Universidad de Salamanca

Departamento de Estadística

Doctorado en Estadística Multivariante Aplicada

Tesis Doctoral

VNiVERSiDAD
D SALAMANCA

Análisis de Datos Funcionales (ADF) Aplicado a las Principales Causas de mortalidad en el Ecuador 1997-2021: Estudio Demográfico

AUTOR: *Juan Tarquino Calderón Cisneros*

DIRECTOR: *José Luis Vicente Villardón*

2021

Departamento de Estadística

José Luis Vicente Villardón

Profesor Titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICA:

Que **Don Juan Tarquino Calderón Cisneros** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo para optar al título del Doctorado en Estadística Multivariante Aplicada que presenta con el de **Análisis de Datos Funcionales (ADF) Aplicado a las Principales Causas de mortalidad en el Ecuador 1997-2021: Estudio Demográfico**, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca a 5 de octubre de 2021.

Director: José Luis Vicente Villardón

Agradecimientos

Nuestro agradecimiento al Vicerrectorado Académico de la Escuela Superior Politécnica del Litoral (ESPOL-Guayaquil), en la persona de la **Dra. Cecilia Paredes Verduga** por haber brindado su apoyo económico y logístico al **PROGRAMA DE FORMACIÓN DE TALENTOS ESTADÍSTICA MULTIVARIANTE APLICADA**, dirigido a la capacitación a nivel de Maestría y Doctorado tanto de profesores de la ESPOL, como otras instituciones del ámbito público y privado de la Provincia del Guayas.

A la **Dra. María Purificación Galindo-Villardón**, Directora del Departamento de Estadística de la Universidad de Salamanca (España) y Coordinador de los Programas de Análisis Avanzado de Datos Multivariantes y del Doctorado en Estadística Multivariante Aplicada en el momento del comienzo de esta tesis, por su valioso apoyo y dedicación en la formación de talentos estadística multivariante aplicada, además de todas sus orientaciones que nos permitieron culminar con éxito este proyecto académico emprendido.

A la **UNIVERSIDAD ESTATAL DE MILAGRO (UNEMI)**, **(Rector, Vicerrectorado Académico y Talento Humano)**, por el apoyo, respaldo y oportunidades de crecimiento profesional para lograr tan anhelado proyecto académico.

Dedicatoria

A **DIOS** por todas sus bendiciones

A **Luis Tarquino Calderón Inca** (†) y **Olga Cisneros Cordero**, mis padres, que por sus enseñanzas y sabidurías impartidas estoy cada día logrando mis sueños.

A mis bellos hijos que son la razón de que cada día tengo que salir adelante por ellos y para ellos.

A **Luis Arturo Calderón Cisneros**, mi Hermano, por su lucha constante y superación diaria.

A mi gran hermano de crianza y compadre **Carlos Esteban Alcívar Trejo**. Por su apoyo, enseñanzas y su aguante en este proceso académico.

A **José Luis Vicente-Villardón**, mi tutor y profesor del Departamento de Estadística de la Universidad de Salamanca, que con su paciencia y empeño logro guiarme en este trabajo de investigación.

Indice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción y objetivos

1.1. Introducción

Los estudios demográficos se han reformado absolutamente, a esto se debe por los problemas existentes en la actualidad, tales como crisis tanto sociales y definitivamente económicas. La disminución de la mortalidad ha estado muy relacionada con las mejoras en los sistemas de salud pública, condiciones de vida, que posibilitaron la reducción de la mortalidad y precipitan el comienzo del análisis demográfico, los patrones de salud y las "enfermedades degenerativas", donde la mortalidad presenta un comportamiento en aumento, entre estas encontramos las enfermedades cardiovasculares, cáncer, tumores, VIH y malformaciones.

La OMS señala que "las principales causas de mortalidad en el mundo son la cardiopatía isquémica y el accidente cerebrovascular, que ocasionaron 15,2 millones de defunciones en 2016 y han sido las principales causas de mortalidad durante los últimos 15 años. Sobre la enfermedad pulmonar obstructiva crónica (EPOC) causó tres millones de fallecimientos en 2016, mientras que el cáncer de pulmón, junto con los de tráquea y de bronquios, se llevó la vida de 1,7 millones de personas. La cifra de muertes por diabetes, que era inferior a un millón en 2000, alcanzó los 1,6 millones en 2016. Las muertes atribuibles a la demencia se duplicaron con creces entre 2000 y 2016, lo cual hizo que esta enfermedad se convierta en la quinta causa de muerte en el mundo en 2016" [77].

Las infecciones de las vías respiratorias inferiores continúan siendo la enfermedad

transmisible más letal; en 2016 causaron tres millones de defunciones en todo el mundo. La tasa de mortalidad por enfermedades diarreicas, que se redujo casi un millón entre 2000 y 2016, fue de 1,4 millones de muertes en 2016 [**?**].

Los cánceres causados por infecciones v´ıricas, tales como las causadas el por virus de las hepatitis B (VHB) y C (VHC) o por el virus del papiloma humano (PVH), son responsables de hasta un 20 % de las muertes por cáncer en los países de ingresos bajos y medios. Más del 60 % de los nuevos casos anuales totales del mundo se producen en África, Asia, América Central y Sudamérica. Estas regiones representan el 70 % de las muertes por cáncer en el mundo. El estudio de datos que son funciones o que pueden representarse mediante funciones ha adquirido nuevo interés recientemente.

Según cifras del censo del año 2010 por el Instituto Nacional de Estadísticas y Censos (INEC), Ecuador registra una población de 14 millones de habitantes, de los cuales el 66 % es población urbana. El crecimiento de la población se ha visto afectada por la reducción de la tasa bruta de natalidad de 32,4 a 11,4 nacimientos por 1000 habitantes entre 1981 y 2010, la disminución de la tasa de mortalidad de 6,7 muertes por 1000 habitantes en 1981 a 4,3 en 2008 y la tasa de mortalidad infantil en 2009 fue de 20 por 1000 nacidos vivos.

Aproximadamente un 30 % de las muertes por cáncer corresponden a cinco factores de riesgo conductuales y dietéticos: índice de masa corporal elevado, ingesta reducida de frutas y verduras, falta de actividad física, consumo de tabaco y consumo de alcohol. El consumo de tabaco es el factor de riesgo más importante, y es la causa más del 20 % de las muertes mundiales por cáncer en general, y alrededor del 70 % de las muertes mundiales son producidos por cáncer de pulmón.

También ha disminuido “el número de muertes por tuberculosis durante el mismo periodo, pero esta enfermedad continúa siendo una de las 10 principales causas de muerte, con 1,3 millones de fallecimientos. En cambio, la infección por el VIH/sida ya no figura entre las 10 primeras causas; para el año 2000, fallecieron 1,5 millones de personas por esta causa, para el año 2016 esta cifra se redujo hasta los 1,1 millones. Los accidentes de tránsito se cobraron 1,4 millones de vidas en 2016; alrededor de tres cuartas partes de las v´ıctimas (el 74 %) fueron varones” (OPS 2017).

Hay que destacar que estos umbrales “reflejan múltiples características intr´ınsecas

de un país y de una población, desde los comportamientos y contactos sociales hasta la propia estructura de la atención primaria y atención hospitalaria, por lo que umbrales epidemiológicos de disparo en un país concreto, no tienen necesariamente por qué ser aplicables en nuestro país. Por esta razón, desarrollar umbrales o niveles de alerta propios podría evitar que las autoridades sanitarias tomasen decisiones erróneas como, por un lado, confinamientos innecesarios (falso positivo) y, por otro, decidir no confinar cuando esta acción es necesaria (falso negativo)".

La mortalidad como concepto que expresa la acción de muerte, es una de las componentes demográficas básicas, determinantes del tamaño, de la composición por tipo de muerte, lugar de residencia, género, edad de una población determinada, se establece como el registro estadístico de hechos vitales, que "corresponden a los hechos de nacidos vivos, defunciones generales, defunciones fetales, ocurridos en el Ecuador. El INEC mediante el aprovechamiento de los registros administrativos de las diversas instituciones públicas, presenta a continuación los principales resultados de las estadísticas vitales de nacidos vivos y defunciones"(INEC, 2017).

Las realizaciones de la mortalidad se basan en la observación de las defunciones que ocurren en una poblacion durante un tiempo determinado y la evolución de las mismas. La baja originada en la mortalidad por medio de políticas públicas en el área de salud, las mejoras en las condiciones de vida, a pesar que la mortalidad es un hecho inevitable, su comportamiento presenta diferencias muy importantes entre países, regiones, clases sociales. (IARC, 2014).

Los datos de mortalidad constituyen uno de los elementos fundamentales para cuantificar los problemas de salud. A partir de ellos se han desarrollado numerosos métodos estadísticos para el análisis de este fenómeno. Desde los métodos clásicos a los del análisis multivariante pasando por los modelos de series temporales.

Este tipo de datos se presentan de manera discreta en el tiempo, por lo cual, es necesario transformarlos en curvas que permiten suavizar la información con el fin de aplicar la Metodología funcional [85]

El análisis convencional de datos multivariantes permite el estudio de observaciones que constituyen un conjunto finito de números; sin embargo, en los casos reales aparecen situaciones donde los datos que se estudian son procesos continuos [103]. En algunos contextos los datos deben ser tratados como funciones, en lugar de una

cadena de valores. Este tipo de datos, los llamados datos funcionales, aparecen en muchas áreas como meteorología, economía o medicina. El estudio de datos funcionales es un campo reciente en la investigación estadística, en desarrollo durante los últimos años. Para una revisión acerca de este tema se pueden ver las monografías en [89] o [29].

Actualmente, el análisis de datos funcionales "ha surgido como un área importante en la estadística y ha logrado avances significativos respecto de su base teórica. Es por esto que, en comparación con los datos tradicionales, que consisten en observaciones puntuales, los datos funcionales pueden contener información más detallada sobre el sistema subyacente que los genera, debido a que se toma en cuenta la curva y sus derivadas" [91].

Para el análisis de datos funcionales (FDA), donde se "proporcionan un enfoque de predicción y modelado relativamente novedoso para estimar las tendencias de incidencia, sin embargo, hasta la fecha, su aplicación a los datos epidemiológicos, de mortalidad y fertilidad ha sido limitada. Resulta interesante e importante "identificar los patrones o tendencias de estos indicadores epidemiológicos medidos en las áreas geográficas, grupos sociales, al determinar hallazgos y analizar estos patrones, siendo una herramienta que contribuye a determinar los puntos críticos y falencias a ser superadas; así como, un aporte significativo e importante en la Epidemiología y salud pública" ([61]; [113]).

El FDA, "es una rama de la estadística, donde la unidad básica de información es la función completa más que un conjunto de valores. En general, cualquier observación que var´ıe en un continuo se puede considerar como dato funcional" (Ramsay & Dalzell, 1991). Los avances tecnológicos "han hecho que el FDA se convierta en una disciplina emergente de la estadística con multitud de estudios y publicaciones en diferentes revistas de gran impacto. Las técnicas de FDA más utilizadas son el Análisis de Componentes Principales (ACPF), ANOVA y los modelos de regresión funcional, aunque en las últimas décadas se ha incursionado en técnicas de detección de atípicos, clúster, kriging, entre otros" [**?**].

El coronavirus denominado SARS-CoV-2 "causante de la enfermedad COVID-19 fue reportado por primera vez en diciembre de 2019 en la ciudad de Wuhan, en la P.R. China" [111] y "fue declarado pocos meses después pandemia por la Organización

Mundial de la Salud (OMS). Este virus para principios de mayo de 2020 ya había causado más de 4.000.000 de casos confirmados y más de 250.000 muertes a nivel mundial" [95].

La pandemia por SARS-CoV-2 "es el mayor desafío sanitario en los últimos 100 años, ocasionando el mayor exceso de mortalidad no bélico en este período en el mundo occidental. Ante una enfermedad de elevada contagiosidad y asintomática en un tercio de los casos, es fundamental disponer de modelos que predigan su evolución" [74].

Para finales del año 2019, "tras un fallo colosal de los mecanismos de detección, alarma y control de la enfermedad, agudizado por la falta de pruebas diagnósticas, comenzó la transmisión comunitaria en la mayor parte de los países, lo que obligó a tomar medidas excepcionales de salud pública, como el confinamiento forzoso de la mayor parte de la población para cortar las cadenas de transmisión del SARS-CoV-2" [95]. El confinamiento se aplicó a nivel mundial, "pero con una situación epidemiológica muy desigual entre ciertos países de América Latina. Para el caso de Ecuador en algunas provincias este confinamiento llegó tarde y no se pudo evitar el colapso de nuestro sistema de salud pública, mientras que, en otras, debido a la baja transmisión comunitaria, el SARS-CoV-2 causó un impacto en hospitalizaciones y fallecimientos mucho menor" [27].

El confinamiento "logró progresivamente el control de la enfermedad, pero también un indudable impacto económico. Disponer de umbrales epidemiológicos de fácil interpretación y que, con pocos datos, permitan predecir la evolución de la pandemia en áreas pequeñas (a nivel de provincia o incluso de áreas más pequeñas), podría permitir a las autoridades sanitarias actuar de forma más eficiente a través del establecimiento de medidas de salud pública más o menos drásticas ante aumentos de incidencia que conlleven transmisión comunitaria no controlada" [74].

En los FDA, los datos en lugar de ser un conjunto de vectores, como en análisis multivariante clásico, son un conjunto de curvas. En la mayoría de las aplicaciones, las curvas muestrales proceden de la observación de un proceso estocástico en tiempo continuo. El "gran desarrollo que está experimentando el análisis de datos funcionales ha ocasionado que muchos estudios con datos longitudinales, planteados desde un punto de vista multivariante, ahora sean analizados en base a su naturaleza funcional.

Las técnicas estadísticas de FDA se han desarrollado en los últimos 20 años como una generalización de las técnicas de análisis multivariante y de regresión al caso en el que las observaciones son curvas en lugar de vectores".

Hablar de cómo manejamos la información y cuál sería el tratamiento adecuado de los datos tiene mucho que ver con qué modelo estadístico pretendemos usar, que definición multivariante para poder discernir mejor dichos datos y definir el mejor modelo a explicar en cualquier ámbito de estudio, es bien cierto que desde hace mucho tiempo, los aspectos funcionales en las estadísticas se han investigado en una referencia del componente Probabilístico y Estadístico, tal como lo expone del Instituto Matemático de Toulouse, donde señala que históricamente, tiene nacimiento de dos investigaciones muy famosas que señalan que el principal trabajo desarrollado fue el enfoque funcional del análisis multivariante proporcionado por [19], donde su investigación se basó en los análisis factoriales con una función ambiente.

Para este tipo de datos de alta dimensión son los llamados "Datos funcionales" necesita una atención especial, y una nueva funcionalidad en tanto a las herramientas estadísticas teniendo en cuenta tales datos. Pero el verdadero punto de partida de los métodos estadísticospara datos funcionales fue en la década de 1990 con el trabajo de [90].

Dentro de los esquemas tradicionales el desarrollar nuevas herramientas estadísticas tanto en análisis descriptivo y multivariante con datos paramétricos, como no paramétrico, estos métodos al trabajarlos con funciones y no solamente con los datos como era costumbre, tiene un tratamiento con una atención especial para el fondo matemático funcional [100]. Para datos dependientes del tiempo, las observaciones se pueden ver como realizaciones de una función, ha sido suavizada con el factor tiempo en la cual se ha medido y puede definirse un error en puntos de tiempo específicos, pero que podrían haberse medido en cualquier momento.

Basados en revisiones bibliográficas, los datos funcionales se han analizado normalmente utilizando técnicas multivariantes, el analizar los datos funcionales como funciones tiene varias ventajas, ya que las funciones, a diferencia de los datos sin procesar, se pueden evaluar en cualquier "momento". Este tratamiento de la información es importante porque permite el uso de estadísticas multivariantes, donde se que requieren mediciones espaciadas uniformemente y permite extrapolación para

su uso en predicciones o decisiones de tratamiento [105]. Para los análisis de datos funcionales aplicado a varias disciplinas todas las observaciones son funciones, por lo cual constituyen realizaciones de variables funcionales, que se caracterizan por la evolución de una variable a lo largo del tiempo (proceso estocástico). Estas observaciones se denominan datos funcionales [21].

Cuando los datos se han obtenido en intervalos irregulares, en diferentes momentos o tiempos y sobre diferentes temas, podemos aplicar métodos funcionales (por ejemplo, componentes principales funcionales, correlación canónica funcional) ya que los análogos de estos, en el análisis clásico, son inapropiados o no están disponibles [25].

Los datos tomados como medidas son, a menudo, tratados mejor como funciones, incluso en casos donde las medidas se han realizado en un número relativamente pequeño de puntos [105].

Los modelos estadísticosque tratan con datos funcionales son llamadas Análisis de Datos Funcionales o FDA por su abreviatura en inglés. Estos métodos toman en cuenta el carácter funcional de los datos en ciertos escenarios los datos funcionales son tratados usando Metodologías propias del análisis multivariante; las cuales están en estudio todavía ya que algunas presentan ciertos errores al manejar este tipo de datos [6].

Las propiedades de las funciones han sido objeto de estudio de los matemáticos dentro del análisis funcional con un enfoque eminentemente teórico, actualmente, las funciones modelan muchos procesos importantes, y extraer adecuadamente la información que contienen las curvas puede ser extremadamente útil [11].

El Ecuador durante los últimos diez años ha tenido una trasformación, en varios puntos estratégicos y vulnerables dentro de la sociedad, el gobierno le ha dado prioridad a manejo de la información, mediante decretos ejecutivos y la creación de nuevas instituciones, con el objetivo principal de ofrecer un marco sociopolítico para vincular la dinámica demográfica, los procesos poblacionales y la formulación de políticas públicas (Plan del Buen Vivir, 2016).

El presente trabajo estudia la mortalidad bajo el esquema de número de muertes (t) según agrupamiento de causas (Lista de las principales causas de muerte Becker), Periodo 1997 - 2019, utilizando la información de las bases de datos públicas del

Instituto Nacional de Estadísticas y Censos. Los resultados se analizan utilizando el método estadístico de Análisis de Datos funcional, citado también como FDA y son contrastados con los descritos en la literatura mundial de acuerdo con la Clasificación de la OMS.

Se ha introducido también un análisis de las muertes por COVID-19 para situar la posición del Ecuador en comparación con el resto de países de América ya que ha sido una de las principales causas de muerte en el último año para el que aun no están disponibles los datos de las causas tradicionales.

En este trabajo se plantea una estrategia novedosa para estudiar el "comportamiento temporal de la mortalidad en Ecuador, permitiendo hacer una caracterización en el tiempo de desarrollo de los diferentes tipos de muertes bajo unos supuestos obtenidos empíricamente de los resultados internacionales de organización mundial de la salud. 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 su proceso de prevención y sus políticas de ámbito en salud pública" [72].

1.2. Hipótesis y Objetivos

1.2.1. Hipótesis

Según las ideas de [29] , una variable aleatoria toma valores en un espacio de funciones, como un espacio infinito dimensional. Así , una observación $f(t)$ de la variable aleatoria se denomina dato funcional en un instante t, entonces el modelo nos permite describir el comportamiento de las variables asociadas con la mortalidad y, en particular, las relacionadas con las causas de muerte en el Ecuador, con la evolución de las tasas de mortalidad por COVID-19, tanto en los países de América como en las distintas provincias de Ecuador.

1.2.2. Objetivo General

Estudiar la mortalidad y sus principales causas registradas en las regiones del Ecuador en el periodo 1997-2017, junto con la evolución de la mortalidad por COVID,

tanto en los países del entorno como en las provincias del Ecuador, a través del método Multivariante de Análisis de Datos Funcionales (ADF) y complementar el análisis con nuevas herramientas gráficas.

1.2.3. Objetivos Específicos

- Efectuar una revisión bibliográfica sobre los principales métodos estadísticosutilizados en estudios de mortalidad.
- Describir y aplicar algunos de los métodos clásicos del Análisis multivariante en el contexto de la salud y la mortalidad.
- Ilustrar la utilidad de los métodos biplot en el estudio de datos demográficos.
- Describir los métodos de Análisis de Datos Funcionales (ADF) y proponer su uso en el contexto de estudios epidemiológicos de mortalidad.
- Proponer métodos estadísticosno tradicionales para desarrollar las capacidades del análisis de los datos funcionales, concretamente una representación biplot asociada al análisis de componentes principales para datos funcionales.
- Ilustrar la utilidad del método de Análisis de Datos Funcionales (ADF) y su extensión, en el análisis de la mortalidad y sus principales causas registradas en las regiones del Ecuador en el periodo 2010-2021.
- Ilustrar la utilidad del método de Análisis de Datos Funcionales (ADF) y su extensión, en el análisis de la mortalidad por COVID-19 en América y la posición relativa del Ecuador en relación al resto de los países.
- Ilustrar la utilidad del método de Análisis de Datos Funcionales (ADF) y su extensión, en el análisis de la mortalidad por COVID-19 en las distintas regiones del Ecuador.

1.3. Organización de la memoria

Además de este capítulo de introducción,

En el capítulo 2 introduciremos los datos de salud y mortalidad ue ilustraran tanto los métodos más estándar como las propuestas realizadas en es tesis, a saber:

- Tasa de mortalidad en Ecuador clasificadas por causas de muerte.
- Datos sobre la incidencia de *Helicobacter pilory*
- Evolución de las tasas de mortalidad por COVID en los países del entorno del Ecuador
- Evolución de las tasas de mortalidad por COVID en las distintas regiones del Ecuador

El capítulo 3 describe el Análisis de Componentes principales y la aplicación a los datos de las causas de muerte.

En el capítulo 4 describimos los métodos biplot junto con una aplicación a un estudio sobre *Helicobacter pilory* en Milagro (Ecuador).

El capítulo 5 describe el Análisis de datos funcionales.

El capítulo 6 describe la propuesta fundamental de la tesis que consiste en la extensión de los métodos biplot a los datos funcionales.

El capítulo 7 muestra las aplicaciones de las principales propuestas a los datos de mortalidad.

Capítulo

2 Datos

2.1. Introducción

A lo largo de la memoria utilizaremos diversos conjuntos de datos para aplicar las distintas Metodologías que vayamos describiendo. Hemos de hacer notar que, todas las aplicaciones que presentamos no son meras ilustraciones sino que tienen interés en sí mismas. Se trata de obtener resultados interesantes a partir de los datos utilizando técnicas novedosas o incluso desarrollando nuevas propuestas que ayuden a comprender mejor la estructura de la información observada.

Usaremos 4 conjuntos de datos diferentes:

- Evolución de la mortalidad por causas de muerte desde 2010 a 2019.
- Prevalencia de la bacteria *Helicobacter pylori*.
- Tasas de mortalidad acumulada por COVID-19 en las distintas provincias del Ecuador.
- Tasas de mortalidad acumulada por COVID-19 en América para estudiar la posición relativa del Ecuador.

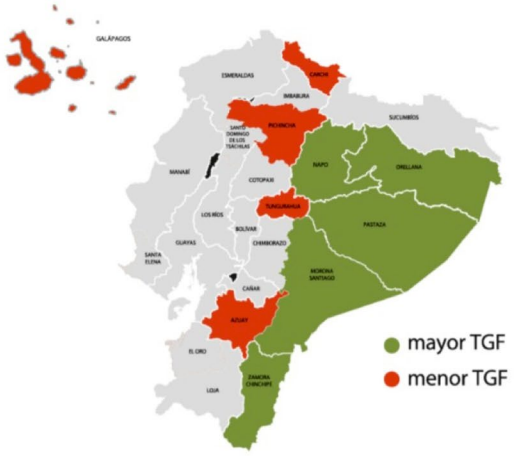
2.2. Mortalidad en Ecuador

2.2.1. La mortalidad

En general, "la demografía parte del hecho común de que las fuentes de información contienen errores de captación que comprometen los niveles y tendencias de los fenómenos. Por lo cual, se deben desarrollar y aplicar métodos basados en técnicas matemáticas que coadyuven a ofrecer un escenario próximo a la realidad. Sin embargo, a excepción de que se tengan datos empíricos, se debe tener especial cuidado en respetar la información básica, ya que de otra forma el estudio no reflejar´ıa la realidad y sería poco útil para resolver o prever situaciones que pueden alterar de forma negativa los ritmos de crecimiento poblacional" [53] (p.12).

La mortalidad es uno de los tres factores de cambio demográfico, sin embargo, a diferencia de los otros, para su estudio solo de dispone de registros administrativos, a excepción de la mortalidad en la infancia. Sin embargo, no contar con al menos una encuesta que señale los porcentajes de cobertura de la fuente, obliga que el estudio consista en confrontar la información levantada en los ejercicios de enumeración poblacional con la disponible en estadísticas vitales. Por lo anterior, "las fuentes de información son los Censos de Población y Vivienda, además de los datos sobre defunciones captados en registros administrativos disponibles a la fecha de elaboración para el periodo 1990-2010. Para las mismas, se consideró como premisa la posibilidad de que su cobertura y declaración de la edad no fuera adecuada. Con base en ello, el proceso inicio con la distribución de los no especificados por edad y sexo, seguido de la corrección de la declaración de la edad y el traslado de las poblaciones" [52] (p.11).

Al tratar el tema sobre defunciones "se dispone de información anual, para la población se tiene únicamente información decenal, por lo que surgió la necesidad de reconstruir la población para cada año calendario"[52].



Mayor TGF	AÑO			
	2010	2012	2015	2020
Morona Santiago	4,7	4,4	3,9	3,3
Napo	3,9	3,8	3,6	3,3
Zamora Chinchipe	3,9	3,7	3,5	3,1
Orellana	3,9	3,6	3,3	2,8
Pastaza	3,7	3,6	3,3	3,0
Menor TGF	**AÑO**			
	2010	2012	2015	2020
Pichincha	2,2	2,2	2,1	2,0
Tungurahua	2,5	2,4	2,3	2,2
Galápagos	2,5	2,5	2,4	2,4
Azuay	2,5	2,4	2,2	2,1
Carchi	2,6	2,5	2,4	2,3

Figura 2.1: Tasa global de fecundidad por provincias período 2010-2020(Instituto Nacional de Estadísticas y Censos, 2018)

En el caso de las defunciones de mayores de un año se "formuló como guía de corrección que los registros vitales tienen errores por declaración de la edad, además de adolecer de subestimación. Para el primer aspecto se utilizó el mismo procedimiento que el aplicado a la población, es decir, se suavizó con la adecuación sugerida a la técnica de Alan Gray. Mientras que el segundo punto implica que los montos de defunciones corregidos deben ser mayores que los registrados en la fuente. Bajo esta situación se recuperaron premisas de la información básica para forjar el procedimiento útil en el análisis de la mortalidad" [53](p25).

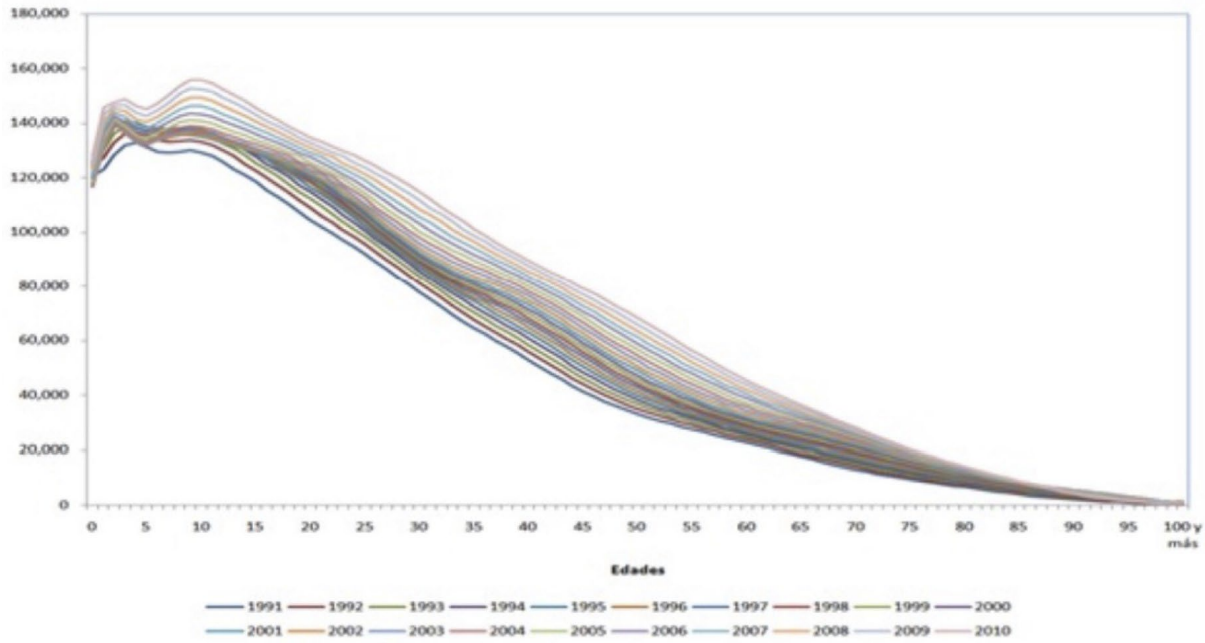


Figura 2.2: Estructuras de las poblaciones anuales interpoladas por edad desplegada 1991-2010, mujeres (INEC, 2021)

Sin embargo, “a pesar de correcciones previas, los indicadores mantenían fluctuaciones e irregularidades que no corresponden a características reales de la población, sino más bien al procedimiento de interpolación y a errores de diverso tipo de la información básica. En particular, las relaciones de supervivencia en las primeras iteraciones mostraron inconsistencias, para edades mayores a los 80 años. Por lo que se decidió suavizarlas con la fórmula de Whitaker-Henderson, al tener la ventaja de que no sigue una curva predeterminada, sino que permite regular la graduación dentro de un margen amplio que va desde la reproducción exacta de los valores observados, hasta valores que siguen una línea recta. De esta manera el procedimiento permitió suavizar la información y mantener la concordancia con los valores originales” [53](p.23).

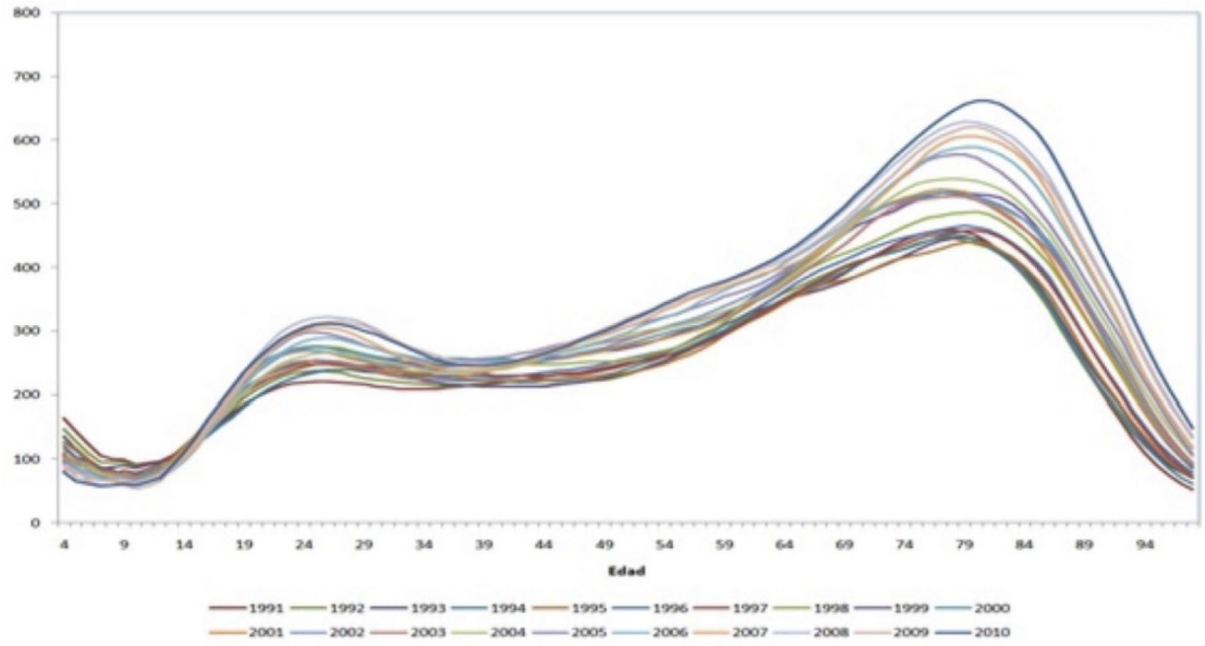


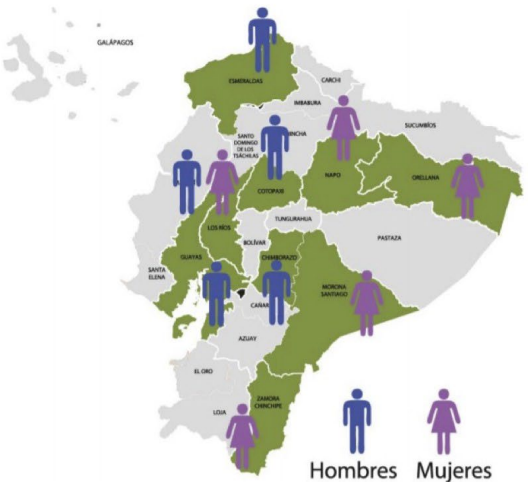
Figura 2.3: Estructura de las defunciones suavizadas por mala declaración de la edad, 1991-2010, hombres (INEC, 2021)

La figura 2.3 se "observan las estructuras de las defunciones masculinas suavizadas por efecto de la inexacta declaración de la edad para cada uno de los años del periodo 1991-2010, se tienen las mismas estructuras, pero corregidas por subestimación. Se observa que el efecto de la corrección se centra principalmente en edades comprendidas entre los 45 y 89 años, lo que implica presumir que en estas edades se tienen mayores problemas producto de la no declaración del evento muerte, situación coherente con la premisa de que en edades mayores la intensidad del evento es mayor, por lo que la probabilidad de no captación también incrementa. Otro punto a recalcar está en las defunciones en edades de entre 15 y 39 años, ya que después de suavizar el efecto de la declaración de la edad y corregir por subestimación, se conserva el comportamiento observado, con lo cual se asevera que el procedimiento de las políticas de salud pública"(Instituto Nacional de Estadística y Censo, 2020: p. 34).

Ecuador tiene 17,7 millones de habitantes en el 2017, en el 2050 "llegaremos a 23'4 millones de habitantes, según las proyecciones poblacionales presentadas por

el Instituto Nacional de Estadística y Censos (INEC). Las proyecciones son un instrumento indispensable para llevar a cabo la planificación demográfica, económica, social y política del país y permite establecer posibles escenarios y prever acciones".

Los niveles y estructuras estimadas para 2010 en el ejercicio de conciliación, se aplica a un modelo de regresión log´ıstica para estimar los niveles de mortalidad para cada ano de la proyección con una cota superior de 77.6 de esperanza de vida para con lo cual el indicador se estima en el año 2020 a 80.1 para mujeres y 74.5 años en promedio en hombres (Ver Figura 2.4).



Hombres	Año			
	2010	2012	2015	2020
Los Ríos	68,6	69,1	69,8	71,0
Esmeraldas	69,1	69,5	70,2	71,3
Chimborazo	70,0	70,8	71,9	73,4
Guayas	70,2	70,6	71,3	72,3
Cotopaxi	70,4	71,2	72,4	73,9
Mujeres	**Año**			
	2010	2012	2015	2020
Los Ríos	75,0	75,5	76,3	77,6
Zamora Chinchipe	76,0	76,4	77,1	78,2
Morona Santiago	76,0	76,4	77,1	78,2
Napo	76,0	76,4	77,1	78,2
Orellana	76,0	76,4	77,1	78,2

Figura 2.4: Las 5 provincias con menor esperanza de vida por sexo período 2010-2020 (Instituto Nacional de Estadísticas y Censos, 2018)

La estimación "para el periodo intermedio que completa el periodo 2000-2010 se hizo mediante interpolación de los montos totales para cada provincia, manteniendo las estructuras por edad, sexo y distribución de los eventos del fenómeno (inmigraciones y emigraciones), a partir de la experiencia registrada por los censos de 2001 y 2010" (Instituto Nacional de Estadística y Censo, 2020).

2.2.2. Tipos de Mortalidad en Ecuador

El hablar de las estadísticas de defunciones, entre ellas, el número de muertes y tasas de mortalidad por sexo y por tipo (maternas/infantiles). Estas estadísticas se

construyen a partir de las inscripciones de los hechos vitales (defunciones) proporcionadas por el Registro Civil, y se alimentan también de la información de la red pública y complementaria de los establecimientos de salud.

La incidencia de las enfermedades cardiovasculares (ECV) está aumentando en todo el mundo en vías de desarrollo, causando más de 16 millones de muertes al año, el 80 % de las cuales se producen en países con ingresos bajos y medios [62]. La actividad física regular reduce el riesgo de mortalidad cardiovascular [14]. Además, la inactividad física (OPS/OMS, 2020) representa entre el 1 % y el 3 % de los costes sanitarios, excluyendo los costes asociados a la salud mental y a las afecciones musculo esqueléticas (Organización Panamericana de la Salud / Organización Mundial de la Salud (OPS/OMS), 2020), y contribuye al 6 % de la carga de mortalidad de las enfermedades coronarias y al 10 % del cáncer de mama y de colon [64].

El comportamiento sedentario, a diferencia de la actividad física, abarca una amplia gama de comportamientos que implican una postura sentada o reclinada y que no aumentan el gasto energético por encima de 1,5 equivalentes metabólicos durante el tiempo de vigilia [73]. Los comportamientos sedentarios se asocian con la ECV, el cáncer y la mortalidad por cualquier causa, independientemente de la actividad física [81]. Las directrices actuales sobre actividad física no prescriben una directriz cuantitativa sobre el tiempo que se pasa sentado [60].

El mayor número de defunciones se registra en la región costa con 35.901 defunciones representando el 51,8 % del total de defunciones. El mayor número de defunciones se presenta en personas de 65 años y más, en hombres y mujeres con 21.413 y 20.942 respectivamente. Durante el 2017, “15.788 defunciones generales ocurrieron en establecimientos pertenecientes al Ministerio de Salud Pública, lo cual representa el 22,8 % sobre el total de defunciones”.

Una de las causas de mortalidad que está presente es la diabetes, que es una carga para la salud pública en todo el mundo asociada a una mayor morbilidad relacionada con sus complicaciones, un exceso de discapacidad los costes de la atención en salud pública y la mortalidad prematura [78]. Según el informe mundial sobre la diabetes de la Organización Mundial de la Salud, el número de personas con diabetes ha aumentado de forma constante en las últimas décadas, debido al crecimiento de la población, el aumento de la edad media edad media de la población y el aumento de

las tasas de sobrepeso y obesidad. Además, las tasas de prevalencia de la diabetes han aumentado más rápidamente en los países de bajo y mediano presupuesto a diferencia de los países más ricos [23].

Se estima que la incidencia de la diabetes aumenta con la edad hasta los 65 años, aproximadamente, tras lo cual tanto la incidencia como la prevalencia parecen estabilizarse. Como resultado, los adultos mayores con diabetes pueden tener una diabetes de larga duración que se inició en la mediana edad o antes.Para [5], en su estudio previo realizado para examinar la asociación entre factores sociodemográficos y de estilo de vida y de estilo de vida y las enfermedades crónicas autodeclaradas, prevalec´ıa en el 13,1 % de los adultos de 60 años o más.

Cabe destacar la alta prevalencia del síndrome metabólico, un conjunto de factores de riesgo cardio metabólicos asociados a un mayor riesgo de diabetes, también se describió entre los ecuatorianos mayores [45]. Asimismo, las mujeres mayores definidas con obesidad abdominal tenían 2 veces más probabilidades de tener diabetes que las que no la tenían [4].

A pesar de estos hechos, existen escasos datos epidemiológicos sobre la prevalencia de la diabetes, particularmente entre los adultos mayores en Ecuador, sin embargo, “el mayor porcentaje de las defunciones ocurren en casa. En el año 2017, la enfermedad isquémica del corazón es la principal causa de muerte en hombres y mujeres con 7.404 defunciones. En el año 2017, la enfermedad isquémica del corazón es la principal causa de muerte en los hombres con 4.230 defunciones, seguido de los accidentes de transporte terrestre con 2.419 defunciones”.

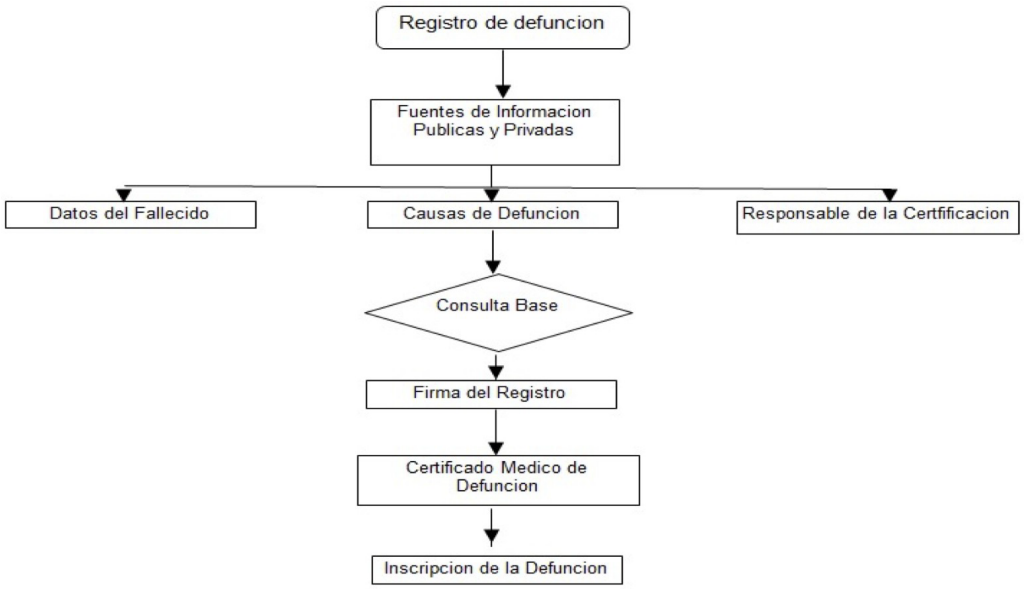


Figura 2.5: Diagrama del Registro de Defunción(INEC, 2017)

En el año 2017, "la enfermedad isquémica del corazón es la principal causa de muerte en las mujeres con 3.174 defunciones, seguida de la diabetes mellitus con 2.606 defunciones. En el año 2017, la tasa de mortalidad más alta se presenta en la provincia de Chimborazo con 5,0 muertes por cada 1.000 habitantes de esta provincia".

Figura 2.6: Componentes del Registro de Defunción. (INEC, 2017)

Seguido se encuentra Bolívar con "una tasa de 4,7 muertes por cada 1.000 habitantes de esta provincia. En el año 2017, la tasa de mortalidad materna más alta

se presenta en la provincia de Pastaza con 111,28 muertes por cada 100.000 nacidos vivos de esta provincia. Las principales temáticas investigadas se relacionan con" (INEC, 2016).

Figura 2.7: Causas Principales de Muerte año 2000 al 2016 (Who, 2017)

De los 56,4 millones de defunciones registradas en el mundo en 2016, "más de la mitad (el 54 %) fueron consecuencia de las 10 causas que se indican a continuación. Las principales causas de mortalidad en el mundo son la cardiopatía isquémica y el accidente cerebrovascular, que ocasionaron 15,2 millones de defunciones en 2016 y han sido las principales causas de mortalidad durante los últimos 15 años"(World Health Organization, 2018).

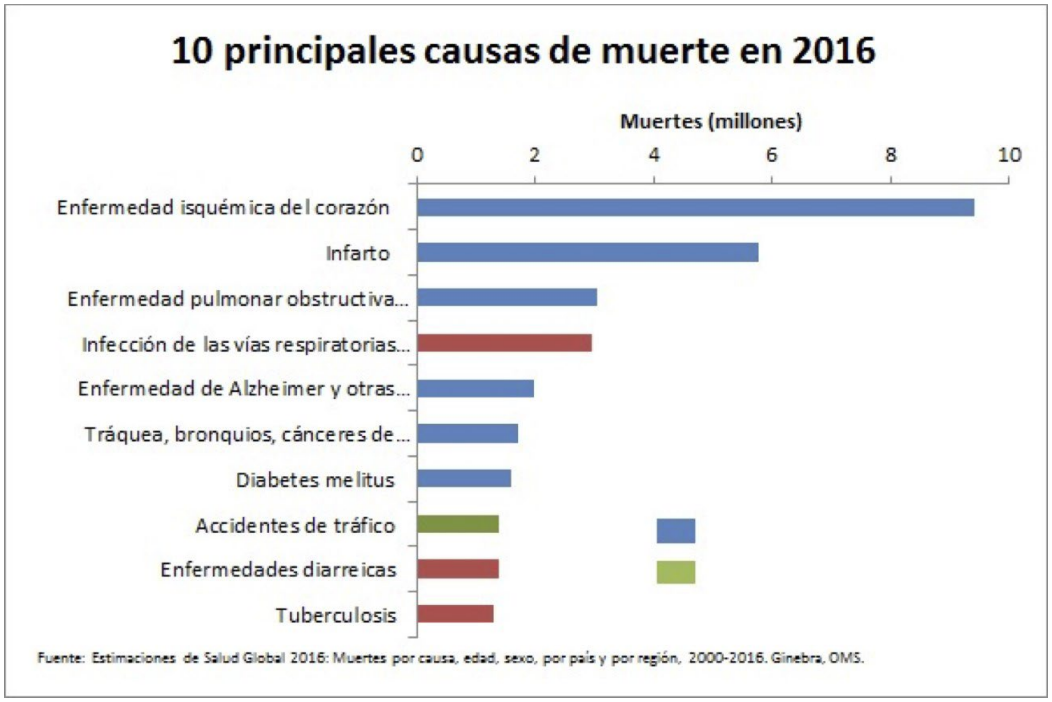


Figura 2.8: Causas Principales de Muerte año 2016.(WHO, 2017)

Por lo tanto con la pandemia del covid 19 hace que en la mayoría de los casos, las personas tenga presente una obesidad abdominal que es un importante problema de salud pública asociado con la resistencia a la insulina, la diabetes de tipo 2, la hipertensión, la dislipidemia, discapacidad y aumento de la mortalidad entre los adultos mayores [78]. Se ha informado de que los adultos mayores con obesidad abdominal tienen un de riesgo de padecer trastornos cardiovasculares, independientemente de la categoría del índice de masa corporal [76].

Además, el envejecimiento se caracteriza por un aumento de la adiposidad y una redistribución del patrón de adiposidad. En general, el volumen de grasa subcutánea disminuye primero, seguido mucho más tarde por la pérdida de grasa en los depósitos viscerales, lo que resulta en redistribución de la grasa de los depósitos subcutáneos a los viscerales desde de la mediana edad hasta los 800 años o más. Las investigaciones anteriores también han demostrado que el aumento del per´ımetro de la cintura con el envejecimiento causados principalmente por un aumento desproporcionado de la adiposidad visceral frente a la grasa subcutánea abdominal [78].

De acuerdo a la teoría de la "transición demográfica, la tasa bruta de mortalidad se verá disminuida con el transcurso de los años, esto como resultado de las mejoras en tecnología, medicina y en general cobertura de salud" [70]. Según [84], a partir de la "década del cincuenta, América Latina empezó a tener un descenso en las tasas de mortalidad, como resultado del incremento en la esperanza de vida en la región".

El Ecuador presenta en el año 1960 una alta "tasa de mortalidad, como era de esperarse, esta presenta un valor aproximado de 15,55 muertes por cada mil habitantes. Esta cifra, en un período de 50 años, decreció a 5,18 muertes por cada mil habitantes en 2010; según datos del Banco Mundial, la tasa de mortalidad de Ecuador se encuentra por debajo del promedio mundial, que fue de 7,95 muertes por cada mil habitantes en 2010"[78], para el año 2020. en este año "se registraron más de 41 mil muertes en exceso con respecto al 2019. Es importante aclarar que, en el contexto de la pandemia, hubo imprecisiones en las actas de defunción. Por tanto, el INEC hizo un trabajo de cruce de datos con el Registro Civil y el Ministerio de Salud Pública para determinar la causa principal del fallecimiento y así mejorar la calidad del registro"(Instituto Nacional de Estadística y Censo, 2020).

Como resultado del trabajo técnico se determinó que, "del total de muertes en exceso, 23.793 son casos confirmados o sospechosos de la COVID-19 y las 17.284 defunciones restantes están asociadas principalmente con enfermedades respiratorias como influenza y neumonía, enfermedades isquémicas del corazón y diabetes"(INEC, 2021).

2.2.3. Definiciones de Defunciones

Defunción : "es la desaparición total y permanente de todo signo de vida en un momento cualquiera posterior al nacimiento, sin posibilidad de resurrección".

Defunción fetal : "la muerte de un producto de la concepción, antes de su expulsión o su extracción completa del cuerpo de su madre, independientemente de la duración del embarazo; la muerte está indicada por el hecho de que después de la separación, el feto no respira ni da ninguna otra señal de vida, como latidos del corazón, pulsaciones del cordón umbilical o movimientos efectivos de los músculos de contracción voluntaria".

Mortalidad infantil : "las defunciones ocurridas en niños que no han cumplido

un año de vida".

Causa básica de defunción : "son todas aquellas enfermedades, estados morbosos o lesiones que produjeron la muerte o que contribuyeron a ella y las circunstancias del accidente o de la violencia que produjo dichas lesiones".

2.2.4. Datos de causas de muerte entre los años 2010 y 2010

Como datos finales para el estudio disponemos del número de defunciones para los años desde el 2010 hasta el 2020 clasificadas de acuerdo con las causas de muerte. Disponemos también de la población de cada uno de los años de forma que podemos convertir los datos en tasas de defunción para cada una de las causas de muerte.

Queremos observar como han evolucionado las tasas para las distintas causas a lo largo del tiempo y cuales de ellas tienen perfiles de evolución similares.

2.2.5. Representatividad de los Datos de Estudio

"La información estadística oficial es esencial para el desarrollo del Ecuador, y en particular para mejorar las condiciones de vida de la población, en la medida que permite sustentar el diseño y formulación de las políticas públicas y decisiones privadas"[55]. Por tal motivo es "preciso que todas las variables que contiene el Informe de Defunción General deben estar correctamente registradas y que las mismas no vengan omitidas para la investigación con fines Estadísticos.

La producción de las estadísticas de defunciones se enmarca en el ámbito de las estadísticas vitales, investigación que comenzó desde el año 1954, clasificando las defunciones por fecha de ocurrencia y por lugar de residencia habitual de los fallecidos. A partir de este año, la investigación se ha ejecutado de forma continua, con modificaciones en coberturas temáticas de la investigación, cobertura geográfica, formas de presentación de publicaciones, tabulados y productores de la información" [55].

Desde 1976, el Instituto Nacional de Estadística y Censos, con "la cooperación de las oficinas dependientes de la Dirección General de Registro Civil, procesa y publica de manera anual y continúa la información de Defunciones Generales, con una cobertura a nivel nacional. Además, en ese año, se elabora el primer anuario de

nacimientos y defunciones investigando las siguientes variables: fecha de inscripción, fecha de nacimiento del fallecido, edad al fallecer, sexo, fecha de acaecimiento, estado civil del fallecido, lugar del fallecimiento o ubicación (establecimiento de Salud, domicilio, otros)"[52].

Con las sucesivas revisiones en la Clasificación Internacional de Enfermedades (CIE), "ésta se fue ampliando de manera notable, enriquecida con información más detallada en muchas categorías, además de la incorporación de nuevas categorías, como consecuencia de la mejora del conocimiento sobre muchas enfermedades y el descubrimiento de otras" [52].

Uno de los cambios más importantes ocurrió de la Quinta a la Sexta Revisión, cuando la CIE pasó de una Clasificación de causas de muerte, con unas doscientas categorías, para una Clasificación de enfermedades y causas de muerte, con más de mil categorías. Es así que de la Clasificación original de Bertillon, de 1893 (la CIE-Cero), que tenía un total de 161 categorías y ofrecía un total de 200 codigos posibles, se pasó "a la CIE-101, implantada en el mundo a partir de 1994, que contiene actualmente un total de 12.421 códigos distribuidos en 2.036 categorías. Este mayor grado de detalle permite profundizar los análisis de enfermedades, ya sea individualmente o en grupos específicos, tanto con datos de mortalidad como de morbilidad. Por otro lado, dificulta la tabulación completa de datos para obtener una visión panorámica de la situación de salud, para la identificación de los problemas más relevantes y la definición de prioridades"[1].

Por tal razón, la misma CIE en "general ofrece listas más cortas, a partir de la agregación de varias categorías en un único grupo. La Clasificación de Bertillon, por ejemplo, era presentada en tres "nomenclaturas": una corta de 44 categorías, una intermedia de 99 y la más detallada de 161. La CIE-10 está estructurada en 21 Capítulos, los cuales se desglosan en 261 grupos, que contienen un total de 2036 categorías. Además, la CIE-10 ofrece cuatro listas especiales (cortas o abreviadas) para la tabulación de mortalidad y una para morbilidad"[50].

El primero, las continuas "revisiones a la Clasificación Internacional de Enfermedades (CIE), herramienta que se utiliza para codificar y tabular los datos sobre causas de muerte y que está sujeta a revisiones periódicas con el fin de reflejar los progresos en el campo sanitario. La segunda, derivada de la anterior, se concreta en

los cambios en la codificación de las sucesivas CIES, que tienen como resultado una ampliación y especificación de las causas de defunción así como un trasvase en los diagnósticos debido a los avances científico-médicos. Como consecuencia, con cada nueva revisión de la CIE las series temporales de mortalidad por causa de muerte específica se interrumpen"[71].

El tercero "viene derivado de la dificultad en el diagnóstico de los decesos, lo que provoca un elevado porcentaje de defunciones clasificadas como desconocidas o mal definidas, se trata de un método para reclasificar las causas de muerte entre dos revisiones de la CIE, tomando como referencia la más actual"[13].

Esta Metodología consta de tres fases principales: "la primera, el establecimiento de una tabla de correspondencia de enfermedades entre las sucesivas CIEs; la segunda, la construcción de las asociaciones elementales cuyo objetivo es la coherencia médica y numérica (número de muertes); y la tercera, el cálculo de los coeficientes de transición que permite la distribución de los óbitos y construcción de las series de datos"[9].

Actualmente en Ecuador se "utiliza las dos vías para el registro de defunciones generales de manera física y electrónica, para poder realizar la publicación se unifica las bases de datos de la información de los registros recolectados de manera física y digital, posterior se realizan cotejamientos de información con otras fuentes como registros de muertes violentas del Ministerio del Interior, base de cedulados de DIGERCIC, base de muertes maternas del MSP, base de egresos hospitalarios, base de nacidos vivos, base de defunciones fetales con el objetivo de mejorar la causa de muerte y la cobertura"[54]

La base de datos que se utilizó para el análisis de este documento, es el resultado de la labor conjunta realizada entre instituciones estatales y de publicación publica que son efectuadas por el INEC [51] , en conjunto con otras oficinas como la Dirección Nacional de Registro Civil, Identificación y Cedulación y del Ministerio de Salud Pública, integrantes del Sistema Estadístico Nacional que constituyen la fuente de información y está disponible en el sitio web de ecuador en cifras. Sobre mortalidad, existen dos instancias: la primera que se refiere al aspecto médico-estadístico que se encarga de certificar el acontecimiento del hecho y la segunda que se refiere a la institución pública, al de Registro Civil, Identificación y Cedulación que se encarga de

Figura 2.9: Componentes del Registro de Defunción. Fuente: (Instituto Nacional de Estadística y Censos (INEC), 2017

la inscripción y legalización del hecho vital, en la Figura 25, se muestra el diagrama.

Actualmente en el INEC se "puede tener acceso a las Estadísticas Vitales (Nacimientos y Defunciones) esta información es sistemática y continúa, se refiere a la inscripción de nacimientos y defunciones. Se revisó las bases de datos de las defunciones generales desde el año 1997 hasta el 2020, de todas las provincias del Ecuador, la cantidad de variables presentes en cada una de las bases es 268, basados en nuestro objetivo general se verifico los datos contenidos en la matriz de principales causas de mortalidad tomando los 20 últimos años (1997-2020p*), las causas de fallecimiento consideradas son solo las relacionadas a nuestro estudio, con esta premisa la base global de todos los 20 años"(Instituto Nacional de Estadística y Censo, 2020), con un total de 1248750 registros de defunción con 69 variables de interés como se muestra en la tabla siguiente.

1. Enfermedades infecciosas intestinales.
2. Tuberculosis.
3. Enfermedades transmitidas por vectores y rabia.
4. Enfermedades inmunoprevenibles.
5. Meningitis.
6. Septicemia.
7. Enfermedad por virus de la inmunodeficiencia (VIH).

8. Neoplasia maligna del esófago.
9. Neoplasia maligna del estómago.
10. Neoplasia maligna del colon, sigmoide, recto y ano.
11. Neoplasia maligna del hígado y de las vías biliares.
12. Neoplasia maligna de la vesícula biliar y de otras.
13. Neoplasia maligna de páncreas.
14. Neoplasia maligna de laringe.
15. Neoplasia maligna de la tráquea, bronquios y pulmón.
16. Melanoma y otra neoplasia maligna de la piel.
17. Neoplasia maligna de la mama.
18. Neoplasia maligna del útero.
19. Neoplasia maligna del ovario.
20. Neoplasia maligna de la próstata.
21. Neoplasia maligna del riñ ón, excepto pelvis renal.
22. Neoplasia maligna de la vejiga urinaria.
23. Neoplasia maligna del encéfalo.
24. Neoplasia maligna del tejido linfático, hematopoyético y afines.
25. Neoplasias benignas *in situ* y de comportamiento incierto.
26. Diabetes Mellitus.
27. Desnutrición y anemias nutricionales.
28. Trastornos de los líquidos, electrolitos, y del equilibrio á cido básico.

29. Demencia y enfermedad de Alzheimer.
30. Trastornos mentales y del comportamiento por uso de sustancias psicoactivas.
31. Enfermedad de Parkinson.
32. Epilepsia y estado de mal epiléptico.
33. Enfermedades reumáticas crónicas.
34. Enfermedades hipertensivas.
35. Enfermedades isquémicas del corazón.
36. Enfermedad cardiopulmonar y enfermedad de la circulación pulmonar.
37. Trastornos valvulares no reumáticos.
38. Cardiomiopatía.
39. Paro cardíaco.
40. Arritmias cardíacas.
41. Insuficiencia cardíaca, complicaciones y enfermedades mal definidas.
42. Enfermedades cerebrovasculares.
43. Aterosclerosis.
44. Aneurisma y disección aórticos.
45. Enfermedades respiratorias agudas excepto influenza y neumonía.
46. Influenza y neumonía.
47. Enfermedades crónicas de las vías respiratorias inferiores.
48. Edema pulmonar y otras enfermedades respiratorias que afectan al intersticio.
49. Insuficiencia respiratoria.

50. Apendicitis, hernia y obstrucción intestinal.
51. Cirrosis y otras enfermedades del hígado.
52. Enfermedades del sistema osteomuscular y tejido conjuntivo.
53. Enfermedades del sistema urinario.
54. Embarazo, parto y puerperio.
55. Ciertas afecciones originadas en el período prenatal.
56. Malformaciones congénitas, deformidades y anomal´ıas cromosómicas.
57. Accidentes de transporte terrestre.
58. Caidas accidentales.
59. Disparo de arma de fuego no intencional.
60. Ahogamiento y sumersión accidentales.
61. Accidentes que obstruyen la respiración.
62. Envenenamiento accidental.
63. Lesiones autoinflingidas intencionalmente (Suicidio).
64. Agresiones (Homicidios).
65. Eventos de intención no determinada.
66. Resto de causas. (88 en los gráficos)
67. Causas mal definidas. (99 en los gráficos)
68. COVID-19, virus identificado.
69. COVID-19, virus no identificado.

2.2.6. 4.2. Depuración de la Base de datos

El análisis comparativo de los datos de la estadística de Defunciones Generales se la realiza cotejando los resultados obtenidos en el "año de estudio con los resultados de la investigación del año anterior, con la finalidad de detectar inconsistencias en el aumento o disminución de la cobertura del año que se está investigando. Para medir la cobertura que se ha obtenido del registro estadístico se realizan dos comparaciones la primera de los datos obtenidos en INEC con los datos que mantiene DIGERCIC, y la segunda comparando los registros obtenidos en INEC frente a las estimaciones de nacimientos del INEC y de la OMS" [51].

La importancia del sistema de códigos CIE-10 puede "evaluarse a partir de su aplicación en varios ámbitos de la gestión de la calidad, la asistencia sanitaria, la tecnología de la información y la salud pública"[63].

El sistema de códigos CIE-10 "ofrece códigos de procedimiento precisos y actualizados para mejorar el costo de la atención médica y garantizar políticas de reembolso justas. Los códigos actuales ayudan específicamente a los proveedores de atención médica a identificar a los pacientes que necesitan un manejo inmediato de enfermedades y a diseñar programas de manejo de enfermedades efectivos" [96].

La CIE-10 se ha adoptado "internacionalmente para facilitar la implementación de una atención médica de calidad, así como su comparación a escala mundial" [51], Los códigos CIE-10 tienen una "importancia particular en la investigación, ya que el análisis de códigos es un componente esencial de la investigación y el desarrollo. El sistema de código y la lógica permiten menos errores de codificación que, en última instancia, se benefician en los análisis de investigación y desarrollo"[51].

Después de la revisión de cada una de las bases de datos por cada año de estudio, "se procedió a definir los datos que se usaron en la presente investigación que es el número de muertes (t) según agrupamiento de causas (Lista de las principales causas de muerte Becker) del Periodo 1997 – 2020 con las variables de interés, causas de muerte de la lista corta CIE-10 y los años de estudio, es importante considerar que existen criterios para la elaboración de las listas de principales causas de muerte siendo los más importantes su relación con la CIE-10, criterios epidemiologicos, importancia en salud pública y el uso específico de la lista"[1].

2.3. COVID-19

2.3.1. Introducción

Los primeros casos de la enfermedad por el virus de Corona 2019 (COVID-19) surgieron en Wuhan, China, a finales de diciembre de 2019 [109]. Un mes después, la presencia de esta enfermedad fue declarada emergencia de salud pública de interés internacional por el director general de la Organización Mundial de la Salud (OMS). Hasta el 30 de abril de 2020, la OMS [24] registró un total de 3.090.445 casos confirmados de COVID-19 en todo el mundo y 217.769 muertes confirmadas. En total, 213 países, zonas o territorios estaban afectados por el COVID-19 para entonces.

El virus del COVID-19 , "que en un corto periodo de tiempo logro convertirse en un brote a nivel mundial fue considerado como emergencia de salud pública por la OMS en enero del 2020, al fallar su contención y al tener características infectocontagiosas fue propagándose a nivel mundial por lo que se denominó una pandemia" [110].

La mayor parte de los estudios reportados sobre el impacto del COVID-19 "en la salud mental se han realizado en China, lo cual es comprensible ya que este fue el primer país afectado por la pandemia. A pesar de la utilidad de estos datos, las características demográficas e interindividuales son determinantes en las respuestas psicológicas de las diferentes poblaciones a un evento estresante a gran escala como el COVID-19 "[75].

La rápida propagación del virus por todo el mundo obligó a las administraciones locales y nacionales a tomar medidas sin precedentes para reducir el impacto de esta pandemia [101]. Las medidas incluyeron la vigilancia activa de los casos sospechosos, el autoaislamiento o el distanciamiento social, las restricciones a los viajes y al transporte, e incluso el cierre de las fronteras de los países [38].

Por esta razón, es "fundamental aumentar el volumen de estudios que evalúen el impacto de esta pandemia en cada uno de los países afectados, especialmente en países como Ecuador, que ha sido uno de los más afectados por la pandemia, al presentar en abril de 2020 las tasas de mortalidad más altas de América Latina"[42].

El impacto que este género en "nuestro país fue demasiado grave, desde colapso en los centros de salud, a su vez el desgaste sufrido por el personal de salud, para

entender la situación se debe tomar en cuenta diferentes variantes que influyen de manera negativa en las limitaciones que se presentan en el sistema de salud, abarcando desde la principal que fue la falta de información para proceder de manera correcta y el incumplimiento de los planes propuestos"[65].

Esta clase de virus se la presenta como el responsable de causar desde un resfriado común hasta enfermedades respiratorias graves, "al ser una nueva enfermedad toda medida o estrategia de prevención se fue añadiendo o descubriendo sobre la marcha, sus síntomas que suelen varias desde los más leves a graves son similares a los de una patología normal lo cual influyo en disfrazar esta enfermedad, a su vez para poder tratar a los contagiados graves se necesitaba de las instalaciones con el equipamiento indicado, al carecer de un número suficiente afecto de sobre medida la atención a las personas lo que en algunos casos genero el agravamiento de la patología o hasta la muerte"[22].

Otra situación presentada en los "sistemas de corrupción son los actos de corrupción, que han causado daños incalculables al momento de presentarse la pandemia, el desvió de fondos causo un retraso que el personal tenga los equipos necesarios de bioseguridad, todos estos actos tomando de escusa la nueva emergencia sanitaria"[93].

Hemos reportado brevemente una comparación de síntomas de depresión y ansiedad entre personas confirmadas y sospechosas de COVID-19 en Ecuador [67], sin "embargo no hemos profundizado en el análisis de las variables relacionadas con la presencia de estos síntomas. La importancia de analizar esos datos para evaluar los síntomas de salud mental y las variables sociodemográficas y conductas asociadas durante el confinamiento de las personas que formaron parte del programa de vigilancia epidemiológica del COVID-19 establecido por el Ministerio de Salud Pública del Ecuador durante marzo y abril de 2020"[59].

Al transcurrir la pandemia los profesionales y autoridades respectivas "aprendieron a enfrentarse al virus y el sistema de salud, aunque colapsado empezó acostumbrarse y a dar resultados en el cuidado del paciente, la prevención en el personal de salud que se encuentra al frente de esta emergencia además de ayudar a presentar las ideas y medidas necesarias que se pueden implementar para manejar esta situación, tanto dentro de los hospitales como en el diario vivir, permitió reducir los casos de contagios intrahospitalario reduciendo el riesgo dentro de los mismos hospitales"[18].

A pesar de todo el virus sigue afectando de manera directa al “sistema de salud porque actualmente sigue sin presentar los recursos suficientes para poder abastecer las diferentes zonas del país, esto se suma a que las medidas implementada y consejos dados a la población no son acatados, lo que conlleva un grave problema y solo genera aumento de casos en vez de reducirlos”[80].

Ecuador, país ubicado en Sudamérica “dividido en cuatro regiones principales 1) la región de la Costa, 2) la región de la Sierra, 3) la región de la Amazon´ıa y 4) la región Insular. La región costera presenta tasas de mortalidad más altas que las demás regiones como consecuencia de la tard´ıa aplicación de medidas restrictivas de distanciamiento social y la limitada capacidad de los servicios sanitarios” [16]. La población de Ecuador se estimó en 17.510.643 habitantes según las últimas proyecciones disponibles para 2020 .

Este seguimiento continuo por parte del “departamento de salud fue el resultado de la estrategia de rastreo de contactos desplegada por el Ministerio de Salud Pública (MSP) entre los pacientes confirmados o sospechosos de COVID-19 en Ecuador. El objetivo del presente estudio fue identificar el impacto que a causado la propagación del COVID-19 en los sistemas de salud de Ecuador.

Según estudios este virus se “transmite a través de got´ıculas de infecciones emitidas al hablar, toser o estornudar directamente sobre una superficie mucosa o conjuntiva de una las gotas flugge es otro medio de transmisión. Este virus posee alta tasa de transmisión fácil y de alto riesgo de contagio, las vías de transmisión son muchas y se incluyen en la motricidad diaria. tenemos a continuación caracter´ısticas de transmisión del virus SARS” [110], las cuales son muy importantes para el conocimiento de las personas que están dentro y fuera de un sistema sanitario:

La propagación del coronavirus puede ocurrir a través del contacto directo con las secreciones respiratorias y la persona infectada puede penetrar en el medio ambiente hablando, tosiendo o estornudando.

Los estudios han demostrado que los virus pueden permanecer en la superficie de los objetos durante horas o incluso días.

Algunos estudios “han establecido que la infección ocurre independientemente si el paciente presenta síntomas de infección. En otras palabras, los casos asintomáticos también pueden transmitir coronavirus porque se encuentra en el cuerpo y se puede

encontrar en sus secreciones, incluso si no causa síntomas de infección”[110]. .

2.3.2. Epidemiología

El cálculo global de la infección por el COVID-19 “ha superado los 85,7 millones de casos confirmados en todo el mundo, mientras que el número de defunciones ronda entre los 1.890.824. A inicios de la pandemia China era el país con más contagio, sin embargo, Estados Unidos el que tiene el mayor número de casos reportados. Italia, España, Alemania, Francia, Irán, Reino Unido o Suiza son otros países gravemente afectados” [82]. .

Ecuador: Según datos del “7 de enero del 2021, más de 217.377 casos acumulados por COVID-19 fueron registrados en Ecuador. Esta enfermedad causada por el virus SARS-CoV-2 fue detectada por primera vez en territorio ecuatoriano el 1 de marzo de 2020. En tanto, la primera muerte ligada al nuevo tipo de coronavirus fue reportada el 14 de marzo de 2020. Diez meses después, el número de personas fallecidas a causa de esta enfermedad ya supera las 14.146” [82]. .

Estados Unidos: Estados Unidos contabilizo este jueves 7 de enero de 2021 más “de 21.4 millones de personas confirmadas de coronavirus y es el país con más casos de coronavirus confirmados. Para interpretar estos datos, conviene saber que Estados Unidos, con 327.352.000 de habitantes, es el tercer país más poblado del mundo, Es por aquello que podemos interpretar mediante los datos estadísticosque la tasa de pacientes confirmados de coronavirus es de 6.359,25 por cada cien mil habitantes, así pues, tiene una alta tasa de confirmados de coronavirus si la comparamos con la del resto de los países. En este momento hay más de 361.483 personas fallecidas por coronavirus, en la última jornada 3.865 personas han muerto, una cifra que supera a lo registrado el martes, con 3 775 muertes” [24]. .

Wuhan: El virus del COVID-19 “fue notificado por primera vez en Wuhan capital de la provincia de Hubei en China el 31 de diciembre de 2019; Donde el número de casos registrados el 7 de enero del 2021 es de 68.149. mientras que el número de personas fallecidas a causa de esta enfermedad ya supera las 4.512. A pesar de la gran cantidad de contagiados al comienzo, China ahora lleva meses con la epidemia prácticamente bajo control, sin presentar ningún fallecimiento desde mediados de mayo” (Wallis et al., 2020) .

2.3.3. Intervención del personal de salud

El COVID-19 ha causado enfermedad en las personas, “las mismas que han acudido a los diferentes centros de salud que posee cada país, esto con la finalidad de poder sobrellevar los signos y síntomas que esta patología puede causar, por este motivo que estos centros de salud se han llegado a saturar en cuanto a su atención debido a que en las salas de emergencias y UCI se han llegado a quedar sin espacio para ingresar a más personas con síntomas de COVID-19”[109].

La causa para esta saturación en la “atención de salud fue debido a que las personas acuden a estos solo por presentar dolor de cabeza con fiebre ya que piensan que son síntomas de COVID-19 , esta situación ha hecho que las personas salgan de sus casas y se expongan a en realidad contraer el virus, ya que a esto hay que sumarle la irresponsabilidad que alguno cometen como es el de salir de casa sin el uso de mascarillas y de no mantener distancia entre personas” [101].

Es por ello por lo que el personal de enfermería actúa desde la promoción de la salud hasta la primera línea hospitalaria y en los distintos niveles asistenciales. Sin embargo, con la realidad impuesta por la pandemia del COVID-19, se ha planificado intervenir en la mejora de las condiciones, fortalecimiento de la formación y desarrollo de los profesionales de enfermería, con el enfoque de liderazgo. Cabe destacar que el proceso enfermero, es sistemático, clínico y racional que permite planificar y brindar cuidados de enfermería al paciente de forma integral e individualizada de calidad, desarrollándose así en cinco fases; tales como:

Valoración aquí es donde se va a comprometer la variabilidad del cuidado, “Diagnóstico donde con los conocimientos y guías de NANDA, NIC, NOC se estandariza un diagnóstico de enfermería, Planificación da como resultado lo que se desea obtener con el plan de cuidados, Ejecución aquí se aplica las actividades e intervenciones planificadas y reconocidas para cada usuario y terminando con Evaluación, donde se va a corroborar si las actividades que realizó tuvo o no eficacia para resolver el conflicto o problema que presentó la persona” [101].

Es así como la enfermería se considera como una “ciencia humanística, sistemática, cuyas prácticas se basan en los fundamentos científicos y la mejor evidencia. A medida de la población y medios han valorado a las enfermeras como héroes de blanco por su continuo trabajo frente a la pandemia en todos los países, ya que se

enfrentan a la angustia, ansiedad, depresión, estrés y soledad, tanto como de los usuarios que se encuentran en las instalaciones de salud, como los mismos profesionales que conforman el equipo de salud, principalmente relacionado a la pérdida de control de la situación donde a diario fallecen decenas de persona con COVID-19" [38], por la rápida propagación del virus, poniendo en riesgo a la misma familia, auto aislándose para otorgarles mayor seguridad en sus hogares. De los cuales también necesitan apoyo emocional y sentirse protegido en su entorno de trabajo.

En relación con los "profesionales de salud públicos y privados se garantiza el acceso a los Equipos de Protección Individual (EPI) adecuados, debate más presente en el primer período de la pandemia - y las regulaciones emitidas (que cambian según el avance de la pandemia), no siempre son viables para su ejecución en los servicios de salud. Situaciones como la falta de espacio y recursos adecuados para la atención, y la infradimensionamiento de los equipos, demuestran la limitada capacidad de respuesta del sistema de salud ante una enfermedad altamente transmisible, en la que hay casos que progresan a una forma grave de la enfermedad" [22].

La calidad y cantidad de EPI ha sido cuestionada diariamente por profesionales que trabajan en la atención directa de sospechosos y pacientes. Los equipos de salud enfrentan dudas sobre el manejo de casos sospechosos y, a menudo, se encuentran perdidos y sin tiempo de trabajo para estudiar notas técnicas, boletines epidemiológicos y recomendaciones.

Según datos del Consejo Federal de Enfermería, "la cantidad de profesionales de enfermería infectados es alta, con más de 20 mil bajas por enfermedad acumuladas hasta junio de 2020, y se sabe que la cantidad de casos está infrarreportada. Es una realidad que genera indignación y tristeza, ya que el cuidado es central en nuestro trabajo, que es fundamental para salvar vidas" (Loveday et al., 2020).

2.3.4. Medidas de prevención básicas

Incluso ahora que existe una vacuna para prevenir la enfermedad del coronavirus, se deben tomar medidas preventivas para evitar contagiar a los demás y contagiarse como son las siguientes: Lavarse las manos frecuentemente: Se recomienda lavarse las manos antes de comer o preparar la comida, tocarse la cara, después de ir al baño, salir de lugares públicos, después de sonarse la nariz, toser o estornudar, entre otras.

Otra opción factible es usar alcohol al 70 Evite el contacto directo: Se debe mantener una distancia de al menos 1 metro entre cada persona cuando se frecuente un lugar concurrido por muchas personas.

Cubrirse la boca y nariz: Es importante utilizar Mascarillas como las quirúrgicas y N95 y el constante lavado de manos como medidas. Es importante que la mascarilla una vez colocada no sea retirada hasta no tenerlas manos limpias como medida de prevención, se debe utilizar en ámbitos públicos y en especial cuando se encuentran en lugares es difícil mantener el distanciamiento.

Cúbrase la Nariz y la boca al toser y estornudar: Es importante cubrirse la nariz y la boca al toser y estornudar con un pañuelo desechable o cúbrase con la parte interna del codo y no escupa. Una vez utilizado el pañuelo deséchelos a la basura y lávese las manos inmediatamente con agua y jabón por al menos 20 segundos. Limpie y desinfecte: Limpiar diariamente los materiales que se toca diariamente. Utilice agua y detergente antes de desinfectarlas. Luego, use un desinfectante de uso doméstico para así mantener las zonas de uso diario sucias y evitar contagio.

Lavar bien los productos: Se recomiendo lavar los alimentos que se traen del exterior, esto se realiza ya que estos alimentos son manipulados por varias personas antes de que el consumidor lo adquiera.

Monitoree su salud a diario: Las personas debes mantenerse atentas a los síntomas que presentan y tratar de dejar la vida sedentaria y empezar a tener una vida activa realizando actividad física y comiendo saludablemente ya que estos factores ayudarán a que la persona tenga un sistema inmunológico activo y fuerte.

Evitar Viajar: Cabe destacar que viajar aumenta el riesgo de contagio. Realizar pruebas de diagnóstico antes y después del viaje puede reducir el riesgo de propagar COVID-19 . Después de viajar, considere volver a realizar la prueba de diagnóstico.

Informarse sobre el COVID-19: Se recomienda informarse acerca de la situación del COVID-19 , asegurándose que la información venga de fuentes confiables como la agencia de salud pública locales o nacional, la página oficial de la OMS o OPS, en caso de no contar con internet o dispositivos informarse de los comunicados por medios de comunicación, ya que así se mantendrá al d´ıa con información relevante.

Es importante implementar las medidas necesarias para poder minimizar el contagio por el contacto entre los trabajadores y los clientes o público que puedan

concurrir en el lugar de trabajo. Establecer planes de continuidad de la actividad ante un aumento de las bajas laborales del personal o también en un contexto que presente incremento de riesgo de transmisión en el lugar de trabajo.

Se puede considerar el realizar 2 veces a la semana trabajo presencial y los demás días llevar a cabo el teletrabajo para evitar estar en contacto con otras personas del entorno laboral. En los establecimientos de atención al público es necesario tomar en consideración que el aforo de personas sea el dictado por las autoridades sanitarias, tomar la temperatura a cada persona que ingrese y exigir el uso de la mascarilla, manteniendo la distancia de 1 metro de distancia, en caso de que existan sillas dentro del establecimiento las personas deben sentarse saltando una silla como m´ınimo para evitar el contacto con otra persona. Se debe informar claramente al público sobre las medidas organizativas y su obligacion de cooperar en su cumplimiento. Medidas de protección colectiva Se recomienda implantar barreras físicas de separación como lo son: mamparas, ventanillas, cortinas transparentes, cinta para que no se sienten las personas juntas en transportes públicos. Delimitación y mantenimiento de distancia de mostradores, ventanillas de atención, etc. Medidas de protección personal El implemento de los Equipos de Protección Personal es lo más recomendado para evitar el contagio. La aplicación de todas estas medidas puede proporcionar un grado adicional de protección. De acuerdo con datos que son proporcionados por cada uno de los representantes de cada país han demostrado que el coronavirus es una patología que ha ido en aumento notablemente y junto a la mortalidad en gran cantidad, ese aumento notable de contagios se debe a que los ciudadanos no han tomado conciencia, algunos no tienen el conocimiento suficiente acerca de las medidas preventivas y las medidas de bioseguridad que deberían de realizar y acoplarlo en la vida cotidiana de cada persona.

Basándose en los “datos de casos confirmados, recuperados y fallecimientos centrándose en Ecuador, EE, UU y Wuhan, China. En donde se observa que el país con más contagios y considerados uno de los países con más contagios es EE. UU. Aunque Ecuador y Wuhan, China son considerados datos altos en donde depende su nivel poblacional para ello el gobierno de cada país tiene sus medidas de prevención contra el COVID-19 y que la población no siga contagiando, aunque eso no fue suficiente el personal médico sufrió graves consecuencias a inicios de la pandemia en donde se

registran casos confirmados, recuperados y hasta de muertes por la falta de experiencia y la saturación de personal medico" (Rodler, Apfelbeck, Stief, Heinemann, & Casuscelli, 2020).

En base a este aumento descontrolado de contagios que se ha dado a nivel mundial se demuestra la importancia que tienen las estrategias de prevención para combatir y disminuir el contagio por Covid- 19, cabe destacar que si se logra hacer que las personas hagan uso de estas medidas durante su vida diaria se va a lograr obtener resultados favorables como la disminución de los contagios e incluso disminuirá la mortalidad en las personas.

Se debe destacar que para la disminuir la transmisión de persona a persona de este virus se debe lavar las manos, usar mascarilla, mantener distancia social con otras personas, realizar actividad física, comer saludable y acatar las disposiciones que dictamina el gobierno a través del MSP.Datos estadísticosactuales de la cantidad de contagiados en los siguientes países: Ecuador, EE. UU y Wuhan, China

Ecuador: Según datos del 21 de enero del 2021, se registró más de 243.377 casos acumulados por COVID-19 . Después de 10 meses de la aparición de los primeros casos, el número de personas fallecidas a causa de esta enfermedad ya supera las 14.437. Además. de los casos recuperados se registran aproximadamente 199.001. Estados Unidos: Estados Unidos contabilizo este jueves 21 de enero de 2021 más de 24.5 millones de personas confirmadas de coronavirus y es el país con más casos de coronavirus confirmados. Para interpretar estos datos, conviene saber que Estados Unidos, con 327.352.000 de habitantes, es el tercer país más poblado del mundo, Es por aquello que podemos interpretar mediante los datos estadísticosque la tasa de pacientes confirmados de coronavirus es de 6.359,25 por cada cien mil habitantes, así pues, tiene una alta tasa de confirmados de coronavirus si la comparamos con la del resto de los países. En este momento hay más de 361.483 personas fallecidas por coronavirus, en la última jornada 3.865 personas han muerto, una cifra que supera a lo registrado el martes, con 3 775 muertes. Aunque, se ha registrando3.021.252 casos recuperados.

Wuhan: El virus del COVID-19 fue notificado por primera vez en Wuhan capital de la provincia de Hubei en China el 31 de diciembre de 2019; donde el número de casos registrados el 21 de enero del 2021 es de 88.701. Mientras que el número de

personas fallecidas a causa de esta enfermedad ya supera las 4.635. Cabe destacar que a pesar de que China fue uno de los epicentros de COVID-19 en los últimos meses ha presentado prácticamente bajo control, sin registrar además ningún fallecimiento desde mediados de mayo. Incluyendo 82.468 de casos recuperados por COVID-19 .

Por lo que podemos observar en la tabla el país con más casos confirmados de COVID-19 es EE. UU ya que es contado como uno de los países con más casos confirmados en comparación de Ecuador y China no tiene cifras tan elevadas.

Datos estadísticosde personal de salud muertos o contagiados por COVID-19 en ecuador Los datos recolectados son hasta abril del año pasado en donde se registra los datos estadísticosde contagios y muertes del personal de salud, El dato muestra que 2.469 son de los casos confirmados de COVID-19 en general del personal de salud, 3.495 casos descartados y 21 fallecidos registrados hasta el mes de abril del año 2020 de los que se han podido registrar.

La labor del profesional de enfermería y de todo el personal de salud es indispensable para el control de la propagación del COVID-19 y la educación que debe dar a las personas por medio de la promoción del uso de los EPP, pero pese a su ardo trabajo y al de los gobiernos por contener la enfermedad, hay millones de casos alrededor del mundo y cada día suman más las muertes por esta enfermedad.

En los resultados de la investigación se pudo apreciar que los datos del 7 de enero del 2021 se reflejan más de 217.377 casos acumulados por COVID-19 y el número de personas fallecidas a causa de esta enfermedad ya supera las 14.146 en Ecuador.

Por otra parte, Estados Unidos contabilizo en esta misma fecha 21.4 millones de personas confirmadas de coronavirus. En este momento hay más de 361.483 personas fallecidas por coronavirus. Mientras que en Wuhan capital de la provincia de Hubei en China, donde el virus del COVID-19 fue notificado por primera vez el 31 de diciembre de 2019, el número de casos registrados el 7 de enero del 2021 es de 68.149. mientras que el número de personas fallecidas a causa de esta enfermedad ya supera las 4.512. pero en las últimas semanas a pesar de algunos rebrotes esporádicos, China afirma que lleva meses con la epidemia prácticamente bajo control, según sin registrar además ningún fallecimiento desde mediados de mayo. Todo esto nos lleva a la siguiente interrogante ¿Son reales las cifras de contagio y número de muertos por coronavirus en Wuhan?

El desconocimiento que se genera ante una nueva patología es un factor que influye en la propagación de dicha patología. Precisamente, este fue el caso del COVID-19, que generó que la poblacion impulsivamente realice compras masivas, consuma los recursos de salud pública, y aumente el contagia. Además, que la falta de protocolos en los Estados, y el desconocimiento del personal de salud acerca de cómo enfrentar esta enfermedad, agravó más la situación.

La información oportuna y correcta a la población, permite que la sociedad actúe eficazmente ante cualquier situación. Por consiguiente, para que la población contribuya a la contención de la enfermedad es preciso que el Estado le brinde la información necesaria que permita que los ciudadanos actúen preventivamente.

Asimismo, las medidas preventivas, no están exclusivamente limitadas al área de salud. Más bien, es un conjunto de políticas, protocoles, y medidas del Estado y de la Ciudadanía, en que ambos actúan de forma responsable y con conciencia. Por consiguiente, la función del Estado frente a la emergencia sanitaria es guiadora, directiva, organizativa, informativa y administrativa, mientras los ciudadanos deben tomar sus propias medidas preventivas y acatar lo más posible de las disposiciones del Estado.

Por lo tanto, para prevenir y disminuir los contagios del COVID-19 , el Estado y la Ciudadanía, deben de colaborar entre sí. Los protocolos de prevención a nivel mundial (OMS) y nacional (MSP), son fundamentales para alcanzar una contención total de la enfermedad.

2.3.5. Datos de mortalidad acumulada por COVID en los países de América

Tomaremos ahora la tasa acumulada de muertes por COVID en los países de Latinoamérica. Tomamos la tasa acumulada porque la forma de las curvas es más suave que la de la tasa diaria que presenta muchas variaciones locales debido a los procedimientos de recogida de datos en cada uno de los paises.

Los datos de han tomado del repositorio de la Universidad Johns Hopkins delos Estados Unidos de América. Los datos disponibles en este repositorio pueden encontrarse en la siguiente dirección https://github.com/CSSEGISandData/COVID-19.

Dentro de este repositorio hemos seleccionado las datos de fallecimientos acumu-

lados para cada uno de los países. Para el cálculo de las tasas aproximadas hemos usado la población en 2019 tal y como aparece en el anexo sobre población mundial en la Wikipedia.

Tras limpiar las bases de datos y hacer coincidir los paises en ambas bases, hemos seleccionado los países de América para hacer el estudio. Se ha utilizado como fecha final el 29 de Agosto aunque el script de cálculo esta pensado para leer hasta el día en el que se procesa. Para el estudio inicial hemos seleccionado solamente aquellos países que tienen más de 5 millones de habitantes. Hemos eliminado también los primeros 60 días ya que la mayor parte de los países no presentaban todavía ningún caso.

2.3.6. Datos de mortalidad acumulada por COVID en las regiones del Ecuador

Como segundo conjunto de datos para el análisis usaremos la tasa de mortalidad acumulada en las distintas regiones del Ecuador.

Los datos se han obtenido del repositorio Ecuacovid que según su creador se trata de un proyecto que te proporciona un conjunto de datos sin procesar extraído de los informes de la situación nacional frente a la Emergencia Sanitaria por el COVID-19 del Servicio Nacional de Gestión de Riesgos y Emergencias del Ecuador (SNGRE).

La página del proyecto donde pueden descargarse los datos y una descripción del mismo es https://github.com/andrab.

En el inicio del proyecto encontramos la siguiente descripción.

Un proyecto que proporciona un conjunto de datos sin procesar extraído de los informes de la situación nacional frente a la Emergencia Sanitaria por el COVID-19 del Servicio Nacional de Gestión de Riesgos y Emergencias del Ecuador (SNGRE), Ministerio de Salud Pública del Ecuador (MSP), y Registro Civil del Ecuador.

Junto con los datos de mortalidad, podemos encontrar muchos otros como los relacionados con la incidencia, vacunas, etc...

Los datos iniciales son el número acumulado de muertes por día hasta el 23 de Agosto de 20021.

Antes del procesamiento de los datos se ha obtenido la tasa de muertes por cada 100000 habitantes en cada una de las provincias.

2.4. Helicobacter

2.4.1. Introducción

Los resultados de este estudio están publicados en [15].

El bacilo Helicobacter pylori, ha sido establecido como el principal causante de la aparición de la gastritis, las úlceras gastroduodenales, la pérdida del apetito y el cáncer gástrico en un gran número de pacientes, estas patologías aparecen por los daños que causa este bacilo al revestimiento mucoso que protege el estómago y el duodeno. El *H. pylori* se caracteriza por ser móviles por flagelos polares monótricos o lofótricos, Gram negativas, microaerofílicas (lo que les permite existir dentro de la mucosa gástrica ya que tiene baja concentración de oxígeno) y una temperatura óptima de crecimiento de 37°C [3]. Es por ello que las apariciones de patologías asociadas a este bacilo han ido aumentando paulatinamente en los últimos años, hasta el punto de que su prevalencia en los casos de cáncer y ulcera gástricas es notoria y motivo de preocupación por parte de las autoridades sanitarias.

Es por ello que de acuerdo al lugar de la aparición del bacilo *H. pylori*,[44], establecieron su clasificaron de acuerdo a su ubicación en, estableciendo especies gástricas y no gástricas, encontradas estas últimas en intestino e hígado. Es por ello que las bacterias Helicobacter de ubicación gástrica, se caracterizan por la producción de ureasa. Debido a la acción de esta enzima, se genera amoniaco, originando una cubierta alcalina alrededor de la bacteria, lo que le permite sobrevivir en el medioambiente ácido del estómago [41] . Por los cual, este bacilo no solo está presente en las enfermedades gástricas, sino que su asociación con patologías hepáticas, lo cual complica aún más la acción de este bacilo en el cuerpo humano y amplia el rango de acción de este microorganismo y la afección que este le produce a diferentes órganos del cuerpo.

De acuerdo con [87] el bacilo *H. pylori* es un organismo microaerofílico, gram negativo, ureasa, catalasa y oxidasa positivos, de crecimiento lento, en forma de espiral, con dos a seis flagelos que le dan la motilidad necesaria para soportar el peristaltismo gástrico y penetrar en lamucosa del estómago, y está asociada al desarrollo de diferentes enfermedades gastroduodenales. De acuerdo a esta definición, este bacilo solo afecta básicamente a la mucosa gástrica, desconociendo el impacto que este pre-

senta sobre el hígado y el intestino, por lo cual para este autor el mayor impacto lo realiza este bacilo sobre los órganos gástricos y puede desencadenar en la generación de ulceras y cáncer gástrico.

Sin embargo, para [94] definen al *H. pylori* como un bacilo Gram negativo que tiene forma de espiral el cual ha colonizado la mucosa gástrica de los humanos, aunque se ha visto que no todas las personas que se encuentran infectadas por esta bacteria desarrollan síntomas gástricos, este microorganismo se le relaciona con el desarrollo de gastritis, ulceras pépticas, adenocarcinoma gástrico y linfoma tipo MALT debido a esto es de gran importancia detectar la presencia de esta bacteria.

Con lo cual queda establecido que el bacilo *H. pylori*, presenta por su condición de bacteria gram negativa, espirilada y microaerofílica, una acción que condiciona la presencia de la infección bacteriana crónica más común que evidencia la población en la actualidad, motivado a que se presenta aquejando al 60 % de los ciudadanos de las naciones desarrollados y un aproximado del 80 % de la población de los países del llamado tercer mundo. Esta situación se encuentra más incuestionablemente en los países en proceso de desarrollo motivado a las condiciones sanitaria que estos presentan, tanto en sus viviendas como en los puestos de trabajo.

Otra situación que propicia su propagación y la supervivencia del bacilo en el organismo humano lo indica [69] quien agregan que su forma espiral y sus múltiples flagelos le facilitan su penetración y movimiento dentro de la capa mucosa, permitiéndole al organismo escapar del pH extremadamente bajo y de los movimientos peristálticos. A si mismo, [104] señala que estos bacilos producen enzimas proteinasas y lipasas, que les permite obtener nutrientes para su desarrollo, reducir la viscosidad del mucus gástrico y facilitar su movimiento flagelar. Es por ello que el Helicobacter pylori presentan una alta incidencia en pacientes mayores de 20 años, siendo este bacilo uno de los factores que posee una mayor incidencia en la presencia de la gastritis y las úlceras crónicas en esta población, la cual posee como punto común su alto consumo de alimentos en espacios públicos comunes.

Es por ello que se ha determinado un conjunto de factores de riesgo para la población, dentro de los cuales se halla el *H. pylori*, que se considera uno de los más importantes, especialmente en la patogenia de la gastritis crónica con hiperplasia folicular [66]. Se ha establecido que la propagación del bacilo *H. pylori*, se produce

por medio del contacto directo ingresando al organismo por la vía oral-oral o fecal-oral, es decir, por la inadecuada realización de las medidas de higiene y saneamiento ambiental lo cual se ha evidenciado por el aislamiento de la bacteria desde saliva y heces. Lo cual refuerza la teoría de que la no aplicación de las medidas sanitarias en los locales ambulantes por parte de las autoridades en cuanto a las medidas adecuadas de manipulación de alimentos, por ejemplo, generan situaciones comunitarias que atentan contra la salud pública de la Ciudadanía.

Una de las patologías que se presentan en la población por la penetración del bacilo es el cáncer gástrico, que es la cuarta neoplasia maligna más común en el mundo y la segunda causa de muerte por cáncer anualmente, totalizando más de un millón de defunciones por año, siendo el adenocarcinoma del estómago el tumor más frecuente (95 %) [97]. Otra patología que es asociada a este bacilo es la gastritis a lo cual [40] indican que se ha sugerido como un factor etiológico de gastritis crónica. Esta patología aparece por la pérdida de apetito y por ende la ausencia de alimentación regular que debe realizar toda persona, produciendo un aumento de los jugos gástricos y el aumento de la presencia de este bacilo.

En cuanto a las formas de detectar el *H. pylori*, en la actualidad se cuenta con una gran cantidad de pruebas o test, las cuales se encuentran clasificadas en dos grandes grupos que son las pruebas convencionalmente, entre las cuales se encuentran los test de muestras microbiológicas, las histológicas, las pruebas rápidas de aliento de urea, entre otras. Sin embargos estas pruebas presentan una dificultad como lo establecen [8] quienes indican que estos estudios pueden arrojar falsos positivos o en otro caso, cuando la colonización por esta bacteria no es en una proporción elevada estos métodos diagnósticos pueden arrojar falsos negativos. Debido a esto se tornan dispendiosas y la sensibilidad puede ser variable.

Los métodos alternativos utilizados para la detección del bacilo *H. pylori*, son aquello que no son invasivo para los pacientes, creando de esta manera una tranquilidad al momento de la toma de muestra y ejecución del estudio. Es por ello que para [32] establecen que el uso de la técnica de inmunocromatografía en heces para detectar la presencia de este agente, esta prueba no es invasiva y logra identificar de manera cualitativa antígenos del *H. pylori* en las heces de laspersonas. En este sentido One Step (s/f) indica que la prueba casett del laboratorio SD BIOLINE el

cual según su reporte presenta una sensibilidad del 98.4 % y una especificidad del 100 %, siendo catalogado como uno de los métodos menos invasivo y con una buena efectividad con lo cual llega a ser este uno de los métodos de diagnóstico de gran utilidad para la identificar el Helicobacter pylori, por cuanto esta presenta un mejor confort para los pacientes y arroja un resultado confiable de forma acelerada.

Muchas de las enfermedades gástricas generan el adenocarcinoma gástrico (ADCA), el cual es definido por [17] como una de las pocas neoplasias malignas para la cual se ha establecido que agentes infecciosos tienen un reconocido e importante rol etiológico. Mientras que [97] indican que el cáncer gástrico es la cuarta neoplasia maligna más común en el mundo y la segunda causa de muerte por cáncer anualmente, totalizando más de un millón de defunciones por año, siendo el adenocarcinoma del estómago el tumor más frecuente (95 %), esta situación coloca de manifiesto la gravedad de la prevalencia de cáncer gástrico a nivel global.

La revisión sistemática de estudios de casos y controles revela que aproximadamente 65 a 80 % de casos de ADCA no cardial (del estómago distal) son atribuidos a la infección por Helicobacter pylori13. (Talley et al. Ob cit.) En un estudio prospectivo realizado en Taiwán, con un seguimiento de 6.3 años, el cáncer gástrico se desarrolló en 1.3 % de pacientes infectados por *H. pylori* y 0 % en no infectados. [47] . En este sentido se establece que la carcinogénesis gástrica no puede ser sólo explicada por la infección por el *H. pylori*. Existe una marcada variación individual del resultado de la infección por esta bacteria en los pacientes, por lo cual no puede ser atribuida solo a la presencia de este vacilo.

La infección por *H. pylori* se asocia a una compleja interacción de factores genéticos, del medio ambiente (alimentarios) y bacterianos que explican los diferentes resultados a los que se llega con la infección. Un ejemplo de ello, es que existen algunos países con altas prevalencias de *H. pylori* que tienen una baja prevalencia de cáncer gástrico (Hsu Ping et al. ob cit.). En este contexto [98] indica que la infección prolongada por *H. pylori* puede causar cambios irreversibles en la mucosa gástrica, caso en el que puede desarrollarse cáncer gástrico sin la presencia de la bacteria; por lo que sería óptimo erradicar la bacteria antes de la producción de dichas lesiones.

Las enfermedades gástricas producto del bacilo *H. pylori*, se ha convertido en un creciente problema de salud pública para el Ecuador, es por ello que el presente

artículo tiene como fin el determinar el efecto del *H. pylori* en la salud gástrica a los habitantes de la ciudadela Cristo del Consuelo segunda etapa de la cuidad de Milagro.

2.4.2. Metodología

Basado en un estudio transversal, descriptiva y de tipo observacional, para el desarrollo de la misma se contó con una población integrada por 230 personas, quienes son residentes de la ciudadela Cristo de Consuelo Milagro, siendo la población de estudio de tipo finita por lo cual la muestra queda constituida por la misma población. Para la ejecución de la investigación se utilizó como técnica de recolección de datos la encuesta y la observación directa, para luego obtener las muestras biológicas, la detección de *H. pylori* se realizó mediante el método de Elisa en muestras de suero y heces. Como técnica de análisis de datos se empleó la tabulación y el procesamiento de los datos se realizó mediante el programa MultBiplot, SPSS statistics 22 y Microsoft Excel.

Para la obtención de la muestra, se seleccionó a población de estudio y se realizaron reuniones informativas a la cual asistieron los residentes de la ciudadela Cristo de Consuelo Milagro, quienes aceptaron a participar en el estudio. Con estos residentes se procedió a socializar e informar a los procedimientos a realizar tomando en cuenta las indicaciones para una correcta obtención de la muestra.

Así mismo en las reuniones se facilitaron los recolectores de heces y se estableció el lugar y fecha para la obtención y recolección de las muestras de sangre y heces respectivamente, que fueron dos tomas de muestra a la semana por 3 meses en horarios de 8am – 10 am en la plaza de la ciudadela Cristo de Consuelo Milagro.

Detección del antígeno de *H. pylori* en materia fecal, es un inmunoensayo enzimático de fase sólida basado en el principio del sándwich. La placa de microtitulación se cubre con anticuerpos de H. pylori Durante el examen los antígenos son extraídos con la solución extractiva y añadidos a los anticuerpos recubiertos en la placa de micro titulación con los anticuerpos a *H. pylori* del conjugado-enzimático y luego se incuban. Si la muestra contiene antígenos de *H. pylori*, se unirán a los anticuerpos recubiertos en la placa de micro titulación y simultáneamente se unirán al conjugado para formar complejos antígeno-conjugado de anticuerpos inmovilizados

de *H. pylori*. Se lava la placa de micro titulación para retirar los materiales que no se han unido (In Control, 2011).

Ensayo de anticuerpos IgG para Helicobacter pylori, se encuentra basada en la reacción de los anticuerpos IgG de la muestra con el antígeno unido a la superficie de poliestireno. Si la muestra contiene anticuerpos IgG *a H. pylori*, éstos se unirán a los antígenos cubiertos en la placa de microtitulación para formar complejos. Después de la incubación inicial se lava la placa de microtitulación para remover los materiales que no se han ligado. Se añade el conjugado-enzimático de anticuerpos anti-humano IgG y luego se incuba (In Control, 2014).

La revisión sistemática de estudios de casos y controles revela que aproximadamente 65 a 80 % de casos de ADCA no cardial (del estómago distal) son atribuidos a la infección por *H pylori* (Talley et al. ob cit.). En un estudio prospectivo realizado en Taiwán, con un seguimiento de 6.3 años, el cáncer gástrico se desarrolló en 1.3 % de pacientes infectados por *H pylori* y 0 % en no infectados [47].

De los infectados por esta bacteria solo un m´ınimo porcentaje desarrollan cáncer gástrico (2-5 %). La mayoría presentan lesiones benignas. Existe pues una marcada variación individual del resultado de esta infección en los pacientes [88].

2.4.3. Descripción de los datos

La detección bacilo *H. pylori,* actividad realizada en la ciudadela el Cristo de Consuelo El Milagro, este diagnóstico se ejecutó por medio del análisis del antígeno fecal y la identificación de la presencia de Anticuerpo IgG en suero, con estos resultados se establece la relación de la presencia del bacilo *H. pylori* con la incidencia de los casos de cáncer gástricos que padece esta población.

En cuanto a la condición de la presencia del bacilo *H. pylori* en la población de la ciudadela, se estableció que a la presencia del antígeno fecal resultaron positivos el 71 % de la población, este porcentaje indica que 164 pobladores se encuentran infectados por el bacilo *H. pylori,* de los cuales 99 de las personas contagiadas son del sexo femenino, lo que representa el 60 % de los casos, mientras que 65 de los contagiados son del sexo masculino, lo que representa el 40 % de la población contagiada. Con respecto a la presencia anticuerpo IgG en suero, los resultados indican que el 75 % de la población de la ciudadela el Cristo de Consuelo El Milagro lo cual indica que 173

habitantes se encuentran infectados con el bacilo, de los cuales 107 personas son del sexo femenino, lo que representa el 62 % de los casos positivos, mientras 66 personas del sexo masculino arrojaron positivo al examen, siendo el 38 % de los casos registrados. Para determinar el efecto del *H. pylori* en la salud gástrica a los habitantes de la ciudadela Cristo del Consuelo segunda etapa de la cuidad de Milagro, se tomará como la prevalencia de contagio a los detectados por la presencia del anticuerpo IgG en la sangre de los pobladores, el cual es un indicador confiable de la infección que presentan la persona por el bacilo *H. pylori.*

La prevalencia estratificada por variables sociodemográficas, se encuentra establecida por rango de edad de la siguiente manera: que presenta una mayor incidencia en el contagio con el *H. pylori,* es el comprendido entre los 18-30 años, en los cuales los habitantes de la ciudadela del Cristo de los Milagros poseen el 55 % de los casos detectados, siendo distribuido en los casos presentado por los pobladores del sexo femeninos con un 57 % de los mismos, mientras que los pobladores masculinos poseen el 51 % de los casos registrados, seguido la incidencia por el rango comprendido entre los 31-50 años con un 23 % de los casos, estando integrado por el 24 % detectados en los pobladores de sexo femenino y el 21 % encontrados en los pobladores del sexo masculino, finalizando con el grupo etario comprendido de edades desde los 51 años en adelante, los cuales presentan 21 % de los casos, siendo distribuido en un 27 % de los detectados en los pobladores de sexo masculino y un 17 % de pobladores femeninos que se encuentran contagiados .

De acuerdo, a las variables sociodemográficas de la población de la ciudadela el Cristo de Consuelo El Milagro, los habitantes de esta población se encuentra constituido por individuos de raza mestiza en un 78,61 %, de lo cual los pobladores del sexo femenino representan el 7,43 % de esta raza y mientras que los pobladores masculinos conforman el 77.27 % de los habitantes, siendo los pobladores de raza indígena el 12 % de los contagiados con el bacilo, siendo los pobladores femeninos el 10 % de los casos y los pobladores masculinos poseen el 15 % de las prevalencias, mientras que la raza negra poseen el 9 % de los casos registrados.

La variable estado civil de la población de la ciudadela el Cristo de Consuelo El Milagro, indica que el 43.93 % de los habitantes se encuentran casados, de estos pobladores el 45.79 % son del sexo femeninos y mientras que el 40,9 % corresponde

al sexo masculino, en cuanto a los que se encuentran solteros representan el 34.68 % de la muestra, estando constituido por un 38.87 % del sexo masculino y mientras el 32,71 % corresponde al sexo femenino, a la condición de unión libre le corresponde el 7,51 % de los pobladores, estando integrado por el 9.09 % de los habitantes del sexo masculino y con el 6.54 % de las habitantes son del sexo femenino, los habitantes divorciado integra el 5.2 % de los habitantes, siendo los pobladores son del sexo femenino el 5.6 % de la muestra y los habitantes masculinos son el 4.54 % de los estudiados, mientras que los habitantes viudos integran el 8.67 % de la muestra de estudio, con un 5.57 % de los pobladores del sexo masculino y el 9.34 % son del sexo femenino.

La zona de procedencia de los habitantes de la ciudadela el Cristo de Consuelo El Milagro, quedo establecida con un 80.34 % de los pobladores provienen de zonas urbanas, estando integrada por un 72.72 % de los habitantes del sexo masculino y el 85.04 % corresponde a los habitantes del sexo femenino, mientras que el 19.66 % proviene de zonas rurales; correspondiendo el 27.27 % a los habitantes de sexo masculino y el 14.96 % a los pobladores de sexo femenino.

En cuanto al nivel de escolaridad de los habitantes de la ciudadela el Cristo de Consuelo El Milagro, el nivel educativo quedo establecido de la siguiente manera: un 52.60 % de los pobladores, poseen un grado de instrucción de nivel superior, siendo el 50 % de los mismos pobladores del sexo masculino y un 54.2 % de los habitantes del sexo femenino, con respecto al nivel de educación primaria esta lo integran el 14.45 %, siendo el 12.12 % habitantes del sexo masculino y el 14.45 % de los pobladores son del sexo femenino, con respecto a la educación secundaria, este nivel educativo lo poseen el 31.21 % de los pobladores la poseen, de los cuales el 36.36 % son habitantes del sexo masculino y el 28.03 % le corresponde a los pobladores del sexo femenino.

Variable		Sexo				Total			
		Masculino		Femenino					
		N	%	N	%	N	%	IC75 %	
Edad	18 – 30	34	51,51	62	57,94	96	55,49	51,51	57,94
	31 – 50	14	21,21	26	24,29	40	23,12	21,21	24,29
	51 o más	18	27,27	19	17,75	37	21,39	17,75	27,27
Raza	Mestiza	51	77,27	85	79,43	136	78,61	77,27	79,43
	Indígena	10	15,15	11	10,28	21	12,13	10,28	12,13
	Negra	5	7,57	11	10,28	16	9,24	7,57	10,28
Estado Civil	Soltero/a	25	38,87	35	32,71	60	34,68	32,71	38,87

Variable		Sexo				Total			
	Casado/a	27	40,90	49	45,79	76	43,93	40,90	45,79
	Uni´on libre	6	9,09	7	6,54	13	7,51	6,54	9,09
	Divorciado/a	3	4,54	6	5,60	9	5,20	4,54	5,60
	Viudo/a	5	5,57	10	9,34	15	8,67	5,57	9,34
Procedencia	Urbana	48	72,72	91	85,04	139	80,34	60,02	62,79
	Rural	18	27,27	16	14,96	34	19,66	14,96	27,27
Escolaridad	Analfabeta	1	1,51	2	1,86	3	1,73	1,51	1,86
	Primaria	8	12,12	17	15,88	25	14,45	12,12	15,88
	Secundaria	24	36,36	30	28,03	54	31,21	28,03	36,36
	Superior	33	50,00	58	54,20	91	52,60	50,00	54,20

Cuadro 2.1: Prevalencia estratificada por variables sociodemogr´aficas de los habitantes de la ciudadela Cristo del Consuelo segunda etapa de la cuidad de Milagro

Con respecto a los hábitos higiénicos de los pobladores de la ciudadela Cristo de Consuelo Milagro, se establece que en cuanto al consumo de agua potable el 80.92 % de los pobladores la consumen directamente de la tubería, de los cuales los pobladores de sexo masculino representa el 81.81 % de los casos, constituyendo los pobladores de sexo femenino el 80.37 % de los contagios, los pobladores que consumen el agua hervida, está establecido en un 11.56 % de los casos, estado compuesta por un 12.12 % de los pobladores de sexo masculino y los pobladores de sexo femenino constituye el 11.21 % de los casos, y los que consumen agua mineral habitualmente son el 7.51 % de los contagiados, siendo la población de sexo masculino el 6.06 % de los casos y los pobladores femeninos quedo establecido en 8.41 % de los contagios.

En cuanto al hábito de lavarse las manos antes de comer, la población contesto que si se lavas las manos en un 53.75 % de los casos, siendo los habitantes del sexo femenino un 63.55 % de los contagios, mientras que los habitantes de sexo masculinos son el 38.87 % de los casos reportados. Los pobladores que contestaron que a veces lo hacen son el 41.19 %, siendo los casos de los pobladores masculinos el 57.57 %, mientras que los casos de los habitantes femeninos son el 32.71 % de los reportes y los habitantes que no se lavan las manos antes de comer son el 4.04 % de los casos, de los cuales los pobladores masculinos son el 4.54 % de los reportes y los pobladores femeninos son el 3.73 % de los casos.

Mientras que a la pregunta sobre el lugar donde realiza su alimentación regularmente los habitantes de la ciudadela Cristo de Consuelo Milagro, contentaron en

un 58.38 % que lo hacen de manera ambulante, siendo los pobladores femeninos un 65.42 % de los casos y los pobladores masculinos representan un 46.96 % de los reportes, los habitantes que se alimentan en su casa son el 23.12 % de los casos, siendo los habitantes masculinos el 30.30 % de los casos y los pobladores femeninos el 18.69 % de los casos y los que se alimentan en restaurante son el 18.49 % de los casos, siendo las damas el mayor número con un 65.42 % de los casos.

Variable		Sexo				Total			
		Masculino		Femenino					
		N	%	N	%	N	%	IC75 %	
Agua de consumo	Servicio	54	81,81	86	80,37	140	80,92	80,37	81,81
	Hervida	8	12,12	12	11,21	20	11,56	11,21	12,12
	Embotellada	4	6,06	9	8,41	13	7,51	6,06	8,41
Lavado de frutas	Si	25	38,87	68	63,55	93	53,75	38,87	63,55
	No	3	4,54	4	3,73	7	4,04	3,73	4,54
	A veces	38	57,57	35	32,71	73	41,19	32,71	57,57
Consumo de alimentos	Casera	20	30,30	20	18,69	40	23,12	18,69	30,30
	Restaurant	15	22,72	17	15,88	32	18,49	15,88	22,72
	Ambulante	31	46,96	70	65,42	101	58,38	46,96	65,42

Cuadro 2.2: H´abitos higi´enicos de los habitantes de la ciudadela Cristo del Consuelo segunda etapa de la cuidad de Milagro para prevenir el contagio del *H. pylori*

Las afecciones que produce la presencia del *H. pylori*, en los habitantes de la ciudadela Cristo de Consuelo Milagro, se encuentra la pérdida del apetito como el de mayor ocurrencia con un 14.45 % de afecciones en la población, siendo los caballeros los que la sufren de manera mayoritaria con un 8.67 % de los casos, mientras los pobladores del sexo femenino lo sufren en un 5.78 % de la ocurrencia, la gastritis aparece en el 11.55 % de los casos, siendo las mujeres las que la padecen en mayor grado con un 6.93 % de la prevalencia, mientras que los hombres la padecen en un 4.62 % de los casos, las ulceras gástricas aparecen en el 4.32 % de los casos de contagio con el bacilo, siendo los hombres los que presentan una mayor incidencia con un 3.75 % de los casos y las mujeres un 0.57 % de las veces, mientras que el cáncer gástrico se presenta en el 6.35 % de las veces, siendo los habitantes del sexo masculino los que padecen mayormente con un 4.62 % de las ocasiones y los pobladores femeninas solo un 1.73 % de las prevalencias.

En la parte correspondiente al biplot haremos un análisis multivariante de los

Patología	Sexo	Frecuencia	Prevalencia %
Pérdida de apetito	Masculino	15	8,67
	Femenino	10	5,78
Gastritis	Masculino	8	4,62
	Femenino	12	6,93
Úlcera gástrica	Masculino	3	3,75
	Femenino	1	0,57
Cáncer gástrico	Masculino	8	4,62
	Femenino	3	1,73

Cuadro 2.3: Afecciones que produce el bacilo H. pylori

datos obtenidos como ilustración del método.

Capítulo 3

Análisis de Componentes Principales

3.1. Introducción

El Análisis de Componentes Principales, cuyos orígenes se remontan a [83] y [46], es probablemente la técnica multivariante descriptiva más utilizada en la práctica debido a que es cada vez más recuente encontrar grandes conjuntos de datos, con un número elevado de variables, que necesitan una reducción drástica de la dimensión reteniendo la mayor parte de la información.

Aunque su desarrollo se hizo hace muchos años, no ha sido hasta la aparición de los ordenadores que se utilizado de forma masiva. Una búsqueda rápida en google scholar proporciona un total de 1030000, de las cuales 13900 son del último año.

Hay una amplia literatura sobre el tema aunque quizás los libros clásicos más populares son [56] o [57].

Las aplicaciones son innumerables y se extienden por prácticamente todos los campos de la ciencia, en particular en demografía.

3.2. Definiciones básicas

Disponemos de una matriz $\mathbf{X}_{n\times p}$ que contiene las medidas de p variables tomadas sobre n individuos. Para simplificar el resto de la exposición supondremos, sin pérdida

de generalidad, que las columnas de $\mathbf{X}$ tienen media cero, es decir que se le ha restado la media.

Todas las variables tienen el mismo papel, es decir, el conjunto no se divide en variables dependientes e independientes como en el caso de la regresión.

El Análisis de Componentes principales consiste en encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables incorreladas, denominadas *Componentes Principales*, que se obtienen en orden decreciente de importancia.

Las componentes son combinaciones lineales de las variables originales y se espera que, solo unas pocas (las primeras) recojan la mayor parte de la variabilidad de los datos, obteniéndose una reducción de la dimensión en los mismos. Luego el propósito fundamental de la técnica consiste en la reducción de la dimensión de los datos con el fin de simplificar el problema en estudio.

Se trata de una técnica orientada a las variables, suponemos que las p columnas de X generan un espacio p dimensional, de forma que los n individuos pueden representarse en dicho espacio en lo que llamaremos una hipernube. La transformación es, de hecho, una rotación en el espacio p-dimensional. El espacio generado por las primeras q componentes es entonces, un subespacio vectorial q-dimensional del espacio p-dimensonal original.

Cuando el valor de q es pequeño, por ejemplo 2, es posible una representación gráfica directa de los individuos que nos ayudará ainterpretar las similitudes entre los mismos.

El ACP puede entenderse también como la búsqueda del subespacio de mejor ajuste.

Una de las diferencias fundamentales con el Análisis Factorial es que el ACP explica variabilidad en lugar de correlaciones, aunque para obtener una reducción efectiva de la dimensión es necesario que las variables estén correlacionadas. En otras palabras, si las variables están altamente correlacionadas, tienen información común y la dimensión real de los datos es menor que p.

En muchas ocasiones es difícil encontrar el significado de las componentes, como variables compuestas, por lo que el uso principal de la técnica es la reducción de la dimensión como paso previo a la aplicación de otros análisis posteriores, por ejemplo,

un diagrama de dispersión de las primeras componentes con el objeto de encontrar "clusters" en los datos o con el objeto de contrastar similitudes o diferencias entre los individuos.

El ACP es una técnica que no necesita que se especifique un modelos concreto para explicar el "error", en particular, no se hace ninguna suposición sobre la distribución de probabilidad de las variables originales, aunque si se supone que es normal multivariante es posible obtener algunos resultados inferenciales adicionales.

En algunos textos se hacen diferencias entre las CP poblacionales y muestrales, aquí entenderemos la técnica como un método descriptivo, libre de distribución, y trabajaremos directamente con los datos muestrales.

3.3. Obtención de las Componentes Principales

La obtención de las CP puede realizarse por varios métodos alternativos:

1. Buscando aquella combinación lineal de las variables que maximiza la variabilidad. (Hottelling).
2. Buscando el subespacio de mejor ajuste por el método de los Mínimos cuadrados. (Minimizando la suma de cuadrados de las distancias de cada punto al subespacio). (Pearson).
3. Minimizando la discrepancia entre las distancias euclídeas entre los puntos calculadas en el espacio original y en el subespacio de baja dimensión. (Coordenadas principales, Gower).
4. Mediante regresiones alternadas (métodos Biplot)

Describiremos aquí la primera opción aunque más tarde lo haremos utilizando el método 4. En todas se obtienen resultados equivalentes.

3.3.1. Obtención de las CP mediante la maximización de la variabilidad

Denotaremos con $(X_1, \ldots, X_p)$ las variables originales y con $(Y_1, \ldots, Y_tp)$ las componentes. En principio, podemos obtener tantas componentes como variables origi-

nales. X denotará el vector de variables originales e Y el de componentes.

$\mathbf{X}$ es la matriz de datos originales, que supondremos centrada por columnas, y $\mathbf{S}$ es la matriz de covarianzas entre las variables:

$$\mathbf{S} = (n-1)^{-1}\mathbf{X}^{\prime}\mathbf{X}$$

Buscamos combinaciones lineales de las variables observadas

$$Y_j = \upsilon_{1j}X_1 + \ldots + \upsilon_{pj}X_p$$

que sean incorreladas y con varianzas progresivamente decrecientes. En forma matricial

$$Y_j = X\mathbf{v}_j$$

donde $\mathbf{v}_j = (\upsilon_{1j}, \ldots, \upsilon_{pj})^{\prime}$ es el vector de coeficientes de la j-ésima componente. Para los datos muestrales

$$\mathbf{y}_j = \mathbf{X}\mathbf{v}_j$$

donde $\mathbf{y}_j = (y_{1j}, \ldots y_{nj})^{\prime}$ es el vector que contiene las puntuaciones (coordenadas) de cada uno de los individuos de nuestra muestra sobre la componente.

Si colocamos los vectores $\mathbf{y}_j, j = (1 \ldots, p)$ como columnas de una matriz $\mathbf{Y}$ y los vectores $\mathbf{v}_j, j = (1 \ldots, p)$ como columnas de una matriz $\mathbf{V}$, tenemos

$$\mathbf{Y} = \mathbf{X}\mathbf{V}$$

Y_1 será aquella componente que explique la mayor parte de la variabilidad, Y_2 será ortogonal a Y_1 y explicará la mayor parte de la variabilidad restante y así sucesivamente.

Buscamos entonces la componente Y_1 que haga máxima la varianza

$$Var(Y_1) = Var(\mathbf{X}\mathbf{v}_1) = \mathbf{v}_1^{\prime}\,\mathbf{S}\mathbf{v}_1$$

En las ecuaciones tenemos un factor de escala arbitraria, es decir, si multiplicamos $\mathbf{v}_1$ por un escalar, podemos hacer la varianza arbitrariamente grande, es por lo que

imponemos la restricción

$$\mathbf{v}_1{}^{l}\mathbf{v}_1 = 1$$

es decir, $\mathbf{v}_1$ es un vector unitario en la dirección de la componentes principal.

Utilizando el método de los multiplicadores de Lagrange para tener en cuenta la restricción, podemos escribir la función a maximizar como

$$L(\mathbf{v}_1) = \mathbf{v}_1{}^{l}\mathbf{S}\mathbf{v}_1 - \lambda(\mathbf{v}_1{}^{l}\mathbf{v}_1 - 1)$$

Derivando e igualando a cero

$$\frac{\partial L(\mathbf{v}_1)}{\partial \mathbf{v}_1} = 2\mathbf{S}\mathbf{v}_1 - 2\lambda\mathbf{v}_1 = 0$$

es decir

$$\mathbf{S}\mathbf{v}_1 = \lambda\mathbf{v}_1$$

lo que quiere decir que $\mathbf{v}_1$ debe ser un vector propio de $\mathbf{S}$ de valor propio λ. Pero $\mathbf{S}$ tiene p valores propios $\lambda_1, \ldots, \lambda_p$ que supondremos distintos y ordenados en orden decreciente $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$.

Teniendo en cuenta que

$$Var(\mathbf{X}\mathbf{v}_1) = \mathbf{v}_1{}^{l}\mathbf{S}\mathbf{v}_1 = \mathbf{v}_1{}^{l}\lambda\mathbf{v}_1 = \lambda$$

λ debe ser λ_1 el primer valor propio y $\mathbf{v}_1$ el vector propio asociado.

La segunda componente principal $Y_2 = \mathbf{X}\mathbf{v}_2$, se obtiene con un procedimiento análogo pero añadiendo la restricción adicional de que Y_1 e Y_2 sean incorreladas.

$$Cov(Y_1, Y_2) = \mathbf{v}_2{}^{l}\mathbf{S}\mathbf{v}_1 = 0$$

o una condicion equivalente más simple $\mathbf{v}_2{}^{l}\mathbf{v}_1 = 0$ ya que $\mathbf{S}\mathbf{v}_1 = \lambda\mathbf{v}_1$

Utilizando de nuevo el método de los multiplicadores de Lagrange, podemos escribir

$$L(\mathbf{v}_2) = \mathbf{v}_2{}^{l}\mathbf{S}\mathbf{v}_2 - \lambda(\mathbf{v}_1{}^{l}\mathbf{v}_1 - 1) - \delta\mathbf{v}_2{}^{l}\mathbf{v}_1$$

Derivando e igualando a cero se obtiene

$$\frac{\partial L(\mathbf{v}_2)}{\partial \mathbf{v}_2} = 2\mathbf{S}\mathbf{v}_2 - 2\lambda\mathbf{v}_2 - \delta\mathbf{v}_1{}^{\prime}$$

premultiplicando por $\mathbf{v}_1{}^{\prime}$,

$$2\mathbf{v}_1{}^{\prime}\mathbf{S}\mathbf{v}_2 - 2\lambda\mathbf{v}_1{}^{\prime}\mathbf{v}_2 - \delta\mathbf{v}_1{}^{\prime}\mathbf{v}_1{}^{\prime} = 0$$

tenemos

$$2\mathbf{v}_1{}^{\prime}\mathbf{S}\mathbf{v}_2|\delta = 0$$

como $\mathbf{v}_1{}^{\prime}\mathbf{S}\mathbf{v}_2 = 0$, entonces $\delta = 0$ en el punto estacionario, de forma que ,

$$\mathbf{S}\mathbf{v}_2 = \lambda\mathbf{v}_2$$

con lo que λ es el segundo valor propio λ_2 y $\mathbf{v}_2$ es el segundo vector propio. Siguiendo con el mismo argumento, podemos obtener las sucesivas componentes principales a partir de los correspondientes valores y vectores propios.

Entonces, si

$$\mathbf{S} = \mathbf{V}^{\prime}\mathbf{\Lambda}\mathbf{V}$$

donde $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$, es la descomposición espectral de la matriz de covarianzas $\mathbf{S}$, los coeficientes de las combinaciones lineales que definen las componentes principales son las columnas de $\mathbf{V}$, es decir los vectores propios de la matriz de covarianzas.Seleccionando q componentes, las puntuaciones (coordenadas) de los individuos en las componentes están dadas por

$$\mathbf{Y}_{(q)} = \mathbf{X}\mathbf{V}_{(q)}$$

donde el subíndice (p) significa las q primeras columnas de la matriz correspondiente.

Ahora

$$\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} \lambda_j = traza(\mathbf{\Lambda})$$

y

$$traza(\mathbf{\Lambda}) = traza(\mathbf{V}^{\prime}\mathbf{S}\mathbf{V}) = traza(\mathbf{S}\mathbf{V}^{\prime}\mathbf{V}) = traza(\mathbf{S}) = \sum_{j=1}^{p} Var(X_j)$$

es decir, la varianza total de las componentes y de las variables observadas es la misma. Este resultado nos permite calcular la proporción de varianza absorbida por cada componente como

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

o acumulada para un subespacio de dimensión q

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

Es posible también calcular las componentes principales a partir de $\mathbf{X}^{\prime}\mathbf{X}$ con $\mathbf{X}$ centrada ya que se obtienen los mismos vectores propios aunque los correspondientes valores propios aparecieran multiplicados por $n - 1$, lo cual no influye en la variabilidad absorbida.

Como se trata de una aproximación en dimensión reducida es necesario decidir si la aproximación es satisfactoria definiendo la bondad del ajuste de la aproximación.

Las coordenadas de las proyecciones sobre el subespacio de las componentes en el sistema de referencia original son

$$\hat{\mathbf{X}} = \mathbf{Y}\mathbf{V}_{(q)}^{\prime} = \mathbf{X}\mathbf{V}_{(q)}\mathbf{V}_{(q)}^{\prime}$$

Estas coordenadas pueden entenderse también como la aproximación de los valores iniciales en dimensión reducida. La discrepancia con los valores originales en $\mathbf{X}$ y los valores esperados $\hat{\mathbf{X}}$ en el subespacio se puede medir como la suma de cuadrados de

$(\mathbf{X} - \hat{\mathbf{X}})$, es decir, como

$$traza[(\mathbf{X} - \hat{\mathbf{X}})^{\prime}(\mathbf{X} - \hat{\mathbf{X}})]$$

o en forma relativa

$$\frac{traza[(\mathbf{X} - \hat{\mathbf{X}})^{\prime}(\mathbf{X} - \hat{\mathbf{X}})]}{traza[\mathbf{X}^{\prime}\mathbf{X}]}$$

luego, una medida de la bondad del ajuste puede ser

$$1 - \frac{traza[(\mathbf{X} - \hat{\mathbf{X}})^{\prime}(\mathbf{X} - \hat{\mathbf{X}})]}{traza[\mathbf{X}^{\prime}\mathbf{X}]} \times 100$$

que puede interpretarse como el porcentaje de la variabilidad de los datos explicado por las componentes principales.

Teniendo en cuenta las propiedades de la traza, la bondad del ajuste puede escribirse también como

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{i=1}^{r} \lambda_i} \times 100$$

3.4. Algunas propiedades de las componentes principales

- La matriz de vectores propios V define un cambio de base del espacio $I\!R^p$ en el que se ha representado la matriz de datos originales.
- Las q primeras columnas de $\mathbf{V}$ definen la proyección de los puntos en $I\!R^q$ sobre el subespacio q-dimensional de mejor ajuste.
- Los elementos de $\mathbf{V}$ son los cosenos de los ángulos que forman las variables originales y las componentes principales.
- Las coordenadas de los individuos en el nuevo sistema de referencia son de la forma $\mathbf{Y} = \mathbf{XV}$.
- Las coordenadas sobre las primeras componentes principales permiten interpretar las similaridades entre individuos con pérdida de información mínima.

- El ACP utiliza la información redundante, a través de las correlaciones entre las variables, para reducir la dimensión.
- La matriz de covarianzas entre las componentes es $\mathbf{\Lambda}$.
- Las componentes principales son variables incorreladas y, por tanto contienen aspectos independientes de la información.
- La varianza de las componentes principales es λ_i.
- Si se trabaja con datos brutos, la primera componente principal suele mostrar la traslación de la nube de puntos con respecto al origen.
- Si las variables están centradas, las componentes se calculan a partir de la matriz de covarianzas y las componentes estarán dominadas por las variables con escala de medida mayores. Es adecuado cuando las escalas de medida de las variables son comparables.
- Si se trabaja con datos estandarizados, las componentes principales se obtienen de la diagonalización de la matriz de correlaciones. Se utilizarán datos estandarizados cuando las escalas de medida de las variables sean muy diferentes.

3.5. interpretación del Análisis de Componentes Principales

A continuación mostramos algunas reglas de interpretación del análisis de componentes principales

- Diagramas de dispersión que representan los valores de los individuos en las primeras componentes principales.
- Interpretación de distancias en términos de similitud.
- Búsqueda de clusters (grupos) y patrones en los individuos.
- Interpretación de las componentes utilizando las correlaciones con las variables originales. Las posiciones de los individuos se interpretan después en relación a la interpretación dada a las componentes.

A los vectores escalados de la forma:

$$\mathbf{c}_j = \lambda_j \mathbf{v}_j$$

o en forma matricial

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}^{1/2}$$

se les denomina factores de carga (**C**)

Cuando las componentes principales se calculan usando la matriz de correlaciones, la matriz **C** contiene las correlaciones entre las variables originales y las componentes.

Para las componentes calculadas a partir de la matriz de covarianzas, los factores de carga dependen de la escala de medida de las variables por lo que son difíciles de interpretar.

Los factores de carga suelen representarse en un gráfico que permite la interpretación visual de las relaciones. Cuando son correlaciones se obtiene el denominado círculo de correlaciones que contiene información sobre la estructura de las componentes.

En cualquiera de los casos podemos calcular también la correlación al cuadrado entre las componentes y las variables y las componentes. A dichas correlaciones al cuadrado se las denomina contribuciones relativas del factor al elemento y miden la proporción de la variabilidad de las variables explicadas por cada componente. Esta cantidad puede utilizarse para interpretar las componentes.

3.6. Ejemplo: Causas de Muerte en el Ecuador

Como ilustración al ACP usaremos los datos de causas de muerte en el Ecuador descritos antes. Haremos solamente una breve descripción de los resultados ya que los mismos datos serán analizados después mediante técnicas funcionales. Los describimos aquí a efectos de comparación.

disponemos de una matriz de datos con 69 filas, correspondientes a cada una de las causas de muerte descritas en el capítulo de datos y 22 columnas correspondientes a los años desde 1997 a 2020. Los datos del último año eran provisionales y presentaban

algunas irregularidades sí que fue finalmente eliminado del estudio.

Sobre los datos se realizó un ACP estandarizando por columnas que es lo más habitual, si bien en este caso no sería completamente necesario ya que los datos son comparables y medidos en la misma escala.

La tabla siguiente contiene la variabilidad explicada por las dos primeras componente principales.

	Eigenvalue	Exp. Var	Cummulative
1	1409.33	92.84	92.84
2	77.68	5.12	97.96

Cuadro 3.1: Varianza explicada por las dos primeras componentes

La proyección de las causas de muerte sobre el primer plano principal se muestra en la figura 3.1. Debido a la diferente magnitud de las distintas causas de muerte, las mayoritarias se distinguen dentro del gráfico y se encuentran en la parte negativa de la primera componente. A la simple vista de los resultados parece que la primera componente es de tamaño, está negativamente correlacionada con la magnitud de las causas. Curiosamente la clasificación más frecuente es *"99 Causas mal definidas"* y *"88 Resto de las causas"* los que significa que la mayor parte de los registros de defunción no son muy específicos.

Si seguimos observando las proyecciones sobre la primera componente vemos que las causas *"26 Diabetes Mellitus"*, *"42 Enfermedades cardiovasculares"*, *"35 Enfermedades isquémicas del corazón"*, *"46 Influenza y Neumonía"*, *"34 Enfermedades hipertensivas"* y *"57 Accidentes de transporte terrestre"* son, al menos en media, las causas más frecuentes.

Le sigue un grupo formado por las causas, *"41 Insuficiencia cardíaca, complicaciones y enfermedades mal definidas"*, *"64 Agresiones (Homicidios)"*, *"55 Afecciones originadas en el periodo prenatal"*, *"51 Cirrosis y otras enfermedades del hígado"*, *"53 Enfermedades del sistema urinario"* y *"9 Neoplasia maligna de estómago"*. Observando las mencionadas la mayor parte están relacionadas con enfermedades del corazón, hipertensión y diabetes. Cabe destacar la posición relativamente elevada de los homicidios y de los accidentes de tráfico.

El resto de las causas aparecen juntas en el gráfico y cercanas al centro de gravedad.

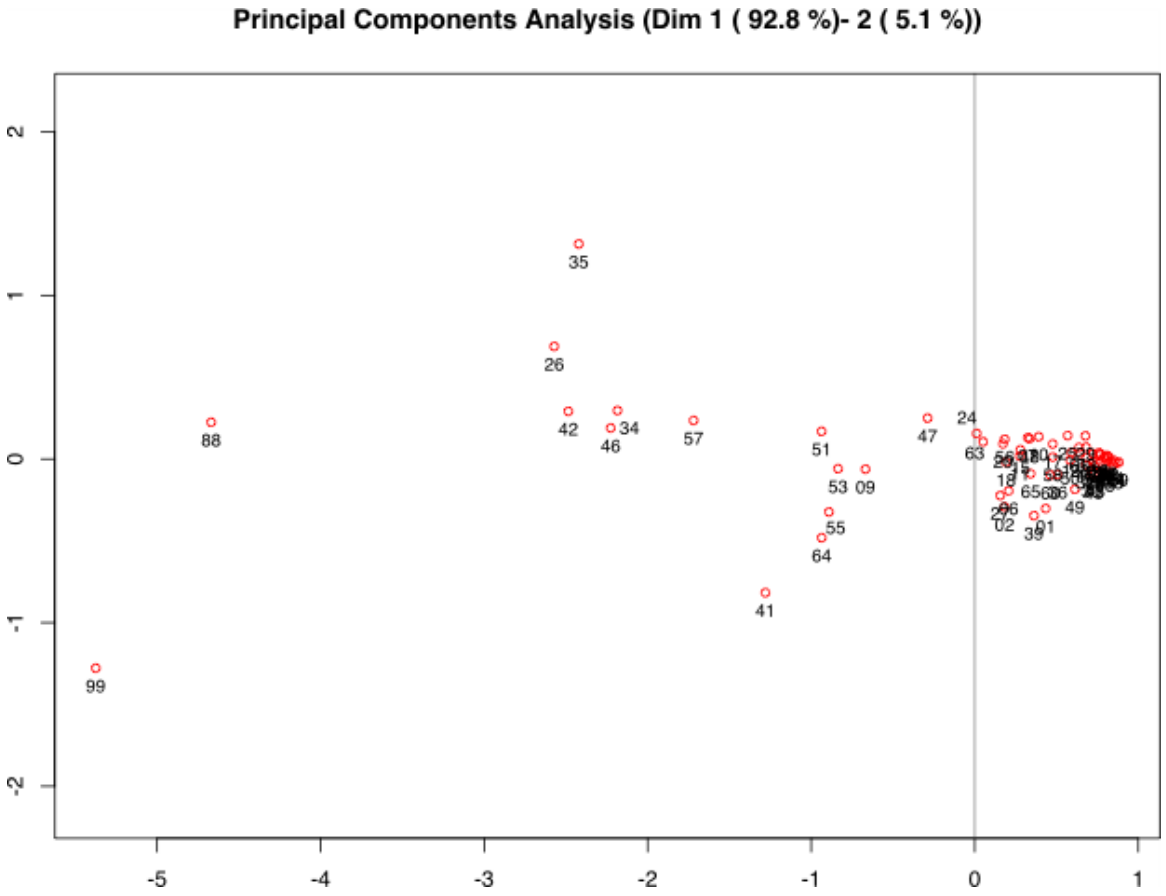


Figura 3.1: Causas de Muerte en el Ecuador: Proyección de las causas sobre las componentes principales

La figura 3.2 muestra una ampliación de la zona central del gráfico. Se trata de causas menos frecuentes y no las describiremos aquí con detalle.

La tabla 3.2 muestra las correlación de las variables (años) con las componentes principales y la figura 3.3, la misma información en forma gráfica. La primera componente principal está altamente correlacionada con todos los años de forma negativa, se puede interpretar como la magnitud media de las distintas causas, por tanto las tasas por las distintas causas son similares a lo largo del tiempo. Esto quiere decir que todos los coeficientes del primer vector propio tienen el mismo signo lo que traduce el hecho de que todas las correlaciones entre los años son positivas.

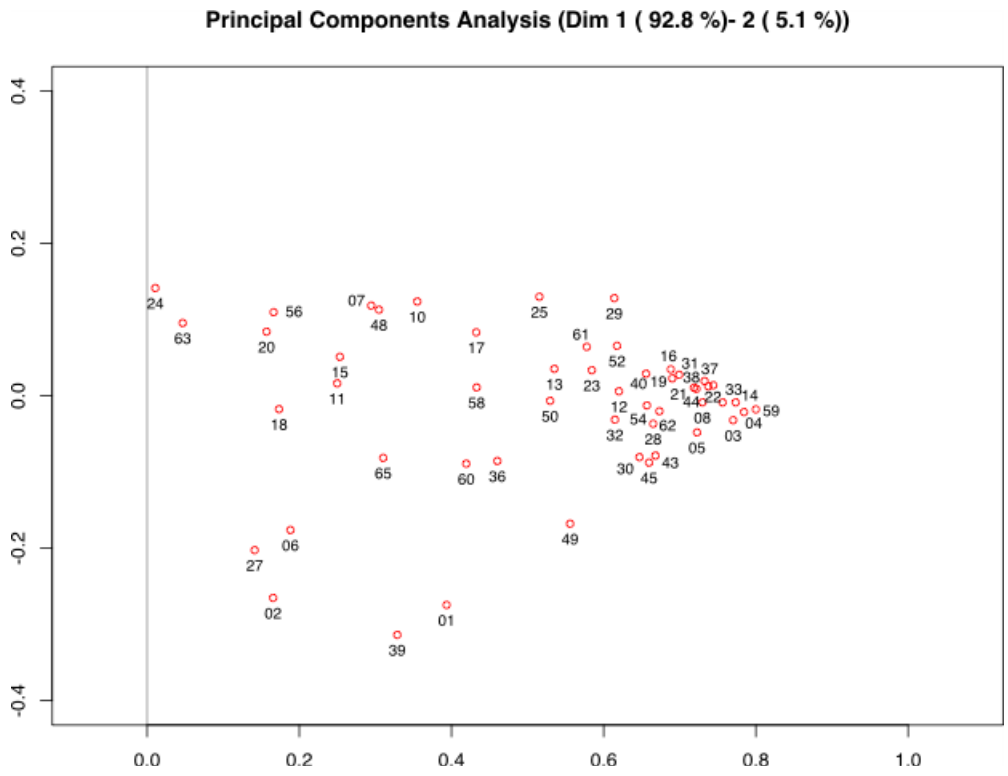


Figura 3.2: Causas de Muerte en el Ecuador: Proyección de las causas sobre las componentes principales

La correlación con la segunda componente es mucho menor, es positiva con los último años analizados, negativa con los primeros y casi nula con los centrales. Esto corrobora la afirmación anterior de que la primera componente está relacionada con la magnitud de la media. La componente muestra las diferencias (pequeñas) a lo largo del tiempo. Aquellas causas de muerte con valores más altos (y positivos) en la segunda componente serán las que han aumentado su importancia en los últimos años mientras que, las que tengan valores negativos, serán las que han disminuido su importancia.

	Dim 1	Dim 2
1997	-0.96	-0.26
1998	-0.96	-0.25
1999	-0.96	-0.26
2000	-0.96	-0.24
2001	-0.97	-0.23
2002	-0.96	-0.23
2003	-0.97	-0.21
2004	-0.98	-0.16
2005	-0.99	-0.13
2006	-0.99	-0.10
2007	-0.99	-0.04
2008	-0.99	-0.03
2009	-0.99	0.01
2010	-0.98	0.00
2011	-0.97	0.04
2012	-0.96	0.07
2013	-0.98	0.13
2014	-0.97	0.20
2015	-0.96	0.25
2016	-0.93	0.34
2017	-0.92	0.37
2018	-0.91	0.39
2019	-0.89	0.43

Cuadro 3.2: Correlaciones con las componentes principales

Así , las causas mal definidas parecen haber disminuido su importancia relativa debido seguramente a que el registro de Ecuador ha aumentado su eficiencia.

Causas como las *”41 Insuficiencia cardíaca, complicaciones y enfermedades mal definidas”*, *”64 Agresiones (Homicidios)”* y la *”55 Afecciones originadas en el periodo prenatal* parecen haber disminuido, especialmente la primera. Es posible que, como la 41 está relacionada con enfermedades del corazón mal definidas, haya mejorado la clasificación en algún. momento. Es importante que los homicidios hayan disminuido en los últimos años.

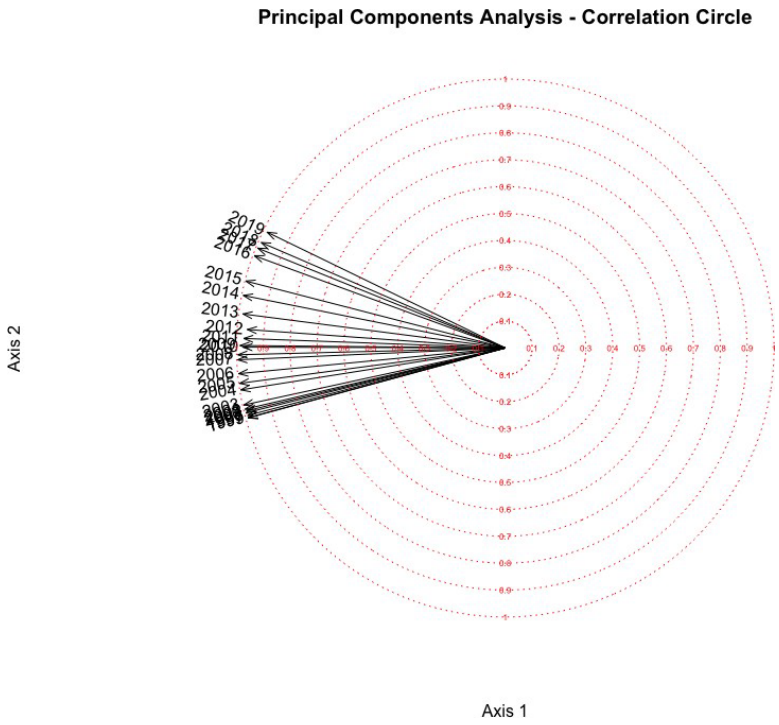


Figura 3.3: Causas de Muerte en el Ecuador: Círculo de correlaciones para las componentes principales

Las enfermedades que parecen haber aumentado más con el paso del tiempo son la *"35 Enfermedades isquémicas del corazón* y la *"26 Diabetes Mellitus*, ambas probablemente relacionadas con un aumento de la obesidad.

Capítulo 4

Biplot para datos multivariantes

4.1. Introducción

El origen de los Biplot se remonta a la década de los 70, cuando Gabriel en 1971 [31] los introduce con el objetivo principal de describir aproximadamente una matriz rectangular utilizando una representación gráfica en baja dimensión, que permita visualizar las interrelaciones entre individuos, entre variables y, además, de las relaciones entre ambos conjuntos. Un biplot es, normalmente, una representación gráfica plana o tridimensional.

Un Biplot [31] es una representación gráfica de datos multivariantes. De la misma manera que un diagrama de dispersión muestra la distribución conjunta de dos variables, un Biplot representa tres o más variables. [30].

El Biplot aproxima la distribución de una muestra multivariante en un espacio de dimensión reducida, normalmente de dimensión dos, y superpone sobre la misma representaciones de las variables sobre las que se mide la muestra [36]. Las representaciones de las variables son normalmente vectores, y coinciden con las direcciones en las que mejor se muestra el cambio individual de cada variable.

El *bi* en la palabra biplot se refiere al hecho de que en ese gráfico existen dos tipos de marcadores correspondientes a los dos tipos de información: los marcadores para los individuos o filas y los marcadores para las variables o columnas.

Si consideráramos las proyecciones de estos marcadores en un plano, el biplot quedar´ıa ahora representado en un espacio de dimensión 2. Los biplots permiten, por

inspección visual, identificar relaciones entre variables, relaciones entre individuos y relaciones entre variables e individuos.

Desde el punto de vista del usuario, los biplots serán importantes porque su interpretación se basa en conceptos geométricos sencillos, que forman parte de la cultura matemática de los potenciales usuarios, a saber,

- La similitud entre individuos es una función inversa de la distancia entre los mismos, sobre la representación biplot.
- En determinados tipos, las longitudes y los ángulos de los vectores que representan a las variables, se interpretan en términos de variabilidad y covariabilidad respectivamente.
- Las relaciones entre individuos y variables se interpretan en términos de producto escalar, es decir, en términos de las proyecciones de los puntos ¨ındividuo"sobre los vectores "variable".

Algunas referencias importantes a lo largo de la historia son las siguientes:

[7] , "al presentar el análisis factorial de correspondencias, pensado para tablas de contingencia de dos vías, relacionando las categorías de dos variables cualitativas, sintetiza las conclusiones de esos análisis en gráficos planos designados representations simultanées ". Esos gráficos tienen marcadores para las categorías en filas y marcadores para las categorías en columnas; conociendo las coordenadas de los marcadores de las filas es posible calcular las coordenadas de los marcadores de las columnas usando las fórmulas de transición [12].

[39], "en la obra que presenta a los lectores anglosajones el Análisis Factorial de Correspondencias de Benzécri para los investigadores de habla inglesa, introduce entre muchas otras innovaciones, el concepto de descomposición en valores singulares generalizados (GSVD), base para el concepto de biplot generalizado".

[108],"generaliza el concepto de representación simultánea creando un nuevo tipo de biplot - el HJ Biplot - que se aplica a todo conjunto de datos de dos vías y permite representar los individuos y las variables con igual calidad de representación" - lo que no ocurre con los biplots clásicos.

[106], "demuestra el papel central del concepto de biplot en análisis de datos multivariantes y presenta una generalización: el biplot generalizado".

[99], “estudia las relaciones entre los biplots y los métodos de análisis cluster, desarrollando un método de análisis cluster basado en el concepto de inercia”.

[68], “generaliza a los biplots de Gabriel, HJ-biplot y biplot generalizado, los métodos de integración de análisis en componentes principales”.

[2], “realiza una investigación teórica de las propiedades de los MANOVA Biplot en el contexto del MODELO LINEAL GENERAL MULTIVARIANTE, desarrollando métodos de interpretación de los MANOVA-BIPLOTS”.

Podemos encontrar muchos otros trabajos relacionados con el tema, hemos listado solamente aquellos que consideramos importantes en el desarrollo de los biplots para datos numéricos.

Describiremos primero el biplot clásico relacionado directamente con las componentes principales y en segundo lugar el biplot de regresión que nos permitirá proyectar variables sobre otro ya existente. Utilizamos aquí la notación más habitual en los artículos sobre el tema que adaptaremos después para el caso funcional.

Como libros completos que tratan el tema podemos encontrar diferentes puntos de vista en [36], [37] o [112].

4.2. Biplot clásico

4.2.1. Biplot general

Disponemos de una matriz de datos $\mathbf{X}_{n\times n}$ con las medidas de una variable en p ocasiones sobre n individuos.

Suponemos la matriz centrada por columnas y en muchos casos también estandarizada. Si llamamos

$$\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^t = \frac{1}{n}\mathbf{X}\mathbf{1}_n$$

donde $\mathbf{1}_n$ es un vector de n unos, al vector que contiene las medias de las p variables, el proceso de centrado consiste en sustituir la matriz $\mathbf{X}$ por

$$\mathbf{X} \leftarrow \mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}} = \mathbf{H}\mathbf{X} \tag{4.1}$$

donde $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$ es la matriz de centrado.

El vector que contiene las desviaciones t´ıpicas $\mathbf{s} = (s_1, \ldots, s_p)^l$ de cada variable, entonces

$$\mathbf{s}^2 = (\underset{1}{s^2}, \ldots, \underset{p}{s^2})^l = diag\left(\frac{1}{n}\mathbf{X}^l\mathbf{X}\right)$$

El proceso de estandarizado consiste en sustituir la matriz $\mathbf{X}$ por

$$\mathbf{X} \leftarrow \mathbf{X}\mathbf{D}_s^{-1} \tag{4.2}$$

donde $\mathbf{D}_s = diag(s_1, \ldots, s_p)$. El procedimeinto de estandarizado es particularmente útil cuando las variables están medidas en escalas muy diferentes.

Un Biplot para una matriz de datos $\mathbf{X}$ es una representación gráfica mediante marcadores (vectores) $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ para las filas y $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p$ para las columnas, de forma que el producto escalar $\mathbf{a}_i^l\mathbf{b}_j$ aproxime el elemento x_{ij} de $\mathbf{X}$ tan bien como sea posible, es decir

$$x_{ij} \approx \mathbf{a}_i^l\mathbf{b}_j \tag{4.3}$$

o también

$$x_{ij} = \mathbf{a}_i^l\mathbf{b}_j + e_{ij} \tag{4.4}$$

donde e_{ij} es el error cometido en la aproximación para cada elemento de la matriz. Como el biplot aproxima los elementos de la matriz, se le denomina normalmente *biplot de predicción.*

Normalmente los marcadores se toman en dimensión 2, pero podrían tomarse en cualquier otra dimensión q. La ventaja de que la dimensión sea 2 o 3 es que es posible visualizarlos en la pantalla del ordenador.

Los marcadores fila se suelen representar como puntos y los marcadores fila como vectores.

El aspecto típico de un biplot bidimensional se muestra en la figura 4.1.

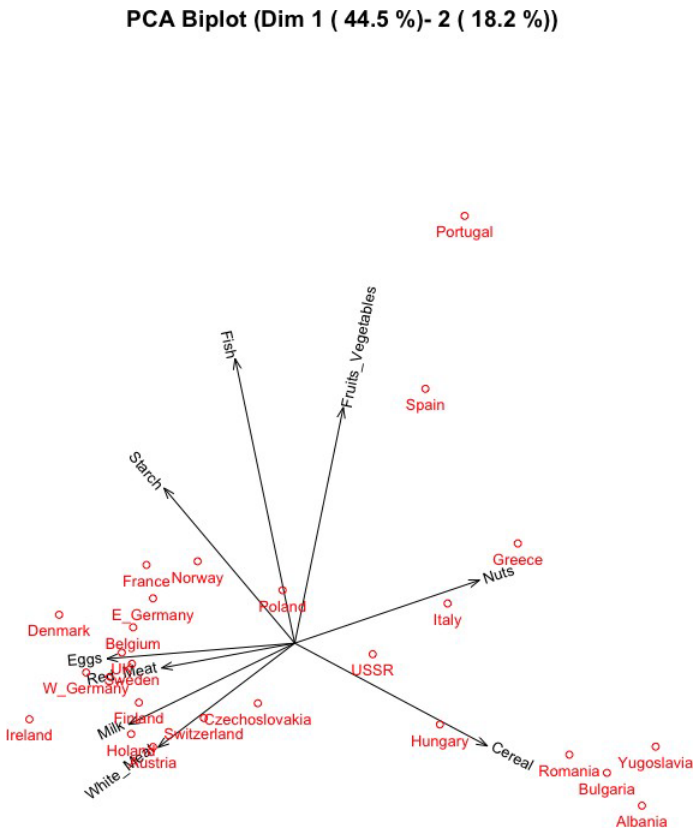


Figura 4.1: Un biplot típico

Desde el punto de vista geométrico la situación puede verse en la figura 4.2a. La aproximación de un elemento de la matriz mediante el producto escalar es

$$x_{ij} \approx \mathbf{a}_i^{\dagger}\mathbf{b}_j = Proy(\mathbf{a}_i/\mathbf{b}_j)@||\mathbf{b}_j|| \tag{4.5}$$

donde @ es el producto con signo.

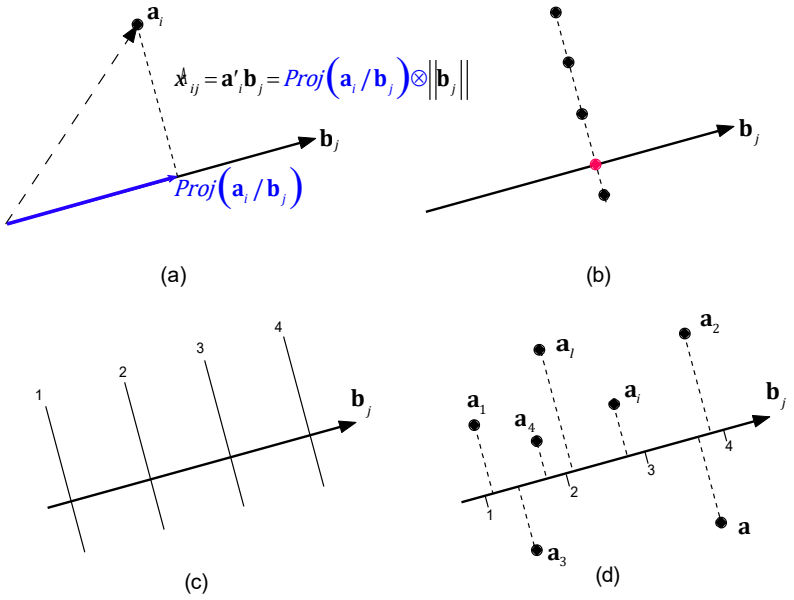


Figura 4.2: Aproximación biplot: (a) Producto escalar de los marcadores fila y columna. (b) El conjunto de puntos que predicen el mismo valor está en una línea recta perpendicular a la dirección definida por el marcador columna $\mathbf{b}_j$. (c) Los puntos que predicen diferentes valores están en líneas paralelas. (d) Sobre la dirección de la variable se pueden añadir escalas para obtener visualmente la predicción.

De la ecuación 4.5 se deduce que todos los puntos que tienen el mismo producto escalar (predicen el mismo valor), están en una línea recta perpendicular a la dirección definida por el marcador columna $\mathbf{b}_j$ (Figura 4.2, (a)-(b)). Tenemos entonces que diferentes valores se predicen mediante líneas paralelas (Figura 4.2, (c)). De esta forma podemos seleccionar, sobre la dirección del biplot, las líneas que predicen distintos valores y sus proyecciones sobre la dirección para mostrar escalas graduadas que se utilizan como las escalas de los ejes de cualquier gráfico cartesiano (Figura 4.2, (d)).

Los cálculos para colocar las escalas son sencillos. Para encontrar el marcador correspondiente a un valor fijo μ, buscamos el punto (x, y) que predice μ y está en la dirección de $\mathbf{b}_j$, es decir, sobre la línea que une los puntos $(0, 0)$ y $\mathbf{b}_j = (b_{j1}, b_{j2})$, esto es

$$y = \frac{b_{j2}}{b_{j1}} x$$

La predicción verifica también que

$$\mu = b_{j1}x + b_{j2}y$$

Entonces, obtenemos

$$x = \frac{\mu\, b_{j1}}{b_{j1}^2 + {}_{j2}^2}; \qquad y = \frac{\mu\, b_{j2}}{b_{j1}^2 + {}_{j2}^2} \tag{4.6}$$

La figura 4.3 contiene una representación biplot con escalas para todas las variables y contodos los individuos proyectados sobre una de ellas.

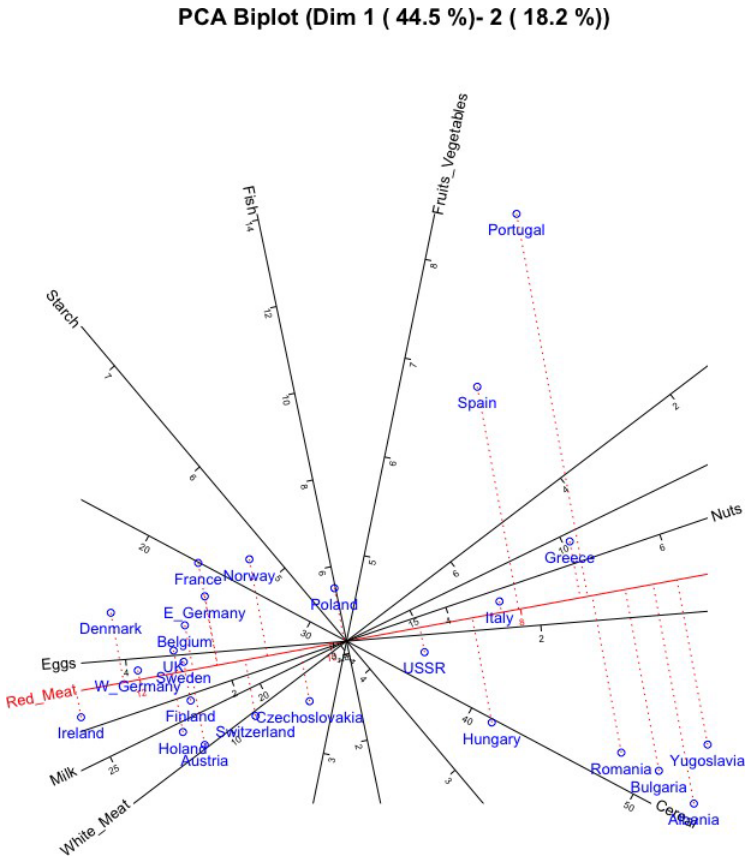


Figura 4.3: Un biplot con escalas graduadas para cada variable y todos los puntos fila proyectados sobre una variable

Una propiedad importante es que todo lo expuesto no depende de la elección particular de los marcadores, es decir, es posible colocar las escalas graduadas en cualquier biplot sea cual sea la forma en que se han obtenido los marcadores. Veremos en las secciones siguientes que hay diversas elecciones posibles.

Si colocamos los marcadores, que suponemos en dimensión q, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ en una matriz $\mathbf{A}$ y los de las columnas $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p$ en una matriz $\mathbf{B}$, un biplot en

dimensión q es una descomposición de la matriz $\mathbf{X}$ de la forma

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\intercal} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \tag{4.7}$$

donde $\mathbf{A}$ y $\mathbf{B}$ son matrices de rango q con n y p filas respectivamente, y $\mathbf{E}$ es una matriz $n \times p$ que contiene lo errores or residuales. Los valores ajustados de $\mathbf{X}$, pueden escribirse como

$$E[\mathbf{X}] = \hat{\mathbf{X}} = \mathbf{A}\mathbf{B}^{\intercal} \tag{4.8}$$

4.2.2. Bondad de ajuste para un biplot general

Como hemos visto, el objeto de un biplot es aproximar (predecir) los valores de la matriz de datos en dimensión reducida, lo mejor posible. teniendo en cuenta 4.7, los valores ajustados en el biplot (las predicciones) serán

$$\hat{\mathbf{X}} = \mathbf{A}\mathbf{B}^{\intercal} \tag{4.9}$$

La bondad de ajuste global, es decir, la cantidad de variabilidad explicada por la representación en dimensión reducida, podemos escribirla como

$$\rho^2 = tr(\hat{\mathbf{X}}^{\intercal}\hat{\mathbf{X}})/tr(\mathbf{X}^{\intercal}\mathbf{X}) \tag{4.10}$$

o bien

$$\rho^2 = 1 - tr(\hat{\mathbf{E}}^{\intercal}\hat{\mathbf{E}})./tr(\mathbf{X}^{\intercal}\mathbf{X}) \tag{4.11}$$

Es conocido que una buena bondad de ajuste global no implica un buen ajuste de todas las filas y las columnas de la matriz. Incluso en algunas disciplinas, como la genómica, en el que el número de columnas es muy grande y la mayor parte no tienen información, puede que la bondad de ajuste global sea irrelevante. Sería interesante saber cuales son las variables cuya variabilidad está correctamente recogida en la representación

La bondad de ajuste para una columna podríamos expresarla como

$$\rho^2_{Cj} = \frac{\sum_{i=1}^{n} \hat{x}_{ij}^2}{\sum_{i=1}^{n} x_i^2} \tag{4.12}$$

El vector que contiene las bondades de ajuste para cada columna separada lo podríamos expresar como

$$\rho^2_C = diag(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})./diag(\mathbf{X}^\top \mathbf{X}) \tag{4.13}$$

donde ./ significa la operación de división elemento a elemento.

ρ^2_{Cj} puede interpretarse como la variabilidad de la j-ésima variable explicada en el biplot o incluso el R^2 de la regresión de la columna sobre las coordenadas de biplot. En la literatura se denomina también *Calidad de representación*, *Predictividad* o *cosenos al cuadrado*. Es también la suma de las correlaciones al cuadrado entre la variable y cada uno de las componentes (ejes) de la representación y la correlación al cuadrado entre los valores observados y esperados (cuando los datos están centrados).

El vector que contiene las bondades de ajuste para cada fila separada lo podríamos expresar como

$$\rho^2_{Ri} = \frac{\sum_{j=1}^{p} x_{ij}^2}{\sum_{j=1}^{p} x_{ij}^2} \tag{4.14}$$

El vector que contiene las bondades de ajuste para cada fila separada lo podríamos expresar como

$$\rho^2_R = diag(\hat{\mathbf{X}}\hat{\mathbf{X}}^\top)./diag(\mathbf{X}\mathbf{X}^\top) \tag{4.15}$$

A estas medidas también se las denomina *calidad de representación*, *predictividad* o *cosenos al cuadrado*.

Obsérvese que las medidas son independientes de la descomposición particular, si bien es posible que la interpretación sea diferente en algunas de ellas.

4.2.3. Biplot de interpolación

Todo lo desarrollado para los biplots generales se refiere a lo que se denomina un *biplot de predicción* ya que el objeto es aproximar (predecir) lo mejor posible los valores observados en la matriz de datos. Hay otro tipo de biplot que sirve para interpolar nuevos individuos sobre una representación existente de forma geométrica. Supongamos que tenemos un nuevo individuo nuevo que tiene valores en las p variables $\mathbf{x}_h = (x_{h1}, \ldots, x_{hp})$ y queremos proyectarlo (interpolarlo) sobre el biplot. En otros entornos se conoce como *individuo suplementario.*

En primer lugar realizamos las transformaciones iniciales oportunas como en 4.1 y 4.2

$$\mathbf{x}_h \leftarrow (\mathbf{x}'_h - \bar{\mathbf{x}}')\mathbf{D}_s^{-1}$$

Buscamos las coordenadas $\mathbf{a}_h$ del individuo sobre el biplot.

Si en la ecuación

$$\mathbf{X} \approx \mathbf{A}\mathbf{B}' \tag{4.16}$$

Multiplicamos ala derecha por $\mathbf{B}$, tenemos

$$\mathbf{X}\mathbf{B} \approx \mathbf{A}\mathbf{B}'\mathbf{B} \tag{4.17}$$

multiplicando ahora por $(\mathbf{B}'\mathbf{B})^{-1}$, se obtiene

$$\mathbf{A} \approx \mathbf{X}\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1} = \mathbf{X}\mathbf{C} \tag{4.18}$$

donde $\mathbf{C} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}$. Las coordenadas del nuevo punto son entonces

$$\mathbf{a}_h = \mathbf{x}'_h\mathbf{C} = \sum_{j=1}^{p} x_{hj}\mathbf{c}_j \tag{4.19}$$

donde $\mathbf{c}_j$ es la *j-ésima* fila de $\mathbf{C}$, es decir, las coordenadas son una suma ponderada de los vectores $\mathbf{c}_j$ cada uno de ellos ponderados por los correspondientes valores de la variable.

A la representación de $\mathbf{A}$ y $\mathbf{C}$ la denominamos *biplot de interpolación* ya que nos

permite la interpolación gráfica de nuevos puntos como en la figura 4.4.

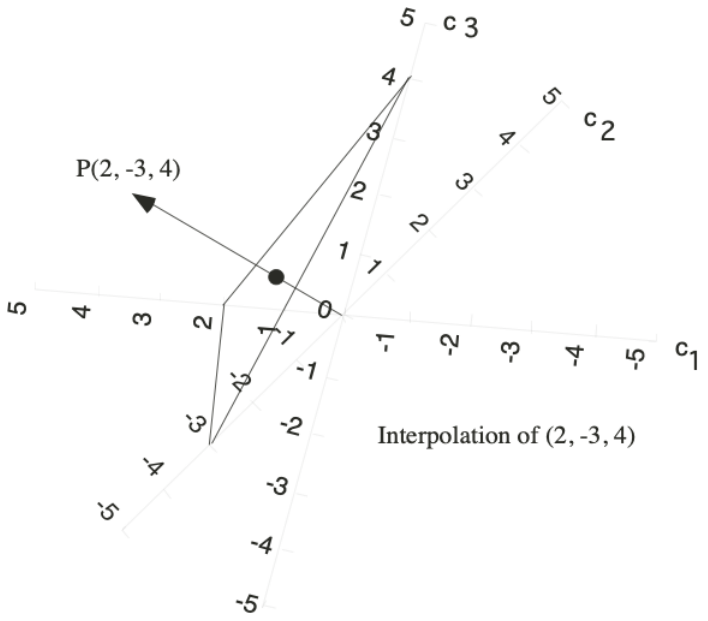


Figura 4.4: Escalas graduadas para interpolación

Para encontrar las escalas graduadas para la interpolación buscamos, para un marcador fijo μ, el punto (x, y)

$$x = \mu\ c_{j1}; \qquad y = \mu\ c_{j2} \tag{4.20}$$

Finalmente la interpolación se realiza como suma de vectores seleccionando los valores adecuados en las escalas graduadas. Obsérvese que las direcciones del biplot de interpolación son los coeficientes de regresión resultantes de poner cada variable en función de las dimensiones del biplot.

4.2.4. Biplot basado en la descomposición en valores singulares

Hay diversas formas de obtener una descomposición en la forma 4.7 aunque la más útil será aquella que proporcione la mejor aproximación en dimensión reducida. Es bien conocido que podemos reproducir exactamente la matriz en dimensión

r, donde r es el rango de $\mathbf{X}$, utilizando su descomposición en valores y vectores singulares (DVS),

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^{l} = \sum_{k=1}^{r} \lambda_k \mathbf{u}_k \mathbf{v}_k \tag{4.21}$$

donde $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$ y $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ son las matrices que contienen los vectores singulares por la izquierda y por la derecha respectivamente y Λ la matriz diagonal R-dimensional que contiene los valores singulares en orden decreciente $\lambda_1 \geq \ldots \geq \lambda_r > 0$. Sabemos también que $\mathbf{U}$ son los vectores propio de $\mathbf{X}\mathbf{X}^{l}$, $\mathbf{V}$ los vectores propios de $\mathbf{X}^{l}\mathbf{X}$ y Λ las ra´ıces cuadradas de los valores propios no nulos de ambas matrices, que sen iguales. Obsérvese que la primera descomposición está relacionada con los productos escalares euclídeos y por tanto con las Coordenadas Principales [35] mientras que la segunda está relacionada con las componentes principales tal y como las describimos antes.

Vemos también que la aproximación consiste en la suma de varias matrices de rango 1.

Es también conocido que la mejor aproximación en rango reducido p ($p << r$) para $\mathbf{X}$ se obtiene de

$$\hat{\mathbf{X}} = \mathbf{U}_{(q)}\Lambda_{(q)}\mathbf{V}_{(q)}^{T} = \mathbf{A}\mathbf{B}^{T} = \sum_{k=1}^{q} \lambda_k \mathbf{u}_k \mathbf{v}_k \tag{4.22}$$

donde el subíndice (q) significa *las primeras q columnas* de la matriz correspondiente. La aproximación en rango q es también la suma de q matrices de rango 1.

Podemos definir un biplot como en la ecuación 4.7 tomando

$$\mathbf{A} = \mathbf{U}_{(q)}\Lambda_{(S)}^{\gamma} \tag{4.23}$$

y

$$\mathbf{B} = \mathbf{V}_{(q)}\Lambda_{(K)}^{(1-\gamma)} \tag{4.24}$$

con $0 \leq \gamma \leq 1$.

Una versión particularmente interesante para nuestro caso es aquella en la que $\gamma = 1$ de forma que

$$\mathbf{A} = \mathbf{U}_{(q)}\Lambda_{(S)} \tag{4.25}$$

and

$$\mathbf{B} = \mathbf{V}_{(q)} \tag{4.26}$$

De esta forma, las coordenadas de las filas se corresponden con la coordenadas sobre las componentes principales (scores) tal y como fueron descritas en el apartado correspondiente, que se corresponden también con las coordenadas principales. Esta representación es conocida como JK-Biplot or RMP-Biplot, donde RMP significa *Row Metric Preserving*, es decir se trata de un biplot que preserva la métrica de las filas en el sentido de que la distancia eúclidea entre dos puntos fila, en el espacio de dimensión reducida, aproxima la distancia euclídea en el espacio completo.

Una descomposición alternativa con $\gamma = 0$ sería

$$\mathbf{A} = \mathbf{U}_{(q)} \tag{4.27}$$

and

$$\mathbf{B} = \mathbf{V}_{(q)}\Lambda_{(q)} \tag{4.28}$$

Este biplot estaría más relacionado con la solución de componentes principales para el modelo factorial. Si los datos están estandarizados, las coordenadas de las columnas son las correlaciones de las variables observadas y los factores latentes mientras que, las coordenadas de las filas son las puntuaciones estandarizadas sobre los factores. Esta representación es conocida como GH-Biplot or CMP-Biplot, donde RMP significa *Row Metric Preserving*, es decir se trata de un biplot que preserva la métrica de las columnas en el sentido de que el coseno del ángulo entre dos puntos columna, en el espacio de dimensión reducida, aproxima la covarianza (correlación) en el espacio completo (el coseno del ángulo en el espacio completo).

Todo lo dicho sobre la calidad de representación para un biplot general.

4.3. Biplot de regresión

Desde otro punto de vista, si consideramos los valores en **A** como fijos, los marcadores columna **B** pueden calcularse mediante regresiones multivariantes

$$\mathbf{B}^{\intercal} = (\mathbf{A}^{\intercal}\mathbf{A})^{-1}\mathbf{A}^{\intercal}\,\mathbf{X} \tag{4.29}$$

De la misma manera, fijando **B**, **A** pueden obtenerse como:

$$\mathbf{A}^{\intercal} = (\mathbf{B}^{\intercal}\mathbf{B})^{-1}\mathbf{B}^{\intercal}\mathbf{X} \tag{4.30}$$

Alternando ambas ecuaciones se puede demostrar que el procedimiento converge a la misma solución que la DVS.

Es inmediato comprobar que si fijamos las coordenadas de las filas como en **??** y usamos la regresión en 4.29, obtenemos el mismo resultado. Usando las propiedades de la DVS

$$\mathbf{B}^{\intercal} = (\mathbf{A}^{\intercal}\mathbf{A})^{-1}\mathbf{A}^{\intercal}\,\mathbf{X} = (\Lambda_{(q)}\mathbf{U}^{\intercal}_{(q)}\mathbf{U}_{(q)}\Lambda_{(q)})^{-1}\Lambda_{(q)}\mathbf{U}^{\intercal}_{(q)}\mathbf{X} = \Lambda^{-1}_{(q)}\mathbf{U}_{(q)}\mathbf{X} = \mathbf{V}^{\intercal}_{(q)} \tag{4.31}$$

Simplemente teniendo en cuenta que $\mathbf{V}_{(S)}\Lambda_{(S)} = \mathbf{X}^{\intercal}\mathbf{U}_{(S)}$ y entonces $\mathbf{V}_{(S)} = \mathbf{X}^{\intercal}\mathbf{U}_{(S)}\Lambda^{-1}_{(S)}$.

La ecuación 4.29 puede servir, por ejemplo, para construir biplots lineales aproximados a partir de las coordenadas principales o de la solución de un escalado multidimensional. También puede servir para proyectar variables externas sobre un biplot ya construido. A este tipo de biplots los llamamos *biplots externos* porque las coordenadas de las filas se construyen mediante procedimientos separados. Un ejemplo de este tipo de biplots externos es el denominado *biplot logísticoexterno para datos binarios* en el que se usa una regresión logística, en lugar de una regresión lineal, para proyectar las variables sobre la solución obtenida de un Análisis de Coordenadas Principales.

La ecuación 4.30 puede servir para proyectar individuos nuevos sobre el biplot como alternativa a las fórmulas de interpolación. Hay que tener en cuenta que si los datos están centrados y estandarizados, antes de proyectar un individuo nuevo hay que centrarlo y estandarizarlo usando las mismas medias y las mismas desviaciones típicas.

4.4. HJ-Biplot

4.4.1. Introducción

Como hemos comprobado en apartados anteriores, las representaciones son asimétricas en el sentido de que no obtienen la misma calidad de representación para las filas y para las columnas de la matriz de datos. Cuando el propósito es la aproximación de los elementos de la matriz original, los biplots presentados son óptimos, además en cada uno de ellos es posible representar con mejor calidad las características de las filas o de las columnas, cuando se quieren interpretar por separado.

Cuando las filas y las columnas son importantes en si mismas, y se quieren interpretar las características de ambas manteniendo cierta relación entre las mismas, son más útiles las interpretaciones basadas en representaciones simétricas como el Análisis Factorial de Correspondencias en el que se interpretan las posiciones de las filas, las posiciones de las columnas y las relaciones fila-columna a través de los factores, es decir se realiza una interpretación factorial.

El problema es que, el Análisis de Correspondencias está pensado solamente para matrices de frecuencias. Sería interesante disponer de una técnica simétrica similar, pero aplicable a cualquier conjunto de datos.

(1986)[108] propone el que denomina HJ-biplot que responde a las características descritas en los párrafos anteriores.

4.4.2. Definición y selección de marcadores

Un HJ-Biplot para una matriz de datos $\mathbf{X}$ es una representación gráfica multivariante mediante marcadores (vectores) $\mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_n$ para las filas y $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p$ para las columnas de $\mathbf{X}$, elegidos de forma que ambos marcadores puedan superponerse en el mismo sistema de referencia con máxima calidad de representación.

Partimos, también de la descomposición en valores singulares

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}' \tag{4.32}$$

elegimos como marcadores en dimensión q

$$\mathbf{A} = \mathbf{U}_{(q)}\Lambda_{(q)} \tag{4.33}$$

and

$$\mathbf{B} = \mathbf{V}_{(q)}\Lambda_{(q)} \tag{4.34}$$

4.4.3. Propiedades

1.- Los marcadores fila y columna se pueden representar en el mismo sistema de referencia.

En el contexto de las Correspondencias, [39] basa esta afirmación en que ambas nubes están referidas a los mismos valores propios y al hecho de que ambas nubes están relacionadas.

El que las nubes están referidas a los mismos valores propios es obvio, ya que los valores propios de $\mathbf{X}^{\prime}\mathbf{X}$ y $\mathbf{X}\mathbf{X}^{\prime}$ son los mismos.

Las relaciones entre las nubes son las relaciones baricéntricas similares a las del Análisis Factorial de Correspondencias. Teniendo en cuenta que a partir de la DVS se tiene que $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{D}^{-1}$ y $\mathbf{V} = \mathbf{X}^{\prime}\mathbf{U}\mathbf{D}^{-1}$, entonces tenemos

$$\mathbf{A} = \mathbf{U}\mathbf{D} = \mathbf{X}\mathbf{V} = \mathbf{X}\mathbf{X}^{\prime}\mathbf{U}\mathbf{D}^{-1} = \mathbf{X}\mathbf{B}\mathbf{D}^{-1} \tag{4.35}$$

$$\mathbf{B} = \mathbf{V}\mathbf{D} = \mathbf{X}^{\prime}\mathbf{U} = \mathbf{X}^{\prime}\mathbf{X}\mathbf{V}\mathbf{D}^{-1} = \mathbf{X}^{\prime}\mathbf{A}\mathbf{D}^{-1} \tag{4.36}$$

Es decir, las coordenadas para las filas son medias ponderadas, salvo un factor de escala relacionado con los valores singulares, de las coordenadas de las columnas, donde las ponderaciones son los valores originales en la matriz $\mathbf{X}$. Lo mismo ocurre con las coordenadas de las columnas respecto de las de las filas.

2.- Las calidades de representación de filas y columnas son las mismas. La aproximación de los productos escalares entre filas y entre columnas tienen la misma bondad de ajuste.

3.- Las propiedades del HJ-Biplot son las de los marcadores $\mathbf{A}$ del JK-biplot y $\mathbf{B}$ del GH-Biplot detalladas en apartados anteriores.

4.5. Aplicación del biplot a los datos de *Helicobacter Pylori*

4.5.1. Resultados

Las variables hábitos de higiene y Afecciones que produce el bacilo *H. pylori* analizadas se representan mediante vectores, mientras que las zonas de muestreo se pueden identificar mediante puntos; sus colores varían según el lugar de muestreo. La Figura 4.5 muestra la representación del plano factorial 1-2 que resultó del HJ Biplot, lo que explica el 38,12 % de la varianza.

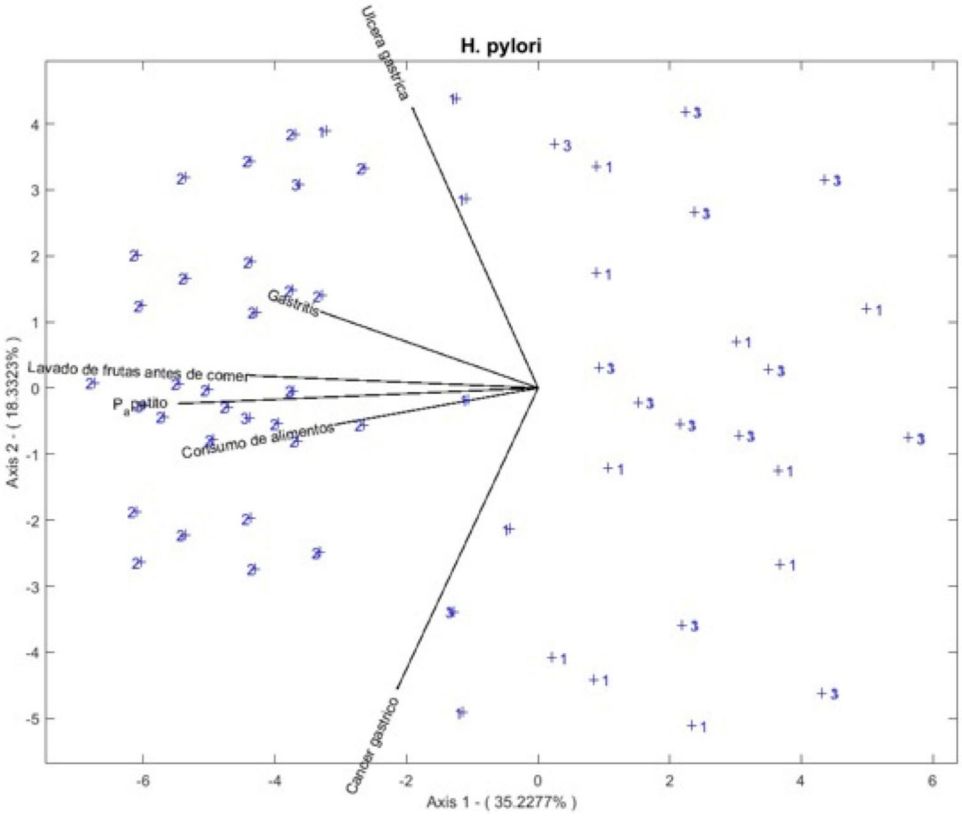


Figura 4.5: Representación de HJ-Biplot en el primer plano principal.

En el primer eje, las variables con mayor variabilidad fueron el cáncer gástrico, consumó de alimentos, pérdida de apetito. Esta variabilidad entre Agua de consumo, Lavado de frutas antes de comer, Gastritis, Ulcera gástrica podría explicarse por un

gradiente estacionario relacionado con el comienzo de la estación de lluvias; mientras que la variabilidad en la Gastritis y Ulcera gástrica puede estar asociada a los cambios en la temperatura y el viento.

El análisis de conglomerados se llevó a cabo sobre la base de las coordenadas obtenidas del HJ-Biplot (método K-means, coseno). Se formaron dos cúmulos con los diferentes puntos de muestra. En la representación gráfica pueden observarse los cúmulos identificados a través de las líneas de Cascos Convexos (Figura 3). Este análisis permitió identificar las variables físico-químicas y biológicas que influyeron en los conjuntos entre los diferentes puntos de muestra.

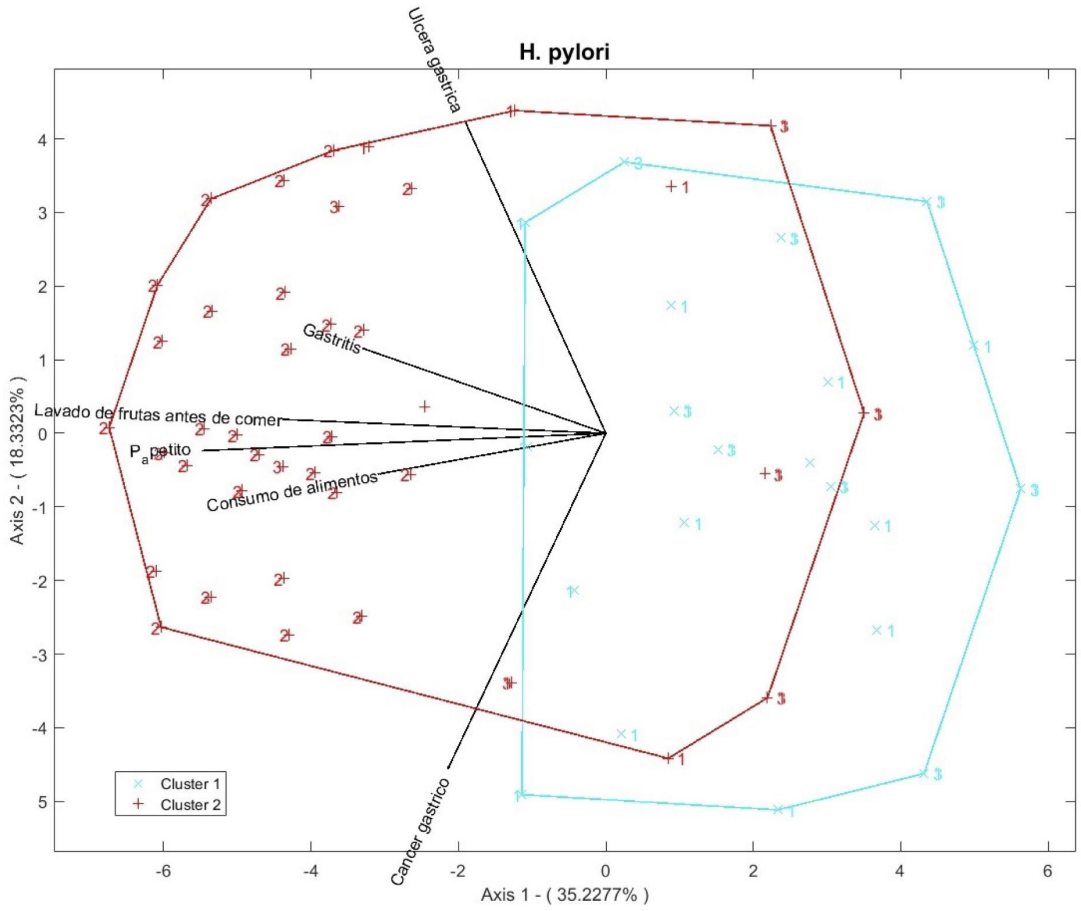


Figura 4.6: Representación de HJ-Biplot en el primer plano principal con clusters superpuestos.

Este método también ofrece la ventaja de permitir una representación de los puntos de muestreo (filas) y hábitos de higiene y Afecciones que produce el bacilo H. pylori (columnas) simultáneamente en el mismo sistema de coordenadas. Además, este método proporciona una alta calidad de representación tanto de los puntos de muestreo como de las variables hábitos de higiene y Afecciones que produce el bacilo H. pylori. Sin embargo, este método presenta una limitación en la representación del elemento original Xij de la matriz de datos. Por lo tanto, estos métodos son útiles para la evaluación de la calidad del agua, hábitos de higiene y para detectar patrones multivariantes en un grupo complejo de datos con diferentes variables.

4.5.2. Discusión

La ocurrencia del contagio con el bacilo del *H. pylori,* presenta múltiples factores, siendo la más frecuente el incumplimiento de las normas sanitaria y los hábitos alimenticios, por cuanto la población en estudio se determinó que la mayor prevalencia de los casos de contagio estuvo presente en el sexo femenino, las cuales mantienen una alimentación en puestos ambulantes, los cuales no presentan unas condiciones de sanidad reglamentarias, ocasionando la presencia del bacilo en el organismo y las consecuencia que trae su prolongada permanencia en el organismo. En este sentido Ram´ırez-Ramos y Sánchez-Sánchez ob cit. indica que pese a la erradicación de la infección puede todavía producirse un cáncer gástrico debido a la continua progresión de las lesiones pre-cancerosas. Esto podría deberse a que toma un buen tiempo para que el cáncer gástrico adquiera suficiente tamaño para ser reconocido endoscópicamente. Por lo cual el personal contagiado por el bacilo puede desarrollar en un futuro el cáncer gástrico.

En este estudio, un novedoso enfoque, basado en el desarrollo de la técnica HJ-Biplot, ha permitido generar una nueva interpretación, conocimiento y evaluación sobre la calidad de hábitos de higiene y afecciones que produce el bacilo H. pylori. En este sentido, facilitó la interpretación de las similitudes/desimilitudes entre los individuos (puntos de muestreo), la covariación entre las variables hábitos de higiene y afecciones que produce el bacilo H. pylori y, lo que es más importante, determinó qué variables indican las diferencias entre los distintos puntos de muestreo. El HJ-Biplot tiene un carácter descriptivo que ha permitido ver en un plano, de manera

simple y clara, la representación conjunta entre las variables hábitos de higiene y afecciones que produce el bacilo H. pylori a lo largo de las diferentes zonas de estudio rural y urbana.

En cuanto al grupo etario que más afección con el contagio del bacilo del H. pylori, fue población joven, es la que registro la mayor cantidad de caso, en esta sección de la población se estableció que no mantiene una atención dela infección por el bacilo, siendo este un factor para la aparición de las enfermedades gástricas, en la cual si mantienen esta posición las mismas pueden generar el cáncer gástrico como consecuencia de la afección por el H. pylori, ya que la afectación no se nota de inmediato sino con el pasar de los años. Por lo cual Ram´ırez-Ramos y Sánchez-Sánchez ob cit. establece que la aparición del cáncer gástrico no es de manera inmediata, sino que puede durar hasta décadas y que durante este período tiene lugar un prolongado proceso pre-canceroso representado por una cascada de eventos, histopatológicamente secuenciales: gastritis crónica activa no atrófica; gastritis atrófica multifocal; metaplasia intestinal (completa y luego incompleta); displasia y carcinoma invasivo.

El HJ-Biplot demuestra ser una herramienta eficaz para los estudios de calidad y afecciones que produce el bacilo H. pylori. En este caso, demostró que el nivel de casos de afecciones por la presencia del bacilo H. pylori en la población estudiada fue de 75 % en el estudio del suero (anticuerpos IgG) y 71 % en los estudios de heces. Estos resultados son muchos mayores a los obtenido por Matta-deGarc´ıa (2015) quien realizó un estudio similar en la ciudad de Guatemala en el año 2015 el cual arrojó la presencia de un 56,2 % de anticuerpos IgG de H. pylori, variando para el antígeno fecal con un 30,9 %, reflejando así un mayor porcentaje de anticuerpos IgG, debido a que estos permanecen elevados por un tiempo prolongado (6 meses o más) e incluso después de que el tratamiento haya sido efectivo. Además, señala que el antígeno fecal tiene una mínima concordancia con la biopsia (gold standar), debido a que ambas pruebas diagnostican directamente la presencia de la bacteria a diferencia de los anticuerpos

Se estableció que el nivel de casos de afecciones por la presencia del bacilo *H. pylori* en la población estudiada fue de 75 % en el estudio del suero (anticuerpos IgG) y 71 % en los estudios de heces. Estos resultados son muchos mayores a los obtenido por Matta-de-Garc´ıa (2015) quien realizó un estudio similar en la ciudad

de Guatemala en el año 2015 el cual arrojo la presencia de un 56,2 % de anticuerpos IgG de *H. pylori*, variando para el antígeno fecal con un 30,9 %, reflejando así un mayor porcentaje de anticuerpos IgG, debido a que estos permanecen elevados por un tiempo prolongado (6 meses o más) e incluso después de que el tratamiento haya sido efectivo. Además, señala que el antígeno fecal tiene una mínima concordancia con la biopsia (gold standar), debido a que ambas pruebas diagnostican directamente la presencia de la bacteria a diferencia de los anticuerpos.

En el presente estudio la prevalencia del cáncer gástrico fue la tercera en ocurrencia, pero la de mayor importancia por sus consecuencias, y a que, aunque se haya eliminado el bacilo del *H. pylori,* está ya ha causado las lesiones que perjudicaran al organismo. Se ha demostrado que esta infección desempeña un papel importante en la gastritis, ulcera gástrica y duodenal, carcinoma gástrico y Maltoma3. También se ha postulado asociación a enfermedades extraintestinales, aunque la evidencia en este aspecto es aún insuficiente (Ramírez-Ramos y Gilman, 2004)

Aun cuando la prevalencia del cáncer gástrico, es motivado a múltiples factores, el bacilo del *H. pylori,* favorece la aparición de este carcinoma gástrico motivado a que afecta la mucosa gástrica aumentando la posibilidad de la aparición de esta patología, aunado al descuido que realiza el portador de este bacilo y a los efectos que este causa aun cuando haya sido eliminado del organismo y la recurrencia de su posible contagio.

Capítulo 5

Análisis de Datos Funcionales

5.1. Introducción

El objeto de estudio para este capítulo es la presentación del Análisis de Datos Funcionales, citado también como FDA por sus siglas en inglés. Se expone la bibliografía consultada para poder explicar el campo de acción de este modelo Estadístico, que estudia y analiza la información contenida en curvas, superficies, o cualquier elemento que varía sobre un dato continuo, que por lo general es una línea de tiempo. Este es el caso de los datos demográficos, objeto de estudio en este trabajo.

Los datos en muchos campos nos llegan a través de un proceso descrito naturalmente como funcional en el que se consideran funciones que generan cada una de las observaciones, es decir, cada individuo u observación está representado por una función o curva en lugar de por un vector como en los métodos multivariantes tradicionales. El tratamiento de estas observaciones como funciones, hace que surja todo un aparato de definiciones, teoremas, axiomas, demostraciones y herramientas para poder estudiarlas correctamente [91].

Una cuestión importante del enfoque del FDA es que las funciones que se trabajen deben ser suaves, es decir, se trabajan con datos menos distorsionados en comparación con el comportamiento real inicial. En caso de que no lo sean, éstas deberán ser sometidas a un suavizamiento antes de realizar el tratamiento con FDA.

Los objetivos del análisis de datos funcionales son esencialmente los mismos que los de cualquier otra rama de la estadística, estudia y analizar la información conte-

nida en curvas, superficies, o cualquier otro elemento, que generalmente varía en el tiempo, y que surgen, de manera natural, en varias áreas donde se trabaja con grandes conjuntos de datos. En particular, trataremos de representar los datos de manera que ayuden a un mejor análisis, mostrar los datos para resaltar sus características, estudiar fuentes importantes de patrones y variaciones entre observaciones, explicar la variación en un resultado o variable dependiente mediante el uso de variables independientes, etc ... Estos puntos concentran la mayor parte de la investigación en el análisis de datos funcionales [91]. De entre toda la literatura disponible sobresalen como referencias básicas los libros [89], [91] o [28] que tratan muchos de los problemas básicos de la estadística funcional, incluso desde el punto de vista no-paramétrico. En el libro [92] se centra en los aspectos computacionales en R y MATLAB del FDA.

Lo que describimos en la memoria se basa en [89] ya que se trata de una obra clave en el tema.

5.2. Naturaleza funcional de los datos demográficos

Como ya hemos visto en apartados anteriores, con datos demográficos nos referimos a datos socio-económicos tales como población, raza, ingresos, natalidad, mortalidad o desempleo, por ejemplo, y que están asociados normalmente con una localización particular y el tiempo. Este tipo de datos es recogido normalmente por la correspondiente agencia gubernamental de cada país. Normalmente se refieren a fechas concretas aunque se supone que el proceso que los genera es continuo por naturaleza. Po ejemplo, los datos de fallecimientos por COVID se refieren a días completos pero pudieron ser generados en cualquier momento del d´ıa.

Los datos demográficos tienen una naturaleza esencialmente temporal y pueden considerarse como generados por una función que depende del tiempo así que, las técnicas para el análisis de datos funcionales pueden ser útiles en este tipo de información [79].

En una revisión sistemática del año 2013 [102] ya se citan algunas aplicaciones a datos demográficos de fertilidad y mortalidad ([49]), tasas de mortalidad por edades ([48]), tasas específicas por edades de mortalidad por cáncer ([26]) entre otros.

Más recientemente [79] describen datos demográficos y económicos mediante técnicas funcionales.

En relación a la mortalidad por COVID-19, [10] caracterizan las distintas curvas para las regiones italianas mediante análisis de datos funcionales. En este mismo sentido [58] utilizan el Análisis de Componentes Principales Funcionales (ACPf) para la clasificación de las curvas de incidencia y mortalidad en distintos países del mundo.

Parece claro, entonces, que el FDA es útil para la modelización tanto de datos demográficos de mortalidad (como en el caso de las causas de muerte en Ecuador) como de datos de incidencia y mortalidad por COVID-19.

La figura siguiente muestra la evolución de las tasas acumuladas de muertes por COVID-19 en España y Ecuador, que pueden tratarse como datos funcionales.

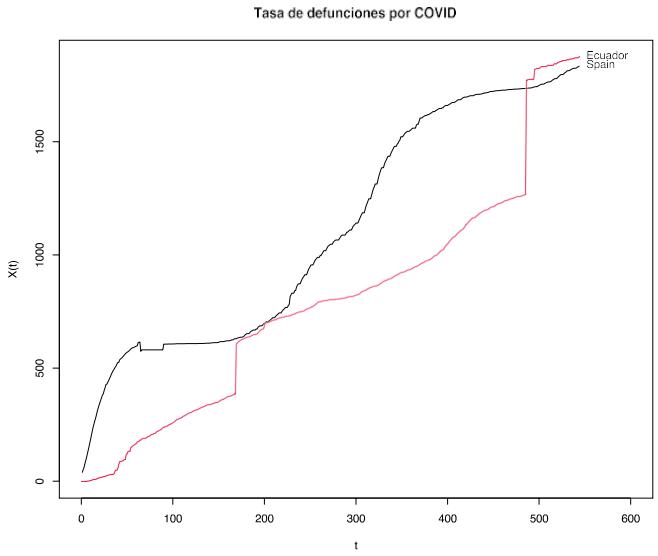


Figura 5.1: Tasa de defunciones por COVID-19 en cada 100000 habitantes para España y Ecuador

En la figura observamos el comportamiento de ambos países. Cabe destacar que, en España, se ha producido un ajuste de los datos después de la primera ola que ha resultado en un pequeño escalos en la curva. Para el caso de Ecuador se ha producido un fuerte incremento el 7 de Septiembre de 2020 y otro el 20 de Julio de 2021. Suponemos que estos datos obedecen a un problema de actualización de los

datos y no a un aumento súbito real. Como veremos después, el suavizado de los datos puede paliar este problema.

5.3. Definiciones Básicas Análisis de Datos Funcionales

Para [48], "el utilizar datos funcionales conlleva unas particularidades que hacen que los métodos tradicionales no sirvan o queden cortos". Esto es debido principalmente a tres de sus características: la alta dimensión, la alta correlación y el trabajar en espacios funcionales.

En el contexto multivariante, los datos para una réplica concreta vienen en un vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ que contiene los valores para la réplica i en p variables, es decir, suponemos que tenemos una muestra de una variable aleatoria multivariante $X = (X_1, \ldots, X_p)$. En el caso de los datos funcionales tenemos una muestra de funciones reales independientes $(X_1(t), \ldots X_p(t))$ en un intervalo compacto de la línea real $[1 \ldots, T]$. Las funciones pueden entenderse como realizaciones de un proceso estocástico $X(t)$. En la práctica el vector resultante es el mismo, aunque en el caso funcional suponemos que es una realización para un conjunto discreto de valores de t en lugar de un conjunto de variables aleatorias. Esto define una característica importante de los datos funcionales que implica que la escala de todas las variables es similar.

Por ejemplo, supongamos que tenemos la información de un electrocardiograma o el clima con las temperaturas de una ciudad. En la práctica, estos sucesos son recogidos por dispositivos electrónicos que toman muestras de una determinada variable aleatoria en distintos instantes de tiempo dentro de un cierto rango $(t_{min}, \ldots, t_{max})$ De este modo, una observación puede expresarse mediante un conjunto de datos $X(t_j)$. En análisis funcional se asume que las muestras son observaciones de un conjunto de datos continuos, por lo cual podríamos considerar estas muestras como observaciones del conjunto de datos continuo $\mathbf{x} = \{X(t); t \in (t_{min}, \ldots, t_{m\acute{}ax})\}$.

Así , una Variable Funcional es aquella variable aleatoria X que toma valores en un espacio funcional de dimensión infinita (o espacio funcional). Una observación $\mathbf{x}$ de X se llama dato funcional [29].

Tomando en cuenta esta definición nos muestra el caso más sencillo es el unidimensional en el que la función toma valores en $T = \mathbf{R}$, pero también puede ser $T = \mathbf{R}^2$, en imágenes, u otras expresiones para casos más complejos.

De la misma forma que en el caso multivariante, denotaremos a nuestra matriz de datos con $\mathbf{X}$ que ahora tiene las medidas de n réplicas o individuos en los que medimos una única variable (tasa de mortalidad, tasa de incidencia del COVID-19, etc) en p momentos distintos $t \in (1 \ldots, p)$. La característica fundamental es que la función generadora los datos está definida no solo en los tiempos medidos sino en cualquier otro tiempo intermedio.

El problema de la representación es muy importante en FDA y hay much´ısimo material al respecto. En próximas secciones se presentarán los métodos más utilizados con sus ventajas e inconvenientes.

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 FDA 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 FDA".

En resumen, trabajaremos con datos demográficos a los que dotaremos de una estructura funcional para acceder a las ventajas que tiene el análisis de este tipo de datos.

Los objetivos del análisis de datos funcionales son esencialmente los mismos que los de cualquier otra rama de la estadística.

- Representar los datos de manera que ayuden a un análisis adicional.
- Mostrar los datos para resaltar varias características.
- Estudiar importantes fuentes de patrón y variación entre los datos.

5.4. Representación de los datos: Suavizado e interpolación

5.4.1. Muestras de datos funcionales

Hemos visto que, para cada individuo i, disponemos de un vector discreto de valores medidos en distintos momentos del tiempo $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. La primera tarea será convertir este conjunto de valores discretos en una función $x_i(t)$ calculable para cualquier valor del argumento t y no solamente para los valores que hemos medido. Si suponemos que el proceso se mide sin error el procedimiento para hacerlo es la *interpolación* mientras que, si suponemos que las observaciones se miden con error usaremos algún tipo de suavizado.

En todo el proceso supondremos que las funciones que utilizamos para describir los datos son *suaves*. Por suave entendemos que la función tenga una o más derivadas de distintos órdenes. Los datos observados puede que no sean *suaves* debido al posible error de medida de los mismos. En muchos casos es difícil, si tenemos un alto grado de variabilidad alrededor de cada x_i, obtener un estimador estable $x(t_i)$ o de sus derivadas.

Para un vector de observaciones $\mathbf{x} = (x_1, \ldots, x_p)$ no es obvio inferir las propiedades de suavidad de la función latente x debido al ruido introducido por el proceso de medida.

En general, tendremos que

$$x_j = x(t_j) + E_j$$

donde el término de error E_j contribuye a que los datos brutos no sean tan suaves como deberían. Una de las principales tareas al trabajar con este tipo de datos sería intentar filtrar el ruido tan eficientemente como sea posible. En otros casos podemos dejar el ruido en los datos y requerir la suavidad en los resultados del análisis en lugar de en los datos.

En notación vectorial podemos escribir

$$\mathbf{x} = x(\mathbf{t}) + \mathbf{e}$$

donde $\mathbf{x}$, $x(\mathbf{t})$ y $\mathbf{e}$ son vectores columna de longitud p.

5.4.2. Representando funciones mediante bases

Un sistema base es un conjunto de funciones conocidas φ_k que son matemáticamente independientes unas de otras y tienen la propiedad de que cualquier función puede ser aproximada tomando una suma ponderada o combinación lineal de un número suficientemente grande K de estas funciones. La base más conocida es la colección de monomios

$$\varphi_0(t) = 1, \varphi_1(t) = t, \varphi_2(t) = t^2, \varphi_3(t) = t^3, \ldots, \varphi_k(t) = t^k, \ldots$$

Otra base conocida es el sistema de series de Fourier

$$\varphi_0(t) = 1, \varphi_{2r-1}(t) = sin(r\omega t), \varphi_{2r}(t) = cos(r\omega t), (r = 1, 2, \ldots)$$

Esta base es periódica, y el parámetro ω determina el período $2\pi/\omega$. Hay muchas otras bases de funciones que se pueden utilizar en este contexto basadas, por ejemplo, en B-splines.

Las bases pueden expresar una función x como una expansión lineal

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

Si denotamos con $\mathbf{c}$ el vector de longitud k con coeficientes c_k y con φ al vector funcional cuyos elementos son las funciones de la base φ_k, podemos expresar la ecuación anterior en notación matricial como:

$$x = \mathbf{c}^l \varphi = \varphi^l \mathbf{c}$$

La dimensión de la expansión es K aunque el problema es potencialmente infinitodimensional. Aunque seleccionemos una aproximación discreta el problema no es equivalente al análisis multivariante tradicional ya que depende, en gran medida, de la base seleccionada φ.

Una representación exacta o *interpolación* se consigue cuando $K = p$, en el

sentido de que podemos elegir coeficientes c_k para los que $x(t_j) = x_j$ para cada j. Por tanto, el grado en el que se suavizan los datos x_j está determinado por el número de funciones en la base K de forma que esta cantidad puede considerarse como un parámetro que depende de las características de los datos y no como un número fijado *a priori*.

Debe tratarse de elegir una base que refleje las características de los datos. Normalmente elegimos un número relativamente pequeño de funciones en la base. De acuerdo con [89], cuanto más pequeño sea el número de funciones, mejor se reflejan ciertas características de los datos, por ejemplo

- Tenemos más grados de libertad para contrastar hipótesis y calcular intervalos de confianza adecuados
- Se requieren menos cálculos
- Es más probable que los propios coeficientes sean descriptores interesantes de los datos desde una perspectiva aplicada

En general no hay una base de funciones que funcione mejor en todas las aplicaciones. Como regla general, utilizaremos bases de Fourier cuando los datos son periódicos y bases formadas por B-splines para datos no periódicos.

5.4.3. El sistema de Fourier para datos periódicos

Como ya hemos mencionado antes, uno de los sistemas de funciones más conocidos es el formado por las series de Fourier, en el que la representación suavizada de una función es de la forma

$$\hat{x}(t) = c_0 + c_1 sin(\omega t) + c_2 cos(\omega t) + c_3 sin(2\omega t) + c_2 cos(2\omega t) + \dots$$

definido por la base

$$\varphi_0(t) = 1, \varphi_{2r-1}(t) = sin(r\omega t), \varphi_{2r}(t) = cos(r\omega t), (r = 1, 2, \dots)$$

Si los valores de t_j están igualmente espaciados en $\mathbf{T}$ y el periodo es igual a la longitud del intervalo $\mathbf{T}$, entonces la base es ortogonal en el sentido de que la

matriz de productos cruzados $\mathbf{\Phi}^{'}\mathbf{\Phi}$ es diagonal. Es posible encontrar los coeficientes de forma eficiente cuando p es un múltiplo de 2, lo que hace que este tipo de bases se utilicen con frecuencia cuando tenemos series de tiempo largas, aunque técnicas más novedosas como los B-splines o wavelets pueden incluso superar su eficiencia computacional. Otra ventaja de este tipo de funciones es que es muy fácil obtener sus derivadas cuando es necesario.

Las series de Fourier son ampliamente conocidas por lo que es necesario detallas sus problemas ya que ningun sistema de funciones debe elegirse sin un estudio crítico previo.

Estas bases son especialmente útiles para funciones que son extremadamente estables en el sentido de que no presentan fuertes variaciones locales y la curvatura tiende a ser del mismo rango de orden en todas partes. La periodicidad de las series debe reflejarse de alguna manera en los datos como ocurre, por ejemplo, con datos climáticos.

Las bases no son apropiadas para datos en los que se sospecha que reflejan discontinuidades en la propia función o en sus derivadas.

5.4.4. El sistema de splines para datos no periódicos

En datos no periódicos las funciones basadas en splines son la elección más común sustituyendo a las bases de polinomios. Los splines combinan la facilidad de cálculo de los polinomios con una flexibilidad mucho mayor que se consigue con solamente un pequeño número de funciones en la base.

El primer paso para definir un spline es dividir el intervalo en el que se va a aproximar la función en L subintervalos separados por valores τ_l, $l = 1, \ldots, L$ y que los llamaremos *puntos de corte* o *nodos*. Para cada intervalo, un *spline* es un polinomio de orden m. El orden del polinomio es el número de constantes que se necesitan para definirlo. Los polinomios adyacentes se juntan de forma suave en los puntos de corte que los separan para polinomios de grado mayor de 1, así los valores de la función se restringen para que que sean iguales en las uniones. También las derivadas hasta el orden $m - 2$ son iguales en las uniones. Si no hay puntos de corte interiores, el spline es simplemente un polinomio.

En resumen, una función spline está determinada por dos cosas:

- El orden de los segmentos polinómicos.
- La secuencia de nodos τ

Construimos a continuación sistema de funciones base basado en splines especificando las funciones $\varphi_k(t)$ que tendrán las siguientes propiedades:

- Cada función de la base $\varphi_k(t)$ es un spline definido por un orden m y una secuencia de nodos τ.
- Como un múltiplo de un spline es también un spline, y como las sumas y diferencias también lo son, cualquier combinación lineal de estas funciones base es también un spline
- Cualquier spline definido por m y τ puede expresarse como una combinación lineal de las funciones de la base.

El sistema de B-splines más popular es el desarrollado por [20] que, en su libro, muestra otros tipos como los M-splines o los splines naturales. La figura (5.2) muestra una base de splines de orden 4 con 15 nodos interiores.

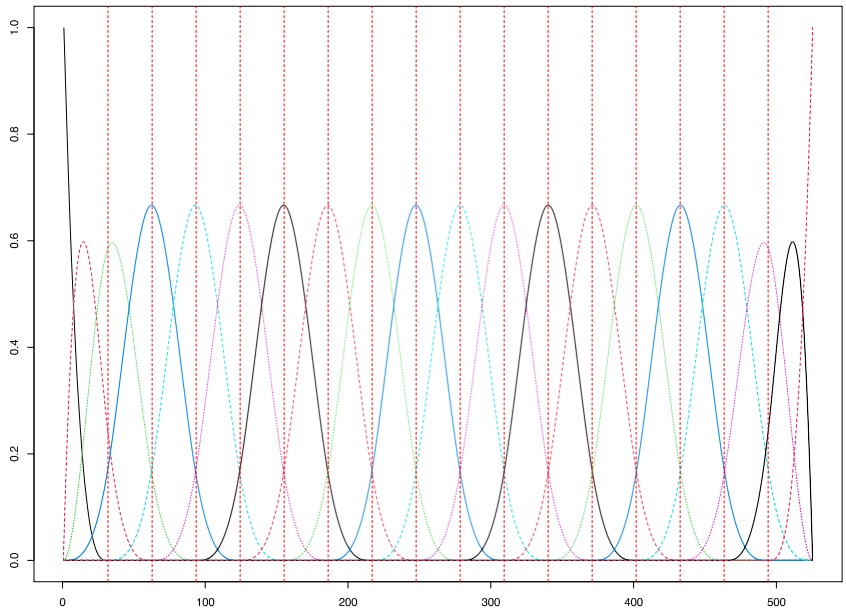


Figura 5.2: Base de B-splines de orden 4 con 15 nodos interiores

En la figura siguiente mostramos la aproximación usando B-splines de los datos mostrados antes.

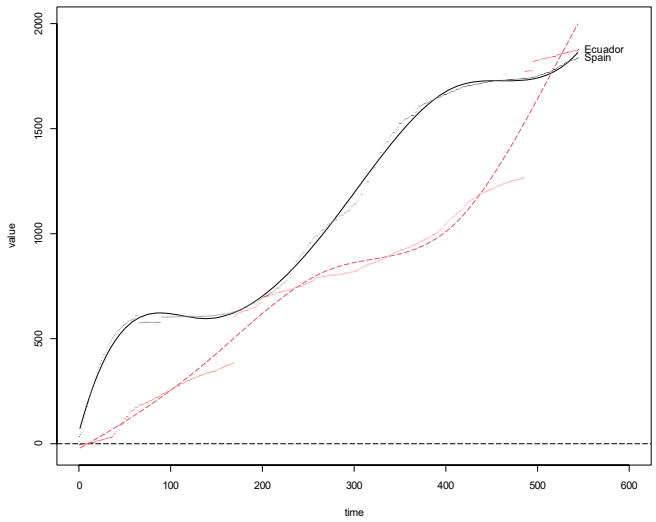


Figura 5.3: Aproximación usando B-splines de la tasa de muertes por COVID-19 en España y Ecuador. Splines de orden 4 y 7 nodos.

Podemos observar como los saltos se han suavizado, mostrando de forma adecuada las tendencias de cada una de las curvas. Para el caso de España se ha introducido una bajada que no debería estar ya que se trata de la incidencia acumulada, si bien refleja el comportamiento de los datos disponibles. También para España los datos se han suavizado de forma que se ha eliminado prácticamente el efecto de la segunda y tercera ola. Para el caso del Ecuador, se han eliminado los saltos bruscos de forma que la curva de evolución obtenida parece más plausible que los datos observados inicialmente.

El suavizado obtenido para los datos depende en gran medida del número de nodos utilizados y del grado de los polinomios. Probablemente a menor número de nodos se obtiene un suavizado mayor a costa de perder algunas de las características de la curva, como el efecto de las dos olas en el caso de España. Por ejemplo, si utilizamos 20 nodos (Figura 5.4) la curva suave está más cerca de los datos.

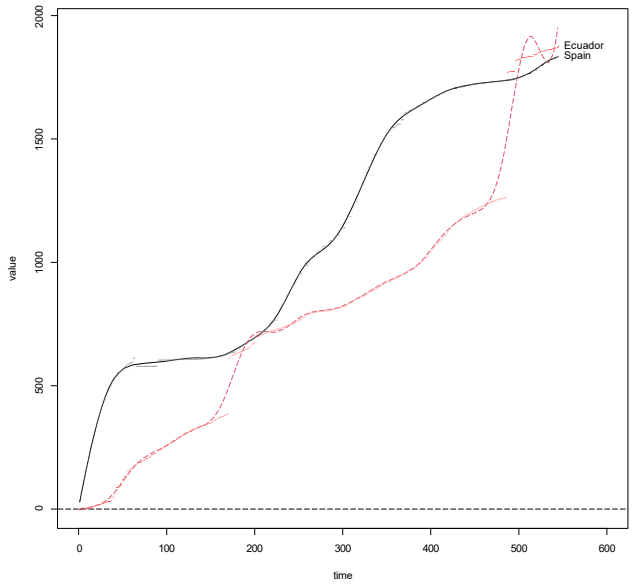


Figura 5.4: Aproximación usando B-splines de la tasa de muertes por COVID-19 en España y Ecuador. Splines de orden 4 y 20 nodos.

Los datos de España están muy bien aproximados, no tanto los de Ecuador que presenta un efecto no deseado en el final de la curva.

5.4.5. Otros sistemas de base útiles

Hay muchos otros sistemas de bases que pueden ser útiles en el análisis de los datos tales como los wavelets o bases exponenciales, polinómicas, poligonales, etc ...

Wavelets

5.5. Suavizado de los datos mediante Mínimos cuadrados

Antes hemos presentado algunas curvas suavizadas basadas en distintas bases. Presentamos ahora las técnicas para ajustar dichas curvas, concretamente en el método de los Mínimos cuadrados.

Recordemos que nuestro propósito es ajustar el conjunto discreto de observaciones $x_j, j = 1, \ldots, p$ usando un modelo de la forma $x_j = x(t_j) + \mathbb{E}_j$ y que vamos a usar una base para $x(t)$ de la forma

$$x(t) = \sum_{k}^{K} c_k \varphi_k(t) = \mathbf{c}^{\prime} \varphi$$

El vector $\mathbf{c}$ de longitud K contiene los coeficientes c_k. Definamos la matriz $\mathbf{\Phi}$ de orden $p \times K$ que contiene los valores de $\varphi_k(t_j)$.

5.5.1. Mínimos cuadrados ordinarios

Un suavizados lineal simple puede obtenerse si determinamos los coeficientes de la expansión c_k minimizando la función

$$SCR(\mathbf{x}|\mathbf{c}) = \sum_{j=1}^{p} \left(x_j - \sum_{k}^{K} c_k \varphi_k(t_j) \right)^2$$

o, en forma matricial

$$SCR(\mathbf{x}|\mathbf{c}) = (\mathbf{x} - \mathbf{\Phi c})^{\prime}(\mathbf{x} - \mathbf{\Phi c})$$

Tomando la derivada con respecto a $\mathbf{c}$ tenemos

$$2\mathbf{\Phi}\mathbf{\Phi}^{\prime}\mathbf{c} - 2\mathbf{\Phi}^{\prime}\mathbf{x} = 0$$

Resolviendo para $\mathbf{c}$ obtenemos que el estimador mínimo cuadrático $\hat{\mathbf{c}}$ es

$$\hat{\mathbf{c}} = (\mathbf{\Phi}^{\prime}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\prime}\mathbf{x}$$

El vector $\hat{\mathbf{y}}$ de valores ajustados es

$$\hat{\mathbf{y}} = \mathbf{\Phi}\hat{\mathbf{c}} = \mathbf{\Phi}(\mathbf{\Phi}^{\prime}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\prime}\mathbf{x}$$

Usaremos los Mínimos cuadrados ordinarios cuando suponemos 2ue los residuales $\mathbb{E}_j$ son independientes e idénticamente distribuidos con media cero y varianza

constante.

5.5.2. Mínimos cuadrados ponderados

Si tenemos errores autocorrelacionados los Mínimos cuadrados ordinarios no son adecuados. Necesitamos ponderar los residuales de forma diferencial de la forma siguiente.

$$SCR(\mathbf{x}|\mathbf{c}) = (\mathbf{x} - \mathbf{\Phi c})^{l}\mathbf{W}(\mathbf{x} - \mathbf{\Phi c})$$

Donde **W** es una matriz simétrica, definida positiva que permite la ponderación diferencial de los residuales. Si conocemos la matriz de vaianzas-covarianzas $\mathbf{\Sigma}_e$ entre los residuales, entonces podemos tomar

$$\mathbf{W} = \mathbf{\Sigma}_e^{-1}$$

Cuando la desconocemos y la estimación no es posible normalmente tomamos una matriz diagonal con los recíprocos de la varianza del error asociados con los x_j's. Si $\mathbf{W} = \mathbf{I}$ el procedimiento coincide con el de Mínimos cuadrados ordinarios.

5.5.3. Elección del número K de funciones en la base

En principio, como ya apuntábamos en los ejemplos anteriores, cuanto más grande sea el orden K mejor es el ajuste pero al precio de ajustar ruido que desearíamos eliminar. Por otra parte, si elegimos un valor demasiado pequeño, podemos ignorar aspectos importantes de la función que deseamos tener en cuenta.

Para valores grandes de K y p, el sesgo para estimar $x(t)$,

$$Sesgo[\hat{x}(t)] = x(t) - E[\hat{x}(t)]$$

es pequeño. Si $K = p$ el sesgo es cero.

No solamente estamos interesados en el *sesgo* sino también en la *varianza* del estimador.

$$Var[\hat{x}(t)] = E[[\hat{x}(t) - E[\hat{x}(t)]]^2]$$

y en el caso de que $K = n$ es muy alta. Para reducir la varianza hay que tomas valores menores de K pero no tanto que el sesgo sea inaceptable.

Una forma de expresar lo que queremos alcanzar realmente es el error cuadrático medio

$$ECM[\hat{x}(t)] = E[[\hat{x}(t) - x(t)]^2]$$

que se suele llamar también la función de pérdida $\mathbf{L}^2$.

Teniendo en cuenta que

$$ECM[\hat{x}(t)] = Sesgo[\hat{x}(t)]^2 + Var[\hat{x}(t)]$$

Debemos tolerar un pequeño sesgo si se produce una disminución importante en la variabilidad.

Podríamos desarrollar procedimientos para la elección de K basados en conceptos similares a los que usamos para seleccionar modelos en regresión.

5.5.4. Cálculo de la variabilidad maestral y Límites de confianza

De la misma manera que en los modelos lineales generales, el estimador del vector de coeficientes $\mathbf{c}$ de la expansión de la base $x = \mathbf{c}^{\prime}\varphi$ es un estimador lineal que define una aplicación lineal del vector de datos observado $\mathbf{x}$ en el estimado. Sabemos que si una variable aleatoria X tiene una distribución normal con matriz de varianzas-covarianzas Σ_x, entonces la variable aleatoria $\mathbf{Ax}$ definida por cualquier matriz $\mathbf{A}$ tiene matriz de varianzas-covarianzas

$$\mathbf{Ax} = \mathbf{A}\Sigma_x\mathbf{A}^{\prime}$$

El modelo para el vector de datos $\mathbf{x}$, $x(\mathbf{t})$, puede entenderse como un efecto fijo con varianza cero. Entonces, la matriz de varianzas-covarianzas de x usando el modelo $\mathbf{x} = x(\mathbf{t}) + E$ es la matriz de varianzas-covarianzas Σ_e del vector de residuales E. Tenemos que usar la información de los residuales calculados para reemplazar la matriz poblacional Σ_e por un estimador muestral razonable $\hat{\Sigma}_e$.

En nuestro caso

$$\mathbf{A} = (\mathbf{\Phi}^{\prime}\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\prime}\mathbf{W}$$

obtenemos

$$Var[\mathbf{c}] = (\mathbf{\Phi}^{\prime}\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\prime}\mathbf{W}\Sigma_e\mathbf{W}\mathbf{\Phi}(\mathbf{\Phi}^{\prime}\mathbf{W}\mathbf{\Phi})^{-1}$$

Para el modelo estándar suponemos que $\Sigma_e = \sigma^2\mathbf{I}$ y obtenemos el resultado simple que aparece en los libros de análisis de regresión

$$Var[\mathbf{c}] = \sigma^2(\mathbf{\Phi}^{\prime}\mathbf{\Phi})^{-1}$$

Para el caso de los datos funcionales, la interpretación del vector de coeficientes $\mathbf{c}$ no suele ser de interés. Estamos interesados mas bien en el conocimiento de la varianza muestral de algunas cantidades calculadas a partir de estos coeficientes. Por ejemplo, podemos querer conocerla varianza muestral del ajuste a los datos definido por $x(t) = \varphi(t)^{\prime}\mathbf{c}$ que es

$$Var[x(t)] = \varphi(t)^{\prime}Var[\mathbf{c}]\varphi(t)$$

y las varianzas de todos los valores ajustados correspondientes a los valores muestrales t_j están en la diagonal de la matriz

$$Var[\hat{\mathbf{x}}] = \mathbf{\Phi}Var[\mathbf{c}]\mathbf{\Phi}^{\prime}$$

que en el modelo estándar ajustado mediante Mínimos cuadrados no ponderados se reduce a

$$Var[\hat{\mathbf{x}}] = \sigma^2\mathbf{\Phi}(\mathbf{\Phi}^{\prime}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\prime}$$

5.5.5. Estimación de Σ_e

Para una buena estimación de las varianzas muestrales necesitamos buenos estimadores de las varianzas y covarianzas entre los residuales $\mathbb{E}_j$. Si aceptamos el modelo estándar, el estimador de σ^2 es

$$s^2 = \frac{1}{n-K} \sum_{j=1}^{p} (x_j - \hat{x}_j)^2$$

Usando este estimador se puede desarrollar una estrategia para elegir K consistente en añadir funciones a la base hasta que s^2 se reduzca sustancialmente.

Otra estrategia para estimar Σ_e para N pequeño, o incluso $N = 1$ es suponer una estructura autorregresiva para los residuales.

Cuando disponemos de un número sustancial de curvas podemos intentar buscar estimadores más sofisticados de Σ_x, por ejemplo, Estimar la matriz completa de covarianzas a partir de la matriz $\mathbf{E}_{N\times p}$ de residuales mediante

$$\hat{\Sigma}_e = (N-1)^{-1}\mathbf{E}^t\mathbf{E}$$

Aunque para esta estimación completa el tamaño muestral tiene que ser muy grande.

5.5.6. Límites de confianza

Como en muchas otras situaciones, es posible calcular intervalos de confianza sumando y restando múltiplos de los errores estándar que son las raíces cuadradas de las varianzas del ajuste obtenido. Sabemos que los intervalos al 95 % se corresponden con aproximadamente dos veces el error estándar. Estos se denominan intervalos puntuales ya que reflejan regiones de confianza para valores concretos de t y no para la curva completa.

5.6. Funciones restringidas

Hasta el momento, la única condición que hemos puesto a las funciones es que sean suaves pero en algunos casos es necesario que tengan algunas restricciones adicionales, por ejemplo, que sean estrictamente crecientes o que sean positivas. A continuación describimos someramente estos procedimientos para poner de manifiesto que es posible utilizarlos en la práctica.

5.6.1. Ajuste de funciones positivas

En muchos casos, como el de la tasa acumulada de muertes por COVID que mostrábamos antes, los datos se corresponden con funciones estrictamente positivas. Aunque los datos pueden ser cero, se puede entender que lo que ocurre es que la función toma valores muy pequeños en ese punto.

Puede definirse una función de suavizado positiva x como la exponencial de una función no restringida W,

$$x(t) = e^{W(t)}$$

de forma que W es el logaritmo x. Pueden usarse bases diferentes para el logaritmo en lugar de e, por ejemplo 2 o 10.

Como $W(t)$ puede ser positiva o negativa y no está restringida de ninguna manera, es razonable expandir W en términos de un conjunto de funciones base

$$W(t) = \sum_k c_k \varphi_k(t)$$

Podemos usar las bases que hemos descrito en los puntos anteriores.

El problema fundamental, que no tratamos aquí con detalle, es que para ajustar los coeficientes se necesitan métodos numéricos. Este tipo de métodos parten de un valor inicial de $W(t)$ y decrecen hasta la convergencia. En la mayor parte de los casos comenzar con $W = 0$ funciona correctamente.

5.6.2. Ajuste de funciones estrictamente monótonas

Para una función x estrictamente monótona, se verifica que la primera derivada Dx (velocidad) se supone que es positiva. Podemos usar

$$x(t) = e^{W(t)}$$

expresando Dx como la exponencial de una función no restringida W para obtener

$$Dx(t) = e^{W(t)}$$

Integrando ambos lados de la ecuación tenemos

$$x(t) = C + \int_t^{-t} e^{W}$$

donde C es es una constante que debemos de estimar de los datos.

La solución al problema puede hacerse usando ecuaciones diferenciales que no describimos aquí con detalle.

5.7. Estadística descriptiva para datos funcionales

De la misma manera que en las estadística tradicional es posible usar estadísticosde resumen con datos funcionales. Existen análogos funcionales no solamente de los principales estadísticosdescriptivos sino también de mochas de las técnicas de la Estadística Inferencial como los Métodos de Regresión, o el Análisis de la Varianza, entre otros.

5.7.1. Media y Varianza

Podemos calcular la función media como

$$\hat{x}(t) = n^{-1} \sum_{i=1}^{n} x_i(t)$$

que es la media de las funciones en cada punto.

De la mima manera podemos calcular la función varianza

$$Var_X(t) = (n-1)_{-1} \sum_{i=1}^{n} [x_i(t) - \hat{x}(t)]^2$$

La desviación típica o estándar será la raíz cuadrada de la varianza.

5.7.2. Covarianza y correlación

La función *covarianza* resume la dependencia de los registros para diferentes valores de los argumentos y se calcula para todos los valores de t_1 y t_2 mediante la

siguiente expresión

$$Cov_{X,Y}(t_1, t_2) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t_1) - \hat{x}(t_1)][x_i(t_2) - \hat{x}(t_2)]$$

La función de correlación asociada es, entonces

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

Ambos son los análogos funcionales de las matrices de covarianzas y correlaciones en el análisis multivariante tradicional.

5.8. Análisis de Componentes Principales (ACP) para datos funcionales

De la misma forma que en Análisis Multivariante clásico, es posible definir un Análisis de Componentes Principales para datos funcionales que nos permite una exploración inicial de los datos.

Por muchas razones, el análisis de componentes principales (PCA) de datos funcionales es una técnica clave a considerar. En primer lugar, la bibliografía consultada indica que, después de los pasos preliminares de registro y visualización de los datos, el usuario quiere para explorar esos datos para ver los rasgos que caracterizan las funciones típicas.

En primer lugar reformularemos el análisis de Componentes principales descrito antes para ponerlo en la forma que usaremos después para los datos funcionales para pasar a la generalización para este tipo de datos.

5.8.1. ACP para datos multivariantes

Disponemos de una matriz de datos multivariantes $\mathbf{X} = (x_{ij})$ un concepto clave utilizado en muchas técnicas multivariantes es el de combinaciones lineales de valores de las variables

$$f_i = \sum_{j=1}^{p} \beta_j x_{ij}, \; i = 1, \ldots, n$$

donde β_j es el coeficiente de ponderación aplicado a los valores x_{ij} de la j-ésima variable. Podemos expresar también esto en forma compacta como

$$f_i = \beta^{l} \mathbf{x}_i, \; i = 1, \ldots, n$$

donde β es el vector $(\beta_1, \ldots, \beta_p)^{l}$ y $\mathbf{x}_i$ es el vector $(x_{i1}, \ldots, x_{ip})$

En el caso multivariante buscamos ponderaciones para destacar las mayores fuentes de variación presentes en los datos. Buscamos un conjunto de ponderaciones normalizadas que maximicen la varianza de los f_i's. El procedimiento sería como sigue:

- Se trata de buscar el vector $\xi_1 = (\xi_{11}, \ldots, \xi_{p1})^{l}$ para el que la combinación lineal de los valores

$$f_{i1} = \sum_{j} \xi_{j1} x_{ij} = \xi_1^{l} \mathbf{x}_i$$

tenga la mayor media cuadrática $n^{-1} \sum_i f_{i1}^2$ y sujeto a la restricción

$$\sum_{j} \xi_{j1}^2 = ||\xi_1|| = 1$$

- Llevar a cabo los pasos siguientes hasta obtener el número deseado de componentes. Eln el paso m.ésimo, calcular un nuevo vector ξ_m de componentes ξ_{jm} y nuevos valores $f_{im} = \xi_m^{l} \mathbf{x}_i$, de forma que tengan máxima media cuadrática, sujeto a la restricción de que $||\xi_m||^1 = 1$ y a las $m - 1$ restricciones adicionales de que

$$\sum_{j} \xi_{jk} \xi_{jm} = \xi_k^{l} \xi_m = 0$$

La motivación para el primer paso es que, al maximizar el cuadrado medio, estamos identificando el modo de variación más fuerte e importante en las variables. La restricción de suma de cuadrados unitaria para los pesos es esencial para poder definir bien el problema; sin ella, los cuadrados medios de los valores de la combinación

lineal podrían hacerse arbitrariamente grandes.

En segundo lugar, buscamos los modos de variación más importantes en la parte no explicada, pero añadiendo la restricción de que los pesos de las nuevas componentes sean ortogonales a los identificados anteriormente, de modo que están indicando caracter´ıaticas nuevas de los datos. Por supuesto, la cantidad de la variación recogida por cada componente, medida en términos de $n^{-1} \sum_i \gamma_i^2$, disminuirá en cada paso. El número de componentes interesantes estará, normalmente, muy por debajo del número de variables p, es decir, conseguiremos una considerable reducción de la dimensión.

Es conocido también que la restricción de longitud 1 en los vectores que definen las componentes, los hace únicos salvo el signo, es decir, podemos cambiar el signo de todos sin que cambie la varianza recogida.

Para más detalles ver el capítulo correspondiente a las Componentes Principales descrito anteriormente.

5.8.2. Generalización del ACP para datos funcionales

Definición de ACP para datos funcionales:

Suponemos que ya tenemos los datos en forma de funciones, es decir, el índice discreto j en x_{ij} se reemplaza por el índice continuo s en $x_i(s)$. Tenemos entonces un conjunto de réplicas de la función $x_i(s)$; $i = 1, \ldots, n$. Para vectores, la forma correcta de combinar el vector de pesos β con el vector de datos $\mathbf{x}$ era el producto escalar

$$\beta^t \mathbf{x} = \sum_j \beta_j x_j$$

Cuando tanto β como x son funciones $\beta(s)$ y $x(s)$, las sumas se reemplazan por integrales obteniendo,

$$\beta x = \int \beta(s) x(s) ds$$

Entonces, en lugar de coeficientes tenemos ahora funciones para definir las componentes principales. Las puntuaciones sobre las componentes principales son ahora:

$$f_i = \int \beta x_i = \int \beta(s) x_i(s) ds$$

En el primer paso buscamos una función $\xi_1(s)$ que maximiza

$$n^{-1} \sum_i f_{i1}^2 = n^{-1} \sum_i (\int \xi_1(s) x_i(s))^2 = n^{-1} \sum_i (\int \xi_1 x_i)^2$$

sujeto a

$$\int \xi_1(s)^2 ds = 1$$

que es el análogo continuo de la restricción de suma de cuadrados unidad.

Para las siguientes componentes, la función ξ_m tiene que satisfacer también las restricciones de ortogonalidad

$$\int \xi_k \xi_m = 0, k < m$$

Como antes, cada componente recoge la máxima variabilidad teniendo en cuenta que es ortogonal a las anteriores y por tanto, recoge aspectos nuevos de la información. Las componentes tienen órdenes de importancia decrecientes y solamente unas pocas son necesarias para explicar las características más relevantes de los datos. De nuevo, las componentes están definidas salvo un cambio de signo.

Definición de ACP como una base ortonormal empírica óptima:

Otra forma posible de definir las componentes principales es mediante la búsqueda de un conjunto de K funciones ortonormales ξ_m de forma que las expansión de cada curva en términos de estas funciones base aproxima la curva lo mejor posible. La expansión con estas funciones base ortonormales será de la forma

$$\hat{x}_i(t) = \sum_{k=1}^{K} = f_{ik} \xi_k(t)$$

donde f_{ik} es la puntuación sobre la componente principal $intx_i\xi_k$. Como criterio

de ajuste de una curva individual, tenemos el error cuadrático integrado

$$||x_i - \hat{x}_i||^2 = \int [x(s) - \hat{x}(s)]^2$$

y como medida de la aproximación global

$$SCEPACP = \sum_{i=1}^{n} ||x_i - \hat{x}_i||^2$$

La base que minimiza este criterio es precisamente la formada por las componentes principales tal y como las hemos definido antes en términos de la variabilidad.

ACP y valores y vectores propios:

Suponemos que nuestros valores observados, tanto en el caso multivariante x_{ij} como en el funcional $x(_it)$, han sido centrados restando la media de forma que las nuevas medias son cero.

Es bien conocido que, en el caso multivariante las componentes principales se obtienen dela descomposición en valores y vectores propios de la matriz de covarianzas o correlaciones. Sean $\mathbf{X}_{(n\times p)}$, la matriz que contiene los valores x_{ij}, y el vector $\xi(p)$ que contiene los coeficientes de la combinación lineal. El criterio de los cuadrados medios para encontrar la primera componente principal puede escribirse como

$$\max_{\xi} \xi' \mathbf{X}' \mathbf{X} \xi$$

ya que el vector de puntuaciones sobre la componente principal f_i puede escribirse como $\mathbf{X}\xi$.

Si llamamos $\mathbf{V}$ a la matriz $(p \times p)$ que contiene las varianzas y covarianzas muestrales $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$, el criterio puede expresarse como

$$\max_{\xi} \xi' \mathbf{V} \xi$$

El problema se resuelve encontrando el mayor valor propio ρ de la ecuación

$$\mathbf{V}\xi = \rho\xi$$

Cada uno de los valores propios está asociado con una delas componentes principales como vimos en el capítulo correspondiente.

Para la versión funcional del ACP definimos, en primer lugar, la función de covarianza

$$v(s,t) = n^{-1} \sum_{i=1}^{n} x_i(s)x_i(t)$$

Las funciones que definen las componentes principales $\xi_j(s)$, satisface la ecuación

$$\int v(s,t)\xi(t)dt = \rho\xi(s)$$

para un valor propio apropiado ρ. El lado izquierdo de la ecuación es una trasformación integral $\mathbf{V}$ de la función de pesos ξ definida por

$$V\xi = \int v(.,t)\xi(t)dt \tag{5.1}$$

Esta transformación integral se denomina el operador de covarianza $\mathbf{C}$. Por consiguiente podemos expresar la ecuación directamente

$$V\xi = \rho\xi \tag{5.2}$$

donde ξ es una función propia en lugar de un vector propio.

Al contrario que en el caso multivariante en el que el posible número de valores propios distintos de 0 es el mínimo entre $n-1$ y p, en el caso funcional es normalmente $n-1$ ya que cada función puede tener infinitos valores. Vemos que tanto en el caso multivariante como el funcional, los pasos a seguir para la construcción de las componentes son los mismos.

Visualización del ACP:

La visualización de las componentes puede realizarse representando las funciones que las definen, como en la figura 5.5 que corresponde a los datos de la evolución de las tasas acumuladas de fallecimientos por COVID en países latinoamericanos.

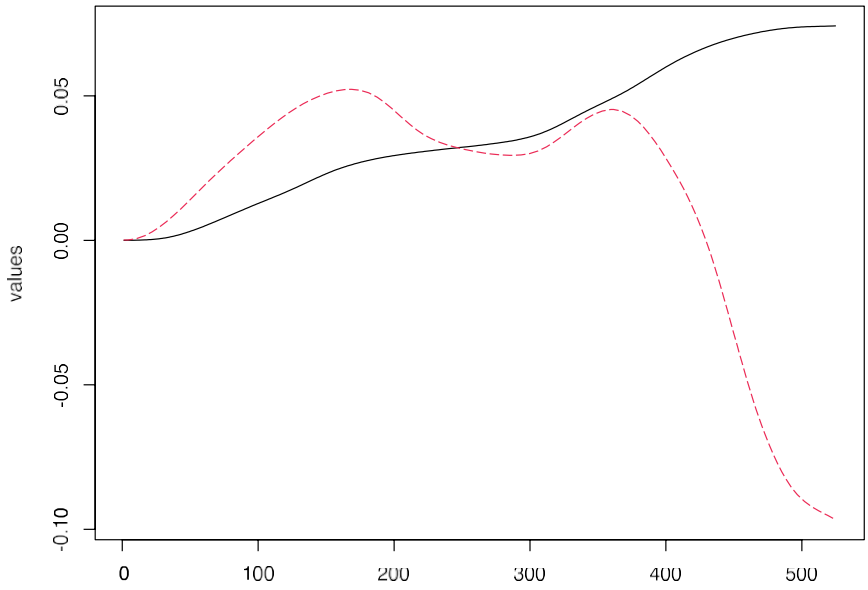


Figura 5.5: Representación de las dos primeras componentes principales de la evolución de las tasas acumuladas de fallecimientos por COVID en países americanos.

La primera componente (en negro) muestra el aumento general de las tasas de mortalidad a lo largo del tiempo. La segunda muestra un crecimiento al principio de la pandemia para mostrar un descenso acusado al final.

La representación más habitual en este contexto es la que se muestra en la figura 5.6. Representa los dibujos de la media global y las funciones obtenidas sumando y restando un múltiplo conveniente de la función que define la componente principal en cuestión. La primera componente muestra el aumento progresivo alrededor de la media. Los países con mayores puntuaciones en esta componente serán aquellos que han presentado mayores aumentos en las tasas. Para la segunda componente vemos que los valores positivos serán para aquellos países con un incremento menor en la última parte de la pandemia mientras que, los que tengan valores negativos serán aquellos que han crecido más en la segunda parte.

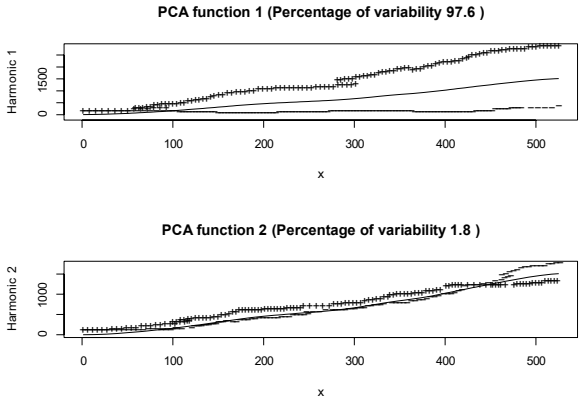


Figura 5.6: Representación de las dos primeras componentes principales de la evolución de las tasas acumuladas de fallecimientos por COVID en países americanos.

Podemos representar también las puntuaciones de los países sobre las dos primeras componentes.

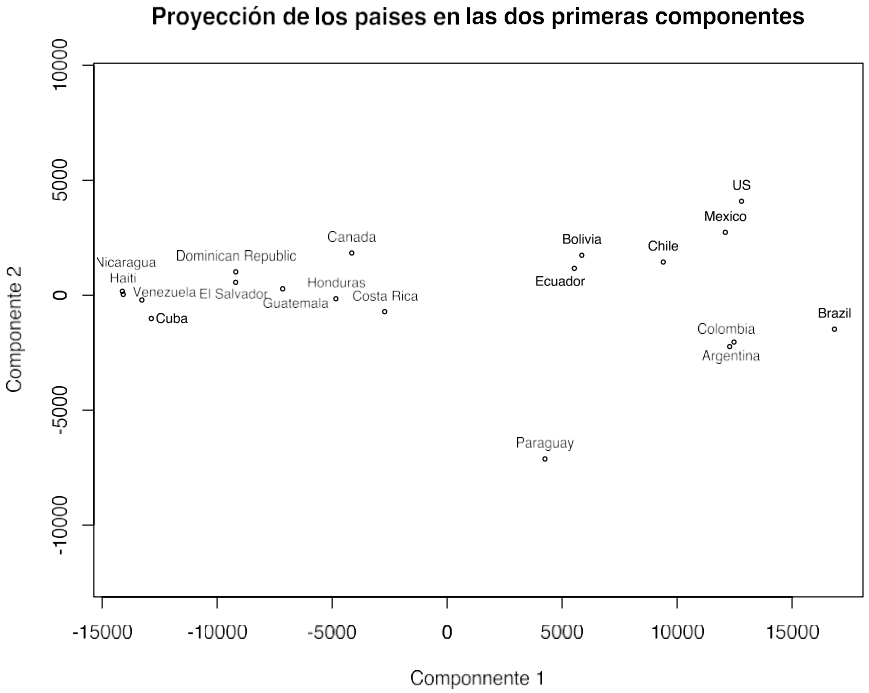


Figura 5.7: Puntuación de los países americanos en las dos primeras componentes principales de la evolución de las tasas acumuladas de fallecimientos por COVID.

La interpretación del gráfico es similar al caso de datos multivariantes. La distancia entre países se interpreta en términos de similitud. Dos países próximos tienen trayectorias de evolución similar, por ejemplo, le evolución de Estados Unidos (US) y México ha sido similar con relación a las tasas de fallecidos por cada 100000 habitantes. Teniendo en cuenta que la primera componente representa el aumento en la mortalidad, en la parte derecha del gráfico se situarán aquellos países que han tenido una mayor mortalidad. En este caso se eliminó Perú ya que tenía una mortalidad mucho mayor que el resto. Exceptuando Perú, la mortalidad ha estado liderada por Brasil seguido de Argentina, Colombia, los Estados Unidos, México, Chile, Bolivia, Ecuador y Paraguay. Los países de Centroamérica y el Caribe parecen haber tenido incidencias y trayectorias similares.

En el gráfico observamos que Paraguay aparece alejado de los demás países con valores más bajos (negativos) en la segunda lo que significa que su mortalidad ha aumentado considerablemente en los últimos meses.

Rotación de las componentes principales:

Vimos antes que los funciones ξ_m 'pueden entenderse como un conjunto de K ortonormal para expandir las curvas que además minimiza el criterio de la suma de cuadrados. Esto no significa que no haya otro conjunto de funciones ortonormales que funcione igual de bien, de hecho, si usamos ξ para referirnos al vector de funciones $(\xi_1, \ldots, \xi_K)^|$, un conjunto ortonormal igual de bueno está definido por

$$\psi = \mathbf{T}\xi$$

donde υ es una matriz ortogonal de orden K, es decir, $\mathbf{T}^|\mathbf{T} = \mathbf{T}\mathbf{T}^| = \mathbf{I}$. Desde el punto de vista geométrico, *psi* es una rotación r´ıgida de ξ. Las primeras componentes ya no tienen porqué recoger la misma cantidad de variabilidad, pero la suma de todas ellas es la misma y la capacidad de la base ortonormal $\psi_1, \ldots, \psi_K$ para aproximar las curvas es la misma.

Sea $\mathbf{B}$ una matriz $K \times p$ que representa las funciones de las K primeras componentes principales $\xi_1, \ldots, \xi_K$. Por el momento supongamos que $\mathbf{B}$ tiene como m-ésima fila los valores $\xi_m(t_1), \ldots, \xi_m(t_p)$ para p valores del argumento en el intervalo T igualmente espaciados. La matriz correspondiente $\mathbf{A}$de valores de las funciones rotadas

$\psi = \mathbf{T}\xi$ estará dado por

$$\mathbf{A} = \mathbf{TB}$$

La estrategia VARIMAX para elegir la rotación ortogonal $\mathbf{T}$ es simplemente maximizar la variación de los valores a^2_{mj} colocados en un único vector. Como $\mathbf{T}$ es una matriz de rotación, la suma de cuadraros total será la misma sea cual sea la rotación aplicada, es decir,

$$\sum_m \sum_j a^2_{mj} = tr\mathbf{A}^\intercal\mathbf{A} = tr\mathbf{B}^\intercal\mathbf{T}^\intercal\mathbf{TB} = tr\mathbf{B}^\intercal\mathbf{B}$$

De esta forma la varianza de los a^2_{mj} es máxima si cada uno de ellos tiende ser grande o a ser cero, con lo que los coeficientes son más fáciles de interpretar. Los métodos para llevar a cabo la rotación VARIMAX pueden encontrarse en cualquier libro clásico de Análisis Factorial, como por ejemplo [43]. Además de este criterio, pueden usarse muchos otros que encontraremos también en los libros clásicos de Análisis Multivariante.

5.8.3. Métodos de cálculo para en ACP funcional

Supongamos ahora que tenemos n curvas x_i para las que hemos hecho los suavizados necesarios y la resta de la media. Sea $v(s, t)$ la función de covarianza muestral calculada de los datos. Consideraremos, a continuación, diversas estrategias para resolver el problema de los valores y vectores propios. En todos los casos, vamos a convertir el problema funcional continuo en una tarea de valores y vectores propios equivalente para una matriz.

Discretización de las funciones:

Una forma simple de resolver el problema es discretizar las funciones x_i en una cuadr´ıcula de p valores igualmente espaciados s_j en el intervalo T. Esto nos proporciona una matriz de datos $\mathbf{X}$ de tamaño $(n \times p)$ sobre la que se puede utilizar un procedimiento estándar de componentes principales calculando los valores y vectores

propios de la matriz de covarianzas

$$\mathbf{Cu} = \lambda\mathbf{u}$$

A partir de aquí tenemos que transformar las componentes principales en términos funcionales.

La matriz de varianzas-covarianzas muestral $\mathbf{C} = n^{-1}\mathbf{X}'\mathbf{X}$ tendrá elementos $c(s_j, s_k)$ donde $c(s, t)$ es la función de covarianza muestral. Dada una función ξ, sea $\tilde{\xi}$ el vector de p valores $\xi(s_j)$. Sea $w = T/p$ donde T es la longitud del intervalo $\mathbf{T}$. Entonces para cada s_j,

$$V\xi(s_j) = \int v(s_j, s)\xi(s)ds \approx w \sum v(s_j, s_k)\tilde{\xi}_k$$

de forma que la ecuación funcional $Cxi = \rho\xi$ tiene una forma aproximada discreta

$$w\mathbf{V}\tilde{\xi} = \rho\tilde{\xi}$$

Las soluciones de esta última se corresponden con las de anterior tomando $\rho = w\lambda$. La aproximación discreta a la normalización $\int \xi(s)^2 ds = 1$ es $||\tilde{\xi}||^2 = 1$, de forma que $\tilde{\xi} = w^{-1/2}\mathbf{u}$ si $\mathbf{u}$ es un vector propio normalizado de $\mathbf{V}$. Para obtener todos los valores de la función podríamos utilizar un método de interpolación. Si los valores de s_j están muy cercanos, el método de interpolación no tiene una influencia importante en el proceso.

Expansión de las funciones mediante la base:

Otra forma de reducir el problema a una forma discreta o matricial es expresar cada función x_i como una función de las bases conocidas φ_k. Como vimos antes, el número de funciones base K depende de varias consideraciones normalmente rela- cionadas con el tipo de suavizado que se desea realizar.

Supongamos que cada función tiene la expansión en la base

$$x_i(t) = \sum_{k=1}^{K} c_i k\varphi_k(t)$$

Podemos escribirla de forma más compacta definiendo el vector de funciones $\mathbf{x}$ que tiene componentes $x_1, \ldots, x_n$ y el vector φ que tiene como componentes $\varphi_1, \ldots, \varphi_K$, de forma que podemos expresar simultáneamente la expresión de todas las n curvas como

$$\mathbf{x} = \mathbf{C}\varphi$$

donde la matriz de coeficientes $\mathbf{C}$ es $n \times K$. En términos matriciales, la la función de varianza y covarianza se puede escribir como

$$v(s, t) = n^{-1}\varphi(s)^{\prime}\mathbf{C}^{\prime}\mathbf{C}\varphi(t)$$

Definimos la matriz simétrica $\mathbf{W}$ de orden $K \times K$ con valores

$$w_{k_1,k_2} = \int \varphi_{k_1}\varphi_{k_2}$$

ó

$$\mathbf{W} = \int \varphi\varphi^{\prime}$$

Para algunas bases los valores de $\mathbf{W}$ se calculan inmediatamente, para otras será necesario algún tipo de integración numérica. Supongamos que la función propia tenga una expansión

$$\xi(s) = \sum_{k=1}^{K} b_k\varphi_k(s)$$

o, en notación matricial

$$\xi(s) = \varphi(s)^{\prime}\mathbf{b}$$

.

Esto nos lleva a

$$\int v(s, t)\xi(t)dt = \int n^{-1}\varphi(s)^{\prime}\mathbf{C}^{\prime}\mathbf{C}\varphi(t)\varphi(t)^{\prime}\mathbf{b} = n^{-1}\varphi(s)^{\prime}\mathbf{C}^{\prime}\mathbf{C}\mathbf{W}\mathbf{b}$$

Luego la ecuación ara el cálculo de las componentes queda como

$$n^{-1}\varphi(s)^{\prime}\mathbf{C}^{\prime}\mathbf{C}\mathbf{W}\mathbf{b} = \rho\varphi(s)^{\prime}\mathbf{b}$$

Como debe verificarse para todo s, implica una ecuación puramente matricial

$$n^{-1}\mathbf{C}^{\prime}\mathbf{C}\mathbf{W}\mathbf{b} = \rho\mathbf{b}$$

Pero, nótese que $||\xi|| = 1$ implica que $\mathbf{b}^{\prime}\mathbf{W}\mathbf{b} = 1$ y, de la misma manera, dos funciones ξ_1 y ξ_2 serán ortogonales si y solo si los correspondientes vectores de coeficientes verifican que $\mathbf{b}^{\prime}{}_1\mathbf{W}\mathbf{b}_2 = 0$. Para obtener las componentes principales, definimos $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{b}$ y resolvemos el problema simétrico equivalente

$$n^{-1}\mathbf{W}^{1/2}\mathbf{C}^{\prime}\mathbf{C}\mathbf{W}^{1/2}\mathbf{u} = \rho\mathbf{u}$$

y calcular $\mathbf{b} = \mathbf{W}^{-1/2}\mathbf{u}$

Si la base es ortonormal, lo que quiere decir que $\mathbf{W} = \mathbf{I}$, el problema se reduce a un ACP estándar sobre la matriz de coeficientes $\mathbf{C}$ y solamente necesitamos la descomposición en valores y vectores propios de la matriz simétrica $n^{-1}\mathbf{C}^{\prime}\mathbf{C}$ de orden K.

Cuadratura numérica más general:

En los casos anteriores, hemos utilizado una estrategia de discretización para aproximar la integral $\int x_i(s)\xi(s)ds$ mediante una suma de valores discretos. La mayor parte de los procedimientos de integración numérica o cuadratura, implican una aproximación de la forma

$$\int f(s)ds \approx \sum_{j=1}^{n} w_j f(s_j) \tag{5.3}$$

Nos centramos ahora en los esquemas de cuadratura de esta forma. Hay tres aspectos de la aproximación que pueden manipularse para cumplir varios objetivos:

- p, el número de valores discretos s_j del argumento
- s_j, los valores del argumento, denominado puntos de cuadratura

- w_j, las ponderaciones, llamadas ponderaciones de cuadratura, asociadas a cada valor de la función en la suma

Aplicando los esquemas de cuadratura del tipo 5.3 al operador V en 5.1, produce una aproximación discreta

$$V\xi \approx \mathbf{V}\mathbf{W}\tilde{\xi} \tag{5.4}$$

donde, como antes, la matriz $\mathbf{V}$ contiene los valores $v(s_j, s_k)$ de la función de covarianza en los puntos de la cuadratura, y $\tilde{\xi}$ es un vector de orden n que contiene los valores $\xi(s_j)$. la matriz $\mathbf{W}$ es una matriz diagonal con las ponderaciones de la cuadratura w_j.

El problema de valores propios equivalente es, entonces

$$\mathbf{V}\mathbf{W}\tilde{\xi} = \rho\tilde{\xi}$$

donde la restricción de ortogonalidad es ahora

$$\tilde{\xi}_m^t \mathbf{W}\tilde{\xi}_m = 1$$

y

$$\tilde{\xi}_{m_1} \mathbf{W}\tilde{\xi}_{m_2} = 0, m_1 \quad m_2$$

Como la mayor parte de los esquemas de cuadratura tienen ponderaciones positivas, podemos poner la ecuación en la forma

$$\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}\mathbf{u} = \rho\mathbf{u}$$

donde $\mathbf{u} = \mathbf{W}^{1/2}\tilde{\xi}$ y $\mathbf{u}^t\mathbf{u} = 1$. Entonces, el procedimiento completo es como sigue:

1. Elegir n, los s_j's y los s_j's.
2. Calcular los valores propios ρ_m y los vectores propios $\mathbf{u}_m$ de $\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}$
3. Calcular $\tilde{\xi}_m\mathbf{W}^{-1/2}\mathbf{u}_m$ Usar una técnica de interpolación para convertir cada vector $\tilde{\xi}_m$ en una función ξ_m

Como es habitual en el contexto de las componentes principales, solamente unas cuantas serán necesarias para explicar el comportamiento de los datos.

Capítulo 6

Biplot para datos funcionales

6.1. Aproximación de los datos observados mediante las componentes principales

Para el caso multivariante es posible aproximar la matriz de datos $\mathbf{X}$ por una de rango menor $\hat{\mathbf{X}}$ usando las componentes principales. Si denotamos con $\mathbf{f}_i$ el vector $(f_{i1}, \ldots, f_{iK})'$ el vector que contiene las puntuaciones del i-ésimo individuo en las K primeras componentes y con ξ_j el vector $(\xi_{j1}, \ldots, \xi_{jK})'$ de los coeficientes de la j-ésima variable en las K componentes seleccionadas, cada elemento x_{ij} de la matriz de datos se puede aproximar como

$$x_{ij} \approx \hat{x}_{ij} = \mathbf{f}_i' \xi_j = \sum_{k=1}^{K} f_{ik} \xi_{jk}$$

Si colocamos todos los vectores $\mathbf{f}_i$ como filas de una matriz $\mathbf{F}_{(n \times K)}$ y los vectores ξ_j como filas de una matriz $\mathbf{\Xi}_{p \times K}$, podemos aproximar la matriz completa como

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{F}\mathbf{\Xi}'$$

Obteniendo la mejor aproximación de la matriz a bajo rango de la misma manera que a partir dela conocida como Descomposición en Valores Singulares (DVS) tal y como la definíamos en capítulos anteriores para la obtención del biplot. (Solamente hemos cambiado la notación para adaptarla a este caso).

Entonces, si $K = 2$ y representamos los $\mathbf{f}_i$ y los ξ_j como puntos en un espacio bidimensional, tenemos un JK-Biplot tal y como fue propuesto por [31] y que puede encontrase también en [33]. De forma gráfica, la aproximación de un elemento de la matriz se realizará mediante el producto escalar de un punto (vector) que representa a una fila $\mathbf{f}_i$ por el vector que representa a una columna ξ_j.

6.2. Construccion del biplot para datos funcionales

Para el caso funcional la generalización es inmediata.

El vector $\mathbf{f}_i = (f_{i1}, \ldots, f_{iK})^\prime$ es igual al anterior. Para un valor t podemos colocar los valores de cada una de las K funciones principales en un vector $\xi(t)$ el vector $(\xi_1(t), \ldots, \xi_K(t))^\prime$ de forma que el valor $x_i(t)$ se puede aproximar como

$$x_i(t) \approx \hat{x}_i(t) = \mathbf{f}_i^\prime \xi(t) = \sum_{k=1}^{K} f_{ik}\xi_k(t)$$

En particular, si $\mathbf{X}$ contiene los datos observados en momentos discretos t_j, $(j = 1, \ldots, p)$ podemos colocar los p vectores $\xi(t_j)$, $(j = 1, \ldots, p)$ como filas de una matriz $\mathbf{\Xi}_{p\times K}$, obtenemos una aproximación de la matriz observada similar al caso multivariante.

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{F}\mathbf{\Xi}^\prime$$

De la misma manera que antes, si $K = 2$ podemos representar los vectores $\mathbf{f}_i$ como puntos en el espacio bidimensional y los vectores $\xi(t)$ para cualquier valor de t para obtener un Biplot. Aunque podemos limitarnos al conjunto de momentos observados t_j, $(j = 1, \ldots, p)$, podríamos hacerlo para cualquier valor de t. Podemos, entonces, representar una trayectoria continua sobre el espacio bidimensional uniendo los puntos $(\xi_1(t), \xi_2(t))$ para distintos valores de t, observados o no, tal y como se muestra en la figura 6.1.

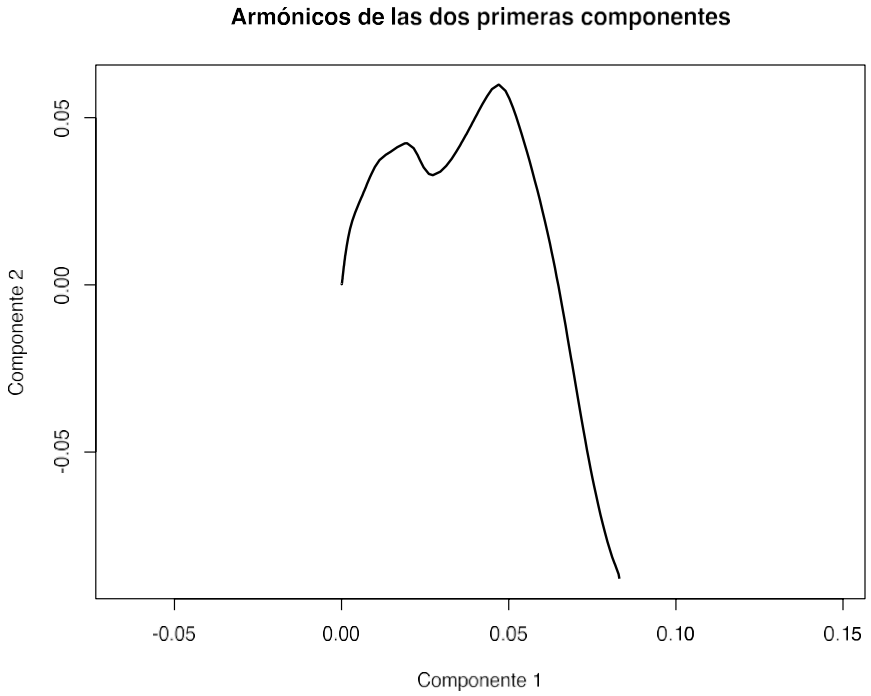


Figura 6.1: Trayectorias resultantes de la representación de los armónicos en el espacio bidimensional.

Para obtener el biplot en cualquier momento t bastaría con unir el punto correspondiente de la trayectoria con el origen de coordenadas y proceder como en el caso del biplot clásico.

De esta forma, hemos definido un biplot para datos funcionales que no ha sido previamente descrito en la literatura.

Lo mismo en el biplot clásico que en el que acabamos de definir para datos funcionales, podemos tener un problema con la magnitud de la escala para los puntos que representan a las filas y los que representan a las columnas ya que, mientras que los coeficientes de las componentes en Ξ están normalizados por columnas y, por tanto, toman valores pequeños, los puntos que representan a las filas dependen de la escala de medida de las variables originales, especialmente en el caso funcional en el que normalmente no se estandarizan los datos. El biplot propuesto sería una combinación de los gráficos de las figuras 5.7 y 6.1 que, como podemos observar, están en escalas muy diferentes. Mientras que el gráfico de las puntuaciones tiene escalas entre -15000 y 15000, el gráfico de los armónicos las tiene entre -0.05 y 0.05.

Si los representamos juntos, no podríamos visualizar las trayectorias.

Para visualizar mejor las relaciones entre filas y columnas puede que sea necesario modificar la escala de los marcadores (puntos fila y columna) sin modificar el producto escalar de los mismos para no cambiar la interpretación. Esto se consigue multiplicando unos por un escalar y dividiendo los otros por el mismo. Si llamamos a a dicho escalar, tenemos

$$\mathbf{X} \approx \hat{\mathbf{X}} = (\frac{1}{a}\mathbf{F})(a\mathbf{\Xi})'$$

El cálculo del valor de a no es inmediato. En el paquete MultBiplotR ([107]) el valor de la constante utilizado es

$$a = \frac{n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} f_{ik}^2}{\bar{p}^{\,1} \sum_{j=1}^{p} \sum_{k=1}^{K} \xi_{ik}^2}^{\frac{1}{4}}$$

No hay un razonamiento teórico claro en la propuesta, pero funciona en la mayor parte de los casos. La propuesta es similar a la que se utiliza en [34].

La representación final sería como aparece en la figura 6.2.

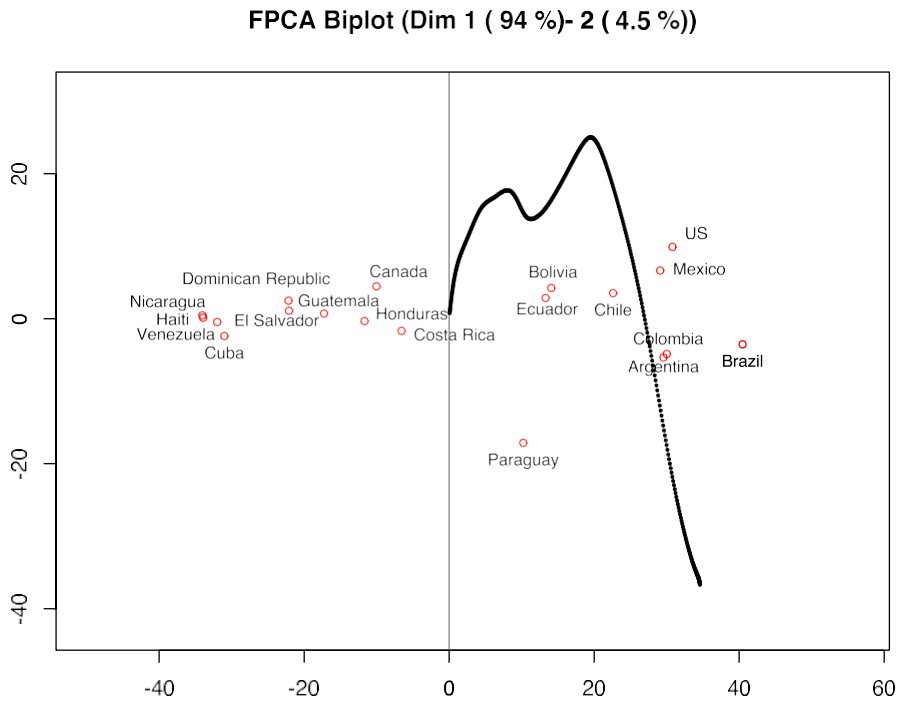


Figura 6.2: Representación Biplot funcional para la evolución de las tasas de mortalidad por COVID en América

El inicio de la trayectoria está en el origen. El final de la trayectoria corresponde con el último tiempo analizado. Observe como ha cambiado la escala del gráfico para poder representar simultáneamente las filas y las columnas. En cualquier caso, las escalas de los ejes no son necesarias para la interpretación del gráfico y se podrían eliminar. Observe también que las escalas del eje X y del eje Y son las mismas, es decir, una unidad en un eje es igual a una unidad en el otro. Esto es así para que la interpretación de la distancia entre puntos sea correcta. La representación sin escalas en los ejes se muestra en la figura 6.3.

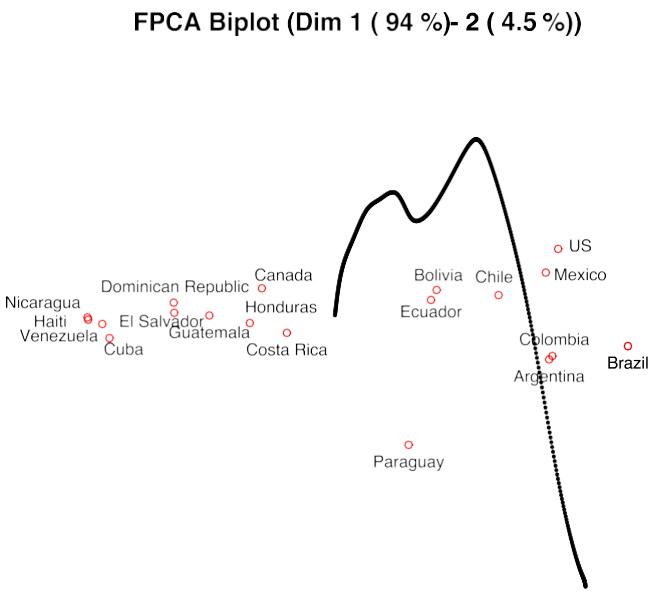


Figura 6.3: Representación Biplot funcional para la evolución de las tasas de mortalidad por COVID en América. Sin las escalas de los ejes

6.3. Algunas propiedades geométricas: Interpretación del biplot

El que hemos definido hasta el momento es el que se conoce como *biplot de predicción* ya que representa, de forma gráfica, una aproximación de los valores en la matriz de datos mediante la proyección de los puntos fila sobre las direcciones que representan a las columnas. Para interpretar un momento determinado bastaría con trazar la línea imaginaria entre el punto de la trayectoria y el origen de coordenadas y proyectar los puntos fila sobre la misma.

Por ejemplo, para el d´ıa 14 de Junio de 2021, que es el día 450 de los analizados, los resultados se muestran en la figura 6.4.

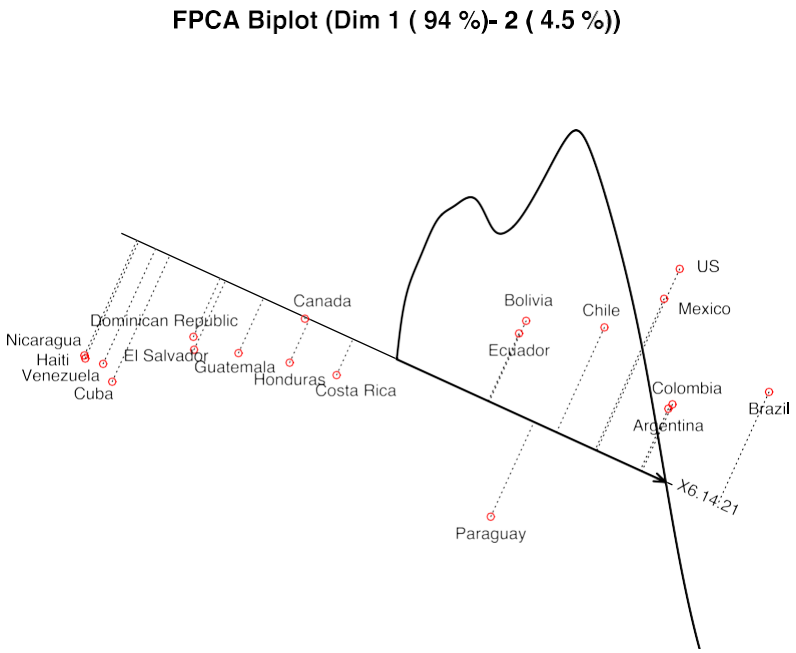


Figura 6.4: Representación Biplot funcional para la evolución de las tasas de mortalidad por COVID en América. Tasas estimadas el 14 de Junio de 2021

Para este día, la tasa más alta corresponde a Brasil, seguido de Colombia y Argentina, con tasas similares. Tras ellos aparecen US y México y así sucesivamente. Los países con tasas más bajas son Nicaragua, Haiti, Venezuela y Cuba.

Podríamos complementar la dirección del biplot con escalas graduadas para tener una idea aproximada de cuales son las tasas estimadas reales. El resultado se muestra en la figura 6.5.

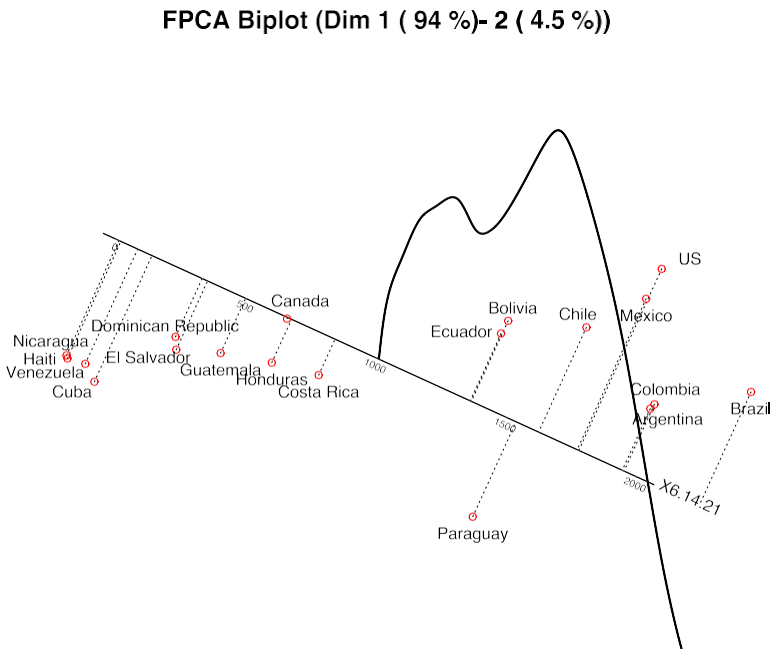


Figura 6.5: Representación Biplot funcional para la evolución de las tasas de mortalidad por COVID en América. Tasas estimadas el 14 de Junio de 2021, con escalas graduadas

Los cálculos para obtener los marcadores de la escala son simples. Para encontrar el marcador en un valor fijado μ buscamos el punto (x, y) que predice el valor μ y que está en la dirección de $\xi(t)$, es decir, en la línea que une los puntos $(0, 0)$ y $\xi(t) = (\xi_1(t), \xi_2(t))$, esto es, $y = \frac{\xi_2(t)}{\xi_1(t)}x$. La predicción verifica también $\mu = \xi_1(t)x + \xi_2(t)y$. Entonces obtenemos que las coordenadas del punto son:

$$x = \mu \frac{\xi_1(t)}{\xi_1(t)^2 + \xi(t)^2}$$

e

$$y = \mu \frac{\xi_2(t)}{\xi_1(t)^2 + \xi_2(t)^2}$$

6.4. Biplot de regresión aproximado para datos funcionales

Supongamos ahora que $\mathbf{A}_f$ son las coordenadas de las filas de la matriz sobre las componentes principales funcionales (CPF) en lugar de las componentes clasicas, podemos aproximar un biplot mediante los resultados anteriores en lugar de utilizar el equivalente a los vectores propios en el caso funcional. Se trataría de una aproximación discreta de un función continua en dos dimensiones. Como para el cálculo de las CPF hemos ajustado una función continua a cada una de las filas de la matriz $\mathbf{X}$, utilizaremos los valores ajustados $\tilde{\mathbf{X}}$ resultantes de los ajustes previos.

Las coordenadas del biplot se calcularán, entonces, como

$$\mathbf{B}_\mathbf{f}^{]} = (\mathbf{A}_\mathbf{f}^{]}\mathbf{A}_\mathbf{f})^{-1}\mathbf{A}_\mathbf{f}^{]}; \tilde{\mathbf{X}} \tag{6.1}$$

Las filas de $\mathbf{B}_f$ se representarán junto con las de $\mathbf{A}_f$ para obtener un biplot. En lugar de representarlas de la forma habitual mediante vectores, haremos una línea o trayectoria uniendo todos los puntos $(\mathbf{b}^{]}_{f1}, \ldots, \mathbf{b}^{]}_{fJ})$, haciendo referencia a que se trata de una línea continua en lugar de un conjunto de puntos.

La forma tradicional de un biplot puede obtenerse uniendo el origen con cualquier punto de la línea resultante, tanto para los observados como para los puntos intermedios obtenidos por interpolación. Incluso podríamos colocar escalas graduadas sobre las direcciones observadas para la estimación de los valores originales. Sobre esta dirección pueden proyectarse los puntos fila para obtener valores aproximados de la variable.

La interpretación general del biplot es la habitual. La distancia entre dos filas la interpretamos en términos de similitud, el ángulo entre dos direcciones como correlación y la proyección de un punto en la dirección como la aproximación del valor original.

Las posiciones de los puntos fila son estáticas y muestran la similitud en los perfiles pero no muestran claramente el carácter funcional de los datos, por ejemplo, si se trata de datos de evolución en el tiempo, no muestran esa evolución.

Proponemos, a continuación, la definición de una trayectoria para cada fila sobre el espacio de las componentes que permite ver la evolución de cada una de las filas

de la matriz. La trayectoria estará formada por las proyecciones del punto fila sobre todas las direcciones correspondientes a los puntos observados, es decir, para una fila i la trayectoria está formada por la proyección del punto

$$\mathbf{a}_f$$

sobre todos las direcciones ($\mathbf{b}'_{f1}, \ldots, \mathbf{b}'_{fJ}$). Obsérvese que esta trayectoria podría definirse también sobre cualquier otro biplot, incluido el clásico, siempre que tenga sentido su interpretación.

6.5. Calidad de representación y predictividad

La bondad del ajuste global es la cantidad de variabilidad explicada por la predicción de la matriz completa, es decir,

$$\rho^2 = tr(\hat{\mathbf{X}}'\hat{\mathbf{X}})/tr(\mathbf{X}'\mathbf{X}) \tag{6.2}$$

Incluso en los casos para los que se obtiene un buen ajuste global, puede que algunas de las filas o de las columnas de la matriz original no estén bien ajustadas.

La bondad del ajuste para cada columna es

$$\rho_j^2 = diag(\hat{\mathbf{X}}'\hat{\mathbf{X}}) \div diag(\mathbf{X}'\mathbf{X}) \tag{6.3}$$

donde $\div$ significa la operación elemento a elemento. ρ_j^2 is like the R-Squared of the regression of each column of $\mathbf{X}$ on $\mathbf{A}$. We call that quantity *quality of representation* of the variable in analogy with the terminology of Correspondence Analysis ([**?**], [**?**]). The term *predictiveness* of the column is used in ([**?**]). The measures are used to identify which variables are most related to the representation. The goodness of fit for each row is

$$\rho_i^2 = diag(\hat{\mathbf{X}}\hat{\mathbf{X}}') \div diag(\mathbf{X}\mathbf{X}') \tag{6.4}$$

This measures are also called *quality of the representation* or *predictiveness*. The measures separated for each dimension are also called *Contributions of the Factor to the Element (row or column)* or *Squared Cosines*. The measures are used to identify

which dimensions are useful to differentiate the individual from the rest. Individuals with low representation qualities are usually placed around the origin.

Capítulo 7

Aplicaciones

7.1. Causas de muerte en el Ecuador

Trabajamos ahora con el número de muertes clasificadas por año y causas de muerte.

Hemos buscado también la población en cada uno de los años para calcular las tasas de cada una de las causas de muerte por cada 100000 habitantes. Disponemos desde los años 1997 al 2020, los datos del último año son provisionales. Después de observar la forma, hemos decidido eliminar este último año ya que parece presentar datos extraños con subidas y bajadas en algunas de las cusas que no parecen seguir la tendencia general.

En primer lugar vamos a dibujar las tasas para cada una de las causas, después de convertirlas en un objeto de datos funcionales.

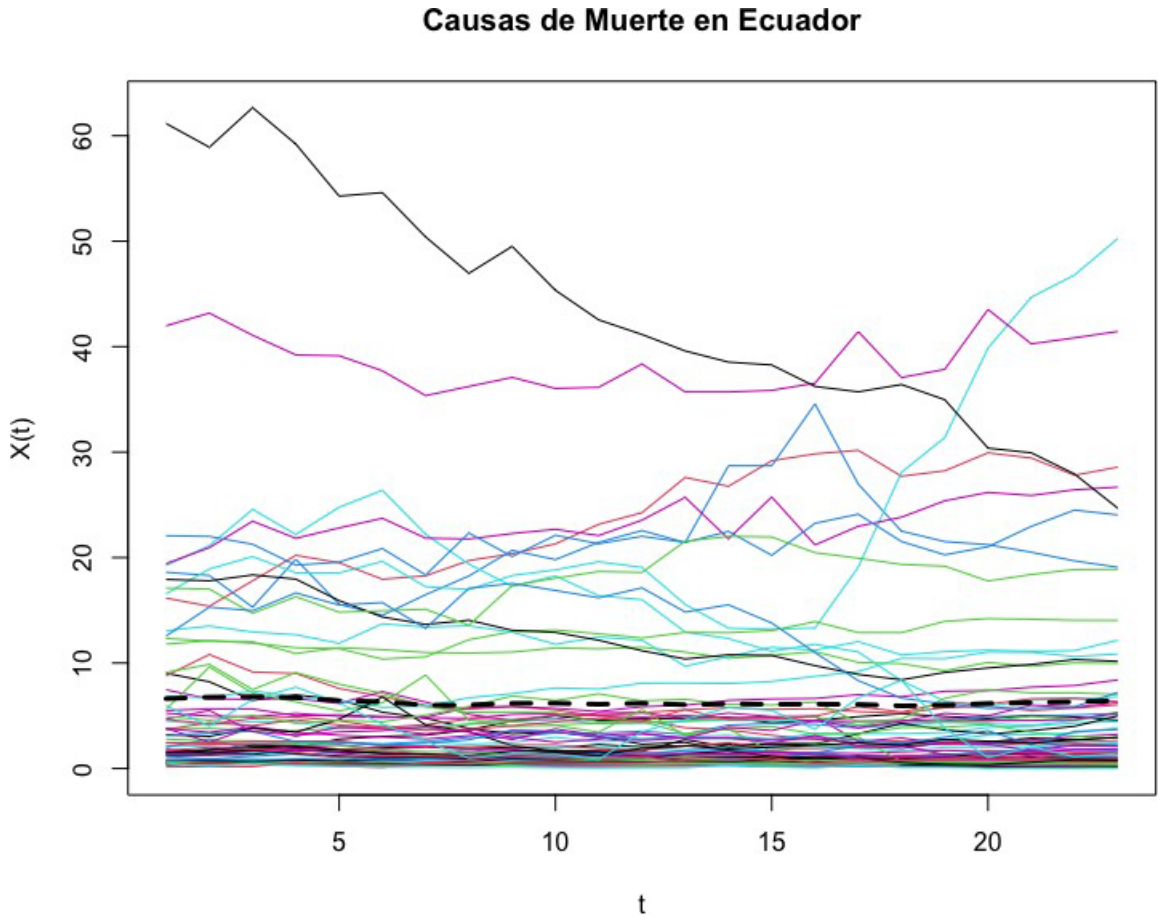


Figura 7.1: Causas de Muerte en el Ecuador: Datos observados

Observamos que la curvas son, de alguna manera, irregulares. Las dos que tienen los valores más altos son, precisamente las causas mal definidas (azul) y el resto de las causas (magenta). La primera va disminuyendo con el tiempo, es decir, parece que se ha mejorado en la definición de las causas, mientras que la segunda parece más o menos constante a lo largo del tiempo. La línea azul que corresponde e las muertes por diabetes parece que se ha incrementado bastante en los últimos años.

La línea punteada muestra la tasa media que permanece más o menos constante a lo largo del tiempo.

El gráfico es difícil de visualizar y no se ve claro la similitud y la diferencia entre

las distintas causas por lo que haremos un Análisis de Componentes Principales para Datos Funcionales (ACPDF) que nos ayude a resumir la información.

Antes de proceder al ACPDF tenemos que suavizar las curvas. Utilizaremos una base basada en splines. Las curvas suavizadas se muestran en la figura siguiente.

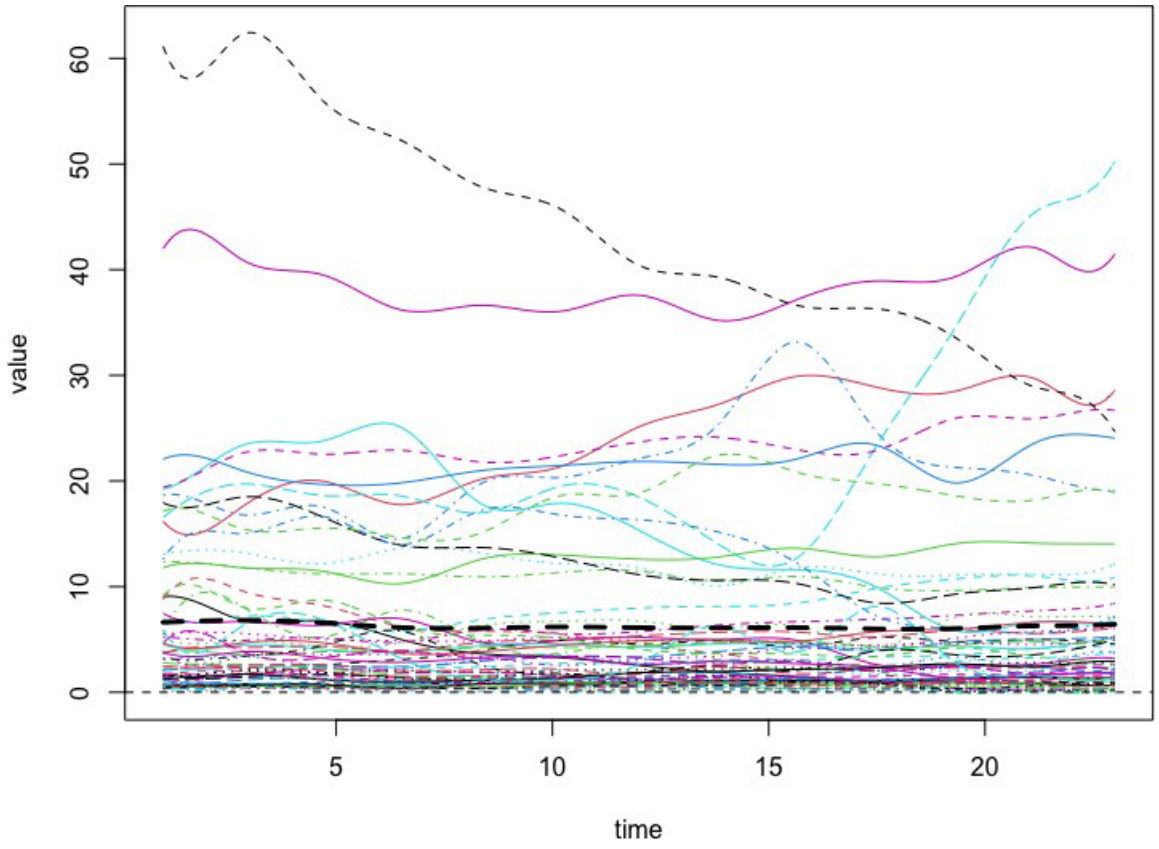


Figura 7.2: Causas de Muerte en el Ecuador: Datos suavizados

La tendencia de las curvas es similar a las originales pero ahora están suavizadas.

Calculamos las covarianzas y las representamos en función del tiempo.

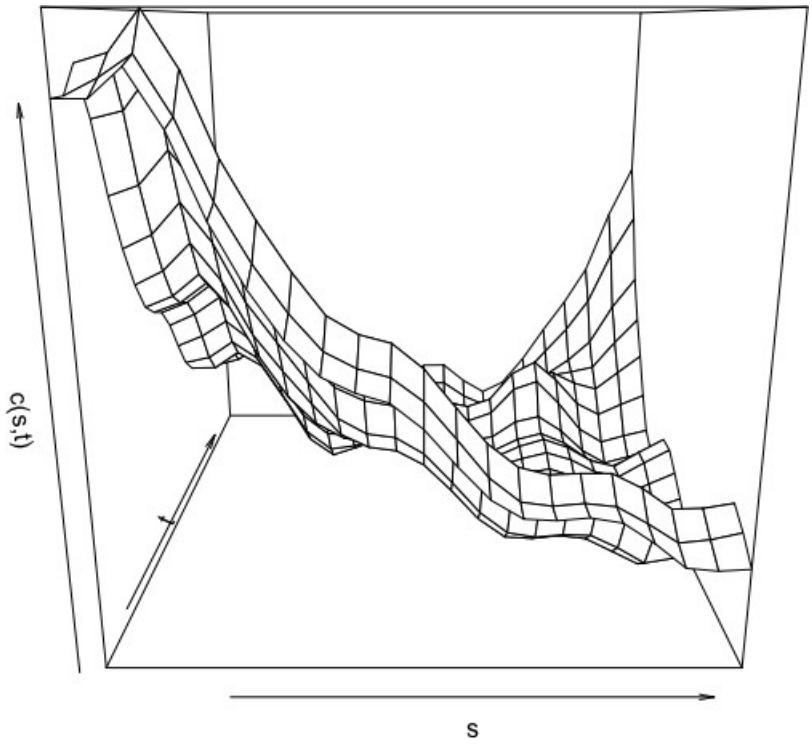


Figura 7.3: Causas de Muerte en el Ecuador: Covarianzas

También con las correspondientes curvas de nivel

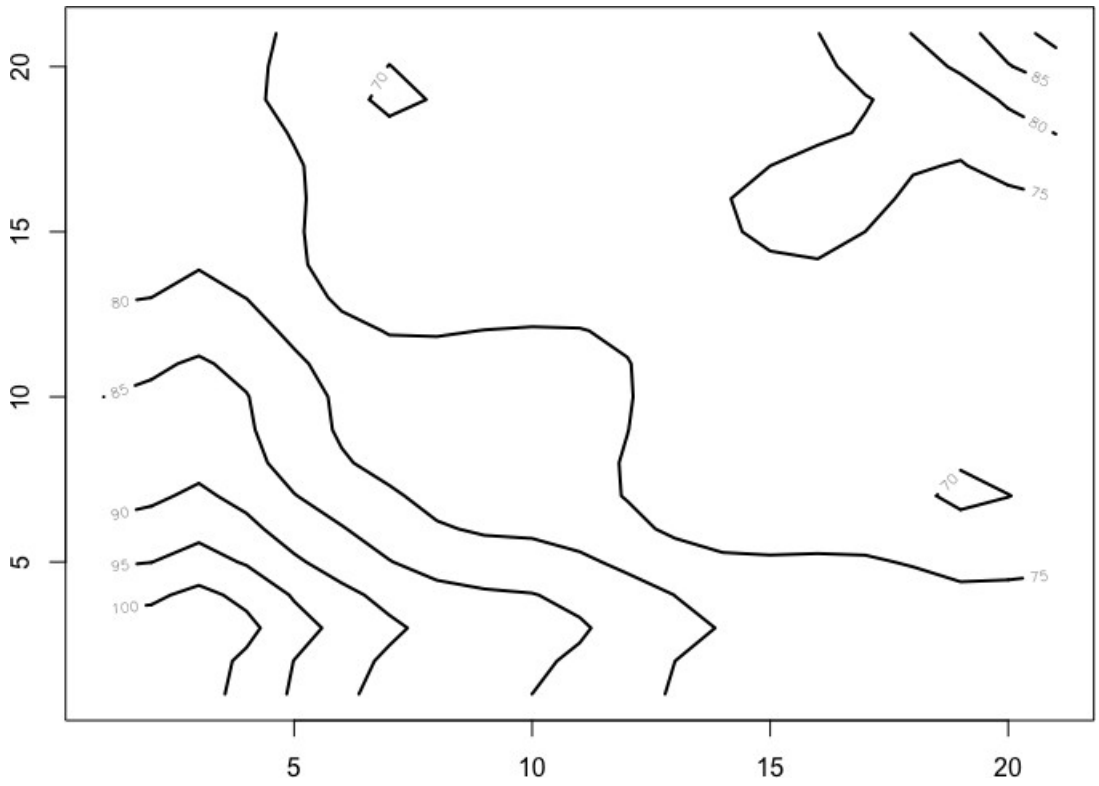


Figura 7.4: Causas de Muerte en el Ecuador: Curvas de nivel de la función de covarianzas

Las covarianzas suelen ser mayores entre momentos cercanos del tiempo, si bien van disminuyendo y son algo más pequeñas en los últimos años lo que quiere decir que se ha modificado probablemente la composición de las causas de muerte.

Estamos ya en disposición de realizar las componentes principales. La tabla siguiente muestra la proporción de la variabilidad recogida por cada componente.

	Proporción de varianza	Acumulada
Componente 1	93.07	93.07

	Proporción de varianza	Acumulada
Componente 2	5.08	98.16

Entre las dos primeras componentes recogen el 98.16 % de la variabilidad por lo que son suficientes para representar adecuadamente los datos.

Las componentes principales tienen que ser entendidas como funciones, en este caso dependientes del tiempo. Se representan en la figura siguiente.

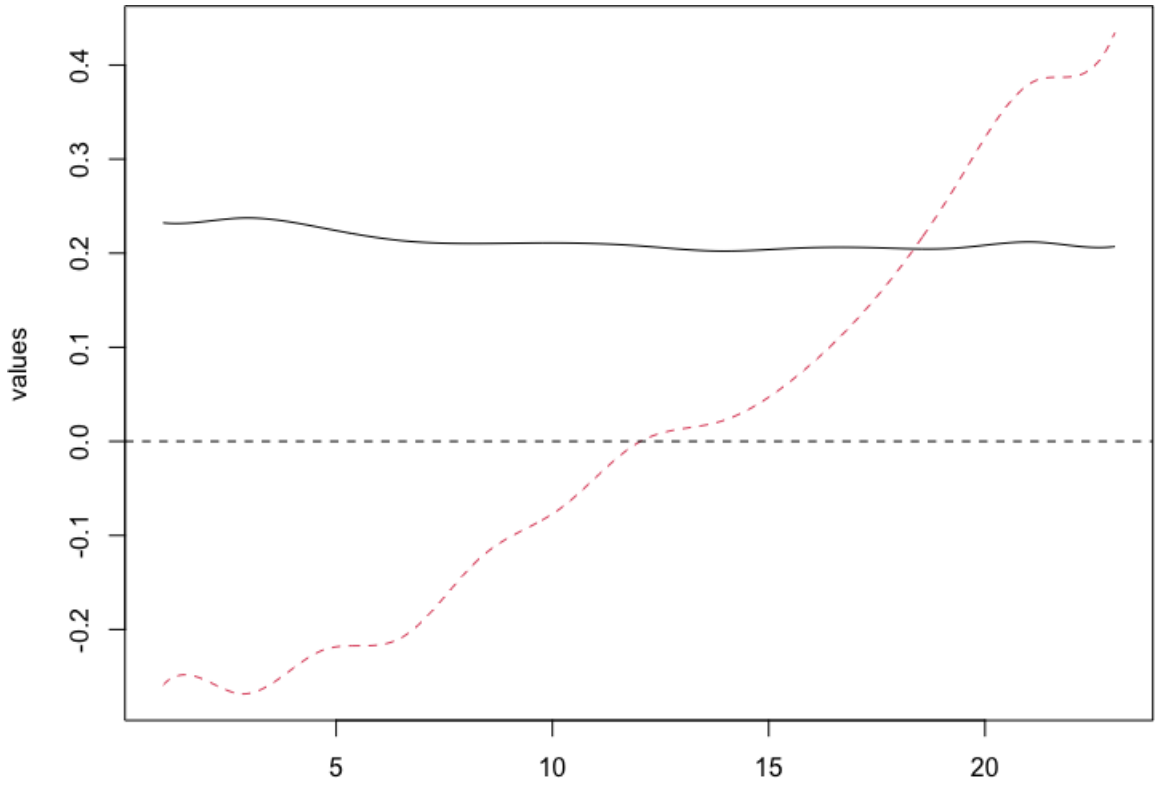


Figura 7.5: Causas de Muerte en el Ecuador: Funciones que definen las dos primeras componentes principales

Estas funciones son el equivalente a los vectores propios en los métodos clásicos. La primera componente, en negro en el gráfico, es una componente de tamaño igual que en la aplicación de los métodos clásicos. Muestra, entonces, la diferencia de tamaño entre las distintas causas de muerte y recoge la parte constante de cada una de ellas a lo largo del tiempo. De esta forma, las causas (que veremos en una representación posterior) se ordenaran sobre la primera componente de acuerdo con un gradiente de maño, las más frecuentes a un lado y las menos frecuentes en el lado opuesto.

La segunda componente tiene que ver con la forma de las curvas, es decir, con la variación, en relación a la parte constante, a lo largo del tiempo. De un lado se situarán las causas que han aumentado con el tiempo y de otro las que han disminuido. Las causas cuyas tasas han permanecido constantes a lo largo del tiempo tomarán valores alrededor del 0 para la segunda componente.

Debido a la alta absorción de varianza de la primera componente, parece que la mayor parte de las tasas, en general, muestran poca variabilidad a lo largo del tiempo.

Aunque no es estándar en la literatura, vamos a hacer a hacer un gráfico bidimensional evaluando loas armónicos en los momentos de tiempo observados que nos dará una idea adicional de como se relaciona el tiempo con las componentes.

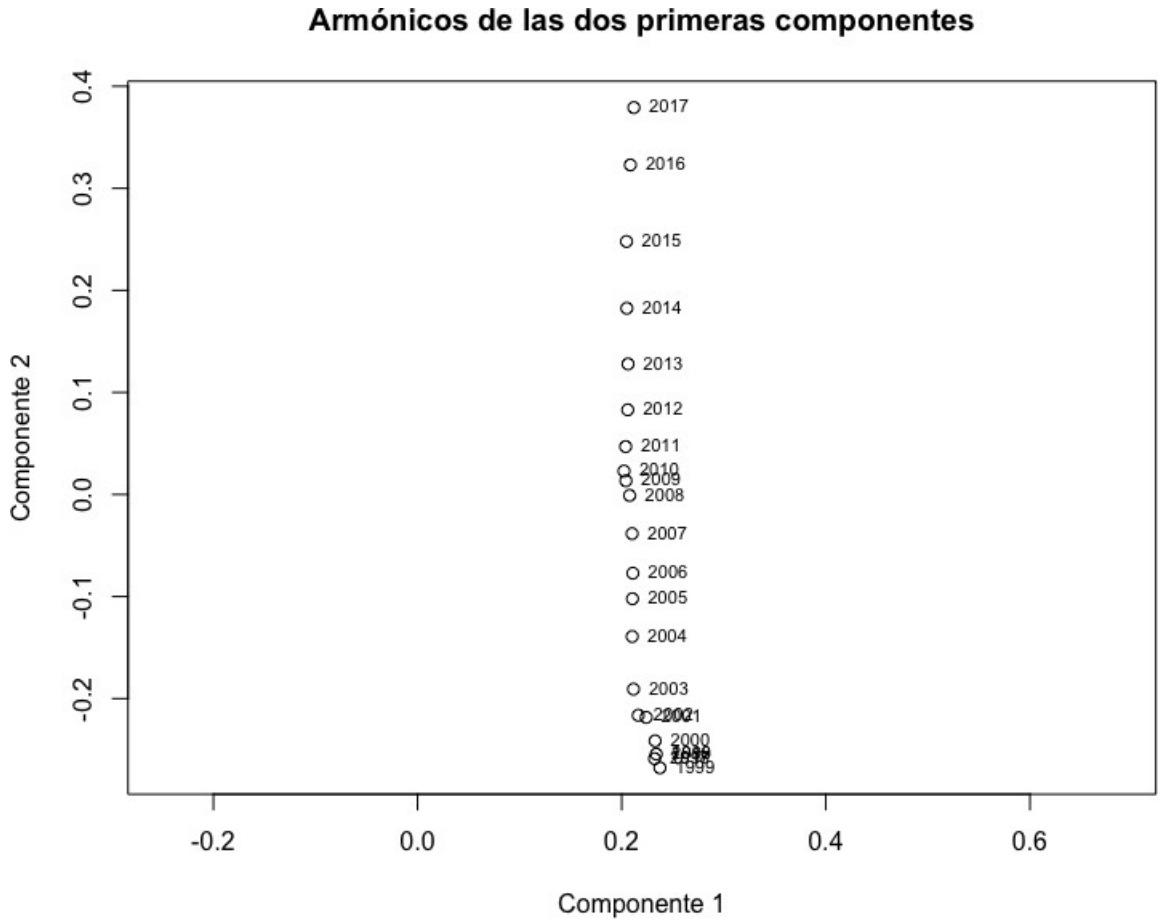


Figura 7.6: Causas de Muerte en el Ecuador: Representación las dos primeras componentes principales

La contribución de todos los tiempos a la primera componente principal es muy similar para todos ellos, lo que es otro indicador de que se trata de una componente de tamaño. Hacia la derecha, en la parte positiva de la componente, se colocaran las causas mayoritarias y en la parte negativa las minoritarias.

La segunda componente muestra un gradiente temporal y, por tanto, la variación de las causas de muerte a lo largo del tiempo. Las causas con mayor puntuación en esta componte son las que han aumentado en el transcurso de los años, las que tengan valores negativos habrán disminuido.

Finalmente representamos las puntuaciones de cada una de las causas sobre las componentes.

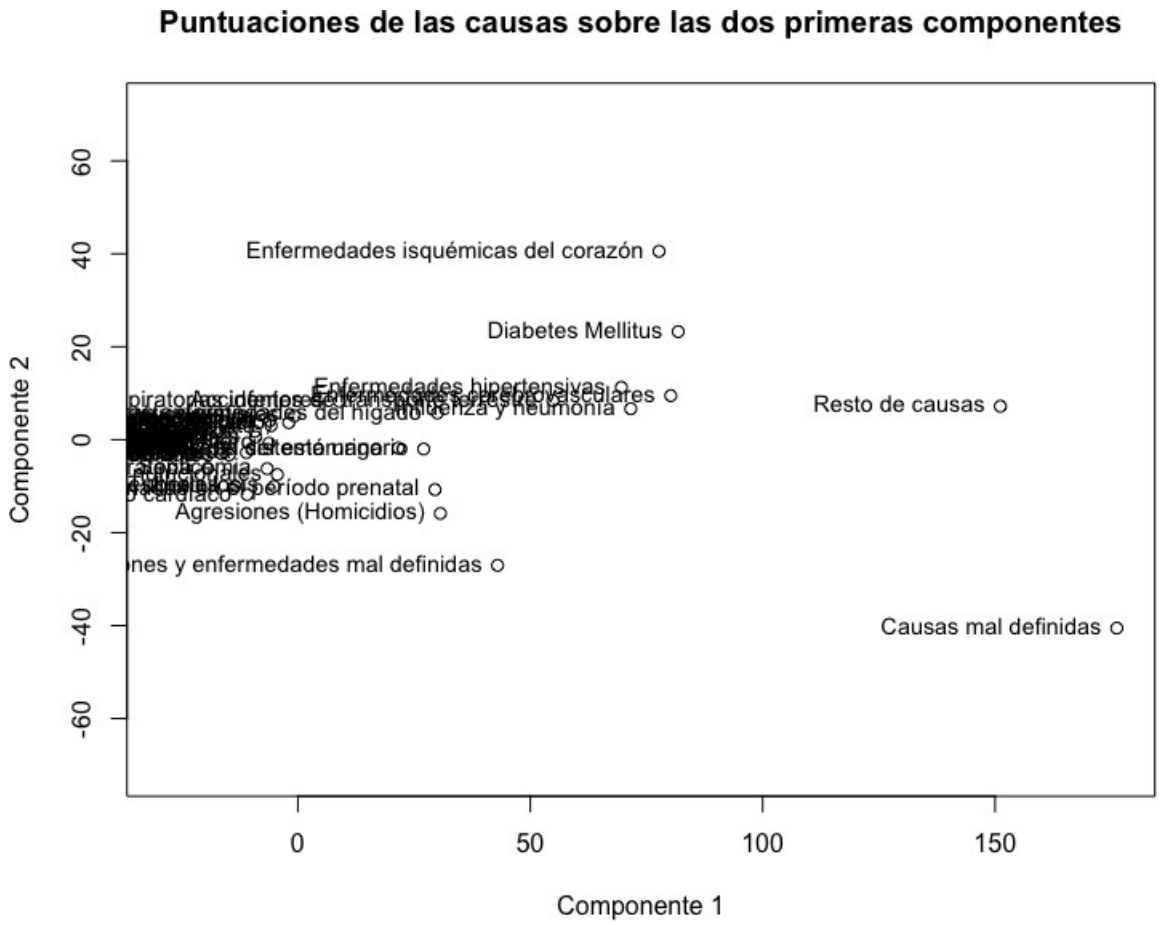


Figura 7.7: Causas de Muerte en el Ecuador: Representación las puntuaciones de las causas de muerte sobre las dos primeras componentes principales

Hay una concentración alta de puntos en la izquierda del gráfico que no permite ver las etiquetas. Vamos a ampliarlo para separarlas aunque perdamos la escala adecuada.

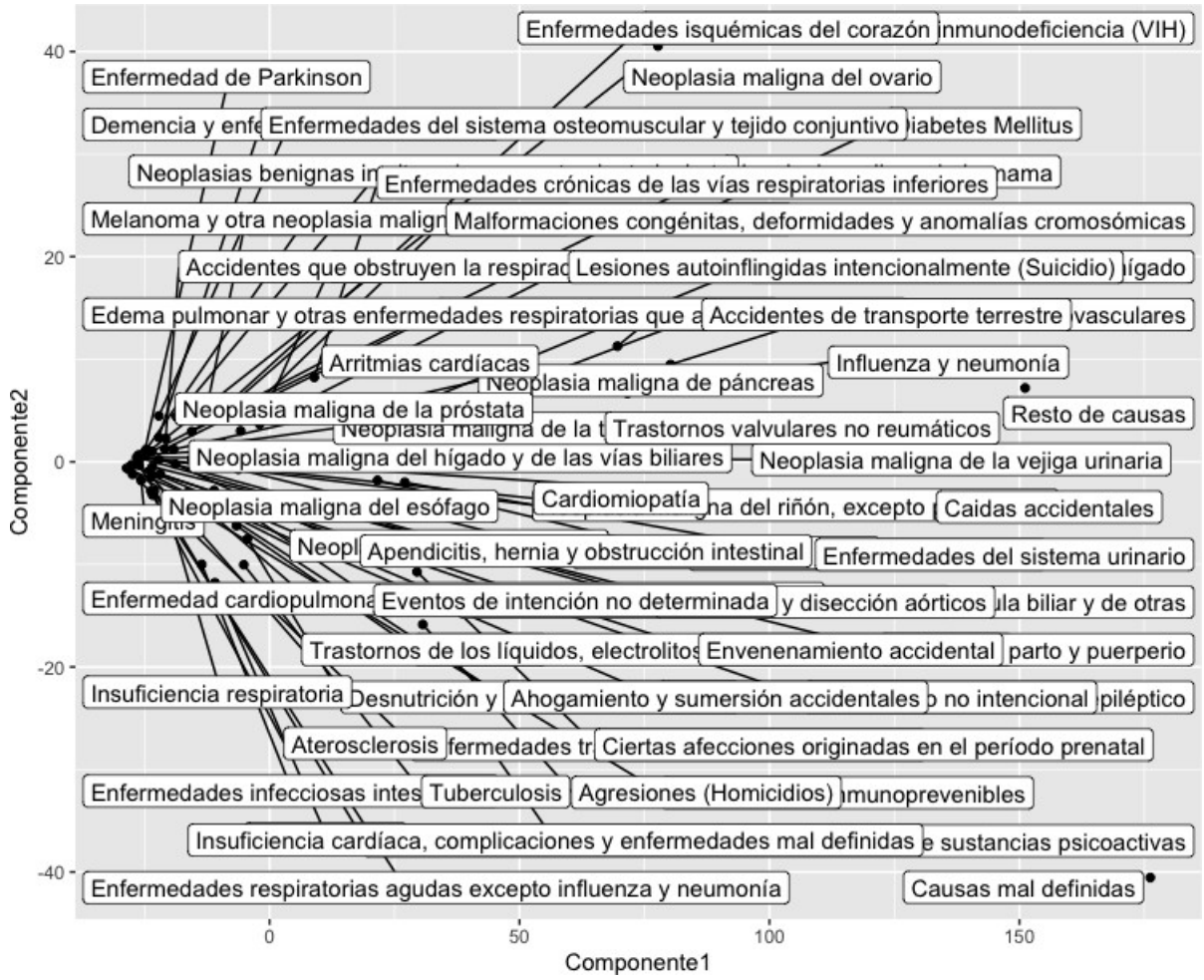


Figura 7.8: Causas de Muerte en el Ecuador: Representación las puntuaciones de las causas de muerte sobre las dos primeras componentes principales (ggplot2)

Como ya habíamos intuido del gráfico inicial con las curvas, la categoría dominante son las causas mal definidas que tiene la puntuación más alta sobre la primera componente. Tiene también la puntuación negativa en la segunda por lo que ha ido disminuyendo con el paso del tiempo, es decir, la administración ecuatoriana ha mejorado su eficiencia en relación a la definición y clasificación correcta de las causas de muerte.

La segunda tasa más alta es el resto de las causas que se supone que incluye muchas otras causas minoritarias agregadas. Esta categoría ha aumentado un poco

a lo largo de los años.

A continuación tenemos un grupo que es también frecuente y que ha aumentado con el tiempo. Este grupo incluye Diabetes, Enfermedades isquémicas del corazón, enfermedades relacionadas con la hipertensión, accidentes de transporte, accidentes cerebrovasculares e influenza y neumonía. Todas ellas tienen perfiles de evolución similares. Cabe destacar un aumento importante en las enfermedades isquémicas del corazón y en la diabetes.

La insuficiencia cardíaca también es frecuente aunque ha disminuido con el tiempo.

A continuación aparece un grupo de enfermedades menos frecuentes que incluye la cirrosis hepática, neoplasias estomacales, enfermedades del sistema urinario, homicidios y "ciertas afecciones??". De ellas, la cirrosis parece haber aumentado algo en el tiempo mientras que los homicidios han disminuido.

Finalmente ajustamos el biplot aproximado como describíamos antes.

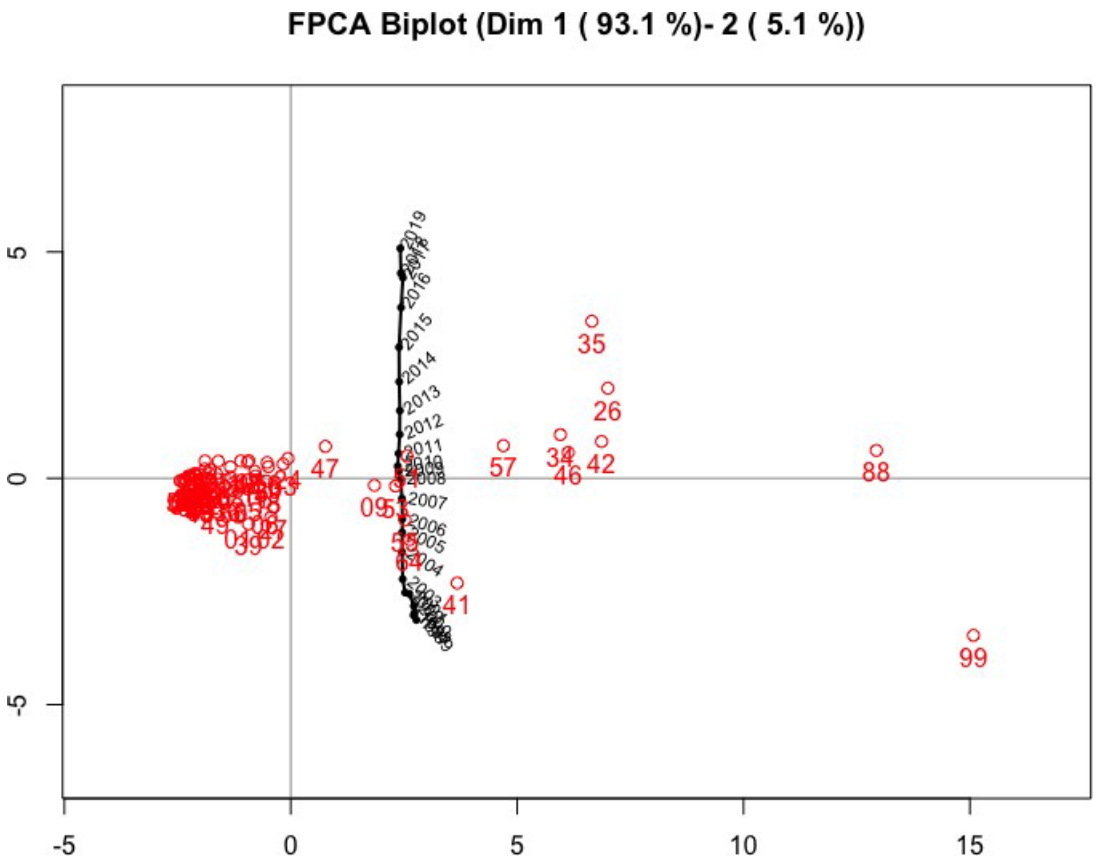


Figura 7.9: Causas de Muerte en el Ecuador: Representación Biplot funcional

Hemos proyectado las causas sobre la dirección del último año de la representación (2019) de forma que podemos observar cuales son las causas de muerte mayoritarias al final del periodo.

A parte de las englobadas como "Resto de las causas", las más frecuentes son las enfermedades isquémicas seguidas de la diabetes. Después las enfermedades relacionadas con la hipertensión, accidentes cerebrovasculares, influenza y neumonía y accidentes de tráfico.

En el gráfico siguiente hemos añadido también la proyección sobre el primer año en estudio. De esta forma podemos ver como han variadao las causas de muerte a lo

largo del periodo estudiado

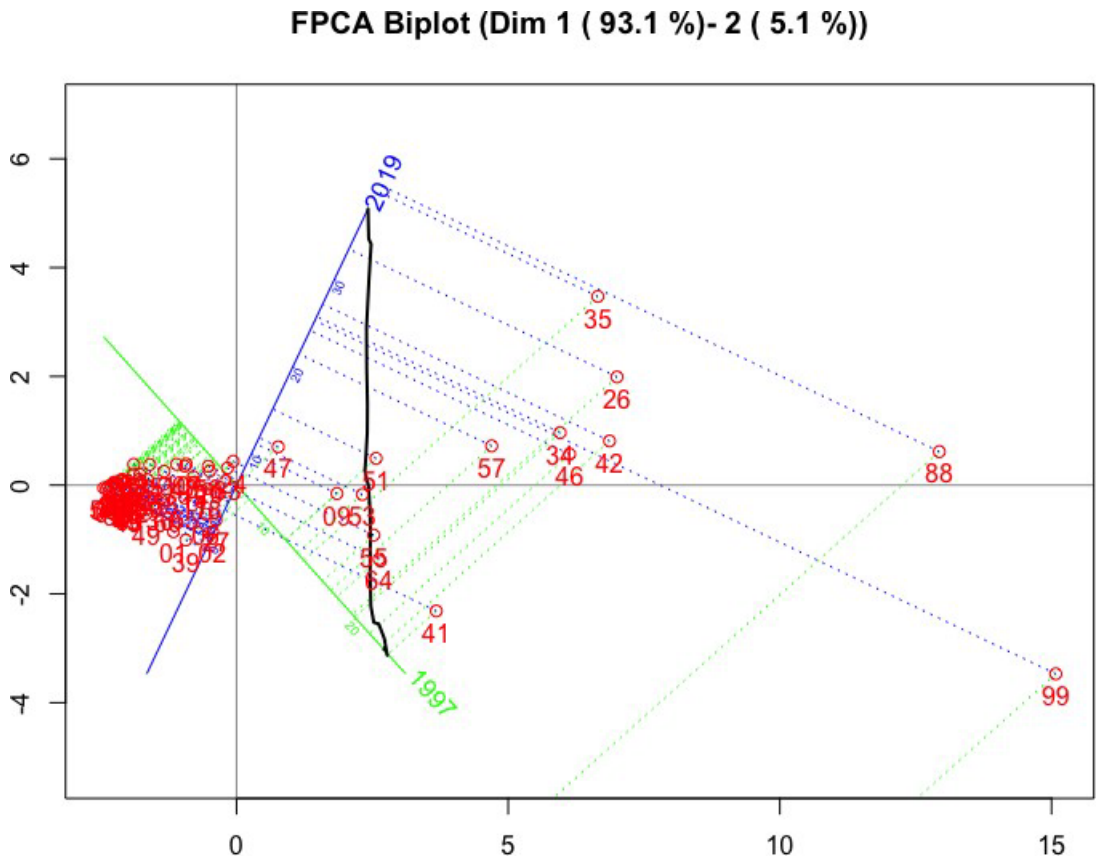


Figura 7.10: Causas de Muerte en el Ecuador: Representación Biplot funcional con proyecciones sobre el primer año en estudio

La causa 99 (Causas mal definidas) era la más frecuente al principio mientras que ya no lo es al final. Como vimos antes, las causas que quedan por encima del primer eje (tienen coordenadas positivas en el segundo) son aquellas que han aumentado con el paso del tiempo mientras que las que quedan por debajo son las que han disminuido, tal y como se muestra en los armónicos (vectores propios).

Las calidades de representación de las variables son todas muy altas como se muestra en la tabla siguiente.

Cuadro 7.2: Calidad de representación de las variables

	Dim 1	Dim 2
1997	91.954	98.212
1998	92.441	98.433
1999	92.332	98.761
2000	93.618	99.128
2001	94.399	99.304
2002	93.587	98.704
2003	94.746	98.956
2004	97.239	99.574
2005	98.068	99.338
2006	98.608	99.333
2007	99.135	99.318
2008	98.879	98.879
2009	97.778	97.800
2010	95.670	95.735
2011	93.146	93.411
2012	93.119	93.938
2013	95.261	97.256
2014	94.879	98.973
2015	92.097	99.453
2016	87.601	99.055
2017	83.674	98.280
2018	81.949	97.465
2019	77.792	96.493

Colocamos las variables solas en el gráfico para observar el comportamiento de las escalas de predicción.

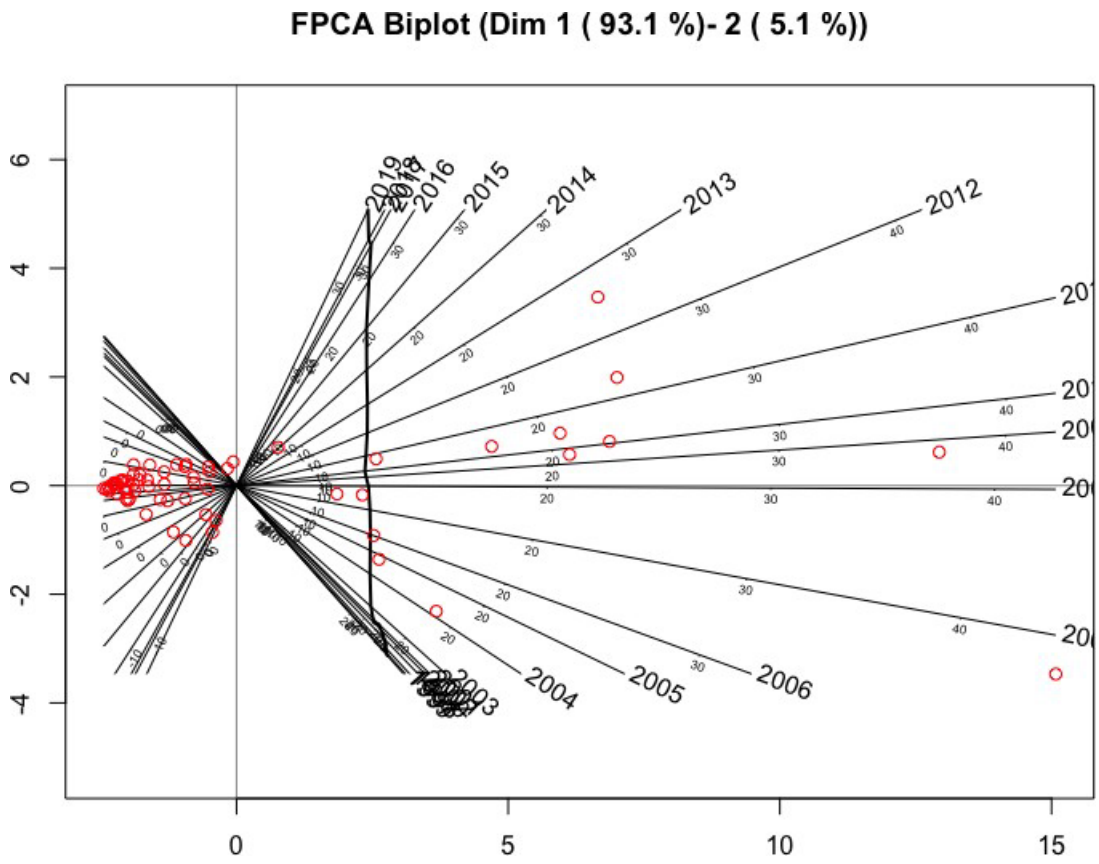


Figura 7.11: Causas de Muerte en el Ecuador: Representación de las variables

Obsérvese que los valores de las predicciones en las escalas de las variables forman aproximadamente un círculo, lo que quiere decir que las escalas en casa una de las variables son comparables y van a servir para mostrar la evolución de las causas como veremos después.

Para ver mejor las causas de muerte, podemos visualizar una zona parcial del gráfico cambiando los Límites de la representación. Se han añadido las proyecciones sobre el primer y último año.

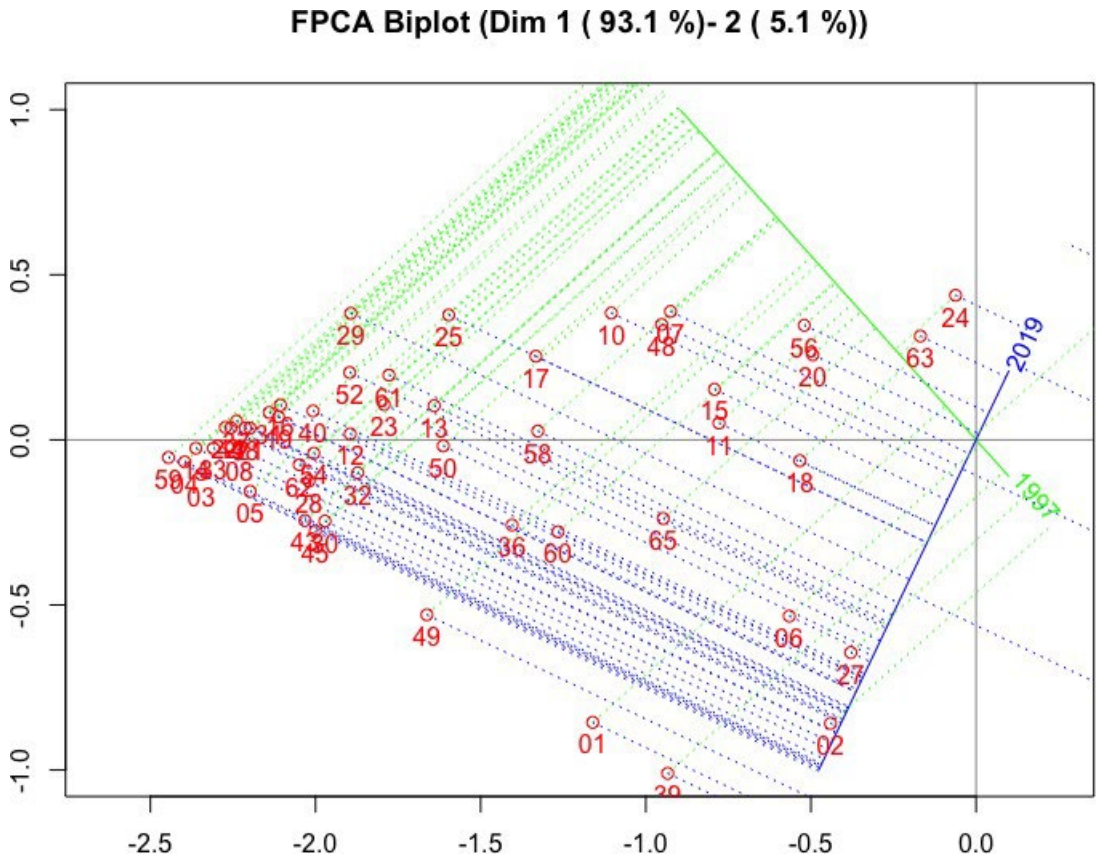


Figura 7.12: Causas de Muerte en el Ecuador: Representación parcial del biplot

Podemos ver la evolución de cada una de las causas mostrando las trayectorias que marcan las proyecciones de cada una de ellas sobre todos los años.

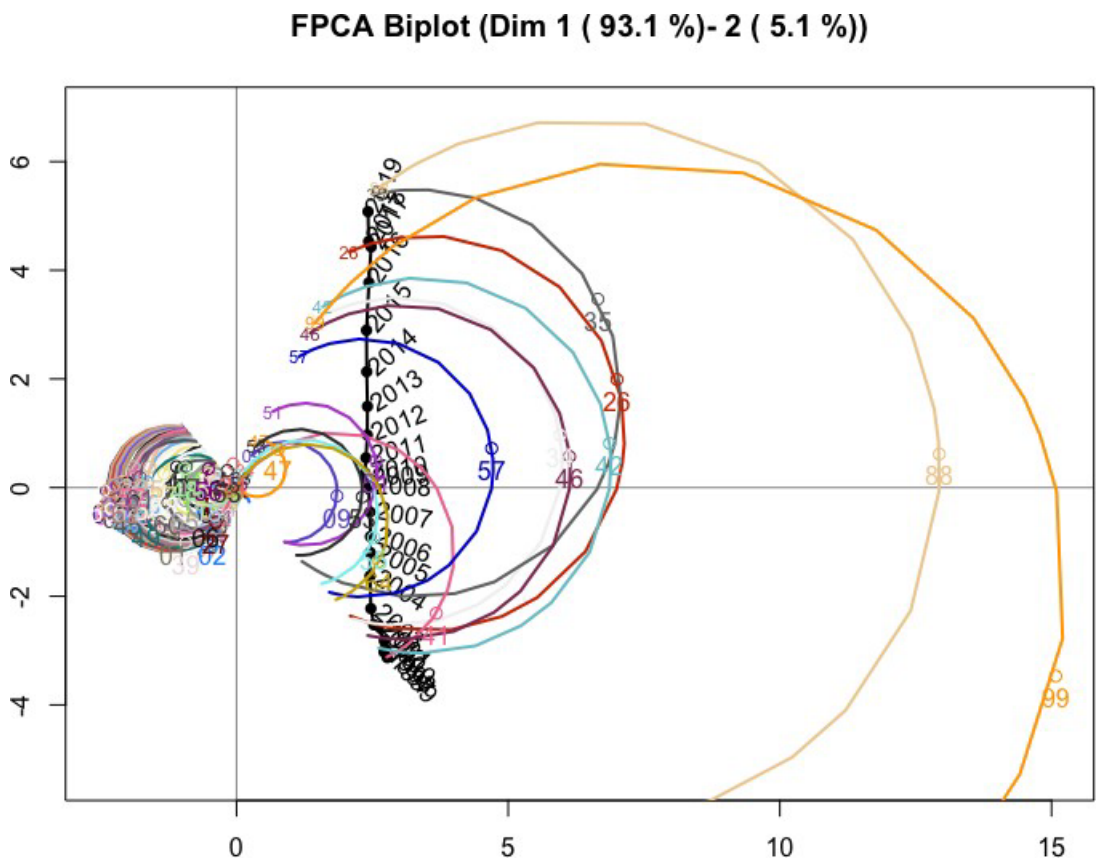


Figura 7.13: Causas de Muerte en el Ecuador: Trayectorias de las causas

Mostramos también el detalle de las causas que aparecen juntas.

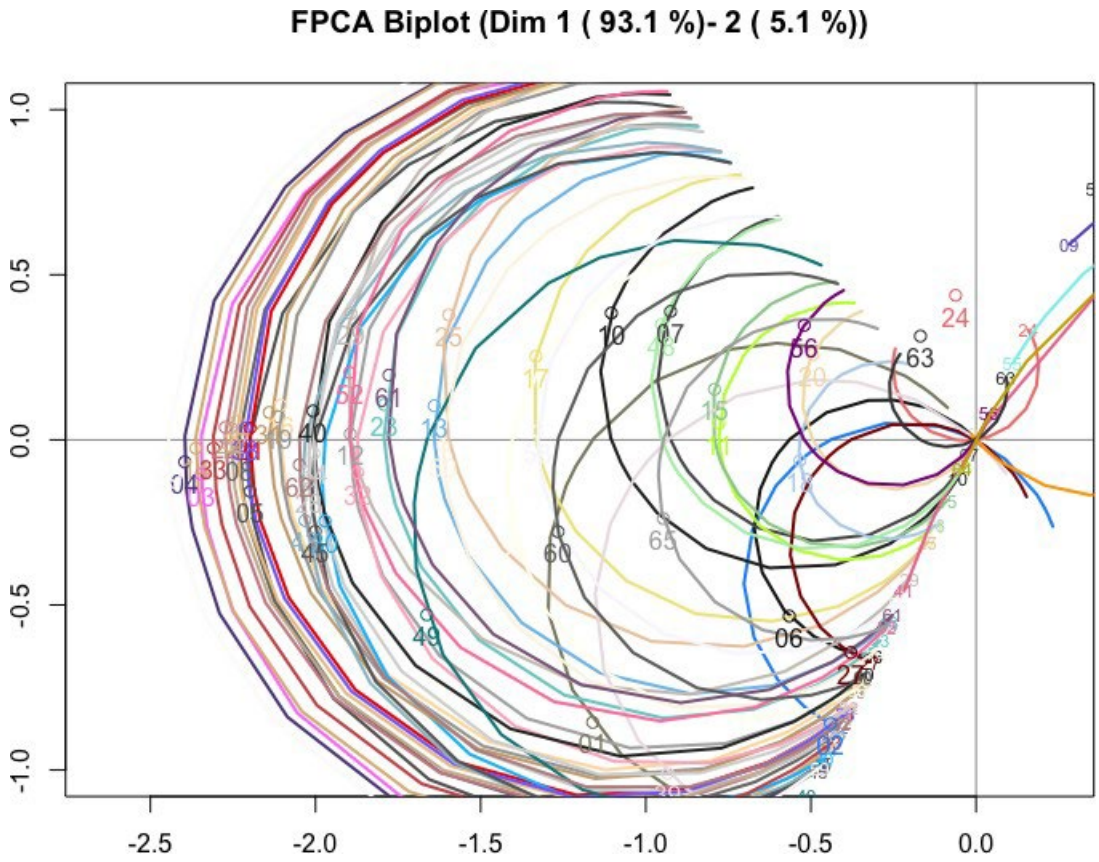


Figura 7.14: Causas de Muerte en el Ecuador: Trayectorias de las causas (vista parcial)

Todas las trayectorias son prácticamente circulares debido al tipo de datos que estamos trabajando. Si lo hacemos para un biplot general, no tendrían porqué ser así .

Dos trayectorias serán casi paralelas cuando la diferencia entre las tasas de las dos causas se han mantenido constantes a lo largo del tiempo. Si se cruzan, significa que se han intercambiado los órdenes. Por ejemplo, para las trayectorias de las causas 99 (mal definidas) y 88 (Otras causas) las diferencias eran a faor de la primera al inicio del periodo mientras que son a favor de la segunda en el final. El cruce se produce

alrededor del a˜no 2013.

las causas 42 y 46 son aproximadamente paralelas y casi circulares por lo que se han mantenido constantes en el tiempo.

7.2. Evolución de las tasas de mortalidad por COVID en los países de América.

7.2.1. Descripción y suavizado de los datos

Los datos originales se muestran en la figura 7.15.

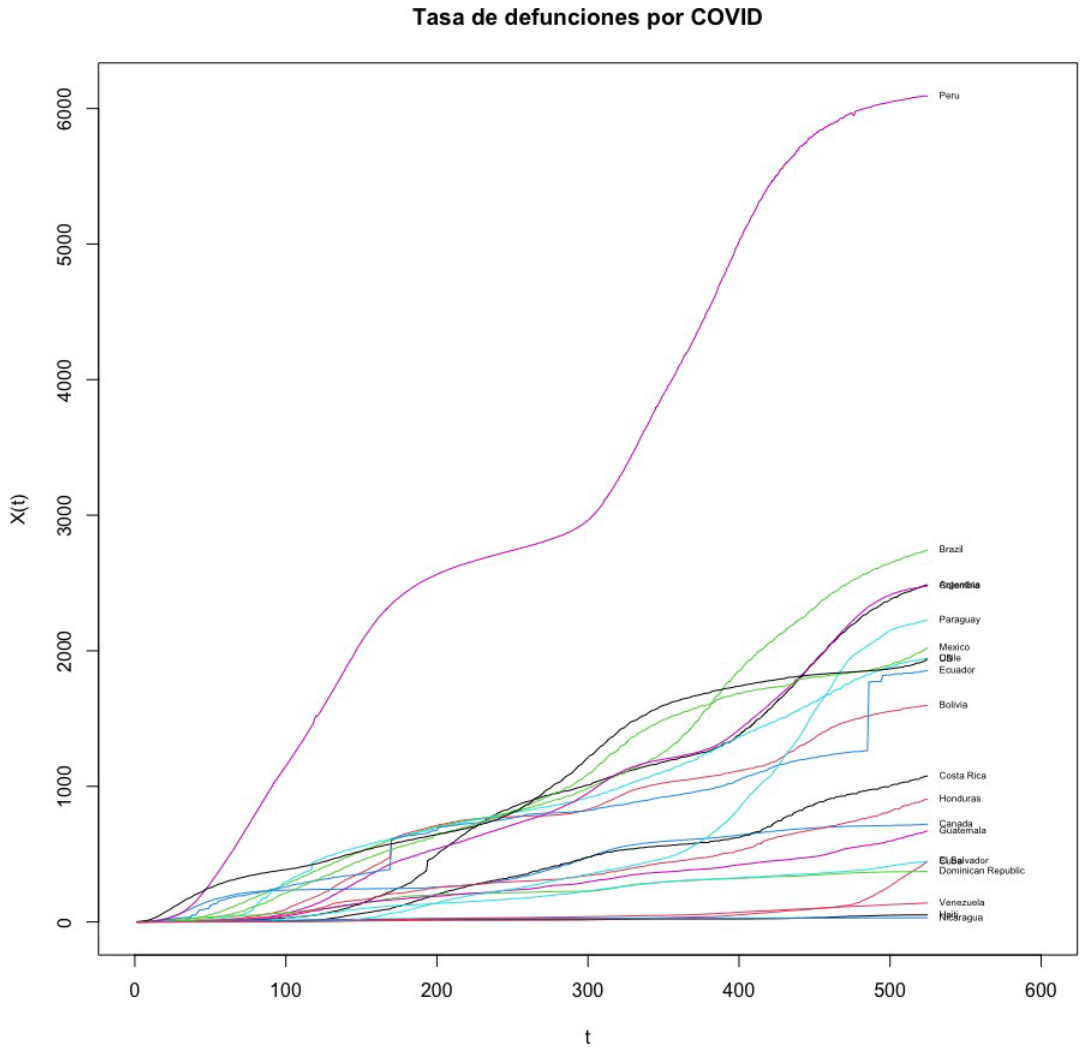


Figura 7.15: Causas de Muerte en el Ecuador: Datos observados

El país que aparece por encima de los demás es Perú que parece haber tenido

unas tasas de mortalidad muy superiores al resto durante todo el periodo de la pandemia. En Ecuador hay un comportamiento extraño en dos momentos en los que se produce un incremento brusco en las tasas. Esto se debe, probablemente a que los datos se actualizaron en esos dos momentos contabilizando los datos que no se habían registrado hasta entonces. En los datos regionales, que veremos más tarde, se produce un fenómeno similar.

Como en el caso anterior, construimos una base y suavizamos las curvas de crecimiento de las tasas. La figura 7.16 muestra las curvas suavizadas.

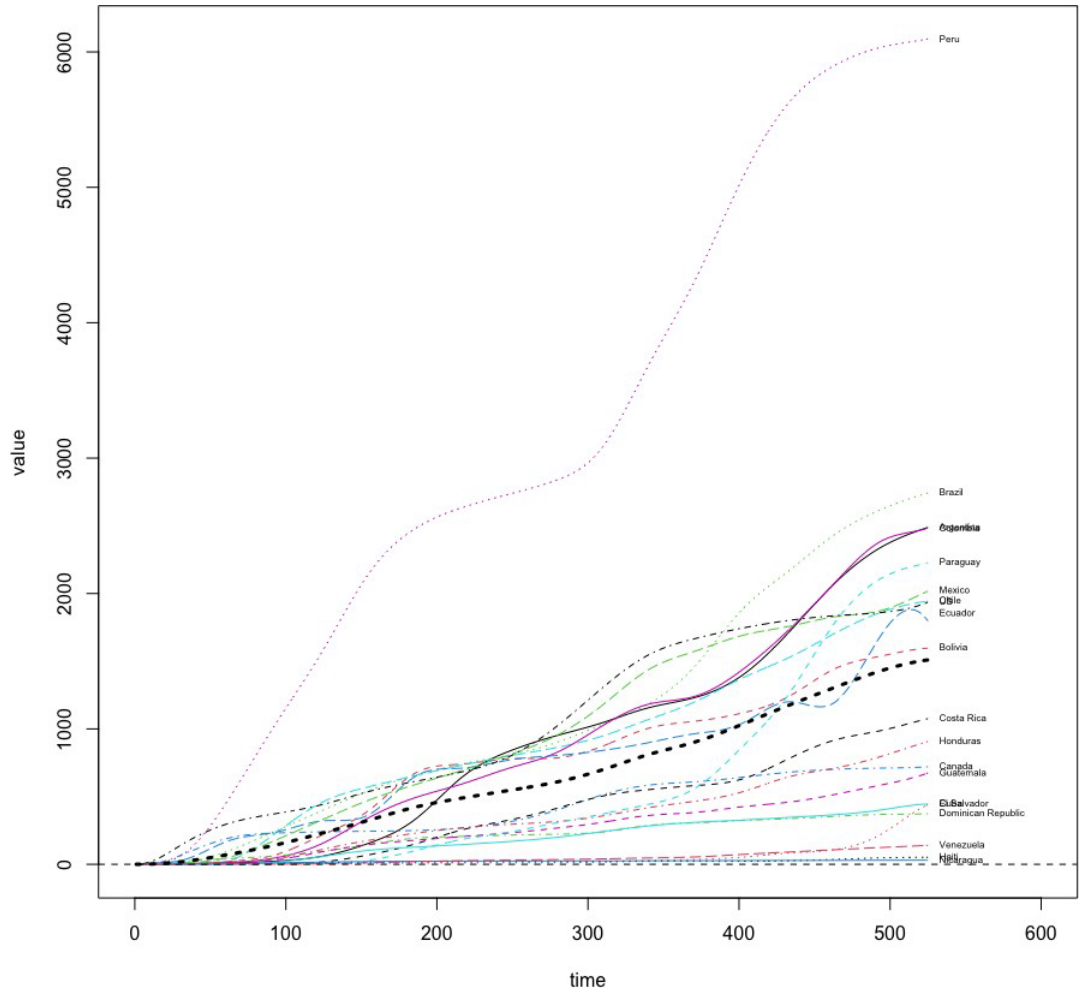


Figura 7.16: Causas de Muerte en el Ecuador: Datos suavizados mediante B-Splines

En todos los casos se muestra un aumento de las tasas con el tiempo si bien el aumento es mucho más acusado en unos países que en otros. La linea gruesa punteada muestra la media de todos que también crece en el tiempo.

Exploramos las covarianzas y representamos la función de covarianza en función del tiempo en la figura 7.17.

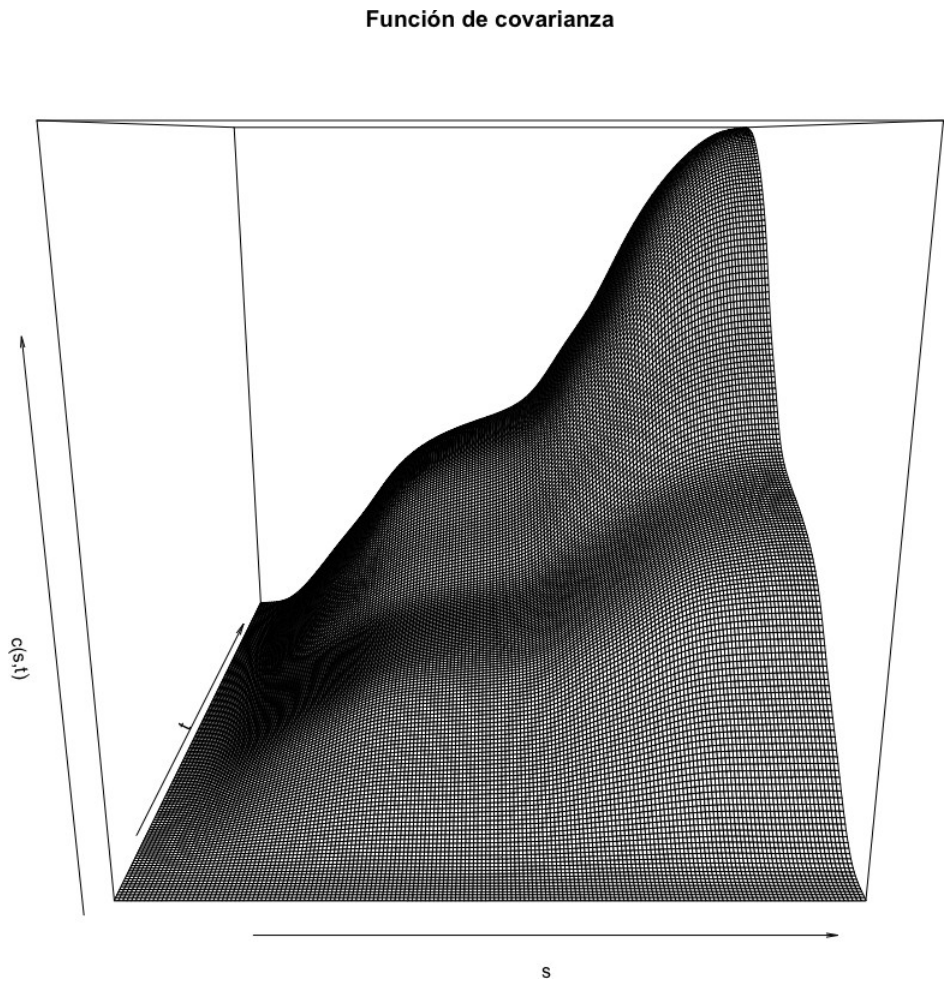


Figura 7.17: Causas de Muerte en el Ecuador: Función de covarianza

Como cabe esperar, las varianzas y las covarianzas van aumentando con el tiempo ya que se trata de curvas crecientes. La misma información se muestra la figura 7.18 que contiene las curvas de nivel de la función de covarianza.

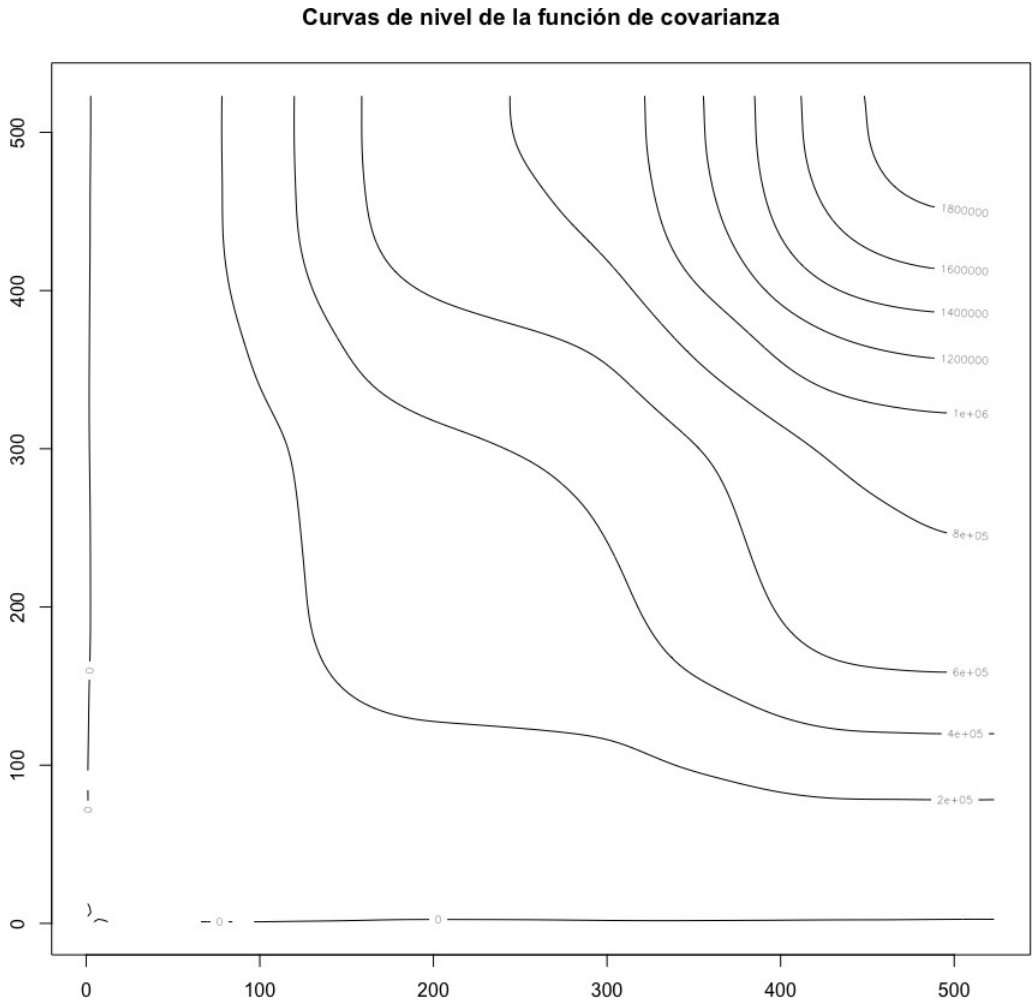


Figura 7.18: Causas de Muerte en el Ecuador: Curvas de nivel de la función de covarianza

Probablemente serán más informativas las correlaciones. La figura 7.19 muestra las correlaciones para cada pareja de tiempos.

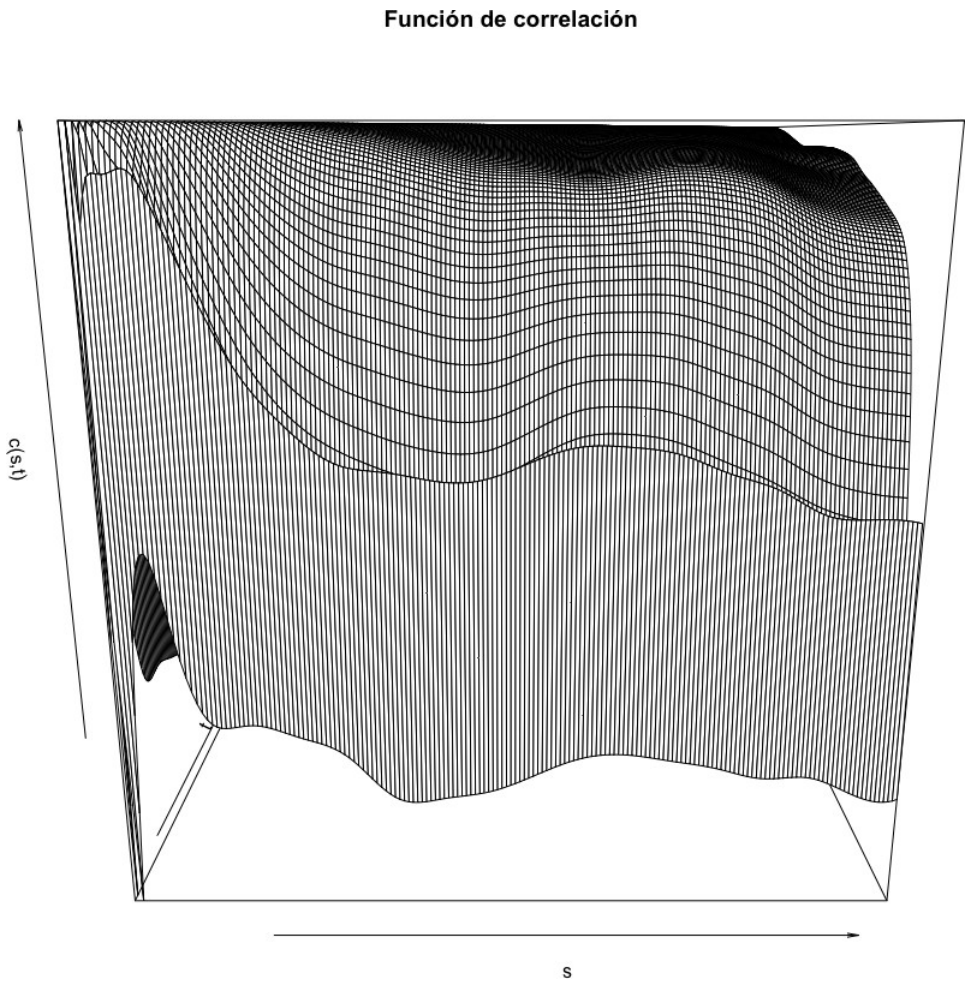


Figura 7.19: Causas de Muerte en el Ecuador: Función de correlación

La figura no es concluyente aunque parece observarse que, a medida que avanza la pandemia las correlaciones son más altas. Las correspondientes curvas de nivel se muestran en la figura **??**.

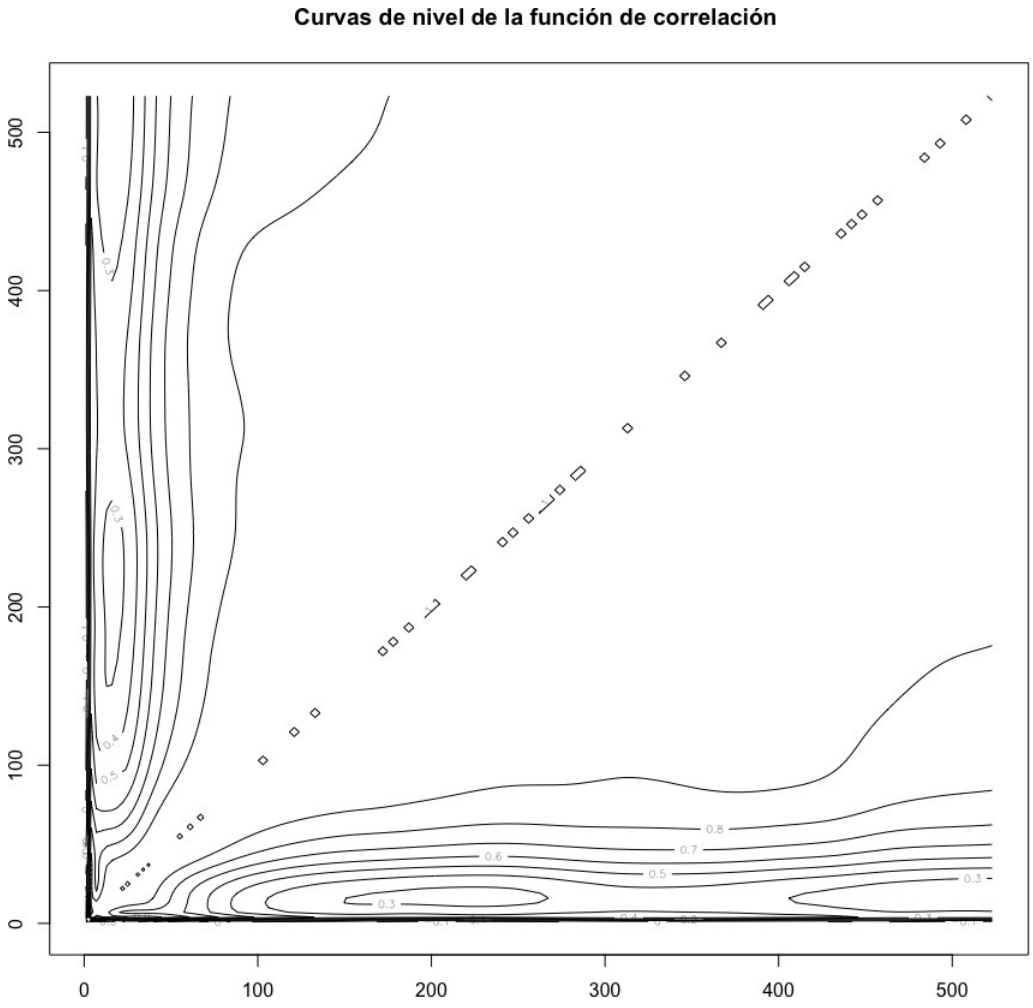


Figura 7.20: Causas de Muerte en el Ecuador: Curvas de nivel de la función de correlación

Como cabía esperar, las correlaciones entre momentos cercanos del tiempo es muy alta y va disminuyendo a medida que nos alejamos en tiempo, si bien a medida que avanza la pandemia, las correlaciones entre momentos más alejados son mayores.

7.2.2. Análisis de Componentes Principales Funcionales

Ajustamos las componentes principales para estudiar la estructura de los datos. La variabilidad explicada por cada una de las componentes se muestra en la tabla siguiente.

	Proporción de varianza	Acumulada
Componente 1	97.55	97.55
Componente 2	1.78	99.33
Componente 3	0.34	99.67
Componente 4	0.20	99.87

Cuadro 7.3: Variabilidad de los datos de COVID en América explicada por las componentes principales funcionales

Las primera componente recoge un 97.55 % de la variabilidad, las dos primeras componentes el 99.33 que es suficiente para explicar correctamente el comportamiento de nuestros datos.

Dibujamos los armónicos (vectores propios) en función del tiempo 7.21. Recordemos que las componentes principales también pueden considerarse como funciones en el tiempo.

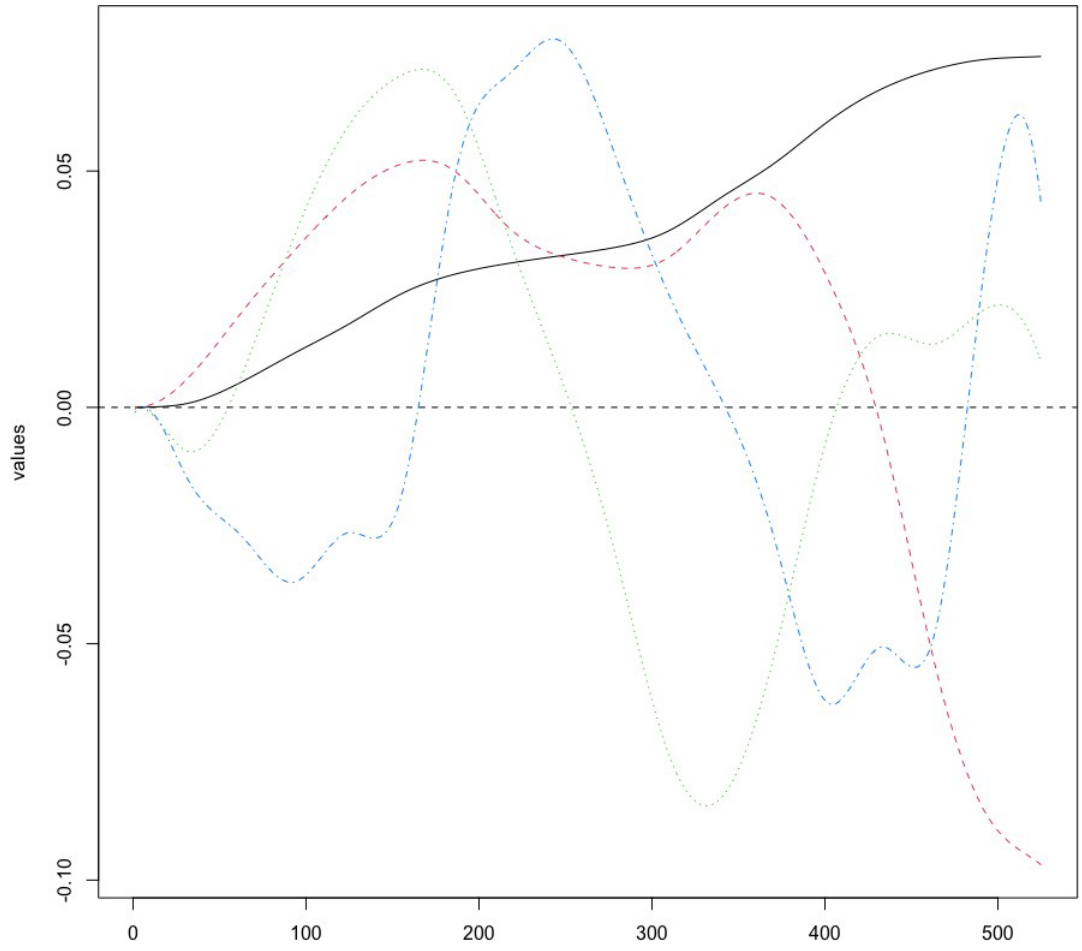


Figura 7.21: Representación de las componentes principales para los datos de COVID en América

La primera componente principal (en negro) recoge el incremento (medio) de la incidencia acumulada que es el comportamiento general de todos los países. Obsérvese que la curva tiene dos oscilaciones que probablemente se corresponden con las dos olas de la epidemia y que tienen dos periodos de crecimiento más rápido y de decrecimiento que se corresponde con la estabilización de la curva acumulada. La segunda componente (en rojo) muestra dos picos con un decrecimiento al final. ES la com-

ponente de “forma”. Valores altos de un país en esta componente significan decrecimientos importantes en el final del periodo mientras que valores bajos (negativos) significan incrementos al final del periodo estudiado.

La tercera componente (en verde) tomará valores elevados en aquellos países que comenzaron la primera ola más pronto, la superaron y tuvieron una segunda ola menos marcada. La cuarta componente (en azul) clasificará los países que tuvieron una segunda ola más tarde y siguen aun en el pico.

Podemos dibujarlos por separado como variaciones alrededor de la media, que es la práctica habitual en este contexto.

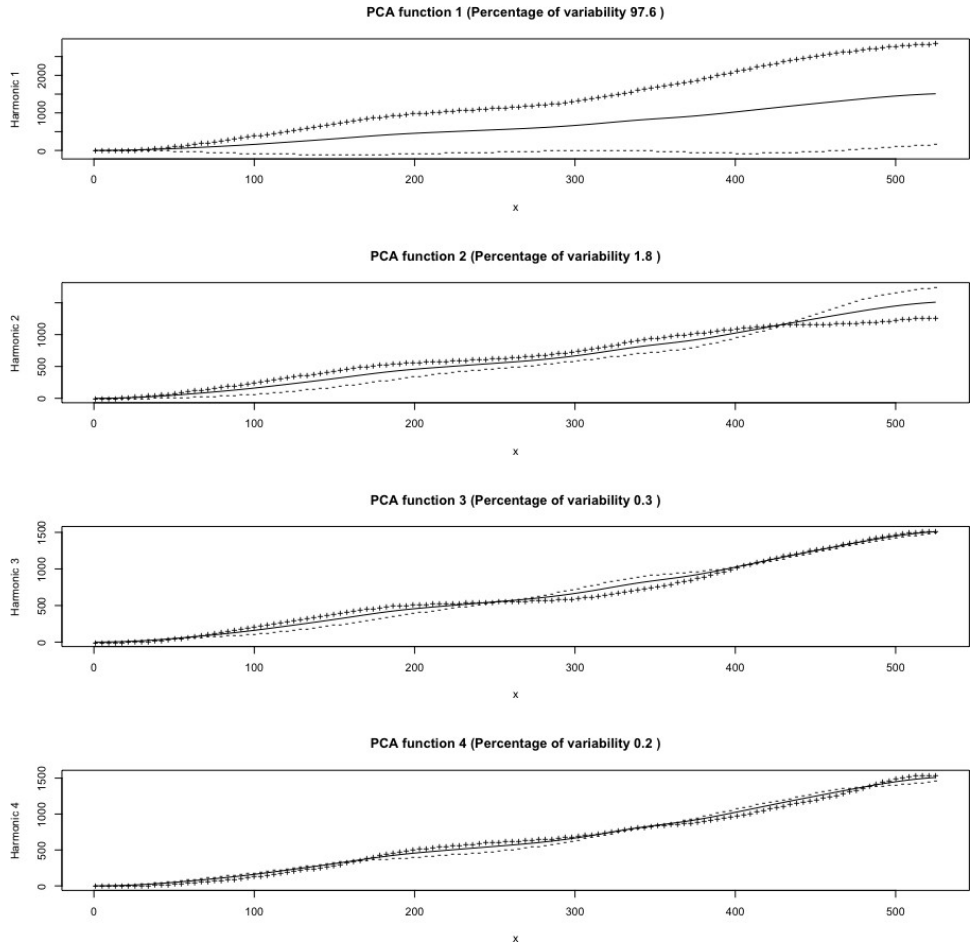


Figura 7.22: Representación de las componentes principales para los datos de COVID en América

Las dos primeras componentes son suficientes para explicar el comportamiento de los datos. La primera muestra el incremento general para los distintos países de forma que, aquellos que presenten mayores puntuaciones sobre la misma serán aquellos para los que se ha producido un mayor incremento en el tiempo, sería una componente de tamaño del crecimiento. La segunda componente sería de forma, aquellos países que presenten puntuaciones más altas en la misma serían aquellos que han moderado su crecimiento con el paso del tiempo mientras que, aquellos que presenten puntuaciones más bajas (probablemente negativas) serán aquellos que han incrementado más us tasa en relación al resto.

La representación conjunta de los dos primeros armónicos se muestra en la figura siguiente.

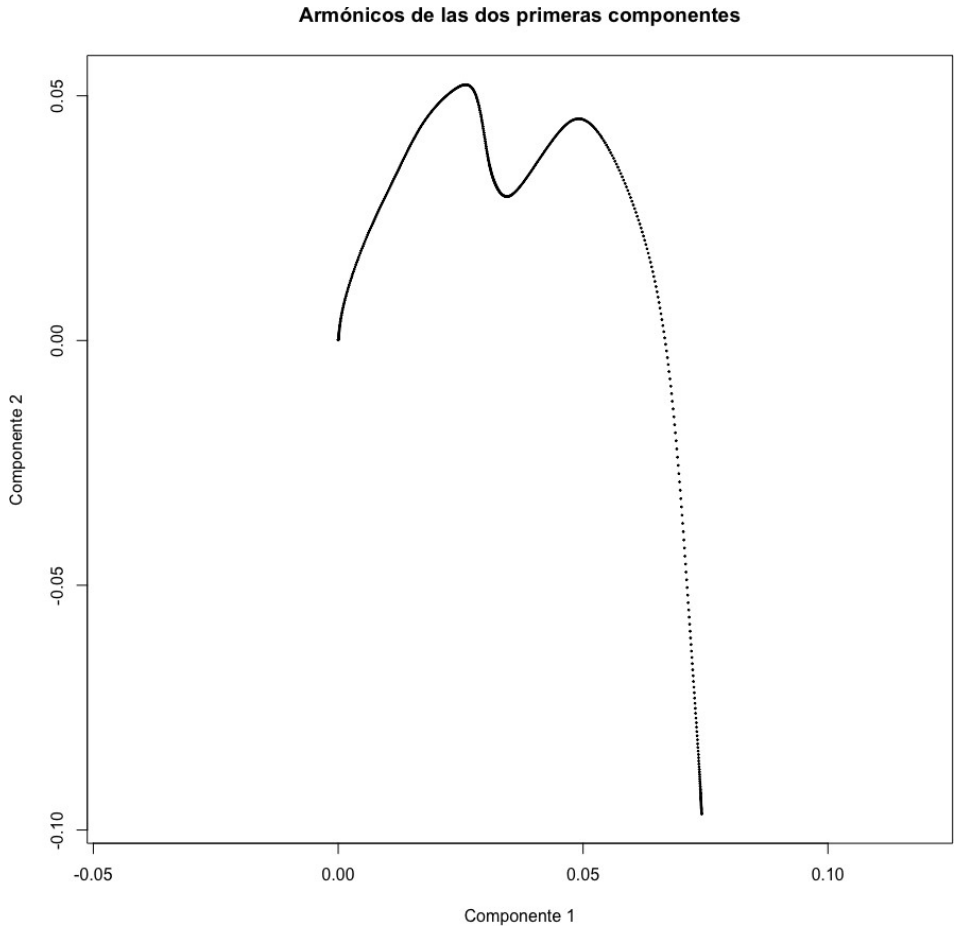


Figura 7.23: Representación conjunta de los armónicos de las dos primeras componentes principales para los datos de COVID en América

La interpretación es un poco más complicada que la del caso de las causas de muerte. Observamos como se produce un rápido crecimiento al principio hasta llegar al pico de la primera ola, después un decrecimiento para volver a crecer hasta el pico de la segunda seguido de un decrecimiento más lento.

Representamos las puntuaciones de cada uno de los países sobre las dos primeras componentes en la figura 7.24

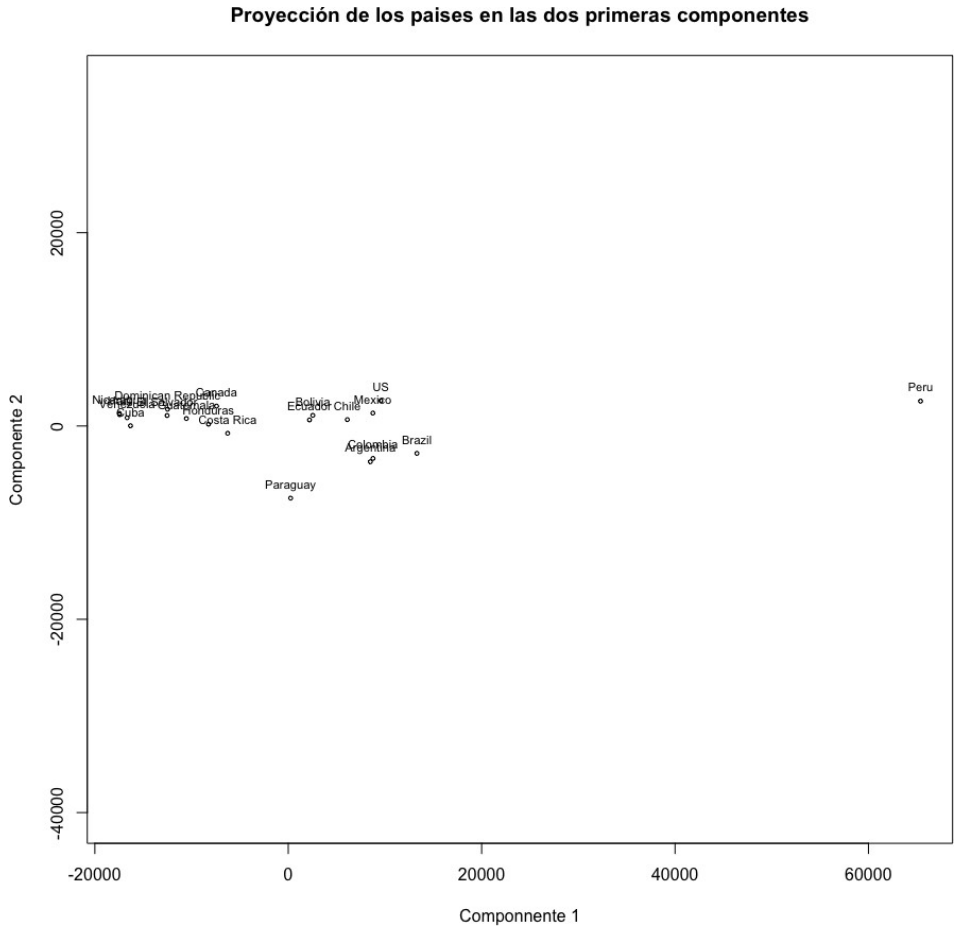


Figura 7.24: Representación de las puntuaciones de los países sobre las dos primeras componentes principales para los datos de COVID en América

Observamos como Perú aparece bastante apartado de los demás y con un valor alto en la primera componente. Eso significa que la incidencia es la más alta de todas y que, además ha aumentado casi de manera constante a lo largo de todo el periodo. El hecho de que los puntos aparezcan en una banda estrecha obedece a que la magnitud de ambas componentes es muy distinta.

La figura 7.25 muestra la misma representación anterior pero sin Perú, con el objeto de ver mejor la representación.

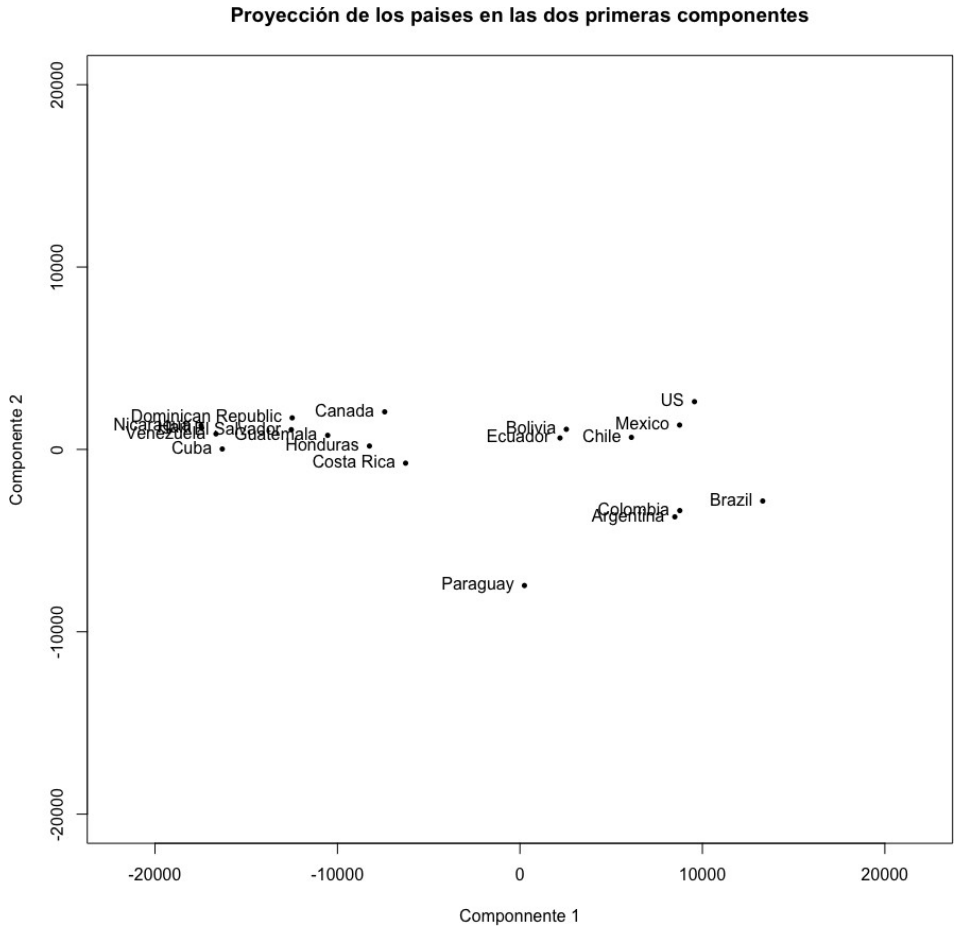


Figura 7.25: Representación de las puntuaciones de los países sobre las dos primeras components (sin Perú)

Hay dos grupos de países con incidencias altas. El primer grupo de países está formado por US, México, Chile, Bolivia y Ecuador. Todos ellos presentan puntuaciones positivas en la componente 2 lo que significa que comenzaron con incidencias más altas que se han ido estabilizando con el tiempo, es decir, tuvieron incidencias más altas en la primara ola. El otro grupo, formado por Argentina, Colombia y Brasil, ha tenido mayor incidencia que el resto en la segunda ola. Paraguay aparece separado de estos grupos y tiene la puntuación más baja de todas en la segunda componente lo que significa que ha tenido un incremento más acentuado en la última parte del

periodo estudiado.

7.2.3. Biplot para datos funcionales

Construimos el biplot y lo representamos.

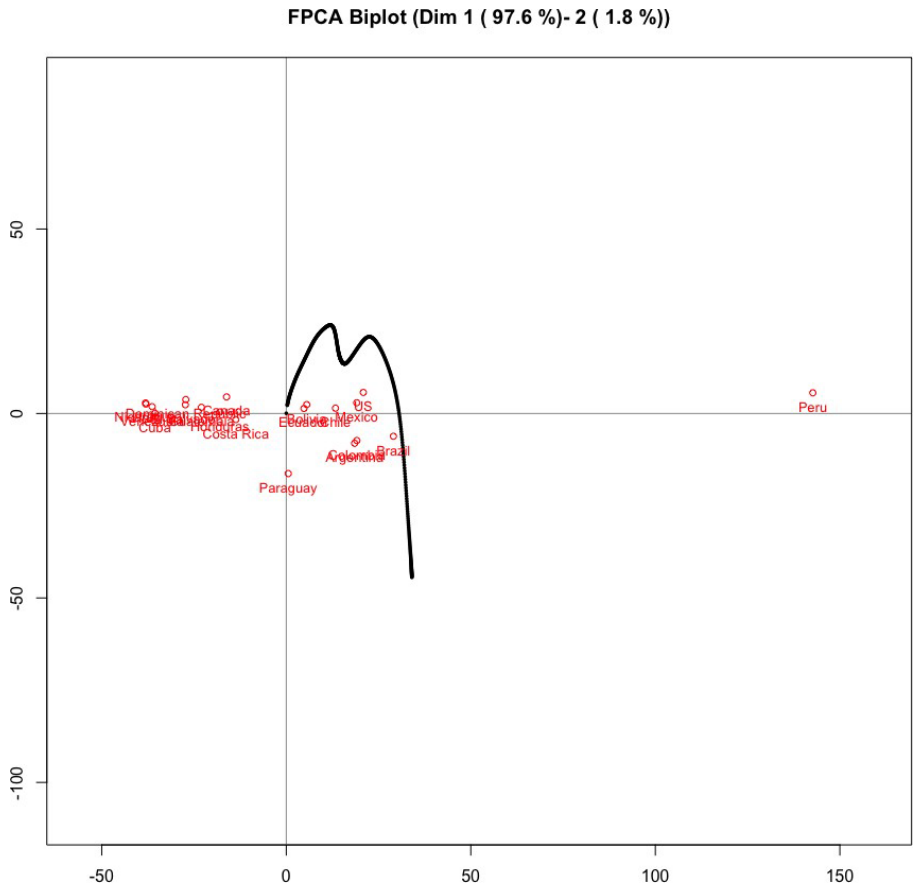


Figura 7.26: Representación de las puntuaciones de los países sobre las dos primeras components (sin Perú)

El aspecto es similar al gráfico que vimos antes. Para interpretar las incidencias al final del periodo proyectamos sobre la última fecha.

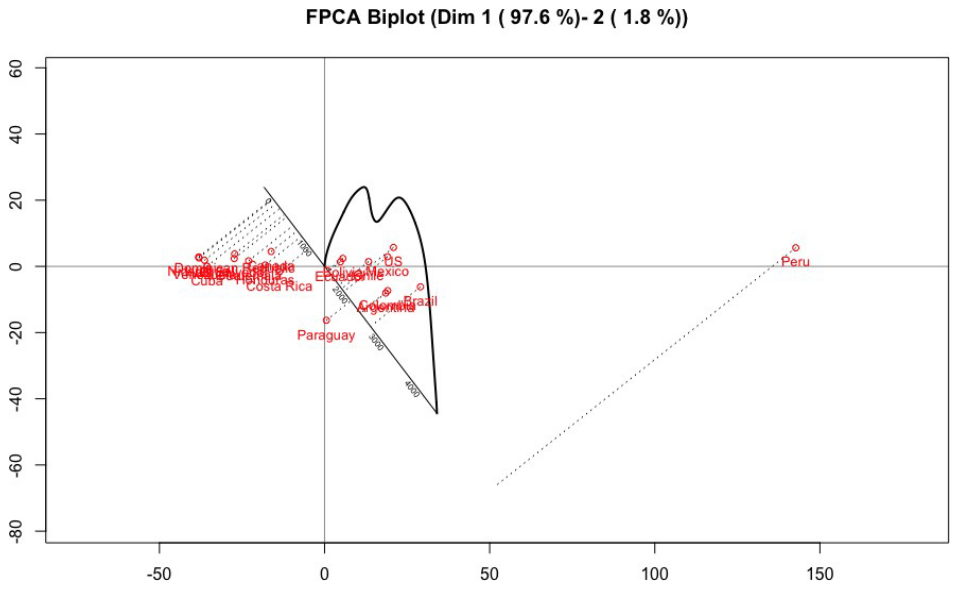


Figura 7.27: Representación de las puntuaciones de los países sobre las dos prime- ras components (sin Perú)

Vemos como Perú dobla la tasa del siguiente país que es Brasil. Para verlo mejor ampliamos el grupo de países que tenemos en el centro.

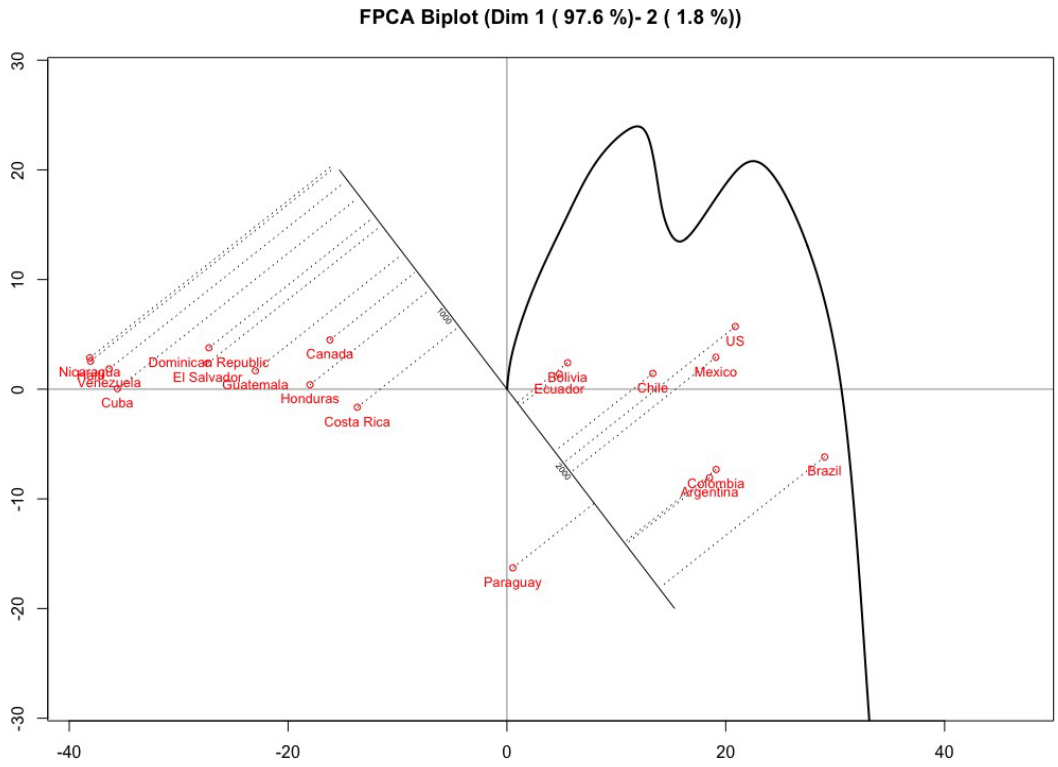


Figura 7.28: Representación de las puntuaciones de los países sobre las dos primeras components (sin Perú)

Después de Perú, el país con mayor incidencia al final del periodo corresponde a Brasil, seguido de Colombia y Argentina que presentan perfiles muy similares. Después encontramos a Paraguay. Todos estos países tienen puntuaciones negativas en la segunda componente por lo que parece que han alcanzado la incidencia final con un incremento en el último periodo.

Sigue el grupo formado por los Estados Unidos, México y Chile y, a continuación encontramos a Ecuador y a Bolivia que presentan perfiles muy similares. Todos estos países tienen puntuaciones positivas en la segunda componente por lo que tuvieron una mayor incidencia en el principio de la pandemia.

El resto de los países presentan incidencias menores y comportamiento más estable a lo largo del tiempo.

Representamos las trayectorias de los países que se han etiquetado en el punto final.

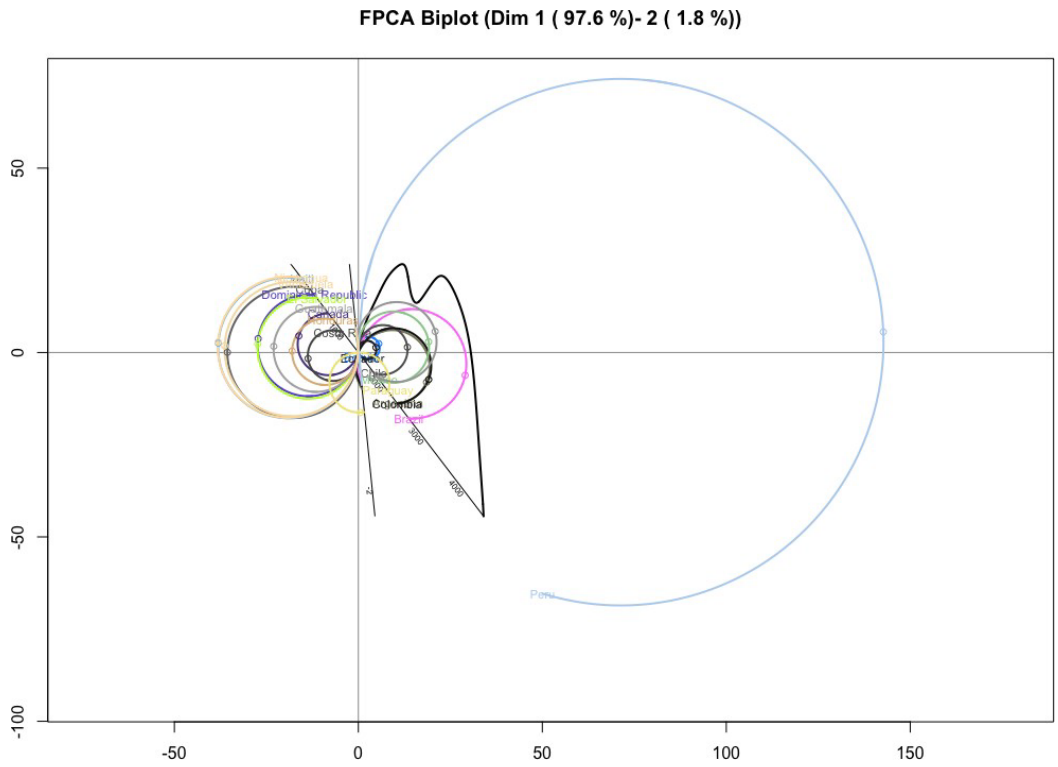


Figura 7.29: Representación de las puntuaciones de los países sobre las dos primeras components (Trayectorias)

Observamos que Perú se ha mantenido con incidencias más altas desde el inicio de la pandemia. Ampliamos el resto para verlos mejor.

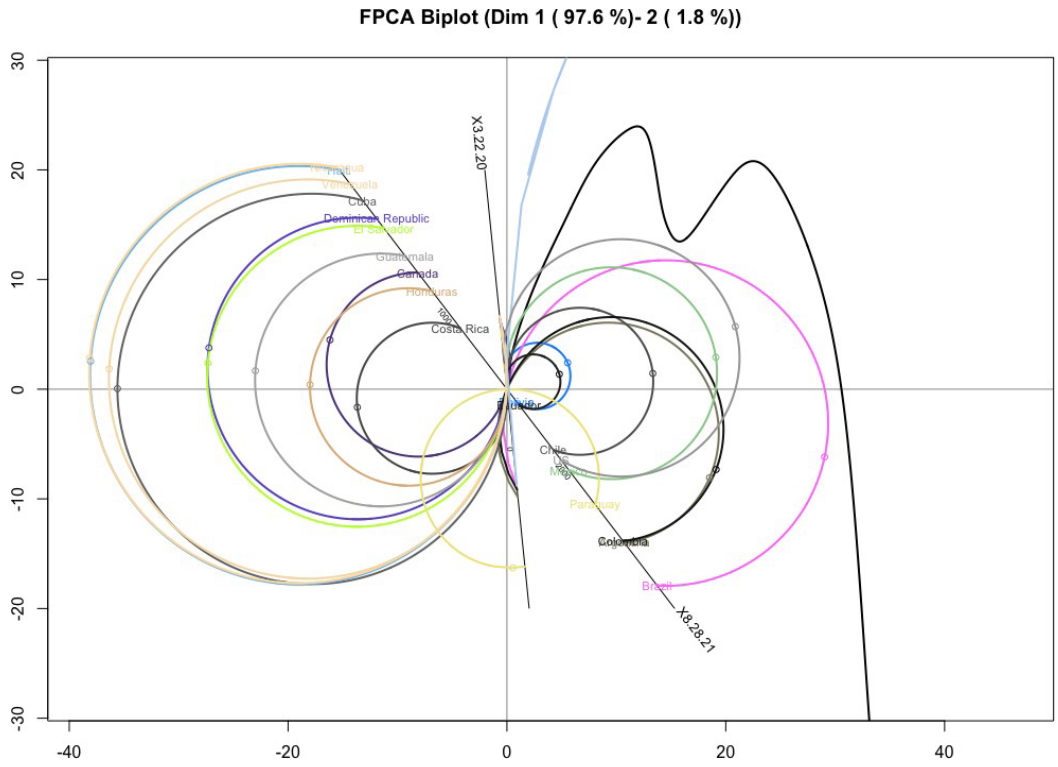


Figura 7.30: Representación de las puntuaciones de los países sobre las dos primeras components (Trayectorias, sin Perú)

Como especificamos en la teoría, las trayectorias muestran la evolución de cada uno de los países. Todos ellos parten del centro, lo que indica que, en las primeras fechas de la pandemia todos se encontraban en situaciones similares con tasas de mortalidad baja. En el gráfico hemos situado una de las primeras fechas del estudio y la final, de forma que podemos comprobar el punto de partida y el final. Teniendo en cuenta que el origen de coordenadas coincide con el centro de gravedad, las trayectorias que aparecen a la izquierda del gráfico son las de aquellos países que se han mantenido por debajo de la media mientras que, las que aparecen a la derecha son las que han estado por encima de la media.

Las mortalidades más bajas se han producido en Panamá, Haiti, Venezuela y Cuba y, en general, en todos los países de Centro América y Canadá.

Alrededor de la media tenemos a Ecuador y Bolivia, que se han mantenido en esa posición a lo largo de toda la pandemia. El resto de los países ha terminado por encima de la media, especialmente Perú que ha tenido mortalidades más altas desde el inicio.

Los cruces en las trayectorias también tienen información importante sobre la evolución. Por ejemplo, Brasil era superado por muchos países al principio y fue superándolos a todos (excepto a Perú) para situarse a la cabeza de las mortalidad. A continuación tenemos a Colombia y Argentina que, en el último periodo también han sufrido un incremento importante en la mortalidad.

El caso de México y Estados Unidos ha sido diferente, mientras que empezaron siendo superados solamente por Perú en el inicio y por Brasil al final de la primera hora, han disminuido sus mortalidades en relación a otros países como Colombia y Argentina.

Vemos, entonces, que la representación biplot produce información adicional a la que proporciona el ACP funcional y junto con las trayectorias propuestas, permite la caracterización de la evolución de la pandemia en los distintos países de América.

7.3. COVID-19 en la regiones del Ecuador

7.3.1. Suavizado de los datos

Disponemos de los datos de mortalidad por COVID-19 en las provincias del Ecuador desde el principio de la pandemia hasta el día 23 de Agosto. Los datos se encuentran en dos archivos diferentes, el primero incluye los datos hasta el final de 2020 y el segundo los datos de 2021. Loas archivos contienen también la población de cada una de las provincias de forma que calcularemos las tasas por cada 10000 habitantes.

En la figura 7.31 mostramos el mapa que contiene la tasa de muertes por COVID por cada 10000 habitantes en la fecha de final de los datos de los que disponemos (23 de Julio).

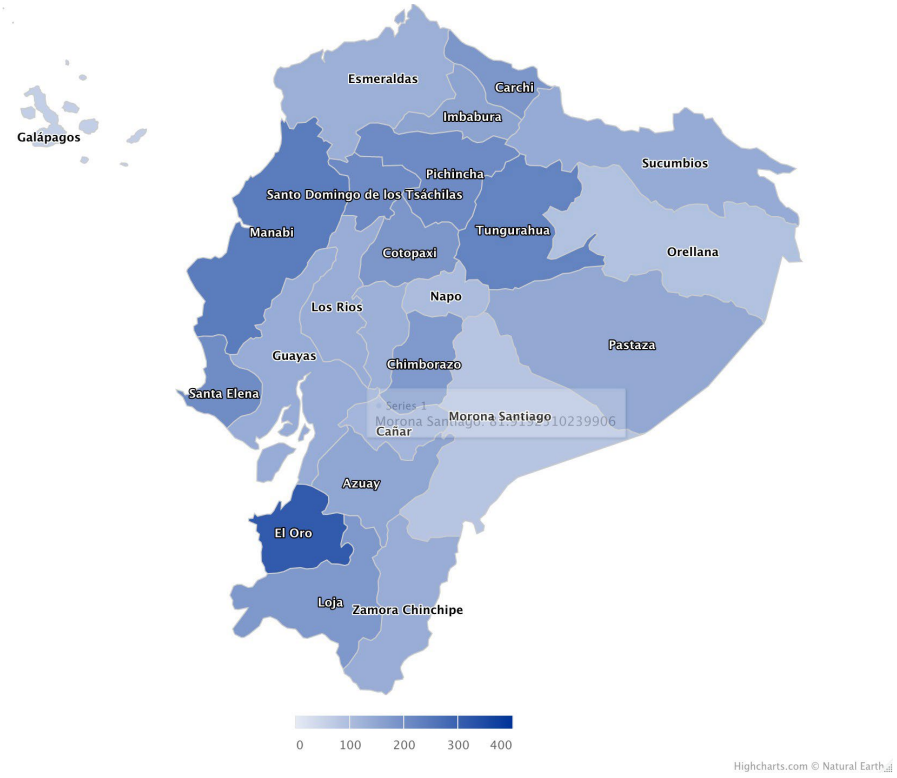


Figura 7.31: Tasa de mortalidad en las provincias del Ecuador a 23 de Julio de 2021 (mapa)

Los datos iniciales se muestran en la figura 7.32.

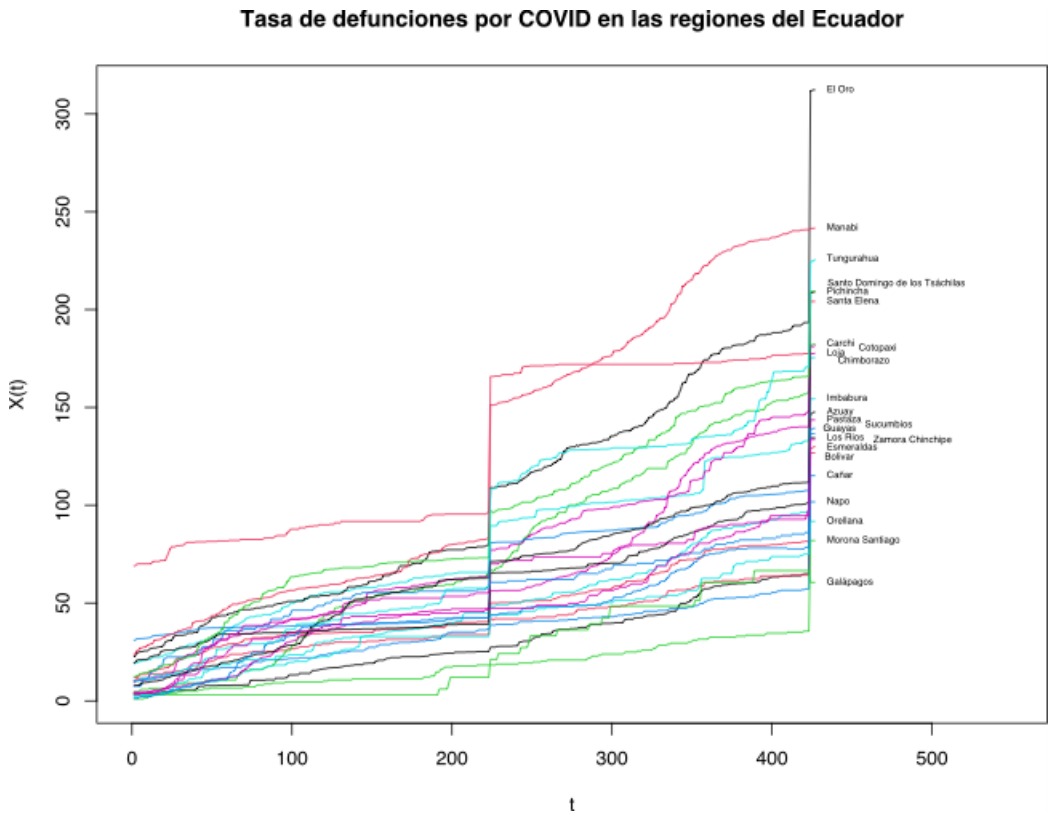


Figura 7.32: Tasa de mortalidad en las provincias del Ecuador hasta el 23 de Julio de 2021.

Observamos en los datos que hay dos saltos importantes que, en algunos casos, duplican las tasas. Los saltos se producen los días 1 de Enero y 20 de Julio de 2021 y se deben probablemente a la actualización de los datos no reportados con anterioridad. En estas circunstancias el suavizado será más difícil. Utilizaremos una base con pocas funciones en la base y conseguir un suavizado mayor que elimine los saltos bruscos. Cuanto mayor sea el número de funciones en la base, la curva suavizada se ajusta mejor a los datos manteniendo los saltos bruscos, por tanto, con mayor posibilidad de sobreajuste.

Los curvas suavizadas se muestran en la figura 7.33.

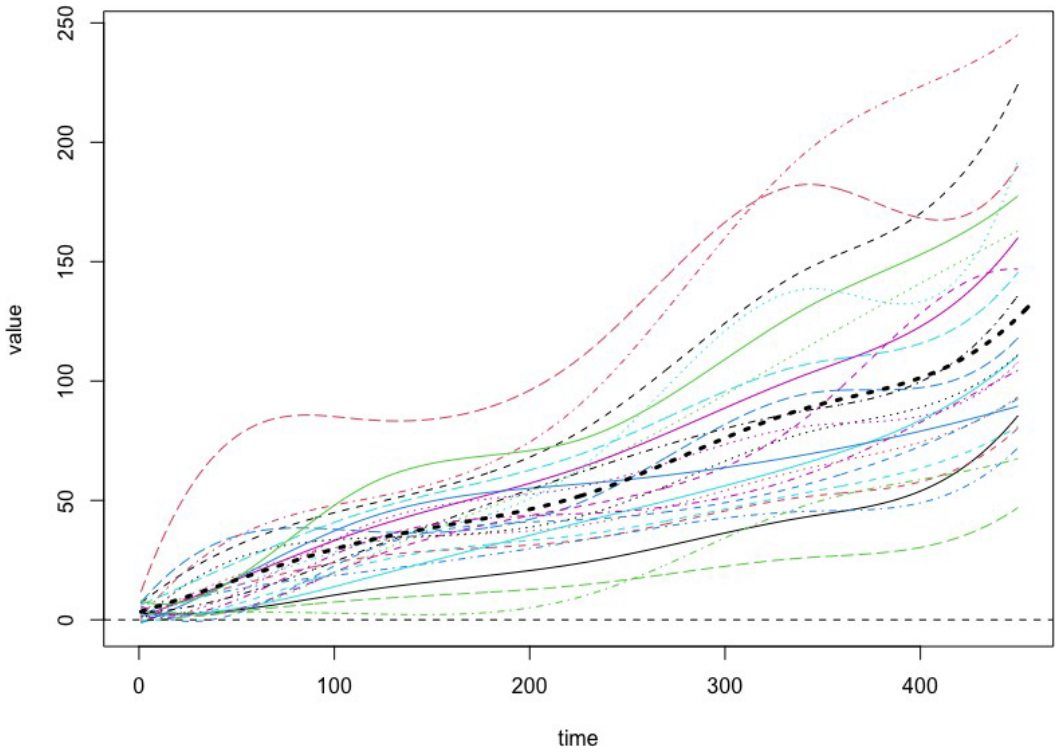


Figura 7.33: Tasa de mortalidad suavizada en las provincias del Ecuador hasta el 23 de Julio de 2021.

Prácticamente todas las curvas se han suavizado correctamente. Solamente dos de ellas parecen comportarse de forma no completamente adecuada. hemos probado otros ajustes y finalmente hemos decidido quedarnos con este suavizado aunque se desv´ıe un poco de los esperado.

7.3.2. Componentes Principales

la figura 7.34 muestra los armónicos de las componentes principales.

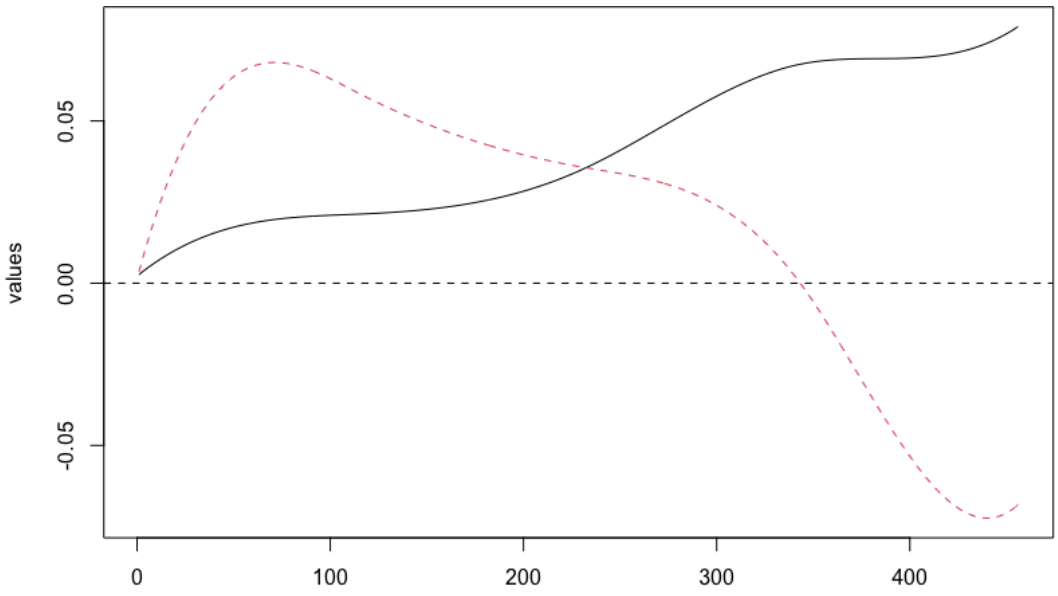


Figura 7.34: Armónicos de las dos primeras componentes principales para los datos de las regiones del Ecuador

Lo mismo que para el caso de los países, la primera componente principal muestra el incremento general en las tasas. La segunda componente muestra la forma de las curvas, las regiones con valores bajos en la componente serán aquellas que tuvieron incrementos al principio y las que tengan valores altos serán aquellas que disminuyeron al final.

La proporción de varianza se muestra en la tabla siguiente.

	Proporción de varianza	Acumulada
Componente 1	92.47	92.47
Componente 2	5.20	97.67

Las dos primeras componentes explican conjuntamente el 97 % de la variabilidad. Las coordenadas de los países sobre las dos primeras componentes se muestran en la

figura 7.35.

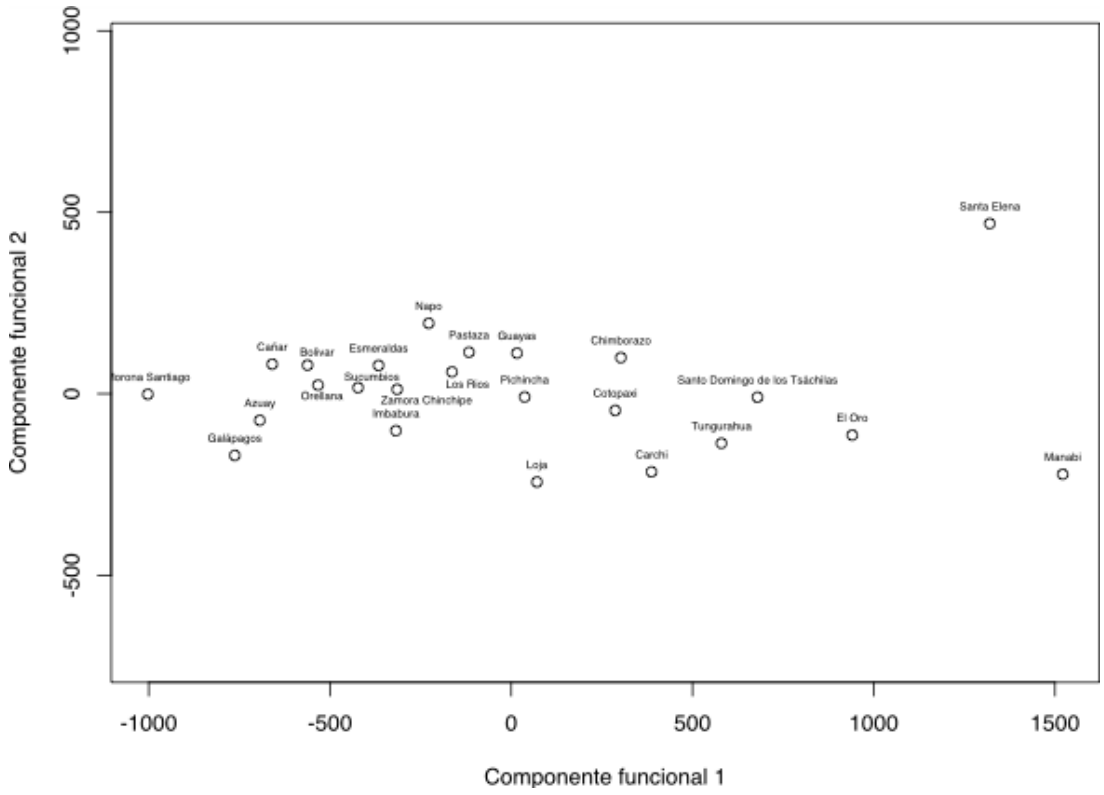


Figura 7.35: Proyección de las regiones sobre las dos primeras componentes principales

La proyección de las regiones sobre la primera componente mostrará, a la derecha las que han tenido mayores tasas y a la izquierda las que menos. Los que estén en la parte positiva de la segunda componente tendrán tasas altas al principio mientras que los que se sitúen en la parte negativa habrán aumentado sus tasa más que el resto en la parte final de la pandemia.

7.3.3. Biplot

Ajustamos el biplot correspondiente que mostramos en la figura 7.36.

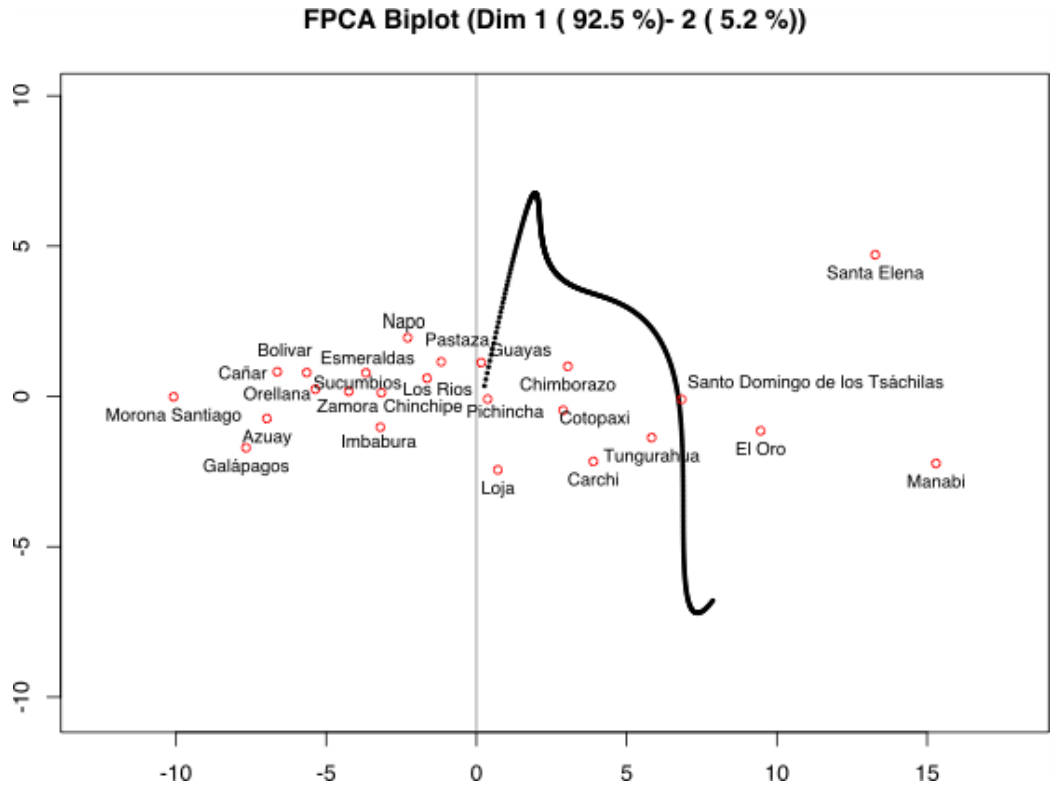


Figura 7.36: Biplot para los datos de las regiones del Ecuador

El biplot nos permite la misma interpretación que las CP añadiendo información sobre el tiempo. Como ya vimos con las componentes, las regiones con coordenadas positivas en el eje 2 como, por ejemplo, *Santa Elena* son las que comenzaron con tasas altas y las disminuyeron con el tiempo mientras que, las regiones situadas en la parte negativa como, por ejemplo *Manabí*, comenzaron con tasas más bajas y las han aumentado con el tiempo con respecto a las demás. Los que tienen puntuaciones cercanas al cero en esta componente se han mantenido más o menos conla misma posición relativa respecto al resto. Ilustramos estas afirmaciones proyectando los puntos sobre el día 20 de los analizados (principio de la pandemia) y sobre el último, en la figura 7.37.

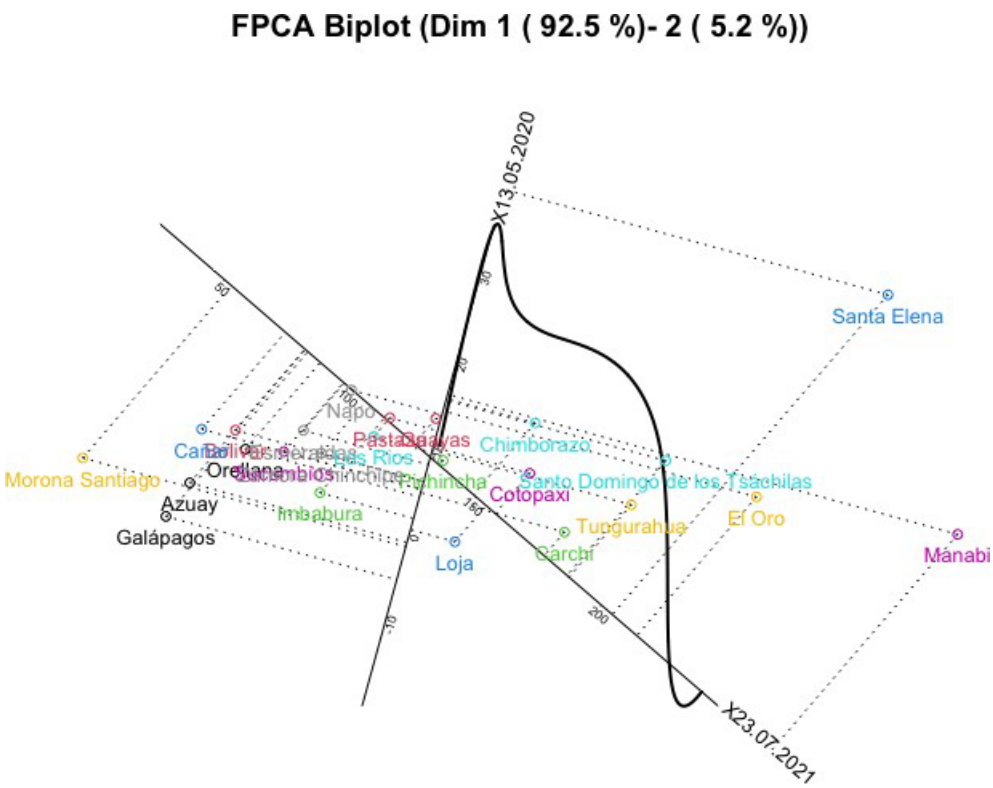


Figura 7.37: Biplot para los datos de las regiones del Ecuador con proyección de las regiones sobre dos de las fechas

Observamos, por ejemplo, como *santa Elena* y *Manabí* han intercambiado sus posiciones desde el principio hasta el final de la pandemia.

La pequeña curva que se observa al final de la trayectoria del biplot puede deberse al incremento inusual que se produce por el efecto de la actualización de los datos.

En la figura 7.38 hemos etiquetado algunas de las fechas. Podemos observar que, en las primeras fechas de la pandemia los ángulos entre los vectores son mucho más pequeños lo que indica unas correlaciones más altas. En el periodo intermedio y final

las flechas se separan, indicando una correlación más baja debida, probablemente, a una evolución desigual de las distintas regiones.

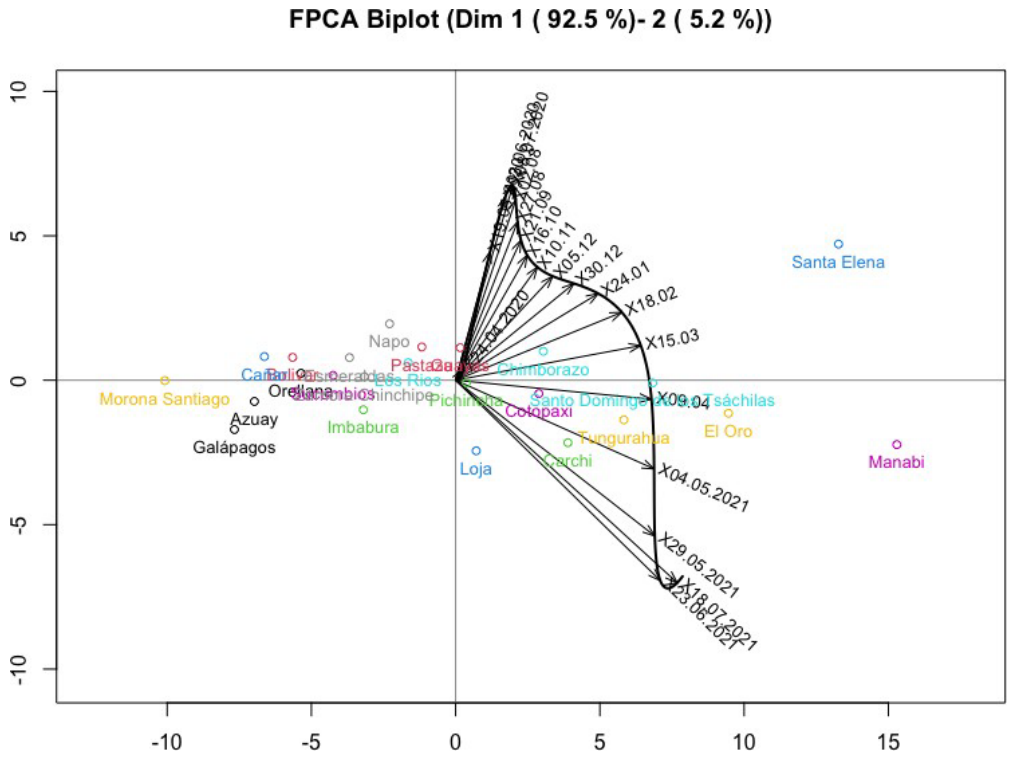


Figura 7.38: Biplot para los datos de las regiones del Ecuador con algunas fechas etiquetadas

El software utilizado permite limpiar el gráfico colocando las etiquetas de las variables fuera del espacio ocupado por los puntos. Mostramos el resultado en la figura 7.39.

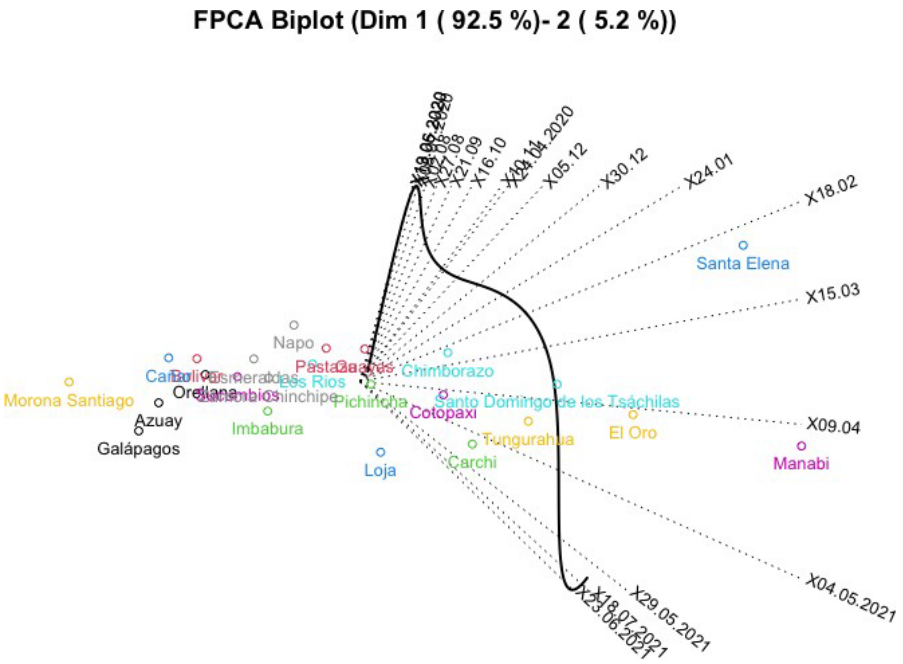


Figura 7.39: Biplot para los datos de las regiones del Ecuador con algunas fechas etiquetadas

Finalmente colocamos las trayectorias en la figura 7.40.

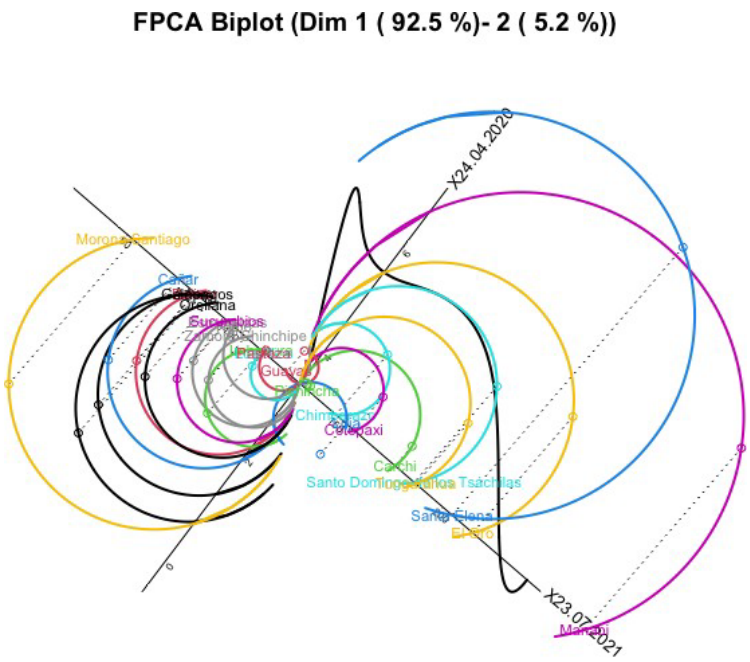


Figura 7.40: Biplot para los datos de las regiones del Ecuador con trayectorias para las regiones

Dejamos la interpretación como ejercicio para el lector.

7.4. Nota de Software

Todos los procedimientos de esta parte se han llevado a cabo usando el paquete *MultBiplotR* [107] de R [86].

El paquete ha incorporado una función que convierte un objeto resultante de las componentes funcionales en un biplot para datos continuos y las herramientas necesarias para implementar las nuevas propuestas como las trayectorias.

Conclusiones

1. La mayor parte de los estudios sobre demografía utilizan solamente técnicas estadísticas descriptivas. La técnica Multivariante más utilizada es el Análisis de Componentes Principales.

2. Las Componentes Principales clásicas siguen siendo útiles como técnicas de reducción de la dimensión para el análisis de datos demográficos y de salud.

3. Los métodos Biplot clásicos son útiles para el análisis de matrices de datos demográficos y de salud aunque hasta la fecha, no han sido muy populares para este tipo de datos. El uso del biplot complementa las componentes principales y mejorar´ıa sustancialmente las aplicaciones de éstas.

4. Los métodos de Análisis de Datos Funcionales (ADF) son útiles en estudios epidemiológicos de mortalidad, capturando algunas características de los datos que no reflejan los métodos multivariantes tradicionales.

5. Se propone la extensión de los métodos biplot para datos funcionales y algunas herramientas gráficas como el análisis de trayectorias para su interpretación.

6. Los métodos de Análisis de Datos Funcionales (ADF) y su extensión biplot, han permitido el análisis de la mortalidad y sus principales causas registradas en las regiones del Ecuador en el periodo 2010-2021, permitiendo estudiar, no solamente la similitud entre los perfiles de las distintas causa sino también su evolución en el tiempo.

7. Los métodos de Análisis de Datos Funcionales (ADF) y su extensión biplot, han sido muy útiles en el análisis de la mortalidad por COVID-19 en América y la

posición relativa del Ecuador en relación al resto de los países. Han permitido no solamente analizar la similitud de las evoluciones de la pandemia en los distintos países sino también establecer la evolución en el tiempo.

8. Los métodos de Análisis de Datos Funcionales (ADF) y su extensión biplot, han sido muy útiles en el análisis de la mortalidad por COVID-19 en las distintas regiones del Ecuador. Han permitido caracterizar las distintas regiones, sus similitudes y su evolución.

Bibliografía

[1] Mario Almagro, Raquel Martínez Unanue, Víctor Fresno Fernández, y Soto Montalvo Herranz. Estudio preliminar de la anotación automática de códigos cie-10 en informes de alta hospitalarios. 2018.

[2] Isidro Rafael Amaro, José Luis Vicente-Villardón, y María Purificación Galindo-Villardón. Manova biplot para arreglos de tratamientos con dos factores basado en modelos lineales generales multivariantes. *Interciencia*, 29(1):26–32, 2004.

[3] Edilia Andrews G, H Fernández, et al. Género helicobacter: una entidad taxonómica en expansión, de características zoonóticas. *Rev. chil. cienc. méd. biol*, págs. 17–24, 1997.

[4] Pamela Apablaza, Néstor Soto, Rossana Román, et al. Nuevas tecnologías en diabetes. *Revista Médica Clínica Las Condes*, 27(2):213–226, 2016.

[5] Lilian Bitencourt Alves Barbosa, Ana Leticia Carnevalli Motta, y Zélia Marilda Rodrigues Resck. Los paradigmas de la modernidad y posmodernidad y el proceso de cuidar en enfermería. *Enfermería Global*, 14(37):335–341, 2015.

[6] Ildar Batyrshin y Michael Wagenknecht. Towards a linguistic description of dependencies in data. *International Journal of Applied Mathematics and Computer Science*, 12:391–401, 2002.

[7] Jean-Paul Benzécri et al. *Lánalyse des données*, tomo 2. Dunod Paris, 1973.

[8] Alejandra Berroteran, Marianella Perrone, María Correnti, María Eugenia Cavazza, Claudio Tombazzi, Lecuna Vicente, y Rosa Goncalvez. Prevalencia de

helicobacter pylori en el estómago y placa dental de una muestra de la población en Venezuela. *Acta Odontológica Venezolana*, 39(2):35–41, 2001.

[9] Aina Faus Bertomeu y Rosa María Gómez Redondo. La reconstrucción de las causas de muerte por el método modicod en el análisis demográfico-sanitario. de la cie-9 a la cie-10 a un nivel de cuarto dígito. *Empiria: Revista de Metodología de ciencias sociales*, (40):167–195, 2018.

[10] Tobia Boschi, Jacopo Di Iorio, Lorenzo Testa, Marzia A Cremona, y Francesca Chiaromonte. Functional data analysis characterizes the shapes of the first covid-19 epidemic wave in italy. *Scientific reports*, 11(1):1–15, 2021.

[11] Nicoleta Breaz. Numerical experiments with least squares spline estimators in a parametric regression model. *Acta Universitatis Apulensis*, (8):50–59, 2004.

[12] Olesia Cárdenas, Purificación Galindo, et al. Los métodos biplot: evolución y aplicaciones. *Revista Venezolana de Análisis de Coyuntura*, 13(1):279–303, 2007.

[13] Ligia Elena Chicaiza Claudio. Nivel de conocimiento y de percepción sobre la utilidad, aplicabilidad y cumplimiento del sistema de clasificación internacional de enfermedades cie-10, en profesionales del ámbito prehospitalario de la ciudad de quito julio-agosto 2020. 2021.

[14] Anthipa Chokesuwattanaskul, Ploypin Lertjitbanjong, Charat Thongprayoon, Tarun Bathini, Konika Sharma, Michael A Mao, Wisit Cheungpasitporn, y Ronpichai Chokesuwattanaskul. Impact of obstructive sleep apnea on silent cerebral small vessel disease: a systematic review and meta-analysis. *Sleep medicine*, 68:80–88, 2020.

[15] Juan Tarquino Calderón Cisneros, Vilma Raffo Babici, Carlos Adeodato Ricaurte Guerrero, y José Luis Vicente Villardón. Análisis multivariado hj-biplot de la ocurrencia de helicobacter pylori como riesgo para cáncer gástrico, en la ciudadela el cristo de consuelo, milagro ecuador. *Boletín de Malariología y Salud Ambiental*, 60(2), 2020.

[16] COVIDSurg Collaborative. Global guidance for surgical care during the COVID-19 pandemic. *The British journal of surgery*, 2020.

[17] Pelayo Correa y R Goldberg. Pathology and molecular pathogenesis of gastric cancer. *Up to Date*, 2008.

[18] Giuseppe Curigliano. How to guarantee the best of care to patients with cancer during the COVID-19 epidemic: The Italian experience. *The oncologist*, 25(6):463, 2020.

[19] J Dauxois y A Pousse. Une extension de l'analyse canonique. quelques applications. En *Annales de l'IHP Probabilités et statistiques*, tomo 11, págs. 355–379. 1975.

[20] Carl De Boor y Carl De Boor. *A practical guide to splines*, tomo 27. springer-verlag New York, 1978.

[21] Jan de Leeuw. Multivariate analysis with linearizable regressions. *Psychometrika*, 53(4):437–454, 1988.

[22] Filippo de Marinis, Ilaria Attili, Stefania Morganti, Valeria Stati, Gianluca Spitaleri, Letizia Gianoncelli, Ester Del Signore, Chiara Catania, Cristiano Rampinelli, y Emanuela Omodeo Salè. Results of multilevel containment measures to better protect lung Cancer patients from COVID-19: the IEO model. *Frontiers in oncology*, 10:665, 2020. ISSN 2234-943X.

[23] Gerson Luis de Moraes Ferrari, Irina Kovalskys, Mauro Fisberg, Georgina Gómez, Attilio Rigotti, Lilia Yadira Cortés Sanabria, Martha Cecilia Yépez García, Rossina Gabriella Pareja Torres, Marianella Herrera-Cuenca, Ioná Zalcman Zimberg, et al. Original research socio-demographic patterning of self-reported physical activity and sitting time in latin american countries: Findings from elans. *BMC Public Health*, 19(1):1–12, 2019.

[24] Elissa Driggin, Mahesh V Madhavan, Behnood Bikdeli, Taylor Chuich, Justin Laracy, Giuseppe Biondi-Zoccai, Tyler S Brown, Caroline Der Nigoghossian, David A Zidar, y Jennifer Haythe. Cardiovascular considerations for patients,

health care workers, and health systems during the COVID-19 pandemic. *Journal of the American College of Cardiology*, 75(18):2352–2371, 2020. ISSN 1558-3597.

[25] N Dyn, D Levin, y I Yad-Shalom. Conditions for regular-spline curves and surfaces. *ESAIM: Mathematical Modelling and Numerical Analysis*, 26(1):177–190, 1992.

[26] Bircan Erbas, Muhammed Akram, Dorota M Gertig, Dallas English, John L Hopper, Anne M Kavanagh, y Rob Hyndman. Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the united states and england–wales. *Journal of epidemiology*, 20(2):159–165, 2010.

[27] Esmeralda Escobar-Muciño, Estrella Escobar-Muciño, y Adriana Gamboa-Pérez. Descripción del sistema crispr-cas y su aplicación como Metodología de punto de cuidado en la detección del sars-cov-2. 2020.

[28] Frédéric Ferraty. *Recent advances in functional data analysis and related topics*. Springer Science & Business Media, 2011.

[29] Frédéric Ferraty y Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[30] K Ruben Gabriel y Charles L Odoroff. Biplots in biomedical research. *Statistics in medicine*, 9(5):469–485, 1990.

[31] Karl Ruben Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[32] Carmen Gloria González, Carolina Serrano, y Paul R Harris. Diagnóstico de la infección poe helicobacter pylori en niños mediante la detección de antígenos en deposiciones. *Revista médica de Chile*, 135(2):182–188, 2007.

[33] Hand D. J. Gower, J.C. *Biplots.* Chapman and Hall, 1996.

[34] J. C. Gower, S. Gardner-Lubbe, y N. le Roux. *Understanding Biplots*. Wiley, 2011.

[35] John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

[36] John C Gower y David J Hand. *Biplots*, tomo 54. CRC Press, 1995.

[37] John C Gower, Sugnet Gardner Lubbe, y Niel J Le Roux. *Understanding biplots*. John Wiley & Sons, 2011.

[38] Dylan Graetz, Asya Agulnik, Radhikesh Ranadive, Yuvanesh Vedaraju, Yichen Chen, Guillermo Chantada, Monika L Metzger, Sheena Mukkada, Lisa M Force, y Paola Friedrich. Global effect of the COVID-19 pandemic on paediatric cancer care: a cross-sectional study. *The Lancet Child & Adolescent Health*, 2021. ISSN 2352-4642.

[39] Michael J Greenacre. *Theory and applications of correspondence analysis*. Academic Press, 1984.

[40] WG Guilford y DR Strombeck. Chronic gastric diseases. *Strombeck's small animal gastroenterology*, 3:275–302, 1996.

[41] Marja-Liisa Hänninen, I Happonen, S Saari, y K Jalava. Culture and characteristics of helicobacter bizzozeronii, a new canine gastric helicobacter sp. *International Journal of Systematic and Evolutionary Microbiology*, 46(1):160–166, 1996.

[42] Amer Harky, Chun Ming Chiu, Thomas Ho Lai Yau, y Sheung Heng Daniel Lai. Cancer patient care during COVID-19. *Cancer Cell*, 37(6):749–750, 2020. ISSN 1535-6108.

[43] Harry H Harman. *Modern factor analysis*. University of Chicago press, 1976.

[44] W Hermanns, K Kregel, W Breuer, y J Lechner. Helicobacter-like organisms: histopathological examination of gastric biopsies from dogs and cats. *Journal of comparative pathology*, 112(3):307–318, 1995.

[45] Eu Pilar Hevia. Educación en diabetes. *Revista Médica Clínica Las Condes*, 27(2):271–276, 2016.

[46] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[47] Ping-I Hsu, Kwok-Hung Lai, Ping-Ning Hsu, Gin-Ho Lo, Hsien-Chung Yu, Wen-Chi Chen, Feng-Woei Tsay, Hui-Chen Lin, Hui-Hwa Tseng, Luo-Ping Ger, et al. Helicobacter pylori infection and the risk of gastric malignancy. *Official journal of the American College of Gastroenterology— ACG*, 102(4):725–730, 2007.

[48] Rob J Hyndman y Han Lin Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010.

[49] Rob J Hyndman y Md Shahid Ullah. Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956, 2007.

[50] INEC. Evolución de las variables investigadas en los censos de población y vivienda del Ecuador 1950, 1962, 19749 1982, 1990, 2001 y 2010. *Inec*, 2010. URL http://www.ecuadorencifras.gob.ec/documentos/web-inec/Publicaciones/Evolucion_variables_1950_2010_24_04_2014.pdf.

[51] Inec. Proyecciones De La Población De La Rep. Del Ecuador. Inf. téc., 2020. URL http://www.inec.gob.ec/proyecciones_poblacionales/metodologia.pdf%5Cnhttp://www.ecuadorencifras.gob.ec/documentos/web-inec/Poblacion_y_Demografia/Proyecciones_Poblacionales/metodologia.pdf.

[52] Instituto Nacional de Estadística y Censos INEC. COMPENDIO ESTADÍSTICO 2015 INEC. Inf. téc., 2015. URL http://www.ecuadorencifras.gob.ec/documentos/web-inec/Bibliotecas/Compendio/Compendio-2015/Compendio.pdf.

[53] Instituto Nacional de Estadística y Censos INEC. Boletín Técnico de Defunciones Generales. Inf. téc., 2021. URL https://www.ecuadorencifras.

gob.ec/documentos/web-inec/Poblacion_y_Demografia/Defunciones_Generales_2020/boletin_tecnico_edg_2020_v1.pdf.

[54] INEC: Instituto Nacional de Estadísticas y Censos. Registro Estadístico De Defunciones Generales. Inf. téc., 2018. URL http://www3.inegi.org.mx/sistemas/sisept/Default.aspx?t=mdemo125&s=est.

[55] Instituto Nacional de Estadística y Censos (INEC). Nacimientos y Defunciones 2017. Inf. téc., 2017. URL http://www.ecuadorencifras.gob.ec/nacimientos_y_defunciones/.

[56] J Edward Jackson. *A user's guide to principal components*, tomo 587. John Wiley & Sons, 2005.

[57] Ian T Jolliffe. *Principal components in regression analysis*. Springer, 1986.

[58] V Kumar, Ashish Sood, Shaphali Gupta, y Nitish Sood. Prevention-versus promotion-focus regulatory efforts on the disease incidence and mortality of covid-19: A multinational diffusion study using functional data analysis. *Journal of International Marketing*, 29(1):1–22, 2021.

[59] Alexander Kutikov, David S Weinberg, Martin J Edelman, Eric M Horwitz, Robert G Uzzo, y Richard I Fisher. A war on two fronts: cancer care in the time of COVID-19. 2020.

[60] Luis Ávila Lachica, Javier Sangrós González, Antonio García Ruiz, Francisco Javier García-Soidán, José Manuel Millaruelo Trillo, Daniel Bordonaba Bosque, y Juan Mart´ınez Candela. Coste del tratamiento farmacológico de los factores de riesgo cardiovascular en población diabética anciana según género (estudio escadiane). *Atención Primaria Práctica*, 1(5):73–79, 2019.

[61] Sungbok Lee, Dani Byrd, y Jelena Krivokapić. Functional data analysis of prosodic effects on articulatory timing. *The Journal of the acoustical society of America*, 119(3):1666–1671, 2006.

[62] Annette Leibing. The turn towards prevention–moral narratives and the vascularization of alzheimer's disease. *New Genetics and Society*, 39(1):31–51, 2020.

[63] Andrew E Libby, Elise S Bales, Jenifer Monks, David J Orlicky, y James L McManaman. Perilipin-2 deletion promotes carbohydrate-mediated browning of white adipose tissue at ambient temperature. *Journal of lipid research*, 59(8):1482–1500, 2018.

[64] Shiliang Liu, Wee-Shian Chan, Joel G Ray, Michael S Kramer, KS Joseph, y Canadian Perinatal Surveillance System (Public Health Agency of Canada). Stroke and cerebrovascular disease in pregnancy: incidence, temporal trends, and risk factors. *Stroke*, 50(1):13–20, 2019.

[65] Chey Loveday, Amit Sud, Michael E Jones, John Broggio, Stephen Scott, Firza Gronthound, Beth Torr, Alice Garrett, David L Nicol, y Shaman Jhanji. Prioritisation by FIT to mitigate the impact of delays in the 2-week wait colorectal cancer referral pathway during the COVID-19 pandemic: a UK modelling study. *Gut*, 2020. ISSN 0017-5749.

[66] Diklar Makola, David A Peura, y Sheila E Crowe. Helicobacter pylori infection and related gastrointestinal diseases. *Journal of clinical gastroenterology*, 41(6):548–558, 2007.

[67] Camille Maringe, James Spicer, Melanie Morris, Arnie Purushotham, Ellen Nolte, Richard Sullivan, Bernard Rachet, y Ajay Aggarwal. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. *The lancet oncology*, 21(8):1023–1034, 2020. ISSN 1470-2045.

[68] Jesús Martın-Rodrıguez, Ma Purificación Galindo-Villardón, y José L Vicente-Villardón. Comparison and integration of subspaces from a biplot perspective. *Journal of Statistical Planning and Inference*, 102(2):411–423, 2002.

[69] CATHERINE C McGowan, TIMOTHY L Cover, y MARTIN J Blaser. Heli-

cobacter pylori and gastric acid: biological and therapeutic implications. *Gastroenterology*, 110(3):926–938, 1996.

[70] Matthew McKeever y Nicholas H Wolfinger. Reexamining the economic costs of marital disruption for women. *Social Science Quarterly*, 82(1):202–217, 2001.

[71] France Meslé y Jacques Vallin. Reconstructing long-term series of causes of death: the case of france. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29(2):72–87, 1996.

[72] Osval Antonio Montesinos-López y Carlos Moisés Hernández-Suárez. Modelos matemáticos para enfermedades infecciosas. *Salud pública de México*, 49(3):218–226, 2007.

[73] Jakub Morze, Carolina Schwedhelm, Aleksander Bencic, Georg Hoffmann, Heiner Boeing, Katarzyna Przybylowicz, y Lukas Schwingshackl. Chocolate and risk of chronic disease: a systematic review and dose-response meta-analysis. *European journal of nutrition*, 59(1):389–397, 2020.

[74] Sergio Ignacio Muñoz-Fernández, Diana Molina-Valdespino, Rosalba Ochoa-Palacios, Oscar Sánchez-Guerrero, y Juan Antonio Esquivel-Acevedo. Estrés, respuestas emocionales, factores de riesgo, psicopatología y manejo del personal de salud durante la pandemia por covid-19. *Acta Pediátrica de México*, 41(S1):127–136, 2020.

[75] N Ngoi, J Lim, S Ow, W Ying Jen, M Lee, W Teo, J Ho, R Sundar, M L Tung, y Y M Lee. A segregated-team model to maintain cancer care during the COVID-19 outbreak at an academic center in Singapore. *Annals of Oncology*, 2020. ISSN 0923-7534.

[76] Carlos H Orces, Martha Montalvan, y Daniel Tettamanti. Prevalence of abdominal obesity and its association with cardio metabolic risk factors among older adults in ecuador. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 11:S727–S733, 2017.

[77] World Health Organization. Las 10 principales causas de defunción. URL https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[78] Esteban Ortiz-Prado, Katherine Simbaña-Rivera, Lenin Gómez Barreno, Ana Maria Diaz, Alejandra Barreto, Carla Moyano, Vannesa Arcos, Eduardo Vásconez-González, Clara Paz, Fernanda Simbaña-Guaycha, et al. Epidemiological, socio-demographic and clinical features of the early phase of the covid-19 epidemic in ecuador. *PLoS Neglected Tropical Diseases*, 15(1):e0008958, 2021.

[79] Adrián Padilla-Segarra, Mabel González-Villacorte, Isidro R Amaro, y Saba Infante. Brief review of functional data analysis: A case study on regional demographic and economic data. En *Conference on Information and Communication Technologies of Ecuador*, págs. 163–176. Springer, 2020.

[80] Vinidh Paleri, John Hardman, Theofano Tikka, Paula Bradley, Paul Pracy, y Cyrus Kerawala. Rapid implementation of an evidence-based remote triaging system for assessment of suspected referrals and patients with head and neck cancer on follow-up after treatment during the COVID-19 pandemic: Model for international collaboration. *Head and neck*, 42(7):1674–1680, 2020. ISSN 1043-3074.

[81] Hui Pang, Qiang Fu, Qiumei Cao, Lin Hao, y Zhenkun Zong. Sex differences in risk factors for stroke in patients with hypertension and hyperhomocysteinemia. *Scientific reports*, 9(1):1–9, 2019.

[82] Clara Paz, Guido Mascialino, Lila Adana-Díaz, Alberto Rodríguez-Lorenzana, Katherine Simbaña-Rivera, Lenin Gómez-Barreno, Maritza Troya, María Ignacia Paez, Javier Cárdenas, y Rebekka M Gerstner. Behavioral and sociodemographic predictors of anxiety and depression in patients under epidemiological surveillance for COVID-19 in Ecuador. *PLoS One*, 15(9):e0240008, 2020. ISSN 1932-6203.

[83] Karl Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

[84] Dudley L Poston y Michael Micklin. *Handbook of population*. Springer, 2005.

[85] Hartmut Prautzsch, Wolfgang Boehm, y Marco Paluszny. *Bézier and B-spline techniques*, tomo 6. Springer, 2002.

[86] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[87] Paola Ramírez-Bulla, Marcela Mercado-Reyes, Alba Alicia Trespalacios-Rangel, Jenny Avila-Coy, y William Otero-Regino. Estado actual de la resistencia de helicobacter pylori a tetraciclina: revisión sistemática de la literatura. *Universitas scientiarum*, 17(2):216–229, 2012.

[88] Alberto Ramírez Ramos y Rolando Sánchez Sánchez. Helicobacter pylori y cáncer gástrico. *Revista de Gastroenterologia Del Peru*, 28(3):258–266, 2008.

[89] James Ramsay y Bernard W Silverman. *Functional data analysis*. 2005.

[90] James O Ramsay y Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363, 1998.

[91] James O Ramsay y Bernard W Silverman. *Applied functional data analysis: methods and case studies*, tomo 77. Springer, 2002.

[92] JO Ramsay, Giles Hooker, y Spencer Graves. *Functional data analysis with R and MATLAB*. Springer, 2008.

[93] Severin Rodler, Maria Apfelbeck, Christian Stief, Volker Heinemann, y Jozefina Casuscelli. Lessons from the coronavirus disease 2019 pandemic: Will virtual patient management reshape uro-oncology in Germany? *European journal of cancer*, 132:136–140, 2020. ISSN 0959-8049.

[94] Sebastián Rojas-Lara, Carlos Eduardo Barragán, Martín Alonso Bayona-Rojas, Ricardo Oliveros, y Andrés Julián Gutiérrez-Escobar. Detección de helicobacter pylori por pcr del gen 16s en biopsias gástricas colectadas en la ciudad de bogotá: estudio preliminar. *Medicina*, 37(3):215–222, 2015.

[95] Alfonso Ruiz-Bravo y María Jiménez-Valera. Sars-cov-2 y pandemia de síndrome respiratorio agudo (covid-19). *Ars Pharmaceutica (Internet)*, 61(2):63–79, 2020.

[96] Karen K Stout, Curt J Daniels, Jamil A Aboulhosn, Biykem Bozkurt, Craig S Broberg, Jack M Colman, Stephen R Crumb, Joseph A Dearani, Stephanie Fuller, Michelle Gurvitz, et al. 2018 aha/acc guideline for the management of adults with congenital heart disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 73(12):e81–e192, 2019.

[97] Nicholas J Talley, Kwong Ming Fock, y Paul Moayyedi. Gastric cancer consensus conference recommends helicobacter pylori screening and treatment in asymptomatic persons from high-risk populations to prevent gastric cancer. *Official journal of the American College of Gastroenterology— ACG*, 103(3):510–514, 2008.

[98] Jun Tashiro, Jun Miwa, Takashige Tomita, Yasuo Matsubara, y Yasuhiko Oota. Gastric cancer detected after helicobacter pylori eradication. *Digestive Endoscopy*, 19(4):167–173, 2007.

[99] Santiago Vicente Tavera. *Las técnicas de representación de datos multidimensionales en el estudio del Índice de Producción Industrial (IPI) en la CEE*. Tesis Doctoral, Universidad de Salamanca, 1992.

[100] Wesley K Thompson. The statistical analysis of functional mri data by lazar, na. 2009.

[101] Masumi Ueda, Renato Martins, Paul C Hendrie, Terry McDonnell, Jennie R Crews, Tracy L Wong, Brittany McCreery, Barbara Jagels, Aaron Crane, y David R Byrd. Managing cancer care during the COVID-19 pandemic: agility and collaboration toward a common goal. *Journal of the National Comprehensive Cancer Network*, 18(4):366–369, 2020. ISSN 1540-1405.

[102] Shahid Ullah y Caroline F Finch. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):1–12, 2013.

[103] Mariano J Valderrama, Francisco A Ocaña, y Ana M Aguilera. Forecasting pc-arima models for functional data. En *Compstat*, págs. 25–36. Springer, 2002.

[104] A Valdés. Helicobacter spp.¿ nuevo patógeno en caninos y felinos. *Monogr. Med. Vet*, 20:117–123, 2000.

[105] Eeke Van der Burg, Jan De Leeuw, y Garmt Dijksterhuis. Overals: Nonlinear canonical correlation with k sets of variables. *Computational Statistics & Data Analysis*, 18(1):141–163, 1994.

[106] JL Vicente-Villardón. *Una alternativa a los métodos factoriales clásicos basada en una generalización de los métodos biplot*. Tesis Doctoral, 1992.

[107] Jose Luis Vicente-Villardon. *MultBiplotR: Multivariate Analysis Using Biplots in R*, 2021. URL https://CRAN.R-project.org/package=MultBiplotR. R package version 1.6.14.

[108] María Purificación Galindo Villardón. Una alternativa de representacion simultanea: Hj-biplot. *Qüestiió: quaderns d'estadística i investigació operativa*, págs. 13–23, 1986.

[109] Christopher J D Wallis, James W F Catto, Antonio Finelli, Adam W Glaser, John L Gore, Stacy Loeb, Todd M Morgan, Alicia K Morgans, Nicolas Mottet, y Richard Neal. The Impact of the COVID-19 Pandemic on Genitourinary Cancer Care: Re-envisioning the Future. *European urology*, 2020. ISSN 0302-2838.

[110] Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, y Yong Xiong. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *Jama*, 323(11):1061–1069, 2020. ISSN 0098-7484.

[111] Xintian Xu, Ping Chen, Jingfang Wang, Jiannan Feng, Hui Zhou, Xuan Li, Wu Zhong, y Pei Hao. Evolution of the novel coronavirus from the ongoing wuhan outbreak and modeling of its spike protein for risk of human transmission. *Science China Life Sciences*, 63(3):457–460, 2020.

[112] Weikai Yan y Manjit S Kang. *GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists*. CRC press, 2002.

[113] J Zhang. Analysis of variance for functional data. *Monographs on statistics and applied probability*, 127:127, 2014.