ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

ESCUELA DE FÍSICA Y MATEMÁTICA

ANÁLISIS ESTADÍSTICO MULTIVARIANTE PARA EL ESTUDIO DE LOS FACTORES QUE INFLUYEN EN LA PRODUCCIÓN DEL PLÁTANO EN EL ECUADOR, PERIODO 2014-2016

TRABAJO DE TITULACIÓN

TIPO: PROYECTO DE INVESTIGACIÓN

Trabajo de titulación presentado para optar al grado académico de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

AUTOR: SEGUNDO EDUARDO GUAMÁN DAQUILEMA

TUTOR: Ing. HÉCTOR SALOMÓN MULLO GUAMINGA

Riobamba-Ecuador

2018

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

ESCUELA DE FÍSICA Y MATEMÁTICA

El Tribunal del Trabajo de Titulación certifica que: El trabajo de investigación: "**ANÁLISIS ESTADÍSTICO MULTIVARIANTE PARA EL ESTUDIO DE LOS FACTORES QUE INFLUYEN EN LA PRODUCCIÓN DEL PLÁTANO EN EL ECUADOR, PERIODO 2014-2016"**, de responsabilidad del señor Segundo Eduardo Guamán Daquilema, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación, quedando autorizada su presentación.

	FIRMA	**FECHA**
Ing. Héctor Mullo Guaminga **DIRECTOR DEL TRABAJO DE TITULACIÓN**	_____	_____
Ing. Isabel Escudero Villa **MIEMBRO DEL TRIBUNAL**	_____	_____

Yo, Segundo Eduardo Guamán Daquilema, declaro que el presente trabajo de titulación es de mi autoría y que los resultados del mismo son auténticos y originales. Los textos constantes en el documento que provienen de otra fuente están debidamente citados y referenciados. Como autor asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación.

_____

Segundo Eduardo Guamán Daquilema

060449955-8

DEDICATORIA

A Dios, por bendecirme con la vida y la salud, además guiándome por el camino correcto para culminar mi carrera con éxito.

A mi hijo Edward Leonel, lo más hermoso que me ha podido pasar en la vida, la fuente de mi inspiración, la fortaleza de mi vida, la luz de mis ojos y mi todo.

A mi madre Paula, que a pesar de todas las dificultades siempre estaba animándome y apoyándome en cada momento de mi vida para alcanzar mi meta propuesta.

A mi esposa Rocío Alexandra, que con su cariño, amor y apoyo incondicional ha sido posible seguir adelante y lograr este objetivo.

A mis hermanas y hermanos como son: Ángela, Rosario, Luis y Pedro, que con sus palabras de aliento y ayuda ha sido posible obtener este título.

A mi suegra María, que también ha sido de gran ayuda en este proceso.

Eduardo Guamán

AGRADECIMIENTO

El agradecimiento más grande va dirigido a Dios porque este ser supremo es quien bendice y guía nuestras vidas, a mi familia quienes formaron parte fundamental en esta etapa estudiantil, apoyándome en las buenas y en las malas para conseguir esta meta.

El más sincero agradecimiento a la Escuela Superior Politécnica de Chimborazo, por abrir sus puertas y darme la oportunidad de obtener una profesión de éxito, a todos los docentes de la Escuela de Física y Matemática quienes compartieron sus conocimientos.

De manera especial al Ing. Héctor Mullo director de trabajo de titulación y a la Ing. Isabel Escudero como miembro, quienes, con sus conocimientos, sus experiencias han dedicado paciencia, tiempo y su invaluable colaboración para poder terminar exitosamente mi trabajo de investigación.

Eduardo Guamán

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE FIGURAS

INDICE DE GRÁFICOS

RESUMEN

La presente investigación tuvo por objetivo identificar los factores que influyen en la producción del plátano (Musa AAB) en el Ecuador periodo 2014-2016, se estudió variables agronómicas generales consideradas en la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC), el estudio fue de alcance descriptivo y correlacional causal porque se aplicó el análisis factorial de datos mixtos (AFDM) y el análisis de regresión lineal múltiple con variables dummy (ARLMVD), todo esto usando el software estadístico R versión 3.4.2 con la ayuda de la hoja de cálculo Excel 2016. El estudio también utilizó métodos estadísticos descriptivos y bivariados, en donde se determinó que la mayor producción del plátano en el país se da entre 0.23 y 183 toneladas métricas, por otro lado, la variable producción del plátano tiene una relación lineal positiva perfecta con la variable ventas, una relación lineal positiva muy alta con las variables: superficie plantada, superficie en edad productiva y superficie cosechada. En lo referente al AFDM que permitió agrupar las variables en dos factores que influyen en la producción del plátano, de las cuales el factor superficie está formado por las variables: superficie cosechada, superficie en edad productiva, superficie plantada y ventas; el factor uso y cuidado está formada por las variables: uso de fitosanitarios, uso de plaguicida químico y uso de fertilizante químico. Además, con referente al ARLMVD que se aplicó para las variables que conforman los dos factores que se determinó en el AFDM permitió determinar que la variable (uso de fertilizante químico) que conforma el factor uso y cuidado no es significativo en la variable dependiente (producción del plátano), lo cual permitió realizar un modelo restringido excluyendo dicha variable, el modelo restringido fue capaz de explicar el 99.43% de la variabilidad observada en la producción del plátano.

Palabras Claves: <ESTADÍSTICA>, <ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE CON VARIABLES DUMMY (ARLMVD)>, <ANÁLISIS FACTORIAL DE DATOS MIXTOS (AFDM)>, <ENCUESTA DE SUPERFICIE Y PRODUCCIÓN AGROPECUARIA CONTINUA (ESPAC)>, <PLÁTANO *(Musa ABB)*>.

ABSTRACT

The present investigation had the objective to identify the factors that influence plantain production (Musa AAB) in Ecuador during the period 2014-2016, general agronomic variables considered in the Survey of Surface and Continuous Agricultural Production (ESPAC, by its acronym in Spanish) were studied, the research was descriptive and causal correlational in scope because the factor analysis of mixed data (FAMD) and the multiple linear regression analysis with dummy variables (MLRADV) were applied, all this using the R statistical software version 3.4.2 with the help of the Excel 2016 spreadsheet. The study also used descriptive and bivariate statistical methods, where it was determined that the highest plantain production in the country is between 0.23 and 183 metric tons, on the other hand, the plantain production variable has a perfect positive linear relationship with the sales variable, a very strong positive linear relationship with the variables: planted surface, surface in productive age and harvested surface. With regard to the FAMD that allowed grouping the variables into two factors that influence plantain production, which the surface factor is formed by the variables: harvested surface, surface in productive age, planted surface and sales; the use and care factor is formed by the variables: use of phytosanitary products, use of chemical pesticide and use of chemical fertilizer. In addition, with regard to the MLRADV that was applied for the variables that make up the two factors that were determined in the FAMD, it allowed to determine that the variable (use of chemical fertilizer) that makes up the use and care factor is not significant on the dependent variable (plantain production), which allowed to execute a restricted model excluding this variable, the restricted model was able to explain 99.43% of the variability observed in plantain production.

Key words: <STATISTICS>, <MULTIPLE LINEAR REGRESSION ANALYSIS WITH DUMMY VARIABLES (MLRADV)>, <FACTOR ANALYSIS OF MIXED DATA (FAMD)>, <SURVEY OF SURFACE AND CONTINUOUS AGRICULTURAL PRODUCTION (ESPAC, by its acronym in Spanish)>, <PLANTAIN (*Musa ABB*)>.

INTRODUCCIÓN

El plátano (Musa AAB[1]) es una fruta tropical como alimento básico de la población ecuatoriana que forma parte de la canasta básica familiar, al ser la materia prima de deliciosos platos tradicionales sobre todo de la región costa, desde el punto de vista socioeconómico es la principal fuente de ingreso y de empleo para miles de ecuatorianos. Además, Ecuador exporta una gran cantidad de productos a varios países del mundo los cuales sustentan la economía del país, el plátano es el segundo producto más exportado, después del petróleo.

Durante años la producción del plátano es una alternativa para el agricultor ya que beneficia al sector campesinos dedicado a esta actividad. Sin embargo, la disminución en la producción del plátano es uno de los problemas que presenta Ecuador en los últimos años.

En el Ecuador son escasos o pocos los estudios realizados sobre los factores que influyen en la producción del plátano. Por otra parte, la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC) y el Instituto Nacional de Estadística y Censos (INEC) se encarga de satisfacer la demanda de información agropecuaria realizando algunas publicaciones de estadísticas descriptivas las cuales ayudan a la toma de decisiones, sin embargo, los datos generados por estas instituciones están subutilizadas y no representan la multidimensionalidad de la problemática de la producción del plátano. Este trabajo pretende llenar este vacío mediante el uso del Análisis Multivariante (AM), específicamente el Análisis Factorial de Datos Mixtos (AFDM) y Análisis de Regresión Lineal Múltiple con Variables Dummy (ARLMVD) para determinar los factores que influyen en la producción del plátano en el Ecuador. Todo esto utilizando datos secundarios de acceso público obtenidos dentro de estadísticas agropecuarias del INEC y mediante la utilización del software estadístico R versión 3.4.2.

[1] El grupo del genoma AAB o tipo “manzano” comprende todos los cultivares que tienen dos conjuntos de cromosomas donados por Musa acuminata y uno por Musa balbisiana. Algunos de los cultivares, como los plátanos, son tipos de almidón que generalmente se cocinan, mientras que otros, como la seda, son tipos dulces de postre que se consumen crudos.

CAPITULO I

1. MARCO REFERENCIAL

1.1. Antecedentes

El plátano es uno de los alimentos básicos de la dieta de la población ecuatoriana, especialmente del litoral y oriente, desde el punto de vista socioeconómico es el principal componente de la mayoría de los sistemas de producción, siendo la principal fuente de ingreso y de empleo para miles de ecuatorianos (Orellana *et al.*, 2002). El plátano forma parte de la canasta básica familiar, al ser la materia prima de deliciosos platos tradicionales sobre todo de la región costa, las plantaciones de plátano se pueden divisar por todo el territorio ecuatoriano, gracias a que el clima de Ecuador es beneficioso para el cultivo de plátano y otras frutas exóticas (PRO ECUADOR, 2015). Ademán según COVECA citado González (2012) el plátano o banano es uno de los cultivos más importantes en la agricultura.

(Jeproll, 2009) afirma que el Ecuador es un productor de vanguardia de plátanos a escala mundial. Por otro lado, según la página web Sinmiedosec (2015) y OEC, Ecuador exporta una gran cantidad de productos a varios países del mundo los cuales sustentan la economía del país, el plátano es el segundo producto más exportado, después del petróleo.

El Sistema de Información Nacional de Agricultura, Ganadería, Acuacultura y Pesca (SINAGAP) del Ecuador menciona que la producción mundial de plátano en el año 2014 incrementó 1.93% con respecto al año 2012, en el año 2015 incrementó 4% respecto al año 2013 y en el año 2016 incrementó 1% con relación al año 2015 (SINAGAP, 2014), (SINAGAP, 2015) y (SINAGAP, 2015).

La producción nacional de plátano en el año 2014 incrementó en 27.36% con respecto al año 2013 (SINAGAP, 2014). En el año 2015 disminuyó en 11% respecto al año 2014 y debido a esto, las exportaciones también descendieron en 2% (SINAGAP, 2015). Durante el periodo 2016 la producción presenta un decremento de 10% respecto al año 2015 y las exportaciones también descendieron en 2%. Sin embargo, Ecuador se mantiene entre los principales exportadores de este producto a nivel mundial (SINAGAP, 2016).

Con respecto al método empleado en la presente investigación tenemos datos que corresponden a variables tanto cualitativas y cuantitativas, por lo tanto, es preciso emplear el método de AFDM

y ARLMVD. Según Pagès (2014) el ***primer método*** fue propuesto para encontrar relaciones entre grupos de variables y, dentro de tales grupos, variables cuantitativas y cualitativas simultáneamente. El AFDM conjuga las técnicas del Análisis de Componentes Principales (ACP) y Análisis de Correspondencias Múltiples (ACM), ampliamente usadas en estudios de variabilidad genética (Franco y Hidalgo, 2003) y según la revisión literaria el ***segundo método*** se emplea cuando se pretenda analizar la relación existente entre una variable dependiente cuantitativa y un conjunto de variables independientes mixtas, las variables cualitativas son codificadas numéricamente con 0's y 1's, y para evitar la trampa de las variables dummy se analiza restando 1 a la cantidad total de categorías de una variable dummy ($m-1$). Estas técnicas son muy utilizadas en los últimos años ya que es frecuente analizar variables cuantitativas y cualitativas al mismo tiempo.

En relación a algunas investigaciones que han utilizado el ADFM y el ARLMVD tenemos: García et al. (2017) realizaron la investigación ***"Caracterización de los Sistemas Lecheros en San Joaquín de Tuís, Turrialba, Costa Rica"*** y al aplicar el AFDM identificaron dos grupos de producción representativos. Pacheco *et al.* (2009) enfocaron el AFDM en la investigación sobre ***"Clasificación de 85 accesiones de arveja (Pisum sativum L.), de acuerdo con su comportamiento agronómico y caracteres morfológicos"*** para la identificación de progenitores y selección de variables, en donde el AFDM permitió agrupar las 85 accesiones en tres grupos de variación cualitativa y cuantitativa.

Mamuye realizó una investigación dirigida al ***"Análisis estadístico de los factores que afectan en la producción de plátano en el distrito de Gamo Gofa, sur de Etiopía".*** El estudio utiliza métodos estadísticos descriptivos como tabla de distribución de frecuencia, medidas de resumen y métodos estadísticos inferenciales, principalmente el análisis de regresión múltiple de la función de producción de Cobb-Douglas con transformación logarítmica, utilizando la técnica de mínimos cuadrados ordinarios (MCO), esto debido a que dicha investigación afirma que no existe una relación lineal directa entre la variable respuesta y las variables independientes. Obteniendo como resultado que la edad de las plantas del plátano, el tamaño de la familia, la edad de los agricultores y la cantidad de mano de obra que se usa para la granja bananera resultaron ser predictoras estadísticamente significativos de la producción de plátano en la región. También en este estudio, factores como el género, el nivel de educación de los agricultores, la fertilidad del suelo agrícola y la cantidad de fertilizante que se usa en la finca bananera no tienen un impacto estadísticamente significativo en la producción de banano. Finalmente, en esa investigación indica, que los investigadores recomendaron que es deber y responsabilidad de la oficina agrícola introducir nuevas variedades de plátano y crear conciencia sobre la producción de plátano a los agricultores para aumentar la productividad de la planta (Mamuye, 2016).

1.2. Formulación del problema

Ecuador, es el principal productor y exportador a nivel mundial de plátano, siendo esta una fuente de ingreso y de empleo para miles de ecuatorianos. Se percibe una aparente disminución en la producción del plátano a partir del año 2014, por lo que es de interés identificar los factores que están influyendo en este fenómeno.

Aportando con una visión general a las organizaciones de interés, que les permita tomar decisiones en beneficio de los agricultores y así, fortalecer la matriz productiva del país al incrementar los ingresos, generando estabilidad económica a los productores minoristas y mayoristas del país.

1.2.1. Forma interrogativa

¿Qué factores influyen en la producción del plátano en el Ecuador con base a la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC) 2014-2016?

1.2.2. Sistematización del problema

¿Existe un cambio o disminución de la producción del plátano en los años 2014, 2015 y 2016?

¿Cuáles son los principales factores o variables que intervienen en la producción del plátano en el Ecuador?

1.3. Justificación del trabajo de titulación

1.3.1. Justificación teórica

Debido a que no existe un estudio que identifique los factores que influyen en la producción del plátano en el Ecuador surge la necesidad de realizar esta investigación con el propósito de aportar a los productores y al país en general. Los resultados podrán ser incorporado como conocimiento ya que se identificará los factores influyentes en este fenómeno.

1.3.2. Justificación aplicativa

La ESPAC tiene como finalidad satisfacer la demanda de información agropecuaria, facilitando de esta manera el monitoreo permanente del sector, también realiza algunas publicaciones de

estadísticas descriptivas conjuntamente con el INEC. Estas publicaciones si bien es cierto ayuda a la toma decisiones, pero no muestran particularidades que surgen al momento de estudiar conjuntamente toda la información, por lo tanto, con este trabajo se pretende realizar un análisis más profundo que ayude a los productores a obtener más productividad, más ganancias y menos pérdida de esta fruta. Además, este rubro es de vital importancia para el Ecuador como ya se menciona anteriormente y dar una información de este tipo permitirá al estado ecuatoriano y las organizaciones de interés diseñar sus políticas económicas ya que el plátano es el segundo producto más exportado, después del petróleo a nivel mundial.

1.4. Objetivos

1.4.1. Objetivo general

Estudiar los factores que influyen en la producción del plátano en el Ecuador con base a la ESPAC 2014-2016 mediante técnicas estadísticas multivariantes.

1.4.2. Objetivos específicos

- Identificar los factores que intervienen en la producción del plátano, utilizando métodos estadísticos univariantes y multivariantes.
- Generar un modelo de predicción adecuado para la producción del plátano.

CAPITULO II

2. MARCO TEÓRICO

2.1. El plátano

El plátano es una fruta tropical originada en el Sudoeste Asiático procedente del árbol que recibe el mismo nombre o banano, perteneciente a la familia de las musáceas (es un híbrido triploide de Musa acuminata y Musa balbisiana) (IICA, MAGFoR y jica, 2004). Tiene forma alargada o ligeramente curvada, de 100-200 g de peso. La piel es gruesa, de color amarillo y fácil de pelar, y la pulpa es blanca o amarillenta y carnosa. Aunque en numerosas ocasiones se ha citado América Central como el lugar de origen del plátano, la mayoría de los autores opinaron que esta fruta es originaria del sudeste asiático, concretamente de la India, siendo conocida en el Mediterráneo después de la conquista de los árabes en el año 650 D.C. La especie llegó a Canarias en el siglo XV y desde allí fue llevada a América en el año 1516 (Ávila et al., 2007). A pesar que su origen es del Sudoeste Asiático, a lo largo de los años su cultivo se ha extendido a Centroamérica, Sudamérica, y África Subtropical.

2.2. Producción Nacional

La producción y las ventas de esta fruta se ha disminuido en los últimos años, según datos de la ESPAC, a continuación, en la Tabla 1-2 se presenta la evolución de las ventas y el volumen de producción de los años 2014, 2015 y 2016.

Tabla 1-2: Evolución de la producción y ventas en los años 2014, 2015 y 2016

Año	SUPERFICIE (Has.)		PRODUCCIÓN	VENTAS
	Plantada	Cosechada	(Tm.)	(Tm.)
2014	131340	104574	761226	641305
2015	123355	105817	675538	582460
2016	110110	94911	610413	530668

Fuente: ESPAC, INEC
Realizado por: Eduardo Guamán 2018

Según datos de la misma fuente, el porcentaje de producción de plátano ecuatoriano por región y provincia para el año 2014, 2015 y 2016 se distribuye como se puede observar en la Figura 1-2, 2-2 y 3-2 respectivamente.

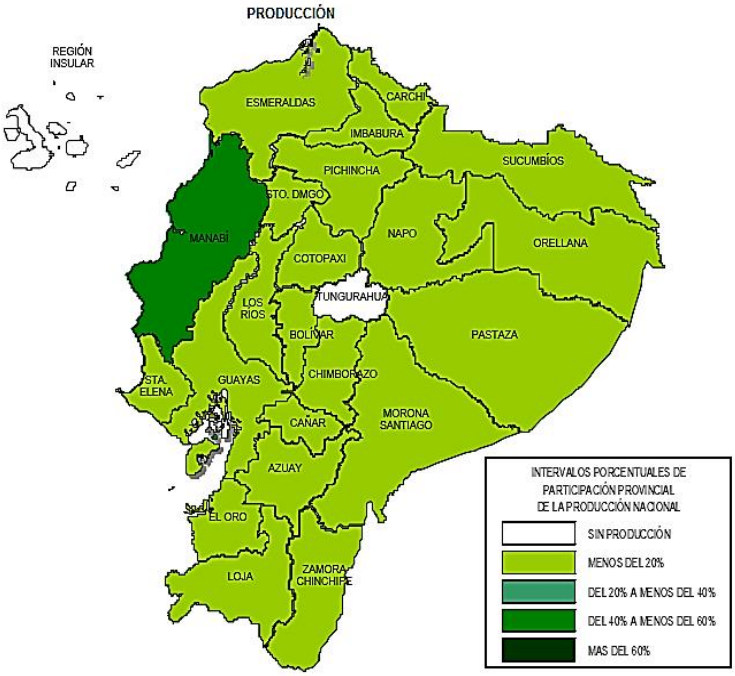


Figura 1-2: Porcentaje de producción, según región y provincia, 2014
Fuente: ESPAC, INEC. 2014

De acuerdo a Figura 1-2, para el año 2014, se observa que en la provincia de Tungurahua y en la Región Insular no hubo producción, mientras que en la provincia de Manabí hubo del 40% a menos del 60% de producción y en el resto de las provincias se presenció menos del 20% de producción.

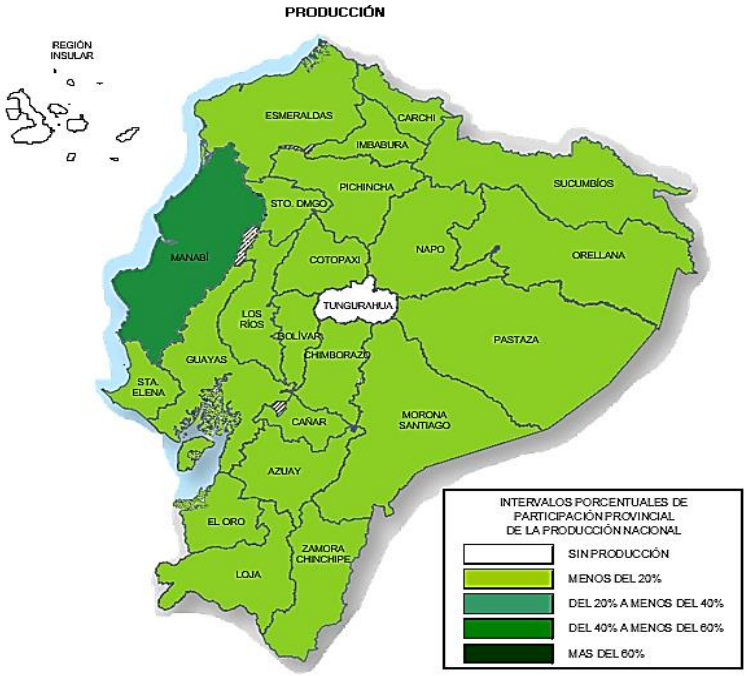


Figura 2-2: Porcentaje de producción, según región y provincia, 2015
Fuente: ESPAC, INEC. 2015

En la Figura 2-2 de la misma manera para el año 2015, se observa que en la provincia de Tungurahua y en la Región Insular no hubo producción, mientras que en la provincia de Manabí hubo del 40% a menos del 60% de producción y en el resto de las provincias se presenció menos del 20% de producción.

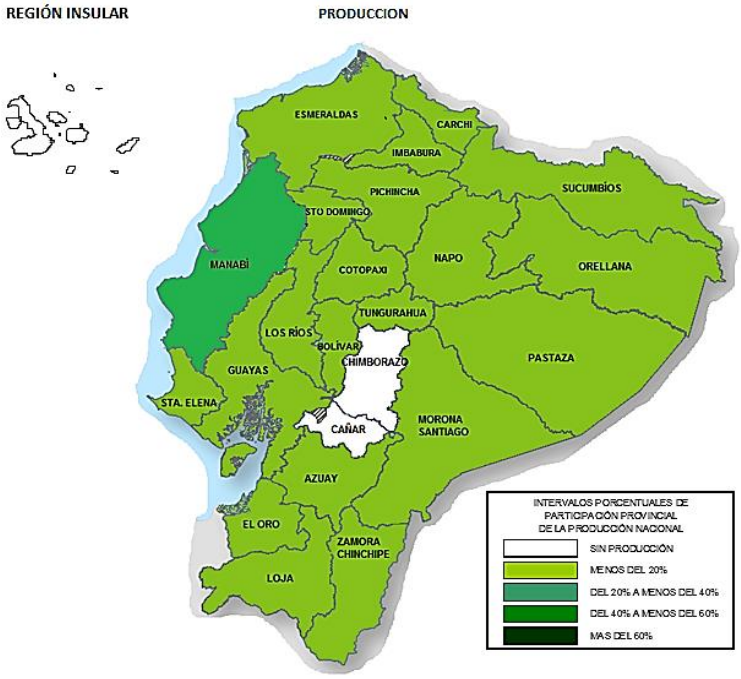


Figura 3-2: Porcentaje de producción, según región y provincia, 2016
Fuente: ESPAC, INEC. 2016

La Figura 3-2 para el año 2016, se observa que, en las provincias de Chimborazo, Cañar y en la Región Insular no hubo producción, mientras que en la provincia de Manabí hubo del 20% a menos del 40% de producción y en el resto de las provincias se presenció menos del 20% de producción.

En general del año 2016 con respecto a los años 2014 y 2015 podemos decir que no hubo producción en dos provincias como es Chimborazo, Cañar y en la Región Insular ya que en los años 2014 y 2015 que no hubo producción tan solo en la provincia de Tungurahua y en la Región Insular. Por último, en el año 2016 en la provincia de Manabí hubo un decremento en el porcentaje de producción en comparación a los años 2014 y 2015.

2.3. Variables en la producción del plátano

Para la producción del plátano se toman en cuenta una gama de variables, entre estos están:

2.3.1. Variables ambientales o climáticos

- Altitud
- Temperatura
- Precipitación
- Humedad relativa
- Agua
- Luz
- Viento

2.3.2. Variables edáficas

- Condiciones o tipos de suelo

2.3.3. Variables generales (considerados por la ESPAC)

Las variables que se estudiarán en esta esta investigación (variables con datos completos) y consideradas por la ESPAC según la fuente INEC y ANDA (2010) son:

Provincia: Es la provincia donde se encuentra sembrado el cultivo.

Condición de cultivo: Es la condición del cultivo permanente que han sido sembrados de un solo tipo, de dos o más clases, o si estos se encuentran en invernaderos.

Cultivo con condición de Sólo: Es el área que se encuentra plantada o sembrada por un solo cultivo, sea que esté en campo abierto o bajo invernadero.

Cultivo con condición de Asociado: Es el área que se encuentra plantada o sembrada en forma intercalada con dos o más cultivos.

Cultivo con condición bajo Invernadero: Es el área que se encuentra plantada o sembrada sea en condición de solo o asociado bajo construcciones con soporte de madera, metal o mixto, cubiertas con un material plástico que permite el paso de la luz solar, con la finalidad de obtener condiciones climáticas y ambientales que favorezcan el desarrollo de las plantas en su interior.

Edad de la plantación: Es la edad que tiene la plantación del cultivo permanente.

Semilla de más uso: Plantas procreadas o multiplicadas por dos individuos de distintas especies, es decir el resultado de todo lo que es producto de especies o variedades distintas. Estos pueden ser a través de fecundación de las flores de un determinado cultivar con polen de un cruzamiento, o como portainjertos. Puede haber semillas híbridas nacionales e internacionales, estas últimas son importadas de otros países.

Común: Plantas que no han recibido tratamiento genérico alguno.

Mejorada: Corresponde a las plantas que han sido mejoradas genéticamente, con el fin de aumentar la capacidad productiva, resistencia a enfermedades, plagas, sequías o para que adquiera otras características deseables.

Híbrida nacional: Corresponde a las semillas propias del país

Híbrida internacional: Corresponde a las semillas exportadas de otros países.

Superficie plantada en hectáreas: Es la superficie que ocupan los diferentes árboles o plantas con distancias establecidas en hectáreas, que le permita el desarrollo suficiente de la planta, permitiendo la libre circulación del aire y la luz.

Superficie en edad productiva en hectáreas: Es la edad que ha alcanzado o debe alcanzar un árbol o una planta para entrar en el período de producción y poder obtener cosechas del (la) mismo(a).

Superficie cosechada en hectáreas: Es la superficie que está ocupada por un determinado cultivo y está lista para la recolección o cosecha manual o mecánica de los frutos, los mismos que deben alcanzar un determinado grado de desarrollo y de madurez para su comercialización o conservación.

Uso de riego: Es la práctica de suministrar deliberadamente agua a la tierra para la producción y mejoramiento de los cultivos permanentes.

Uso de fertilizantes: Es la utilización o no de fertilizantes en los cultivos permanentes.
Fertilizantes: es cualquier sustancia añadida al suelo que sirve para aumentar los nutrientes de las plantas, mejorar su crecimiento e incrementar la productividad.

Uso de fitosanitarios: Es la aplicación de insecticidas, fungicidas y control biológico que se realizan en los cultivos con el fin de combatir las plagas y enfermedades y evitar daños en el desarrollo biológico de los mismos.

Producción cantidad cosechada: Es la cantidad de frutos cosechados en un tiempo determinado de acuerdo al ciclo de producción de cultivos permanentes, el mismo que está destinado para su comercialización o autoconsumo.

Producción cantidad cosechada equivalente en libras: Es la cantidad de productos cosechados en un tiempo determinado de acuerdo al ciclo de producción de cada cultivo permanente equivalente en libras.

Cantidad vendida: Es la cantidad vendida de la producción de cultivos permanentes.

Cantidad vendida equivalencia en libras ventas: Es la cantidad vendida de la producción de cultivos permanentes equivalente en libras.

Producción en toneladas métricas: Cantidad de frutos cosechados en un tiempo determinado de acuerdo al ciclo de producción de cada cultivo, el mismo que está destinado para su comercialización o autoconsumo en toneladas.

Ventas en toneladas métricas: Es la cantidad de frutos vendida en un tiempo determinado de acuerdo al ciclo de producción de cada cultivo, el mismo que está destinado para su comercialización en toneladas métricas.

Uso de fertilizante orgánico: La utilización o no de fertilizante orgánico.

Uso de fertilizante químico: Es la utilización o no de fertilizante químico.

Uso de plaguicida orgánico: Es la utilización o no de plaguicida orgánico. Plaguicida, o pesticida, es cualquier sustancia destinada a prevenir, destruir, atraer, repeler o combatir cualquier plaga, incluidas las especies indeseadas de plantas o animales, durante la producción, almacenamiento, transporte, distribución y elaboración de alimentos, productos agrícolas o alimentos para animales, o que pueda administrarse a los animales para combatir ectoparásitos.

Uso de plaguicida químico: Utilización o no de plaguicida químico.

Los conceptos anteriores descritos para cada una de las variables o factores considerados en el estudio y la ESPAC son según la fuente de la INEC y Archivo Nacional de Datos y Metadatos Estadísticos (ANDA) (INEC y ANDA, 2010).

2.4. Técnicas Multivariadas

El AM es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos. Las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás. La información estadística en AM es de carácter multidimensional, por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariantes juegan un papel fundamental. La información multivariante es una ***matriz de datos***, pero a menudo, en AM la información de entrada consiste en matrices de distancias o similaridades, que miden el grado de discrepancia entre los individuo (Cuadras, 2014, p. 13).

Si se observa p variables numéricas en un conjunto de n elementos, cada una de estas p variables se denomina una variable escalar o univariante y el conjunto de las p variables forman una variable vectorial o multivariante. Los valores de las p variables escalares en cada uno de los n elementos pueden representarse en una matriz $\mathbf{X}$ de dimensiones $(n \times p)$ que llamaremos matriz de datos. Se denota por x_{ij} al elemento genérico de esta matriz, que representa el valor de la variable escalar *j* sobre el individuo i, es decir, datos x_{ij} donde $i = 1, \dots, n$ representa el individuo y $j = 1, \dots, p$ representa la variable.

La matriz de datos $\mathbf{X}$, se representa:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x'}_1 \\ \vdots \\ \vdots \\ \mathbf{x'}_n \end{bmatrix}$$

Donde cada variable $\mathbf{x'}_i$ es un vector fila $p \times 1$ que representa los valores de las p variables sobre el individuo i, alternativamente se representa la matriz $\mathbf{X}$ por columnas: $\mathbf{X} = [\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(p)}]$, donde ahora cada variable $\mathbf{x}_{(j)}$ es un vector columna $n \times 1$ que representa la variable escalar x_j medida en los n elementos de la población. Se llamará $\mathbf{x} = (x_1, \dots, x_p)'$ a la variable multivariante formada por las p variables escalares que toma los valores particulares $\mathbf{x_1}, \dots, \mathbf{x_p}$, en los n elementos observados (Peña, 2002, pp. 67-69).

2.4.1. Clasificación de las técnicas multivariantes

Los enfoques de dependencia y el de interdependencia cobijan la mayoría de metodologías multivariadas, en éste estudio no se explica a profundidad cada uno de los métodos, sin embargo, se presenta un esquema general (Ver Figura 4-2).

2.4.1.1. Técnicas de análisis de dependencias

Las técnicas de análisis de dependencias buscarán la existencia o ausencia de relaciones entre los dos grupos de variables. Si el investigador, basándose en un experimento controlado o gracias a una base teórica previa, clasifica los dos grupos de variables en dependientes e independientes, entonces el objetivo de las técnicas de dependencia será establecer si el conjunto de variables independientes afecta al conjunto de dependientes de manera conjunta o individual (Aldás y Uriel, 2017, p. 23-24). (Ver Figura 4-2).

2.4.1.2. Técnicas de análisis de interdependencia

Sin embargo, se puede encontrar ante un problema en el que sea imposible distinguir conceptualmente entre variables dependientes e independientes. Interesa simplemente saber cómo se relacionan entre sí todas las variables del problema, es decir, interesa determinar cómo y por qué las variables están correlacionadas entre ellas (Aldás y Uriel, 2017, p. 24). (Ver Figura 4-2)

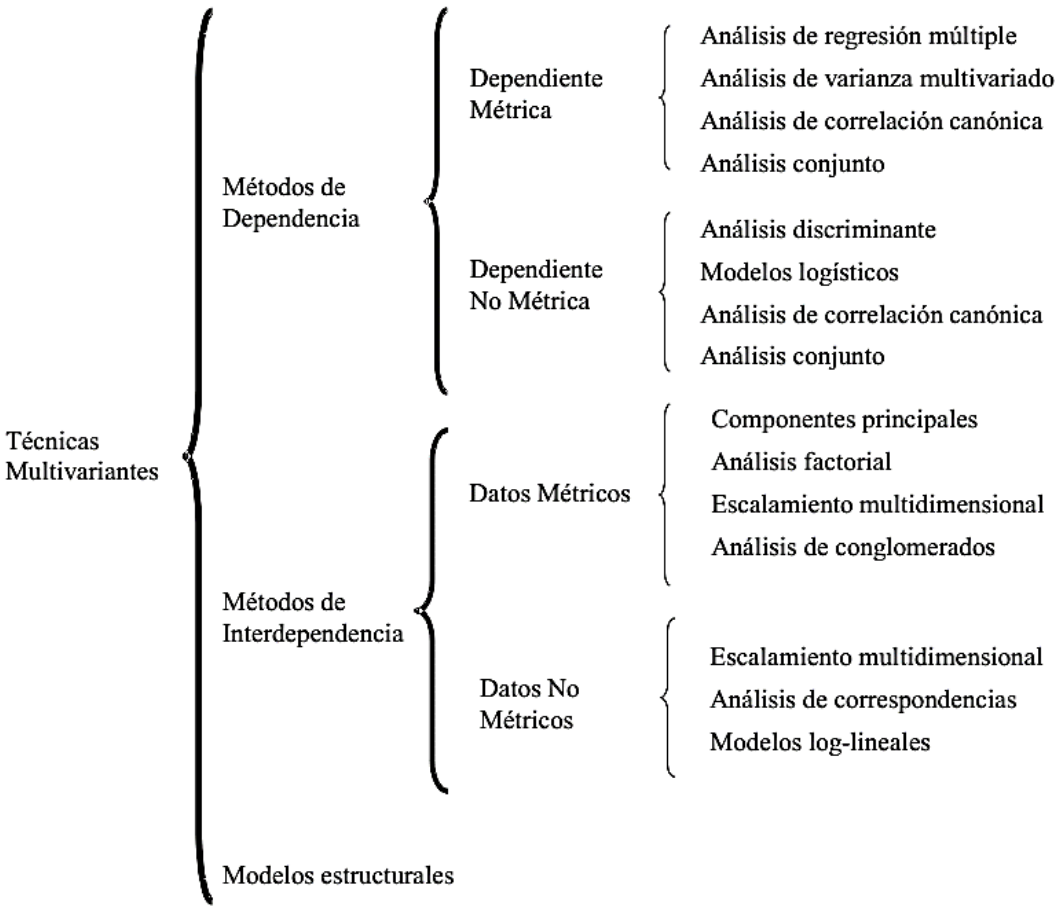
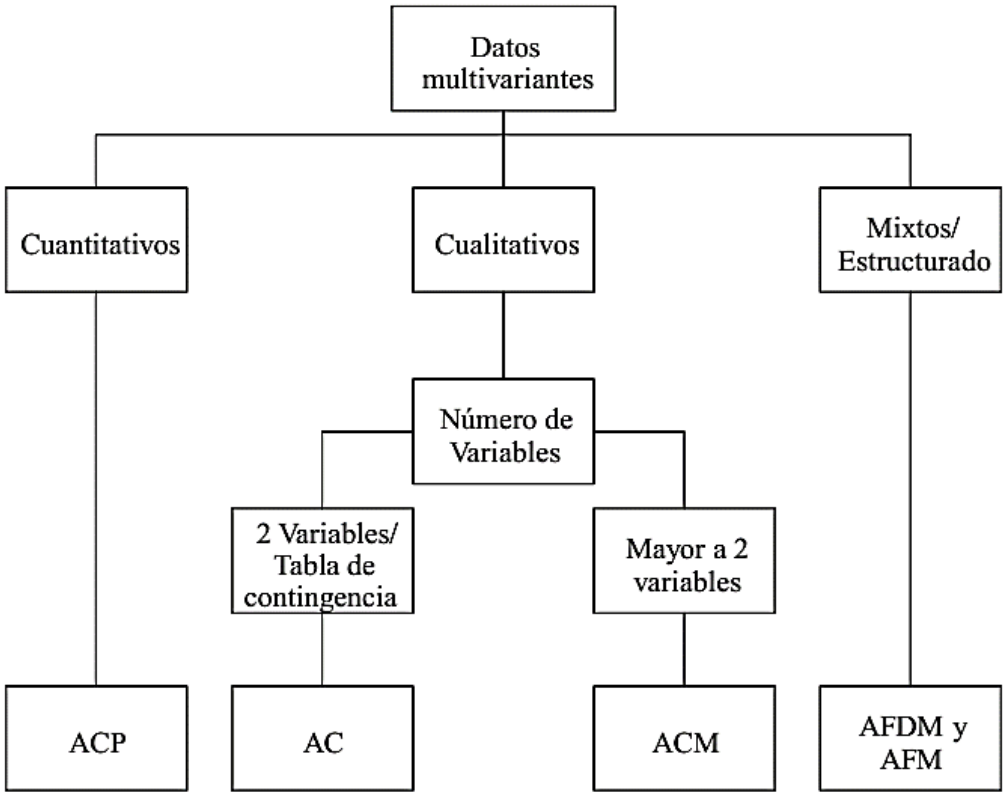


Figura 4-2: Principales técnicas multivariadas
Fuente: (Díaz y Morales, 2012, p. 20)
Realizado por: Eduardo Guamán 2018

Según Kassambara clasifica los métodos de componentes principales desde un enfoque exploratorio y descriptivo, además afirma que el tipo de métodos de componentes principales a utilizar depende de los tipos de variables obtenidas en el conjunto de datos (Kassambara, 2017, pp. vi-vii). (Ver Figura 5-2).



- ACP: Análisis de Componentes Principales
- AC: Análisis de Correspondencias
- ACM: Análisis de Correspondencias Múltiples
- AFDM: Análisis Factorial de Datos Mixtos
- AFM: Análisis Factorial Múltiple

Figura 5-2: Métodos de componentes principales
Fuente: (Kassambara, 2017, p. vi)
Realizado por: Eduardo Guamán 2018

2.5. Análisis factorial de datos mixtos (AFDM)

El autor Pagès en el año 2015 en su libro escrito sobre "Multiple Factor Analysis by Example Using R" explica de manera profunda sobre dicho método, afirmando que la necesidad de introducir simultáneamente variables cuantitativas y cualitativas (conocidas como datos mixtos) como elementos activos de un análisis factorial es común. La metodología habitual es transformar las variables cuantitativas en variables cualitativas, descomponer su intervalo de variación en clases y sometiendo la tabla homogénea resultante a un análisis de ACM. Esta metodología es relativamente fácil de implementar y se usa siempre que haya suficientes individuos; generalmente más de 100, un límite por debajo del cual los resultados del ACM no son muy estables.

En dos casos, hay ventajas para conservar las variables cuantitativas:

1. Cuando el número de variables cualitativas es muy bajo en comparación con las variables cuantitativas: por lo que se debe pensar dos veces antes de codificar 20 variables cuantitativas con el único objetivo de introducir una única variable cualitativa.
2. Cuando solo hay un pequeño número de individuos.

El método que se presenta en esta investigación proviene de dos orígenes diferentes. En 1979, Brigitte Escofier sugirió introducir variables cuantitativas en el ACM (gracias a la codificación apropiada). En 1990, Gilbert Saporta sugirió introducir variables cualitativas en el ACP gracias a una métrica específica. En realidad, estos dos enfoques diferentes producen los mismos resultados. El análisis factorial resultante presenta un número suficiente de propiedades positivas y potencial de aplicación para justificar el estado de un método separado, esto se logra con el análisis factorial de datos mixtos (AFDM) (Pagès, 2015, p. 67).

2.5.1. Datos, notaciones

Se tiene I individuos, a cada individuo i se le atribuye un peso p_i como $\sum_i p_i = 1$, para simplificar las cosas, excepto cuando se indique explícitamente, bajo la suposición de que los individuos tienen el mismo peso, por lo tanto $p_i = 1/I \ \forall_{\boldsymbol{i}}$. Estas personas se describen por:

- K_1 variables cuantitativas {k = 1, K_1}, estas variables están estandarizadas (centradas y reducidas), esto no es solo por conveniencia sino que es necesario debido a la presencia de dos tipos de variables.
- Q variables cualitativas $\{q = 1, Q\}$, la q-ésima variable presenta las categorías $K_q\{k_q = 1, K_q\}$, el número total de categorías es $\sum_q K_q = K_2$ y se denota con p_{k_q} la proporción de individuos que poseen la categoría k_q.

Sea $K = K_1 + K_2$ el número total de variables cuantitativas y variables indicadoras. Estas anotaciones se pueden observar en la Figura 6-2 en la cual las variables cualitativas aparecen tanto en su forma resumida como en su forma disyuntiva completa.

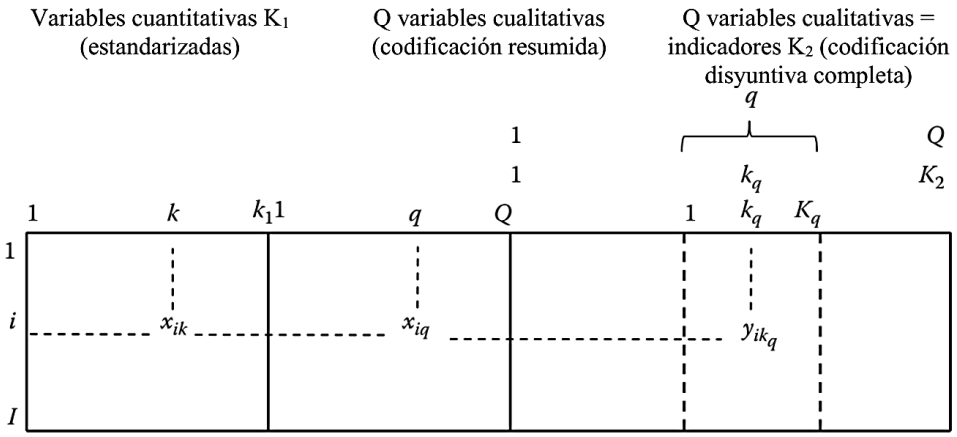


Figura 6-2: Estructura de datos y notaciones principales
Fuente: (Pagès, 2015, p. 68)
Realizado por: Eduardo Guamán 2018

Donde x_{ik} es el valor de i para la variable (centrado-reducido) k, x_{iq} es la categoría de i para la variable q y y_{ik_q} es 1 si i posee k_q de la variable q y en caso contrario, 0 (Pagès, 2015, pp. 67-68).

2.5.2. Representación de variables

Permite que $R^{\boldsymbol{I}}$ sea el espacio de funciones en I. Este espacio está dotado con la métrica diagonal de los pesos de los individuos, denominada D:

$$D(i,j) = \begin{cases} 0 \; si \; j \neq i \\ p_i \; si \; j = 1 \end{cases}$$

En general, los individuos tienen los mismos pesos: $D = (1/I)I_d$ (donde I_d es la matriz de identidad de dimensiones apropiadas).

Al igual que en el ACP estandarizada, las variables cuantitativas están representadas por vectores con una longitud 1.

Como el ACM, la variable q está representada por la nube N_q de sus indicadores centrados K_q, esta nube genera subespacio E_q de dimensión $K_q - 1$, donde E_q es el conjunto de funciones centradas constantes en las clases de la partición definida por q. Para que N_q posea las mismas propiedades de inercia que en un ACM, si se realiza un ACP no estandarizado sobre esto, el indicador k_q debe dividirse por p_{k_q} y se le atribuye un peso p_{k_q} (obteniendo la inercia exacta del ACM que requiere el peso p_{k_q}/J. Dividir por J 'medias' las inercias de acuerdo con el número de variables, lo cual es indeseable en este caso, ya que las variables cualitativas se enfrentan con variables cuantitativas, cuyas inercias no se promedian).

Específicamente, al proceder de esta manera, se obtiene una propiedad fundamental del ACM: la inercia proyectada de N_q en una variable centrada y es igual a la razón de correlación cuadrada $\eta^2(q, y)$ entre q e y.

Al buscar la dirección v de $R^{\boldsymbol{I}}$ que maximiza la inercia proyectada de la nube N_k (compuesta por las variables cuantitativas y los indicadores), se maximiza el criterio:

$$\sum_{k \in K_1} r^2(k, v) + \sum_{q \in Q} \eta^2(q, v)$$

Este es el punto de partida del método propuesto por Gilbert Saporta en 1990. Geométricamente, como las k variables están estandarizadas, la coordenada de proyección de la variable k en v vale $\cos(\theta_{kv}) = r(k, v)$ donde θ_{kv} es el ángulo entre los vectores k y v. Del mismo modo, como v está centrado, $\eta^2(q, y) = \cos^2(\theta_{qv})$ donde θ_{qv} es el ángulo entre v y su proyección en E_q. El criterio se expresa así:

$$\sum_{k \in K_1} \cos^2(\theta_{kv}) + \sum_{q \in Q} \cos^2(\theta_{qv})$$

Este siendo el punto de partida del método propuesto por Brigitte Escofier en 1979.

La influencia de una variable debe explicarse de acuerdo con la dimensión del subespacio que genera. Por lo tanto, en el espacio $R^{\boldsymbol{I}}$:

- Una variable cuantitativa está representada por un vector asociado con una inercia de 1.
- Una variable cualitativa con categorías K_q está representada por vectores K_q que generan un subespacio E_q de dimensión $K_q - 1$, todos asociadas con una inercia de $K_q - 1$.

Al igual que en el ACM, la inercia total de una variable cualitativa aumenta cuantas más categorías haya. Sin embargo, cuando se proyecta en cualquier dimensión de E_q, esta inercia vale 1, de esta manera, cuando se buscan direcciones de inercia máxima, estos dos tipos de variables se equilibran, las cuales son resaltadas por una u otra de las dos expresiones del criterio a continuación (Pagès, 2015, pp. 68-69).

2.5.3. Representación de individuos

Las dimensiones del espacio $\mathbb{R}^{\mathrm{K}}$ son las variables cuantitativas K_1 y los indicadores K_2, su métrica Euclidiana diagonal es la ponderación de las columnas (1 para las variables cuantitativas y p_{k_q} para las categorías).

La distancia entre los individuos i y l se expresa:

$$d^2(i,l) = \sum_{k \in K_1} (x_{ik} - x_{lk})^2 + \sum_{q \in Q} \sum_{k \in K_q} p_{k_q} \left(\frac{y_{ik_q}}{p_{k_q}} - \frac{y_{lk_q}}{p_{k_q}} \right)^2$$

Las variables cuantitativas contribuyen a esta distancia exactamente de la misma manera que en un ACP sobre estas variables por sí solas; las variables cualitativas contribuyen a esta distancia (hasta el coeficiente $1/Q$) como lo hacen en el ACM de estas variables por sí solas. Un caso importante específico es el de la distancia entre un individuo y el centro de gravedad de la nube. Este centro de gravedad se encuentra en el origen O cuando las variables están centradas, del mismo modo para las variables cuantitativas. Para los indicadores codificados por el ACM, que representan la división por $\mathrm{p_{k_q}}$, la media de la columna k_q vale 1. Finalmente, obtenemos:

$$d^2(i,O) = \sum_{k \in K_1} x_{ik}^2 + \sum_{q \in Q} \frac{1 - p_q(i)}{p_q(i)}$$

donde denotamos $q(i)$ la categoría de la variable q poseída por i, y $\mathrm{p_q(i)}$ la proporción asociada con $q(i)$.

Es necesario garantizar el equilibrio entre la influencia de los dos tipos de variables en estas relaciones. Es natural medir la influencia de una variable por su contribución a la inercia de todos los puntos. Las consideraciones establecidas en $\mathbb{R}^{\mathrm{I}}$ son transpuestas en $\mathbb{R}^{\mathrm{K}}$ por dualidad. Particularmente, en el subespacio de $\mathbb{R}^{\mathrm{K}}$ generado por las categorías K_q de la variable q, la proyección de la nube de individuos tiene una inercia de $K_q - 1$ distribuida isotrópicamente (igualdad) en todas las direcciones de este subespacio de dimensión $K_q - 1$.

Como en todos los análisis factoriales se representa:

- La nube de individuos por su proyección sobre sus ejes de inercia (denotamos $F_s(i)$ la proyección del individuo i sobre el eje de rango s).
- Las variables cuantitativas por su coeficiente de correlación con los factores $\mathrm{F_s}$.
- Las categorías de variables cualitativas por los centros de gravedad de los individuos correspondientes, denotamos con $F_s(k_q)$ la coordenada de la proyección en el eje de rango s del centro de gravedad de los individuos que poseen la categoría k de la variable q (Pagès, 2015, pp. 69-70).

2.5.4. Relaciones de transición

Aquí se aplica las fórmulas generales para el ACP.

Relaciones de $\mathbb{R}^K$ hacia $\mathbb{R}^I$

Que $G_s(k)$ sea la coordenada de la columna k en el eje de rango s.
Caso de una variable cuantitativa:

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i p_i x_{ik} F_s(i) = r(k, F_s)$$

Caso de una categoría k_q de variable q con una frecuencia relativa de p_{k_q}:

$$G_s(k_q) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{p_{k_q}}} \sum_i p_i y_{ik_q} F_s(i) = \frac{1}{\sqrt{\lambda_s}} F_s(k_q)$$

donde $F_s(k_q)$ es la coordenada a lo largo del eje de rango s del centro de gravedad de los individuos con categoría (k_q), al igual que en el ACM hasta el coeficiente $1/\sqrt{\lambda_s}$ la coordenada de una categoría como indicador (es decir, en $\mathbb{R}^I$) es igual a la del baricentro de los individuos que la poseen (en $\mathbb{R}^K$).

Relación de $\mathbb{R}^I$ hacia $\mathbb{R}^K$

Esta relación es fundamental en ACM ya que expresa la posición de un individuo según las categorías que posee. Rara vez es explícito en el ACP, pero subyace en la interpretación. Para el AFDM, se expresa:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K_1} x_{ik} G_s(k) + \frac{1}{\sqrt{\lambda_s}} \sum_{k_q \in K_2} p_{k_q} \left(\frac{y_{ik_q}}{p_{k_q}} - 1 \right) G_s(k_q)$$

El primer miembro es del ACP. Expresa que un individuo se encuentra del lado de las variables para las cuales tiene un valor superior al promedio, y las variables opuestas para las cuales tiene un valor inferior al promedio. El segundo miembro es del ACM, hasta el coeficiente $1/Q$. Se puede expresar de acuerdo con $F_s(k_q)$, gracias a la ecuación anterior que relaciona $G_s(k_q)$ con $F_s(k_q)$:

$$\frac{1}{\lambda_s} \sum_{k_q \in K_2} \left(y_{ik_q} - p_{k_q} \right) F_s(k_q) = \frac{1}{\lambda_s} \sum_{k_q \in K_2} y_{ik_q} F_s(k_q)$$

La última ecuación expresa que un individuo es hasta el coeficiente de λ_s, en el baricentro de las categorías que posee (con estas categorías en sí mismas los baricentros de los individuos) (Pagès, 2015, pp. 70-71).

Observación

En la relación de transición que expresa la coordenada de un individuo de acuerdo con los de las categorías, el coeficiente es:

- $\sqrt{\lambda_s}$ si las categorías están representadas por la proyección de los indicadores (en $\mathbb{R}^I$)
- λ_s si las categorías están representadas por los centros de gravedad de los individuos que poseen la misma categoría (en $\mathbb{R}^K$)

Finalmente, un individuo se encuentra tanto en el lado de las variables cuantitativas para el cual tiene un alto valor, como en el lado de las categorías que posee (Pagès, 2015, pp. 71-72).

2.6. Análisis de Regresión Lineal Múltiple

De acuerdo a Ximénez y San Martín en el año 2013, la regresión simple (RS) explica los valores que toma la variable dependiente (Y_i) a partir de los de una sola variable independiente (X_i). La regresión múltiple (RM) tiene por objeto combinar p variables independientes $(X_1, X_2, \ldots, X_p)$ de tal modo que pronostiquen con la mayor precisión los valores que toma la variable dependiente (Y). La RM permite analizar tanto las contribuciones individuales como las colectivas del conjunto de variables independientes en los cambios que se producen en la variable dependiente (Ximénez y San Martín, 2013, p. 49).

2.6.1. El modelo lineal general

Un modelo es una afirmación algebraica sobre cómo se relacionan dos o más variables. Los modelos lineales establecen una hipótesis sobre la relación entre dos tipos de variables: las dependientes y las independientes. La estructura de la relación entre ambas constituye su forma funcional, que incluye la relación entre las principales variables, el tipo de distribución de probabilidad de las variables aleatorias y los parámetros de las ecuaciones del modelo.

Expresado formalmente, si Y_i es la medida en la variable dependiente para el sujeto i, el modelo lineal descompone las puntuaciones en Y_i como el resultado de la suma ponderada de los siguientes componentes:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

donde $X_{1i} + X_{2i} + \cdots + X_{pi}$ son las p variables independientes incluidas en el modelo para explicar el comportamiento de la variable dependiente. Se consideran variables fijas. Los $\beta_1 + \beta_2 + \cdots + \beta_p$ son los p parámetros que se necesita estimar para decidir sobre la importancia de cada una de las variables presentes en la ecuación. $\beta_0 X_{0i}$ representa el conjunto de efectos debidos a variables mantenidas constantes (donde X_{0i} toma el valor 1 para todos los sujetos). Por último, ε_i es el efecto debido al conjunto de variables no incluidas en el modelo. Se denomina error aleatorio y se supone varía aleatoriamente con media 0 y varianza σ^2.

Según estas especificaciones, el modelo lineal general asume que hay n observaciones en p variables no correlacionadas tal que:

$$E(Y) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

$$Var(Y) = \sigma^2$$

Si se consideran p variables independientes ($X_1 + X_2 + \cdots + X_p$), el modelo de regresión para predecir los valores de la variable dependiente Y_i en n ensayos es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

que puede ser escrita en matrices y sistema de ecuaciones como se observa a continuación:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_p X_{p1} + \varepsilon_1 \\ \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_p X_{p2} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_p X_{pn} + \varepsilon_n \end{bmatrix};$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \ddots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \vdots & X_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

de modo más compacto en forma matricial es:

$$\mathbf{Y}_{\boldsymbol{n}\times 1} = \mathbf{X}^*_{\mathbf{n}\times(\mathrm{p}+1)} \boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\varepsilon}_{(n\times 1)}$$

donde **Y** es un vector columna n dimensional, $\mathbf{X}^*$ es una matriz $\mathbf{n} \times (\mathrm{p} + 1)$, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión a ser estimados, su dimensión es p y $\boldsymbol{\varepsilon}$ es un vector columna aleatorio de dimensión n.

Además de los ya mencionados, otros supuestos del modelo lineal general son los siguientes:

a) $E(\varepsilon) = E \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

b) $E(\varepsilon\varepsilon') = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$

c) $Cov(\boldsymbol{\varepsilon}, \mathbf{X}) = 0$

d) $r(\mathbf{X}) = p\ \left(no\ multicolinealidad\ =\ Las\ X_j\ son\ independientes\right)$

e) Adicionalmente, puede asumirse que $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, aunque no es imprescindible. Si se asume, puede utilizarse el método de estimación de máxima verosimilitud y llevarse a cabo las pruebas de significación (Ximénez y San Martín, 2013, pp. 49-51).

2.6.2. Estimación de parámetros

El modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$ puede expresarte mediante:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi} + e_i$$

donde $b_0 + b_1 + b_2 + \cdots + b_p$ son los estimadores de los parámetros $\beta_0 + \beta_1 + \beta_2 + \cdots + \beta_p$ y e_i es el estimador de ε_i.

De modo más compacto la ecuación estimada puede expresarse mediante:

$$\mathbf{Y} = \mathbf{X}^*\mathbf{b} + \mathbf{e}$$

Existen diferentes métodos para estimar los parámetros. Aquí se expone el más utilizado, el de mínimos cuadrados, aunque también se comenta brevemente el de máxima verosimilitud.

- **Método de estimación por mínimos cuadrados**

Con el cálculo de los estimadores de los parámetros se pretende estimar la ecuación de regresión que mejor se ajusta a los datos empíricos. El procedimiento matemático para estimar dicha ecuación consiste en calcular la recta (en regresión simple) o el plano (en regresión múltiple) cuya distancia vertical a los distintos valores de **Y** sea mínima.

Si $\widehat{\mathbf{Y}}$ es el valor predicho mediante las variables independientes en **Y**, se tiene que:

En puntuaciones directas: $\mathbf{Y} = \mathbf{X}^*\mathbf{b} + \mathbf{e}$; $\widehat{\mathbf{Y}} = \mathbf{X}^*\mathbf{b}$

En puntuaciones diferenciales: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$; $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$

En puntuaciones típicas: $\mathbf{z_y} = \mathbf{z_x}\mathbf{b}^* + \mathbf{e}^*$; $\hat{\mathbf{z}}_\mathbf{y} = \mathbf{z_x}\mathbf{b}^*$

El error obtenido en el pronóstico será:

En puntuaciones directas: $\boldsymbol{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}^*\mathbf{b}$

En puntuaciones diferenciales: $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{X}\mathbf{b}$

En puntuaciones típicas: $\boldsymbol{e} = \mathbf{z_y} - \hat{\mathbf{z}}_\mathbf{y} = \mathbf{z_y} - \mathbf{z_x}\mathbf{b}^*$

La mejor predicción de las **Y** a partir de las $\widehat{\mathbf{Y}}$ es aquella en que el valor de los errores sea lo más pequeño posible. Aplicando el método de mínimos cuadrados:

$$min: \sum e^2 = \mathbf{e'e}$$

La expresión que permite calcular el vector de parámetros de la ecuación de regresión que hace mínima $\mathbf{e'e}$ es:

En puntuaciones directas: $\mathbf{b} = (\mathbf{X^{*\prime}X^{*}})^{-1}\mathbf{X^{*\prime}Y}$

En puntuaciones diferenciales: $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$

En puntuaciones típicas: $\mathbf{b} = (\mathbf{Z'Z})^{-1}\mathbf{Z'Z_y} = (\mathbf{R_{xx}})^{-1}\mathbf{R_{xy}}$

Propiedades de los estimadores.

1. La estimación de los elementos del vector **b** es lineal, insesgada y eficiente, es decir:

$$E(\mathbf{b}) = \boldsymbol{\beta}$$

$$Var(\mathbf{b}) = \sigma^2(\mathbf{X'X})^{-1} < Var(\mathbf{c})$$

2. Este método de estimación no exige normalidad, es decir, los ε_i pueden tener cualquier tipo de distribución con $E(\boldsymbol{\varepsilon}) = 0$ y $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$
3. $cov(\varepsilon_i, \varepsilon_j) = 0$ que indica independencia entre los errores o ausencia de autocorrelación, es decir, el error que se comete en $\boldsymbol{i}$ no debe tener ninguna relación con el que se comete en j.
4. $cov(\varepsilon_i, X_j) = 0$. Los errores deben ser aleatorios, no debe haber errores sistemáticos.
5. $cov(\varepsilon_i, \hat{y}_j) = 0$ (independencia).

- **Método de estimación de máxima verosimilitud**

El anterior procedimiento es válido independientemente de la distribución de los errores. Si se asume que los errores son normales el modelo de regresión viene dado por:

$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$,

$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde, $r(\boldsymbol{\Sigma}) = p$ lo que implica que $r(\mathbf{X}) = p$.

En este caso se puede utilizar el método de estimación de máxima verosimilitud. Se trata de estimar los valores del vector de parámetros $\boldsymbol{\beta}$ que hagan más probable el valor de los datos observados.

Asumiendo normalidad:

$$f(Y|X) = \frac{1}{\sigma_Y\sqrt{2\pi}\sqrt{1-\rho^2}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)\right)\right]^2}$$

Los estimadores máximos verosímiles se obtienen maximizando la función de verosimilitud:

$$L = \prod_{i=1}^{n} f(Y|X) = \left(\frac{1}{\sigma_Y^2 2\pi(1-\rho^2)}\right)^{n/2} e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\sum_{i=1}^{n}\left[y_i-\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)\right)\right]^2}$$

En la práctica se toman logaritmos pues queda una expresión más sencilla:

$$\log L = \frac{n}{2}\log\left(\frac{1}{\sigma_Y^2 2\pi(1-\rho^2)}\right) - \frac{1}{2\sigma_Y^2(1-\rho^2)}\sum_{i=1}^{n}\left[y_i - \left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)\right)\right]^2$$

Para obtener el estimador máximo verosímil de $\boldsymbol{\beta}$ se iguala la primera derivada de $\ln L$ a cero. Mediante este procedimiento se llega a lo siguiente:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

En el caso de que la variable Y sea normal la estimación por mínimos cuadrados y máxima verosimilitud proporcionan resultados idénticos (Ximénez y San Martín, 2013, pp. 51-57).

2.6.3. Verificación del modelo

Una vez estimado el modelo hay que valorar si constituye una buena o mala aproximación a nuestro conjunto de datos. Es decir, cabe preguntarse: ¿En qué medida es posible predecir los valores de **Y** a partir de los de **X** con el modelo? Una representación gráfica de los datos empíricos y el modelo estimado puede proporcionar una primera aproximación al problema de la verificación del modelo.

Asimismo, hay que valorar en qué medida el modelo se ajusta a los datos empíricos y la contribución de las variables independientes en los cambios que se producen en la variable dependiente. A todo esto, se le denomina bondad de ajuste.

- **Medidas de bondad de ajuste**

1. Descomposición de la varianza

Una parte de la variación de los datos puede explicarse mediante el modelo de regresión ($\hat{\mathbf{y}}$). Sin embargo hay otra parte que queda sin explicar ($\mathbf{e}$), es decir:

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$$

calculando la suma de cuadrados de **y**

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

en términos de análisis de varianza (o ANOVA):

$$SCT = \sum y_i^2 = \mathbf{y}'\mathbf{y}$$

$$SCT = \sum \hat{y}_i^2 = \hat{\mathbf{y}}'\hat{\mathbf{y}} = (\mathbf{X}\mathbf{b})'(\mathbf{X}\mathbf{b}) = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

$$SCE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

2. Coeficiente de determinación

Informa sobre el grado de ajuste de los puntos a la recta o al plano de regresión. Es la bondad del modelo de regresión y se calcula mediante el índice estadístico R^2:

Si: $R_{y(x_1,x_2,\ldots,x_p)} = r_{y\hat{y}} = \frac{\sum y_i \hat{y}_i}{\sqrt{\sum y_i^2}\sqrt{\sum \hat{y}_i^2}} = \frac{\mathbf{y}'\hat{\mathbf{y}}}{\sqrt{\mathbf{y}'\mathbf{y}}\sqrt{\hat{\mathbf{y}}'\hat{\mathbf{y}}}} = \frac{(\hat{\mathbf{y}}+\mathbf{e})'\hat{\mathbf{y}}}{\sqrt{(\mathbf{y}'\mathbf{y})(\hat{\mathbf{y}}'\hat{\mathbf{y}})}} = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\sqrt{(\mathbf{y}'\mathbf{y})(\hat{\mathbf{y}}'\hat{\mathbf{y}})}}$

entonces: $R^2 = \frac{(\hat{\mathbf{y}}'\hat{\mathbf{y}})^2}{(\mathbf{y}'\mathbf{y})(\hat{\mathbf{y}}'\hat{\mathbf{y}})} = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$

R^2 oscila entre 0 y 1 y es la proporción de varianza de Y que queda explicada por las X_j.

3. Coeficiente de determinación corregido

El coeficiente R^2 viene afectado por un cierto efecto inflacionista sobre el grado de ajuste. Esta inflación se origina en dos hechos: el tamaño muestral (n) y el número de predictores (p). Por tanto, es necesario introducir un factor corrector. El procedimiento consiste en corregir las sumas de cuadrados:

$$\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-p)}{\mathbf{y}'\mathbf{y}/(n-1)} = 1 - \frac{SCE/(n-p)}{SCT/(n-1)} = 1 - \frac{SCE}{SCT}\frac{n-1}{n-p}$$

Con lo que se llega a: $\bar{R}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$

Para $p > 1$, $\bar{R}^2 < R^2$ y esta diferencia aumenta a medida que aumenta también el número de variables independientes. Si el modelo no incluye el término b_0, el numerador es n en lugar de $n-1$ y $\bar{R}^2$ puede ser menor que 0, cosa que nunca puede ocurrir con R^2. En regresión múltiple es más apropiado utilizar $\bar{R}^2$, sobre todo si el tamaño muestral es pequeño y si se desea comparar distintos modelos para pronosticar los valores de una misma variable dependiente.

➢ Contraste de hipótesis

Los coeficientes obtenidos en la ecuación de regresión son estimadores de los parámetros del modelo. Por ello es necesario realizar una prueba de significación para contrastar si su valor es 0 en la población y calcular los intervalos de confianza de los coeficientes de la regresión. Pueden llevarse a cabo tres tipos de contrastes, para lo cual es necesario que los errores se distribuyan normalmente con media $\mathbf{0}$ y varianza $\sigma^2\mathbf{I}$:

1. $\mathbf{H_0}: \boldsymbol{\beta_0} = \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \cdots = \boldsymbol{\beta_j} = \cdots = \boldsymbol{\beta_p} = \mathbf{0}$ o bien $\mathbf{H_0}: \boldsymbol{\beta} = [\mathbf{0}]$

Una de las hipótesis a contrastar es si los elementos del vector $\boldsymbol{\beta}$ son nulos, es decir, la hipótesis sobre linealidad. Para ello se calcula el estadístico F utilizando el formato ANOVA:

FV	SC	gl	MC	F	$F \sim F_{p,(n-p-1)}$
Regresión	$\hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$	p	$\hat{\mathbf{y}}'\hat{\mathbf{y}}/p$	MCR/MCE	
Error	$\mathbf{e}'\mathbf{e}$	$n-p-1$	$\mathbf{e}'\mathbf{e}/(n-p-1)$		
Total	$\mathbf{y}'\mathbf{y}$	$n-1$			

2. $\boldsymbol{H_0}: \boldsymbol{\rho_m} = \mathbf{0}$ (Correlación múltiple)

Otra forma de determinar si existe relación lineal es si el coeficiente de determinación (R^2) es significativo. Se calcula el estadístico F:

$$F = \frac{n-p-1}{p}\frac{R^2}{1-R^2} \sim F_{p,(n-p-1)}$$

3. $H_0: \beta_j = 0$

Las anteriores pruebas de significación son un indicador de la bondad de ajuste global del modelo. Para comprobar la significación de cada uno de los coeficientes b_j se calcula el estadístico T:

$$T = \frac{b_j}{\hat{\sigma}\sqrt{c_{ii}}} \sim t_{n-p-1}$$

donde $\hat{\sigma}^2 = \frac{\mathbf{e'e}}{n-p-1} = MCE$; c_{ii} = i-ésimo elemento de la matriz $(\mathbf{X'X})^{-1}$.

Dado el valor de b_j también se puede estimar el intervalo de confianza de su verdadero valor en la población mediante: $b_j \pm t_{1-\alpha/2;n-p-1}\hat{\sigma}\sqrt{c_{ii}}$ (Ximénez y San Martín, 2013, pp. 57-61).

2.6.4. Análisis del cumplimiento de los supuestos

Además de preguntarse si el modelo obtiene un buen ajuste, es necesario preguntarse: ¿Es el modelo correcto? Para que la respuesta sea afirmativa se requiere el cumplimiento de ciertas condiciones de aplicación: que la relación entre las variables independientes y la dependiente sea lineal, que los residuos sean independientes, homogéneos y normales, y que no haya colinealidad entre las variables independientes. A continuación, se comenta cada uno de estos supuestos y su procedimiento de comprobación.

1. Linealidad de la relación

La relación entre cada una de las variables independientes incluidas en el modelo y la variable dependiente ha de ser lineal. Los gráficos parciales entre cada variable independiente y la variable dependiente permiten detectar el tipo de relación entre ambas.

En regresión múltiple la representación gráfica de los residuos ayuda en esta detección. Hay que elaborar los diagramas de dispersión de los residuos que resultan de la regresión de cada variable independiente sobre las restantes y la regresión de la variable dependiente sobre la variable independiente.

Este supuesto puede incumplirse cuando se omiten variables independientes importantes, la relación entre éstas y la variable dependiente no es lineal, los parámetros no son constantes o se da aditividad, es decir, alguna variable independiente interactúa con otra. En estos casos se puede utilizar otro tipo de regresión diferente a la lineal o efectuar alguna transformación en las variables que permita linealizar el modelo.

2. Independencia

Los residuos se comportan como una variable aleatoria. Por tanto, han de ser independientes entre sí, de las variables independientes y de los pronósticos. En caso de no cumplirse este supuesto, se produce el problema de la autocorrelación.

La prueba de Durbin-Watson permite conocer el grado de independencia entre los residuos:

$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$ Donde: $0 \leq DW \leq 4$

si los residuos son independientes $DW = 2$. Se puede asumir independencia entre residuos si $1.50 \leq DW \leq 2.50$.

3. Homocedasticidad

La variación de los residuos debe ser uniforme a lo largo de los valores pronosticados ($\hat{y}_i$). Esto implica que el tamaño de los residuos es independiente del de los valores pronosticados.

Para comprobar el cumplimiento de este supuesto se elabora el diagrama de dispersión entre los pronósticos y los residuos tipificados y se comprueba que no existe relación lineal entre las variables.

4. Normalidad

Si se asume, para cada valor de la variable independiente, los residuos se distribuyen normalmente con media cero y varianza σ^2. Hay tres formas de comprobar este supuesto:

La primera elaborar el histograma de los residuos tipificados para observar el grado de alejamiento de su distribución con respecto a la distribución teórica normal.

La segunda, elaborar el gráfico P-P de probabilidad normal que permite comparar la probabilidad acumulada observada y la esperada según la curva normal. La discrepancia mayor o menor es un indicador del mayor o menor alejamiento de los residuos a la normalidad, este tipo de gráficos no son muy informativos a no ser que el tamaño muestral sea suficientemente grande ($n \geq 20$).

Por último, también se puede emplear el test de normalidad Kolmogorov-Smirnov y comprobar que no sea significativo.

5. Ausencia de colinealidad

Las variables independientes no deben tener correlaciones demasiado altas. Cuando se incumple este supuesto se dice que existe colinealidad.

La existencia de colinealidad entre las variables puede originar diversos problemas: si la colinealidad es perfecta, no se pueden estimar los coeficientes de la ecuación de regresión, si es parcial, aumenta el tamaño de los residuos tipificados y las estimaciones de los coeficientes son muy inestables y difíciles de interpretar.

Para detectar el problema de la colinealidad entre variables independientes se puede observar si se da alguno de los siguientes indicadores:

a) El estadístico F del modelo es significativo, pero ninguno de los coeficientes de regresión parcial lo es y los coeficientes de correlación son muy grandes.
b) Los coeficientes de regresión parcial tipificados están fuera del rango $1 < b^*_j < -1$.
c) Los valores de la tolerancia de las X_j, que se calculan mediante la expresión: $1 - R^2_{j(1,2,\ldots,p)}$, son menores de 0.01 y los factores de inflación de la varianza (FIV), los inversos de la tolerancia, son grandes.
d) En el análisis de componentes principales realizado sobre la matriz estandarizada de productos cruzados entre las variables independientes hay varios autovalores próximos a cero. Un componente explica mucha varianza de los coeficientes de dos o más variables.

Si se detecta la existencia de colinealidad, para corregirla, se puede aumentar el tamaño muestral, generar nuevas variables en base a combinaciones lineales de las variables altamente correlacionadas, o bien utilizar un procedimiento jerárquico a la hora de incluir las variables en la ecuación. Esta es una forma de selección de variables que permite elegir sólo aquellas que expliquen una parte de varianza distinta a la de las variables ya incluidas en el modelo (Ximénez y San Martín, 2013, pp. 61-66).

2.6.5. Regresión con variables dummy: variables categóricas

Cuando una variable cualitativa asume solamente dos valores es llamada variable indicadora, variable binaria o variable "dummy". Estas variables son codificadas numéricamente con 0's y 1's. En un problema de regresión debe haber por lo menos una variable predictora cuantitativa. Si todas las variables predictoras fueran cualitativas entonces el problema se convierte en uno de diseños experimentales (Acuña, 2007, p. 108).

Los autores Arcarons y Calonge indican que la información cualitativa se introduce en el modelo de regresión lineal múltiple mediante la utilización de las variables ficticias, también conocidas como variables dummy (ies), y captan el electo de una determinada característica o atributo de los datos sobre la variable dependiente, que no puede interpretarse o medirse cuantitativamente (Arcarons y Calonge, 2007, p. 111).

2.6.5.1. Construcción de las variables dummy

De acuerdo a Rodríguez y González, para estructurar las variables dummy, es indispensable tener en cuenta que:

- Identifique y establezca cuales son las variables cuantitativas y cualitativas.

- Determine si desea realizar un análisis con variables de tipo dicotómicas, tricomicas o policotomicas. Usualmente en éstos estudios se usan análisis con variables de tipo dicotómicas (0,1) para determinar las respuestas, independiente del número de sub variables que existan. Estas variables tienen características de ser excluyentes con las demás categorías.
- Para adecuar que características corresponden a qué valor, la forma más útil de establecer estos valores es usar el azar o usar la conveniencia de la respuesta, por ejemplo, si se tiene una variable cualitativa de género, a quien se le establece el valor 0 o 1, ¿al hombre o a la mujer"; esto depende estrictamente de lo que el investigador desee averiguar.
- Cree o establezca las variables que conoceremos como dummy que vendrán de la variable cualitativa original. Si una variable cualitativa como el estrato social tiene seis estratos, entonces se crearán seis nuevas variables que representarán todos los estratos sociales.
- Finalmente, para asignar los valores a las características, use procesos dicotómicos a modo de switch, es decir, el encendido es 1 y el apagado es 0. Recuerde que este sistema es totalmente excluyente, si una característica esta, activa o encendida, el resto estarán apagadas (Rodríguez y González, 2017, pp. 53-54).

2.6.5.2. El modelo de regresión con una sola variable cualitativa

Las variables ficticias toman generalmente el valor 1 o 0 en función que se dé o no la característica que se pretende recoger,

$$D_i = \begin{cases} 1 & si\ se\ cumple\ la\ característica \\ 0 & en\ caso\ contrario \end{cases}$$

consideremos un modelo de regresión con una sola variable cualitativa D y una variable cuantitativa X, es decir:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \mathrm{D} + \varepsilon_i$$

Si $\mathrm{D} = 0$ se obtiene el modelo lineal simple

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

Si $\mathrm{D} = 1$ se obtiene el modelo

$$Y_i = (\beta_0 + \beta_2) + \beta_1 X_{1i} + \varepsilon_i$$

Notar que las líneas estimadas de los modelos $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$ y $Y_i = (\beta_0 + \beta_2) + \beta_1 X_{1i} + \varepsilon_i$ serán paralelas (igual pendiente). El valor estimado de β_2 representa el cambio promedio en la variable de respuesta al cambiar el valor de la variable "dummy" (Acuña, 2007, pp. 108-109).

2.6.5.3. El modelo de regresión con múltiples variables cualitativas

Un factor politómico se puede ingresar en una regresión codificando un conjunto de 0/1 regresores ficticios, se analiza el número de categorías del factor menos uno (evitar la trampa de las variables

ficticias). La categoría "omitido", codificada como 0 para todos los regresores ficticios en el conjunto, sirve como base para comparar las demás categorías. El modelo representa superficies de regresión paralelas, una para cada categoría del factor (Fox, 2016, p. 136).

La forma general del modelo con una variable independiente cualitativa en m niveles (Sahay, 2016, pp. 153-154):

$$y = b_0 + b_1 x_{11} + b_2 x_{22} + \cdots + b_{m-1} x_{m-1}$$

donde x_i es la variable ficticia para el nivel (i + 1) y

$$x_i = \begin{cases} 1 & si\ y\ es\ observada\ en\ el\ nivel\ (i+1) \\ 0 & en\ caso\ contrario \end{cases}$$

Siguiendo el sistema de codificación tenemos:

$$\begin{aligned} \mu_A &= b_0 \\ \mu_B &= b_0 + b_1 \\ \mu_C &= b_0 + b_2 \\ \mu_D &= b_0 + b_3 \\ &\vdots \end{aligned}$$

y

$$\begin{aligned} b_1 &= \mu_B - \mu_A \\ b_2 &= \mu_C - \mu_A \\ b_3 &= \mu_D - \mu_A \\ &\vdots \end{aligned}$$

Una vez que se crea un conjunto de variables ficticias, en forma general, para m niveles de variables cualitativas, necesitamos $(m - 1)$ variables ficticias. Tenemos que dejar uno de los niveles fuera del modelo de regresión para evitar la multicolinealidad perfecta (también conocida como singularidad o redundancia).

La generalización de los resultados anteriores conduce a combinar en la matriz de regresores: variables cuantitativas, variables exógenas habituales, otros factores con un patrón no cuantitativo y las variables ficticias. Ello da lugar a tres distintas especificaciones conocidas como: el modelo aditivo, el modelo multiplicativo y el modelo mixto (Arcarons y Calonge, 2007, p. 117).

CAPITULO III

3. MARCO METODOLÓGICO

3.1. Hipótesis general

Existen factores en la base de datos de la ESPAC y estos influyen sobre la producción del plátano en el Ecuador para los años 2014, 2015 y 2016.

3.2. Identificación de variables

Las variables que se estudiaron en esta investigación (variables con datos completos) y consideradas por la ESPAC son en total 20 variables, de las cuales 10 son numéricas y 10 son categóricas (Ver Tabla 1-3). Se aplicó la técnica del AFDM y ARLMVD utilizando el software libre R versión 3.4.2.

3.3. Población y muestra

La ESPAC, utiliza la metodología del muestreo de marcos múltiples (MMM), que consiste en la combinación del muestreo de marco de áreas (MMA) con el marco de lista (MML), este método estadístico se lleva a cabo con el fin de seleccionar unidades de investigación a partir del MMA y MML.

La metodología del marco de áreas (MA) es un procedimiento estadístico que contempla la segmentación de la superficie total del país por estratos basados en intensidad de actividad agropecuaria, los cuales son divididos en Segmentos de Muestreo (SM), cuya superficie varía de acuerdo al estrato.

El marco de lista (ML) es un directorio preparado por el INEC, en donde constan las principales explotaciones dedicadas a un determinado cultivo, los que son investigados con el fin de mejorar la calidad de las estimaciones.

La unidad de observación son todos los terrenos que se encuentran en los segmentos seleccionados para la etapa de campo. Los segmentos muestrales tienen límites geométricos,

puesto que fueron construidos en función a los mapas de uso y cobertura de suelo. A partir del marco de áreas se selecciona una muestra aleatoria de segmentos.

3.4. Recolección de Información

Los datos para el estudio se obtuvieron dentro la página web sobre estadísticas agropecuarias en el INEC, es decir, se trabajó con información secundaria de acceso público, comprendidos entre el año 2014 – 2016.

3.5. Operacionalización de variables

Siguiendo la estructura de la ESPAC y basado en la necesidad de este estudio, para conseguir los objetivos planteados de la investigación se trabajó con las variables presentadas en la Tabla 1-3.

Tabla 1-3: Cuadro categórico de variables

Ítem	Campo codificada por la ESPAC	Variable	Descripción	Tipo	Escala	Categoría o Intervalo
1	cp_prod	Producción	Es la cantidad de frutos cosechados en un tiempo determinado de acuerdo al ciclo de producción del plátano en toneladas métricas.	Cuantitativa: Continua	Razón	[0.23,2386]
2	cp_k406	Edad de la plantación	Es la edad (años) de la plantación del plátano.	Cuantitativa: Continua	Intervalo	[1,99]
3	cp_k409h	Superficie plantada	Es la superficie en que ocupan las plantas con distancias establecidas en hectáreas.	Cuantitativa: Continua	Razón	[0.01;280]
4	cp_k410h	Superficie en edad productiva	El cultivo del plátano requiere alcanzar cierta edad para entrar en el período de producción. La superficie (en hectárea) de esta variable, corresponde a aquella que ocupan las plantas que han alcanzado esta edad.	Cuantitativa: Continua	Razón	[0.01;280]
5	cp_k411h	Superficie cosechada	Es la superficie del plátano que está lista para la recolección o cosecha en hectáreas.	Cuantitativa: Continua	Razón	[0.01;280]
6	cp_k416	Cantidad de producción	Es la cantidad de frutos cosechados en un tiempo determinado de acuerdo al ciclo de producción del plátano.	Cuantitativa: Discreta		[20,175000]
7	cp_k418	Cantidad de producción en libras	Es la cantidad de frutos cosechados en un tiempo determinado de acuerdo al ciclo de producción del plátano equivalente en libras.	Cuantitativa: Continua	Razón	[2.2,2200]

8	cp_k422	Cantidad vendida	Es la cantidad de frutos vendida de la producción del plátano.	Cuantitativa: Discreta		[2,173290]
9	cp_k424	Cantidad vendida en libras	Es la cantidad de frutos vendida de la producción del plátano equivalente en libras.	Cuantitativa: Continua	Razón	[2.20;2200]
10	cp_vent	Ventas	Es la cantidad de frutos vendida de la producción del plátano en toneladas métricas.	Cuantitativa: Continua	Razón	[0.02;2363.05]
11	cp_k404	Condición de cultivo	Es la condición del cultivo para el plátano que han sido sembrados.	Cualitativa: Politómica	Nominal	1: Sólo 2: Asociado 3: Invernadero
12	cp_k408	Semilla de más uso	Plantas procreadas o multiplicadas por dos individuos de distintas especies, es decir el resultado de todo lo que es producto de especies o variedades distintas.	Cualitativa: Politómica	Nominal	1: Común 2: Mejorada 3: Hibrida nacional 4: Hibrida internacional
13	cp_k413	Uso de riego	La utilización o no de riego.	Cualitativa: Dicotómica	Nominal	1: Si 2: No
14	cp_k414	Uso de fertilizantes	La utilización o no de fertilizantes.	Cualitativa: Dicotómica	Nominal	1: Si 2: No
15	cp_k415	Uso de fitosanitarios	La aplicación de insecticidas, fungicidas y control biológico.	Cualitativa: Dicotómica	Nominal	1: Si 2: No
16	cp_forg	Uso de fertilizante orgánico	La utilización o no de fertilizante orgánico.	Cualitativa: Dicotómica	Nominal	1: Si 2: No
17	cp_fqui	Uso de fertilizante químico	La utilización o no de fertilizante químico.	Cualitativa: Dicotómica	Nominal	1: Si 2: No
18	cp_porg	Uso de plaguicida orgánico	La utilización o no de plaguicida orgánico.	Cualitativa: Dicotómica	Nominal	1: Si 2: No
19	cp_pqui	Uso de plaguicida químico	La utilización o no de plaguicida químico.	Cualitativa: Dicotómica	Nominal	1: Si 2: No
20	part_prov	Participación provincial	Es el intervalo porcentual de participación provincia de la producción nacional.	Cualitativa: Politómica	Ordinal	1: Sin producción 2: Menos del 20% 3: [20%;40%) (Manabí 2016) 4: [40%;60%) (Manabí 2014 y 2015) 5: Más del 60%

Fuente: ESPAC, INEC
Realizado por: Guamán Eduardo 2018

3.6. Alcances de la Investigación

El estudio tiene alcance Descriptiva y Correlacional Causal. Descriptiva porque identifica y describe los factores influyentes en la producción del plátano, y Correlacional/Causal porque describe relaciones de tipo causa efecto.

3.7. Análisis de datos

El presente estudio se realizó utilizando el software estadístico R versión 3.4.2 y la ayuda de la hoja de cálculo Excel 2016. Es importante señalar que las variables: cantidad de producción (cp_k416), cantidad de producción en libras (cp_k418), cantidad vendida (cp_k422), cantidad vendida en libras (cp_k424) y uso de fertilizantes (cp_k414) no se ingresaron al análisis debido a que por su definición comparten información con la variable producción (cp_prod), ventas (cp_vent), uso de fertilizante orgánico (cp_forg) y uso de fertilizante químico (cp_fqui).

CAPITULO IV

4. RESULTADOS Y DISCUSIÓN

4.1. Análisis exploratorio de datos

Tabla 1-4: Distribución estadística de frecuencia (D.e.f.) de la variable producción del plátano (cp_prod)

Límite inferior	Límite superior	Marca de clase	n_i	N_i	f_i (%)	F_i (%)
0.23	183	91.62	3204	3204	97.98	97.98
184	367	275.50	48	3252	1.47	99.45
368	551	459.50	7	3259	0.21	99.66
552	735	643.50	6	3265	0.18	99.85
736	919	827.50	3	3268	0.09	99.94
920	1099	1009.50	1	3269	0.03	99.97
2210	2389	2299.50	1	3270	0.03	100

Realizado por: Eduardo Guamán 2018

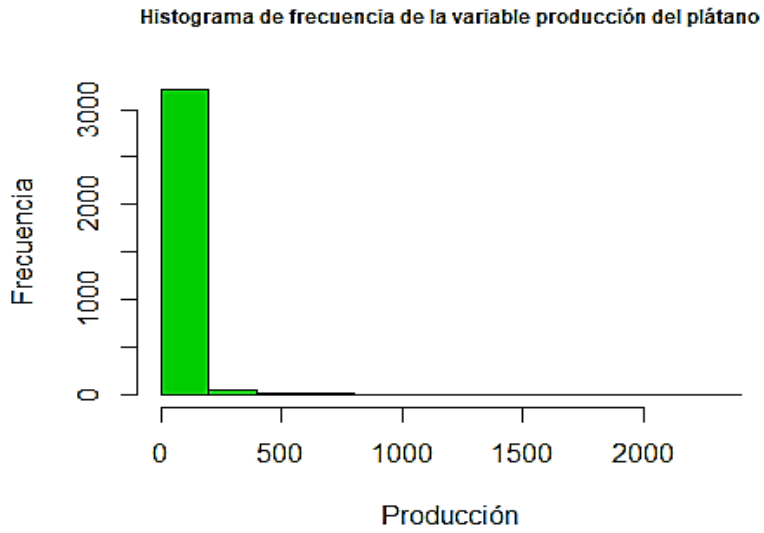


Gráfico 1-4: Histograma de frecuencia de la variable producción del plátano
Realizado por: Eduardo Guamán 2018

El 97.98% de los terrenos tuvieron una producción del plátano entre 0.23 y 183 toneladas métricas, mientras que el 2.01% de los terrenos presentaron una producción de plátano entre 184 y 2389 toneladas métricas.

Tabla 2-4: D.e.f. de la variable edad de la plantación (cp_k406)

Límite inferior	Límite superior	Marca de clase	n_i	N_i	f_i (%)	F_i (%)
1	8	4.50	2184	2184	66.79	66.79
9	16	12.50	642	2826	19.63	86.42
17	24	20.50	189	3015	5.78	92.20
25	32	28.50	142	3157	4.34	96.54

33	40	36.50	66	3223	2.02	98.56
41	48	44.50	13	3236	0.4	98.96
49	56	52.50	24	3260	0.73	99.69
57	64	60.50	3	3263	0.09	99.79
65	72	68.50	1	3264	0.03	99.82
97	104	100.50	6	3270	0.18	100

Realizado por: Eduardo Guamán 2018

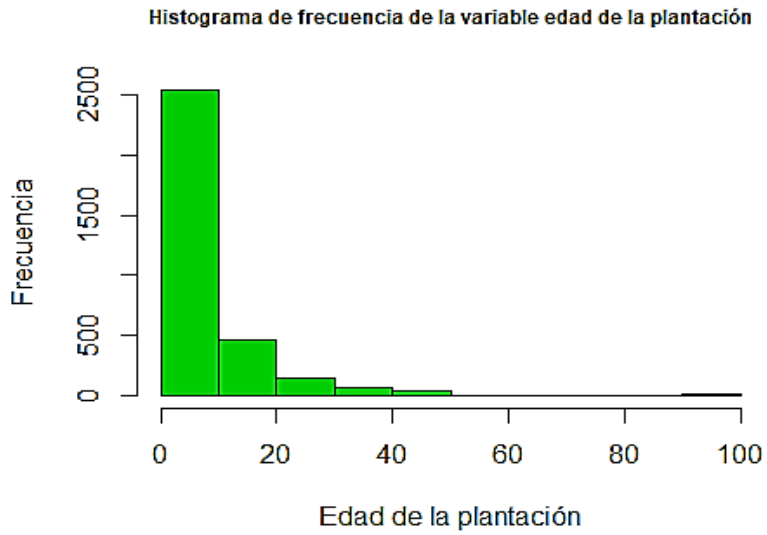


Gráfico 2-4: Histograma de frecuencia de la variable edad de la plantación
Realizado por: Eduardo Guamán 2018

El 66.79% de los terrenos productores del plátano tuvieron una edad de la plantación entre 1 y 8 años, el 19.63% de los terrenos productores del plátano presentaron una edad intermedia de la plantación entre 9 y 16 años, finalmente el 13.57% de los terrenos productores del plátano tuvieron una edad alta de la plantación entre 17 y 104 años.

Tabla 3-4: D.e.f. de la variable superficie plantada (cp_k409h)

Límite inferior	Límite superior	Marca de clase	n_i	N_i	f_i (%)	F_i (%)
0.01	21	10.51	3213	3213	98.26	98.26
22	43	32.50	38	3251	1.16	99.42
44	65	54.50	12	3263	0.37	99.79
66	87	76.50	2	3265	0.06	99.85
88	109	98.50	2	3267	0.06	99.91
154	175	164.50	1	3268	0.03	99.94
264	285	274.50	2	3270	0.06	100

Realizado por: Eduardo Guamán 2018

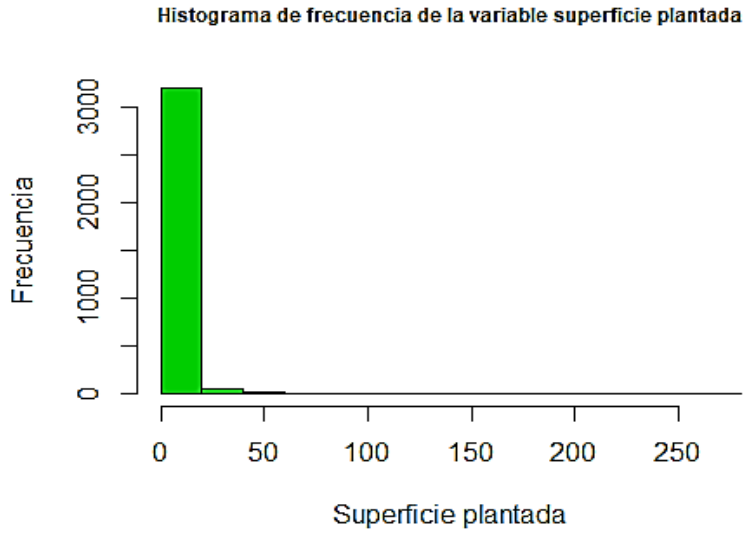


Gráfico 3-4: Histograma de frecuencia de la variable superficie plantada
Realizado por: Eduardo Guamán 2018

El 98.26% de los terrenos presentaron una superficie plantada del plátano entre 0.01 y 21 hectáreas, mientras que tan solo el 1.74% de los terrenos tuvieron una superficie plantada del plátano entre 22 y 285 hectáreas.

Tabla 4-4: D.e.f. de la variable superficie en edad productiva (cp_k410h)

Límite inferior	Límite superior	Marca de clase	n_i	N_i	f_i (%)	F_i (%)
0.01	21	10.51	3219	3219	98.44	98.44
22	43	32.50	34	3253	1.04	99.48
44	65	54.50	11	3264	0.34	99.82
66	87	76.50	1	3265	0.03	99.85
88	109	98.50	2	3267	0.06	99.91
154	175	164.50	1	3268	0.03	99.94
264	285	274.50	2	3270	0.06	100

Realizado por: Eduardo Guamán 2018

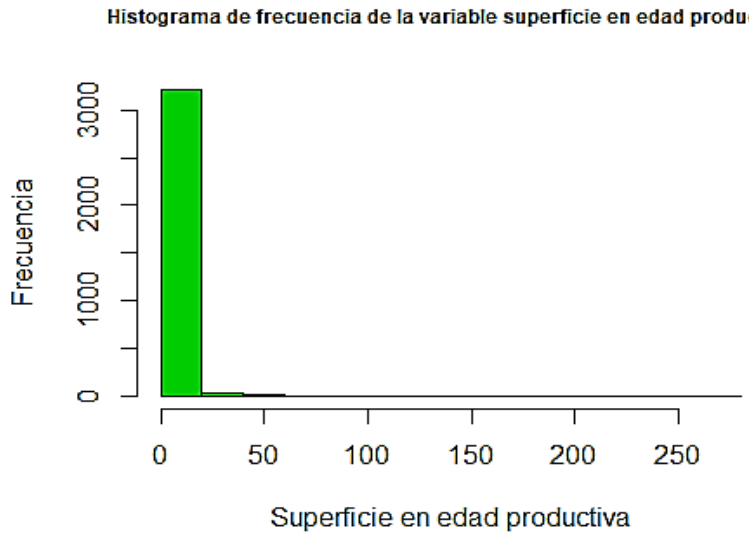


Gráfico 4-4: Histograma de frecuencia de la variable superficie en edad productiva
Realizado por: Eduardo Guamán 2018

El 98.44% de los terrenos con producción del plátano contaron con una superficie en edad productiva entre 0.01 y 21 hectáreas, mientras que tan solo el 1.56% de los terrenos con producción del plátano presentaron una superficie en edad productiva entre 22 y 285 hectáreas.

Tabla 5-4: D.e.f.de la variable superficie cosechada (cp_k411h)

Límite inferior	Límite superior	Marca de clase	n_i	N_i	f_i (%)	F_i (%)
0.01	21	10.51	3221	3221	98.5	98.5
22	43	32.50	34	3255	1.04	99.54
44	65	54.50	10	3265	0.31	99.85
66	87	76.50	2	3267	0.06	99.91
88	109	98.50	1	3268	0.03	99.94
264	285	274.50	2	3270	0.06	100

Realizado por: Eduardo Guamán 2018

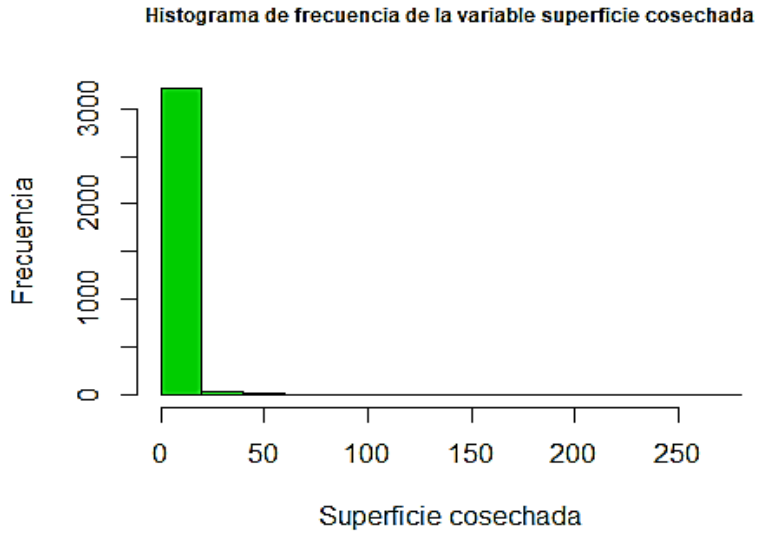


Gráfico 5-4: Histograma de frecuencia de la variable superficie cosechada
Realizado por: Eduardo Guamán 2018

El 98.5% de los terrenos tuvieron una superficie cosechada del plátano entre 0.01 y 21 hectáreas, mientras que tan solo el 1.5% de los terrenos tuvieron una superficie cosechada del plátano entre 22 y 285 hectáreas.

Tabla 6-4: D.e.f. de la variable ventas (cp_vent)

Límite inferior	Límite superior	Marca de clase	n_i	N_i	f_i (%)	F_i (%)
0.02	181	90.51	3206	3206	98.04	98.04
182	363	272.5	47	3253	1.44	99.48
364	545	454.5	7	3260	0.21	99.69
546	727	636.5	5	3265	0.15	99.85
728	909	818.5	3	3268	0.09	99.94
910	1089	999.5	1	3269	0.03	99.97
2180	2369	2274.5	1	3270	0.03	100

Realizado por: Eduardo Guamán 2018

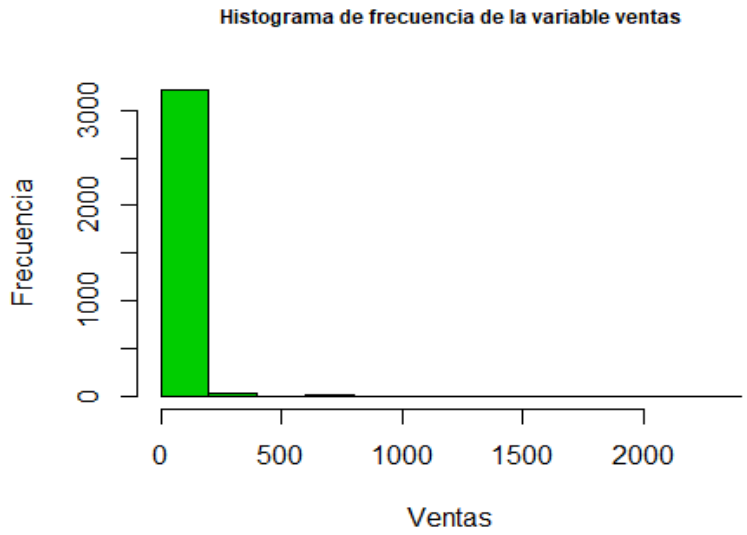


Gráfico 6-4: Histograma de frecuencia de la variable ventas
Realizado por: Eduardo Guamán 2018

El 98.4% de la cantidad cosechada del plátano en el Ecuador se ha vendido una cantidad entre 0.02 y 181 toneladas métricas, mientras que tan solo el 1.95% de la cantidad cosechada del plátano en el Ecuador se ha vendido una cantidad entre 182 y 2369 toneladas métricas.

Tabla 7-4: Resumen estadístico de las variables cuantitativas

Medidas		Producción (p_prod)	Edad de la plantación (cp_k406)	Superficie plantada (cp_k409h)	Superficie en edad productiva (cp_k410h)	Superficie cosechada (cp_k411h)	Ventas (cp_vent)
Medidas de tendencia central	Media	22.94	8.73	3.37	3.25	3.10	21.60
	Mediana	8.00	5.00	1.48	1.25	1.00	6.82
	Moda	3.00	2.00	1.00	1.00	1.00	2.73
Medidas de colocación	25%	4.00	3.00	1.00	1.00	1.00	3.25
	50%	8.00	5.00	1.48	1.25	1.00	6.82
	75%	17.04	10.00	3.00	3.00	3.00	15.99
Medidas de dispersión	Varianza	5065.93	102.29	89.12	85.09	76.28	4859.49
	Desviación típica	71.18	10.11	9.44	9.22	8.73	69.71
	Coeficiente de variación	3.10	1.16	2.80	2.84	2.82	3.23
Medidas de forma	Coeficiente de asimetría	15.71	3.05	17.70	18.71	20.23	16.14
	Coeficiente de curtosis	413.19	14.93	451.35	494.22	585.10	432.10

Realizado por: Eduardo Guamán 2018

En la Tabla 7-4, se observa que todas las variables siguen una distribución asimétrica positiva, es decir, existe mayor ubicación de los datos al lado izquierdo de su media. La variable con menor coeficiente de asimetría (3.05) es la edad de la plantación y la variable con mayor coeficiente 20.23 es la superficie cosechada. Además, todas las variables siguen una distribución leptocúrtica o más apuntada que la normal, es decir, con mayor grado de concentración de los datos alrededor de la media. La variable con menor coeficiente de curtosis (14.93) es la edad de la plantación y la

variable con mayor coeficiente de asimetría (585.1) es la superficie cosechada, esta última variable contiene datos de dos o más poblaciones diferentes.

Observando la varianza y la desviación típica de la misma Tabla 7-4, también es importante mencionar que los datos de todas las variables están muy separados de la media, por lo tanto, existe mayor dispersión entre los valores de la distribución y la media aritmética. La variable con menor desviación típica (8.73 hectáreas) es la superficie cosechada y la variable con mayor desviación típica (71.18 hectáreas) es la producción.

Tabla 8-4: Resumen de la D.e.f. para las variables cualitativas

Variable cualitativa	Categoría	n_i	f_i (%)
Condición de cultivo (cp_k404)	ASOCIADO	988	30.21
	SOLO	2282	69.79
Semilla de más uso (cp_k408)	COMUN	3073	93.98
	HIBRIDA INTERNACIONAL	3	0.09
	HIBRIDA NACIONAL	21	0.64
	MEJORADA	173	5.29
Uso de riego (cp_k413)	NO	2845	87.00
	SI	425	13.00
Uso de fitosanitarios (cp_k415)	NO	1996	61.04
	SI	1274	38.96
Uso de fertilizante orgánico (cp_forg)	NO	3146	96.21
	SI	124	3.79
Uso de fertilizante químico (cp_fqui)	NO	2412	73.76
	SI	858	26.24
Uso de plaguicida orgánico (cp_porg)	NO	3251	99.42
	SI	19	0.58
Uso de plaguicida químico (cp_pqui)	NO	2004	61.28
	SI	1266	38.72
Participación provincial (part_prov)	[20%;40%) (Manabí, 2016)	308	9.42
	[40%;60%) (Manabí, 2014 y 2015)	739	22.60
	Menos del 20%	2223	67.98

Realizado por: Eduardo Guamán 2018

El 69.79% de los terrenos con sembríos del plátano fueron SOLOS, el 93.98% fueron con semilla COMUN, el 87% NO fueron con el uso de riego, el 61.04% NO fueron con el uso de fitosanitarios, el 96.21% NO fueron con el uso de fertilizante orgánico, el 73.76% NO fueron con el uso de fertilizante químico, el 99.42% NO fueron con el uso de plaguicida orgánico, el 61.28% NO fueron con el uso de plaguicida químico y el 67.98% de los sembríos del plátano fueron de participación provincial en la producción nacional menos del 20%, es decir todas las provincias excepto Manabí (porque esta provincia representa del [20%;40%) para el año 2016 y del [40%;60%) en los años 2014 y 2015) participaron con un 67.98% en la producción nacional.

De manera general, la mayor producción del plátano en el Ecuador se da entre 0.23 y 183 toneladas métricas, con una edad de la plantación entre 1 y 8 años, una superficie plantada entre 0.01 y 21 hectáreas, una superficie en edad productiva entre 0.01 y 21 hectáreas, una superficie cosechada entre 0.01 y 21 hectáreas, presenta mayores ventas entre 0.02 y 181 toneladas métricas. Además, la producción del plátano se da con condición de cultivo sólo, con semilla común, sin riego, sin fitosanitarios, sin fertilizante químico y orgánico, sin plaguicida químico y orgánico, con una participación provincial del menos 20%, esto representaron a todas las provincias excepto Manabí.

4.2. Análisis bivariado de datos

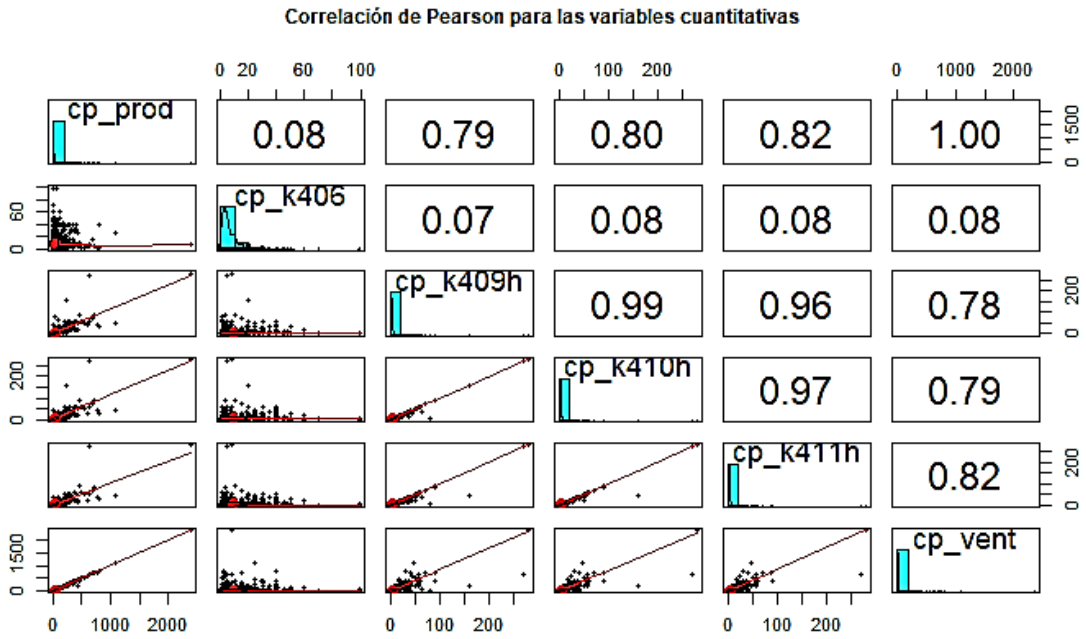


Gráfico 7-4: Correlación de Pearson para las variables cuantitativas
Realizado por: Eduardo Guamán 2018

El Gráfico 7-4 indica que la producción del plátano tiene una relación lineal positiva perfecta (correlación positiva perfecta) con las ventas, además una relación lineal positiva muy alta con las variables: superficie plantada, superficie en edad productiva y superficie cosechada. De otra manera se puede decir, a medida que la producción aumenta las ventas, la superficie plantada, la superficie en edad productiva y la superficie cosechada también aumentan. Por otro lado, la edad de la plantación tiene una correlación positiva muy baja con el resto de variables. La superficie plantada presenta una relación lineal positiva muy alta con las variables superficie en edad productiva y superficie cosechada, además una correlación positiva alta con las ventas. De la misma manera la superficie en edad productiva presenta una relación positiva muy fuerte con la superficie cosechada y una relación positiva fuerte con las ventas. Por último, la superficie cosechada tiene una relación lineal muy fuerte con las ventas.

A continuación, en la Tabla 9-4 se realiza un análisis de correlaciones con la prueba de Pearson y sus respectivos valores p.

Tabla 9-4: Test sobre la correlación de Pearson para variables cuantitativas

Variable	cp_prod	cp_k406	cp_k409h	cp_k410h	cp_k411h	cp_vent
cp_prod	1 0	-	-	-	-	-
cp_k406	0.08 0.00	1 0	-	-	-	-
cp_k409h	0.79 0.00	0.07 0.00	1 0	-	-	-
cp_k410h	0.80 0.00	0.08 0.00	0.99 0.00	1 0	-	-
cp_k411h	0.82 0.00	0.08 0.00	0.96 0.00	0.97 0.00	1 0	-
cp_vent	1 0	0.08 0.00	0.78 0.00	0.79 0.00	0.82 0.00	1 0

Realizado por: Eduardo Guamán 2018

Según la prueba de correlación de Pearson afirma que estadísticamente todas las variables están correlacionadas porque todos los p-valores son aproximadamente nulos y por ende son menores que cualquier nivel de significancia.

En la Tabla 10-4 se realiza un análisis de independencia, específicamente la prueba de Chi-Cuadrado χ^2 o el test de exacto de Fisher cuando los tamaños de las muestras en el análisis de tablas de contingencia son pequeños (y frecuencias esperadas menores a 5), estas pruebas consisten en comprobar si dos características cualitativas están relacionadas entre sí. En esta Tabla 10-4 se presenta los p-valores resultantes de dicho análisis.

Tabla 10-4: Valores-p resultantes del análisis de las tablas de contingencia

Variable	cp_k404	cp_k408	cp_k413	cp_k415	cp_forg	cp_fqui	cp_porg	cp_pqui	part_prov
cp_k404	0	-	-	-	-	-	-	-	-
cp_k408	0.170	0	-	-	-	-	-	-	-
cp_k413	0.000	0.000	0	-	-	-	-	-	-
cp_k415	0.663	0.004	0.000	0	-	-	-	-	-
cp_forg	0.685	0.000	0.044	0.000	0	-	-	-	-
cp_fqui	0.474	0.000	0.000	0.000	0.215	0	-	-	-
cp_porg	0.167	0.004	0.000	0.000	0.000	0.001	0	-	-
cp_pqui	0.640	0.004	0.000	0.000	0.000	0.000	0.138	0	-
part_prov	0.004	0.000	0.000	0.000	0.080	0.002	0.438	0.000	0

Realizado por: Eduardo Guamán 2018

La variable condición de cultivo no está relacionada con las variables: semilla de más uso, uso de fitosanitarios, uso fertilizante orgánico y químico, uso de plaguicida orgánico y químico. Además, el uso de fertilizante orgánico no está relacionada con el uso de fertilizante químico. Y finalmente

el uso de plaguicida orgánico no está relacionado con las variables uso de plaguicida químico y la participación provincial.

4.3. Análisis multivariado de datos

Es importante mencionar que para realizar el análisis multivariado se recodificó el nombre de las variables. (Ver Tabla 11-4).

Tabla 11-4: Variables recodificadas para el análisis multivariante

Variables recodificadas
cp_prod = Y
cp_k406 = X_2
cp_k409h = X_3
cp_k410h = X_4
cp_k411h = X_5
cp_vent = X_{10}
cp_k404 = X_{11}
cp_k408 = X_{12}
cp_k413 = X_{13}
cp_k415 = X_{15}
cp_forg = X_{16}
cp_fqui = X_{17}
cp_porg = X_{18}
cp_pqui = X_{19}
part_prov = X_{20}

Realizado por: Eduardo Guamán 2018

Esta recodificación se realizó para una mejor visualización de las variables en los gráficos.

4.3.1. Primer análisis factorial de datos mixtos

La primera dimensión del AFDM explica el 22.8% de la variabilidad total, la segunda dimensión explica el 13.3% y las dos primeras dimensiones del AFDM explican tan solo el 36.12% de la variabilidad total. (Ver Gráfico 8-4).

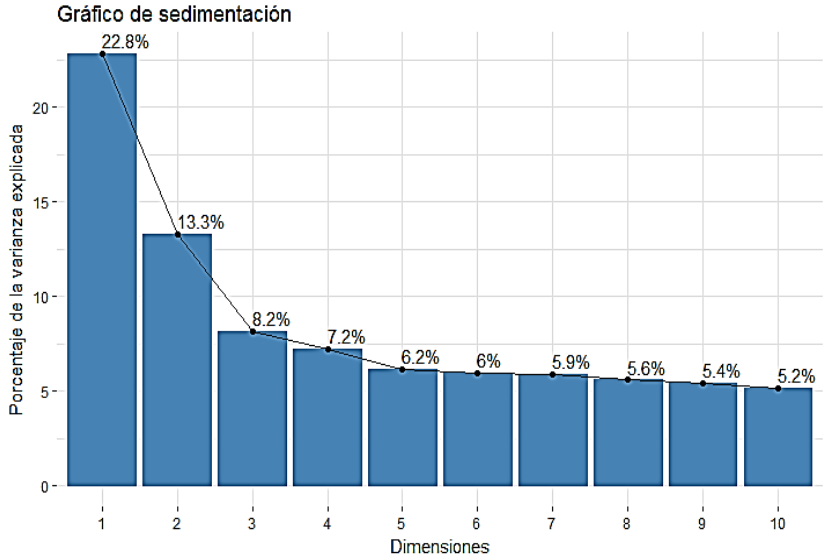


Gráfico 8-4: Gráfico de sedimentación en el primer AFDM
Realizado por: Eduardo Guamán 2018

En el Gráfico 9-4 se observa que las variables superficie cosechada, superficie en edad productiva, superficie plantada y ventas presenta una mayor correlación con la primera dimensión, mientras que las variables uso de fitosanitarios y uso de plaguicida químico presenta una mayor correlación con la segunda dimensión, por otro lado, la variable uso de fertilizante químico tiene una baja correlación con la segunda dimensión del primer AFDM.

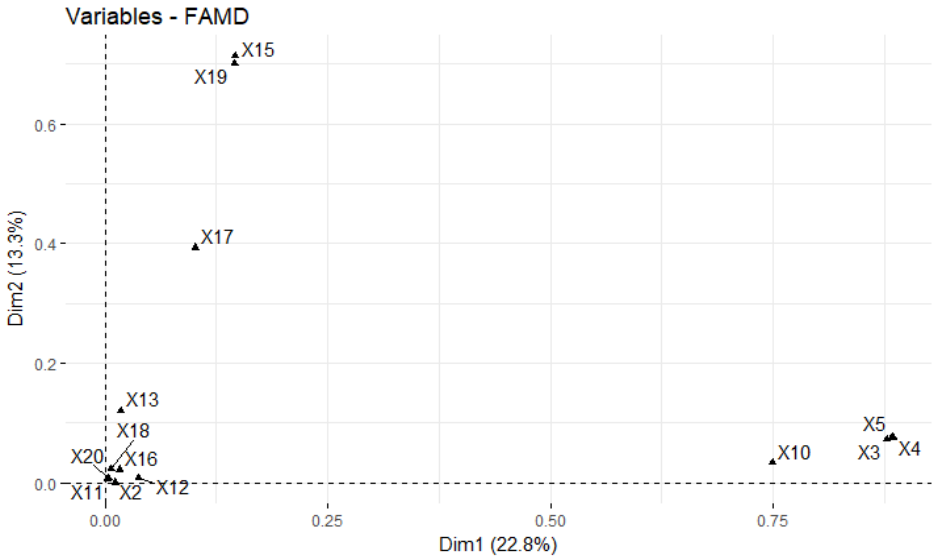


Gráfico 9-4: Representación de variables en el primer AFDM
Realizado por: Eduardo Guamán 2018

A continuación, se estudia la calidad de representación y las contribuciones de las variables a las dos primeras dimensiones del ADFM.

La calidad de representación se estudia con el cos2 (coseno cuadrado, coordenadas cuadradas), para una variable dada la suma del cos2 en todas las dimensiones es igual a uno, si una variable está bien representada solo por dos dimensiones la suma del cos2 de estas dos dimensiones serán cercano a uno. Las contribuciones en una dimensión dada se expresan en porcentajes y para una variable determinada se calcula (por ejemplo para la Dim1 y Dim2) como contrib = [(C1 * Eig1) + (C2 * Eig2)]/(Eig1 + Eig2) sabiendo que C1 y C2 son las contribuciones de la variable en Dim1 y Dim2, respectivamente, Eig1 y Eig2 son los valores propios de la Dim1 y Dim2, respectivamente, cuanto mayor sea el valor de la contribución, más contribuirá la variable a las dimensiones, finalmente es importante mencionar que la contribución promedio esperada (límite) de una variable para la Dim1 y Dim2 es: [(1/longitud(variables) * Eig1) + (1/longitud(variables) * Eig2)] / (Eig1 + Eig2).

Las variables superficie cosechada, superficie en edad productiva y superficie plantada están bien representadas en el mapa factorial debido a que la suma del cos^2 de las dos primeras dimensiones se aproxima a 1 (muy cercana a 0.8), por otro lado, se puede decir que también las variables ventas, uso de fitosanitarios y uso de plaguicida químico están representadas de una forma moderada ya que la suma del cos2 en las dos primeras dimensiones sobrepasan el 0.5. (Ver Gráfico 10-4 y 11-4).

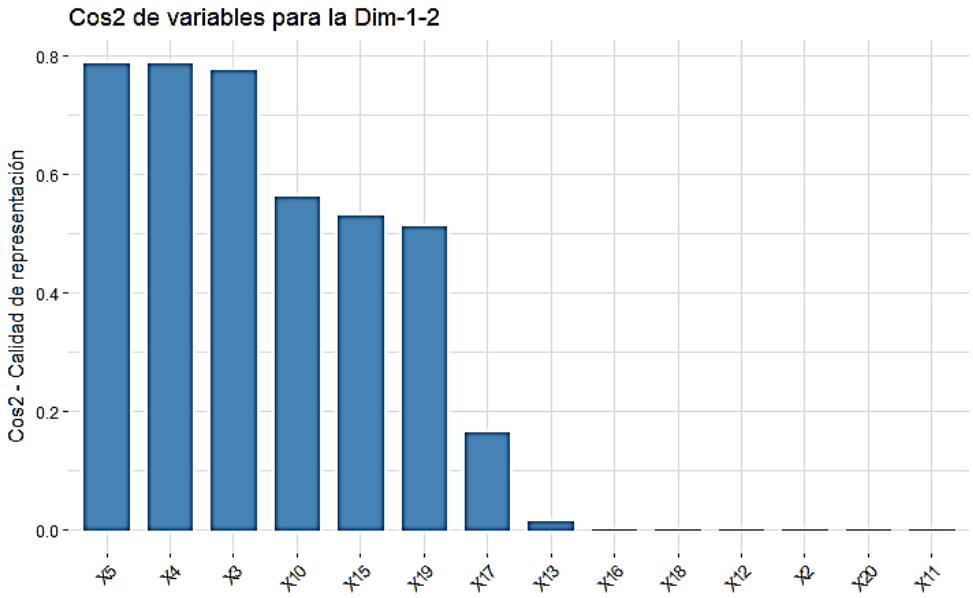


Gráfico 10-4: Cos2 de variables para las dos primeras dimensiones en el primer AFDM
Realizado por: Eduardo Guamán 2018

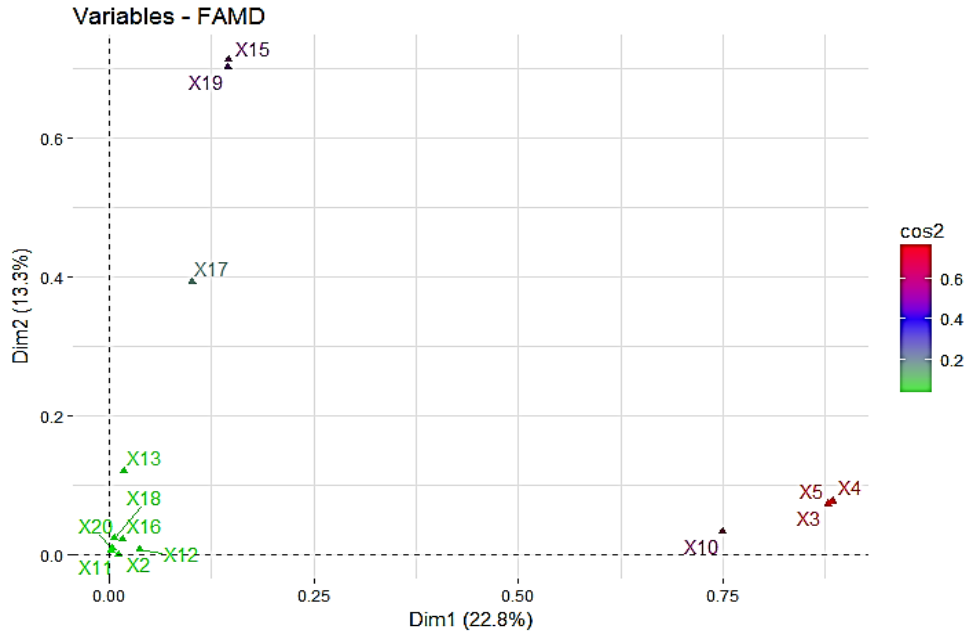


Gráfico 11-4: Calidad de representación (cos2) en el plano factorial en el primer AFDM
Realizado por: Eduardo Guamán 2018

Las variables más importantes (o contribuyentes) a las dos primeras dimensiones del ADFM son: superficie en edad productiva, superficie cosechada, superficie plantada, uso de fitosanitarios, uso de plaguicida químico, vetas y uso de fertilizante químico (Ver Gráfico 12-4 y 13-4).

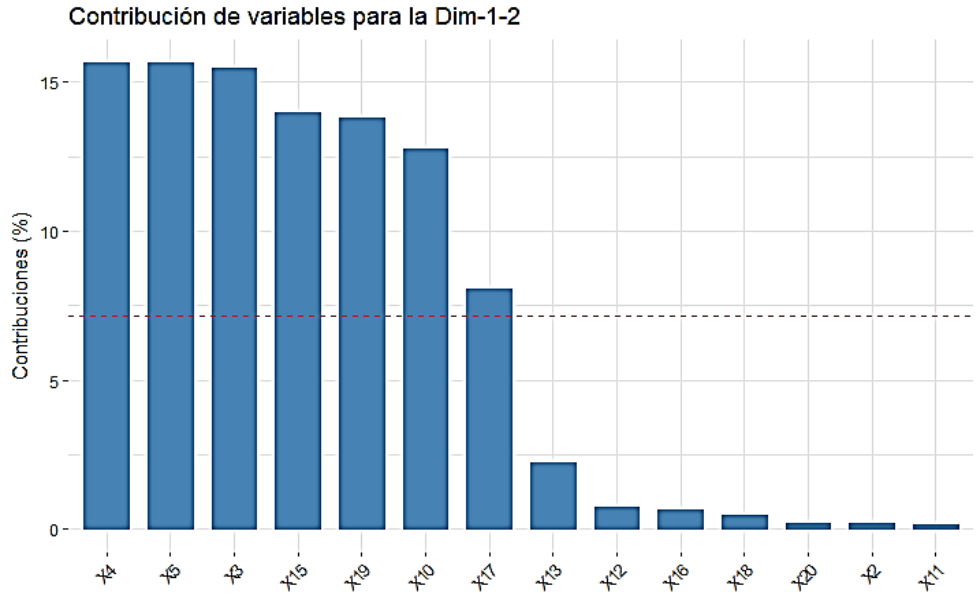


Gráfico 12-4: Contribución de variables para las dos primeras dimensiones en el primer ADFM
Realizado por: Eduardo Guamán 2018

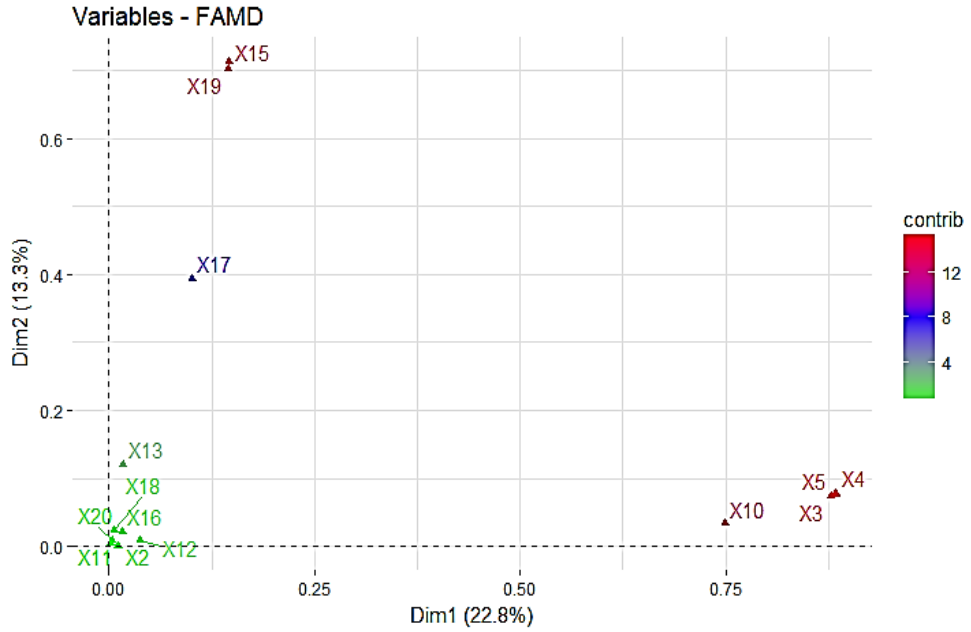


Gráfico 13-4: Contribución de las variables en el plano factorial del primer AFDM
Realizado por: Eduardo Guamán 2018

Luego de analizar la calidad de representación y las contribuyentes de las variables a las dos primeras dimensiones del primer AFDM se encontraron con algunas variables que no están bien representadas y a su vez no contribuyen a las dos primeras dimensiones, por lo tanto, estas variables se pueden separar para realizar un nuevo AFMD, lo cual nos permitirá mejorar la variabilidad total explicada por las dos primeras dimensiones y una mejor interpretación de los resultados en el segundo AFDM.

4.3.2. Segundo análisis factorial de datos mixtos

En este análisis se consideraron solamente las variables superficie en edad productiva, superficie cosechada, superficie plantada, uso de fitosanitarios, uso de plaguicida químico, vetas y uso de fertilizante químico.

En el segundo AFDM se presencia que la variabilidad total explicada mejora considerablemente porque la primera dimensión del AFDM explica el 54.5% de la variabilidad total, la segunda dimensión explica el 30.8% y las dos primeras dimensiones del segundo AFDM explican el 85.32% de la variabilidad total. (Ver Gráfico 14-4).

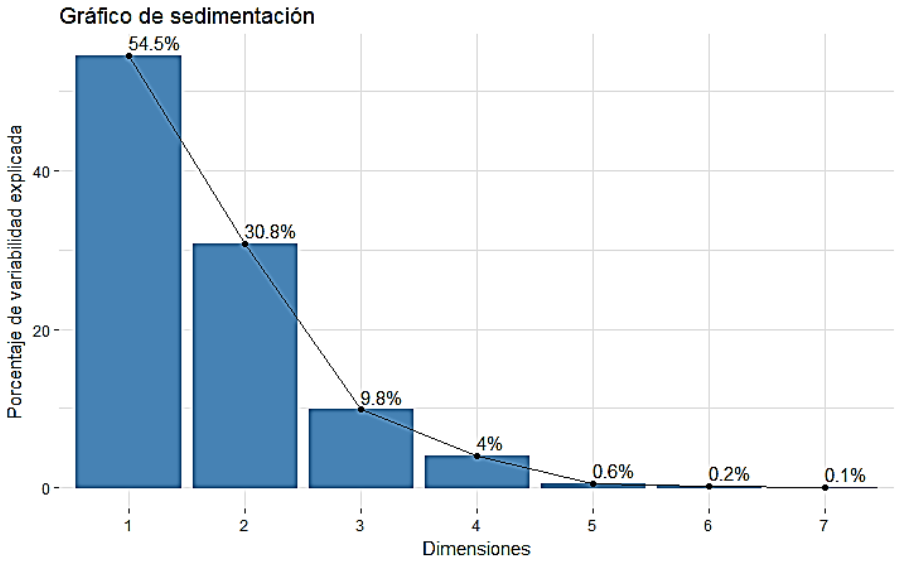


Gráfico 14-4: Gráfico de sedimentación en el segundo AFDM
Realizado por: Eduardo Guamán 2018

Con una buena variabilidad explicada por las dos primeras dimensiones, en el segundo análisis también se observa que las variables superficie cosechada, superficie en edad productiva, superficie plantada y ventas presenta una mayor correlación con la primera dimensión mientras que las variables uso de fitosanitarios y uso de plaguicida químico presenta una mayor correlación con la segunda dimensión, por otro lado, la variable uso de fertilizante químico tiene una baja correlación con la segunda dimensión del primer AFDM. (Ver Tabla 12-4 y Gráfico 15-4).

Tabla 12-4: Coordenadas del segundo AFDM

Variable	Dim.1	Dim.2
X_3	0.90	0.06
X_4	0.91	0.06
X_5	0.91	0.06
X_{10}	0.75	0.03
X_{15}	0.13	0.80
X_{17}	0.09	0.35
X_{19}	0.13	0.80

Realizado por: Eduardo Guamán 2018

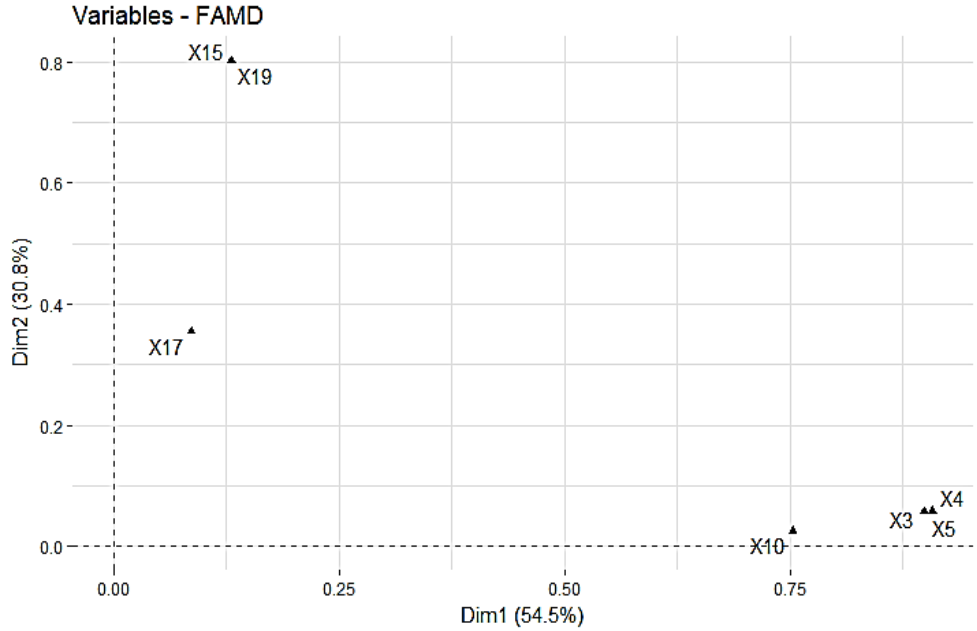


Gráfico 15-4: Representación de variables del segundo AFDM
Realizado por: Eduardo Guamán 2018

Del resultado del AFDM se puede realizar un análisis solo para las variables cuantitativas y por otro lado solo para las variables cualitativas.

Análisis de las variables cuantitativas

Las variables superficie plantada, superficie en edad productiva, superficie cosechada y ventas están altamente correlacionadas entre sí y con la primera dimensión, ya que todos los vectores están juntos y tienden a una sola dirección. Esto indica que a mayor producción del plátano mayor es la superficie cosechada, mayor es la superficie en edad productiva, mayor es la superficie plantada y obviamente las ventas son mayores. (Ver Tabla 13-4 y Gráfico 16-4).

Tabla 13-4: Coordenadas de las variables cuantitativas en el AFDM

Variables	Dim.1	Dim.2
X_3	0.95	0.24
X_4	0.95	0.24
X_5	0.95	0.24
X_{10}	0.87	0.16

Realizado por: Eduardo Guamán 2018

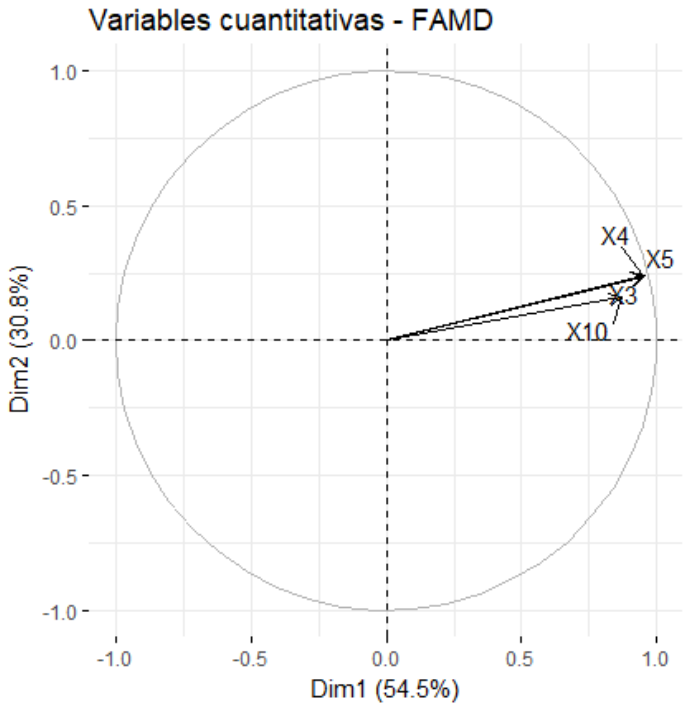


Gráfico 16-4: Representación de las variables cuantitativas en el plano factorial

Realizado por: Eduardo Guamán 2018

Análisis de las variables cualitativas

Es evidente al observar la Tabla 14-4 y el Gráfico 17-4 que la producción del plátano, el no usar fitosanitarios corresponde o está relacionada principalmente al no usar fertilizante químico y al no usar plaguicida químico, por otro lado, el usar fitosanitarios corresponde principalmente el uso de fertilizante químico y el uso de plaguicida químico. También podemos observar que en la producción del plátano se da mayor uso de fitosanitarios, fertilizante químico y plaguicida químico.

Tabla 14-4: Coordenadas de las variables cualitativas en el AFDM

Categorías	Dim.1	Dim.2
NO_X15	-0.56	1.05
SI_X15	0.88	-1.65
NO_X17	-0.34	0.52
SI_X17	0.96	-1.47
NO_X19	-0.56	1.05
SI_X19	0.89	-1.65

Realizado por: Eduardo Guamán 2018

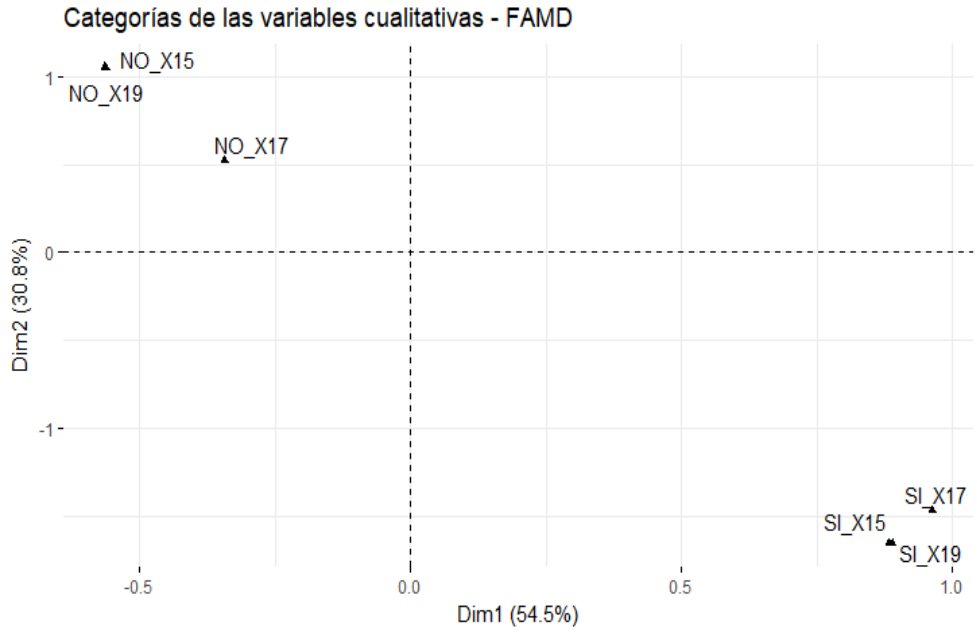


Gráfico 17-4: Representación de las variables cualitativas en el plano factorial
Realizado por: Eduardo Guamán 2018

De maneara general, el primera AFDM clasificó las variables agronómicas sobre la producción del plátano en 3 grupos, el segundo AFDM se realizó excluyendo las variables no contribuyentes y aquellas que no estuvieron bien representadas en las dos primeras dimensiones del primer AFDM. La variabilidad explicada por las dos primeras dimensiones en el segundo AFDM mejoró considerablemente y clasificó a las variables agronómicas sobre la producción del plátano en 2 factores. Por lo tanto, el primer factor que influye en la producción del plátano está formado por las variables: superficie cosechada, superficie en edad productiva, superficie plantada y ventas (al primer factor se le dio el nombre de "superficie"); el segundo factor que influye en la producción del plátano está formada por las variables: uso de fitosanitarios, uso de plaguicida químico y uso de fertilizante químico (al segundo factor se le dio el nombre de "uso y cuidado"), la variable uso de fertilizante químico del factor uso y cuidado tiene una baja influencia en la producción del plátano por su calidad de representación y contribución en las dos primeras dimensiones del AFDM.

En comparación con el análisis bivariado, las variables que forman el primer factor que influye en la producción del plátano están altamente correlacionadas entre sí y la prueba de correlación indica que estas variables son altamente significativas, las variables formadas por el segundo factor que influye en la producción del plátano de la misma manera están relacionadas entre sí según la prueba de Chi-Cuadrado.

4.3.3. Análisis de regresión lineal múltiple con variables dummy

El ARLMVD se realizó solo a las variables que conforman los dos factores que influyen en la producción del plátano. Esto para obtener más información acerca de las variables.

1. Análisis de la correlación entre cada par de variables cuantitativas y diferencias del valor promedio entre las categóricas

Análisis de la correlación entre cada par de variables cuantitativas

Las variables que tienen una mayor relación lineal con la producción del plátano son todas aquellas conformadas por el primer factor. Se observa que existe una alta correlación entre las variables predictoras por lo que posiblemente no sea útil introducir algunos predictores en el modelo, por otro lado, tanto la variable respuesta y predictoras cuantitativas muestran una distribución exponencial, por lo tanto, una transformación logarítmica posiblemente haría más normal su distribución. (Ver Gráfico 18-4).

Es importante recordar que el ARLMVD se realizó para obtener más información de las variables que conforman los factores que influyen en la producción del plátano, debido a esto no se realiza dicho análisis mencionado anteriormente.

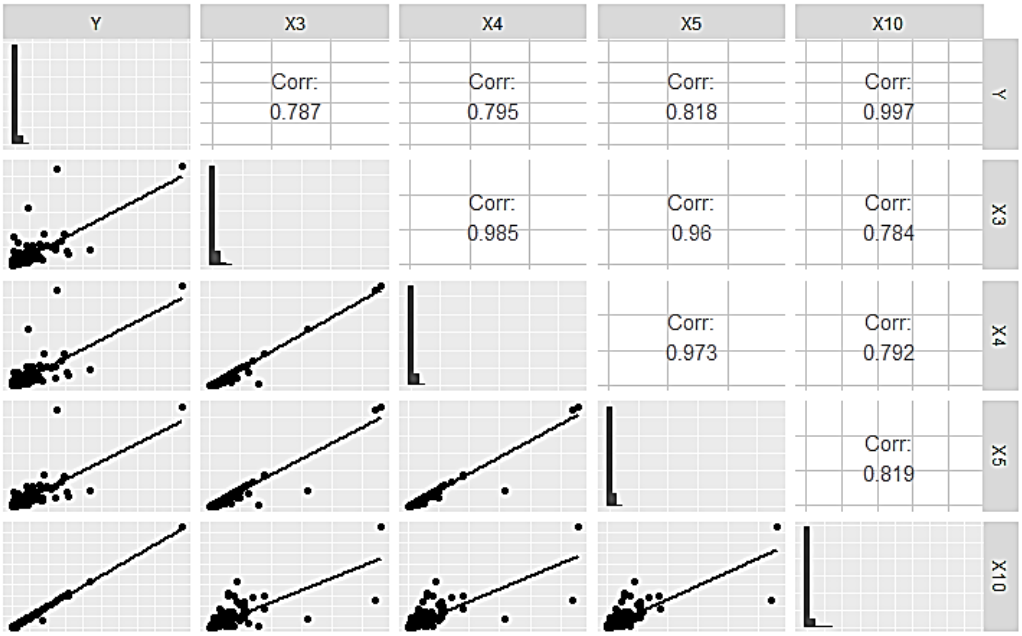


Gráfico 18-4: Análisis preliminar de las variables cuantitativas
Realizado por: Eduardo Guamán 2018

Análisis de las diferencias del valor promedio entre las categóricas

Las variables cualitativas del factor uso y cuidado (X_{15}, X_{17} y X_{19}) parecen influir de forma significativa en la variable respuesta producción del plátano. (Ver Gráfico 19-4).

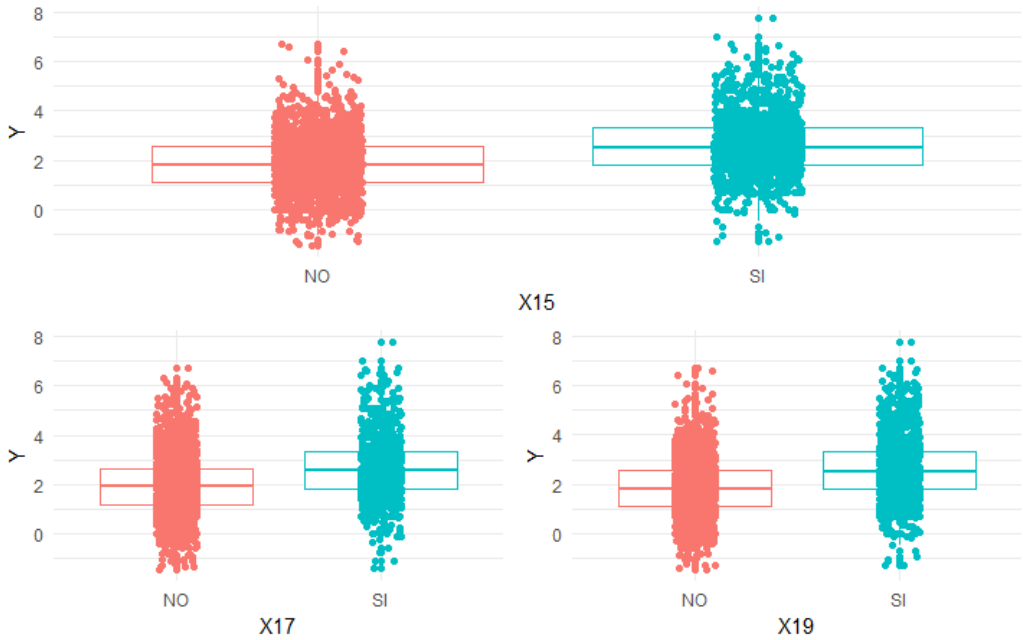


Gráfico 19-4: Análisis preliminar de las variables cualitativas
Realizado por: Eduardo Guamán 2018

Con el análisis preliminar se puede decir que las variables que conforman los dos factores que influyen en la producción del plátano pueden ser buenos predictores en un modelo ARLMVD para la variable dependiente.

2. Generación del modelo ARLMVD y selección de los mejores predictores

En el modelo general (1) se observar que la variable que conforma el segundo factor (uso de fertilizante químico) no influye en la producción del plátano para ningún nivel de significancias habituales que se conoce (0.01, 0.05 y 0.10), por lo tanto, se realiza el modelo restringido (2) excluyendo dicha variable, este modelo también es la resultante del mejor modelo del proceso de selección con el método paso a paso y el criterio de información de Akaike (AIC). En el modelo restringido se observa la buena significatividad individual y conjunta de los parámetros estimados (Valores p de la T de Student y la F de Fisher menores que 0.05 y 0.01), el R^2 ajustado tiene un valor muy aceptable (99.4% de explicatividad). (Ver Tabla 15-4).

Tabla 15-4: Modelo general y restringido para la producción del plátano

		Variable dependiente:	
		Y	
		Modelo general (1)	Modelo restringido (2)
X_3	Coeficiente	0.143	0.142
	Estadístico t-Student	2.436	2.426
	Valor-p	0.01**	0.02**
X_4		0.543	0.544
		7.413	7.421
		0.00***	0.00***

	-0.686	-0.686
X_5	13.693	13.691
	0.00***	0.00***
	1.016	1.016
X_{10}	427.476	428.470
	0.00***	0.00***
	3.305	3.273
X_{15}SI	1.727	1.713
	0.08*	0.09*
	-0.081	
X_{17}SI	-0.336	
	0.74	
	-3.453	-3.453
X_{19}SI	-1.804	-1.805
	0.07*	0.07*
	0.935	0.927
Constante	7.456	7.527
	0.00***	0.00***
Observaciones	3270	3270
R^2	0.994	0.994
R^2 Ajustado	0.994	0.994
Error Std. Residual	5.395 (df = 3262)	5.394 (df = 3263)
Estadístico F-Fisher	80819.960 (df = 7; 3262)	94315.580 (df = 6; 3263)
Valor-p	0.00***	0.00***
Nota:	*p<0.1; **p<0.05; ***p<0.01; df = grados de libertad	

Realizado por: Eduardo Guamán 2018

El modelo ARLMVD $Producción = 0.927 + 0.142 SuperficiePlantada + 0.544 SuperficieEdadProductiva - 0.686 SuperficieCosechada + 1.016 Ventas + 3.273 X_{15} Si - 3.453 X_{19} Si$, es capaz de explicar el 99.4% de la variabilidad observada en la producción del plátano. El test F es altamente significativo.

$\beta_0 = 0.927$: Indica que cuando no se usa fitosanitarios, no se usa plaguicida químico y cuando las variables superficie plantada, superficie en edad productiva, superficie cosechada y ventas se mantienen constantes la producción media del plátano es 0.927 toneladas métricas.

$\beta_3{}^* = 0.142$: Por cada incremento de una hectárea de la superficie plantada, aumentará en 0.142 toneladas métricas la producción del plátano en el Ecuador.

$\beta_4 = 0.544$: Por cada incremento de una hectárea de la superficie en edad productiva, aumentará en 0.544 toneladas métricas la producción del plátano en el Ecuador.

$\beta_5 = -0.686$: Por cada incremento de una hectárea de la superficie cosechada, disminuirá en 0.686 toneladas métricas la producción del plátano en el Ecuador.

$\beta_{10} = 1.016$: Por cada incremento de una tonelada métrica de las ventas de plátano, aumentará en 1.016 toneladas métricas la producción del plátano en el Ecuador.

Para las variables cualitativas se puede interpretar de la siguiente manera y estas son:

* Los coeficientes se codificaron de esa forma para que sea correspondiente a las variables.

$\beta_{15}Si = 3.273$: En promedio los terrenos que, si usan fitosanitarios, tienen una producción del plátano de 4.2 (0.927+3.273) toneladas métricas, cuando no se usan fitosanitarios en promedio la producción del plátano es de 0.927 toneladas métricas; o que en promedio la producción del plátano con el uso de fitosanitarios es 3.273 toneladas métricas más que al no usar fitosanitarios.
$\beta_{19}Si = -3.453$: En promedio la producción del plátano con el uso de plaguicida químico es 3.453 toneladas métricas menos que al no usar plaguicida químico.

3. Validación de condiciones para la regresión lineal múltiple con variables dummy

La validación de condiciones para el ARLMVD se realizó a la regresión restringida.

No multicolinealidad (No colinealidad)

Se detecta mediante los factores de inflación de la varianza (FIV) que deben ser menores que 10. Era claro que existiría multicolinealidad porque en el análisis preliminar de las variables cuantitativas, en el análisis de correlación de Pearson y en el análisis de las tablas de contingencia se observó que había una alta correlación entre las variables predictoras cuantitativas y una relación entre las variables cualitativas, finalmente con el análisis de los FIV podemos afirmar que hay problemas de multicolinealidad (Ver Tabla 16-4).

Tabla 16-4: Factores de inflación de la varianza del modelo ARLMVD

Variables	FIV
X_3	34.36
X_4	51.35
X_5	21.50
X_{10}	3.07
X_{15}	97.60
X_{19}	97.62

Realizado por: Eduardo Guamán 2018

Linealidad

Esta condición se puede validar bien mediante diagramas de dispersión entre la variable dependiente y cada uno de los predictores o con diagramas de dispersión entre cada uno de los predictores y los residuos del modelo. Si la relación es lineal, los residuos deben distribuirse aleatoriamente en torno a 0 con una variabilidad constante a lo largo del eje X. Esta última opción suele ser más indicada por los analistas ya que permite identificar posibles datos atípicos. Otra forma de contrastar la linealidad es realizando el gráfico de componentes más residuos (residuos parciales).

En el Gráfico 20-4 se observa que todos los gráficos se ajustan a una recta, por otra parte, el Gráfico 21-4 también se ajusta a una recta. Y finalmente el contraste RESET de Ramsey afirma que no existe problemas de linealidad ya que el estadístico es 229.27 y el valor-p es aproximadamente cero (2.20e-16).

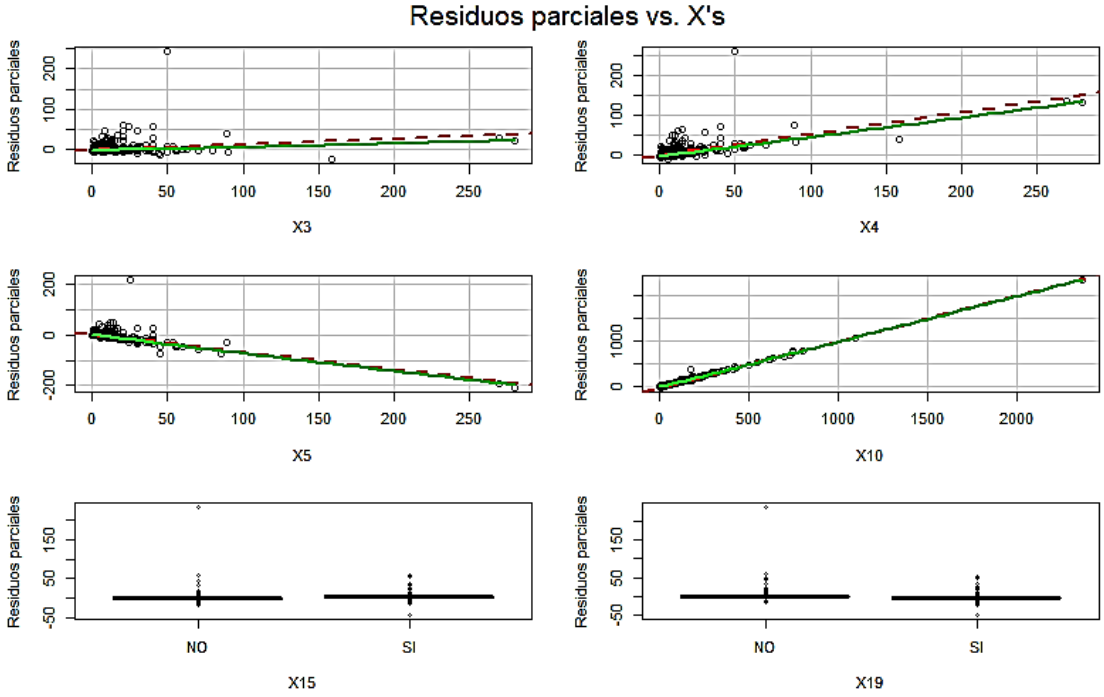


Gráfico 20-4: Residuos parciales contra variables independientes para contrastar la linealidad del modelo
Realizado por: Eduardo Guamán

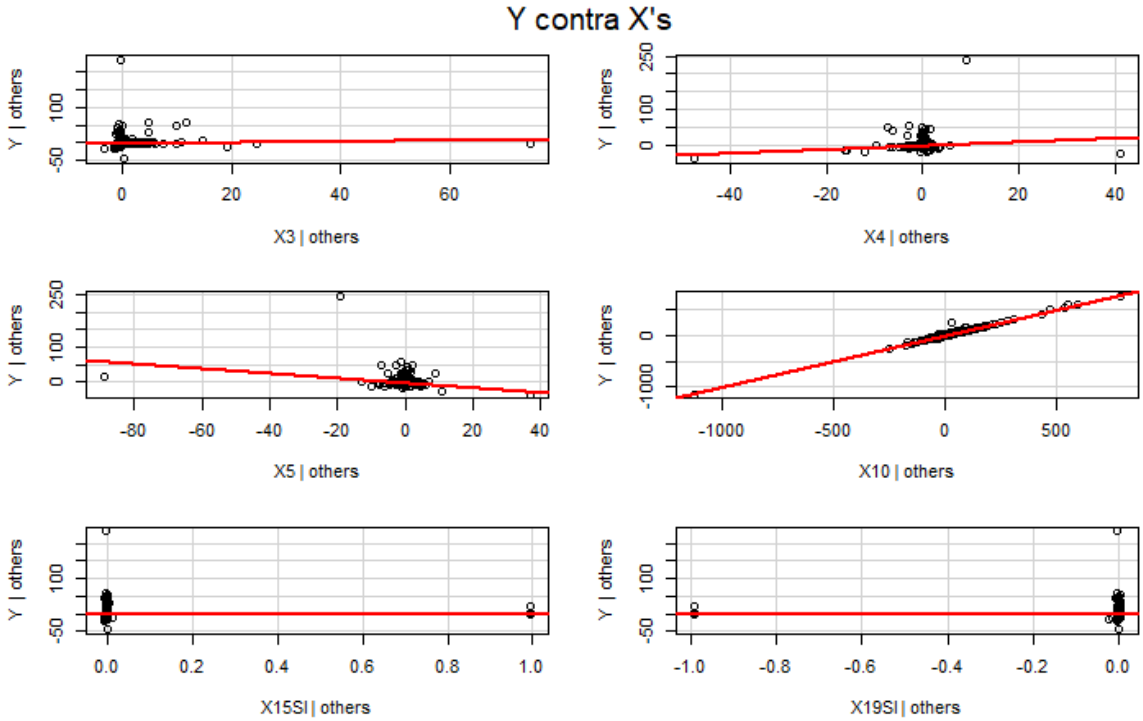


Gráfico 21-4: Variable dependiente contra las variables independientes
Realizado por: Eduardo Guamán 2018

Normalidad

En el Gráfico 22-4 se observa que los residuos no se ajustan a la línea recta, también existe dos valores atípicos residuales (se recomienda separar) y además la prueba de normalidad de Shapiro-Wilk presenta un valor-p aproximado a cero (2.2e-16), esto indica que hay problemas de normalidad.

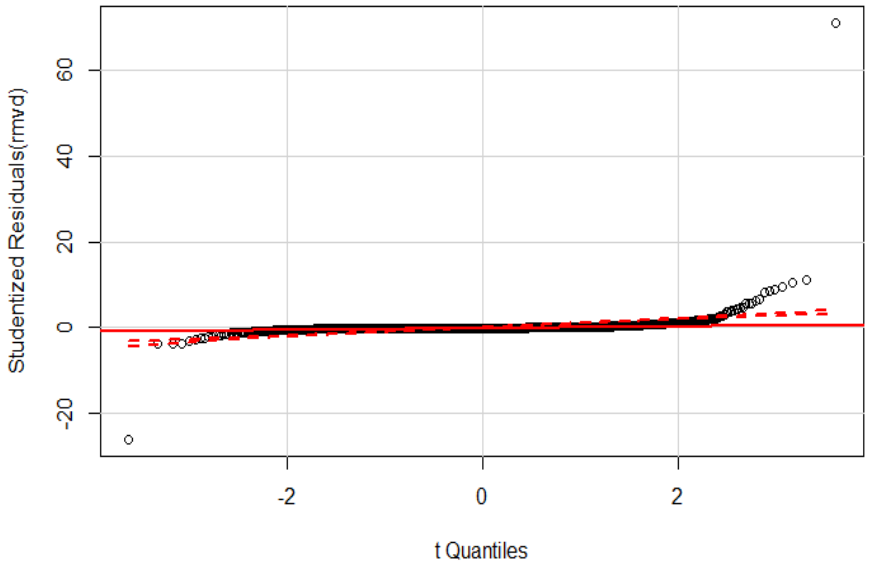


Gráfico 22-4: QQ de cuantiles para analizar la normalidad de los residuos en el ARLMVD

Realizado por: Eduardo Guamán 2018

Homocedasticidad

En el Gráfico 23-4 se observa que no hay problemas de homocedasticidad y el contraste de Goldfeld-Quandt afirma que los residuos no tienen problemas de homocedasticidad, ya que el valor-p es igual a 1.

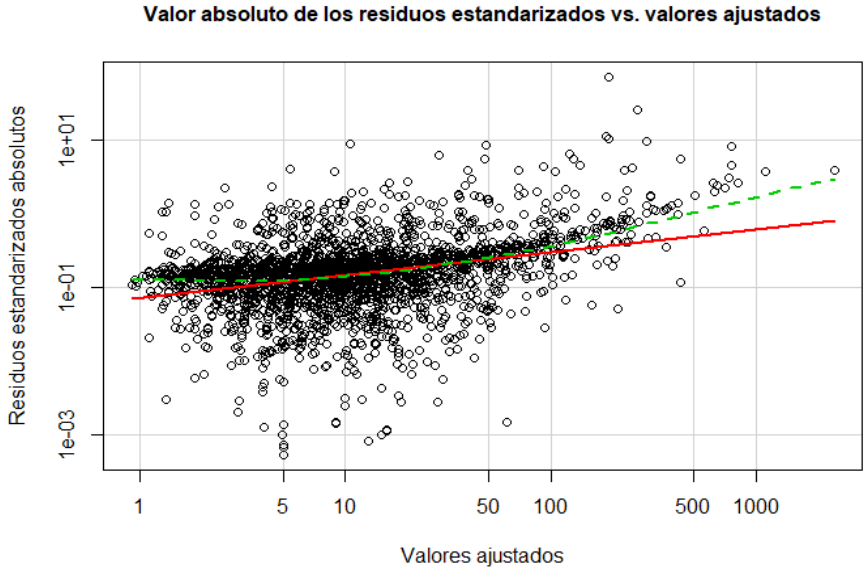


Gráfico 23-4: Valores absolutos de los residuos estandarizados contra los valores ajustados

Realizado por: Eduardo Guamán 2018

Autocorrelación

Según Ximénez y San Martín (2013) en la página 62 afirma que se puede asumir independencia entre residuos si $1.50 \leq DW \leq 2.50$, en nuestro caso el estadístico de Durbin-Watson tiene un valor de 1.85 y esto indica que el modelo restringido no tiene problemas de autocorrelación, es decir los residuos del modelo restringido son independientes.

Es cierto que es importante validar las condiciones para la regresión lineal múltiple, y uno de los supuestos más fuertes es la multicolinealidad, luego la linealidad, normalidad, homocedasticidad y autocorrelación. Sin embargo, en este estudio no se dio mayor importancia al supuesto de multicolinealidad ya que de antemano se sabe que habrá un problema de multicolinealidad debido al uso de las variables dentro de los dos factores encontrados en el AFMD. La regresión múltiple persigue dos objetivos importantes, (1) estudiar la estructura de la relación entre las variables independientes y dependiente, y (2) estimar la variable dependiente; con respecto al primer objetivo la falta de cumplimiento de los supuestos viene a ser un problema, y para el segundo objetivo no necesariamente es importante la validación del modelo. En esta investigación el ARLMVD más bien sirve para estimar la variable dependiente e interpretar los resultados por el hecho de que cumple con el supuesto de linealidad, homocedasticidad e independencia.

Se puede corregir cada uno de los problemas del no cumplimento de los supuestos en regresión lineal múltiple pero hay casos que no es posible solucionar el problema, por lo tanto como una alternativa y recomendación es realizar un análisis mediante modelos lineales generalizados, como por ejemplo la regresión Gamma (cuando la variable dependiente es cuantitativa, y las variables independientes solo cuantitativas o mixtas), también se podría categorizar la variable dependiente y estudiar como una regresión logística, la regresión robusta o entre otros análisis que no necesariamente requiere la validación de los supuestos.

CONCLUSIONES

- El análisis exploratorio de datos permitió determinar que la mayor producción del plátano en el país se da entre 0.23 y 183 toneladas métricas, con una edad de la plantación entre 1 y 8 años, una superficie plantada, superficie en edad productiva, superficie cosechada entre 0.01 y 21 hectáreas, la mayor venta del plátano se da entre 0.02 y 181 toneladas métricas, Además, los cultivos de plátanos en el país son con condición de cultivo sólo, con semilla común, sin riego, sin fitosanitarios, sin fertilizante químico, sin fertilizante orgánico, sin plaguicida químico, sin plaguicida orgánico y la producción mayoritaria (67.98%) se da en aquellas provincias del primer quintil de producción. Por otra parte, con dicho análisis se determinó que la variable producción conjuntamente con el resto de variables cuantitativas (edad de la plantación, superficie plantada, superficie en edad productiva, superficie cosechada y ventas) siguen una distribución asimétrica positiva y leptocúrtica, finalmente la variable con mayor coeficiente de asimetría es la superficie cosechada, lo cual indica que contiene datos de dos o más poblaciones diferentes.
- En cuanto al análisis bivariado de datos se obtuvo que la variable producción del plátano tiene una relación lineal positiva perfecta con la variable ventas, una relación lineal positiva muy alta con las variables: superficie plantada, superficie en edad productiva y superficie cosechada. Según la prueba de correlación de Pearson todas las variables cuantitativas están correlacionadas y conforme la prueba Chi-Cuadrado no todas las variables cualitativas son independientes.
- El análisis multivariante muestra que las dos primeras dimensiones del AFDM explican el 85.32% de la variabilidad total, las variables superficie cosechada, superficie en edad productiva, superficie plantada y ventas tiene una mayor correlación con la primera dimensión, las variables uso de fitosanitarios y uso de plaguicida químico presenta una mayor correlación con la segunda dimensión y la variable uso de fertilizante químico tiene una baja correlación con la segunda dimensión.
- Respecto a los factores que influyen en la producción del plátano se halló dos factores influyentes, el primer factor (factor superficie) que influye en la producción del plátano está formado por las variables: superficie cosechada, superficie en edad productiva, superficie plantada y ventas, el segundo factor (factor uso y cuidado) que influye en la producción del plátano está formada por las variables: uso de fitosanitarios, uso de plaguicida químico y uso de fertilizante químico, finalmente la variable uso de fertilizante químico del segundo factor tiene una baja influencia en la producción del plátano por su calidad de representación y contribución en las dos primeras dimensiones del AFDM.
- Con base a la regresión lineal múltiple con variables dummy, en la cual se trabajó con las variables que conforman los dos factores que influyen en la producción del plátano se encontró

que la variable que conforma el segundo factor (uso de fertilizante químico) no es significativo en la variable dependiente producción del plátano, lo cual permitió realizar un modelo restringido excluyendo dicha variable.

- El modelo restringido es altamente significativo presentando la siguiente ecuación $Producción = 0.927 + 0.142SuperficiePlantada + 0.544SuperficieEdadProductiva - 0.686SuperficieCosechada + 1.016Ventas + 3.273X_{15}Si - 3.453X_{19}Si$, esta ecuación es capaz de explicar el 99.43% de la variabilidad observada en la producción del plátano, finalmente en la validación del modelo se encontró que el modelo cumple con la linealidad, homogeneidad e independencia.

RECOMENDACIONES

- Socializar los resultados del estudio para que los productores e instituciones (MAGAP, SINAGAP, etc.) interesadas y dedicadas a la producción de dicho rubro puedan tomar decisiones y además tener una mejor perspectiva de la producción.
- Se recomienda realizar y aplicar las técnicas usadas en este estudio para otros productos del país y de esta manera tomar buenas decisiones de acuerdo a los resultados con el fin de mejorar la economía del país.
- La técnica del AFDM si bien es cierto es muy usadas en los últimos años debido a que trabajan simultáneamente con variables mixtas, sin embargo, en la provincia su estudio no es generalizado, por lo tanto, se recomienda incluir en la lista de electivas de la carrera de ingeniería en estadística informática.

BIBLIOGRAFÍA

Acuña, E., *Análisis de regresión* [en línea]. S.l.: s.n. 2007. Disponible en: https://es.scribd.com/doc/89707341/ML-Analisis-de-regresion-Edgar-Acuna.

Aldás, J. y Uriel, E., *Análisis multivariante aplicado con R*. Segunda. S.l.: Paraninfo. 2017. ISBN 8428329699.

Arcarons, J. y Calonge, S., *Microeconometría: introducción y aplicaciones con software econométrico para Excel*. S.l.: Delta. 2007. ISBN 9788496477940.

Ávila et al., *La alimentación española características nutricionales de los principales alimentos de nuestra dieta* [en línea]. S.l.: Fundación Española de la Nutrición. 2007. [Consulta: 7 junio 2017]. ISBN 978-84-491-0805-1. Disponible en: http://www.fen.org.es/mercadoFen/mercadofen_ajus_General.html.

Cuadras, C.M., *Nuevos Métodos De Análisis Multivariante* [en línea]. CMC Editio. S.l.: s.n. 2014. Disponible en: http://www.ub.edu/stat/personal/cuadras/nuevosmetodos.pdf.

Díaz, L. y Morales, M., *Estadistica multivariada: inferencia y métodos*. 3a ed. Bogotá: s.n. 2012. ISBN 978-958-701-195-1.

Fox, J., *Applied regression analysis and generalized linear models*. Third Edition. S.l.: SAGE. 2016. ISBN 9781452205663.

Franco, T.L. y Hidalgo, R., *Análisis estadístico de datos de caracterización morfológica de recursos fitogenéticos* [en línea]. S.l.: Instituto Internacional de Recursos Fitogenéticos. 2003. [Consulta: 17 abril 2018]. ISBN 9290435437. Disponible en: https://www.bioversityinternational.org/e-library/publications/detail/analisis-estadistico-de-datos-de-caracterizacion-morfologica-de-recursos-fitogeneticos/.

García et al., Caracterización de los Sistemas Lecheros en San Joaquín de Tuís, Turrialba, Costa Rica. *Revista de Investigación* [en línea], 2017. p. 45-51. [Consulta: 26 febrero 2018]. Disponible en: http://revistasnicaragua.net.ni/index.php/revinvucc/article/view/2991.

González, M., *Mercado del plátano en México 1971-2010, un modelo econométrico.* [en línea]. S.l.: Colegio de Postgraduados. 2012. [Consulta: 7 junio 2017]. Disponible en: http://colposdigital.colpos.mx:8080/jspui/handle/10521/1681?show=full.

Iica, Magfor y Jica, *Cadena agroindustrial plátano*. [en línea]. Nicaragua: 2004. [Consulta: 20 julio 2017]. Disponible en: http://repiica.iica.int/docs/B0030e/B0030e.pdf.

INEC y ANDA, *Ecuador - Encuesta de superficie y producción agropecuaria continua 2010*. [en línea]. 2010. [Consulta: 12 noviembre 2017]. Disponible en: http://anda.inec.gob.ec/anda/index.php/catalog/266/vargrp/VG15.

Jeproll, S., Exportación e Importación de plátanos, mango, piña, papaya y otras frutas exóticas de Ecuador en la Unión Europea, en los demás países de Europa y en Rusia. [en línea]. 2009. [Consulta: 24 octubre 2017]. Disponible en: http://www.jeproll.com/platanos-ecuatorianos.php.

Kassambara, A., *Practical guide to principal component methods in R*. Edition 1. S.l.: s.n. 2017. ISBN 1975721136.

Mamuye, N., Statistical Analysis of Factor Affecting Banana Production in Gamo Gofa District, Southern Ethiopia. [en línea], 2016. p. 5. [Consulta: 25 octubre 2017]. DOI 10.11648/J.EAS.20160101.12. Disponible en: http://article.sciencepublishinggroup.com/html/10.11648.j.eas.20160101.12.html.

OEC, *Ecuador (ecu) exportaciones, importaciones, y socios comerciales*. [en línea]. [sin fecha]. [Consulta: 3 julio 2017]. Disponible en: http://atlas.media.mit.edu/es/profile/country/ecu/.

Orellana, J., Unda, J. y Analuisa, P., Estudio de comercialización del plátano en la zona norte del trópico húmedo ecuatoriano. [en línea]. S.l.: Santo Domingo, EC, INIAP, Estación Experimental Santo Domingo, 2002. 2002. [Consulta: 7 junio 2017]. Disponible en: http://repositorio.iniap.gob.ec/handle/41000/3546.

Pacheco, C., Vergara, M. y Ligarreto, G., Clasificación de 85 accesiones de arveja (Pisum sativum L.), de acuerdo con su comportamiento agronómico y caracteres morfológicos. *Agronomía Colombiana* [en línea], 2009. p. 323-332. [Consulta: 26 febrero 2018]. Disponible en: https://revistas.unal.edu.co/index.php/agrocol/article/view/13275.

Pagès, J., *Analyse factorielle de données mixtes : principe et exemple d'application*. [en línea], 2014. [Consulta: 20 julio 2017]. Disponible en: http://www.agro-montpellier.fr/sfds/CD/textes/pages1.pdf.

Pagès, J., *Multiple factor analysis by example using R*. S.l.: CRC Press. 2015. ISBN 9781482205473.

Peña, D., *Análisis de datos multivariantes*. *Book* [en línea], 2002. p. 515. ISSN 1098-6596. DOI 8448136101. Disponible en: http://www.mhe.es/universidad/ciencias_matematicas/pena/index.html.

Pro Ecuador, Análisis sectorial Plátano 2015. *Instituto de promoción de exportaciones e inversiones* [en línea]. 2015. [Consulta: 25 junio 2017]. Disponible en: http://www.proecuador.gob.ec/wp-content/uploads/2015/06/PROEC_AS2015_PLATANO1.pdf.

Rodríguez, D. y González, G., *Principios de Econometría* [en línea]. S.l.: Instituto Tecnológico Metropolitano. 2017. [Consulta: 22 mayo 2018]. ISBN 9789585414181. Disponible en: http://fondoeditorial.itm.edu.co/libros-electronicos/principios-de-econometria/detalle-libro.html.

Sahay, A., *Applied regression and modeling: A Computer Integrated Approach*. S.l.: s.n. 2016. ISBN 9781631573309.

Sinagap, *Boletín situacional plátano*. [en línea]. 2014. [Consulta: 7 junio 2017]. Disponible en: http://sinagap.agricultura.gob.ec/phocadownloadpap/cultivo/2014/nboletin-situacional-de-platano-2014-actualizado.pdf.

Sinagap, *Boletín situacional plátano*. [en línea]. 2015. [Consulta: 7 junio 2017]. Disponible en: http://sinagap.agricultura.gob.ec/phocadownloadpap/cultivo/2016/boletin_situacional_platano_2015.pdf.

Sinagap, *Boletín situacional plátano*. [en línea]. 2016. [Consulta: 12 mayo 2018]. Disponible en: http://sipa.agricultura.gob.ec/biblioteca/boletines_situacionales/2016/boletin_situacional_platano_ 2016.pdf.

Sinmiedosec, *Lista de productos que exporta ecuador*. [en línea]. 2015. [Consulta: 3 julio 2017]. Disponible en: http://sinmiedosec.com/lista-de-productos-que-exporta-ecuador/.

Ximénez, M.C. y San Martín, R., *Fundamentos de las técnicas multivariantes*. S.l.: Uned. 2013. ISBN 9788436267983.

ANEXOS

Anexo A. Script de R

ANÁLISIS EXPLORATORIO DE DATOS

Cargar las librerías

```
Library("e1071")
library("prettyR")
library("modeest")
library("modes")
library("sjPlot")
```

Lectura de la base de datos

```
dir()
bd<-read.csv(file="BD141516.csv",header=T,sep=";",dec=".")
# str(bd1)
dim(bd)
```

VARIABLES CUANTITATIVAS

Variable estadística cp_prod = "Producción"

```
tabla.def<-sjmisc::frq(bd$cp_prod,auto.grp=nclass.Sturges(bd$cp_prod))
tabla.def
write.csv(tabla.def,file="tabla.def1.csv")
hist(bd$cp_prod,breaks=nclass.Sturges(bd$cp_prod),
    col="green3",
    main="Histograma de frecuencia de la variable producción del plátano",
    xlab="Producción",ylab="Frecuencia",cex.main=0.75)
```

Medidas de posición

```
# Medidas de posición central (Medidas de tendencia central o de centralización)
Media<-round(mean(bd$cp_prod),2)
Mediana<-median(bd$cp_prod)
Moda<-Mode(bd$cp_prod)
MMM<-c(Media,Mediana,Moda)
mmm<-cbind(MMM)
rownames(mmm)<-c("Media","Mediana","Moda")
mmm
# Medidas de posición no central (Medidas de colocación o de posición relativa)
Cuartiles<-quantile(bd$cp_prod, c(.25,.50,.75))
cbind(Cuartiles)
```

Medidas de dispersión (Medidas de variabilidad)

```
# Medidas de dispersión absolutas
# Rango<-range(bd$cp_prod)
# IQR<-(quantile(bd$cp_prod,.75,names=F)-quantile(bd$cp_prod,.25,names =F))
Varianza<-round(var(bd$cp_prod),2)
Desviacion<-round(sd(bd$cp_prod),2)
# Medidas de dispersión relativa
CV<-round(sd(bd$cp_prod)/mean(bd$cp_prod),2)
Dispersion<-c(Varianza,Desviacion,CV)
dispersion<-cbind(Dispersion)
rownames(dispersion)<-c("Varianza","Desviación Típica","Coeficiente de Variación")
dispersion
```

Medidas de forma

```
# Asimetría
asimetria<-round(skewness(bd$cp_prod,finite=T),2)
# Curtosis
curtosis<-round(kurtosis(bd$cp_prod,finite=T),2)
Forma<-c(asimetria,curtosis)
forma<-cbind(Forma)
rownames(forma)<-c("Coeficiente de asimetría","Coeficiente de curtosis")
forma
```

VARIABLES CUALITATIVAS

Variable estadística cp_k404 = "Condición de cultivo"

```
sjmisc::frq(bd$cp_k404)
```

Nota: Los códigos mencionados anteriormente solo es para la primera variable, si deseamos obtener resultados de otras variables se puede cambiar fácilmente con tan solo modificar la parte de los puntos de esta pequeña parte del código bd**$.....**

ANÁLISIS BIVARIADO DE DATOS

Cargar las librerías

```
library("psych")
```

Lectura de la base de datos

```
bd<-read.csv(file="BD141516.csv",header=T,sep=";",dec=".")
# str(bd1)
dim(bd)
```

Gráfico de la correlación de Pearson para las variables cuantitativas

```
pairs.panels(bd[,c(1:6)],method="pearson")
```

Contraste de correlación de Pearson para las variables cuantitativas

```
r.test <-function(m,method){
  n<-0
  # Matriz que contendrá los p-value
  p<-matrix(rep(0,ncol(m)^2),nc=ncol(m),nr=ncol(m))
  colnames(p)<-rownames(p)<-colnames(m)
  # Matriz que contendrá las correlaciones
  r<-matrix(rep(1,ncol(m)^2),nc=ncol(m),nr=ncol(m))
  colnames(r)<-rownames(r)<-colnames(m)
  for (i in 1:(ncol(m)-1)) {
    for (j in (i+1):ncol(m)) {
      n <- n+1
      test<-cor.test(m[,i],m[,j],method=method)
      p[i,j]<-p[j,i]<-test$p.value
      r[i,j]<-r[j,i]<-test$estimate
    }
  }
  return(list("Method"=test$method,"r"=r,"p-value"=p))
}
# Se hará una correlación entre todas la variables
testp<-r.test(bd[,c(1:6)],method="pearson")
print(testp)
write.csv(testp,file="testpearson.csv")
```

Análisis de independencia para datos cualitativos

```
tabla1.2<-table(bd$cp_k404,bd$cp_k408)
tabla1.2
fisher.test(tabla1.2)
tabla1.3<-table(bd$cp_k404,bd$cp_k413)
tabla1.3
chisq.test(tabla1.3)
```

ANÁLISIS FACTORIAL DE DATOS MIXTOS

Cargar las librerías

```
library("FactoMineR")
library("factoextra")
library("corrplot")
```

```
library("dplyr")
```

Lectura de la base de datos

```
bd<-read.csv(file="BD141516.csv",header=T,sep=";",dec=".")
# str(bd1)
dim(bd)
```

Recodificar las variables para evitar errores

```
bd.r<-mutate(bd,X13=recode(X13,"SI"="SI_X13","NO"="NO_X13"),
             X15=recode(X15,"SI"="SI_X15","NO"="NO_X15"),
             X16=recode(X16,"SI"="SI_X16","NO"="NO_X16"),
             X17=recode(X17,"SI"="SI_X17","NO"="NO_X17"),
             X18=recode(X18,"SI"="SI_X18","NO"="NO_X18"),
             X19=recode(X19,"SI"="SI_X19","NO"="NO_X19"))
```

Primer análisis factorial de datos mixtos

```
res.famd<-FAMD(bd.r,graph=F,ncp=17)
print(res.famd)
# Gráfico de sedimentación
fviz_screeplot(res.famd,addlabels=TRUE)
# Resultados del AFDM para las variables
var<-get_famd_var(res.famd)
var
# Gráfico de variables
fviz_famd_var(res.famd,"var",col.var="black",shape.var=15,
              repel=TRUE)
# Diagrama de barras del cos2 de variables en las dos primeras dimensiones
fviz_cos2(res.famd,choice="var",axes=1:2)
# Cos2: calidad de representación en el mapa de factores
fviz_famd_var(res.famd,"var",col.var="cos2",
              gradient.cols=c("green","blue","red"),
              repel=T)
# Diagrama de barras de las contribuciones de las variables en las dos primeras dimensiones
fviz_contrib(res.famd,choice="var",axes=1:2,top=14)
# Contrib: contribuciones de las variables en el mapa de factores
fviz_famd_var(res.famd,"var",col.var="contrib",
              gradient.cols=c("green","blue","red"),
              repel=T)
```

Recodificar las variables para el segundo AFDM

```
bd.r<-mutate(bd,X15=recode(X15,"SI"="SI_X15","NO"="NO_X15"),
             X17=recode(X17,"SI"="SI_X17","NO"="NO_X17"),
             X19=recode(X19,"SI"="SI_X19","NO"="NO_X19"))
```

Segundo análisis factorial de datos mixtos

```
res.famd<-FAMD(bd.r,graph=F,ncp=17)
print(res.famd)
# Gráfico de sedimentación
fviz_screeplot(res.famd,addlabels=TRUE)
# Resultados del AFDM para las variables
var<-get_famd_var(res.famd)
var
# Gráfico de variables
fviz_famd_var(res.famd,"var",col.var="black",shape.var=15,
              repel=TRUE)
# Diagrama de barras del cos2 de variables en las dos primeras dimensiones
fviz_cos2(res.famd,choice="var",axes=1:2)
# Cos2: calidad de representación en el mapa de factores
fviz_famd_var(res.famd,"var",col.var="cos2",
              gradient.cols=c("green","blue","red"),
              repel=T)
# Diagrama de barras de las contribuciones de las variables en las dos primeras dimensiones
fviz_contrib(res.famd,choice="var",axes=1:2,top=14)
# Contrib: contribuciones de las variables en el mapa de factores
fviz_famd_var(res.famd,"var",col.var="contrib",
              gradient.cols=c("green","blue","red"),
              repel=T)
# Análisis de las variables cuantitativas
quanti.var<-get_famd_var(res.famd,"quanti.var")
quanti.var
# Representar las variables cuantitativas (círculo de correlaciones)
fviz_famd_var(res.famd,"quanti.var",repel=T,
              col.var="black")
# Análisis de las variables cualitativas
quali.var<-get_famd_var(res.famd,"quali.var")
quali.var
##########################################
# visualizar variables cualitativas
```

```
# Representar las categorias de las variables cualitativas en el mapa factorial
fviz_famd_var(res.famd,"quali.var",repel=TRUE,
              col.var="black")
```

ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE CON VARIABLES DUMMY

Cargar las librerías

```
library("psych")
library("GGally")
library("car")
library("lmtest")
library("dplyr")
```

Lectura de la base de datos

```
bd<-read.csv(file="BD141516.csv",header=T,sep=";",dec=".")
# str(bd1)
dim(bd)
```

1. Analizar la correlación entre cada par de variables cuantitativas y diferencias del valor promedio entre las categóricas

```
# Analizar la relación entre variables cuantitativas
ggpairs(bd[,c(1:5)],lower=list(continuous="smooth"),
        diag=list(continuous="bar"),axisLabels="none")
# Analizar las diferencias del valor promedio entre las categóricas
ggplot(data=bd,mapping=aes(x=X15,y=Y,color=X15))+
geom_boxplot()+geom_jitter(width=0.1)+theme_bw()+theme(legend.position="none")

ggplot(data=bd,mapping=aes(x=X17,y=Y,color=X17))+
geom_boxplot()+geom_jitter(width=0.1)+theme_bw()+theme(legend.position="none")

ggplot(data=bd,mapping=aes(x=X19,y=Y,color=X19))+
geom_boxplot()+geom_jitter(width=0.1)+theme_bw()+theme(legend.position="none")
```

2. Generar el modelo

```
arlmvd<-lm(Y~.,data=bd)
summary(arlmvd)
Anova(arlmvd)
```

3. Selección de los mejores predictores

```
step(object=arlmvd,direction="both",trace=1)
# El mejor modelo resultante del proceso de selección ha sido:
```

```
arlmvd<-lm(Y~X3+X4+X5+X10+X15+X19,data = bd)
summary(arlmvd)
Anova(arlmvd)
```

4. Validación de condiciones para la regresión múltiple lineal

```
# 1. Relación lineal entre los predictores numéricos y la variable dependiente:
# Grafico de componentes más residuos para detectar la falta de linealidad
crPlots(arlmvd)
# Grafico de la variable dependiente contra las variables independientes para detectar
problemas de linealidad
avPlots(arlmvd)
# Contraste RESET de Ramsey
resettest(Y~X3+X4+X5+X10+X15+X19,power=2:3,type="regressor",data=bd)
# 2.Distribución normal de los residuos:
# Grafico QQ de los residuos para comprobar la normalidad de los residuos
qqPlot(arlmvd)
# Test de Shapiro-Wilk para contrastar la normalidad
shapiro.test(arlmvd$residuals)
# 3. Variabilidad constante de los residuos (Homocedasticidad):
# Grafico del valor absoluto de los residuos estandarizados contra los valores ajustados
spreadLevelPlot(arlmvd)
# Goldfeld-Quandt Test
gqtest(arlmvd)
# 4. No multicolinialidad:
# Analsis de los factores de inflacion de la varianza (FIV)
vif(arlmvd)
# 5. Autocorrelación:
# Estadístico de Durbin Watson para detectar la presencia de autocorrelación
plot(residuals(arlmvd))
dwtest(arlmvd,alternative="two.sided")
```

Anexo B. Pasos para acceder a los datos

1. Ingresar al sitio web: www.ecuadorencifras.gob.ec/ en donde se visualiza el inicio del Instituto nacional de estadística y censos INEC. (Ver Figura 1-B).

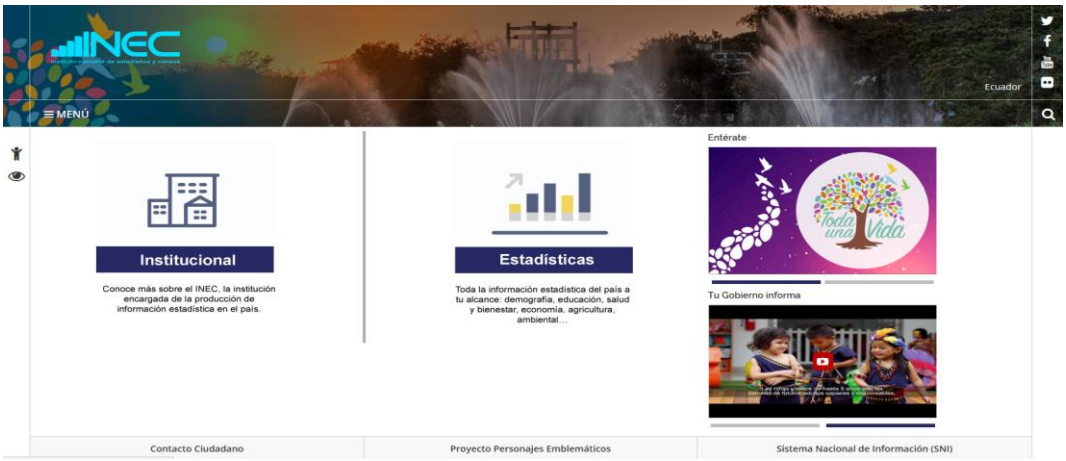


Figura 1-B: Inicio del sitio web INEC

2. Hacer clic en “Estadísticas” para visualizar la Figura 2-B.

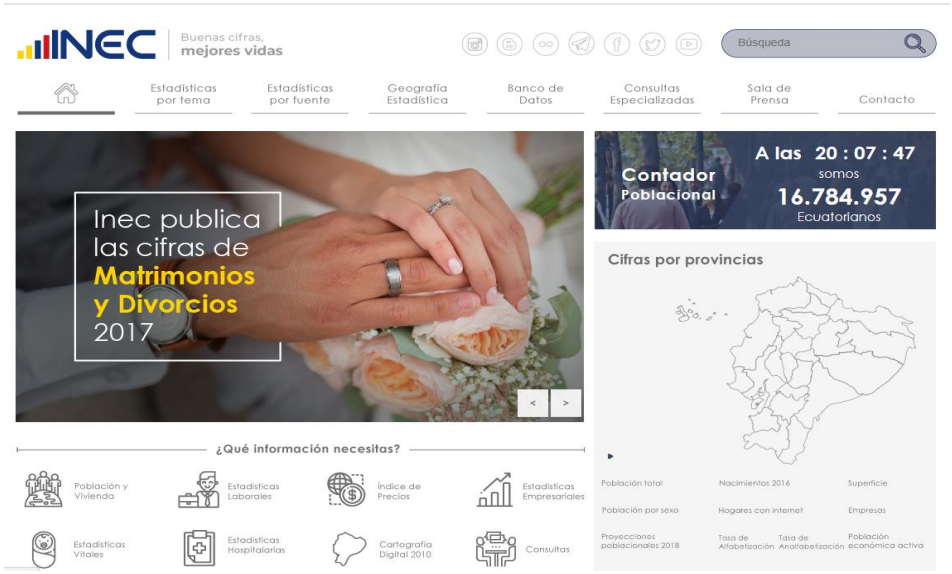


Figura 2-B: Sitio una vez hecho clic en “Estadísticas”

3. Hacer clic en la opción “Estadísticas por fuente” y luego en “Estadística de Información Ambiental Económica en GAD Municipales”, para visualizar la Figura 3-B.

Figura 3-B: Sitio una vez hecho clic del paso 2

4. Por último, clic en “Encuesta de Superficie y Producción Agropecuaria” para poder acceder a los datos (Ver Figura 4-B).

Figura 4-B: Acceso a los datos de la ESPAC