
# Estudio sobre tendencias en líneas de investigación en los trabajos de grado del Programa de Estadística de la Universidad del Valle

**Davinson Alexander Arteaga Bermúdez**
**Gian Alepsi Mendoza Oviedo**

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2023

# Estudio sobre tendencias en líneas de investigación en los trabajos de grado del Programa de Estadística de la Universidad del Valle

**Davinson Alexander Arteaga Bermúdez**
**Gian Alepsi Mendoza Oviedo**

Trabajo de grado presentado como requisito parcial para optar al título de:
**Estadístico**

Director:
Wilmar Sepúlveda
Codirectora:
MSc. Jennyfer Portilla
Asesor:
PhD. Jaime Mosquera

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2023

# Agradecimientos

Agradecemos a nuestras familias por ser la mayor fuente de apoyo en todo momento y sobretodo por ser la motivación para alcanzar nuestras metas. Agradecemos a nuestros amigos y compañeros de la Universidad que nos han dado compañía y apoyo en nuestras vidas académicas.

Damos gracias a nuestros directores Wilmar Sepúlveda y Jennyfer Portilla por guiarnos en la elaboración de este trabajo, como también al profesor Jaime Mosquera por su gran ayuda a lo largo de este proceso.

Expresamos nuestros agradecimientos a todos los profesores y a la Universidad del Valle como institución académica por habernos instruido y guiado para convertirnos en profesionales y personas capaces de afrontar diversas situaciones.

# Resumen

En este trabajo de grado se realizó un estudio para identificar las tendencias temáticas que caracterizan los trabajos de grado del pregrado de Estadística de la Universidad del Valle presentados entre el año 2002 y el 2022. Se emplearon técnicas de Procesamiento de Lenguaje Natural (NLP), aprendizaje de maquina (Machine Learning) y teoría estadística, donde todo se resume al uso del Modelado de tópicos mediante el Modelo de Asignación Latente de Dirichlet (LDA). Se encontró que los trabajos de grado se enfocan principalmente en temas de Salud, Análisis Multivariado, Control de la calidad y varios subtemas, Regresión y Datos Funcionales, Diseño de Experimentos y Series de Tiempo. También se obtuvo que gran parte de los trabajos de grado tienden a ser dirigidos por pocos directores.

**Palabras clave: Modelado de tópicos, LDA, NLP, Minería de texto**

# Abstract

In this undergratuated thesis, a study was carried out to identify the thematic tendencies that characterize the undergratuated thesis in Statistics at the Universidad del Valle presented between 2002 and 2022. Natural Language Processing (NLP) techniques, Machine Learning and statistical theory were used, where everything boils down to the use of Topic Modeling through the Latent Dirichlet Allocation (LDA) model. It was found that the undergratuated thesis focus mainly on topics of Health, Multivariate Analysis, Quality Control and various subtopics, Regression and Functional Data, Experiments Design and Time Series. It was also obtained that a large part of the undergratuated thesis are usually directed by few directors.

**Keywords: Topic Modeling, LDA, NLP, Text Mining**

# Contenido

# Lista de Figuras

# Lista de Tablas

# Declaración

Nos permitimos afirmar que hemos realizado el presente Trabajo de Grado de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el propio trabajo. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los hemos reconocido. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de Tesis o Trabajo de Grado.

Igualmente declaramos que los datos utilizados en este trabajo están protegidos por las correspondientes cláusulas de confidencialidad.

Santiago de Cali, Julio 2023

_____

Davinson Alexander Arteaga Bermúdez

_____

Gian Alepsi Mendoza Oviedo

# 1 Introducción

Las técnicas estadísticas han tenido una variedad de aplicaciones a lo largo de su existencia, teniendo dificultades en ciertas situaciones debido a la naturaleza de los datos estudiados, por ejemplo, datos no estructurados como textos e imágenes. Por lo tanto, con el paso del tiempo y mejoras en la tecnología se han creado distintas metodologías especializadas en el manejo de estas situaciones, una de estas es la Minería de Datos y sus distintas aplicaciones (Coenen 2011), las cuales pueden dar solución a escenarios donde se deba manejar grandes volúmenes de datos.

La Minería de Textos es una rama específica de la Minería de Datos que busca extraer información del contenido de los textos, los cuales son clasificados como datos no estructurados (Hotho et al. 2005). Esta clasificación se da por el hecho de que un texto no es tan sencillo de analizar como una hoja de calculo donde los datos están organizados por filas y columnas, los cuales reciben el nombre de datos estructurados.

Las tendencias temáticas en un grupo de textos puede ser información de interés para tener una idea de su contenido general. Aplicaciones de la minería de textos con este fin se han visto en distintos campos de estudio, desde investigaciones en negocios como en el estudio de (Vukanti & Jose 2021), hasta el campo de la salud, con el estudio de (Cho et al. 2020).

Un caso particular de estudio en el que se pueden aplicar estas técnicas de minería de texto, es en la identificación de aquellos temas que desarrollan los estudiantes en su trabajo de grado. Por lo tanto, surgen las preguntas: ¿Cuáles son los temas desarrollados en los trabajos de grado del Pregrado en Estadística en la Universidad del Valle?, ¿Estos temas corresponden a los enseñados durante la formación académica o algunos están ausentes?. Esto permitiría saber si los temas enseñados durante la formación académica de los estudiantes están siendo aplicados en su totalidad o si existe alguna preferencia por los temas.

Para responder a las preguntas anteriores se empleará el ajuste del modelo de Asignación Latente de Dirichlet, que en este caso, hace uso de los trabajos de grado del Pregrado en Estadística de la Universidad del Valle que fueron publicados entre el año 2002 a 2022.

# 2 Planteamiento del problema

La antes conocida como Escuela de Ingeniería Industrial y Estadística con su programa de Pregrado en Ingeniería Industrial y Estadística, desde el año 2012 es conocida como la Escuela de Estadística junto a su Pregrado en Estadística y han estado operando por varios años, hasta el día de hoy. Durante este tiempo los estudiantes vinculados al Pregrado han realizado una gran variedad de trabajos de grado con el fin de completar su carrera universitaria y cumplir con los requisitos necesarios para graduarse como Estadísticos.

Estos trabajos de grado se han enfocado en diferentes técnicas, métodos y análisis estadísticos en una gran variedad de tipos de datos, como también de distintos contextos intelectuales o situaciones problema. Estos trabajos de grado han sido dirigidos por los profesores de planta que ha tenido el programa y también por profesores contratistas, los cuales a lo largo de su carrera profesional se han enfocado en ciertos campos de la Estadística para especializarse, lo cual puede implicar que al momento de dirigir estos trabajos de grado influencien en las técnicas utilizadas, priorizando aquellas en las que tienen mayor conocimiento. Esta influencia puede provocar el estancamiento en las técnicas estadísticas utilizadas, llevando a la reiteración de métodos.

De esta manera surge la pregunta de ¿Cuáles son las técnicas estadísticas utilizadas por los estudiantes para el desarrollo de los trabajos de grado?

Como no existe un estudio que haya abordado la identificación de las lineas de investigación en algún área, en este estudio se busca demostrar las posibles tendencias de línea investigación que tiene el programa de pregrado de Estadística y analizar la variedad o la carencia de temas en los trabajos de grado producidos por el programa durante el periodo 2002 a 2022, pues el objetivo es analizar todos los trabajos de grado que posee el programa y los suministrados se contenían en estas fechas.

# 3 Justificación

La estadística es una disciplina fundamental y transversal en la investigación de múltiples áreas, ya que proporcionan herramientas y métodos necesarios para recopilar datos, analizar, interpretar y tomar decisiones. Es importante que los métodos utilizados en estas investigaciones vayan acorde al tiempo y nuevas metodologías que se van desarrollando con el paso del tiempo. Por lo tanto, es crucial identificar y explorar las diferentes lineas de investigación que se han desarrollado a lo largo de un periodo de tiempo, en especial desde el año 2002 a 2022. Se estudia este periodo porque es aquí donde se publicaron todos los trabajos de grado que la dirección del programa de estadística tiene almacenados de forma digital ya sea en formato pdf o word, pues el objetivo es realizar el análisis con todos los trabajos de grado en formato digital debido al procesamiento mediante software.

Al realizar un estudio de líneas de investigación de trabajos de grado en especial en Estadística, se pueden identificar y explorar áreas emergentes que aún no han sido exploradas en profundidad, y también identificar las áreas de mayor fortaleza de la escuela. Esto permitiría que futuros estudiantes puedan enfocar sus trabajos de grado a estas áreas y contribuir al conocimiento actual y adicional a esto, garantizar que sean trabajos relevantes y de gran impacto tanto para el ámbito académico como para la sociedad en general.

# 4 Objetivos

## 4.1. Objetivo general

Identificar las líneas de investigación de los trabajos de grado del programa de Estadística de la Universidad del Valle presentados entre el año 2002 a 2022 mediante el modelo de Asignación Latente de Dirichlet (LDA).

## 4.2. Objetivos específicos

- Recolectar y procesar mediante técnicas de Procesamiento de Lenguaje Natural (NLP) los trabajos de grado del programa de Estadística de la Universidad del Valle.
- Ajustar un modelo basado en la Asignación Latente de Dirichlet (LDA), el cual permita identificar y agrupar los temas de investigación de los trabajos de grado de estadística de la Universidad del Valle.
- Identificar las debilidades y fortalezas de la escuela en los diferentes campos de investigación a nivel de pregrado.

# 5 Antecedentes

La aplicación de técnicas de minería de textos tienen la posibilidad de aplicarse en distintos contextos técnicos y situaciones problema debido a que la mayoría de la información se suele guardar de manera textual en escritos.

Por ejemplo, en el ambiente investigativo de analítica empresarial, tendríamos el caso presentado por (Vukanti & Jose 2021) donde se desarrolla un marco a seguir con este tipo de técnicas en el contexto en el que se mueven los investigadores. En esta investigación se trata el problema que suele presentarse al intentar modelar escritos digitales que tratan sobre diversos temas económicos y cómo mediante la Asignación Latente de Dirichlet es posible encontrar tópicos en los 2.400 artículos con los que cuentan los investigadores. Mediante LDA, los investigadores lograron crear quince tópicos representativos del corpus utilizado y mediante estos tópicos se buscaba clasificar estos textos según el tópico asignado.

Se puede destacar otra situación ejemplo de la aplicación de la metodología NLP y del modelo LDA, la cuales son abordadas en el estudio realizado por (Cho et al. 2020). En este estudio, se analizan y comparan las tipologías presentes en los textos académicos de salud citados en el pasado y presente en artículos recientes. Mediante el modelo LDA los investigadores buscaban tendencias en los escritos académicos citados, esto con el fin de identificar la evolución en los temas de los textos. Con el objetivo de saber en qué han avanzado los temas según las citaciones que tienen los escritos académicos recientes en frente los escritos académicos del pasado.

Finalmente, otra situación que ejemplifica la utilidad de las técnicas NLP y es similar a lo que se quiere obtener en este escrito, es la abordada por (Marcos 2017). En ese trabajo de grado, se implementa el modelado de tópicos LDA para realizar el perfilado de Blogs. El investigador denota la aplicabilidad del LDA, para la categorización de textos que pueden presentar una gran variabilidad de temas posibles, pero debido a esta variabilidad se tuvo que apoyar de la lectura al azar de unos cuantos Blogs según los tópicos asignados para realizar la caracterización de los tópicos debido a que el LDA arrojaba resultados difíciles de comprender. Pero incluso con las posibles fallas, el investigador logró demostrar la utilidad de LDA para la caracterización de escritos cuando no se sabe de antemano las posibles temáticas a encontrar.

# 6 Marco Teórico

En este apartado se muestran los conceptos necesarios para la correcta comprensión de este estudio. Además , se presentan las definiciones requeridas para el uso apropiado del modelo de Asignación Latente de Dirichlet.

## 6.1. Marco Conceptual

### 6.1.1. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (NLP) es el uso de técnicas computacionales para analizar el lenguaje humano escrito (Miller 1981). Estas técnicas se plantearon a partir del año 1954, donde uno de los primeros avances fue la traducción de oraciones de idioma Ruso a Inglés mediante una máquina haciendo uso de unas cuantas reglas gramaticales (Hutchins 2004). También se tuvo un gran avance entre 1964 y 1966 gracias al Instituto Tecnológico de Massachusetts y la creación del programa ELIZA, el cual hacía posible la comunicación entre una persona y una computadora a partir de reglas de descomposición gramatical y adecuadas transformaciones de textos (Weizenbaum 1966). Con el pasar de los años se desarrollaron diferentes propuestas de *chabot* (aplicación que permite interacción entre una persona y la máquina a través de texto o audio), hasta llegar a métodos de NLP relativamente recientes basados en aprendizaje automático (machine learning) y redes neuronales (neural networks).

### 6.1.2. Aprendizaje automático (Machine learning)

El aprendizaje automático se basa en la creación e implementación de algoritmos computacionales cuyo fin es imitar o emular la inteligencia humana mediante repeticiones, partiendo de un conjunto de datos y alcanzando un objetivo deseado sin estar programado para generar un resultado en particular (Naqa et al. 2015). Un claro ejemplo de esto es tomar fotografías de animales y que mediante aprendizaje automático, se clasifiquen de acuerdo a nombres de animales (distinguir entre gatos y perros) o mediante la creación de grupos que contienen animales similares pero sin tener un nombre asignado a estos grupos (agrupar animales por colores, por forma de las orejas, etc.), esto gracias a que existen diferentes tipos de aprendizaje automático.

Los tipos de algoritmos de aprendizaje automático son:

- **Aprendizaje supervisado**
  Consiste en definir ciertas características como por ejemplo el tamaño, la forma o el color, y que según la combinación de estas, se asocie una etiqueta como el nombre de una fruta o de un animal.

- **Aprendizaje no supervisado**
  En este caso, no se asocian características a etiquetas sino que el algoritmo encuentra patrones a base de prueba y error, como el camino adecuado que debe seguir un automóvil para estacionarse sin golpear nada a su alrededor o la agrupación de elementos que son similares pero que el algoritmo no sabe exactamente lo que son.

- **Aprendizaje semi-supervisado**
  Siendo una especie de combinación entre los dos tipos de aprendizaje automático mencionados, se basa en que una parte de los datos cuenta con etiquetas y otra no, de tal manera que se usan estas etiquetas conocidas para aprender sobre el apartado sin etiquetar.

### 6.1.3. Minería de datos

Con el paso del tiempo surgen diversas maneras de obtener mediciones, lo que aumenta la cantidad de información que se recolecta y a su vez, se almacenan demasiados datos que sin un correcto procedimiento o análisis, es información que se pierde o que no se aprovecha. Claramente, esto va ligado al uso de herramientas de cómputo que cada vez permiten una mayor capacidad de procesamiento de los datos. La minería de datos se compone de una serie de metodologías que posibilitan el análisis y transformación de grandes volúmenes de datos en información útil para la toma de decisiones, identificar relaciones o caracterizar individuos (Santos-Pereira et al. 2022).

### 6.1.4. Minería de texto

De acuerdo con (Miller 1981), además de ser una de las ramas de la minería de datos, es el proceso de analizar y obtener información a partir de textos, determinando patrones o correlaciones entre los términos para descubrir relaciones que pueden estar ocultas. Estos procesos pueden ser aplicados en la clasificación de spam de los correos electrónicos, en el análisis del tipo de publicaciones que se realizan en redes sociales, también para resumir una gran cantidad de informes médicos y demás elementos que cuenten con texto. Todo esto se logra gracias a la combinación de técnicas de Procesamiento de Lenguaje Natural, métodos estadísticos y el uso de aprendizaje de máquina.

### 6.1.5. Datos estructurados y no estructurados

Los datos estructurados son aquellos que pueden encajar en campos y columnas como por ejemplo, en una hoja de cálculo, pues están bien organizados. Mientras que los no estructurados pueden ser los videos, audios, textos, imágenes satelitales, etc. (Miller 1981).

### 6.1.6. Tópicos

Se refiere a un conjunto de palabras que suelen aparecer juntas en los mismos contextos y que además, nos permiten observar relaciones que seríamos incapaces de observar a simple vista, se refiere de alguna manera a los temas (Amorrosta 2021).

### 6.1.7. Modelado de tópicos

Es una técnica del procesamiento de lenguaje natural que se basa en identificar las temáticas de textos (Amorrosta 2021), resumiendo de alguna manera el contenido de una gran cantidad de documentos. Esta herramienta estadística permite extraer variables latentes o no observadas de grandes bases de datos no estructurados (Vayansky & Kumar 2020).

Los modelos de tópicos se basan en la premisa de que en estos grupos de textos existen diversos temas que no están explícitos y que a su vez, permiten de alguna manera la organización de los documentos debido a estos tópicos (Luque et al. 2021).

### 6.1.8. Lematización

Busca reducir inflexiones o derivaciones de las palabras encontradas en un grupo de textos. Es decir, busca reducir las palabras a una palabra en concreto que sirva como palabra representativa de todas las posibles inflexiones o derivaciones que pueda tener esta palabra, esta palabra es conocida como lema (Lovins 1968).

### 6.1.9. Tokenización

Consiste en dividir un texto en las unidades que lo conforman, siendo estas unidades los elementos más sencillos con un significado propio para el análisis en cuestión (Benchimol et al. 2022). Por ejemplo, tomando la oración: EL JOVEN SABE LEER, con la tokenización el resultado sería "$EL$" "$JOVEN$" "$SABE$" "$LEER$".

### 6.1.10. Stopwords

Son las palabras que suelen ser las más frecuentes en un idioma, pero carecen de significado por sí solas (Rajaraman & Ullman 2011). Por ejemplo "*a*", "*de*", "*el*", "*si*". Sin embargo,

según los documentos y el análisis que se esté realizando, también se pueden considerar stopwords aquellas palabras que no tengan un buen aporte a los resultados o que se repitan mucho debido a su naturaleza en el entorno que se está analizando.

### 6.1.11. Corpus

Generalmente hace referencia a una colección de documentos o archivos que serán objeto de estudio. Suele ser un conjunto de datos que pueden estar relacionados como por ejemplo del entorno bancario, de salud o deportes para realizar un análisis en dicho campo, pero también pueden ser datos de diferentes entornos cuyo propósito es clasificarlos de acuerdo a su contenido.

## 6.2. Marco Teórico Estadístico

### 6.2.1. Distribución Multinomial

Sea $X_i \in \{0, \cdots, n\}$ una variable aleatoria que indica la cantidad de veces que ha ocurrido el resultado $i$ sobre los $n$ sucesos, entonces el vector $X = (X_1, \cdots, X_n) \in \mathbb{R}^k$ tiene una distribución Multinomial con parámetros $n$ y $k$ (donde $n$ es el número de intentos y $k$ la cantidad de posibles resultados) y su probabilidad está dada por (Siegrist 2022):

$$P(X_1 = x_1, \cdots, X_n = x_n) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \tag{6-1}$$

Donde $n > 0$, $\sum X_i = n$ y $\sum p_i = 1$

La esperanza matemática del suceso $i$ observado en $n$ pruebas y su varianza son:

$$E(X_i) = np_i \tag{6-2}$$

$$Var(X_i) = np_i(1 - p_i) \tag{6-3}$$

### 6.2.2. Distribución Dirichlet

La distribución Dirichlet con parámetros $k$ y $\alpha = (\alpha_1, \cdots, \alpha_K)$ (donde $k$ es el número de categorías y $\alpha$ son los parámetros de concentración en dichas categorías), tiene una función de densidad que está dada por (Kotz et al. 2005):

$$f(x_1, \cdots, x_k) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \tag{6-4}$$

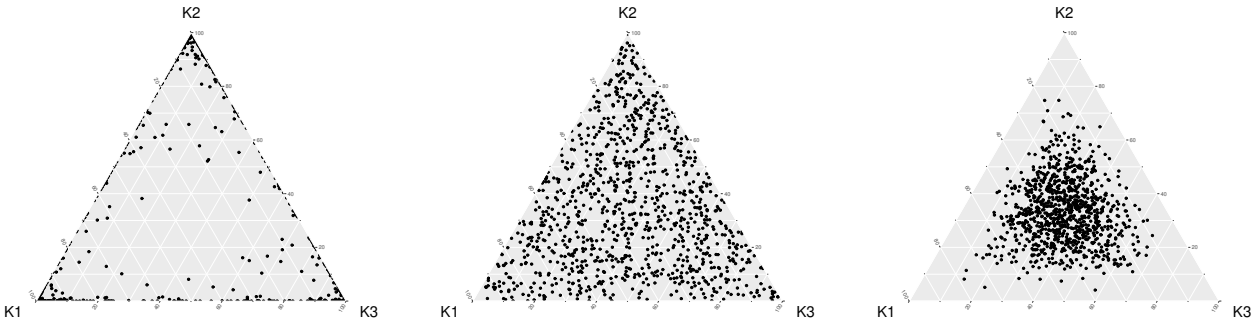
Donde $\sum_{i=1}^{k} x_i = 1$, $x_1, \cdots, x_k > 0$ y $\alpha_1, \cdots, \alpha_k \geq 0$

La esperanza y varianza de la variable aleatoria $X_i$ es:

$$E[X_i] = \frac{\alpha_i}{\sum_{i=1}^{K} \alpha_k} \tag{6-5}$$

$$Var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}; \;\; \alpha_0 = \sum_{i=1}^{K} \alpha_i \tag{6-6}$$

En la Figura **6-1** se ve la forma en que se distribuyen las 1.000 observaciones cuando se definen tres categorías (las categorías serían las puntas de los triángulos) y cuando se varía su parámetro de concentración $\alpha$. De modo que en **6-1**a se evidencia que si $\alpha < 1$, entonces las observaciones se concentran más en las categorías o puntas. En **6-1**b se muestra que cuando $\alpha = 1$, las observaciones no parecen concentrarse en sitios específicos sino que de alguna manera están esparcidos uniformemente. Por último, en **6-1**c se observa que cuando $\alpha > 1$ las observaciones no se concentran en las categorías sino, en el centro del espacio.




(a) $\alpha = (0{,}08, 0{,}08, 0{,}08)$ (b) $\alpha = (1, 1, 1)$ (c) $\alpha = (5, 5, 5)$

**Figura 6-1**: Distribución de 1.000 observaciones aleatorias Dirichlet en las categorías según el parámetro de concentración.

### 6.2.3. Matriz término-documento (DTM)

Es la representación matricial de un conjunto de documentos o archivos de texto. En esta matriz las filas representan los términos, y las columnas los documentos. Cada celda de esta

matriz indica la frecuencia de aparición del $i$-ésimo término en el $j$-ésimo documento. En la Ecuación 6-7 se muestra una DTM para cuando se tienen M documentos.

$$
\begin{array}{cc}
 & \text{Documentos} \\
 & \begin{array}{ccccccc} d_1 & d_2 & d_3 & \cdots & d_M \end{array} \\
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_i \end{array} & \begin{bmatrix} 1 & 1 & 0 & \cdots & 2 \\ 0 & 2 & 3 & \cdots & 0 \\ 3 & 0 & 5 & \cdots & 3 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & 3 & \cdots & 2 \end{bmatrix}
\end{array}
\quad \textit{Términos} \tag{6-7}
$$

### 6.2.4. Asignación Latente de Dirichlet

La Asignación Latente de Dirichlet o LDA (Latent Dirichlet allocation), es un algoritmo o técnica para el modelado de tópicos, el cual consiste en una distribución de probabilidades de aparición de las distintas palabras del vocabulario (Amorrosta 2021).

Este método fue introducido por Blei, Ng and Jordan en 2003 (Blei et al. 2003) y asume que cada documento puede ser representado como una distribución probabilística de temas o tópicos latentes, esta distribución de los tópicos en cada documento es obtenida mediante la distribución Dirichlet. Por otro lado, cada uno de estos tópicos es representado como una distribución probabilística de palabras, donde la distribución de las palabras en cada tópico también sigue una distribución Dirichlet (Jelodar et al. 2018).

Dado un corpus o conjunto de textos $D$ que contiene $M$ documentos, donde cada documento $j$ tiene $N_j$ palabras ($j \in 1, ..., M$), LDA modela el corpus $D$ de acuerdo al siguiente proceso:

- Elegir una distribución multinomial $\varphi_i$ para el tópico $i$ ($i \in \{1, ..., K\}$) de una distribución Dirichlet con parámetro $\beta$
- Elegir una distribución multinomial $\theta_j$ para el documento $j$ ($j \in \{1, ..., M\}$) de una distribución Dirichlet con parámetro $\alpha$
- Para una palabra $w_t$ ($t \in \{1, ..., N_j\}$):
  - Seleccionar un tópico $z_t$ de $\theta_j$
  - Seleccionar una palabra $w_t$ de $\varphi_{z_t}$

En el proceso anterior, las palabras en los documentos solo son variables observadas, mientras que las otras son variables latentes ($\varphi$ y $\theta$) e hiperparámetros ($\alpha$ y $\beta$).

La idea del modelo es generar la misma cantidad de $M$ documentos pero siguiendo el proceso mencionado, de tal manera que se haría una comparación de este corpus generado por el modelo y el corpus original. De este modo, se pueden generar muchos modelos y cada modelo va a generar un corpus diferente, entonces es de interés aquel modelo que genera el corpus más parecido al corpus original.

La probabilidad de obtener el corpus $D$ con el modelo, se obtiene de la siguiente manera:

$$P(W, Z, \theta, \varphi, \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\varphi_{Z_{j,t}}) \tag{6-8}$$

Donde $\alpha$ y $\beta$ son parámetros de la distribución Dirichlet a nivel de corpus, que se supone que se muestrean una vez en el proceso de generación de un corpus. Además, $K$ es el número de tópicos, $M$ es el número de documentos y $N$ es el número de palabras según el documento. Las variables $\theta_j$ son variables a nivel del documento muestreadas por documento. Las variables $Z_{j,t}$ y $W_{j,t}$ son variables a nivel de las palabras y son muestreadas para cada palabra en cada documento.

A continuación se explica con más detalle las partes que componen la ecuación 6-8.

Partiendo de $K$ tópicos que se quieran obtener:

$$\prod_{j=1}^{M} P(\theta_j; \alpha) \tag{6-9}$$

La primera parte (Ecuación 6-9) hace referencia a que para cada documento $j$, se elige aleatoriamente la distribución de los tópicos en dicho documento mediante una distribución Dirichlet con $\alpha < 1$ para que el documento resulte en la medida de lo posible, representado con un tópico o pocos. Un ejemplo de esto se muestra en la Figura **6-2**, donde se han definido $K = 3$ tópicos, el documento resulta con un 60 % del Topico 1, 20 % del Tópico 2 y un 20 % del Tópico 3. Estos porcentajes representan la probabilidad de que dicho Tópico se encuentre en determinado documento.

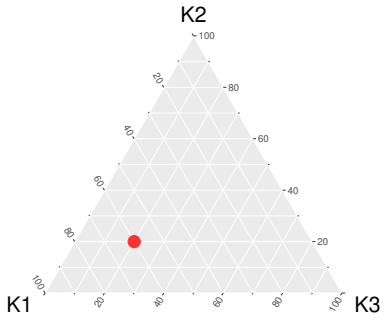



**Figura 6-2**: Distribución de 3 tópicos en un trabajo de grado.

$$\prod_{i=1}^{K} P(\varphi_i; \beta) \tag{6-10}$$

La segunda parte (Ecuación 6-10) hace referencia a que para cada tópico $i$ se elija aleatoriamente la distribución de las palabras haciendo uso de una distribución Dirichlet con $\beta < 1$. Entonces de manera similar al caso anterior, si por ejemplo sólo se contara con tres palabras, el Tópico 1 podría resultar con 50 % de la Palabra 1, 30 % de la Palabra 2 y 20 % de la Palabra 3. Estos porcentajes hacen referencia a la probabilidad de que la palabra pertenezca a cierto Tópico, de tal manera que una palabra puede estar en varios Tópicos pero con mas probabilidad de pertenecer a un Tópico que a los otros.

$$\prod_{t=1}^{N} P(Z_{j,t}|\theta_j) \tag{6-11}$$

La tercera parte (Ecuación 6-11) hace referencia a que según la distribución de los tópicos en los documentos (Ecuación 6-9), se extrae una muestra aleatoria de los tópicos teniendo en cuenta la proporción de cada tópico en el documento. Con la Tabla **6-1** se muestra que si por ejemplo se tuviera la misma distribución de tópicos del ejemplo de la Ecuación 6-9 (60 %, 20 %, 20 %), se podría obtener una muestra de 10 tópicos donde se obtiene cinco veces el Tópico 1, tres veces el Tópico 2 y dos veces el Tópico 3.

**Tabla 6-1**: Ejemplo de muestra de tópicos de la Ecuación 6-11

| TÓPICOS | MUESTRA DE TÓPICOS |
|---|---|
| Tópico 1 | Tópico 1 |
| Tópico 1 | Tópico 1 |
| Tópico 1 | Tópico 1 |
| Tópico 1 | Tópico 2 |
| Tópico 1 | Tópico 1 |
| Tópico 1 | Tópico 3 |
| Tópico 2 | Tópico 3 |
| Tópico 2 | Tópico 2 |
| Tópico 3 | Tópico 2 |
| Tópico 3 | Tópico 1 |

$$P(W_{j,t}|\varphi_{Z_{j,t}}) \tag{6-12}$$

La cuarta parte (Ecuación 6-12) hace referencia a que según la distribución de las palabras en los tópicos (Ecuación 6-10) y según la muestra de los tópicos (Ecuación 6-11), se toma cada tópico de la muestra y se extrae una palabra al azar de acuerdo con la distribución de las palabras en dicho tópico. Siguiendo con la muestra de topicos del ejemplo de la Ecuación 6-11, se obtendría una lista de 10 palabras donde cada una proviene de la distribución de palabras según el tópico de donde se extrae.

De esta manera se genera un documento aleatorio y se hace lo mismo para cada uno de los documentos, por lo que luego de todo el proceso se obtiene un corpus generado que se compara con el corpus original. Entonces se elige aquel modelo que genere el corpus más similar al corpus original, es decir, aquel que maximice la Ecuación 6-8. Por último, los tópicos que resultan serán los asociados al mejor modelo obtenido.

# 7 Metodología

En este apartado son detallados y explicados los pasos necesarios llevados a cabo con el fin de dar solución a los objetivos planteados de este estudio.

## 7.1. Recopilación de los Documentos

Los Documentos fueron suministrados por la dirección del programa de pregrado de Estadística de la Universidad del Valle, la dirección cuenta con una recopilación de estos trabajos desde el año 2002 hasta el año 2022. Estos trabajos de grado en su mayoría se encontraban en el formato de archivo tipo PDF. En esta investigación se prefirió continuar con este tipo de archivo y para los trabajos de grado que no se encontraban en este formato se realizó su debida transformación a archivos tipo PDF. Al finalizar este proceso se obtuvo cantidad total de 290 documentos.

Los trabajos de grado se agruparon en dos vías: año de publicación del trabajo de grado y el director de este. Esta agrupación se realizó con el fin de encontrar el comportamiento de los tópicos sobre estas dos categorías para hallar las posibles tendencias en líneas de investigación a través de los años o según los directores de trabajos de grado.

## 7.2. Procesamiento del texto

Con todos los documentos de los trabajos de grado obtenidos se formó el corpus, el cual es el objeto principal de estudio, en este corpus se tendrá en cuenta la totalidad de texto de cada documento. A estos documentos se les aplica un proceso de limpieza, el cual consiste en transformar cada documento de manera que todo el texto se encuentre escrito en minúsculas, se hace la eliminación de caracteres especiales como los números y signos de puntuación, también se eliminan las stopwords, se reduce la cantidad de espacios en blanco y se realiza la tokenización de cada documento

Después de completar el proceso de limpieza, se da paso a la realización del proceso de lematización, pero como en este proceso pueden aparecer de nuevo palabras que fueron eliminadas en la limpieza inicial de los datos, se realiza una segunda limpieza para obtener el corpus con el cual trabajar.

## 7.3. Representación de los datos

Los datos para el modelo LDA consisten en las frecuencias de las palabras o términos, por lo que es necesario construir una matriz de términos-documentos (DTM), donde las filas representan los documentos, y las columnas los términos, de tal manera que cada celda de esta matriz indica la frecuencia de aparición del $j$-ésimo término en el $i$-ésimo documento.

Como esta matriz es de frecuencias, puede darse el caso de que hayan muchas celdas vacías o con un valor muy bajo, pues hay palabras que pueden aparecer sólo en un documento o incluso aparecer solo una vez en todo el corpus. Es por esto que se hace un proceso de eliminación de palabras tal que solo queden aquellas palabras que se encuentren en el 98 % de los documentos, este porcentaje es elegido con el fin de encontrar un equilibrio entre eliminar palabras que aparecen en pocos documentos y mantener una buena cantidad de palabras para el ajuste del modelo.

## 7.4. Selección del número de tópicos

Para este estudio se ajustaron varios modelos LDA donde la diferencia entre ellos era la cantidad de tópicos. Esto con el fin de analizar los tópicos obtenidos en cada modelo y elegir aquel modelo en el que cada tópico tuviera combinaciones de palabras que se pudieran asociar de manera coherente a un tema o en este caso, a una línea de investigación. Bajo este método se determinó considerar un modelo que contara con seis tópicos.

Sin embargo, existen medidas como la perplejidad y la coherencia que sirven para elegir un número adecuado de tópicos, pero cuando se probaron en este estudio no se obtuvo un resultado muy bueno, pues ambas métricas indicaban que entre más tópicos se ajustaran en el modelo, era mejor. Pero, cuando se consideran muchos tópicos es complejo identificar el contexto de cada uno de ellos, pues muchas palabras se van a repetir en los tópicos y se dificulta el proceso de asociar un contexto o tema a cada tópico.

## 7.5. Ajuste del modelo

Haciendo uso del software R (R Core Team 2022) y la función "LDA" del paquete *topicmodels* (Grün & Hornik 2023), se realiza la construcción del modelo LDA basado en la bolsa de palabras o DTM obtenida con anterioridad. Los resultados obtenidos del LDA serán vectores de probabilidad, los cuales se tienen combinaciones de palabras con su respectiva probabilidad de presencia en el vector que la contiene. Estos vectores serán nuestros tópicos y como paso final de la metodología sería la asociación de estos tópicos a un tema en especifico según el contexto en el que se maneja el corpus, teniendo en cuenta el contexto de esta

investigación se debe asociar cada tópico a un o varios métodos estadísticos. Con los tópicos identificados se realiza su análisis frente a los directores de trabajo de grado y el año de publicación, buscando posibles relaciones entre los directores de trabajo de grado, el año de publicación y el tópico encontrado.

# 8 Resultados

En esta sección se desarrolla un análisis descriptivo de los trabajos de grado utilizados, así como los resultados del ajuste del modelo de Asignación Latente de Dirichlet y tipificaciones encontradas al respecto de los tópicos obtenidos con LDA.

Al recopilar los Documentos, la Dirección del Programa del Pregrado en Estadística de la Universidad del Valle contaba con 290 trabajos de grado de estudiantes, los cuales fueron publicados entre el año 2002 al año 2022. La distribución de trabajos de grado a través de los años se puede ver en la Figura **8-1**, donde es posible evidenciar una tendencia positiva en la cantidad de trabajos publicados a través de los años, pero con momentos destacables de baja cantidad de trabajos entregados. También destacan ciertos momentos en los que cambia la cantidad y esto puede ser debido a la acumulación de estudiantes de diferentes cohortes o a causa de eventos históricos que ocasionaron los cambios en la cantidad de trabajos de grado que fueron entregados cada año (como por ejemplo la reforma curricular del 2015, el Paro estudiantil ocurrido el año 2015 al 2016 o la pandemia del Covid-19 que tuvo sus inicios en el año 2020).

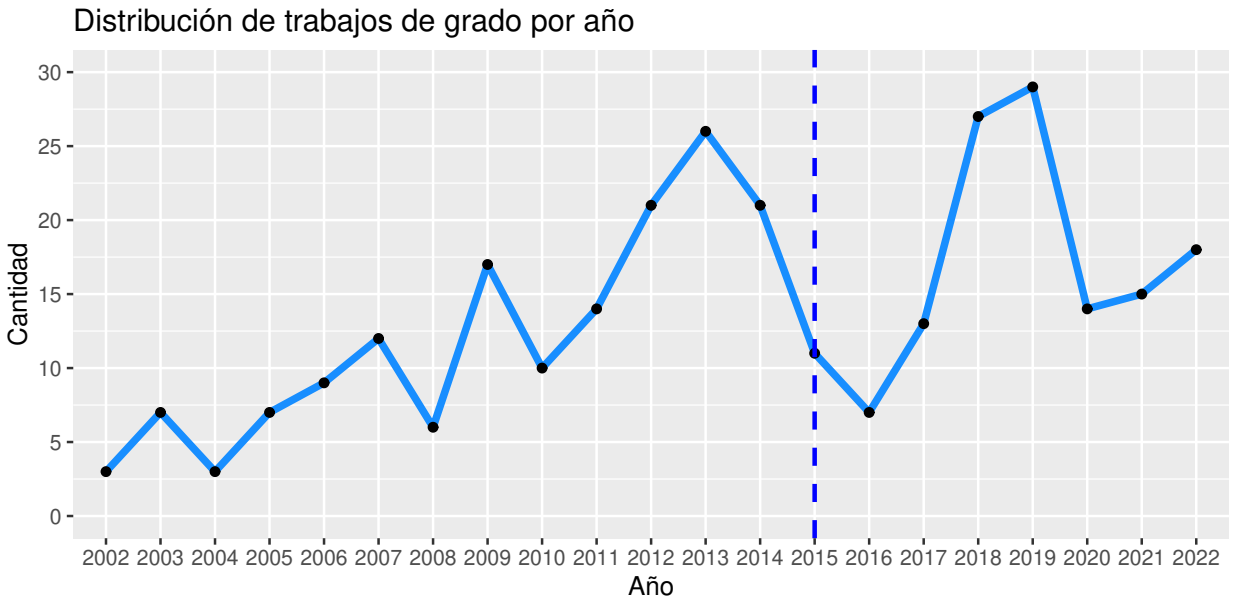



**Figura 8-1**: Distribución de la cantidad de trabajos de grado por año.

Antes del procesamiento del corpus se procede a verificar rápidamente las palabras más frecuentes de todos los documentos como es posible apreciar en la nube de palabras presentada en la Figura **8-2**. Aquí se puede ver la presencia de números, signos de puntuación y signos de ortografía como las tíldes o las virgulillas, así como también las mencionadas palabras vacías por no tener un significado propio (stop words) y que son muy frecuentes, como por ejemplo palabras como "los", "por", "que". Estos elementos no permiten visualizar el contenido esencial del texto, es por esto que se requiere realizar la limpieza del corpus.


**Figura 8-2**: Nube de palabras del corpus inicial.

En el procesamiento del corpus se pasa por varias etapas en las que partiendo de los archivos originales, se hace una primera limpieza de stop words, luego se lematiza y se realiza una segunda limpieza para eliminar los residuos generados en este paso. Por último, se hace una reducción de la DTM. La métrica con que se mide la limpieza o reducción, es la cantidad de palabras únicas o que no se repiten, es decir, las palabras sin importar su frecuencia.

De acuerdo con lo anterior, en la Figura **8-3** se muestra que el corpus inicial cuenta con 374.872 palabras únicas y que luego de todo el proceso se logra una reducción del 97 %, terminando con 11.001 palabras únicas, las cuales son el objeto de insumo para el ajuste del

modelo LDA. Cabe recordar que el ajuste del modelo depende principalmente del número de tópicos que se quieran obtener, por lo que a mayor cantidad de tópicos y sobretodo a mayor cantidad de palabras, se extiende el tiempo de ajuste de todos los modelos para elegir el mejor, de modo que esta es una de las motivaciones para realizar una buena limpieza de los documentos.

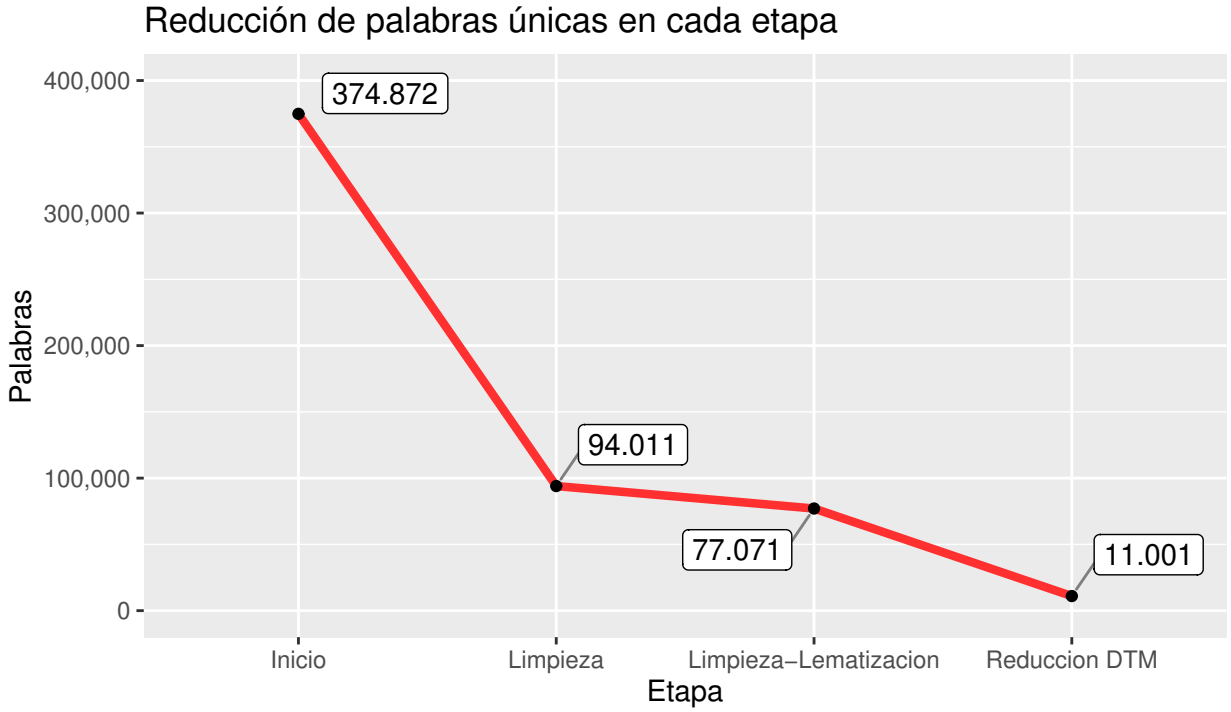



**Figura 8-3**: Reducción de palabras en las etapas de procesamiento.

Luego de realizar el procesamiento del texto requerido, procedemos a verificar la calidad de la limpieza de los datos mediante la nube de palabras presentadas en la Figura **8-4**, a grandes rasgos y basándonos en que la frecuencia de las palabras está relacionada con el tamaño de las mismas, se podría pensar que los trabajos de grado abordan temas relacionados con el tiempo, muestras, regresión, contaminación e incluso con la salud.


**Figura 8-4**: Nube de palabras del corpus procesado.

Los resultados del modelo LDA son los tópicos que se muestran en la Figura **8-5**. En esta se presentan las palabras con mayor probabilidad de pertenecer a un tópico en específico, de los seis tópicos obtenidos del modelo LDA aplicado al corpus de Trabajos de Grado de pregrado en Estadística de la Universidad del Valle.

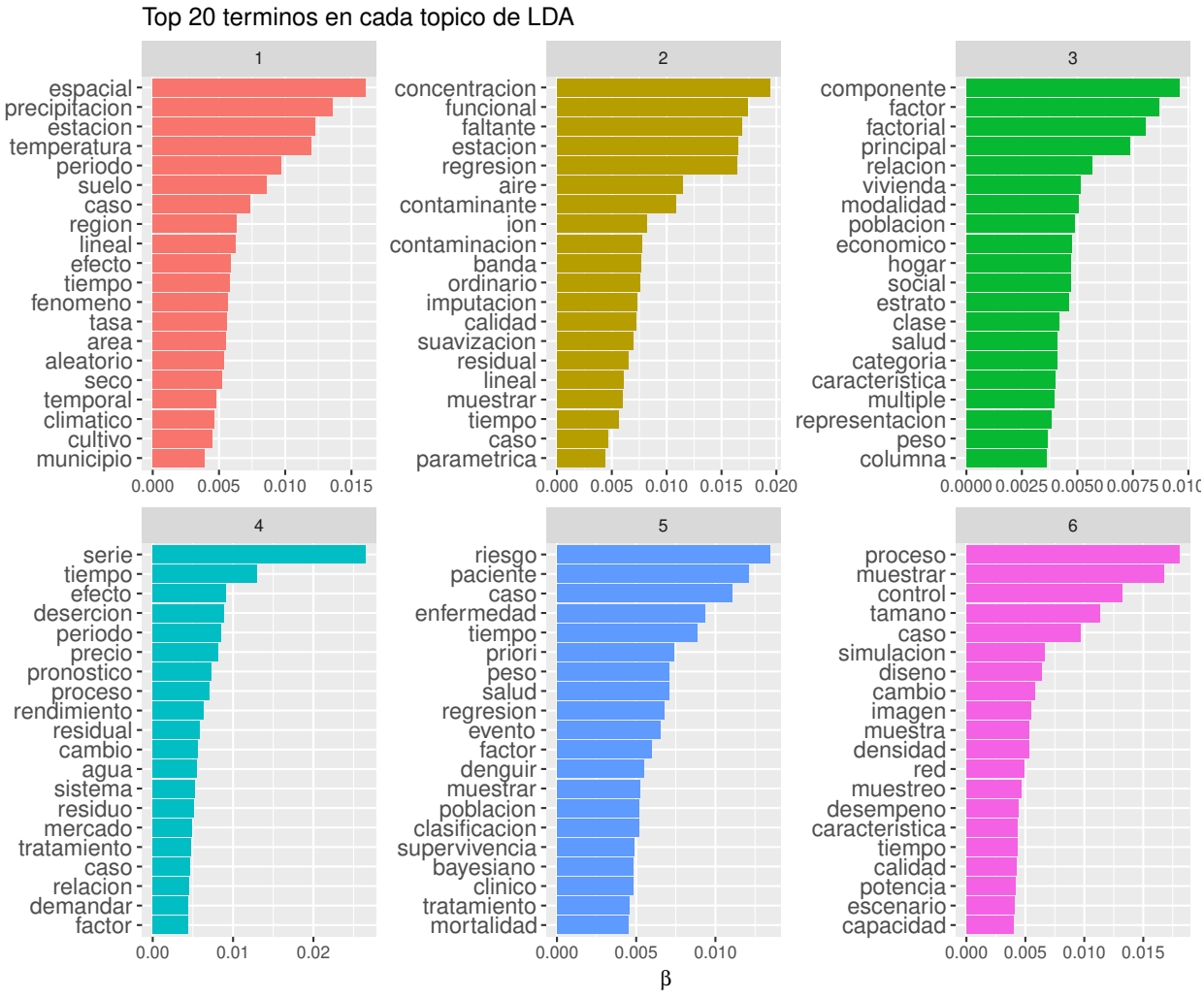



**Figura 8-5**: Distribución de las palabras en los tópicos.

El objetivo de analizar la Figura **8-5** es identificar de alguna manera el contexto de cada uno de los tópicos según las palabras y la relación entre ellas mismas, en pocas palabras se resume a nombrar los tópicos. Partiendo de esto, se buscará relacionar las palabras de los tópicos obtenidos con metodologías específicas de la estadística, mediante el conocimiento previo de las posibles técnicas que aprende un estudiante en su pregrado y busca aplicar en su trabajo de Grado.

En el Tópico 1 se observan palabras como “espacial”, “suelo”, “efecto”, “aleatorio” y “cultivo”, que da lugar a que este tópico pueda estar asociado al Diseño de Experimentos, debido a que en este campo de la estadística se suele aplicar en contextos de agricultura y se suele buscar principalmente el efecto que tienen los tratamientos en la población estudio.

En cuanto al Tópico 2, están presentes las palabras “concentración”, “funcional”, “faltante”, “regresión”, “contaminación” y “suavización”, entonces es posible atribuirle la Regresión con datos funcionales por el hecho de que al analizar datos de contaminación suele haber presencia de datos faltantes y además, es usual el uso de técnicas de suavizado de curvas en este entorno.

Cuando se analiza el Tópico 3, se aprecia que contiene palabras como “componente”, “factor”, “principal”, “social” y “vivienda”, entonces es posible una asociación con la temática estadística de Análisis Multivariante, ya que en este conjunto de técnicas se suele buscar una reducción de dimensionalidad de las variables a estudiar a una cantidad pequeña de factores y se suele aplicar estas técnicas en estudios sociodemográficos.

Pasando al Tópico 4, es importante destacar que son pocas las palabras con una alta probabilidad de pertenecer a este tópico, por lo que parece condensarse un tema muy en específico como lo son las Series Temporales y sus aplicaciones en el ámbito del análisis de precios, pronósticos, análisis de mercado y el efecto de diferentes factores en dichos casos.

Continuando con el Tópico 5, es entendible como este tópico hace referencia a la aplicación de Técnicas estadísticas en el entorno de la Salud. Esta conclusión es obtenida debido a que las primeras palabras de mayor probabilidad del tópico hacen referencia a términos que se suelen usar al tratar de enfermedades como “riesgo”, “paciente” y “caso”, ya refiriéndose a que metodologías estadísticas en especial parece ser que el tópico agrupa técnicas de distintas líneas de investigación de la estadística.

Por último, un caso particular es el del Tópico 6, pues con palabras como “proceso”, “muestra”, “tamaño”, “simulacion”, “desempeño”, “tiempo”, “calidad”, “potencia” y “capacidad”, parece que es un tópico asociado al Control de la Calidad donde claramente está implícito el muestreo, pero que también puede contener otros temas que se pueden denotar como Varios.

Con la Figura **8-6** se puede ver la manera en que se nombraron los tópicos obtenidos y también la proporción de trabajos de grado en cada uno de estos tópicos, donde si se observa la proporción más alta, se determina que el 21 % de trabajos están asociados a la Estadística en salud, acompañado de un 19.7 % de trabajos para el Análisis Multivariado, mientras que los temas de Series de tiempo y Diseños de experimentos parecen ser los temas menos abordados, acumulando un 13.1 % cada uno.

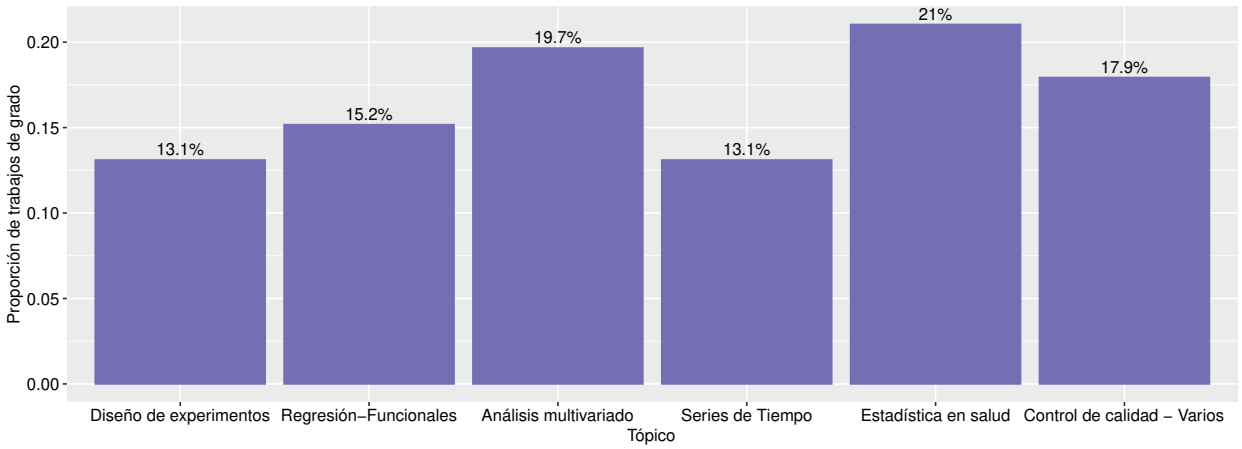



**Figura 8-6**: Proporción de trabajos de grado en cada tópico.

Anteriormente se hizo el análisis de la distribución de las palabras en los tópicos de acuerdo con la Figura **8-5**, pero también es de interés saber la manera en que se distribuyen estos tópicos en los documentos. Es por esto que en la Tabla **8-1**, a manera de ejemplo se presenta la distribución de los tópicos en siete documentos, de tal manera que por ejemplo el 73.6 % del Documento 1 está asociado al Diseño de experimentos pero también tiene 10.1 % de Análisis multivariado y 10.2 % de Estadística en la salud, aunque el tema dominante es el mencionado primero. Desde el Documento 2 hasta el Documento 6 se presenta esta situación en la que predomina por mucho uno de los tópicos, a excepción del Documento 7, pues es un trabajo de grado que usa diversos temas o enfoques, puesto que el 45.4 % del documento está asociado al Análisis multivariado y el 48.8 % está relacionado con la Estadística en la salud.

**Tabla 8-1**: Distribución de tópicos en algunos trabajos de grado

| Documento | Diseño | Regresión | Multivariado | Series | Salud | Ctrl. calidad |
|---|---|---|---|---|---|---|
| **1** | 73.6 % | 1.7 % | 10.1 % | 1.5 % | 10.2 % | 3.0 % |
| **2** | 0.4 % | 73.2 % | 0.5 % | 0.8 % | 0.7 % | 24.4 % |
| **3** | 1.5 % | 0.7 % | 77.6 % | 0.4 % | 16.9 % | 3.0 % |
| **4** | 0.6 % | 0.7 % | 1.7 % | 94.6 % | 0.5 % | 1.8 % |
| **5** | 2.1 % | 0.8 % | 3.9 % | 1.6 % | 86.9 % | 4.6 % |
| **6** | 0.4 % | 1.0 % | 0.7 % | 0.7 % | 0.5 % | 96.7 % |
| **7** | 1.2 % | 0.9 % | 45.4 % | 0.3 % | 48.8 % | 3.5 % |

Analizando la información presente en la Figura **8-7** destaca cómo más de la mitad de todos los trabajos de grado que han sido desarrollados a lo largo del periodo de tiempo del corpus, 2002 a 2022, han sido publicados bajo la dirección de 5 directores en específico. Los directores que destacan y sus porcentajes de trabajos de grado bajo su dirección respectivamente son: el director 19 con 19.93 % de los trabajos publicados bajo su dirección, el director 23 con el 15.73 %, el director 17 con 11.11 %, el director 29 con 7.7 % y finalmente el director 36 con 7.7 %.


**Figura 8-7**: Cantidad de trabajos de grado por Director.

En la Figura **8-8** de la cantidad de trabajos de grado por Director y por tópico asignado es posible destacar el comportamiento esperado de que cada director va a tender a dirigir trabajos de grado en los cuales maneja temas en los que se ha especializado a lo largo de su carrera profesional como estadístico. Por ejemplo, destaca como el director 23 tiene gran mayoría de los trabajos de grado bajo su dirección asignados al tópico 5, el cual fue definido como el tópico de Estadística aplicada a la Salud. Otro caso destacable, es el del director con mayor cantidad de documentos, el director 19 presenta una clara preferencia al tópico 2, el cual fue definido como el tópico de Regresión y datos funcionales. También se tiene el caso del director 29, donde se demuestra una clara preferencia por el tópico 1 referente a Diseño experimental. El director 36 presenta una preferencia al tópico 3 de análisis multivariante y finalmente el director 17 con una preferencia al tópico 6 de Calidad de datos y temas varios.

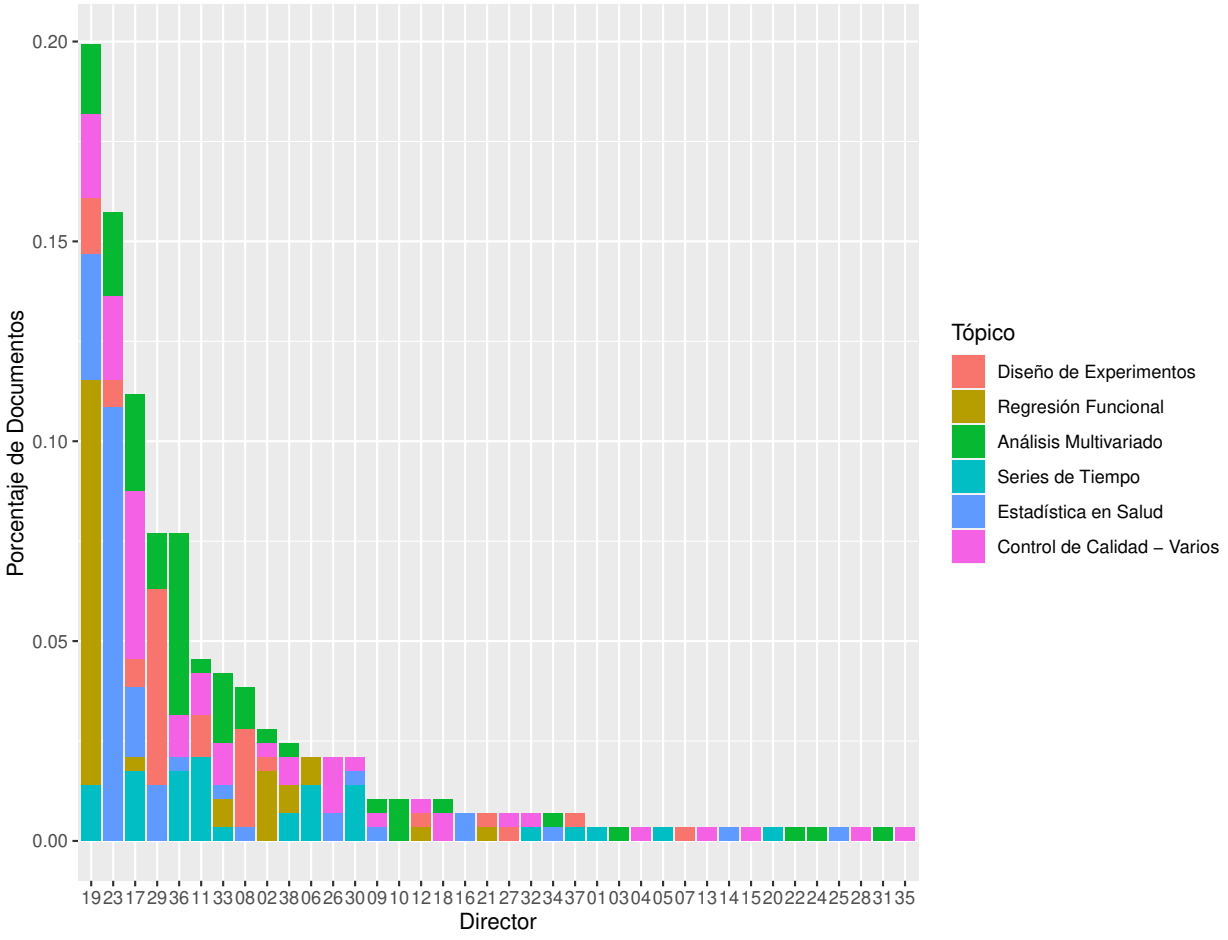



**Figura 8-8**: Proporción de trabajos de grado por Director y por tópico asignado.

Analizando el comportamiento de la proporción de documentos según su tópico asignado a lo largo del tiempo en la Figura **8-9**, es destacable el comportamiento de picos para el Tópico 5 de Estadística en la Salud en los años 2017 al 2019 que recordando la Figura **8-1** estos fueron los años con mayor cantidad de Trabajos de grado publicados, también destaca como este tópico aparente ser el de mayor popularidad en los últimos años. Por otra parte, es destacable el comportamiento del tópico referente al Análisis Multivariado teniendo mayor proporción al inicio y final del tiempo limitado en el corpus. Pero en general, no se presenta un comportamiento del tipo que demuestre que se ha obtenido mayor preferencia por cierto tema de los obtenidos con el modelo LDA a lo largo del tiempo, ya que parece ser que cada tópico tiene cierto pico de proporcionalidad en distintos años representando que esto se deba posiblemente a tiempos en que los directores que mostraban mayor cantidad de Trabajos de grado bajo su dirección se encontraban con mayor cantidad de trabajo.

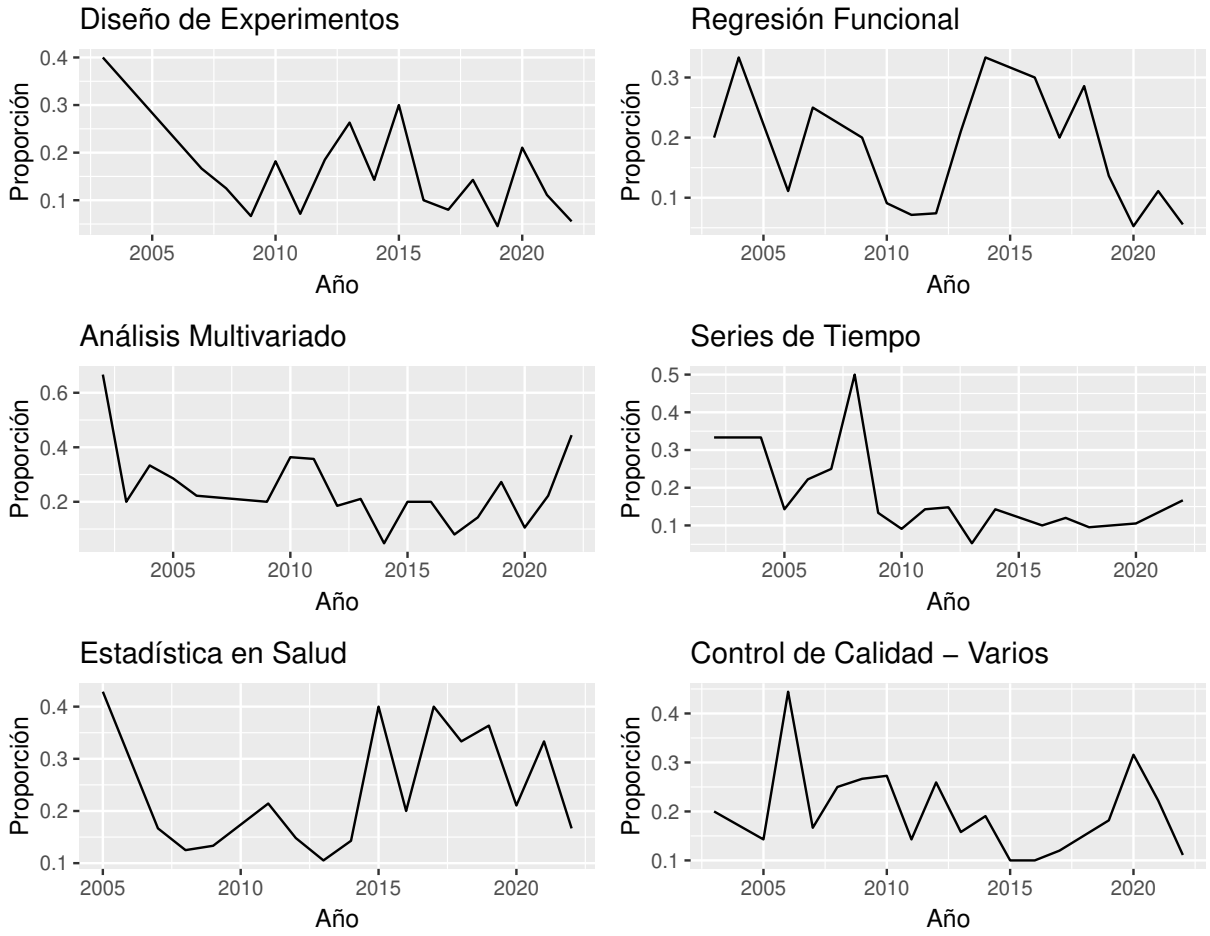



**Figura 8-9**: Proporción de tópicos en el tiempo.

# 9 Conclusiones

Los resultados del estudio realizado, permitieron identificar las tendencias en lineas de investigación que los estudiantes del pregrado en Estadística de la Universidad del Valle optaron por abordar en sus trabajos de grado durante el periodo de tiempo del año 2002 hasta el año 2022. Los tópicos obtenidos con el ajuste del Modelo LDA hicieron posible determinar los temas en los que trabajaron los estudiantes, pues se encontró que el 21 % de los trabajos se enfoca en la Estadística Aplicada en Salud, el 19.7 % en el Análisis Multivariado, el 17.9 % en lo que comprende el Control de la calidad, Muestreo y Simulación, el 15.2 % en todo lo que abarca la Regresión y el Análisis de Datos Funcionales, y por último, el Diseño de Experimentos y las Series de Tiempo contenían cada uno un 13.1 % de todos los trabajos de grado.

La identificación de estos tópicos da lugar a que se pueda saber qué temas son los que más se suelen tratar, los que menos se usan y sobretodo aquellos que no se usan, esto sirve para tratar de incentivar las aplicaciones de las diferentes áreas de investigación o directamente empezar a aplicar más metodologías o temas que no son muy frecuentes o que no se suelen usar. Esto con el fin de aumentar las áreas de aplicación y las técnicas o métodos estadísticos utilizados en general.

Además de identificar los temas de los trabajos de grado, también se pudo conocer que el 62.17 % de estos trabajos fueron dirigidos por cinco directores que representan solo un 13 % de todos los directores. Lo cual es una cifra notable, que puede estar relacionada con el tiempo que estos profesores llevan ejerciendo en la Universidad del Valle, puesto que a mayor tiempo, podría esperarse que dirijan muchos trabajos de grado y que a su vez, estos trabajos aborden temas con los que estos profesores se han especializado a lo largo de su carrera. Esto se pudo evidenciar cuando se analizaron los tópicos de los trabajos dirigidos por cada uno de estos directores, pues para cada director siempre había un tópico que era mucho más frecuente que los demás. De estos Directores con mayor cantidad de trabajos de grado bajo su dirección también se encontró la clara preferencia que tienen ellos a ciertos temas, dando a entender que estos temas presentan cierto manejo debido a que son las áreas en que se especializan los directores con mayor presencia referente a la dirección de trabajos de grado.

Finalmente, después de analizar al comportamiento de los tópicos encontrados al paso de los años en los que se encuentra limitado el corpus. Se encontraron que ciertos tópicos tuvieron momentos de pico en tendencia, pero que últimamente se están dejando de utilizar, como es el caso para los tópicos de Series de Tiempo, Diseño de Experimentos y Regresión Funcional. Por otra parte los tópicos de Estadística en Salud y el Análisis Multivariado están presentando una alza en proporción en los últimos años, en especial el de Análisis Multivariado siendo el tópico de mayor proporción en el año 2022.

# 10 Recomendaciones

Para un próximo estudio y si se continúa con el uso del modelo LDA, se recomienda incluir los nuevos trabajos de grado que se vayan publicando y ajustar modelos con diferente número de tópicos para identificar si las áreas de investigación y temas se han ampliado o siguen igual. También se propone en futuras aplicaciones del modelo hacer uso de las posibles métricas que ayuden en el proceso de selección de la cantidad de tópicos.

Con fines de comparación y de medir la eficiencia del modelo, se puede proponer el uso de un método de agrupación de los trabajos de grado para determinar si estos grupos se asemejan a los grupos de documentos que formaría cada uno de los tópicos o si se da el caso de que se forma un número de grupos diferente a seis.

# Bibliografía


Amorrosta, S. (2021), 'Introducción al topic modeling con gensim (i): fundamentos y preprocesamiento de textos', `https://elmundodelosdatos.com/topic-modeling-gensim-fundamentos-preprocesamiento-textos/`.

Benchimol, J., Kazinnik, S. & Saadon, Y. (2022), 'Text mining methodologies with r: An application to central bank texts', **8**, 100286.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *J. Mach. Learn. Res.* **3**, 993–1022.

Cho, S. M., ung Park, C. & Song, M. (2020), 'The evolution of social health research topics: A data-driven analysis', *Social Science  Medicine* **265**, 113299.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0277953620305189*

Coenen, F. (2011), 'Data mining: Past, present and future', *Knowledge Eng. Review* **26**, 25–29.

Grün, B. & Hornik, K. (2023), *topicmodels: Topic Models*. R package version 0.2-14.
**URL:** *https://CRAN.R-project.org/package=topicmodels*

Hotho, A., Nürnberger, A. & Paaß, G. (2005), 'A brief survey of text mining', *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology* **20**(1), 19–62.
**URL:** *http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf*

Hutchins, W. J. (2004), The georgetown-ibm experiment demonstrated in january 1954, *in* R. E. Frederking & K. B. Taylor, eds, 'Machine Translation: From Real Users to Research', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 102–114.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L. (2018), 'Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey', *Multimedia Tools and Applications* **78**, 15169 – 15211.

Kotz, S., Balakrishnan, N. & Johnson, N. (2005), *Continuous Multivariate Distributions, Models and Applications: Second Edition*.

Lovins, J. B. (1968), 'Development of a stemming algorithm', *Mech. Transl. Comput. Linguistics* **11**, 22–31.

Luque, C., Galvis, J., Rubriche, J. & Sosa, J. (2021), 'Modelamiento de tópicos para identificar patrones en la investigación científica del covid-19', *Comunicaciones en Estadística* **14**, 48–66.

Marcos, J. V. (2017), 'Modelado de tópicos para perfilado de blogs'.
**URL:** *https://e-archivo.uc3m.es/handle/10016/28247*

Miller, L. A. (1981), 'Natural language programming: Styles, strategies, and contrasts', *IBM Systems Journal* **20**(2), 184–215.

Naqa, I. E., Li, R. & Murphy, M. J. (2015), 'Machine learning in radiation oncology'.

R Core Team (2022), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *https://www.R-project.org/*

Rajaraman, A. & Ullman, J. D. (2011), *Mining of Massive Datasets*, Cambridge University Press.

Santos-Pereira, J., Gruenwald, L. & Bernardino, J. (2022), 'Top data mining tools for the healthcare industry', *Journal of King Saud University - Computer and Information Sciences* **34**(8, Part A), 4968–4982.
**URL:** *https://www.sciencedirect.com/science/article/pii/S131915782100135X*

Siegrist, K. (2022), 'The multinomial distribution', `https://stats.libretexts.org/@go/page/10237`.

Vayansky, I. & Kumar, S. (2020), 'A review of topic modeling methods', *Information Systems* **94**, 101582.

Vukanti, V. & Jose, A. (2021), 'Business analytics: A case-study approach using lda topic modelling', *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* pp. 1818–1823.

Weizenbaum, J. (1966), 'Eliza—a computer program for the study of natural language communication between man and machine', *Commun. ACM* **9**(1), 36–45.
**URL:** *https://doi.org/10.1145/365153.365168*