UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

SCORE PARA LA EVALUACIÓN Y GESTIÓN DE LA CONDUCCIÓN RIESGOSA EN CONDUCTORES PROFESIONALES

POR

EVELYN ALEJANDRA PAFIÁN MARTÍNEZ

Tesis presentada a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción para optar al título profesional de Ingeniera Civil Matemática

Supervisión por

Prof. Guía. Bernardo Moisés Lagos Álvarez

Prof. Co-guía. Fernando Viego Campillos

Concepción, Septiembre de 2023

© 2023 Evelyn Alejandra Pafián Martínez
Se autoriza la reproducción total o parcial, con fines académicos, por cualquier medio o procedimiento, incluyendo la cita bibliográfica del documento.

UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

SCORE PARA LA EVALUACIÓN Y GESTIÓN DE LA CONDUCCIÓN RIESGOSA EN CONDUCTORES PROFESIONALES

POR

EVELYN ALEJANDRA PAFIÁN MARTÍNEZ

Comisión Evaluadora

Bernardo Lagos Álvarez
Departamento de Estadística
Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción.

Fernando Viego Campillos
Gerente de Productos, Wisetrack Corp.
Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile.

Luisa Rivas Calabrán
Departamento de Estadística
Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción.

Patricio Salas Fernández
Departamento de Estadística
Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción.

Septiembre 2023

Resumen

El objetivo es desarrollar/aplicar mejoras al *Índice de Conducción Riesgosa (ICR)*, el cual tiene el fin de puntuar el riesgo entre los diferentes viajes realizados por los diferentes tipos de automóviles, usando siete comportamientos riesgosos de conducción.

Cada etapa del proceso de construcción de un *índice* es extremadamente importante, por lo que se prestó especial atención en que las metodologías establecidas en cada etapa sean afines entre sí. Inicialmente se verifica el rol de cada *indicador* usado para la construcción del *índice*, así como también la *normalización* adecuada, y métodos de *ponderación* y *agregación* compatibles, pues ambas son etapas entrelazadas. Para obtener ponderaciones se usa el *Análisis de Componentes Principales*, *Análisis Factorial* y *Benefit of the Doubt*. Dado que el *Análisis de Componentes Principales* y *Benefit of the Doubt* en su metodología de ponderación preestablecían una agregación lineal, para la etapa de agregación se usa *lineal* y además, *media cóncava*, la cual tiene por objetivo "penalizar" los *desbalances* entre las diferentes *dimensiones*.

Con la ayuda del *Análisis de Sensibilidad* y *Robustez*, se aprecia si los *índices* obtenidos aportan validez y resultados congruentes. Entre los que se destacan el obtenido por *Benefit of the Doubt* con agregación *lineal* con coeficiente de variación menor en comparación de los otros índices obtenidos, logrando una ordenación más robusta, pues al introducir perturbaciones a los indicadores elementales, el índice obtenido se ve menos afectado por estas, posiblemente por el propio procedimiento de construcción, lo que permitiría "absorber" en cierta medida el efecto de la variabilidad inducida.

Keywords: *risk index, index, composite indicator, aggregation method, weighting methods, performance index*

AGRADECIMIENTOS

Doy gracias a mi papá y mi mamá por su apoyo y paciencia durante estos años mientras cursaba cada asignatura. A mi hermana por su compañía, apoyo y alegría, y a mi abuela por su cariño incondicional.

A mis compañeros, nuestro "Equipo" como Coni acostumbra a llamarnos, por las pequeñas alegrías y desgracias que compartimos y espero sigamos compartiendo.

A los profesores que compartieron sus conocimientos, entre los cuales, quiero destacar al Prof. Freddy Paiva quien estuvo dispuesto a aconsejar, así como a explicar y motivar; al profesor que fue guía de mi memoria, el Dr. Bernardo Lagos, por su paciencia y disposición por ayudar. Al Ing. Fernando Viego, quien intenta sacar lo mejor de las personas, apoyando con sus conocimientos y disposición.

Para concluir, a Wisetrack Corp., quienes me acogieron al realizar mi práctica y memoria de título y donde tuve la oportunidad de conocer personas con las que aún mantengo contacto.

Evelyn Pafián Martínez
Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción
Concepción, Chile
Agosto 2023

Índice general

Índice de Figuras

Índice de Tablas

Índice de Abreviaturas

AHP	Analytic Hierchy Processes
BA	Budget Allocation
BoD	Benefit of the Doubt
BP	Breakdown Point (Punto de ruptura)
CA	Conjoint Analysis
CP	Componente Principal
DEA	Data Envelopment Analysis
DF	Grados de Libertad
DS	Duration Score
EL	Expected Loss (Pérdida Esperada)
EW	Equal Weight
HDI	Índice de Desarrollo Humano
ICR	Índice de Conductor Riesgoso
MCD	Minimun Covariance Determinant
MSD	Stahel Donoho Estimator
ORQ	Ordered Quantile
PCA	Principal Component Analysis
pdf	Función de Densidad de Probabilidad
RM	Risk Matrix (Matriz de Riesgo)
SA	Análisis de Sensibilidad
SS	Severity Score
SVD	Singular Value Descomposition
UA	Análisis de Incertidumbre
UAF	Unbalance-Adjusted Function

VS Violation Score

WAM Media Aritmética Ponderada

Capítulo 1

Introducción

La Comisión Nacional de Seguridad de Tránsito (CONASET) menciona que "durante el año 2021 se registraron 80.751 siniestros de tránsito y 1.688 personas perdieron la vida", cifras que tuvieron un aumento del 13,7 % respecto a los fallecidos informados el año 2020. Por tanto, para mejorar la seguridad vial, es imprescindible comprender el comportamiento de los conductores e influir en él [2], pues estos tienen gran impacto sobre todos los usuarios de las carreteras y calles. En particular, para el transporte comercial, el comportamiento riesgoso de los conductores tiene un impacto directo sobre la empresa u organización, como en los costos de combustible, costes de neumáticos y otros costes de explotación de la flota, así como su imagen corporativa [2]. En otras palabras, si no se gestionan los comportamientos riesgosos, las empresas se exponen a grandes riesgos y costes innecesarios.

Una de las principales formas de describir y comunicar el nivel de riesgo es a través de *índices de riesgo* [3] [12], que resumen el riesgo de un evento utilizando números o valores categóricos como palabras, letras o colores. Con el fin de identificar y comparar entre diferentes riesgos, entender como éste está cambiando a lo largo del tiempo, y apoyar la toma de decisiones [2] [3]. Los *índices*, o también llamados *indicadores compuestos* ([12], [13], [14], [15]) son usados en diferentes disciplinas y con diversos objetivos. Algunos ejemplos son el *Índice de Desarrollo Humano (HDI)*, el *Índice de Logros Tecnológicos* y el *Índice de Sostenibilidad Ambiental* [12].

Esencialmente, un *índice* puede describir un fenómeno global complejo formado por numerosas "componentes" [14] (información de varias dimensiones subyacentes [12]), con la

finalidad de facilitar su comprensión, y no interpretar cada una de estas "componentes" por separado [14] [12]. Hay muchas etapas en el proceso de construcción de un *índice* y si el procedimiento seguido no está claro y razonablemente justificado para todos los especialistas involucrados, la bibliografía relacionada de [14] (Grupp y Mogee 2004; Grupp y Schubert 2010), menciona que existe un margen considerable para la manipulación del resultado. Además, elecciones "incompatibles" o "ingenuas" (es decir, sin conocer las consecuencias reales) en las etapas de *ponderación* y *agregación* pueden dar lugar a una medida sintética "sin sentido"[14]; y por ende, las conclusiones extraídas serán erróneas. Por esta razón, a pesar de su popularidad creciente, los *índices* deben interpretarse con extrema cautela, debido a que su validez está intrínsecamente ligada a su construcción y, no hay ningún elemento en su construcción que esté por encima de la crítica [14]. Cada uno de los enfoques tiene sus ventajas y sus inconvenientes [14], e indenpendiente del objetivo del *índice*, las medidas agregadas deberían someterse a pruebas para comprobar su solidez en conjunto. Este proceso es una herramienta de "garantía de calidad" que ilustra lo sensible que es el *índice* a los cambios en los pasos seguidos para construirlo y reducirá en gran medida las posibilidades de transmitir un mensaje engañoso [14]. Por tanto, los *índices de riesgo* exigen un estudio cuidadoso de los *índices de riesgo* y de las mejores formas posibles para su construcción [3].

1.1. Revisión Bibliográfica

Los comportamientos riesgosos claves identificados en [2], inicialmente en base a la revisión de la literatura, son: consumo de alcohol y drogas, exceso de velocidad, cambios de carril frecuentes o rápidos, falta de señal, seguimiento demasiado de cerca, uso del cinturón de seguridad, fatiga del conductor, tomar las curvas bruscamente, frenar y acelerar de forma insegura, actitud del conductor, incumplimiento de las señales de tráfico, y distracción o falta de atención del conductor (por ejemplo, uso del teléfono móvil o pasajeros en el vehículo).

Se suele usar *índice de riesgo* y *medida de riesgo* como sinónimos. Sin embargo, en [3] hacen diferenciación entre éstos términos y ofrece recomendaciones para la construcción de un *índice de riesgo* para cualquier disciplina. Los tipos de *medida de riesgo* descritos son: momentos de una distribución de probabilidad, cuantiles de una distribución,

funciones de desutilidad, y funciones de indicadores (o factores). Dada la distinción entre *medida de riesgo* e *índice de riesgo*, una medida se puede considerar un *índice*. Sin embargo, transformar la escala de la medida puede facilitar la comprensión y su uso, por lo que la forma final del índice va ligada a los objetivos específicos y al público al que va dirigido. Por otra parte, el uso más popular para resumir el riesgo en forma de *índice* es la *Matriz de Riesgo (RM)* codificada por colores con métodos de puntuación aditiva. Método que no tiene en cuenta la incertidumbre de forma matemáticamente determinista, basándose en significados ambiguos de palabras como "muy frecuente" y "alto impacto", creando consecuencias imprevistas basadas en sistemas de puntuación y clasificación arbitrarios [3]. Por tanto, mediante el desarrollo de pasos concretos basados en los principios del análisis de decisiones y riesgos, [3] se esfuerza por alejarnos del uso de las *matrices de riesgo*.

Siguiendo con las *RM*s, existe numerosa literatura ([4], [5], [6], [7], y [8]) que enfatiza los problemas de usar este método. En [4] se examinan algunas propiedades matemáticas de la *RM* y señalan las siguientes limitaciones: pobre resolución pues solo pueden comparar correctamente y sin ambigüedades una pequeña fracción de pares de peligros seleccionados al azar, además de asignar calificaciones idénticas a riesgos cuantitativamente muy diferentes; asignación errónea de calificaciones cualitativamente más altas a riesgos cuantitativamente más pequeños; asignación subóptima de recursos; e inputs y outputs ambiguos, pues los inputs y outputs requieren una interpretación subjetiva, por ende, diferentes usuarios pueden obtener calificaciones opuestas de los mismos riesgos cuantitativos. Se enuncia, que una *RM*, como mínimo debe cumplir 3 axiomas y un teorema. En [5] resumen los defectos ya mencionados en [4] y además identifican otros no discutidos: ranking arbitrario, inestabilidad por la categorización, y distancia relativa distorsionada.

En [6] examinan la fiabilidad y utilidad de las *RM*s para clasificar los riesgos en el contexto de las actividades de ocio público, incluidos los viajes, donde nuevamente se constatan las distintas puntuaciones a los riesgos asignadas por distintos evaluadores. En [7] se propone un enfoque llamado enfoque de actualización secuencial (SUA por sus siglas en inglés) para diseñar *RM*s, debido a que la mayor parte de la literatura relacionada está centrada en la aplicación de una *RM*. Además, se propone un algoritmo para desarrollar el diseño en base a los 3 principios de Cox mencionados en [4].

1.2. Objetivos

1.2.1. Objetivo General

Como se puede observar, el modo en que se construye un *índice* para medir el riesgo de un evento es de suma importancia para lograr comunicar información certera, existiendo múltiples opciones como parte de este proceso. En consecuencia, esta Memoria de Título se define, en colaboración con Wisetrack Corp; empresa especialista en control, gestión y monitoreo de flotas; a fin de desarrollar/aplicar las mejoras necesarias a un índice en particular, a saber, **Índice de Conductor Riesgoso (ICR)** utilizando el software estadístico ***R project***. Teniendo como objetivo, mayor precisión y rigurosidad al momento de puntuar el riesgo durante los diferentes viajes, y tomar las medidas pertinentes o necesarias. Este índice, puntúa el riesgo de cada viaje realizado utilizando 7 comportamientos riesgosos: Aceleración, Desaceleración, Exceso de Velocidad, Salida de Carril, Colisión Frontal, Fatiga del Conductor, y Distracción del Conductor.

1.2.2. Objetivos Específicos

- Determinar el número adecuado de *indicadores individuales* considerando la correlación que puede existir entre ellos para la construcción de un índice apropiado.
- Usar *Métodos de Ponderación* para determinar ponderaciones que sean adecuadas para obtener un *índice* idóneo al objetivo general.
- Comparar el *overall scoring* entre *ICR* inicial y el *ICR* basado en las diferentes metodologías establecidas para su construcción/modificación, con la finalidad de obtener un *índice* certero y robusto.

Inicialmente, se realiza una revisión y comprensión del *ICR* usado y de los resultados que entrega, para luego realizar una investigación sobre las metodologías existentes, tanto como de sus ventajas y desventajas entre sí, así como de la compatibilidad entre ellas, pues cada etapa va "ensamblada" con las demás etapas existentes. En efecto, los indicadores usados deben ser capaces de captar el fenómeno multidimensional que se quiere representar con el *índice*; por tanto, para lograr el primer objetivo específico, un *análisis multivariado* de los datos e indicadores es fundamental, pues si los indicadores son en

extremo correlacionados, puede que alguno de ellos no sea necesario para la construcción del *índice*, o si los datos con los que se construyen los indicadores tienen valores atípicos en exceso, se puede ver perjudicado el desarrollo de construcción. Otro aspecto, es la *normalización*; determinar si es o no necesaria para el caso en particular, y cuál emplear. Por otra parte, para los *métodos de ponderación*, las ponderaciones derivadas de estos métodos pueden tener un significado diferente a "importancia"; por ende, previo al desarrollo de las metodologías es necesario seleccionar las más adecuadas o factibles, pues además varios de los métodos se pueden considerar subjetivos, porque son basados en el criterio de varios expertos y no en base a datos. Con las metodologías establecidas para esta etapa, se considera también, las modificaciones/adaptaciones efectuadas (como en *DEA*[1] y para *PCA*[2]. Finalmente, para los *esquemas de agregación*, existen diversos métodos que van desde una agregación *lineal*, hasta otros más sofisticados que tienen como objetivo evitar la *compensabilidad* existente entre los *indicadores*. Para esta etapa las elecciones son en base a los requisitos de los *esquemas de ponderación* usados (*lineal*) y otro método con *penalización* intermedia.

Este documento está organizado en 7 *Capítulos*. En el *Capítulo 1* se presentó una breve introducción al tema, junto a los objetivos. El *Capítulo 2* corresponde a los fundamentos básicos sobre el *riesgo* e *índices*. Seguidamente, en el *Capítulo 3*, se introduce el *índice* objetivo a mejorar, *ICR*. El *Capítulo 4*, contiene las etapas elementales para la construcción de un *índice* y las metodologías utilizadas en cada una de éstas, como los conceptos y métodos necesarios para mayor comprensión. El *Capítulo 5* contiene la introducción de los datos disponibles, así como el análisis de éstos. Para seguir en el *Capítulo 6* con la construcción/modificación del *índice* usando las metodologías establecidas previamente, además de un *Análisis de Sensibilidad* y *Robustez* al *índice sintético* obtenido, con el fin de obtener una cercioración sobre el sentido y robustez del mismo. Finalmente, el *Capítulo 7* presenta conclusiones e ideas a resaltar.

[1] Por sus siglas en inglés *Data Envelopment Analysis*.
[2] Por sus siglas en inglés *Principal Component Analysis*.

Capítulo 2

Fundamentos Teóricos sobre el Riesgo

En esta sección se presentan algunas definiciones claves relacionadas a la gestión del riesgo.

2.1. Riesgo

Definición 1. ***Riesgo***. Pérdida potencial, catástrofe u otro acontecimiento indeseable medido con probabilidades asignadas a pérdidas de diversas magnitudes [8].

El acontecimiento indeseable puede ser un concepto abstracto, como catástrofes naturales, la retirada de un producto importante, el impago de un deudor importante, la liberación de datos sensibles de clientes por parte de piratas informáticos, la inestabilidad política en torno a una oficina extranjera, accidentes laborales con resultado de lesiones o un virus de gripe pandémica que interrumpa las cadenas de suministro. También podría significar desgracias personales, como un accidente de coche de camino al trabajo, la pérdida de un empleo, un ataque al corazón, etc. Casi todo lo que puede salir mal, dentro de un contexto, es un *riesgo* [8].

Existen variadas definiciones de *riesgo*. Sin embargo, Hubbard en [8], menciona que la definición dada (Def. 1) representa mejor a la utilizada por tratamientos matemáticos

bien establecidos, así como también cualquier diccionario en inglés o incluso cómo el público no especializado utiliza el término.

La forma de mitigar los *riesgos* se basa en el coste de esas mitigaciones y en el efecto que se espera que tengan sobre los *riesgos*. En otras palabras, incluso la elección de las mitigaciones implica otra capa de *análisis de riesgos* [8]. No obstante, estos *riesgos*, independiente del ámbito, pueden revelarse sólo después de un desastre importante en una empresa, en un programa gubernamental o incluso en la vida personal. Por tanto, una medición errónea de estos riesgos acarrearía graves problemas [8].

2.2. Métodos de Evaluación de Riesgo

La mayoría de los métodos de *evaluación de riesgos* corresponden a *Métodos cualitativos* y a *RM*s, ya sea en la gestión de riesgo empresarial o áreas específicas de la gestión de riesgos, como la ciberseguridad y la gestión de riesgos de proyectos ([8], p. 174).

2.2.1. Métodos Cualitativos de puntuación

Los *Métodos cualitativos de puntuación* del *riesgo* son fáciles de elaborar y no requieren una formación especial ni siquiera una investigación previa. Cualquiera puede desarrollar su propio método de puntuación para casi cualquier atributo. Casi todos ellos utilizan algún tipo de la escala ordinal simple, es decir, una escala que indica un orden relativo de lo que se está evaluando, no unidades de medida reales. Pueden utilizar valores numéricos o simplemente etiquetas como “alto”, “medio” o “bajo” [8].

Existe un gran número de métodos de puntuación, pero todos ellos pueden agruparse en dos grandes categorías: las *puntuaciones ponderadas aditivas* y las *puntuaciones multiplicativas* [8]. Las *aditivas* pueden incluir varias escalas ordinales que pretenden ser *indicadores de riesgo*, que se ponderan y suman de alguna manera. Y en las *multiplicativas*, las diferentes puntuaciones se multiplican entre sí en lugar de ponderarse y sumarse. Entre las *puntuaciones multiplicativas*, se encuentra la *Matriz de Riesgo*.

Definición 2. ***Matriz de Riesgo (RM)***. [1] Es una representación gráfica de la *Probabilidad* (*posibilidad*, o *frecuencia*) de que se produzca un resultado y del *Impacto*

[1] Puede ser de tipo *Cualitativo*, *Semi-Cuantitativo*, y *Cuantivativo*. Para más detalles revisar [7].

(consecuencias) en caso de que se produzca dicho resultado [5]. *Probabilidad* tiene varias categorías para sus filas (o columnas) e *Impacto* tiene varias categorías para sus columnas (o filas, respectivamente). Asocia un nivel recomendado de *riesgo* (urgencia, prioridad o acción de gestión) con cada par fila-columna, es decir, con cada celda [4].

La Figura 2.1 ilustra una *RM semi-cuantivativa de* 5×5, ya que una matriz de este tipo es la usada en el *índice de conductor riesgoso* (índice a mejorar/modificar) para obtener las ponderaciones.

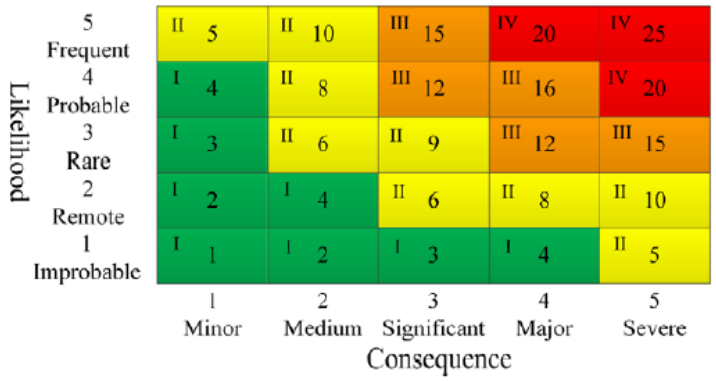

FIGURA 2.1: Categorización de riesgos en una *matriz de riesgos* semi-cuantitativa. (Fuente: How to Design Rating Schemes of Risk Matrices: A Sequential Updating Approach [7])

En el contexto de las *RM*s, esta multiplicación entre *Probabilidad* e *Impacto* da como resultado la *Pérdida Esperada* (EL[2]); término que se utiliza para hacer referencia a las consecuencias negativas esperadas asociadas riesgo [5].

2.2.2. Problemas con los Métodos Cualitativos de Puntuación

Como norma general, no es buena idea tratar las escalas ordinales como si fueran medidas de distancia o masa. En realidad, no suman ni multiplican como otras medidas. Aun así, casi todos los sistemas de puntuación utilizados en las *evaluaciones de riesgos* suman y multiplican valores en escalas ordinales [8].

Dado a que no se han desarrollado como resultado de una investigación rigurosa, ninguno de estos métodos de puntuación tiene en cuenta las cuestiones relativas a la percepción de los riesgos y las incertidumbres; no hay pruebas empíricas de que estos métodos mejoren

[2]Por su nombre en inglés *Expected Loss*.

las decisiones en absoluto. De hecho, ni siquiera se plantea la cuestión de si las decisiones mejoran de forma medible [8]. Además, las descripciones cualitativas de la *probabilidad* son entendidas y utilizadas de forma muy diferente por distintas personas, incluso cuando se toman medidas deliberadas para normalizar los significados ([8], p. 179). Finalmente, las correlaciones entre diferentes *riesgos* y diferentes factores es importante para todos los análisis de riesgos, pero universalmente ignorada en los modelos de puntuación [8]. Por tanto, las características arbitrarias de estos sistemas de puntuación añaden sus propias fuentes de error como resultado de las consecuencias imprevistas de su estructura [8].

Específicamente, para la *Matriz de Riesgo* han sido identificados varios problemas.

2.2.2.1. Problemas con la Matriz de Riesgo

En [4] se enumeran las siguientes limitaciones:

1. Poca resolución. Las *RM*s típicas sólo pueden comparar correctamente y sin ambigüedades una pequeña fracción (por ejemplo, menos del 10 %) de pares de peligros seleccionados al azar. Pueden asignar calificaciones idénticas a riesgos cuantitativamente muy diferentes ("compresión de rango"). El uso del mecanismo de puntuación integrado en las *RM*s comprime la gama de resultados y, por tanto, comunica erróneamente la magnitud relativa de las *Consecuencias* y las *Probabilidades*. El hecho de que la *RM* no consiga transmitir esta distinción parece socavar su beneficio, comúnmente declarado, de mejorar la comunicación [5].

2. Errores. Las *RM*s pueden asignar erróneamente calificaciones cualitativas más altas a riesgos cuantitativamente más pequeños. Para riesgos con *Frecuencias* y *Consecuencias* correlacionados negativamente; es decir, riesgos cuyo suceso tenga una *probabilidad* muy baja de ocurrir, pero *consecuencias* altas (o viceversa) puede dar resultados inútiles.

3. Inputs y Output ambiguos. Los inputs de las matrices de riesgo (por ejemplo, las categorizaciones de *Probabilidad* e *Impacto*) y los outputs resultantes (las calificaciones de riesgo) requieren una interpretación subjetiva, y diferentes usuarios pueden obtener calificaciones opuestas de los mismos riesgos cuantitativos. Las categorías de gravedad no pueden hacerse de forma objetiva para las *consecuencias* inciertas.

Tony Cox, el autor de [4], sugiere que para diseñar *RM*s y garantizar que la *pérdida esperada (EL)* en la región verde sea consistentemente más pequeña que la *EL* en la región roja, la *RM* debe cumplir con la *consistencia débil*, es decir, que los riesgos en la categoria cuantitativamente [3] mayor sean mayores que los riesgos de la categoría cualitativa inferior; *beetwenness*, es decir, que cada segmento de línea con pendiente positiva que se encuentre en una celda verde en su extremo inferior y en una celda roja en su extremo superior pase por al menos una celda intermedia; *coloración consistente*, es decir, que riesgos cuantitativamente iguales tengan igual color de casilla, y un teorema que establece que si se cumple la *consistencia débil*, *beetwenness*, y la *coloración consistente*, entonces todas las celdas de la fila inferior y más a la izquierda son verdes y que todas las celdas de la segunda columna desde la izquierda y de la segunda fila desde abajo son no rojas.

El propósito principal de la región amarilla es separar la región verde y la región roja en las *RM*s, no categorizar los resultados [4]. La *RM* es inconsistente si la *EL* en la región amarilla puede ser mayor que en cualquiera de las celdas rojas o menor que en cualquiera de las celdas verdes [4].

Existen otros fallos identificados, y no mencionados en [4], que se pueden revisar en [5].

2.3. Índices

Una de las principales formas de describir y comunicar el nivel de *riesgo* es a través de los *índices de riesgo* [3]. Estos *índices* se utilizan para comunicar los *riesgos* al público, entender cómo está cambiando el *riesgo* a lo largo del tiempo, comparar entre diferentes *riesgos* y apoyar la toma de decisiones [3]. Cada vez son más utilizados por las oficinas de estadística y las organizaciones nacionales o internacionales para transmitir información sobre el estado de los países en campos como el medio ambiente, la economía, la sociedad o el desarrollo tecnológico [15].

Definición 3. ***Índice de Riesgo***. Resume el *riesgo* de un evento o situación utilizando números o categorías como palabras, letras o colores, con el fin de identificar y comparar *riesgos* [3].

[3] Riesgo cuantitativo se define como el producto de *frecuencia* y *consecuencia* [4].

Un *índice de riesgo* generalmente intenta cumplir al menos uno de los siguientes objetivos [3]:

- Describir el *riesgo* de un evento con precisión.
- Comunicar el nivel de *riesgo*.
- Comparar entre diferentes *riesgos*.
- Identificar el *riesgo* más grave.
- Traducir cómo cambia el *riesgo* con el tiempo.
- Medir la eficacia de las estrategias de reducción de *riesgos*.
- Recomendar acciones para un nivel de *riesgo* dado.

Definición 4. ***Medida de Riesgo***. Resumen numérico del *riesgo* en un número real o vector de números reales [3]; como la media o un cuantil, o los dos primeros momentos de una distribución o tres cuantiles diferentes, en caso de ser un vector numérico [3].

Se considera que las *medidas de riesgo* son un subconjunto de los *índices de riesgo*, porque un *índice de riesgo* además de un número real (o vector de números reales) también puede resumir en letras, palabras o colores, ampliando así los medios de comunicación de *riesgo* al usuario [3].

Dado que un paso importante para construir un *índice de riesgo* es la medición precisa del mismo [3], a continuación se enumeran las *medidas de riesgo financiero* más conocidas, clasificadas en cuatro categorías generales: *momentos de la distribución de probabilidad de* X, *cuantiles de la distribución* X, *funciones de desutilidad*, y *combinación de Factores (o Indicadores)*; donde X es una variable aleatoria que representa las consecuencias de un evento cuyo riesgo interesa medir[4] [3]. La *función de Factores* consiste en la consideración de varios *factores* relacionados con el *riesgo* del evento, que a menudo se agregan mediante una combinación lineal ponderada [3].

A continuación se entregan dos definiciones del término *indicador*:

Definición 5. ***Indicador***. Correlato empírico de las variables que se intentan medir, son su expresión concreta, práctica, medible. Es decir, son las propiedades manifiestas

[4] Si $X < 0$ implica consecuencias negativas o adversas, por ejemplo, pérdida de dinero en una inversión, muertes derivadas de un suceso.

que se hallan empíricamente relacionadas con una propiedad latente o no observable de modo directo [9].

Definición 6. ***Indicador***. Puede definirse como una medida cuantitativa o cualitativa que se deduce de una serie de hechos observados para revelar las posiciones relativas de los objetos en un área determinada ([13] y [12]).

Los *indicadores* reciben este nombre porque indican o son indicios de otras variables más generales, y por ello de su existencia se puede inferir la concurrencia de dichas variables más abstractas de las que son signo y con las que están relacionadas (Sierra Bravo, 1989: 112), bibliografía relacionada de [9]. Una de las características útiles de un *indicador* es que puede representar grandes cantidades de información de forma sencilla [12]; además, pueden utilizarse para varios objetivos, como el seguimiento del rendimiento, la identificación de tendencias, la predicción de problemas, la evaluación del impacto de las políticas, la priorización de medidas, la evaluación comparativa, etc. (Litman, 2007; Sharpe, 2004), bibliografía relacionada de [12].

Definición 7. ***Índice (Indicador Compuesto o Índice sintético)***. Combinación de *subindicadores* (o *indicadores individuales* [13]) bien elegidos en uno solo [15].

Definición 8. ***Dimensión***. Cada uno de los diferentes aspectos que componen el concepto teórico que integra una multiplicidad de aspectos [9].

Definición 9. ***Polaridad***. Es el signo de la relación entre el *indicador* y el concepto que debe medirse. Los *indicadores individuales* con polaridades iguales deben estar correlacionados positivamente, mientras que los *indicadores individuales* con polaridades opuestas deben estar correlacionados negativamente [19].

Esencialmente, un *índice* podría reflejar un "sistema complejo" formado por numerosas "componentes" (información de varias *dimensiones* subyacentes [12]) [14], con el fin de facilitar la comprensión y no analizar cada "componente" de manera individual. Esta "complejidad" es una característica universal e interdisciplinar, (Rosen 1991), bibliografía relacionada de [14].

2.3.1. Ventajas y Desventajas

Los principales pros y contras de la utilización de indicadores compuestos (*índices*) se han debatido en los servicios de la Comisión Europea, Instituto para la Protección y

Seguridad del Ciudadano, Ispra, Italia. La discusión la resumen Saisana y Tarantola (2002), donde citan:

Pros:

- Los *indicadores compuestos* pueden utilizarse para resumir cuestiones complejas o multidimensionales, con vistas a apoyar a los responsables de la toma de decisiones.
- Los *indicadores compuestos* ofrecen una visión de conjunto. Pueden ser más fáciles de interpretar que intentar encontrar una tendencia en muchos *indicadores* separados. Desde el estudio en áreas complejas al nivel del mundo, facilitan la tarea de clasificar unidades en estas cuestiones complejas.
- Los *indicadores compuestos* pueden ayudar a atraer el interés del público al proporcionar una cifra resumida con la que comparar los resultados de los distintos países y sus progresos a lo largo del tiempo.
- Los *indicadores compuestos* pueden ayudar a reducir el tamaño de una lista de indicadores o a incluir más información dentro del límite de tamaño existente.

Contras:

- Los *indicadores compuestos* pueden enviar mensajes políticos engañosos y poco sólidos si están mal construidos o mal interpretados. El *Análisis de Sensibilidad (SA)* puede servir para comprobar la solidez de los *indicadores compuestos*.
- Los resultados simples "a grandes rasgos" que muestran los *indicadores compuestos* pueden invitar a los políticos a extraer conclusiones políticas simplistas. Los *indicadores compuestos* deben utilizarse en combinación con los *subindicadores* para extraer conclusiones políticas sofisticadas.
- La construcción de *indicadores compuestos* implica etapas en las que hay que hacer juicios de valor: la selección de *subindicadores*, la elección del modelo, la ponderación de los *indicadores* y el tratamiento de los valores que faltan, etc. Estos juicios deben ser transparentes y basarse en principios estadísticos sólidos.
- Podría haber mayor flexibilidad y discreción para los Estados Miembros, países o naciones que son miembros de la Unión Europea, al seleccionar y aplicar *indicadores compuestos*, a comparación que con *indicadores individuales*. Sin embargo, la

selección de los *indicadores individuales* que componen los *indicadores compuestos* pueden ser objeto de discusiones y debates políticos, ya que la elección de *subindicadores* y las ponderaciones asignadas pueden tener implicaciones en la manera en que se evalúa y se comparan entidades o situaciones.

- Los *indicadores compuestos* aumentan la cantidad de datos necesarios porque se requieren datos para todos los *subindicadores* y para un análisis estadísticamente significativo (es decir, poco probable que el resultado sea fruto del azar).

Capítulo 3

Índice de Conductor Riesgoso

El *Índice de Conductor Riesgoso (ICR)*, usado por Wisetrack Corp., tiene por objetivo determinar qué tan riesgoso es un viaje. A continuación se esbozan los detalles básicos, extraídos de [1], para la compresión de este *índice*.

El *Índice de Conductor Riesgoso (ICR)* se obtiene en función de las infracciones (alertas) descritas en la Tabla 3.1:

Tabla 3.1: Tipo Alertas

i	Alerta
1	Aceleracion
2	Posible Colisión Frontal
3	Distración
4	Fatiga
5	Salida de Carril
6	Exceso Velocidad Máxima
7	Desaceleración

3.1. Score

La puntación de un viaje se basa en los eventos adversos (infracciones) originados por los sistemas avanzados de ayuda a la conducción (sistema ADAS) durante el transcurso del viaje. La generación de los eventos se produce cuando un valor asociado medido

alcanza y supera un valor de activación preestablecido. Los valores de activación para la aceleración, frenado y exceso de velocidad se ilustran en la Tabla 3.2.

Tabla 3.2: Valores preestablecidos de activación

Tipo Vehículo	Aceleración (km/hr/s)	Desaceleración (km/hr/s)	Exceso Velocidad (km/hr)
Liviano	12	15	120
Pesado	10	12	85

3.1.1. Descripción score

Las puntuaciones se comunican en forma de porcentaje y por viaje, siendo la puntuación de 100 % una puntuación perfecta. Las infracciones para un viaje se puntúan en función de la duración, severidad y peso.

A continuación, se entregan definiciones esenciales para el cálculo del score (puntuaciones):

1. **Weight Factor (WF_i)**: Ponderación que se le asigna a la infracción $(0 - 99)$. Se obtiene de una *Matriz de Riesgo* de 5×5 según qué tan riesgoso se considere el tipo de la infracción ($Impacto \times Probabilidad$).
2. **Duration Factor (DF_i)**: Peso que se le asigna a la duración de la infracción (0 % - 100 %).
3. **Severity Factor (SF_i)**: Indica el efecto que tiene la severidad en la infracción (100 %- 112 %).
4. **Driver Scoring**: valor numérico obtenido del ICR.

Donde $i = 1, 2, 3, 4, 5, 6, 7$. Es decir, para cada tipo de Alerta.

3.1.2. Cálculo del score

Se calculan los indicadores *Duration Score (DS)* y *Severity Score (SS)* por tipo de infracción registrada durante el viaje, tal como se ilustra a continuación:

3.1.2.1. Duration Score

1. Para $i = 1, 2, 3, 6, 7$:

$$DS_i = \begin{cases} 0 & , \quad si \quad timeinfr_i \geq DF_i \cdot timetot \\ \left(1 - \dfrac{timeinfr_i}{timetot}\right) \cdot 100, & e.o.c \end{cases} \tag{3.1}$$

2. Para $i = 4$ (*Fatiga*):

$$DS_4 = \begin{cases} 100 - 50 \cdot n_4, & si \quad n_4 < 2 \\ 0 & , \quad e.o.c \end{cases} \tag{3.2}$$

3. Para $i = 5$ (*Salida de Carril*):

$$DS_5 = \begin{cases} 100 - 20 \cdot n_5, & si \quad n_5 < 5 \\ 0 & , \quad e.o.c \end{cases} \tag{3.3}$$

donde $timeinfr_i$ corresponde al tiempo total de la infracción i en minutos ocurridas durante el viaje, $timetot$ corresponde al tiempo que dura el viaje (minutos), n_j corresponde al número de ocurrencias de alertas del tipo $j = 4, 5$ durante el viaje.

3.1.2.2. Severity Score

1. Para $i = 1, 7$ (*Aceleración* y *Desaceleración*):

$$SS_i = \begin{cases} 0 & , \quad si \quad a_{infr_i} \geq SF_i \cdot a_{p_i} \\ \left(1 - \dfrac{a_{infr_i} - a_p}{SF_i \cdot a_{p_i} - a_{p_i}}\right) \cdot 100, & e.o.c \end{cases} \tag{3.4}$$

2. Para $i = 2, 3, 4, 5$ (*Posible Colisión Frontal, Distracción, Fatiga*, y *Salida de Carril*):

$$SS_i = 0 \tag{3.5}$$

3. Para $i = 6$ (*Exceso de Velocidad*):

$$SS_6 = \begin{cases} 0 & , \quad si \quad v_{infr} \geq SF_6 \cdot v_p \\ \left(1 - \dfrac{v_{infr} - v_p}{SF_6 \cdot v_p - v_p}\right) \cdot 100, & e.o.c \end{cases} \tag{3.6}$$

donde a_{infr} y a_p corresponde a la aceleración de la infracción realizada, y a la aceleración permitida, respectivamente. Y v_{infr} y v_p corresponde a la velocidad de la infracción, y a la velocidad permitida, respectivamente (Tabla 3.2).

3.1.2.3. Violation Score

Para $i = 1, 2, 3, 4, 5, 6, 7$:

$$VS_i = \begin{cases} DS_i & , \quad si \quad SS_i = 0 \\ SS_i & , \quad si \quad DS_i = 0 \\ \dfrac{DS_i + SS_i}{2} & , \quad e.o.c \end{cases} \tag{3.7}$$

3.1.2.4. Score

$$Score = \sum_{i=1}^{7} WF_i \cdot VS_i \tag{3.8}$$

Sin embargo, en la práctica, solo se obtienen los indicadores DS_i, $i = 1, 2, 3, 4, 5, 6, 7$, dado a que la variable a_{infr} no es accesible, y solo se registra la infracción. Siendo obtenido el score final de la suma ponderada de DS_i, y no con VS_i.

Capítulo 4

Marco Teórico

4.1. Construcción de Índices

La calidad de un *índice*, así como la solidez de los mensajes que transmite, dependen no sólo de la metodología utilizada en su construcción, sino principalmente de la calidad del marco teórico y de los datos utilizados [13].

Cada paso es extremadamente importante, pero la coherencia en todo el proceso es igualmente vital. Las elecciones hechas en un paso pueden tener implicaciones importantes para los demás; por lo tanto, el constructor de *índices* no solo tiene que tomar las decisiones metodológicas más apropiadas en cada paso, sino también identificar si encajan bien entre sí [13].

Las etapas de construcción se ilustran en la Figura 4.1 y se describen a continuación:

1. ***Marco Teórico***. Es el punto de partida para construir *índices*; por ende, debe definir claramente el fenómeno que se va a medir y sus subcomponentes, seleccionando *indicadores individuales* y *ponderaciones* que reflejen su importancia relativa y las dimensiones del compuesto global. Lo ideal es que este proceso se base en lo que es deseable medir y no en los indicadores disponibles [13]. Por otra parte, debido a que no todos los conceptos multidimensionales tienen una base teórica y empírica tan sólida, es recomendable primeramente [13]: definir el concepto, determinar subgrupos, e identificar los criterios de selección de los indicadores subyacentes.

2. ***Selección de Datos***. Los puntos fuertes y débiles de los *índices* se derivan en gran medida de la calidad de las variables subyacentes. Lo ideal es seleccionar las variables en función de su pertinencia, solidez analítica, actualidad, accesibilidad, etc. [13]. No obstante, el proceso de selección de datos puede ser bastante subjetivo, ya que puede no haber un único conjunto definitivo de indicadores. La falta de datos relevantes también puede limitar la capacidad para construir indicadores compuestos sólidos. Dada la escasez de datos cuantitativos (duros) comparables a nivel internacional, los indicadores compuestos suelen incluir datos cualitativos (blandos) procedentes de encuestas o revisiones de políticas [13].

   El *marco teórico* no afecta solo a la pertinencia del *índice*, sino también a su credibilidad e interpretabilidad. La *imputación de datos*, afecta a la precisión y a la credibilidad; es más, un uso excesivo de las técnicas de imputación puede socavar la calidad general del *índice*. La etapa de *normalización* es crucial tanto para la precisión como para la coherencia de los resultados finales, ya que un procedimiento de normalización inadecuado puede dar lugar a resultados poco fiables o sesgados. La calidad de los *indicadores elegidos* para construir el índice afecta en gran medida a su precisión y credibilidad. En consecuencia, el uso del *Análisis Multivariante* para identificar la estructura de los datos puede aumentar tanto la precisión como la interpretabilidad de los resultados finales; siendo fundamental para identificar redundancias entre los fenómenos seleccionados y evaluar posibles lagunas en los datos básicos. La etapa correspondiente a la elección de esquemas de *ponderación* y *agregación*, es un aspecto clave en la construcción del *índice*, dado que casi todas las dimensiones de la calidad se ven afectadas por esta elección, especialmente la precisión, la coherencia y la interpretabilidad, por tanto, se debe prestar especial atención para evitar contradicciones internas y errores al ponderar y agregar los indicadores individuales. Para más detalles revisar ([13], p. 46).

3. ***Imputación de Missing Values***. Los valores perdidos, de forma aleatoria o no aleatoria, suelen dificultar la elaboración de índices sólidos.

   Existen tres métodos generales para tratar este problema: *eliminación de casos*, *imputación única*, e *imputación múltiple* [13]. Sin embargo, no se detallarán aquí, porque en este trabajo no se usó ningún método de imputación. Para ver detalladamente los métodos mencionados se puede recurrir a [13].

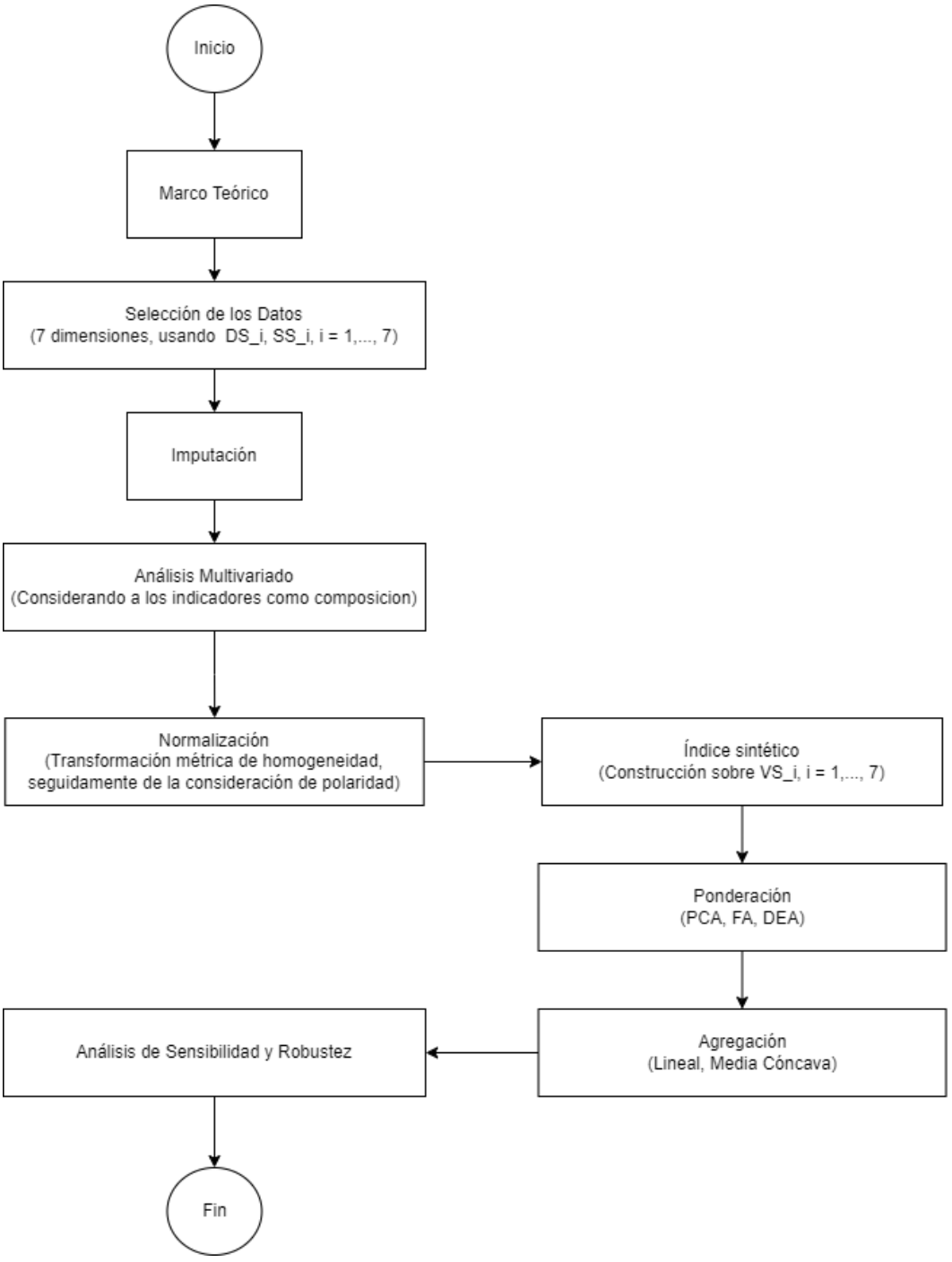

FIGURA 4.1: Diagrama de flujo para las etapas de construcción de un *índice*. En paréntesis se indica las opciones metodológicas usadas para el *ICR*

Las etapas siguientes de construcción, *Análisis Multivariado*, *normalización*, *ponderación* y *agregación*, así como los *Análisis de Sensibilidad y Robustez*, se tratan de manera más detallada, debido a su extensión.

4.1.1. Análisis Multivariado

Los *indicadores individuales* se seleccionan a veces de forma arbitraria, prestando poca atención a las interrelaciones entre ellos. Esto puede dar lugar a *índices* que abruman, confunden y engañan a los responsables de la toma de decisiones y al público en general ("rico en indicadores pero pobre en información") [13]. La naturaleza subyacente de los datos debe analizarse cuidadosamente antes de construir un *índice*. Este paso preliminar es útil para evaluar la idoneidad del conjunto de datos y permitirá comprender las implicaciones de las opciones metodológicas, como la *Ponderación* y la *Agregación*, durante la fase de construcción del *índice (indicador compuesto)* [13]. La información puede agruparse y analizarse a lo largo de al menos dos dimensiones del conjunto de datos: *indicadores individuales* y unidades [13].

En esta etapa, para el *Análiis Multivariado* de los indicadores se usa *PCA* sobre ellos, pero considerando su cáracter composicional (Sec. 4.3). La razón de esto, es porque los indicadores individuales (DS_i, SS_i, $i = 1, \ldots, 7$) usados para el *ICR*, están expresados en porcentajes, y en conjunto evalúan qué tan riesgoso es un viaje; así, cada *indicador* es una componente del riesgo durante la conducción, y este carácter *composicional* se debe tener en cuenta en esta etapa.

4.1.2. Normalización

A menudo los *indicadores* de un conjunto de datos suelen tener diferentes unidades de medida y/o polaridades diferentes, por ende, es necesaria la normalización de éstos antes de cualquier *agregación* de datos, pues así se pretende que éstos sean comparables [13] [19]. Los indicadores normalizados se calculan transformando los indicadores individuales en números puros, sin dimensiones, con *polaridad* positiva (Def. 9) [19].

Existen diferentes métodos, como *ranking*, *estandarización* (o *z-scores*), *distancia a una medida de referencia*, *escalas categóricas*, *indicadores por sobre o bajo la media*, mín-máx, etc. Todas estas descritas en [13]. Cada una de éstas, puede ser más adecuada que otra, según las propiedades de los datos, así como los objetivos del indicador compuesto.

Otra transformación de los datos, utilizada a menudo para reducir la asimetría de los datos (positivos), es la *transformación logarítmica*. Cuando el rango de valores del indicador es amplio o está sesgado positivamente, la *transformación logarítmica* reduce el lado derecho de la distribución.

La secuencia de transformaciones seguidas en esta Memoria, son las siguientes [23]:

1. Para mejorar la homogeneidad métrica, es decir, para que los incrementos aditivos similares de los valores transformados independientemente de su nivel inicial tengan significados similares para el aspecto descrito por el indicador. Si la variable (indicador) solo puede tomar valores entre dos constantes fijas l y L, normalmente un porcentaje es de este tipo con $l = 0$ y $L = 100$, entonces:

$$x_i = \begin{cases} \log\left(\dfrac{x_i - l + d}{L - x_i}\right), & si\ x_i = l \\ \log\left(\dfrac{x_i - l}{L - x_i + d}\right), & si\ x_i = L \\ \log\left(\dfrac{x_i - l}{L - x_i}\right) \end{cases} \tag{4.1}$$

   donde x_i corresponde a la unidad i del indicador que se está transformando, y $d = 0.001$ fijado arbitrariamente para evitar los infinitos.
2. Sobre el resultado de la operación anterior se realiza una inversión de signo (se multiplica por -1) si se considera que el indicador tiene una *correlación negativa ex ante* con el concepto que se está midiendo.
3. Se puede realizar una transformación (globalmente) facultativa, que tiene por objetivo "enderezar" la función cuantil de la variable, con el fin de atenuar sus dos colas y el efecto de los valores atípicos en los próximos procesos de síntesis. Se aplica a la variable transformada, según las indicaciones anteriores, una que haga que la función cuantil del resultado se parezca más a una función afín (línea recta); o dicho de otro modo, que minimice el cociente de la suma de cuadrados residual (la menor suma entre esa función cuantil y una línea recta) sobre la suma de cuadrados de regresión (la suma de cuadrados de las diferencias de la función cuantil con su media). Para ello, se aplica la función afín $f(t) = \dfrac{t - \min t}{\max t - \min t}$, de la función rectilínea que mejor se ajusta, a la función cuantil de la variable.

4.1.3. Ponderación y Agregación

Existe una bibliografía mucho más amplia sobre los *Métodos de agregación* que sobre los *Métodos de Ponderación*. No obstante, tanto los *Métodos de Ponderación*, como los *Métodos de Agregación* son controversiales [14] [13]. Con respecto a la *Ponderación*, la razón es porque según el valor asignado a ciertos indicadores, se obtiene una clasificación, pero al variar la ponderación, la clasificación obtenida de las unidades evaluadas puede cambiar [14]. Con respecto a la Agregación, las ponderaciones pueden tener el significado de *Compensatorias* o *Coeficientes de Importancia*, desprendiéndose de aquí la importancia de usar métodos compatibles, ya que ambas etapas están relacionadas y entrelazadas.

Esto supone un enorme reto en la construcción de un *índice*, a menudo denominado el "problema del índice" (Rawls 1971). Básicamente, aunque se llegue a un acuerdo sobre los *indicadores* que se van a utilizar; lo siguiente, y lo más "pernicioso" (Freudenberg 2003), es cómo puede lograrse un esquema de ponderación. La literatura intenta resolver este rompecabezas de varias maneras [14].

4.1.3.1. Ponderación

El significado de la *Ponderación* en la construcción de *índices* (indicadores compuestos) es doble: "importancia explícita" e "importancia implícita" [13] [14]. Para la primera, una ponderación puede considerarse como una especie de coeficiente que se asigna a un criterio, mostrando su importancia en relación con el resto de los criterios. Mientras, que el segundo significado, se refiere a la importancia implícita de los atributos, tal y como muestra el "equilibrio" entre los pares de criterios en un proceso de agregación [14].

Los diferentes métodos podrían clasificarse en "subjetivos" y "objetivos" [13] [14]. Debido a que hay esquemas de ponderación participativos, donde el desarrollador o el grupo de expertos pondera, y luego según el enfoque elegido se llega a un consenso [14]. Para esta Memoria, no se tratan estos métodos. Sin embargo, a continuación se enumeran los métodos de este tipo: *Budget Allocation*[1] *(BA)*, *Analytic Hierchy Processes (AHP)*, y *Conjoint Analysis (CA)* (todos descritos en [14]). Por otra parte, los métodos "objetivos" surgen de los propios datos bajo una función matemática específica, argumentándose

[1] Descrito e implementado en [12].

que estos métodos no sufren problemas de manipulación de resultados y no dependen del criterio subjetivo de los responsables de tomar las decisiones. No obtante, también tienen crítica [14]. Estos métodos son: *Análisis de Componentes Principales (PCA)*, *Factor Analysis (FA)*, *Data Envelopment Analysis (DEA)*, *Regresión lineal Múltiple*, y *Análisis de Correlación*. No obstante, los Métodos tratados aquí excluyen la *Regresión lineal Multiple* y el *Análisis de correlación* (para ponderar). Otro método existente es la *Igual Ponderación*[2] (*EW*), que asigna ponderaciones iguales a todos los indicadores individuales; sin embargo, para más detalles se puede revisar [14] y [13].

4.1.3.1.1 Análisis de Componentes Principales (PCA)

Este método multivariado sigue siendo uno de los más utilizados en la literatura empírica, posiblemente por su sencillez [18].

Si el conjunto tiene Q variables, $x_1, \ldots, x_Q$, gran parte de la variación de los datos puede explicarse a menudo con $P < Q$ variables $Z_1, Z_2, \ldots, Z_Q$ (*Componentes Principales* o *combinaciones lineales* de las variables originales) no correlacionadas:

$$Z_j = a_j' x, \quad j = 1, \ldots, P, \ldots, Q \tag{4.2}$$

donde x denota al vector $Q \times 1$ de variables observables, y a_j se consigue de manera a tener mínima varianza.

La primera *Componente Principal (CP)* capta la mayor fracción de la varianza de las variables originales; la segunda explica la mayor parte de la varianza restante, y así sucesivamente.

Sea $\mathbf{\Sigma}$ la *Matriz de Covarianza* de x, λ_j, $j = 1, \ldots, Q$ los *valores propios* de $\mathbf{\Sigma}$, y a_j , $j = 1, \ldots, Q$ los *vectores propios* correspondientes. Dado que $\mathbf{\Sigma}$ es simétrica y definida positiva, se tienen $\boldsymbol{A}$ y $\mathbf{\Theta}$, donde $\boldsymbol{A} = [a_1, \ldots, a_Q]$ y $\mathbf{\Theta} = diag(\lambda_j)$, $j = 1, ..., Q$ con los λ_j's ordenados en orden decreciente según su valor. Se tiene $\mathbf{\Sigma}^{-1} = \boldsymbol{A}\mathbf{\Theta}^{-1}\boldsymbol{A}'$. La varianza de cada una de las *CP*'s es igual al valor propio correspondiente, es decir $V(Z_j) = \lambda_j$, $\forall j$.

[2] Igual Ponderación no significa sin ponderación. Revisar referencias para más detalles.

Los elementos de los *vectores propios* de la matriz de covarianza se llaman *factor loadings*, *pesos* o *componentes* y cumplen con $a_{j_1}^2 + a_{j_2}^2 + \ldots + a_{j_Q}^2 = 1$, $j = 1, \ldots, Q$.

Para los indicadores derivados de las *CP*, hay dos métodos más utilizados y los presentados aquí están descritos en base a [18]:

1. Tomar la primera *CP* (la que corresponde al mayor valor propio λ_j) $Z_1 = a_1' x$ como un índice agregado. Entonces $V(Z_1) = \lambda_1$.

2. Una media ponderada de todas las *CP*'s, en el que las ponderaciones w_j vienen dadas por la proporción de la varianza total explicada por cada *CP*. Es decir,

$$\hat{\boldsymbol{H}} = \sum_{j=1}^{Q} w_j Z_j \tag{4.3}$$

donde $w_j = \dfrac{\lambda_j}{\sum_{j=1}^{Q} \lambda_j}$. Entonces, usando que $V(Z_j) = \lambda_j$, su varianza se puede escribir como $V(\hat{\boldsymbol{H}}) = \omega' \boldsymbol{\Theta} \omega$, donde $\omega' = [w_1, \ldots, w_Q]$.

4.1.3.1.1.1. Ventajas y Desventajas

El *PCA* tiene una serie de excelentes propiedades matemáticas, las que se enumeran a continuación:

1. La propiedad más importante es que el índice obtenido a partir de la primera *CP* explica la mayor parte de la varianza total de los indicadores individuales. Por tanto, el primer factor estará correlacionado con al menos algunos de los indicadores individuales, generalmente, la mayoría de ellos [19].

2. Computacionalmente es simple [19] [18].

3. Dado que los resultados se definen en función del conjunto de datos y no apriori, el *PCA* puede considerarse una herramienta de análisis de datos adaptable (Jolliffe y Cadima 2016, referencia relacionada de [19]).

Sin embargo, que la naturaleza de este enfoque se base en las propiedades estadísticas de los datos, puede considerarse tanto una ventaja como una desventaja. A continuación se enumeran algunas de estas desventajas:

1. La primera *CP* representa una parte limitada de la varianza de los datos, por lo que se puede perder una cantidad consistente de información [19] [18].

2. El *índice* basado en *PCA* puede ser "elitista". Con una fuerte tendencia a representar los indicadores altamente intercorrelacionados y a descuidar los demás, independientemente de su posible importancia contextual. En consecuencia, muchos indicadores de gran importancia pero escasamente intercorrelacionados pueden quedar sin representación en el índice compuesto [19].

3. Método ciegamente empirista, basado en las correlaciones observadas e ignora la polaridad de los indicadores individuales. Por tanto, si los indicadores normalizados no están todos positivamente intercorrelacionados, el índice basado en *PCA* no es correcto, ya que los indicadores individuales se resumen sin considerar las polaridades adecuadas [19].

4. El significado de las ponderaciones es claro desde un punto de vista matemático, pero tiene poco sentido en relación con el objetivo a medir. Esto, porque los factores encontrados por el *PCA* son dimensiones empíricas (basadas en la variabilidad), y no son dimensiones teóricas (basadas en un marco conceptual). Las dimensiones empíricas y la teórica a menudo no coinciden, lo que podría dificultar la asignación de un significado claro a los factores [19].

5. El *PCA* no permite realizar comparaciones interespaciales (para diferente grupos de unidades) o intertemporales (para distintos momentos), ya que la cantidad de varianza contabilizada y las ponderaciones calculadas por *PCA* cambian para cada matriz de datos, y entonces los resultados de diferentes análisis no son fácilmente comparables.

6. El *PCA* puede ser poco robusto y muy sensible a la inclusión o exclusión de un indicador individual. Cuanto menor sea la correlación del indicador con los demás, la solidez de los resultados es menor.

7. El uso de *PCA/FA* implica los supuestos de tener indicadores continuos y relacionados linealmente entre ellos [14].

8. Dado que las ponderaciones se asignan endógenamente, no corresponden necesariamente a los vínculos reales entre los indicadores (en particular los estadísticos) [14].

4.1.3.1.2 Análisis Factorial

El *Análisis Factorial (FA)* suele utilizarse para reducir las dimensiones de un problema. Existen varias directrices para evaluar el número óptimo de factores al que puede reducirse el problema, y una representación gráfica puede ser útil [12]. Considerando p factores (p valor a decidir), el siguiente paso es la rotación para mejorar la interpretabilidad. Esto resulta, en que cada indicador (i) tiene una gran puntuación factorial (a_{ij}) sólo en uno de los factores (j). Las ponderaciones de los indicadores pueden deducirse de estas cargas factoriales rotadas mediante un cálculo relativamente limitado.

El siguiente procedimiento está basado en [12]: Sean a_{ij}, $i = 1, \ldots, l$; $j = 1, \ldots, p$ los factores rotados. Definiendo $u_{ij} = \dfrac{a_{ij}^2}{\sum_{m=1}^{l} \sum_{n=1}^{p} a_{mn}^2}$, el peso preliminar del indicador i, $u_i = \max_j u_{ij}$. Sin embargo, $U = \sum_i u_i < 1$, debido a la reducción del problema, dejando una pequeña parte de la varianza sin explicar. El peso final para cada indicador i es igual a $w_i = \dfrac{u_i}{U}$. Por construcción, $W = \sum_i w_i = 1$.

4.1.3.1.2.2. Ventajas y Desventajas

Entre sus ventajas se encuentra que puede resumir un conjunto de indicadores individuales conservando la máxima proporción posible de la variación total del conjunto de datos original [13].

Desventajas:

1. El inconveniente más importante es que las ponderaciones se basan en correlaciones que no corresponden necesariamente a los vínculos reales entre los fenómenos que se miden [12].

2. Deducir las ponderaciones a partir del análisis factorial requiere un cierto nivel de correlación (para reducir el problema en un número de factores), una selección justificada del número óptimo de factores (ya que las ponderaciones dependen del número de factores elegido) y unos resultados de rotación claros (porque sólo se utilizan las cargas factoriales rotadas más altas en el cálculo de las ponderaciones) [12].

4.1.3.1.3 Análisis Envolvente de los Datos

Análisis Envolvente de los Datos (DEA), desarrollado por Charnes (1978), es una técnica de medición del rendimiento que puede utilizarse para evaluar la eficiencia relativa de las *unidades de toma de decisiones (DMU)*. Para cada *DMU*, la *eficiencia* se define como la relación entre la suma ponderada de los productos y la suma ponderada de los insumos. De este modo, se determina un conjunto de ponderaciones que dan como resultado la mejor puntuación posible para la unidad. Esto implica que las dimensiones en las que esta unidad obtiene unos resultados relativamente buenos tienen una mayor ponderación. Trasladando el *DEA* original al contexto de indicador compuesto, los *inputs* no se consideran y cada indicador es un *output*. En [20], se considera al valor bruto y_{ij} del indicador j, $j = 1, \ldots, l$ para la unidad i, $i = 1, \ldots, n$. Por tanto, el problema de maximización restringida puede escribirse[3]:

$$
\begin{aligned}
CI_i \quad = \quad & \text{máx} \sum_{j=1}^{l} y_{ij} w_j \\
& \textit{sujeto a:} \\
& \sum_{j=1}^{l} y_{kj} w_j \leq 100 \quad \forall k = 1, \ldots, n \\
& w_j \geq 0 \quad \forall j = 1, \ldots, l
\end{aligned}
\tag{4.4}
$$

Esta metodología es una aplicación del *DEA*, llamada *Beneficio de la Duda (BoD)* [4].

Todas las ponderaciones son generadas por el propio modelo (Ec. 4.4) y no se imponen restricciones exógenas a las ponderaciones. No obstante, es posible que las ponderaciones de algunos indicadores sean iguales a cero, y no sean considerados finalmente en la agregación; también es posible que exista información adicional que se puede incorporar a las ponderaciones. Por ende, se describe el siguiente modelo usado en esta Memoria, el cual impone límite superior e inferior para la contribución[5] de los indicadores individuales, con la finalidad de evitar las ponderaciones cero y evitar que uno o varios indicadores

[3] El modelo original restringe la suma ponderada a 1, y no a 100. Se hizo así para mantener concordancia con los datos.

[4] Por sus siglas en Ingles *Benefit of the Doubt*.

[5] La condición agregada corresponde a una restricción proporcional sobre la cuota de los indicadores, propuesta por Wong y Beasley (1990), y descrita en [20], donde además se mencionan otro tipo de restricciones adicionales para los indicadores individuales, y que "la plena libertad de los resultados, pueden contradecir las opiniones previas sobre las ponderaciones", y que por tanto, "se debe garantizar el establecimiento de un esquema de ponderación adecuado".

puedan tener una ponderación extremadamente alta:

$$\begin{aligned} CI_i \quad = \quad & \text{máx} \sum_{j=1}^{l} y_{ij} w_j \\ & \textit{sujeto a:} \\ & \sum_{j=1}^{l} y_{kj} w_j \leq 100 \quad \forall k = 1, \ldots, n \\ & L_j \leq \frac{w_j y_{ij}}{\sum_{j=1}^{l} y_{ij} w_j} \leq U_j \quad \forall j = 1, \ldots, l \\ & w_j \geq 0 \quad \forall j = 1, \ldots, l \end{aligned} \tag{4.5}$$

La construcción resultante del indicador compuesto sigue siendo invariable con respecto a las unidades de medida [20].

4.1.3.1.3.3. Ventajas y Desventajas

Como ventajas se tienen las siguientes:

1. Las ponderaciones se determinan de manera endógena y se derivan de los datos [12].

2. Al tener cada unidad una ponderación específica, estas ponderaciones se encuentran con el fin de que el valor de su índice final sea lo más alto posible [20]. Básicamente, se concede a cada unidad el beneficio de la duda a la hora de asignar ponderaciones dado que se desconocen las ponderaciones específicas [20].

3. Si se dispone de información adicional sobre las ponderaciones, éstas pueden agregarse al problema [20].

4. Una característica importante del *DEA*, y por tanto, también de *BoD*, es su invarianza unitaria, pues el índice resultante es independiente de las unidades de medida de los indicadores individuales [6] [20].

Y como desventajas:

[6] La razón fundamental de esta invarianza se remonta a la característica de que los pesos son endógenos. En (Cooper 2000, p.39), bibliografía relacionada de [20], se proporcionan detalles de esta afirmación.

1. De la no normalización de los indicadores, se desprende el hecho de que las ponderaciones no suman uno, lo que hace poco práctica la comparación de las ponderaciones con otros métodos de ponderación [12].

2. Que entre las restricciones solo se tengan que los pesos sean positivos y que el índice no supere la unidad (100 en este caso), se permite estimar libremente las ponderaciones para maximizar la puntuación. Sin embargo, también existe la posibilidad de que algunos indicadores obtengan ponderación con valor cero. Y por tanto, el índice está siendo evaluado solo con una fracción de los indicadores y su rendimiento final no es tan perfecto como lo parece [20].

3. Asume completa *compensabilidad* entre los indicadores dada su naturaleza de agregación *lineal* [14].

4. Sensible a valores atípicos [14].

Tal como se aprecia, no hay sistema de ponderación exento de críticas. Cada uno tiene sus ventajas e inconvenientes, y no hay ninguno que sea mejor que otros o una solución única sin importar el método. En consecuencia, hay que elegir el sistema de ponderación que mejor se adapte al propósito definido [14].

4.1.3.2. Agregación

En esta etapa se elige el procedimiento de agregación que producirá, para cualquier unidad dada, el valor del *índice* en función del vector $\boldsymbol{z} = (z_1, \ldots, z_n)$ de los *indicadores individuales* [22].

Los *métodos de agregación* se pueden dividir en *Lineal*, *Geométrica*, y *Multicriterio* [13]. Sin embargo, la agregación *lineal* y *geométrica* se incluyen dentro del *Análisis Multicriterio*. Otra forma de clasificar los métodos de agregación, es en *Compensatorios* y *No Compensatorios* [14] [22]). Donde, la *compensabilidad* entre las variables se define como la posibilidad de compensación de cualquier déficit en una *dimensión* (Def. 8) con un excedente adecuado en otro [22]. Según esta clasificación, la aproximación *multicriterio*[7] es un enfoque no compensatorio/no lineal. En esta Memoria se trata una función de agregación que es un caso intermedio de *compensación* y *no compensación*.

[7]Para más detalles se puede revisar [21][14].

A continuación, se presentan dos sistemas de agregación; cada uno corresponde al enfoque *compensatorio* y *no compensatorio*:

1. *Media Arimética Ponderada (WAM)* [8]. Enfoque *compensatorio* más sencillo, no trivial y ampliamente usado [22]. Donde:

$$F(\boldsymbol{z}) = \sum_{i=1}^{n} w_i z_i, \quad w_i > 0 \quad \forall i = 1, \ldots, n : \sum_{i=1}^{n} w_i = 1 \tag{4.6}$$

   En *WAM*, para cada i, j, existe una constante *trade-off* $c_{ij} = w_i/w_j$, tal que cualquier cantidad A dada del indicador z_i puede ser reemplazada por la cantidad $B = c_{ij}A$ del indicador z_j sin variar el valor resultante del índice [22]. Ejemplo: Índice de Desarrollo Humano de las Naciones Unidas (HDI).

2. *Función Mínimo*. Enfoque *no compensatorio* común y simple que utiliza el mín$(z_1, \ldots, z_n)$, es decir que el valor global del índice es igual al valor del indicador con peores resultados, lo que implica la máxima penalización de los desequilibrios entre indicadores [22].

Por otra parte, se introduce la siguiente definición:

Definición 10. ***Desbalance***. Es un desequilibrio entre las variables (indicadores individuales) que se utilizan para construir un índice compuesto dado. Por ejemplo, en el caso de tener solo dos indicadores X e Y, convenientemente normalizados, cuyos valores están entre cero y uno, entonces, un balance perfecto es $X = Y$, mientras que un desbalance máximo ocurre cuando $X = 1$ e $Y = 0$, o viceversa [22].

Varios *índices* (*indicadores compuestos*) se construyen sin considerar el *desbalance*; en otras palabras, son neutrales a éstos, y la mayoría utiliza *WAM*, posiblemente con pesos, que ignora los desequilibrios [22].

El ajuste de desbalances puede ser útil, pues incluso cuando se aborda este problema, los aspectos normativos derivados de los diferentes métodos utilizados no suelen explicarse satisfactoriamente. En general, la falta de claridad se refiere a los procedimientos de *ponderación*, pero correspondientemente a menudo, se refiere a todos los *métodos de agregación* relacionados con el problema del equilibrio [22]. La perspectiva del ajuste de

[8] Por sus siglas en infles: *Weighted Arithmetic Mean*.

desbalances es que la *compensabilidad* entre indicadores puede ser posible, pero su costo aumenta con el *desbalance* [22].

4.1.3.2.1 Ajuste de Desbalances

Suponiéndose que cualquier *índice* se obtiene aplicando al vector $\boldsymbol{z} = (z_1, \ldots, z_n)$ una función de agregación $F(\boldsymbol{z})$ llamada *Función Ajustada por Desequilibrio* (*UAF*[9]). Se espera, que este esquema de agregación que ajuste desequilibrios, cumpla las propiedades, descritas a continuación, y extraídas de [22]:

1. **Monotonicidad positiva.**[10] Si $t > 0$, entonces $F(z_1, \ldots, z_j, \ldots, z_n) \leq F(z_1, \ldots, z_j + t, \ldots, z_n)$.
2. **Quasi-concavity**. $F(\lambda z + (1-\lambda)z') \geq \text{mín}(F(z), F(z'))$, $0 < \lambda < 1$.
3. **Quasi-convexity**. $F(\lambda z + (1-\lambda)z') \leq \text{máx}(F(z), F(z'))$, $0 < \lambda < 1$.
4. **Dominio sin restricciones**. La función F está definida en $\mathbb{R}^n$. En otras palabras, la función debe estar definida en cada posible n-tupla de variables, no solo en las entradas dentro de un intervalo específico como $[0, 1]$ [11].
5. **Continuidad**. La función F es continua sobre su dominio.
6. **Idempotencia**. Si $z_1 = \ldots = z_n$, entonces $F(z_1, \ldots, z_n) = z_1$.
7. **Estabilidad para traslaciones**. $\forall \epsilon$, $F(z_1 + \epsilon, \ldots, z_n + \epsilon) = F(z_1, \ldots, z_n) + \epsilon$.

La *Función Mínimo* y *WAM* son casos extremos de *UAF*, con penalización máxima y mínima (cero), respectivamente [22]. Por tanto, todas las funciones de agregación compensatorias con penalización por desequilibrio son casos intermedios entre estos 2 casos mencionados [22].

4.1.3.2.1.1. Media cóncava

Propuesto en [23] en base a los trabajos de Palazzi y Lauri (1998). Se da la posibilidad de diferentes ponderaciones para los indicadores, con la finalidad de evaluar exógena

[9] Por sus siglas en inglés: *Unbalance-Adjusted Function.*

[10] La función F es débilmente creciente con respecto a cada variable, es decir, si el valor de cada indicador individual para la unidad A es mayor o igual que el valor para la unidad B, entonces el valor del índice en A es mayor o igual que el de B.

[11] Con esta propiedad, se puede utilizar la normalización que mejor se ajuste a los objetivos del análisis, características de los datos, etc., dado que el tipo de normalización no está restringida por el dominio ([22] , p. 29).

y subjetivamente la relevancia de los indicadores individuales. Así, para cada grupo (*dimensión* (Def. 8)) $k = 1, \ldots, K$, se asigna una ponderación p_k, con el fin de calibrar las influencias relativas de los grupos en el *índice*. Luego, denotando al *indicador* i-ésimo del grupo k-ésimo por $x_{i,k}$, $i = 1, ..., N_k$, $k = 1, ..., K$, se le asigna la ponderación $v_{i,k} > 0$. Una asignación de ponderaciones diferentes dentro de un grupo puede usarse para reflejar la importancia relativa de algunas variables con respectos a otras. Finalmente, la ponderación $w_{i,k}$ de un indicador con respecto al conjunto de todas los $N = \sum_{k=1}^{K} N_k$ *indicadores* se define como el producto de las dos ponderaciones anteriores dividido por la suma de las ponderaciones de las variables del mismo grupo. Es decir:

$$w_{i,k} = \frac{p_i v_{i,k}}{\sum_{i'=1}^{N_k} v_{i',k}} \tag{4.7}$$

A partir de este punto, simplemente se denota $w_{i,k}$ por w_i, $i = 1, \ldots, N$ suprimiéndose la indexación del grupo.

Cada variable y_i, $i = 1, \ldots, N$, resultante de la transformación descrita en la Sec. 4.1.2; es decir, los 3 pasos indicados, se "penaliza". La función $f(t) = t - ae^{-bt}$ se aplica a cada variable, posiblemente con a y b diferentes para cada variable. Ésta es definida y suave en todo el eje real, estrictamente creciente, estrictamente cóncava y asintótica con respecto a t a medida que $t \to \infty$. Por tanto, la función de N variables que devuelve el índice sintético propuesto en términos de las variables individuales transformadas viene dada por la media ponderada

$$F(y_1, \ldots, y_N) = \frac{\sum_{i=1}^{N} w_i (y_i - a_i e^{-b_i y_i})}{\sum_{i=1}^{N} w_i} \tag{4.8}$$

donde a_i, b_i son parámetros relacionados a la intensidad de penalización del *desbalance* y de la complementariedad entre factores. Este índice es definido y suave en todo $\mathbb{R}^N$, estrictamente creciente en cada variable por separado, estrictamente cóncava y asintótica a $\frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}$ para $y_1, \ldots, y_N$ grandes.

4.1.4. Análisis de Sensibilidad y Robustez

En la Sec. 2.3.1 ya se han mencionado las desventajas de un *índice*; entre ellas, los juicios de valor, como la selección de *indicadores*, la normalización de los datos, las ponderaciones, y los métodos de agregación, que pueden llevar a una forma sintética inadecuada. No obstante, es ventajoso tener en una sola cifra el desempeño de un proceso complejo [15]. Una combinación de *Análisis de Incertidumbre (UA)* y *Análisis de Sensibilidad (SA)* puede ayudar a calibrar la solidez del *índice* y mejorar la transparencia del uso de estos [13].

El *UA* se centra en cómo la incertidumbre en los *input factores* (las opciones consideradas en cada etapa de construcción como la selección de indicadores individuales, calidad de los datos, normalización, ponderación, método de agregación, etc. [13]) se propaga a través de la estructura del *indicador compuesto* y afecta a sus valores [15]. El *Análisis de Sensibilidad* evalúa la contribución de cada fuente de incertidumbre a la varianza de los resultados [15]. Aunque el *UA* se utiliza con más frecuencia que el *SA* y casi siempre se trata por separado, el uso iterativo del *Análisis de Incertidumbre* y *Sensibilidad* durante el desarrollo de un *índice* podría mejorar su estructura [15].

En esta Memoria no se realiza el *Análisis de incertidumbre*, debido a la extensión del proceso; pues, a cada *input factor* se le asigna una *funcón de densidad de probabilidad (pdf)*, y luego usando la *aproximación Monte Carlo*[12] se realizan múltiples evaluaciones del modelo [13]:

$$IC_c = f_{rs}(I_{1c}, I_{2c}, \ldots, I_{lc}, w_{s1}, w_{s2}, \ldots, w_{sl}) \tag{4.9}$$

donde IC_c (*índice sintético*) es el *índice* para la *unidad*[13] c, $c = 1, ..., n$, según el modelo de ponderación f_{rs}, donde r se refiere al *esquema de agregación*, y s se refiere al *esquema de ponderación*. El *índice* es basado sobre los $l = 7$ *indicadores individuales* normalizados $I_{1c}, \ldots, I_{lc}$ para la unidad (viaje) c y ponderaciones $w_{s1}, \ldots, w_{sl}$ dependientes del esquema. Por lo que este análisis se propone para ser realizado a futuro.

[12] En [16] se explica más a detalle este método.
[13] En este caso, cada viaje corresponde a una unidad.

4.1.4.1. Análisis de Sensibilidad

Dados los $K = 7$ indicadores (VS_j, $j = 1, \ldots, 7$), se realizaron K réplicas eliminando cada vez un indicador diferente y calculando el valor del *índice sintético* a partir de los $K - 1$ indicadores restantes. Para cada repetición, se construyeron las clasificaciones de los viajes según los distintos métodos y, para cada viaje, se calcularon las *diferencias absolutas* de rango entre la posición en la clasificación original y la posición en la clasificación para los $K - 1$ indicadores [24]; es decir:

$$D_c = |ICR_c - IC_c| \tag{4.10}$$

donde ICR_c es el *índice de conducción riesgosa* para la unidad c e IC_c es la puntuación asignada por el *índice* con $K - 1$ indicadores para la unidad (viaje) c.

4.1.4.2. Análisis de Robustez

Se realiza un *Análisis de Robustez* de las estimaciones del *índice sintético* obtenidas a partir de las metodologías (métodos de ponderación y agregación) consideradas, centrándose no tanto, en la correspondencia entre el *índice sintético* y el fenómeno a estudiar, sino más bien en la modificación más o menos extensa de los resultados cuando se añade una perturbación aleatoria. De este modo, los métodos en cuestión se verifican en función de su robustez intrínseca y, por tanto, de su fiabilidad "estadística" más o menos amplia. Se procede de la siguiente manera [24]:

1. Para cada indicador k elemental (DS_i, SS_i para $i = 1, \ldots, 7$), se añadió a los datos originales una perturbación aleatoria uniforme, expresada mediante la fórmula:

$$I_k^N = I_k + Unif(-\alpha, \alpha) \cdot \overline{I_k}, \quad \forall k = 1, \ldots, K \tag{4.11}$$

   con α inicialmente igual a 0.05 e $\overline{I}_k$ valor medio del indicador k[14].
2. Para cada extracción aleatoria, se calcularon los K indicadores finales (VS_i, $i = 1, \ldots, 7$) con valores en torno a los originales y, a continuación, a partir de los enfoques considerados, se obtuvieron los *índices sintéticos* correspondientes.

[14] Si el valor I_k^N es negativo, se fija en cero, y si es mayor a 100, se fija en 100.

3. Se calculó el *coeficiente de variación muestral* en las distintas pruebas para cada unidad (viaje), dado por $cv = \frac{s_x}{|\overline{x}|}$, con $\overline{x} \neq 0$.

4.2. Preprocesamiento de Datos

En esta sección se describen los Métodos usados para tratar las variables usadas para el cálculo de los indicadores individuales.

4.2.1. Métodos de Normalización

Hay muchos casos en los que se puede querer *normalizar* una variable. En primer lugar, está el supuesto, a menudo problemático, de *normalidad* del resultado (condicional a las covariables[15]) en el problema clásico de regresión lineal. Aunque se han utilizado muchos métodos para relajar este supuesto (modelos lineales generalizados, regresión cuantil, modelos de supervivencia, etc.), una técnica que sigue siendo popular es transformar los datos para que los datos parezcan normales, mediante transformaciones tan sencillas como una *logarítmica*, o tan compleja como una *transformación de Yeo-Johnson*. De hecho, muchos métodos de normalización complejos se diseñaron expresamente para encontrar una transformación que pudiera hacer que los residuos de regresión fueran gaussianos. Aunque quizás no sea la solución más elegante al problema, a menudo esta técnica funciona bien como solución rápida [33].

Existen muchos métodos, como por ejemplo: *Transformación Box-Cox*, *Transformación Yeo-Johnson*, *Transformación Lambert* $W \times F$, *Logarítmica*, *raíz cuadrada*, *exponencial*, y *arcseno* (todas estas descritas en [33]). Y otras como: *Box-Cox Modificado*, *Exponential de Manly*, *Módulo de John/Draper*, y *Box-Cox modificado de Bickel/Doksum*. A continuación se describe la *Técnica de normalización ORQ*, método usando en esta Memoria, ya que usando el package de *R* llamado *bestNormalize*, se determinó que para los datos en cuestión, éste método es más adecuado.

[15] Variables que posiblemente predice el resultado bajo estudio.

4.2.1.1. Técnica de Normalización Cuantiles Ordenados

La *Técnica de normalización ORQ*, por sus iniciales inglés (Ordered quantile), se basa en la transformación (Ec. 4.12), originalmente discutida, en (Bartlett, 1947) [35] y desarrollada en Van der Waerden (1952) [33].

El procedimiento de normalización ORQ es un enfoque *semiparamétrico* que utiliza los valores originales de una muestra, los rangos correspondientes, la interpolación y la extrapolación no lineal para estimar una función de transformación normalizadora que pueda aplicarse fácilmente a los nuevos datos.

Formalmente, $\boldsymbol{x}$ se refiere a los datos originales (un vector de longitud n), y $\boldsymbol{z}$ se refiere a los rangos de $\boldsymbol{x}$ (ordenados de manera coherente; es decir, que los rangos $\boldsymbol{x}$ se correspondan adecuadamente con los datos originales $\boldsymbol{x}$). Luego, denotando a los valores individuales dentro del vector $\boldsymbol{z}$, indexados por i como z_i y x_i, y a x^* como a una nueva observación que puede o no estar representada entre los $\boldsymbol{x}$ originales. Entonces, se define $f(x_i) = \Phi^{-1}\left(\dfrac{(z_i-1)/2}{n}\right)$, función que se puede interpretar como la función de distribución normal inversa evaluada en el percentil estimado de x_i, donde $p_i \in (0,1)$ denota al percentil muestral de $\dfrac{(z_i-1)/2}{n}$, el cual es un percentil estimado de x_i.

La *transformación normalizadora ORQ* se define:

$$g(x^*|\boldsymbol{x}) = \begin{cases} f(x^*), & \text{si } x^* \in \{\boldsymbol{x}\} \\ \dfrac{f(x_u)-f(x_l)}{x_u - x_l}, & \text{si } x^* \notin \{\boldsymbol{x}\} \text{ y } \min \boldsymbol{x} < x^* < \max \boldsymbol{x} \\ r(x^*, \boldsymbol{x}), & \text{si } x^* < \min \boldsymbol{x} \text{ o } x^* > \max \boldsymbol{x} \end{cases} \tag{4.12}$$

donde, x_l y x_u se refieren a los puntos más cercanos a x^* que aparecían en los datos originales $\boldsymbol{x}$, definidos como:

$$\begin{aligned} x_l &= \max\{a \in \boldsymbol{x} | a < x^*\} \\ x_u &= \min\{a \in \boldsymbol{x} | a > x^*\} \end{aligned}$$

Y, $r(x^*, \boldsymbol{x})$ es una función de extrapolación, que se determina primero ajustando un modelo lineal generalizado (logit-link) con parámetros β_0, β_1 de la siguiente estructura: $logit(\pi_i) = \beta_0 + \beta_1 x_i$, donde, π_i denota el percentil poblacional correspondiente. Por

otra parte, para ajustar este modelo, se emplea una función objetivo basada en la forma de la log-verosimilitud para un modelo de regresión logística derivado de la distribución binomial:

$$\sum_{i=1}^{n}[p_i n(\beta_0 + \beta_1 x_i) - n\log(1 + \exp(\beta_0 + \beta_1 x_i))]$$

Con este modelo ajustado, se utilizan las estimaciones ($\hat{\beta}_0$, $\hat{\beta}_1$) para informar a las extrapolaciones ORQ. A continuación, dejando,

$$l(a) = \Phi^{-1}\left(\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 a)}\right)$$

que se refiere al cuantil normal de la predicción del modelo para una cantidad a, se puede definir $r(x^*, \boldsymbol{x})$ como:

$$r(x^*, \boldsymbol{x}) = \begin{cases} l(x^*) + \text{mín}_i[f(x_i)] - \text{mín}_i[l(x_i)], & \text{si } x^* < \text{mín}\,\boldsymbol{x} \\ l(x^*) + \text{máx}_i[f(x_i)].\,\text{máx}_i[l(x_i)], & \text{si } x^* > \text{máx}\,\boldsymbol{x} \end{cases} \tag{4.13}$$

El resultado es la aproximación logit "desplazada" de la transformación no paramétrica de los datos originales. El desplazamiento garantiza que la transformación sea suave y unívoca cuando la función de extrapolación se encuentra con el dominio original.

La transformación ORQ puede considerarse *semiparamétrica*; es *no paramétrica* a lo largo del dominio original de $\boldsymbol{x}$, pero es *paramétrica* fuera del dominio original de $\boldsymbol{x}$ [34]. Al igual que ocurre con otros procedimientos basados en rangos, durante el proceso de transformación se pierde cierta información [34].

La *transformación ORQ* es reversible (es decir, uno a uno), lo que permite una interpretación directa; cualquier análisis realizado sobre los datos normalizados puede interpretarse utilizando las unidades originales [34].

La técnica ORQ no garantizará una distribución normal en presencia de empates, pero aun así podría producir la mejor transformación normalizadora en comparación con los otros enfoques posibles [33].

4.2.2. Métodos Multivariados para detección de Valores atípicos

La eliminación de valores atípicos evita estimaciones erróneas, disminuye el riesgo de divulgación y también ayuda a obtener tablas estadísticas de buena calidad. Por tanto, para detectar los valores atípicos es necesario un método multivariante basado en las relaciones entre variables [29].

4.2.2.1. Métodos Multivariantes Robustos

Teniendo en cuenta el aspecto multivariante de los datos, la periferia de las observaciones puede medirse mediante la *distancia de Mahalanobis*. Para evitar el efecto de *enmascaramiento*, es necesario que las estimaciones de estos parámetros sean sólidas y que tengan un punto de ruptura positivo. Las estimaciones del *vector de localización multivariante* μ y de la *matriz de dispersión* Σ son también una piedra angular en el análisis de datos multidimensionales, ya que constituyen la entrada de muchos métodos multivariantes clásicos, es decir. las estimaciones MLE. Estas estimaciones son óptimas si los datos proceden de una distribución normal multivariante, pero son extremadamente sensibles a la presencia de incluso unos pocos valores atípicos en los datos, pues estos influyen en las estimaciones de la media muestral $\overline{x}$ y matriz de covarianza muestral $\boldsymbol{S}$, y, en consecuencia, empeorarán el rendimiento del Procedimiento multivariante clásico basado en estas estimaciones. [31].

En esta Memoria se describen *Stahel Donoho Estimator (MSD)* y *MCD*[16] *estimator*. El primero corresponde al primer estimador propuesto de este tipo, y el *MCD estimator* corresponde al estimador de desglose alto más utilizado [31]. Para más detalle de otros métodos, en [31] se proporciona un conjunto casi completo de estimadores para localización multivariante y dispersión con alto punto de ruptura.

4.2.2.1.1 Estimadores Stahel-Donoho modificados (MSD)

Estimadores SD fue propuesto por Stahel (1981) y Donoho (1982), corresponde al primer estimador equivariante multivariante de localización y dispersión con punto de ruptura alto [31], con algunas propiedades obtenidas por Maronna y Yohai (1995) [30]; que estima

[16] Por sus siglas en inglés: *Minimun Covariance Determinant*.

de forma *robusta* el *vector medio* y la *matriz de varianza/covarianza* proyectando los datos sobre líneas rectas en varias direcciones y ponderando cada punto de datos en función de su desviación respecto al centro en la línea, es decir, que tolera la inclusión de muchos valores atípicos [30]. Se recomienda para conjuntos de datos pequeños [31].

Estimadores Stahel-Donoho modificados (MSD) fue propuesto por Patak (1990) y usado en (Franklin 2000). No obstante, en (Béguin 2003) se presenta otra modificación para *MSD*, que tiene en cuenta los procedimientos anteriores ya existentes. En [30] se describen ambos procedimientos (*MSD* basado en Franklin 2000 y *MSD* basado en Béguin 2003). A continuación el método descrito corresponde a *MSD* basado en Béguin 2003 que ilustran en [30].

Sea el conjunto de datos $\boldsymbol{X}_{n\times p}$, es decir, n observaciones y p variables, y μ y σ^2 estimadores univariantes equivariantes afines de localización y dispersión:

1. Asignar ponderaciones fijas iniciales en uno: $u_i = 1$, para $i = 1, \ldots, n$.
2. Para $k = 1, \ldots m$, con $m = \frac{\exp(2{,}1328+0{,}8023\cdot p)}{p}$. En dimensiones elevadas se puede elegir un m mucho menor (Béguin 2003):
   - *a*) Generar aleatoriamente un vector unitario $v_1 \in \mathbb{R}^p$ utilizando una distribución uniforme en la esfera unitaria en $\mathbb{R}^p$.
   - *b*) Calcula $v_2, \ldots, v_p$ de tal manera que los v_i formen una base ortonormal de $\mathbb{R}^p$ (mediante *Ortogonalización Gram-Schmidt*).
   - *c*) Con cada punto de datos z_i proyectado sobre una recta estirada por el j-ésimo vector base v_j, $j = 1, \ldots, p$ (perpendicular al vector), calcular la longitud del vector de proyección $v_j^T z_i$. Luego, suponiendo una distribución normal, la desviación mediana absoluta (mad) dividida por 0,674 da una estimación de la desviación típica, y la desviación absoluta mediana es el valor mediano (med) de cada punto de datos menos la mediana de la muestra. Así, para $i = 1, \ldots n$ y $j = 1, \ldots, p$ se computa:

$$r_{ij} = \frac{|v_j^T z_i - med(v_j^T z)|}{mad(v_j^T z)/0{,}674} \quad , \quad \tilde{r}_{ij} = \begin{cases} r_{ij}, & \text{si } 0 \le r_{ij} \le c \\ \frac{c^2}{r_{ij}}, & \text{si } c \le r_{ij} \end{cases} \tag{4.14}$$

con $c = \sqrt{\chi^2_{p;0{,}95}}$.

Luego,

$$u_i^k = \prod_{j=1}^{p} \frac{\tilde{r}_{ij}}{r_{ij}} \tag{4.15}$$

Nótese que $\tilde{r}_{ij}/r_{ij}$ da una ponderación de 1 para los puntos de datos situados cerca del centro de los datos no podados, pero una ponderación menor para los puntos de datos situados más lejos del centro de los datos muy podados.

d) Si $u_i^k < u_i$ entonces $u_i = u_i^k$.

3. Calcular las estimaciones ponderadas de localización $\hat{u}$ y dispersión $\hat{V}$ utilizando las ponderaciones u_i.

$$\hat{u} = \frac{\sum_{i=1}^{n} u_i x_i}{\sum_{i=1}^{n} u_i} \quad , \quad \hat{V} = \frac{\sum_{i=1}^{n} (x_i - \hat{u})(x_i - \hat{u})^T u_i^2}{\sum_{i=1}^{n} u_i^2} \tag{4.16}$$

4. Rehaga el bucle 2. pero esta vez sustituyendo la base ortogonal aleatoria (puntos a) y b)) por el cálculo de los componentes principales de la matriz de covarianza ponderada actual.Y compare la ponderación primaria con la ponderación secundaria para cada punto de datos y la ponderación menor se toma como ponderación final.
5. Calcule la *distancia de Mahalanobis* al cuadrado. Utilizando las ponderaciones finales, estimar nuevamente la *matriz de varianza/covarianza* y el *vector medio*, y a partir de estos valores obtener la *distancia de Mahalanobis* al cuadrado $D^2(x_i)$:

$$D^2(x_i) = (x_i - u)^T V^{-1} (x_i - u) \tag{4.17}$$

6. Identificar valores atípicos. El estadístico de prueba F_i para la *distancia de Mahalanobis* al cuadrado $D^2(x_i)$ puede obtenerse según la distribución F con p y $n-p$ grados de libertad:

$$F_i = \frac{(n-p)n}{(n^2-1)p} D^2(x_i) \tag{4.18}$$

El criterio para que el estadístico de prueba F_i se considerase un valor atípico fue el valor del 99,9 %, siguiendo a Franklin y Brodeur (1997).

Wada (2010) implementó los estimadores *MSD* original y mejorado, y corroboró que

las sugerencias de Beguin y Hulliger (2003) mejoran el rendimiento de los estimadores originales adoptados por Statistics Canada. Sin embargo, la aplicación de la versión mejorada se limita a conjuntos de datos con un número reducido de variables, como se describe en la sección anterior [29].

4.2.2.1.2 Determinante de Covarianza Mínimo (MCD)

Propuesto por Rousseeuw (1985), quien también introdujo el *estimador elipsoide de volumen mínimo (MVE)*. El estimador *MCD* es afín equivariante [29], es estadísticamente más eficiente que *MVE* debido a su normalidad asintótica, y tiene una tasa de convergencia más alta que *MVE*. Además, es el estimador de desglose más utilizado [31], y que cuenta con un algoritmo de cálculo rápido (Rousseeuw y Van Driessen 1999) llamado *MCD-Fast* [29], que es el utilizado normalmente en la práctica [31].

El método que se describe es el usado en la función *covMcd* del package *rrcov* del software *R*, preparado por Valentín Todorov basándose en el código de S-plus para el algoritmo *Fast-MCD* (implementado por Rousseeuw y Driessen en 1999) con el paso de corrección de muestras pequeñas de Pison et al. (2002) añadido:

El *estimador MCD* para un conjunto de datos $\{x_1, ..., x_n\} \in \mathbb{R}^p$ se denota por aquel subconjunto $\{x_{i_1}, ..., x_{i_h}\}$ de h observaciones cuya *matriz de covarianza* tiene el determinante más pequeño entre todos los subconjuntos posibles de tamaño h. La ubicación *MCD* y la estimación de dispersión $\boldsymbol{T}_{MCD}$ y $\boldsymbol{C}_{MCD}$ se dan entonces como la media aritmética y un múltiplo de la matriz de covarianza de la muestra de ese subconjunto:

$$\boldsymbol{T}_{MCD} = \frac{1}{h}\sum_{j=1}^{h} x_{i_j} \tag{4.19}$$

$$\boldsymbol{C}_{MCD} = c_{ccf}\, c_{sscf} \frac{1}{h-1}\sum_{j=1}^{h}(\boldsymbol{x}_{i_j} - \boldsymbol{T}_{MCD})(\boldsymbol{x}_{i_j} - \boldsymbol{T}_{MCD})^T \tag{4.20}$$

c_{ccf} y c_{sscf} se denominan *factor de corrección de consistencia* y *factor de corrección para muestras pequeñas*, respectivamente, y se seleccionan de forma que $\boldsymbol{C}$ sea consistente en el modelo normal multivariante e insesgado en muestras pequeñas. Una elección recomendable es $h = \lfloor \frac{(n+p+1}{2} \rfloor$ porque entonces el *punto de ruptura* (*BP*, por su nombre en inglés) del *MCD* se maximiza, pero puede elegirse cualquier entero $h \in [\frac{(n+p+1)}{2}, n]$.

Aquí $\lfloor z \rfloor$ denota la parte entera de z que no es menor que z. Si $h = n$ entonces la ubicación *MCD* y la estimación de dispersión $\boldsymbol{T}_{MCD}$ y $\boldsymbol{C}_{MCD}$ se reducen a la *media muestral* y la *matriz de covarianza* del conjunto de datos completo.

El algoritmo se basa en el C-step, nombrado así porque "C" significa "concentración", y se está buscando una matriz de covarianza más "concentrada" con un determinante más bajo; pasa de una aproximación $(\boldsymbol{T}_1, \boldsymbol{C}_1)$ de la *estimación MCD* de un conjunto de datos $\boldsymbol{X} = \{x_1, ...x_n\}$ a la siguiente $(\boldsymbol{T}_2, \boldsymbol{C}_2)$ con un determinante posiblemente menor $det(\boldsymbol{C}_2) \leq det(\boldsymbol{C}_1)$ calculando las distancias d_i, $i = 1, \ldots, n$ respecto a $(\boldsymbol{T}_1, \boldsymbol{C}_1)$, es decir,

$$d_i = \sqrt{(\boldsymbol{x}_i - \boldsymbol{T}_1)^T \boldsymbol{C}_1^{-1} (\boldsymbol{x}_i - \boldsymbol{T}_1)} \tag{4.21}$$

y luego calcular $(\boldsymbol{T}_2, \boldsymbol{C}_2)$ para aquellas observaciones h que tienen distancias más pequeñas.

Rousseeuw y Van Driessen (1999) han demostrado un teorema según el cual el proceso de iteración dado por el C-step converge en un número finito de pasos a un mínimo (local). Dado que no hay garantía de que se alcance el mínimo global de la función objetivo *MCD*, la iteración debe iniciarse muchas veces a partir de subconjuntos iniciales diferentes, para obtener una solución aproximada. El procedimiento es muy rápido para conjuntos de datos pequeños, pero para que sea realmente "rápido" también para conjuntos de datos grandes se utilizan varias mejoras computacionales.

1. *Subconjuntos iniciales.* Es posible reiniciar las iteraciones a partir de subconjuntos de tamaño h generados aleatoriamente, pero para aumentar la probabilidad de extraer subconjuntos sin valores atípicos, se seleccionan aleatoriamente $p+1$ puntos. Estos $p+1$ puntos se utilizan para calcular $(\boldsymbol{T}_0, \boldsymbol{C}_0)$. A continuación, se calculan las distancias d_i, $\forall i = 1, \ldots, n$ y se ordenan en orden creciente. Por último, se seleccionan las primeras h para formar el h-subconjunto inicial H_0.
2. *Reducción del número de C-steps.* El C-step, que implica el cálculo de la *matriz de covarianza*, su determinante y las distancias relativas, es la parte del algoritmo que requiere más cálculos. Por lo tanto, en lugar de iterar hasta la convergencia para cada subconjunto inicial, sólo se realizan dos C-step y se mantienen los 10

subconjuntos con el determinante más bajo. Sólo se itera hasta la convergencia con estos subconjuntos.

3. *Particionamiento*. Dividir el conjunto de datos en un máximo de cinco subconjuntos de aproximadamente el mismo tamaño (pero no más de 300, por ejemplo) e iterar en cada subconjunto por separado. Se conservan las diez mejores soluciones para cada conjunto de datos y, finalmente, sólo se itera sobre el conjunto de datos completo.
4. *Anidación*. Puede reducirse aún más el tiempo de cálculo para conjuntos de datos con $n > 1.500$, por ejemplo, extrayendo 1.500 observaciones sin reemplazo y realizando los cálculos (incluida la partición) en este subconjunto. Sólo las iteraciones finales se realizan en el conjunto de datos completo. El número de iteraciones depende del tamaño real del conjunto de datos.

El *estimador MCD* no es muy eficiente en modelos normales, especialmente si h se selecciona de forma que se consiga el máximo *BP*. Para superar la baja eficiencia del *estimador MCD*, puede utilizarse una versión reponderada.

Según [29], *MSD* es un mejor estimador con datos sesgados a comparación con *MCD*.

4.3. Datos Composicionales

Definición 11. ***Composición***. Se define a menudo como un vector de D *componentes* o partes tal que,

$$\boldsymbol{x} = (x_1, ..., x_D), \quad x_i > 0, \quad i = 1, ..., D; \quad \sum_{i=1}^{D} x_i = k \tag{4.22}$$

k fijada típicamente igual a 1 (porciones), 100 (porcentajes) o 10^6 (ppm) [44].

Sin embargo, esta definición es engañosa, porque muchos conjuntos de datos composicionales no la satisfacen aparentemente [44].

Definición 12. ***Subcomposición***. Una *composición* que sólo representa algunas de las *componentes* posibles [44].

Se considera un vector como una *composición* siempre que sus *componentes* representen el peso relativo o la importancia de un conjunto de partes que forman un todo [44]. Luego, el reescalado de las composiciones se puede formalizar mediante la operación de *cierre* [44] [47].

Definición 13. ***Cierre***. Operación que fuerza a los vectores de datos composicionales a compartir la misma suma total k. Es decir, para cada vector positivo $\boldsymbol{x} = (x_1, ..., x_D)$,

$$\mathcal{C}_k(\boldsymbol{x}) = \left(\frac{k \cdot x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{k \cdot x_D}{\sum_{i=1}^{D} x_i} \right) \tag{4.23}$$

El *cierre* suele ser una manipulación controvertida, ya que al cerrar una *subcomposición*, parece que se pierde cierta información sobre la masa total presente en las partes consideradas, y aparentemente sustituimos algunos valores medidos por otros calculados [44].

Definición 14. ***Simplex***. Puede considerarse una generalización de la noción de triángulo o tetraedro a dimensiones superiores [47].

Definición 15. ***Simplex de D partes (S^D)***. Subconjunto de $\mathbb{R}^D$ definido por [47],

$$S^D = \left\{ \boldsymbol{x} = (x_1, ..., x_D)' \in \mathbb{R}^D : x_i > 0, \sum_{i=1}^{D} x_i = k \right\} \tag{4.24}$$

Nótese que las partes compositivas son positivas. Esto es una limitación, pero no implica que el análisis de datos composicional no pueda tratar ceros. La definición con valores estrictamente positivos es más bien una conveniencia para un tratamiento metodológico estándar basado en los *logratios*, donde los ceros conducirían a valores mal definidos [47].

Definición 16. ***Composiciones equivalentes***. Dos *composiciones* $\boldsymbol{x}$ e $\boldsymbol{y}$, es decir, vectores en $\mathbb{R}^D_+$, que tengan cualquier suma (posiblemente diferente) de partes compositivas, reescaladas con igual suma k, es decir, $\mathcal{C}_k(\boldsymbol{x})$ y $\mathcal{C}_k(\boldsymbol{y})$, tal que $\mathcal{C}_k(\boldsymbol{x}) = \mathcal{C}_k(\boldsymbol{y})$. Entonces, $\boldsymbol{x}$ e $\boldsymbol{y}$ son "iguales" y difieren solo en la constante (factor escala) y son llamadas *composiciones equivalentes* [47].

El énfasis en la restricción de la suma constante (Ec. 4.24) puede llevar incluso a confusión, ya que da la impresión de que todos los conceptos para el análisis de *datos composicionales* sólo son válidos para el caso de suma constante de las *partes composicionales*.

Con el fin de evitar esta confusión, se entrega una nueva definición [47]:

$$\tilde{S}^D = \{\boldsymbol{x} = (x_1, ..., x_D)' \in \mathbb{R}^D : x_i > 0, \forall k \quad \exists! \, \lambda > 0 : \boldsymbol{x} = \lambda \mathcal{C}_k(\boldsymbol{x})\} \tag{4.25}$$

La constante k no importa en absoluto. En otras palabras, el espacio $\tilde{S}^D$ se refiere a $\mathbb{R}^D_+$, descompuesto según clases de equivalencia de vectores *composicionalmente equivalentes* [47].

4.3.1. Principios del Análisis Composicional

Cualquier metodología estadística objetiva debería dar resultados equivalentes cuando se aplica a dos conjuntos de datos que sólo difieren en detalles irrelevantes [44]. Por tanto, se generan cuatro principios de invarianza, que sustentan el análisis de datos composicionales [44]:

- ***Escala Invariante***. Un análisis composicional sensato debería proporcionar la misma respuesta independientemente del valor de o incluso independientemente de si se ha aplicado el *cierre* o de si los vectores de datos suman valores diferentes [44]. (Aitchison 1986), bibliografía relacionada de [44], demostró que todas las funciones invariantes de escala de una composición pueden expresarse como funciones de relaciones logarítmicas $\ln(x_i/x_j)$.
- ***Invarianza de la Perturbación***. Los datos de composición se pueden presentar en muchas unidades diferentes, e incluso cuando se dan en porciones, sigue siendo relevante en qué cantidades físicas se midieron originalmente las componentes (g, toneladas, porcentaje en masa, porcentaje en volumen, etc.) [44]. Dado que, es evidente, que los análisis estadísticos aplicados deben dar los mismos resultados cualitativos, independientemente de las unidades elegidas, siempre que contengan la misma información, es decir, siempre que podamos transformar las unidades entre sí [44]. Esta operación se realiza mediante una *perturbación* con una composición que contiene como entradas los factores de conversión para cada *componente*. Pero como nunca se sabe qué tipo de cuantificaciones podrían considerarse, se debería solicitar la invariancia del análisis con respecto a la *perturbación* con todos los factores de ponderación posibles [44].

Este principio es particularmente crítico cuando se trata de conjuntos de datos de unidades mixtas (composición con componentes en diferentes unidades). Por tanto, cuando se exige la *invarianza de perturbación*, siempre que exista una *perturbación* que lleve los datos al mismo sistema de unidades, incluso si no se sabe, estos datos podrán ser analizados significativamente en un marco común [44].

- ***Coherencia Subcomposicional***. Las *subcomposiciones* desempeñan el mismo papel con respecto a las composiciones que los *marginales* en el análisis multivariante real convencional: representan subespacios de dimensión inferior donde los datos pueden proyectarse para su inspección [44]. Esto tiene varias implicaciones enumeradas a continuación, que Aitchison (1986) denomina conjuntamente coherencia subcomposicional:

  1. Si se mide la distancia entre dos composiciones de $D-$partes, ésta debe ser mayor cuando se mide con todos los componentes D que cuando se mide en cualquier subcomposición.
  2. La dispersión total de un conjunto de datos composicional de $D-$partes debe ser mayor que la dispersión en cualquier subcomposición.
  3. Si se ajusta un modelo significativo a un conjunto de datos de composición de $D-$partes, el resultado no debería cambiar si incluimos una nueva componente no informativa (por ejemplo, aleatorio) y se trabajara con la composición $(D+1)-$parte resultante.

- ***Invarianza de Permutación***. los resultados de cualquier análisis no deberían depender de la secuencia en la que se dan las componentes en el conjunto de datos [44]. Para el enfoque *logratio*, también es un principio muy importante [44], por ejemplo, la *alr-transfomación* (Sec. 4.3.2.2) no es *invariante de permutación*, y por tanto no debería utilizarse para el *Análisis de Conglomerados* [44].

4.3.2. Representaciones de coordenadas de composiciones

Frecuentemente se asocian los datos composicionales con primero aplicar una transformación adecuada, y a continuación, emplear la metodología estadística estándar como de costumbre [47]. Sin embargo, a pesar de ser práctico, luego la dificultad está en la interpretación de resultados, porque al aplicar una transformación, ya no se trabaja con las composiciones originales.

Dado que la *Geometría de Aitchison* constituye una estructura de *espacio euclídeo* $(D-1)-$dimensional sobre el *Simplex*, se puede trasladar prácticamente cualquier cosa definida para vectores reales a *composiciones*, ya que un espacio euclídeo es siempre equivalente al espacio real. Esta equivalencia se consigue mediante una *isometría*, es decir, una transformación del *Simplex* al espacio real que mantiene *ángulos* y *distancias* [44]. Por tanto, el foco en esta sección es en las coordenadas *logratios* que expresan datos composicionales, impulsada por la *Geometría de Aitchison*, en la geometría euclidiana habitual del espacio real [47].

4.3.2.1. Coeficientes de logratio centrado (clr)

Una composición $\boldsymbol{x} \in \tilde{S}^D$ es expresada por un vector $\boldsymbol{y} \in \mathbb{R}^D$, con:

$$\boldsymbol{y} = clr(\boldsymbol{x}) = \left(\ln \frac{x_i}{g(\boldsymbol{x})}\right)_{i=1,\ldots,D} \quad con \quad g(\boldsymbol{x}) = \sqrt[D]{x_1 \cdot x_2 \cdots x_D} \tag{4.26}$$

o dicho de manera compacta, $clr(\boldsymbol{x}) = \ln(\frac{\boldsymbol{x}}{g(\boldsymbol{x})})$, donde el *logratio* del vector es aplicado por componentes [44]. Además, las componentes *clr-transformadas* suman cero (revisar [47], p. 46). En efecto, la imagen de *clr* es un hiperplano del espacio real $\mathbb{H} \subset \mathbb{R}^D$ ortogonal al vector $\mathbf{1}_D$ [44]. Finalmente, cabe mencionar que, $g(\boldsymbol{x})$ corresponde a la *media geométrica* de $\boldsymbol{x}$ [47].

4.3.2.2. Coordenadas del coeficiente aditivo (alr)

Mapeo de $\tilde{S}^D$ a $\mathbb{R}^{D-1}$, y el resultado para $\boldsymbol{x} \in \tilde{S}^D$ son coordenadas $\boldsymbol{x}^{(j)} \in \mathbb{R}^{D-1}$ con,

$$\boldsymbol{x}^{(j)} = alr_j(\boldsymbol{x}) = (x_1^{(j)}, \ldots, x_{D-1}^{(j)})' = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j}\right)' \tag{4.27}$$

Si $\boldsymbol{X}_{n\times D}$ es una matriz de datos de composición, con las *composiciones* $x_i' = (x_{i1}, \ldots, x_{iD})$ en las filas de $\boldsymbol{X}$, para $i = 1, \ldots, n$, entonces la matriz de coordenadas *alr* está formada por las filas,

$$\left(\boldsymbol{x}_i^{(j)}\right)' = (alr_j(\boldsymbol{x}_i))' = \left(\ln \frac{x_{i1}}{x_{ij}}, \ldots, \ln \frac{x_{i,j-1}}{x_{ij}}, \ln \frac{x_{i,j+1}}{x_{ij}}, \ldots, \ln \frac{x_{iD}}{x_{ij}}\right) \tag{4.28}$$

El índice $j \in \{1, \ldots, D\}$ se refiere a la variable que se elige como variable de relación en las coordenadas. Esta elección usualmente depende del contexto, pero además de la idoneidad de los resultados para la visualización y exploración de datos [47].

Por último, el análisis de datos composicionales en el enfoque original de Aitchison (1986) se basaba en la *alr-transformación* [44].

4.3.2.3. Logratio isométrico (ilr)

La *ilr-transformación* tiene por objeto construir una base ortonormal en el hiperplano $(D-1)-$dimensional de $\mathbb{R}^D$, formada por los coeficientes *clr*, y expresar en él la *composición*, donde el resultado es un vector $\boldsymbol{z} \in \mathbb{R}^{D-1}$. Dado que existen infinitas posibilidades para definir dicho sistema de bases ortonormales, *ilr* es considerado como una clase de coordenadas, siendo común referirse a esto como *coordenadas ortonormales (logratio)* [47].

Una elección particular de una base conduce a $ilr(\boldsymbol{x}) = \boldsymbol{z} = (z_1, \ldots, z_{D-1})'$, con ((Fišerová and Hron 2011), bibliografía relacionada de [47]):

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \quad para \quad j = 1, \ldots, D-1 \tag{4.29}$$

Una parte (aquí x_1) se establece como *pivote*, la cual está contenida sólo en la primera coordenada. Por tanto, a estas *ilr coordenadas* se le denominan *coordenadas pivote (logratio)* [47]. Esta elección tiene también una importancia primordial para el sistema de coordenadas en su conjunto [47].

Para una data composicional $\boldsymbol{X}_{n\times D}$, con $\boldsymbol{x}' = (x_{i1}, \ldots, x_{iD})$, $i = 1, \ldots n$, en sus filas, la $\boldsymbol{Z}_{n\times(D-1)}$ matriz de *coordenadas pivote* es formada por los elementos con índice (i, j),

$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_{ij}}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_{ik}}} \tag{4.30}$$

4.3.3. Análisis de Componentes Principales (PCA)

Es la principal herramienta para el análisis de exploración multivariado. Tiene como objetivo reducir la dimensionalidad del conjunto de datos input mediante un conjunto de datos con nuevas coordenadas llamadas *Componentes Principales (CP)*, que buscan la mayor variabilidad explicada posible [49]. Pueden derivarse usando una *Descomposición de Valor Singular (SVD)* de una matriz de datos, o mediante una descomposición de *valores propios* de la matriz de covarianza para obtener los *loadings* (coeficientes de base) y las puntuaciones (coordenadas) de los componentes principales [49].

4.3.3.1. Estimación de Componentes Principales

El *PCA* define nuevas variables, consistentes en combinaciones lineales de las originales, de tal forma que el primer eje se encuentra en la dirección que contiene la mayor variación. Cada nueva variable posterior es ortogonal a las anteriores, pero de nuevo en la dirección que contiene la mayor parte de la variación restante. Las nuevas variables se denominan *Componentes Principales (CP)* [49]. Desde un punto de vista práctico, el *PCA* proporciona un mapeo directo de los datos originales, posiblemente de alta dimensión, en un espacio de menor dimensión que captura la mayor parte de la información contenida en los datos originales [49].

4.3.3.2. Estimación por SVD

Para un conjunto de datos composicionales, la *Descomposición de Valor Singular (SVD)* de una data centrada *clr-transformada* $\boldsymbol{X}^*$ es un producto de tres matrices:

$$clr(\boldsymbol{X}^*) = \boldsymbol{U} \cdot \boldsymbol{D} \cdot \boldsymbol{V}^t \tag{4.31}$$

Donde:

- $\boldsymbol{V}$. Las filas identifican las variables y sus columnas son vectores ortonormales de D componentes, llamados *vectores singulares derechos*, *loadings*, o *componentes principales (CP)*; ellos definen una base ortonormal del *Simplex*.

- $\boldsymbol{D}$ es una matriz diagonal, con D valores positivos ordenados en orden decreciente, llamados *valores singulares*. Son interpretados como la desviación estándar de las coordenadas del conjunto de datos sobre la nueva base.
- $\boldsymbol{U}$. Las filas identifican las observaciones, y las columnas son vectores ortonormales de N componentes, llamados *vectores singulares izquierdos* o *scores*. Las filas de $\boldsymbol{U}$ son interpretadas como las coordenadas estandarizadas del conjunto de datos sobre la nueva base.

La *clr-transformación* debe aplicarse para que los *vectores singulares izquierdos* y *derechos* reproduzcan la escala relativa de los datos de composición [46].

La mejor aproximación de rango r a la $\boldsymbol{X}^*$, en el sentido de mínimos cuadrados, se obtiene seleccionando los r primeros *valores singulares* y los r primeros *vectores singulares izquierdos* y *derechos* asociados, y calculando con ellos:

$$\boldsymbol{X}_r = \boldsymbol{U}_r \boldsymbol{D}_r \cdot \boldsymbol{V}_r^t \tag{4.32}$$

La calidad global de esta aproximación puede medirse por la proporción de varianza preservada:

$$\pi_r = \frac{\sum_{i=1}^{r} d_{ii}^2}{\sum_{j=1}^{D} d_{jj}^2} \tag{4.33}$$

La importancia de esta descomposición radica en el hecho de que permite una representación óptima de un conjunto de datos multidimensional en una dimensión inferior [46].

4.3.3.3. Biplot

Un *biplot* es una representación gráfica del caso $r = 2$ (o en 3D, con $r = 3$), donde se muestran las variables y las observaciones, en una especie de proyección simultánea de la nube de datos y los ejes de las variables en un gráfico bidimensional (o 3D) [46].

Para construir un *biplot*, las observaciones se representan típicamente como puntos, en las posiciones dadas por las filas de $\boldsymbol{U}_r \cdot \boldsymbol{D}_r^{(1-\alpha)}$, y las variables se grafican como flechas

desde el centro del gráfico hasta las filas de $\boldsymbol{V}_r\boldsymbol{D}_r^{\alpha}$, con $\alpha \in (0,1)$, y 1 es el valor predeterminado en R. Si $\alpha = 0$, se denomina *biplot de forma*, y si $\alpha = 1$ se denomina *biplot de covarianza* [46].

Cuando se interpreta un *biplot*, ya sea composicional o clásico, es muy importante recordar que la calidad de representación es controlada por la proporción de variabilidad conservada [46].

En el *biplot de forma*, la proporción de variabilidad de una variable viene dada por la longitud de su rayo: así, una variable perfectamente representada tiene una longitud de una unidad, y una variable mal representada tiene un rayo muy corto [46]. Esta proporción de variabilidad explicada para cada componente es llamada *comunalidad* [46] (p. 183).

En un *biplot de covarianza* no composicional, las longitudes de flecha son proporcionales a las varianzas de cada variable, y el coseno del ángulo entre dos flechas es el *coeficiente de correlación* entre esas dos variables. Para los datos de composición, se debe recordar que el *SVD* se aplicó a los datos *clr-transformados*. Esto implica que no se pueden interpretar directamente los rayos de un *biplot composicional*: cada uno representa la *varianza clr* de una parte y tiene una relación bastante compleja con todas las partes originales. En su lugar, los vínculos entre las puntas de flecha representan los *logratios* entre las dos partes implicadas y están íntimamente relacionados con la *matriz de variación* [46].

Definición 17. ***Matriz de Variación***. La matriz tiene D^2 componentes, cada una definida por [45]:

$$\tau_{ij} = var\left(\ln \frac{x_i}{x_j}\right) \tag{4.34}$$

y son estimados por $\hat{\tau}_{ij} = \dfrac{1}{N-1}\displaystyle\sum_{n=1}^{N} \ln^2 \frac{x_{ni}}{x_{nj}} - \ln^2 \frac{\overline{x}_i}{\overline{x}_j}$. Además, es una matriz simétrica, dado que $\ln\left(\frac{a}{b}\right) = -\ln\left(\frac{b}{a}\right)$ y $var(-c) = var(c)$. Un valor pequeño de $\tau_{ij} = \tau_{ji}$ implica una varianza pequeña de $\ln(x_i/x_j)$, por tanto una "buena proporcionalidad" $x_i \propto x_j$.

La estructura de composición de dependencia en el *biplot de covarianza* puede interpretarse como sigue [46]:

- Dos variables proporcionales tienen una relación logarítmica casi constante (una pequeña entrada en la matriz de variación); por lo tanto, su enlace debería ser muy corto, y las puntas de flecha se encuentran juntas.
- Si un enlace es muy largo, la relación logarítmica de las dos partes implicadas es muy variante (una entrada grande en la matriz de variación). Si vemos tres rayos muy largos que apuntan hacia direcciones diferentes (a 120° aproximadamente), un diagrama ternario de estas tres partes tendrá una dispersión elevada, porque sus enlaces mutuos también son muy largos.
- El ángulo entre dos enlaces debe aproximarse al coeficiente de correlación entre las dos relaciones logarítmicas:
    - Dos relaciones logarítmicas no correlacionadas proporcionan rayos ortogonales.
    - Tres o más partes situadas sobre una línea común tienen enlaces que forman 0° o 180°, estando así perfectamente correlacionadas (directa o inversamente); en este caso, la subcomposición formada por estas partes debería mostrar un patrón unidimensional de variación; es decir, esta subcomposición es probablemente colineal.
    - Dos conjuntos de subcomposiciones colineales cuyos rayos forman 90° están (posiblemente) no correlacionados.

4.3.3.4. Scree Plot

El *Scree Plot* es un bar plot (o line plot) de la varianza de las componentes principales, es decir, el número d_{ii}^2: un buen *biplot* es asociado con un *scree plot* donde las primeras dos columnas son muy grandes en comparación con las barras restantes [46].

4.3.3.5. Loadings

Reformulando el *SVD* en términos de las operaciones internas del simplex, la perturbación y la potenciación; los vectores de *loadings* pueden (y deben) interpretarse como composiciones, transmitiendo solo información sobre incrementos o decrementos de una parte con respecto a otra [46]. Esto significa que no se puede interpretar los *loadings* de forma aislada, pero al menos sí, en pares [46].

4.4. Comparación de Muestras de datos usando un Método no Paramétrico Multivariante

A diferencia del *Análisis multivariante clásico de la varianza*, no se requiere normalidad multivariante para los datos [40].

El package *npmv* de *R* proporciona una aproximación no paramétrica completa, cuyo método se describe a continuación en base a [40].

4.4.1. Modelo Multivariado no paramétrico

El modelo no paramétrico subyacente al paquete *npmv* de *R* establece para las a muestras (*niveles de factor*) de vector de observaciones de p variables, es decir las p *variables de respuesta*, con muestras individuales de tamaño $n_1, \ldots, n_a$, respectivamente, y un total muestral de $N = \sum_{i=1}^{a} n_i$, que los vectores de observación multivariantes $\boldsymbol{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(p)})^T$ son independientes y que dentro del mismo nivel de factor i, siguen la misma distribución p-variable: $\boldsymbol{X}_{ij} \sim F_i$. Denotando a las diferentes variables por $k = 1, \ldots, p$, a las diferentes condiciones (tratamientos, subpoblaciones, niveles de factor) por $i = 1, \ldots, a$, y dentro de cada condición, los n_i sujetos (unidades experimentales), sobre los que se realizan las observaciones de las p-variables, se indexan por $j = 1, ..., n_i$. Se asume implícitamente, que se observan las mismas p variables de respuesta en cada uno de los a niveles de factor, y estas p variables pueden ser dependientes. No es necesario especificar la estructura de dependencia [40]. Además, las distribuciones marginales pueden, por supuesto, ser diferentes para las distintas variables de respuesta.

4.4.2. Hipótesis Estadísticas Globales

Las Hipótesis Estadísticas Globales típicas en este contexto son las siguientes: "¿Proceden las a muestras de la misma población (distribución multivariante)?" o "¿Tienen los a tratamientos el mismo efecto?". Esto puede formularse como $H_0 : F_1 = \ldots = F_a$.

Para probar la hipótesis nula general de que las distribuciones multivariantes F_i, $i = 1, \ldots, a$, no difieren entre los niveles de los factores, se emplean estadísticos de prueba que utilizan sumas de cuadrados y productos cruzados basados en rangos. En este caso,

los rangos se toman en función de las variables. En consecuencia, los estadísticos de prueba resultantes son *invariantes* bajo transformaciones estrictamente monótonas de las variables de respuesta individuales. Se trata de una propiedad importante y deseable, ya que, por ejemplo, cambiar la escala de una variable de porcentaje a proporción o de unidades métricas a imperiales[17], o utilizar un conjunto de números diferente para una característica ordinal, no debería cambiar los resultados de la prueba.

En el Apéndice A de [40], se resume brevemente cómo se construyen las cuatro pruebas estadísticas análogas basadas en rangos de tipo *ANOVA*, tipo *Lambda de Wilks*, tipo *Lawley Hotelling* y tipo *Bartlett Nanda Pillai*, además de las aproximaciones de la distribución F. En total, hay ocho pruebas (cuatro tipos, cada uno con aproximación F y como prueba de permutación). Ninguna de ellas es uniformemente mejor que las demás.

El paquete *npmv*, permite que las variables de respuesta sean métricas, ordinales, incluso binarias [40].

[17] Sistema de medida utilizado en el Reino Unido y otros países de la Commonwealth.

Capítulo 5

Análisis de datos

En esta sección se detalla el preprocesamiento de los datos realizado, ilustrando el conocimiento obtenido de éstos a partir de este análisis, con el objetivo de tomar las elecciones correctas para el proceso de construcción y análisis del *índice de Conductor Riesgoso (ICR)*. Todo el proceso fue realizado en el software estadístico *RStudio* versión 4.2.2.

5.1. Conjunto de Datos

Se utilizan dos conjutos de datos, *viajes* con 1.684.002 observaciones (es decir, 1.684.002 registros de viajes diferentes), desde periodo de 2021-11-28 17:22:53 hasta 2022-02-27 23:57:50, y los datos *alertas* con 267072 observaciones (registros) ocurridas en el mismo intervalo de tiempo que *viajes*. Dado que las alertas registradas no indican a que viaje pertenecen, sino que solo indican la patente (*Movil*) y el lapsus de tiempo en el que se registraron, se realiza un match considerando estas variables, y la fecha-hora y patente registradas para los viajes. El resultado es un conjunto de datos con 1.684.002 registros de viajes contenidos en 1.783.860 observaciones, compuesto por las variables de ambos conjuntos: *viajes* y *alertas*. Por otra parte, se cuenta con otro conjunto de datos que contiene 47050 viajes en 196651 observaciones con sus respectivas alertas, desde 2022-04-20 00:00:20 a 2022-05-05 23:37:21. Así, el conjunto de datos unificado queda en 178021 registros de viajes diferentes en 427480 observaciones.

Se debe mencionar que previamente se eliminaron los viajes que no contenían alertas registradas.

A continuación se describen las variables disponibles:

Tabla 5.1: Descripción Variables Disponibles

Variable	Descripción	Unidad	Missing Values (n)	Missing Values (%)
ID	Identificador único para cada viaje registrado	-	0	0
Movil	Patente del Móvil para viaje registrado	-	0	0
InicioFecha	Fecha y Hora del viaje	*AAA-MM-DD H:M:S*	0	0
FinFecha	Fecha y Hora de finalización del viaje	*AAA-MM-DD H:M:S*	0	0
Duración	Duración viaje	*Minutos*	0	0
Recorrido	Recorrido viaje	*Kms*	0	0
Transportista	Empresa	-	0	0
TipoMovil	Tipo de móvil en que se realiza el viaje	-	2466	1.4
Conductor	Nombre Conductor Registrado en el viaje	-	166920	93.8
Reg	Identificador único para cada alerta registrada	-	0	0
idalerta	Tipo de Alerta Registrada	-	82410	19.3
Aler_Fecha_Ini	Fecha y Hora de la Alerta Registrada	*AAA-MM-DD H:M:S*	0	0
Aler_Fecha_Fin	Fecha y Hora de finalización de la Alerta	*AAA-MM-DD H:M:S*	201528	47.1
Aler_Velocidad	Velocidad del Móvil al registrar la Alerta	*Kms/Hr*	116	0.03
Aler_Duración	Duración de la Alerta	*seg*	230829	54.0
Aler_Lat	Latitud al Momento de Registrarse la Alerta	-	196651	46.0
Aler_Lon	Longitud al Momento de Registrarse la Alerta	-	196651	46.0
TipoVehiculo	Clasificación del Móvil (Liviano, Pesado)	-	2466	1.4

5.1.1. Data Cleaning

5.1.1.1. Imputación

Los Tipos de Alerta $i = 1, 2, 3, 4, 5, 7$ tienen una duración predefinida de $2s$ para la variable *Aler_Duracion*. Por ende, se imputó con este valor a todos los registros de alertas ($i \neq 6$) cuando el valor estaba ausente.

5.1.1.2. Valores Atípicos

Los valores atípicos multivariantes son observaciones que se consideran extrañas no por el valor que toman en una determinada variable, sino en el conjunto de aquellas. Son más difíciles de identificar que los valores atípicos unidimensionales [28]. Su presencia tiene

efectos todavía más perjudiciales que en el caso unidimensional, porque distorsionan no sólo los valores de la *medida de posición (media)* o de *dispersión (varianza)*, sino muy especialmente, las *correlaciones* entre las variables [28].

TABLA 5.2: Estadísticos descriptivos de las variables numéricas

Est.	*Duración*	*Recorrido*	*Aler_Velocidad*							*Aler_Duracion*
			Tipo Alerta							Tipo Alerta
			1	2	3	4	5	6	7	6
Mínimo	1	0	1	0	-127	0	0	0	0	0.1
Q_1	28	11	50	2	35	39	29	43	0	2
Mediana	53	32	50	23	50	50	50	81	5	8
Mean	67.04	2182	50.5	29.02	51.93	56.67	52.28	66.7	13	36.68
Q_3	84	51	50	50	69	77	83	82	23	32
Máximo	1439	8623563	273	201	129	129	133	9991	97	2005
skew	4.01	51.27	2.54	0.87	0.53	0.42	0.03	97.5	1.2	7.37
Kurtosis	35.48	3527.56	19.45	0.07	-0.09	-0.61	-1.03	16788	0.4	78.97
Trimmed	56.6	34.08	49.43	25.12	50.66	55.57	51.94	68.31	10.3	17.21
n	178021	178021	3718	104016	17662	1814	11171	191974	14715	175460
NA's	0	0	0	0	0	0	0	0	0	16514

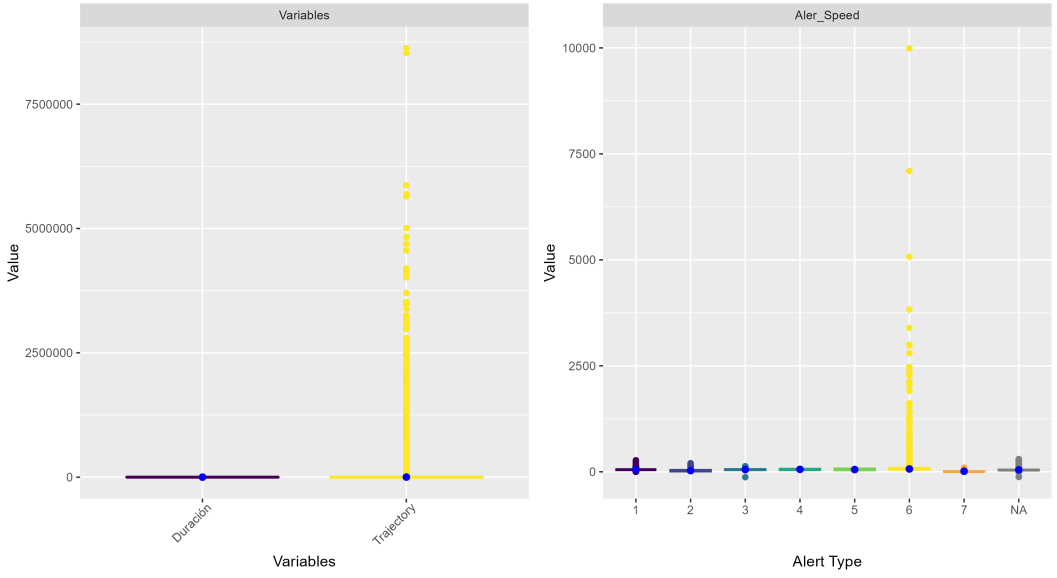

FIGURA 5.1: *Boxplot* Inicial de las variables numéricas

TABLA 5.3: Correlación entre Variables numéricas

Tipo Alerta	Variables		n	Correlación	p value
-	Duración	Recorrido	178021	0.013	8.191617×10^{-8}
6	Aler_Duración	Aler_Velocidad	427480	0.017	3.162359×10^{-12}

5.1.1.2.1 Variables: *Duracion*, *Recorrido*

Para ambas variables se nota la presencia de valores "sin sentido" en sus valores máximo y mínimo, especialmente en la variable *Recorrido* (Tabla 5.2 y Figura 5.1). Y tal como se menciona previamente, la media se ve afectada por los posibles valores atípicos presentes en el conjunto de datos, ya que efectivamente, se aprecia una diferencia bastante amplia, entre *trimed*[1] y *mean*. Además, que la correlación entre estas variables es significativa para $\alpha = 0.05$[2] (Tabla 5.3). Sin embargo, es bastante pequeña a comparación de lo que indicaría la lógica, pues comúnmente se espera que mientras más dure un viaje en tiempo, mayor sea su recorrido realizado. Pero la correlación de Pearson es sensible a observaciones influyentes, y en ocaciones determinados puntos de datos pueden ejercer un efecto desproporcionado a la hora de estimar la significación de una correlación, hasta el punto de que su eliminación del análisis conduce a un resultado no significativo [25], o incluso, que una correlación no significativa se convierta en significativa tras eliminar puntos influyentes [25]. Es más, el signo de una correlación estadísticamente significativa también puede verse afectado por puntos influyentes [25].

De la Tabla 5.2, se aprecia que ambas variables presentan sesgo positivo y una curtosis elevada, por lo que es necesario transformar estas variables antes de enfocarse en la búsqueda de valores atípicos, pues deben aproximarse a una *Distribución Elíptica* [29]. No obstante, para muestras grandes, por ejemplo 1000 observaciones o más, la prueba de normalidad puede concluir que una pequeña desviación de la normalidad es significativa [27], tal como sucede en este caso, ya que al realizar la prueba con el Test Kolmogorov-Smirnov (Tabla 5.4) usando un nivel de significancia $\alpha = 0.05$, ambas variables rechazan la Hipótesis Nula H_0 de distribución normal. Por tanto, lo recomendable es mirar el gráfico *Q-Q normal* para ver si la desviación de la normalidad es realmente significativa [27] [26]. En consecuencia, de las Figuras 5.2 y 5.3, se confirma la necesidad de transformar estas variables.

[1] *Trimmed*, o *media recortada*, se refiere a la media obtenida luego de recortar el 10 % de los datos por ambos lados; o dicho de otra manera, es la media obtenida a partir del 80 % central de los datos.

[2] Se rechaza la hipótesis $H_0 : \rho = 0$.

TABLA 5.4: Test de normalidad Kolmogorov-Smirnov

Variable	statistic	p.value
Duración	0.1615579	< 2.2e-16
Recorrido	0.5074526	< 2.2e-16

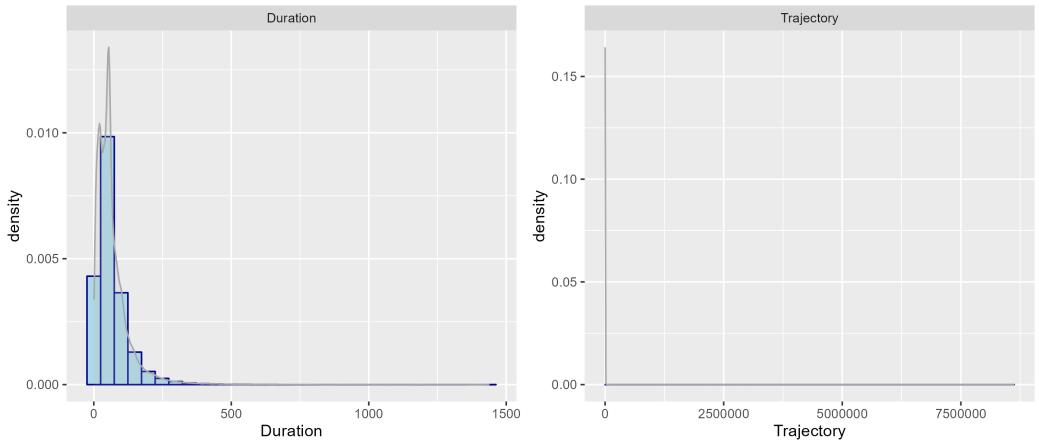

FIGURA 5.2: Histograma de las variables

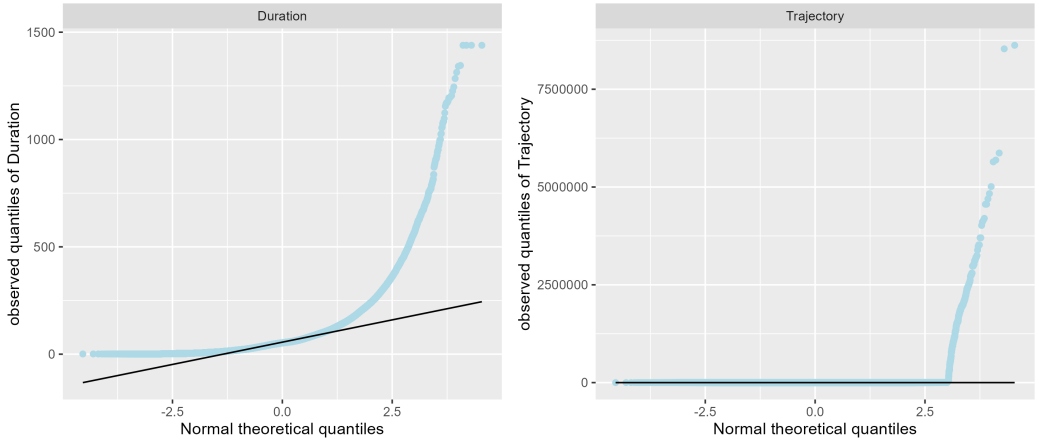

FIGURA 5.3: Gráfico Q-Q de las variables

5.1.1.2.1.1. Transformación Variables

Existen varias funciones para transformar variables. Para este caso se usó el paquete *best-Normalize* [33] de *R* para determinar que transformación usar. El paquete está diseñado para estimar la mejor transformación normalizadora para un vector de forma coherente y precisa. Implementa la *transformación Box-Cox*, la *transformación Yeo-Johnson*, tres

tipos de *transformaciones Lambert* $W \times F$ y la *transformación de normalización cuantil ordenada (ORQ)*.

La función *bestNormalize* selecciona la mejor transformación de acuerdo con una estimación extra-muestra del estadístico P de *Pearson* dividido por sus *grados de libertad (DF)* [33]. Este estadístico P se define como:

$$P = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{5.1}$$

donde O_i es el número observado y E_i es el número esperado (bajo la hipótesis de normalidad) de caer en el *intervalo* i. Los intervalos (o *clases*) se construyen de forma que las observaciones entren en cada uno de ellos con la misma probabilidad según la hipótesis de normalidad. Existen diversas pruebas de normalidad alternativas, pero ésta en concreto es relativamente interpretable como una prueba de bondad de ajuste, y la relación P/DF puede compararse entre transformaciones como una medida absoluta de desviación de la normalidad. Concretamente, si los datos en cuestión siguen una distribución normal, esta relación será cercana a 1 o inferior. La transformación que produce datos con el estadístico de normalidad más bajo es, por tanto, la más eficaz para normalizar los datos, y es seleccionada por *bestNormalize*.

Para ambas variables las transformaciones más adecuadas son la *Transformación OrderNorm* y la *Transformacióon Yeo-Johnson* (Tabla 5.5), coincidiendo ambas en la *Transformación OrderNorm (ORQ)*, descrita en la Sec. 4.2.1.1, como la más apta entre las mencionadas.

Tal como se aprecia en la Tabla 5.6 y en los *histogramas* y *gráficas Q-Q normal* (Figuras 5.4 y 5.5), se puede concluir que ahora las variables transformadas por *ORQ* tienen distribución normal.

TABLA 5.5: Estadístico de normalidad estimada (Pearson P/DF) por bestNormalize

Variable	orderNorm	Box-Cox	Yeo-Johnson
Duración	18.4257	26.7923	27.5077
Recorrido	11.3722	-	208.9191

TABLA 5.6: Descripción de las variables transformadas

Variable	mean	sd	median	trimmed	mad	min	max	skew	kurtosis	se
Duración	0	1.00	0.01	0	0.99	-2.97	4.24	0.01	-0.04	0
Recorrido	0	0.99	0.00	0	1.00	-2.33	4.54	0.04	-0.13	0

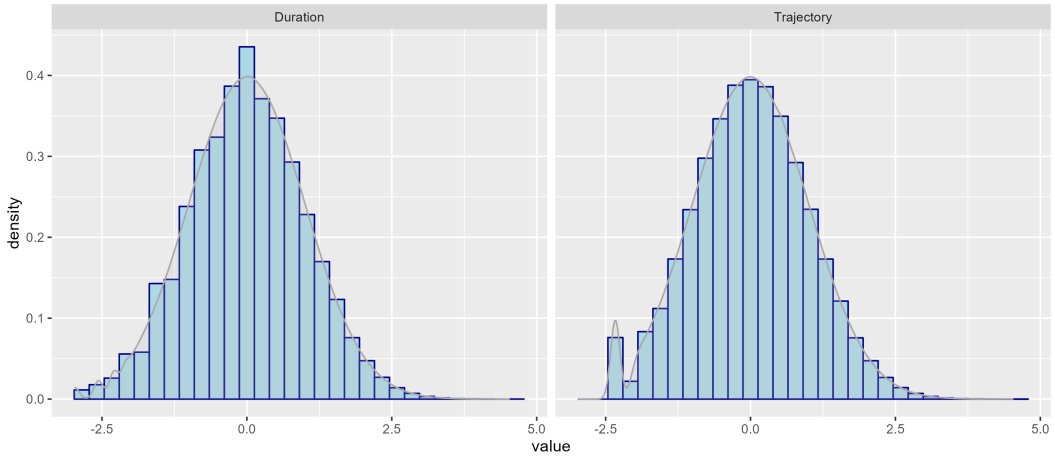

FIGURA 5.4: Histograma de las variables transformadas por *ORQ*

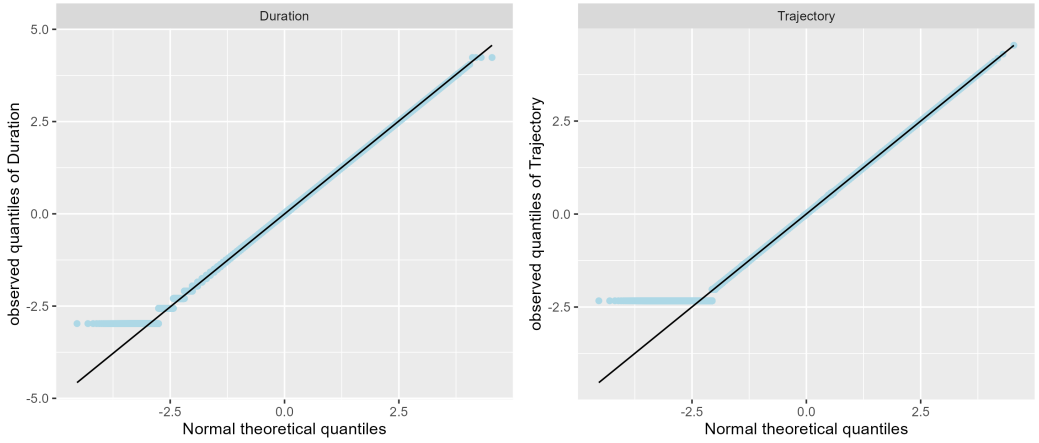

FIGURA 5.5: Gráfico Q-Q de las variables transformadas por *ORQ*

5.1.1.2.1.2. Métodos Multivariados Robustos para detección de Valores atípicos

Se eligen dos métodos multivariados robustos para la detección de valores atípicos: *MCD-Fast* y *MSD* (Sec. 4.2.2), en base a que en [29], se evalúa el rendimiento de cuatro

métodos tolerantes a la asimetría, entre ellos *MCD-Fast* y *MSD*, en datos con distribución asimétrica multivariante, y concluyen que *MSD* es el candidato más prometedor en términos de capacidad de detección de valores atípicos para datos multivariantes, pero que es un algoritmo de fuerza bruta que es costoso computacionalmente. Por otra parte, en (Wada 2004), bibliografía relacionada de [29], *MCD-Fast* tiende a mantener su robustez con datos de cola pesada. Entonces, se usan ambos métodos, con la finalidad de comparar los resultados obtenidos y escoger el más conveniente para este caso.

De la Tabla 5.7, se aprecia que la menor cantidad de observaciones detectadas como valores atípicos en los datos transformados es con el *Método MSD* mediante la Técnica *ORQ*.

Tras la eliminación de las observaciones detectadas como valores atípicos en el conjunto de datos transformados por *ORQ*, se obtienen correlaciones similares, y para ambas no hay evidencia suficiente para decir que no existe asociación lineal entre las variables con un nivel de significancia $\alpha = 0.05$ (Tabla 5.8); e *histogramas* y *gráficas Q-Q normal* son similares (Figura 5.6). Por esta razón, se eligen los resultados obtenidos para el conjunto de datos transformados por *ORQ* y con detección de valores atípicos por *MSD*, pues no existen grandes diferencias entre los resultados obtenidos y *MSD* detecta menos observaciones como atípicas.

Tabla 5.7: Observaciones detectadas como valores atípicos

Método	n	Sin transformación (%)	*ORQ* (%)
MCD-Fast	175554	21.23	11.30
MSD	175554	16.61	7.47

Tabla 5.8: Correlación univariada de las variables transformadas sin valores atípicos

Método	n_{ORQ}	correlación	p value
MCD-Fast	155718	0.91	0
MSD	162448	0.88	0

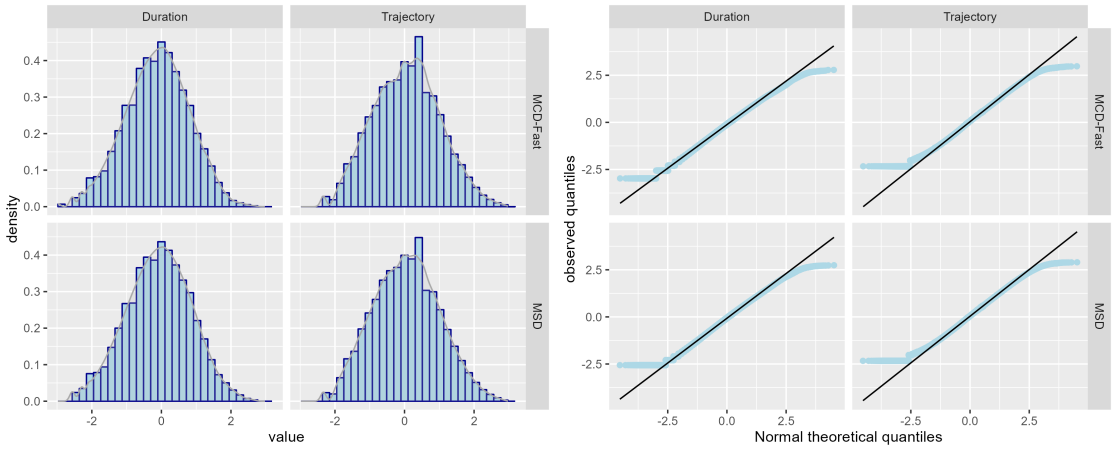

FIGURA 5.6: Histograma y Gráfica Q-Q normal de las variables transformadas sin valores atípicos

5.1.1.2.2 Variables: *Aler_Duracion, Aler_Velocidad*

Para cada Tipo de Alerta $i = 1, \ldots, 7$ se tiene dos variables, *Aler_Duracion* y *Aler_Velocidad*, que fueron tomadas simultáneamente al ocurrir la alerta mientras transcurre el viaje. Para estas variables se usan *Métodos Robustos* para la detección de valores atípicos.

- Para el Tipo de Alerta $i = 6$ (Exceso de Velocidad) se usa *MCD-Fast* y *MSD* (Sec. 4.2.2), con la intención de eligir el que entregue mejores resultados.
- Para el Tipo de Alerta $i = 1, 2, 3, 4, 5, 7$, la variable *Aler_Duración*, en cada observación toma el valor de $2s$. Por tanto, se tiene solo la variable *Aler_Velocidad*, y para ésta se usa el *Método Univariante Robusto*: *Box-plot* [30]. Este método considera como valores atípicos aquellos que están fuera del siguiente intervalo: $(Q_3 + 1{,}5 \cdot IQR, Q_1 - 1{,}5IQR)$, donde $IQR = Q_3 - Q_1$.

Como ya se ha mencionado previamente, las variables deben asemejarse a una *distribución elíptica*, y en este caso, las variables presentan sesgo positivo y una curtosis elevada (Tabla 5.9). En la Tabla 5.10 se presentan las transformaciones recomendadas en cada caso por la función *bestNormalize* de *R*. Así, la *transformación ORQ* es la usada para $i = 6$, y para $i \neq 6$ (caso univariante) se usará la transformación a raíz cuadrada, por simplicidad. A continuación:

- Para el Tipo Alerta $i = 6$, no se perciben grandes diferencias entre los resultados obtenidos por cada uno de los métodos (Tabla 5.11), por tanto nos quedamos

con los resultados obtenidos con *MSD*, simplemente porque consigue resultados similares eliminando menos observaciones.

- Un 1.9 % de las observaciones con Tipo de Alerta $i \neq 6$ se considera valores atípicos. Donde los Tipo de Alerta con mayor cantidad de valores atípicos son: 1 (50 %), 5 (46.76 %), 3 (2.69 %), y 4 (0.04 %), en orden decreciente. Destacándose que el Tipo de Alerta $i = 2$ no prensenta valores atípicos.

Tabla 5.9: Estadísticos descriptivos de las variables alertas

Estadisticos	*Aler_Velocidad*							*Aler_Duración*
	Tipo Alerta							Tipo Alerta
	1	2	3	4	5	6	7	6
Mínimo	1	0	-127	0	0	0	0	0.1
Q_1	50	3	36	41	31	44	0	2
Mediana	50	26	50	50	50	81	8	8
Mean	52.63	30.39	52.62	57.97	53.25	67.4	14.6	36.87
Q_3	52	50	71	77	83	82	24	32
Máximo	273	157	129	129	133	9991	97	2005
skew	3.2	0.8	0.5	0.34	-0.02	102.07	1.1	7.35
Kurtosis	23.21	-0.06	-0.11	-0.65	-1.02	18721.04	0.1	78.13
Trimmed	50.9	26.77	51.44	57.11	53.31	69.21	11.9	17.21
n	2834	93251	16270	1605	10302	185757	12822	170685
NA's	0	0	0	0	0	0	0	15072

Tabla 5.10: Estadístico de normalidad estimada (*Pearson P/DF*) por *bestNormalize*

Tipo Alerta	Variable	*ORQ*	$\sqrt{x}$	Yeo-Johnson
6	*Aler_Duracion*	668.7	666.1	679.4
	Aler_Velocidad	1881.3	1904	1899.2
1, 2, 3, 4, 5, y 7	*Aler_Velocidad*	591.3	617	609.1

Tabla 5.11: Resumen comparativo de ambos Métodos

Método	Observaciones atípicas (%)	Datos limpios	
		Correlación	p value
MCD-Fast	3.7	0.09	0
MSD	1.3	0.05	0

5.1.1.3. Eliminación Missing Values

1. *TipoMovil*: se eliminaron las observaciones con valores perdidos en esta variable.
2. *idalerta*: se eliminaron las observaciones que no tienen un Tipo Alerta registrado.

El conjunto de datos queda compuesto por las variables mencionadas previamente en la Tabla 5.1 con 109295 registros de viajes en 301294 observaciones. No obstante, se conservaron variables con valores perdidos (Tabla 5.12).

Tabla 5.12: Conteo de variables con valores perdidos

Variable	Missing Values (%)	Valores disponibles (n)
Aler_Longitud	62.2 %	113856
Aler_Latitud	62.2 %	113856
Conductor	94.1 %	6506

5.2. Análisis Multivariado

La razón principal por la que se debe analizar un conjunto de datos multivariantes utilizando métodos multivariantes en lugar de examinar cada variable por separado utilizando uno u otro método univariante conocido es que cualquier estructura o patrón en los datos es tan probable que esté implícito en las "relaciones" entre las variables o en la "proximidad" relativa de las distintas unidades como en los distintos valores de las variables; en algunos casos, quizá en ambos [36].

5.2.1. Descripción de Variables

5.2.1.1. Variables numéricas

Se presentan los *boxplot* correspondientes en la Figura 5.7.

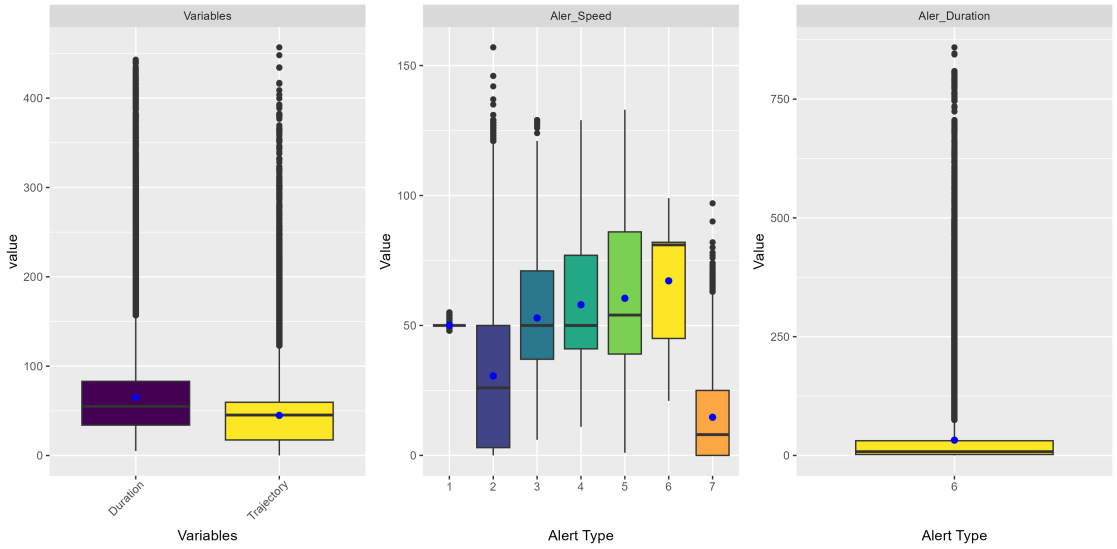

FIGURA 5.7: *Box-plot* de los datos sin valores atípicos.

5.2.1.2. Variables categóricas

- *Tipo Móvil*: los registros de viajes cuentan con 19 tipo diferentes de móvil, categorizados en 2 grandes grupos: Pesados (56.54 %) y Livianos (43.46 %). Hay 5 Tipos que abarcan el 96.74 % de los registros (Tabla 5.13 y Figura 5.8).
- *Móvil*: Hay 2468 patentes diferentes. Un móvil (patente) realiza en promedio 44.5 viajes.
- *Conductor*: Hay 1382 conductores diferentes registrados, donde solo 5.9 % de los viajes cuentan con conductor registrado.
- *idalerta*: son 7 Tipos diferentes, donde del total de registros de alertas (301294), el 0.5 %, 30.6 %, 5.36 % , 0.53 % , 3.01 %, 55.78 %, y 4.23 % corresponden a los tipos 1,2,3,4,5,6 y 7, respectivamente.

5.2.1.3. Variables geoespaciales

- *Aler_Lat* y *Aler_Lon*: de las alertas, el 113856 (62.2 %) tiene las coordenadas registradas. Las restantes alertas (187438) no posee las coordenadas registradas (Tabla 5.12).

Tabla 5.13: Tipo Móvil

Tipo Móvil	n	Tipo Vehículo	Porcentaje
Tractocamión	43517	Pesado 53.46 %	96.7 %
Bus	10607		
Camión	2558		
Minibus	1752		
Camioneta	47273	Liviano 43.25 %	
Bus 2P	892	Pesado 3.08 %	3.11 % (viajes <1 %)
Camión Pluma	706		
Camión Tolva	475		
Furgón	359		
Camión Combustibles	264		
Camión Aljibe	244		
Taxibus	143		
Camión 3/4	131		
Camión Ampliroll	83		
Ambulancia	59		
Camión Llubricador	38		
Camión Baranda	14		
Automóvil	107	Liviano 0.17 %	
Station Wagon	73		

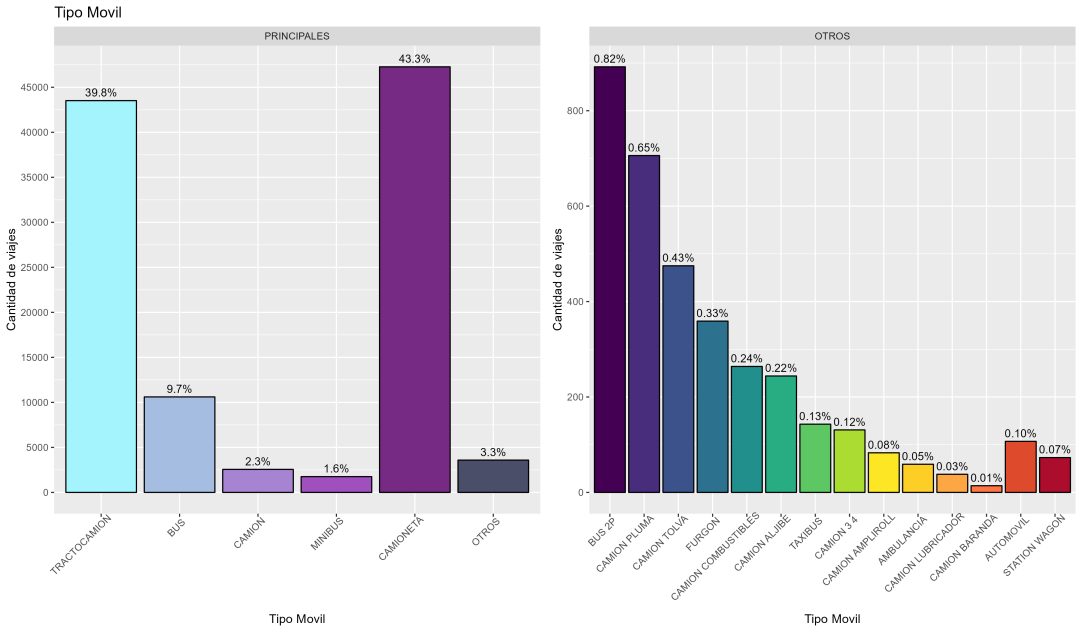

Figura 5.8: Registros de viajes según Tipo de Móvil

5.2.1.4. Variables generadas

- *Velocidad* (*kms/hr*): Velocidad del viaje estimada a partir de las variables *Duración* y *Recorrido*.

- *timea* (*min*): Minutos a los que ocurre la alerta desde que inicia el viaje.
- *Aceleración* (*km/h/s*): Aceleración/Desaceleración estimada mediante la Velocidad estimada obtenida, la Velocidad a la que se registró la Alerta (*Aler_Velocidad*) y la Duración de la Alerta (*Aler_Duración*).
- *Número de alertas*: número de alertas registradas en un viaje.

En la Tabla 5.14 se presenta un resumen descriptivo de estas variables numéricas. De donde, se destaca el hecho de que las infracciones comienzan a ocurrir en un promedio de 35 *min* (desde que se inició el viaje).

TABLA 5.14: Estadísticos descriptivos de las variables generadas

Estadístico	*Velocidad*	*timea*	*Aceleración*	*Número Alertas*
Mínimo	0.2727	0	-490.79	1
Q_1	26.40	13.13	-2.41	2
Mediana	41.49	24.80	1.23	5
Media	40.63	35.36	15.88	9.033
Q_3	54.33	44.62	9.65	13
Máximo	193.29	429.13	731.1	89

5.2.2. Análisis Exploratorio

5.2.2.1. Correlación

La *correlación* es independiente de la escala en la que se miden las variables [36] e indica la fuerza y dirección lineal de la relación lineal y la proporcionalidad entre las dos variables estadísticas. Para este caso, las correlaciones obtenidas (Tablas 5.15 y 5.16) son todas significativas al nivel de $\alpha = 0.05$ (todas con un p value cero). Se destaca la correlación positiva entre *Duración* y *Recorrido*, *Recorrido* y *Velocidad*, *Aceleración* y *Aler_Velocidad* , y *timea* y *Aler_Velocidad*.

TABLA 5.15: Matriz de Correlación

	Duración	*Recorrido*	*Velocidad*	*Núm. Alertas*
Duración	1.000	0.798	0.053	0.052
Recorrido	0.798	1.000	0.529	0.074
Velocidad	0.053	0.529	1.000	0.098
Núm. Alertas	0.052	0.074	0.098	1.000

TABLA 5.16: Matriz de correlación

	timea	*Aler_Velocidad*	*Aceleración*
timea	1.000	-0.128	-0.038
Aler_Velocidad	-0.128	1.000	0.307
Aceleración	-0.038	0.307	1.000

5.2.2.2. Normalidad

Las Figuras 5.9 y 5.10, corresponden a los *gráficos de densidad* de las variables numéricas. De éstas y de los *Boxplot* previos (Figura 5.7) se vislumbra que las variables no tienen distribución normal.

Si los datos tienen más de un "pico" (moda), se considera una *distribución multimodal* [37]. En tal caso, se puede suponer que debe haber otra variable que ayude a explicar por qué de la existencia de estos "picos" [37]. En este caso, quizá las variables categóricas, como Tipo de vehículo, o Tipo de Alerta, pueden ayudar a explicar.

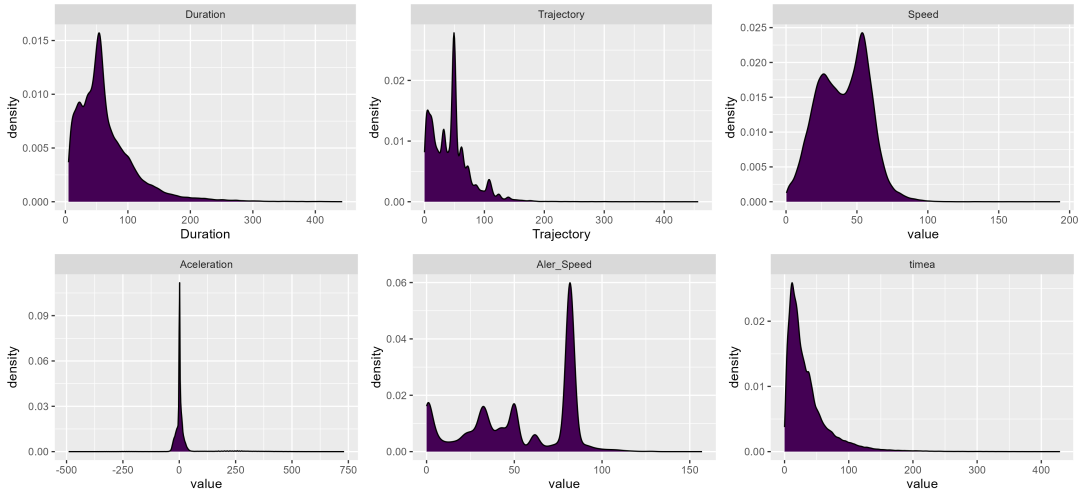

FIGURA 5.9: *Gráfico de Densidad* de las variables numéricas

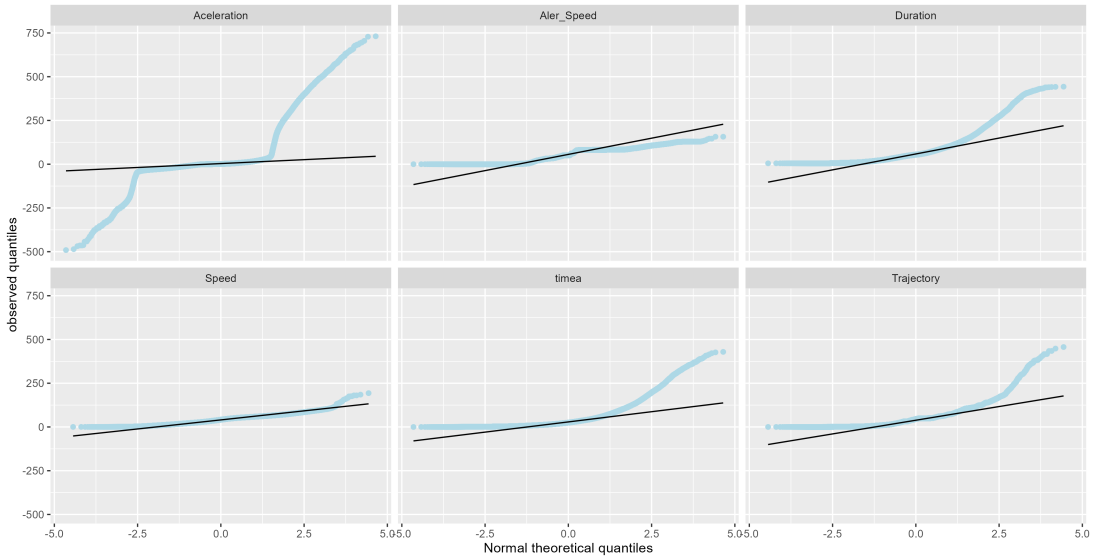

FIGURA 5.10: *Gráficas Q-Q normal* de las variables numéricas

5.2.3. Análisis para Grupos

Considerando que el conjunto de datos presenta variables categóricas (*Tipo Vehículo*, *Tipo Móvil* y *Tipo Alerta*) que dividen las observaciones en diferentes grupos, ya sea con respecto a los diferentes viajes, o las distintas alertas que ocurren, se procede a inspeccionar estos grupos, con la finalidad de comprender mejor las variables numéricas. Sin embargo, se deben considerar los siguientes aspectos previamente [38]:

- Que los grupos sean comparables, es decir, grupos lo suficientemente grandes para ser comparados, de igual tamaño aproximadamente, recogidos más o menos al mismo tiempo, y además que las características que no interesan tengan una distribución aproximadamente igual en todos los grupos.
- Tipo de datos: Distribución normal, o no tienen distribución normal.
- Número de grupos a comparar.
- Tipo de grupos: emparejados o no emparejados.

5.2.3.1. Tipo de Vehículo

Una pregunta natural es si los distintos Tipos de Vehículos pueden diferenciarse a través de las variables *Duración del viaje*, *Recorrido del viaje*, *Velocidad*, y *número de alertas del viaje*, y además saber si difieren entre sí y con respecto a cuál de las variables.

Con respecto, a las consideraciones previas, la variable *Tipo de Vehículo* divide a los datos en liviano (43.42 %) y pesado (56.58 %) (Tabla 5.13); siendo, subconjuntos de tamaño similar. De los gráficos de *densidad* y *boxplot* (Figura 5.11), se aprecia que ambos, digamos grupos, difieren entre sí (exceptuando *Número Alertas*, que hasta la mediana coincide y parecen similares). Los Vehículos pesados, tienen una *Duración*, *Recorrido* y *Velocidad*, mayor que los Vehículos livianos, y tambien levemente, mayor cantidad de alertas por viajes. Todo concuerda con lo que se podía presuponer, pues los Vehículos pesados son constituidos en parte por *Tractocamiones*, *Buses*, *Camiones*, *Minibuses*, etc. con los cuales se puede esperar viajes más largos, y posiblemente recorrido en carreteras. Además, se observa visualmente que éstos dos grupos, no tienen distribución normal para las variables correspondientes (Figura reffig.005:08-09˙densidad˙boxplot˙tveh). Por ende, es adecuado utilizar pruebas no paramétricas. Por otra parte, estos datos son no emparejados; es decir, las observaciones de cada grupo no están relacionadas entre sí (las observaciones de cada grupo proceden de individuos diferentes [38]).

Debido a que los datos son no normales, no se puede implementar el *Análisis de Varianza Multivariante (MANOVA)*, pues los supuestos necesarios no se cumplen (Normalidad Multivariada y Homogeneidad de las matrices de varianza-covarianza); y por tanto, se recurre a la *Inferencia no Paramétrica Multivariada*, la cual a diferencia del *Análisis Multivariado Clásico*, no requiere normalidad multivariante para los datos [40]; las diferentes variables de respuesta pueden incluso medirse en diferentes escalas (binaria, ordinal, cuantitativa), los valores p se calculan para pruebas globales (pruebas de permutación y aproximaciones F) y, se identifican subconjuntos significativos de variables de respuesta y niveles de factor utilizando algoritmos de pruebas múltiples que controlan la tasa de error familiar [40].

Una de las principales razones por las que el *MANOVA clásico* apenas se utiliza, es debido a que los supuestos necesarios son bastante restrictivos y difíciles de verificar [40]. Otra limitación es que, incluso cuando se cumplen los supuestos de normalidad

multivariante, las pruebas *MANOVA* suelen proporcionar respuestas que no son útiles en la práctica: sólo hacen una afirmación global sobre la significación. Los procedimientos *MANOVA clásicos* no proporcionan información coherente sobre qué subgrupos de variables de respuesta o niveles de factor son responsables de la significación global [40].

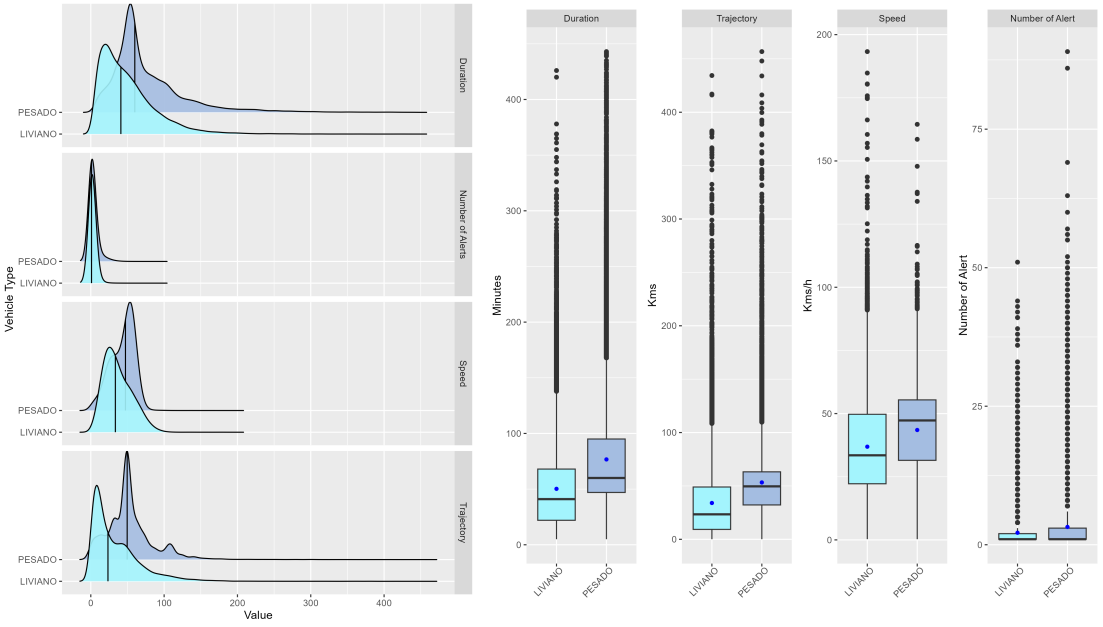

FIGURA 5.11: *Gráfico de Densidad* y *Boxplot* para los *Tipo Vehículo*

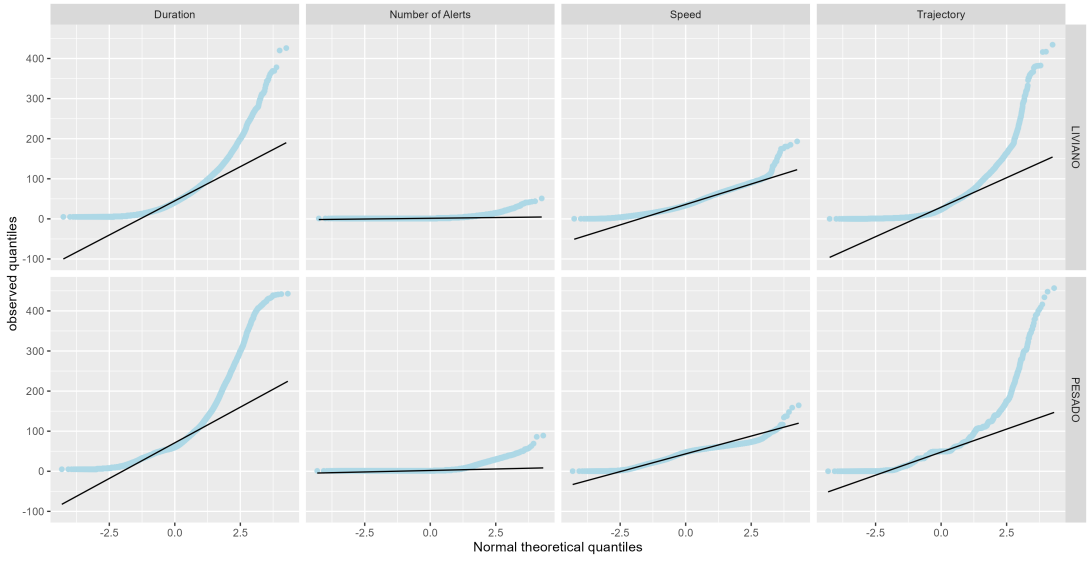

FIGURA 5.12: *Gráfico Q-Q Normal*

5.2.3.1.1 Multivariado no paramétrico

Considerando el modelo no paramétrico, proporcionado por el package *npmv* de *R*, y descrito en la Sec. 4.4, se tiene para este caso, *niveles de factor* $a = 2$ (*liviano* y *pesado*), con *variables de respuesta* $p = 4$ (*Duración*, *Recorrido*, *Velocidad*, y *Número de Alertas durante el viaje*). Usando una muestra aleatoria del 30 % del conjunto de datos, se tienen muestras de tamaño individual de $n_1 = 14306$ y $n_2 = 18489$, respectivamente, y un total muestral de $N = 32795$. Las diferentes variables se denotan por $k = 1, \ldots, p$, las muestras (tratamientos, subpoblaciones, niveles de factor) por $i = 1, \ldots, a$, los n_i sujetos (unidades experimentales), sobre los que se realizan las observaciones de p-variables, se indexan por $j = 1, ..., n_i$. Además, se establece que los vectores de observación multivariantes $\boldsymbol{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(p)})^T$ son independientes y que dentro del mismo nivel de factor i, siguen la misma distribución p-variable: $\boldsymbol{X}_{ij} \sim F_i$.

- Hipótesis Estadísticas Globales. Plantear la pregunta de si los Tipos de Vehículos: *Pesados* y *Livianos*, proceden de la misma población (distribución multivariante), se formula: $H_0 : F_1 = F_2$.

  La Tabla 5.17 muestra los resultados de los estadísticos de prueba no paramétricos elegidos con sus estadísticos de prueba, los grados de libertad del numerador y el denominador, y los valores p para cada estadístico utilizando tanto el método de *aproximación F* como el de *permutación* (aleatorización). Concluyéndose que las diferencias entre las distribuciones multivariantes son significativas con un $\alpha = 0.05$.

- Efectos Relativos no paramétricos. Si se rechaza una hipótesis global, es esencial saber cuál de todas las p variables mostraron diferencias significativas, y cuál de los niveles de factor contribuyó al resultado significativo. Los efectos relativos cuantifican las tendencias observadas en los datos en términos de probabilidades. En general, los *efectos mínimos* y *máximos* posibles para el grupo i son $\frac{n_i}{2N}$ y $1 - \frac{n_i}{2N}$, respectivamente [40].

  La Tabla 5.18 enumera los efectos relativos empíricos no paramétricos y el efecto mínimo posible de los Tipo de Vehículo. Notándose que el Tipo *liviano* cuenta con el menor *efecto mínimo* posible.

Cada una de las variables discrimina, exceptuando *número de alertas*, entre *liviano* y *pesado*. Por ejemplo, la probabilidad de que una medida elegida al azar de *Recorrido* del Tipo *pesado* sea mayor que una observación elegida al azar de la muestra completa, incluyendo al tipo *pesado*, es 0.7, que es muy cercano al máximo efecto posible para este grupo.

Tabla 5.17: Estadísticos Descriptivos de la Hipótesis Global

	Test Statistic	df1	df2	P-value Permutation	Test p-value
ANOVA type test p-value	2479.732	2.492	80376.01	0	0
McKeon approx. for the Lawley Hotelling Test	1202.879	4.000	32790.00	0	0
Muller approx. for the Bartlett-Nanda-Pillai Test	1202.842	4.000	32790.00	0	0
Wilks Lambda	1202.879	4.000	32790.00	0	0

Tabla 5.18: Efectos Relativos

Tipo Vehiculo	*Duración*	*Recorrido*	*Velocidad*	*Núm. Alertas*	*Efecto Mínimo*
Liviano	0.31327	0.29751	0.37803	0.46415	0.2181125
Pesado	0.68673	0.70249	0.62197	0.53585	0.2818875

A pesar de que Tipo de Vehículo discrimina bien entre un tipo de viaje registrado, aún las variables *Duración*, *Recorrido*, y *Velocidad*, no son normales en cada grupo (*liviano*, *pesado*), lo que siguiere que aún puede existir otra variable afectando, o quizá que está clasificación se puede mejorar, o subdividir.

5.2.3.2. Tipo Alerta

Para esta variable, los diferentes grupos $i = 1, \ldots, 7$, no son de tamaño similar (las alertas del Tipo $i = 2, 6$ corresponden al 30 %, y 55.78 % de todas las alertas registradas). Sin embargo, el paquete *npmv* de *R* no menciona como restricción la diferencia de tamaño. Por tanto, estableciendo el modelo paramétrico con *niveles de factor* $a = 7$, y *variables de respuesta* $p = 3$ (*Aler_Velocidad*, *timea*, y *Aceleración*) sobre una muestra estratificada (*Duración*) del 50 %, se obtienen muestras de tamaño individual de $n_1 = 1506$, $n_2 = 92184$, $n_3 = 16148$, $n_4 = 1602$, $n_5 = 9058$, $n_6 = 168064$, y $n_7 = 12732$, respectivamente, y un total muestral de $N = 301294$. Estableciendo nuevamente, que

los vectores multivariantes son independientes y que dentro del mismo *nivel de factor* i siguen la misma distribución p-variable: $\boldsymbol{X_{ij}} \sim F_i$, se tiene:

1. Hipótesis Globales: $H_0 : F_1 = \ldots = F_7$. Los resultados obtenidos, indican diferencias entre las distribuciones multivariantes a unn nivel de significancia $\alpha = 0.05$. Rechazándose la hipótesis Nula (Tabla 5.19).
2. Efectos Relativos no paramétricos. En este caso, las variables *timea*, y *Aceleración*, no muestra muchas diferencias para todos los diferentes *niveles de factor*. Por ejemplo, la probabilidad de que una medida elegida al azar de *Aceleración* del *Tipo* 4, 5 y 6 sea mayor que una observación elegida al azar de la muestra completa, incluyendo al Tipo 4, 5, y 6, respectivamente, es 0.6 (Tabla 5.20).

Tabla 5.19: Estadísticos Descriptivos de la Hipótesis Global

	Test Statistic	df1	df2	P-value Permutation	Test p-value
ANOVA type test p-value	2056.044	13.78	72303.97	0	0
McKeon approx. for the Lawley Hotelling Test	10043.466	18	602564	0	0
Wilks Lambda	8951.082	18	852163.15	0	0

Tabla 5.20: Efectos Relativos

Tipo Alerta	*Aler_Velocidad*	*timea*	*Aceleración*	*Efecto Mín.*
1	0.47367	0.45734	0.67664	0.00250
2	0.29948	0.58295	0.36909	0.153
3	0.49208	0.50337	0.56436	0.0268
4	0.53282	0.60977	0.62096	0.00266
5	0.57385	0.61068	0.63406	0.0150
6	0.63270	0.45185	0.58499	0.279
7	0.15660	0.44323	0.11286	0.0211

La Figura 5.13, tiene la función de densidad de las variables *Aler_Velocidad* y *aceleración (kms/hr/s)* para cada tipo de alerta ($i = 1, \ldots, 7$), de donde se destaca que las variables no tienen distribución normal (razón de usar el método no paramétrico). Una razón, puede deberse a que estos tipos de alertas, no son registrados en función de las variables mencionadas, a excepción del exceso de velocidad ($i = 6$), sino en función de otros factores, por ejemplo la distancia a un vehículo sin movimiento (para $i = 2$), la perdida

de detección facial (para $i = 3$), aceleración/desaceleración (para $i = 1, 7$), donde la variable *aceleración* para estos dos tipo de alertas, se observa que a pesar de no ser normales, sino más leptocúrticas ambos tipos se diferencian bien.

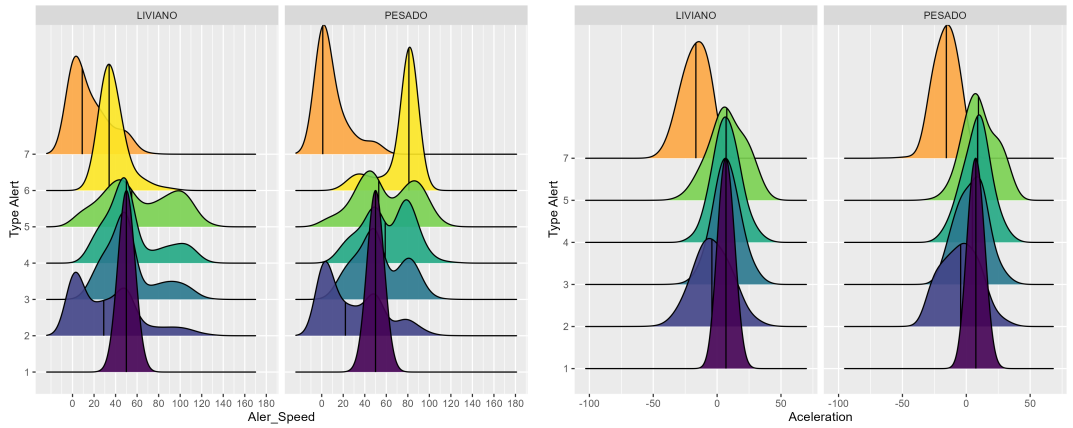

FIGURA 5.13: Gráfico de Densidad: Tipo Alerta

5.3. Indicadores Individuales

Los *indicadores individuales* para el *ICR* son *Duration Score* y *Severity Score* (DS_i y SS_i, $i = 1, \ldots, 7$), de los cuales se obtiene el *indicador individual Violation Score* (VS_i) (Sec. 3.1). Dado que se debe decidir, si la estructura anidada del *índice (indicador compuesto)* está bien definida y si el conjunto de *indicadores individuales* disponibles es suficiente y/o adecuado para describir el fenómeno, en esta sección se revisan primeramente los resultados entregados por los *indicadores individuales*, y luego se realiza *PCA*, con tal de analizar su estructura y correlación entre ellos. La decisión se puede basar en la opinión de los expertos y en la estructura estadística del conjunto de datos [13].

En la práctica, el *ICR* se obtiene de la suma de los indicadores DS_i ponderados. Por lo que, en las Tablas 5.21 y 5.22, se muestran solo los resultados para este *indicador individual* en los diferentes Tipos de Alerta. Se destaca que para los diferentes viajes, la mayoría de estos obtiene un DS_i alto, donde mientras mas cercano a 100[3] , más seguro, y más cercano a 0, es un viaje más riesgoso. Solo los DS_j, $j = 4, 5$ están haciendo distinción entre los viajes según el número de alertas del tipo $i = 4, 5$ registradas.

[3]Cuando se tiene DS_i, SS_i, y VS_i para $i = 1, \ldots, 7$ igual a 100, es porque el tipo de alerta i no fue registrada durante el viaje.

TABLA 5.21: Resultados para DS_i, $i = 1, 2, 3, 6, 7$.

DS_i	Resultados	Viajes (n)	Viajes (%)
1	0	5	0.01
	(99,100]	109290	99.99
2	0	1	0.01
	(93,100]	109294	99.99
3	[98.1,100]	109295	100
6	0	4943	4.5
	(89,96]	3653	3.3
	(96,100]	100699	92.1
7	0	26	0.024
	(99,100]	109269	99.9

TABLA 5.22: Resultados para DS_i, $i = 4, 5$.

Viajes	DS_4			DS_5					
	0	50	100	0	20	40	60	80	100
n	135	1282	107878	336	183	325	872	3206	104373
%	0.12	1.17	98.7	0.31	0.17	0.3	0.8	2.93	95.5

5.3.1. Indicadores individuales como datos composicionales

Tal como se menciona en la Sec. 4.1 (Etapa: *Análisis Multivariado*), para realizar *PCA* sobre los *indicadores* (DS_i, $i = 1, \ldots, 7$), éstos se consideran como *datos composionales* (descritos en la Sec. 4.3). Esta característica añade una complejidad adicional a las cuestiones de escala, ya que son objetos multivariables en sí mismos. En una *composición*, no se puede tratar una variable aislada porque sus valores sólo tienen sentido en relación con las cantidades de las demás variables. Así pues, toda la *composición* debe considerarse como un objeto único con alguna escala con propiedades conjuntas, como sumar 1 o 100 %, ser positiva y obedecer a una ley de razón de diferencia [44].

Cada *indicador* puede tomar valores desde 0 % a 100 %, donde un 100 % corresponde a la no ocurrencia del tipo de alerta correspondiente durante el viaje. Entonces, la suma de los 7 indicadores DS_i, $i = 1, \ldots, 7$ toma como máximo 700 %, y como mínimo 0 %, siendo todos solo valores positivos. No obstante, la suma de cada viaje no es constante. Por ende, según las directrices para la elección de escala adecuada sugeridas en [44],

a consecuencia de la no importancia de la suma total y escala relativa, la escala es *Composicional de Aitchison (acomp)*. Además, por ser proporciones, se debe forzar la *cerradura* (Def. 13) [42].

El package *compositions* de *R* fue usado para analizar los datos como composiciones.

5.3.1.1. PCA sobre los indicadores

El *PCA* es una interpretación de la *Descomposición de Valores Singulares (SVD)* (Sec. 4.3.3.2) de la matriz de datos *clr-transformada*. Con el comando *princomp* del package *compositions* sobre el conjunto de datos con la *escala composicional de Aitchison (acomp)* se obtiene el *PCA*, donde el objeto contiene los elementos: loading (la matriz $\boldsymbol{V}$), sdev (vector con los elementos de la diagonal de $\boldsymbol{D}$ en orden decreciente de magnitud), y scores (la matriz $\boldsymbol{U}$).

Por otra parte, al obtener la correlación[4] de DS_i, $i = 1, \ldots, 7$, considerando su carácter composicional, se nota que la asociación lineal de los indicadores 1, 2, 3 y 7 es la máxima posible. Y que estos mismos, tienen correlación negativa con los *indicadores* 4 y 5, y correlación positiva con el indicador 6. Además, los indicadores 4 y 5 tienen correlación negativa con el indicador 6 (Tabla 5.23).

Tabla 5.23: Matriz de Correlación y Matriz de Variación

DS_i	Correlación							Variación						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	1	1	1	-0.4	-0.71	0.79	1	0	0	0	0.006	0.011	0	0
2	1	1	1	-0.4	-0.71	0.78	1	0	0	0	0.006	0.011	0	0
3	1	1	1	-0.4	-0.71	0.79	1	0	0	0	0.006	0.011	0	0
4	-0.4	-0.4	-0.4	1	-0.34	-0.36	-0.4	0.006	0.006	0.006	0	0.016	0.006	0.006
5	-0.71	-0.71	-0.71	-0.34	1	-0.62	-0.71	0.011	0.011	0.011	0.016	0	0.011	0.011
6	0.79	0.78	0.79	-0.36	-0.62	1	0.79	0	0	0	0.006	0.011	0	0
7	1	1	1	-0.4	-0.71	0.79	1.00	0	0	0	0.006	0.011	0	0

Si las variables originales no están correlacionadas, el análisis carece de valor. En cambio, puede obtenerse una reducción significativa cuando las variables originales están muy correlacionadas, positiva o negativamente [13].

[4] La correlación se basa en varianzas y covarianzas que se definen para el espacio euclídeo y no para el Simplex, por lo que el análisis de correlación estándar para datos composicionales conduce a propiedades indeseables como la dependencia de la escala y la incoherencia subcomposicional [50]. Para más detalles revisar [50], [51], y [52].

En la Tabla 5.24 se presenta la desviación estándar, varianza y varianza acumulativa de cada componente.

TABLA 5.24: PCA: Resumen de las Componentes Principales (CP)

CP	Standard deviation	Proportion of Variance (%)	Varianza Acumulativa (%)
1	0.0967	65.7	65.7
2	0.0687	33.2	98.92
3	0.0123	1.06	99.98
4	0.00113	0.009	99.99
5	0.0004845	0.0016	99.99
6	0.0002911	0.0005	100

5.3.1.1.1 Biplots

Los *biplots* de *varianza* y *forma* en la Figura 5.14, conservan un 98.92 % de variabilidad explicada (las primeras 2 *CP*) (Tabla 5.24), donde el *biplot de covarianza* puede interpretarse como sigue (Sec. 4.3.3.3):

- Dado que las puntas de flechas de DS_i, $i = 1, 2, 3, 6, 7$ están superpuestas, y el largo del enlace es "corto", se puede decir que estos indicadores son proporcionales, con una relación logarítmica casi constante (lo que en la matriz de variación (Tabla 5.23) se traduce como una entrada de pequeño valor). Es decir, forman una subcomposición con baja varianza métrica. En efecto, la varianza métrica total es de 0.01423, y la varianza métrica de la subcomposición mencionada es de 0.00014 (1 % de la varianza total). Además, la varianza métrica de la subcomposición con los DS_i, $i = 4, 5$ es de 0.008 (55.7 % de la varianza total).
- No hay enlaces de un gran largo; es decir, gran variación en la *matriz de variación*, pero hay 3 rayos que apuntan a diferentes direcciones: DS_i, $i = 1, 2, 3, 5$, DS_4, y DS_5.
- Dado que el ángulo entre dos flechas se aproxima al valor de la correlación entre las dos relaciones logarítmicas:

  1. DS_i, $i = 1, 2, 3, 6, 7$ tienen un ángulo casi ortogonal con DS_4, es decir, son mínimamente correlacionadas.

2. DS_i, $i = 1, 2, 3, 6, 7$ están situados uno sobre otro (situados sobre una línea común); es decir, tienen correlación perfecta, y se espera que la subcomposición formada por estas partes muestre un patrón unidimensional de variación, lo que efectivamente se puede comprobar de la Tabla 5.21, pues los resultados por estos indicadores son todos cercanos al 99-100 %.

Para el *biplot de forma*, los indicadores mejor representados son DS_i, $i = 4, 5$, pues la longitud es cercana a la unidad. Lo que tiene sentido, pues ambos indicadores por definición, aseguran tomar 3 y 6 valores distintos, respectivamente (Tabla 5.22).

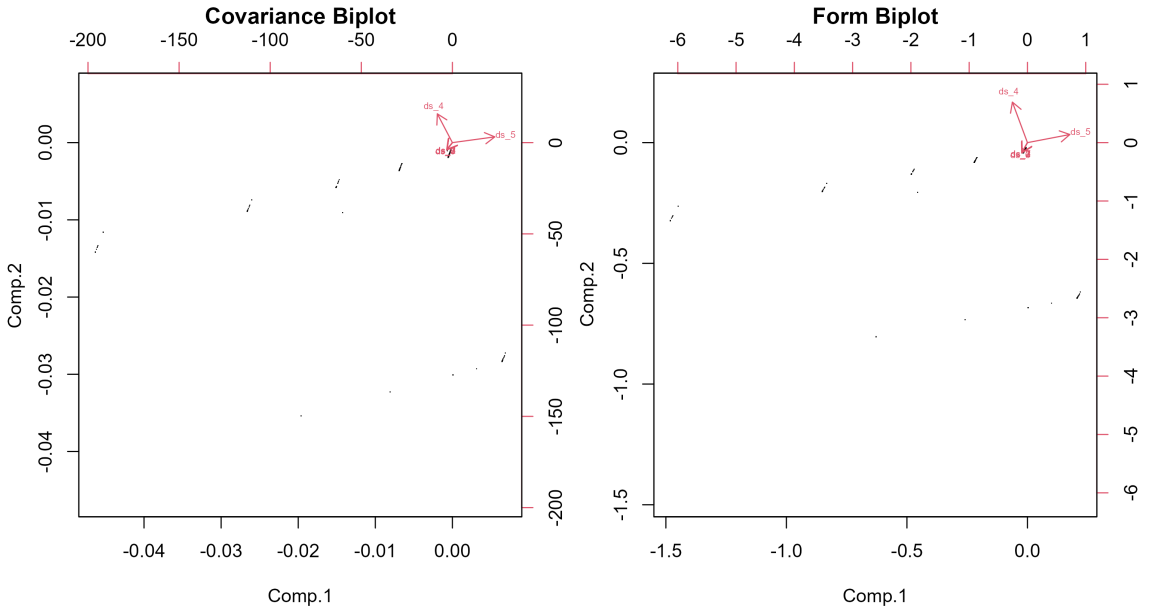

FIGURA 5.14: Biplots composicionales

Capítulo 6

Construcción del Índice

En este capítulo se exponen los resultados obtenidos para los indicadores *Duration Score* y *Severity Score* en las diferentes simulaciones propuestas y modificaciones sugeridas. Seguidamente, de las etapas de *Normalización*, *Ponderación* y *Agregación*, y *Análisis de Sensibilidad* y *Robustez*.

6.1. Modificaciones Establecidas a ICR

De los resultados previos, es evidente que se necesita comprender por qué estos indicadores DS_i no están evaluando distintivamente entre los diferentes viajes, pues a pesar de que los viajes tienen diferente número de infracciones se sigue obteniendo valores cercanos a 100 %.

6.1.1. Duration Score

El fallo evidente de los indicadores DS_i, $i = 1, 2, 3, 6, 7$ (Ec. 3.1) es que cada viaje obtiene un porcentaje cercano a 100 %. En otras palabras, $timeinfr_i/timetot$ es un valor demasiado pequeño. Efectivamente, la media de la duración total de las infracciones del mismo tipo en un viaje corresponden al 0.11 %, 0.13 %, 0.1 %, 6 %, y 0.11 %, respectivamente para $i = 1, 2, 3, 6, 7$. En consecuencia, se propone la siguiente modificación para

este indicador:

$$DS_i = \begin{cases} 0 & , \quad si \quad timeinfr_i \geq DF_i \cdot timetot \\ \left(1 - \dfrac{timeinfr_i}{DF_i \cdot timetot}\right) \cdot 100, & e.o.c \end{cases} \tag{6.1}$$

para $i \neq 4, 5$.

En otras palabras, con la finalidad de que a la unidad se le reste un valor más grande, se propone usar un porcentaje de la duración del viaje. Y tal como se establecía originalmente, conservar la condición de $\dfrac{timeinfr_i}{DF_i \cdot timetot} < 1$, y así que $DS_i \in [0, 100)$. Por tanto, es fundamental la elección del valor DF_i (*Duration Factor* para la infracción del tipo i).

En la Tabla 6.1 se ilustran los valores testeados. Nótese, que $DF1$ corresponde a la media de $timeinfr/timetot$.

Tabla 6.1: Factor de Duración (DF)

Tipo Alerta	$DF1$ (%)	$DF2$ (%)	$DF3$ (%)
Aceleración	0.11	0.3	0.4
Posible Colisión Frontal	0.13	0.3	0.6
Distracción	0.1	0.4	0.5
Fatiga	-	-	-
Salida de Carril	-	-	-
Exceso de Velocidad	6	8	10
Desaceleración	0.11	0.3	0.58

Dado que DS_4, el indicador correspondiente a la Fatiga (Ec. 3.2), solo toma 3 valores (0, 50, 100), la siguiente modificación se realiza con el objetivo de ampliar los valores que puede tomar este indicador, pues un viaje con 2 alertas de este tipo está siendo catalogado igual que un viaje con 3, 4, o 5 alertas.

En general, un viaje tiene entre 0-2 alertas de este tipo (Tabla 6.2), y considerando solo los viajes con esta alerta registrada, la media de sucesos registrados es 1.13 en un viaje, por lo que más de 3 alertas se debería considerar muy riesgoso. Por otra parte, suponer que la fatiga del conductor aumente mientras más amplio sea el lapsus de duración del viaje, se respalda con los datos, pues al obtener la media de la duración de los viajes según el número de fatigas de registradas (Tabla 6.2) se nota que con el aumento de

fatigas, también aumenta el promedio del viaje, exceptuando desde la cuarta fatiga en adelante.

TABLA 6.2: Cantidad de Alertas del Tipo $i = 4$

Cantidad	Viajes (n)	Viajes (%)	Duración (mean)
0	107878	98.7	65
1	1282	1.17	76.3
2	99	0.091	88.1
3	26	0.024	94.7
4	7	0.006	136
5	2	0.002	48.5
6	1	0.001	189

Se propone el siguiente indicador DS_4:

$$DS_4 = \begin{cases} \dfrac{100 - 30 \cdot n_4}{2}, & si \quad n_4 < 4 \\ 0 \qquad\quad , & e.o.c \end{cases} \tag{6.2}$$

Así, cada viaje con $n_4 \geq 4$ fatigas, tendrá un $DS_4 = 0$. Recuérdese que n_4 corresponde al número de fatigas registradas en un viaje.

Para DS_5 no se propone ninguna modificación, pues toma 6 valores posibles, donde el 95.5 %, 2.93 %, 0.8 %, 0.3 %, 0.17 %, y 0.31 % de los viajes tiene $DS_5 = 100$, 80, 60, 40, 20, y 0, respectivamente (Tabla 5.22).

6.1.2. Severity Score

Este indicador (Ecs. 3.4 y 3.6) se deberían obtener según lo establecido para las alertas del tipo $i = 1, 6, 7$. Sin embargo, en la práctica, este indicador no se determina numéricamente, debido a que los sistemas de registro no entregan la aceleración de la infracción realizada, sino solo la velocidad a la que ocurre la infracción.

Usando la variable *aceleración*, generada de la velocidad estimada ("promedio") del viaje, la velocidad de la infracción, y un tiempo establecido de $2s$ para las alertas del tipo $i = 1, 7$ se chequea cómo evalúa este indicador a los viajes. Lo primero a destacar, de los resultados entregados en la Tabla 6.3, es que SS_i, $i = 1, 7, 6$ obtiene valores

mayores a 100 %. La razón de este hecho, es por la fórmula de los indicadores, pues si $v_{infr_6} \in (v_{p_6}, SF_6 \cdot v_{p_6})$, y $a_{infr_i} \in (a_{p_i}, SF_i \cdot a_{p_i})$, $i = 1, 7$ se obtiene un valor distinto de cero y menor a cien. Sin embargo, existen registros de velocidades/aceleraciones de menor valor a la velocidad/aceleración permitida (Tabla 3.2).

De la Tabla 6.3, nótese que para las infracciones del tipo $i = 1, 7$ sólo el 0.9 % y 4.09 % de los viajes tiene aceleraciones para las infracciones menores a las permitidas, pero que estas aceleraciones son bastante menores, pues $SS_i > 100$, $i = 1, 7$. Para el caso de las infracciones $i = 6$, el porcentaje de viajes con velocidad de infracciones menores es mayor.

TABLA 6.3: Resultados SS_i, $i = 1, 6, 7$ para los viajes

Tipo Alerta i	SS_i	Viajes (n)	Viajes (%)
1	[0, 100]	108310	99.1
	(100, 1765]	985	0.9
6	[0, 100]	74933	68.6
	(100, 787.5]	34362	31.4
7	[0, 100]	104823	95.9
	(100, 1766.7]	4472	4.09

Para evitar el problema que se da cuando $v_{infr_6} \leq v_{p_6}$ y $a_{infr_i} \leq a_{p_i}$ para $i = 1, 7$, se sugiere el siguiente *Severity Score*:

$$SS_6 = \begin{cases} 0 & , \quad si \quad v_{infr} \geq SF_6 \cdot v_p \\ \left(1 - \dfrac{v_{infr}}{SF_6 \cdot v_p}\right) \cdot 100, & e.o.c \end{cases} \tag{6.3}$$

$$SS_2 = \begin{cases} 100 & , \quad si \quad v_{infr} < v \\ 0 & , \quad si \quad v_{infr} \geq SF_2 \cdot v_p \\ \left(1 - \dfrac{v_{infr}}{SF_2 \cdot v_p}\right) \cdot 100, & e.o.c \end{cases} \tag{6.4}$$

con $v = 20\ km/h$.

Para las Alertas del Tipo $i = 1, 7$ se sugieren dos opciones:

$$SS1_i = \begin{cases} 0 & , \quad si \quad a_{infr_i} \geq SF_i \cdot a_{p_i} \vee a_{infr_i} = 0 \\ \left(1 - \dfrac{a_{infr_i}}{SF_i \cdot a_{p_i}}\right) \cdot 100, & e.o.c \end{cases} , \quad i = 1, 7 \quad (6.5)$$

$$SS2_i = \begin{cases} 0 & , \quad si \quad |v_{infr_i} - v_v| \geq SF_i \\ \left(1 - \dfrac{|k \cdot v_{infr_i} - v_v|}{SF_i}\right) \cdot 100, & e.o.c \end{cases} , \quad i = 1, 7 \quad (6.6)$$

donde v_v corresponde a la velocidad llevada antes del evento de activación y $k > 1$ para evitar los problemas dados cuando $v_{infr_i} = v_v$. Así, se estableció $k = 1.01$.

La razón de sugerir dos alternativas ($SS1_i$ y $SS2_i$) para las infracciones $i = 1, 7$, es dado a que la aceleración es la diferencia entre las velocidades divida por el tiempo que se mantuvo este cambio de velocidad, por lo que es más simple obtener un SS_i, $i = 1, 7$ a partir de las velocidades entregadas por el sistema ADAS, y evitar el exceso de cálculos. Así, $SS2_i$, $i = 1, 7$ sugiere penalizar según la diferencia entre las velocidades; así mientras mayor diferencia, se obtendrá un menor $SS2_i$. Ahora, para $SS1_1$, $i = 1, 7$ mientras más próxima este la aceleración de la infracción a $SF \cdot a_{infr}$, entonces mayor es el descuento realizado, considerando una aceleración pequeña, como menos grave aunque se aleje más de la a_p.

Nótese además que se sugiere un *Severity Score* para la infracción $i = 2$. La razón de esto es lograr una mayor variación en la información entregada para este tipo de alerta, pues no se diferencia en gran medida del tipo $i = 3$. Por otra parte, la condición de $SS_2 = 100$ cuando $v_{infr} < 20\ km/h$ se debe a que la posible colisión frontal ocurre cuando la distancia limite establecida a un vehículo sin movimiento se sobrepasa, y a velocidades tan bajas, se puede deber a sucesos registrados a la salida de un estacionamiento o similares. La estructura de este indicador es similar al SS_6 sugerido. Es decir, mientras más cercano este v_{infr_i} a $SF_i \cdot v_{infr_i}$, $i = 2, 6$, mayor es el descuento realizado al indicador.

Finalmente, hay que destacar que se debe establecer nuevos valores para los *Severity Factors* SF_i, $i = 1, 2, 6, 7$. Así, los valores iniciales para SF_i, es decir $SF1_i$ (primer grupo testeado), corresponden:

1. A la media de v_{infr_i}/v_{p_i} para $i = 2, 6$.
2. A la media de a_{infr_i}/a_{p_i} para $i = 1, 7$ en $SS1$.
3. A la media de $|v_{infr_i} - v_v|$ para $i = 1, 7$ en $SS2$.

En la Tabla 6.4 se muestran los valores de *Severity Factor* testeados para cada indicador SS_i con $i = 2, 6$, y $SS1_i$ y $SS2_i$ con $i = 1, 7$. Nótese que $SF2_i$ y $SF3_i$ son incrementos de $SF1_i$ basados en la *desviación estándar* de:

1. v_{infr_i}/v_{p_i} para $i = 2, 6$.
2. a_{infr_i}/a_{p_i} para $i = 1, 7$ en $SS1$.
3. $|v_{infr_i} - v_v|$ para $i = 1, 7$ en $SS2$.

obtenidos sobre el conjunto de datos. Por ende, cuando la *desviación estándar* es pequeña, el incremento para $SF2_i$ y $SF3_i$ es mayor, dado que en caso contrario, los resultados sobre los indicadores SS_i con $i = 2, 6$, $SS1_i$ con $i = 1, 7$, y $SS2_i$ con $i = 1, 7$ no presentan diferencias notorias.

Tabla 6.4: Factor de Severidad (SF)

i	Tipo Alerta	Indicador	$SF1$ (%)	$SF2$ (%)	$SF3$ (%)
2	Posible Colisión Frontal	SS	102	110	112
6	Exceso de Velocidad	SS	77	80	81
1	Aceleración	SS1	61	65	76
7	Desaceleración	SS1	118	120	124
1	Aceleración	SS2	1430	1500	1580
7	Desaceleración	SS2	3460	3480	3600

6.2. Simulaciones

A continuación, se ilustran los resultados obtenidos para DS_i, $i = 1, 2, 3, 4, 6, 7$ ($i \neq 5$) (Ecs. 6.1 y 6.2), y *Severity Score* (es decir, SS_i con $i = 1, 2$, y $SS1_i$, $SS2_i$ con $i = 1, 7$) para los diferentes valores propuestos de *Factor de Duración (DF)* (Tabla 6.1), y *Factor de Severidad (SF)* (Tabla 6.4), respectivamente.

6.2.1. Duration Score

De la Tabla 6.5, que muestra la cantidad y porcentaje de viajes con DS_i, $i = 1, 2, 3, 6, 7$ entre 0, (0, 50], (50, 100), y 100 (cuando el viaje no registra la alerta del tipo i), se aprecia que, con las modificaciones establecidas al indicador *Duration Score* DS_i:

- Ahora hace distinción entre los diferentes viajes.
- A medida que se va aumentando *Duration Factor DF*, los viajes con $DS_i = 0$, $i = 1, 2, 3, 6, 7$ disminuyen, pero sin desaparecer la distinción entre los viajes, pues, de la Figura 6.1, se concluye que aquellos con un $DS_i \in (0, 50]$ tienen una media mayor de ocurrencias de infracciones del tipo $i = 1, 2, 3, 6, 7$ en relación a viajes con $DS_i \in (50, 100)$. Es más, particularmente para DS_1, la media de duración de los viajes para $DS_1 = 0$, $DS_1 \in (0, 50]$, y $DS_1 \in (50, 100)$ con *DF1* es de 26, 54.67, y 109 minutos, respectivamente; con *DF2* es de 12.8, 27, y 80 minutos, respectivamente; y con *DF3* es de 10, 20.4, y 76.2 minutos respectivamente. En consecuencia, DS_1 está dando un menor valor a los viajes de poca duración y muchas infracciones del tipo $i = 1$. Que efectivamente, es el objetivo buscado. Y este fenómeno ocurre también con las infracciones del tipo restantes.
- Para DS_4 (Ec. 6.2) se destaca el hecho de que los viajes con $DS_4 = 0$ obtenidos con DS_4 original (Ec. 3.2) se subdividieron en viajes con $DS_4 = 0, 5, 20$, tal como se esperaba (Tabla 6.6).

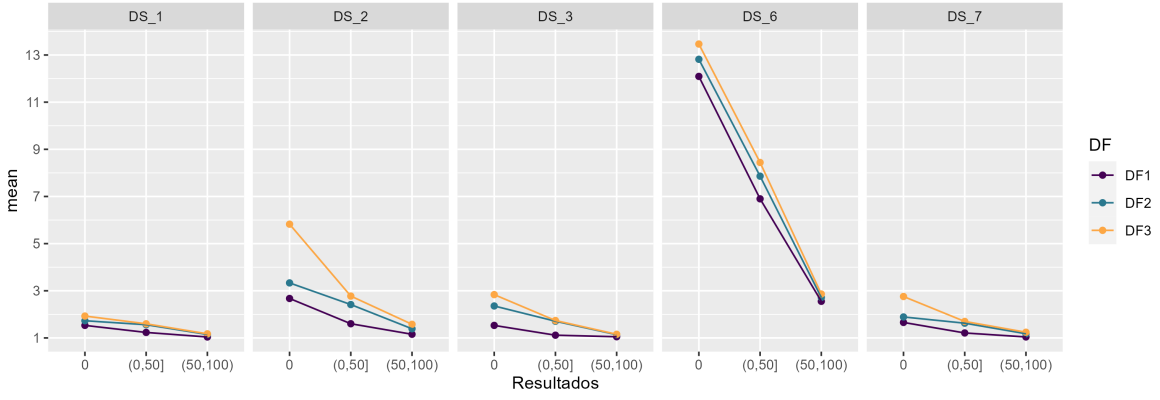

FIGURA 6.1: Media de alertas registradas por clasificación de DS_i, $i = 1, 2, 3, 6, 7$ para los diferentes *DF*.

TABLA 6.5: Resultados DS_i, $i = 1, 2, 3, 6, 7$ para los diferentes DF.

Tipo Alerta	DS_i	DF1		DF2		DF3	
		Viajes		Viajes		Viajes	
		(n)	$(\%)$	(n)	$(\%)$	(n)	$(\%)$
1	0	342	0.31	85	0.08	42	0.04
	(0,50]	377	0.34	148	0.14	110	0.1
	(50,100)	497	0.45	983	0.9	1064	0.97
	100	108079	98.9	108079	98.9	108079	98.9
2	0	16415	15.0	5385	4.93	986	0.9
	(0,50]	14957	13.7	8664	7.93	4399	4.02
	(50,100)	21053	19.3	38376	35.1	47040	43.0
	100	56870	52.0	56870	52.0	56870	52.0
3	0	3450	3.16	295	0.27	135	0.12
	(0,50]	1053	0.96	195	0.18	160	0.15
	(50,100)	485	0.44	8951	8.19	10004	9.15
	100	95841	87.7	95841	87.7	95841	87.7
6	0	6952	6.4	5790	5.3	4954	4.53
	(0,50]	2829	2.6	2806	2.6	27335	2.5
	(50,100)	25253	23.1	26438	24.2	27347	25
	100	74261	68.0	74261	68.0	74261	68.0
7	0	3070	2.81	725	0.66	136	0.12
	(0,50]	3298	3.02	1322	1.21	593	0.54
	(50,100)	3501	3.2	7822	7.16	9140	8.36
	100	99426	91.0	99426	91.0	99426	91.0

TABLA 6.6: Resultados DS_4.

Viajes	DS_4				
	0	5	20	35	100
n	10	26	99	1282	107878
$(\%)$	0.009	0.024	0.091	1.17	98.7

6.2.2. Severity Score

Las Tablas 6.7 y 6.8, presentan la cantidad y porcentaje de viajes con valores entre 0, (0,50], (50, 100), y 100 (cuando la alerta del tipo i no ocurre durante el viaje), para los *indicadores* SS_i, $i = 2, 6$ (Ecs. 6.3, 6.3) y $SS1_i$, $SS2_i$, con $i = 1, 7$ (Ecs. 6.5, 6.6). De donde se destaca:

- Los *indicadores* entregan valores en el intervalo $[0, 100]$.
- A medida que se varia (aumenta) el *Severity Factor*, el indicador da valores menos severos a los viajes, pues la cantidad de estos con $SS_i = 0$, $SS1_i = 0$, y $SS2_i = 0$ disminuye.

TABLA 6.7: Resultados SS_i, $i = 2, 6$ para SF1, SF2, y SF3

Tipo Alerta	SS	SF1		SF2		SF3	
		Viajes		Viajes		Viajes	
		(n)	$(\%)$	(n)	$(\%)$	(n)	$(\%)$
2	0	7341	6.72	6861	6.28	6748	6.17
	(0,50]	19184	17.6	18614	17	18440	16.9
	(50,100)	6780	6.2	7830	7.16	8117	7.43
	100	75990	69.5	75990	69.5	75990	69.5
6	0	13075	12	12967	11.9	12933	11.8
	(0,50]	10739	9.83	10243	9.37	10064	9.21
	(50,100)	11220	10.3	11824	10.8	12037	11.0
	100	74261	67.9	74261	67.9	74261	67.9

TABLA 6.8: Resultados de $SS1_i$ y $SS2_i$, $i = 1, 7$ para SF1, SF2, y SF3.

Severity Score		SF1		SF2		SF3	
		Viajes		Viajes		Viajes	
		(n)	$(\%)$	(n)	$(\%)$	(n)	$(\%)$
$SS1_1$	0	588	0.54	548	0.5	430	0.4
	(0,50]	333	0.31	350	0.32	402	0.4
	(50,100)	295	0.27	318	0.3	384	0.4
	100	108079	98.9	108079	98.9	108079	98.9
$SS2_1$	0	606	0.55	575	0.53	560	0.51
	(0,50]	334	0.31	348	0.32	357	0.33
	(50,100)	276	0.25	293	0.27	299	0.27
	100	108079	98.9	108079	98.9	108079	98.9
$SS1_7$	0	5470	5	5384	4.9	5213	4.8
	(0,50]	2276	2.1	2314	2.1	2391	2.2
	(50,100)	2122	1.9	2170	2	2264	2.1
	100	99427	91.0	99427	91.0	99427	91.0
$SS2_7$	0	4383	4.01	4349	3.98	4144	3.79
	(0,50]	3324	3.04	3345	3.06	3448	3.15
	(50,100)	2162	1.98	2175	1.99	2277	2.08
	100	99426	91.0	99426	91.0	99426	91.0

Los valores entregados por el indicador SS_i, $i = 2, 6$ para los viajes, parecen sensatos, ya que:

- Para SS_2 (Posible Colisión Frontal), los viajes con mayor velocidad de infracción tienen un $SS_2 = 0$ (comparando con los viajes con $SS_2 \neq 0$), los viajes con una velocidad promedio a 50 kms/hr obtienen un $SS_2 \in (0, 50]$, y aquellos con menor velocidad de infracción obtienen un *severity score* alto (Figura 6.2).
- De manera similar, para SS_6 (Exceso de Velocidad), los viajes con mayor velocidad de infracción tienen un *Severity score* más bajo que aquellos con velocidad más baja (Figura 6.2).

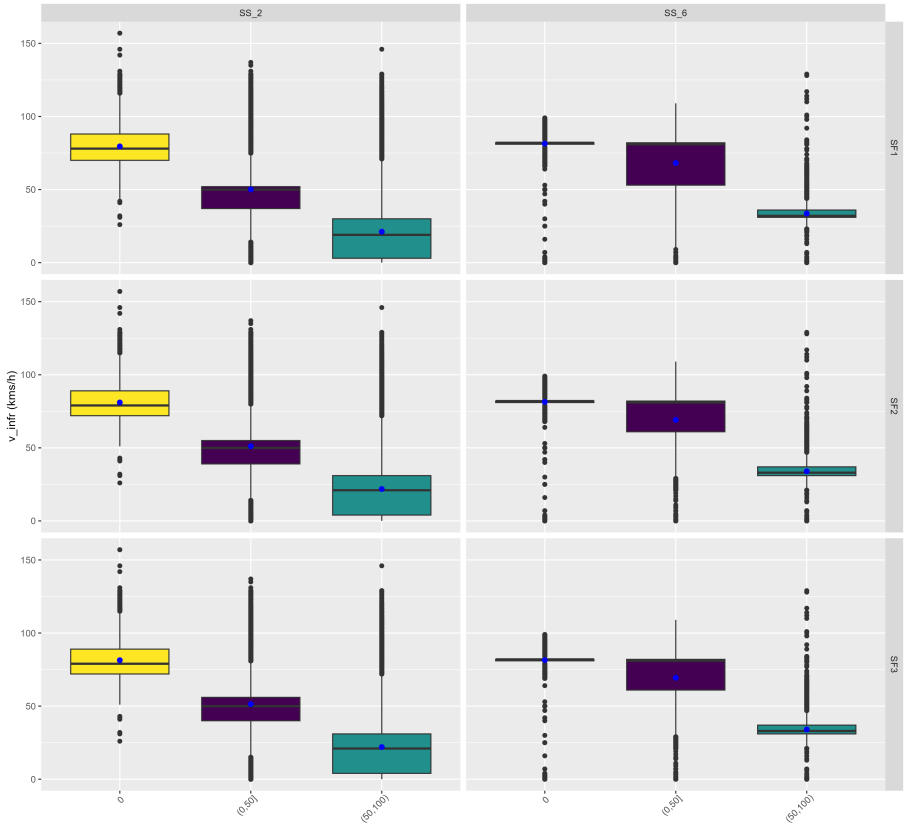

FIGURA 6.2: Clasificación de los viajes con SS_i, $i = 2, 6$.

Ahora, para las infracciones $i = 1, 7$, cuya severidad no se puede determinar directamente de la velocidad de infracción, se propusieron dos alternativas:

- Para $SS1$, se nota una clara distinción entre la agrupación de viajes según el valor de $SS1$, pues mientras más alta la aceleración/desaceleración más bajo su *Severity Score*.
- Para $SS2$, ocurre lo mismo; mientras mayor sea la diferencia entre la velocidad de infracción y la velocidad del viaje (en este caso), el valor de su *Severity Score* es menor.

Ambos entregan resultados coherentes, no obstante, hay que destacar, que $SS2$ es más simple.

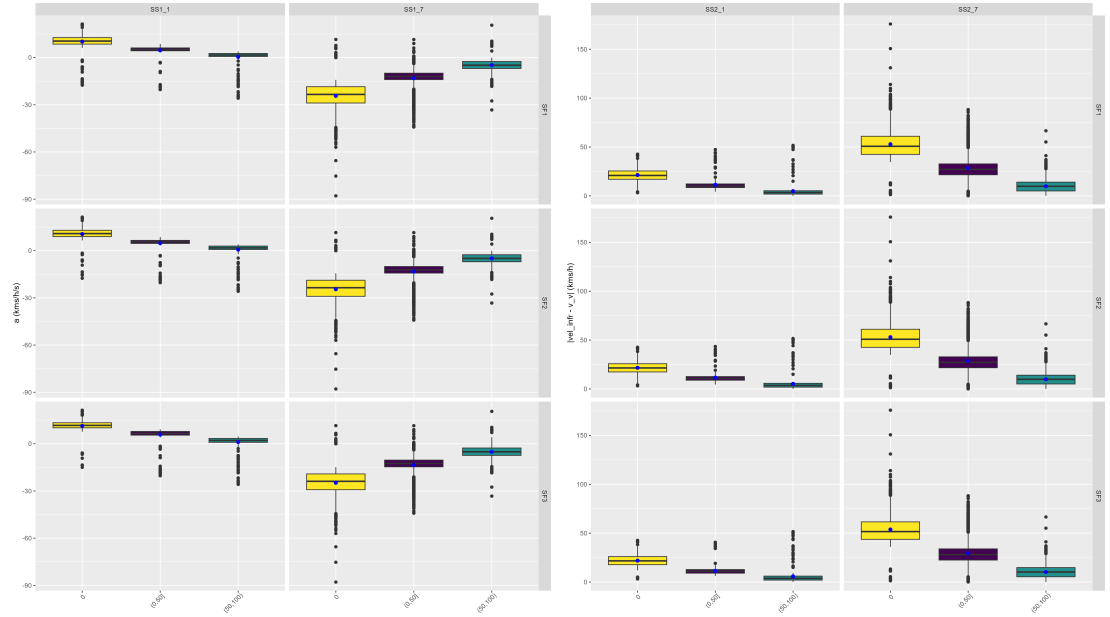

FIGURA 6.3: Clasificación de los viajes con $SS1_i$ y $SS2_i$, $i = 1, 7$.

TABLA 6.9: Velocidad media de viajes para $SS1_1$ y $SS2_1$ con los diferentes *Severity Factor*

Severity	$SS1_1$			$SS2_1$			$SS1_7$			$SS2_7$		
Factor	0	(0,50]	(50,100)	0	(0,50]	(50,100)	0	(0,50]	(50,100)	0	(0,50]	(50,100)
SF1	28.39	39.43	46.49	28.58	39.89	46.71	55.9	41.3	35.9	59	42	36.1
SF2	27.87	38.77	46.20	28.19	39.37	46.52	56.2	41.4	36	59.1	42	36.1
SF3	26.32	36.76	45.43	27.98	39.16	46.45	56.6	41.6	36	59.7	42.6	36.1

Por otra parte, de la Tabla 6.9 se puede observar:

1. Que los viajes con menor valor en su *Severity Score* para la infracción del tipo $i = 1$, tienen menor velocidad media. Es decir, estos viajes con velocidades más bajas realizan mayor aceleración.
2. Que los viajes con menor valor en su *Severity Score* para la infracción del tipo $i = 7$, tienen mayor velocidad media. Es decir, estos viajes con velocidades más altas realizan mayores desaceleraciones.

Se podría concluir que los viajes con menor *Severity Score* tienen una velocidad más inconsistente durante la duración del viaje. Es más, si estos viajes con velocidad inconsistente, ocurren durante un mayor flujo vehicular, se pueden ocasionar maniobras riesgosas por parte de los otros conductores para esquivar a este tipo de conductor inconsistente.

Tras todas estas observaciones/conclusiones sobre los testeos para *Duration Factor* y *Severity Factor* en *Duration Score* ($DS_i, i = 1, 2, 3, 6, 7$) y *Severity Score* (SS_i con $i = 2, 6$, y $SS1_i$, $SS2_i$ con $i = 1, 7$), respectivamente, dado que para todos los teteos solo disminuye la "severidad" para la evaluación de los viajes, la elección para continuar el proceso de construcción del *Índice de Conducción Riesgosa (ICR)* es $SF3$ y $SF3$; es decir:

1. $DF3_i$ tiene los valores de 0.4 %, 0.6 %, 0.5 %, 11 %, 0.58 % para $i = 1, 2, 3, 6, 7$ respectivamente.
2. $SF3_i$ tiene los valores de 112 % y 81 % para $i = 2, 6$ respectivamente en SS_i, y 1550 % y 3600 % para $i = 1, 7$ respectivamente en $SS2_i$.

Nótese que para lo que resta del documento, $SS2_i$, $i = 1, 7$ será referido simplemente como SS_i, $i = 1, 7$, pues esta fórmula con distancia será la utilizada en adelante.

6.3. Normalización

En esta etapa de construcción se transforman los datos con la finalidad de homogeneizar la métrica, debido a que generalmente los diferentes indicadores vienen en diferentes unidades. En este caso, todos los indicadores están en porcentajes (0-100 %) y además la mayoría de las observaciones están mas proximas a ser valores altos (cercanos a 100 %),

se sigue la transformación descrita en la Sec. 4.1.2, la cual corresponde a la recomendada por Casadio en [23] para los indicadores en porcentajes y con sesgo.

Dado que cada *indicador*, al tener mayor valor contribuye a que el *índice* también tenga un score mayor, se considera la *correlación ex-ante* (*polaridad*) como positiva, y en consecuencia no se le invierte el signo a ningún *indicador*. Finalmente, los indicadores quedan en un rango de [0,1].

Usar estos valores brindan la ventaja de mitigar el efecto de los valores atípicos en el resescalado [23].

6.4. Ponderación

Para obtener las ponderaciones para los diferentes *indicadores*, se realizan 3 esquemas: *PCA*, *FA*, y *DEA (BoD)* (Sec. 4.1.3.1). Los métodos *PCA* y *FA* se efectuaron sobre los datos transformados, mientras que para *DEA* se realiza sobre los valores brutos de los indicadores.

Se tiene un número de 109295 viajes; cantidad que resulta demasiada intensa para el algoritmo de *DEA*, pues cada viaje (función objetivo) está sujeto a un número de $N + 3l$ restricciones, con N el número de viajes del conjunto y l el número de indicadores. Por ende, se extrae una muestra aleatoria estratificada del 50 % del conjunto de datos, con el objetivo de que la muestra represente lo mejor posible todo el conjunto de datos, asegurándose que la variable *Duración* (en este caso) mantenga su distribución (Figura 6.4). En la Tabla 6.10 se puede corroborar que las cantidades de los diferentes Tipos de Alertas, en la muestra extraída, mantienen su porcentaje de viajes. Nótese, además, que los porcentajes no suman 100 %, pues hay viajes que presentan más de un Tipo de Alerta. Explícitamente, el conjunto de datos presenta 7.71 % de viajes con más de un Tipo de Alerta; y en la muestra, corresponde a 7.86 %. Con esta muestra estratificada se ejecutan los tres *Métodos de Ponderación*.

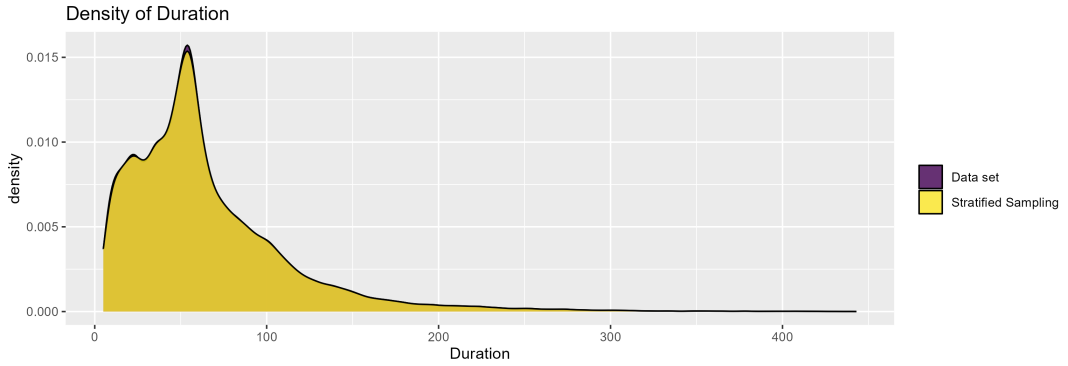

FIGURA 6.4: Función de Densidad estimada de la variable *Duración*

TABLA 6.10: Tipos de Alertas presentes en el conjunto de datos y muestra estratificada.

Tipo	Conjunto de Datos		Muestra estratificada	
Alerta	n	(%)	n	(%)
1	1216	1.11	636	1.16
2	52425	48.0	26319	48.16
3	13454	12.3	6799	12.44
4	1417	1.3	711	1.30
5	4922	4.5	2442	4.47
6	35034	32.0	17391	31.82
7	9869	9.03	4946	9.05

6.4.1. PCA

Se ejecutan dos *PCA*, uno sobre el conjunto tal como está (muestra estratificada); es decir, donde los tipos de alertas $i = 2, 6$ predominan por sobre el conjunto de viajes, y otro sobre el conjunto balanceado. Balancear las clases, fue un cuestionamiento, porque se puede alterar la naturaleza de los datos, pero dado que la mayoria de los viajes tienen registros de solo un tipo de infracción y la mayoría de estos registros son de dos tipos, los otros indicadores tienen poca variabilidad y los resultados pueden verse afectados y no reflejar adecuadamente el riesgo de todos los tipos de infracciones.

6.4.1.1. PCA 1

Se efectua *PCA* sobre los indicadores VS_i, $i = 1, \ldots, 7$ (Ec. 3.7) ya calculados a partir de los DS_i y SS_i, con $i = 1, \ldots, 7$, usando la primera *CP* (descrito en la Sec. 4.1.3.1.1),

que capta una variabilidad de 47.8 % para obtener las ponderaciones. Se denomina *PCA 1*.

Las ponderaciones obtenidas, escaladas para que la suma al cuadrado sea 1 [17], se ilustran en la Tabla 6.12. Sin embargo, se debe mencionar que, a partir de Tabla 6.11, que muestra los *eigenvalue*, *varianza*, y *varianza acumulada* para cada *CP*, debe notarse que los *eigenvalues* son menor a 1, lo que indica que la varianza explicada por cada *CP* es menos de lo que se esperaría al azar; efectivamente, para explicar al menos el 90 % de la varianza en el conjunto de datos de todos los indicadores, se necesitan 4 *CP*. Este aspecto indica, que el fenómeno descrito por el conjunto de los 7 indicadores es bastante multidimensional. Es más, una mayor correlación entre los indicadores habría dado lugar a menos componentes necesarias [15].

Tabla 6.11: Summary del PCA efectuado sobre los indicadores.

CP	PCA 1		
	Eigenvalue	Varianza (%)	V. Acumulada (%)
1	0.0704	47.7625	47.7625
2	0.0320	21.6919	69.4545
3	0.0195	13.2348	82.6893
4	0.0126	8.5451	91.2344
5	0.0069	4.7078	95.9422
6	0.0035	2.3609	98.3031
7	0.0025	1.6969	100

La Figura 6.5 corresponde a la correlación de los *indicadores* normalizados con la transformación de 3 pasos. Se aprecia que, los indicadores que tienen mayor correlación negativa entre ellos son VS_2 con VS_j, $j = 3, 6$, y VS_3 con VS_6. Pero nótese, que esto puede ser reflejo de que la mayoría de los viajes registran alertas de un solo tipo, y sólo un 7.86 % de la muestra tiene infracciones de Tipo diferente (con máximo dos tipos), siendo además estas asociaciones lineales coherentes con la lógica, ya que mientras más infracciones de posible colisión frontal, también hay más infracciones de exceso de velocidad registradas en un viaje, o mientras mayor registro de infracciones de distracción, mayor registro de posibles colisiones frontales, etc. Pero debido a que las ponderaciones se basan en correlaciones en lugar de vínculos reales entre los indicadores evaluados, se pueden producir resultados impredecibles [12].

Dado que la correlación entre los indicadores es bastante pequeña, y además la variabilidad de la primera *CP* en el *PCA* efectuado es menor a 50 %, las ponderaciones obtenidas por estos métodos parecen poco factibles. Pues, tanto el *PCA* como el *FA* pueden asignar ponderaciones más bajas a una dimensión crucial simplemente porque está débilmente correlacionada con otras dimensiones. Además, las ponderaciones no se corresponden con la importancia relativa de las dimensiones en el mundo real, asi que las ponderaciones derivadas mediante ACP/FA pueden no ser válidas [12].

Nótese además que las ponderaciones obtenidas con mayor porcentaje, son para aquellos tipos de infracciones con más registros en el conjunto de datos usado (la muestra estratificada), es decir, para los tipos $i = 2, 6$.

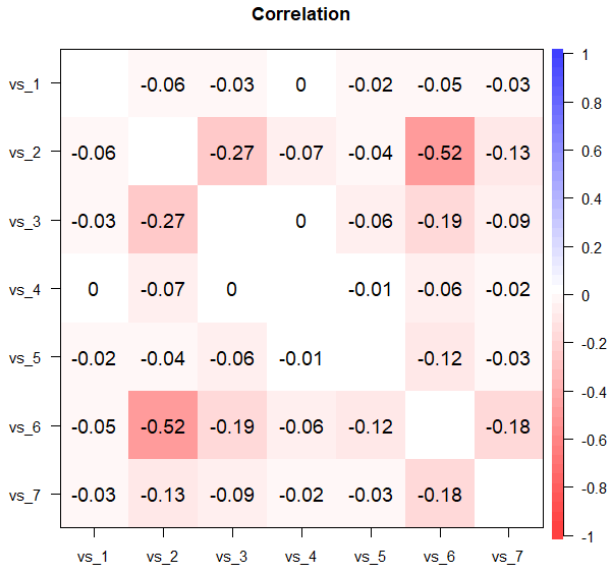

FIGURA 6.5: Correlación de los indicadores

6.4.1.2. PCA 2

El balance del conjunto de datos se realiza con el *Algoritmo Smote*[1] que consiste en la combinación de *Submuestreo y Sobremuestreo Aleatorio Simple*. La elección de una combinación de submuestreo y sobremuestreo fue debido a que las clases minoritarias y mayoritarias son demasiado dispar en tamaño.

Para evaluar la sensibilidad del conjunto balanceado, y ver como afecta a las ponderaciones obtenidas con *PCA*, se ejecuta 100 veces el *Algortimo Smote*, y para cada uno

[1]Disponible en el package *UBL* de *R*, con el comando *SmoteClassif*, y la opción *balance*.

de estos conjuntos balanceados de 54645 observaciones, se obtienen las poderaciones mediante *PCA*, como describe la Sec. 4.1.3.1.1.

La Figura 6.6 representa las ponderaciones obtenidas en cada conjunto balanceado, de donde se aprecia que a pesar de que hay variaciones en algunas ponderaciones ($i = 2, 4, 5$), en general oscilan entre el mismo rango de valores. Además:

1. La *desviación estándar* de las ponderaciones de cada indicador VS_i, $i = 1, \ldots, 7$, en los diferentes muestreos para balancear los datos, es de 0.4 %, 7.3 %, 1.0 %, 5.8 %, 3.1 %, 0.2 %, y 0.2 %, respectivamente; y por tanto, las ponderaciones no se alejan en gran medida de la media.
2. El *Rango Intercuartil (IQR)* de las ponderaciones de cada indicador VS_i, $i = 1 \ldots 7$ es de 0.2 %, 1.3 %, 0.5 %, 2.1 %, 1.3 %, 0.1 %, y 0.1 %, respectivamente; es decir, que se reafirma el hecho de que las ponderaciones no tienen gran variabilidad.
3. La variabilidad captada por la primera *CP* para las 100 réplicas corresponde a 22 %, y el *eigenvalue* es menor a 1 (aproximandamente a 0.06 en todas las iteraciones).

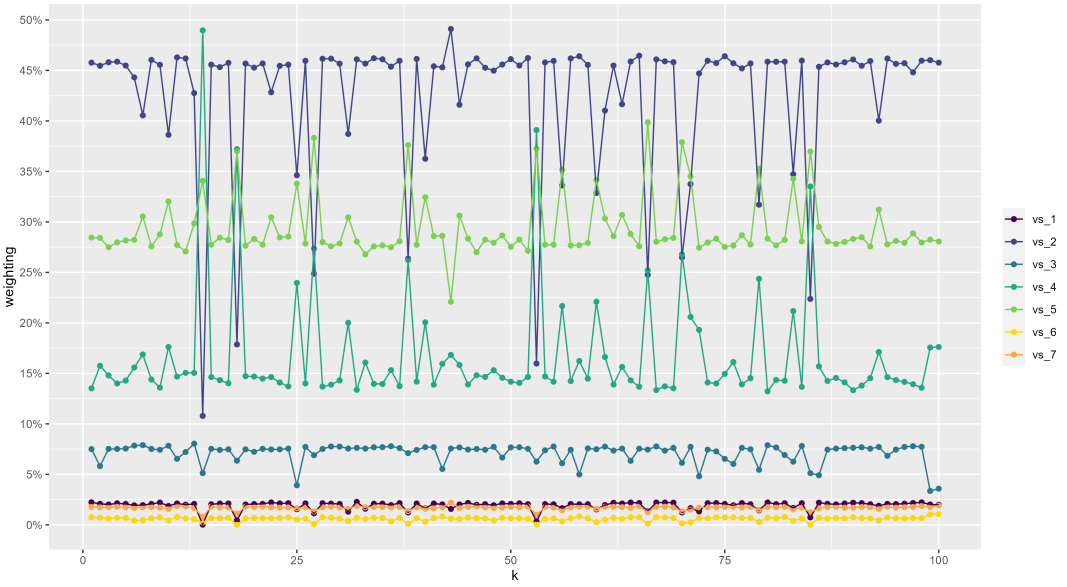

FIGURA 6.6: Ponderaciones obtenidas en los 100 conjuntos balanceados aleatoriamente

Finalmente, las ponderaciones para cada indicador corresponden a la media de las ponderaciones en las 100 réplicas, y se ilustran en la Tabla 6.12.

6.4.2. FA

Para determinar el número de factores a extraer, se usó el análisis del *Scree Plot*, que se basa en la magnitud de los *eigenvalues* pero a partir de la tendencia que se observa en el *Scree Plot*. Para este caso, se determinó que la extracción seria de 3 factores, pues son los que tienen un valor mayor en consideración a los demás siguientes, es decir el punto de inflexión en la curva del *scree plot* (Figura 6.7).

Rotando la matriz factorial con *varimax*, cuyo método de rotación ortogonal minimiza el número de variables que tienen saturaciones altas en cada factor y simplifica la interpretación de los factores; se extraen los factores, y se siguen los pasos ilustrados en la Sec. 4.1.3.1.2, obteniéndose las ponderaciones dadas en la Tabla 6.12 junto con observaciones. De donde, se destaca que los indicadores con mayor ponderación, son aquellos que tienen mayor correlación con algún indicador; por ejemplo VS_2 con VS_6 y VS_2 con VS_3 (Tabla 6.5).

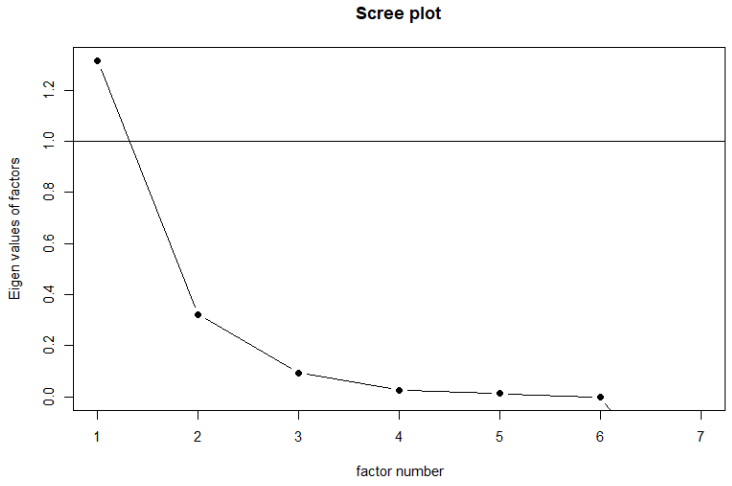

FIGURA 6.7: Número de factores a extraer.

6.4.3. DEA

En una primera instancia se ejecutó el algoritmo de *DEA* descrito en la Ec. 4.4, el cual solo impone la no negatividad de las ponderaciones (y que la puntuación final no supere el 100) y permite estimar libremente las ponderaciones para maximizar la puntuación final, pero para algunos indicadores se obtuvieron ponderaciones de valor cero y el índice finalmente entregaba para el viaje correspondiente un resultado de 100 % a

pesar de tener infracciones. Para evitar estos resultados poco razonables, puntualmente las ponderaciones de valor cero, se propone una cuota mínima proporcional para los indicadores; y una cuota máxima proporcional, para evitar que un indicador domine en exceso al índice. Así, el modelo propuesto y usado corresponde a la Ec. 4.5. Para este caso, se establece $L_j = 0.05$, $\forall j = 1, \ldots, 7$, y $U_j = 0.3$, $\forall j = 1, \ldots, 7$ (Es decir, 5 % y 30 % respectivamente).

Como las ponderaciones se derivan de los datos, sólo se ven influidas por los valores de los indicadores y las restricciones impuestas [12]. Por lo tanto, merece la pena restringir la flexibilidad de las ponderaciones.

Dado que *DEA* construye un *índice* para cada viaje, al obtener las ponderaciones de todos los viajes (7 ponderaciones por viaje), las ponderación preliminar del indicador j corresponde a $u_j = \frac{\sum_{i=1}^{n} w_{ij}}{n}$, con n el número de viajes y w_{ij} la ponderación dada por *DEA* para el viaje i en el indicador j. Las ponderaciones definitivas son:

$$\hat{u}_j = \frac{u_j}{\sum_{j=1}^{l} u_j} \tag{6.7}$$

donde $l = 7$ es el número de indicadores.

Así, las ponderaciones obtenidas por este método se ilustran en la Tabla 6.12.

De los resultados obtenidos para las ponderaciones se destaca:

1. La *desviación estándar* de las ponderaciones de cada indicador VS_i, $i = 1, \ldots, 7$ para el conjunto de datos usado es 5 %, 8 %, 8 %, 3 %, 1 %, 4 % y 7 % respectivamente. Es decir, que las ponderaciones en los diferentes n viajes no se alejan en gran medida de la media (ponderaciones usadas).
2. El *Rango Intercuartil (IQR)* de las ponderaciones de cada indicador VS_i, $i = 1, \ldots, 7$ para el conjunto de datos es 1 %, 13.5 %, 14.7 %, 0.1 %, 0.1 %, 0.2 %, 1.2 %, respectivamente; es decir, que las ponderaciones en los diferentes n viajes no tienen gran variabilidad, ya que el 50 % de las ponderaciones se encuentra dentro del *IQR*. Por ejemplo, para el indicador VS_1, el 50 % de las ponderaciones obtenidas se encuentra dentro de un rango de $\pm$ 0.5 alrededor de la mediana (Q_2).
3. Las ponderaciones dadas para VS_j, $j = 1, 7$ son altas con respecto a las demás. Esto se podría explicar con el hecho de que el objetivo del *DEA (BoD)* es maximizar la

suma ponderada de cada viaje, y dado que la muestra cuenta con solo un 1.16 % y 9.05 % de alertas del tipo $i = 1, 7$, respectivamente (Tabla 6.10); es decir que la mayoría de los viajes para estos indicadores corresponden al máximo posible (100), se podría presuponer que las ponderaciones de ambos indicadores sean altas. No obstante, VS_j, $j = 4, 5$ tienen un porcentaje de registros menor a 9.05 % y sin embargo, estas ponderaciones son pequeñas (6 % y 5 %). Estos resultados pueden deberse a:

a) Interacciones complejas entre variables; es decir, que en lugar de que una variable actúe de manera independiente, su influencia puede estar relacionada con las combinaciones de valores de otras variables; y así, que la importancia relativa de las variables no se refleje.

b) Que la restricción de cuota mínima y máxima proporcional a pesar de ser bastante amplia puede haber tenido un efecto significativo en las ponderaciones resultantes, en conjunto con el orden inicial de los indicadores.

c) Tamaño del conjunto de datos, ya que a pesar de haber consistido del 50 % del conjunto original, aún sigue siendo bastante grande, y las interacciones y patrones se vuelven más complejas.

6.4.4. Resultados

La Tabla 6.12 tiene las ponderaciones obtenidas para el indicador *Violation Score* por los Métodos de ponderación: *PCA 1*, *PCA 2*, *FA* y *DEA*.

Tabla 6.12: Ponderaciones obtenidas por los diferentes métodos ejecutados.

Indicador	PCA 1 (%)	PCA 2 (%)	FA (%)	DEA (%)
VS_1	0.001	2	1	29
VS_2	59.875	42	28	14
VS_3	0.420	7	35	13
VS_4	0.002	17	1	6
VS_5	0.023	29	2	5
VS_6	39.677	1	27	6
VS_7	0.002	2	6	27

Tal como se destacó previamente, las ponderaciones obtenidas mediante *PCA 1* y *FA*, parecen poco factibles, sobre todo por la predominancia de los tipos de alerta $i = 2, 6$ (poca variabilidad para los indicadores restantes) y por la poca correlación existente entre los indicadores. *PCA 2* tiene ponderaciones aparentemente más congruentes con el riesgo, pero la poca variabilidad captada por la primera *CP*, aún hace dudosa la factibilidad de estas. Por otra parte, *DEA*, parece aportar ponderaciones aceptables y hasta cierto punto realistas; sin embargo, aún falta chequear la fiabilidad del *índice* usando estas ponderaciones.

Entre las ponderaciones de *PCA 1* y *PCA 2*, las primeras son descartadas, por el exceso de ponderación para los indicadores VS_i, $i = 2, 6$.

6.5. Agregación

En esta sección se realizan las agregaciones correspondientes según las metodologías establecidas previamente. Específicamente, se establece la agregación *Media Cóncava* y *lineal* descritas con mayor precisión a continuación.

6.5.1. Media Cóncava

Los *indicadores* transformados a una vecindad del intervalo real [0, 1], mediante los 3 pasos indicados en la Sec. 4.1.2, se transforman con una *función cóncava* adecuada cuyos parámetros a y b pueden calibrarse. A mayor b, mayor *penalización*, mientras que a mayor a aumenta la diferencia entre la penalización de las unidades [23], en este caso de los viajes más riesgos y menos riesgos.

Usando los *indicadores transformados* y *penalizados* se realiza la agregación tal como indica la Ec. 4.8, con las ponderaciones *PCA 2*, y *FA* sobre los *indicadores* VS_i, $i = 1, \ldots, 7$.

Las Figuras 6.8 y 6.9, muestran la varianza y media del conjunto de viajes puntuados según las diferentes ponderaciones y agregación *media cóncava* para diferentes valores de a y b. Efectivamente, se aprecia que a mayor b la varianza aumenta (media disminuye); mientras, que para mayor a la varianza (media) de los viajes con [1, 10) alertas registradas aumenta (disminuye), pero en menor medida que para los viajes con 10 y

más alertas registradas. Se establece como parámetros de la *función cóncava*, $a = 1$ y $b = 1$ para todos los *indicadores*, con el objetivo de que la "penalización" no sea tan severa, y que los scores obtenidos de los *índices sintéticos* no decrezcan tanto.

Dado que el *índice* obtenido con *media cóncava* tiene dominio sobre todo $\mathbb{R}$, es necesario reescalar a un intervalo de [0, 100], para facilitar la comparación entre los diferentes viajes y entre los diferentes *índices sintéticos*. Es importante que el orden relativo de los valores originales se mantenga al realizar la transformación; en otras palabras, que, si un viaje tiene mayor score que otro, después de la transformación la relación debe mantenerse. Una función que mapea los valores originales a un intervalo de [0, 1] y que cumple con los objetivos mencionados es la *función sigmoide suavizada*:

$$f(x) = \left(\frac{1}{1 + e^{-cx}} \right) \tag{6.8}$$

donde el parámetro c se puede calibrar para ajustar la pendiente de la curva sigmoide y la suavidad de la transición. A mayor c, mayor pendiente (Figura 6.10). Además, como se quiere que el intervalo sea de [0, 100], los valores transformados por la función 6.8 se multiplican por 100.

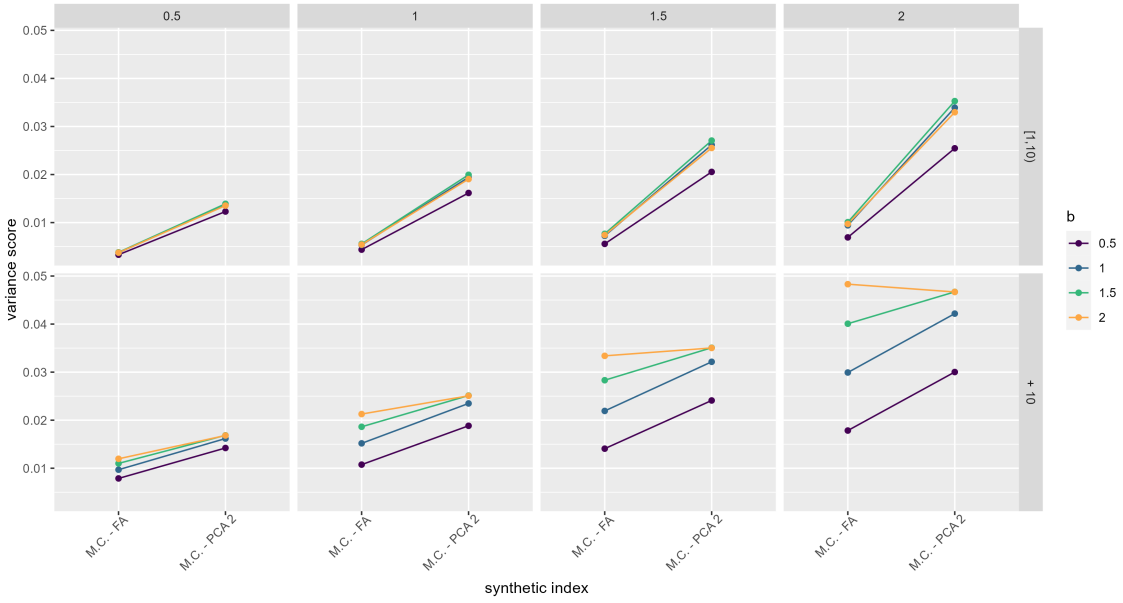

FIGURA 6.8: *Varianza* de las puntuaciones obtenidas para los viajes con alertas de [1, 10) y viajes con 10 y más alertas registradas, usando los diferentes esquemas de ponderación y agregación *media cóncava*, variando los valores de a (columna) y b.

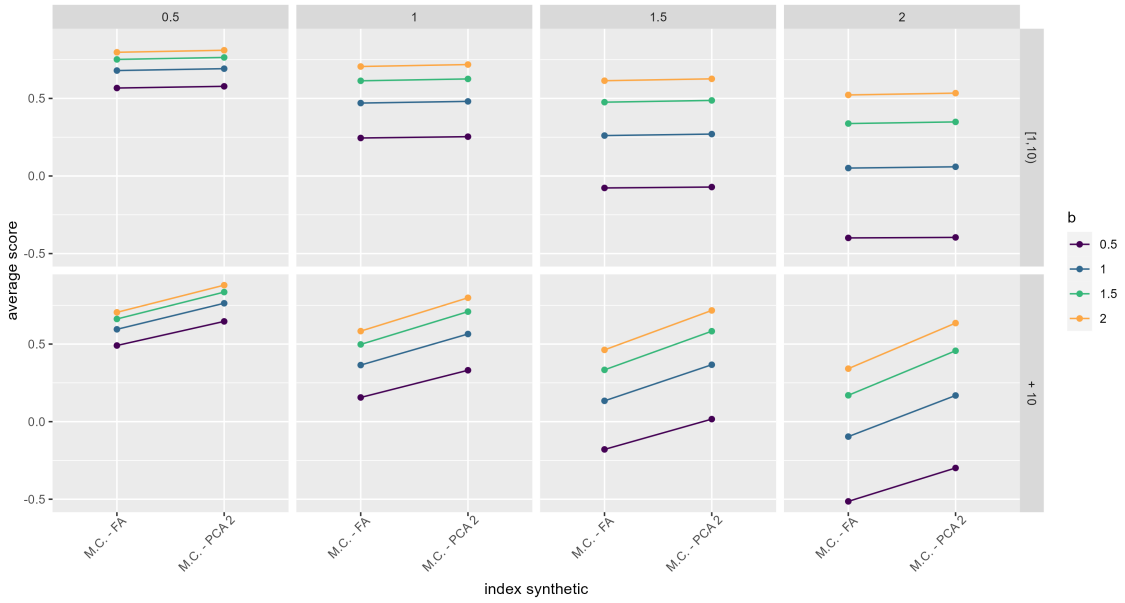

FIGURA 6.9: *Media* de los puntuaciones obtenidas para los viajes con alertas de [1, 10) y viajes con 10 y más alertas registradas, usando los diferentes esquemas de ponderación y agregación *media cóncava*, variando los valores de a (columna) y b.

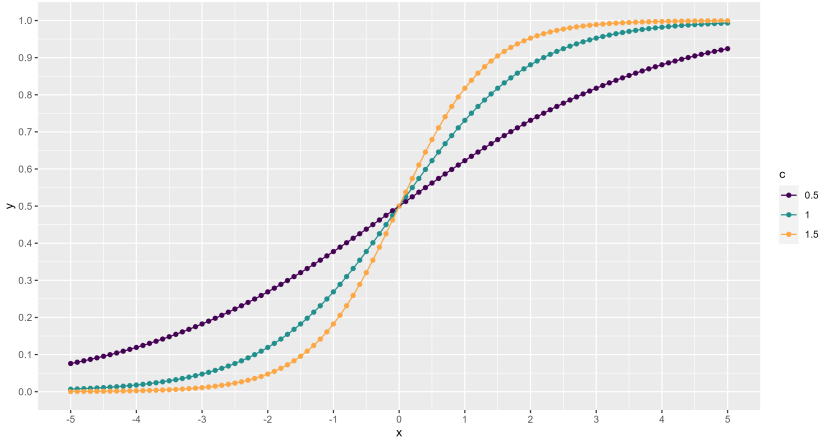

FIGURA 6.10: Función Sigmoide Suavizada.

TABLA 6.13: Estadísticos descriptivos de los *índices sintéticos* obtenidos con agregación media cóncava y parámetros de penalización $a = 1$ y $b = 1$

Estadísticos	M.C. - PCA 2	M.C. - FA
Mínimo	-0.2498	-0.2262
Q_1	0.3519	0.4290
Mediana	0.4520	0.4592
Media	0.4854	0.4645
Q_3	0.6314	0.4996
Máximo	0.6321	0.6321

De la Figura 6.10 y de los estadísticos descriptivos de los *índices sintéticos* obtenidos con *media cóncava*, se establece $c = 1$.

6.5.2. Lineal

Se realiza esta agregación con las ponderaciones obtenidas con *DEA* sobre los valores brutos de los indicadores correspondientes, y también con las ponderaciones *PCA 2* sobre los valores de los indicadores transformados mediante los 3 pasos indicados. La agregación *lineal* con *PCA 2* se realiza, con el fin de comparar la *media cóncava* con *PCA 2*.

6.5.3. Resultados

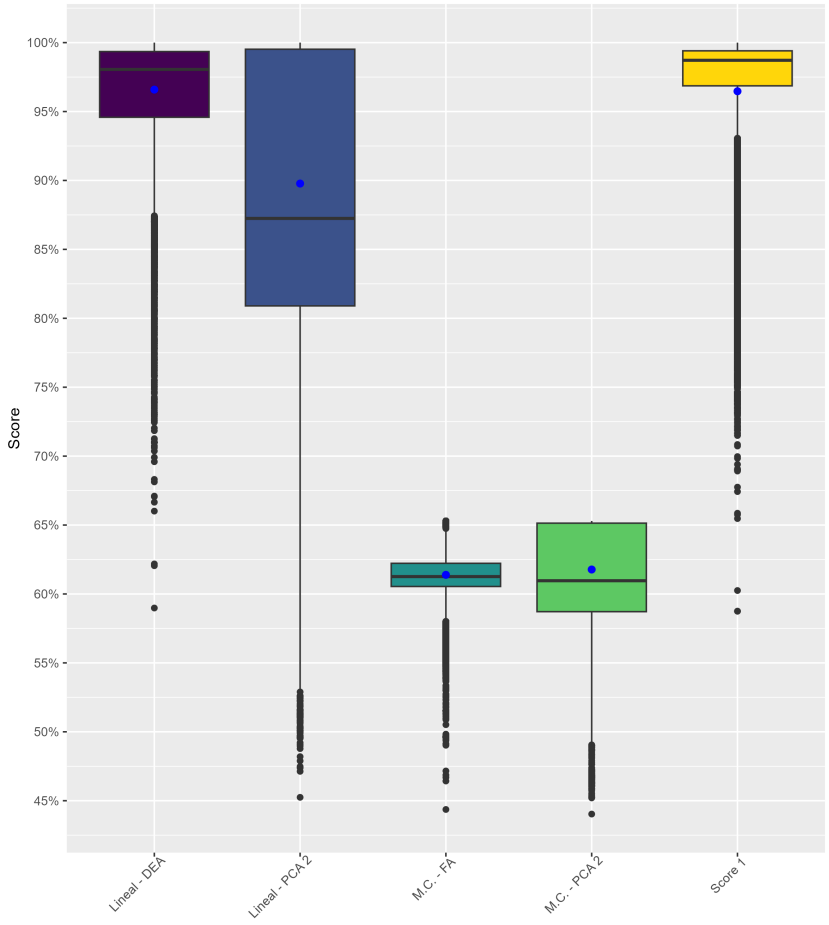

FIGURA 6.11: *Boxplot* de los scores obtenidos usando los diferentes esquemas.

En la Figura 6.11 se ilustran los *boxplots* de las puntuaciones obtenidas para los diferentes viajes, usando las diferentes metodologías de *ponderación* y *agregación*. Aquí, *Score 1*, simula al *score* usado en la práctica (solo con DS_i, $i = 1, \ldots, 7$) con las ponderaciones dadas originalmente en una agregación *lineal* sobre los valores brutos de los indicadores, pero con sus fórmulas modificadas, ya que los DS_i originales entregan valores solo entre 99-100 %. Es esta Figura, se aprecia que los viajes obtienen menor puntuación y menor dispersión cuando se usa la *media cóncava* como agregación; que *DEA* con agregación *lineal* puntúa con valores más altos a los viajes en relación a los otros métodos (lo cual tiene sentido, porque *DEA (BoD)* tiene como objetivo maximizar la suma ponderada lineal); *PCA 2* con agregración *lineal* difiere en gran medida visualmente con *PCA 2* con agregación *media cóncava*; y finalmente, *Score 1* presenta valores similares con *lineal-DEA*, pero con un Rango Intercuartílico menor.

6.6. Análisis de Sensibilidad y Robustez

En esta sección se realiza un *análisis de sensibilidad* y *robustez*, con la finalidad de minimizar los riesgos de producir índices sintéticos sin sentido. Este tipo de análisis puede mejorar la precisión, credibilidad e interpretabilidad de los resultados finales [24].

Se realiza sobre el otro 50 % del conjunto de datos[2].

6.6.1. Análisis de Sensibilidad

El objetivo es estudiar si las clasificaciones de los viajes cambian, y con qué intensidad, tras la eliminación de uno de los siete indicadores finales usados (VS_i, $i = 1, \ldots, 7$).

La Tabla 6.14 muestra el examen de los resultados, donde claramente el *índice sintetico* con *DEA (BoD)* tiene mayor sensibilidad a los indicadores, pues a pesar de que la media es de 1.11, se muentra que la extracción de un indicador afecta a las posiciones de los viajes, ya sea hasta 2.4 posiciones o 0.22 posiciones. Las ponderaciones de *FA* con agregación *media cóncava* y *Score 1*, también presentan sensibilidad con la extracción de cualquier indicador. En cambio, *Media Cóncava - PCA 2*, presenta rangos de variación

[2]Recuérdese que para la ponderación y agregación se usó una muestra aleatoria estratificada del 50 % del tamaño original.

TABLA 6.14: Media, Varianza y rango de variación de las diferencias, tras la eliminación de un indicador, para los métodos de síntesis.

Indicador	Media Concava		Lineal		
Eliminado	PCA 2	FA	PCA 2	DEA	Score 1
Mean					
VS_1	0.08	0.15	0.22	2.42	0.85
VS_2	3.24	2.69	9.39	2.00	1.52
VS_3	0.37	2.26	1.07	0.73	0.94
VS_4	0.77	0.15	2.17	0.26	2.25
VS_5	1.54	0.06	4.31	0.22	0.99
VS_6	0.08	2.17	0.22	0.55	3.04
VS_7	0.09	0.34	0.27	1.59	1.47
mean	**0.88**	**1.12**	**2.52**	**1.12**	**1.58**
varianza	**1.36**	**1.41**	**11.38**	**0.77**	**0.65**
Varianza					
VS_1	0.01	0.00	0.04	3.10	0.29
VS_2	9.27	2.52	77.21	6.65	1.35
VS_3	0.11	2.10	0.95	1.26	2.77
VS_4	0.54	0.00	4.20	0.26	5.61
VS_5	2.20	0.01	16.71	0.20	0.33
VS_6	0.00	2.03	0.031	1.07	26.96
VS_7	0.01	0.04	0.06	6.76	0.89
mean	**1.73**	**0.96**	**14.17**	**2.76**	**5.45**
varianza	**11.67**	**1.41**	**808.89**	**7.987**	**95.212**
Rango					
VS_1	0.74	0.24	2.00	27.99	11.00
VS_2	16.63	10.93	42.00	14.00	7.00
VS_3	2.63	13.92	7.00	13.00	16.00
VS_4	6.52	0.21	17.00	6.00	23.00
VS_5	11.34	0.90	29.00	5.00	7.00
VS_6	0.37	10.39	1.001	6.00	24.00
VS_7	0.74	2.25	2.090	27.00	11.00
mean	**5.57**	**5.55**	**14.29**	**14.14**	**14.14**
varianza	**39.79**	**35.29**	**253.89**	**95.76**	**50.14**

mucho más extremos, pero casi nula variación al eliminar el indicador VS_i, $i = 1, 6, 7$, y varianzas mayores al *índice* obtenido con *Media Cóncava - FA*.

El análisis de la varianza y de rango de las diferencias absolutas confirman la mayor variabilidad de los desplazamientos de rango que se producen con *Media Cóncava - PCA 2*, *lineal - PCA 2*, y *Score 1*.

En general, el uso de la "penalización" en *media cóncava - PCA 2*, parece reducir la sensibilidad en términos de variabilidad de las distribuciones, pues *PCA 2* con agregación *lineal* sobre los datos transformados, pero no penalizados (se usa en media cóncava),

muestra mayor varianza y rango de diferencias absolutas, como también mayor sensibilidad a la extracción de los indicadores VS_i, $i = 1, 6, 7$.

6.6.2. Análisis de Robustez

En este apartado se realiza un *Análisis de Robustez* de las estimaciones del *índice sintético*, a partir de los siete indicadores VS_i, $i = 1, \ldots, 7$ obtenidos de los indicadores DS_i y SS_i, $i = 1, \ldots, 7$ perturbados, tal como se indica en la Sec. 4.1.4.2, y aplicando los esquemas de ponderación y agregación respectivos.

El objetivo, es describir en qué medida, la variabilidad aumenta al variar la perturbación añadida.

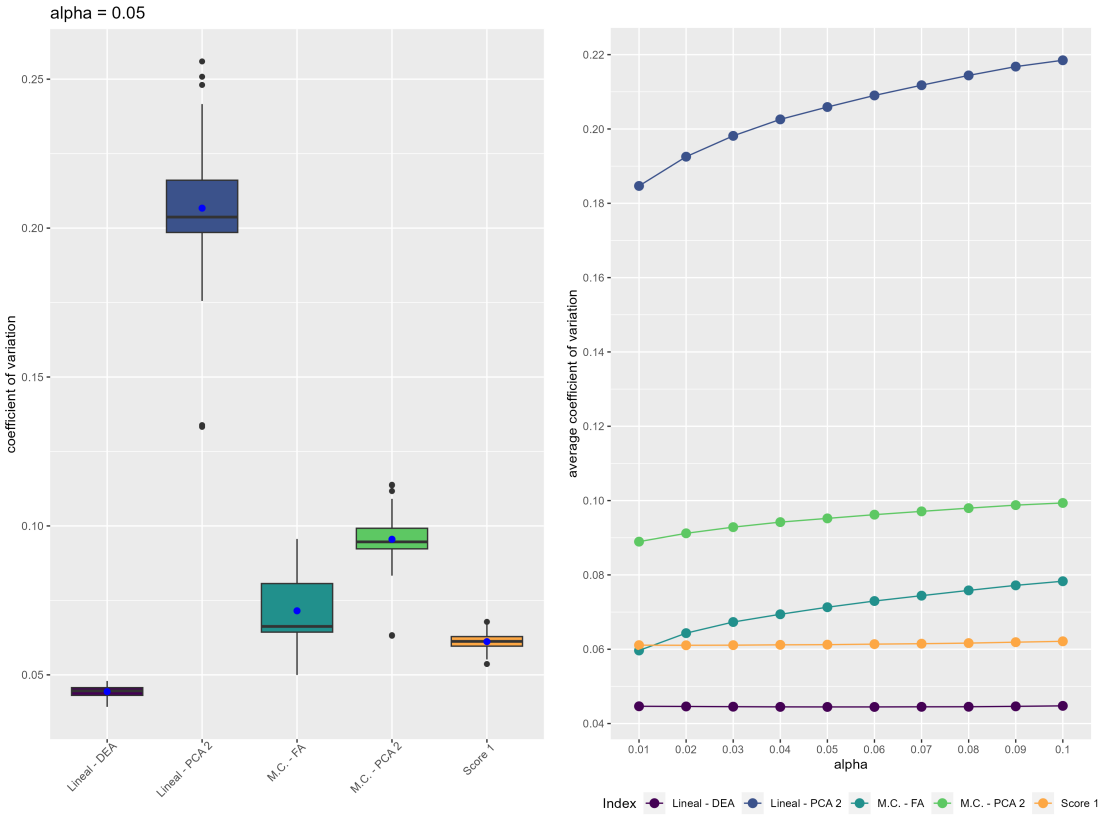

FIGURA 6.12: Izquierda: *Boxplot* de los coeficientes de variación de los índices sintéticos con las metodologías utilizadas. Derecha: Media del Coeficiente de variación de los *índice sintético* al variar la perturbación aleatoria usando 100 muestras aleatorias para cada α.

Usando $\alpha = 0.05$ y 100 muestras aleatorias, del examen (Figura 6.12, izquierda), se aprecia que el índice *Lineal - DEA* exhibe valores de coeficientes de variación significativamente más pequeños que los otros índices. Por otra parte, al comparar *Lineal - PCA 2* con *M.C. - PCA 2*, este último muestra menor coeficiente de variación.

De la Figura 6.12 (derecha) se deduce que las metodología *Lineal-DEA* y *Score 1* son las más estables en términos absolutos en comparación con los otros métodos. La *Media Cóncava* con los 2 tipos de ponderación también muestran estabilidad, sin embargo existe una mayor variabilidad en los resultados. No obstante, se aprecia que la penalización que se realiza con *Media Cóncava* es bastante efectiva, pues *Lineal-PCA 2* constrasta en gran medida con *M.C. - PCA 2* a medida que se aumenta α. Por otra parte, *Score 1*, a pesar de tener mayor promedio de coeficiente de variación, éste sigue siendo menor que con *Media Cóncava - PCA 2*, y *Lineal - FA*.

El aspecto más interesante del análisis radica en si una mayor variabilidad de los *índices* (obtenidos con las diferentes metodologías) es capaz o no de influir en la ordenación de las unidades, es decir, si una mayor variabilidad de los datos se corresponde con

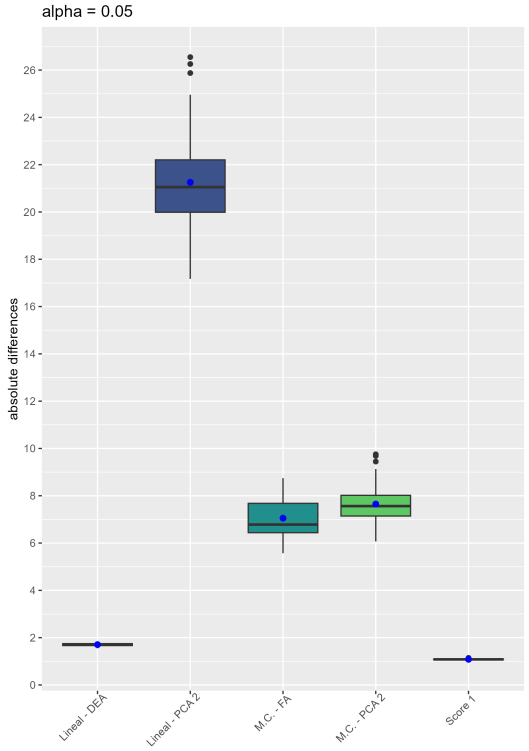

FIGURA 6.13: *Boxplot* de las medias de las diferencias absolutas en las diferentes metodologias.

mayores diferencias absolutas medias de rango con respecto a las puntuaciones relativas inalteradas. Variando los datos de los indicadores simples ($\alpha = 5\,\%$), se obtiene, para cada uno de los 100 muestreos aleatorios y para cada unidad (viaje), las diferencias de clasificación absolutas con respecto a la ordenación obtenida en el caso no perturbado y, por tanto, para cada unidad, la media en las distintas pruebas de las diferencias de clasificación absolutas (Figura 6.13).

De la Figura 6.13; en primer lugar, hay que señalar que los enfoques con *media cóncava* tienen una movilidad similar entre ellas y más amplia que *DEA*. *Score 1* presenta menor diferencia absoluta, seguido por *DEA*, que presenta una ordenación robusta, pues se ve menos afectado por las perturbaciones simuladas en los indicadores elementales DS_i y SS_i (Figuras 6.13 y 6.14). Una de las razones de este comportamiento, puede residir en el hecho de que el rango de variación del indicador, debido al propio procedimiento de su construcción, ya es más amplio en la ordenación original (sin perturbaciones) que en los otros métodos, lo que permitiría "absorber" en cierta medida el efecto de la variabilidad inducida por las simulaciones realizadas.

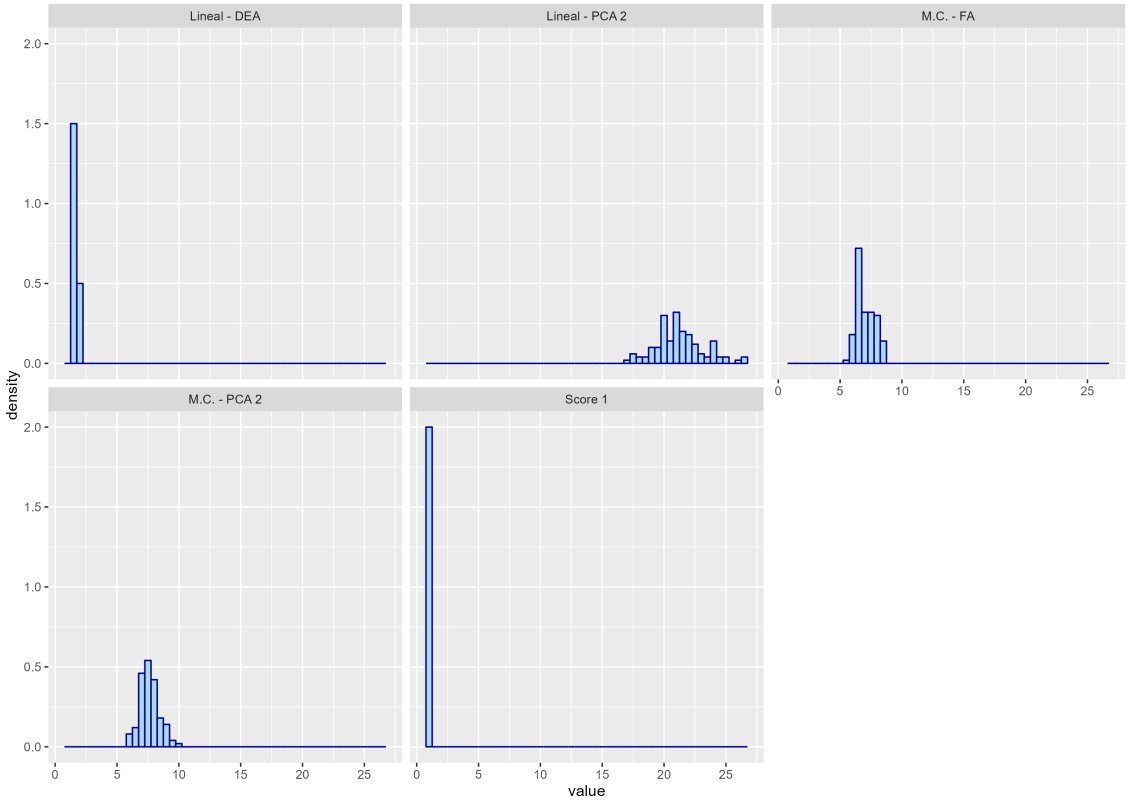

FIGURA 6.14: Histograma de las medias de diferencias absolutas de los índices sintéticos con las metodologias utilizadas.

6.6.3. Resultados

Una rápida visión de las diferencias entre las mediciones puede obtenerse mediante la representación gráfica de los *índices*. El gráfico de la Figura 6.15 compara todos los *índices sinteticos*, ya que toma un viaje cualquiera de una duración, digamos por ejemplo 20 minutos, y lo puntúa usando todos los índices disponibles. La representación gráfica pone de manifiesto los siguientes aspectos:

1. Se confirman nuevamente los numerosos cambios en la clasificación y las variaciones significativas de los valores relativos de los *índices* obtenidos con los distintos métodos de medición, a excepción de *Lineal-DEA* y *Score 1*.
2. Se aprecia como al aumentar la duración del viaje, la puntuación obtenida por *Lineal-DEA*, *Score 1* disminuye en menor intensidad y son casi similares. No obstante, *Lineal-DEA* puntúa con mayor severidad, posiblemente por el cálculo de *Severity Score*, que descuenta puntuación dependiendo de la intensidad de la infracción. Este hecho se puede confirmar al ver como son puntuados viajes con alertas solo del tipo $i = 3, 4, 5$ (curvas similares y casi superpuestas, a excepción para los viajes con alertas $i = 4$, donde se aprecia un desplazamiento vertical, seguramente por la ponderación otorgada), pues estos tipos de alerta por definición tienen *Severity Score* con valor cero. Y por tanto, *Violation Score* corresponde a los valores de *Duration Score*.
3. Otro aspecto a recalcar, según lo observado en secciones anteriores, es que los *índices* que usan las ponderaciones obtenidas mediante *PCA 2* y *FA*, para los tipos de alertas $i = 1, 3, 4, 5, 7$ ($i \neq 2, 6$) ls variación es casi nula (linea horizontal), especialmente para *Lineal- FA*.
4. De los Análisis de Sensibilidad y Robustez, se aprecia que *Score 1*, resulta ser un *índice* casi a la par que *Lineal-DEA*, destacándose el hecho de que el primero fue construido por un conjunto de 20 expertos, y que *Lineal-DEA* logre una robustez superior le da aún más validez.

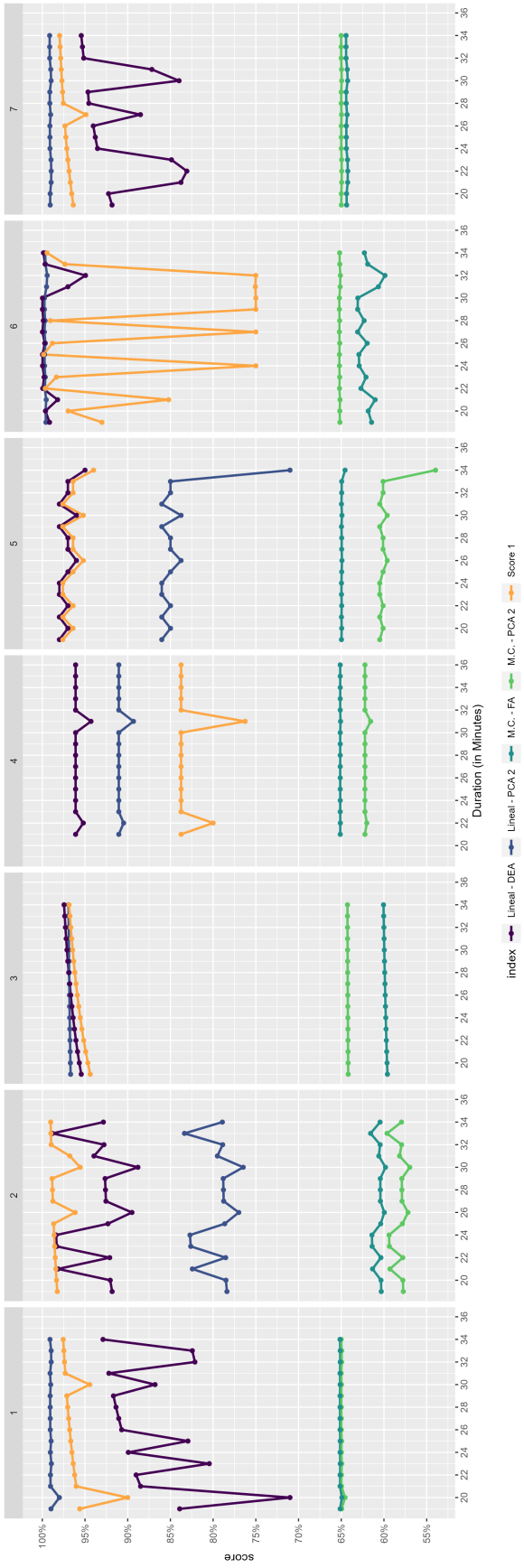

FIGURA 6.15: Viajes puntuados con los diferentes *índices sintéticos*: viajes solo con un tipo de alerta y ordenados en orden creciente de Duración (en minutos).

Capítulo 7

Conclusiones

El objetivo de esta Memoria de Título fue proponer mejoras para el *índice de conducción riesgosa* denominado *ICR*, considerando diferentes metodologías investigadas, especialmente esquemas de *ponderación* y *agregación*.

De los *métodos de ponderación*, *DEA (BoD)*, entrega ponderaciones que puntúan a los viajes lo mejor posible, pero donde cada tipo de infracción resultó ser relevante, con la ventaja de usar los valores brutos de los *indicadores Violation Score*. De este modo, se obtienen ponderaciones aceptables y realistas para cada indicador. Como las ponderaciones se derivan de los datos, sólo se ven influenciadas por los valores de los indicadores y las restricciones impuestas. Con *PCA 1*, *PCA 2*, y *FA* no ocurre lo mismo, pues indicadores como VS_1, VS_3, VS_4 y VS_5 tienen una ponderación muy cercana a cero, especialmente en *PCA 1*. La conclusión respecto a este hecho, es porque la correlación entre los indicadores no es muy alta, y hay registros de alertas de tipos específicos (tipo $i = 2, 6$, es decir, *posible colisión frontal* y *exceso de velocidad*) que predominan sobre todo el conjunto de datos e influyen en la variabilidad de los indicadores.

Para los *métodos de agregación* se usó la agregación *lineal*, debido a que *DEA (BoD)* lo impone en su forma de construcción, pues es la suma ponderada la que busca ser maximizada en cada viaje (unidad), pero también se puso en práctica un método de cálculo que *penaliza* el desequilibrio entre las *dimensiones*; es decir, bajo este contexto, mientras mayor número de registros de alertas en *dimensiones* de manera desigual, mayor "penalización". Este método se denomina *Media Cóncava*. Es un método propuesto por

Casadio en [23], fácil de aplicar aunque haya un número elevado de *indicadores* a usar para determinar el *índice*.

Dada a la relevancia de tener certeza de la calidad de los *índices* obtenidos, y determinar cuál se ajusta de mejor manera al contexto y objetivo buscado, se compararan los resultados de estos *índices* obtenidos con las diferentes ponderaciones y ambos métodos de agregación (*lineal* y *media cóncava*).

De los *índices* obtenidos, *Lineal-DEA* tiene menor coeficiente de variación, y al aumentar la perturbación de los indicadores elementales, es más estable (las diferencias absolutas entre el *índice* obtenido de los indicadores con perturbación y el *índice* obtenido de los indicadores sin perturbación, es menor). *Score 1* también presenta diferencias absolutas pequeñas (menor diferencias absolutas entre todos los otros *índices*), pero con mayor coeficiente de variación que *Lineal-DEA*. Por lo tanto, de la comparación entre *Lineal-DEA* y *Score 1*, queda en evidencia que la inclusión del cálculo de *Severity Score*, aporta congruencia y sólidez a los resultados, incluso a pesar de que se usaron las diferencias entre la velocidad permitida y la velocidad de infracción para las infracciones del tipo $i = 1, 7$. En consecuencia, implementar este *índice sintético*, permitirá obtener resultados más fiables y consistentes, con la ventaja de que su implementación es simple, ya que solo es una suma ponderada, y además no es necesaria la normalización de los indicadores para su cálculo.

En cuanto, a la comparación de resultados para los diferentes viajes, se desprende claramente que los distintos *índices* arrojan clasificaciones de viajes (unidades) significativamente diferentes. No obstante, los *índices* que usaron como ponderación *PCA 2* y *FA*, no califican de manera sensata a los viajes, en especial si los viajes registran alertas del tipo que tiene ponderaciones entre 1 %-2 %; pero este resultado ya era esperado, solo fue confirmado. Por tanto, una conclusión evidente de esto es que los métodos basados en correlación no son recomendables para este caso, porque un conductor si tiene infracciones, en general tendrá solo de un tipo; solo aproximadamente un 7 % de los móviles registraban más de un tipo. Lo que induce a pensar que los conductores tienen hábitos de conducción (que se pueden modificar con capacitación e incentivos) y niveles de atención (desigual a la extensión del viaje), que quizá en las rutas que deben seguir se encuentran con diferentes condiciones de tráfico y/o tipos de carreteras, que pueden

depender del objetivo del viaje; o sea que viajes cortos en rutas urbanas pueden presentar infracciones características diferentes a viajes largos en carreteras. Yendo al contexto de flotas de vehículos, observar este fenómeno puede ser resultado de la cultura de conducción establecida por la empresa para mantener estándares de seguridad y eficiencia en la conducción; es decir, son conductores capacitados y educados para mantener la seguridad vial, y dependiendo del tipo de vehículo y rutas, ciertas infracciones podrían ser más comunes.

Por otra parte, en el proceso de elaboración de un *índice*, además de la *ponderación* y *agregación*, y de las otras etapas de construcción de un *índice*, como la selección de indicadores, la imputación, y la normalización usada, hay que prestar atención a otros aspectos, pues lo que se pretende es obtener un *índice* fiable. Llevando las consideraciones mencionadas a este contexto, dado que algunos de los comportamientos riesgosos ($i = 2, 3, 4$) se obtienen en función de otras variables y no de las usadas, posiblemente diseñar alguno de los *indicadores* con su variable correspondiente, implicaría mejoras y/o más simplicidad en el cálculo del indicador en concreto, y en consecuencia, simplicidad al cálculo del *índice sintético*. Es más, otro aspecto a considerar, y que puede pulir aún más los resultados, es la clasificación del tipo de Vehículo. Actualmente, las velocidades y aceleraciones permitidas son en base a los dos tipos de Vehículos (liviano y pesado); sin embargo, agrupar de otra manera o subdividir alguno de los grupos, puede implicar mejoras; al menos valdría la pena chequear si es así.

Otro aspecto que puede ser interesante de abordar, pero que no se hizo porque los objetivos y foco central eran mejorar el *índice*, y dado que se tiene disponibilidad de la longitud y latitud de las infracciones registradas, sería realizar un análisis de patrones para determinar si hay infracciones ocurriendo en ubicaciones específicas; es más, si hay tipos de alertas específicos ocurriendo en zonas específicas; y de ser así, analizar las causas posibles. Estos factores quizá pueden ser considerados como condiciones del entorno, que están fuera del control del conductor, y se puede agregar "flexibilidad" al *índice* o construir un *índice* que evalúe no solo si la conducción es riesgosa, sino si el viaje es riesgoso y considerar factores externos de las rutas establecidas. Finalmente, la clave sería encontrar un equilibrio entre incentivar un comportamiento de conducción segura y que esta puntuación sea "justa", reconociendo que hay circunstancias que pueden estar más allá del control del conductor.

Apéndice A

Algoritmo DEA (BoD)

Se presenta el algoritmo construido para obtener las ponderaciones. Las variables de entrada de la función *mylp_2* son:

- *outputs*: *data.frame* de l columnas y n filas, donde l es el número de *indicadores* a ponderar, y n el número de unidades (bajo este contexto, n viajes).
- *lower*: cota inferior. Por ejemplo si es 5 %, ingresar 0.05.
- *upper*: cota superior. Por ejemplo si es 30 %, ingresar 0.3.

Nótese que cada fila corresponde a una unidad diferente. La salida del algoritmo es un *data.frame* de $l+1$ columnas y n filas; es decir, el conjunto de ponderaciones para cada indicador en cada unidad, junto con el valor que se obtiene de la suma ponderada (score de la unidad).

```
#### RESOLUCIÓN PROBLEMA LP ######
# Fecha: 07/07/2023
# indicadores (columnas): m = 1 to M
# unidades (filas): k = 1 to n

# max_{w_{k1}, w_{k2}, ... w_{kM}} \sum_{m=1}^M w_{km} y_{km}
# s.a.:
#                        \sum_{m = 1}^M w_{km} y_{jm} <= 100  \forall j = 1, ..., k                    (con 1)
#                                                w_{km} >= 0    \forall m = 1, ...., M                 (con 2)
# \dfrac{ w_{km}y_{jm} }{ \sum{m = 1}^M w_{km}y_{jm}} >= lower\forall m = 1, ... , M,\forall j = 1, ...,k(con 3)
# \dfrac{ w_{km}y_{jm} }{ \sum{m = 1}^M w_{km}y_{jm}} <= upper\forall m = 1, ... , M,\forall j = 1, ...,k(con 4)

mylp_2 = function(outputs, lower, upper){
```

```
 library(readxl)
 library(lpSolve)
 #library(rJava) # fuera de la function
 library(WriteXLS)
 inputs = matrix(0, nrow(outputs), ncol(outputs)) %>% as.data.frame()
 # num de outputs
 s = ncol(outputs)
 # num de DMU
 N = nrow(outputs)
 # defining right hand side, directions
 f.dir = c(rep("<=", N), rep(">=", s), rep("<=", s), rep("=", s))
 f.rhs = c(rep(100, N), rep(0, s), rep(0, s), rep(0, s))
 for (j in 1:N) {
  # defining objective, and matrix of coefficient
  f.obj = cbind(inputs[j, ], outputs[j, ]) %>% as.matrix()
  con1 = cbind(inputs, outputs)
  con2 = cbind(diag(outputs[j, ]),
               (rep(-lower*outputs[j, ],
               ncol(outputs)) %>% as.vector() %>% matrix(nrow = ncol(outputs), byrow = T) %>% as.data.frame()) )
  names(con2) = names(con1)
  con3 = cbind(diag(outputs[j, ]),
               (rep(-upper*outputs[j, ],
               ncol(outputs)) %>% as.vector() %>% matrix(nrow = ncol(outputs), byrow = T) %>% as.data.frame()) )
  names(con3) = names(con1)
  con4 = cbind(diag(s), -diag(s)) %>% as.data.frame()
  names(con4) = names(con1)
  f.con = rbind(con1, con2, con3, con4)
  # solving model
  results = lp ("max", f.obj, f.con, f.dir, f.rhs, scale = 0, compute.sens = F)
  if (j==1) {
   weights = c(results$solution[seq(1, s)])
   effvrss = results$objval
  } else {
   weights = rbind(weights, c(results$solution[seq(1,s)]))
   effvrss = rbind(effvrss, results$objval) }
 }
 weights = weights %>% as.data.frame()
 weights = cbind(effvrss, weights)
 colnames(weights)[2:(ncol(outputs) + 1)] = names(outputs)
 return(weights)
}
```

Bibliografía

[1] De Bruin J., Buma B., Masimira P., Waelbers K., Vos B., *(2020) Advanced Driver Assist Systems BI Metrics and Business Rules.* AngloAmerican

[2] Luke, R., & Heyns, G. J., *(2014). Reducing risky driver behaviour through the implementation of a driver risk management system.* Journal of Transport and Supply Chain Management, 8(1). doi:10.4102/jtscm.v8i1.146

[3] MacKenzie, C. A., *(2014) Summarizing Risk Using Risk Measures and Risk Indices.* Risk Analysis, 34(12), 2143–2162. doi: 10.1111/risa.12220

[4] Anthony (Tony) Cox, L., *(2008). What's Wrong with Risk Matrices?* Risk Analysis, 28(2), 497–512. doi:10.1111/j.1539-6924.2008.01030.x

[5] Thomas, P., Bratvold, R. B., & Bickel, J. E., *(2014). The Risk of Using Risk Matrices.* SPE Economics & Management, 6(02), 056–066. doi:10.2118/166269-pa

[6] Ball, D. J., & Watt, J., *(2013). Further Thoughts on the Utility of Risk Matrices.* Risk Analysis, 33(11), 2068–2078. doi:10.1111/risa.12057

[7] Li, J., Bao, C., & Wu, D., *(2017). How to Design Rating Schemes of Risk Matrices: A Sequential Updating Approach.* Risk Analysis, 38(1), 99–117. doi:10.1111/risa.12810

[8] Hubbard, D. W., *(2020). The Failure of Risk Management, Second Edition.* Wiley Editorial. https://doi.org/10.1002/9781119521914.

[9] Batthyány, K., Cabrera, M., Alesina, L., Bertoni, M., Mascheroni, P., Moreira, N., Picasso, F.,Ramirez, J., & Rojo, V., *(2011). Metodología de la investigación para las ciencias sociales: apuntespara un curso inicial.* Universidad de la República. https://repositorio.minedu.gob.pe/handle/20.500.12799/4544

[10] Arango Serna, M. D., Ruiz Moreno, S., Ortiz Vásquez, L. F., & Zapata Cortes, J. A., *(2017). Indicadores de desempeño para empresas del sector logístico: Un enfoque desde el transporte de carga terrestre.* Desalination, Ingeniare. Revista Chilena de Ingeniería, 25(4), 707–720. doi:10.4067/s0718-33052017000400707

[11] Schöner, H.-P., Pretto, P., Sodnik, J., Kaluza, B., Komavec, M., Varesanovic, D., Chouchane H. & Antona-Makoshi, J., *(2021). A safety score for the assessment of driving style.* Traffic Injury Prevention, 22(5), 384–389. doi:10.1080/15389588.2021.1904508

[12] Hermans, E., Van den Bossche, F., & Wets, G., *(2008). Combining road safety information in a performance index.* Accident Analysis & Prevention, 40(4), 1337–1344. doi:10.1016/j.aap.2008.02.004

[13] Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E., *(2005). Handbook on Constructing Composite Indicators: Methodology and User Guide.* Organisation for Economic Co-operation and Development.

[14] Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G., *(2018). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness.* Social Indicators Research. doi:10.1007/s11205-017-1832-9

[15] Saisana, M., Nardo, M., & Saltelli, A., *(2005). Uncertainty and sensitivity analysis of the 2005 environmental sustainability index.* In D. Esty, T. Srebotnjak, & A. de Sherbinin (Eds.), Environmental sustainability index: Benchmarking national environmental stewardship (pp. 75–78). New Haven: Yale Center for Environmental Law and Policy

[16] Zhang, J. , *(2020). Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey.* WIREs Computational Statistics, 13(5). doi:10.1002/wics.1539

[17] Ram, R. *(1982). Composite indices of physical quality of life, basic needs fulfilment, and income.* Journal of Development Economics, 11(2), 227–247. doi:10.1016/0304-3878(82)90005-0

[18] Krishnakumar, J., & Nagar, A. L. *(2007). On Exact Statistical Properties of Multidimensional Indices Based on Principal Components, Factor Analysis, MIMIC and Structural Equation Models.* Social Indicators Research, 86(3), 481–496. doi:10.1007/s11205-007-9181-8

[19] Mazziotta, M., & Pareto, A., *(2018). Use and Misuse of PCA for Measuring Well-Being.* Social Indicators Research. doi:10.1007/s11205-018-1933-0

[20] Cherchye, L., Moesen, W., Rogge, N., & Puyenbroeck, T. V., (2006). *An Introduction to "Benefit of the Doubt" Composite Indicators.* Social Indicators Research, 82(1), 111–145. doi:10.1007/s11205-006-9029-7

[21] Munda, G. *2011). Choosing Aggregation Rules for Composite Indicators.* Social Indicators Research, 109(3), 337–354. doi:10.1007/s11205-011-9911-9

[22] Casadio Tarabusi, E., & Guarini, G. *(2012). An Unbalance Adjustment Method for Development Indicators.* Social Indicators Research, 112(1), 19–45. doi:10.1007/s11205-012-0070-4

[23] Casadio Tarabusi, E., & Palazzi, P. *(2004). An index for sustainable development.* BNL Quarterly Review, 229, 185–206; Italian transl., Un indice per lo sviluppo sostenibile, Moneta e Credito 226, 123–149.

[24] Mazziotta, C., Mazziotta, M., Pareto, A., & Vidoli, F. *(2010). La sintesi di indicatori territoriali di dotazione infrastrutturale: Metodi di costruzione e procedure di ponderazione a confronto.* Review of Economics and Statistics for Regional Studies, 1, 7–33.

[25] Bu, K., Wallach, D. S., Wilson, Z., Shen, N., Segal, L. N., Bagiella, E., & Clemente, J. C., *(2022). Identifying correlations driven by influential observations in large datasets.* Briefings in bioinformatics, 23(1), bbab482. doi:10.1093/bib/bbab482

[26] Gregory (Greg) L. Snow Ph.D., *(2007) [R] normality test for large sample sizes.* Recuperado de: https://stat.ethz.ch/pipermail/r-help/2007-April/129620.html

[27] *Testing the assumption of normality.* Recuperado en 2008 de: https://analyse-it.com/blog/2008/8/testing-the-assumption-of-normality

[28] Muñoz, J. A. & Amón, I., *(2013). Técnicas para detección de outliers multivariantes*. Recuperado de: http://hdl.handle.net/20.500.11912/6582.

[29] Wada, K., Kawano, M., & Tsubaki, H. , *(2020). Comparison of Multivariate Outlier Detection Methods for Nearly Elliptical Distributions*. Austrian Journal of Statistics, 49(2), 1–17. doi: 10.17713/ajs.v49i2.872

[30] Wada, K., *(2010) Detection of Multivariate Outliers: Modified Stahel-Donoho Estimators (in Japanese)*. Research Memoir of Official Statistics, 67, 89-157. URL http://www.stat.go.jp/training/2kenkyu/pdf/ihou/67/wada1.pdf

[31] Todorov, V., & Filzmoser, P., *(2009). An Object-Oriented Framework for Robust Multivariate Analysis*. Journal of Statistical Software, 32(3), 1–47. doi: 10.18637/jss.v032.i03

[32] Korkmaz, S., Goksuluk, D., & Zararsiz, G., *(2014). MVN: An R Package for Assessing Multivariate Normality*. The R Journal, 6(2), 151-162. doi:10.32614/RJ-2014-031

[33] Peterson, Ryan .A., *(2021). Finding Optimal Normalizing Transformations via bestNormalize*. R J., 13, 310.

[34] Peterson, R. A., & Cavanaugh, J. E., *(2019). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era*. Journal of Applied Statistics, 1–16. doi:10.1080/02664763.2019.1630372

[35] Bartlett, M. S., *(1947). The Use of Transformations*. Biometrics, 3(1), 39. doi:10.2307/3001536

[36] Everitt, B., & Hothorn, T., *(2011). An Introduction to Applied Multivariate Analysis with R*. doi:10.1007/978-1-4419-9650-3

[37] Dauber Daniel, *Sources of bias: Outliers, normality and other "conundrums"*. Recuperado de: https://bookdown.org/daniel_dauber_io/r4np_book/sources-of-bias.html#bootstrapped-regression

[38] Dauber Daniel, *Comparing groups*. Recuperado de: https://bookdown.org/daniel_dauber_io/r4np_book/comparing-groups.html

[39] Stephanie Glen, *"Box's M Test: Definition"* From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/boxs-m-test/

[40] Burchett, W. W., Ellis, A. R., Harrar, S. W., & Bathke, A. C., *(2017). Nonparametric Inference for Multivariate Data: The R Package npmv.* Journal of Statistical Software, 76(4), 1–18. doi: 10.18637/jss.v076.i04

[41] Gero Szepannek., *(2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R.* The R Journal, 10(2), 200-208. doi: 10.32614/RJ-2018-048

[42] Van den Boogaart, K. G., & Tolosana-Delgado, R., *(2008). "compositions": A unified R package to analyze compositional data.* Computers & Geosciences, 34(4), 320–338. doi:10.1016/j.cageo.2006.11.017

[43] Van den Boogaart, K. G., & Tolosana-Delgado, R., *(2013). Introduction.* Analyzing Compositional Data with R, 1–12. doi:10.1007/978-3-642-36809-7_1

[44] Van den Boogaart, K. G., & Tolosana-Delgado, R., *(2013). Fundamental Concepts of Compositional Data Analysis.* Analyzing Compositional Data with R, 13–50. doi:10.1007/978-3-642-36809-7_2

[45] Van den Boogaart, K. G., & Tolosana-Delgado, R., *(2013). Descriptive Analysis of Compositional Data.* Analyzing Compositional Data with R, 51–71. doi:10.1007/978-3-642-36809-7_4

[46] Van den Boogaart, K. G., & Tolosana-Delgado, R., *(2013). Multivariate Statistics.* Analyzing Compositional Data with R, 177–207. 10.1007/978-3-642-36809-7_6

[47] Filzmoser, P., Hron, K., & Templ, M., *(2018). Geometrical Properties of Compositional Data.* Applied Compositional Data Analysis, 107–130. doi:10.1007/978-3-319-96422-5_3

[48] Filzmoser, P., Hron, K., & Templ, M., *(2018). Cluster Analysis.* Applied Compositional Data Analysis, 107–130. doi:10.1007/978-3-319-96422-5_6

[49] Filzmoser, P., Hron, K., & Templ, M., *(2018). Principal Component Analysis.* Applied Compositional Data Analysis, 131–148. doi:10.1007/978-3-319-96422-5_7

[50] Filzmoser, P., & Hron, K., *(2008). Correlation Analysis for Compositional Data*. Mathematical Geosciences, 41(8), 905–919. doi:10.1007/s11004-008-9196-y

[51] Kynčlová, P., Hron, K., & Filzmoser, P., *(2017). Correlation Between Compositional Parts Based on Symmetric Balances*. Mathematical Geosciences, 49(6), 777–796. doi:10.1007/s11004-016-9669-3

[52] Long, W., & Wang, Q., *(2013). Two Methods of Correlation Coefficient on Compositional Data*. Procedia Computer Science, 18, 1757–1763. doi:10.1016/j.procs.2013.05.344