^{-1}] = (X'X)^{-1} X'E(uu')X(X'X)^{-1} = \\ &= (X'X)^{-1} X'(\sigma^2 I_n)X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \end{aligned} \quad (3.20)$$

4. Cada una de las variables explicativas es ortogonal al vector de residuos mínimo cuadráticos, es decir:

$$X'e = X'(y - X\hat{\beta}) = X'y - X'X(X'X)^{-1}X'y = X'y - X'y = 0_k \quad (3.21)$$

5. Partiendo de la expresión (3.21), si hay término independiente o intercepto en la regresión, entonces la suma de los residuos mínimo cuadráticos es cero:

$$x_{1n} = 1 \quad \forall n \Rightarrow \sum_1^n e_n = 0$$

6. La combinación lineal $c'\hat{\beta}$ es el estimador lineal insesgado de mínima varianza de la combinación lineal de parámetros $c'\beta$.
7. **Teorema de Gauss-Markov:** El estimador MCO es el estimador lineal insesgado óptimo, en el sentido de que cualquier otro estimador lineal e insesgado tiene una matriz de covarianzas “mayor” que la del estimador MCO.

Demostración:

Sea $\tilde{\beta} = \tilde{A}y$ un estimador lineal de β , donde \tilde{A} es una matriz $k \times n$.

Denotemos por A (matriz $k \times n$) la diferencia $A = \tilde{A} - (X'X)^{-1}X'$, de modo que:

$$\tilde{\beta} = [A + (X'X)^{-1}X']y = [A + (X'X)^{-1}X'](X\beta + u) = AX\beta + \beta + [A + (X'X)^{-1}X']u$$

y, por tanto, $E(\tilde{\beta}) = AX\beta + \beta$. El estimador $\tilde{\beta}$ será insesgado sólo si la matriz A es tal que $AX = 0_{k \times k}$. Con esta condición, el estimador $\tilde{\beta}$ resulta:

$$\tilde{\beta} = \beta + [A + (X'X)^{-1}X']u$$

y su matriz de covarianzas será:

$$\begin{aligned} Cov(\tilde{\beta}) &= E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] = E\{([A + (X'X)^{-1}X']u)([A + (X'X)'X']u)'\} = \\ &= \sigma^2 AA' + \sigma^2 (X'X)^{-1} \end{aligned}$$

donde se ha utilizado la condición de ausencia de sesgo $AX = 0_{k \times k}$. Como la matriz AA' es semidefinida positiva, se concluye que la diferencia entre las matrices de covarianzas de $\tilde{\beta}$ y $\hat{\beta}$ es una matriz semidefinida positiva, por lo que la primera es igual, si no mayor, que la segunda.

8. El vector de residuos mínimo-cuadráticos es una transformación lineal del vector término de error. En efecto:

$$e = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = My = Mu \quad (3.22)$$

donde $M = I_n - X(X'X)^{-1}X'$ es una matriz singular, simétrica e idempotente, y donde la última igualdad viene de $MX = 0_{n \times k}$.

De esta propiedad se deduce inmediatamente otra expresión para la suma residual:

$$SR = e'e = y'My = (X\beta + u)'M(X\beta + u) = u'Mu$$

ya que $MX = 0_{n \times k}$

9. Si $u \sim N_n(0_n, \sigma^2 I_n)$ entonces $\hat{\beta} \sim N_k(\beta, \sigma^2 (X'X)^{-1})$

3.1.5.2. Suma de cuadrados y su descomposición

La expresión (3.21) de donde se derivan las propiedades 4 y 5 de los estimadores MCO permite la descomposición de la suma de cuadrados de los valores de la variable endógena, siempre que exista un término independiente en la regresión. Descomponiendo el vector y en una parte explicada por la regresión y una parte no explicada,

$$y = \hat{y} + e = X\hat{\beta} + e$$

se tiene que:

$$y'y = (y + e)'(y + e) = y'y + e'e = \hat{\beta}'X'X\hat{\beta} + e'e$$

donde $y'y = \sum_1^n y_i^2$ es la suma de cuadrados de los valores actuales de la variable endógena. Sin embargo, el interés normalmente se centra en analizar la variación en y medida por la suma del cuadrado de sus desviaciones con respecto a la media muestral, es decir:

$$\sum_1^n (y_i - \bar{y})^2 = \sum_1^n y_i^2 - n\bar{y}^2$$

por lo que, substrayendo $n\bar{y}^2$ de cada lado de la descomposición previa llegamos a la expresión:

$$(y'y - n\bar{y}^2) = (\hat{\beta}'X'X\hat{\beta} - n\bar{y}^2) + e'e$$

$$ST = SE + SR$$

donde ST indica la suma total de cuadrados en y , SE la explicada por la regresión y SR la suma de cuadrados de los residuos (variación no explicada).

El resultado es interesante puesto que es conforme a la intuición acerca de la utilización de un modelo econométrico: se pretende explicar la variable y ; el modelo es capaz de generar la variable \hat{y} , que dictamina la parte de y que el modelo explica, mientras que el vector de error e determina el tamaño de la componente de y no explicada por el modelo. Pues bien, cuando hay un término independiente en el modelo econométrico, entonces las sumas explicada y no explicada (o residual) se agregan hasta ser exactamente igual a la suma a explicar (suma total).

Para mostrar el grado de poder explicativo de un modelo lineal se define el **coeficiente de determinación múltiple** como la cantidad:

$$R^2 = 1 - \frac{SR}{ST} \quad (3.23)$$

mientras que su raíz cuadrada positiva, cuando existe, se denomina **coeficiente de correlación lineal** entre y y \hat{y} . El coeficiente de determinación siempre es menor o igual que 1 y cuando una de las variables explicativas es constante, entonces $R^2 \geq 0$ y el estadístico puede expresarse como:

$$R^2 = \frac{SE}{ST}$$

Cuando no hay un término independiente en el modelo, entonces el coeficiente de determinación puede tomar cualquier valor, siempre menor o igual a 1, pero incluyendo en su rango todos los reales negativos. Por supuesto que cuando el coeficiente de determinación es negativo, no cabe hablar del coeficiente de correlación lineal, definido como $\sqrt{R^2}$.

En la literatura econométrica se ha sugerido tradicionalmente la utilización del estadístico conocido R^2 **corregido**, que se define como:

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k} \right) (1 - R^2)$$

El interés de este estadístico reside en que, cuando el número de variables explicativas k aumenta, la fracción $\frac{n-1}{n-k}$ también aumenta, mientras que $1-R^2$ disminuye, ya que el coeficiente de determinación R^2 aumenta. Como en la definición del \bar{R}^2 aparece el producto de estos dos factores, la idea es que ambos efectos, el creciente y el decreciente, se compensen aproximadamente, por lo que este estadístico es una medida de la bondad de ajuste de un modelo econométrico con la propiedad de ser neutral frente a la introducción de variables adicionales. Este estadístico está concebido, por tanto, para la comparación de modelos anidados.

3.1.5.3. Estimación de la varianza del error

La matriz de varianzas-covarianzas del estimador MCO es importante por las siguientes razones:

1. Debido a que sólo disponemos de una muestra para llevar a cabo la estimación del vector β entonces es crucial conocer la matriz de varianzas-covarianzas del estimador para poder juzgar la exactitud

con la que la estimación obtenida se aproxima a su esperanza matemática $E(\hat{\beta}) = \beta$.

2. Por otra parte, la matriz es necesaria para conocer las varianzas y covarianzas de cada elemento del vector $\hat{\beta}$, de modo que podamos hacer contrastes de hipótesis acerca de valores individuales, o de varios coeficientes, o de combinaciones lineales entre ellos.

Sin embargo, al ser el parámetro σ^2 desconocido, la matriz de covarianzas de β también es desconocida. Sólo podemos aspirar a estimar dicha matriz, estimando previamente el parámetro σ^2 y multiplicando cada estimación por la matriz $(X'X)^{-1}$.

Es razonable elaborar un estimador de la varianza del error sobre la base de la suma de residuos al cuadrado de la regresión. Siguiendo la expresión (3.22), tenemos que $e = My = M(X\beta + u) = Mu$ y como $MX = 0$, entonces:

$$E(e'e) = E(u'M'Mu) = E(u'Mu)$$

Utilizando el hecho de que la traza de un escalar es un escalar, se puede expresar:

$$\begin{aligned}
 E(u'Mu) &= E[tr(u'Mu)] \\
 &= E[tr(uu'M)] \\
 &= \sigma^2 tr(M) \\
 &= \sigma^2 trI - \sigma^2 tr[X(X'X)^{-1}X'] \\
 &= \sigma^2 trI - \sigma^2 tr[(X'X)^{-1}(X'X)] \\
 &= \sigma^2(n-k)
 \end{aligned}$$

por lo que:

$$s^2 = \frac{e'e}{n-k} \quad (3.24)$$

define un estimador insesgado de σ^2 . La raíz cuadrada s^2 es la desviación estándar de los valores de y . A menudo s es denominado error estándar de la estimación o **error estándar de la regresión**.

3.1.5.4. Inferencia en el modelo lineal

Una vez establecidas las propiedades del estimador MCO para β , queda mostrar la utilización de estos estimadores para probar varias hipótesis sobre los parámetros desconocidos β . Se consideran las siguientes pruebas de hipótesis como las más comunes:

- i. $H_o : \beta_i = 0$. Esta hipótesis expresa que el regresor X_i no tiene influencia sobre y . Este tipo de prueba es la más aplicada y es descrita muchas veces como una **prueba de significancia**.
- ii. $H_0 :$

$$\begin{bmatrix} \beta_2 \\ \beta_3 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

Esta hipótesis expresa que el conjunto completo de variables explicativas no tiene efecto sobre y . Esta hipótesis pone a prueba la significancia global del modelo. El término del intercepto no está

incluido en la hipótesis debido a que el interés se centra en la variación de y alrededor de su media.

Todas estas hipótesis pueden especificarse en una sola estructura lineal general:

$$R\beta = r \quad (3.25)$$

donde R es una matriz $q \times k$ de constantes conocidas con q menor a k , y r es un vector de dimensión $q \times 1$ de constantes conocidas. Cada hipótesis nula determina los elementos relevantes en R y r . Así, para las hipótesis anteriores tenemos:

- i.* $R = [0 \ \dots \ 0 \ 1 \ 0 \ \dots \ 0]$; $r = 0$; $q = 1$; con el valor de 1 en la posición i .
- ii.* $R = [0 \ I_{k-1}]$; $r = 0$; $q = k - 1$, donde 0 es un vector de $k - 1$ ceros.

La manera eficiente para proceder es derivar un procedimiento para una prueba de hipótesis lineal general:

$$H_0 : R\beta - r = 0$$

Esta prueba o “test” general puede utilizarse para cualquier aplicación específica. Dado un estimador MCO, establecemos el vector $(R\hat{\beta} - r)$. Este vector mide la discrepancia entre lo observado y su valor esperado. Si esta discrepancia es, en cierto sentido, “grande”, ésta reflejará dudas sobre la hipótesis nula y, por el contrario, si es “pequeña”, no contradice la hipótesis nula. Como en procedimientos convencionales, la distinción entre grande y pequeña es determinada mediante una distribución de muestreo específica bajo la hipótesis nula, en este caso, la distribución de $R\hat{\beta}$ cuando $R\beta = r$.

De la expresión en (3.19) se tiene directamente que:

$$E(R\hat{\beta}) = R\beta \tag{3.26}$$

y de la expresión en (3.20) se tiene además que:

$$\begin{aligned} \text{Var}(R\hat{\beta}) &= E\left[R(\hat{\beta} - \beta)(\hat{\beta} - \beta)'R'\right] \\ &= R\text{Var}(\hat{\beta})R' \end{aligned}$$

$$= \sigma^2 R(X'X)^{-1}R' \quad (3.27)$$

Por lo que, se ha encontrado la media y varianza del vector $R\hat{\beta}$; sin embargo, se requiere de un supuesto adicional para determinar la forma de la distribución de muestreo. Debido a que $\hat{\beta}$ es una función del vector u , la distribución de muestreo de $R\hat{\beta}$ está determinada por la distribución de u . Los supuestos hechos hasta ahora sobre u están dados por las expresiones (3.9) y (3.10); y, ahora, adoptando el supuesto adicional de que los errores estocásticos u_i están normalmente distribuidos, es posible combinar los tres supuestos en una sola proposición:

$$u \sim N(0_n, \sigma^2 I) \quad (3.28)$$

Debido a que la combinación lineal de variables normales también está normalmente distribuida, se sigue directamente que:

$$\hat{\beta} \sim N[\beta, \sigma^2 (X'X)^{-1}] \quad (3.29)$$

luego,

$$R\hat{\beta} \sim N[R\beta, \sigma^2 R(X'X)^{-1}R'] \quad (3.30)$$

por lo que,

$$R(\hat{\beta} - \beta) \sim N[0, \sigma^2 R(X'X)^{-1}R'] \quad (3.31)$$

Si la hipótesis $R\beta = r$ es verdadera, entonces

$$(R\hat{\beta} - r) \sim N[0, \sigma^2 R(X'X)^{-1}R'] \quad (3.32)$$

Esta expresión nos da la distribución de muestreo de $R\hat{\beta}$ de la que se puede derivar la variable χ^2 siguiente:

$$(R\hat{\beta} - r)' [\sigma^2 R(X'X)^{-1}R']^{-1} (R\hat{\beta} - r) \sim \chi^2(q) \quad (3.33)$$

donde q es el número de restricciones.

El único problema con la expresión anterior es la presencia del parámetro desconocido σ^2 . Sin embargo, se puede derivar que:

$$\frac{e'e}{\sigma^2} \sim \chi^2(n-k) \quad (3.34)$$

y que este estadístico está distribuido independientemente de $\hat{\beta}$. De aquí que, las expresiones (3.33) y (3.34) se pueden combinar en un solo estadístico el cual tendrá una distribución F bajo la hipótesis nula, es decir:

$$\frac{(R\hat{\beta} - r)[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{e'e/(n-k)} \sim F(q, n-k) \quad (3.35)$$

Luego, el procedimiento de prueba es rechazar la hipótesis $R\beta = r$ si el valor obtenido para el estadístico F excede el valor crítico seleccionado con anterioridad. Con estas condiciones, es posible introducir este método de prueba para los casos expuestos anteriormente.

Para algunas aplicaciones es más conveniente rescribir la expresión (3.35) como:

$$(R\hat{\beta} - r) \left[s^2 R(X'X)^{-1} R' \right]^{-1} (R\hat{\beta} - r) / q \sim F(q, n - k) \quad (3.36)$$

donde s^2 se definió en la expresión (3.24). De aquí que, $s^2(X'X)^{-1}$ es la **matriz estimada de varianzas-covarianzas** del vector $\hat{\beta}$. Si representamos con c_{ij} al elemento (i,j) de la matriz $(X'X)^{-1}$, entonces:

$$s^2 c_{ii} = \text{Var}(\hat{\beta}_i) \quad \text{y} \quad s^2 c_{ij} = \text{Cov}(\hat{\beta}_i, \hat{\beta}_j) \quad \text{para } i, j = 1, 2, \dots, k$$

En cada aplicación la forma específica de R y r se substituyen en las expresiones (3.35) o (3.36).

- i. $H_0 : \beta_i = 0$. De $R\hat{\beta}$ se escoge a $\hat{\beta}_i$ y de $R(X'X)^{-1}R'$ se escoge a c_{ii} , el elemento i ésimo en la diagonal de $(X'X)^{-1}$. De aquí que, la expresión (3.35) pasa a ser:

$$F = \frac{\hat{\beta}_i^2}{s^2 c_{ii}} = \frac{\hat{\beta}_i^2}{\text{Var}(\hat{\beta}_i)} \sim F(1, n - k)$$

o tomando la raíz cuadrada,

$$t = \frac{\hat{\beta}_i}{s\sqrt{c_{ii}}} = \frac{\hat{\beta}_i}{d.e.(\hat{\beta}_i)} \sim t(n-k)$$

Como resultado, la hipótesis de que una variable explicativa X_i no tenga asociación con y es probada dividiendo el i -ésimo coeficiente estimado por la estimación de su desviación estándar.

- ii. $H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$. Cada uno de los primeros cuatro casos involucran una hipótesis simple. Este caso implica una hipótesis compuesta sobre los $k-1$ coeficientes de los regresores. Ahora, de $R(X'X)^{-1}R'$ se escoge la submatriz cuadrada de orden $k-1$ en el extremo inferior derecho de $(X'X)^{-1}$. Para evaluar esta submatriz, se particiona a la matriz X en una matriz $[i \ X_2]$ donde X_2 es la matriz de observaciones sobre los $k-1$ regresores. Como resultado:

$$X'X = \begin{bmatrix} i' \\ X_2' \end{bmatrix} [i \ X_2] = \begin{bmatrix} n & i'X_2 \\ X_2' & X_2'X_2 \end{bmatrix}$$

Utilizando la propiedad de la inversa de matrices particionadas, se puede expresar a la submatriz requerida como:

$$\left[X_2' X_2 - X_2' i n^{-1} i' X_2 \right]^{-1} = \left[X_2' A X_2 \right]^{-1} = \left[X_*' X_* \right]^{-1}$$

donde la matriz A se define como

$$A = I_n - \left(\frac{1}{n} \right) i i' ; \quad i \text{ es un vector columna de unos}$$

Esta matriz transforma las observaciones en desviaciones con respecto a su media, y $X_* = A X_2$. $R \hat{\beta} = b_2$, el cual es un vector de estimadores MCO de los coeficientes de los $k-1$ regresores. Aparte del divisor q , el numerador de la expresión (3.35) es entonces $b_2' X_*' X_* b_2$, el cual corresponde a la suma de cuadrados de la regresión (SE) expresada en términos de las desviaciones con respecto a la media. De aquí, el estadístico F para probar la significancia global del conjunto completo de regresores es:

$$F = \frac{SE/k-1}{SR/n-k} \sim F(k-1, n-k) \quad (3.37)$$

el cual, utilizando (3.23), también puede ser expresado como:

$$F = \frac{R^2 / (k-1)}{(1-R^2) / (n-k)} \sim F(k-1, n-k) \quad (3.38)$$

Esta prueba reconocida como **análisis de varianza** o **ANOVA** esencialmente determina si la media cuadrática de la regresión es significativamente mayor que la media cuadrática del error.

3.1.6. Heteroscedasticidad y Autocorrelación

Se han analizado y expuesto las propiedades del estimador MCO de los coeficientes de un modelo lineal cuyo término del error tiene una matriz de varianzas-covarianzas escalar: todos sus elementos son cero, excepto los de la diagonal principal los cuales son todos iguales a σ^2 . Sin embargo, existen dos situaciones en las que dicha matriz de covarianzas tiene una estructura más compleja, en cuyo caso las propiedades analizadas anteriormente para los estimadores MCO podrían dejar de ser válidas.

Una de tales situaciones se produce cuando la varianza del término de error es distinta de unas observaciones a otras. Con datos de series temporales, esta situación se produce cuando $Var(u_t) = \sigma_t^2$, con $\sigma_t^2 \neq \sigma_s^2$ para dos instantes $t \neq s$. En el caso de datos de sección cruzada, esta situación se produce cuando $Var(u_i) = \sigma_i^2$, con $\sigma_i^2 \neq \sigma_j^2$ para dos sitios $i \neq j$. En ambos casos, los elementos de la diagonal principal de la matriz de covarianzas no serán iguales entre sí. A esta situación se la denomina **heteroscedasticidad**, a diferencia del caso en el que la varianza de u_i para todo periodo es constante, situación que denominamos **homoscedasticidad**.

Una segunda situación ocurre cuando, disponiendo de datos de series temporales, los términos del error correspondientes a distintos períodos no son independientes entre sí, es decir, $Cov(u_t, u_{t-k}) = E(u_t u_{t-k}) \neq 0$ para algún k mayor que cero. Ello hace que los elementos de fuera de la diagonal en la matriz de varianzas-covarianzas no sean todos nulos, por lo que ésta no será diagonal. A esta situación se la denomina **autocorrelación**, para reflejar el hecho de que el término del error está correlacionado consigo mismo a través del tiempo.

Aunque esta propiedad está naturalmente asociada a los datos de series temporales, se presenta también en muestras de sección cruzada si la ordenación de las observaciones en la muestra sigue unos criterios “*geográficos*”, como pueden ser las distintas áreas de un país, las familias de un vecindario, o los lotes de un cultivo; entonces, los términos del error correspondientes a observaciones sucesivas pueden estar correlacionados entre sí; esta situación se denomina ***correlación espacial***, y se tratará de forma extensa en las secciones posteriores.

La primera cuestión que debe plantearse es acerca de las propiedades del estimador MCO en presencia de heteroscedasticidad y/o autocorrelación. Como se expuso anteriormente, cuando la matriz de varianzas-covarianzas del término del error es escalar, dicho estimador es óptimo, sin embargo, ¿es el estimador MCO insesgado cuando el término del error tiene una matriz de covarianzas que no es escalar?. Si es insesgado, ¿es óptimo (es decir, de mínima varianza entre todos los estimadores lineales e insesgados), al igual que ocurría con una matriz de varianzas-covarianzas escalar?. Si no es óptimo, ¿existe algún otro estimador con menor varianza, es decir, más eficiente?. Más

generalmente, ¿en qué medida dependen las propiedades del estimador MCO de la estructura de matriz de covarianzas del vector u ?.

La definición del estimador MCO en la expresión (3.17) no depende de la estructura de la matriz de varianzas-covarianzas del término de error del modelo econométrico. Por consiguiente, es posible redefinir las propiedades del estimador MCO para una situación más general en la que la matriz de covarianzas no es escalar:

Dado el modelo $y = X\beta + u$, donde y , u son $n \times 1$ y X la matriz $n \times k$ de observaciones de las variables explicativas, con $E(u) = 0$ y $Var(u) = \sigma^2 \Omega$, el estimador MCO del vector de parámetros β es una solución al sistema de ecuaciones normales $(X'X)\hat{\beta} = X'y$.

Cuando la matriz $X'X$ es invertible, entonces dicho sistema tiene una solución única, dada por $\hat{\beta} = (X'X)^{-1}X'y$, y al igual que probamos al demostrar la propiedad 2 (Ec. 3.18), se tiene:

$$\hat{\beta} = \beta + (X'X)^{-1}X'u \quad (3.39)$$

Cuando $\Omega = I$, la matriz de varianzas-covarianzas es escalar y los resultados acerca del estimador MCO se reducen a los anteriormente expuestos.

Una implicación de la expresión (3.39) es que las condiciones para que el estimador MCO sea insesgado cuando $Var(u) = \sigma^2\Omega$ son las mismas que eran precisas para que dicho estimador fuese insesgado cuando $Var(u) = \sigma^2I$, es decir:

$$E[(X'X)^{-1}X'u] = 0_k \quad (3.40)$$

Lo que se traduce en que si las variables explicativas del modelo son deterministas y $Var(u) = \sigma^2\Omega$, entonces el estimador de mínimos cuadrados ordinarios es insesgado.

$$E[(X'X)^{-1}X'u] = (X'X)^{-1}X'E(u) = (X'X)^{-1}X'0_n = 0_k$$

Por otro lado, si $Var(u) = E(uu') = \sigma^2\Omega$ y las variables explicativas son deterministas, entonces la matriz de covarianzas del estimador MCO es:

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= E\left[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'\right] = \\
&= E\left[(X'X)^{-1} X'uu'X(X'X)^{-1}\right] = \\
&= (X'X)^{-1} X'E(uu')X(X'X)^{-1} = \\
&= \sigma^2 (X'X)^{-1} X'\Omega X(X'X)^{-1} \tag{3.41}
\end{aligned}$$

Una vez obtenidas las primeras propiedades del estimador MCO cuando la matriz de varianzas-covarianzas del término del error no es escalar, conviene hacer algunas observaciones importantes:

1. La expresión (3.41) es muy distinta en apariencia de la matriz de covarianzas $\sigma^2(X'X)^{-1}$ del estimador MCO cuando $\text{Var}(u) = \sigma^2 I$. Sin embargo, es fácil ver que cuando $\Omega = I_n$, la matriz de varianzas-covarianzas de la expresión (3.41) se reduce a la matriz $\sigma^2(X'X)^{-1}$. En consecuencia, no hay ninguna razón por la que deba imponerse como matriz de varianzas-covarianzas del estimador MCO una expresión restringida como es $\sigma^2(X'X)^{-1}$, sino que más bien debería utilizarse (3.41) y permitir así, que en las circunstancias apropiadas, es decir, en ausencia de heteroscedasticidad y autocorrelación, dicha matriz se reduzca numéricamente a $\sigma^2(X'X)^{-1}$.

2. Los estimadores MCO son *ineficientes* y utilizar $\sigma^2(X'X)^{-1}$ como matriz de varianzas-covarianzas de éstos cuando puede existir autocorrelación y heteroscedasticidad es simplemente *incorrecto*. Las pruebas estadísticas descritas anteriormente basadas sobre estos estimadores son *inválidas*.
3. Como muestra la expresión (3.39), el estimador MCO está definido en este caso más general por la misma expresión que cuando $\Omega = I_n$ y es, por tanto, una función lineal del término de error u . Por ello, sigue siendo cierto que si las variables X son deterministas y el término del error tiene una distribución normal, entonces el estimador MCO también se distribuye normalmente, con el vector de esperanzas y matriz de covarianzas:

$$\hat{\beta}_{MCO} \sim N\left(\beta, \sigma^2(X'X)^{-1} X' \Omega X (X'X)^{-1}\right)$$

3.1.7. Estimación de Mínimos Cuadrados Generalizados (MCG)

Debido a que el término de error ya no posee una matriz de varianzas-covarianzas escalar en presencia de heteroscedasticidad y/o autocorrelación, sería interesante poder transformar el modelo econométrico para lograr este propósito garantizando que los coeficientes del modelo transformado fuesen los mismos del modelo original. Si ello fuese posible, entonces se podría utilizar el estimador MCO del modelo transformado y que por los resultados anteriormente descritos sería eficiente, es decir, de mínima varianza entre los estimadores lineales insesgados.

Para ello, se premultiplica al modelo por una matriz P de dimensiones $n \times n$ para obtener:

$$Py = (PX)\beta + Pu \quad (3.42)$$

Lo que en realidad se hace con la transformación es un cambio de variables donde se define: $X_* = PX$, $y_* = Py$ y $u_* = Pu$.

$$y_* = X_*\beta + u_* \quad (3.43)$$

La linealidad del modelo permite que los coeficientes β del modelo transformado sean precisamente los mismos que los del modelo original.

La matriz de covarianzas del nuevo término del error es:

$$\text{Var}(u_*) = \text{Var}(Pu) = \sigma^2 P\Omega P' \quad (3.44)$$

Debido a que la matriz Ω es simétrica definida positiva, su inversa es también definida positiva, por lo que siempre existe una matriz cuadrada, no singular, V de modo que:

$$\Omega = VV'$$

De aquí que $V^{-1}\Omega(V^{-1})' = I$ y por lo tanto, $\Omega^{-1} = (V^{-1})V^{-1}$. Basta comparar estas expresiones con $\sigma^2 P\Omega P'$ para notar que si se utiliza precisamente la matriz V^{-1} como matriz de transformación P , entonces el término del error resultante tiene matriz de varianzas-covarianzas escalar.

En particular, el estimador de MCO de los parámetros del modelo (3.43) es:

$$\hat{\beta}_{MCG} = (X_*'X_*)^{-1} X_*'y_* = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}y \quad (3.45)$$

Es importante recordar que este estimador no es equivalente a aplicar MCO directamente al modelo original, puesto que las variables que en él aparecen son distintas de las del modelo (3.43) que se ha estimado.

Por obtenerse el estimador MCG mediante la aplicación del procedimiento MCO al modelo transformado, se pueden asegurar algunas de sus propiedades, como:

a) El estimador MCG es insesgado puesto que:

$$\hat{\beta}_{MCG} = \beta + (X_*'X_*)^{-1} X_*'u_*, \text{ y como } E(u_*) = 0, \text{ entonces } E(\hat{\beta}_{MCG}) = \beta.$$

b) Matriz de covarianzas del estimador MCG:

$$\text{Var}(\hat{\beta}_{MCG}) = \sigma^2 (X_*'X_*)^{-1} = \sigma^2 (X'\Omega^{-1}X)^{-1} \quad (3.46)$$

- c) Ecuaciones normales: al igual que el estimador MCO, el estimador MCG también satisface un sistema de ecuaciones normales. En efecto, puesto que este estimador no es sino el estimador MCO del modelo transformado, se tiene $(X_*'X_*)\hat{\beta}_{MCG} = X_*'y^*$ que es lo mismo a:

$$(X'\Omega^{-1}X)\hat{\beta}_{MCG} = X'\Omega^{-1}y$$

que es sistema de ecuaciones normales del estimador MCG.

- d) Eficiencia: El estimador MCG es el estimador lineal insesgado de mínima varianza del vector de parámetros β pues, debido a que satisface los supuestos de la sección 3.1.4; entonces, por el teorema de Gauss-Markov, el estimador MCO de (3.43) es el estimador lineal insesgado de mínima varianza del vector de parámetros β .

El problema fundamental en la obtención del estimador MCG es que tanto σ^2 como Ω son desconocidos, y se requiere una estimación previa de los mismos. Sin embargo, al igual que lo que ocurría con el estimador MCO, no es preciso conocer el parámetro σ^2 para obtener el estimador

$\hat{\beta}_{MCG}$, aunque sí para calcular su matriz de covarianzas, de acuerdo con la expresión (3.46).

La estimación de mínimos cuadrados $\hat{\sigma}_{MCG}^2$ se obtiene a partir del modelo (3.43), del modo que se vio en la sección 3.1.5.3; es decir:

$$\hat{\sigma}_{MCG}^2 = \frac{e_*'e_*}{n-k} = \frac{e_{MCG}'\Omega^{-1}e_{MCG}}{n-k}$$

donde $e_{MCG} = y - X\hat{\beta}_{MCG}$

3.1.8. Contraste de heteroscedasticidad de White

Este contraste asintótico no requiere especificar la forma que puede adoptar la heteroscedasticidad. Simplemente se realiza una regresión auxiliar de los residuos MCO al cuadrado sobre una constante, los regresores del modelo original, sus cuadrados y productos cruzados de segundo orden.

Bajo el supuesto de homoscedasticidad y al aumentar el tamaño muestral, el producto nR^2 , donde n es el tamaño de la muestra y R^2 el coeficiente de determinación entre los residuos y las nuevas variables, sigue una distribución chi-cuadrado con $p - 1$ grados de libertad, donde p es el número de regresores en el modelo sobre el cuadrado de los residuos mínimo-cuadráticos.

$$nR^2 \stackrel{d}{\sim} \chi^2(p - 1) \quad (3.42)$$

Mientras el tamaño muestral crezca con el número de observaciones, el coeficiente de determinación tenderá a cero bajo la hipótesis nula de homoscedasticidad. Sólo cuando la varianza del término del error depende de las variables explicativas del modelo, el coeficiente no tenderá a cero. En tal caso el producto nR^2 permanecerá a un cierto nivel, lejos de cero, y es de esperar que supere el valor de las tablas de la distribución chi-cuadrado. Cuando se rechaza homoscedasticidad, no hay indicador de la forma de heteroscedasticidad presente por lo que no podemos aplicar la estimación de MCG.

3.1.8. Multicolinealidad

La **multicolinealidad** aparece cuando las variables explicativas de un modelo econométrico están correlacionadas entre sí, y tiene implicaciones negativas cuando se pretende estimar un modelo lineal por mínimos cuadrados.

Como ya se ha mencionado, si existe la inversa de la matriz $X'X$, el estimador MCO en la expresión (3.17) viene dado por $\hat{\beta}_{MCO} = (X'X)^{-1} X'y$, y su matriz de varianzas-covarianzas por $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$.

Si se hace el supuesto de que una de las variables explicativas x_{ti} tiene una correlación elevada con respecto a las demás variables explicativas del modelo, es decir, que la regresión lineal

$$x_{ti} = \delta_1 + \delta_2 x_{2i} + \dots + \delta_{t-1} x_{t-1,i} + \delta_t x_{t+1,i} + \dots + \delta_{k-1} x_{ki} + v_i \quad (3.43)$$

tiene un coeficiente de determinación alto. En tales condiciones, la variable x_{ti} podría aproximarse por una combinación lineal de las restantes variables explicativas, dada precisamente por la regresión

estimada en la expresión (3.43). Como consecuencia, una de las columnas de la matriz X es aproximadamente una combinación lineal de las demás, y por tanto, la matriz $X'X$ será aproximadamente singular.

En tanto en cuanto el determinante de $X'X$ no sea numéricamente igual a cero, existirá la inversa $(X'X)^{-1}$ y, por tanto, también existirá el estimador MCO. Este será todavía el estimador de mínima varianza entre los estimadores lineales e insesgados, pero se tendrían varias consecuencias:

1. Aún cuando los estimadores MCO son de mínima varianza entre los estimadores lineales e insesgados, éstos presentan varianzas y covarianzas grandes, que hacen difícil la estimación precisa.
2. Debido a la consecuencia 1, los intervalos de confianza tienden a ser mucho más amplios conduciendo a un más fácil no rechazo de la hipótesis nula de que el parámetro estimado es igual a cero (es decir, que el verdadero coeficiente poblacional es cero).
3. También debido a la consecuencia 1, la razón t de uno o más coeficientes tiende a ser estadísticamente no significativa.

4. Aún cuando la razón t de uno o más coeficientes sea estadísticamente no significativa, el R^2 , la medida global de bondad de ajuste, puede ser muy alto.
5. Los estimadores MCO y sus errores estándar pueden ser sensibles a pequeños cambios en la información.

Para tener una mejor idea de cómo aumentan las varianzas y covarianzas de los estimadores de los parámetros, en el modelo de tres variables expresadas en desviaciones con respecto a sus medias muestrales.

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + e_i$$

las varianzas y covarianzas de $\hat{\beta}_2$ y $\hat{\beta}_3$ están dadas por:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

$$\text{Var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

$$Cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}}$$

donde r_{23} es el coeficiente de correlación entre X_2 y X_3 .

Es aparente que a medida que r_{23} tiende a 1, es decir, a medida que la colinealidad aumenta, las varianzas de los dos estimadores aumentan y, en el límite cuando $r_{23} = 1$, son infinitas. Es igualmente claro que a medida que r_{23} aumenta hacia 1, la covarianza de los dos estimadores también aumenta en valor absoluto.

La velocidad con la cual las varianzas y covarianzas se incrementan puede cuantificarse mediante el **factor inflador de varianza (FIV)** que se define como:

$$FIV = \frac{1}{(1-r_{23}^2)} \quad (3.44)$$

El FIV muestra la forma como la varianza de un estimador es inflada por la presencia de la multicolinealidad. A medida que r_{23}^2 se acerca a 1, el

FIV se acerca a infinito. Es decir, a medida que el grado de colinealidad aumenta, la varianza de un estimador aumenta y, en el límite, se puede volver infinita. Como puede verse fácilmente, si no hay colinealidad entre X_2 y X_3 , el FIV será 1. Utilizando esta definición, se pueden expresar las varianzas de los estimadores como

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} FIV$$

$$Var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2} FIV$$

lo que muestra que las varianzas de $\hat{\beta}_2$ y $\hat{\beta}_3$ son directamente proporcionales al FIV.

Para el modelo de regresión con k variables, la varianza de un coeficiente de regresión parcial puede ser expresada como:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \left(\frac{1}{1 - R_j^2} \right) = \frac{\sigma^2}{\sum x_j^2} FIV \quad (3.45)$$

donde β_j es el coeficiente de regresión (parcial) del regresor X_j , R_j^2 es el R^2 en la regresión (auxiliar) de X_j sobre los restantes $k - 2$ regresores y FIV_j es el primer factor de inflación de varianza. A medida que R_j^2 aumenta hacia la unidad, es decir, a medida que aumenta la colinealidad de X_j con los demás regresores, el FIV también aumenta y en el límite puede ser infinito. Como regla general, si el FIV de una variable es superior a 10 (esto sucederá si R_j^2 excede a 0.90), se dice que esta variable es altamente colineal.

En resumen, si la multicolinealidad es perfecta, los coeficientes de regresión de las variables X son indeterminados y sus errores estándar son infinitos. Si la multicolinealidad es menos que perfecta, los coeficientes de regresión estimados, aunque sean determinados, poseen grandes errores estándar lo cual significa que los coeficientes no pueden ser estimados con precisión o exactitud.

3.2. EL MODELO ESPACIAL DE REGRESIÓN LINEAL MÚLTIPLE

El término ***econometría espacial*** fue acuñado por Jean Paelink a comienzos de la década de los setenta para referirse al conjunto de métodos que tratan de manera adecuada las características especiales de los datos georeferenciados y de los modelos de economía espacial. Anselin (1988), la referencia más citada en los trabajos de econometría espacial, la define como “la colección de técnicas que lidian con las peculiaridades causadas por el espacio en el análisis estadístico de los modelos de la ciencia regional”. Once años más tarde, Anselin (1999) extiende la definición diciendo que:

La econometría espacial es una rama de la econometría que se preocupa del tratamiento adecuado de la interacción espacial (autocorrelación espacial) y la estructura espacial (heterogeneidad espacial) en modelos de regresión con datos de corte transversal (sección cruzada) y de datos de panel.

La creciente importancia que están teniendo estos métodos en las diferentes áreas de investigación se debe, principalmente, a las siguientes razones:

1. El renovado interés por investigar el rol del espacio y de la interacción espacial en las ciencias sociales (especialmente en la economía) y, actualmente, con mayor frecuencia, en el análisis de datos de AP.
2. La creciente disponibilidad de grandes bancos de datos con observaciones georeferenciadas por parte de agencias oficiales de gobiernos y otras instituciones privadas.
3. El desarrollo de una tecnología computacional eficiente y de bajo costo para manejar observaciones espacialmente referenciadas, como son los sistemas de información geográficos (GIS) y software para análisis de datos espaciales.

3.2.1. Expresión formal de los efectos espaciales: Conectividad en el Espacio.

Uno de los aspectos operacionales cruciales de la econometría espacial es el problema de expresar formalmente la manera en la cual la estructura de dependencia espacial es incorporada en un modelo. En contraste al análisis de series de tiempo, donde la noción de un rezago temporal se presenta sin ambigüedades, en el contexto espacial el análisis se complica considerablemente.

La noción de la existencia de dependencia espacial implica la necesidad de determinar qué otras unidades en el sistema espacial tienen influencia sobre una unidad particular bajo consideración. Formalmente, esto es expresado mediante las nociones topológicas de vecindad y vecinos más cercanos.

Considérese un sistema S de n unidades espaciales, $i = 1, 2, \dots, n$, y una variable x con valores observados para cada una de estas unidades espaciales. Un conjunto de vecinos para una unidad espacial i está definido como la colección de aquellas unidades j

para las cuales x_j está contenida en la forma funcional de la probabilidad condicional de x_i , dados los valores de x en otras localidades. Formalmente, esta definición describe el conjunto de vecinos para la unidad i como J , para el cual:

$$P[x_i/x] = P[x_i/x_j]$$

donde x_j es el vector de observaciones para $x_j \forall j \in J$, y x es el vector que contiene a todos los valores de la variable x en el sistema. De forma alternativa, y de manera menos estricta, un conjunto de vecinos j para i pueden ser tomados como:

$$\{j / P[x_i] \neq P[x_i/x_j]\}$$

o, como aquellas localidades para las que la probabilidad condicional marginal de x_i no es igual a su probabilidad marginal incondicional. Nótese que ninguna de estas definiciones incluye información acerca de la ubicación relativa de las dos unidades espaciales, sino que sólo trata una posible influencia vía probabilidades condicionales.

Para introducir un aspecto espacial en estas definiciones de vecinos, se sugiere la siguiente definición alternativa:

$$\{j / P[x_i] \neq P[x_i / x_j] \text{ y } d_{ij} < \varepsilon_i\}$$

donde d_{ij} es una medida de la distancia entre i y j en un espacio propiamente estructurado, y ε_i es un valor crítico o umbral para cada unidad espacial i , y posiblemente para todas las demás unidades espaciales. La métrica de distancia sobre d_{ij} , en el caso más general, es la distancia Euclidiana.

Este concepto alternativo de vecindad introduce una estructura adicional en el conjunto espacial de datos, combinando una noción de dependencia estadística (relacionando magnitudes) con una noción de espacio (distancia y posición relativa). Aunque la definición no impide que las unidades espaciales j que no satisfacen el criterio de distancia tengan influencia sobre la probabilidad condicional de x_i , éstas no son consideradas como vecinos más cercanos sino que pueden incluirse como vecinos de orden superior, es decir, que su influencia sobre la unidad i se presenta vía otras unidades

espaciales. La definición basada sobre probabilidades condicionales no permite hacer esta distinción entre vecinos de primer orden y vecinos de orden superior.

El conjunto de vecinos resultante para cada unidad espacial puede representarse mediante un gráfico o mapa y asociarse a una matriz de conectividad, la cual se discutirá a continuación.

3.2.1.1. Matriz de conectividad espacial

La principal característica de los datos georeferenciados es que están asociados con una localización y por lo tanto se pueden visualizar en un mapa. En este espacio, los datos representan información sobre objetos que pueden ser puntos como ciudades, almacenes, crímenes o accidentes y que están referenciados por coordenadas en el plano; o, los objetos podrían ser polígonos que podrían representar provincias, regiones o tipos de suelo en un cultivo.

Una de las formas más comunes de representar la vecindad de un conjunto de polígonos es a través de una **matriz de conectividad**. Esta es una matriz cuadrada que tiene el mismo número de columnas que el número de polígonos independientes del mapa en estudio y que por convención se escribe como ***W***.

Los valores utilizados para representar vecindad son variados. La formulación más simple es una matriz de conectividad binaria donde la estructura de vecinos es expresada mediante ceros y unos (0 y 1). Si dos unidades espaciales tienen un borde en común de longitud diferente de cero, éstas son consideradas como vecinos más cercanos y el valor de uno es asignado, caso contrario, se asigna el valor de cero. La Figura 3.1 muestra un arreglo espacial de nueve polígonos y su matriz de conectividad asociada.

Esta matriz de conectividad tiene ceros en la diagonal principal porque se asume que un polígono no puede ser vecino consigo mismo. Adicionalmente, en la práctica esta matriz se estandariza por filas, es decir, se divide cada componente de las filas de la matriz por la suma de los elementos de la fila respectiva, de modo que la suma de los

elementos de la fila sea uno, esta forma es muy útil para crear los rezagos espaciales.

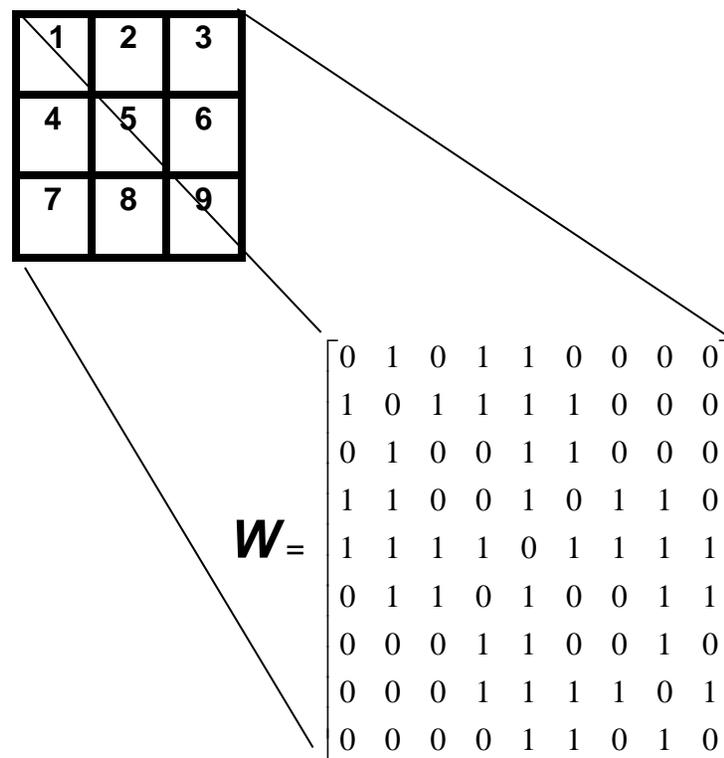


Figura 3.1. Matriz de conectividad espacial no estandarizada
Elaboración: Xavier Vítores, 2004

3.2.2. Taxonomía de los modelos espaciales de regresión lineal múltiple

En esta sección se presenta una especificación general, la cual se considera como el punto de partida para varias especificaciones de interés en la econometría espacial. Se derivan modelos específicos mediante la imposición de varias restricciones sobre los parámetros de la formulación general. Consecuentemente, en los procedimientos de estimación y especificación que siguen en la siguiente sección, muchos resultados para estos modelos serán encontrados de manera simple como casos especiales de la derivación que parte del modelo general. El modelo de partida consiste en la siguiente expresión:

$$\begin{aligned}
 y &= \rho W y + X \beta + u \\
 u &= \lambda W_2 u + \varepsilon \\
 \varepsilon &\sim N(0, \Omega)
 \end{aligned}
 \tag{3.46}$$

y los elementos de la diagonal de la matriz de varianzas-covarianzas del error Ω como:

$$\Omega_{ii} = h_i(z' \alpha) \quad h_i > 0$$

En esta especificación espacial general (Ec. 3.46), β es un vector de dimensiones $k \times 1$ de los parámetros asociados con las variables exógenas o regresores (es decir, sin el rezago espacial) en la matriz X de dimensiones $n \times k$, ρ es el coeficiente del rezago espacial de la variable dependiente, y λ es el coeficiente de la estructura autoregresiva espacial para el error u .

El vector u acoge el supuesto de estar normalmente distribuido con una matriz general de varianzas-covarianzas Ω . Los elementos de la diagonal acogen la presencia de heteroscedasticidad como una función de $p+1$ variables exógenas z , las que incluyen un término constante. Los p parámetros en α están asociados con los términos no-constantas, de manera que, para $\alpha = 0$, se tiene:

$$h = \sigma^2$$

es decir, la clásica situación de homoscedasticidad.

Las dos matrices de dimensiones $n \times n$, W_1 y W_2 , son matrices de conectividad espacial estandarizadas o no estandarizadas,

respectivamente asociadas con un proceso autoregresivo espacial en la variable dependiente y en el término del error. Con todo, el modelo tiene $3 + k + p$ parámetros desconocidos:

$$\theta = [\rho \quad \beta' \quad \lambda \quad \sigma^2 \quad \alpha']' \quad (3.47)$$

Varios modelos espaciales se derivan de la expresión (3.46) cuando algunos subvectores del vector de parámetros (3.47) se consideran como vectores nulos. Específicamente, las siguientes situaciones corresponden a los cuatro modelos autoregresivos espaciales tradicionales:

1. MODELO DE REGRESIÓN LINEAL CLÁSICO: Para $\rho = 0$, $\lambda = 0$, $\alpha = 0$ ($p + 2$ restricciones), tenemos,

$$y = X\beta + u \quad (3.48)$$

2. MODELO MIXTO REGRESIVO-ESPACIAL AUTOREGRESIVO

(SAR): Para $\lambda = 0$ y $\alpha = 0$ ($p+1$ restricciones), tenemos,

$$y = \rho W_1 y + X\beta + u \quad (3.49)$$

3. MODELO ESPACIAL AUTOREGRESIVO DEL ERROR (SEM):

Para $\rho = 0$ y $\alpha = 0$ ($p+1$ restricciones), tenemos,

$$y = X\beta + (I - \lambda W_2)^{-1} \varepsilon \quad (3.50)$$

4. MODELO ESPACIAL GENERAL (SAC): Para $\alpha = 0$ (

p restricciones), tenemos,

$$y = \rho W_1 y + X\beta + (I - \lambda W_2)^{-1} \varepsilon \quad (3.51)$$

Cuatro especificaciones más pueden obtenerse permitiendo una forma específica de heteroscedasticidad.

Esta taxonomía está enfocada principalmente en la especificación de la dependencia espacial. Las especificaciones de la heterogeneidad espacial pueden incorporarse mediante procedimientos tradicionales. De hecho, las formas más comunes de heterogeneidad espacial se especifican mediante la variación de parámetros por submuestras; y, desde que esta situación resulta comúnmente en heteroscedasticidad de alguna forma, la heterogeneidad puede incorporarse especificando los términos apropiados para $h(z'\alpha)$. Consecuentemente, muchas situaciones de interés pueden derivarse de la expresión general en (3.46).

3.2.3. Limitaciones de la estimación MCO en los modelos de procesos espaciales

La dependencia espacial en los varios modelos autoregresivos espaciales muestra muchas similitudes a la ya muy familiar dependencia temporal. Por esto, se podría esperar que las propiedades de la estimación MCO para modelos con un rezago de la variable dependiente y/o correlación serial en el error se trasladen

directamente al caso espacial. Sin embargo, esto no ocurre. La falta de una analogía se da, en primer lugar, debido a la naturaleza bidimensional y multidireccional de la dependencia en el espacio. A continuación se ilustrará este hecho para la situación de un rezago espacial de la variable dependiente y el caso de autocorrelación espacial en el error.

3.2.3.1. Estimadores MCO en presencia de un rezago espacial de la variable dependiente

Es un resultado muy bien conocido en econometría que los estimadores MCO permanecen consistentes (ver el Anexo 4 sobre teoría asintótica) aún cuando un rezago de la variable dependiente está presente, siempre y cuando el término del error estocástico no presente autocorrelación serial (ver detalles en el Anexo 5). Consecuentemente, aunque las propiedades para muestras pequeñas de los estimadores se ven afectadas (dejan de ser insesgados), éstos pueden aún ser utilizados como base para realizar inferencia asintótica.

Para modelos espaciales autoregresivos, este resultado no se mantiene, independientemente de las propiedades del término del error estocástico.

Considérese el siguiente modelo espacial autoregresivo de primer orden:

$$y = \rho W y + u \quad (3.52)$$

donde W es la usual matriz de conectividad espacial, ρ es el coeficiente espacial autoregresivo, la variable dependiente y está expresada en desviaciones con respecto a su media y u es el vector que contiene a los errores estocásticos independientes e idénticamente distribuidos. Aunque el modelo es extremadamente simple, éste captura todos los efectos de la presencia de un rezago espacial de la variable dependiente sobre el estimador MCO, y puede ser utilizado sin pérdida de generalidad.

El estimador MCO para ρ , denotado como $\hat{\rho}$ es:

$$\hat{\rho} = (y_L' y_L)^{-1} y_L' y \quad (3.53)$$

con $y_L = Wy$ como el rezago espacial de la variable dependiente.

Substituyendo la expresión para y en los parámetros poblacionales de (3.52) en (3.53) tenemos:

$$\hat{\rho} = \rho + (y_L' y_L)^{-1} y_L' u$$

Similar a la situación en series de tiempo, el valor esperado del segundo término no es igual a cero, y, consecuentemente, el estimador MCO es sesgado. Esto se debe a la compleja naturaleza estocástica del término a la inversa. Éste tendrá elementos que serán función de los valores de la variable dependiente (y consecuentemente de u) para cada observación. Como resultado, este término estará correlacionado con u . Además, mientras que en las series de tiempo $E(y_L' u) = 0$ si no existe correlación serial del

error, este no es el caso en el espacio. En el modelo espacial

$$E(y_L \hat{u}) = E[u'W(I - \rho W)^{-1}u] \text{ el cual es cero sólo cuando } \rho = 0 .$$

Asintóticamente, la consistencia del estimador MCO depende de las dos condiciones siguientes:

$$plim n^{-1}(y_L' y_L) = Q, \text{ una matriz finita y no singular}$$

$$plim n^{-1}(y_L' \hat{u}) = 0$$

donde *plim* simboliza la convergencia en probabilidad definida en el Anexo 4 como:

$$\lim_t P[|X_t - X| < \varepsilon] = 1, \forall \varepsilon > 0$$

para una sucesión de variables aleatorias $X_1, X_2, \dots, X_t, \dots$

Considerandose que la primera condición puede satisfacerse mediante las restricciones apropiadas sobre el valor de ρ y la estructura de la matriz de conectividad, la segunda condición no se mantiene en el caso espacial. De hecho,

$$plim n^{-1}(y_L \hat{u}) = plim n^{-1}u^T W(I - \rho W)^{-1}u$$

La presencia de la matriz de conectividad espacial en esta expresión resulta en una forma cuadrática del término del error. Por lo que, excepto en el caso trivial cuando $\rho = 0$, la probabilidad límite de esta expresión no será igual a cero.

Consecuentemente, el estimador MCO será **sesgado** así como **inconsistente** para el parámetro ρ del modelo espacial, independientemente de las propiedades del término del error.

3.2.3.2. Estimadores MCO en presencia de un rezago espacial del error estocástico

Los efectos de la autocorrelación espacial del error estocástico sobre las propiedades de los estimadores MCO son muy similares a los resultados obtenidos para las series de tiempo. Los estimadores de los parámetros aún serán insesgados pero ineficientes debido a la estructura no diagonal de la matriz de varianzas-covarianzas del error.

El modelo espacial puede ser considerado como un caso especial del modelo de regresión lineal con una matriz general parametrizada del término del error. De aquí que, las propiedades usuales de los estimadores MCO y MCG pueden aplicarse. Sin embargo, en el caso espacial, la naturaleza multidireccional de la dependencia espacial podrían limitar el tipo de procedimientos para obtener estimadores MCG consistentes.

3.2.4. Estimación de Máxima Verosimilitud (MV)

Debido a lo inapropiado que resulta la aplicación de los estimadores MCO para los parámetros de los modelos espaciales que incorporan la dependencia espacial, la atención se ha dirigido hacia los estimadores de máxima verosimilitud (MV).

Sea $y' = [y_1 \ y_2 \ \dots \ y_n]$ un vector de valores pertenecientes a una muestra de tamaño n , dependiente de un vector de k parámetros desconocidos $\theta' = [\theta_1 \ \theta_2 \ \dots \ \theta_k]$. Sea $f(y; \theta)$ la función de densidad de los valores de la muestra que indica la dependencia

sobre el vector θ . Esta densidad puede ser interpretada de dos maneras distintas. Para un vector θ esta indica la probabilidad de una muestra particular. De manera alternativa, esta puede ser interpretada como una función de θ condicional a una muestra particular. La definición formal es:

$$\text{Función de verosimilitud} = L(\theta; y) = f(y; \theta)$$

Es costumbre invertir el orden de los símbolos de la función de verosimilitud para enfatizar el nuevo centro de interés. Maximizando la función de verosimilitud con respecto a θ nos permite encontrar un valor específico, digamos $\hat{\theta}$, que maximiza la probabilidad de obtener los valores de una muestra particular. Luego, se define a $\hat{\theta}$ como el estimador de máxima verosimilitud (MV) del vector de parámetros desconocidos θ .

En la mayoría de las aplicaciones es menos complicado maximizar el logaritmo de la función de verosimilitud. Denotamos la función log-verosimilitud como:

$$l = \ln L$$

Luego

$$\frac{\partial l}{\partial \theta} = \frac{1}{L} \frac{\partial L}{\partial \theta}$$

y el vector $\hat{\theta}$ que maximiza l también maximizará L . La derivada de l con respecto a θ se conoce como “**score**”, $s(\theta; y)$. El estimador MV, $\hat{\theta}$, se obtiene igualando el score a cero, esto es, encontrando el valor de θ que resuelve:

$$s(\theta; y) = \frac{\partial l}{\partial \theta} = 0$$

En general, la amplia utilización de los estimadores MV se debe a sus interesantes propiedades asintóticas, las cuales se exponen en la siguiente sección.

3.2.4.1. Propiedades de los estimadores MV

La mayoría de las propiedades de los estimadores MV son asintóticas o para muestras grandes:

1. Consistencia

$$plim(\hat{\theta}) = \theta$$

2. Normalidad asintótica

$$\hat{\theta} \overset{d}{\sim} N(\theta, I^{-1}(\theta))$$

donde se simboliza a la convergencia en distribución definida en el Anexo 4 como $\overset{d}{\sim}$ o $\overset{d}{\rightarrow}$.

Esta propiedad quiere decir que la distribución asintótica de $\hat{\theta}$ es normal con media θ y varianza dada por la inversa de $I(\theta)$. $I(\theta)$ es la **matriz de información** y está definida de dos formas equivalentes:

$$I(\theta) = E \left[\left(\frac{\partial l}{\partial \theta} \right) \left(\frac{\partial l}{\partial \theta} \right)' \right] = -E \left[\frac{\partial^2 l}{\partial \theta \partial \theta'} \right]$$

3. Eficiencia asintótica. Si $\hat{\theta}$ es el estimador MV de un parámetro particular θ , esta propiedad significa que:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

para alguna constante finita σ^2 . Si $\tilde{\theta}$ denota algún otro estimador consistente y asintóticamente normal, entonces $\sqrt{n}\tilde{\theta}$ tiene una distribución normal asintótica cuya varianza es mayor o igual a σ^2 . El estimador MV tiene mínima varianza dentro de la clase de los estimadores consistentes y asintóticamente normales. El término **varianza asintótica** se refiere a la varianza de una distribución en el límite. De aquí que, la varianza asintótica de $\sqrt{n}\hat{\theta}$ es σ^2 . Sin embargo, el término también se utiliza para describir la varianza de la aproximación asintótica de la distribución desconocida de muestras finitas. Por lo que, una proposición equivalente es que la

varianza asintótica de $\hat{\theta}$ es σ^2/n . Cuando θ es un vector de parámetros y $\hat{\theta}$ es el vector de estimadores MV

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$$

para alguna matriz positiva V . Si \tilde{V} denota la matriz de varianzas-covarianzas de cualquier otro estimador consistente y asintóticamente normal, entonces $\tilde{V} - V$ es una matriz semidefinida positiva.

- 4. Invarianza.** Si $\hat{\theta}$ es el estimador MV de θ y $g(\theta)$ es una función continua de θ , entonces $g(\hat{\theta})$ es el estimador MV de $g(\theta)$.
- 5. El score tiene media cero y varianza $I(\theta)$.** Para demostrar que el valor de la media es cero debe notarse que integrando la función de densidad

$$\int \cdots \int f(y_1, y_2, \dots, y_n; \theta) dy_1 \cdots dy_n = \int \cdots \int L dy = 1$$

Diferenciando ambos lados con respecto a θ obtenemos

$$\int \cdots \int \frac{\partial L}{\partial \theta} dy = 0$$

Luego

$$\begin{aligned} E(s) &= \int \cdots \int \frac{\partial l}{\partial \theta} L dy \\ &= \int \cdots \int \frac{\partial L}{\partial \theta} dy = 0 \end{aligned}$$

Entonces la varianza del score es

$$\text{Var}(s) = E(ss') = E \left[\left(\frac{\partial l}{\partial \theta} \right) \left(\frac{\partial l}{\partial \theta} \right)' \right] = I(\theta)$$

3.2.5. La función de verosimilitud y el Jacobiano del modelo espacial general

Tal como se expuso en la Sección 3.2.2, la especificación para el modelo espacial general SAC es:

$$y = \rho W_1 y + X\beta + u \quad (3.54)$$

$$u = \lambda W_2 u + \varepsilon \quad (3.55)$$

$$\varepsilon \sim N(0, \Omega) \quad (3.56)$$

y los elementos de la diagonal de la matriz de varianzas-covarianzas del error Ω como:

$$\Omega_{ii} = h_i(z'_i \alpha) \quad h_i > 0 \quad (3.57)$$

En total, el modelo tiene $3 + k + p$ parámetros desconocidos, o, en forma de vector:

$$\theta = [\rho \quad \beta' \quad \lambda \quad \sigma^2 \quad \alpha']' \quad (3.58)$$

Este modelo también puede estar expresado en forma no lineal facilitándose la ilustración de algunos resultados relevantes. Se utiliza la siguiente simplificación notacional:

$$A = I - \rho W_1$$

$$B = I - \lambda W_2$$

la cual, para (3.54) y (3.55):

$$Ay = X\beta + u$$

(3.59)

$$Bu = \varepsilon \tag{3.60}$$

Además, debido a que la matriz de varianzas-covarianzas del error

$$E[\varepsilon\varepsilon'] = \Omega$$

es diagonal, existe un vector aleatorio de errores homoscedásticos v , como

$$\boldsymbol{v} = \boldsymbol{\Omega}^{-1/2} \boldsymbol{\varepsilon} \quad (3.61)$$

o, de forma alternativa,

$$\boldsymbol{\varepsilon} = \boldsymbol{\Omega}^{1/2} \boldsymbol{v} \quad (3.61)$$

y los errores en (3.60) se pueden escribir como

$$\boldsymbol{u} = \boldsymbol{B}^{-1} \boldsymbol{\Omega}^{1/2} \boldsymbol{v} \quad (3.62)$$

Sustituyendo (3.62) en (3.59) obtenemos:

$$\boldsymbol{A}\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{-1} \boldsymbol{\Omega}^{1/2} \boldsymbol{v}$$

o, la forma alternativa,

$$\boldsymbol{\Omega}^{-1/2} \boldsymbol{B}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{v} \quad (3.63)$$

En esta expresión no lineal (no lineal en los parámetros), \boldsymbol{v} es un vector de errores independientes y normalmente distribuidos con media igual a cero y varianza igual a uno. Consecuentemente, (3.63)

nos lleva a la expresión usual para la forma implícita de modelos no lineales

$$f(y, X, \theta) = \nu$$

donde f es una forma funcional no lineal que relaciona a y , X y a un vector de parámetros θ , y ν es el término de error.

Aunque el término de error ν tiene una distribución conjunta identificable, esta no puede ser observada, y la función de verosimilitud debe basarse en y . De aquí que, es necesario introducir el concepto de Jacobiano, el cual permite derivar la distribución conjunta de y a partir de la distribución conjunta de ν a través de la relación funcional expresada en (3.63).

El Jacobiano para la transformación del vector de variables aleatorias ν a el vector de variables aleatorias y es:

$$J = \left| \frac{\partial \nu}{\partial y} \right|$$

el cual, utilizando la expresión (3.63), es:

$$|\Omega^{-1/2}BA| = |\Omega^{-1/2}||B||A| \quad (3.64)$$

Basándose en la distribución normal estándar conjunta para el término v , y utilizando la expresión en (3.64), la función log-verosimilitud para la distribución conjunta de y se obtiene de la siguiente manera:

$$L = -(n/2)\ln(\pi) - (1/2)\ln|\Omega| + \ln|B| + \ln|A| - (1/2)v'v \quad (3.65)$$

con

$$v'v = (Ay - X\beta)'B'\Omega^{-1}B(Ay - X\beta) \quad (3.66)$$

como la suma de cuadrados de la correspondiente transformación de los términos del error.

El componente esencial de la función log-verosimilitud consiste en una forma cuadrática del término de error, la cual conduce a un problema de optimización no muy complejo. Sin embargo, los determinantes $|\Omega|$

, $|A|$ y $|B|$ en (3.65) pueden causar problemas en cierto sentido. De hecho, las propiedades asintóticas para los estimadores MV sólo se mantendrán si algunas condiciones de regularidad para la función de log-verosimilitud se satisfacen. Dentro de este contexto, ambos términos A y B pueden conducir a un comportamiento explosivo para valores particulares que puedan tomar ciertos parámetros, y Ω puede dejar de ser definida positiva.

Por esto, es necesario asegurar que la siguiente condición general se mantenga para el Jacobiano:

$$|\Omega^{-1/2}AB| > 0 \quad (3.67)$$

la cual se satisface con los requerimientos parciales:

$$|I - \rho W_1| > 0 \quad (3.68)$$

$$|I - \lambda W_2| > 0 \quad (3.69)$$

$$h_i(z'\alpha) > 0, \quad \forall i \quad (3.70)$$

Las restricciones (3.68) y (3.69) resultan en restricciones sobre los valores que pueden tomar los coeficientes autoregresivos espaciales. Para una matriz de conectividad espacial estandarizada esto significa que el valor para el parámetro debe estar entre menos uno y uno.

3.2.6. Condiciones de primer orden para los estimadores MV en el modelo general.

Las condiciones de primer orden para los estimadores MV en el modelo (3.63) se obtienen tomando derivadas parciales de la función log-verosimilitud (3.65) con respecto al vector de parámetros. El vector de primeras derivadas parciales resultante, el vector score, se iguala a cero y necesita ser resuelto para los valores de los parámetros:

$$d = (\partial L / \partial \theta) = 0$$

y como elementos de d:

$$\partial L / \partial \beta = v'(\Omega^{-1/2} BX) \quad (3.71)$$

$$\partial L / \partial \rho = -trA^{-1}W_1 + v'\Omega^{-1/2}BW_1y \quad (3.72)$$

$$\partial L / \partial \lambda = -trB^{-1}W_2 + v'\Omega^{-1/2}W_2(Ay - X\beta) \quad (3.73)$$

$$\partial L / \partial \alpha_p = -(1/2)tr\Omega^{-1}H_p + (1/2)v'\Omega^{-3/2}H_pB(Ay - X\beta) \quad (3.74)$$

$$p = 1, \dots, P$$

Este sistema de ecuaciones no lineales no tiene solución analítica y necesita ser resuelto por medio de métodos numéricos.

3.2.7. La matriz asintótica de varianzas del modelo general

Bajo las condiciones usuales de regularidad, los estimadores MV resultado de la solución del sistema de ecuaciones (3.71)-(3.74) serán asintóticamente eficientes. Esto significa que alcanzan la cota inferior de Rao-Cramer, dada por la inversa de la matriz de información:

$$[I(\theta)]^{-1} = -E[\partial^2 L / \partial \theta \theta']^{-1}$$

Los elementos de la matriz de información se derivan tomando las segundas derivadas parciales con respecto a los elementos del vector de parámetros θ y derivando los valores esperados relevantes mediante la utilización de la estructura de los términos de error dados en (3.55) y el sistema (3.71)-(3.74).

Para la variedad de combinaciones de parámetros, se obtienen los siguientes resultados:

$$I\beta\beta' = X'B'\Omega^{-1}BX$$

(3.75)

$$I\beta\rho = (BX)'\Omega^{-1}BW_1A^{-1}X\beta \quad (3.76)$$

$$I\beta\lambda = 0 \quad (3.77)$$

$$I\beta\alpha' = 0 \quad (3.78)$$

$$I\rho\rho = tr(W_1A^{-1})^2 + tr\Omega(BW_1A^{-1}B^{-1})'\Omega^{-1}(BW_1A^{-1}B^{-1}) \\ + (BW_1A^{-1}X\beta)'\Omega^{-1}(BW_1A^{-1}X\beta) \quad (3.79)$$

$$I\rho\lambda = tr(W_2B^{-1})'\Omega^{-1}BW_1A^{-1}B^{-1}\Omega + trW_2W_1A^{-1}B^{-1}$$

(3.80)

$$I\rho\alpha_p = tr\Omega^{-1}H_pBW_1A^{-1}B^{-1} \quad (3.81)$$

$$I\lambda\lambda = tr(W_2B^{-1})^2 + tr\Omega(W_2B^{-1})'\Omega^{-1}W_2B^{-1} \quad (3.82)$$

$$I\lambda\alpha_p = tr\Omega^{-1}H_pW_2B^{-1} \quad (3.83)$$

$$I\alpha_p\alpha_q = (1/2)tr\Omega^{-2}H_pH_q \quad (3.84)$$

La matriz de varianzas asintótica se obtiene sustituyendo los estimadores MV por los parámetros en las expresiones (3.75)-(3.84) y tomando la inversa de la matriz de información. La matriz de varianzas-covarianzas asintótica puede ser utilizada como la base para varias pruebas de hipótesis. Discutiremos este punto en la siguiente sección.

3.2.8. Pruebas de hipótesis basadas en el principio de MV

Dada la extensa aplicación del procedimiento de estimación de MV en la especificación de modelos de procesos espaciales, la mayoría de las pruebas de hipótesis sobre los parámetros de estos modelos también están basadas en consideraciones asintóticas. De las tres principales pruebas asintóticas, es decir, la prueba de Wald (W), la prueba de Razón de Verosimilitud (LR, del inglés: Likelihood Ratio), y

la prueba del Multiplicador de Lagrange (LM, del inglés: Lagrange Multiplier), las primeras dos han recibido mayor atención en la econometría espacial.

3.2.8.1. Principios generales

Las pruebas W, LR y LM se basan en las propiedades óptimas de los estimadores MV. Más específicamente, estas propiedades resultan de las distribuciones normales asintóticas de los estimadores y de funciones de los estimadores.

Formalmente, con h como el estimador MV del vector de parámetros θ :

$$n^{1/2}(h - \theta) \xrightarrow{d} N\left[0, \lim_{n \rightarrow \infty} (I(\theta) / n)^{-1}\right]$$

es decir, la diferencia entre el estimador y el parámetro poblacional converge en distribución a una distribución normal con un vector nulo

como media y una matriz de varianzas-covarianzas que corresponde a la inversa de la matriz de información.

La mayoría de las pruebas de hipótesis pueden ser formuladas como pruebas sobre funciones de los parámetros del modelo:

$$H_0 : g(\theta) = 0$$

$$H_1 : g(\theta) \neq 0$$

donde g es una función matricial de dimensión q lineal o no lineal en los elementos del vector de parámetros θ . Como caso especial, esta incluye pruebas de significancia de los coeficientes individuales: la función g se simplifica en un vector con un uno en el coeficiente de interés y ceros en las posiciones restantes.

En el modelo espacial general, el interés se centra en la significancia de los procesos autoregresivos (es decir, $H_0 : \rho = 0$, o $H_0 : \lambda = 0$), en los parámetros de los regresores β , y en pruebas para la presencia de heteroscedasticidad (es decir, con $H_0 : \alpha = 0$).

En términos generales, las tres pruebas asintóticas están basadas en diferentes medidas de la distancia entre el estimador sin restringir y el estimador que satisface las restricciones impuestas por la hipótesis nula, es decir, el estimador restringido. Por ejemplo, si el vector de parámetros θ se particiona en $\theta' = [\theta_1' \theta_2']$, la hipótesis nula puede ser de la forma:

$$H_0 : \theta_1 = 0$$

El estimador restringido θ_R consiste en los estimadores para θ_2 dejando los parámetros en θ_1 en cero. El estimador sin restricciones consiste en los estimadores para el vector completo θ . Luego, las pruebas estarán basadas en medidas de la diferencia entre los estimadores del vector completo θ y el estimador restringido θ_R . Intuitivamente, si la distancia entre los dos resultados es muy “grande”, las restricciones no se mantendrán y la hipótesis nula deberá rechazarse.

La manera en la cual las pruebas miden la distancia entre los estimadores restringidos y sin restringir también tiene implicaciones

respecto al tipo de estimadores necesarios para llevar a cabo las pruebas. Para la prueba de Wald se necesita estimar el modelo completo, es decir, se utilizan los parámetros sin restricciones. Para la prueba LM los coeficientes restringidos son el punto de partida, es decir, el modelo es estimado con las restricciones impuestas por la hipótesis nula. Para la prueba LR se necesitan ambos estimadores, restringidos y sin restringir.

El más familiar de los tres procedimientos de prueba es sin duda la prueba de Wald que está expresada en su forma más general como:

$$W = g'[GVG']^{-1}g$$

con

- g como un vector de dimensiones $q \times 1$ de valores que resultan cuando las restricciones son evaluadas para los estimadores MV de los parámetros;

- G como una matriz de dimensiones $qx(3+k+p)$ de derivadas parciales $\partial g'(\theta)/\partial\theta$ evaluadas para los estimadores de los parámetros;
- V como la matriz asintótica de varianzas estimada de dimensiones $3+k+p$.

El estadístico de Wald está asintóticamente distribuido como una variable χ^2 con q grados de libertad, donde q corresponde al número de restricciones. Debido a que la raíz cuadrada del estadístico corresponde a una variable aleatoria normal estándar, esta está asintóticamente distribuida como una variable t de Student. Consecuentemente, en el contexto de probar la significancia de los parámetros del modelo, la prueba de Wald es comúnmente conocida como la prueba t -asintótica.

La prueba LR se basa en la diferencia entre la log-verosimilitud para el modelo sin restricciones (con θ) y el modelo restringido (con θ_R):

$$LR = 2[L(\theta) - L(\theta_R)]$$

donde L son las correspondientes log-verosimilitudes. La prueba LR también está asintóticamente distribuida como una χ^2 con q grados de libertad.

La prueba LM, también conocida como la prueba del score, está basada en un procedimiento de optimización, o más específicamente, en las condiciones de primer orden para la optimización de una función lagrangiana en la log-verosimilitud:

$$f = L(\theta) + \eta'g(\theta)$$

donde f es el lagrangiano, L es la log-verosimilitud, y η es un vector de multiplicadores de lagrange correspondientes a las q restricciones $g(\theta)$. La prueba en sí está expresada en términos de los coeficientes restringidos, es decir, está basada en la estimación del modelo simple:

$$LM = d_R' I(\theta_R)^{-1} d_R$$

con

- d_R como el vector score para el modelo, $\partial L / \partial \theta$, evaluado en la hipótesis nula, es decir, con las restricciones impuestas;
- $I(\theta_R)$ es un estimador consistente de la matriz de información, también evaluado en la hipótesis nula.

El estadístico de prueba LM está también distribuido asintóticamente como una χ^2 con q grados de libertad.

3.2.8.2. La prueba de Wald para modelos de procesos espaciales

Las pruebas de mayor relevancia dentro de la econometría espacial son muy probablemente aquellas que tratan sobre la significancia individual de los coeficientes del modelo o sobre la significancia global del vector completo de parámetros. En ambos casos las restricciones sobre los parámetros son lineales, lo que simplifica las expresiones para la prueba de Wald.

Por ejemplo, considérese la prueba de significancia sobre el coeficiente autoregresivo espacial ρ en el modelo espacial. Las restricciones correspondientes pueden ser expresadas como:

$$H_0 : [1 \ 0'] [\rho \ \beta' \ \lambda \ \sigma^2 \ \alpha']' = \rho = 0$$

donde $0'$ es un vector fila de ceros de dimensión $2+k+p$. Consecuentemente, la derivada parcial de la restricción sobre ρ es igual al vector $[1 \ 0']$, y la prueba de Wald se reduce a:

$$\begin{aligned} W &= \hat{\rho} \{ [1 \ 0'] V [1 \ 0'] \}^{-1} \hat{\rho} \\ &= \hat{\rho}^2 / V_{11} \sim \chi^2(1) \end{aligned}$$

con $\hat{\rho}$ como el estimador MV para ρ , y V_{11} como el elemento diagonal correspondiente para ρ en la matriz de varianzas-covarianzas de los parámetros. Por otra parte,

$$\hat{\rho} / se_{11} \sim N(0,1)$$

con se_{11} como la desviación estándar estimada.

Extensiones de la hipótesis que consisten tanto en combinaciones lineales como no lineales de los parámetros del modelo pueden obtenerse con el mismo procedimiento. En todos los casos, la prueba de Wald necesita la estimación de modelo completo sin restricciones.

3.2.8.3. La prueba LR para modelos de procesos espaciales

Para obtener una prueba LR sobre la significancia de los parámetros del modelo, dos estimaciones son necesarias: una con el coeficiente en cuestión incluido, otra con el coeficiente excluido.

Por ejemplo, la prueba LR sobre la significancia del parámetro autoregresivo ρ en el modelo general estará basada en la log-verosimilitud del modelo completo y la log-verosimilitud del modelo sin el rezago espacial para la variable dependiente. Formalmente, como en (3.65):

$$L(\theta) = -(n/2) \ln(\pi) + \ln |\Omega^{-1/2} BA| - (1/2) v'v$$

La log-verosimilitud restringida es:

$$L(\theta_R) = -(n/2)\ln(\pi) + \ln|\Omega_R^{-1/2}B_R| - (1/2)v'v$$

con $v'v = (y - X\beta_R)'B_R'\Omega_R^{-1}B_R(y - X\beta_R)$

La prueba LR resultante no sólo consiste en la usual diferencia de cuadrados de los residuos ($v'v$ y $v'v$) sino, además, de la diferencia entre los determinantes de los Jacobianos:

$$L_R = (v'v - v'v) + 2\left\{\ln|\Omega^{-1/2}BA| - \ln|\Omega_R^{-1/2}B_R|\right\}$$

con los coeficientes evaluados respectivamente en sus estimadores restringidos y no restringidos. El estadístico resultante está asintóticamente distribuido como una χ^2 con un grado de libertad, o, de forma alternativa, su raíz cuadrada está distribuida como una variable normal estándar.

Expresiones similares pueden derivarse para los otros parámetros en el modelo, así como para los modelos más simples dentro de la especificación general.

3.2.8.4. La prueba LM para modelos de procesos espaciales

A diferencia de las pruebas de Wald y LR, las pruebas basadas en el principio del Multiplicador de Lagrange no necesitan la estimación del modelo más complejo. Para la especificación más general considerada aquí, esto implica que la estimación MCO puede ser suficiente para la mayoría de hipótesis de interés.

Más específicamente, considérese una partición del vector de parámetros del modelo general:

$$\theta' = [\rho \quad \lambda \quad \alpha' \quad / \quad \sigma^2 \quad \beta']$$

donde α ha sido normalizado para que la varianza común del error σ^2 pueda ser aislada. La hipótesis nula espacial más general está relacionada con ambos tipos de dependencia espacial (ρ y λ) así como con la heterogeneidad espacial (α), es decir:

$$H_0 : [\rho \quad \lambda \quad \alpha'] = 0$$

La prueba LM para esta hipótesis puede ser:

$$LM = d' I^{-1} d \sim \chi^2(2+P)$$

En esta expresión, d es el vector score con los elementos (3.71)-(3.74), evaluados bajo la hipótesis nula, es decir, con ρ , λ y α iguales a cero, y σ^2 y β estimados por medio de MCO. De hecho, bajo H_0 , el modelo general se reduce a una regresión lineal simple, con $A = B = I$ y $\Omega = \sigma^2 I$.

En muchas situaciones en la econometría tradicional, la matriz de información tiene una estructura diagonal por bloques en los coeficientes de interés, lo que facilita mucho la derivación de la matriz de información particionada I^{11} . La presencia de un rezago espacial para la variable dependiente, y la covarianza entre los estimadores de ρ y λ (y entre ρ y β) impide esta situación en el caso espacial. Como consecuencia, la expresión para la inversa I^{11} es más compleja.

3.2.9. Dependencia espacial término en el del error

Cuando los términos del error en un modelo de regresión muestran dependencia espacial no puede adoptarse el supuesto clásico de una matriz diagonal de varianzas-covarianzas del error. En lugar de esto, la matriz será de la forma general $\Omega(\theta)$, típicamente parametrizada en función de un pequeño número de coeficientes. En la econometría espacial, estos coeficientes están asociados con un patrón de interacciones espaciales o estructura espacial que se asume causa la dependencia.

Los supuestos más comúnmente utilizados para la forma de la dependencia espacial es una especificación espacial autoregresiva.

Para el modelo de regresión lineal simple

$$y = X\beta + u$$

tenemos para el vector de errores u :

$$u = \lambda W u + \varepsilon$$

donde λ es un coeficiente autoregresivo espacial, W es la usual matriz de conectividad espacial, y ε es un término que satisface los supuestos clásicos de independencia e igual distribución con varianza constante σ^2 . La correspondiente matriz de varianzas-covarianzas para el término del error es de la forma:

$$\Omega = E[uu'] = E\left\{[(I - \lambda W)^{-1} \varepsilon][(I - \lambda W)^{-1} \varepsilon]'\right\}$$

o también

$$\sigma^2 \Omega(\lambda) = \sigma^2 (I - \lambda W)^{-1} [(I - \lambda W)^{-1}]'$$

3.2.9.1. Pruebas para la autocorrelación espacial del error basadas en el principio de MV

Tal como se expuso en detalle en la Sección 3.2.8, para el modelo general espacial SAC, las pruebas de Wald, LR y LM son procedimientos asintóticos basados en el principio de Máxima Verosimilitud (MV). Las pruebas para la presencia de autocorrelación espacial en los residuos pueden formularse dentro de este contexto como casos especiales del modelo general.

De hecho, las tres pruebas pueden considerarse como diferentes tipos de procedimientos que tratan el problema de variables omitidas. Formalmente, la hipótesis nula y alternativa son:

$$H_0 : \lambda = 0$$

$$H_1 : \lambda \neq 0$$

El modelo de regresión con un término autoregresivo espacial del error es un caso especial del modelo espacial general con parámetros $\rho = 0$ y $\alpha = 0$. Formalmente, la función de log-verosimilitud para este modelo es:

$$L = -(n/2)\ln(\pi) - (n/2)\ln(\sigma^2) + \ln|B| - (1/2)\sigma^{-2}(y - X\beta)'B'(y - X\beta) \quad (3.85)$$

Bajo la hipótesis nula, es decir, con $\lambda = 0$, y de aquí también con $B = I$, tenemos la usual log-verosimilitud en el modelo de regresión lineal:

$$L_0 = -(n/2)\ln(\pi) - (n/2)\ln(\sigma^2) - (1/2)\sigma^{-2}(y - X\beta)'(y - X\beta) \quad (3.86)$$

La prueba LR para la autocorrelación espacial de los residuos está basada en la diferencia entre (3.85) y (3.86). Cuando los coeficientes en la log-verosimilitud son reemplazados por sus estimadores MV, el

resultado es una expresión mucho más simple para la log-verosimilitud:

$$LR = n[\ln(\sigma_0^2) - \ln(\sigma_1^2)] + 2 \ln |I - \lambda W| \sim \chi^2(1) \quad (3.87)$$

donde σ_0^2 es la varianza estimada del error para el modelo bajo la hipótesis nula (sin autocorrelación espacial, es decir, una regresión simple) y σ_1^2 es la varianza estimada del error para el modelo espacial. La presencia del determinante del Jacobiano en (3.87) diferencia este resultado del de autocorrelación serial para series de tiempo.

Para derivar la prueba de Wald es necesaria la matriz asintótica de varianzas de los estimadores MV en el modelo completo. Para el caso especial considerado aquí, la expresión presentada en la Sección 3.2.7 se simplifica considerablemente. De hecho, la matriz de información resultante es diagonal en bloques en los coeficientes de regresión y en los parámetros relacionados al error σ^2 y λ . La parte relacionada al error consiste en la siguiente matriz de dimensiones 2x2:

$$\begin{bmatrix} n/2\sigma^4 & \sigma^{-2}trWB^{-1} \\ \sigma^{-2}trWB^{-1} & tr(WB^{-1})^2 + tr(WB^{-1})'(WB^{-1}) \end{bmatrix}$$

Una inversión particionada de esta matriz para el elemento correspondiente a λ nos da:

$$Var(\lambda) = [t_2 + t_3 - (1/n)(t_1)^2]^{-1}$$

con la siguiente simplificación notacional:

$$\begin{aligned} t_1 &= trWB^{-1} \\ t_2 &= tr(WB^{-1})^2 \\ t_3 &= tr(WB^{-1})'(WB^{-1}) \end{aligned}$$

Consecuentemente, la prueba de Wald para la autocorrelación espacial del error es:

$$W = \lambda^2 [t_2 + t_3 - (1/n)(t_1)^2] \sim \chi^2(1) \quad (3.88)$$

donde el coeficiente λ se reemplaza por su estimador MV. De forma alternativa, la raíz cuadrada de esta expresión puede considerarse como una variable normal estándar.

Tal como se expresó anteriormente, la prueba LM se basa únicamente en la estimación bajo la hipótesis nula. Esto resulta en un estadístico de fácil implementación derivado de los residuos MCO y algunos cálculos adicionales de trazas de matrices de conectividad. Como se expuso en la Sección 3.2.8.4, este estadístico es de la forma:

$$LM = (1/T)[e'We / \sigma^2]^2 \sim \chi^2(1) \quad (3.89)$$

donde $T = \text{tr}\{(W + W')W\}$ y e es el vector de residuos MCO.

3.2.9.2. Prueba para la autocorrelación espacial del error en un modelo SAR

El tratamiento en esta sección difiere de las pruebas unidireccionales discutidas en la Sección 3.2.8.4. Allí, los otros parámetros en el

modelo espacial son a priori igualados a cero (es decir, $\rho = 0$, $\alpha = 0$). Aquí, algunos de estos parámetros están incluidos en la estimación del modelo, es decir, se asume que son diferentes de cero. El caso considerado en esta sección consiste en la prueba para la autocorrelación espacial del error en presencia de un rezago espacial para la variable dependiente. En términos de los coeficientes del modelo general, esto significa que $\rho \neq 0$, y $\alpha = 0$.

En la notación de la Sección 3.2.5, el modelo bajo consideración puede expresarse como:

$$B(Ay - X\beta) = \varepsilon \quad (3.90)$$

con

$$E[\varepsilon\varepsilon'] = \Omega = \sigma^2 I$$

donde A y B están asociadas respectivamente con el rezago espacial para la variable dependiente (parámetro ρ) y la dependencia espacial en el término de error (parámetro λ). Además, las matrices de conectividad W_1 y W_2 en A y B no tienen que ser iguales.

Siguiendo el procedimiento usual para la prueba LM, el vector de coeficientes se particiona en términos de los parámetros considerados en la hipótesis nula y los otros parámetros. Para el modelo (3.90), se tiene:

$$\theta = [\lambda \quad / \quad \rho \quad \beta' \quad \sigma^2]$$

Consecuentemente, bajo la hipótesis nula, con $\lambda = 0$, la especificación resulta:

$$Ay - X\beta = \varepsilon$$

es decir, un modelo mixto regresivo-espacial autoregresivo (SAR) con términos estándar para el error.

La prueba LM se basa en el vector score y la matriz de información particionada bajo la hipótesis nula. Las expresiones relevantes pueden encontrarse como casos especiales de la formulación general en la Sección 3.2.7 imponiendo las restricciones $\lambda = 0$, $B = I$, y $\Omega = \sigma^2 I$. Puede derivarse que:

$$\partial L / \partial \lambda = \sigma^{-2} (Ay - X\beta)' W_2 (Ay - X\beta)$$

(3.91)

y

$$I\lambda\lambda = [T_{22} - (T_{21A})^2 \text{var}(\rho)]^{-1}$$

con la siguiente simplificación notacional:

$$T_{22} = \text{tr}\{W_2 W_2 + W_2' W_2\}$$

$$T_{21A} = \text{tr}\{W_2 W_1 A^{-1} + W_2' W_1 A^{-1}\}$$

y $\text{Var}(\rho)$ como la varianza estimada de ρ en el modelo bajo la hipótesis nula.

Utilizando el mismo procedimiento general descrito con anterioridad, el estadístico LM para $H_0 : \lambda = 0$ puede expresarse como:

$$(e' W_2 e / \sigma^2)^2 \{T_{22} - (T_{21A})^2 \text{var}(\rho)\}^{-1} \sim \chi^2(1) \quad (3.92)$$

o, de forma equivalente, como:

$$(e'W_2e / \sigma^2) \{T_{22} - (T_{21A})^2 \text{var}(\rho)\}^{-1/2} \sim N(0,1)$$

Este estadístico puede obtenerse fácilmente de la estimación de MV del modelo SAR con la evaluación adicional de un producto cruzado y las trazas relevantes.

3.2.10. Selección de modelos

En esta sección, la validación estadística de modelos econométricos está basada en técnicas de discriminación. En este contexto, la selección de un modelo entre varios modelos alternativos está considerada como un problema de decisión. Más específicamente, la atención se centra en la naturaleza de la compensación entre el

número de parámetros incluidos en el modelo (principio de parsimonia) y el ajuste del mismo, en el sentido de que tan bien el modelo estima las observaciones de la muestra.

3.2.10.1. Medidas de ajuste en los modelos espaciales

La evaluación del ajuste empírico de un modelo estimado es un aspecto importante en el análisis econométrico. En la econometría espacial esta tarea es un poco más compleja debido a la falta de una medida estándar como el coeficiente de determinación en la econometría tradicional. Aunque esta medida es utilizada rutinariamente en el trabajo empírico, su interpretación en presencia de los efectos espaciales conduce a conclusiones erróneas, y no tiene una relación directa sobre la prueba de la significancia global del modelo estimado (prueba ANOVA). Dos situaciones diferentes pueden distinguirse, una cuando la dependencia espacial está presente en el término del error, la otra cuando se incluye un rezago espacial de la variable dependiente en el modelo.

En el primer caso, se obtienen las mismas características para un modelo con una matriz no escalar de varianzas-covarianzas del error. Específicamente, debido a que la estimación está basada en el principio de MV o en MCG, los residuos del modelo estimado no tendrán media cero y no puede efectuarse la descomposición estándar de la variabilidad observada en variabilidad explicada y variabilidad residual. Adicionalmente, la equivalencia del coeficiente de determinación a la correlación al cuadrado entre los valores observados y estimados no tiene validez.

Un número de pseudo valores R^2 pueden reportarse como alternativas que reproducen ciertos aspectos de la medida estándar. Por ejemplo, el coeficiente de correlación al cuadrado entre los valores observados y estimados de la variable dependiente aún provee una medida de asociación lineal (o capacidad para predecir en la muestra) que se encuentra entre cero y uno, aunque no se relaciona con la descomposición de varianza.

Cuando un rezago espacial de la variable dependiente es incluido, el procedimiento de estimación utilizado es el de MV y la medida

estándar del coeficiente de determinación es inválida. Una medida de ajuste más apropiada está basada en la log-verosimilitud maximizada, aunque un pseudo R^2 , en la forma de un coeficiente de correlación lineal entre los valores observados y estimados de la variable dependiente, puede ser utilizado.

Desde el punto de vista del análisis econométrico, uno de los problemas con estas medidas es su falta de ajuste con respecto a la inclusión de variables adicionales, pues estas tienden a inflar el indicador. Algunos criterios de selección sugieren tomar en cuenta explícitamente una compensación entre el ajuste del modelo a los valores de la muestra y el número de parámetros utilizados. Estos criterios pueden distinguirse por la función objetivo utilizada como la base para la selección, es decir, por la función de pérdida o riesgo asociada. Algunos ejemplos son el coeficiente de determinación ajustado \bar{R}^2 y el coeficiente C_p de Mallow, los cuales se presentan rutinariamente en los paquetes de econometría para computadora. Sin embargo, estas medidas de ajuste necesitan interpretarse con precaución debido a que pueden conducir a conclusiones erróneas en presencia de los efectos espaciales.

3.2.10.2. Criterios de información para la discriminación de modelos

Las medidas teóricas de información están directamente relacionadas al concepto de **entropía**. La utilización de la teoría de información como una herramienta en la construcción de modelos y validación de los mismos en la econometría se enfoca principalmente en la obtención de una medida de la “cercanía” de un modelo asumido del modelo verdadero pero desconocido.

Operacionalmente, los modelos alternativos bajo consideración están conceptualizados como funciones de densidad probabilística en la variable dependiente observada. Se realiza una distinción entre la densidad para el verdadero pero desconocido modelo $g(y)$ y el modelo postulado $h(y/X, \theta)$. Como es usual, el último es condicional sobre las variables explicativas o regresores y un conjunto de parámetros.

Una medida teórica de información para el ajuste del modelo propuesto es el criterio de información de Kullback-Leibler o KLIC. El KLIC es similar a un valor esperado del ratio de las log-verosimilitudes del modelo verdadero y el modelo postulado, evaluado con respecto a la verdadera pero desconocida distribución subyacente G :

$$KLIC = E_G \{ \log [g(y) / h(y / X, \theta)] \}$$

El negativo del KLIC, el cual es equivalente a la entropía del modelo postulado con respecto al verdadero, forma una medida natural de ajuste. Debido a que el valor esperado en el KLIC, o entropía, implica una distribución desconocida, un criterio práctico debe basarse sobre un estimador apropiado. El criterio de información de Akaike (AIC, del inglés: Akaike Information Criterion) es un estimador de este valor esperado que maximiza la entropía.

En terminos generales, este medida puede expresarse formalmente como:

$$AIC = -2L + q(K)$$

donde L es la función log-verosimilitud maximizada, K es el número de parámetros desconocidos y q es un factor de corrección. El factor de corrección varía entre las diferentes versiones de la medida de información. En la formulación original de Akaike, el factor multiplicativo simple $q = 2K$ es utilizado. La idea general es que el AIC corrige o penaliza la evaluación del ajuste del modelo dado por la log-verosimilitud maximizada por un factor que refleja el número de parámetros.

CAPÍTULO 4

4. ESPECIFICACIÓN DE UN MODELO ECONONOMÉTRICO PARA LA PRODUCCIÓN DE CAÑA DE AZÚCAR EN EL INGENIO SAN CARLOS.

En este capítulo se desarrolla el análisis econométrico de la producción de caña de azúcar en el ISC con respecto a aquellos factores naturales y de manejo de los que se pudo obtener la mayor cantidad de información en una muestra de 871 lotes de los 1309 establecidos como resultado de la adopción de la AP en el ingenio. El Capítulo 2 describe a la producción y a cada uno de sus factores determinantes de manera individual respondiendo a la pregunta: ¿cuánto varían las características medidas del suelo y del cultivo?; en este capítulo el principal objetivo es responder a la pregunta: ¿cuánto afectan estas variaciones a la producción?.

La tarea fundamental es determinar la mejor especificación para un modelo de la respuesta sitio-específica del cultivo a sus factores determinantes (relación causa-efecto) con la finalidad de establecer un diagnóstico de la situación del cultivo en el período 2002-2003, y así, disminuir la brecha de información existente entre el análisis de datos y las recomendaciones sitio-específicas.

Se formulan dos especificaciones, una tradicional utilizando estimadores MCO y otra espacial utilizando estimadores MV. De acuerdo al marco teórico del Capítulo 3, se escogerá la mejor especificación mediante pruebas de diagnóstico y medidas de ajuste.

4.1. DEFINICIÓN DE VARIABLES

El primer paso en la formulación de una especificación de un modelo econométrico para la producción de caña de azúcar es representar a cada uno de los factores determinantes mediante variables cualitativas y cuantitativas. Las variables cuantitativas toman valores sobre una escala continua y las variables cualitativas se representan mediante variables ficticias o “*dummies*” que toman valores sobre una escala discreta. Para la investigación, las variables ficticias

toman el valor de 1 si el lote presenta cierta característica cualitativa y 0 en caso contrario. De esta manera subdividimos a la muestra de 871 lotes en varias submuestras dependiendo de una determinada categoría. Las variables cualitativas y cuantitativas que representan a los factores de producción en los modelos econométricos se muestran en el Cuadro 4.1.

Por otra parte, además de las variables presentadas en el Cuadro 4.1, se tomaron en cuenta otras adicionales que representan las siguientes interacciones entre factores: 1) tercio de cultivo y textura de suelos; 2) variedades de caña y textura de suelos; y, 3) variedades y tercio de cultivo. Por ejemplo, para representar la interacción de la textura franco arenosa con la variedad PR-671070 se diseñó la variable cualitativa FARE*PR.

Cuadro 4.1. Definición de variables para los modelos econométricos

Factor	Variables Cualitativas	Lotes
Textura de suelos	F: T. franco	155
	FARC: T. franco arcillosa	272
	FARE: T. franco arenosa	138
	ARC: T. arcillosa	198
	OTX: Otras texturas (arenosa, arenosa limosa, franco limosa, f. arcillosa limosa, f. arcillosa arenosa, arcillosa limosa)	108
Variedad de caña	RAGNAR	444
	PR-671070	86
	CC-8592	64
	CR-74250	176
	OTV: Otras variedades (MISCELANEA, MCO-310, BJ-7046, B-74132)	101
Sector	S1: sector 1	124
	S2: sector 2	134
	S3: sector 3	140
	S4: sector 4	197
	S5: sector 5	164
	S6: sector 6	112
Tercio de cultivo	T1: tercio 1	310
	T2: tercio 2	308
	T3: tercio 3	253
Factor	Variables cuantitativas	Lotes
Análisis de suelos	P: Fósforo	871
	K: Potasio	871
	Ca: Calcio	871
	Mg: Magnesio	871
	MO: Materia orgánica	871
	CE: Conductividad eléctrica	871
	pH: Potencial hidrógeno	871
Características fisiológicas	Edad cultivo	871
	Edad cosecha	871
Número de riegos	RIEGOS	871

Elaboración: Xavier Vítóres, 2004

4.2. ESPECIFICACIÓN ESPACIAL MV

El primer aspecto que se consideró para la formulación de la especificación general espacial SAC (modelo MV) del modelo de producción, mostrado en el Cuadro 4.2, es la incorporación de los efectos espaciales de los datos. La dependencia espacial de las observaciones debe considerarse, en primer lugar, debido a que se espera que los valores de ciertas variables medidos en los lotes se parezcan a medida que la distancia entre ellos disminuya; es decir, se espera la existencia de autocorrelación espacial. Las principales variables en las que se espera este comportamiento espacial son la producción y algunos parámetros del análisis de suelos (K, P, Ca y Mg).

En segundo lugar, los errores de medición en unidades espaciales contiguas (lotes) también presentarán cierta relación de tipo espacial en la medida en que el fenómeno bajo estudio no este correctamente representado por el nivel de agregación de datos, es decir, en la medida en que el diseño de los lotes no represente de manera adecuada el sistema productivo.

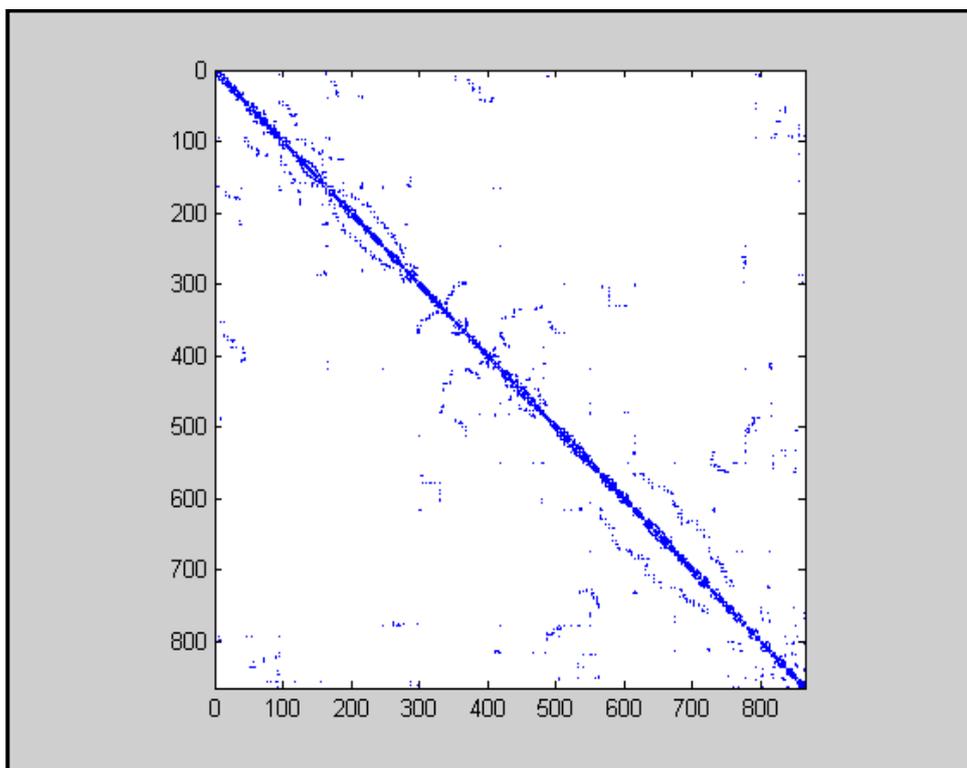


Figura 4.1. Matriz de conectividad para la muestra de lotes del ISC

Para incorporar la estructura de dependencia espacial de los datos en el modelo es necesario el diseño de una matriz de conectividad espacial W (Figura 4.1). Esta es una matriz cuadrada simétrica con 871 filas y columnas. Cada punto en la matriz indica que lotes son vecinos.

Cuadro 4.2. Especificación espacial MV

MODELO MV	Coeficientes		t- asintótica	Sig.	Intervalo 95% de confianza para B	
	B	D.E.			L.I.	L.S.
C	50,582085	3,308383	15,289065	0,000000	44,097654	57,066516
S3	9,103924	2,047037	4,447366	0,000009	5,091731	13,116117
RIEGOS	0,749208	0,254038	2,949196	0,003186	0,251293	1,247123
T1	22,428621	1,608758	13,941578	0,000000	19,275456	25,581786
T2	14,195213	1,184136	11,987821	0,000000	11,874306	16,516120
T1*OTV	-6,118555	2,060145	-2,969963	0,002978	-10,156440	-2,080670
T3*PR	7,531252	3,303388	2,279857	0,022616	1,056612	14,005892
T1*FARC	3,589242	1,546870	2,320326	0,020323	0,557377	6,621107
FCC	7,400152	3,185313	2,323210	0,020168	1,156938	13,643366
FARE*CC	10,250813	4,685499	2,187774	0,028686	1,067235	19,434391
FARE*PR	8,222838	2,701358	3,043965	0,002335	2,928177	13,517499
Edad cultivo	-1,999399	0,230529	-8,673109	0,000000	-2,451235	-1,547563
CC-8592	4,465887	1,747494	2,555595	0,010601	1,040799	7,890975
CR-74250	9,217573	1,197478	7,697489	0,000000	6,870516	11,564630
PR-671070	-6,329794	1,473990	-4,294325	0,000018	-9,218815	-3,440773
F	3,459848	1,068216	3,238904	0,001200	1,366145	5,553551
FARC	1,900156	1,065426	1,783471	0,074510	-0,188079	3,988391
K	3,382351	1,696099	1,994194	0,046131	0,057996	6,706706
pH	-1,884619	0,692733	-2,720557	0,006517	-3,242375	-0,526863
CE	-0,026363	0,010294	-2,561036	0,010436	-0,046539	-0,006187
Edad cosecha	2,114313	0,505999	4,178493	0,000029	1,122555	3,106071
Ca	0,363828	0,122349	2,973679	0,002943	0,124023	0,603633
Mg	-0,426553	0,139906	-3,048856	0,002297	-0,700769	-0,152337
Rho	-0,090999	0,019685	-4,622799	0,000004	-0,129581	-0,052417
lambda	0,611000	0,033143	18,435097	0,000000	0,546039	0,675961

La incorporación de la heterogeneidad espacial se lleva a cabo mediante el diseño de regímenes espaciales que permitan el cambio estructural de la especificación de acuerdo a subconjuntos de datos. Específicamente, se estudia el grado en que las diferentes regiones que representan a diferentes texturas, variedades, sectores y tercios

de cultivo afectan la magnitud, la significancia y el signo de los coeficientes estimados en el modelo. Esta tarea se realiza mediante la utilización de las variables cualitativas.

La primera columna del Cuadro 4.2 corresponde a las variables definidas en la sección anterior; la segunda columna de coeficientes muestra los estimadores para los parámetros de la FRP (B) con sus desviaciones estándar (D.E.); la tercera columna muestra el estadístico de prueba t-asintótica para la hipótesis nula de que el parámetro asociado a la respectiva variable es igual a cero; la cuarta columna muestra el nivel de significancia para el estadístico de prueba; y, la quinta columna muestra un intervalo del 95% de confianza para los parámetros: L.I. es el límite inferior y L.S. es el límite superior. Como puede observarse, todos los coeficientes estimados, incluyendo los de los rezagos espaciales, son significativos al nivel 0,05 a excepción del coeficiente estimado para la variable FARC que es significativo al nivel 0,1.

Debido a la presencia de un rezago espacial para la variable dependiente (producción) y para el error estocástico, no todas las medidas de ajuste del Cuadro 4.3 resultan adecuadas. La medida más apropiada de ajuste debe basarse en el valor de la función

de log-verosimilitud maximizada, es decir, -2404,6969; o, en el valor del estadístico AIC que es 4859,3938. Estas medidas son utilizadas para realizar comparaciones entre modelos. El coeficiente de correlación (R) entre los valores reales y los estimados es de 0,861. La Figura 4.2 muestra una representación gráfica de los valores reales, los estimados y los residuos.

Cuadro 4.3. Medidas de ajuste del modelo MV

R	R ²	R ² Ajust.	Log-verosimilitud	AIC	D.E. Estimación
0,861	0,7427	0,7348	-2404,6969	4859,3938	9,18923

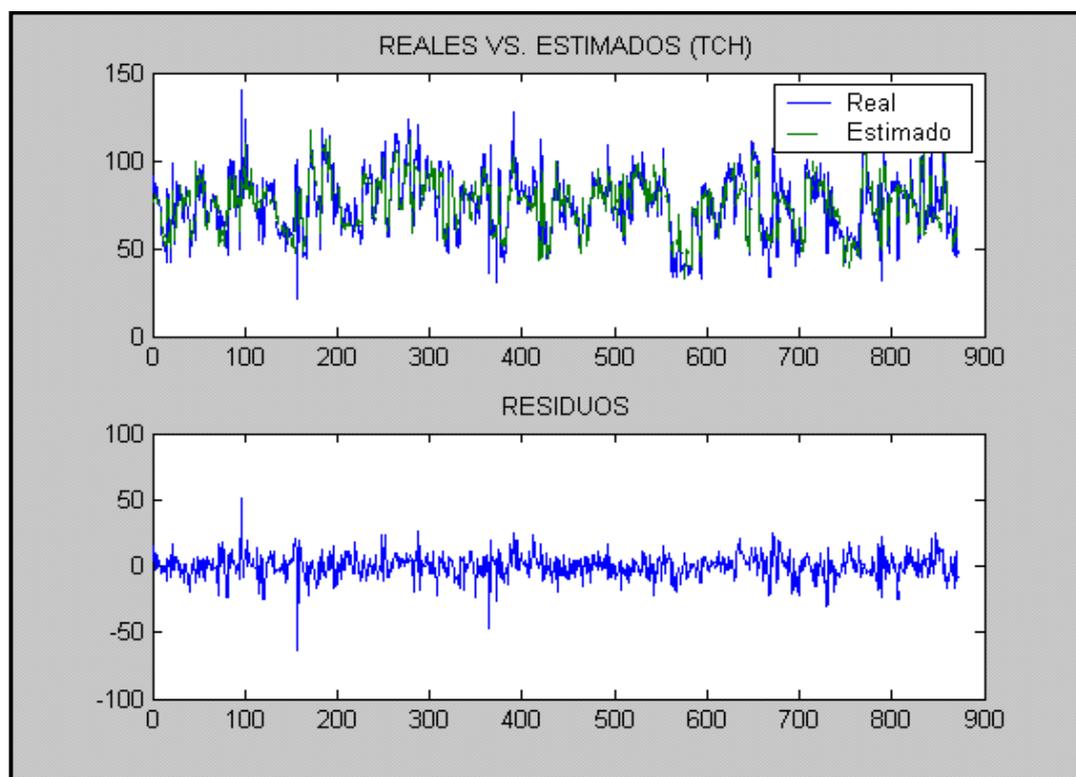


Figura 4.2. Representación gráfica de los valores reales, estimados y residuos del modelo MV

4.2.1. Análisis de los residuos

Los residuos de un modelo se constituyen en un elemento importante para describir que tan bien se ajusta la especificación formulada a la realidad. La estadística descriptiva de los residuos del modelo espacial MV se muestra en el Cuadro 4.4.

Cuadro 4.4. Estadística descriptiva de los residuos del modelo MV

N		871
Media		,0000
Mediana		,3876
Desviación Estándar		9,22546
Varianza		85,109
Rango		114,42
Mínimo		-63,14
Máximo		51,28
Percentiles	25	-5,0859
	50	,3876
	75	5,3539

La media de los residuos es cero con una desviación estándar o error estándar de la estimación de 9,22546. El diagrama de caja de la Figura 4.3 muestra los percentiles 25 y 75, mediana (percentil 50) y los valores extremos o atípicos. El 50% de los residuos se encuentran entre $-5,0859$ y $5,3539$. El máximo es $51,28$ y el mínimo es $-63,14$.

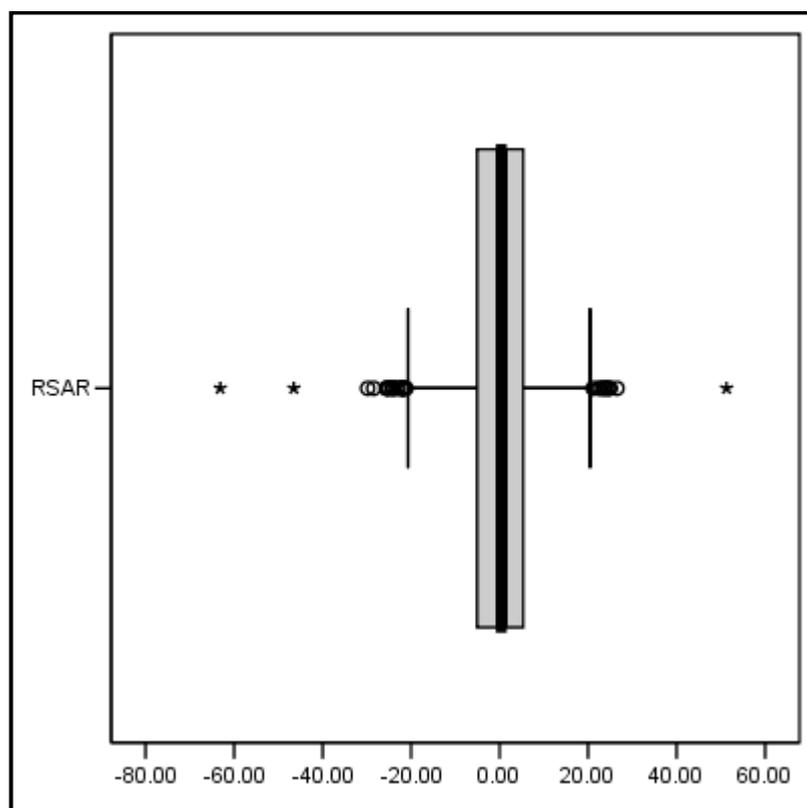


Figura 4.3. Diagrama de caja de los residuos del modelo MV

El histograma de los residuos de la Figura 4.4 muestra que la distribución de los residuos se aproxima a una distribución normal (línea acampanada). El proceso de contrastar la normalidad del componente de error basándose en los residuos es importante para realizar inferencias sobre la significancia estadística de los coeficientes estimados. En la siguiente sección se llevarán a cabo algunas pruebas de diagnóstico que permitirán concluir sobre este supuesto.

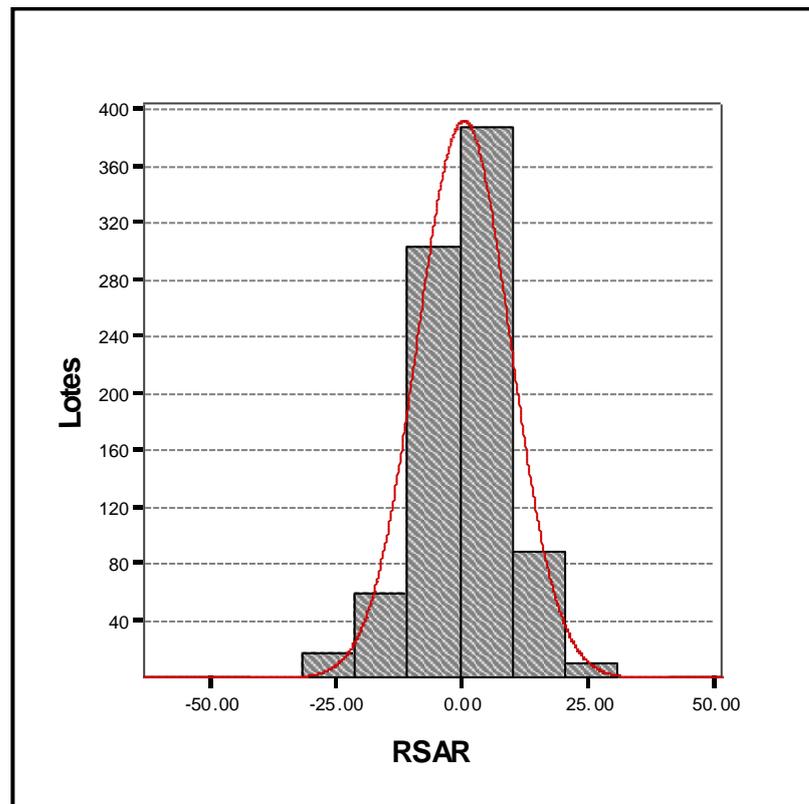


Figura 4.4. Histograma de los residuos del modelo MV

4.3. ESPECIFICACIÓN NO ESPACIAL MCO

La especificación no espacial que se muestra en el Cuadro 4.5 corresponde a la estimación de MCO del modelo para la producción de caña de azúcar. La primera columna del cuadro corresponde a las variables; la segunda columna muestra los estimadores para los parámetros de la FRP (B) con sus desviaciones estándar (D.E.); la tercera columna muestra el estadístico de prueba t para la hipótesis

nula de que el parámetro es igual a cero; la cuarta columna muestra el nivel de significancia del estadístico de prueba; la quinta columna muestra un intervalo del 95% de confianza para los parámetros: L.I. es el límite inferior y L.S. es el límite superior; y, la sexta columna muestra los valores del factor de inflación de varianza (FIV) de cada variable.

Los valores del FIV no indican un serio problema de multicolinealidad (no es mayor que 10 en ningún caso). Las medidas de ajuste del Cuadro 4.6 muestran un coeficiente de correlación (R) de 0,8221, un coeficiente de determinación (R^2) de 0,6774, un coeficiente de determinación ajustado (R^2 Ajustado) de 0,665, Log-verosimilitud de $-3771,090$, AIC de 6588,18 y una desviación estándar de la estimación (D.E. Estimación) de 10,48521.

Cuadro 4.5. Especificación no espacial MCO

MODELO MCO	Coeficientes		t	Sig.	Intervalo 95% confianza para B		FIV
	B	D.E.			L.I.	L.S.	
Variables							
C	33,911	10,106	3,356	,001	14,076	53,746	
S3	8,453	1,275	6,627	,000	5,949	10,956	1,738
RIEGOS	,553	,220	2,515	,012	,121	,984	2,787
T1	22,548	1,477	15,268	,000	19,649	25,446	3,961
T2	15,213	1,085	14,023	,000	13,083	17,342	2,131
T1*OTV	-5,652	2,145	-2,635	,009	-9,862	-1,441	1,173
T3*PR	6,867	3,441	1,996	,046	,114	13,620	1,274
T1*FARC	3,062	1,595	1,919	,055	-,069	6,193	2,511
F*CC	8,291	3,435	2,414	,016	1,549	15,034	1,479
FARE*CC	12,202	5,537	2,204	,028	1,335	23,069	1,110
FARE*PR	12,756	3,068	4,158	,000	6,735	18,777	1,262
Edad cultivo	-1,913	,222	-8,629	,000	-2,348	-1,478	1,868
CC-8592	2,916	1,828	1,595	,111	-,672	6,504	1,802
CR-74250	6,553	1,163	5,633	,000	4,270	8,836	1,729
PR-671070	-8,829	1,511	-5,843	,000	-11,795	-5,863	1,610
F	5,708	1,065	5,358	,000	3,617	7,798	1,315
FARC	4,414	1,075	4,107	,000	2,304	6,523	1,965
K	4,157	1,653	2,515	,012	,912	7,402	1,366
pH	-,714	1,062	-,673	,501	-2,798	1,369	1,600
CE	-,030	,011	-2,627	,009	-,052	-,008	1,364
Edad cosecha	2,349	,626	3,751	,000	1,120	3,578	1,324
Ca	,521	,117	4,469	,000	,292	,750	2,921
Mg	-,732	,143	-5,114	,000	-1,013	-,451	3,219

Cuadro 4.6. Medidas de ajuste del modelo MCO

R	R ²	R ² Ajust.	Log-verosimilitud	AIC	D.E. Estimación
0,821	0,674	0,665	-3271,090	6588,18	10,48521

La significancia global del modelo se determina mediante el análisis de varianza (ANOVA) del Cuadro 4.7.

Cuadro 4.7. Análisis de varianza del modelo MCO

Fuente de variación	Suma de cuadrados	g.l.	Cuadrados medios	F	Sig.
Regresión	192630,165	22	8755,917	79,643	,000
Residual	9322,8731	848	109,940		
Total	285858,896	870			

4.4. EVALUACIÓN DE LAS ESPECIFICACIONES

El Cuadro 4.8 muestra las pruebas de diagnóstico más importantes para las especificaciones MCO y MV. En primer lugar, las pruebas de presencia de heteroscedasticidad de White tanto para términos no cruzados como cruzados en el modelo MCO no son determinantes en la medida en que la primera rechaza el supuesto de homoscedasticidad (Sig. 0,034735) y la otra no (Sig. 0,528846). Luego, las pruebas LR, Wald y LM para determinar la presencia de autocorrelación espacial del término de error en el modelo no espacial MCO indican claramente la existencia del problema con niveles de significancia prácticamente iguales a cero (Sig. igual a 0,000000 en todas las pruebas). La prueba LMSAR para determinar la existencia de autocorrelación espacial del error en un modelo SAR indica la presencia de autocorrelación de este tipo a pesar de haber

incluido un rezago espacial para la variable dependiente en el modelo. Finalmente, la prueba de Kolmogorov-Smirnov para los residuos en el modelo MV (SAC) permite la posibilidad de adoptar el supuesto de normalidad del termino del error (Sig. 0,051).

Cuadro 4.8. Pruebas de diagnóstico

Pruebas de diagnóstico	g.l.	valor	Sig.
Heteroscedasticidad White (términos no cruzados MCO)	30	45,48967	0,034735
Heteroscedasticidad White (términos cruzados MCO)	275	195,8999	0,528846
LR (error MCO)	1	129,2769	0,000000
Wald (error MCO)	1	1914,697	0,000000
LM (error MCO)	1	137,5619	0,000000
LMSAR(error SAR)	1	391,8294	0,000000
Kolmogorov-Smirnov Z (error est. SAC)		1,355	0,051000

Por otra parte, de acuerdo a las medidas de ajuste, el valor del estadístico AIC, que es el más apropiado para realizar comparaciones entre las dos especificaciones, es menor en la especificación espacial con un valor de 4859,3938 que en la especificación no espacial con un valor de 6588,18. Los coeficientes de correlación lineal (R) entre los valores reales y los estimados también muestran diferencias: el valor del estadístico para el modelo espacial es de 0,861 que es mayor a 0,821, el valor para el modelo no espacial.

En conclusión, los resultados de las pruebas de diagnóstico y de ajuste indican que la mejor especificación es la espacial MV o SAC, la cual incorpora los efectos espaciales de los datos: autocorrelación espacial, incluyendo dos rezagos espaciales tanto para la variable dependiente como para el término del error estocástico; y, la heterogeneidad espacial con cambios estructurales en la especificación en forma de regímenes espaciales (diferentes coeficientes en subconjuntos espaciales del conjunto de datos) por sectores, variedades, texturas de suelo y tercios de cultivo. Los estimadores MCO de la especificación no espacial son sesgados, inconsistentes e ineficientes. Además, los valores para las desviaciones estándar del Cuadro 4.5 no corresponden a la correcta estructura de la matriz de varianzas y covarianzas de los estimadores MCO mostrada en la expresión 3.41 del Capítulo 3. Los estimadores MV de la especificación espacial SAC son consistentes, asintóticamente insesgados y son asintóticamente eficientes si adoptamos el supuesto de homoscedasticidad del término del error.

4.5. ANÁLISIS DE SENSIBILIDAD

El análisis de sensibilidad en esta sección tiene como objetivo determinar que tanto se ven afectadas las conclusiones expuestas en la sección anterior después de haber eliminado los valores extremos encontrados en el análisis de los residuos de la especificación MCO. Adoptando el supuesto de que los residuos estandarizados del modelo tienen una distribución normal estándar y conociendo que la probabilidad de encontrar valores alejados en más de 3 unidades en esta distribución es muy pequeña (aproximadamente 0,0027) se especula que un residuo es extremo o atípico si su valor estandarizado es mayor que 3 en valor absoluto. Los valores extremos se muestran en el Cuadro 4.9.

Cuadro 4.9. Residuos estandarizados extremos o atípicos

Lote	Residuo Estandarizado	TMC/HA	Valor Estimado	Residuo
022101	5,407	140,24	83,5455	56,69471
023603	-5,719	20,89	80,8543	-59,96745
051702	-4,619	36,29	84,7197	-48,42602
060404	3,124	127,57	94,8157	32,75392
101002	-3,257	47,35	81,4939	-34,14608

Los cambios más importantes que pueden observarse para la especificación no espacial MCO sin atípicos que se muestra en el Cuadro 4.10 están en las medidas de ajuste.

Cuadro 4.10. Especificación no espacial MCO sin valores atípicos

MODELO MCO	Coeficientes		t	Sig.	Intervalo 95% de confianza para B		VIF
	B	D.E.			L.I.	L.S.	
Variables							
C	32,925	9,503	3,465	,001	14,273	51,577	
S3	8,335	1,198	6,956	,000	5,983	10,687	1,739
RIEGOS	,587	,206	2,845	,005	,182	,992	2,778
T1	22,411	1,387	16,161	,000	19,689	25,133	3,944
T2	15,516	1,020	15,217	,000	13,515	17,517	2,117
T1*OTV	-5,582	2,015	-2,771	,006	-9,536	-1,627	1,174
T3*PR	7,100	3,230	2,198	,028	,759	13,440	1,274
T1*FARC	2,756	1,506	1,831	,067	-,199	5,711	2,518
F*CC	7,531	3,234	2,329	,020	1,184	13,878	1,486
FARE*CC	11,145	5,203	2,142	,032	,932	21,358	1,112
FARE*PR	12,855	2,880	4,463	,000	7,201	18,508	1,262
Edad de cultivo	-1,838	,208	-8,824	,000	-2,247	-1,429	1,865
CC-8592	3,942	1,733	2,275	,023	,541	7,343	1,810
CR-74250	6,865	1,099	6,246	,000	4,707	9,022	1,718
PR-671070	-8,886	1,419	-6,263	,000	-11,671	-6,101	1,609
F	5,632	1,001	5,628	,000	3,668	7,597	1,315
FARC	4,278	1,017	4,204	,000	2,281	6,275	1,977
K	3,688	1,553	2,374	,018	,639	6,737	1,366
pH	-1,061	,999	-1,063	,288	-3,022	,899	1,599
CE	-,029	,011	-2,666	,008	-,050	-,008	1,363
Edad de cosecha	2,509	,589	4,261	,000	1,353	3,664	1,326
Ca	,597	,110	5,435	,000	,382	,813	2,926
Mg	-,792	,135	-5,873	,000	-1,056	-,527	3,228

En el Cuadro 4.11 podemos observar el aumento en los valores para R de 0,821 a 0,838; para R^2 de 0,674 a 0,70; para R^2 Ajustado de 0,665 a 0,694; para Log-verosimilitud de -3271,090 a -3197,626; la disminución del valor de AIC de 6588,18 a 6441,252; y, la disminución del error de estimación de 10,48521 a 9,84416. Estos cambios indican un mejor ajuste del modelo a los datos; sin embargo, debemos recordar que en la sección anterior se concluyó que la mejor especificación es la espacial MV (SAC) por lo que debemos realizar nuevas pruebas de diagnóstico y determinar los valores de ajuste para el modelo espacial sin los valores atípicos en los residuos (Cuadro 4.12).

Cuadro 4.11. Medidas de ajuste del modelo MCO sin valores atípicos

R	R^2	R^2 Ajust.	Log-verosimilitud	AIC	D.E. Estimación
0,838	0,701	0,694	-3197,626	6441,252	9,84416

Las pruebas de diagnóstico del Cuadro 4.13 muestran los cambios más importantes: las pruebas de heteroscedasticidad en el modelo MCO sin valores atípicos nos indican que debemos rechazar el supuesto de homoscedasticidad; y, la prueba de normalidad del error en el modelo SAC sin residuos atípicos (espacial MV sin atípicos) nos permite acoger el supuesto de normalidad (Sig. 0,116).

Por otra parte, las pruebas LR, Wald y LM rechazan el supuesto de no autocorrelación espacial del error en el modelo MCO sin atípicos y la prueba LMSAR rechaza el supuesto de no autocorrelación espacial del error en el modelo espacial SAR sin atípicos.

Cuadro 4.12. Especificación espacial MV sin valores atípicos

MODELO MV	Coeficientes		t- asintótica	Sig.	Intervalo 95% de confianza para B	
	B	D.E.			L.I.	L.S.
C	49,720970	6,715686	7,403707	0,000000	36,55823	62,88371
S3	9,063767	1,955741	4,634440	0,000009	5,230515	12,89702
RIEGOS	0,755541	0,237496	3,181277	0,003076	0,290049	1,221033
T1	22,458608	1,47811	15,194142	0,000000	19,56151	25,3557
T2	14,728067	1,107814	13,294706	0,000000	12,55675	16,89938
T1*OTV	-6,523056	1,906608	-3,421288	0,002878	-10,26	-2,7861
T3*PR	7,601292	3,059206	2,484727	0,021516	1,605248	13,59734
T1*FARC	3,278325	1,435087	2,284409	0,020310	0,465554	6,091096
FCC	6,838927	2,953246	2,315732	0,019000	1,050565	12,62729
FARE*CC	8,835868	4,321316	2,044717	0,025686	0,366089	17,30565
FARE*PR	9,077831	2,490497	3,644988	0,002225	4,196457	13,95921
Edad cultivo	-1,973171	0,211159	-9,344462	0,000000	-2,38704	-1,5593
CC-8592	4,931704	1,621975	3,040555	0,010401	1,752633	8,110775
CR-74250	9,572814	1,121016	8,539407	0,000000	7,375623	11,77001
PR-671070	-6,382360	1,363585	-4,680572	0,000010	-9,05499	-3,70973
F	3,573488	0,995626	3,589188	0,000100	1,622061	5,524915
FARC	1,959744	1,003798	1,952329	0,051510	-0,0077	3,927188
K	2,762846	1,57086	1,758811	0,044431	-0,31604	5,841732
pH	-1,987404	0,73377	-2,708484	0,007117	-3,42559	-0,54921
CE	-0,026420	0,009517	-2,776051	0,010436	-0,04507	-0,00777
Edad cosecha	2,184333	0,474616	4,602315	0,000029	1,254086	3,11458
Ca	0,437749	0,113901	3,843252	0,002333	0,214503	0,660995
Mg	-0,457255	0,129998	-3,517408	0,002217	-0,71205	-0,20246
Rho	-0,096996	0,013339	-7,271361	0,000000	-0,12314	-0,07085
lambda	0,635000	0,03335	19,040756	0,000000	0,569634	0,700366

El análisis de varianza del Cuadro 4.14 indica la significancia global del modelo MCO sin residuos atípicos (Sig. 0,000).

Cuadro 4.13. Pruebas de diagnóstico para las especificaciones sin atípicos

Pruebas de diagnóstico	g.l.	valor	Sig.
Heteroscedasticidad White (términos no cruzados MCO)	30	55,97728	0,002747
Heteroscedasticidad White (términos cruzados MCO)	275	298,2366	0,000005
LR (error MCO)	1	149,2742	0,000000
Wald (error MCO)	1	2454,095	0,000000
LM (error MCO)	1	159,6861	0,000000
LMSAR(error SAR)	1	603,0439	0,000000
Kolmogorov-Smirnov Z (error est. SAC)		1,193	0,116000

Cuadro 4.14. Análisis de varianza del modelo MCO sin atípicos

Fuente de variación	Suma de cuadrados	g.l.	Cuadrados medios	F	Sig.
Regresión	191953,649	22	8725,166	90,036	,000
Residual	81693,060	843	96,908		
Total	273646,710	865			

Las medidas de ajuste relevantes del modelo espacial MV sin atípicos del Cuadro 4.15 muestran los siguientes cambios con respecto al modelo espacial MV: aumento en el valor del coeficiente de correlación lineal (R) de 0,861 a 0,878; aumento del valor de Log-verosimilitud de $-2404,6969$ a $-2326,6506$; y, disminución del valor AIC de 4859,3938 a 4703,3012.

Cuadro 4.15. Medidas de ajuste del modelo espacial MV sin atípicos

R	R ²	R ² Ajust.	Log-verosimilitud	AIC	D.E. Estimación
0,878	0,7710	0,7650	-2326,6506	4703,3012	8,5073

El análisis de los residuos del modelo MV sin atípicos muestra que su distribución empírica (Figura 4.5) se aproxima a una distribución normal. Además, se contrasta el supuesto de normalidad con la prueba de Kolmogorov-Smirnov del Cuadro 4.13. La estadística descriptiva de los residuos se muestra en el Cuadro 4.16.

Cuadro 4.16. Estadística descriptiva de los residuos del modelo MV sin atípicos

N		866
Media		,0000
Mediana		,4139
Desviación Estándar		8,51222
Varianza		72,4579
Rango		54,26
Mínimo		-28,72
Máximo		25,54
Percentiles	25	-5,0890
	50	,4139
	75	5,3086

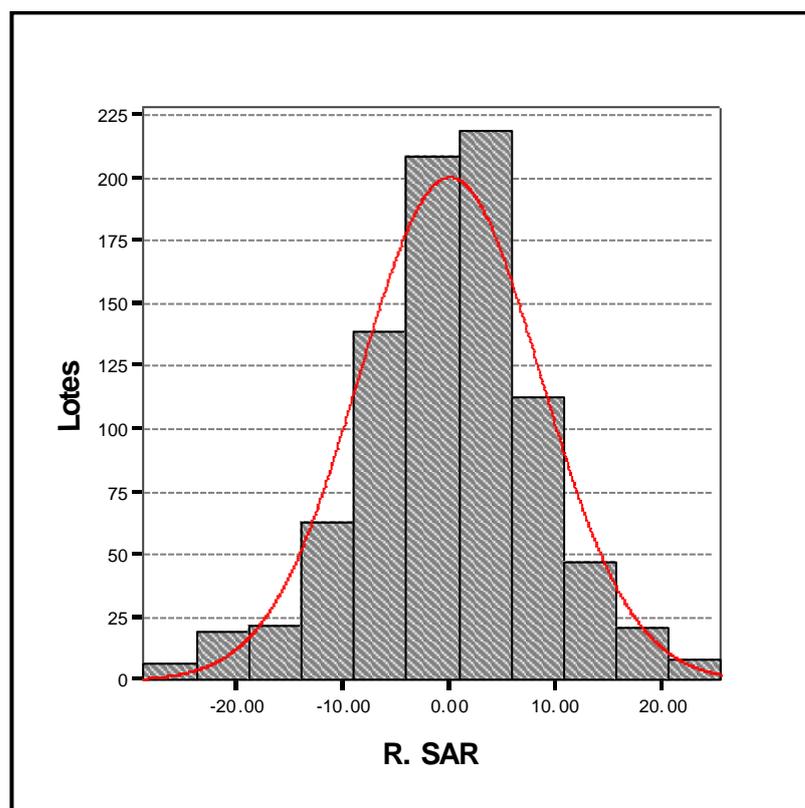


Figura 4.5. Histograma de los residuos del modelo MV sin atípicos

En conclusión, el modelo espacial MV sin residuos atípicos tiene el mejor ajuste entre las especificaciones formuladas. Se ratifican los supuestos de autocorrelación espacial en la variable dependiente y el error estocástico; sin embargo, es muy posible que los estimadores no sean asintóticamente eficientes debido a un problema de heteroscedasticidad.

4.6. INTERPRETACIÓN DE RESULTADOS

En esta sección se interpretan los coeficientes estimados desde una perspectiva agronómica. Entre los modelos evaluados se utilizará el modelo espacial MV o SAC debido a sus mejores propiedades estadísticas y a que las estimaciones de los parámetros no difieren en gran medida de las estimaciones del modelo espacial sin atípicos. Cada coeficiente representa el cambio esperado (promedio) en la producción por cambio unitario en el factor cuantitativo (cambios parciales o efecto neto) asumiendo todos los demás factores constantes. Para los factores cualitativos, cada coeficiente representa el cambio promedio en la producción cuando partimos de un nivel base a otro nivel alternativo permaneciendo todos los demás factores constantes.

4.6.1. Condiciones climáticas

Para el ciclo de cultivo 2002-2003, las precipitaciones fueron de 892 mm, y, considerando que el promedio de precipitaciones del ingenio es de aproximadamente 1600 mm al año, se puede considerar que el año 2003 fue un año seco. Es decir, solamente llovió el 56% de lo

que llueve en un año típico. Este es un factor que afecta directamente a la producción ya que en años como este (2003) las precipitaciones no son suficientes para cumplir con el requerimiento de agua del cultivo. La heliofanía (horas sol), fue otro parámetro cuyo valor fue inferior al promedio anual de 734 horas sol: el valor fue de 629 horas sol. Considerando que la heliofanía es baja en el ISC para los requerimientos de sol de la caña de azúcar, se puede calificar como negativo para el crecimiento del cultivo cualquier valor que esté por debajo del promedio.

Cuadro 4.17. Condiciones climáticas generales del ciclo 2002-2003

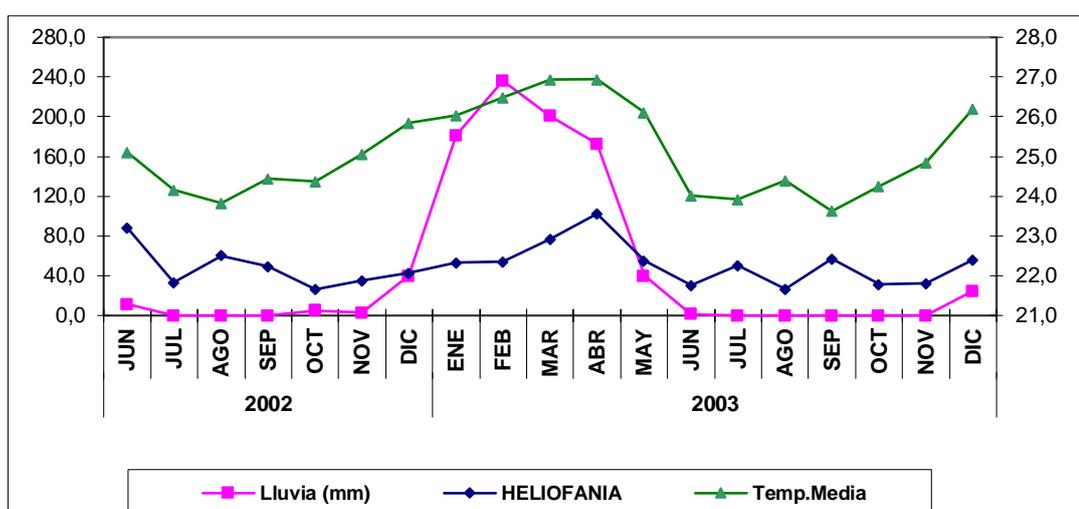
Factor	Valor
Total precipitaciones (mm)	891,79
Total horas sol	629,80
Temperatura Media (°C)	25,28
Temperatura máxima (°C)	28,98
Temperatura mínima (°C)	21,54
Prom. evaporación diaria (mm)	2,88
Prom. evaporación mensual (mm)	87,45

Elaboración: Xavier Vítóres, 2004

Las condiciones generales de temperatura, horas sol, precipitaciones, se muestran en el Cuadro 4.17. Las condiciones específicas (totales y promedio por mes) del clima para el ciclo de cultivo 2002-2003 se muestran en el Cuadro 4.18 y Figura 4.6.

Cuadro 4.18. Condiciones climáticas por mes del ciclo 2002-2003

AÑO	2002							2003											
	JUN	JUL	AGO	SEP	OCT	NOV	DIC	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Lluvia (mm)	11.7	0.0	0.0	0.0	5.3	3.0	39.6	181.1	236.2	200.9	173.0	39.9	1.5	0.0	0.0	0.0	0.0	0.3	24.6
HELIOFANIA	88.2	33.7	60.4	49.4	26.3	35.6	42.6	52.8	54.4	77.0	102.4	55.0	30.5	50.3	26.5	56.8	31.4	32.6	56.5
Temp.Media	25.1	24.1	23.8	24.4	24.4	25.1	25.8	26.0	26.5	26.9	26.9	26.1	24.0	23.9	24.4	23.6	24.3	24.8	26.2

**Figura 4.6.** Condiciones específicas del ciclo 2002-2003

4.6.2. Tercio de cultivo

Para interpretar los coeficientes del modelo, en primer lugar, se escoge a una de las tres variables que representan al tercio de cultivo como base para realizar comparaciones. Se decidió escoger T3 (tercer tercio) como base. El primer tercio (T1) tuvo un

coeficiente estimado de 22.4 TMC/ha., lo que indica que este tercio es en promedio 22.4 TMC/ha. superior al tercer tercio. El segundo tercio (T2) alcanzó 14.2 TMC/ha. sobre el tercero. Esto indica que en ambos tercios se obtiene una producción significativamente mayor al tercer tercio (T3) manteniendo los demás factores constantes.

La gran diferencia que existe en producción de caña entre las tres épocas de inicio del cultivo (T1, T2 y T3) se debe a la interacción existente entre las condiciones del clima y el estado fisiológico de la planta que comprende: germinación (un mes), macollamiento y cierre (dos meses), rápido crecimiento (siete meses) y maduración (dos meses). Para la caña cosechada en el primer tercio (julio-agosto), existen muy buenas condiciones para el desarrollo vegetativo: para la caña de julio (Figura 4.7), desde los 4 hasta los 11 meses de edad y para la caña de agosto desde los 3 a los 10 meses de edad.

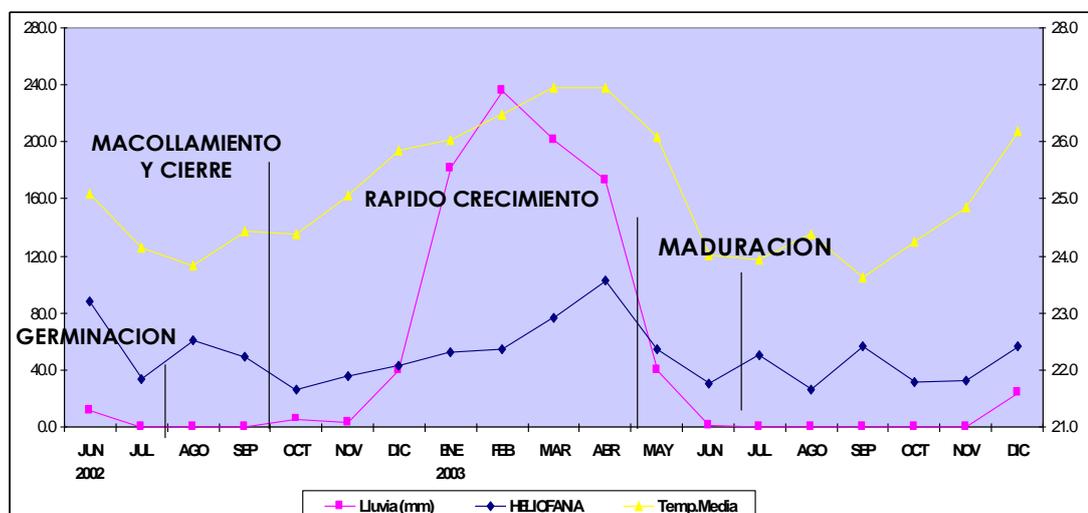


Figura 4.7. Etapas de crecimiento en el primer tercio (julio)

La caña de julio recibe buenas condiciones climáticas en 6 de los 7 meses correspondientes al periodo de rápido crecimiento: temperatura, precipitaciones y horas sol en aumento a partir el mes de noviembre. La caña de agosto también presenta buenas condiciones para 6 de los 7 meses. Este hecho hace que se obtenga una alta producción de biomasa en el primer tercio. Sin embargo, el rendimiento (lb.az./tmc) en esta época es bajo porque la caña no tiene tiempo para secarse y acumular sacarosa (maduración) durante los dos últimos meses antes de ser cosechada. Este efecto es más notorio en la caña de julio que recibe precipitaciones en mayo.

Para las cañas del segundo tercio (septiembre-octubre) las condiciones de buen crecimiento se presentan en 6 de los 7 meses correspondientes al periodo de máximo crecimiento para la caña de septiembre (condiciones ideales desde los 3 a los 9 meses de edad) y en 5 de los 7 meses para la caña de octubre (condiciones ideales desde los 2 meses a los 8 meses). De inmediato siguen condiciones adecuadas para la maduración: disminución de temperatura media y mayor oscilación entre las temperaturas máxima y mínima diarias. Por esta razón se espera una menor cantidad de biomasa en este periodo comparada a la del primer tercio y un mejor equilibrio entre producción de caña y concentración de sacarosa.

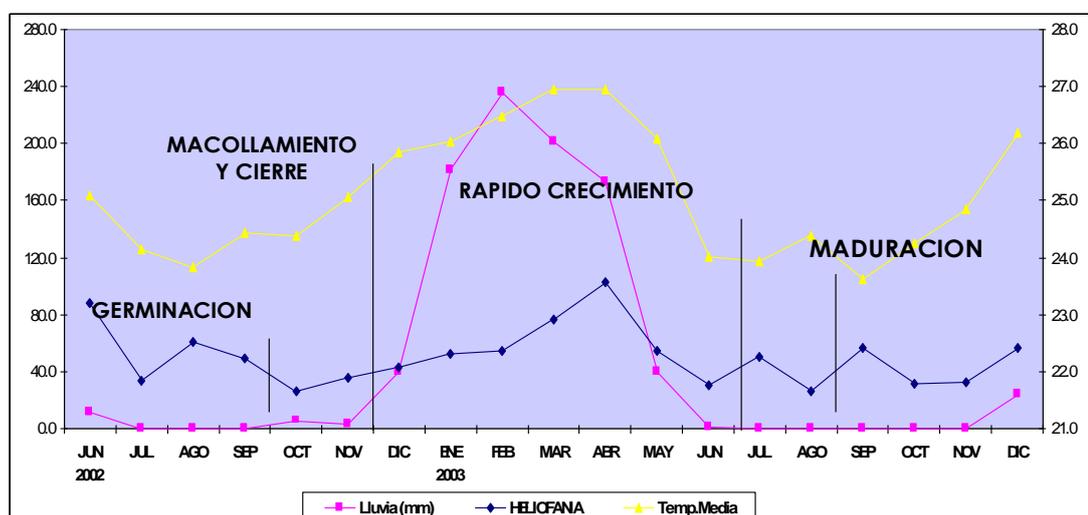


Figura 4.8. Etapas de crecimiento en el segundo tercio (septiembre)

Para las cañas del tercer tercio (noviembre-diciembre) las condiciones de buen crecimiento dependen, principalmente, de la magnitud del invierno. Cuando es pobre (precipitaciones < 800 mm) por lo general se tienen buenas condiciones climáticas desde el inicio del cultivo hasta los 6 a 7 meses de edad (Figura 4.9), pero no se tiene suficiente agua para riego. En el caso contrario, cuando el invierno es normal o fuerte (precipitaciones > 1500 mm) las condiciones son extremas al presentarse estrés hídrico en la planta por sobresaturación del suelo especialmente en la etapa de germinación. Por otra parte, además de estos dos posibles escenarios, en el momento en que termina el invierno (6 a 7 meses de edad de la caña) la temperatura media y las horas sol comienzan a descender, condiciones que no son favorables para el desarrollo vegetativo y por lo tanto, perjudican aún más al cultivo, pues baja su tasa de crecimiento a una edad en la que aún tiene potencial para seguir creciendo: la caña de noviembre recibe buenas condiciones en 4 de los 7 meses de rápido crecimiento y la caña de diciembre en 3 de los 7 meses. Estas se constituyen en las razones principales por las que en este tercio (T3) se alcanzan los promedios de productividad de caña y azúcar más bajos.

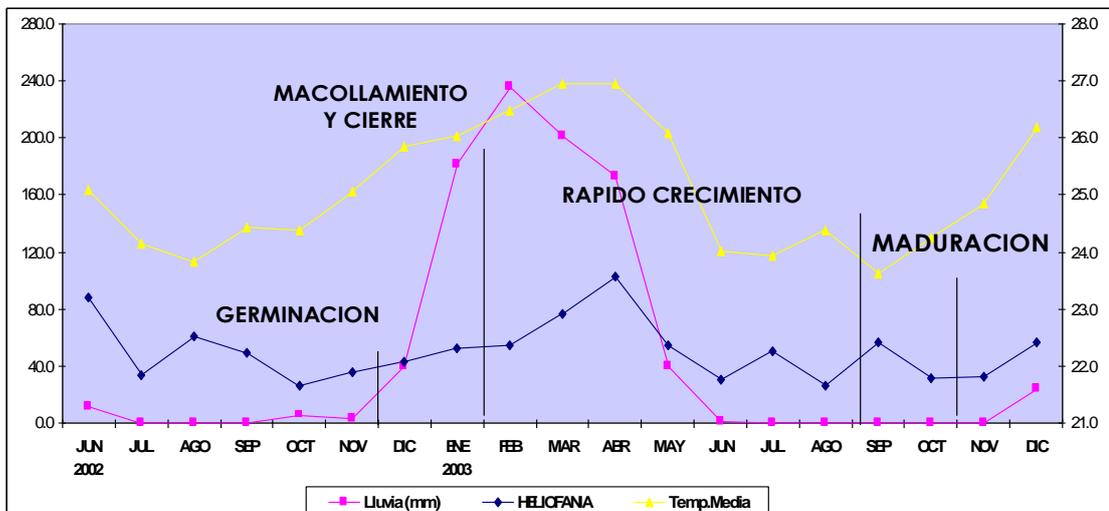


Figura 4.9. Etapas de crecimiento en el tercer tercio (noviembre)

4.6.3. Número de riegos

La variable número de riegos se incluyó en el modelo con el fin de determinar de cierta manera la influencia del suministro de agua al cultivo. Esta variable no considera todas las formas de suministro de agua, la cantidad exacta de aplicación, ni el momento en que se lo hace, por lo tanto no se la debe considerar un indicador absoluto de la efectividad del suministro de agua. Sin embargo, se la considera útil para mostrar si existe alguna tendencia en la producción con respecto al número de riegos aplicados.

En el modelo, el coeficiente para la variable número de riegos fue positivo con un valor de 0.75 TMC/ha.. Este valor indica una tendencia a obtener una mayor producción de caña cuando se aplican más riegos. Esta situación podría estar indicando que están faltando riegos durante el periodo de rápido crecimiento de la caña (3 – 10 meses) que es cuando la planta necesita agua para producir biomasa.

El intervalo de número de riegos aplicados en el ciclo de cultivo está entre 2 y 16; por lo que, en este rango, el potencial de productividad de la caña disminuye en promedio 0.75 TMC/ha. por cada riego que se deja de aplicar, manteniendo todos los demás factores constantes. Los datos de un estudio realizado por el Glenn Shura¹, sobre la influencia del riego y la precipitación en la productividad de caña y el rendimiento (lb.az./tmc) indican que en la zafra 2003 solamente se cumplió con el 82% del requerimiento de agua, mientras que en los tres años anteriores se cumplió con más del 90% del requerimiento y se obtuvieron productividades de caña (78-84 TMC/ha.) más altas. En el ciclo 2002-2003, además de que no se dieron los riegos suficientes durante el verano, las precipitaciones fueron insuficientes (892 mm), esta situación causó que el nivel de

¹ Trabajo del Dep. de Campo, Ing. Glenn Shura, Asesor Agrícola del ISC, 2004

cumplimiento del requerimiento de agua sea bajo (82%) y muy probablemente este hecho se constituye en una de las principales causas de la baja producción de caña se obtuvo (77 TMC/ha.).

4.6.4. Variedades de caña

Las variedades sembradas en el ISC son: Ragnar, CR-74250, CC-8592, PR-671070, NC0-310, BJ-7046 y B-74132. A los lotes con menos del 75% de una variedad específica se les asigna la variedad ficticia MISCELÁNEA, que puede interpretarse como la agrupación de dos o más variedades dentro del lote. A las variedades predominantes se les asignó una variable cualitativa: CR-74250, CC-8592 y PR-671070 tomando como base la Ragnar; mientras que, a los lotes que tuviesen cualquiera de las variedades restantes se les asignó la variable cualitativa OTV que quiere decir “otras variedades”.

Las variedades que obtuvieron una producción promedio de caña significativamente mayor a la alcanzada con la variedad Ragnar, PR-671070 y OTV (manteniendo los demás factores constantes) fueron la CC-8592 y la CR-74250, con 4.5 y 9.22 TMC/ha.,

respectivamente. La variedad PR-671070 presentó un valor significativamente menor a la Ragnar, con -6.3 TMC/ha.. Estos resultados hacen pensar que es una decisión acertada continuar con la propagación de la variedad CC-8592 y con la eliminación de la variedad PR-671070. Sin embargo, más adelante se analiza la interacción variedad-textura de suelo, donde se puede observar que la mayoría del área sembrada con PR-671070 no se encuentra en el tipo de suelo en el que mejor se adapta, lo cual hace que promediando el valor de la producción para todos los lotes se obtenga una baja producción con esta variedad.

Es de gran importancia definir cuáles son las variedades de mejor desempeño bajo diferentes condiciones de suelo, época de inicio (tercio), entre otros factores. Los programas de fitomejoramiento (nuevas variedades) concentran sus criterios de selección principalmente en parámetros de producción dejando para los estados posteriores los temas como adaptación a tipos de suelo y a épocas de cosecha. Las interacciones que trataremos posteriormente serán: textura-variedad, tercio-variedad y textura-tercio. Los modelos se constituyen en una herramienta útil para conocer el tipo de suelo y la época de siembra a la que mejor se adapta cada variedad.

4.6.5. Textura del suelo

Las texturas de suelo predominantes en el ISC son la arcillosa, franco, franco-arcillosa y franco arenosa. Se representó a cada una de estas texturas por medio de una variable cualitativa. Las texturas restantes como la arcillo-limosa, areno-limosa, arenosa, franco-arcillosa-arenosa, franco-arcillosa-limosa y franco-limosa se agruparon en una sola variable denominada OTX, que puede interpretarse como "otras texturas".

Para este factor se utilizó la textura arcillosa como variable base para realizar comparaciones. En el modelo la textura franco (F) resultó en promedio 3,46 TMC/ha. superior a la arcillosa, franco-arenosa (FARE) y OTX; y, la textura franco-arcillosa (FARC) superior en 1,9 TMC/ha. a la arcillosa, franco-arenosa y OTX. Estos resultados confirman que, en promedio, las producciones más altas se obtienen en suelos de texturas medias, específicamente, la textura franco y franco-arcillosa.

4.6.6. Sectores

El ISC está dividido administrativamente en seis sectores. A cada sector se le asignó una variable cualitativa S1, S2, S3, S4, S5 y S6. Se tomó al sector seis como base para realizar comparaciones.

En el modelo el sector 3 resultó significativamente superior que el resto de sectores con 9.1 TMC/ha.. En este sector se ha estado regando desde hace aproximadamente veinte años (Figura 4.10) con el agua residual de la fábrica que contiene cachaza y la vinaza de la destilería de alcohol; estos subproductos son ricos en nutrientes como fósforo (P) y potasio (K) y mejoran las características químicas del suelo. Cuando se inició el riego con esta agua, el sector 3 presentaba los valores de productividad más bajos, sin embargo, después de algunos años de riego la producción comenzó a incrementarse cada año hasta convertirse en el sector de mayor producción. Los cuadros sombreados en amarillo de la Figura 4.10 corresponden a los sectores con la mejor producción por año desde el año 1985 al 2003.

Productividad por sector en toneladas de caña por hectárea (TCH) de 1985 al 2003																			
SECTOR	Año																		
	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
1	66	79	83	67	85	79	79	90	75	55	64	66	118	110	64	84	77	83	78
2	59	66	76	68	73	74	85	96	76	58	65	62	139	111	83	81	80	80	76
3	58	61	77	60	81	88	82	106	87	65	72	79	146	134	91	92	85	89	86
4	70	73	83	62	76	67	80	92	73	62	69	72	103	114	69	70	73	80	69
5	71	66	88	84	88	75	86	98	72	63	76	66	125	101	67	82	74	88	78
6							73	95	87	55	66	67	127	120	87	85	82	81	75

Figura 4.10. Mejores producciones por sectores desde 1985 al 2003

4.6.7. Edad del cultivo

En el modelo el coeficiente para la variable edad de cultivo puede interpretarse como el cambio promedio en la producción por cada año del cultivo; es decir, por cada año transcurrido la producción decrece en promedio 2 TMC/ha. (debido al signo negativo).

Las principales causas por las cuales existe una disminución de la producción cada año son: la despoblación, es decir, el daño de las cepas causado por los equipos que entran durante la cosecha y despacho de la caña; y, el incremento de organismos patógenos en el suelo por efecto del monocultivo constante. Errores en labores como las aplicaciones de herbicidas no selectivos, aplicaciones de

madurantes y pueden también causar despoblación y por lo tanto reducir la productividad del año siguiente.

4.6.8. Edad a la cosecha

En el modelo el coeficiente para la variable edad a la cosecha puede interpretarse como el cambio promedio en TMC/ha. por cada mes adicional en la edad del lote a la cosecha; es decir, por cada mes que demore la cosecha (caña rezagada) la producción aumenta en promedio 2.11 TMC/ha.. Este resultado debe interpretarse con precaución debido al intervalo de valores que toma la muestra. El mínimo mes de cosecha es de 10.13 meses y el máximo 16.11 meses, por lo que el aumento en la producción debe interpretárselo dentro de este rango.

4.6.9. Parámetros del análisis de suelos

Los parámetros del Cuadro 4.19 que se miden en los análisis de suelos y que resultaron significativos en el modelo fueron: potasio

(K), potencial hidrógeno (pH), conductividad eléctrica (CE), calcio (Ca) y magnesio (Mg).

Cuadro 4.19. Estadística descriptiva de los parámetros del análisis de suelos

	pH	CE (mmho/cm)	MO (%)	P (ppm)	K (cmol/L)	Ca (cmol/L)	Mg (cmol/L)	
N	871	871	871	871	871	871	871	
Media	6.2721	72.0103	1.8236	15.5029	.4015	16.8093	8.2450	
Mediana	6.2600	63.0000	1.8200	12.0000	.3200	16.1600	7.0000	
Desviación Estándar	.42354	36.40454	.45456	12.71228	.25132	5.20702	4.45470	
Mínimo	4.77	25.00	.35	1.00	.06	6.76	1.08	
Máximo	8.17	334.00	3.55	108.00	1.54	32.16	27.84	
Percentiles	25	5.9800	50.0000	1.5200	7.0000	.2300	12.6000	4.9600
	50	6.2600	63.0000	1.8200	12.0000	.3200	16.1600	7.0000
	75	6.5300	83.0000	2.1400	19.0000	.4900	20.5200	10.6400

La CE y el pH tuvieron coeficientes negativos en el modelo. Esto indica una disminución promedio de 0.03 TMC/ha. y 1.9 TMC/ha. por unidad de CE y pH, respectivamente. Los valores altos de estos parámetros están asociados con problemas de salinidad en el suelo y de restricción en las disponibilidades de fósforo y micronutrientes como el Zn, Cu, Mn, B, Fe y Cl, lo que quiere decir que posiblemente existe este tipo de problemas en los lotes donde se obtuvieron los valores más altos para los parámetros.

De existir un exceso de sales en los suelos del ingenio, esto podría ser el efecto de las bajas precipitaciones del invierno ya que no se lavan las sales de los suelos como posiblemente sucede en los inviernos con precipitación normal. Este resultado habría que confirmarlo aplicando el modelo para diferentes años y así determinar si el problema de salinidad se repite o es propio de años de baja precipitación.

Cuadro 4.20. Porcentaje de lotes por niveles en los parámetros del análisis de suelos

Parámetros	Niveles		
	Bajo	medio	Alto
pH	2.76	70.84	26.41
MO	66.48	33.52	-
P	44.66	32.38	22.96
K	16.19	66.36	17.45
Ca	-	20.67	79,33
Mg	-	25.6	74,4
Relación	Desbalance	Balance	Desbalance
Ca/Mg	38	60.73	1.26
Mg/K	1.15	32.15	66.7
Ca/K	-	18.83	81.17

Elaboración: Xavier Vítores, 2004.

Los elementos K y Ca tuvieron en el modelo los coeficientes positivos 3.38 y 0.36, respectivamente; lo que significa que en promedio la producción de caña se incrementa en 3.38 TMC/ha. cuando el nivel de K en el suelo aumenta en una unidad (cmol/L). Esto es evidencia de que existe potencial para incrementar la

producción con un aumento en la dosis de K en la fertilización; sin embargo, se puede observar en el Cuadro 4.20 que únicamente el 16,19% de los lotes de la muestra presentan niveles bajos de potasio considerando los niveles establecidos en el Cuadro 2.5 del Capítulo 2.

Para el Ca el coeficiente indica que se obtiene un incremento promedio de 0.38 TMC/ha. por cada cmol/L de incremento del elemento en el suelo a pesar de que el 79% de los lotes analizados presentan contenidos altos de Ca (más de 12 cmol/L). El coeficiente de Mg resultó negativo lo que indica que por cada cmol/L de Mg en el suelo la producción de caña disminuye en promedio 0.43 TMC/ha.

Las condiciones para el potasio parecen no ser desfavorables para la planta debido a que no existe un déficit del elemento en el suelo en la mayoría de los lotes (83.81% de los lotes de la muestra), sin embargo, debemos recordar que tanto el K como el Ca y el Mg (bases intercambiables) deben estar balanceados en el suelo ya que la planta no diferencia entre ellos; es decir, debe existir un equilibrio en el suelo de los elementos para no afectar su disponibilidad al cultivo y, consecuentemente, no ser un factor limitante de la producción. Por esto, los balances nutricionales en el suelo son muy

importantes y no debe considerarse únicamente el nivel del elemento en el suelo con respecto a valores críticos preestablecidos. Las relaciones a analizar son Ca/Mg, Mg/K y Ca/K.

El análisis de la relación Ca/Mg indica que el 61% de los lotes analizados presentó un balance en esta relación, mientras que un 38% de los lotes estuvo en desbalance por exceso de Mg y un 1% por exceso de Ca. Estos resultados indican que es necesario incrementar el contenido de Ca en el suelo o disminuir el de Mg para balancear la relación. El modelo sustenta esta hipótesis mostrando un coeficiente positivo para el Ca y negativo para el Mg. En la relación Mg/K el 32,15% de los lotes están en balance mientras que el 66,7% están en desbalance por exceso de Mg y el 1.15% están en desbalance con respecto al K; estos resultados indican que para tener una relación Mg/K adecuada es necesario aumentar el nivel de K en el suelo. El modelo sustenta la observación presentando coeficientes positivos para el K y negativos para el Mg.

Con respecto a la relación Ca/K, el 18.83% de los lotes presentaron un balance en la relación, mientras que el 81% estuvo en desbalance por exceso de Ca. En ningún lote se presentó desbalance por exceso de K. Para balancear esta relación es

necesario aumentar el nivel de K en el suelo. Este resultado también sustenta el hecho de que el coeficiente del K sea positivo en el modelo y con mayor magnitud que el de Ca.

4.6.10. Interacciones

La interacción entre factores ocurre cuando su actuación no es independiente; es decir, ocurre cuando la respuesta del cultivo a un factor determinado es diferente para distintos niveles de otro factor. En esta investigación los factores para los que podría esperar interacción son: 1) textura de suelo y variedad; 2) tercio de cultivo y variedad; y, 3) tercio de cultivo y textura de suelos.

4.6.10.1. Textura del suelo y variedad

En el modelo econométrico las interacciones entre textura de suelos y variedad que resultaron significativas fueron: textura franco arenosa y la variedad CC-8592 (FARE*CC), textura franco arenosa y variedad PR-671070 (FARE*PR) y textura franco y variedad CC-8592 (F*CC). Estos resultados indican que en los suelos franco

arenosos, la variedad CC-8592 alcanza una producción promedio de 10.25 TMC/ha. sobre las variedades Ragnar y CR-74250; y, la variedad PR-671070 alcanza una producción promedio de 8.22 TMC/ha. sobre las variedades Ragnar y CR-74250. Este resultado confirma la adaptación de estas dos variedades a suelos livianos con buen drenaje. En suelos francos, la variedad CC-8592 alcanzó una producción promedio de 7.4 TMC/ha. sobre las variedades Ragnar, CR-74250 y PR-671070.

Los resultados de las interacciones son similares a aquellos que se esperan obtener en un diseño experimental con los factores variedad y textura de suelos. Es importante mencionar en este punto que comparando el efecto sin interacción, la variedad PR-671070 es la de menor producción (coeficiente negativo); sin embargo, incorporando el factor texturas se obtienen resultados diferenciados. Por esto, es muy probable que una de las posibles causas de la baja producción promedio de la variedad PR-671070 al considerar todos los lotes sin diferenciar por texturas es que ha sido sembrada en su mayoría en lotes donde su textura no le favorece. Esto puede evidenciarse sin dificultades en el Cuadro 4.21 donde se observa que la variedad se encuentra ubicada en su mayoría en lotes con suelos arcillosos (28 lotes) y franco-arcillosos (18 lotes).

Cuadro 4.21. Distribución de lotes por textura de suelos y variedad

TEXTURAS	VARIEDAD				
	RAGNAR	OTRAS	PR-671070	CC-8592	CR-74250
	Lotes	Lotes	Lotes	Lotes	Lotes
FRANCO	92	19	17	14	13
ARCILLOSO	66	7	28	15	82
F. ARENOSO	77	31	15	4	11
F. ARCILLOSO	140	30	18	25	59
OTRAS	69	14	8	6	11

Por otra parte, la variedad CC-8592 esta sembrada en su mayoría en texturas arcillosas y franco-arcillosas; y, en el modelo, su efecto global es positivo; sin embargo, debe investigarse su potencial en suelos franco-arenosos pues parece diferenciarse de las demás variedades en esta textura aunque sólo se cuenta con 4 observaciones para la interacción.

La variedad CR-74250 no se diferencia en ninguna interacción, sin embargo es importante notar que esta variedad se encuentra sembrada en su mayoría en suelos de texturas arcillosas y franco-arcillosas.

4.6.10.2. Tercio de cultivo y variedad

De las interacciones entre el tercio de cultivo y las variedades, en el modelo resultaron significativas: primer tercio y otras variedades (T1*OTV) y tercer tercio y variedad PR-671070 (T3*PR). El coeficiente para la variable T1*OTV no tiene clara interpretación debido a que es el conjunto de todas las variedades que se encuentran distribuidas en menor proporción en el ingenio en el primer tercio. La interacción entre el tercer tercio y la variedad PR-671070 resultó significativa con un coeficiente de 7.53 TMC/ha., lo que implica que en el tercer tercio la variedad PR-671070 obtuvo una producción promedio de 7.5 TMC/ha. sobre las alcanzadas por las demás variedades. Este resultado ocurre posiblemente debido a la mejor adaptación de la variedad PR-671070 a un tercer tercio bajo condiciones de bajas precipitaciones, tal como las del ciclo 2003.

4.6.10.3. Tercio de cultivo y textura de suelo

De las interacciones entre los tercios de cultivo y las texturas, en el modelo resultó significativa la interacción entre el primer tercio y la textura franco arcillosa; es decir, para el primer tercio en la textura

franco arcillosa se obtuvo una producción promedio de 3.59 TMC/ha. sobre las demás texturas. Este resultado puede ser consecuencia de las condiciones de bajas precipitaciones en el ciclo 2002-2003. La textura franco arcillosa favorece en la retención de humedad para los lotes del primer tercio.

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

Por lo general, las investigaciones de campo están dirigidas hacia un factor específico, llegando muchas veces a la conclusión de que otros factores que no fueron analizados debieron tomarse en cuenta. El tipo de análisis realizado en esta investigación toma en cuenta desde un principio a la mayoría de los factores involucrados en el sistema productivo tratando de que no se presenten omisiones graves ni pérdida de tiempo al momento de obtener los resultados y conclusiones que al final se traducen en recomendaciones de manejo. Consecuentemente, se llega a una mejor comprensión del comportamiento del cultivo en relación con sus factores determinantes de producción, se pueden comprobar hipótesis agronómicas y conocer cuáles son las interacciones entre los factores que favorecen una mejor productividad.

A continuación se presentan las principales conclusiones y recomendaciones de esta investigación, las cuales, se exponen desde dos diferentes contextos: 1) el análisis estadístico de la información; y, 2) el manejo sitio-específico del cultivo.

5.1. Conclusiones

5.1.1. Análisis de información

1. El método de regresión tradicional de Mínimos Cuadrados Ordinarios (MCO) proporciona coeficientes estimados inconsistentes, sesgados e ineficientes para el modelo econométrico de la producción de caña de azúcar. Esto se debe, principalmente, a que no podemos rechazar mediante pruebas de diagnóstico (Capítulo 4) la hipótesis sobre la no existencia de autocorrelación espacial de los errores en el modelo MCO y SAR; y, a la alta significancia estadística de los coeficientes para los rezagos espaciales de la variable dependiente y el término del error en el modelo SAC.

2. Las pruebas de heteroscedasticidad en el modelo MCO no son concluyentes en un principio, sin embargo, en el análisis de sensibilidad, las pruebas de White en el modelo MCO sin residuos atípicos exponen evidencia de la posible presencia de heteroscedasticidad.

3. Los valores de las desviaciones estándar del modelo no espacial MCO con y sin atípicos no corresponden a la correcta estructura de la matriz de varianzas y covarianzas de los estimadores MCO mostrada en la expresión 3.41 del Capítulo 3. La matriz de varianzas-covarianzas de $\hat{\beta}_{MCO}$ no es $\sigma^2(X'X)^{-1}$ sino $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$; la primera expresión es una estimación sesgada de la segunda, sin que exista necesariamente ninguna ordenación entre ambas en términos de sus tamaños. Es decir, la varianza de cada parámetro podría ser inferior o superior, si se utilizase la expresión (incorrecta) $\sigma^2(X'X)^{-1}$ para la matriz de varianzas covarianzas de $\hat{\beta}_{MCO}$ que si se utilizase la expresión correcta de dicha matriz de covarianzas.

4. El método de estimación de Máxima Verosimilitud (MV) proporciona coeficientes estimados consistentes y con una

distribución normal asintótica con media igual al vector de parámetros del modelo SAC. No puede comprobarse la eficiencia asintótica de los estimadores debido a la posible presencia de heteroscedasticidad en el término del error del modelo.

5. El Criterio de Información de Akaike o estadístico AIC es menor en el modelo SAC que en el modelo MCO. Esto indica que el grado de entropía del modelo postulado SAC con respecto al verdadero pero desconocido modelo es menor comparado al grado de entropía del modelo postulado MCO. Este resultado junto con las conclusiones 1, 2, 3 y 4 sobre las propiedades de los estimadores de los modelos espacial SAC (MV) y no espacial (MCO) indican que la mejor especificación para la respuesta sitio-específica del cultivo de caña de azúcar con respecto a sus factores de producción es el modelo propuesto SAC y que el procedimiento de estimación más adecuado es el de MV.

6. El aspecto clave a tener en cuenta para una estimación estadística robusta de un modelo econométrico de la respuesta sitio-específica de un cultivo es que el análisis

debe incorporar los efectos espaciales de los datos con referencia espacial, es decir, la autocorrelación y la heterogeneidad espacial.

7. Mediante los coeficientes estimados se pudieron contrastar algunos supuestos de la teoría agronómica relacionados a la influencia de los parámetros del análisis de suelos y las texturas de suelos en la respuesta del cultivo. Específicamente, se pudieron comprobar y cuantificar los efectos de las texturas de suelos y los desbalances de K, Mg y Ca en el suelo sobre la respuesta esperada del cultivo.
8. Mediante los coeficientes estimados se pudieron contrastar supuestos que se adoptaron mediante la experiencia en campo. Específicamente, se pudieron cuantificar en términos de producción esperada los efectos de las variedades, sectores administrativos, tercio de cultivo y edad del cultivo. Este tipo de factores son inherentes al sitio-específico de la plantación, es decir, corresponden a características que no se encuentran en otros lugares donde se cultiva caña de azúcar. Las conclusiones con respecto a

estos factores están vinculadas al concepto de que la AP es también sitio-específica.

9. Mediante los coeficientes estimados se pudo investigar la respuesta del cultivo a diferentes factores sobre los cuales se tiene mucho interés económico pero se necesita aún más análisis. Específicamente se pudieron cuantificar en términos de producción esperada los efectos de la interacción entre variedades y texturas de suelos, interacción entre variedades y tercios de cultivo, riego y edad a la cosecha.

10. Todas las conclusiones específicas mencionadas con anterioridad y los resultados de esta investigación presentados en el Capítulo 4 nos llevan hacia la conclusión más importante vinculada con el objetivo general de esta tesis: Resulta factible disminuir la brecha existente entre el análisis de información georeferenciada y las recomendaciones sitio-específicas del cultivo de caña de azúcar mediante técnicas estadísticas robustas que incorporen los efectos espaciales de los datos de AP.

Manejo sitio-específico del cultivo

- 1. Tercio de cultivo:** Los resultados para la variable tercio de cultivo son: el primer tercio (T1) tuvo un coeficiente estimado de 22.4 TMC/ha., lo que indica que este tercio es en promedio 22.4 TMC/ha. superior al tercer tercio. El segundo tercio (T2) alcanzó 14.2 TMC/ha. sobre el tercero. Esto indica que en ambos tercios se obtiene una producción significativamente mayor al tercer tercio (T3) manteniendo los demás factores constantes.
- 2. Número de riegos:** El resultado obtenido con la variable número de riegos, indica que por cada riego que se dejó de aplicar, la productividad potencial de la caña se redujo en promedio 0.75 TMC/ha.. Esto es un indicador de que se pudo obtener una productividad de caña más alta suministrando un mayor número de riegos al cultivo. Por otro lado, mediante el cálculo de la evapotranspiración, sabemos que se ha cumplido solamente con 82% del requerimiento de agua y que en los tres años anteriores, cuando se logró cumplir con el 90-100%, se obtuvieron productividades más altas.

- 3. Variedad de caña:** Las variedades más productivas que la variedad predominante Ragnar son la CR-74250 y la CC-8592. Sin embargo, teniendo en cuenta la interacción de las variedades con las texturas de suelos, la variedad PR-671070 en la textura franco arenosa se diferencia de la RAGNAR y CR-74250 con una mayor producción.
- 4. Sectores:** Existe una diferencia significativa en producción del sector 3 con respecto a los demás sectores que demuestra el beneficio que se puede obtener sobre la productividad al aplicar materiales orgánicos al suelo como la cachaza y vinaza, confirmándose así la información que se encuentra en la literatura sobre experiencias similares con aplicación de materiales orgánicos.
- 5. Edad de cultivo:** El coeficiente estimado para la edad del cultivo indica que por cada año de edad que tenga el cultivo, en promedio la producción disminuye 2 TMC/ha.
- 6. Edad a la cosecha:** El resultado obtenido para esta variable sugiere que para efectos de obtener mayor productividad es conveniente incrementar la edad de cosecha de los lotes.

Por cada mes que se retrase la cosecha la producción aumenta en promedio 2.11 TMC/ha..

7. Análisis de suelos: Los resultados indican un serio desbalance en las relaciones Mg/K y Ca/K por exceso de Mg y Ca por lo que la disponibilidad del potasio puede verse muy afectada. Los coeficientes del modelo confirman la situación. La importancia del modelo en la interpretación del análisis de suelos es que permite vincular los parámetros analizados con la productividad demostrándose así las implicaciones agronómicas de los desbalances en el cultivo.

5.2. Recomendaciones

5.2.1. Análisis de información

1. Los datos sobre la producción y sus factores serán más confiables si provienen de un diseño experimental. De esta manera podremos evaluar de manera más eficiente las diferentes tecnologías y metodologías de la AP, particularmente, en lo relacionado al manejo diferencial de insumos y prácticas agrícolas.
2. La recolección de datos debe efectuarse de la manera más detallada posible y debe hacerse conciencia de los potenciales beneficios que una buena organización de la información junto con las metodologías de análisis adecuadas tienen sobre la productividad.
3. Se recomienda realizar nuevos modelos con la inclusión de variables adicionales, especialmente, aquellas vinculadas directamente con la variabilidad en la producción inducida por manejo agrícola: aplicación de insumos y prácticas agrícolas. Dentro de la aplicación de insumos se pueden

incluir variables relacionadas a las dosis de fertilizantes, herbicidas, madurantes, entre otros. Dentro de la evaluación de prácticas de manejo agrícola se pueden incluir variables relacionadas a la utilización de maquinaria agrícola y riego. Estas variables representan importantes costos en los que debe incurrir el productor por lo que resulta importante optimizar sus valores dentro del sistema productivo.

4. Es muy probable que la respuesta sitio-específica del cultivo varíe de año en año dependiendo del clima, por lo que se recomienda realizar el análisis para cada ciclo del cultivo con la finalidad de examinar el comportamiento de los coeficientes estimados dentro del modelo a través del tiempo.
5. Aunque la mayoría de las veces resulta una tarea muy compleja y tediosa, es recomendable investigar sobre las posibles fuentes de heteroscedasticidad del término del error en el modelo econométrico. Esto nos permitirá obtener estimadores asintóticamente eficientes con los que podamos obtener resultados de mayor exactitud. La

heteroscedasticidad puede modelarse como una función del término del error en relación con las variables explicativas o alguna otra variable que se considere determinante.

5.2.2. Manejo sitio-específico del cultivo

1. **Tercio de cultivo:** Es recomendable que se organicen las actividades de cosecha (logística de cosecha) de tal forma que la mayor área de inicio de cultivo corresponda a los tercios primero y segundo, periodos en los que existe gran potencial para obtener mayor biomasa. Para el tercer tercio es recomendable tener una mayor eficiencia en las actividades correspondientes al manejo del drenaje y riego del cultivo.
2. **Número de riegos:** Es recomendable que la programación de riegos calculada mediante el balance hídrico es la mejor herramienta de la que se dispone para regar de forma tal que se cumpla con la necesidad de agua del cultivo y por lo tanto evitar una disminución del potencial de crecimiento de la caña.

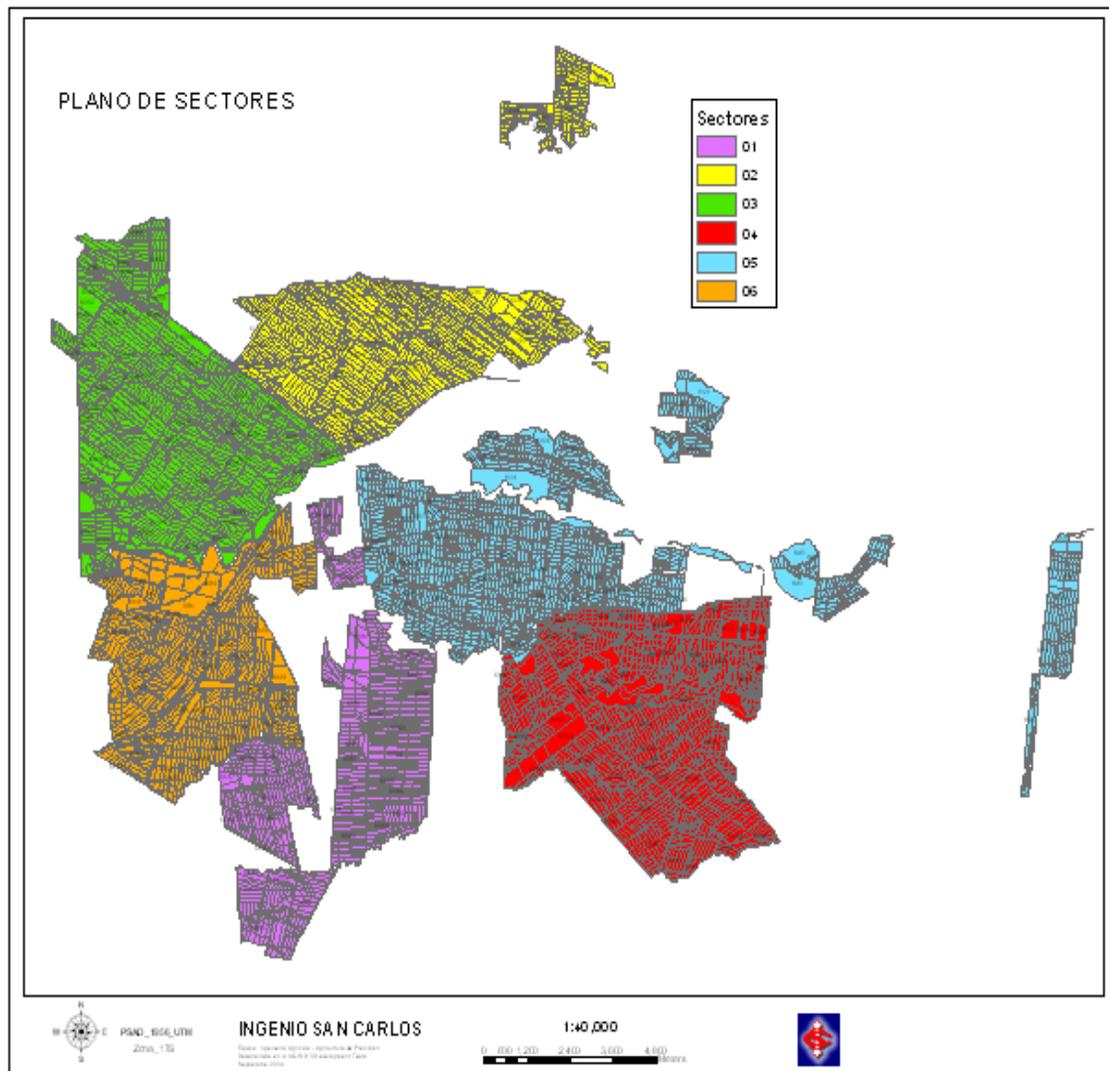
- 3. Variedades de caña:** Se recomienda continuar con la siembra de las variedades CC-8592 y CR-74250; y, sembrar la variedad PR-671070 en texturas franco arenosas. Por otra parte debe investigarse más la respuesta de la variedad PR-671070 en el tercer tercio.
- 4. Sectores:** Se recomienda la aplicación de materiales orgánicos a la mayor cantidad de área posible.
- 5. Edad del cultivo:** La tendencia a disminuir la producción a medida que transcurren los años de cultivo continuo de caña de azúcar conduce a la recomendación de evaluar y supervisar de mejor manera las prácticas relacionadas con algún tipo de mecanización en campo que inciden sobre la magnitud de la despoblación del cultivo. Mejores prácticas mecanizadas pueden conducir a disminuir la tasa de decrecimiento de la producción.
- 6. Edad a la cosecha:** Es recomendable evaluar la respuesta de esta variable por varios años para confirmar si el resultado obtenido se repite o si fue algo propio del ciclo de

cultivo 2002-2003 debido a una posible interacción entre la edad de cosecha y las condiciones climáticas, específicamente, las precipitaciones.

7. **Análisis de suelos:** Para corregir los desbalances se recomienda incrementar el contenido de K en el suelo. Además, debe tenerse en cuenta que el K es el nutriente extraído en mayor cantidad por la caña de azúcar. Estos argumentos respaldan la hipótesis de que se puede obtener una mayor productividad aplicando dosis de K más altas que las actuales en la mayoría de lotes. A pesar de que el contenido de potasio se lo puede considerar alto en muchos lotes (80%), es necesario aumentar el contenido de K en el suelo para balancear las relaciones Mg/K, Ca/K. Este hecho respalda las recomendaciones de la fertilización calculada del programa de AP en el ISC.

ANEXO 1

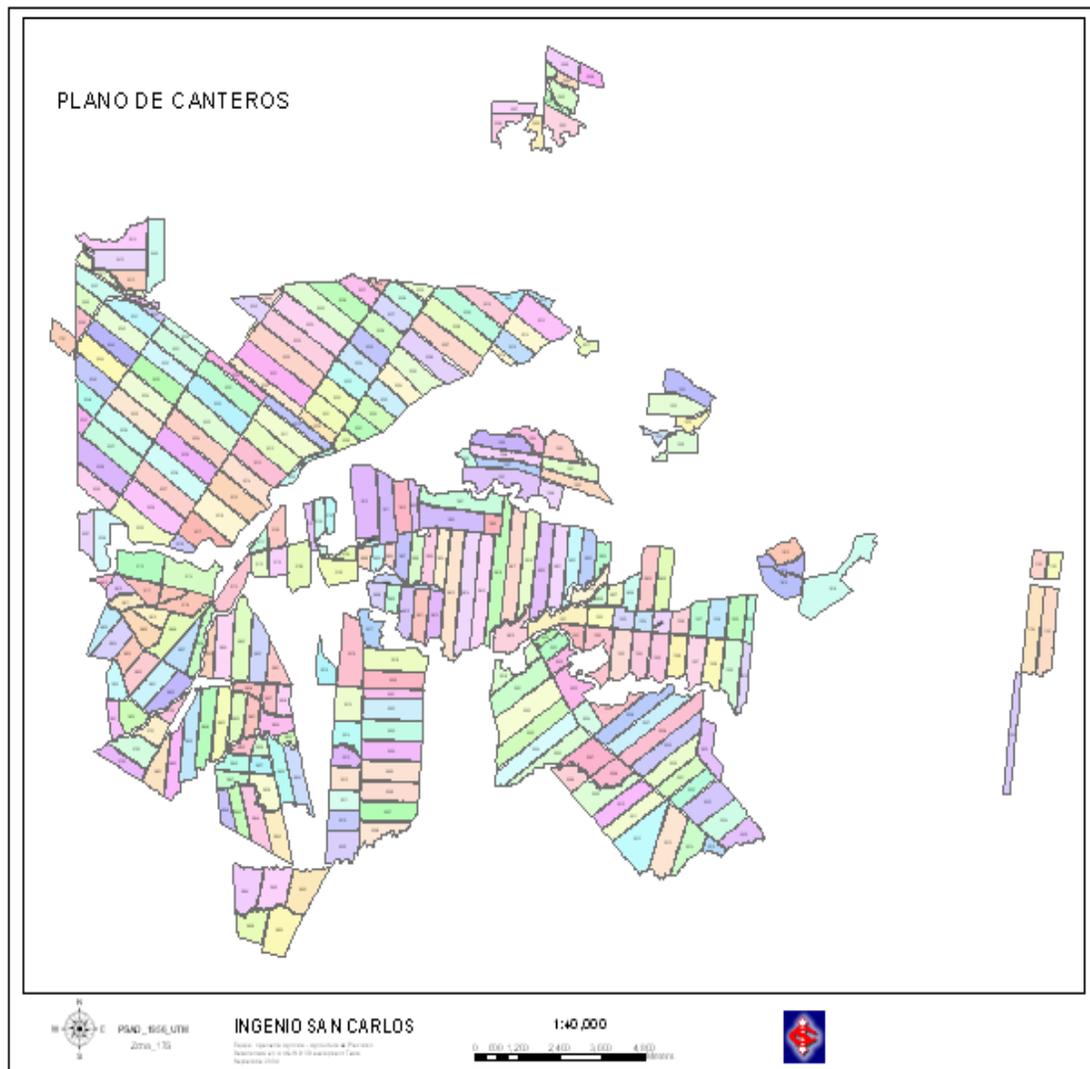
DIVISIÓN POR SECTORES DEL INGENIO SAN CARLOS



Fuente: Dep. Campo, ISC

ANEXO 2

DIVISIÓN POR CANTEROS DEL INGENIO SAN CARLOS



Fuente: Dep. Campo, ISC

ANEXO 3

DIVISIÓN POR LOTES DEL INGENIO SAN CARLOS



Fuente: Dep. Campo, ISC

ANEXO 4

TEORÍA ASINTÓTICA

1. Convergencia en probabilidad

Definición 1. Se dice que la sucesión de variables aleatorias $X_1, X_2, \dots, X_t, \dots$ converge en probabilidad a la variable aleatoria X si:

$$\forall \varepsilon > 0, \lim_t P[|X_t - X| < \varepsilon] = 1$$

La convergencia en probabilidad de la sucesión anterior se indica mediante la notación $\text{plim } X_t = X$. Una sucesión de variables aleatorias interesantes es la formada por los estimadores $\hat{\theta}_t$ de un parámetro desconocido θ que van obteniéndose según aumenta el tamaño muestral. Así, $\hat{\theta}_t$ denota el estimador de θ obtenido con las primeras t observaciones muestrales, $\hat{\theta}_{t+1}$ denota el estimador obtenido con las primeras $t+1$ observaciones, y así sucesivamente. Si la sucesión $\hat{\theta}_t$ converge en probabilidad al verdadero

valor (desconocido) del parámetro θ , entonces se dice que el estimador $\hat{\theta}$ es **consistente**. Es decir:

Definición 2. El estimador $\hat{\theta}$ del parámetro θ es *consistente* si al obtener la sucesión $\{\hat{\theta}_t\}_{t=1}^{\infty}$ del modo descrito se tiene:

$$\forall \varepsilon > 0, \lim_t P[|\hat{\theta}_t - \theta| < \varepsilon] = 1$$

Cuando un estimador no es consistente, a la diferencia $\text{plim } E(\hat{\theta}_t) - \theta$ se le denomina *sesgo asintótico*.

Es preciso hacer notar que:

- I. La definición de convergencia en probabilidad se generaliza inmediatamente al caso de una sucesión de vectores aleatorios de dimensión k fija, sin más que sustituir el valor absoluto en la definición anterior por la norma euclídea del vector diferencia $\hat{\theta}_t - \theta$.

- II. La variable aleatoria límite X en la definición pudiera tener una distribución de probabilidad degenerada en un punto, es decir, ser una constante. De hecho, en aplicaciones econométricas le preocupa al investigador saber si el estimador que utiliza es consistente, lo que equivale a averiguar si su límite en probabilidad coincide con el vector constante de parámetros desconocidos.

Una condición suficiente para la convergencia en probabilidad de una sucesión de variables aleatorias a una constante viene dada por la:

Proposición 1. Si $E(X_t) = c$ para todo t y si $\lim_t \text{Var}(X_t) = \lim \sigma_t^2 = 0$, entonces $\text{plim } X_t = c$.

Proposición 2. Si $\lim_t E(X_t) = c$ para todo t y si $\lim_t \text{Var}(X_t) = \lim \sigma_t^2 = 0$, entonces $\text{plim } X_t = c$.

2. Convergencia en distribución

Definición 3. Sea $X_1, X_2, \dots, X_t, \dots$ una sucesión de variables aleatorias con funciones de distribución $F_1, F_2, \dots, F_t, \dots$. Supongamos $F_t(x) \rightarrow F(x)$ en todos los puntos de continuidad de la función $F(x)$ y que, en ellos, $F(x)$ es una función de distribución. Entonces se dice que la sucesión de variables aleatorias X_t converge en distribución a X , donde X es una variable aleatoria con función de distribución F , y se representa por $X_t \xrightarrow{d} X$.

La definición de convergencia en distribución se extiende, sin ninguna dificultad, al caso de vectores aleatorios.

3. Teorema Central del Límite. Sea $X_1, X_2, \dots, X_t, \dots$ una sucesión de vectores aleatorios de dimensión k , independientes entre sí e idénticamente distribuidos, con $E(X_i) = u$ y $Var(X_i) = \Sigma$. Entonces se tiene:

$$\sqrt{T} \left[\frac{1}{T} \sum_{i=1}^T X_i - u \right] \xrightarrow{d} N_k(0, \Sigma)$$

4. Teorema de Mann-Wald. Sea X una matriz $T \times k$ y u un vector de dimensión

T tales que:

i. $E(u) = 0, E(uu') = \sigma_u^2 I_T.$

ii. $E(X_i' u) = 0, i = 1, 2, \dots, k,$ donde X_i es la columna i -ésima de la matriz $X.$

iii. $plim\left(\frac{X'X}{T}\right) = \Sigma_{xx} < \infty,$ donde Σ_{xx} es una matriz simétrica, definida positiva.

Entonces se tiene que:

a) $plim\left(\frac{X'u}{T}\right) = 0_k$

b) $\frac{X'u}{\sqrt{T}} \xrightarrow{d} N(0, \sigma_u^2 \Sigma_{xx})$

ANEXO 5

ESTIMACIÓN CON RETARDOS DE LA VARIABLE ENDÓGENA CUANDO EL TÉRMINO DE ERROR NO TIENE AUTOCORRELACIÓN SERIAL

Si apareciesen valores retardados de la variable endógena como variables explicativas, entonces dejaría de cumplirse uno de los supuestos bajo los que se desarrollan las teorías de estimación e inferencia del modelo econométrico, pues algunas de las variables explicativas serían ahora variables aleatorias (ya que por ejemplo un rezago y_{t-1} lo es). Sin embargo, si el término de error no tiene autocorrelación, el problema de estimación no es muy importante.

Consideremos, por ejemplo, el modelo:

$$y_t = \beta y_{t-1} + u_t, \quad |\beta| < 1 \quad (1)$$

donde u_t es un proceso de ruido blanco. El estimador de mínimos cuadrados del parámetro β es:

$$\hat{\beta}_{MCO} = \frac{\sum_2^T y_t y_{t-1}}{\sum_2^T y_{t-1}^2} = \sum_2^T \frac{(\beta y_{t-1} + u_t) y_{t-1}}{\sum_2^T y_{t-1}^2} = \beta + \frac{\sum_2^T y_{t-1} u_t}{\sum_2^T y_{t-1}^2}$$

de modo que el estimador será insesgado si y sólo si se cumple:

$$E \left(\frac{\sum_2^T y_{t-1} u_t}{\sum_2^T y_{t-1}^2} \right) = 0 \quad (2)$$

Si la distribución de u_t fuese independiente de y_s para todo par (t,s) , entonces se tendría para $s = 2, \dots, T$

$$E \left(\frac{y_{s-1} u_s}{\sum_2^T y_{t-1}^2} \right) = E \left(\frac{y_{s-1}}{\sum_2^T y_{t-1}^2} \right) E(u_s) = 0 \quad (3)$$

por lo que la condición (2) se cumpliría y el estimador MCO sería insesgado. Sin embargo, (3) muestra que las distribuciones de y_t y u_s no son independientes, puesto que si el valor absoluto de β es inferior a la unidad, entonces:

$$y_t = \sum_{s=0}^{\infty} \beta^s u_{t-s}$$

por lo que y_t depende de u_t y de valores retardados de u_t . Por consiguiente, y puesto que u_t es ruido blanco, y_t será independiente de u_{t+s} , para $s \geq 1$; en particular, $E(y_{t-1}u_t) = 0$, por lo que el numerador de la fracción en (2) tendrá esperanza cero. Lo que ocurre es que el denominador incluye valores de y_t posteriores a u_s , por lo que no podemos efectuar el desarrollo en (3). El estimador MCO del modelo (1) será, en general, sesgado. Sin embargo, bajo determinadas condiciones, dicho sesgo tiende a cero, por lo que el estimador MCO de (1) es *consistente*.

Consideremos ahora el modelo:

$$y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_t u_t, \quad |\beta_2| < 1$$

cuyas variables explicativas y término del error satisfacen las siguientes propiedades:

- a) $E(u) = 0_T$, $E(uu') = \sigma_u^2 I_T$, es decir, no existe autocorrelación.
- b) $E(x_t u_t) = 0$ para todo t , ya que x_t es determinista.
- c) $E(y_{t-1} u_t) = 0$ pues aunque y_{t-1} es estocástica, si $|\beta_2| < 1$, y_{t-1} depende de u_{t-1}, u_{t-2}, \dots , pero no de u_t , y si este proceso es un ruido blanco, entonces se tiene el resultado citado.
- d) $\text{plim}\left(\frac{X'X}{T}\right) = \Sigma_{xx}$, matriz simétrica, definida positiva, donde:

$$X'X = \begin{pmatrix} T-1 & \sum_2^T y_{t-1} & \sum_2^T x_t \\ \sum_2^T y_{t-1}^2 & \sum_2^T y_{t-1} x_t & \\ & & \sum_2^T x_t^2 \end{pmatrix}$$

que a pesar de ser denotada por $X'X$ incluye también, en este caso, valores de la variable endógena, por aparecer un retardo suyo como variable explicativa. Esta condición d) se satisface, en general, bajo el supuesto $|\beta_2| < 1$, siempre que existan las varianzas y covarianzas de las variables explicativas x_t e y_{t-1} . Con estos supuestos, el teorema de Mann-Wald (véase Anexo 4) asegura que:

$$plim\left(\frac{X'u}{T}\right) = 0_k$$

y que:

$$\frac{X'u}{\sqrt{T}} \xrightarrow{d} N(0, \sigma_u^2 \Sigma_{xx})$$

de modo que se tiene:

$$plim \hat{\beta}_{MCO} = plim \left[\beta + \left(\frac{X'X}{T} \right)^{-1} \left(\frac{X'u}{T} \right) \right] =$$

$$= \beta + \text{plim} \left[\left(\frac{X'X}{T} \right)^{-1} \right] \text{plim} \left(\frac{X'u}{T} \right) = \beta + \sum_{xx}^{-1} 0_k = \beta$$

y el estimador de mínimos cuadrados ordinarios es consistente.

ABREVIATURAS

AP	Agricultura de Precisión
SSCM	Manejo Sitio Específico del Cultivo
GPS	Sistema de Posicionamiento Global
DGPS	Sistema de Posicionamiento Global Diferencial
GIS	Sistema de Información Geográfica
VRT	Tecnologías de Dosis Variable
MDI	Manejo Diferencial de Insumos
MCO	Mínimos Cuadrados Ordinarios
MCO	Mínimos Cuadrados Generalizados
MV	Máxima verosimilitud
ISC	Ingenio San Carlos
TMC/ha.	Toneladas métricas de caña por hectárea
lb. Az./tmc	Libras de azúcar sobre tonelada métrica de caña
ppm	partes por millón
cmol(+)/L	centimoles por litro de suelo
MO	Materia Orgánica
CE	Conductividad Eléctrica
ST	Suma total de cuadrados de la variable endógena

SE	Suma explicada
SR	Suma residual
d.e.	Desviación estándar
D.E.	Desviación estándar
g.l.	Grados de libertad
Ec.	Ecuación
Sig.	Significancia
t-asintótica	Distribución t de Student asintótica
L.I.	Límite inferior
L.S.	Límite Superior
ANOVA	Análisis de varianza
FIV	Factor de inflación de varianza
SAC	Modelo Espacial General
SAR	Modero Mixto Regresivo Espacial-Autoregresivo
SEM	Modelo Espacial Autoregresivo del Error
plim	probabilidad límite
W	Prueba de Wald
LR	Prueba de Razón de Verosimilitud
LM	Prueba de Multiplicador de Lagrange

KLIC

Criterio de Información de Kullback-Leibler

AIC

Criterio de Información de Akaike

SIMBOLOGÍA

N	Nitrógeno
K	Potasio
P	Fósforo
Mg	Magnesio
Ca	Calcio
S	Azufre
Zn	Zinc
Cu	Cobre
Mn	Manganeso
B	Boro
Mo	Molibdeno
Fe	Hierro
Cl	Cloro
pH	Potencial Hidrógeno
Al	Aluminio
X_{ij}	Variable que representa al factor k natural o de manejo j

y	Variable dependiente
X	Matriz de variable explicativas
$E[x]$	Valor esperado de la variable aleatoria x
$E(y/x)$	Valor esperado de y condicional a la variable explicativa x
$E(y/X)$	Valor esperado de y condicional a la matriz de variables explicativas X
u	Error estocástico
β_i	Coefficiente de regresión poblacional i
f	Función
\hat{y}	Estimador de $E(y/X)$
$\hat{\beta}$	Estimador de β
$\hat{\beta}_{MCG}$	Estimador de MCG de β
e	Vector de residuales
$p(X)$	Rango de la matriz X
k	Número de regresores
n	Tamaño de la muestra
$N(0, \sigma^2 I)$	Distribución Normal con media 0 y matriz de varianzas covarianzas $\sigma^2 I$

$\chi^2_\alpha(q)$	Distribución Chi-cuadrado con nivel α de significancia y q g.l.
$F_\alpha(q, n-k)$	Distribución F con nivel α de significancia q y $n-k$ g.l.
$t_\alpha(n-k)$	Distribución t de Student con nivel α de significancia y con $n-k$ g.l.
I	Matriz identidad de dimensión $n \times n$
σ^2	Varianza
$Var(x)$	Varianza de x
$Cov(x, y)$	Covarianza de x y y
R^2	Coefficiente de determinación múltiple
\bar{R}^2	Coefficiente de determinación múltiple corregido
$tr(X)$	Traza de la matriz X
s^2	Estimador de σ^2
s	Error estándar de la estimación
H_0	Hipótesis nula
H_1	Hipótesis alterna
ρ	Coefficiente autoregresivo espacial para la variable dependiente y

λ	Coeficiente autoregresivo espacial para el error estocástico u
Ω	Matriz de varianzas-covarianzas del modelo SAC
$h_i(z'\alpha)$	Forma funcional de heteroscedasticidad
W	Matriz de conectividad espacial
θ	Vector de parámetros desconocidos de un modelo
$\tilde{\theta}$	Estimador de θ
$\hat{\theta}$	Estimador de θ
θ_R	Estimador restringido de θ
$L(\theta, y)$	Función de verosimilitud para un vector y y θ
$s(\theta, y)$	Función score
$\overset{d}{\sim}$	Convergencia en distribución
$\overset{d}{\rightarrow}$	Convergencia en distribución
$I(\theta)$	Matriz de información de θ

BIBLIOGRAFÍA Y REFERENCIAS

- [1] Presentación propuesta para el IX Congreso de la SECS, Simposio Manejo de Recursos y Agricultura de Precisión, (Loja, Octubre 6-8, 2004)
“Aplicación práctica y rentable de la Agricultura de Precisión en la Agricultura Tropical: Caso caña de azúcar en el Ingenio San Carlos, Ecuador, por L. Estrada y E. Spaans”
- [2] G. Cox, H. Harris y D. Cox, Application of Precision Agriculture to Sugar Cane. (University of Southern, Queensland, Australia, 1997). En [1].
- [3] M. Bragachini, A. Von Martín y A. Méndez, Agricultura de Precisión en Argentina (Proyecto de Agricultura de Precisión, INTA Manfredi, Argentina, 2000)
- [4] Proceedings of the 10th Australian Agronomy Conference, (Australia, 2001) “Is agronomy being left behind by precision agriculture?, por S.E. Cook and R.G. Bramley”.
www.regional.org.au/au/asa/2001/plenary/2/cook.htm. En [1].

- [5] Bramley, R. y Quabba R, Opportunities for improving the management of sugarcane production through the adoption of precision agriculture – an Australian perspective (Australia, 2001). En: [1]
- [6] A.B. McBratney, Precision Agriculture Course: An Introduction to Concepts, Analysis and Interpretation (Australian Centre for Precision Agriculture, Australia, 2001), www.usyd.edu.au/su/agric/acpa.htm.
- [7] R. Bongiovanni, “La Agricultura de Precisión en la Cosecha”, Trabajo preparado para la Revista IDIA del INTA (Diciembre, 2003), www.agriculturadeprecision.org.
- [8] J. Lowenberg-DeBoer, Camino Sinuoso en la Adopción de Agricultura de Precisión. Traducción al español por R. Bongiovanni, (INTA Manfredi, Argentina, 1988).
- [9] J. Hatfield, Precision Agriculture and Environmental Quality: Challenges for Research and Education (National Soil Tilth Laboratory, Agricultural Research Service, USDA, Ames, Iowa, 2000). En: [7].

- [10] Proceedings of the 5th International Conference on Precision Agriculture and Other Resource Management, (Bloomington, MN, julio 16-19, 2000)
“A Management opportunity index for precision agriculture, por A.B. McBratney, B.M. Whelan, J.A. Taylor y M.J. Pringle.”
- [11] Proceedings of the 6th International Conference on Precision Agriculture, (Bloomington, MN, julio 14-17, 2002) “Spatial Regression, an Alternative Statistical Analysis for Landscape Scale On-farm trials: Case Study of Variable Rate Nitrogen Application in Argentina, por D. Lambert, J. Lowenberg-DeBoer y R. Bongiovanni.”
- [12] Tercer Taller Internacional de Agricultura de Precisión del Cono Sur de América, (Carlos Paz, Córdoba, Argentina, diciembre 17-19, 2002)
“Econometría Espacial: Una herramienta clave para el manejo sitio-específico de insumos, por R. Bongiovanni.”
- [13] L. Anselin, Spatial Econometrics: methods and models. (London, Kluwer Academic Publishers, 1988)

- [14] W. Tobler, Cellular Geography. En Philosophy in Geography (Editado por S. Gale y G. Olsson). En [13].
- [15] Invited presentation for the ENC-GNSS, session "Users at Word", (Rotterdam, The Netherlands, mayo 16-19, 2004) "Sense and non-sense of satellite navigation for precision agriculture in the tropics, por E. Spaans y L. Estrada."
- [16] L. Estrada, Agricultura de Precisión en el Ingenio San Carlos: Lotización (Departemento de Campo, ISC, 2003)
- [17] F. Subirós Ruiz, El cultivo de la Caña de Azúcar (1^{ra} edición, Costa Rica, San José, Universidad Estatal a Distancia San José, 1995)
- [18] J. S. Torres Aguas, R. Cruz Valderrama, F. Villegas Trujillo, Avances Técnicos para la programación y el manejo del riego en caña de azúcar. (Centro de Investigación de la Caña de Azúcar de Colombia: CENICAÑA, Cali, Colombia, Serie técnica No. 33, 2004)

- [19] R. Quintero Durán, Interpretación del análisis de suelo y recomendaciones de fertilizantes para la caña de azúcar. (Centro de Investigación de la Caña de Azúcar de Colombia: CENICAÑA, Cali, Colombia, Serie técnica No.14, 1993)
- [20] C. Pérez López, Técnicas de Muestreo Estadístico. Teoría, práctica y aplicaciones informáticas. (Madrid, España, Alfaomega, 2000)
- [21] J. Johnston y J. DiNardo, Econometric Methods (4^{ta} edición, New York, McGraw-Hill, 1997)
- [22] A. Novales, Econometría (2^{da} edición, Madrid, McGraw-Hill, 1993)
- [23] Gujarati, Damodar N., Econometría (3^{ra} edición, McGraw-Hill, 1997)