ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

IDENTIFICACIÓN DE CLÚSTERS ESPACIALES DE LOS RAYOS GAMMA EN LA PROVINCIA DE CHIMBORAZO

Trabajo de Titulación

Tipo: Proyecto de Investigación

Presentado para obtener el grado académico de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

AUTOR:

WILSON PAUL ERAZO SALAO

Riobamba – Ecuador

2022

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

IDENTIFICACIÓN DE CLÚSTERS ESPACIALES DE LOS RAYOS GAMMA EN LA PROVINCIA DE CHIMBORAZO

Trabajo de Titulación

Tipo: Proyecto de Investigación

Presentado para obtener el grado académico de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

AUTOR: ERAZO SALAO WILSON PAUL

DIRECTORA: Ing. AMALIA ISABEL ESCUDERO VILLA MSc.

Riobamba – Ecuador

2022

Yo, WILSON PAUL ERAZO SALAO, declaro que el presente Trabajo de Titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autor asumo la responsabilidad legal y académica de los contenidos de este Trabajo de Titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 14 de diciembre de 2022

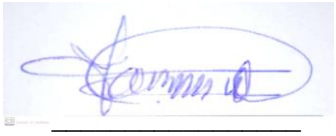
Wilson Paul Erazo Salao
060420979-1

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

El Tribunal del Trabajo de Titulación, certifica que: El Trabajo de Titulación: Tipo: Proyecto de Investigación, **IDENTIFICACION DE CLÚSTERS ESPACIALES DE LOS RAYOS GAMMA EN LA PROVINCIA DE CHIMBORAZO**, realizado por el señor: **WILSON PAUL ERAZO SALAO**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

	FIRMA	**FECHA**
Ing. Johanna Enith Aguilar Reyes MSc. **PRESIDENTE DEL TRIBUNAL**		2022-12-14
Ing. Amalia Isabel Escudero Villa MSc. **DIRECTORA DEL TRABAJO DE TITULACIÓN**		2022-12-14
Dra. Jheny Del Carmen Orbe Ordoñez PhD. **ASESORA DEL TRABAJO DE TITULACIÓN**		2022-12-14

DEDICATORIA

Este trabajo está dedicado a:

La forjadora de mi camino mi madre Nancy, por el apoyo moral y económico, por cada consejo y la confianza que me brindo a lo largo de mi vida. Con valores y principios la que me permitió luchar por mis sueños y metas.

A mi hermano Erick, que con su ejemplo y experiencia han sabido forjarme por el buen camino.

Y demás familiares, que me brindaron su apoyo incondicional, los cuales permitieron no decaer, y seguir adelante en el cumplimiento de mi meta académica.

Con todo amor gracias a todos.

Wilson

AGRADECIMIENTO

Agradecer a:

Toda la planta docente de la Carrera de Estadística, que semestre a semestre me fueron impartiendo conocimientos sólidos. Los conocimientos adquiridos me han servido para crecer como persona y profesional.

La Ing. Isabel Escudero tutora de mi tesis y docente de varias cátedras a lo largo de mi vida estudiantil, y a la Dra. Jheny Orbe miembro del trabajo de titulación que ha guiado de manera óptima para culminación de esta investigación.

Al centro de investigaciones de la Escuela Superior Politécnica de Chimborazo GIDAC por el apoyo brindado y el apoyo en todas las dudas y necesidades que presente.

Wilson

ÍNDICE DE CONTENIDO

CAPÍTULO III

CAPÍTULO IV

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE ANEXOS

RESUMEN

En la presente investigación se realizó un análisis estadístico de las tasas de dosis de radiación gamma (tdrg) presentes en la provincia de Chimborazo. Se realizó un análisis exploratorio tradicional y espacial, así como también un multivariado. Los datos fueron muestreados y facilitados por Grupo de Investigación y Desarrollo para el Ambiente y Cambio Climático (GIDAC). El análisis exploratorio se centró en las medidas de tendencia central, de posición, boxplot, y en el enfoque espacial el índice de Morán y mapa de densidades de la tdrg. Se realizó la comparación de las tdrg entre cantones, periodos temporales (mañana y tarde) y zonas (urbana y rural), en las que se identificó diferencias significativas al 95% de confiabilidad. Se utilizó herramientas estadísticas multivariantes como: el cubo de homogeneidad, dendrogramas, componentes principales, identificación de los puntos clusterizados y construcción de clúster. Se observó dos conglomerados espaciales homogéneos agrupados respecto a la variable altitud. Se realizo interpolaciones IDW y Kriging de la tdrg al 95% de confiabilidad, mostrando una relación positiva con la latitud y mayor eficiencia con el IDW según el análisis de errores. Se concluye según el análisis descriptivo de la radiación gamma en los diferentes cantones de la provincia de Chimborazo presentó de manera similar, con un índice de radiación en un intervalo de 0.032 a 0.079 Sv, este fenómeno se provoca por la emanación de radiación constante que mantiene el país, siendo una de las razones el estar dentro de la cordillera andina o línea de fuego, por lo que se determinó que presentan de manera similar. Se recomienda preservar el apoyo y guía por parte del GIDAC para futuros proyectos de investigación evidenciando la funcionalidad y procesos estadísticos en diferentes áreas de investigación y desarrollo en beneficio de la sociedad.

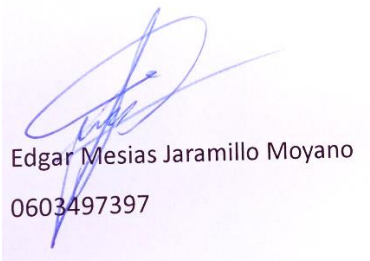
Palabras clave: <CLÚSTER ESPACIALES>, <RAYOS GAMMA>, <GEOESTADÍSTICA>, <PATRONES DE COMPORTAMIENTO>, <CORRELACIÓN ESPACIAL>.

0176-DBRA-UPT-2023

ABSTRACT

In this investigation, a statistical analysis of the gamma radiation dose rates (grdr) present in the province of Chimborazo was done. A traditional and spatial exploratory analysis as well as a multivariate analysis were performed. The data was sampled and provided by the Research and Development Group for the Environment and Climate Change (GIDAC). The exploratory analysis focused on the measures of central tendency, position, boxplot, and in the spatial approach the Moran's index and the density map of the grdr. The comparison of the grdr between cantons, time periods (morning and afternoon) and zones (urban and rural) was made, in which significant differences were identified at 95% reliability. Multivariate statistical tools were used, such as: the homogeneity cube, dendrogram, principal components, identification of clustered points, and cluster construction. Two homogeneous spatial conglomerates grouped with respect to the altitude variable were observed. IDW and Kriging interpolations of the grdr were performed at 95% reliability, showing a positive relationship with latitude and greater efficiency with IDW according to error analysis. It is concluded according to the descriptive analysis of gamma radiation in the different cantons of the province of Chimborazo presented in a similar way, with a radiation index in an interval of 0.032 to 0.079 Sv, this phenomenon is caused by the emanation of constant radiation that maintains the country, since it is within the Andean Mountain range or line of fire, which is one reason why it presents in a similar way. It is recommended to preserve the support and guidance from the GIDAC for future research projects evidencing the functionality and statistical processes in different areas of research and development for the benefit of society.

Keywords: <SPATIAL CLUSTER>, <GAMMA RAYS>, <GEOSTATISTICS>, <BEHAVIOR PATTERNS>, <SPATIAL CORRELATION>.



INTRODUCCIÓN

Durante las últimas dos décadas el desarrollo teórico como observacional de los rayos gamma, ha permitido tener una nueva visión de esta variable, ya que son radiaciones electromagnéticas producidas por la desintegración radiactiva de los núcleos atómicos. (Planas, 2019). A través de los años la recolección de radiación electromagnética ha mejorado; sin embargo, los registros en los que se basa la meteorología aplicada en radiación gamma aún son deficientes, tanto en calidad como en cantidad. La presencia de altas dosis de radiación y astro partículas en la atmosfera en todo el mundo, ha afectado la salud de los seres humanos , generando por ejemplo, presencia de cáncer en la piel, foto envejecimiento, hiperplasia, melanogénesis, entre otros (González. et. al 2009, p. 70). Por tal motivo se ha incrementado el interés del estudio de esta variable mediante técnicas estadísticas espaciales como: análisis de clúster, análisis de componentes principales, conglomerados jerárquicos, variograma, interpolaciones, mapeos geográficos, entre otras, que coadyuvan al estudio de su comportamiento.

En este trabajo de investigación se pretende identificar patrones de comportamiento de las tasas de dosis de rayos gamma en la provincia de Chimborazo mediante clústers espaciales. Para ello se dispuso de las longitudes, latitudes, cantones, zonas, periodo temporal de la variable en estudio.

Esta tesis se compone en cuatro capítulos que se describen a continuación:

En el primer capítulo se describe los aspectos metodológicos de la investigación que explican: el planteamiento del problema, justificación, objetivos. Se detalla de manera objetiva el enfoque de la investigación, para ello primero se realizó un bosquejo bibliográfico para estudios semejantes que presenten técnicas estadísticas acordes al tema de estudio.

En el segundo capítulo se detalla el sustento teórico sobre las tasas de dosis de rayos gamma (tdrg), así como también los principios fundamentales de explicación conceptual de técnicas estadísticas, como análisis multivariado, componentes principales, análisis de conglomerados con un enfoque espacial; el conocimiento sobre cartografía y generación de mapas de clasificación e interpolación en datos espaciales permitió una descripción gráfica y analítica del método IDW y Kriging que dio realce al objeto en estudio.

En el tercer capítulo se describe los aspectos metodológicos (enfoque, nivel, diseño y tipo de estudio de la investigación). El cuarto capítulo se detalla los resultados obtenidos, tanto en el análisis exploratorio, la heterogeneidad espacial con una exploración geográfica para identificar patrones de comportamiento mediante el uso de clústeres espaciales jerárquicos aglomerativos, aplicando técnicas multivariadas y mapas. Finalmente se hace algunas conclusiones y recomendaciones.

Antecedentes

Los rayos gamma son las radiaciones electromagnéticas producidas por la desintegración radiactiva de los núcleos atómicos. (Planas, 2019). Fueron descubiertos poco después del descubrimiento de los rayos X. En 1896, el científico francés Henri Becquerel descubrió que los minerales de uranio podrían exponer una placa fotográfica a través de otro material, por otra parte, en estudios posterior en 1914 se observó que los rayos gamma se reflejaban en las superficies de los cristales, lo que demuestra que deben ser radiaciones electromagnéticas, pero con mayor energía (mayor frecuencia y longitudes de onda más cortas). (Connor, 2020).

La geoestadística se centra en el estudio de variables que cambian en el espacio de manera continua, como las observaciones en unos pocos puntos de medición a partir de las cuales se intenta estimar la variable de interés en otros puntos de la región de estudio (Rubio, 2019). Hasta la actualidad se han desarrollado múltiples aplicaciones en diferentes campos, por ejemplo, Banda (2010) caracteriza los rayos gamma en el universo externo mediante la interpolación basada en distancias más lejanas, puesto que el método Hubble presento inconsistencias por falta de muestra. (Banda, 2010). Aplicaciones de clúster espaciales en precipitaciones registradas en 150 estaciones meteorológicas localizadas en el departamento del valle del Cauca, Colombia, en el cual manejaron técnicas de encadenamiento simple, Ward y centroide, como métodos jerárquicos de aglomeración y la distancia euclídea al cuadrado (DEC) como medida de similitud (Heredia, et. al 2012, p. 11). Mapeo de rayos gamma para identificar el rango de energía con el propósito de evaluar su comportamiento a través de la interpolación Kriging. (Tarela, P; Mariscotti, M; Perone, E, 1993).

Simulación de Monte Carlo para la medición de la intensidad y detección de rayos gamma a partir de la cual obtuvieron algunos parámetros de referencia, también determinaron estrategias de selección de la señal de rayos gamma y lo validaron mediante extensas pruebas, haciendo uso de electrones produciendo rayos gamma en un blanco (Sevilla, Ignacio, 2008). En la universidad Militar Nueva Granada realizaron mapeos en entornos altamente geológicos que brindan información primordial tanto a la parte de suelo como al lecho de roca, y en áreas con situaciones que presentan dificultades como son la emisión de rayos gamma.

Para este propósito, se discute el uso del método de mapeo espacial en la región de Midland, norte de Irlanda utilizando datos de rayos gamma de alta resolución adquiridos en el 2005, en grillas interpoladas. Los mapas de clasificación supervisada resultantes fueron considerados como modelos espaciales. (Martinez Norma, 2017); entre otros múltiples ejemplos más. La unidad académica de estudios nucleares ha desarrollado un mapeo radiológico ambiental mediante espectrometría gamma. En este mapeo incluyeron sitios conurbados, rurales y arqueológicos. Aún

y cuando no se han encontrado valores que sobrepasen los límites radiológicos establecidos, vieron necesaria la realización de un análisis de correlaciones que muestren posibles regularidades entre los mismos o que permitan contrastarlos desde el punto de vista estadístico en la cual, se le dieron enfoque en el análisis estadístico de componentes principales (PCA), análisis de clúster, mostraron resultados óptimos y un comportamiento con una distribución normal (Juan Lopez, 2017).

En Ecuador existe una literatura limitada de estudios de rayos gamma por ello el GIDAC promueve la importancia de estudio de esta variable mediante el proyecto denominado "Evaluación de elementos radioactivos de la serie de Uranio 238 en el ambiente de pacientes de cáncer", las tasas de dosis a las cuales estamos expuestos ya que por generaciones que ha pasado debido a los factores naturales nos hemos visto a la necesidad de realizar estudios de indagación mucho más profunda mediante técnicas y métodos estadísticas espaciales para caracterizar su comportamiento.

CAPÍTULO I

1. PROBLEMA DE INVESTIGACIÓN

1.1. Planteamiento del problema

Una de las principales preocupaciones de los físicos y biofísicos es contar con información depurada y confiable para el conocimiento de la física de astro partículas. La aparición de este campo fue parcialmente promovida por el descubrimiento de la oscilación de los neutrinos, teniendo un rápido desarrollo, tanto en la parte teórica como en la experimental desde principios de 2000. Sin embargo, en el Ecuador es un área muy poco investigada, especialmente en zonas como la región interandina. Sumado a esto la instalación de estaciones con dispositivos (sensores) para la recolección, procesamiento y validación de datos astrofísicos, requiere una significativa inversión por parte del estado. Por lo que, existen limitaciones para su investigación, particularmente de la presencia de astro partículas como los rayos gamma. Sin embargo, a pesar de la reducida literatura, el GIDAC-ESPOCH suma esfuerzos en el estudio en la provincia de Chimborazo, con el fin inicial de caracterizarlo. Motivados por estas razones se da lugar a este trabajo de investigación, para indagar herramientas, métodos y técnicas estadísticas que permitan analizar los datos disponibles, fortalecer las metodologías empleadas para la toma y almacenamiento de datos, y proveer información que evoque en investigaciones más amplias a futuro.

1.2. Formulación (Incógnita)

¿La aplicación de clúster espaciales permitirá identificar patrones de comportamiento de astro partículas (rayos gamma) en la provincia de Chimborazo?

1.3. Objetivos

1.3.1. Objetivo general

- Identificar patrones de comportamiento en los rayos gamma de la provincia de Chimborazo, mediante un análisis de clúster espacial.

1.3.2. Objetivos específicos

- Realizar un análisis estadístico exploratorio de los datos de rayos gamma.
- Realizar un análisis descriptivo espacial de los datos.
- Analizar la auto correlación espacial de los rayos gamma.
- Identificar patrones de comportamiento mediante clúster espaciales.
- Realizar al menos un mapa aproximado de la presencia de rayos gamma en la provincia de Chimborazo.

1.4. Justificación

El flujo de partículas secundarias, producto del astro partículas es dependiente de las variables atmosféricas y geofísicas como: presión, temperatura, densidad, altura y campo geomagnético local. La densidad de la atmósfera refleja cambios en las concentraciones de sus especies químicas, y esto produce una variación en la interacción del astro partículas con los núcleos atmosféricos, reflejada en la tasa de partículas secundarias al nivel del detector.

Al haber menos núcleos atómicos para interaccionar disminuye el flujo de radiación gamma secundaria y viceversa. De igual manera la interacción del astro partículas depende de la intensidad de campo magnético local y rigidez de corte. Esto quiere decir pueden causar grave daño al núcleo de las células, por lo cual se usan para esterilizar equipos médicos y alimentos, pero también en elevadas concentraciones puede causar dato a los seres vivos.

La finalidad de esta investigación es identificar los patrones de comportamiento de los rayos gamma en la provincia de Chimborazo según su propagación, mediante la utilización de métodos y técnicas estadísticas, el análisis exploratorio, correlación y clúster espaciales. Resultados que proporcionarán información relevante para el GIDAC e investigaciones posteriores, así como también a la sociedad en general.

CAPÍTULO II

2. MARCO TEÓRICO REFERENCIAL

2.1. Bases teóricas

2.1.1. Astro partículas

El astro partículas, son partículas neutras como: neutrinos (ν), rayos gamma (γ), o partículas cargadas como: protones, electrones y átomos de elementos pesados provenientes de fuentes galácticas y extra galácticas. Estas interaccionan con los elementos presentes en la atmósfera alta de la Tierra llamada exósfera. En la exósfera, aproximadamente el 2 % son electrones y el 98 % son protones y núcleos, de los cuales aproximadamente el 87 % son protones, el 12 % son núcleos de helio y el 1 % restante son núcleos más pesados. Las principales interacciones que se presentan en la exósfera son básicamente de dos tipos: electromagnética y hadrónica. Estas producen cascadas de partículas secundarias llamadas Extensive Air Showers (EASs, por sus siglas en inglés). Las EASs electromagnéticas se forman cuando un rayo γ o electrón impacta con la exósfera, y las EASs hadrónicas se forman cuando un núcleo de hidrógeno, núcleo de helio o núcleo pesado impacta con la exósfera (Gutiérrez, 2020, p. 10).

2.1.2. Fuentes de Energía

2.1.2.1. Sol

Principal fuente de radiación que la tierra ha estado expuesta y que produce ondas electromagnéticas en casi todos los rangos y que son altamente energéticas debido a las explosiones y ráfagas solares que posee. El flujo de partículas de esta fuente llega constantemente a la tierra, donde está, es protegida por su campo magnético (Correa, 2020, p. 8).

2.1.2.2. Supernovas

Colapso gravitacional de estrellas cuando finaliza su tiempo de vida (cesan las reacciones nucleares) debido a un desbalance entre la gravedad y la presión de radiación, el material que sale disparado por esta explosión llamado “remanente” genera radiaciones altamente energéticas y enriquece todo el medio estelar para la formación de nuevas estrellas. Aunque a veces se ocultan en el polvo estelar, la gran luminosidad que producen ha servido para el descubrimiento de estas (Correa, Iván, 2020, p. 8).

2.1.2.3. Cuásares

Fuentes de radio y ópticas más brillantes del cielo, pero con grandes desplazamientos al rojo. Se les asocia con agujeros negros entre $10^6 - 10^9$ masas solares en los centros galácticos y tiene luminosidades 10^{13} > que el sol (1026 W) (Correa, Iván, 2020, p. 9).

2.1.2.4. Núcleos activos de galaxias (AGNs)

Asociados con los cuásares, se dice que los agujeros negros de estas galaxias consumen materia (estrellas, gas, polvo) formando un disco de acreción resultado de la disminución y conservación del momento angular, y los mismos forman jets de energía. Cuando el material se agota, la AGN se "apaga" y solo queda un agujero negro supermasivo, por eso la luminosidad varia con el tiempo (Correa, Iván, 2020, p. 9).

2.2. Bases conceptuales

2.2.1. Clasificación de Variables

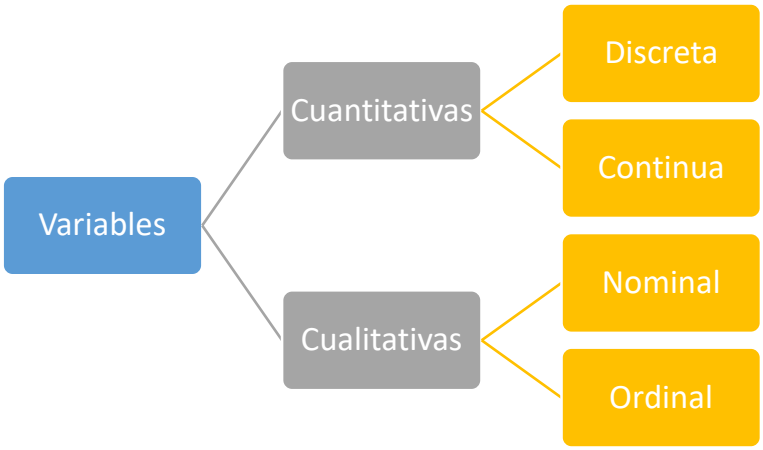


Ilustración 1-2: Clasificación de las variables

Fuente: (Congacha, 2012)

2.2.2. Análisis exploratorio de datos

- Medidas de tendencia central

La media $(\bar{x})$, la mediana (Me), la moda (Mo) y el rango son estadísticos, también conocidos como estadísticos de centralización. La media se calcula sumando todos los valores de la variable y luego dividiendo por el número de observaciones. Su característica primordial es que es estable

en el muestreo, es más uniforme de muestra a muestra que los otros estadígrafos de posición (Estrella 2008, p. 8).

$$\bar{x}_i = \frac{\sum x_i}{n}$$

La mediana es el valor de la variable que deja el mismo número de datos antes y después que él, una vez ordenados éstos (Estrella 2008, p. 9).

$$\tilde{x} = x + \frac{n+1}{2}$$

La moda es el valor que cuenta con una mayor frecuencia en una distribución de datos. Es un estadígrafo de posición que se define como el valor más frecuente (Estrella 2008, p. 10).

- Medidas de dispersión

Entregan información sobre la variación de la variable. Pretenden resumir en un solo valor la dispersión que tiene un conjunto de datos. Las medidas de dispersión más utilizadas son: rango de variación, varianza, desviación estándar (Ricardi 2011, p. 2). El rango de variación se define como la diferencia entre el mayor y menor valor de la variable.

$$Rango\ de\ variación\ =\ Máximo - Mínimo$$

La varianza poblacional se representa con σ^2 y la muestral con s^2. La desviación estándar la raíz cuadrada de la varianza. La varianza se expresa en unidades de variable al cuadrado y la desviación estándar simplemente en unidades de variable (Ricardi 2011, p. 2).

$$\sigma^2 = \frac{\sum_i^n (x_i - \mu)^2}{N} \sim Varianza$$

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \mu)^2}{N}} \sim desviación\ estandar$$

2.2.3. Análisis de varianza (ANOVA)

El Analysis Of Variance (ANOVA) según la terminología inglesa es uno de los aspectos más interesantes dentro del tema de las pruebas de hipótesis, por el ingenio desplegado en su desarrollo y, quizás, por las variadas formas que puede tomar. El ingenioso fue Ronald Aylmer Fisher (1890-1962), un genetista que fue uno de los estadísticos más influyentes del siglo XX. No existe un

solo ANOVA, sino que el este comprende una serie de técnicas cuya aplicación particular depende del diseño experimental. Para sus usos más complejos. El ANOVA permite analizar la variación en una variable de respuesta (variable continua aleatoria) medida en circunstancias definidas por factores discretos (variables de clasificación). Se usa un ANOVA en cuatro situaciones (Dagnino 2014, p. 306).

1. Cuando hay más de dos grupos que necesitan ser comparados. El ANOVA también puede ser usado para comparar solamente dos grupos; de hecho, el test t de Student es un caso especial de ANOVA de una vía.
2. Cuando hay mediciones repetidas en más de dos ocasiones o cuando hay dos o más grupos en quienes se hacen mediciones repetidas en dos ocasiones.
3. Cuando los sujetos pueden variar en una o más características que afectan el resultado y se necesita ajustar su efecto.
4. Cuando se desea analizar simultáneamente el efecto de dos tratamientos diferentes, cuando el efecto de cada uno por separado y su posible interacción es importante.

- Anova de una vía

En el ANOVA, el término factor se refiere a la variable que determina los grupos del estudio, o sea, la variable independiente o predictora. El número de grupos definido por un factor se conoce como el número de niveles del factor, estos corresponden habitualmente a los tratamientos que se comparan. Cuando hay un solo factor de clasificación de los datos, se habla de un ANOVA de un factor, de una vía o de un sentido (Dagnino 2014, p. 362).

$$F = \frac{S_{entre}}{S_{dentro}}$$

2.2.4. Geografía

Según Habegger y Mancila, la geografía es un "procedimiento que permite obtener datos sobre el trazado de un territorio, para su posterior representación técnica y artística en un mapa como sistema predominan-te de comunicación". La cartografía se concibió desde las características socioespaciales de un territorio (Sánchez, 2011, p. 32).

La geografía habilita un escenario para la construcción de conocimiento colectivo y, a partir de allí, posibilita una acción transformadora del territorio. El ejercicio de dibujar la realidad pone en un mismo lenguaje saberes, imaginarios y deseos subjetivos, que, al socializarse a través de la

conversación, las tecnologías de la información y la comunicación (TIC), como el Flickr, y retroalimentados con fotografías, dibujos, dan paso a una construcción de nuevo territorio: La cartografía es una herramienta que nos permite ganar consciencia sobre la realidad, los conflictos y las capacidades individuales y colectivas. Abre caminos desde la reflexión compartida para consolidar lecturas y visiones frente a un espacio y tiempo específicos, para generar complicidades frente a los futuros posibles en donde cada uno tiene un papel que asumir (Sánchez, 2011, p. 32).

2.2.5. Los mapas y su clasificación

Los diferentes tipos de mapas que existen dan cuenta de la complejidad de la organización territorial de los seres humanos y de las regiones que se habita. Estos elementos relacionados con la cartografía pueden adoptar las formas más insospechadas, dependiendo de los criterios que hayan sido utilizados para diseñarlos (Torres, 2015, p. 2).

- Mapa político

Este es uno de los tipos de mapa en los que no se representan elementos físicos, sino que solo aparecen territorios políticos y sus límites: las fronteras. En ellos aparecen Estados o regiones con una cierta soberanía y autogobierno. Independientemente de la escala de lo que se quiere representar, ya sea una comarca o una organización supraestatal, se pone énfasis en las nociones de "dentro" y "fuera" (Torres, 2015, p. 2).

En los casos en los que exista un conflicto territorial que cree discrepancias entre muchas partes involucradas, las fronteras pasan a ser representadas con líneas discontinuas, recurso de representación que muchas veces también se utiliza para marcar los bordes de las aguas territoriales de un país (Torres, 2015, p. 2).

Por supuesto, como en un mapa político aparecen principalmente constructos teóricos, para realizar uno es necesario que existan ciertos consensos sobre los límites de cada entidad política representada (Torres, 2015, p. 3).

- Mapa geológico

Este tipo de mapa puede resultar similar al topológico, ya que en él se representan elementos naturales, pero en este caso no se pone tanto énfasis en el relieve y en la forma de la superficie terrestre, y se remarca más el tipo de minerales que componen el terreno. Esto último se expresa

muchas veces utilizando iconos como simbología para los distintos minerales y formaciones naturales como manantiales, volcanes, vetas de minerales especiales, y similares.

Así, las variaciones del territorio que aparecen tienen que ver con las características de la distribución de minerales y la forma de las placas tectónicas. Se trata de dar una imagen tanto de lo que hay en la superficie como de lo que se encuentra bajo tierra (Torres, 2015, p. 2).

- Mapa climático

Es este caso, se trata de remarcar las diferencias climáticas entre regiones. Eso es hecho coloreando de una manera homogénea cada zona que comparte un mismo tipo de clima, en ocasiones creando zonas de solapamiento (en las que se mezclan varios colores utilizando patrones de franjas finas (Torres 2015, p. 2).

- Mapa urbano

Los mapas urbanos ponen énfasis en los elementos propios de las zonas urbanizadas, es decir, las construcciones hechas por el ser humano y las vías de comunicación para peatones y vehículos, hasta el punto de que en muchos casos no aparece nada más que eso, exceptuando elementos naturales como costas y ríos (Torres 2015, p. 3)

Así pues, normalmente en ellos se representa solamente el espacio ocupado por una ciudad, distrito o barrio, todo a escala. El tipo de elementos gráficos utilizados para ello suelen ser sencillos y de estilo minimalista, en la mayoría de las ocasiones utilizando tan solo polígonos.

En ocasiones se utilizan cambios de color para señalar la presencia de diferentes tipos de espacios: centro antiguo, parques, playas, etc.

- Mapa de tránsito

Esta es una variación del mapa urbano en la que se representa casi exclusivamente el trazado de las rutas de transporte público de una ciudad, de manera muy simplificada. Los recorridos de los autobuses, trenes, redes de metro, y tranvías son representados con líneas de colores, y las estaciones son marcadas para los principales medios de transporte (Torres, 2015, p. 3)

- Mapa meteorológico

Este es el soporte utilizado para mostrar cuáles son o van a ser meteorológicos en cada región, siendo representados con símbolos que representan lluvia, tormenta, nublado, etc. Como consecuencia, la representación del territorio suele ser simple, creada solo para que cada región pueda ser reconocida fácilmente sin necesidad de incluir más información no relacionada con la meteorología, algo que saturaría visualmente la imagen (Torres 2015, p. 4).

- Mapas cualitativos

Estos mapas expresan variables de carácter nominal u ordinal y normalmente se utilizan para representar características del paisaje tales como uso-cobertura del suelo, geología, geomorfología o suelos (Fallas, 2003, p. 15).

- Mapas cuantitativos de superficie

Los mapas cuantitativos de superficie proporcionan tanto información cuantitativa de estudio, como sobre su distribución espacial. La información se mapea utilizando líneas de igual valor denominadas isopletas, isoaritmas o isolíneas o valores medios por unidad de área (coropletas) (Fallas, 2003, p. 16).

- Los mapas coropléticos muestran valores por unidad de área y se utilizan frecuentemente con unidades administrativas tales como fincas, distritos, cantones, provincias o países (unidades estadísticas). Los mapas coropléticos exhiben las características del área en forma simple y concisa y tienen como objetivo transmitir una impresión concreta de la realidad a partir del mapa (Fallas, 2003, p. 16).
- Los mapas isopléticos se elaboran a partir de puntos o centros de observación y muestran líneas con un valor constante. El valor de cada línea es estimado utilizando técnicas estadísticas tales como la interpolación lineal, el inverso cuadrático de la distancia o Kriging y su trazado se genera de forma manual o asistido por programas de computación o módulos específicos en los sistemas de información geográfica. Cuando se elaboren mapas que muestren densidades por unidad de superficie o relaciones entre atributos debe ponerse especial cuidado en la distribución espacial de la variable a mapear. El investigador debe asegurarse mediante un sistema de muestreo apropiado que los valores puntales a partir de los cuales se realiza la interpolación representan a cabalidad la realidad (Fallas, 2003, p. 16).

2.2.6. Técnicas estadísticas para datos espaciales

Las capacidades de cálculo y representación gráfica de los ordenadores actuales permiten de una forma sencilla, la obtención de una amplia variedad de gráficos y estadísticos diferentes y han hecho posible la aparición de una nueva filosofía en los estudios estadísticos: el análisis exploratorio de datos, introducido por Tukey (Batanero, et. al 1991, p. 2).

Con anterioridad a este enfoque, el análisis de datos se basaba fundamentalmente en el cálculo de estadísticos, conduciendo a dos consecuencias: En primer lugar, se disminuía la importancia visual de la representación de los datos, dándosela exclusivamente a los cálculos y en segundo se equiparaba el análisis con el modelo confirmatorio. En este tipo de análisis el conjunto de valores de las variables observadas se supone que se ajusta a un modelo preestablecido, calculando los estadísticos para aceptar o no una hipótesis, que es previa a la toma de las observaciones, las cuales han sido recogidas con el único propósito de poner tal hipótesis a prueba. Al contemplar solamente dos alternativas, confirmación o no de la hipótesis, los datos no se suelen explorar para extraer cualquier otra información que pueda deducirse de los mismos.

Para entender los principios por los que se guía el análisis exploratorio, se ha de tener en cuenta que los datos están constituidos por dos partes: la "regularidad" y las "desviaciones". La regularidad indica la estructura simplificada de un conjunto de observaciones (en una nube de puntos, por ejemplo, es la recta a la cual se ajusta). Las diferencias de los datos con respecto a esta estructura representan las desviaciones o residuos de los datos, que usualmente no tienen por qué presentar una estructura determinada. Tradicionalmente el estudio se ha concentrado en la búsqueda de un modelo que exprese la regularidad de las observaciones (Batanero, et. al 1991, p. 2).

- Características del análisis exploratorio de datos

Esta filosofía consiste en el estudio de los datos desde todas las perspectivas, y con todas las herramientas posibles, incluso las ya existentes. El propósito es extraer cuanta información sea posible, generar hipótesis nuevas, en el sentido de conjeturar sobre las observaciones de las que se dispone (Batanero, et. al 1991, p. 3).
El análisis espacial de datos tiene las siguientes características:

- Fuerte apoyo en representaciones gráficas: "Una idea fundamental del análisis exploratorio de datos es que al usar representaciones múltiples de los datos se convierte en un medio de desarrollar nuevos conocimientos y perspectivas. Esto puede ejemplificarse al pasar de tablas a gráficos, de lista de números a representaciones como la del "tronco", reduciendo los

números a una variedad discreta en un mapa estadístico para facilitar la exploración de la estructura total, construyendo gráficos, como el de la "caja" que hace posible la comparación de varias muestras".

- Empleo preferente de los estadísticos de orden, porque son sensibles a la mayor parte de los datos y con ellos se disminuye el efecto producido por los valores atípicos, escasos y muy alejados de la norma.
- No necesita una teoría matemática compleja, "Como el análisis de datos no supone que estos se distribuyen según una ley de probabilidad clásica (frecuentemente la normal, no utiliza sino nociones matemáticas muy elementales y procedimientos gráficos fáciles de realizar.
- Según las características anteriores es bastante parecida a la estadística descriptiva tradicional, pero se aleja de ella por su intención. Pues, al contrario que en ella, la representación o el cálculo no son en el análisis exploratorio de datos un fin, sino un medio de descubrir la información oculta en los mismos".

Uso de diferentes escalas o re-expresión: La escala en la que una de las variables es observada y registrada no es única. A veces, transformando los valores originales de la variable a una nueva escala se puede lograr que dichos valores sean más manejables. De este modo se incluye también el empleo de otros contenidos matemáticos, especialmente los referidos al concepto de función y el estudio de las propiedades de las funciones elementales (Batanero, et. al 1991, p. 4).

2.2.7. Variables Espaciales

Se debe considerar que el dato espacial es un dato cualquiera sin ninguna peculiaridad supone no realizar sobre él un análisis óptimo. Las características propias de los datos espaciales dotan a estos de una gran potencialidad de análisis, al tiempo que condicionan o limitan otras operaciones. Asimismo, estas particularidades son el origen de una gran parte de los retos aún existentes dentro del análisis geográfico, y por sus implicaciones directas no se desestiman sin más. Su conocimiento es, por tanto, imprescindible para todo tipo de análisis espacial (Vayá, 2000, p.33).

El carácter especial del dato espacial deriva de la existencia de posición. Esta posición se ha de entender tanto en términos absolutos (posición de una entidad en el espacio expresada por sus coordenadas) como relativos (relación con otras entidades también en dicho espacio). Las consecuencias de que todo dato espacial se halle por definición localizado a través de coordenadas son diversas, y deben enfocarse desde los distintos puntos de vista del análisis espacial (Vayá, 2000, p.33).

2.2.8. Estadísticos Espaciales

- Autocorrelación espacial

La medición de la correlación que una misma variable tiene en diferentes unidades espaciales contiguas en una perspectiva horizontal da lugar a una de estas tres posibilidades (Celemín, 2009, p. 3).

- Autocorrelación espacial positiva: las unidades espaciales vecinas presentan valores próximos. Indica una tendencia al agrupamiento de las unidades espaciales.
- Autocorrelación espacial negativa: las unidades espaciales vecinas presentan valores muy disímiles. Indica una tendencia a la dispersión de las unidades espaciales.

- Sin autocorrelación: no ocurre ninguna de las dos situaciones anteriores. Por lo tanto, los valores de las unidades espaciales vecinas presentan valores producidos en forma aleatoria.

Una forma de visualizar estos tres eventos se encuentra en la ilustración 2-2.

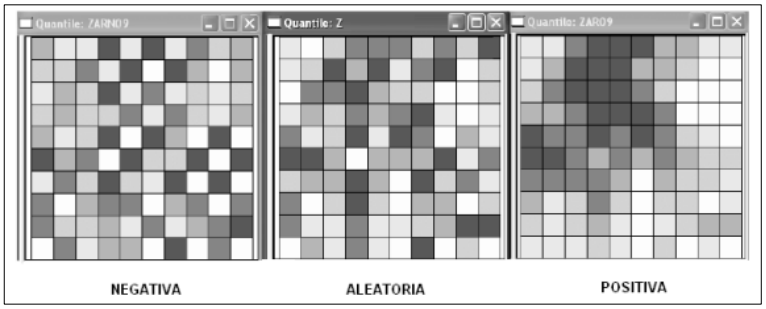


Ilustración 2-2: Tipos de autocorrelación espacial.
Fuente: (Anselin, 2003)

Los índices de AE permiten relacionar en forma conjunta la dependencia entre localizaciones y valores de variables o atributos que interesan y resultan muy adecuados para observar la configuración espacial fragmentada propia de nuestros tiempos. Existen distintos tipos de estadísticos o índices que permiten medir la AE cuya estructura general es la siguiente (Celemín, 2009, p. 4).

$$\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}C_{ij}$$

donde n es el total de lugares del mapa, W_{ij} son los elementos de una matriz (matriz de conexiones, contigüidad o de pesos espaciales) cuyos valores son una función de alguna medida de contigüidad en la matriz de datos originales (Rook, Bishop o Queen). El valor C_{ij} es una medida de la proximidad (distancia) de los valores $i\ y\ j$ en alguna dimensión (por ejemplo distancia euclídea, esférica, de Manhattan, etc.), o cualquier distancia definida por el usuario (Celemín, 2009, p. 4).

Generalmente en la mayoría de los análisis de AE se consideran las relaciones de vecinos próximos. Si se piensa que las áreas espacialmente referenciadas que se están analizando son cuadradas, habrá por lo menos cuatro vecinos que comparten un borde en cada lado del cuadrado. Asimismo, podría haber ocho datos espaciales para cada observación si se consideran adicionalmente aquellos puntos que limitan con los vértices del cuadrado en cuestión. Cuando se tienen en cuenta los cuatro elementos que comparten borde se habla de contigüidad tipo Rook. En el caso de los ocho vecinos se habla de contigüidad tipo Queen y si se toman solamente los vecinos contiguos por el vértice se denomina Bishop. Estos nombres corresponden al movimiento que realizan la torre, la reina y el alfil en un tablero de ajedrez (ver ilustración 3-2). El método Rook, por su simplicidad es el más utilizado (Celemín, 2009, p. 4).

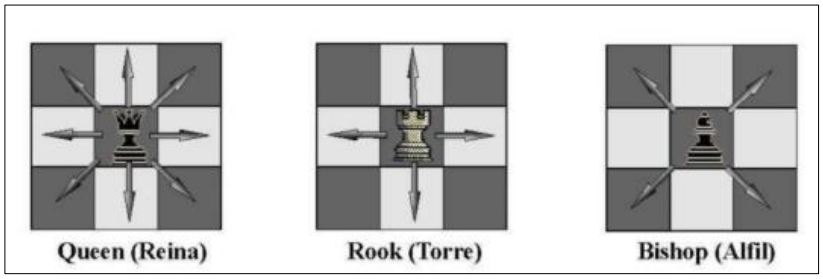


Ilustración 3-2: Tipos de contigüidad

Fuente: (Celemín, 2009)

- Heterogeneidad espacial

La heterogeneidad espacial es uno de los efectos espaciales que está relacionado con la diferenciación espacial o regional de las unidades geográficas. Se trata de un concepto que viene definido por la ausencia de estabilidad en el espacio del comportamiento humano o de otras relaciones en estudio. Esto implica que, en los modelos espaciales, las formas funcionales y los parámetros varían con la localización geográfica no siendo homogéneos para toda la matriz de datos. Esto es lo que ocurre, por ejemplo, en los modelos econométricos estimados con datos de corte transversal procedentes de unidades espaciales no similares, como es el caso de regiones ricas del norte y regiones pobres del sur. A diferencia de lo que sucede con la dependencia

espacial, el problema causado por la heterogeneidad espacial podría en gran parte ser resuelto mediante procedimientos de la econometría estándar (como el análisis clúster). Sin embargo, en algunos casos, la compleja interacción resultante de la estructura y los flujos espaciales pueden generar dependencia espacial combinada con heterogeneidad espacial, haciéndose altamente complicado distinguir entre ambos efectos (Yrigoyen, 2004, p. 1).

La heterogeneidad espacial surge cuando se trabaja con unidades espaciales (países, regiones, municipios, secciones censales) en las que se distribuye de manera distinta sobre el espacio, lo que suele ocurrir con situaciones del tipo centro periferia, norte-sur, este-oeste, etc. Según Anselin, la heterogeneidad espacial puede ser definida como "inestabilidad estructural en forma de varianza no constante de los residuos de una regresión (heteroscedasticidad) o en los coeficientes del modelo, que es posible abordar mediante técnicas de econometría tradicional o con herramientas propias de econometría espacial" (Yrigoyen, 2004, p. 2).

Hay tres razones por las que se debería analizar este efecto de heterogeneidad a través de técnicas propias de econometría espacial:

1. En primer lugar, la estructura que subyace en la inestabilidad espacial es de carácter geográfico, en el sentido de que la localización de las observaciones es fundamental para determinar la forma o especificación de dicha variabilidad. Éste sería, por ejemplo, el caso de la heteroscedasticidad de grupos ("groupwise"), que podría ser modelizada a través de tantos valores de la varianza de la perturbación aleatoria como distintos grupos geográficos compactos puedan derivarse de los datos (Yrigoyen, 2004, p. 2).

2. En segundo lugar, dado que la estructura es espacial, la heterogeneidad suele producirse conjuntamente con el problema de autocorrelación espacial, no siendo ya adecuadas las herramientas de la econometría tradicional, dado que los contrastes habituales de heteroscedasticidad pueden estar sesgados en un contexto espacial (Yrigoyen 2004, p. 2).

3. En tercer lugar, en un modelo de regresión de corte transversal, ambos efectos de autocorrelación y heterogeneidad espacial pueden ser, desde una óptica meramente observacional, totalmente equivalentes. Así, por ejemplo, un "clúster" o agrupamiento espacial (observado en localizaciones muy próximas) de los residuos con valores extremos podría ser interpretado como un problema de heterogeneidad espacial (heteroscedasticidad de grupos o "groupwise"), o también como un efecto de autocorrelación espacial. Por eso, deben estructurarse perfectamente ambos efectos espaciales para identificar correctamente

los parámetros de un modelo con estos problemas y nunca considerar un aspecto independientemente del otro (Yrigoyen 2004, p. 2).

- Autocorrelación espacial en el modelo de regresión

El modelo de regresión lineal (múltiple) tiene como forma funcional:

$$Y = \sum_{q=1}^{Q} X_q \beta_q + \varepsilon$$

y su versión muestral

$$y_i = \sum_{q=1}^{Q} x_{iq} \beta_q + \varepsilon_i \qquad i = 1, \ldots, n$$

donde y_i es una observación de la variable dependiente o de interés, x_{iq} es una observación en una variable explicativa con $q = 1, \ldots, Q$, β_q es el coeficiente de regresión que mide la influencia por si sola de la q-variable explicativa en la variable dependiente, es decir, mide el cambio en Y por cada cambio unitario en X_q manteniendo las restantes variables explicativas constantes. El término ε_i es el error aleatorio, que puede ser debido a variables no controladas y a la variabilidad muestral. Para estos términos de error se asume que ε_i son variables independientes e idénticamente distribuidas a una variable normal con media cero y varianza una constante σ^2, esto es, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, donde E $(\varepsilon_i \cdot \varepsilon_i)$ = E (ε_i)·E (ε_i)= 0. En notación matricial este modelo de regresión lineal puede ser escrito de la siguiente forma: (Borrego, 2018, p. 36).

$$Y = X\beta + \varepsilon$$

donde Y es un vector $n \times 1$ en el que se encuentran las n observaciones de la variable dependiente, X es una matriz $n \times Q$ que muestra las observaciones de las variables explicativas, β es el vector $Q \times 1$ de parámetros de regresión asociados a dichas variables explicativas y el vector ε de dimensión $n \times 1$ de términos de error. Este modelo es válido para entender la relación funcional entre la variable dependiente y las variables explicativas y estudiar cuáles pueden ser las causas de la variación de Y . Para ello, a partir de la información muestral, se obtiene un estimador de β $(\hat{\beta})$ para así obtener una predicción de Y a partir de las variables explicativas $X_1 \cdot \cdot \cdot, X_Q$ (Borrego, 2018, p. 37).

Bajo la hipótesis de que las observaciones son independientes, el modelo se simplifica notablemente. Sin embargo, a la hora de estudiar los datos espaciales, esta simplificación puede desembocar a unos resultados parciales inconsistentes debido a la dependencia espacial. Esta dependencia espacial puede estar presente en variables explicativas, variable dependiente o en los residuos (términos de error). Cuando la dependencia espacial se encuentra en la variable dependiente los modelos se denominan modelos de retardo espacial mientras que si está en los residuos se denominan modelos de error espacial. Cuando está presente en las variables explicativas se llaman modelos de regresión cruzada o modelos X espacialmente retardados, pero, en contraste con los otros dos modelos, no precisan de procedimientos especiales para la estimación (Borrego, 2018, p. 37).

2.2.9. Autocorrelación mediante el método de Moran

La herramienta auto correlación espacial (I de Moran global) mide, simultáneamente, la auto correlación espacial basada en las ubicaciones y los valores de las entidades. Dado un conjunto de entidades y un atributo asociado, evalúa si el patrón expresado está agrupado, disperso o es aleatorio. La herramienta calcula el valor del índice I de Moran y una puntuación z. El estadístico de prueba I de Moran para contrastar la autocorrelación espacial es el estimador de la pendiente de la regresión por mínimos cuadrados ordinarios. Para construir el diagrama de dispersión de Moran de una variable específica, es necesario rezagar espacialmente la variable en cada observación; este proceso, consiste en calcular un parámetro w y multiplicarlo por una observación i de la variable en cuestión, donde el parámetro w se obtiene al promediar los valores de la variable vecinos a i, en el orden de contigüidad especificado. Quedando W_i (Corso Sicilia, Rivera 2017, p. 101).

$$Y_i * W_i$$

$$W_i = (Y_i + Y_k + Y_l + \ldots + Y_n)/n$$

Donde $Y_i, Y_k, Y_l, \ldots, Y_n$ son las zonas contiguas a la región Y_i. Los valores rezagados de la variable se ubican en el eje Y, los valores normales de la variable se ponen en el eje X.

2.2.10. Métodos de interpolación espacial

La estadística aplicada a datos geoespaciales o geoestadística es una técnica estadística usada para la estimación, predicción y simulación de datos correlacionados espacialmente, que se ha conocido como el arte de modelar datos espaciales. Su importancia radica en que permite describir la continuidad espacial de las variables y estimar valores muy cercanos a los reales en puntos desconocidos.

La geoestadística tuvo su origen en procesos de búsqueda y exploración de minerales, es por esto la asignación del prefijo GEO para referirse a ciencias de la tierra. A lo largo de su evolución se han identificado cuatro generaciones (Sotter et al. 2002, p. 31).

- **Lineal:** Dedicado a la teoría de funciones aleatorias.
- **No Lineal:** Dedicada a la aplicación minera y a la gran difusión de la ciencia.
- **Tercera generación:** Dedicado al desarrollo de diferentes tipos de Kriging.
- **Cuarta generación:** La cual utiliza algoritmos geoestadísticos a través de herramientas computacionales.

2.2.10.1. Características de interpolación espacial

La interpolación tiene como objetivo estimar, a partir de una muestra, valores de Z para un set de puntos (X, Y). La interpolación puede utilizarse para cumplir tres funciones (Fallas 2007, p. 4):

- Estimar valores de Z para ubicaciones particulares (X, Y);
- Estimar valores de Z para una cuadrícula rectangular.
- Cambiar la resolución de la cuadrícula en un archivo ráster (método conocido como remuestreo).

La clasificación de los métodos de interpolación puede realizarse en base a múltiples criterios (Fallas 2007, p. 4).

- Globales o locales, según si utilizan todos los valores del área evaluada o sólo una parte de ella (subconjunto).
- Graduales o abruptas, según la continuidad y suavidad de la superficie resultante.
- Exactos o aproximados, según si respetan los valores de mediciones exactas de entrada para la interpolación o si, por el contrario, pueden ser alterados o suavizados para ajustarlos al modelo del conjunto.
- Univariantes o multivariantes, según si admiten o no valores de múltiples variables de entrada para generar el modelo y la superficie de interpolación. En GIS, generalmente la distancia es la variable admitida para métodos de interpolación univariantes.
- Determinísticos o estocásticos, según si incorporan o no variaciones aleatorias (incertidumbre) en la superficie interpolada. Los métodos determinísticos son aplicables cuando hay mediciones suficientes para describirla matemáticamente, mientras que los estocásticos incorporan el concepto de aleatoriedad por una insuficiencia de ellas.

2.2.10.2. Vecino más cercano (nearest neighbor)

La interpolación del vecino más cercano se basa en la construcción del polígono de Voronoi. Los polígonos de Voronoi son el método de interpolación vectorial más sencillo y sencillo. Este método se basa únicamente en la distancia euclidiana, ignorando cualquier tipo de valores asignados a los puntos de muestreo (Estévez, 2019, p.31).
Este método de interpolación divide el espacio en zonas de equivalencia o regiones de influencia para cada punto de medición de entrada. Los polígonos de Voronoi o Thiessen están definidos por los límites del punto más cercano. El perímetro de cada área se crea equidistante de todos los puntos adyacentes (Estévez, 2019, p.31).

Finalmente, el método asigna a cada polígono el valor del punto que contiene ya partir del cual se generó. Dado que este es un método de solo distancia, las variables interpoladas pueden ser cualitativas o cuantitativas (Estévez, 2019, p.31).

2.2.10.3. TIN (Triangulated Irregular Network)

Este método de interpolación devuelve una superficie de triángulos formada a partir de la localización de una serie de vértices cuyos valores son conocidos. Los vértices se conectan mediante aristas para generar dicha red triangular (Estévez, 2019, p.32).

El resultado obtenido, la superficie TIN, es una malla o red de triángulos interconectados, donde cada uno de ellos representa una zona homogénea en lo que a la variable estudiada se refiere. El método TIN tratará, por tanto, de generar un conjunto de triángulos sobre el espacio que maximicen la relación área/perímetro (Estévez, 2019, p.32).

Es muy habitual su uso sobre todo para modelos del terreno en base a mediciones de elevación conocida, aunque puede aplicarse a otras mediciones cuantitativas de distintas variables ambientales (Estévez, 2019, p.32).

2.2.10.4. Interpolación IDW (Inverse Distance Weighting)

Mediante el método de interpolación IDW los puntos de muestreo se ponderan durante la interpolación. De esta manera, la influencia de un punto en relación con otros se reduce o disminuye a medida que aumenta la distancia entre ellos (Estévez, 2019, p.33).

En el método de interpolación IDW puede establecerse un valor de potencia, denominado coeficiente P de distancia que por defecto es 2. A mayor valor de P, mayor énfasis o peso asignado a los puntos cercanos a evaluar, resultando en una superficie estadística más abrupta. A menor valor de P, mayor énfasis en el conjunto de la muestra de valores, dando como resultado superficies más suavizadas (Estévez, 2019, p.33).

Generalmente se utiliza en procesos de interpolación donde el conjunto de datos disponible para la interpolación es abundante, se reparte homogéneamente por el espacio y no existen grandes distancias entre sus localizaciones.

Su fórmula:

$$\hat{Z}_j = \sum_{i=1}^{N} k_{ij} * S_j$$

2.2.10.5. Interpolación mediante spline

La herramienta Spline utiliza un método de interpolación que estima valores usando una función polinómica que minimiza la curvatura general de la superficie, lo que resulta en una superficie suave que pasa exactamente por los puntos de entrada. Este método, juntamente con Kriging, es uno de los métodos exactos de interpolación existentes que no admite aproximaciones o suavizados de los valores de entrada. Matemáticamente, la herramienta Spline emplea funciones polinómicas distintas más acordes para cada tramo, adecuándose así a una superficie más suave, menos abrupta y uniforme (Estévez, 2019, p.34).

2.2.11. Componentes principales

El ACP es un método algebraico estadístico que trata de sintetizar y dar una estructura a la información contenida en una matriz de datos. El procedimiento consiste en homologar dicha matriz a un espacio vectorial tratando de encontrar en él unos ejes o dimensiones que, siendo combinación lineal de las variables introducidas (Lozares y López 1991, p. 33).

- No pierdan la información inicial al conservar la varianza total.
- No tengan correlación entre ellos, esto es, sean linealmente independientes, lo que asegura la estructuración de las variables iniciales.
- Tengan una importancia diferencial y conocida en la explicación de la varianza total.

Realizadas estas exigencias, el objetivo básico consiste en reducir el número de variables introducidas. Para ello se toman como nuevas variables los ejes o componentes hallados, eligiendo un número y peso de los mismos suficiente para que la pérdida de varianza total sea la conveniente, llenando así las finalidades del método, esto es, las de simplificar, reducir y estructurar la información inicial (Lozares y López 1991, p. 33).

En la definición de las componentes Y interviene el conjunto de las variables X contribuyendo cada una en su totalidad, sin diferenciar en cada variable una parte común y otra específica que no intervenga en la creación de los ejes. Idénticamente para dar cuenta de cada variable no se supone la existencia de una parte de esta explicada por la comunalidad conjunta de las variables y otra parte inexplicada por ella, o sea especifica de la variable. Todas las variables contribuyen a dar cuenta de todas (Lozares y López 1991, p. 36).

- Explicación geométrica del PCA

Antes de aplicar un PCA, las observaciones tienen que moverse al centro del eje de coordenadas, esto es, centrarlas para que tengan media 0, para así eliminar posibles vías en las mediciones. Los datos también se escalan a una varianza unitaria para eliminar el efecto de las distintas unidades en las que puedan estar medidas los datos. Se traza la línea que mejor se ajusta a los datos centrados y escalados (menor error residual), que explicará mejor dichos datos cuanta más correlación exista entre ellos. Esta línea, que será la primera variable latente o primera componente principal (Z_i), pasará en la dirección de máxima varianza de las proyecciones de las observaciones sobre dicha línea (Martinez, 2018, p.32).

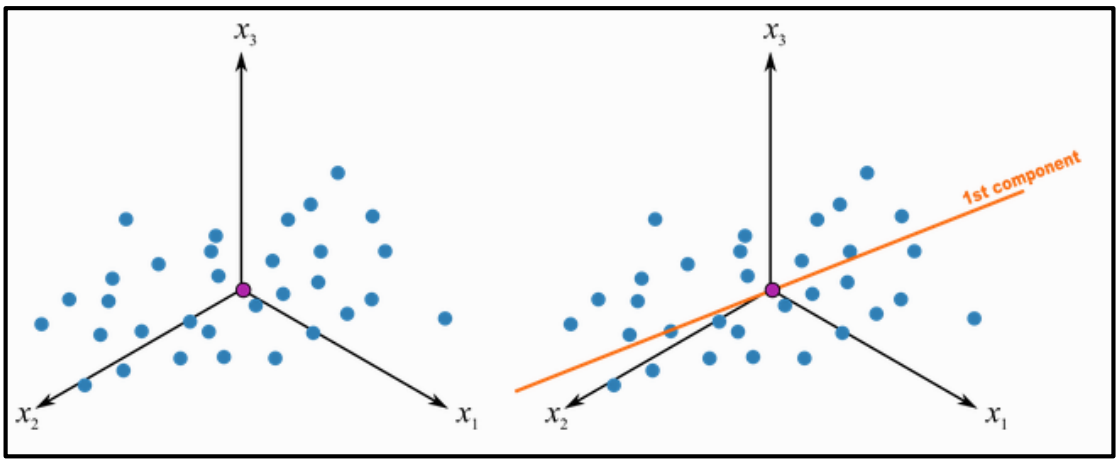


Ilustración 4-2: Componentes principales

Fuente: (Martínez, 2018)

2.2.12. Análisis de clúster

E1 Análisis de Clúster (AC, en adelante) es el nombre genérico otorgado a una gran variedad de técnicas que tienen como objetivo primordial la búsqueda de grupos en un conjunto de individuos. En líneas generales, todo método de clasificación parte de un conjunto de elementos singulares que deben ser clasificados en un número reducido de grupos o clúster, obtenidos por particiones sucesivas del conjunto original y en los que se respete la estructura relacional que en el mismo se mantenga. Las leyes matemáticas por las que se rigen estos métodos reciben el nombre de Taxonomía Numérica. Este concepto, algo confuso, puede quedar aclarado al delimitar las propiedades de los clústeres (Santana, 1991, p. 66).

- Densidad: esta primera propiedad define un clúster como un conglomerado espacial de puntos relativamente compacto en comparación con otras áreas de ese espacio que tienen menos o ningún punto.
- Varianza: grado de dispersión de los puntos de cada conglomerado en el espacio.
- Forma: configuración espacial de los puntos (redondeada-hiperesfera alargada, etc.).
- Separación: grado de solapamiento o de separación entre los clústeres.

En los últimos años, el AC ha ido ganando popularidad, en parte porque su objetivo es muy apetecido y fácil de entender, y en parte porque siempre se consigue un resultado interpretable. Es innegable la enorme importancia que tiene el descubrimiento de tipologías en el ámbito de las ciencias sociales y, especialmente, en el marketing. El problema principal de su utilización radica en el amplio espectro de controversias relativas a sus modos de aplicación. En efecto, los postulados teóricos y metodológicos sobre los que se asienta el AC no están tan sólidamente fundamentados como, por ejemplo, los del Análisis Factorial. Existen diferentes tipos de algoritmos de clúster explicados a continuación (Santana, 1991, p. 66).

- Algoritmo para la formación de grupos espaciales homogéneos

En los algoritmos de clustering tradicionales (jerárquicos o no), cuando se agrupan unidades geográficas, los grupos homogéneos creados no necesariamente están formados por ciudades estrictamente vecinas (Carvalho et al. 2009, p. 416).

- Algoritmos de agrupamiento jerárquico espacial

Dada la gran aplicabilidad de los métodos de agrupamiento, la literatura en el área ha evolucionado notablemente, por lo que a medida que se crean más algoritmos, brindan mayor

eficiencia y más idoneidad para diferentes situaciones. Hastie, Tibishirani y Friedman (2001), Jain y Dubes (1988), Jain et al. (1999), discuten estos algoritmos. Los algoritmos de agrupamiento se pueden dividir en tres categorías principales (Carvalho et al. 2009, p. 416).

- Algoritmos combinatorios.
- Modelos mixtos.
- Búsqueda de modo

Las últimas dos categorías se basan en alguna forma de modelos probabilísticos para el proceso de generación de datos. Los algoritmos combinatorios, por otro lado, se basan en reglas heurísticas para buscar las mejores agrupaciones, tratando de minimizar algunos criterios de variabilidad general. Los siguientes pasos describen la modificación del algoritmo jerárquico, para satisfacer la restricción de vecindad entre las unidades de cada grupo homogéneo (Carvalho et al. 2009, p. 417).

- Average linkage (vinculación media)

De acuerdo con el método de vinculación media, también conocido como método de McQuitty, la distancia entre dos conglomerados es la distancia media entre todos los pares de vectores variables extraídos de los dos conglomerados. Para el agrupamiento de aglomeraciones no espaciales, este método tiende a unir clústeres con baja varianza y está ligeramente sesgado hacia la producción de clústeres con la misma varianza. La expresión para la medida de la distancia entre los conglomerados K y L viene dada por (Carvalho et al. 2009, p. 418).

$$D_{K,L} = \frac{1}{N_K N_L} \sum_{i=C_K} \sum_{i=C_L} d(x_i, x_j)$$

Donde $d(x_i, x_j)$ (que contienen las características de los polígonos x_i y x_j. Otra forma de implementar el algoritmo de agrupamiento jerárquico es a través de la actualización de la matriz de proximidad (o distancia) entre cada vez que se crea un nuevo conglomerado CM mediante la unión de los conglomerados C_L y C_K existentes, se actualiza la matriz de proximidades para considerar las distancias al nuevo conglomerado, esta actualización se puede realizar directamente a través de las distancias de la matriz anterior, utilizando fórmulas combinatorias. Considere un grupo C_J. Para el método de enlace promedio, la distancia entre cualquier conglomerado C_J y el nuevo conglomerado C_M se puede obtener a partir de las distancias anteriores, usando la siguiente expresión (Carvalho et al. 2009, p. 419).

$$D_{J,M} = \frac{1}{2} D_{J,K} + \frac{1}{2} D_{J,L}$$

El método de enlace promedio considera un promedio de todos los miembros de los conglomerados cuya distancia se está calculando. Por lo tanto, este método está menos influenciado por los valores extremos, como es el caso de los métodos de enlace simple y enlace completo (Carvalho et al. 2009, p. 419).

- Single linkage (Método del vecino más cercano)

De acuerdo con el método de vinculación promedio, también conocido como método de McQuitty, la distancia entre dos grupos es la distancia promedio entre todos los pares de vectores variables extraídos de los dos grupos. Para el agrupamiento de aglomeraciones no espaciales, este método tiende a unir clústeres con baja varianza y está ligeramente sesgado hacia la producción de clústeres con la misma varianza. La expresión para la medida de la distancia entre los conglomerados K y L viene dada por (Carvalho et al. 2009, p. 419).

$$D_{K,L} = min_{i=C_K j=C_L}\, d(x_i, x_j)$$

La medida de distancia entre un grupo C_J y el nuevo grupo C_M se puede actualizar mediante la fórmula combinatoria:

$$D_{J,M} = \left|\frac{1}{2}D_{J,K} + \frac{1}{2}D_{J,I} - \frac{1}{2}\left|D_{J,K} - D_{J,I}\right|\right|$$

Dado que no impone restricciones en la forma de los racimos, este método sacrifica la posibilidad de obtener grupos compactos, con la ventaja de permitir no sean irregulares o alargados. En la agrupación de aglomeraciones no espaciales, el método de enlace único también tiende a cortar las colas de las distribuciones antes de separar las agrupaciones principales (Carvalho et al. 2009, p. 419).

- Complete linkage method (método de enlace completo)

Según el método de enlace completo, la distancia entre dos conglomerados es el máximo de las distancias entre todos los pares de vectores variables extraídos de los dos conglomerados (Carvalho et al. 2009, p. 420).

$$D_{K,L} = max_{i=C_K j=C_L}\, d(x_i, x_j)$$

La medida de la distancia entre un grupo C_J y el nuevo grupo C_M se puede obtener mediante la fórmula combinatoria:

$$D_{J,M} = \frac{1}{2}D_{J,K} + \frac{1}{2}D_{J,I} + \frac{1}{2}\left|D_{J,K} - D_{J,I}\right|$$

Está fuertemente sesgado hacia la producción de cúmulos compactos con diámetros similares y puede distorsionarse severamente por valores atípicos moderados. Es un método que asegura que todos los elementos de un grupo estén a una distancia mínima entre sí (Carvalho et al. 2009, p. 420).

- Ward's Minimal Variance Method (Método varianza mínima de Ward)

El método de varianza mínima de Ward está sesgado hacia la generación de conglomerados del mismo tamaño. El método se basa en la suma de los cuadrados de los errores (SSE) de cada conglomerado (suma de los cuadrados de las desviaciones del centroide del conglomerado). Se agrega los SSE de todos los clústeres G, generando el TSSE. El método consiste en analizar todos los posibles pares de conglomerados unidos, identificando qué conjunto produce el menor incremento en SSE. En este método, la distancia entre dos grupos es la suma de cuadrados ANOVA entre dos grupos para todas las variables. En cada generación, minimiza la suma de cuadrados intra-cluster que se puede obtener mediante la unión de dos clústers. Con frecuencia, se recomienda utilizar la relación SQE/SQET sobre la SQE absoluta. La medida de distancia entre los conglomerados C_K y C_L se define como (Carvalho et al. 2009, p. 421).

$$D_{K,L} = \frac{d(\overline{x_k}, \overline{x_L})^2}{\left(\frac{1}{N_K} + \frac{1}{N_L}\right)}$$

donde $\overline{x_k}$ y $\overline{x_L}$ son los vectores medios dentro de los grupos C_K y C_L respectivamente. Es un método que tiene como objetivo maximizar la verosimilitud en cada nivel de jerarquía bajo las hipótesis de mezcla de distribuciones normales multivariadas, matrices esféricas de igual covarianza e iguales probabilidades muestrales. En la agrupación de aglomeración no espacial, tiende a unir grupos con un número pequeño de observaciones y está fuertemente sesgado hacia la producción de grupos de la misma forma y número de observaciones. También está influenciado por valores atípicos (Carvalho et al. 2009, p. 421).

- Centroid method (método centroide)

Desarrollado por Sokal y Michener en 1958, el método de enlace centroide considera la medida de la distancia entre dos conglomerados como el cuadrado de la distancia entre los centroides de los conglomerados (Carvalho et al. 2009, p. 422).

$$D_{K,L} = d(\overline{x_k}, \overline{x_L})^2$$

Al ser una comparación de promedios, los valores atípicos tienen poca influencia. En otros aspectos, puede ser menos eficiente que otros métodos como el enlace medio y Ward. El mayor de dos grupos unidos tiende a dominar el nuevo grupo (Carvalho et al. 2009, p. 422).

- Average linkage weigthed (enlace medio ponderado)

El método de vinculación promedio ponderado difiere del método de vinculación promedio original debido a los diferentes pesos insertados en la fórmula combinatoria. La nueva fórmula es (Carvalho et al. 2009, p. 422).

$$D_{J,M} = \frac{n_K}{n_L + n_K} D_{J,K} + \frac{n_L}{n_L + n_K} D_{J,L}$$

donde n_K y n_L son los números de observaciones en los conglomerados C_K y C_L respectivamente.

- Median Method (Método de la mediana)

El método de la mediana tiene la fórmula combinatoria para actualizar la matriz de distancia dada por (Carvalho et al. 2009, p. 422).

$$D_{J,M} = \frac{1}{2} D_{J,K} + \frac{1}{2} D_{J,L} - \frac{1}{4} D_{K,L}$$

Este método fue desarrollado por Gower (1967, p.34).

- Método Elbow

En este método se tiene que calcular la distorsión promedia del clúster, que es la distancia promedia del centroide a todos los puntos del clúster y se obtiene con el algoritmo *K-Means* en función del número de clúster. Así, cuando se va de una situación en la que el número de *clúster* es inferior al correcto a una situación en la que el número es el adecuado, el valor de la dispersión disminuye bruscamente, mientras que, si aumenta el número de *clúster* al adecuado, el valor de la dispersión se reducirá más lentamente, formando un codo en la figura 5-2 (Pulido, 2019, p.46).

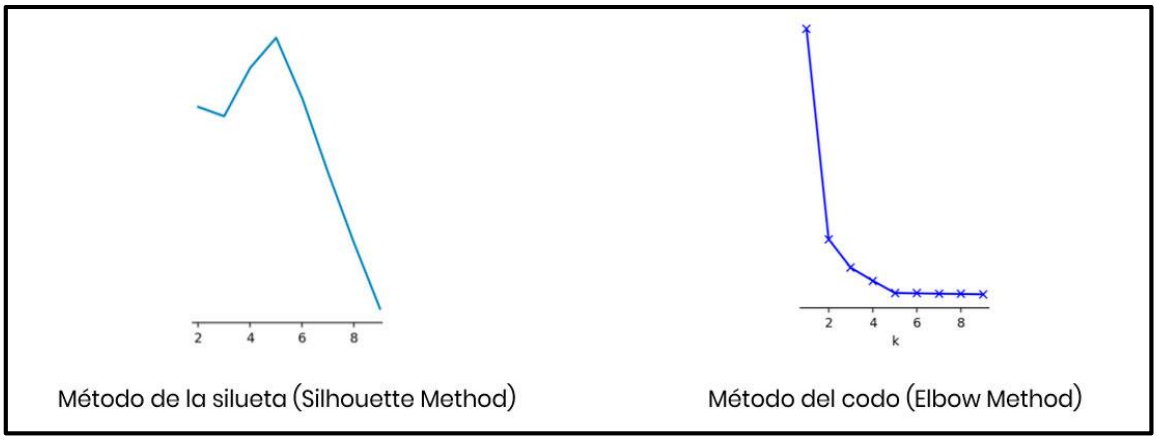


Ilustración 5-2: Métodos de análisis de clúster.

Fuente: (Pulido, 2019)

El método Elbow o codo ayuda a elegir el número óptimo de clúster, se buscó hacer clasificación en un conjunto de datos. Para hacer uso de este método se parte del cálculo de la distorsión promedio de cada clúster, esto es la distancia de cada elemento con su centroide correspondiente. Para el calcular la distorsión se usó:

$$distorsión = \frac{\sum_i^N \|x_i - centroide\|^2}{N}$$

Dónde: N es el número de elementos

- Average silhouette method (método de silueta promedio)

El método de la silueta promedio (Average Silhouette Method) tiene el mismo propósito que el método del codo propuesto por Liu y Sarkar. Se basa en calcular la silueta promedio de las observaciones para diferentes valores. De lo contrario, la diferencia entre las distancias que tiene un objeto a otros objetos en el mismo grupo y la distancia que tiene en otros grupos. Para un objeto en una colección, el ancho de la silueta (SW) se define mediante la siguiente ecuación (Abdelilah ,et. al 2020, p. 5).

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

El SW se considera como un indicador de rendimiento que varía de -1 a +1, donde el número óptimo de clústeres corresponde al valor más alto. El diagrama de flujo del método del codo y el procedimiento de la silueta promedio se ilustran en la Figura 5-2 (Abdelilah ,et. al 2020, p. 5).

2.2.12.1. Validación de clúster

Existen varios métodos para validación de clúster después de decidir cuál es la mejor partición para los datos iniciales, queda la tarea de validar esta decisión. Esta labor puede satisfacerse por alguno de estos cinco criterios:

- Coeficiente de correlación cophenttico
- Coeficiente de pertenencia
- Replicación
- Simulaciones Monte Carlo.
- Interpretabilidad teórico-práctica

- Coeficiente de correlación cophenttico

Este coeficiente es válido solo para los métodos jerárquicos aglomerativos. Representa una medida de ajuste entre los datos de partida y la estructura del dendrograma, esto es, el grado en que la partición reproduce el armazón de distancias entre los individuos. En líneas generales, se suele decir que una buena partición debe tener una correlación cophenttica (coeficiente de correlación) (Santana 1991, p. 74).

- Coeficiente de pertenencia

Esta es una medida para ver cuán diferentes son los conglomerados, en función de los ítems contenidos en cada uno de ellos. A tal fin se calcula un cociente cuyo numerador expresa la media de la inter-correlación entre los sujetos dentro de un mismo clúster, y el denominador es igual a la media de la inter-correlación de pares de ítems en donde una añadidura en cada par pertenece al grupo de interés. Si el conglomerado está bien elegido, el numerador será superior a la unidad, sugiriéndose que para un valor del coeficiente igual o superior a 1.3, se puede considerar que un clúster ha sido identificado. Con el manejo de estos coeficientes (también llamados (coeficientes B)) no es necesario recurrir a la visualización gráfica de los resultados del análisis, ya que los factores B elevados representarán los grupos más significativos (Santana, 1991, p. 74).

- Replicación

La táctica de la replicación consiste en repetir el AC para diferentes submuestras (dos o más) de la población total, a fin de ver si las particiones resultantes mantienen un cierto nivel de consistencia interna. El fallo de este criterio es que solo cuando los resultados son muy diferentes

cabe sospechar algún problema en la partición original: su (éxito) no es un indicador claro de que la solución clúster establecida sea la más apropiada (Santana, 1991, p. 74).

- Simulaciones Monte Carlo

Esta aproximación es raramente utilizada a causa de su costo y su complejidad. Los procedimientos Monte Carlo se basan en generadores de números aleatorios (*Random Number Generato*), para crear una nueva matriz de datos cuyas características generales queden apareadas con las características generales de la matriz original de datos, pero sin contener ningún clúster. Entonces se lleva a cabo un AC idéntico al realizado con los datos originales y se comparan los resultados de los dos AC. Esta comparación, que puede fundamentarse, por ejemplo, en realizar Análisis de Varianza para cada solución, tendrá como fruto la validación o invalidación de la partición analizada (Santana, 1991, p. 75).

- Interpretabilidad teórico-práctica

En último término, todos los criterios de selección de la partición idónea y de su posterior validación carecen de sentido si la clasificación no tiene sentido teórico o es difícilmente interpretable. Es por ello se considera este criterio como el que, en última instancia, supera a todos los anteriores (Santana, 1991, p. 76).

- Dendrograma

El dendrograma, gráfico más representativo de este tipo de análisis, asume la forma de un árbol de clasificación en el que es posible observar con toda claridad la forma y el número de los grupos que se van formando. En este gráfico es el eje de ordenadas el que adquiere verdadero protagonismo pues representa los distintos niveles de similaridad en torno al cual se han ido agrupando las unidades de análisis en función de la medida elegida. Por su parte, en el eje de abscisas únicamente se identifican los casos u observaciones. El problema de esta representación gráfica es que sólo se puede emplear cuando el número de casos es reducido ($n<200$). Constituye un resumen de la información original presente en la matriz de distancias o similaridades y la información que presenta será más útil cuanto más agrupado sean los datos que represente (Rodríguez y Mora, 2001, p. 148).

- Medidas de similitud

Uno de los factores que ocasionan mayores divergencias en los resultados es la elección de la medida de similitud (destinada a cuantificar la separación entre las unidades de análisis). Las más utilizadas son las medidas de distancia, a pesar de que también existen los coeficientes de correlación, las medidas de asociación y los coeficientes de similitud probabilística. Uno de los aspectos que en última instancia son más importantes en este tipo de decisiones es la disponibilidad de un paquete de programas estadísticos u otro (Santana 1991, p. 68). Dados dos vectores x_i, x_j pertenecientes a $\mathbb{R}^k$, se ha establecido una distancia entre ellos si se define una función d con las propiedades siguientes:

- $d : \mathbb{R}^k \, x \, \mathbb{R}^k \longrightarrow \mathbb{R}^+$, es decir $d\left(x_i , x_j\right) \geq 0$
- $d\left(x_i , x_i\right) = 0 \; \forall i$, la distancia entre un elemento y sí mismo es cero
- $d\left(x_i , x_j\right) = d\left(x_j , x_i\right)$, la distancia es simétrica
- $d\left(x_i , x_j\right) \leq d\left(x_i , x_p\right) + d\left(x_p , x_j\right)$, la distancia verifica la propiedad triangular

Estas propiedades generalizan la noción intuitiva de distancia euclídea entre dos puntos

- Espacios métricos

Se introduce ahora la notación básica para el problema de satisfacer consultas por proximidad. El conjunto U denotara el universo de objetos válidos. Un subconjunto finito de él, S, de tamaño $n = |S|$, es el conjunto de objetos donde se busca S será llamado el diccionario, base de datos o simplemente nuestro conjunto de objetos o elementos (Reyes, 2002, p. 8).
La función:

$$d: U \, x \, U \rightarrow \mathbb{R}$$

denotará una medida de "distancia" entre objetos (es decir, mientras más pequeña es la distancia, más cercanos o similares son los objetos). Las funciones de distancia tienen las siguientes propiedades:

- Positividad
- Simetría
- Reflexividad

Las propiedades enumeradas de la función de similaridad soló aseguran su definición consistente y no pueden ser usadas para ahorrarse comparaciones en una consulta por proximidad. Si es en

verdad una métrica, es decir si satisface la desigualdad triangular, entonces el par (U_1, d) se denomina un espacio métrico (Reyes 2002, p. 8).

2.2.13. Software libre R

R es un software libre que permite realizar análisis estadísticos y el más usado en la comunidad científica. Este programa está disponible en la página web: http://www-r-project.org y consta de una aplicación central y de librerías de multitud de temas que se pueden instalar según necesidad. R es un programa de instrucciones, y por tanto, no resulta del todo "amigable" para los usuarios que no están acostumbrados a este tipo de manejo (Botella, et. al 2018, p. 3).

2.2.14. **Librerías para clúster**

2.2.14.1. Library(factoextra)

- Factoextra: es un paquete de R que facilita la extracción y visualización de los resultados de los análisis exploratorios de datos multivariados Kassambara (2020), que incluyen:
- Análisis de componentes principales (PCA), que se utiliza para resumir la información contenida en datos multivariados continuos (es decir, cuantitativos) al reducir la dimensionalidad de los datos sin perder información importante.
- Análisis de correspondencias (CA), que es una extensión del análisis de componentes principales adecuado para analizar una gran tabla de contingencia formada por dos variables cualitativas (o datos categóricos).
- Análisis de Correspondencia Múltiple (MCA), que es una adaptación de CA a una tabla de datos que contiene más de dos variables categóricas.
- Análisis de factores múltiples (MFA), dedicado a conjuntos de datos donde las variables se organizan en grupos (variables cualitativas y/o cuantitativas).
- Análisis jerárquico de factores múltiples (HMFA), es una extensión de MFA en una situación en la que los datos se organizan en una estructura jerárquica.
- Análisis factorial de datos mixtos (FAMD), un caso particular del MFA, dedicado a analizar un conjunto de datos que contiene variables tanto cuantitativas como cualitativas.
- Hay varios paquetes de R que implementan métodos de componentes principales. Estos paquetes incluyen: FactoMineR, ade4, stats, ca, MASS y ExPosition.
- Sin embargo, el resultado se presenta de forma diferente según los paquetes utilizados.
- FactoMineR es un paquete elegante creado por Sebastián Le en el 2008 que permite calcular PCA, MCA, FAMD y MFA.

CAPÍTULO III

3. MARCO METODOLÓGICO

3.1. Enfoque de investigación

El método es cuantitativo, ya que, la información se concentra en la variable "dosis de rayos gamma" de la provincia de Chimborazo, con el objetivo de caracterizar dicha variable e identificar clústeres espaciales. Los resultados obtenidos proporcionan información de la variable estudiada en la provincia en lugares específicos a través técnicas estadísticas y geoestadísticas.

3.2. Nivel de investigación

El nivel de investigación es de tipo exploratorio, ya que, la información se concentra en la variable estadística rayos gamma de la provincia de Chimborazo concentrada en los 10 cantones, siendo Riobamba el más grande.
Los clústeres espaciales proporcionan una inferencia inductiva, de las dosis de astro partículas monitoreadas por el GIDAC.

3.3. Diseño de la investigación

Es una investigación no experimental ya que, la información se concentra en la variable estadística cuantitativa "dosis de rayos gamma" de la provincia de Chimborazo, que no permite manipular las variables y obsérvala en su ambiente natural, buscando los factores que influyen en la localización de las dosis de rayos gamma según la demografía de los mismo. De acuerdo con el periodo de estudio es de tipo transversal, ya que la información analizada está tomada en un periodo de tiempo determinado. (Berger et al. 2018, p.34).

3.4. Tipo de estudio

El tipo de estudio es de carácter cuantitativo, ya que los datos analizados son tasas de dosis de rayos gamma, con el propósito de caracterizar su comportamiento en la provincia de Chimborazo.

3.4.1. Población y planificación, selección y cálculo del tamaño de la muestra

La población son las dosis de rayos gamma presentes en la provincia de Chimborazo, el proceso de muestreo fue realizado por el GIDAC mediante un muestreo aleatorio simple en cada cantón de cual se obtuvo 407 localizaciones en la cual se midió la variable en estudio.

3.4.2. Métodos, técnicas e instrumentos de investigación

El GIDAC realizo la medición de las dosis de rayos gamma mediante el uso de espectrómetro modelo AT6102 series. La recolección de información secundaria se realizó mediante entrevistas a los delegados y coordinadores del proyecto de "Evaluación de elementos radioactivos de la serie de Uranio 238 en el ambiente de pacientes de cáncer". Los datos fueron facilitados mediante una hoja de Excel. Se realizo el respectivo análisis exploratorio de datos proporcionados. Para ello de utilizo estadísticos descriptivos, identificación de datos atípicos, anovas, correlación espacial, componentes principales, cubo de homogeneidad, estandarización de los datos, método silueta y clústeres espaciales.

3.4.3. Mapa de Chimborazo

Ilustración 1-3: Provincia de Chimborazo.

Fuente: (R-STUDIO, 2022)

La provincia de Chimborazo se encuentra ubicada en el centro sur del país con respectiva capital cabecera la ciudad de Riobamba, sus limitaciones son al norte se encuentra la provincia de Tungurahua, al sur la provincia del Cañar, al este la provincia de Morona Santiago y al oeste la provincia de Bolívar.

3.4.4. Población de estudio

La población de estudio son las dosis de rayos gamma que ingresan en la capa de ozono en la provincia de Chimborazo de acuerdo con la distancia en la que ese encuentra con respecto al sol.

3.4.5. Método de muestreo

El método de muestreo usado por el grupo de investigación GIDAG fue aleatorio simple en las zonas rurales y sistemático en las zonas urbanas, en los diferentes cantones de la provincia de Chimborazo.

3.4.6. Tamaño de la muestra

El GIDAC midió las tasas de dosis de rayos gamma en 407 ubicaciones geográficas en la provincia de Chimborazo.

3.4.7. Técnica de recolección de datos

La técnica de recolección primaria se realizará mediante el espectrómetro de rayos gamma dentro del proyecto de “Evaluación de elementos radioactivos de la serie de Uranio 238 en el ambiente de pacientes de cáncer”.
Como técnica secundaría mediante la revisión de los datos registrados en un archivo Excel en cual fue por el GIDAC.

3.4.8. Identificación de variables

La variable de estudio principal son las dosis de rayos gamma medidos en una ubicación geográfica especifica (latitud y longitud), también se tomó en cuenta el tiempo (mañana y tarde) y los sectores (urbano y rural).

3.4.9. Modelo estadístico

Dentro del análisis estadístico se realiza como primera instancia el análisis exploratorio de las variables de rayos gamma en la provincia de Chimborazo, así como también de las posiciones geográficas en la zona de la investigación, después se analizará la correlación espacial de los rayos gama, mediante el índice de moran, para posteriormente analizar los datos mediante clúster e identificar la formación de clúster más idónea que refleje dicha variable y finalmente se realizará al menos un mapeo de una primera aproximación del comportamiento de la tdrg en la zona de estudio. El análisis de datos se realizará mediante el software libre R-Studio.

CAPÍTULO IV

4. MARCO DE ANÁLISIS Y DISCUSIÓN DE RESULTADOS

El país está distribuido en las 24 provincias, siendo Chimborazo la novena con un 2.69% de la población total del país. Con las mediciones de tasas de dosis rayos (tdrg)gama proporcionada por el GIDAC para la presente investigación, se consideró un análisis exploratorio (tradicional y espacial), y un clúster jerárquico aglomerativo. La matriz de información consta de 407 observaciones de tasas de dosis de radiación gamma con 4 tipos de variables cuantitativas: temperatura, altitud, coordenadas de latitud y longitud, y con 2 variables cualitativas: parroquia y manzana, datos tomados de la información proporcionada por el GIDAC.

4.1. Análisis Exploratorio

4.1.1. Tradicional

Tabla 1-4: Análisis descriptivo de la tdrg en la provincia de Chimborazo.

Descriptivos	Media	Mediana	Mínimo	Máximo
Radiación gamma	0.06	0.06	0.03	0.08

Realizado por: Erazo, Wilson, 2022.

El valor promedio de la tdrg en la provincia de Chimborazo es de 0.06 Sv con una radiación mínima de 0.03 y una permanencia máxima de 0.08 Sv, ubicando el 50% sobre 0.06. El 25% de sus datos esta explicado por 0.05 radiaciones y el 75% por 0.07 radiaciones.

Tabla 2-4: Promedio de la tdrg por cantones en la provincia de Chimborazo.

Cantón	Promedio de radiación gamma	Frecuencia	Porcentaje
Alausí	0.06	80	20%
Chambo	0.04	11	3%
Chunchi	0.06	36	9%
Colta	0.06	45	11%
Cumandá	0.06	15	4%
Guamote	0.06	27	7%
Guano	0.07	8	2%
Pallatanga	0.06	12	3%
Penipe	0.07	6	1%
Riobamba	0.06	168	41%

Realizado por: Erazo, Wilson, 2022.

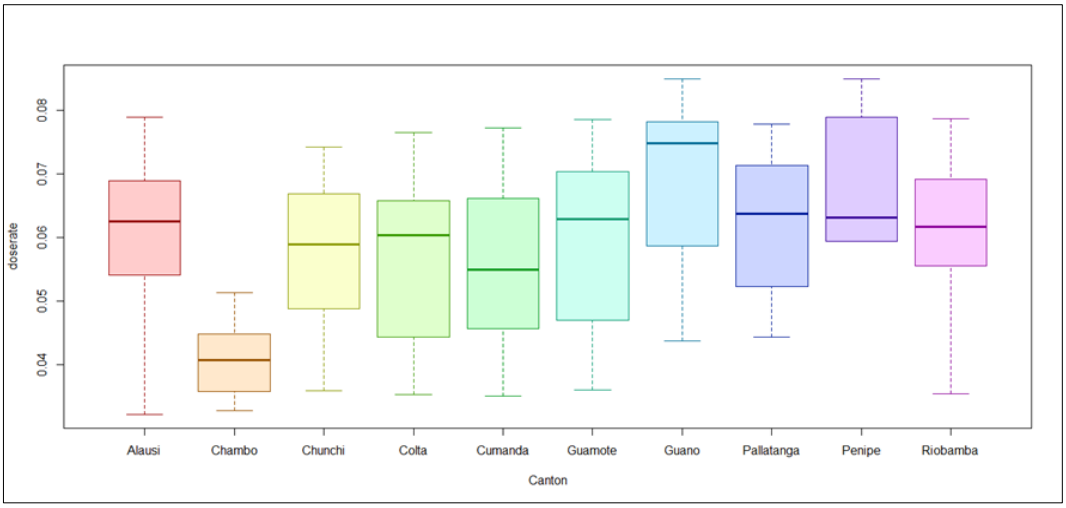


Ilustración 1-4: Boxplot de la tdrg por cantones de la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

La ilustración 1-4 indica los estadísticos de las tdrg por cantones de la provincia, en el cual se identifica diferencias significativas entre varios cantones para corroborar lo antes mencionada se realiza un análisis de varianza (ANOVA), detallado en la tabla 3-4, con un valor p de 0.000 es decir que se confirma la diferencia significativa ente los tdrg de los diferentes cantones.

Tabla 3-4: Análisis de varianza de la radiación gamma de los cantones de la provincia de Chimborazo.

Fuente de Variación	G. L	SC	CM	F	p-value
Cantón	9	0.007	0.00	6.045	0.000
Errores	398	0.047	0.00		

Realizado por: Erazo, Wilson, 2022.

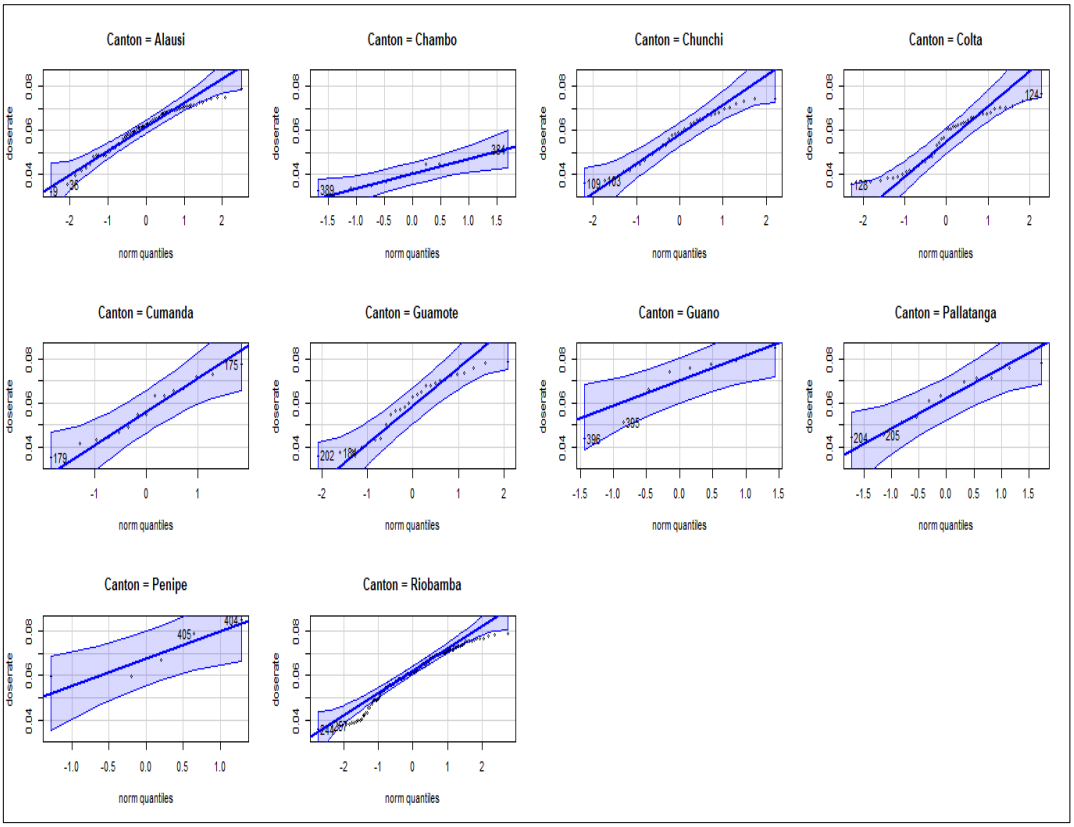


Ilustración 2-4: Distribución en cuartiles de la tdrg por cantones.

Realizado por: Erazo, Wilson, 2022.

La ilustración 2-4 muestra la distribución de los rayos gamma dentro de la provincia, y siguen una distribución normal con un 95% de confiabilidad.

Tabla 4-4: Promedio temporal de la tdrg por en la provincia de Chimborazo.

Clasificación	Radiación gamma
Mañana	0.06
Tarde	0.06

Realizado por: Erazo, Wilson, 2022.

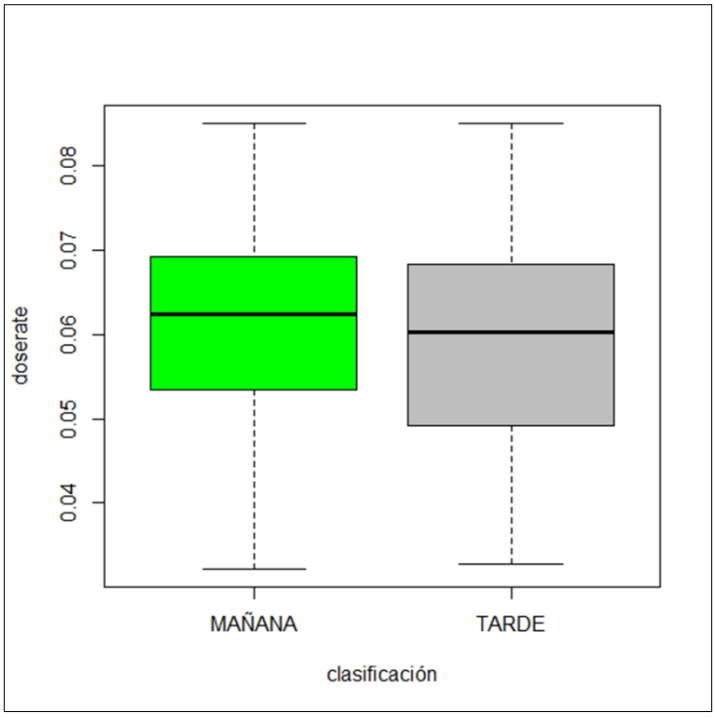


Ilustración 3-4: Boxplot de la tdrg en la mañana y tarde.

Realizado por: Erazo, Wilson, 2022.

La ilustración 3-4 muestra la que no existe diferencia significativa entre la tdrg de la mañana y tarde, lo que se corrobora en la tabla 5-4 ya que se obtuvo un valor p de 0.07 al 95% de confiabilidad.

Tabla 5-4: Análisis de varianza de la tdrg de Chimborazo.

Fuente de Variación	G. L	SC	CM	F	p-value
Clasificación	1	0.00	0.00	3018	0.07
Errores	406	0.05	0.00		

Realizado por: Erazo, Wilson, 2022.

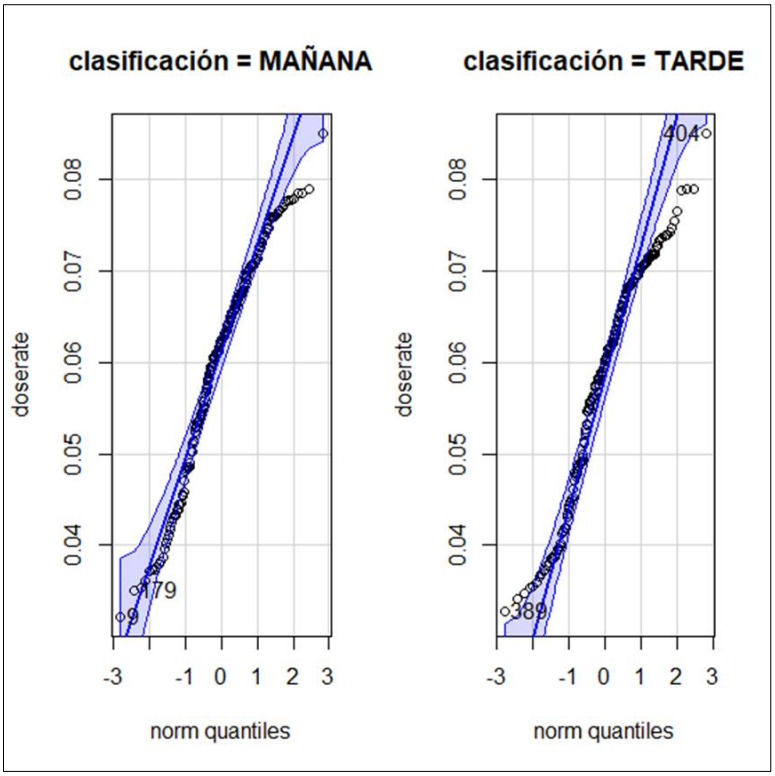


Ilustración 4-4: Distribución en cuartiles de la tdrg por horario.

Realizado por: Erazo, Wilson, 2022.

La ilustración 4-4 muestra la distribución de la tdrg en la mañana y en la tarde el cual indica que siguen distribución normal. Además, que los valores promedios según el periodo temporal son iguales (tabla 4-4).

Tabla 6-4: Promedio de la tdrg según las zonas urbana y rural.

Zona	Promedio de radiación gamma
Rural	0.06
Urbana	0.06

Realizado por: Erazo, Wilson, 2022.

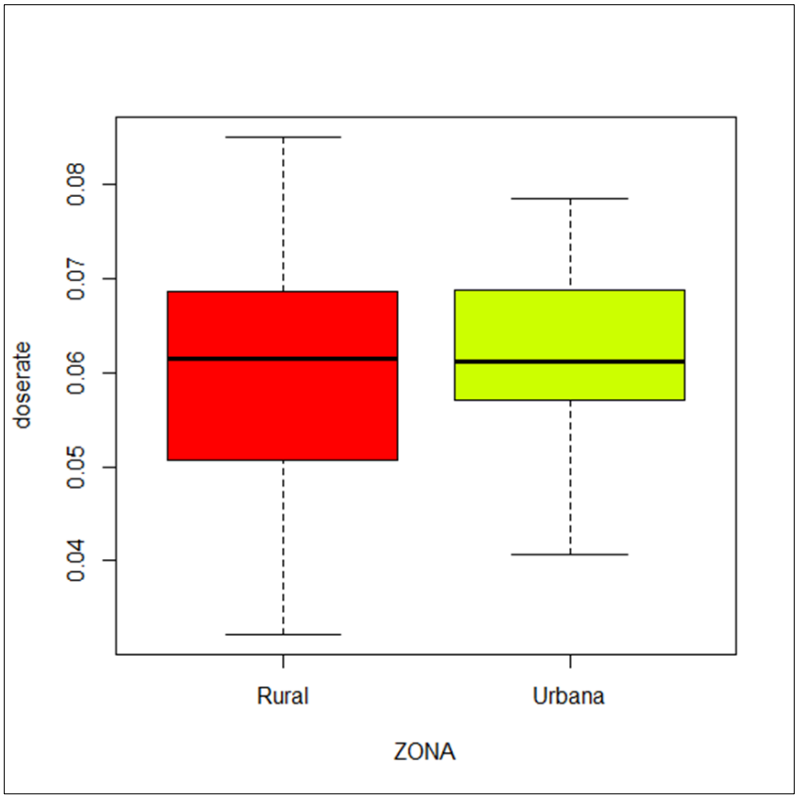


Ilustración 5-4: Boxplot de la tdrg por zonas.

Realizado por: Erazo, Wilson, 2022.

La ilustración 5-4 muestra mayor dispersión de la tdrg en la zona rural, sin embargo, no existe diferencia significativa al 95% de confiabilidad, ya que se obtuvo en la tabla 7- 4 un valor de p de 0.18.

Tabla 7-4: Análisis de varianza de la zona de radiación gamma en la provincia de Chimborazo.

Fuente de Variación	G. L	SC	CM	F	p-value
Zona	1	0.00	0.00	1.82	0.18
Errores	406	0.05	0.00		

Realizado por: Erazo, Wilson, 2022.

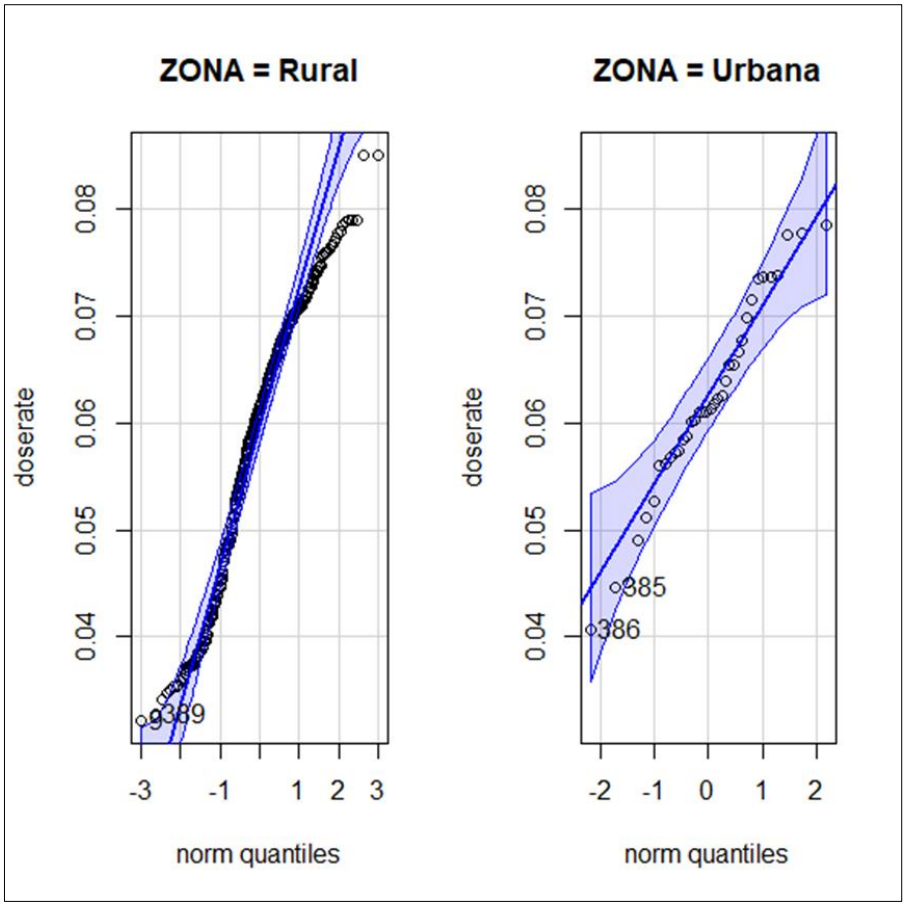


Ilustración 6-4: Distribución de cuartiles de la tdrg por zona.

Realizado por: Erazo, Wilson, 2022.

La ilustración 6-4 muestra la distribución de la tdrg en la zona rural y urbana, lo que indica que siguen una distribución normal.

4.1.2. Espacial

La información proporcionada por el GIDAC (Grupo de Investigación-Desarrollo para el Ambiente y Cambio Climático) presentó variaciones en la tdrg en la provincia, por lo que, se realizó el análisis espacial y de conglomerados exclusivamente sobre la provincia de Chimborazo.

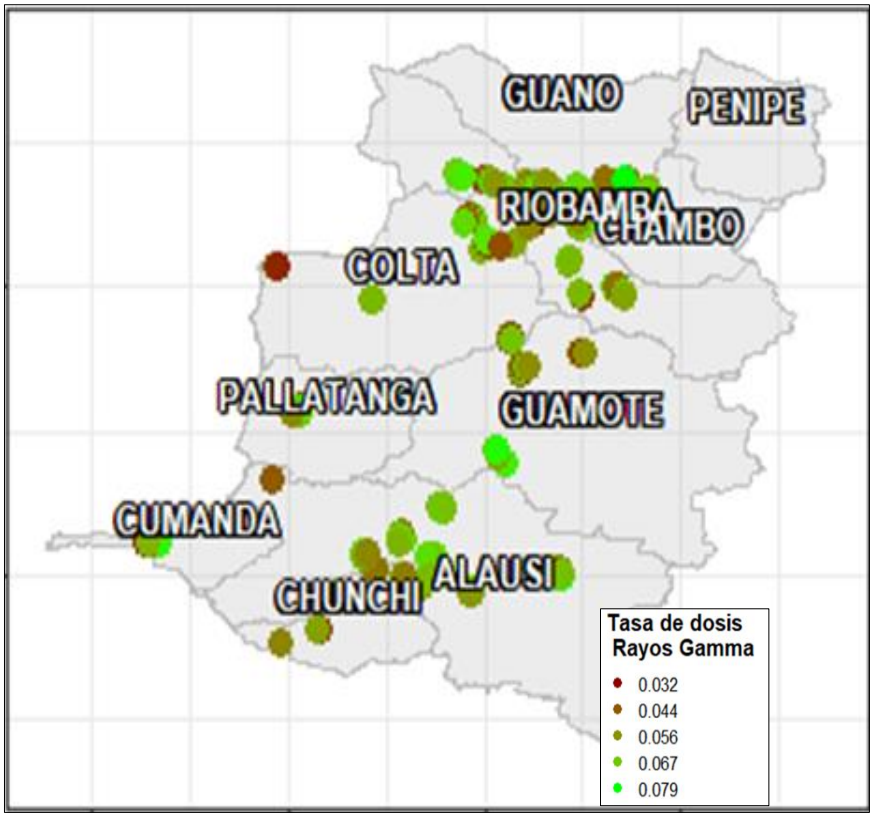


Ilustración 7-4: Distribución de la tdrg en la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

La ilustración 7-4 muestra la ubicación geográfica de las ubicaciones muestreadas en donde se midió la tdrg en la provincia de Chimborazo.

Tabla 8-4: Índice de autocorrelación de Morán.

Variable	Índice de Moran	Valor de Z	p- value
Dosis de radiación gamma	0.1269	4.289	0.000018

Realizado por: Erazo, Wilson, 2022.

Mediante el análisis de autocorrelación espacial con el uso del índice de Morán, se determinó que la tdrg en la provincia de Chimborazo tienen una autocorrelación positiva perfecta, con un nivel de significancia del 5%, es decir presentó un equilibrio entre los valores positivos y negativos del estudio espacial. Esto quiere decir que, mientras más pequeño sea el valor de p mejor autocorrelación presentan los datos, lo que indica que la tdrg es clusterizable.

4.2. Análisis de clustering jerárquico

Para aplicar la metodología de clúster jerárquico aglomerativo es fundamental estandarizar la matriz de información.

4.2.1. Análisis de conglomerados de la radiación gamma en la provincia de Chimborazo

Se realiza el análisis clúster de las tasas de dosis de rayos gamma. Para la cual, se hace un análisis de componentes principales, indagación del número de clúster óptimos, identificación detallada del agrupamiento mediante el dendrograma, construcción del clúster y finalmente identificación de las tdrg clusterizados en la provincia de Chimborazo y sus cantones.

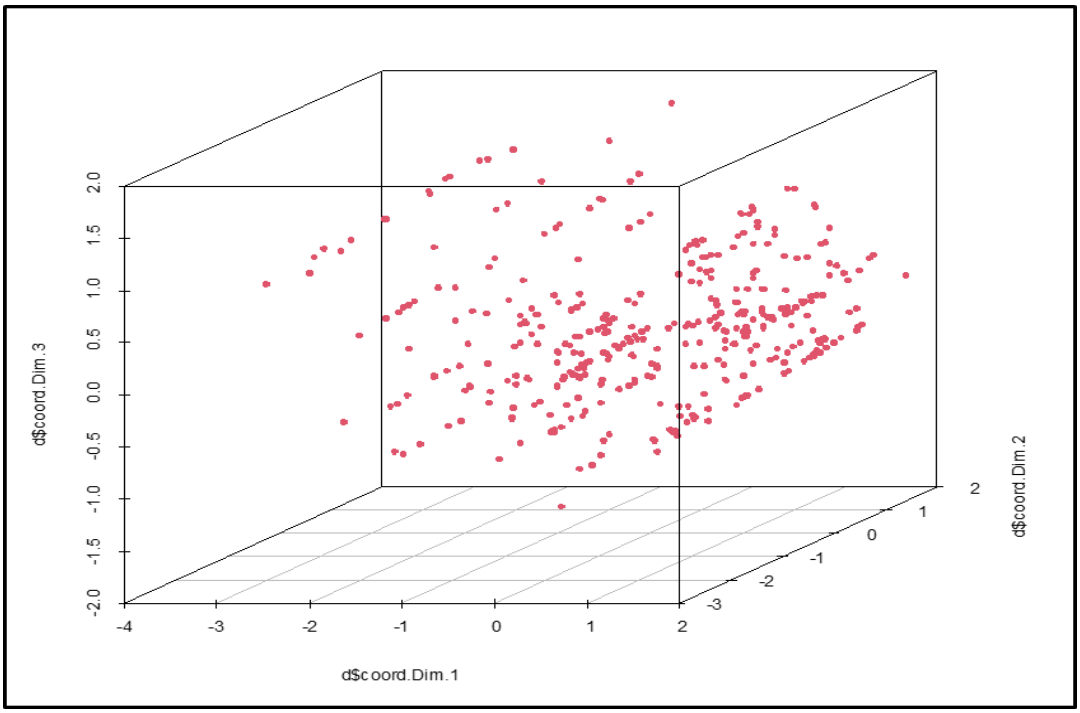


Ilustración 8-4: Cubo de homogeneidad de la tdrg en la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 8-4 se observa, que la tdrg en la provincia de Chimborazo tiene una distribución homogénea.

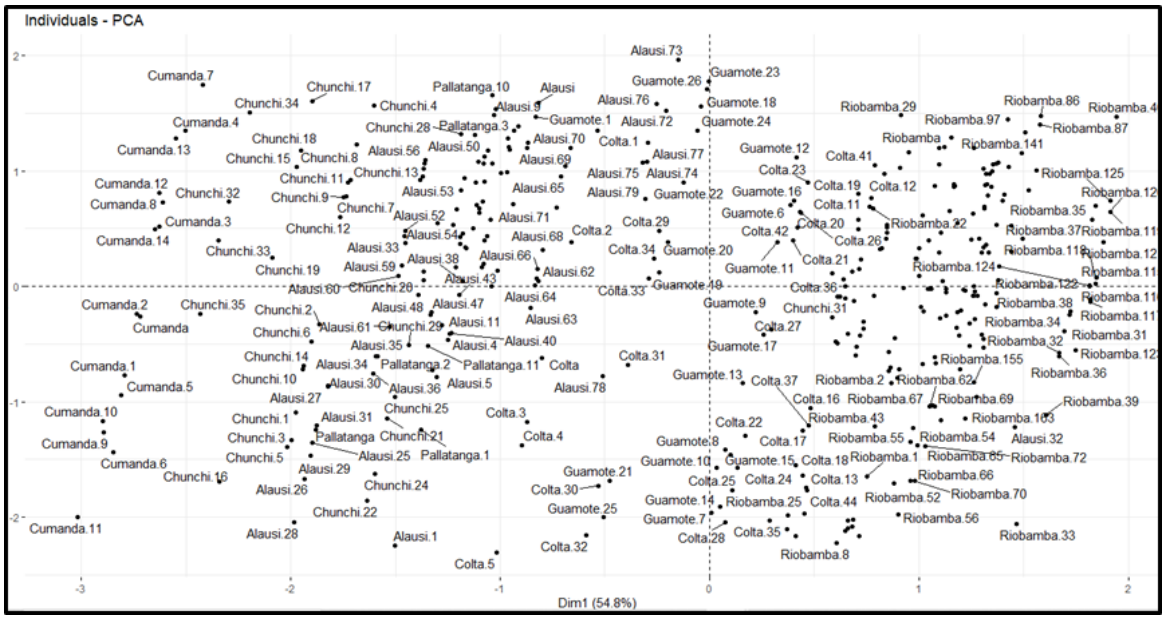


Ilustración 9-4: Distribución de la tdrg en la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 9-4 se observa la aglomeración de la información en los cuatro cuadrantes, indica la existencia de asociación a través de clúster. El coeficiente estadístico de Hopkins con un índice de 0.55 evidencia que la base de datos es clusterizable.

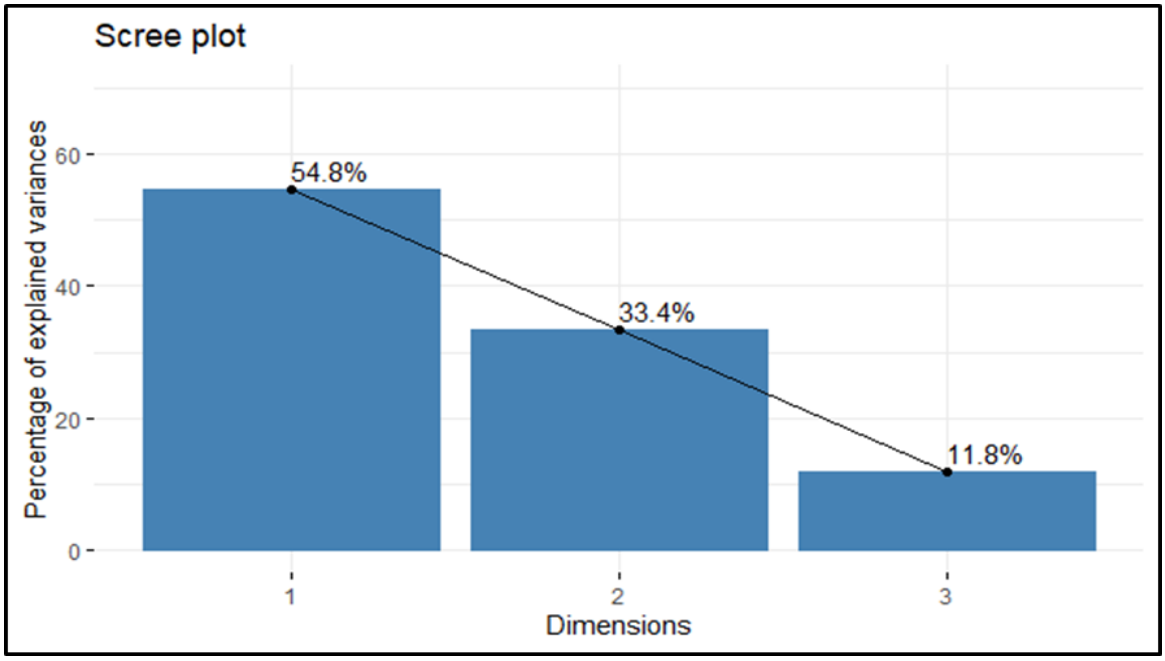


Ilustración 10-4: Porcentaje de la varianza explicada de las PCA en Chimborazo.

Realizado por: Erazo, Wilson, 2022.

La ilustración 10-4 se determina que las 2 primeras componentes principales explican un 88.20% de variabilidad total de la tdrg en la provincia de Chimborazo.

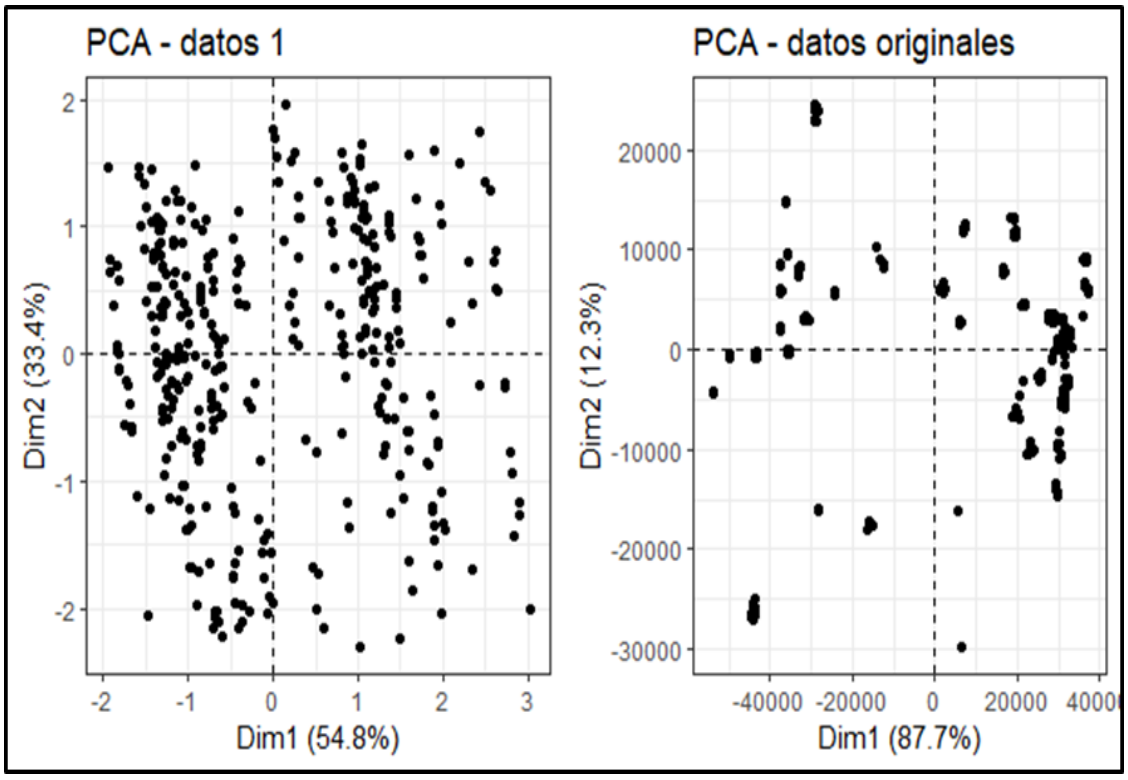


Ilustración 11-4: Estandarización de la tdrg en la provincia de Chimborazo.

Realizado por: *Erazo, Wilson, 2022.*

Mediante la ilustración 11-4 se observa las tdrg estandarizadas, en el cual se identifica la asociación entre los datos. Además, el factor del estadístico de Hopkins con un índice de 0.55 que evidencia una vez más que la base de datos son clusterizables.

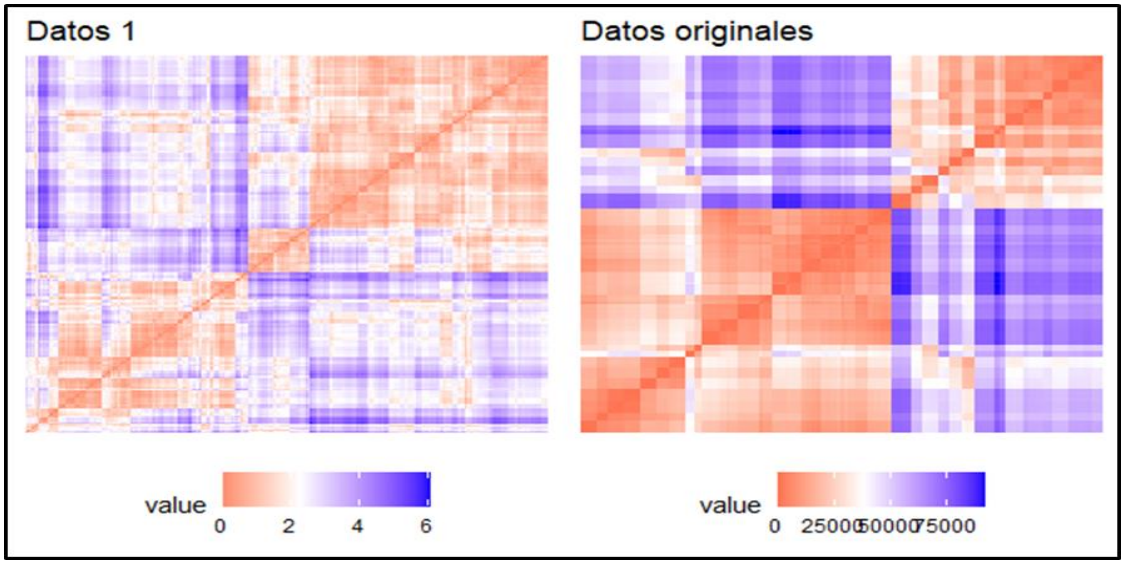


Ilustración 12-4: Relación de la tdrg en la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 12-4 muestra una autocorrelación positiva perfecta entre la ubicación geográfica y la presencia de la tdrg en la provincia de Chimborazo. También se identifica diferencia significativa entre la autocorrelación de datos estandarizados y originales, lo cual permite apreciar que la estandarización mejora la visualización de la autocorrelación de forma gráfica.

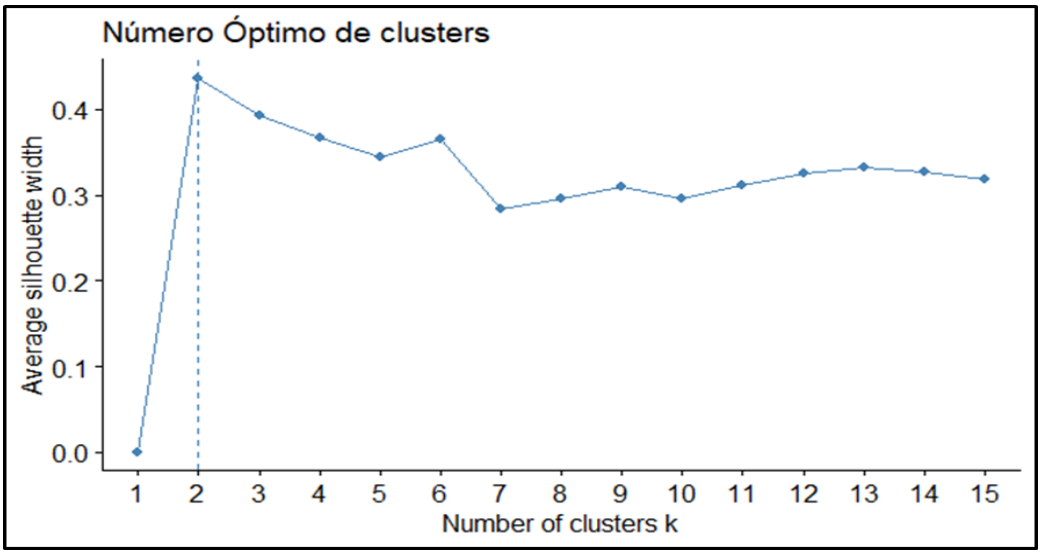


Ilustración 13-4: Número de clúster de tdrg por el método de silueta promedio.

Realizado por: Erazo, Wilson, 2022.

A través del método de silueta promedio que maximiza la media representada en la ilustración 13-4, se identifica que 2 es el número óptimo de clúster para la tdrg en la provincia de Chimborazo, ya que se concentran de manera homogénea las características geográficas relacionadas con la tdrg que ingresa en la atmosfera en la provincia de Chimborazo en los diferentes cantones que contiene esta región.

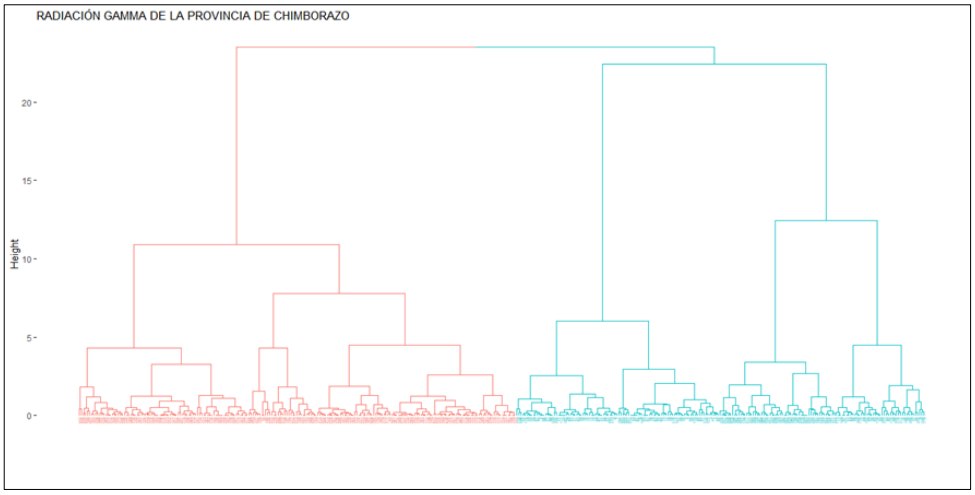


Ilustración 14-4: Dendrograma de la tdrg en la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

La ilustración 14-4 explica la distribución de 2 clúster. Se observa mayor concentración de tdrg en el segundo aglomerado distribuido en los diferentes cantones de la provincia de Chimborazo. Además, muestra similaridad entre las tdrg de los diferentes cantones con cortes proporcionados entre las dos agrupaciones, siendo la razón por la que se intersecan varios puntos ya que presentan características semejantes de ubicación geográfica y de la tdrg.

Tabla 9-4: Clasificación de clúster por cantones de la provincia de Chimborazo.

Cantón	Clúster	
	1	2
Alausí	76	3
Chambo	10	1
Chunchi	34	2
Colta	16	29
Cumandá	15	0
Guamote	17	10
Guano	1	7
Pallatanga	4	8
Penipe	0	6
Riobamba	23	145
Total	**196**	**211**

Realizado por: Erazo, Wilson, 2022.

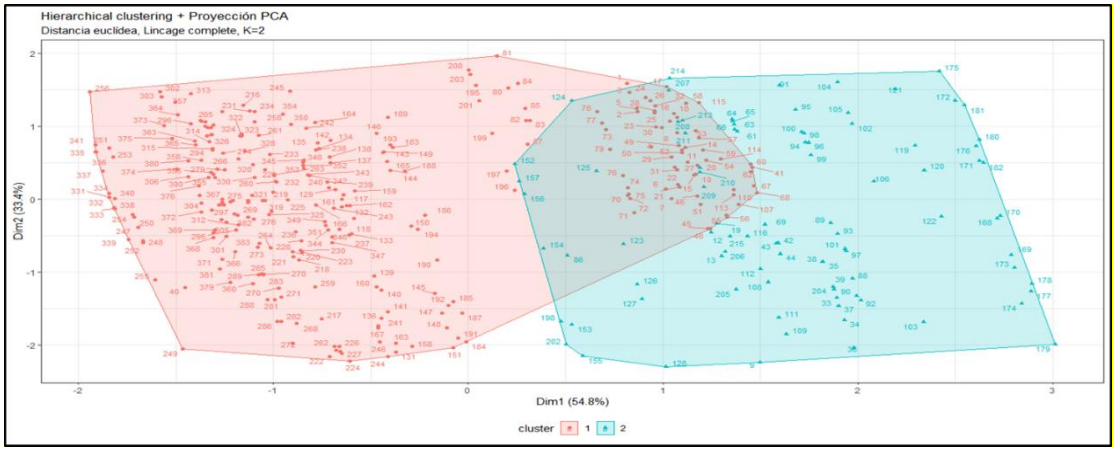


Ilustración 15-4: Clusterización de la tdrg en la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 15-4 se expone los clústeres de la tdrg tomados por GIDAC, con un coeficiente de afluencia del 0.70. Se identifico 2 clúster expuestos en la tabla 9-4 según los cantones dentro de la provincia, con una mayor incidencia en el segundo clúster (211 dosis). La mayoría de las tdrg se concentran en el clúster 1.

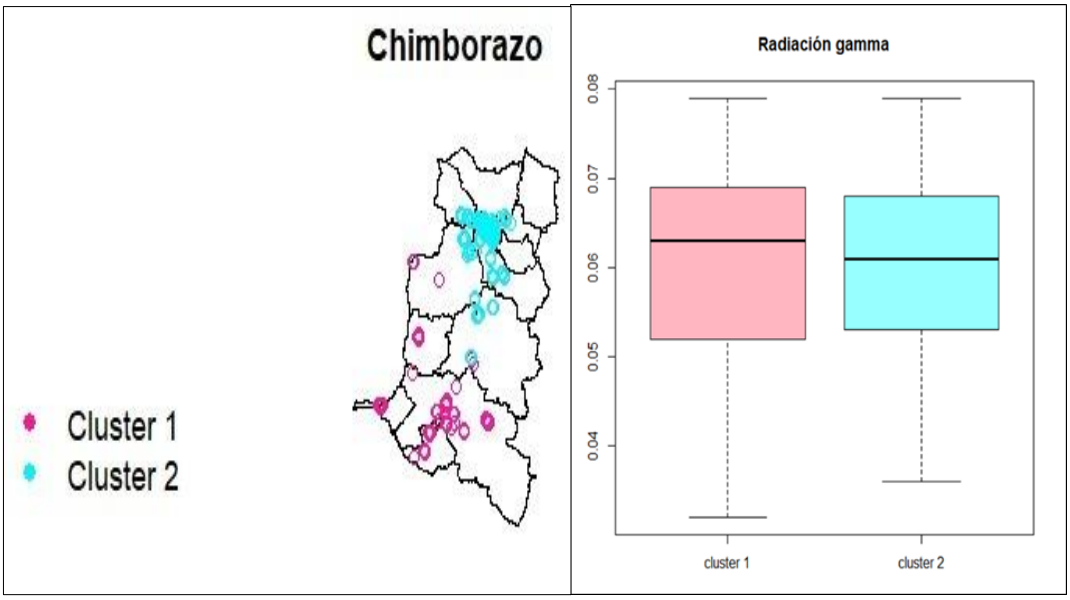


Ilustración 6-4: Mapa de clasificación de clúster y promedios.

Realizado por: Erazo, Wilson, 2022.

La ilustración 16-4 muestra la ubicación geográfica de las tdrg clusterizadas en los cantones de la provincia de Chimborazo. La ubicación geográfica del clúster 1 está en el sur de la provincia mientras que el clúster 2 se concentra en el norte de la provincia. Los diagramas de caja indican que no existe diferencia significativa entre los tdrg del clúster 1 y el clúster 2, aunque geográficamente estén bien definidos.

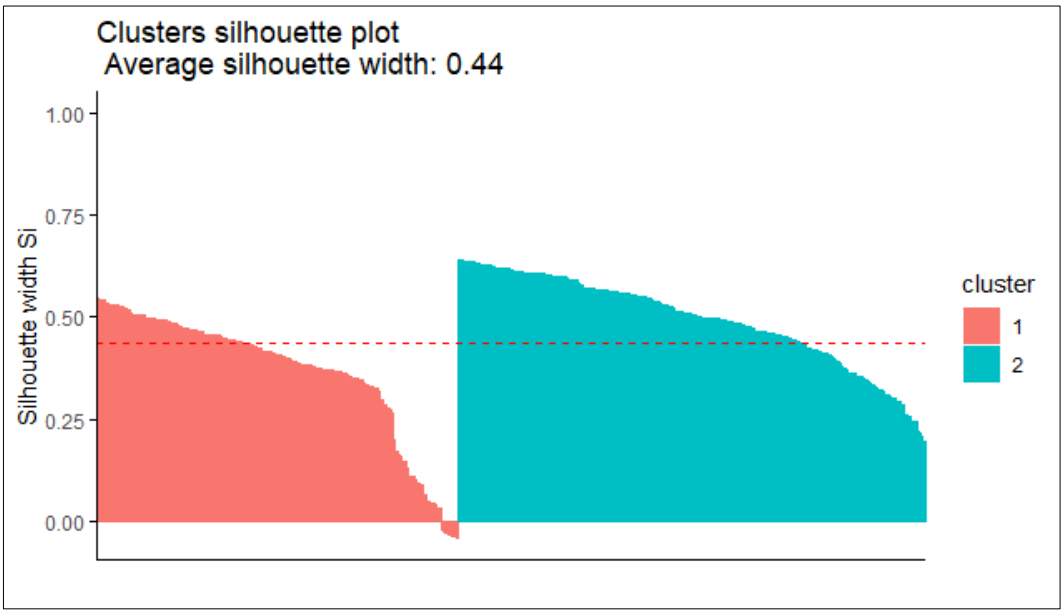


Ilustración 17-4: Aglomeración de la tdrg por clúster.

Realizado por: Erazo, Wilson, 2022.

El método de silueta promedio que se muestra en la ilustración 17-4 identifica un 47% de la tdrg está dentro del primer clúster aglomerativo y el 53% en el segundo, de esta manera se identifica la mayor concentración de tdrg en el segundo clúster, especialmente en los cantones Alausí, Colta y Guamote, lo que se cree que es debido a las bajas temperaturas presentes en dichos cantones.

4.2.2. Clúster de los cantones

4.2.3. Alausí

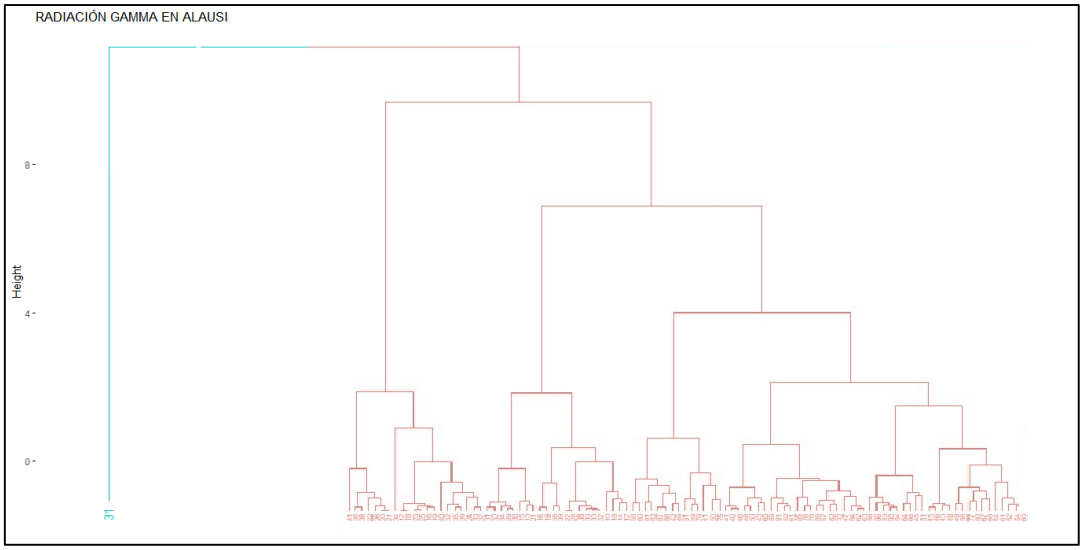


Ilustración 18-4: Dendrograma del cantón Alausí.

Realizado por: Erazo, Wilson, 2022.

La ilustración 18-4 explica 2 clústers que tiene una gran concentración de tdrg en el clúster N°1 en el cual se distribuye en el cantón Alausí. Además, muestra cortes simétricos dando paso a la intersección de varios puntos.

Ilustración 19-4: Clúster del cantón Alausí.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 19-4 se puede observar los clústeres de la tdrg del cantón Alausí con un coeficiente de aflujo del 98.75 % en el clúster N°1, ya que los datos fueron tomados en distintos lugares del cantón por lo cual se puede ver la cantidad de tdrg mediante el método jerárquico aglomerativo y los puntos se dividen y se agrupan de esta manera.

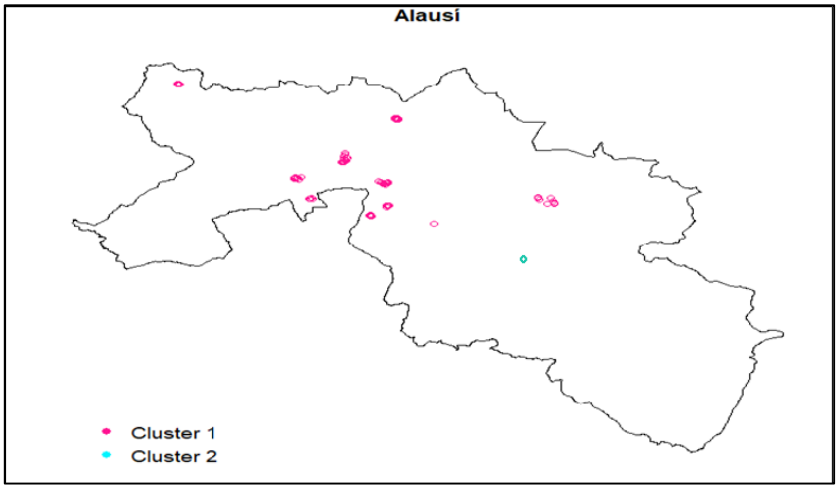


Ilustración 20-4: Mapa del cantón Alausí.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 20-4 se muestra el mapa del cantón Alausí teniendo en cuenta que se dividen los clústeres de esa manera ya que los datos fueron tomados en distintos lugares del cantón y fueron enfocados en el nivel de tdrg, por lo cual se puede ver la influencia que tiene la ubicación geografía en el clúster mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de esta manera.

4.2.4. Chunchi

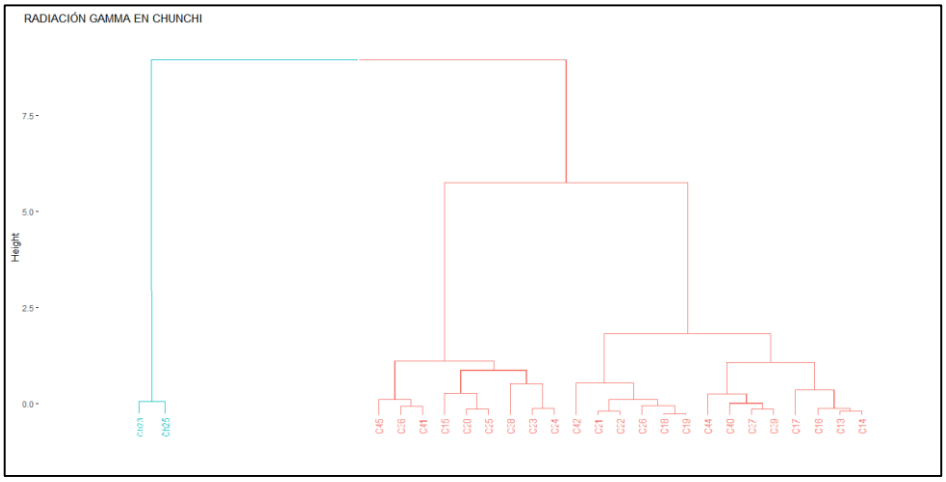


Ilustración 21-4: Dendrograma del cantón Chunchi.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 21-4 se explica la distribución de 2 clúster. En la cual su mayor concentración de tdrg se encuentra en el clúster N°1 con cortes simétricos siendo la razón por la que se intersecan varios puntos ya que presentan características similitud en la ubicación geográfica.

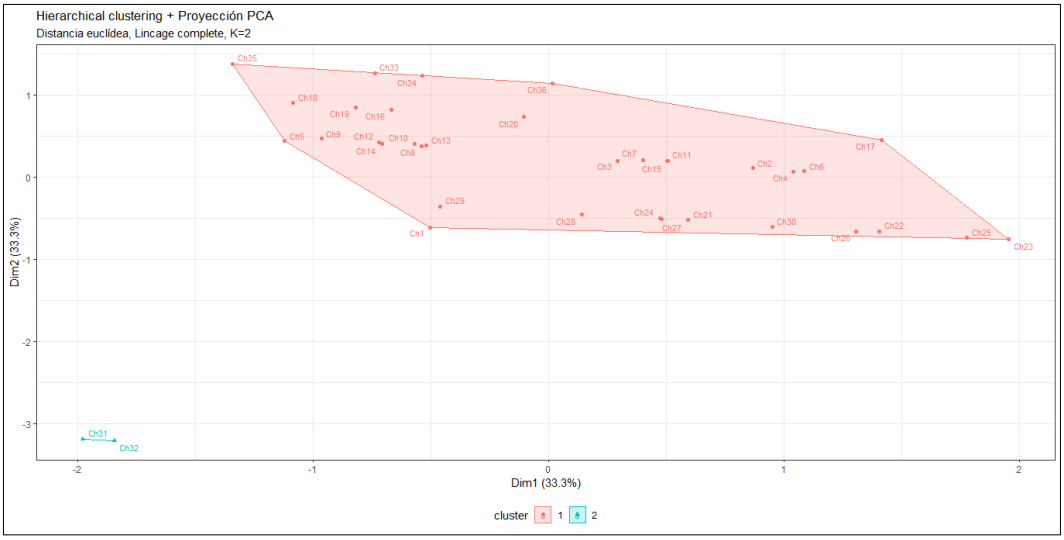


Ilustración 22-4: Clúster del cantón Chunchi.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 22-4 que muestran los clústers de la tdrg del cantón Chunchi teniendo más afluencia el clúster N°1 con un coeficiente de aglomeración del 0.98 y con una cantidad de 34 tdrg y por otra parte el clúster N°2 con una cantidad de 2 tdrg.

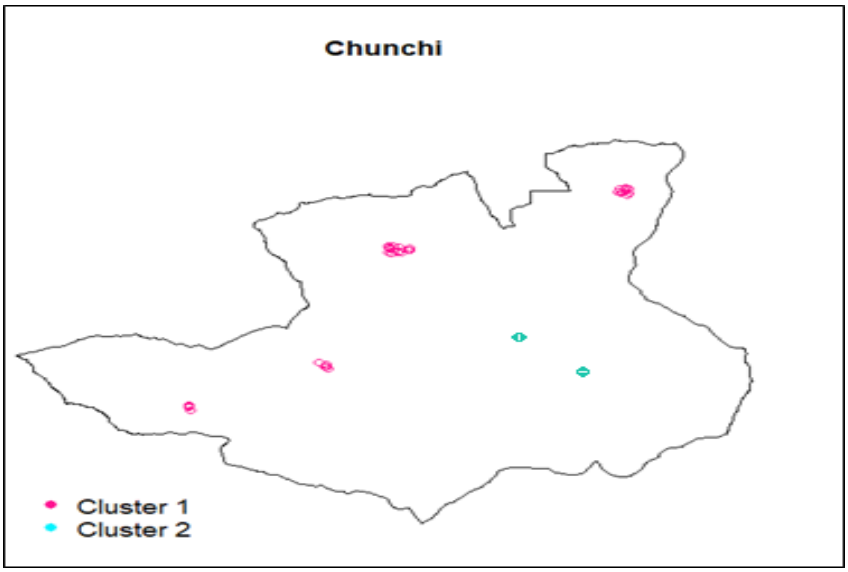


Ilustración 23-4: Mapa del Cantón Chunchi.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 23-4 que muestra el mapa se dividen los clústeres de esa manera ya que los datos fueron tomados en distintos lugares del cantón por lo cual se puede ver la influencia que tiene la ubicación geográfica, mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de esta manera.

4.2.5. Colta

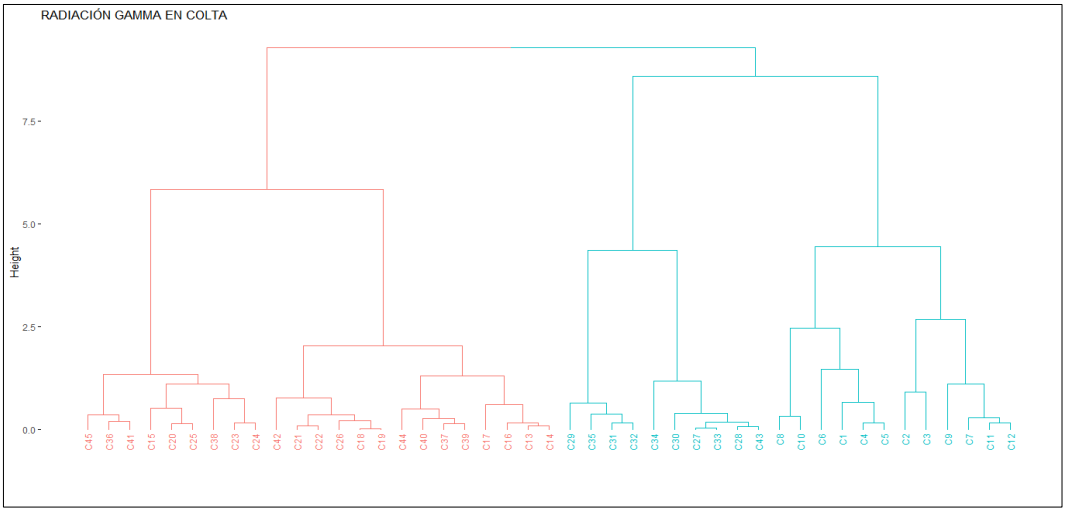


Ilustración 24-4: Dendrograma del cantón Colta.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 24-4 se explica la distribución de 2 clúster con una mayor concentración de tdrg ubicado en el clúster N°1. Además, se observa cortes simétricos por esta razón se intersecan varios puntos.

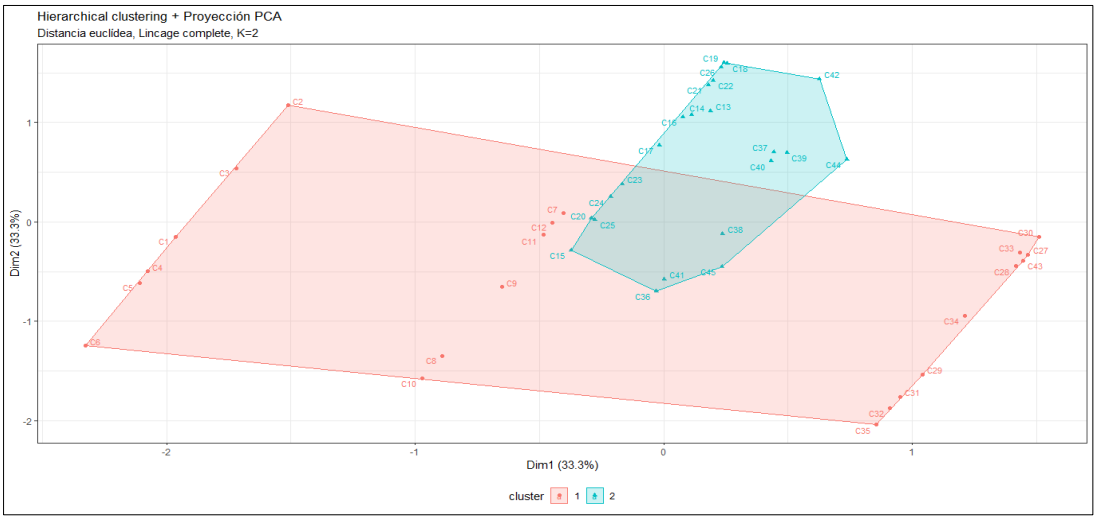


Ilustración 25-4: Clúster del cantón Colta.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 25-4 que muestran los clústeres de tdrg se muestra que existe un 0.51% de aglomeración en el clúster N°1 (23 dosis) y por otra parte con un 0.49% en el clúster N°2 (22 dosis).

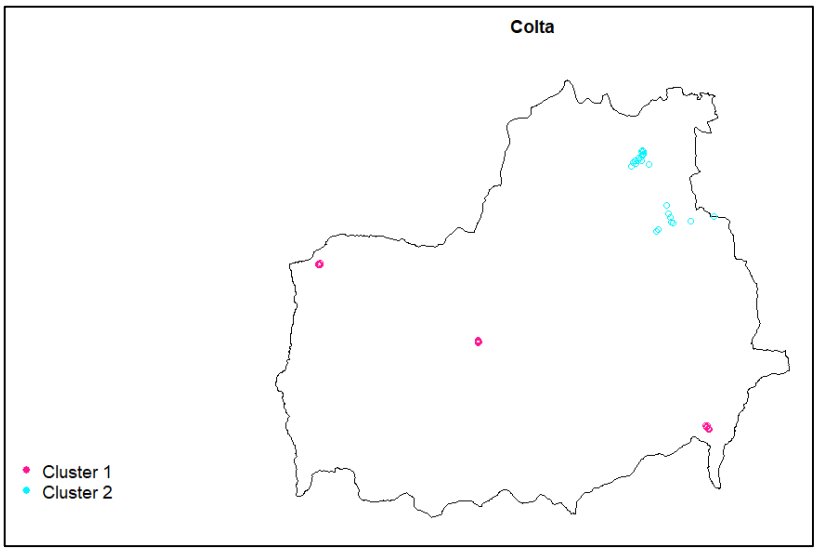


Ilustración 26-4: Mapa del Cantón Colta.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 26-4 que muestra el mapa está dividido en los clústeres de esa manera ya que los datos fueron tomados en distintos lugares del cantón por lo cual se puede ver la influencia que tiene la ubicación geográfica, mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de esta manera los mismo que están dispersos en el cantón Colta.

4.2.6. Cumandá

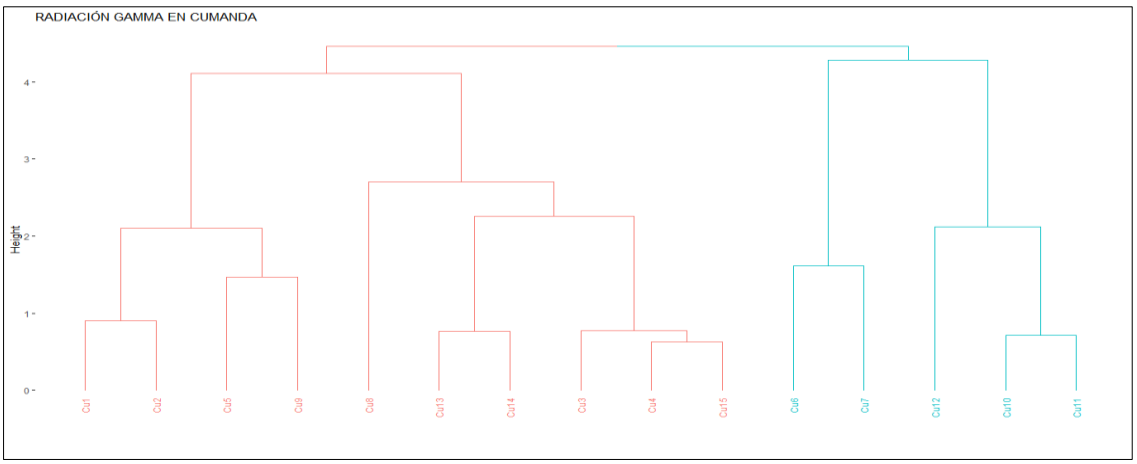


Ilustración 27-4: Dendrograma del Cantón Cumandá.

Realizado por: Erazo, Wilson, 2022.

La ilustración 27-4 explica la distribución de 2 clúster con una mayor concentración de tdrg en el primer conglomerado. Además, se observa cortes simétricos por esta razón se intersecan varios puntos de tdrg en estudio.

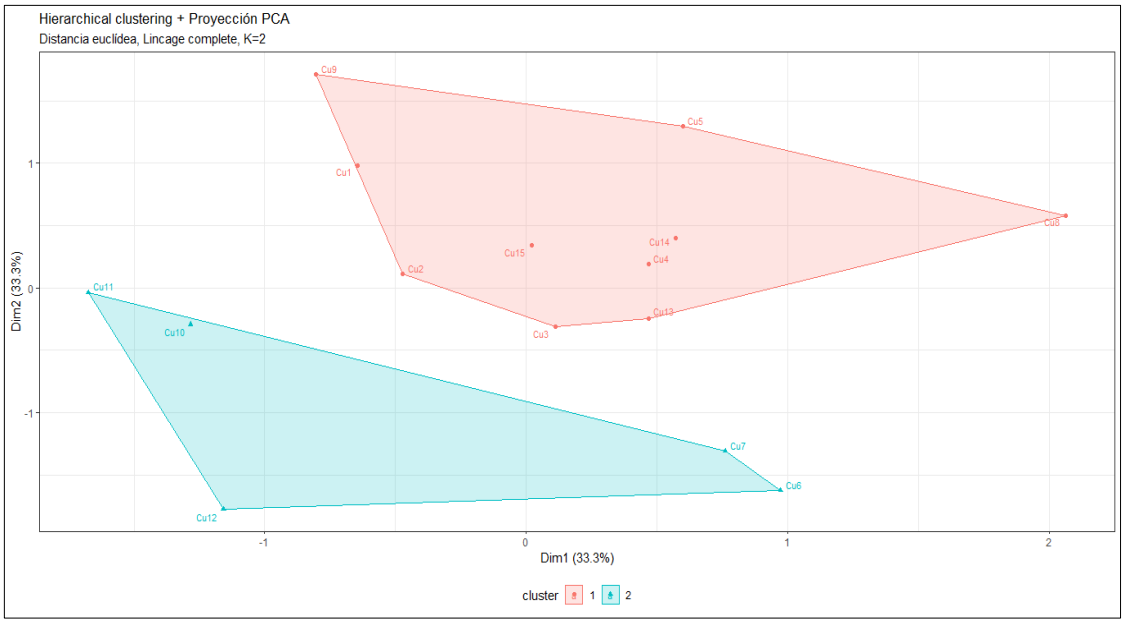


Ilustración 28-4: Clúster del cantón Cumandá.

Realizado por: Erazo, Wilson, 2022.

La ilustración 28-4 que muestran los clústeres de la tdrg en el cual se observa un coeficiente de aglomeración del 0.67 en el primer clúster (10 dosis) y del 0.33 en el segundo clúster (5 dosis).

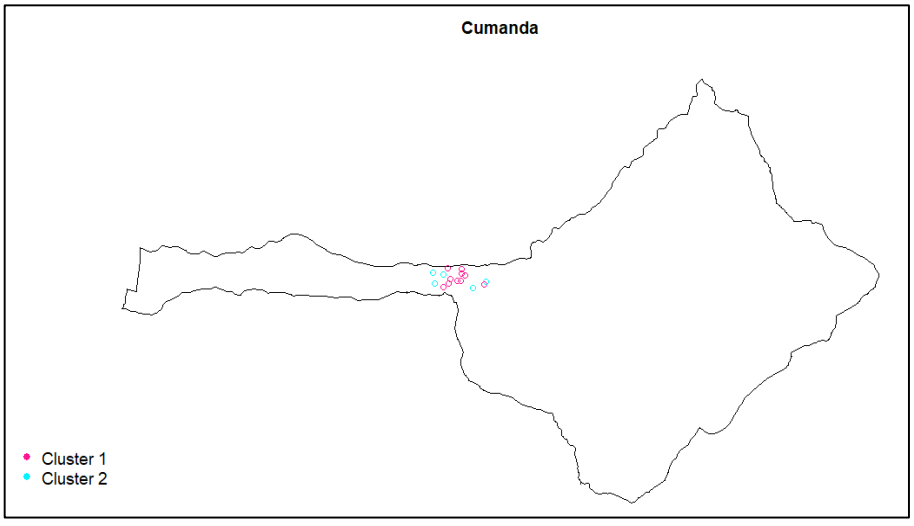


Ilustración 29-4: Mapa del cantón Cumandá.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 29-4 que muestra el mapa se dividen los clústeres de esa manera ya que los datos fueron tomados en lugares cercanos del cantón por lo cual se puede ver la influencia que tiene la ubicación geografía, mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de esta manera.

4.2.7. Guamote

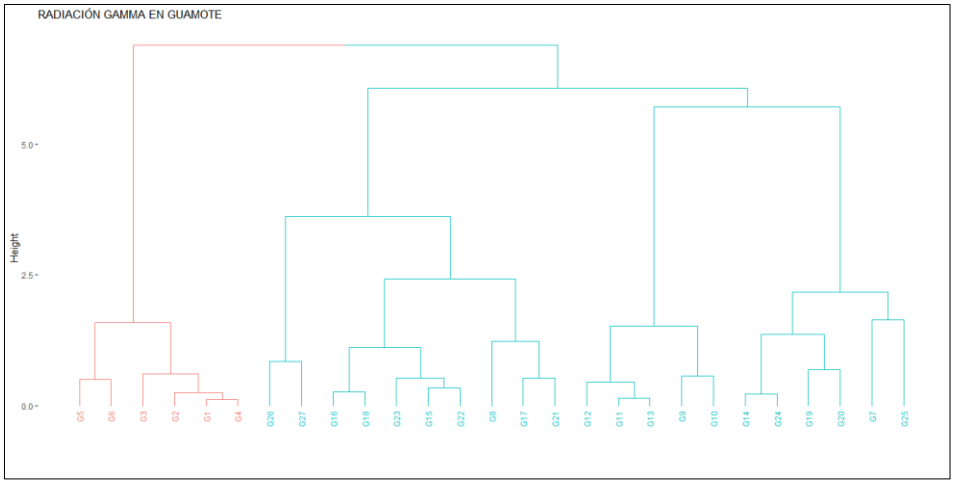


Ilustración 30-4: Dendrograma del cantón Guamote.

Realizado por: Erazo, Wilson, 2022.

La ilustración 30-4 explica la distribución de 2 clúster. Se observa una mayor concentración de tdrg en el segundo conglomerado. Además, muestra similaridad entre las tdrg de los datos en estudio, evidenciando cortes proporcionados entre las dos agrupaciones, siendo la razón por la que se intersecan varios puntos ya que presentan características semejantes de ubicación geográfica y de la tdrg.

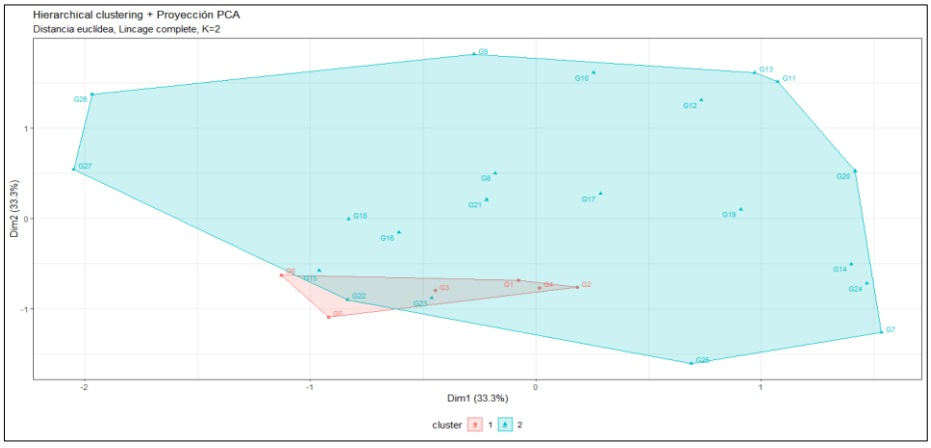


Ilustración 31-4: Clúster del cantón Guamote.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 31-4 se expone los clústeres de tdrg en la cual se aprecia un coeficiente de aglomeración del 0.78 en el segundo clúster (21 dosis) y del 0.22 en el primer clúster (6 dosis), siendo la mayor concentración en el clúster N°2.

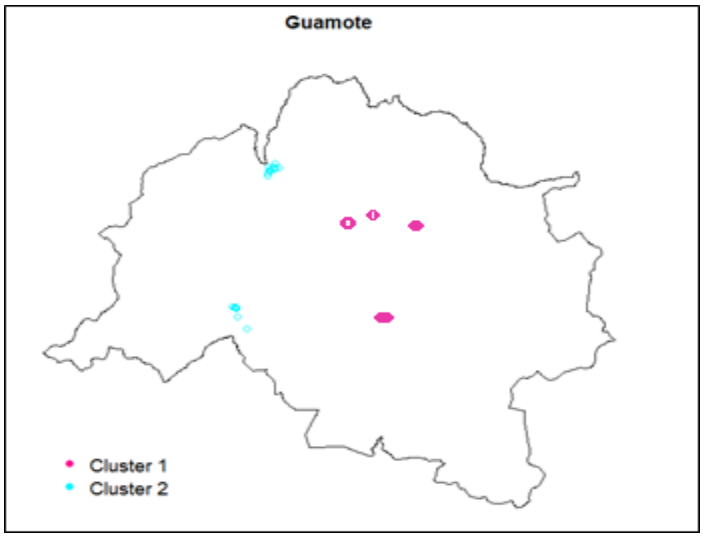


Ilustración 32-4: Mapa del Cantón Guamote.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 32-4 que muestra el mapa esta divido en los clústeres de esa manera ya que los datos fueron tomados en distintas ubicaciones geográficas del cantón en la cual se puede ver la influencia que tiene la tdrg, mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de dicha manera.

4.2.8. Pallatanga

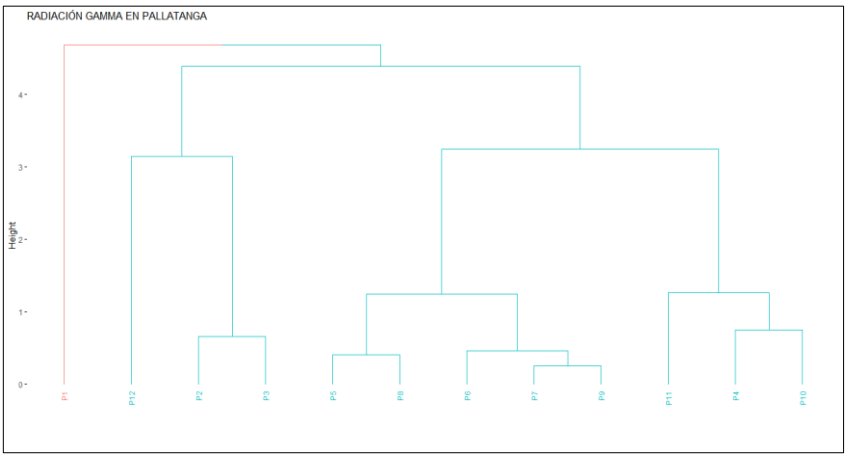


Ilustración 33-4: Dendrograma del cantón Pallatanga.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 33-4 se explica la distribución de 2 clúster. Se aprecia una mayor concentración de tdrg en el segundo conglomerado y una mínima en el primero. Además, muestra similaridad en los tdrg con cortes proporcionados entre las dos agrupaciones por su ubicación geográfica y la tdrg.

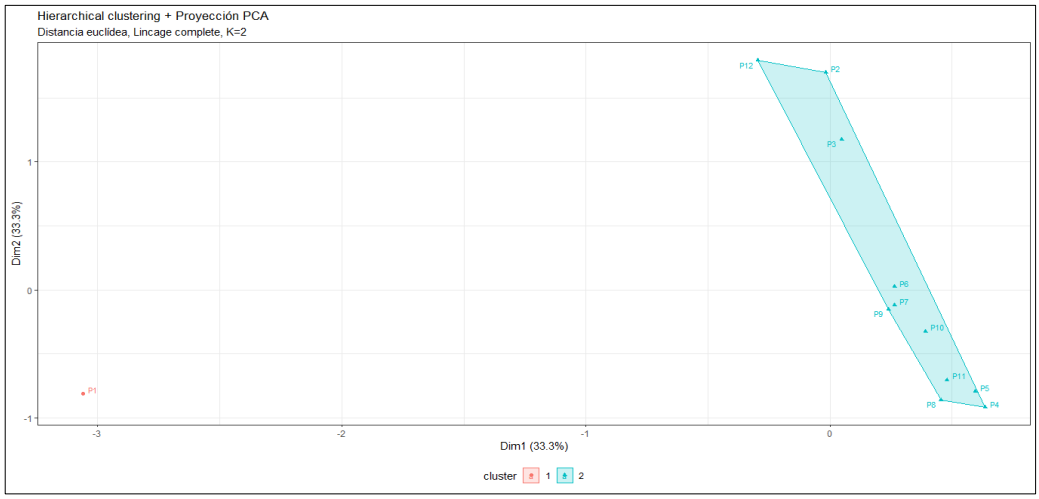


Ilustración 34-4: Clúster del cantón Pallatanga.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 34-4 se muestran los clústeres de tdrg con un coeficiente de aglomeración del 0.92 en el segundo clúster (11 dosis) y del 0.08 en el primer clúster (1 dosis), siendo su mayor aglomeración en el clúster N°2.

Ilustración 35-4: Mapa del Cantón Pallatanga.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 35-4 que muestra el mapa en el cual está dividido los clústeres de esta manera, ya que los datos fueron tomados en un lugar en específico del cantón por lo cual se puede la ubicación geografía en el clúster mediante el método jerárquico aglomerativo.

4.2.9. Riobamba

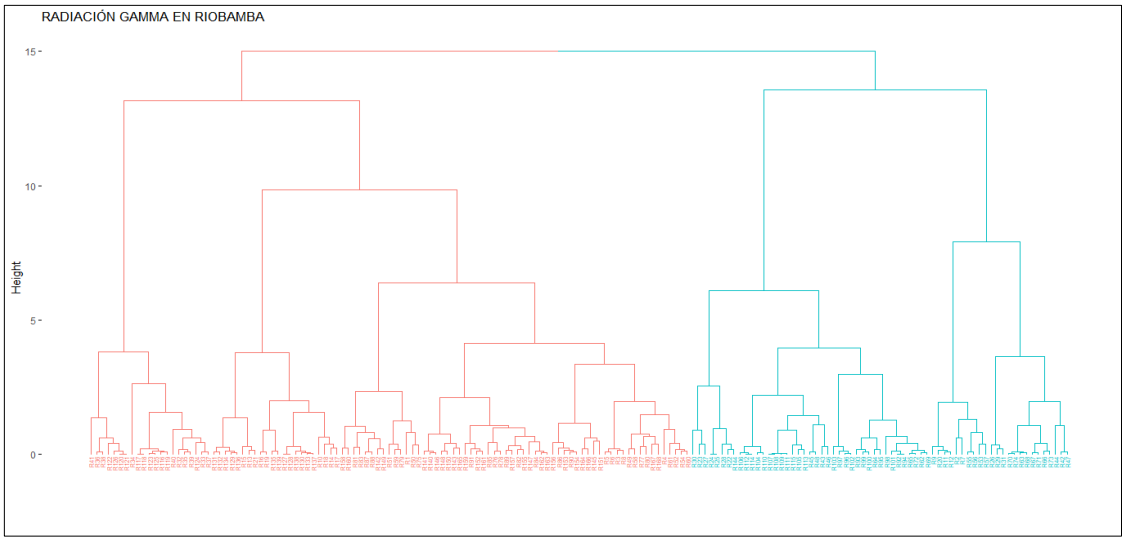


Ilustración 36-4: Dendrograma del cantón Riobamba.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 36-4 explica la distribución de 2 clúster, con una mayor concentración de tdrg en el primer conglomerado. Además, existe similaridad en los tdrg con cortes proporcionados entre los dos clústeres por ubicación geográfica y la tdrg.

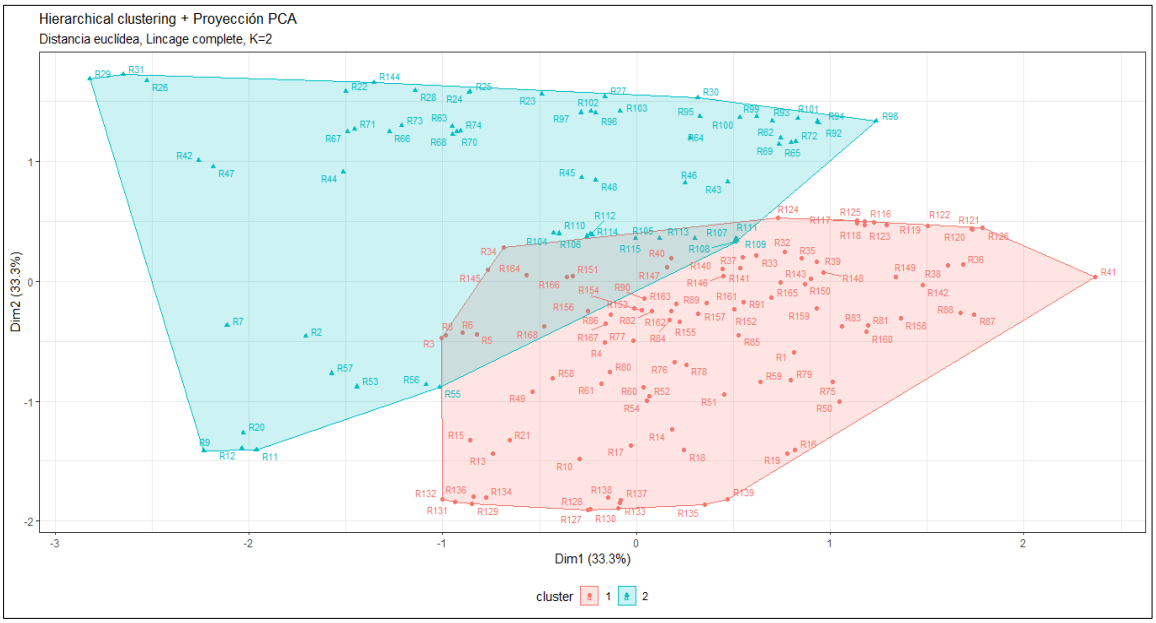


Ilustración 37-4: Clúster del cantón Riobamba.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 37-4 se muestran los clústeres de tdrg con un coeficiente de 0.60 en el primer clúster y del 0.40 en el segundo, siendo su mayor aglomeración en el clúster N°1, en la cual se observa la dispersión de los datos tomados en distintos lugares del cantón. Además, se observa la cantidad de tdrg mediante el método jerárquico aglomerativo. Debido a que el mayor número de muestras tomadas en el canton Riobamba pertenecen a la ciudad se mapea específicamente aquella zona.

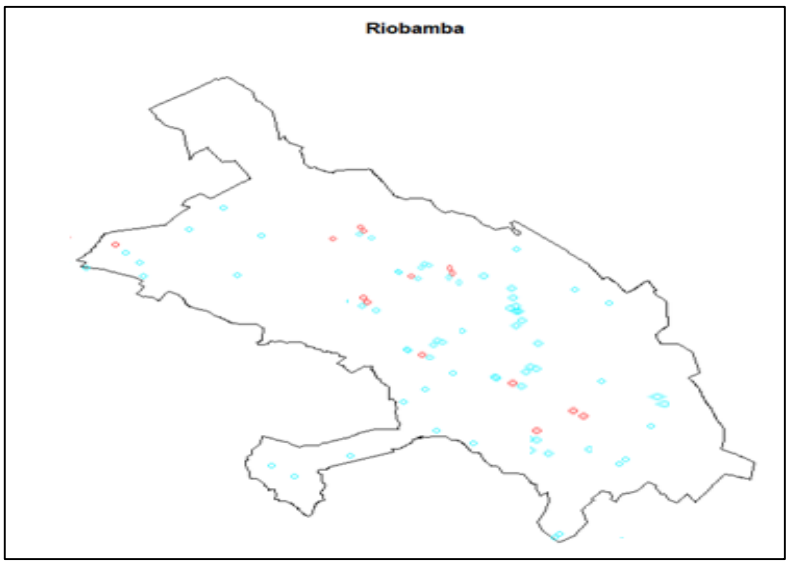


Ilustración 38-4: Mapa del cantón Riobamba.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 38-4 que muestra el mapa en el cual está dividida los clústeres de esta manera ya que los datos fueron tomados en distintos lugares de la ciudad de Riobamba, en el cual no está claramente definida los clústers definida.

4.2.10. Chambo

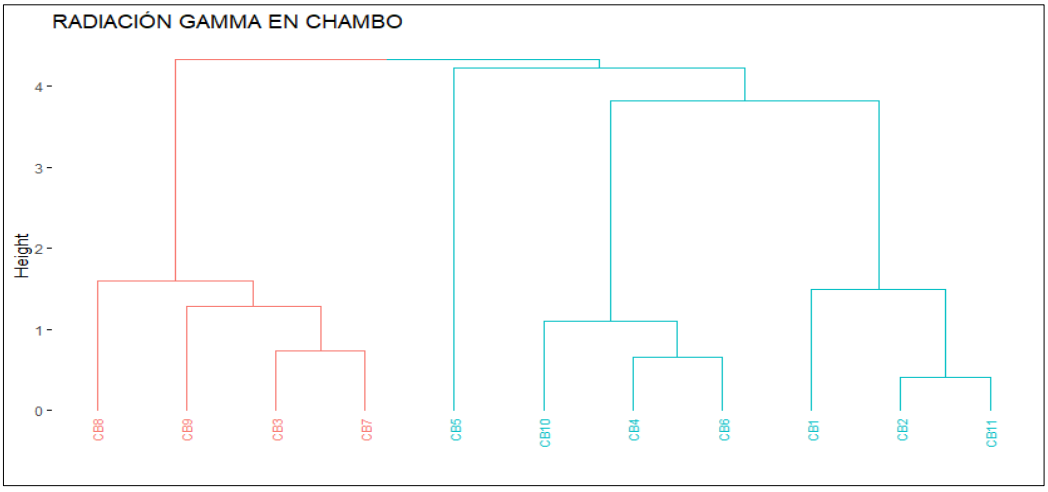


Ilustración 39-4: Dendrograma del cantón Chambo.

Realizado por: Erazo, Wilson, 2022.

La ilustración 39-4 explica la distribución de 2 clúster. Se observa una mayor concentración de en el segundo conglomerado. Además, muestra similaridad entre las tdrg de los datos en estudio, razón por la cual se intersecan varios puntos ya que presentan características semejantes de ubicación geográfica y de la tdrg.

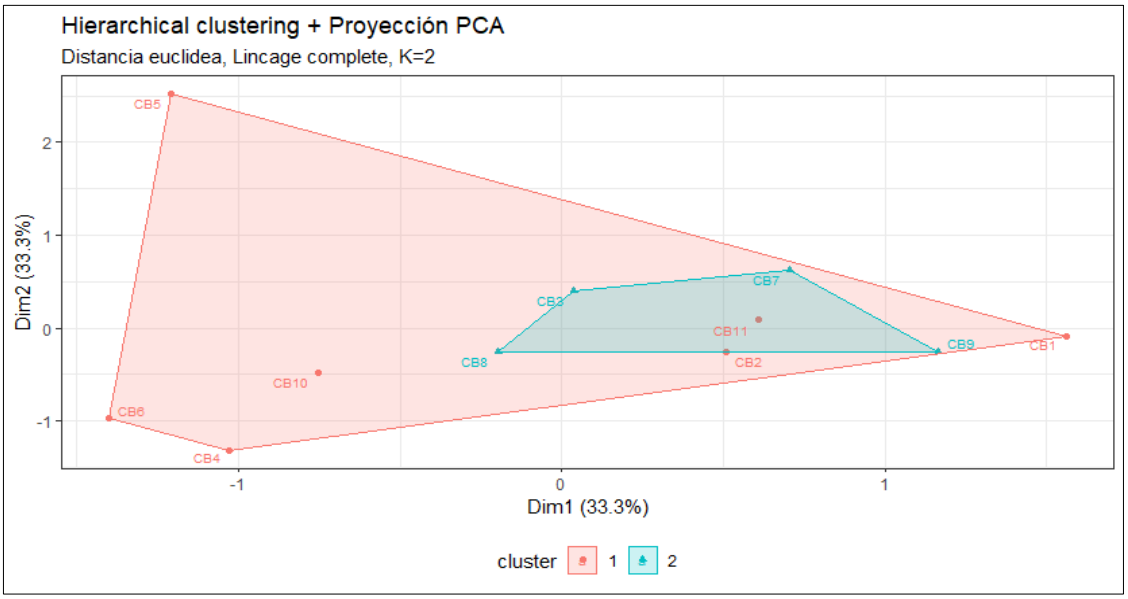


Ilustración 40-4: Clúster del cantón Chambo.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 40-4 que muestran los clústeres de la tdrg en la cual se observa un coeficiente de aglomeración del 0.82 en el primer clúster (7 dosis) y del 0.12 en el segundo clúster (4 dosis). La mayoría de las tdrg se concentran en el clúster 1.

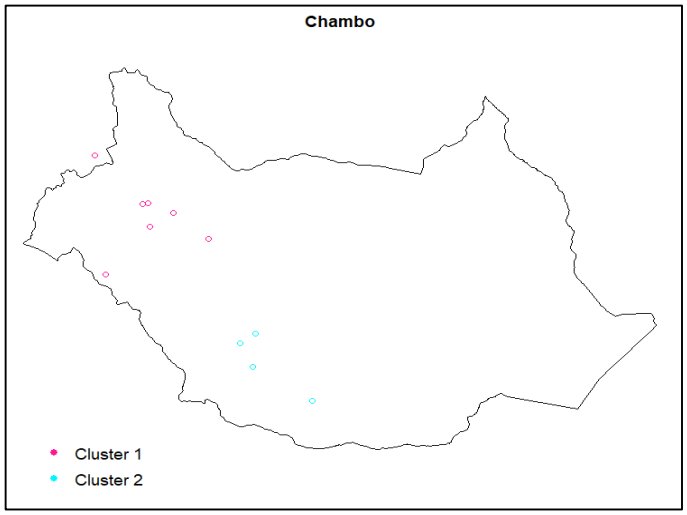


Ilustración 41-4: Mapa del cantón Chambo

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 41-4 que muestra el mapa en el cual está dividido los clústeres de esta manera ya que los datos fueron tomados en distintos lugares del cantón por lo cual se puede ver la influencia que tiene la ubicación geografía en los clústeres mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de esta manera.

4.2.11. Guano

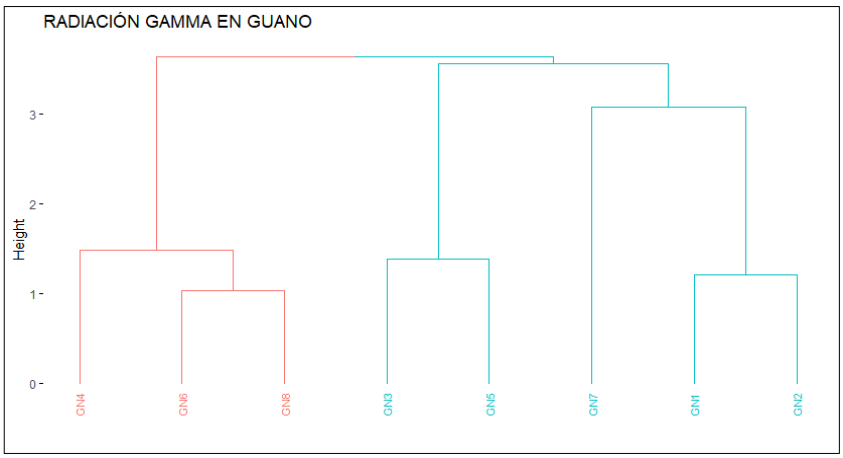


Ilustración 42-4: Dendrograma del cantón Guano.

Realizado por: Erazo, Wilson, 2022.

La ilustración 42-4 explica la distribución de 2 clúster. Se observa una mayor concentración de tdrg en el segundo conglomerado. Además, se observa similaridad entre las tdrg de los datos en estudio, evidenciando cortes proporcionados entre las dos agrupaciones, siendo la razón por la que se intersecan varios puntos ya que presentan características semejantes de ubicación geográfica y de la tdrg.

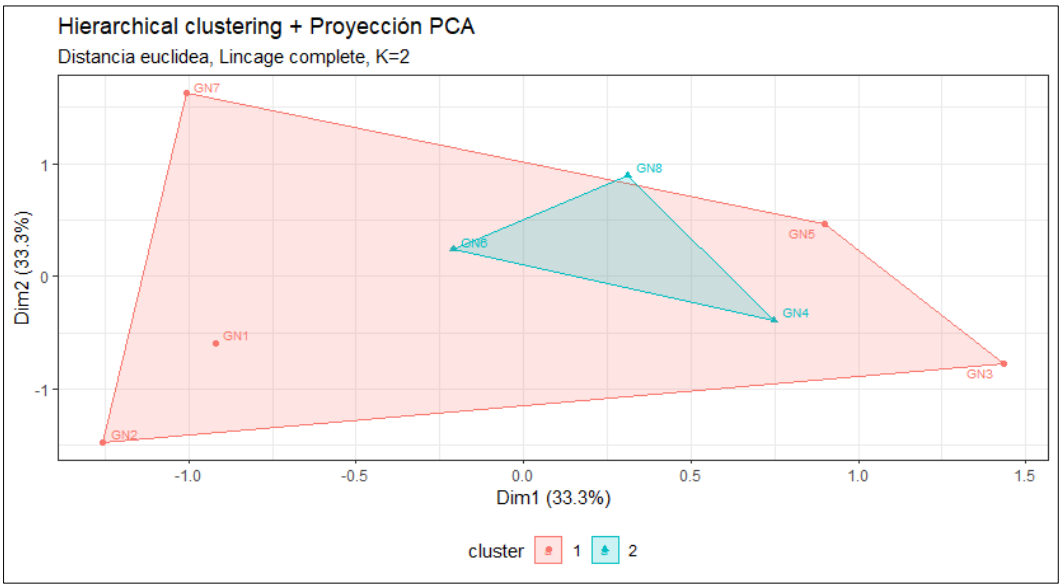


Ilustración 43-4: Clúster del cantón Guano.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 43-4 se muestran los clústeres de tdrg de los datos en estudio el cual muestra un coeficiente de aglomeración del 0.38 en el primer clúster (3 dosis) y del 0.62 en el segundo (5 dosis), en la cual se observa que existe intersección varios puntos ya que presentan características semejantes de ubicación geográfica y de la tdrg.

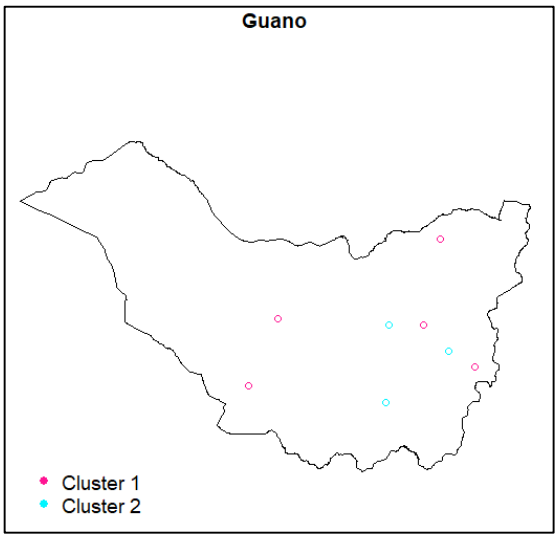


Ilustración 44-4: Mapa del cantón Guano.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 44-4 que muestra el mapa en el cual está dividido los clústeres de esta manera, ya que los datos fueron tomados en distintos lugares del cantón por lo cual se puede ver la influencia que tiene la ubicación geografía en los clústeres mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de esta manera.

4.2.12. Penipe

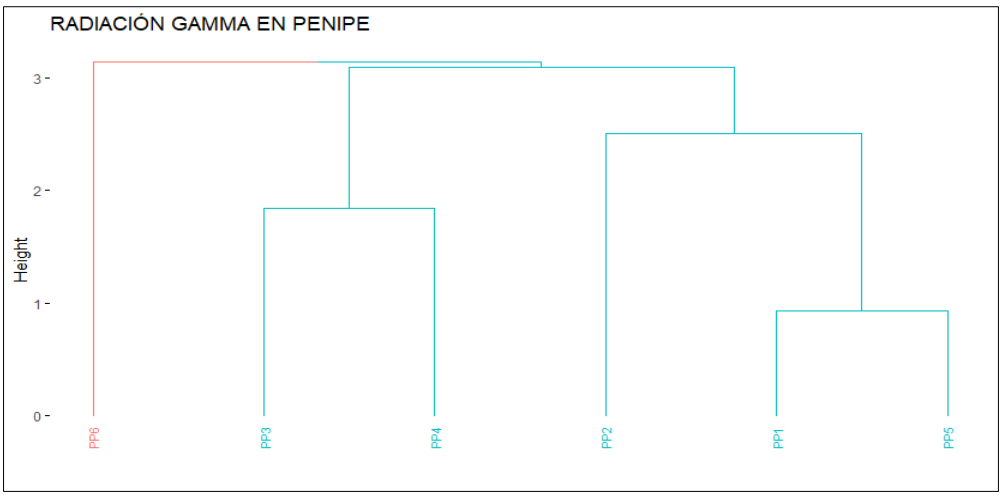


Ilustración 45-4: Dendrograma del cantón Penipe.

Realizado por: Erazo, Wilson, 2022.

La ilustración 45-4 explica la distribución de 2 clúster. Se observa una mayor concentración de tdrg en el segundo conglomerado. Además, se puede observar que existe similaridad entre las tdrg de los datos en estudio, evidenciando cortes proporcionados entre las dos agrupaciones, siendo la razón por la que se intersecan varios puntos ya que presentan características semejantes de ubicación geográfica y de la tdrg.

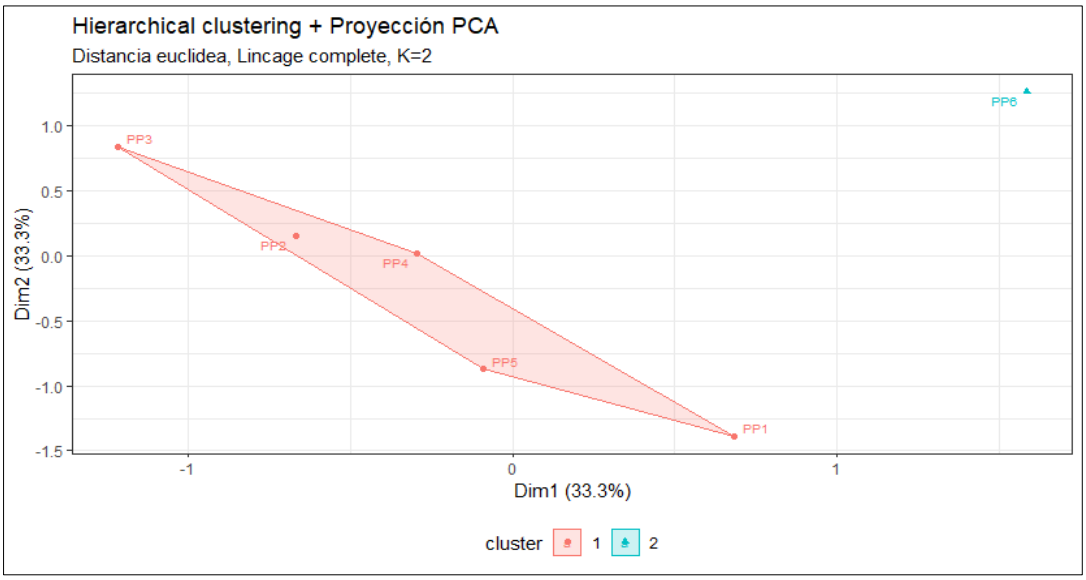


Ilustración 46-4: Clúster del cantón Penipe.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 46-4 que muestran los clústeres de tdrg de los datos en estudio en el cual se observa un coeficiente de aglomeración del 0.17 en el primer clúster (1 dosis) y del 0.83 en el segundo (5 dosis), en la cual se observa que existe intersección varios puntos ya que presentan características semejantes de ubicación geográfica y de la tdrg.

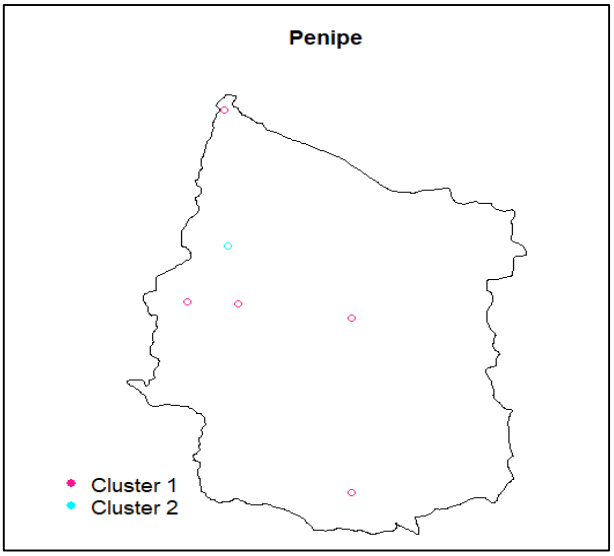


Ilustración 47-4: Mapa del cantòn Penipe.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 47-4 que muestra el mapa en el cual está dividido los clústeres de esa manera, ya que los datos fueron tomados en distintos lugares del cantón por lo cual se puede ver la influencia que tiene la ubicación geografía en los clústeres mediante el método jerárquico aglomerativo los puntos se dividen y se agrupan de esta forma.

4.3. Mapeo de la tdrg en la provincia de Chimborazo

4.3.1. Variograma

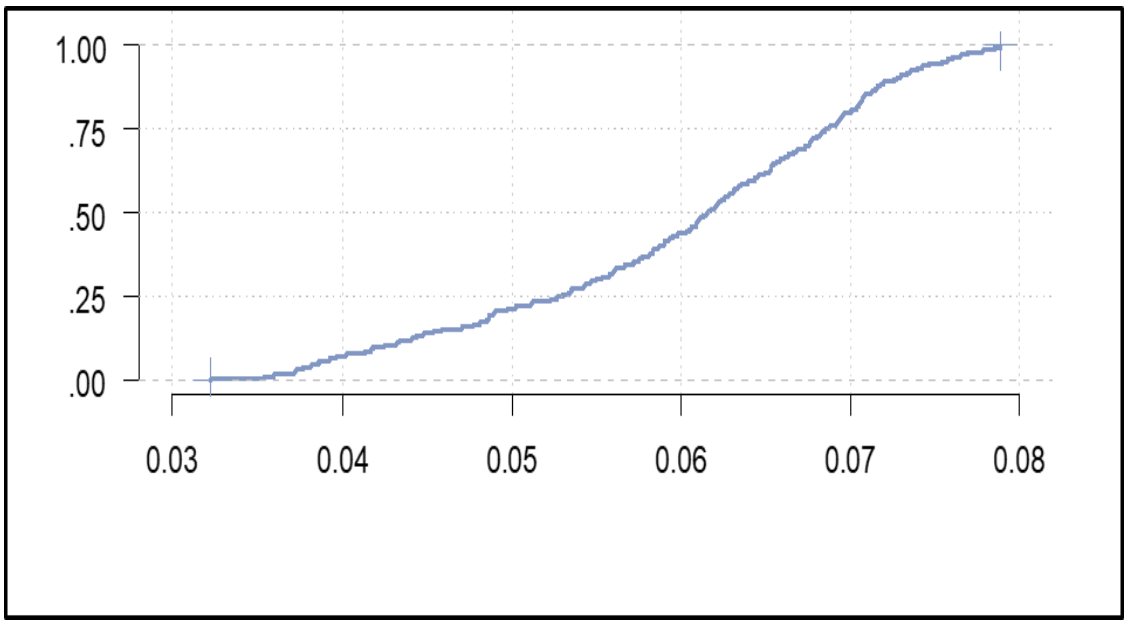


Ilustración 48-4: Variograma

Realizado por: Erazo, Wilson, 2022.

La ilustración 48-4 muestra el variograma con el fin de identificar el método óptimo para la interpolación.

4.3.2. Interpolación de la tdrg en la provincia de Chimborazo

- **Provincial**

Se realizó la interpolación IDW y Kriging con el propósito para identificar el menor error en las estimaciones de las tdrg.

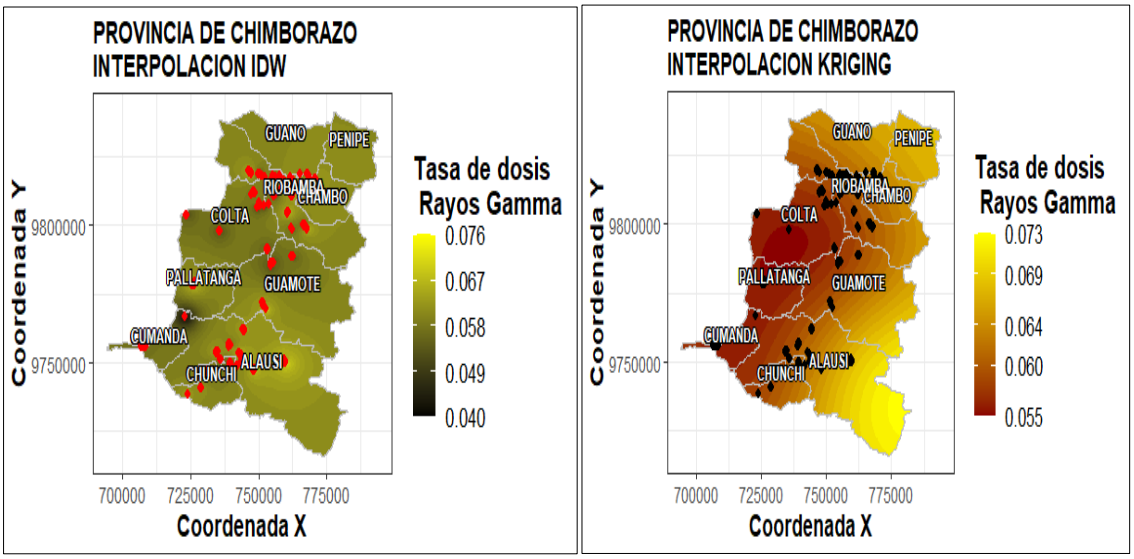


Ilustración 49-4: Interpolación IDW y Kriging de la tdrg en la provincia de Chimborazo.

Realizado por: Erazo, Wilson, 2022.

La ilustración 49-4 muestra la interpolación de la tdrg con la Distancia Inversa Ponderada (IDW), se comporta de manera homogénea y uniforme mientras que con el Kriging se identifica que la distribución de la tdrg es heterogénea en ciertos sectores de la provincia de Chimborazo, siendo Chambo el cantón con los valores más altos.

4.4. Cantonal

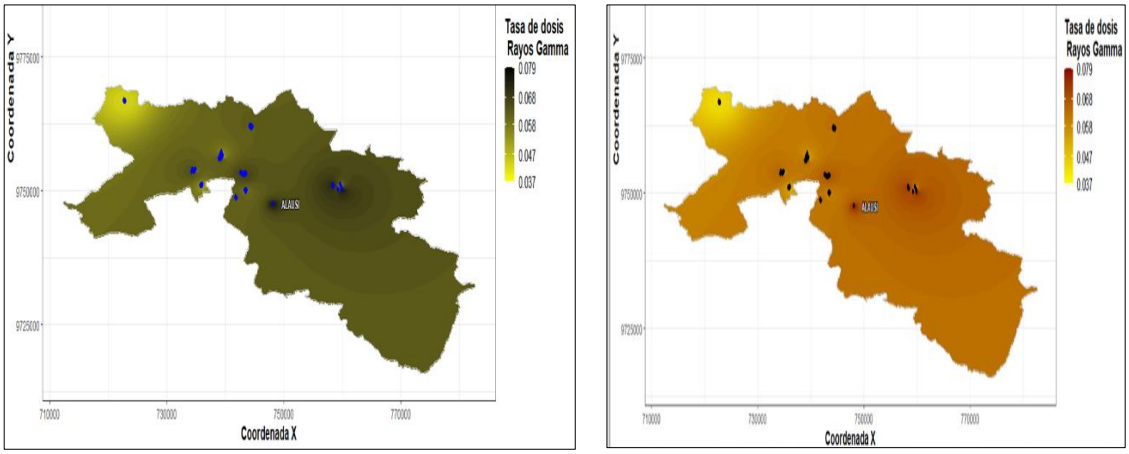


Ilustración 50-4: Interpolación IDW y Kriging de la tdrg en el cantón Alausí.

Realizado por: Erazo, Wilson, 2022.

La ilustración 50-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg, la cual se comporta de manera homogénea y uniforme con un valor mínimo de 0.037 y máximo de 0.079 Sv en el cantón Alausí, mientras que en la parte derecha se encuentra la interpolación de Kriging donde esta es heterogénea en ciertos sectores del cantón, determinando que Kriging expone mejores resultados basados en probabilidades.

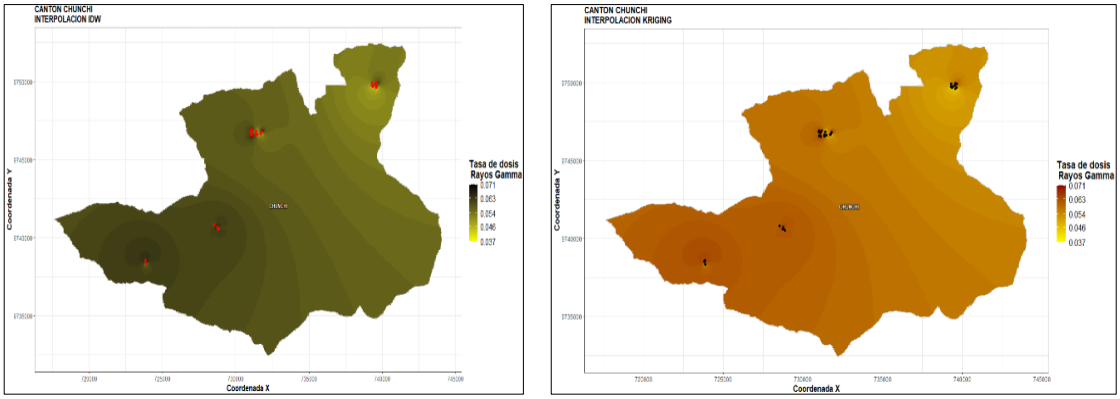


Ilustración 51-4: Interpolación IDW y Kriging de la tdrg en el cantón Chunchi.

Realizado por: *Erazo, Wilson, 2022.*

Mediante la ilustración 51-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg se comporta de manera homogénea y uniforme el cual presenta un valor mínimo de 0.037 y máximo de 0.071 Sv en el cantón Chunchi, mientras que en la parte derecha se encuentra la interpolación de Kriging donde esta presenta tonalidades uniformes, determinando que Kriging expone resultados óptimos para el estudio.

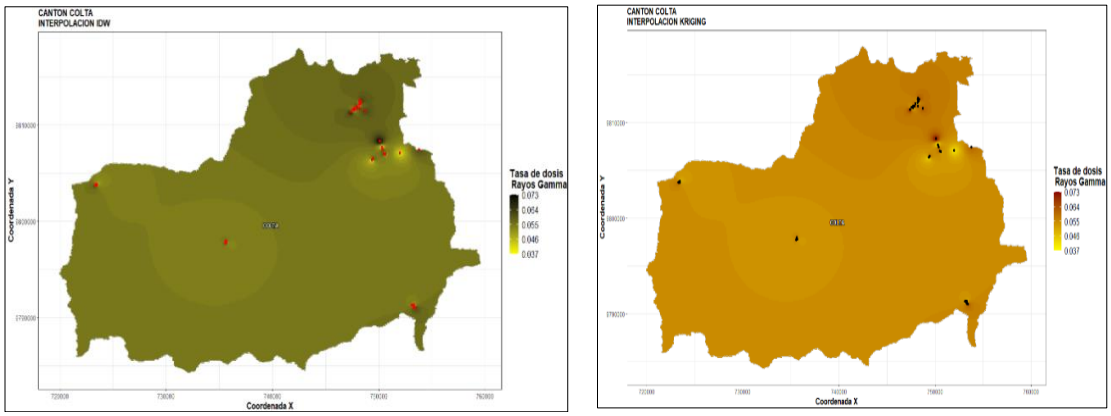


Ilustración 52-4: Interpolación IDW y Kriging de la tdrg en el cantón Colta.

Realizado por: Erazo, Wilson, 2022.

La ilustración 52-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg en la cual se aprecia observar de manera heterogénea en varios sectores del cantón con un valor mínimo de 0.037 y máximo de 0.073 Sv en el cantón Colta, mientras que en la parte derecha se encuentra la interpolación de Kriging esta presenta tonalidades uniformes dentro del mapa la cual expone mejores resultados.

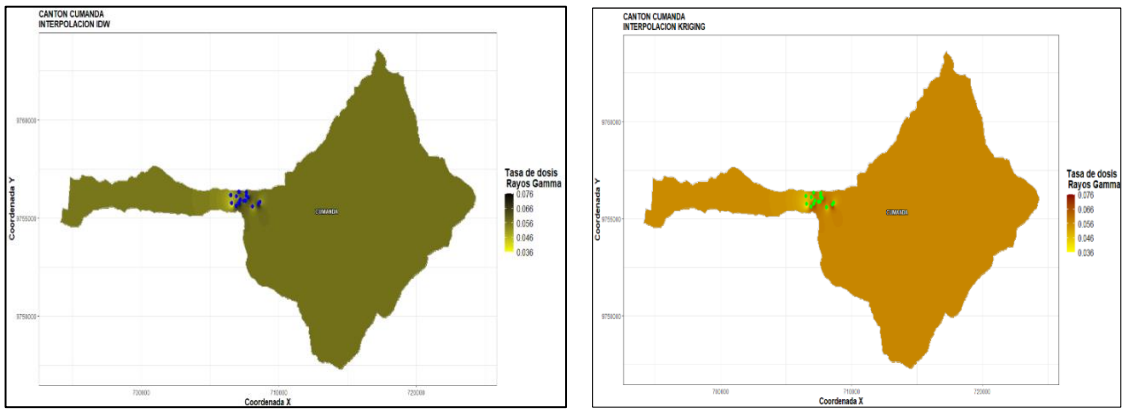


Ilustración 53-4: Interpolación IDW y Kriging de la tdrg en el cantón Cumandá.

Realizado por: Erazo, Wilson, 2022.

La ilustración 53-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg, siendo su comportamiento de manera homogénea y uniforme que presenta un valor mínimo de 0.036 y máximo de 0.076 Sv en el cantón Cumandá, mientras que en la parte derecha se encuentra la interpolación de Kriging el cual presenta uniformidad dentro del mapa.

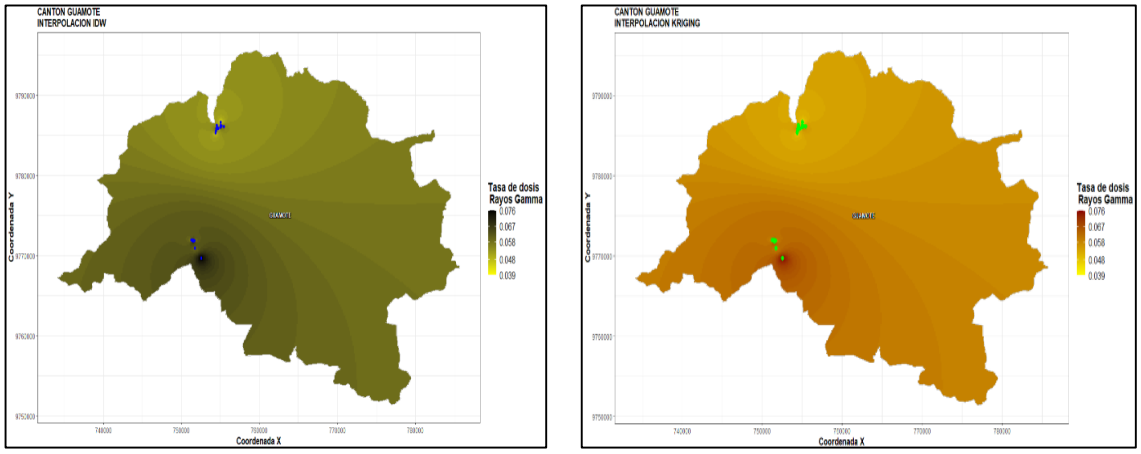


Ilustración 54-4: Interpolación IDW y Kriging de la tdrg en el cantón Guamote.

Realizado por: Erazo, Wilson, 2022.

La ilustración 54-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg permite apreciar su comportamiento de manera homogénea y uniforme el cual presenta un valor mínimo de 0.039 y máximo de 0.076 Sv en el cantón Guamote, mientras que en la parte derecha se encuentra la interpolación de Kriging el cual no se aprecia mejores resultados.

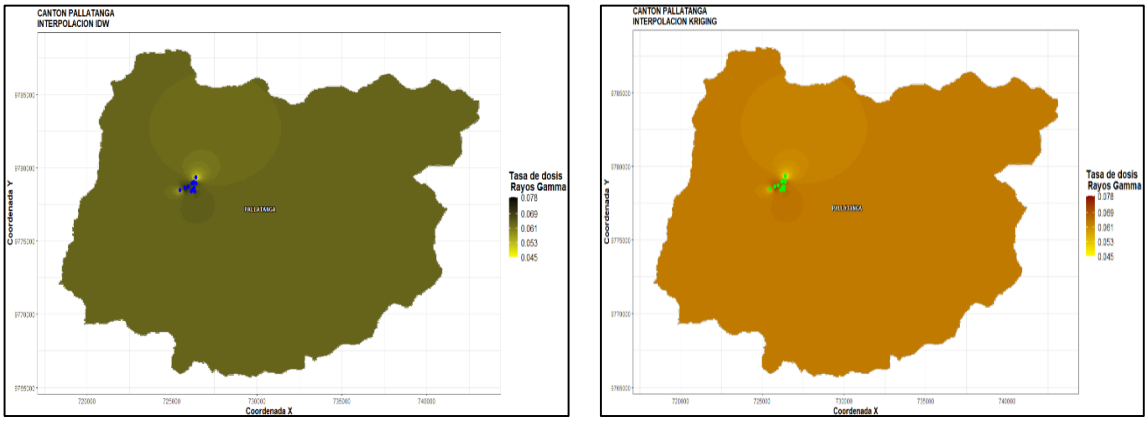


Ilustración 55-4: Interpolación IDW y Kriging de la tdrg en el cantón Pallatanga.

Realizado por: Erazo, Wilson, 2022.

La ilustración 55-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg en la cual se visualiza de manera homogénea y uniforme que presentan un valor mínimo de 0.045 y máximo de 0.078 Sv en el cantón Pallatanga, mientras que en la parte derecha se encuentra la interpolación de Kriging donde las tonalidades dentro del mapa presentan uniformidad dando así resultados mucho más confiables de manera probabilística.

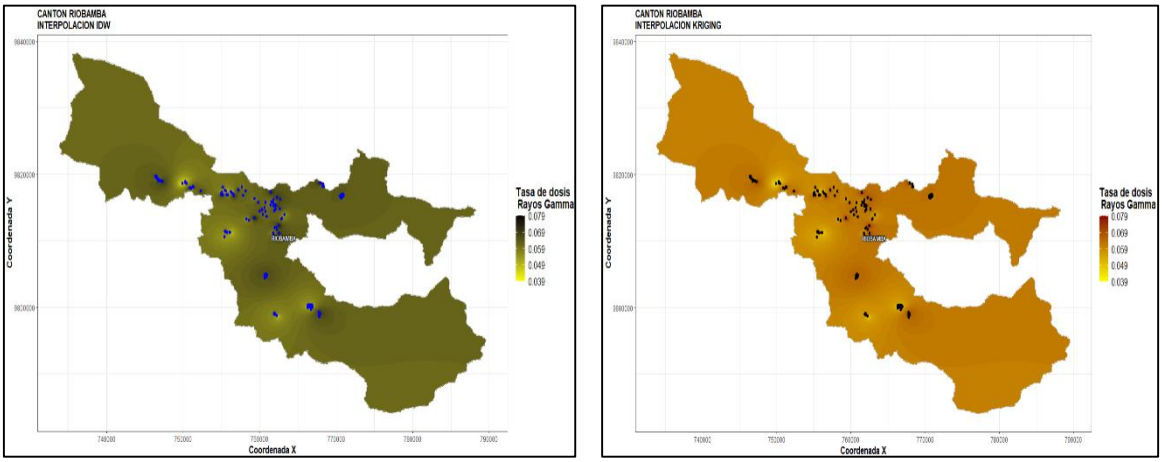


Ilustración 56-4: Interpolación IDW y Kriging de la tdrg en el cantón Riobamba.

Realizado por: Erazo, Wilson, 2022.

La ilustración 56-4 muestra en la parte izquierda la interpolación con la Distancia Inversa Ponderada (IDW) de la tdrg, la cual se comporta de manera homogénea y uniforme que presentan una tdrg mínima de 0.039 y máxima de 0.079 Sv en el cantón Riobamba, mientras que en la parte derecha se encuentra la interpolación de Kriging donde las tonalidades dentro del mapa presentan uniformidad, determinando que Kriging expone mejores resultados basados en probabilidades.

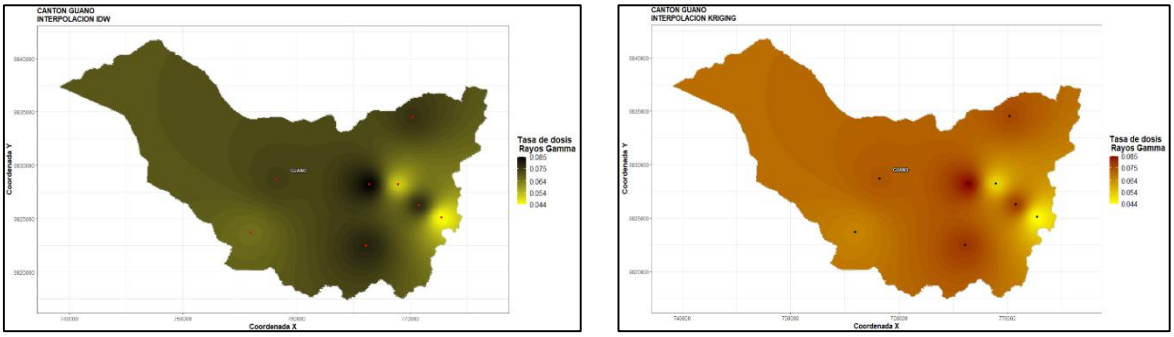


Ilustración 57-4: Interpolación IDW y Kriging de la tdrg en el cantón Guano.

Realizado por: Erazo, Wilson, 2022.

Mediante la ilustración 57-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg, se comporta de manera heterogénea en varios sectores del cantón, con un valor mínimo de 0.044 y máximo de 0.085 Sv en el cantón Guano, mientras que en la parte derecha se encuentra la interpolación de Kriging el cual presenta uniformidad dentro del mapa, arrojando mejores basados en probabilidades.

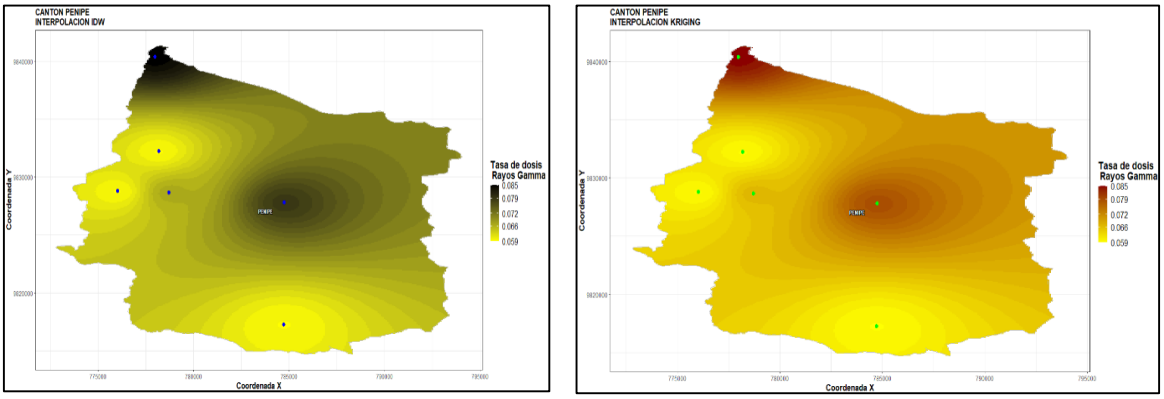


Ilustración 58-4: Interpolación IDW y Kriging de la tdrg en el cantón Penipe.

Realizado por: Erazo, Wilson, 2022.

En la ilustración 58-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) de la tdrg permite visualizar de manera heterogénea en varios sectores del cantón, con un valor mínimo de 0.059 y máximo de 0.085 Sv en el cantón Penipe, mientras que en la parte derecha se encuentra la interpolación de Kriging donde las tonalidades dentro del mapa presentan una estimación optima, determinando que Kriging expone mejores resultados basados en probabilidades.

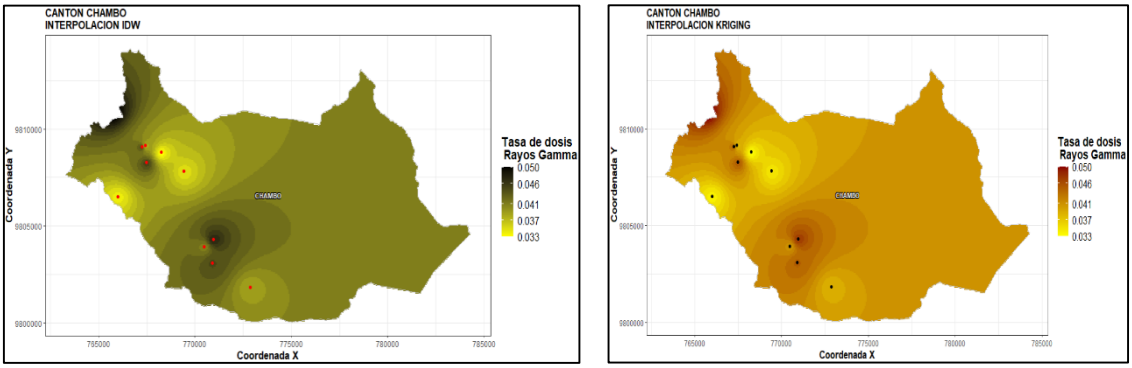


Ilustración 59-4: Interpolación IDW y Kriging de la tdrg en el cantón Chambo.

Realizado por: Erazo, Wilson, 2022.

La ilustración 59-4 muestra en la parte izquierda la interpolación Distancia Inversa Ponderada (IDW) la cual se comporta de manera heterogénea en ciertos sectores del cantón, con un valor mínimo de 0.033 y máximo de 0.055 Sv en el cantón Chambo, mientras que en la parte derecha se encuentra la interpolación de Kriging se comporta de igual manera al IDW, determinando que IDW expone mejores resultados.

Tabla 10-4: Errores de métodos de interpolación de la provincia de Chimborazo

CANTON	IDW	KRIGING	DECISIÓN
ALAUSI	-5.83E-04	-5.83E-04	KRIGING
CHUNCHI	6.73E-04	2.12E-05	KRIGING
COLTA	9.37E-04	3,26E-06	KRIGING
CUMANDA	-1.57E-03	-1.44E-04	KRIGING
GUAMOTE	-4.42E-04	3.07E-05	IDW
PALLATANGA	-1.48E-03	-2.71E-04	KRIGING
RIOBAMBA	2.01E-04	-1.27E-05	KRIGING
GUANO	1.15E-03	1.35E-04	KRIGING
PENIPE	2.75E-03	-1.21E-17	KRIGING
CHAMBO	-2.75E-04	-2.59E-05	IDW

Realizado por: Erazo, Wilson, 2022.

En la tabla 10-4 se compara los residuos de las interpolaciones IDW y Kriging de los cantones de la provincia de Chimborazo, en la cual se refleja como resultado, que el método de Kriging presenta menor error, siendo el óptimo para la interpolación de la tdrg de nuestro estudio.

CONCLUSIONES

- En el análisis exploratorio, mediante la estadística tradicional se encontró que las tasas de dosis de radiación gamma en la provincia de Chimborazo están en un intervalo de 0.032 a 0.079 Sv, con una media de 0.06 Sv y no se identificaron datos atípicos. No existe diferencia significativa entre la tdrg medidas en la mañana y en la tarde, lo mismo sucede en las zonas urbanas y rural; y el espacial mostro la existencia de autocorrelación.
- El índice de Moran determino que las tasas de dosis de radiación gamma en la provincia de Chimborazo se distribuyen aleatoriamente, con presencia de autocorrelación espacial positiva perfecta, dando lugar al análisis de clúster.
- En el análisis de clustering jerárquico se observó que la tdrg en la provincia de Chimborazo tiene una distribución homogénea, corroborando con el estadístico de Hopkins se dedujo que los datos utilizados son clusterizables. Las dos primeras componentes principales explicaron un 88.20 % de variabilidad total. La estandarización optimizo el estudio de los clústers. El método silueta identifico que los tdrg en la provincia de Chimborazo se clasifican en dos clústers, siendo su mayor concentración de tdrg en el segundo aglomerado (211 dosis) concentrado al norte.
- En el análisis de clúster jerárquico cantonal los dendrogramas, clusterización y mapeo geográfico identificaron la dependencia espacial, lo que permite formar grupos de acuerdo con las similaridades en función de la ubicación geográfica y tdrg expuesta en la región. Se identifico clústers de las tdrg bien definidas en Chunchi, Colta, Guamote, Chambo, Alausi, Pallatanga y Penipe mientras que Cumandá, Guano y en la ciudad de Riobamba no se identificaron características exclusivas en sus clústers.
- En los cantones Alausi, Chunchi, Colta, Cumandá, Pallatanga, Riobamba, Guano y Penipe el método Kriging optimizó la interpolación de las tdrg, mientras que en los cantones Guamote y Chambo se apreció que el método optimo es el IDW.

RECOMENDACIONES

- Para futuros estudios sería interesante comparar las diferentes técnicas de manejo de clúster para proporcionar información sobre la técnica que mejor se ajuste a las tasas de dosis de rayos gamma.
- Que el GIDAC continue apoyando a los futuros tesistas de la carrera de Estadística en sus diferentes proyectos.
- Buscar estrategias para incorporar análisis estadísticos espaciales en las líneas de investigación de la carrera de Estadística.

BIBLIOGRAFÍA

ABDELILAH, M. "Faults Detection for Photovoltaic Field Based". [En línea], 2020. [Consulta: 29 julio 2022]. Disponible en: https://scholar.google.com.ec/scholar?q=Faults+Detection+for+Photovoltaic+Field+Based+on+K-Means,+Elbow&hl=es&as_sdt=0&as_vis=1&oi=scholart

BATANERO, C y GODINO, J. "Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria". *Suma*, vol. 9, no. 25-31(1991).

BENÍTEZ, L y PRECIADO, R. El clúster: una alternativa para la competitividad de las pymes de banano orgánico en Ecuador. *Conference Proceedings*, UTMACH, (2017).

BORREGO, J. Modelos de regresión para datos espaciales. nº 1 (2018), (Chile) pp. 1-20.

BOTELLA, P y MARTÍNEZ, M. "Instalación e introducción al software estadístico R y la librería R-Commander. Estadística descriptiva". [En línea], 2018. Disponible en: https://www. uv. es/∼ mamtnez/IRCED. pdf.

BUITRAGO, J. "Visualización de clustering espaciotemporal, un entorno interactivo para el aprendizaje no supervisado de datos". 2019.

CARVALHO, A y ALBUQUERQUE, P. Spatial Hierarchical Clustering. *Revista Brasileira de Biometria*. vol. 27, pp. 411-442. (2009).

CELEMÍN, J. Autocorrelación espacial e indicadores locales de asociación espacial: Importancia, estructura y aplicación. *Revista Universitaria de Geografía*. vol. 18, no. 1, pp. 11-31 (2009).

CORREA, P y ISER, I. "Caracterización del flujo de partículas secundarias provocadas por protones y rayos gamma en la atmósfera, para determinar el tipo de detectores de astropartículas adecuados a la región de Riobamba-Ecuador.". 2020.

DAGNINO, J. Análisis de varianza. *Revista chilena de anestesia*. 2014. vol. 43, no. 4, pp. 306-310.

DE CORSO SICILIA, G y RIVERA, M. Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas. *Cuadernos latinoamericanos de administración*. 2017. Vol. 13, no. 25, pp. 92-104.

DONAIRE, J y BLASCO, B. "Pautas espaciales en la variabilidad de las precipitaciones españolas/Special guidelines in the variability of the Spanish rainfall". *Anales de geografía de la Universidad Complutense*. Universidad Complutense de Madrid. 2006. pp. 227.

ESTÉVEZ, R. Interpolación espacial en QGIS: métodos, procesos y evaluación. *geomapik*. [En línea], 2022. [Consulta: 29 julio 2022]. Disponible en: http://www.geomapik.com/analisis-gis/como-realizar-interpolacion-espacial-qgis-metodos/

ESTRELLA, S. Medidas de tendencia central en la enseñanza básica en Chile: análisis de un texto de séptimo año. *Revista Chilena de Educación Matemática (RECHIEM)*. 2008. vol. 4, no. 1, pp. 20-32.

FALLAS, J. Conceptos básicos de cartografía. *Programa Regional en Manejo de Vida Silvestre y Escuela de Ciencias Ambientales. Universidad Nacional. Heredia. Costa Rica*. 2003.

GONZÁLEZ, T y MARIOLY, V. La radiación ultravioleta. Su efecto dañino y consecuencias para la salud humana. *Theoria*. 2009. vol. 18, no. 2, pp. 69-80.

GUTIÉRREZ, O. Diseño y simulación de un detector de astropartículas en la sierra ecuatorial con geant4 para la determinación del área efectiva en función de la energía. Quito, 2020.

HEREDIA, L y DÍAZ, J. Análisis clúster como técnica de análisis exploratorio de registros múltiples en datos meteorológicos. *Ingeniería de Recursos Naturales y del Ambiente*. 2012. no. 11, pp. 11-20.

LOZARES, C y LÓPEZ, P. "El análisis de componentes principales: aplicación al análisis de datos secundarios". *Papers: revista de sociologia*. 1991. no. 37, pp. 031-063.

OSORIO, H y SÁNCHEZ, E. La cartografía como medio investigativo y pedagógico. *Dearq. Revista de Arquitectura*. 2011. no. 9, pp. 30-47.

PRECIADO, G y LARIOS, R. Estrategia Didáctica para el Análisis e Interpretación de un Diagrama de Caja mediante Simulación en MATLAB. *XXIII Semana de Investigación y Docencia en Matemáticas*. 2013. pp. 77.

REYES, N. Índices dinámicos para espacios métricos de alta dimensionalidad. *Universidad Nacional de San Luis, Argentina*. 2002.

RICARDI, Q. Medidas de tendencia central y dispersión. *Revista Biomédica Revisada Por Pares*. 2011. pp. 1-8.

RUEDA, B y GELLES, C. Identificación de patrones de variabilidad climática a partir de análisis de componentes principales, Fourier y clúster k-medias. *Tecnura*. 2016. vol. 20, no. 50, pp. 55-68.

SANTANA, F. El análisis de cluster: aplicación, interpretación y validación. *Papers: revista de sociologia*. 1991. pp. 65-76.

SOTTER, A y SÁNCHEZ, V. Geoestadística aplicada a estudios de contaminación ambiental. *Ingeniería*. 2002. vol. 7, no. 2, pp. 31-38.

TORRES, A. "Los 8 tipos de mapas principales, y sus características". [En línea], 2015. [Consulta: 26 julio 2022]. Disponible en: https://psicologiaymente.com/cultura/tipos-de-mapas

VIDAL, R y FAVARO, P. Low rank subspace clustering (LRSC). *Pattern Recognition Letters*. 2014. vol. 43, pp. 47-61.

YRIGOYEN, C. Modelos de heterogeneidad espacial. *University Library of Munich, Germany, Tech. Rep*. 2004.

ANEXOS

ANEXO A: CÓDIGO DE CLÙSTER

```
library(cluster)
library(stats)
library(MASS)
library(ggplot2)
library(FactoMineR)
library(corrplot)
library(readxl)
library(PerformanceAnalytics)
library(psych)
library(factoextra)
library(DataExplorer)
library(scatterplot3d)
par(mfrow=c(1,1))
datos<-read_xlsx("C:/Users/USER/Desktop/escritorio/ANEXOS
PAUL/Analisis_completo/EXCEL/canton/alausi.xlsx"
                 , sheet = "alausi")
Datoss=datos[-1]
Dat<- scale(Datoss)
row.names(Dat)=datos$Canton
## Test de esferecidad
cortest.bartlett(Dat,n=length(Datoss$doserate))
## Se rechaza que la matriz de correlacion sea la identidad, es factible realizar un pca
## Estandarizacion
pca2<- PCA(X = Dat, scale.unit = TRUE, graph = F)
## Autovalores, varianza acumulada
pca2$eig


## Proporcion explicada
```

```
fviz_screeplot(pca2, addlabels = TRUE, ylim = c(0, 70))

## Las dos primeras componentes explican el 74.71%

###### Variables PCA

fviz_pca_var(pca2, col.var="contrib",

             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),

             repel = TRUE, # Avoid text overlapping

             axes = c(1, 2) # choose PCs to plot

)

fviz_pca(pca2)

fviz_pca_biplot(pca2, label = "var")

## Grafica de las pronvincias con 2 componentes

fviz_pca_ind(

  pca2, col.var="contrib",

  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),

  repel = TRUE, # Avoid text overlapping

  axes = c(1, 2) # choose PCs to plot

)

## Primer componente

d=data.frame(pca2$ind)

d$coord.Dim.1

d$coord.Dim.2

d$coord.Dim.3

## 3D

scatterplot3d(x = d$coord.Dim.1,

              y = d$coord.Dim.2,

              z = d$coord.Dim.3,

              pch = 20, cex.lab = 0.8,

              grid = TRUE, box = TRUE,

              color = "2")

text(d[1:3], labels =(datos$Canton),

     cex= 0.4, col = "steelblue")

## Por componentes
```

```
dat=cbind(d$coord.Dim.1,d$coord.Dim.2,d$coord.Dim.3)
colnames(dat)=c("Componente1","Componente2","Componente3")
dat=as.data.frame(dat)
row.names(dat)=datos$Canton
View(dat)
dat$Componente1<-(dat$Componente1 - mean (dat$Componente1)) / sd (dat$Componente1)
dat$Componente2<-(dat$Componente2 - mean (dat$Componente2)) / sd (dat$Componente2)
dat$Componente3<-(dat$Componente3 - mean (dat$Componente3)) / sd (dat$Componente3)
aa=dist(dat, method = "euclidean")
a=hclust(aa,method = "ward.D2")
a$order
library(factoextra)
set.seed(101)
fviz_dend(a, cex = 0.6, k = 2, horiz = FALSE, main="RADIACI?N GAMMA EN ALAUSI")
clust <- cutree(a, k = 2)
fviz_cluster(list(data = dat, cluster = clust),ellipse.type = "convex", repel = TRUE, show.clust.cent = FALSE,
             labelsize = 8, )+
  labs(title = "Hierarchical clustering + Proyecci?n PCA",
       subtitle = "Distancia eucl?dea, Lincage complete, K=2") +
  theme_bw() +
  theme(legend.position  =  "bottom")
plot(a, cex=.35)
Numgrupos <- 2
rect.hclust(a, k = Numgrupos, border = "red")
g<- cutree(a, Numgrupos)
View(g)
GG = data.frame(g)
View(GG)
library(openxlsx)
write.xlsx(GG, "RY_ALAUSI.xlsx")
Grupo=as.factor(g)
View(Grupo)
```

ANEXO B: CÓDIGO DE MAPAS

```
library(readxl)
library(sp)
library(rgdal)
library(mapview)
library(tidyverse)
library(raster)
library(sf)
#mapa Chimborazo
setwd("~/CLUSTER GAMMA TESIS")
setwd("~/CLUSTER GAMMA TESIS/tesis calculos/Analisis_completo/cantones")
chimb<-st_read("cantones.shp")
setwd("~/CLUSTER GAMMA TESIS/tesis calculos/Analisis_completo/clúster_cantones/cluster1")
c1 <- st_read("cluster1.shp")
setwd("~/CLUSTER GAMMA TESIS/tesis calculos/Analisis_completo/cluster_cantones/cluster2")
c2 <- st_read("cluster2.shp")
plot(st_geometry(chimb),col="white", main="Chimborazo")
plot(st_geometry(c1),col="#FF1493", border="#FF1493", add=TRUE)
plot(st_geometry(c2),col="#00F5FF", border="#00F5FF", add=TRUE)
legend("bottomleft”, pch=16, inset = 0.05,
      box.lwd = 0.05,bty = "n", legend = c("Cluster 1", "Cluster 2"),
      cex = 1.15,
      col = c("#FF1493","#00F5FF"),
      horiz = FALSE,
      xpd = TRUE)
```

ANEXO C: CÓDIGO DE MAPAS DE INTERPOLACIÓN

```
 library(openxlsx)
library(tidyverse)
library(viridis)
library(shadowtext)
library(rgdal)
library(ggplot2)
library(rgdal)
library(gstat)
setwd("~/CLUSTER GAMMA TESIS/tesis calculos/Analisis_completo/bases_interpolacion/canton")
dir()
datosal <- read.xlsx(xlsxFile = "alausi.xlsx",
                     sheet = "alausi", detectDates = TRUE)
#dataframe de los datos
datos2 <- datosal
coordinates(datosal) <- ~x+y#asignar coordenadas espaciales
utm <- "+proj=utm +zone=17 +south +ellps=WGS84 +datum=WGS84 +units=m +no_dsef"
#Asignar sistema de referencia
proj4string(datosal) <- utm
#spatialdataframe de datos
datosal
#||||||||||||||||||||||||cargar mapas base de los cantones de Chimborazo||||||||||||||||||||
#leer un shapefile y convertirlo en spatialdataframe
setwd("~/CLUSTER GAMMA TESIS/tesis calculos/Analisis_completo/Chimborazo_cantones")
dir()
Chimborazo_cantones<-readOGR("Alausi.shp")
# Filtrar dtos de los mapas para la provincia Chimborazo
#cantones <- subset(cantoness, DPA_DESPRO=='CHIMBORAZO')
#head(Alausi@data["DPA_DESCAN"])
Chimborazo_cantones@data["DPA_DESCAN"]
```

```
#convertir el spatialdataframe a dataframe

spdf_fortified1 <- fortify(Chimborazo_cantones, region = "DPA_DESCAN")

#unir el dataframe de las observaciones y el dataframe de los cantones

spdf_fortified1 = spdf_fortified1 %>%

  left_join(. , datosal@data, by=c("id"="Parroquia"))

#cambiar de nombre la base

map_base <- spdf_fortified1

#filtar nombres y centroides de los cantones

namescan <- data.frame(caname=unique(Chimborazo_cantones@data$DPA_DESCAN))

dtcan <- merge(Chimborazo_cantones,namescan,by.x="DPA_DESCAN",by.y="caname")

dtcan <-dtcan[is.na(dtcan$DPA_DESPRO)==FALSE,]

head(dtcan@data)

#dt <- dtcan[,c(1,8:length(dtcan))]

dt <- cbind(namescan,coordinates(dtcan))

colnames(dt)[2:3] <- c("x","y")

est_base <- dt

est_base

#nombres de los cantones

dt <- est_base

dbbp <- dt

dbbp <- dbbp%>%

  mutate( name=factor(caname, unique(caname))) %>%

  #mutate( mytext=paste(ESTACION, "\n",varprom, sep="")

  mutate( mytext=paste(caname, "\n",sep="")

  )

point_names <- dbbp

point_names

spdf_fortified <- map_base

dbbp <- point_names#nombres de las observaciones

#unidades <- un_med

scpts <- seq(min(datos2$doserate), max(datos2$doserate), length.out = 5)

scpts <- as.numeric(round(scpts,2))
```

```
unidades <- "Tasa de dosis\n Rayos Gamma"
titulo <- "CANTON ALAUSI"
p <- ggplot() +
  geom_polygon(data = spdf_fortified, aes(x=long, y = lat, group=group), color="grey",fill="grey", alpha=0.3) +
  #geom_point(data= datos2,aes(x=x, y=y, size=doserate),color="brown",stroke=FALSE) +
  geom_point(data= datos2,aes(x=x, y=y, color=doserate),size=3,stroke=FALSE) +
  #scale_fill_continuous(low = "green", high = "red",breaks=scrtr, guide="colorbar",na.value="white") +
  scale_size_continuous(breaks=scpts) +
  scale_alpha_continuous(breaks=scpts) +
  #scale_alpha_continuous(breaks=scrtr) +
  scale_color_gradient(breaks=scpts,low = "green", high = "darkred") +
  guides( colour = guide_legend()) +
  geom_shadowtext(data= dbbp,aes(x=x, y=y, label=mytext),
                  fontface = "bold",size=3.2) +
  theme_bw() +
  labs(title= titulo, x="Coordenada X", y="Coordenada Y",
       color=paste0(unidades),
       size=paste0(unidades),
       fill=paste0(unidades),
       alpha=paste0(unidades)) +
  theme(text = element_text(size=10),
        title = element_text(face="bold",size=rel(1.2)),
        axis.title.x = element_text(face="bold",size=rel(1.2)),
        axis.title.y = element_text(face="bold",size=rel(1.2)),
        legend.title = element_text(size=rel(1.3)),
        legend.text = element_text(size=rel(1.3)),
        axis.text.x = element_text(size=rel(1.3)),
        axis.text.y = element_text(size=rel(1.3)))
p
#crear regillas de forma aleatoria para toda la provincia
#sptdf <- sptdfMean
# grid para prediccion
```

```
#sptdf_grid = spsample(parro_chimb, type = "regular", cellsize = c(1000,1000))

sptdf_grid1  <- spsample(Chimborazo_cantones, "regular", n=50000)

gridded(sptdf_grid1) = TRUE

proj4string(sptdf_grid1) <- proj4string(datosal)

grid_pred1 <-sptdf_grid1

head(datos2)

#IDW <- reactive({

#sptdf <- sptdfMean

sptdf_grid1 <- grid_pred1

idw_p1 <- idw(doserate ~ 1, datosal, sptdf_grid1)

IDW1 <- idw_p1

IDW1

view(IDW1)

#output$IDWPlot <- renderPlot({

#spdf_fortified <- map_base

#unidades <- un_med

#dbbp <- point_names

idw.d1<- IDW1

df1 <- data.frame(idw.d1)

head(df1)

#redondea las predicciones para formar isolineas (capas)

#ya que los valores estan por debajo de 1 podria redondearse a 3 decimales

predaprox <- round(df1$var1.pred,digits = 3)

df1$predaprox <- predaprox

#secuencia de datos para la leyenda (5 intervalos)

scpts <- seq(min(datos2$doserate), max(datos2$doserate), length.out = 5)

scpts <- as.numeric(round(scpts,3))

scrtr <- seq(min(df1$var1.pred), max(df1$var1.pred), length.out = 5)

scrtr <- as.numeric(round(scrtr,3))

titulo <- "CANTON ALAUSI\nINTERPOLACION IDW"

unidades <- "Tasa de dosis\n Rayos Gamma"

#Interpolacion IDW sin redondear debido a que los valores son menores a 1
```

```
#Nota: las isolineas se aprecian mejor redondenando las predicciones a 3 decimales
idwplot1 <-  ggplot() + geom_raster(data=df1, aes(x = x1, y = x2, fill = predaprox)) +
  scale_fill_continuous(low = "yellow", high = "black",breaks=scrtr, guide="colorbar",na.value="white") +
  scale_alpha_continuous(breaks=scrtr) +
  geom_polygon(data = spdf_fortified1, aes(x=long, y = lat, group=group), color="grey",fill="grey", alpha=0) +
  #los puntos iniciales ya no van parece
  geom_point(data= datos2,aes(x=x, y=y),size=2,color="blue",stroke=FALSE) +
  scale_size_continuous(breaks=scpts) +
  #scale_alpha_continuous(breaks=scpts) +
  scale_color_gradient(breaks=scpts,low = "green", high = "red") +
  guides( colour = guide_legend()) +
  geom_shadowtext(data= dbbp,aes(x=x, y=y, label=mytext),
                  fontface = "bold",size=3.2)+
  #scale_fill_gradient(high = "yellow", low = "white")+
  #scale_color_viridis(option=sample(LETTERS[2:5],1)) +
  theme_bw() +
  #theme_void()+
  labs(title= titulo, x="Coordenada X", y="Coordenada Y",
       color=paste0(unidades),
       size=paste0(unidades),
       fill=paste0(unidades),
       alpha=paste0(unidades)) +
  theme(text = element_text(size=10),
        title = element_text(face="bold",size=rel(1.2)),
        axis.title.x = element_text(face="bold",size=rel(1.2)),
        axis.title.y = element_text(face="bold",size=rel(1.2)),
        legend.title = element_text(size=rel(1.3)),
        legend.text = element_text(size=rel(1.3)),
        axis.text.x = element_text(size=rel(1.3)),
        axis.text.y = element_text(size=rel(1.3)))
idwplot1
#interpolacion Kriging
```

```
auto <- autoKrige(doserate ~ 1, datosal, sptdf_grid1)

auto$krige_output

#spdf_fortified <- map_base

#unidades <- un_med

#dbbp <- point_names

ok.d<- auto$krige_output

df1 <- data.frame(ok.d)

predaprox <- round(df1$var1.pred,digits = 3)

#varaprox <- round(df1$var1.var,digits = 3)

df1$predaprox <- predaprox

#df1$varaprox <- varaprox

scpts <- seq(min(datos2$doserate), max(datos2$doserate), length.out = 5)

scpts <- as.numeric(round(scpts,3))

scrtr <- seq(min(df1$var1.pred), max(df1$var1.pred), length.out = 5)

scrtr <- as.numeric(round(scrtr,3))

titulo <- "CANTON ALAUSÍ\nINTERPOLACION KRIGING"

unidades <- "Tasa de dosis\n Rayos Gamma"

#ag1 <- "Predicciones de\n"

ag1 <- "Valor"

input <- input$var_sel2

ag2 <- "\n(en "

#unidades <- c("C","%","KJ","Kj","C","cm/h","grados")

ag3 <- ")"

ag4 <- "Valores de\n"

ag5 <- "Intensidad de\n"

krigingplot1 <- ggplot() + geom_raster(data=df1, aes(x = x1, y = x2, fill = predaprox)) +

  scale_fill_continuous(breaks=scrtr, low = "yellow", high = "darkred",guide="colorbar",na.value="white") +

  #scale_alpha_continuous(breaks=scrtr) +

  geom_polygon(data = spdf_fortified, aes(x=long, y = lat, group=group), color="grey",fill="grey", alpha=0) +

  geom_point(data= datos2,aes(x=x, y=y),size=2,color="black",stroke=FALSE) +

  scale_size_continuous(breaks=scpts) +

  scale_alpha_continuous(breaks=scpts) +
```

```
  scale_color_gradient(breaks=scpts,low = "blue", high = "green") +
  guides( colour = guide_legend()) +
  geom_shadowtext(data= dbbp,aes(x=x, y=y, label=mytext),
                  fontface = "bold",size=3.2)+
  #scale_fill_gradient(high = "yellow", low = "white")+
  #scale_color_viridis(option=sample(LETTERS[2:5],1)) +
  theme_bw() +
  #theme_void()+
  labs(title= titulo, x="Coordenada X", y="Coordenada Y",
       color=paste0(unidades),
       size=paste0(unidades),
       fill=paste0(unidades)) +
  theme(text = element_text(size=10),
        title = element_text(face="bold",size=rel(1.2)),
        axis.title.x = element_text(face="bold",size=rel(1.2)),
        axis.title.y = element_text(face="bold",size=rel(1.2)),
        legend.title = element_text(size=rel(1.3)),
        legend.text = element_text(size=rel(1.3)),
        axis.text.x = element_text(size=rel(1.3)),
        axis.text.y = element_text(size=rel(1.3)))
krigingplot1
```

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 23 / 01 / 2023

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: Wilson Paul Erazo Salao
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Estadística
Título a optar: Ingeniero en Estadística Informática
f. Analista de Biblioteca responsable: Ing. Rafael Inty Salto Hidalgo