Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Estadística espacial: una aplicación para identificar los focos de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo tecnológico

TESIS

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Frank Josué CASTILLO ISISOLA

ASESOR

Mg. María Estela PONCE ARUNERI

Lima, Perú

2019



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Castillo, F. (2019). *Estadística espacial: una aplicación para identificar los focos de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo tecnológico*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Frank Josué Castillo Isisola
Tipo de documento de identidad	DNI
Número de documento de identidad	46936667
URL de ORCID	
Datos de asesor	
Nombres y apellidos	María Estela Ponce Aruneri
Tipo de documento de identidad	DNI
Número de documento de identidad	10010612
URL de ORCID	https://orcid.org/0000-0002-3091-5741
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Ysela Dominga Agüero Palacios
Tipo de documento	DNI
Número de documento de identidad	09059647
Miembro del jurado 1	
Nombres y apellidos	Emma Norma Cambillo Moyano
Tipo de documento	DNI
Número de documento de identidad	15377390
Datos de investigación	
Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad (Empresas, Instituciones, Poblaciones locales, regionales y nacionales).

Grupo de investigación	No aplica
Agencia de financiamiento	Perú. Universidad Nacional Mayor de San Marcos. Vicerrectorado de Investigación y Posgrado. Programa de Promoción de Tesis de Pregrado. B17140924a-PTPGRADO.
Ubicación geográfica de la investigación	Edificio: Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Calle: Av. Universitaria cruce con Av. Venezuela cuadra 34 Latitud: -12.056423 Longitud: -77.084332
Año o rango de años en que se realizó la investigación	Marzo 2016 - octubre 2018
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03 Agricultura https://purl.org/pe-repo/ocde/ford#4.01.01

Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

FACULTAD DE CIENCIAS MATEMÁTICAS
ESCUELA PROFESIONAL DE ESTADÍSTICA

ACTA DE SUSTENTACIÓN DE TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA

En la Ciudad Universitaria, Facultad de Ciencias Matemáticas, siendo las 17:00 horas del día 14 de MARZO del año 2019, se reunieron los docentes designados como miembros del Jurado:

Mg. Emma Norma Cambillo Moyano	(Presidenta)
Mg. Ysela Dominga Agüero Palacios	(Miembro)
Mg. María Estela Ponce Aruneri	(Miembro Asesor)

Para la sustentación de la Tesis intitulada **"ESTADÍSTICA ESPACIAL: UNA APLICACIÓN PARA IDENTIFICAR LOS FOCOS DE LOS PEQUEÑOS PRODUCTORES AGROPECUARIOS PERUANOS CON BAJO NIVEL DE DESARROLLO TECNOLÓGICO"**, presentada por el Bachiller FRANK JOSUÉ CASTILLO ISISOLA, para obtener el Título Profesional de Licenciado en Estadística.

Luego de la exposición de la Tesis, la Presidenta invitó al expositor a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del jurado, el expositor mereció la aprobación. SOBRESALIENTE., con un calificativo promedio de DIECISIETE (17) (letras y números).

A continuación los miembros del jurado, dan manifiesto que el participante Bachiller FRANK JOSUÉ CASTILLO ISISOLA, en virtud de haber aprobado la sustentación de su tesis, será propuesto para que se le otorgue el Título Profesional de Licenciado en Estadística.

Siendo las 18:10 horas, se levantó la Sesión, firmando para constancia la presente Acta en tres (3) copias originales.

Mg. Ysela Dominga Agüero Palacios
Miembro

Mg. María Estela Ponce Aruneri
Miembro Asesor

Mg. Emma Norma Cambillo Moyano
Presidenta

UNIVERSIDAD NACIONAL MAYOR DE
SAN MARCOS
Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, María Estela Ponce Aruneri. en mi condición de asesor acreditado con RD N° 0282/FCM-D/2018 de la tesis, cuyo título es "*Estadística Espacial Una aplicación para identificarlos focos de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo tecnológico - 2026*" presentado por el bachiller *Frank Josué Castillo Isisola* para optar el título profesional de Licenciado en Estadística.

CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 11% de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional.**
Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título profesional correspondiente.

Lima,23 de setiembre 2026

Firma del Asesor: 

DNI : 10010612

DEDICATORIA

Al ser supremo, por darme la oportunidad de ver la luz cada día.

A mi madre, por su amor incondicional y su apoyo incansable

A Yuan, por su amor, paciencia y comprensión durante este proceso

A mi tía, abuela y primo, por siempre alentarme en cada paso.

A mis amigos, por su incondicional apoyo.

AGRADECIMIENTOS

A la Mg. María Estela Ponce Aruneri por su asesoramiento y paciencia en la realización de la tesis.

A la Mg. Rosario Bullón, por sus consejos y apoyo en mi desarrollo profesional.

A cada uno de mis profesores por sus enseñanzas a lo largo de la carrera.

Al Vicerrectorado de Investigación y Posgrado por el financiamiento que hizo posible la realización del trabajo,

RESUMEN

“ESTADISTICA ESPACIAL: UNA APLICACIÓN PARA IDENTIFICAR LOS FOCOS DE LOS PEQUEÑOS PRODUCTORES AGROPECUARIOS PERUANOS CON BAJO NIVEL DE DESARROLLO TECNOLOGICO”

Br. Frank Josué Castillo Isisola

Octubre 2018

Asesora: Mg. María Estela Ponce Aruneri

Título: Licenciado en Estadística

Las comunidades rurales del país se concentran mayormente en la sierra y selva del país y constituyen la población más vulnerable debido a las desventajas que enfrentan como la pobreza, carencia de servicios básicos y de apoyo por parte del estado, aunque son parte importante en nuestro país ya que nos proporcionan los alimentos que consumimos diariamente. La principal actividad económica de estas comunidades rurales es la agropecuaria; sin embargo, esta no se ha desarrollado lo suficiente como para que las comunidades salgan de la condición de pobreza en que se encuentran. El objetivo de esta investigación es identificar en qué zonas del país se presenta mayor concentración de pequeños productores agropecuarios con bajos niveles de desarrollo mediante el cálculo de un Índice de Adopción Tecnológica utilizando la técnica estadística de análisis de correspondencias múltiples.

Para determinar los focos pequeños productores agropecuarios sea aplicaron densidades focales de kernel tomando como referencia los Sectores de Enumeración Agropecuaria los cuales tienen agricultores con adopción tecnológica muy baja y condiciones vulnerables según el agrupamiento mediante redes neuronales de Kohonen. Se obtuvo un total de ocho focos de alta densidad de pequeños productores agropecuarios según la estimación de densidad con el Kernel cuártico y su principal característica es el bajo nivel de adopción tecnológica debido al manejo de tecnología convencional como uso de fertilizantes e insecticidas químicos, vacunación artificial y dosificación de animales. Sus principales carencias son la falta de acceso a la información, capacitación y asistencia técnica.

Palabras Clave: *Densidades Focales de Kernel, Índice de Adopción Tecnológica, Redes Neuronales de Kohonen**,** Análisis de Correspondencias, pobreza rural,*

ABSTRACT

"SPATIAL STATISTICS: AN APPLICATION TO IDENTIFY THE HIGH-DENSITY MAP OF SMALL PERUVIAN FARMERS AND LIVE STOCKERS WITH LOW LEVELS OF TECHNOLOGICAL DEVELOPMENT"

Br. Frank Josué Castillo Isisola

October 2018

Advisor: Msc. María Estela Ponce Aruneri

Degree: Graduate in Statistics

The country's rural communities are mostly concentrated in the highlands and jungle of the country and they constitute the most vulnerable population due to the disadvantages they face such as poverty, lack of basic services and support from the state although they are an important part of our country because they provide us the food that we consume daily. The main economic activity of these rural communities is the agriculture, but this activity has not been developed enough to get out of the poverty in which these rural communities find themselves. The research goal is to identify in which country areas are a greater concentration of small farmers with low development levels by calculating a Technological Adoption Index using the multiple correspondence analysis technique.

To determinate the small farmers spotlights the Kernel focal densities was applied taking as reference the agricultural enumeration sector which have farmers with very low adoption of technology and vulnerable conditions according to the clustering by Kohonen self-organized maps. Eight high density spotlights of small farmers were obtained according to the density estimation with the quartic kernel and its main characteristic is the low level of technological adoption due to the use of conventional technology such as the use of chemical fertilizers and insecticides, artificial vaccination and dosage of animals. Its main shortcomings are the lack of access to information, training and technical assistance.

Palabras Clave: *Kernel Density Map, Technological Adoption Index, Self-Organized Kohonen Map**,** Correspondence Analysis, Rural poverty.*

ÍNDICE

CAPÍTULO I: INTRODUCCIÓN 2
1.1. Situación problemática 2
1.2. Formulación del problema 3
1.2.1. Problema General 3
1.2.2. Problemas Específicos 3
1.3. Objetivos de la investigación 3
1.3.1. Objetivo general 4
1.3.2. Objetivos específicos 4
1.4. Justificación de la investigación 4
1.5. Originalidad de la Investigación 5
1.6. Limitaciones de la investigación 6
1.7. Hipótesis 6
CAPÍTULO II: MARCO TEÓRICO 8
2.1. Antecedentes de la investigación 8
2.1.1. Antecedentes Nacionales 8
2.1.2. Antecedentes Internacionales 10
2.2. Bases teóricas 15
2.2.1. Análisis de Correspondencias 15
2.2.1.1. Introducción 15
2.2.1.2. Análisis de Correspondencias Múltiples 15
2.2.2. Definición del Índice Sintético 19
2.2.3. Índice de Adopción Tecnológica 19
2.2.4. Análisis de Agrupamientos 20
2.2.4.1. Determinación del número óptimo de Agrupamientos 20
2.2.4.2. Validación estadística de los grupos 22
2.2.4.3. Validación interna de Grupos 22
2.2.5. Mapas Auto Organizados de Kohonen. 25
2.2.5.1. Arquitectura de los Mapas Auto organizados de Kohonen 26
2.2.5.2. Algoritmo de aprendizaje incremental de los Mapas Autoorganizados 26
2.2.5.3. Preprocesamiento de datos 30
2.2.6. Función de Densidad Kernel 31
2.2.7. Estadística Espacial 34
2.2.8. Patrones de Puntos Espaciales 36

2.2.8.1. Estacionariedad e isotropía 38
2.2.8.2. Procesos de puntos estacionarios 38
2.2.8.3. Intensidad 38
2.2.8.4. Aleatoriedad Espacial Completa. 40
2.2.8.5. Efecto Frontera 41
2.2.8.6. Distancia al vecino más cercano 41
2.2.8.7. Estimación de la intensidad 42
2.2.8.8. Estimación de la función G 43
2.2.8.9. Estimación de la función L 44
2.2.9. Definiciones Técnicas de Agricultura. 46
CAPITULO III: METODOLOGÍA 48
3.1. Fuente de datos 48
3.2. Población y unidad de análisis 48
3.3. Variables 49
3.4. Procedimiento de análisis 50
3.5. Software utilizado 51
CAPITULO IV: RESULTADOS Y DISCUSIÓN 52
4.1. Índice de Adopción Tecnológica 52
4.2. Agrupamiento de los pequeños productores Agropecuarios 58
4.3. Determinación de los focos de alta densidad 63
4.4. Focos de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo tecnológico 66
4.5. Hipótesis 69
CONCLUSIONES 72
RECOMENDACIONES 74
REFERENCIAS BIBLIOGRÁFICAS 75
ANEXOS 80
Anexo 1. Glosario de términos 80
Anexo 2. Matriz de consistencia 82
Anexo 3. Matriz de operacionalización de variables. 83
Anexo 4. Análisis geométrico de datos 86
Anexo 5. Análisis de Correspondencias Simples 91
Anexo 6. Métodos de agrupamiento 94
Anexo 7. MANOVA (Análisis multivariante de la varianza) 95
Anexo 8. Cálculos de resultados mostrados en subsección 4.5 97

Anexo 9. Verificación de veracidad del índice. 99
Anexo 10. Codificación en R..... 100

LISTA DE FIGURAS

Figura 1: Estructura del Mapa Autoorganizado 26
Figura 2: Cuadrícula hexagonal de clasificación de modelos de espectro del habla natural del idioma finlandés. 27
Figura 3: Ejemplos de vecindad topológica. 28
Figura 4: Representación de neuronas de salida luego de 20000 iteraciones. (a) Vectores de entrada normalizados (b) vectores de entrada sin normalizar. 31
Figura 5: Estimación de la función de densidad Kernel Gaussiano para 6 puntos 32
Figura 6: Estimación mediante Kernel Gaussiano mostrando Kernels individuales con h=0.4 (izquierda) y h=0.2 (derecha) 33
Figura 7: Representación de la aplicación del Kernel bivariado sobre una región R. 34
Figura 8: Enfoques de la Estadística Espacial (de izquierda a derecha): Geoestadística, Enmallados y Patrones Puntuales. 35
Figura 9: Patrones de puntos espaciales: aleatorio (1), regular (2) y agrupado (3). 37
Figura 10: (a) Función K estimada de patrones regulares de grupos, (b) Diagrama de patrones regulares de grupos, (c) Función K estimada de grupos de patrones regulares, (d) Diagrama de grupos de patrones regulares. 45
Figura 11: Contribuciones de las variables sobre cada dimensión 54
Figura 12: Nube de puntos de modalidades (a) y nube de variables (b) en base a las dimensiones 1 y 2 56
Figura 13: Función Densidad del Índice de Adopción Tecnológica de los productores agropecuarios. CENAGRO 2012 57
Figura 14: Gráfico del ancho de silueta para determinar el número óptimo de clústeres 59
Figura 15: Mapa Autoorganizado de Kohonen dividido según número óptimo de clúster 60
Figura 16: Mapa de las ubicaciones de los SEA de pequeños productores agropecuarios con bajo nivel de desarrollo 63
Figura 17: Ubicación de los eventos en ventana reducida (a), Estimación de densidad intensidad mediante kernel cuártico en tres dimensiones (b), estimación de la intensidad mediante kernel cuártico en dos dimensiones (c), gráfico de contorno de densidad (d). 64
Figura 18: Función G estimada (a) y Función L estimada (b) 65
Figura 19: Mapa del Perú identificando los focos de alta densidad de SEA con agricultores que tienen muy bajos niveles de adopción de tecnologías y baja calidad de vida. 66
Figura 20: Enfoque del Análisis Geométrico de Datos 87
Figura 21: Metodología del análisis geométrico de datos para análisis de datos estructurados 90
Figura 22: Metodología del uso de la estadística inferencial en el AGD. 90

LISTA DE TABLAS

Tabla 1: Resultados del ACM: valores propios, porcentaje de varianza y porcentaje de varianza acumulada 53
Tabla 2: Categorías del nivel de Adopción Tecnológica 58
Tabla 3: Principales características de los clústeres generados 62
Tabla 4: Resultados de la prueba Z para la proporción de productores con bajo desarrollo tecnológico ..... **¡Error! Marcador no definido.**
Tabla 5: Resultados de la prueba Z para la proporción de productores con bajo desarrollo tecnológico ..... **¡Error! Marcador no definido.**
Tabla 6: Tabla de contingencia de las variables X e Y. 91

CAPÍTULO I: INTRODUCCIÓN

1.1. Situación problemática

En abril del 2014, al menos el 85% de los 570 millones de granjas agrícolas del mundo tenían una administración individual o familiar y dependían sobre todo de mano de obra de los miembros de la familia. La gran mayoría de las granjas del mundo son pequeñas o muy pequeñas y de menos de 2 hectáreas representando el 84% de todas las granjas y controlando solamente el 12% de las tierras agrícolas. Las explotaciones familiares producen más del 80% de los alimentos en el mundo en cuanto al valor, confirmando la vital importancia de la agricultura familiar en la seguridad alimentaria mundial actual y para las futuras generaciones. (Lowder, Skoet, & Singh, 2014).

La política agraria sobre las agriculturas familiares ha reconocido la diversidad y complejidad de los retos que aquejan a este grupo cuyo desarrollo es clave para acabar con la pobreza rural, desnutrición crónica, calidad educativa y lograr sistemas agrícolas más eficientes que permitan el crecimiento sostenible (Meynard Vivar, 2015).

En Perú, el IV Censo Nacional Agropecuario 2012 contó con una población de pequeños productores agropecuarios de aproximadamente 1’800’000. El Censo impulso iniciativas de investigación cuyos resultados buscaron contribuir en la formulación de las políticas, planes y programas de desarrollo del sector agropecuario en nuestro país. (Escobal & Armas, 2015)

El desarrollo agropecuario es la base del desarrollo en las áreas rurales de nuestro país y su principal fuente de ingresos. El sector agropecuario representó el 6.42% del PBI Nacional para el año 2016, cifra que año a año crece y decrece, a pesar de que su valor real en soles aumenta (INEI, 2018)

Uno de los objetivos específicos de la política Nacional Agraria, presentado por el Ministerio de agricultura y Riego en el 2016 es incrementar la competitividad agraria y la inserción a los mercados, con énfasis en el pequeño productor agrario. (MINAGRI, 2016)

1.2. Formulación del problema

1.2.1. Problema General

¿Cuáles son los focos y patrones de comportamiento de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo tecnológico en el año 2012?

1.2.2. Problemas Específicos

- ¿Cuál es nivel de adopción tecnológica de los pequeños productores agropecuarios en el año 2012?
- ¿Cuáles son los focos de concentración de los pequeños productores agropecuarios en el año 2012?
- ¿Cuáles son las características de los pequeños productores agropecuarios de acuerdo con su nivel de adopción tecnológica en focos de alta densidad en el año 2012?

1.3. Objetivos de la investigación

1.3.1. Objetivo general

Identificar los focos de pequeños productores agropecuarios peruanos con bajo nivel de desarrollo tecnológico y sus patrones de comportamiento en el año 2012.

1.3.2. Objetivos específicos

- Calcular el nivel de adopción tecnológica de los pequeños productores agropecuarios a nivel nacional utilizando Análisis de Correspondencias Múltiples.

- Identificar los focos de los pequeños productores agropecuarios para el año 2012 mediante densidades focales de Kernel.

- Describir las características de pequeños productores agropecuarios en el 2012 según su nivel de adopción tecnológica en focos de alta densidad.

1.4. Justificación de la investigación

El desarrollo agrícola es la base del desarrollo rural en nuestro país y en los últimos años la pobreza existente ocasiona déficits nutricionales y educativos en esta zona del país.

La propuesta es social, metodológica y práctica por lo siguiente:

Existen en el país focos de pequeños productores agrícolas con bajo nivel de adopción tecnológica y que están presentes en distintos puntos del territorio nacional. Estas familias son las más vulnerables de padecer los efectos de los cambios climáticos

–sobre todo en la sierra del Perú–, pero estos focos de agricultores no han sido identificados, ni ubicados, ni mucho menos caracterizados. Existen investigaciones anteriores que muestran indicadores de adopción tecnológica, pero no se encontró en la literatura revisada el método que se propone: combinación de estadística multivariada, minería de datos y estadística espacial.

La importancia del Índice de adopción tecnológica planteado en la presente investigación es que, además de identificar el nivel de adopción tecnológica por productor agropecuario, permitirá realizar una caracterización y agrupación según las tendencias que muestre el productor agropecuario.

Además, el presente trabajo se alinea con los objetivos políticos planteados por el Ministerio Nacional de Agricultura en el Eje de Innovación y Tecnificación Agraria el cual indica que las prácticas de innovación y tecnificación contribuyen al fortalecimiento de los productores agrícolas mejorando la producción. El acceso a nuevos mercados, logrando mejores tasas de rentabilidad, competitividad agrícola y propiciando mejor uso de los recursos naturales; por lo que se tiene la necesidad de realizar investigaciones que contribuyan a la toma de decisiones en el sector agrícola.

1.5. Originalidad de la Investigación

La originalidad de la investigación está dada por tres aspectos:

El primero es el uso de la estadística multivariada en un estudio social con la finalidad de la obtención de un índice para la adopción tecnológica que permita realizar comparaciones y ponderar correctamente cada una de las variables que conforman el índice.

El segundo, el uso de la minería de datos para realizar un agrupamiento que permita identificar los pequeños productores agropecuarios con bajo índice

de adopción tecnológica quienes se encuentran en situación de vulnerabilidad. Estos métodos requieren menor tiempo computacional, dado que para esta investigación el número de observaciones para analizar es bastante grande.

El tercero es la aplicación del análisis estadístico espacial para la identificación de los agrupamientos, lo que permitió encontrar los grupos de pequeños productores agropecuarios que además de tener características similares, tienen ubicaciones cercanas, lo que permitiría tomar acciones y políticas de estados en las distintas áreas geográficas donde se ubican, considerando además el requerimiento de sus necesidades previamente identificadas.

1.6. Limitaciones de la investigación

La investigación se ha trabajado con los datos del Censo Nacional Agropecuario 2012, la primera limitación es que desde la fecha en que los datos fueron publicados (2014) no se ha realizado otro Censo Agropecuario, hasta la redacción de este trabajo han pasado 6 años y las condiciones de los pequeños productores agropecuarios deben haber cambiado; una segunda limitación es la georreferenciación de las unidades agropecuarias de los pequeños productores, puesto que únicamente se cuenta con la georreferenciación de los SEA (Sectores de Enumeración Agrícola) los cuales agrupan un número de aproximadamente 30 productores agropecuarios; para este estudio se consideran las SEA donde existen al menos 80% de agricultores con bajo nivel de desarrollo económico.

1.7. Hipótesis

Para obtener información que permita a las autoridades del Ministerio de Agricultura y Riego, apoyar prioritariamente a los pequeños productores agropecuarios con bajo nivel de desarrollo tecnológico proponemos las siguientes hipótesis:

- La mayoría de productores agropecuarios cuentan con niveles muy bajos de adopción tecnológica.
- Existe al menos un foco de alta densidad en donde se concentran productores con bajos nivel de desarrollo.
- Existen diferencias en las características de los productores en los distintos focos de alta densidad.

CAPÍTULO II: MARCO TEÓRICO

2.1. Antecedentes de la investigación

Se realizó una revisión de la literatura nacional e internacional para darle un sustento científico a cada una de las cuestiones tratadas en la presente investigación. A continuación, se presentarán los principales aportes, empezando por la literatura nacional y siguiendo con la internacional.

2.1.1. Antecedentes Nacionales

Se hizo una revisión de la literatura que trabajo con datos del IV CENAGRO con el interés de realizar una caracterización del nivel tecnológico de los pequeños agricultores lo cual nos dio precedentes de diversas técnicas y variables utilizadas para resolver este problema.

Kim (2015), en su trabajo "Difusión tecnológica agropecuaria en el Perú: causas y tendencias en el periodo intercensal 1994-2012" trabajo un índice de tecnología con el cual clasificó los distritos según su nivel tecnológico; para la creación de su índice utilizó la metodología del Programa de las Naciones Unidas para el Desarrollo (PNUD) para la construcción del Índice de Desarrollo Humano (IDH o HDI –en inglés-). Los resultados del índice fueron utilizados para aplicar un modelo econométrico que permita explicar su comportamiento con la presencia de variables independientes (también extraídas del CENAGRO 2012) como porcentaje de superficie

cultivada en el distrito destinado a venta, porcentaje de mujeres productoras agropecuarias en el distrito, entre otras sobre el comportamiento del indicador. Una de las limitaciones principales del índice fue la ponderación que debía tener cada dimensión, así como también la heterogeneidad del País, junto con el problema que no todos los productores se dedican a la misma actividad productiva, algunos solo se dedican a la agricultura o a la ganadería, lo que origina una subestimación de la adopción tecnológica en su área.

Escobal (2014), realizó una investigación de caracterización espacial de la adopción de tecnologías agrarias. Su investigación tuvo como objetivos principales identificar los factores determinantes de la adopción de tecnologías agrarias, así como también identificar las zonas específicas del país que por sus características puedan ser priorizadas por un programa de difusión de innovaciones. Como parte de su investigación estimó las tasas de adopción de riego tecnificado, semilla o plantones mejorados (o certificados), abono orgánico, control biológico de plagas, crianza de ganado mejorado, prácticas pecuarias adecuadas y mecanización. Para la caracterización espacial hizo uso del Índice de Correlación de Moran y los complementó con el Indicador de asociaciones locales y globales de Getis y Ord. Para la caracterización utilizó modelos econométricos MCO y Tobit. Los principales resultados de la investigación mostraron patrones espaciales bien identificados; por ejemplo, en la región de la costa se encontró un despliegue de tecnologías con el uso de semilla certificada, así como también de la mecanización; por otro lado, en el caso del riego tecnificado se encontraron patrones mucho más completos puesto que están presentes en las tres regiones naturales.

Maletta & De la Torre Ugarte (2015), presentan una investigación en la cual realizan una tipificación de los pequeños productores del sector agropecuario. Para ello en uno de los puntos de su investigación realizan el cálculo de un índice de nivel tecnológico en dos distintos niveles: tecnología de cultivo y tecnología pecuaria. Para construir la medida del nivel tecnológico agrícola hicieron uso del Análisis de Componentes Principales, con la cual

redujeron a tres dimensiones el cálculo de cada índice. Entre sus principales resultados encontraron que un 70% de las unidades estaban en un nivel tecnológico que calificó como “realmente bajo” (un rango de 0 a 1 con un máximo de 6 puntos) para los pequeños agricultores.

2.1.2. Antecedentes Internacionales

Devendra (2002), en su informe “Crop–animal systems in Asia: importance of livestock and characterisation of agro-ecological zones", indica la importancia de los pequeños agricultores dentro los países en desarrollo; y aumentar la productividad de estos aliviaría la pobreza y mejoraría sus condiciones de vida, en su caso explora como sería afectada su sostenibilidad en cuanto a la degradación del suelo.

Andrade Limas, Faz Cano, & García Fernandez (s/a) en su publicación “Farmer classification. An approach to soil qualityin Murcia, SE of Spain” realizaron una clasificación de los suelos agrícolas en base al conocimiento de los agricultores. En el trabajo se concluyó que dentro de las políticas de desarrollo se debería tener en cuenta el conocimiento de los agricultores respecto a los problemas técnicos, ecológicos y de gestión.

Escobar & Henao (1983), en su trabajo “Assessing adoption of technology among small farmers: classification and profile analysis” realizaron una clasificación de los agricultores con un análisis discriminante de 4 niveles en base al número de prácticas agrícolas tecnológicas adoptadas. Los cuatro niveles fueron los siguientes: (1) No adoptantes, (2) Adoptantes bajos, (3) Adoptantes medianos y (4) Adoptantes de todas las prácticas. Esto se hace con la finalidad de encontrar una función que permita obtener grupos de pertenencia y predecir a qué grupo pertenecería un agricultor de acuerdo con las características que presenta frente a la adopción tecnológica.

De Cos Guerra (2004), hizo uso del método de Estimadores Focales de Kernel para estudiar la Evolución demográfica en España encontrando las siguientes ventajas en el método: Produce un suavizado masivo de los resultados de modo que se deja de lado las limitaciones municipales con el incremento y decrecimiento de la densidad. El método de Kernel permite apreciar patrones espaciales que permiten comprender los fenómenos evolutivos con mayor facilidad y facilita la comparación de resultados cartográficos en distintos ámbitos. También encontró desventajas en el método de densidad de Kernel desde el punto de vista en que este puede señalar zonas densas en ciudades con límites costeros en donde la alta presencia de población carece de sentido, así como también en las zonas con presencia de montañas, lagos, etc.; sin embargo, esta limitación puede asumirse desde que se tiene en cuenta que la cartografía generada no pretende en ningún caso la exactitud.

Seaman & Powell (1996), trabajaron las estimaciones de densidad de Kernel con la finalidad de medir su exactitud en base a los diferentes niveles del ancho banda, estimada por los métodos del Kernel Normal bivariado según validación cruzada y un ancho de banda referencial en comparación con las estimaciones por media armónica. Se encontró que las estimaciones mediante el método de Kernel con validación cruzada tiene una mayor exactitud.

Mwangi & Kariuki (2015), en su publicación hacen una extensiva revisión bibliográfica sobre el papel que tiene la adopción tecnológica en pequeños agricultores de países en desarrollo y los factores asociados a su adopción. En su investigación sustentan que la adopción tecnológica juega un importante rol en el crecimiento económico, la reducción de pobreza y el desarrollo rural, puesto que se mejora el uso de los recursos como el agua, aumenta la productividad de la finca, mejora el estatus nutricional, disminuye el costo de producción, aumenta las posibilidades de empleo y, por último, permite cubrir la creciente demanda alimenticia. Además, los factores

asociados a la adopción tecnológica más importantes fueron el tamaño de la finca, el nivel educativo, el género y la edad del productor agrícola.

Jaller (2010), en su informe sobre la sostenibilidad de la actividad de los pequeños productores agropecuarios indica que los mismos sufren las consecuencias de ser parte de grandes cadenas de producción lo cual hace difícil acceder al comprador final y, por tanto, limitan su desarrollo. Las principales soluciones a esas problemáticas planteadas en su informe son la adopción tecnológica, así como la incorporación de las tecnologías de información y comunicación (TIC), facilitar el acceso al crédito, fomentar el emprendimiento agroempresarial, fomentar la asociatividad entre pequeños productores rurales y diseñar estrategias comerciales que permitan la diversificación de los canales de distribución para dejar de depender de los mayoristas.

Graziano Da Silva y otros (2010), en su artículo publicado para la revista austral de Ciencias Sociales, indican que el Perú tiene una concentración de crecimiento agrícola únicamente en la costa donde en su mayoría son grandes y medianos agricultores. Además, encontraron evidencia que en Perú hubo una reducción de la pobreza asociada con el crecimiento agrícola, pero más que nada en la costa Sur del país. También recalcan que sigue existiendo concentración de pobreza en la Sierra peruana encontrando contradicciones en los programas de desarrollo, puesto que se está incentivando el abandono de la agricultura en lugar de crear programas integrales que complementen está con actividad con otra forma de generar ingresos.

Mendola (2007), estudia el impacto potencial de la adopción tecnológica sobre la reducción de la pobreza mediante un modelo Propensity Score Matching (PSM) con una base de datos de una encuesta realizada a 3800 hogares rurales con producción agrícola. Para el modelo se consideraron únicamente 2562 productores agrícolas entre pequeños, medianos y grandes que habían adoptado o no tecnología. Concluyó que existe una relación causal

–y compleja– entre la adopción tecnológica y la reducción de pobreza. Evidenció que la reducción de pobreza rural está directamente relacionada con la adopción de tecnología en pequeños productores agrícola, puesto que estos cuentan con tierras limitadas y adoptar tecnologías les permitiría tener mayor capacidad productiva.

Lowder, Skoet, & Singh (2014) en su informe para la FAO (Food and Agriculture Organization) hacen una estimación del número y distribución de granjas familiares y no familiares en todo el mundo determinando que hay predominancia a nivel mundial de granjas familiares de menos de 2 hectáreas, mientras que en Perú la gran mayoría tiene menos de 5 hectáreas; además, en el informe el Perú es catalogado como un país de ingresos agrícolas medianos y altos -contradiciendo los grandes niveles de pobreza en el sector agrícola-. Vale recalcar que en el informe se consideran granjas pequeñas aquellas que tienen una dimensión máxima de 5 hectáreas.

Asfaw, Kassie, & Lipper (2012), evalúan el impacto de la adopción tecnológica sobre los gastos de consumo y el estado de probreza medidos mediante el índice de recuento, índice de brecha de pobreza e índice de severidad de pobreza en zonas rurales de Tanzania. La primera conclusión indica que los adoptantes y los no adoptantes de tecnología agricola tienen caracteristicas diferentes, también indican que la adopción de la tecnología agrícola contribuye a la mejor calidad de vida de los hogares rurales, ya que mayores ingresos por el uso de la tecnología significan menos pobreza. Las variables utilizadas en el análisis fueron el área total cultivada, años de estudio, agricultura como actividad económica principal, distancia a una cuidad principal, tenencia de la granja, entre otras.

Liu (2013), investigó sobre el riesgo de la adopción de nuevas tecnologías en el uso de nuevas semillas de algodón en China. Dentro de sus conclusiones encontramos que a pesar de que se ha demostrado que la adopción de tecnologías mejora la acumulación de riqueza y el acceso a nuevos

mercados, los productores agricolas de China demoraron cerca de 10 años para tecnificarse. Los principales factores que determinaron la resistencia a la adopción tecnológica fueron la edad del agricultor, los años de estudio, la pertenencia de la finca y el número de adultos mayores de 28 años en el hogar.

Kassie, Shiferaw & Muricho (2011), estudiaron el efecto de la adopción de nuevas tecnológicas sobre los ingresos de los agricultores por la producción en sus tierras y el estatus de pobreza, en particular para agricultores que usaron variedades mejoradas de maní. Encontraron que la pobreza se redujo entre un 7% y 9% en los pequeños agricultores adoptantes de semillas de maní mejoradas. Usaron el modelo Propensity Score Matching (PSM) para estudiar cómo los ingresos cambiaban de acuerdo a los años de estudio y los tamaños de la granja.

Ezzrari & Verme (2013), calcularon un Indice de Bienestar Compuesto (CWI: Composite Welfare Indicator) como médida de pobreza multidimencional en Marruecos para evaluar los cambios entre los años 2001 y 2007 mediante el uso del Análisis de correspondencias múltiples. Evaluaron dos modelos para zonas rurales y urbanas, confirmando que existen dos brechas bien marcadas entre ambas zonas, ya que a pesar de la reducción del índice a lo largo de los años estudiados, esta reducción mostró mayor fuerza en la zona rural, pero aún estaba lejos de alcanzar las mismas condiciones que las zonas urbanas.

Decancq & Lugo (2013), en su artículo hacen una revisión de las distintas ponderaciones dadas a las variables en el cálculo de distintos índices multidimensionales en las Ciencias Sociales. Dentro de la revisión de los métodos estadísticos hicieron referencia al Análisis de Correspondencias Múltiples y Modelos de Ecuaciones estructurales afirmando que el problema principal de estos enfoques es que asignan menos pesos a las dimensiones menos correlacionadas o con menor número de observaciones. También

argumentan que el problema de los enfoques basados en datos pueden cambiar en distintas ediciones y carecen de comparabilidad a lo largo del tiempo.

2.2. Bases teóricas

2.2.1. Análisis *de Correspondencias*

En esta sección se hizo una revisión de la teoría necesaria para comprender el análisis de correspondencias múltiples, es importante conocer el análisis geométrico de datos que explica la naturaleza de la representación de variables categóricas mediante nube de puntos en dos o más dimensiones (ver Anexo 6). Se hizo una revisión del análisis de correspondencias que fue tomado como base para comprender a profundidad el análisis de correspondencias múltiples.

2.2.1.1. Introducción

El análisis de correspondencias es también llamado mapeo de correspondencias, mapeo perceptual, análisis del espacio social, análisis factorial de correspondencias, análisis de correspondencias principales de datos cualitativos y escalamiento dual son algunos de los sinónimos más usuales para este método.

2.2.1.2. Análisis de Correspondencias Múltiples

El Análisis de Correspondencias Múltiples puede ser definido de dos maneras: (i) como el análisis de correspondencias de la matriz binaria o (ii) como el análisis de correspondencias de la matriz de Burt (Greenacre, 2017). Además, es definido como una forma más de reducción de un espacio de

propiedades basado en variables cualitativas y como posible método alternativo al corrientemente utilizado en la construcción de índices. (Baranger, 2009). El análisis de correspondencias múltiples permite realizar análisis con mapas de categorías, mapas de variables y mapas de individuos. (Husson, Lê, & Pagès, 2017).

Matriz Binaria. La matriz binaria es una matriz $N \times J$ de individuos por variables en el cual los individuos representan las filas y las variables se encuentran en las columnas con cada una de sus categorías (o modalidades) dicotomizadas, el número de columnas en la matriz binaria es igual al número de preguntas por el número de categorías; es decir, $J = \sum_q J_q$, donde J_q representa el número de categorías en la j-ésima variable, suponiendo que se tienen Q variables y q categorías en cada variable.

Sea la matriz binaria Z con J columnas, es una matriz que está formada por tablas Z_q agrupadas lateralmente. La inercia total de una matriz binaria es igual al promedio de las tablas binarias que la componen. Debido a su naturaleza, cada una de las matrices binarias tiene un solo 1 en cada fila y los demás valores cero y, por tanto, las inercias de todos los ejes principales son iguales a 1. Por tanto, la inercia de cada tabla binaria Z_q es igual a su dimensionalidad; es decir $J_q - 1$. (Greenacre M. , Multiple Correspondence Analysis, 2017)

$$inercia(Z) = \frac{1}{Q}\sum_q inercia(Z_q) = \frac{1}{Q}\sum_q (J_q - 1) = \frac{J - Q}{Q}$$

El análisis de correspondencias en la matriz binaria es utilizado calcular las puntuaciones de cada individuo en cada una de las nuevas dimensiones. Las puntuaciones son utilizadas para obtener el índice de adopción tecnológica en la presenta investigación. La metodología para el cálculo de las puntuaciones (Greenacre M. , 1991) se presenta a continuación:

(i) $P = \left(\frac{1}{NQ}\right) Z$, es la "matriz de correspondencias", donde NQ es el total de la matriz Z.

(ii) Un perfil de Z se refiere a una fila o columna de Z (o equivalente a P) expresado en función al total de sus elementos.

(iii) El vector $r = P1$ es el vector de masas de las filas o el perfil promedio de las columnas con el total de elementos equivalente a $1/N$ para el caso especial de la matriz indicadora Z.

(iv) El vector $c = P^T 1$ es el vector de masas de las columnas o el perfil promedio de las filas con elementos iguales a las frecuencias marginales divididas por NQ.

(v) D_r y D_c son las diagonales de la matriz de masas.

(vi) La distancia Chi-cuadrado entre dos perfiles fila (o columna) es una distancia euclidiana ponderada definida por ${D_c}^{-1}$ (respectivamente ${D_r}^{-1}$).

(vii) El análisis consiste en realizar una Descomposición de Valores Singulares (DVS), de:

$$E = {D_r}^{-\frac{1}{2}}(P - rc^T){D_c}^{-1} = \left(\frac{1}{N^{1/2}Q}\right)(Z - Q1c^T){D_c}^{-1/2}$$

y con las K^* componentes principales $UD_\lambda V^T$; en la práctica debido a que N es mucho más grande que J, es conveniente realizar la descomposición mediante el cálculo de los autovalores y auto vectores de la matriz $E^T E$ de rango $J \times J$.

(viii) Las inercias principales se encuentran en la diagonal de $D_\lambda \equiv {D_\alpha}^2$; existen $J - Q$ valores singulares distintos de cero en la DVS de E y la inercia total está dada por $\frac{J}{Q} - 1$.

(ix) Digamos que el espacio total de perfiles tiene dimensionalidad $K = J - Q$ y el espacio reducido tiene K^* dimensiones.

(x) Las coordenadas de filas y columnas en los espacios reducidos son dados por las siguientes matrices:

Coordenadas principales:

$$F(N \times K^*) = {D_r}^{-1/2} U {D_\lambda}^{-1/2} \text{ (filas)},$$
$$G(J \times K^*) = {D_c}^{-1/2} V {D_\lambda}^{-1/2} \text{ (columna)}$$

Coordenadas estándar:

$$\Phi(N \times K^*) = {D_r}^{-1/2} U \text{ (filas)},$$
$$\Gamma(J \times K^*) = {D_c}^{-1/2} V \text{ (columna)}$$

Las coordenadas principales permiten realizar un gráfico conjunto entre F y G, conocido como mapa simétrico, en el cual se puede interpretar la distancia entre los perfiles fila y los perfiles columnas como una aproximación de la distancia chi cuadrado.

Las coordenadas estándar permiten realizar un gráfico asimétrico, en el cual se gráfica las filas en coordenadas principales y las columnas en coordenadas estándar o viceversa y también es válido interpretarlo como una aproximación de las distancias chi cuadrado.

Matriz de Burt. La matriz de Burt es una tabla resultante de la unión de todas las tablas resultantes de cruzar las variables dos a dos. La matriz de Burt se relaciona directamente con la tabla binaria de la siguiente manera $B = Z^T Z$. Aplicar el análisis de correspondencias múltiples es realizar los pasos mencionados antes, pero a la matriz de Burt. La inercia de la matriz de Burt es calculada como el promedio de las inercias de las QxQ tablas de contingencia de la matriz de burt.

Relación entre la matriz de Burt y la matriz binaria se da a notar en las inercias, puesto que las inercias principales del análisis de Burt son el cuadrado de los de la matriz binaria. Dado que las inercias principales son menores de 1, sus cuadrados serán valores aún más pequeños; en consecuencia, los porcentajes de inercia serán siempre mayores en el análisis de Burt. (Greenacre M. , 2017)

2.2.2. Definición del Índice Sintético.

Una vez calculadas las coordenadas principales mediante el análisis de correspondencias múltiples (o cualquier otro método de reducción de dimensiones) se calcula el índice sintético de la forma siguiente: se multiplica la puntuación de cada coordenada principal por la raíz cuadrada del valor singular asociado y se divide entre la suma de la raíz de cada valor singular (Escobar Jaramillo, 2008). La expresión matemática está dada por:

$$Indice_{\ j} = \frac{\sum_{i=1}^{k} z_{ij}\sqrt{\lambda_i}}{\sum_{i=1}^{k}\sqrt{\lambda_i}} \tag{2.1}$$

donde $j = 1,2,\dots,n\ individuos$ e $i = 1,2,\dots,K\ componentes$

Siendo z_{ij} la puntuación del componente i-ésimo para la observación j-ésima y $\sqrt{\lambda_i}$ la raíz cuadrada del autovalor asociado a la i-ésima componente, lo cual garantiza que las componentes que tengan mayor varianza explicada tengan una ponderación mayor en la calificación del índice.

2.2.3. Índice de Adopción Tecnológica

El índice de adopción tecnológica tiene origen en el año 2015, cuando Galileu Kim realizó una investigación en la cual calculó el Indicador de Nivel Tecnológico haciendo uso de 23 variables entre pecuarias, agrícolas, uso de energía, acceso a crédito y capacitación. Este indicador fue creado para calcular el nivel tecnológico a nivel distrital, mediante un promedio ponderado en función del número de agricultores que adoptaban cierta cantidad de tecnologías (Kim, 2015).

De manera similar al funcionamiento que tiene en Indicador de Desarrollo Humano (IDH), el índice de adopción tecnológica asigna un valor

basado en las variables relacionadas con las tecnologías aplicadas en las prácticas agropecuarias.

El cálculo del Índice de Adopción Tecnológica se realiza según la expresión (2.1) en el que se define el índice sintético y se realiza un re-escalamiento restándole el mínimo y dividiéndolo entre el rango, por 100.

2.2.4. Análisis de Agrupamientos

El análisis de agrupamientos es una técnica exploratoria cuyo objetivo es encontrar grupos de individuos u objetos internamente homogéneos y heterogéneos entre ellos.

El análisis de agrupamientos mide la similaridad entre observaciones y las agrupa en base a sus distancias para lo cual utiliza la distancia euclidiana u otra según la naturaleza de la variable. La fortaleza del análisis de agrupamientos es su naturaleza exploratoria y es conocido, en *Machine Learning*, como una técnica de aprendizaje no supervisado (Koch, 2014).

En algunos métodos de agrupación -como el de k medias- se requiere determinar el número de agrupamientos a priori, pero esto no es fácil, puesto que su número puede cambiar por completo el sentido e interpretación de cada agrupamiento. Para garantizar que el número de agrupaciones sea el correcto se usaron técnicas que determinan el número óptimo de clúster y su validación.

2.2.4.1. Determinación del número óptimo de Agrupamientos

Determinar el número óptimo de agrupamientos es un problema fundamental cuando se requiere hacer particiones mediante métodos de agrupamiento no jerárquico y no se conoce a priori el número de grupos. Una forma popular de realizarlo es mediante el dendograma que sugiere el número de clústeres a usar en el agrupamiento jerárquico; sin embargo, su enfoque es

subjetivo porque el número de agrupamientos dependerá del punto de corte que elija el investigador.

Se va a presentar a continuación métodos que pueden ser usados tanto para agrupación jerárquica, como no jerárquica. (Kassambara, 2017). Estos métodos incluyen los siguientes:

- *Métodos directos*: Consiste en un criterio de optimización que usa como medida de comparación la suma de cuadrados intra-agrupaciones o la silueta promedio. Estos métodos son llamados Elbow y Silueta respectivamente.
- *Métodos de pruebas estadísticas*: Consiste en comparar evidencia contra una hipótesis nula especificada; por ejemplo, la estadística Gap.

Método Elbow. La media de la suma de cuadrados intra-grupos mide que tan compacto es el agrupamiento y para ello se requiere que esta medida sea lo más pequeña posible. El método Elbow toma el total de la suma de cuadrados intra-grupos en función del número de grupos. Se elige el número de grupos de tal manera que al cambiar este número no mejore la medida de la suma de cuadrados intra-grupos. La suma de cuadrados intra- grupos (SCIG) es definida por la fórmula:

$$SCIG = \sum_{x_j \in C} \left(x_j - \bar{x}_c\right)^2$$

Donde, x_j representa a cada observación en el grupo C y es el promedio de las observaciones del grupo C.

Método de Silueta promedio. El enfoque de la silueta promedio mide la calidad de la agrupación; es decir, determina cuan bien clasificada se encuentra cada observación en su clúster y un alto valor de silueta promedio indica un buen agrupamiento. El valor de la silueta promedio es calculado para distinto valores de k (número de clústeres) y el número óptimo de clústeres es la silueta promedio máxima entre los valores de k.

Método de la estadística Gap. La estadística de Gap compara la varianza intra-clúster para diferentes valores de k con sus valores esperados, bajo la hipótesis nula específica de una distribución de los datos. El número óptimo de clúster será el valor que minimice la estadística de Gap.

2.2.4.2. Validación estadística de los grupos

La validación de los grupos hace referencia a la bondad de ajuste del resultado del algoritmo de agrupamiento. Esto es importante para no caer en el error de encontrar patrones en datos aleatorios y para realizar comparaciones de algoritmos (Kassambara, 2017).

La validación estadística puede ser categorizada en 3 clases:

Validación Interna. Es una medida de información interna para evaluar la bondad del ajuste de la estructura del clúster sin tener como referencia la información externa; también es utilizado para calcular el número de agrupamientos.

Validación Externa. Consiste en comparar el resultado del ajuste del clúster con información externa conocida (como clases dadas a priori). Este método es usado cuando se quiere decidir que algoritmo usar en una base de datos de acuerdo con el número de observaciones que fueron bien agrupadas.

Validación Relativa. Evalúa la estructura del clúster variando los valores de los parámetros para el mismo algoritmo. Es usualmente usado en la determinación del número de agrupamientos.

2.2.4.3. Validación interna de Grupos

En la validación interna se evalúa la compacidad, la conectividad y la separación de los agrupamientos (Kassambara, 2017).

Compacidad. Mide cuan cerca están los objetos dentro del mismo grupo y los índices que miden la compacidad se basan en la distancia entre los objetos de un mismo agrupamiento.

Conectividad. Corresponde a la cantidad de elementos que se colocan en un mismo grupo y la relación con sus vecinos más cercanos en el conjunto de datos. La conectividad tiene un valor entre cero e infinito y debe ser minimizado.

Separación. Consiste en medir cuán lejos está un clúster de otro y se mide respecto a la distancia entre sus centros y la distancia entre sus elementos por pares.

Generalmente, los índices de validación interna comparan la medida de compacidad con la medida de separación.

Se hizo la revisión de algunos índices de validación interna.

i. Coeficiente Silueta. El coeficiente silueta mide cuan bien agrupada está una observación estimando la distancia promedio entre agrupaciones. El gráfico silueta muestra que tan cerca está un punto de un grupo con un punto del grupo vecino. Para cada observación i, el ancho de silueta s_i se calcula de la siguiente manera:

   a. Para cada observación se calcula la medida promedio de disimilaridad a_i entre i y todos los otros puntos de su grupo.
   b. Para los otros agrupamientos C a los cuales i no pertenece, se calcula la medida de disimilaridad $d(i, C)$ de i con todas las observaciones de C. La más pequeña medida de disimilaridad definida como $b_i = min_i[d(i, C)]$, entonces b_i puede ser visto como la medida de disimilaridad entre i y sus agrupaciones vecinas; es decir, las agrupaciones más cercanas a los cuales i no pertenece.

c. Finalmente, el ancho de silueta para la observación i es definida como $S_i = (b_i - a_i)/\max(a_i, b_i)$

El ancho de silueta se interpreta según el tamaño de S_i, mientras más cerca de 1, mejor está la agrupación. Un valor del coeficiente cercano a cero significa que debe haber solamente 2 agrupamientos y un valor del ancho de silueta negativo, indica un mal agrupamiento.

ii. Índice de Dunn

El índice de Dunn se calcula como sigue:

a. Para cada agrupamiento, calcular las distancias entre cada una de las observaciones del agrupamiento y las observaciones de otros agrupamientos.
b. Usar la mínima separación entre agrupamientos como medida de referencia. $(sep.mínima)$.
c. Para cada agrupamiento, calcular la distancia máxima entre cada una de las observaciones.
d. Usar la máxima distancia intra-agrupamiento calculada $(sep.máxima)$.

   Calcular el índice de Dunnet como $D = \frac{sep.mínima}{sep.máxima}$

Se espera que la separación entre agrupamientos sea alta y la separación entre los elementos del mismo grupo pequeña, así que se espera un valor de Dunnet maximizado.

2.2.5. Mapas Auto Organizados de Kohonen.

Los mapas auto organizados de Kohonen son un tipo de red neuronal que son entrenados mediante el aprendizaje no supervisado y que son muy usados para clasificación debido a su naturaleza. Fueron presentados en 1982 por Teuvo Kohonen, basado en ciertas evidencias de procesos que suceden en el cerebro para reconocer patrones; por ejemplo, cuando se hace un reconocimiento del rostro de un conocido, el cerebro procesa la información en mapas que se encuentran en las cortezas cerebrales donde las neuronas están ordenadas topológicamente de forma que la información captada del entorno mediante los sentidos se representa internamente en capas bidimensionales. En base a esto los mapas auto organizados de Kohonen pretenden imitar de forma simplificada la capacidad del cerebro de formar mapas topológicos a partir de las señales recibidas del exterior (Kohonen, 2001).

El mapa de autoorganización es un principio de mapeo computacional que forma una proyección ordenada no lineal de datos de entrada de grandes dimensiones en unas cuadriculas regulares de baja dimensión (generalmente en dos dimensiones). Los puntos que están en la cuadrícula son llamados nodos y la visualización producida por estos nodos puede ser considerado como un gráfico de similitud (Natschläger, Ruf, & Schmitt, 2013).

De este modo, es capaz de convertir relaciones estadísticas complejas no lineales entre elementos de alta dimensión en una imagen de baja dimensión. A medida que comprime la información conserva las relaciones topológicas y métricas más importantes de los datos en la imagen de salida que a su vez produce un tipo de abstracción. Estos dos aspectos: abstracción y visualización se puede realizar para solucionar tareas complejas como para realizar análisis de procesos, comunicación, etc. El modelo organiza automáticamente en un orden bidimensional en el que los nodos similares están cerca los uno de los otros en la cuadrícula en comparación con los dispares, en ese sentido el mapa

auto organizado es un gráfico de similitudes y de agrupación a la vez. El cálculo se realiza mediante un proceso de regresión recursivo no paramétrico. (Kohonen, 1990).

2.2.5.1. Arquitectura de los Mapas Auto organizados de Kohonen

Un modelo de Mapas Auto organizados está formado por dos capas de neuronas: la capa de entrada (formada por n neuronas -una por cada variable-) la cual se encarga de recibir y enviar la información del exterior a la capa de salida. La capa de salida en la cual se procesa la información y forma el mapa de rasgos (ver figura 1). La conexión entre la capa de entrada y la capa de salida se da mediante una matriz de pesos que se van actualizando de acuerdo con el avance del aprendizaje (Becerra Guevara, 2014).

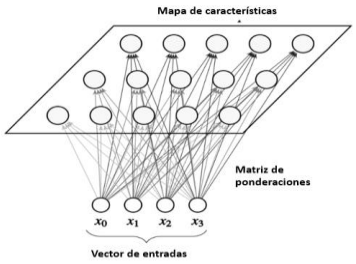

Figura 1: Estructura del Mapa Autoorganizado
Fuente: *Kohonen (2001)*

2.2.5.2. Algoritmo de aprendizaje incremental de los Mapas Autoorganizados

Este algoritmo puede ser visto como un proceso de regresión recursiva especial en el cual solo un grupo de modelos son procesados en cada paso.

Supongamos que se tiene datos de entrada en un espacio R^n los cuales, mediante el algoritmo de las redes neuronales de Kohonen, serán proyectados en un arreglo de nodos bidimensional. Para cada nodo i está asociado un vector modelo paramétrico, llamado también vector de referencia, $m_i = [u_{i1}, u_{i2}, \dots, u_{in}]^T \in R^n$ el cual debe ser inicializado antes del proceso recursivo. Se seleccionan números aleatorios para las componentes de m_i y a lo largo del aprendizaje m_i alcanzará valores representados en dos dimensiones como efecto de la auto organización. El tipo de arreglo de salida puede ser definido como rectangular, hexagonal o irregular, aunque por temas de visualización es más popular la forma hexagonal (ver figura 2). (Kohonen, 2001). En la figura 2, se muestra un ejemplo en el cual se busca clasificar los modelos de espectros del habla del idioma finlandés y se puede ver que un espectro representado en una neurona de salida es muy parecido al espectro representado en sus neuronas vecinas.

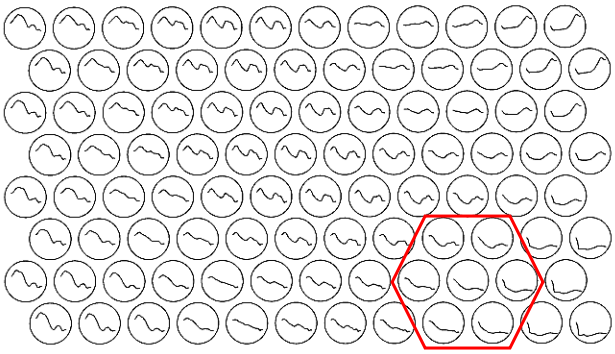

Figura 2: Cuadrícula hexagonal de clasificación de modelos de espectro del habla natural del idioma finlandés.
Fuente: *Kohonen (2001)*

En un caso más simple, un vector de entrada $x = [\xi_1, \xi_2, \dots, \xi_n]^T \in R^n$ está conectado a todas las neuronas en paralelo a través de pesos escalares

variables u_{ij}, las cuales son, en general, diferentes para cada diferente neurona. En otras palabras, el vector x se va comparando mediante procesos de cálculos paralelos con todos los m_i y se elige como respuesta a la mejor coincidencia a partir de alguna medida de distancia.

Sea $x \in R^n$ un vector estocástico, entonces se podría decir que los mapas auto organizados son una proyección no lineal de la función de densidad de probabilidad del vector de datos de entrada x en la gráfica bidimensional.

El vector x puede ser comparado con m_i en cualquier medida de distancia, en muchas aplicaciones prácticas se suele usar la más pequeña de las distancias euclidianas $||x - m_i||$ la cual puede ser definida como el "nodo ganador" denotado por c:

$$c = argmin_i\{||x - m_i||\},$$

que es lo mismo que:

$$||x - m_c|| = min_i\{||x - m_i||\}.$$

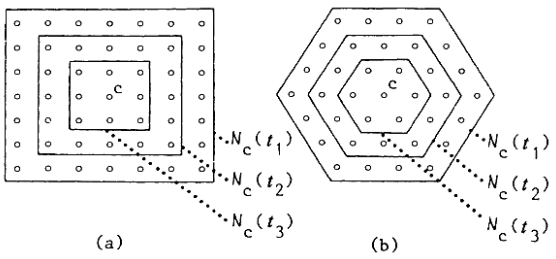

Figura 3: Ejemplos de vecindad topológica.
Fuente: *Kohonen (2001)*

Durante el proceso aquellos nodos que están ubicados topológicamente hasta cierta distancia geométrica se activarán mutuamente para aprender de la misma entrada x (figura 3). Esto genera un efecto de suavizamiento local sobre los vectores de pesos de las neuronas en este vecindario que en aprendizaje continuo conduce al ordenamiento global. Considerando los límites de la convergencia del proceso de aprendizaje, los valores iniciales de $m_i(0)$ puede ser arbitrarios. Sea

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)], \text{ donde } t = 0,1,2, \ldots$$

En el proceso de suavizamiento, la función $h_{ci}(t)$, llamada también función de vecindad es una función de suavización de Kernel sobre la rejilla de puntos (neuronas de salida) tiene un rol importante, puesto que para la convergencia es necesario que $h_{ci}(t) \to 0$, cuando $t \to \infty$.

Usualmente, $h_{ci}(t) = h(||r_c - r_i||, t)$ donde $(r_c, r_i) \in R^2$ y son los vectores de la ubicación de los nodos c e i en el arreglo, respectivamente. Con el incremento de $||r_c - r_i||$, $h_{ci} \to 0$. El ancho y forma promedio de h_{ci} definen la "rigidez" de la "superficie elástica" para ajustarse a los datos.

La elección de la función $h_{ci}(t)$ ocurre comúnmente, en la literatura, por dos razones: (a) El más simple de ellos hace referencia a una vecindad matricial (ver fig. 13 (a)) alrededor del nodo c (nodo ganador). Denotemos $N_c = N_c(t)$ como una función de tiempo, con lo cual $h_{ci}(t) = \alpha(t)$ si $i \in N_c$ y $h_{ci}(t) = 0$ si $i \notin N_c$. El valor de $\alpha(t)$ es conocido como tasa de aprendizaje $(0 < \alpha(t) < 1)$. Tanto $\alpha(t)$ como el radio de $N_c(t)$ usualmente decrecen monótonamente en el tiempo (durante el proceso de ordenamiento - aprendizaje-).

Otra función Kernel más fácil de usar y ampliamente aplicada se puede escribir en términos de la función Gaussiana.

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}\right)$$

$\alpha(t)$ es una tasa de aprendizaje con valor escalar y el parámetro $\sigma(t)$ determina el ancho del Kernel, este último determina el radio de $N_c(t)$ (ver figura 13). Ambos $\alpha(t)$ y $\sigma(t)$ son funciones que decrecen monótonamente en el tiempo. Es importante tener cuidado con la elección del tamaño de $N_c = N_c(t)$, si la vecindad de inicio es demasiado pequeña el mapa no podrá realizar el ordenamiento global.

2.2.5.3. Preprocesamiento de datos.

El preprocesamiento de los datos es un tema importante que se debe considerar a la hora de realizar mapas auto organizados de Kohonen, puesto que el mapa final dependerá del tipo de preprocesamiento que se aplique a los datos. Cuando las variables de entrada tienen rangos muy diferentes (por ejemplo: edad e ingresos anuales) la escala lineal es una solución rápida, ya que todas las variables de entradas deben estar en el mismo rango; por ejemplo, el intervalo [0,1], pero si existen varios valores que son más altos o más pequeños en comparación con los demás, podría sesgar la escala y la resolución del resto de valores podría perderse. Otra posibilidad es la normalización de todas las variables x a otra variable z de media cero y varianza uno. Con la ecuación de la forma $z = (x - x_m)/\sigma$ donde x_m es la media de la variable x y σ su desviación estándar. Esta estrategia es comúnmente usada en estadística, ya que mantiene a todas las variables en un mismo rango con una desviación de una unidad y además la importancia de la normalización gaussiana es que cada variable se mide en cuanto a su desviación respecto a la media, de esta forma la nueva variable transformada indica la importancia de un valor respecto al resto. (Martin-del-Brio & Serrano-Cinca, 1993).

El preprocesamiento de los datos es uno de los temas más complicados de tratar cuando se diseña una herramienta mediante redes neuronales; por

ejemplo, hay distintos tipos de preprocesamiento, dos ejemplos de ellos son la normalización y la parametrización de los datos, pero luego está la cuestión de cómo se hace, si tengo que usar una red neuronal u otra, y si la transformación de normalización debe ser lineal o gaussiana. En general, se debe usar una forma de normalización independiente del modelo de red neuronal a usarse o de los datos. Para los mapas auto organizados es muy común en la práctica hacer uso de la normalización gaussiana para los vectores de entrada (Eberhart & Dobbins, 1990). Por otro lado, Kohonen argumentó que la normalización de los vectores de entrada es suficiente para obtener la convergencia de la red (ver figura 4), puesto que los pesos se mueven muy de cerca con los valores de entrada normalizados y se normalizan automáticamente (Blayo, 1992).

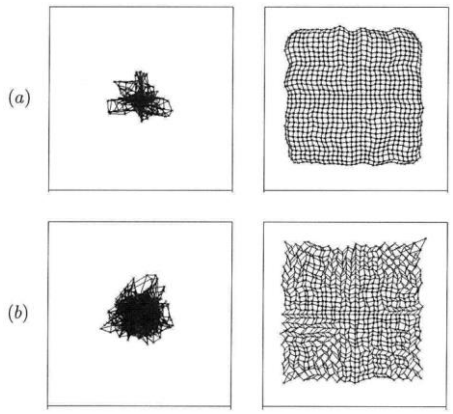

Figura 4: Representación de neuronas de salida luego de 20000 iteraciones. (a) Vectores de entrada normalizados (b) vectores de entrada sin normalizar.
Fuente: *Blayo (1992)*

2.2.6. Función de Densidad Kernel

La función de densidad Kernel permite estimar la función de densidad de un conjunto de datos en función de su propio comportamiento. Desde un

punto de vista matemático se tiene las siguientes propiedades. (Moreno J., 1991)

a. Funciones centradas en cero y su integral es la unidad
b. Como Kernel es una función de densidad, $\hat{f}$ también lo es.
c. En la mayoría de los casos se trata de funciones continuas y diferenciables dependiendo del tipo de Kernel.
d. Según su definición los Kernel son positivos.

Estimador univariado de la función de densidad de Kernel. Para el caso univariado del estimador de la función de densidad Kernel se define de la siguiente manera:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

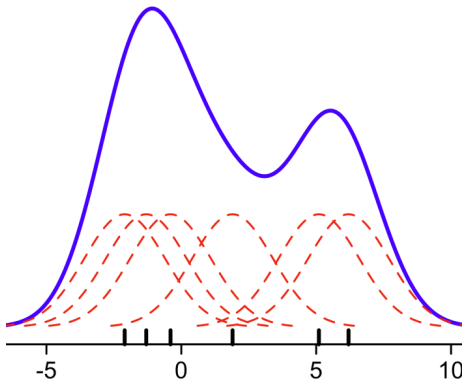

Figura 5: Estimación de la función de densidad Kernel Gaussiano para 6 puntos
Fuente: *Elaboración propia*

Donde, x es el punto sobre el cual se estima la densidad, x_i el valor de la variable para $i = 1,2, \dots, 3, K$, donde K es el símbolo Kernel y h es la anchura

de la ventana o parámetro de suavización. La función Kernel puede tomar distintas expresiones, pero para este trabajo se usó el Kernel cuartico, puesto que es el Kernel que tiene menor pérdida de calidad con solo 1% de diferencia respecto de la solución óptima (Moreno J., 1991). El kernel cuartico sigue la siguiente expresión:

$$k(u) = \frac{3}{\pi}(1-u^2)^2, \quad u^2 \leq 1$$

Elección del ancho de banda. La elección de un adecuado ancho de banda es incluso más delicada que el tipo de Kernel a usar, puesto que un cambio en el valor podría alterar completamente el resultado y la interpretación (ver fig. 6). Enfoques para la elección del ancho de banda:

a. El primero es un enfoque que trata de probar distintos tamaños del ancho de banda y observar su representación cartográfica hasta encontrar una adecuada y que tenga sentido para la investigación.
b. El segundo es suponer una distribución subyacente de los datos y comparar distintos tamaños de ventana que minimiza el MISE (Mean Integrated Square Error) que no es más que la diferencia entre los valores obtenidos con la función Kernel y la función teórica asumida.

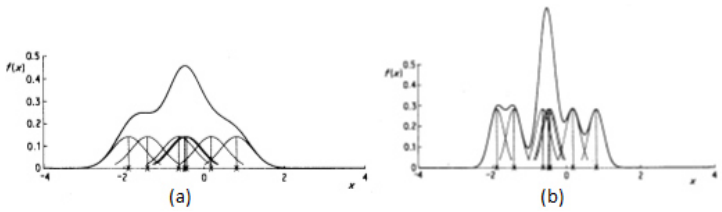

Figura 6: Estimación mediante Kernel Gaussiano mostrando Kernels individuales con h=0.4 (izquierda) y h=0.2 (derecha)
Fuente: *Moreno (1991)*

Estimación kernel multivariada. Si se tiene un vector $\tilde{x} = (x_1, x_2, \dots, x_r)$ como el punto sobre el cual se desea estimar la densidad, el valor de la función es dado por:

$$\hat{f}(\tilde{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{r}\frac{1}{h_j}\right)K\left(\frac{x_1 - x_{i1}}{h_1}, \dots, \frac{x_d - x_{ir}}{h_r}\right)$$

En el cual h_i es el ancho de banda de la ventana para cada dimensión; sin embargo, se puede usar un mismo ancho de banda para todas las dimensiones.

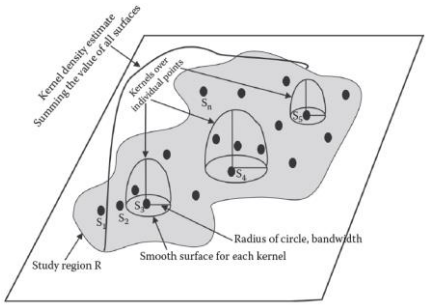

Figura 7: Representación de la aplicación del Kernel bivariado sobre una región R.
Fuente: *Oyana & Margai (2016)*

2.2.7. Estadística Espacial

En esta subsección se definirá el concepto de estadística espacial con énfasis en el análisis espacial de patrones puntuales. Finalmente se definirá el estimador de densidad espacial Kernel y el estimador multivariado que será usado en la presente tesis.

La estadística espacial es un conjunto de metodologías que permite analizar datos que corresponden a la medición de variables en puntos, polígonos o enmallados en el espacio. (Giraldo, 2002).

Formalmente, se puede definir como el análisis de las realizaciones del proceso estocástico $\{Z(s): s \in A\}$ donde s que pertenece al espacio R^d, $Z(s)$ una variable aleatoria en la ubicación s y s varía sobre $A \subset R^d$.

La estadística espacial considera tres distintos enfoques los cuales cambian de acuerdo con la naturaleza del espacio D en el cual varían los puntos.

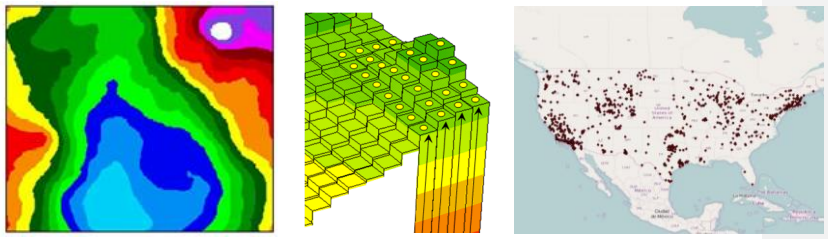

Figura 8: Enfoques de la Estadística Espacial (de izquierda a derecha): Geoestadística, Enmallados y Patrones Puntuales.
Fuente: *Elaboración Propia*

Geoestadística: El espacio A es de naturaleza continua y fija (definido por el investigador) cuyo objetivo es realizar interpolaciones para lo cual la continuidad espacial es importante, ya que si esta no ocurre se podrían realizar predicciones sin sentido. Por ejemplo, si se quiere calcular el nivel contaminación del suelo agrícola en una cierta área y se toman muestras en distintos puntos, se podría predecir una alta contaminación en un área urbana de no existir una delimitación adecuada que garantice la continuidad espacial.

Enmallados: El espacio A es de naturaleza discreta y fija. En este caso el objetivo es analizar indicadores o variables aleatorias en un espacio discreto y no en puntos. Como ejemplo: el promedio de ventas de aceite en distintas manzanas del distrito de San Isidro.

Patrones de puntos espaciales: El espacio A es continuo y aleatorio, puesto que el análisis se da en una región determinada, pero los puntos están ubicados de acuerdo con su propia naturaleza. En la presente tesis se utilizará este enfoque.

2.2.8. Patrones de Puntos Espaciales

Un patrón de puntos espaciales es el comportamiento de una serie de puntos $s_i = (s_{i1}, s_{i2})$ dentro de una región continua R que pueden ser un conjunto de puntos distribuidos de forma aleatoria o puntos que forman conjuntos de puntos. Estos puntos s_i son las ubicaciones en donde se ha producido un evento (fenómeno). La naturaleza de los patrones de los puntos dependerá de la naturaleza del evento que se está estudiando. Una forma sencilla de clasificar estos patrones es denominarlos como regulares, aleatorios o agrupados (ver Fig. 9). Los patrones de puntos espaciales no solo se limitan a analizar un evento, si no también varios eventos que ocurren al mismo tiempo dando origen al análisis de patrones de puntos espaciales multivariados y, si se incluye la variable temporal en una misma región, se da paso a los patrones de puntos espaciotemporales. (Diggle, 2014).

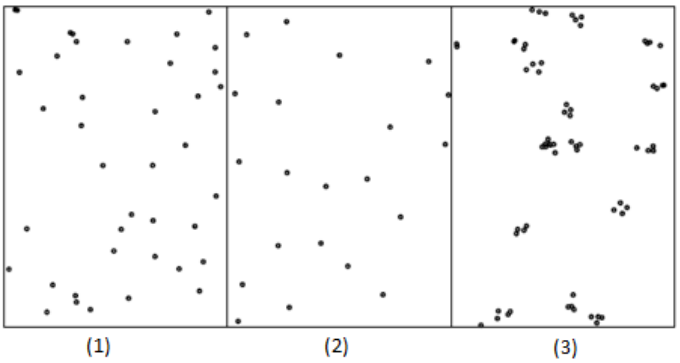

Figura 9: Patrones de puntos espaciales: aleatorio (1), regular (2) y agrupado (3).
Fuente: *Illian, Penttinen, Stoyan, & Stoyan (2008)*

Un aspecto importante a la hora de realizar un análisis de patrones de puntos espaciales es definir la ventana de observación o el área de estudio A. En algunos casos, A está definido a priori por la naturaleza del fenómeno y cuando este fenómeno es local el área de estudio es finita, pero lo más común es estar frente a un conjunto A tan grande que tomar cualquier muestra adecuada permite hacer un análisis estadístico. (Illian, Penttinen, Stoyan, & Stoyan, 2008)

A continuación, se presentan algunas definiciones necesarias para el análisis de patrones puntuales espaciales

Sea $B \in R^d$ un conjunto de puntos y se define la función N sobre el conjunto B tal que $N(B)$ es el número aleatorio de puntos en B y se asume que $N(B) < \infty$ para todo B. Se deduce que $N(B_1 \cup B_2) = N(B_1) + N(B_2)$, debido a que B_1 y B_2 son disjuntos. N también puede denotar el conjunto de todas las observaciones $x_1, x_2, \ldots$ que son parte del fenómeno de estudio. Se denota como:

$$N = \{x_i : i = 1,2,3, \ldots\} \text{ o } N = \{x_1, x_2, \ldots\}$$

2.2.8.1. Estacionariedad e isotropía

La estacionariedad indica que un punto y otro están relacionados únicamente por la distancia que hay entre ellos y no por la presencia del evento en sí; por otro lado, isotropía se refiere a que la relación entre dos puntos en el espacio depende únicamente de la distancia entre ellos y no de la dirección u orientación entre ellas. (Waller & Gotway, 2004)

2.2.8.2. Procesos de puntos estacionarios

Los procesos de puntos estacionarios hacen referencias a los procesos puntuales que asumen condiciones de estacionariedad u homogeneidad y que, además, son subconjuntos de patrones de puntos infinitos.

Un proceso $N = \{x_1, x_2, \dots\}$ en $\mathbb{R}^d$ es estacionario si N y el proceso puntual trasladado $N_x = \{x_1 + x, x_2 + x, \dots\}$ tienen la misma distribución para todo x.

2.2.8.3. Intensidad

La intensidad se define como el número promedio de observaciones en una unidad de área y su uso es similar a la del promedio muestral en la estadística clásica. Cuando la intensidad es no constante permite construir mapas de intensidad en donde se remarcan las zonas en las cuales es mayor. (Illian, Penttinen, Stoyan, & Stoyan, 2008)

El promedio del número de puntos de B satisface:

$$E\big(N(B)\big) = \lambda \cdot v(B) \tag{2.2}$$

Donde λ es la intensidad y B es un subconjunto de $\mathbb{R}^d$. La interpretación de λ es sencilla, puesto que si consideramos $v(B) = 1$ en la ecuación (2.2) el

número de puntos esperados en B es simplemente λ. Por lo tanto, la intensidad es interpretada como el número de puntos esperados por unidad de área.

La ecuación (2.2) puede ser generalizada para el comportamiento promedio de procesos puntuales no necesariamente estacionarios con la siguiente expresión:

$$\Lambda(B) = E\big(N(B)\big) = \text{Número promedio de puntos de } N \text{ en } B.$$

Si la distribución de N coincide con el proceso puntual trasladado N_x, tenemos:

$$E\big(N(B)\big) = E\big(N_x(B)\big),$$

y desde que $N_x(B) = N(B - x)$

$$E\big(N(B)\big) = E\big(N(B - x)\big).$$

Como x ha sido elegido arbitrariamente, podemos decir que: $E\big(N(B)\big) = E\big(N(B_x)\big)$ con lo que:

$$\Lambda(B) = \Lambda(B_x) \text{ para todo conjunto } B \text{ y } x \in \mathbb{R}^d.$$

Un teorema de la medida indica que la medida de traslación invariante Λ debe ser múltiplo de la medida de Lebesgue o de su área o de su volumen (Illian, Penttinen, Stoyan H., & Stoyan D., 2008, p. 175); es decir:

$$\Lambda(B) = \lambda \cdot v(B)$$

La intensidad es una propiedad de primer orden. Las propiedades de primer y segundo orden son importantes para realizar el análisis de patrones de puntos espaciales y ambos brindan información diferente, pero complementaria para entender el tipo de comportamiento de los puntos espaciales estudiados (Waller & Gotway, 2004).

La propiedad de primer orden mide la distribución de eventos en una región: intensidad y dependencia espacial. La propiedad de segundo orden mide la tendencia de eventos para aparecer agrupados, independientemente o espaciados regularmente (Cressie, 1993). La estacionariedad de segundo orden está definida como la densidad del producto $\varrho(x_1, x_2)$ y depende únicamente de la diferencia de $h = x_1 - x_2$, puesto que por estacionariedad:

$$\varrho(x_1, x_2) = \varrho(x_1 + x, x_2 + x) \text{ para todo } x, x_1 \text{ y } x_2$$

Si N es invariante al movimiento, entonces la densidad del producto depende solo de la distancia r entre x_1 y x_2 o la longitud de h, $r = \| h \|$, entonces la notación se simplifica a $\varrho(r)$ con r positivo. (Illian, Penttinen, Stoyan H., & Stoyan D., 2008, p. 176).

2.2.8.4. Aleatoriedad Espacial Completa.

El modelo teórico que sirve de base para el análisis de los patrones de puntos espaciales es la aleatoriedad espacial completa (*Complete Spatial Randomness: CSR)* que tiene las siguientes características (Diggle, 2014):

- El número de eventos en una región B con área $v(B)$ sigue una distribución de Poisson con media $\lambda v(B)$, donde λ es la intensidad constante.

- Dado n eventos x_i en una región B, los eventos x_i son independientes entre sí y están uniformemente distribuidos sobre la región B.

Para demostrar que existe dependencia espacial o patrones puntuales es necesario probar la hipótesis de aleatoriedad espacial completa para determinar si los eventos espaciales son regulares o agrupados (Diggle, 2014).

2.2.8.5. Efecto Frontera

Al igual que como en el análisis de sobrevivencia se tienen observaciones censuradas, cuando se trabaja con patrones de puntos espaciales es importante tener en cuenta que hay puntos fuera de la región de estudio (y muy cerca a los limites) que están afectando al comportamiento de los puntos más cercanos dentro de la región y este es un problema común en la estadística espacial, puesto que el estadístico que se use puede no captar el patrón subyacente en la región; en consecuencia, es importante hacer correcciones cuando se trabaja en regiones pequeñas. Si la región de estudio y el número de puntos son lo suficientemente grandes, la corrección por el efecto frontera puede ser ignorada. (Illian, Penttinen, Stoyan, & Stoyan, 2008)

2.2.8.6. Distancia al vecino más cercano

Para n eventos en la región B, sea y_i la distancia desde el i-ésimo evento al evento más cercano también en B. La distancia y_i es llamada la distancia al vecino más cercano. (Diggle, 2014, p. 24).

La distribución teórica de la distancia al vecino más cercano y bajo aleatoriedad espacial completa depende de n y de B y no es expresable en forma cerrada por el efecto frontera; sin embargo, una aproximación que ignora el efecto frontera se obtiene al expresar el área de B como $v(B)$, entonces $\pi y^2 v(B)^{-1}$ es la probabilidad bajo aleatoriedad espacial completa que un evento arbitrario esté dentro un evento específico y la de la distancia y. Dado que los eventos se ubican independientemente bajo aleatoriedad espacial completa, la función de distribución aproximada de Y es

$$G(y) = 1 - (1 - \pi y^2 v(B)^{-1})^{n-1}.$$

Y de la ecuación (2.2), una aproximación para n grande, se puede expresar como $\lambda = n \cdot v(B)^{-1}$, entonces:

$$G(y) = 1 - \exp(-\lambda\pi y^2)\,;\;\; y \geq 0 \qquad (2.3)$$

2.2.8.7. Estimación de la intensidad

Supongamos que la región A es dividida en subregiones cuadradas de tamaño $a \times a$ (de un tamaño no tan grande para no destruir el patrón espacial), se cuenta el número de eventos que cayeron dentro de cada cuadrante y de esa manera resumir el patrón espacial. (Cressie, 1993)

La estimación de la intensidad se puede realizar mediante el método no paramétrico de Kernel.
Sea $s_1, s_2, \ldots, s_n$ ubicaciones de $n = N(B)$ eventos en una región B, entonces el estimador de la intensidad de Kernel tiene la siguiente forma:

$$\hat{\lambda}_h(s) = \frac{1}{p_h(s)}\left\{\sum_{i=1}^{n} \frac{1}{h^2} k\left(\frac{s - s_i}{h}\right)\right\}.$$

donde, el valor de $p_h(s)$ es el factor de corrección por el efecto frontera denotado de la siguiente igualdad:

$$p_h(s) = \int_B \frac{1}{h^2} k\left(\frac{s-u}{h}\right) du.$$

Según lo mencionado en la sección anterior, no es necesario el uso del factor de corrección por el efecto frontera. La estimación de densidad Kernel tendrá un ancho de banda $h = 0.68n^{-0.2}$, que es el valor recomendado cuando A es el cuadrado unitario. Debido a que la estimación de densidad se realizará mediante el kernel cuartico, este será reemplazado en la ecuación de la estimación de densidad y se obtiene:

$$\hat{\lambda}_h(s) = \sum_{d_i \leq h} \frac{1}{\pi h^2} \left(1 - \frac{{d_i}^2}{h^2}\right)^2 \quad (2.4)$$

donde, d_i representa la distancia entre la localización s_i y el punto s.

2.2.8.8. *Estimación de la función G*

Se puede estimar una función de distribución empírica de las distancias de los vecinos más cercanos de una región B con la siguiente expresión:

$$\hat{G}(y) = n^{-1} \cdot \#(y_i \leq y) \quad (2.5)$$

Donde, "#" representa la expresión "el número de" y n es el número de eventos en la región B.

La expresión en (2.5) puede ser nombrada como $\hat{G}_1(y)$, la cual puede ser comparada con valores superiores e inferiores de simulaciones de n eventos independientes y uniformemente distribuidos en B denotados como $\hat{G}_i(y)$ con $i = 1,2, \dots, s$. Para contrastar $\hat{G}_1(y)$ se usa la siguiente medida de discrepancia:

$$u_i = \int \left\{\hat{G}_i(y) - \bar{G}_i(y)\right\}^2 \partial y \quad (2.6)$$

Donde:

$$\bar{G}_i(y) = (s-1)^{-1} \sum_{j \neq i} \hat{G}_j(y)$$

La función de densidad estimada $\hat{G}(y)$ se puede comparar con la aproximación teórica de la expresión (2.3) o mediante las comparaciones de la expresión (2.6) para ver si estamos frente a un proceso con estacionariedad espacial completa. Si la curva estimada, está por encima de la teórica, se dice que no hay estacionariedad espacial completa.

2.2.8.9. Estimación de la función L

De igual manera como se estudió la concentración de los datos mediante la propiedad de primer orden es importante realizar el análisis para medidas de segundo orden que estudian los patrones entre eventos. Para ello se define la función K que es la más común. (Waller & Gotway, 2004)

$$K(h) = \frac{E[Número\ de\ eventos\ a\ distancia\ h\ de\ un\ evento\ cualquiera]}{\lambda}$$

El enunciado que define a $K(h)$ asume un proceso isotrópico y estacionario, así que se espera que λ tenga un valor constante para cualquier punto elegido. Bajo la hipótesis de que se trate de un proceso de aleatoriedad espacial completa se espera que $K(h) = \pi h^2$. Entonces para un proceso regular se esperaría que $K(h) < \pi h^2$, mientras que para un proceso agrupado se espera $K(h) > \pi h^2$. (Illian, Penttinen, Stoyan, & Stoyan, 2008)

El estimador de la función $K(h)$ tiene la siguiente forma:

$$\widehat{K}(h) = \frac{1}{N\hat{\lambda}} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \delta(d(i,j) < h)\,.$$

donde, N es el número de eventos que se realizaron, $d(i,j)$ denota la distancia entre los eventos i y j; y $\delta(d(i,j) < h)$ es una función indicadora que toma el valor de 1 si $d(i,j) < h$ y 0 caso contrario. Sin embargo, esta medida no es del todo robusta cuando los puntos se encuentran cerca de los límites del área en estudio y por ello se propone el estimador de $K(h)$ con un factor de corrección por efecto frontera. (Illian, Penttinen, Stoyan, & Stoyan, 2008)

$$\widehat{K}_{ec}(h) = \frac{1}{\hat{\lambda}} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij}{}^{-1} \delta(d(i,j) < h)\,.$$

Cuando $d(i,j)$ es menor a la distancia entre i y el borde del área de estudio $w_{ij} = 1$. Para realizar el diagnóstico con la función K, se realizar una transformación para realizar gráficas que permitan realizar una acertada conclusión. Como teníamos que $K(h) = \pi h^2$, entonces:

$$\sqrt{\frac{K(h)}{\pi}} = h.$$

Se define

$$\hat{L}(h) = \sqrt{\frac{\hat{K}_{ec}(h)}{\pi}}.$$

Se gráfica h versus $\hat{L}(h) - h$ mediante el método de simulación de Montecarlo se hace la estimación de $\hat{L}(h)$ para distintos valores de h y se espera que la curva empírica esté por encima de la curva teórica para dar evidencias de la presencia de patrones espaciales (Ver fig. 10)

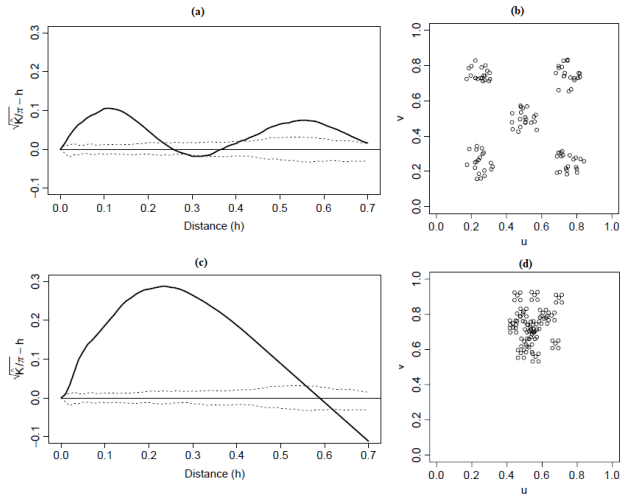

Figura 10: (a) Función K estimada de patrones regulares de grupos, (b) Diagrama de patrones regulares de grupos, (c) Función K estimada de grupos de patrones regulares, (d) Diagrama de grupos de patrones regulares.

Fuente: *Waller & Gotway (2014)*

2.2.9. Definiciones Técnicas de Agricultura.

Adopción Tecnológica. Se define tecnología que es un conjunto de herramientas, conocimientos o servicios que hacen que ciertas tareas sean más fáciles, rápidas y sencillas de usar ahorrando tiempo y recursos. Por otro lado, adopción es definida como un proceso mental en el cual se pasa de escuchar y conocer sobre innovación a implementarla en su día a día, Entonces la adopción tecnológica es el proceso por el cual el agricultor se informa sobre la tecnología y posteriormente la usa en sus cultivos como una práctica adquirida. (Mwangi & Kariuki, 2015).

Productores agropecuarios. El término agropecuario viene de combinar los términos agricultura y pecuario. El primero se refiere a personas que realizan actividades de producción de material vegetal o animal para alimentos, fibras y otros usos. El segundo se refiere a la actividad ganadera para la producción de carne o derivados. (Yunlong & Smit, 1994)

Pequeño productor agropecuario. Categorizar a un productor agropecuario como pequeño ha entrado en debate por mucho tiempo, puesto que hay muchas características que pondrían hacer que un productor pertenezca a esta categoría. La primera es por la extensión de tierras cultivadas y se considera un productor pequeño aquel que tiene 5 hectáreas o menos, otra definición es que se consideran pequeño productor agropecuario a quienes tienen las actividades agropecuarias como principal fuente de ingreso familiar y destinan la mayor parte de la producción para su subsistencia. (Macías, 2016).

Pobreza agropecuaria. La pobreza agropecuaria es medida de acuerdo con la productividad agrícola, condiciones de vida y porcentaje de la producción destinada a la venta y es considerada como una de las principales razones de la pobreza rural (De Janvry & Sadoulet, 2009).

Sector de Enumeración Agropecuaria. El sector de enumeración agropecuaria (SEA) que corresponde a una porción de territorio en la cual se localizan en promedio 100 unidades agropecuarias en la que se aplicó la entrevista durante el proceso de empadronamiento.

CAPITULO III: METODOLOGÍA

3.1. Fuente de datos

La fuente de datos de esta investigación procede del Censo Nacional Agropecuario realizado por el Instituto Nacional de Estadística e Informática (INEI) en el año 2012.

La investigación es de tipo cuantitativa descriptiva. Los datos proceden de una fuente secundaria.

3.2. Población y unidad de análisis

La población para la cual realizará el estudio corresponde a todos los pequeños productores agropecuarios del Perú que fueron encuestados en el Censo Nacional Agropecuario (CENAGRO) 2012 realizado por el INEI, incluyendo a productores agropecuarios que contaban con una superficie agrícola menor o igual a 5 hectáreas y con registros completos. La población fue de 1’529,571 pequeños productores agropecuarios. La unidad de análisis para la creación del índice de adopción tecnológica es cada uno de los pequeños productores agropecuarios, mientras que la

unidad de análisis para el cálculo de las densidades focales de Kernel son los Sectores de Enumeración Agrícola (SEA).

3.3. Variables

Se consideraron 15 variables para el cálculo del índice de adopción tecnológica, posteriormente junto con otras 6 variables que se refieren a las características del hogar, se construirán los grupos de pequeños productores agropecuarios.

Las variables para el Índice de Adopción Tecnológica se refieren al uso en las parcelas de:

- Vacunación animal
- Dosificación
- Dieta Balanceada
- Capacitación
- Asistencia técnica
- Inseminación artificial
- Semillas artificiales
- Fertilizantes químicos
- Insecticidas químicos
- Fungicidas
- Herbicidas
- Energía eléctrica
- Tractor
- Fumigadores
- Molino eléctrico
- Trilladora
- Motor bomba
- Camión

- Riego por aspersión
- Riego por goteo
- Riego por exudación
- Control biológico

Para la creación de agrupamientos se consideró la información del agricultor:

- Índice de Adopción Tecnológica
- Sabe leer y escribir
- Grado de instrucción
- Uso de cocina mejorada en la vivienda
- Tipo de servicio higiénico en la vivienda
- Cuenta con computadora
- Tenencia propia de la vivienda

3.4. Procedimiento de análisis

El procedimiento de análisis consistió en los siguientes pasos:

i. Calcular el índice de adopción tecnológica mediante el análisis de correspondencias múltiples.
ii. Seleccionar a los productores agropecuarios con un nivel de adopción tecnológico muy bajo.
iii. Crear agrupaciones con los pequeños productores agropecuarios seleccionados en el paso (ii) con la técnica de redes neuronales de Kohonen.
iv. Agrupar las neuronas de salida del paso (iii) en base al número óptimo de agrupaciones.
v. Seleccionar las agrupaciones con productores agropecuarios que tengan características de menor desarrollo.
vi. Identificar los Sectores de Enumeración Agropecuaria a los cuales pertenecen los productores seleccionados en el paso (v)

vii. Realizar el análisis espacial en los Sectores de Enumeración Agropecuaria identificados en (vi) y encontrar los focos de densidad formados
viii. Caracterizar los focos de densidad en función a las características de los productores agropecuarios.

3.5. Software utilizado

El software utilizado para el análisis de los datos es R Project for Statistical Computing con el IDE RStudio los cuales son de código abierto.

CAPITULO IV: RESULTADOS Y DISCUSIÓN

4.1. Índice de Adopción Tecnológica

El cálculo del índice se realizó en el software R y se siguieron las sugerencias de codificación y uso de gráficos de Kassambara (2017, pp. 83 - 106).

La tabla 2 muestra los valores propios, el porcentaje de varianza y el porcentaje de varianza acumulada para cada una de las dimensiones. Los valores propios sugieren que se elijan las 10 primeras dimensiones, puesto que -según el criterio de Greenacre (1984)- se toman en cuenta un número de dimensiones igual al número de valores propios que cumplan con la condición de ser mayores a 1/K (donde K es el número total de variables en el análisis de correspondencias). En nuestro caso k = 22 y, por tanto, se conservarán las dimensiones cuyos autovalores tengan un valor propio asociado mayor o igual a 0.045 que resultan ser las 9 primeras dimensiones. La varianza acumulada de las 9 dimensiones elegidas representa un 57.858% de la varianza total de los datos.

Sin embargo, las 9 primeras dimensiones no fueron suficientes para explicar el comportamiento de todas las variables (ver fig. 11 donde el color azul más oscuro indica que hay una mayor contribución total a la dimensión asociada).

Tabla 1: Resultados del ACM: valores propios, porcentaje de varianza y porcentaje de varianza acumulada

Dimensiones	Valores Propios	Porcentaje de varianza	Porcentaje de varianza acumulado
Dim.1	0.143	14.281	14.281
Dim.2	0.079	7.872	22.153
Dim.3	0.067	6.685	28.837
Dim.4	0.058	5.806	34.643
Dim.5	0.05	4.955	39.598
Dim.6	0.047	4.731	44.329
Dim.7	0.046	4.574	48.904
Dim.8	0.045	4.485	53.389
Dim.9	0.045	4.469	57.858
Dim.10	0.044	4.354	62.212
Dim.11	0.042	4.246	66.458
Dim.12	0.041	4.065	70.523
Dim.13	0.04	4.008	74.531
Dim.14	0.038	3.773	78.304
Dim.15	0.036	3.649	81.953
Dim.16	0.036	3.581	85.534
Dim.17	0.033	3.308	88.841
Dim.18	0.028	2.764	91.606
Dim.19	0.025	2.475	94.081
Dim.20	0.024	2.418	96.499
Dim.21	0.019	1.916	98.415
Dim.22	0.016	1.585	100

Fuente: *Elaboración Propia*

Se usaron un total de 13 dimensiones que explican el 72% de la varianza de los datos. Las dimensiones elegidas fueron la dimensión 1 explica el uso de semillas artificiales, fertilizantes químicos, insecticidas químicos, herbicidas y fungicidas que hace referencia a los cuidados de los cultivos; la dimensión 2 explica el uso de asistencia técnica, capacitaciones e inseminación artificial.

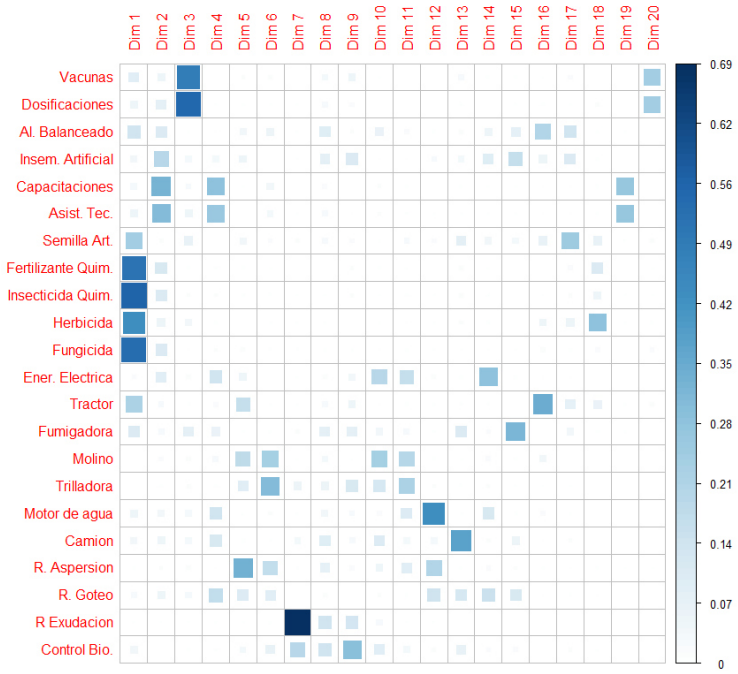

Figura 11: Contribuciones de las variables sobre cada dimensión

Fuente: *Elaboración propia*

La dimensión 3 corresponde al uso de vacunas y dosificaciones; la dimensión 4 que aporta al riego por goteo; la dimensión 5 aporta principalmente al riego por aspersión; la dimensión 4 hace referencia al uso de trilladora y molino; la dimensión 7 explica el uso de riego por exudación; la dimensión 9 aporta al uso de control biológico del cultivo; las dimensiones 12 y 13 explican el uso de motor de agua y camión respectivamente.; la dimensión 15 aporta al uso de fumigadora; la dimensión 16, al uso de tractor y la dimensión 17 aporta al uso de alimento balanceado.

La nube de puntos de variables y modalidades nos ayuda a armar perfiles de las preferencias de los productores agropecuarios y como se asocian las variables entre sí.

La gráfica de variables (ver fig.12(2)) indica que existe una fuerte asociación entre el uso de tractor y semillas artificiales, también entre el uso de energía eléctrica, dosificación animal; por otro lado, se puede ver una fuerte asociación entre uso de riego por goteo, camión de cultivo y motor de agua para bombeo. También se ve la existencia de la asociación entre quienes participan en programas de capacitación y cuentan con asistencia técnica, del mismo modo el uso de fertilizantes está asociado al uso de fungicidas, insecticidas y herbicidas; sin embargo, la gráfica de variables no ayuda a entender si está asociación es debido al uso o no de la tecnología; por ello, se utiliza la nube de puntos de modalidades (categorías de las variables).

La gráfica de modalidades indica que hay fuerte relación entre quienes usan asistencia técnica, inseminación artificial y energía eléctrica; por otro lado, existe un perfil de productores que se caracterizan por el no uso de fungicidas, herbicidas, insecticidas y fertilizantes químicos. También se puede ver disimilaridades entre los distintos tipos de riego, sobre todo el riego por goteo que presenta una distancia mayor con respeto a los demás. Se observa que la mayoría de todas las modalidades que se refieren al uso de tecnologías están en los cuadrantes positivos de la dimensión 1, mientras que las modalidades que corresponden a los no usos de tecnología están en el lado izquierdo de la dimensión 1, lo cual muestra perfiles totalmente opuestos entre los adoptantes y no adoptantes de tecnología. A pesar de que la gráfica de variables y la de modalidades muestren la posible aparición de celdas discordantes que podrían estar afectando en los resultados del indicador, en el Anexo 10 se muestra la correlación positiva existente entre las puntuaciones totales de las variables y el indicador de adopción tecnológica demostrando que sus resultados son correctos.

El Índice de Adopción Tecnológica se calculó a partir de las 13 dimensiones elegidas siguiendo la metodología explicada en la subsección 2.2.2 del presente trabajo (página 20).

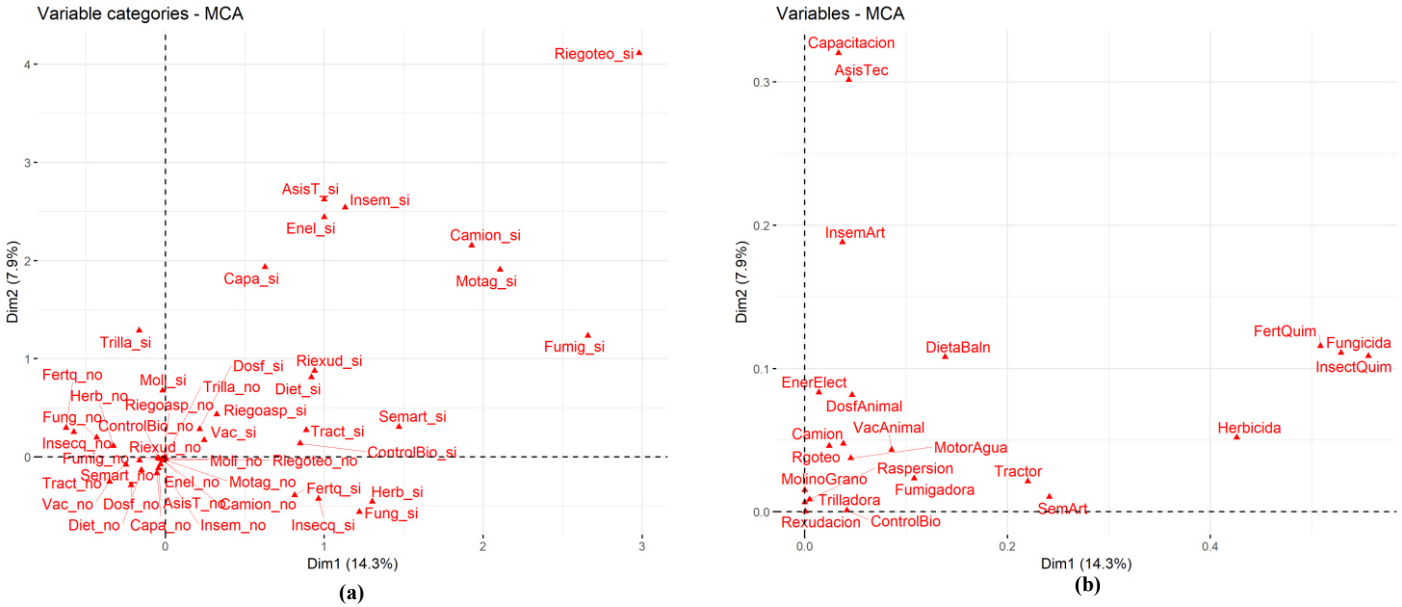

Figura 12: Nube de puntos de modalidades (a) y nube de variables (b) en base a las dimensiones 1 y 2

Fuente: *Elaboración propia*

El indicador de adopción tecnológica varía entre 0 y 100 puntos y evidencia que el 88% de los productores agropecuarios está por debajo de una puntuación de 10 con la mayoría con puntuaciones entre 2 y 4 (ver figura 13). Lo cual indica condiciones de adopción tecnológica realmente bajas.

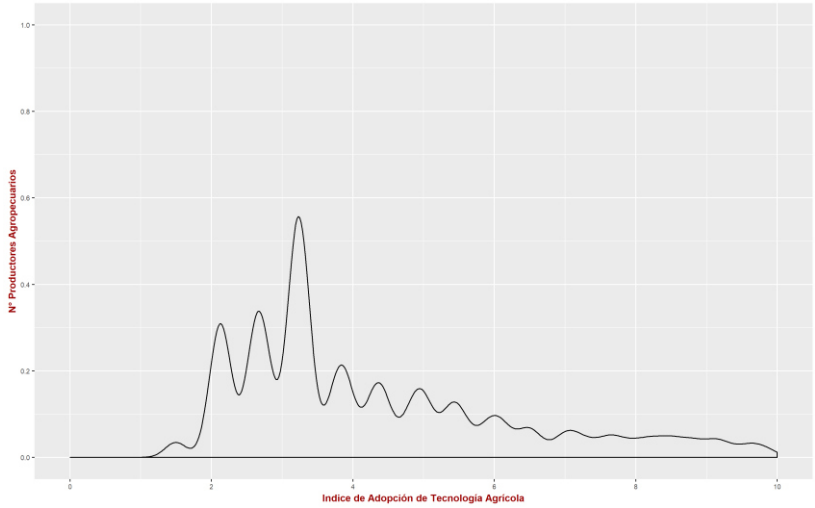

Figura 13: Función Densidad del Índice de Adopción Tecnológica del 88% de los productores agropecuarios. CENAGRO 2012
Fuente: *Elaboración propia*

El índice de adopción tecnológica se categorizó en 5 niveles que van desde muy bajo hasta muy alto (ver tabla 3) en donde se puede ver que el 98.1% de los productores agropecuarios están en la categoría más baja. Es importante resaltar que únicamente 31 productores han tenido niveles de adopción tecnológico muy alto. Se hizo una selección de los productores agropecuarios que tienen un nivel de adopción tecnológica muy bajo para realizar una agrupación de ellos.

Tabla 2: Categorías del nivel de Adopción Tecnológica

Rango	Nivel de Adopción Tecnológico	N° de pequeños productores agropecuarios	Porcentaje
0 – 20	Muy bajo	1501173	98.143
20 – 40	Bajo	26027	1.702
40 – 60	Medio	2085	.136
60 – 80	Alto	255	.017
80 – 100	Muy alto	31	.002
	Total	1529571	100.0

Fuente: *Elaboración propia*

4.2. Agrupamiento de los pequeños productores Agropecuarios

Se realizó el análisis de agrupamientos con el método de redes neuronales de Kohonen. Se trabajó con un total de 1’182,459 productores agropecuarios con un nivel de adopción tecnológica muy bajo (ver Tabla 3). Para realizar el análisis de agrupamientos se usó el Índice de Adopción Tecnológica y con las variables de condiciones de vida tales como: poseer servicio higiénico, saber leer y escribir, tener computadora en casa, tener acceso a internet, contar con cocina mejorada, la condición de tenencia de la vivienda propia, el nivel educativo (que fue categorizado en 6 niveles), si la actividad agropecuaria le da ingresos suficientes y si trabaja en actividades distintas a la actividad agropecuaria. Esto permitió hacer una reducción de 1.2 millones de observaciones a 100 (neuronas).

El uso de las redes neuronales de Kohonen tiene dos ventajas: La primera es que permite encontrar a los productores agropecuarios con bajo nivel de desarrollo, puesto que solo con el índice de adopción tecnológica no es posible hacer una agrupación debido a factores que no son considerados por él mismo; la segunda se refiere a la ganancia computacional, ya que se está trabajando con más de 1.5 millones de datos y hacer una clasificación mediante los métodos tradicionales toma demasiado tiempo y, en ocasiones, es imposible realizarse porque no se puede guardar la matriz

de distancias debido a que esta pesa aproximadamente 625 GB debido al gran número de observaciones. Los grupos que determinen los mapas auto organizados de Kohonen permitirán encontrar los grupos que serán tomados en cuenta para la aplicación de las densidades focales de Kernel.

Las redes neuronales de Kohonen fueron aplicadas sobre aquellos agricultores que tenían un Índice de Adopción Tecnológica muy bajo con la finalidad de encontrar entre ellos a los que son más vulnerables, puesto que muchos de ellos a pesar de tener bajos niveles tecnológicos, tienen otras actividades económicas que les permite compensar las carencias económicas. Además, los métodos de agrupamiento nos permitirán identificar a aquellos productores que tienen buenas condiciones de vida a pesar de no tener una adopción tecnológica alta.

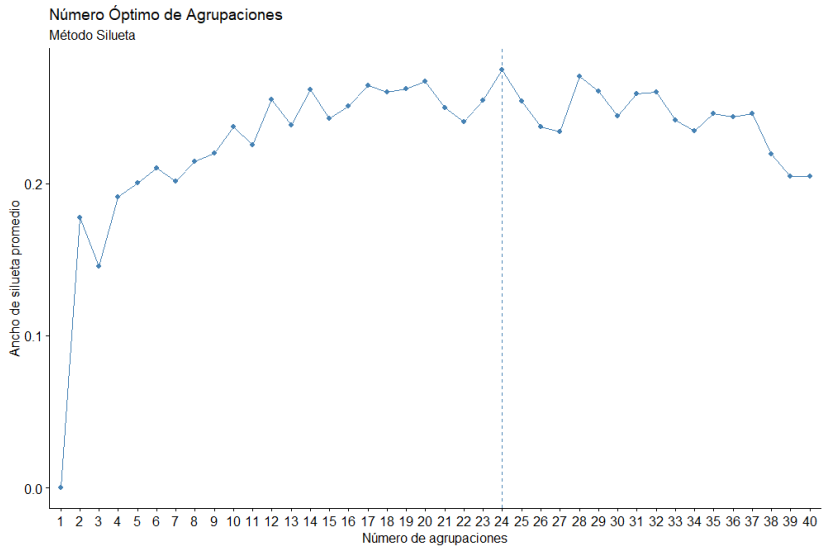

Figura 14: Gráfico del ancho de silueta para determinar el número óptimo de clústeres

Fuente: *Elaboración propia*

Al aplicar la técnica del ancho de silueta sobre las 100 neuronas obtenidas mediante los mapas auto organizados de Kohonen se obtiene que el número óptimo de agrupaciones es igual a 24 y las 100 neuronas son divididas en 24 agrupaciones distintas. Cada variable tuvo una diferente influencia en cada agrupación y está reflejada en el mapa auto organizado de Kohonen; por ejemplo, el grupo rojo (ver fig. 15) se caracteriza porque agrupa a los productores agropecuarios que saben leer y escribir y tienen en su mayoría tierras propias, mientras que la agrupación color salmón es principalmente caracterizada por productores que tienen ingresos suficientes por las actividades agropecuarias.

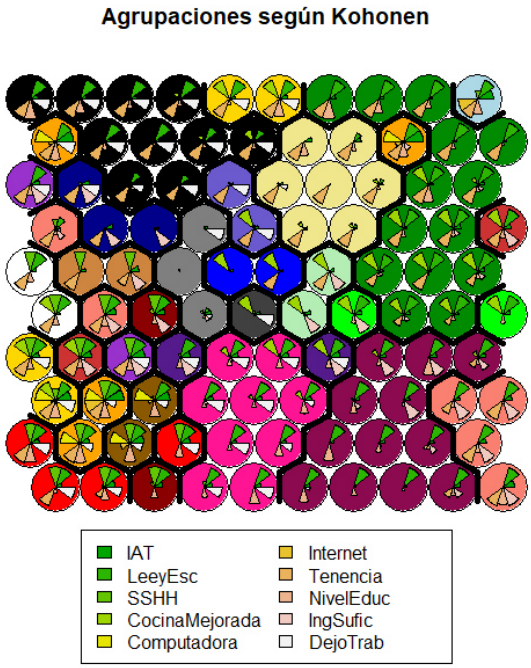

Figura 15: Mapa Autoorganizado de Kohonen dividido según número óptimo de clúster
Fuente: *Elaboración propia*

La Tabla 4 muestra las distintas características de las 24 agrupaciones que se formaron a raíz de las redes neuronales de Kohonen. De las 24 agrupaciones que se

muestran, se eligieron únicamente 16 que, según sus características, son aquellos grupos que contienen productores agropecuarios con menor desarrollo. Estas agrupaciones tienen la característica que al menos el 50% de los productores agropecuarios indicaron no tener ingresos suficientes con la actividad agropecuaria y además no tienen acceso a tecnología: ni computadora, ni internet. Los grupos descartados del mapeo espacial son aquellos en la que el 100% de los productores agropecuarios indican que la actividad agropecuaria es suficiente para cubrir sus gastos. También se dejaron fuera aquellos grupos en los cuales, a pesar de que pocos productores agropecuarios indicaron que el dinero proveniente de la agricultura no cubre sus gastos, pero cuentan con computadora y/o internet y tienen trabajo fuera de las actividades agropecuarias.

Tabla 3: Principales características de los clústeres generados

Segmento	Número de productores	Sabe leer y escribir	Cuenta con ss.hh. dentro de la vivienda	Cocina mejorada en la vivienda	Tiene computadora en casa	Acceso a Internet	Tenencia Tierras propias	Indicador de Adopción Tecnológica	Grado de instrucción	Superficie Agrícola (ha.)	Ingresos Suficientes	Trabajo fuera de las parcelas
1	57976	100.0%	100.0%	0.0%	0.0%	0.1%	74.4%	5.76	4	1.01	0.0%	100.0%
2	18804	100.0%	100.0%	25.4%	1.6%	0.1%	0.0%	5.97	3	0.96	31.3%	0.0%
3	102948	100.0%	4.9%	28.5%	0.0%	0.0%	0.0%	5.34	3	1.16	0.0%	100.0%
4	113633	100.0%	0.0%	0.0%	0.2%	0.0%	0.0%	4.82	3	1.39	27.0%	0.0%
5	126457	100.0%	12.4%	0.0%	0.6%	0.0%	100.0%	5.81	3	1.62	100.0%	0.0%
6	19568	99.2%	60.1%	3.6%	100.0%	42.1%	99.3%	7.37	5	1.23	16.8%	59.5%
7	7282	99.7%	65.4%	19.3%	100.0%	40.2%	0.0%	7.30	5	1.06	11.9%	67.3%
8	89145	100.0%	19.0%	100.0%	4.5%	1.2%	100.0%	5.26	3	1.16	1.2%	100.0%
9	59130	100.0%	12.6%	100.0%	1.0%	0.1%	100.0%	5.68	3	1.48	100.0%	23.2%
10	35974	100.0%	17.3%	0.0%	1.9%	0.0%	100.0%	6.17	3	1.59	100.0%	100.0%
11	17367	100.0%	15.7%	28.4%	1.0%	0.0%	0.0%	6.29	3	1.41	100.0%	100.0%
12	45802	100.0%	100.0%	25.5%	1.4%	0.0%	100.0%	5.46	3	1.11	0.0%	0.0%
13	33680	0.0%	8.5%	0.0%	0.2%	0.1%	0.0%	4.42	1	1.00	0.0%	29.1%
14	3610	0.0%	11.8%	100.0%	0.4%	0.1%	0.0%	5.05	1	0.88	0.0%	100.0%
15	14083	0.0%	7.5%	100.0%	0.3%	0.1%	75.7%	5.14	1	1.20	100.0%	13.8%
16	35404	100.0%	0.0%	100.0%	0.6%	0.1%	0.0%	5.37	3	1.19	32.7%	0.0%
17	317814	100.0%	0.0%	26.0%	0.2%	0.0%	100.0%	4.90	3	1.23	0.0%	0.0%
18	9525	0.0%	100.0%	0.0%	1.6%	0.7%	92.1%	5.09	1	0.89	31.6%	1.1%
19	29843	0.0%	7.3%	100.0%	0.3%	0.1%	78.3%	4.47	1	0.98	0.0%	0.0%
20	38853	0.0%	1.0%	0.0%	0.2%	0.1%	78.5%	5.04	1	1.20	100.0%	11.9%
21	43064	0.0%	10.3%	25.9%	0.2%	0.1%	100.0%	4.76	1	0.96	0.0%	100.0%
22	74355	0.0%	0.0%	0.0%	0.1%	0.0%	100.0%	4.49	1	0.92	0.0%	0.0%
23	206823	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	5.30	3	1.21	0.0%	100.0%
24	33	100.0%	48.5%	0.0%	0.0%	100.0%	100.0%	6.79	5	1.70	0.0%	0.0%

4.3. Determinación de los focos de alta densidad

Se analizó un total de 720 Sectores de Enumeración Agropecuaria con al menos el 80% pequeños productores agropecuarios con bajos niveles de desarrollo que se distribuyen a lo largo del país -principalmente en la sierra- (ver figura 16). Sobre las ubicaciones de cada uno de los Sectores de Enumeración Agropecuaria se estudiaron las propiedades espaciales de primer y segundo orden para saber si existen patrones de agrupamiento entre las ubicaciones.

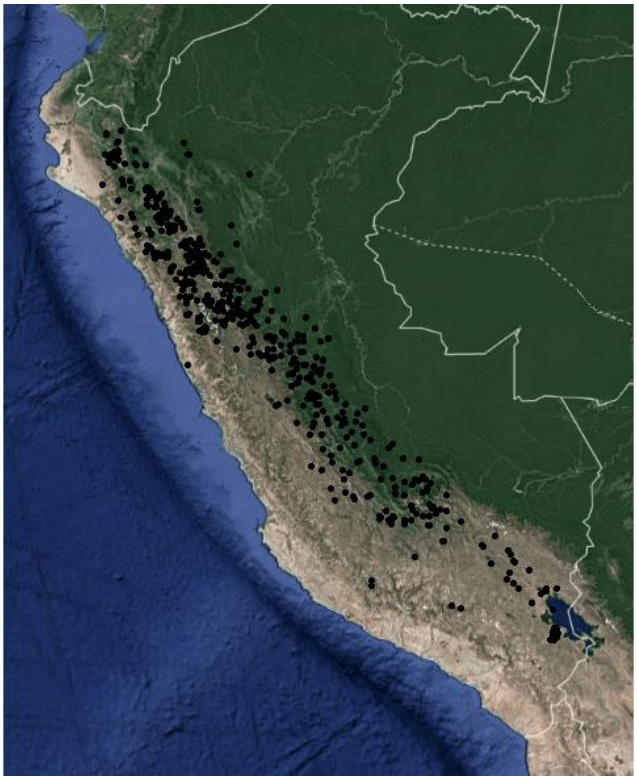

Figura 16: Mapa de las ubicaciones de los SEA de pequeños productores agropecuarios con bajo nivel de desarrollo
Fuente: *Elaboración propia*

Los resultados del análisis de patrones puntuales indican que, a lo largo de la ventana o región en la cual ocurren los eventos (ver fig. 17(a)), no estamos frente a un proceso de aleatoriedad espacial completa porque podemos encontrar un patrón de aglomeración de eventos y picos en el mapa de densidad (ver fig. 17(b)). Resaltan dos patrones bien definidos, uno en la zona norte y otro en la zona sur (cerca al lago Titicaca).

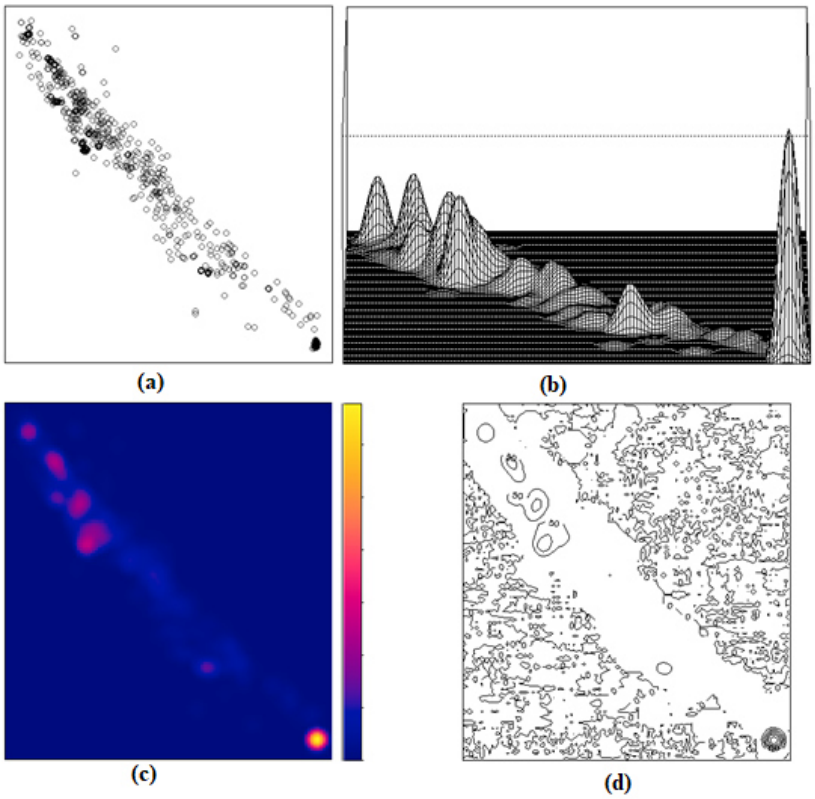

Figura 17: Ubicación de los eventos en ventana reducida (a), Estimación de densidad intensidad mediante kernel cuártico en tres dimensiones (b), estimación de la intensidad mediante kernel cuártico en dos dimensiones (c), gráfico de contorno de densidad (d).
Fuente: *Elaboración propia*

Al analizar la propiedad de primer orden estudia la dependencia espacial mediante la función G estimada la cual crece rápidamente y está por encima de la curva teórica (ver fig. 18(a)) indicando que hay muchos puntos que están cerca uno del otro con lo cual es suficiente para comprobar que no estamos frente a un proceso de aleatoriedad espacial completa. Además, se analizó la propiedad de segundo orden mediante la estimación de la función L (ver fig. 18(b)), en donde encontramos que esta crece rápidamente por encima de la curva teórica y confirma la presencia de patrones regulares de agrupamientos.

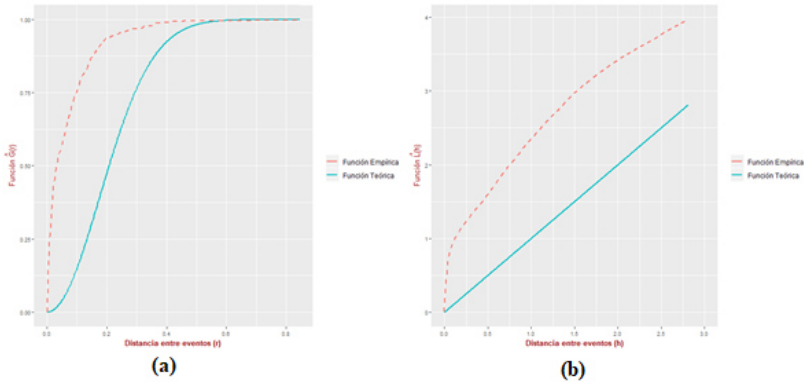

Figura 18: Función G estimada (a) y Función L estimada (b)
Fuente: *Elaboración propia*

Una vez que se demostró que no existe aleatoriedad espacial completa se realizó la estimación de densidad Kernel haciendo uso del Kernel cuartico. Se identificaron 8 focos de densidad en los cuales se apreció una alta densidad espacial de Sectores de Enumeración Agropecuarios con al menos un 80% de pequeños productores agropecuarios con bajos niveles de desarrollo. Los focos están ubicados en los departamentos de Puno, Piura, Cajamarca, La Libertad, Apurímac y tres focos en el departamento de Ancash (ver fig. 19).

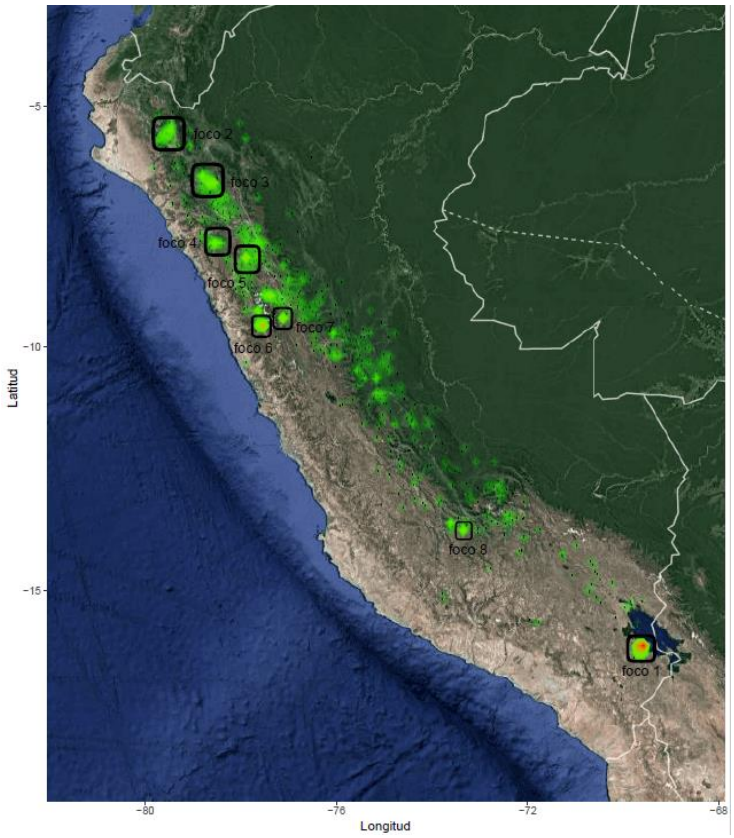

Figura 19: Mapa del Perú identificando los focos de alta densidad de SEA con agricultores que tienen muy bajos niveles de adopción de tecnologías y baja calidad de vida.
Fuente: *Elaboración propia*

4.4. Focos de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo tecnológico

El primer foco está ubicado en el distrito de Ilave, en la provincia de El Collao en el departamento de Puno, y está conformado por 578 pequeños productores agropecuarios donde el 100% afirma que el dinero que obtiene con la actividad

agropecuaria no es suficiente para cubrir sus gastos y las tierras de por lo menos el 60% tiene como máximo una hectárea. El 97% de productores de este foco tienen al Aymara como su idioma origen y únicamente el 7% de ellos cuenta con estudios superiores. No cuentan con cocina mejorada, ni con servicios higiénicos dentro de la vivienda, ni con computadora, tampoco tienen acceso a internet Su Índice de Adopción Tecnológico promedio es de 2.1 y las únicas tecnologías aplicadas son el uso de insecticidas y fertilizantes químicos.

El foco 2 está ubicado en el distrito de Huamarca, provincia de Huancabamba, departamento de Piura y está conformado por 575 pequeños productores agropecuarios. Todos los productores hablan castellano y 77% de los productores tienen estudios de primaria incompleta o completa. Un 88% indica que la actividad agropecuaria no les da ingresos suficientes para subsistir y el 45% de los productores afirma que la mayor parte de la producción es para auto consumo. Debido a ello el 40% de los productores se ven obligados a buscar trabajos fuera de la actividad agropecuaria para poder subsistir. El Índice de Adopción Tecnológico promedio de este foco es de 3.7 debido a la práctica de la dosificación y vacunación de animales y el uso de fertilizantes e insecticidas químicos.

En el distrito de Chota, provincia de Chota en Cajamarca encontramos el tercer foco de densidad con 372 pequeños productores agropecuarios de los cuales 46% tienen cocinas mejoradas en sus hogares, pero no tienen servicios higiénicos conectados a la red pública, dentro de la vivienda. Los productores agropecuarios de este foco cuentan primaria completa e incompleta (61% aproximadamente) y todos hablan castellano. Todos los productores de este foco afirman que los ingresos provenientes de la actividad agropecuaria son insuficientes para cubrir sus gastos y el 26.3% de ellos buscan trabajo fuera de la parcela para generar ingresos extras. Además, se debe considerar que los productores de este foco siempre siembran el mismo cultivo y su principal razón de producción es el autoconsumo. El Índice de Adopción Tecnológico promedio es de 4.67 y las tecnologías que adoptan son el uso de fungicidas, fertilizantes e insecticidas químicos para la agricultura y la vacunación y dosificación animal para las prácticas pecuarias.

El foco 4 está ubicado en el distrito y provincia Otuzco en el departamento de La Libertad con 192 pequeños productores agropecuarios y puntuación de 2.41 de Índice de Adopción Tecnológica teniendo en cuenta que todos los productores aseguran que sus ingresos por la actividad agropecuaria no les alcanzan para sus gastos. Además de fertilizantes, fungicidas e insecticidas hacen uso de la vacunación animal. El 50% de los productores cuentan solo con primaria incompleta. Solo un 10% cuenta con cocina mejorada y servicio higiénico dentro de la vivienda y el 72% de los productores tienen un régimen de tenencia propia con tierras de menos de 1 hectárea.

El foco 5 está ubicado en el distrito de Pampas, provincia de Pallasca en el departamento de Ancash con un Índice de Adopción Tecnológica de 5.26 con el 93% de ellos con ingresos insuficientes, con un 12.5% de ellos con lengua materna quechua, 47% de pequeños productores agropecuarios sin ningún nivel de instrucción o primaria incompleta. Este foco o grupo está conformado por 208 pequeños productores agropecuarios, quienes solo hacen uso de tecnologías en prácticas agrícolas como: fungicidas, herbicidas, insecticidas, fertilizantes químicos y el uso de tractor para trabajar las tierras. El 17% de su producción está destinada a la venta, el 50% para auto consumo y lo restante como alimento de sus animales. Sus condiciones de vida no son adecuadas, puesto que solo cuentan con servicios higiénicos y cocina mejorada el 8% de los pequeños productores, el 40% solo cocina mejorada o servicios higiénicos dentro de la vivienda y los demás productores no cuenta con ninguna de las dos.

El foco 6, ubicado en la provincia de Huaraz, departamento de Ancash cuenta con 501 pequeños productores agropecuarios divididos en 3 distritos distintos colindantes: Huaraz, Independencia y Pira. El 86% de los productores agropecuarios son propietarios de las tierras en las que trabajan y 94% son quechua hablantes, el 70% tiene como máximo nivel de instrucción primaria completa. El Índice de Adopción Tecnológico promedio es de 4.02 debido al uso de tractor, fertilizantes químicos, insecticidas químicos, fungicidas, vacunación y dosificación animal. Ningún productor en este grupo considera que los ingresos por la agricultura son suficientes para cubrir sus gastos y debido a esto el 47% tiene que trabajar fuera de la parcela para conseguir ingresos extras.

El séptimo foco está ubicado en el distrito de Huachis, en la provincia de Huari del departamento de Ancash con un índice de adopción tecnológica de 3.53 por actividades como dosificación, vacunación animal y fertilizantes químicos. El foco cuenta con 130 productores agropecuarios cuya actividad principal es la producción agropecuaria; sin embargo, el 100% indica que esta actividad no le es suficiente para generar ingresos que cubran sus gastos y el 78% trabajan fuera de la parcela, principalmente en construcción. Su lengua materna es el quechua y 10% de los productores agropecuarios que tienen estudios superiores. Solo el 16% de los productores cuentan con servicio higiénico dentro de la vivienda y cocina mejorada y el 26% no cuenta con ninguno.

El octavo y último foco está ubicado en el distrito de Andahuaylas, provincia de Andahuaylas en el departamento de Apurímac. Este foco cuenta con 362 productores agropecuarios y un Índice de Adopción Tecnológica promedio de 6.92 debido a la práctica de la vacunación animal, dosificación, fertilizantes químicos, insecticidas y fungicidas; además, se caracterizan por usar tractor para trabajar la tierra y utiliza control biológico en cultivos. La lengua materna de este foco es el quechua, con un 80% como máximo tienen nivel de instrucción primaria y el 71% son propietarios de las tierras. Solo el 10% cuenta con cocina mejorada o servicio higiénico dentro de la vivienda e indican que los ingresos de la actividad agropecuaria no son suficientes para cubrir sus gastos a pesar de que el 50% indica que la mayoría de su producción está destinada a la venta porque siembran siempre el mismo cultivo y el 20% de los productores trabaja fuera de la parcela en otras actividades.

4.5. Hipótesis

Los resultados nos permiten demostrar la veracidad de cada una de las hipótesis planteadas en la investigación:

- Primera Hipótesis: Se encontró que aproximadamente el 94% de los productores agropecuarios cuentan con un nivel de desarrollo

tecnológico muy bajo, con lo cual queda demostrada la hipótesis que la gran mayoría de los productores tienen bajos niveles de adopción tecnológica.

- Segunda Hipótesis: Como se ha demostrado en los resultados y con los estadísticos de análisis espacial, existen focos de alta densidad en la distribución espacial de los productores agropecuarios con un bajo nivel de desarrollo.

- La hipótesis 3 hace referencia a que no existe diferencia entre las características de los pequeños productores agropecuarios en focos de alta densidad. Las variables relacionadas con esta hipótesis fueron convertidas a numérica y normalizadas de la misma forma que en el análisis de Redes Neuronales de Kohonen. Las variables fueron

  Y_1: Índice de adopción tecnológica
  Y_2: Grado de instrucción
  Y_3: Régimen de tenencia propio
  Y_4: Servicios higiénicos dentro de la vivienda
  Y_5: Sabe leer y escribir
  Y_6: Si tiene ingresos insuficientes
  Y_7: Si realizó trabajos fuera de la parcela
  X: Foco al cual pertenece (1: Primer foco, 2: Segundo foco, …, 8: Octavo foco)

Para responder esta hipótesis se hizo uso de la prueba MANOVA.

$$H_0: \boldsymbol{\mu}_i = \boldsymbol{\mu}_j;\ i,j = 1,\dots,8$$
$$H_1: \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j;\ para\ algún\ i \neq j$$

Comentado [MEPA1]: Cuáles son las características que vas a comparara?????????

El estadístico de prueba obtenido es

$$F_{aproximado} = 52.99$$

Para un nivel de significancia del 5%, se rechaza la hipótesis nula, debido a que la prueba nos da un valor p de 0.000. Por tanto, con un nivel de significancia del 5% existe un par de focos cuyas características promedio difieren de los demás focos.

CONCLUSIONES

1’501,173 pequeños productores agropecuarios es decir el 98.1% presenta un nivel de desarrollo muy bajo de adopción tecnológica, 26027 es decir el 1.702% presenta un índice bajo, 2085 (0.136%) presenta un índice medio, 255 (0.017%) tiene un índice de adopción tecnológica alto y sólo 31 (0.002%) un índice muy alto de adopción tecnológica.

Se identificaron 8 distintos focos de pequeños productores ubicados principalmente en el norte y solo dos en la zona sur del país, uno identificado muy cerca al lago Titicaca. Estos focos están conformados por los pequeños productores agropecuarios con muy bajo nivel de índice de adopción tecnológico.

Las principales características de estos focos son:

El primer foco ubicado en el distrito de Ilave, provincia del Collao, departamento de Puno, conformado por 578 pequeños productores agropecuarios, con un índice de adopción tecnológica promedio de 2.1, el 97% de los productores son Aymara hablantes. El segundo foco está ubicado en el distrito de Huamarca, provincia de Huancabamba y departamento de Piura, con 575 pequeños productores y con un índice de adopción tecnológico promedio de 3.7. El tercer foco se encuentra en el distrito de Chota, provincia de Chota y departamento de Cajamarca, con 372 pequeños productores y con un índice de adopción tecnológico promedio 4.67. El cuarto foco se encuentra en el distrito de Otuzco, provincia de Otuzco y departamento de La Libertad, con 192 pequeños productores y con un índice de adopción tecnológico promedio 2.41,

todos los pequeños productores aseguran que sus ingresos no les alcanza para cubrir sus gastos. El quinto foco se encuentra en el distrito de Pampas, provincia de Pallasca y departamento de Ancash, con 208 pequeños productores y con un índice de adopción tecnológico promedio 5.26, el 12.5% habla quechua. El sexto foco se encuentra en el departamento de Ancash, provincia de Huaraz, con 501 pequeños productores en los distritos de Huaraz, Independencia y Pira con un índice de adopción tecnológico promedio 4.02, el 94% habla quechua; el 100% afirma que sus ingresos son suficiente para cubrir sus gastos. El séptimo foco se encuentra en el distrito de Huachis, provincia de Huari en Ancash con un índice de adopción tecnológico promedio de 3.53, con 103 pequeños productores, todos los productores agropecuarios indican que no tienen ingresos suficientes para cubrir sus gastos y todos son quechua hablantes. El octavo foco está ubicado en el distrito de Andahuaylas, provincia de Andahuaylas en el departamento de Apurímac y tienen un índice de adopción tecnológico de 6.92 con 392 productores agropecuarios que tienen como idioma materno al quechua y el 50% indica que no tiene ingresos suficientes para cubrir sus gastos.

Todos los focos tienen en común que hay ausencia de acceso a información por vías de internet, teléfono u otro medio; tampoco cuentan con el apoyo de asistencia técnica ni capacitaciones en temas agropecuarios. Esto ocasiona que no puedan tecnificarse debido a que no tienen acceso a información; por ello, no se encontró en los focos que tengan riego tecnificado en sus cultivos.

En cuanto a las hipótesis estadísticas planteadas, más del 75% de pequeños productores agropecuarios tienen un nivel de adopción tecnológico muy bajo, y menos del 30% cuentan con energía eléctrica. Por último, se obtuvo que existen diferencias en las características que presentan al menos un par de focos de pequeños productores agropecuarios.

RECOMENDACIONES

Considerando que la adopción tecnológica es el principal factor relacionado con la calidad de vida y los ingresos de los pequeños productores agropecuarios, es importante considerar que se debería hacer uso de políticas estatales para apoyar a la adopción de tecnologías más sofisticadas como riego tecnificado, uso de herramientas a base de energía eléctrica, puesto que las tecnologías adoptadas en estos focos de alta densidad son de fácil alcance, pero no son suficientes para incrementar significativamente sus ingresos.

Un segundo aspecto importante que debería tomarse en cuenta para mejorar las condiciones de vida de los productores agropecuarios mediante la adopción tecnológica es promover la asociación entre productores, ya que la asociación ayuda a que compartan información, disminuyan los costos en la producción, tengan acceso a nuevos mercados y la adopción de tecnologías.

REFERENCIAS BIBLIOGRÁFICAS

Andrade Limas, E., Faz Cano, A., & García Fernandez, G. &. (2001). Farmer classification: An approach to soil quality in Murcia. *European Soil Bureau*(7), 157-163.

Asfaw, S., Kassie, M. S., & Lipper, L. (2012). Poverty reduction effects of agricultural technology adoption: a micro-evidence from rural Tanzania. *Journal of Development Studies, 48*(9), 1288-1305.

Baranger, D. (2009). Análisis Factorial de Correspondencias. En *Construcción y Análisis de datos. Introducción al uso de técnicas cuantitativas en la investigación social.* (págs. 97-114). Posadas.

Becerra Guevara, R. L. (2014). *Modelo Neuronal de Demanda de Transporte en Redes Viales Urbana (Tesis de Maestría).* Universidad Nacional de Ingeniería, Lima.

Blayo, F. (1992). Kohonen Self-Organizing Maps: Is the Normalization Necessary?'. *Complex Systems, 6*(6), 105-123.

Cressie, N. (1993). Spatial Point Patterns. En *Statistics for Spatial Data* (págs. 577-650). Iowa: John Wiley & Sons.

Cuadras, C. M. (2014). MANOVA de un factor. En *Nuevos métodos de Análisis Multivariante.* España: CMC Editions.

De Cos Guerra, O. (2004). Valoración del método de densidades focales (kernel) para la identificación de los patrones espaciales de crecimiento de la población en españa. *GeoFocus: Revista Internacional de Ciencia y Tecnología de la Información Geográfica, 4*(16), 136-165.

De Janvry, A., & Sadoulet, E. (2009). Agricultural growth and poverty reduction: Additional evidence. *The World Bank Research Observer, 25*(1), 1-20.

Decancq, K., & Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews, 32*(1), 7-34.

Decreto Supremo Nº 002-2016. (18 de marzo de 2016). 1099-1114. Lima, Peru: Diario El Peruano.

Devendra, C., & Thomas, D. (2002). Crop–animal systems in Asia: importance of livestock and characterisation of agro-ecological zones. *Agricultural Systems, 71*(1-2), 5-15.

Diggle, P. J. (2014). Spatial Point Process. En *Statistical Analysis of Spatial and Spatial-Temporal Point Patterns* (págs. 55-79). Lancaster: CRC Press.

Eberhart, R. C., & Dobbins, R. W. (1990). Implementatios. En *Neural network PC tools: a practical guide* (págs. 49-58). Academic Press.

Escobal, J. (2014). *Análisis Espacial de la Adopción Tecnologías Agrarias en el Perú: Una mirada desde el Censo Nacional Agropecuario 2012.* Organización de las Naciones Unidas para la Alimentación y la Agricultura, Lima, Perú.

Escobal, J., & Armas, C. (2015). El uso de Encuestas y Censos Agropecuarios para desarrollar una tipología de pequeña y mediana agricultura familiar en el Perú. En GRADE, J. Escobal, R. Fort, & Z. Eduardo (Edits.), *Agricultura Peruana: Nuevas miradas desde el Censo Agropecuario* (págs. 15-61). Lima: GRADE.

Escobar Jaramillo, L. A. (2008). Indicadores Ambientales Sintéticos: Una aproximación conceptual desde la estadística multivariante. *Gestion y ambiente, 11*(1), 121-140.

Escobar, G., Henao, J., & Shenk, M. (1983). *Assessing adoption of technology among small farmers: classification and profile analysis.* Documento presentado en Third Annual System Research Symposium, Kansas State University, Manhattan.

Ezzrari, A., & Verme, P. (2013). A multiple correspondence analysis approach to the measurement of multidimensional poverty in Morocco 2001–2007. En V. Berenger, & F. & Bresson, *Poverty and social exclusion around the Mediterranean Sea* (págs. 181-209). New York: Springer.

Giraldo, R. (2002). Datos espaciales y Análisis exploratorio. En *Introducción a la Geoestadística: Teoría e Introducción* (págs. 8-16). Bogotá: Universidad de Colombia.

Graziano Da Silva, J., Gómez, S., & Castañeda, R. (2010). Boom agrícola y persistencia de la pobreza rural en América Latina: Algunas reflexiones. *Revista Austral de Ciencias Sociales*(18), 5-10.

Greenacre, M. (1991). Interpreting multiple correspondence analysis. *Aplied Stochastics Models in Business and Industry, 7*(2), 195-210.

Greenacre, M. (2017). Multiple Correspondence Analysis. En *Correspondence Analysis in Practice* (3 ed., págs. 137-144). Florida: CRC Press.

Greenacre, M. (2017). Theory of Correspondence Analysis. En *Correspondence Analysis in Practice* (págs. 241-254). Florida: CRC Press.

Greenacre, M. J. (1984). Multiple Correspondence Analysis. En *Theory and applications of correspondence analysis.* (págs. 126-162). Orlando: Academic Press Inc.

Härdle, W., & Simar, L. (2007). Correspondence Analysis. En *Applied multivariate statistical analysis* (págs. 341-372). Berlin: Springer.

Husson, F., Lê, S., & Pagès, J. (2017). Multiple Correspondence Analysis. En *Exploratory Multivariate Analysis by Example Using R* (págs. 131-172). Chapman and Hall/CRC.

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). Statistical analysis and modelling of spatial point patterns. En *Statistical analysis and modelling of spatial point patterns* (págs. 173-282). London: John Wiley & Sons.

INEI. (10 de 6 de 2018). *Instituto Nacional de Estadística e Informática del Perú.* Obtenido de https://www.inei.gob.pe/estadisticas/indice-tematico/economia/

Jaller R., S. (2010). *Vinculación de los pequeños productores a los sistemas de abastecimiento y distribución de los alimentos.* Proyecto de Investigación, Organización de las Naciones Unidas para la Agricultura y la Alimentación FAO, Bogotá.

Kassambara, A. (2017). Cluster Validation. En *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning* (1 ed., págs. 117-160). STHDA.

Kassie, M., Shiferaw, B., & Muricho, G. (2011). Agricultural technology, crop income, and poverty alleviation in Uganda. *World Development, 39*(10), 1784 - 1795.

Kim, G. (2015). La Difusión Tecnológica Agropecuaria en el Perú: causas y tendencias en el período intercensal 1994-2012. *Economía y Sociedad*(85), 38-39.

Koch, I. (2014). Clsuter Analysis. En *Analysis Multivariate and High Dimensional Data* (págs. 183-220). New York: Cambridge University Press.

Kohonen, T. (1990). The self-organizing map. *Proceeding of the IEEE, 78*(9), 1464-1480.

Kohonen, T. (2001). The Basic SOM. En T. S. Huang, T. Kohonen, & M. Schroeder (Edits.), *Self-Organizing Maps* (3 ed., págs. 105-176). Espoo, Finlandia: Springer.

Le Roux, B., & Rouanet, H. (2004). Overview of Geometric Data Analysis. En *Geometric data analysis: from correspondence analysis to structured data analysis* (1 ed., págs. 1-20). Dordrecht: Springer Science & Business Media.

Le Roux, B., & Rouanet, H. (2010). The Method of Multiple Correspondence Analysis. En *Multiple correspondence analysis* (págs. 34-67). California: Sage Publications.

Liu, E. M. (2013). Time to change what to sow: Risk preferences and technology adoption decisions of cotton farmers in China. *Review of Economics and Statistics, 95*(4), 1386-1403.

Lowder, S., Skoet, J., & Singh, S. (2014). What do we really know about the number and distribution of farms and family farms in the world. *Background paper for the State of Food and Agriculture, 8*, 1-45.

Macías, A. M. (2016). Introducción. Los pequeños productores agrícolas en México. *Carta económica regional, 25*(111-112), 7-18.

Maletta, H., & De la torre Ugarte, D. (2016). Una nueva forma de medir y analizar la pequeña agricultura familiar en el país. *Economía y Sociedad*(88), 47-42.

Martin-del-Brio, B., & Serrano-Cinca, C. (1993). Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing & Applications, 1*(3), 193-206.

Mendola, M. (2007). Agricultural technology adoption and poverty reduction: A propensity-score matching analysis for rural Bangladesh. *Food Policy, 32*(3), 372-393.

Meynard Vivar, F. (2015). Perú: A la espera de políticas específicas para la agricultura familiar. En E. Sabourin, M. Samper, & O. Sotomayor, *Políticas públicas y agriculturas familiares en América Latina y el Caribe: Nuevas Perspectivas* (págs. 294-311). San José: IICA.

MINAGRI. (2016). *Política Nacional Agraria.* Lima.

Moreno J., A. (1991). Modelización cartográfica de densidades mediante estimadores de Kernel. *Treballs de la Societat Catalana de Geografia, 30*, 155-170.

Mwangi, M., & Kariuki, S. (2015). Factors determining adoption of new agricultural technology by smallholder farmers in developing countries. *Journal of Economics and sustainable development, 6*(5), 208 - 216.

Natschläger, T., Ruf, B., & Schmitt, M. (2013). Unsupervised Learning and Self-Organization in Networks of Spiking Neurons. En U. Seiffert, & L. C. Jain, *Self-organizing neural networks: Recent advances and applications* (págs. 45-74). Berlin: Physica-Verlag Heidelberg.

Oyana, T., & Margai, F. (2016). Engaging in Point Pattern Analysis. En *Spatial Analysis: Statistics, Visualization and Computational Methods* (págs. 151-181). CRC Press.

Seaman, D. E., & Powell, R. A. (1996). An evaluation of the accuracy of Kernel Density Estimators for home range analysis. *Ecology, 77*(7), 2075-2085.

Waller, L. A., & Gotway, C. A. (2004). Analysis of Spatial Point Patterns. En *Applied Spatial Statistics for Public Health Data* (págs. 118-154). New Jersey: John Wiley & Sons.

Yunlong, C., & Smit, B. (1994). Sustainability in agriculture: a general review. *Agriculture, Ecosystems & Environment*, 299-307.

ANEXOS

Anexo 1. Glosario de términos

- Análisis multivariante. Método estadístico por el cual se realiza el análisis simultáneo de varias variables medidas en un mismo sujeto.

- Aprendizaje supervisado. Es un aprendizaje en el cual hay un agente exterior que influenciará en la respuesta y pesos hasta encontrar el resultado deseado.

- Aprendizaje no supervisado. Este tipo de aprendizaje se caracteriza por usar medidas de similitud entre los distintos inputs con la finalidad de encontrar patrones.

- Clúster. Hace referencia a agrupación o conjunto de individuos u observaciones de similares características.

- Datos georreferenciados. Los datos georreferenciados son datos que poseen información de su localización geográfica en algún sistema de coordenadas el cual es único y bien definido.

- Estimador. Un estadístico, en función de la muestra, que permite estimar el valor aproximado de un parámetro desconocido de la población.

- Kernel. Es una función de ponderación utilizada en técnicas de estimación no paramétricas de funciones de densidad de variables aleatorias.

- medida de disimilitud. Son medidas que miden la distancia entre los objetos y mientras más distantes sean las observaciones más diferentes son entre sí.

- Método de Montecarlo. Es un método estadístico numérico para realizar aproximaciones de expresiones matemáticas difíciles de calcular y con altos costos computacionales.

- Normalización. Proceso por el cual se centran los datos alrededor de la media y se reduce la variabilidad a uno.

- Proceso estocástico. Un proceso estocástico es un conjunto o familia de variables aleatorias que varían en el tiempo, espacio o en ambos; es decir, tomando como ejemplo el tiempo; en distintos momentos del tiempo se puede tener distintas variables aleatorias.

- Valor propio. El escalar asociado al vector propio que ayuda a cambiar la dirección de este.

- Vector propio. Es el vector no nulo asociado a un operador lineal da lugar a un múltiplo escalar de sí mismo con lo cual no cambia su dirección.

Anexo 2. Matriz de consistencia

"MINERÍA DE DATOS ESPACIAL: UNA APLICACIÓN PARA IDENTIFICAR LOS FOCOS DE LOS PEQUEÑOS PRODUCTORES AGROPECUARIOS PERUANOS CON BAJO NIVEL DE DESARROLLO A NIVEL NACIONAL"

PROBLEMA GENERAL	OBJETIVO GENERAL
¿Cuáles son los focos de pequeños productores agropecuarios peruanos, sus patrones y su categorización según el nivel de adopción tecnológica?	Identificar los focos de pequeños productores agropecuarios peruanos con bajo nivel de desarrollo a nivel nacional y sus patrones con el método de densidades focales de Kernel.
PROBLEMAS ESPECÍFICOS	**OBJETIVOS ESPECÍFICOS**
¿Cuáles son los focos de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo a nivel nacional?	Identificar los focos de los pequeños productores agropecuarios peruanos con bajo nivel de desarrollo a nivel nacional mediante el método de densidades focales de Kernel.
¿Cómo se puede medir en Índice de Adopción tecnológica de los productores agropecuarios?	Calcular el índice de adopción tecnológica de los productores agropecuarios peruanos a nivel nacional mediante el Análisis de Correspondencias Múltiples
¿Cuáles son las características de los productores agropecuarios según su nivel de adopción tecnológica?	Describir las características de los productores agropecuarios según su nivel de adopción tecnológica en los focos de alta densidad calculados mediante el método de densidades focales de Kernel.

Anexo 3. Matriz de operacionalización de variables.

Variable	Definición	Tipo de variable	Medición
Uso de semillas artificiales	El agricultor usa semillas artificiales y/o plantones certificados	Nominal	0: No 1: Si
Uso de fertilizantes químicos	El agricultor usa fertilizantes químicos en el cultivo	Nominal	0: No 1: Si
Uso de Insecticidas químicos	El agricultor usa insectisidas químicos durante el cultivo	Nominal	0: No 1: Si
Uso de fungicidas	El agricultor hace uso de fungicidas para el cultivo	Nominal	0: No 1: Si
Uso de herbicidas	El agricultor hace uso de herbicidas para el cultivo	Nominal	0: No 1: Si
Uso de energía eléctrica	El agricultor hace uso de energía eléctrica para trabajar la tierra	Nominal	0: No 1: Si
Uso de Tractor	El agricultor hace uso de tractor para trabajar la tierra	Nominal	0: No 1: Si
Uso de fumigadores	El agricultor hace uso de fumigadores eléctricos durante el cultivo	Nominal	0: No 1: Si
Uso de molino eléctrico	El agricultor hace uso de molino eléctrico para granos	Nominal	0: No 1: Si
Uso de trilladora	El agricultor hace uso de energía eléctrica para trabajar la tierra	Nominal	0: No 1: Si
Uso de motor bomba	El agricultor hace uso de motor bomba para extraer agua	Nominal	0: No 1: Si

Uso de camión	El agricultor hace uso de camión durante el cultivo	Nominal	0: No 1: Si
Riego por aspersión	El agricultor tiene en alguna de sus parcelas el sistema de riego por aspersión	Nominal	0: No 1: Si
Riego por goteo	El agricultor tiene en alguna de sus parcelas el sistema de riego por goteo	Nominal	0: No 1: Si
Riego por exudación	El agricultor tiene en alguna de sus parcelas el sistema de riego por exudación	Nominal	0: No 1: Si
Índice de Adopción Tecnológica	Indicador de adopción tecnológica con valores entre 0 y 100 calculado a partir de el análisis de correspondencias múltiples.	Ordinal	1: Muy bajo 2: Bajo 3: Medio 4: Alto 5: Muy Alto
Sabe leer y escribir	El agricultor sabe leer y escribir	Nominal	0: No 1: Si
Uso de cocina mejorada	El agricultor tiene cocina mejorada dentro o fuera de la vivienda	Nominal	0: No 1: Si
Tipo de servicio higiénico	El agricultor tiene servicio higiénico conectada a una red de agua	Nominal	0: No 1: Si
Cuenta con computadora	El agricultor tiene computadora o hace uso de una	Nominal	0: No 1: Si

Nivel educativo	Nivel educativo del agricultor	Ordinal	1: Sin Nivel 2: Primaria incompleta 3: Primaria completa 4: Secundaria incompleta 5: Secundaria completa 6: Estudios superiores
Tenencia de las parcelas	La parcela en la que trabaja le pertenece al agricultor	Nominal	0: No 1: Si

Anexo 4. Análisis geométrico de datos

> *"Es un hecho que la gran mayoría de métodos estadísticos multivariados puede ser enfocadas desde el punto de vista del Análisis Geométrico de Datos"* (Le Roux & Rouanet, 2010)

Este concepto engloba y da un enfoque distinto al análisis multivariado, fue enunciado por primera vez por Karl Pearson en 1901 como "Goemetric Display of Data", pero su popularidad no se dio hasta los años ochenta en los cuales se presentaron las primeras publicaciones en inglés, debido a que su desarrollo original fue en Francia. Entre las primeras publicaciones que se realizaron en inglés se tiene la de Greenacre (1984) que desarrollo el análisis de correspondencias, Lebart & al (1984) con "Análisis Descriptivo Multivariado". Desde la fecha se han popularizado aún más estos métodos y desarrollado más investigaciones que en la actualidad brindan una amplia gama de técnicas derivadas del análisis de correspondencias múltiples y Escalamiento Dimensional.

El análisis geométrico de datos se enfoca en la representación de datos multivariados mediante *nubes de puntos* (similar a los gráficos de dispersión) formadas a partir de nuevas componentes (ver figura 1) y las interpretaciones se basan en la información que brinden las nubes de puntos. Este enfoque es muy utilizado en las investigaciones sociales, permite realizar una interpretación adecuada de las relaciones existentes entre las distintas categorías de las variables, permitiendo realizar inferencias y conclusiones adecuadas. (Greenacre M. J., 1984)

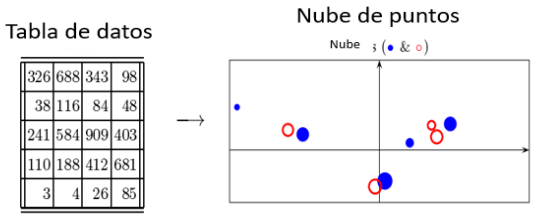

Figura 20: Enfoque del Análisis Geométrico de Datos
Fuente: *Adaptado de Le Roux & Rouanet (2004)*

El enfoque principal del análisis multivariante se basa en tres pilares: (Le Roux & Rouanet, 2004)

Cuantitativo: Los resultados numéricos son los más importantes tanto durante el procedimiento como a la hora de tomar una decisión.

Enfoque Matricial: Los procedimientos de cálculo y las demostraciones son realizados exclusivamente mediante operaciones matriciales.

Orientado a muestras: Las bases de datos son tratadas como si fueran una muestra y el análisis se realiza como si fueran muestras: se realizan ajustes, se estiman modelos y se hacen prueba de hipótesis.

Por otro lado, el enfoque del análisis geométrico de datos se basa en lo que se conoce como "*Las tres ideas claves del AGD"* (Three Key Ideas of GDA por sus siglas en inglés) que se presentan a continuación: (Le Roux & Rouanet, 2004)

Modelamiento Geométrico de Datos: Consiste en la representación de los datos mediante nubes de puntos (ver figurar 1), que es la principal característica del AGD. Desde la construcción de nubes de puntos para tablas de doble entrada

hasta tablas con mayores dimensiones (como la matriz de Burt). Una vez construida la nube de puntos se realiza el análisis e interpretación de esta, que tiene como principal característica la dimensionalidad (porque se usa un número determinado de dimensiones). Sin embargo, el objetivo de realizar el análisis geométrico no es excluir el análisis numérico, puesto que dentro del mismo se incluyen cálculos como: autovalores, coordenadas principales, etc. Teniendo en cuenta que para las interpretar resultados de un análisis geométrico de datos se hace uso de cálculos como las *contribuciones*.

"Algunas veces el objetivo de una investigación es la construcción de índices numéricos; para ello, las principales variables del AGD, con sus propiedades de escalas óptimas, proporcionan mejores resultados que calcular únicamente índices a priori". (Citado en Le Roux & Rouanet, 2004: 7).

Enfoque Formal y Algebra Lineal Abstracta: Existe una relación estrecha entre las dos ideas principales: La geometría y formalidad son esenciales en el análisis geométrico de datos que a su vez es la base de la estadística multivariante. Este punto es crucial en las dos fases del análisis geométrico de datos: (i) Construcción de nubes y especificación métrica y (ii) Descomposición espectral. En otras palabras, el análisis geométrico de datos sigue la formalidad teórico-matemática de la estadística que permite fácilmente realizar pruebas de los procedimientos.

Filosofía del Análisis de Datos: Se refiere a la clave del análisis geométrico de datos que consiste en la estadística descriptiva como base del análisis. Hay que tener en cuenta que la estadística no se limita solo a la inferencia, sino que también podemos obtener mucha información a partir de las nubes de puntos calculadas con el análisis geométrico de datos. Es importante aclarar que el hecho de que la base fundamental del análisis geométrico de datos sea la estadística descriptiva, esta no omite la estadística inferencial porque se habla de prioridad, pero no de exclusividad.

El Análisis Geométrico de Datos también presenta tres paradigmas asociado con tres tipos de tablas de datos. (Le Roux & Rouanet, 2004)

- Análisis de Correspondencias**:** Basado en la medida de distancias Chi Cuadrado y se presenta a partir de tablas de contingencia o de doble entrada.
- Análisis de Componentes Principales: Proviene de tablas de individuos y variables numéricas (aunque ahora hay métodos derivados que incluyen también variables categóricas) y hay distintos métodos para calcular distancias.
- Análisis de Correspondencias Múltiples: Proviene de tablas de individuos y variables categóricas (a partir de 3 variables categóricas).

La esencia metodológica del análisis geométrico de datos ha mostrado un fuerte cambio en las ciencias sociales con respecto a la estadística clásica. Mientras que en la estadística clásica la prioridad son los resultados numéricos y sus valores *p* asociados, para el análisis geométrico de datos, el foco se centra en las nubes de puntos. Este contraste ayuda a que no solo se realice un análisis social de las variables, si no a crear un espacio social en el cual las variables permitan desprender ideas o diagnósticos para un posterior análisis inferencial. Para que el análisis geométrico de datos cumpla con sus objetivos se debe contar con una base de datos relevante y para ello y el análisis geométrico de datos tiene dos condiciones que deben ser estrictamente cumplidas: (i) Homogeneidad: Todas las variables de la tabla deben tener la misma naturaleza. (ii) Exhaustividad: Los datos debe contener exhaustiva o al menos suficiente información de la naturaleza del problema de investigación. (Le Roux & Rouanet, 2004)

El Análisis Geométrico de datos ha traído controversias a la hora de su uso dentro del análisis de datos estructurados como cuando se hace el análisis de una encuesta social en la cual no se tienen únicamente las características de los individuos, si no otras variables como edad, años de estudio, etc. Entonces

surge la siguiente pregunta: ¿Es adecuado el uso del análisis geométrico de datos? Pues la respuesta es sí y la metodología a seguir será similar a la que se usa cuando se realiza un análisis de regresión o un análisis de varianza (ver figura 2), puesto que se puede trabajar los resultados del análisis geométrico de datos y realizar comparaciones o buscar relaciones con las variables que no entraron en el análisis. (Le Roux & Rouanet, 2004)



Figura 21: Metodología del análisis geométrico de datos para análisis de datos estructurados
Fuente: *Adaptado de Le Roux & Rouanet (2004)*

Otra duda, muy común es que se cree que, el análisis geométrico de datos excluye el análisis inferencial, pero esto no es cierto, si no que se debe seguir una metodología adecuada, así como cuando se hace un análisis en la estadística convencional: primero se realiza la estadística descriptiva y con ello se plantean hipótesis que son probadas con la estadística inferencial. Por tanto, si es posible aplicar la estadística inferencial en un análisis geométrico de datos, pero este debe ser tratado con mucho cuidado a la hora de ser realizado, puesto que su objetivo es principalmente responder preguntas, ya que no toda investigación tiene que tener dentro su análisis el uso de la estadística inferencial. (Le Roux & Rouanet, 2004)



Figura 22: Metodología del uso de la estadística inferencial en el AGD.
Fuente: *Adaptado de Le Roux & Rouanet (2004)*

Anexo 5. Análisis de Correspondencias Simples

El análisis de correspondencias simples permite analizar la asociación entre filas y columnas de una tabla de contingencia de dos variables cualitativas. La idea principal es calcular índices simples que nos permitan analizar la relación entre las categorías de las filas y columnas de la tabla de contingencia, mostrando cuales categorías de las columnas tienen más peso sobre las categorías de las filas y viceversa. El análisis de correspondencias múltiples es muy similar al análisis de componentes principales con la diferencia de que este caso se trabaja con variables categóricas y para calcular las distancias utilizamos la estadística chi-cuadrado en lugar de la distancia Euclidiana. (Härdle & Simar, 2007).

Las definiciones que a continuación se mostrarán han sido extraídas de la literatura desarrollada por Greenacre en Correspondence Analysis in Practice (2017) y Cuadras (2014).

Supongamos que tenemos dos variables cualitativas X e Y, sea I el número de categorías de la variable X y el i-ésimo término será representado por X_i. Se denota como J el número de categorías de la variable B y el j-ésimo término es representado por B_j. La tabla de contingencia nos da un resumen de ambas variables.

Tabla 4: Tabla de contingencia de las variables X e Y.

Fila\Columna	Y_1	Y_2	…	Y_j	…	Y_J
X_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}
X_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2J}
⋮	⋮	⋮		⋮		⋮
X_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iJ}
⋮	⋮	⋮		⋮		⋮
X_I	n_{I1}	n_{I2}	…	n_{Ij}	…	n_{IJ}

Fuente: Elaboración Propia

Se define como N la matriz $I \times J$ con las frecuencias de la tabla de contingencia. Se define P como la matriz de correspondencias

$$\boldsymbol{P} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{N}$$

Sea r el vector dimensión $I \times 1$ con los totales marginales de las filas de la matriz P y el vector c de dimensión $J \times 1$ con los totales marginales de las columnas de la matriz P. Se calculan de la siguiente manera:

$$\boldsymbol{r} = \boldsymbol{P}\mathbf{1}_{\boldsymbol{J}}$$

$$\boldsymbol{c} = \boldsymbol{P}^{\boldsymbol{T}}\mathbf{1}_{\boldsymbol{I}}$$

donde $\mathbf{1}$ es el vector de unos con la dimensión apropiada para realizar la operación con la matriz P. Se definen las matrices diagonales $\boldsymbol{D_r}$ y $\boldsymbol{D_c}$ como las matrices en las cuales sus diagonales son los vectores r y c respectivamente.

El objetivo principal del análisis de correspondencias es reducir la dimensionalidad para poder observar mejor las categorías de las variables para ello es necesario calcular las coordenadas principales de filas y columnas y la inercia total para conocer la contribución de cada dimensión.

Para encontrar las coordenadas principales de las filas y de las columnas se calcula la matriz rectangular S de orden $I \times J$

$$S = D_r^{-\frac{1}{2}}(P - rc')D_c^{-\frac{1}{2}} \quad (1)$$

aplicando la descomposición en valores singulares en (1) obtenemos

$$S = UD_\lambda V^t \quad (2)$$

donde D_λ es la matriz diagonal de valores singulares ordenados de forma decrecientes y $U^TU = V^TV = I$. Las coordenadas principales de las filas (también llamadas como coordenadas de los individuos u observaciones) se enuncian de la siguiente manera:

$$F = D_r^{-\frac{1}{2}}UD_\lambda$$

Las coordenadas principales de las columnas (también llamadas como coordenadas de las variables) se calculan de la siguiente manera:

$$G = D_C^{-\frac{1}{2}} U D_\lambda$$

Una vez que se tienen las coordenadas principales para filas y columnas en cada una de las dimensiones, es importante hacer una evaluación de la contribución de cada dimensión sobre la inercia total.

La inercia total se define de la siguiente manera:

$$\phi^2 = \sum_{k=1}^{K} \lambda_k^2, \text{ donde } K = \min\{I, J\}$$

La inercia total también coincide con el estadístico Chi cuadrado dividida entre el número total de datos.

$$\phi^2 = \sum_{k=1}^{K} \frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{f_{i.} f_{.j}}$$

La contribución de la i-ésima dimensión se calcula dividiendo la inercia total entre el i-ésimo autovalor.

Anexo 6. Métodos de agrupamiento

Clúster Jerárquico

Dentro del análisis jerárquico se pueden ver dos enfoques complementarios:

- *Agrupamiento aglomerativo*. Empieza con un grupo de una observación y se van agrupando los datos en grupos más grandes.
- *Agrupamiento divisivo*. Empieza con un único clúster y se divide en clústeres más pequeños hasta encontrar la distancia suficiente entre los grupos.

Clúster No jerárquico

El análisis no jerárquico es un método por el cual se busca agrupar a los individuos en base a un número determinado de clústeres establecido a priori. La formación de estos grupos se hace a partir de puntos semilla (centro de los clústeres definidos a priori) y se van actualizando los centros de acuerdo con el ingreso de cada observación a un grupo medido por la distancia que este tiene con respecto al centro definido. Dentro de los métodos usados en el análisis de clúster no jerárquico, tenemos k-medias, PAM y CLARA. En la presente investigación se aplicará el clúster no jerárquico, puesto que se determinará a priori el número de clústeres a usar.

Anexo 7. MANOVA (Análisis multivariante de la varianza)

Un modelo multivariado para comparar las medias de g poblaciones, donde a cada observación se miden p características (variables dependientes, métricas en escala de intervalo o razón), se representa como:

$$\boldsymbol{y_{irj} = u_{ij} + e_{irj} \quad i = 1, 2, \ldots, g \quad j = 1, 2, \ldots, p \quad r = 1, 2 \ldots, n_i}$$

Matricialmente dicho modelo puede representarse como:

$$\boldsymbol{Y_{ir} = U_i + \varepsilon_{ir}, \qquad tal\ que\ \epsilon \sim N_p(0, \Sigma)}$$

Se busca verificar si los vectores de las medias de todos los grupos son iguales, es decir provienen de la misma población:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_g$

$H_1: \mu_j \neq \mu_k, \quad para\ algún\ j \neq k$

Estimación del Modelo Manova

Utilizaremos el principio de razón de verosimilitud. Por el criterio de razón de verosimilitud se tiene que

$$\Lambda = \frac{|W|}{|T|}$$

Donde:

$$W = \sum_{i=1}^{g} \sum_{r=1}^{n_i} (Y_{ir} - \bar{Y}_i)(Y_{ir} - \bar{Y}_i)$$

$$T = \sum_{i=1}^{g}\sum_{r=1}^{n_i}(Y_{ir} - \bar{Y})(Y_{ir} - \bar{Y})$$

W: Matriz de suma de cuadrados y producto cruzado de errores, dentro de grupos.

T: Suma de cuadrados totales y productos cruzados.

Con T = W + B, donde:

$$B = \sum_{i=1}^{g} n_i(\bar{Y}_i - \bar{Y})(\bar{Y}_i - \bar{Y})$$

Es la suma de cuadrados de los productos cruzados entre grupos. Bajo la hipótesis nula, se tiene que:

$$W \sim W_p(\Sigma, n-k)$$

$$B \sim W_p(\Sigma, k-1)$$

Además, W y B son independientes. La tabla MANOVA queda como sigue:

MANOVA de un factor

	g.l.	Matriz Wishart	Lambda de Wilks
Entre grupos	k-1	**B**	$\Lambda = \lvert W\rvert/\lvert T\rvert \sim \Lambda(\mathrm{p}, \mathrm{n}-\mathrm{k}, \mathrm{k}-1)$
Dentro grupos	n-k	**W**	
Total	n-1	**T**	

El criterio de decisión es que se rechaza H_0 si $\Lambda < \Lambda_\alpha$ donde $P[\Lambda(p, n-k, k-1) < \Lambda_\alpha] = \alpha$

Anexo 8. Cálculos de resultados mostrados en subsección 4.5

Para contrastar las hipótesis se seleccionó una muestra probabilística que contiene a 380 productores agropecuarios con bajos niveles de desarrollo tecnológico.

$$n = \frac{p \times (1-p) \times Z^2_{1-\alpha/2}}{e^2} = \frac{0.09 \times 0.81 \times 1.96^2}{0.01^2} = 380$$

donde:

p: proporción de característica principal de interés igual a 0.99

Z: Nivel de confianza del 95%.

e: Margen de error de 0.01

- Cálculos de la primera hipótesis

1. Hipótesis

H_0: $p \leq 0.75$

H_1: $p > 0.75$

2. Estadístico para la prueba

$$Z_c = \frac{0.994 - 0.75}{\sqrt{\frac{0.75(0.25)}{380}}} = 10.98$$

3. Decisión

El valor crítico es $Z_{1-\alpha} = Z_{1-0.05} = 1.644$ y se rechaza la hipótesis nula si el valor Z_c es mayor a 1.644. Entonces, como el valor de $Z_c = 10.98$ es mayor a 1.644, se rechaza la hipótesis nula.

- Cálculos de la segunda hipótesis:

1. Hipótesis

H_0: $p \geq 0.30$

H_1: $p < 0.30$

Comentado [MEPA2]: Indicar los valore de *p*, *q* y *e*. y colocar el tamaño de la muestra

Comentado [MEPA3]: Indicar los valore de *p*, *q* y *e*. y colocar el tamaño de la muestra

Comentado [MEPA4]: Realizar los cálculos y tomar la decisión

2. Estadístico para la prueba

$$Z_c = \frac{0.013-0.30}{\sqrt{\frac{0.30(0.7)}{380}}} = -12.2$$

Comentado [MEPA5]: Realizar los cálculos y tomar la decisión

3. Decisión

El valor crítico es $Z_\alpha = Z_{0.05} = -1.644$ y se rechaza la hipótesis nula si el valor Z_c es menor a 1.644. Entonces, como el valor de $Z_c = -12.2$ es menor a -1.644, se rechaza la hipótesis nula.

Anexo 9. Verificación de veracidad del índice.

Con la finalidad de verificar la veracidad del indicador, se llevo a realizar una comprobación mediante la cual se mide la correlación existente entre el total de las variables y el indicador para los individuos analizados.

Tabla 5: Índice de correlación de Pearson entre el total de las variables y el Indicador de adopción tecnológica

Correlación de Pearson	Sig.
0.843	0.031

Fuente: Elaboración: Propia

La tabla 10 muestra que existe una relación directa entre la puntuación total de las variables y el indicador propuesto; por tanto, podemos decir que el indicador está realizando una medición adecuada de la adopción tecnológica.

Anexo 10. Codificación en R

```
# Librerias ------------------------------------------------------------------
library(data.table)
library(foreign)
library(tidyverse)
library(FactoMineR)
library(factoextra)
library(corrplot)
library(kohonen)
library(NbClust)
library(DescTools)
library(ggmap)
library(maps)
library(lubridate)
library(ggplot2)
library(ggrepel)
library(data.table)
library(tidyverse)
library(spatstat)
library(mapview)
library(plotly)
library(ggthemes)
library(viridis)
library(rgl)
library(raster)
library(readxl)

#Lectura de datos ------------------------------------------------------------
df<-read.dta("DatosTesis.dta",convert.factors = F)


# Filtro y Selección de variables ---------------------------------------------
df<- df %>% filter(supagc <= 5)

names(df)<-c("Departamento", "Provincia", "Distrito", "SEA", "NumCedula", "VacAnima","DosfAni
mal", "DietaBaln", "InsemArt", "Capacitacion", "AsisTec", "SemArt","FertQuim", "InsectQuim",
"Herbicida", "Fungicida", "EnerElect", "Tractor","Fumigadora", "MolinoGrano", "Trilladora", "
BombaPozo", "MotorAgua", "Camion","Raspersion", "Rgoteo", "Rexudacion", "SupAgc", "SupNoAgc",
"SupOtros","LongSea", "LatSea",  "LeeyEsc", "CondJur", "SSHH", "CocinaMejorada","Computadora"
, "Internet", "Tenencia", "Ubigeo", "Resultado", "RegionNat","PisoAlti", "RegHidro", "CuencaH
idro", "TamHogar", "Sexo", "Edad", "NivelEduc", "Idioma", "prov", "DestinoProd", "ControlBio"
, "UsoAnimAgr", "GanadoVac", "ExisteAnim", "InfoTelef", "InfoRadio", "InfoTv", "InfoInter", "
InfoPublic", "SolCred", "ObtuvoCred", "TrabPer", "TrabEven", "Asociado", "RazonProd", "IngSuf
ic", "DejoTrab", "ActPrincipal", "HorasCap", "DemoraCapital")


# Eliminacion y Tratamiento de NA's -------------------------------------------
df <- df %>% filter(SemArt != "NA")
df <- df %>% select(-BombaPozo)
df <- df %>% filter(NivelEduc != "NA")
df <- df %>% filter(ControlBio!= "NA")

df$DestinoProd[is.na(df$DestinoProd)] <- 5

df$VacAnimal[is.na(df$VacAnimal)] <- 0
df$VacAnimal[df$VacAnimal==2] <- -1

df$DosfAnimal[is.na(df$DosfAnimal)] <- 0
df$DosfAnimal[df$DosfAnimal==2] <- -1

df$DietaBaln[is.na(df$DietaBaln)] <- 0
df$DietaBaln[df$DietaBaln==2] <- -1
```

```r
df$InsemArt[is.na(df$InsemArt)] <- 0
df$InsemArt[df$InsemArt==2] <- -1

for (i in 10:23) {
  df[,i][df[,i] == 2] <- 0
}

df <- df %>% filter(IngSufic != "NA")
df <- df %>% filter(DejoTrab != "NA")

df$ObtuvoCred[is.na(df$ObtuvoCred)] <- 3
df$TrabPer[is.na(df$TrabPer)] <- 9999
df$TrabEven[is.na(df$TrabEven)] <- 9999
df$RazonProd[is.na(df$RazonProd)] <- 9999
df$ActPrincipal[is.na(df$ActPrincipal)] <- 9999
df$HorasCap[is.na(df$HorasCap)] <- 9999
df$DemoraCapital[is.na(df$DemoraCapital)] <- 9999

df <- df %>% filter(VacAnimal != 0)

df$FertQuim[df$FertQuim == 0] <- 1
df$FertQuim[df$FertQuim == 3] <- 0

df$SSHH<-if_else(df$SSHH==1 | df$SSHH==2,1,0)
df$LeeyEsc<-if_else(df$LeeyEsc==1,1,0)
df$CocinaMejorada<-if_else(df$CocinaMejorada==1,1,0)
df$Computadora<-if_else(df$Computadora==1,1,0)
df$Internet<-if_else(df$Internet==1,1,0)
df$IngSufic[df$IngSufic==2] <- 0
df$DejoTrab[df$DejoTrab==2] <- 0

# recodificación del nivel educativo.
# 1: sin nivel e inicial
# 2: Primaria incompleta
# 3: Primaria completa
# 4: Secundaria incompleta
# 5: Secundaria completa
# 6: Supeior
df$NivelEduc[df$NivelEduc==10] <- 6


# Eliminación de datos duplicados-----------------------------------------------
df <- df[!duplicated(df$NumCedula),]

# Preparación de datos para creación de indicador -----------------------------
Indicador <- df[,c(5:26, 52)]
table(Indicador$VacAnimal)
Indicador$VacAnimal <- if_else(Indicador$VacAnimal == -1, "Vac_no", "Vac_si")
Indicador$VacAnimal <- as.factor(Indicador$VacAnimal)

# Creación del indicador ------------------------------------------------------
names(Indicador[,2:23])
res.mca<-MCA(Indicador[,2:23],ncp = 20,graph = TRUE)
eig.val <- round(get_eigenvalue(res.mca),3)
fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0, 20), ncp = 20)

fviz_mca_var(res.mca, choice = "mca.cor",
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())
ggsave("mca.png",width = 7,height = 6)

fviz_mca_var(res.mca, choice = "var.cat",
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())


var <- get_mca_var(res.mca)
cors <- var$cos2[seq(2,44,2),]
```

```r
rownames(cors) <- c("Vacunas","Dosificaciones","Al. Balanceado","Insem. Artificial",
                    "Capacitaciones","Asist. Tec.","Semilla Art.","Fertilizante Quim.",
                    "Insecticida Quim.", "Herbicida", "Fungicida","Ener. Electrica",
                    "Tractor","Fumigadora","Molino","Trilladora","Motor de agua",
                    "Camion","R. Aspersion","R. Goteo","R Exudacion", "Control Bio.")

corrplot(cors,
         is.corr=FALSE,
         method = "square")


# Calculo de puntuaciones del IAT ------------------------------------------------

pesos <- as.matrix(res.mca$eig[c(1:7,9,12:13,15:17), 1])
ind<-res.mca$ind$coord[,c(1:7,9,12:13,15:17)]

iat <- (ind %*% sqrt(pesos))/sum(sqrt(pesos))
iat <- (iat - min(iat))/(max(iat) - min(iat))
iat <- as.data.frame(iat)
names(iat) <- c("IAT")
iat$IAT <- iat$IAT*100

ggplot(iat,aes(x=IAT))+geom_density(adjust = 1.4)+
  xlab("Indice de Adopción de Tecnología Agrícola") +
  ylab("N° Productores Agropecuarios") +
  theme(axis.title = element_text(face="bold", colour="#990000", size=10),
        axis.text = element_text(size=7))+
  scale_x_continuous(limits = c(0,10),breaks = seq(0,10,2))+
  scale_y_continuous(limits = c(0,1),breaks = seq(0,2,0.2))
ggsave("IAT.png",width = 11.63,height = 7.29,units = "in")


#agregación del indicador a la base antes
df <- as.data.frame(cbind(df, IAT = iat$IAT))


#compartiendo codigo en df original
df[,c(6:26)] <- Indicador[,2:22]
df[,52] <- Indicador[,23]

#Correla matrix para comprobar la relación directa entre indicador y valores elegidos
C <- Indicador[,2:23]
C <- as.data.frame(C)
cat <- names(C)
C <- data.frame(sapply(C[,cat], as.numeric ))
C <- C - 1
C <- as.data.frame(cbind(x = rowSums(C),y = iat$IAT))
cor(C$x,C$y) #0.6247


# Categorizacion del IAT ---------------------------------------------------------
df <- df %>% mutate(codIAT = cut(IAT,c(-0.01,20,40,60,80,100)))
levels(df$codIAT) <- c("Muy bajo","Bajo","Medio","Alto","Muy alto")
round(prop.table(table(df$codIAT))*100,digits = 3)

#seleccion de productores agropecuarios con IAT muy bajo
df2 <- df %>% filter(codIAT == "Muy bajo")
df2$codIAT <- droplevels(df2$codIAT)


# Agrupamiento -------------------------------------------------------------------
cluster<-df2 %>% select(IAT,LeeyEsc,SSHH,CocinaMejorada,Computadora,Internet,Tenencia,
                        NivelEduc,IngSufic,DejoTrab)

#Acondicionamiento de variables para Kohonen
cluster$IAT<-cluster$IAT/100
cluster$NivelEduc<-(cluster$NivelEduc - 1)/5


#redes neuronales de kohonen
```

```r
set.seed(666)
cluster <- as.matrix(cluster)
data.som <- som(X = cluster, grid = somgrid(10, 10, "hexagonal"))

plot(data.som,main = "Niveles de Desarrollo")
plot(data.som,type = "counts",main = "Conteo")
plot(data.som,type = "changes",main = "Cambios")
plot(data.som,type = "quality",main = "Calidad")
plot(data.som,type = "dist.neighbours", main = "distancia al vecino mas cercano")

codes<-as.data.frame(data.som$codes)

fviz_nbclust(codes, kmeans, method = "silhouette",k.max = 40)+
  labs(subtitle = "Método Silueta",title = "Número Óptimo de Agrupaciones")+
  xlab(label = "Número de agrupaciones") +
  ylab(label = "Ancho de silueta promedio")


cores <- c("red","red4","deeppink","deeppink4","salmon","orange","orange4","gold",
           "brown3","darkorchid","purple4","white","gray50","gray25","darkseagreen2",
           "green","green4","tan3","blue","darkblue","slateblue","khaki","black",
           "lightblue")


som_cluster <- cutree(hclust(dist(codes)), 24)
plot(data.som, bgcol = cores[som_cluster], main = "Agrupaciones según Kohonen")
add.cluster.boundaries(data.som, som_cluster)
ggsave("kohonen.png",width = 7,height = 7)


#asignacion de cluster
neuronas <- cbind(neu=seq(1,100,1),cluster=som_cluster)
segmento <- c()
for(i in 1:dim(df2)[1]){
  x <- data.som$unit.classif[i]
  segmento[i] <- neuronas[x,2]
}

df2 <- cbind(df2, segmento = segmento)

grupo <- df2 %>% group_by(segmento) %>% summarise(n=n(),LeeyEsc=mean(LeeyEsc),SSHH=mean(SSHH)
,
                                        cocina=mean(CocinaMejorada),Comp=mean(Computadora),
                                        Internet=mean(Internet),Tenencia=mean(Tenencia),
                                        IAT=mean(IAT),Educ=median(NivelEduc),
                                        sup=mean(SupAgc),
                                        Ing = mean(IngSufic),
                                        Trab = mean(DejoTrab),
                                        Dist = mean(HorasCap))
write.csv(grupo,"grupos.csv")

df2 <- df2 %>% filter(segmento %in% c(1:4, 8, 12:14, 16:19, 21:24))


df <- df %>% left_join(df2[,c(5,74)],by = "NumCedula")
df <- df %>% mutate(filtro = if_else(is.na(segmento), 0, 1))

# Densidades de Kernel ----------------------------------------------------------------


sea<- df %>% group_by(SEA) %>% summarise(long = mean(LongSea),
                                         lat = mean(LatSea),
                                         filter = mean(filtro))
sea <- sea %>% filter( filter >= 0.8)



ggplot(sea, aes(x = long, y = lat)) +
  geom_point()+
  coord_map()
```

```
map <- get_googlemap(center = c(-74.971860, -10.682910), zoom = 5, maptype = "terrain",source
= "google",style = 'feature:administrative.country|element:labels|visibility:off')
p <- ggmap(map)
p <- p + scale_x_continuous(limits = c(-81.414020,-68.529701)) + scale_y_continuous(limits =
c(-18.445906,-2.919910)+c(0,-0.1))
p <- p + xlab("Longitud") + ylab("Latitud")
p <- p + geom_point(data = sea, aes(x = long, y = lat), size = 0.1)
ggplotly(p)
ggsave("peru.png",p)



0.68*(704^(-0.2))
x<-sea$long
y<-sea$lat
sp <- ppp(x, y, c(min(x)-0.5, max(x)+0.5), c(min(y)-0.5, max(y)+0.5))


plot(sp)
plot(density(sp,0.1832291))
contour(density(sp,0.1832291), axes = FALSE)
persp(density(sp,0.1832291))
densidad <- density(sp,0.1832291)


nbcol <- 100
color <- rev(rainbow(nbcol, start = 0/6, end = 6/6))
zcol  = cut(densidad$v, nbcol)
persp3d(x = densidad$xcol,y = densidad$yrow  , z=densidad$v,
        theta=50, phi=25, expand=0.75, col=color[zcol],
        ticktype="detailed", xlab="", ylab="time", zlab="",axes=TRUE)



#gráficas de criterio espacial #

#Funcion G estimada
G<-Gest(sp,correction="none")
ggplot(data=G,aes(x=G$r))+
  geom_line(aes(y=G$theo,colour="Función Teórica"),linetype="solid",size=1)+
  geom_line(aes(y=G$raw,colour="Función Empírica"),linetype="dashed",size=1)+
  theme(axis.title = element_text(face="bold", colour="#990000", size=10),axis.text = element
_text(size=7))+
  xlab("Distancia entre eventos (r)")+
  ylab(expression(paste("Función ",hat(G),"(r)")))+
  guides(colour=guide_legend(title=NULL))

#Funcion L estimada
L<-Lest(sp,correction="none",nlarge=1000)
ggplot(data=L,aes(x=r))+
  geom_line(aes(y=theo,colour="Función Teórica"),linetype="solid",size=1)+
  geom_line(aes(y=un,colour="Función Empírica"),linetype="dashed",size=1)+
  theme(axis.title = element_text(face="bold", colour="#990000", size=10),axis.text = element
_text(size=7))+
  xlab("Distancia entre eventos (h)")+
  ylab(expression(paste("Función ",hat(L),"(h)")))+
  guides(colour=guide_legend(title=NULL))+
  scale_x_continuous(limits = c(0,3),breaks = seq(0,3,0.5))



# Estimación de segundo orden
d<-density(sp,window="kernel",kernel = "quartic",0.1832291)
par(mfrow = c(1,1))
plot(sp,title=NULL)
plot(d,title = NULL)
persp(d,title = NULL)
```

```
contour(d, title = NULL)
summary(sp)
plot(layered(d,sp))


sea$SEA<-as.numeric(sea$SEA)
map <- get_googlemap(center = c(-74.971860, -10.682910), zoom = 5, maptype = "terrain",source
= "google",style = 'feature:administrative.country|element:labels|visibility:off')
p <- ggmap(map)
p <- p + scale_x_continuous(limits = c(-81.414020,-68.529701)) + scale_y_continuous(limits =
c(-18.445906,-2.919910)+c(0,-0.1))
p <- p + xlab("Longitud") + ylab("Latitud")

# Reconocimientos de SEA
p<-p+geom_point(data=sea,aes(x=long,y=lat))+ggthemes::theme_map()
ggplotly(p)


ggplotly(p+
           geom_point(data=sea,
                      aes(x = long, y = lat, text = paste("sea: ",SEA)),
                      size = 5,
                      shape = ".",
                      color="black")+
           stat_density2d(
           aes(x = long, y = lat, fill = ..level.., alpha = ..level..),
           h = 0.1832291,
           data = sea,
           geom = "polygon",
           bins = 150)+
           scale_fill_gradient(low = "green", high = "red")+
           theme(legend.position = 'none')
)


# Especificación de zonas ######
df2$Provincia <- paste(df2$Departamento,df2$Provincia,sep = "")
df2$Distrito <- paste(df2$Provincia,df2$Distrito)
df2$Distrito <- str_replace_all(df2$Distrito,fixed(" "),"")

dep<-read.dta("Ubig_Dep.dta",convert.factors = F)
prov<-read.dta("Ubig_Prov.dta",convert.factors = F)
dist<-read.dta("Ubig_Dist.dta",convert.factors = F)

f <- read_xlsx("focos.xlsx", sheet = "Hoja1")
foco <- f %>% filter(Grupo == "foco 7")

df3 <- df2 %>% filter(SEA %in% foco$SEA)

for (i in 1:dim(df3)[1]) {
  a <- df3$Departamento[i]
  df3$Departamento[i] <- dep$var2[dep$p001==a]
}


for (i in 1:dim(df3)[1]) {
  a <- df3$Provincia[i]
  df3$Provincia[i] <- prov$var3[prov$prov==a]
}


for (i in 1:dim(df3)[1]) {
  a <- df3$Distrito[i]
  df3$Distrito[i] <- dist$var4[dist$ubigeo == a]
}

df3 %>% group_by(Provincia) %>% summarise(n=n(),LeeyEsc=mean(LeeyEsc),SSHH=mean(SSHH),
                                                    cocina=mean(CocinaMejorada),Comp=mean(Compu
tadora),
                                                    Internet=mean(Internet),Tenencia=mean(Tenen
```

```
cia),
                                                IAT=mean(IAT),Educ=median(NivelEduc),
                                                sup=mean(SupAgc),
                                                Ing = mean(IngSufic),
                                                Trab = mean(DejoTrab),
                                                Dist = mean(HorasCap),
                                                edad = mean(Edad))

round(prop.table(table(df3$SSHH,df3$CocinaMejorada))*100,2)
prop.table(table(df3$NivelEduc))
prop.table(table(df3$Idioma))
summary(df3$Edad)
prop.table(table(df3$ActPrincipal))
prop.table(table(df3$DejoTrab))
prop.table(table(df3$IngSufic))
prop.table(table(df3$RazonProd))prop.table(table(df3$ExisteAnim))
prop.table(table(df3$GanadoVac))
prop.table(table(df3$UsoAnimAgr))
prop.table(table(df3$DestinoProd))
summary(df3[,c(6:25,52)])



# Análisis inferencial  ----------------------------------------------------------------

#Tamaño de muestra
inf <- read.csv("datatesisexfinal.csv", sep = ",", header = TRUE)
table(inf$codIAT)
N <- dim(inf)[1]

n <- (1.96*sqrt(0.99*0.01)/0.01)^2
n <- round(n,0)

set.seed(111)
index <- sample(x = seq(1:N),size = n, replace = FALSE)


#Hipotesis 1
cod <- as.data.frame(inf$codIAT[index])
names(cod) <- "cod"
head(cod)
cod <- cod %>% mutate(indicador = ifelse(cod == "Muy bajo",1,0))
table(cod$indicador)
prop.test(x = 378, n =380, p = 0.75, conf.level = 0.95, alternative = "greater")

#Hipotesis2
energia <- as.data.frame(inf$EnerElect[index])
names(energia) <- "energia"
head(energia)
energia <- energia %>% mutate(indicador = ifelse(energia == "Enel_si",1,0))
table(energia$energia)
prop.test(x = 5, n =380, p = 0.3, conf.level = 0.95, alternative = "less")

#hipótesis 3

foco <- f %>% filter(Grupo == "foco 8")
foco8 <- df2 %>% filter(SEA %in% foco$SEA)
foco8$foco <- "foco8"
focos <- rbind(foco1,foco2,foco3,foco4,foco5,foco6,foco7,foco8)
h3 <- manova(cbind(IAT,NivelEduc,Tenencia,SSHH,LeeyEsc,IngSufic,DejoTrab)~foco,data = focos)
summary(h3)
focos$foco <- as.factor(focos$foco)
names(focos)
```