Aplicación de técnicas de minería de datos geo-referenciados en los circuitos de comercialización alternativa de productos agrícolas en Ecuador.

Autor:

Washington Raúl Padilla Arias

En cumplimiento parcial de los requisitos para el grado de Doctor en Ciencia y Tecnología Informática

Universidad Carlos III de Madrid

Directores:

Dr. D Jesús García Herrero

Dr. D José Manuel Molina López

Tutor:

Dr. D Jesús García Herrero

Colmenarejo, mayo 2019

Agradecimiento

Hace un tiempo escuche una frase que me llamo la atención "Somos el resultado del esfuerzo de muchos", que me parece muy oportuna en este momento que nos encontramos en la etapa final de este proyecto.

Y un poco para hacer alusión al trabajo desarrollado, las personas que mas cercanas se encuentran son las que mas han aportado con su ayuda, así debo agradecer en distintos ámbitos: familiar, académico, institucional.

Comenzando por el circulo mas cercano, mi reconocimiento la familia que ha tenido que sacrificar muchas horas de compartir actividades, a mi esposa Irlanda, mis hijos Sharon y Daniel, que siempre estuvieron alentándome desde que esta aventura fue solo un sueño.

A Jésus García que se constituyo en la piedra angular de esta construcción, un agradecimiento no solo personal sino de toda mi familia, tienes un sitio especial en nuestros corazones.

A José Manuel Mólina que siempre estaba listo a iluminar el sendero manteniendo firme el propósito emprendido.

En la parte académica un agradecimiento a la Universidad Politécnica Salesiana del Ecuador y en especial a su rector Dr. Javier Herran por haber

confiado y autorizado la ayuda económica y laboral para desarrollar este programa.

A los integrantes de los grupos de investigación IDEIA GEOCA de la UPS en Ecuador y GIAA de la UC3M en España, sus observaciones y sugerencias contribuyeron a mejorar la tarea emprendida.

A los mencionados un profundo agradecimiento y hasta siempre su re-cuerdo.

W Padilla.

CONTENIDOS PUBLICADOS Y PRESENTADOS

El material de esta fuente incluido en la tesis no está señalado por medios tipográficos ni referencias.

Evento/URL	**Titulo**	**Autor**	**Con-tribución**
International Conference on Practical Applications of Agents and Multi-Agent Systems 2016 https://link.springer.com/chapter/10.1007/978-3-319-39387-2_26	CIALCO: Alternative marketing channels [116]	Washington R Padilla, H Jesús García	Incluida parcialmente cap: III, IV, VI
2018 21st International Conference on Information Fusion (FUSION) https://ieeexplore.ieee.org/abstract/document/8455594	Model Learning and Spatial Data Fusion for Predicting Sales in Local Agricultural Markets[117]	Washington R Padilla; García H. Jesús ; Jose M. Molina	Incluida parcialmente cap: III, V, VI,VII,VIII

Evento/URL	Titulo	Autor	Con-tribución
International Conference on Practical Applications of Agents and Multi-Agent Systems 2018 https://link.springer.com/chapter/10.1007/978-3-319-94779-2_21	Information Fusion and Machine Learning in Spatial Prediction for Local Agricultural Markets [118]	Washing-ton R. Pa-dilla, Jesús García, José M. Molina	Incluida parcial-mente cap: III, VI, VII
International Conference on Hybrid Artificial Intelli-gence Systems 2018 https://link.springer.com/chapter/10.1007/978-3-319-92639-1_40	Improving forecast-ing using infor-mation fusion in lo-cal agricultural markets[119]	Washing-ton R Pa-dilla, Jesús García, José M Molina	Incluida parcial-mente cap: III, IV, VII, VIII

Evento/URL	Titulo	Autor	Con-tribución
Revista Sensors 2019 https://www.mdpi.com/1424-8220/19/2/286	Knowledge Extraction and Improved Data Fusion for Sales Prediction in Local Agricultural Markets[120]	WR Padilla, J García, JM Molina	Incluida parcialmente cap; I, III, V,VI,VII,VII
Revista Computational intelligence and neuroscience 2018 https://www.hindawi.com/journals/cin/2018/6587049/abs/	Data Association Methodology to Improve Spatial Predictions in Alternative Marketing Circuits in Ecuador[121]	WR Padilla, J García	Incluida parcialmente cap: III, V, VI,VII,VIII

Evento/URL	Titulo	Autor	Con-tribución
Libro PUBLICACIONES ALTARIA, S.L. www.publicacionesaltaria.com gerencia@altariaeditorial.com C/ Córcega, 60 08029 – Barcelona Tel. 647827278 ISBN – 978-84-947319-6-9	*CIENCIA DE DATOS*. MINERÍA DE DATOS. BIG DATA *Técnicas analíticas de aprendizaje estadístico. Un enfoque práctico*.	Jesús García, José M. Molina, Antonio Berlanga, Miguel A. Patricio, Álvaro L. Bustamante y Washington R. Padilla	Incluida parcialmente cap: V, VII

Resumen

A nivel mundial se utilizan sistemas de información para realizar el seguimiento y optimización de la producción agrícola.

En el Ecuador el ministerio de Agricultura y Ganaderia (MAG), tiene un programa orientado a fortalecer la asociación de productores agrícolas familiares que comercializan sus productos de manera directa con el consumidor, en un denominado circuito alternativo de comercialización (CIALCO).

A la información recolectada por el MAG, de ferias tipo Cialco, ubicadas en las provincias de Tungurahua y Chimborazo, se aplican técnicas de minería de datos descriptivas y predictivas, para descubrir patrones de comportamiento que permitan optimizar la utilización del suelo y mejorar los ingreso en la comercialización de productos agrícolas de este sector.

En la parte descriptiva, basados en la inducción de reglas de asociación, generadas utilizando los algoritmos A priori y FP-growth con parámetros mínimos de soporte y confianza, se genera un conjunto que se compone de todos los elementos resultado de obtener las mejores reglas.

El conjunto asociativo resultante se integra por los productos cebolla blanca, tomate de árbol, zanahoria, brócoli y tomate riñón.

En la parte predictiva se busca realizar una estimación pronóstica utilizando dos dimensiones: tiempo y ubicación geográfica.

Con un solo predictor, se genera una serie de tiempo utilizando el algoritmo SMOReg, para realizar una extrapolación pronostica con la que se encuentra valores de comercialización de productos agrícolas fuera del periodo de registro de información.

Adicionando coordenadas geográficas a la información inicial se ubican espacialmente las ferias en la región de estudio, compuesto por las provincias de Tungurahua y Chimborazo, para utilizar la dimensión espacial y en base a procesos de kriging realizar interpolación pronóstica para estimar valores de comercialización en lugares donde no se tiene información.

Una vez desarrollados estos tres procesos de minería de datos se propone una metodología qué, utilizando el conjunto asociativo como predictor, vuelve a calcular la estimación pronostica para la dimensión tiempo y la dimensión espacio.

La comparación de resultados con un solo predictor frente a los resultados de estimación pronóstica utilizando el conjunto asociativo como predictor indican que los porcentajes de error en la estimación pronostica multivariable disminuyen de manera considerable.

Para validar los resultados obtenidos de mejora de estimación pronostica, se crean dos modelos de datos utilizando variables externas al proceso de comercialización población y piso climático.

En los resultados finales, se aprecia que las dos variables de forma independiente muy poco aportan en la disminución del error de estimación, mientras que si se las hace interactuar con el conjunto asociativo se vuelve a encontrar una disminución en el error de estimación pronóstica obtenido.

Palabras clave: Reglas de Asociación, Series de tiempo, estimación pronostica, extrapolación pronostica, Regresión lineal, variable aleatoria, variograma, kriging, cokriging, fusión de datos.

Abstract

At the world level, information systems are used to monitor and optimize agricultural production.

In Ecuador, the Ministry of Agriculture and Livestock has a program aimed at strengthening the association of family agricultural producers, who market their products directly with the consumer, in a so-called alternative marketing circuit (CIALCO).

To the information collected from Cialcos-type fairs, located in the provinces of Tungurahua and Chimborazo, descriptive and predictive data mining techniques are applied.

To discover patterns of behavior that allow to optimize the use of the soil and improve the income in the commercialization of agricultural products.

In the descriptive part, based on the induction of association rules, generated using the Apriori and FP-growth algorithms with minimum support and Confidence parameters, a set is generated that consists of all the elements resulting from obtaining the best rules.

The resulting associative set is integrated by the products white onion, tree tomato, carrot, broccoli and tomato kidney.

The resulting associative set is integrated by the products: white onion, tree tomato, carrot, broccoli and tomato kidney.

In the predictive part, a prognostic estimation is sought using two dimensions: time and geographic location.

With a single predictor, a series of time is generated using the SMOReg algorithm, to perform a forecast extrapolation with which commercialization values of agricultural products are found outside the period of information registration.

By adding geographical coordinates to the initial information, the fairs are located spatially in the study region, composed of the provinces of Tungurahua and Chimborazo, to use the spatial dimension and based on kriging processes to perform prognostic interpolation to estimate marketing values in places where you do not have information.

Once these three processes of data mining have been developed, it is proposed to establish a methodology that, using the associative set as a predictor, recalculates the forecast forecast for the time dimension and the space dimension.

The comparison of results with a single predictor versus the results of prognostic estimation using the associative set as a predictor they indicate that the

percentages of error in the multivariable forecast estimate decrease considerably.

In order to validate the results obtained from improvement of forecast estimation, two data models are created using variables external to the population and climatic floor marketing process.

In the final results, it can be seen that the two variables inde- pendently contribute very little in the reduction of the estimation error, whereas if they are made to interact with the associative set, they will find a decrease in the error obtained.

Índice general

Índice de Figuras

Índice de Tablas

I. Parte

La presentación de este trabajo se la realiza en tres partes: La primera enfoca el problema planteado, la segunda parte se encarga del desarrollo teórico de técnicas de minería de datos y la tercera parte la aplicación de la metodología propuesta que mejora la estimación pronóstica utilizando las dimensiones: tiempo y espacio.

Esta primera parte se divide en dos capítulos: el primero, denominado Introducción, da a conocer esfuerzos a nivel mundial, regional y nacional por monitorear y establecer estimaciones a futuro sobre procesos de producción agrícola, para el Ecuador es muy importante realizar un seguimiento sobre la agricultura familiar campesina y sus procesos de asociación para comercialización de productos en circuitos alternativos de comercialización denominado Cialco.

Además, establece las técnicas de minería de datos aplicadas en la búsqueda de patrones de comportamiento, relaciona los temas de asociación, temporalidad y ubicación geográfica para presentar una propuesta metodológica que mejora los procesos de estimación pronostica en comercialización de productos agrícolas en centros alternativos de comercialización tipo feria

El capítulo II presenta la motivación para su desarrollo, los objetivos generales y específicos de la investigación, el aporte realizado y la estructura general de este documento.

1. Introducción

Un gran esfuerzo en la producción agrícola global sostenible, Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas[1], se ha llevado a cabo en los últimos años, especialmente en cumplir el objetivo 2: "Acabar con el hambre, lograr la seguridad alimentaria y mejorar la nutrición, y promover una agricultura sostenible", pero muchos factores interactúan, cambios climáticos, aumento de la población y la riqueza, escasez de agua y aumento de los costos de energía entre otros.

Se utilizan dos perspectivas complementarias para enfrentar de mejor manera las interrupciones en el suministro de alimentos y las fluctuaciones de precios en el mercado mundial de cultivos:

1. Un sistema de monitoreo que garantice información actual, oportuna y precisa sobre la producción de alimentos

2. Técnicas de pronóstico mejoradas permiten una comprensión de los riesgos clave en el suministro de alimentos

Existe varios sistemas de monitoreo agrícola a nivel nacional, regional y mundial, que han funcionado por décadas, los principales sistemas mencionados en [2] son:

• Sistema mundial de información y alerta temprana (Global Information and Early Warning System GIEWS) de la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO Food and Agriculture Organization).

• Red de sistemas de alerta temprana contra el hambre (Famine Early Warning Systems Network FEWSNET) de la Agencia de los Estados Unidos para el Desarrollo Internacional (USAID United States Agency for International Development)

• Sistema de monitoreo de la agricultura con teledetección (Monitoring Agriculture with Remote Sensing MARS) de la Comisión Europea (European Commission CE).

• CropWatch dirigido por el Instituto de Teledetección y Tierra Digital en la Academia de Ciencias de China, evalúa la producción de cultivos a nivel nacional y mundial.

• GEOGLAM (GEO GLobal Agricultural Monitoring), es una iniciativa emblemática de GEO (Group on Earth Observations GEO), que fue aprobada por el G20 en 2011, para compartir información a nivel internacional y producir dos boletines regulares (uno para el Sistema de Información del Mercado Agrícola (AMIS) y otro para Alerta Temprana que cubre aproximadamente el 95% de las tierras de cultivo del mundo).

Para América Latina y el Caribe, la FAO, presenta la iniciativa regional 2 en busca de mejorar los accesos del agricultor familiar campesino a recursos y servicios para fortalecer su actividad[3], se enmarca en los Objetivos de Desarrollo Sostenible (ODS) y apoya el principal acuerdo regional de erradicación del hambre: el Plan para la Seguridad Alimentaria, Nutrición y Erradicación del Hambre 2025 de la Comunidad de Estados Latinoamericanos y Caribeños (CELAC).

En esta región, la Agricultura Familiar es la principal fuente de empleo agrícola y rural abarcando el 80% de explotaciones agropecuarias que representan alrededor de 60 millones de personas. Este tipo de modelo productivo conjuga a la agricultura, ganadería, silvicultura, pesca, acuacultura y pastoreo dentro de la misma explotación y provee en promedio entre el 27 y el 67% del total de la producción alimentaria para cada país en la región. [4]

En Ecuador, cada grupo familiar se encarga de mantener diversidad de cultivos[5] en pequeña superficie de terreno, destinando la producción al autoabastecimiento y la comercialización de excedentes ya sea por el intercambio monetario o por otros productos generados por grupos similares (Trueque) [6].

En respuesta a las limitaciones en la comercialización ha surgido desde la asociación de agricultores familiares, la iniciativa para crear espacios que fomentan el encuentro directo con los consumidores, en lo que se ha denominado Circuitos Alternativos de Comercialización (CIALCO), en el Ecuador el Ministerio de Agricultura y Ganadería (MAG) tiene por objetivo "Impulsar la agricultura familiar garantizando la soberanía alimentaria." [7], por lo que se encarga de generar políticas agrícolas para un desarrollo equitativo y sostenido del sector [8] para realizar su monitoreo.

La información que posee la Coordinación General de Redes de Comercialización del Ministerio de Agricultura y Ganadería, se obtienen como resultado de un proyecto conjunto con la FAO el 2014 "Año Internacional de la Agricultura Familiar", los datos se registran físicamente en el lugar que se generan y detalla el comportamiento de ventas de productos agrícolas realizados en circuitos tipo feria, donde intervienen asociaciones de productores agrícolas campesinos ubicados en la sierra central ecuatoriana en las provincias de Tungurahua y Chimborazo.

Los registros incluidos en un archivo, presenta los siguientes atributos: nombre de la feria, nombre del responsable, cantidad de productos comercializados, fecha de adquisición, ubicación geográfica de ferias y volumen de ventas, los grupos de productos como hortalizas, legumbres, cárnicos, lácteos, frutas, tubérculos y productos procesados.

A nivel mundial han desarrollado varios modelos y algoritmos para predecir el rendimiento de las producciones agrícolas, la predicción utiliza las propiedades del suelo y las condiciones ambientales como datos para encontrar correlaciones utilizando varias técnicas:

• Correlación lineal del rendimiento con las propiedades del suelo y las condiciones ambientales [9].

• Métodos lineales, especialmente regresiones lineales múltiples para predecir los rendimientos utilizando las propiedades del suelo[10].

• Métodos no lineales (ANN y lógicas difusas) para la predicción del rendimiento [11] [12] .

En general, en este tipo de estudios, los autores trabajan con diferentes factores, principalmente las propiedades del suelo y los insumos agrícolas. Varios factores no controlados podrían afectar la producción agrícola; por lo tanto, incluso los modelos complejos y matemáticos no pueden dar los resultados precisos. Uno de los principales factores está relacionado con la variabilidad climática [13].

A nivel nacional la información recolectada sobre venta de productos en circuitos de comercialización tipo feria presenta atributos como: nombre del producto, fecha de venta, valor de venta y ubicación geográfica, cada

campo reúne características que permiten aplicar técnicas de minería de datos descriptivas y predictivas.[14]

Transformando el valor de venta de cada producto a un valor binario que indique la presencia o ausencia en un proceso de comercialización, más el nombre del producto se aplica técnicas de cobertura mínima para generar el aprendizaje de reglas de asociación, el resultado de este proceso permite conocer productos asociados en el proceso de comercialización.[15], [16], [17].

Por otro lado, el atributo fecha de venta incorpora un componente de temporalidad, que permite establecer una serie de tiempo, en conjunto con nombre del producto y valor de venta se utilizan técnicas de regresión, y se genera un modelo de datos para estimar a futuro el comportamiento de ventas basado en los valores de comercialización anteriores.

Utilizando un discriminante lineal[18], como una máquina de vector soporte (SVM), se realiza el análisis [19] para las muestras semanales de venta de productos del año 2014, que permite conocer un modelo de datos y obtener estimaciones a futuro utilizando un número limitado de muestras.[20]

Para medir la validez de las estimaciones se divide al conjunto total de datos en dos subconjuntos: el setenta por ciento de las muestras se incluyen en el conjunto de entrenamiento(training) y lo restante en el conjunto de prueba(test), la calidad del modelo generado para estimar el valor de venta

a futuro utilizando la serie de tiempo se mide con la relación entre el valor de venta predicho menos en valor de venta real utilizando métricas de error absoluto medio (MAE) y error cuadrático medio (RMSE).

La ubicación geográfica[21] de una feria agrega el componente espacial, que permite utilizar algoritmos de geoestadística, apoyados en el atributo de valor de venta de un producto en ferias cercanas (vecinas) [22], se estiman los valores de venta de productos agrícolas en lugares donde al momento no existe un circuito de comercialización tipo feria en funcionamiento, [23], [24].

Existen trabajos[25] que tratan de relacionar las reglas de Asociación y las series de tiempo, [26], explican cómo se pueden extraer reglas de asociación en base a una temporalidad continua, los resultados se enfoca en el número de reglas descubiertas. [27]

Se encuentran tareas investigativas sobre la asociatividad en la parte espacial [28], orientado a la generación de reglas de asociación tomando en cuenta el componente espacial. [29]

Las investigaciones analizadas parten de utilizar técnicas de minería de datos predictivas como series de tiempo y geoestadísticas que inciden en mejorar los resultados de utilizar técnicas descriptivas como la generación de reglas de asociación. El trabajo propuesto se basa en utilizar técnicas de

minería de datos de tipo descriptivo para establecer su incidencia en técnicas de tipo predictivo, en particular se plantea una metodología que utiliza reglas de asociación para encontrar el conjunto de productos asociados con mayor demanda en el proceso de comercialización en ferias, utilizando este conjunto se ejecutan dos tareas de tipo predictivo: estimaciones a futuro con series de tiempo y estimación de ventas en ferias utilizando el componente espacial.

Los resultados de las estimaciones utilizando productos asociados (multivariable) se comparan con las estimaciones de predicción utilizando un solo producto.

Adicionalmente utilizando dos variables externas al proceso de comercialización: población y piso climático, se aplica la metodología propuesta para conseguir un nuevo modelo predictivo.

A cada modelo: el que utiliza una variable, al modelo multivariable en base a la utilización de reglas de asociación y al modelo con variables externas de población y piso climático se realiza una evaluación mediante validación cruzada, con el procedimiento LOOCV (Leave one out cross validation), que evita un sobre o sub-ajuste, los residuos obtenidos, presentan una disminución considerable en la varianza del calculo de error, en los casos que se utiliza el conjunto de productos asociados.

La utilización de técnicas de minería de datos como la inducción de reglas de asociación, establecer pronósticos en base a series de tiempo y la utilización de un componente espacial permiten realizar un monitoreo del comportamiento de comercialización de los productos en circuitos tipo feria, aplicada la metodología propuesta se advierte la disminución de la varianza de los valores de error , todos estos resultados se enfocan en ayudar a reducir la inseguridad alimentaria y proponer mejores técnica de predicción.

2. Motivación y Objetivos

A nivel mundial el sector agrícola representa un eje fundamental en la provisión alimentaria, de ahí que organismos como las Naciones Unidas lo incluyan en sus objetivos de desarrollo y se enfoca en conseguir sistemas productivos sostenibles.

La organización de las Naciones Unidas para la alimentación y agricultura (FAO) a nivel de América Latina y el Caribe pone especial énfasis en fortalecer al sector productor agrícola.

En el Ecuador se encuentran en funcionamiento circuitos alternativos de comercialización que son lugares físicos establecidos donde periódicamente se realizan actividades de comercialización de productos agrícolas, se caracteriza por tener una relación directa entre productores y consumidores.

La información registrada sobre CIALCO se limita a ferias que funcionan en las provincias de Tungurahua y Chimborazo en el año 2014, lo que no permite ser parte de un monitoreo constante ni establecer sistemas de predicción de la producción agrícola familiar, toda la información generada no puede ser considerada en el contexto nacional.

El problema planteado tiene una serie de componentes que pocas veces se pueden conjugar. Por un lado, la necesidad de establecer escenarios a futuro que ayuden a generar políticas públicas en beneficio de un sector que socialmente se considera vulnerable como es el agricultor campesino ecuatoriano, el enfoque para mejorar los ingresos familiares y disminuir con la migración de sectores rurales a grandes centros de desarrollo poblacional, la conservación de saberes ancestrales, así como el trabajo colaborativo en una familia y el asociativo entre grupos cercanos, además de mejorar la provisión y soberanía alimentaria crean un entorno de desarrollo social.

Por otro lado, la información recogida el 2014 declarado "Año Internacional de la Agricultura Familiar", en las provincias de Tungurahua y Chimborazo, permite descubrir patrones de comportamiento en ventas, aplicando técnicas de minería de datos de tipo predictivo y descriptivo, por ejemplo, es interesante conocer que grupos de productos son más comercializados o dependiendo de su estacionalidad que productos son más requeridos por los consumidores en una parte del año, también podemos conocer el comportamiento de un producto en base a la ubicación geográfica.

Estos elementos conjugan objetivos de tipo social, cultural, económico, tecnológico en busca de proveer un fortalecimiento al sector de la agricultura familiar campesina en la región andina del Ecuador.

La mayor limitante encontrada es la escasez de información, se dispone de datos, pertenecientes a un solo año, lo que dificulta la extracción de patrones con técnicas estadísticas tradicionales por lo que se recurre a implementar procesos de aprendizaje en base a técnicas de minería de datos complejas como series de tiempo y geoestadística.

No tener un seguimiento continuo a los datos de comercialización generados por el sistema de producción de CIALCO tipo feria en las provincias de Tungurahua y Chimborazo, lo aisla del contexto nacional haciendo más difícil su monitoreo y predicción de comportamiento a futuro, lo que disminuye la eficiencia productiva.

2.1 Titulo de la Investigación

Esta investigación lleva por título:

"Aplicación de técnicas de minería de datos y georeferenciados en los circuitos de comercialización alternativa de productos agrícolas en Ecuador".

2.1.1 Objetivo General

Analizar la información de circuitos alternativos de comercialización tipo feria de las provincias de Tungurahua y Chimborazo del año 2014, para encontrar modelos de comportamiento en comercialización de productos, utilizando reglas de asociación, series de tiempo y datos espaciales, el modelo

generado permite establecer escenarios a futuro como herramienta en la toma de decisiones o generación de políticas que favorezcan a la agricultura familiar, el desarrollo territorial y la soberanía alimentaria en el Ecuador.

Basado en la aplicación de estos tres algoritmos de aprendizaje se propone una nueva metodología: utilizar el conjunto de productos integrado por los productos resultado de inferir reglas de asociación, como un predictor multivariable.

Con este nuevo predictor, se vuelve a ejecutar los procesos de minería de datos para la extrapolación pronostica utilizando series de tiempo y la interpolación pronóstica utilizando el componente espacial.

2.1.2 Objetivos Específicos

Los objetivos específicos son:

1. Realizar procesos de limpieza y validación de la información de ferias de las provincias de Tungurahua y Chimborazo
2. Utilizar tareas descriptivas para crear el conjunto de datos asociados utilizando índices de soporte y confianza para conocer los productos más comercializados en estas ferias.
3. Utilizar técnicas de minería de datos de tipo predictivas, para establecer estimaciones a futuro de producción y comercialización de productos en ferias utilizando series de tiempo.

4. Con la información georreferenciada, calcular el valor estimado de venta de productos en un CIALCO por establecer
5. Proponer una metodología para mejorar la la estimación de comercialización a futuro utilizando un conjunto asociativo multivariable y aplicandolo técnicas de serie de tiempo y datos espaciales.
6. Establecer un nuevo modelo de datos en base a variables externas al proceso de venta, aplicar la metodología propuesta y medir el impacto del modelo fusionando con el modelo asociativo para estimaciones de ventas en ferias.
7. Generar escenarios que permitan realizar un monitoreo y sirvan de apoyo en la toma de decisiones para el desarrollo de los centros de comercialización alternativa (CIALCO).

2.2 Contribución

Este proyecto tiene definido dos áreas: en la parte social permite conocer el mejor aprovechamiento de las granjas de agricultura familiar campesina determinando la mejor combinación de cultivos, la época del año más apropiada para producir o posibles niveles de comercialización donde no existen ferias.

A nivel técnico la utilización de elementos de minería de datos en CIALCO tipo feria que extraen patrones de comportamiento, en base a técnicas

descriptivas el conjunto de productos comercializados con mayor asociatividad, utilizando técnicas predictivas como la regresión lineal se establece predicción de ventas aplicando series temporales y datos espaciales

La contribución de esta tesis es presentar una metodología que establece que las estimaciones a futuro aplicando series de tiempo y geoestadística mejoran cuando se utiliza el conjunto de productos resultado de la inferencia de reglas de asociación.

2.3 Organización de la Investigación

El trabajo presentado consta de tres partes generales: la primera contiene la introducción y los objetivos, la segunda parte se encarga del estado del arte de las técnicas de minería de datos empleadas, la tercera parte: "Propuesta, desarrollo y caso de estudio", presenta el trabajo realizado en el conjunto de datos y en su procesamiento. En la parte final se encuentra conclusiones y expectativa de trabajo a futuro.

El capítulo uno da una visión general del problema alimentario presentándolo como un objetivo de desarrollo, identifica las distintas Instituciones que se encargan del monitoreo de la producción alimentaria en el mundo y en la región de América Latina y el Caribe, presenta al Ministerio de Agricultura como la entidad encargada de la Agricultura Familiar en el Ecuador

La información inicial se refiere a ferias que son un tipo de Circuito alternativo de comercialización ubicadas en las provincias de Tungurahua y Chimborazo en la zona centro sur de la región andina del Ecuador, consta de ubicación geográfica, nombre del responsable de la actividad comercial, nombre del producto, fecha de comercialización, unidad de comercialización y valor unitario, organizadada en registros del año 2014 con periodicidad semanal, sobre esta información se propone aplicar tres tipos de técnicas de minería de datos: reglas de asociación, series de tiempo y datos espaciales.

Finalmente, en este capitulo se presenta el desarrollo de una metodología que utilizando un conjunto de productos resultado de obtener las mejores reglas de asociación como predictor, permite disminuir el error de estimación en procesos de predicción temporal y geoestadístico.

El capítulo dos presenta la motivación por la que se realiza la investigación, identifica la connotación social y tecnológica del proyecto, se da a conocer el nombre de la investigación, así como la descripción detallada de los objetivos general y específicos, y una descripción general de la estructura de investigación presentada.

La segunda parte se enfoca en la descripción conceptual y algorítmica de cada una de las técnicas de minería de datos empleadas como son la búsqueda de patrones de comportamiento para la estimación a futuro

tomando en cuenta la asociatividad, temporalidad y ubicación geográfica de los datos.

En los tres capítulos que conforman esta segunda parte se sigue una misma estructura, aborda formalmente la concepción teórica, realiza la presentación de algoritmos utilizados en los procesos de minería de datos para la búsqueda de patrones de comportamiento a futuro, finalmente se evidencia la literatura consultada sobre la resolución de problemas que tienen elementos en común o utilizan el mismo tipo de técnicas o algoritmos planteados en este trabajo.

El capítulo tres está centrado en la generación de reglas de asociatividad validas, teniendo como base los índices de soporte y confianza elementos determinantes en los algoritmos utilizados Apriori y Fp-growth, las reglas generadas son del tipo $A \rightarrow B$ que indica una ocurrencia múltiple, es decir que si compro el producto que se encuentra en el lado izquierdo (A) de la regla va acompañado de la compra del producto de la derecha (B), pero también puede ser interpretado como si no compro B tampoco A, esta relación se da no como una casualidad sino como la ocurrencia en simultaneo.

Además, se realiza una presentación del estado del arte con un detalle de la aplicación de reglas de asociación en diferentes actividades comenzando con la ubicación de productos en un supermercado, o la búsqueda

de comportamientos de un visitante en una página web, así como la implementación de mejoras o nuevas alternativas para conseguir reglas de asociación válidas para distintos tipos de problemas.

El capítulo cuatro describe la búsqueda de patrones basadas en la temporalidad utilizando series de tiempo para la elaboración de pronósticos a futuro, con base a la información registrada en el pasado. Comienza con la descripción de los procesos de tipo estadístico para la formulación de pronosticos utilizando series de tiempo y presenta el desarrollo teórico de algoritmos que se desprenden de la utilización de SVM (Support Vector Machine) específicamente SMOreg (Sequential Minimal Optimization for Regression).

La evaluación de pronósticos se realiza utilizando dos métricas: el error medio absoluto (Mean absolute error) y el error medio cuadrático (Root mean squared error), la parte final explora trabajos realizados por distintos autores, sobre el tema de pronosticos utilizando series de tiempo.

El capítulo cinco se encarga de la descripción teórica de los procesos de Geoestadística, presenta conceptos de variable regionalizada, función aleatoria y variograma para establecer la diferencia entre dos puntos geográfico.

Basado en los conceptos de variograma experimental y variograma de modelo ajustado, utiliza técnicas de Kriging y Cokriging para interpolar atributos

en una región para obtener una estimación de comportamiento de una variable donde no existe valores conocidos.

La evaluación de la calidad de la predicción se obtiene mediante validación cruzada LOOCV.

En la segunda parte del capitulo, se analizan las distintas formas de realizar una fusión del modelo de datos desde el punto de vista de las fuentes de recolección de información.

Para mantener la misma estructura de los capítulos anteriores, al finalizar se presentan una serie de trabajos relacionados.

La tercera parte presenta la metodología propuesta y el detalle de los pasos realizados para su aplicación, explica los resultados obtenidos y las conclusiones.

El capítulo seis se centra en la preparación de datos, limpieza y estandarización de información, describe a detalle los procesos aplicados para transformar la información en archivos discretizados que permitan obtener reglas de asociación válidas, describe la manera de temporalizar la información para ser utilizados por las tareas de estimación a futuro utilizando series de tiempo y los requisitos necesarios para obtener información que incluya el componente espacial.

Para finalizar este capítulo, se adiciona datos externos al proceso de compra de productos, que se los utiliza para realizar la fusión de modelos.

El capítulo siete presenta el trabajo de aplicación de la metodología propuesta, detalla los procesos para generar reglas de asociación y conseguir el conjunto de productos que presentan la mejor asociatividad en el proceso de comercialización en circuitos alternativos de comercialización en las provincias de Tungurahua y Chimborazo.

La siguiente sección en este capítulo desarrolla las series de tiempo empezando desde un enfoque probabilístico describe los elementos de Tendencia, Ciclicidad, Estacionalidad y Aleatoriedad que resultan de la descomposición de una serie temporal y los procesos para obtener una estimación a futuro asumiendo una distribución normal.

Utilizado los algoritmos SMOreg, genera las estimaciones a futuro, este proceso se lo realiza en dos ocasiones: la primera utilizando una variable como predictor y la segunda calcula las estimaciones a futuro utilizando series de tiempo con los productos resultantes de obtener las mejores reglas de asociación como predictor.

Con las dos estimaciones a futuro se realiza una comparación de los resultados obtenidos en base a las metricas de error.

De manera similar se procede con los archivos de tipo espacial, se genera el variograma modelo y la estimación por medio de Kriging para una variable predictiva, en una segunda instancia se procede a generar un modelo de variograma multivariable, obteniendo valores de estimación para lugares donde no se tiene una medida real por medio del proceso Cokriging.

Se evalúa las predicciones realizadas utilizando los valores residuales encontrados al aplicar validación cruzada con el método Loocv.

Utilizando un procedimiento de multivariable, se inserta las variables de población y piso climático, calificadas como externas al proceso de comercialización de productos agrícolas y se verifica la validez de la metodología propuesta.

En el capitulo ocho se presenta un resumen de los valores obtenidos. para cada caso de estimación: utilizando tanto series de tiempo con una o múltiples variables, así como la estimación espacial con los dos casos de estudio una variable y el conjunto de productos asociados como predictor.

Se incluye el cálculo de pronósticos para variables externas al proceso de venta de productos, utilizando la técnica de coKrigin entre una variable asociada con cada una de estas nuevas variables externas.

Finalmente se incluyen los resultados y se describe las relaciones de mejora y los casos en que ha sido factible encontrar una variación significativa, los

casos en los que no se tiene una variación significativa y los casos en los que no se tienen ningún tipo de inferencia en los resultados presentados en el modelo fusionado.

Parte II. Estado del Arte

La búsqueda de sistemas predictivos para garantizar la provisión de alimentos ha llevado a plantear la utilización de técnicas de minería de datos descriptivas y predictivas.

De los datos registrados por el Ministerio de Agricultura y Ganadería del Ecuador, en el año 2014, para circuitos alternativos de comercialización ubicados en las provincias de Tungurahua y Chimborazo se extrae conocimientos aprovechando procesos de asociatividad de productos de comercialización, temporalidad y ubicación geográfica.

Esta segunda parte está conformada por los capítulos tres, cuantro y cinco, el capítulo tres aborda los procesos asociativos que se aplican en la inferencia de reglas de asociación para encontrar una relación de asociatividad entre productos comercializados, basa su busqueda en índices de soporte y confianza, el conjunto de productos asociados está compuesto por los productos resultantes de las mejores reglas encontradas utilizando dos algoritmos de aprendizaje Apriori y FP-growth.

Se realiza la descripción de los dos algoritmos y su aplicación en investigaciones en distintos ámbitos como son el marketing, el comportamiento humano, procesos de selección de estudiantes [30], hidrología.

La presentación de las relaciones por asociación permite encontrar patrones que no presenten una correlación de manera inmediata.

El capítulo cuatro topa temas relacionados con temporalidad, analiza dos formas de estudiar series temporales: una con procedimientos estadísticos y la segunda utilizando SVM (máquinas de vector soporte), utiliza metricas de evaluación para establecer la validez de los pronosticos, en la parte final presenta una serie de trabajos relacionados.

La predicción del comportamiento en procesos de comercialización se realiza utilizando las condiciones de temporalidad encontrada en el campo fecha de transacción, la baja cantidad de datos y el único periodo anual de información impide utilizar procedimientos estadisticos tradicionales y es necesario utilizar algoritmos de inteligencia artificial.

Se presenta la predicción de series temporales como una herramienta para descubrir patrones de comportamiento a futuro utilizando SVM (Support Vector Machine) y la función SVR(Support Vector Machine Regression). Se selecciona dos métricas de evaluación MAE (Mean absolute error) y RMSE (Root mean squared error) que se toman como referencia para establecer la disminución del error en los conjuntos de entrenamiento(training) y prueba (test) que establecen el nivel de éxito en la predicción numérica.

Las series de tiempo permiten el descubrimiento de patrones en diversos campos como son la hidrología, ventas, medicina, agricultura, etc y se basa

en generar un modelo de regresión en función de una variable que se encuentre relacionada con una secuencia de periodicidad, que puede ser semanal, mensual o anual. El desarrollo de este modelo determina valores desconocidos en función de los datos proporcionados por una secuencia conocida.

El capitulo cinco utiliza datos con una componente espacial que le permite realizar tareas de geoestadística, presenta el concepto de variograma y proceso kriging para establecer estimaciones de comportamiento utilizando una variable, en la siguiente sección presenta el variograma para modelos multivariables y procesoso de cokriging.

Establece la forma de evaluar las estimaciones realizadas utilizando validación cruzada.

En la parte final revisa los conceptos de fusión de modelo de datos para establecer nuevos modelos.

3. Reglas basadas en Asociación

De momento hemos referenciado a tareas de minería de datos de tipo descriptiva y predictiva, en este capítulo se abordan tareas de tipo descriptiva y la manera de generar reglas de asociación utilizando dos algoritmos: A priori y FP-growth.

3.1 Primeros conceptos

La obtención de patrones de comportamiento en base a reglas de asociación busca conjuntos de ítems que suceden frecuentemente y cumplen con una cobertura mínima, el proceso de encontrar elementos en conjuntos se basa en la probabilidad de ocurrencia de un elemento que implique el aparecimiento de otro elemento,[31].

Originalmente este método se utilizó para optimizar la ubicación de productos en los supermercados,[32]: se considera un conjunto de elementos (items) $I=\{i_1,i_2,\ldots.i_n\}$, y este conjunto se identifica el nombre de los productos

Sobre este conjunto de elementos se tiene el conjunto de transacciones $T=\{t_1,t_2,\ldots..,t_n\}$ $t_n \in T$, $t \subseteq I$, un valor identifica si en la t_n, existe el producto I,

Así en una tabla de tipo I x T se representa la presencia de un producto en una transacción determinada utilizando un valor binario [v] o la ausencia con un valor [f] Tab 3-1.

	I_1	……..	I_n
T_1	v	…….	v
….	f	…….	f
T_n	v	…….	v

Tab. 3-1 Tabla Productos Transacciones

Una regla de asociación se representa de la forma A $\rightarrow$B que se interpreta como la existencia de los elementos del conjunto A implica la existencia de los elementos del conjunto B.

"A" se le conoce como antecedente y significa que si aparece uno o varios elementos del conjunto A en el conjunto de transacciones, también aparece uno o varios elementos del conjunto B, ambos subconjuntos del conjunto de transacciones[33].

La calidad de una regla se evalúa en base al cumplimiento de índices de medida de una regla de asociación, como son el soporte, confianza, sustentación y convicción, descritos en el siguiente apartado.

3.2 Índices de medida de reglas de asociación

Una regla se considera valida o de suficiente calidad si cumple con los umbrales especificados para los índices de soporte y confianza, aunque también se la puede evaluar utilizando los índices de sustentación o convicción. Una regla generada es de calidad si cumple con el mínimo establecido para cada índice especificado.

3.2.1 Soporte (Support)

El soporte o también llamando cobertura, indica el número de veces que se da la regla dentro del conjunto de transacciones,

$$supp(A \rightarrow B) = Prob(A \cap B) = \frac{card(A \cap B)}{card(T)},$$

donde T es el conjunto de transacciones y card expresa el número de veces que se da el antecedente A

3.2.2 Confianza (Confidence)

La confianza o precisión indica el porcentaje de veces que la regla se cumple cuando se aplica

$$conf(A \rightarrow B) = \frac{supp(A \rightarrow B)}{supp(A)} = \frac{card(A \cap B)}{card(A)}$$

expresa la precisión de la regla al representar el porcentaje de aciertos de las predicciones.

3.2.3 Sustentación (lift)

Establece el nivel de dependencia entre el antecedente y el consecuente, mide en que grado la parte del consecuente (B) de una regla tiende a ser frecuente cuando aparece el antecedente (A), medido con relación al caso en que fuesen independientes.

$$lift = \frac{supp(A \rightarrow B)}{supp(A)\ x\ supp(B)}, = \frac{card(A \cap B)}{card(A).card(B)}$$

$$\text{lift} = \begin{cases} < 1, & B\ influye\ la\ no\ ocurrencia\ de\ A \\ = 1, & A, B\ independientes\ (A\ no\ depende\ de\ B) \\ > 1, & B\ influye\ la\ ocurrencia\ de\ A \end{cases}$$

3.2.4 Convicción (Conviction)

Indica la frecuencia con la que puede ocurrir una estimación no acertada

$$conv(A \rightarrow B) = \frac{1 - supp(B)}{1 - conf(A \rightarrow B)}$$

Evalúa el grado que A influye en la ocurrencia de B

En base a los parametros de soporte y confianza se presentan dos algoritmos para generar reglas de asociación Apriori y FP-growth

3.3 Algoritmo Apriori

En este algoritmo de aprendizaje la obtención de reglas de Asociación se realiza en dos etapas:

1.- Extraer conjuntos de ítems (itemset) que cumpla con el soporte mínimo especificado (minsup), un itemset es frecuente si cada uno de los ítems es frecuente por si solo

2.- A partir de los conjuntos se generan las reglas de asociación

El desarrollo del algoritmo está basado en el conocimiento previo o "apriori" de los conjuntos frecuentes y se define de la siguiente manera [34]

```
Paso 1

Generar todos los itemsets L con un único elemento

Se toman todos los posibles pares cuyo Sup sea igual a minsup

Este conjunto se utiliza en una nueva instancia para formar
uno nuevo con un elemento adicional

Paso 2

Por cada itemset frecuente L' encontrado

     Por cada subconjunto J de L'
```

Determinar todas las reglas de asociación de la forma:

Si L'-J→J

Seleccionar aquellas reglas cuya confianza sea mayor o igual a minconf

Se repite el paso uno, incluyendo otro elemento a L.

3.4 Algoritmo Fp_growth

Otro algoritmo para encontrar reglas de asociación se denomina FP-gro

wth [35] ,se basa en los mismos parametros de soporte y confianza pero genera reglas de asociación sin obtener elementos candidatos

La propuesta y algoritmo [36] se basa en que un conjunto de datos comparte ítems frecuentes, para lo cual crea un árbol de patrones frecuentes Fp-tree (Frequent Pattern Tree), esta estructura tiene un nodo principal con valor cero y un conjunto de sub árboles que se refieren a las transacciones, además crea una tabla que se refiere a la frecuencia de los ítems (header table).

Cada nodo del subárbol posee el nombre del ítem y la frecuencia que aparece en las transacciones y un apuntador al siguiente sub árbol que almacena el mismo ítem.

Al comenzar la construcción del árbol recorre el conjunto de transacciones en busca de ítems frecuentes y registra el número de ocurrencias, el resultado se ordena de mayor frecuencia a menor.

Los ítems de mayor frecuencia de transaccionalidad se ubican en los nodos de menor nivel, se vuelve a recorrer el conjunto de transacciones, pero solo tomando en cuenta a los ítems más frecuentes

Algoritmo [37]

1.- Explorar la fuente de datos primera vez, en busca de conjuntos frecuentes de un ítem (patrón de un solo elemento cumple minsuport)

2.- Crear lista-f ordenada descendentemente con los artículos frecuentes

3.- Construir FP-tree utilizando la fuente de datos

Construir FP-tree

Comenzando el encabezado del arbol FP-tree, con el primer elemento en la lista-f.

Construya el árbol de FP siguiendo el enlace de cada elemento frecuente p

Descomponga todas las rutas de prefijo transformadas del elemento p para formar p's nodos que constituyen el patrón-base condicional

Para cada patrón-base

Acumula el conteo para cada elemento en la fuente de datos

Construya el árbol FP-Tree para los elementos frecuentes del patrón base

3.5 Trabajos relacionados

El aprendizaje de reglas de asociación ha sido aplicado en muchas áreas del conocimiento, especialmente en las que permiten utilizar atributos de tipo nominal, por lo que en muchos casos es necesario discretizar los atributos de tipo numérico en un proceso inicial, las primeras aplicaciones para descubrir patrones asociativos se las realiza tomando en cuenta los productos de una cesta de supermercado el área de interés se centra en el conjunto de ítems discretizados escogidos en una transacción[38],[39], [40].

Otro campo para la inducción de reglas de asociación es la búsqueda de patrones en páginas web así se puede encontrar patrones de comportamiento de navegación de un visitante en base a los clics de ratón realizado sobre los ficheros disponibles en [41] se indica que cuando se tiene patrones

secuenciales el algoritmo “Apriori” puede ser ineficiente por lo que presenta un algoritmo mejorado.

Otro tipo de búsqueda de comportamientos utilizando reglas de asociación se encuentra en el comportamiento generado por un sensor para encontrar diferentes alarmas [42].

Basados en el algoritmo "Apriori" propone un nuevo algoritmo de minería en redes de sensores, los autores proponen un algoritmo eficiente de minería de datos para generar patrones de comportamiento de sensores utilizando PLT (Árbol Lexicográfico Posicional), que es una nueva estructura de representación para almacenar los datos de comportamiento de los sensores.

En [43] se presenta una interesante descripción para la consecución de reglas de asociación que presentan en especial dos limitaciones: por un lado no se refiere a atributos cuantitativos, por otro lado trata cada elemento con el mismo significado, muchas transacciones reales tienen atributos cuantitativos. Este trabajo presenta algoritmos para inducir reglas de asociación utilizando datos cuantitativos

En [44], se presenta el problema de patrones secuenciales y tres algoritmos para resolver el problema, AprioriSome, Aprioriall, Dynamicsome, el algoritmo planteado, de las propuestas realizadas AprioriSome tiene un mejor desempeño para casos que tienen un patrón secuencial bajo

Como una alternativa diferente, se discuten el paradigma de las reglas de asociación temporal pero de diferentes autores, considerando reglas dentro de marcos temporales, también basados en el esquema de minería Apriori extendido para considerar soporte temporal [45], o incluso meta-reglas que describen cómo las relaciones varían en el tiempo [46],[47].

EO-ARM[48] es un algoritmo más eficiente para encontrar reglas de asociación optimizadas al introducir una medida de correlación que elimina las reglas menos interesantes, EO-ARM se ha implementado utilizando un conjunto de datos de transacciones binarias, un soporte(supp) óptimo y un umbral de confianza.

En [49] se trata el tema de generación de reglas positivas y negativas, las positivas indican la presencia de un atributo, mientras que las negativas su ausencia, presenta la idea que se puede obtener más propiedades asociativas basados en reglas negativas.

Además en trabajos como [50] plantea un método para elegir los umbrales mínimos y máximos del soporte (supp) utilizando una distribución binomial, [51].

Utilizando los algoritmos Apriori y FP-growth se realiza un análisis de patrones de uso de un sitio web y descubre las características del conocimiento del comportamiento de usuarios,[52] realiza un análisis de una cesta de supermercado utilizando el uso frecuente de patrones.

En [53] se presenta una descripción del funcionamiento de un conjunto de algoritmos que nos permiten encontrar reglas de asociación.

Los elementos presentados en este capitulo se refieren a la forma de inducir reglas de asociación y los algoritmos mas conocidos para generarlas. En particular, para el dominio agrícola motivo de esta investigación, se busca el conjunto de productos resultado de encontrar las mejores reglas que asociación basados en la comercialización, del año 2014, que permita establecer una mejora en la precisión de técnicas predictivas como parte de de una nueva metodología propuesta.

4. Series de Tiempo

Una serie de tiempo S se describe como una secuencia de datos $(x_1, x_2, \ldots\ldots x_n)$, medidos en un cierto momento y ordenados cronológicamente, el análisis realizado con estas series permite con diferente grado de confianza extrapolar sobre los datos fuente, para estimar el comportamiento de la serie S en el futuro (extrapolación pronóstica).

Este capítulo presenta la parte conceptual para establecer estimación pronóstica a futuro basado en el análisis de una serie de datos con un componente de temporalidad con dos enfoques: la modelización estadística que necesita de una gran cantidad de datos para su análisis y técnicas basadas en estimador núcleo (kernel) que utilizan técnicas de regresión basadas en aprendizaje automático.

Las series temporales predicen comportamientos a futuro en base a datos reales y han sido de utilidad en varios campos de la ciencia como por ejemplo estimaciones sobre tasa de nacimientos, la cantidad de vehículos que circulan por una determinada área en una ciudad en un horario por la mañana, comportamiento de tasas de interés al largo de un periodo, mejorando los procesos de toma de decisiones. Así, en un hospital se puede administrar de mejor forma los recursos para recién nacidos, en una ciudad controlar de manera óptima el flujo vehicular en una hora establecida, o

esperar el momento más oportuno para realizar una compra o venta en el Mercado de Valores[54], el análisis de las series de tiempo evalúa patrones de causalidad que miden el efecto de las variables.

La regresión lineal se constituye en una técnica para realizar tareas de predicción tanto en procesos de modelización estadística como en las basadas en aprendizaje automático.

4.1 Regresión Lineal modelo estadístico

Un modelo de regresión es una función que permite establecer el comportamiento de una variable en función del conocimiento de otras, la variable resultado (y) a explicar el comportamiento, se la calcula en base a las variables de entrada o predictivas (x). En los problemas de regresión es suficiente que las variables predictivas tengan una relación con la variable de respuesta más el término de error

$$y_i = r(x_{il}, \dots, x_{ip}) + \varepsilon_i$$

Se dice que el valor esperado de ε_i es cero, cuando se trata de un valor centrado

$$E[\varepsilon_i]=E[y_i - r(x_{il}, \dots, x_{ip})]=0$$

Se habla de un modelo de regresión si todas las variables son de tipo cuantitativo, la función r se estima minimizando el promedio de la desviación al cuadrado respecto a los puntos.

El mínimo de la función corresponde a la media condicional de y_i, con relación al valor de las variables explicativas.

$$r(x_{i1}, \ldots\ldots, x_{ip}) = E[y_i | x_{i1}, \ldots\ldots, x_{ip}]$$

La función más simple de regresión es la lineal que se conoce como

$$E[y_i | x_{il}, \ldots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Las variables explicativas pueden ser transformaciones presentadas como funciones o como polinomios que ajustan a la función r no lineal, si los predictores son modificados por una función se convierte en una regresión lineal de la siguiente forma

$$y = c_0 + f_1(x_1) + \cdots + f_n(x_n)$$

Se supone que el termino de error ε tiene una influencia muy pequeña sobre la variable de respueta y estan poco relacionados entre si y sigue aproximadamente una distribución normal centrada

$$E[\varepsilon] = 0$$

Estan generados con la misma varianza para todos los individuos que se desprende de la función de probabilidad

$$var(\varepsilon_i) = \sigma^2$$

Los valores de error entre individuosson independientes, que bajo la condición de normalidad equivale a

$$cor(\varepsilon_i, \varepsilon_{i'}) = 0$$

El valor del error es independiente de los valores que tomen las variables explicativas

$$cor(\varepsilon_i, x_j) = 0$$

Desde el punto de vista estadístico la serie temporal se encuentra constituida por cuatro elementos tendencia, variación, estacionalidad y aleatoriedad, el estudio de estos cuatro elementos se conoce como análisis de series temporales

4.1.2 Tendencia de una serie

Indica el comportamiento de la serie en un lapso largo, permitiendo establecer su tendencia al alza, baja o estable, para extraer una tendencia es necesario conseguir minimizar el impacto de los demás componentes, lo que se consigue mediante el cálculo del valor de la media en un intervalo,

esta técnica permite disminuir los saltos bruscos entre x valores, utilizando la siguiente relación

$$y_i' = \frac{y_{\left(i-\frac{(x-1)}{2}\right)} + y_{\left(i-\frac{(x-1)}{2}+1\right)}^{+\cdots+y_{(i-1)}+y+y_{(i+1)}+} y_{\left(i-\frac{(x-1)}{2}-1\right)} + y_{\left(i-\frac{(x-1)}{2}\right)}}{x+1}$$

4.1.3 Variaciones Cíclicas de una serie

Identifica los ciclos que presenta una serie, que pueden ser de tipo periódico y se identifican con las fluctuaciones que presenta una serie sobre la línea de tendencia, la ciclicidad se obtiene dividiendo el valor original para el valor estimado, el valor igual a uno indica ausencia de ciclicidad

$$C = \frac{y}{y'}$$

Tanto la tendencia como la ciclicidad se las calcula en periodos mayores a un año

4.1.4 Estacionalidad de una serie

La característica estacional de una serie se mide gráficamente como un comportamiento repetido en k elementos, un movimiento estacional de una serie es una dependencia correlacional con un orden de nivel k entre los i-esimos y el (k+1)-esimo elemento de la serie.

Se conoce como retraso o lag el valor que tome k, así para analizar la autocorrelación entre valores adyacentes se toma k=1, para k=n se busca la autocorrelación entre los n elementos anteriores con el actual.

El valor de la autocorrelación para un k-lag se mide por:

$$r_k = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Donde

$$\bar{y} = \frac{\sum_{i}^{n-k}(x_{i+k})}{k}$$

y

$$\bar{x} = \frac{\sum_{i}^{n}(x_i)}{k}$$

4.1.5 Aleatoriedad o irregularidad en una serie

Identifica el comportamiento de la serie debido a factores imprevistos que afecta a una serie, existen herramientas de análisis de series como ARIMA, que es una metodología que realiza análisis de series de tiempo en el pasado para extrapolar su evolución a futuro.

El punto central que diferencia los problemas de series de tiempo de la mayoría de otros problemas estadísticos es que, en una serie de tiempo, las observaciones no son mutuamente independientes. Un solo evento ocurrido al

azar puede afectar a todos los puntos de datos posteriores. Esto hace que el análisis de series de tiempo sea muy diferente de otras áreas de la estadística.

4.2 Regresión lineal con estimadores núcleo (kernel)

La función de regresión encontrada en la modelización estadística puede ofrecer una buena aproximación en base a las medias móviles, pero siempre va a tener una discontinuidad especialmente en el extremo de los intervalos.

Para evitar este inconveniente se debe tomar un intervalo especifico de la variable explicativa centrado en un valor t y que tome valores cercanos a t, ponderando el peso de la observación (x_i,y_i) sea continua y tienda a cero.

La signación de pesos (x_i,y_i)se la realiza mediante la función núcleo (kernel) K, y la estimación r(t) es

$$w_i = w(t, x_i) = \frac{K(\frac{x_j - t}{h})}{\sum_{j=1}^{n} K(\frac{x_j - t}{h})'}$$

h controla la concentración del peso alrededor de t y se le conoce con el nombre de parámetro de suavizado, ahora volvemos al problema de minimos cuadrados ponderados

$$\min_{a,b} \sum_{i=1}^{n} w_i(y_i - (a + b(x_i - t))^2$$

La recta de regresión alrededor de t es

$$l_t(x) = a(t) + b(t)(x - t)$$

La regresión polinomca ponderada se consigue a partir de una regresión lineal donde la variable x es reemplazada por el polinomio $\beta_0 + \beta_1 x + \cdots + \beta_q x^q$, con q regresores

El estimador r(t), por polinomios locales se contruye de la siguiente manera

$$\min_{\beta_0,,\beta_q} \sum_{i=1}^{n} w_i(y_i - (\beta_0 + \beta_1(x_i - t) + \cdots + \beta_q(x_i - t)^q))^2$$

4.2.1 SVM (Support Vector Machine)

El análisis de series temporales se ha dado como resultado de la aplicación exitosa de los fundamentos de máquinas de vector soporte (Support Vector Machine SVM), [55] que son parte de los clasificadores lineales que inducen hiperplanos, que se generan por funciones llamadas kernel (núcleo), poseen un sesgo inductivo para maximizar el margen de separación entre distintas clases.

Un hiperplano D dimensiones $\Re$D se expresa como

h(x)=$\langle w.x\rangle + b$, w∈$\Re$D es un vector ortogonal al hiperplano, b es un número y $\langle -.-\rangle$ expresa el producto escalar

La regla para un clasificador binario es f(x)= signo(h(x))

Signo(x)=$\begin{cases} +1\ si\ x \geq 0 \\ -1\ si\ x < 0 \end{cases}$

x∈$\Re$D es un vector, a la combinación wx se le conoce como vector de pesos e indica su importancia y contribución en la clasificación, b es el umbral de decisión

Un SVM [56], es muy utilizado en problemas relacionados con clasificación y regresión, son algoritmos de aprendizaje supervisado que encuentra un tipo especial de modelo lineal: el hiperplano de máximo margen, y se utilizan para solucionar problemas de clasificación y predicción numérica. Se basan en un algoritmo que define un modelo lineal.

Para visualizar un hiperplano de máximo margen, se puede imaginar un conjunto de datos de dos clases △ y ○ (cuyas clases son linealmente separables, es decir, hay un hiperplano en el espacio instancia que clasifica todas las instancias de forma correcta.

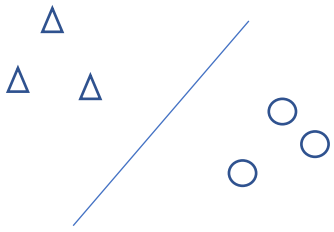


Fig. 4-1 Hiperplano

El hiperplano de máximo margen es el que da la mayor separación entre las clases. Técnicamente, la envolvente convexa de un conjunto de puntos es el más apretado polígono convexo que los encierra:

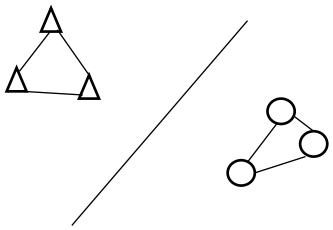
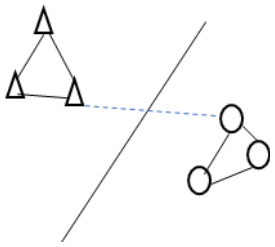


Fig. 4-2 Poligono convexo

Entre todos los hiperplanos que separan las clases, el hiperplano de máximo margen es el que está tan lejos como sea posible de ambos cascos convexos, representada como la línea mediatriz más corta que conecta los polígonos



$$\vec{w}.\vec{x} + b = 0$$

Fig. 4-3 Mediatriz conecta los pol[igonos

El hiperplano que guarda la misma distancia a las dos clases se encuentra de la siguiente manera

Maximizar $\frac{1}{\|w\|}$

Sujeto a: $y_i(\langle w, x_i \rangle + b) \geq 1 \quad 1 \leq i \leq N$

Función SVR (Support Vector Machine – Regression) [57], explica la manera de llegar a establecer un modelo lineal SVR de la forma:

$$y = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*).\langle x_i, x\rangle + b$$

que nos permite definir el modelo para la predicción de datos

Llamado Optimización mínima secuencial SMO para resolver problemas de regresión lineal usando SVM, es una extensión del algoritmo SVM, mejora la velocidad computacional y facilidad de implementación son algunas de sus características

4.3 Métricas de evaluación

Las medidas de evaluación[58], se utilizan para verificar el éxito de la predicción numérica, los valores obtenidos de la predicción se denominan p_n, los valores reales son denominados a_n

Para este caso en particular se manejan dos conjuntos de datos denominados entrenamiento (training) y prueba (test)

Los datos del conjunto entrenamiento son utilizados para realizar el aprendizaje y los datos del conjunto de test se utiliza para el cálculo de la tasa de error [59]

En este trabajo para evaluar el error de predicción en el conjunto de test en la serie de datos se utiliza el Error medio absoluto (Mean absolute error) y la raíz de error cuadrático medio (Root mean squared error).

Mean absolute error definido por la expresión $\frac{|p_1-a_1|+\cdots..+|p_n-a_n|}{n}$,

El Root mean squared error definido por la expresión $\sqrt{\frac{(p_1-a_1)^2+\cdots...+(p_n-a_n)^2}{n}}$,

4.4 Trabajos relacionados

En muchas aplicaciones, los patrones varían en el tiempo y, a veces son periódicos, el problema en este caso es descubrir los patrones a partir de los datos y tener en cuenta los tributos temporales para describir cómo varían en el tiempo. Existe bibliografía para descubrir patrones temporales en bases de datos de secuencias, generalmente tratadas como minería de secuencias (o búsqueda de patrones frecuentes) y asociación de datos temporales.

Los trabajos presentados a continuación están organizados en tres partes, la primera establece criterios para encontrar modelos estadisticos para realizar estimaciones a futuro utilizando un modelo estadístico, en la segunda parte, presenta trabajos que utilizan en conjunto algoritmos asociativos y temporales para mejorar o encontrar relación entre estos y finalmente se presenta trabajos que utilizan algoritmos de tipo kernel polinomial

En [60] utiliza un modelo estadístico para realizar un pronostico utilizando series de tiempo de la demanda hospitalaria desde una red de tensión media, aplicando el modelo ARIMA predice la demanda de carga eléctrica esperada.

En [61] se presenta una metodología que consta de tres pasos para predecir el precio de las acciones en un nuevo mercado, en la primera etapa se realiza un preprocesamiento preferencial para obtener el porcentaje de cambio en el precio de las acciones, a estas series de tiempo se realiza pruebas para determinar su estacionalidad, en la tercera etapa utiliza el método de series de tiempo para estimar la probabilidad del porcentaje de cambio a futuro

Trabajos como [62] busca revelar coherencias interesantes en una y entre pares de series de tiempo, aplicando este prototipo a los datos de medición del nivel del río.

En [63], extrae patrones frecuentes en la información de múltiples series de tiempo utilizando el algoritmo secuencial 'a priori', este enfoque secuencial ayuda en la poda de los patrones que no son frecuentes en las series de tiempo anteriores, ya que no serán frecuente para las próximas iteraciones.

En [64] proponen dos estrategias para demostrar que la asociación gobierna en una serie de tiempo.

[65] propone crear un modelo de previsión de ventas utilizando series temporales con datos reales , para apoyar la construcción de la estrategia de marketing y plan de cultivo, [66] establece un método mejorado de reglas de asociación cuantitativa de minería de datos para implementar el análisis

hidrológico que indica la factibilidad y viabilidad para utilizar las series temporales., [67] la identificación de los atributos predictores más importantes , y la extracción de un conjunto de reglas de asociación que pueden ser usados para predecir el comportamiento de series de tiempo en el futuro, [68] es un trabajo de predicción en base a un conjunto de reglas de asociación, [69], realiza una propuesta de análisis en base a una asociación de series de tiempo, [70] transforman las asociaciones extraídos de las bases de datos de series de tiempo en reglas de inferencia.

En [71] se encuentra una descripción general del funcionamiento de un SVM y sus posibles ineficiencias presentando un algoritmo que permita mejorar sus procesos de clasificación.

En lo referente a la utilización del algoritmo SMOreg se puede encontrar trabajos como [72], donde se presenta una comparación entre tres algoritmos para predicción uno de estos el Poly Kernel Regression.

Nuevos trabajos como [73], donde se calcula el IDH (Índice de desarrollo humano), utilizando SMOReg por ser considerado el de mejor rendimiento en el aprendizaje.

Investigaciones como la realizada por [74] se puede apreciar la evaluación del rendimiento de SMOreg en el pronóstico de carga para el consumo de electricidad.

Este capitulo ha presentado el desarrollo teoríco para realizar estimaciones a futuro utilizando series de tiempo con dos enfoques igualmente válidos, el primero un modelamiento estadístico y la segunda forma utilizando estimadores núcleo y ajuste local de polinomios.

En la parte de trabajos relacionados además de explorar temas que utilizan estos dos enfoques se presenta una serie de trabajos que utilizan un desarrollo asociativo para mejorar la parte temporal y viceversa, lo que permite establecer un aspecto directamente relacionado con la metodología propuesta de partir de un modelo asociativo para mejorar los niveles de error en estimaciones a futuro utilizando series de tiempo

5. Datos espaciales y Fusión de modelos de datos

Este capitulo trata el desarrollo teórico que permite establecer procesos predictivos en base a la ubicación geográfica, partiendo de los postulados realizados por el geografo Waldo Tobler conocidas como la primera y segunda ley de geográfia [75].

Cualquier área por pequeña que sea en la superficie de la Tierra, tiene todas las condiciones para describir su comportamiento, siendo este el principio utilizado para los métodos de minería de datos espaciales, desde modelos que únicamente utilizan la distancia "física", (IDW Inverse Distance Weighting IDW) o los que aplican modelos de variables regionalizadas para aplicar kriging [76].

El proceso predictivo espacial en este trabajo, inicia utilizando la característica espacial de cercanía de puntos con información (IDW), el siguiente nivel explota el concepto de variable regionalizada y continuidad de la función generando por el variograma experimental y ajustando al variograma teórico con lo que se utiliza procesos de Kriging (predicción).

Tambien se presenta el desarrollo de predicciones en base a la utilización de multiples variables y la estimación de valores por medio del Co-kriging, que puede verse como un caso particular de fusión de información georeferenciada.

A partir de aquí, se presenta la integración de fuentes de información complementarias como un proceso de fusión de información.

5.1 Distancia Inversa Ponderada

Esta es una forma bastante sencilla computacionalmente hablando de estimar el valor de un punto desconocido en base a las características que presentan otros puntos espacialmente distribuidos [77], este método denominado Distancia Inversa Ponderada (IDW), asigna a cada dato una ponderación inversamente proporcional a la distancia.

Este proceso establece valores en puntos desconocidos apoyándose en valores de puntos conocidos en base a una combinación lineal considerando a la distancia para realizar la ponderación [78],[79]. Con esta consideración se asume que los valores más cercanos a la muestra desconocida son más influyentes por lo que le otorga un mayor peso

$$d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

Donde (x,y) son coordenadas Cartesianas conocidas, (x_i, y_i) puntos de interpolación

La interpolación en base a distancia y pesos se realiza utilizando la siguiente relación:

$$f(x,y) = \frac{\sum_{i=1}^{n} d_i^{-d_{exp}} P_i}{\sum_{i=1}^{n} d_i^{-d_{exp}}}$$

Donde n son los puntos conocidos, P_i, es el valor medido en un punto conocido y d_{exp} es el exponente de la distancia, generalmente igual a 2.

Si la potencia es cercana a cero, la distancia elevada a esa potencia es cercana a uno, con lo que se asigna una misma ponderación a todos los datos.

5.2 Estimación Espacial

Como se menciona en la primera ley de la geografía [80] "en el espacio geográfico todo se encuentra relacionado con todo, pero los espacios más cercanos están más relacionados entre sí". Desde este principio es difícil considerar los datos medidos como variables aleatorias independientes

A la medida de las variables numéricas a ser estudiadas en este espacio geográfico se denomina variables regionalizadas. Se conoce como campo de la variable al dominio limitado D donde se realiza el estudio de la variable regionalizada.

La geoestadística [81] utiliza el concepto de función aleatoria para presentar los valores no determinísticos distribuidos sobre una región D. Si x recorre la región, se obtiene una serie de variables aleatorias, Z = {Z(x), x ∈ D}, que

constituyen una función aleatoria, y presentan una correlación entre ellas que refleja la continuidad en la región geográfica.

Si consideramos que una función aleatoria generada en $\{x_1, \ldots., x_k\}$ sitios, entrega el conjunto de variables aleatorias $\{Z(x_1), \ldots., Z(x_k)\}$ en D, que posee una característica generada por una función de probabilidad multivariable.

$$F_{x1,\ldots,xk} = Prob\{Z(x_1) < z_1, \ldots., Z(x_k) < z_k\}\ \forall z_1, \ldots z_k \in \mathbb{R}$$

Al conjunto de las funciones de distribución para los enteros k y todas las elecciones posibles de x_k en D se le denomina distribución espacial, en base a los datos disponibles se infiere las distribuciones en las que se basa un modelo de distribución espacial.

Si k=1 hablamos de una distribución univariable de la siguiente manera

$$F_{x_1}(z_1) = Prob\{Z(x_1) < z_1\}$$

Que corresponde con la función de distribución de $Z(x_1)$ ubicada en las coordenadas x_1

Si k=2 se refiere a una distribución bivariable

$$F_{x_1,x_2}(z_1, z_2) = Prob\{Z(x_1) < z_1, Z(x_2) < z_2\}\}$$

Corresponde a la distribución conjunta de las variables $\{Z(x_1), Z(x_2)\}$ ubicadas en x_1, x_2

Al considerar la función aleatoria en estos dos casos particulares se simplifica algunos parametros descriptivos o momentos de las distribuciones univariables y bivariables que permiten conocer la información de mayor importancia y se presentan de la siguiente manera

- Esperanza o momento de primer orden, entrega el valor esperado

  M(x)=E[Z(x)], $E[Z(x)] = \frac{\sum_i Z(x_i)}{n}$

  representa la media (esperanza) alrededor de la cual se distribuyen los valores tomados por la función aleatoria.

- Varianza o momento, cuantifica el carácter aleatorio de una función, está definido por:

  σ^2(x)= var[Z(x)]

  =E{[Z(x)-m(x)]^2}

  =E[Z(x)^2]-m(x)^2,

la varianza y su raíz cuadrada llamada desviación estándar constituyen medidas de dispersión de Z(x) en torno a su valor medio,

- la covarianza centrada entre dos variables aleatorias está dada por la relación:

$$C(x_1, x_2) = E[Z(x_1)Z(x_2)] - m(x_1)m(x_2)$$

y nos entrega una vision elemental de la interacción que existe entre $Z(x_1)\ y\ Z(x_2)$,

- el semi variograma entre dos variables aleatorias $Z(x_1)$ y $Z(x_2)$, viene dado por la expresión:

$$\gamma(x_1, x_2) = \frac{1}{2} var[Z(x_1) - Z(x_2)]$$

y refleja la forma en que un punto tiene influencia sobre otro punto a diferentes distancias.

5.2.1 Inferencia estadística

A partir de los datos es necesario determinar la distribución espacial de la función aleatoria de la variable regionalizada, que solo se conoce de manera fragmentada en los sitios de muestreo.

Para esto se recurre a la noción de estacionaridad, suponiendo que los valores que se encuentran en distintas secciones del campo presentan iguales características pudiendo considerarse como realizaciones distintas de un mismo proceso.

La hipótesis de estacionaridad indica que la distribución espacial de una función aleatoria es invariante por traslación, que indica las propiedades del conjunto de datos dependen de sus posiciones relativas, que permite realizar las simplificaciones presentadas a continuación.

Los momentos de orden uno (esperanza y varianza) que representan a la distribución univariable no depende de la ubicación geográfica considerada

$$F(z_1) = Prob\{Z(x_1) < z_1\}$$

Y, por tanto, son constantes en el espacio.

$$m = E[Z(x)]$$

$$\sigma^2 = var[Z(x)]$$

La función de distribución bivariable depende solamente de la separación entre los sitios determinados

$$F_h(z_1, z_2) = Prob\{Z(x+h) < z_1, Z(x) < z_2\}$$

los momentos de orden dos como la covarianza(C) y variograma (γ) dependen de la separación de los sitios geográficos considerados

$$C(h) = cov[Z(x+h), Z(x)]$$

$$\gamma(h) = \frac{1}{2} var[Z(x+h) - Z(x)]$$

Siendo el variograma una síntesis de la continuidad espacial, indica que tan desemejantes son los valores de la variable entre dos sitios.

Suponiendo que la esperanza no depende de la localización x, sino que depende únicamente de la separación se tiene:

$$\gamma(h) = \frac{1}{2} E\{[Z(x+h) - Z(x)]^2\}$$

5.2.2 Variograma Experimental.

Los momentos de orden dos permiten conocer una descripción de la continuidad espacial de una variable regionalizada entre dos puntos.

Si consideramos la variable regionalizada z conocida en n sitios, $\{x_1, \ldots, x_n\}$, el estimador para un vector de separación h, queda definido de la siguiente manera

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} [z(x_\propto) - z(x_\beta)]^2$$

$N(h) = \{(\propto, \beta) tq. x_\propto - x_\beta = h\}$;

$|N(h)|$ número de pares contenidos en N(h)

El estimador así definido toma por nombre variograma experimental y se interpreta como la medición de la distancia promedio entre parejas, se parte del criterio que a menor distancia existe una mayor correlación espacial

Ejemplo: Dado el conjunto de datos espaciales separados cada 50 metros

4	3	7	4	6	1
1	2	5	5	1	2
6	1	4	1	6	1
3	1	1	7	3	1
1	2	5	3	6	1
6	1	2	7	4	3

El variograma experimental para este caso se construye de la siguiente manera

$$\hat{\gamma}(50m) = \frac{1}{2x5}((4-3)^2 + (3-7)^2 + (7-4)^2 + (4-6)^2 + (6-1)^2)$$

= 5,5

Variograma Izquierda-Derecha			
$\hat{\gamma}(50)$	$\hat{\gamma}(100)$	$\hat{\gamma}(150)$	$\hat{\gamma}(200)$
5,5	2,5	7,5	2
2,7	6,25	4,33333333	0
9,3	1	9,83333333	0
6	10	3,33333333	0
4,8	2,75	6	6,5
6,1	9	1,83333333	2

Variograma Superior-Inferior			
$\hat{\gamma}(50)$	$\hat{\gamma}(100)$	$\hat{\gamma}(150)$	$\hat{\gamma}(200)$
7,2	5,25	0,16666667	8,5
0,4	0,75	0,66666667	0,5
3,9	3,375	6,66666667	3,25
8,5	2,125	8,16666667	1,25
7,2	0,625	6,33333333	2,25
0,6	0,625	0,83333333	0,25

Para solucionar el problema de predicción espacial es necesario conocer la autocorrelación para cualquier distancia posible en el área de estudio,

por lo que se hace necesario el ajuste de modelos que generalicen a cualquier distancia Fig 5-1.

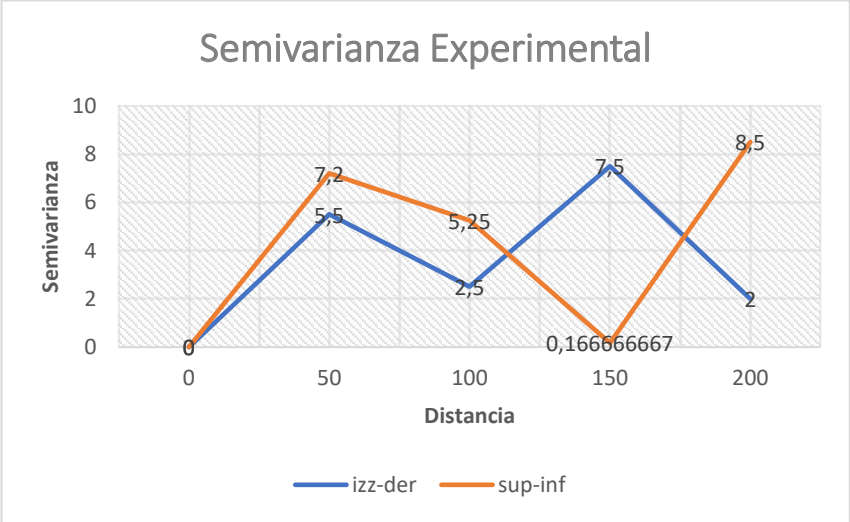


Fig. 5-1 Variograma dos direcciones

El variograma experimental es un estimador insesgado del variograma teórico

$$E[\hat{\gamma}(h)] = \gamma(h)$$

5.2.3 Variograma Modelado.

Un variograma experimental no puede utilizarse directamente para la estimación espacial, porque está definido solo para ciertas distancias y direcciones contenidas en el conjunto de datos disponible. Por tanto, para interpretar la continuidad espacial de la variable de estudio, se debe ajustar a

un modelo teórico que cumpla con ciertas propiedades como son: matemáticas, comportamiento en el origen y comportamiento para distancias grandes.

5.2.3.1 Propiedades matemáticas

La función variograma debe cumplir con

1.- $\gamma(h) = \gamma(-h)$

2.- $\gamma(0) = 0$

3.- $\gamma(h) \geq 0$

4.- $\lim_{|h|\to+\infty} \frac{\gamma(h)}{|h|^2} = 0$

5.- Negativo Condicional $\forall k \in \mathbb{N}^*, \forall \lambda_1, \ldots., \lambda_k \in \mathbb{R}\ tales\ que\ \sum_{i=1}^{k} \lambda_i = 0, \forall x_1, \ldots, x_k \in D,$

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i\, \lambda_j \gamma(x_i - x_j) \leq 0$$

5.2.3.2 Comportamiento en el origen

Dependiendo de su comportamiento en el origen se puede intuir el comportamiento de la variable regionalizada, se distingue tres tipos de comportamientos:

1.- Variograma parabólico, caracteriza a una variable regionalizada regular en el espacio

2.- Variograma lineal, corresponde a una variable regionalizada continua no regular

3.- Variograma discontinuo, que está asociado a un comportamiento errático de la variable regionalizada a estos cambios repentinos se les denomina efecto pepita cuando las variaciones son muy pequeñas y no se aprecia fácilmente su continuidad.

5.2.3.3 Comportamiento para grandes distancias:

Por lo general el variograma crece desde el origen y se estabiliza en una distancia a, donde comienza una meseta, de manera que las dos variables aleatorias Z(x) y Z(x+h) estarán correlacionadas si la longitud del vector de separación h es inferior a la distancia a, denominada alcance o zona de influencia. Más allá de |h|=a, el variograma es constante e igual a su meseta.

$$\gamma(h) = \quad C(0) - C(h)$$

$$\downarrow_{|h|\to\infty} \qquad \downarrow_{|h|\to\infty}$$

$$\gamma(\infty) = C(0) = \sigma^2$$

En lo referente a su comportamiento direccional, un variograma, $\Upsilon(h)$ es isótropo si tiene el mismo comportamiento en todas las direcciones del espacio, si no depende de la orientación del vector h sino de su modulo |h| caso

contrario hay anisotropía, que caracteriza a un fenómeno que se extiende de preferencia en una dirección.

5.2.3.4 Modelos de Variogramas

Para que γ sea el variograma de una función aleatoria debe cumplir con la propiedad de tipo negativo condicional, al ser esta propiedad difícil de conseguir se recurre a funciones que cumplen con esta restricción como son entre otras:

- Un modelo esférico de alcance a y meseta C se define como

$$\gamma(h) = \begin{cases} C\left\{\frac{3}{2}\frac{|\mathrm{h}|}{a} - \frac{1}{2}\left(\frac{|\mathrm{h}|}{a}\right)^3\right\} \; si \; |\mathrm{h}| \leq \mathrm{a} \\ C \; en \; caso \; contrario \end{cases}$$

- Un modelo de tipo exponencial con parámetro a y meseta C se define:

$$\gamma(h) = C(1 - \exp\left(-\frac{|h|}{a}\right))$$

El variograma experimental mide la desemejanza promedio entre dos datos en función de su separación, a menudo presenta cambios de pendiente, que indican un cambio en la continuidad espacial a partir de ciertas distancias, el variograma puede modelarse como la suma de varios modelos elementales denominados modelos anidados o estructuras anidadas $\gamma(h) = \gamma_1(h) + \gamma_2(h) + \cdots + \gamma_s(h)$

El ajuste a un modelo no se hace considerando solamente el variograma experimental, sino que se debe tomar en cuenta toda la información disponible sobre la variable regionalizada.[82]

5.3 Estimación con Kriging ordinario (media desconocida)

El método de kriging es considerado como una predicción lineal, con un estimador lineal insesgado.

Se desea predecir el valor de una variable en el punto x_0 que no tiene mediciones, para esto el kriging ordinario propone que puede predecirse como una combinación lineal

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

$$\lambda_i, pesos\ de\ valores\ originales$$

Debe cumplir con las restricciones indicadas en las siguientes secciones

5.3.1 Restriccion Lineal.

El estimador debe ser una combinación lineal ponderada de los datos:

$$\boldsymbol{Z^*(x_0)}\textbf{=a+}\ \boldsymbol{\sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha)}$$

$\boldsymbol{x_0}$ es el sitio donde se establece una estimación, {x_α, α=1,…,n}, son los sitios con datos conocidos, $\{\boldsymbol{\lambda_\alpha}, \alpha = 1, \dots, n\}$ son los ponderadores que en conjunto con '**a**' son las incognitas

5.3.2 Restricción de Insesgado. - Se expresa que el error de estimación tiene esperanza nula

$$E[Z^*(x_0) - Z(x_0)] = 0$$

$$= a + \sum_{\alpha=1}^{n} \lambda_\alpha \, E[Z(x_\alpha)] - E[Z(x_0)]$$

$$= a + m\left(\sum_{\alpha=1}^{n} \lambda_\alpha - 1\right)$$

5.3.3 Restricción de optimalidad. - Busca ponderadores que minimizan la varianza del error de estimación

$$var[Z^*(x_0) - Z(x_0)] \; es \; mínima$$

$$= \sum\nolimits_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \, \lambda_\beta \, C(x_\alpha - x_\beta) + C(0) - 2 \sum\nolimits_{\alpha=1}^{n} \lambda_\alpha \, C(x_\alpha - x_\beta)$$

Siendo el variograma una herramienta equivalente a la covarianza a partir de la relación

$$\gamma(h) = C(0) - C(h)$$

El calculo del kriging se realiza de la siguiente manera

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta \gamma(x_\alpha - x_\beta) - \mu = \gamma(x_\alpha - x_0) \ \ \forall \alpha = 1 \dots n \\ \sum_{\alpha=1}^{n} \lambda_\alpha = 1 \end{cases}$$

$$\begin{pmatrix} \gamma(x_1 - x_1) & \cdots \gamma(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots \\ \gamma(x_n - x_1) & \dots \gamma(x_n - x_n) & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix}$$

La varianza del error cometido en el punto x_0, se expresa así:

$$\sigma_{KO}^2(x_0) = \sigma^2 - \sum_{\alpha=1}^{n} \lambda_\alpha C(x_\alpha - x_0) - \mu$$

Por lo cual el kriging ordinario sigue aplicable aun cuando el variograma no presenta meseta

Aplicación de Kriging

Para ilustrar el calculo del Kriging, supongase que se solicita encontrar el valor del punto x_0, que se encuentra separado de los puntos S_1,..,S_4, como se indica Fig 5-2.

El variograma modelo es una función de correlación exponencial $\gamma(h) = 10(1 - \exp\left(-\frac{3h}{10}\right)))$, pepita ($C_0$) = 0, meseta ($C_0$+$C_1$) y rango (a) = 10. La distancia entre todos los puntos se detalla en la tab 5-1

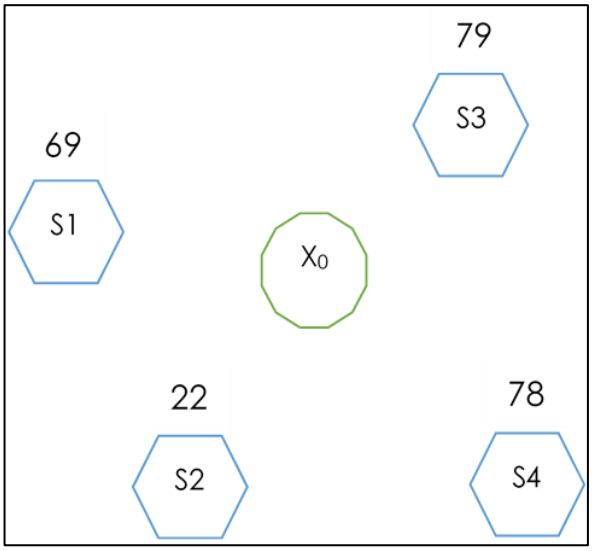


Fig. 5-2 Ilustración Kriging ordinario

Sitio	S1	S2	S3	S4	X_0
S1	0	11.05	10.5	16.97	3.61
S2		0	15	11	8.06
S3			0	13.15	8.94
S4				0	13.45

Tab. 5-1 Matriz de distancia entre puntos

La matriz C_{ij}=C(0)-C(h) es la siguiente

10	0,3633	0,4285	0,06	1
0,3633	10	0,111	0,3688	1
0,4285	0,111	10	0,1935	1
0,06	0,3688	0,1935	10	1
1	1	1	1	0

Tab. 5-2 matriz Cij=C(0)-C(h)

C_{ij}^{-1} es la siguiente

0,07749692	-0,0264133	-0,0273451	-0,0237386	0,2477684
-0,0264133	0,07743204	-0,0239953	-0,0270234	0,24790528
-0,0273451	-0,0239953	0,07685514	-0,0255148	0,2507665
-0,0237386	-0,0270234	-0,0255148	0,07627679	0,25355982
0,2477684	0,24790528	0,2507665	0,25355982	-2,690415

Tab. 5-3 (C_ij)$^{-1}$

λ= $C_{ij}^{-1}.C_{io}$

C_{io}
3,38
0,8909
0,6842
0,17
1

0,46343135	λ_1
0,20660098	λ_2
0,18520957	λ_3
0,1447581	λ_4
-1,4174194	μ

$Z_0^* = \sum_{i=1}^{4} \lambda_i Z_i$ =(0,4634)(69)+(0,2066)(22)+(0,1852)(79)+(0,1447)(78)=62,4372

La varianza de error se calcula de la siguiente manera

$\sigma_k^2 = \sigma^2 - \sum_{i=1}^{4} \lambda_i c_{io} - \mu =$

10-[(0,4634*3,38)+(0,2026*0,8909)+(0,1852*0,6842)+(0,1447*0,17)]+1,417= 9,5

El valor estimado para el punto X_0 es 62,43, utilizando la función de correlación definida , la distancia euclidea entre puntos utilizando el proceso de kriging ordinario, la varianza de error es 9,5.

5.4 Co-kriging

Es la versión multivariable del kriging que busca estimar el valor de una variable tomando en cuenta el valor de otras variables correlacionadas en el mismo punto, requiriendo los modelos variograficos de cada variable, así la principal herramienta para estimar semivarianzas entre diferentes variables es el variograma cruzado definido como:

$$\gamma_{ij}(h) = \frac{1}{2}\mathrm{E}[(Z_i(s) - Z_i(s+h))(Z_j(s) - Z_j(s+h))]$$

Múltiples variables pueden tener correlación cruzada, lo que significa que la variabilidad espacial de la variable A esta correlacionada con la variable B y viceversa. Las medidas de las dos variables son tomadas en un conjunto limitado de localidades y la interpolación se realiza a un número ilimitado de localidades.

El variograma cruzado entre dos variables (Z_1, Z_2) se define de la siguiente forma:

$$\gamma_{12}(h) = \frac{1}{2} cov\{Z_1(x+h) - Z_1(x), Z_2(x+h) - Z_2(x)\}$$

y se puede conseguir a partir de los datos disponibles:

$$\hat{\gamma}_{12}(h) = \frac{1}{2|\mathrm{N(h)}|} \sum_{N(h)} [Z_1(x_\alpha) - Z_1(x_\beta)] [Z_2(x_\alpha) - Z_2(x_\beta)]$$

Donde N(h)={α, β tal que $x_\alpha - x_\beta = h$}

$siendo\ ambas\ variables\ Z_{1\,y}\ Z_2\ medidas\ en\ x_\alpha\ y\ x_\beta$ }

5.5 Validación Cruzada (Cross Validations).

La validación cruzada permite medir la diferencia entre el valor estimado frente al valor verdadero, para conseguir este propósito considera estimar sucesivamente mediante kriging cada dato considerando los datos restantes

En este trabajo se utiliza la validación cruzada dejando uno fuera, Leave-one-out cross-validation, por ser uno de los procesos que presenta el mas bajo error, se realiza tantas interaciones como datos (N) tenga el conjunto, el resultado final es la media aritmetica de los N resultados de error obtenidos

$E = \frac{1}{N} \sum_{i=1}^{n} e_i$

La validación cruzada por lo general proporciona una estimación pesimista del rendimiento (sesgo), ya que la mayoría de los modelos mejorarían si el

conjunto de entrenamiento fuera más grande. Por esta razón, LOOCV tiene el sesgo más bajo, ya que el conjunto de entrenamiento contiene todo el conjunto de datos, excepto un dato. Por otro lado, algunos autores señalan que el error estimado por LOOCV puede tener mayor varianza que la validación cruzada de k-fold, con k << n, ya que el tamaño de los conjuntos de datos es mayor y la estimación es más suave. Sin embargo, esto está abierto a discusión, como se indica [83], ya que la validación cruzada K-fold produce errores de prueba dependientes, y sus correlaciones no pueden estimarse de manera imparcial. Como se indica, en [84], en problemas de aprendizaje que emplean modelos con inestabilidad moderada / baja (como problemas de regresión lineal), la LOOCV a menudo tiene una variabilidad baja tanto en el sesgo como en la varianza.

En cualquier caso, para situaciones con conjuntos de datos pequeños, la varianza en el ajuste del modelo tiende a ser mayor, lo que implica que es probable que la validación cruzada de K-k tenga una varianza alta (así como un sesgo mayor) con respecto a LOOCV. Esta es la razón por la que LOOCV suele ser la mejor opción con cantidades limitadas de datos disponibles, como el estudio de caso en este trabajo, para obtener el máximo uso de los datos para comparar el rendimiento de estructuras de aprendizaje alternativas.

5.6 Fusión de modelo de datos

También se puede realizar la fusión o combinación de distintos algoritmos de aprendizaje [85] o la combinación de distintas hipótesis, la calidad del modelo combinado depende de la precisión y diversidad de los componentes del conjunto.

Una nueva técnica de combinación de modelos se basa en la composición de partes obtenidas de técnicas de aprendizaje básicas, este nuevo modelo aglutina las características de los modelos participantes[86] otra manera de conseguir un nuevo modelo es alterando el conjunto de atributos de entrada.

En este contexto se tiene la fusión de modelos como la integración de distintas fuentes de información complementarias, externas a un proceso especifico como por ejemplo datos meteorológicos o de población.

Esta información se la puede obtener de datos que se alimentan directamente como son los sensores de hardware o software.

Como se describe [87], [88] los sensores se definen como fuentes de información del entorno. En un entorno real, los sensores están compuestos por hardware y software específicos para tomar medidas desde una variable física (como ubicación, tamaño, temperatura, presión, etc.) [89]. En

entornos virtuales como redes sociales, webs, etc., los sensores son herramientas de software específicas que extraen información de una fuente digital. [90] Estos dos tipos diferentes de sensores podrían denominarse "sensores de hardware" y "sensores de software", etiquetados como "h", "s", y ambos tipos de sensores generan información sobre el entorno, pero esta definición permite distinguir entre un ambiente real y otro digital. [91]

Un sensor se puede definir de la siguiente manera:

S_i^t: Sensor i de tipo t, t∈ $\{h, s\}$.

Además, varios sensores de diversos tipos podrían colocarse juntos en una plataforma de sensores para aprovechar la medición de algunas variables de forma coordinada. Por ejemplo, una estación meteorológica está compuesta por varios sensores de hardware (temperatura, presión atmosférica, altitud, etc.) y puede incluir sensores de software (observaciones humanas y reportes de pronóstico) o un sistema de vigilancia podría estar compuesto por un conjunto de sensores duros (radar) , cámara y cámara infrarroja alineadas para detectar cualquier objeto en el campo de búsqueda) y pueden incluir entradas humanas de operadores humanos.

Una plataforma de sensores se define de la siguiente manera

$P_j = \{S_{ij}^t\}_{i=1}^n$: Plataforma j compuesta por n sensores de diferentes tipos.

En el general, las plataformas se deben colocar distribuidas en el entorno para cubrir un área grande y, en muchos casos, las áreas de cobertura se superponen para mejorar las detecciones en los límites, como se muestra en la Fig. 5-3 Una red desplegada que incluye sensores duros y blandos. La información capturada por las plataformas se envía a un centro de fusión donde se integra utilizando la información espacial y temporal.

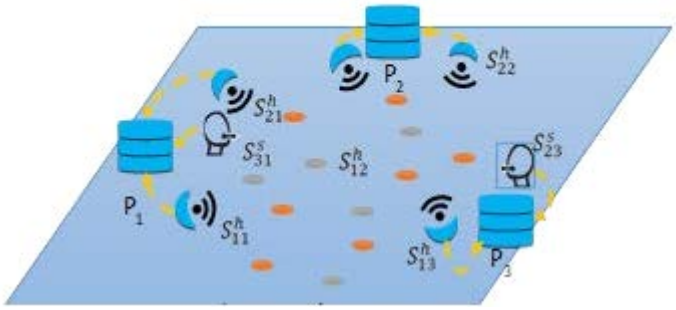


Fig. 5-3 Una red desplegada que incluye sensores duros y blandos

La información fusionada se utiliza para predecir situaciones futuras. Esta predicción depende de un modelo que podría basarse en medidas de una sola fuente, por ejemplo, las posiciones actuales y pasadas de un objetivo para predecir posiciones futuras basadas en medidas y el modelo de movimiento. En algunas situaciones, la información del conjunto de sensores de la plataforma podría integrarse y la información complementaria ser parte de un nuevo modelo de datos.

5.7 Trabajos relacionados

Las investigaciones sobre temas con componente espacial se encuentran en variados campos, entre lo más destacado podemos comentar los siguientes:

En [92] se realiza una explicación del desarrollo matemático del Kriging y del Cokriging basado en modelos de sustitución dentro del marco de optimización, [93] plantean mejorar la construcción del variograma utilizando información de magnitud y dirección aplicado a datos de la Red Nacional de los Observatorios Geomagnéticos de China.

En [94] se utiliza cuatro distintos tipos de interpolación entre los que se menciona las técnicas de IDW, Kriging ordinario y cokriging ordinario que permite confirmar la valides del estudio planteado en este trabajo, en la misma línea investigativa se presenta [95] donde se llega a la conclusión que el cokriging es superior al pronóstico IDW.

En [96] trata la misma temática del manejo de distintos procesos de interpolación aplicado a la erosión pluvial.

En lo referente a fusión de datos existen varios trabajos que combinan el factor temperatura con la predicción utilizando técnicas de IDW[97], Krigin [98]y cokriging como se menciona en, [99], [100]

5.8 Visión general del estado del arte.

Se ha presentado dos tipos de procesos para realizar el minado de datos: descriptivas y predictivas.

En la parte descriptiva, utilizando inferencia de reglas de asociación se establece la relación basado en un atributo en común, como puede ser el proceso de comercialización de productos agricolas.

En lla parte predictiva se ha presentado la utilización de dos dimensiones para el proceso de búsqueda de patrones el primero el tiempo y el segundo la ubicación espacial.

Para la utilización de las dos dimensiones se describe la utilización de regresión lineal, al utilizar series de tiempo se encuentra estimaciones de comportamiento a futuro utilizando estimadores núcleos que asignan un peso a todas las variables de su entorno creando un modelo de datos para estimar el comportamiento futuro basado en datos presentes.

En la parte espacial, de la misma manera se utiliza un estimador de tipo lineal, para este caso los pesos de los componentes regionalizados permiten encontrar un modelo de datos que indica valores en lugares donde no existen mediciones.

Tambien se han encontrado procesos donde la temporalidad y la ubicación geográfica son utilizados como elementos para mejorar la producción de

reglas de asociación, se parte de una predicción realizada desde un proceso correlacional y se aplica a un proceso asociativo para mejorar la producción de reglas de asociación.

Esta investigación parte de criterios de asociatividad generados por el agricultor familiar campesino y este resultado se aplica para mejorar los procesos predictivos correlacionales como son las series temporales y los geoestadísticos.

Parte III.- Propuesta, desarrollo y caso de estudio

La información de comercialización de productos registrada en el año 2014 para los circuitos alternativos de comercialización (CIALCOs) de las provincias de Tungurahua y Chimborazo permite aplicar mineria de datos de tipo descriptivo y predictivo, utilizando varios procesos como son la asociatividad, la temporalidad y datos espaciales.

Por un lado, realizando la inferencia de reglas de asociación se obtiene un conjunto de productos que tienen una mayor relación en la comercialización, estableciendo correlaciones de tipo descriptivo.

Utilizando la dimensión de temporalidad se puede predecir temporadas de mayor comercialización de un producto y establecer programas de cultivo para optimizar la producción de granjas agrícolas.

La dimensión espacial interpola niveles de producción en un lugar geografico desconocido basado en los registros de comercialización de productos ofertados en ferias cercanas.

Cada dimensión escogida genera escenarios que pueden evaluar el impacto de la implementación de políticas orientadas a fortalecer el sector agrícola.

De acuerdo con el proceso de aprendizaje aplicado, a la información, en un caso debe ser discretizada, en otro utiliza la variable de temporalidad, o en base a la ubicación geográfica, para estos tres distintos tipos de aprendizaje se propone una metodología de tratamiento, que establece una mejora en los procesos de estimación pronóstica.

Metodología Propuesta

Esta investigación realiza la fusión de dos modelos de aprendizaje, el primero de tipo descriptivo, utiliza el descubrimiento de patrones en base a la inferencia de reglas de asociación y un segundo modelo de aprendizaje de tipo predictivo qué basado en el resultado del modelo asociativo, genera predicciones multivariable utilizando métodos lineales de regresión, el desarrollo de la metodología se la resume en los siguientes puntos:

1. Generar el conjunto inicial de productos para el estudio.
2. Preparar formato de datos apropiado a cada proceso de minado.
3. Establecer un conjunto de productos asociados (atributos nominales).
4. Generar estimaciones a futuro utilizando atributos correlacionales para un predictor

5. Medir el impacto resultante en la utilización de técnicas de predicción en conjuntos correlacionales fusionados con resultado de técnicas asociativas (multivariable).

La Fig. III-1 Metodología Propuesta, describe gráficamente la metodología propuesta: de izquiera a derecha comienza por realizar los procesos para preparar información y adecuarla a los formatos que requiere cada algoritmo de aprendizaje aplicados. Los exagonos siguientes representan las técnicas de minería de datos aplicadas: descriptivas y predictivas.

Utilizando técnicas de tipo descrptiva, se generan reglas de asociación y se constituye el conjunto multivariable.

Al aplicar técnicas predictivas se utiliza la dimensión temporal y la dimensión espacial para obtener la estimación pronóstica utilizando algoritmos de:

1. Series de tiempo
2. Geoestadística.

A cada uno de los modelos de datos de aprendizaje correlacionales obtenidos, se fusiona el conjunto resultante de la inferencia de reglas de asociación, al conjunto multivariable se le aplica nuevamente los procesos de mineria de datos predictiva.

Se evalua los resultados de estimación pronóstica obtenidos utilizando un solo predictor, con los resultados utilizando el conjunto asociativo como predictor.

Finalmente utilizando la segunda ley de la geografía que indica: cualquier actividad externa al área de interés afecta lo que sucede al interior, se recolecta información de población y piso climático.

Estas dos variables externas al proceso de comercialización de circutos alternativos de comercialización se utilizan para generar dos nuevos modelos, a los que se aplica la metodología propuesta y se compara sus resultados.

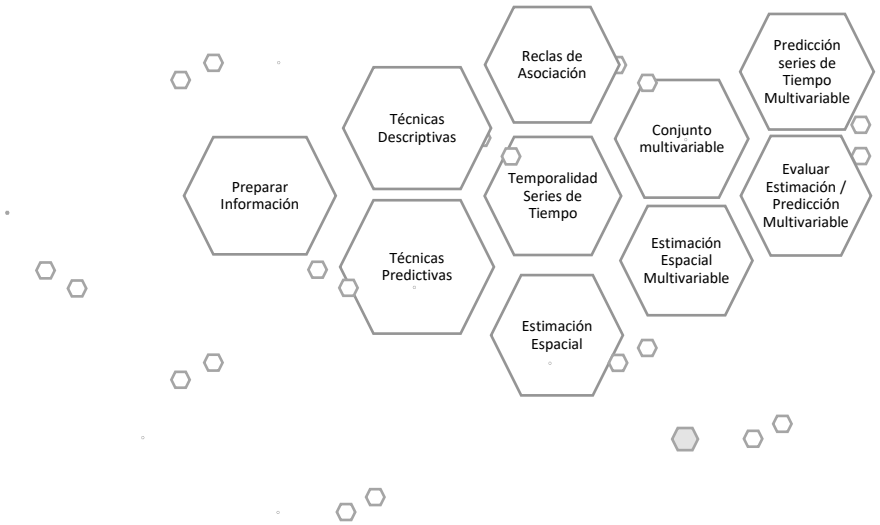


Fig. III-1 Metodología Propuesta

6. Conjunto de datos

Este capítulo presenta los procesos y tareas utilizadas para la generación del conjunto inicial de productos y la transformación a estructuras de datos útiles para la aplicación de algoritmos de minería de datos.

Para conseguir el conjunto inicial de estudio se realiza la evaluación de los datos registrados por el Ministerio de Agricultura y Ganadería del Ecuador, verificando la calidad de la información, poniendo enfasís en la estandarización de nombres de productos y unidades de medida.

Posteriormente se registran los procesos de transformación que agregan características de discretización, temporalidad y ubicación geográfica en tres tipos de estructuras independientes.

Finalmente se realiza la creación de un nuevo modelo de predicción basado en dos variables externas: población y piso climático.

6.1 Fuente de datos original

En el proyecto CIALCO, una feria consta de varios productores que se asocian para realizar actividades de comercialización sobre ítems que producen o que elaboran.

Los datos de la comercialización del año 2014 han sido registrados en forma mensual, el formato original del registro se presenta en Fig. 6-1 Datos originales. La información consta de nombre de la organización, sector geográfico de funcionamiento, nombre del productor, identificación del producto comercializado, volumen de productos vendidos, valor unitario, ingreso por ventas.

FERIA CIUDADANA PUYO

FECHA: 06 DE JULIO DEL 2014

HORTALIZAS Y LEGUMBRES

ORDEN	NOMBRE DE LA ORGANIZACIÓN/ ASOCIACION	SECTOR	NOMBRES COMPLETOS	PRODUCTOS	VOLUMEN DE PRODUCTOS VENDIDOS	VALOR UNITARIO	INGRESO POR VENTAS
1	FERIAS CIUDADANAS	CHAMBO RIOBAMBA	ANITA CONDO	PAPAS	2 QQ	18.00	120.00
				TOMATE DE CARNE	10 FUNDAS	1.00	
				HIGOS	4 FUNDAS	1.00	
				FREJOL SECO	4 FUNDAS	1.00	
				BABACOS	8 UNIDADES	1.00	
				COL	1 UNIDAD	0.50	
				CHOCLO	1/2 SACO	1.00	
				ZAPALLO	1 UNIDAD	4.00	
				SAMBO	6 UNIDADES	1.00	
				HABAS	5 FUNDAS	1.00	
				HUEVOS	20 CUBETAS	3.50	
2	FERIAS CIUDADANAS	HUICHI RIOBAMBA	BLANCA GUAMAN	FREJOL SECO	10 FUNDAS	1.00	130.00
				TOMATE DE ARBOL	15 FUNDAS	1.00	
				MANZANA	9 BALDES	2.00	
				LECHUGA DE HOJA	20 UNIDADES	0.25	
				LECHUGA DE REPOLLO	10 UNIDADES	0.25	
				MASHUA	3 FUNDAS	0.50	
				AJO	5 ATADOS	1.00	
				BABACO	17 BABACOS	0.50	
				LIMON	5 FUNDAS	1.00	
				HIGO	4 FUNDAS	1.00	
				COL	12 UNIDADES	0.50	
				SAMBOS	4 UNIDADES	0.75	

FERIA-ENERO | FERIA-FEBRERO | FERIA-MARZO | FERIA-ABRIL | FERIA-MAYO | FERIA-JUNIO | FERIA-JULIO

Fig. 6-1 Datos originales

La tarea de captura de datos no cuenta con los suficientes controles para garantizar la calidad de la información, como consecuencia se encuentra una serie de inconsistencias, y se mencionan a continuación las mas relevantes:

- El mismo producto se lo encuentra con distintas denominaciones.

- El valor ingreso por ventas se encuentra consolidado como el sumatorio de la venta de productos individuales.
- Las unidades de comercialización de productos son distintas por ejemplo: atados, sacos, quintales, para un mismo producto.

El número total de datos encontrados (17.259) se desglosa mensualmente como se indica en la Fig 6-2.

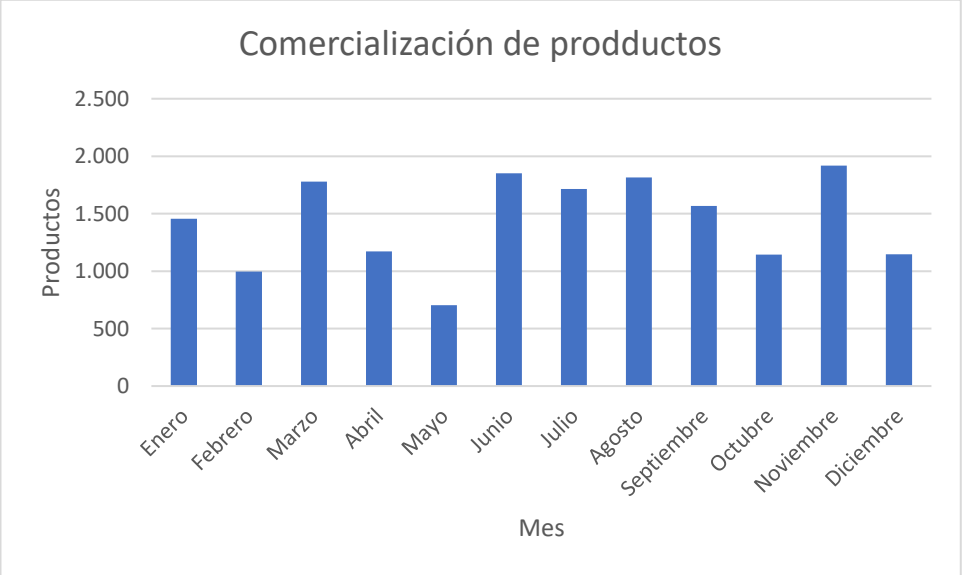


Fig. 6-2 Comercialización mensual

El valor medio de ítems se ubica en 1.400 datos/mes de distintos grupos como son hortalizas y legumbres, cárnicos y lácteos, frutas y tubérculos, productos procesados, comida preparada y varios.

La información mensual se encuentra detallada en 47 periodos semanales, los procesos de limpieza y calidad de datos se concentran en estandarizar

el nombre de los productos transformando todas las variantes a un solo identificador, por ejemplo, el producto arveja tiene cuatro identificativos ("Alberja", "Alberjas", "Alverja"," Arbeja")

En la Tab 6-1, se presenta una muestra de la calidad en la información en relación con el nombre: no existe uniformidad, tampoco se encuentra estandarizada en las unidades de venta del producto y en otros casos no existe ni la cantidad ni el valor de venta.

Producto	cantidad	valor unitario
ARVEJA	1 SACO	27.00
ALVERJA	1 SACO	10.00
ARVEJA	15 Lbs	1.00
ARVEJAS	1/2 SACO	13.00
ARVEJA	1/2 SACO	6.00
ALVERJA	10 FUNDAS	1.00
ARVEJA		
ARVEJA	1 SACO	27.00
ALVERJA	1 SACO	10.00
ARVEJA	15 Lbs	1.00
ARVEJAS	1/2 SACO	13.00
ARVEJA	1/2 SACO	6.00
ALVERJA	10 FUNDAS	1.00
ARVEJA	1 SACO	
ARVEJA	1/2 SACO	6.00
ARVEJA	1 SACO	27.00
ALVERJA	1 SACO	10.00
ARVEJA	15 Lbs	1.00
ARVEJA	1/2 SACO	6.00
ALVERJA	10 FUNDAS	1.00
ARVEJA	1 SACO	

Producto	cantidad	valor unitario
ARVEJA	1 SACO	27.00
ALVERJA	1 SACO	10.00
ARVEJA	15 Lbs	1.00
ALVERJA	10 FUNDAS	1.00
ARVEJAS	15 FUNDAS	1.00
ARVEJA	15 FUNDAS	1.00
ARVEJAS	1/2 SACO	7.00

Tab. 6-1 Datos producto Arveja

Las acciones realizadas sobre este archivo se enfocan en los siguientes pasos:

a.- Para cada registro verificar la calidad de los datos

1.- Eliminando los registros sin información

2.- Estandarizar nombres de productos

3.- Establecer unidades de medida únicas

b.- Generar el conjunto de productos de estudio.

Del conjunto total de datos entregados, este estudio se limita al subconjunto de hortalizas y legumbres que es el más representativo del sector agrícola familiar. Los productos que presentan una mayor transaccionalidad a lo largo del periodo 2014 se muestra en la tab 6-2.

Producto	Nombre Cientifico
Acelga	Beta vulgaris var. cicla
Ajo	Allium sativum
Arveja	Pisum sativum
Babaco	Carica pentagona
Brócoli	Brassica oleracea italica
Cebolla Blanca	Allium fistulosum
Cebolla Paiteña	Allium fistulosum
Choclo	Zea mays
Col	Brassica oleracea
Col verde	Brassica oleracea var. Sabellica
Coliflor	Brassica oleracea var. Botrytis
Espinaca	Spinacia oleracea
Frejol	Phaseolus vulgaris
Frutilla	Fragaria
Habas	Vicia faba
Hierbas	Coriandrum sativum, Petroselinum crispum
Lechuga	Lactuca sativa
Melloco	Ullucus tuberosus
Nabo	Brassica rapa
Papas	Solanum tuberosum
Pepinillo	Cucumis sativus
Pepino	Cucumis sativus

Producto	Nombre Cientifico
Pimiento	Capsicum annuum
Rábano	Raphanus sativus
Remolacha	Beta vulgaris
Tomate de árbol	Solanum betaceum
Tomate Riñón	Solanum lycopersicum
Vainita	Phaseolus Vulgaris L
Zanahoria	Daucus carota
Zapallo	Cucurbita máxima

Tab. 6-2 Productos seleccionados

Este conjunto con treinta productos es la base para aplicar varias transformaciones que le otorguen la característica de asociatividad, temporalidad y ubicación geográfica para encontrar patrones de comportamiento al aplicar técnicas de minería de datos.

6.2 Datos Discretizados

Con la información originalmente registrada no es posible inferir reglas de asociación, la asociatividad se presenta en la presencia o ausencia de un producto en una transacción de venta, por esta razón se convierte a un conjunto de datos de tipo binario [101], que permita la infererencia de reglas asociativas

La asociatividad entre productos comercializados necesita de una estructura de datos que identifique a un producto como parte de una transacción, cuando esto sucede se asigna una variable "s" y para el caso contrario "n", estas variables son reemplazadas por el valor de venta Tab. 6-3 Archivo Discretizado

Producto	10/5/2014	10/5/2014	10/5/2014	10/5/2014	10/5/2014	10/5/2014	10/5/2014	10/5/2014	10/5/2014	10/5/2014	10/5/2014
Acelga	s	s	n	n	s	n	s	n	s	n	s
Ajo	n	n	n	s	n	s	s	s	n	n	n
Arveja	n	s	n	n	s	s	s	n	n	s	n
Babacos	n	n	n	n	n	n	n	s	n	n	n
Brocoli	s	s	n	s	n	s	s	s	s	n	n
Cebolla blanca	n	n	s	s	s	n	s	s	s	s	s
Cebolla paiteña	s	n	s	n	s	n	n	n	s	s	n
Choclo	s	s	s	n	s	s	s	n	n	s	n
Col	s	n	s	s	s	s	s	n	s	n	s
Coliflor	n	n	n	s	n	s	n	n	n	n	n
Espinaca	n	n	n	n	n	n	n	n	n	n	n
Frejol	n	s	n	s	s	s	n	n	n	s	n
Frutilla	n	n	n	s	n	n	s	s	n	s	n
Habas	n	n	n	s	n	s	s	s	n	s	s
Hierbas	s	n	n	n	s	n	n	n	n	s	n
Lechuga	s	n	n	s	s	s	s	n	s	n	n
Melloco	n	s	n	n	n	n	s	n	n	n	n
Nabo	n	n	n	n	n	n	n	n	n	n	n
Papas	n	s	s	s	s	s	s	n	s	n	n
Pepinillo	n	n	s	s	n	n	n	n	s	s	n
Pepino	n	n	n	n	n	n	s	n	n	n	n
Pimiento	n	s	s	s	s	s	s	s	n	s	n
Rabano	n	n	s	s	s	n	n	s	n	n	s
Remolacha	n	n	n	s	n	n	s	n	s	n	n

Tab. 6-3 Archivo Discretizado

En base a los umbrales mínimos de soporte y confianza, cada producto tiene dos valores por para cada índice: el primero provisto para la variable "s" y el segundo generado por la variable "n", estos índices se calculan en base

a la cantidad de veces que aparecer juntos o que no aparecen con respecto al total de transacciones realizadas.

Con este sistema de discretización del conjunto inicial, se encuentra conjuntos de ítems frecuentes para las dos variables, al final las reglas de asociación generadas sirven para determinar si un producto aparece o si este no aparece, en ambos casos cumple con los umbrales mínimos.

Para disminuir el número de reglas y centrarnos solo en los productos que son parte de una transacción se utiliza un identificador binario (t,"“), cuando aparece el producto en una transacción se utiliza “t”, en caso de no aparecer se considera un espacio “”, este cambio permite reducir el número de reglas validas considerando solamente las que son parte de una transacción de compra del producto Fig 6-3)

No.	1: ACELGA Nominal	2: Aguacate Nominal	3: Ajo Nominal	4: Arveja Nominal	5: Babacos Nominal	6: BROCOLI Nominal	7: Camote Nominal	8: CEBOLLA BLANCA Nominal	9: CEBOLLA PAITE¾A Nominal	10: Choclo Nominal	11: COL Nominal	12: COL MORADA Nominal
1			t	t		t		t	t	t	t	
2	t	t	t			t		t	t	t	t	
3	t		t	t	t	t			t	t		
4	t		t			t		t	t			
5	t		t			t		t	t	t	t	
6	t	t		t		t		t		t	t	
7	t		t	t	t	t		t	t	t	t	
8	t					t			t	t		
9	t	t		t		t		t		t	t	
10	t	t	t			t		t		t	t	
11	t		t	t	t	t			t	t		
12	t					t		t	t	t	t	
13		t			t	t				t	t	
14		t	t	t		t		t	t		t	
15		t		t		t		t	t	t	t	
16	t	t				t		t		t	t	t
17	t	t		t		t		t		t	t	
18	t	t	t			t		t	t	t	t	
19	t		t	t	t	t			t	t		
20	t					t		t		t	t	
21	t	t		t	t			t				
22	t	t	t			t		t	t	t	t	
23		t		t		t		t	t	t	t	
24	t		t			t		t		t	t	
25		t			t							
26		t	t	t	t	t		t	t	t	t	
27		t				t		t	t	t	t	
28				t		t		t	t	t	t	
29		t	t		t				t	t	t	
30	t	t	t	t		t		t	t	t	t	
31	t		t	t		t		t	t	t	t	
32	t		t	t		t				t		t
33	t			t		t		t	t	t		
34	t	t				t		t	t	t	t	
35	t		t	t	t	t			t	t		

Fig. 6-3 Archivo formato final para generar reglas de asociación válidas (Informacion2014.arff)

6.3 Series de Datos Temporales

Una serie de datos temporales permite, buscar patrones de comportamiento en base a un atributo de temporalidad. En nuestro caso, la venta de productos tiene el registro de fecha, este campo permite utilizar series de tiempo para explicar a futuro el comportamiento de las transacciones.

Otro elemento que se considera para la generación de una serie de tiempo es el valor de comercialización de un producto, la unidad de medida encontrada para la comercialización de productos no es estándar, por lo que es necesario unificar la medida de venta y calcular el valor agregado para cada producto, generando un archivo con cuarenta y tres registros que

representan a las semanas del año que contienen la información de comportamiento comercial de los treinta productos según se muestra en la Fig. 6-4 con ventas semanales calculadas a partir del conjunto inicial.

No.	1: fecha	2: ACELGA	3: AJO	4: ARVEJA	5: BABACOS	6: BROCOLI	7: CEBOLLA BLANCA	8: CEBOLLA PAITEÑA	9: CHOCLO
	Date	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
17	2014-06-15	4.5	7.0	17.0	36.25	17.25	78.6	47.0	75.0
18	2014-06-22	13.25	96.0	24.0	15.0	36.45	197.0	77.5	52.0
19	2014-06-29	13.25	96.0	24.0	15.0	36.45	190.75	77.5	52.0
20	2014-07-20	33.5	76.4	112.5	28.9	47.8	183.05	71.75	82.0
21	2014-07-06	12.75	66.0	39.0	26.5	32.2	46.0	100.75	53.66
22	2014-07-13	12.5	18.0	13.0	6.0	26.4	82.6	36.0	245.0
23	2014-07-27	33.5	76.4	112.5	28.9	47.8	228.05	71.75	90.0
24	2014-08-03	7.25	55.0	40.0	10.0	25.95	54.0	97.0	69.66
25	2014-08-10	7.75	53.0	17.0	10.0	36.45	81.1	70.0	107.0
26	2014-08-17	7.0	53.0	24.0	10.0	47.9	95.7	97.0	112.0
27	2014-08-24	27.25	67.4	112.5	28.9	49.5	188.15	65.0	70.0
28	2014-08-31	7.0	53.0	24.0	10.0	47.9	90.7	97.0	112.0
29	2014-09-07	20.0	45.4	89.5	14.4	49.95	80.1	77.0	57.5
30	2014-09-14	19.75	45.4	89.5	14.4	49.95	121.9	77.0	57.5
31	2014-09-21	14.25	68.0	84.5	12.5	59.05	114.6	110.0	70.0
32	2014-09-28	13.75	55.4	52.0	6.4	68.95	91.4	65.0	25.0
33	2014-10-05	21.25	45.4	70.5	6.4	43.7	56.9	89.0	69.0
34	2014-10-12	29.75	57.4	144.5	20.9	42.35	249.8	113.0	70.0
35	2014-10-19	28.25	45.0	132.0	20.9	38.45	245.0	113.0	60.0
36	2014-11-02	12.25	52.4	52.0	6.4	66.45	86.4	71.0	15.0
37	2014-11-09	12.25	52.4	52.0	6.4	66.45	86.4	71.0	15.0
38	2014-11-16	28.25	57.4	117.0	21.3	54.85	232.3	110.0	60.0
39	2014-11-23	12.25	68.0	84.5	12.5	59.05	137.1	86.0	70.0
40	2014-11-30	7.0	53.0	24.0	10.0	47.9	90.7	73.0	58.0
41	2014-12-07	19.75	45.4	89.5	14.4	49.95	121.9	77.0	58.0
42	2014-12-14	14.25	55.4	52.0	6.4	68.95	93.5	65.0	25.0
43	2014-12-21	28.25	57.4	117.0	21.3	54.85	232.3	98.0	60.0

Fig. 6-4 Registro de ventas año 2014

6.4 Datos Espaciales

Una tercera forma de buscar patrones de comportamiento se obtiene utilizando el componente de ubicación geográfica de ferias donde se realizan las transacciones de productos.

Se debe considerar que una estructura de tipo espacial tiene dos componentes:

- Bounding box(@bbox), es el delimitador de la ubicación espacial
- Coordinate Reference System (CRC), indica las coordenadas utilizadas

Al tener varias fuentes de datos, como son:

- La región que ocupan las provincias de Tungurahua y Chimborazo.
- Ubicación de circuitos alternativos de comercialización tipo feria
- Area a considerar para realizar la interpolación pronóstica
- Valores piso climático

Cada elemento posee sus propias características espaciales, en unos caos ya son establecidas de forma previa, en otros casos deben ser creados o transformados cuando utilizan distintos sistemas de coordenadas.

Esta sección detalla la manera que se adiciona o transforma el componente espacial de cada fuente de datos y de su integración en una estructura de datos que permita realizar interpolación pronóstica.

Una coordenada geográfica se representa por medio de un punto con latitud que es una referencia a la línea ecuatorial o paralelo 0 y la longitud que referencia al meridiano 0 conocido como Meridiano de Greenwich como se muestra en la Fig 6-5.

El mapa de cualquier país del mundo se puede encontrar en [102], para este caso utilizamos uno del país Ecuador de nivel 2, con denominación ECU_ADM2.RDS, que presenta la ubicación de provincias y cantones

Para el manejo de la información geográfica en esta investigación se ha tomado como herramienta de apoyo el lenguaje R y su framework de desarrollo R Studio[103],[104].

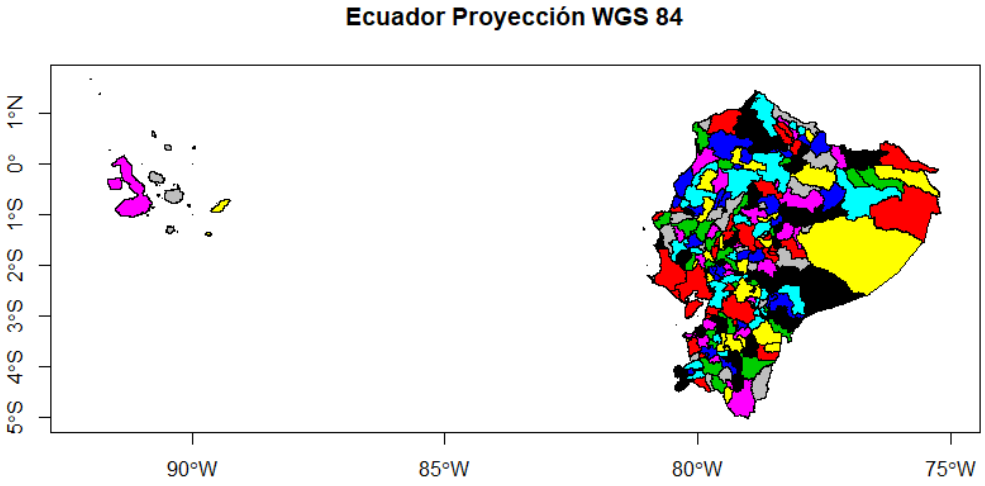


Fig. 6-5 Ubicación del Ecuador Latitud y Longitud

El Ecuador se encuentra situado en las coordenadas geográficas 1.8312° S, 78.1834° W, las provincias de Tungurahua y Chimborazo se ubican en la zona Andina central y la ubicación geográfica de cada feria en estas provincias,

así como su denominación se encuentran descritas en la Tab. 6-4 Descripción geográfica de Ferias

Feria	x	y
Colta	-78.7238	-1.888
Tisaleo	-78.6923	-1.40951
RIOBAMBA6	-78.6737	-1.66153
RIOBAMBA7	-78.673	-1.66089
RIOBAMBA8	-78.6687	-1.66025
RIOBAMBA4	-78.6588	-1.67626
Cevallos	-78.6566	-1.25714
HUICHI RIOBAMBA	-78.6558	-1.6871
RIOBAMBA5	-78.6538	-1.67599
RIOBAMBA3	-78.65	-1.67803
SAN FARNCISCO RIOBAMBA	-78.6497	-1.67468
RIOBAMBA1	-78.6462	-1.68942
CHAMBO RIOBAMBA	-78.6077	-1.7303
Pillaro	-78.551	-1.32836

Tab. 6-4 Descripción geográfica de Ferias

Para la creación del componente espacial de una feria, se inserta el valor de coordenadas geográficas utilizando el sistema geodésico mundial (WGS 84) [105].

El proceso completo para la trasformación a un archivo de tipo espacial se encuentra en [106], un objeto de la clase Spacial, se constituye principalmente por dos "slots":

1.- Bounding box, que permite conocer el cuadro delimitador de la ubicación espacial (@bbox).

El área del Ecuador @bbox se encuentra en la Tab 6-5

@bbox Ecuador	Min	Max
X	-92.008904	-75.200043
y	-5.017375	1.681548

Tab. 6-5 @bbox Ecuador

El valor de la componente para las provincias de Tungurahua y Chimborazo que se extrae de los datos del mapa del país Ecuador se presenta en Tab 6-6

@bbox Tungurahua/Chimborazo	Min	Max
X	-79.133499	-78.109627

@bbox Tungura-gua/Chimborazo	Min	Max
Y	-2.556218	-0.970623

Tab. 6-6 @bbox Tungurahua y Chimborazo

2.- CRC (Coordinate Reference System) indica la proyección utilizada para ambos casos es WGS 84 (@proj4string)[107], `+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0`

La estructura de tipo espacial resultante para la ubicación de Cialco tipo feria se encuentra en la Tab 6-7, `crs= +proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0`

@bbox CIALCO Fe-rias	Min	Max
x	-78.723827	-78.551033
Y	-1.888003	-1.257143

Tab. 6-7 @bbox Cialco tipo feria

La representación geográfica resultado de la trasformación se encuentra en la Fig 6-6

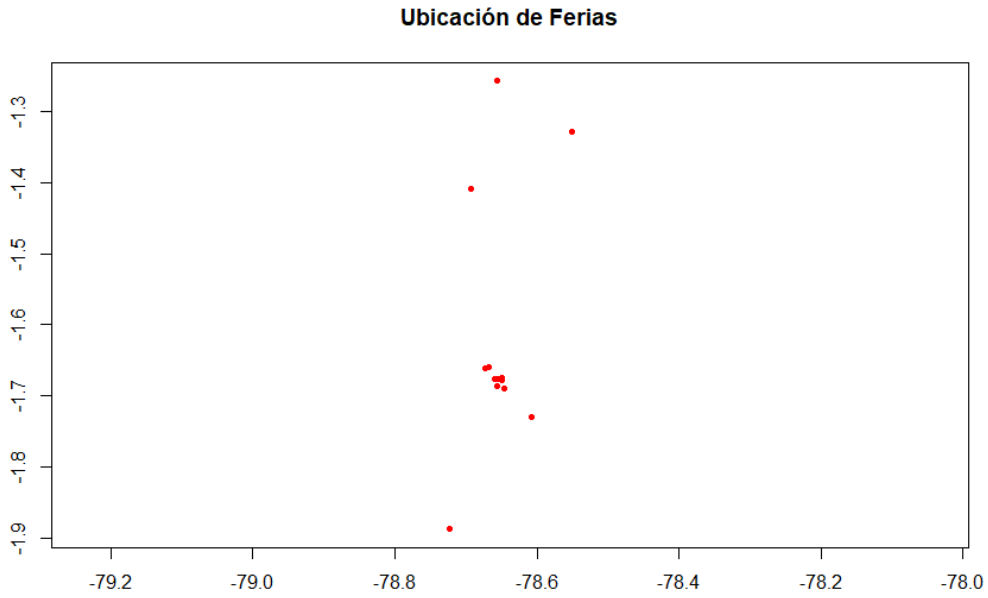


Fig. 6-6 Ubicación geográfica de Ferias

Al integrar la información geográfica de provincias con la ubicación de las ferias se visualiza su ubicación Fig 6-7 , en un mapa de la región.

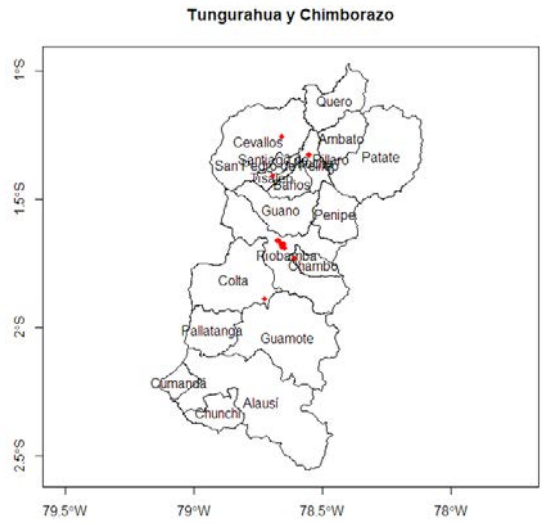


Fig. 6-7 Representación geográfica de Ferias

A la estructura espacial de circuitos alternativos de comercialización se adiciona los atributos correspondientes a valores agregados de ventas de productos del conjunto de ferias de 2014.

La distribución espacial de la variable permite hacer predicciones y fusión a nivel de datos con otras fuentes de información.

6.5 Fusión de datos externos

Hasta el momento se ha generado información con características propias que permiten encontrar asociatividad, temporalidad y ubicación geográfica entre productos comercializados en mercados locales, con el fin de validar la metodología propuesta se adicionan datos externos al proceso de comercialización como población y piso climático.

La segunda fuente de datos es entregada por el Instituto Nacional de Meteorología e Hidrología del Ecuador[108], con los pisos climáticos para la región geográfica de las ferias.

La información se encuentra en el formato EPSG:3857 que es la proyección del globo terráqueo a un plano, la información hasta el momento generada tiene el formato EPSG: 4326[109] que es la proyección de coordenadas a una elipsoide, para unificar los tipos de coordenadas se realiza la conversión al tipo 4326

En la Fig 6-8, se puede observar tres capas de datos la primera presenta los pisos climáticos en variados colores, una segunda capa en color negro que representa la malla o área de calculo de estimaciones a futuro y finalmente una tercera capa en puntos de color rojo que representan la ubicación de las catorce ferias.

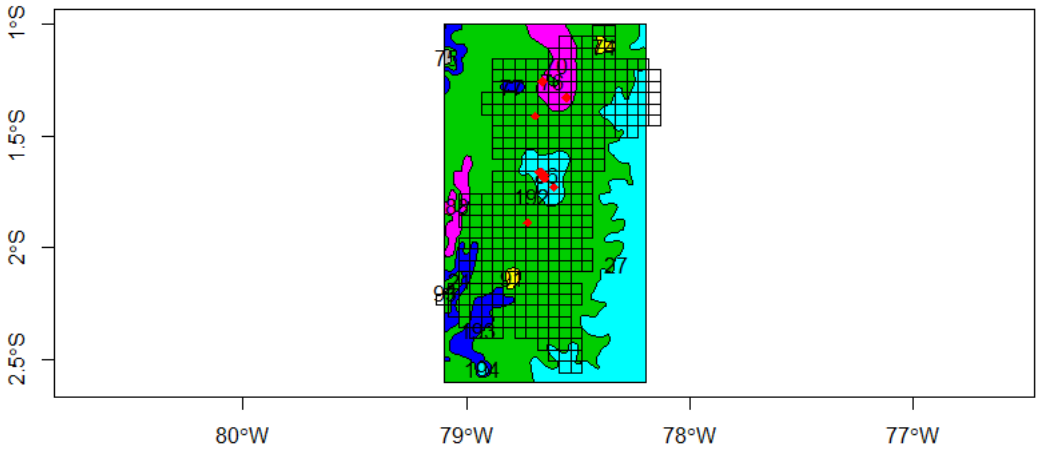


Fig. 6-8 Presentación datos fusionados

Además, el Instituto Nacional de Estadísticas y Censos de Ecuador[110] proporciona una tercera fuente de datos, con la población estimada en cada uno de los cantones de las provincias.

En la Fig 6-9, se presenta la información que corresponde al conjunto de datos fusionados de diversas fuentes de la siguiente manera:

Identificación de cada feria en la columna uno, las columnas dos a seis contiene los elementos asociados, las columnas siete y ocho las coordenadas geográficas, la columna nueve los datos de población del cantón donde se ubica cada feria y en las siguientes columnas se incluyen para el piso climático sus elementos de precipitación, temperatura, humedad y evaporación.

Feria	BROCOLI	CEBOLLA E	TOMATE	TOMATE D	ZANAHOR	x	y	Poblacion	pclimatico	precip	temp	hum	evap
Colta	6	60	26	30	30	-78.7238	-1.888	44.701	192	4	18	-16.5	64
Tisaleo	37.9	36	88.25	185	47.5	-78.6923	-1.40951	10.565	192	4	18	-16.5	64
RIOBAMBA	8	9.6	18	0	0	-78.6737	-1.66153	193.315	86	6.5	18	-16.5	64
RIOBAMBA	26	42.5	60	15	17.5	-78.673	-1.66089	193.315	86	6.5	18	-16.5	64
RIOBAMBA	0	10	8	8	5	-78.6687	-1.66025	193.315	86	6.5	18	-16.5	64
RIOBAMBA	2.5	74	60	35	29.5	-78.6588	-1.67626	193.315	86	6.5	18	-16.5	64
Cevallos	9	9.6	51	45	20	-78.6566	-1.25714	6.873	70	6.5	18	-16.5	64
RIOBAMBA	7.2	50	30	35	20	-78.6558	-1.6871	193.315	86	6.5	18	-16.5	64
RIOBAMBA	6	65	42	27	29	-78.6538	-1.67599	193.315	86	6.5	18	-16.5	64
RIOBAMBA	2	0	10	0	5	-78.65	-1.67803	193.315	86	6.5	18	-16.5	64
RIOBAMBA	6	26.5	33	0	36	-78.6497	-1.67468	193.315	86	6.5	18	-16.5	64
RIOBAMBA	2.5	0	6.9	0	3	-78.6463	-1.68942	193.315	86	6.5	18	-16.5	64
CHAMBO	21	50	30	20	20	-78.6077	-1.7303	10.541	86	6.5	18	-16.5	64
Pillaro	20.1	141.5	92	78	40	-78.551	-1.32836	34.925	70	6.5	18	-16.5	64

Fig. 6-9 Archivo datos fusionados

En el área que corresponde a las catorce ferias, hay un tipo de clima mesotérmico con una temperatura promedio de 18 ° C durante todo el año, la precipitación con un promedio de 650 mm, el índice de humedad que está entre -16.5 y 10, y la evaporación potencial con valores que varían entre 64 y 106 mm.

7. Trabajo realizado

Este capitulo presenta la implementación de la metodología propuesta, yl se divide en dos secciones:

La primera parte presenta la implementación de tres procesos de minería de datos: el primer proceso de minería de datos descriptiva se encarga de la generación de reglas de asociación, para obtener un conjunto multivariable de productos, el segundo proceso utiliza la minería de datos predictiva y la utilización de la dimensión temporal para generar series de tiempo, con las que se encuentra estimaciones de venta de productos agrícolas a futuro y el tercero la utilización de la dimensión espacial para establecer el valor de venta de productos donde no se tenga una medida real.

La utilización de modelos lineales de regresión se aplica primero a la dimensión tiempo, con el modelo generado se realizar una extrapolación a la serie de tiempo y se obtiene un valor a futuro.

Para la dimensión espacio, el modelo lineal de regresión generado se aplica interpolación para establecer valores de comercialización en sitio que no se posee información.

En la segunda parte del capitulo, se realiza una integración de datos utilizando los elementos del conjunto de productos asociados y se establece el valor de la predicción para la dimensión temporal como para la espacial.

Para completar la metodología propuesta se evalua los dos procesos: el que utiliza un solo producto predictor y el proceso de integración de productos asociados como predictor, con los resultados se establece la variación de la estimación pronóstica.

7.1 Conjunto de productos asociados

Para generar el conjunto con productos que presenten una mayor asociatividad basados en la comercialización, se utiliza algoritmos de la herramienta WEKA [111], específicos para inducir reglas de asociación.

Del conjunto inicial de 30 productos se han registrado 549 transacciones, para encontrar relaciones de asociatividad, aplicando el algoritmo no supervisado "Apriori" al archivo discretizado generado en la sección 6.2 Datos Discretizados, utilizando dos parámetros: Soporte igual a 0.4 (220 apariciones) y Confianza igual a 0.8 el conjunto resultante de las mejores reglas de asociación de la forma $A \rightarrow B$ se presenta en la Tab 7-1

Producto A	aciertos	Producto B	aciertos	Confi-anza	Sus-tentación	Apalan-camiento
Cebolla blanca	338	Tomate Ri-ñon	293	0,87	1,1	0.05
Tomate de árbol	312	Tomate Ri-ñon	268	0,86	1,09	0,04
Zanahoria	374	Tomate Ri-ñon	311	0,83	1,06	0,03
Brócoli	327	Tomate Ri-ñon	269	0,82	1,05	0,02

Tab. 7-1 Reglas de Asociación Algoritmo Apriori

Los productos que participan en la generación de las mejores reglas de asociación constituyen el conjunto mulivariable siguiente:

A={Tomate Riñón, Cebolla Blanca, Tómate de Árbol, Zanahoria, Brócoli }

Este conjunto debe estar integrado por el mayor número de productos que tengan una relación, por este motivo se utiliza un segundo algoritmo, denominado fp growth, que basa su búsqueda en parámetros mínimos de soporte y confianza, los resultados obtenidos se muestran en la tab 7-2

Producto A	# aciertos	Producto B	# aciertos	Confianza
Cebolla blanca	338	Tomate Riñon	293	0,87
Tomate de árbol	312	Tomate Riñon	268	0,86
Zanahoria	374	Tomate Riñon	311	0,83
Brócoli	327	Tomate Riñon	269	0,82

Tab. 7-2 Reglas de Asociación Algoritmo FP-Growth

Utilizando los dos algoritmos se encuentra los mismos elementos participantes en las reglas de asociación inducidas, las reglas presentan una sustentación (lift) mayor a 1 que indica la existencia de una relación y no solo una ocurrencia aleatoria.

El apalancamiento (leverage) en todos los casos es superior a cero y la convicción indica que los cuatro productos influyen en la compra del tomate riñon.

Utilizando los dos algoritmos de forma independiente sobre el mismo conjunto de datos se concluye que los elementos resultante de escoger las mejores reglas de asociación conforman el conjunto multivariable A={Tomate Riñón, Cebolla Blanca, Tómate de Árbol, Zanahoria, Brócoli}, que se utilizá en los procesos para búsqueda de patrones multipredictor.

7.2 Asociación en series de Tiempo

Con la información del archivo generado en la sección 6.3 Series de datos temporales se han aplicado dos tipos de análisis: el probabilístico y el basado en aprendizaje automático.

7.2.1 Modelo Probabilistico

Para el análisis de datos utilizando un modelo probabilístico se considera que una serie de tiempo debe tener cuatro componentes: Tendencia, Cíclico, Estacional, Aleatoriedad [28]

La tendencia de la serie se obtiene como el desplazamiento de la serie en periodos de tiempo largos promediando las variaciones a corto plazo, el componente cíclico está dado por todos los puntos que se encuentren por encima o por debajo de una línea de tendencia en un periodo establecido y el componente irregular o aleatoriedad mide la variabilidad de la serie cuando los otros componentes se eliminan o no existen

Usando técnicas estadísticas como las mencionadas en [67], no es posible determinar una función de predicción, la información disponible se limita a un año, la descomposición de la serie de tiempo se la presenta Fig 7-1

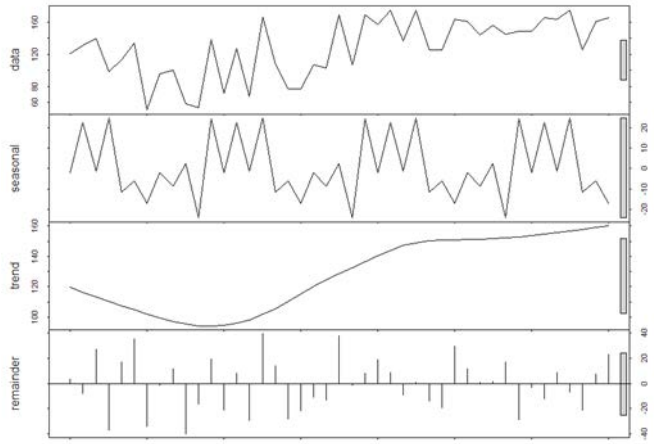


Fig. 7-1 Descomposición serie del Tomate

Al descomponer las series de datos, podemos ver una estacionalidad de tipo trimestral, en lo que tiene que ver con la tendencia en el primer trimestre del gráfico, que muestra una disminución que se compensa a lo largo del año y termina con una tendencia al crecimiento, el componente de aleatoriedad es bastante alto.

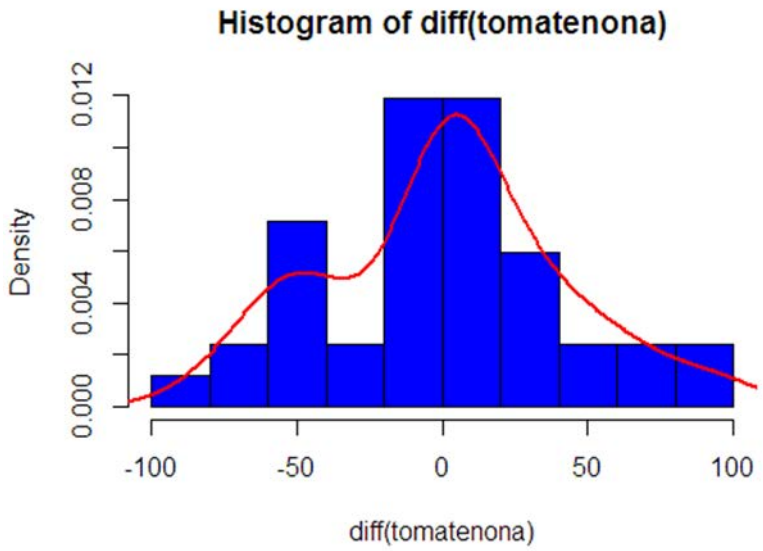


Fig. 7-2 Prueba estadística Distribución Normal

Con base en el gráfico de la prueba de normalidad de los residuos, Fig. 7-2 Prueba estadística Distribución Normal no se aprecia una tendencia a una distribución normal, que se puede confirmar con los gráficos presentados en la Fig. 7-3 Curvas distribución Normal.

La línea roja representa el gráfico de los residuos, mientras que la línea azul es el gráfico de una distribución normal simulada, lo que no permite establecer una similitud entre la distribución normal simulada con la distribución normal de los residuos.

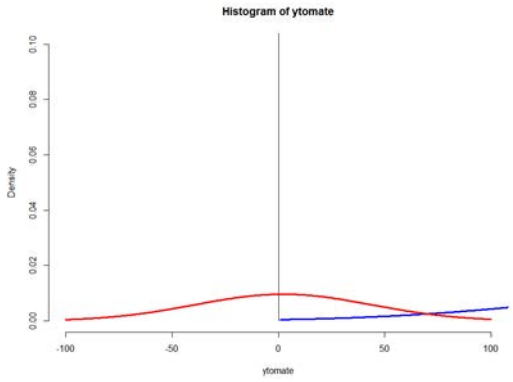


Fig. 7-3 Curvas distribución Normal

El estudio indica que estadísticamente no es posible realizar una predicción de consumo utilizando esta metodología, especialmente por la poca cantidad de datos encontrados que no permite establecer claramente la distribución de los residuos.

7.2.2 Modelo estimadores núcleo (kernel)

Si se tiene un conjunto de datos pero se desconoce su distribución se puede utilizar una función de aproximación [112], una opción es la utilización de un método kernel, que son funciones que se asocian con cada uno de los datos, la suma ponderada de estas funciones se constituye en el estimador para aproximar la función de densidad.

Un estimador kernel realiza una expansión de la dimensionalidad para obtener una solución lineal en este espacio expandido. [113]

Este trabajo utiliza el algoritmo SMOImproved[65], que es un SVM para calcular regresión (un caso particular de SVR Support Vector Regression), donde el punto de dato es un ejemplo independiente del conjunto del que hay que aprender y el orden de los puntos de datos dentro de un conjunto de datos no es relevante [66]

Los algoritmos utilizados se encuentran en la herramienta weka paquete Forecast.

La representación gráfica para el archivo serietiempofinal.arff que contiene la serie de datos para los productos asociados se muestra Fig 7-4.

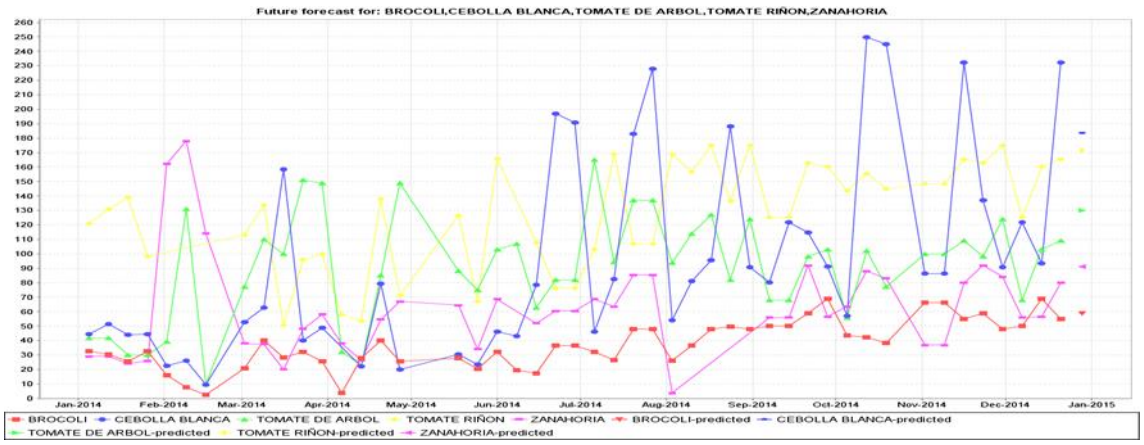


Fig. 7-4 Serie de tiempo productos asociados

Cada punto resaltado identifica la fecha y el valor del producto comercializado

Definido el conjunto de productos asociados A (sección 7.1), se realiza la extrapolación pronóstica a futuro utilizando una serie de tiempo, con la siguiente metodología

1.- Establecer estimación pronostica de una variable seleccionada utilizando un predictor.

2.- Establecer estimación pronostica de una variable utilizando el conjunto asociativo como predictor

3.- Medir las diferencias de estimación pronostica.

En las siguientes secciones, se detalla la ejecución de cada punto

7.2.3 Predicción a futuro variable Tomate: un predictor

El calculo de la serie de tiempo se basa en la creación de variables retrasadas (lagged) que son el principal mecanismo para establecer una relación entre los valores pasados y actuales de una serie para utilizar algoritmos de aprendizaje.

Por otro lado un elemento que siempre debe tomarse en cuenta es el tiempo como variable independiente. En el entorno de WEKA forecast, la trasformación de la variable fecha a un número secuencial se establece con la relación: (fecha/(1000*60*60*24*365.25) + 21)*12.

Por otro lado, para el algoritmo de estimación SMOreg, el kernel polinomial que se ha utilizado es $K(x,y)=(<x,y>+1)^p$ [114]

De las variables disponibles, el producto para predecir es tomate riñón (tomate) y se realiza una estimación de un periodo adelante utilizando todo el conjunto de datos para entrenamiento (training).

Las variables de retardo generadas son:

TOMATE RIÑON
fecha-remapped
Lag_TOMATE RIÑON-1
Lag_TOMATE RIÑON-2
Lag_TOMATE RIÑON-3
Lag_TOMATE RIÑON-4
Lag_TOMATE RIÑON-5
Lag_TOMATE RIÑON-6
Lag_TOMATE RIÑON-7
Lag_TOMATE RIÑON-8
Lag_TOMATE RIÑON-9
Lag_TOMATE RIÑON-10
Lag_TOMATE RIÑON-11
Lag_TOMATE RIÑON-12
fecha-remapped^2
fecha-remapped^3
fecha-remapped*Lag_TOMATE RIÑON-1
fecha-remapped*Lag_TOMATE RIÑON-2
fecha-remapped*Lag_TOMATE RIÑON-3
fecha-remapped*Lag_TOMATE RIÑON-4
fecha-remapped*Lag_TOMATE RIÑON-5
fecha-remapped*Lag_TOMATE RIÑON-6
fecha-remapped*Lag_TOMATE RIÑON-7

fecha-remapped*Lag_TOMATE RIÑON-8
fecha-remapped*Lag_TOMATE RIÑON-9
fecha-remapped*Lag_TOMATE RIÑON-10
fecha-remapped*Lag_TOMATE RIÑON-11
fecha-remapped*Lag_TOMATE RIÑON-12

El modelo de datos para la estimación a futuro generado consta de la fecha recalculada (remapped), doce variables de retardo para tomate, la fecha recalculada elevada al cuadrado, la fecha recalculada elevada al cubo y la fecha recalculada por cada una de las variables de retardo.

El modelo de datos generado de grado tres para el calculo de estimación a futuro por un periodo del producto tomate utilizando el algoritmo SMOreg es:

+ 0.2704 * (normalized) fecha-remapped - 0.0819 * (normalized) Lag_TOMATE RIÑON-1
- 0.0234 * (normalized) Lag_TOMATE RIÑON-2 + 0.086 * (normalized) Lag_TOMATE RIÑON-3
+ 0.1293 * (normalized) Lag_TOMATE RIÑON-4 - 0.1522 * (normalized) Lag_TOMATE RIÑON-5
- 0.046 * (normalized) Lag_TOMATE RIÑON-6 - 0.0425 * (normalized) Lag_TOMATE RIÑON-7
+ 0.0798 * (normalized) Lag_TOMATE RIÑON-8 - 0.2687 * (normalized) Lag_TOMATE RIÑON-9
+ 0.0852*(normalized) Lag_TOMATE RIÑON-10 + 0.1948 *(normalized) Lag_TOMATE RIÑON-11
+ 0.2034 * (normalized) Lag_TOMATE RIÑON-12 + 0.0165 * (normalized) fecha-remapped^2
- 0.2971 * (normalized) fecha-remapped^3
- 0.0503 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-1
+ 0.0129 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-2
+ 0.3388 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-3
+ 0.0563 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-4
- 0.0855 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-5
+ 0.2634 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-6

+ 0.1612 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-7
+ 0.1624 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-8
+ 0.1046 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-9
- 0.2224 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-10
- 0.2899 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-11
+ 0.0725 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-12
+ 0.2401

7.2.4.- Predicción a futuro variable Tomate predictor: conjunto asociativo

Este segundo modelo utiliza el conjunto de productos asociados A={ tomate riñón, brócoli, cebolla blanca, tomate de árbol y zanahoria}, como predictores para estimar valores a futuro de la variable tomate, al igual que la sección anterior se realiza una estimación a un paso adelante para todo el conjunto de datos para el proceso de entrenamiento.

El modelo de datos para la estimación a futuro multivariable incluye cuatro variables adicionales al generado en la sección anterior: Brocoli, Cebolla Blanca, Tomate de árbol y Zanahoria quedando de la siguiente manera:

+ 0.0976 * (normalized) BROCOLI- 0.1374 * (normalized) CEBOLLA BLANCA
+ 0.2727 * (normalized) TOMATE DE ARBOL + 0.3097 * (normalized) ZANAHORIA
+ 0.0772 * (normalized) fecha-remapped + 0.0682 * (normalized) Lag_TOMATE RIÑON-1
+ 0.0762 * (normalized) Lag_TOMATE RIÑON-2 + 0.1079 * (normalized) Lag_TOMATE RIÑON-3
+ 0.2096 * (normalized) Lag_TOMATE RIÑON-4 - 0.0366 * (normalized) Lag_TOMATE RIÑON-5
- 0.2354 * (normalized) Lag_TOMATE RIÑON-6 - 0.1308 * (normalized) Lag_TOMATE RIÑON-7
+ 0.0867 * (normalized) Lag_TOMATE RIÑON-8 - 0.0455 * (normalized) Lag_TOMATE RIÑON-9
+ 0.1036 * (normalized) Lag_TOMATE RIÑON-10 +0.3452*(normalized) Lag_TOMATE RIÑON-11
+ 0.243 * (normalized) Lag_TOMATE RIÑON-12 - 0.1284 * (normalized) fecha-remapped^2
- 0.3918*(normalized)fecha-remapped^3
- 0.1335*(normalized)fecha-rmapped*Lag_TOMATE RIÑON-1

- 0.0588 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-2
+ 0.4899 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-3
+ 0.0954 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-4
+ 0.0548 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-5
+ 0.4146 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-6
+ 0.1854 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-7
+ 0.0636 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-8
- 0.0807 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-9
- 0.0885 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-10
- 0.4103 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-11
- 0.0083 * (normalized) fecha-remapped*Lag_TOMATE RIÑON-12
- 0.2791

7.2.5 Medir diferencias pronósticas: Series de Tiempo

La Fig 7-5 realiza la representación gráfica de la estimación del producto tomate utilizando un solo predictor

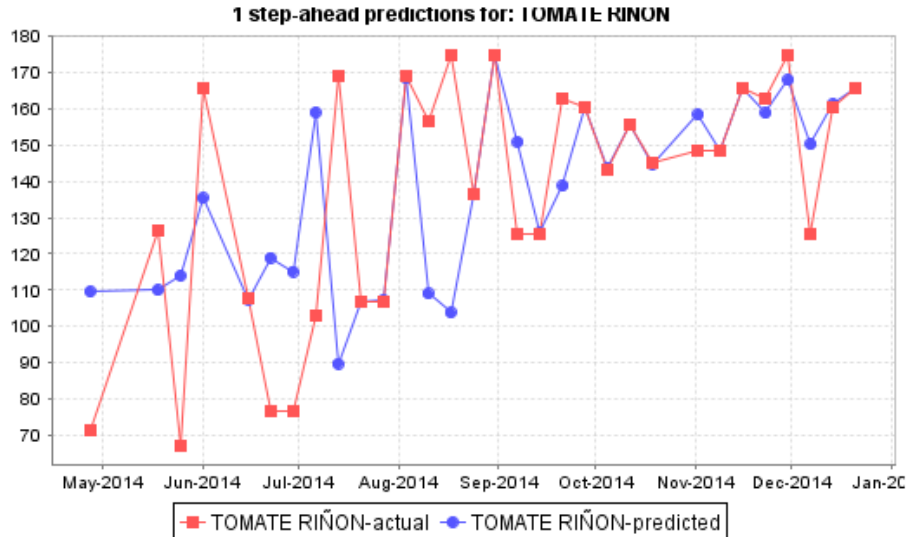


Fig. 7-5 Datos estimados a futuro variable Tomate un predictor

El grafico de la estimación a futuro por un periodo utilizando el conjunto de productos asociados como predictor, se puede apreciar en Fig 7-6

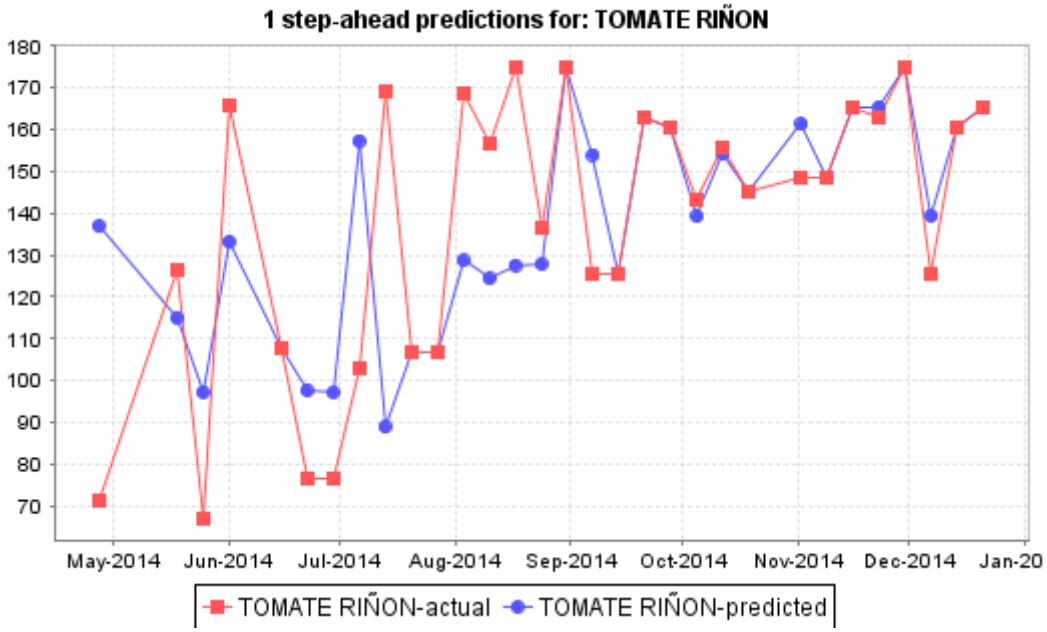


Fig. 7-6 Datos estimados a futuro multivariable

Los resultados obtenidos para cada caso se muestran en la tab 7-3.

Producto Predictor	Tomate	Conjunto asociativo A
Producto estimado a futuro	Tomate	Tomate
MAE	18.302	16.4369
RMSE	29.4848	27.1553

Tab. 7-3 Estimación futura Tomate.

La medida del error Mae (Mean absolute error) y Rmse (Root mean squared error), tiene una mejora cuando se utiliza el conjunto asociativo A como predictor, como se presentan en Fig 7-4

Valor error medido	Val mult/val univar	% de mejora
Mean absolute error	0.898181441	10.18185594
Root mean squared error	0.92099319	7.900681029

Tab. 7-4 Medición error Series de tiempo

En los dos casos el nivel de error de la predicción disminuye, en el Mae hay una mejora de un 10% , para el Rmse se puede apreciar una mejora de alrededor de un 8%

7.3 Asociación en Geoestadística

Al igual que la sección anterior se define la metodología para utilizar datos con ubicación geográfica, que se resume en cuatro pasos siguientes:

1.- Preparar formatos de datos con coordenadas geográficas

2.- Establecer los valores de estimación a futuro para una variable

3.- Utilizar el conjunto de productos asociados para conseguir mejorar las predicciones espaciales con un modelo multivariable

4,- Medir resultados interpolación predictiva y validar los resultados utilizando validación cruzada

A continuación, se explica a detalle cada uno de los pasos.

7.3.1 Preparar Información espacial.

Las provincias de Tungurahua y Chimborazo representan la región a interpolar valores predictivos, la primera actividad consiste en delimitar este sector creando una malla que lo cubra.

Cada celda de la malla, tiene una dimensión de 0,05 x 0,05 y en total el área de las dos provincias posee 21x32 celdas, sus coordenadas de inicio y fin se presentan en la tab 7-5.

@bbox A. Predicción	Min	Max
X	-79.133499	-78.109627
Y	-2.556218	-0.970623

Tab. 7-5 Area para interpolación

El país Ecuador se encuentra atravesado por el paralelo 0 o Línea Ecuatorial, en el sector del planeta en el que un grado de longitud equivale a 111,32 Km. Por lo tanto, la distancia ocupada en longitud por las dos provincias es de 1,049 grados o 116 km, la distancia entre celdas es el equivalente a 5.84Km, Fig 7-7.

El área de predicción tiene una dimensión de 116Km x 187Km o 21.692 Km^2.

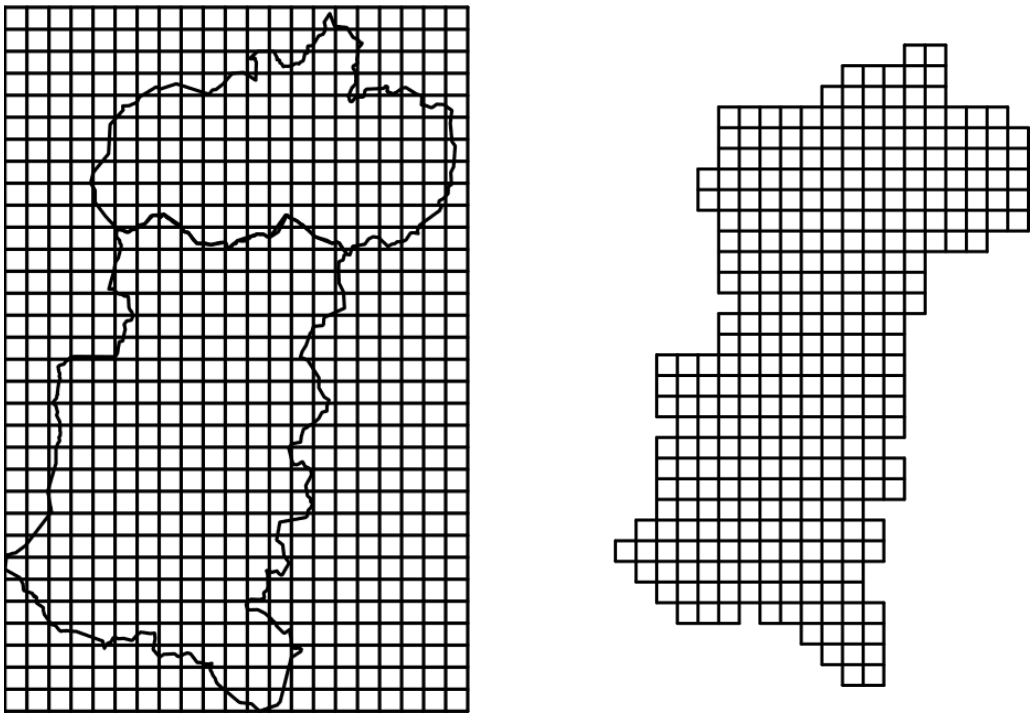


Fig. 7-7 Área para realizar pronóstico

7.3.2 Establecer estimación con un predictor

El proceso para conseguir la estimación utilizando la dimensión espacial se resume en los siguientes pasos:

- Encontrar los variogramas experimental y teórico en base a los datos generados en la sección 6.4
- Calcular la estimación utilizando kriging

7.3.2.1 Calculo de variograma experimental y teórico

El variograma permite analizar el comportamiento de la variable tomate sobre el área de predicción, en la Fig. 7-8 Variograma Ajustado se aprecia la distribución de puntos al calcular el variograma experimental con los datos de las ferias que aportan datos significativos para el mes de Julio del año 2014. La distancia entre los puntos está expresada en décimas de grado, y entre cada salto existe una distribución en promedio de dos ferias. El variograma modelo (m) utilizado es una función de tipo esférica m <- vgm(2151, "Sph", 0.473)

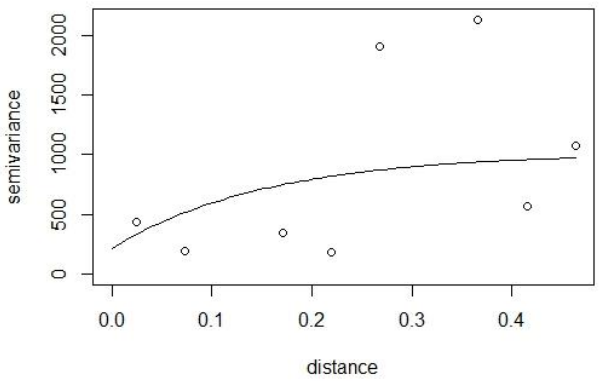


Fig. 7-8 Variograma Ajustado

7.3.2.2 Kriging

Utilizando la función continua del variograma ajustado se obtiene los valores de predicción de consumo del tomate en base a la distancia y la correlación espacial de la siguiente manera:

krige(TOMATE~1, FJespacial, spgrid, model = m), el ~ 1 define un único predictor constante.

Basados en el método de kriging, los valores encontrados en la interpolación varían especialmente en dos focos sobre los cuales se genera las predicciones. Los valores más cercanos a los puntos focales de los que se posee información son de más influenciados que los que se encuentran alejados, los valores encontrados para la estimación futura se encuentran en Tab 7-6

[1] 53.51724 53.17939 54.76818 55.08448 55.00930 54.61879 54.05662
[8] 55.47078 56.58642 57.07643 56.82893 56.08847 55.15481 53.63914
[15] 54.07811 54.31013 54.41375 55.17990 57.35292 59.45051 60.09012
[22] 59.35250 57.94507 56.43149 55.08112 53.97633 53.11356 54.76428
[29] 55.61047 56.11085 55.87149 55.77412 60.47876 64.25577 64.59281
[36] 62.59732 60.03706 57.72975 55.87791 54.46444 53.41289 52.64192
[43] 56.13752 57.79761 59.35441 60.16816 60.89148 67.00942 72.24208
[50] 70.60947 65.97080 61.87284 58.74218 56.44350 54.78268 53.59225
[57] 52.74266 55.38555 57.43928 60.22484 63.69215 67.12672 69.83976
[64] 74.15514 81.72601 74.86977 67.70443 62.64851 59.09047 56.58994
[71] 54.83339 53.59917 52.73133 55.62703 58.09523 61.78589 67.22813
[78] 74.19763 75.68254 75.23533 75.54771 71.81865 66.33699 61.84237
[85] 58.55191 56.21246 54.56453 53.40696 52.59402 57.56855 61.20076
[92] 66.65137 73.87333 74.05992 71.12487 69.08280 66.31421 62.88508
[99] 59.72324 57.21437 55.34641 53.99346 53.02559 52.33703 55.78336

[106] 58.32051 61.62539 64.66629 65.15567 63.75594 62.20227 60.54862
[113] 58.71196 56.91743 55.37079 54.13808 53.19800 53.16402 54.21275
[120] 55.27005 55.86306 55.69421 55.24886 54.99569 54.80235 54.44812
[127] 53.92418 50.27473 49.91161 49.06783 47.59171 46.25206 46.42477
[134] 47.79996 49.32740 50.44540 51.08342 46.07677 43.76772 40.23274
[141] 36.36505 37.44381 41.13078 44.57437 47.04804 48.66065 45.29107
[148] 43.06712 40.00185 36.00179 33.12451 29.59589 36.35818 41.24496
[155] 44.61094 43.53963 40.89549 37.62066 33.71579 29.07356 29.89261
[162] 35.15009 39.83213 43.32021 47.62034 46.35903 44.61732 42.28634
[169] 39.35333 36.09668 33.22810 31.73663 32.93354 36.29557 39.95652
[176] 43.02853 47.32981 45.96754 44.08715 41.55108 38.29565 34.69132
[183] 32.55528 33.12219 35.15199 37.93758 40.83526 43.38581 47.25429
[190] 45.89678 44.03067 41.49719 38.11571 33.71564 31.17577 34.28234
[197] 37.01468 39.59387 41.99203 44.08420 46.13070 44.45121 42.23634
[204] 39.42881 36.31677 35.03500 36.83271 39.13933 41.30413 43.25081
[211] 44.93702 47.66080 46.59713 45.22229 43.51962 41.59346 39.87855
[218] 39.24322 40.00470 41.45925 43.03792 44.52292 48.03894 47.18860
[225] 46.14265 44.93578 43.70188 42.72676 42.36466 42.73043 43.59327
[232] 44.65431 45.72293 46.70457 48.45235 47.80447 47.04588 46.22342
[239] 45.44146 44.86167 44.64267 44.83542 45.35133 46.04239 46.78298
[246] 47.49334 48.85417 48.37680 47.84190 47.29018 46.79140 46.43502
[253] 46.29819 46.40628 46.72114 47.16796 49.51659 49.21600 48.87200
[260] 48.50065 48.13220 47.81077 47.58640 47.49911 47.56240 47.75843
[267] 48.04805 48.38680 49.93449 49.74514 49.52523 49.28087 49.02511
[274] 48.77893 48.56966 48.42589 48.36936 48.40761 48.53176 48.72059
[281] 48.94803 49.93873 49.77993 49.60791 49.43242 49.26749 49.13000
[288] 49.03662 48.99959 49.02327 49.10300 49.22696 49.98419 49.86378
[295] 49.74351 49.63263 49.54158 49.48027 49.45580 49.47073 49.52254
[302] 49.60446 50.06079 49.97836 49.90354 49.80223 49.78593 49.79549
[309] 49.82949 49.88397 49.95334 50.03098 50.05347 50.08990 50.13686
[316] 50.21660 50.24109 50.27296 50.35178 50.37349

Tab. 7-6 Valores estimación futura kriging Tomate

La representación gráfica de estos valores se encuentra Fig. 7-9 Predicción solo variable tomate establece la estimación del comportamiento de venta

del producto tomate utilizada como predictor en los lugares cercanos a los que se encuentra información registrada.

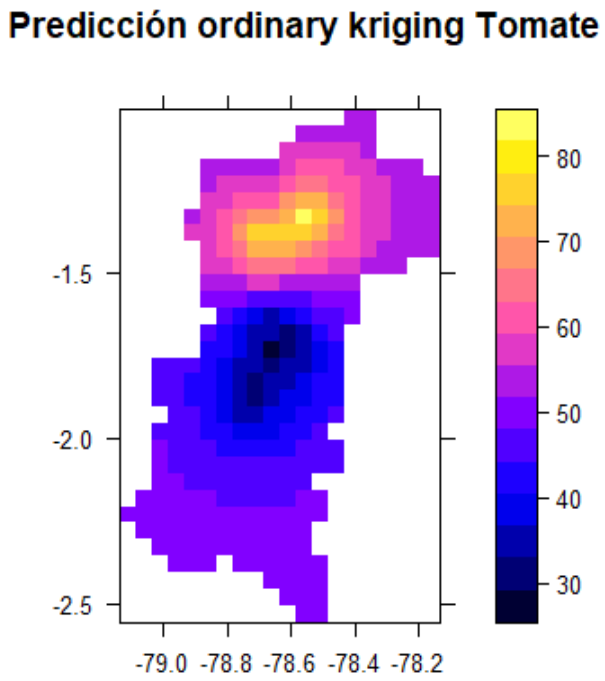


Fig. 7-9 Predicción solo variable tomate

7.3.3 Establecer estimación con predictor multivariable

Para establecer la estimación a futuro multivariable, es necesario crear una función que utilice como predictores a los datos del conjunto asociativo, esta función se genera de la siguiente forma:

```
g <- gstat(NULL, "TOMATE", TOMATE ~ 1, FJespacial)
```

g <- gstat(g, "CEBOLLA BLANCA", CEBOLLA.BLANCA ~ 1, FJespacial)

g <- gstat(g, "BROCOLI", BROCOLI ~ 1, FJespacial)

g <- gstat(g, "TOMATE DE ARBOL", TOMATE.DE.ARBOL ~ 1, FJespacial)

g <- gstat(g, "ZANAHORIA", ZANAHORIA ~ 1, FJespacial)

El modelo (m) utilizado es m <- vgm(2151, "Sph", 0.473) y el modelo de la función g se puede apreciar en

Fig. 7-10 Variograma Multivariable al igual que la relación que se establece con cada variable

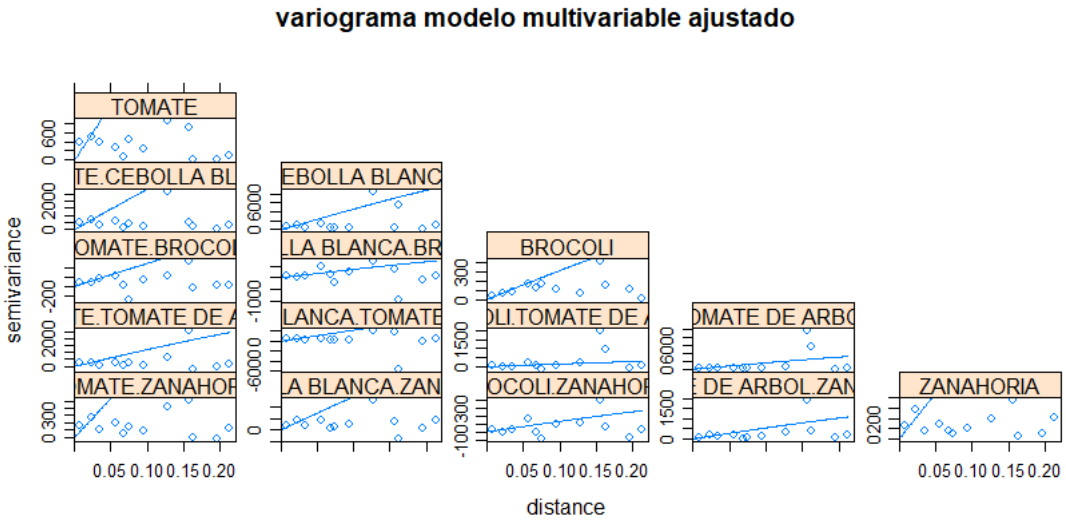


Fig. 7-10 Variograma Multivariable

Los valores de predicción multivariable son los siguientes Tab 7-7:

[1] 49.96975 51.03030 46.88724 49.22357 50.89769 51.91892 52.53923
[8] 47.53265 51.14504 53.89549 55.28937 55.53393 55.11951 45.34511
[15] 45.50144 45.55059 45.70778 47.50848 52.33829 57.63352 60.80101
[22] 61.50519 60.53247 58.76790 56.88508 55.26308 53.15299 49.93229
[29] 50.44310 51.00794 50.37865 50.31940 58.90078 66.82372 69.97792
[36] 69.22751 66.37910 62.94017 59.61801 56.85397 54.29862 51.42137
[43] 54.88449 57.28758 58.78220 58.94023 59.48861 69.34002 78.41567
[50] 79.59656 76.26683 71.86297 66.88760 62.18244 58.27861 55.13817
[57] 51.88998 56.47221 59.51761 62.97394 66.94544 70.79396 74.59439
[64] 81.17059 90.61724 86.20119 79.86513 74.19821 69.10695 64.01657
[71] 59.23411 55.50545 52.04670 58.61336 62.10239 66.94465 73.45537
[78] 81.29227 84.24524 85.58261 87.31972 84.55475 79.08652 73.56874
[85] 68.51651 63.97939 59.41772 55.27411 51.85048 62.55983 67.52247
[92] 74.22331 82.36760 83.90207 82.29209 81.07496 78.56671 74.67263
[99] 70.30665 66.01711 62.00958 58.47174 54.40226 51.35282 60.49148
[106] 64.07388 68.40469 72.32217 73.62644 72.98561 71.98808 70.46659
[113] 68.11689 65.20545 62.05700 58.90423 55.90296 56.28322 57.85445
[120] 59.42068 60.34556 60.35722 60.26057 60.60258 60.87170 60.45335
[127] 59.15975 51.00142 50.59550 49.73841 48.03027 45.95875 46.02664
[134] 48.34711 50.89297 52.47778 52.88559 43.67208 41.09647 37.41010
[141] 31.13499 31.70995 37.17218 42.11735 45.30453 47.00298 40.85183
[148] 37.92766 34.77539 31.37057 44.15198 20.09457 30.55035 36.18021
[155] 39.80263 37.28306 33.72600 30.40828 27.93872 23.84518 29.75629
[162] 30.34762 33.42270 36.77927 46.13995 42.82872 39.16647 35.30420
[169] 31.45512 28.09278 25.88652 25.59490 28.29739 30.78565 33.46482
[176] 36.87759 46.25989 43.20318 39.32862 35.07443 30.87050 27.23366
[183] 25.54231 26.61901 29.28469 32.36125 35.66235 39.36033 46.46740
[190] 43.90620 40.53166 36.60493 32.41098 28.26629 26.29688 28.86419
[197] 31.97758 35.33457 38.75248 42.12474 44.70572 42.00962 38.80232
[204] 35.40516 32.36036 31.27782 32.98472 35.68744 38.68972 41.73323
[211] 44.64303 46.53027 45.37914 43.50104 41.19283 38.83759 37.01037
[218] 36.47825 37.49627 39.50087 41.92638 44.43085 47.04340 45.64108
[225] 44.64297 43.25990 41.85431 40.85620 40.66481 41.40703 42.86474
[232] 44.69605 46.59879 48.32590 47.80704 46.55139 45.24453 44.54869
[239] 43.94328 43.55771 43.63465 44.25745 45.32669 46.63418 47.94579
[246] 49.03676 48.56613 47.63451 46.59987 45.59783 45.05173 44.89349

[253] 45.20891 45.78071 46.50375 47.28224 49.59973 49.20245 48.59114
[260] 47.87535 47.17104 46.59349 46.24277 46.23853 46.59565 47.14934
[267] 47.76747 48.37914 49.69611 49.69611 49.60614 49.30337 48.88074
[274] 48.43642 48.06203 47.83235 47.79330 47.95273 48.27868 48.70615
[281] 49.14865 49.69611 49.69611 49.66798 49.51518 49.30155 49.10143
[288] 48.97298 48.95081 49.04074 49.21889 49.43540 49.69611 49.69611
[295] 49.69611 49.68424 49.63779 49.59712 49.58947 49.61940 49.66819
[302] 49.69611 49.69611 49.69611 49.69611 49.69611 49.69611 49.69611
[309] 49.69611 49.69611 49.69611 49.69611 49.69611 49.69611 49.69611
[316] 49.69611 49.69611 49.69611 49.69611 49.69611

Tab. 7-7 Valores predicción multivariable

La representación gráfica de la estimación multivariable se presenta en Fig 7-11

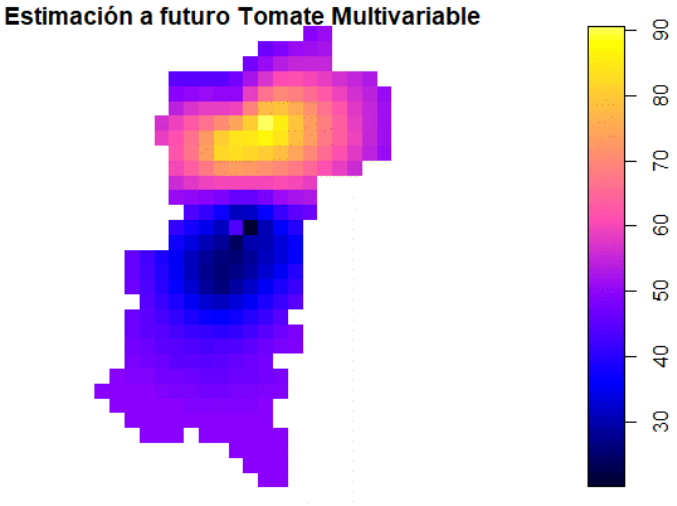


Fig. 7-11 Cokriging Tomate

Establecidos los valores para la predicción de la variable tomate y de la variable tomate en conjunto con las variables resultantes de las reglas de asociación se compara los residuos.

7.3.4. Comparación de los resultados de estimación espacial

Para la medición en los procesos de estimación se trabaja con los valores residuales obtenidos de aplicar a los dos procesos (univariable y multivariable), se establece validación cruzada LOOCV, se calcula el valor estimado en un punto y se compara con el valor real estableciendo el nivel de error cometido.

Los valores correspondientes a los residuos para el caso de un solo predictor se encuentran en la Tab 7-8

```
cvkritomate <- krige.cv(TOMATE ~ 1, FJespacial, v.fit, nfold =14, verbose = FALSE)

summary(cvkritomate$residual)

   Min.        1st Qu.    Median    Mean     3rd Qu.     Max.

-35.41000 -23.71000  -8.65900   0.01276  19.45000  56.42000
```

Tab. 7-8 C.V. Tomate

El valor residual del procesos de validación cruzada dejando fuera un elemento para el proceso multivariable encuentra en Tab. 7-9 C.V. Multivariable Tomate

```
xra <- variogram(g)

g.fit = fit.lmc(xra, g,m)

outtom = gstat.cv(g.fit, nmax = 14, nfold = 14)

summary(outtom$residual)

   Min.      1st Qu.   Median Mean    3rd Qu.  Max.

-34.1200 -11.1400  -3.1510  -0.9209   6.2990  29.2200
```

Tab. 7-9 C.V. Multivariable Tomate

Comprando las muestras se aprecia que: los valores residual obtenidos de aplicar la referencia cruzada a la serie del tomate multivariable son menores, lo que indica que las estimaciones a futuro tienen una disminución del error si se utilizan las reglas de asociación para predecir las estimaciones de comportamiento en lugares que no se conoce la medida real esto se aprecia en la fig 7-12

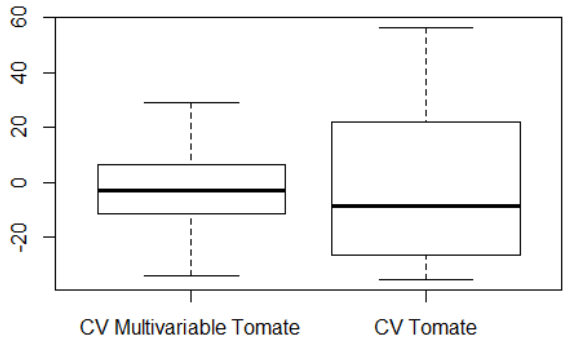


Fig. 7-12 Resultado Residuals Validación Cruzada

7.4 Asociación en Fusión de Datos Externos

Con el propósito de validar los resultados preliminares obtenidos en las secciones anteriores, se aplica la metodología propuesta a un nuevo modelo que presenta dos variables externas al conjunto transaccional analizado.

Estas dos variables son: La primera variable proviene de la información de piso climático de la zona donde se ubican las provincias de Tungurahua y

Chimborazo, conformada por cuatro atributos: precipitación ambiental, temperatura, humedad y evaporación.

La segunda variable es la población existente en cada cantón. La información detallada se encuentra en Tab 7-10 que, contiene los datos de Nombre de la feria y su ubicación geográfica, el valor de comercialización de los productos asociados, datos de población y características del piso climático.

Feria	BROCOLI	CEBOLLA BLANCA	TOMATE	TOMATE DE ARBOL	ZANAHORIA	x	y	Población	precip	temp	hum	evap
Colta	6	60	26	30	30	-78.7238	-1.888	44.701	4	18	-17	64
Tisaleo	37.9	36	88.25	185	47.5	-78.6923	-1.41	10.565	4	18	-17	64
RIOBAMBA1	8	9.6	18	0	0	-78.6737	-1.662	193.32	6.5	18	-17	64
RIOBAMBA2	26	42.5	60	15	17.5	-78.673	-1.661	193.32	6.5	18	-17	64
RIOBAMBA3	0	10	8	8	5	-78.6687	-1.66	193.32	6.5	18	-17	64
RIOBAMBA4	2.5	74	60	35	29.5	-78.6588	-1.676	193.32	6.5	18	-17	64
Cevallos	9	9.6	51	45	20	-78.6566	-1.257	6.873	6.5	18	-17	64
RIOBAMBA5	7.2	50	30	35	20	-78.6558	-1.687	193.32	6.5	18	-17	64
RIOBAMBA6	6	65	42	27	29	-78.6538	-1.676	193.32	6.5	18	-17	64
RIOBAMBA7	2	0	10	0	5	-78.65	-1.678	193.32	6.5	18	-17	64
RIOBAMBA8	6	26.5	33	0	36	-78.6497	-1.675	193.32	6.5	18	-17	64
RIOBAMBA9	2.5	0	6.9	0	3	-78.6463	-1.689	193.32	6.5	18	-17	64
CHAMBO	21	50	30	20	20	-78.6077	-1.73	10.541	6.5	18	-17	64
Pillaro	20.1	141.5	92	78	40	-78.551	-1.328	34.925	6.5	18	-17	64

Tab. 7-10 Datos fusionados

La generación de un nuevo modelo predictivo inicia verificando la existencia de correlación entre las variables externas con el producto tomate.

La correlación entre las variables climáticas solo se nuestra significativa para el caso de la precipitación, tabla 7-11.

	TOMATE	Precip	Temp	Hum	Evap
TOMATE	1.0000000	-0.2681709	NA	NA	NA
Precip	-0.2681709	1.0000000	NA	NA	NA
Temp	NA	NA	1	NA	NA
Hum	NA	NA	NA	1	NA
Evap	NA	NA	NA	NA	1

Tab. 7-11 Relación variable heterogénea

Para obtener una estimación a futuro se utiliza el siguiente proceso:

- Crear una nueva función constituida con las variables tomate y precipitación atmosférica
- Utilizando el modelo de variograma m
- Realizar la predicción multi variable (Tomate/Precipitación),
- Utilizando validación cruzada encontrar los valores residuales

Los resultados se muestran en la fig 7-13.

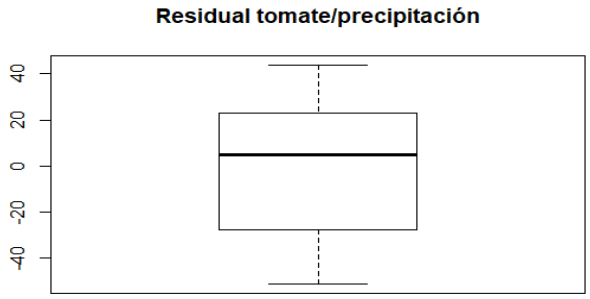


Fig. 7-13 Residual Precipitación

De similar manera se procede para la variable Tomate Población, los valores residuales resultantes se encuentran en Fig 7-14

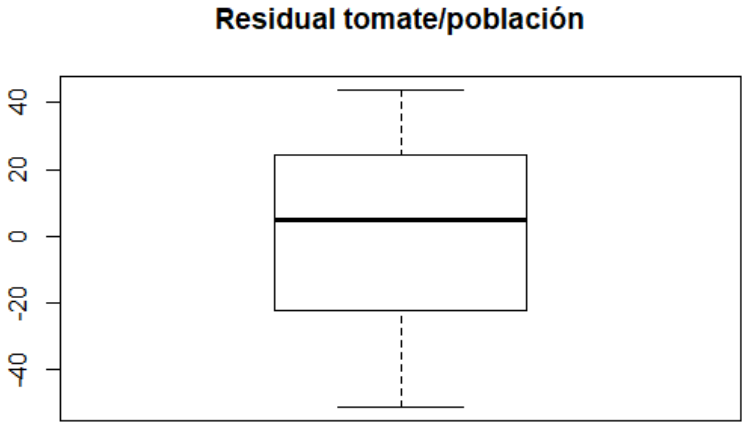


Fig. 7-14 Residual Población

7.5 Comparación resultados: Fusión de datos.

En las secciones anteriores se ha trabajado sobre cuatro modelos que utilizan distintos predictores para obtener la interpolación pronostica del tomate.

La línea base de comparación de las estimaciones pronosticas se establece con los valores obtenidos de aplicar un proceso IDW, cada algoritmo utiliza como datos de entrada los detallados en la Tab 7-10, el proceso realizado es el siguiente:

- Obtener el conjunto de interpolación pronostica
- Realizar la validación cruzada, para obtener los valores residuales
- Comparar los valores de error obtenidos

Los nombres identificativos de cada experimento y los procesos utilizados se encuentran en la Tab 7-12.

Nombre Experimento	Predictor	Método Utilizado
IDW	Tomate	IDW
OT	Tomate	Kriging
TPop	Tomate/Población	Co-kriging
Prec	Tomate/Precipitación	Co-kriging

Nombre Experimento	Predictor	Método Utilizado
TAR	Conjunto asociativo A	Co-kriging
TALL	Conjunto Asociativo, Población, Precipitación	Co-kriging

Tab. 7-12 Procesos Utilizados

Los resultados residuales de los seis modelos son presentados en la Tab 7-13, ordenados en base al proceso aplicado.

IDW	**-40,16**	**-21,95**	**1,6690**	**2,259**	**23,6100**	**43,760**
OT	**-51,23**	**-21,1400**	**4,8030**	**-0,01941**	**23,5700**	**43,9400**
Tpop	**-51,23**	-21,1500	4,8070	-0,09446	23,3600	43,9400
Prec	**-51,22**	-26,4700	4,8470	-0,7310	23,1000	43,9400
TAR	**-34,12**	-11,1400	-3,1510	-0,9209	6,2990	29,2200
TALL	**-36,83**	-9,4600	1,3220	-0,1151	5,6000	34,6000

Observando la tabla 7-14, se aprecia que los valores residuales tienen diferencias dependiendo del proceso que se utiliza, al comparar la columna del tercer cuadrante con los valores del primer cuadrante se encuentra:

- El proceso IDW tiene valores residuales menores sobre los resultados obtenidos de los procesos OT, Tpop, Prec.
- TAR presenta valores mas bajos que los encontrados utilizando la línea base (IDW)
- TALL indica valores menores que IDW, OT, Tpop y Prec

	Q3-Q1	Max-Min
IDW	63,770	83,920
OT	74,8000	95,1700
Tpop	74,5900	95,1700
Prec	74,3200	95,1600
TAR	40,4190	63,3400
TALL	42,4300	71,4300

Tab. 7-13Diferencias Residuales

Los modelos OT, Tpop y Prec, presentan un comportamiento muy similar al IDW por lo que se puede considerar la equivalencia de resultados utilizando cualquiera de estos métodos

La utilización del conjunto de productos asociados TAR, presenta una disminución de los valores residuales, con respecto al modelo base IDW.

El proceso para estimar comportamiento utilizando el proceso TALL, presenta valores residuales menores que utilizar los procesos IDW, OT,Tpop y Prec.

La Fig. 7-15 presenta de forma gráfica los valores residuales obtenidos de cada proceso.

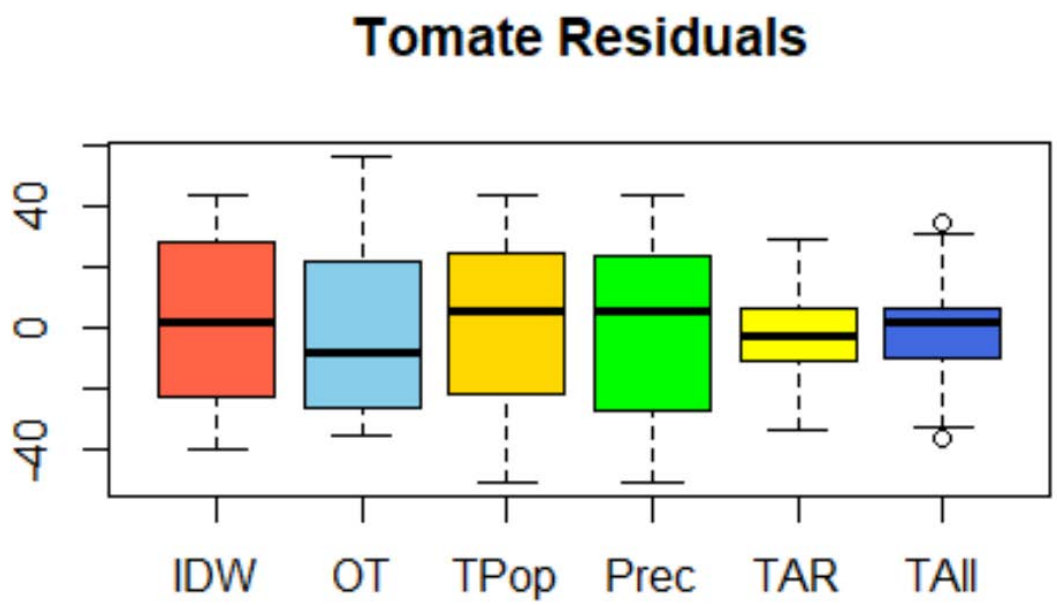


Fig. 7-15 Valor Residual todas las series

El menor valor residual se establece utilizando como predictor el conjunto asociativo,

8. Discusión y Conclusiones

8.1 Impacto de resultados obtenidos

El trabajo realizado se ha centrado en el campo de la minería de datos y específicamente en buscar modelos que permitan extraer patrones para predecir comportamientos, utilizando técnicas descriptivas y predictivas. [72].

El conjunto de datos está constituido por treinta productos pertenecientes al grupo de las hortalizas y verduras comercializados por asociacion de productores artesanales agrícolas en circuitos alternativos de comercialización (CIALCO), ubicados en diferentes cantones de las provincias de Tungurahua y Chimborazo en el centro sur de la región andina en el país Ecuador.

Esta información ha sido depurada, especialmente se ha homologado para cada producto:

- nombre
- unidades de medida y
- valor de venta.

Sobre esta información, se desarrolla tres propuestas para la búsqueda de patrones de comportamiento:

- La primera tarea de tipo descriptivo busca la inferencia de reglas de asociación.
- Las dos restantes de tipo predictivo realizan tareas de extrapolación pronostica utilizando series de tiempo e interpolación pronostica en base a la ubicación espacial.

La obtención de patrones en cada uno de los procesos se realiza basado en características de los datos, así para generar las reglas de asociación es necesario una estructura de archivo que identifique si un producto es parte de una transacción, lo que se consigue discretizando la información de venta de productos.

La parte predictiva utiliza modelos de regresión que son aplicados dependiendo de la dimensionalidad buscada: tiempo o espacio.

La estimación a futuro utilizando una serie de tiempo, se realiza en base a los atributos de tiempo y valor de venta.

La búsqueda de patrones de comportamiento espacial se basa en la ubicación geográfica de los lugares de comercialización, la especificación de coordenadas geográficas es incorporada en los archivos a utilizar.

La tesis propuesta es:

Establecer si: patrones de comportamiento, utilizando atributos nominales (asociativos) en una fusión de datos, tienen incidencia sobre los patrones de comportamiento correlacionales (predictivos).

En base a esta premisa se plantea la siguiente metodología.

1.- Generar el conjunto de productos asociados.

2.- Trasformar archivo para incluir atributos requeridos por el proceso.

2.- Búsqueda de patrones de comportamiento utilizando un predictor.

3.- Fusión de datos y búsqueda de patrones utilizando el conjunto asociativo.

4.- Medición de resultados obtenidos.

La metodología se aplica de manera independiente a cada proceso de estimación realizado: extrapolación utilizando series de tiempo e interpolación utilizando la ubicación geográfica.

Los procesos de búsqueda de reglas de asociación y series de tiempo se han realizado utilizando los algoritmos incluidos en la herramienta Weka toda la parte de geoestadística utiliza el lenguaje R y su interfaz de desarrollo R Studio.

La aplicación de la metodología comienza generando el conjunto de productos asociados como se detalla en la siguiente sección.

8.1.1.- Conjunto de productos asociados

Al conjunto discretizado de 549 transacciones y treinta productos, se aplica los dos algoritmos "Apriori" y "Fp-growth", con índices de soporte 0,4 y confianza 0,8.

De las cuatro mejores reglas de asociación resultantes se obtienen los productos que mas se relacionan en los procesos de comercialización, constituyendo el conjunto de productos asociados A= {Tomate Riñón, Cebolla Blanca, Tomate de Árbol, Zanahoria, Brócoli}

Analizando la composición de las reglas de asociación los productos Cebolla Blanca, Tomate de Árbol, Zanahoria y Brócoli aparecen en el antecedente de la regla y solo el tomate en el consecuente por esta razón todos los procesos de minería de datos de tipo predictivo realizados tienen relación con la búsqueda de patrones de comportamiento de la variable tomate.

Generado el conjunto de productos asociados, este se utiliza para encontrar la estimación pronostica en la dimensión tiempo y espacio de manera independiente.

Utilizando el atributo fecha de venta se realiza la estimación a futuro, utilizando series de tiempo, como se detalla en la siguiente sección

8.1.2.- Extrapolación Pronostica: Series de Tiempo

8.1.2.1 Búsqueda de patrones de comportamiento utilizando un predictor.

El archivo utilizado para aplicación del proceso de extrapolación pronostica, se basa en el atributo fecha de venta y valor de venta agregado, tiene registrada la información de cuarenta y tres semanas del año 2014.

Utilizando estimadores núcleos con el algoritmo SMOReg, el modelo de datos contiene un polinomio de grado tres que consta de las siguientes variables: fecha recalculada (remapped), doce variables de retardo para el tomate, la fecha recalculada elevada al cuadrado y la fecha recalculada elevada al cubo además la fecha recalculada por cada una de las variables de retardo.

La información de las cuarenta y tres semanas constituye el conjunto de entrenamiento y el error de la estimación a futuro en base al modelo generado para una semana adelante es la registrada en la tabla 8-1.

Estimación series de tiempo: un predictor	Variable Tomate
Mean absolute error	18.3002
Root mean squared error	29.4848

Tab. 8-1 Estimación error Series de Tiempo: Un predictor

8.1.2.2.- Búsqueda de patrones de comportamiento, Predictor conjunto asociativo.

El conjunto multivariable A sirve como predictor para establecer el modelo de datos y generar la extrapolación pronostica utilizando series de tiempo, las variables adicionales que aparecen en este modelo son: BROCOLI, CEBOLLA BLANCA, TOMATE DE ARBOL, ZANAHORIA.

La evaluación realizada sobre el mismo conjunto de entrenamiento de la sección anterior entrega los resultados siguientes: Tab 8-2:

Estimación series de tiempo: conjunto asociativo	Variable Tomate
Mean absolute error	16.4369
Root mean squared error	27.1553

Tab. 8-2 Estimación Series de Tiempo: predictor conjunto asociativo

8.1.3.- Interpolación Pronostica: Ubicación geografica

La segunda dimensión para realizar la estimación pronostica es la ubicación geográfica.

El Ecuador se encuentra atravesado por el paralelo 0 o Línea Ecuatorial ubicado en las coordenadas geográficas 1.8312S, 78.1834W, sobre el archivo original de información se establece las coordenadas geográficas a cada una de las ferias y la conversión del archivo a un objeto de tipo espacial.

El área ocupada por las provincias de Tungurahua y Chimborazo sirve como base para establecer la superficie para realizar la interpolación pronostica que tiene una dimensión de 21.692 Km^2.

Sobre esta superficie se aplica la metodología propuesta encontrando los siguientes resultados.

8.1.3.1 Búsqueda de patrones de comportamiento utilizando un predictor

El primer caso de interpolación pronostica utilizando coordenadas geográficas se realiza utilizando una variable como predictor, que es la variable tomate.

El procedimiento utilizado es kriging ordinario y validación cruzada con el método LOOCV, para establecer el valor de los residuos, encontrando los siguientes resultados Tab 8-3

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Kriging Ordinario	-51,23	-21,1400	4,8030	-0,01941	23,5700	43,9400

Tab. 8-3 Kriging Ordinario (Residuals)

8.1.3.2 Búsqueda de patrones utilizando el conjunto asociativo

La segunda parte de la metodología establece encontrar la interpolación pronostica utilizando como predictor el conjunto de productos asociados.

Aplicando el proceso Co-kriging y el método LOOCV, se genera los siguientes resultados Tab 8-4

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Q3-Q1	Max-Min
TAR	-34,12	-11,1400	-3,1510	-0,9209	6,2990	29,2200	40,4190	63,3400

Tab. 8-4 Co kriging (Residuals)

8.1.4 Línea base y datos externos

Aplicando el método IDW, se establece la línea base de comparación con los valores obtenidos de aplicar la metodología propuesta.

Los valores resultantes al utilizar la distancia inversa ponderada son Tab 8-5

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
IDW Tom	-40,16	-21,95	1,6690	2,259	23,6100	43,760

Tab. 8-5 Estimación Pronostica IDW: Tomate (Residuals)

Con el fin de validar los resultados de la metodología propuesta se adicionan dos variables externas al proceso de comercialización, pero de incidencia en el área de interpolación pronostica: Población y Precipitación, los valores residuales obtenidos son Tab 8-6:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Población/Tom	-51,23	-21,1500	4,8070	-0,09446	23,3600	43,9400
Prec/Tom	-51,22	-26,4700	4,8470	-0,7310	23,1000	43,9400

Tab. 8-6 Estimación pronostica: Población, Precipitación (Residuals)

8.1.5 Fusión de datos

Con el fin de establecer el nivel de impacto que tiene la utilización de el conjunto asociativo frente a los modelos que utilizan variables externas en la

interpolación pronostica, se realiza una fusión de modelos, en uno que considera todas las variables utilizadas: Población, Precipitación y Conjunto asociativo.

El resultado de los residuos de este nuevo modelo se presenta en la Tab 8-7

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Fusion de datos (TALL)	-36,83	-9,4600	1,3220	-0,1151	5,6000	34,6000

Tab. 8-7 Estimación Pronostica Fusión de Datos (Residuals)

En la Tab 8-8, se puede apreciar un consolidado de todos los valores residuales obtenidos

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
IDW	-40,16	-21,95	1,6690	2,259	23,6100	43,760
OT	-51,23	-21,1400	4,8030	-0,01941	23,5700	43,9400
Tpop	-51,23	-21,1500	4,8070	-0,09446	23,3600	43,9400
Prec	-51,22	-26,4700	4,8470	-0,7310	23,1000	43,9400
TAR	-34,12	-11,1400	-3,1510	-0,9209	6,2990	29,2200
TALL	-36,83	-9,4600	1,3220	-0,1151	5,6000	34,6000

Tab. 8-8 Consolidado Valores Residuals

8.1.4.- Resultados obtenidos

La presentación de resultados se la divide en dos partes: pronostico utilizando series de tiempo y pronostico utilizando geoestadística.

8.1.4.1 Pronostico series de tiempo

Los valores para considerar utilizando series de tiempo son las estimaciones de error: Mean absolute error (Mae) y el cálculo del Root mean squared error (Rmse).

La relación a considerar es el valor obtenido para la estimación a futuro utilizando el conjunto multivariable A dividido para el valor de error obtenido para la estimación de la variable única tomate.

Para el Mae, se encuentra una mejora del 10%, el valor del Rmse muestra una mejora de alrededor del 8% cuando se utilizan el conjunto A de variables asociadas Tab 8-9.

Valor error medido	Val mult/val univar	% de mejora
Mean absolute error	0.898181441	10.18185594
Root mean squared error	0.92099319	7.900681029

Tab. 8-9 Porcentajes disminución error

8.1.4.2 Pronostico geoestadística

Los resultados del error medio cuadratico encontrado para cada proceso de predicción espacial se presenta en la tabla 8-10

Modelo	Valor RMSE
IDW	27.804
Kriging Tomate	29.44047
Cokriging Tomate/Población(Tpop)	29.37169
Cokriging Tomate/Precipitación(Prec)	29.3256
Cokriging Tomate/Reg. Asociación(TAR)	16.55579
Cokriging Tomate/Toda Variable(TALL)	19.41976

Tab. 8-10 Error medio Observado/Predicción

El menor valor de error RMSE presentan los procesos en los que se utiliza el conjunto multivariable generado por la inferencia de reglas de asociación, TAR, TALL.

8.2 Conclusiones

A la información inicial registrada de comercialización se realizó un proceso de preparación de los datos para mejorar su calidad y aplicar procedimientos de minería de datos.

Sobre el subconjunto con la información de hortalizas y legumbres se realizan las transformaciones para aplicar técnicas de minera de datos descriptiva y predictiva.

Dentro del tema descriptivo se aplicaron los algoritmos Apriori y Fp-Tree para generar reglas de asociación validas utilizando los índices de Soporte y Confianza, el conjunto de productos asociados en la comercialización es {Tomate, cebolla blanca, Tomate de árbol, zanahoria, brócoli}.

En la parte predictiva se maneja dos dimensiones tiempo y espacio.

Con la dimensión temporal se aplica procesos de regresión no lineal, utilizando estimadores kernel y el algoritmo SMOReg que calcula estimaciones a futuro utilizando un solo predictor para la variable tomate.

De la misma manera se aplicó un proceso de predicción multivariable, que se constituye por el conjunto de productos asociados y se realiza la estimación a futuro.

Al comparar las dos estimaciones realizadas, se encuentra que el conjunto multivariable como un predictor genera un error mas pequeño y por lo tanto las predicciones a futuro mejoran al utilizar el conjunto multivariable.

En cuanto a la dimensión espacial Fig 8-1, se realiza el mismo procedimiento, calcular valores estimados utilizando método de Kriging para un solo

predictor y en una segunda parte medida utilizando el conjunto multivariable y aplicando el proceso de Cokriging.

Tanto en el caso temporal como en el espcial se verifica que cuando se utiliza un conjunto multivariable existe una dsiminución de los valores de error.

En consecuencia, se ha probado la hipótesis que utilizar el conjunto predictor multivariable basado en los resultados de reglas de asociación permite disminuir significativamente los niveles de error tanto si se utiliza series temporales como si se utiliza predicción espacial.

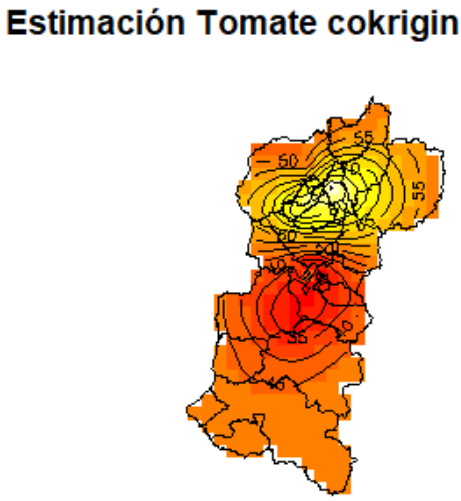


Fig. 8-1 Estimación variable Tomate

8.3 Trabajo Futuro

En la actualidad el Ministerio de Agricultura y Ganadería del Ecuador posee una base de datos de tipo transaccional donde registra todos los circuitos alternativos de comercialización, su ubicación y producción.

Se debe considerar la implementación de estimaciones de series de tiempo que puedan ser alimentadas directamente por el sistema transaccional tan pronto se genere un dato para establecer las diferencias entre valores predichos y valores reales lo que permitirá ir mejorando los procesos de predicción un paso adelante.

De igual manera con la dimensión espacial es necesario encontrar la manera de alimentar la información tan pronto se produce y alimentar al sistema de predicción espacial para mejorar sus estimaciones

8.4 Publicaciones realizadas

Hay distintas publicaciones realizadas que se refieren al tema tratado en esta tesis, en la tab 8-11 se presenta el detalle

Evento	Titulo	Autor
International Conference on Practical Applications of Agents and Multi-Agent Systems 2016	CIALCO: Alternative marketing channels [116]	Washington R Padilla, H Jesús García
2018 21st International Conference on Information Fusion (FUSION)	Model Learning and Spatial Data Fusion for Predicting Sales in Local Agricultural Markets[117]	Washington R Padilla ; García H. Jesús ; Jose M. Molina
International Conference on Practical Applications of Agents and Multi-Agent Systems 2018	Information Fusion and Machine Learning in Spatial Prediction for Local Agricultural Markets [118]	Washington R. Padilla, Jesús García, José M. Molina

Evento	Titulo	Autor
International Conference on Hybrid Artificial Intelligence Systems 2018	Improving forecasting using information fusion in local agricultural markets[119]	Washington R Padilla, Jesús García, José M Molina
Revista Sensors 2019	Knowledge Extraction and Improved Data Fusion for Sales Prediction in Local Agricultural Markets[120]	WR Padilla, J García, JM Molina
Revista Computational intelligence and neuroscience 2018	Data Association Methodology to Improve Spatial Predictions in Alternative Marketing Circuits in Ecuador[121]	WR Padilla, J García

Tab. 8-11 Publicaciones Realizadas

8.5 Proyecto Relacionado.

La investigación presentada en este documento se la realiza como parte del convenio para fortalecer el desarrollo de circuitos alternativos de comercialización firmado entre el Ministerio de Agricultura y Ganadería y la Universidad Politécnica Salesiana en el Ecuador

Los datos utilizados son proporcionados por la Coordinación General de Redes de Comercialización, en la cláusula 3.1.f se permite a la Universidad Politécnica Salesiana, la publicación de los resultados.

Por parte de la Universidad Carlos III se ha contado con el apoyo de los proyectos del Ministerio de Economía y Competitividad TEC2017-88048-C2-2-R, TEC2014-57022-C2-2-R, para la preparación y difusión de los resultados obtenidos en esta investigación.

8.6 Bibliografía

[1] «Post 2015 process .:. Sustainable Development Knowledge Platform». [En línea]. Disponible en: https://sustainabledevelopment.un.org/post2015. [Accedido: 06-ene-2019].

[2] S. Fritz *et al.*, «A comparison of global agricultural monitoring systems and current gaps», *Agric. Syst.*, vol. 168, pp. 258-272, ene. 2019.

[3] «Agricultura familiar y desarrollo territorial rural en América Latina y el Caribe | Oficina Regional de la FAO para América Latina y el Caribe | Organización de las Naciones Unidas para la Alimentación y la Agricultura». [En línea]. Disponible en: http://www.fao.org/americas/prioridades/agricultura-familiar/es/. [Accedido: 06-ene-2019].

[4] Michel Leporati, Salomón Salcedo, Byron Jara, Verónica Boero, y Mariana Muñoz, «La agricultura familiar en cifras», en *Recomendaciones de Política*, FAO 2014., p. 486.

[5] «Clasificación de los sistemas de producción agrícola.», *Parlamento Científico de Jóvenes*, 29-oct-2014. .

[6] Raúl Contreras, Ekaterina Krivonos, y Luis Sáez, «Mercados locales y ferias libres: El caso de Chile», en *Recomendaciones de Política*, FAO 2014, p. 370.

[7] «El Ministerio – Ministerio de Agricultura y Ganadería». [En línea]. Disponible en: https://www.agricultura.gob.ec/el-ministerio/. [Accedido: 06-ene-2019].

[8] S. Salcedo, Ana Paula De La O, y Lya Guzmán, «El concepto de agricultura familiar en América Latina y el Caribe», en *Recomendaciones de Política*, FAO 2014., p. 486.

[9] T. Gemtos, A. Markinos, y T. Nassiou, «Cotton lint quality spatial variability and correlation with soil properties and yield», ene. 2019.

[10] O. Wendroth, P. Jürschik, A. Giebel, y D. R. Nielsen, «Spatial Statistical Analysis of On-Site Crop Yield and Soil Observations for Site-Specific Management», *Precis. Agric.*, vol. acsesspublicati, n.º precisionagric4a, pp. 159-170, 1999.

[11] E. I. Papageorgiou, A. T. Markinos, y T. A. Gemtos, «Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application», *Appl. Soft Comput.*, vol. 11, n.º 4, pp. 3643-3657, jun. 2011.

[12] E. I. Papageorgiou, K. D. Aggelopoulou, T. A. Gemtos, y G. D. Nanos, «Yield prediction in apples using Fuzzy Cognitive Map learning approach», *Comput. Electron. Agric.*, vol. 91, pp. 19-29, feb. 2013.

[13] J. N. Brown, Z. Hochman, D. Holzworth, y H. Horan, «Seasonal climate forecasts provide more definitive and accurate crop yield predictions», *Agric. For. Meteorol.*, vol. 260-261, pp. 247-254, oct. 2018.

[14] J. Hernandez-Orallo, *Introducción a la Minería de Datos*, 2004.ª ed. Madrid: Pearson Prentice Hall, 2004.

[15] Z. Chun-Sheng y L. Yan, «Extension of local association rules mining algorithm based on apriori algorithm», en *2014 IEEE 5th International Conference on Software Engineering and Service Science*, Beijing, China, 2014, pp. 340-343.

[16] R. Sumithra y S. Paul, «Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery», en *2010 Second International conference on Computing, Communication and Networking Technologies*, Karur, India, 2010, pp. 1-5.

[17] «An Efficient Frequent Patterns Mining Algorithm Based on Apriori Algorithm and the FP-Tree Structure - IEEE Conference Publication». [En línea]. Disponible en: https://bibliotecas.ups.edu.ec:2095/document/4682180. [Accedido: 27-dic-2018].

[18] B. Li, X. Ding, W. Han, L. Cheng, y M. Yu, «The Number of Passengers Forecast in Emergency Situation of Airport Based on Time Series Analysis», en *2009 Second International Symposium on Knowledge Acquisition and Modeling*, Wuhan, China, 2009, pp. 369-371.

[19] V. Prakaulya, R. Sharma, U. Singh, y R. Itare, «Railway passenger forecasting using time series decomposition model», en *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, 2017, pp. 554-558.

[20] T.-M. Choi, C.-L. Hui, y Y. Yu, «Intelligent time series fast forecasting for fashion sales: A research agenda», en *2011 International Conference on Machine Learning and Cybernetics*, Guilin, China, 2011, pp. 1010-1014.

[21] Y. Shang, J. Zhang, Y. Wang, J. Hou, y K. Ma, «Analysis of heavy metal pollution based on kriging interpolation in the suburb of Yidu», en *Proceeding of the 11th World Congress on Intelligent Control and Automation*, Shenyang, China, 2014, pp. 120-123.

[22] Suprayitno y J.-C. Yu, «Evolutionary algorithm using progressive Kriging model and dynamic reliable region for expensive optimization problems», en *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary, 2016, pp. 004383-004388.

[23] G. L. Wei Xue, «Forest Fires Regional Distribution and Numerical Simulation Based on Universal Kriging Algorithm - IEEE Conference Publication». [En línea]. Disponible en: https://bibliotecas.ups.edu.ec:2095/document/5360648. [Accedido: 27-dic-2018].

[24] J. Ma, X. Yu, G. Chen, J. Wang, y Y. Pu, «Research on urban accessibility distribution areal model by Floyd algorithm and Kriging interpolation», en *2010 18th International Conference on Geoinformatics*, Beijing, China, 2010, pp. 1-4.

[25] T. Schlüter y S. Conrad, «About the analysis of time series with temporal association rule mining», en *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2011, pp. 325-332.

[26] G. N. Pradhan y B. Prabhakaran, «Association Rule Mining in Multiple, Multidimensional Time Series Medical Data», *J. Healthc. Inform. Res.*, vol. 1, n.º 1, pp. 92-118, jun. 2017.

[27] J. Li, G. Xia, y X. Shi, «Association Rules Mining from Time Series Based on Rough Set», en *Sixth International Conference on Intelligent Systems Design and Applications*, 2006, vol. 1, pp. 509-516.

[28] W. Dong *et al.*, «Discovery of generalized spatial association rules», en *Proceedings of 2012 IEEE International Conference on Service Operations and Logistics, and Informatics*, Suzhou, China, 2012, pp. 60-65.

[29] A. Kondaveeti, H. Liu, G. Runger, y J. Rowe, «Extracting geographic knowledge from sensor intervention data using spatial association rules», en *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, Fuzhou, China, 2011, pp. 127-130.

[30] R. V. Mane y V. R. Ghorpade, «Predicting student admission decisions by association rule mining with pattern growth approach», en *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, 2016, pp. 202-207.

[31] «DIA_PinhoLucasJ_Métodosdeclasificación.pdf». .

[32] «Agrawal et al. - Mining Association Rules between Sets of Items in .pdf». .

[33] «La contribución de las reglas de asociación a la minería de datos | De Moya Amaris | Tecnura». [En línea]. Disponible en: https://revistas.udistrital.edu.co/ojs/index.php/Tecnura/article/view/6175. [Accedido: 15-oct-2018].

[34] S. D. Patil, R. R. Deshmukh, y D. K. Kirange, «Adaptive Apriori Algorithm for frequent itemset mining», en *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, 2016, pp. 7-13.

[35] J. H. J. Pei, «Mining Frequent Patterns without Candidate Generation», p. 12.

[36] J. Han, J. Pei, y Y. Yin, «Mining Frequent Patterns Without Candidate Generation», en *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2000, pp. 1–12.

[37] S. Sidhu, U. Kumar Meena, A. Nawani, H. Gupta, y N. Thakur, «FP Growth Algorithm Implementation», *Int. J. Comput. Appl.*, vol. 93, pp. 6-10, may 2014.

[38] R. Agrawal, T. Imielinski, y A. Swami, «Database mining: a performance perspective», *IEEE Trans. Knowl. Data Eng.*, vol. 5, n.º 6, pp. 914-925, dic. 1993.

[39] W. B. Zulfikar, A. Wahana, W. Uriawan, y N. Lukman, «Implementation of association rules with apriori algorithm for increasing the quality of promotion», en *2016 4th International Conference on Cyber and IT Service Management*, 2016, pp. 1-5.

[40] S. Asadifar y M. Kahani, «Semantic association rule mining: A new approach for stock market prediction», en *2017 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*, 2017, pp. 106-111.

[41] J. Yang, H. Huang, y X. Jin, «Mining Web Access Sequence with Improved Apriori Algorithm», en *22017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, Guangzhou, China, 2017, pp. 780-784.

[42] A. Das, «A novel association rule mining mechanism in wireless sensor networks», en *2012 3rd National Conference on Emerging Trends and Applications in Computer Science*, Shillong, India, 2012, pp. 274-276.

[43] A. Gosain y M. Bhugra, «A comprehensive survey of association rules on quantitative data in data mining», en *2013 IEEE Conference on Information Communication Technologies*, 2013, pp. 1003-1008.

[44] R. Agrawal y R. Srikant, «Mining sequential patterns», en *Proceedings of the Eleventh International Conference on Data Engineering*, 1995, pp. 3-14.

[45] C.-C. Chang, Y.-C. Li, y J.-S. Lee, «An efficient algorithm for incremental mining of association rules», en *15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications (RIDE-SDMA'05)*, 2005, pp. 3-10.

[46] T. Abraham y J. F. Roddick, «Incremental meta-mining from large temporal data sets», en *Advances in Database Technologies*, Springer, 1999, pp. 41–54.

[47] M. Spiliopoulou y J. F. Roddick, *Higher Order Mining: Modelling and Mining the Results of Knowledge Discovery*. 2000.

[48] C. Ravi y N. Khare, «EO-ARM: An efficient and optimized k-map based positive-negative association rule mining technique», en *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]*, 2014, pp. 1723-1727.

[49] S. Naredi y R. A. Deshmukh, «Improved extraction of quantitative rules using Best M Positive Negative Association Rules Algorithm», en *2015 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2015, pp. 1-5.

[50] M. M. Hasan y S. Zaman Mishu, «An Adaptive Method for Mining Frequent Itemsets Based on Apriori And FP Growth Algorithm», en *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, Rajshahi, 2018, pp. 1-4.

[51] K. Dharmaraajan y M. A. Dorairangaswamy, «Analysis of FP-growth and Apriori algorithms on pattern discovery from weblog data», en *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Coimbatore, India, 2016, pp. 170-174.

[52] L. Yongmei y G. Yong, «Application in Market Basket Research Based on FP-Growth Algorithm», en *2009 WRI World Congress on Computer Science and Information Engineering*, Los Angeles, California USA, 2009, pp. 112-115.

[53] J. Hipp, U. Güntzer, y G. Nakhaeizadeh, «Algorithms for association rule mining --- a general survey and comparison», *ACM SIGKDD Explor. Newsl.*, vol. 2, n.º 1, pp. 58-64, jun. 2000.

[54] Richard Darlington, «A regression approach to time series analylsis». [En línea]. Disponible en: http://node101.psych.cornell.edu/Darlington/series/series0.htm. [Accedido: 12-dic-2018].

[55] V. N. Vapnik, «An overview of statistical learning theory», *IEEE Trans. Neural Netw.*, vol. 10, n.º 5, pp. 988-999, sep. 1999.

[56] J. Han, J. Pei, y M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[57] A. J. Smola y B. Schölkopf, «A tutorial on support vector regression», *Stat. Comput.*, vol. 14, n.º 3, pp. 199-222, ago. 2004.

[58] G. Aburiyana y M. E. El-Hawary, «An overview of forecasting techniques for load, wind and solar powers», en *2017 IEEE Electrical Power and Energy Conference (EPEC)*, Saskatoon, SK, 2017, pp. 1-7.

[59] Guo Jianhua, «Performance Measures for Short Term Traffic Condition Forecasting Algoithms», en *2013 Fifth International Conference on Measuring Technology and Mechatronics Automation*, Hong Kong, 2013, pp. 796-798.

[60] «Load Forecasting Using Statistical Time Series Model in a Medium Voltage Distribution Network - IEEE Conference Publication». [En línea]. Disponible en: https://bibliotecas.ups.edu.ec:2095/document/8592891. [Accedido: 22-feb-2019].

[61] M. Li y D. Quan, «A time series approach for risk forecasting of individual stocks in the new three board market», en *2017 4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS)*, Kyoto, Japan, 2017, pp. 1-5.

[62] T. Schlüter y S. Conrad, «About the analysis of time series with temporal association rule mining», en *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2011, pp. 325-332.

[63] G. N. Pradhan y B. Prabhakaran, «Association Rule Mining in Multiple, Multidimensional Time Series Medical Data», *J. Healthc. Inform. Res.*, vol. 1, n.º 1, pp. 92-118, jun. 2017.

[64] J. Li, G. Xia, y X. Shi, «Association Rules Mining from Time Series Based on Rough Set», en *Sixth International Conference on Intelligent Systems Design and Applications*, 2006, vol. 1, pp. 509-516.

[65] T. Uetake, X. Ma, M. Horikawa, T. Takeno, y M. Sugawara, «Development of sales management support system for agricultural produce using sales forecasting model», en *Computers and Industrial Engineering (CIE), 2010 40th International Conference on*, 2010, pp. 1–6.

[66] D. Wan, Y. Zhang, y S. Li, «Discovery Association Rules in Time Series of Hydrology», en *2007 IEEE International Conference on Integration Technology*, 2007, pp. 653-657.

[67] M. Last, Y. Klein, y A. Kandel, «Knowledge discovery in time series databases», *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 31, n.º 1, pp. 160-169, feb. 2001.

[68] S. Guirguis, K. M. Ahmed, N. M. E. Makky, y A. M. Hafez, «Mining the Future: Predicting Itemsets' Support of Association Rules Mining», en *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, 2006, pp. 474-478.

[69] Y. Zhi-Hong *et al.*, «Operation rule extracting based on time series association analysis in transient stability study», en *2014 International Conference on Power System Technology*, 2014, pp. 461-466.

[70] I. Z. Batyrshin y L. B. Sheremetov, «Perception Based Associations in Time Series Data Bases», en *NAFIPS 2006 - 2006 Annual Meeting of the North American Fuzzy Information Processing Society*, 2006, pp. 655-660.

[71] H.-H. Huang, Z. Wang, y W. Chung, «Efficient parameter selection for SVM: The case of business intelligence categorization», en *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, China, 2017, pp. 158-160.

[72] S. Arevalos, F. Lopez-Pires, y B. Baran, «A Comparative Evaluation of Algorithms for Auction-Based Cloud Pricing Prediction», en *2016 IEEE International Conference on Cloud Engineering (IC2E)*, Berlin, 2016, pp. 99-108.

[73] C. Bilynkievycz dos Santos, B. Pedroso, A. Margarete Guimaraes, D. Ribeiro Carvalho, y L. Alberto Pilatti, «Forecasting of Human Development Index of Latin American Countries Through Data Mining Techniques», *IEEE Lat. Am. Trans.*, vol. 15, n.º 9, pp. 1747-1753, 2017.

[74] E. Becirovic y M. Cosovic, «Machine learning techniques for short-term load forecasting», en *2016 4th International Symposium on Environmental Friendly Energies and Applications (EFEA)*, Belgrade, Serbia, 2016, pp. 1-4.

[75] W. R. Tobler, «A Computer Movie Simulating Urban Growth in the Detroit Region», *Econ. Geogr.*, vol. 46, n.º sup1, pp. 234-240, jun. 1970.

[76] K. K. Kemp, Sage Publications, y Sage eReference (Online service), *Encyclopedia of geographic information science*. 2008.

[77] M. Villatoro, C. Henríquez, y F. Sancho, «COMPARACIÓN DE LOS INTERPOLADORES IDW Y KRIGING EN LA VARIACIÓN ESPACIAL DE pH, Ca, CICE y P DEL SUELO», *Agron. Costarric.*, p. 12, 2008.

[78] E. Oktavia, Widyawan, y I. W. Mustika, «Inverse distance weighting and kriging spatial interpolation for data center thermal monitoring», en *2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, 2016, pp. 69-74.

[79] M. Villatoro, C. Henríquez, y F. Sancho, «Comparación de los interpoladores IDW Y Kriging en la variación espacial de pH, Ca, CICE y P del suelo», *Agron. Costarric.*, vol. 32, n.º 1, pp. 95-105, 2008.

[80] J. P. Celemín, «Autocorrelación espacial e indicadores locales de asociación espacial: Importancia, estructura y aplicación», *Rev. Univ. Geogr.*, vol. 18, n.º 1, pp. 11-31, 2009.

[81] R. S. Bivand, E. Pebesma, y V. Gómez-Rubio, *Applied Spatial Data Analysis with R*. New York, NY: Springer New York, 2013.

[82] R. S. Bivand, E. Pebesma, y V. Gómez-Rubio, *Applied Spatial Data Analysis with R*. New York, NY: Springer New York, 2013.

[83] Y. Bengio y Y. Grandvalet, «No Unbiased Estimator of the Variance of K-Fold Cross-Validation», p. 17.

[84] Y. Zhang y Y. Yang, «Cross-validation for selecting a model selection procedure», *J. Econom.*, vol. 187, n.º 1, pp. 95-112, 2015.

[85] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 1.ª ed. Chapman and Hall/CRC, 2012.

[86] P. Pumpuang, A. Srivihok, y P. Praneetpolgrang, «Comparisons of classifier algorithms: Bayesian network, C4.5, decision forest and NBTree for Course Registration Planning model of undergraduate students», en *2008 IEEE International Conference on Systems, Man and Cybernetics*, 2008, pp. 3647-3651.

[87] J. Llinas, C. Bowman, G. Rogova, A. Steinberg, E. Waltz, y F. White, «Revisiting the JDL data fusion model II», SPACE AND NAVAL WARFARE SYSTEMS COMMAND SAN DIEGO CA, 2004.

[88] «Distributed Data Fusion for Network-Centric Operations», *CRC Press*, 29-mar-2017. [En línea]. Disponible en: https://www.crcpress.com/Distributed-Data-Fusion-for-Network-Centric-Operations/Hall-Chong-Llinas-II/p/book/9781138073838. [Accedido: 31-ene-2018].

[89] *Handbook of Multisensor Data Fusion: Theory and Practice, Second Edition*. 2008.

[90] E. Blasch, E. Bosse, y D. A. Lambert, *High-level Information Fusion Management and Systems Design*. Artech House, 2012.

[91] «Context-Enhanced Information Fusion - Boosting Real-World Performance with Domain Knowledge | Lauro Snidaro | Springer». [En línea]. Disponible en: https://www.springer.com/gp/book/9783319289694. [Accedido: 20-nov-2018].

[92] K. S. Won y T. Ray, «Performance of kriging and cokriging based surrogate models within the unified framework for surrogate assisted optimization», en *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)*, 2004, vol. 2, pp. 1577-1585 Vol.2.

[93] D. Chen, D. Liu, Y. Li, L. Meng, y X. Yang, «Improve Spatiotemporal Kriging with Magnitude and Direction Information in Variogram Construction», *Chin. J. Electron.*, vol. 25, n.º 3, pp. 527-532, 2016.

[94] X. Luo, Y. Xu, y Y. Shi, «Comparison of interpolation methods for spatial precipitation under diverse orographic effects», en *2011 19th International Conference on Geoinformatics*, Shanghai, China, 2011, pp. 1-5.

[95] W. Gan, X. Chen, X. Cai, J. Zhang, L. Feng, y X. Xie, «Spatial interpolation of precipitation considering geographic and topographic influences - A case study in the Poyang Lake Watershed, china», en *2010 IEEE International Geoscience and Remote Sensing Symposium*, Honolulu, HI, USA, 2010, pp. 3972-3975.

[96] L. Ma, X. Chi, y C. Zuo, «Evaluation of interpolation models for rainfall erosivity on a large scale», en *2012 First International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, Shanghai, China, 2012, pp. 1-5.

[97] X. Li, J. Yang, G. Zheng, G. Han, L. Ren, y J. Wang, «Data fusion analysis of Sea Surface Temperature from HY-2A satellite radiometer», en *2016 Progress in Electromagnetic Research Symposium (PIERS)*, Shanghai, China, 2016, pp. 4312-4315.

[98] Q. P. Zhang y L. L. Lai, «Intelligent fusion for meteorological nephogram processing», en *2008 International Conference on Machine Learning and Cybernetics*, Kunming, China, 2008, pp. 3874-3877.

[99] S. Bagis, B. Berk Ustundag, y E. Ozelkan, «An adaptive spatiotemporal agricultural cropland temperature prediction system based on ground and satellite measurements», en *2012 First International Conference on Agro- Geoinformatics (Agro-Geoinformatics)*, Shanghai, China, 2012, pp. 1-6.

[100] L. Dong y Z. Hu, «The preliminary research on the multidimensional poor information spatial processing and measuring», en *2013 21st International Conference on Geoinformatics*, Kaifeng, China, 2013, pp. 1-10.

[101] M. E. D. M. Amaris y J. E. R. Rodríguez, «La contribución de las reglas de asociación a la minería de datos», *Tecnura*, vol. 7, n.º 13, pp. 94–109, 2003.

[102] «Global Administrative Areas | Boundaries without limits», 30-ene-2018. [En línea]. Disponible en: http://gadm.org/. [Accedido: 29-ene-2018].

[103] «RStudio», *RStudio*, 18-sep-2017. .

[104] «Introduction to R — R Spatial». [En línea]. Disponible en: http://rspatial.org/intr/. [Accedido: 02-dic-2018].

[105] «Using PROJ.4 — PROJ.4 5.0.0 documentation», 08-dic-2017. [En línea]. Disponible en: http://proj4.org/usage/index.html. [Accedido: 08-dic-2017].

[106] G. H. Jesus, Jose Molina, Antonio Berlanga, MIguel A Patricio, Alvaro Bustamante, y Washington Padilla, *Ciencia de Datos*. Altaria, 2018.

[107] «Códigos EPSG de Sistemas de Referencia :: Red de Información Ambiental de Andalucía :: Consejería de Medio Ambiente y Ordenación del Territorio :: Junta de Andalucía »:, 27-ene-2018. [En línea]. Disponible en:

http://www.juntadeandalucia.es/medioambiente/site/rediam/menuitem.04dc44281e5d53cf8ca78ca731525ea0/?vgnextoid=2a412abcb86a2210VgnVCM1000001325e50aRCRD&lr=lang_es. [Accedido: 26-ene-2018].

[108] I. N. de M. e Hidrología, «Instituto Nacional de Meteorología e Hidrología » Geoinformación Hidrometeorológica», *Instituto Nacional de Meteorología e Hidrología*, 13-mar-2018. .

[109] «WGS 84: EPSG Projection -- Spatial Reference». [En línea]. Disponible en: http://spatialreference.org/ref/epsg/wgs-84/. [Accedido: 24-dic-2018].

[110] I. N. de E. y Censos, «Población y Demografía», *Instituto Nacional de Estadística y Censos*, 14-mar-2018. [En línea]. Disponible en: http://www.ecuadorencifras.gob.ec/censo-de-poblacion-y-vivienda/. [Accedido: 13-mar-2018].

[111] «Weka 3 - Data Mining with Open Source Machine Learning Software in Java», 28-sep-2017. [En línea]. Disponible en: http://www.cs.waikato.ac.nz/ml/weka/. [Accedido: 28-sep-2017].

[112] L. Rodríguez Ojeda, «Construcción de kernels y funciones de densidad», ene. 2014.

[113] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, y K. R. K. Murthy, «Improvements to the SMO algorithm for SVM regression», *IEEE Trans. Neural Netw.*, vol. 11, n.º 5, pp. 1188-1193, sep. 2000.

[114] A. J. Smola y B. Schölkopf, «A tutorial on support vector regression», *Stat. Comput.*, vol. 14, n.º 3, pp. 199-222, ago. 2004.

[115] «Essentials of Management Information Systems». [En línea]. Disponible en: https://www.goodreads.com/work/best_book/20511372-essentials-of-management-information-systems-organization-and-technolog. [Accedido: 18-dic-2018].

[116] W. R. Padilla y H. J. García, «CIALCO: Alternative Marketing Channels», en *Highlights of Practical Applications of Scalable Multi-Agent Systems. The PAAMS Collection*, 2016, pp. 313-321.

[117] W. R. Padilla, G. H. Jesús, y J. M. Molina, «Model Learning and Spatial Data Fusion for Predicting Sales in Local Agricultural Markets», en *2018 21st International Conference on Information Fusion (FUSION)*, 2018, pp. 1-5.

[118] W. R. Padilla, J. García, y J. M. Molina, «Information Fusion and Machine Learning in Spatial Prediction for Local Agricultural Markets», en *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection*, 2018, pp. 235-246.

[119] W. R. Padilla, J. García, y J. M. Molina, «Improving Forecasting Using Information Fusion in Local Agricultural Markets», en *Hybrid Artificial Intelligent Systems*, 2018, pp. 479-489.

[120] W. R. Padilla, J. García, y J. M. Molina, «Knowledge Extraction and Improved Data Fusion for Sales Prediction in Local Agricultural Markets», *Sensors*, vol. 19, n.º 2, p. 286, ene. 2019.

[121] W. R. Padilla y J. García, «Data Association Methodology to Improve Spatial Predictions in Alternative Marketing Circuits in Ecuador», *Computational Intelligence and Neuroscience*, 2018. [En línea]. Disponible en: https://www.hindawi.com/journals/cin/2018/6587049/abs/. [Accedido: 07-mar-2019].