Metodología para la construcción de índices compuestos aplicado a la economía.

Facultad de Ciencias
Instituto de Estadística

Trabajo de titulación presentado por:

Hans Eduardo Leiva Frost

Bajo la supervisión de: Dr. Carlos Henríquez.

Valparaíso, 2020.

Índice general

Dedicado a

... mi familia y
a todos quienes me apoyaron durante mi carrera.
Muchas Gracias.

Resumen

Este documento explica el proceso de elaboración de un índice compuesto, empezando desde la definición de un índice simple, hasta la elaboración de metodologías para la creación de un índice compuesto o sintético, utilizando como insumo índices elaborados periódicamente para medir la economía chilena.

En el capítulo 1, se explica la definición índices, y qué métodos son los más frecuentemente usados para su construcción.

En el capítulo 2, se hablará de los índices compuestos, detallando los requerimientos técnicos para su contrucción, además de las ventajas y desventajas que conlleva el trabajo con índices de esta clasificación.

A continuación, en el capítulo 3, se expone la metodología del proceso de construcción del indicador compuesto, desde el desarrollo del marco conceptual, selección de los indicadores, análisis multivariado, imputación de datos perdidos, normalización de los datos, ponderación de la información, para finalizar con el análisis de robustez y sensibilidad.

En el capítulo 4, se aplicará la metodología a una serie de indices económicos, el cual tendrá como índice de referencia al IMACEC (indicador mensual de activicad económica) índice coyuntural presentado por el Banco Central con frecuencia mensual. La idea principal es obtener un índice compuesto que refleje la economía del país y con esta herramienta poder predecir el comportamiento del IMACEC.

Finalmente, en el capítulo 5 se concluye con la importancia de desarrollar índices compuestos que puedan aportar mejoras en la producción estadística, con el objetivo de evaluar el área de estudio y tener capacidad para tomar mejores decisiones tanto para políticas públicas como privadas.

Lista de abreviaturas

ACP	Análisis de componentes principales
AF	Análisis factorial
ANOVA	Analysis of variance
CE	Comisión Europea
CEPAL	Comisión Económica para America Latina y el Caribe
EM	Expectation Maximixation
GSA	Global Sensitivy Analysis
HDMR	High Dimensional Model Representation
IAC	Índice de actividad de comercio
IC	Índice compuesto
ICT	Índice de costos del transporte
IICOM	Índice de inventario del comercio
IIMCU	Índice de inventario de la minería del cobre
IINVMAN	Índice de inventarios de la industria manufacturera
IMACEC	Indicador mensual de actividad económica
INE	Instituto Nacional de Estadísticas
IPC	Índice de precios del comsumidor
IPEGA	Índice de producción de electricidad, gas y agua
IPMAN	Índice de producción manufacturera
IPMIN	Índice producción minera
IPP	Índice de precios del productor
IPSA	Índice precios selectivo de acciones
IR	Índice nominal de remuneraciones
ISUP	Índice de supermercados
IVAAER	Índice de ventas de actividades artísticas, de entretenimiento y recreativas
IVAASC	Índice de ventas de actividades de alojamiento y de servicio de comidas
IVAI	Índice de ventas de actividades inmobiliarias
IVAPCT	Índice de ventas de actividades profesionales, científicas y técnicas
IVASAA	Índice de ventas de actividades de servicios administrativos y de apoyo
IVIC	Índice de ventas de información y comunicaciones
IVOAS	Índice de ventas de otras actividades de servicios
IVTA	Índice de ventas de transporte y almacenamiento
MAR	Missing at randon
MC	Matriz de covarianza

MCAR	Missing completely at randon
MVA	Missing value analysis
NMAR	Not missing at randon
OCDE	Organización para la Cooperación y el Desarrollo
ONU	Organización de Naciones Unidas
PCA	Principal Component Analysis
PIB	Producto Interno Bruto
RMS	Root Mean Squared
SPSS	Statistical Package for the Social Science
UF	Unidad de fomento

Introducción

En la actualidad, el uso de la información es de vital importancia en la toma de decisiones a nivel político, social, ambiental y empresarial. Es por esto que los indicadores compuestos o sintéticos son tan populares en el mundo desarrollado y se han convertido en una referencia ineludible para los agentes privados y públicos que dan seguimiento a la coyuntura social, ambiental y económica.

Por iniciativa de la Comisión Económica para América Latina y el Caribe (CEPAL) de Naciones Unidas , y con el apoyo de la Organización para la Cooperación y el Desarrollo Económico (OCDE) y la Comisión Europea (CE), se han impulsado metodologías para la construcción de indicadores compuestos, definiendo los pasos a seguir para su construcción, Schuschny y Soto (2009).

Esta investigación pretende explicar uso de estas metodologías y cómo se aplican para la construcción de un indicador compuesto que mida la actividad económica de forma coyuntural y anticipada, considerando la sintetización de índices simples que puedan medir el sector observado.

Lo primero será definir el marco muestral, el cual estará definido por el sector evaluado. A través de técnicas estadísticas se hará un análisis exploratorio de los datos para filtrar los índices que pueden aportar a la construcción del índice compuesto. Se revisará si existe algún dato faltante y *outlier* que tratar. Una vez definidos los índices que sintetizarán el índice compuesto, se procederá a calcular las ponderaciones para la construcción del índice final, en este caso se obtendrán ponderaciones por tres métodos y se decidirá cual de ellos es más aplicable para este cálculo. Una vez obtenidos los índices compuestos a través de los tres métodos, se compararán con el índice de referencia y se eligirá el mejor. Para concluir, se procede a hacer análisis de robustez y sensibilidad.

Objetivo

El objetivo general del trabajo de titulación, es investigar la metodología detrás de la construcción de indicadores compuestos que permitan medir conceptos multidimensionales relacionados con el área económica del país. La finalidad es construir un indicador que resuma la información entregada por indicadores simples, en un solo indicador.

El trabajo se resume en los siguientes objetivos específicos:

(1) Adaptar las directrices metodológicas en la construcción de indicadores compuestos otorgados por la CEPAL, para el análisis multidimensional de indicadores económicos.

(2) Estudiar la aplicación de técnicas estadísticas en cada paso de la construcción de indicadores compuestos.

Capítulo 1

Índices

En el mundo de la ciencia, es muy importante el análisis de una o más variables, y mucho más cuando se necesitan comparar. En el caso de la comparación de variables existen dos métodos, por diferencia o por cociente. Esta última elimina el problema de unidades de medida, aunque no deja de estar afectado por otros factores, como el de elegir el periodo de referencia para realizar las comparaciones.

"En general diremos que un número índice es aquella medida estadística que permite estudiar las fluctuaciones o variaciones de una sola magnitud o de más de una en relación al tiempo o al espacio" Sánchez, J. (2004).

A continuación se mostrarán los distintos tipos de números índices, (Rojo, J. *et al.*, 2018):

1.1. Tipos de números índices

Según se quiera registrar la evolución de una o más magnitudes, a través del tiempo:

- Índices simples: recogen la evolución del precio, la cantidad o el valor de un único bien o producto.

- Índices compuestos, complejos o sintéticos: recogen la evolución conjunta de los precios, las cantidades o los valores de k bienes o productos. A su vez, los índices complejos se clasifican como:

  - Sin ponderar: todas las magnitudes o componentes tiene la misma importancia, es decir, los mismos pesos. Los k bienes o productos se consideran con el mismo peso.

  - Ponderados: cada magnitud o componente tiene un peso diferente asignado en función de diversos criterios. Los k bienes o productos se consideran con distinto peso, los cuales señalan

la importancia relativa de cada uno de los bienes.

Según su magnitud:

- Índices de precios: estudian la evolución de los precios de un bien o de un conjunto de bienes.
- Índices de cantidades: estudian la evolución de la cantidad producida o consumida de un bien o de un conjunto de bienes.
- Índices de valores: estudian la evolución del valor de un bien o de un conjunto de bienes.

1.2. Números índices simples

Según (Rojo, J. *et al.*, 2018) un índice simple es el cociente entre la magnitud en el período corriente y la magnitud en el período base. Generalmente, se multiplica por cien y se lee en porcentaje. No presenta gran utilidad en sí mismo y su interés radica en que son el punto de partida de la construcción de los índices complejos y en que algunas de sus propiedades sirven para evaluar la bondad de éstos. Se considera la magnitud X en distintos períodos de tiempo. El índice simple de la magnitud X en el período t con respecto al período 0 será:

$$I_{t/0} = \frac{X_t}{X_0} \times 100 \qquad .$$

Se interpreta como la variación en tanto por ciento, experimentada por la magnitud X entre el periodo 0 y el periodo t. Con todo, en los desarrollos y propiedades de los números índices ha de considerarse la primera de las expresiones.

Los índices simples pueden recoger la evolución de los precios de un bien, de su producción (cantidad) o de sus valores.

1.3. Números índices complejos

Para (Rojo, J. *et al.*, 2018) frecuentemente, el interés no está en comparar precios, cantidades o valores de un único bien, sino en conocer la evolución conjunta de esas magnitudes para un grupo más o menos numeroso de bienes. Para ello, se intenta resumir la información suministrada por los índices simples en un único índice que denominaremos compuesto, complejo o sintético. El propósito es obtener un número índice sencillo pero que reúna la mayor cantidad de información posible. Según

prime la sencillez o la conservación de la máxima información se tendrá dos tipos de índices complejos: sin ponderar y ponderados.

Sea la magnitud X (precios, cantidades, valores, ...) relativa a n bienes $X_1, X_2, ..., X_n$. Los valores de la magnitud para los distintos bienes en los distintos períodos de tiempo se recogen en la siguiente tabla:

Periodo	X_1	$\cdots$	X_i	$\cdots$	X_n
0	X_{10}	$\cdots$	X_{i0}	$\ldots$	X_{n0}
1	X_{11}	$\cdots$	X_{i1}	$\cdots$	X_{n1}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
t	X_{1t}	$\cdots$	X_{it}	$\cdots$	X_{nt}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
T	X_{1T}	$\cdots$	X_{iT}	$\cdots$	X_{nT}

A partir de los índices simples de X para cada uno de los bienes,

$$I_{t/0}(X_1) = \frac{X_{1t}}{X_{10}}, ..., I_{t/0}(X_i) = \frac{X_{it}}{X_{i0}}, ..., I_{t/0}(X_n) = \frac{X_{nt}}{X_{n0}} \qquad .$$

Se podrá obtener un índice complejo para X utilizando un promedio, índice complejo que resume la información proporcionada por los índices simples. Los más habituales son los que se obtienen a partir de medias aritméticas o medias agregativas. También se pueden aplicar medias geométricas o medias armónicas.

Índices complejos sin ponderar: todos los índices simples, y por tanto todas las componentes, tienen el mismo peso.

Media aritmética

$$\begin{aligned} I_{t/0}(X) &= \frac{I_{t/0}(X_1) + ... + I_{t/0}(X_i) + ... + I_{t/0}(X_n)}{n} \\ &= \frac{\sum_{i=1}^{n} I_{t/0}(X_i)}{n} \qquad . \end{aligned}$$

Media agregativa

$$I_{t/0}(X) = \frac{X_{1t} + ... + X_{it} + ... + X_{nt}}{X_{10} + ... + X_{i0} + ... + X_{n0}} = \frac{\sum_{i=1}^{n} X_{it}}{\sum_{i=1}^{n} X_{i0}} \quad .$$

Índices complejos ponderados: en este caso, los índices simples tienen distinto peso, es decir, se asignan diferentes ponderaciones a las componentes o magnitudes. Sean estos pesos $\alpha_1, ..., \alpha_i, ..., \alpha_n$.

Media aritmética

$$I_{t/0}(X) = \frac{I_{t/0}(X_1)\alpha_1 + ... + I_{t/0}(X_i)\alpha_i + ... + I_{t/0}(X_n)\alpha_i}{\alpha_1 + ... + \alpha_i + ... + \alpha_n} = \frac{\sum_{i=1}^{n} I_{t/0}(X_i)\alpha_i}{\sum_{i=1}^{n} \alpha_i} \quad .$$

Los índices complejos de media aritmética sin ponderar se pueden considerar un caso particular de éste, cuando todas las ponderaciones son iguales a 1.

Media agregativa

$$I_{t/0}(X) = \frac{X_{1t}\alpha_1 + ... + X_{it}\alpha_i + ... + X_{nt}\alpha_n}{X_{10}\alpha_1 + ... + X_{i0}\alpha_i + ... + X_{n0}\alpha_n} = \frac{\sum_{i=1}^{n} X_{it}\alpha_i}{\sum_{i=1}^{n} X_{i0}\alpha_i} \quad .$$

Los índices complejos de media agregativa sin ponderar se pueden considerar un caso particular de éste, cuando todas las ponderaciones son iguales a 1. Además, las medias agregativas se pueden expresar como medias aritméticas ponderadas. Así, para la media agregativa sin ponderar resulta:

$$I_{t/0}(X) = \frac{I_{t/0}(X_1)X_{10} + ... + I_{t/0}(X_i)X_{i0} + ... + I_{t/0}(X_n)X_{n0}}{X_{10} + ... + X_{i0} + ... + X_{n0}}$$
$$= \frac{\sum_{i=1}^{n} I_{t/0}(X_i)X_{i0}}{\sum_{i=1}^{n} X_{i0}} \quad .$$

Y para la media agregativa ponderada:

$$I_{t/0}(X) = \frac{I_{t/0}(X_1)X_{10}\alpha_1 + ... + I_{t/0}(X_i)X_{i0}\alpha_i + ... + I_{t/0}(X_n)X_{n0}\alpha_n}{X_{10}\alpha_1 + ... + X_{i0}\alpha_i + ... + X_{n0}\alpha_n}$$
$$= \frac{\sum_{i=1}^{n} I_{t/0}(X_i)X_{i0}\alpha_i}{\sum_{i=1}^{n} X_{i0}\alpha_i} \quad .$$

Por tanto, la media aritmética ponderada de los índices simples es la forma más general de agregar índices simples para obtener un índice complejo:

$$I_{t/0}(X) = \frac{I_{t/0}(X_1)\alpha_1 + ... + I_{t/0}(X_i)\alpha_i + ... + I_{t/0}(X_n)\alpha_n}{\alpha_1 + ... + \alpha_i + ... + \alpha_n}$$
$$= I_{t/0}(X_1)A_1 + ... + I_{t/0}(X_i)A_i + ... + I_{t/0}(X_n)A_n \quad ;$$

donde son las ponderaciones normalizadas,

$$A_i = \frac{\alpha_i}{\sum_{i=1}^{n} \alpha_i}, \quad (\sum_{i=1}^{n} A_i = 1) \quad .$$

Habitualmente estas ponderaciones se expresan en tanto por ciento o en tanto por mil.

En la actualidad uno de los índices más importante es el índice de precios del consumidor (IPC) que mide la relación entre precio y cantidad en un periodo, en general mide cómo evoluciona el gasto de una familia media.

Los métodos más utilizados para el cálculo de precios son el índice de Laspeyres y el índice de Paasche, como se definen a continuación.

1.3.1. Índice de Laspeyres

Según de la Fuente (2013), se analizan las varaciones debidas a los cambios en los precios de un conjunto de artículos ponderándolos siempre por las mismas cantidades.

El índice de Laspeyres se define como la media aritmética ponderada de los índices simples de precios. El criterio de ponderación es $p_{i0} \times q_{i0}$, con lo cual:

$$IP_l = \frac{\sum_{i=1}^{n} p_{it} \times q_{i0}}{\sum_{i=1}^{n} p_{i0} \times q_{i0}} \times 100 \qquad .$$

Los criterios para elección del período base son variados, fundamentalmente se requiere que sea un año no irregular o normal.

El incoveniente del índice de Laspeyres es que supone que siempre se adquieren las mismas cantiades que en el período base.

1.3.2. Índice de Paasche

Siguiendo a de la Fuente (2013), el índice de Laspeyres se cuestiona en ocaciones, ya que, parece poco realista suponer que las cantidades compradas o adquiridas en el año de referencia, no varian en el tiempo.

Como ejemplo, no parece muy realista la hipótesis de que en años de sequía, y en consecuencia, de subidas importantes de los precios de los productos agrarios, las cantidades demandadas sean iguales.

Se planteó la necesidad de disponer de otros índices que, con la finalidad de medir la variación de precios de determinado conjunto de artículos, no estuviera sujeto a la restricción de suponer que siempre se adquirían las mismas cantidades que en el período base.

El índice de Paasche se define como media aritmética ponderada de los indices simples de precios. El criterio de ponderación es $p_{i0} \times q_{it}$, con lo cual:

$$IP_p = \frac{\sum_{i=1}^{n} p_{it} \times q_{it}}{\sum_{i=1}^{n} p_{i0} \times q_{it}} \times 100 \qquad .$$

El cálculo del índice de Paasche es laborioso, ya que, exige calcular las ponderaciones $p_{it} \times q_{it}$ para cada período corriente.

Otro inconveniente adicional, el índice de precios de cada año, solo se puede comparar con el del año base.

Capítulo 2

Índice compuesto

El indicador compuesto, también llamados índice, es una función simplificada que buscar explicar un concepto multidimensional en una sola variable o índice simple. Estos pueden ser de forma cuantitativa o cualitativa, dependerá de los requerimientos del analista.

De forma técnica, *"Un indicador se define como una función de una o más variables, que conjuntamente miden una característica o atributo de los individuos en estudio"* Schuschny y Soto (2009).

Cabe recordar que para la construcción de un indicador compuesto hay que tener claro dos condiciones básicas:

1.- La definición clara del atributo que se desea medir.

2.- La existencia de información confiable para poder realizar la medición.

La primera definición otorgará un "sustento conceptual" mientras que la segunda otorgará "validez". Para la generación de un indicador compuesto, se debe tener claro el objetivo por el cual se está creando. Generalmente, el indicador compuesto se construye con la finalidad de medir la unidad de análisis de algún área en particular, el cual puede ser utilizado como punto de partida para observar la tendencia del fenómeno no directamente detectable de un conjunto de variables en una sola unidad.

Estos indicadores son herramientas eficaces en la contribución de políticas públicas o privadas, y logran llamar la atención del público dando lugar a la creación de argumentos convincentes y contribuyendo en los debates de políticas integradas que promueven el tema de interés. La mayoría de los estudios de indicadores compuestos miden conceptos multidimensionales de países para su posterior comparación. De esta forma parece más fácil comparar indicadores compuestos, que observar los diferentes indicadores simples entre países, Schuschny y Soto (2009).

En la actualidad, está en incremento el uso de indicadores compuestos, dando lugar a la importancia de crear este tipo de indicador, donde la información se ve resumida en un valor. Desde la

OCDE, ONU y la CEPAL entregan diferentes metodologías y guías para su construcción Schuschny y Soto (2009).

2.1. Requerimientos técnicos

Siguiendo a Castro Boñano (2002), es válido definir algunas condiciones que a priori deberían exigirse a un indicador compuesto:

- Existencia y determinación: la función matemática que define el indicador debe existir y tener solución perfectamente determinada.

- Exhaustividad: el indicador debe ser tal que aproveche al máximo, sin redundancia y en forma útil, la información suministrada por los indicadores y variables que la componen.

- Monotonía: el indicador ha de responder positivamente al cambio positivo de las componentes y viceversa. Ello obliga, en algunos casos, a cambiar el signo de las variables que lo componen cuyas correlaciones pudieran estar invertidas.

- Unicidad: el indicador compuesto ha de ser único para una situación dada.

- Invariancia: el indicador debe ser invariante frente a cambios de origen o de escala de sus componentes.

- Homogeniedad: la función matemática que define al indicador compuesto: $I = f(x_1, ..., x_p)$ debe ser homogénea de grado 1, es decir debe cumplirse que:

$$f(\alpha x_1, ..., \alpha x_p) = \alpha f(x_1, ..., x_p) \qquad .$$

- Transitividad: si (a), (b) y (c) son tres situaciones distintas que dan lugar a tres indicadores, debería verificarse que:

$$si \quad I(a) > I(b) > I(c) => I(a) > I(c) \qquad .$$

2.2. Ventajas y desventajas de construcción de indicadores compuestos

En particular para la construcción de indicadores compuestos, se pretende reducir las limitantes por medio de una construcción metodológica adecuada.

A continuación se mostrará ventajas y desventajas que se obtienen a través de la construcción de indicadores compuestos, Bas (2014).

2.2.1. Ventajas

- Sirven para resumir temas complejos y multidimensionales. Integran y sintetizan diferentes dimensiones del concepto que miden. Facilitan la evaluación de la eficacia de las políticas y la rendición de cuentas.
- Permiten disponer de una "imagen de contexto"(o imagen general) que facilita la interpretación de temas complejos. Por el contrario, encontrar una tendencia en cada uno de los indicadores por separado es mucho más costoso y complicado. Facilitan, por tanto, la tarea de construcción de rankings de unidades de análisis en temas multidimensionales.
- Atraen el interés público por su capacidad de proporcionar una comparación y evolución entre diferentes unidades de análisis.
- Ayudan a reducir el tamaño de la lista de indicadores simples.

2.2.2. Desventajas

- Pueden transmitir un mensaje confuso si están mal construidos o son mal interpretados.
- La "imagen general"que transmiten puede llevar a conclusiones simplistas en la opción de los políticos. Los indicadores compuestos se deben utilizar en combinación con los indicadores simples para extraer conclusiones políticas adecuadas.
- Su construcción requiere de faces en las que se deben hacer juicios de valor. Para una correcta construcción, los juicios de valor deben ser transparentes y deben basarse en principios estadísticos siempre que sea posible. Además, deben ir acompañados de un análisis de sensibilidad y robustez de los resultados.
- Su diseño debe realizarse a partir de un conjunto de información medible. Ello requiere que los datos estén ampliamente disponibles, sus frecuencias de muestreo sean razonables en relación a los objetivos que se plantean y las unidades de análisis hayan consensuado un nivel tolerable de armonización sobre las estadísticas e indicadores a utilizar.

Como se observa, en la realización de un indicador compuesto se encontrarán ventajas y desventajas. La literatura dice que para construir un indicador compuesto robusto y confiable, se debe

tener claridad del marco conceptual del indicador y además encontrar información confiable para la construcción de este.

Para mejorar la metodología de construcción de indicadores compuestos hay que tener en claro tres pasos:

- Un marco conceptual sólido (estructura del fenómeno a medir)
- Un proceso sólido de construcción (selección de indicadores, normalización y agregación)
- Buena calidad de los datos subyacentes.

Las críticas más comunes en la construcción de indicadores compuestos son la generalización excesiva, la selección de indicadores simples, la comparabilidad de indicadores compuestos en situaciones diversas, el contrapeso que se produce entre indicadores de naturaleza muy diversa u otros. Otras críticas que comparten los especialistas sobre la construcción de indicadores compuestos es que no tienen una adecuada y justificada base teórica a partir de la cual el análisis y las técnicas de agregación utilizadas para sintetizar los indicadores simples en un solo indicador.

Capítulo 3

Metodología

Para la construcción de un índice compuesto, se debe tener en claro la metodología involucrada en su desarrollo. Para esto se verá algunas experiencias de expertos en el tema.

Construcción por etapas

Siguiendo la experiencia realizada por Nardo *et al.* (2005a), a lo largo del proceso de construcción de un indicador compuesto se debe seguir una serie de etapas minuciosas. Estas etapas son:

1.- Desarrollo de un marco conceptual.
2.- Selección de indicadores.
3.- Imputación de datos perdidos.
4.- Análisis multivariado.
5.- Normalización de los datos.
6.- Ponderación de la información.
7.- Agregación de la información.
8.- Análisis de robustez y sensibilidad.

La construcción del marco es primordial, ya que alimenta de manera conceptual la generación del indicador compuesto. Es a partir de aquí que se justifica la construcción del indicador, proporcionándonos el camino definido, para posteriormente analizar los indicadores que serán parte del indicador compuesto.

Ya definido el marco conceptual, se evaluarán los indicadores que se desean incorporar. Este proceso previo, consiste en una búsqueda de los indicadores que, dentro del marco definido puedan ser construidos o utilizados si es que ya existen, para posteriormente ser incorporados en un indicador compuesto.

Elegidos los indicadores y variables que formarán parte del indicador compuesto, se hará un análisis exploratorio para evaluar si efectivamente los datos con la información seleccionada están en

concordancia con las ideas que dieron forma al marco conceptual. Este es el primer proceso de validación de los indicadores seleccionados, en donde puede haber problemas de valores faltantes o falta de información. Esto puede conducir a problemas en las etapas posteriores , dado que puede generar errores en los análisis que conduzcan al final a conclusiones incorrectas, lo que hace recurrir a las metodologías de imputación de datos perdidos o faltantes.

Los indicadores o variables seleccionados, generalmente, estarán medidos en distintas escalas, por lo que es necesario normalizarlos para que puedan ser agregados de manera comparable. Ya hecho esto, es necesario definir un factor de peso o ponderaciones para cada indicador o variable, donde finalmente se generará el agregado del indicador compuesto. Concluido el proceso de construcción del indicador compuesto, será necesario presentarlo en formato claro y entendible, ya sea de manera gráfica o tabular.

Finalizado el indicador compuesto, como un medio de validación, se procede a hacer un análisis de sensibilidad, el cual consiste en evaluar si pequeñas variaciones en los datos contenidos en los indicadores y variables que se incluyen en la agregación conducen efectivamente a pequeñas variaciones en el valor de indicador compuesto, lo cual no está garantizado, pero es requerido como un elemento de robustez.

Cabe destacar que todas las etapas mencionadas anteriormente pueden ser realizadas de distintos modos. Si bien, no en todas las etapas se requerirá hacer uso de alguna metodología, por ejemplo, no hay datos faltantes o outlier que tratar, es importante considerarlas.

3.1. Desarrollo de un marco conceptual

"Lo que está mal definido será erróneamente medido"(Nardo et al., 2008)

El marco conceptual debe ser lo más claro y detallado posible para disponer de una mejor definición del índice compuesto y de las relaciones entre los indicadores simples que lo componen. Para ello, es necesario tener categorizado, de forma amplia, el contexto de análisis y tener comprensión del fenómeno a medir. El marco conceptual teórico debe ser basado en lo que se desea medir y no en lo que está disponible para medir. Es conveniente que en esta etapa participen expertos en el área de estudio o grupos de interés, para tener en cuenta múltiples puntos de vista y aumentar la solidez del marco conceptual y del conjunto de indicadores.

Según Bas (2014), no todos los conceptos multidimensionales poseen fundamentos teóricos sólidos y empíricos a partir de los cuales se encamina un marco conceptual teórico bien definido. Los indicadores compuestos construidos en el ámbito político y en la actualidad, como la competitividad, el desarrollo sostenible u otros, podrán ser muy discutidos, puesto que la investigación de dichos campos todavía está en desarrollo. A continuación, se verán los pasos a seguir en el desarrollo de un marco conceptual teórico (Nardo *et al.*, 2008):

- Definición del concepto multidimensional. La definición del concepto multidimensional a medir debe proporcionar una idea clara y concisa de lo que se quiere evaluar mediante el índice compuesto. Sin embargo, algunos conceptos complejos son muy difíciles de definir y medir con precisión, estos pueden ser objeto de controversia entre las partes interesadas. Finalmente, los usuarios de los índices compuestos deben evaluar su calidad y relevancia.
- Clasificación en subgrupos o dimensiones. Es recomendable dividir el concepto multidimensional en subgrupos o dimensiones. No es necesario que las dimensiones sean estadísticamente independientes entre sí, pero los vínculos existentes deben describirse teórica y empíricamente cuando la consistencia lógica en relación con los fenómenos considerados favorezca a la organización conceptual y no a la puramente estadística. Esta división en dimensiones facilita la asignación de pesos en los diferentes factores. Se debe hacer referencia al marco conceptual conectando los diferentes subgrupos o dimensiones con los indicadores simples.
- Identificación de los criterios de selección para el conjunto de indicadores subyacentes. Los criterios de selección se utilizan con una guía para la decisión de la inclusión o exclusión de un indicador simple en el indicador compuesto global. Deben ser lo más claro y concisos posible para facilitar la elección del conjunto de indicadores relevantes que va a ser objeto de estudio en posteriores análisis.

3.2. Selección de los indicadores

"La calidad de un indicador compuesto es consecuencia directa de los indicadores simples que lo definen" (Nardo *et al.*, 2008).

Por ello, los indicadores simples deben seleccionarse sobre la base de su calidad, su relevancia, su disponibilidad y la frecuencia con la que se muestrean.

En este caso, los indicadores deben cumplir ciertas propiedades que permitan medir el grado de cumplimiento de los objetivos planteados en el estudio.

Según Bas (2014), existen diferentes criterios para la selección de los indicadores simples que aseguran la calidad de un sistema de indicadores para la construcción de un indicador compuesto. Un ejemplo, basado en los cuatro criterios siguientes, se propuso para la construcción del índice de innovación de las empresas e industrias por la Comisión Europea (*European Commission* - DG MARKT, 2001):

- Relevancia política: se deben seleccionar los indicadores simples que resulten relevantes en la toma de decisiones políticas.
- Redundancia: si dos indicadores aportan la misma información se recomienda seleccionar solo uno de ellos.
- Correlación: si dos indicadores están muy correlacionados, pero ambos transmiten mensajes políticos fuertes se pueden incluir en la lista final de indicadores relevantes.

– Disponibilidad: se recomienda utilizar indicadores que están disponibles para un gran número de unidades de análisis y que puedan obtener regularidad de una base de datos actualizada.

Para obtener los indicadores que alimenten el indicador compuesto, primero se tiene que definir bien las dimensiones a medir. Dentro de estas dimensiones, se identifican grandes procesos y a su vez, se obtiene una serie de indicadores asociados que son considerados en primera instancia como relevantes. En este primer proceso de selección, se realiza sobre técnicas de consenso, donde finalmente se obtienen aquellos indicadores factibles para el indicador compuesto, de forma que su medición resulte viable.

Cabe destacar que la elección de indicadores debe estar guiada por el marco conceptual teórico, sin embargo, el proceso de selección puede ser un tanto subjetivo. Es aquí donde deben actuar expertos y partes interesadas, por lo tanto, puede ser que no haya un conjunto definido y único de indicadores.

Por otro lado, el proceso de selección debe estar documentado mediante la construcción de metadatos donde se especifiquen las características de los indicadores su disponibilidad, fuentes responsables de su cálculo, el tipo de indicador, las unidades de medida con las que se expresa, entre otros.

Es importante mencionar la mayor limitación que presenta el diseño de un indicador compuesto y que consiste en la ausencia de información estadística en la que se basa el indicador. Para ello, existen métodos de imputación de datos faltantes.

Con todo lo anterior para finalizar esta etapa de la construcción de un indicador compuesto, se concluye que, si no se realiza una selección correcta de los indicadores simples o si esta no abarca las principales dimensiones del objeto de estudio, difícilmente el índice desarrollado mostrará algo representativo acerca del concepto que se desea estudiar.

3.3. Imputación de datos perdidos

3.3.1. Detección de valores atípicos y datos faltantes

Según Bas (2014), los valores atípicos son observaciones muy diferentes al resto de los datos. Su presencia puede tener un gran impacto en posteriores análisis, provocando sesgos indeseables si se trata de datos atípicos problemáticos (valores mal recogidos en la base de datos, errores de transcripción u otros). Por otra parte, existen casos atípicos que no son problemáticos a pesar de ser diferentes a la mayor parte de la muestra o la población. Estos son, por ejemplo, datos reales con comportamientos anómalos.

Se pueden detectar valores atípicos desde una perspectiva univariante o multivariante (Hair *et al.*, 2007):

– Nivel univariante: en este caso se examina la distribución de las observaciones detectando como

valores atípicos aquellos que caen fuera de los rangos de la distribución. Se recomienda estandarizar los datos (con media cero y desviación estándar uno) para poder realizar comparaciones entre los indicadores simples. Para muestras pequeñas (de 80 observaciones o menos) se suelen considerar como valores atípicos aquellos que tienen un valor, en valor absoluto de 2,5 o superior. Para muestras grandes el umbral se sitúa entre 3 o 4. También se pueden detectar valores atípicos analizando la asimetría y la curtosis de las distribuciones de los indicadores simples. Si la asimetría, en valor absoluto, es mayor a 1 o la curtosis mayor a 3,5 se considera que existe algún caso atípico en el indicador simple evaluado (Saisana, 2010). Al realizar un histograma del indicador simple también se puede detectar el caso atípico. En este caso, el analista debe decidir si mantenerlo o excluirlo del análisis, o simplemente ajustar el valor para obtener una asimetría y una curtosis dentro de los rangos establecidos.

– Nivel multivariante: se pueden identificar los valores atípicos multivariantes con la medida D^2 de Mahalanobis o mediante la T^2 de Hotelling. Sin embargo, en el ámbito de los índices compuestos se suelen detectar posibles valores atípicos a nivel univariante (Nardo *et al.*, 2008; Saisana, 2010; Annoni, 2010). No obstante, en este estudio se ha considerado interesante nombrar las técnicas multivariantes para la detección de valores atípicos.

Una vez identificado los valores atípicos, se debe decidir si incluir o no el valor juzgando no solo las características del caso atípico, sino también los objetivos del análisis.

Cabe mencionar que después de identificar los valores atípicos Hair *et al.* (2007) recomienda realizar un análisis exploratorio para evaluar si efectivamente, los datos están en concordancia con las ideas que dieron lugar a su elección. Esta es la primera etapa de validación de la utilidad de los indicadores seleccionados, el cual puede manifestar problemas de ausencia parcial de información y con ello puede conducir a problemas en las siguientes etapas. Por lo tanto, es necesario utilizar técnicas de imputación para datos perdidos o faltantes.

A menudo, la falta de información obstaculiza el desarrollo de un índice compuesto robusto. Esto se debe a factores de procedimientos como errores en la recogida de datos, fallos al completar cuestionarios y traspasos de información. Y otras, se deben netamente con el encuestado, quien a veces se niega a responder ciertas preguntas. Para evitar estos casos de datos ausentes el analista debe cerciorarse de la buena calidad de la recogida de los datos y de la respuesta en las encuestas durante el diseño de la investigación.

Para poder aplicar una solución a la ausencia de datos se debe averiguar el grado de aleatoriedad en los datos faltantes. Para ello, supóngase que se observan dos indicadores I_1 e I_2 , siendo I_1 el indicador sin datos ausentes e I_2 el indicador con datos ausentes. A continuación, se plantean tres patrones de comportamiento a los que pueden obedecer los datos ausentes según grado de aleatoriedad (Hair *et al.*, 2007):

– Perdida de datos completamente al azar (*missing complety at randon* – MCAR). Cuando hay un mayor nivel de aleatoriedad el proceso es completamente aleatorio. En este caso, los valores observados de I_2 son una muestra aleatoria de los valores de I_2 sin un proceso subyacente que

tiende a sesgar los datos observados. Si los datos ausentes siguen este patrón, cualquier solución se podría aplicar sin tener en cuenta el impacto de cualquier otro indicador o proceso de datos ausentes.

- Perdida de datos al azar (*missing at randon* –MAR). Los datos ausentes siguen un patrón MAR si los valores ausentes de I_2 dependen de I_1, pero no en I_2. Es decir que los valores observados para I_2 representan una muestra de los valores reales de I_2 para cada valor de I_1, pero los datos observados para I_2 no representan necesariamente una muestra verdaderamente aleatoria para todos los valores de I_2. Aunque el proceso de datos ausentes es aleatorio en la muestra, sus valores no son generalizables para la población.
- Perdida de datos sistemática (*not missing at randon* –NMAR). Si se encuentra un proceso de datos ausentes entre I_1 e I_2, donde existen diferencias significativas para los casos de I_2 con datos válidos y datos ausentes en función de los valores de I_1, entonces los datos ausentes no son aleatorios. En la práctica es común que se presenten situaciones en las que los datos faltantes no siguen un patrón completamente aleatorio (MCAR) y tampoco aleatorio (MAR).

Desafortunadamente, no hay ningún método estadístico para tratar la perdida de datos sistemática y, a menudo, no se puede diferenciar si los datos faltantes se han producido por perdida ocasional o sistemática. La mayoría de los métodos que se utilizan para la imputación de datos requieren un mecanismo de perdida de los datos completamente aleatoria MCAR o al menos aleatoria MAR.

3.3.2. Técnicas de imputación

A continuación, se veran algunos tipos de tratamiento para datos faltantes. Cabe destacar que la literatura es muy extensa en este ámbito y está en actual desarrollo, vease (Litle y Rubin, 1987; Little 1988; Hair *et al.*, 2007; Medina y Galván, 2007), a continuación algunos métodos de imputación:

Imputación por eliminación

En este procedimiento se omiten los valores ausentes para análisis posteriores. En la construcción del indicador compuesto significa omitir un indicador para todas las unidades de análisis u omitir una unidad de análisis completa, la cual implica descartar datos que pueden haber sido costosos de obtener. Este tipo de imputación no es muy utilizada, puesto que el objetivo del indicador compuesto es proporcionar una puntuación para todas las unidades de análisis manteniendo el mayor número posible de indicadores subyacentes.

Por otra parte, los errores estándar serán más grandes en muestras reducidas dado que se utiliza menos información. Como regla general, si un indicador tiene más de un 5 % de datos faltantes los casos no se eliminan (Little y Rubin, 1987)

Imputación simple

La imputación simple consiste en el método efectivo de sustitución de los valores ausentes por datos estimados sobre la base de información disponible en la muestra. Esta medida puede realizarse de muchas maneras, que van desde una sustitución directa de valores hasta procesos de imputación basados en relaciones entre indicadores. A continuación, se van a nombrar los métodos más ampliamente utilizados, aunque existen otras formas de imputación simple (Little y Rubin, 1987; Hair *et al.*, 2007):

Modelización implícita

Según Little y Rubin (1987), esta técnica de imputación la atención se centra en un algoritmo con supuestos implícitos que deben evaluarse. Además de la necesidad de verificar cuidadosamente si las suposiciones implícitas son razonables y se ajustan a la cuestión tratada, el riesgo de este tipo de imputación de datos faltantes es considerar al conjunto de datos resultante como completo, olvidándose que se realizó una imputación. La modelización implícita incluye:

- Imputación *Hot deck*: en este método de imputación se llenan los vacíos de información a partir de unidades con comportamiento similar. Por ejemplo, en una encuesta se podría agregar la información que respondieron los encuestados a los que no respondieron, pero con similares características.
- Imputación por sustitución de caso: en este método las observaciones con datos ausentes se sustituyen por otras observaciones no muéstrales. Un ejemplo común es reemplazar un encuestado que está en la muestra, pero que ha sido dificil de contactar o que tiene gran cantidad de datos ausentes por otro encuestado que no está en la muestra, preferiblemente muy similar al de la observación original. Este método es el que más se utiliza para sustituir las observaciones con los datos ausentes completos, aunque también puede emplearse para reemplazar observaciones con menor cantidad de datos ausentes.
- Imputación *cold deck*: este método el analista sustituye los datos ausentes por un valor constante derivado de fuentes externas o investigación previa.

Modelización explícita

Para Little y Rubin (1987), la modelización explícita se realiza considerando un modelo estadístico. La modelización explicíta incluye:

- Imputación por el método de la media/mediana/moda no condicionada: este método consiste en sustituir los valores ausentes por un indicador cuyo valor medio se calcula sobre todas las respuestas disponibles. Así, las respuestas de la muestra disponible se usan para calcular el valor de sustitución (media/mediana/moda). Este tipo de imputación es uno de los métodos más empleados.

  La sustitución de los valores ausentes por la media es extensamente utilizada, sin embargo, presenta tres inconvenientes. En primer lugar, invalida las estimaciones de la varianza derivadas de las fórmulas estándar de la varianza para conocer la verdadera varianza de los datos. Por la manera en que se realiza la sustitución de los datos, la suma de cuadrados de las desviaciones de

las observaciones respecto de la media permanece inalterada, pero se incrementa el tamaño de la muestra, lo cual origina que la varianza del indicador disminuya y se generen, de forma artificial, intervalos de confianza más estrechos. En segundo lugar, la distribución real de los valores se encuentra distorsionada por la sustitución de los datos ausentes por la media. Finalmente, este método modifica la correlación observada puesto que todos los datos ausentes tendrán un valor único constante. Sin embargo, tiene la ventaja de poderse llevar a cabo fácilmente y proporcionar una información completa para todos los casos.

A continuación, se va a formalizar en términos matemáticos el procedimiento de imputación por el método de medias no condicionadas. Sea X_q la variable aleatoria asociada al indicador simple q con $q = 1, \ldots, Q$ y X_{qc} el valor observado de X_q para la unidad de análisis c, con $c = 1, ..., M$. Para algunas unidades de análisis c, sea m_q el número de valores registrados en X_q, y $M - m_q$ el número de valores ausentes. La media no condicional se calcula de la siguiente forma (Nardo *et al.*, 2008):

$$\bar{X}_q = \frac{1}{m_q} \sum_{registrados} x_{qc} \qquad .$$

De forma similar, se calcula la mediana (el valor que divide en dos partes iguales la distribución de la variable aleatoria) y la moda (el valor con mayor frecuencia) de la distribución sobre la muestra de datos disponible con el fin de sustituir los valores faltantes por estos valores.

– Imputación por regresión: en este método se usa el análisis de regresión para predecir los valores de un indicador basándose en su relación con otros indicadores del conjunto de datos. La ventaja de este procedimiento es el uso que se le da a las relaciones ya existentes en la muestra como base de predicción. Sin embargo, la aplicación de este método i) refuerza las relaciones ya existentes en los datos, por lo que los datos resultantes finalmente son más característicos de la muestra y menos generalizables, ii) se subestima la varianza de la distribución (menos cuando se añaden valores estocásticos a los valores estimados), iii) se supone que el indicador con los datos ausentes tiene correlaciones sustanciales con otros indicadores, en caso contrario este método no es preferible como método de imputación y iv) los valores estimados puede que no pertenezcan a los rangos validos de los indicadores, requiriendo por tanto alguna forma de ajuste adicional.

A pesar de los inconvenientes que presenta este método, se utiliza mucho cuando se presentan niveles moderados de dispersión de los datos ausentes y cuando las relaciones entre indicadores son suficientemente significativas.

Para describir el método supóngase que se tiene un conjunto de datos de $h - 1 < Q$ indicadores con datos completos $(x_1, \ldots, x_{h-1})$ y un indicador x_h observado en r unidades de análisis, pero con datos perdidos de *M-r* de ellas (Nardo *et al.*, 2008). Este método realiza una regresión de x_h con $(x_1, \ldots, x_{h-1})$ usando las r observaciones completas de tal manera que la imputación se haga a partir de la predicción:

$$\hat{X}_{ih} = \hat{\beta}_0 + \sum_{j=1}^{h-1} \hat{\beta}_j x_{ij}, \qquad i = 1, ..., M - r \qquad .$$

Por lo general, la estrategia para definir la mejor regresión se basa en un procedimiento que presenta dos fases. En la primera, todos los subconjuntos diferentes de predictores se tienen en cuenta en la regresión. En la segunda fase, el mejor subconjunto se determina usando uno de los siguientes criterios:

- El valor de R^2.
- El valor del cuadrado medio residual (RMS)
- El valor de la C_k de Mallows.
- Regresión por pasos (regresión *"stepwise"*)

Una variante de este procedimiento es la imputación por "regresión estocástica", en la cual los datos faltantes se obtienen con un modelo de regresión más un valor aleatorio asociado al término de error.

$$\hat{X}_{ih} = \hat{\beta}_0 + \sum_{j=1}^{h-1} \hat{\beta}_j x_{ij} + \varepsilon_i, \qquad i = 1, ..., M - r \sim N(0, \sigma^2) \qquad ;$$

en que σ^2 es la varianza residual de la regresión de x_h con $(x_1, \ldots, x_{(h-1)})$ basada en los r casos completos.

Este procedimiento garantiza variabilidad en los valores imputados y contribuye a reducir el sesgo en la varianza y en el coeficiente de determinación del modelo.

- Algoritmo *Expectation-Maximization* (EM): Es un algoritmo propuesto por Dempster, Laird y Rubin (1977) que presenta una técnica iterativa general para realizar una estimación de máxima verosimilitud de parámetros de problemas con datos ausentes. En el algoritmo intervienen dos etapas (las etapas"E" y "M"). La etapa de *Expectation*, "E", realiza las mejores estimaciones posibles de los datos ausentes mientras que en la etapa de *Maximization*, "M", realiza estimaciones de los parámetros (medias, desviaciones típicas o correlaciones) con la suposición de reemplazamiento de todos los datos ausentes. El proceso continúa realizando las dos etapas hasta obtener una diferencia despreciable en los valores estimados con respecto a etapas anteriores y reemplazar todos los datos ausentes.

  Sea x el conjunto de datos. Para describir el algoritmo EM supóngase que los datos se generan por un modelo descrito por la función de distribución $f(x|\theta)$, donde $\theta \, \epsilon \, \Omega_\theta$ es el vector de parámetros desconocido del espacio de parámetros. La función $f(x|\theta)$ captura la relación entre el

conjunto de datos y los parámetros del modelo de los datos. Como se desconocen los parámetros, pero se conocen los datos tienen sentido estimar la probabilidad de observar cierto conjunto de parámetros dados los datos, es decir la función de verosimilitud.

El algoritmo *EM* alterna etapas de expectación en la que se calcula la verosimilitud esperada mediante la inclusión de variables latentes como se fueran observables y una etapa de maximización, donde se calculan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso anterior. Los parámetros que se encuentran en la etapa de maximización se usan para comenzar la nueva etapa de expectación y así, el proceso se repite recursivamente. Dado x, sea la función de verosimilitud $L(x|\theta)$ proporcional a $f(x|\theta)$:

$$L(x|\theta) = k(x)f(x|\theta) \quad con \quad k(x) > 0$$

Para M observaciones $(x_1, \ldots, x_M)$ consideradas independientes e idénticamente distribuidas conforme a la distribución normal $N(0, \sigma^2)$ se tiene que la función de densidad conjunta es:

$$f(x|\mu, \sigma^2) = (\frac{1}{2\Pi\sigma^2})^{\frac{-M}{2}} exp(-\frac{1}{2}\sum_{i=1}^{M}\frac{(x_i - \mu)}{\sigma^2})$$

Por tanto, el logaritmo de la función de verosimilitud será:

$$\begin{aligned} l(\mu, \sigma_2|x) &= ln[L(\mu, \sigma_2|x)] \\ &= ln[k(x)] - \frac{M}{2}log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{M}\frac{(x_i - \mu)}{\sigma^2} \end{aligned}$$

La condición de primer orden para la maximización de esta función es:

$$\frac{\partial \quad lnL(\theta|x_{obs})}{\partial\theta} = 0$$

es decir:

$$\frac{\partial \quad lnL(\mu, \sigma^2|x_{obs})}{\partial\mu} = 0, \quad \frac{\partial \quad lnL(\mu, \sigma^2|x_{obs})}{\partial\sigma^2} = 0 \quad .$$

De esta forma se buscan aquellos valores de $\theta \, \epsilon \, \Omega_\theta$ que más se acomodan a la muestra de datos x. Dado que los datos perdidos forman parte de x, el algoritmo debe estimar tanto μ y σ^2 como valores perdidos. Para llegar a esta solución se suele proceder iterativamente. En la etapa de

maximización se estiman los parámetros como si no hubiera datos perdidos (estos son reemplazados por estimaciones) y en la etapa de expectación se estiman datos perdidos a partir de los conocidos y los previamente estimados. Luego se establece un ciclo repitiendo estas etapas hasta alcanzar un cierto criterio de convergencia preestablecido, como por ejemplo la ausencia de cambios significativos de los valores estimados. El resultado final de la obtención de un máximo local de la función de verosimilitud (Nardo *et al.*, 2008).

El algoritmo *EM* puede aplicarse en muchas situaciones en las que se desea estimar un conjunto de parámetros que describen una distribución de probabilidad subyacente con solo una parte observada de los datos completos producidos por la distribución. Además, es muy simple de construir tanto en el aspecto conceptual como en el práctico, cada etapa tiene una interpretación estadista y siempre converge. El único inconveniente que presenta este método de imputación es que la convergencia puede resultar de larga duración cuando se parte un conjunto de datos con muchos valores ausentes.

Imputación múltiple

– Imputación múltiple: la imputación múltiple utiliza métodos de simulación de Monte Carlo vía cadenas de Markov y sustituye los datos faltantes a partir de un número de simulaciones (normalmente ese número se ubica entre 3 y 10) (Rubin, 1987). La metodología consta de varias fases y en cada cadena una de las simulaciones realizadas se aplica métodos estadísticos convencionales para analizar la matriz de datos completa. Posteriormente, se combinan los resultados para generar estimadores robustos, su error estándar de confianza (Medina y Galván, 2007).

A pesar de minimizar los problemas que pueden surgir en cualquier método de imputación simple, no se recomienda que se apliquen los métodos de imputación múltiple como la mejor opción estadística para la sustitución de datos faltantes (Little, 1988; Robins, Rotnitzky y Zhao, 1994). Cada situación es diferente y dependiendo del indicador que se analice, del porcentaje de respuesta y de su patrón de comportamiento es probable que se presenten situaciones en las que se obtengan mejores resultados con los métodos de imputación simple que con el que se acaba de nombrar. Por tanto, no se va a describir con más detalle este procedimiento de imputación.

Hay que tener en cuenta que el uso de técnicas de imputación no puede sustituir por completo la información perdida, por lo que en primer lugar se debe tratar de recuperar la información de los datos ausentes desde las fuentes originales. También comentar que el abuso de métodos de imputación puede llevar a conclusiones que no reflejan la realidad de lo que se está midiendo.

3.4. Análisis multivariado

"Analizar la estructura subyacente de los datos sigue siendo un arte" (Nardo et al.,2008)

En la última década la construcción de indicadores compuestos ha ido en aumento, diseñados

principalmente por diversos organizamos nacionales o internacionales. Generalmente, se seleccionan indicadores simples de forma aleatoria, prestando poca atención a las posibles relaciones entre ellos. Lo que puede llevar a la construcción de indicadores compuestos confusos y con un mensaje erróneo para el público. Es por esto, que se debe analizar la naturaleza de los datos subyacentes con mucho cuidado previamente a la construcción del índice compuesto. Es aquí donde el análisis multivariado permite evaluar la capacidad del conjunto de datos y facilitar la comprensión de las elecciones metodológicas tomadas en el proceso de construcción del índice.

"La información debe agruparse y analizarse como mínimo en función de dos dimensiones del conjunto de datos: los indicadores individuales y las unidades de análisis que corresponden a cada una de las observaciones sobre las cuales se miden los indicadores simples definidos" (Nardo *et at.*, 2008). En las siguientes secciones se describen las técnicas más frecuentes en el contexto de indicadores compuestos.

3.4.1. Información agrupada con respecto a los indicadores individuales

En primer lugar, el analista debe decidir si la estructura conceptual del indicador compuesto está bien definida (Etapa 1) y si el conjunto de indicadores individuales disponible es apropiado para describir el fenómeno a medir (Etapa 2). Esta decisión puede tomarse con la ayuda de la opinión de expertos y de la estructura estadística del conjunto de datos. Diferentes procedimientos estadísticos se pueden utilizar para explorar si las dimensiones del fenómeno están, desde un punto de vista estadístico, bien equilibradas en el indicador compuesto. En caso de no ser así sería necesaria una revisión de los indicadores individuales o del marco conceptual teórico (esto sería necesario si existe una justificación teórica de dicha agrupación de indicadores en dimensiones).

3.4.2. Análisis factorial

El análisis factorial es una técnica estadística de modelación de datos cuya idea principal es explicar la variabilidad de Q indicadores observados en términos de un número menor m de variables no observadas llamadas factores, cuya influencia queda matizada por unos pesos o cargas incluyendo un término de error.

Los indicadores observados se modelan como combinaciones lineales de factores más expresiones de error (Uriel, 1995).

$$\begin{aligned} I_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + ... + \alpha_{1m}F_m + e_1 \\ I_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + ... + \alpha_{2m}F_m + e_2 \\ &\vdots \\ I_Q &= \alpha_{Q1}F_1 + \alpha_{Q2}F_2 + ... + \alpha_{Qm}F_m + e_Q \qquad , \end{aligned}$$

donde I_1 son los indicadores observados que se consideran tipificados o estandarizados, $i = 1, \ldots, Q$.

$\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{im}$ son las cargas factoriales o saturadas del indicador I_i en los factores $F_1, F_2, \ldots, F_m$ e $i = 1, \ldots, Q$.

$F_1, F_2, \ldots, F_m$ son los factores no correlacionados, cada uno de ellos con media cero y varianza 1.

e_i son los errores independientes e idénticamente distribuidos con media cero e $i = 1, \ldots, Q$.

Existen muchos procedimientos para resolver el modelo expuesto: factores de máxima verosimilitud, mínimos cuadrados no ponderados, mínimos cuadrados generalizados, u otros. Pero el procedimiento más usado para extraer los m factores es el análisis de componentes principales (ACP), puesto que permite la construcción de pesos que representan la información contenida en los indicadores simples. Este método de extracción es el que más se aplica en la construcción de indicadores compuestos y, por tanto, el que se va a describir a continuación. Nótese que los diferentes procedimientos para la resolución del modelo implican diferentes valores de los factores extraídos y diferentes pesos para los indicadores simples influyendo, por tanto, en la puntuación del índice compuesto final.

3.4.3. Análisis de componentes principales (*Principal Component Analysis–PCA*)

La técnica de análisis de componentes principales fue descrita por Karl Pearson en 1901. Una descripción de su metodología fue introducida más tarde por Hotelling en 1933. El objetivo de esta técnica es explicar la mayor parte de la variabilidad total observada en un conjunto de variables con el menor número de componentes posible (Uriel, 1995). Esto es posible trasformando las variables correlacionadas en un nuevo conjunto de variables no correlacionadas, denominadas factores o componentes principales, relacionadas con las variables originales mediante una transformación lineal y ordenadas de forma decreciente según el porcentaje de variabilidad que explican. Dicho de otro modo, la técnica utilizada en el análisis de componentes principales consiste en proyectar la nube de observaciones sobre un subespacio afín de dimensión menor, determinado de tal manera que la nube proyectada se deforme lo menos posible.

El análisis de componentes principales está relacionado con el análisis factorial, pero existen ciertas diferencias i) los componentes principales se construyen para explicar las varianzas, mientras que los factores se construyen para explicar covarianzas o correlaciones entres las variables, ii) el análisis de componentes principales es una técnica descriptiva, mientras que el análisis factorial presupone un modelo estadístico formal de generación de datos como se ha descrito anteriormente (Peña, 2002). Para describir la técnica de análisis de componentes principales (ACP) supóngase que se tienen Q indicadores en el análisis I_i , $i = 1, \ldots, Q$,medidos sobre n unidades de análisis (Uriel, 1995; Nardo *et al.*, 2005). Sea X la forma matricial que representan los datos del estudio:

$$X = \begin{pmatrix} I_{11} & \ldots & I_{Q1} \\ \vdots & \ddots & \vdots \\ I_{1n} & \ldots & I_{Qn} \end{pmatrix} \epsilon \quad R^{n \times Q} \qquad .$$

La matriz de covarianza muestral CM de los datos originales es:

$$CM = E[(X - E[X])(X - E[X])'] = \begin{pmatrix} \sigma_1^2 & \dots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_n^2 \end{pmatrix} \epsilon \quad R^{n \times n} \qquad .$$

Para evitar que algún indicador tenga una influencia indebida en las componentes principales se suele estandarizar la matriz de variables originales. En este caso, la matriz de varianzas-covarianzas se convierte en la matriz de correlaciones:

$$R = \begin{pmatrix} 1 & \dots & r_{1Q} \\ \vdots & \ddots & \vdots \\ r_{Q1} & \dots & 1 \end{pmatrix} \epsilon \quad R^{QxQ} \quad con \quad r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}}, \quad 1 \leq i, \quad j \leq Q \qquad .$$

Las componentes principales pueden estimarse a partir de cualquiera de estas dos matrices que son las que proporcionan información acerca de la relación en la variabilidad observada en las variables cuando son tomadas de dos en dos.

Las componentes principales son un conjunto de variables Z_j, $j = 1, \dots, Q$, ortogonales entre si que surgen de una combinación lineal de las variables originales con la propiedad de contener en conjunto la misma varianza total que el conjunto original.

$$\begin{aligned} Z_1 &= \alpha_{11} I_1 + \alpha_{12} I_2 + ... + \alpha_{1Q} I_Q \\ Z_2 &= \alpha_{21} I_1 + \alpha_{22} I_2 + ... + \alpha_{2Q} I_Q \\ &\vdots \\ Z_Q &= \alpha_{Q1} I_1 + \alpha_{Q2} I_2 + ... + \alpha_{QQ} I_Q \end{aligned}$$

La primera componente principal retendrá la máxima porción de la varianza del conjunto de las variables originales, la segunda retendrá el máximo de la varianza restante y así sucesivamente hasta la última componente principal que contendrá el resto de la varianza no incluida en las componentes principales antecesoras.

La primera componente se expresará como la combinación lineal siguiente:

$$\begin{pmatrix} Z_{11} \\ \vdots \\ Z_{1n} \end{pmatrix} = \begin{pmatrix} I_{11} & \dots & I_{Q1} \\ \vdots & \ddots & \vdots \\ I_{1n} & \dots & I_{Qn} \end{pmatrix} \begin{pmatrix} a_{11} \\ \vdots \\ a_{1Q} \end{pmatrix} \quad ; \quad Z_1 = X * a_1$$

El vector a_1 se obtiene maximizando la varianza de Z_1:

$$MAX \quad VAR(Z_1)) \quad con \quad VAR(Z_1) = \frac{\sum_{i=1}^{n} Z_{1i}^2}{n}$$
$$= \frac{1}{n} Z_1' Z_1$$
$$= \frac{1}{n} a_n' X' X a_1 = a_1' [\frac{1}{n} X' X] a_1$$

$$s.a. \sum_{i=1}^{Q} a_{1j}^2 = 1$$

- Si las variables están normalizadas, $[\frac{1}{n} X' X] = R$
- Si las variables están expresadas como desviaciones típicas alrededor de la media, $[\frac{1}{n} X^{'} X] = CM$

Sin pérdida de generalidad, supóngase la segunda situación. Por tanto, para maximizar la varianza de Z_1, se construye el lagrangiano:

$$\mathcal{L} = a_1^{'} * CM a_1 - \lambda(a_1^{'} a_1 - 1)$$

Cuya condición de primer orden es:

$$\frac{\partial \quad \mathcal{L}}{\partial a_1} = 2CM * a_1 - 2\lambda a_1$$
$$= 0, \quad (CM - \lambda I) * a_1 = 0$$

Y dado que a_1 es un vector no nulo, se tiene que λ es el autovalor de la matriz de covarianzas y a_1 su autovector.

El resto de las componentes se obtienen aplicando el mismo procedimiento, pero añadiendo una nueva restricción de ortogonalidad respecto de las componentes anteriores, ya calculadas.

En resumen, el ACP trata de encontrar los autovalores λ_j de la matriz de covarianzas (MC) de los datos originales que son las varianzas de las componentes principales. Además, se cumple:

$$\lambda_1 + \lambda_2 + ... + \lambda_Q = \sigma_1^2 + \sigma_2^2 + ... + \sigma_Q^2 \qquad .$$

Con esto se obtiene Q componentes principales, tantas como variables del análisis. El siguiente paso es seleccionar $P < Q$ componentes que conserven la mayor cantidad de varianza acumulada de

los datos originales.

Los coeficientes de correlación entre las componentes principales Z_j y los indicadores I_i se llaman cargas factoriales o puntos variables. En el caso que los indicadores no estén correlacionados, las cargas factoriales coinciden con los pesos a_{ij}. Cuanto mayor sea la carga factorial de un indicador con respecto a una componente significa que la relación entre ambos es alta.

Análogamente al coeficiente de correlación de Pearson, el cuadrado de la carga factorial del indicador I_i, a_{ij}^2, se denomina comunalidad e indica la proporción de la variabilidad del indicador i que queda explicada por la componente principal j. De esta forma, la comunalidad del indicador I_i , h_i^2 , se define como la suma de todas las comunalidades de cada factor respecto al indicador (Cuadras, 2006):

$$h_1^2 = a_{i1}^2 + a_{i2}^2 + ... + a_{iQ}^2 \qquad .$$

Las puntuaciones de cada una de las observaciones en cada componente principal se llaman puntos-individuo. Los puntos-individuo para una observación particular respecto a un componente principal se calculan estandarizando el valor para cada indicador, multiplicándolo por la correspondiente carga factorial de ese componente principal y sumando los productos.

La falta de correlación en las componentes principales indica que cada una de ellas mide "dimensiones estadísticas" diferentes en los datos. No siempre la aplicación de ACP reduce el número de indicadores originales en un número de variables latentes menor. Esto último ocurre cuando los indicadores originales no están correlacionados. Se conseguirá reducir notablemente el número de componentes principales cuando los indicadores originales estén altamente correlacionados, ya sea positiva como negativa.

Una herramienta que se suele emplear después de seleccionar el número de componentes y que mejora la interpretación de los resultados es la rotación de factores. Esta herramienta se basa en girar los ejes de referencia hasta alcanzar una determinada posición. La rotación de factores hace que se redistribuya la varianza de los primeros factores a los restantes. De esta forma se consigue un patrón de factores más simple y más fácilmente interpretable (Hair *et al.*, 2007) . Existen varios métodos de rotación, pero según la literatura, los métodos más comunes son la rotación ortogonal "Varimax" y la rotación oblicua "Oblimin" (Nardo *et al.*, 2005; Hair *et al.*, 2007). No existen reglas concretas que indiquen la selección de una técnica de rotación. La rotación afecta a las cargas factoriales de las variables y puede ocurrir que los grupos obtenidos con ACP sin aplicar rotación no sean los mismos que al aplicar rotación.

En resumen, los pasos a seguir para aplicar un ACP/AF como un análisis exploratorio son:

1.- Cálculo de la matriz de correlaciones: si las correlaciones entre indicadores simples son bajas es muy probable que estos no compartan factores comunes.
2.- Identificar el número de factores necesarios para representar el conjunto de datos y el método para

calcularlos.
3.- Aplicar, en el caso que sea necesario, una rotación sobre los factores para facilitar interpretación de los resultados.

Para poder aplicar ACP/AF se deben cumplir una serie de supuestos que se van a enumerar a continuación:

Para aplicar el ACP se debe primero partir de los siguientes supuestos (Nardo *et al*, 2005a):

Poseer un número de casos suficientemente grande. Esto supondrá la adopción de alguna de las siguientes reglas empíricas planteadas por varios autores:

Regla del 10: Disponer de al menos 10 casos por cada variable.
El 3 a 1: que el número de casos sea el triple que el de variables.
El 5 a 1: Otros autores plantean respetar una relación 5 a 1 entre casos y variables.
Regla del 100: El número de casos debería ser 5 veces el número de variables y superiores a 100.
Regla del 150: Disponer de más de 150 cuando hay muy pocas variables correlacionadas.
Regla del 200: Tener más de 200 casos, sin importar el número de variables.
Regla de la significancia: Tener 51 casos más que el número de variables, con el fin de poder realizar la prueba chi-cuadrado.

Nótese la dispersión que poseen estas reglas empíricas. La elección de aquella a aplicar dependerá de la relación que se establezca entre la disponibilidad de información y el número de variables involucradas, así como del nivel de robustez pretendido.

Que no haya sesgos de selección de las variables. La exclusión de variables relevantes junto con la inclusión de otras irrelevantes afectará ciertamente a la matriz de covarianza y por lo tanto la representatividad del resultado que se obtenga.

Que no haya datos atípicos (*outliers*). Como en el caso de otras técnicas estadísticas, la presencia de datos atípicos puede afectar las interpretaciones que devienen de un análisis de componentes principales.

- Linealidad: El análisis de componentes principales es una técnica basada en el uso del álgebra lineal y por eso es claramente conveniente que la relación entre las variables sea lineal.

- Normalidad multivariada: este supuesto es conveniente validar si se busca realizar pruebas de contraste estadístico. Si se supone que las variables están distribuidas a partir de funciones de distribución diferentes, será más complicado hacer dichas pruebas pues generalmente las herramientas de software sólo contemplan las basadas en el supuesto de normalidad.

- Correlación fuerte entre variables: Al aplicar ACP/AF a una matriz de covarianza con correlaciones

bajas se obtienen tantos factores como variables originales se tengan y, por lo tanto, no se podrá reducir la dimensionalidad del conjunto inicial de datos. Pruebas para comprobar la correlación entre variables: media de Kaiser-Meyer-Olkin, test de esfericidad de Bartlett, correlación observada y reproducida, matriz de correlaciones, matriz de correlaciones anti-imagen, entre otros.

3.4.4. Coeficiente de Crombach

Otro de los métodos utilizados para el análisis multivariado de los indicadores, es el Coeficiente Alfa de Crombach (1951). Este coeficiente mide la consistencia interna, es decir, indica que tan bien está representada la información de múltiples variables en un solo indicador. Este coeficiente toma valores entre 1 y 0, y sirve para comprobar si un indicador que se está evaluando, recopila información de las variables que lo componen que es defectuosa y por lo tanto llevaría a conclusiones equivocadas o si se trata de un instrumento fiable que mide lo que dice. El valor del coeficiente será mayor cuanto mayor sea la correlación entre las variables. Cuanto más se acerque el índice al extremo 1, mejor es la fiabilidad de la selección de variables propuesta, considerando una fiabilidad respetable a partir de 0,70.

El coeficiente de Combrach se puede calcular como:

$$\alpha = \frac{p}{p-1}\left(\frac{\sigma_I^2 - \sum_{i=1}^{p}\sigma_{xi}^2}{\sigma_I^2}\right) \quad .$$

Donde σ_I es la varianza del indicador y σ_{xi} es la de cada una de las p variables. Para construir este estimador, se supone que el indicador se calcula como la suma simple de todos los sub-indicadores; así mismo, vale estimar el coeficiente para cada unidad de análisis. El estimador mide la fracción de la variabilidad total de la muestra de variables debido a su correlación. Si no hay correlación y las variables son independientes entre sí, su valor es nulo, mientras que cuando la correlación es total, valdrá la unidad. Por eso, un valor cercano a uno indicará que las variables consideradas miden correctamente el fenómeno latente que se desea representar.

Un ejercicio empleado es observar cómo varía el coeficiente incluyendo y excluyendo indicadores. Esto ayuda a detectar existencia de dimensiones en los indicadores simples. Si su valor se incrementa con la exclusión se puede afirmar que el indicador eliminado no está muy correlacionado con el resto.

3.4.5. Análisis de conglomerados (*Clustering Analysis*)

El análisis de conglomerados es una herramienta que sirve para observar similitudes entre variables y crear grupos o clúster. El objetivo de este método es clasificar las unidades de análisis las cuales son homogéneas entre sí y heterogéneas entre los grupos.

Este método se puede utilizar para la construcción de indicadores compuestos como:

- Un método estadístico de agregación de indicadores.
- Una herramienta de diagnóstico para explorar cada elección de distintas alternativas al momento de construir el indicador compuesto.
- Un procedimiento para agrupar unidades de análisis por su similitud y a partir de allí imputarle a algunos de ellos datos perdidos con el fin de disminuir la dispersión de la información.
- Una técnica de análisis de los resultados.

Algunos métodos nombrados por Bas (2014), para variables o indicadores cuantitativos, las distancias más utilizadas son:

Distancia euclidiana

$$d(x,y) = \left(\sum_{i=1}^{Q}(x_i - y_i)^2\right)^{\frac{1}{2}} \quad .$$

x,y representan dos unidades de análisis.

Una ventaja de utilizar este método es que la distancia no se ve afectada por la inclusión de nuevas unidades de análisis, por lo tanto. No se ve afectada por *outlier*. Al contrario, su valor depende de la escala o unidad de medida.

Distancia euclidiana al cuadrado

$$d(x,y) = \sum_{i=1}^{Q}(x_i - y_i)^2 \quad .$$

Al no calcularse la raíz cuadrada las variables más alejadas tendrán más peso.

Distancia euclidiana al cuadrado normalizada

$$d(x,y) = \sum_{i=1}^{Q}\left(\frac{x_i - y_i}{\sigma_i}\right)^2 \quad .$$

Distancia de Chebychev

$$d(x,y) = máx|x_i - y_i| \qquad .$$

Medida útil cuando se desea definir elementos como diferentes si hay una diferencia apreciable en cualquiera de los mismos.

Distancia de Manhattan (*city-block* o de Hamming)

$$d(x,y) = \sum_{i=1}^{Q} |x_i - y_i| \qquad .$$

Distancia que atenúa la presencia de valores atípicos.

Distancia de Mahalanobis

$$D^2 = (x_i - x_j)' V^{-1} (x_i - x_j) \qquad .$$

Siendo V la matriz de varianzas y covarianzas de las variables que intervinieron.

Una ventaja de este método, es la distancia que tiene en cuenta la correlación de las variables. Al contrario, no cumple la desigualdad triangular que es una condición que debe cumplir cualquier distancia definida.

Distancia de Minkowski

$$d(x,y) = \left(\sum_{i=1}^{Q} (x_i - y_i)^s \right)^{\frac{1}{r}} \qquad .$$

Distancia ventajosa cuando se desea incrementar o atenuar el valor de una variable, el valor s mide las diferencias de las variables mientras que r controla el peso de la distancia entre unidades de análisis diferentes.

Proporción de discrepancias

$$d(x,y) = \#veces\ que\ x_i \neq y_i \qquad .$$

Distancia útil cuando se trabaja con información discreta o categorizada.

En el caso de variables cualitativas, se pueden calcular medidas de similitud a través de variables binarias. 1 en caso favorable y 0 en caso no favorable. El coeficiente de similitud da una medida de semejanza entre dos objetos en relación con los indicadores.

El análisis puede ser jerárquico, dando lugar a una estructura en forma anidada de árbol, o no jerárquico si se establece un número de clases predeterminado. La principal ventaja de los métodos jerárquicos es la facilidad de interpretación del árbol taxonómico que resulta.

Para realizar un análisis de agrupamientos es necesario definir una distancia. Toda definición de distancia debe satisfacer ciertas condiciones definidas en un espacio métrico. Sean X, Y y Z tres vectores definidos en el espacio de las p variables. Entonces la distancia entre ellos es un número real que debe satisfacer las condiciones de:

1.- No negatividad:

$$d(x,y) \begin{cases} = 0 & \text{sí} \quad x = y \\ > 0 & \text{sí} \quad x \neq y \end{cases} \qquad .$$

2.-Conmutatividad:

$$d(x,y) = d(y,x) \qquad .$$

3.-Desigualdad triangular:

$$d(x,z) < d(x,y) + d(y,z) \qquad .$$

Existen numerosas medidas de distancia que satisfacen estas condiciones. Una vez tomada la decisión acerca de qué medida de distancia entre elementos utilizar, se debe elegir el tipo de algoritmo que se utilizará para realizar el análisis de agrupamiento. Esto significa que se debe determinar una metodología de cálculo de la distancia entre grupos, propiamente dicha. Siguiendo a Spath (1980), las más comunes son:

-Distancia mínima (enlace simple): la distancia entre dos grupos se determina por la distancia entre dos vectores (o unidades de análisis) más cercanos pertenecientes a grupos distintos. En términos matemáticos, dados dos clústeres C_i, C_j,

$$\delta(C_i, C_j) = \text{mín}\{d(x,y), \quad x \quad \epsilon \quad C_i, \quad y \quad \epsilon \quad C_j\} \qquad .$$

-Distancia máxima (enlace completo): la distancia entre dos grupos se determina por la mayor distancia entre dos unidades de análisis perteneciente a grupos distintos. Método útil cuando las unidades de análisis forman claramente grupos definidos. Dados dos clústeres C_i, C_j,

$$\delta(C_i, C_j) = \text{máx}\{d(x, y), \quad x \quad \epsilon \quad C_i, \quad y \quad \epsilon \quad C_j\} \qquad .$$

-Método del centroide: calcula la proximidad entre dos conglomerados como la distancia entre sus centroides. Si $\bar{x_i} = \frac{1}{n_i} \sum_{x \quad \epsilon \quad C_i} x$; $\bar{y_i} = \frac{1}{n_j} \sum_{y \quad \epsilon \quad C_j} y$, entonces:

$$\delta(C_i, C_j) = d(\bar{x_i}, \bar{y_i}) \qquad .$$

-Distancia media entre dos grupos (enlace promedio ponderado): la distancia entre dos grupos se define como el promedio de las distancias entre todos los pares de elementos de ambos grupos usando el tamaño de los grupos como peso,

$$\delta(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \,\epsilon\, C_i; y \,\epsilon\, C_j} d(x, y) \qquad .$$

-Método de Ward (Ward, 1963; Pérez, 2005): para este método se considera la distancia euclidiana al cuadrado como medida de disimilitud. Sea $d(x_i, x_j)^2 = \|x_i - x_j\|^2$ la distancia entre los puntos x_i y x_j. La varianza total del conjunto de puntos se define como $I = \sum_i m_i \|x_i - x_j\|^2$, siendo G el centro de gravedad de los puntos dados con masa respectiva m_i. Sea G_q el centro de gravedad del q-ésimo conglomerado al dividirse el conjunto de individuos en que q grupos y m_q su respectiva masa. La varianza total o inercia se puede descomponer de la siguiente forma:

$$I = \sum_q m_q \|G_q - G\|^2 + \sum_q \sum_{i \,\epsilon\, q} m_i \|x_i - G\|^2 \qquad .$$

Sean x_i y x_j dos elementos de masa m_i y m_j respectivamente que se unen en un elemento x cuya masa es $m = m_i + m_j$, con $x = \frac{(m_i x_i + m_j x_j)}{(m_i + m_j)}$. Se descompone la varianza I_{ij} de x_i y x_j con respecto a G por la ecuación:

$$I_{ij} = m_i \|x_i - x\|^2 + m_j \|x_j - x\|^2 + m \|x - G\|^2 \qquad .$$

La reducción de la varianza se calcula reemplazando x por su valor como la función de x_i y x_j:

$$\triangle I_{ij} = \frac{(m_i m_j)}{(m_i + m_j)} \|x_i - x_j\|^2 \qquad .$$

Por lo tanto, el objetivo del método es encontrar los individuos x_i y x_j con la condición de que hagan mínima $\triangle I_{ij}$.

3.5. Normalización de los datos

Antes de ocupar algún tipo de integregación de indicadores, se recomienda normalizar los datos. La idea de normalizar los indicadores es que todos estén en la misma escala al momento de agregar la información. Algunos objetivos de las técnicas de normalización son:

- Ajustar para que los datos no tengan diferentes unidades de medida.
- Ajustar para que los datos no tengan diferentes rangos de variación.
- Ajustar en el caso que los datos sigan una distribución asimétrica o ante la presencia de datos atípicos.

Para estos casos existen diferentes tipos de normalización de los datos (Freundenberg, 2003; Jacobs, Smith y Goddar, 2004). La elección de una u otra metodología dependerá de las características de cada indicador y del juicio experto del analista. A continuación, se va a considerar la siguiente notación:

x_{qc}^t: puntuación del indicador simple q para la unidad de análisis c en el momento de tiempo t para $q = 1, \ldots, Q$ y $c = 1, \ldots, M$, donde Q es el número de indicadores simples y M el número de unidades de análisis/observaciones.

3.5.1. Ranking

$$I_{qc}^t = Ranking(x_{qc}^t) \qquad .$$

Esta es una de las técnicas de normalización más simples. Se puede usar tanto en datos cualitativos (ordinales) como con datos cuantitativos. En este método, los indicadores tienen las mismas unidades de medida y además no se ve afectado por valores atípicos. La desventaja que tiene este método es que se pierde información de las diferencias entre unidades de análisis cuando se agrega a nivel absoluto.

3.5.2. Estandarización (método *z-score*)

$$I_{qc}^t = \frac{x_{qc}^t - x_{qc=\bar{c}}^t}{\sigma_{qc=\bar{c}}^t} \qquad .$$

Donde $x_{qc=\bar{c}}^t$ representa la media del indicador q para todas las unidades de análisis en el momento t y $\sigma_{qc=\bar{c}}^t$ la desviación típica del indicador q para todas las unidades de análisis en el momento t.

Esta técnica se puede usar solo con indicadores cuantitativos. Transforma los indicadores a una escala adimensional con media 0 y desviación típica 1 manteniendo las distancias relativas, puesto que

se trata de una transformación lineal. También cabe destacar que los indicadores con valores extremos tendrán un mayor efecto sobre el índice compuesto. Esto puede que sea deseable si la intención es premiar el comportamiento excepcional de los indicadores, es decir, si se considera mejor resultados cuando el valor de un indicador es muy alto respecto a la media de las puntuaciones de todos los indicadores. Este efecto se puede corregir en la fase de agregación, bien excluyendo la mejor y la peor puntuación de los indicadores simples en el indicador compuesto o asignando ponderaciones diferentes basadas en la conveniencia de las puntuaciones de los indicadores simples.

También cabe mencionar que un método alternativo es la normalización geométrica y la transformación de Box-Cox (1964).

3.5.3. Re-escalamiento (método Mín- Máx)

$$I_{qc}^{t} = \frac{x_{qc}^{t} - \text{mín}_c(x_q^t)}{\text{máx}_c(x_q^t) - \text{mín}_c(x_q^t)} \qquad ;$$

donde $\text{máx}_c(x_q^t)$ y $\text{mín}_c(x_q^t)$ es el máximo y el mínimo del valor de x_q^t obtenido para todas las unidades de análisis en el momento t.

Este método se puede utilizar tanto en datos cuantitativos como cualitativos. Transforma los indicadores a una escala adimensional manteniendo las distancias relativas. Normaliza los indicadores para obtener un rango de variación entre 0 y 1. Además, los valores atípicos pueden distorsionar el indicador transformado. Esta normalización puede ampliar el rango de indicadores que están dentro de un mismo intervalo pequeño aumentando más el efecto sobre el indicador compuesto que la transformación *z-score*.

3.5.4. Distancia a una unidad de análisis referencial

$$I_{qc}^{t} = \frac{x_{qc}^{t}}{x_{qr}^{t_0}} \qquad ;$$

donde r es una unidad considerada como referencia y t_0 el año base de estudio.

Se usa solamente para valores cuantitativos. Mide la posición de un indicador dado un punto referencial. Este punto referencial podría ser un objeto a alcanzar en un marco temporal dado una unidad promedio. La unidad de análisis referencial también podría ser el líder del grupo en el que la principal unidad de análisis recibe una puntuación de y los otros se le asigna una puntuación porcentual según la distancia a la unidad de análisis líder. Este procedimiento, sin embargo, está basado en valores extremos que podrían ser valores atípicos no fiables.

3.5.5. Categorización de escalas

Ejemplo:

$$I_{qc}^{t} = \begin{cases} 0 & \text{sí} \;\; x_{qc}^{t} < p_{15} \\ 20 & \text{sí} \;\; p_{15} \leq x_{qc}^{t} < p_{25} \\ 40 & \text{sí} \;\; p_{25} \leq x_{qc}^{t} < p_{65} \\ 60 & \text{sí} \;\; p_{65} \leq x_{qc}^{t} < p_{85} \\ 80 & \text{sí} \;\; p_{85} \leq x_{qc}^{t} < p_{95} \\ 100 & \text{sí} \;\; x_{qc}^{t} \leq p_{95} \end{cases}$$

Este método se utiliza tanto en indicadores cualitativos (ordinales) como para indicadores cuantitativos. Ajusta el rango de variación de 0 a 100 para cada uno de los indicadores. También asigna una puntuación a cada indicador. La asignación puede ser categórica: uno, dos, tres estrellas, o cualitativa como "bueno", "regular" y "malo". Normalmente a cada categoría se le asigna un rango de valores basados en los percentiles de la distribución del indicador a lo largo de las unidades de análisis, p^i (i-ésimo percentil de la distribución del indicador) por ejemplo el top 5 % de las unidades de análisis recibe una puntuación de 100, las unidades entre el percentil 85 y 95 reciben 80 puntos, los valores entre el percentil 65 y 85 reciben 60 puntos y así sucesivamente. De esta forma se premia a las unidades de análisis con mejores resultados y se penaliza a aquellas con peores resultados.

3.5.6. Categorización de valores por encima y por debajo de la media

$$I_{qc}^{t} = \begin{cases} 1 & \text{sí} \;\; w > (1-p) \\ 0 & \text{sí} \;\; (1-p) \leq w \leq (1+p) \;\; donde \;\; w = \dfrac{x_{qc}^{t}}{x_{qc=\bar{c}}^{t_0}} \\ -1 & \text{sí} \;\; w < (1-p) \end{cases}$$

Este método no se ve afectado por valores atípicos y es solo para datos cuantitativos. Esta transformación asigna valores alrededor de la media valores 0, mientras que los valores que están por encima o pode debajo de un cierto umbral p se les asigna puntuación de 1 y -1, cabe destacar que la arbitrariedad del umbral p y la omisión de un nivel absoluto de información es criticable.

3.5.7. Método de normalización para indicadores cíclicos

$$I_{qc}^{t} = \frac{x_{qc}^{t} - E_t(x_q^t)}{E_t(\mid (x_q^t) - E_t(x_q^t) \mid)} \qquad ;$$

donde $E_t(x_q^t)$ indica la media de los valores de x_{qc}^t a lo largo de un cierto periodo de tiempo t.

Este método implícitamente, le asigna menos peso a las series más irregulares en el movimiento cíclico del índice compuesto.

3.5.8. Porcentaje de diferencias anuales en años consecutivos

$$I_{qc}^t = \frac{x_{qc}^t - x_{qc}^{t-1}}{x_{qc}^t} \quad .$$

Representa el porcentaje de crecimiento del año anterior. Esta transformación solo se usa cuando los indicadores están disponibles en diferentes años.

Para la construcción de indicadores compuesto la selección de métodos no es trivial. El método adecuado debe ir en concordancia con las propiedades de los índices. Es por esto que, las técnicas de robustez y sensibilidad son necesarios para evaluar el impacto en los resultados.

3.6. Ponderación de la información

Esta etapa es crucial para la construcción del indicador compuesto, ya que, consiste en asignar pesos a los indicadores simples, para sucesivamente integrarlos y formar un único valor que represente nuestro índice global.

Esta asignación puede hacerse de forma equitativa o bien estableciendo diferentes pesos a cada indicador, representando la importancia, la significancia y fiabilidad de este en el índice final.

El método elegido en esta etapa tiene un gran impacto sobre el índice final, por lo que se requiere que sea lo más explícito, transparente y justificado posible. Cabe destacar que no existe un método de ponderación objetiva, común y única para la construcción de los indicadores compuestos, por lo cual los pesos deben seleccionarse de acuerdo con el marco conceptual que define el concepto que se está midiendo.

3.6.1. Métodos de ponderación equitativa

Asignación de Pesos Iguales

Este es el método de ponderación más sencillo, el cual, implica que cada índice simple tiene el mismo peso o la misma importancia sobre el índice compuesto. Con este método se corre el riesgo de

que combinando indicadores con un elevado grado de correlación ciertos aspectos tengan un elevado peso en el índice compuesto.

3.6.2. Métodos de ponderación basados en modelos estadísticos

Correlación simple

Un criterio para seleccionar los pesos de cada uno de los indicadores se basa en el análisis de la correlación entre cada uno de éstos y una variable que registra la evolución del conjunto que queremos medir, por ejemplo, el indicador mensual de actividad económica (IMACEC) o el producto interno bruto (PIB), ya sea mensual, trimestral o anual. Si r_i es el coeficiente de correlación entre el indicador i y la variable de referencia, el peso de cada uno de los indicadores en la definición del índice compuesto será:

$$w_i = \frac{r_i}{\sum_{i=1}^{n} r_i} \quad ,$$

en que n será el número de indicadores utilizados.

Análisis factorial y componentes principales

Este método puede ser útil cuando existe alta correlación entre los indicadores simples, sin embargo, se debe tener en cuenta que la agrupación de variables se basa en propiedades estadísticas y no en el plano interpretativo del concepto que se desea medir.

Una forma de obtener los pesos mediante esta técnica de reducción de dimensión fue propuesta por Nicoletti, Scarpetta y Boylaud (2000). Este enfoque consiste en agrupar los indicadores simples con las mayores saturaciones en índices compuestos intermedios. Los pasos son los siguientes:

1.-Obtención de las variables latentes, de acuerdo a las diferentes reglas de extracción.
2.-Cálculo de la varianza de todos los indicadores explicada por cada factor. Para obtener la varianza explicada se calcula la suma de los cuadrados de las cargas factoriales de los indicadores.
3.-Normalización de los cuadrados de las cargas factoriales. Para ello se calcula el cuadrado de la carga factorial dividida por la varianza de los indicadores explicada por el factor.
4.-Obtención de los pesos de cada factor mediante el cociente entre la varianza de los indicadores explicada por cada factor y la varianza del total de los factores retenidos.
5.-Cálculo del peso de cada indicador ponderando la máxima variabilidad de este explicada por un factor por el peso del factor. Los pesos obtenidos se normalizan dividiendo por la suma de todos ellos de forma que sume la unidad.

Se debe tener en cuenta que la correlación no necesariamente implica redundancia. Si no existe correlación significativa entre los indicadores simples no se sugiere que los pesos se estimen con esta técnica.

Método de regresión lineal

Los modelos de regresión lineal pueden proporcionar información valiosa sobre el vínculo existente entre un conjunto de variables independientes $I_{1c}, I_{2c}, \ldots, I_{Qc}$ y una variable dependiente $\hat{Y}_c$. Supóngase que las variables independientes son indicadores simples que se han definido para la construcción del indicador compuesto y la variable dependiente representa el objetivo global que debe alcanzar cada unidad de análisis. En muchos estudios sobre índices compuestos la variable dependiente es otro índice compuestos muy relacionado con el que se quiere comparar (Saisana, 2008):

$$\hat{Y}_c = \hat{\alpha} + \hat{\beta}_1 I_{1c} + ... + \hat{\beta}_Q I_Q c \quad con \;\; C = 1, ..., M \qquad ;$$

donde $\hat{\alpha}$ es la constante estimada, M es el número de observaciones y $\hat{\beta}_1, ..., \hat{\beta}_Q$ son los coeficientes de regresión asociados a los indicadores simples $I_{1c}, I_{2c}, \ldots, I_{Qc}$ que pueden ser considerados, una vez estandarizados, como factores de ponderación.

Se trata de un procedimiento que, aunque es adecuado para un número elevado de variables de diferentes tipos, implica la suposición de que los indicadores simples tienen un comportamiento lineal en relación con el objetivo planteado y que deben ser independientes entre sí, puesto que si existe multicolinealidad el análisis se torna deficiente. No obstante, el uso de modelos de regresión lineal puede ser útil para cuantificar el efecto relativo de cada objetivo de política, representado por cada variable, y los objetivos globales a ser alcanzados, así como para validar un conjunto de factores de ponderación calculados a partir de otra técnica.

3.6.3. Métodos de ponderación basados en modelos participativos

De forma alternativa se nombraran algunas otras técnicas de ponderación basadas en modelos participativos, en la que se pide la opinión de expertos en el tema en cuestión, sobre la importancia que debe tener cada indicador simple en el índice global.

Método de asignación presupuestaria

Esta técnica de ponderación consiste en repartir una cantidad de 100 puntos en cada indicador simple, con el fin de asígnarle más puntos a un índice más importante y menos puntos a alguno con menos importancia. Todo esto se hace en consenso con todos los expertos del tema. En el caso de no llegar a consenso este método no sería adecuado.

Este método es óptimo para un número máximo de 12 indicadores simples. Si se dispone de más indicadores, este método puede producir confusión y estrés a los expertos (Nardo *et al.*, 2005).

Opinión pública

Este método consiste en ponderar los índices en base a las urgencias políticas y no tanto a la importancia relativa que tiene cada índice. En este caso se realiza una encuesta para que el público participe en la asignación de los pesos de cada uno de los indicadores simples.

Análisis conjunto

Esta es una técnica estadística utilizada en ciencias sociales aplicadas, generalmente en el marketing, la administración del producto y la investigación operativa, que trata de entender como los encuestados desarrollan preferencias acerca de productos o servicios (Hair *et al.*, 2007). Se trata de una metodología de carácter participativo que se basa en la evaluación por parte de expertos.

El análisis conjunto es el único entre los métodos multivariantes en el que el analista deber reformular el problema identificado los atributos a considerar. En el caso de los índices compuestos, los atributos son los indicadores y sus posibles niveles.

3.7. Agregación de la información

Una de las partes más polémicas en la construcción de índices compuestos es la agregación de la información. Muchos autores están en contra de agregar toda la información disponible obtenida mediante las etapas anteriores en un único valor que representa la unidad de análisis, ya que puede perderse mucha información. Por otro lado, es muy difícil interpretar e identificar diferencias entre las unidades de análisis si no se realiza una agregación de los indicadores simples.

A continuación, se verán diferentes técnicas de agregación de los indicadores simples para la obtención del índice compuesto, aunque en la practica la que más se suele utilizar es la agregación lineal ponderada (véase Nardo *et al.*, 2008).

3.7.1. Métodos aditivos de agregación lineal

Suma de rankings

Es el método más simple de agregación consiste en sumar, para cada unidad de análisis c con $c = 1, \ldots, M$ el orden o ranking que posee cada uno de los indicadores simples en relación con el resto de las unidades de análisis:

$$IC_c = \sum_{q=1}^{Q} Ranking_{qc} \quad , con \qquad C = 1, ..., M \qquad .$$

Se trata de un método que no se ve afectado por valores atípicos. La principal desventaja al utilizar este método es que se pierde mucha información al calcular el ranking en los indicadores simples.

Agregación lineal ponderada

Se trata de la agregación lineal más utilizado en la construcción de indicadores compuestos:

$$IC_c = \sum_{q=1}^{Q} w_q I_{qc} \qquad ;$$

con $\sum_{q=1}^{Q} w_q = 1$ y $0 \leq w_q \leq 1$ y I_{qc} el valor normalizado de la unidad de análisis c respecto al indicador q, para $q = 1, \ldots, Q$ y $c = 1, \ldots, M$. La obtención de los pesos w_q debe quedar clara en la etapa anterior.

Agregación Geométrica

Un comportamiento no deseado de las técnicas de agregación lineal descritas anteriormente es la compensación total entre indicadores, de tal forma que si existe un rendimiento bajo en algunos indicadores este comportamiento se compensa por altos valores en el resto de indicadores. Los empleados en las técnicas de agregación compensatoria se tratan como factores de escala o *trade offs* (Munda, 2008).

La agregación geométrica es una solución intermedia entre la compensación total y la no compensación entre indicadores proporcionada por las técnicas de agregación multi-criterio no compensatorias.

Este tipo de agregación es similar a la agregación lineal ponderada, pero considerando la media geométrica:

$$IC_c = \prod_{q=1}^{Q} (I_q c)^w q \qquad ;$$

con $\sum_{q=1}^{Q} w_q = 1$ y $0 \leq w_q \leq 1$ y I_{qc} el valor normalizado de la unidad de análisis c respecto al indicador q, para $q = 1, \ldots, Q$ y $c = 1, \ldots, M$. La obtención de los pesos w_q debe quedar clara en la etapa anterior.

3.8. Análisis de robustez y sensibilidad

La construcción de un índice compuesto requiere de etapas subjetivas en las que se emiten juicios de valor para la elección de los diferentes supuestos. Ejemplos de estas etapas son: la elección de los indicadores simples, la imputación de datos faltantes, la asignación de pesos para cada indicador y dimensión, la elección del método de agregación, entre otros. Estas elecciones subjetivas forman la base de la construcción del índice compuesto, por lo tanto, para incrementar la transparencia y evaluar la calidad de su diseño es imprescindible aplicar un análisis de sensibilidad y de incertidumbre.

En general, las fuentes de incertidumbre asociados al diseño de índices compuestos dependen de muchos factores, entre los que destacan (Saltelli *et al.*, 2008):

- La inclusión o exclusión de indicadores simples en la construcción del índice global.
- Tratamiento previo de los datos (detección de valores atípicos, detección de distribuciones asimétrica en los indicadores, entre otros).
- Elección del método apropiado de imputación de datos faltantes.
- Elección de pesos asignados a cada indicador simple y, si se da el caso, a las dimensiones.
- Elección del método de agregación de los indicadores simples y, si se da el caso, de las dimensiones.

Todas las premisas o supuestos descritos anteriormente pueden influenciar en la puntuación final de cada unidad de análisis del estudio, por lo que deben tener en cuenta en la construcción de índices compuestos, ya que diferentes metodologías pueden originar diferentes índices compuestos. No obstante, se debe elegir aquel índice que sea válido y que cumpla que pequeños cambios en su arquitectura den lugar a pequeñas variaciones en la puntuación final.

La construcción de un índice compuesto puede identificarse con el desarrollo de un modelo. Supóngase que se asume un modelo representado por una función de la siguiente forma:

$$\begin{aligned} Y &= f(X_1, X_2, ..., X_n) \\ &= f(X) \qquad . \end{aligned}$$

donde Y es la variable *output* del modelo y $X_1, X_2, \ldots, X_n$ son los factores *imput* del modelo.

El análisis de sensibilidad trata de estudiar como la variación en la salida del modelo puede distribuirse cuantitativamente a diferentes fuentes de variación en los supuestos del modelo (Nardo *et al.*, 2008).

El análisis de sensibilidad está ligado al análisis de incertidumbre. Este último trata de cuantificar la incertidumbre del *output* provocada por diferentes fuentes de incertidumbre de los factores *input*, mientras que el análisis de sensibilidad trata de identificar de donde proviene dicha incertidumbre, es

decir, cuáles son los factores *input* que más afectan a la variación del *output*. Una combinación del análisis de incertidumbre y sensibilidad ayuda a evaluar la robustez y calidad del índice compuesto y a aumentar la transparencia de su construcción.

No obstante, se debe tener cuidado y ser objetivo en la decisión sobre el grado de incertidumbre que se debe asignar a los factores *input* para comprobar la variación del factor *output*. Si la variación asignada a los factores *input* es grande, el modelo predictivo tendrá tanta variabilidad que el análisis no será útil.

En la construcción de índices compuestos los factores *input* suelen identificarse con las posibles metodologías que se pueden aplicar en cada etapa de su construcción, mientras que el factor *output* se identifica con el objetivo del estudio, es decir, con el índice compuesto construido para cada una de las unidades de análisis.

Por lo tanto, el objetivo de esta etapa es realizar un análisis de incertidumbre con el fin de cuantificar la incertidumbre del índice compuesto provocada por las diferentes fuentes de variación del espacio de supuestos e identificar mediante un análisis de sensibilidad los supuestos que más afectan a la incertidumbre del índice global.

A continuación, se describen los procedimientos de análisis de incertidumbre y sensibilidad que suelen aplicar a los índices compuestos para el estudio de su robustez y calidad (Saisana, Saltelli y Tarantola, 2005; Saltelli *et at.*, 2008). El ejemplo que se expone para el desarrollo de ambos análisis es ilustrativo, puesto que en la práctica la elección del espacio de supuestos y de las fuentes de incertidumbre depende del objetivo del problema planteado.

Supóngase que el modelo planteado se considera tres fuentes de incertidumbre:

1.− Técnica de normalización.

2.− Técnica de ponderación.

3.− Técnica de agregación.

Sea IC_j el valor del índice para las unidades de análisis j con $j = 1, \ldots, M$,

$$IC_j = f_{rs}(I_{1,j}, I_{2,j}, ..., I_{Q,j}, w_{s,1}, w_{s,2}, ..., w_{s,Q}) \qquad .$$

Calculado con el método de agregación $r = 1, 2, 3, \ldots$ y el método de ponderación $s = 1, 2, 3, \ldots$, con un número finito de posibilidades para r y s, y donde los números se identifican con distintas técnicas de agregación y ponderación. Los $I_{Q,j}$ representan los indicadores simples utilizados para la construcción del índice una vez normalizados mediante una técnica de normalización y los $w_{s,Q}$ los pesos de los indicadores simples calculados según modelo de ponderación s. Puede ocurrir que alguna

de las combinaciones de los distintos métodos no se puedan aplicar.

Un procedimiento utilizado para cuantificar el output del modelo es calcular la posición que ocupa cada unidad de análisis en relación al resto de las unidades, $Ranking(IC_j)$, y estimar la discrepancia respecto a una metodología de referencia, $Ranking_{referencial}(IC_j)$:

$$\bar{R}_s = \frac{1}{M}\sum_{j=1}^{M} |Ranking_{referencial}(IC_j) - Ranking(IC_j)| \qquad .$$

En este caso, diferentes supuestos en la construcción del índice compuesto pueden producir una variación en las variables Ranking IC_j y $\bar{R}_s$. Estas dos variables se identifican como variables output del modelo y constituyen las medidas de interés en el análisis de sensibilidad e incertidumbre. Se pueden utilizar medidas de interés según convenga.

3.8.1. Análisis de sensibilidad global basado en cálculo de varianzas

El análisis de sensibilidad de un modelo es un procedimiento relevante, ya que permite identificar los factores *input* que más afectan a la incertidumbre del *output*. Existen diversos métodos de análisis de sensibilidad clasificados en dos grandes grupos: análisis de sensibilidad local y análisis de sensibilidad global.

El análisis de sensibilidad local está formado por métodos locales basados en el cálculo de derivadas parciales de los factores *output* con respecto a los factores *input*. Se trata de métodos simples de evaluar y con bajo coste computacional. La principal limitación de este tipo de análisis es que proporciona información únicamente en el punto base donde las derivadas son calculadas, sin explorar todo el intervalo de variación de los factores *input*. Además, cuando el modelo contiene discontinuidades de las derivadas parciales no pueden ser calculadas. En el análisis de sensibilidad global (GSA, por sus siglas en ingles), que supera las limitaciones del análisis de sensibilidad local, se han desarrollado métodos como los coeficientes de regresión estandarizados, la prueba de efectos elementales, el filtrado de Monte Carlo, los métodos basados en el cálculo de varianzas, u otros (véase Saltelli, Chan y Scott (2000), Saltelli *et al.* (2004) y Saltelli *et al.* (2008)).

Los métodos de análisis de sensibilidad global más usados en la literatura se basan en el cálculo de varianzas, puesto que presentan las siguientes ventajas (Nardo et al. (2008); Saltelli *et al.* (2008)):

- Se pueden aplicar independientemente de la forma funcional del modelo.
- Permiten distinguir, mediante el uso de coeficientes de sensibilidad, los principales factores que afectan a la sensibilidad del índice compuesto.
- Permiten distinguir los efectos principales de primer orden de los efectos de orden superior de los factores *input*.

– Permiten captar la influencia de todo rango de variación de cada factor *input*.

– Permiten tratar a los factores *input* de forma agrupada.

– Son fáciles de interpretar.

A continuación, se describe el procedimiento para la obtención de medidas de sensibilidad de los factores input del modelo a partir del cálculo de varianzas (Saltelli *et al.*, 2008).

Supónganse el siguiente modelo genérico:

$$Y = f(X_1, X_2, ..., X_n) \qquad .$$

Sean X_i, $i = 1, \ldots, n$, los factores *input* del modelo con un rango de variación o de incertidumbre no nulo e Y el factor que puede ser $Ranking(IC_j)$, $\bar{R}_s$ o cualquier otra medida de interés.

Supóngase que se fija un factor X_i en un punto determinado x_i^*. Sea $V_{X_{\sim i}}(Y|X_i = x_i^*)$ la varianza resultante de Y sobre $X_{\sim i}$ (la notación " $\sim i$" significa que en X se están considerando todos los factores menos el factor o indicador X_i. Este término se llama varianza condicionada, puesto que se calcula al fijar X_i en el valor x_i^*. Cabe esperar que al fijar una fuente de variación X_i, la varianza resultante $V_{X_{\sim i}}(Y|X_i = x_i^*)$ sea inferior que la varianza total incondicionada $V(Y)$. Por lo tanto, se puede utilizar $V_{X_{\sim i}}(Y|X_i = x_i^*)$ como una medida de importancia relativa de X_i sobre Y, de tal forma que cuanto menor sea $V_{X_{\sim i}}(Y|X_i = x_i^*)$, mayor será la influencia de X_i sobre Y.

Sin embargo, se han detectado dos limitaciones con este procedimiento. En primer lugar esta medida de importancia depende de la posición del punto x_i^* para cada factor *input*. En segundo lugar, se puede definir un modelo tal que para ciertos X_i y puntos fijos x_i^* ocurre que $V_{X_{\sim i}}(Y|X_i = x_i^*) > V(Y)$. Para resolver estas limitaciones se calcula el promedio de la medida definida sobre todos los posibles valores de x_i^* , de tal forma que la dependencia con x_i^* desaparece. El promedio $E_{X_i}(V_{X_{\sim i}}(Y|X_i))$ siempre es inferior a $V(Y)$ puesto que:

$$V_{X_{\sim i}}(E_{X_i}(Y|X_i)) + E_{X_i}(V_{X_{\sim i}}(Y|X_i)) = V(Y) \qquad ;$$

en que al primer término se le llama efecto principal de la variable X_i sobre el *output* Y, y el segundo término es el residuo.

Un valor para el residuo $E_{X_i}(V_{X_{\sim i}}(Y|X_i))$ o un valor grande para el efecto principal $V_{X_{\sim i}}(E_{X_i}(Y|X_i))$ significa que X_i es un factor importante sobre Y. También se cumple por la ecuación anterior que $V_{X_{\sim i}}(E_{X_i}(Y|X_i)) \leq V(Y)$. La varianza condicionada $V_i = V_{X_{\sim i}}(E_{X_i}(Y|X_i))$ es un valor comprendido entre cero (cuando X_i no contribuye en la formación de Y) y la varianza no condicional de Y, $V(Y)$ (cuando el resto de los factores no influyen en la creación de Y).

A partir de la varianza condicional, se define el coeficiente de sensibilidad S_i conocido como el coeficiente de sensibilidad de primer orden de X_i sobre Y[2] (Saltelli *et al.*, 2008):

$$S_i = \eta^2 = \frac{V_{X_i}(E_{X_{\sim i}}(Y|X_i))}{V(Y)} \qquad . \tag{3.1}$$

Los coeficientes S_i cuantifican la importancia de un factor de entrada X_i sobre el factor de salida Y.

La esperanza condicional de $E_{X_{\sim i}}(Y|X_i)$ del numerador de la expresión (3.1) puede ser una función lineal o no en X_i. En el caso particular en que $E_{X_{\sim i}}(Y|X_i)$ es una función lineal en X_i:

$$E_{X_{\sim i}}(Y|X_i) = a_i + b_i X_i$$

Mediante el procedimiento de mínimos cuadrados se estima $b_i = \dfrac{cov(Y, X_i)}{V(X_i)}$.

Sea $\hat{\beta}_i = b_i(\sqrt{V(X_i)}/\sqrt{V(Y)})$ el coeficiente de regresión estandarizado de b_i. Por lo tanto, el coeficiente de sensibilidad de primer orden de X_i sobre Y coincide con el coeficiente de regresión estandarizado al cuadrado de la regresión lineal de Y en X_i:

$$\begin{aligned} S_i &= \frac{V_{X_i}(E_{X_{\sim i}}(Y|X_i))}{V(Y)} \\ &= \frac{b_i^2 V(X_i)}{V(Y)} \\ &= \frac{cov^2(Y, X_i)}{V(X_i)V(Y)} \\ &= \hat{\beta}^2 \qquad . \end{aligned}$$

En la práctica, el modelo de $E_{X_{\sim i}}(Y|X_i)$ no suele ser una función lineal en X_i, por lo tanto, el coeficiente de regresión estandarizado no recoge toda la reducción en varianza del indicador compuesto al fijar cada uno de los factores *input*. Es por eso que para el cálculo de los coeficientes S_i resulta conveniente el uso de un modelo GSA basado en el cálculo de varianzas aplicable a cualquier modelo independiente de su forma funcional.

Sacks *et al.* (1989) y Sobol (1993) definieron los coeficientes de análisis de sensibilidad global S_i a partir de la descomposición de f en un conjunto de funciones de creciente dimensionalidad, conocida como *High Dimensional Model Representation* (HDMR). A continuación, se describe su procedimiento analítico:

Sea $Y = f(X) = f(X_1, X_2, \ldots, X_n)$ una función definida en un compacto unitario $K^n \equiv [0, 1]^n$ donde X_i son los factores *input* con dominio de variabilidad U y Y es el factor *output*.

La función $f(X)$ se puede descomponer mediante la representación HDMR como sigue (Rabitz *et al.*, 1999):

$$\begin{aligned} f(X) &= f(X_1, X_2, \ldots, X_n) \\ &= f_0 + \sum_i f_i(X_i) + \sum_{i<j} f_{i,j}(X_i, X_j) + ... + f_{1,2,...,n}(X_1, X_2, ..., X_n) \end{aligned} \tag{3.2}$$

donde cada término de la descomposición es la función de los factores *input*, es decir, $f_i(X_i)$ denota el efecto del factor *input* X_i sobre f cuando actúa de forma independiente al resto de factores *input*, la función $f_{i,j}(X_i, X_j)$ describe el efecto de interacción de los *inputs* X_i, X_j sobre f y el resto de los términos de mayor grado describen el efecto conjunto de los factores *input* que actúan en el término correspondiente sobre Y. Finalmente, el último término $f_{1,2,...,n}(X_1, X_2, \ldots, X_n)$ indica la independencia residual de todos los factores input fijados de forma que tienen un efecto conjunto sobre el modelo f.

Si las interacciones entre los *input* no tienen efecto sobre el modelo, la descomposición solo se define mediante el término de orden cero f_0 y los términos de primer orden $f_i(X_i)$ que son funciones de los factores *input* X_i. En este caso, el modelo se llama aditivo.

El objetivo, en principio, es buscar descomposiciones que den una buena aproximación a la función $f(*)$ en la norma L_2:

$$\zeta \equiv \int_{K^n} \left[f(X) - f_0 - \sum_i f_i(X_i) - \sum_{i<j} f_{i,j}(X_i, X_j) - ... - \sum_{i_1<...<i_s} f_{i_1,...,i_s}(X_{i_1}, ..., X_{i_s}) \right]^2 \partial X \qquad ,$$

con $s \leq I$.

Sin embargo, la forma de las funciones de la descomposición (3.2) no es única.

Cuando los factores *input* del modelo son independientes y se cumple la siguiente condición de partida:

$$\int_0^1 f_{i_1,...,i_s}(X_{i_1}, ..., X_{i_s}) \partial X_j = 0, \quad j \,\epsilon\, \{i_1, i_2, ..., i_s\},\ s \,\epsilon\, \{1, ..., I\}$$

Entonces la descomposición HDMR cumple las siguientes propiedades:

1) $\int_{K^n} f(X) \partial X = E(Y) = f_0 \rightarrow f_0$ =Valor medio de la función $f(X)$.
2) Todo par de términos de la descomposición HDMR son ortogonales:
   $\int_{K^n} f_{i_1,...,i_s}(X_{i_1}, ..., X_{i_s}) f_{j_1,...,j_s}(X_{j_1}, ..., X_{j_s}) \partial X = 0$ con $\{i_1, i_2, ..., i_s\} \neq \{j_1, j_2, ..., j_s\}$.

3) La descomposición HDMR es única.

En este caso, la descomposición HDMR de la función $f(*)$ viene dada de la siguiente forma (Rabitz *et al.*, 1999):

$$\begin{aligned} f_0(X_0) &= \int_{K^n} f(X)\partial X \\ &= E(Y) \qquad , \end{aligned}$$

$$\begin{aligned} f_i(X_i) &= \int_{K^{n-1}} f(X) \prod_{j \neq i} \partial X_j - f_0 \\ &= E_{X_{-i}}(Y|X_i) - f_0 \qquad , \end{aligned}$$

$$f_{i,j}(X_i, X_j) = \int_{K^{n-2}} f(X) \prod_{k \notin \{j,i\}} \partial X_k - f_i(X_i) - f_j(X_j) - f_0 \qquad ,$$

$$= E_{X_{-ij}}(Y|X_i, X_j) - E_{X_{-i}}(Y|X_i) - E_{X_{-j}}(Y|X_j) - f_0 \qquad ,$$

$$\vdots$$

$$f_{i_1,...,i_r}(X_{i_1}, ..., X_{i_r}) = \int_{K^{n-r}} f(X) \prod_{k \notin \{j,i\}} \partial X_k - \sum_{j_1 < ... < j_{r-1} \subset \{i_1, i_2, ..., i_r\}} f_{j_1,...j_{r-1}}(X_{j_1}, ..., X_{j_{r-1}})$$

A esta descomposición única se le conoce con el nombre de ANOVA-HDMR (Archer y Satelli, 1997; Rabitz *et al.*, 1999), puesto que es la definición de la descomposición ANOVA (análisis de varianza) del factor *output*. Se trata de la descomposición que mejor aproxima en norma L_2 a la función $f(*)$.

La aplicación de la descomposición ANOVA-HDMR es muy útil para medir el efecto de la varianza de cada uno de los factores *input* de la función $f(*)$ en la varianza total del factor *output* Y.

Considerando que los factores *input* son independientes, el esquema de descomposición de la varianza incondicional del *output* Y, equivalente a la descomposición *HDMR*, se define como sigue (Sobol, 1993):

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{i,j} + ... + V_{1,2,...,n} \qquad , \tag{3.3}$$

donde:

$$V_i = V_{X_i}[E_{X_{\sim i}}(Y \backslash X_i)],$$

$$V_{i,j} = V_{X_i X_j}[E_{X_{\sim ij}}(Y \backslash X_i, X_j)] - V_i - V_j,$$

$$V_{i,j,l} = V_{X_i X_j X_l}[E_{X_{\sim ijl}}(Y \backslash X_i, X_j, X_l)] - V_{i,l} - V_{i,l} - V_{j,l} - V_i - V_j - V_l \qquad .$$

El término $V_{X_i X_j}[E_{X_{\sim ij}}(Y \backslash X_i, X_j)]$ mide el efecto conjunto del par de factores *input* (X_i, X_j) sobre el *output* Y. de forma análoga, se pueden calcular los términos de interacción de órdenes superiores. La notación $E_{X_{\sim i}}$ en $V_{X_i}[E_{X_{\sim i}}(Y \backslash X_i)]$ indica la matriz de todos los factores *input* exceptuando X_i. El operador $E_{X_{\sim i}}$ denota la esperanza del factor *output* Y sobre todos los posibles valores de $X_{\sim i}$ manteniendo fijo X_i. La varianza se calcula sobre todos los posibles valores de X_i.

Cuckier *et al.* (1973) y Sobol (1993) estiman de forma analítica los términos de esta descomposición para el cálculo de los coeficientes de sensibilidad. Normalizado por $V(Y)$ la ecuación (3.3) se obtiene la siguiente identidad:

$$1 = \sum_I S_i + \sum_i \sum_{j>i} \sum_{i,j} S_{i,j} + ... + S_{1,2,...,n} \tag{3.4}$$

Si además de cumplirse la independencia entre los factores *input* no existen interacciones entre ellos en el modelo, se cumple $\sum_{i=1}^{n} V_i = V(Y)$ y, por tanto $\sum_{i=1}^{n} S_i = 1$.

Si el modelo no es aditivo, se calculan el resto de coeficientes de orden superior. Los coeficientes que miden el efecto total de variación del *output* Y debido a un factor *input* X_i, es decir, debido a su coeficiente de primer orden y a los efectos de mayor orden provocados por las interacciones, se llaman coeficientes de sensibilidad total. Por ejemplo, supóngase que se dispone de $n = 3$ factores independientes, los tres coeficientes de sensibilidad total se calculan como sigue:

$$S_{T1} = S_1 + S_{1,2} + S_{1,3} + S_{1,2,3} \qquad , \tag{3.5}$$

donde $S_1 + S_{1,2} + S_{1,3} + S_{1,2,3}$ es la suma de todos los términos de la ecuación (3.4) para este caso.

De forma análoga,

$$S_{T2} = S_2 + S_{1,2} + S_{1,3} + S_{1,2,3}$$

$$S_{T3} = S_3 + S_{1,3} + S_{2,3} + S_{1,2,3}$$

Sin embargo, no se suelen estimar los coeficientes de orden superior calculando todos los términos de la fórmula (3.4) puesto que el modelo con n factores el número total de coeficientes a estimar es $2^n - 1$ (incluidos los de primer orden S_i).

Por lo tanto, se define el coeficiente de sensibilidad total del factor X_i como sigue (Saltelli *et al.*, 2008):

$$\begin{aligned} S_{T_i} &= \frac{E_{X_{\sim i}}(V_{X_i}(Y|X_{\sim i}))}{V(Y)} \\ &= 1 - \frac{V_{X_{\sim i}}(E_{X_i}(Y|X_{\sim i}))}{V(Y)} \end{aligned}$$

Para n factor dado X_i, toda diferencia significativa entre S_i y S_{T_i} indica que la interacción entre factores *input* es importante para el factor *output* Y. El cálculo de S_{T_i} se puede aplicar a cualquier modelo independiente del grado de correlación entre los factores *input*, sin embargo, la ecuación (3.5) solo es válida cuando existe independencia de factores *inputs*.

Para un modelo no aditivo en el que existen interacciones entre los factores *input*, los coeficientes de sensibilidad de primer orden, o efectos principales, no recogen la variabilidad del *output* del modelo Y y, por lo tanto $\sum_{i=1}^{n} S_i \leq 1$.

Los coeficientes de sensibilidad definidos mediante el cálculo de varianzas satisfacen el objetivo del GSA que es identificar cuáles son los factores *input* con mayor efecto sobre la construcción del factor *output* Y.

3.8.2. Análisis de incertidumbre

En primer lugar, para poder aplicar un análisis de incertidumbre se deben identificar las fuentes de incertidumbre en los *inputs* y simular diferentes escenarios según estas. Para realizar el análisis de incertidumbre en el campo de los indicadores compuestos se sugiere la aplicación de la técnica Monte Carlo (Nardo *et al*, 2008) que consiste en perturbar todas las fuentes de incertidumbre y analizar los efectos derivados de dichas variaciones sobre el modelo, de tal forma que sea posible estimar una función de distribución para $Ranking(IC_j)$ y/o $\bar{R}_s$. La técnica de Monte Carlo se basa en los siguientes pasos:

1) Sin pérdida de generalidad, supóngase que se plantean las siguientes tres fuentes de variabilidad (*input*) con el número finito de alternativas a considerar:

a) X_1: técnica de normalización de datos.

b) X_2: técnica de ponderación de los indicadores simples.

c) X_3: técnica de agregación.

En el primer lugar, se considera una función de densidad de probabilidad para cada factor *input* según sus posibles alternativas. Supóngase que el factor X_i, pudiendo ser i una de las tres fuentes de variabilidad anteriores, se compone de m alternativas a ser elegidas, por lo tanto $X_i \sim U(0,1)$. Sea $\alpha \epsilon [0,1]$ el número aleatorio del que se partirá el algoritmo. Por lo tanto, se seleccionan una de las n opciones según α quede incluido dentro de los intervalos siguientes:

2) Se genera un número N de muestras compuestas por las combinaciones de factores X^k, $1 \leq k \leq N$, con $X^k = X_1^k, X_2^k, X_3^k$ siendo X_i^k la opción obtenida para X_i en la simulación k. de tal forma, se obtiene cada muestra el valor del indicador compuesto y se calcula el escalar Y^k que puede ser cualquiera de las variables *output* del análisis de incertidumbre, $Ranking(IC_j)$ y/o $\bar{R}_s$.

3) Se calcula Y^k, $1 \leq k \leq N$ para todas las combinaciones. La secuencia de Y^k, $1 \leq k \leq N$ proporciona la distribución de probabilidad estimada de la variable *output*. Las características de esta distribución, tales como la media, la varianza y momentos de orden superior se pueden estimar con un nivel arbitrario de precisión relacionado con el tamaño de la simulación N. Estos valores son los que se analizan para constatar el grado de incertidumbre del índice compuesto frente a los cambios considerados.

La generación de muestras puede realizarse empleando procedimientos tales como el muestreo aleatorio simple, el muestreo cuasialeatorio, el muestreo estratificado u otro que se considere conveniente (Saltelli *et al.*, 2008).

Capítulo 4

Aplicación

Este capítulo mostrará los resultados de la construcción del índice compuesto, cabe destacar que se analizaron un total de treinta y tres índices, obtenidos de diferentes fuentes, entre ellas el Instituto Nacional de Estadísticas (INE), Banco Central, Adimark y otros. Estos índices en su mayoría son de frecuencia mensual y se encuentran disponibles periódicamente en la red.

El marco conceptual del índice compuesto, es medir la economía del país. Es por esto que un indicador referencial de nuestro índice compuesto es el Indicador Mensual de Actividad Económica (IMACEC), este índice es un indicador coyuntural proporcionado por el Banco Central, alimentado de índices de diferentes áreas como comercio, construcción, minería, transporte, servicios y otros.

Para seleccionar los índices que van a formar parte del indicador compuesto, se hizo un análisis multivariado de conglomerados como se muestra en la figura a continuación:

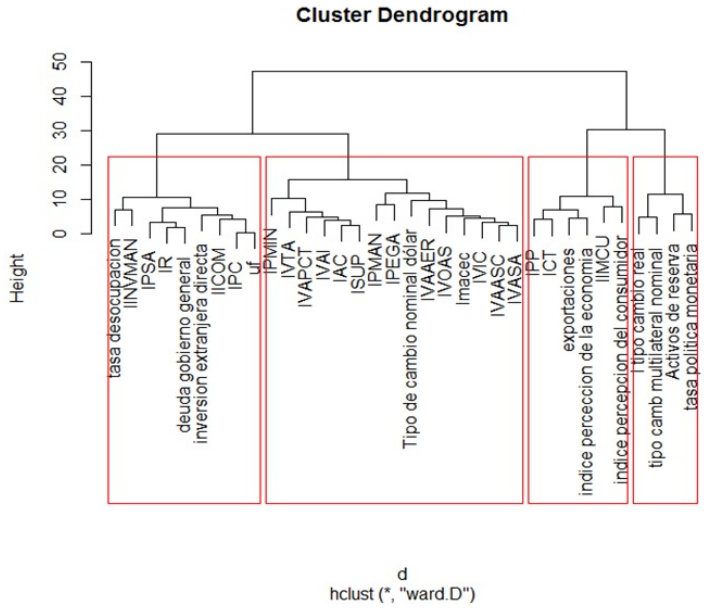

Figura 4.1: Análisis de conglomerados indices observados

Como se aprecia en la figura 4.1, el análisis de conglomerados formó 4 grupos, donde el segundo grupo se puede ver contenido al IMACEC. Cabe mencionar que la literatura dice, las variables seleccionadas deben tener una alta correlación con el índice de referencia, en este estudio, con repecto al IMACEC. A continuación, se detalla el análisis de correlación IMACEC vs indicadores:

Índice	Correlación
Tasa desocupación	0,11
Exportaciones	0,40
IAC	0,90
IPMAN	0,60
IINVMAN	0,38
IPMIN	0,44
IIMCU	-0,23
IPEGA	0,45
ISUP	0,83
IICOM	0,56
IPC	0,58
IR	0,61
IPP	0,28
ICT	0,45
Índice tipo cambio real	-0,44
Tipo cambio multilateral nominal	-0,28
IPSA	0,50
Activos de reserva	-0,18
Deuda gobierno general	0,58
Inversión extranjera directa	0,48
UF	0,59
Tasa política monetaria	-0,46
Índice percepción del consumidor	-0,24
Tipo de cambio nominal dólar	0,33
Índice percepción de la economía	0,15
IVTA	0,64
IVAASC	0,83
IVIC	0,85
IVAI	0,81
IVAPCT	0,77
IVASAA	0,89
IVAAER	0,71
IVOAS	0,79

Tabla 4.1: Análisis de correlación IMACEC vs indicadores

Se puede ver que hay una fuerte similitud entre la correlación de los índices con el IMACEC y el segundo grupo que se formó con los conglomerados. Es por esto que los índices que van a alimentar el índice compuesto serán aquellos que tienen mayor correlación con el IMACEC, superior a 0.51.

A continuación, la tabla de los 16 índices seleccionados con definición, fuente y frecuencia:

Índice	Definición	Fuente	Frecuencia
IAC (Índicador de Actividad de Comercio)	El objetivo del Índice de Actividad del Comercio (IAC) es estimar la evolución mensual del volumen de producción del comercio, sección G de la CIIU4.CL 2012, (comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas) mediante la utilización de las ventas a precios constantes de las empresas que se clasifican en esta actividad, como aproximación al valor agregado del sector.	INE	Mensual
IPMAN (Índice de Producción Manufacturera)	El objetivo del Índice de Producción Manufacturera es estimar la evolución mensual del volumen de producción de la industria manufacturera sección C de la CIIU4.CL 2012, (industrias manufactureras), utilizando principalmente, según tipo de actividad, variables de producción física, valor bruto de la producción (VBP), y en menor medida, variables de costo de avance deflactado y cantidad de horas-persona utilizadas en los procesos productivos de los distintos establecimientos que componen el sector.	INE	Mensual
IICOM (Índice de Inventarios del Comercio)	El objetivo del Índice de Inventarios del Comercio, que contiene el Índice de Inventarios de Supermercados, es estimar mensualmente la evolución, a precios corrientes, del nivel de los inventarios valorados contablemente del sector Comercio, con el fin de entregar antecedentes que faciliten un análisis más acabado de la economía nacional.	INE	Mensual
ISUP (Índice de Supermercados)	El objetivo del índice es estimar en el corto plazo la evolución de la actividad de los supermercados, a través de las ventas netas totales a precios constantes de los establecimientos de supermercados con tres o más cajas instaladas. Este índice tiene representatividad nacional y regional.	INE	Mensual
IPC (Índices de Precios del Consumidor)	El índice de Precios al Consumidor, es un indicador económico que mide mes a mes la variación de los precios de una canasta de bienes y servicios representativa del consumo de los hogares urbanos del conjunto de las capitales regionales y sus zonas conurbadas dentro de las fronteras del país.	INE	Mensual
UF (Unidad de Fomento)	Unidad de Fomento es una unidad de cuenta usada en Chile, reajustable de acuerdo con la inflación.	Banco Central	Diario
IR (Índice de Remuneraciones)	El Índice Nominal de Remuneraciones (IR) mide la evolución mensual de las remuneraciones ordinarias por hora ordinaria pagada a los trabajadores.	INE	Mensual
Deuda Gobierno General	La deuda pública corresponde a las obligaciones financieras (bonos, préstamos) contraídas por el gobierno, a través de las cuales se compromete a pagar intereses y el préstamo original en ciertas fechas determinadas. La deuda pública del Gobierno Central incluye la deuda de Tesorería y Corto.	Banco Central	Mensual
IVTA (Índice de Ventas de Transporte y Almacenamiento)	El objetivo de los Índices de Ventas de Servicios (IVS) nominales es estimar la evolución coyuntural del total de ventas a precios corrientes correspondientes a las Secciones de Servicios de mercado no financieros en medición, tomando en consideración a las empresas que se clasifican en estas categorías de actividad. La variable de seguimiento son las ventas netas a precios corrientes (ventas sin IVA y/o exentas de IVA), en pesos chilenos, incluyendo las ventas de actividades secundarias del giro.	INE	Mensual
IVAASC (Índice de Ventas de Actividades de Alojamiento y de Servicio de Comidas)		INE	Mensual
IVIC (Índice de Ventas de Información y Comunicaciones)		INE	Mensual
IVAI (Índice de Ventas de Actividades Inmobiliarias)		INE	Mensual
IVAPCT (Índice de Ventas de Actividades Profesionales, Científicas y Técnicas)		INE	Mensual
IVASAA (Índice de Ventas de Actividades de Servicios Administrativos y de Apoyo)		INE	Mensual
IVAAER (Índice de Ventas de Actividades Artísticas, de Entretenimiento y Recreativas)		INE	Mensual
IVOAS (Índice de Ventas de Otras Actividades de Servicios)		INE	Mensual

Tabla 4.2: Índices seleccionados para la construcción del índice compuesto

En esta investigación no fue necesario imputar datos, ya que la serie de datos parte desde 2014 al 2018. Se observaron y analizaron los datos y no hubo valores perdidos que tratar. Para ver si existía algún *outlier* se trabajó a nivel univariante donde se estandarizaron los datos, también se obtuvieron los coeficientes de curtosis y asimetría. Cabe señalar que hubo valores que sobrepasaron el umbral, pero al revisar el sector los valores puntuales que se detectaron como potencialmente atípicos, son valores propios del sector. Esto se da en sectores que son estacionales, por ejemplo comercio, que tiene valores muy altos en marzo y diciembre que fueron detectados como *outlier* pero que son propios del sector. En estos casos no se imputaron datos.

Para la normalización de los indicadores, se ocupó la transformación *z-score* y la transformación mínimo máximo. Al revisar que no había diferencia entre una transformación y otra, se optó por trabajar con la transformación *z-score*, donde se aplicaron una serie de técnicas para sacar los ponderados.

Para sacar las ponderaciones del índice compuesto, se ocupó la correlación simple como primer método, análisis de componentes principales y regresión lineal, con el fin de observar que método de ponderación ajusta y predice mejor el índice de referencia.

A continuación, la tabla de ponderaciones y los distintos métodos:

	Ponderaciones usadas		
Índice	CS	CP	RL
IAC	0,08	0,03	-0,29
IPMAN	0,05	0,30	0,27
ISUP	0,07	0,03	0,26
IICOM	0,05	0,03	-0,03
IPC	0,05	0,04	-0,44
IR	0,05	0,03	0,00
deuda gobierno general	0,05	0,04	0,17
uf	0,05	0,04	0,04
IVTA	0,06	0,10	0,04
IVAASC	0,07	0,03	0,04
IVIC	0,07	0,04	0,08
IVAI	0,07	0,03	0,13
IVAPCT	0,07	0,06	-0,01
IVASAA	0,08	0,04	0,79
IVAAER	0,06	0,12	-0,05
IVOAS	0,07	0,05	-0,01
Suma	1,00	1,00	1,00

CS: correlación simple, CP: componentes principales y RL: regresión lineal

Tabla 4.3: Ponderadores

Como se observa en la tabla 4.3, se sacaron ponderadores para cada método. Con el método de CS (correlación simple), se obtuvieron las ponderaciones de forma proporcional al total de la corre-

lación entre indicadores e IMACEC. En el método de CP (componentes principales) se trabajó con las primeras 4 componentes, las cuales acumularon el 92,17 % de la varianza, posteriormente se procedió a sacar el cálculo de la varianza de todos los indicadores explicada por cada componente. Para obtener la varianza explicada, se calcula la suma de cuadrados de los componentes de los indicadores. Se normalizan los cuadrados de los componentes, para esto se calcula el cuadrado del componente dividido por la varianza de los indicadores explicada por el componente. Se obtuvieron los pesos de cada componente mediante el cociente entre la varianza de los indicadores, explicada por cada componente y la varianza total de los componentes retenidas. Se calculan los pesos de cada indicador ponderando la máxima variabilidad de éste, explicada por el peso del componente. Finalmente, los pesos se normalizan dividiendo por la suma de todos ellos, de tal forma que la suma sea 1. En el caso del método RL (regresión lineal), se obtuvieron los pesos a través de los estimadores de la regresión para cada índice, los cuales se normalizaron sumando la unidad.

Ya con las ponderaciones sacadas, se procede a la obtención de un índice compuesto, en este caso ocuparemos la agregación lineal de la siguiente forma:

$$IC = \sum_{q=1}^{16} w_q I_q$$

donde $\sum_{q=1}^{16} w_q = 1$ y $0 \leq w_q \leq 1$ y I_q son los indicadores elegidos estandarizados.

A continuación, las series compuestas y comparadas con respecto al IMACEC:

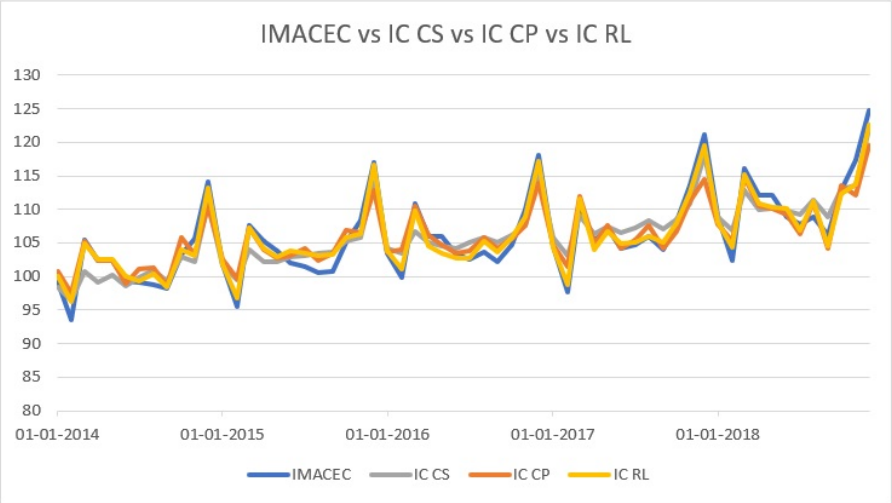

Gráfico 4.1: IMACEC vs índices compuestos

Como se observa en el gráfico 4.5, la serie que parece ajustar mejor al IMACEC, es el indicador compuesto con ponderaciones extraídas a través de regresión lineal (IC RL), seguida por las pondera-

ciones obtenidas a través de componentes principales (IC CP) y finalmente las ponderaciones sacadas por correlación simple (IC CS).

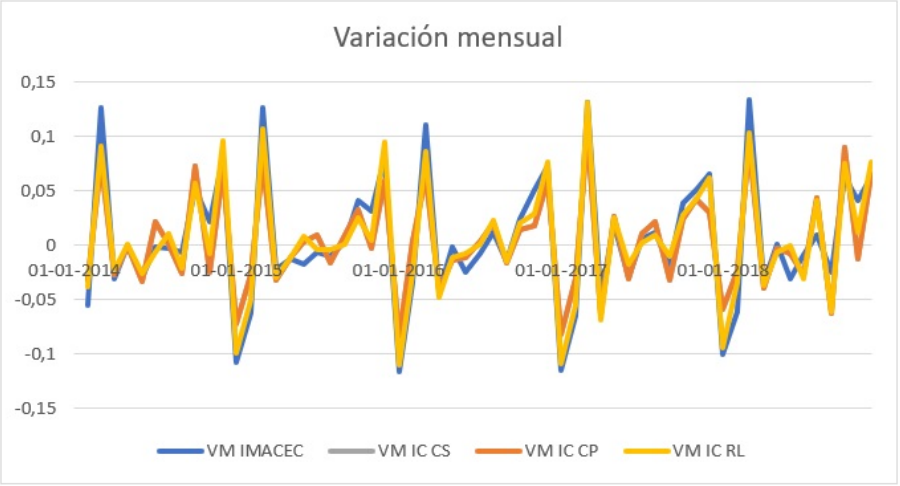

Gráfico 4.2: variación mensual IMACEC vs índices compuestos

En la figura 4.6, también se puede apreciar una gran similitud entre la variación mensual del IMACEC con respecto al indicador compuesto por regresión lineal (IC RL).

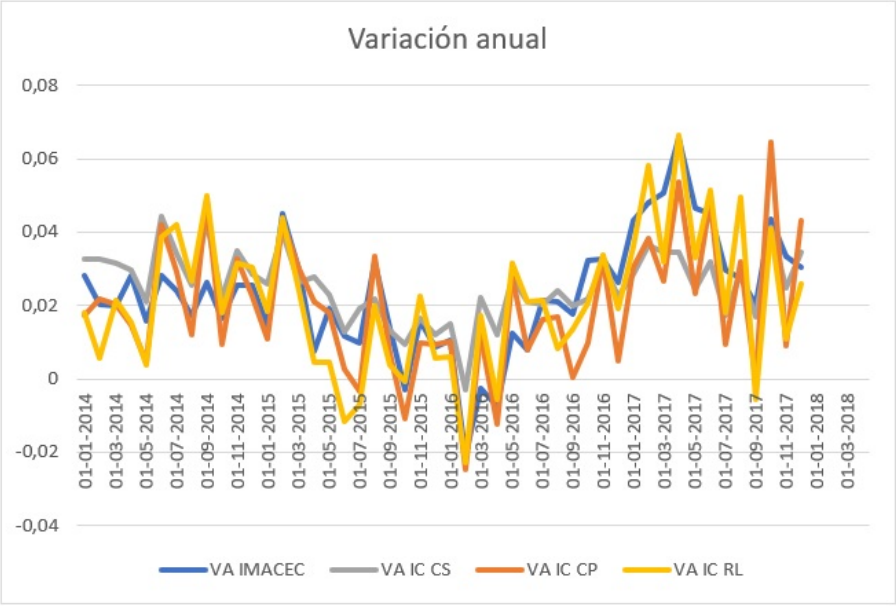

Gráfico 4.3: Variación anual IMACEC vs índices compuestos

En el gráfico 4.3 de variaciones anuales, se puede apreciar que hay mayor variación en los indicadores compuestos con respecto al IMACEC. Pero siguen la misma tendencia que la serie de referencia.

En los gráficos mostrados anteriormente, se puede ver que el indicador compuesto con ponderaciones extraídas a través de regresión lineal múltiple (IC RM) parece ajustarse mejor que el indicador compuesto con ponderaciones por componentes principales (IC CP) y correlación (IC CS). Para ver que indicador compuesto se ajusta mejor, se analizó la correlación entre el IMACEC y cada uno de los índices compuestos.

Método	Correlación
IMACEC/IC CS	0,92
IMACEC/IC CP	0,96
IMACEC/IC RL	0,98

Tabla 4.4: Correlación índices compuestos respecto a IMACEC

Para corroborar que método ajusta mejor al IMACEC, se sacó el error cuadrático medio definido por $ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ donde $\hat{Y}_i$ es el vector de los valores estimados e Y_i es el vector de los verdaderos valores. Para cada método, los resultados son los siguientes:

Método	ECM
IMACEC/IC CS	6,31
IMACEC/IC CP	5,42
IMACEC/IC RL	1,95

Tabla 4.5: Error cuadrático medio de índices compuestos respecto IMACEC

Como muestra la tabla 4.5, el indicador compuesto por ponderadores obtenidos a través de regresión lineal múltiple parece ajustar mejor que el indicador con ponderaciones obtenidas por componentes principales y por correlación simple.

A continuación, se proyectarán los datos de cada uno de los métodos aplicados y se compararán con los datos reales del IMACEC.

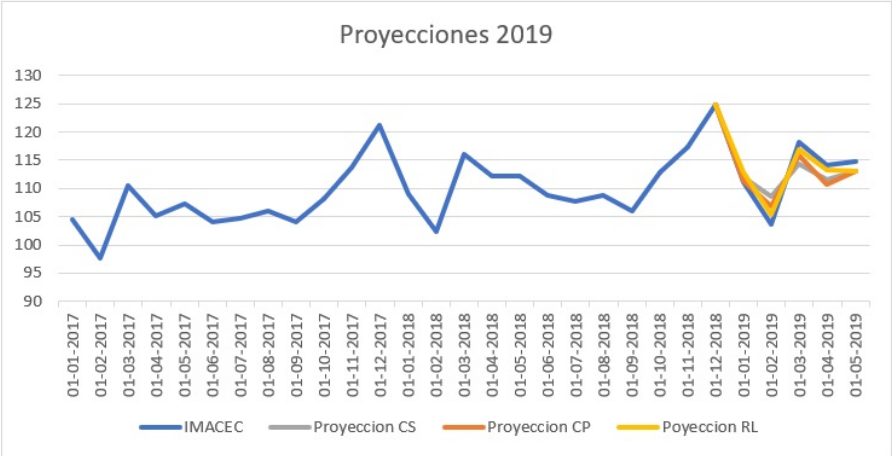

Gráfico 4.4: Proyecciones índices compuestos vs IMACEC 2019

Como se muestra en el gráfico 4.4, en las proyecciones se puede apreciar una gran similitud entre los valores del IMACEC y el indicador compuesto con ponderaciones obtenidos a través de regresión lineal.

A continuación, se detalla la tabla de proyecciones:

Fecha	IMACEC	Proyección CS	Proyección CP	Proyección RL
01-01-2019	111,1	112,1	111,3	112,8
01-02-2019	103,6	108,6	106,9	105,4
01-03-2019	118,1	114,3	115,9	116,9
01-04-2019	114,2	111,6	110,8	113,2
01-05-2019	114,8	113,0	113,1	113,1

Tabla 4.6: Proyecciones índices compuestos vs IMACEC 2019

Para ver que método ajusta mejor los datos del IMACEC en sus proyecciones, se sacó el error cuadrático medio como se muestra a continuación:

Método	ECM
IMACEC/IC CS	10,17
IMACEC/IC CP	6,07
IMACEC/IC RL	2,29

Tabla 4.7: Error cuadrático medio índices compuestos vs IMACEC 2019

Como se aprecia el método de ponderadores a través de regresión lineal, es el que mejor ajusta los datos del IMACEC, seguido por el método de componentes principales y finalmente correlación simple.

El mejor método para ajustar y predecir el IMACEC es el índice compuesto ponderado a través de regresión lineal, en éste caso se hará un análisis de sensibilidad global para el modelo elegido.

A continuación, la tabla con los factores de primer orden y total del análisis de sensibilidad.

Parámetros	Primer orden	Total	St-Si
X1_IAC	0,077518	0,079030	0,001512
X2_IPMAN	0,063416	0,064895	0,001479
X3_ISUP	0,061988	0,063464	0,001476
X4_IICOM	0,000658	0,001993	0,001335
X5_IPC	0,173192	0,174920	0,001728
X6_IR	0,000004	0,001343	0,001339
X7_deuda gobierno general	0,026395	0,027792	0,001397
X8_uf	0,001303	0,002647	0,001344
X9_IVTA	0,001828	0,003172	0,001344
X10_IVAASC	0,001373	0,002716	0,001343
X11_IVIC	0,005922	0,007273	0,001351
X12_IVAI	0,015491	0,016856	0,001365
X13_IVAPCT	0,000153	0,001494	0,001341
X14_IVASAA	0,566161	0,567531	0,001370
X15_IVAAER	0,001910	0,003252	0,001342
X16_IVOAS	0,000109	0,001451	0,001342

Tabla 4.8: Análisis de sensibilidad

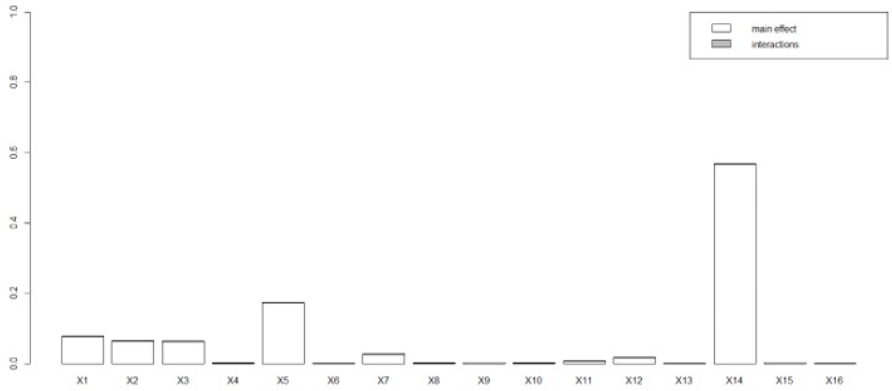

Gráfico 4.5: Análisis de sensibilidad

Como se observa en la tabla 4.8 y el gráfico 4.5 de sensibilidad, los factores más importantes o que aportan más al modelo son los parámetros asociados al IVASAA (Índice de Ventas de Actividades de Servicios Administrativos y de Apoyo) y al IPC (índice precios del consumidor), son los parámetros con mayor peso en el modelo. También se observo que los efectos principales son muy parecidos a las iteraciones entre parámetros y no hay diferencias significativas, esto quiere decir que los factores

input del modelo afectan de igual medida a los factores *output*, esto se debe a la alta correlación de las variables.

Capítulo 5

Conclusión

Los indicadores compuestos pueden ser útiles herramientas del proceso de toma de decisiones en el ámbito de diseño, implementación y evaluación de políticas públicas, siempre y cuando se cumplan los métodos expuestos en este documento.

El elemento básico que establece la construcción del indicador compuesto requiere una necesidad explícita previa que justifique su construcción. El presente documento ha descrito las más recientes metodologías aplicadas al procesamiento, cálculo y análisis de indicadores, del cual se han considerado los principales aspectos metodológicos estadísticos involucrados en el proceso como análisis exploratorio, normalización, ponderación, agregación, robustez y sensibilidad.

A través de las metodologías expuestas en el documento, se logró construir un indicador compuesto en referencia al indicador económico coyuntural IMACEC. Definido el marco conceptual, se logró obtener una serie de indicadores, los cuales a través de técnicas estadísticas fueron seleccionados. Una vez definidos los indicadores que alimentaron el índice compuesto, se normalizaron las variables y se sacaron tres tipos de ponderadores, para luego agregarlos y compararlos con el índice de referencia IMACEC. Como resultado se obtuvieron los errores cuadráticos medios y sus correlaciones. Se sacaron las proyecciones para el año 2019. Al obtener todos estos resultados, el método de ponderación por regresión lineal fue el método que más se acercó a IMACEC. Finalmente, se hizo un análisis de sensibilidad para los parámetros de este último método.

Cabe señalar este documento como una guía práctica para obtener indicadores compuestos y de qué forma se pueden utilizar.

Bibliografía

[1] Annoni, P. (2010). *Transformations.* En: Constructing composite indicators: From theory to practice. 20-21 de mayo. Ispra –Italia.

[2] Archer, G. y Saltelli, I. (1997). Sensitivity measures, anova-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation*, 58(1), 99 − 120 .

[3] Bas, M. (2014). *Estrategias Metodológicas para la Construcción de Indicadores Compuestos para la Gestión Universitaria.* Valencia: Universitat Politécnica de Valéncia.

[4] Castro Bonaño, J. M. (2002) *Indicadores de Desarrollo Sostenible Urbano: Una Aplicación para Andalucía.*Tesis Doctoral, Universidad de Málaga, http://www.eumed.net/tesis/jmc/.

[5] Cuadras, C. M. (2008). *Modelos estadísticos multivariantes.* Barcelona: Promociones y Publicaciones Universitarias.

[6] Cukier, R. I., Fortuin, C., Schuler, K. E., Petscheck, A. G. y Schabaibly, J. (1973). Study of the sensitivity of coupled reaction system to uncertainties in rate coefficients. I Theory. *The Journal of Chemical Physics*, 59(8), 3873 − 3878.

[7] de la Fuente, S. F. (2013) *Estadística descriptiva: números índices* Edición electrónica en http://www.fuenterrebollo.com/Economicas2013/indices-teoria.pdf.

[8] Dempster, A. P., Laird, N. M. y Rubin, D. B. (1977) Maximun Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B,* 39(1), 1 − 38.

[9] Frenderberg, M. (2003). *Composite Indicators of Country Performance: A Critical Assessment.* OECD STI Working Paper, DSTI/DOC(2003)16, OECD Publishing, Paris.

[10] Hair, J. F., Anderson, R., Tatham, R. y Black, W. (2007). *Análisis Multivariante.* Madrid: Prentice Hall.

[11] Jacobs, R., Smith, P., y Goddar, M. (2004). *Measuring Performance: An Examination of Composite Perfomance Indicators.* CHE Technical Paper Series 29, Centre for Health Economic, University of York.

[12] Little, R. J. A. y Rubin, D.B. (1987). *Statistical Analysis with Missing Data, Second Edition.* New Jersey: Jhon Wiley and Sons, Inc.

[13] Little, R. J. A. (2004). A Test of Missing Completely at Random for Multivariate Data with Missing Values.*Journal of the American Statistical Association*, 83(404), 1198 − 1202.

[14] Medina, F. y Galván, M. (2007) *Imputación de datos: teoría y práctica,* CEPAL- Serie Estudios estadísticos y prospectivos N° 54, División de Estadística y Proyecciones Económicas, Santiago de Chile.

[15] Munda, G. (2008). *Social Multi-Criteria Evaluation for a Sustantinable Economy.* Berlen Heidelberg: Springer-Verlag.

[16] Nardo, M., Saisana, M., Saltelli, A. y Tarantola, S., Hoffman, a. y Giovannini, E. (2005a). *Handbook on constructing composite indicators: Methodology and user guide,* OECD Statistics Working Paper, STD/DOC(2005)3.

[17] Nardo, M., Saisana, M., Saltelli, A., y Tarantola, S., Hoffman, A., y Giovannini, E. (2008). *Handbook on constructing composite indicators: methodology and user guide,* OECD Statistics Working Paper, STD/DOC (2005)3, OCDE Publishing, Paris.

[18] Nicoletti, G., Scarpetta, S., y Boylaud. O. (2000). *Summary indicators of product market regulation with an extension to employment protection legislation,* OECD Econimics Departament Working Papers ECO/WKP(99)18, OECD Publishing, Paris.

[19] Peña, D. (2002). *Análisis de datos multivariantes* Madrid: McGraw Hill.

[20] Pérez, C. (2005). *Métodos estadísticos avanzados con SPSS.* Madrid: Thomson Paraninfo.

[21] Rabitz, H., Alis, O.F., Shorter, J. y Shim, K. (1999) Efficient input-output model representations. *Computer Physics Communications*,117(1 − 2), 11 − 20.

[22] Robins, J. M., Rotnitzky, A. y Zhao, L. P. (1994). Estimation of Regression Coeficients when some Regressors are not always observed. *Journal of the American Statistical Association*,89(427), 846 − 866.

[23] Rojo, J.L., Gómez, B. F., Fernández-Abascal, H.,T., Fernández, J. y Sanz, J. A. (2018). *Materiales escritos y multimedia para la docencia de la Estadística orientada a la Economía y a la Empresa.*Edición electrónica en http://www5.uva.es/estadmed/datos/indices/indices.htm.

[24] Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* . New Jersey. John Wiley and Sons, Inc.

[25] Sacks, J., Welch, W., Mitchell, T. y Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science,* 4(4), 409 − 435.

[26] Saisana, M., Saltelli, A. y Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tolos for the quality assessment of composite indicators. *Journal of Royal Statistical Society A,* 168(2), $1 - 17$.

[27] Saisana, M. (2008). *2007 Composite Learning Index: Robustness Issues and Critical Assessment,* EUR 23274 EN, JRC Scientific and Technical Reports (EUR collection), Italy.

[28] Saisana, M. (2010). *Imputation of missing data* En: Constructing composite indicators: From theory to practise. $20 - 21$ Mayo.

[29] Sánchez, J. (2004). *Introducción a la Estadística Empresarial.* Edición electrónica en http://www.eumed.net/cursecon/libreria/index.htm.

[30] Saltelli, A., Chan, K. P. S. y Scott, M. (2000). *Sensitivity analysis.* EChichester: John Wiley and Sons, Ltd.

[31] Saltelli, A., Tarantola, S., Campolongo, F. y Ratto, M. (2004). *Sensitivity Analysis in practice: a guide to assessong scientific models.* New York: John Wiley and Sons, Ltd.

[32] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. y Tartantola, S. (2008). *Global Sensitivity Analisis. The primer.* Chichester; John Wiley and Sons, Ltd.

[33] Schuschny, A. y Soto, H. (2009). *Guía metodológica Diseño de indicadores compuestos de desarrollo sostenible.* Comisión Económica para América Latina y el Caribe (CEPAL).

[34] Sobol, I. M. (2003). Sensitivity analysis for non –linear mathematical models. *Mathematical Modelling and Computational Experiment,* 1(4), $407 - 414$.

[35] Spath, H. (1980). *Cluster Analysis Algorithms.* Chichester, England: Ellis Horwood.

[36] Uriel, E. (1980). *Análisis de datos. Series temporales y Análisis multivariante.* Madrid: Editorial AC.

[37] Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objetive Funtion. *Jorunal of the American Statistical Association,* 58(301), $236 - 244$.