



UNIVERSIDAD NACIONAL DE COLOMBIA

# Muestreo de Estructuras de Redes en Datos no Estructurados

Luis David Velásquez Tafur

Universidad Nacional de Colombia  
Facultad de Ciencias  
Departamento de Estadística - Perfil de profundización  
Bogotá, Colombia  
2023

# Muestreo de Estructuras de Redes en Datos no Estructurados

Luis David Velásquez Tafur

Trabajo de grado presentado como requisito parcial para optar al título de:  
**Magíster en Estadística - Profundización**

Director:

Leonardo Trujillo Oyola, PhD

Codirector:

Joaquin Guillermo Ramirez Gil, PhD

Universidad Nacional de Colombia

Facultad de Ciencias

Departamento de Estadística - Perfil de profundización

Bogotá, Colombia

2023

## Dedicatoria

*A mis padres y a mi hermana.*

# Agradecimientos

Agradezco de corazón a mi mamá, Ángela, y a mi hermana, Karen, por su apoyo incondicional y comprensión a lo largo de esta emocionante travesía académica. Gracias por ser mi fuente de inspiración y por alentarme a alcanzar mis metas en todo momento. Sus palabras de aliento y motivación han sido mi impulso para seguir adelante, incluso en los momentos más desafiantes. Agradezco profundamente su presencia constante y su cariño que han hecho posible el desarrollo de este trabajo.

También quiero agradecer a mi padre por todo su apoyo incondicional mientras nos acompañó terrenalmente. Espero que desde donde esté, pueda ver el nuevo logro en mi vida profesional y se sienta orgulloso.

A mi pareja, Vanessa, gracias por ser mi refugio durante este camino. Tu apoyo inquebrantable, paciencia y comprensión han sido fundamentales para mantener el equilibrio. Cada palabra de aliento ha sido una inspiración para seguir adelante y no rendirme en esta etapa de crecimiento académico y profesional.

Agradezco a mis directores, Leonardo Trujillo Oyola y mi Codirector Joaquín Guillermo Ramírez Gil por su valiosa orientación y guía a lo largo de este proceso de investigación. Sus conocimientos, experiencia y consejos han sido fundamentales en la elaboración de este trabajo. Agradezco la confianza que depositaron en mí y por aceptar la invitación a esta aventura que representó esta investigación de gran interés.

También quiero agradecer a todas las personas que facilitaron la realización de este trabajo, tales como el grupo de investigación del profesor Joaquín, en especial, Viviana Rodríguez, estudiante de la Maestría en Ciencias-Climatología, por su apoyo clave en la recolección de datos para el desarrollo de este trabajo, y también al grupo técnico de Fedearroz que nos permitió usar la información que tenían disponible para poner a prueba todas las ideas que teníamos en mente.

Como se puede ver, este no es un trabajo que haya nacido en solitario, es el resultado de muchos esfuerzos y colaboraciones para buscar una solución a un problema conjunto. Esto hace que esta investigación sea mucho más especial, puesto que no se quedará solamente escrito en este papel, sino que trascenderá a unas aplicaciones prácticas donde se pueda solucionar los problemas existentes y pueda dar cabida a la creación de nuevas preguntas.

Finalmente, quiero expresar mi gratitud a la Universidad Nacional de Colombia, la institución que me acogió durante el pregrado y ahora la maestría. A todos los profesores y compañeros que allí conocí, gracias por contribuir a mi formación y por hacer de este periodo de estudio una experiencia enriquecedora y significativa.

Por último y no menos importante, agradezco a la Dirección de Investigación y Extensión de la sede Bogotá, DIEB, por la financiación del proyecto "Herramientas de Big Data Aplicadas al Análisis Epidemiológico en Sistemas de Producción de Arroz Bajo Condiciones Climáticas Adversas" de Fedearroz, y a mis distintos empleadores como el Banco Itaú y Esri-Colombia que me apoyaron tanto económicamente como con el tiempo para cumplir con todos los requerimientos de este programa académico.

A todos, mi más profundo agradecimiento. Este logro es también de ustedes y llevaré en mi corazón cada enseñanza, cada palabra de aliento y cada apoyo recibido dentro de mi camino profesional.

# Resumen

Teniendo en cuenta el creciente uso de la tecnología para una mejor lectura del mundo que nos rodea, el presente trabajo busca dar respuesta a un problema práctico en la industria fitosanitaria de la producción de arroz mediante la implementación de una metodología de muestreo estadístico en redes. Para esto, se aborda dicha problemática desde una investigación y análisis detallado de diversos métodos para realizar muestreo en redes y, por otro lado, la aplicación de dichos métodos enfocados a la optimización de muestreo fitosanitario en el cultivo de arroz. Dentro de los métodos estudiados para la selección de muestras, se destacaron algunos métodos tradicionales como el muestreo aleatorio simple y el estimador Horvitz Thompson. Adicionalmente, se analizaron diferentes herramientas estadísticas como los métodos de clasificación no supervisados y el método de estimación Monte Carlo que son clave en el desarrollo de este estudio. Del mismo modo, este análisis inicial contempló los muestreos de redes, donde se pueden seleccionar nodos o conexiones específicas para la recolección de datos y se contemplaron los métodos de muestreo basados en caminatas aleatorias y muestreos para grafos conectados que usan información de enlaces para seleccionar la muestra de un manera más eficiente. Posteriormente, dichos conceptos de corte estadístico se utilizaron en la optimización de muestreo fitosanitario en los cultivos de arroz. Se trabajó con los datos suministrados por Fedearroz para poder hacer la aplicación de los diferentes métodos estadísticos estudiados y la creación de redes respectivas para posteriormente contrastar los resultados basados en listas que es la forma tradicional. Esta comparación usó muestreos basados en tipo red aplicados en un escenario de muestreo estratificado, con estratos creados bajo una metodología de aprendizaje no supervisado utilizando variables relevantes tales como el rendimiento por hectárea, el uso de fertilizantes, las condiciones climáticas y las características edáficas del cultivo. Con base en las muestras recopiladas a través de diferentes métodos de muestreo tradicionales y de grafos, se logró realizar un estimado sobre la incidencia de la enfermedad de mayor relevancia en el cultivo de arroz en Colombia, la Piricularia (*Pyricularia Oryzae*). Asimismo, se exploraron las relaciones entre las variables recopiladas para comprender mejor los factores que influyen en la fitosanidad de los cultivos. Los resultados obtenidos en la aplicación práctica demostraron la efectividad de utilizar métodos de muestreo estadístico en redes para la optimización de muestreo e inferencia de parámetros poblacionales asociados a la Piricularia. Esta aproximación permitió obtener información representativa y confiable que, a su vez, es útil para tener una visión más completa de la situación de los cultivos de arroz, lo que facilita la toma de decisiones informadas en el ámbito agrícola incurriendo en menores costos a los incurridos actualmente mediante el uso de censos agrícolas.

**Palabras clave:** muestreo de grafos, redes, cultivos de arroz, muestreo de caminatas aleatorias, muestreo basado en nodos.

# Abstract

## Sampling of Network Structures in Unstructured Data

This paper deals with the topic of statistical sampling in networks, focusing on two fundamental aspects: in the first chapter, various methods for sampling in networks were investigated and analyzed, while in the second chapter an application focused on the optimization of phytosanitary sampling in rice cultivation was carried out. In the first chapter, the different approaches and techniques used for sample selection were discussed, highlighting traditional methods such as simple random sampling and the Horvitz Thompson-estimator. In addition, other tools used in the work such as the non-supervised classification methods, the Monte Carlo estimation method, which are of vital importance in the application performed, are discussed. Also, network sampling, where specific nodes or connections can be selected for data collection, sampling methods based on random walks and sampling for connected graphs, which use link information to select the sample in a more efficient way, were explored. In the second chapter, the concepts of statistical sampling in networks were applied in the optimization of phytosanitary sampling in crops. An application scenario was presented in which data were collected from different rice fields supplied by the Fedearroz entity, carrying out the process of creating the networks and later their analysis comparing the methods based on traditional lists with the sampling based on network type having as a basis the stratified sampling, with strata created under a method of unsupervised learning, in this method relevant variables such as yield per hectare, use of fertilizers, climatic conditions and soil characteristics were used. Based on the different samples collected in different methods, estimates were made on the intensity measures of the most relevant disease of rice cultivation in Colombia, pyricularia (*Pyricularia Oryzae*), the disease that most affects rice crops. Also, relationships between the variables collected were explored to better understand the factors that influence crop phytosanitation. The results obtained in the practical application demonstrated the effectiveness of using statistical sampling methods in networks for the optimization of sampling and inference of population parameters associated with pyricularia. This approach allowed obtaining representative and reliable information, providing a more complete vision of the situation of rice crops and facilitating informed decision making in the agricultural field at a lower cost than the traditional way, which is through censuses.

**Keywords:** graph sampling, networks, rice crops, random walk sampling, node-based sampling )

# Contenido

<b>Agradecimientos</b>	<b>IV</b>
<b>Resumen</b>	<b>VI</b>
<b>Consideraciones generales</b>	<b>VII</b>
<b>Lista de figuras</b>	<b>IX</b>
<b>Lista de tablas</b>	<b>x</b>
<b>1 Consideraciones Generales</b>	<b>1</b>
1.1 Introducción . . . . .	1
1.2 Planteamiento del problema . . . . .	3
1.3 Justificación . . . . .	4
1.4 Objetivos . . . . .	5
1.5 Hipótesis . . . . .	6
<b>2 Muestreo y Estructuras de Redes</b>	<b>7</b>
2.1 Muestreo probabilístico . . . . .	7
2.1.1 Diseño muestral . . . . .	8
2.1.2 Estimación de totales . . . . .	11
2.2 Teoría de redes . . . . .	13
2.3 Muestreo en redes . . . . .	15
2.3.1 El estimador HT en muestreos de redes basados en caminatas aleatorias	18
2.3.2 Métodos de muestreo en redes . . . . .	20
2.3.3 Métodos de muestreo en redes conectadas . . . . .	27
2.4 Aprendizaje no supervisado . . . . .	31
2.4.1 Análisis de métodos de clusterización . . . . .	32
2.4.2 $k$ -Means, y algoritmos relacionados . . . . .	36
2.4.3 Índice de silueta . . . . .	39
2.5 Método de estimación Monte Carlo . . . . .	40

---

<b>3</b>	<b>Aplicaciones de herramientas de muestreo al monitoreo de problemas fitosanitarios del cultivo de arroz en Colombia</b>	<b>42</b>
3.1	Introducción . . . . .	42
3.1.1	El cultivo de arroz mundialmente . . . . .	42
3.1.2	El cultivo de arroz en Colombia . . . . .	43
3.1.3	El papel de las enfermedades en los cultivos de arroz . . . . .	44
3.1.4	Piricularia en los cultivos de arroz . . . . .	45
3.1.5	El muestreo en las poblaciones de patógenos en plantas . . . . .	46
3.1.6	El muestreo de redes aplicado al monitoreo fitosanitario en cultivos de arroz . . . . .	47
3.2	Materiales y métodos . . . . .	48
3.2.1	Origen y procesamiento de los datos . . . . .	48
3.2.2	Metodología de muestreo . . . . .	50
3.3	Resultados . . . . .	54
3.3.1	Creación de los estratos . . . . .	54
3.3.2	Creación de la red . . . . .	56
3.3.3	Estimaciones . . . . .	59
3.4	Discusión . . . . .	61
<b>4</b>	<b>Conclusión y Recomendaciones</b>	<b>63</b>
	<b>Bibliografía</b>	<b>65</b>

# Lista de Figuras

<b>2-1</b>	Tipos de enlaces . . . . .	15
<b>2-2</b>	Muestreo de rama aleatoria - tomado de Trujillo et al. (2016) . . . . .	29
<b>3-1</b>	Distribución de clústeres - lotes por departamento - Autoria Propia . . . . .	55
<b>3-2</b>	Representación gráfica Cluster0 - Autoría propia . . . . .	56
<b>3-3</b>	Representación gráfica Cluster1 - Autoría propia . . . . .	57
<b>3-4</b>	Representación gráfica Cluster2 - Autoría propia . . . . .	57
<b>3-5</b>	Representación gráfica Cluster4 - Autoría propia . . . . .	58
<b>3-6</b>	Representación gráfica red completa - Autoría propia . . . . .	58
<b>3-7</b>	Representación gráfica CV por tamaño de muestra y método- Autoría propia	60

# Lista de Tablas

<b>3-1</b>	Distribución de clústeres - lotes por clúster . . . . .	54
<b>3-2</b>	Estimaciones asociadas a cada método . . . . .	59
<b>3-3</b>	Sesgo asociado a cada método . . . . .	60
<b>3-4</b>	Coefficiente de variación asociado a cada método . . . . .	60



# 1 Consideraciones Generales

## 1.1. Introducción

La investigación en redes representa un reto en estadística, puesto que son estructuras complejas de analizar dado el gran volumen de información que pueden llegar a contener; en ocasiones, entre miles y millones de nodos y enlaces. Esto da lugar a conjuntos de datos masivos que requieren técnicas de procesamiento eficientes y escalables (Leskovec et al. 2007). Sin embargo, con el transcurso del tiempo y la madurez en el uso de estas metodologías, ha sido posible corroborar su gran utilidad, ya que, como lo mencionaba (Qi 2022) facilitan el estudio de una amplia variedad de campos tales como caminos de redes (Xie & Levinson 2007), comunicación mediante redes (Shimbel 1953; Gupta, Jain & Vaszkun 2016), redes sociales y profesionales (Ahn et al. 2007); redes de citación (Portenoy, Hullman & West, 2017; McLaren & Bruner 2022), redes colaborativas (Newman 2001) y redes biológicas (Charitou, (Bryan & Lynn, 2016; Zhang & Itan 2019).

Históricamente, uno de los mayores retos en el análisis del muestreo por redes consistía en la dificultad para la recolección y el procesamiento de altos volúmenes de información, razón por la cual, este tipo de metodologías llegó a caer en desuso por parte de algunos investigadores (Zhang & Patone 2017). Adicionalmente, esta metodología dificulta la construcción de estimadores tanto de totales como de la varianza estimada del estimador impidiendo así la creación de expresiones comunes para las probabilidades de segundo orden.

No obstante, gracias a los últimos avances tecnológicos y la mayor capacidad computacional con la que contamos hoy en día, el interés investigativo que usa el modelamiento por redes ha resurgido, ya que se pudo subsanar el mayor inconveniente que existía en el uso de este tipo de metodología. Además con la implementación de nuevas metodologías provenientes de *Machine Learning*, los análisis de redes han asumido un papel protagónico en la construcción de nuevos métodos de estimación y en el desarrollo de herramientas computacionales que cada vez cobran mayor importancia como los modelos de interacción y recomendación, detección de comunidades y desarrollos en procesamiento de lenguaje natural.

A pesar de los grandes adelantos en el procesamiento de datos de la actualidad, el muestreo de redes continúa representando gran complejidad teórica, ya que, como se mencionaba anteriormente, dificulta la construcción de expresiones teóricas para la estimación de totales. Por

tal razón, se han adelantado diferentes desarrollos para suministrar algoritmos que permitan la recolección de datos de manera más sencilla sin dejar de lado el muestreo tradicional basado en listas. Un claro ejemplo de esto se puede observar en las metodologías de muestreo adaptativo.

Dentro del contexto estadístico anteriormente citado, los escenarios de aplicación estadística en el campo de la agricultura cobran especial relevancia en términos de realización de muestreos, ya que, en general existe la posibilidad de conocer toda la población a muestrear. En Colombia, por ejemplo, Fedearroz censa los lotes cultivados y posteriormente muestrea las plantas de cada uno. Estos muestreos son esenciales a lo largo del ciclo del cultivo, tomando en consideración que la productividad de los cultivos de arroz puede variar por factores tales como enfermedades y cambios climáticos, entre otros. De ahí su importancia para la formulación de medidas preventivas y de manejo del cultivo.

Sin embargo realizar muestreos en todos los lotes de arroz a nivel nacional puede ser una tarea compleja y costosa, especialmente en países con desafíos logísticos y de infraestructura de transporte como lo es Colombia. Por lo tanto, es crucial proponer métodos que permitan inferir conclusiones precisas a un costo más bajo, sin sacrificar la representatividad de la muestra.

En este sentido, para dar respuesta a las necesidades del sector agrícola, también se ha trabajado con muestreos basados en listas; sin embargo, esta metodología también presenta desafíos importantes, ya que la recolección de variables auxiliares de cada lote puede ser incluso más compleja que la realización de un censo. En tales muestreos por listas con variables auxiliares, el costo de los implementos para medir variables atmosféricas o de suelo es muy alto. Adicionalmente, los muestreos que no recogen ninguna variable auxiliar, como el muestreo aleatorio simple y sin reemplazo, pueden ocasionar una pérdida de información relevante.

En consecuencia, el presente trabajo propone un enfoque innovador mediante la implementación de un muestreo basado en redes. Para su elaboración es necesario contar con el trabajo realizado previamente en cuanto a la estructuración de la red de lotes de arroz suministrado por Fedearroz, donde la correcta definición de los nodos y enlaces se convierte en un aspecto clave para superar las limitaciones de los muestreos tradicionales. Así, las diferentes técnicas de muestreo funcionarán como base para obtener una muestra que refleje las conexiones y relaciones entre los lotes de arroz, garantizando una inferencia eficiente y precisa.

Por otro lado, para abordar el inconveniente teórico de los muestreos en redes, se realizarán estimaciones de varianza del estimador de manera numérica teniendo en cuenta el escenario particular en el que se conoce toda la población a muestrear.

Así, teniendo en cuenta lo anterior, este trabajo espera contribuir al avance en la aplicación de técnicas estadísticas de muestreo en un contexto específico como el de la agricultura, a la vez que espera brindar a entidades de prevención y control, como Fedearroz, una herramienta pertinente y eficaz para la toma de decisiones informadas que faciliten la creación de estrategias de manejo y prevención de enfermedades. Asimismo, este trabajo pretende profundizar el conocimiento existente en el campo de la investigación estadística de redes y su aplicabilidad a diversos sectores controlados donde se tengan poblaciones completas y se busque estimar una mejor relación costo-beneficio de modo tal que haya una mejor comprensión de las estructuras complejas presentes en un campo de aplicación real y la obtención de estimadores que permitan llegar a conclusiones más fácilmente.

Este documento está estructurado de la siguiente manera, en el primer capítulo se trataron las consideraciones generales del trabajo además de esta introducción se encontrará el planteamiento del problema, justificación, objetivos e hipótesis. En el segundo capítulo se abordará todo el marco teórico de la investigación desde la parte de muestreo en listas, muestreo en redes, métodos de muestreo en redes y elementos estadísticos importantes para la investigación como el aprendizaje no supervisado y el método Monte Carlo. En el tercer capítulo se abordará una aplicación de estos métodos al monitoreo de problemas fitosanitarios del cultivo de arroz en Colombia. Por último en el cuarto capítulo se encuentran las conclusiones y recomendaciones obtenidas tanto por la construcción del marco teórico del segundo capítulo, como de la aplicación del tercer capítulo.

## 1.2. Planteamiento del problema

El presente trabajo busca utilizar el muestreo estadístico en un contexto de redes en el que se pueda obtener estimaciones precisas y representativas de las variables de interés presentes en una población, mediante la selección adecuada de muestras. El muestreo en redes representa gran complejidad en su aplicación dada la naturaleza de las relaciones entre los diferentes elementos de la red. Por ende, el presente trabajo busca responder a los siguientes aspectos:

- Optimización del muestreo en redes: la selección de muestras en redes puede ser más compleja que en diseños de muestreo tradicionales, ya que se debe considerar la interconexión entre nodos y el efecto que estos tienen en la distribución de las variables de interés. El desafío consiste en desarrollar métodos de muestreo estadístico que permitan maximizar la precisión y representatividad de las estimaciones obtenidas, considerando la estructura de la red y evitar sesgos en la selección de muestras.
- Diseño muestral eficiente y representativo: el diseño muestral juega un papel fundamental en la calidad de los resultados obtenidos. Se debe determinar la mejor estrategia de muestreo para garantizar que todas las partes de la red tengan la oportunidad de

ser seleccionadas en la muestra y de que la información recopilada sea representativa de toda la población de interés. Esto implica tener un equilibrio entre el tamaño de la muestra, la precisión de las estimaciones y la creación de la red.

- Generalización de resultados: al obtener estimaciones en redes, se busca generalizar los resultados a toda la población objetivo. Sin embargo, debido a la complejidad de las relaciones en la red, es importante evaluar la validez y la precisión de las estimaciones para asegurar que sean aplicables a la población en su conjunto.

En resumen, el problema de esta investigación desde un punto de vista estadístico radica en cómo realizar un muestreo óptimo en redes para obtener estimaciones precisas, y que representen a toda la población. Esto implica superar los desafíos asociados a la complejidad de las redes y desarrollar estrategias de muestreo adecuadas que aseguren la validez y eficiencia del estudio estadístico en un contexto específico. Al abordar estas problemáticas, se podrá mejorar la calidad de la investigación y garantizar que los resultados obtenidos sean confiables y aplicables en diversas áreas de estudio y para la toma de decisiones.

### **1.3. Justificación**

El presente trabajo tiene una relevancia significativa en el estado del arte de los muestreos modernos. A continuación, se presentan las principales justificaciones que respaldan la importancia y el valor de este estudio:

- En la actualidad, el análisis de redes se ha vuelto fundamental en diversos campos como redes sociales, epidemiología, agricultura y ciencias ambientales. Al explicar adecuadamente los métodos de muestreo estadístico en redes, se generan procesos de selección de muestras con el fin de obtener datos de manera distinta, estos métodos tienen en cuenta la interconexión entre los individuos y que dependiendo el escenario aplicado pueden tener mejores resultados que los muestreos basados en listas, además de darle otra perspectiva al problema para los investigadores, puesto que en el proceso de creación de la red se pueden identificar características importantes de los individuos.
- Los recursos en la investigación agrícola son valiosos y limitados por la complejidad en su recolección. El uso de muestreo en redes permite recolectar datos de manera eficiente, evitando gastos excesivos y minimizando el tiempo necesario para recopilar información en comparación a los censos. Esto facilita la realización de estudios a gran escala y mejora la capacidad de realizar seguimientos a lo largo del tiempo a un costo menor.

- La inferencia en sistemas agrícolas desempeña un papel fundamental en el monitoreo y manejo de problemas fitosanitarios, es decir, enfermedades, plagas y otros agentes que afectan la salud y productividad de los cultivos, puesto que permite la detección temprana de los mismos, una estimación de la magnitud y gravedad del problema fitosanitario, la identificación de distintos factores de riesgo que contribuyen a la propagación de la enfermedad, y en relación al punto anterior la optimización de recursos para combatirla.
- Por otra parte este trabajo busca no solamente explicar de manera teórica los conceptos de redes y muestreos en redes, si no que busca llevar esta teoría a un escenario práctico en este caso de cultivos de arroz, pero se busca que la teoría descrita en este trabajo pueda ser aplicada a otros escenarios y realizar una comparación de resultados con muestreos usuales.

Este trabajo tiene un enfoque en optimizar el proceso de selección de muestras para proporcionar análisis precisos y eficientes, tiene un impacto significativo en la calidad de la investigación y en el desarrollo de estrategias para mejorar la productividad agrícola y la toma de decisiones informadas. Además, su aplicabilidad en diversas áreas de estudio amplía su relevancia y potencial impacto en la comunidad científica estadística y en la sociedad en general.

## 1.4. Objetivos

### Objetivo General

Estudiar y analizar por medio de un diseño muestral para redes los distintos tipos de muestreos para redes para obtener formas particulares de estimadores, varianzas y poder hacer un cálculo de los mismo a través de herramientas computacionales.

### Objetivos Específicos

- Observar los distintos tipos de redes y observar las diferentes características para los cuales se puede aplicar determinado diseño muestral.
- Demostrar la eficiencia que representan los diseños muestrales basados en redes en comparación de los diseños muestrales tradicionales que no usan las conexiones de enlaces de manera que sirvan para obtener información auxiliar.

- Evaluar y comparar diferentes métodos de muestreo en redes y uno de los métodos tradicionales más usados, tal como el diseño estratificado en donde a cada estrato se aplica un muestreo aleatorio simple y sin reemplazo, para determinar cual proporciona estimaciones más precisas y representativas de variables agrícolas en los cultivos de arroz.

## 1.5. Hipótesis

Tomando como punto de partida la aplicación del muestreo estadístico en redes agrícolas para el estudio de aspectos fitosanitarios en los cultivos de arroz colombianos, se plantea la siguiente hipótesis:

- El muestreo en redes agrícolas proporciona estimaciones más precisas y representativas de variables agrícolas en cultivos de arroz en comparación con los métodos de muestreo tradicionales.

En otras palabras, este trabajo sugiere que el muestreo en redes para escenarios agrícolas tiene una ventaja estadísticamente significativa sobre los métodos tradicionales, ya que proporciona estimaciones más precisas y representativas a menor costo del actualmente incurrido en la industria fitosanitaria de los cultivos de arroz en Colombia.

Para probar esta hipótesis, se realizará un análisis comparativo de los resultados obtenidos mediante ambas metodologías de muestreo para así evaluar la validez y aplicabilidad de dichos resultados en el contexto particular ya mencionado.

## 2 Muestreo y Estructuras de Redes

### 2.1. Muestreo probabilístico

De acuerdo a Särndal, Swensson y Wretman (1992), las condiciones para que una muestra sea considerada probabilística son las siguientes:

- Se puede definir el conjunto de muestras  $Q = \{s_1, s_2, s_3, \dots, s_q\}$  que es posible obtener con el procedimiento de muestreo.
- Se asocia una probabilidad de selección  $p(s)$  con cada posible muestra  $s$ .
- El procedimiento otorga a todos los elementos  $i = 1, \dots, N$  de la población una probabilidad  $\pi_i$  distinta a cero de ser incluidos en la muestra
- La selección de una muestra  $s$  mediante un mecanismo aleatorio que reproduce exactamente la probabilidad  $p(s)$  y los individuos  $i$  en la población que reciben aproximadamente la probabilidad  $\pi_i$  de ser incluidos en la muestra.

Sea  $Y$  una variable de tal forma que  $y_i$  represente el valor de la variable  $Y$  para el  $i$ -ésimo elemento poblacional. Por defecto asumimos que los elementos  $y_i$ , para todo  $i \in U$  son desconocidos.

Para estimar el total poblacional de  $Y$ , dado  $t = \sum_{i=1}^N y_i$  observamos  $y$  un subconjunto de  $U$ , en vez de la población que generalmente resultaría demasiado costoso o poco práctico (Särndal et al. 1992). Se denomina una muestra al subconjunto de la población que se selecciona donde el valor  $y_i$  es observado para cada elemento  $i$  que son usados para la estimación de  $t$ .

Así, por ejemplo, consideremos  $U$  como una población de lotes agrícolas donde  $Y$  es la variable "medida de la incidencia de la enfermedad". En este contexto,  $y_i$  cuantifica la incidencia de la enfermedad presente en el lote  $i$ -ésimo.

La incidencia de una enfermedad, se refiere a la cantidad de casos nuevos de una enfermedad en una población definida durante un período de tiempo específico (Porta 2014). Es una medida importante para comprender cómo cambia la dinámica de una enfermedad en una población y como medida de intensidad de la enfermedad en un período determinado.

Ahora, si asumimos que el interés está en estimar el total poblacional de  $Y$  que en este caso representa la incidencia de producción de arroz de la población  $U$ , la cantidad llamada  $t$  es desconocida y sí misma constituirá un parámetro de una población finita.

### 2.1.1. Diseño muestral

La selección de la muestra se realiza mediante un diseño muestral que se define como una distribución de probabilidad definida sobre un soporte  $Q$ , tal que  $p(s) > 0$  para todo  $s \in Q$  y

$$\sum_{s \in Q} p(s) = 1$$

Cassel, Särndal y Wretman (1976) afirman que dado un soporte  $Q$ , un diseño de muestreo puede ser sin reemplazo o con reemplazo y de tamaño de muestra fijo o aleatorio. Del mismo modo, explican que la forma de identificar cada una de todas las posibles muestras que pertenecen al soporte  $Q$  es un factor crucial que permite designar un conjunto de muestras a las cuales se les asigna una probabilidad positiva de selección para así distribuir la totalidad de la masa de probabilidad entre los miembros de  $Q$ .

La inclusión del elemento  $i$ -ésimo en una muestra  $s$  particular es un evento aleatorio definido por la función indicadora  $I_i$ , que está dada por Särndal (1992):

$$I_i = \begin{cases} 1 & \text{si } i \in s \\ 0 & \text{si } i \notin s \end{cases}$$

Bajo un diseño de muestreo  $p(\cdot)$ , se asigna una probabilidad de inclusión a cada elemento de la población para indicar la probabilidad de que el elemento pertenezca a la muestra. Para el elemento  $i$ -ésimo de la población, dicha probabilidad está dada por Särndal et al. (1992):

$$\pi_i = P(i \in s) = P(I_i = 1) = \sum_{s \ni i} p(s)$$

En donde el subíndice  $s \ni i$  se refiere a la suma sobre todas las muestras que contienen al elemento  $i$ -ésimo. Análogamente,  $\pi_{il}$  se conoce como la probabilidad de inclusión de segundo orden y denota la probabilidad de que los elementos  $i$  y  $l$  pertenezcan a la muestra, y esta probabilidad está dada por Särndal et al. (1992):

$$\pi_{il} = P(i \in s \text{ y } l \in s) = P(I_i I_l = 1) = \sum_{s \ni iyl} p(s)$$

El objetivo de la investigación por muestreo es estudiar una característica de interés  $Y$  que se encuentre asociada a cada unidad de la población (Cassel et al. 1976). Es importante notar que los  $y_i$  no se consideran variables aleatorias, sino cantidades fijas. Por tanto, la notación de éstas se hace con una letra minúscula  $y$ . El objetivo del muestreo es estimar una función de interés  $T$ , denominada parámetro, de la característica de interés dentro de la población Särndal et al. (1992).

$$T = f(y_1, \dots, y_k, \dots, y_N).$$

Algunos de los parámetros de interés más comunes son:

- El total poblacional

$$t_y = \sum_{i \in U} y_i$$

- La media poblacional

$$\bar{y}_U = \frac{\sum_{i \in U} y_i}{N} = \frac{t_y}{N}$$

Donde los  $y_i$  son los valores observados de la variable  $y$  para todos los elementos de la población  $U$  y  $N$  corresponde al tamaño de la población.

Así mismo, una estadística  $\hat{T}(S)$  se define como una función (que toma valores reales) de la muestra aleatoria  $S$  y una realización de tal estadística sólo puede ser calculada una vez que  $s$  sea seleccionada  $S$ . Siendo  $\hat{T}$  una estadística, sus propiedades están determinadas por el diseño de muestreo. Es decir, dada la probabilidad de selección de cada muestra  $s \in Q$ , la esperanza, la varianza y otras propiedades de interés están definidas a partir de  $p(s)$  (Särndal et al. 1992). Así, la esperanza de una estadística se define como:

$$E(\hat{T}) = \sum_{s \in Q} p(s) \hat{T}(s)$$

Por otro lado, la varianza de la estadística  $\hat{T}$  se define como:

$$\begin{aligned} \text{Var}(\hat{T}) &= E[\hat{T} - E(\hat{T})]^2 \\ &= \sum_{s \in Q} p(s) [\hat{T}(s) - E(\hat{T})]^2. \end{aligned}$$

Donde  $\hat{T}(s)$  es el valor real que toma la estadística  $\hat{T}$  en la muestra seleccionada (realizada)  $s$ . Cuando una estadística se construye con la intención de estimar un parámetro, recibe el nombre de estimador. Así, las propiedades más comúnmente utilizadas de un estimador  $\hat{T}$  de un parámetro de interés  $T$  son el sesgo, definido por Särndal et al. (1992):

$$B(\hat{T}) = E(\hat{T}) - T$$

Mientras que el error cuadrático medio está dado por:

$$\begin{aligned} \text{ECM}(\hat{T}) &= E \left[ (\hat{T} - T)^2 \right] \\ &= \text{Var}(\hat{T}) + B^2(\hat{T}). \end{aligned}$$

Si el sesgo de un estimador es cero, se dice que el estimador es insesgado y cuando esto ocurre, el error cuadrático medio se convierte en la varianza del estimador. Este enfoque inferencial se conoce como inferencia basada en el diseño de muestreo, ya que el parámetro es una constante desconocida. Bajo esta inferencia, las estimaciones de los parámetros de interés y sus propiedades dependen directamente de la medida de probabilidad discreta inducida por el diseño de muestreo escogido para seleccionar la muestra y no considera las propiedades de la población finita (Cassel et al. 1976).

Cassel, Särndal y Wretman (1976) afirman que el objetivo en un estudio de muestreo es estimar uno o más parámetros poblacionales. En este sentido, las decisiones más importantes a la hora de abordar un problema de estimación por muestreo son:

- La elección de un diseño de muestreo y un algoritmo de selección que permita implementar el diseño
- La elección de una fórmula estadística o estimador que calcule una estimación del parámetro de interés en la muestra seleccionada.

Cabe mencionar que las decisiones anteriores no son de carácter independiente. Esto significa que la selección de un estimador dependerá, usualmente, del diseño de muestreo utilizado.

### Definición

Siendo  $\hat{T}$  un estimador de un parámetro  $T$  y  $p(\cdot)$  un diseño de muestreo definido sobre un soporte  $Q$ , se define una estrategia de muestreo como la dupla  $(p(\cdot), \hat{T})$  (Särndal et al. 1992).

La anterior definición parece ser estándar en la literatura. Sin embargo, en algunos textos clásicos como (Cochran 1954; Wiegand & Kish 1965), el término diseño muestral incluye

tanto la forma de muestreo como el método de estimación. En el caso particular de este trabajo, se prefiere usar el término estrategia de muestreo como la combinación de los dos elementos, diseño muestral y estimador .

### 2.1.2. Estimación de totales

Para un universo  $U$ , se quiere estimar el total poblacional  $t_y$  de la característica de interés. Luego, se define el estimador Horvitz-Thompson (HT, Horvitz & Thompson 1952) como:

$$\hat{t}_{y,\pi} = \sum_s \frac{y_i}{\pi_i} = \sum_s d_i y_i$$

Donde,  $d_i = 1/\pi_i$  es el inverso de la probabilidad de inclusión y es conocido como el factor de expansión.

Este estimador también se conoce como el  $\pi$ -estimador y está motivado, como Brewer (2002) lo indica, en el principio de representatividad, que afirma que cada elemento incluido en una muestra se representa a sí mismo y a un grupo de unidades que no pertenecen a la muestra seleccionada cuyas características son cercanas a las del elemento incluido en la muestra.

El factor de expansión indica cuántas veces representa un elemento de la muestra a un elemento de la población. Así, para una población de tamaño  $N = 10$ , al utilizar un diseño de muestreo aleatorio simple sin reemplazo de tamaño  $n = 2$ , el factor de expansión es  $d_k = 10/2 = 5$ . En otras palabras, el elemento incluido se representa a sí mismo y a cuatro elementos más. Los siguientes resultados dan cuenta de las propiedades de este estimador bajo un muestreo probabilístico.

#### Propiedad 1

Si todas las probabilidades de inclusión de primer orden son mayores a cero ( $\pi_i > 0$  para todo  $i$ ), el estimador de Horvitz-Thompson es insesgado para el total poblacional. Por tanto, se tiene que

$$E(\hat{t}_{y,\pi}) = t_y$$

#### Propiedad 2

La varianza del estimador de Horvitz-Thompson está dada por la siguiente expresión:

$$\text{Var}_1(\hat{t}_{y,\pi}) = \sum \sum_U \Delta_{il} \frac{y_i}{\pi_i} \frac{y_l}{\pi_l}$$

En donde  $\Delta_{il} = \text{Cov}(I_i, I_l) = \pi_{il} - \pi_i \pi_l$ . Por otra parte, se tiene el siguiente resultado cuando el diseño de muestreo es de tamaño fijo.

### Propiedad 3

Si el diseño  $p(\cdot)$  es de tamaño de muestra fijo, entonces la varianza del estimador de Horvitz-Thompson se escribe como

$$\text{Var}_2(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum \sum_U \Delta_{il} \left( \frac{y_i}{\pi_i} - \frac{y_l}{\pi_l} \right)^2$$

Es posible construir dos estimadores insesgados para las expresiones dadas anteriormente. Para esto, se requiere que todas las probabilidades de inclusión de segundo orden sean estrictamente positivas ( $\pi_{il} > 0$  para todo  $i$  y  $l$ ). Con el anterior supuesto, se tienen los siguientes resultados.

### Propiedad 4

Si todas las probabilidades de inclusión de segundo orden son mayores que cero estrictamente, un estimador insesgado para la varianza está dado por

$$\widehat{\text{Var}}_1(\hat{t}_{y,\pi}) = \sum \sum_s \frac{\Delta_{il}}{\pi_{il}} \frac{y_i}{\pi_i} \frac{y_l}{\pi_l}$$

### Propiedad 5

Si el diseño de muestreo es de tamaño de muestra fijo y si todas las probabilidades de inclusión de segundo orden son mayores que cero estrictamente, un estimador insesgado para la varianza está dado por

$$\widehat{\text{Var}}_2(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum \sum_s \frac{\Delta_{il}}{\pi_{il}} \left( \frac{y_i}{\pi_i} - \frac{y_l}{\pi_l} \right)^2$$

## 2.2. Teoría de redes

Las redes han sido usadas para modelar varios de los problemas del mundo real, tales como aplicaciones industriales, modelos químicos, redes sociales, sensores remotos, entre otros problemas, dado que para obtener su solución, el problema puede ser transformado en términos de nodos y enlaces (Leskovec et al. 2007). En consecuencia, la teoría de redes comienza a ser usada en las áreas de aprendizaje automático y de reconocimiento de patrones para extraer conocimiento de representaciones dadas por una red (Zhang & Patone 2017). Formalmente, una red simple no dirigida se define como un par ordenado  $G = (V, E)$ , donde  $V$  es un conjunto de elementos denominados nodos, y  $E$  es un conjunto de pares no ordenados  $\{v, u\}$  denominados enlaces (Gilbert 1959).

A partir de ahora usaremos redes o grafos indistintamente como el mismo concepto, así mismo para nodos o vértices y para enlaces o aristas.

Sea  $\psi_G(e)$  una función de incidencia que asocia a cada par de nodos de  $G$ , si  $e$  es un enlace, y  $u$  y  $v$  son nodos tales que la función  $\psi_G(e) = \{u, v\}$ , se dice que  $e$  es una unión de  $u$  y  $v$  o los enlaces de la red. A partir de lo comentado en Biggs, Lloyd y Wilson (1986), se expondrán las siguientes definiciones.

### Nodos adyacentes

Dos nodos  $v$  y  $u$  se denominan adyacentes si hay un enlace  $\{v, u\} \in E$ , que corresponde a todo el conjunto de enlaces en el grafo  $G$ .

### Vecindad

La vecindad de un nodo  $v$  en un grafo  $G = (V, E)$  es  $N(v) = \{\forall u \in V \mid \{v, u\} \in E\}$ . Es decir,  $N(v)$  es el conjunto de todos los nodos adyacentes a  $v$  sin el mismo. Por lo tanto, los nodos no vecinos a un nodo  $v$  son todos aquellos que no comparten enlaces con  $v$ .

### Grado de un nodo

Dado un grafo  $G = (V, E)$ , el grado de un nodo  $v \in V$ , denotado como  $\delta(v)$ , es  $|N(v)|$ . Es decir, el número de enlaces en las cuales  $v$  incide.

### Subgrafo

Dado un subconjunto de nodos  $S \subseteq V$ , el subgrafo de  $G$  denotado como  $G \mid S$  tiene un conjunto de nodos  $S$  y un conjunto de enlaces tal que,  $E(G \mid S) = \{\{u, v\} \in E : u, v \in S\}$ . Por lo que,  $G \mid S$  se denomina el subgrafo de  $G$  inducido por  $S$ . Se escribe  $G - S$  para denotar el grafo  $G \mid (V - S)$ . El subgrafo inducido por  $N(v)$  es denotado como  $H(v) = G \mid N(v)$ , el cual tiene al conjunto  $N(v)$  como conjunto de nodos y todos los enlaces sobre ellos.

Dado un subgrafo  $H \subseteq G$ , para cada nodo  $u \in V(H)$ , dado  $\delta_H(u)$  el grado de  $u$  en un subgrafo inducido  $H$  de  $G$ , si  $H = G$ , entonces  $\delta_G(u) = \delta(u)$  y  $E_H(u) = \{\{u, v\} \in E(G) : v \in H\}$ . De manera similar,  $N_H(u)$  denota el conjunto de nodos de  $H$  adyacentes a  $u$ . Para cualquier subgrafo  $H \subseteq G$ ,  $\delta_G(H) = \sum_{u \in H} \delta_G(u)$ . Si  $H$  es un conjunto independiente de  $G$ , entonces  $\delta_G(H)$  es el número de enlaces de  $G$  incidentes a cualquier nodo de  $H$ .

### Camino

Un camino del nodo  $v$  a un nodo  $u$  en un grafo, es una secuencia de enlaces:

$v_0 - v_1, v_1 - v_2, \dots, v_{n-1} - v_n$ , tal que,  $v = v_0, v_n = u, v_k$  es adyacente a  $v_{k+1}$  y la longitud del camino es  $n$ . Un camino simple, es un camino tal que  $v_0, v_1, \dots, v_{n-1}, v_n$  donde todos son distintos.

### Ciclo

Un ciclo, con al menos tres nodos, es un camino no vacío cuyos nodos pueden ser organizados en una secuencia cíclica, es decir, un nodo inicial y final se unen por un enlace.

### Árbol

Es un grafo sin ciclos, es decir un grafo  $G$  tal que, para cualquier par de nodos en  $G$  hay un solo camino que los une.

### Árbol de expansión

Un árbol de expansión  $T$  contiene todos los nodos del grafo original sin enlaces que formen ciclos.

### Grafo conectado

Es un grafo  $G = (V, E)$  si cada par de nodos en  $G$  tienen un camino entre ellos. Si el grafo es no conectado, cada pieza conectada máxima se denomina componente.

### Grafo completo

Es un grafo no dirigido en el cual cualquier par de nodos está conectado por un único enlace.

### Grafo ponderado

Un grafo ponderado denotado como  $G_w = (V, E)$ , es un grafo donde cada enlace  $e \in E$  tiene asociado un número real  $w(e)$  denominado peso. La matriz de adyacencia de un grafo ponderado  $G_w$  es una  $V \times V$  matriz, tal que  $M_G = (w_{vu})$ , donde cada elemento  $(v_i, v_j)$  contiene un peso  $w(e)$  asignado al enlace  $e = v_i, v_j$  considerando si los nodos  $v_i$  y  $v_j$  son adyacentes o no en el grafo.



Figura 2-1: Tipos de enlaces

## 2.3. Muestreo en redes

Ove Frank es reconocido como la figura más destacada en la contribución a la teoría del muestreo de grafos existente. Se pueden encontrar ejemplos de sus trabajos en el resumen proporcionado por Frank (Frank 1977c, 1979, 1980b, 1981, 2011). Sin embargo, sus numerosos trabajos están dispersos a lo largo de varias décadas y no son fácilmente comprensibles en su totalidad. Por ejemplo, Frank ha presentado resultados para diversas muestras de nodos (Frank 1971, 1977c, 1994), díadas (Frank 1971, 1977a, 1977b, 1979) o tríadas (Frank, 1971, 1979), pero nunca ha propuesto una definición formal del muestreo en grafos que unifique todas estas muestras.

Otra perspectiva de los estudios de Frank se centra en varias características de un grafo, como su orden (Frank 1971, 1977c, 1994), tamaño (Frank 1971, 1977a, 1977b, 1979), distribución de grados (Frank 1971, 1980), conectividad (Frank 1971, 1978), entre otros. Sin embargo, no ha proporcionado una estructura de parámetros de grafos que permita clasificar y comparar los diferentes aspectos de estudio. Además, Frank no ha explorado explícitamente la relación entre la teoría del muestreo de grafos y ciertos métodos de muestreo de grafos comunes por ejemplo, (Birnbbaum & Sirken 1965; Thompson 2006; Lavallée 2007), que aunque no se

plantean como problemas de grafos, pueden abordarse de esta manera (Frank,1977c).

Por otra parte tal como comenta Frank (1978), los conceptos de muestreo básico son de gran importancia para la construcción de la teoría necesaria para el muestreo en redes. dentro de estos conceptos se incluye la selección de la muestra, el diseño muestral, los esquemas de observación y variables de interés o auxiliares definidas para los individuos.

Sin embargo, para realizar un muestreo en redes, la población a muestrear debe tener una estructura relacional entre sus unidades, usualmente ésta se dá como una variable binaria definida entre pares ordenados de unidades, y esto se puede representar mediante un grafo donde los vértices representan las unidades y los enlaces indican los pares ordenados que están relacionados Frank (1978).

Según Lavallée (2007), la población en la que se extrae la muestra puede llegar a ser distinta de la población objetivo, esto puede pasar en escenarios donde la misma es de difícil acceso pero se puede relacionar a otra población mediante un enlace, esto hace que las características de la población original sean aproximadas por la población a la que esta conectada.

Así mismo, otra consideración importante es que las redes pueden cambiar con el tiempo o espacio; por esta razón, es importante definir los parámetros de manera previa tal como se ha identificado la relación entre nodos.

Ahora, bajo la suposición de que existe una población de  $N$  unidades denotada por  $U = \{1, 2, \dots, N\}$ , las muestras se obtendrán de este universo  $U$  como secuencias con o sin repeticiones y formando subconjuntos de  $U$ . En este caso, si se toman  $n$  unidades a muestrear, se genera el subconjunto de  $U$  denotado por  $u = \{u_1, u_2, \dots, u_n\}$ , donde los elementos pertenecientes a esta muestra obtenida, dependen de sus probabilidades de inclusión dadas a partir del diseño muestral.

Este esquema de muestreo nos permite indicar cuales variables serán conocidas u observables y estas serán de vital ayuda, dependiendo el diseño muestral, tendrán ciertos usos, en el caso de muestreo en redes, las variables son especificadas para cada nodo y se tendrá una o más variables que servirán de enlaces entre los mismos.

La definición de una variable en un rango  $R$  viene definida por  $y = \{(i, y_i) : i \in U\}$ , el cual es un conjunto de pares  $(i, y_i)$  que asigna un valor  $y_i$  en  $R$  a cada unidad  $i$  en  $U$ . Si existe una especie de orden en las unidades muestreadas, x podría representarse como una secuencia de valores ordenados  $y_{(n)} = \{y_{(1)}, y_{(2)}, y_{(3)} \dots, y_{(n)}; n = 1, 2, \dots, N\}$  (Frank 1978).

Los diseños muestrales pueden ser probabilísticos como por ejemplo aleatorio simple, sis-

temático, estratificado, conglomerado, o no probabilístico como muestreo por conveniencia, incidental o voluntario. Teniendo en cuenta el tipo de muestreo a realizar se procede a la estimación del parámetro de interés junto a su varianza para posterior uso en inferencia con respecto al parámetro poblacional. Un enfoque similar tiene el muestreo en redes solo que se tienen algunas diferencias marcadas con el muestreo tradicional como la presencia de una variable auxiliar para todos los individuos que permite un enlace y debe estar presente en todos los individuos, además de diferencias importantes en los diseños muestrales tales como la selección de la muestra y la estimación tanto del parámetro poblacional como de la varianza del estimador (Frank 1978).

Según lo discutido en Biggs et al. (1986), se define un grafo como un par ordenado  $G(V, E)$ , en el que  $V$  representa el conjunto de vértices y  $E$  el conjunto de aristas (nodos y enlaces, respectivamente). Para denotar el número de enlaces y nodos en el grafo, se utilizan las variables  $n$  y  $m$ . Cada vértice en la red representa a un individuo  $u_i$ , donde  $i = 1, 2, \dots, N$ . Esta identificación es única para cada nodo. Además, si se considera que  $G$  es un grafo simple y no dirigido, implica que la relación entre dos individuos  $u_i$  y  $u_j$  es única y no tiene dirección. Esto significa que  $u_i$  está conectado a  $u_j$  y  $u_j$  está conectado a  $u_i$ , lo que indica una relación simétrica. Por lo tanto, se define la arista  $(i, j)$  para representar la conexión entre los individuos  $u_i$  y  $u_j$  a través de una única arista que los une.

De aquí podemos definir el primer muestreo que surge de forma natural en esta metodología el cual consiste en la obtención de un subgrafo del grafo  $G$  original denotado por  $G^*$ , con lo cual se define un subconjunto de nodos  $V^*$  y enlaces de este subconjunto de nodos  $E^*$ . Este tipo de muestreo llamado muestreo inducido de redes fue el primer propuesto para abordar este tipo de problemas, sin embargo se noto que tenia muchos problemas debido al sesgo natural que produce tomar subconjuntos de nodos de manera tradicional en muestreo, puesto que si no se tiene apoyo en los enlaces de estos nodos, se produce una gran pérdida de información en los enlaces correspondientes a los nodos que no serian parte de la muestra del subgrafo (Frank 1971).

De aquí surgen distintos métodos para solucionar este sesgo, cada uno tiene en cuenta el contexto donde esta desarrollado el grafo, la estructura de conexiones que tiene y la cantidad de información adicional que tiene cada nodo, de estos métodos surgen diversos diseños muestrales que se abordaran más adelante.

### 2.3.1. El estimador HT en muestreos de redes basados en caminatas aleatorias

El estimador HT se utiliza ampliamente en la estimación de totales para los métodos basados en caminatas aleatorias tal como se comenta en (Zhang & Patone 2017).

A continuación sera resumido de manera general sus características referentes a parámetro objetivo, muestreo y estimador.

#### Parámetro objetivo

Sea  $M_k$  un subconjunto de  $U$ , donde  $|M_k| = k$ . Sea  $\mathcal{C}_k$  el conjunto de todos los posibles  $M_k$ , donde  $|\mathcal{C}_k| = \frac{N!}{[k!(N-k)!]}$ . Sea  $G(M_k)$  el subgrafo inducido por  $M_k$ . Sea  $y(G(M_k))$ , o simplemente  $y(M_k)$ , una función de valor entero o real. El total de grafo de  $k$ -ésimo orden correspondiente se define como (Zhang & Patone 2017):

$$\theta = \sum_{M_k \in \mathcal{C}_k} y(M_k)$$

Se observa que los parámetros de orden refieren a las funciones de totales de grafo, dependiendo de los elementos involucrados en estos totales se asigna el orden de los mismos, por ejemplo, la suma de una variable auxiliar asociada a nodos muestreados corresponde a un total de orden 1, cuando el calculo del total implica la relación entre dos nodos se conoce como un total de orden 2 o una diada, y por ultimo cuando implica un camino entre 3 nodos se conoce como un total de orden 3 o triada (Zhang & Patone 2017).

En todos los métodos de muestreo de redes mencionados anteriormente, el parámetro objetivo es el total de un valor asociado a cada nodo del grafo, denotado por  $y_i$  para  $i \in U$ , que puede denominarse total del grafo de orden 1,  $\theta = \sum_{i \in U} y_i$  (Zhang & Patone 2017).

Esto no difiere de lo que ocurre cuando se aplican métodos de muestreo tradicionales con el mismo fin, tales como el muestreo aleatorio simple y sin reemplazo, estos métodos de muestreo de redes sólo se han usado hasta ahora para superar ciertas deficiencias del marco o la falta de eficacia de los métodos de muestreo tradicionales, como se expone a continuación en términos de muestreo y estimador, pero no para estudiar totales verdaderos o parámetros de redes, que son de órdenes de magnitud superiores a uno (Zhang & Patone 2017).

#### Muestreo

Denotemos por  $s_1$  una muestra inicial de nodos, donde  $s_1 \subseteq U$ . Bajo un diseño de probabilidad,  $\pi_i$  y  $\pi_{ij}$  representan las probabilidades de inclusión, de primer y segundo orden de nodos

en  $s_1$ . Una característica definitoria del muestreo en grafos es que se utilizan los enlaces para seleccionar el grafo de muestra, que denotaremos como  $G_s$ . Dado  $s_1$ , los enlaces relevantes se encuentran en  $\alpha(s_1) = \bigcup_{i \in s_1} \alpha_i$ . Se debe especificar un procedimiento de muestreo para los enlaces, y los enlaces observados pueden expresarse en términos de un conjunto de referencia de pares de nodos, denotado como  $s_2$ , donde  $s_2 \subseteq U \times U$ , bajo la convención de que el conjunto de enlaces  $A_{ij}$  se observa siempre que  $(ij) \in s_2$ . Denotemos por  $A_s = A(s_2)$  el conjunto de enlaces inherentes a  $s_2$ , y por  $U_s = s_1 \cup \text{Inc}(A_s)$  la unión de  $s_1$  y aquellos nodos que son incidentes a  $A_s$ . El grafo de muestra resultante es  $G_s = (U_s, A_s)$  (Zhang & Patone 2017).

El marco de muestreo de  $s_1$  puede ser directo o indirecto, directo cuando las unidades a muestrear en  $s_1$  son las mismas de la población de estudio, en el caso indirecto las unidades de muestra no son las mismas que la población de interés sino que refieren a unidades de otra población que esta conectada mediante enlaces a la población de interés y que generalmente es más fácil de muestrear, por ejemplo los casos de transmisión de enfermedades, un marco de muestreo directo solamente muestrearía a personas enfermas, mientras que un marco de muestreo indirecto además de muestrear personas enfermas, seleccionaría a los que estén conectados al mismo mediante un enlace pero podrían no tener la enfermedad (Zhang & Patone 2017).

Esto puede ser necesario porque un marco de la población no se pueden muestrear tal como pasa en escenarios donde se quieren identificar grupos muy pequeños, sin embargo estos elementos que no son identificables sean relacionados mediante un enlace a la muestra inicial.

### Estimadores para parámetros de primer orden del grafo

El estimador usual de primer orden para una población de elementos que son representados a través de un grafo  $G = (U, A)$ , es el Estimador HT visto anteriormente, el mismo es definido para los nodos presentes en el grafo resultante de la muestra  $G_s = (U_s, A_s)$ . Sea  $F$  la lista de nodos que se incluyeron en la muestra, donde  $l \in F$  tiene inclusión de probabilidad  $\pi_l$ . Se tiene que

$$\sum_{l \in F} z_l = \sum_{l \in F} \left( \sum_{i \in U} w_{li} y_i \right) = \sum_{i \in U} y_i \sum_{l \in F} w_{li} = \sum_{i \in U} y_i = \theta,$$

donde  $z_l = \sum_{i \in U} w_{li} y_i$  es el valor de la variable de interés en los nodos muestreados, en caso de que el grafo sea ponderado se tiene que sea  $w_{ji}$  el peso del enlace  $i, j$  se debe cumplir que  $\sum_{i \in F} w_{ji} = 1$  tal como lo comento Birnbaum y Sirken (1965). El siguiente estimador HT es insesgado para el parámetro poblacional  $\theta$  esta dado por:

$$\hat{\theta}_{HT} = \sum_{l \in s_1} z_l / \pi_l = \sum_{l \in F} z_l \delta_l / \pi_l,$$

donde  $\delta_l = 1$  si  $l \in s_1$  y 0 en otro caso. Para asegurar que  $z_l$  sea calculado sin importar la muestra inicial  $s_1$ , los pesos  $w_{li}$  no deben depender de la muestra inicial  $s_1$ . Un enfoque común y el que trataremos en la aplicación de este trabajo es el grafo donde todos los enlaces presentan la misma ponderación y es  $w_{li} = 1/m_i$ , donde  $l$  es una unidad muestreada en  $s_1$  que esta conectada con  $i$ , y  $m_i$  es el numero de todas las unidades de la muestra en  $F$  que dan lugar a la observación  $i$ , para todo  $i \in U$  (Zhang 2021).

El número  $m_i$  es conocido también como la multiplicidad del elemento (Birnbaum & Sirken 1965). La observación de  $m_i$  para cada elemento de la muestra es el equivalente a la probabilidad de inclusión que conocemos en los métodos tradicionales, por lo que se observa que los métodos basados en redes poseen una forma de estimación similar a los métodos tradicionales.

Esto haría pensar que los métodos de redes son fácilmente aplicables en cualquier campo de muestreo y que con la capacidad computacional que se tiene en estos momentos se pueden replicar en cualquier escenario, sin embargo tienen una gran deficiencia en realizar estimaciones en ordenes superiores a 1 por la imposibilidad de construir estimadores para la varianza, por lo que el campo de aplicación de estos métodos requiere características muy específicas.

### 2.3.2. Métodos de muestreo en redes

#### Muestreo caminata aleatoria simple

El término **Caminata aleatoria** fue acuñado por primera vez por Pearson (1905) y ha sido objeto de estudio durante varias décadas debido a su diversidad de significados en diferentes contextos. Por lo tanto, la caminata aleatoria continúa siendo un tema de investigación relevante en los últimos años. El modelo más estudiado es la caminata aleatoria en retículas, que incluye tanto la caminata aleatoria unidimensional como el de dimensiones superiores. Además, la caminata aleatoria en redes es una variante de este modelo.

En un grafo  $G = (V, E)$ , una caminata aleatoria simple es un proceso en el que se elige de forma uniforme uno de los vecinos del nodo actual como el siguiente nodo. Si especificamos el número de pasos  $t$ , la caminata aleatoria de  $t$  pasos es un proceso que involucra variables aleatorias  $Y_1, Y_2, \dots, Y_t$ , donde  $Y_i \in V$  y  $Y_{i+1}$  es un nodo elegido al azar de los vecinos de  $Y_i$ . En el muestreo de redes, la caminata aleatoria simple recopila todos los nodos y enlaces alcanzados durante todo el proceso, creando así un grafo de muestra que es un subconjunto definido del grafo original  $G$  en términos de conjuntos de nodos y enlaces (Qi 2022). A continuación se presenta un pseudocódigo para la caminata aleatoria simple (Qi 2022).

**Algoritmo**

- 1 Selecciona un nodo  $v_0$  del conjunto de nodos  $V$  como nodo inicial
- 2 Mientras  $i < \text{total de la muestra}$ 
  - 2-1 Encontrar los vecinos del nodo actual  $v_i$
  - 2-2 Elegir un nodo  $u$  de forma uniforme entre los vecinos de  $v_i$ .
  - 2-3  $v_{i+1} \leftarrow u$
  - 2-4  $i = i + 1$
- 3 **terminar**

La caminata aleatoria posee una teoría completa como proceso de Markov. Un proceso de Markov es una clase de proceso estocástico en el cual el estado siguiente depende únicamente del estado actual y es independiente de la historia pasada (es decir, los estados anteriores desde 0 hasta  $t - 1$ ). Por lo tanto, esto puede expresarse en la teoría de las cadenas de Markov de la siguiente manera (Biggs et al. 1986):

$$P(X_{t+1}|X_t) = P(X_{t+1}|X_t, X_{t-1}, \dots, X_0)$$

Consideramos todos los nodos del conjunto  $V$  como  $N$  estados diferentes en el espacio muestral  $\Omega$ , y cada enlace comparte la misma probabilidad en el tiempo  $t$ , en este escenario  $t$  corresponderá al orden del  $i$ -ésimo elemento de la muestra, ya que los estados siguientes se eligen de forma uniforme entre los vecinos del estado actual (nodo). Este mecanismo de muestreo hace que el caminata aleatoria simple sea claramente Markoviano, ya que la probabilidad de seleccionar el siguiente nodo solo depende del grado de salida del nodo actual. Específicamente, en redes no dirigidos, esto solo depende del grado del nodo actual (Frank 1978).

Como cadena de Markov, las propiedades de Markov de la caminata aleatoria simple en redes también han sido ampliamente estudiadas en las últimas décadas. Según la probabilidad de transición de la caminata aleatoria simple, la distribución estacionaria es proporcional a los grados de los nodos, mientras que para los enlaces, la distribución estacionaria es uniforme (Zhang 2021). Por lo tanto, la caminata aleatoria simple puede utilizarse para simular el muestreo uniforme de nodos cuando no se dispone de un muestreo no sesgado o este es demasiado costoso de realizar en algunas situaciones.

La caminata aleatoria simple es el algoritmo de caminata aleatoria más básico en esta revisión y es la base de todos los algoritmos basados en caminata aleatoria. Podemos observar que todas las variantes consisten en modificar la selección del mecanismo de vecinos o filtrar la muestra recopilada por la caminata aleatoria simple (Qi 2022).

Para asegurarse de que los muestreos basados en caminatas aleatorias son aleatorios, es necesario que los pasos sean tomados al azar y que no haya patrones o sesgos en la selección de los pasos. Según Zhang & Patone (2017) los muestreos basados en caminatas aleatorias son aleatorios si se cumplen las siguientes condiciones:

- El punto de partida es aleatorio.
- La dirección de cada paso es aleatoria.
- La longitud de cada paso es aleatoria.

En otras palabras, el proceso de selección de los pasos debe ser completamente aleatorio para que los muestreos basados en caminatas aleatorias sean aleatorios. Además, es importante que los pasos sean independientes entre sí, es decir, que la dirección y la longitud de un paso no dependan de los pasos anteriores.

Los procedimientos de muestreo nos proporcionan muestras, pero también es necesario un marco de estimación si deseamos estimar propiedades específicas del grafo original. En cuanto al marco de estimación, dado que las distribuciones estacionarias son conocidas o se pueden calcular cuando se diseñan las caminatas aleatorias, se conoce la probabilidad de ser elegido y, por lo tanto, el peso de la muestra (Zhang 2021).

Un principio comúnmente utilizado para construir estimadores imparciales es emplear el estimador de Horvitz y Thompson (1952), para estimar el parámetro de interés (Zhang 2021). Este tema se abordará más adelante.

La caminata aleatoria simple tiene dos desventajas principales en el muestreo de redes. La primera es que tiende a seleccionar nodos con un grado más alto debido a su distribución estacionaria, lo que hace que las muestras estén sesgadas hacia nodos de mayor grado, lo que conduce a una baja precisión (Qi 2022).

La segunda desventaja es que puede quedar atrapado en un grupo estrechamente unido al nodo inicial. Para superar estas desventajas, se proponen la caminata aleatoria - Metropolis Hasting y la caminata aleatoria con escape para corregir las limitaciones mencionadas anteriormente (Qi 2022).

### **Muestreo caminata aleatoria con escape**

La caminata aleatoria con escape es una variante ampliamente utilizada de la caminata aleatoria simple puesto que emplea una forma adicional de saltar a un nodo aleatorio arbitrario

en el conjunto de nodos.

En la caminata aleatoria simple, la caminata elige un vecino del nodo actual. Pero en este caso, la caminata elige un nodo con probabilidad  $\alpha$  para saltar a un nodo que no sea vecino y  $1 - \alpha$  para continuar con un vecino actual. Si el caminante elige continuar, entonces se elige un vecino del nodo actual de la misma manera que la caminata aleatoria simple, si se elige el salto aleatorio, se elige un nodo aleatoriamente de todo el grafo (Qi 2022).

La caminata aleatoria con escape se originó originalmente a partir del algoritmo PageRank de Google (Langville & Meyer, 2006) para dos propósitos:

- 1) Hacer que la cadena sea tanto aperiódica como irreducible
- 2) Ampliar el espacio espectral para que la cadena pueda converger más rápido.

Los algoritmos basados en caminatas aleatorias se utilizan para explorar redes desconocidas, pero es imposible alcanzar todo el grafo cuando utilizamos la caminata aleatoria o sus variantes. Además, a veces la caminata aleatoria con escape no puede evitar que el algoritmo salte al mismo grupo en la que solía estar atrapado, especialmente cuando el mismo es grande y está muy unido.

El pseudocódigo detallado se encuentra a continuación (Qi 2022).

### Algoritmo

- 1- Escoge un nodo  $v_0$  de manera uniforme  $V$  como semilla
- 2- Selecciona la probabilidad de salto  $\alpha$
- 3-  $i \leftarrow 0$
- mientras**  $i < n$  **entonces**
  - 3-1 De manera aleatoria seleccione un número  $r \sim U(0, 1)$
  - si**  $r > \alpha$  **entonces**
    - 3-2 Escoga un nodo  $u$  uniformemente de  $V$
    - 3-3  $v_{i+1} \leftarrow u$
  - en cambio**
    - 3-4 Busque los vecinos del nodo actual  $v_i$
    - 3-5 Selecciona un nodo  $u$  uniformemente de  $N(v_i)$
    - 3-6  $v_{i+1} \leftarrow u$
  - 3-7  $i \leftarrow i + 1$
- 4-terminar

### Muestreo caminata aleatoria - Metropolis Hasting

La caminata aleatoria Metropolis-Hasting agrega un filtro en cada paso de selección de vecinos de la caminata aleatoria simple. Similar al algoritmo de Metropolis-Hasting en el conjunto de números reales, primero elegimos un vecino del nodo actual como el nodo candidato, luego comparamos la cantidad de un número aleatorio y la proporción de grados entre el nodo actual y el nodo candidato para decidir si aceptar o no el nodo candidato (Qi 2022). El pseudocódigo detallado se muestra a continuación (Qi 2022).

#### Algoritmo

1. Elegir un nodo  $v_0$  del conjunto de nodos  $V$  como semilla
2. Mientras  $i < n$ :
  - a) Encontrar los vecinos del nodo actual  $v_i$
  - b) Elegir un nodo  $u$  uniformemente de entre esos vecinos de  $v_i$
  - c) Elegir un número aleatorio  $r \sim U(0,1)$
  - d) Si  $r < \frac{d(u)}{d(v_i)}$ ; donde  $d(i)$  corresponde al grado del nodo  $i$ .
    - $v_{i+1} \leftarrow u$
  - e) Sino:
    - $v_{i+1} \leftarrow v_i$
  - f)  $i \leftarrow i + 1$
  - g) **terminar**

Teóricamente, el algoritmo Metropolis-Hasting, como una técnica de MCMC, puede simular cualquier distribución. Sin embargo, en el área de muestreo de redes, se prefieren muestras uniformes, es decir, cada nodo en el grafo original tiene igual probabilidad de ser elegido a través de un recorrido aleatorio. Por lo tanto, los investigadores suelen utilizar el término caminata uniforme al mencionar y/o emplear la caminata aleatoria Metropolis-Hasting (Thompson 2006).

### Muestreo de caminata aleatoria simple con reemplazo

Como la caminata aleatoria simple presenta inconvenientes los cuales fueron comentados anteriormente, se idearon combinaciones simples para superar los mismos.

A partir de estos inconvenientes se propuso la caminata aleatoria simple con reemplazo para calcular la afinidad entre nodos, y en Leskovec, Kleinberg & Faloutsos (2007) lo implementaron en la tarea de muestreo de redes.

Suponiendo que el punto de partida de la caminata aleatoria es el nodo  $v_0$ , la caminata aleatoria simple con reemplazo establece una probabilidad no nula y fija para que la caminata aleatoria regrese al nodo de inicio  $v_0$  con una probabilidad  $p$ , la diferencia clave con la caminata aleatoria simple puesto que implica que el algoritmo pueda volver a su punto de origen. (Qi 2022). El pseudocódigo detallado se muestra a continuación (Qi 2022).

### Algoritmo

**1** Elegir un nodo  $\mathbf{v}_0$  del conjunto de nodos  $\mathbf{V}$  como semilla.

Mientras  $\mathbf{i} < \mathbf{n}$ :

**2-1** Muestrear un número aleatorio  $\mathbf{r} \sim \mathbf{Bernoulli}(\mathbf{p})$ .

Si  $\mathbf{r} = \mathbf{0}$ :

**2-2** Encontrar los vecinos del nodo actual  $\mathbf{v}_i$ .

**2-3** Elegir un nodo  $\mathbf{u}$  uniformemente de entre esos vecinos de  $\mathbf{v}_i$ .

$\mathbf{v}_{i+1} \leftarrow \mathbf{u}$ .

Si  $\mathbf{r} = \mathbf{1}$ :

$\mathbf{v}_{i+1} \leftarrow \mathbf{v}_0$ .

**3-  $\mathbf{i} \leftarrow \mathbf{i} + \mathbf{1}$ .**

**4-terminar**

La caminata aleatoria simple con reemplazo puede explorar el grafo de manera más exhaustiva y producir un subgrafo conectado, ya que todos los nodos muestreados están conectados al punto de inicio. Se observa que la caminata aleatoria simple con reemplazo es un caso específico de la caminata aleatoria con escape puesto que reinicia el muestreo en cualquier nodo del grafo, mientras que este método lo hace en el nodo  $v_0$ .

### Muestreo bola de nieve

A continuación se define otro tipo de muestreo en redes comúnmente usado el cual corresponde al muestreo bola de nieve, el cual consiste en un muestreo de varias etapas. Que se realiza de la siguiente manera, se toma una muestra inicial de nodos con alguna probabilidad de selección usual sin reemplazamiento  $A$ , en la siguiente etapa de muestreo los nodos que no fueron seleccionados en  $A$  o sea en el subconjunto  $G \cap A^c$ , sus probabilidades de selección en la segunda etapa de este muestreo serán recalculadas teniendo en cuenta los enlaces a los nodos seleccionados en la primera etapa, con esto se generaría otro subconjunto de elementos  $B$  y este procedimiento se realizaría iterativamente hasta conseguir el número de elementos a muestrear deseados, definición dada a partir de Carrington, Scott & Wassermann (2005).

En Lavallée (2007) se comenta que el muestreo bola de nieve y muestreos basados en el mismo podían también ser vistos como muestreos probabilísticos si se tenía en cuenta la aleatoriedad que genera la cantidad de individuos que se muestrean en etapas siguientes a la primera, esta aleatoriedad hace que el muestreo bola de nieve se vea como probabilístico y por lo tanto sea posible obtener las expresiones mostradas a continuación.

Tal como se comenta en Trujillo, Niño & Hernández (2016), se tiene la siguiente expresión para el total de la variable, total estimado y varianza estimada, para un muestreo bola de nieve con 2 etapas de selección de nodos y enlaces.

$$\begin{aligned}
T^B &= \sum_{i=1}^{N^s} y_i = \sum_{i=1}^{N^s} \left( \sum_{j=1}^{N^A} \tilde{\theta}_{ji}^{AB} \right) y_i \\
&= \sum_{i=1}^{N^s} \sum_{j=1}^{N^A} \tilde{\theta}_{ji}^{AB} y_i \\
&= \sum_{j=1}^{N^A} \sum_{i=1}^{N^s} \tilde{\theta}_{ji}^{AB} y_i \\
&= \sum_{j=1}^{N^A} \left( \sum_{i=1}^{N^s} \tilde{\theta}_{ji}^{AB} y_i \right)
\end{aligned}$$

Donde  $y$  refiere a la variable de interés medida tanto en  $A$  como en  $B$ ,  $\tilde{\theta}_{ji}^{AB}$  se refiere al peso ponderado entre el conjunto  $A$  y  $B$  definida por Lavallée (2007),  $\pi_i^A$  es igual a la probabilidad de inclusión de primer orden del elemento  $i$  bajo la muestra obtenida en el subconjunto  $A$ ,  $\pi_{ij}$  se define como la probabilidad de inclusión de segundo orden de los elementos  $i, j$ , bajo la muestra obtenida en el subconjunto  $A$ ,  $N^A$  corresponde al tamaño del subconjunto  $A$ ,  $N^B$  al tamaño del subconjunto  $B$ ,  $Z_j$  se define como una variable indicadora que toma el valor de 1 si el elemento  $j$  pertenece a la muestra, 0 en caso contrario.

$$\begin{aligned}
\hat{T}^B &= \sum_{j=1}^{N^A} \frac{y_j}{\pi_j^A} \left( \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i \right) \\
&= \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \frac{y_j}{\pi_j^A} \tilde{\theta}_{ji}^{AB} y_i = \sum_{i=1}^{N^B} \sum_{j=1}^{N^A} \frac{y_j}{\pi_j^A} \tilde{\theta}_{ji}^{AB} y_i \\
&= \sum_{i=1}^{N^B} w_i y_i
\end{aligned}$$

$$\hat{V}(\hat{T}^B) = \sum_{j=1}^{N^A} \sum_{i=1}^{N^A} \frac{\pi_{ji}^A - \pi_j^A \pi_i^A}{\pi_j^A \pi_i^A} Z_j Z_i y_j y_i$$

### 2.3.3. Métodos de muestreo en redes conectadas

#### Muestreo dirigido por el encuestado (Respondent-Driven Sampling)

Otro tipo de diseño muestral usado para el muestreo en redes corresponde a una variante del muestreo bola de nieve, y corresponde al muestreo dirigido por el encuestado, este tipo de diseño muestral es usado comúnmente en poblaciones de difícil acceso que tienen individuos con características difíciles de encontrar, los cuales en ciertos contextos pueden corresponder a miembros de poblaciones ilegales o estigmatizados tales como por ejemplo, personas con cierta enfermedad extraña, o consumidores de droga (Trujillo et al. 2016).

Esta técnica consiste en tomar un muestreo en varias etapas tal como el muestreo bola de nieve, con la diferencia es que en la primera etapa seleccionamos un número  $i$  de individuos que cumplen la característica que estamos analizando y posteriormente en base a sus enlaces y variables auxiliares se toma un determinado número  $j$  de individuos que estén en lanzados con los  $i$  individuos de la primera muestra, de aquí se evalúa si el individuo posee la característica si esto ocurre continua en la muestra, en caso contrario el mismo es eliminado de la muestra, así se continua sucesivamente hasta obtener el tamaño muestral deseado para los individuos con la característica deseada (Trujillo et al. 2016).

El estimador propuesto para la estimación de totales bajo este diseño muestral fue propuesto por Salganik & Heckathorn (2004), basado en el concepto de caminatas aleatorias de la misma forma que en la sección 2.3.2 de este documento, donde la probabilidad de inclusión es calculada mediante los nodos que dan origen al nodo  $i$ , o de forma más concreta con el factor de expansión de cada nodo esta dado por su grado (Gile et al. 2018).

$$\hat{T} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{m_i}$$

Donde  $y_i$  es la variable de interés del  $i$ -ésimo nodo,  $m_i$  es el grado del nodo  $i$ -ésimo y  $N$  es el tamaño de la población.

Con la varianza se tiene el mismo inconveniente que el mencionado en la sección 2.3.2 puesto que no se pueden construir estimadores para ordenes superiores a 1, y la varianza al asociar dos nodos en su concepción, se concibe como una estimación de orden 2 (Zhang & Patone 2017).

### **Muestreo trazo de ruta (Trace-Route Sampling)**

Por ultimo de los diseños muestrales más usados en redes se encuentra el muestreo trazo de ruta, el cual consiste tal como su nombre lo indica en trazar una ruta entre los individuos muestreados, y su procedimiento se explica a continuación (Trujillo et al. 2016):

- 1 Se extrae una muestra de nodos
- 2 Se definen unos nodos objetivos, los cuales cumplen la característica de interés.
- 3 Para cada par de individuos de la muestra y de los nodos objetivos, se establece una ruta a través de otros nodos a partir de los cuales se pueda conectar el nodo inicial de la muestra y el nodo objetivo, a partir de aquí todos los nodos en esta ruta estarán en la muestra.

Este muestreo es comúnmente usado en problemas de redes computacionales y su complejidad incurre en todos los posibles caminos que puedan existir para llegar de un nodo inicial a un nodo objetivo, esto recurre en problemas de optimización y de cálculos computacionales complejos, por lo cual este tipo de diseño no es muy usado en la practica de redes para individuos y se usa más que todo en problemas computacionales donde se tienen redes unidireccionales y una cantidad de nodos reducida (L'heureux et al. 2017).

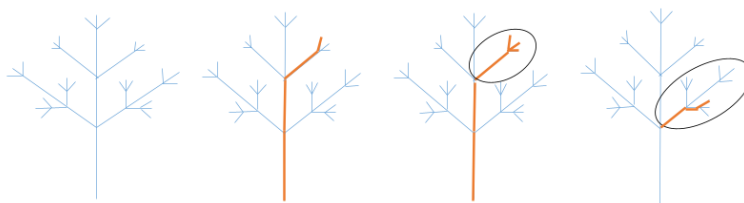
### Muestreo de rama aleatoria (Randomized Branch Sampling)

Es un muestreo similar al visto anteriormente muestra trazo de ruta, puesto que los nodos y enlaces generan un camino desde una unidad inicial hasta una posible unidad final, el mismo se puede ver como un diagrama de árbol, donde la unidad inicial es la base y luego se va pasando a través de distintas ramas hasta llegar al punto objetivo (Jessen, 1955, Gregoire & Valentine, 2007; Trujillo et al, 2016).

La terminología en este documento sigue a Gregoire & Valentine (2007). Se define una rama como el tallo completo que se desarrolla a partir de una parte lateral del árbol que es generada de un nodo final, y se define un segmento como una parte de una rama entre dos nodos consecutivos. El extremo inferior del tallo principal de un árbol se considera un nodo y el árbol se considera una rama. Los laterales del árbol se consideran tanto ramas como segmentos de rama. Por lo tanto, cualquier árbol o rama se puede definir como una población de segmentos de rama.

Se define camino como varios segmentos de rama conectados. Un camino puede extenderse desde el extremo inferior del tallo principal hasta un nodo final, en cuyo caso el número de caminos posibles es igual al número de nodos terminales. Sin embargo, el término no necesita ser un brote terminal y el punto de inicio de un camino no necesita ser el extremo inferior del tallo principal (Gregoire & Valentine 2007).

En la siguiente figura se mostrarán diagramas de cómo se puede ver representado el muestreo de rama aleatoria y las distintas formas de muestreo - caminos entre ramas para llegar a un nodo, teniendo en cuenta que se puede iniciar el camino desde la base del árbol o desde una rama en específico, esto influirá en la estimación a posteriori (Trujillo et al. 2016).



**Figura 2-2:** Muestreo de rama aleatoria - tomado de Trujillo et al. (2016)

En donde se define  $q_i$  la probabilidad de selección del nodo  $i$ ,  $Q_r$  la probabilidad de selección del segmento o rama  $r$  y  $Q_{ir}$  como la probabilidad de selección del nodo  $i$  mediante la rama  $r$ .

A partir de aquí se trata la etapa de selección del camino, para esto se tiene que la probabilidad de seleccionar una rama determinada depende del nodo de donde provenga y la rama asignada anteriormente, por lo que seleccionar una rama se convierte en una probabilidad

condicional dependiendo de la rama anteriormente seleccionada, por lo que seleccionar un camino desde el punto inicial al punto final se convertirá en la multiplicación de múltiples probabilidades condicionales asociados a las ramas a evaluar, por ejemplo si se desea estimar la probabilidad de que la  $R$ -ésima rama sea seleccionada la probabilidad se vera como a continuación (Trujillo et al. 2016).

$$Q_r = \prod_{k=1}^r q_k, \quad r = 1, \dots, R$$

$$Q_1 = q_1$$

$$Q_2 = q_1 q_2 = q_2 Q_1$$

$$\dots$$

$$Q_r = q_1 q_2 \dots q_R = q_R Q_{R-1}$$

Ahora según la definición dada en (Zhang 2021), se tiene que el estimador bajo el muestreo de rama aleatoria viene dado para una variable  $y_{ir}$  para el  $r$ -ésimo segmento del  $i$ -ésimo nodo y  $t_y$  la variable objetivo a estimar, se tiene que un estimador insesgado para la variable objetivo en esa rama vendría dado por

$$\hat{t}_{yQ_i} = \sum_{r=1}^R \frac{y_{ir}}{Q_{ir}}$$

y para tomar el estimador para  $m$  un número entero, mayor a 2 ramas el mismo se vería reflejado como un promedio de los totales por rama, el cual daría una estimación al total del camino elegido, el estimador viene dado por.

Se define ahora la varianza de este estimador, como :

$$V(\hat{t}_{yQ}) = \frac{1}{m} \left[ \sum_{i=1}^M Q_{ir} (\hat{t}_{yQ_i} - t_y)^2 \right]$$

Con su estimador de varianza definido , como:

$$\hat{V}(\hat{t}_{yQ}) = \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{t}_{yQ_i} - \hat{t}_{yQ})^2$$

## 2.4. Aprendizaje no supervisado

En esta sección se tratará las definiciones principales del aprendizaje no supervisado y adicionalmente los distintos métodos que pueden ser aplicados en el mismo, puesto que esta metodología es usada como una parte fundamental en el capítulo de aplicación de este documento, debido a que con el aprendizaje no supervisado se crearán conglomerados, que tomaremos en el muestreo como estratos.

El término aprendizaje no supervisado se asocia genéricamente con la idea de utilizar una colección de observaciones  $y_1, \dots, y_n$  muestreadas a partir de una distribución  $p(\mathbf{Y})$  para describir propiedades de  $p(\mathbf{Y})$  Michalski, Carbonell & Mitchell (2013).

El objetivo del aprendizaje no supervisado es identificar patrones, estructuras ocultas, relaciones o características intrínsecas en los datos sin saber de antemano cuáles son. Los algoritmos de aprendizaje no supervisado exploran el conjunto de datos y, a menudo, tratan de agrupar puntos de datos similares en conglomerados o realizar reducción de dimensionalidad para representar los datos de una manera más manejable.

En la práctica, el término aprendizaje supervisado se refiere a Najafabadi et al. (2015):

- **closterización:** Procedimientos que identifican grupos en los datos.
  
- **Análisis de componentes principales:** Método estrictamente relacionado con la descomposición en valores singulares y la transformada de Karhunen-Loève, que trata de encontrar variables no correlacionadas obtenidas como combinaciones lineales de las variables originales, los campos relacionados son el análisis factorial y el análisis de la varianza (ANOVA).
  
- **Modelos de asociación:** Técnica de minería de datos que encuentra colecciones de atributos (mejor, valores de atributos) que aparecen juntos con frecuencia en las observaciones.
  
- **Escalado multidimensional:** Proceso que consiste en identificar un espacio euclidiano de dimensiones reducidas, y una cartografía posiblemente no lineal del espacio original al nuevo espacio, de forma que la distancia entre pares de puntos de entrenamiento en el espacio original sea casi igual a las distancias entre sus proyecciones (Najafabadi et al. 2015).

Al igual que aprendizaje no supervisado, closterización es un término complejo de definir, algunas de las definiciones asociadas al mismo son:

- El proceso de encontrar grupos en los datos.
- El proceso de dividir los datos en grupos homogéneos.
- El proceso de dividir los datos en grupos, donde los puntos dentro de cada grupo son cercanos (o similares) entre sí.
- El proceso de dividir los datos en grupos, donde los puntos dentro de cada grupo están cerca (o son similares) entre sí, y donde los puntos de diferentes grupos están lejos (o no son similares) entre sí.
- El proceso de dividir el espacio de características en regiones con una densidad de puntos relativamente alta, separadas por regiones con una densidad de puntos relativamente baja.

Estas definiciones no son equivalentes. Por ejemplo, encontrar grupos en el conjunto de datos no es lo mismo que dividir el conjunto de datos en grupos homogéneos (Najafabadi et al. 2015).

Además, las definiciones son genéricas: no especifican qué son los términos grupo, qué significan realmente homogéneo, cercano, lejano, densidad relativamente alta, densidad relativamente baja.

Esto incurre en que existan problemas al plantear y entender los algoritmos usados tradicionalmente, puesto que en general, no se tiene idea de lo que hace un método de agrupación hasta que se ve su especificación formal y algorítmica.

### 2.4.1. Análisis de métodos de closterización

Las distintas definiciones aquí descritas fueron definidas por (Michalskin et al. 2013; Najafabadi et al 2015; Grolinger, Eliyamany & Capretz (2017); Dangeti 2017) con algunos cambios dados en este documento. Con estas definiciones se pueden reconocer una variedad de características que ayudan a describir los algoritmos de closterización.

- **Métodos jerárquicos vs. particionales:**

Los algoritmos de closterización jerárquico inducen sobre los datos una estructura de closterización parametrizada por un parámetro de similitud. Una vez finalizada la fase de aprendizaje, el usuario puede obtener inmediatamente diferentes agrupaciones de datos especificando diferentes valores del índice de similitud. Los métodos particionales producen esencialmente una partición de los datos en conglomerados.

- **Métodos aglomerativos vs Métodos divisivos**

Los métodos aglomerativos comienzan asignando cada muestra a su propio conglomerado y proceden a fusionar los conglomerados. Los métodos divisivos comienzan asignando todas las muestras a un único conglomerado y proceden a dividir los conglomerados.

- **Métodos monotéticos vs Métodos politéticos**

Los métodos monotéticos aprenden los conglomerados utilizando una característica cada vez. Los métodos politéticos utilizan colecciones de características como por ejemplo mediante combinaciones lineales. La gran mayoría de los métodos de closterización son politéticos.

- **closterización duro vs closterización difuso**

En una closterización duro los puntos pertenecen a un único conglomerado, mientras que en la closterización difuso existe la posibilidad de que un punto pertenezca a uno o más conglomerados.

- **Uso de algoritmos deterministas vs algoritmos estocásticos:**

Si todos los pasos del algoritmo de agrupación son deterministas, el método es determinista. Algunos métodos utilizan pasos aleatorios y corresponden a métodos estocásticos.

- **closterización incremental vs. closterización no incremental:**

Algunos métodos necesitan todos los puntos de muestra desde el principio. Otros pueden iniciarse con menos muestras y perfeccionarse de forma incremental, estos últimos métodos son más adecuados para grandes conjuntos de datos.

Como ya se ha mencionado, el término métodos jerárquicos es muy general. En la práctica, se utiliza para designar una clase muy específica de métodos que funcionan de la siguiente manera asignan cada muestra a un conglomerado distinto, luego tal como se comenta en Michalski et al. (2013)

- Agrupar el par de conglomerados para el que un criterio de similitud es más pequeño.
- Asignar a la fusión el valor de la similitud entre los conglomerados fusionados.

- Iterar los 2 pasos anteriores hasta que todos los datos se agrupen en un único conglomerado.

El proceso de agrupación puede representarse mediante un árbol denominado dendrograma. Un dendrograma difiere de un árbol normal en que la dirección desde la raíz hasta las hojas se interpreta como un eje de similitud. Las fusiones se dibujan entonces en la coordenada correspondiente al valor de similitud asignado.

Una ventaja del dendrograma es que se puede dividir el árbol en un valor de similitud deseado y obtener un árbol con nodos interiores correspondientes a valores de similitud menores que el especificado. Por lo tanto, se puede construir una gran variedad de árboles cortando el dendrograma completo.

En Duda, Hart y Stork (2000) mencionan que a partir del dendrograma se puede determinar si los grupos son naturales: si las divisiones en el dendrograma se extienden por todos los valores de la escala de similitud, entonces no hay un número natural de conglomerados.

La forma de leer esta afirmación es la siguiente: una vez que seleccionamos una métrica y un índice de distancia entre conglomerados, hemos definido implícitamente lo que es un conglomerado.

Por lo tanto, el hecho de que el dendrograma sugiera que existe un número natural de conglomerados para un problema específico no tiene por qué coincidir con el juicio de un usuario experto que analice los datos Hastie, Tibshirani y Friedman (2009).

Una afirmación similar, pero más precisa, aparece en Hastie et al. (2009). Aquí los autores mencionan una forma algo objetiva de juzgar lo bien que el algoritmo de agrupación jerárquica representa los datos y es basado en el coeficiente de correlación cofenética definido por primera vez en Farris (1969).

El coeficiente de correlación cofenética es una medida estadística utilizada para evaluar la calidad de la representación de los datos obtenida mediante un algoritmo de agrupación jerárquica.

Este coeficiente mide la correlación entre las distancias originales entre las muestras y las distancias cofenéticas, que son las disimilitudes obtenidas a través del dendrograma resultante del proceso de agrupación. El procedimiento para calcular el coeficiente de correlación cofenética implica los siguientes pasos (Farris 1969):

- Se calculan todas las distancias entre pares de muestras originales. Esto se hace midien-

do la disimilitud o similitud entre cada par de muestras según una métrica específica (por ejemplo, distancia euclidiana o coeficiente de correlación).

- Se realiza el proceso de agrupación jerárquica utilizando el algoritmo adecuado. El resultado es un dendrograma que muestra la jerarquía de conglomerados y cómo las muestras se agrupan entre sí.
- Se obtienen las distancias cofenéticas para cada par de muestras del dendrograma. Estas distancias representan la altura del nodo en el que las muestras se fusionan en el mismo conglomerado.
- Se crea un diagrama de dispersión de 2 dimensiones con los puntos correspondientes a los pares de muestras. Cada punto tiene dos coordenadas: la disimilitud original entre las muestras y su distancia cofenética.
- Finalmente, se calcula el coeficiente de correlación entre las dos coordenadas de los puntos del diagrama de dispersión. Este coeficiente indica qué tan bien la jerarquía de conglomerados captura las distancias originales entre las muestras. Un coeficiente de correlación alto sugiere una mejor representación y similitud entre las distancias originales y las cofenéticas, mientras que un coeficiente bajo indica una representación deficiente.

Cuando aplicamos la agrupación jerárquica a un conjunto de datos, generamos una jerarquía de conglomerados basada en las similitudes o disimilitudes entre las muestras. El coeficiente de correlación cofenética nos permite evaluar si esta jerarquía captura adecuadamente las relaciones de similitud originales presentes en los datos (Farris 1969).

Un valor alto del coeficiente de correlación cofenética indica que la jerarquía de conglomerados representa de manera fiel las distancias originales entre las muestras. Es decir, el dendrograma captura adecuadamente las relaciones de similitud entre los datos (Farris 1969).

Por otro lado, un valor bajo del coeficiente de correlación cofenética indica que la jerarquía de conglomerados no representa de manera precisa las distancias originales entre las muestras. Esto puede deberse a que el algoritmo de agrupación jerárquica no logra capturar las relaciones de similitud entre los datos de manera efectiva (Farris 1969).

En resumen, el objetivo de utilizar el coeficiente de correlación cofenética es proporcionar una medida objetiva de qué tan bien el algoritmo de agrupación jerárquica representa las relaciones de similitud originales entre los datos. Esto es esencial para evaluar la calidad y eficacia de la técnica de agrupación y para tomar decisiones informadas sobre la adecuación del método para analizar el conjunto de datos específico.

### 2.4.2. $k$ -Means, y algoritmos relacionados

El algoritmo  $k$ -Means se asocia comúnmente con la minimización de un criterio de error cuadrático. Más concretamente, el error cuadrático de un conjunto de datos  $\mathcal{D}$  de  $N$  puntos con respecto a un conglomerado.

$$e^2(\mathcal{D}, \text{cal}C) = \sum_{k=1}^K \sum_{i=1}^N 1_{\{\mathbf{y}_i \in C_k\}} \|\mathbf{y}_i - \mathbf{c}_k\|^2$$

Donde  $\|\mathbf{y}_i - \mathbf{c}_k\|^2$  es la distancia euclídeana al cuadrado entre el dato ( $\mathbf{y}_i$ ) y la muestra representativa  $\mathbf{c}_k$  del  $k$  grupo, y  $1_{\{\mathbf{y}_i \in C_k\}}$  es igual a 1 si  $\mathbf{y}_i$  pertenece al conglomerado  $C_k$ , y a cero en caso contrario (Michalski et al. 2013).

L'heureux et al. (2017) cita dos criterios de optimalidad que debe satisfacer un algoritmo de agrupación de esta clase para producir una solución que sea un mínimo local de la función objetivo:

- **Criterio 1**

La muestra representativa  $c$  de un conglomerado  $\mathcal{C}$  debe ser el punto que minimiza

$$\sum_{\mathbf{y} \in \mathcal{C}} \|\mathbf{y} - \mathbf{c}\|^2$$

De forma más general, minimizar la función objetivo sobre las muestras del conglomerado (Michalski et al. 2013).

- **Criterio 2**

Un punto  $\mathbf{y}$  debe asignarse al conglomerado  $\mathcal{C}^*$  cuyo centroide  $\mathbf{c}^*(\mathbf{y})$  sea el más cercano a  $\mathbf{y}$ .

$$\mathbf{c}^*(\mathbf{y}) = \arg \min_k \|\mathbf{y} - \mathbf{c}_k\|^2$$

De de forma más general, minimizar la función objetivo para la muestra  $\mathbf{y}$  (Michalski et al. 2013).

El algoritmo  $k$ -Means realiza una búsqueda heurística satisfaciendo ambos requisitos. El algoritmo funciona del siguiente modo (Michalski et al. 2013):

1. Realiza una selección inicial de centroides.
2. Iterar sobre las muestras, para cada muestra encontrar el centroide más cercano y asignar la muestra al conglomerado correspondiente, además de recalcular el centroide de ese conglomerado.
3. Repita el paso 2 hasta que se cumpla un criterio de convergencia.

Los criterios de convergencia típicos son comentados a continuación:

- La finalización del proceso después de una iteración en la que la pertenencia a los conglomerados de las muestras no cambia.
- La finalización del proceso después de una iteración en la que los centroides no cambian.
- La finalización del proceso después de una iteración en la que el valor de la función objetivo no cambia.
- La finalización del proceso después de alcanzar un número máximo de iteraciones.

Cuando el conjunto de datos es grande la convergencia puede ser algo lenta, y los criterios de finalización anteriores se sustituyen por criterios de umbral (por ejemplo, terminar después de una iteración en la que la función objetivo disminuye menos de  $\epsilon$ , etc.).

Un método que se parece mucho al algoritmo  $k$ -Means es el algoritmo de Cuantización Vectorial (VQ) de Linde, Buzo & Gray (1980), también conocido como LBG, o como Algoritmo Lloyd Modificado:

- 1. Realizar una selección inicial de centroides.
- 2 Para cada muestra, encontrar el centroide más cercano, asignar la muestra al conglomerado correspondiente.
- 3. Después de iterar sobre todas las muestras, volver a calcular los centroides utilizando la nueva asignación de puntos a los conglomerados.
- 4. Repita el paso 2 hasta que se cumpla un criterio de convergencia.

Se puede observar aquí una similitud paralela; el  $k$ -Means y el LBG tienen la misma relación que el algoritmo de entrenamiento original Perceptron (Duda et al. 2000). Como en ese caso, es difícil decir si el original  $k$ -Means funciona mejor que el algoritmo LBG. LBG tiene una clara ventaja computacional; el cálculo de las distancias entre puntos y centroides puede hacerse de forma más eficiente, porque ni los puntos ni los centroides cambian durante el cálculo (Duda et al. 2000).

### Variaciones del algoritmo $k$ -Means

Aquí describimos algunas extensiones de los algoritmos  $k$ -Means.

La primera intenta modificar el algoritmo para adaptarlo más a la cuarta definición de closterización, permitiendo la división y fusión de conglomerados. Al final de cada iteración, calculamos la dispersión dentro de un conglomerado como por ejemplo, la distancia cuadrática media de los puntos a sus centroides y las medidas de dispersión entre conglomerados por ejemplo, la distancia cuadrática media entre los puntos de un conglomerado y el centroide de otro conglomerado (Michalski et al. 2013).

Los conglomerados con una dispersión interna cercana a la dispersión media entre conglomerados se declaran demasiado dispersos y se dividen. Los pares de conglomerados en los que la dispersión dentro del conglomerado está cerca de la dispersión entre conglomerados correspondiente (es decir, en los que los puntos de un conglomerado no están demasiado lejos del centroide del otro) se fusionan (Michalski et al. 2013).

Otra clase de métodos intenta mejorar los  $k$ -Means seleccionando adecuadamente los valores iniciales de los centroides. De este modo, se espera garantizar una convergencia más rápida hacia una mejor solución. Por ejemplo, se podría entrenar un cuantificador vectorial estructurado en árbol y utilizar los centroides producidos como semillas para el algoritmo  $k$ -Means; la esperanza en este caso es que los puntos de inicio se adapten mejor a la distribución de los datos que los seleccionados al azar (Michalski et al. 2013).

En la práctica, es discutible que estos enfoques funcionen realmente como se desea. Si observamos lo que ocurre durante las primeras iteraciones del  $k$ -Means original, con centroides iniciales seleccionados al azar, veremos que las posiciones de los centroides varían mucho y que la función objetivo disminuye rápidamente. Los centroides finales suelen ser muy diferentes de los iniciales. Tras las primeras iteraciones, el método identifica finalmente un mínimo local, y a partir de este punto observamos una mejora ordenada pero más lenta de la función objetivo mientras el algoritmo converge.

También suele observarse el mismo comportamiento en los otros métodos descritos anteriormente: los centroides finales suelen ser completamente distintos de los iniciales. Es concebible, sin embargo, que una selección apropiada de centroides pueda prevenir la convergencia a mínimos locales particularmente malos que resultarían de selecciones aleatorias de centroides errados (Michalski et al. 2013).

Por último, mencionamos la posibilidad de utilizar diferentes clases de funciones objetivo. Por ejemplo, la minimización de la distancia euclidiana media produce el enfoque  $k$ -Median o  $k$ -Medoid.

### 2.4.3. Índice de silueta

El índice de silueta es una métrica utilizada para evaluar la calidad de los agrupamientos (conglomerados) obtenidos en un conjunto de datos. Proporciona una medida de cuán bien cada punto de datos se ajusta a su propio conglomerado en comparación con los conglomerados vecinos más cercanos. El índice de silueta varía entre -1 y 1, donde valores más cercanos a 1 indican una mejor separación de los conglomerados, valores cercanos a 0 indican una superposición entre conglomerados y valores cercanos a -1 indican que los puntos podrían estar asignados al conglomerado incorrecto (Rousseeuw 1987).

La fórmula general para calcular el índice de silueta para un punto de datos  $i$  es la siguiente (Kaufman & Rousseeuw 1990):

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Donde:

- $a(i)$  es la distancia promedio entre el punto  $i$ -ésimo y todos los demás puntos en el mismo conglomerado (medida de cohesión).
- $b(i)$  es la distancia promedio entre el punto  $i$ -ésimo y todos los puntos en el conglomerado vecino más cercano (medida de separación).
- $\max(a(i), b(i))$  se utiliza para asegurarse de que el índice esté en el rango entre -1 y 1.

Un índice de silueta alto indica que el punto está bien asignado a su conglomerado y que su conglomerado está bien separado de los conglomerados vecinos. Por otro lado, un índice de silueta bajo indica que el punto podría estar más cerca de los puntos de otro conglomerado, sugiriendo una asignación de conglomerado deficiente (Hastie et al. 2009).

## 2.5. Método de estimación Monte Carlo

En esta sección se abordará el método de estimación Monte Carlo puesto que será usado en las estimaciones finales distintos muestreos que se abordaran en la parte de aplicación.

El método Monte Carlo es una técnica estadística que se utiliza para obtener resultados aproximados mediante la generación de múltiples muestras aleatorias. Se basa en la idea de realizar una simulación estadística repetida para obtener una estimación de parámetros o realizar análisis de incertidumbre (Harrison 2010).

En el contexto del método Monte Carlo, se generan numerosas muestras aleatorias con base en distribuciones de probabilidad apropiadas. Cada muestra se considera como una realización posible del fenómeno en estudio. Luego, se calculan los resultados para cada muestra y se obtiene una distribución de los resultados.

El objetivo del método Monte Carlo es proporcionar una aproximación numérica o estadística de una variable o función desconocida. Esto se logra mediante la generación de muestras aleatorias que siguen ciertas distribuciones de probabilidad conocidas o asumidas. Estas muestras aleatorias se utilizan para calcular estadísticas descriptivas o estimadores, como promedios, varianzas o percentiles (Harrison 2010).

El método Monte Carlo es particularmente útil cuando no es posible obtener una solución analítica exacta o cuando se tienen múltiples variables aleatorias interrelacionadas. Permite analizar sistemas complejos, realizar pronósticos o simulaciones, y evaluar el impacto de la incertidumbre en los resultados (Harrison 2010).

Para aplicar el método Monte Carlo, se sigue generalmente el siguiente proceso (Harrison 2010):

1. Definir el problema y establecer los parámetros relevantes.
2. Seleccionar las distribuciones de probabilidad apropiadas para las variables aleatorias involucradas.
3. Generar muestras aleatorias a partir de estas distribuciones de probabilidad.
4. Realizar los cálculos necesarios utilizando las muestras aleatorias generadas.
5. Repetir los pasos 3 y 4 un número suficiente de veces para obtener una estimación estable de los resultados.

6. Analizar los resultados obtenidos, calcular estadísticas descriptivas y/o inferenciales, y obtener conclusiones sobre el fenómeno en estudio.

Es importante destacar que el método Monte Carlo se basa en la aleatoriedad y la generación de muestras suficientemente grandes para obtener resultados confiables. A medida que aumenta el número de muestras generadas, se espera que las estimaciones se aproximen más a los valores verdaderos o teóricos.

En resumen, el método Monte Carlo es una técnica estadística que utiliza muestras aleatorias para aproximar y analizar variables o funciones desconocidas. Es una herramienta poderosa y versátil en la investigación científica, la modelización de sistemas complejos y la toma de decisiones en presencia de incertidumbre (Harrison 2010).

# 3 Aplicaciones de herramientas de muestreo al monitoreo de problemas fitosanitarios del cultivo de arroz en Colombia

## 3.1. Introducción

### 3.1.1. El cultivo de arroz mundialmente

Según el informe del Departamento de Agricultura de Estados Unidos (USDA) y Durand-Morat & Bairagi (2021), la producción mundial de los principales cereales (trigo, maíz y arroz) fue de aproximadamente 2.656 millones de toneladas en 2020, con un crecimiento promedio del 2.1 % entre 2000 y 2020. En cuanto al arroz, representó en promedio el 31 % de la producción conjunta de estos cereales, pero su participación disminuyó al 28 % en 2020, favoreciendo al maíz. La producción mundial de arroz paddy seco fue estimada en 751 millones de toneladas en 2020.

En términos regionales, Asia lidera la producción mundial de arroz paddy seco con el 90 %, seguida por África con el 5 % y América con el 4.8 %. En América, la producción alcanzó 37.6 millones de toneladas en 2020, siendo Brasil y Estados Unidos los principales productores.

En cuanto a la productividad del arroz, Australia tiene el mayor rendimiento con 10.4 toneladas por hectárea, mientras que Colombia se ubica en el puesto 21 con 5.78 toneladas por hectárea en riego y en el 35 con 4.42 toneladas por hectárea en secano. En el grupo 4 de países (producciones entre 2 y 10 millones de toneladas) donde se encuentra Colombia, la productividad promedio colombiana se ubicaría en el sexto lugar entre los 17 países. En América, Colombia se clasifica en la posición 12 entre 26 países, ocupando la décima posición al considerar la producción de riego y el puesto 15 en rendimientos en secano (Durand-Morat & Bairagi 2021).

### 3.1.2. El cultivo de arroz en Colombia

Según el Censo Nacional Agropecuario DANE (2014), aproximadamente el 19 % del área agrícola en Colombia se utilizó para sembrar cereales, de los cuales el 38 % correspondió al cultivo de arroz. La siembra de arroz representa alrededor del 7.3 % del total de área cultivada en el país, pero aumenta al 12.6 % si se excluyen otros cultivos agroindustriales. Datos del Ministerio de Agricultura y Desarrollo Rural en 2019 indican que el cultivo de arroz ocupó el 35 % del total de cultivos de ciclo semestral.

El cultivo de arroz en Colombia se realiza de forma mecanizada y manual, siendo la producción manual apenas el 1 % del total y se destina principalmente para consumo propio. El cultivo mecanizado se divide en sistemas de riego y secano, siendo el riego más común en las zonas arroceras del Centro, Costa Norte y Santanderes, mientras que el sistema de secano se practica principalmente en los Llanos Orientales y el Bajo Cauca, dependiendo del régimen de lluvias y la disponibilidad de agua.

Las áreas de siembra de arroz están concentradas en los departamentos de Casanare, Tolima, Meta, Sucre, Huila y Norte de Santander, que representan conjuntamente el 80.5 % del total. Sin embargo, la distribución de productores y Unidades Productoras Agropecuarias (UPA) no sigue la misma proporción, con Tolima destacándose con el 20 % del total nacional de productores y el 29 % de las UPA.

El país se divide en cinco zonas de producción arroceras: Centro, Llanos, Costa Norte, Bajo Cauca y Santanderes. Los Llanos han experimentado un crecimiento significativo del 75 % en el área cultivada en los últimos años debido a mejoras en infraestructura y bajos costos de producción. Por otro lado, la Costa Norte ha experimentado una disminución del 39 % en el área sembrada, y el área en la zona Centro ha permanecido constante.

En conclusión, el cultivo de arroz en Colombia ha mostrado una expansión en zonas de secano como los Llanos y el Bajo Cauca, mientras que las áreas de riego han disminuido. La siembra se realiza en dos períodos durante el año, con la mayor cantidad de hectáreas sembradas en el primer semestre bajo el sistema de secano. El segundo semestre concentra la siembra en zonas de riego, representando aproximadamente el 35 % del área total cultivada durante todo el año, estos datos son soportados por FEDEARROZ-DANE en el censo nacional arroceros (2016).

### 3.1.3. El papel de las enfermedades en los cultivos de arroz

Las enfermedades en el cultivo de arroz representan uno de los factores más limitantes debido a los diversos impactos negativos que generan como los descritos por Fedearroz (2021):

- Pérdidas en rendimiento: Las enfermedades pueden afectar el crecimiento y desarrollo de las plantas de arroz, lo que resulta en una disminución del rendimiento de la cosecha.
- Necesidad de aplicar fungicidas: Para controlar y prevenir la propagación de enfermedades, los agricultores a menudo deben aplicar fungicidas. Esta medida tiene costos asociados y puede ser un desafío para algunos agricultores con recursos limitados.
- Reducción en la calidad del arroz: Algunas enfermedades afectan la calidad del grano de arroz, lo que puede disminuir su valor comercial y afectar la aceptación del producto en el mercado.
- Aumento de los costos de producción: La aplicación de fungicidas, así como otras prácticas de manejo para controlar enfermedades, puede incrementar los costos de producción del cultivo de arroz.
- Impacto ambiental: El uso excesivo de fungicidas puede tener consecuencias negativas para el medio ambiente, contaminando suelos y cuerpos de agua.
- Riesgo de resistencia: El uso continuo y excesivo de fungicidas puede llevar al desarrollo de resistencia en los patógenos, lo que dificulta su control en el futuro.

Para abordar estos desafíos, es importante adoptar estrategias de manejo integrado de plagas (MIP). El MIP combina diversas prácticas, como el uso de variedades resistentes, rotación de cultivos, monitoreo temprano de enfermedades. Aquí es donde el muestreo juega un papel importante, debido a la capacidad de prevención que le da a los agricultores de la detección de una enfermedad en una etapa temprana del cultivo, entre más completo sea el diseño muestral que se use en la estimación de la enfermedad, tendremos mejores proyecciones de la enfermedad a largo plazo y un costo menor que hacer un estudio a nivel censo de todos los cultivos.

### 3.1.4. Piricularia en los cultivos de arroz

La enfermedad más significativa del cultivo de arroz a nivel mundial es la quemazón del arroz, también conocida como piriculariosis. Su impacto se debe tanto a su amplia distribución como a los daños que causa. Los ataques en la panícula son especialmente importantes, ya que afectan negativamente la rentabilidad del cultivo debido a la disminución en la calidad y cantidad de la cosecha. Sin embargo, las infecciones en las hojas también pueden afectar los rendimientos, llegando incluso a causar la muerte parcial o total de las mismas.

El manejo de esta enfermedad requiere la implementación de métodos de prevención que abarcan aspectos genéticos (cultivo de plantas resistentes o tolerantes), químicos (uso de fungicidas) y prácticas culturales. A continuación, se enumeran una serie de factores que contribuyen al desarrollo de la enfermedad, tomado de IFAPA (2004):

- Susceptibilidad de las variedades de arroz.
  
- En siembras tardías, la fase final del ciclo vegetativo de la planta coincide con condiciones ambientales propicias para el desarrollo de la enfermedad. En siembras tempranas, la etapa de formación del grano ocurre a temperaturas inferiores a las óptimas para el crecimiento del hongo.
  
- Densidades excesivas de plantas, que reducen la ventilación y aumentan la humedad relativa, además de prolongar el ciclo del cultivo.
  
- Aplicación excesiva de fertilizantes nitrogenados: el exceso de nitrógeno debilita las células de la epidermis de la planta, facilitando la penetración del hongo. El ataque de la enfermedad es más severo cuando se utilizan fertilizantes nitrogenados de acción rápida, como el sulfato de amonio, aplicados como cobertura.
  
- Periodos de alta humedad o presencia de rocío. Los periodos prolongados de 12-14 horas con una humedad relativa superior al 93 % favorecen el desarrollo de las esporas del hongo *Pyricularia oryzae*.
  
- Rango de temperaturas óptimas para el desarrollo del hongo. La temperatura adecuada para el crecimiento del hongo se encuentra entre los 15 y 35 °C, siendo la óptima entre los 24 y 28 °C.

- Presencia de lloviznas o nieblas prolongadas.
- Vientos suaves que facilitan la dispersión de las esporas.
- El estrés por sequía, común en otras regiones arroceras donde se somete al cultivo a períodos prolongados sin agua o donde se retira el agua demasiado pronto para la cosecha, también incrementa la susceptibilidad a esta enfermedad. Esto se debe a que el agua acumulada en el campo retiene calor durante las horas de sol y lo libera a la atmósfera durante la noche, lo que retrasa la formación de rocío, un elemento crucial para la infección.

Las infecciones en el cuello de la panícula suelen ser las más perjudiciales, ya que pueden provocar, al igual que los ataques en los nudos, una disminución en el peso del grano e incluso, en casos de infecciones tempranas y severas, la aparición de panículas blancas y erectas con granos vacíos. La infección en el cuello y las ramas de la panícula (raquis, ramas primarias y secundarias), así como en los pedicelos que sostienen las espiguillas (granos), puede ocurrir de manera simultánea o no, dependiendo de ciertas condiciones ambientales y genéticas, aunque generalmente las ramas se ven afectadas más tarde, durante la etapa de llenado del grano. La cascarilla del grano también puede ser afectada, cubriéndose de manchas de color marrón oscuro. Tomado de (IFAPA 2004)

### **3.1.5. El muestreo en las poblaciones de patógenos en plantas**

El objetivo del muestreo de enfermedades en plantas aplicado en distintos momentos y lugares es recopilar datos que permitan tomar decisiones de manejo en el campo. Esto se logra mediante el uso de métodos de estadística inferencial y modelos epidemiológicos de pronóstico (Madden & Hughes 1999; Garrett, Madden, Hughes & Pfender 2004; Madden, Hughes & van der Bosch 2007).

A través del muestreo, se utilizan estimadores muestrales que buscan inferir los parámetros de la población (Cochran 1977). En el caso de las enfermedades de las plantas, estos parámetros incluyen medidas de intensidad como presencia-ausencia, incidencia y/o severidad de pronóstico (Madden & Hughes 1999; Garrett et al. 2004; Madden et al. 2007).

El muestreo permite realizar estimaciones poblacionales y determinar parámetros fitosanitarios importantes, como especies presentes, densidad y distribución en el campo. Además, ayuda a comprender el potencial impacto en términos de lesiones, daños y pérdidas, pronosticar dinámicas futuras y diseñar estrategias de manejo (Binns, Nyrop & Van der Warf, 2000).

La inferencia poblacional precisa a partir de los estimadores muestrales está relacionada con el intervalo de confianza y el nivel de precisión (Cochran 1977). Una mayor cantidad y densidad de muestras resulta en una mayor precisión, mientras que una menor cantidad de muestras puede llevar a una subestimación o sobreestimación de la variable de interés (Cochran 1977). Otros factores importantes al realizar y diseñar un muestreo en sistemas agrícolas son el costo operativo, la aplicabilidad en campo, la capacidad de inferencia y el tiempo computacional de análisis. Por lo tanto, una de las estrategias más buscadas es optimizar el muestreo, buscando un equilibrio entre la precisión y el costo del muestreo (Hu & Wang 2011).

### **3.1.6. El muestreo de redes aplicado al monitoreo fitosanitario en cultivos de arroz**

Los campos agrícolas presentan un contexto especial en el que la recolección de muestras se vuelve esencial a lo largo de todo el ciclo de cultivo. En este entorno, las enfermedades tienen un impacto significativo en la productividad, por lo que se requiere tomar medidas preventivas. Como resultado, se llevan a cabo numerosos muestreos en diferentes etapas para detectar posibles enfermedades en una fase temprana.

Sin embargo, en un país como Colombia y para una entidad como Fedearroz que posee una gran cantidad de lotes a nivel nacional, realizar muestreos en todos, se convierte en una tarea compleja y costosa, esto debido a las distintas problemáticas que enfrenta el país como la inseguridad, o el estado de las vías. Por lo que proponer métodos que permitan hacer inferencia a un costo más bajo, se vuelven muy atractivos.

Los muestreos tradicionales basados en listas, presentan dificultades en estos contextos, puesto que la recolección de variables auxiliares del lote puede ser incluso más complejo que el tomar muestras del mismo, y los muestreos que no recogen ninguna variable auxiliar tal como en el muestreo aleatorio simple y sin reemplazo presentan una pérdida de información que puede ser aprovechable.

Aquí es donde la propuesta del trabajo se vuelve un punto fundamental puesto que consiste en implementar un muestreo distinto que no es basado en listas sino que es basado en redes, lo que implica un gran trabajo previo al muestreo puesto que dar la estructura de red a un listado de elementos, es una tarea que se vuelve compleja y de mucho conocimiento técnico agronómico, puesto que la correcta definición de los nodos y enlace es primordial.

En este trabajo uno de los objetivos fundamentales es construir toda la metodología desde 0 en la aplicación de muestreos de redes en el escenario agronómico donde no se han con-

templado estas iniciativas previamente, para el mismo se realizará una construcción de una red agronómica de lotes de arroz suministrados por Fedearroz. Posteriormente realizar la comparación de distintos métodos de muestreo en redes con el muestreo aleatorio simple y sin reemplazo.

Uno de los factores más decisivos en la presencia y gravedad de la enfermedad es su incidencia. Esta se refiere a la cantidad de plantas infectadas en relación al total de plantas muestreadas. Si bien es importante considerar la presencia de la enfermedad, la incidencia tiene aún más relevancia, ya que a mayor valor, se esperan mayores pérdidas, principalmente económicas, ya que indica que la producción final se verá significativamente afectada. Por lo tanto, la variable  $Y$  representada por la incidencia de la piricularia en los lotes se convierte en nuestra variable objetivo de estudio.

A medida que los diferentes actores involucrados en el proceso del cultivo de arroz realicen un muestreo adecuado que les permita prevenir y monitorear el desarrollo de enfermedades de manera rentable, se logrará un ahorro logístico y económico en toda la cadena de producción.

En la actualidad, Fedearroz lleva a cabo una labor importante de recolección de muestras en todos los lotes a manera de censo con el fin de conocer el estado de las principales enfermedades que afectan al cultivo incluida la piricularia.

En este trabajo en particular, se destaca una característica importante; se conoce el marco poblacional en su totalidad. No se trata de una población infinita, sino con un escenario específico donde el marco poblacional es finito y conocido; esto nos permite realizar estimaciones sencillas y teóricas de la varianza y el sesgo de manera numérica, con esto solucionaríamos el gran inconveniente que presentan los métodos de muestreo en redes, que se discutieron en el capítulo 1.

## **3.2. Materiales y métodos**

### **3.2.1. Origen y procesamiento de los datos**

En particular para este trabajo se cuenta con un marco de 230 lotes de arroz que hacen parte del programa de monitoreo de Fedearroz, se ha adelantado gracias al grupo de investigación Biocomputación los cuales han sido facilitados por la entidad bajo un marco normativo de cooperación entre Fedearroz y la Facultad de Ciencias Agrarias de la Universidad Nacional de Colombia sede Bogotá.

Los lotes son manejados de manera comercial por cada uno de sus agricultores, es decir, son

sistemas de producción a los cuales se les realiza aplicaciones de fungicidas para el control de enfermedades de acuerdo al plan de mitigación de enfermedades suministrado por Fedearroz. La base de datos se relaciona con el monitoreo de lotes sensores, los cuales son lotes distribuidos en las cuatro zonas arroceras del país, en los cuales se hace el seguimiento fitosanitario en las etapas de desarrollo del cultivo como: emergencia, inicio de macollamiento, inicio de primordio floral, inicio de embuchamiento, inicio de floración, grano pastoso y maduración. Esta base de datos contiene información sobre la incidencia y severidad de diferentes enfermedades en el cultivo del arroz en cada una de las etapas fenológicas, fechas de siembra y evaluaciones, datos de ubicación de los lotes como municipio, finca, vereda, así como datos de coordenadas y polígonos de los lotes monitoreados. Esta base de datos ha sido limpiada, organizada y enriquecida con variables edáficas y climáticas por el grupo de trabajo.

Para cada uno de los lotes se registra información relevante tal como ubicación geográfica, longitud, latitud, región, variedad de arroz, departamento y la zona según tipo de suelo, etapas fenológicas donde se realizaban los muestreos, todas estas variables serán aprovechadas en el estudio, tanto en la metodología de closterización que se realizará para la creación de estratos como en la creación de la red, con el fin de reducir la variabilidad que afecta la presencia de la enfermedad.

Se utilizaron las etapas 7 y 8, dado que la enfermedad estudiada presenta un alto impacto potencial en la productividad cuando afecta esta etapa fenológica, por lo cual hace parte del target u objetivo de muestreo.

Como se mencionó anteriormente, la piricularia es una enfermedad que presenta diferentes factores que influyen en su aparición, y estos factores pueden variar entre los lotes, como diferentes tipos de suelos, variedad de arroz, zonas climáticas y condiciones.

Para la consolidación de la información en una sola medida, se probaron diferentes alternativas. Una opción fue tomar la máxima presencia de piricularia registrada durante todo el periodo. Sin embargo, se observó que factores externos podían afectar esta medición y no ser completamente confiable, debido a características particulares como errores humanos en las mediciones o la imposibilidad de realizar las aplicaciones necesarias en el cultivo para controlar la enfermedad. Por lo tanto, se optó por usar una medida más robusta a valores atípicos como la mediana para resumir la susceptibilidad de cada lote a la enfermedad.

Con esto se obtuvo una base consolidada con los 230 lotes con una única medida que reflejaba el comportamiento histórico de la enfermedad en etapas fenológicas 7 y 8, junto a las covariables mencionadas anteriormente, a partir del mismo esta base se constituyó como la población completa del estudio, y con esto se obtuvo una medición de la piricularia histórica a nivel general, el cual será el parámetro poblacional que buscaremos estimar a través de

distintos métodos.

### 3.2.2. Metodología de muestreo

Según se mencionó en el marco teórico de este documento, los métodos de muestreo que se utilizarán no son los tradicionales. Los muestreos tradicionales se basan en listas de elementos sin mapear las relaciones entre individuos en la muestra. En cambio, los escenarios propuestos en este documento se refieren a escenarios de tipo red, donde cada individuo en el marco muestral tiene al menos una relación con otro individuo. Es importante proporcionar el marco muestral, que en este caso son los nodos, así como los enlaces entre nodos. La correcta determinación de los enlaces puede favorecer significativamente la estimación del parámetro y la metodología de muestreo.

En la construcción de los enlaces se utilizan medidas de asociación naturales más comunes. Por ejemplo, en redes de transmisión de enfermedades sexuales, el contacto sexual entre individuos se convierte en el principal vínculo de asociación. Otro ejemplo son las redes sociales, donde los enlaces son explícitos en términos de seguidores, amigos y otros. En este caso, podemos encontrar individuos que se relacionan con muchas personas, como los creadores de contenido. Estos individuos son puntos focales que, por ejemplo, en campañas publicitarias, pueden llegar a un público muy específico. Estas son las ventajas que brindan las estructuras de tipo red y una correcta definición de los nodos y enlaces.

No obstante, nos enfrentamos a una gran variedad de características presentes en los lotes debido a principalmente dos factores, en primer lugar las diferencias climáticas entre lotes de distintas regiones y en segundo lugar las diversas variedades de arroz.

Para abordar el primer factor, se decidió aplicar un método de aprendizaje no supervisado, en concreto la construcción de clústeres. Para este proceso se utilizó los valores de las variables climáticas a nivel de lote, obtenido mediante sus coordenadas geográficas. La clusterización se realizó usando el algoritmo  $k$ -Means espacial según las indicaciones y el paso validado e implementado por el equipo de trabajo, específicamente por (Rodríguez et al., en progreso, Tesis de maestría en Ciencias-Climatología).

Este método nos permitirá crear distintos conglomerados que reducirán la variabilidad del clima entre grupos y generarán una diferenciación significativa con los pertenecientes a otro grupo; además, estos conglomerados serán incluidos en el muestreo como estratos, que estarán presentes en el marco muestral

El objetivo principal del muestreo estratificado es mejorar la precisión de las estimaciones y reducir la variabilidad, al tener en cuenta las diferencias internas dentro de la población

objetivo. En lugar de seleccionar muestras al azar de toda la población, se eligen muestras más pequeñas pero representativas de cada estrato, lo que permite obtener información más precisa sobre cada grupo en particular.

Es importante que cada unidad de la población pertenezca a un solo estrato, y que en conjunto los estratos abarquen toda la población.

El muestreo estratificado resulta especialmente útil cuando existen diferencias significativas entre los subgrupos de la población objetivo y se desea obtener estimaciones precisas para cada estrato. Al asignar los recursos de manera más eficiente y garantizar una representación adecuada de cada estrato, el muestreo estratificado mejora la calidad de los resultados y permite realizar inferencias más precisas sobre la población en general.

En el caso específico de la construcción de este muestreo estratificado para el cultivo de arroz, se utilizaron distintas variables relacionadas con el clima en los diferentes lotes, como la ubicación geográfica (latitud y longitud), la evapotranspiración potencial, las temperaturas máximas y mínimas durante un período de 5 años, la precipitación, el déficit de presión de vapor, el índice de aridez, el índice térmico, entre otras variables.

A partir de estos datos, se realizó un análisis de componentes principales de las variables mencionadas anteriormente con el objetivo de reducir la cantidad de variables y permitir que los algoritmos funcionen de manera más eficiente. Además, se decidió utilizar el método K-Means y optimizar el índice de silueta para la construcción de los clústeres, generando como resultado 3 clústeres donde se agruparan todos los lotes.

Estos clústeres representarán los 3 estratos en los que se agruparán todos los lotes. Sin embargo, también se optó por crear un clúster adicional donde se incluyan todos los lotes sin presencia histórica de la enfermedad, y donde se podrán agregar nuevos lotes que Fedearroz desee incluir en la población en futuras aplicaciones de esta metodología. En total, se generarán 4 estratos que abarcarán los 230 lotes del marco muestral.

Una vez obtenidos los estratos, se procede a construir la red. Recordemos que teníamos dos problemáticas a abordar que afectan de manera diferencial a la enfermedad en los distintos lotes: la variabilidad del clima y la variedad del arroz cultivado en cada lote.

La variabilidad climática se abordó mediante la construcción de los estratos mencionados anteriormente. Por otro lado, la variedad del arroz se presenta como un enlace natural para la construcción de la red, por lo que cada lote es un nodo y está conectado con otro nodo si y solo si ambos lotes presentan la misma variedad de arroz.

Se construyó un grafo con 230 nodos enlazados por la misma variedad de arroz, lo que resulta en un grafo no conectado debido que podemos encontrar nodos que no estén conectados, por ejemplo dos lotes con distinta variedad de arroz. Esta distribución genera un tipo de subgrafos específicos donde todos los individuos están conectados entre sí, pero no se presenta una conectividad completa. Esta característica es importante, ya que determinará los métodos de muestreo en redes que pueden ser aplicados.

Es de vital importancia recordar los métodos definidos en el capítulo 1 de este documento, donde teníamos ciertas particularidades de los mismos y escenarios donde pueden ser aplicados unos u otros. En este específico problema, se realizó un descarte automático de los métodos trazo de ruta, dirigido por el encuestado y rama aleatoria, debido a que son aplicables solamente en grafos conectados. Para aplicar estos métodos es necesario que entre 2 nodos siempre debe existir un camino; en este caso, no es posible.

Por otra parte, los muestreos de caminata aleatoria no tienen la necesidad de que el grafo esté completamente conectado, por lo que, junto al muestreo bola de nieve, serán los escogidos para este estudio.

Solamente se realizará la excepción de la caminata aleatoria simple, puesto que al tener posiblemente subgrafos conectados con pocos nodos, la mejor alternativa es la caminata aleatoria simple con escape por las eficiencias comentadas en el capítulo 1.

Por lo tanto, los métodos de muestreo en redes que serán testeados en esta aplicación serán caminata aleatoria con escape, caminata aleatoria Metropolis-Hasting, caminata aleatoria con reemplazo, y bola de nieve.

Adicionalmente, junto con estos métodos de muestreo en redes, se empleará el muestreo aleatorio simple y sin reemplazo, que es el método de muestreo más comúnmente utilizado en Fedearroz para otros procesos y además representa los métodos tradicionales basados en listas.

Para realizar la comparación entre los métodos, aprovechando que tenemos el marco completo y un parámetro poblacional establecido, evaluaremos la varianza, el sesgo y el coeficiente de variación de manera numérica para cada método en distintos tamaños de muestra.

Los tamaños de muestra considerados serán: 50, que representa el 20 % de la población; 70, que representa el 30 % de la población; 100, que representa el 43 % de la población; 120, que representa el 52 % de la población; 170, que representa el 70 % de la población; y 200, que representa el 86 % de la población.

Estos tamaños de muestra fueron definidos en colaboración con un experto que afirmó que una reducción de al menos el 10% de la muestra de lotes implicaría un ahorro logístico y económico significativo para Fedearroz. Por lo tanto, en este escenario, el objetivo es encontrar la combinación ideal de método y tamaño de muestra que asegure un coeficiente de variación consistente.

Para llevar a cabo este ejercicio de manera concisa y precisa, se decidió realizar un estudio de simulación con cada una de las combinaciones de tamaños de muestra y métodos de muestreo. Esto implica generar estimaciones del estimador, varianza, sesgo y coeficiente de variación utilizando el método Monte Carlo mencionado anteriormente con 1000 iteraciones. Estas simulaciones permitirán contrastar aún más las mediciones en cada combinación de método y tamaño de muestra.

El método Monte Carlo también validará los supuestos de insesgamiento del estimador HT bajo los métodos de muestreo en redes que se comentaron en el capítulo 1. Dado que estos métodos no tienen una probabilidad de inclusión de segundo orden fácilmente definida, la varianza estimada de este estimador solo se puede calcular de forma numérica.

El software usado en este documento es una combinación de los dos programas más usados en estadística y ciencia de datos, los cuales son Python y R. Todo el proceso de limpieza, preprocesamiento y creación de la estructura de los datos para que sean de tipo red fue mediante R. La creación de los clústeres, la aplicación de los métodos de muestreo fueron realizados en Python.

Para la creación de los algoritmos de muestreo, se realizaron programaciones orientadas a objetos en Python, basadas en los códigos tomados de (Ashish 2020).

Se añaden los códigos usados en este trabajo en el siguiente repositorio de GitHub

[https://github.com/ldvelasquez/Network\\_Sampling](https://github.com/ldvelasquez/Network_Sampling).

### 3.3. Resultados

#### 3.3.1. Creación de los estratos

En esta sección, se abordará cómo se realizó la creación de los estratos, tal como se comentó anteriormente. La creación de los clústeres se hace debido a las características que hacen más propensa la presencia en los cultivos de la enfermedad. Estas características podemos dividirlos en dos: su ubicación geográfica que reúne toda la parte de clima y suelos, y por otra parte la variedad de arroz cultivado.

Para tratar con el primer factor, se realizó un ejercicio de clusterización primero con una aplicación de análisis de componentes principales. Se generaron dos componentes. En la primera componente, se resumía toda la información geográfica, tales como el tipo de suelo y su ubicación. Por otra parte, en la segunda componente se encontraban agrupadas las variables referentes al clima.

Mediante el algoritmo *k*-Means, con el que, a través de las variables descritas en la sección anterior, se obtuvieron 3 clústeres en los que se pudieron clasificar cada uno de los lotes. El índice de silueta para esta clusterización fue de 0,74, por lo que se interpreta como una buena creación de clústeres, puesto que el índice de silueta clasifica de -1 a 1, entre más cercano a 1 el valor, mejor es la representación de la clusterización.

Sin embargo, en el ejercicio, observamos que 121 de los 230 lotes, lo que corresponde al 52% de los lotes, nunca habían presentado la enfermedad. Por lo que es posible que, al aplicar el muestreo estratificado, estos salieran en la muestra de todos los estratos. Por lo tanto, se decidió crear un clúster adicional que agrupaba todos los lotes que nunca habían presentado la enfermedad. Además, este clúster funcionará para cuando Fedearroz desee incluir nuevos lotes a estudiar, puesto que serán asignados a este clúster automáticamente.

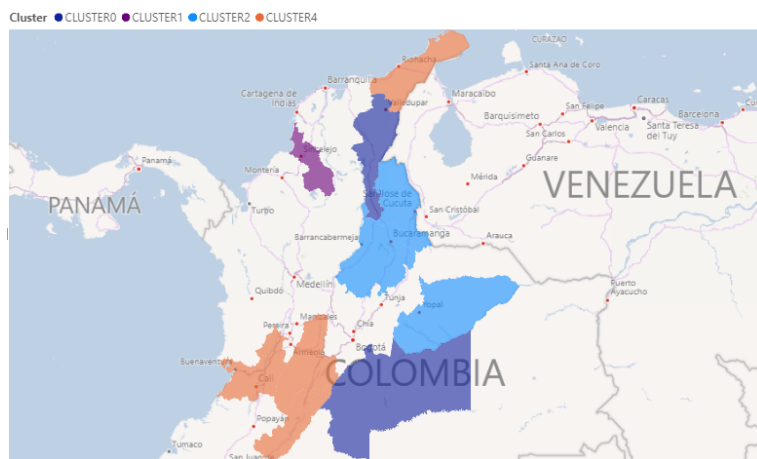
Por lo que al final del ejercicio, se encuentran 4 distintos clústeres. A continuación, se mostrará la distribución de los 230 lotes en los clústeres mencionados.

Clúster	Cantidad
CLUSTER0	24
CLUSTER1	21
CLUSTER2	64
CLUSTER4	121
<b>Total general</b>	<b>230</b>

**Tabla 3-1:** Distribución de clústeres - lotes por clúster

Cada uno de estos clústeres tiene una distribución particular de lotes que poseen características similares. Una de las características principales se ve reflejada en su distribución geográfica. Aquí, la ubicación del lote presenta una importancia principal, puesto que dependiendo de la zona, se tiene un clima característico y esto se ve, por ejemplo, en las diferencias de climas entre La Guajira y Meta. También, el suelo es distinto, por lo que es natural que las distribuciones de los clústeres estén acompañadas de zonas predominantes por departamentos.

Los lotes presentes en Fedearroz se encuentran ubicados en los departamentos de Casanare, Cesar, Guajira, Huila, Meta, Norte de Santander, Santander, Sucre, Tolima y Valle del Cauca. A continuación, se mostrará la distribución de los clústeres por departamento, donde se representa el color predominante del clúster en los departamentos donde Fedearroz tiene lotes de producción arrocerá.



**Figura 3-1:** Distribución de clústeres - lotes por departamento - Autoría Propia

Aquí se ve que el clúster 0 es predominante en el departamento de Meta, el clúster 1 es predominante en el departamento de Sucre, el clúster 2 es predominante en la zona de Norte de Santander y Casanare, y el clúster 4 es predominante en las zonas costeras en los departamentos de Valle del Cauca y Guajira.

Por lo tanto, estos clústeres constituirán los 4 estratos que serán usados en la metodología. El objetivo de estos estratos es la reducción en la variabilidad teórica sobre la incidencia de la Piricularia en los cultivos. Con esto, la única variabilidad teórica que se debe mapear con respecto a lotes que aumente la posibilidad de que la enfermedad esté presente es la variedad de arroz, cuyos resultados serán tratados en la siguiente sección.

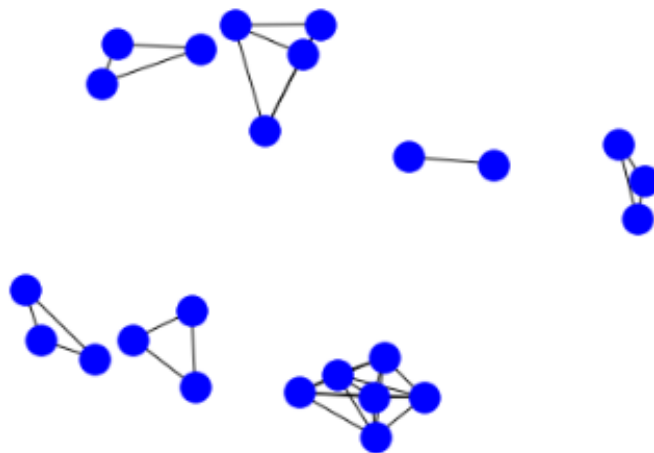
### 3.3.2. Creación de la red

Teniendo en cuenta que el primer factor que influye en la presencia e incidencia de la piricularia tratado en la sección anterior ya fue parcialmente controlado, otro causante de la piricularia, es la variedad del arroz, y esto es debido a que la enfermedad dependiendo las características del cultivo en particular puede ser más propensa a darse.

La siembra de distintas variedades de arroz se realiza para adaptarse a las condiciones locales, tanto en las características propias del cultivo tales como el riego, la ubicación geográfica, como las distintas preferencias del mercado.

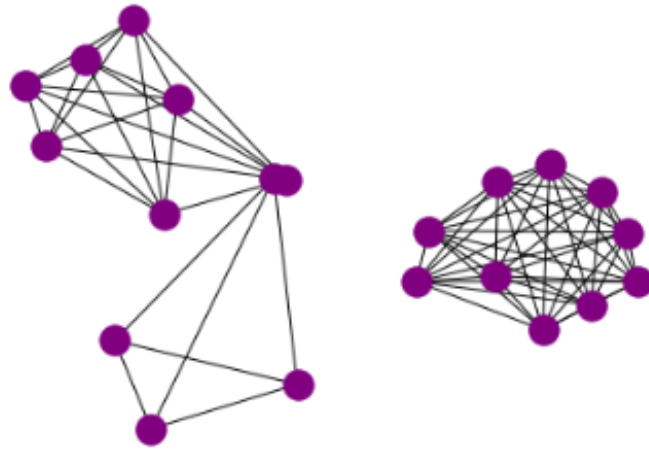
Como mencionamos anteriormente, utilizaremos esta variable como el vínculo principal para los lotes. En consecuencia, la red se definirá exclusivamente como un grafo de lotes, donde cada lote será un nodo, y los lotes estarán conectados si comparten la misma variedad de arroz.

Este procedimiento sería aplicado para los cuatro estratos de la misma manera, en las figuras 3-2, 3-3, 3-4, 3-5 y 3-6 se muestra la representación gráfica de los mismos.



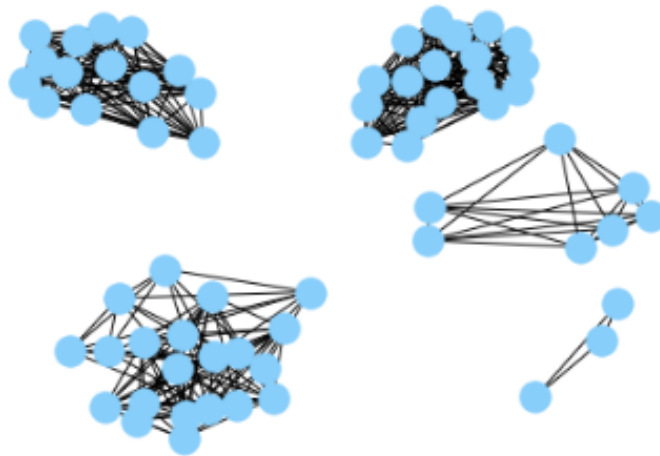
**Figura 3-2:** Representación gráfica Cluster0 - Autoría propia

Se observa que en la red generada por el clúster 0, se encuentran presentes 7 variedades de arroz en estos lotes que han presentado piricularia en alguna ocasión, se observa que la red no es densa pero así mismo como se comentó anteriormente no está conectada entre sí.



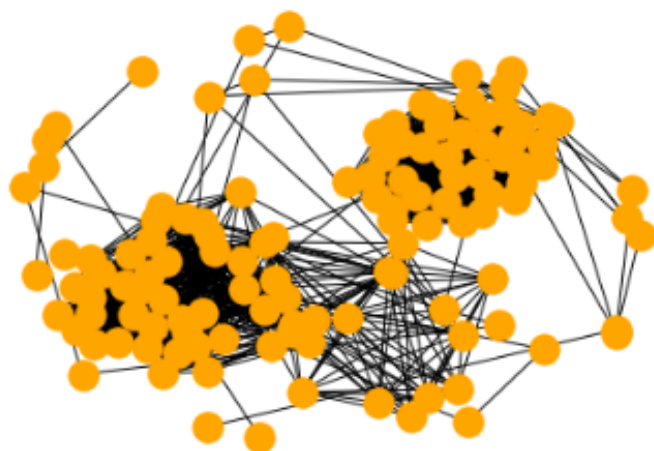
**Figura 3-3:** Representación gráfica Cluster1 - Autoría propia

Mientras que en el clúster 1 se aprecian 3 variedades de arroz en lotes que presentaron la enfermedad, pero se ve que cada variedad tiene una presencia mayor de lotes respecto al clúster 0.



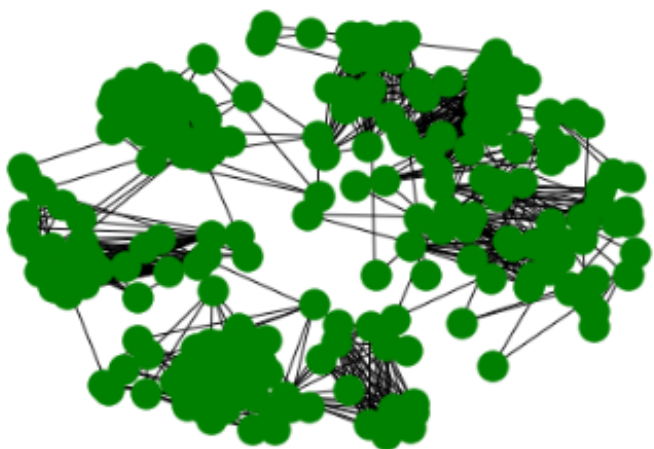
**Figura 3-4:** Representación gráfica Cluster2 - Autoría propia

En el clúster 2 se observan 5 variedades de arroz, pero al tener muchos más lotes, los pequeños subgrafos generados por cada variedad son más densos, lo que implica es que los muestreos en particular se enfocaran en estas variedades más grandes.



**Figura 3-5:** Representación gráfica Cluster4 - Autoría propia

En este clúster se aprecian una gran tipos de variedades de arroz presentes y conexiones que no son densas puesto que es necesario recordar la creación de este clúster, como el resultado de los lotes de los tres clústeres anteriores que no presentaran históricamente piricularia.



**Figura 3-6:** Representación gráfica red completa - Autoría propia

### 3.3.3. Estimaciones

De acuerdo a lo comentado en la sección correspondiente a metodología de muestreo, se procedió a realizar la aplicación de los distintos algoritmos que son adecuados para el tipo de grafos que se tienen presentes en esta aplicación, los cuales son los métodos basados en caminatas aleatorias y el muestreo bola de nieve.

Teniendo en cuenta la sección anterior todos los métodos serán aplicados mediante un muestreo estratificado, por lo que el tamaño de muestra total será distribuido proporcionalmente al tamaño de cada uno de los estratos, con esto se obtendrá el tamaño de muestra por cada estrato donde será aplicado el método.

Luego se procederá a replicar la selección de la muestra y cálculo tanto de la estimación, varianza, sesgo y coeficiente de variación 1000 veces, con lo cual estaríamos aplicando la estimación Monte Carlo explicada en el capítulo 1.

Para realizar el cálculo de la varianza y sesgo, es necesario ir a la definición numérica de los mismos, donde es necesario tomar el parámetro poblacional calculado a partir de los 230 lotes, el cual indica que hay una incidencia de piricularia de **1,95 %**, este se convertirá en el parámetro que deseamos aproximar mediante los distintos algoritmos de muestreo y tamaños de muestra.

Método-estimación	50	70	100	120	150	200
Caminata aleatoria simple con escape	0,49 %	0,46 %	0,97 %	1,30 %	1,21 %	1,82 %
Caminata aleatoria simple con reemplazo	0,43 %	0,44 %	0,91 %	1,18 %	1,48 %	1,63 %
Bola de nieve	0,57 %	0,77 %	0,76 %	1,05 %	1,22 %	1,56 %
Caminata aleatoria Metropolis-Hasting	0,46 %	0,62 %	0,77 %	0,85 %	1,33 %	1,79 %
Aleatorio simple y sin reemplazo	0,30 %	0,41 %	0,63 %	0,68 %	1,19 %	1,59 %

**Tabla 3-2:** Estimaciones asociadas a cada método

A partir de las estimaciones presentadas en la tabla **3-2** se observa una tendencia general esperada de acercarse al valor objetivo respecto a un mayor tamaño de muestra, de igual manera otra tendencia relevante es que para todos los tamaños de muestra asociados los métodos basados en redes se acercan más al parámetro objetivo.

Método-sesgo	50	70	100	120	150	200
Caminata aleatoria simple con escape	0,00022255	0,00023165	0,00010349	0,00004740	0,00006001	0,00000264
Caminata aleatoria simple con reemplazo	0,00024204	0,00023824	0,00011569	0,00006510	0,00002550	0,00001240
Bola de nieve	0,00020140	0,00014641	0,00014903	0,00008660	0,00005913	0,00001013
Caminata aleatoria Metropolis-Hasting	0,00023254	0,00018539	0,00014652	0,00012938	0,00004277	0,00000370
Aleatorio simple y sin reemplazo	0,00028286	0,00024856	0,00018449	0,00016989	0,00006348	0,00001521

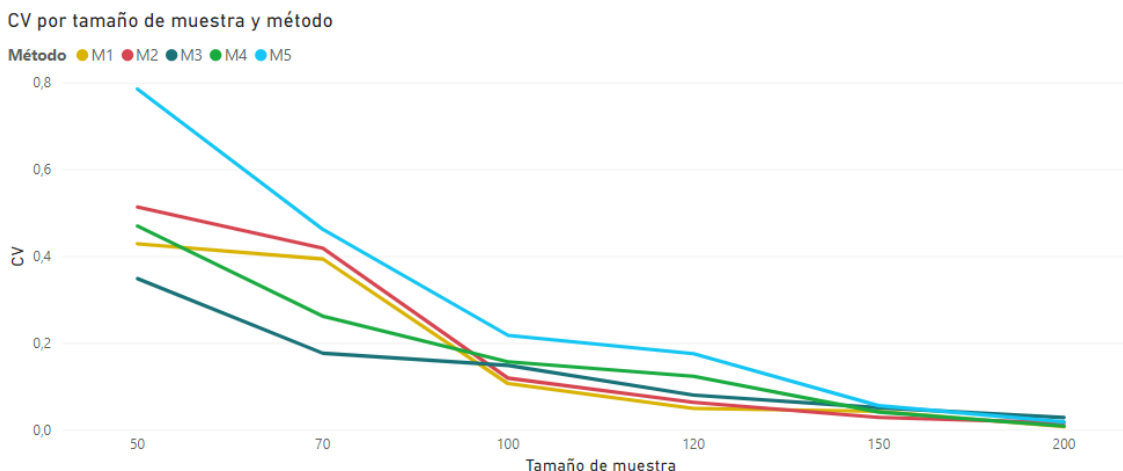
**Tabla 3-3:** Sesgo asociado a cada método

Uno de los puntos a contrastar en el trabajo es la efectividad de los métodos con respecto al sesgamiento que se puede observar en la tabla **3-3** que podía llegar a presentar el estimador HT bajo los muestreos en redes, el mismo se calculo con el sesgo de cada una de las iteraciones para cada combinación de método-tamaño.

Método-CV	50	70	100	120	150	200
Caminata aleatoria simple con escape	42,79 %	39,30 %	10,56 %	4,86 %	4,15 %	0,63 %
Caminata aleatoria simple con reemplazo	51,27 %	41,80 %	11,82 %	6,25 %	2,78 %	1,52 %
Bola de nieve	34,78 %	17,56 %	14,77 %	7,92 %	5,01 %	2,75 %
Caminata aleatoria Metropolis-Hasting	46,88 %	26,11 %	15,63 %	12,25 %	4,01 %	0,75 %
Aleatorio simple y sin reemplazo	78,48 %	46,15 %	21,67 %	17,46 %	5,47 %	1,72 %

**Tabla 3-4:** Coeficiente de variación asociado a cada método

Por ultimo la medida que nos permite realizar comparaciones directas entre muestreos es el coeficiente de variación presente en la tabla **3-4** aqui se observan dos tendencias en particular y es que la estimación mejora a medida que se aumenta el tamaño de muestra algo que es intuitivo en los muestreos, pero adicionalmente lo que nos permite contrastar nuestra hipótesis y es que el coeficiente de variación de los métodos basados en redes superan ampliamente a la estimación obtenida mediante el muestreo aleatorio simple y sin reemplazo.



**Figura 3-7:** Representación gráfica CV por tamaño de muestra y método- Autoría propia

Esta gráfica describe los cambios a través de cada uno de los tamaños de muestra y métodos, donde se visualiza la disminución del cv a medida de aumentar el tamaño de muestra en general, también se observa como el muestreo aleatorio simple y sin reemplazo tiene el cv más alto en cada uno de los tamaños de muestra, los métodos presentados en la gráfica se ven en el mismo orden como se refieren en las tablas anteriormente mostradas.

### 3.4. Discusión

Uno de los métodos más usados para comparar distintos muestreos es el coeficiente de variación del estimador, puesto que nos permite evaluar la dispersión o variabilidad de los datos en relación con su tamaño promedio. Al comparar el coeficiente de variación del estimador en diferentes muestras, podemos determinar cuál muestra tiene una mayor o menor variabilidad relativa en relación con su promedio.

Este método es el utilizado en esta aplicación para realizar la comparación y probar la efectividad de los muestreos. Esta medida indica que entre más cercano esté el coeficiente de variación del estimador a 0, es mejor, ya que indica una mayor homogeneidad en los resultados de la muestra.

Un menor coeficiente de variación del estimador indica que los datos tienden a agruparse más cerca del valor promedio, lo que es especialmente valioso cuando se busca minimizar la incertidumbre y obtener estimaciones más confiables. Por otro lado, un coeficiente de variación del estimador más alto sugiere que los datos tienen una mayor dispersión en relación con su media, lo que indica una mayor variabilidad o heterogeneidad en los valores de la muestra. En estos casos, la precisión de las estimaciones puede ser menor, ya que los datos están más dispersos y pueden variar significativamente entre observaciones.

Teniendo esto en mente y analizando los resultados de la tabla **3-4**, se observa que los métodos basados en redes presentan un coeficiente de variación menor que el muestreo aleatorio simple y sin reemplazo.

Sin embargo, se observa que los coeficientes de variación para las muestras de tamaño 50 y 70 son muy altos para lo acostumbrado en los estudios de aplicación, puesto que una condición empírica que se usa para determinar cuando el muestreo es apropiado es que el coeficiente de variación sea máximo del 15 %.

Respecto a los métodos basados en redes, se observa que los métodos de caminatas aleatorias simples con escape funciona de mejor manera con los tamaños de muestra 100 y 120, con reemplazamiento funcionan de mejor manera en las muestras 150 y 200, por último para muestras pequeñas de tamaños 50 y 70 el método que mejor funciona es el muestreo bola de

nieve.

Uno de los motivos que puede causar que el muestreo aleatorio simple y sin reemplazo no funcione de la mejor manera es que los lotes seleccionados en los estratos 0, 1 y 2 donde se presenta la enfermedad, sean seleccionados de manera completamente aleatoria, incluyendo lotes donde la presencia de la enfermedad no es nula pero es muy baja al no tener en cuenta la variedad del arroz. Esto, sumado a los lotes donde la enfermedad no se ha presentado, hace que las estimaciones estén muy por debajo de la estimación del parámetro poblacional que se definió.

También se observa que en estratos 0, 1 y 2 se encuentran estructuras de sub-grafos donde está concentrada la enfermedad. Esto nos haría indicar que estos lotes son de especial cuidado, puesto que indicarían una correlación entre la variedad y sus características climáticas y de suelo que hacen más propensa la enfermedad.

Por lo tanto, la estructura de tipo red no solamente nos serviría para obtener una estimación, sino que el trabajo adelantado a partir de la creación de los estratos y de la red sería de vital ayuda para detectar variedades que presenten problemas en ciertos ambientes.

También se puede observar en estos resultados que los métodos de redes propuestos se comportan en términos de sesgo de la misma forma como el muestreo aleatorio simple y sin reemplazo que se conoce desde el enfoque teórico es insesgado. Por lo que la eficiencia de los muestreos de redes queda corroborada, al menos para estimaciones de primer orden de los grafos.

Se puede observar que los muestreos de 100, 120, 150 y 200 lotes pueden presentar una mejora económica y logística significativa para Fedearroz y no compromete en gran medida la estimación de la enfermedad presente en el territorio nacional.

Por lo tanto, en escenarios donde se conoce todo el marco y se busca reducir de un censo a una muestra, comprometiendo poco la eficacia del muestreo, los métodos de muestreo en redes se convierten en una alternativa muy interesante a explorar.

Estas características son especialmente comunes en escenarios agronómicos, por lo que un estudio similar podría adelantarse en otro tipo de cultivos, teniendo siempre en cuenta el estudio teórico que se debe hacer del cultivo para definir correctamente los nodos y enlaces para la creación de la red.

## 4 Conclusión y Recomendaciones

En conclusión, el presente trabajo de maestría abordó de manera exitosa la investigación sobre muestreos en redes y su aplicación al campo agronómico, con énfasis en el estudio de cultivos de arroz. A lo largo de este estudio, se ha logrado demostrar que el enfoque de muestreo basado en redes bajo escenarios particulares ofrecen ventajas significativas sobre los métodos tradicionales de muestreo basados en listas.

La metodología ha demostrado ser más eficiente y precisa para estimar características clave de los cultivos de arroz, como la incidencia de la piricularia. Esta mayor eficiencia se debe en gran parte a la capacidad del muestreo en redes para capturar las interconexiones y relaciones entre los lotes de arroz, permitiendo una inferencia más acertada y representativa de toda la población a un costo menor que hacer un censo.

Asimismo, se destaca la utilidad de distintas herramientas estadísticas tales como el aprendizaje no supervisado y la integración del muestreo estratificado al muestreo en redes, así como el análisis de los diversos métodos que permiten su implementación, para mejorar y enriquecer la investigación en este campo. Estas herramientas han permitido un mayor entendimiento de las estructuras complejas presentes en los cultivos de arroz y han facilitado la obtención de estimadores precisos y confiables.

Los resultados de este trabajo son de gran relevancia para el campo agronómico, en particular para entidades como Fedearroz, ya que brindan una herramienta eficaz y sólida para la toma de decisiones informadas. La detección temprana de enfermedades y la implementación de estrategias de prevención basadas en los resultados de los muestreos en redes pueden contribuir significativamente a mejorar la productividad y la calidad de los cultivos de arroz.

En el ámbito estadístico, este trabajo ha aportado valiosos conocimientos y ha ampliado la comprensión de cómo abordar muestreos desde un punto de vista distinto a los muestreos basados en listas. Se espera que estos resultados impulsen futuras investigaciones en esta área y motiven a otras personas a explorar y aplicar métodos de muestreo en redes en diversos campos.

En la aplicación de este trabajo se tomó una definición de red dada de manera particular por los lotes y su interconexión dada por la variedad sin embargo, se podrían cambiar las formas

de construir la red por ejemplo para que la misma este completamente conectada y aprovechar todos los métodos descritos en este trabajo, también como se tienen datos históricos sobre el clima en estas regiones se puede pensar en redes dinámicas que vayan cambiando sus enlaces en tiempo real, u otras dinámicas como redes ponderadas y redes direccionadas.

Así mismo otro gran reto en esta línea de investigación es la construcción de expresiones teóricas para la varianza del estimador HT bajo muestreos basados en caminatas aleatorias, puesto que si se logra esta construcción no serían necesarios métodos iterativos para demostrar su efectividad y podrían ser aplicados a otros escenarios donde no se conozca toda la población.

En escenarios futuros se puede pensar en trabajar una metodología más a detalle puesto que los muestreos agronómicos incluyen dentro de cada lote un muestreo de plantas, por lo que se puede pensar en desarrollar muestreos en redes con dos o más etapas que permitan el mapeo completo desde el muestreo del lote a el muestreo de plantas, por otra parte este estudio fue basado en la piricularia puesto que es la enfermedad que más impacta los cultivos de arroz, se podría pensar en metodologías que permitan recoger múltiples enfermedades, y la estimación sea más completa e informativa.

# Bibliografía

- Agrama, H. A., Yan, W., Jia, M., Fjellstrom, R., McClung, A. M. et al. (2010), ‘Genetic structure associated with diversity and geographic distribution in the usda rice world collection’, *Natural Science* **2**(04), 247.
- Ahn, Y.-Y., Han, S., Kwak, H., Moon, S. & Jeong, H. (2007), Analysis of topological characteristics of huge online social networking services, *in* ‘Proceedings of the 16th international conference on World Wide Web’, pp. 835–844.
- Ashish (2020), ‘Graph sampling’.  
**URL:** [https://github.com/Ashish7129/Graph\\_Sampling](https://github.com/Ashish7129/Graph_Sampling)
- Baños, R.A. A.A., . (2020), ‘Induced random walk sampling: a new methodology for social network analysis’, *Quality Quantity*, *54*(5), pp.1371-1387. DOI .
- Biggs, N., Lloyd, E. K. & Wilson, R. J. (1986), *Graph Theory, 1736-1936*, Oxford University Press.
- Binns, M. (2000), ‘Sampling and monitoring in crop protection: The theoretical basis for developing practical decision guides. by mr binns, jp nyrop and w. van der werf. wallingford, uk: Cabi publishing (2000), pp. 284,£ 49.95. isbn 0-85199-347-8.’, *Experimental Agriculture* **37**(1), 125–134.
- Birnbaum, Z. W. & Sirken, M. G. (1965), *Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates*, number 1000, Vital Health Statistics, 2(11), pp. 1-14. National Center for Health Statistics.
- Bloemena, A. (1964), ‘Sampling from a graph’, *MC Tracts* .
- Brewer, K. (2002), ‘Combined survey sampling inference: Weighing basu’s elephants’, *Arnold Publishers* .
- Carrington, P. J., Scott, J. & Wasserman, S. (2005), *Models and Methods in Social Network Analysis*, Vol. 28, Cambridge university press.
- Cassel, C. M., Särndal, C. E. & Wretman, J. H. (1976), ‘Some results on generalized difference estimation and generalized regression estimation for finite populations’, *Biometrika* **63**(3), 615–620.

- Charitou, T., Bryan, K. & Lynn, D. J. (2016), 'Using biological networks to integrate, visualize and analyze genomics data', *Genetics Selection Evolution* **48**(1), 1–12.
- Cochran, W. G. (1954), 'The combination of estimates from different experiments', *Biometrics* **10**(1), 101–129.
- Cochran, W. G. (1977), *Sampling Techniques*, John Wiley & Sons New, York, USA.
- DANE (2014), 3er censo nacional agropecuario: Hay campo para todos, Technical report, Departamento Administrativo Nacional de Estadística. Bogotá, Colombia.
- Dangeti, P. (2017), *Statistics for Machine Learning*, Packt Publishing Ltd.
- Duan, Y. & Lu, F. (2014), 'Robustness of city road networks at different granularities', *Physica A: Statistical Mechanics and its Applications* **411**, 21–34.
- Duda, R., Hart, P., Stork, D. & Ionescu, A. (2000), 'Pattern classification, chapter nonparametric techniques'.
- Durand-Morat, A. & Bairagi, S. (2021), 'International rice outlook: International rice baseline projections 2020-2030'.
- Farris, J. S. (1969), 'On the cophenetic correlation coefficient', *Systematic Zoology* **18**(3), 279–285.
- Fedearroz (2021), 'Cultivo de arroz en colombia 1998-2016: Cambios espaciales', *División de Investigaciones Económicas* .
- Frank, O. (1971), 'Statistical inference in graphs', *Försvarets forskningsanstalt* .
- Frank, O. (1977a), 'Estimation of graph totals', *Scandinavian Journal of Statistics* pp. 81–89.
- Frank, O. (1977b), 'A note on bernoulli sampling in graphs and horvitz-thompson estimation', *Scandinavian Journal of Statistics* pp. 178–180.
- Frank, O. (1977c), 'Survey sampling in graphs', *Journal of Statistical Planning and Inference* **1**(3), 235–264.
- Frank, O. (1978), 'Estimation of the number of connected components in a graph by using a sampled subgraph', *Scandinavian Journal of Statistics* pp. 177–188.
- Frank, O. (1979), 'Sampling and estimation in large social networks', *Social networks* **1**(1), 91–101.
- Frank, O. (1980), 'Estimation of the number of vertices of different degrees in a graph', *Journal of Statistical Planning and Inference* **4**(1), 45–50.

- Frank, O. (1980b), 'Sampling and inference in a population graph', *International Statistical Review/Revue Internationale de Statistique* pp. 33–41.
- Frank, O. (1981), 'A survey of statistical methods for graph analysis', *Sociological methodology* **12**, 110–155.
- Frank, O. (2011), 'Survey sampling in networks', *The Sage handbook of social network analysis* pp. 389–403.
- Frank, O. & Snijders, T. (1994), 'Estimating the size of hidden populations using snowball sampling', *Journal of Official Statistics-Stockholm-* **10**, 53–53.
- Garrett, K., Madden, L., Hughes, G. & Pfender, W. (2004), 'New applications of statistical tools in plant pathology', *Phytopathology* **94**(9), 999–1003.
- Gilbert, E. N. (1959), 'Random graphs', *The Annals of Mathematical Statistics* **30**(4), 1141–1144.
- Gile, K. J., Beaudry, I. S., Handcock, M. S. & Ott, M. Q. (2018), 'Methods for inference from respondent-driven sampling data', *Annual Review of Statistics and Its Application* **5**, 65–93.
- Gregoire, T. G. & Valentine, H. T. (2007), *Sampling Strategies for Natural Resources and the Environment*, CRC Press.
- Gupta, L., Jain, R. & Vaszkun, G. (2015), 'Survey of important issues in uav communication networks', *IEEE communications surveys & tutorials* **18**(2), 1123–1152.
- Harrison, R. L. (2010), Introduction to monte carlo simulation in aip conference proceedings, Vol. 1204, American Institute of Physics, pp. 17–21.
- Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Vol. 2, Springer.
- Horvitz, D. G. & Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American statistical Association* **47**(260), 663–685.
- Hu, M.-G. & Wang, J.-F. (2011), 'A spatial sampling optimization package using msn theory', *Environmental Modelling & Software* **26**(4), 546–548.
- IFAPA (2004), 'Comportamiento de *pyricularia oryzae* en las marimas del Guadalquivir. eficacia fungicida frente al patógeno', *Junta de Andalucía. Consejería de Agricultura y Pesca* .
- Jessen, R. J. (1955), 'Determining the fruit count on a tree by randomized branch sampling', *Biometrics* **11**(1), 99–109.

- Kaufman, L. & Rousseeuw, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons.
- Langville, A. N. & Meyer, C. D. (2006), *Google's PageRank and beyond: The science of Search Engine Rankings*, Princeton university press.
- Lavallée, P. (2007), 'Gwsm and calibration', *Indirect Sampling* pp. 121–150.
- Leskovec, J., Kleinberg, J. & Faloutsos, C. (2007), 'Graph evolution: Densification and shrinking diameters', *ACM transactions on Knowledge Discovery from Data* **1**(1), 2–es.
- Linde, Y., Buzo, A. & Gray, R. (1980), 'An algorithm for vector quantizer design', *IEEE Transactions on Communications* **28**(1), 84–95.
- L'heureux, A., Grolinger, K., Elyamany, H. F. & Capretz, M. A. (2017), 'Machine learning with big data: Challenges and approaches', *IEEE Access* **5**, 7776–7797.
- Madden, L. & Hughes, G. (1999), 'Sampling for plant disease incidence', *Phytopathology* **89**(11), 1088–1103.  
**URL:** [arxiv.org/pdf/physics/0603229.pdf](https://arxiv.org/pdf/physics/0603229.pdf)
- Madden, L. V., Hughes, G. & Van Den Bosch, F. (2007), *The Study of Plant Disease Epidemics*.
- McLaren, C. D. & Bruner, M. W. (2022), 'Citation network analysis', *International Review of Sport and Exercise Psychology* **15**(1), 179–198.
- Michalski, R. S., Carbonell, J. G. & Mitchell, T. M. (2013), *Machine Learning: An Artificial Intelligence Approach*, Springer Science & Business Media.
- Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R. & Muharemagic, E. (2015), 'Deep learning applications and challenges in big data analytics', *Journal of big data* **2**(1), 1–21.
- Newman, M. E. (2001), 'The structure of scientific collaboration networks', *Proceedings of the national academy of sciences* **98**(2), 404–409.
- Pearson, K. (1905), 'The problem of the random walk', *Nature* **72**(1865), pp. 294.
- Porta, M. (2014), *A Dictionary of Epidemiology*, Oxford university press.
- Portenoy, J., Hullman, J. & West, J. D. (2017), 'Leveraging citation networks to visualize scholarly influence over time', *Frontiers in Research Metrics and Analytics* **2**, 8.
- Qi, X. (2022), 'A review: Random walk in graph sampling'.  
**URL:** [arxiv.org/abs/2209.13103](https://arxiv.org/abs/2209.13103)

- Rojas, H. (2009), *Estrategias de muestreo. Diseño de Encuestas y Estimación de Parámetros*, Ediciones de la U.  
**URL:** <https://books.google.com.co/books?id=yiV8esNE9v4C>
- Rousseeuw, P. J. (1987), ‘Silhouettes: A graphical aid to the interpretation and validation of cluster analysis’, *Journal of Computational and Applied Mathematics* **20**, 53–65.
- Salganik, M. J. & Heckathorn, D. D. (2004), ‘Sampling and estimation in hidden populations using respondent-driven sampling’, *Sociological methodology* **34**(1), 193–240.
- Särndal, C.-E., Swensson, B. & Wretman, J. (2003), *Model Assisted Survey Sampling (2nd edition)*, Springer Science & Business Media.
- Särndal, C., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer series in statistics, Springer-Verlag.  
**URL:** <https://books.google.com.co/books?id=MWCzngEACAAJ>
- Shimbel, A. (1953), ‘Structural parameters of communication networks’, *The bulletin of mathematical biophysics* **15**, 501–507.
- Thompson, S. K. (2006), ‘Adaptive web sampling’, *Biometrics* **62**(4), 1224–1234.
- Trujillo, L., Niño, J. & G, H. (2016), ‘Latinamerican congress of probability and mathematical statistics’, *CLAPEM, San José, Costa Rica* .
- Van den Bos, W., Crone, E. A., Meuwese, R. & Güroğlu, B. (2018), ‘Social network cohesion in school classes promotes prosocial behavior’, *PLoS One* **13**(4), e0194656.
- Wiegand, H. & Kish, L. (1965), ‘Survey sampling’.
- Xie, F. & Levinson, D. (2007), ‘Measuring the structure of road networks’, *Geographical analysis* **39**(3), 336–356.
- Zhang, L.-C. (2021), *Graph Sampling*, CRC Press.
- Zhang, L.-C. & Patone, M. (2017), ‘Graph sampling’, *Metron* **75**, 277–299.
- Zhang, P. & Itan, Y. (2019), ‘Biological network approaches and applications in rare disease studies’, *Genes* **10**(10), 797.