CIMAT

Centro de Investigación en Matemáticas, A.C.

Aplicación de técnicas de reducción de dimensionalidad sobre datos de la ENA2019

TRABAJO FINAL

Que para obtener el grado de

Maestro en Análisis Estadístico y Computación

CIMAT-INEGI

Presenta

Andrés Lira Tiscareño

Director de proyecto:

Dra. Lilia Leticia Ramírez Ramírez

Autorización de la versión final

7 de octubre de 2022

Resumen

El Instituto Nacional de Estadística y Geografía (INEGI) proporciona información estadística básica y derivada, relacionada con los sectores: población, economía, agropecuario entre otros. Dicha información es obtenida a partir de eventos como censos, encuestas y registros administrativos.

Un censo o encuesta agropecuaria proporciona bastante información, a través de un cuestionario que es elaborado con preguntas bien seleccionadas, de tal manera que se mantenga informado al país sobre la situación de éste en el tema del sector agropecuario y con la finalidad de que aporte elementos para la toma de decisiones gubernamentales, de empresarios o para caracterizar las unidades de producción agropecuarias entre otras cosas. Así pues, un evento de este tipo recolecta tanta información que el extraerla y analizarla en conjunto puede representar un reto.

Se observa que la información estadística básica proveniente de un censo o encuesta no es suficiente para comprender o entender los datos, lo cual motiva para aplicar técnicas estadísticas multivariadas utilizando datos agropecuarios. Así pues, el objetivo general de este trabajo es aplicar métodos estadísticos como análisis de factores, escalamiento multidimensional, análisis de correspondencia, árboles de clasificación y bosques de clasificación para explotar la información de la Encuesta Nacional Agropecuaria 2019 (ENA 2019) y caracterizar las entidades federativas e incluso los municipios si fuese necesario, a través de las unidades que producen ciertos cultivos como maíz, frijol, chile por mencionar algunos, o unidades que crían alguna especie ganadera como bovinos, porcinos o aves de corral, entre otras. Además de representar al país o al estado por medio de las características generales de la unidad de producción agropecuaria de la ENA 2019, con la intención de prepararse para replicar los métodos multivariados mencionados, pero ahora a los datos del Censo Agropecuario 2022 (CA 2022).

Índice

Motivación

El sector agropecuario es la unión de dos subsectores que forman parte del sector primario: El subsector agrícola en el cual se consideran todas las actividades relacionadas con la agricultura y el subsector pecuario cuyas actividades están relacionadas con la ganadería.

El Instituto Nacional de Estadística y Geografía (INEGI) es una de las instituciones que proporciona información del sector agropecuario en México, a través de proyectos como censos o encuestas agropecuarias. Es importante mencionar que en el año 1930 dio inicio el levantamiento de los censos agropecuarios, de esta manera, los siguientes censos se llevaron a cabo en los años 1940, 1950, 1960, 1970, 1981, 1991 y 2007. Desafortunadamente, una de las principales causas por las que no se ha realizado otro censo agropecuario es debido al presupuesto, es decir, no se cuenta con los recursos económicos para realizar un evento tan grande. Debido a ésto y con la intención de que el INEGI proporcione información agropecuaria actualizada del país, se han realizado Encuestas Nacionales Agropecuarias (ENA) en los años 2012, 2014, 2017 y 2019. Más aún, cabe resaltar que se pensó en realizar cada dos años este tipo de encuestas, sin embargo, con la intención de ya realizar un censo, en el año 2016 se llevó a cabo una Actualización al Marco Censal Agropecuario (AMCA 2016), con la finalidad de crear un directorio (listado) de terrenos y así prepararse para que en 2017 se realizará el levantamiento de información para un posible censo agropecuario 2017. Sin embargo, no hubo presupuesto suficiente, por lo cual se optó por hacer la ENA 2017. La siguiente encuesta realizada fue la ENA 2019, que después seguiría la ENA 2021, pero ésta última no se lleva a cabo ya que en 2021 se autoriza, finalmente, el presupuesto para realizar el Censo Agropecuario 2022 (CA 2022). Así pues, de septiembre a noviembre de 2022 se llevará a cabo el levantamiento de la información del CA 2022 considerando que la publicación de la información será en el año 2023 y ésta hace referencia al periodo de octubre 2021 a septiembre de 2022.

Una encuesta agropecuaria, como las ENA que se han realizado en años pasados proporcionan bastante información de cada una de las unidades de producción que son seleccionadas y son parte de una muestra representativa, donde la **unidad de observación** es la unidad de producción, la cual se considera como la unidad económica conformada por uno o más terrenos ubicados en un mismo municipio, en donde al menos en alguno de ellos se realizan actividades agropecuarias o forestales, bajo la administración de un mismo productor. Sin embargo, un censo proporciona más

información que una ENA, por ejemplo, las ENA proporcionan información de los principales productos agropecuarios del país y para una muestra de unidades de producción, la cual, se expande al universo, mientras que un censo no se limita a los principales productos si no que éste capta la información de todos los productos agropecuarios en todas las unidades de producción.

La unidad de observación de la Encuesta Nacional Agropecuaria 2019 fue la unidad de producción agropecuaria, y que suman 29 productos agropecuarios los cuales son considerados principales en el país. Algunos productos de interés fueron el aguacate, la alfalfa, el arroz, el maíz, el frijol, los bovinos, los porcinos, las aves de corral, entre otros. Más aún, por mencionar algunas variables de las cuales se obtuvo información son: superficie de la unidad de producción (en hectáreas), superficie sembrada y cosechada de los distintos cultivos de interés, destino de la producción, uso de fertilizantes, tecnologías en agricultura o en ganadería, maquinaria, mano de obra, entre otras. Además, su objetivo principal fue conocer las características de las unidades de producción agrícolas y ganaderas, ofreciendo datos numéricos y categóricos de la producción en zonas rurales.

El contar con demasiadas variables en un censo o en una encuesta complica el análisis para conocer cómo se puede caracterizar las zonas de México, por lo cual, las técnicas estadísticas de reducción de dimensionalidad ayudan a analizar, entender y poder realizar clasificaciones de zonas en el país, como, por ejemplo, identificar en qué estados se produce la fresa y con cuales otros productos tienden a producirse. Con la finalidad de analizar y clasificar la información que se obtendrá en el CA 2022, se utiliza información de la ENA 2019 para aplicar técnicas estadísticas como análisis de factores, escalamiento multidimensional, análisis de correspondencia, árboles de clasificación y bosques de clasificación, que ayuden a analizar, clasificar y entender mejor la información.

El objetivo primordial de este trabajo es utilizar la información de algunas variables de la ENA 2019, con la finalidad de aplicar las técnicas estadísticas mencionadas anteriormente e interpretar los resultados obtenidos para que, en un futuro próximo, se repliquen dichas técnicas en los datos del CA 2022, pero ahora con la información que se obtendrá en el censo agropecuario. De esta forma, as su vez se pueden realizar comparativos entre lo obtenido con la ENA 2019 y el CA 2022, sin dejar de considerar que ambos eventos cuentan con información referenciada a distinto año, en cuyo caso, existen factores, como climáticos, por mencionar alguno, que afectan a los

productos agropecuarios, o factores que benefician como los programas de gobierno (federal, estatal o municipal).

Cuando se llevan a cabo proyectos como las encuestas nacionales agropecuarias o los censos agropecuarios, se obtiene demasiada información en una gran cantidad de variables, por lo que, para entender mejor ésta es importante realizar distintos análisis reduciendo el número de variables, para utilizar aquellas que aporten más información y entenderla mejor, de tal manera que se obtengan y se interpreten más fácilmente los resultados importantes de las encuestas o los censos, como por ejemplo, qué es lo que caracteriza a una unidad de producción en el país o en cada entidad federativa, incluso en cada municipio, que es el nivel de desagregación más bajo que puede tener la unidad de producción agropecuaria. De esta manera, unas de las características que distinguen a una unidad de producción pueden ser, por ejemplo, las existencias totales de ganado bovino, porcino o aves de corral con las que se cuenta a nivel nacional y a nivel entidad, la superficie sembrada y cosechada en hectáreas de cada cultivo, la producción obtenida en toneladas de cada cultivo así como el tipo de tecnología, maquinaria y sistemas de riego que utilizan las unidades de producción, los tipos de estructura que se utilizan y cultivos que se siembran en la agricultura protegida, entre otras cosas importantes.

Antecedentes

El INEGI genera estadísticas de distintos proyectos (censos o encuestas), como, por ejemplo, el censo de población y vivienda que se realiza cada 10 años, los censos económicos que se llevan cada cinco años, el censo agropecuario, y en cuestión de encuestas se pueden mencionar algunas como la Encuesta Nacional de Hogares (ENH), Encuesta Nacional Agropecuaria (ENA) entre otras. También, genera estadísticas a través de registros administrativos, además de publicar distintos índices como el índice de precios, el índice global de personal y remuneraciones de los sectores económicos entre otros.

Toda la información generada y publicada por el INEGI se encuentra en la dirección https://www.inegi.org.mx/ y en la revista internacional de estadística y geografía realidad, datos y espacio cuya página web es https://rde.inegi.org.mx/. Ahora bien, realizando una búsqueda de

información para identificar los avances y/o antecedentes que se tienen en el sector agropecuario (censos y encuestas agropecuarias) sobre cómo y de qué manera se han aplicado las técnicas estadísticas: análisis factorial, escalamiento multidimensional, análisis de correspondencia, árboles de clasificación y bosques de clasificación; se observó que no existen publicaciones (o al menos no se encontraron) que expliquen las aplicaciones de estas técnicas a la información agropecuaria, por tal motivo se realizó una segunda búsqueda para encontrar la aplicación de dichas técnicas pero en otro tipo de información (encuestas y/o censos ajenos al agropecuario). A continuación, se mencionan los trabajos identificados.

En la **conformación del marco nacional de viviendas en hogares** se tiene conocimiento que se utilizó el análisis de factores, mientras que el análisis de correspondencias múltiples se menciona su aplicación en un artículo publicado en la revista de INEGI, el cual se denomina "**Análisis comparativo de metodologías utilizadas para la medición de la corrupción**" y se encuentra ubicado en https://rde.inegi.org.mx/index.php/2019/04/23/analisis-comparativo-de-metodologias-utilizadas-para-la-medicion-de-la-corrupcion/, en cuyo caso se utiliza la técnica estadística para comparar varios índices que miden la corrupción, como el índice de percepción de la corrupción (IPC), el estimado de control de la corrupción (ECC) del banco mundial, el índice de fuentes de soborno (IFS), el índice global de la competitividad (IGC) entre otros. Además, otro artículo de nombre "**Construcción de un índice compuesto y aproximación para medir los cambios en el tiempo**" y publicado en la revista del INEGI en la dirección: https://rde.inegi.org.mx/index.php/2014/05/05/construccion-de-un-indice-compuesto-y-aproximacion-para-medir-los-cambios-en-el-tiempo/, menciona que se aplicaron técnicas como análisis de componentes principales y el análisis factorial con la finalidad de construir un índice compuesto que resuma la información contenida en fenómenos de naturaleza multidimensional.

También, en la página de INEGI estaba una herramienta llamada **Estratificador INEGI** que permitía hacer la reducción de dimensiones utilizando la información del censo de población de 1990 y el censo de población de 2000, junto con los métodos estadísticos como "k-medias" y análisis de conglomerados. El Estratificador INEGI considera actualmente la información del censo de población 2020.

Además, es importante mencionar que el INEGI cuenta con el **atlas de género** cuya ubicación está en la página web http://gaia.inegi.org.mx/atlas_genero/, y el cual utiliza información de población en general, educación, salud, trabajo, toma de decisiones, pobreza, emprendimiento, violencia y

población indígena, con la finalidad de estratificar el país de México, desconociendo los métodos estadísticos que hacen posible dicha estratificación, ya que no se encontró publicada.

Más aun, la página web https://gaia.inegi.org.mx/scince2020/ utiliza información del censo de población 2020, como población, vivienda, fecundidad, mortalidad, migración, etnicidad, discapacidad, educación, características económicas, servicios de salud, situación conyugal, hogares censales y religión. A través de dicha información y con ayuda de técnicas estadísticas multivariadas, se hace reducción de dimensionalidad estratificando el país México, por entidad federativa, por municipio o localidad urbana.

Sin embargo, no se encontró información publicada donde se mencione que hayan aplicado las técnicas de escalamiento multidimensional, árboles de clasificación y bosques de clasificación, pues posiblemente si se han utilizado dichas técnicas en distintos eventos del INEGI (censos o encuestas) pero no se han publicado los resultados obtenidos a través de éstas, o también, debido a la inmensa información con la que se cuenta en la página de INEGI y la gran cantidad de artículos publicados en la revista internacional de INEGI, no se encontró la información que se buscaba.

Descripción de datos

Para levantar la información de la Encuesta Nacional Agropecuaria 2019 (ENA2019) se diseñó un cuestionario que cuenta con poco más de 600 preguntas, las cuales captan tanto información en variables numéricas como en categóricas (ordinales y no ordinales). La documentación de la ENA 2019 como el cuestionario, la presentación de resultados, la metodología y algunos resultados generales (mini monografías); además de los tabulados publicados, microdatos y datos abiertos se pueden consultar en la dirección: https://www.inegi.org.mx/programas/ena/2019/ .

Es importante recalcar que la unidad de observación de la encuesta es la Unidad de Producción (UP), la cual se considera como la unidad económica conformada por uno o más terrenos ubicados en un mismo municipio, en donde al menos en alguno de ellos se realizan actividades agropecuarias o forestales, bajo la administración de un mismo productor. De esta manera, un productor puede tener una o más unidades de producción. Además, en la ENA 2019 se captó información general para una muestra representativa de 60 mil UP aproximadamente sobre los

temas: categoría jurídica, organización y apoyo, clasificación de la unidad de producción, características generales de los terrenos, uso de suelo, sistemas de riego calidad y origen del agua, agricultura, cría y explotación de animales, tractores maquinaria y vehículos, mano de obra y remuneraciones, crédito y seguro, tecnologías informáticas y de comunicación, problemática, medio ambiente, características sociodemográficas del productor y algunos datos de identificación.

Una unidad de producción (UP) es manejada por un productor, donde éste puede ser una persona física, una persona moral (empresa, grupo, sociedad, asociación o unión), el gobierno u otro tipo de organización, considerando que los datos sociodemográficos sólo se captan para las personas físicas. Más aún, la información de la unidad de producción está referenciada a dos fechas; por ejemplo, la información de organización y apoyo de la UP, el uso de suelo, la información de los cultivos, la venta de bovinos, porcinos y aves de corral, la tecnología de bovinos, porcinos y aves de corral, los tractores, maquinaria y vehículos, la mano de obra y remuneraciones entre otras está referenciada a octubre del año 2018 y septiembre de 2019; mientras que las existencias de ganado bovino, porcino y aves de corral están referenciadas a dos días: el 31 de marzo de 2019 y el 30 de septiembre de 2019.

La información captada puede registrarse en distintas unidades de medida, y uno de los procesos que se realiza antes de publicarla, es codificar y normalizar a unas cuantas unidades de media, por ejemplo, la clave de los cultivos se codifica y los distintos tipos de superficie se normalizan a hectáreas, la producción obtenida de un cultivo a cielo abierto se normaliza a toneladas, sin embargo, la producción de cultivos en agricultura protegida (cultivos producidos en invernaderos, viveros, malla sombra, macro túnel entre otros tipos de instalaciones) se normaliza en varias unidades como toneladas, plantas, piezas, costales, cajas, etc. También, el precio de venta de bovinos y porcinos se da por cabeza, canal, kilogramo, peso, etc. Sin embargo, con mayor frecuencia el precio es por cabeza.

Para el análisis de la información, dada la situación y la relevancia de las variables captadas, se optó por hacer un filtrado de información y combinación de algunas de ellas, seleccionando aquellas variables que puedan ser de mayor interés para la sociedad y que caractericen a las unidades de producción, así como obtener una mejor representación de los resultados, eligiendo un total de 25 nuevas variables, las cuales se menciona en la tabla 1. Consideremos el ejemplo, la superficie agrícola (SUP_AGRICOLA) es una nueva variable formada a partir de las variables del

cuestionario (ver Anexos) que contienen la información de superficie de cultivos que duren menos de un año (US111_02) más la superficie de cultivos que duren más de un año (US111_03) más la superficie dedicada a la agricultura pero que no se sembró (US112_01). También, es importante mencionar que de las 25 variables sólo la cantidad de ganado bovino "BOV", el rendimiento de leche "RENDIMIENTO_LECHE", el ingreso obtenido al vender la leche "INGRESO_LECHE" y la cantidad de ganado porcino "PORC" contienen información referenciada a un día (30 de septiembre de 2019), mientras que la información del resto de variables está referenciada a un año (octubre de 2018 a septiembre de 2019).

Tabla 1. Descripción de 25 variables formadas para análisis.

CLAVE DE VARIABLE	DESCRIPCIÓN DE VARIABLE	VARIABLES DEL CUESTIONARIO INVOLUCRADAS
SUP_AGRICOLA	SUPERFICIE AGRICOLA (EN HECTÁREAS)	US111_02, US111_03, US112_01
OTRAS_SUP	OTRAS SUPERFICIES (EN HECTÁREAS)	US211, US313, US412
SUP_RIEGO	SUPERFICIE DE RIEGO (EN HECTÁREAS)	AR111_02
CULTIVOS_CA	CANTIDAD DE CULTIVOS OBJETO DE INTERES Y EN CIELO ABIERTO	CONTEO DE AA111_02
RENDIMIENTO_CA	RENDIMIENTO (PRODUCCION TOTAL (TON) / SUPERFICIE COSECHADA TOTAL (HA)) CONSIDERANDO CULTIVOS OBJETO DE INTERES Y EN CIELO ABIERTO	AA111_17, AA111_13
VALOR_PROD_CA	VALOR DE LA PRODUCCIÓN AGRICOLA=PRODUCCIÓN TOTAL * PRECIO PROMEDIO DE VENTA	AA111_17, DA132_09
TECNOLOGIA_CA	NÚMERO DE TECNOLOGIAS AGRICOLAS UTILIZADAS EN CULTIVOS A CIELO ABIERTO	AT111_42, AT111_27_01, AT111_27_02, AT111_28_01, AT111_28_02, AT111_30_01, AT111_30_02, AT111_32, AT111_33, AT111_34, AT111_35, AT112_04, AT112_06, AT112_11, AT112_08, AT112_02, AT112_03, AT112_09, AT112_13, AT112_14, AT112_15, AT112_05, AT112_07, AT112_10, AT112_16, AT112_99
BOV	CANTIDAD DE GANADO BOVINO	CB112
RENDIMIENTO_LECHE	RENDIMIENTO=PRODUCCION DE LECHE/ VACAS QUE ORDEÑO	CB122_01, CB121_061
INGRESO_LECHE	INGRESO_LECHE=PRODUCCION DE LECHE*PRECIO DE VENTA	CB122_01, CB122_05
INGRESOS_VTA_BOV	INGRESO_VENTA_BOV=BOVINOS VENDIDOS*PRECIO PROMEDIO DE VENTA DE ANIMALES POR CABEZA	CB172_01, CB173
PORC	CANTIDAD DE GANADO PORCINO	CP112
INGRESOS_VTA_PORC	INGRESO_VENTA_PORC=PORCINOS VENDIDOS*PRECIO PROMEDIO DE VENTA DE ANIMALES POR CABEZA	CP172_01, CP173
TECNOLOGIA_BOV_PORC	CANTIDAD DE TECNOLOGIAS EN BOVINOS Y PORCINOS	CB141_272, CB141_99, CB141_01, CB141_02, CB141_03, CB141_04, CB141_05, CB141_07, CB141_08, CB141_09, CB141_10, CB141_12, CB141_121, CB141_13, CB141_14, CB141_16, CB141_17, CB141_18,

CLAVE DE VARIABLE	DESCRIPCIÓN DE VARIABLE	VARIABLES DEL CUESTIONARIO INVOLUCRADAS
		CB141_19, CB141_20, CB141_21, CB141_22, CB141_23, CB141_24, CB141_25, CB141_26, CB141_271, CP141_01, CP141_02, CP141_03, CP141_04, CP141_05, CP141_06, CP141_07, CP141_08, CP141_09, CP141_99
NUM_PROBLEMATICAS	NÚMERO DE PROBLEMATICAS REPORTADAS	PP111
ORGANIZACION_PROD	ORGANIZACIÓN CON OTROS PRODUCTORES PARA LA OBTENCIÓN DE APOYOS O SERVICIOS (SI/NO)	OP111
APOYO_PROD	RECIBIR APOYOS ECONÓMICOS PARA LA PRODUCCIÓN (SI/NO)	PO111
APOYO_SINIESTRO	APOYO ECONÓMICO RECIBIDO POR EL GOBIERNO FEDERAL POR CAUSA DE ALGÚN SINIESTRO (SI/NO)	SE116
ACCIONES_MEDIO_AMBIENTE	NÚMERO DE ACCIONES REALIZADAS PARA PROTECCION DEL MEDIO AMBIENTE	ME111
MANO_D_OBRA	CANTIDAD TOTAL DE MANO OBRA EMPLEADA	MO111, MO121, MO114, MO115, MO115_02, MO120
TRAC_MAQ_VEHI	CANTIDAD DE TRACTORES, MAQUINARIA Y VEHICULO PROPIOS DE LA UNIDAD DE PRODUCCIÓN (UP)	TR115_01, VE112_01, VE112_02, CONTEO DE MA114
SEXO_PROD	SEXO (HOMBRE/MUJER) DEL PRODUCTOR	SD113
EDAD_PROD	EDAD DEL PRODUCTOR (EN AÑOS)	SD114
ANIOS_EN_ACTIVIDAD	AÑOS QUE EL PRODUCTOR TIENE REALIZANDO LA ACTIVIDAD	SD122
NIVEL_D_ESTUDIOS	NIVEL DE ESTUDIOS DEL PRODUCTOR	SD118

Debido a la complejidad de información de la ENA 2019 y con la finalidad de restringir y caracterizar el universo de unidades de producción a trabajar, se observó que el 95.35% de las UP de la muestra seleccionada son manejadas por una persona física. De esta manera, considerando sólo las UP que son manejadas por una persona física y haciendo un poco más de análisis, se opta por restringir el universo de estudio agregando algunas otras caracterizaciones de las unidades de producción, seleccionando la información del 87.87% de las UP total de la muestra seleccionada de la ENA 2019.

Las características de las UP del universo de estudio se describen a continuación.

- UP donde el productor es una persona física.
- UP sólo con cultivos objeto de interés en agricultura a cielo abierto (cultivos que se producen sin protección como en los invernaderos, viveros, etc.) y que hayan reportado superficie cosechada y producción mayor a cero, o con al menos un bovino o al menos un porcino.

- Para el precio promedio de cultivos a cielo abierto, sólo se consideran las UP-cultivo que vendieron.
- Para los ingresos de venta de bovinos o porcinos sólo se consideran los vendidos por cabeza.

Por tanto, el universo de estudio queda constituido con el 87.87% de UP de la muestra total y con información de 25 variables generadas cuya información se utiliza sin ser estandarizada, además de utilizar la entidad (E03) donde se encuentra la unidad de producción y la llave que la hace única, que en este caso se formó utilizando un consecutivo.

Además, del universo de estudio, se construye una tabla de frecuencias para los 25 cultivos de interés en las 32 entidades federativas, con el propósito de observar en qué entidades se produce cierto cultivo a través de la cantidad de unidades de producción. A manera de un ejemplo, la tabla 2 muestra la información del número de unidades de producción que cultivan el Aguacate.

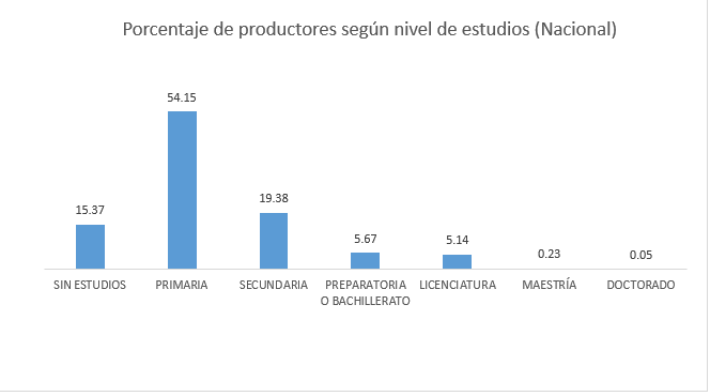
Tabla 2. Distribución de unidades de producción que cultivan Aguacate en cada Entidad.

Número de UP que producen Aguacate.			
ENTIDAD	NÚM DE UP	ENTIDAD	NÚM DE UP
Total	1035		
Aguascalientes	1	Morelos	57
Baja California	0	Nayarit	140
Baja California Sur	3	Nuevo León	1
Campeche	6	Oaxaca	10
Coahuila de Zaragoza	0	Puebla	118
Colima	22	Querétaro	2
Chiapas	6	Quintana Roo	18
Chihuahua	0	San Luis Potosí	1
Ciudad de México	0	Sinaloa	3
Durango	2	Sonora	0
Guanajuato	3	Tabasco	4
Guerrero	24	Tamaulipas	0
Hidalgo	3	Tlaxcala	2
Jalisco	51	Veracruz de Ignacio de la Llave	13
México	79	Yucatán	59
Michoacán de Ocampo	407	Zacatecas	0

Estadística descriptiva

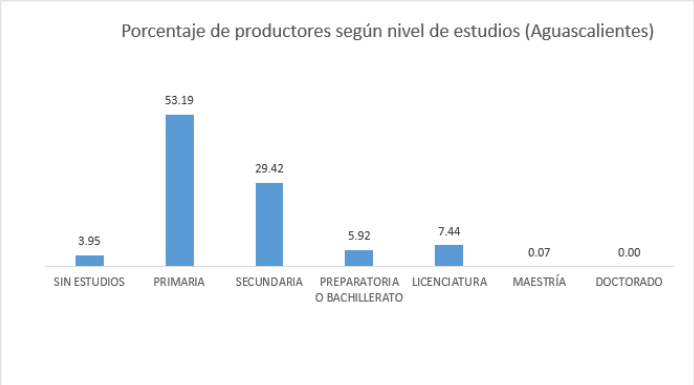
A continuación, se realiza un análisis descriptivo de las variables mencionadas en la sección anterior.

Si consideramos la variable de nivel de estudios que tiene el productor que está a cargo de una o varias unidades de producción agropecuarias, y éste la representamos tanto a nivel nacional como a nivel estatal, podemos observar los siguientes histogramas.

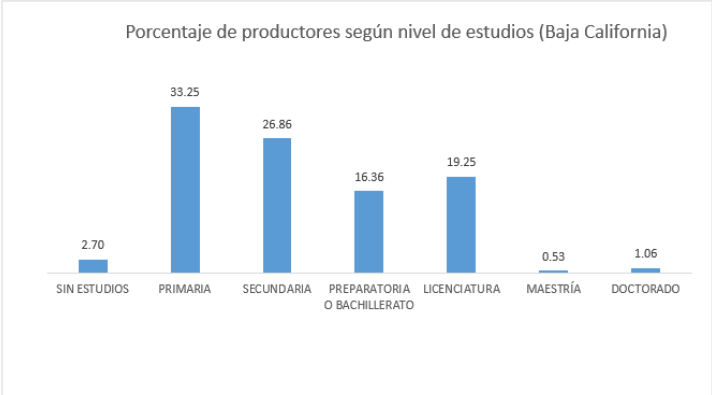


En el gráfico anterior se observa que, a nivel nacional, el 54.15% de los productores dijeron tener estudios de sólo primaria, seguido de éste el 19.38% dice haber asistido hasta la secundaria, sin embargo, es importante hacer notar que el 15.37% de los productores no asistieron a la escuela, es decir, no tienen estudios. Por tal motivo, 88.90% de los productores sólo alcanzó un nivel de estudios básico, mientras que el 11.10% tiene un nivel de estudios medio o superior. Así pues, es interesante observar que la mayoría de las personas que se dedican a la actividad agropecuaria son aquellas que posiblemente no tuvieron oportunidad de estudiar y por tal motivo se dedicaron a dicha actividad, que seguramente es la actividad realizada por sus padres.

Ahora bien, se analiza el porcentaje de nivel de estudios en el estado de Aguascalientes, el cual se muestra en el siguiente gráfico, observando casi la misma similitud que la información nacional, es decir, la mayoría de los productores tiene un nivel de estudios alcanzado como nivel básico, sin embargo, se observa que ya empieza a aumentar el porcentaje de productores que tiene un poco más de preparación alcanzando estudios de preparatoria, e incluso de licenciatura.



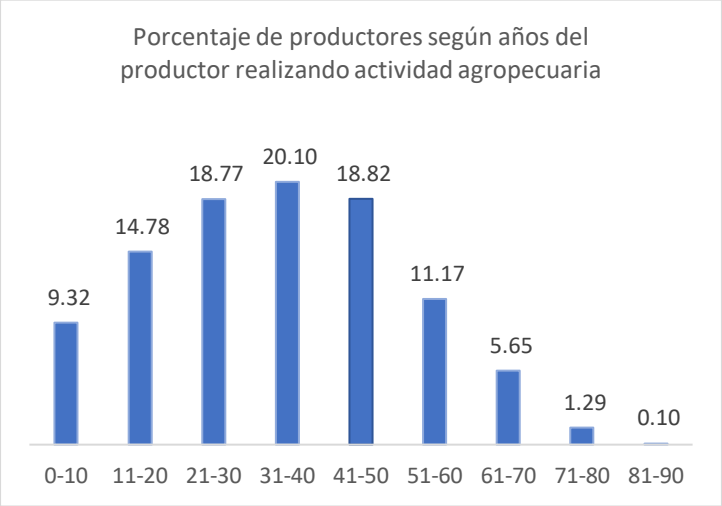
La entidad de Baja California Norte destaca sobre la nacional, al mostrar que los productores dedicados a la actividad agropecuaria ya están mejor preparados, pues el porcentaje de éstos disminuye teniendo una preparación de nivel básico y aumentando una preparación a nivel medio superior o superior, incluso se observa que la mayoría de productores que tienen nivel medio o superior entran en el grupo de aquellos que estudiaron una licenciatura, además de notar que posiblemente éstos se sigan preparando ya que empieza a haber un porcentaje mínimo de productores que tienen nivel de estudios de doctorado.



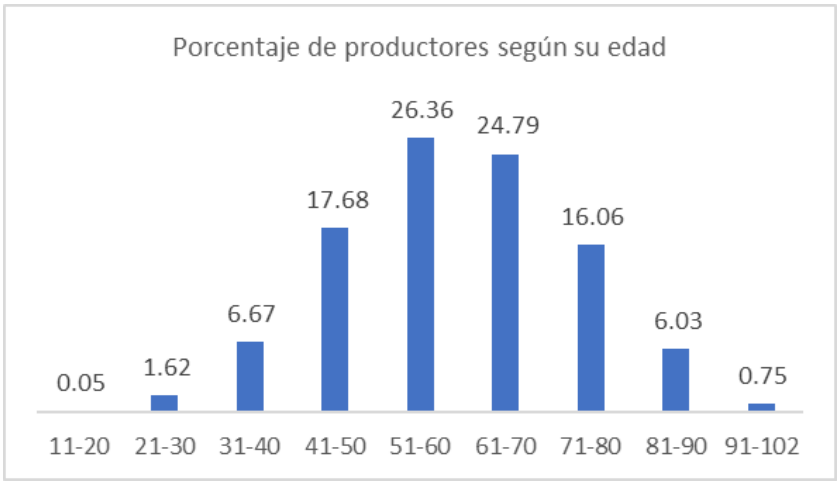
También, se realizaron gráficos del porcentaje de productores según su nivel de estudios para el resto de las entidades, los cuales se pueden consultar en la parte de Anexos.

Por otro lado, analizando la variable años que lleva el productor realizando la actividad agropecuaria, se observó que la mayoría de productores lleva entre 21 y 50 años realizando dicha actividad, según lo muestra el histograma siguiente, además de haber productores que tienen más de 50 años en la actividad agropecuaria, indicando que prácticamente toda su vida se han dedicado a la agricultura o a la ganadería, es decir, desde muy pequeños y seguramente por las

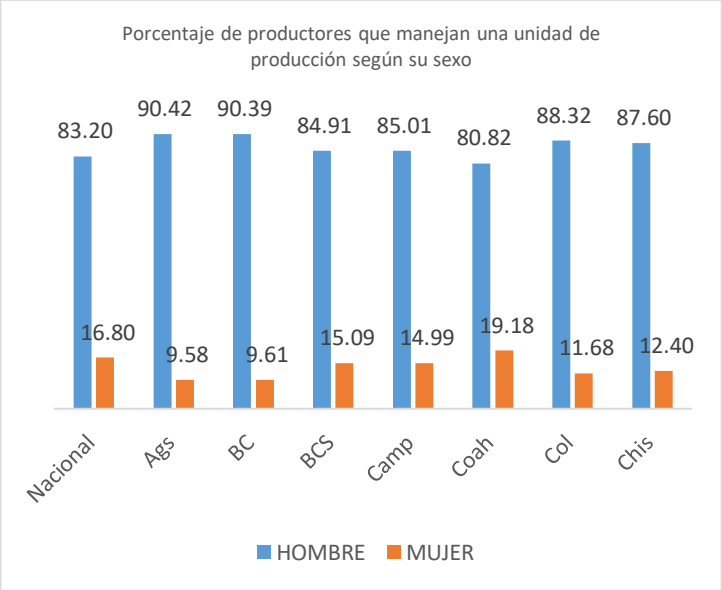
carencias o necesidades familiares se vieron obligados a realizar este tipo de actividades, que al final de cuentas la realizan con mucho cariño.



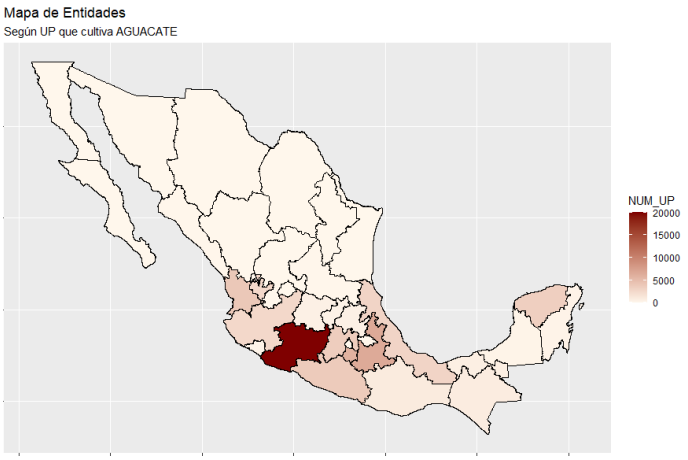
Ahora bien, considerando la edad en años que tiene el productor al momento de ser captada la información, en el siguiente histograma, se observa que la mayoría de los productores declararon tener entre 51 y 70 años, reforzando la información que proporcionó el gráfico anterior, es decir, que la mayoría de los productores han entregado su vida entera a las labores del campo realizando actividades agrícolas y pecuarias. Más aún, el histograma muestra un comportamiento parecido a una distribución normal.



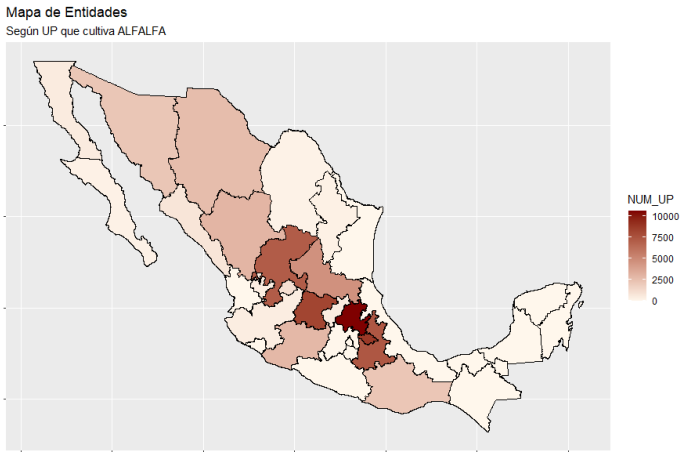
Además, el 83.20% de las unidades de producción son trabajadas por un productor de sexo masculino, mientras que el 16.80% es manejada por un productor de sexo femenino. Estos datos según se ilustra en el siguiente gráfico, son a nivel nacional, pues considerando una cierta entidad estos datos cambian relativamente poco, mostrando en cada entidad que la mayoría de productores son hombres.



A continuación, se muestran mapas con las entidades federativas para cada uno de los cultivos de interés de la ENA2019, ilustrando primero los cultivos de aguacate, alfalfa, café y en Anexos el resto de los cultivos (amaranto, arroz, cacao, calabacita, calabaza, caña de azúcar, cebolla, chile, fresa, frijol, jitomate “tomate rojo”, limón, maíz, mango, manzana, naranja, plátano, sorgo, soya, trigo y uva). En dichos mapas se observa qué estado es el que contiene la mayor cantidad de unidades de producción agrícolas que producen cierto cultivo, es decir, se calcula una distribución de frecuencias de unidades de producción por entidad federativa para cada cultivo y ésta se grafica en el mapa. Por ejemplo, en el siguiente mapa se observa que la entidad de Michoacán es la que contiene la mayor cantidad de unidades de producción que cultiva el aguacate y, por ende, se pudiera decir que Michoacán es la que produce más aguacate.

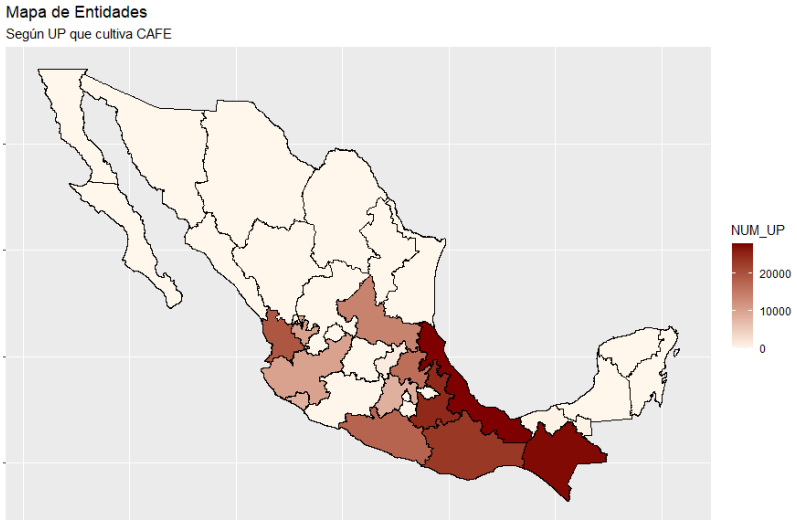


La principal ventaja de presentar los datos de producción estatales en estos mapas es que es fácilmente identificable factores como los geográficos, climáticos y económicos que impactan positivamente la producción. Así pues, es interesante resaltar que el estado con mayor número de unidades de producción que cultivan la alfalfa es Hidalgo, y seguido de éstos aparecen las entidades de Tlaxcala, Puebla, Guanajuato y Zacatecas, sin embargo, el mapa ilustra muy bien que aunque el número de unidades de producción de alfalta ya no es tan marcado en Sonora, Chihuahua, Durango, San Luis Potosí, Michoacán y Oaxaca, éstas sí producen dicho cultivo, pero en menor número de unidades de producción.



El café es un cultivo que es muy consumido por los mexicanos, por lo cual, suele ser importante identificar qué entidades son aquellas que tienen mayor número de unidades de producción agrícolas que lo producen. Como se ilustra en el siguiente mapa, las entidades de Veracruz y Chiapas son las que cuentan con un número mayor de unidades que producen café, seguidas de las entidades Puebla, Oaxaca y Nayarit, y aunque con menor frecuencia de unidades de

producción agrícolas, pero no por eso dejan de ser importantes las entidades Guerrero, México, Hidalgo, San Luis Potosí, Jalisco y Colima, mientras que el resto de los estados no cultivan el café.



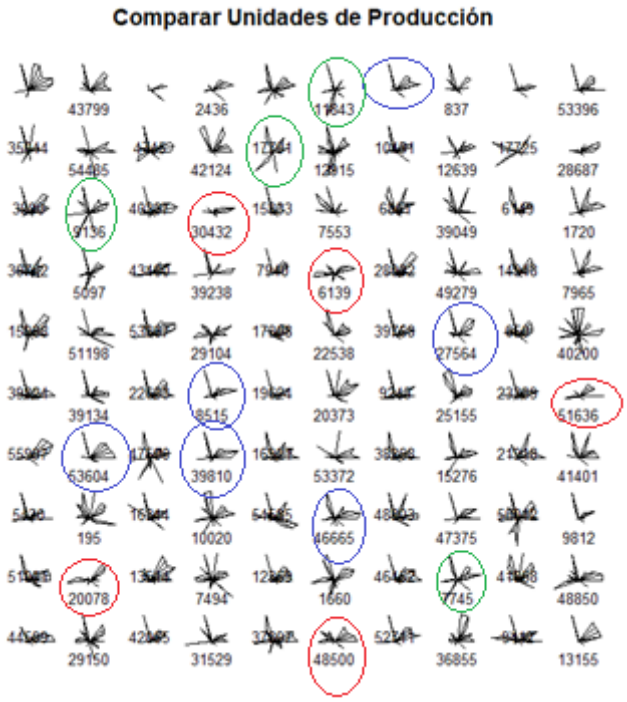
Por otro lado, con la finalidad de ilustrar e identificar si las unidades de producción agropecuarias son parecidas entre ellas, claro, considerando sus características descritas por las 25 variables que se mencionan en la sección anterior (Descripción de datos) se realiza un gráfico de estrellas, que es uno de los recursos básicos para identificar comunidades. En nuestro caso buscamos identificar unidades de producción que por la producción que presentan, puedan parecerse. Esto sin considerar su cercanía geográfica y otras características, sino meramente por la distribución de sus productos. Para fines ilustrativos consideramos algunas unidades de producción muestreados aleatoriamente y obtenemos el diagrama de estrellas. En dicho gráfico se encierran algunas estrellas que tienen cierto parecido en distintos colores agrupando aquellas que se parecen un poco más entre sí, ya sea por la forma general de cada estrella, como el que esté más aplanada en alguna dirección, etc.

Ahora, comparando estos tres grupos formados (estrellas en círculos azules, en círculos rojos y en círculos verdes) y claro, con el resto de las estrellas, se observa que:

1) Las estrellas en círculos azules representan unidades de producción en los cuales el productor cuenta con relativamente poca tecnología de cultivos en cielo abierto, poca tecnología para bovinos y porcinos, poca maquinaria, tractores y vehículos, y más aún relativamente poca mano de obra, lo cual impacta en que la superficie de riego es

pequeña así como las otras superficies, por lo cual, la cantidad de bovinos es mínima y en consecuencia éstos deben ser bovinos destinados para la carne ya que no se observa rendimiento de leche pero si ingresos por la venta de bovinos.

2) Las estrellas en círculos rojos se caracterizan por ser unidades de producción agropecuarias donde el productor recibió apoyos económicos para la producción de cultivos y/o especies ganaderas, además de manejar un promedio de 16 personas como mano de obra para las labores agropecuarias y tener poca tecnología para cultivos en cielo abierto, poca tecnología para bovinos y porcinos, y en consecuencia un promedio de 5 hectáreas de riego lo cual implica tener un promedio de rendimiento de cultivos en cielo abierto alto y valor de la producción alto.
3) Las estrellas en círculos verdes representan a las unidades de producción cuyos productores varían en su educación (nivel de estudios), es decir, hay una mezcla de ellos, sin embargo, estas unidades son muy caracterizadas por tener una considerable cantidad en tecnologías de bovinos, además de tener una considerable cantidad de hectáreas de otras superficies que en promedio es de 134 ha. No obstante, se observa que dichas unidades también cuentan con una cantidad considerable de cabezas de bovinos y que ellos seguramente son vacas lecheras pues el rendimiento de leche es alto, así como los ingresos por ésta.



Método 1 Análisis de Factores/ Descripción, Resultados con gráficas y discusión

El origen del Análisis factorial, también conocido como análisis de factores, se le atribuye a Charles Spearman (1904), en su clásico trabajo sobre inteligencia, donde distingue un factor general (factor G) y cierto número de factores específicos. Es una técnica estadística que sirve para realizar reducción de datos con la finalidad de explicar las correlaciones entre variables observadas en términos de un número menor de variables no observadas denominadas factores. Las variables observadas son modeladas como combinaciones lineales de factores considerando un error. Se puede decir que existen dos tipos de análisis factorial, es decir, el análisis factorial exploratorio (AFE) que de alguna manera se considera cuando se analiza un conjunto de datos sin tener hipótesis previa, y el análisis de factores confirmatorio (AFC) en el que se ha planteado al menos algunas hipótesis bien especificadas, las cuales se ponen a prueba evaluando el ajuste de un modelo. Más aún, se puede decir que el análisis de factores es una generalización de las componentes principales para reducir la dimensionalidad de los datos, ya éste último se centra en encontrar componentes o factores que expliquen la mayor parte de la varianza total del conjunto

de datos, mientras que, el análisis factorial pretende encontrar un conjunto de factores que expliquen la mayor parte de la varianza común (parte de la variación de la variable que es compartida con las otras variables), donde, se puede considerar la varianza total como la suma de la varianza común y la varianza única (parte de la variación de la variable que es propia de ella). Este análisis es usado en las ciencias sociales, marketing entre otras.

A manera de resumen, podemos decir que el análisis de factores es una técnica de reducción de dimensionalidad cuyo objetivo es encontrar grupos de variables homogéneas a partir de un conjunto numeroso de variables, de tal manera que dichos grupos sean lo más independiente posibles entre ellos, es decir, el objetivo es obtener un número mínimo de dimensiones con las cuales sea posible explicar la máxima información de los datos. Así pues, el análisis de componentes principales es una herramienta descriptiva, mientras que el análisis factorial presupone un modelo estadístico formal de generación de datos.

Con la intención de realizar un análisis de factores exploratorio, se consideraron 18 variables de las 25 seleccionadas y mencionadas en la Tabla 1 del apartado "**Descripción de los datos**", de tal manera que se desea observar si se puede reducir la dimensionalidad. Cabe mencionar que las 18 variables son consideradas debido a que son numéricas, mientras que el resto son variables categóricas como el "sexo del productor", el "nivel de estudios del productor", entre otras, lo cual implicaría tener más cuidado en el análisis al considerar variables dicotómicas, sin embargo, la idea en esta ocasión es considerar sólo variables numéricas y ver la posibilidad de reducir la dimensión con éstas. Las variables utilizadas se mencionan a continuación:

- ANIOS_EN_ACTIVIDAD, EDAD_PROD, MANO_D_OBRA, TRAC_MAQ_VEHI
- TECNOLOGIA_CA, TECNOLOGIA_BOV_PORC, SUP_AGRICOLA, SUP_RIEGO
- OTRAS_SUP, BOV, RENDIMIENTO_LECHE, INGRESO_LECHE, INGRESOS_VTA_BOV
- PORC, INGRESOS_VTA_PORC, CULTIVOS_CA, RENDIMIENTO_CA, VALOR_PROD_CA

Para asegurar que la matriz que conforma la información con las 18 variables es factorizable, se realizan 3 pruebas:

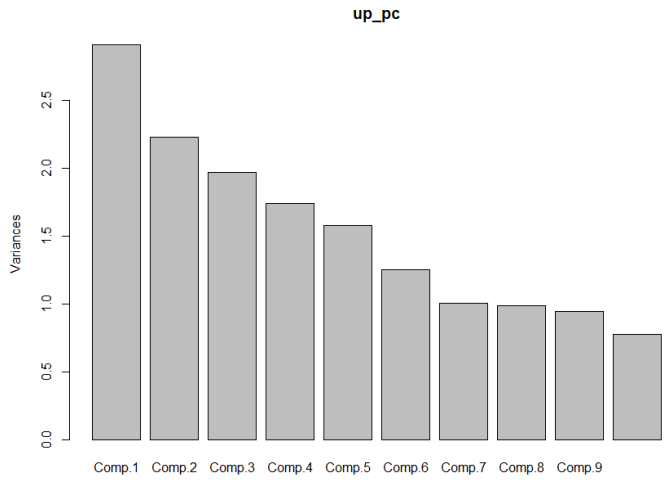
1) En primera instancia se calcula la matriz de correlación y a ésta se le obtiene su determinante, el cual es de 0.0027 aproximadamente, cuyo valor es muy próximo a cero, aunque siendo pequeño no es igual a cero, por lo que podríamos decir que la matriz de correlación no es la matriz identidad, y en consecuencia existe correlación entre las

variables. Más aún, el determinante es menor o igual a 1. Cuando éste es igual a 1 se dice que la matriz de correlaciones es la matriz identidad y en consecuencia no hay correlación entre las variables. Cuando el determinante es menor a 1, pero distinto de cero entonces se dice que hay correlación entre variables. Por último, cuando el determinante es cero entonces la matriz de correlaciones es una matriz singular.

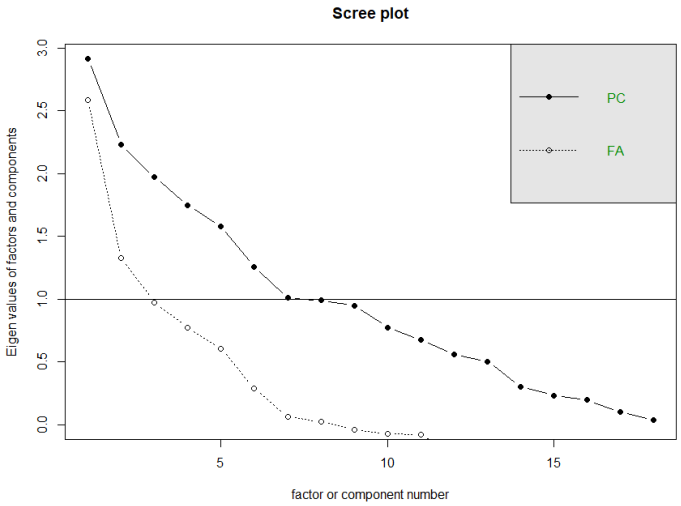
2) Se aplica la prueba de esfericidad de Bartlett, el cual consiste en una estimación de ji-cuadrado a partir de una transformación del determinante de la matriz de correlaciones, obteniendo un p-valor de 0, es decir, menor a 0.05, lo cual indica que podemos rechazar la hipótesis de que la matriz de correlación es la matriz identidad.
3) Se aplica la prueba de Kaiser-Meyer-Olkin (KMO)[1] obteniendo un valor de 0.6 indicando un valor mediocre según los valores definidos por Kaiser (0.00 a 0.49 inaceptable; 0.50 a 0.59 miserable; 0.60 a 0.69 mediocre; 0.70 a 0.79 medio; 0.80 a 0.89 meritorio; 0.90 a 1.00 maravilloso).

Derivado de los tres puntos anteriores, se decide continuar con el análisis factorial utilizando el método de componentes principales para estimar las cargas.

En el siguiente gráfico se observa la varianza de cada una de las componentes.



[1] http://www.uco.es/zootecniaygestion/img/pictorex/16_11_30_11_Factorial.pdf



Observando los dos gráficos anteriores, se sugiere que se utilicen 5 factores. Así pues, se elige un modelo con 5 factores. Luego, calculando la matriz de cargas y extrayendo la información de las 5 componentes e interpretando los factores en términos de sus cargas se observa lo siguiente:

Factor 1: Está compuesto por las variables MANO_D_OBRA, TRAC_MAQ_VEHI, SUP_AGRICOLA, SUP_RIEGO, VALOR_PROD_CA, lo cual se puede considerar como un factor agrícola.

Factor 2: Está compuesto por las variables TECNOLOGIA_BOV_PORC, BOV, RENDIMIENTO_LECHE, INGRESO_LECHE, INGRESOS_VTA_BOV, así pues, se puede considerar como un factor de bovinos.

Factor 3: Está compuesto por las variables PORC, INGRESOS_VTA_PORC, así pues, se puede considerar como un factor de porcinos.

Factor 4: Está compuesto por las variables ANIOS_EN_ACTIVIDAD, EDAD_PROD, lo cual podría decirse que es un factor de experiencia del productor.

Factor 5: Está compuesto por las variables TECNOLOGIA_CA, OTRAS_SUP, CULTIVOS_CA, RENDIMIENTO_CA, lo cual se diría que es un factor de cultivos en cielo abierto.

Para identificar cómo se forman los factores anteriores, es decir, que variables conforman a cada factor, se hace uso de la matriz de cargas, la cual aparece en la tabla 3, donde se observa en color naranja aquellas cargas mayores en valor absoluto para cada una de las variables, por ejemplo, la carga con mayor valor absoluto para la variable “ANIOS_EN_ACTIVIDAD” es de 0.56124311 y

conforma al factor 4. Otro ejemplo es la variable “MANO_D_OBRA” cuya carga mayor en valor absoluto es de 0.11493188 y en consecuencia ésta es parte del factor 1.

Tabla 3. Matriz de cargas obtenida con el método de componentes principales.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
ANIOS_EN_ACTIVIDAD	0.01353438	0.10070094	0.02035183	**0.56124311**	0.41334098
EDAD_PROD	0.01943508	0.08886393	0.03124476	**0.56096835**	0.39126402
MANO_D_OBRA	**-0.11493188**	-0.00790288	0.00522664	-0.08581842	0.09987049
TRAC_MAQ_VEHI	**-0.26460127**	0.12169795	0.00762355	-0.09604797	0.20342302
TECNOLOGIA_CA	-0.13727158	-0.10346262	-0.03067379	-0.32247914	**0.49181256**
TECNOLOGIA_BOV_PORC	-0.07179321	**0.25794386**	0.02093338	0.20167357	-0.17306941
SUP_AGRICOLA	**-0.49447584**	-0.22791335	0.02332021	0.13530458	-0.113838
SUP_RIEGO	**-0.4993207**	-0.25219352	0.0233372	0.12482401	-0.11377647
OTRAS_SUP	-0.02632269	0.12115923	0.00186363	0.12406055	**-0.21305473**
BOV	-0.24364815	**0.52230798**	-0.05413997	-0.09957961	-0.01916461
RENDIMIENTO_LECHE	-0.1216523	**0.25410232**	-0.02225918	0.0178673	0.03230117
INGRESO_LECHE	-0.22262243	**0.45119713**	-0.0559552	-0.13925186	0.04645342
INGRESOS_VTA_BOV	-0.15696072	**0.38620683**	-0.04217823	-0.10124814	-0.00439751
PORC	-0.00544274	0.05116919	**0.70203676**	-0.05659023	0.0182428
INGRESOS_VTA_PORC	-0.00450684	0.04963878	**0.70214307**	-0.05630613	0.01814625
CULTIVOS_CA	-0.05887007	-0.11459919	-0.03342411	-0.22953837	**0.4617342**
RENDIMIENTO_CA	-0.05608998	-0.06963901	-0.02041145	-0.22757144	**0.24102221**
VALOR_PROD_CA	**-0.49262049**	-0.21552602	0.01795046	0.09517982	-0.07770058

Además, la siguiente imagen muestra la desviación estándar explicada de cada una de las componentes principales.

```
Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9   Comp.10
1.7063303 1.4927016 1.4034455 1.3202998 1.2566203 1.1208271 1.0043387 0.9938236 0.9728692 0.8800045
  Comp.11   Comp.12   Comp.13   Comp.14   Comp.15   Comp.16   Comp.17   Comp.18
0.8217048 0.7459567 0.7090548 0.5507547 0.4784340 0.4416938 0.3168693 0.1807698
```

Ahora, se realiza un análisis de factores utilizando máxima verosimilitud para estimar los parámetros del modelo (las cargas y las varianzas específicas). Por default el análisis de factores se realiza sobre los datos estandarizados y utilizando la rotación varimax. En este caso, se utiliza la

función “fa” de la librería “psych” del programa estadístico R, cuyos parámetros son: la matriz de correlación, la cantidad de factores, el método ml (que indica que se hará un análisis factorial de máxima verosimilitud) de regresión con una rotación varimax y el número de observaciones.

Se eligen 5 factores para comparar los resultados obtenidos con el método anterior, de componentes principales, obteniendo que éstos se componen de la siguiente forma.

Factor 1: SUP_AGRICOLA, SUP_RIEGO, VALOR_PROD_CA

Factor 2: BOV, RENDIMIENTO_LECHE, INGRESO_LECHE, INGRESOS_VTA_BOV, TECNOLOGIA_BOV_PORC, OTRAS_SUP

Factor 3: PORC, INGRESOS_VTA_PORC

Factor 4: ANIOS_EN_ACTIVIDAD, EDAD_PROD

Factor 5: TRAC_MAQ_VEHI, TECNOLOGIA_CA, CULTIVOS_CA, RENDIMIENTO_CA

En la tabla 4 se observan sólo las cargas con mayor valor absoluto para cada una de las variables y en cada factor, las cuales son arrojadas al utilizar la función “fa” de la librería “psych” en R. Por ejemplo, el factor 4 está conformado por las variables “ANIOS_EN_ACTIVIDAD” y “EDAD_PROD”, pues éstas contienen su mayor carga en dicho factor.

Tabla 4. Matriz de cargas con mayor valor absoluto, obtenidas a partir de máxima verosimilitud.

	Factor1	Factor2	Factor3	Factor4	Factor5
ANIOS_EN_ACTIVIDAD				0.996	
EDAD_PROD				0.684	
MANO_D_OBRA					
TRAC_MAQ_VEHI	0.178	0.215			0.306
TECNOLOGIA_CA					0.951
TECNOLOGIA_BOV_PORC		0.189			-0.121
SUP_AGRICOLA	0.898				
SUP_RIEGO	0.978				
OTRAS_SUP		0.14			-0.116
BOV		0.996			
RENDIMIENTO_LECHE		0.194			
INGRESO_LECHE		0.731			
INGRESOS_VTA_BOV		0.557			
PORC			0.997		
INGRESOS_VTA_PORC			0.97		
CULTIVOS_CA					0.395
RENDIMIENTO_CA					0.264
VALOR_PROD_CA	0.851				

Sin embargo, es importante observar que la variable mano de obra “MANO_D_OBRA” no es representativa en ningún factor, posiblemente quedando en un sexto factor, según lo muestra la tabla anterior.

Más aún, en la tabla 5 se muestra la desviación estándar de cada factor, así como la proporción de varianza explicada por cada uno de los factores y la proporción de varianza explicada acumulada, observando en la última columna (Factor5) y último renglón (Cumulative Var) que la varianza explicada por los 5 factores es del 51.2 por ciento.

Tabla 5. Desviación estándar, proporción de varianza explicada y acumulada de cada factor.

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	2.535	1.988	1.937	1.479	1.277
Proportion Var	0.141	0.11	0.108	0.082	0.071
Cumulative Var	0.141	0.251	0.359	0.441	0.512

Además, observando la tabla anterior, se tiene que el porcentaje de varianza explicada por el modelo es de un 51.2 por ciento.

En conclusión, se observa que los resultados obtenidos por el método de componentes principales y el método de máxima verosimilitud son muy parecidos, salvo algunas pequeñas diferencias, lo cual da seguridad en los resultados observados, de tal manera que se puede decir los siguiente:

1) El método de componentes principales indica que el factor 1 está compuesto por las variables MANO_D_OBRA, TRAC_MAQ_VEHI, SUP_AGRICOLA, SUP_RIEGO, VALOR_PROD_CA, lo cual tiene mucho sentido pues si el productor tiene una superficie agrícola grande y si la superficie de riego también lo es, entonces eso indica que para trabajar dicha superficie requiere de maquinaria o mano de obra, obteniendo un valor de la producción alto. Ahora, observando el resultado obtenido con el método de máxima verosimilitud éste se arroja que el factor 1 está compuesto por las variables SUP_AGRICOLA, SUP_RIEGO, VALOR_PROD_CA, sin embargo, no se indica la mano de obra ni los tractores, maquinaria y vehículos. De esta manera, dicho factor se puede considerar como un factor agrícola.
2) Para el factor 2 el método de componentes principales indica que está formado por las variables BOV, RENDIMIENTO_LECHE, INGRESO_LECHE, INGRESOS_VTA_BOV, TECNOLOGIA_BOV_PORC, mientras que el método de máxima verosimilitud indica que se forma con las variables BOV, RENDIMIENTO_LECHE, INGRESO_LECHE, INGRESOS_VTA_BOV, TECNOLOGIA_BOV_PORC, OTRAS_SUP. En este caso, los dos métodos arrojan los mismos resultados salvo la variable de otras superficies que arroja el método de máxima verosimilitud, que de alguna manera tiene cierto sentido, es decir, si este factor se considera como un factor de bovinos, éste tiene sentido porque al tener existencias de reses entonces puede tener un rendimiento de leche e incluso ingresos por la leche o ingresos por la venta de bovinos, pero para eso, también ocupa tener tecnología

para los bovinos y en consecuencia el productor necesita tener otras superficies como establos o un agostadero donde pastar los animales.

3) Para el factor 3, los dos métodos utilizados arrojan que este representa a las variables PORC, INGRESOS_VTA_PORC, por lo cual se puede decir éste es un factor de porcinos, ya que al tener una cantidad de porcinos considerable tendrán un ingreso por la venta de puercos.
4) Los dos métodos utilizados arrojan que el factor 4 representa a las variables ANIOS_EN_ACTIVIDAD, EDAD_PROD, por lo cual, se puede considerar como un factor de experiencia del productor, y suena lógico ya que en el apartado anterior de Estadística descriptiva se observó que la mayoría de los productores inician desde muy jóvenes a realizar la actividad agropecuaria, además, de que dichos productores no están cambiando de actividad, por lo cual resulta ser que las variables tienen una proporción directa, es decir a mayor cantidad de años dedicados a la actividad agropecuaria, mayor años de edad tiene el productor.
5) Por último, para el factor 5 el método de componentes principales indica que representa a las variables TECNOLOGIA_CA, OTRAS_SUP, CULTIVOS_CA, RENDIMIENTO_CA, mientras que el método de máxima verosimilitud indica que representa a las variables TRAC_MAQ_VEHI, TECNOLOGIA_CA, CULTIVOS_CA, RENDIMIENTO_CA, por tal caso podemos decir que es un factor de cultivos en cielo abierto, lo cual tiene sentido porque se tiene cultivos en cielo abierto en una cantidad considerable entonces se espera que tenga un cierto rendimiento (superficie cosechada entre producción en toneladas) y para esto tuvo que tener tecnología para realizar las distintas labores de los cultivos.

Método 2 Escalamiento multidimensional/ Descripción, Resultados con gráficas y discusión

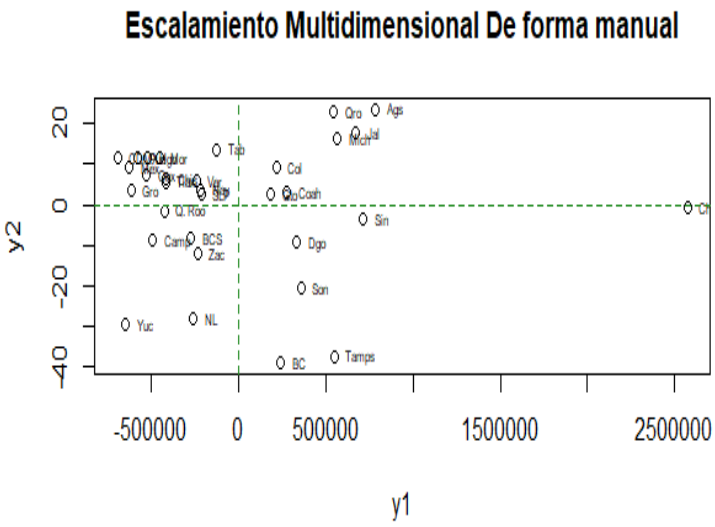
La técnica de escalamiento multidimensional, también conocido como MultiDimensional Scaling (MDS), surge a principios del siglo XX en la Psicología, cuando se pretende estudiar la relación que existía entre la intensidad física de ciertos estímulos con su intensidad subjetiva. Se puede decir que esta técnica multivariante tiene como enfoque exploratorio ayudar a que a partir de una

matriz de disimilaridad se encuentre un conjunto de variables que identifica la dimensión de las disimilaridades, con otras palabras, es una técnica de representación espacial que trata de visualizar sobre un mapa un conjunto de estímulos cuya posición relativa se desea analizar. Además, el propósito del MDS es transformar los juicios de similitud o preferencia llevados a cabo por una serie de individuos sobre un conjunto de objetos o estímulos en distancias susceptibles de ser representadas en un espacio multidimensional. Existen dos tipos de modelos básicos de escalamiento multidimensional que son: el modelo de escalamiento métrico y el modelo de escalamiento no métrico. El primero de ellos parte de la idea de que las distancias son una función de disimilaridades, de hecho, se parte del supuesto que la relación entre las disimilaridades y las distancias es de tipo lineal, mientras que el MDS no métrico no presupone una relación lineal entre las distancias y las disimilaridades, sino que establece una relación monótona creciente entre ambas.

En resumen, el MDS es una técnica que se utiliza para reducir las dimensiones de un conjunto de datos ya que éstos últimamente resultan ser demasiado grandes, así pues, esta técnica trata de representar en un espacio geométrico, por lo regular de dos dimensiones, las proximidades existentes entre un conjunto de datos.

Para aplicar la técnica de escalamiento multidimensional, se decide considerar las 8 variables numéricas, y no considerar variables categóricas, pues con la información de las variables se calcularán promedios: ANIOS_EN_ACTIVIDAD, EDAD_PROD, MANO_D_OBRA, TRAC_MAQ_VEHI, SUP_AGRICOLA, SUP_RIEGO, CULTIVOS_CA, VALOR_PROD_CA. Así pues, calculando promedios de cada variable por entidad, y formando la tabla a trabajar, para después de ésta calcular una matriz de distancias utilizando el método euclidiano.

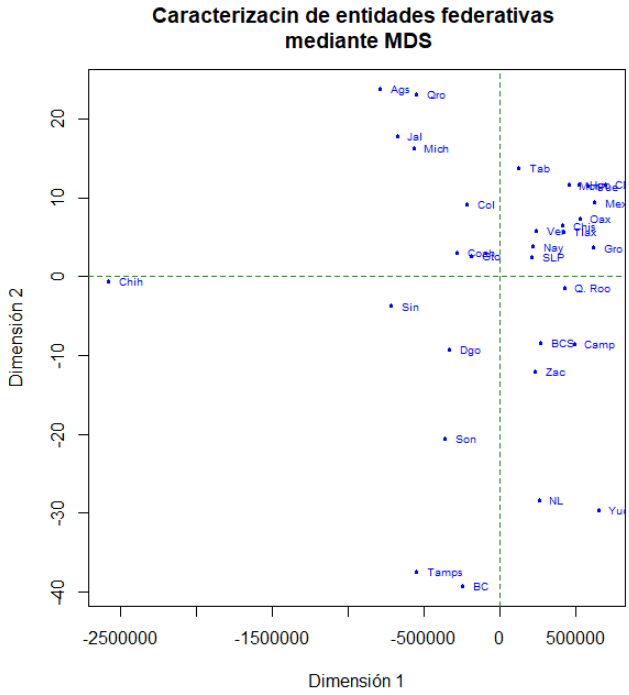
Ahora, haciendo uso de la matriz de distancias calculada anteriormente se obtiene el MDS de manera manual, que al graficarlo en 2 dimensiones queda de la siguiente forma:



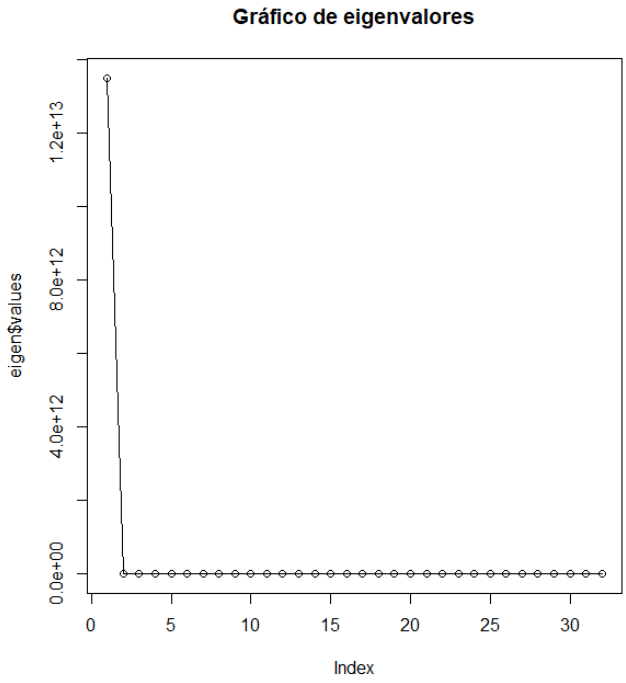
Sin embargo, al calcular la proporción de la varianza explicada por las dos dimensiones se obtiene un valor muy cercano al 100%, ya que con tan sólo la primera dimensión se explica casi el 100% mientras que la segunda dimensión tiene una proporción explicada del 6.094077e-10, lo cual es evidente según lo muestra la siguiente imagen de los eigenvalores.

```
> eigen$values
 [1]  1.349521e+13  8.224083e+03  1.109950e+03  5.045119e+02  1.426673e+02
 [6]  5.097720e+01  1.524446e+01  1.819247e+00  2.520716e-03  1.549296e-03
[11]  1.172953e-03  3.568048e-04  2.184961e-04  2.170974e-04  1.785635e-04
[16]  1.437631e-04  1.015686e-04  9.929173e-05  9.013054e-05  1.368123e-05
[21]  1.085956e-05 -3.235806e-05 -3.588677e-05 -5.601086e-05 -6.467651e-05
[26] -7.153034e-05 -8.974984e-05 -1.761133e-04 -3.238821e-04 -8.910727e-04
[31] -9.440082e-04 -1.266773e-03
```

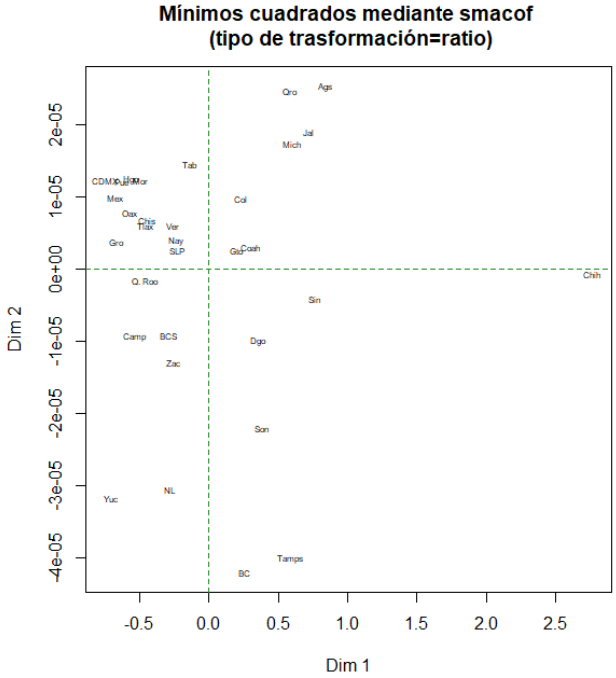
Ahora, realizando el análisis con la función cmdsscale del programa R y graficando la configuración solución obtenida mediante MDS clásico.

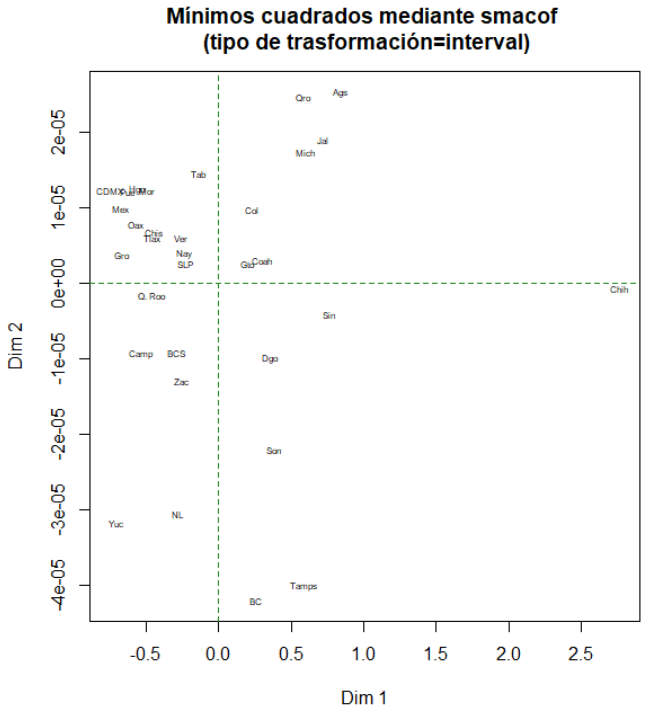


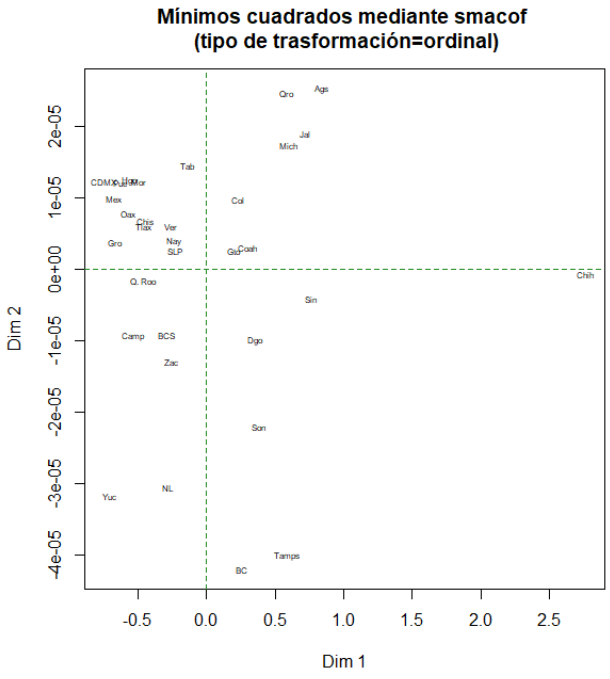
También, se observa que la proporción de la varianza explicada por la primera dimensión es casi el 100% y que la proporción de la varianza explicada por la segunda dimensión es mínima, por lo cual, ya no tiene caso hacer análisis con 3 dimensiones. Esto se respalda al graficar los eigenvalores, los cuales se muestran a continuación.



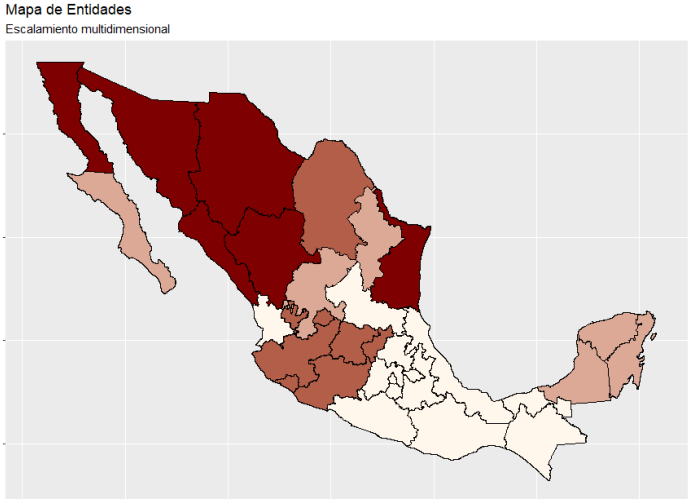
Ahora se utiliza el enfoque de mínimos cuadrados mediante la función smacof y utilizando las transformaciones ratio, Interval y ordinal.







En conclusión, si se grafican los estados obtenidos en las imágenes anteriores y considerando los cuatro cuadrantes, pero en un mapa de entidades, entonces se obtiene la siguiente imagen.



Como las variables estudiadas son: ANIOS_EN_ACTIVIDAD, EDAD_PROD, MANO_D_OBRA, TRAC_MAQ_VEHI, SUP_AGRICOLA, SUP_RIEGO, CULTIVOS_CA y VALOR_PROD_CA. Entonces se podría decir que las unidades de producción agropecuarias que se encuentran en las entidades federativas de Baja California, Sonora, Sinaloa, Chihuahua, Durango y Tamaulipas se caracterizan por la superficie agrícola, la superficie de riego y por los tractores, maquinaria y vehículos utilizados para las labores agropecuarias. Mientras que algunos estados del sur se caracterizan por la mano de obra, la edad del productor y los años que éste lleva realizando la actividad agropecuaria.

Método 3 Análisis de correspondencia/ Descripción, Resultados con gráficas y discusión

El análisis de correspondencia se utiliza haciendo uso de datos categóricos que se presentan en una tabla de contingencia con la finalidad de identificar las dimensiones subyacentes de los datos, como enfoque exploratorio. Dicho de otra manera, es una técnica descriptiva que resume gráficamente la información contenida en una tabla dinámica. Es decir, es una técnica descriptiva o exploratoria cuyo objetivo es resumir una gran cantidad de datos en un número reducido de dimensiones, con la menor pérdida de información posible. Se considera como análisis de correspondencia simple aquel cuya representación de datos se puede presentar en forma de

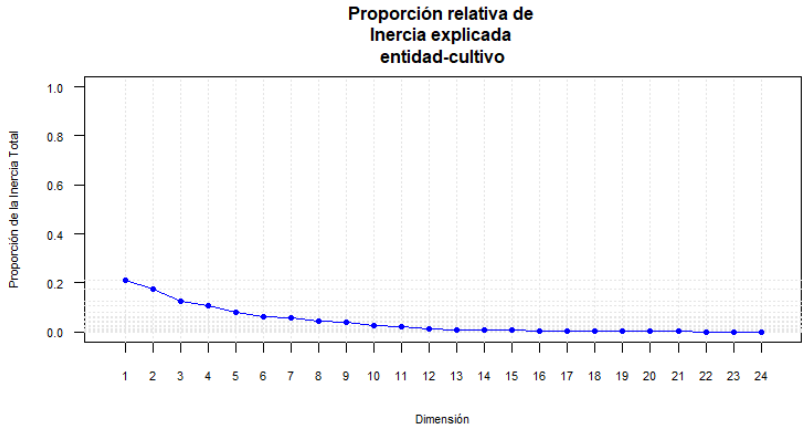
tablas de contingencia de dos variables categóricas para resumir la información que se presente en filas y columnas de manera que pueda proyectarse sobre un subespacio reducido, y representarse simultáneamente los puntos fila y los puntos columna, pudiéndose obtener conclusiones sobre relaciones entre las dos variables nominales u ordinales de origen. La extensión del análisis de correspondencias simples al caso de varias variables categóricas (tablas de contingencia multidimensionales) se denomina Análisis de Correspondencia Múltiple.

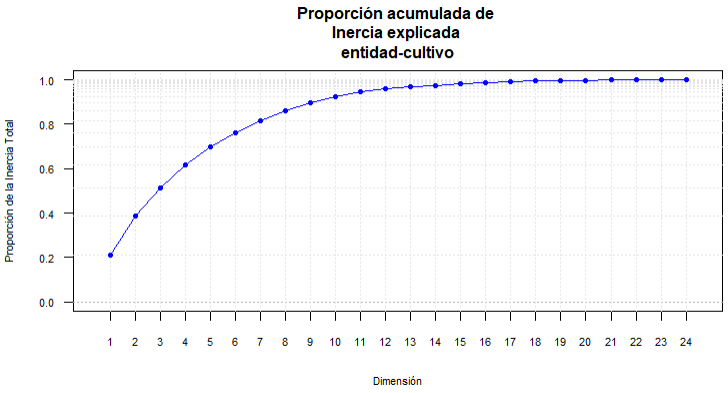
Con la intención de aplicar el método del análisis de correspondencia simple a la información del sector agropecuario, se utiliza la tabla de frecuencias de unidades que producen alguno de los 25 cultivos en cualquiera de las entidades federativas.

Al realizar el ejercicio y calcular la inercia explicada relativa por una dimensión se obtiene que ésta es del 21.29 por ciento, mientras que si se consideran dos dimensiones entonces la inercia explicada relativa es del 38.64 por ciento, sin embargo, si se quisiera una inercia explicada por lo menos del 89.87 por ciento entonces se requieren 9 dimensiones. Lo mencionado en este párrafo se muestra en la siguiente imagen, recordando que la inercia es la que cuantifica la cantidad de información retenida en cada dimensión:

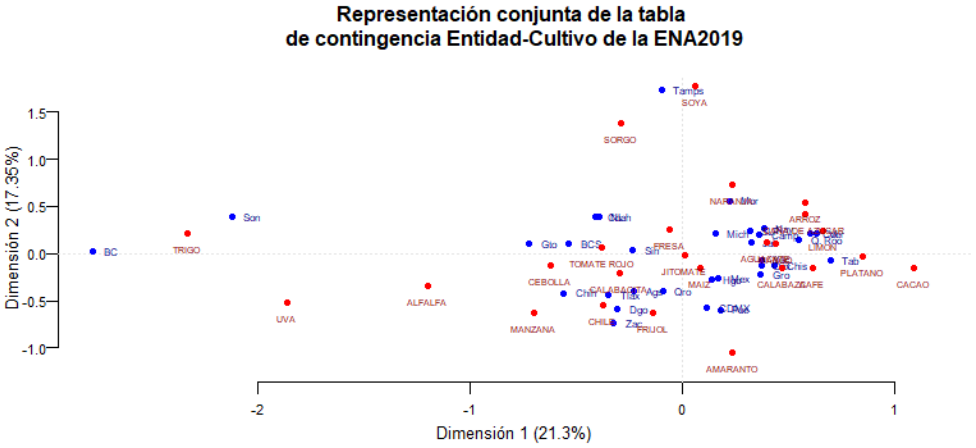
```
> iner_expl_acum
 [1] 0.2129602 0.3864350 0.5108397 0.6161971 0.6970659 0.7592501 0.8164701 0.8600575
 [9] 0.8987420 0.9245332 0.9449295 0.9584118 0.9677584 0.9751768 0.9816547 0.9867908
[17] 0.9916108 0.9939962 0.9961298 0.9975989 0.9987404 0.9994321 0.9998558 1.0000000
```

Más aún, los siguientes dos gráficos muestran la proporción relativa de la inercia explicada, así como la proporción acumulada de la inercia explicada.





Ahora, graficando la representación conjunta de los renglones y las columnas en el espacio de dos dimensiones, sin perder de vista que con ellas sólo se explica el 38.64 por ciento de la inercia total.



Ahora, si se consideran 9 dimensiones las cuales explican el 89.87 por ciento de la inercia total, y haciendo uso de la distancia euclidiana para calcular la menor distancia que existe de un cultivo a las entidades federativas entonces podemos ver qué entidad es la que produce cierto cultivo. A manera de ejemplo, se utiliza el cultivo de Aguacate, el cual es representado por un punto con 9 dimensiones, y se calcula la distancia que existe entre éste y cada uno de los puntos que representan a cada entidad, se obtienen las distancias que a continuación se muestran, representando la distancia que existe del Aguacate a las entidades "Ags", "BC", "BCS", "Camp", "Coah", "Col", "Chis", "Chih", "CDMX", "Dgo", "Gto", "Gro", "Hgo", "Jal", "Mex", "Mich", "Mor", "Nay", "NL", "Oax", "Pue", "Qro", "Q. Roo", "SLP", "Sin", "Son", "Tab", "Tamps", "Tlax", "Ver", "Yuc", "Zac".

```
> distancia_ent
 [1] 2.048444 3.772549 2.078068 1.763160 2.188448 1.507655 2.100240 2.469629
 [9] 2.000783 2.413752 2.089133 1.936496 2.100293 1.669708 1.787139 0.593095
[17] 1.634000 1.499305 2.389655 1.927015 2.205219 2.023819 1.925710 2.311051
[25] 1.872598 3.159080 3.008741 2.836166 2.000662 2.157378 2.211470 2.450927
```

De lo anterior, se observa que la distancia mínima es de 0.593095 y corresponde a la distancia entre el Aguacate y el estado de Michoacán. Por tanto, podemos decir que el Aguacate es un cultivo representativo de Michoacán.

Considerando el proceso anterior, calculando la distancia euclidiana mínima que existe entre cada uno de los cultivos y las 32 entidades se obtiene que:

- El aguacate, el arroz y la fresa son representativos de Michoacán
- La alfalfa es representativa de Guanajuato y Chihuahua
- El amaranto es representativo de Puebla
- El cacao es representativo de Tabasco
- El café es representativo de Chiapas y Oaxaca
- La calabacita y el tomate rojo son representativos de Baja california sur
- La calabaza es representativa de Yucatán
- La caña de azúcar es representativa de Colima
- La cebolla es representativa de Guanajuato
- El chile es representativo en Chihuahua y Zacatecas
- El frijol es representativo de Durango y Zacatecas
- El jitomate es representativo de Sinaloa y Nayarit
- El limón es representativo de Quintana Roo
- El maíz es representativo de Querétaro e Hidalgo
- El mango es representativo de Nayarit
- La manzana es representativa de Chihuahua
- La naranja es representativa de Nuevo León
- El plátano es representativo de Chiapas y Veracruz
- El sorgo y la soya son representativos de Tamaulipas
- El trigo y la uva son representativos de Sonora

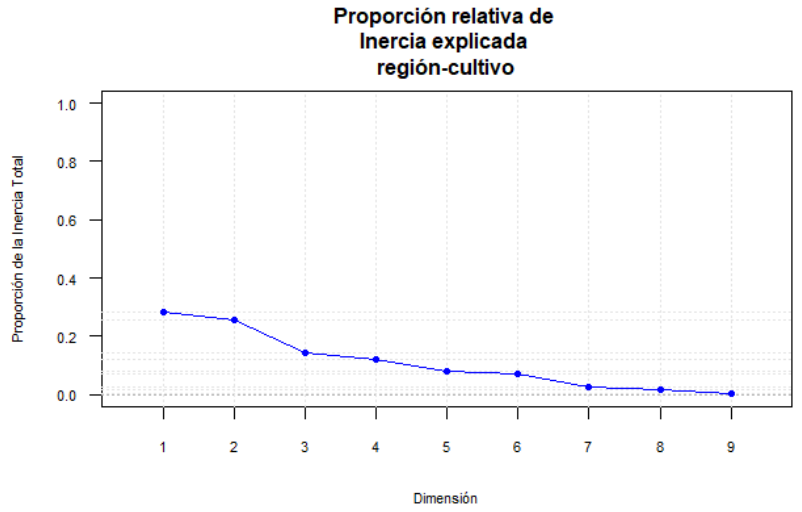
Con la intención de tener un ejercicio de análisis de correspondencia en el cual se explique un poco más de la inercia total en dos dimensiones, se considera la tabla de frecuencias de unidades que producen alguno de los 25 cultivos en 10 regiones, las cuales se clasifican según se muestran en la tabla 6:

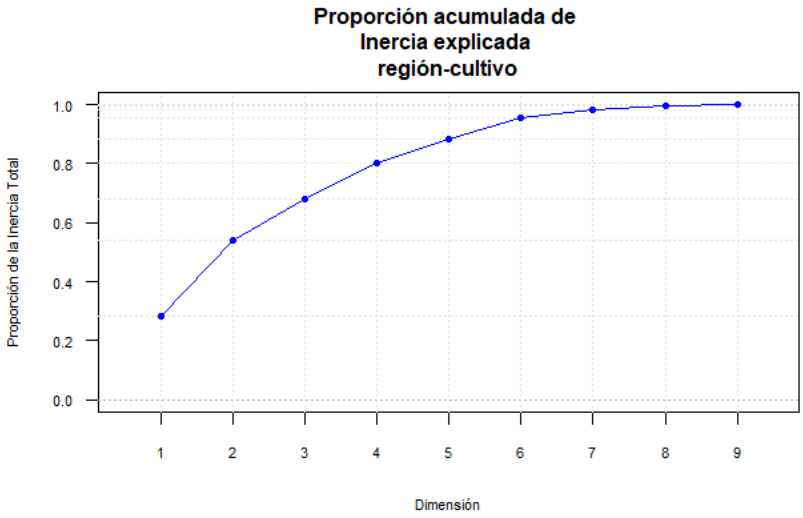
Tabla 6. Regiones conformadas por Entidades Federativas

REGIÓN	ENTIDADES (siglas)			
NOROESTE	BC	BCS	Sin	Son
NORESTE	Coah	NL	Tamps	
NORTE	Chih	Dgo	Zac	
CENTRO NORTE	Ags	Gto	Qro	SLP
OCCIDENTE	Col	Jal	Mich	Nay
CENTRO SUR	Gro	Mex	Mor	
ORIENTE	Hgo	Pue	Tlax	Ver
SUR	Chis	Oax	Tab	
SURESTE	Camp	Q. Roo	Yuc	
CENTRO	CDMX			

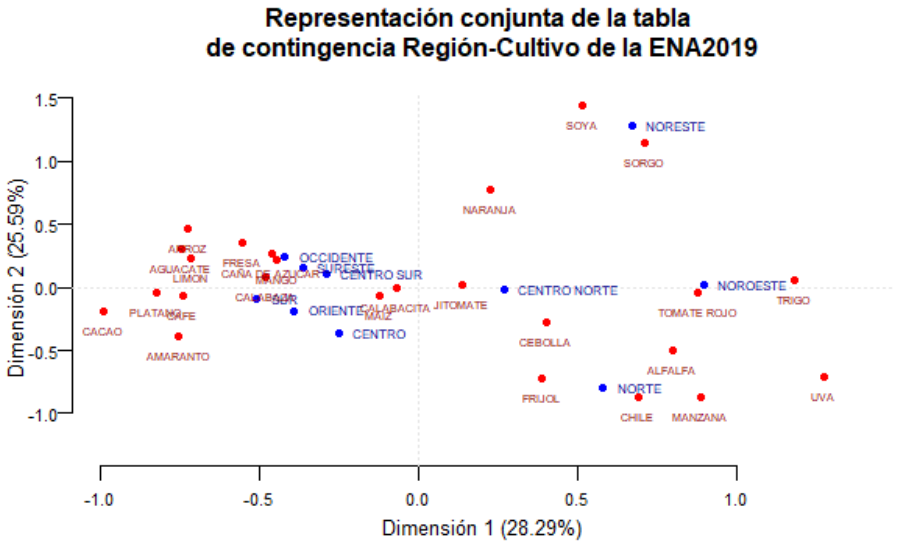
Cabe mencionar que las regiones es una clasificación (estratificación) que es definida y utilizada por el INEGI.

La inercia total es calculada con la ayuda del programa R, obteniendo un valor del 0.9140822; Así pues, en los dos siguientes gráficos se observa que ahora la proporción explicada de la inercia total con una dimensión es del 28.29 por ciento, mientras que si se consideran dos dimensiones (coordenadas) es del 53.88 por ciento, es decir, la inercia total se explica mejor para dos dimensiones cuando se utilizan regiones que cuando se usan entidades federativas.





Por tanto, graficando la representación conjunta de los renglones y de las columnas en el espacio de dos dimensiones, se obtiene el siguiente gráfico.



Si se considera la representación de cultivos por región de acuerdo a la proyección conjunta de renglones y columnas, entonces del gráfico “Representación conjunta de la tabla de contingencia región-cultivos de la ENA 2019”, se concluye lo siguiente:

Los cultivos representativos por región son:

- Noreste: Sorgo, Soya y Naranja
- Noroeste: Trigo y Tomate rojo
- Norte: Frijol, Chile, Manzana, Uva y Alfalfa
- Sur: Plátano, Café y Cacao
- Occidente: Arroz, Fresa, Aguacate, Caña de Azúcar, Limón, Mango y Calabaza
- Sureste: Arroz, Fresa, Aguacate, Caña de Azúcar, Limón, Mango y Calabaza
- Centro Sur: Arroz, Fresa, Aguacate, Caña de Azúcar, Limón, Mango y Calabaza
- Oriente: Amaranto, Maíz y Calabacita
- Centro: Maíz y Calabacita
- Centro Norte: Cebolla y Jitomate

Además, es fácil ver que las regiones Occidente, Sureste y Centro Sur son muy parecidas y en consecuencia los cultivos representativos son los mismos.

Las regiones donde se producen cierto cultivo son:

- Soya: Noreste (Coahuila, Nuevo León y Tamaulipas)
- Sorgo: Noreste (Coahuila, Nuevo León y Tamaulipas)
- Naranja: Noreste (Coahuila, Nuevo León y Tamaulipas)
- Jitomate: Centro Norte (Aguascalientes, Guanajuato, Querétaro y San Luis Potosí)
- Trigo: Noroeste (Baja California, Baja california Sur, Sinaloa, Sonora)
- Calabaza: Occidente (Colima, Jalisco, Michoacán, Nayarit), Sureste (Campeche, Quintana Roo, Yucatán) y Centro Sur (Guerrero, México, Morelos) es un cultivo que si se puede dar en todos estos estados.
- Mango: Occidente (Colima, Jalisco, Michoacán, Nayarit), Sureste (Campeche, Quintana Roo, Yucatán) y Centro Sur (Guerrero, México, Morelos) Son entidades tropicales.
- Caña de azúcar: Occidente (Colima, Jalisco, Michoacán, Nayarit), Sureste (Campeche, Quintana Roo, Yucatán) y Centro Sur (Guerrero, México, Morelos)
- Fresa: Occidente (Colima, Jalisco, Michoacán, Nayarit), Sureste (Campeche, Quintana Roo, Yucatán) y Centro Sur (Guerrero, México, Morelos)
- Aguacate: Occidente (Colima, Jalisco, Michoacán, Nayarit), Sureste (Campeche, Quintana Roo, Yucatán) y Centro Sur (Guerrero, México, Morelos)
- Limón: Occidente (Colima, Jalisco, Michoacán, Nayarit), Sureste (Campeche, Quintana Roo, Yucatán) y Centro Sur (Guerrero, México, Morelos)
- Arroz: Occidente (Colima, Jalisco, Michoacán, Nayarit), Sureste (Campeche, Quintana Roo, Yucatán) y Centro Sur (Guerrero, México, Morelos)
- Calabacita: Oriente (Hidalgo, Puebla, Tlaxcala, Veracruz) y Centro (Ciudad de México) es representativa en todas las entidades.
- Maíz: Oriente (Hidalgo, Puebla, Tlaxcala, Veracruz) y Centro (Ciudad de México) sin embargo,
  el maíz es un cultivo que es representativo a nivel nacional (en todas las regiones)
- Plátano: Sur (Chiapas, Oaxaca, Tabasco)
- Café: Sur (Chiapas, Oaxaca, Tabasco)
- Cacao: Sur (Chiapas, Oaxaca, Tabasco)
- Amaranto: Oriente (Hidalgo, Puebla, Tlaxcala, Veracruz)
- Tomate rojo: Noroeste (Baja California, Baja california Sur, Sinaloa, Sonora) este cultivo es de exportación y tiene mucho sentido que se cultive en la región noroeste.
- Cebolla: Centro Norte (Aguascalientes, Guanajuato, Querétaro y San Luis Potosí)

- Alfalfa: Norte (Chihuahua, Durango, Zacatecas)
- Uva: Norte (Chihuahua, Durango, Zacatecas) y Noroeste (Baja California, Baja california Sur, Sinaloa, Sonora)
- Frijol: Norte (Chihuahua, Durango, Zacatecas)
- Chile: Norte (Chihuahua, Durango, Zacatecas)
- Manzana: Norte (Chihuahua, Durango, Zacatecas)

Por otro lado, si consideramos 5 dimensiones entonces la inercia total explicada es del 88.38 por ciento, además, utilizando la distancia euclidiana para calcular la distancia mínima de cada una de las regiones a los 25 cultivos, se puede decir que:

- La región noroeste (Baja California, Baja california Sur, Sinaloa, Sonora) produce los cultivos tomate rojo y trigo.
- La región noreste (Coahuila, Nuevo León y Tamaulipas) produce los cultivos de sorgo, naranja y soya.
- La región norte (Chihuahua, Durango, Zacatecas) produce los cultivos de chile y frijol.
- La región centro norte (Aguascalientes, Guanajuato, Querétaro y San Luis Potosí) produce los cultivos de cebolla y maíz
- La región occidente (Colima, Jalisco, Michoacán y Nayarit) produce los cultivos de el mango, aguacate y caña de azúcar.
- La región centro sur (Guerrero, México y Morelos) produce los cultivos caña de azúcar y maíz.
- La región oriente (Hidalgo, Puebla, Tlaxcala, Veracruz) produce los cultivos de café y el maíz.
- La región sur (Chiapas, Oaxaca y Tabasco) produce los cultivos de plátano y caña de azúcar.
- La región sureste (Campeche, Quintana Roo y Yucatán) produce los cultivos de limón y la calabaza.
- La región centro (Ciudad de México) produce los cultivos de calabacita y el maíz.

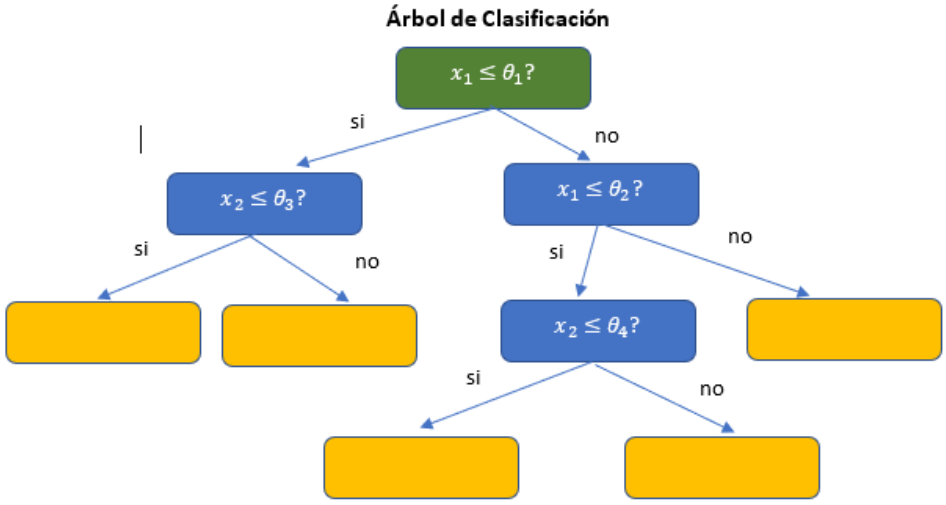
Método 4 Árboles de clasificación/ Descripción, Resultados con gráficas y discusión

Los árboles de clasificación y regresión (Breiman et al., 1984), también conocidos como CART (Classification and Regression Tree) es un método no paramétrico cuya intención es clasificar y/o discriminar la información de un conjunto de datos, en otras palabras, es el resultado de preguntar una secuencia ordenada de preguntas, donde el tipo de pregunta que se contesta en cada paso de la secuencia depende de las respuestas de las preguntas previas en la secuencia anterior. Además, la secuencia termina en la predicción de la clase. Una ventaja del método es que puede destacar su robustez a outliers (datos atípicos).

Se dice que son árboles de clasificación cuando la variable dependiente es de tipo categórica, pues si ésta resulta ser de tipo numérica entonces se dice que son árboles de regresión. Además, las variables independientes de un árbol de clasificación pueden ser categóricas o numéricas.

Un árbol de clasificación es representado por nodos y ramas, donde cada nodo simboliza una cuestión o decisión sobre una de las características (variables). Los nodos se clasifican como nodo raíz, nodo intermedio y nodo terminal. Sin embargo, sólo del nodo raíz que es el nodo inicial o del nodo intermedio pueden salir dos o más ramas, las cuales dependen de si la respuesta a la cuestión es binaria o no.

La siguiente figura muestra un ejemplo de árbol de clasificación, en el cual, el rectángulo verde indica el nodo raíz, los rectángulos azules muestran los nodos intermedios y los rectángulos anaranjados son los nodos terminales. Además, en este caso, en la figura se ilustran sólo dos variables (x1 y x2) y cuatro parámetros representados por la letra griega theta y un subíndice, los cuales se ajustarán con los datos para poder hacer una clasificación. A manera de ejemplo, sólo se ilustraron 2 variables y cuatro parámetros, sin embargo, pueden existir más variables y parámetros.



Para construir un árbol de decisión, el conjunto de datos se particiona en dos, un conjunto de entrenamiento que se sugiere sea al menos el 70 por ciento de los datos, y el conjunto de prueba con los datos restantes. Se trabaja con los datos de entrenamiento y se construye el árbol siguiente con los pasos que se repiten recursivamente:

1) Cada nodo parte en función de una prueba planteada sobre el valor de alguna de las características de los datos. En el caso binario puede ser verdadero o falso, o representado por valores cero y uno, así pues, los datos que cumplen cierta condición se asignan a uno de los nodos hijos (intermedios) y los restantes al otro. Además, es importante mencionar que cuando se parte un nodo, éste pasa a ser un nodo intermedio. Ahora bien, de entre todas las posibles particiones que se puedan hacer se tendrán que elegir aquellas que lleven a un mejor resultado obteniendo una mayor homogeneidad o pureza de los nodos hijos con respecto al nodo padre, para tal caso, se utilizan la entropía, el índice de Gini o el mínimo error (error de Bayes), los cual establece una medida de impureza para la construcción del árbol de clasificación. Sin embargo, la medida de entropía es la más utilizada.
2) La condición de parada detiene el proceso de partición de nodos, cuando un nodo cumple esta condición se nombra a éste como nodo terminal. Se detiene la división de los nodos cuando éstos sean puros, es decir, cuando todos los datos del nodo terminal son de la misma clase, pero también se detiene cuando su tamaño sea inferior a un determinado umbral, o superen un determinado nivel de pureza. Así pues, si el nodo padre se divide en dos nodos hijos entonces la pureza de éstos debe ser mayor que la del nodo padre, o lo

que es lo mismo, la impureza de los nodos hijos debe ser menor que la del nodo padre. Por último, es importante mencionar que también se puede hacer una poda, de esta manera también se detiene el proceso de partición de nodos. Recordando que una poda consiste en eliminar divisiones de nivel inferior que no contribuyen significativamente a la precisión del árbol y, ésta se realiza con la intención de simplificar un árbol para facilitar su interpretación.

Además, es importante mencionar que en cada nodo se busca la mejor división entre todas las variables que son objeto del análisis, sin embargo, una variable puede ser utilizada en varias divisiones, sólo una vez o nunca. Así pues, se buscan divisiones que disminuyan la pureza, lo cual implicaría aumentar la homogeneidad de los nodos que se obtienen, siendo un nodo perfectamente homogéneo si y sólo si las observaciones están contenidas en la misma clase.

Con la intención de aplicar el método de árboles de clasificación a la información de la Encuesta Nacional Agropecuaria 2019, se utiliza la función rpart en el programa R, así como la información de 13 variables que a continuación se describen y de las cuales, las primeras 4 son variables categóricas y el resto son variables numéricas: “nivel de estudios”, “edad del productor”, “sexo del productor”, “apoyo económico que recibe el productor para la producción”, “número de problemáticas reportadas en la unidad de producción”, “número de acciones realizadas para protección del medio ambiente”, “mano de obra”, “tractores, maquinaria y vehículos”, “tecnología a cielo abierto”, “superficie agrícola de la unidad de producción”, “cantidad de ganado bovino”, “cantidad de ganado porcino” y “cultivos en cielo abierto”. Cabe mencionar que la información de dichas variables es considerada a nivel nacional. Más aún, como el nivel de estudios es una variable categórica que puede tomar 7 valores distintos según se muestra en la tabla 7, y considerando que en el apartado de estadística descriptiva se observó que son relativamente pocos los productores que tienen un nivel de estudios de preparatoria, licenciatura, maestría o doctorado, entonces se decide agrupar éstos quedando una variable con cuatro categorías.

Tabla 7. Clasificación del nivel de estudios.

CLAVE	CLAVE2	DESCRIPCIÓN
1	1	SIN ESTUDIOS
2	2	PRIMARIA
3	3	SECUNDARIA
4	4	PREPARATORIA O BACHILLERATO
5	4	LICENCIATURA
6	4	MAESTRÍA
7	4	DOCTORADO

También, se estratificó la edad del productor considerando lo observado en el apartado estadística descriptiva, quedando de la siguiente manera.

Tabla 8. Clasificación de la edad del productor.

CLAVE	DESCRIPCIÓN
1	MENOR A 40
2	DE 40 Y MENOR A 60
3	DE 60 Y MENOR A 80
4	DE 80 Y MÁS

Así pues, considerando un 80 por ciento de los datos como entrenamiento y un 20 por ciento como prueba, se realiza el ejercicio con la ayuda de la función rpart y el conjunto de entrenamiento, donde el método utilizado es class de clasificación, y recordando que en dicha función el parámetro cp es el parámetro de complejidad para la poda del árbol de tal forma que si cp=1 entonces se genera un árbol sin divisiones, pero si cp=0 entonces se genera un árbol de profundidad máxima. Además, de tomar en cuenta que la variable “nivel de estudios” será la dependiente y que para cada ejercicio realizado se considera la semilla set.seed(12345), de tal manera que se pueda encontrar un modelo que prediga el nivel de estudios de un productor y que éste se pueda replicar.

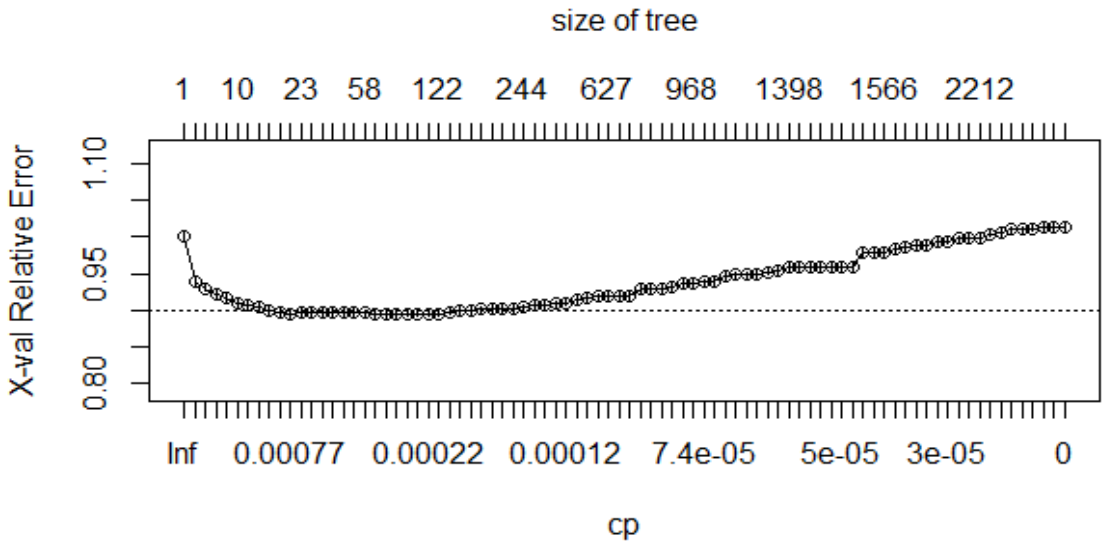
Al buscar un árbol de clasificación sin realizar podas, es decir, encontrar el árbol máximo, el programa R hace los cálculos encontrando este árbol con más de 2 212 nodos terminales, por lo

cual, no es posible calcular las más de 2 212 reglas ya que R muestra una limitante de 1 000 reglas. Sin embargo, si es posible graficar dicho árbol, pero debido a la gran cantidad de nodos que este contiene (nodo raíz, nodos intermedios y nodos terminales), el programa tarda varios minutos en generarlo y al terminar se observa que el gráfico no es nada apreciable, es decir, no se aprecian los resultados obtenidos y se da un amontonamiento de nodos.

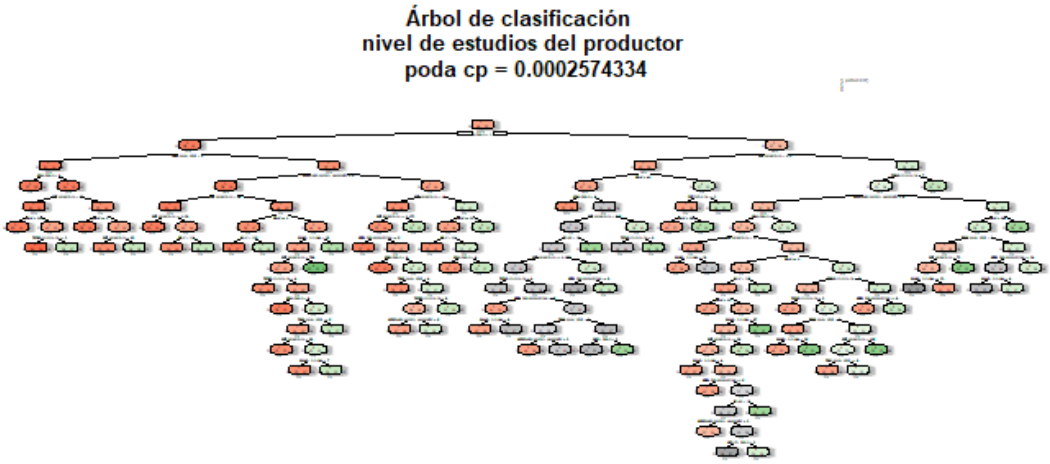
Sin embargo, al utilizar los datos de prueba para predecir con el árbol de clasificación máximo (sin podar), se obtiene la siguiente matriz de confusión obtenida con la instrucción confusionMatrix, en la cual se observa que el 47.57 por ciento de los datos de prueba coinciden con los datos predichos por el modelo. Este porcentaje es obtenido al realizar el cociente entre la suma de los datos de la diagonal de la matriz de confusión y el total de datos; que, en este caso, es el cociente entre 5 338 y 11 222.

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1  123  378   64   32
         2  712 3614 1151  959
         3   68  716  565  488
         4   59  715  542 1036

Overall Statistics

               Accuracy : 0.4757
                 95% CI : (0.4664, 0.485)
    No Information Rate : 0.4832
    P-Value [Acc > NIR] : 0.9469

                  Kappa : 0.1775

 Mcnemar's Test P-Value : <2e-16
```

Además, la siguiente imagen, en el eje y aparece el error relativo de validación cruzada, mientras que en la parte inferior del eje x se muestra el valor del parámetro de penalización, es decir, el valor de poda (cp), y en la parte superior del eje x aparece el número de nodos terminales. De esta manera, en la figura se observa que el árbol máximo, es decir, el árbol sin poda (cp=0) tiene más de 2 212 nodos terminales y en consecuencias más de esta cantidad como reglas, aunado a esto, el error relativo es el error máximo que tiene un árbol de clasificación.



Es obvio que el árbol máximo no es el óptimo, pues aquel que lo sea es el que refleje un buen equilibrio entre el sesgo y la varianza. Así pues, con la ayuda de R se obtiene que el árbol óptimo se da realizando una poda con valor cp= 0.0002574334, observando en la figura anterior que es donde el error relativo es mínimo, además de tener alrededor de 76 nodos terminales, por lo cual, al graficarlo este no es muy apreciable, como lo muestra la siguiente figura.



Sin embargo, al utilizar el árbol óptimo y haciendo predicciones con él, mediante la función predict se observa en la siguiente imagen la matriz de confusión, donde se aprecia que no se realizaron predicciones para productores sin estudios pero si para el resto, obteniendo un 54.14 por ciento de los casos en los cuales la predicción realizada con el modelo coincide con el dato real, es decir, si sumamos los valores de la diagonal de la matriz de confusión se obtiene un total de 6 076 casos

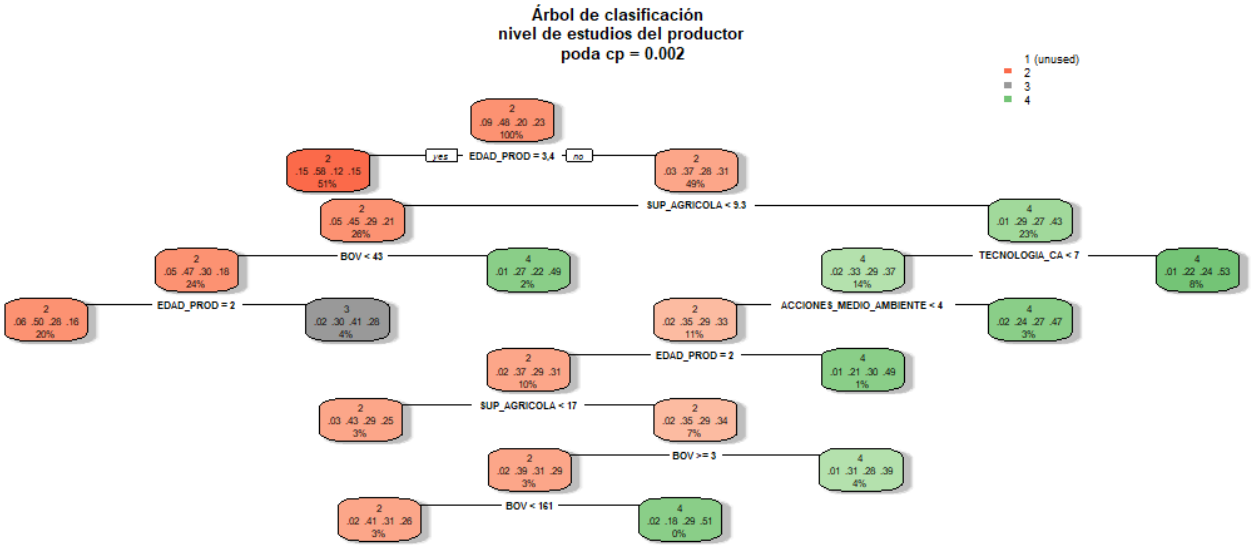
que se predicen y coinciden con la realidad (los datos de prueba), pero los datos fuera de la diagonal suman 5 146 datos que el modelo no predice bien ya que no coinciden con los datos de prueba, de tal manera que el 54.14 por ciento de un total de 11 222 casos se predicen coincidiendo con los datos de prueba.

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1    0    0    0    0
         2  919 4716 1539 1229
         3   11  123  176  102
         4   32  584  607 1184

Overall Statistics

               Accuracy : 0.5414
                 95% CI : (0.5322, 0.5507)
    No Information Rate : 0.4832
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2127

 Mcnemar's Test P-Value : < 2.2e-16
```

Ahora, con la intención de tener un tercer árbol y realizando una poda para este con valor de cp=0.002, de tal manera que se compare la información obtenida en este caso con los dos árboles anteriores (árbol máximo y árbol óptimo), se observa en la siguiente imagen que el 53.37 por ciento de los datos se predicen coincidiendo con los datos reales, que es muy parecido al porcentaje predicho con el árbol óptimo.

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1    0    0    0    0
         2  927 4775 1542 1376
         3   12  133  189  114
         4   23  515  591 1025

Overall Statistics

               Accuracy : 0.5337
                 95% CI : (0.5244, 0.5429)
    No Information Rate : 0.4832
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1926

 Mcnemar's Test P-Value : < 2.2e-16
```

Además, el árbol en este caso se muestra a continuación, el cual tiene una mejor apreciación que los árboles anteriores, es decir, se aprecian los resultados obtenidos en dicho árbol de clasificación.



Más aún, para dicho árbol se obtiene 11 nodos, lo cual implica el tener las siguientes 11 reglas.

NIVEL_D_ESTUDIOS 1 2 3 4

2 [.02 .41 .31 .26] when EDAD_PROD is 2 & SUP_AGRICOLA>= 17.1 & TECNOLOGIA_CA < 7 &

ACCIONES_MEDIO_AMBIENTE < 4 & BOV is 3 to 161

2 [.03 .43 .29 .25] when EDAD_PROD is 2 & SUP_AGRICOLA is 9.3 to 17.1 & TECNOLOGIA_CA < 7 &
ACCIONES_MEDIO_AMBIENTE < 4

2 [.06 .50 .28 .16] when EDAD_PROD is 2 & SUP_AGRICOLA < 9.3 & BOV < 43

2 [.15 .58 .12 .15] when EDAD_PROD is 3 or 4

3 [.02 .30 .41 .28] when EDAD_PROD is 1 & SUP_AGRICOLA < 9.3 & BOV < 43

4 [.01 .31 .28 .39] when EDAD_PROD is 2 & SUP_AGRICOLA>=17.1 & TECNOLOGIA_CA<7 &
ACCIONES_MEDIO_AMBIENTE < 4 & BOV < 3

4 [.02 .24 .27 .47] when EDAD_PROD is 1 or 2 & SUP_AGRICOLA >=9.3 & TECNOLOGIA_CA < 7 &
ACCIONES_MEDIO_AMBIENTE >= 4

4 [.01 .21 .30 .49] when EDAD_PROD is 1 & SUP_AGRICOLA >= 9.3 & TECNOLOGIA_CA < 7 &
ACCIONES_MEDIO_AMBIENTE < 4

4 [.01 .27 .22 .49] when EDAD_PROD is 1 or 2 & SUP_AGRICOLA < 9.3 & BOV >= 43

4 [.02 .18 .29 .51] when EDAD_PROD is 2 & SUP_AGRICOLA >=17.1 & TECNOLOGIA_CA < 7 &
ACCIONES_MEDIO_AMBIENTE < 4 & BOV >= 161

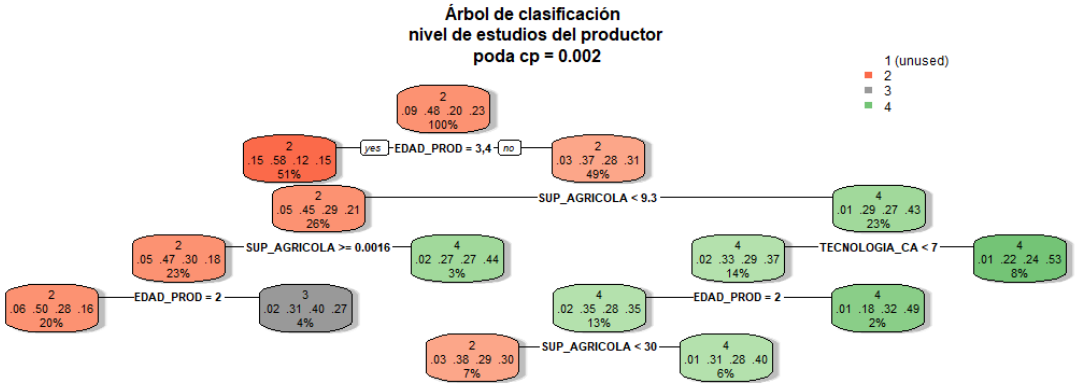
4 [.01 .22 .24 .53] when EDAD_PROD is 1 or 2 & SUP_AGRICOLA >= 9.3 & TECNOLOGIA_CA >= 7

Se observa que al menos en el árbol de clasificación obtenido con poda cp=0.002, no todas las 13 variables son consideradas en los resultados, además, que el gráfico de los tres árboles calculados anteriormente no son muy apreciables, y que en las reglas del tercer árbol obtenido (con poda cp=0.002) las variables que aparecen con mayor frecuencia son la “edad del productor”, “superficie agrícola” y “tecnologías en agricultura a cielo abierto”. Así pues, se decide obtener un cuarto árbol de clasificación considerando como variables independientes las tres mencionadas anteriormente, y como variable dependiente el “nivel de estudios del productor”, además de considerar un 80 por ciento de los datos como entrenamiento y un 20 por ciento como prueba, y una poda con valor del cp=0.002

De esta manera, se obtiene la matriz de confusión siguiente en la cual se observa que el 53.02 por ciento de los datos predichos coincide con los datos reales, recordando que el total de éstos es 11 222 unidades de producción agropecuaria.

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1    0    0    0    0
         2  928 4761 1554 1408
         3   11  135  190  108
         4   23  527  578  999

Overall Statistics

               Accuracy : 0.5302
                 95% CI : (0.5209, 0.5395)
    No Information Rate : 0.4832
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1855

 Mcnemar's Test P-Value : < 2.2e-16
```

Además, el gráfico del cuarto árbol se muestra en la siguiente figura, en la cual, los resultados obtenidos se aprecian mejor, más aún, se observa que hay ocho nodos terminales y que sigue sin considerarse los productores sin estudios.



Las ocho reglas se muestran en la siguiente imagen.

```
> rpart.rules(arbol)
 NIVEL_D_ESTUDIOS    1   2   3   4
               2 [.03 .38 .29 .30] when EDAD_PROD is      2 & SUP_AGRICOLA is 9.2750 to 29.9000 & TECNOLOGIA_CA <  7
               2 [.06 .50 .28 .16] when EDAD_PROD is      2 & SUP_AGRICOLA is 0.0016 to  9.2750
               2 [.15 .58 .12 .15] when EDAD_PROD is 3 or 4
               3 [.02 .31 .40 .27] when EDAD_PROD is      1 & SUP_AGRICOLA is 0.0016 to  9.2750
               4 [.01 .31 .28 .40] when EDAD_PROD is      2 & SUP_AGRICOLA >=           29.9000 & TECNOLOGIA_CA <  7
               4 [.02 .27 .27 .44] when EDAD_PROD is 1 or 2 & SUP_AGRICOLA <  0.0016
               4 [.01 .18 .32 .49] when EDAD_PROD is      1 & SUP_AGRICOLA >=            9.2750 & TECNOLOGIA_CA <  7
               4 [.01 .22 .24 .53] when EDAD_PROD is 1 or 2 & SUP_AGRICOLA >=            9.2750 & TECNOLOGIA_CA >= 7
```

Derivado del análisis observado con los cuatro árboles de clasificación distintos, obtenidos con información de variables de la Encuesta Nacional Agropecuaria 2019, se puede decir que, al menos para estos casos es mejor utilizar una menor cantidad de variables, ya que no todas serán tomadas

en cuenta, además de obtener árboles de clasificación con un cierto valor de poda, pues el programa R tiene sus limitantes al calcular las reglas y al generar los gráficos que representan al árbol de clasificación, lo cual provoca que dicho gráfico no se aprecie ocasionando que éste no sea fácil de interpretar, pues, los gráficos siempre son de gran ayuda visualmente.

La tabla 9 muestra una breve descripción de los cuatro árboles de clasificación que se analizaron, observando el error de prueba y empírico para cada uno de ellos, así como el porcentaje de predicción de cada árbol. De esta manera, la diferencia absoluta máxima entre los errores de prueba y empírico es detectada para el árbol sin poda, mientras que la diferencia mínima es detectada para el árbol con poda cp=0.002 y considerando las 13 variables. Además, los porcentajes de predicción de los árboles con poda mayor a cero son muy parecidos.

Tabla 9. Descripción y algunas características de los árboles analizados

Número	Descripción del árbol de clasificación	Error de prueba	Error empírico	Diferencia absoluta entre errores	Porcentaje de predicción
1	Árbol máximo con cp=0 y 13 variables	0.52433	0.34808	0.1762	47.57
2	Árbol óptimo cp=0.0002574334 y 13 variables	0.45856	0.45005	0.0085	54.14
3	Árbol con cp=0.002 y 13 variables	0.46632	0.46823	0.0019	53.37
4	Árbol con cp=0.002 y 4 variables	0.46979	0.47358	0.0038	53.02

Ahora, si comparamos las predicciones realizadas por el árbol óptimo con las predicciones de los otros árboles de clasificación se observa que:

1) Árbol óptimo vs árbol máximo: En promedio el 33.10 por ciento de las predicciones es diferente.
2) Árbol óptimo vs árbol con cp=0.002 y 13 variables: En promedio el 7.93 por ciento de las predicciones es diferente.
3) Árbol óptimo vs árbol con cp=0.002 y 4 variables: En promedio el 11.42 por ciento de las predicciones es diferente.

Por lo cual, se puede elegir el modelo obtenido con el tercer árbol, derivado de lo observado en los puntos anteriores, así como en la tabla 9 y en las matrices de confusión.

Método 5 Bosques de clasificación/ Descripción, Resultados con gráficas y discusión

La técnica bosques de clasificación, también conocida como **Random Forest**, es un método de aprendizaje estadístico introducido por Leo Breiman en 2001 y consiste en una combinación de árboles de clasificación, cuyo objetivo es agrupar un conjunto de predictores, los cuales, no necesariamente deben ser árboles óptimos, pues su finalidad no es como los árboles de clasificación que buscan un árbol óptimo. Además, es importante decir que cada árbol del bosque es construido a partir de una muestra Bootstrap, recordando que una muestra Bootstrap es aquella formada a partir de un conjunto de tamaño n, extrayendo aleatoriamente n observaciones con reemplazo teniendo cada observación una probabilidad 1/n de ser seleccionada. Más, aún, la generalización del error para los bosques es convergente a un límite si el número de árboles en el bosque es grande.

Para realizar el modelo Random Forest en el programa R, se utiliza la función randomForest, la cual hace uso del método bootstrap, con el propósito de entrenar cada árbol que será agregado en el bosque. En dicho proceso es conveniente dejar aproximadamente un tercio de los casos de la muestra; a los casos que no son considerados para entrenar el árbol se les llama out-of-bag (OOB). Con ellos se puede estimar un error insesgado de clasificación y también se pueden utilizar para hacer una estimación de la importancia de las variables. El funcionamiento de esta lógica es la siguiente:

1) Se escoge el error de clasificación out-of-bag (error OOB), seguido de esto es tomada una variable de forma aleatoria y se permutan sus valores dentro de los datos de entrenamiento, ocasionando que dicha variable escogida no correlacione lo aprendido por el modelo.
2) Se vuelve a calcular el error OOB, para después realizar la comparación con el error calculado inicialmente. En consecuencia, por lógica, si el error cambia, se afirma que dicha variable es importante.
3) Este proceso se repite con todas las variables y luego estas se ordenan de acuerdo con los cambios que produjeron cada una en los errores OOB.

El lector interesado en el tema puede checar el libro: Genuer, R. & Poggi, J-M. (2020). *Random Forests with R*. Springer

Otra manera de estimar la importancia de las variables en el modelo de bosques de clasificación (bosques aleatorios) es utilizando el criterio de Gini que consiste en seleccionar la variable en cada partición en la construcción de los árboles de clasificación y que corresponde a una disminución de esta medida. La importancia de una variable en un árbol se mide como la suma de los decrementos atribuidos a esa variable y la importancia final, como la media en todos los árboles.

Ahora, considerando las 13 variables utilizadas en el análisis realizado para árboles de clasificación, las cuales, a manera de recordatorio se mencionan a continuación: “nivel de estudios”, “edad del productor”, “sexo del productor”, “apoyo económico que recibe el productor para la producción”, “número de problemáticas reportadas en la unidad de producción”, “número de acciones realizadas para protección del medio ambiente”, “mano de obra”, “tractores, maquinaria y vehículos”, “tecnología a cielo abierto”, “superficie agrícola de la unidad de producción”, “cantidad de ganado bovino”, “cantidad de ganado porcino” y “cultivos en cielo abierto”. Además, de hacer uso de una semilla para poder replicar el análisis, la función randomForest, el conjunto de datos partido en un 80 por ciento como base de entrenamiento y un 20 por ciento como base de prueba, y con 500 árboles e indicando la importancia de las variables (importance=TRUE), se obtiene un modelo con el valor del OOB del 46.12 por ciento, lo cual quiere decir que 53.88 de los datos de entrenamiento el modelo lo predijo.

En la siguiente imagen, se observa que el número de árboles de clasificación es de 500, que el método de bosques aleatorios es de clasificación, que el número de variables probadas en cada división fue de 3, además, la estimación de la tasa de error del OOB es de 46.12 por ciento. También, aparece la matriz de confusión que distribuye a los datos de entrenamiento, en los cuales, en la diagonal se encuentran los datos que en su predicción coincidieron con los datos de entrenamiento indicando un total de 24 186 cuyo porcentaje es del 53.88 por ciento del total de datos de entrenamiento. Más aún, la columna “class.error” muestra el porcentaje de los datos por renglón que no fueron predichos por el modelo, por ejemplo, para el primer renglón se tiene que el 96.45 por ciento de las unidades de producción agropecuaria que son manejadas por productores que tienen nivel de estudios de primaria terminada (2), secundaria terminada (3) y preparatoria, licenciatura, maestria o doctorado terminado (4), se predijo como unidades que son manejadas por productores sin estudios (1). De la misma manera, en el renglón 2 se observa que

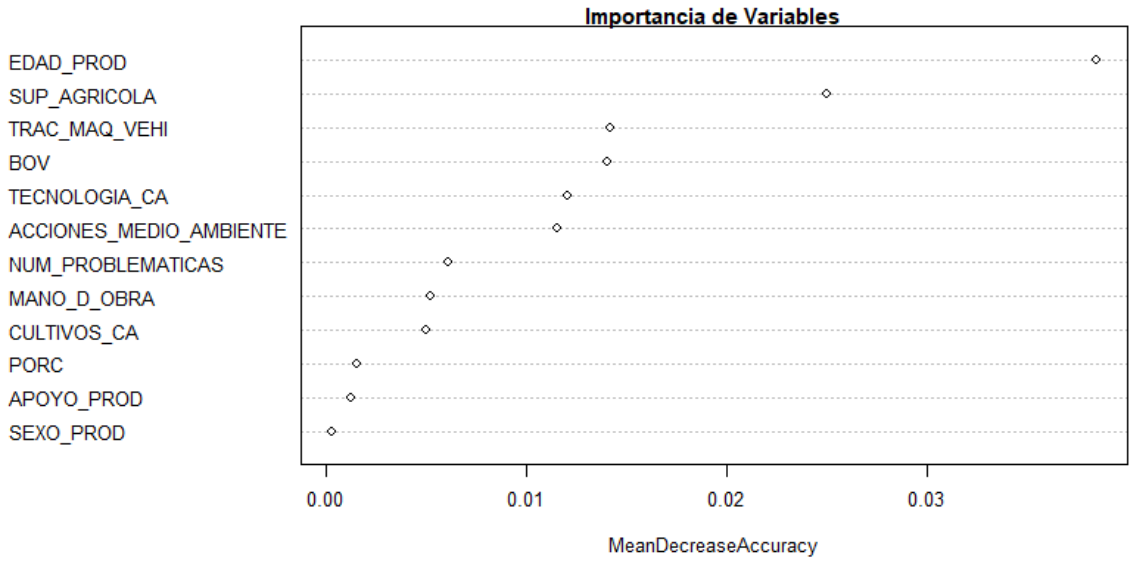
se predijo un 16.49 por ciento como unidades manejadas por productores que terminaron la primaria, cuando realmente son productores que no tienen estudios, terminaron la secundaria o terminaron la preparatoria, licenciatura, maestria o doctorado. Así pues, el modejo predijo un 86.98 por ciento de unidades manejadas por productores que terminaron la secundaria, cuando realmente son productores que no estudiaron, terminaron la primaria o terminaron la preparatoria, licenciatura, maestria o doctorado. Por último, el modelo predijo un 52.72 por ciento de los datos de entrenamiento como productores que terminaron la preparatoria, licenciatura, maestria o doctorado, cuando realmente éstos no estudiaron, terminaron la primaria o terminaron la secundaria.

```
Call:
 randomForest(formula = NIVEL_D_ESTUDIOS ~ ., data = bd_entrenamiento,      ntree = 500,
 importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 46.12%
Confusion matrix:
    1     2    3    4 class.error
1 145  3734   80  127   0.9645130
2 193 18019 1209 2158   0.1649752
3  20  5675 1164 2085   0.8698569
4  10  4461  948 4858   0.5272940
```

Por otro lado, al calcular las predicciones utilizando el conjunto de prueba y comparando éstas obteniendo una matriz de confusión (siguiente imagen), se observa que de un total de 11 222 datos, el modelo predice bien en un 54.84 por ciento de los datos del conjunto de prueba, cuyo valor absoluto es 6 154, el cual es la suma de los datos que se encuentran en la diagonal de la matriz de confusión. Así pues, comparando dicho porcentaje con los obtenidos en los primeros tres modelos (47.57%, 54.14% y 53.37% respectivamente) del apartado de árboles de clasificación, éste mejora. Sin embargo, al calcular el error de prueba y el error empírico estos son del 0.4517911 y 0.04337655 respectivamente.

```
p      1    2    3    4
  1   19   36    3    1
  2  890 4600 1437 1091
  3   19  275  334  222
  4   34  512  548 1201
```

Ahora, el siguiente gráfico muestra la importancia de las variables, considerando aquellas como más importantes: la edad del productor, la superficie agrícola, los tractores, maquinaria y vehículos que se utilizan para las actividades de la unidad de producción, las cabezas de bovinos y la tecnología en cultivos a cielo abierto. Y como variables menos importantes: las existencias de porcinos, el apoyo económico que recibe el productor para la producción y el sexo del productor.



Por otro lado, si se considera las cuatro variables: nivel de estudios, edad del productor, superficie agrícola y tecnologías utilizadas en cultivos a cielo abierto, para aplicar el método de bosques de clasificación con un total de 500 árboles, la misma base de entrenamiento y prueba (del 80 y 20 por ciento respectivamente) y considerando la importancia de las variables, se obtiene una tasa de error del OOB del 47.21 por ciento, indicando que el modelo no predice en este porcentaje a los datos de entrenamiento, los cuales se encuentran fuera de la diagonal de la matriz de confusión que aparece en la siguiente imagen.

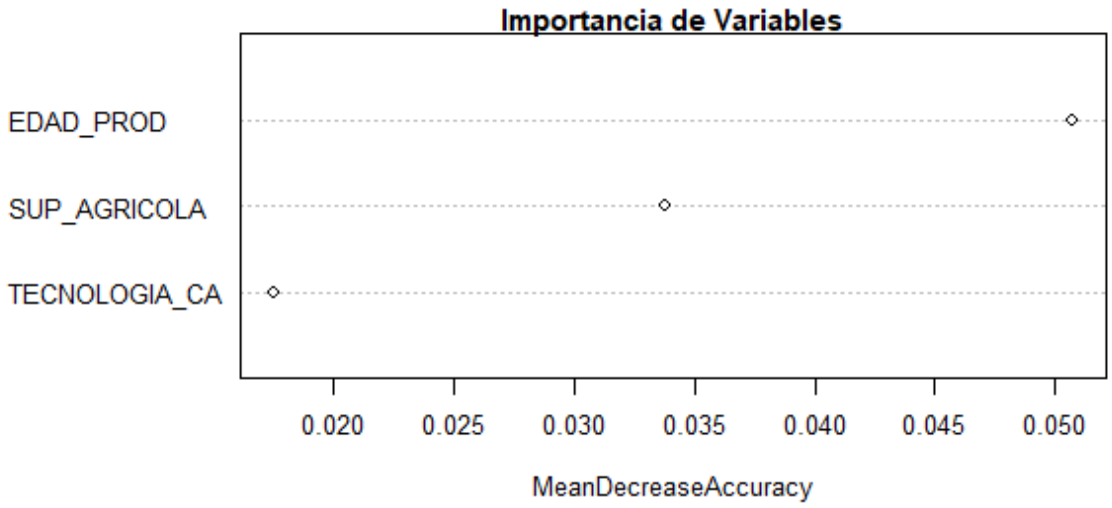
```
Call:
 randomForest(formula = NIVEL_D_ESTUDIOS ~ EDAD_PROD + SUP_AGRICOLA +      TECNOLOGIA_CA,
 data = bd_entrenamiento, ntree = 500, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 47.21%
Confusion matrix:
  1     2   3    4 class.error
1 0  3968  31   87   1.0000000
2 0 19066 491 2022   0.1164558
3 0  6161 633 2150   0.9292263
4 0  5861 418 3998   0.6109760
```

Sin embargo, aplicando los datos de prueba al modelo obtenido, se obtiene la siguiente matriz de confusión en la que se observa que el modelo no predijo ningún productor sin estudios, cuando realmente si los hay. El modelo predice bien 53.10 por ciento del total de datos, ya que son los que coinciden con los datos de prueba, pero el 46.90 por ciento representa a los datos de prueba que no fueron predichos, por lo cual, se encuentran fuera de la diagonal de la matriz de confusión. Así pues, comparando el 53.10 por ciento obtenido con bosques de clasificación y el porcentaje del 53.02 por ciento obtenido del cuarto modelo de árboles de clasificación, se observa que la mejora fue muy poca.

p	1	2	3	4
1	0	0	0	0
2	927	4778	1588	1439
3	11	131	196	91
4	24	514	538	985

El error de prueba en este caso es de 0.4689895, mientras que el error teórico es de 0.4665375.

La importancia de las variables se muestra en el siguiente gráfico, donde se observa que tienen el mismo orden de importancia que en el modelo anterior.



En conclusión, de los dos modelos obtenidos con bosques de clasificación, el mejor es el obtenido con las 13 variables, y más aún, es un mejor modelo que los obtenidos con el método árboles de clasificación.

Discusión de todos los resultados

La Encuesta Nacional Agropecuaria 2019 aporta bastante información del sector agropecuario, en una gran variedad de variables que caracterizan a las unidades de producción agropecuaria. Así pues, considerando la información de 25 variables numéricas y categóricas, se procede a realizar análisis univariado describiendo éstas tanto a nivel nacional como a nivel entidad. También, se aplican técnicas multivariadas para analizar la información agrupada de algunas variables.

Cabe mencionar que se realizó una distinta selección de variables objeto de análisis, para aplicarse en cada uno de los métodos estadísticos trabajados, observando que, por ejemplo, al realizar el análisis univariado georreferenciando a un cierto cultivo y utilizando la frecuencia de unidades de producción, los resultados fueron congruentes con los obtenidos al aplicar la técnica de análisis de correspondencia en su forma simple y considerando la información de las unidades que producen cierto cultivos.

Una unidad de producción agropecuaria es manejada por un productor, que por ciertas razones tiene un nivel de estudios clasificado como: sin estudios, primaria terminada, secundaria terminada, preparatoria o bachillerato terminado, licenciatura terminada, maestría o doctorado terminado. Analizando la información de esta variable de forma unidimensional y a nivel nacional, se observa que la mayoría de los productores tiene primaria terminada, es decir, que sólo estudiaron hasta sexto grado de primaria, después, le siguen aquellos productores que estudiaron un poco más llegando a terminar sus estudios en tercero de secundaria, es decir, terminaron la secundaria. Más aún, existe un porcentaje considerable de productores que dice no tener estudios, ya que nunca asistieron a la escuela. También, existen muy pocos productores que dicen tener un nivel de estudios superior al básico (preparatoria en adelante). Sin embargo, al analizar la información a nivel entidad federativa, se observa que los estados del norte tienen un porcentaje mayor de nivel de estudios superior al básico que los estados del sur, y de alguna manera parece que esto está ligado con lo observado en el escalamiento multidimensional pues en éste se observó que algunos estados del norte tienen mayor maquinaria, tractores y vehículos que utilizan para las labores agropecuarias, además, de observarse mayor superficie agrícola y superficie de riego, mientras que para el sur parece haber más mano de obra, así como poca superficie de riego. Lo mencionado anteriormente suena lógico pues se puede decir que entre más estudios tenga el productor, éste piensa en utilizar más tecnología que le ayude a rendir las labores agropecuarias y de esta manera que también le sea menos costoso.

También, al analizar por separado las variables de edad del productor y años que éste tiene realizando actividades agropecuarias, se observa que tiene una distribución parecida a la normal y que la mayoría de los productores son de 50 años o más, además de que la mayoría de éstos lleva la mayor parte de su vida dedicándose a las actividades agropecuarias. Más aún, al realizar el análisis factorial, se observó que estas dos variables están muy relacionadas, lo cual tiene mucho sentido.

Un porcentaje alto de productores a nivel nacional es de sexo masculino, pero al analizar esta variable a nivel entidad se observa que el porcentaje disminuye un poco en algunas de las entidades del sur, indicando que las mujeres ganan un poco de peso en dichos estados.

Por otro lado, se calculó la frecuencia de unidades de producción que producen cierto cultivo según la entidad federativa, de tal manera que se georreferencia a cada uno de estos cultivos observando que, por ejemplo, el aguacate es representativo de Michoacán o, por ejemplo, el chile

es un cultivo representativo del estado de Zacatecas. Lo anterior mencionado se realizó para los 25 cultivos objeto de interés de la ENA 2019, obteniendo como resultado la formación de grupos de estados en los cuales se cultiva cierto producto, o también, los grupos de cultivos formados que son producidos por cierto estado.

La técnica de análisis factorial aplicada a algunas variables de la ENA 2019, arrojó buenos resultados observando que en la vida real tiene sentido las agrupaciones que hizo para formar los factores observados, aunque el porcentaje de pérdida de información es considerable, ya que la varianza se explica en un 51.2 por ciento, sin embargo, esto implica que se puede seguir jugando con distintas variables para que dicho porcentaje de varianza explicada sean mejor.

El método de escalamiento multidimensional agrupó muy bien las entidades federativas a partir de la información de las variables ANIOS_EN_ACTIVIDAD, EDAD_PROD, MANO_D_OBRA, TRAC_MAQ_VEHI, SUP_AGRICOLA, SUP_RIEGO, CULTIVOS_CA y VALOR_PROD_CA, de hecho, con tan sólo la primera componente se obtuvo un porcentaje considerable y cercano al 100%, y con una segunda componente se acerca más al 100%, observando que algunos estados del norte son caracterizados por la maquinaria, los tractores y vehículos utilizados para las actividades agropecuarias, así como tener una superficie agrícola y una superficie de riego mayor, que por otro lado algunos estados del sur se caracterizan por su mano de obra, la edad del productor y los años que el productor lleva haciendo la actividad agropecuaria.

Con el análisis de correspondencia se formaron grupos de estados que producen cierto cultivo, y de igual forma grupos de cultivos que se producen en ciertos estados, que en su mayoría comparándolos con la vida real dichos grupos son muy congruentes, es decir, si representan la realidad.

Por último, se utilizó la información de las variables nivel de estudios y edad del productor, así como la superficie agrícola y las existencias de ganado bovino en cada unidad de producción agropecuaria para ilustrar los métodos de árboles de clasificación y bosques de clasificación, considerando sólo estas variables, pues la complejidad aumenta al agregar más variables y el costo computacional de graficar un árbol de clasificación es muy grande. Y como se esperaba, el modelo obtenido por bosques de clasificación fue mejor que por árboles de clasificación, sin embargo, el costo computacional utilizando el programa R si fue una limitante.

Conclusiones, recomendaciones y trabajos futuros

En base a las técnicas estadísticas multivariadas aplicadas y con los resultados obtenidos al ejecutarlos sobre la información de las 25 variables de la ENA 2019, se observa que, a grandes rasgos, la información está relacionada con la ubicación geográfica de las unidades de producción agropecuaria, es decir, es posible caracterizar en áreas geográficas (entidades federativas o municipios) la información de las unidades de producción.

También, se observó que esta relación geográfica permitió encontrar agrupaciones de entidades que bajo las distintas técnicas estadísticas, no necesariamente reflejan los mismos resultados.

Para aplicar los métodos estadísticos a la información de la ENA 2019 y realizar los análisis correspondientes, se apoyó del software estadístico R, obteniendo en general excelentes resultados, salvo al aplicar las técnicas de árboles de clasificación y bosques aleatorios, pues éstos se pudieron aplicar en su totalidad siembre y cuando se realicen podas, ya que al no realizarlas dicho software empieza a tener problemas como el no poder generar todas las reglas para el árbol de clasificación y es un poco tardado para generar bosques aleatorios con una cantidad de variables considerables, por lo cual se recomienda hacer uso de otros software que no tengas dichas limitantes, tal vez, Python.

En general, el objetivo establecido se cumplió de manera satisfactoria caracterizando la información agropecuaria en el país. Además, como trabajos futuros se espera aplicar las técnicas utilizadas, pero ahora a la información del censo agropecuario 2022, que se publicará a finales del año 2023. De esta manera, al aplicar dichas técnicas a la información del CA 2022 permitirá realizar comparativos con los resultados obtenidos al utilizar la información de la ENA 2019, con la finalidad de verificar si la caracterización que se realizó geográficamente se comporta de manera similar en la mayoría de las variables.

Por último, es importante mencionar algunas relaciones que pueden ser de interés para usuarios del sector agropecuario, en las cuales se puede hacer uso de la información del Censo Agropecuario 2022 y aplicando técnicas estadísticas para resumir la información:

1) Apoyos recibidos por parte del gobierno a través de diferentes programas gubernamentales (PO113_01), y niveles de productividad en las secciones de producción

agrícola, pecuaria y forestal (toneladas de cultivos, producción y venta de leche y carne, producción en metros cúbicos de productos maderables, etc.)

2) Establecimiento de la diferenciación proporcional de la producción agropecuaria en relación con el tipo de tenencia de la Unidad de Producción (UP).
3) Establecimiento de la diferenciación proporcional de la producción agropecuaria en relación con el tipo de derecho sobre la tierra de la Unidad de Producción.
4) Establecimiento de la diferenciación proporcional de la producción agrícola en relación con el tipo de régimen hídrico de la UP.
5) Relación de la superficie no sembrada y en descanso de acuerdo a los tipos de cultivos de la UP, cuando en ella se tienen dos o más cultivos, con el fin de identificar y ratificar de acuerdo al tipo de cultivo la necesidad del descanso, dada la diferenciación de consumo y aportación de nutrientes entre los diferentes cultivos.
6) La relación entre la superficie reportada de agostadero y la carga animal en la UP con el propósito de verificar el coeficiente de agostadero.
7) Clasificación y análisis del destino de la producción agrícola, en función del tamaño en hectáreas de la UP.
8) Análisis de las causas de las mermas de la producción agrícola en función de la producción vendida.
9) Relaciones analíticas entre el uso de tecnologías agrícolas y rendimientos agrícolas obtenidos.
10) Análisis de la producción obtenida de cultivos en agricultura protegida en función del tipo de instalación utilizada.
11) Análisis sobre el tipo de manejo del ganado bovino (CB161) utilizado en la UP en función del tamaño en Hectáreas (Ha) y el total de cabezas de ganado.
12) Análisis sobre el tipo de calidad del ganado bovino (CB131) utilizado en la UP en función del tamaño en Ha y el total de cabezas de ganado.

Claro que las relaciones mencionadas anteriormente, son solo algunas y que son de interés seguramente para un usuario que tiene el perfil de agrónomo, sin embargo, el Censo Agropecuario aporta demasiada información de tal manera que cualquier tipo de usuario (agrónomo, veterinario, científico, etc.) puede plantear sus propias preguntas y con ella iniciar distintos análisis, dependiendo de los objetivos de cada usuario.

Referencias

Vences, J.. (1999). *Estadística Multivariada, Análisis Factorial*. Estado de Aguascalientes: Instituto de Educación de Aguascalientes.

Johson, R.A. & Wichern, D.W.. (2007). *Applied multivariate statistical analysis*. 6th: Ed. Prentice-Hall.

Izenman, J.. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. Springer.

Genuer, R. & Poggi, J-M. (2020). *Random Forests with R*. Springer

Mejía, J.. (2017). *Las Ciencias de la Administración y el Análisis Multivariante, Tomo II Las Técnicas Interdependientes*. Zapopan, Jalisco, México: Universidad de Guadalajara.

INEGI. (2020). *Encuesta Nacional Agropecuaria 2019*. ENA. Metodología. 17 de mayo de 2021, de INEGI Sitio web: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825197315.

INEGI. (2020). *Encuesta Nacional Agropecuaria 2019 Cuestionario*. 15 de abril de 2021, de INEGI Sitio web: https://www.inegi.org.mx/contenidos/programas/ena/2019/doc/cuest ena19.pdf

INEGI. (2020). *Encuesta Nacional Agropecuaria. Presentación*. 28 de abril de 2021, de INEGI Sitio web: https://www.inegi.org.mx/contenidos/programas/ena/2019/doc/rrdp ena2019.pdf

De la Fuente S. (2011). *Análisis Factorial*. 2011, de Universidad Autónoma de Madrid (UAM) Sitio web: https://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/FACTORIAL/analisis-factorial.pdf

López-Roldan, P., & Fachelli, S.. (2015). *Metodología de la Investigación Social Cuantitativa*. febrero de 2015, de Universidad Autónoma de Barcelona Sitio web: https://ddd.uab.cat/pub/caplli/2015/142928/metinvsoccua_cap3-11a2016v3.pdf

Tapia G. & García J.. (2001). *Análisis Factorial y Componentes Principales: su Uso para Modelos Macroeconométricos de la Economía Mexicana*. octubre de 2001, de Profesores Investigadores de la Facultad de Economía "Vasco de Quiroga" de la UMSNH. Sitio web: Análisis Factorial y Componentes Principales – Dialnet https://dialnet.unirioja.es› descarga › articulo

Mejía, J.. (2017). *Las Ciencias de la Administración y el Análisis Multivariante, Tomo II Las Técnicas Interdependientes*. 2017, de Universidad de Guadalajara Sitio web: http://dca.cucea.udg.mx/sites/default/files/adjuntos/2017_tomo_ii_las_ciencias_de_la_administracion_y_el_analisis_multivariante.pdf

Detrinidad, E.. (2016). *Análisis Factorial Exploratorio y Confirmatorio aplicado al modelo de secularización propuesto por Inglehart-Norris. Periodo 2010-2014 (Estudio de caso España, Estados Unidos, Alemania, Holanda) WSV.* . Julio de 2016, de Universidad de Granada Sitio web: https://masteres.ugr.es/moea/pages/curso201516/tfm1516/detrinidad_barquero_tfm/!

Cardona N.. (2019). *Predicción y selección de variables con bosques aleatorios en presencia de variables correlacionadas.* septiembre de 2019, de Universidad Nacional de Colombia Sitio web: https://repositorio.unal.edu.co/bitstream/handle/unal/75561/8063120.2019.pdf?isAllowed=y&sequence=1

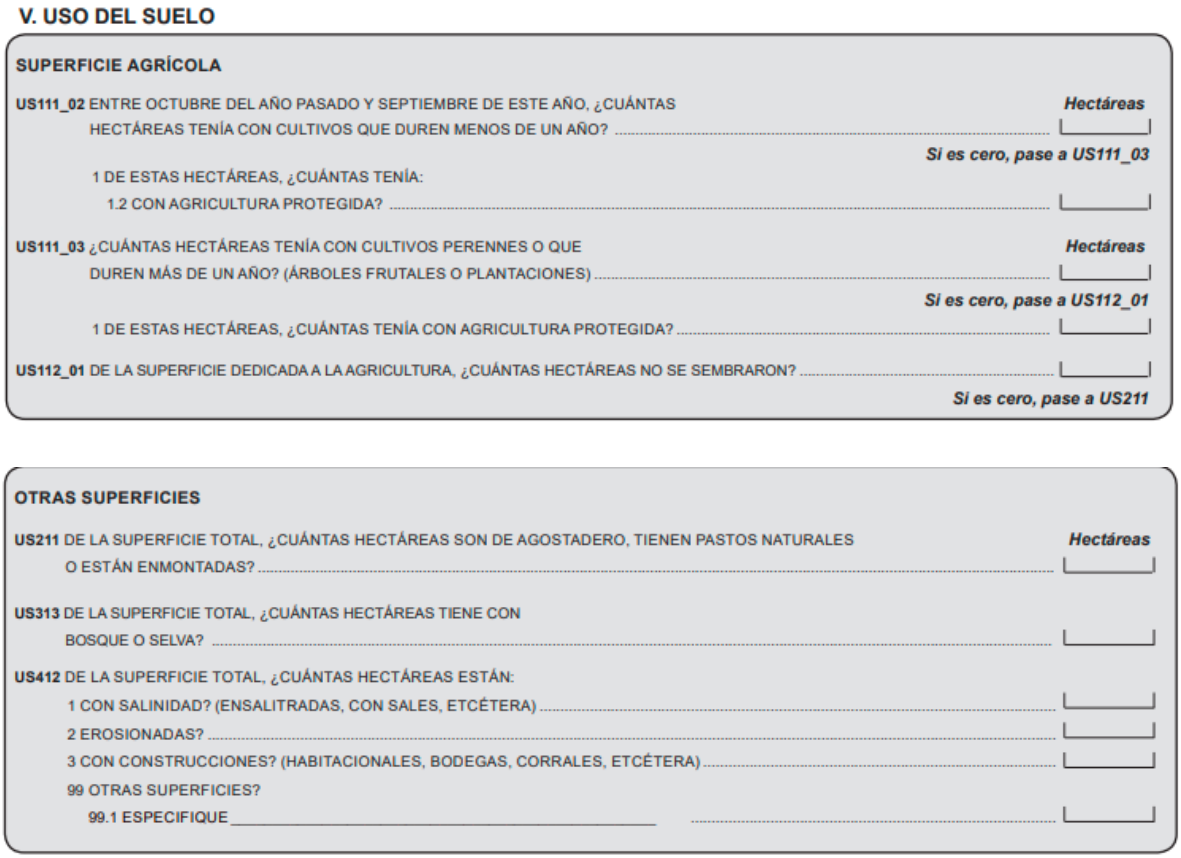
Greenacre M.. (2008). *La práctica del análisis de correspondencia.* julio 2008, de Fundación BBVA Sitio web: https://www.fbbva.es/wp-

content/uploads/2017/05/dat/DE_2008_practica_analisis_correspondencias.pdf

Anexo A

Algunas imágenes del Cuestionario de la ENA2019

A continuación, se muestran algunas imágenes del cuestionario de la ENA2019, con la intención de observar qué variables son las utilizadas para formar las 25 variables objeto de análisis que son mencionadas en la tabla 1. Para más detalle, se puede consultar el cuestionario completo en la liga https://www.inegi.org.mx/contenidos/programas/ena/2019/doc/cuest_ena19.pdf



V. USO DEL SUELO

SUPERFICIE AGRÍCOLA

US111_02 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, ¿CUÁNTAS HECTÁREAS TENÍA CON CULTIVOS QUE DUREN MENOS DE UN AÑO? ..... *Hectáreas* ⌊___⌋

Si es cero, pase a US111_03

1 DE ESTAS HECTÁREAS, ¿CUÁNTAS TENÍA:

1.2 CON AGRICULTURA PROTEGIDA? ..... ⌊___⌋

US111_03 ¿CUÁNTAS HECTÁREAS TENÍA CON CULTIVOS PERENNES O QUE DUREN MÁS DE UN AÑO? (ÁRBOLES FRUTALES O PLANTACIONES) ..... *Hectáreas* ⌊___⌋

Si es cero, pase a US112_01

1 DE ESTAS HECTÁREAS, ¿CUÁNTAS TENÍA CON AGRICULTURA PROTEGIDA? ..... ⌊___⌋

US112_01 DE LA SUPERFICIE DEDICADA A LA AGRICULTURA, ¿CUÁNTAS HECTÁREAS NO SE SEMBRARON? ..... ⌊___⌋

Si es cero, pase a US211

OTRAS SUPERFICIES

US211 DE LA SUPERFICIE TOTAL, ¿CUÁNTAS HECTÁREAS SON DE AGOSTADERO, TIENEN PASTOS NATURALES O ESTÁN ENMONTADAS? ..... *Hectáreas* ⌊___⌋

US313 DE LA SUPERFICIE TOTAL, ¿CUÁNTAS HECTÁREAS TIENE CON BOSQUE O SELVA? ..... ⌊___⌋

US412 DE LA SUPERFICIE TOTAL, ¿CUÁNTAS HECTÁREAS ESTÁN:

1 CON SALINIDAD? (ENSALITRADAS, CON SALES, ETCÉTERA) ..... ⌊___⌋

2 EROSIONADAS? ..... ⌊___⌋

3 CON CONSTRUCCIONES? (HABITACIONALES, BODEGAS, CORRALES, ETCÉTERA) ..... ⌊___⌋

99 OTRAS SUPERFICIES?

99.1 ESPECIFIQUE _____ ..... ⌊___⌋

VI. SISTEMAS DE RIEGO, CALIDAD Y ORIGEN DEL AGUA

ii. Entrevistador: Si la suma de US111_02, US111_03 y US112_01 es mayor a cero, continúe, si no, pase a CB112.

AR111_01 DE LA SUPERFICIE QUE DEDICÓ A LA AGRICULTURA: ***Hectáreas***

1 ¿CUÁNTAS HECTÁREAS SON DE TEMPORAL? ..... ⌊___⌋

Si es cero, pase a AR111_02

1.1 DE ESTA SUPERFICIE, ¿CUÁNTAS HECTÁREAS SON DE JUGO O HUMEDAD? ..... ⌊___⌋

AR111_02 ¿CUÁNTAS HECTÁREAS SON DE RIEGO? ..... ⌊___⌋

Si es cero, pase a nota iii

VII. AGRICULTURA

AGRICULTURA A CIELO ABIERTO (CULTIVOS, ÁRBOLES FRUTALES O PLANTACIONES)

iii. Entrevistador: Si contestó que tiene cultivos anuales a cielo abierto o superficie con cultivos perennes (árboles frutales o plantaciones) continúe, si no, pase a AGRICULTURA PROTEGIDA (página 9).

iv. Entrevistador: Si sembró el mismo cultivo más de una vez en el periodo de referencia, o el cultivo es de riego y temporal regístrelos por separado en una hoja anexa. Para cada cultivo o plantación realice las siguientes preguntas:

		AA111_05 ¿QUÉ MODALIDAD ES?
AA111_02 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, ¿QUÉ CULTIVO SEMBRÓ O QUÉ ÁRBOL FRUTAL O PLANTACIÓN TIENE?	***Nombre del cultivo o plantación***	1 RIEGO 2 TEMPORAL
	_____	⌊__⌋
		Mes ***Año***

PRODUCCIÓN

Hectáreas

AA111_13 DE ESTE CULTIVO, ¿CUÁNTAS HECTÁREAS COSECHÓ O COSECHARÁ? ..... ⌊___⌋

Si es cero, pase a nota vii

Toneladas

AA111_17 ¿CUÁNTAS TONELADAS OBTUVO O ESPERA OBTENER? ..... ⌊___⌋

COMERCIALIZACIÓN

vi. Entrevistador: Si DA115 es mayor a cero, continúe, sino pase a nota vii.

Clave

DA122 PRINCIPALMENTE, ¿A QUIÉN LE VENDIÓ O ESPERA VENDER? ..... ⌊___⌋

1. CENTRO DE ACOPIO SEGALMEX	*2. DIRECTAMENTE AL CONSUMIDOR*	*3. INTERMEDIARIO (COYOTE)*	*4. CENTRAL DE ABASTOS*
5. CENTRO COMERCIAL O SUPERMERCADO	*6. EMPACADORA O USO INDUSTRIAL (INGENIO, PROCESADORA, ETCÉTERA)*	*7. DIRECTAMENTE A OTRO PAÍS*	*8. BAJO CONTRATO*
9. BODEGA, ALMACÉN O CENTRO DE ACOPIO	*99. OTRO COMPRADOR*		

Pesos

DA132_09 ¿CUÁL FUE EL PRECIO QUE LE PAGARON O PAGARÁN POR TONELADA EN LA ÚLTIMA VENTA? ..... ⌊_____⌋

TECNOLOGÍA EN AGRICULTURA A CIELO ABIERTO

AT112 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA REALIZAR LAS ACTIVIDADES AGRÍCOLAS: ***Marque con "x" la respuesta***

AT111_42 ¿USÓ COA O AZADÓN?	SÍ ☐ 1	NO ☐ 2
AT112_04 ¿USÓ ANIMALES DE TIRO O YUNTA?	SÍ ☐ 1	NO ☐ 2
AT112_06 ¿REALIZÓ ROTACIÓN DE CULTIVOS?	SÍ ☐ 1	NO ☐ 2
AT111_27 ¿USÓ HERBICIDAS:		
1 QUÍMICOS?	SÍ ☐ 1	NO ☐ 2
2 ORGÁNICOS?	SÍ ☐ 1	NO ☐ 2
AT111_28 ¿USÓ INSECTICIDAS:		
1 QUÍMICOS?	SÍ ☐ 1	NO ☐ 2
2 ORGÁNICOS?	SÍ ☐ 1	NO ☐ 2
AT111_30 ¿USÓ FUNGICIDAS (CONTRA LA CENICILLA, ROYA, MILDIU, ETCÉTERA):		
1 QUÍMICOS?	SÍ ☐ 1	NO ☐ 2
2 ORGÁNICOS?	SÍ ☐ 1	NO ☐ 2
AT112_11 ¿REALIZÓ INJERTO DE ÁRBOLES?	SÍ ☐ 1	NO ☐ 2
AT112_08 ¿REALIZÓ PODAS?	SÍ ☐ 1	NO ☐ 2
AT112_02 ¿USÓ SEMBRADORAS?	SÍ ☐ 1	NO ☐ 2
AT112_03 ¿USÓ COSECHADORA, TRILLADORA O COMBINADA?	SÍ ☐ 1	NO ☐ 2
AT111_32 ¿USÓ DESGRANADORAS?	SÍ ☐ 1	NO ☐ 2
AT112_09 ¿REALIZÓ CONTROL BIOLÓGICO DE PLAGAS?	SÍ ☐ 1	NO ☐ 2
AT112_13 ¿USÓ SENSOR DE HUMEDAD?	SÍ ☐ 1	NO ☐ 2
AT112_14 ¿USÓ SENSOR DE NITRÓGENO, COLORACIÓN O VERDOR?	SÍ ☐ 1	NO ☐ 2

Marque con "x" la respuesta

AT112_15 ¿USÓ MEJORADORES DE SUELO?	SÍ ☐ 1	NO ☐ 2
AT111_33 ¿USÓ POLINIZACIÓN CONTROLADA?	SÍ ☐ 1	NO ☐ 2
AT112_05 ¿REALIZÓ LABRANZA DE CONSERVACIÓN DE SUELOS?	SÍ ☐ 1	NO ☐ 2
AT111_34 ¿CUMPLIÓ ALGUNA NORMA OFICIAL DE SANIDAD VEGETAL?	SÍ ☐ 1	NO ☐ 2
AT111_35 ¿HA REALIZADO ANÁLISIS DE SUELOS?	SÍ ☐ 1	NO ☐ 2
AT112_07 ¿REALIZÓ QUEMAS CONTROLADAS?	SÍ ☐ 1	NO ☐ 2
AT112_10 ¿RECIBIÓ ASISTENCIA TÉCNICA PARA LA PRODUCCIÓN?	SÍ ☐ 1	NO ☐ 2
AT112_16 ¿RECIBIÓ OTRO TIPO DE ASISTENCIA TÉCNICA? (COMERCIALIZACIÓN, TRÁMITE DE CRÉDITO U OTRA)	SÍ ☐ 1	NO ☐ 2
AT112_99 ¿UTILIZÓ ALGUNA OTRA TECNOLOGÍA DIFERENTE A LAS ANTERIORES?	SÍ ☐ 1	NO ☐ 2

99.1 ESPECIFIQUE _____

VIII. CRÍA Y EXPLOTACIÓN DE ANIMALES

BOVINOS

xvi. ***Entrevistador: Recuerde que las siguientes preguntas se refieren al municipio del que estamos hablando.***

CB112 EL 30 DE SEPTIEMBRE DE ESTE AÑO, CONSIDERANDO A LOS RECIÉN NACIDOS, ¿CUÁNTAS RESES TENÍA EN TOTAL EN ESTE MUNICIPIO? (NO CONSIDERE AL GANADO DE LIDIA) ***Cabezas*** ☐☐☐☐

CB121 EL 30 DE SEPTIEMBRE DE ESTE AÑO, DEL TOTAL DE SUS RESES, ¿CUÁNTAS ERAN: *Cabezas*

1 ANIMALES PARA TRABAJO? (TIRO O YUNTA) ____

Si CB112 es menor a 5, pase a CB161

2 VAQUILLAS PARA REEMPLAZO? ____

Meses

2.1 ¿A QUÉ EDAD TIENEN SU PRIMER PARTO? ____

Cabezas

3 ANIMALES EN ENGORDA? ____

Kilogramos

3.2 ¿CUÁL ES EL PESO PROMEDIO DE VENTA? ____

Cabezas

4 SEMENTALES? ____

5 VACAS SOLO PARA LA CRÍA DE BECERROS? ____

6 VACAS SOLO PARA LA PRODUCCIÓN DE LECHE? ____

6.1 DE ESTAS, ¿CUÁNTAS VACAS ORDEÑÓ? ____

Litros

CB122_01 EN PROMEDIO, ¿CUÁNTOS LITROS DE LECHE OBTUVO AL DÍA? ____

1 DE ESTA CANTIDAD, ¿CUÁNTOS LITROS DESTINÓ AL CONSUMO DE SU FAMILIA? ____

2 DE ESTA CANTIDAD, ¿CUÁNTOS LITROS VENDIÓ O ESPERA VENDER? ____

Pesos

CB122_05 ¿A CÓMO LE PAGARON EL LITRO DE LECHE? ____

VOLUMEN DE VENTAS DEL GANADO BOVINO

Marque con "x" la respuesta

CB171 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, ¿VENDIÓ RESES? ..... SÍ ☐ 1 NO ☐ 2

Pase a CP112

CB172 ¿QUÉ ANIMALES VENDIÓ? (*1)	CB172_01 ¿CUÁNTAS CABEZAS VENDIÓ?	CB172_02 ¿A QUIÉN LE VENDIÓ? (*2)	CB173 ¿CUÁL FUE EL PRECIO QUE LE PAGARON POR CABEZA EN LA ÚLTIMA VENTA?	
Clave	*Cabezas*	*Clave*	*Pesos*	*Unidad de medida*

**1) 1. BECERROS 2. PIE DE CRÍA 3. VAQUILLAS PARA REEMPLAZO 4. ANIMALES EN DESARROLLO*
5. ANIMALES EN ENGORDA 6. SEMENTALES 7. VACAS SOLO PARA LA CRÍA DE BECERROS 8. VACAS SOLO PARA LA
9. VACAS PARA LA CRÍA DE BECERROS Y ORDEÑA (DOBLE PROPÓSITO) 10. ANIMALES DE DESECHO 11. ANIMALES PARA TRABAJO PRODUCCIÓN DE LECHE

**2) 1. CENTRO DE ACOPIO SEGALMEX 2. INTERMEDIARIO (COYOTE) 3. DIRECTAMENTE AL CONSUMIDOR 4. CENTRAL DE ABASTOS*
5. CENTRO COMERCIAL O SUPERMERCADO 6. RASTRO TIF (TIPO INSPECCIÓN FEDERAL) 7. RASTRO MUNICIPAL 8. RASTRO PRIVADO
9. DIRECTAMENTE A OTRO PAÍS 99. OTRO COMPRADOR

PORCINOS

CP112 EL 30 DE SEPTIEMBRE DE ESTE AÑO, CONSIDERANDO A LOS RECIÉN NACIDOS, ¿CUÁNTOS MARRANOS O MARRANAS TENÍA EN TOTAL EN ESTE MUNICIPIO? ____ *Cabezas*

VOLUMEN DE VENTAS DEL GANADO PORCINO

Marque con "x" la respuesta

CP171 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, ¿VENDIÓ MARRANOS O MARRANAS? ..... SÍ ☐ 1 NO ☐ 2

Pase a CA112

CP172 ¿QUÉ ANIMALES VENDIÓ? (*1)	CP172_01 ¿CUÁNTAS CABEZAS VENDIÓ?	CP172_02 ¿A QUIÉN LE VENDIÓ? (*2)	CP173 ¿CUÁL FUE EL PRECIO QUE LE PAGARON POR CABEZA EN LA ÚLTIMA VENTA?	
Clave	*Cabezas*	*Clave*	*Pesos*	*Unidad de medida*

**1) 1. SEMENTALES 2. VIENTRES 3. LECHONES 4. ANIMALES EN DESARROLLO O ENGORDA*
5. ANIMALES DE DESECHO 6. ANIMALES FINALIZADOS

**2) 1. CENTRO DE ACOPIO SEGALMEX 2. INTERMEDIARIO (COYOTE) 3. DIRECTAMENTE AL CONSUMIDOR 4. CENTRAL DE ABASTOS*
5. CENTRO COMERCIAL O SUPERMERCADO 6. RASTRO TIF (TIPO INSPECCIÓN FEDERAL) 7. RASTRO MUNICIPAL 8. RASTRO PRIVADO
9. DIRECTAMENTE A OTRO PAÍS 99. OTRO COMPRADOR

TECNOLOGÍA EN BOVINOS

CB141 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA EL MANEJO DE SUS RESES: *Marque con "x" la respuesta*

1 ¿LAS DESPARASITÓ INTERNAMENTE?	SÍ ☐ 1	NO ☐ 2
1.1 ¿CUÁNTAS VECES AL AÑO?	*Cantidad* ☐☐	
2 ¿LAS VACUNÓ?	SÍ ☐ 1	NO ☐ 2
2.1 ¿CUÁNTAS VECES AL AÑO?	*Cantidad* ☐☐	
3 ¿LAS DESPARASITÓ EXTERNAMENTE? (BAÑO CONTRA LA GARRAPATA, SARNA, PIOJOS U OTROS PARÁSITOS)	SÍ ☐ 1	NO ☐ 2
3.1 ¿CUÁNTAS VECES AL AÑO?	*Cantidad* ☐☐	
4 ¿REALIZÓ ROTACIÓN DE POTREROS?	SÍ ☐ 1	NO ☐ 2
5 ¿LES DIO ALIMENTO BALANCEADO O ALIMENTO PREPARADO?	SÍ ☐ 1	NO ☐ 2
7 ¿LAS LLEVÓ A PASTOREAR EN POTREROS CON PASTO NATIVO?	SÍ ☐ 1	NO ☐ 2
8 ¿LAS LLEVÓ A PASTOREAR EN POTREROS CON PASTO INDUCIDO?	SÍ ☐ 1	NO ☐ 2
9 ¿LES DIO RASTROJO O ESQUILMO?	SÍ ☐ 1	NO ☐ 2
10 ¿USÓ LA MONTA CONTROLADA?	SÍ ☐ 1	NO ☐ 2
12 ¿LAS INSEMINÓ ARTIFICIALMENTE?	SÍ ☐ 1	NO ☐ 2
12.1 ¿USÓ SEMEN SEXADO?	SÍ ☐ 1	NO ☐ 2
13 ¿APLICÓ ALGÚN PROGRAMA DE MEJORAMIENTO GENÉTICO?	SÍ ☐ 1	NO ☐ 2
14 ¿LES APLICÓ HORMONAS?	SÍ ☐ 1	NO ☐ 2
16 ¿LES TRANSFIRIÓ EMBRIONES?	SÍ ☐ 1	NO ☐ 2
17 ¿USÓ TANQUE ENFRIADOR?	SÍ ☐ 1	NO ☐ 2
18 ¿LES DIO SALES MINERALES?	SÍ ☐ 1	NO ☐ 2
19 ¿USÓ LA MONTA DIRIGIDA?	SÍ ☐ 1	NO ☐ 2
20 ¿RECIBIÓ ASISTENCIA TÉCNICA PARA LA PRODUCCIÓN?	SÍ ☐ 1	NO ☐ 2
21 ¿RECIBIÓ OTRO TIPO DE ASISTENCIA TÉCNICA? (COMERCIALIZACIÓN, TRÁMITE DE CRÉDITO U OTRA)	SÍ ☐ 1	NO ☐ 2
22 ¿USÓ FUEGO PARA EL CONTROL DE MALEZAS?	SÍ ☐ 1	NO ☐ 2
23 ¿USÓ FUEGO PARA MEJORAR EL REBROTE DE LOS PASTOS?	SÍ ☐ 1	NO ☐ 2
24 ¿LLEVÓ A CABO ALGUNA PRÁCTICA DE CONSERVACIÓN DE FORRAJE?	SÍ ☐ 1	NO ☐ 2
25 ¿UTILIZÓ ALGÚN PROGRAMA DE BIOSEGURIDAD?	SÍ ☐ 1	NO ☐ 2
26 ¿CUMPLIÓ ALGUNA NORMA OFICIAL DE SANIDAD ANIMAL?	SÍ ☐ 1	NO ☐ 2
27 ¿QUÉ TIPO DE SISTEMA DE IDENTIFICACIÓN DE GANADO UTILIZÓ:		
27.1 ARETADO?	SÍ ☐ 1	NO ☐ 2
27.2 MARCA CON FIERRO?	SÍ ☐ 1	NO ☐ 2
99 ¿UTILIZÓ ALGUNA OTRA TECNOLOGÍA DIFERENTE A LAS ANTERIORES?	SÍ ☐ 1	NO ☐ 2

TECNOLOGÍA EN PORCINOS

CP141 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, A SUS MARRANOS O MARRANAS: *Marque con "x" la respuesta*

1 ¿LOS DESPARASITÓ INTERNA O EXTERNAMENTE?	SÍ ☐ 1	NO ☐ 2
1.1 ¿CUÁNTAS VECES AL AÑO?	*Cantidad* ☐☐	
2 ¿LOS VACUNÓ?	SÍ ☐ 1	NO ☐ 2
2.1 ¿CUÁNTAS VECES AL AÑO?	*Cantidad* ☐☐	
3 ¿LES DIO ALIMENTO BALANCEADO?	SÍ ☐ 1	NO ☐ 2
3.1 EN PROMEDIO, POR CABEZA, ¿CUÁNTOS KILOGRAMOS DE ALIMENTO BALANCEADO LES DIO AL DÍA?	*Kilogramos* ☐☐	
3.2 ¿USTED ELABORÓ EL ALIMENTO BALANCEADO?	SÍ ☐ 1	NO ☐ 2
4 ¿RECIBIÓ ASISTENCIA TÉCNICA PARA LA PRODUCCIÓN?	SÍ ☐ 1	NO ☐ 2
5 ¿UTILIZÓ INSEMINACIÓN ARTIFICIAL?	SÍ ☐ 1	NO ☐ 2
6 ¿APLICÓ ALGÚN PROGRAMA DE MEJORAMIENTO GENÉTICO?	SÍ ☐ 1	NO ☐ 2
7 ¿RECIBIÓ OTRO TIPO DE ASISTENCIA TÉCNICA? (COMERCIALIZACIÓN, TRÁMITE DE CRÉDITO U OTRA)	SÍ ☐ 1	NO ☐ 2
8 ¿CUENTA CON ALGÚN SISTEMA DE IDENTIFICACIÓN DE GANADO? (ARETADO, MUESCAS)	SÍ ☐ 1	NO ☐ 2
99.1 ESPECIFIQUE _____		
9 ¿CUMPLIÓ ALGUNA NORMA OFICIAL DE SANIDAD ANIMAL?	SÍ ☐ 1	NO ☐ 2
99 ¿UTILIZÓ ALGUNA OTRA TECNOLOGÍA DIFERENTE A LAS ANTERIORES?	SÍ ☐ 1	NO ☐ 2
99.1 ESPECIFIQUE _____		

XIII. PROBLEMÁTICA

PP111 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA LAS ACTIVIDADES AGROPECUARIAS, ¿QUÉ PROBLEMAS SE LE PRESENTARON: *Marque con "x" la respuesta*

1 ALTOS COSTOS DE INSUMOS Y SERVICIOS?	SÍ ☐ 1	NO ☐ 2
2 PÉRDIDA DE LA COSECHA O ANIMALES POR CAUSAS CLIMÁTICAS, COMO:		
2.1 SEQUÍAS?	SÍ ☐ 1	NO ☐ 2
2.2 EXCESO DE HUMEDAD?	SÍ ☐ 1	NO ☐ 2
2.3 INUNDACIONES?	SÍ ☐ 1	NO ☐ 2
2.4 HELADAS?	SÍ ☐ 1	NO ☐ 2
2.5 BAJAS TEMPERATURAS?	SÍ ☐ 1	NO ☐ 2
2.6 VIENTOS?	SÍ ☐ 1	NO ☐ 2
2.7 GRANIZO?	SÍ ☐ 1	NO ☐ 2
2.8 INCENDIOS NATURALES?	SÍ ☐ 1	NO ☐ 2
3 FALTA DE CAPACITACIÓN Y ASISTENCIA TÉCNICA?	SÍ ☐ 1	NO ☐ 2
4 PÉRDIDA DE FERTILIDAD DEL SUELO?	SÍ ☐ 1	NO ☐ 2
5 INFRAESTRUCTURA INSUFICIENTE PARA LA PRODUCCIÓN?	SÍ ☐ 1	NO ☐ 2
6 DIFICULTAD PARA LA COMERCIALIZACIÓN DEBIDO A LA EXISTENCIA DE INTERMEDIARIOS? (COYOTES)	SÍ ☐ 1	NO ☐ 2
7 DIFICULTAD PARA LA COMERCIALIZACIÓN DEBIDO A LOS PRECIOS BAJOS?	SÍ ☐ 1	NO ☐ 2
8 PRODUCTOR DE EDAD AVANZADA O ENFERMO?	SÍ ☐ 1	NO ☐ 2
9 FALTA DE ORGANIZACIÓN PARA LA PRODUCCIÓN?	SÍ ☐ 1	NO ☐ 2
10 NO PODER OBTENER EL CRÉDITO?	SÍ ☐ 1	NO ☐ 2
11 FALTA DE DOCUMENTACIÓN PARA ACREDITAR LA POSESIÓN DE LA TIERRA?	SÍ ☐ 1	NO ☐ 2
12 LITIGIO O INVASIÓN DE LA TIERRA?	SÍ ☐ 1	NO ☐ 2
13 INSEGURIDAD?	SÍ ☐ 1	NO ☐ 2
14 FALTA DE INFORMACIÓN DE LOS PRECIOS DE LOS PRODUCTOS?	SÍ ☐ 1	NO ☐ 2
16 PÉRDIDA DE LA COSECHA O ANIMALES POR CAUSAS BIOLÓGICAS, CÓMO:		
16.1 PLAGAS?	SÍ ☐ 1	NO ☐ 2
16.2 ENFERMEDADES?	SÍ ☐ 1	NO ☐ 2

II. ORGANIZACIÓN Y APOYO

ORGANIZACIÓN DE LOS PRODUCTORES

OP111 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, ¿SE ORGANIZÓ CON OTROS PRODUCTORES PARA OBTENER ALGÚN APOYO O SERVICIO? *Marque con "x" la respuesta*

SÍ ☐ 1 NO ☐ 2

Pase a PO111

APOYOS PARA LA PRODUCCIÓN

PO111 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, ¿RECIBIÓ ALGÚN APOYO ECONÓMICO PARA LA PRODUCCIÓN? *Marque con "x" la respuesta*

SÍ ☐ 1 NO ☐ 2

Pase a SE116

1 ¿FUE PRIVADO?	SÍ ☐ 1	NO ☐ 2
2 ¿FUE POR PARTE DEL GOBIERNO?	SÍ ☐ 1	NO ☐ 2

Pase a SE116

SE116 POR CAUSA DE ALGÚN SINIESTRO O DESASTRE, ¿RECIBIÓ APOYO ECONÓMICO DE ALGÚN PROGRAMA DEL GOBIERNO? *Marque con "x" la respuesta*

SÍ ☐ 1 NO ☐ 2

99.1 ESPECIFIQUE _____

XIV. MEDIO AMBIENTE

PROTECCIÓN DEL MEDIO AMBIENTE

ME111 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA LAS ACTIVIDADES AGROPECUARIAS, REALIZÓ ALGUNA DE LAS SIGUIENTES ACCIONES PARA PROTEGER EL MEDIO AMBIENTE EN SUS TERRENOS, ¿COMO: *Marque con "x" la respuesta*

1 DISMINUIR EL CONSUMO DE ENERGÍA ELÉCTRICA?	SÍ ☐ 1	NO ☐ 2
2 UTILIZAR ENERGÍAS ALTERNATIVAS? (SOLAR, EÓLICA, HÍDRICA)	SÍ ☐ 1	NO ☐ 2
3 DISMINUIR EL CONSUMO DE AGUA?	SÍ ☐ 1	NO ☐ 2
4 TRATAMIENTO DE EXCRETAS O AGUAS RESIDUALES?	SÍ ☐ 1	NO ☐ 2
5 VERIFICAR LOS VEHÍCULOS?	SÍ ☐ 1	NO ☐ 2
6 TENER UN LUGAR PARA GUARDAR LOS EMPAQUES Y ENVASES DE LOS AGROQUÍMICOS O PRODUCTOS BIOLÓGICOS QUE UTILIZÓ?	SÍ ☐ 1	NO ☐ 2
7 PLANTAR O MANTENER CERCOS VIVOS PARA DISMINUIR LA EROSIÓN?	SÍ ☐ 1	NO ☐ 2
8 PRODUCCIÓN DE COMPOSTA CON RESIDUOS ORGÁNICOS?	SÍ ☐ 1	NO ☐ 2
9 LA PREVENCIÓN DE INCENDIOS?	SÍ ☐ 1	NO ☐ 2
10 MONITOREAR PRESENCIA DE PLAGAS Y ENFERMEDADES?	SÍ ☐ 1	NO ☐ 2
11 APLICAR OBRAS DE CONSERVACIÓN Y RESTAURACIÓN DE SUELOS?	SÍ ☐ 1	NO ☐ 2
12 RECIBIR CAPACITACIÓN AMBIENTAL?	SÍ ☐ 1	NO ☐ 2
13 CAPTACIÓN DE AGUA?	SÍ ☐ 1	NO ☐ 2
99 OTRA ACTIVIDAD DIFERENTE A LAS ANTERIORES?	SÍ ☐ 1	NO ☐ 2

99.1 ESPECIFIQUE _____

X. MANO DE OBRA Y REMUNERACIONES

MANO DE OBRA, SUELDOS Y SALARIOS

	TOTAL	¿CUÁNTAS SON MUJERES?
MO111 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA LAS LABORES AGROPECUARIAS, SIN CONTAR AL PRODUCTOR (A) NI A LOS JORNALEROS, ¿CUÁNTOS FAMILIARES PARTICIPARON SIN RECIBIR UN SUELDO O SALARIO?	____	**MO112_01** ____
MO121 ¿CUÁNTOS PRACTICANTES, BECARIOS O PRESTADORES DE SERVICIO PARTICIPARON SIN RECIBIR SUELDO O SALARIO?	____	**MO121_01** ____
MO114 ¿CUÁNTAS PERSONAS SE CONTRATARON POR 6 MESES O MÁS?	____	**MO114_01** ____
MO115 ¿CUÁNTAS PERSONAS SE CONTRATARON POR MENOS DE 6 MESES, SIN CONTAR A LOS JORNALEROS?	____	**MO115_01** ____

xxvi. Entrevistador: Si la suma de MO114 y MO115 es cero, pase a MO118.

JORNALEROS Y PAGO POR JORNALES

	TOTAL	¿CUÁNTAS SON MUJERES?
MO115_02 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA REALIZAR LAS LABORES AGROPECUARIAS, ¿CUÁNTOS JORNALEROS CONTRATÓ?	____	**MO115_03** ____

Si es cero, pase a MO120

1 LOS JORNALEROS PROVENÍAN DE: *Marque con "x" la respuesta*

1.1 ¿LOS ALREDEDORES O ZONAS CERCANAS?	SÍ ☐ 1	NO ☐ 2
1.2 ¿OTRA PARTE DEL MISMO ESTADO?	SÍ ☐ 1	NO ☐ 2
1.3 ¿OTRO ESTADO O ENTIDAD FEDERATIVA?	SÍ ☐ 1	NO ☐ 2
1.4 ¿OTRO PAÍS?	SÍ ☐ 1	NO ☐ 2

99.1 ESPECIFIQUE _____

	TOTAL	¿CUÁNTAS SON MUJERES?
MO120 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA REALIZAR LAS LABORES AGROPECUARIAS, ¿CUÁNTAS PERSONAS TRABAJARON EN LOS TERRENOS O CON SUS ANIMALES Y QUE FUERON PROPORCIONADOS POR OTRA RAZÓN SOCIAL?	____	**MO120_01** ____

IX. TRACTORES, MAQUINARIA Y VEHÍCULOS

xxiv. Entrevistador: Si se registró al menos un producto agrícola objeto de estudio de la Encuesta o existencias de bovinos o porcinos mayores a cero o existencias de aves de corral mayores a 10,000 continúe, si no, pase a Datos de identificación.

TRACTORES

TR110 ENTRE OCTUBRE DEL AÑO PASADO Y SEPTIEMBRE DE ESTE AÑO, PARA LAS ACTIVIDADES AGROPECUARIAS, ¿UTILIZÓ TRACTORES? ..... SÍ ☐ 1 NO ☐ 2

Marque con "x" la respuesta

Pase a nota xxv

TR111 LOS TRACTORES, ¿ERAN: ***Marque con "x" la respuesta***

1 PRESTADOS? ..... SÍ ☐ 1 NO ☐ 2

2 RENTADOS? ..... SÍ ☐ 1 NO ☐ 2

TR115 PROPIOS? ..... SÍ ☐ 1 NO ☐ 2

Pase a nota xxv

Cantidad

TR115_01 EL 30 DE SEPTIEMBRE DE ESTE AÑO, ¿CUÁNTOS TRACTORES PROPIOS TENÍA EN TOTAL? ..... ☐☐☐

VEHÍCULOS

VE111 EL 30 DE SEPTIEMBRE DE ESTE AÑO, PARA LAS ACTIVIDADES AGROPECUARIAS, ¿TENÍA EN PROPIEDAD CAMIONES O CAMIONETAS? ..... SÍ ☐ 1 NO ☐ 2

Marque con "x" la respuesta

Pase a MO111

	Cantidad	**VE112_05** EN PROMEDIO, ¿QUÉ CAPACIDAD DE CARGA TIENEN? ***Toneladas***	**VE119_01** EN PROMEDIO, ¿CUÁNTOS AÑOS DE USO TIENEN? ***Años***
VE112_01 ¿CUÁNTAS CAMIONETAS TENÍA?	☐☐☐	☐☐☐	☐☐☐
VE112_02 ¿CUÁNTOS CAMIONES TENÍA?	☐☐☐	☐☐☐	☐☐☐

MAQUINARIA Y EQUIPO

xxv. Entrevistador: Si los campos llegaran a ser insuficientes debe utilizar el anexo y llenarse con la información faltante según corresponda.

MA111 EL 30 SEPTIEMBRE DE ESTE AÑO, PARA LAS ACTIVIDADES AGROPECUARIAS, ¿TENÍA EN PROPIEDAD TRILLADORAS, MOTOGRÚAS, SEMBRADORAS O ALGUNA OTRA MAQUINARIA O EQUIPO? ..... SÍ ☐ 1 NO ☐ 2

Marque con "x" la respuesta

Pase a VE111

MA114 ¿QUÉ TIPO DE MAQUINARIA O EQUIPO TENÍA? (UNO POR RENGLÓN) ***Tipo***	**MA115** ¿CUÁNTOS AÑOS DE USO TIENE? ***Años***	**MA115_99** ***No sabe***
_____	☐☐☐	☐ 3
_____	☐☐☐	☐ 3
_____	☐☐☐	☐ 3

XV. CARACTERÍSTICAS SOCIODEMOGRÁFICAS DEL PRODUCTOR

xxvii. Entrevistador: Si la respuesta de la pregunta JU111 fue que el productor(a) o su familia maneja los terrenos, continúe, en caso contrario pase a Datos de identificación.

xxviii. Entrevistador: Recuerde que las siguientes preguntas son referentes al Productor.

SD120 NOMBRE DEL PRODUCTOR: _____

Marque con "x" la respuesta

SD113 ¿CUÁL ES SU SEXO? ..... HOMBRE ☐ 1 MUJER ☐ 2

Años

SD114 ¿QUÉ EDAD TIENE? (EN AÑOS CUMPLIDOS) ..... ⌊____⌋

Si es menor a 65 años, pase a SD122

Marque con "x" la respuesta

SD123 ¿RECIBE APOYO ECONÓMICO POR SER ADULTO MAYOR? ..... SÍ ☐ 1 NO ☐ 2

Años

SD122 ¿CUÁNTOS AÑOS TIENE REALIZANDO ACTIVIDADES AGRÍCOLAS, GANADERAS O FORESTALES? ..... ⌊____⌋

Horas

PA112_01 EN PROMEDIO, ¿CUÁNTAS HORAS AL DÍA DEDICA A ESTA ACTIVIDAD? ..... ⌊____⌋

PA114 ¿SE DEDICA A OTRA ACTIVIDAD DIFERENTE A LA AGROPECUARIA O FORESTAL? ..... SÍ ☐ 1 NO ☐ 2

99.1 ESPECIFIQUE _____

SD115_01 ¿HABLA ALGÚN DIALECTO O LENGUA INDÍGENA? ..... SÍ ☐ 1 NO ☐ 2

99.1 ESPECIFIQUE _____

SD115 DE ACUERDO CON SU CULTURA, ¿SE CONSIDERA INDÍGENA? ..... SÍ ☐ 1 NO ☐ 2

Clave ***Grado***

SD118 ¿CUÁL ES SU ÚLTIMO NIVEL DE ESTUDIOS ALCANZADO? ..... ⌊____⌋ ⌊____⌋

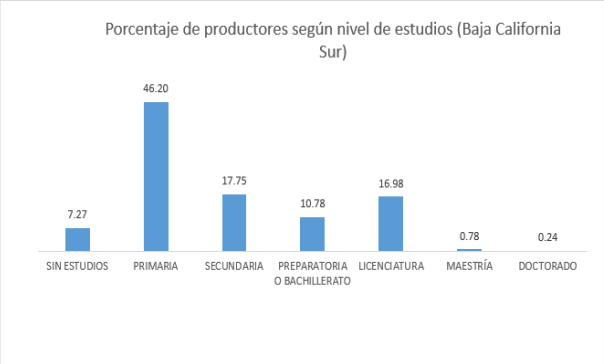
1. PREESCOLAR *2. PRIMARIA* *3. SECUNDARIA* *4. BACHILLERATO O PREPARATORIA*
5. CARRERA TÉCNICA *6. LICENCIATURA O INGENIERÍA* *7. POSGRADO* *8. NO SABE*
99. OTRO *99.1 ESPECIFIQUE* _____

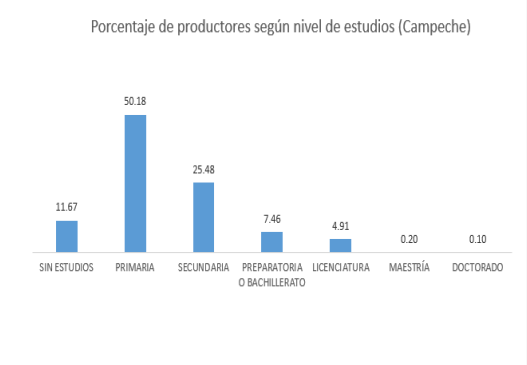
LISTA DE PRODUCTOS AGROPECUARIOS DE INTERÉS PARA LA ENCUESTA NACIONAL AGROPECUARIA 2019

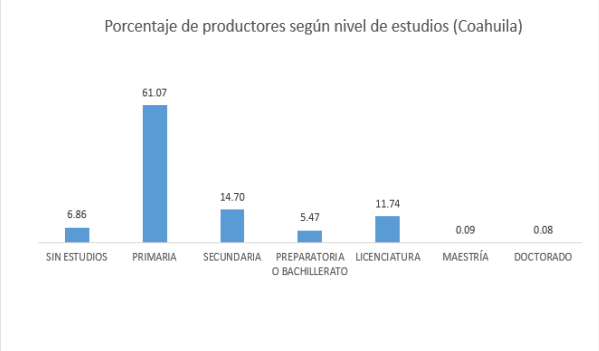
AGUACATE
ALFALFA
AMARANTO
ARROZ
CACAO
CAFÉ
CALABAZA/CALABACITA
CAÑA DE AZÚCAR
CEBOLLA
CHILE
FRESA
FRIJOL
JITOMATE (TOMATE ROJO)
LIMÓN
MAÍZ
MANGO
MANZANA
NARANJA
PLÁTANO
SORGO
SOYA
TRIGO
UVA
BOVINOS
LECHE
PORCINOS
AVES DE CORRAL
HUEVO

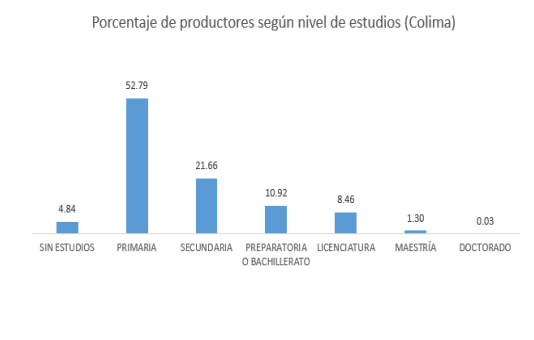
Gráficos porcentaje según nivel de estudios

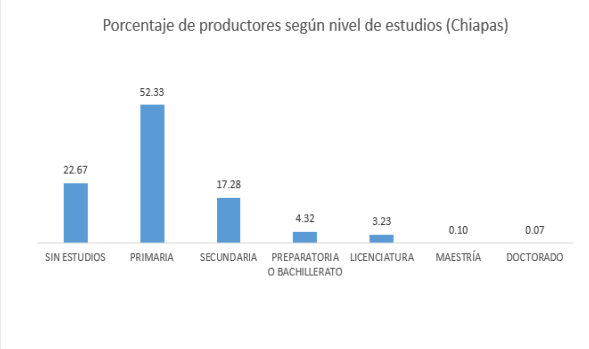
A continuación, se presentan los gráficos de porcentaje de productores según el nivel de estudios y en cada entidad, considerando a éstas como Baja California sur, Campeche, Coahuila, Colima, Chiapas, Chihuahua, Ciudad de México, Jalisco, México, Michoacán, Morelos, Querétaro, Quintana Roo, San Luis Potosí, Sinaloa, Veracruz, Yucatán, Zacatecas.

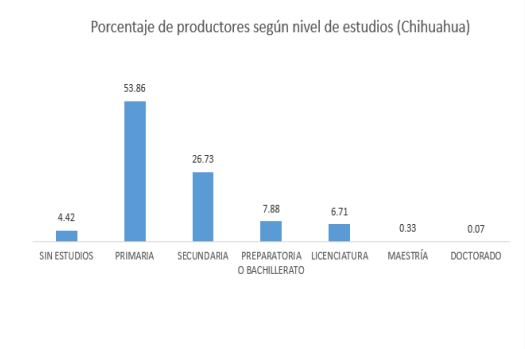


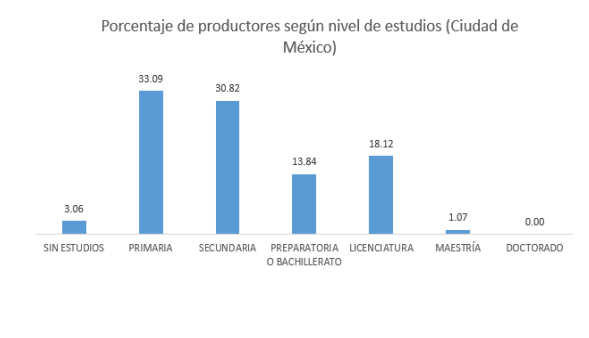


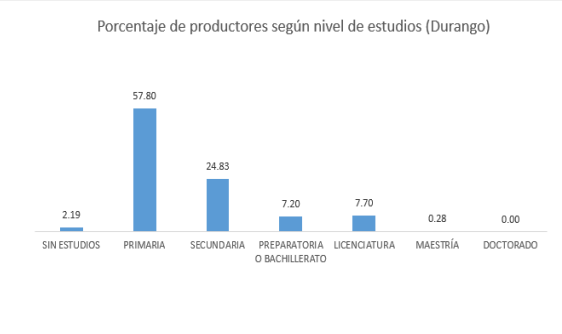


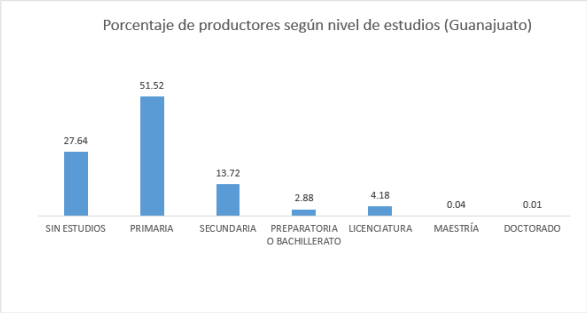


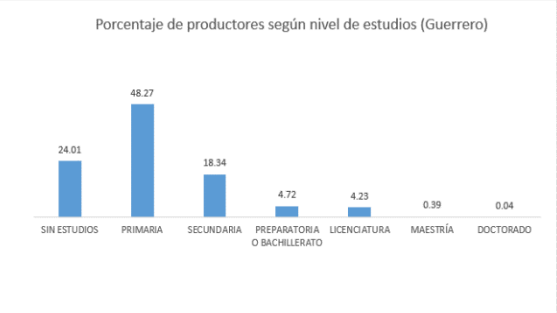


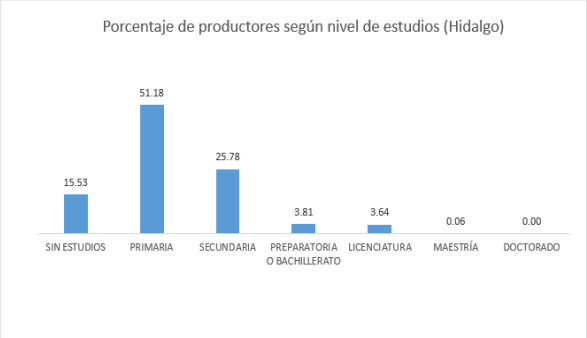


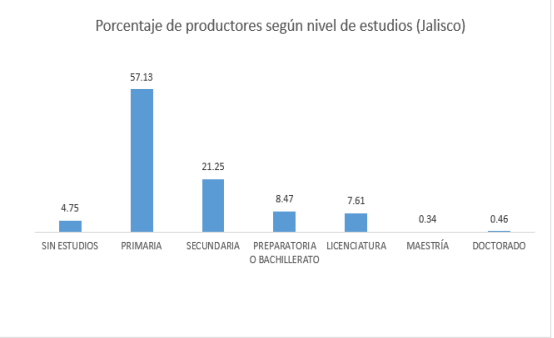

Porcentaje de productores según nivel de estudios (Ciudad de México)
3.06
33.09
30.82
13.84
18.12
1.07
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

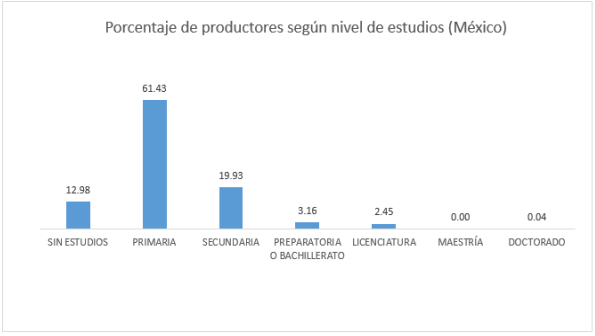

Porcentaje de productores según nivel de estudios (Durango)
2.19
57.80
24.83
7.20
7.70
0.28
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

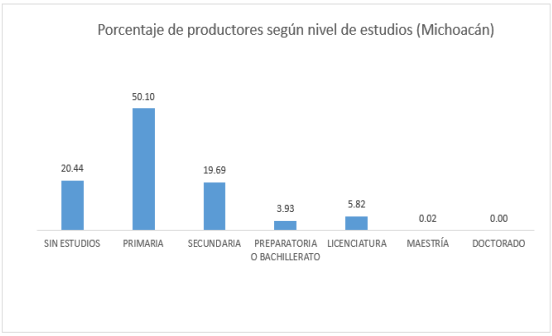

Porcentaje de productores según nivel de estudios (Guanajuato)
27.64
51.52
13.72
2.88
4.18
0.04
0.01
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

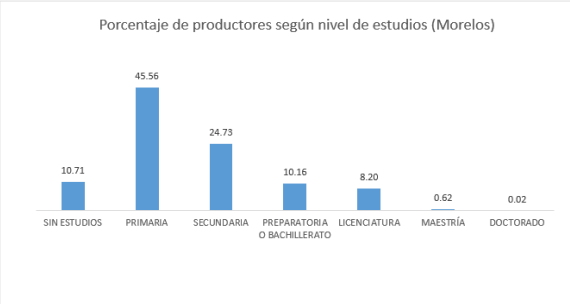

Porcentaje de productores según nivel de estudios (Guerrero)
24.01
48.27
18.34
4.72
4.23
0.39
0.04
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

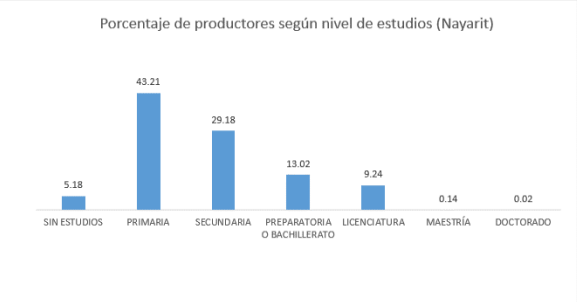

Porcentaje de productores según nivel de estudios (Hidalgo)
15.53
51.18
25.78
3.81
3.64
0.06
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

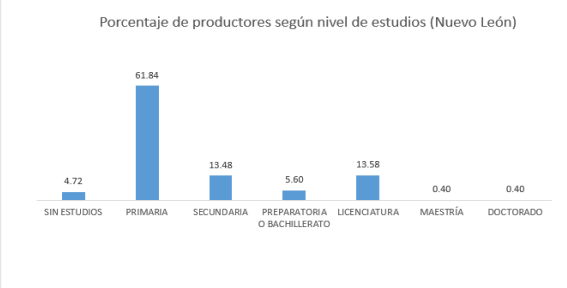

Porcentaje de productores según nivel de estudios (Jalisco)
4.75
57.13
21.25
8.47
7.61
0.34
0.46
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

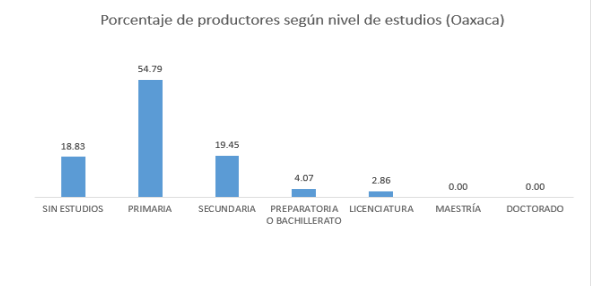

Porcentaje de productores según nivel de estudios (México)
12.98
61.43
19.93
3.16
2.45
0.00
0.04
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

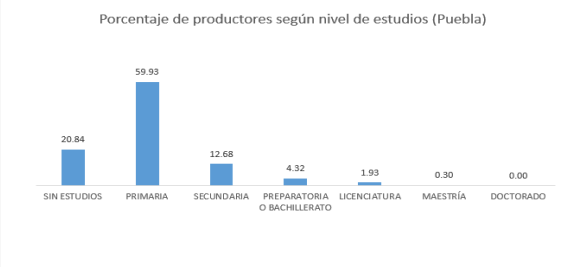

Porcentaje de productores según nivel de estudios (Michoacán)
20.44
50.10
19.69
3.93
5.82
0.02
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO


Porcentaje de productores según nivel de estudios (Morelos)
10.71
45.56
24.73
10.16
8.20
0.62
0.02
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

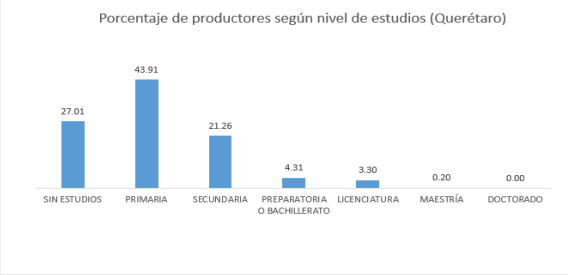

Porcentaje de productores según nivel de estudios (Nayarit)
5.18
43.21
29.18
13.02
9.24
0.14
0.02
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

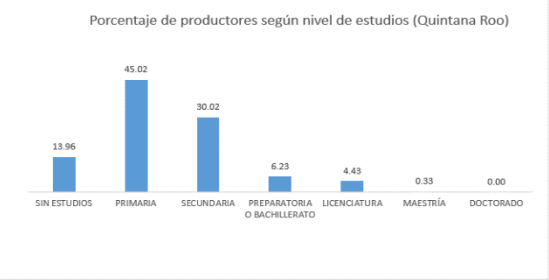

Porcentaje de productores según nivel de estudios (Nuevo León)
4.72
61.84
13.48
5.60
13.58
0.40
0.40
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

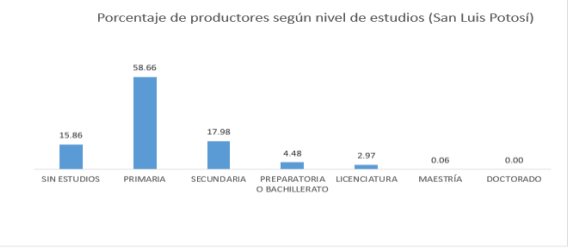

Porcentaje de productores según nivel de estudios (Oaxaca)
18.83
54.79
19.45
4.07
2.86
0.00
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

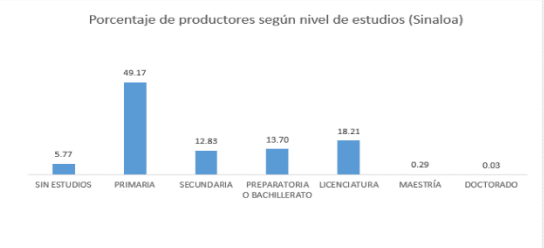

Porcentaje de productores según nivel de estudios (Puebla)
20.84
59.93
12.68
4.32
1.93
0.30
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

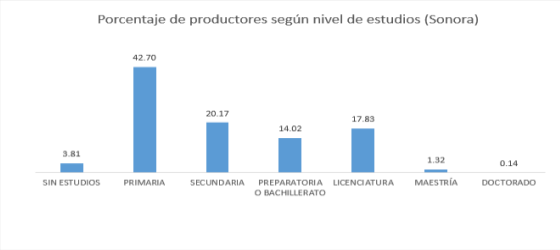

Porcentaje de productores según nivel de estudios (Oaxaca)
18.83
54.79
19.45
4.07
2.86
0.00
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

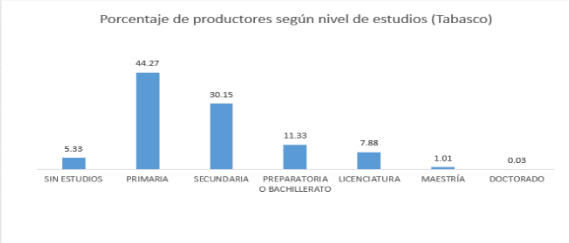

Porcentaje de productores según nivel de estudios (Querétaro)
27.01
43.91
21.26
4.31
3.30
0.20
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

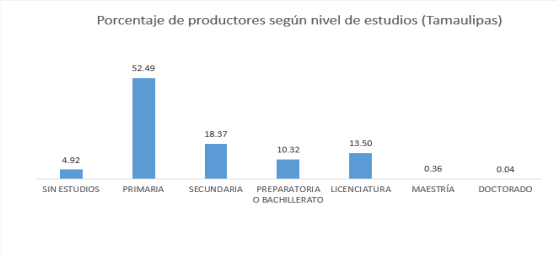

Porcentaje de productores según nivel de estudios (Quintana Roo)
13.96
45.02
30.02
6.23
4.43
0.33
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

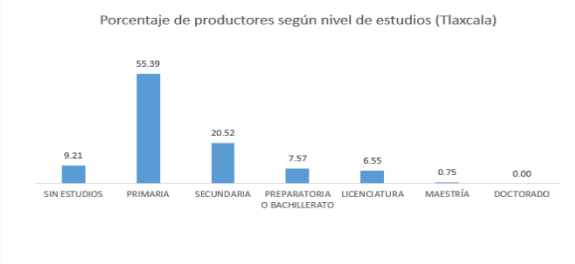

Porcentaje de productores según nivel de estudios (San Luis Potosí)
15.86
58.66
17.98
4.48
2.97
0.06
0.00
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

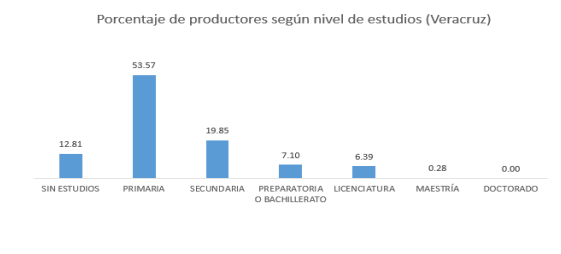

Porcentaje de productores según nivel de estudios (Sinaloa)
5.77
49.17
12.83
13.70
18.21
0.29
0.03
SIN ESTUDIOS
PRIMARIA
SECUNDARIA
PREPARATORIA O BACHILLERATO
LICENCIATURA
MAESTRÍA
DOCTORADO

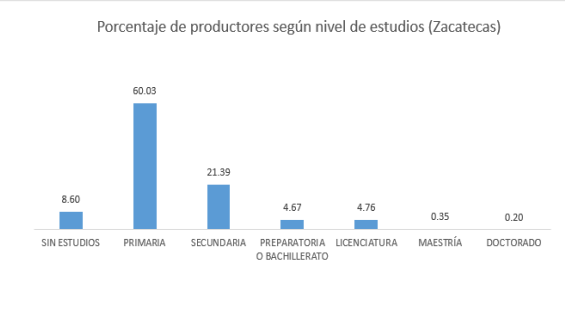






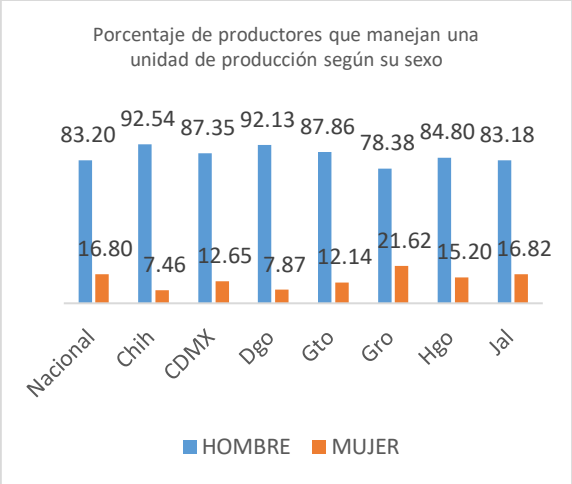


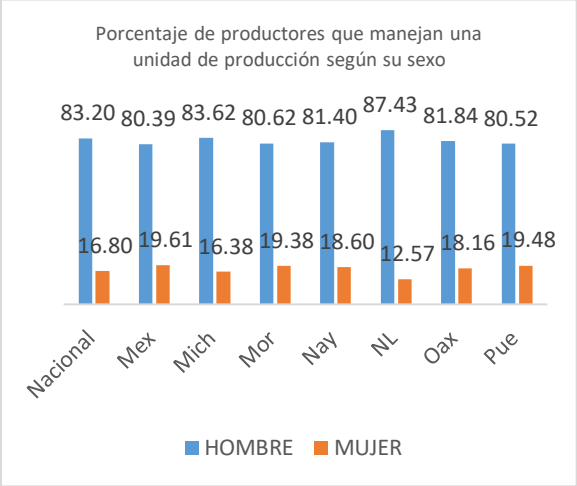


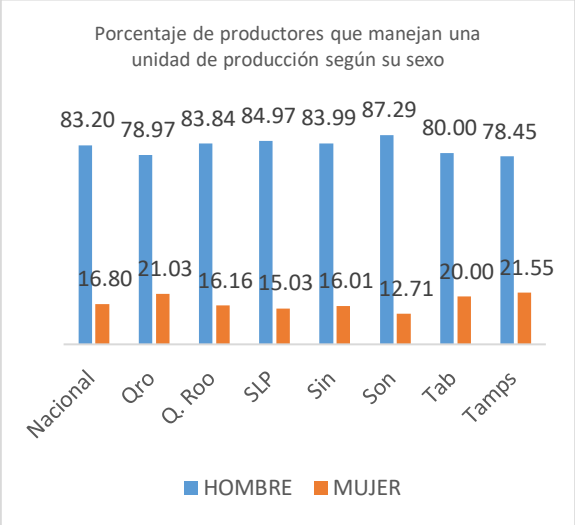


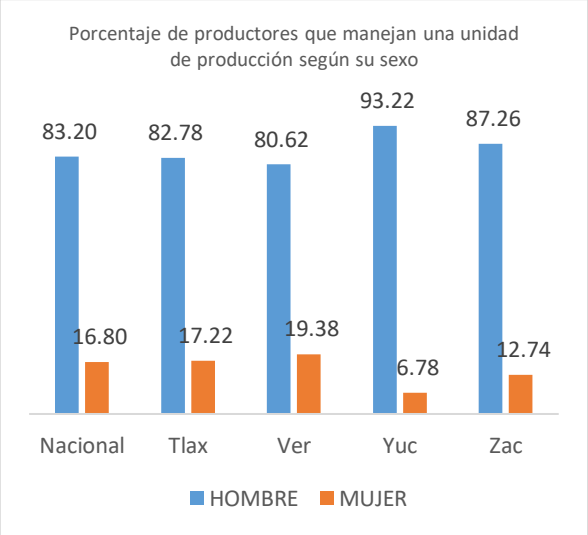
Gráficos porcentaje de productores según el sexo

A continuación, se muestran algunos gráficos que ilustran el porcentaje de productores que manejan una unidad de producción agropecuaria según el sexo (Hombres o Mujeres) y en distintas entidades, comparándolo con la nacional.





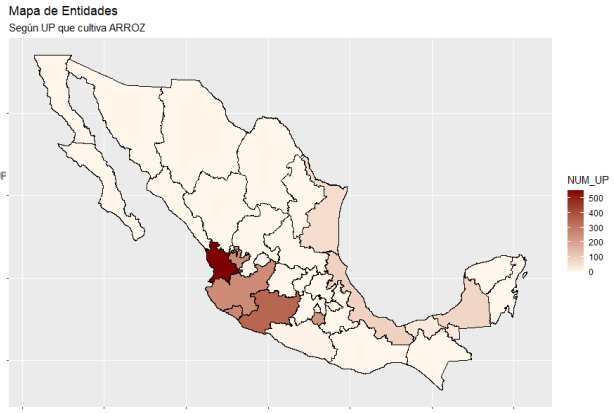




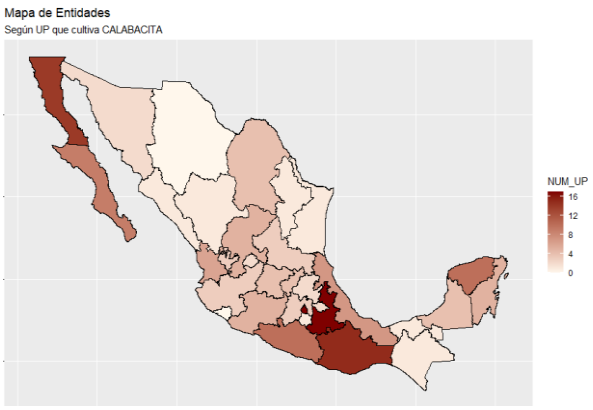
Mapas de entidades según el producto cultivado

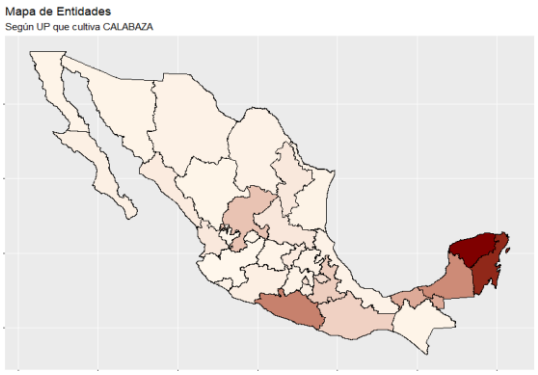
En los siguientes mapas, se observa con un color rojo fuerte los estados con mayor cantidad de unidades de producción agrícolas que producen cierto cultivo como: amaranto, arroz, cacao, calabacita, calabaza, caña de azúcar, cebolla, chile, fresa, frijol, jitomate (tomate rojo), limón, maíz, mango, manzana, naranja, plátano, sorgo, soya, trigo y uva, mientras que los estados que no tienen unidades de producción que produzcan dichos cultivos o éstas sean mínimas, se muestran en un color rojo débil, parecido al blanco.

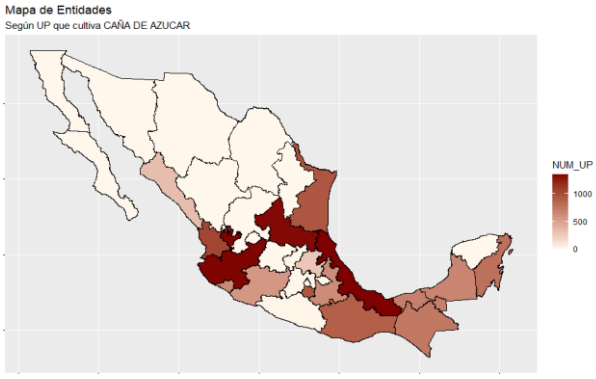

Mapa de Entidades
Según UP que cultiva AMARANTO
NUM_UP
90
60
30
0

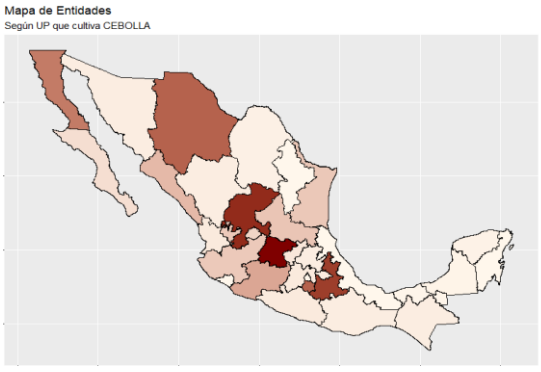

Mapa de Entidades
Según UP que cultiva ARROZ
NUM_UP
500
400
300
200
100
0

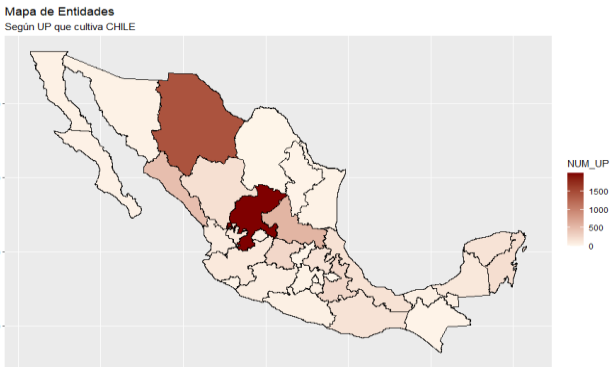

Mapa de Entidades
Según UP que cultiva CACAO
NUM_UP
9000
6000
3000
0

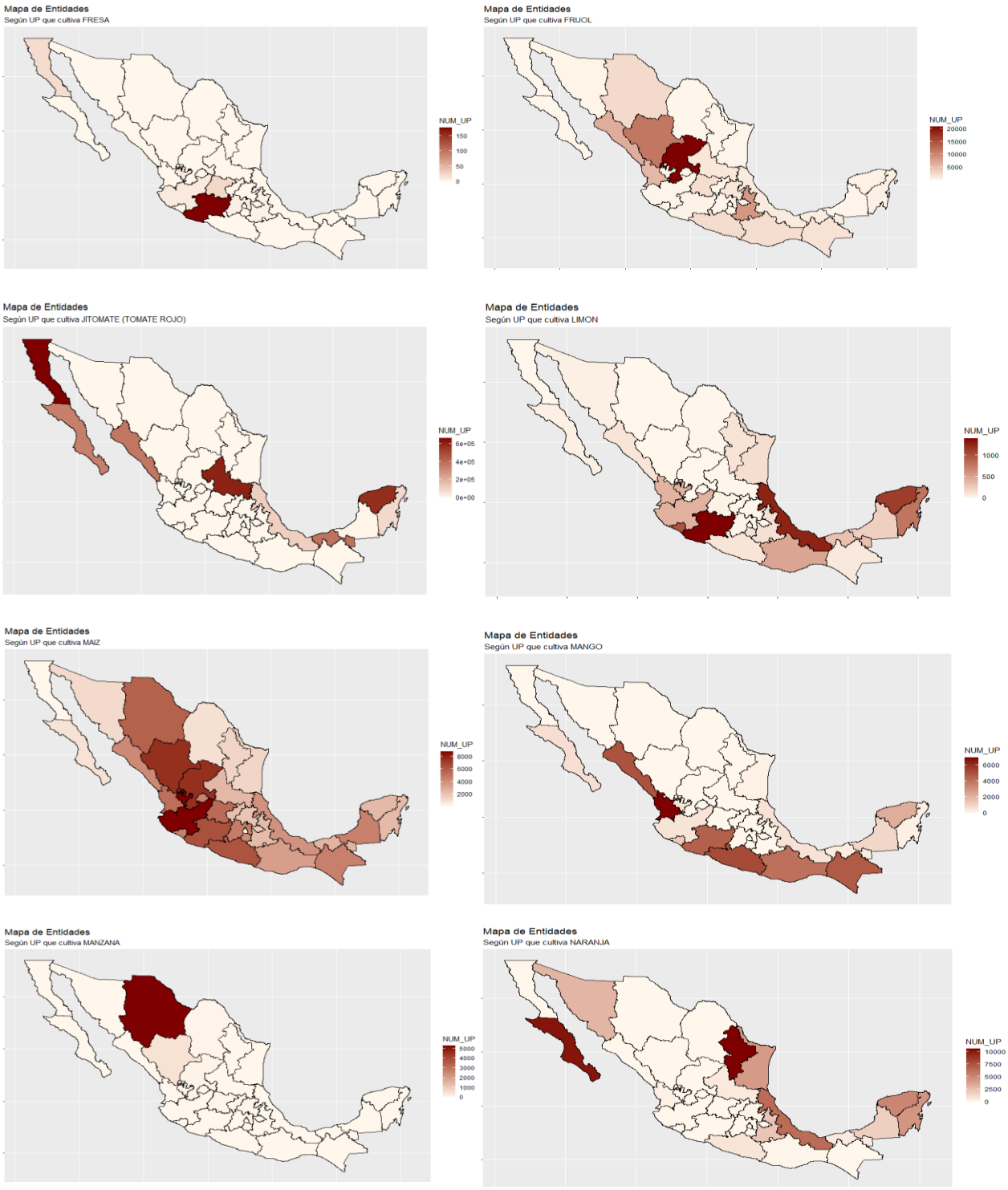

Mapa de Entidades
Según UP que cultiva CALABACITA
NUM_UP
16
12
8
4
0

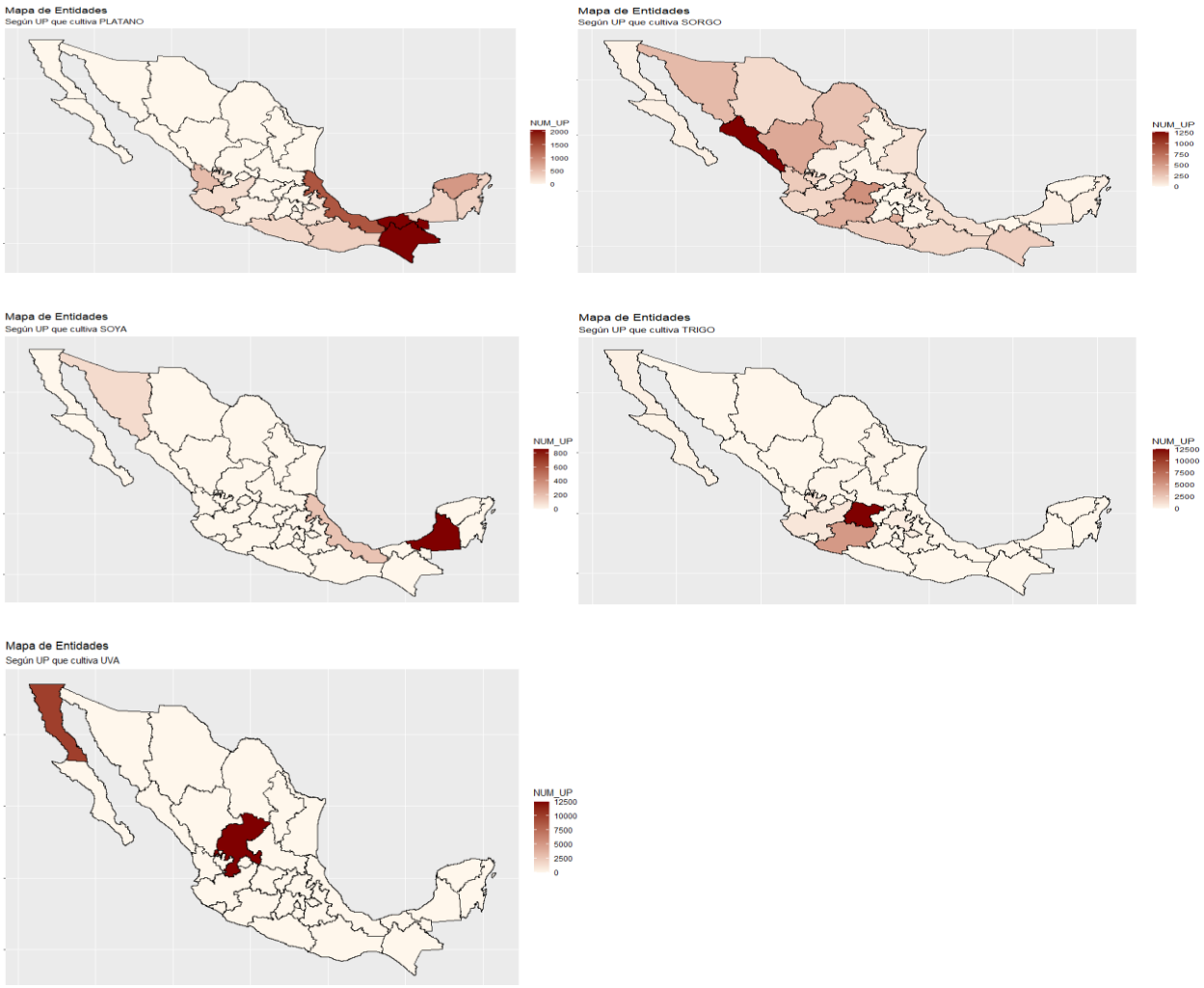

Mapa de Entidades
Según UP que cultiva CALABAZA
NUM_UP
500
400
300
200
100
0


Mapa de Entidades
Según UP que cultiva CAÑA DE AZUCAR
NUM_UP
1000
500
0


Mapa de Entidades
Según UP que cultiva CEBOLLA
NUM_U
100
50
0


Mapa de Entidades
Según UP que cultiva CHILE
NUM_UP
1500
1000
500
0


Mapa de Entidades
Según UP que cultiva FRESA
NUM_UP
150
100
50
0
Mapa de Entidades
Según UP que cultiva FRIJOL
NUM_UP
20000
15000
10000
5000
Mapa de Entidades
Según UP que cultiva JITOMATE (TOMATE ROJO)
NUM_UP
6e+05
4e+05
2e+05
0e+00
Mapa de Entidades
Según UP que cultiva LIMON
NUM_UP
1000
500
0
Mapa de Entidades
Según UP que cultiva MAIZ
NUM_UP
8000
6000
4000
2000
Mapa de Entidades
Según UP que cultiva MANGO
NUM_UP
6000
4000
2000
0
Mapa de Entidades
Según UP que cultiva MANZANA
NUM_UP
5000
4000
3000
2000
1000
0
Mapa de Entidades
Según UP que cultiva NARANJA
NUM_UP
10000
7500
5000
2500
0



Anexo B

Código de R utilizado para Análisis de Factores

```
library("MASS")

#library("alr3")

library("stats")

library("scatterplot3d")

library("psych") #prueba de esfericidad de bartlett
```

```
#-----------------------------------------------------------------
--

#CARGAR RUTA

setwd("D:/andres.lira/Desktop/RR/13
Maestria_Análisis_Estadístico_y_Computo/TRABAJO FINAL/AF")

# datos expandidos

bd<-read.csv("BASE.csv",header=TRUE)

# trabajar con la información de las 25 variables

var_up<-cbind(bd[,4:5],bd[,12:27])

names(var_up)

#resto_var_up<-cbind(bd[,1:3],bd[,6:11])

# revisar la correlación de las variables

var_up_corr <-cor(var_up)

det_v_up<-det(var_up_corr)

det_v_up

#Prueba de esfericidad de bartlett para probar la hipotesis nula
de que las variables

#no estan correlacionadas. La idea es rechazar la hipotesis nula
para proseguir con un

#analisis de factores

#names(bd_sin_factores)

cortest.bartlett(var_up)

# Se aplica la prueba de Kaiser-Meyer-Olkin (KMO)

KMO(var_up)

#Se realiza el analisis de factores utilizando componentes
principales para estimar las cargas

up_pc<-princomp(var_up, cor=TRUE)

summary(up_pc)

screeplot(up_pc)
```

```
scree(var_up)

# por lo tanto elegimos un modelo con m=5 factores como máximo

#se obtienen las cargas (matriz L)

cargas_up<-up_pc$loadings

#elegimos las primeras 5 columnas de la matriz de carga(L),
correspondiente a m=5

cargas_5up<-cargas_up[,1:5]

cargas_5up

#el siguiente paso seria interpretar los factores en terminos de
sus cargas

#por ultimo calculamos los factor scores

up_pc_scores<-princomp(var_up, cor=TRUE, score=TRUE)

up_pc_scores

scores_up<-up_pc_scores$scores

scores_up_5comp<-scores_up[,1:5]

scores_up_5comp

#up_pc_scores$sdev

#-----------------------------------------------------------------
--

#  Se realiza un analisis de factores utilizando maxima
verosimilitud para estimar los parametros del modelo

#(las cargas y las varianzas especificas).

#Por default el analisis de factores se realiza sobre los datos
estandarizados z y utilizando la rotación varimax

#se prueba la solucion con un factor (m=1)

up_fa1<-  factanal(var_up,factors=1)  #funcion  que  realiza  el
analisis de factores

CARGAS1<-up_fa1$loadings #cargas estimadas

VAR_ESP1<-up_fa1$uniquenesses #varianzas especificas estimadas
```

```
prueba_hipo1<-up_fa1$PVAL #prueba de hipotesis para determinar si
un factor es adecuado



comun1<-1-VAR_ESP1

#se prueba la solucion con un factor (m=2)

up_fa2<-  factanal(var_up,factors=2)  #funcion  que  realiza  el
analisis de factores

CARGAS2<-up_fa2$loadings #cargas estimadas

VAR_ESP2<-up_fa2$uniquenesses #varianzas especificas estimadas

prueba_hipo2<-up_fa2$PVAL #prueba de hipotesis para determinar si
un factor es adecuado

comun2<-1-VAR_ESP2

#se prueba la solucion con un factor (m=3)

up_fa3<-  factanal(var_up,factors=3)  #funcion  que  realiza  el
analisis de factores

CARGAS3<-up_fa3$loadings #cargas estimadas

VAR_ESP3<-up_fa3$uniquenesses #varianzas especificas estimadas

prueba_hipo3<-up_fa3$PVAL #prueba de hipotesis para determinar si
un factor es adecuado

comun3<-1-VAR_ESP3

#se prueba la solucion con un factor (m=4)

up_fa4<-  factanal(var_up,factors=4)  #funcion  que  realiza  el
analisis de factores

CARGAS4<-up_fa4$loadings #cargas estimadas

VAR_ESP4<-up_fa4$uniquenesses #varianzas especificas estimadas

prueba_hipo4<-up_fa4$PVAL #prueba de hipotesis para determinar si
un factor es adecuado

comun4<-1-VAR_ESP4

up_fa5<- factanal(var_up,factors=5) #ya no se puede optimizar a
partir de este valor
```

```
#Se calcula la diferencia entre las correlaciones observadas y las
predichas para m=3 factores

#Primero se obtiene la estimacion de la matriz de correlaciones
(matriz reproducida)

pred3_vc<- CARGAS3%*%t(CARGAS3)+diag(VAR_ESP3)

#se calcula la diferencia entre la matriz de correlacion observada
y la matriz reproducida

round(var_up_corr-pred3_vc,digits=3)

#Se calcula la diferencia entre las correlaciones observadas y las
predichas para m=4 factores

#Primero se obtiene la estimacion de la matriz de correlaciones
(matriz reproducida)

pred4_vc <-CARGAS4%*%t(CARGAS4)+diag(VAR_ESP4)

#se calcula la diferencia entre la matriz de correlacion observada
y la matriz reproducida

round(var_up_corr-pred4_vc,digits=3)

# por lo anterior el modelo de 4 factores produce una matriz
residual cercana a cero. Por tanto

#un modelo con 4 factores puede ser suficiente para explicar la
variabilidad de los datos originales

#Interpretación de los factores de acuerdo a los valores de sus
cargas

#----------------------------------------------------------------
--

#  Se realiza un analisis de factores utilizando maxima
verosimilitud para estimar los parametros del modelo

#(las cargas y las varianzas especificas).

#Por default el analisis de factores se realiza sobre los datos
estandarizados z y utilizando la rotación varimax

#modelo

fit.ml<-
fa(var_up_corr,nfactors=5,fm="ml",rotate="varimax",n.obs=56108)
#otra función que calcula fa, esta en la libreria "psych"
```

```
# error cuadrático medio

fit.ml$rms

# error cuadrático medio ajustado según los grados de libertad.

fit.ml$crms

# ¿Qué tan bien reproduce el modelo factorial la matriz de
correlación?

fit.ml$fit

# ¿Qué tan bien se reproducen los elementos fuera de la diagonal?

fit.ml$fit.off

# Si n.obs > 0, ¿cuál es la probabilidad de observar un chi-
cuadrado tan grande o más grande?

fit.ml$PVAL

# El índice Tucker Lewis de confiabilidad de factorización,
también conocido como índice de ajuste no normado

fit.ml$TLI

# Las correlaciones de las estimaciones de la puntuación factorial
utilizando el modelo especificado, si se encontraran. La
comparación de estas correlaciones con las de las puntuaciones en
sí mostrará, si se utiliza una estimación alternativa de las
puntuaciones de los factores (por ejemplo, el método de tenBerge),
el problema de la indeterminación de los factores. Porque estas
correlaciones no serán necesariamente las mismas.

fit.ml$r.scores

# una rotación varimax

fit.ml$rotation

#Estimaciones de comunalidad para cada elemento. Estos son
simplemente la suma de las cargas factoriales al cuadrado para ese
elemento.

fit.ml$communality

# Si se usa el análisis factorial minrank, estas son las
comunalidades que reflejan la cantidad total de varianza común.
Excederán la comunalidad (arriba) que es la varianza común
estimada del modelo
```

```
fit.ml$communalities

#Valores propios de la solución del factor común

fit.ml$values

# Valores propios de la matriz original

fit.ml$e.values

# Una matriz de carga de elementos por factor (patrón) de la clase "cargas" Adecuada para usar en otros programas (p. ej., rotación de GPA o factor2cluster). Para mostrarlos por orden, use print.psych con sort=TRUE

fit.ml$loadings

fit.ml$model
```

Código de R utilizado para Escalamiento Multidimensional

```
library("MASS")

#library("alr3")

library("stats") #esta libreria incluye la funcion que realiza el MDS clásico

#library("scatterplot3d")

library("smacof") #Libreria incluye las funciones que realizan el MDS de minimos cuadrados usando SMACOF

library("ggplot2")

#CARGAR RUTA

setwd("D:/andres.lira/Desktop/RR/13 Maestria_Análisis_Estadístico_y_Computo/TRABAJO FINAL/MDS")

bd_prom<-read.csv("BASE_PROMEDIO.csv",header=TRUE)

names(bd_prom)

# se crea una matriz de distancias

m_distancia<-dist(bd_prom[,2:9],method = "euclidian", diag=TRUE, upper=TRUE)

#class(m_distancia)

m2_distancia<-as.matrix(m_distancia)
```

```
class(m2_distancia)

cmd_dist_entidad<-cmdscale(dist_entidad)

cmd_dist_entidad<-as.data.frame(cmd_dist_entidad)

ggplot(cmd_dist_entidad,aes(x=V1,

                      y=V2,

                      label=rownames(bd_prom)))+

  geom_text(alpha=0.8,size=3,col="salmon")

n<-32

D2<-m2_distancia**2

unos<-matrix(1,n,1)

id<-diag(1,n)

H<-id-((1/n)*(unos%*%t(unos)))

B<-(-1/2)*(H%*%D2%*%H)

eigen<-eigen(B)

#1ERA DIMENSIÓN

lambda1<-sqrt(eigen$values[1])

eigv1<-eigen$vectors[,1]

y1<-lambda1*eigv1

#2DA DIMENSIÓN

lambda2<-sqrt(eigen$values[2])

eigv2<-eigen$vectors[,2]

y2<-lambda2*eigv2

label<-t(bd_prom[,1])

plot(y1,y2,main="Escalamiento Multidimensional De forma manual")

text(y1,y2,labels=label,pos=4,cex=0.5)

# %de ajuste

print("La propoción de la varianza explicada por
```

```
           las dos dimensiones es:" )
print((eigen$values[1]+eigen$values[2])/sum(abs(eigen$values)))
print("La propocin de la varianza explicada por
           la  1er dimensión es:")
PCN1<-(eigen$values[1])/sum(abs(eigen$values))
print(PCN1)
print("La propocin de la varianza explicada por
           la 2da dimensión es:")
PCN2<-(eigen$values[2])/sum(abs(eigen$values))
print(PCN2)


# Se realiza el  análisis con la funcion cmdsscale
resulmdsclas<-cmdscale(m_distancia, k=2, eig=TRUE,
                                        add=FALSE,x.ret= FALSE)
Y<-resulmdsclas$points
#se grafica la configuracion solucion obtenida mediante mds
clásico
label<-t(bd_prom[,1])
plot(Y[,1],Y[,2],   lwd=2,   main="Caracterizacion   de   entidades
federativas \n mediante MDS", col ="blue",
         cex =0.4, xlab="Dimensión 1", ylab="Dimensión 2")
text(Y[,1],Y[,2], labels=label , pos = 4 , cex =0.6, col="blue")
abline(h=0, col="forestgreen", lty=2)
abline(v=0,col="forestgreen", lty=2)
# Se obtiene la proporción de la varianza total explicada
#por las dos dimensiones
print("La propoción de la varianza explicada por las
           dos dimensiones es:")
```

```
print(resulmdsclas$GOF)


plot(eigen$values,main="Gráfico de eigenvalores")

lines(eigen$values)

R<-cor(cbind(bd_prom[,2:9],y1,y2))

colnames(R)

col<-colorRampPalette(c("blue","white","darkgreen"))(20)

heatmap(R,col=col,symm=TRUE)
```

Código de R utilizado para Análisis de Correspondencia

```
# ----ejercicio para entidad-cultivo

# Se carga la tabla de contingencia

tabla<-t(tabla)

# se le da nombre a las filas y a las columnas

row.names(tabla)                                                <-
c("Ags","BC","BCS","Camp","Coah","Col","Chis","Chih","CDMX","Dgo",
"Gto","Gro","Hgo","Jal","Mex","Mich","Mor","Nay","NL","Oax","Pue",
"Qro","Q.
Roo","SLP","Sin","Son","Tab","Tamps","Tlax","Ver","Yuc","Zac")

colnames(tabla)<-
c("AGUACATE","ALFALFA","AMARANTO","ARROZ","CACAO","CAFE","CALABACI
TA","CALABAZA","CAÑA                                            DE
AZUCAR","CEBOLLA","CHILE","FRESA","FRIJOL","JITOMATE","LIMON","MAI
Z","MANGO","MANZANA","NARANJA","PLATANO","SORGO","SOYA","TOMATE
ROJO","TRIGO","UVA")

chiR <- chisq.test(tabla); chiR

n<-sum(tabla)

# se calcula la matriz F de correspondencia

renglones<-nrow(tabla);        columnas<-ncol(tabla);    Matriz_F<-
(tabla)/n

#calcular frecuencias relativas de filas y de columnas de F o
```

```
#tambien llamadas masas (las ri y las cj)

rtot<-apply(Matriz_F,1,sum);  ctot<-apply(Matriz_F,2,sum)

# vectores obtenidos anteriormente se colocan en matrices
diagonales

Dr<-diag(rtot);  Dc<-diag(ctot)

# calcular matriz de perfiles por fila (R)

tabla_R<-Matriz_F/rtot  #matriz R de renglones

# calcular matriz de perfiles por columna (Rc)

tabla_C<- t(t(Matriz_F)/ctot)  #matriz R de columnas

# calcular chi cuadrada para probar la independencia de renglones
y

#columnas de la tabla de contingencia y la inercia total (medida
de la

#variabilidad total de los datos en la tabla)

tabla_suma_R<-n*rtot;  tabla_suma_C<-n*ctot

tabla_esperadas<-tabla_suma_R%o%tabla_suma_C/n      #   matriz   de
frecuencias esperadas

chi2<- sum((tabla-tabla_esperadas)^2/tabla_esperadas); chi2

inetot<-chi2/n; inetot

# calcular la matriz Z

Z<-(sqrt(solve(Dr)))%*%Matriz_F%*%(sqrt(solve(Dc)))

#obtener vectores propios y valores propios de Z

dvalsing<-svd(Z); dvalsing$d

###################### considerar 2 dimensiones

#Se obtienen las representaciones de las filas y columnas en un
espacio de

#dos dimensiones considerando los vectores propios ai y bi
obtenidos de la dvs

ind<-c(2,3)

#Representacion de las filas en dos dimensiones
```

```
Cr<-(sqrt(solve(Dr)))%*%Z%*%dvalsing$v[,ind]; Cr

#head(dvalsing$v)

#Representacion de las columnas en dos dimensiones

Cc<-(sqrt(solve(Dc)))%*%t(Z)%*%dvalsing$u[,ind]; Cc

#calcular proporción de la inercia explicada por las dos

#dimensiones (asociadas a valores propios mas grandes distintos de
uno)

vp<-(dvalsing$d)^2

vp_dist1<-vp[-1]; vp_dist1

# Inercia explicada relativa

iner_expl_rel <- vp_dist1/sum(vp_dist1); iner_expl_rel

# Inercia explicada acumulada

iner_expl_acum <- iner_expl_rel

for(i in 1:length(iner_expl_rel))

  iner_expl_acum[i] <- sum(vp_dist1[1:i])/sum(vp_dist1)

iner_expl_acum

# Gráfica

# Inercia explicada relativa

plot(factor(1:length(iner_expl_rel)),iner_expl_rel,lwd=0.5,col="wh
ite",pch=20,cex=0,

     main="Proporción relativa de \nInercia explicada \nentidad-
cultivo",type="n",lty=0,


ylim=c(0,max(iner_expl_acum)),xlab="Dimensión",ylab="Proporción  de
la Inercia Total",bty="n",

     cex.main=1,cex.axis=.7,cex.lab=.7,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,col="grey70",lty=3)

abline(h=iner_expl_rel,v=1:length(iner_expl_rel),col="grey90",lty=
3)
```

```
lines(iner_expl_rel,col="blue",pch=20)

points(iner_expl_rel,col="blue",pch=20)

# Inercia explicada acumulada

plot(factor(1:length(iner_expl_acum)),iner_expl_acum,lwd=0.5,col="
white",pch=20,cex=0,

     main="Proporción acumulada de \nInercia explicada \nentidad-
cultivo",type="n",lty=0,


ylim=c(0,max(iner_expl_acum)),xlab="Dimensión",ylab="Proporción  de
la Inercia Total",bty="n",

     cex.main=1,cex.axis=.7,cex.lab=.7,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,col="grey70",lty=3)

abline(h=iner_expl_acum,v=1:length(iner_expl_acum),col="grey90",lt
y=3)

lines(iner_expl_acum,col="blue",pch=20)

points(iner_expl_acum,col="blue",pch=20)

# Graficamos la representación conjunta de los renglones y
columnas en el

# espacio de dos dimensiones

# Con estas dos dimensiones solo se explica el 38.64% de la
inercia total.

#iner_expl_rel[1]*100 + iner_expl_rel[2]*100

plot(Cr[,1],Cr[,2],xlim=range(Cr[,1],Cc[,1],Cr[,1]+0.5),ylim=range
(Cr[,2],Cc[,2]),

     main="Representación de la tabla \nde contingencia por
Entidad",

     col="white",pch=20,type="p",lwd=1,

     xlab=paste("Dimensión                                    1
(",round(iner_expl_rel[1]*100,2),"%)",sep=""),

     ylab=paste("Dimensión                                    2
(",round(iner_expl_rel[2]*100,2),"%)",sep=""),
```

```
bty="n",cex=1,cex.main=1,cex.axis=.7,cex.lab=.8,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,v=0,lty=3,col="grey90")

points(Cr[,1],Cr[,2],col="blue",pch=20,cex=1)

text(Cr[,1],Cr[,2],labels=c("Ags","BC","BCS","Camp","Coah","Col","
Chis","Chih","CDMX","Dgo","Gto","Gro","Hgo","Jal","Mex","Mich","Mo
r","Nay","NL","Oax","Pue","Qro","Q.
Roo","SLP","Sin","Son","Tab","Tamps","Tlax","Ver","Yuc","Zac"),

     pos=4,cex=0.5,col="darkblue")

plot(Cr[,1],Cr[,2],xlim=range(Cr[,1],Cc[,1],Cr[,1]+0.5),ylim=range
(Cr[,2],Cc[,2]),

     main="Representación de la tabla \nde contingencia por
Cultivos de la ENA2019",

     col="white",pch=20,type="p",lwd=1,

     xlab=paste("Dimensión                                    1
(",round(iner_expl_rel[1]*100,2),"%)",sep=""),

     ylab=paste("Dimensión                                    2
(",round(iner_expl_rel[2]*100,2),"%)",sep=""),


bty="n",cex=1,cex.main=1,cex.axis=.7,cex.lab=.8,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,v=0,lty=3,col="grey90")

points(Cc[,1],Cc[,2],col="red",pch=20,cex=1)

text(Cc[,1],Cc[,2],labels=names(tabla)[-
1],pos=4,cex=0.5,col="darkred")

text(Cc[,1],Cc[,2],labels=c("AGUACATE","ALFALFA","AMARANTO","ARROZ
","CACAO","CAFE","CALABACITA","CALABAZA","CAÑA                    DE
AZUCAR","CEBOLLA","CHILE","FRESA","FRIJOL","JITOMATE","LIMON","MAI
Z","MANGO","MANZANA","NARANJA","PLATANO","SORGO","SOYA","TOMATE
ROJO","TRIGO","UVA"),

     pos=1,cex=0.45,col="darkred")

plot(Cr[,1],Cr[,2],xlim=range(Cr[,1],Cc[,1],Cr[,1]+0.5),ylim=range
(Cr[,2]-0.5,Cc[,2]),
```

```
     main="Representación conjunta de la tabla \nde contingencia
Entidad-Cultivo de la ENA2019",

     col="white",pch=20,type="p",lwd=1,

     xlab=paste("Dimensión                                    1
(",round(iner_expl_rel[1]*100,2),"%)",sep=""),

     ylab=paste("Dimensión                                    2
(",round(iner_expl_rel[2]*100,2),"%)",sep=""),


bty="n",cex=1,cex.main=1,cex.axis=.7,cex.lab=.8,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,v=0,lty=3,col="grey90")

points(Cr[,1],Cr[,2],col="blue",pch=20,cex=1)

text(Cr[,1],Cr[,2],labels=c("Ags","BC","BCS","Camp","Coah","Col","
Chis","Chih","CDMX","Dgo","Gto","Gro","Hgo","Jal","Mex","Mich","Mo
r","Nay","NL","Oax","Pue","Qro","Q.
Roo","SLP","Sin","Son","Tab","Tamps","Tlax","Ver","Yuc","Zac"),

     pos=4,cex=0.5,col="darkblue")

points(Cc[,1],Cc[,2],col="red",pch=20,cex=1)

text(Cc[,1],Cc[,2],labels=names(tabla)[-
1],pos=4,cex=0.5,col="darkred")

text(Cc[,1],Cc[,2],labels=c("AGUACATE","ALFALFA","AMARANTO","ARROZ
","CACAO","CAFE","CALABACITA","CALABAZA","CAÑA                    DE
AZUCAR","CEBOLLA","CHILE","FRESA","FRIJOL","JITOMATE","LIMON","MAI
Z","MANGO","MANZANA","NARANJA","PLATANO","SORGO","SOYA","TOMATE
ROJO","TRIGO","UVA"),

     pos=1,cex=0.45,col="darkred")

# A continuación, se calcula las distancias chi cuadrada

# de cada renglón de R a su centroide: (R-c)'Dc^-1(R-c)

# y calculando la inercia de cada entidad

dist.chicua.ren   <-   sqrt(apply((t(tabla_R)-ctot)^2/ctot,2,sum));
dist.chicua.ren

inercia.ren <- ((dist.chicua.ren)^2) *rtot; t(inercia.ren)

# Ahora, se calculan las distancias chi cuadrada de cada
```

```
# columna de R a su centroide: (C-r)'Dr^-1(C-r)

# y la inercia de cada columna

dist.chicua.col   <-   sqrt(apply(((tabla_C)-rtot)^2/rtot,2,sum));
dist.chicua.col

inercia.col <- ((dist.chicua.col)^2) *ctot; t(inercia.col)

###################### considerar 9 dimensiones

#Se obtienen las representaciones de las filas y columnas en un
espacio de

#dos dimensiones considerando los vectores propios ai y bi
obtenidos de la dvs

ind<-c(2,3,4,5,6,7,8,9,10)

#Representacion de las filas en 9 dimensiones

Cr<-(sqrt(solve(Dr)))%*%Z%*%dvalsing$v[,ind]; Cr

#head(dvalsing$v)

#Representacion de las columnas en 9 dimensiones

Cc<-(sqrt(solve(Dc)))%*%t(Z)%*%dvalsing$u[,ind]; Cc

class(Cr)

euclidiana <- function(a, b) sqrt ( sum((a - b) ^ 2))

distancia_ent<-rep(0,32)

for (i in 1:32){  distancia_ent[i]<-euclidiana(Cc[25,],Cr[i,])}

distancia_ent

min(distancia_ent)==distancia_ent


# ----ejercicio para región-cultivo

# Se carga la tabla de contingencia

tabla<-t(tabla)

# se le da nombre a las filas y a las columnas
```

```
row.names(tabla)      <-      c("NOROESTE","NORESTE","NORTE","CENTRO
NORTE","OCCIDENTE","CENTRO
SUR","ORIENTE","SUR","SURESTE","CENTRO")

colnames(tabla)<-
c("AGUACATE","ALFALFA","AMARANTO","ARROZ","CACAO","CAFE","CALABACI
TA","CALABAZA","CAÑA                                                DE
AZUCAR","CEBOLLA","CHILE","FRESA","FRIJOL","JITOMATE","LIMON","MAI
Z","MANGO","MANZANA","NARANJA","PLATANO","SORGO","SOYA","TOMATE
ROJO","TRIGO","UVA")

chiR <- chisq.test(tabla); chiR

n<-sum(tabla)

# se calcula la matriz F de correspondencia

renglones<-nrow(tabla);        columnas<-ncol(tabla);    Matriz_F<-
(tabla)/n

#calcular frecuencias relativas de filas y de columnas de F o

#tambien llamadas masas (las ri y las cj)

rtot<-apply(Matriz_F,1,sum);  ctot<-apply(Matriz_F,2,sum)

# vectores obtenidos anteriormente se colocan en matrices
diagonales

Dr<-diag(rtot);  Dc<-diag(ctot)

# calcular matriz de perfiles por fila (R)

tabla_R<-Matriz_F/rtot  #matriz R de renglones

# calcular matriz de perfiles por columna (Rc)

tabla_C<- t(t(Matriz_F)/ctot)  #matriz R de columnas

# calcular chi cuadrada para probar la independencia de renglones
y

#columnas de la tabla de contingencia y la inercia total (medida
de la

#variabilidad total de los datos en la tabla)

tabla_suma_R<-n*rtot;  tabla_suma_C<-n*ctot

tabla_esperadas<-tabla_suma_R%o%tabla_suma_C/n      #   matriz   de
frecuencias esperadas
```

```
chi2<- sum((tabla-tabla_esperadas)^2/tabla_esperadas); chi2

inetot<-chi2/n; inetot


# calcular la matriz Z

Z<-(sqrt(solve(Dr)))%*%Matriz_F%*%(sqrt(solve(Dc)))

#obtener vectores propios y valores propios de Z

dvalsing<-svd(Z); dvalsing$d

#Se obtienen las representaciones de las filas y columnas en un
espacio de

#dos dimensiones considerando los vectores propios ai y bi
obtenidos de la dvs

ind<-c(2,3)

#Representacion de las filas en dos dimensiones

Cr<-(sqrt(solve(Dr)))%*%Z%*%dvalsing$v[,ind]; Cr

#Representacion de las columnas en dos dimensiones

Cc<-(sqrt(solve(Dc)))%*%t(Z)%*%dvalsing$u[,ind]; Cc

#calcular proporcion de la inercia explicada por las dos

#dimensiones (asociadas a valores propios mas grandes distintos de
uno)

vp<-(dvalsing$d)^2

vp_dist1<-vp[-1]; vp_dist1

# Inercia explicada relativa

iner_expl_rel <- vp_dist1/sum(vp_dist1); iner_expl_rel

# Inercia explicada acumulada

iner_expl_acum <- iner_expl_rel

for(i in 1:length(iner_expl_rel))

  iner_expl_acum[i] <- sum(vp_dist1[1:i])/sum(vp_dist1)

iner_expl_acum
```

```
# Gráfica

# Inercia explicada relativa

plot(factor(1:length(iner_expl_rel)),iner_expl_rel,lwd=0.5,col="wh
ite",pch=20,cex=0,

     main="Proporción relativa de \nInercia explicada \nregión-
cultivo",type="n",lty=0,


ylim=c(0,max(iner_expl_acum)),xlab="Dimensión",ylab="Proporción de
la Inercia Total",bty="n",

     cex.main=1,cex.axis=.7,cex.lab=.7,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,col="grey70",lty=3)

abline(h=iner_expl_rel,v=1:length(iner_expl_rel),col="grey90",lty=
3)

lines(iner_expl_rel,col="blue",pch=20)

points(iner_expl_rel,col="blue",pch=20)

# Inercia explicada acumulada

plot(factor(1:length(iner_expl_acum)),iner_expl_acum,lwd=0.5,col="
white",pch=20,cex=0,

     main="Proporción acumulada de \nInercia explicada \nregión-
cultivo",type="n",lty=0,


ylim=c(0,max(iner_expl_acum)),xlab="Dimensión",ylab="Proporción de
la Inercia Total",bty="n",

     cex.main=1,cex.axis=.7,cex.lab=.7,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,col="grey70",lty=3)

abline(h=iner_expl_acum,v=1:length(iner_expl_acum),col="grey90",lt
y=3)

lines(iner_expl_acum,col="blue",pch=20)

points(iner_expl_acum,col="blue",pch=20)
```

```
# Graficamos la representación conjunta de los renglones y
columnas en el

# espacio de dos dimensiones

# Con estas dos dimensiones solo se explica el 53.88% de la
inercia total.

#iner_expl_rel[1]*100 + iner_expl_rel[2]*100

plot(Cr[,1],Cr[,2],xlim=range(Cr[,1],Cc[,1],Cr[,1]+0.5),ylim=range
(Cr[,2],Cc[,2]),

     main="Representación de la tabla \nde contingencia por
Región",

     col="white",pch=20,type="p",lwd=1,

     xlab=paste("Dimensión                                    1
(",round(iner_expl_rel[1]*100,2),"%)",sep=""),

     ylab=paste("Dimensión                                    2
(",round(iner_expl_rel[2]*100,2),"%)",sep=""),


bty="n",cex=1,cex.main=1,cex.axis=.7,cex.lab=.8,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,v=0,lty=3,col="grey90")

points(Cr[,1],Cr[,2],col="blue",pch=20,cex=1)

text(Cr[,1],Cr[,2],labels=c("NOROESTE","NORESTE","NORTE","CENTRO
NORTE","OCCIDENTE","CENTRO
SUR","ORIENTE","SUR","SURESTE","CENTRO"),

     pos=4,cex=0.5,col="darkblue")

plot(Cr[,1],Cr[,2],xlim=range(Cr[,1],Cc[,1],Cr[,1]+0.5),ylim=range
(Cr[,2],Cc[,2]),

     main="Representación de la tabla \nde contingencia por
Cultivos de la ENA2019",

     col="white",pch=20,type="p",lwd=1,

     xlab=paste("Dimensión                                    1
(",round(iner_expl_rel[1]*100,2),"%)",sep=""),

     ylab=paste("Dimensión                                    2
(",round(iner_expl_rel[2]*100,2),"%)",sep=""),
```

```
bty="n",cex=1,cex.main=1,cex.axis=.7,cex.lab=.8,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,v=0,lty=3,col="grey90")

points(Cc[,1],Cc[,2],col="red",pch=20,cex=1)

text(Cc[,1],Cc[,2],labels=names(tabla)[-
1],pos=4,cex=0.5,col="darkred")

text(Cc[,1],Cc[,2],labels=c("AGUACATE","ALFALFA","AMARANTO","ARROZ
","CACAO","CAFE","CALABACITA","CALABAZA","CAÑA                    DE
AZUCAR","CEBOLLA","CHILE","FRESA","FRIJOL","JITOMATE","LIMON","MAI
Z","MANGO","MANZANA","NARANJA","PLATANO","SORGO","SOYA","TOMATE
ROJO","TRIGO","UVA"),

     pos=1,cex=0.45,col="darkred")

plot(Cr[,1],Cr[,2],xlim=range(Cr[,1],Cc[,1],Cr[,1]+0.5),ylim=range
(Cr[,2]-0.5,Cc[,2]),

     main="Representación conjunta de la tabla \nde contingencia
Región-Cultivo de la ENA2019",

     col="white",pch=20,type="p",lwd=1,

     xlab=paste("Dimensión                                        1
(",round(iner_expl_rel[1]*100,2),"%)",sep=""),

     ylab=paste("Dimensión                                        2
(",round(iner_expl_rel[2]*100,2),"%)",sep=""),


bty="n",cex=1,cex.main=1,cex.axis=.7,cex.lab=.8,mgp=c(1.5,.5,0),

     col.main="black", las=1)

abline(h=0,v=0,lty=3,col="grey90")

points(Cr[,1],Cr[,2],col="blue",pch=20,cex=1)

text(Cr[,1],Cr[,2],labels=c("NOROESTE","NORESTE","NORTE","CENTRO
NORTE","OCCIDENTE","CENTRO
SUR","ORIENTE","SUR","SURESTE","CENTRO"),

     pos=4,cex=0.5,col="darkblue")

points(Cc[,1],Cc[,2],col="red",pch=20,cex=1)
```

```
text(Cc[,1],Cc[,2],labels=names(tabla)[-
1],pos=4,cex=0.5,col="darkred")

text(Cc[,1],Cc[,2],labels=c("AGUACATE","ALFALFA","AMARANTO","ARROZ
","CACAO","CAFE","CALABACITA","CALABAZA","CAÑA                    DE
AZUCAR","CEBOLLA","CHILE","FRESA","FRIJOL","JITOMATE","LIMON","MAI
Z","MANGO","MANZANA","NARANJA","PLATANO","SORGO","SOYA","TOMATE
ROJO","TRIGO","UVA"),

     pos=1,cex=0.45,col="darkred")

# A continuación, se calcula las distancias chi cuadrada

# de cada renglón de R a su centroide: (R-c)'Dc^-1(R-c)

# y calculando la inercia de cada renglón

dist.chicua.ren   <-   sqrt(apply((t(tabla_R)-ctot)^2/ctot,2,sum));
dist.chicua.ren

inercia.ren <- ((dist.chicua.ren)^2) *rtot; t(inercia.ren)

# Ahora, se calculan las distancias chi cuadrada de cada

# columna de R a su centroide: (C-r)'Dr^-1(C-r)

# y la inercia de cada columna

dist.chicua.col    <-    sqrt(apply(((tabla_C)-rtot)^2/rtot,2,sum));
dist.chicua.col

inercia.col <- ((dist.chicua.col)^2) *ctot; t(inercia.col)

#5 dimensiones considerando los vectores propios ai y bi obtenidos
de la dvs

ind<-c(2,3,4,5,6)

#Representacion de las filas en 5 dimensiones

Cr<-(sqrt(solve(Dr)))%*%Z%*%dvalsing$v[,ind]; Cr

#Representacion de las columnas en 5 dimensiones

Cc<-(sqrt(solve(Dc)))%*%t(Z)%*%dvalsing$u[,ind]; Cc

distancia_cult<-rep(0,25)

for (i in 1:25){  distancia_cult[i]<-euclidiana(Cc[i,],Cr[10,])}

distancia_cult

min(distancia_cult)==distancia_cult
```

Código de R utilizado para Árboles de clasificación

```
library(caret)

library(dplyr)

# cargar librarias de analisis

library(tidyverse)

# cargar librerias para clasificacion

# install.packages('rpart')

library(rpart)

#install.packages('rattle')

library(rattle)

# install.packages('rpart.plot')

library(rpart.plot)

#-----------------------------------------------------------------
---

########### inicia ejercicio

# cargar base de datos

#CARGAR RUTA

setwd("D:/andres.lira/Desktop/RR/13
Maestria_Análisis_Estadístico_y_Computo/TRABAJO FINAL/AF")

var_bd<-read.csv("BASE_ARBOL.csv")

#cambiar tipo de variable

var_bd$NIVEL_D_ESTUDIOS<-as.factor(var_bd$NIVEL_D_ESTUDIOS)

var_bd$SEXO_PROD<-as.factor(var_bd$SEXO_PROD)

var_bd$APOYO_PROD<-as.factor(var_bd$APOYO_PROD)

var_bd$EDAD_PROD<-as.factor(var_bd$EDAD_PROD)

##------------------ árbol sin poda y con las 13 variables

set.seed(12345)
```

```
nobs <- nrow(var_bd)

itrain <- sample(nobs, 0.8 * nobs)

bd_entrenamiento <- var_bd[itrain, ]

bd_prueba <- var_bd[-itrain, ]

arbol_max <- rpart(formula = NIVEL_D_ESTUDIOS ~., data =
bd_entrenamiento,method = 'class',cp = 0)

arbol_max

plotcp(arbol_max)

rpart.plot(arbol_max, fallen.leaves = FALSE,

           main = "Árbol de clasificación \n nivel de estudios del
productor \n poda cp = 0",

           shadow.col = "gray")

#rpart.rules(arbol)

# predicción

prediccion <- predict(arbol_max, newdata = bd_prueba, type =
"class")

# matriz de confusión

confusionMatrix(prediccion, bd_prueba[["NIVEL_D_ESTUDIOS"]])

#error de la prueba

errTest_arbol_max <- mean(

  predict(arbol_max, bd_prueba, type = "class") !=
bd_prueba$NIVEL_D_ESTUDIOS)

errTest_arbol_max

# error empírico

errEmp_arbol_max <- mean(

  predict(arbol_max, bd_entrenamiento, type = "class") !=
bd_entrenamiento$NIVEL_D_ESTUDIOS)

errEmp_arbol_max

#------------arbol óptimo con las 13 variables-------------------
```

```
cp_Opt <- arbol_max$cptable[which.min(arbol_max$cptable[, 4]), 1]

arbol_Opt <- prune(arbol_max, cp = cp_Opt)

#plot(arbol_Opt)

#text(arbol_Opt, xpd = TRUE, cex = 0.8)

rpart.plot(arbol_Opt, fallen.leaves = FALSE,

           main = "Árbol de clasificación \n nivel de estudios del
productor \n poda cp = 0.0002574334",

           shadow.col = "gray")

rpart.rules(arbol_Opt)

# predicción

prediccion <- predict(arbol_Opt, newdata = bd_prueba, type =
"class")

# matriz de confusión

confusionMatrix(prediccion, bd_prueba[["NIVEL_D_ESTUDIOS"]])

#error de la prueba

errTest_arbol_Opt <- mean(

  predict(arbol_Opt, bd_prueba, type = "class") !=
bd_prueba$NIVEL_D_ESTUDIOS)

errTest_arbol_Opt

# error empírico

errEmp_arbol_Opt <- mean(

  predict(arbol_Opt, bd_entrenamiento, type = "class") !=
bd_entrenamiento$NIVEL_D_ESTUDIOS)

errEmp_arbol_Opt

#-------------arbol con las 13 variables y con poda cp=0.002------
-------------

set.seed(12345)

arbol_3 <- rpart(formula = NIVEL_D_ESTUDIOS ~., data =
bd_entrenamiento,method = 'class',cp = 0.002)
```

```
rpart.plot(arbol_3, fallen.leaves = FALSE,

           main = "Árbol de clasificación \n nivel de estudios del
productor \n poda cp = 0.002",

           shadow.col = "gray")

rpart.rules(arbol_3)

# predicción

prediccion <- predict(arbol_3, newdata = bd_prueba, type =
"class")

# matriz de confusión

confusionMatrix(prediccion, bd_prueba[["NIVEL_D_ESTUDIOS"]])

#error de la prueba

errTest_arbol_3 <- mean(

  predict(arbol_3,     bd_prueba,     type     =     "class")     !=
bd_prueba$NIVEL_D_ESTUDIOS)

errTest_arbol_3

# error empírico

errEmp_arbol_3 <- mean(

  predict(arbol_3,   bd_entrenamiento,   type   =   "class")   !=
bd_entrenamiento$NIVEL_D_ESTUDIOS)

errEmp_arbol_3

#######----------arbol con 4 variables y poda cp=0.002

set.seed(12345)

nobs <- nrow(var_bd)

itrain <- sample(nobs, 0.8 * nobs)

bd_entrenamiento <- var_bd[itrain, ]

bd_prueba <- var_bd[-itrain, ]

arbol_4   <-   rpart(formula   =   NIVEL_D_ESTUDIOS   ~EDAD_PROD
+SUP_AGRICOLA  +TECNOLOGIA_CA,  data  =  bd_entrenamiento,method  =
'class',cp = 0.002)

arbol_4
```

```
rpart.plot(arbol_4, fallen.leaves = FALSE,

           main = "Árbol de clasificación \n nivel de estudios del
productor \n poda cp = 0.002",

           shadow.col = "gray")

rpart.rules(arbol_4)

# predicción

prediccion <- predict(arbol_4, newdata = bd_prueba, type =
"class")

# matriz de confusión

confusionMatrix(prediccion, bd_prueba[["NIVEL_D_ESTUDIOS"]])

#error de la prueba

errTest_arbol_4 <- mean(

  predict(arbol_4,     bd_prueba,     type     =     "class")     !=
bd_prueba$NIVEL_D_ESTUDIOS)

errTest_arbol_4

# error empírico

errEmp_arbol_4 <- mean(

  predict(arbol_4,   bd_entrenamiento,   type   =   "class")   !=
bd_entrenamiento$NIVEL_D_ESTUDIOS)

errEmp_arbol_4

#  --  compara  predicciones  entre  dos  árboles  (árbol_max  y
arbol_Opt)

mean(predict(arbol_max, bd_prueba, type = "class") !=

       predict(arbol_Opt, bd_prueba, type = "class"))

# -- compara predicciones entre dos árboles (árbol_3 y arbol_Opt)

mean(predict(arbol_3, bd_prueba, type = "class") !=

       predict(arbol_Opt, bd_prueba, type = "class"))

# -- compara predicciones entre dos árboles (árbol_4 y arbol_Opt)

mean(predict(arbol_4, bd_prueba, type = "class") !=
```

```
       predict(arbol_Opt, bd_prueba, type = "class"))
```

Código de R utilizado para Bosques de clasificación

```
# Cronstruir un bosque de decisión

#install.packages("randomForest")

library(randomForest)

#creamos el modelo

# modelo considerando las 13 variables

set.seed(12345)

bd_RandomF <- randomForest(NIVEL_D_ESTUDIOS ~ .,

                           data=bd_entrenamiento,

                           ntree=500,

                           importance=TRUE)

print(bd_RandomF)

p <- predict(bd_RandomF, bd_prueba)

table(p,bd_prueba$NIVEL_D_ESTUDIOS)

#traza el error OOB en función del número de árboles del bosque:

# errores de prueba y de aprendizaje obtenidos mediante Bagging

errTest_RF <- mean(

  predict(bd_RandomF, bd_prueba) != bd_prueba$NIVEL_D_ESTUDIOS)

errEmp_RF <- mean(

  predict(bd_RandomF,              bd_entrenamiento)              !=
bd_entrenamiento$NIVEL_D_ESTUDIOS)

errTest_RF

errEmp_RF

# importancia de las variables

varImpPlot(bd_RandomF, type = 1, scale = FALSE,

           n.var = ncol(bd_entrenamiento)-1, cex = 0.8,
```

```
           main = "Importancia de Variables")

# modelo considerando 4 variables

set.seed(12345)

bd_RandomF    <-    randomForest(NIVEL_D_ESTUDIOS    ~    EDAD_PROD
+SUP_AGRICOLA +TECNOLOGIA_CA,

                           data=bd_entrenamiento,

                           ntree=500,

                           importance=TRUE)

print(bd_RandomF)

p <- predict(bd_RandomF, bd_prueba)

table(p,bd_prueba$NIVEL_D_ESTUDIOS)

#traza el error OOB en función del número de árboles del bosque:

# errores de prueba y de aprendizaje obtenidos mediante Bagging

errTest_RF <- mean(

  predict(bd_RandomF, bd_prueba) != bd_prueba$NIVEL_D_ESTUDIOS)

errEmp_RF <- mean(

  predict(bd_RandomF,              bd_entrenamiento)              !=
bd_entrenamiento$NIVEL_D_ESTUDIOS)

errTest_RF

errEmp_RF

# importancia de las variables

varImpPlot(bd_RandomF, type = 1, scale = FALSE,

           n.var = ncol(bd_entrenamiento) - 10, cex = 0.8,

           main = "Importancia de Variables")
```