INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS A.C.

Metodología para la estratificación de unidades de producción agropecuaria del Estado de Tlaxcala e implementación de un modelo probabilístico para la asignación de nuevos integrantes

TESIS:

Que para obtener el título de Maestro en Ciencias en Estadística Oficial presenta:

ERIK FERNANDO LIMÓN HERNÁNDEZ

Presidente del Jurado**:**

Dr. Rogelio Ramos Quiroga

Octubre 2011

Resumen

Los productores agropecuarios del estado de Tlaxcala presentan diferencias considerables entre sí, debido a su condición técnica y económica, y a la diversidad climática y ecológica prevalecientes en sus explotaciones.

En el año 2004 se llevó a cabo el Padrón de Productores Agropecuarios del Estado de Tlaxcala, con el propósito de obtener información básica de las Unidades de Producción Agropecuaria (UP), e información estructural del sector agropecuario y forestal, así como generar marcos de muestreo a través del inventario estatal de terrenos, además de la generación del directorio de productores en el Estado.

El presente estudio tiene como objetivo principal desarrollar una metodología que permita obtener la división por estratos de las unidades de producción agropecuaria del estado de Tlaxcala, que contribuya a la toma de decisiones en cuanto a programas o actividades vinculadas directamente con el sector primario en el gobierno estatal y/o federal enfocado a mejorar la situación económica y laboral de los productores. Así mismo, obtener un modelo probabilístico que determine el estrato correspondiente a cada nuevo productor que se registre, y le permita heredar las características o beneficios que conlleva esta asignación.

En general, la metodología propuesta para cumplir con este proyecto consta de:

- Identificación, análisis de temas principales para la integración de la base de datos a trabajar.
- Exploración descriptiva de la base.
- Aplicación de Componentes Principales.
- Método de K-Medias.
- Análisis discriminante.

- Presentación de los resultados preliminares.

Se propone por parte del Grupo de Trabajo de Información Agropecuaria del Comité Técnico Especializado de Información Estadística y Geográfica del Estado de Tlaxcala (CTEIEG), con base en el conocimiento del total de unidades de producción y las características particulares del sector agropecuario en la Entidad, la creación de cinco estratos representativos, homogéneos internamente y heterogéneos entre sí, lo cual permite una correcta descripción propia de cada uno y facilita el trabajo de evaluación con base en sus características.

Palabras Clave: Estratificación, Análisis Multivariado, Sector Primario, Análisis Discriminante.

CONTENIDO

PLANTEAMIENTO

Introducción

En el año 2004 se llevó a cabo el Padrón de Productores Agropecuarios del Estado de Tlaxcala, con el propósito de obtener información básica de las unidades de producción e información estructural del sector agropecuario y forestal, así como generar marcos de muestreo a través del inventario estatal de terrenos, además de la generación del directorio de productores en el Estado.

Cabe mencionar, que el último Censo Agropecuario en el país, antes del año 2004, fue realizado en 1991, y por tanto se tenía la necesidad de contar con información reciente al respecto, solicitada principalmente por el Gobierno del Estado y la Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación, (SAGARPA), siendo Tlaxcala el único estado de la república donde se llevó a cabo un proyecto con estas características.

Los objetivos del levantamiento fueron establecidos por el Instituto Nacional de Estadística y Geografía (INEGI) en coordinación con la SAGARPA, las demás etapas del proyecto fueron desarrolladas y ejecutadas por el INEGI. Algunos productos obtenidos fueron una base de datos y los tabulados básicos.

Es notable que en 2007 se llevó a cabo el VIII Censo Agropecuario a nivel nacional. La idea del proyecto es tomar como base la información del padrón de productores del 2004 pero con la facilidad de poder actualizarse con bases de datos nuevas. Así, cada vez que haya un levantamiento reciente se podrá adaptar la información a los estudios ya realizados.

Definición

Los productores agropecuarios del estado de Tlaxcala presentan diferencias considerables entre sí, debido a su condición técnica y económica, y a la diversidad climática y ecológica prevalecientes en sus explotaciones.

El tener un padrón actualizado de ellos resulta trascendental para promover el fortalecimiento de este sector en diversas áreas, pero sobre todo, en lo que concierne a los apoyos que se les podría proporcionar para la mejora en sus procesos y calidad de sus productos.

Además de contar con un registro completo, resulta conveniente considerar agrupaciones de productores, que tomen en cuenta propiedades en las que son similares, dado que para programas masivos, estudios de regiones, entre otros, no es factible estar analizando a cada productor, sino el grupo al que pertenecen según características propias.

Esta labor de estratificación resulta un tanto laboriosa al identificar el número de integrantes que componen el universo en la entidad, y sobre todo al llevar a cabo un análisis conjunto de variables representativas del sector primario para formar los grupos o estratos.

Con una serie de métodos estadísticos que se proponen, se logrará un agrupamiento de los productores basado en las características más representativas que sugieren los representantes de las instituciones que participan en el Grupo de Trabajo de Información Agropecuaria del CTEIEG.

Por la naturaleza de las técnicas a aplicar, se espera que los registros pertenecientes a un estrato sean lo más similares posibles, y lo más diferentes a los de otro conjunto, quedando con esto bien definida la separación, la identificación y por tanto la descripción de cada uno de los estratos.

Así mismo, una vez establecidas las agrupaciones de las unidades de producción, será necesario encontrar la manera de asignar a una de éstas cada nuevo productor que se registre en el padrón, y prácticamente de manera inmediata. Esta situación también contempla implementar el uso de otros métodos estadísticos descritos más adelante.

Adicionalmente, aunque no contemplado para este trabajo, se propone la automatización del registro, consulta y modificación de las Unidades de Producción con sus respectivos programas de apoyo según al grupo al que pertenecen, así como del método de asignación de estrato de forma inmediata. Con lo que se amplía el horizonte del proyecto dada la inclusión de herramientas informáticas con todas las bondades que esto conlleva.

Objetivo General

Desarrollar una metodología que permita obtener la división por estratos de las unidades producción agropecuaria del estado de Tlaxcala, que contribuya a la toma de decisiones en cuanto a programas o actividades vinculadas directamente con el sector primario en el gobierno estatal y/o federal enfocado a mejorar la situación económica y laboral de los productores. Así mismo, obtener un modelo probabilístico que determine el estrato correspondiente a cada nuevo productor que se registre, y le permita heredar las características o beneficios que conlleva esta asignación.

Objetivos Específicos

- Tener un registro confiable y oportuno de los productores agropecuarios.

- Contribuir con el Gobierno del Estado proporcionándole una herramienta para la toma de decisiones en cuanto a programas de apoyo o cualquier otra actividad vinculada con el campo que se considere pertinente.

- Establecer una metodología con técnicas estadísticas que permita el análisis de las UP y la separación de ellas en grupos homogéneos al interior.

- Contar con las bases para el desarrollo de un sistema informático que permita la administración completa del padrón de unidades de producción agropecuaria y programas de apoyo al sector primario, y su vez, pueda implementar la función probabilística para la asignación de estrato a productores de reciente registro.

- Tener una clara identificación de las características generales de las unidades de producción pertenecientes a un estrato.

- Identificar la distribución en el espacio de las unidades de producción por estrato.

Hipótesis de Investigación

Entre los supuestos principales planteados para este proyecto se tiene:

- Existe un vínculo entre el comportamiento de las unidades de producción con su ubicación geográfica. Por lo que se espera que los estratos queden relativamente definidos en un mapa estatal. Pudiendo con esto agregar factores geográficos a las características de cada uno de los grupos.

- Se supone encontrar una estratificación bien definida, ya que se sabe que los productores agropecuarios del estado de Tlaxcala presentan diferencias considerables entre sí, debido a su condición técnica y económica, y a la diversidad climática y ecológica prevalecientes en sus explotaciones.

- A partir de la estratificación establecida bien definida, se espera encontrar un modelo probabilístico de asignación con alto porcentaje de efectividad.

- Dado que componentes principales es una herramienta base en este trabajo, se experimentarán modelos de predicción con diferentes cantidades de componentes esperando encontrar el mejor modelo con una menor cantidad de componentes que los que se generan en total cuando se abarca el 100% de la variabilidad de los datos.

DESARROLLO

Integración de la base de datos y análisis descriptivo

El proyecto de levantamiento incluyó una gran cantidad de información organizada por diversos temas. En principio, ya se han estudiado los tópicos a incluir con la colaboración del Grupo de Trabajo de Información Agropecuaria del Comité Técnico Especializado de Información Estadística y Geográfica del Estado de Tlaxcala (CTEIEG), cuyos integrantes provienen de la SAGARPA, el INEGI y el Gobierno del Estado a través de la SEFOA (Secretaría de Fomento Agropecuario). Con esto se cuenta con un enfoque directo sobre los temas más relevantes para estas instituciones y sobre los cuales se parte para la estratificación.

Identificación, análisis de temas principales y variables

Cabe resaltar, que no todas las variables pertenecientes a un tema son incluidas en los métodos estadísticos aquí planteados, debido a la naturaleza de las mismas, como el que unas son categóricas y otras escalares. Misma situación que fue tratada en el Comité y se determinó la obtención de las variables más representativas para sus intereses por cada tema y a su vez, la creación de indicadores mediante operaciones aritméticas, con lo cual se fortalece la comparabilidad de los valores entre las unidades de producción. Esta creación de nueve indicadores se detalla en los cuadros siguientes.

Las variables que se proponen para el cálculo, destacan por su poco movimiento en el tiempo, por lo que el comportamiento de la información agropecuaria en estas variables, del Padrón de Productores de 2004 al VIII censo Agropecuario en 2007 es muy semejante, esto nos hace suponer que la metodología propuesta podrá ser aplicable a bases de datos más recientes, prácticamente de forma directa.

Los tópicos propuestos con sus respectivas variables son los siguientes:

Nivel Tecnológico: • Superficie de riego • Superficie de temporal • Sistemas de riego • Fuerza de tracción empleada • Semilla certificada • Fertilizantes o abonos • Herbicidas y/o insecticidas • Existencia y funcionamiento de tractores	**Producción Pecuaria** • Total de cabezas de ganado • Total de reses • Total de cerdos • Total de chivos • Total de borregos
Superficie y Tenencia de la Tierra • Superficie ejidal • Superficie comunal • Superficie municipal • Superficie propiedad privada	**Características de la UP** • Superficie total de la UP • Total de ingresos de la UP
Sector Agrícola • Superficie agrícola de la UP • Superficie total sembrada • Superficie cultivada • Producción total • Volumen cosechado en terreno de vivienda del productor • Principal cultivo • Promedio de producción	**Formas de organización y comercialización de la UP** • Integrantes del grupo u organización de la UP • Personas adicionales que trabajan para la UP • Principal forma de comercialización

Y los nueve indicadores con base en los cuales se aplicarán las técnicas son:

Indicador	Descripción	Unidad de medida	Conformación
RSupagricola	Razón de superficie agrícola de la unidad productora.	Hectáreas	Suma de la superficie agrícola en cada terreno, incluyendo donde se encuentra la vivienda del productor dividido por la superficie total de la UP
RSuptemp	Razón de superficie de temporal de la unidad productora	Hectáreas	Superficie de temporal en la UP dividido por su superficie total de la UP
RSupriego	Razón de superficie de riego de la unidad productora	Hectáreas	Superficie de riego en la UP dividido por su superficie total de la UP
RTotsupsembrada	Razón de superficie total sembrada	Hectáreas	Suma de las hectáreas sembradas en la UP durante el año dividido por la superficie total de la UP
RSupfertilizada	Razón de superficie en la que utilizó fertilizante	Hectáreas	Superficie en la que se usó fertilizante dividido por la superficie total de la UP
RSupsemcert	Razón de superficie sembrada con semilla certificada	Hectáreas	Superficie donde se uso semilla certificada dividida por la superficie total de la UP
Rproduccion	Producción total agrícola por hectárea en la UP	Ton/Htas.	Suma de todos los productos agrícolas dividido por la superficie total de la UP
Ringresos	Ingresos promedio del productor provenientes de la actividad agropecuaria por hectárea	Pesos/htas.	Suma de las ganancias obtenidas por agricultura, ganadería y aprovechamiento forestal en el año dividido por la superficie total de la UP
Rcabezas	Promedio de cabezas de ganado de la unidad productora por hectárea	Cabezas/htas.	Suma de reses, cerdos, chivos y borregos dividido por la superficie total de la UP

Con esto se establece una base definitiva conformada por 33 variables, dos de ellas para identificación de la unidad de producción, 9 de cálculo, y 22 variables de descripción general, con 53967 registros (unidades de producción).

Análisis descriptivo de las variables

Algunos estadísticos descriptivos de la tabla propuesta son:

Descriptivos

	Estadístico						
	Media	**Media recortada al 5%**	**Mediana**	**Varianza**	**Desv. típ.**	**Mínimo**	**Máximo**
Ragricola	81.97	85.52	97.09	958.19	30.95	.00	100.00
RTemp	72.44	74.93	95.56	1459.02	38.20	.00	100.00
R_SRiego	9.53	5.15	.00	698.63	26.43	.00	100.00
RSembrada	83.70	84.12	97.22	1410.64	37.56	.00	600.00
RFertilizada	63.00	63.98	86.96	1806.60	42.50	.00	200.00
RSemilla_Certif	7.35	2.61	.00	608.80	24.67	.00	200.00
Promedio_Producc	3.69	1.86	1.22	188.93	13.75	.00	370.29
Ringresos	4970.98	472.14	.00	7424292714.32	86164.34	.00	12222222.22
RGanado	21.61	1.65	.00	345812.56	588.06	.00	90000.00

	Percentiles						
	5	**10**	**25**	**50**	**75**	**90**	**95**
Ragricola	.00	9.52	83.33	97.09	99.34	100.00	100.00
RTemp	.00	.00	52.94	95.56	99.17	100.00	100.00
R_SRiego	.00	.00	.00	.00	.00	48.36	93.13
RSembrada	.00	15.91	76.92	97.22	99.72	100.00	107.82
RFertilizada	.00	.00	.00	86.96	98.60	99.84	100.00
RSemilla_Certif	.00	.00	.00	.00	.00	.00	92.25
Promedio_Producc	.00	.00	.50	1.22	2.82	6.05	7.00
Ringresos	.00	.00	.00	.00	685.71	2620.11	5013.01
RGanado	.00	.00	.00	.00	1.68	9.41	25.81

Destaca la poca variabilidad en los indicadores de superficie, contrastando con el comportamiento de los ingresos, la producción y las cabezas de ganado.
Esto sustenta la creación de razones en las variables originales para la aplicación de las técnicas estadísticas. A su vez, con la generación de unidades de medida distintas, y la gran diferencia numérica entre la variable ingresos con las demás, se propone usar las variables estandarizadas.

Ahora se verifican las correlaciones:

	RSupagricola	RSuptemp	RSupriego	Rtotsupsembrada	Rsupfertilizada	RSupsemcert	RProducción	Ringresos	Rcabezas
RSupagricola	1.000								
RSuptemp	.727	1.000							
RSupriego	.121	-.594	1.000						
Rtotsupsembrada	.601	.392	.137	1.000					
Rsupfertilizada	.588	.426	.074	.517	1.000				
RSupsemcert	.137	.086	.036	.139	.179	1.000			
RProducción	.041	-.145	.257	.179	-.050	.004	1.000		
Ringresos	-.097	-.076	-.004	-.030	-.060	-.008	.019	1.000	
Rcabezas	-.077	-.058	-.007	-.023	-.047	-.010	.016	.270	1.000

Destaca, como es de esperarse, que exista alta correlación entre la superficie agrícola con la superficie de temporal, y el total de superficie sembrada con superficie fertilizada, y otras marcadas en la tabla.

El que exista alta correlación entre algunas variables representa cierta dificultad para los métodos de conglomeración, pues al tener un comportamiento semejante las variables será más complicado identificar separaciones para la creación de grupos. Dada esta situación y la propuesta de los participantes en el comité por incluir los 9 indicadores, se obtendrá una matriz representativa de los datos mediante componentes principales.

Estratificación de las Unidades de Producción

El procedimiento general para llevar a cabo la estratificación consta de: generación de componentes principales con el fin de obtener nuevas variables linealmente independientes y reducción de la información, incluyendo los registros que llegaran a considerarse atípicos, ya que posteriormente se generarán los grupos con la técnica de k-medias sobre los scores obtenidos por componentes principales y no sobre los valores reales de las variables, pero sobre todo, en la generación del modelo probabilístico cuyas entradas también serán los scores, por lo que es necesario aplicar las técnicas sobre todos los registros desde un inicio.

Componentes principales

El Análisis de Componentes Principales es una técnica estadística de síntesis de la información, o reducción de la dimensión. Es decir, ante un banco de datos con varias variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Esto se logra mediante la creación de combinaciones lineales de las variables originales. Estas nuevas variables *no correlacionadas* son llamadas componentes principales.

Algebraicamente, son combinaciones lineales de las p variables aleatorias $X_1,...,X_p$. Geométricamente, estas combinaciones lineales representan la selección de un nuevos sistema de coordenadas obtenido por la rotación del sistema original con $X_1,\ldots,X_p$ como los ejes coordenados. Los nuevos ejes representan las direcciones con máxima variabilidad y proporcionan una descripción más simple de la estructura de covarianza.

Sea el vector aleatorio X' = [*X1, . . . ,Xp*] con matriz de covarianza Σ con autovalores $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_p \geq 0$, y autovectores $e_1, e_2, ..., e_p$, respectivamente.

Entonces la *i-´*esima componente principal es dada por:

$$Y_i = \boldsymbol{e}_i'\boldsymbol{X} = e_{i1}X_1 + \cdots + e_{ip}X_p \quad i = 1, \ldots, p$$

Con:

$$\begin{aligned} Var(Y_i) &= \boldsymbol{e}_i'\Sigma\boldsymbol{e}_i = \lambda_i \qquad i = 1, \ldots, p \\ Cov(Y_i, Y_k) &= \boldsymbol{e}_i'\Sigma\boldsymbol{e}_k = 0 \quad i \neq k \end{aligned}$$

Si algunas λ_i son iguales, entonces las respectivas Y_i no son únicas.

Otra característica es:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

Lo que indica que:

$$\begin{aligned} \text{Varianza total de la población} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \cdots + \lambda_p \end{aligned}$$

Por tanto:

$$\left(\begin{array}{l}\text{Proporción de la varianza}\\ \text{poblacional total}\\ \text{debida a la k−ésima}\\ \text{componente principal}\end{array}\right) = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

Para cada $k = 1, \ldots, p.$

Como se observó en el análisis descriptivo, existe una alta correlación entre algunas variables propuestas para cálculo, pero a la vez impera la inclusión de todas en el estudio. En caso de trabajarlas como tales, se estaría afectando una característica importante de los métodos de conglomeración, que es la no correlación de las variables para la generación de grupos mejor definidos, es decir, más homogéneos al interior y heterogéneos entre sí.

La solución propuesta a esta situación es convertir los 9 indicadores originales en 9 combinaciones lineales de sí mismas no correlacionadas, con base en el análisis de su variabilidad, lo cual no es otra cosa más que crear los componentes principales de la tabla original explicando el 100% de la varianza.

Varianza total explicada

Componente	eigenvalor	% de la varianza	% acumulado
1	2.709	30.104	30.104
2	1.604	17.821	47.926
3	1.256	13.951	61.877
4	.972	10.804	72.681
5	.848	9.426	82.107
6	.730	8.110	90.217
7	.452	5.027	95.244
8	.428	4.756	100.000
9	.000	.000	100.000

Matriz de componentes

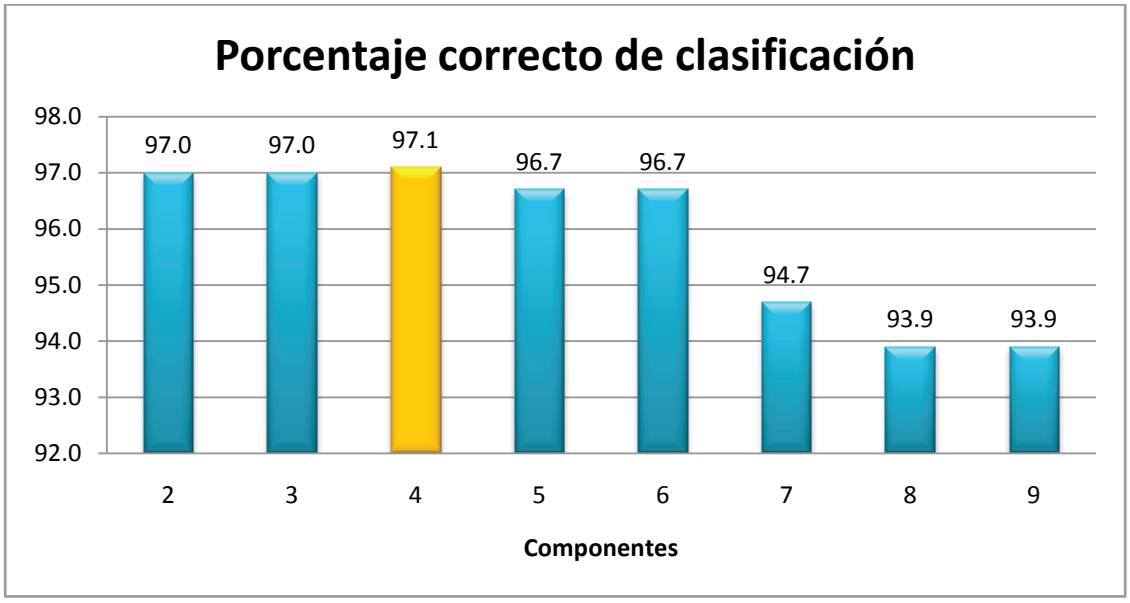
	Componente								
	1	2	3	4	5	6	7	8	9
RSupagricola	.894	.131	.018	-.108	-.083	.001	.366	.173	.000
RSuptemp	.808	-.511	.077	-.136	.179	.001	.149	.087	.000
RSupriego	-.120	.891	-.091	.071	-.356	.000	.213	.077	.000
Rtotsupsembrada	.745	.348	.098	-.123	.015	-.017	-.087	-.539	.000
Rsupfertilizada	.765	.161	.046	.101	-.286	-.011	-.478	.257	.000
RSupsemcert	.253	.141	.056	.884	.357	.020	.051	-.009	.000
RProducción	-.012	.603	.037	-.359	.684	.017	-.105	.165	.000
Ringresos	-.147	.034	.780	.004	-.031	-.606	.026	.028	.000
Rcabezas	-.123	.025	.785	-.017	-.068	.602	.016	.014	.000

Se calculan las puntuaciones o scores y se verifica que no existe dependencia lineal entre ellas:

Matriz de covarianza de las puntuaciones de las componentes									
Componente	1	2	3	4	5	6	7	8	9
1	1.0	.0	.0	.0	.0	.0	.0	.0	.0
2	.0	1.0	.0	.0	.0	.0	.0	.0	.0
3	.0	.0	1.0	.0	.0	.0	.0	.0	.0
4	.0	.0	.0	1.0	.0	.0	.0	.0	.0
5	.0	.0	.0	.0	1.0	.0	.0	.0	.0
6	.0	.0	.0	.0	.0	1.0	.0	.0	.0
7	.0	.0	.0	.0	.0	.0	1.0	.0	.0
8	.0	.0	.0	.0	.0	.0	.0	1.0	.0
9	.0	.0	.0	.0	.0	.0	.0	.0	1.0

Con lo que se cumple el objetivo de utilizar esta técnica previamente a la conglomeración.

Respecto a la cantidad de componentes principales a retener, y adelantando un poco las siguientes técnicas, se propone realizar una separación por conglomerados y un modelo discriminante verificando el porcentaje correcto de asignación al retener dos, tres y hasta nueve componentes. El comparativo general es el siguiente:



Con este procedimiento se concluye una mejor asignación con cuatro componentes.

Técnica de K-Medias

Se cuenta con *n* variables en vectores $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_n}$, donde cada x, está representado en un espacio *m* dimensional y se sabe que están agrupados en *k* cúmulos (k<n). Se define m_j como la media del j-ésimo cúmulo. Si los cúmulos están bien separados, se puede usar una mínima distancia de clasificación para separarlos. Esto es, $\boldsymbol{x_i}$ esta en el j-ésimo cúmulo si $\| \boldsymbol{x_i} - \boldsymbol{m_j} \|$ es el mínimo con respecto a los k cúmulos. Esto sugiere el siguiente algoritmo para encontrar las k-medias:

- Hacer una estimación inicial para la k medias $m_1, m_2, ..., m_k$.

- Mientras no cambie alguna media:

    - Usar la media estimada para clasificar los datos en cúmulos. ***b(i,j) = 1*** si el i-ésimo dato, es el más cercano a la j-ésima media.

    - Para cada uno de los cúmulos
        - Calcular la nueva media m_j, utilizando la nueva clasificación

$$\boldsymbol{m_j} = \frac{\sum_{i=1}^{n} b(i,j) * x_i}{\sum_{i=1}^{n} b(i,j)}$$

    - fin
- fin

Este algoritmo tiene algunas debilidades.

- La manera de inicializar no se especifica. Una forma común es comenzar seleccionado aletoriamente k medias de la muestra.

- Los resultados dependerán del valor inicial de las medias y frecuentemente pasa que particiones “subóptimas” son encontradas. La solución estándar es calculada atizando diferentes puntos de arranque.

- Los resultados dependen de la métrica utilizada para medir || $\boldsymbol{x}$ - $\boldsymbol{m}_i$ ||. Una solución es la estandarización de las variables.

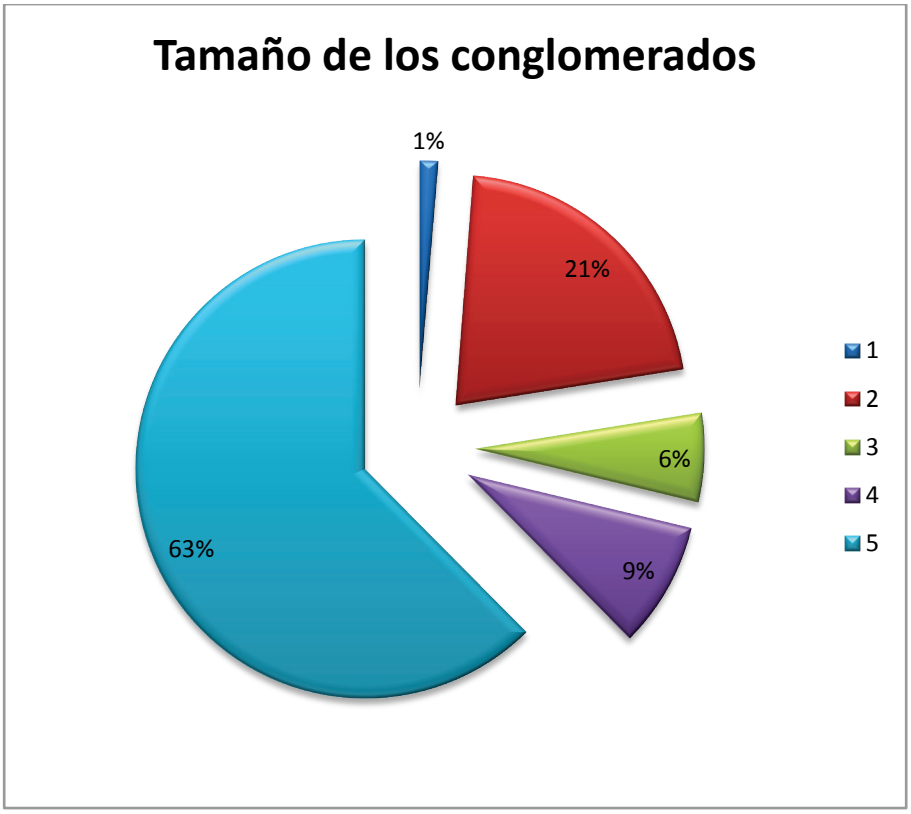
- La solución depende del número de cúmulos seleccionado.

- El último problema es particularmente pesado, normalmente no se sabe cuántos cúmulos existen, para el caso del estudio y la necesidad organizacional, se plantean cinco.

Aplicación a la base de unidades de producción

Tomando como entrada los cuatro componentes principales obtenidos en el método anterior, usando k medias con centros iniciales al azar, y distancia de Mahalanobis, se generan los siguientes resultados de conglomeración:

Distribución de conglomerados			
		N	% del total
Tamaño de conglomerados			
	1	657	1.22
	2	11480	21.27
	3	3365	6.24
	4	4774	8.85
	5	33691	62.43
Total		53967	100



La verificación de estos grupos con la información original de las variables agropecuarias se presenta al final del documento en la sección de resultados generales preliminares.

Implementación del Modelo Probabilístico de Asignación

Una vez que se tiene el total de registros ya con un grupo asignado, se plantea el estudio de esta base para la creación de un modelo matemático con alta probabilidad de correcta asignación de estrato para cada nueva unidad de producción agropecuaria que se registre, incluso se plantea que podría utilizarse para estratificar un nuevo levantamiento de las unidades del estado de Tlaxcala, puesto que las características del padrón con el que actualmente se cuenta y lo obtenido en un nuevo registro arrojaría resultados no muy distantes. Para cumplir con este objetivo, se implementa con la información la siguiente técnica.

Análisis Discriminante

El Análisis Discriminante es una técnica estadística multivariada cuya finalidad es analizar si existen diferencias significativas entre grupos de objetos respecto a un conjunto de variables medidas sobre los mismos para, en el caso de que existan, explicar en qué sentido se dan y proporcionar procedimientos de clasificación sistemática de nuevas observaciones de origen desconocido en uno de los grupos analizados. Por tanto, los objetivos del Análisis Discriminante pueden sintetizarse en dos:

1) *Descriptivo*. Analizar si existen diferencias entre los grupos en cuanto a su comportamiento con respecto a las variables consideradas y averiguar en qué sentido se dan dichas diferencias

2) *Predictivo*. Elaborar procedimientos de clasificación sistemática de individuos de origen desconocido, en uno de los grupos analizados.

Este segundo enfoque es el requerido para este estudio dado el planteamiento de obtener un modelo matemático que con cierta probabilidad genere el estrato correspondiente a una unidad de producción de reciente ingreso al sistema.

Modelo matemático

A partir de *q* estratos donde se asignan a una serie de objetos y de *p* variables medidas sobre ellos ($x_1, \ldots, x_p$), se trata de obtener para cada objeto una serie de puntuaciones que indican el grupo al que pertenecen ($y_1, \ldots, y_m$), de modo que sean funciones lineales de $x_1, \ldots, x_p$:

$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1p}x_p + a_{10} \\ &\cdots\cdots\cdots \\ y_m &= a_{m1}x_1 + \cdots + a_{mp}x_p + a_{m0} \end{aligned}$$

donde m = q − 1, tales que discriminen o separen lo máximo posible a los q grupos. Estas combinaciones lineales de las p variables deben maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos. Geométricamente, el análisis discriminante busca ejes que separen lo más posible los centros de los grupos. Maximizando:

$$\frac{\text{Variabilidad entre grupos}}{\text{Variabilidad intra grupos}}$$

El modelo matemático generado, por tanto, radica en conseguir a partir de esta maximización *los coeficientes de las funciones lineales discriminantes* (y_i's). Se trata de pesos o ponderaciones discriminantes y son determinados por la estructura de varianza de las variables originales a través de los grupos de la variable dependiente (estrato). Las variables independientes con un poder discriminante grande por lo general presentan pesos grandes y las que tienen poco poder discriminante presentan pesos pequeños. Aunque la existencia de multicolinealidad entre las variables independientes puede conducir a una excepción de la regla.

Es necesario considerar una serie de restricciones o supuestos para esta técnica:

Se tiene una variable categórica y el resto de variables son de intervalo o de razón y son independientes respecto de ella. Situación que se cumple pues la variable categórica estará definida por el estrato correspondiente y las demás serán las puntuaciones de los componentes principales obtenidos.

Es necesario que existan al menos dos grupos, y para cada grupo se necesitan dos o más casos.

Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes. Esto se consigue con la obtención de los componentes principales, dado que resultan linealmente independientes.

Normalidad multivariada. Se consigue una aproximación de esta distribución mediante la obtención de los componentes principales, aún así cabe resaltar la robustez del método ante la falta de este supuesto.

Una vez establecido el modelo y aplicado a los datos de la base de componentes con la que se cuenta, se verificará la efectividad del mismo con base en la correcta clasificación comparando con la variable estrato obtenida en el método de conglomerados jerárquicos.

Se aplica el método a la base de puntuaciones de componentes obteniendo los siguientes resultados:

Coeficientes de las funciones canónicas discriminantes

	Función			
	F1	F2	F3	F4
CP1	1.005	.371	-.363	-.299
CP2	.707	.715	.443	.483
CP3	.251	-.050	-1.011	.405
CP4	.211	.095	.621	.207
(Constante)	.000	.000	.000	.000

La tabla anterior muestra las 4 funciones discriminantes, en otras palabras, los 4 ejes que separan los 5 grupos de forma analítica. Cada valor es el coeficiente correspondiente a las variables trabajadas, en este caso, los componentes principales.

A continuación se evalúan las funciones en los promedios de cada grupo, con el fin de obtener las puntuaciones, o coordenadas de los centroides correspondientes. Así, mediante la distancia más corta de Mahalanobis entre las puntuaciones de cada unidad de producción y los centroides se puede determinar el estrato a asignar.

	Función			
Estrato	1	2	3	4
1	7.366	2.241	-.696	-2.676
2	.411	.574	-1.502	.357
3	2.190	-3.617	.253	.310
4	1.453	1.551	1.989	.639
5	-.708	-.098	.218	-.191

Se aplica el modelo con los datos de la base de componentes para comprobación de la capacidad de predicción y se obtiene:

Resultados de la clasificación

	Estrato	Grupo de pertenencia pronosticado					Total
		1	2	3	4	5	1
Recuento	1	365	25	40	207	20	657
	2	0	10091	0	115	1274	11480
	3	0	40	3258	19	48	3365
	4	0	154	0	4620	0	4774
	5	0	636	0	0	33055	33691
%	1	55.56	3.81	6.09	31.51	3.04	100.0
	2	0.00	87.90	0.00	1.00	11.10	100.0
	3	0.00	1.19	96.82	0.56	1.43	100.0
	4	0.00	3.23	0.00	96.77	0.00	100.0
	5	0.00	1.89	0.00	0.00	98.11	100.0

Clasificados correctamente el 97.1% de los casos agrupados originales.

ALGORITMO GENERAL

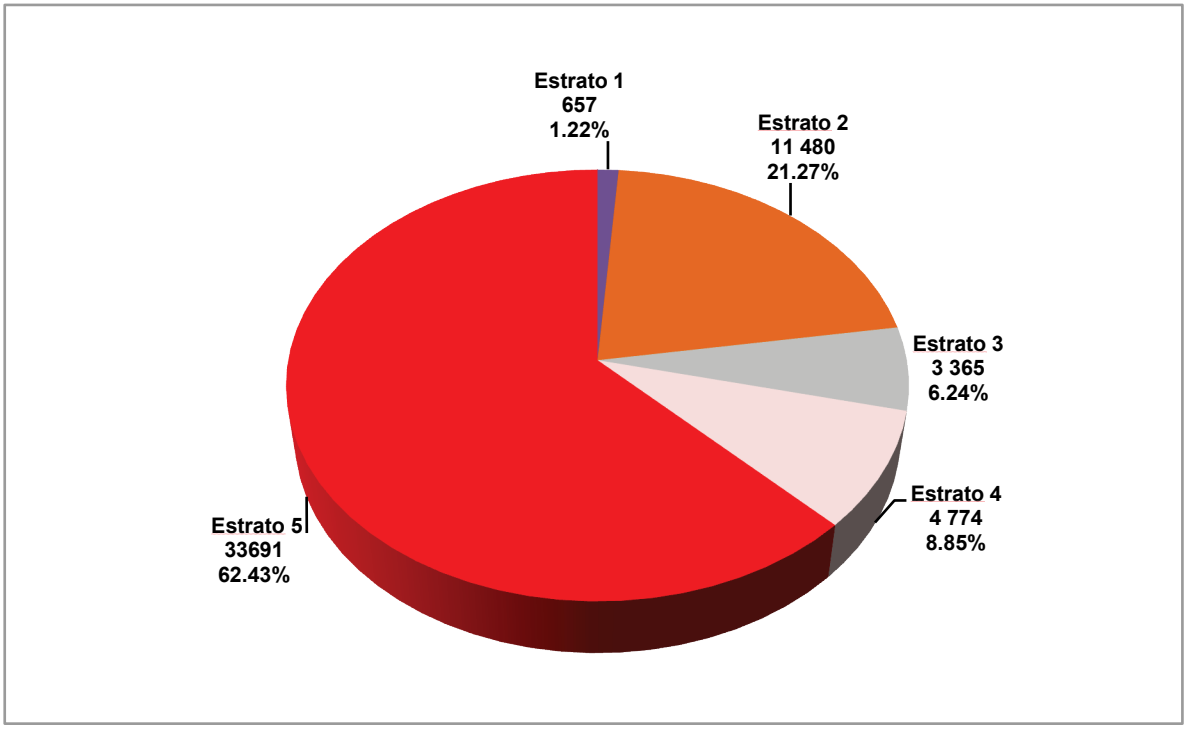
Una vez implementadas las técnicas, se puede resumir la obtención de estrato para una unidad de producción de la siguiente manera:

- Captura de las variables agropecuarias de la unidad de producción.
- Generación de los 9 indicadores propuestos para asignación de estrato.
- Obtención de las cuatro puntuaciones (scores) en los Componentes Principales obtenidos.
- Aplicación de las funciones discriminantes a los scores.
- Obtención del estrato para la Unidad de producción con el modelo probabilístico generado.

RESULTADOS GENERALES PRELIMINARES

La distribución de unidades de producción, quedó definida de la siguiente forma:

	Cantidad de UP	Proporción de UP por estrato (%)
Estrato 1	657	1.22
Estrato 2	11480	21.27
Estrato 3	3365	6.24
Estrato 4	4774	8.85
Estrato 5	33691	62.43
Estatal	53967	100



Descriptivos por estrato

Estrato 1, *Unidades de Producción grandes y muy grandes*

- Contiene apenas el 1.22% del total de UP, pero son responsables del 21.9% de la superficie agrícola de la Entidad.

- Son UP grandes y muy grandes, el 50% de ellas tienen una superficie total entre 15.8 y 71.7 hectáreas; el 20% entre 71.7 y 254.5; y el 5% tienen una superficie de 254.5 o más hectáreas. La mediana en este grupo es de 40.1ha.

- La superficie es predominantemente de temporal (85.51%) siendo relativamente representativa la superficie de riego (14.49%).

- El rendimiento promedio total es de 9.06 ton/ha y es el segundo más alto de los cinco estratos.

- El promedio de cabezas de ganado es de 60.18 animales que incluyen ganado bovino, porcino, ovino y caprino.

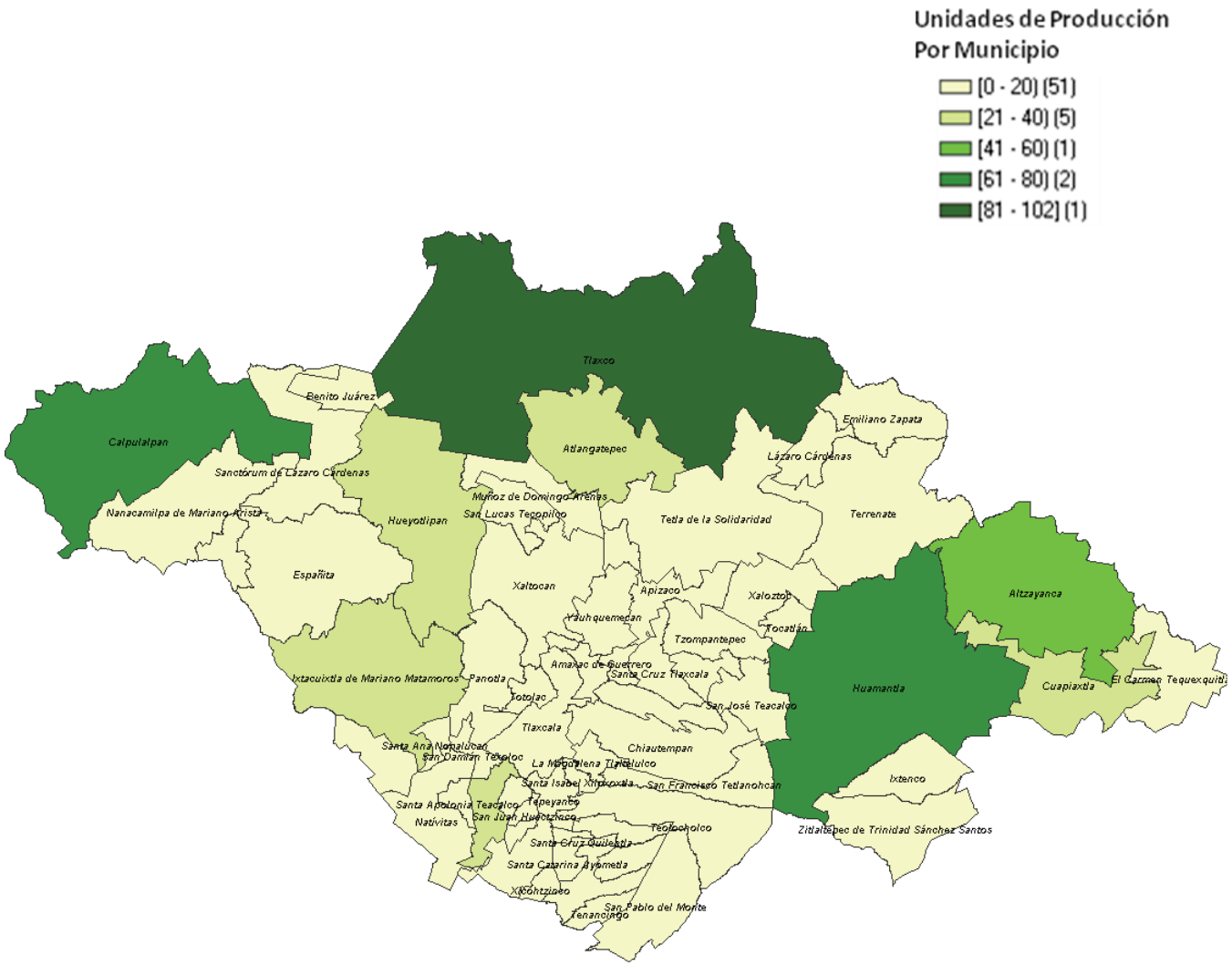
- El 46.88% de los responsables de las UP cuentan con al menos un tractor.

- En este estrato sólo 8.07% de los productores son mujeres.

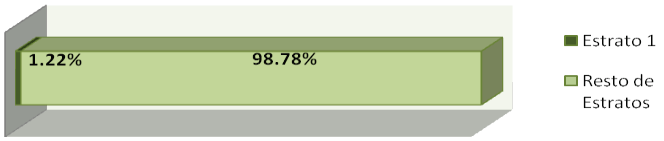
- El promedio de ingresos anuales por la actividad agropecuaria y forestal es el más alto (167762 pesos).

- La mayor cantidad de productores de este estrato se localiza en los municipios: Altzayanca, Calpulalpan, Huamantla y Tlaxco.

Distribución del Estrato 1 según el número de Unidades de Producción por Municipio



Distribución de unidades de producción del estrato 1 según porcentaje que representa en el Estado.



Estrato 2, *Unidades de Producción con alta representatividad en la superficie agrícola estatal*

- Contiene el 21.27% del total de unidades de producción y son responsables del 29.01% de la superficie agrícola de la Entidad, quedando por arriba de las grandes unidades de producción del estrato 1, situación que se explica al ser el segundo estrato con más UP.

- Son UP pequeñas, el 50 % de ellas tienen una superficie total entre 0.61 y 6.50 ha; siendo el promedio de 4.3 ha.

- La superficie es casi totalmente de temporal (99.73 %).

- El rendimiento promedio es de 2.6 ton/ha. Ubicándose en el cuarto lugar en este rubro, pues es mayor el aprovechamiento en los estratos 1,4 y 5, a pesar de tener casi el 30% de la superficie agrícola del Estado.

- El promedio de cabezas de ganado es de 11.28 animales que incluyen ganado bovino, porcino, ovino y caprino.

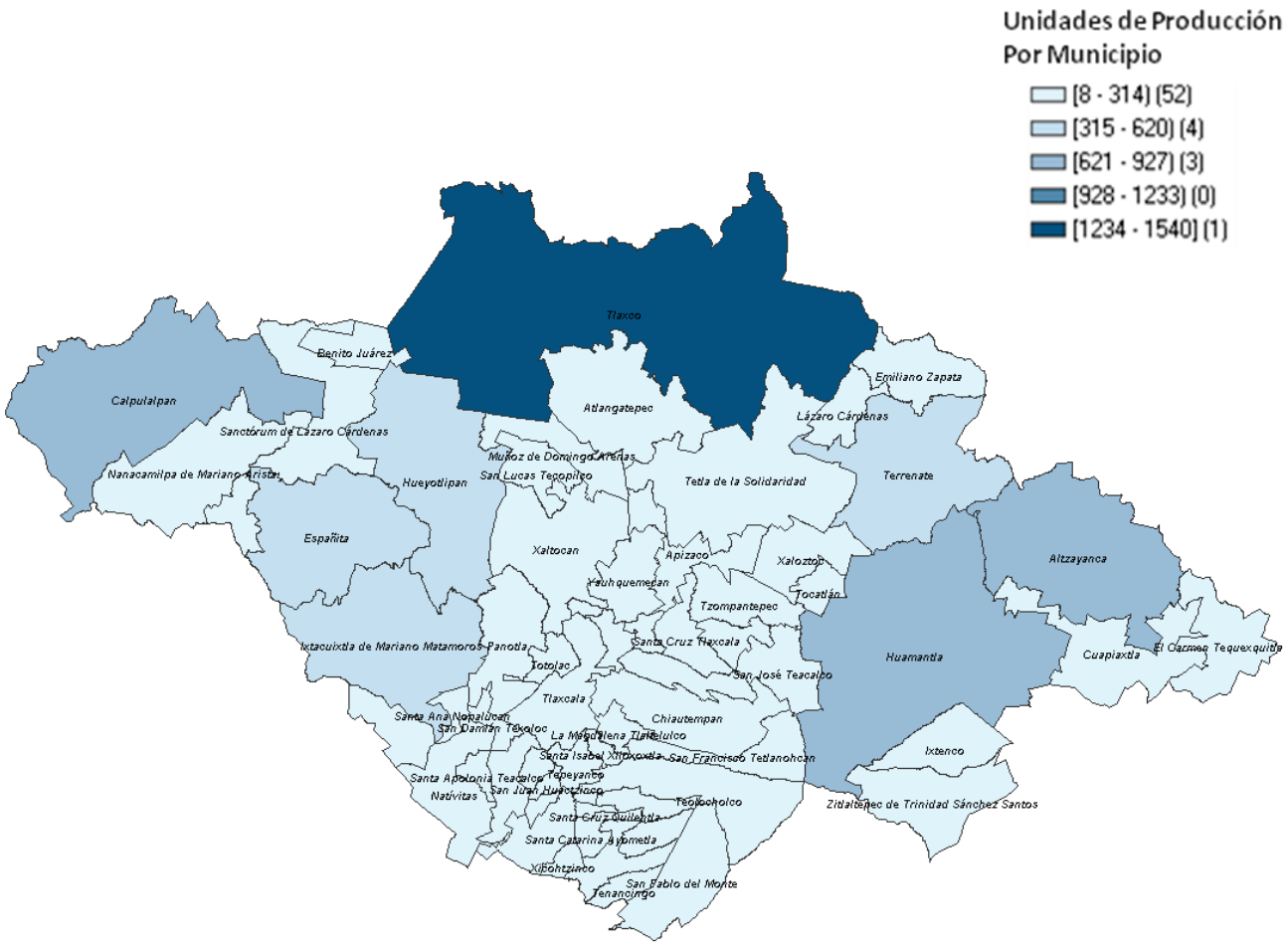
- El 9.72 % de los responsables de las UP cuentan con al menos un tractor.

- En este estrato el 12.87 % de los productores son mujeres.

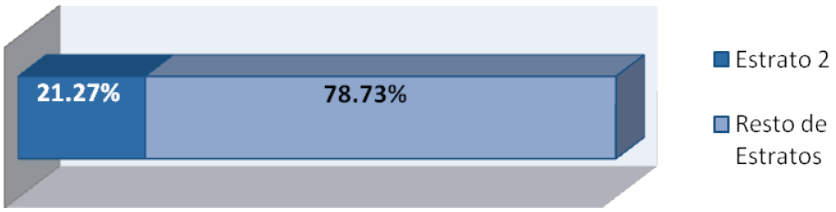
- El promedio de ingresos anuales por la actividad agropecuaria y forestal es de 6962 pesos.

- La mayor cantidad de productores de este estrato se localiza en los municipios: Altzayanca, Calpulalpan, El Carmen Tequexquitla, Españita, Huamantla, Hueyotlipan, Ixtacuixtla, Nanacamilpa, Terrenate, Tetla de la Solidaridad, Tetlatlahuca y Tlaxco.

Distribución del Estrato 2 según el número de Unidades de Producción por Municipio



Distribución de unidades de producción del Estrato 2 según porcentaje que representa en el Estado.



Estrato 3, *Unidades de Producción con el menor rendimiento promedio de producción*

- Contiene el 6.24 % del total de UP y son responsables del 7.57 % de la superficie agrícola de la Entidad.

- Son UP pequeñas, el 50 % de ellas tienen una superficie total entre 1.75 y 5.2 has.; siendo el promedio de 4 ha.

- La superficie es casi totalmente de temporal (99.8 %).

- El rendimiento promedio es de 1.18 ton/ha. Este es el más bajo de los cinco estratos**.** También es el estrato donde menos se usa fertilizante en proporción con su superficie sembrada y el segundo más bajo en el uso de semilla certificada.

- El promedio de cabezas de ganado es de 4.04 animales que incluyen ganado bovino, porcino, ovino y caprino.

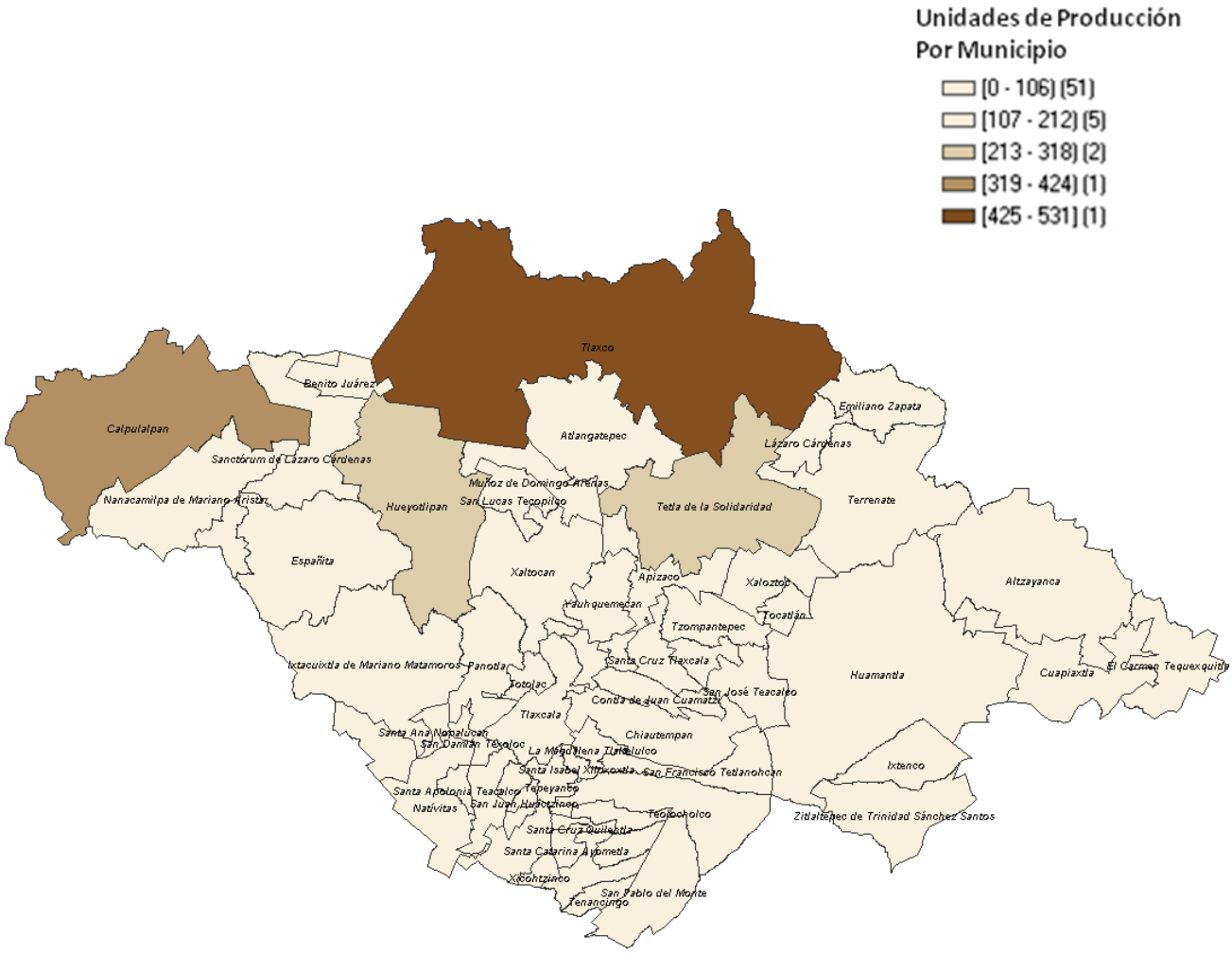
- Apenas el 5.08 % de los responsables de las UP cuentan con al menos un tractor.

- En este estrato el 18.1 % de los productores son mujeres, que es el segundo porcentaje más alto.

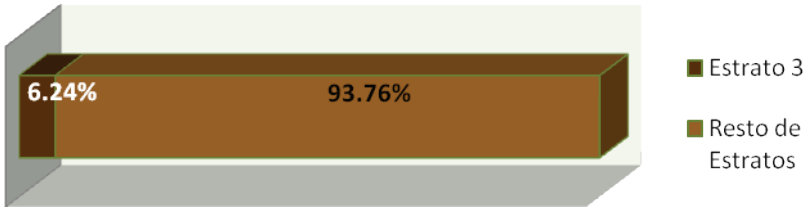
- El promedio de ingresos anuales por la actividad agropecuaria y forestal es de 2654 pesos. Este es el segundo más bajo de los cinco estratos.

- Los municipios con mayor cantidad de productores de este estrato son: Altzayanca, Calpulalpan, Hueyotlipan, Tetla de la Solidaridad, y Tlaxco.

Distribución del Estrato 3 según el número de Unidades de Producción por Municipio



Distribución de unidades de producción del estrato 3 según porcentaje que representa en el Estado.



Estrato 4, *Unidades de Producción con predominio de superficie de riego y mejor aprovechamiento promedio*

- Contiene el 8.85 % del total de UP y son responsables del 8.34 % de la superficie agrícola de la Entidad.

- Son UP pequeñas, el 50 % de ellas tienen una superficie total entre 1.02 y 3.75 ha; siendo el promedio de 3 ha.

- Es el único estrato donde predomina la superficie de riego (51.65 %) sobre la de temporal (48.35 %).

- Este estrato también destaca en el rendimiento promedio que es de 16.73 ton/ha. El más alto de los cinco.

- El promedio de cabezas de ganado es de 4.74 animales que incluyen ganado bovino, porcino, ovino y caprino.

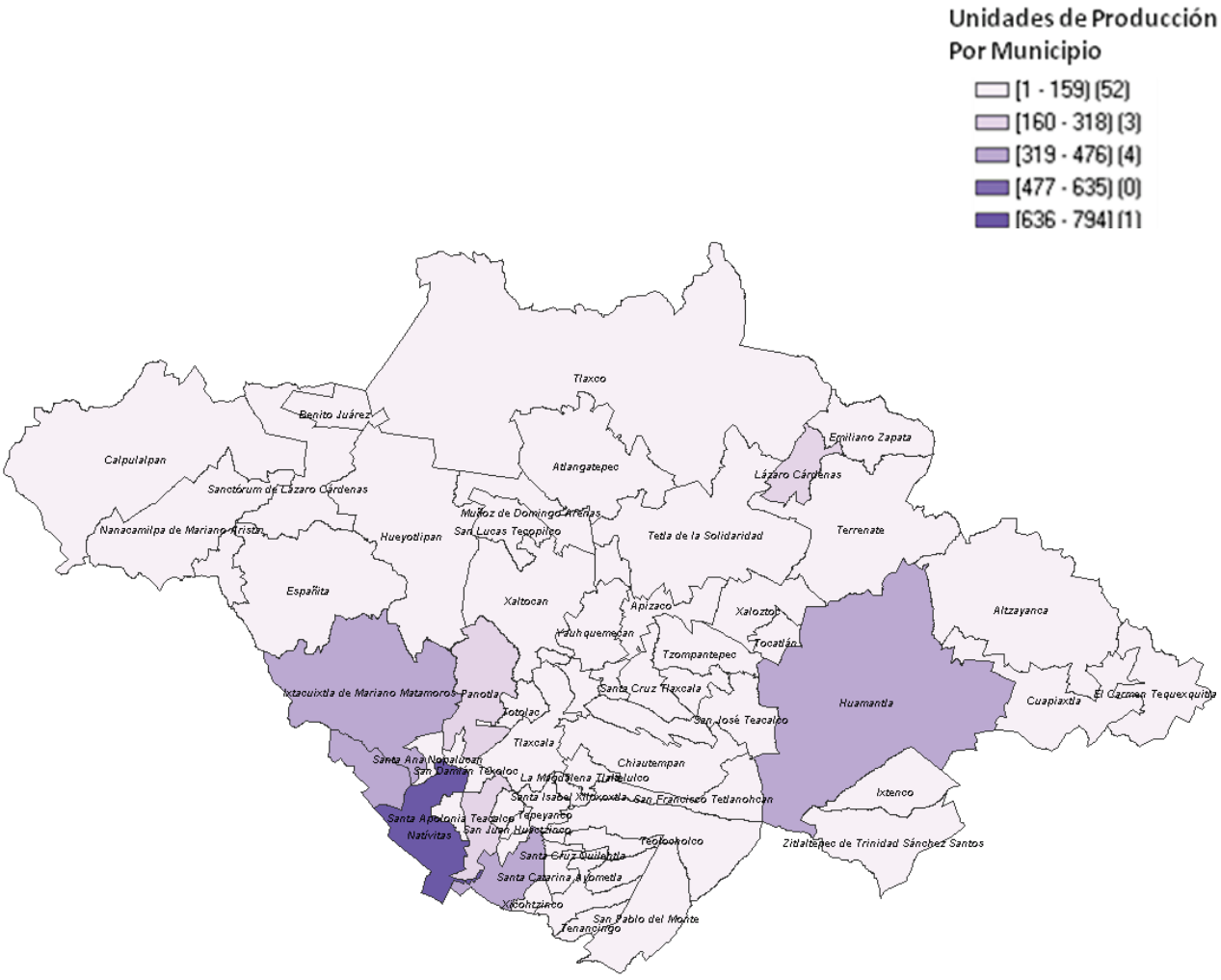
- Sólo 8.21 % de los responsables de las UP cuentan con al menos un tractor.

- En este estrato el 14.52 % de los productores son mujeres.

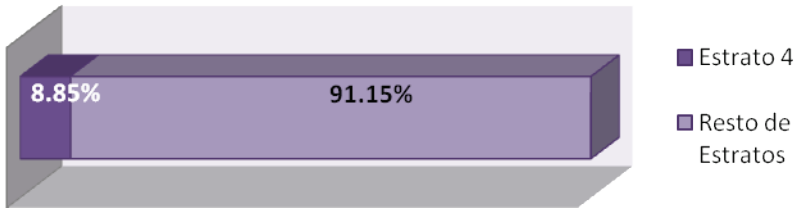
- El promedio de ingresos anuales por la actividad agropecuaria y forestal es de 5377 pesos.

- Los municipios con mayor cantidad de productores de este estrato son: Huamantla, Ixtacuixtla de Mariano Matamoros, Tepetitla de Lardizabal, Natívitas, Tetlatlahuca y Zacatelco.

Distribución del Estrato 4 según el número de Unidades de Producción por Municipio



Distribución de unidades de producción del estrato 4 según porcentaje que representa en el Estado.



Estrato 5, *Predominante en la Entidad*, las *Unidades de Producción más pequeñas*

- Este estrato contiene el 62.43 % del total de UP. A pesar de ser el más grande, los productores de este estrato sólo son responsables del 33.18 % de la superficie agrícola de la Entidad.

- Son las UP más pequeñas, el 50 % de ellas tienen una superficie total entre 0.23 y 2.55 ha; siendo el promedio de 1.7 ha.

- La superficie es casi totalmente de temporal (99.04 %).

- El rendimiento promedio es de 2.93 ton/ha.

- El promedio de cabezas de ganado es el más bajo de los estratos (0.36 animales, que incluyen ganado bovino, porcino, ovino y caprino).

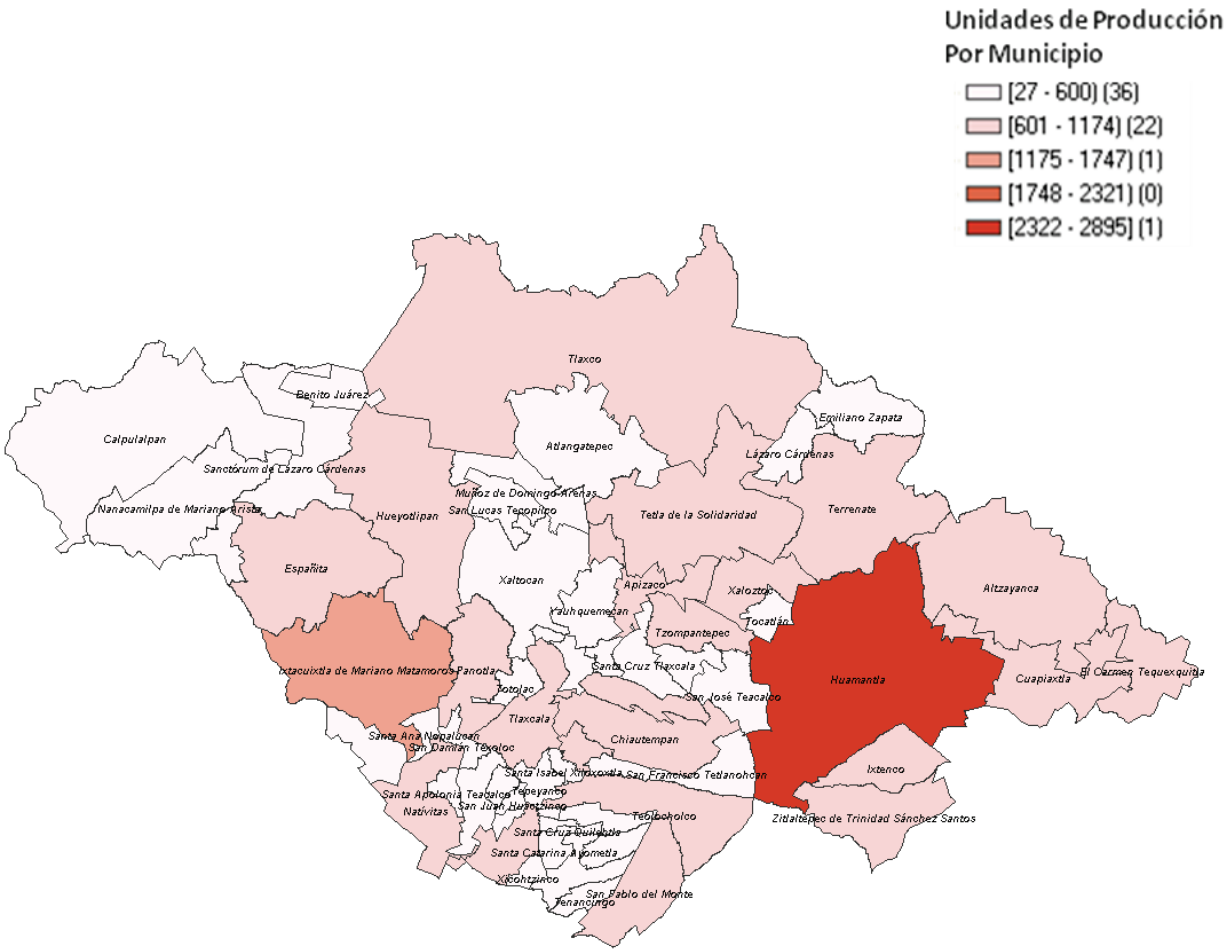
- El porcentaje de responsables de las UP que cuentan con al menos un tractor es el más bajo de los estratos (1.63 %).

- El estrato tiene el porcentaje más alto de mujeres responsables de las UP (19.8 %).

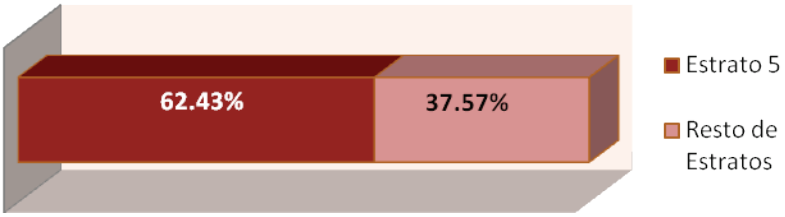
- El promedio de ingresos anuales por la actividad agropecuaria y forestal es de 879 pesos. Es el más bajo de los cinco estratos.

- Es el estrato predominante en 90 % de los municipios de la Entidad.

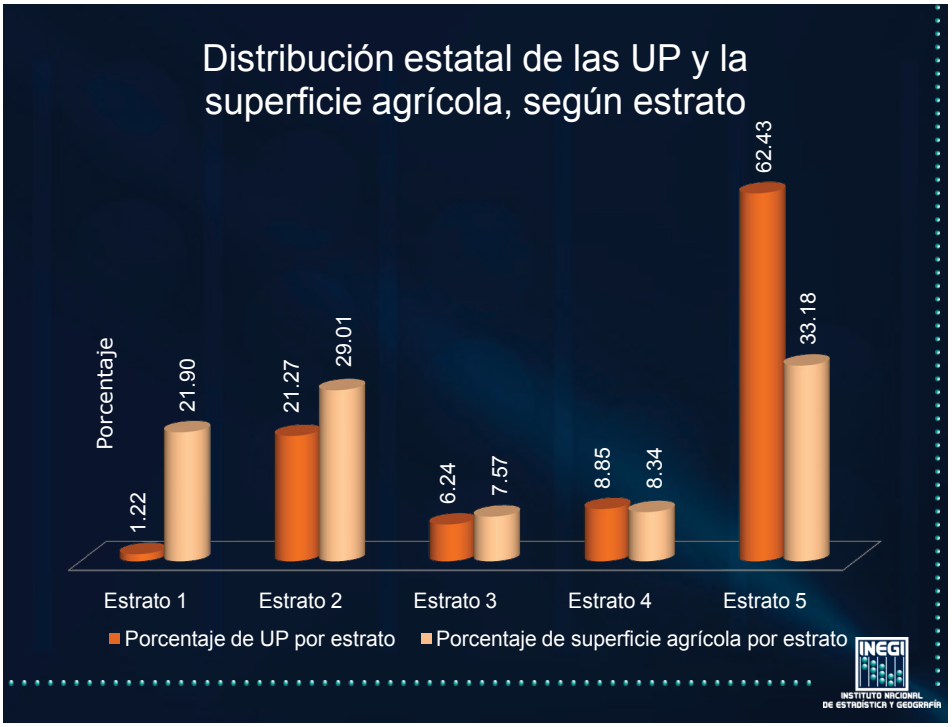
Distribución del Estrato 5 según el número de Unidades de Producción por Municipio

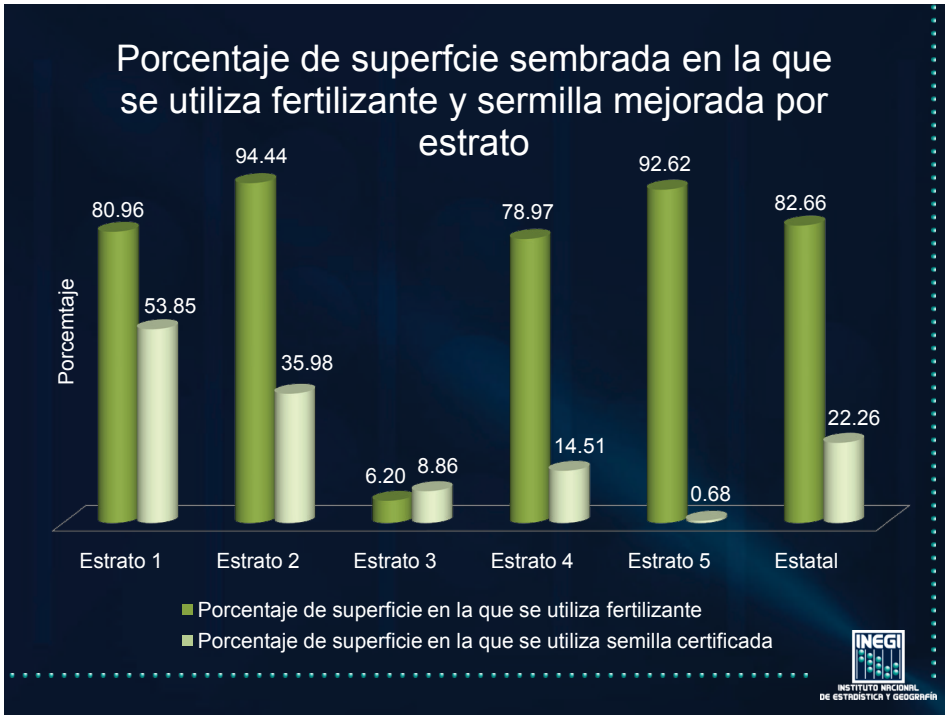


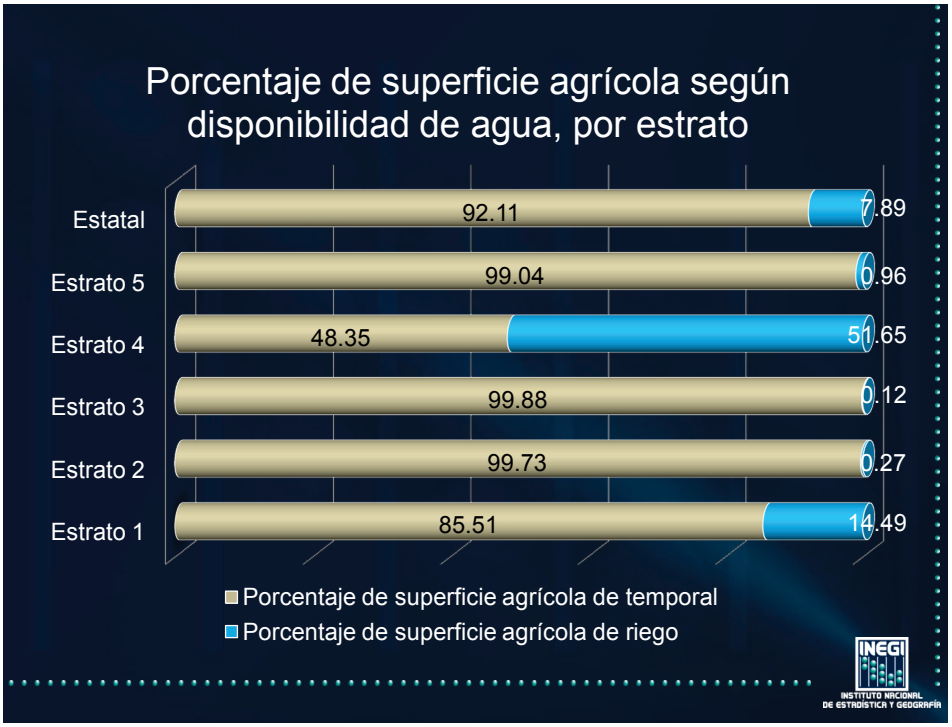
Distribución de unidades de producción del Estrato 5 según porcentaje que representa en el Estado.



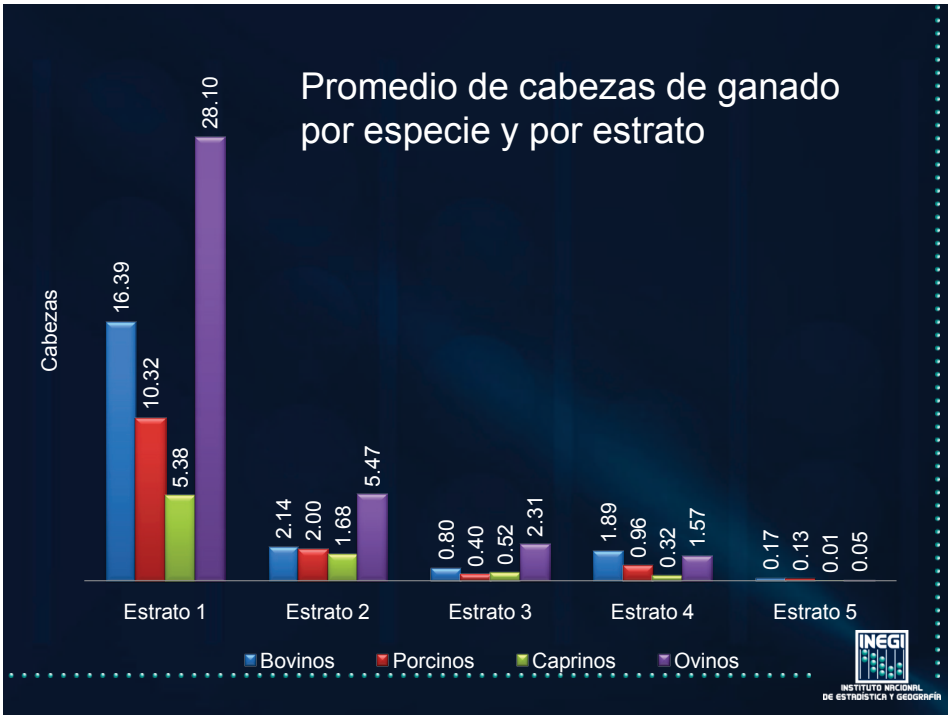
Gráficos comparativos por estrato


Distribución estatal de las UP y la superficie agrícola, según estrato
Porcentaje
1.22
21.90
21.27
29.01
6.24
7.57
8.85
8.34
62.43
33.18
Estrato 1
Estrato 2
Estrato 3
Estrato 4
Estrato 5
Porcentaje de UP por estrato
Porcentaje de superficie agrícola por estrato
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA


Porcentaje de superfcie sembrada en la que se utiliza fertilizante y sermilla mejorada por estrato
Porcemtaje
80.96
53.85
94.44
35.98
6.20
8.86
78.97
14.51
92.62
0.68
82.66
22.26
Estrato 1
Estrato 2
Estrato 3
Estrato 4
Estrato 5
Estatal
Porcentaje de superficie en la que se utiliza fertilizante
Porcentaje de superficie en la que se utiliza semilla certificada
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA


Porcentaje de superficie agrícola según disponibilidad de agua, por estrato
Estatal
Estrato 5
Estrato 4
Estrato 3
Estrato 2
Estrato 1
92.11
7.89
99.04
0.96
48.35
51.65
99.88
0.12
99.73
0.27
85.51
14.49
Porcentaje de superficie agrícola de temporal
Porcentaje de superficie agrícola de riego
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA


Rendimiento promedio de las unidades de producción por estrato
ton/ha
9.06
2.60
1.18
16.73
2.93
Estrato 1
Estrato 2
Estrato 3
Estrato 4
Estrato 5
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA


Promedio de cabezas de ganado por especie y por estrato
Cabezas
16.39
10.32
5.38
28.10
2.14
2.00
1.68
5.47
0.80
0.40
0.52
2.31
1.89
0.96
0.32
1.57
0.17
0.13
0.01
0.05
Estrato 1
Estrato 2
Estrato 3
Estrato 4
Estrato 5
Bovinos
Porcinos
Caprinos
Ovinos
INEGI
INSTITUTO NACIONAL DE ESTADISTICA Y GEOGRAFIA


Promedio de ingresos anuales que provienen de la actividad agropecuaria por estrato
Pesos
167 762
6 962
2 654
5 377
879
Estrato 1
Estrato 2
Estrato 3
Estrato 4
Estrato 5
INEGI
INSTITUTO NACIONAL DE ESTADISTICA Y GEOGRAFIA

Disponibilidad de tractor en las UP según estrato

Estrato	UP que no tienen tractor		UP con al menos un tractor	
	Cantidad	%	Cantidad	%
1	349	53.12	308	46.88
2	10 364	90.28	1 116	9.72
3	3 194	94.92	171	5.08
4	4 382	91.79	392	8.21
5	33 142	98.37	549	1.63
Estatal	**51 431**	**95.30**	**2 536**	**4.70**

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Sexo del responsable de la UP por estrato

Estrato	Responsables Hombres		Responsables Mujeres	
	Cantidad	%	Cantidad	%
1	604	91.93	53	8.07
2	10 002	87.13	1 478	12.87
3	2 756	81.90	609	18.10
4	4 081	85.48	693	14.52
5	27 019	80.20	6 672	19.80
Estatal	**44 462**	**82.39**	**9 505**	**17.61**

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

CONCLUSIONES

La diversidad del sector rural demostrada en este trabajo, requiere del diseño de políticas públicas que tomen en cuenta las características de su población objetivo para proporcionar a los distintos tipos de beneficiarios los bienes y servicios que puedan tener mayor incidencia en su desarrollo y con ello maximizar el impacto de los recursos públicos.

Ante la diversidad de esta población objetivo, resulta imprescindible que los gobiernos estatales tomen decisiones con respecto a los apoyos que se otorgarán a cada tipo de beneficiario, a fin de maximizar la rentabilidad económica y social de los recursos públicos. Esto es, diseñar esquemas de apoyos diferenciados según las capacidades y características de producción de cada estrato generado.

Durante el estudio se logró obtener una herramienta estadística, e informática en una siguiente etapa, que cumple con la necesidad propuesta de estratificar los productores agropecuarios del estado y a partir de esto generar un modelo de asignación automática, el cual resultó muy efectivo al pronosticar correctamente el 93.4% de los casos, trabajando con una menor cantidad de información al tomar sólo cuatro componentes principales de la misma.

El comportamiento de aglomeración, responde muy cercanamente a la ubicación geográfica de las unidades de producción, situación que de antemano se planteó puesto que unidades con condiciones geográficas semejantes tendrán condiciones climáticas parecidas y esto hará que su cantidad y calidad de producción se asemeje. El uso de la metodología estadística, y las herramientas de espacialización confirmaron este planteamiento. A su vez, es factible el uso del desarrollo hasta aquí establecido en un proyecto total de empadronamiento de productores agropecuarios con la asignación de estrato de forma inmediata.

BIBLIOGRAFÍA

RICE J A. (2007). Mathematical Satatistics and Data Analysis, Belmont CA. ThompsonBrooks/Cole Press.

RENCHER A C.(2002). Methods of Multivariate Analysis (second edition). USA. Wiley interscience.

JOHNSON R – Wichern D (2007). Applied Multivariate Statistical Analysis (6th Edition). Prentice Hall.

BISHOP C.(2006) Pattern Recognition and Machine Learning. New York. Springer

CHRISTENSEN R(1997). Linear Models for Multivariate, Time Series, and Spatial Data. Springer.

AGRESTI A (2002). Categorical Data Analysis(second edition). Wiley-Interscience.

TABACHNICK B, FIDELL L(2006) Using Multivariate Statistics (5th Edition). Allyn & Bacon

FIELD A(2005) . Discovering Statistics Using SPSS(2nd edition). Sage Publications Ltd

CRAWLEY M(2007). The R Book. Wiley

DUDA R, HART P, STORK D(2000). Pattern Classification (2nd Edition). Wiley-Interscience.