Modelo de datos funcionales para la caracterización de la propagación de una epidemia

Ana María Ortiz Másmela

Universidad ECCI
Departamento de Ciencias Básicas
Bogotá, Colombia

2020

Modelo de datos funcionales para la caracterización de la propagación de una epidemia

Ana María Ortiz Másmela

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Profesional en Estadística

Director:

MSc. en Estadística, José Alexander Fuentes Montoya

Codirector:

MSc. en Ingeniería electrónica y de computadores, Luis Enrique Romero Medina

Universidad ECCI

Departamento Ciencias Básicas

Bogotá, Colombia

2020

Título en español

Modelo de datos funcionales para la propagación de una epidemia.

Title in English

Functional data model of epidemic spread.

Resumen: Las epidemias han acompañado a la humanidad desde las primeras civilizaciones y cuando entran en escena, crean una huella imborrable. Esto se evidencia en el número de víctimas mortales y en el alcance geográfico de propagación, dejando en su camino consecuencias graves y devastadoras. Por ello, la comunidad internacional ha asignado parte de sus recursos en la realización de estudios de epidemias, con el objetivo de contrarrestar los efectos nocivos que producen en las personas y sociedades. En el presente trabajo de investigación, se desarrolla la modelización de una epidemia utilizando Datos Funcionales, los cuales se generan a través de simulaciones que proporcionan diferentes escenarios de propagación de un brote y su respectivo análisis.

Abstract: Epidemics have accompanied the humanity since the first civilizations and when they enter the scene, they create an indelible mark. This is evidenced in the number of fatalities and their geographical scope of spread, leaving severe and devastating consequences in their way. For this reason, the international community has allocated part of its resources on conducting studies of epidemics with the purpose to counter the harmful effects that they produce on people and societies. In the present work presented in this investigation, the modeling of an epidemic is developed using Functional Data, which is generated through simulations that provide different scenarios of the spread of an outbreak and their respective analysis.

Palabras claves: Epidemias, Datos Funcionales.

Keywords: Epidemics, Functional Data.

Dedicatoria

Esta tesis la dedico a toda mi familia, pero en especial a mi madre, mi hermana y Nanita, quienes estuvieron conmigo apoyándome en cada momento y brindándome la fuerza para jamás rendirme. ¡Las amo!

"Sólo hay una cosa que hace que un sueño sea imposible de alcanzar: el miedo al fracaso"

Paulo Coelho

Agradecimientos

Primero que todo, agradezco a mi familia, quiénes fueron parte fundamental para que culminar este proceso. Siempre estuvieron para darme fuerza, brindarme una mano y para mostrarme que soy más fuerte de lo que creo. A mi madre, quién siempre estuvo conmigo, dándome fuerzas de aliento cuando quise rendirme y quién me brindó los abrazos más sinceros cuando pensé que ya no podía más; a mi Nanita por sus valiosos consejos, por siempre escucharme, entenderme, por su preocupación y su amor incondicional; a mi hermana Sofi, quién me decía *¡Usted puede, usted es inteligente!* Seguido de un abrazo. Son cosas que llenan mi corazón de agradecimiento y a Danna quién siempre creyó en mí y me apoyó incondicionalmente.

En segunda instancia, agradezco enormemente a mi director de tesis, Alexander Fuentes. Sin sus enseñanzas, su paciencia, humildad y apoyo no hubiese podido terminar este proyecto; aunque no todo fue color de rosas y existieron situaciones difíciles, jamás dejó de creer en mí y pudimos sacar este proyecto adelante.

Por último y no menos importante, a mis compañeros de carrera Sergio Cabrera y Yamith Corredor, estuvieron conmigo siempre, enseñándome, guiándome, ayudándome y mostrando que su amistad es única, valiosa e incondicional; inicié mi carrera con ellos y su apoyo fue fundamental para finalizar esta etapa de mi vida. Ellos dos junto con Fabio Yomayusa y Karen Ortiz, siempre me dieron una razón para continuar con mis sueños, brindando siempre palabras de aliento justo en el momento perfecto.

¡Gracias a todos ustedes por mostrarme que los sueños se pueden cumplir, que por más difícil y oscuro que se vea el panorama habrá una luz y en ella estarán ustedes para festejar conmigo mis triunfos!

1. Índice general

2. Índice de tablas

3. Índice general de Figuras

4. Introducción

El estudio de las epidemias en la actualidad ha tomado un papel crucial en el desarrollo de las sociedades, ya que estas afectan de manera circunstancial dejando una huella imborrable con su paso. Por este motivo se hace importante realizar análisis para poder comprender de alguna manera el comportamiento de las misma y a su vez poder tomar acciones para su control y/o extinción.

Por la pandemia del covid-19 los gobiernos no han escatimado esfuerzos en realizar diferentes estudios de prevención y control, dadas las más recientes afectaciones a la salud y su economía, afectando el desarrollo en el transcurso de este 2020. Por esta razón se hace necesario estudiar el desarrollo de la epidemia entre muchas otras, con el fin de adelantar acciones de control e identificar patrones de riesgo.

En este trabajo se plantea una estrategia novedosa para estudiar el comportamiento temporal de una epidemia, permitiendo hacer una caracterización en el tiempo de desarrollo de la epidemia bajo unos supuestos obtenidos empíricamente de los resultados internacionales de propagación. Con la alternativa planteada se facilita la interpretación de los resultados, puesto que se hace una estimación del comportamiento global de las distintas realizaciones de la propagación de la epidemia.

La metodología del análisis de datos funcionales ha sido utilizada en la actualidad en un sin número de estudios, como se puede ver en metrología y en análisis de series temporales, entre otros. En el estudio de epidemias, éstas se pueden tratar como curvas o funciones en el tiempo facilitando el alcance a la hora de realizar análisis con grande volumen de información, caracterizando patrones y fluctuaciones.

Las realizaciones de la epidemia se presentan de manera discreta en el tiempo, por lo cual, es necesario transformarlos en curvas a través de los B-splines que permiten suavizar la información y representarla por curvas, con el fin de aplicar la metodología funcional.

El trabajo está organizado de la siguiente manera: en la primera parte se realiza una revisión de conceptos sobre epidemiología, su comportamiento, división y características fundamentales para su estudio, la importancia de su modelamiento y de algunos métodos matemáticos (SIR) y estadísticos (Bioestadística) empleados para este fin. En el segundo capítulo (capítulo 2) se presenta un marco teórico referente a datos funcionales y funciones B-splines. En el tercer capítulo se detalla la información estadística obtenida mediante las realizaciones del modelo SIR, se realiza un análisis descriptivo de las variables en estudio y por último se aplica la metodología del análisis de datos funcionales. Finalmente se incluye una sección de conclusiones y posible trabajo futuro.

Capítulo 1

1 La epidemiología

1.1 Epidemiología

La Organización Mundial de la Salud (OMS) define la epidemiología como "*El estudio de la distribución y los determinantes de estados o eventos (en particular de enfermedades) relacionados con la salud y la aplicación de esos estudios al control de enfermedades y otros problemas de salud*" [1]. Además, también puede ser entendida como una ciencia (rama de la medicina) que estudia la aparición de enfermedades y características de salud que se pueden presentar tanto en la población humana como en población de animales; con esto, se busca analizar la relación causa-efecto entre la exposición y la enfermedad.

El interés principal de la epidemiología es conocer características de los grupos afectados; la distribución geográfica, la frecuencia en que se presenta la enfermedad y las posibles causas o factores que la generan. En epidemiología, la información de las condiciones externas e internas del área, de la población y el período de tiempo que se desea estudiar son imprescindibles para analizar, explicar, modelar, modificar, prevenir y/o frenar el fenómeno epidemiológico que se pueda presentar.

Además, obtener dicha información permite conocer la distribución de enfermedades infecciosas en la población y los posibles factores que influyen en su desarrollo. Es importante mencionar que tales enfermedades, no tienen el mismo comportamiento ni la misma frecuencia. Cierta enfermedad puede variar en el mismo individuo, así como en el área en que se encuentra. La epidemiología busca explicar los factores determinantes de las enfermedades infecciosas y las diferencias que existen entre los individuos, permitiendo predecir el comportamiento de la enfermedad tanto a nivel general como individual.

Este tipo de análisis se puede realizar mediante las series temporales, caracterizando el comportamiento de la enfermedad. En esta, se evidencian los períodos en que se presenta mayor incidencia y las posibles variaciones estacionales; además, si las condiciones iniciales de la patología se mantienen, se puede realizar una proyección de su comportamiento.

Según Colimon (1990), la epidemiología se usa principalmente para "*Medir la naturaleza y magnitud de los problemas causados por las enfermedades en la comunidad, lo mismo que la variación de la patología según tiempo y lugar*" [2]. El estudio de la epidemiología se puede clasificar en: observacionales y experimentales, los cuales serán explicados y desglosados en el capítulo **1.5**.

1.2 Formas de propagación de una enfermedad

La OMS define la enfermedad como "*Alteración o desviación del estado fisiológico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por síntomas y signos característicos, y cuya evolución es más o menos previsible*" (como se citó en Herrero, 2016) [3]. A lo largo de la historia, el concepto de "enfermedad" ha estado en constante cambio y aún no hay una definición única, ya que abarca un

sinfín de significados, desde aspectos teológicos hasta aspectos científicos. Un claro ejemplo se evidencia en las civilizaciones primitivas, las cuales creían que las enfermedades eran maldiciones, hechizos y/o castigos de dioses o demonios. En este sentido, nace la idea de hechiceros o chamanes quiénes eran los encargados de curar los diversos tipos de enfermedades o "maldiciones" de la época.

Otra de las acepciones del concepto "enfermedad" se puede entender como una alteración leve o severa del funcionamiento del organismo de un ser vivo por causas internas o externas. Es importante decir, que la salud de un ser vivo se puede ver afectada por: aspectos ambientales, interacción con seres humanos o animales, entre otros factores. Las enfermedades se clasifican según su duración (agudas, subagudas y crónicas); según su distribución (esporádica, endemia, epidemia y pandemia); y según su causa (endógenas, genéticas, degenerativas, autoinmunes, mentales, entre otras).

Este trabajo de investigación se enfoca en caracterizar el comportamiento de la distribución de alguna enfermedad, es decir, cómo se distribuye a nivel geográfico. La clasificación se da de la siguiente forma: **Esporádica**: este tipo de enfermedad es poco usual, afecta a una cantidad mínima de la población y se presenta de forma ocasional. Un ejemplo de este tipo de enfermedad es el accidente cerebrovascular conocido como ACV, por su siglas. **Endemia:** se define cuando una enfermedad tiene una incidencia normal, común o esperada dentro de un área geográfica o población. Un ejemplo claro de este tipo de enfermedad es la fiebre amarilla. **Epidemia:** es el incremento significativo de una enfermedad en una población específica o en un tiempo determinado, un ejemplo de esta es la malaria. **Pandemias:** según la OMS

(2010), es “la propagación mundial de una nueva enfermedad” [4], un ejemplo de esta es la peste negra.

1.3 Epidemias y pandemias más letales

La historia de la humanidad se ha visto afectada por enfermedades que han tenido un gran impacto, tanto por la evolución de las mismas, como por la mortandad que han generado. La mayoría de las enfermedades se han transmitido por medio de animales, consecuencia de la domesticación o la ingesta de los mismos, entre otras causas.

Como se sabe, según la OMS (2010) una pandemia “es la propagación de una enfermedad a gran escala” [4] y se da en diferentes zonas geográficas; la mayoría de las personas afectadas por las pandemias no son inmunes a ella, por ende pueden llegar a ser letales dependiendo de su gravedad.

Las pandemias que han marcado, en mayor medida, la historia de la humanidad han sido: “*La peste negra”*, se dio en el siglo XIV en Europa, aproximadamente un tercio de la población europea murió a causa de esta bacteria. *“La viruela”* también fue una de las pandemias más graves, se dio en el siglo XVIII y ha causado muchas muertes. Aunque esta enfermedad fue controlada y actualmente hay vacunas y medicamentos con los que se trata, aún no posee cura. En añadidura, se encuentra “*La Gripe Española”* o más conocida como “Gripe de 1918”, fue una de las pandemias más letales y mortíferas de la historia, con un número aproximado de muertes desde los 20 hasta los 200 millones de personas [5], “*la mayoría de los enfermos en horas de la mañana presentaban los primeros síntomas y en la tarde ya habían muerto*” [6]. A principios del año 2020, se presenta una epidemia la cual avanza rápidamente conocida como el “Coronavirus” o en su nombre científico “COVID-19”, es una

enfermedad que se transfiere de animales a humanos como la mayoría de las enfermedades; se conoce que ha afectado a gran parte de la población de China, Italia, España y Estados Unidos y se ha expandido a la mayoría de países alrededor del mundo.

1.4 Variable epidemiológica

Las variables epidemiológicas permiten analizar, describir y determinar un problema que esté afectando la salud de una población. Es importante estudiar estas variables para poder conocer el origen de una enfermedad y saber qué poblaciones e individuos se encuentran expuestos al riesgo. En este sentido, se debe generar una hipótesis, una predicción y por último, analizar el comportamiento de la patología.

La causalidad se define como el estudio de la relación entre el elemento que ayuda al desarrollo de una enfermedad y la exposición. Las fuentes de variabilidad epidemiológica están dadas por la fuente biológica y la fuente que se da por error de muestreo. Entre las biológicas se encuentran el factor de riesgo, el efecto, factores asociados y de confusión y las características y atributos de persona, de tiempo y lugar [6]; y en las del error de muestreo se deben al investigador, al sujeto investigado y al instrumento de medición.

Según la OMS un factor de riesgo es definido como "*cualquier rasgo, característica o exposición de un individuo que aumente su probabilidad de sufrir una enfermedad o lesión*" [7]; también son conocidos como eventos o fenómenos a los cuales está expuesto según el ambiente en el que se encuentre el individuo, esto puede provocar un efecto o enfermedad. Los factores de riesgo pueden ser internos y externos, el papel que tienen éstos permiten la explicación del por qué algunos individuos expuestos a un factor de riesgo desarrollan una enfermedad mientras que otros

expuestos al mismo factor no la desarrollan. Así como el factor de riesgo, también existe un indicador de riesgo, el cual permite conocer si la presencia de la enfermedad es tardía o temprana.

La enfermedad o también conocida como "efecto" se presenta por un factor de riesgo preexistente, ésta puede variar dependiendo de su ubicación geográfica, condiciones climáticas y otros factores que se puedan generar en el ambiente. Es por esto que "*es importante tener en cuenta la distribución de frecuencia de una enfermedad, tanto su incidencia como su prevalencia*" [8].

1.5 Tipos de estudios epidemiológicos

Los estudios epidemiológicos permiten al investigador profundizar y obtener información, con el fin de encontrar posibles causas que puedan generar o propagar una enfermedad determinada; estos pueden ser de carácter experimental y observacional. La diferencia entre ambos es: el experimental permite que el investigador haga intervención en el objeto de estudio; en el observacional, el investigador realiza seguimiento sin necesidad de intervenir.

El estudio experimental se divide en las siguientes categorías: ensayo clínico, en donde la población de estudio son los enfermos. Ensayo de campo preventivo, cuya población de estudio son los pacientes sanos. Ensayo comunitario de intervención, el cual se realiza sobre poblaciones. Para el caso de estudios observacionales se tiene: cuando el objeto observado es expuesto a una enfermedad, este estudio es de cohortes; en los estudios de casos y controles, se seleccionan dos grupos, el primero está compuesto por sujetos que poseen una enfermedad y el segundo, por sujetos que no la poseen; allí se investiga si estuvieron o no expuesto a un factor de riesgo.

Existe otro tipo de estudio que busca estimar la enfermedad y la exposición simultáneamente, éste se conoce como transversal [8]; por último, siguiendo [9] se realiza una muestra, la cual permite estudiar exposiciones en una zona geográfica.

Para realizar un estudio epidemiológico es importante tener clara la pregunta a investigar, esto con el fin de plantear un sistema de hipótesis preciso para poder llegar a conclusiones particulares. En este caso, conocer la relación que existe entre el factor de riesgo y el efecto o enfermedad. Probar la hipótesis permite, por medio de la muestra analizada, sacar conclusiones y realizar afirmaciones.

1.6 Modelos epidemiológicos

Los modelos epidemiológicos buscan describir, explicar y predecir el comportamiento de una enfermedad. Los individuos a estudiar pueden presentar varios estados o categorías las cuales son: los susceptibles (S), infectados (I) o removidos (R); según [10], los modelos epidemiológicos más importantes son: *SI*, *SIS*, *SIR* y *SIRS*; estos se pueden modelar de forma estocástica o determinista.

Los modelos matemáticos que se implementan en la epidemiología son herramientas indispensables para el estudio de las enfermedades, mostrando propiedades de dispersión de una enfermedad y prediciendo su comportamiento de manera exacta; controlando los factores que intervienen en el proceso. Sin embargo, tienen diversas limitaciones. Una de ellas es que en estos modelos no se pueden incluir todos los factores que intervienen en la propagación de la enfermedad [11], debido a que puede ser más complejo. Adicionalmente, existen factores imposibles de controlar como las variaciones climatológicas. La modelación epidemiológica es diferente para cada caso, es importante tener en cuenta las características de la enfermedad como: propagación, inmunidad, transmisión, entre otras.

La modelación que se utilizará en este documento será el modelo SIR, éste es simple pero que abarca muchas características representativas de un brote epidémico. Las ecuaciones del modelo se representan de la siguiente manera:

$$S' = -\beta SI,$$
$$I' = (\beta S - \nu)I,$$
$$R' = \nu I,$$

donde, β es la probabilidad de pasar de susceptible a infectado, ν es la probabilidad de pasar de infectado a recuperado; los parámetros expuestos serán positivos. Para mayor información consultar en [21].

Existen valores críticos de gran importancia, ya que al conocerlos se puede tener control de una enfermedad, epidemia y/o pandemia; estos valores son conocidos como *número reproductivo básico*, *número de contactos* y *número de reemplazo*; éstos deben salir del límite para que ocurra un brote epidémico o una enfermedad.

Definición 1 (Tomada de [11, 12]): El número reproductivo básico, denotado como R_0 es definido como:

$$R_0 = \frac{\beta}{\mu} S'(0).$$

Es decir, corresponde al número promedio de casos secundarios producidos por un individuo infectado en una población susceptible.

El número reproductivo básico (R_0) "*es un parámetro teórico que proporciona cierta información acerca de la velocidad con que una enfermedad puede propagarse en una población determinada*" Éste nos indica si una enfermedad puede afectar a cierta población. De esta manera tenemos:

$R_0 > 1,$ *La transmisión de la enfermedad se dará en varios individuos.*

$$R_0 < 1, \qquad la\ enfermedad\ se\ extingue.$$

Es definido como el número promedio de casos secundarios producidos por un individuo infectado en una población susceptible.

Definición 2 (Tomada de [11]): El número de contactos (σ) es el número promedio de contactos suficiente para la transmisión de un virus durante el período infeccioso. Esto quiere decir que el contacto se dará entre un individuo infectado y uno susceptible [21].

$$\sigma = \frac{\beta}{\mu}$$

Definición 3 (Tomada de [11]): El número de reemplazos (R) se denomina como "*número promedio de casos secundarios producidos por un individuo infeccioso durante todo el período infeccioso"*. Es posible obtener R por medio del número de contactos (σ) por el número de individuos susceptibles en un tiempo t.

$$R = \sigma S'(t)$$

Capítulo 2

2 Marco Teórico

2.1 Análisis de Datos Funcionales (ADF)

Este capítulo abordaremos el Análisis de Datos Funcionales (ADF). Se desarrolló en 1905 por Georg Hamel (Hamel 1905) y Giuseppe Peano (Peano 1915) desarrollado en los libros de Ramsay y Silverman (1997).

El ADF es la parte de la estadística que se dedica a trabajar con muestras de funciones aleatorias. Éstas, van a permitir simplificar la información colectada y observar las principales características de los datos.

Con el ADF, es posible manipular funciones por medio de métodos estadísticos ya conocidos; es importante aclarar que las funciones que se trabajen deben ser suaves, es decir, se trabajan con datos menos distorsionados en comparación con el comportamiento real de los datos. En caso de que no lo sean, éstas deberán ser sometidas a un suavizamiento para luego realizar el tratamiento de datos con ADF. En añadidura, permite identificar patrones dentro de un conjunto de datos.

Se pueden aplicar operaciones sobre los datos como derivadas, integrales u otras funciones, esto se hace con el fin de poder visualizar la información de manera sencilla. El ADF también explica el comportamiento de la(s) variable(s) respuesta;

estudia el sistema dinámico en el caso de las series temporales funcionales y estima la dimensión finita del problema.

La idea básica del Análisis de Datos Funcionales es considerar que cada sujeto es representado por una función. Los datos funcionales son generalmente observados y registrados discretamente.

A continuación, se darán unas definiciones formales adaptadas de [13] y [16]

Definición 1: Una variable aleatoria X es llamada variable funcional (v.f) si asume valores en un espacio de dimensión infinita (o espacio funcional). Una observación x de X se llama dato funcional.

Definición 2: El espacio funcional es un conjunto de funciones entre dos conjuntos fijos.

El espacio funcional hace referencia a todas las funciones que se pueden construir en un dominio real. Sin embargo, la base del espacio funcional es infinita. Al tomar una base de esas características realiza una expansión del espacio euclidiano al espacio de Hilbert

Los espacios funcionales son espacios de Hilbert y permiten representar de forma matemática la información.

Cuando se dice que una observación es un dato funcional, se refiere a que una función suave genera estos valores. La suavidad es un indicador fuerte para aplicar el ADF en lugar de utilizar otras técnicas estadísticas. Sin embargo, si la función no tiene el comportamiento deseado, se realiza un proceso de suavizado (interpolación) para que se pueda hacer uso del ADF.

El que la función tenga un carácter suave, no indica que sea una condición necesaria para la observación original $\boldsymbol{y} = (\boldsymbol{y_1}, \dots, \boldsymbol{y_n})$, ya que puede estar sujeta la observación a un ruido. Su notación es:

$$y_j = x(t_j) + \varepsilon_j,$$

donde x es la función suavizada y ε es el ruido.

Las funciones utilizadas en el ADF pueden ser representadas por combinaciones lineales de funciones base. Funciones ϕ_k conocidas, independientes entre sí y que satisfacen la propiedad de la combinación lineal de las funciones que se generan representan la función original de manera arbitraria. Las funciones originales, denotadas por x, son representadas por

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t),$$

Siendo K el número de funciones bases utilizadas, c es un vector de dimensión K que contiene los coeficientes c_k. El número de bases K permite evidenciar el grado de suavidad de la función resultante; cuanto mayor sea K, más suave será la función obtenida.

"*El objetivo esencial en el ADF es la obtención de una función suave con determinadas características semejantes encontradas en la función original. Sin embargo, no hay necesidad de obtener una equivalencia exacta entre los valores de la función*" [16].

Las funciones abordadas en ADF poseen una dimensión infinita, esto debido a que los valores *a priori* no se encuentran definidos anteriormente; se hace necesario que todos los valores de la función, para cada t, se conozcan, obteniendo así un conjunto

de valores infinito. Es importante mencionar, que el uso de un conjunto limitado de valores es suficiente para determinar la función más un error. Siendo así, para cada caso se tiene un valor ideal para la dimensión de cada función.

2.2 Datos discretos y funcionales

Conocer la forma funcional de x supone la existencia de una función x, fundamentada en los datos observados. Esto implica la posibilidad de evaluar x en cualquier punto t, y permite evaluar a su vez las derivadas $D^m(t)$ existentes en t. No obstante, es importante aclarar que el ADF aplica para datos continuos, siendo así, si los datos obtenidos son valores discretos, es necesario aplicar métodos de suavizamiento para poder evaluar $x(t)$ y $D^m(t)$ en cualquier valor de t.

Una función x es suave, de modo que un par de valores de datos adyacentes mediante segmentos de línea recta se encuentran vinculados y es poco probable que sean diferentes. Es importante aclarar que, si la propiedad de suavidad no se aplica, los datos no podrán ser tratados de manera funcional.

Por lo general, los datos discretos y_j, $j = 1,2, \dots, n$ se usan para estimar la función x y al mismo tiempo, varias de sus derivadas.

2.3 De datos reales a datos funcionales

Cuando se hace recolección de la información, generalmente los datos se toman a través del tiempo o el espacio y se observan de la siguiente manera (datos tomados de la librería fda de R, la base se llama pinch):

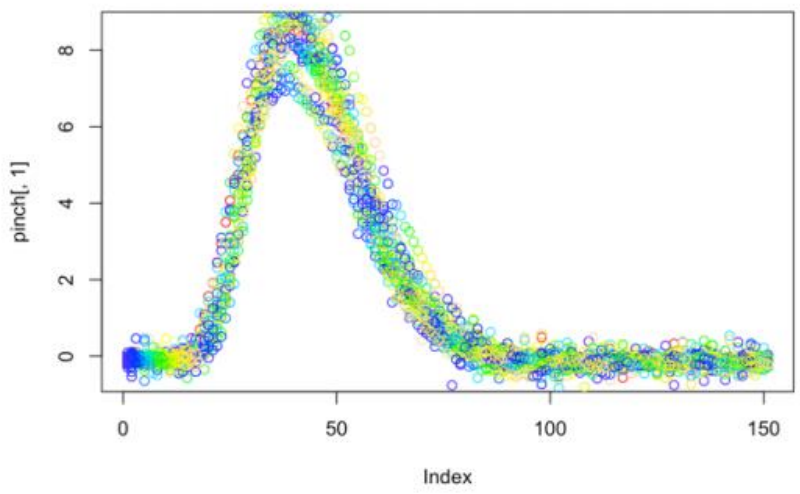

Figura 1. Datos de la base Pinch de R, forma original de los datos.

Como se observa en la Figura 1, los datos se obtienen en forma de puntos, y es necesario transformar la información en curvas; sin embargo, estas curvas deben tener ciertas características: dicha curva debe permitir evaluar el registro en cualquier punto; se debe poder evaluar la tasa de cambio; debe reducir el ruido y debe permitir que el registro en una escala de tiempo común.

Por tanto, se deben aplicar técnicas de suavizamiento con el fin de reducir la cancelación de la variación aleatoria que existe al recopilar la información. Para esto, se necesita una base que será explicada en siguiente sección.

2.4 Función base

Definición 3: [13] Una base es un conjunto de funciones conocidas e independientes, denotadas $\{\phi_k\}_{k\in\mathbb{N}}$, donde ϕ es una función conocida tal que $\forall\, k \in \mathbb{N};\ \phi_k \in L^2$, tales que cualquier función puede ser aproximada, tan bien como se quiera, mediante una combinación lineal de ellas con k suficientemente grande.

La base permitirá que los datos $x_n(t)$ tal que

$$\{\boldsymbol{x_n(t) = t \in [T_1, T_2]}, \qquad \boldsymbol{n = 1, 2, \dots, N}\}$$

puedan ser representados por medio de una función.

El intervalo de espacio o tiempo de estudio es $[T_1, T_2]$. N es el número de réplicas y $x_n(t)$ es la n-ésima replica de la realización en un tiempo t.

Se requiere un conjunto de bases de construcción funcional básicos que van a compilarse unos a otros para obtener así las características necesarias.

Así, se construye una función $\boldsymbol{X_n(t)}$ usando K bases; tal que:

$$\boldsymbol{X_n(t)} = a_1\phi_1(t) + a_2\phi_2(t) + \cdots + a_k\phi_k(t). \qquad (2.1)$$

La función construida es una combinación lineal y ésta va a representar el conjunto de datos. Las funciones bases pueden ser tanto finitas como infinitas y existen diversas bases:

- Constante: $\phi_k(t) = 1$
- Polinomios: $1,\ x,\ x^2, \dots$
- Potencias: $t^{\lambda_1}, t^{\lambda_2}, t^{\lambda_3}, \dots$
- Exponencial: $exp^{t\lambda_1}, exp^{t\lambda_2}, exp^{t\lambda_3}, \dots$
- Fourier: $1, \sin(wt), \cos(wt), \sin(2wt), \cos(2wt)$

Este trabajo se va a enforcar en un tipo de base polinomial llamado *B-splines*.

2.5 Función B-*splines*

Las funciones Spline son utilizadas con el propósito de aproximar una serie de datos no periódicos. Estas funciones son polinomios que se encuentran ajustados en un espacio o periodo T. La función Spline se encuentra determinada por el orden de segmentos polinomiales y la secuencia de los nodos τ, [17].

La información dada a continuación fue tomada de [18] y [20].

Para poder definir B-Splines, sea (a_i) una secuencia de nodos binfinitos y estrictamente crecientes por simplicidad, es decir $a_i < a_{i+1}$, para todo i.

Definición 4: Si $x \in X^Z$, entonces $x = \cdots x_{-1}x_0x_1 \ldots$ es una secuencia numerable bi-infinita, donde los índices $i < 0$ denotan el pasado de la secuencia, y los $i \geq 0$ el futuro, particularmente el índice $i = 0$ es el primer símbolo desconocido de la secuencia. Definición tomada de [19].

Se define los B-splines N_i^n con estos nodos por la fórmula de recursión:

$$N_i^0(u) = \{1, \quad si \quad u \in [a_i, a_{i+1}) \ 0, \ si \quad e.o.c$$

y

$$N_i^n(u) = \alpha_i^{n-1} N_i^{n-1}(u) + (1 - \alpha_{i+1}^{n-1}) N_{i+1}^{n-1}(u),$$

donde

$$\alpha_i^{n-1} = \frac{(u - a_i)}{(a_{i+n} - a_i)}$$

α_i^{n-1} es el parámetro local con respecto al soporte de N_i^{n-1}. Ahora, para el caso de nodos múltiples, los B-spline $N_i^n(u)$ están definidos por la misma fórmula de recursión y convención

$$N_i^{r-1} = \frac{N_i^{r-1}}{(a_{i+r} - a_i)} = 0 \qquad si \ a_i = a_{i+r}$$

Las siguientes propiedades de B-splines que son evidentes:

1. $N_i^n(u)$ *es un polinomio por partes y es de grado* n
2. $N_i^n(u)$ *es positivo en* (a_i, a_{i+n+1}),
3. $N_i^n(u)$ *es cero cuando está fuera del intervalo cerrado* $[a_i, a_{i+n+1}]$,
4. $N_i^n(u)$ *es continuo hacia la derecha.*

Observación 1: Si, en particular, $a_1 = a_2 = \cdots = a_n = 0$ y $a_{n+1} = \cdots = a_{2n} = 1$, entonces la fórmula de recursión anterior para $N_0^n, \ldots, N_n^n$ y $u \in [0,1)$ coincide con la fórmula de recursión de los polinomios de Bernstein. Por lo tanto, se tiene:

$$N_i^n(u) = B_i^n(u) \qquad \text{para} \quad i = 0, \ldots, n \quad \text{y} \quad u \in [0,1).$$

Como se mencionó anteriormente, los B-*Splines* están constituidos por piezas de polinomios que se encuentran unidas de una forma especial con ciertos valores llamados nodos. La ventaja que posee este tipo de información es que su computación es rápida y flexible.

A continuación, se muestra la construcción planteada en [20] y para mayor profundización consultar en los libros de Carl de Boor (1978) y [18].

Sea una partición o secuencia de nodos, lo que es, una secuencia no decreciente $t = (t_i)$ que tiene $N + 2$ valores reales conocidos como nodos, donde $N \geq 0$ se designan como "nodos interiores" y hay una existencia de dos puntos finales, t_0 y t_{N+1}. De modo que:

$$t_0 \leq t_1 \leq \cdots \leq t_{N+1}.$$

El conjunto de nodos aumentados se define:

$$t_{-k} = \cdots = t_0 \leq t_1 \leq \cdots \leq t_N \leq t_{N+1} = \cdots = t_{(N+1)+k} \ ,$$

Esto indica que los nodos límites t_0 y t_{N+1} se van a repetir k veces, k indicará el grado de la base B-spline. Se puede reajustar el índice i de los nodos, empezando de último t_{-k}, entonces, los $N + 2(k + 1)$ nodos aumentados t_i serán identificados por $i = 0, \ldots, N + 2k + 1$.

Para cada nodo t_i, se define de manera recursiva un conjunto de valores reales de las funciones $B_{ij}(t)$, $j = 0, \dots, k$, donde k es el grado de base B-spline, esto es descrito:

Para el caso de un B-spline de grado 0, la secuencia de nodos son las funciones:

$$B_{i0}(t) = X_i(t) = \{\ 1,\ \ si\ t_i \leq t < t_{i+1}\ 0, \qquad si\ \ e.o.c$$

Las funciones expuestas son continuas por derecha y cumplen con:

$$\sum_i \quad B_{i0}(t) = 1, \quad para\ todo\ t.$$

En particular,

$$t_i = \ t_{i+1} \qquad implica\ \ B_{i0}(t) = X_i(t) = \ 0.$$

Con lo anterior, se puede obtener el B-spline de un grado mayor mediante recurrencia:

$$B_{ik}(t) = w_{ik}(t)B_{i,k-1}(t) + (1 - \ w_{i+1,k}(t))B_{i+1,k-1}(t),$$

Donde,

$$w_{ik}(t) = \{\frac{t - t_i}{t_{i+k} - t_i} \qquad si\ t_i \neq t_{i+k}\ 0 \quad si\ e.o.c$$

Siendo así, el B-spline de primer orden se expresa:

$$B_{I1} = w_{i1}(t)X_i(t) + \Big(1 - w_{i+1,1}(t)\Big)X_{i+1}(t),$$

Aquí se puede evidenciar que son dos piezas lineales unidas para formar una pieza continua en un intervalo $[t_i, t_{i+2})$, por esta razón $B_{i1}(t)$ es conocido como B-spline lineal.

Después de $k-1$ pasos de recurrencia, se obtiene que la forma de $B_{ik}(t)$ es:

$$B_{ik}(t) = \sum_{j=i}^{i+k} b_{jk}(t) X_j(t),$$

Donde cada $b_{jk}(t)$ es un polinomio de grado k, entonces, el B-spline de grado k consiste en la unión de los nodos de polinomios de grado k, el cual es cero fuera del intervalo $[t_i, t_{i+k})$.

Una particularidad de $B_{ik}(t)$ es una función cero en el caso en que $t_i, = t_{i+k}$ ya que, si el primer nodo es igual al último, los demás serán iguales. Esta función también será positiva en el intervalo (t_i, t_{i+k}).

Una función B-spline de grado k es una curva paramétrica, está compuesta de una combinación lineal de bases *B-Spline* $B_{i,k}(t)$ de grado k, se obtiene de la siguiente manera:

$$B(t) = \sum_{i=0}^{N+k} c_i B_{i,k}(t), \quad t \in [t_0, t_{N+1}],$$

Donde c_i se conocen como puntos de control o puntos de Boor. Para el *B-spline* de grado k con N nodos interiores existen $M = N + K + 1$ puntos de control [20].

2.6 Suavizamiento

La primera fase del análisis de datos funcionales se basa en definir una función $X_n(t)$, en este caso (2.1), este proceso es conocido como suavizamiento de una función.

Lo que busca el suavizamiento es eliminar ruido en los datos con el fin de conservar la forma correcta. Sin embargo, al aplicarlo, es importante que

1. No puede ser muy suave

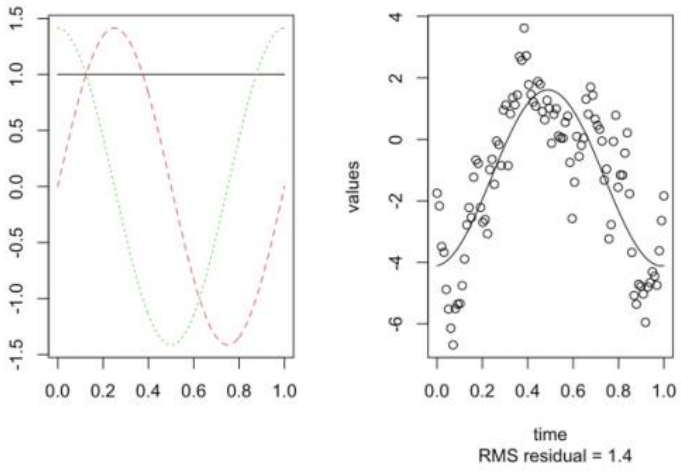

Figura 2. Función con gran suavidad.

Como se observa en la figura 2, la función es muy suave y cuando esto ocurre la curva no cubre por completo la información.

2. No puede ser muy estricta

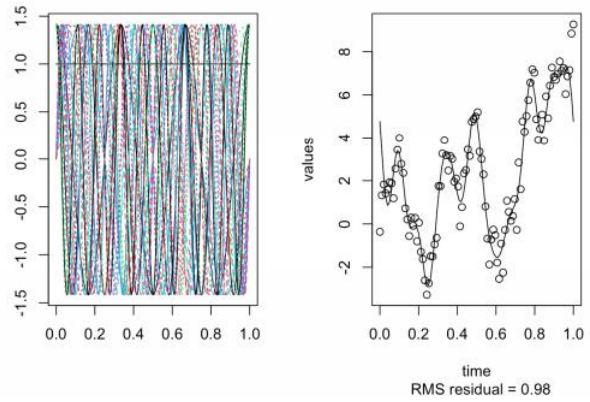

Figura 3. Suavizado estricto.

Si el suavizado realizado es muy estricto, como sucede en la figura 3, puede llegar a dificultar el análisis.

Lo que se desea con el suavizado es que la curva sea suficientemente suave y que recupere gran parte de la información como se observa en la figura 4.

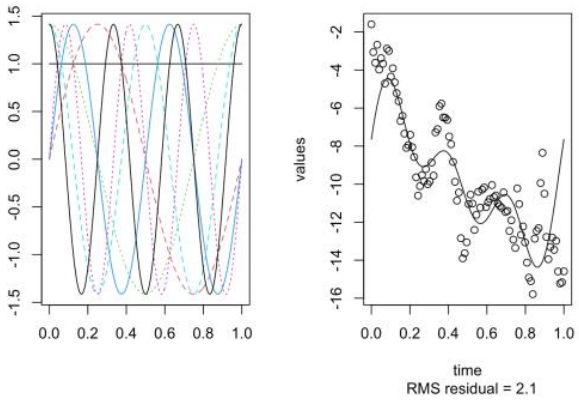

Figura 4. Suavizado estable.

2.7 Estadística descriptiva para datos funcionales

La teoría que se presenta a continuación fue adaptada de [16].

2.7.1 Media y varianza funcional

Las estadísticas descriptivas básicas conocidas para datos univariados son aplicadas de igual forma para los datos funcionales.

Definición 5: sean $x_1, x_2, \dots, x_n$ funciones definidas para una muestra de n datos funcionales (es decir, n funciones), la media funcional es dada por:

$$\underline{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t). \qquad (2.2)$$

Definición 6: la varianza funcional es definida como:

$$Var_x(t) = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i(t) - \underline{x}(t)\right)^2, \qquad (2.3)$$

En cuanto a la función de desviación estándar, se expresa como la raíz cuadrada de la varianza funcional.

2.7.2 Covarianza y correlación funcional

La covarianza funcional resume la dependencia de las observaciones de pares diferentes y se calcula para todo t_1 y t_2 a través de la siguiente fórmula:

$$Cov_x(t_1, t_2) = \frac{1}{n-1}\sum_{i=1}^{n} \left[x_i(t_1) - \underline{x}(t_1)\right]\left[x_i(t_2) - \underline{x}(t_2)\right]. \qquad (2.4)$$

La función de correlación asociada es dada por:

$$Corr_x(t_1, t_2) = \frac{Cov_x(t_1, t_2)}{\sqrt{Var_x(t_1)Var_x(t_{2)}}} \qquad (2.5)$$

De forma análoga para el análisis multivariado, se obtienen matrices de correlación, varianza y covarianza.

2.7.3 Covarianza y correlación cruzada funcionales

En general, cuando se observan funciones pares (x_i, y_i), la forma en que estás depende una de la otra, pueden cuantificarse mediante la función de covarianza cruzada:

$$Cov_{X,Y}(t_1, t_2) = \frac{1}{n-1}\sum_{i=1}^{n} \left[x_i(t_1) - \underline{x}(t_1)\right]\left[y_i(t_2) - \underline{y}(t_2)\right]. \qquad (2.6)$$

O por la función de correlación cruzada:

$$Corr_{X,Y}(t_1, t_2) = \frac{Cov_{X,Y}(t_1, t_2)}{\sqrt{Var_X(t_1)Var_Y(t_{2)}}} \qquad (2.7)$$

Capítulo 3

3 Aplicación de datos funcionales

3.1 Aplicación de datos funcionales a una epidemia

En esta sección se hace una breve descripción de los métodos y técnicas computacionales utilizados en la implementación de la simulación de la epidemia, en el análisis estadístico de datos funcionales (descritos en el marco teórico), partiendo de un desarrollo descriptivo básico de las variables de interés obtenidas en el modelo simulado.

Para dar inicio con la aplicación, se realizó la simulación de una epidemia con características similares al COVID-19. Para este trabajo, se tomó como base de la simulación el modelo SIR (Susceptibles-Infectados-Recuperados) [22].

Suponga que se tiene una superficie de área 21×21, cada una de sus intersecciones tiene un individuo el cual entra en interacción con sus compañeros, es decir, una grilla o *lattice* que representa el comportamiento de una epidemia; el tamaño de ésta es el área mencionada anteriormente, cuenta con 500 iteraciones, las cuales corresponden a 500 días transcurridos.

En principio, un individuo tendrá ocho vecinos, por tanto, si éste se encuentra infectado podrá propagar o no la enfermedad a cualquiera de estos, aquí es donde se definen los susceptibles, infectados y recuperados que definen el modelo SIR.

Para conocer el estado del individuo, se designaron números de la siguiente manera: el individuo susceptible se conoce como (-1); una vez éste pasa a ser infectado toma el valor de (14), y en cada iteración empieza a disminuir este número, haciendo la cuenta regresiva de los días que dura contagiado el sujeto. Por consiguiente, se sabe que el individuo está recuperado cuando éste toma el valor de (0).

El número reproductivo básico elegido para la simulación de la epidemia es $R_0 = 2.0$, este valor se tomó con el fin de obtener un efecto similar al del COVID-19. Diferentes artículos científicos indican que en esta enfermedad el R_0 se encuentra en el intervalo de $1{,}4 - 5{,}7$, si desea profundizar esta información consultar en [24, 25, 26, 27].

Partiendo que $R_0 = 2.0$, se puede decir que un infectado contagiará a 2.0 individuos en promedio, durante el periodo de contagio; la transmisión de la enfermedad se dará durante los 14 días que permanece infectado el sujeto; adicional, el individuo estará en contacto con 8 vecinos. Conociendo esta información, la probabilidad de contagiar a una persona en un día se va a expresar:

$$P(I \rightarrow S) = \frac{R_0}{D*8} = \frac{2.0}{14*8} = 0.018$$

Es decir, cuando un individuo infectado interactúa con otro sujeto, lo puede infectar con una probabilidad de 0.018 y así la probabilidad de no contagio es 0.98 lo cual supone que una persona susceptible no se infecta.

A continuación, se realiza un ejemplo que explica la mecánica de la simulación

	3	10	8	
	-1	12	-1	
	4	0	0	

Tabla 1. Ejemplo de grilla.

El programa ubica en la grilla a los $x > 0$, es decir, a los individuos infectados. Observe la Tabla 1 el dato encontrado 12 resaltado en la grilla, éste hace una interacción con los 8 vecinos; por tanto, los individuos que se encuentren infectados y/o recuperados no se tendrán en cuenta a la hora de la interacción. Cuando se halla un vecino susceptible — expresado como (-1) — éste se puede contagiar con una probabilidad de $P = 0.018$.

Se genera de forma uniforme un número aleatorio que toma los valores $0 < r < 1$, entonces, si $r < P$ el individuo se contagia. Otros valores observados en la Tabla 1 tales como (3, 10, 8, 4) hacen referencia a los días faltantes para la recuperación, que como se menciona anteriormente, el sujeto recuperado es representado con un 0.

Una vez terminado el proceso, es decir, $S = 0$ o $I = 0$, se guarda esta información y se genera de nuevo una epidemia, es decir, se realiza de nuevo este proceso simultáneamente, obteniendo diferentes escenarios de la epidemia. Cada uno de estos escenarios tiene una base de datos con la cantidad de susceptibles (S), infectados (I) y recuperados (R) por día, permitiendo así el desarrollo del estudio para aplicar la metodología de datos funcionales a los diferentes escenarios de la epidemia.

3.2 Análisis descriptivo de los datos

En esta sección se desarrolla un resumen descriptivo de los datos obtenidos después de las realizaciones de la epidemia, con el fin de conocer grosso modo el comportamiento de estos. En principio, se desea conocer el comportamiento en los diferentes escenarios obtenidos en las simulaciones realizadas según las variables; tiempo de duración, pico de infección y el total de infectados. Los datos mencionados anteriormente son de gran importancia en la realización de un estudio epidemiológico.

En la *Tabla 2* se muestran algunas estadísticas básicas de la variable Infectados (I) con respecto al tiempo de duración de la epidemia.

	Infectados (I)
Min	14
Máx	464
Mean	243
Desv	58

Tabla 2. Resumen estadístico del tiempo de duración de la epidemia.

Los resultados de la Tabla 2 indican que, de los diferentes escenarios de la epidemia, la infección en promedio dura 243 días (observe Figura 5) y su desviación es de 58 días; el periodo de infección más alto fue de 464 días y el de menor duración fue de 14 días, es decir, existió un contagiado, pero éste no propagó la enfermedad, por tanto, no se desarrolló una epidemia.

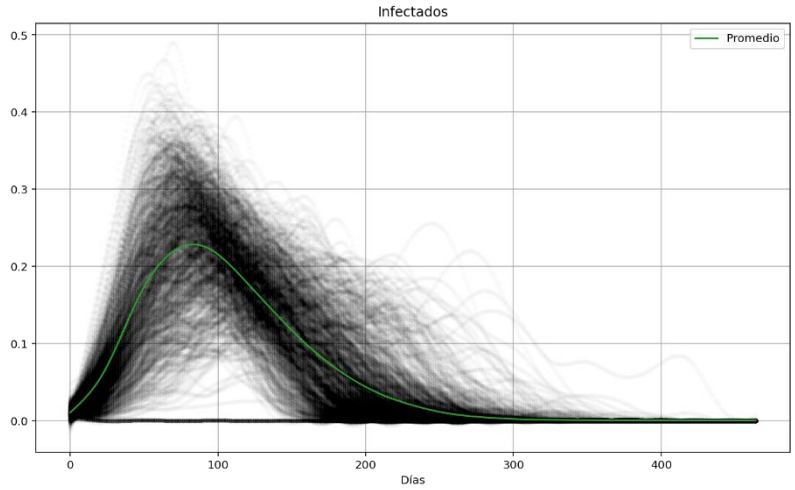

Figura 5. Curva media de los escenarios de la epidemia de Infección.

La Figura 6 permite visualizar de manera clara y sencilla el comportamiento de los diferentes escenarios de la epidemia, mostrando así donde se centra la mayor parte de información, el dato máximo y mínimo.

Es de suma importancia conocer la distribución de la variable a estudiar, las medidas como asimetría y curtosis permite estudiar las características de simetría y homogeneidad de la información; en este caso se obtuvo un coeficiente de asimetría negativo de -1.001, lo que nos indica que la distribución se encuentra sesgada a la derecha, es decir, los valores que se encuentran en la cola derecha están más alejados de la media.

Por su parte la curtosis nos indica que tan escarpada o achatada se encuentra la distribución; el valor obtenido de curtosis fue de 4.66 lo que denota que hay gran concentración de la información alrededor de la media, al obtener una $Curtosis > 3$ se dice que su distribución es leptocúrtica. En el gráfico, se puede confirmar lo dicho con antelación.

Según el Teorema de Chebyshev [27], al menos el 75% del número de contagios registrados se encuentran en el intervalo $[150 - 359]$ días, adicional, se encuentra un pico de infección que excede a 350 infectados.

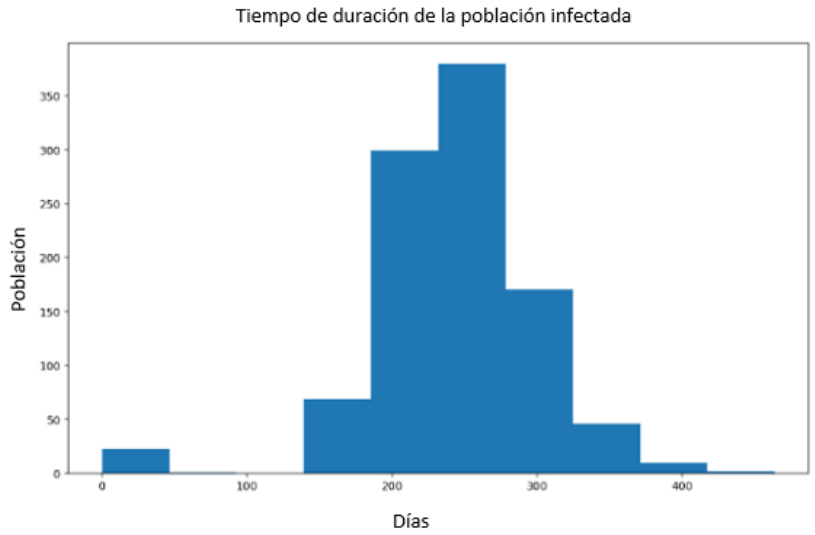

Figura 6. Histograma del tiempo de duración de la epidemia.

La Figura 7 deja en evidencia la presencia de datos atípicos los cuales pueden llegar a afectar los resultados, además de mostrar la información anteriormente expuesta. Cabe resaltar que, en este proceso de análisis descriptivo no se realiza el tratamiento los datos atípicos. Este proceso se desarrolla en el capítulo 3.2 siguiendo la metodología de análisis de datos funcionales.

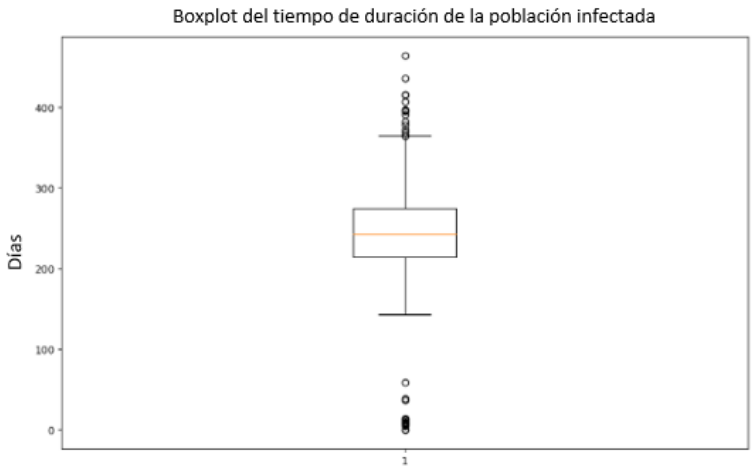

Figura 7. Boxplot del tiempo de duración de la infección.

En la *Tabla 3* se muestran los resultados estadísticos obtenidos de la variable Infectados (I) con respecto al pico de infección de la epidemia.

	Infectados (I)
Min	1
Máx	222
Mean	124
Desv	32

Tabla 3. Resumen estadístico del pico de infección de la epidemia.

Los resultados expuestos en la Tabla 3 muestran que, de los diferentes escenarios de la epidemia, el pico de infección llega en promedio con 124 individuos contagiados en un día y su desviación es de 32 individuos. El pico de infección máximo se dio con 222 sujetos contagiados en un día y el valor mínimo del pico de infección fue 1 individuo, este último no transmitió la infección.

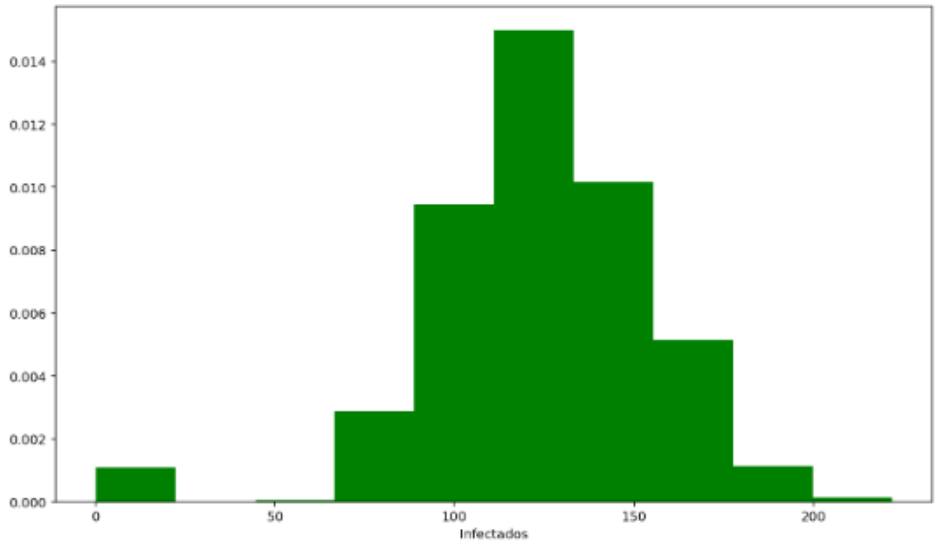

Figura 8. Histograma del pico de infección de la epidemia.

En la Figura 8 podemos observar que según [26], el pico de infección se encuentra en el intervalo de $[60 - 188]$ contagios registrados en un día, lo que representa aproximadamente el 75% de la información. La asimetría del gráfico es negativa, se puede evidenciar un leve sesgo hacia la izquierda, el valor obtenido del coeficiente de asimetría fue -0.98. Por otra parte, la curtosis dio un valor de 3.23, la distribución obtenida será leptocúrtica, como muestra el gráfico, la mayor parte de información se encuentra alrededor su media.

La Figura 9 representa información general y permite evidenciar los datos atípicos.

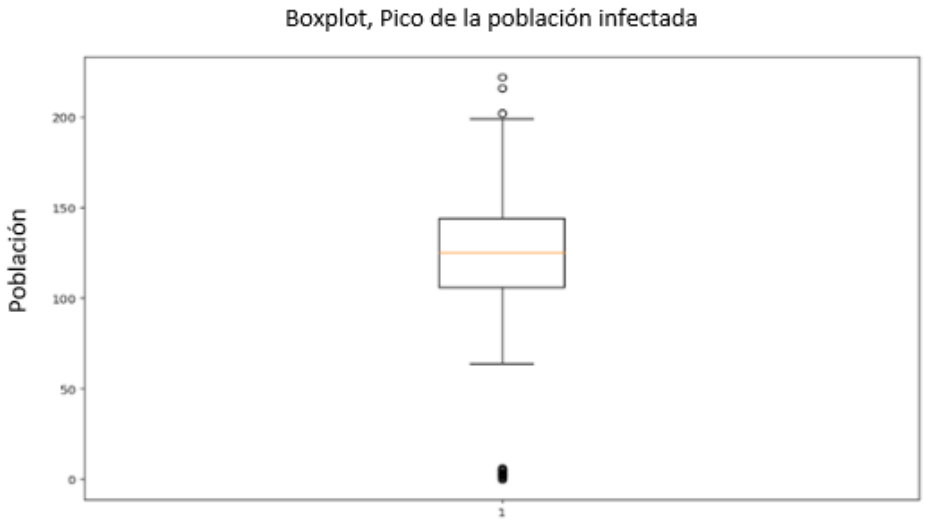

Figura 9. Boxplot del pico de la infección de la epidemia.

Por último, la Tabla 4 presenta el resumen estadístico del total de infectados que hubo durante la epidemia.

	Infectados (I)
Min	0
Máx	439
Mean	415
Desv	65

Tabla 4. Resumen estadístico del total de infectados.

Estos resultados arrojan que, de los diferentes escenarios de la epidemia, en promedio se infectaron 415 individuos y su respectiva desviación fue de 65 individuos. El mayor caso de infección fue de 439 sujetos y el menor caso de infección fue 0.

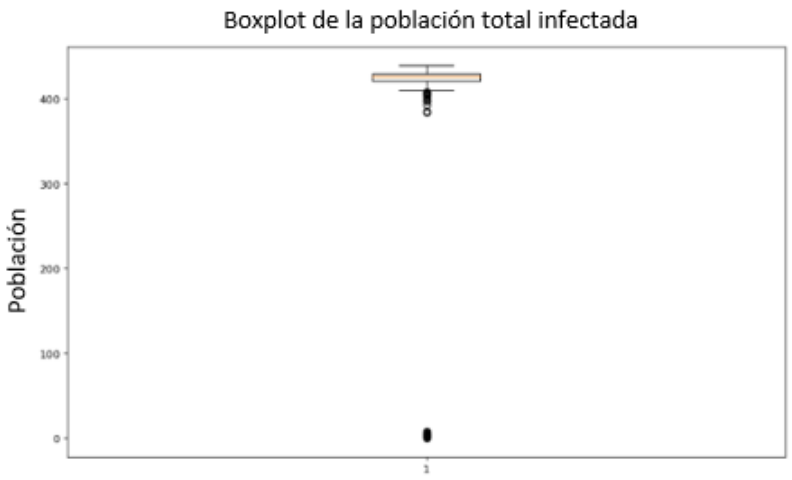

Figura 10. Boxplot de total de infectados.

La Figura 10 es la representación gráfica de lo expuesto, muestra los datos atípicos que pueden llegar a afectar los resultados, adicionalmente, se puede ver asimetría.

3.3 Análisis funcional

Al haber realizado los diferentes escenarios de la realización de una epidemia, los resultados obtenidos son valores discretos, lo que representan los valores de los Infectados (I), Recuperados (R), Susceptible (S), día a día, que es lo natural al realizar un estudio epidemiológico.

Para realizar el análisis de datos funcionales, por la metodología propuesta en el capítulo 2, es necesario expresar las realizaciones de la epidemia en funciones. A continuación, se muestra la nube de puntos de los diferentes escenarios de la epidemia.

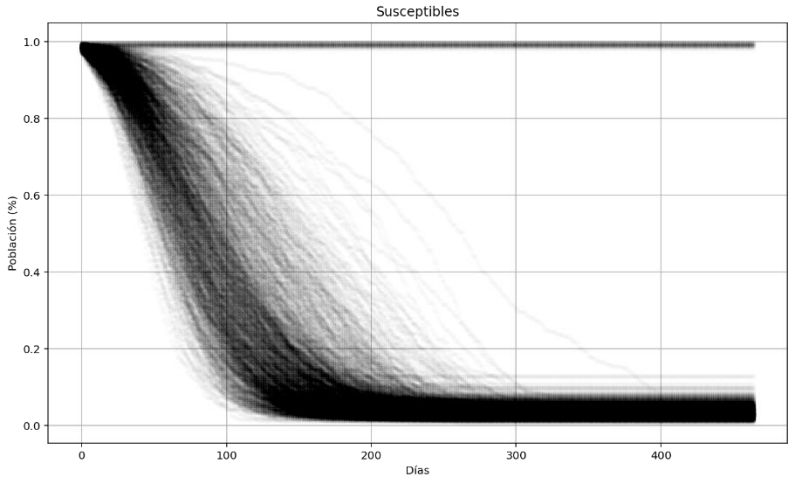

Figura 11. Datos originales porcentaje de susceptibles por día.

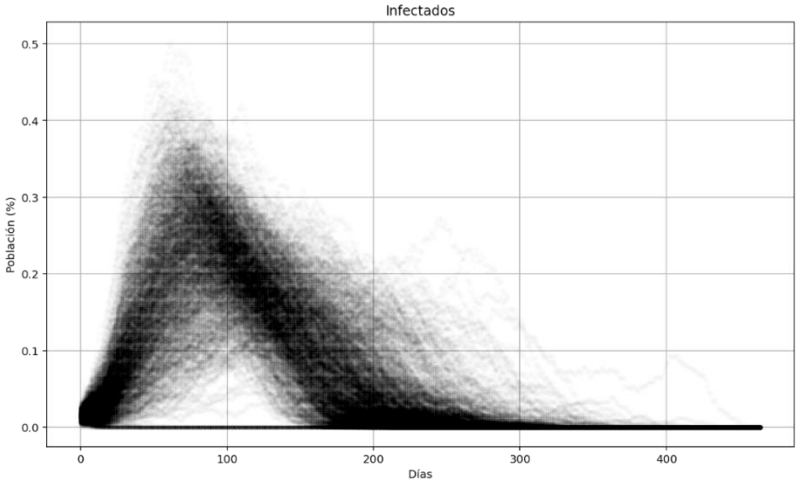

Figura 12. Datos originales del porcentaje de infectados por día.

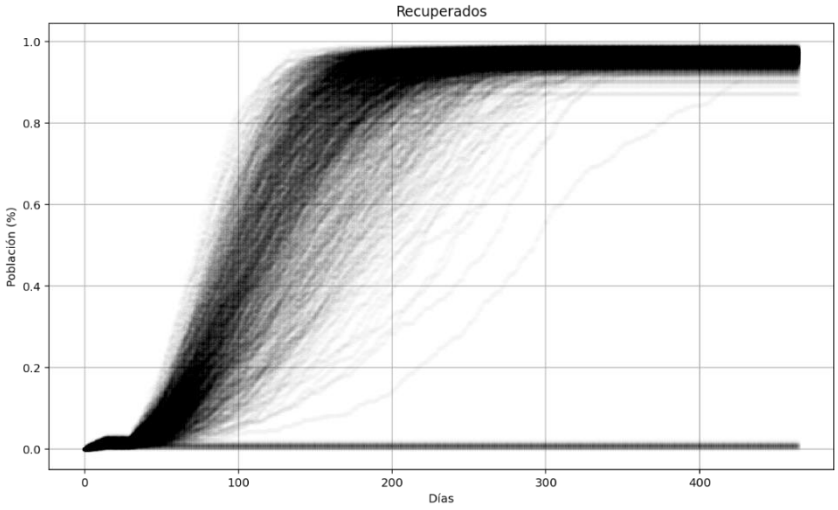

Figura 13. Datos originales del porcentaje de recuperados por día.

En las figuras 11, 12 y 13 se muestran los datos en su forma original, los cuales son discretos. Sin embargo, es importante transformarlos en curvas, con el fin de poder aplicar la metodología del análisis de datos funcionales.

3.4 Representación Funcional

Para poder iniciar la transformación de los datos se debe escoger una base adecuada que permita realizar el modelamiento de los mismos. Para este proceso se aplicó la metodología *B-spline*, con el fin de realizar el suavizamiento de los datos y así poder emplearlos como funciones. Es importante mencionar que esta metodología se utilizó por la misma naturaleza de las realizaciones, las cuales, no se presentan como funciones periódicas. A continuación, se mostrará los resultados obtenidos aplicando *B-splines* a los diferentes escenarios de la epidemia:

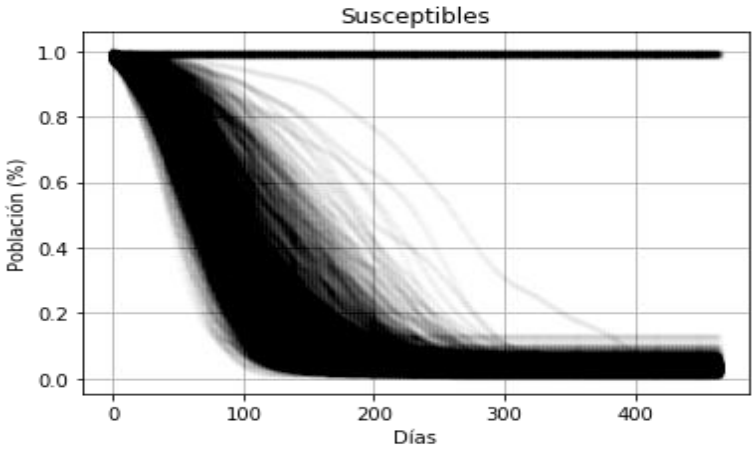

Figura 14. Aplicación de B-splines a porcentaje de susceptibles por día.

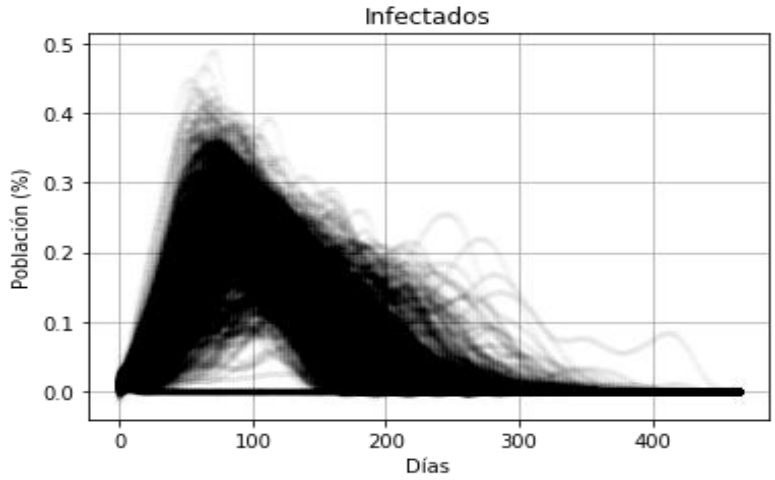

Figura 15. Aplicación de B-splines a porcentaje de infectados por día.

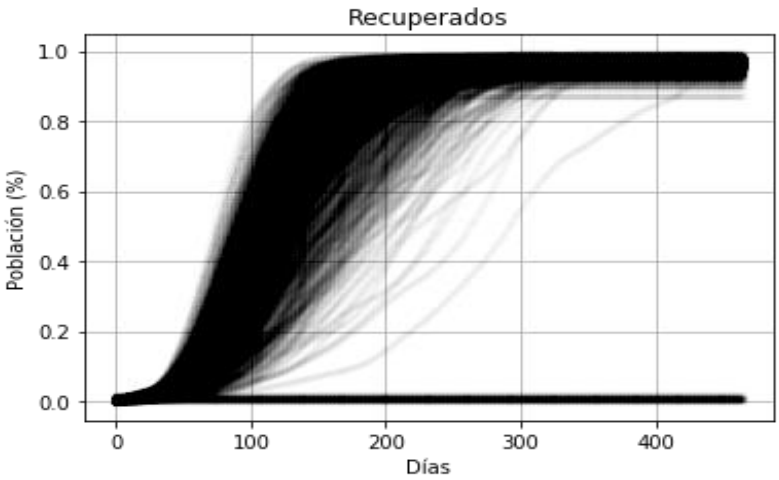

Figura 16. Aplicación de B-splines a porcentaje de recuperados por día.

Se puede observar que las Figuras (14, 15, 16) presentan suavidad, esto fue obtenido al aplicar la metodología B-splines, lo que era necesario en la realización del proceso de suavizamiento que permita realizar un tratamiento como funciones en el análisis de datos funcionales.

Los B-splines que se aplicaron a cada uno de los escenarios serán las funciones que se trabajarán como datos funcionales. X_t representa la realización de la epidemia en el instante t. Para la realización de la estadística descriptiva funcional, como se

explicó en el capítulo 2, cada observación es considerada como una curva o función, es por esto que, primero se calcula la curva media y curva de varianza para cada una de las variables estudiadas.

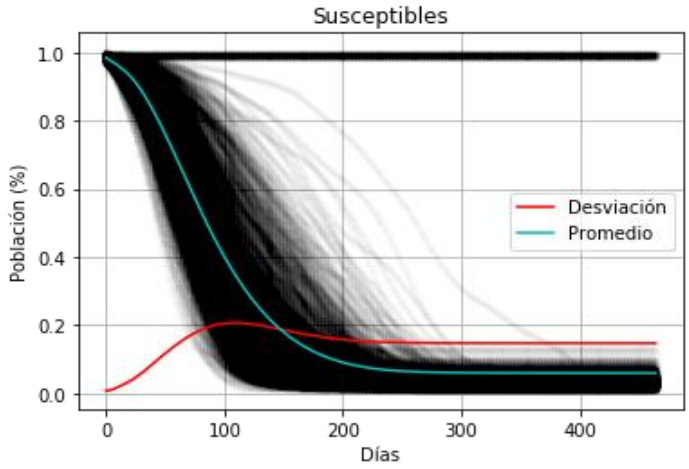

Figura 17. Curvas de desviación y media de la población de susceptibles.

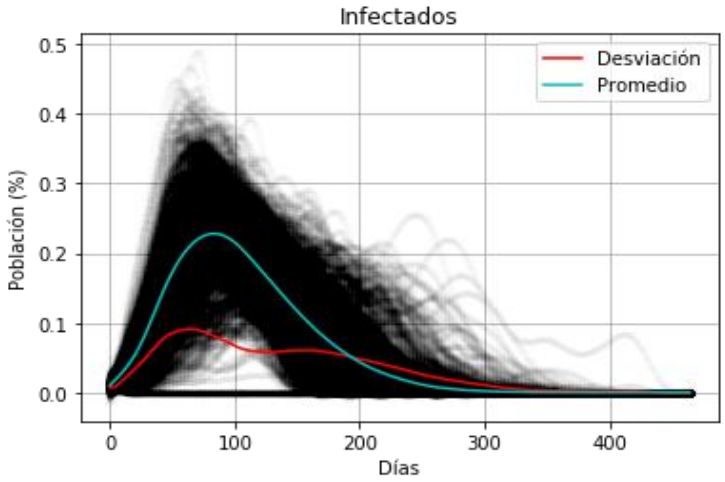

Figura 18. Curvas de desviación y media de la población de infectados.

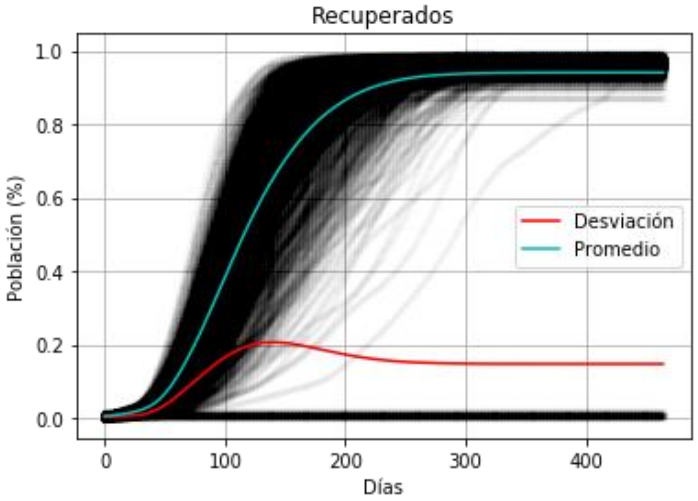

Figura 19. Curvas de desviación y media de la población de recuperados.

Las gráficas 17, 18 y 19 muestran el comportamiento que tiene la curva media y su respectiva curva de desviación en cada variable. La curva de color azul representa la media y caracteriza a la mayor cantidad de curvas de cada una de las variables en los diferentes escenarios de la epidemia. Sin embargo, es importante analizar que los datos atípicos para este caso son curvas y aunque no afectan la curva promedio, sí se puede observar su afectación en la curva de desviación. La desviación por su parte muestra qué tan dispersos se encuentran los datos con respecto a la media.

5. Resultados

Una buena metodología para realizar el análisis descriptivo de cualquier tipo de datos es por medio de la gráfica Boxplot. Para este estudio se realiza por medio del *boxplot* funcional, por el cual se puede realizar: análisis de los cuartiles, análisis de concentración de información y detección de datos atípicos y se muestran los diferentes *boxplot funcionales* obtenidos.

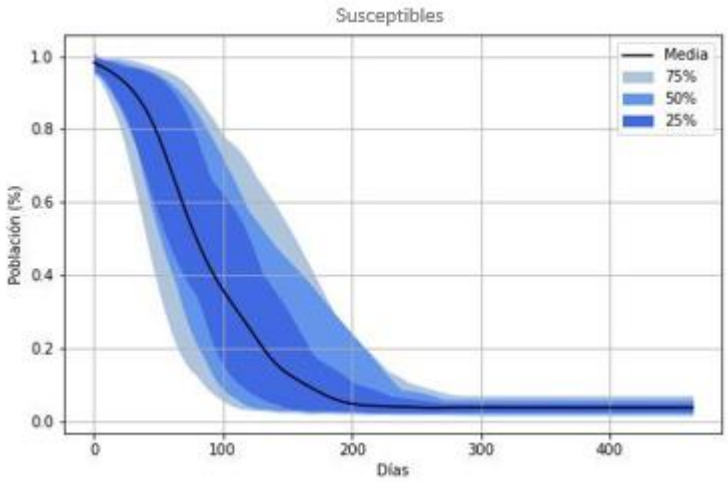

Figura 20. Boxplot funcional de la población de susceptibles.

Como se puede observar en la Figura 20, al inicio de la epidemia, el 100% de los sujetos son susceptibles. En el tercer cuartil, donde se encuentra el 75% de los escenarios de la epidemia, se observa que en menos de 100 días existe de un 10% a un 15% de la población infectada. Sigue descendiendo hasta el rango entre 0% al 20% en donde la mayor parte de susceptibles ha sido infectada en un lapso de 200 a 300 días, después de 300 días se observa que el 10% o menos de la población no sé infectó.

Para el 50% de la población susceptible, se puede evidenciar que el 20% aproximadamente de la población después de 200 días aún sigue siendo susceptible. Esto nos muestra que ya existe aproximadamente un 80% de la población contagiada.

Por otra parte, para la población perteneciente al cuartil uno, expresada en el 25% de la misma, se evidencia que existe entre el 15% y el 20% de la población aún susceptible, mostrando que en menos de 200 días hubo un contagio del 80% de los sujetos.

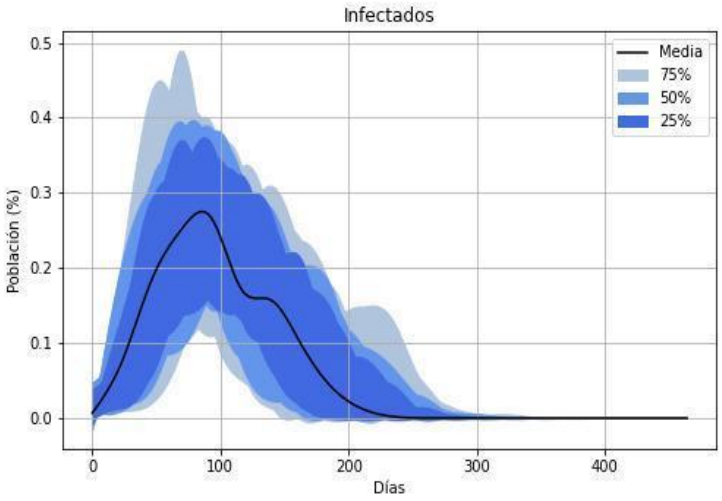

Figura 21. Boxplot funcional de población infectada.

La Figura 21 permite ver el porcentaje de contagios que se dio en los diferentes días. Se puede evidenciar que en el intervalo del 75%, es decir, el tercer cuartil tuvo un máximo de infectados de casi el 50% de la población en menos de 100 días; a su vez, se puede observar que también tuvo un mínimo de infectados que se encuentra entre el 10% y 20% de la población en menos de 250 días aproximadamente.

En el segundo cuartil, donde se encuentra el 50% de los escenarios de la epidemia, se puede evidenciar que su máximo de infectados se dio en el 40% de la población en menos de 100 días; el mínimo de infectados de éste encuentra entre el 0% y el 10% de la población en menos de 300 días.

En el primer cuartil se concentra el 25% de las epidemias, se puede observar en la figura 21 que existen dos picos de infección, el máximo de infectados en las dos

ocasiones se dio en el 40% de la población aproximadamente. No obstante, no llega a este valor. Los días en que tardó en llegar a esos picos máximos fue en un lapso menor a 100 días.

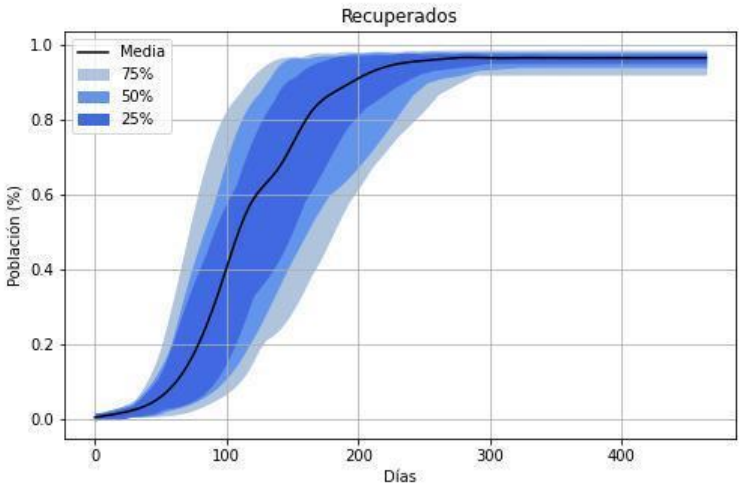

Figura 22. Boxplot funcional de población recuperada.

La figura 22, nos muestra la población de recuperados en los diferentes escenarios de la epidemia. Inicialmente, se observa que en todas las variables inicia con 0 recuperados. En el tercer cuartil, se puede evidenciar que un aproximado de 80% de la población en 100 días se recuperaron; el pico máximo en este cuartil se evidencia en el rango de 90% y 100% de la población, donde la recuperación de los sujetos se dio en menos de 200 días.

Para el segundo cuartil, el 80% de la población se recuperó en un lapso mayor a 100 días; su pico máximo se evidencia en un tiempo menor a 200 días, con un porcentaje entre el 90% y 100% de sujetos recuperados, en los diferentes escenarios de la epidemia.

Por último, en el primer cuartil que se encuentra el 25% de la población de estudio, para el día 100 se evidenció que aproximadamente el 60% de la población se recuperó.

Su pico máximo se encuentra en el rango de los 150 a 200 días con un porcentaje de casi el 90% o 100% de la población recuperada.

De esta manera, se puede concluir que la mayor parte de los escenarios simulados tienen un comportamiento similar. Sin embargo, es importante también mostrar la información atípica que se encontró.

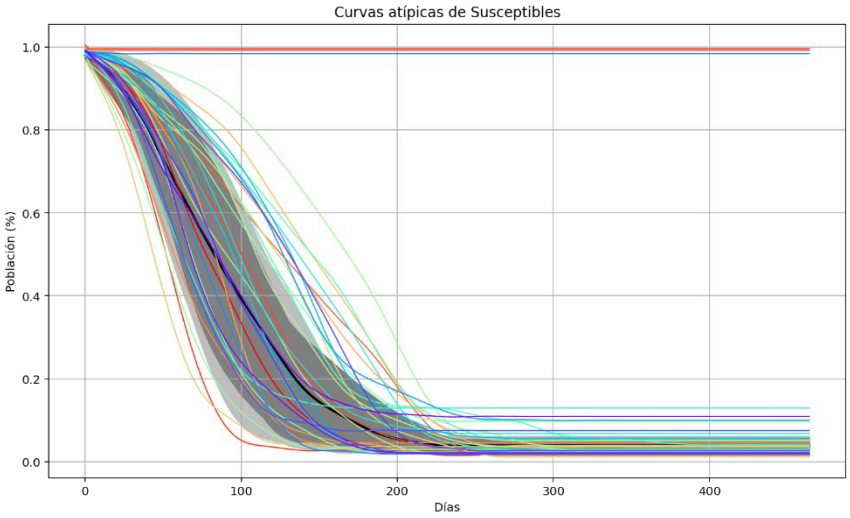

Figura 23. Boxplot funcional con curvas atípicas de la población susceptible.

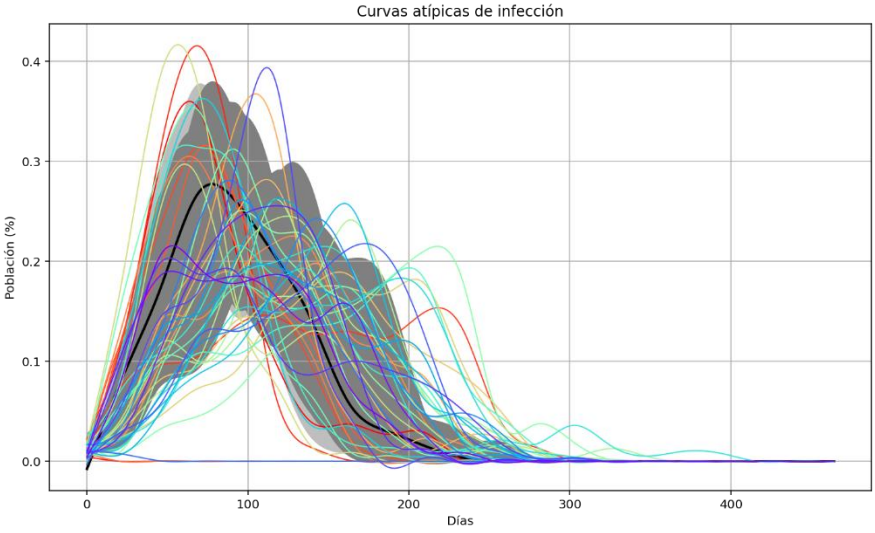

Figura 24. Boxplot funcional con curvas atípicas de la población infectada.

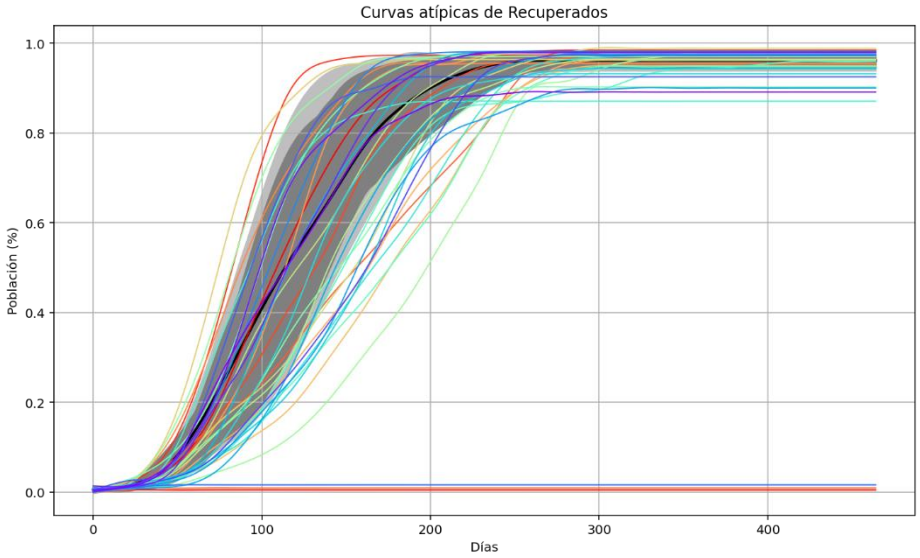

Figura 25. Boxplot funcional con curvas atípicas de la población recuperada.

Las figuras 23, 24 y 25, muestran los datos atípicos de las diferentes realizaciones de la epidemia. La zona sombreada en gris corresponde a los gráficos (20, 21 y 22), la curva negra a su media y las demás son las curvas atípicas, en algunos casos la epidemia no se desarrolla y es por esto que se observa en la gráfica una curva lineal éste es un caso atípico de los más significativos, pero no corresponden ni al 0.1% de las realizaciones.

6. Conclusiones

A lo largo de este trabajo, presentamos una introducción del Análisis de Datos Funcionales (ADF), que consiste en representar un conjunto de datos a través de funciones. El ADF es una alternativa para analizar datos definidos en el tiempo y que presentan características espaciales. Además, se explora un análisis poco convencional en estadística, como el uso de derivadas o integrales de funciones, ya que no es usual que los datos sean representados por funciones.

Debido a que todavía es un tema reciente, muchos aún desconocen el análisis de datos funcionales. Sin embargo, el ADF podría ganar gradualmente su lugar entre la comunidad estadística mundial. Esto se debe a la forma en que analiza los datos, modelando aleatoriedad a través de las características que poseen las funciones.

Las principales limitaciones de este trabajo se encontraron en la literatura. Básicamente, cuando se trata de publicaciones de libros, la literatura de ADF sigue siendo bastante escasa y puramente matemática, en muchos casos, no se presentan aplicaciones directas. Pero pronto, a medida que aumente la divulgación del ADF, también se espera que crezcan las aplicaciones que permitan facilitar el entendimiento. Otra limitante del abordaje llevado a cabo fue que no se realizó análisis inferencial, debido a la poca literatura encontrada. Entonces, como trabajo futuro, se propone hacer una extensión de esta metodología realizando estudios de estimación y predicción desde una perspectiva funcional aplicado a datos reales en Colombia.

7. Referencias

[1] World Health Organization, "Epidemiology". Disponible en: https://www.who.int/topics/epidemiology/es/ - OMS [Consultado el 22 de septiembre del 2019].

[2] Colimon, Martin-Kahl (1990), "Fundamentos de Epidemiología", Medellín, Colombia; Madrid: Díaz de Santos, S.A.

[3] Herrero, J. "Formalización del concepto de salud a través de la lógica: impacto del lenguaje formal en las ciencias de la salud". 2016. Vol. 10

[4]Organización Mundial de la Salúd (2010). *¿Qué es una pandemia?* Disponible en: https://www.who.int/csr/disease/swineflu/frequently_asked_questions/pandemic/es/. [Consultado el 22 de septiembre del 2019].

[5] Murillo-Godínez G. Recordando a la gripe española. Med Int Mex 20911;27(5):463-467.

[6] González BS. La pandemia olvidada de 1918. Revista de estudios Médico-Humanísticos 2005;14(14):123-127

[7] World Health Organization, "Risk Factors". Disponible en: https://www.who.int/topics/risk_factors/es/- OMS. [Consultado el 11 de febrero del 2020].

[8] Bellido, Juan. (2016); "Introducción a los estudios epidemiológicos, 29". Ed. 1; Valencia: Escola Valenciana d'Estudis de la Salut.

[9] Villa, Antonio; Moreno, Laura; García, Gaudalupe. (2011); "Epidemiología y estadística en salud pública"; México D.F.; McGraw-Hill Interamericana Editores, S.A. de C.V

[10] Montesinos-López OA, Hernández-Suárez CM. "Modelos matemáticos para enfermedades infecciosas". Salud Publica Mex 2007;49:218-226

[11] Pliego, Emilene C. (2011), "Modelos epidemiológicos de enfermedades virales infecciosas". Benemérita Universidad Autónoma de Puebla, Facultad de ciencias físico-matemáticas.

[12] Ridenhour B, Kowalik JM, Shay DK. "Unraveling R_0: Considerations for Public Health Applications". Am J Public Health. 2014;104:e32–e41. doi: 10.2105/AJPH.2013.301704.

[13] Ferraty, F. y Vieu, P. (2006). Nonparametric functional data analysis. Berlin: Springer-Verlag.

[14] Ramsay, J., *Functional Data Analysis*, Second Edition ed., John Wiley & Sons, NY, USA, 2006.

[15] Nagy, S., and Riesz, F., *Functional analysis*, New York **3** (1990), 6, 35.

[16] Frozza, Maicom, *Introdução à Análise de Dados Funcionais*, Porto Alegre (2010), 25-32.

[17] Ramsay, J. O. and Silverman, B. W. (2009). *Functional Data Analysis.* New York: Springer-Velarg.

[18] Hartmut, Prautzsch. Wolfgang, Boehm, Marco, Paluszny (2002). *Bézier and B-Spline Techniques.* Springer-Verlag. pp 61

[19] Tornero, Juan (2017), *Machine Learning: Modelos Ocultos de Markov (HMM) y Redes Neuronales Artificiales (ANN)*. Barcelona, Universitat de Barcelona.

[20] Plazola, Rodigo, (2013). *Monitoreo de Datos Funcionales*. Guanajuato, México: Centro de Investigación en Matemáticas, A.C. (CIMAT).

[21] Vargas, Paola. (2013). *Métodos matemáticos para el análisis de epidemias y propagación de gusanos en redes de sensores inalámbricos.* Bogotá, Colombia. Pontificia Universidad Javeriana.

[22] W. O. Kermack & A. G. McKendrick "A contribution to the mathematical theory of epidemics" Proceedings of the Royal Society of London Series A, 1927

[23] Wu, J.T., Leung, K., Bushman, M. et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat Med 26, 506–510 (2020). https://doi.org/10.1038/s41591-020-0822-7

[24] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K., Lau, E., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., … Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. The New England journal of medicine, 382(13), 1199–1207. https://doi.org/10.1056/NEJMoa2001316

[25] Riou, J., & Althaus, C. L. (2020). Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 25(4), 2000058. https://doi.org/10.2807/1560-7917.ES.2020.25.4.2000058

[26] Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerg Infect Dis. 2020 Jul [date cited]. https://doi.org/10.3201/eid2607.200282

[27] Saw, J. G., Yang, M. C., & Mo, T. C. (1984). Chebyshev inequality with estimated mean and variance. The American Statistician, 38(2), 130-132.