Dpto. IO
Instituto de Computación
Facultad de Ingeniería
Universidad de la República

MODELIZACIÓN Y ESTIMACIÓN DEL RENDIMIENTO EN CULTIVOS AGRÍCOLAS

Informe de Proyecto de Grado para la obtención
del título de Ingeniero en Computación

Montevideo, Uruguay, septiembre del 2008.

Autores:
Álvaro Crespi
Enrique Mora
Giovani Sosa

Tutor:
MSc. Ing. Omar Viera – Io Inco Fing
Supervisor:
MSc. Mercedes Berterretche - ICA

Resumen

La Agricultura de Precisión es un concepto ampliamente utilizado en el presente. Se intenta a medida que el tiempo avanza, tratar de complementar cada vez más las prácticas agrícolas con la tecnología, en virtud de todos los beneficios que se pueden obtener. Como antecedente de la aplicación de tecnologías en las prácticas agrícolas, se encuentran los Sistemas de Posicionamiento Global (GPS), los Sistemas de Información Geográfica (SIG), los Monitores de Rendimiento, Manejo de Dosis Variable, etc. La integración de las mencionadas tecnologías es fundamental para poder analizar los datos recolectados y sacar conclusiones provechosas; una de las más importantes es poder predecir el rendimiento de los cultivos. En lo que refiere a la predicción, existen distintos estudios sobre la predicción de los valores que puede tomar una variable en base a datos históricos (en este caso, de corte transversal) de la misma y de factores que influyen en ella. La mayoría de estos estudios utilizan las herramientas que brinda la Econometría Espacial (EE), y atacan problemas puntuales en lo que refiere a las distintas etapas que se deben realizar para llegar a la predicción, pero ninguno de ellos analiza estos problemas en su totalidad. Teniendo en cuenta lo anterior, surge la motivación de dar una solución al problema de predecir el rendimiento de un cultivo para una futura zafra, en base a datos recolectados de una chacra, utilizando para ello las técnicas que brinda la EE, integrando en una única solución la posibilidad de realizar un estudio de las propiedades que cumplen los datos, modelar mediante modelos de regresión el rendimiento en base a los datos recolectados, estimar el modelo obtenido y predecir el rendimiento para la zafra futura. Para esto se plantea una solución al problema mediante una sucesión de etapas a seguir para lograr finalmente el modelo del rendimiento. En términos genéricos las etapas son: el análisis de los datos, la selección del modelo, la estimación del modelo, la validación del modelo y finalmente la predicción del rendimiento para una futura zafra. Cada etapa de la solución se puede ver como un modulo dentro de un Sistema informático, por eso se planteó un diseño y arquitectura para la solución general y finalmente se desarrolló y testeó el modulo de análisis de los datos. Como resultado de todo lo anterior se obtuvo una solución general para el modelado del rendimiento mediante modelos de regresión y una librería informática que represente la solución, no desarrollada completamente pero si diseñada totalmente. En cuanto a los resultados obtenidos de la investigación se integraron en una solución genérica todos los posibles caminos que se pueden tomar para la toma de decisiones a la hora de encontrar el modelado del rendimiento, dejando un gran precedente para posibles investigaciones futuras que tomen a este proyecto como buen punto de partida teniendo en cuenta la ardua investigación llevada a cabo. Como resultado de la ejecución del módulo de análisis de datos, con datos reales de dos potreros para dos zafras otorgados por el cliente, se obtuvieron resultados que indicaban que para uno de los potreros se detectaba autocorrelación espacial en los datos y no se detectaba multicolinealidad. Sin embargo para el otro potrero en sus dos zafras se detectaba autocorrelación espacial en el rendimiento pero además multicolinealidad alta en los datos lo cual indica que algunas de las variables tenidas en cuenta para el modelo están relacionadas linealmente y se deberían de quitar para a posteriori realizar nuevamente el análisis.

Palabras claves: Agricultura de Precisión, Econometría, Econometría Espacial, Modelo de Regresión, Predicción, Autocorrelación, Heteroscedasticidad.

Contenido

Capítulo 1

Introducción

1.1 Contexto Marco

Este trabajo forma parte del proyecto de grado de la carrera Ingeniería en Computación de la facultad de Ingeniería (FING), Universidad de la República (UdeLaR). Tiene como objetivo realizar una investigación y desarrollo de software en el área de la Agricultura más específicamente en la Agricultura de Precisión (AP, [22]), para la empresa consultora ICA. [21]

El principal interés de ICA es poder darle al productor agropecuario herramientas tecnológicas para poder sacar el mejor provecho posible a su chacra enmarcado en todo el contexto que la AP abarca. Con este objetivo una de las respuestas que se desea obtener es cómo se comporta el rendimiento de los cultivos con respecto a los factores que influyen en él. Si se puede plantear el comportamiento del rendimiento de cultivos mediante un modelo matemático, el usuario puede obtener una herramienta robusta para predecir rendimientos futuros y/o saber qué tanto influye un factor en la variación de los rendimientos.

Mediante el planteo de este trabajo de grado se desea desarrollar una actividad de investigación, dada la escasa difusión que existe en esta área en nuestro país, es importante brindar una aproximación a una solución informática para este problema.

1.2 Motivación

Junto al avance de la tecnología y la sofisticación de las prácticas agrícolas, surgen nuevos desafíos y oportunidades, principalmente respecto al concepto del cuidado ambiental, la administración eficiente de los recursos naturales y la economía del proceso de producción. La manera de enfrentar dichos desafíos ha sido la generación de tecnología que permita desarrollar técnicas y metodologías que cuantifiquen y manejen diferenciadamente la variabilidad natural del área productora. El manejo localizado de las prácticas agrícolas, permite principalmente una mayor eficiencia de aplicación y costeo de insumos, reduciendo el impacto sobre el medio ambiente, aprovechando las condiciones agrícolas, geográficas y ecológicas particulares de cada sitio y en consecuencia, disminuyendo los costos y aumentando los rendimientos de la producción de alimentos. A ese conjunto de procesos y sistemas aplicados se los denomina Agricultura de Precisión.

Un desafío importante a la que es expuesta la tecnología, es poder en base a la recolección de datos históricos de un lugar físico dedicado a la agricultura, realizar un análisis consistente de los datos para poder sacar conclusiones que ayuden a la continua superación a todo nivel de la producción agrícola. Si un productor puede predecir los rendimientos de su chacra antes de decidir que cultivo plantar, el mismo podría saber de antemano si le es redituable o no llevar adelante su plantación o seleccionar otro tipo de cultivo de forma tal de maximizar su rendimiento en un momento dado del tiempo para maximizar su rentabilidad.

1.3 Introducción al problema

La definición del problema tuvo su fase inicial en la idea de lo que el cliente quería en base a los conocimientos y los datos recolectados de las chacras que en principio el cliente podía contar. Con la participación en conjunto de los autores, el tutor del proyecto y el cliente se llegó a una definición del problema. Posteriormente y a lo largo de las primeras investigaciones se debió replantear la definición del problema llegando a tener como principales cometidos los siguientes puntos:

- Investigar y documentar los posibles caminos y metodologías para la obtención de un modelo matemático que represente el rendimiento de cultivos en base a datos obtenidos de una determinada chacra para una zafra en particular.
- Una vez obtenido el modelo del rendimiento en función de los factores influyentes, ver las distintas formas o posibilidades de predecir el rendimiento para zafras futuras a corto plazo.

Los datos con los cuales se cuenta son datos de corte transversal[1], este tipo de datos en general posee características especiales determinadas por los llamados efectos espaciales[2]. Estos son tratados en profundidad por la Econometría Espacial la cual proporciona un conjunto de herramientas formales que permiten en primer lugar contrastar la presencia de dichos efectos. En segundo lugar permite realizar una correcta especificación de la realidad mediante modelos de regresión, formalizando su tratamiento, estimación y predicción.

A lo largo de la investigación se ve como resolver el problema propuesto sirviéndose de las herramientas que brinda la Econometría Espacial. [1]

1.4 Objetivos

El principal objetivo es realizar una investigación sobre las diferentes metodologías que se pueden aplicar para la predicción del rendimiento, teniendo como punto de partida los datos recolectados en las chacras. Se intenta mediante este proyecto de grado dejar un buen precedente para posibles aplicaciones de las metodologías y posibles trabajos futuros que tomen a éste como punto de partida. Una vez realizada la investigación se debe dar una solución al

[1] Datos de corte transversal se entiende como valores muestrales de distintas propiedades de elementos que se distribuyen a través del espacio, obtenidos en un mismo instante de tiempo.
[2] Los efectos espaciales son la heterogeneidad y la dependencia espacial. Estos se suelen originar cuando se trabaja con datos de corte transversal.

problema con la mejor metodología a nuestro entender que se pueda aplicar a los datos con los cuales se cuenta.

Ya mencionamos que la Econometría Espacial proporciona una batería de herramientas especializadas para el tratamiento de datos espaciales, por medio de esta se puede modelar la variable rendimiento en función de las variables que en él influyen mediante los modelos de regresión. Una vez concluida la investigación se estará en condiciones de brindar de acuerdo al objetivo, una solución general al problema de modelar el rendimiento, analizando los datos recolectados y tomando decisiones sobre cuales de los modelos de regresión se aplican en base a los contrastes necesarios.

El segundo objetivo es en base a la solución encontrada brindar un diseño informático para la implementación detectando los casos de uso, diseñando la arquitectura e implementando los casos de usos que se mencionaran en el alcance del Sistema.

1.5 Método de solución

Previo a realizar una investigación directa sobre las distintas metodologías que brinden herramientas que permitan realizar una predicción en base a datos históricos (en nuestro caso del rendimiento de un cultivo), se debió realizar una investigación y familiarización sobre el contexto en que se enmarca la investigación (la AP), el producto de esto fue la elaboración de un documento de el estado del arte de la AP. [22]

El segundo paso fue realizar una investigación de las diferentes técnicas y metodologías de predicción y/o descripción de una variable en base a datos históricos de la variable en cuestión y otras que influyen en ella, en particular esta investigación se centra en aquellas metodologías que se basan en Series Temporales[3], ya que el enfoque y los lineamientos del proyecto se inclinan hacia esta metodología. El producto de esto fue la elaboración de un documento del estado del arte de métodos y modelos de predicción temporal. [23]

El tercer paso de la investigación surge por la imposibilidad por parte del cliente de contar con una serie temporal de datos históricos, este plantea la posibilidad de contar con a lo sumo datos históricos de dos zafras (datos de corte transversal). Por lo tanto la investigación sufre un desvío y se debe centrar en el estudio de herramientas que brinda tanto la Econometría[4] como la Econometría Espacial, para poder obtener un modelo econométrico y en base a éste realizar las predicciones del rendimiento. El fruto de esta investigación fue la elaboración de un documento del estado del arte de métodos de análisis espacial (ver ref. [24]). En este documento se compara las posibilidades que existen para tratar datos de corte transversal, la Econometría o la Econometría Espacial; se llegó a la conclusión que la segunda opción es la apropiada para aplicar en el caso de estudio, debido a que incorpora los efectos espaciales en el tratamiento de los datos y en el modelo de regresión formulado.

Finalmente el proceso culmina con la elaboración de una solución al problema de predicción de rendimiento en base a datos de corte transversal. La solución al problema se basa en una herramienta que contiene un conjunto de contrastes de especificación, modelos de

[3] Una serie temporal representa una sucesión de valores de una variable, ordenados en el tiempo.

[4] La Econometría forma parte de la ciencia económica y se encarga de estudiar las matemáticas y estadísticas aplicadas a las teorías económicas, para poder verificar y solucionar los problemas económicos por medio de modelos.

especificación, métodos de estimación, procedimientos de validación del modelo estimado y procedimientos que permiten realizar la preedición; donde la interacción conjunta de estos elementos permite brindar la predicción y/o descripción deseada por el usuario. Luego de formulada la solución se procedió a la implementación de unos de los componentes que componen la solución, este es el Análisis de Datos. Este componente contiene distintos contrastes para chequear los efectos espaciales y otros contrastes para chequear distintas propiedades en los datos necesarias para que el modelo pueda ser estimado mediante los estimadores propuestos.

1.6 Conclusiones

Como conclusión fundamental de este trabajo se tiene que la rama de la Econometría llamada Econometría Espacial (EE) es la que mejor se adapta al problema de predicción que se plantea resolver, debido a que tiene en cuenta los llamados "efectos espaciales".
Las etapas de la solución son: análisis de datos, selección, estimación y validación del modelo, y finalmente predicción del rendimiento, y surgen como conclusión de la investigación de diferentes bibliografías, artículos científicos y estudios académicos.
Ninguna de las fuentes bibliográficas atacan de forma general (siguiendo las distintas etapas antes mencionadas) el problema de predicción de una variable por medio de modelos econométricos espaciales, tampoco fue posible encontrar un estudio que integre todas las posibles soluciones, dependiendo de las propiedades que los datos puedan cumplir.

Como conclusión del objetivo de diseñar una herramienta informática que implemente la solución brindada por la investigación, se diseñó teniendo en cuenta criterios tales como la extensibilidad, mantenibilidad y adaptabilidad.
De los análisis realizados por la biblioteca sobre los archivos de datos se confirmó la presencia de autocorrelación global positiva en los rendimientos de los cultivos. Esto permite concluir que el uso de la EE es adecuado en este caso, dado que de no detectarse autocorrelación la EE no sería necesaria. En cuanto a la autocorrelación en el error, se verificó su presencia, también en forma positiva y global.

En base a la detección de los dos tipos de autocorrelación, se concluye que la forma correcta de modelar sería mediante un modelo de autocorrelación mixta, que tenga en cuenta tanto la autocorrelación en el rendimiento como en el error. Sin embargo, en este trabajo se propone el utilizar la estrategia seguida por Anselin y Florax, vista en la sección 11.23.4.4, entonces el modelo seleccionado es el modelo de ponderación, o el modelo de error espacial.

Dado que como fruto de la investigación se esperaba detectar heteroscedasticidad en los datos y el contraste de White no detectó la presencia de dicho efecto se concluye que no se puede asegurar que el hecho de omitir los insumos en el modelo implique la presencia de este efecto espacial.

Finalmente, siendo este trabajo la primera experiencia en la materia de este grupo en dicha problemática, y estando los resultados obtenidos dentro de los parámetros aceptables, y considerando la importancia a nivel de la Agricultura en forma general, se concluye que es posible, rentable y provechoso el profundizar en el estudio de este problema.

1.7 Estructura del informe

El documento de trabajo se organiza en siete capítulos; en el Capítulo 1 es la introducción, contiene la motivación del proyecto, introducción al problema, los objetivos del proyecto, el método de solución utilizado, los resultados esperados y las conclusiones a las que se llegaron. El Capítulo 2 contiene la definición del problema con mayor detalle, la evolución de los requerimientos a lo largo del proyecto, los requerimientos definidos con el usuario y antecedentes de investigación en el problema propuesto. El Capítulo 3 contiene el marco teórico el cual contiene todos los conceptos relevantes en este proyecto necesarios para comprender el problema, estos son, una visión de la Econometría Espacial enfocado hacia la Agricultura de Precisión, un análisis de los efectos espaciales, definición de la matriz de contigüidad, los modelos de regresión espacial, la forma en que se estima dichos modelos, las formas en que se valida el modelo y el uso del modelo para predecir. El Capítulo 4 contiene la solución propuesta para resolver el problema, este capítulo se encuentra dividido en las distintas etapas que componen la solución. El Capítulo 5 contiene la solución informática diseñada, la cual contiene la especificación funcional, los diagramas de secuencia del sistema, el alcance del sistema, el diseño del sistema y la implementación y testeos realizados. En el Capítulo 6 se expone y analiza los resultados obtenidos. El Capítulo 7 expone la conclusión a la que se llegó por parte de la investigación realizada. El Capítulo 8 contiene los posibles trabajos futuros que se podría realizar, que extiendan o complementen la investigación realizada. Con respecto a la bibliografía utilizada, esta se encuentra al final del documento. Los apéndices que contiene son: el Apéndice A, contiene el análisis de riego del proyecto, Apéndice B, contiene la administración del proceso, el Apéndice C, contiene el índice de figuras, el Apéndice D, contiene la evolución de la investigación justificando los cambios en los requerimientos, el Apéndice E, contiene la especificación de los casos de uso, finalmente el Apéndice F contiene el documento de Especificación Preliminar de los Requerimientos.

Los capítulos de necesaria lectura para comprender el trabajo realizado son, Capítulo 1, Capítulo 2, Capítulo 3, Capítulo 4 y Capítulo 7.

Capítulo 2

Definición del problema

2.1 *Definición*

El concepto de AP implica entre otras cosas utilizar las tecnologías existentes para poder especificar, planificar y tomar decisiones de forma de maximizar la producción y cumplir con prácticas de políticas amigables con el ambiente que nos permitan poder sacar provecho del campo, y que a la vez el mismo sea sustentable.

Si el cometido es explotar al máximo las propiedades de la chacra, es intuitivo pensar que si existen zonas determinadas con diferentes características, sea conveniente aplicar un tratamiento diferencial a cada una de ellas. En la AP esta idea se conoce como zonas de manejo[5].

En base a esta idea, es que se plantea el desafío de contar con una serie de herramientas que apliquen estas técnicas y conocimientos, y que asistan en la toma de decisiones. Dicho marco de trabajo se centra en tres grandes líneas que involucran el determinar las zonas de manejo existentes en una chacra, predecir el rendimiento dado datos históricos, y como último punto brindar prescripciones en cuanto a la cantidad de insumos necesarios para poder obtener los rendimientos estimados anteriormente.

Dicho esto, la línea de trabajo que se siguió en este proyecto es el de predecir el rendimiento de una chacra, en base a datos históricos. Los otros puntos del problema global fueron resueltos por otros dos grupos de proyecto.

Algunos de los problemas al los que nos vamos a enfrentar al querer predecir el rendimiento de cultivos en base a datos históricos recolectados de zafras anteriores, tanto de rendimiento como de variables que influyen en él, es el alto grado de incertidumbre tanto cuantitativo como cualitativo de los datos recolectados, representación del rendimiento en base a los datos que se

[5] Las chacras se dividen en zonas de manejo, donde cada zona se determina por la variación de productividad de la chacra. Cada zona de manejo debe cumplir que a) la variabilidad del rendimiento entre dos zonas debe ser mayor que la variabilidad dentro de la propia zona, y b) los factores que limitan el rendimiento dentro de una zona deben ser los mismos.

tiene, propiedades que los datos cumplen o no cumplen, medición del grado de confiabilidad de la predicción.

Teniendo en cuenta lo anterior se puede definir claramente al problema en cinco puntos fundamentales que son: realizar un *análisis* completo de distintas propiedades espaciales y lineales de los datos que se tienen mediciones, determinar un *modelo* matemático del rendimiento en función de las variables que influyen en él, *estimar* la representación matemática obteniendo resultados concretos para ésta, *predecir* rendimientos futuros en base a las estimaciones obtenidas y de ser posible contrastar las predicciones para *medir* los resultados obtenidos.

2.2 *Evolución de los Requerimientos*

El objetivo inicial del trabajo fue construir un sistema informático que permita predecir el rendimiento de un cultivo en base a datos históricos de zafras anteriores de una chacra y presentar los rendimientos predichos en diferentes zonas de manejo. Específicamente el sistema informático recibía como entrada una serie de tiempo que contiene las zonas de manejo de la chacra, índices de performance (índices que ayudan a decidir el número óptimo de zonas de manejo) y rendimiento para cada una de ellas, además se cuenta con información del clima y distintos atributos del suelo. Estos datos se supone que se obtuvieron en distintas zafras para la misma chacra, tomados en periodos regulares a través del tiempo (cada cierta cantidad de meses, dependiendo de la duración entre una zafra y otra).

El sistema debe proporcionar como salida una predicción de las nuevas zonas de manejo para la zafra siguiente, esto es: si los datos son anuales y llegan hasta el año 2007, se debe predecir las zafras del año 2008. Conjuntamente con cada zona de manejo se debe predecir el rendimiento de cada una de ellas y distintos índices de performance. La figura 1 ilustra la primer versión de los requerimientos, por más detalle ver apéndice F (documento de Especificación Preliminar del Proyecto).

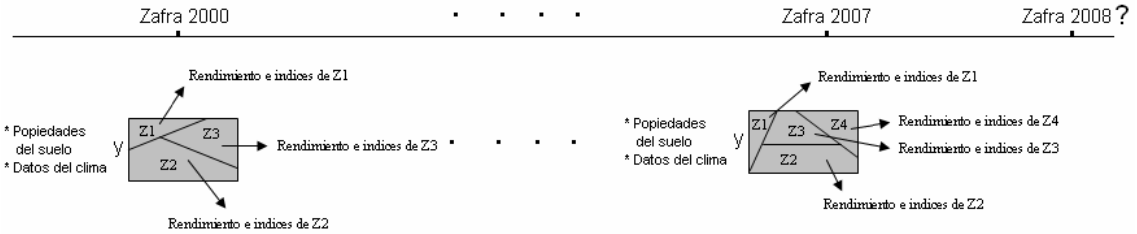


Figura 2.1 Primera versión de los requerimientos

Cada zafra se encuentra dividida en celdas regulares, cada una de ellas contiene una muestra del rendimiento y de distintas propiedades del suelo y del manejo del mismo (muestreo en grilla sistemático[6]), a su vez cada celda está identificada de forma de indicar a la zona de manejo que pertenece. La figura 2 ilustra lo explicado, en la misma las zonas de manejo están identificadas por el color de la celda. La salida del sistema es una predicción de cómo se encontrarán divididas en un futuro las zafras y la predicción del rendimiento en cada celda en la que se encuentra dividida la chacra.

[6] Se detalla en la sección 3.3.

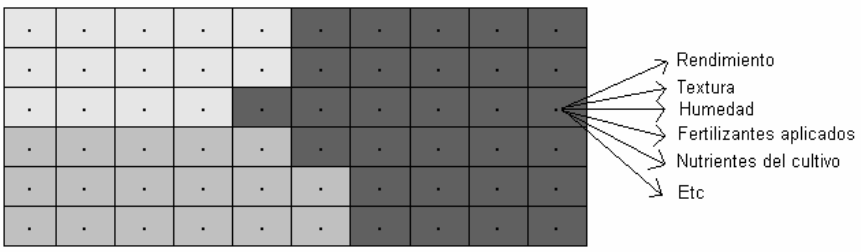


Figura 2.2 Muestreo en grilla sistemático indicando las zonas de manejo.

Luego que se llega a una definición madura de los requerimientos, surge el primer punto de inflexión sobre los mismos. La herramienta a desarrollar debe ejecutar dos algoritmos claramente identificados para producir la salida, un de ellos para obtener la predicción del rendimiento en cada celda en la cual esta dividida la chacra y otro algoritmo para determinar las nuevas zonas de manejo agrupando cada celda en la zona que corresponda. Justamente este ultimo algoritmo está muy relacionado con el objetivo de unos de los grupos (que componen el proyecto global mencionado en la sección 2.1) el cual es obtener las zonas de manejo de la chacra. Se llega a un acuerdo, recortándose el alcance del sistema, se elimina las zonas de manejo como entrada y salida del sistema y el resto de los datos y requerimientos se mantienen en su forma original.

La investigación realizada para resolver el problema anterior se centra en las metodologías capaces de predecir el rendimiento de un cultivo en base a datos históricos del mismo y de distintas variables que se suponen que influyen en el rendimiento. Cabe destacar que la investigación es puramente de carácter temporal, de forma de reducir la complejidad del problema no se tuvo en cuenta las diferentes relaciones espaciales que puedan llegar a tener los datos de la chacra. Esto significa que la predicción del rendimiento en una celda de la chacra se realiza en base a los datos históricos de lo que ocurrió en dicha celda, sin tener en cuenta lo que ocurrió en las celdas vecinas. Se plantea una solución basada en los modelos multidireccionales VAR (ver ref. [23]).

Una restricción importante para aplicar un modelo VAR es que se debe contar con una importante cantidad de datos históricos, como es el caso de cualquier modelo de series temporales. Justamente esta restricción causa un punto de inflexión importante en el desarrollo de la tesis, debido a que el cliente no puede contar con la cantidad de datos históricos acordada, esto llevo a reformular los requerimientos de forma de reencaminar el proyecto. El cliente plantea la posibilidad de contar con a lo sumo datos históricos de tres zafras, por lo tanto ya no se cuenta con datos de corte longitudinal (en el tiempo) pero si se cuenta con datos de corte transversal, pues se tiene muchas muestras del rendimiento del cultivo para una misma zafra.

Teniendo en cuenta este cambio en los datos con los que se cuenta se debió reformular los requerimientos. En base a la entrevista brindada por una experta en el área de la Econometría, nos plantea la posibilidad de aplicar técnicas econométricas, más específicamente el modelo de regresión lineal, el objetivo de aplicar este modelo pueden ser dos:

- Obtener una función que permita predecir el valor del rendimiento una vez conocidos los valores de los factores que influyen en él. En este caso se utiliza el modelo de regresión como un modelo predictivo.
- Conocer la relación que existe entre el rendimiento y los factores que lo limitan. De esta forma el usuario del sistema puede variar el valor de los factores y ver como

responde el rendimiento del cultivo. De esta forma se puede conocer por ejemplo en cuanto puede disminuir la producción si para la próxima zafra diminuyen las precipitaciones. En este caso se utiliza el modelo de regresión como un modelo explicativo.

Teniendo en cuenta los beneficios de aplicar un modelo de regresión lineal, surgen los nuevos requerimientos. Uno de ellos se basa en utilizar el modelo como forma predictiva, esto es poder predecir (para la próxima zafra) el rendimiento del cultivo en cada celda en que fue dividida la chacra, teniendo como entrada los datos del rendimiento, propiedades del suelo y clima de la zafra anterior (divididos por celda), además se debe contar con los nuevos datos del suelo y clima próximos a la nueva zafra, de forma de sustituir estos datos en la función que se obtiene y obtener los rendimientos esperados para la nueva zafra, la siguiente figura ilustra lo explicado.

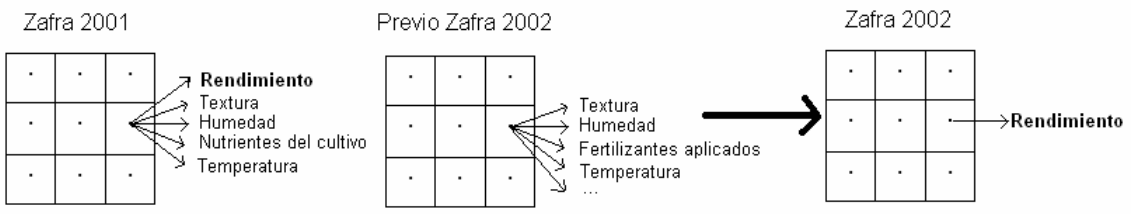


Figura 2.3 Predicción del rendimiento en base a datos de corte transversal

El otro requerimiento se basa en devolver el modelo de manera que permita ser utilizado de forma explicativa. Esto es devolver los parámetros estimados del modelo teniendo como entrada los datos del rendimiento, propiedades del suelo y clima de una determinada zafra. La siguiente figura ilustra lo explicado.

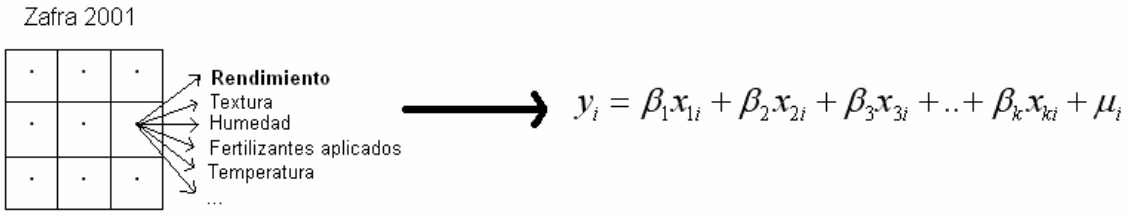


Figura 2.4 Obtención de un modelo de regresión lineal en base a datos de corte transversal

Pero a medida que se profundiza en la investigación sobre el modelo de regresión lineal surge un gran inconveniente sobre la utilización de estos modelos en datos de corte transversal, esto se debe a que estos modelos no obtienen buenos resultados cuando los datos con los que se trabaja (en especial la variable endógena y los errores cometidos por el modelo) se encuentran correlacionados espacialmente. La autocorrelación en el modelo de regresión lineal se trata de forma unidireccional, esto debido a que los datos debieron linealizarse de forma de poder aplicar el modelo. Pero la autocorrelación en datos de corte transversal se da de forma multidireccional y al ser linealizados pierde sentido estudiar esta propiedad en los datos.

Por lo tanto, la investigación se centró en la Econometría Espacial, ésta proporciona las herramientas necesarias para poder modelar y estimar el rendimiento de un cultivo, teniendo en cuenta los efectos espaciales que poseen los datos muestrales. La investigación no fue sencilla debido a la poca difusión que tiene esta área de la Econometría, la mayor parte de los manuales de econometría se limitan a mencionar los problemas de efectos espaciales en el análisis econométrico sin profundizar en ellos, ejemplo de ello son los manuales de econometría general

de Green y Novales [3][20]. De todos modos la investigación permite comenzar a diseñar una solución al problema, la cual debe contar con un conjunto de contrastes de especificación, modelos de especificación, métodos de estimación y procedimientos de validación. La interacción conjunta de estos elementos permite brindar la predicción deseada por el usuario. En cada uno de los aspectos mencionados que componen la solución, las variantes que se pueden aplicar son variadas y de compleja formulación matemática. En muchos de ellos no existe un consenso sobre la metodología a aplicar, de todas formas la solución al problema planteado por el cliente intenta ser lo mas genérica posible, teniendo presente los tiempos del proyecto.

2.3 *Requerimientos*

Se cuenta con los siguientes datos históricos de dos zafras de una chacra con las siguientes variables:

Datos topográficos
- Curvatura
- Elevación
- Orientación
- Pendiente
- Plano de curvatura
- Perfil de curvatura
- Área específica de cuenca
- Índice topográfico SPI
- Índice topográfico STCI
- Índice topográfico WI
- Cuenca

Datos de monitor de rendimiento normalizados
- Rendimiento normalizado
- Humedad normalizada

Datos de monitor de conductividad eléctrica normalizados
- CE a 30 cm.
- CE de 30 a 90 cm.
- CE a 90 cm.
- CE relación 90-30

Datos de imagen satelital
- Reflectancia de diferentes bandas
- Bandas Tasseled cap 1-3
- Índice NDVI
- Índice GNDVI

Datos de suelos
- Índice CONEAT
- Drenaje
- Muestreo de PBray

Se debe investigar la posibilidad de poder inferir un modelo para la función de producción en base a los datos históricos otorgados. Para esto se debe realizar documentación completa de la investigación exponiendo las diferentes metodologías de trabajo que se puedan aplicar. Una vez investigadas las posibilidades, justificar el uso de la metodología utilizada y proponer una solución completa al problema especificado en la sección 2.1. Realizar un Sistema en forma de

librería, especificar el análisis, arquitectura y diseño completo del Sistema. Implementar el Sistema especificado.

Los requerimientos no funcionales propuestos por el cliente consisten en:

- Ser fácilmente integrable.
- Tener una buena mantenibilidad.
- Buena performance en cuanto a los cálculos de predicción.

Capítulo 3

Marco Teórico

3.1 La Econometría Espacial en la agricultura

La Econometría Espacial es una rama de la Econometría, que surge por los llamados "efectos espaciales": la heterogeneidad y la dependencia espacial. Estos se suelen originar cuando se trabaja con datos espaciales (de corte transversal). Anselin, es quien da la primera definición de la Econometría Espacial: "Es un conjunto de métodos y técnicas, que se basan en una representación formal de la estructura de dependencia y heterogeneidad espacial, proporcionando las herramientas para llevar a cabo una correcta especificación, estimación, test de hipótesis y predicción de modelos." [1]

En la anterior definición se encuentra la principal razón por la que el foco de la investigación se centra en la metodología econométrica espacial para llevar a cabo el estudio del rendimiento de un cultivo, esta proporciona las herramientas necesarias para poder modelar y estimar el rendimiento de un cultivo teniendo en cuenta los efectos espaciales que poseen los datos muestrales con los que se cuenta.

La heterogeneidad espacial surge por la falta de estabilidad del fenómeno en estudio a través del espacio, esto significa que el fenómeno no se comporta de forma homogénea. Ésta se manifiesta de dos formas distintas, como heteroscedasticidad o inestabilidad estructural. Por otro lado, la dependencia espacial o autocorrelación espacial, surge cuando el valor de una variable en un lugar del espacio está relacionado con el valor de la variable en otro u otros lugares del espacio. [1]

Los datos que se analizan en el caso de estudio, no escapan de los efectos espaciales, es de esperar que el rendimiento de un cultivo se encuentre correlacionado a través del espacio, esto se debe a que los valores del rendimiento entre puntos vecinos tienden a ser parecidos, este es un caso de autocorrelación positiva. Como se verá posteriormente, la autocorrelación además puede ser negativa.

Con respecto a la heterogeneidad, es común que ésta también se presente, ya que suena ilógico pensar que las propiedades de un terreno sean totalmente homogéneas, teniendo en cuenta la cantidad de factores que influyen en éste: químicos, físicos, fisiológicos y climáticos. Debido a la heterogeneidad espacial de las variables es que uno de los principales objetivos de la agricultura de precisión es poder determinar las zonas de manejo y así encontrar aquellos factores que limitan el rendimiento en determinadas zonas del cultivo. [4]

Por lo general, la heterogeneidad espacial puede ser resuelta mediante las técnicas de econometría estándar, pero no ocurre lo mismo con la autocorrelación espacial, la Econometría Espacial surge principalmente por este último efecto espacial. [1]

Los primeros estudios de Econometría Espacial se remontan a la década de los setenta, Paelinck y Klaassen (1979) denominaron este término a partir de sus investigaciones acerca la autocorrelación espacial en los términos de error producidos por las regresiones. Es en la década de los 80 cuando surgen los principales estudios por parte de Cliff y Ord (1981), Blommestein (1981) y Anselin (1988a), este último es quien da la primera definición de la Econometría Espacial, citados por Anselin, en el estudio sobre Spatial Econometrics: Methods and Models (1998). [1]

Mas allá del desarrollo de la Econometría Espacial, esta tiene menor difusión que la econometría clásica, esta diferencia es más notoria en el aspecto teórico, la mayor parte de los manuales de econometría se limitan a mencionar los problemas de efectos espaciales en el análisis econométrico sin profundizar en ellos, ejemplo de ello son los manuales de econometría general de Green y Novales.

En el aspecto empírico la adopción de la econometría espacial no ha sido la mejor, ya advertido por Anselin y Florax (1995a), "…. a pesar de los importantes desarrollos metodológicos, sería excesivo sugerir que la econometría espacial se ha convertido en una práctica aceptada en la investigación empírica en la ciencia regional y en la economía regional".

3.2 *Los efectos espaciales*

3.2.1 La autocorrelación espacial

En términos generales, la autocorrelación espacial se puede considerar como una extensión sobre la autocorrelación temporal, donde en el primer caso el análisis es sobre dos dimensiones y en el segundo sobre una dimensión. La autocorrelación espacial surge cuando existe una relación funcional entre los valores que puede tomar una variable en un punto del espacio, en relación a los valores que toma en otros puntos cercanos del espacio. Matemáticamente esto se expresa como $E\,[x\,(i),\,x\,(j)] \neq 0$, donde x es la variable en estudio e i y j son dos puntos vecinos del espacio. [2]

Los primeros estudios para detectar la presencia de autocorrelación espacial se remontan a los trabajos de Moran (1948) y Geray (1954). En lo que respecta a los modelos de regresión en

presencia de autocorrelación espacial son Paelinck y Klassen (1979) quienes realizaron los primeros estudios sobre autocorrelación espacial en los términos de perturbación del modelo de regresión. [2]

La autocorrelación espacial es muy similar a la autocorrelación temporal observada en las series de tiempo. Sin embargo, en las series de tiempo este problema econométrico es únicamente unidireccional, es decir que el pasado explica el presente y puede ser corregido simplemente con el operador de rezago. Por su parte, la dependencia espacial es multidireccional, es decir todas las regiones pueden afectarse entre sí. Esto imposibilita la utilización del operador de rezago utilizado en series de tiempo, y motiva la implementación de la matriz de contigüidad espacial (ver sección 3.3). [2]

Teniendo en cuenta la definición de autocorrelación espacial parece lógico pensar que el rendimiento en un cultivo se puede encontrar correlacionado espacialmente, esto se debe a que valores altos o bajos de rendimiento tienden a estar rodeados de valores con las mismas características. Esto implica que el rendimiento en un punto no esta únicamente explicado por los valores de aquellos factores que influyen en el rendimiento, además depende de valores del rendimiento en otros puntos del espacio.[4] [2] [6]

La autocorrelación puede ser positiva o negativa, es positiva cuando un valor alto o bajo de una variable tiende a estar rodeado de valores de la variables altos o bajos respectivamente; por el contrario en la autocorrelación negativa un valor bajo de un variable en un punto del espacio tiende a que los valores de las observaciones cercanas a él sean disímiles. De no darse ninguno de los dos casos anteriores entonces los datos se encuentran distribuidos de forma aleatoria en el espacio. En especial para el caso del rendimiento de un cultivo, es de esperar que la autocorrelación sea positiva y por lo tanto, cumpla una de las leyes de Tobler's, "todo esta relacionado con todo lo demás, pero cosas cercanas están mas relacionadas que cosas distantes" [2] [9]

El tipo de autocorrelación que se da en el rendimiento se puede observar a partir del gráfico de dispersión que se obtiene a partir del calculo de la I de Moran, un ejemplo de esto es la figura 3.1, en el cuadrante I se ubican aquellas zonas con alto rendimiento que además están rodeados por observaciones vecinas con rendimiento alto, en el cuadrante III es igual al caso anterior pero con rendimiento bajo; como el rendimiento posee autocorrelación positiva es por ello que las observaciones se ubicarán en los cuadrante I y III. Si se da un caso de autocorrelación negativa, entonces las observaciones se ubicarán en los cuadrantes II y IV. En caso de no darse ningún tipo de autocorrelación las observaciones se distribuirán en todos los cuadrantes. [10]

II - Bajo - Alto	I + Alto - Alto

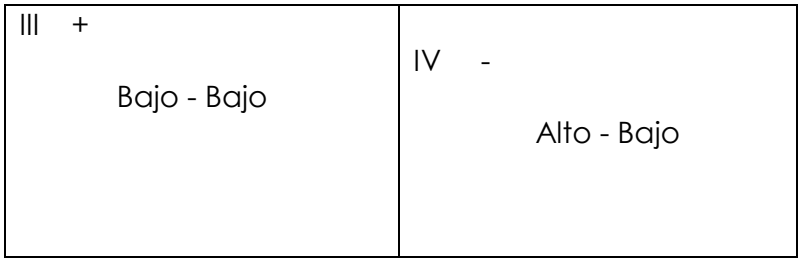


Figura 3.1 Categorías de asociación espacial

La autocorrelación espacial puede surgir por dos razones fundamentales, una de ellas es producto de los errores de medición y la otra es simplemente por fenómenos de interacción espacial. Teniendo en cuenta esto, surge la pregunta, ¿cuáles de estas razones provoca la autocorrelación en el rendimiento? Los fenómenos de interacción espacial son la principal razón, el rendimiento entre otras cosas, depende de las propiedades del suelo, las cuales definen las características de las zonas, y zonas más cercas entre sí tienden a ser más parecidas que otras zonas que se encuentran más separadas. Pero no se debe dejar de lado los errores de medición, mas allá que las herramientas para obtener datos del suelo como los monitores de rendimiento han tenido una gran adopción, éstos no están libres de errores de medición. [2] [6]

Si bien en distintos artículos se menciona que el rendimiento de un cultivo se encuentra correlacionado espacialmente, esta propiedad igualmente debe ser chequeada mediante algún contraste. Moran (1948) y Geary (1954) son quienes realizaron los primeros estudios de contrastes para detectar la autocorrelación espacial, de ellos surgieron dos de los principales índices de contraste, la I de Moran y la C de Geary.

Los contrastes que se analizan en este documento son contrastes a nivel univariante, pues permiten contrastar la presencia de autocorrelación espacial a nivel de una variable, dicho de otra forma, no dicen nada acerca de la posible autocorrelación entre dos o más variables. Estos se pueden dividir en dos grandes grupos, por un lado están los que detectan autocorrelación global y por otro lado los que detectan autocorrelación local. Los contrastes globales se utilizan para detectar autocorrelación a nivel global y así obtener una visión global del fenómeno de interacción espacial, una vez detectada la autocorrelación a nivel global, esta se puede estudiar a nivel local utilizando los contrastes de autocorrelación local, estos contrastes permiten analizar como se da la autocorrelación en distintas zonas del espacio en estudio (ver ref. [24]). [8] [2]

A continuación se presenta los contrastes de I de Moran y de C de Geary, estos son los que se aplicarán en el trabajo, si se quiere ver otros tipos de contrastes de autocorrelación, estos se presentan en el documento Estado del Arte: Métodos de Análisis Espacial (ver ref. [24]).

Contraste I de Moran

El contraste de I de Moran se utiliza para aceptar o rechazar la presencia de autocorrelación espacial en un conjunto de datos tanto a nivel global como local, el que se verá en esta sección se aplica a nivel global, este contraste devuelve un valor numérico que se denota por I, a partir de este valor es que se acepta o rechaza la autocorrelación.

Se elige este contraste para detectar la autocorrelación ya que es uno de los más utilizados en distintos estudios a nivel de análisis espacial y se suele aplicar cuando la muestra es una grilla

regular, los tipos de muestreo se ven en la sección 3.3. El calculo de la I de Moran para chequear la autocorrelación a nivel global se expresa formalmente como:

$$I = \frac{N}{\sum_{i}^{N}\sum_{j}^{N} w_{ij}} \cdot \frac{\sum_{i}^{N}\sum_{j}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i}^{N}(x_i - \bar{x})^2} \qquad i \neq j \qquad [3.1]$$

donde x_i es el valor de la variable en estudio en el punto i (en el caso de estudio es el valor del rendimiento en el punto i de la chacra), $\bar{x}$ es el valor medio de la variable analizada, N es el tamaño de la muestra (en el caso de estudio esta dado por la cantidad de puntos muestrales donde se obtuvo datos mediante el monitor de rendimiento) y finalmente w_{ij} es el elemento i,j de la matriz W de vecindad, que representa el peso de vecindad entre el punto i y el punto j.

La expresión anterior muestra como se obtiene el índice I, pero éste por sí solo no indica nada acerca de la presencia o no de autocorrelación espacial. Para ello se debe realizar una inferencia estadística a partir del índice I, esta se realiza en base al resultado obtenido de I y el teóricamente esperado. La I de Moran estandarizada sigue una distribución asintótica normal cuando el tamaño muestral es suficientemente grande:

$$Z(I) = \frac{I - E(I)}{[V(I)]^{1/2}} \sim N(0,1) \qquad [3.2]$$

Una vez obtenido el valor de *Z(I)* es que se puede saber si los datos cuentan o no con autocorrelación espacial. Teniendo en cuenta que la hipótesis nula de este test es la no existencia de autocorrelación espacial, en base a la siguiente regla se puede determinar el resultado del test:

- Si *Z(I) >0* y significativo, significa que existe autocorrelación espacial positiva
- *Si Z(I)<0* y significativo, significa que existe autocorrelación espacial negativa
- Si *Z(I)* no significativo entonces no se rechaza la hipótesis nula de no autocorrelación espacial. [2] [11]

Para calcular *Z(I)*, se debe tener estimada la esperanza y varianza de I, para el cálculo de éstas hay dos posibilidades de las cuales se puede partir, la primera supone que la distribución de la variable en estudio (rendimiento) sigue una distribución normal; la segunda opción asume que no existe una asunción acerca de la distribución que sigue la variable, esto significa que el valor que tomó la variable en cada punto i muestral, podría haber ocurrido en cualquier otra localización por igual. [11]

Si se tiene en cuenta la primera de las posibilidades la estimación de la esperanza y varianza es como sigue:

$$E(I) = -\frac{1}{N-1} \qquad [3.3]$$

$$Var(I) = \frac{N^2 S_1 - N S_2 + 3 S_0^2}{(N-1)(N+1) S_0^2}$$

[3.4]

donde:

$$S_0 = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} w_{ij}, \quad S_1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (w_{ij} + w_{ji})^2, \quad S_2 = \sum_{k=1}^{N} (\sum_{j=1}^{N} w_{kj} + \sum_{i=1}^{N} w_{ik})^2$$

Por otro lado, si la variable no sigue una distribución normal, entonces el cálculo de la esperanza y varianza es:

$$E(I) = -\frac{1}{N-1}$$

[3.5]

$$Var(I) = \frac{N[(N^2 - 3N + 3) S_1 - N S_2 + 3 S_0^2] - b_2 [(N^2 - N) S_1 - 2N S_2 + 6 S_0^2]}{(N-1)(N-2)(N-3) S_0^2}$$

[3.6]

donde:

$$b_2 = \frac{m_4}{m_2^2}, \qquad m_2 = \frac{\sum_i z_i^2}{N}, \qquad m_4 = \frac{\sum_i z_i^4}{N}, \qquad S_1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (w_{ij} + w_{ji})^2,$$

$$S_2 = \sum_{i=1}^{N} \sum_{j=1}^{N} (w_i + w_i)^2 \quad y \quad w_i = \sum_{j=1}^{N} w_{ij}, \quad [12]$$

Más adelante se verán diferentes contrastes para el chequeo de la distribución normal de la variable.

Contraste C de Geary

Este contraste, al igual que el anterior, es también utilizado para aceptar o rechazar la presencia de autocorrelación espacial en un conjunto de datos tanto a nivel global como local, el mismo devuelve un valor numérico que se denota por la C de Geary, mediante el uso de este valor es que se puede aceptar o rechazar la autocorrelación.

Es un contraste ampliamente usado en la detección de la autocorrelación en distintos estudios a nivel de análisis espacial, al igual que el contraste I de Moran, se suele aplicar cuando la muestra es una grilla regular, los tipos de muestreo se ven en la sección 3.3. El cálculo para chequear la autocorrelación a nivel global se expresa de la siguiente forma:

$$C = \frac{N-1}{2S_0} \cdot \frac{\sum_{i}^{N} \sum_{j}^{N} w_{ij}(x_i - x_j)}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad i \neq j \qquad [3.7]$$

donde x_i es el valor de la variable en estudio en el punto i, $\bar{x}$ es el valor medio de la variable analizada, N es el tamaño de la muestra y w_{ij} es el elemento i,j de la matriz W de vecindad, representa el peso de vecindad entre el punto i y el punto j.

Al igual que en el caso de la I de Moran, la distribución del estadístico tras una estandarización se puede asumir como normal N(0,1):

$$Z(C) = \frac{C - E(C)}{[Var(C)]^{1/2}} \sim N(0,1) \qquad [3.8]$$

$$E(C) = 1$$

$$Var(C) = \frac{(N-1)S_1[N^2 - 3N + 3 - (n-1)b_2] - \frac{1}{4}(N-1)S_2[N^2 + 3N - 6 - (N^2 - N + 2)b_2]}{N(N-2)(N-3)S_0^2} +$$

$$\frac{S_0^2[N^2 - 3 - (N-1)b_2^2]}{N(N-2)(N-3)S_0^2}$$

donde:

$$b_2 = \frac{m_4}{m_2^2}, \qquad m_2 = \frac{\sum_i z_i^2}{N}, \qquad m_4 = \frac{\sum_i z_i^4}{N}, \qquad S_1 = \frac{1}{2}\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (w_{ij} + w_{ji})^2,$$

$$S_2 = \sum_{i=1}^{N} \sum_{j=1}^{N} (w_i + w_i)^2 \quad y \quad w_i = \sum_{j=1}^{N} w_{ij}, \quad [12]$$

Como en el caso del contraste anterior, la hipótesis nula del estadístico C de Geary es la inexistencia de autocorrelación, sin embargo al contrario del test I de Moran, un valor negativo y significativo de la *Z(C)* indicará la existencia de un esquema de dependencia positiva, por el contrario un valor positivo en la *Z(C)* indicará la existencia de autocorrelación negativa.

- Si *Z(C) >0* y significativo, significa que existe autocorrelación espacial negativa
- *Si Z(C)<0* y significativo, significa que existe autocorrelación espacial positiva
- Si *Z(C)* no significativo entonces no se rechaza la hipótesis nula de no autocorrelación espacial. [2] [11]

3.2.2 La heterogeneidad espacial

La heterogeneidad espacial es un efecto que suele surgir cuando se trabaja con datos espaciales, está relacionada con la ausencia de estabilidad espacial en la relación en estudio, esto significa que el fenómeno espacial se comporta de distinta forma sobre el espacio. [2]

La heterogeneidad suele ocurrir especialmente cuando se trabaja con unidades espaciales irregulares, como es el caso de las superficies de una chacra. Debido a que las variables agronómicas y ambientales son espacialmente heterogéneas, es que el rendimiento no escapa de esta propiedad. El manejo de la agricultura depende de una gran cantidad de factores físicos, químicos, fisiológicos tanto de la planta como del suelo y del microclima donde se desarrolla el cultivo. Son tantos los factores que influyen en el rendimiento que hacen que raramente se tenga un suelo que presente uniformidad de nutrientes y agua, es por ello que es imposible pensar que el desarrollo del cultivo pueda ser totalmente homogéneo. [5] [4]

La heterogeneidad espacial en la AP está muy relacionada con la obtención de las zonas de manejo. Son innumerables los factores que influyen en el rendimiento y distintos los factores que pueden influir de forma diferente en distintas zonas de la chacra, bajo esta situación lo ideal es poder encontrar los principales factores que determinan el rendimiento, y a partir de estos obtener las zonas de manejo. [4]

Existe otra razón aparte de la inestabilidad espacial en el comportamiento de una variable que puede causar la heterogeneidad, ésta es la heteroscedasticidad. Se origina por variables que pertenecen al modelo o por variables que fueron omitidas del mismo (ya sea por que no se cuenta con datos muestrales de las mismas o simplemente porque el experto en el tema desconoce la influencia de esta variable en la que se quiere explicar), otra razón es que simplemente este fue mal especificado; la consecuencia de esto es que el término de error posee una varianza no constante, lo que se conoce como heteroscedasticidad.[2]

La posible omisión de variables que influyen en el rendimiento se puede dar perfectamente cuando se quiere modelar un problema relacionado con la agricultura, como se mencionó en párrafos anteriores son tantos los factores que influyen en el rendimiento que hace probable la omisión de algunos de ellos. Básicamente el rendimiento depende de tres grupos de factores: a) factores climáticos, que no pueden ser manejados, b) las características de los suelos y c) los insumos aplicados y la genética del material de siembra. Es muy difícil que no se cuente con datos muestrales sobre los insumos aplicados, pero con los factores climáticos y las características del suelo no ocurren lo mismo, en especial con este último grupo, son tantas las variables que definen las características del suelo y en consecuencia algunas de ellas tienden a ser omitidas. [5]

El tratamiento (aplicación de contrastes, modelaje y estimación) de la heterogeneidad espacial puede ser aplicado mediante las técnicas de Econometría estándar, pero cuando la heterogeneidad se da conjuntamente con la autocorrelación la aplicación de la Econometría estándar no es adecuada debido a que los contrastes para detectar heteroscedasticidad pueden estar sesgados y los modelos que se especifiquen no serán los correctos. En este caso es necesario el uso de las herramientas de la Econometría Espacial. [7]

Como se mencionó en párrafos anteriores, la heterogeneidad se puede dar por la inestabilidad estructural o por la heteroscedasticidad. Existe una batería de contrastes para detectar cualquiera de las dos posibles causas. Como el fenómeno de inestabilidad estructural está relacionado con la obtención de las zonas de manejo, como se explica al inicio de esta sección, y como el tema de detección de zonas de manejo escapa al alcance de este proyecto en particular, es que en este texto se profundiza en aquellos contrastes que detectan la heteroscedasticidad.

Los contrastes se pueden dividir si se sabe que existe o no autocorrelación en los datos. Si no existe autocorrelación en los datos, se pueden aplicar los contrastes que provee la Econometría estándar, en caso contrario, estos ya no son aplicables pues los resultados no serán validos, por ello es necesario recurrir a los contrastes que brinda la Econometría Espacial. [2]

Dentro de la primer rama se encuentra los contrastes de White, de Goldfeld y Quandt, de Breusch y Pagan, de Spearman, entre otros. El primero de ellos es un contraste general, sirve para detectar cualquier forma en que se presente la heteroscedasticidad. Los otros tres se aplican cuando se sospecha la razón por la que se da la heteroscedasticidad, por ejemplo, el de Goldfeld y Quant se utiliza cuando se supone que es producto de una determinada variable (que pertenece o no al modelo). El de Breusch y Pagan se aplica cuando se supone que la heteroscedasticidad es producto de un conjunto de variables omitidas del modelo.[13] [3].

Cuando la autocorrelación espacial se da en los términos de errores producidos por el modelo, la Econometría Espacial nos provee el contraste del límite de Bonferroni, para la detección de la heteroscedasticidad. [2]

A continuación se profundiza con más detalles los diferentes contrastes que se pueden aplicar para la detección de la heteroscedasticidad.

Contraste White

El contraste de White tiene como principal ventaja que permite chequear la heteroscedasticidad sin hacer ningún supuesto previo respecto a la estructura de ésta. Para el contraste de White se deben seguir los siguientes pasos:

a) Como primer paso en este contraste se estiman los errores por Mínimos Cuadrados Ordinarios[7] (MCO) sin tener en cuenta la posible presencia de heteroscedasticidad.
b) Luego de esto se estima mediante MCO una regresión, cuya endógena es el cuadrado de los residuos del paso anterior, y las exógenas son los regresores del modelo original, sus cuadrados y los productos cruzados de segundo orden[8].
c) Cuando el tamaño de la muestra es grande, el producto TR^2 tiende a una distribución chi-cuadrado con p-1 grados de libertad, siendo T el tamaño de la muestra, R^2 el coeficiente de determinación y p el número de regresores estimado en el punto *b*. En base al valor χ^2_{p-1} se rechaza o acepta la presencia de heteroscedasticidad. [3]

Contraste Goldfeld y Quant

Este contraste es utilizado cuando se supone que la heteroscedasticidad se origina por causa de una variable explicativa x_i. En especial este contraste detecta la heteroscedasticidad cuando la dependencia que se da entre la variable x_i y la varianza del término de error σ_i^2 es positiva, esto es que valores grandes de σ_i^2 ocurren en momentos en los que x_i asume valores grandes.

Los pasos del contraste son:

[7] Ver ref. [24].

[8] Si tenemos x_1,x_2,x_3 como variable regresoras solo se tienen en cuenta los cuadrados de cada variable y los productos cruzados de a dos variables x_1x_2 pero no x_1,x_2,x_3.

a) Se ordena de menor a mayor la base de datos de la muestra en función de la variable x_i que se considera produce la heteroscedasticidad.
b) Se omiten *p* observaciones de la mitad de la muestra.
c) Se estima por MCO para las $(T-p)/2$ primeras observaciones y las $(T-p)/2$ finales de la muestra, obteniendo en cada una de ellas la sumatoria del cuadrado de los errores. Notar que las *p* omitidas debe ser lo suficientemente pequeño de modo tal que $(T-p)/2$ sea bastante mayor al número de parámetros del modelo (*T* es el tamaño de la muestra).
d) Obteniendo SR_1 y SR_2 las sumas residuales de las dos regresiones, entonces según el supuesto de homoscedasticidad y normalidad del término de error, el cociente

$$\lambda = \frac{SR_2}{SR_1} = \frac{\sigma_2^2}{\sigma_1^2}$$

sigue una distribución $F_{m,m}$, donde $m = \left({(t-p)}/{2} \right) - k$ (*k* número de parámetros del modelo).

Entonces, si el valor del estadístico λ excede el valor que la tabla de distribución que *F* proporcione, se rechazara la hipótesis nula de ausencia de heteroscedasticidad.[3]

Contraste Breuch y Pagan

Este contraste parte del supuesto de que la heteroscedasticidad se da por causa de un conjunto de variables *Z*. Dicho conjunto tendrá dos restricciones: deberá ser pequeño y solo deberá incluir variables que no estén incluidas en el modelo. Si el contraste llega a dar que la varianza del término de error depende de un conjunto de variables que no fueron incluidas en el modelo puede ser un indicio de que el modelo fue mal especificado y deba especificarse de nuevo. Los pasos del contraste son los siguientes:

a) Se estima el modelo inicial, sobre el que se pretende saber si hay o no heteroscedasticidad, empleando MCO y se obtienen los errores.
b) Se calcula una serie con los errores del modelo anterior al cuadrado estandarizados.

$$\tilde{e}_i^2 = \frac{e_i^2}{\hat{\sigma}^2} \quad \text{con } \hat{\sigma}^2 = \frac{e'e}{n}$$

c) Se estima un modelo de regresión donde $\tilde{e}_i^2$ (calculado en b) es la variable endógena y el conjunto de variables *Z* son las exógenas. A partir de esta estimación se pretende saber si estas variables producen o no heteroscedasticidad en el modelo. Se obtiene R^2 con este modelo y la varianza de la estimada

$$\tilde{e}_i^2 = \alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + ... + \alpha_p z_{pi} + \varepsilon_i$$

$$R_{\tilde{e}}^2$$

d) Se calcula la varianza de la endógena estimada ($S_{\hat{\tilde{e}}^2}^2$), la cual si es muy pequeña, se afirmará que el poder explicativo del conjunto de las variables Z sobre la representación de la varianza de las perturbaciones aleatorias es escaso. Entonces se puede generar un contraste calculado con esta varianza, sabiendo que cuanto más cerca de cero se encuentre, más probabilidades de homoscedasticidad habrá en el modelo. El contraste se basa en que:

$$S_{\hat{\tilde{e}}^2}^2 \Big/ 2$$

se distribuye como una χ_p^2 , por lo tanto si el valor supera al valor de tablas, se rechaza la hipótesis nula; es decir, se acepta que la heteroscedasticidad no es introducida por estas variables. [16]

Contraste Spearman

Al igual que en el contraste de Goldfeld y Quant, aquí también se supone que la heteroscedasticidad se origina por causa de una variable explicativa x_i. Detecta la heteroscedasticidad cuando se da una dependencia positiva entre la variable x_i y la varianza del término de error σ_i^2.

Para el contraste de Spearman se siguen los siguientes pasos:

a) Se ordena de menor a mayor tanto la variable sospechosa x_{ij} como el valor absoluto del residuo $|e_i|$.

b) El cambio de puesto en ambas, para ambas observaciones, deberá ser el mismo número de puestos respecto al orden original de las series. En caso que el cambio de puesto respecto al original no sea el mismo para las dos observaciones ya ordenadas se obtiene el grado de correlación en ese cambio de puesto respecto al inicial de cada una de las variables con la siguiente expresión.

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

c) En base a este *r* calculado en el paso b) se obtiene el siguiente ratio, con una función de distribución conocida bajo hipótesis nula de no significancia

$$\frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \rightarrow t_{n-2}$$

Si el resultado del ratio es superior al valor en la tabla *t* se afirma que la correlación es significativa y que hay indicios de heteroscedasticidad en el modelo provocado por la variable x_{ij}. [3][16]

Contraste Limites de Bonferroni

Este método es utilizado para detectar la heteroscedasticidad como alternativa espacial cuando hay dependencia espacial en los residuos. Los pasos de este contraste son los siguientes:

a) Se realiza una contrastación de la presencia conjunta de ambos efectos espaciales (autocorrelación y heteroscedasticidad) $H_0 : [\beta, \beta']' = 0$ a través de una adaptación del test de multiplicadores de Lagrange a un modelo de regresión con perturbaciones espaciales autorregresivos y heteroscedásticas:

$$LM = \frac{1}{2} f'Z(Z')^{-1} Z'f + \frac{1}{T}\frac{e'We}{s^2} \sim \chi_{p+1}^2$$

$$f_i = \left(s^{-1}e_i\right)^2 - 1 \quad T = tr[W'W + W^2]$$

Siendo *e* el vector de residuos MCO del modelo; s^2 la varianza MV y *Z* una matriz $N \times (p+1)$ formada por un término constante y las variables causantes de la heteroscedasticidad.

b) En caso de ser rechazada la hipótesis nula se deberá contrastar individualmente la presencia de heteroscedasticidad (a través de contrastes de econometría tradicional) y de dependencia espacial residual.[2]

3.3 *Definición de la matriz de pesos W*

La matriz de pesos W se encarga de capturar las relaciones multidireccionales en un contexto espacial. En las secciones anteriores se vio que esta matriz se aplica para realizar los contrastes de autocorrelación, y como se verá en la sección 3.4 también se utiliza para definir los modelos de regresión espacial en presencia de autocorrelación.

La matriz W es una matriz cuadrada no estocástica, donde cada elemento w_{ij} de la misma define la relación de vecindad que existe entre dos regiones i y j. La matriz asigna unos o ceros a sus elementos w_{ij} dependiendo si las regiones i y j son vecinas o no, según el criterio de vecindad que se defina para la matriz. Cada fila de la matriz contiene elementos distintos de cero en aquellas columnas correspondientes a las regiones vecinas de la región que representa la fila. La matriz tiene como característica que es positiva simétrica, w_{ij} =1 si i y j son vecinos, w_{ij} =0 en caso contrario y por definición w_{ii} = 0 (un elemento no es vecino de si mismo). [2]

$$W = \begin{bmatrix} 0 & w_{12} & w_{13} & \cdots & w_{1N} \\ w_{21} & 0 & w_{23} & \cdots & w_{2N} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ w_{N1} & w_{N2} & w_{N3} & & 0 \end{bmatrix}$$

Como se cita en el párrafo anterior el criterio de vecindad que se tome para definir la matriz es variado y depende de cómo es la distribución de los datos sobre el espacio. Moran (1948) utiliza el concepto de contigüidad física de primer orden para definir la matriz, el cual define la vecindad entre dos regiones si son físicamente adyacentes. Para el caso de la contigüidad física de primer orden existen distintos criterios de adyacencia que se pude elegir, los mas conocidos son Rook, Bishop y Queen, la figura 3.2 ilustra cada uno de ellos.

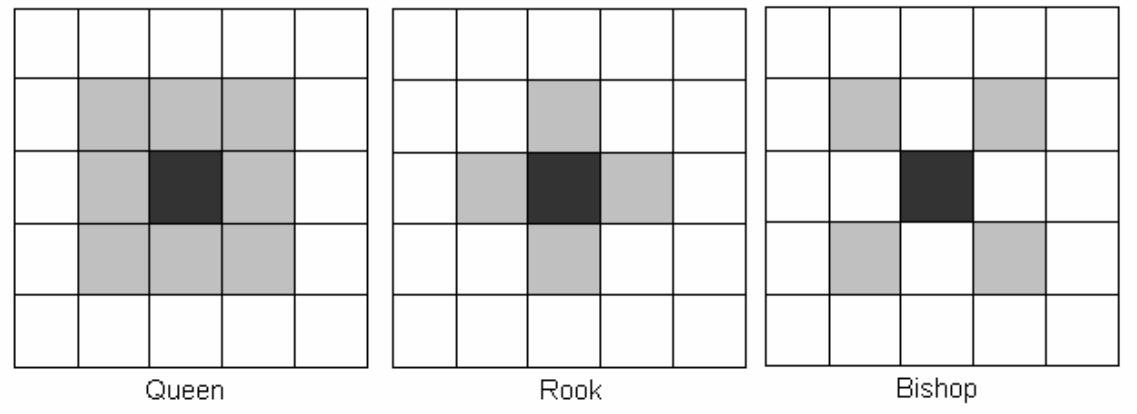


Figura 3.2 Criterios de adyacencia de primer orden

Los anteriores son criterios de primer orden, pero eventualmente el orden puede aumentar hasta donde sea necesario, la siguiente figura explica un criterio de adyacencia de segundo orden. [1]

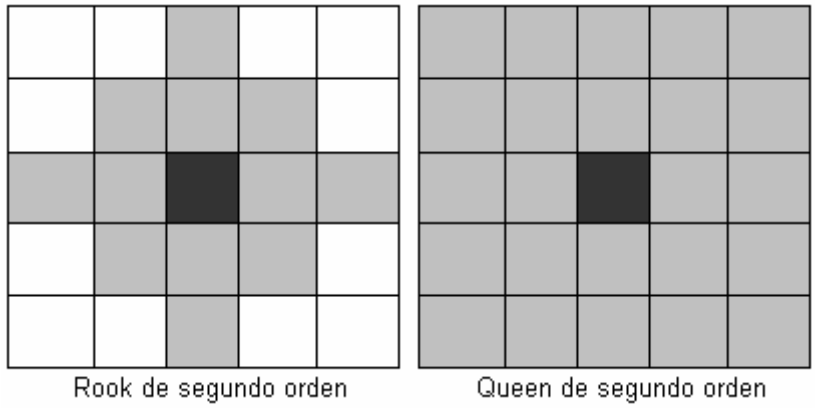
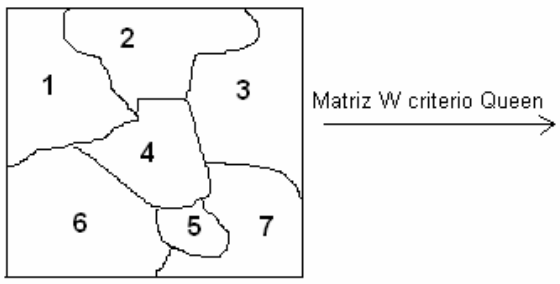


Figura 3.3 Criterios de adyacencia de segundo orden

En la figura 3.4 se pueden observar 7 regiones contiguas y como queda definida matriz W 7x7 de primer orden siguiendo el criterio Queen. [2]



Matriz W criterio Queen

	W_{i1}	W_{i2}	W_{i3}	W_{i4}	W_{i5}	W_{i6}	W_{i7}
W_{1j}	0	1	0	1	0	1	0
W_{2j}	1	0	1	1	0	0	0
W_{3j}	0	1	0	1	0	0	1
W_{4j}	1	1	1	0	1	1	1
W_{5j}	0	0	0	1	0	1	1
W_{6j}	1	0	0	1	1	0	1
W_{7j}	0	0	1	1	1	0	0

Figura 3.4 Matriz W de Queen en un caso de 7 regiones contiguas

Cuando la malla de datos no se distribuye de forma regular se suele utilizar otro criterio para la definición de la matriz. Ckiff y Ord (1973, 1981) definen un criterio que se basa en la distancia entre dos regiones *i* y *j* de la siguiente forma:

$$w_{ij} = d_{ij}^{-a} \beta^{b}_{ij} \qquad [3.9]$$

donde d_{ij} es la distancia entre las regiones i y j, β_{ij} es la longitud relativa de la frontera entre las regiones i y j con respecto al perímetro de i, a y b son parámetros a estimar, se puede reducir la complejidad asumiendo estos parámetros como dados. El anterior no es el único criterio que se basa en la distancia, Dacey (1968) define un criterio similar y Anselin (1980) define una matriz inversa de distancias al cuadrado. [2]

El tipo de criterio que se elija ya sea basado en la contigüidad física o basado en la distancia, depende directamente de cómo se encuentra distribuidos los datos en el espacio. En la agricultura los tipos de muestreos de suelo más conocidos son:

- Muestreo al azar simple y estratificado: La desventaja del muestreo al azar simple es el posible desbalance en la distribución de los puntos de muestreo, en consecuencia los puntos no quedan regularmente distribuidos. Por lo tanto para la definición de la matriz es necesario aplicar el criterio de distancia. El muestreo al azar estratificado mejora el problema de distribución pero no lo elimina, por lo tanto sigue siendo necesario aplicar un criterio de distancia.
- Muestreo en grilla sistemático: Es el tipo más intensivo de muestreo. En él, las muestras son tomadas a intervalos regulares en todas las direcciones, analizándose por separado. En este caso conviene elegir un criterio de contigüidad física, ya que los puntos se encuentran a distancias regulares.
- Muestreo sistemático estratificado desalineado: En este tipo de muestreo se divide la chacra en celdas uniformes y se toma un punto al azar dentro de cada una de las celdas. Si bien la distancia entre los puntos de muestreo no es regular, sí lo es el tamaño de las celdas y por lo tanto al igual que el muestreo en grilla sistemático conviene elegir un criterio de contigüidad física.

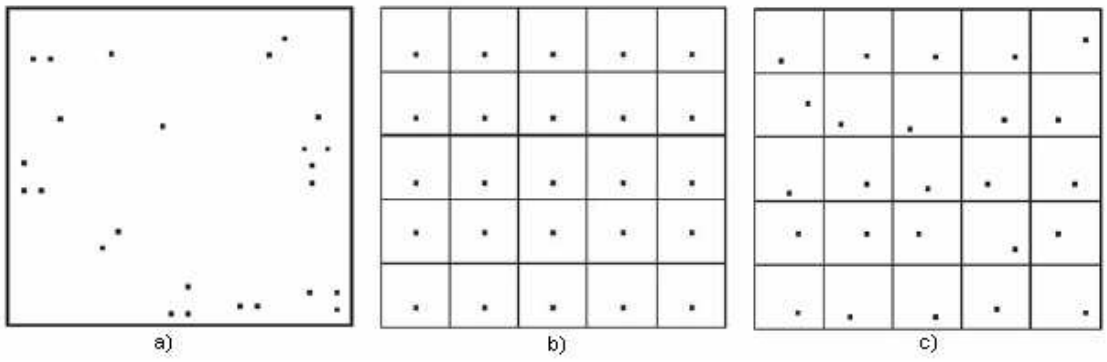


Figura 3.5 Tipos de muestreos: a) al azar simple b) en grilla sistemático c) sistemático estratificado desalineado [4]

La matriz W es una matriz cuadrada no estocástica, donde cada elemento w_{ij} de la misma define la relación de vecindad que existe entre dos regiones i y j. La matriz asigna unos o ceros a sus elementos w_{ij} dependiendo si las regiones i y j son vecinas o no, según el criterio de vecindad que se defina para la matriz.

Los elementos w_{ij} de la matriz W tienen asignados valores 0 o 1 dependiendo si las regiones *i* y *j* son vecinas o no, esta forma de asignar valores a los elementos tiene una desventaja, y es que la matriz queda simétrica, esto hace que no pueda reflejar influencias no recíprocas (la región *i* influye sobre la *j,* pero no al revés). Por ello se suele estandarizar las filas de la matriz, esto significa dividir el peso que recibe un elemento de parte de cada uno de sus vecinos entre la cantidad de vecinos que posee:

$$w_{ij} = \frac{1}{\sum_j w_{ij}}$$

De esta forma la sumatoria de los pesos de los vecinos para cada punto da uno independientemente de la cantidad de vecinos que posea. La estandarización de la matriz no siempre es adecuada, cuando la matriz se define utilizando el criterio de distancia, la estandarización de la misma carece de significado. En la figura 3.7 se puede observar como queda la matriz estandariza para el ejemplo de la figura 3.4.

	W_{i1}	W_{i2}	W_{i3}	W_{i4}	W_{i5}	W_{i6}	W_{i7}
W_{1j}	0	1/3	0	1/3	0	1/3	0
W_{2j}	1/3	0	1/3	1/3	0	0	0
W_{3j}	0	1/3	0	1/3	0	0	1/3
W_{4j}	1/6	1/6	1/6	0	1/6	1/6	1/6
W_{5j}	0	0	0	1/3	0	1/3	1/3
W_{6j}	1/5	0	0	1/5	1/5	0	1/5
W_{7j}	0	0	1/3	1/3	1/3	0	0

Figura 3.6 Matriz estandarizada

3.4 *Modelos de regresión espacial*

La idea que existe detrás de un modelo de regresión es poder relacionar el comportamiento de una variable endógena en base a un conjunto de variables exógenas. En nuestro caso de estudio el modelo representa la respuesta del rendimiento de un cultivo frente a los distintos factores que influyen en éste. En el caso de estudio, el rendimiento del cultivo es la variable endógena del modelo y los distintos factores son conjunto de variables exógenas, que se pueden dividir en dos grupos: los factores climáticos y las características del suelo.

Existen diferentes formas de relacionar la variable endógena con las variables exógenas, esta relación puede llegar a ser lineal o no lineal, teniendo en cuenta diferentes parámetros espaciales que determinen la relación. Algunos estudios indican que el rendimiento de un cultivo se encuentra correlacionado espacialmente, como además las técnicas para modelar que provee la Econometría estándar (modelos de regresión lineal) no tienen en cuenta el efecto de la

autocorrelación a nivel espacial, es necesario recurrir a los modelos de regresión espacial, los cuales pertenecen a la Econometría Espacial.

Los modelos de regresión lineal expresan la relación ente la variable endógena y las variables exógenas mediante una relación lineal, la relación se expresa mediante un vector β de k parámetros que deben ser calculados, este vector es el que define la relación de dependencia entre la variable endógena y las variables exógenas. Formalmente el modelo de regresión lineal múltiple se expresa:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \mu_i \qquad [3.10]$$

donde y_i representa el valor de la variable endógena en la observación i, x_{1i} representa el valor de la variable x_1 en la observación i, los β_j son los parámetros a estimar que representan el impacto producido por cada una de las variables exógenas y μ_i es el error cometido por el modelo en la observación i. Este modelo se puede expresar en forma matricial:

$$y = X\beta + \mu \qquad [3.11]$$

donde, y es un vector de observaciones de la variable dependiente, X es una matriz de observaciones de las variables exógenas, β es un vector con los parámetros del modelo y μ es un vector con los errores cometidos por el modelo. [3]

Los modelos regresivos espaciales incorporan la autocorrelación relacionando el valor de una variable en un punto del espacio con el valor de la misma variable en otros puntos del espacio, como se verá esto se logra mediante el uso de la matriz de vecindad W que permite relacionar los valores de una variable con varios puntos en múltiples direcciones y un operador de retardo ρ el cual indica la fuerza de la relación (que tan autocorrelacionada se encuentra la variable). Formalmente el modelo de regresión espacial es un modelo de regresión lineal con la incorporación de la matriz W y el operador de retardo ρ. [2]

Los modelos regresivos espaciales varían dependiendo si el efecto es la autocorrelación espacial o la heterogeneidad espacial. Como se verá a continuación los modelos en presencia de autocorrelación varían si la misma se detecta en la variable dependiente, el término de error del modelo o ambos. Para el caso de heterogeneidad espacial los modelos regresivos también varían si esta se presenta como heteroscedasticidad o como inestabilidad espacial.

3.4.1 Modelos regresivos de autocorrelación espacial

La autocorrelación espacial se incorpora al modelo de regresión de dos formas básicas. En un modelo, la autocorrelación se limita al término de error cometido por el modelo de regresión, el llamado modelo de error espacial. En el otro caso, la autocorrelación espacial se refiere a la variable dependiente del modelo en si misma y se denomina modelo de ponderación espacial. A continuación se verá más detalladamente cada uno de ellos. [1]

Para tener en cuenta la autocorrelación espacial, al modelo de regresión clásico se le debe agregar la matriz de vecindad W y el operador de retardo ρ. Esta matriz permite incorporar las distintas relaciones espaciales que se presentan en la variable que se detectó la autocorrelación,

y ponderar dichas relaciones. La matriz de vecindad que se utiliza en el modelo es la misma que se utiliza en los contrastes de autocorrelación espacial (ver sección 3.3).

Formalmente y usando la notación dada por Anselin (1988), el modelo de ponderación espacial (en la cual la variable endógena se encuentra autocorrelacionada espacialmente) se expresa formalmente como:

$$y = \rho Wy + X\beta + \mu \qquad [3.12]$$

donde, y es un vector ($N \times 1$) de observaciones de la variable dependiente, X es una matriz ($N \times K$) de observaciones de las variables exógenas, ρ es el operador de retardo (coeficiente autorregresivos), Wy es el vector de variables exógenas ponderado por la matriz de vecindad W, μ es un vector con los errores producidos por el modelo, estos errores siguen una distribución $\mu \sim N(0, \sigma^2 I)$. [2]

Este modelo expresa que el valor que toma la variable y en un punto del espacio, no solo depende de los valores de las variables exógenas X en el mismo punto, si no que además depende de los valores vecinos de la propia variable y, los cuales se expresan mediante el término Wy. El parámetro ρ en esta ecuación expresa el grado de interdependencia entre las observaciones de la variable y, si ρ valiera cero, entonces indica que no existe autocorrelación y el modelo regresivo de autocorrelación espacial se transforma en un simple modelo de regresión lineal. El valor que puede tomar ρ varia entre $1/w_{\min} < \rho < 1/w_{\max}$, dónde $w_{\min}$ y $w_{\max}$ son los valores propios menor y mayor de la matriz W, en caso que esta se encuentre estandarizada, se cumple que $w_{\max}$ vale uno y $w_{\min} \leq -1$. [2]

En caso que la variable endógena se encuentre correlacionada y no se incluya en el modelo el retardo ρ para dicha variable, entonces la autocorrelación se traslada directamente al término de error, este tipo de autocorrelación se conoce como autocorrelación espacial sustantiva y se corrige mediante la inclusión del retardo espacial en la variable endógena como se expresa en el modelo anterior. Pero la anterior no es la única causa que pude producir la autocorrelación en el término de error, si por alguna razón se excluyeron del modelo variables que se suponen que no influyen en la variable endógena y éstas se hallan correlacionadas espacialmente, o se cometieron errores de medida durante la extracción de los datos, entonces puede ocurrir que los errores se encuentren correlacionados, esto se conoce como autocorrelación espacial residual. Formalmente y usando la notación dada por Anselin (1988), el modelo de error espacial se expresar como:

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &= \lambda W\varepsilon + \mu \end{aligned} \qquad [3.13]$$

donde, y es un vector ($N \times 1$) de observaciones de la variable dependiente, X es una matriz ($N \times K$) de observaciones de las variables exógenas, λ es el coeficiente autorregresivo, ε es el vector de errores de muestreo aleatorio que sigue una especificación autorregresiva de muestreo espacial, $W\varepsilon$ es el vector de errores ponderado por una matriz W de N observaciones vecinas y μ son los nuevos errores del modelo, estos errores siguen una distribución $\mu \sim N(0, \sigma^2 I)$.

En el anterior modelo, el vector de errores ε se expresa como la suma de errores espaciales ($W\varepsilon$) y los nuevos errores μ. Donde el primer término expresa una suma ponderada de los errores en las localidades vecinas. La selección de los vecinos se lleva a cabo a través de la matriz de vecindad *W*. [2]

Existen otras variantes de modelos regresivos que incluyen retardos espaciales en distintas variables exógenas, cuando se conoce cuales son las variables exógenas que se encuentran correlacionadas, y distintas combinaciones de modelos. Además los modelos aquí presentados son de primer orden, pero es posible especificar modelos de segundo orden de ser necesario, mediante la incorporación de más retardos espaciales y matrices de ponderación. Todas la variantes anteriores no son tratadas debido a que los modelos más comunes y para las cuales la investigación realizada permitió encontrar una solución a la forma de estimar el modelo son los planteados anteriormente. No obstante si el lector pretende profundizar en otras variantes de modelos puede recurrir al estudio realizado por Moreno Serrano y Vayá Valcarce (2000) (ver referencia [2])

3.4.2 Modelos regresivos de heterogeneidad

La heterogeneidad generalmente puede ser modelada mediante técnicas provistas por la econometría estándar. Cuando el fenómeno en estudio posee una inestabilidad estructural a través del espacio, entonces los parámetros estimados del modelo varían según la estructura espacial. Si la heterogeneidad se presenta en forma de heteroscedasticidad en los residuos del modelo se puede aplicar el modelo de regresión lineal clásico combinado con técnicas de estimación como Mínimos Cuadrados Generalizados. Pero existe una razón para aplicar técnicas de econometría espacial, es cuando la heteroscedasticidad se presenta conjuntamente con la autocorrelación espacial. A continuación se profundizará en cada unos de estos modelos.

El modelo de error heteroscedástico es el más simple de todos, en este caso la heteroscedasticidad se encuentra en los errores cometidos por la regresión, en consecuencia la varianza de los errores es diferente en cada una de las observaciones muestrales *i*, $Var(\mu_i) = \sigma_i^2$. Las razones más habituales que provocan este tipo de heteroscedasticidad en los modelos de regresión son la omisión de variables importantes o simplemente por que éste fue mal especificado. El modelo se expresa formalmente como:

$$\begin{aligned} y &= X\beta + \mu \\ Var(\mu_i) &= \sigma_i^2 I \\ E(\mu\mu') &= \Omega \end{aligned} \qquad [3.14]$$

donde y es un vector ($N \times 1$) de observaciones de la variable dependiente, X es una matriz ($N \times K$) de observaciones de las variables exógenas, μ es un vector ($N \times 1$) de errores aleatorios con varianza no constante para cada observación *i*, y Ω es una matriz de varianzas y covarianzas de μ.

El anterior no es el único modelo de heteroscedasticidad que se puede especificar, si se conoce la forma en que la misma se presenta, el modelo se puede formular con mayor especificidad. Una posibilidad es el modelo de heteroscedasticidad aditiva, que se utiliza cuando se produce por un conjunto de variables explicativas que pueden o no pertenecer al modelo, en este caso la

varianza del término de error se expresa como una función lineal de las variables anteriormente mencionadas, con γ como vector de coeficientes de la función y Z como una matriz $N \times P$ de variables heteroscedásticas, donde N es el tamaño de la muestra y P es la cantidad de variables. [7] [2]

$$Var(\mu_i) = Z \ \gamma \qquad [3.15]$$

Si la heterogeneidad se produce por inestabilidad estructural en el espacio, los parámetros del modelo pueden variar en el espacio conforme varía la estructura según la localización geográfica. Un accidente geográfico como un curso de agua, puede hacer variar el rendimiento de un cultivo, y este se puede comportar de distinta manera según se encuentre cercano o no al curso de agua, la especificación correcta de esta realidad hace que sea necesario especificar un modelo con determinados coeficientes en las zonas de la chacra cercanas al curso de agua y otro en las zonas alejadas, recordar en este punto el concepto de zonas de manejo. [7]

Por lo tanto no parece razonable explicar mediante un único modelo las distintas relaciones que se dan en las distintas zonas espaciales. En estos casos lo ideal es estimar distintos parámetros del modelo dependiendo de las características de la zona que se está modelando. Existen distintas metodologías que permiten especificar la inestabilidad estructural: Variación de coeficientes aleatorios, Switching regressions, Expansión espacial y Regresiones ponderadas geográficamente. Si el lector quiere profundizar en estas metodologías puede hacerlo desde el estudio realizado por Moreno Serrano y Vayá Valcarce (2000). [2]

3.4.3 Modelos regresivos mixtos

Eventualmente la autocorrelación espacial se puede presentar conjuntamente con la heterogeneidad espacial, pues las causas que provocan este último efecto espacial pueden además provocar la aparición de autocorrelación espacial, una de ellas pude ser los errores de medida o de especificación. Por lo tanto se hace necesario especificar un modelo que contemple ambos efectos espaciales, de la siguiente forma:

$$\begin{aligned} y &= \rho W_1 y + X\beta + \varepsilon \\ \varepsilon &= \lambda W_2 \varepsilon + \mu \\ \Omega_{ii} &= h_i(Z\alpha); h_i > 0 \end{aligned} \qquad [3.16]$$

donde, y es un vector ($N \times 1$) de observaciones de la variable dependiente, X es una matriz ($N \times K$) de observaciones de las variables exógenas, λ y ρ son los coeficientes autorregresivos, ε es el vector de errores de muestreo aleatorio que sigue una especificación autorregresiva de muestreo espacial , $W_2\varepsilon$ es el vector de errores ponderado por una matriz W de N observaciones vecinas, μ son los nuevos errores del modelo, estos siguen una distribución $\mu \sim N(0, \Omega)$, Ω es una matriz diagonal de varianzas y covarianzas de los términos de error μ, esta matriz es una función de $P+1$ variables exógenas de Z y $h_i(Z\alpha)$ es una función de $P+1$ variables, donde P es la cantidad de variables exógenas que generan la heteroscedasticidad, el uno es por el término constante del modelo, esta función define los elementos de la diagonal de la matriz Ω, α es un vector de P coeficientes asociados a las variables exógenas. Si $\alpha = 0$ entonces

$h_i = \sigma^2$ (esto significa que la varianza del término de error vuelve a ser constante y en consecuencia el modelo deja de ser heteroscedastico). [1]

En este modelo la cantidad de parámetros a estimar es *3+K+P*, para ser mas específico los parámetros son: los coeficientes autorregresivos λ y ρ; la varianza del término de error σ^2, los K coeficientes β asociados a las variables exógenas y por último los P coeficientes α asociados a las variables exógenas que generan la heteroscedasticidad. [1]

Se puede variar el modelo regresivo haciendo valer cero algunos de los parámetros a estimar, de esta forma:

- Si $\rho = 0, \lambda = 0, \alpha = 0$ el modelo que se obtiene es el clásico modelo de regresión lineal $y = X\beta + \varepsilon$, ya que se elimina la especificación de la autocorrelación y heteroscedasticidad.
- Si $\rho = 0, \alpha = 0$ se obtiene el modelo de error espacial.
- Si $\lambda = 0, \alpha = 0$ se obtiene el modelo de ponderación espacial.
- Si $\alpha = 0$ se obtiene el modelo de error y ponderación espacial. [1] [7]

3.4.4 Selección de un modelo espacial

Los métodos que existen para la selección de un modelo espacial son variados y no existe un consenso acerca de cual es el método adecuado, éste depende del caso que se está estudiando. Los dos métodos principales son propuestos por Anselin y Florax (1995) y por Folmer y Florax (1992), ambos métodos permiten elegir un determinado modelo de regresión en presencia de autocorrelación, vistos en la Sección 3.4.1. Es importante destacar que los contrastes vistos en la sección 3.23.2.1, si bien permiten contrastar la presencia de autocorrelación espacial en los datos, a partir del resultados de éstos no es posible escoger el modelo, para ello existen los métodos propuestos por los autores citados anteriormente, ambos métodos utilizan los contrastes de Multiplicadores de Lagrange (LM)[9] para determinar que tipo de autocorrelación modelar. [2][15]

Unos de los métodos para seleccionar un modelo fue definido por Anselin y Florax (1995), los cuales proponen una forma de elegir el modelo adecuado cuando se dan ambos tipos de autocorrelación espacial (autocorrelación en la variable endógena o en el error). Ya que el modelo mixto es de compleja solución, esta complejidad se reduce cuando se imponen determinadas restricciones en el modelo de regresión, como es el caso del modelo de ponderación espacial o el modelo de error espacial. El criterio que propone para elegir el modelo se basa en aplicar los estadísticos de LM para ver con qué intensidad se da la dependencia espacial tanto en la variable endógena como en el término de error y dependiendo de cual de los estadísticos sea más significativo será el modelo que se elija. Sea LM_λ y LM_ρ, los estadísticos para el término de error y la variable endógena respectivamente, si LM_λ es más significativo que LM_ρ entonces el modelo se define solo teniendo en cuenta la autocorrelación en el término de error, de suceder lo contrario el modelo se define solo teniendo en cuenta la

[9] Ver ref. [24] o referencia [2], en este último se los nombra como LM-ERR y LM-LAG

autocorrelación en la variable endógena. Teniendo en cuenta el criterio anterior Florax (2003) define los siguientes pasos para la elección del modelo adecuado:

1) Se estima el modelo de regresión lineal mediante MCO y se calculan los errores producidos por el modelo.
2) Se contrasta la hipótesis de no autocorrelación espacial tanto en la variable endógena como en los errores cometidos por la estimación MCO, utilizando respectivamente los contrastes LM_λ y LM_ρ. El modelo elegido está dado por alguno de los cuatros siguientes pasos dependiendo del resultado de los contrastes.

3.a) Si no se detecta ningún tipo de autocorrelación espacial, entonces la elección es el clásico modelo de regresión lineal, estimado mediante MCO (ya calculado en el paso 1).

3.b) Si se detecta autocorrelación espacial en la variable endógena como en el término de error, entonces el modelo correcto corresponde al estadístico más significativo.

3.c) Si LM_λ es significativo y LM_ρ no, entonces el modelo correcto es el que incorpora la autocorrelación en el término de error.

3.d) Si LM_ρ es significativo y LM_λ no, entonces el modelo correcto es el que incorpora la autocorrelación en la variable endógena. [15]

El segundo método propuesto por Folmer y Florax (1992) se basa en el método de expansión espacial de variables, estos autores proponen tres métodos, pero dentro de la literatura analizada solo se enfatiza en el método denominado Expansión Espacial de Variables 2 (EEV2). Este método pretende corregir la autocorrelación espacial mediante la inclusión al modelo de aquellas variables que no fueron tenidas en cuenta y que además se encuentran autocorrelacionadas espacialmente, en consecuencia también se debe ingresar al modelo un retardo espacial correspondiente a las variables ingresadas. El primer inconveniente que surge de este método es que la elección de las variables no se realiza de forma mecánica, por el contrario la elección de estas variables tiene que tener un sentido teórico por el cual se le incluye un retardo espacial. El conjunto de pasos para la elección del modelo adecuado es:

1) Se estima el modelo de regresión lineal mediante MCO y se calculan los errores producidos por el modelo.
2) Se selecciona un conjunto S de variables que no pertenecen al modelo, para las cuales la inclusión de un retardo espacial tiene un sentido teórico.
3) Se contrasta la hipótesis de no autocorrelación espacial tanto en la variable endógena como en los errores cometidos por la estimación MCO, utilizando respectivamente los contrastes LM_λ y LM_ρ. El modelo elegido está dado por alguno de los cuatro siguientes pasos dependiendo del resultado de los contrastes.

4.a) Si ninguno de los dos contrastes (LM_λ y LM_ρ) rechazan sus respectivas hipótesis nulas, entonces se agregan al modelo el conjunto de variables S y el retardo espacial para dichas variables.

4.b) Si solo el contraste LM_λ es significativo o los dos test LM_λ y LM_ρ rechazan sus respectivas hipótesis nulas pero la probabilidad del LM_λ es inferior a la del LM_ρ, entonces el modelo correcto es el que incorpora la autocorrelación en los errores.

4.c) Si solo el contraste LM_ρ es significativo o los dos contrastes LM_λ y LM_ρ rechazan sus respectivas hipótesis nulas pero la probabilidad del LM_ρ es inferior a la del LM_λ , entonces el modelo correcto es el que incorpora la autocorrelación en la variable dependiente. Además se debe ir agregando las variables seleccionadas en el paso 2, contrastando la significación del modelo. [2]

Adicionalmente y debido a ciertas desventajas señaladas acerca del método EEV2, los autores recomiendan otra estrategia, la llamada metodología Hendry, no se profundizará en esta metodología, pero sí se indica que tras unos ejercicios de simulación realizados por Florax, no encontró una superioridad clara de una metodología frente a la otra a la hora de seleccionar el modelo correcto. Pero la conclusión a la que se llegó es que cuando la autocorrelación se presenta en algunas de las variables exógenas la metodología Hendry se mostró superior a la EEV2. Pero cuando la autocorrelación se presenta en la variable endógena o el término de error sucede todo lo contrario. [2]

Teniendo en cuenta las tres estrategias mencionadas anteriormente para la elección del modelo espacial adecuado es importante mencionar que ninguna de ellas se refiere a la posible presencia de heteroscedasticidad y la literatura estudiada no menciona que estrategia seguir para la elección del modelo adecuado cuando la autocorrelación se da en conjunto con la heteroscedasticidad, mas allá que Luc Anselin (1988) propone un modelo mixto con presencia de autocorrelación con heteroscedasticidad, el cual se especificó en la sección 3.4.3.

3.5 *Estimación del modelo de regresión*

Estimar un modelo implica asignar valores numéricos a los parámetros desconocidos que componen el modelo, utilizando para esto información muestral de las variables del modelo. Existen distintos estimadores que se pueden elegir, la elección depende de la característica de los datos. Los estimadores se diferencian unos de otros en el modo de resumir los datos para asignar valores numéricos a los parámetros del modelo. [3]

Dentro de la econometría clásica el estimador más conocido es el Mínimos Cuadrados Ordinarios MCO, si los datos poseen heteroscedasticidad se puede utilizar una variante del estimador MCO, que es el estimador Mínimos Cuadrados Ponderados (MCP). Sin embargo en presencia de autocorrelación espacial la estimación del modelo mediante MCO o MCP no obtiene buenos resultados (ver ref. [24]).

Las consecuencias de aplicar el estimador MCO en presencia de autocorrelación espacial son distintas dependiendo del tipo de autocorrelación (autocorrelación espacial en la variable endógena o en el término de error). Si la autocorrelación se produce en el término de error la estimación de los parámetros será insesgada pero no será eficiente pues la varianza de los residuos si será sesgada. Este sesgo tiene otras consecuencias importantes, el resultado de los test de significación de la t-student[10] y el coeficiente de determinación R^2 (ver sección 3.6) también estarán sesgados, con lo cual no se podrá saber con precisión si los resultados son buenos. Si la autocorrelación se produce en la variable endógena la estimación de los parámetros mediante MCO darán sesgadas e inconsistentes. [2]

Ante la presencia de autocorrelación espacial en los datos, Anselin sugiere el uso del método de Máxima Verosimilitud (MV) para estimar el modelo espacial, pero este no es el único método que existe, el Método Generalizado de los Momentos (MGM), (ver ref. [24]), es otra

[10] El test t de Student es utilizado para comparar dos grupos independientes de observaciones con respecto a una variable numérica con el objetivo de determinar si existen diferencias significativas en la variable en estudio entre los dos grupos.

metodología que puede ser utilizada para la estimación en presencia de autocorrelación espacial. Si bien este estimador es más simple y computacionalmente menos costoso que el estimador de MV, tiene una clara desventaja, la matriz de covarianza de los errores debe ser estacionaria[11] (son pocos los procesos espaciales que cumplen esta propiedad), esta restricción limita el uso del estimador MGM a modelos donde los errores cometidos por estos no sean heteroscedasticos. Además de esto, si los datos cumplen con dicha propiedad solamente permite estimar modelos con autocorrelación a nivel de las variables exógenas con el error, pero no autocorrelación en la variable endógena (rendimiento), por tanto solo puede ser aplicado para el modelo de error espacial. [2] [14]

Por lo tanto teniendo en cuenta el amplio uso del método de MV en distintos proyectos de Econometría Espacial, por ser el estimador más flexible, ya que permite estimar modelos que especifican de distinta forma la autocorrelación espacial, es que se profundiza el estudio en este método de estimación.

El fundamento teórico y matemático que hay detrás del método de MV aplicado a modelos de regresión espacial es complejo, por ello en este documento solo se presenta los procedimientos (de forme breve) de estimación para los modelos de ponderación espacial y error espacial, en caso de que el lector quiera profundizar en este tema, se recomienda la lectura del libro, Spatial Econometrics:Methods and Models, de Luc Anselin (1988), si se quiere una análisis menos profundo y en español, se recomienda el libro, Técnicas econométricas para el tratamiento de datos espaciales: La econometría espacial, de Moreno Serrano y Vaya Valcarce (2000).

El primer caso que se analiza, es un modelo que tiene en cuenta la autocorrelación espacial solo en la variable endógena, y además los errores producidos por el modelo no poseen heteroscedasticidad, o sea el modelo de ponderación espacial. Para este caso la función de verosimilitud es:

$$LnL_C = -\frac{N}{2}\ln 2 + \ln\left|I - \rho W\right| - \frac{N}{2}\ln\left|\frac{(e_{0-}\rho e_L)'(e_{0-}\rho e_L)}{N}\right| \qquad [3.17]$$

donde, e_0 y e_L son, respectivamente, los errores cometidos mediante la estimación por MCO de las regresiones lineales de y y Wy sobre X. Como todos los elementos de la función dependen del parámetro autorregresivo ρ, la estimación MV de dicho parámetro se hace a través de una optimización numérica de la función de verosimilitud, según el siguiente procedimiento:

1) Estimación mediante MCO de la regresión $y = X\beta_0 + u$, se obtiene b_0 como la estimación de los parámetros β_0.
2) Estimación mediante MCO de la regresión $Wy = X\beta_L + u$, se obtiene b_L como la estimación de los parámetros β_L.
3) Se obtiene los residuos e_0 y e_L de los pasos 1) y 2) respectivamente.
4) Tras la obtención de los residuos e_0 y e_L se está en condición de estimar mediante MV el parámetro ρ maximizando la expresión.
5) Tras la obtención del valor estimado de ρ, se obtiene b_{MV} como la estimación mediante MV de los parámetros β del modelo de ponderación espacial, a partir de la expresión $b_{MV} = b_0 - \rho_{MV} b_L$. También es necesario obtener la estimación del

[11] Covarianza estacionaria: Una variable posee covarianza estacionaria cuando la covarianza entre dos valores depende únicamente de la distancia entre los valores.

parámetro σ^2 de la matriz de covarianza del residuos, recordar que se está estimando un modelo con residuos homoscedásticos, por lo tanto la matriz solo depende del parámetro σ^2, este también se estima utilizando MV a partir de la expresión $\sigma^2_{MV} = (1/N)(e_0 - \rho_{MV} b_L)'(e_0 - \rho_{MV} b_L)$.

Para el caso de un modelo de error espacial (solo posee autocorrelación en los residuos generados por el modelo), la función de verosimilitud es:

$$LnL_C = -\frac{N}{2}\ln 2 + \ln|I - \lambda W| - \frac{N}{2}\ln\left|\frac{e'(I - \lambda W)'(I - \lambda W)e}{N}\right| \qquad [3.18]$$

Donde $e = y - Xb_{MCGE}$, b_{MCGE} es la estimación mediante Mínimos Cuadrados Generalizados (MCG) del vector de parámetros β. [24]

La expresión a partir de la cual se obtiene la estimación es:

$$b_{MCGE} = [X'(I - \lambda W)'(I - \lambda W)X]^{-1} X'(I - \lambda W)'(I - \lambda W)y \qquad [3.19]$$

A diferencia de la estimación del parámetro autorregresivo ρ para el caso anterior, la estimación por MV del parámetro autorregresivo λ es más compleja ya que es necesario aplicar un proceso iterativo. Esto debido a la interdependencia entre los parámetros y residuos a calcular, para obtener el valor de λ que maximiza la función de verosimilitud dada en [3.18] es necesario obtener previamente los residuos *e*. Para obtener estos residuos mediante la expresión $e = y - Xb_{MCGE}$ es necesario una previa estimación de b_{MCGE}, y para obtener esta estimación mediante la expresión [3.19] es necesario obtener previamente una estimación de λ, con lo cual se cierra el ciclo de dependencia. Por lo tanto es necesario un proceso iterativo que comience con una estimación robusta de los parámetros β y se detenga cuando llegue a una cierta condición de parada. El procedimiento es:

1) Se aplica el estimador MCO sobre la regresión $y = X\beta + u$, sin tener en cuenta la autocorrelación espacial, se obtiene b como el estimado MCO de β.
2) Se obtienen los residuos e_{MCO} de la estimación mediante MCO del paso anterior.
3) A partir de los residuos e_{MCO}, se obtiene el valor de λ que maximiza la función de verosimilitud para el modelo de error espacial.
4) Luego de obtener λ, sustituir esta en la expresión [3.18] y aplicar el estimador MCG para estimar dicha expresión, se obtiene la nueva estimación b_{MCGE} del vector de parámetros β.
5) Se vuelve a obtener los residuos, mediante la expresión $e = y - Xb_{MCGE}$, esta vez utilizando la estimación b_{MCGE} obtenida en el paso anterior.
6) A partir de los residuos obtenidos en el paso 5), volver al paso 3) y repetir el proceso hasta que se alcancen los criterios de convergencia.
7) Una vez obtenidas las estimaciones definitivas de β y de λ, calcular $\sigma^2_{MV} = (1/N)e'B'Be$ donde $B = (I - \lambda_{MV} W)$. [2]

3.6 *Validación del modelo*

Existen distintos métodos para detectar la bondad de un modelo estimado, un método posible es comparar el modelo estimado con un modelo de fácil uso, como es el caso del modelo de regresión lineal múltiple (expresión [3.10]). Si los resultados del modelo espacial estimado no son mejores que la del modelo más sencillo, es probable que el modelo haya sido mal especificado. También se puede comparar con modelos alternativos. Una técnica posible para validar modelos, es dividir la muestra en dos partes, con una de ellas se realiza la estimación y con la otra se valida la eficacia del modelo, llamada validación cruzada.

Otra posibilidad para determinar la confiabilidad del modelo es el cálculo del coeficiente de determinación R^2, este coeficiente se puede interpretar como la proporción de variación de la variable endógena que queda explicada por el modelo de regresión, es decir, qué tanto explican las variables exógenas a la variable endógena según la relación dada por el modelo. El valor de R^2 esta dado por la siguiente expresión:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

dónde, $\hat{y}_i$ e y_i son los valores estimado y real respectivamente, de la variable endógena en la observación *i*, $\bar{y}$ es el valor medio de los valores observados de y_i. Como se cumple que el denominador siempre será mayor o igual que el numerador, entonces R^2 tomará valores dentro del intervalo $0 \le R^2 \le 1$. Cuanto más cercano a uno se encuentre el valor de R^2, mejor explicada se encuentra la variable endógena por el modelo estimado. [13][17]

Una particularidad del coeficiente de determinación es que tiende a aumentar cuando aumenta la cantidad de variables exógenas que se agregan al modelo (aunque las variables que se agreguen no aporten nada en la explicación sobre la endógena). Esto se debe a que al aumentar la cantidad de variables disminuye la variabilidad no explicada. Es por ello que se suele utilizar el coeficiente de de determinación corregido por el número de grados de libertad $\overline{R}^2$, el valor está dado por la siguiente expresión:

$$. \ \overline{R}^2 = 1 - \frac{\frac{1}{n-(k+1)} \sum_{i=1}^{n} (\hat{y}_i - y)^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Observar que en esta expresión se utilizan las varianzas (el numerador es la varianza del error y el denominador es la varianza de la variable endógena), al contrario que la expresión donde se utilizan sumas de cuadrados. Se cumple que $\overline{R}^2 \le R^2$. Si bien a partir del valor de $\overline{R}^2$ no se puede deducir inmediatamente la validez del modelo (debido a que no se pude comparar éste valor con ningún tipo de tabla estadística como se suele hacer con otros índices que se calculan en distintas técnicas econométricas), se suele establecer que para valores de coeficientes mayores a 0.8 se puede considerar que el modelo estimado es bueno. Se debe tener en cuenta que un valor alto del coeficiente no garantiza la validez del modelo, ya que éste valor puede ser casual. Esto significa que un valor alto es un buen indicio pero no se puede tomar ninguna decisión confirmatoria a partir de él. [13][17]

Un contraste muy utilizado para comprobar si un conjunto de variables exógenas influye en la variable endógena, es el contraste de regresión múltiple de la *F*, donde las hipótesis del contraste son las siguientes:

$$H_0 \equiv \beta_1 = \beta_2 = ...\beta_K$$
$$H_1 \equiv \beta_i \neq 0 \ \textit{para algún i}$$

Si se cumple H_0 (todos los coeficientes del modelo valen 0), implica que ninguna de las variables exógenas influyen sobre la endógena (al menos teniendo en cuenta la relación dada por el modelo estimado), en este caso el estadístico *F*, sigue una distribución *F* de Snedecor con *(k-1)* y *(n-k)* grados de libertad, donde *n* es el número de observaciones y *k* es la cantidad de variables exógenas. Este contraste utiliza la varianza residual, para comparar dos modelos, uno de ellos no contiene ninguna restricción (no contiene variables exógenas), el otro es el modelo que se quiere chequear. Si el error del modelo es notoriamente inferior al modelo sin restricciones, entonces se rechaza la H_0. [13][18]

Otra forma de evaluar la capacidad del modelo para describir la realidad, es observar los errores cometidos por el modelo, esto significa ver la diferencia entre el valor real y el estimado de la variable endógena. Existen distintos índices para medir los errores, para realizar el cálculo se utiliza la diferencia (en cada observación i) entre el valor real de la variable endógena (y_i) y el valor estimado $\hat{y}_i$. La siguiente tabla muestra algunos de los índices más importantes. [19]

Nombre	**Fórmula**
Error Medio (EM): La desventaja de este índice es que puede tomar valores pequeños aunque existan grandes desviaciones entre el valor real y el calculado.	$EM = \frac{1}{n}\left\{\sum_{i=1}^{n}(y_i - \hat{y}_i)\right\}$
Error Medio Absoluto (EAM): Este índice no posee la desventaja del anterior, ya que tiene en cuenta el valor absoluto de la diferencia.	$EAM = \frac{1}{n}\left\{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert\right\}$
Raíz del Error Cuadrático Medio (RECM): Se aplica la raíz para que el error quede expresado en la misma unidad que la variable en estudio.	$RECM = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$

La desventaja de estos índices, es que si se produce un cambio de escala sobre la variable endógena, afecta la magnitud de los índices, por ello surgen los índices que se muestran en la siguiente tabla, para evitar tal inconveniente. [19]

Nombre	**Formula**
Porcentaje Medio de Error (MPE)	$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)\times 100$
Error Absoluto Medio del Porcentaje de Error (MAPE)	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}\right)\times 100$

Coeficiente de Desigualdad de Tehil (CDT): Es uno de los índices de error más utilizados en los modelos econométricos. Cuanto más cercano a cero se encuentre el valor de éste índice, mejor es la estimación del modelo, en caso que el valor sea cero, significa que la estimación es perfecta.	$$CDT = \frac{\sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_i - y_i)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}\hat{y}_i^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}y_i^2}{n}}}$$

3.7 *Uso del modelo para predecir*

Después que se obtiene la estimación de los parámetros del modelo, se puede utilizar el modelo con el objetivo de realizar predicciones acerca de los valores futuros del rendimiento para la zafra siguiente, o alguna posterior, siempre y cuando se cuente con datos muestrales de las variables que pertenecen al modelo. Esto último significa que si se quiere realizar una predicción del rendimiento para una zafra futura a realizarse en enero del 2009, entonces previamente se debe haber obtenido valores muestrales en la chacra, sobre las variables explicativas que componen el modelo. [3]

La predicción se realiza simplemente sustituyendo los valores obtenidos para cada una de las variables en el modelo estimado, la salida es el rendimiento futuro para cada celda en la cual se dividió la chacra para obtener los valores muestrales de las variables. Es importante tener en cuenta que aunque se halla estimado un buen modelo respecto a los datos históricos, esto no garantiza una buena capacidad predictiva del modelo respecto al futuro, las razones de esto son explicadas en este capítulo.

Surge un aspecto importante cuando se utiliza la estimación de un modelo de regresión para realizar predicciones, esto es que la relación funcional estimada entre el rendimiento y sus factores, debe mantener una estabilidad para la zafra en que se realiza la predicción. Esto significa que, si para la fecha para la cual se realiza la predicción surge algún suceso que hace que se realice algún cambio estructural importante en la chacra, como puede ser algún hecho meteorológico importante, entonces la predicción realizada no tendrá sentido debido a que los parámetros estimados no reflejan la nueva realidad en que se encuentra la chacra. [3]

El tiempo también juega un rol importante en la predicción, esto se debe a que las variables agronómicas son dinámicas, tanto en el espacio como en el tiempo, en consecuencia las relaciones entre las variables puede cambiar con el tiempo, por lo tanto solo se puede asegurar una buena predicción a corto plazo.

Otro aspecto a tener en cuenta es el tipo de muestreo que se realiza tanto en la obtención de datos, con la que se realiza la estimación del modelo, como en la obtención de datos para la predicción, el muestreo debe ser el mismo en ambos casos, esto se debe a la definición de la matriz de contigüidad. Si para la estimación del modelo se realiza un muestreo simple estratificado entonces la matriz adecuada a utilizar en el modelo es una que utiliza un criterio de distancia, pero si durante la extracción de datos para la predicción se utiliza un muestreo en

grilla sistemático (corresponde un matriz de contigüidad física) y estos datos se sustituyen en un modelo que utiliza una matriz en base a criterio de distancia la consecuencia puede ser una predicción del rendimiento engañosa. Lo mismo ocurre si tanto para la estimación del modelo como para la predicción, los datos se extraen a través de un muestreo en grilla sistemático. Pero en el primer caso el diámetro de la celda es la mitad que en el segundo caso (cuanto mayor es el tamaño de la celda, menor es la cantidad de los puntos muestrales y en consecuencia es mas económico el proceso de muestreo). Entonces en el modelo se utiliza una matriz de contigüidad física de segundo orden, pero cuando se sustituyan los datos en el modelo para realizar la predicción, la distribución de estos no se corresponde con la distribución de los datos con la cual se estimó el modelo y en consecuencia la predicción del rendimiento puede ser engañosa. Estos aspectos pueden ser confusos por el hecho de que se puede llegar a pensar que el modelo estimado se puede aplicar siempre en la chacra y esto no es así, por eso es importante tenerlo en cuenta en el momento de realizar la predicción.

Es importante destacar que el modelo estimado en base a datos de la zafra para una chacra solo se puede utilizar para realizar predicciones del rendimiento (en futuras zafras) para la misma chacra. El modelo estimado no se debe aplicar en otras chacras, debido a que las relaciones entre las variables entre una chacra y otra son distintas, además por otro lado, los factores que influyen en el rendimiento pueden cambiar de una chacra a otra; por todo esto el modelo estimado para una chacra no se puede utilizar en otras.

En lo que se refiere a las variables con las cuales se estima el modelo, en el momento de obtener la muestra de datos para realizar la predicción, se deben obtener valores para cada una de las variables con las cuales se estimó el modelo. Si por alguna razón no se cuenta con datos para todas las variables entonces éstos no pueden ser aplicados al modelo aunque se le asigne el valor cero a las variables para las cuales no se obtuvo datos, en este caso las predicciones obtenidas del rendimiento no serán buenas. Lo que se debe realizar en este caso es volver a estimar el modelo pero solo con las variables para las que se cuenta con valores muestrales en el momento de realizar la predicción.

Se resume en los siguientes puntos las condiciones necesarias para utilizar el modelo como un modelo predictivo:

- Para realizar la predicción se debe contar con datos de todas las variables para la zafra a predecir que se utilizan en el modelo estimado, de lo contrario la predicción no se puede realizar. Eventualmente de contarse con menos variables, se debería estimar nuevamente el modelo con dichas variables.
- El modelo estimado se debe encontrar correctamente especificado.
- El horizonte de predicción no debe ser muy lejano, debido a que la relación entre el rendimiento y las variables sufre cambios, y el modelo estimado no es una representación de la realidad y en consecuencia las predicciones tampoco.
- El modelo estimado solo puede ser utilizado en la chacra de la cual se obtuvo los datos para realizar la estimación.
- Se debe mantener el mismo criterio de muestreo tanto para la obtención de datos para realizar la estimación, y en los datos para realizar la predicción.

Capítulo 4

Solución al problema

A partir de la investigación sobre distintas metodologías que permitan analizar datos de corte transversal con el objetivo de predecir el cultivo de una chacra, se llega a la conclusión que las herramientas que provee la Econometría Espacial son las apropiadas para aplicar a este problema. La solución al problema se basa en una herramienta que contiene un conjunto de contrastes de especificación, modelos de especificación, métodos de estimación y procedimientos de validación; donde la interacción conjunta de estos elementos permite brindar la predicción deseada por el cliente.

Por lo tanto se pueden distinguir cinco componentes generales en los cuales se encuentra dividida la solución, como lo muestra la Fig. 4.1. El primero de ellos es el *análisis y tratamiento* de los valores de las variables que componen el conjunto de datos, este es el principal de todos los componentes ya que la salida de éste determina qué variables formarán parte del modelo y cuáles son las propiedades que cumplen los datos. Mediante la salida del análisis, en donde se presentan las propiedades que los datos cumplen y las que no cumplen se obtiene y define el modelo de regresión apropiado. El segundo componente se trata de la *definición y obtención* del modelo; este componente debe decidir en base al análisis realizado en la etapa anterior cual es el modelo correcto. Los modelos de regresión básicos con los que debería contar la herramienta son: el modelo de regresión lineal, el de ponderación espacial y el de error espacial. El tercer componente es el encargado de seleccionar y aplicar el procedimiento de *estimación* apropiado, en caso que los datos posean algún tipo de autocorrelación los procedimientos de estimación se basan en el método de MV, en caso contrario el procedimiento que se aplica es en base a MCO o algunas de sus variantes dependiendo si los errores producidos por el modelo poseen o no heteroscedasticidad en los datos. El cuarto componente debe contar con un determinado conjunto de *test de validación* que permitan decidir si la estimación realizada en el paso anterior es confiable o no, si el resultado de los test no es satisfactorio se debe volver a la etapa de análisis de datos, ya que probablemente alguna decisión tomada en esta etapa no fue la apropiada. El quinto y último componente realiza la *predicción* de los rendimientos para la zafra futura en base al modelo estimado en los puntos anteriores y los datos recolectados para la chacra que se quiere predecir. Eventualmente este componente debe contar con herramientas que permitan validar que las predicciones obtenidas son confiables o al menos dar algún tipo de intervalo de confianza para que el cliente tome las decisiones pertinentes.

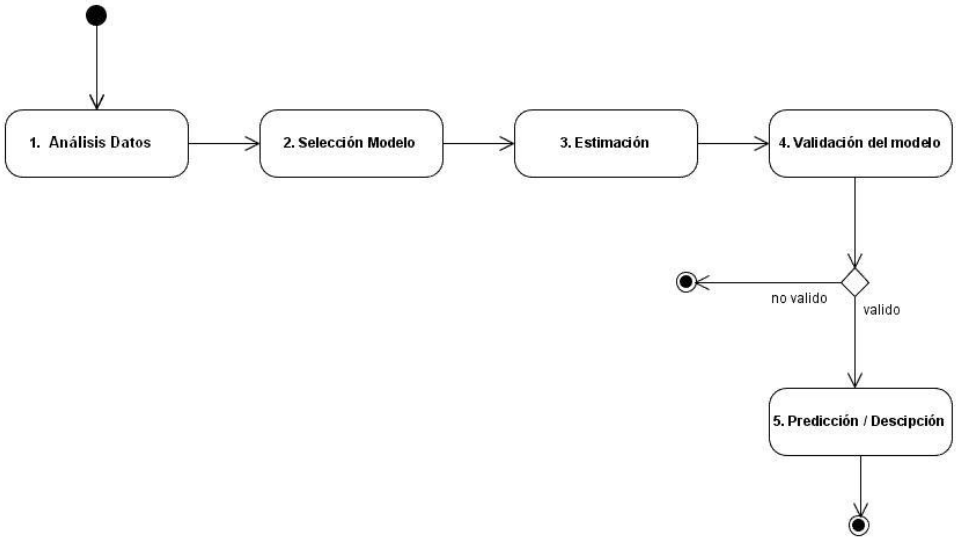


Figura 4.1 Diagrama de etapas de la solución

A su vez cada etapa esta subdividida en más etapas las cuales se muestran en el siguiente diagrama.

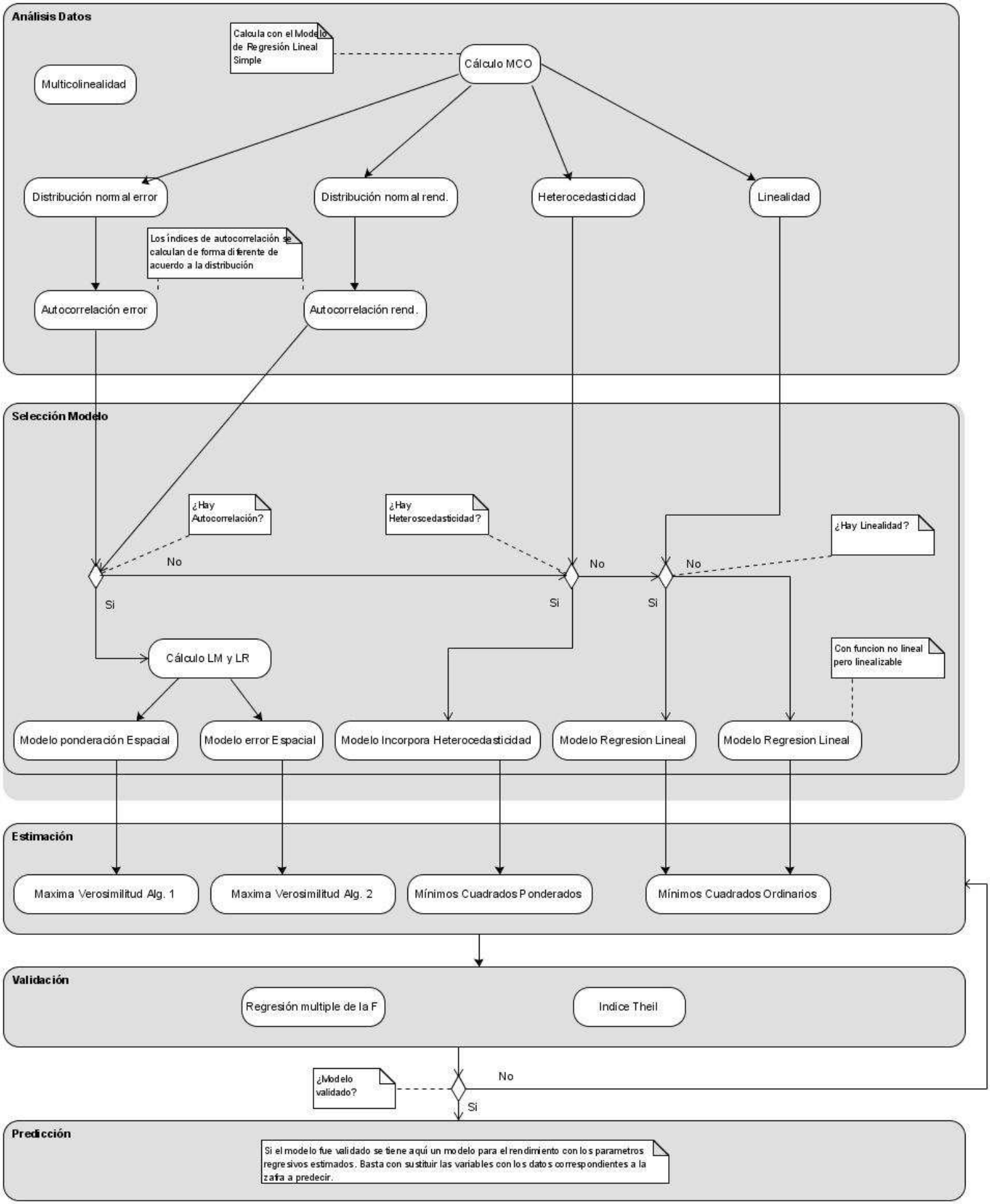


Figura 4.2 Diagrama detallado de etapas de la solución

4.1 Análisis de datos

Esta etapa es la encargada de realizar un análisis de los datos tanto en las variables exógenas como en la endógena. Los análisis se pueden dividir en dos grupos, uno de ellos es el encargado de analizar a nivel lineal la relación entre las variables mediante contrastes de linealidad y multicolinealidad. El segundo grupo es el encargado de determinar si los datos poseen o no los llamados efectos espaciales autocorrelación y heteroscedasticidad. Además de estos dos grandes grupos existen otros tipos de contrastes que sirven de apoyo para los contrastes más importantes y para la selección del modelo final, como es el caso del contraste de normalidad.

Análisis de Multicolinealidad

El estimador de MCO se debe aplicar como estimador sea cual sea el modelo de regresión seleccionado para trabajar. En algunos como el modelo de regresión lineal simple este es el estimador final y en otros este se aplica como paso intermedio por ejemplo para sacar el vector de errores para analizarlo y mejorar el modelo. Como ya se vio el estimador de MCO requiere que la matriz X conformada por los valores de las variables regresoras sea invertible (ver ref. [24]).

En este contexto se pueden dar tres casos entre las variables exógenas:[13]

- Que sean linealmente dependientes lo cual hace que la matriz no sea invertible.
- Que dos variables sean ortogonales lo cual implica que estas variables no aportan nada al modelo, si se agregan o quitan variables ortogonales el aporte será nulo.
- Que exista una situación intermedia entre los dos casos extremos anteriores. Esto es, existe una cierta relación entre las variables explicativas, lo que hace que los coeficientes de regresión estén correlacionados. Si está relación es muy fuerte porque dos o más variables regresoras "están próximas" a una relación entonces estamos frente a un problema de Multicolinealidad.

Lo que realmente nos está diciendo la Multicolinealidad es que es un problema de la muestra, de la que se quiere obtener más información de la que contiene. Una vez detectada la Multicolinealidad se debe omitir del modelo las variables que generan este efecto no deseado. [13]

Lo primero que se debe hacer es detectar la Multicolinealidad, para esto existen cuatro métodos:

E. Gráfico de Dispersión Matricial
Por medio de este grafico se puede apreciar una posible relación lineal entre dos o más variables regresoras. [13]

F. Matriz R de regresión de las variables regresoras
Si existe algún valor dentro de la matriz R fuera de la diagonal $\left(r_{i,j}, i \neq j, proximo\ a\ \pm 1\right)$ podemos inferir una fuerte relación lineal entre las variables $x_i\ y\ x_j$. El problema que tiene este método es que no detecta las relaciones lineales entre una variable y un conjunto de variables. [13]

G. Elementos diagonales de la matriz R^{-1}

Si invertimos la matriz de correlaciones R y algún elemento de la diagonal i es "grande", por ejemplo mayor que 10 podríamos inferir que la variable x_i es la variable

que causa la Multicolinealidad. A raíz de esto deberíamos eliminar esta variable del modelo. La desventaja de este método es que la matriz R^{-1} se calcula con poca precisión (depende mucho de la muestra) cuando la matriz R es casi singular (su determinante es próximo a cero). [13]

H. Calcular valores propios de matriz de correlación R, Índice de Condicionamiento

Cuando se calculan los valores propios de la matriz R y éstos son iguales a uno, se puede verificar que las variables regresoras son todas ortogonales. Pero si alguno de los valores propios es próximo a cero indica la presencia de Multicolinealidad y se puede verificar también que la variable regresora asociada al valor propio es aproximadamente una combinación lineal de las otras variables regresoras. Se define el índice de condicionamiento de la matriz R que es una buena medida de la singularidad de esta matriz y se calcula de la siguiente forma: [13]

$$IC(R) = \sqrt{\lambda_{MAX} / \lambda_{MIN}}$$

El criterio para la detección de Multicolinealidad es el siguiente:

Si $IC(R) \leq 10$, no hay Multicolinealidad.

Si $10 \leq IC(R) \leq 30$, hay Multicolinealidad moderada.

Si $IC(R) > 30$, hay Multicolinealidad alta.

Creemos que este es el método adecuado para la detección de la Multicolinealidad dado que no tienen ningún tipo de restricción y además detecta Multicolinealidad entre una variable y un conjunto de variables.

Una vez detectada la Multicolinealidad se debe realizar un algoritmo para quitar dichas variables del modelo. En el caso de utilizar el método de índice de condicionamiento lo que se debe hacer es ir quitando toda aquella variable cuyo valor propio asociado es el más pequeño y realizar nuevamente el contraste de Multicolinealidad hasta que se alcance el nivel deseado. Esto siempre y cuando nos encontremos frente a un modelo de regresión lineal o multivariante. En el caso de los modelos de regresión espacial no existe una documentación que ataque el tema de la Multicolinealidad especialmente, es por eso que una vez completado la implementación de toda la biblioteca en la cual se seleccione el modelo y se pueda estimar, se debería contrastar los resultados obtenidos teniendo en cuenta el algoritmo que elimina la Multicolinealidad con los resultados obtenidos considerando todas las variables.

Análisis de Linealidad

El objetivo del análisis de linealidad es verificar que la relación entre el rendimiento y las variables es lineal en los coeficientes β, como ocurre en [3.8]. Para determinar si existe o no linealidad es necesario medir de alguna manera que tan fuerte puede llegar a ser la asociación lineal, esto se puede realizar mediante el *coeficiente de correlación múltiple*.

$$R^2 = \frac{\sum (y_i - \bar{y})^2}{\sum (\hat{y}_i - \bar{y})^2}$$

donde, $\hat{y}_i$ representa el valor estimado del rendimiento en cada observación *i*, y_i representa el valor real del rendimiento en cada observación *i*, por último $\bar{y}$ representa el valor medio del rendimiento real. El valor de R^2 varía entre 0 y 1, si vale uno significa que la relación lineal es perfecta. La idea del contraste es ver si existe una correlación ente el rendimiento estimado y el real, $\hat{y}_i$ es estimado partiendo de un modelo de regresión lineal, si llega a estar correlacionado con el rendimiento real y_i , entonces esto significa que la relación entre el rendimiento y las variables es lineal en los coeficientes β. [12]

Este contraste es necesario para la toma de decisiones en la selección del modelo a aplicar, ya que de detectarse que la relación entre las variables y el rendimiento es lineal, el modelo de regresión lineal simple es uno de los posibles modelos a aplicar. Se debe tener en cuenta que este contraste deberá ser analizado conjuntamente con los contrastes de autocorrelación y heteroscedasticidad que pueden llegar a descartar el modelo de regresión lineal simple. Si el contraste determina que no existe una relación lineal entre el rendimiento y las variables, entonces se debe buscar cual es el tipo de relación. En caso que no existan fundamentos teóricos fuertes a priori que nos orienten a la elección de una determinada función de producción que represente la relación entre el rendimiento y las variables, se debe probar con funciones no lineales (que sean linealizables) conocidas, para así poder aplicar el modelo de regresión lineal. [24]

Análisis de la distribución

Para el análisis de distribución la biblioteca cuenta con tres contrastes, los cuales son: el contraste de asimetría, el de curtosis y uno que combina los anteriores. Los tres contrastes persiguen el objetivo de contrastar la normalidad o no de la distribución de los datos con los cuales se cuenta.

En el contraste de asimetría la idea que subyace es que la distribución normal es simétrica, bajo la hipótesis de normalidad el coeficiente de asimetría (CA) toma el valor cero. Los pasos del algoritmo son los siguientes:

1) Se calcula el coeficiente CA de la siguiente forma:

$$CA = \frac{m_3}{s_X^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n s_X^3}$$

2) Una vez obtenido este contraste se lo estandariza obteniendo el coeficiente de asimetría estandarizado (CAS), ya que para muestras grandes ($n \geq 50$) el CAS sigue una distribución N(0,1) y se usa como estadístico para contrastar la hipótesis de que la distribución de la muestra es simétrica.

$$CAS = \sqrt{\frac{n}{6}} CA \sim N(0,1) \qquad [13]$$

El contraste de curtosis o de apuntamiento, sirve para contrastar la hipótesis de que el coeficiente de apuntamiento (*CAp*) es cero. Esta propiedad es usada para verificar la distribución normal. Los pasos del algoritmo son los siguientes:

1) Se calcula el coeficiente *CAp* de la siguiente forma:

$$CAp = \frac{m_4}{s_X^4} - 3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n s_X^4} - 3$$

2) Una vez obtenido este contraste se lo estandariza obteniendo el coeficiente de apuntamiento estandarizado (*CApS*), ya que para muestras grandes ($n \geq 50$) el *CApS* sigue una distribución N(0,1) y se usa como estadístico para contrastar la hipótesis de que la distribución de la muestra es simétrica.

$$CApS = \sqrt{\frac{n}{24}} CAp \sim N(0,1) \qquad [13]$$

Por último se cuenta con el contraste en conjunto, en el cual se combinan los dos contrastes anteriores y se obtiene un estadístico:

$$d = CAS^2 + CApS^2$$

Bajo la hipótesis de normalidad de dicho estadístico se distribuye en forma asintótica como una chi-cuadrado con dos grados de libertad. Entonces se tiene que si *d* toma valores positivos grandes (según una χ_2^2 con dos grados de libertad) se rechaza la hipótesis de distribución simétrica y/o que tiene curtosis nula, por lo tanto se rechaza la hipótesis de normalidad. [13]

Análisis de autocorrelación:

Como se mencionó en la sección 3.23.2.1 los contrastes que se utilizan para determinar la presencia o no de la autocorrelación espacial (tanto en el rendimiento como en el término de error) son el contraste de I de Moran y C de Geary. Se optó por estos contrastes ya que son los más utilizados para detectar la autocorrelación global y son los que se aplican cuando la muestra se obtuvo mediante una grilla regular, el cual es el caso de la muestra del rendimiento del cultivo para el caso en estudio. Otra posibilidad en caso que la muestra sea irregular es utilizar el índice G de Getis y Ord, (ver ref. [24]), es aconsejable que en trabajos futuros la biblioteca cuente con la implementación del índice G de Getis y Ord y de esta manera abarca de forma general cualquier tipo de grilla ya sea regular o irregular.

La solución brinda la posibilidad de aplicar cualquiera de los dos contrastes, pero el usuario debe especificar cual se utiliza. Ambos contrastes retornan el índice y un valor de significancia. Para obtener el valor de significancia se utiliza la tabla de distribución N(0,1) ya que tanto el I de Moran como la C de Geary estandarizados siguen una distribución N(0,1) cuando la muestra es grande. Estos contrastes indican si existe o no autocorrelación espacial, pero solo en base al resultado de ellos no se puede decidir que modelo aplicar, si el modelo de ponderación espacial o, el modelo de error espacial, para ello se debe aplicar el test que se basan en los Multiplicadores de Lagrange. [24]

Los pasos del contraste de I de Moran para detectar autocorrelación en el rendimiento son:

1) Se obtiene la matriz de vecindad W estandarizada, en base a los datos del rendimiento. El criterio de vecindad que se aplique puede ser de los vistos en la sección 3.3, Queen, Rook o Bishop, tanto de primer o segundo orden, según lo que defina el usuario.
2) Se calcula el índice I, en base a la expresión [3.1]
3) Se obtiene los valores estimados de *V(I)* y *E(I)* en base a las expresiones [3.4] y [3.3] respectivamente.
4) Luego de obtenidos los valores de *V(I)* y *E(I)* se calcula el valor de I estandarizado *Z(I)*, expresión [3.2]

5) En base a *Z(I)* se obtiene el valor de significancia utilizando la distribución N(0,1), se compara este valor con una constante ingresada por el usuario que indica el porcentaje de confiabilidad que pretende de parte del índice.

En el paso 5) se compara el nivel de significancia para el índice calculado contra una constante que el usuario define, esta constante típicamente tiene un valor del 95% o 99% de confiabilidad, o sea que el usuario define valores de 0.95 o 0.99 respectivamente. En caso que el usuario defina esta constante como 0.99 y el valor de significancia que retorna el procedimiento es de 0.93, entonces se entiende que no hay autocorrelación en los datos, ya que el porcentaje de confiabilidad no sobrepasa el límite especificado por el usuario.

Los pasos del contraste de I de Moran para detectar autocorrelación en los errores son:

1) Se estima el modelo de regresión lineal utilizando el estimador MCO.
2) A partir del modelo estimado en el paso anterior se obtienen los errores producidos por la estimación.
3) Se obtiene la matriz de vecindad W estandarizada, en base a los errores obtenidos en el paso anterior. El criterio de vecindad que se utilice puede ser de los vistos en la sección 3.3, Queen, Rook o Bishop, tanto de primer o segundo orden, según lo que defina el usuario.
4) Se calcula el índice I, en base a la expresión [3.1].
5) Se obtiene los valores estimados de *V(I)* y *E(I)* en base a las expresiónes [3.4] y [3.3] respectivamente.
6) Luego de obtenidos los valores de *V(I)* y *E(I)* se calcula el valor de I estandarizado *Z(I)*, expresión [3.2]
7) En base a *Z(I)* se obtiene el valor de significancia utilizando la distribución N(0,1), se compara este valor con una constante ingresada por el usuario que indica el porcentaje de confiabilidad que pretende de parte del índice.

Los pasos para calcular el índice de C de Geary son similares al I de Moran, la diferencia pasa por la fórmula que se utiliza para calcular el índice y la estimación de la varianza y esperanza del índice.

Análisis de heteroscedasticidad:

Para determinar la presencia de heteroscedasticidad en los datos se optó por los contrastes de White y de Spearman ya que son dos contrastes con bastante uso y que brindan usos complementarios. Dado que el contraste de White tiene como ventaja que es general por no establecer ninguna suposición en lo previo, y una vez detectada la heteroscedasticidad, mediante el uso de Spearman se puede detectar qué variable o qué variables son las que introducen este fenómeno en los datos.

La solución brinda la posibilidad de aplicar cualquiera de los dos contrastes, o de combinarlos a ambos chequeando primero la existencia o no de heteroscedasticidad, y en caso de verificarse la misma, aplicar el contraste de Spearman con el fin de detectar en donde específicamente se presenta la heteroscedasticidad.

En el contraste de White la idea que subyace es la de intentar determinar si las variables explicativas del modelo, sus cuadrados y todos sus combinaciones posibles no repetidas, sirven

para determinar la evolución del error al cuadrado. Esto es, si la evolución de las variables explicativas, de sus varianzas y de las covarianzas, son significativas para determinar el valor de la varianza muestral de los errores. [3]

Los pasos del contraste de White para detectar heteroscedasticidad son los siguientes:

1) Se estima el modelo de regresión lineal mediante MCO y se calculan los errores producidos por el modelo.
2) Se estima un modelo en el que la variable endógena serían los valores al cuadrado de los errores obtenidos previamente (paso 1) con todas las variables explicativas del modelo inicial, sus cuadrados y sus combinaciones no repetidas.

$$e_i^{\,2} = \alpha_0 + \alpha_1 x_{1i} + \ldots + \alpha_k x_{ki} + \alpha_{k+1} x_{1i}^{\,2} + \ldots + \alpha_{k+k} x_{ki}^{\,2} + \alpha_{k+k+1} x_{1i} x_{2i} + \alpha_{k+k+2} x_{1i} x_{3i} + \ldots + \alpha_{3k+1} x_{2i} x_{3i} + \ldots + \varepsilon_i$$

3) Una vez estimado el modelo del paso 2 y obtenido el vector de los α_i se calcula el valor de la R_e^2 de este segundo modelo (sea $\xi_i = e_i^{\,2}$ del modelo anterior, entonces $R_e^2 = \left(\hat{\xi}_i - \bar{\xi}\right)^2 \Big/ \left(\xi_i - \bar{\xi}\right)^2$). Este valor es usado para la estimación del contraste de la siguiente forma: si obtenemos un valor del producto $n R_e^2$ (con n igual a la cantidad de datos con los que se cuenta) mayor que el reflejado por las tablas de la χ^2_{p-1}, con p igual a la cantidad de elementos del vector de α, se puede afirmar la existencia de heteroscedasticidad, y viceversa, si este valor es más pequeño se afirmaría que se mantiene la homoscedasticidad. [3][16]

El contraste de Spearman se basa en que la variable sospechosa de producir heteroscedasticidad debería provocar un crecimiento del residuo estimado, al mismo ritmo que ella va creciendo. Entonces los pasos para el algoritmo de Spearman son los siguientes:

1) Se ordena de menor a mayor tanto la variable sospechosa x_{ij} como el valor absoluto del residuo $|e_i|$. A continuación se presenta una tabla como ejemplo:

Series originales			Series ordenadas				d	d^2
Puesto	x_{ij}	$\|e_i\|$	x_{ij}	Puesto original	$\|e_i\|$	Puesto original		
1	1.838	1,6	424	2	1,2	3	2-3=-1	1
2	424	1,4	501	3	1,3	4	3-4=-1	1
3	501	1,2	688	5	1,4	2	5-2=3	9
4	2.332	1,3	1.838	1	1,5	5	1-5=-4	16
5	688	1,5	2.332	4	1,6	1	4-1=3	9

Tabla 4.1.1 – contraste de Spearman

2) Entonces el cambio de puesto en ambas, para cada una de las observaciones, debería ser el mismo número de puestos respecto al orden original de las series. Si el cambio de puesto respecto al original no es el mismo para las dos (luego de ordenarlas) se puede hablar de movimientos no correlacionados. Spearman propone determinar el grado de correlación en ese cambio de puesto respecto al inicial de cada una de las variables, este grado de correlación esta dado por la siguiente expresión:

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

3) Por último en base al *r* calculado en el paso anterior se obtiene el siguiente ratio, con una función de distribución conocida bajo hipótesis nula de no significancia:

$$\frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \rightarrow t_{n-2}$$

Se tiene entonces que si el resultado del ratio es superior al valor en la tabla *t* se puede afirmar que la correlación es significativa y que hay indicios de heteroscedasticidad en el modelo provocado por la variable x_{ij}.

4) Este proceso se sigue para cada una de las variables exógenas mientras no se presente heteroscedasticidad, ni bien el algoritmo detecta heteroscedasticidad en una variable, detiene su iteración, obteniéndose la confirmación de este fenómeno en los datos y además la variable en donde este algoritmo detectó la heteroscedasticidad. [3][16]

Definición de la matriz W

Los datos provistos se obtuvieron con distintos métodos de muestreo. Los obtenidos mediante imágenes satelitales vienen en formato de grilla, esta grilla es la que define la grilla final de información entregada por el usuario. El rendimiento y la conductividad eléctrica se obtuvo mediante un monitor de rendimiento de alta densidad, esto significa que se obtiene información de una gran cantidad de puntos del suelo, por lo tanto para obtener un valor del rendimiento por cada celda de la grilla, se realizo por parte del cliente un promedio del rendimiento con los puntos obtenidos por el monitor en cada celda, lo mismo para la conductividad eléctrica. Otros datos del suelo fueron muestreados en forma aleatoria, por lo cual se les realizó por parte del cliente una interpolación con métodos geoestadísticos para completar la información en aquellas celdas donde no se obtuvo información muestral.

En consecuencia, como los datos se encuentran ordenados en una grilla, el criterio de definición que se escoge para la matriz, es en base a la contigüidad física. Solo resta definir que tipo de contigüidad física es más conveniente utilizar, Queen, Rook o Bishop. Intuitivamente parece mas adecuado aplicar el criterio de contigüidad física de Queen, ya que este criterio tiene en cuenta 8 vecinos por punto muestral frente a los otros dos criterios que solo tienen en cuenta 4 vecinos, pero se debe tener sumo cuidado al momento de decidir la cantidad de vecinos, ya que muchos de ellos pueden no aportarle nada al modelo y esto puede traer como consecuencia un modelo mal especificado o un contraste de autocorrelación con resultado erróneo. Como ejemplo se puede citar el caso de las matrices de segundo orden, estas se deben aplicar cuando el valor de la variable en un punto no solo se ve afectado por las unidades inmediatamente contiguas, si no también por las contiguas de segundo orden, pero en caso que el valor de las unidades contiguas de segundo orden no influyan sobre el punto analizado, puede acarrear como consecuencia que los resultados no sean los deseados.

Solo resta definir el orden de la matriz, esto depende del tamaño de la celda ya que no es lo mismo una celda de 5 m a una de 20 m. Para el caso de celdas de 5 m de lado la distancia máxima a una observación de segundo orden es de 14,15 m en cambio para una celda de 20 m de lado la distancia es de 56,6 m, parece intuitivo pensar que las observaciones del rendimiento que se encuentran a 56,6 m influyen en menor medida que aquellas que se encuentran a 14,15 m.

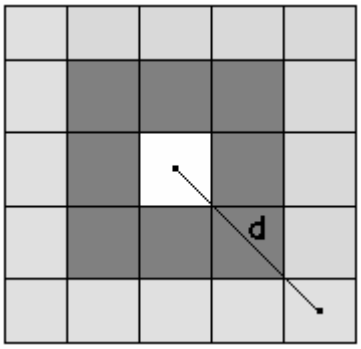


Figura 4.2 Distancia máxima entre dos observaciones de segundo orden.

4.2 Selección del modelo

En la sección 3.4.4 se vieron dos estrategias para seleccionar el modelo regresivo correcto, el primero de ellos propuesto por Anselin y Florax (1995) selecciona uno de los tres siguientes modelos, modelo de regresión lineal, modelo de ponderación espacial y modelo de error espacial, en base a los resultados de los test LM_λ y LM_ρ. La otra estrategia EEV2 propuesto Folmer y Florax (1992), selecciona uno de los siguientes modelos, modelo de error espacial, modelo de ponderación espacial o un modelo que incorpore la autocorrelación en un conjunto de variables que no fueron tenidas en cuenta y de las cuales se posee una justificación teórica por la cual se le asigna un retardo espacial a las mismas (ver sección 3.4.4).

El principal inconveniente que poseen ambas estrategias es que no tienen en cuenta la heteroscedasticidad que se puede producir en los errores cometidos por el modelo. Si se compara ambas estrategias, vemos un punto a favor de la estrategia propuesta por Anselin y Florax (1995) frente a la estrategia de Folmer y Florax (1992), esta es que no siempre se cuenta con un conjunto de variables omitidas del modelo, en principio se incluyen en el modelo todas las variables de las cuales se cuente con datos.

La estrategia EEV2 no es tenida en cuenta en la solución planteada debido a que no existe en principio un subconjunto de variables exógenas excluidas para la selección del modelo, que posteriormente puedan ser incluidas todas o algunas de las variables para mejorar la estimación.

Si bien en la etapa de análisis de datos se puede llegar a excluir variables del conjunto inicial, esto se realiza mediante el contraste de Multicolinealidad (ver sección 4.1), estas variables no puede ser agregadas al modelo en algún paso posterior, debido a que la estimación mediante MCO fallaría, esta estimación es necesaria tanto en la estimación de un modelo de regresión lineal (expresión [3.11]) , o como paso intermedio en los algoritmos de estimación para los modelos de regresión espacial (expresiónes [3.12] y [3.13]).

La estrategia propuesta en la presente investigación para la elección del modelo adecuado pasa por el uso de la estrategia formulada por Anselin y Florax (1995), pero además se incorpora el análisis de heteroscedasticidad realizado sobre los errores cometidos por el modelo de regresión lineal estimado mediante MCO.

Para tener una visión general de la selección del modelo, primero se presenta la salida del modulo análisis de datos:

- Resultados del contraste acerca de la relación lineal (*linealidad*) entre las variables exógenas y la variable endógena.
- Resultados del contraste de I de Moran o C de Geary (dependiendo de la elección del usuario) y verificación de la presencia de *autocorrelación* en el rendimiento y/o en el error.
- Resultado de los contrastes de White y de Spearman para detectar la *heteroscedasticidad*.
- Análisis de *multicolinealidad* entre las variables exógenas, eventualmente puede producirse la exclusión de aquellas variables que provoquen multicolinealidad alta.

Teniendo en cuenta la salida anterior se procede a ejecutar el módulo selección del modelo, según los siguientes pasos:

1) Si el contraste para detectar la autocorrelación realizado en el análisis de datos no detecta autocorrelación tanto en el rendimiento como en los términos de errores, y además los contrastes de heteroscedasticidad no detectan dicho fenómeno en el término de error, entonces el modelo que se debe especificar depende del resultado del contraste de linealidad.
    a. Si el contraste de linealidad da que hay una relación lineal entre el rendimiento y las variables entonces se aplica directamente el modelo de regresión lineal simple.
    b. Si el contraste de linealidad da que no hay una relación lineal entre el rendimiento y las variables, entonces teniendo en cuenta un conjunto de funciones de producción no lineales (pero linealizables) se deben linealizar cada una de ellas por medio de transformaciones algebraicas. Se estima cada una de ellas mediante MCO, luego se valida cada una de las estimaciones mediante el test t-Student o mediante el coeficiente de determinación R^2. Aquél modelo que verifique una mejor estimación, será el seleccionado como el modelo final.
2) Si el contraste para detectar la autocorrelación realizado en el análisis de datos no detecta autocorrelación tanto en el rendimiento como en el término de error y además los contrastes de heteroscedasticidad detectan heteroscedasticidad en el término de error, entonces el modelo correcto es un modelo de regresión lineal que incorpore la heteroscedasticidad en el término de error. Este puede ser el modelo de error heteroscedástico o el modelo de heteroscedasticidad aditiva (expresiónes [3.14] y [3.15] respectivamente), la elección de uno u otra se realiza en base al los resultados de los contrastes de heteroscedasticidad.
    a. Si el contraste de White detecta la heterocedasticidad pero los demás contrastes no la detectan (se detecta la heteroscedasticidad de forma general, pero no se

puede especificar la causa por la que esta se da) entonces el modelo que se especifica es el modelo de error heteroscedastico.

b. Si se detecta heterocedasticidad ya sea por medio del contraste Spearman o por el contraste de Breuch y Pagan, entonces se especifica el modelo de heteroscedasticidad aditiva, teniendo en cuenta si la heteroscedasticidad se produce por una variable que pertenece al modelo (contraste de Spearman), o por un conjunto de variables que pueden o no pertenecer al modelo (contraste de Breuch y Pagan).

3) Si el contraste para detectar la autocorrelación realizado en el análisis de datos detecta autocorrelación, entonces se ejecuta la estrategia propuesta por Anselin y Florax (1995).

4.3 Estimación

Esta etapa de la solución consiste en aplicar el estimador adecuado al modelo seleccionado en la etapa anterior. Se plantea el siguiente cuadro donde se indica que estimador se debe aplicar para cada modelo.

Modelo	**Estimador**
Modelo de regresión lineal. $y = X\beta + \varepsilon$	Estimador MCO (ver ref. [24])
Modelo de error heteroscedastico. $y = X\beta + \mu$ $Var(\mu_i) = \sigma_i^2 I$ $E(\mu\mu') = \Omega$	Estimador MCP (ver ref. [24])
Modelo de heteroscedasticidad aditiva. $y = X\beta + \mu$ $Var(\mu_i) = Z\ \gamma$	Estimador MCP (ver ref. [24])
Modelo de ponderación espacial. $y = \rho Wy + X\beta + \mu$	Estimador de MV (Sección 3.5)
Modelo de error espacial. $y = X\beta + \varepsilon$ $\varepsilon = \lambda W\varepsilon + \mu$	Estimador de MV (Sección 3.5)

En la tabla anterior, no se tuvo en cuenta el modelo de mixto con autocorrelación ya sea en la variable endógena o en el término de error y con heteroscedasticidad, esto debido a que en la bibliografía estudiada no se encontró el método adecuado para estimar este tipo de modelo; si bien mencionan que se estiman mediante el estimador de MV.

4.4 Validación del modelo

En la sección 3.6 se vieron distintos métodos que permiten validar el modelo estimado, la validación de un modelo no pasa por la aplicación de un solo método, si no por la aplicación conjunta de ellos.

Uno de los métodos que se vio es dividir la muestra en dos partes y con una de ellas realizar la predicción, y con la otra se valida el modelo. La dificultad que presenta esta técnica, es el criterio que se aplique para dividir la muestra. Si se aplica un criterio geográfico para dividir la muestra, y se estima el modelo teniendo en cuenta la muestra que se obtuvo de una determinada zona de la chacra, y por otro lado se valida el modelo con los datos muestrales de otra zona de la chacra, es probable que los resultados de la validación rechacen el modelo. La explicación de esto se debe a las propiedades heterogéneas que poseen los suelos, como se vio en la sección 3.23.2.2, la estimación se puede haber realizado teniendo en cuenta determinadas propiedades que tiene el suelo en una determinada zona de la chacra, pero las propiedades de los datos con las que se valida el modelo pueden ser diferentes debido a razones geográficas. La otra posibilidad para dividir una muestra es temporalmente, esto significa que en caso que se cuente con datos muestrales de dos zafras seguidas, se pude realizar la estimación del rendimiento con los datos de la primera zafra y validar el modelo con los datos de la segunda zafra.

Otra de las posibilidades que se vio es la utilización del coeficiente de determinación corregido $\overline{R}^2$, la desventaja de este índice es que a mas allá de que éste de un valor alto, esto no asegura la validez del modelo, por eso no es recomendable que un sistema decida la validez del modelo teniendo en cuenta este índice, aunque sí es recomendable que éste figure como salida del sistema.

Teniendo en cuenta la desventaja de los métodos anteriores, creemos que los métodos más indicados para decidir la validez de un modelo es en base al contraste de regresión múltiple de la F, complementado con el cálculo del coeficiente de desigualdad de Theil. [13][18][19]

4.5 Predicción

La predicción del rendimiento se realiza sustituyendo los valores de las variables exógenas en el modelo estimado en el paso anterior. Para realizar la predicción se debe tener en cuenta los aspectos mencionados en la sección 3.7, aunque estos aspectos no son controlados por el sistema, el usuario debe tenerlos en cuenta. La salida de este paso es el rendimiento futuro para cada punto de la chacra donde se obtuvo datos muestrales de la chacra en estudio.

Capítulo 5

Solución Informática

Luego de haber especificado la solución al problema en el capítulo anterior pasamos a especificar funcionalmente como representar informáticamente dicha solución. Para esto mencionaremos los casos de uso más relevantes de la biblioteca a implementar, así como también las decisiones de diseño, la implementación y los respectivos tests realizados.

5.1 Especificación funcional

Las necesidades que debe cubrir el Sistema son:

- Analizar las propiedades necesarias para la selección del modelo, que tienen los datos obtenidos históricamente. Las propiedades que se analizan para los datos son:
    - Distribución de las variables rendimiento y término de error del modelo de regresión, aplicado al estimador de MCO.
    - Autocorrelación en la variable rendimiento y error por medio de los índices I de Moran y la C de Geary.
    - Heteroscedasticidad en los datos por medio de los contrastes de Spearman y White.
    - Contraste de Multicolinealidad en los datos mediante el índice de condicionamiento.
    - Contraste de Linealidad en los datos mediante el coeficiente de correlación múltiple.
- Obtener un modelo regresivo que represente el rendimiento en base a los datos históricos que se tienen. Esta funcionalidad abarca también la identificación de las propiedades anteriormente mencionadas para seleccionar el modelo correcto. Se hace uso también de los multiplicadores de Lagrange para la selección del modelo.
- Estimar el rendimiento en base a un modelo previamente obtenido y con los datos obtenidos de la chacra para el momento en que se quiere estimar el rendimiento.

De lo anteriormente dicho se desprenden entonces los cuatro casos de uso principales del Sistema, los cuales presentamos en el siguiente diagrama de casos de uso.

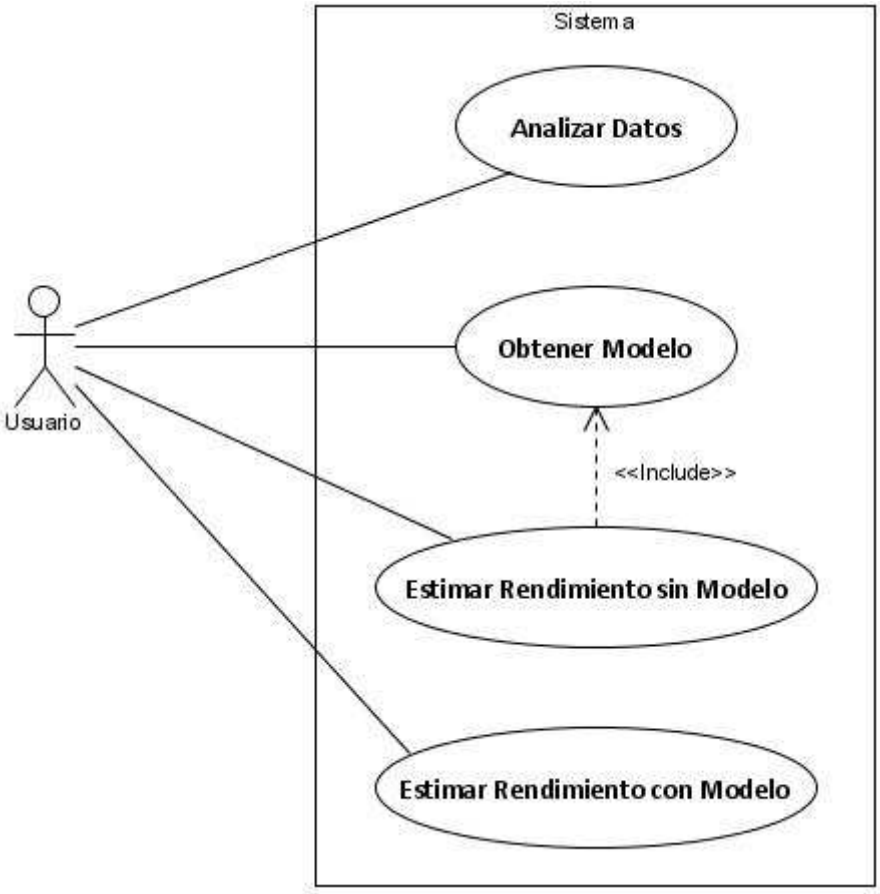


Figura 5.1 Diagrama de casos de usos.

Para mayor detalle del lector en el Apéndice E se detallan los casos de uso con su flujo principal y sus correspondientes flujos alternativos. En particular para cada caso de uso se tiene un Diagrama de Secuencia del Sistema que muestra la interacción entre los usuarios y el Sistema.

5.2 Diagramas de secuencias del Sistema

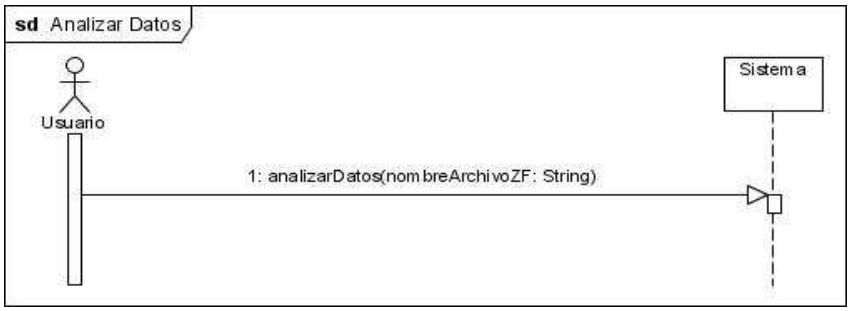


Figura 5.2 DSS Analizar Datos

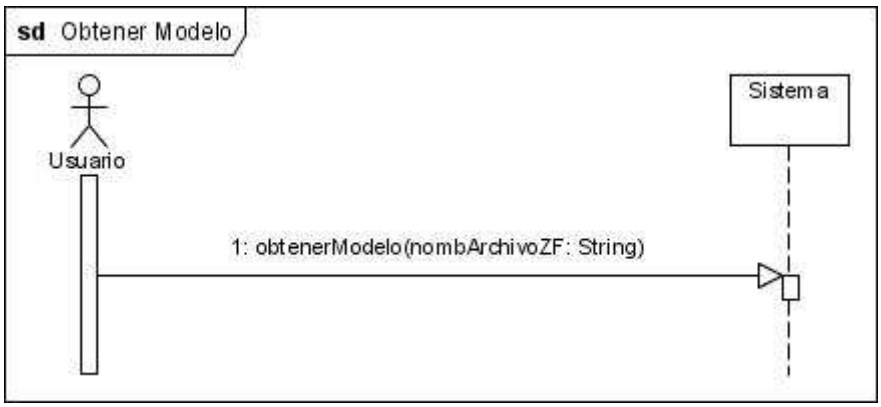


Figura 5.3 DSS Obtener modelo

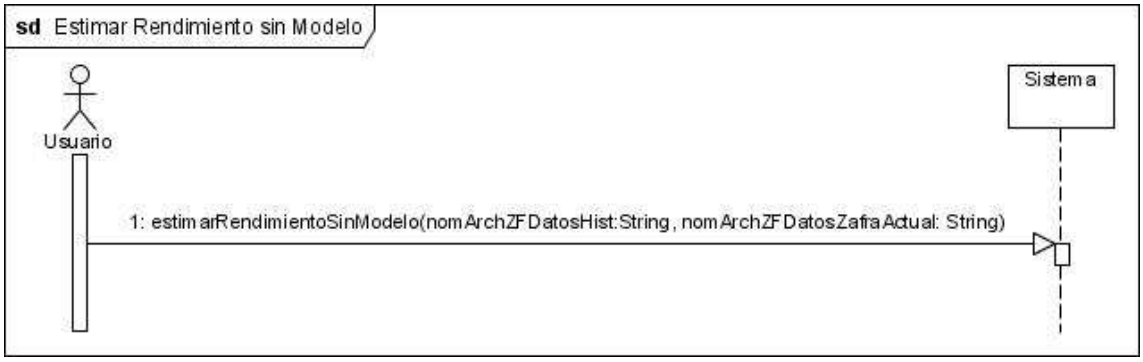


Figura 5.4 DSS Estimar rendimiento sin modelo

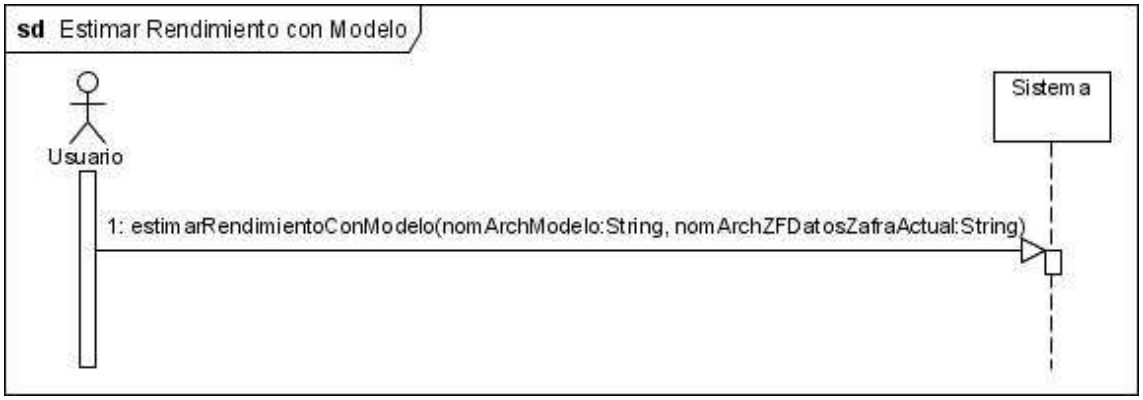


Figura 5.5 DSS Estimar rendimiento con modelo

5.3 Alcance del Sistema

El alcance del Sistema abarca el caso de uso Análisis de Datos. Los demás casos de uso son especificados, así como también tenidos en cuenta en el diseño pero no serán implementados. Las unidades de tiempo del proyecto fueron ampliamente consumidas en la investigación, dejando un escaso tiempo para la implementación teniendo en cuenta los límites del Sistema.

5.4 Diseño

A continuación se presenta el diseño de la biblioteca. Primero se presenta la Arquitectura de la misma, luego se presentan los diagramas de interacción involucrados en los casos de uso y finalmente se presenta el diagrama de clases de la solución.

5.4.1 Modelos regresivos de autocorrelación espacial

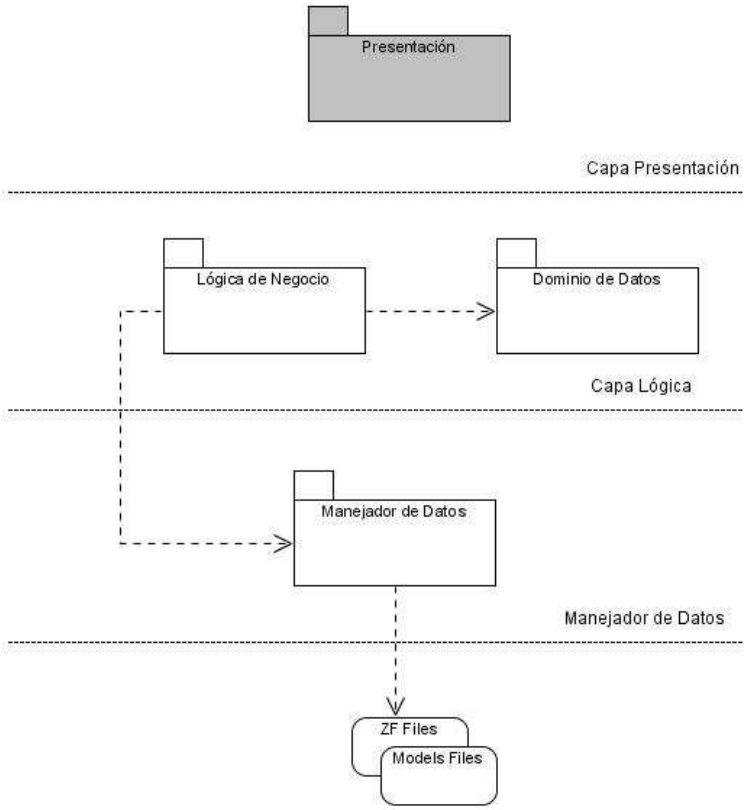


Figura 5.6 Arquitectura del sistema

La Arquitectura de la librería cuenta básicamente con dos capas, Capa Lógica y Capa Manejadora de datos. La capa de presentación que se denota en la figura es simplemente a modo de ilustración de cómo se vería la librería en una arquitectura en la cual fuese invocada simplemente para mostrar los resultados obtenidos. En la capa lógica se tiene dos componentes Lógica de Negocio y Dominio de datos.

Lógica de Negocio: es el encargado de realizar todos los contrastes, manejo de modelos de regresión y sus estimadores, y aplicación de algoritmos como por ejemplo el del cálculo de la matriz de correlación, matriz de vecindad, etc. Cuenta con interfaces para el manejo de cada contraste independiente de forma tal de poder implementar nuevos contrastes o algoritmos que puedan surgir en un futuro para contrastar determinada propiedad de los datos. Este componente es el encargado de manipular los diferentes modelos de regresión que soporta la biblioteca, así como también tiene bien definida una interfaz con los correspondientes estimadores para cada uno de ellos.

Dominio de Datos: Tiene como principal cometido la transformación de los datos en caso de ser necesario. En nuestro caso se le da una sola responsabilidad y es la de filtrar solamente las celdas de la chacra para las cuales se tiene datos para todas las variables declaradas junto con

sus coordenadas, dejando fuera del análisis aquellas celdas en las que no se tiene valores. Cualquier tratamiento previo al análisis de los datos que se deba realizar sobre los mismos se deberá hacer en este componente. Si se desea realizar una interpolación espacial sobre los datos previo al análisis o a la obtención del modelo se deberá implementar aquí.

En la capa Manejadora de datos se tiene un solo componente que es **Manejador de Datos**, este componente es el encargado de realizar la lectura y escritura de archivos. Tiene la capacidad de interpretar tanto los archivos de entrada como de salida y es el único que se encarga de realizar operaciones de entrada y salida de la biblioteca. La forma de comunicarse con él, es por medio de los datatypes especificados en el diagrama BussinesDatatypes, mencionado más adelante.

5.4.2 Diagrama de interacción

Análisis de Datos

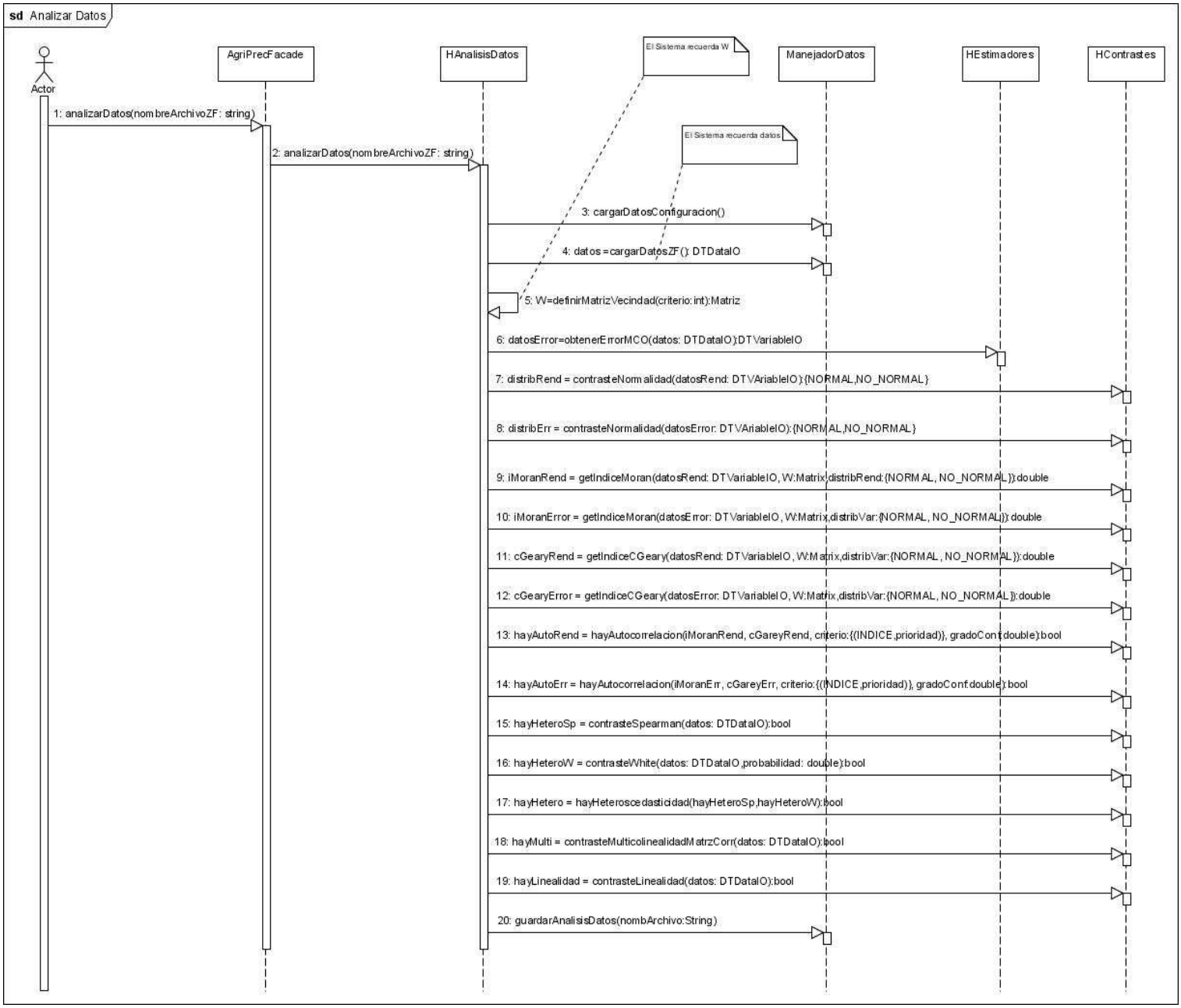


Figura 5.7 Diagrama de Interacción: Análisis de datos

Una vez que el usuario solicita el análisis de datos dando la ruta al archivo ZF con los datos, el Sistema realiza los diferentes contrastes y retorna un archivo con todos los resultados obtenidos.

Obtener Modelo

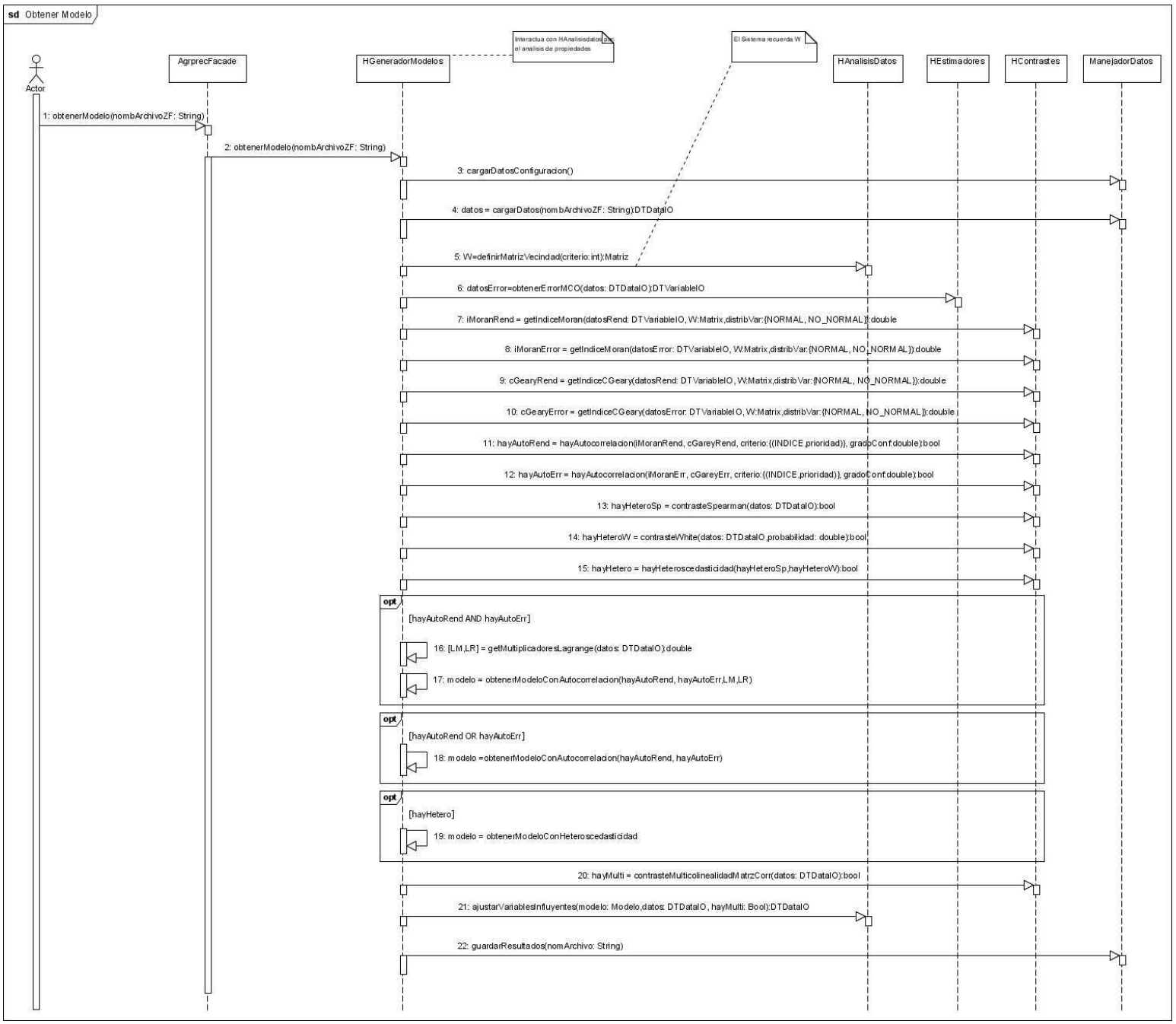


Figura 5.8 Diagrama de Interacción: Obtener modelo

Una vez que el usuario solicita la obtención del modelo para el rendimiento que se ajusta a los datos dando la ruta al archivo ZF con los datos históricos, el Sistema realiza los contrastes correspondientes para la selección del modelo que se ajusta a las propiedades de los mismos. Luego de obtenido el modelo se debe realizar el algoritmo para selección de variables influyentes, de esta forma la estimación del rendimiento solo tendrá en cuenta las variables cuyos valores influyen realmente en el rendimiento y descartará aquellas que no aportan nada al modelo. Finalmente se guardan los datos con los coeficientes de regresión del modelo así como también las variables influyentes y las que son quitadas del modelo.

Estimar Rendimiento Con Modelo

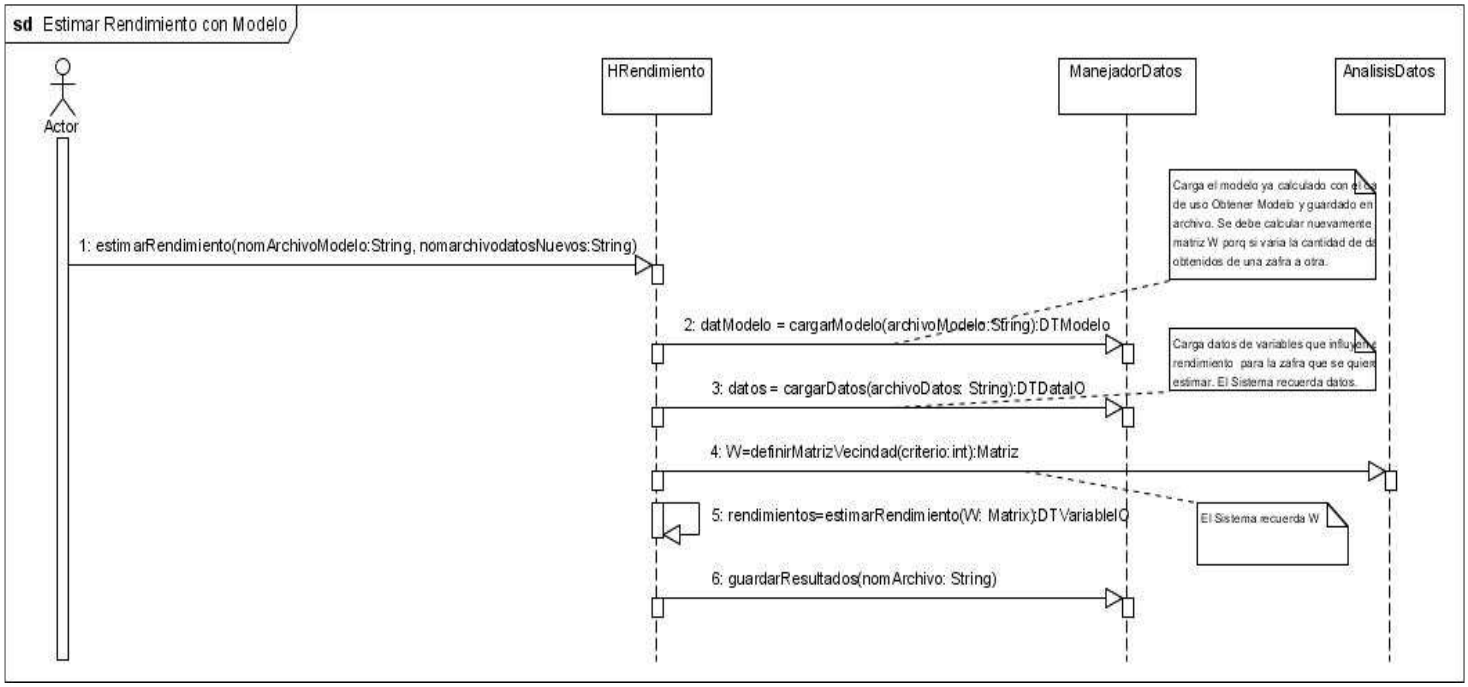


Figura 5.9 Diagrama de Interacción: Estimar rendimiento con modelo

Como alternativa al caso de uso estimar Rendimiento con modelo se tiene el caso de uso estimar rendimientos sin modelo, el cual incluye el caso de uso obtener modelo y luego estima los rendimientos levantando el modelo calculado.

5.4.3 Diagrama de clases

A continuación se presentan los diagramas de clase de cada paquete involucrado en la arquitectura de la biblioteca.

Lógica de Negocio

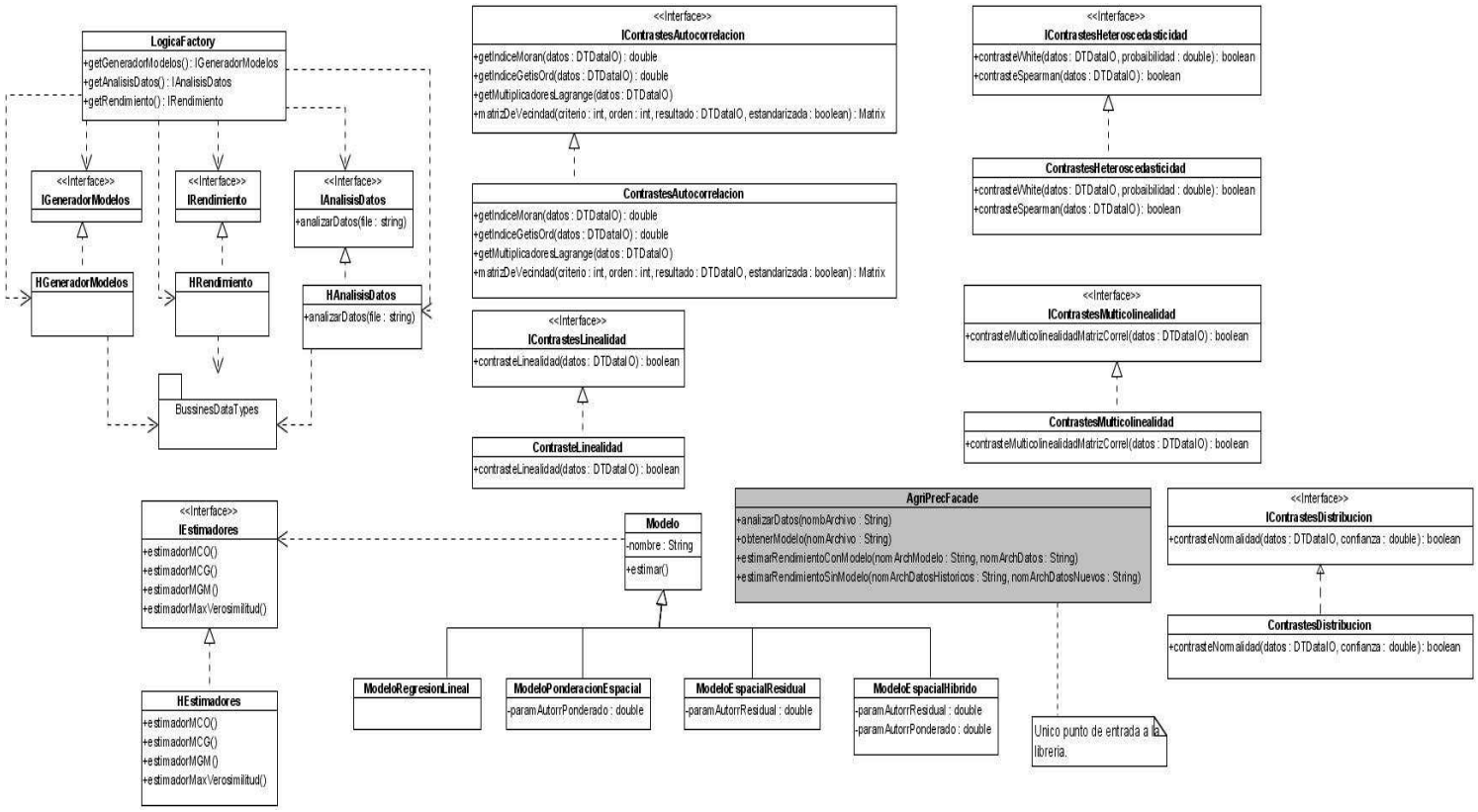


Figura 5.10 Diagrama de Clases - Lógica de Negocio

Dominio de Datos

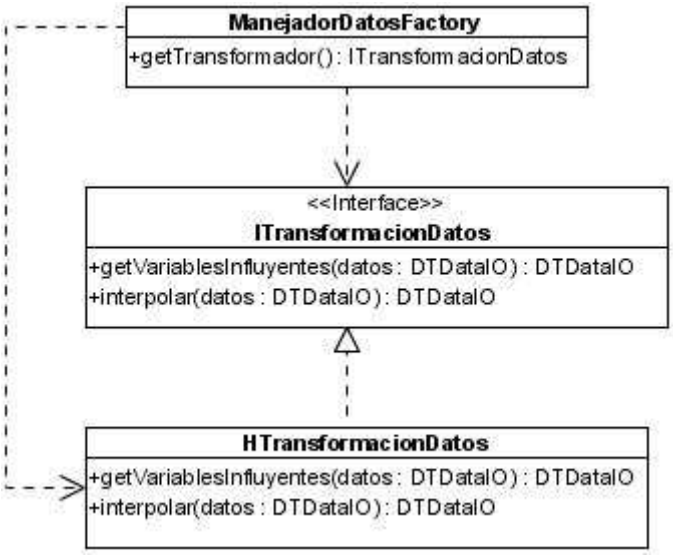


Figura 5.11 Diagrama de Clases – Dominio de Datos

Manejador de Datos

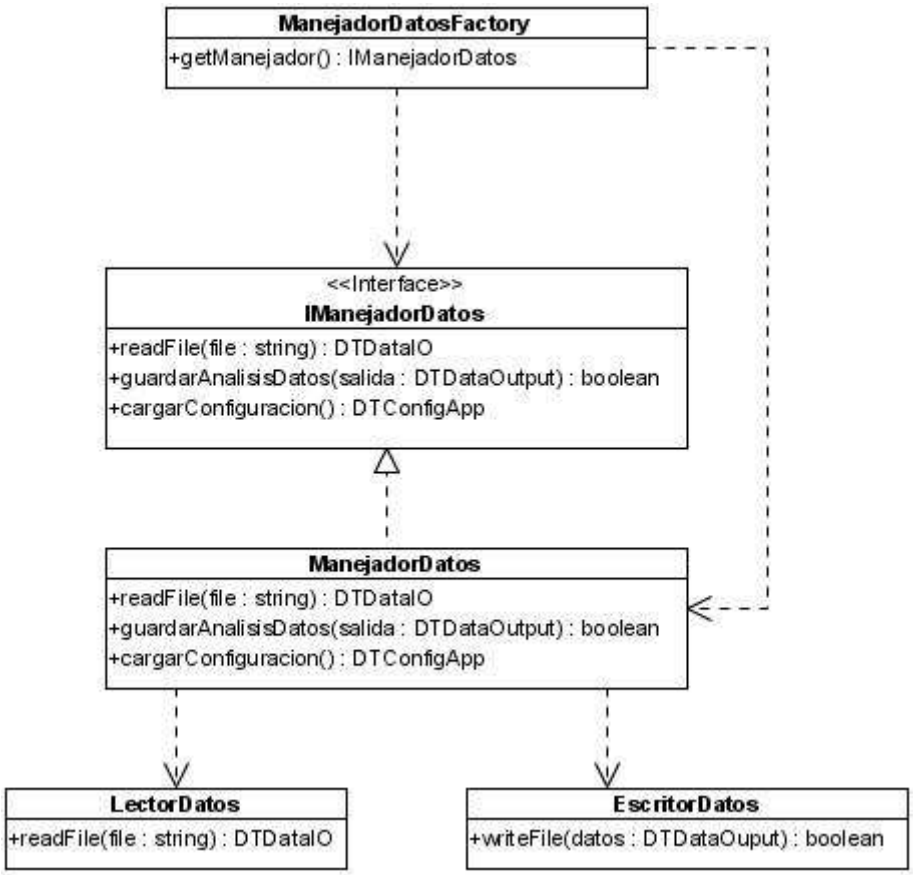


Figura 5.12 Diagrama de Clases – Manejador de Datos

BusinessDatatypes

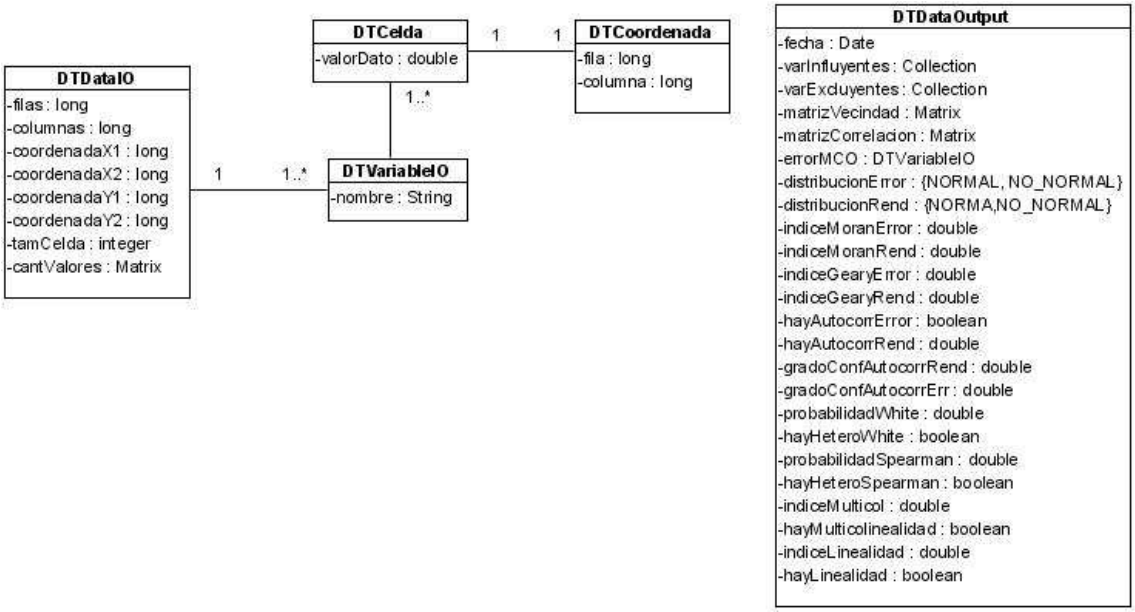


Figura 5.13 Diagrama de Datatypes

5.5 Implementación

Para la implementación de la biblioteca se utilizó el lenguaje de programación C# desarrollado y estandarizado por Microsoft como parte de su plataforma .NET. Este lenguaje es orientado a objetos y por tanto contiene toda la potencia que la programación orientada a objetos posee y la pone dentro de las tecnologías de punta a la hora del desarrollo de software.

Una de las principales ventajas que podemos encontrar en C# es que posee ya desarrollado y disponible las estructuras de datos básicas y que además se puede integrar dentro de cualquier aplicación que se desarrolle bajo la suite .Net de Microsoft. Otra ventaja es que tienen una administración de memoria muy eficiente y transparente para el usuario. Teniendo en cuenta que para el cliente no era importante que la biblioteca sea multiplataforma no se tuvo en cuenta a la hora de decidir en qué lenguaje realizarla.

El objetivo del Sistema es que sea utilizado como subsistema en otro/s sistema/s en donde se pueda invocar las funcionalidades y una vez obtenidos los datos poder mostrarlos de manera estética y amigable para el usuario. Por eso es que tanto la entrada como la salida del Sistema se presentan en archivos de texto plano, de esta forma cualquier sistema puede leer los archivos que cumplen con una determinada estructura y presentar de manera grafica al usuario final los resultados.

Para la implementación de las diferentes funcionalidades del Sistema se deben realizar operaciones con matrices tales como transponer una matriz, invertirla, multiplicarlas, etc. para esto se hace uso de una biblioteca “open source” llamada Math.Net Iridium desarrollada en C#. La versión utilizada de MathNet es la 2008.8.16.470. Esta biblioteca se puede adquirir gratuitamente en la siguiente URL http://www.math-net.de/.

Para presentar el diseño orientado a objetos del Sistema se utiliza el lenguaje UML 2.0, representando los diagramas mediante la herramienta Visual Paradigm 6.0 con la licencia community que es gratis.

Para el control de versionado tanto para el código fuente como para la documentación de la totalidad del proyecto se utiliza el SVN teniendo como servidor de repositorio el proporcionado en forma gratuita por el hosting http://unfuddle.com/. Como cliente SVN se utiliza TortoiseSVN el cual se puede adquirir gratuitamente en la siguiente URL http://tortoisesvn.net/.

5.5.1 Estructura de Datos

Cada archivo de entrada contiene los datos obtenidos para cada celda en la chacra, por lo que se puede ver a cada variable como una matriz en la que en cada celda se tiene o no datos de la misma. Cada celda tiene su correspondiente coordenada x y su coordenada y la cual la identifica como única en la matriz y en la chacra.

Teniendo en cuenta que pueden existir una gran cantidad de celdas sin valor para alguna de las variables es que se decidió tener una estructura propia para representar los datos de cada variable. No se representan los datos mediante matrices por el simple hecho que el 60% o 70% de las celdas no contienen datos para todas las variables a considerar en el análisis, lo cual ocuparía más memoria y dificultaría la realización de las diferentes operaciones y cálculos que realiza el Sistema. La estructura para mantener y trabajar con los datos es la presentada en el diagrama BussinesDatatypes en la que se tiene un datatype con los datos generales de la chacra, una colección de datatypes de tipo DTVariableIO que representan las variables exógenas con su

nombre, y la misma tiene una colección de DTCeldaIO con el valor de la celda y las coordenadas en la matriz original de la chacra.

Existen cálculos en los cuales se necesita tener los datos en forma matricial y realizar operaciones sobre estas matrices, para esto se utiliza la estructura Matrix de la biblioteca MathNet la cual tiene una estructura fácil de utilizar accediendo a cualquier elemento dando el índice de la fila y el de la columna directamente. Las matrices de correlación y de vecindad son almacenadas en este tipo de estructura y permite realizar operaciones como obtener el determinante, valores propios, etc.

5.5.2 Archivos de entrada y salida

5.5.2.1 Archivo de Entrada ZF

Este tipo de archivo es el requerido por el Sistema para leer los datos de entrada. El formato del mismo fue propuesto por el cliente y adaptado a las necesidades de todos los grupos que formaron parte del proyecto. Cualquiera sea la chacra de la cual se recolecten los datos deberá cumplir con el formato expuesto a continuación.

```
[PDT-SIGA: Zone File]
Rows: <Cantidad de filas de la matriz que representa la chacra>
Cols: <Cantidad de columnas de la matriz que representa la chacra >
CoordX: <Coordenada en X>
CoordY: <Coordenada en Y>
CellSize: <Tamaño de la celda>
Year: <Año en que se realiza la recolección de datos>

[Variables]
VarQty: <Cantidad de variables exógenas sumada la endógena>
Var1: <nombre variable endógena>; ENDOGENA
Var2: <nombre variable exógena>; EXOGENA
Var3: <nombre variable exógena>; EXOGENA
.
.
.


[Cells]
valVar1_11;valVar2_11;valVar3_11;valVar4_11;valVar5_11
valVar1_12;valVar2_12;valVar3_12;valVar4_12;valVar5_12
valVar1_13;valVar2_13;valVar3_13;valVar4_13;valVar5_13
valVar1_14;valVar2_14;valVar3_14;valVar4_14;valVar5_14
valVar1_21;valVar2_21;valVar3_21;valVar4_21;valVar5_21
valVar1_22;valVar2_22;valVar3_22;valVar4_22;valVar5_22
valVar1_23;valVar2_23;valVar3_23;valVar4_23;valVar5_23
valVar1_24;valVar2_24;valVar3_24;valVar4_24;valVar5_24
valVar1_31;valVar2_31;valVar3_31;valVar4_31;valVar5_31
valVar1_32;valVar2_32;valVar3_32;valVar4_32;valVar5_32
valVar1_33;valVar2_33;valVar3_33;valVar4_33;valVar5_33
..
..
..
```

La estructura general de los archivos de entrada ZF es la siguiente:

En el tag (`[PDT-SIGA: Zone File]`) se tiene la información del cabezal del archivo, la cantidad de filas, cantidad de columnas de la matriz que representa la chacra, las coordenadas X

e Y, del primer dato superior izquierdo, el tamaño de la celda y el año al cual corresponden los datos.
Ejemplo:

```
[PDT-SIGA: Zone File]
Rows: 5
Cols: 4
CoordX: 279275,985163911
CoordY: 6268178,1109817
CellSize: 10
Year: 2007
```

El tag [Variables] representa el punto de partida para la declaración de las variables a considerarse en el análisis con su respectivo tipo, el cual puede ser EXÓGENA o ENDÓGENA. Notar que la palabra reservada VarQty indica la cantidad total de variables.

Cada variable debe declararse con la palabra reservada Var seguido del numero, su nombre y el tipo de la siguiente manera `VarN : nombreVariable; {EXOGENA | ENDOGENA}`
Tener en cuenta que solo puede existir una variable endógena, que es el rendimiento.
Ejemplo:

```
[Variables]
VarQty: 5
Var1: rendimiento; ENDOGENA
Var2: a_curvature(Banda 1); EXOGENA
Var3: a_dem_corr(Banda 1); EXOGENA
Var4: a_orientac(Banda 1); EXOGENA
Var5: a_pendiente(Banda 1); EXOGENA
```

Finalmente se presentan los datos para cada celda seguidos del tag `[Cells]`. Cada celda queda definida en una línea. En caso de que no exista dato para alguna variable en una celda se debe indicar mediante la palabra reservada NaN (Not a Number)
Ejemplo:

```
[Cells]
valVar1_11;valVar2_11;valVar3_11;valVar4_11;valVar5_11
valVar1_12;valVar2_12;valVar3_12;valVar4_12;valVar5_12
valVar1_13;valVar2_13;valVar3_13;valVar4_13;valVar5_13
valVar1_14;valVar2_14;valVar3_14;valVar4_14;valVar5_14
valVar1_21;valVar2_21;valVar3_21;valVar4_21;valVar5_21
valVar1_22;valVar2_22;valVar3_22;valVar4_22;valVar5_22
valVar1_23;valVar2_23;valVar3_23;valVar4_23;valVar5_23
. . . . . .
```

5.5.2.2 Archivo de Salida ZF

El formato de salida es igual al archivo de entrada solo que en este caso la única variable escrita es el rendimiento resultado de la predicción resultante del caso de uso estimar rendimiento.
Ejemplo:

```
[PDT-SIGA: Zone File]
Rows: 5
Cols: 4
CoordX: 279275,985163911
CoordY: 6268178,1109817
CellSize: 10
Year: 2007

[Variables]
VarQty: 1
Var1: rendimiento; ENDOGENA

[Cells]
valVar1_11
```

```
valVar1_12
valVar1_13
valVar1_14
valVar1_21
valVar1_22
valVar1_23
valVar1_24
valVar1_31
valVar1_32
valVar1_33
valVar1_34
valVar1_41
valVar1_42
valVar1_43
valVar1_44
valVar1_51
valVar1_52
valVar1_53
valVar1_54
```

5.5.2.3 Archivo de Salida de Modelo con Coeficientes

```
[MODELO]
CODIGO MODELO: 2
NOMBRE MODELO: REGRESIÓN ESPACIAL EN EL RENDIMIENTO

[COEFICIENTES REGRESIVOS]
Coef1
Coef2
Coef3
.
.
.
Coefn

[COEFICIENNTE AUTORREGRESIVO ERROR]
NOMBRE: Lambda
VALOR: x.x

[COEFICIENNTE AUTORREGRESIVO RENDIMIENTO]
NOMBRE: Rho
VALOR: y.y

[COEFICIENNTE HETEROSCEDASTICO]
NOMBRE: Alpha
VALOR: z.z
```

Los coeficientes autorregresivos y el heteroscedástico son opcionales dependiendo del modelo seleccionado por el caso de uso seleccionar modelo. Si el modelo es el modelo de regresión lineal clásico estos coeficientes no forman parte del archivo de salida.

5.5.2.4 Archivo de Salida Análisis de Datos

Se presenta la estructura del archivo de salida para el análisis de datos.

```
[General]
Fecha: <Fecha en la que se realiza el analisis>
ArchivoEntrada: <Nombre del archivo de entrada ZF>

[IndiceMoran]
Error : <Índice de Moran calculado para el vector de errores>
Rendimiento: <Índice de Moran calculado para el vector de rendimientos>

[IndiceGeary]
Error: <Índice de Geary calculado para el vector de errores>
```

```
Rendimiento: <Índice de Geary calculado para el vector de rendimientos>

[Autocorrelacion]
Hay Autocorrelacion Error: <True si hay autocorrelación en el error, False en
otro caso>
Hay Autocorrelacion Rendimiento: <True si hay autocorrelación en el
rendimiento, False en otro caso>
GradoConfianzaError: <Grado de confianza del contraste aplicado a los errores>
GradoConfianzaRendim: <Grado de confianza del contraste aplicado a los
rendimientos>

[ContrasteWhite]
Probabilidad: <Grado de confianza del contraste de White>
hayHeterocedasticidad: <True si hay heteroscedasticidad en los datos para el
contraste de White, False en otro caso>

[ContrasteSpearman]
Probabilidad: <Grado de confianza del contraste de Spearman>
hayHeterocedasticidad: <True si hay heteroscedasticidad en los datos para el
contraste de Spearman, False en otro caso>

[Multicolinealidad]
//indice multicolinealidad: < 10 OK, 10-30 moderada, >30 Alta
hayMulticolinealidad: <Indica el grado de multicolinealidad
{INEXISTENTE,MEDIA,ALTA}>
Indice: <Valor calculado del indice de condicionamiento>

[Linealidad]
hayLinealidad: <True si el contraste de linealidad asegura que los datos se
comportan como una función lineal, false en otro caso>
Indice: <Índice calculado del coeficiente de determinación>

[Normalidad]
Error: <Indica si la distribución del vector de errores es NORMAL o NO_NORMAL>
Rendimiento: <Indica si la distribución del vector de rendimientos es NORMAL o
NO_NORMAL>

[MatrizCorrelacion]
Filas: <Cantidad de filas de la matriz de Correlación de las variables>
Columnas: <Cantidad de columnas de la matriz de Correlación de las variables>
Determinante: <Valor calculado del determinante de la matriz de correlacion>

[CellsMC]
<Se presentan en formato matricial los datos de la matriz de correlacion
calculada>

[ErrorVectorMCO]
<Se presentan los valores calculados del error en el estimado de MCO>

[MatrizW]
Filas: <Cantidad de filas de la matriz de vecindad>
Columnas: <Cantidad de columnas de la matriz de vecindad>
Estandard: <True si esta estandarizada o False en caso contrario>

[CellsW]
<Se presentan forma matricial los valores de la matriz de vecindad calculada>
```

5.5.3 Archivo de Configuración

La biblioteca cuenta con un archivo de configuración en lenguaje XML en el cual se setean los parámetros básicos para administrar la ejecución de la misma. Cada tag del archivo tiene su correspondiente comentario para mayor entendimiento, el siguiente archivo es el actual archivo de configuración de la biblioteca pero puede ser modificado por el usuario en cualquier momento.

```
<?xml version="1.0" encoding="UTF-8"?>

<config>
```

```
<matrizVecindad>
       <!-- Criterio para el calculo de la matriz de vecindad. Los valores posibles son:
       QUEEN, BISHOP, ROOK -->
    <criterio>ROOK</criterio>
</matrizVecindad>

<autocorrelacion>
<!-- Coleccion de incices que se pueden  calcular con su criterio de prioridad para la
toma de desicion de si existe AUTOCORRELACION en los datos o no -->
       <indice>
              <nombre>GEARY</nombre>
              <prioridad>1</prioridad>
              <gradoConfianza>0,9</gradoConfianza>
       </indice>
       <indice>
              <nombre>MORAN</nombre>
              <prioridad>2</prioridad>
              <gradoConfianza>0,9</gradoConfianza>
       </indice>
</autocorrelacion>

<heterocedasticidad>
<!-- Coleccion de incices que se pueden  calcular con su criterio de prioridad para la
toma de desicion de si existe AUTOCORRELACION en los datos o no -->
        <contrastes>
              <nombre>SPEARMAN</nombre>
              <gradoConfianza>0.9</gradoConfianza>

              <nombre>WHITE</nombre>
              <gradoConfianza>0.9</gradoConfianza>
       </contrastes>
</heterocedasticidad>

<normalidad>
<!-- nivel de significancia para el contraste de normalidad -->
       <nivel_significancia>0.8</nivel_significancia>
</normalidad>

<linealidad>
       <!-- grado de confianza con elque  se4 -->
        <grado_confianza>0.01</grado_confianza>
</linealidad>

</config>
```

5.5.4 Uso de la biblioteca

Para utilizar la biblioteca se debe referenciar el archivo LibreriaAgriPrec.dll en el proyecto en el cual se desee utilizar la misma. Tener en cuenta que el archivo config.xml debe de estar en la misma carpeta en donde se encuentra el archivo dll de la biblioteca.

Una vez referenciada la biblioteca, dado que por definición de diseño la misma cuenta con una clase Facade que contiene las operación disponibles por la biblioteca, se debe instanciar la misma y luego simplemente servirse de la operaciones disponibles (por mas detalle ver diagrama de clases en donde se detallan las operaciones disponibles en el AgriPrecFacade).

Se presenta a continuación un ejemplo de invocación de la operación analizar datos:

```
AgriPrecFacade argPrec = AgriPrecFacade.getInstance();
resultado = argPrec.analizarDatos(txtFielPath.Text);
```

5.6 Testeo

El objetivo de la siguiente sección es el de realizar las validaciones de los contrastes con los que cuenta la aplicación y de verificar que los distintos contrastes se comporten de manera esperada para datos de entrada que se sabe cumplen con determinadas propiedades. Dichas verificaciones

le dan robustez a los componentes de la biblioteca para cuando la misma sea utilizada con datos reales de las chacras que a priori no se sabe las propiedades con las que cumple.

Se realizó un testeo unitario de cada contraste para la verificación de los mismos. A continuación se mencionan los contrastes que se verificaron, los archivos que se usaron y los resultados obtenidos. No se presentan los juegos de datos propiamente dichos, debido a la extensión de los mismos, mostrarlos aquí no haría más que aumentar sin sentido este documento, por este motivo se adjuntarán por medio electrónico.

5.6.1 Contrastes de Autocorrelación

Contraste I Moran
Realiza el contraste de autocorrelación del índice I de Moran, dado: los datos, la matriz de vecindad W, y el tipo de distribución que siguen los datos (normal o randómico). Retornando la probabilidad de que se verifique la autocorrelación y el índice.

Requerimientos Funcionales

Entrada	Salida esperada	Salida obtenida
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF Se tomó el criterio de vecindad ROOK.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,613981169132151. Prob error: 100%. I de Moran en el Rendimiento: 0,787973974165419. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Pot1Trigo06.ZF Se tomó el criterio de vecindad ROOK.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,710916544253709. Prob error: 100%. I de Moran en el Rendimiento: 0,832095502560679. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Se tomó el criterio de vecindad ROOK.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,156489604300655. Prob error: 90,27639574421481%. I de Moran en el Rendimiento: 0,559057107353452. Prob rendimiento: 99,9875876504636%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF Se tomó el criterio de vecindad QUEEN.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,613981169132151. Prob error: 100%. I de Moran en el Rendimiento: 0,787973974165419. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)

Se utilizó el siguiente archivo: Pot1Trigo06.ZF Se tomó el criterio de vecindad QUEEN.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,638049979158066. Prob error: 100%. I de Moran en el Rendimiento: 0,786184509776972. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Se tomó el criterio de vecindad QUEEN.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,0836059811321127. Prob error: 89,667914129204274 %. I de Moran en el Rendimiento: 0,430446702753972. Prob rendimiento: 99,9992724265968 %. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF Se tomó el criterio de vecindad BISHOP.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,537626862708915. Prob error: 100%. I de Moran en el Rendimiento: 0,741534076876441. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Pot1Trigo06.ZF Se tomó el criterio de vecindad BISHOP.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: 0,563905383000649. Prob error: 100%. I de Moran en el Rendimiento: 0,739468268731901. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)

Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Se tomó el criterio de vecindad BISHOP.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	I de Moran en el Error: -0,0105353654605882. Prob error: 90,27639574421481%. I de Moran en el Rendimiento: 0,26432493014631. Prob rendimiento: 59,077760878765984 %. Se verificó la existencia de autocorrelación solamente en el error.

Contraste C Geary

Realiza el contraste de autocorrelación del índice C de Geary, dado: los datos, la matriz de vecindad W, y el tipo de distribución que siguen los datos (normal o randómico). Retornando la probabilidad de que se verifique la autocorrelación y el índice.

Para el testeo de ambos contrastes se usaron los siguientes archivos:

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: Pot1Trigo06.ZF Se tomó el criterio de vecindad ROOK.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,000167143248930978. Prob error: 100%. C de Geary en el Rendimiento: 0,000208790286183202. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF Se tomó el criterio de vecindad ROOK.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,000769799453899637. Prob error: 100%. C de Geary en el Rendimiento: -0,00000934676404627167. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)

Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Se tomó el criterio de vecindad ROOK.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,029051450658357209. Prob error: 99,278373681926824%. C de Geary en el Rendimiento: 0,001120684277615017. Prob rendimiento: 99,7227669344424%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF Se tomó el criterio de vecindad QUEEN.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,000842473721494632. Prob error: 100%. C de Geary en el Rendimiento: 0,00018193312266957. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Pot1Trigo06.ZF Se tomó el criterio de vecindad QUEEN.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,000872070811054649. Prob error: 100%. C de Geary en el Rendimiento: 0,000367025773874998. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Se tomó el criterio de vecindad QUEEN.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,0733801321940702. Prob error: 99,573448997167036%. C de Geary en el Rendimiento: 0,00197486714182661. Prob rendimiento: 99,822925867713019%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)

Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF Se tomó el criterio de vecindad BISHOP.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,00106109977310399. Prob error: 100%. C de Geary en el Rendimiento: 0,000523105563601834. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: Pot1Trigo06.ZF Se tomó el criterio de vecindad BISHOP.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: 0,0000314335673721034. Prob error: 100%. C de Geary en el Rendimiento: 0,000709817434547774. Prob rendimiento: 100%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Se tomó el criterio de vecindad BISHOP.	Que el contraste verifique la presencia de autocorrelación con una probabilidad del 90%.	C de Geary en el Error: -0,165856230212226. Prob error: 99,542211644540557%. C de Geary en el Rendimiento: 0,00417688580619063. Prob rendimiento: 99,543620517590936%. Se verificó la existencia de autocorrelación en ambas variables (error y rendimiento)

5.6.2 Contrastes de Heteroscedasticidad

Contraste White

Realiza el contraste de heteroscedasticidad de White, dado: los datos, y la probabilidad con la cual se contrastará si se verifica o no la heteroscedasticidad, retornando true en caso que se verifique el contraste y false en caso contrario.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: PotreroDatosTest2.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de White con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest3.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de White con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest4.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de White con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest5.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de White con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest6.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de White con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest7.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de White con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True

Contraste Spearman

Realiza el contraste de heteroscedasticidad de Spearman, dado: los datos, y la probabilidad con la cual se contrastará si se verifica o no la heteroscedasticidad, retornando true en caso que se verifique el contraste y false en caso contrario.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: PotreroDatosTest2.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de Spearman con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest3.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de Spearman con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest4.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de Spearman con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest5.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de Spearman con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest6.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de Spearman con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True
Se utilizó el siguiente archivo: PotreroDatosTest7.ZF	Que el contraste verifique la presencia de heterocedasticidad usando el contraste de Spearman con una probabilidad del 80%.	Probabilidad > 0,8 hayHeterocedasticidad: True

5.6.3 Contrastes de Distribución

Contraste de Asimetría

Realiza el contraste de normalidad teniendo en cuenta el coeficiente de asimetría (CA), en donde dada la variable objeto de estudio, y el nivel de significación simétrica, se obtiene un valor que representa si la distribución es simétrica o no, o lo que es lo mismo si la distribución es normal o no.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: PotreroDatosTest5.ZF	Que el contraste verifique la presencia de una distribución normal en el error con una probabilidad mayor 80%, según el contraste de asimetría.	Probabilidad en el Error: 86,521548427370609%. Se verificó que la distribución del **error** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest3.ZF	Que el contraste verifique la presencia de una distribución normal en el error con una probabilidad mayor 80%, según el contraste de asimetría.	Probabilidad en el Error: 91,020177795501844%. Se verificó que la distribución del **error** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest3.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad mayor 80%, según el contraste de asimetría.	Probabilidad en el Rendimiento: 100%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest8.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad mayor 80%, según el contraste de asimetría.	Probabilidad en el Rendimiento: 100%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.

Contraste Curtosis

Realiza el contraste de normalidad teniendo en cuenta el coeficiente de curtosis (o apuntamiento) (CAp), en donde dada la variable objeto de estudio, y el nivel de significación de curtosis, se obtiene un valor que representa el rechazo o no (dependiendo del nivel de significación de curtosis) de la hipótesis de que "la distribución tiene curtosis cero" (en ese caso la distribución no será normal).

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: PotreroDatosTest6.ZF	Que el contraste verifique la presencia de una distribución normal en el error con una probabilidad menor al 20%, según el contraste de curtosis.	Probabilidad en el Error: 0,1%. Se verificó que la distribución del **error** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest9.ZF	Que el contraste verifique la presencia de una distribución normal en el error con una probabilidad menor al 20%, según el contraste de curtosis.	Probabilidad en el Error: 05,4474407618432696% Se verificó que la distribución del **error** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest9.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad menor al 20%, según el contraste de curtosis.	Probabilidad en el Rendimiento: 05,4474408064377089%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest3.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad menor al 20%, según el contraste de curtosis.	Probabilidad en el Rendimiento: 18,60911454630092%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest8.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad menor al 20%, según el contraste de curtosis.	Probabilidad en el Rendimiento: 19,378650221993654%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.

Realiza el contraste de normalidad teniendo en cuenta los dos contrastes anteriores combinándolos en un estadístico, en donde dada la variable objeto del estudio, y el nivel de significación combinado, se obtiene un valor que representa si se rechaza o no que la distribución es simétrica y/o que tiene curtosis nula, y en consecuencia, se rechaza la hipótesis de normalidad.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: PotreroDatosTest9.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad menor al 80%, según el contraste de combinado.	Probabilidad en el Rendimiento: 100%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest9.ZF	Que el contraste verifique la presencia de una distribución normal en el error con una probabilidad menor al 80%, según el contraste de combinado.	Probabilidad en el Error: 100%. Se verificó que la distribución del **error** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest6.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad menor al 80%, según el contraste de combinado.	Probabilidad en el Rendimiento: 100%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest6.ZF	Que el contraste verifique la presencia de una distribución normal en el error con una probabilidad menor al 80%, según el contraste de combinado.	Probabilidad en el Error: 100%. Se verificó que la distribución del **error** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest10.ZF	Que el contraste verifique la presencia de una distribución normal en el rendimiento con una probabilidad menor al 80%, según el contraste de combinado.	Probabilidad en el Rendimiento: 100%. Se verificó que la distribución del **rendimiento** sigue una distribución normal.
Se utilizó el siguiente archivo: PotreroDatosTest10.ZF	Que el contraste verifique la presencia de una distribución normal en el error con una probabilidad menor al 80%, según el contraste de combinado.	Probabilidad en el Error: 100%. Se verificó que la distribución del **error** sigue una distribución normal.

5.6.4 Contrastes de Multicolinealidad

Contraste Multicolinealidad

Realiza el contraste de multicolinealidad, dado: los datos, y la matriz de correlaciones, retornando un enumerado que indica el nivel de multicolinealidad existente (baja, moderada o alta) y un índice de multicolinealidad.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF	Si el indice multicolinealidad: < 10 No Existe, 10-30 Moderada, >30 Alta	Multicolinealidad: INEXISTENTE Indice: 3,55936988422355
Se utilizó el siguiente archivo: Pot1Trigo06.ZF	Si el indice multicolinealidad: < 10 No Existe, 10-30 Moderada, >30 Alta	Multicolinealidad: ALTA Indice: 34,80191871742
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF	Si el indice multicolinealidad: < 10 No Existe, 10-30 Moderada, >30 Alta	Multicolinealidad: ALTA Indice: 82,8229946783696

5.6.5 Contraste de Linealidad

Contraste Linealidad

Realiza el contraste de linealidad, dado: los datos, y el porcentaje de confianza de linealidad (obtenido del archivo de configuración), retornando el índice de linealidad y in boolean que toma el valor true en caso que se verifique el contraste y false en caso contrario.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: Pot1Trigo06.ZF	Que el contraste verifique la presencia de linealidad usando el contraste de Linealidad con una probabilidad mayor al 90%.	Probabilidad > 0,9 hayLinealidad: True Indice: 0,626566652703956
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF	Que el contraste verifique la presencia de linealidad usando el contraste de Linealidad con una probabilidad mayor al 90%.	Probabilidad > 0,9 hayLinealidad: True Indice: 0,663066444875037
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF	Que el contraste verifique la presencia de linealidad usando el contraste de Linealidad con una probabilidad mayor al 90%.	Probabilidad > 0,9 hayLinealidad: True Indice: 0,907671468505051

Se utilizó el siguiente archivo: PotreroDatosTest2.ZF	Que el contraste verifique la presencia de linealidad usando el contraste de Linealidad con una probabilidad mayor al 90%.	Probabilidad > 0,9 hayLinealidad: True Indice: 0,830042549272938
Se utilizó el siguiente archivo: PotreroDatosTest6.ZF	Que el contraste verifique la presencia de linealidad usando el contraste de Linealidad con una probabilidad mayor al 90%.	NO se detecta linealidad Probabilidad < 0,9 hayLinealidad: False Indice: 0,00104724679130433

5.6.6 Test de Matriz de Vecindad

Matriz de Vecindad

Se realiza el test de la matriz de vecindad dado: los datos, el criterio que seguirá la matriz, el orden de vecindad (en nuestro caso siempre uno) y un parámetro boolean que indica si es estandarizada o no.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Criterio QUEEN Estandarizada	Matriz de 20×20 Estandarizada y siguiendo el criterio QUEEN	Matriz de 20×20 Estandarizada, siguiendo el criterio QUEEN y con los datos correctos.
Se utilizó el siguiente archivo: PotreroDatosTest1.ZF Criterio QUEEN No estandarizada	Matriz de 20×20 NO estandarizada y siguiendo el criterio QUEEN	Matriz de 20×20 NO estandarizada, siguiendo el criterio QUEEN y con los datos correctos.
Se utilizó el siguiente archivo: PotreroDatosTest2.ZF Criterio BISHOP Estandarizada	Matriz de 6×6 Estandarizada y siguiendo el criterio BISHOP	Matriz de 6×6 Estandarizada, siguiendo el criterio BISHOP y con los datos correctos.
Se utilizó el siguiente archivo: PotreroDatosTest2.ZF Criterio BISHOP No estandarizada	Matriz de 6×6 NO estandarizada y siguiendo el criterio BISHOP	Matriz de 6×6 NO estandarizada, siguiendo el criterio BISHOP y con los datos correctos.

Se utilizó el siguiente archivo: PotreroDatosTest7.ZF Criterio ROOK Estandarizada	Matriz de 10×10 Estandarizada y siguiendo el criterio ROOK	Matriz de 10×10 Estandarizada, siguiendo el criterio ROOK y con los datos correctos.
Se utilizó el siguiente archivo: PotreroDatosTest7.ZF Criterio ROOK No estandarizada	Matriz de 10×10 NO estandarizada y siguiendo el criterio ROOK	Matriz de 10×10 NO estandarizada, siguiendo el criterio ROOK y con los datos correctos.

5.6.7 Testing de integración

Testing de Integración

Se realiza un test de integración con todos los contrastes que se hicieron en forma unitaria chequeando la coherencia de los resultados, esto es por ejemplo que si el contraste de heterocedasticidad en White da positivo lo de también el del Spearman; o que si se da autocorrelación aplicando el I de Moran también se cumpla esta propiedad usando el C de Geary.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: Pot1Trigo06.ZF Empleando el criterio QUEEN para autocorrelación.	Se espera que los resultados de los contrastes mantengan una coherencia ya que se trata de datos reales de una chacra, esto es: No Normalidad Autocorrelación Heteroscedasticidad	Se detecta la No Normalidad Autocorrelación Heteroscedasticidad
Se utilizó el siguiente archivo: Pot1Trigo06.ZF Empleando el criterio BISHOP para autocorrelación.	Se espera que los resultados de los contrastes mantengan una coherencia ya que se trata de datos reales de una chacra, esto es: No Normalidad Autocorrelación Heteroscedasticidad	Se detecta la No Normalidad Autocorrelación Heteroscedasticidad
Se utilizó el siguiente archivo: PotreroDatosTest2.ZF	Se espera que los resultados de los dos contrastes de heteroscedasticidad verifiquen dicha propiedad de los datos.	Se detecta la Heteroscedasticidad tanto en contraste de White como de Spearman.
Se utilizó el siguiente archivo: PotreroDatosTest3.ZF	Se espera que los resultados de los dos contrastes de heteroscedasticidad verifiquen dicha propiedad de los datos.	Se detecta la Heteroscedasticidad tanto en contraste de White como de Spearman.

5.6.8 Testing de performance

Testing de performance

Se realiza un test de performance sobre las funcionalidades de la biblioteca recorriendo todos los contrastes implementados. Como entrada se especifica el archivo utilizado y las características del hardware en el cual se probó la biblioteca; se brinda como salida los tiempos utilizados en la ejecución de cada prueba.

Entrada	**Salida esperada**	**Salida obtenida**
Se utilizó el siguiente archivo: Pot1Trigo06.ZF **Cantidad de Celdas: 5609** procesador: Intel(R) Pentium(R) M 1.73 GHz RAM: 504 MB	Se espera que la conclusión de los contrastes implementados de la biblioteca sea efectuada en un tiempo razonable.	tiempo inicial: 18:43:15:500 tiempo final: 18:43:29:546 TOTAL: 14,046875 segundos
Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF **Cantidad de Celdas: 4118** procesador: Intel(R) Pentium(R) M 1.73 GHz RAM: 504 MB	Se espera que la conclusión de los contrastes implementados de la biblioteca sea efectuada en un tiempo razonable.	tiempo inicial: 18:41:34:390 tiempo final: 18:41:50:203 TOTAL: 15,8125 segundos
Se utilizó el siguiente archivo: Pot1Maiz07.ZF **Cantidad de Celdas: 5609** procesador: Intel(R) Pentium(R) M 1.73 GHz RAM: 504 MB	Se espera que la conclusión de los contrastes implementados de la biblioteca sea efectuada en un tiempo razonable.	tiempo inicial: 20:8:46:15 tiempo final: 20:9:0:453 TOTAL: 14,4375
Se utilizó el siguiente archivo: Pot1Trigo06.ZF **Cantidad de Celdas: 5609** procesador: Intel(R) Celeron(R) CPU 2.54 GHz RAM: 192 MB	Se espera que la conclusión de los contrastes implementados de la biblioteca sea efectuada en un tiempo razonable.	tiempo inicial: 17:58:50:203 tiempo final: 17:59:16:937 TOTAL: 26,734375 segundos

Se utilizó el siguiente archivo: Potrero1PrietoNuevo.ZF **Cantidad de Celdas: 4118** procesador: Intel(R) Celeron(R) CPU 2.54 GHz RAM: 192 MB	Se espera que la conclusión de los contrastes implementados de la biblioteca sea efectuada en un tiempo razonable.	tiempo inicial: 18:1:31:953 tiempo final: 18:1:58:984 TOTAL: 27,03125 segundos
Se utilizó el siguiente archivo: Pot1Maiz07.ZF **Cantidad de Celdas: 5609** procesador: Intel(R) Celeron(R) CPU 2.54 GHz RAM: 192 MB	Se espera que la conclusión de los contrastes implementados de la biblioteca sea efectuada en un tiempo razonable.	tiempo inicial: 20:58:50:203 tiempo final: 20:59:17:937 TOTAL: 27,56375 segundos

5.6.9 Origen de los datos del test

Los datos para realizar los test fueron provistos en algunos casos por el cliente, y en la mayoría de los casos extraídos de Internet. A continuación se brinda una descripción detallada del origen de dichos archivos.

Nombre Archivo	Origen
Pot1Trigo06.ZF	Provisto por el cliente
Potrero1PrietoNuevo.ZF	Provisto por el cliente
PotreroDatosTest1.ZF	http://www.monografias.com/trabajos30/regresion-multiple/regresion-multiple.shtml
PotreroDatosTest10.ZF	http://www.ucm.es/info/ecocuan/mjm/ecoaplimj/ archivo Simulados[1].xls
PotreroDatosTest3.ZF	datos extraidos del archivo: sacado de 1_nosferi.pdf
PotreroDatosTest8.ZF	datos extraidos del archivo: sacado de 1_nosferi.pdf
PotreroDatosTest4.ZF	datos extraidos del archivo: sacado de 1_nosferi.pdf
PotreroDatosTest5.ZF	datos extraidos del archivo: 1_est2-practica5R.pdf
PotreroDatosTest9.ZF	http://www.wellesley.edu/Chemistry/stats/lesson4.html archivo bell1.xls
PotreroDatosTest6.ZF	http://www.ucm.es/info/ecocuan/mjm/ecoaplimj/ archivo EstadisticaDescriptiva[1].xls
PotreroDatosTest7.ZF	http://www.monografias.com/trabajos30/regresion-multiple/regresion-multiple.shtml
PotreroDatosTest2.ZF	http://www.monografias.com/trabajos30/regresion-correlacion/regresion-correlacion.shtml

Capítulo 6

Resultados

El proyecto tiene dos objetivos bien definidos ya mencionados en secciones anteriores que son: la investigación de técnicas de predicción que permitan llegar a una solución al problema planteado por el cliente y la definición de una solución informática general que abarca el diseño y la arquitectura en su totalidad y la implementación del caso de uso principal *Análisis de Datos*. Es por esto que dividimos los resultados obtenidos en estos dos puntos.

6.1 Resultados de la Investigación

Se analizaron las diferentes metodologías que se pueden aplicar para predecir rendimientos en una chacra en base a datos históricos recolectados en la misma chacra, para variables que influyen en el rendimiento y del mismo rendimiento.

Como primer resultado se obtuvo un documento de estado del arte para las diferentes metodologías utilizadas para la predicción de rendimiento en base a datos temporales. Los modelos VAR fueron los elegidos para brindar una solución al problema teniendo en cuenta que los datos con los que se iba a contar de las chacras, podían cumplir con los requerimientos que el modelo VAR exigía. Dado que luego se vio que los datos con los que se contaba no eran suficientes para aplicar esta metodología, se debió incurrir en el análisis de datos de corte transversal y no temporal.

En este punto la investigación viró a encontrar metodologías de predicción aplicables a datos de corte transversal en el espacio. Frente a esta problemática se comenzó a investigar las herramientas que la Econometría nos brinda para el análisis espacial de los datos. Dado que la Econometría tradicional no tiene en cuenta determinadas propiedades espaciales que los datos cumplen, se comienza a analizar más profundamente la Econometría Espacial. El resultado de dicha investigación es un documento del estado del arte que detalla las herramientas de la Econometría Espacial y cuales serían necesarias utilizar para solucionar el problema del cliente. Los modelos de regresión son los utilizados para especificar determinada variable (en nuestro caso el rendimiento de cultivos) en base a otras variables que influyen en él.

Es aquí donde se encuentra entonces un resultado final para la investigación dando como salida del mismo una solución al problema conformada por distintas etapas ya detalladas pero las mencionamos nuevamente para mayor entendimiento: análisis de propiedades de los datos, selección del modelo de regresión para el rendimiento que se adapte a los datos, estimación y

validación del modelo de acuerdo al tipo, predicción del rendimiento en base a datos recolectados para la zafra que se quiere predecir y por último medición de confiabilidad de la predicción.

La solución brindada al cliente genera un buen precedente como punto de partida a un problema tan complejo como predecir de manera confiable rendimientos de cultivos. Las diferentes etapas por las cuales se debe pasar antes de llegar a la solución final, que es la predicción, son totalmente detalladas en este documento. De todas formas se cree que existen muchos puntos que pueden parecer ínfimos frente a la solución general, pero pueden llegar a ser determinantes para obtener una buena predicción. No se debe dejar de lado y es muy importante tener en cuenta que este tipo de análisis actualmente no tiene soluciones genéricas sino que se atacan subproblemas definiendo determinadas hipótesis o en los datos o en la chacra, etc. Por tal motivo este proyecto pretende ser un aporte para la integración de las soluciones parciales tratando de marcar una línea de trabajo para una solución general.

6.2 Resultados de la Solución informática

La salida del resultado de la investigación que es la solución general al problema de predicción, es utilizada como entrada y requerimiento para la solución informática. Se determinó la arquitectura y el diseño para la totalidad de la solución. Se definieron los casos de uso principales de la librería que son Analizar Datos, Obtener Modelo, Estimar rendimiento con modelo y Estimar rendimiento sin modelo. Mediante reuniones con el tutor y el cliente, teniendo en cuenta los tiempos del proyecto y la cantidad de tiempo insumida en la investigación de temas totalmente nuevos para los autores, se definió como alcance del Sistema implementar el caso de uso Análisis de Datos.

La arquitectura de la biblioteca se definió teniendo en cuenta la mayor cantidad de casos posibles con los que nos podemos encontrar, pero teniendo en cuenta que pueden existir nuevas investigaciones y algoritmos que puedan aparecer con mejor performance o resultados, la misma es flexible para la extensibilidad. El diseño es claro y no deja dudas para una eventual etapa de desarrollo de los casos de uso no implementados.

Los resultados obtenidos, en lo que tiene que ver con la implementación, se indican buenos tiempos de performance (ver Testing de Performance 5.6.8) en los algoritmos y contrastes utilizados, así como también resultados concretos y correctos en base a conjunto de datos obtenidos en la Web de los cuales se saben las propiedades de los datos de antemano.

Creemos que el análisis de los datos previo a la selección del modelo es muy importante y determinante, dado que si no se realiza un buen análisis obteniendo resultados confiables, la selección del modelo será errónea y por ende la predicción en base a dicho modelo será sesgado con respecto a la mejor solución.

El cliente otorgó cuatro juegos de datos en el formato correspondiente de archivo ZF para los cuales se obtuvieron los siguientes resultados:

Archivo Pot1Maiz07.ZF

Multicolinealidad: Inexistente

Autocorrelación en el Rendimiento: Sí

Autocorrelación en el Error: Sí

Heteroscedasticidad: No

Distribución en el Rendimiento: No Normal

Distribución en el Error: No Normal

Linealidad: Sí

Archivo Pot1Trigo06.ZF

Multicolinealidad: Moderada

Autocorrelación en el Rendimiento: Sí

Autocorrelación en el Error: Sí

Heteroscedasticidad: No

Distribución en el Rendimiento: No Normal

Distribución en el Error: No Normal

Linealidad: Sí

Archivo Pot6Maiz06.ZF

Multicolinealidad: Alta

Autocorrelación en el Rendimiento: Si

Autocorrelación en el Error: No se pudo chequear porque la matriz del MCO no es invertible

Heteroscedasticidad: No se pudo chequear porque la matriz del MCO no es invertible

Distribución en el Rendimiento: No Normal

Distribución en el Error: No se pudo calcular los errores del MCO

Linealidad: No se pudo chequear porque la matriz del MCO no es invertible

Archivo Pot6Soja07.ZF

Multicolinealidad: Alta

Autocorrelación en el Rendimiento: Si

Autocorrelación en el Error: No se pudo chequear porque la matriz del MCO no es invertible

Heteroscedasticidad: No se pudo chequear porque la matriz del MCO no es invertible

Distribución en el Rendimiento: No Normal

Distribución en el Error: No se pudo calcular los errores del MCO

Linealidad: No se pudo chequear porque la matriz del MCO no es invertible

Capítulo 7

Conclusiones

En lo que refiere al objetivo de realizar una investigación que tenga como resultado encontrar una solución al problema de predicción del rendimiento de un cultivo, en base a datos históricos (de corte transversal) del mismo y de distintos factores que influyen en él, se tuvo en cuenta las recomendaciones del tutor y de un experto en el tema de Econometría, los cuales nos orientaron hacia el uso de las distintas técnicas que brinda la misma. La investigación realizada nos llevó a la conclusión que la rama de la Econometría llamada Econometría Espacial (EE) es la metodología que ataca con mayor especificidad el problema de predicción, cuando los datos se encuentran distribuidos espacialmente, debido a que tiene en cuenta los llamados "efectos espaciales", características que se analizaron en los datos para el caso en estudio. En base a las herramientas que brinda la EE, se diseñó una solución genérica para el problema planteado, presentada en una serie de etapas que son: análisis de datos, selección, estimación y validación del modelo, y finalmente predicción del rendimiento. Las etapas mencionadas son una conclusión de la investigación de diferentes bibliografías, artículos científicos y estudios académicos. Ninguna de las fuentes bibliográficas a partir de las cuales se realizó la investigación, atacan de forma general (las distintas etapas mencionadas anteriormente) el problema de predicción de una variable por medio de modelos econométricos espaciales, mucho menos, se encontró un estudio que integre todas las posibles soluciones, dependiendo de las propiedades que los datos puedan cumplir. Creemos que un resultado importante de esta investigación es el de llegar a dicha solución genérica que abarque la mayor cantidad de casos posibles, teniendo en cuenta la heterogeneidad de los datos entre diferentes chacras, ya que éste puede ser el punto de partida para refinar la solución, haciéndola cada vez mas potente a la hora de aplicarla en cualquier tipo de chacra. El hecho de tener modularizadas las etapas, hace que la solución sea fácilmente extensible, teniendo en cuenta los posibles nuevos contrastes que puedan surgir o modelos que se adapten cada vez mejor al problema en cuestión.

Con respecto al segundo objetivo que es el de diseñar una herramienta informática que implemente la solución brindada por la investigación, se diseñó completamente la herramienta teniendo en cuenta criterios fundamentales en este caso, como la extensibilidad, mantenibilidad y adaptabilidad. Concluimos como característica principal del diseño y la arquitectura, la facilidad con la que se pueden agregar nuevos algoritmos para los contrastes, o nuevos contrastes sobre los datos, así como también la interpretación de nuevos modelos con sus correspondientes estimadores. Como resultado importante extraído a partir de los análisis realizados por la biblioteca sobre los cuatro archivos de datos de chacras brindadas por el cliente, se confirmó la presencia de autocorrelación global positiva en los rendimientos de los cultivos. Este resultado confirma que el uso de la EE es adecuado, dado que, de no detectarse

autocorrelación la EE no sería necesaria. Para la detección de la autocorrelación se emplearon indistintamente los contrastes de I de Moran y C de Geary, ambos con resultados similares. A su vez, ambos índices utilizaron en diferentes pruebas distintos criterios de vecindad en la matriz de ponderación de primer orden. En cuanto a la autocorrelación en el error, se verificó su presencia, también en forma positiva y global, empleándose los mismos contrastes que en el caso del rendimiento, así como también los mismos criterios para la matriz de vecindad. Teniendo en cuenta que se detectaron los dos tipos de autocorrelación, la forma correcta de modelar sería mediante un modelo de autocorrelación mixta, que tenga en cuenta tanto la autocorrelación en el rendimiento como en el error. Sin embargo, teniendo en cuenta que la solución propuesta utiliza la estrategia seguida por Anselin y Florax, vista en la sección 3.4.4, entonces el modelo seleccionado es el modelo de ponderación, o el modelo de error espacial. Como se vio en la sección 3.23.2.2, una de las razones que produce la heteroscedasticidad es la omisión de variables que tienen una fuerte influencia en la variable en estudio. Teniendo en cuenta que el rendimiento depende de tres grandes grupos de factores, de los cuales uno de ellos (los insumos aplicados a la chacra) no fue tenido en cuenta, era de esperar que se detectara heteroscedasticidad en los datos. Sin embargo el contraste de White no detectó la presencia de dicho efecto. Por lo tanto podemos concluir, que no se puede asegurar que el hecho de omitir los insumos en el modelo implique la presencia de este efecto espacial.

Finalmente, teniendo en cuenta que se trata de la primera experiencia por parte de este grupo en el abordaje de dicha problemática, estando los resultado obtenidos dentro de los parámetros aceptables y considerando la importancia a nivel de la Agricultura en forma general, de conocer de antemano el rendimiento futuro, se entiende que es posible, rentable y provechoso el profundizar en el estudio de este problema.

Capítulo 8

Trabajos futuros

Los trabajos futuros de interés los podemos dividir en dos grupos. El primer grupo es el que compone principalmente la implementación de los módulos de la solución diseñada pero no implementada. El otro grupo corresponde a la extensión de la investigación realizada en lo que tiene que ver con la Econometría Espacial dado que esta rama de la Econometría es un tema muy amplio de investigación. No dejar de lado el hecho de tener en cuenta otras ramas de la Econometría Espacial que puedan ayudar a mejorar la solución en la medida que se pueda contar con una mayor cantidad de datos históricos. A continuación se listan lo que a nuestro criterio serían trabajos de relevancia:

Implementación de los componentes

Este trabajo consiste en la implementación de los componentes que fueron diseñados pero no implementados, estos son la selección, estimación, validación del modelo y predicción del rendimiento.

Predicción en base a panel de datos

Las características del suelo y del cultivo son dinámicas tanto en el espacio como en el tiempo, esto significa que el rendimiento no solo depende de las distintas relaciones espaciales entre las variables, además depende de las historia del cultivo. El hecho de contar con datos de corte transversal hace que la predicción se realice solo teniendo en cuenta las distintas interacciones espaciales entre el rendimiento y las variables, dejando de lado el aspecto temporal. Es por ello que un trabajo de relevancia pasa por realizar una investigación en aquellas metodologías que tienen en cuenta los dos aspectos para realizar predicción, esto siempre que se cuente con una suficiente cantidad de datos históricos de varias zafras para una misma chacra, este tipo de datos se conoce como datos de panel[12]. La solución a este problema también pasa por el uso de modelos econométricos.

Selección de variables influyentes

[12] Datos de panel de datos son compuestos por datos de corte transversal y de series de tiempo simultáneamente.

En casos donde se cuenta con una gran cantidad de variables exógenas, se debe plantear la pregunta si se debe o no incluir todas las variables al modelo. Intuitivamente se puede pensar el hecho de agregar todas las variables exógenas posibles al modelo permite obtener mejores resultados, pero esto no siempre es así ya que agregar variables que no influyen en el rendimiento hace que aumente la varianza de la estimación del modelo y además el hecho de contar con muchas variables puede producir problemas de multicolinealidad. La solución planteada en este proyecto, toma como punto de partida todas las variables con las que se cuenta, pues a priori se supone que todas influyen en el rendimiento, la única razón por la que se puede excluir una variable del modelo es por causa de multicolinealidad entre algunas de ellas eliminando aquellas que producen este efecto no deseado. Teniendo en cuenta que no todas la variables tienen por qué influir sobre el rendimiento es interesante formular una solución que tome el mejor conjunto de variables que hacen que la predicción realizada sea la mejor dentro de todos los conjuntos posibles. Una posible solución a esto es probar con la combinación de todos los conjuntos posibles, la cual llega al mejor subconjunto, pero computacionalmente es muy costosa. Existen métodos para seleccionar el mejor subconjunto como los procedimientos "paso a paso" (o stepwise)[13] o el algoritmo "branch and bound"[14], pero cualquiera de estas dos opciones se aplican a modelos de regresión lineal, pero una posibilidad seria estudiar la forma (de ser posible) de adaptar estos procedimientos a modelos de regresión espacial. [13]

[13] Los procedimientos "paso a paso" (o setpwise) permiten elegir el subconjunto de variables regresoras que deben entrar en un modelo de regresión lineal.

[14] El algoritmo "branch and bound" permite seleccionar el mejor subconjunto de variables en base a medidas de bondad de ajuste sobre el modelo de regresión estimado.

Referencias

[1] Luc Anselin. Spatial Econometrics: Methods and Models. 1988. ISBN: 9024737354.

[2] Moreno Serrano y Vayá Valcarce. Técnicas econometritas para el tratamiento de datos espaciales: La econometría espacial. 2000 ISBN: 84-8338-224-5.

[3] Alfonso Novales. Econometría Segunda edición. Universidad Complutense Madrid. 1993. ISBN: 84-481-0128-6

[4] Rodolfo Bongiovanni, Evandro Chartuni Mantovani, Stanley Best, Álvaro Roel. Agricultura de precisión: Integrando conocimientos para una agricultura moderna y sustentable. Instituto Interamericano de Cooperación para la Agricultura. 2006. http://www.procisur.org.uy/online/cyber_ficha.asp?grupo=9&doc=135 (12/07/2006)

[5] Rodolfo Bongiovanni. Econometría espacial: Una herramienta clave para el manejo sitio-específico de insumos. Taller Internacional de Agricultura del Cono Sur de America. Diciembre 2002. http://www.agriculturadeprecision.org/cursos/IIITallerInternacional/Bongiovanni%20(INTA%20Manfredi)%20Procisur%2017-19%20Dic%202002.pdf (12/07/2008)

[6] Rodolfo Bongiovanni. Que hacer con los datos de monitor de rendimiento: Analisis Economico. Septiembre 2002. http://www.agriculturadeprecision.org/articulos/analecon.htm (12/07/2008) REF 3

[7] Coro Chasco Yrigoyen. Modelos de heterogeneidad especial. Universidad Autonoma de Madrid. http://129.3.20.41/eps/em/papers/0411/0411004.pdf (10/07/2008)

[8] Amparo Toral Arto. El factor espacial en la convergencia de las regiones de la Unión Europea: 1980-1996. Universidad Pontificia Comillas de Madrid. Octubre 2001. ISBN: 84-689-0568-2. http://www.eumed.net/tesis/ata/index.htm (12/07/2008)

[9] Martha Bohórquez. Introduccion a la Estadistica Espacial. Universidad Nacional de Colombia. http://www.docentes.unal.edu.co/mpbohorquezc/docs/Introducci%3Fn%20a%20la%20Estad%3Fstica%20Espacial.pdf (12707/2008)

[10] Gerson Javier Pérez. Dimensión espacial de la pobreza en Colombia. Banco de la Republica. Enero 2005. ISSN: 1692-3715. http://www.banrep.gov.co/docum/Pdf-econom-region/Documentos/DTSER-54.pdf (12/07/2008)

[11] Simon Sánchez Moral. El estudio econométrico de la concentración espacial de la industria: Ejemplo de aplicación en Madrid, Toledo y Guadalajara. Universidad Complutense Madrid. Septiembre 2004. ISSN: 0211-9803 http://www.ucm.es/BUCM/revistas/ghi/02119803/articulos/AGUC0404110207A.PDF (09/07/2008)

[12] J. Paul Elhorst, Dirk Strijker. Spatial developments of EU agriculture in the post-war period: The case of wheat and tobacco. Agricultural Economics Review. Febrero 2003. http://ageconsearch.umn.edu/bitstream/26417/1/04010063.pdf (13/07/2008)

[13] Juan Vilar. Curso Estadística 2. Universidad de La Coruña – Departamento de Matemáticas. http://www.udc.es/dep/mate/estadistica2/indice_gral.html Ultima fecha de ingreso (10/07/2008)

[14] Luc Anselin. Spatial Econometrics. Bruton Center School of Social Sciences University of Texas at Dallas Richardson. Abril 1999. http://www.csiss.org/learning_resources/content/papers/baltchap.pdf (13/07/2008)

[15] Universidad de las Américas Puebla- Innovación y Servicios de Información. http://catarina.udlap.mx/u_dl_a/tales/documentos/lec/lara_i_gd/apendiceE.pdf (22/06/2008)

[16] Rafael de Arce. Conceptos básicos sobre la Heterocedasticidad en el modelo básico de regresión lineal, tratamiento con E-Views - Universidad Autónoma de Madrid. Abril 2001. http://www.uam.es/departamentos/economicas/econapli/pdf/heterocedasticidad.pdf (23/07/2008)

[17] Angel Alcaide Inchausti, Nelson Alvarez Vázquez. Econometría: Modelos deterministas y estocásticos. 1992. ISBN 8480040491

[18] Rafael de Arce, Ramon Mahia. Contrastes de significacion conjunta en el MBRL. Universidad Autónoma de Madrid. http://www.uam.es/personal_pdi/economicas/rarce/pdf/significatividad.pdf (24/07/2008)

[19] Estudio de la demanda residencial anual de la energia electrica en la Comunidad Autonoma de Andalucia. http://descargas.cervantesvirtual.com/servlet/SirveObras/01604185214584961880035/013293_4.pdf (24/07/2008)

[20] William H. Green. Econometric Analysis (fifth edition). New York University. Pertince Hall. ISBN: 0-13-066189-9

[21] ICA – Ingenieros Consultores Asociados. http://www.ica.com.uy

[22] Estado del Arte: Agricultura de Precisión – Autores: Álvaro Crespi, Enrique Mora, Giovanni Sosa

[23] Estado del Arte: Métodos de Predicción Temporal – Autores: Álvaro Crespi, Enrique Mora, Giovanni Sosa

[24] Estado del Arte: Métodos de Análisis Espacial – Autores: Álvaro Crespi, Enrique Mora, Giovanni Sosa

Apéndice

A. Análisis de riesgos del proyecto

A continuación presentan los riesgos que se pensaron podían surgir en el transcurso del proyecto. Se realizó una lista y un ranking de los mayores riesgos del proyecto, y que es lo que se planea hacer para mitigar cada riesgo**.**

Nombre	Descripción	Probabilidad	Impacto	Plan
Confiabilidad de datos	Escasez de datos confiables para la prueba final del sistema.	Alta	Crítico	Generar datos adicionales manipulándolos para que cubran los flujos incluidos dentro del alcance.
Desvíos en la investigación	Los datos del cliente pueden ser insuficientes para aplicar el modelo investigado.	Alta	Crítico	Desvíos de la investigación realizada en dirección a la nueva realidad planteada. Ejemplo: En principio la investigación se centraba en obtener una predicción realizando un análisis espacio temporal, luego esto cambio a un análisis solamente espacial. En principio la investigación se centraba en series temporales, luego vario a modelos de regresión y análisis de econometría espacial.
Alcance	La totalidad de las funcionalidades requeridas pueden estar por fuera del alcance de lo que el equipo de desarrollo puede entregar en el tiempo disponible.	Alta	Marginal	Asignar niveles de importancia a los casos de uso.
Exceso de especialización de un tema en particular	Demasiada especialización de un tema en particular del sistema por parte de una o más personas del equipo.	Media	Marginal	Guías, clases, y charlas con el experto.
Requerimientos	Requerimientos no están del todo claros al comienzo	Media	Crítico	Los requerimientos serán detallados primero los de más

	del proyecto. El cliente puede no disponer con suficientes recursos como para especificar sus requerimientos.			prioridad de acuerdo a las metas planteadas.
Disponibilidad de Datos	Cambios en la especificación del problema debido a disponibilidad de datos.	Media	Crítico	Aumento del tiempo utilizado en la investigación de soluciones a la nueva realidad. Por consiguiente se ven afectados los tiempos previstos del proyecto, lo que implica una nueva estimación de los tiempos. Ejemplo: Al inicio se arranco con series temporales, para poder predecir con series temporales, la cantidad de variables debe ser mayor que los grados de libertad de la función.
Falta de Datos para pruebas unitarias (datos matemáticos)	Dado que se cuenta con test matemáticos, el hecho de no tener datos confiables para testear el software conlleva a poca confiabilidad al sistema, debido a que esto empobrece la validación y la verificación del sistema.	Media	Media	Generar datos con datos que cumplan con las propiedades necesarias para generar el sistema.
Cambios	Después de los requisitos se han acordado y documentado, se comienza con el desarrollo basados en estos. De cambiar mas adelante los requerimientos se incurriría en un esfuerzo es en vano.	Bajo	Crítico	Un procedimiento de control de cambios es requerido, entonces los cambios solo son hechos cuando el costo del cambio es tolerable.

B. Administración del Proceso

B.1 Metodología

En el transcurso del proyecto se habrá de utilizar un modelo en cascada tal como se muestra en la figura:

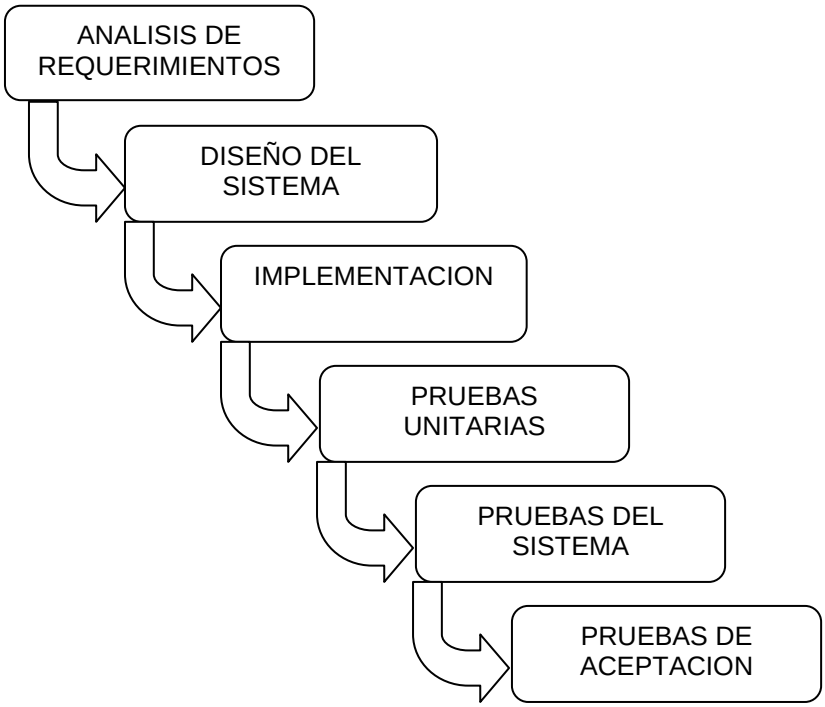


Figura 11.1 Modelo del proceso

El hecho de haber usado un modelo en cascada no implica en este caso que no existan saltos hacia atrás en el proceso en cascada. De hecho la etapa del análisis de requerimientos fue visitada más de una vez ya que se necesitaron hacer ajustes en los requerimientos, como así también en el diseño y en menor medida en la implementación.

B.2 Organización del equipo del proyecto

Se optó por un esquema de asignación de roles flexible dada la escasa cantidad de recursos humanos (solamente 3 personas en el equipo) y el alto involucramiento y conocimiento, de los aspectos mas relevantes de cada área, que demandaba el proyecto por parte de sus integrantes.
No obstante, existieron especialistas en determinadas áreas, por ejemplo: un miembro encargado de la arquitectura que contaba con un conocimiento más detallado de la misma, un encargado en la elaboración del algoritmo general de la aplicación, y un especialista técnico (ya que no todos los componentes usados eran familiares a todos los integrantes).

B.3 Herramientas de desarrollo y colaboración usadas

Las herramientas usadas a lo largo del proyecto fueron las siguientes:

- Visual Studio 2005, C# Framework 2.0
- Biblioteca MathNet.Iridium-2008.8.16.470 que provee funciones matemáticas avanzadas, como las necesarias para operar con matrices, interpolar, o estadísticas. Esta biblioteca es open source con licencia del tipo LGPL (GNU Lesser General Public License).

- SVN Tortoise 1.4.8, para el control de versiones
- Servidor gratuito para alojar las fuentes versionados por el SVN Tortoise.
- Issue Tracking.

B.4 Control de cambios

- Los pedidos de cambios en los requerimientos se registraron en el issue Tracker.
- Los pedidos de cambios fueron realizados a través de reuniones y menor medida mediante el uso del mail, y en cada caso se discutió su viabilidad entre los miembros de equipo, el representante del cliente y el tutor del proyecto.
- La validación de los requerimientos se hizo de acuerdo a los datos disponibles y a la teoría investigada. Una vez validado los requerimientos por parte del cliente no se aceptaran nuevos requerimientos.
- Luego de finalizado el hito código completo, no se hará ningún agregado a la versión.

B.5 Actualizaciones del proyecto

Este plan de proyecto fue actualizado según las necesidades surgidas a lo largo del proyecto. En la tabla al inicio de este mismo documento se especifica cuando se hizo cada cambio y en que consistió el mismo, así como los responsables de cada cambio.

B.6 Diagrama de Gantt

Number	Task	Start	End	Duration
1	Elaboracion del Estado del Arte AP	3/29/2007	7/21/2007	82
2	Elaboracion de la Especificacion de Requerimientos	5/28/2007	12/20/2007	148
3	Elaboracion del Estado del Arte Análisis espacio temporal	6/18/2007	11/15/2007	108
4	Refinamiento del formato de los archivos de intercambio de datos del Software	6/17/2007	2/23/2008	180
5	Elaboracion del Estado del Arte Funciones de Producción AP	10/18/2007	1/14/2008	62
6	Elaboración del Estado del Arte Modelos Análisis Espacial	2/15/2008	4/26/2008	51
7	Implementacion	4/11/2008	6/15/2008	46
8	Testeo de la Aplicacion	5/25/2008	7/20/2008	40
9	Elaboración del Documento Final	6/15/2008	9/20/2008	70

2007: January, February, March, April, May, June, July, August, September, October, November, December

2008: January, February, March, April, May, June, July, August, September, October, November, December

Figura B.1 Diagrama de Gantt

B.7 Work Breakdown Structure (WBS) y estimaciones

Pasos	Descripción	Estimación
1.	Investigación	
1.1.	Elaboración del Estado del Arte AP	465h
1.2.	Elaboración del Estado del Arte Análisis espacio temporal	512h
1.3.	Elaboración del Estado del Arte Funciones de Producción AP	360h
1.4.	Elaboración del Estado del Arte Modelos Análisis Espacial	310h
2.	Análisis y Diseño	
2.1.	Elaboración de la Especificación de Requerimientos	310h
2.2.	Refinamiento del formato de los archivos de intercambio de datos del Software	52h
3.	Implementación	
3.1.	Implementación	215h
4.	Testeo de la aplicación	
4.1.	Pruebas Unitarias	78h
4.2.	Pruebas del Sistema	23h
4.3.	Pruebas de Performance	23h
5.	Documentación Final	
5.1.	Elaboración del Documento Final	476h
	Total	**2825 horas**

C. Índice de Figuras

D. Evolución de la Tesis

En esta sección se da una explicación de los requerimientos y distintos objetivos que se plantearon con el cliente durante el desarrollo de la tesis.

D.1 Predicción en base a datos históricos

El objetivo inicial de la tesis fue construir un sistema informático que permita predecir el rendimiento de un cultivo, en base a datos históricos de zafras anteriores de una chacra y presentar los rendimientos predichos en diferentes zonas de manejo. Este objetivo estaba inmerso dentro de un objetivo más ambicioso enmarcado dentro de la AP, del cual formaban parte tres proyectos, donde la salida que produce la herramienta elaborada en un determinado proyecto formaba parte de la entrada que recibe la herramienta elaborada por otro proyecto. Si bien el objetivo global era ambicioso cada grupo por separado se sometía a un riesgo debido a la interdependencia con los demás proyectos.

La investigación se encuentra inmersa dentro del marco de la AP, por lo tanto teniendo en cuenta la falta de pericia en esta área, fue necesario realizar un estado del arte de la AP (ref. [22]); este documento ayuda a comprender cada uno de los conceptos que forman parte de la AP, como es el caso de las zonas de manejo.

Específicamente el sistema informático recibe como entrada una serie de tiempo que contiene las zonas de manejo de la chacra, índices de performance (índices que ayudan a decidir el número óptimo de zonas de manejo) y rendimiento para cada una de ellas, además se cuenta con información del clima y distintos atributos del suelo. Estos datos se supone que se obtuvieron en distintas zafras para la misma chacra, tomados en periodos regulares a través del tiempo (cada cierta cantidad de meses, dependiendo de la duración entre una zafra y otra).

El sistema debe proporcionar como salida una predicción de las nuevas zonas de manejo para la zafra siguiente, esto es: si los datos son anuales y llegan hasta el año 2007, se debe predecir las zafras del año 2008. Conjuntamente con cada zona de manejo se debe predecir el rendimiento de cada una de ellas y distintos índices de performance. La figura 1 ilustra la primera versión de los requerimientos, por más detalle ver apéndice F (documento: Especificación Preliminar del Proyecto).

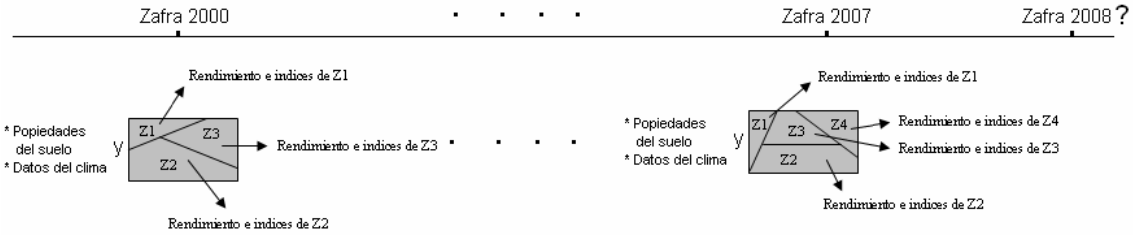


Figura 1 Primera versión de los requerimientos

Los requerimientos definidos llevaron a que se elabore un estado del arte sobre distintos métodos de predicción que utilizan análisis de series de tiempo para formular la predicción, ver

referencia [23]. Tras un par de reuniones los requerimientos se fueron afinando y se fue especificando con que tipos de datos complementarios de la chacra y del clima se contaba, con respecto a estos datos en todo momento se mencionaba por parte del cliente que en un futuro cercano se contaría con una extensa cantidad de datos históricos. Esto traía aparejado un fuerte riesgo sobre la investigación, debido a que se encontraba en una etapa avanzada de la misma y los datos históricos aun no eran confirmados por el cliente.

El refinamiento de los requerimientos llevó a que se especifique con mayor detalle. Luego de este proceso se llego a la siguiente especificación: la chacra esta dividida en celdas regulares, cada una de ellas contiene una muestra del rendimiento y de distintas propiedades del suelo y del manejo del mismo (muestreo en grilla sistemático), a su vez cada celda esta identificada de forma de indicar a la zona de manejo que pertenece, la figura 2 ilustra lo explicado, en la misma las zonas de manejo están identificadas por el color de la celda. Cada zafra de la serie de tiempo contiene los datos según lo expresa la figura y la predicción debe retornar el rendimiento de la zafra futura en cada celda y las nuevas zonas de manejo.

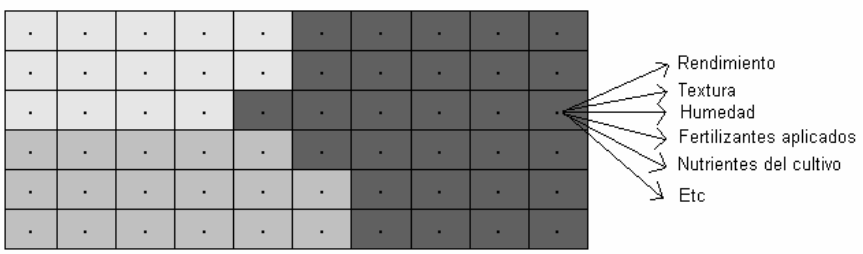


Figura 2 Muestreo en grilla sistemático indicando las zonas de manejo.

Luego de sucesivas reuniones se llega a la estabilidad de los requerimientos, pero surge el primer punto de inflexión sobre los mismos. La herramienta a desarrollar debe ejecutar dos algoritmos claramente identificados para producir la salida, uno de ellos para obtener la predicción del rendimiento en cada celda en la cual esta dividida la chacra, y otro algoritmo para determinar las nuevas zonas de manejo agrupando cada celda en la zona que corresponda. Justamente este último algoritmo esta muy relacionado con el objetivo de unos de los grupos (que componen el proyecto global mencionado anteriormente) el cual es, obtener en base a un determinado índice (p.ej.: rendimiento) las zonas de manejo de la chacra. Se llega a un acuerdo con el cliente y el tutor, recortándose el alcance del sistema. Eliminándose las zonas de manejo como entrada y salida del sistema, y el resto de los datos y requerimientos se mantienen tal cual.

Ahora los requerimientos son mas claros, lo que se quiere es poder predecir el rendimiento en cada celda de la chacra para la futura zafra, teniendo como datos de entrada una serie de tiempo con los rendimientos, las propiedades del suelo, y los datos climáticos como la temperatura (media en la zafra) y las lluvias (promedio anual en milímetros cúbicos) los cuales son constantes en todos los puntos de la chacra, pero varían de una zafra a otra.

Llegado a este punto la investigación se centra en las metodologías que puedan predecir el rendimiento de un cultivo en base a datos históricos del mismo y de distintas variables que se suponen que influyen en el rendimiento. Cabe destacar que la investigación es puramente de carácter temporal, de forma de reducir la complejidad del problema no se tendrá en cuanta las diferentes relaciones espaciales que puedan llegar a tener los datos de la chacra. Esto significa que la predicción del rendimiento en una celda de la chacra se realiza en base a los datos

históricos de lo que ocurrió en dicha celda, sin tener en cuenta lo que ocurrió en las celdas vecinas. La solución que se plantea se basa en los modelos de multidireccionales VAR.

Una restricción importante que plantea este modelo es que se debe contar con una importante cantidad de datos históricos, como cualquier modelo de series temporales. Justamente esta restricción causa un punto de inflexión importante en el desarrollo de la tesis, debido a que el cliente no puede contar con la cantidad de datos históricos acordada, esto llevó a reformular los requerimientos de forma de reencaminar el proyecto. De todas formas antes de mencionar los nuevos requerimientos acordados presentamos la propuesta de solución planteada en ese momento.

Propuesta de Solución

Si bien durante la investigación se abarcó distintas metodologías, las cuales fueron documentadas en el estado del arte Métodos de Predicción Temporal, la solución al problema se plantea mediante el uso de una metodología basada en Series de Tiempo, ya que el enfoque y los lineamientos del proyecto se inclinan hacia esta metodología. La solución planteada pasa por la utilización de los modelos multidireccionales VAR. Se opta por estos modelos pues permiten plantear una ecuación por cada punto muestral (recordar que cada celda se trata por separado), además estos modelos permiten expresar el comportamiento de una variable en base a valores pasados de la misma y a valores pasados de otras variables que se suponen que influyen en la primera. El modelo planteado se expresa formalmente:

$$r_t = \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + ... + \alpha_p r_{t-p} + \beta_1 x_{1,t-1} + ... + \beta_k x_{1,t-p} + ... + \beta_m x_{k,t-1} + ... + \beta_{m+p} x_{k,t-p} + \varepsilon_t$$

Donde r_t es la variable endógena (en el caso de estudio representa el rendimiento del cultivo en la zafra t de la serie de tiempo), $x_{i,t}$ es la variable exógena i en la zafra t (representa las distintas propiedades del suelo, las lluvias y temperaturas, que son constantes en todos los puntos de la chacra, pero no a través de las distintas zafras, o sea del tiempo), α_j y β_q son coeficientes del modelo a estimar, reflejan el impacto de las variables sobre el rendimiento a predecir, ε_t es el error cometido en el tiempo t al estimar el modelo, p es la cantidad de retardos a utilizar, este parámetro expresa cuantos datos históricos hacia atrás utiliza el modelo para realizar la predicción, y k es la cantidad de variables exógenas.

Este modelo se define en cada celda en que esta dividida la chacra y cada uno de ellos se estiman por separado, en consecuencia los parámetros estimados son distintos en cada celda de la chacra. El método de estimación que se aplica es Mínimos Cuadrados Ordinarios (MCO).

El sistema de ecuaciones que resuelve el MCO para cada celda se forma tomando series de tiempo de largo p, $t_1, t_2, .., t_p$, en donde t_i representa la zafra i. Además contiene el valor del rendimiento, las propiedades del suelo, y los datos del clima en dicha zafra para la celda que se esta analizando. Teniendo en cuenta estas aclaraciones el sistema de ecuaciones queda planteado con las siguientes series de tiempo:

ecuación 1: $t_1, t_2, .., t_p$ (comienza en la zafra 1 y termina en la zafra p)

ecuación 2: $t_2, t_3, .., t_{p+1}$ (comienza en la zafra 2 y termina en la zafra $p + 1$)

ecuación 3: $t_3, t_4, .., t_{p+2}$ (comienza en la zafra 3 y termina en la zafra $p + 2$)

.....

.....

ecuación n-p: $t_{n-p}, t_{n-p+1}, .., t_n$ (comienza en la zafra n-p y termina en la zafra n)

Formalmente el sistema de ecuaciones queda determinado de la siguiente forma:

$$r_{t_p} = \alpha_0 + \alpha_1 r_{t_{p-1}} + \alpha_2 r_{t_{p-2}} + ... + \alpha_p r_{t_1} + \beta_1 x_{1,t_{p-1}} + ... + \beta_k x_{1,t_1} + ... + \beta_m x_{k,t_{p-1}} + ... + \beta_{m+p} x_{k,t_1} + \varepsilon_{t_p}$$

$$r_{t_{p+1}} = \alpha_0 + \alpha_1 r_{t_p} + \alpha_2 r_{t_{p-1}} + ... + \alpha_p r_{t_2} + \beta_1 x_{1,t_p} + ... + \beta_k x_{1,t_2} + ... + \beta_m x_{k,t_p} + ... + \beta_{m+p} x_{k,t_2} + \varepsilon_{t_{p+1}}$$

.....

.....

.....

$$r_{t_n} = \alpha_0 + \alpha_1 r_{t_{n-1}} + \alpha_2 r_{t_{n-2}} + ... + \alpha_p r_{t_{n-p}} + \beta_1 x_{1,t_{n-1}} + ... + \beta_k x_{1,t_{n-p}} + ... + \beta_m x_{k,t_{n-1}} + ... + \beta_{m+p} x_{k,t_{n-p}} + \varepsilon_{t_n}$$

Este es un sistema de $n - p$ ecuaciones, en consecuencia para que pueda ser resuelto debe tener a lo sumo $n - p$ parámetros a estimar. Aquí surge el primer inconveniente de este modelo, ya que son tantos los factores que influyen en el rendimiento (físicos, químicos, fisiológicos, climáticos, etc.), que hace que se deba contar con una serie de datos históricos numerosa. Claro que es imposible contar con información de todos los factores, pero en caso que estos sean más numerosos que la serie de tiempo se puede realizar un estudio de cuales son los factores que tiene mayor influencia en el rendimiento y recortar las variables del modelo a aquellos factores con mayor influencia, de forma de tener un sistema con mas ecuaciones que variables.

La cantidad de parámetros a estimar no solo depende de la cantidad de variables, aquí entra en juego el parámetro de rezago p, por cada variable que se incluye en el modelo hay que estimar p parámetros mas, por lo tanto la cantidad de parámetros a estimar depende de la cantidad de variables y del valor del parámetro de rezago p. Formalmente la restricción entre cantidad de variable, largo de la serie y tamaño de rezago se plantea como:

$$n - p >= 1 + p + p * E$$

donde E es la cantidad de variables que se incluyen en el modelo y el término $p * E$ representa la cantidad de parámetros β_i a estimar correspondientes a las variables exógenas del modelo; el término p representa la cantidad de parámetros α_i a estimar correspondiente al rezago del rendimiento y el uno corresponde al parámetro independiente α_0.

Se pude expresar la desigualdad anterior de forma de replantear una restricción en cuanto al largo de la serie temporal $t_1, t_2, .., t_n$ (cantidad de zafras) con respecto a la cantidad de variables exógenas y el valor del parámetro de rezago:

$$n >= 1 + 2p + p * E$$

De la desigualdad anterior se pude deducir que, debe ser mayor la cantidad de datos históricos con la que se cuente respecto a la cantidad de variables, de lo contrario se debe eliminar algunas de las variables exógenas y reducir el valor del parámetro de rezago. Por citar un ejemplo, supongamos que se cuenta con datos históricos de 20 zafras y que la cantidad de variables exógenas es de 10, en este caso el único valor que se le puede asignar al parámetro de rezago p es 1, esto significa que el modelo tan solo utilice el estado de la zafra anterior para predecir el rendimiento de la siguiente zafra, esto va en contra de la idea que hay detrás de los modelos basados en series temporales, donde lo ideal es utilizar la mayor cantidad de datos históricos posibles. Una de las soluciones a este problema pasa por el lado de reducir la cantidad de variables exógenas a aquellas que posean mayor influencia sobre el rendimiento, aunque esta opción no parece ser las mas apropiada ya que se puede omitir alguna variable importante y en consecuencia el modelo queda mal especificado, y las predicciones del rendimiento serán sesgadas. La otra opción es esperar a contar con mayor cantidad de datos históricos para luego si poder aplicar el modelo.

En el caso de estudio se cuenta con información de 30 variables, por lo tanto si $p = 5$, para que todas las variables puedan ser utilizadas se debe contar con datos históricos de por lo menos 161 zafras (de la misma chacra), por cierto que esta cantidad de datos históricos es imposible en cualquier parte del mundo y menos en Uruguay, debido a que la adopción de las técnicas de AP son muy recientes. En consecuencia la solución inevitablemente pasa por reducir la cantidad de variables exógenas, para lo cual se debe realizar un estudio de aquellas variables que influyen en mayor medida en el rendimiento.

Luego de haber obtenido la estimación de los parámetros del modelo para cada punto de la chacra por separado, se llega a una función $f_i(y_{t-1},..,y_{t-p}, x_{1,t-1}..x_{1,t-p}..... x_{k,t-1}..x_{k,t-p})$ para cada punto i. Para realizar la predicción del rendimiento en $t+1$ para el punto i se debe sustituir en la función f_i los últimos p valores del rendimiento y de las variables exógenas del punto i.

D.2 Predicción en base a datos de corte transversal

Debido a la escasez de datos históricos con los cual el cliente cuenta se debe reformular los requerimientos. El cliente plantea la posibilidad de contar con a lo sumo datos históricos de tres zafras, en consecuencia la solución explicada en la sección anterior no puede ser aplicada. Por lo tanto ya no se cuenta con datos de corte longitudinal (en el tiempo) pero si se sigue contando con datos de corte transversal, pues se tiene muchas muestras del rendimiento del cultivo para una misma zafra.

En base a la entrevista brindada por una experta en el área de la econometría, nos plantea la posibilidad de aplicar técnicas econométricas. La nueva solución se focaliza a la utilización del modelo de regresión lineal. Estos modelos permiten relacionar una variable endógena (rendimiento) en base a un conjunto de variables exógenas (propiedades del suelo, manejo del suelo y el clima) mediante una función lineal, de la siguiente forma:

$$y = f(x_1, x_2, x_3, .., x_k, \mu / \beta)$$

donde y es la variable endógena, x_i son las variables exógenas, β es el vector de coeficientes asociados a cada una de las variables exógenas y μ es un vector con los errores cometidos por el modelo.

El modelo de regresión lineal se suele aplicar para datos de series temporales, pero se puede adaptar los datos de corte transversal linealizando todos los puntos de la chacra donde se obtuvo una muestra, de forma de poder aplicar el modelo. El modelo se pude expresar de la siguiente forma:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + .. + \beta_k x_{ki} + \mu_i \quad \text{con } i = 1,2,..,N$$

donde N es el tamaño de la muestra, x_{1i} es el valor que toma la variable x_1 en la celda i y β_i representa el impacto producido por la variable x_i.

El objetivo de aplicar un modelo de regresión lineal en el contexto en el cual se utiliza pueden ser dos:

- Obtener una función que permita predecir el valor del rendimiento una vez conocidos los valores de los factores que influyen en él y que utiliza el modelo. En este caso se utiliza el modelo de regresión como un modelo predictivo.
- Conocer la relación que existe entre el rendimiento y los factores que lo limitan. De esta forma el usuario del sistema puede variar el valor de los factores y ver como responde el rendimiento del cultivo. De esta forma se pude conocer por ejemplo en cuanto puede disminuir la producción si para la próxima zafra diminuye las precipitaciones. En este caso se utiliza el modelo de regresión como un modelo explicativo.

Teniendo en cuenta los beneficios de aplicar un modelo de regresión lineal, surgen los nuevos requerimientos. Uno de ellos se basa en utilizar el modelo como forma predictiva, esto es poder predecir (para la próxima zafra) el rendimiento del cultivo en cada celda en que fue dividida la chacra, teniendo como entrada los datos del rendimiento, propiedades del suelo y clima de la zafra anterior (divididos por celda), además se debe contar con los nuevos datos del suelo y clima próximos a la nueva zafra, de forma de sustituir estos datos en la función que se obtiene y obtener los rendimientos esperados para la nueva zafra, la siguiente figura ilustra lo explicado.

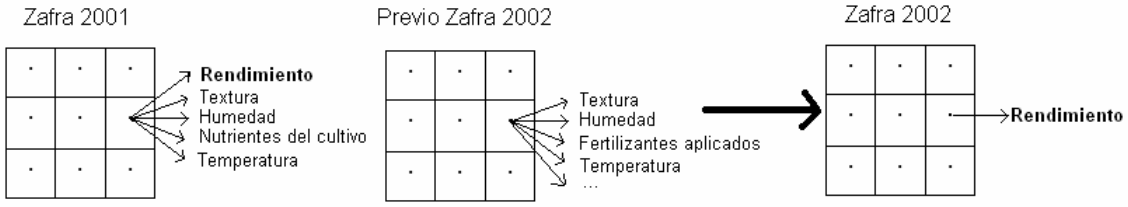


Figura 3 Predicción del rendimiento en base a datos de corte transversal

El otro requerimiento se basa en devolver el modelo de manera que permita ser utilizado de forma explicativa. Esto es devolver los parámetros estimados del modelo teniendo como entrada los datos del rendimiento, propiedades del suelo y clima de una determinada zafra. La siguiente figura ilustra lo explicado.

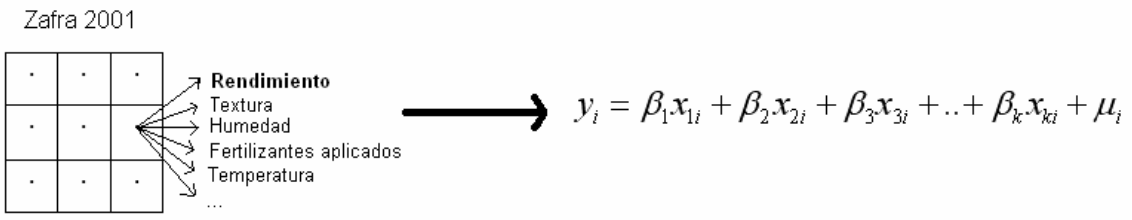


Figura 4 Obtención de un modelo de regresión lineal en base a datos de corte transversal

El modelo de regresión lineal puede ser aplicado en el proyecto en estudio, pues el rendimiento se comporta según una función de producción, dicha función esta definida por parámetros que relacionan el rendimiento en base a distintos factores. El inconveniente que surge es poder determinar a priori el tipo de relación que existe entre el rendimiento y los factores que lo determinan. El modelo de regresión lineal como su nombre lo indica permite obtener una función lineal tanto en las variables como en los parámetros β, pero el modelo también puede ser aplicado a determinadas funciones no lineales pero que son fácilmente linealizables haciendo determinadas transformaciones de sus variables. Por lo tanto, parte de la investigación se centra en conocer que tipo de relación existe entre el rendimiento y sus factores determinantes (ver ref. [23]). Distintos estudios indican que esta no es una relación lineal y depende del tipo de cultivo que se analiza y de los factores que se incluyen (muchos de ellos incluyen el factor capital y trabajo en consecuencia la función de producción es estudiada desde un punto de vista agro-económico), en todos los casos las funciones de producción que se aplican son fácilmente linealizables, las mas aplicadas son las mencionadas en el documento del estado del arte de métodos de análisis temporal (ref. [23]).

Por lo tanto la solución pasa por estimar el modelo utilizando como relación base cada una de las funciones especificadas en el Apéndice antes mencionado, y ver cual de ellas se ajusta de mejor forma a los datos brindados por el cliente, utilizando para ello el modelo de regresión lineal estimado mediante el método de MCO. La estimación mediante MCO exige que se deban realizar determinados estudios a los datos previos a la estimación, los más importantes son chequeo de autocorrelación y heteroscedasticidad en los errores producidos por el modelo y chequeo de autocorrelación en el rendimiento.

Es durante el estudio de los distintos contrastes de autocorrelación donde surge otro punto de inflexión en el proyecto, si bien los requerimientos no cambian lo que si se debe modificar es la metodología a utilizar. La autocorrelación en el modelo de regresión lineal se trata de forma unidireccional, esto debido a que los datos debieron linealizarse de forma de poder aplicar el modelo, por tal motivo los contrastes que brinda la Econometría solo pueden detectar autocorrelación lineal. Pero la autocorrelación en datos de corte transversal se da de forma multidireccional y al ser linealizados pierde sentido estudiar esta propiedad en los datos. Profundizando en este tema se encontraron textos que indican que se debe estudiar diferentes fenómenos espaciales que se dan en este tipo de datos. La Econometría Espacial provee herramientas para este tipo de estudio, integradas a las soluciones para los modelos de regresión, mediante los modelos de regresión espacial. Lo que llevó a que la investigación se centre en esta área de la Econometría.

E. Especificación de casos de uso

E.1 Analizar Datos

Descripción

El usuario solicita realizar un análisis de los datos ingresados.
Se realiza un análisis de las diferentes propiedades que los datos cumplen.
Al finalizar el caso de uso el usuario obtendrá:

1) Análisis de Multicolinealidad.
2) Análisis de Linealidad.
3) Análisis de Distribución del rendimiento.
4) Análisis de Autocorrelación
5) Análisis de Heterosedasticidad.

Pre-condiciones

Debe estar definido el archivo con los datos recolectados en formato ZF.

Flujo de eventos principal

Usuario	Sistema
1. Solicita el análisis de los datos recolectados que se encuentran en el archivo con formato ZF.	
	2. Define la matriz W de vecindad en base a los datos.
	3. Se aplica el estimador MCO sobre las variables exógenas ingresadas y el rendimiento. De aquí se obtiene el error de la estimación.
	4. Se contrasta la normalidad o no normalidad de la distribución del error obtenido en el 3.
	5. Se contrasta la normalidad o no normalidad de la distribución de la variable rendimiento.
	6. Se calculan los índices de Moran para el error y el rendimiento.
	7. Se calculan los índices de Geary para el error y el rendimiento.
	8. En base a los índices calculados para el rendimiento de define si existe autocorrelación en la variable rendimiento.
	9. En base a los índices calculados para el rendimiento de define si existe autocorrelación en el error del modelo.
	10. Se determina mediante el test de Spearman si existe heterosedasticidad.
	11. Se determina mediante el test de White si existe heterosedasticidad.
	12. Se determina en base a los cálculo de 10 y 11

	si existe heterosedasticidad. 13. Se realiza el contraste de Multicolinealidad. 14. Se realiza el contraste de linealidad entre las variables consideradas en los datos de entrada. 15. Se determina el conjunto de las variables que influyen en el rendimiento y las que no. 16. Se persisten todos los datos calculados y los resultados obtenidos en archivo con formato de salida.

E.2 Obtener Modelo

Descripción

El usuario solicita la obtención del mejor modelo que se adapta a los datos disponibles dentro de una serie de modelos predefinidos en el sistema. El modelo retornado es el que adapta mejor a las propiedades que tienen los datos recolectados de la chacra.

Pre-condiciones

Debe estar definido el archivo con los datos recolectados en formato ZF.

Flujo de eventos principal

Usuario	Sistema
1. Solicita el análisis de los datos recolectados que se encuentran en el archivo con formato ZF.	
	2. Define la matriz W de vecindad en base a los datos. 3. Se aplica el estimador MCO sobre las variables exógenas ingresadas y el rendimiento. De aquí se obtiene el error de la estimación. 4. Se contrasta la normalidad o no normalidad de la distribución del error obtenido en el 3. 5. Se contrasta la normalidad o no normalidad de la distribución de la variable rendimiento. 6. Se calculan los índices de Moran para el error y el rendimiento. 7. Se calculan los índices de Geary para el error y el rendimiento. 8. En base a los índices calculados para el rendimiento de define si existe autocorrelación en la variable rendimiento. 9. En base a los índices calculados para el rendimiento de define si existe autocorrelación en el error del modelo. 10. Se determina mediante el test de Spearman si existe heterosedasticidad. 11. Se determina mediante el test de White si existe heterosedasticidad. 12. Se determina en base a los cálculo de 10 y 11

	si existe heterosedasticidad. 13. Se realiza el contraste de Multicolinealidad. 14. Se realiza el contraste de linealidad entre las variables consideradas en los datos de entrada. 15. Se determina que existe autocorrelación en el término de error y en la variable rendimiento. *16.* Se calcula el contraste de los multiplicadores de Lagrange (LM_λ y LM_ρ). 17. LM_λ es más significativo que LM_ρ por lo tanto se define el modelo teniendo en cuenta autocorrelación en el término de error. *18.* Se persisten todos los datos calculados y los resultados obtenidos en archivo con formato de salida para que se pueda cargar el modelo obtenido en el Sistema.

Flujo Alternativo 1

Usuario	Sistema
	1. LM_ρ es más significativo que LM_λ por lo tanto se define el modelo teniendo en cuenta autocorrelación en la variable endógena (rendimiento). ***2.*** Se persisten todos los datos calculados y los resultados obtenidos en archivo con formato de salida para que se pueda cargar el modelo obtenido en el Sistema.

Flujo Alternativo 2

Usuario	Sistema
	1. Se determina que existe autocorrelación en el término de error y no en el rendimiento. 2. Se define el modelo teniendo en cuenta autocorrelación solo en el termino de error. 3. Se persisten todos los datos calculados y los resultados obtenidos en archivo con formato de salida para que se pueda cargar el modelo obtenido en el Sistema.

Flujo Alternativo 3

Usuario	Sistema
	1. Se determina que existe autocorrelación en el rendimiento y no en el término de error. **2.** Se define el modelo teniendo en cuenta autocorrelación solo en la variable rendimiento. **3.** Se persisten todos los datos calculados y los resultados obtenidos en archivo con formato de salida para que se pueda cargar el modelo obtenido en el Sistema.

Flujo Alternativo 4

Usuario	Sistema
	1. Se determina la no existencia de autocorrelación. **2.** Se determina heterosedasticidad en los datos. **3.** Se define el modelo teniendo en cuenta la heterosedasticidad en los datos. **4.** Se persisten todos los datos calculados y los resultados obtenidos en archivo con formato de salida para que se pueda cargar el modelo obtenido en el Sistema.

Flujo Alternativo 5

Usuario	Sistema
	1. Se determina la no existencia de autocorrelación. **2.** Se determina la no existencia de heteroscedasticidad en los datos. **3.** Se define el modelo de regresión lineal simple. **4.** Se persisten todos los datos calculados y los resultados obtenidos en archivo con formato de salida para que se pueda cargar el modelo obtenido en el Sistema.

E.3 Estimar Rendimiento sin Modelo

Descripción

El usuario solicita estimar el rendimiento para una zafra para la cual tiene los datos recolectados de las mismas variables que la zafra con la cual va a obtener el modelo.

Pre-condiciones

Debe estar definido el archivo con los datos recolectados en formato ZF.

Flujo de eventos principal

Usuario	Sistema
1. Solicita la estimación del rendimiento en base a datos (inclusive rendimiento) de la zafra anterior y los datos obtenidos para la nueva zafra.	
	2. Caso de Uso: ***Obtener Modelo***
	3. Se cargan los datos de la nueva zafra.
	4. Se estiman los rendimientos para cada uno de los puntos de la chacra.
	5. Se persisten todos los datos calculados en formato idéntico a los datos de entrada

con la variable rendimiento estimado.

E.4 Estimar Rendimiento con Modelo

El usuario solicita la obtención del mejor modelo que se adapta a los datos

F. Documento: Especificación preliminar del proyecto

Introducción al Problema

Se desea resolver mediante un Sistema Informático la predicción del rendimiento y las nuevas zonas de manejo, en una chacra en particular, en base a datos históricos y zonas de manejo históricas de la misma.

Los datos históricos que se tienen de la chacra pueden variar y serán definidos por el cliente en el documento de requerimientos. Se deberá establecer un formato de normalización para los mismos, dado que el uso que se le dio a la chacra en los distintos puntos del tiempo, puede haber variado; desde cultivar trigo, soja, hasta dejarlo descansar en pastoreo.

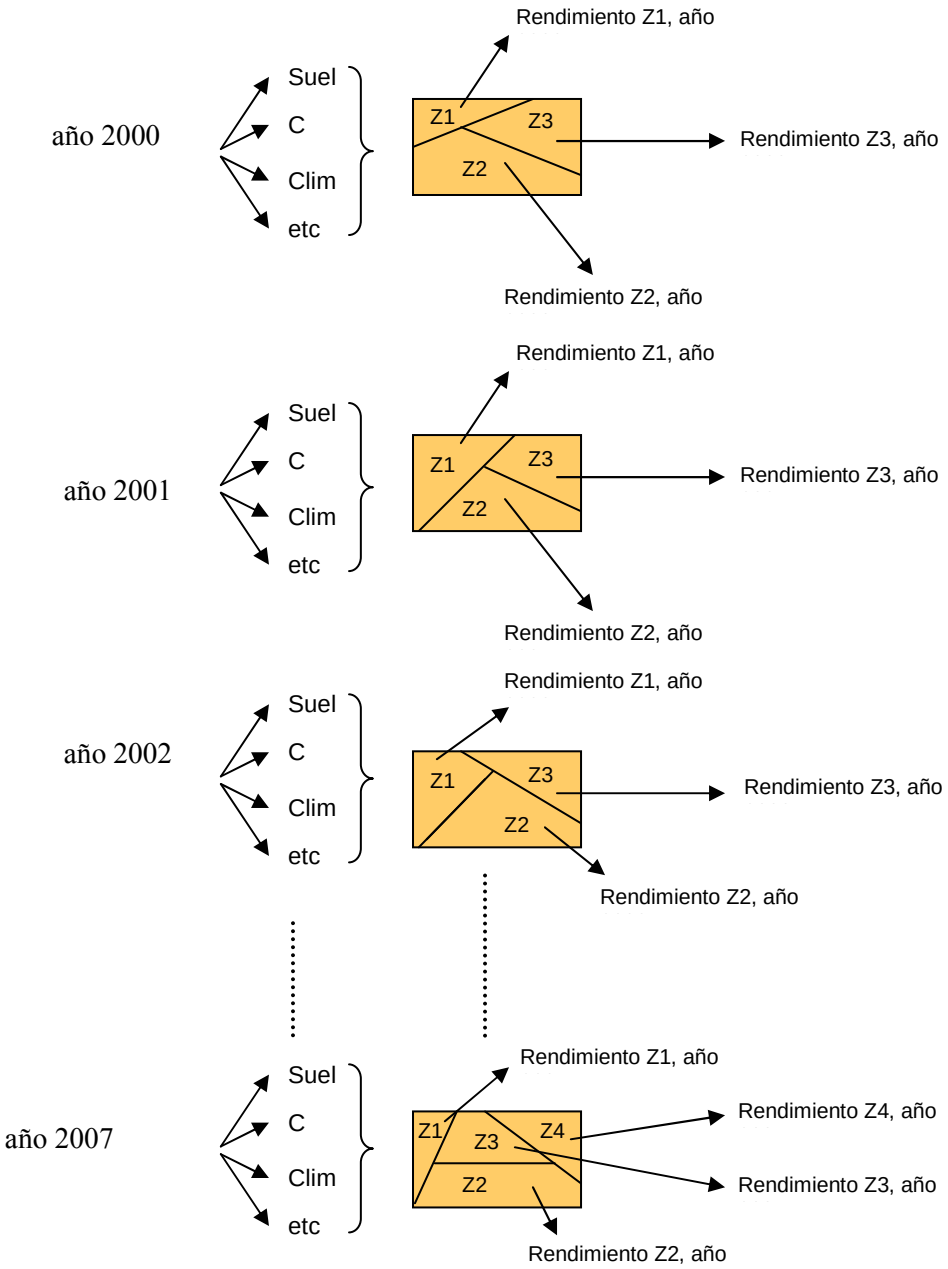
Entradas del Sistema

Dado una chacra 

Se proporcionaran como entrada al Sistema los siguientes datos:

- Series de tiempo con sus respectivas Zonas de Manejo, índices de performance y rendimientos en cada una de ellas. Entendemos por índices de performance a aquellos índices que ayudan a decidir objetivamente el número óptimo de zonas de manejo para la chacra.
- Otros datos, como pueden ser: información del clima, atributos del suelo, etc.

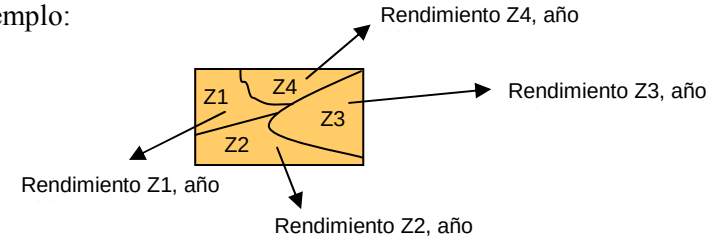
Un ejemplo de datos de entrada seria el siguiente:



Salidas del Sistema

Mediante técnicas de predicción y correlación, el Sistema deberá poder predecir cuales serán las nuevas zonas de manejo definidas para la chacra, para el siguiente ejercicio; por ejemplo para el año siguiente, esto es: si los datos son anuales, y llegan hasta el año 2007, se predecirán los resultados del año 2008. Además para cada zona definida para ese año, se deberá predecir el rendimiento estimado para cada zona de la chacra en ese año. La salida puede tener además otros índices de performance que serán definidos por el cliente a nivel de requerimientos.

Como ejemplo:



Definiciones necesarias a cargo del cliente a nivel de requerimiento

- Estructura y formato de datos de entrada y salida.
- Entrada y Salida persistente o a nivel de memoria.
- Tipo de sistema requerido:
    - Biblioteca
    - Consola
    - Web Service
    - Etc.
- Sistemas o subsistemas con los cuales debe ser integrado.
- Suposiciones que se realizaran con respecto al clima para hallar la salida.
- Se deberá definir si la prescripción para las zonas de manejo de salida son tenidas en cuenta para la predicción

Glosario

Entrada del Sistema: Se considera como entrada de un Sistema informático todos aquellos datos necesarios para producir la salida del Sistema.

Salida del Sistema: Se considera salida de sistema aquellos datos que una vez finalizada la invocación o ejecución del sistema se muestran como resultados.

Documento de requerimientos: Documento en el cual se especifica lo que el cliente pretende que el Sistema Informático realice. Se definen las entradas y las salidas del Sistema así como también otros aspectos técnicos del mismo.

Estructura de datos: Se refiere a cual es el formato en que se presentan los datos tanto a nivel de entrada como de salida en el sistema.

Entrada y/o Salida persistente: Se considera persistencia del Sistema informático a todo aquello que va a ser almacenado en almacenamiento secundario (memoria no volátil) como puede ser un disco duro, disco compacto, cinta, etc.

Entrada y/o Salida en memoria: Se considera como entrada o salida en memoria a todo aquello que no es guardado en almacenamiento secundario pero si se almacena temporalmente en memoria volátil.

Memoria Volátil: Es aquella cuya información se pierde al interrumpirse el flujo de corriente eléctrica, comúnmente conocida como memoria RAM o memoria de acceso aleatorio.

Memoria no volátil: Memoria cuyo contenido no se pierde al interrumpirse el flujo eléctrico que la alimenta.

Biblioteca: Se denominan bibliotecas a sistemas empaquetados de tal forma que puedan ser invocados por otras bibliotecas, sistemas o subsistemas.

Web Service: Es una colección de protocolos y estándares que sirven para intercambiar datos entre aplicaciones. Distintas aplicaciones de software desarrolladas en lenguajes de programación diferente y ejecutada sobre cualquier plataforma pueden utilizar los servicios Web para intercambiar datos en redes de ordenadores como Internet

Interoperacion de Sistemas: Es la condición mediante la cual sistemas heterogéneos pueden intercambiar procesos o datos pudiendo integrarse para interactuar entre si.