



COMILLAS
UNIVERSIDAD PONTIFICIA

ICAI

ICADE

CIHS

MÁSTER EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

**Methodological Improvements for the Construction of
Composite Indicators / Mejoras metodológicas para la
construcción de indicadores compuestos**

Nombre del autor: Gustavo Álvarez Fernández

Nombre del director: David Roch Dupré

Madrid

Enero de 2025

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título
**“Methodological Improvements for the Construction of Composite Indicators /
Mejoras metodológicas para la construcción de indicadores compuestos”**

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2024-2025 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.



Fdo.: Gustavo Álvarez Fernández

Fecha: 20/01/2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO



Fdo.: David Roch Dupré

Fecha: 20/01/2025

RESUMEN

Mejoras metodológicas para la construcción de indicadores compuestos

Autor: Gustavo Álvarez Fernández

Director: David Roch Dupré

Entidad colaboradora: ICAI - Universidad Pontificia de Comillas

Palabras clave: Indicador sintético, análisis factorial (FA), análisis de importancia, imputación

1. Introducción y planteamiento del proyecto

En un mundo globalizado y en constante cambio, medir fenómenos económicos y sociales es clave para orientar estrategias y decisiones políticas. Los indicadores sintéticos permiten sintetizar información compleja y proporcionar una visión integral de sectores o fenómenos específicos. Ejemplos como el Índice de Desarrollo Humano (IDH) y el Índice de Competitividad Global demuestran su impacto en la formulación de políticas y evaluación del progreso económico y social. Sin embargo, estos indicadores enfrentan limitaciones en su adaptabilidad y capacidad para captar la complejidad de la realidad actual.

Este Trabajo de Fin de Máster (TFM) surge de la necesidad de perfeccionar los indicadores compuestos para representar con mayor precisión realidades complejas, como el mercado laboral. La investigación se basa en el Synthetic Labor Market Index (SLMI) (Álvarez, 2022), desarrollado previamente para evaluar el mercado laboral en el TFG, a la par que también ha sido revisitado en un artículo académico en revisión que analiza la relación del SLMI con indicadores económicos clave en España, como el PIB, la Tasa de Desempleo y el IPC.

El interés en este tema surge de la curiosidad por entender las dinámicas del mercado laboral y su impacto en el bienestar social. Utilizar herramientas cuantitativas para resolver problemas reales y generar impacto positivo en la sociedad es altamente motivador. Además, la investigación en este campo puede contribuir a un análisis de la realidad socioeconómica más preciso y acertado de cara a la implantación de políticas que tengan un impacto social.

De hecho, la experiencia adquirida con el SLMI ha evidenciado la importancia de metodologías rigurosas, especialmente en la imputación de datos y la asignación de pesos, factores clave en la validez de los indicadores. La gestión adecuada de datos ausentes es crucial para evitar sesgos y distorsiones en los resultados. Aplicar técnicas avanzadas, como la imputación múltiple, mejorará la representatividad de los datos. Asimismo, una correcta ponderación de las variables garantizará indicadores más válidos y relevantes. Aunque existen técnicas consolidadas para construir indicadores, cada caso requiere un enfoque específico basado en criterios estadísticos y juicio del investigador. Mejorar estas tres etapas críticas permitirá obtener indicadores más precisos y representativos en diversos ámbitos económicos y sociales.

Esta investigación no solo busca fortalecer la metodología del SLMI, sino también mejorar habilidades en la construcción y análisis de indicadores. De hecho, el primer objetivo implica una revisión metodológica de las etapas críticas en la construcción de indicadores sintéticos, comenzando con un análisis general de la bibliografía y la elaboración del marco teórico. Luego, se elegirán las técnicas y los indicadores a estudiar, y se recolectarán y estructurarán los datos en MATLAB para su análisis. El segundo objetivo se enfoca en el análisis metodológico de las técnicas de imputación de información ausente, con una revisión de la literatura sobre cómo diferentes estructuras de datos afectan los resultados de la imputación y un análisis en MATLAB y IBM SPSS comparando varios métodos de imputación.

El tercer objetivo se centra en evaluar la fase de ponderación, examinando la “importancia” proporcionada por el criterio experto y la asignación basada en la variabilidad estadística de los métodos FA. Se revisará la literatura sobre la ponderación y sus posibles mejoras metodológicas, seguido de un análisis en MATLAB para evaluar el impacto de las correlaciones no lineales y su efecto sobre la robustez de los indicadores. Finalmente, el cuarto objetivo se refiere a la extracción de conclusiones para replicar el análisis de importancia y FA en otros indicadores, con una evaluación final de los resultados y la redacción del informe del proyecto.

Este TFM busca así ser un puente para el desarrollo de métricas estadísticamente sólidas y útiles en la toma de decisiones económicas y políticas, siendo una oportunidad de crecimiento académico y profesional en un campo de gran impacto.

2. Metodología

a. Análisis de imputación con diferentes estructuras de datos

El análisis metodológico de las técnicas de imputación de información ausente se centra en entender cómo los diferentes tipos de estructuras de datos, como las series temporales y los datos de panel, influyen en los resultados de la imputación. Entender cómo cada estructura impacta en los resultados de la imputación es esencial para elegir y aplicar las técnicas más adecuadas, lo que permitirá obtener resultados más precisos y fiables en la construcción de indicadores. En el presente estudio se valoran, en línea con el TFG realizado, las imputaciones múltiple, Expectation-Maximization y la regresión lineal.

Aunque la imputación múltiple y el método EM presentan características ventajosas, su implementación puede generar resultados imprevistos. Para minimizar esta incertidumbre, se recomienda realizar un análisis empírico, similar al propuesto por Gold y Bentler (2000). Este método implica la eliminación aleatoria de ciertos datos de un conjunto disponible, seguida de la aplicación de diversas técnicas de imputación. Luego, se compara la diferencia entre los valores imputados y los originales utilizando un indicador, como el error cuadrático medio normalizado (NRMSE, por sus siglas en inglés).

De esta manera, la ecuación del NRMSE para n indicadores es la siguiente:

$$NMRSE = \frac{\sqrt{\frac{\sum_i^N (X_i - \hat{X}_i)^2}{N}}}{\max(X_i) - \min(X_i)}$$

donde X_i es el valor conocido y real de las variables i sin normalizar, \hat{X}_i es el valor imputado de cada variable i sin normalizar.

Una vez explicado el mecanismo para comparar los métodos de imputación, es importante resaltar el verdadero valor del proyecto: la comparación del NRMSE entre las diferentes estructuras de datos. Al imputar los datos de variables que están asociadas a áreas que definen conjuntos sintéticos disponibles en varios países, surge la pregunta de qué estructura de datos utilizar para la imputación: ¿debería hacerse sobre todo el conjunto de datos, incluyendo todas las variables y países a la vez, o debería realizarse variable por variable para capturar mejor las variaciones específicas de cada indicador? ¿O, por el contrario, sería mejor imputar considerando todos los indicadores, pero país por país?

Para aterrizar este análisis metodológico, se estudiará el impacto de imputar a diferentes niveles de agregación del SLMI y de un indicador con estructura de panel, el Senior Economy Tracker (Roch-Dupré et al., 2023), que mide el progreso de 27 economías europeas en su adaptación a un

contexto de transición demográfica a lo largo del tiempo ($N = 18$ años). Su estructura de datos en forma de panel lo hace idóneo para la labor que se procederá a realizar.

Una vez realizados los análisis, se han obtenido los valores de NRMSE incorporando la información cruzada de los países. De estos resultados se deduce que imputar todos los datos juntos, incorporando la información de panel en la imputación el NRMSE se reduce. De hecho, este comportamiento se ratifica también en cuando se imputa por áreas o por subáreas, en tanto que el NRMSE siempre aumenta cuanto menor información se impute de forma conjunta. Los datos mostrados incorporan la información de panel de los 27 países europeos que disponen del indicador.

Sin embargo, el análisis también se ha reproducido obviando la característica de panel sobre el indicador, es decir, imputando las variables de cada país por separado y sumando los NRMSE por comparabilidad en la que la imputación se realiza a lo largo del panel. Comparando las imputaciones con las dos estructuras de datos, se percibe una mejora consistente al ampliar el tamaño de la muestra en todos los niveles, tanto incluyendo variables de áreas diferentes como de países diferentes. Tal y como arguyen Young y Johnson (2015), la imputación con datos de panel aumenta la cantidad de información disponible para el análisis, lo que mejora la eficiencia y reduce los errores estándar.

Esto queda patente en el análisis, en tanto que además el ajuste tiene un mayor impacto dependiendo del método empleado, dándose para la regresión la mayor reducción en el NRMSE. Después de todo, aunque el conjunto de datos original dispone de 18 datos, la imputación solo se puede aplicar sobre un set de datos completo. Esto reduce la mayoría de las variables a aproximadamente 12 datos (ya que la serie temporal debe estar completa para considerarse válida a la hora de imputar), a las que posteriormente han de eliminarse 3 más para evaluar el NRMSE. Por ello, la cantidad de información disponible para el indicador Senior Economy tracker estudiado es realmente baja si no se imputa considerando los países.

Este mismo análisis se ha realizado sobre el SLMI, con la salvedad de que el SLMI no dispone de datos por países, por lo que solo se puede evaluar el impacto de la estructura de datos al incorporar todas las variables frente a la imputación por áreas o variables. Los resultados son también consistentes con lo mostrado para el Senior Economy Tracker, obteniendo un NRMSE de 0.9747 para la técnica EM aplicada sobre todo el conjunto de datos sobre la serie temporal que se empleó en el trabajo (Álvarez, 2022) que además compara muy positivamente frente al 2.564 obtenido al construir el SLMI. Sin embargo, debido a la robustez del indicador, esto apenas goza de impacto sobre el indicador final.

b. Métodos avanzados de ponderación: FA vs análisis de importancia

El análisis factorial es una técnica estadística utilizada para reducir la dimensionalidad de los datos y encontrar estructuras subyacentes en un conjunto de variables observadas. Para asignar los pesos a las variables, sigue varios criterios fundamentales que permiten determinar la influencia de cada variable en los factores extraídos. Estos criterios se basan en la correlación entre variables, la varianza explicada, la rotación factorial, la carga factorial y la interpretación teórica (OCDE, 2008). Cabe resaltar que, en el análisis factorial, se busca reducir el número de variables manteniendo la mayor cantidad posible de información, por lo que la varianza explicada se convierte en el elemento decisor a la hora de elegir el número de factores que definirán dicha realidad. En este sentido, los factores se ordenan según la proporción de varianza que explican del conjunto de datos original.

Por otro lado, el análisis de importancia se basa en comprender las correlaciones entre las variables y evaluar la sensibilidad de cada una en relación con el modelo. A diferencia del análisis

factorial, que se enfoca en identificar estructuras subyacentes, el análisis de importancia compara el índice de sensibilidad de primer orden de cada variable, considerando correlaciones tanto lineales como no lineales. Esto permite ajustar los pesos de manera que se minimice la desviación entre la importancia real de cada variable y su importancia deseada dentro del modelo. Este enfoque es particularmente útil para garantizar que las variables que tienen un papel fundamental en la interpretación del modelo reciban un peso acorde a su impacto real.

En el contexto de indicadores compuestos, aunque la fórmula de agregación suele ser lineal, las correlaciones entre variables pueden introducir no linealidades en la relación entre y y x_i . Para capturar estas dinámicas, se emplea una medida más general llamada S_i , también conocida como índice de sensibilidad de primer orden. Este enfoque mide la influencia potencialmente no lineal de cada variable en el indicador compuesto y puede interpretarse como la reducción esperada en la varianza de y al fijar x_i (Becker et al., 2017).

El índice de sensibilidad se definiría por tanto como:

$$S_i = \frac{Var(E_{x \sim i}[y|x_i])}{Var(y)}$$

donde $E_{x \sim i}[y|x_i]$ es el valor esperado de y para x_i , asumiendo que el resto de variables i son evaluadas constantes. Esto puede interpretarse como una regresión no lineal de y sobre x_i , donde se observa la varianza explicada por la curva ajustada. De forma estimada, S_i puede hallarse como:

$$\hat{S}_i = \frac{\sum_{j=1}^n (\hat{m}_{ij} - \bar{m}_i)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

donde \hat{m}_{ij} representa el valor ajustado de la regresión de y sobre x_i para la observación j , y \bar{m}_i es el promedio de esos valores ajustados. El denominador simplemente muestra la diferencia entre la variable y respecto de su media (Becker et al., 2017).

En resumen, el índice de sensibilidad de primer orden (S_i) proporciona una forma efectiva de evaluar la importancia de las variables de entrada en un indicador compuesto, considerando tanto las relaciones lineales como las no lineales. De hecho, la literatura (Hastie et al., 2008) ya describe los principales métodos de regresión que se pueden aplicar para ejecutar la regresión no lineal de y sobre x : procesos gaussianos y splines penalizados. Este último, que viene a basarse en una regresión con componentes elevados a una potencia, viene definido por:

$$f(x_i) = \beta_0 + \beta_1 \cdot x_i + \sum_t^N \beta_t (x_i - \alpha_t)_+^k$$

Donde la β_0 y β_1 son los coeficientes de la parte lineal y α_t representa los diferentes *knots*, puntos que representan discontinuidades en la regresión del modelo.

En el ámbito del análisis de sensibilidad, se ha observado ampliamente (Xu & Gertner, 2008) que, cuando las variables de entrada están correlacionadas, S_i (también conocida como índice de sensibilidad de primer orden) incluye tanto el efecto directo de la variable x_i como el impacto derivado de su correlación con otras variables. Esto se expresa como:

$$S_i = S_i^{incorrelacionada} + S_i^{correlacionada}$$

donde $S_i^{incorrelacionada}$ representa la parte no correlacionada y $S_i^{correlacionada}$ la parte correlacionada. Este desglose proporciona información adicional sobre la influencia de cada variable.

De esta forma, resulta capital descomponer S_i , para lo que se sigue el enfoque de Xu y Gertner (2008), basado en regresión. Para obtener la parte no correlacionada $S_i^{incorrelacionada}$, primero se

realiza una regresión de x_i sobre el resto de variables, de manera que se obtenga una regresión multivariante para modelar x_i como función de las demás variables:

$$x_i = \beta_0 + \sum_{k \neq i} \beta_k x_k$$

donde los coeficientes β_k se obtienen mediante mínimos cuadrados. De esta manera, se define la varianza de x_i que puede ser explicada a través del resto de variables. Posteriormente, se calculan los residuos \hat{z}_i como:

$$\hat{z}_i = x_i - \hat{x}_i = x_i - (\beta_0 + \sum_{k \neq i} \beta_k x_k)$$

Al restar esta estimación \hat{x}_i del valor real de x_i , se aísla únicamente la parte de x_i que no puede ser explicada por las demás variables. Esto elimina la dependencia lineal entre x_i y x_k , puesto que toda la varianza de x_i que era explicada a través del resto de indicadores ha sido restada de su componente. Por tanto, el residuo \hat{z}_i representa la variabilidad que únicamente x_i puede explicar de y , que sigue siendo el indicador sintético.

Finalmente, se realiza la regresión de y sobre \hat{z}_i para obtener $S_i^{incorrelada}$ acometiendo la siguiente fórmula:

$$S_i^{incorrelada} = \frac{Var(E_{\hat{z}_i}[y|\hat{z}_i])}{Var(y)} = \frac{\sum_{j=1}^n (\hat{y}_j^i - \bar{y}^i)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

De esta manera, habiendo calculado $S_i^{incorrelada}$ y S_i , la obtención de $S_i^{correlada}$ resulta sencillo si recordamos la fórmula antes mencionada:

$$S_i = S_i^{incorrelada} + S_i^{correlada}$$

Una vez comprendido el proceso de cálculo de importancias, una cuestión que surge naturalmente es cómo ajustar los pesos de las variables para lograr una cierta importancia deseada. Esto se puede abordar mediante un algoritmo de optimización que minimice la diferencia entre el índice de sensibilidad normalizado (S_i) y el peso deseado (w_i). El problema de optimización se formularía como:

$$w_{opt} = \underset{w}{\operatorname{argmin}} \sum_i^d (S_i^* - S_i(w))^2$$

donde S_i^* es el índice de sensibilidad objetivo normalizado, y $S_i(w)$ se calcula para cada conjunto de pesos w como el índice de primera sensibilidad mencionado anteriormente (Becker et al., 2017). Nótese que, típicamente, el criterio experto asigna pesos en un intento de reflejar la "importancia" que se desea tengan los indicadores, por lo que dicho índice de sensibilidad objetivo realmente serían los pesos proporcionados por el panel de investigadores o expertos.

Como parte del análisis de importancia, se debe primero comprender la estructura de los datos y la potencial existencia de correlaciones no lineales que puedan no ser captadas por otros métodos de agregación como el Factor Analysis (FA).

Para ello, en primer lugar, se calculan las regresiones lineales y no lineales a través de splines de segundo grado que reflejen dichas no linealidades. En dichos análisis se observa claramente cómo la mayoría de las variables muestran un comportamiento razonablemente lineal, con la salvedad de la variable A.2 donde el spline se desvía en gran medida de la recta de correlación lineal. Si bien otras variables muestran una ligera variación, no se consideran suficientemente reseñables.

Retomando el análisis de importancia, a continuación se presenta el índice de primera sensibilidad para todas las variables de la categoría *Desempleo*, que es un reflejo de la correlación lineal y no lineal respecto del área estudiada, cuantificando el impacto mostrado en las gráficas mostradas anteriormente:

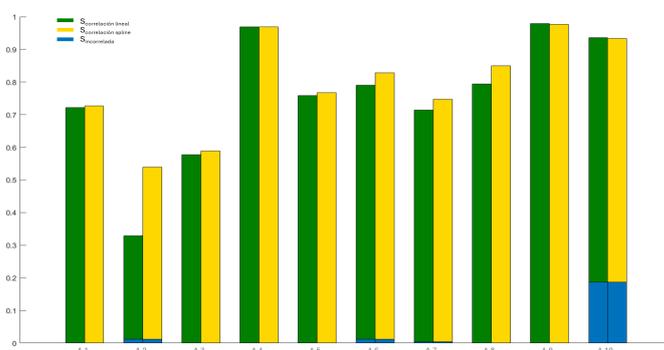


Ilustración 1. Índices de sensibilidad de primer orden sobre la categoría *Desempleo*

De la gráfica, se observa cómo la mayor variación entre la importancia obtenida linealmente y de spline se da precisamente en la variable A.2. Dada esta notable variación, se considerará de ahora en adelante esta estimación por splines penalizados para el área “Desempleo”.

En segundo lugar, lo primordial que ha de valorarse es la existencia de una importancia incorrelada que apenas se muestra en el resto de las variables (0.1898). Nótese que la importancia, en tanto que es definida como el índice de sensibilidad de primer orden de x (indicador base) sobre y (indicador sintético), goza siempre de una componente correlada de x_i explicada a través del resto de variables $x_{\sim i}$ y una componente incorrelada de la variabilidad de y explicada a través de x_i . Esto está basado en el propio concepto de importancia: una variable no solo afecta sobre el conjunto según el peso recibido, sino también a través de aquellos otros indicadores con los que comparta comportamiento y , por tanto, correlación.

Una vez comprendido el proceso a través del cual se han obtenido los valores de la ilustración, se deduce que la variable A.10 puede ser explicada a partir de las demás de forma tan significativa como el resto de las variables, pero goza de una componente de incorrelación entre x_i e y que no aparece en el resto de indicadores. Por ello, de acuerdo con la teoría estudiada (Becker et al., 2017), sobre esta variable tendrá el mayor impacto la variación del peso que se pueda realizar a través de una optimización.

A continuación, se debe comparar dicho índice de sensibilidad con el peso, no tanto en valores absolutos sino relativos para comparar cuán importante es cada variable sobre el área final. Para ello, se lleva a cabo un reescalado a 1 de la importancia obtenida por spline y se compara frente al peso original del SLMi, la varianza explicada por cada indicador base en el SLMi y también frente a la equiponderación con que ha sido construido el nuevo indicador *Desempleo* construido de forma equiponderada (ilustración A2 del anexo 2).

Queda por tanto patentes ciertas desviaciones en las aportaciones de cada variable al resultado final del área “Desempleo”. La mayor desviación se produce en las variables A.2, A.3, A.4, A.9 y A.10. Este problema probablemente surge porque la parte correlacionada de S_i prevalece sobre la parte no correlacionada en todos los indicadores. Esto implica que las correlaciones, más que los pesos, son las que determinan la importancia, como se ha mencionado anteriormente.

Parece que, a mayor correlación, el FA busca asignar un menor peso para compensar aquella variabilidad que ya es explicada por otras variables. No obstante, en cuanto al impacto sobre la importancia de cada variable, la variación del peso solo tendrá un efecto significativo en aquellos casos en los que la importancia no correlacionada goce de entidad, puesto que si no toda la importancia se obtendría a partir de las regresión lineal entre indicadores base y , por tanto, el

resultado de alterar uno de sus pesos será marginal sobre el sintético si el impacto está mayormente definido a través de la variabilidad que es explicada a través del resto de variables.

Siguiendo con el análisis, se procede a optimizar los pesos para obtener la importancia buscada equiponderada:

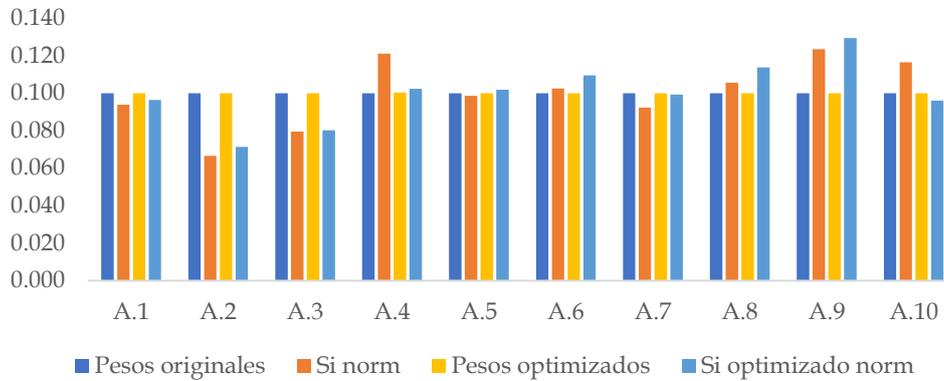


Ilustración 2. Resultados para la categoría Desempleo

A partir de la ilustración mostrada cabe resaltar que el impacto de la optimización sobre la categoría es pequeño, en tanto que los pesos varían ligeramente y logran reducir las desviaciones en las variables A.2, A.3, A.4, A.7 y A.10. De hecho, la variación más notable se da en esta última variable, lo cual resulta consistente con lo explicado anteriormente. Para concluir este análisis, cabe mencionar que la suma de las desviaciones cuadráticas (función objetivo) entre importancia deseada y obtenida ha mejorado de 0.054 frente a la nueva de 0.049.

De igual modo, cabe resaltar la categoría de *Protección a los desempleados*, en la que de forma notable queda patente que este mecanismo habría sido metodológicamente más robusto que la simple equiponderación que se aplicó en la construcción original del SLMI. Por supuesto, si bien el impacto sobre el SLMI será reducido tanto por la escasa desviación frente a la importancia deseada como porque dicha área es la que menos pondera en el segundo nivel de agregación (10% frente al 30% del resto de áreas), desde el punto de vista teórico esto permite extraer ciertas conclusiones al respecto.

A continuación se muestran las importancias para dicha área así como la matriz de correlaciones:

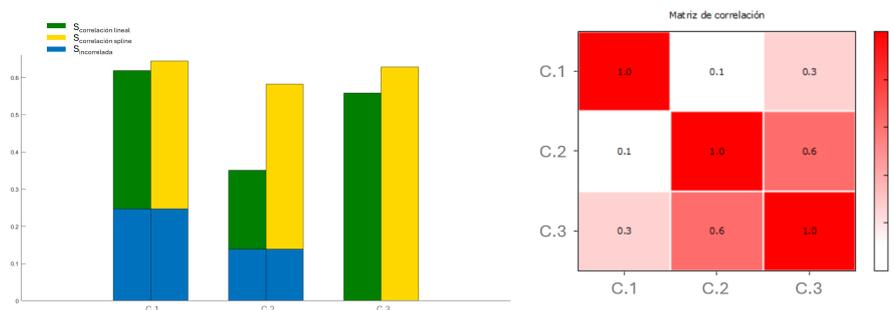


Ilustración 3. Índices de sensibilidad de primer orden y matriz de correlaciones de Protección a los desempleados

Nótese cómo en esta categoría la importancia incorrelada goza de una importancia mucho mayor en las dos primeras variables, lo cual se ve además ratificado por su matriz de correlaciones, donde se observa menor correlación entre las primeras dos variables.

En este caso donde no se ha aplicado FA sobre el set de datos, parece que las dos primeras variables, menos correladas tanto a través de la matriz de correlaciones como del análisis de importancia $S_i^{correlada}$, serán aquellas donde la optimización podrá mejorar de forma más notable

los resultado. Nótese cómo los valores de $S_i^{correlada}$ y las matrices de correlación, tanto en esta categoría como en la anterior, muestran resultados consistentes. Sin embargo, en este caso se denota una componente de $S_i^{incorrelada}$ mayor, lo que mejorará el análisis de importancia ya que los pesos asignados disponen por tanto de un efecto directo sobre la importancia de cada indicador componente. Así, se realiza el análisis mencionado y, efectivamente, se logra exitosamente obtener una importancia exactamente igual a la deseada:

		Equiponderación	Si	Si norm	Pesos optimizados	Si optimizado	Si optimizado norm
Protección a los desempleados - C	C.1	0.333	0.645	0.348	0.332	0.630	0.333
	C.2	0.333	0.582	0.314	0.392	0.630	0.333
	C.3	0.333	0.628	0.339	0.275	0.630	0.333

Tabla 1. Resultados del análisis de sensibilidad de primer orden y su optimización para la categoría C

Además, también se planteó el escenario en que todas las variables sean incorporadas directamente sobre el SLMI, obviando el paso intermedio de subagregación en torno a las áreas mencionadas. Asumiendo que todos los indicadores base gocen del mismo peso en un contexto de equiponderación, cada uno se vería asignado con 0.033 en tanto que el SLMI está conformado por 30 indicadores base. Parece destacarse que, aquellas categorías con un grado alto de correlación en su mayoría lineal, el FA proporciona mayores ventajas que el análisis de importancia, en tanto que puede reducir la dimensionalidad a partir de la varianza explicada (aunque en el SLMI no se aplica esta reducción) y proporciona al autor unos pesos cuando se lidia con un número de variables demasiado elevado. El análisis de importancia no es capaz de proporcionar unos pesos óptimos que de forma notable alteren la estructura del sintético final.

Esto, además, puede verse ratificado por la otra agregación que en el caso del SLMI se realiza con criterio experto: la agregación de las 4 áreas en torno al SLMI, donde se asignan pesos del 30% a *Desempleo, Empleo y Salarios y costes laborales* y un 10% a *Protección a los desempleados*.

A continuación, se proporciona la tabla con la optimización de las áreas frente al SLMI:

	Pesos	Si	Si norm	Pesos opt	Si opt	Si opt norm
Desempleo – A	0.300	0.528	0.344	0.156	0.686	0.256
Empleo – B	0.300	0.284	0.185	0.301	0.741	0.276
Protección a los desempleados – C	0.100	0.383	0.250	0.304	0.565	0.211
Salarios y costes laborales – D	0.300	0.339	0.221	0.239	0.689	0.257

Tabla 2. Resultados de la optimización para las áreas del SLMI

Excepto por el área *Desempleo*, cuya distancia frente a la importancia deseada no varía, el resto de las categorías presentan una importancia optimizada mucho más fiel a los pesos otorgados por el criterio experto. De esta forma, parece que allí donde el criterio experto impera, el análisis de importancia es una herramienta altamente valiosa para la ponderación.

3. Conclusiones

En lo referente a los mejoras metodológicas propuestas para la imputación, trabajar con el conjunto completo de datos para la imputación ofrece varias ventajas clave. Permite capturar relaciones y patrones más completos que podrían perderse con un enfoque local, mejorando la precisión al aprovechar relaciones globales entre variables. Además, reduce la variabilidad entre áreas, generando resultados más estables y menos influenciados por una posible falta de datos

desmesurada en algunos de las categorías. También resulta una herramienta útil para evitar el sobreajuste que puede ocurrir al imputar de manera local con pocos datos.

En cuanto al análisis de ponderación, el FA es útil para reducir variables en estructuras de datos grandes, especialmente cuando las correlaciones son altas, mientras que el análisis de importancia es adecuado cuando se busca asegurar trazabilidad de los datos o cuando se aplican criterios expertos. Aunque en el SLMI no se redujo el número de indicadores utilizando FA en una estructura jerárquica, sí se obtuvieron unos pesos estadísticos y se optó por una agregación intermedia de áreas específicas.

En cambio, el análisis de importancia es más útil para asignar pesos relativos a los indicadores cuando el criterio experto es el elegido por dimensionalidad y utilidad. De hecho, el análisis de importancia, en tanto que depende de unas importancias deseadas que deben ser proporcionadas por un panel de investigadores, sufre las mismas limitaciones que el criterio experto en aplicabilidad.

En cualquier caso, el análisis de importancia ha revelado que, a mayor $S_i^{incorrelada}$ mayor impacto tiene el peso de la propia variable sobre la importancia y, por tanto, más factible es aplicarle cambios u optimizaciones para reflejar mejor los deseos del criterio experto aplicado. De igual manera, cuando el número de variables analizada es muy grande, la $S_i^{correlada}$ domina la importancia, lo que efectivamente desaconseja el uso del propio análisis de importancia en tanto que la optimización de los pesos no logra afectar de forma significativa a los valores obtenidos de importancia.

En cualquier caso, se han abierto ciertas líneas de investigación referidas a, entre otras cosas, la relación entre la asignación de pesos del método FA frente a las correlaciones existentes entre las variables, el impacto de la variación en los *knots* y término de suavizado de la regresión por spline, y también la inclusión de más técnicas de imputación para evaluar el impacto de las estructuras de datos.

4. Referencias

- Álvarez, G. (2022). *Desarrollo de un KPI sintético para medir la recuperación española frente al COVID-19*. Madrid: Universidad Pontificia Comillas.
- Becker, W., Saisana, M., Paruolo, P., & Vandecasteele. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological indicators*, 80, 12-22.
- Gold, M. S., & Bentler, P. M. (2000). Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(3):319–355.
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. Springer Series.
- OCDE. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. París: OECD Publications.
- Roch-Dupré, D., Aracil, E. & Calvo-Bascones, P. (2023). Monitoring socioeconomic readiness for the demographic transition: Introducing the Senior Economy Tracker . *Structural Change and Economic Dynamics*, Vol. 71, 430-443.
- Young, R., & Johnson, D. (2015). Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. *J Marriage Fam*, 77 (1) 277-294

ABSTRACT

Methodological Improvements for the Construction of Composite Indicators

Author: Gustavo Álvarez Fernández

Director: David Roch Dupré

Collaborating entity: ICAI - Universidad Pontificia de Comillas

Key words: Synthetic indicators, factor analysis (FA), importance analysis, imputation

1. Introduction and project outline

In a globalized and constantly changing world, measuring economic and social phenomena is key to guiding strategies and political decisions. Synthetic indicators allow for the synthesis of complex information and provide a comprehensive view of specific sectors or phenomena. Examples like the Human Development Index (HDI) and the Global Competitiveness Index demonstrate their impact on policy formulation and the evaluation of economic and social progress. However, these indicators face limitations in their adaptability and ability to capture the complexity of the current reality.

This Master's Thesis (TFM) arises from the need to improve composite indicators to more accurately represent complex realities, such as the labor market. The research is based on the Synthetic Labor Market Index (SLMI) (Álvarez, 2022), which was previously developed to evaluate the labor market in the Bachelor's Thesis (TFG) and has also been revisited in an academic article under review, analyzing the relationship between the SLMI and key economic indicators in Spain, such as GDP, Unemployment Rate, and CPI.

The interest in this topic stems from a curiosity to understand the dynamics of the labor market and its impact on social well-being. Using quantitative tools to solve real-world problems and generate a positive impact on society is highly motivating. Furthermore, research in this field can contribute to a more precise and accurate analysis of the socio-economic reality, which will inform the implementation of policies that have a social impact.

In fact, the experience gained with the SLMI has highlighted the importance of rigorous methodologies, especially in data imputation and weight assignment, which are key factors in the validity of indicators. Proper management of missing data is crucial to avoid biases and distortions in the results. Applying advanced techniques such as multiple imputation will improve data representativeness. Additionally, correct weighting of variables will ensure more valid and relevant indicators. Although there are established techniques for building indicators, each case requires a specific approach based on statistical criteria and the researcher's judgment. Improving these three critical stages will lead to more accurate and representative indicators in various economic and social fields.

This research not only aims to strengthen the SLMI methodology but also to improve skills in constructing and analyzing indicators. In fact, the first objective involves a methodological review of the critical stages in building synthetic indicators, starting with a general analysis of the literature and the development of the theoretical framework. Then, the techniques and indicators to be studied will be chosen, and data will be collected and structured in MATLAB for analysis. The second objective focuses on the methodological analysis of missing data imputation techniques, with a review of the literature on how different data structures affect imputation results and an analysis in MATLAB and IBM SPSS comparing various imputation methods.

The third objective is to evaluate the weighting phase, examining the "importance" provided by expert criteria and assignment based on statistical variability from FA methods. Literature on weighting and its potential methodological improvements will be reviewed, followed by an analysis in MATLAB to assess the impact of nonlinear correlations and their effect on the robustness of indicators. Finally, the fourth objective involves drawing conclusions to replicate the importance and FA analysis in other indicators, with a final evaluation of the results and the writing of the project report.

Thus, this TFM aims to serve as a bridge for the development of statistically robust and useful metrics in economic and political decision-making, offering an opportunity for academic and professional growth in a field of great impact.

2. Methodology

a. Imputation analysis on different data structures

The methodological analysis of missing data imputation techniques focuses on understanding how different types of data structures, such as time series and panel data, influence the imputation results. Understanding how each structure impacts the imputation results is essential for choosing and applying the most suitable techniques, which will lead to more accurate and reliable results in the construction of indicators. In the present study, in line with the TFG carried out, multiple imputation, Expectation-Maximization (EM), and linear regression imputations are evaluated.

Although multiple imputation and the EM method present advantageous characteristics, their implementation may lead to unforeseen results. To minimize this uncertainty, an empirical analysis is recommended, similar to the one proposed by Gold and Bentler (2000). This method involves the random removal of certain data from an available set, followed by the application of various imputation techniques. The difference between the imputed and original values is then compared using an indicator such as the normalized root mean square error (NRMSE).

Thus, the NRMSE equation for n indicators is as follows:

$$NMRSE = \frac{\sqrt{\frac{\sum_i^N (X_i - \hat{X}_i)^2}{N}}}{\max(X_i) - \min(X_i)}$$

Where X_i is the known and real value of the i -th variable without normalization, and \hat{X}_i is the imputed value of each i -th variable without normalization.

Once the mechanism for comparing imputation methods is explained, it is important to highlight the true value of the project: the comparison of NRMSE across different data structures. When imputing data from variables associated with areas that define synthetic sets available in various countries, the question arises about which data structure to use for imputation: should it be done on the entire data set, including all variables and countries at once, or should it be done variable by variable to better capture the specific variations of each indicator? Or, conversely, would it be better to impute by considering all indicators, but country by country?

To ground this methodological analysis, the impact of imputing at different aggregation levels of the SLMI and an indicator with a panel structure, the Senior Economy Tracker (Roch-Dupré et al., 2023), which measures the progress of 27 European economies in adapting to a demographic transition context over time (N=18 years), will be studied. Its panel data structure makes it ideal for the task at hand.

Once the analyses were performed, NRMSE values were obtained by incorporating cross-country information. The results show that imputing all the data together, incorporating panel information in the imputation, reduces the NRMSE. In fact, this behavior is also confirmed when imputing by areas or sub-areas, as the NRMSE always increases the less information is imputed together. The data shown incorporates panel information from the 27 European countries that have the indicator.

However, the analysis was also reproduced by disregarding the panel characteristic on the indicator, i.e., imputing the variables for each country separately and summing the NRMSE for comparability when the imputation is done across the panel. Comparing imputations with the two data structures, a consistent improvement is seen when the sample size is increased at all levels, both by including variables from different areas and different countries. As argued by Young and Johnson (2015), imputing with panel data increases the amount of information available for analysis, improving efficiency and reducing standard errors.

This is evident in the analysis, as the adjustment has a greater impact depending on the method used, with regression achieving the largest reduction in NRMSE. After all, although the original data set has 18 data points, imputation can only be applied to a complete data set. This reduces most of the variables to approximately 12 data points (since the time series must be complete to be valid for imputation), and 3 more must be removed to evaluate the NRMSE. Therefore, the amount of information available for the Senior Economy Tracker indicator studied is very low if imputation is not performed considering countries.

This same analysis was conducted on the SLMI, with the exception that the SLMI does not have data by countries, so the impact of the data structure can only be assessed by incorporating all variables versus imputing by areas or variables. The results are also consistent with those shown for the Senior Economy Tracker, obtaining an NRMSE of 0.9747 for the EM technique applied to the entire data set over the time series used in the work (Álvarez, 2022), which also compares very positively against the 2.564 obtained when constructing the SLMI. However, due to the robustness of the indicator, this has little impact on the final indicator.

b. Advanced weighting methods: FA vs. importance analysis

Factor analysis is a statistical technique used to reduce the dimensionality of data and find underlying structures in a set of observed variables. To assign weights to the variables, it follows several fundamental criteria that determine the influence of each variable on the extracted factors. These criteria are based on the correlation between variables, the explained variance, the factor rotation, the factor loading, and theoretical interpretation (OCDE, 2008). It is important to highlight that, in factor analysis, the goal is to reduce the number of variables while maintaining as much information as possible. Therefore, the explained variance becomes the key element when choosing the number of factors that will define this reality. In this sense, the factors are ordered according to the proportion of variance they explain from the original data set.

On the other hand, importance analysis is based on understanding the correlations between variables and evaluating the sensitivity of each one in relation to the model. Unlike factor analysis, which focuses on identifying underlying structures, importance analysis compares the first-order sensitivity index of each variable, considering both linear and nonlinear correlations. This allows adjusting the weights in a way that minimizes the deviation between the actual importance of each variable and its desired importance within the model. This approach is particularly useful to ensure that variables with a fundamental role in the interpretation of the model receive a weight corresponding to their actual impact.

In the context of composite indicators, although the aggregation formula is typically linear, correlations between variables can introduce nonlinearities in the relationship between y and x_i . To capture these dynamics, a more general measure called S_i , also known as the first-order sensitivity index, is employed. This approach measures the potentially nonlinear influence of each variable on the composite indicator and can be interpreted as the expected reduction in the variance of y when x_i is fixed (Becker et al., 2017).

Therefore, the sensitivity index would be defined as:

$$S_i = \frac{Var(E_{x_{\sim i}}[y|x_i])}{Var(y)}$$

Where $E_{x_{\sim i}}[y|x_i]$ is the expected value of y for x_i , assuming that the other variables are evaluated as constants. This can be modelled as a nonlinear regression of y on x_i , where the variance explained by the fitted curve is observed. Estimation of S_i can be found as:

$$\hat{S}_i = \frac{\sum_{j=1}^n (\hat{m}_{ij} - \bar{m}_i)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

Where \hat{m}_{ij} represents the adjusted value of the regression of y on x_i for observation j , and \bar{m}_i is the average of those adjusted values. The denominator simply shows the difference between the variable and its mean (Becker et al., 2017).

In summary, the first-order sensitivity index (S_i) provides an effective way to assess the importance of input variables in a composite indicator, considering both linear and nonlinear relationships. In fact, the literature (Hastie et al., 2008) already describes the main regression methods that can be applied to perform the nonlinear regression of y on x : Gaussian processes and penalized splines. The latter, which is based on a regression with components raised to a power, is defined as:

$$f(x_i) = \beta_0 + \beta_1 \cdot x_i + \sum_i^N \beta_i (x_i - \alpha_i)^k$$

Where β_0 and β_1 are the coefficients of the linear part and α_i represents the different knots, points that represent discontinuities in the regression model.

In the field of sensitivity analysis, it has been widely observed (Xu & Gertner, 2008) that, when input variables are correlated, S_i (also known as the first-order sensitivity index) includes both the direct effect of the variable x_i and the impact derived from its correlation with other variables. This is expressed as:

$$S_i = S_i^{uncorrelated} + S_i^{correlated}$$

Where $S_i^{uncorrelated}$ represents the uncorrelated and $S_i^{correlated}$ the correlated part of S_i .

It is crucial to decompose S_i , for which the approach of Xu and Gertner (2008), based on regression, is followed. To obtain the uncorrelated part $S_i^{uncorrelated}$, first, a regression of x_i on the other variables is performed, so that a multivariate regression is obtained to model x_i as a function of the other variables:

$$x_i = \beta_0 + \sum_{k \neq i} \beta_k x_k$$

Where coefficients β_k are obtained through ordinary least squares (OLS). In this way, the variance of x_i that can be explained by the other variables is defined. Thereafter, residuals \hat{z}_i shall be calculated as follows:

$$\hat{z}_i = x_i - \hat{x}_i = x_i - (\beta_0 + \sum_{k \neq i}^n \beta_k x_k)$$

When subtracting this estimate \hat{x}_i from the actual value of x_i , only the part of x_i that cannot be explained by the other variables is isolated. This removes the linear dependence between x_i and x_k , as all the variance of x_i explained by the other indicators has been subtracted from its component. Therefore, the residual \hat{z}_i represents the variability that only x_i can explain in y , which remains the synthetic indicator.

Finally, the regression of y on \hat{z}_i is performed to obtain $S_i^{uncorrelated}$ by applying the following formula:

$$S_i^{uncorrelated} = \frac{Var(E_{\hat{z}_i}[y|\hat{z}_i])}{Var(y)} = \frac{\sum_{j=1}^n (\hat{y}_j^i - \bar{y}^i)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

In this way, having calculated $S_i^{uncorrelated}$ and S_i , obtaining $S_i^{correlated}$ becomes straightforward if we recall the previously mentioned formula:

$$S_i = S_i^{uncorrelated} + S_i^{correlated}$$

Once the process of calculating importances is understood, a natural question that arises is how to adjust the weights of the variables to achieve a certain desired importance. This can be addressed using an optimization algorithm that minimizes the difference between the normalized sensitivity index (S_i) and the desired weight (w_i). The optimization problem can be formulated as follows:

$$w_{opt} = \underset{w}{\operatorname{argmin}} \sum_i^d (S_i^* - S_i(w))^2$$

Where S_i^* is the normalized objective sensitivity index, and $S_i(w)$ is calculated for each set of weights w as the first-order sensitivity index mentioned earlier (Becker et al., 2017). Note that, typically, the expert criterion assigns weights in an attempt to reflect the "importance" that the indicators are supposed to have, so this objective sensitivity index would actually represent the weights provided by the panel of researchers or experts.

As part of the importance analysis, it is essential to first understand the structure of the data and the potential existence of nonlinear correlations that may not be captured by other aggregation methods, such as Factor Analysis (FA).

To do this, linear and nonlinear regressions are first calculated using second-degree splines to capture such nonlinearities. These analyses clearly show how most of the variables exhibit reasonably linear behavior, with the exception of variable A.2, where the spline significantly deviates from the linear correlation line. While other variables show slight variations, they are not considered sufficiently notable.

Returning to the importance analysis, the first-order sensitivity index for all variables in the Unemployment category is presented below. This index reflects the linear and nonlinear

correlations concerning the studied area, quantifying the impact shown in the graphs presented earlier:

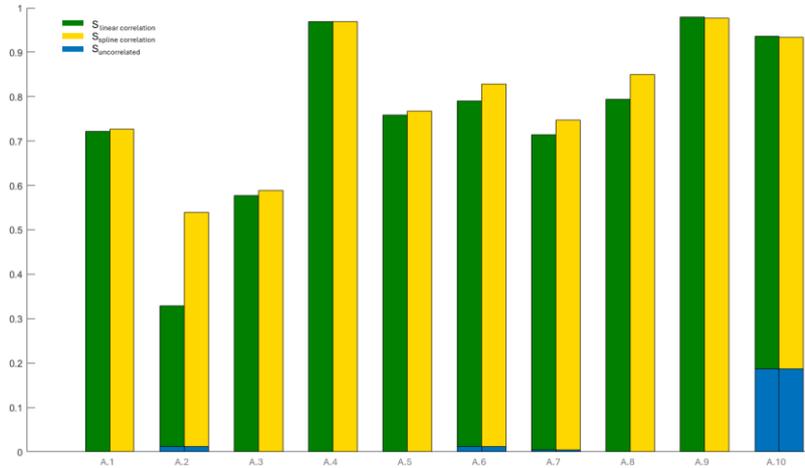


Figure 4. First-order sensitivity indices on category Unemployment

From the graph, it can be seen that the greatest variation between the linearly obtained importance and the spline-based importance occurs precisely in variable A.2. Given this notable variation, the penalized spline estimation will be considered for the "Unemployment" area from now on.

Secondly, it is essential to evaluate the existence of an uncorrelated importance, which is barely shown in the other variables (0.1898). Note that importance, as it is defined as the first-order sensitivity index of x (base indicator) on y (synthetic indicator), always has a correlated component of x_i explained by the other variables x_{-i} , and an uncorrelated component of the variability of y explained by x_i . This is based on the very concept of importance: a variable not only affects the overall outcome based on the weight it receives, but also through other indicators with which it shares behavior and, therefore, correlation.

Once the process through which the values in the illustration were obtained is understood, it can be deduced that variable A.10 can be explained by the others as significantly as the other variables but has a component of uncorrelation between x_i and y that does not appear in the other indicators. Therefore, according to the theory studied (Becker et al., 2017), the weight variation that can be made through optimization will have the greatest impact on this variable.

Next, the sensitivity index should be compared with the weight, not so much in absolute values but in relative terms to compare how important each variable is for the final area. To do this, the spline-based importance is rescaled to 1 and compared with the original weight of the SLM, the variance explained by each base indicator in the SLM, and also against the equal weighting with which the new "Unemployment" indicator has been constructed (figure A2 in annex 2).

Thus, certain deviations in the contributions of each variable to the final result of the "Unemployment" area become apparent. The greatest deviation occurs in variables A.2, A.3, A.4, A.9, and A.10. This problem likely arises because the correlated part of S_i prevails over the uncorrelated part in all the indicators. This implies that correlations, rather than weights, are what determine the importance, as previously mentioned.

It seems that, the higher the correlation, the FA seeks to assign a lower weight to compensate for variability already explained by other variables. However, regarding the impact on the importance of each variable, weight variation will only have a significant effect in cases where the uncorrelated importance has substance, since otherwise, all the importance would come from the linear regression between base indicators. Therefore, altering one of their weights will have a

marginal effect on the synthetic indicator if the impact is largely defined by the variability explained through the other variables.

Continuing with the analysis, the weights are then optimized to obtain the desired equalized importance.

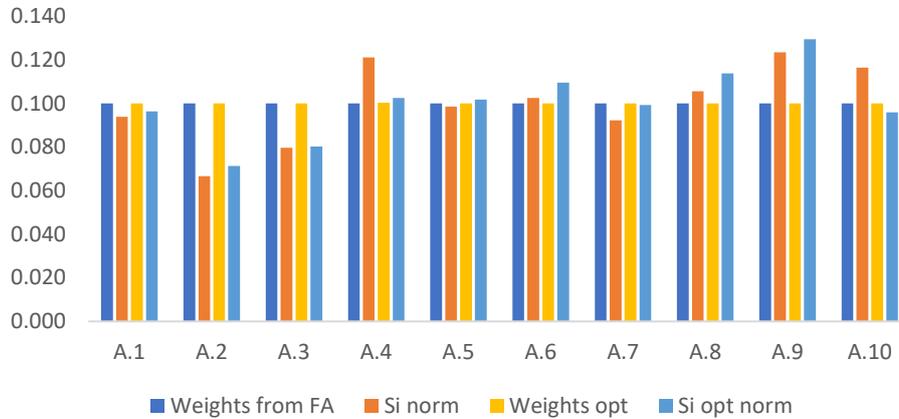


Figure 5. Results for the Unemployment category

From the illustration shown, it is worth noting that the impact of optimization on this category is small, as the weights vary slightly and manage to reduce deviations in variables A.2, A.3, A.4, A.7, and A.10. In fact, the most notable variation occurs in the latter variable, which is consistent with the previous explanation.

To conclude this analysis, it is important to mention that the sum of the squared deviations (objective function) between the desired and obtained importance has improved from 0.054 to the new value of 0.049.

Similarly, the Unemployment Protection category stands out, as it is clear that this mechanism would have been methodologically more robust than the simple equal weighting that was applied in the original construction of the SLMI. Of course, while the impact on the SLMI will be reduced both by the small deviation from the desired importance and because this area carries the least weight in the second aggregation level (10% compared to the 30% of the other areas), from a theoretical point of view, this allows for certain conclusions to be drawn.

Below, the importances for this area are shown, as well as the correlation matrix:

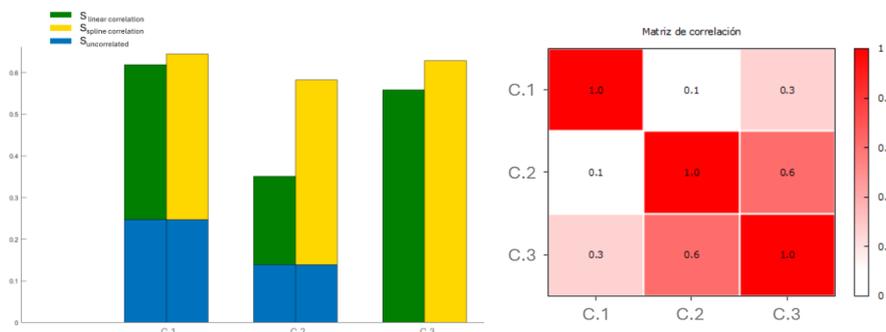


Figure 6. First-order sensitivity indices and correlation matrix of Protection for the unemployed

Note how, in this category, the uncorrelated importance has a much higher value in the first two variables, which is also confirmed by its correlation matrix, where a lower correlation between the first two variables is observed.

In this case, where FA has not been applied to the data set, it seems that the first two variables, less correlated both through the correlation matrix and the $S_i^{correlada}$ importance analysis, will be those where optimization can most notably improve the results. Notice how the values of $S_i^{correlada}$ and the correlation matrices, both in this category and the previous one, show consistent results. However, in this case, a greater component of $S_i^{incorrelada}$ is noted, which will improve the importance analysis since the assigned weights have a direct effect on the importance of each component indicator. Thus, the mentioned analysis is performed, and successfully, the desired importance is achieved exactly:

		Equal weighting	Si	Si norm	Weights opt	Si opt	Si opt norm
Protection for the unemployed - C	C.1	0.333	0.645	0.348	0.332	0.630	0.333
	C.2	0.333	0.582	0.314	0.392	0.630	0.333
	C.3	0.333	0.628	0.339	0.275	0.630	0.333

Table 3. Results of the first-order sensitivity analysis and its optimization for Protection for the unemployed

Additionally, a scenario was considered in which all variables are incorporated directly into the SLMI, bypassing the intermediate subaggregation step around the mentioned areas. Assuming that all base indicators are given the same weight in an equal-weighting context, each would be assigned 0.033, as the SLMI is composed of 30 base indicators. It seems clear that for categories with a high degree of mostly linear correlation, FA offers more advantages than importance analysis, as it can reduce dimensionality based on the explained variance (although this reduction is not applied in the SLMI) and provides weights when dealing with a large number of variables. The importance analysis is unable to provide optimal weights that notably alter the structure of the final synthetic indicator.

This can also be supported by the other aggregation performed in the SLMI using expert criteria: the aggregation of the 4 areas around the SLMI, where weights of 30% are assigned to *Unemployment*, *Employment*, and *Salaries and Labor Costs*, and 10% to *Protection for the Unemployed*.

Below is the table with the optimization of the areas relative to the SLMI:

	Weights	Si	Si norm	Weights opt	Si opt	Si opt norm
Unemployment – A	0.300	0.528	0.344	0.156	0.686	0.256
Employment – B	0.300	0.284	0.185	0.301	0.741	0.276
Protection for the unemployed – C	0.100	0.383	0.250	0.304	0.565	0.211
Salaries and labor costs – D	0.300	0.339	0.221	0.239	0.689	0.257

Table 4. Optimization results for the SLMI areas.

Except for the *Unemployment* area, whose distance from the desired importance remains unchanged, the rest of the categories show an optimized importance much closer to the weights assigned by the expert criterion. Thus, it seems that where the expert criterion prevails, the importance analysis is a highly valuable tool for weighting.

3. Conclusions

Regarding the proposed methodological improvements for imputation, working with the full dataset for imputation offers several key advantages. It allows capturing more comprehensive relationships and patterns that could be missed with a local approach, improving accuracy by leveraging global relationships between variables. Additionally, it reduces variability across areas, resulting in more stable outcomes less influenced by potential data shortages in some

categories. It also serves as a useful tool to avoid overfitting that can occur when imputing locally with limited data.

Regarding the weighting analysis, FA is useful for reducing variables in large data structures, especially when correlations are high, while the importance analysis is suitable when traceability of the data is needed or when expert criteria are applied. Although in the SLMi the number of indicators was not reduced using FA in a hierarchical structure, statistical weights were obtained, and an intermediate aggregation of specific areas was chosen. In contrast, the importance analysis is more useful for assigning relative weights to indicators when expert criteria are chosen due to dimensionality and utility. In fact, the importance analysis, as it depends on desired importances provided by a panel of researchers, suffers from the same applicability limitations as expert criteria.

In any case, the importance analysis has shown that the higher the $S_i^{incorrelada}$, the greater the impact of the variable's weight on the importance, and therefore, it becomes more feasible to apply changes or optimizations to better reflect the applied expert criteria. Similarly, when the number of analyzed variables is very large, $S_i^{correlada}$ dominates the importance, effectively discouraging the use of the importance analysis itself, as optimizing the weights does not significantly affect the obtained importance values. In any case, certain lines of research have emerged, including, among other things, the relationship between the FA method's weight assignment and the existing correlations between variables, the impact of variations in knots and the spline regression smoothing term, as well as the inclusion of more imputation techniques to evaluate the impact of data structures.

4. References

- Álvarez, G. (2022). *Desarrollo de un KPI sintético para medir la recuperación española frente al COVID-19*. Madrid: Universidad Pontificia Comillas.
- Becker, W., Saisana, M., Paruolo, P., & Vandecasteele. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological indicators*, 80, 12-22.
- Gold, M. S., & Bentler, P. M. (2000). Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(3):319–355.
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. Springer Series.
- OCDE. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. París: OECD Publications.
- Roch-Dupré, D., Aracil, E. & Calvo-Bascones, P. (2023). Monitoring socioeconomic readiness for the demographic transition: Introducing the Senior Economy Tracker . *Structural Change and Economic Dynamics*, Vol. 71, 430-443.
- Young, R., & Johnson, D. (2015). Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. *J Marriage Fam*, 77 (1) 277-294.

MEMORIA

Índice

1. Introducción.....	25
a. Estado de la cuestión	25
b. Motivación	29
c. Objetivos del proyecto.....	29
d. Metodología de trabajo.....	30
2. Marco conceptual del análisis	33
a. Introducción al SLMI	33
3. Análisis de imputación con diferentes estructuras de datos.....	38
a. Estado del arte.....	38
b. Resultados	41
4. Métodos avanzados de ponderación: FA vs análisis de importancia	46
a. Estado del arte.....	46
b. Resultados	52
4. Conclusiones	70
5. Referencias.....	72
Anexo 1. Alineación con los ODS	74
Anexo 2. Composición del SLMI.....	75
Anexo 3. Estructura del Senior Economy Tracker	77
Anexo 4. Código fuente	79

Índice de tablas

Tabla 1. Resultados del análisis de sensibilidad de primer orden y su optimización para la categoría C	10
Tabla 2. Resultados de la optimización para las áreas del SLMi	10
Table 3. Results of the first-order sensitivity analysis and its optimization for Protection for the unemployed	19
Table 4. Optimization results for the SLMi areas	19
Tabla 5. Resultados del análisis de imputación considerando todos los países europeos	42
Tabla 6. Resultados del análisis de imputación obviando la dimensionalidad regional	43
Tabla 7. Valores del índice de primera sensibilidad (“importancia”) en la categoría Desempleo	54
Tabla 8. Valores de importancia frente a los valores clave en la construcción de la categoría Desempleo	56
Tabla 9. Valores de importancia frente a los valores clave en la construcción de la categoría Empleo	59
Tabla 10. Valores de importancia frente a los valores clave en la construcción de la categoría Empleo	60
Tabla 11. Valores de importancia frente a los valores clave en la construcción de Protección a los desempleados.....	62
Tabla 12. Resultados del análisis de sensibilidad de primer orden y su optimización para la categoría C	62
Tabla 13. Valores de importancia frente a los valores clave en la construcción de Salarios y costes laborales	65
Tabla 14. Valores de importancia frente a los valores clave en la construcción de la categoría Salarios y costes laborales.....	66
Tabla 15. Resultados para todas las categorías.....	67
Tabla 16. Resultados derivados de aplicar la optimización sobre todas las variables juntas	68
Tabla 17. Resultados de la optimización para las áreas del SLMi	69

Índice de ilustraciones

Ilustración 1. Índices de sensibilidad de primer orden sobre la categoría Desempleo	8
Ilustración 2. Resultados para la categoría Desempleo	9
Ilustración 3. Índices de sensibilidad de primer orden y matriz de correlaciones de Protección a los desempleados.....	9
Figure 4. First-order sensitivity indices on category Unemployment.....	17
Figure 5. Results for the Unemployment category	18
Figure 6. First-order sensitivity indices and correlation matrix of Protection for the unemployed	18
Ilustración 7. Diagrama de Gantt del proyecto	32
Ilustración 8. Marco conceptual del SLMI (Monitor Deloitte, 2022).....	34
Ilustración 9. Estructura de panel del indicador estudiado en la imputación	41
Ilustración 10. Impacto de la imputación simultánea sobre el SLMI.....	45
Ilustración 11. Distribuciones de los indicadores base de Desempleo frente al SLMI y sus ajustes lineal y por spline	53
Ilustración 12. Índices de sensibilidad de primer orden sobre la categoría Desempleo	54
Ilustración 13. Matriz de correlaciones de la categoría Desempleo	56
Ilustración 14. Resultados de la optimización para la categoría Desempleo.....	57
Ilustración 15. Distribuciones de los indicadores base de Empleo frente al SLMI y sus ajustes lineal y por spline	58
Ilustración 16. Índices de sensibilidad de primer orden sobre la categoría Empleo	58
Ilustración 17. Matriz de correlaciones de la categoría Empleo	59
Ilustración 18. Resultados de la optimización para la categoría Empleo.....	60
Ilustración 19. Distribuciones de los indicadores base de Protección a los desempleados frente al SLMI	61
Ilustración 20. Índices de sensibilidad de primer orden sobre la categoría Protección a los desempleados.....	61
Ilustración 21. Matriz de correlaciones de la categoría de Protección a los desempleados.....	62
Ilustración 22. Resultados de la optimización para la categoría Protección a los desempleados.....	63
Ilustración 23. Distribuciones de los indicadores base de Salarios y costes laborales frente al SLMI	63
Ilustración 24. Índices de sensibilidad de primer orden sobre la categoría Salarios y costes laborales	64
Ilustración 25. Matriz de correlaciones de la categoría Salarios y costes laborales	65
Ilustración 22. Resultados de la optimización para la categoría Salarios y costes laborales	66
Ilustración 27. Resultados de la optimización para el SLMI equiponderado a partir de todos los indicadores base	67
Ilustración 28. Diagrama de caja y bigotes sobre las diferentes iteraciones de SLMI obtenidas ..	71

1. Introducción

En un entorno global cada vez más complejo y dinámico, la medición de fenómenos económicos y sociales se vuelve fundamental para orientar decisiones estratégicas y políticas. La creación y utilización de indicadores se ha convertido en una herramienta esencial para comprender la realidad en la que operan las organizaciones y los gobiernos. La importancia de los indicadores sintéticos es evidente en su capacidad para sintetizar información compleja y ofrecer una visión integral del estado de un sector o de un fenómeno particular. Indicadores famosos como el Índice de Desarrollo Humano (IDH) de la ONU y el Índice de Competitividad Global del Foro Económico Mundial sirven como ejemplos de cómo las métricas bien diseñadas pueden influir en la formulación de políticas y en la evaluación del progreso social y económico. Sin embargo, a pesar de su relevancia, muchos de estos indicadores enfrentan limitaciones significativas en términos de adaptabilidad y capacidad para captar los matices de la realidad contemporánea.

Este trabajo busca abordar las limitaciones metodológicas existentes, centrándose en la mejora de la construcción de indicadores sintéticos en etapas tan críticas como la imputación de datos, la asignación de pesos y la agregación. La imputación de datos ausentes, un aspecto crítico en la creación de indicadores precisos, es a menudo una fuente de sesgo que puede distorsionar la interpretación de los resultados. La capacidad de utilizar técnicas avanzadas, como la imputación múltiple, a diferentes bases de datos con dimensionalidades diferentes permitirá comprender de qué manera los datos pueden reflejar con mayor fidelidad las características de los fenómenos estudiados. Al mismo tiempo, la ponderación adecuada de cada variable involucrada en la construcción de un indicador y su agregación en el sintético final son pasos fundamentales para garantizar que la representación del fenómeno sea válida y relevante.

Mi motivación personal para realizar este trabajo proviene del deseo de contribuir al desarrollo de herramientas que permitan a las organizaciones y responsables de políticas evaluar de manera más precisa la realidad económica y social. De hecho, este trabajo se inscribe en un esfuerzo continuo por mejorar la construcción de indicadores sintéticos, partiendo de la experiencia adquirida en mi trabajo de fin de grado (TFG), donde desarrollé el Synthetic Labor Market Index (SLMI), un indicador diseñado para evaluar la salud del mercado laboral en un contexto económico en constante cambio con dos niveles de agregación. A partir de esta base, he ampliado mi investigación a través de un artículo académico que analiza la relación del SLMI con otros factores económicos en España, estableciendo vínculos con indicadores reconocidos como el Producto Interno Bruto (PIB), la Tasa de Desempleo y el Índice de Precios al Consumidor (IPC).

A través de este trabajo, espero no solo avanzar en el campo académico, sino también ofrecer soluciones prácticas que impacten en el ámbito real de la toma de decisiones. La importancia de esta investigación radica en su potencial para influir en la formulación de políticas y mejorar la calidad de la información que guía las decisiones estratégicas. Mi objetivo es impulsar la mejora metodológica en los indicadores estudiados y, a través de la replicabilidad de dichas mejoras en otros indicadores sintéticos, asegurar que aquellos usados a mayor escala puedan dotarse de una mayor robustez, de forma que sean realmente representativos y relevantes en un mundo en constante cambio.

a. Estado de la cuestión

La construcción de indicadores sintéticos es algo muy estudiado y contrastado. Al fin y al cabo, la estructuración y el orden en la presentación de la información siempre ha generado valor *per se*, lo cual se acentúa especialmente en el campo de los indicadores y los datos desagregados por lo mencionado en la introducción. Además, su presencia se extiende por multitud de sectores y entidades, tanto en la vida pública como en la privada.

No obstante, la complejidad de los mismos aumenta enormemente cuando se trata de describir el entorno en que se encuentran. Por ello, tanto empresas privadas como organismos públicos de todas las esferas han contribuido con el paso de los años a estandarizar los mecanismos que permiten la construcción de indicadores sintéticos con fiabilidad y trazabilidad. En este sentido, la OCDE (2008) ha preparado un manual que aspira a establecer un marco común en el proceso de creación de indicadores sintéticos. Así, no solo aquellos deseosos de elaborar uno tendrán a su alcance los pasos bien detallados, sino que los interesados en comprender la lógica subyacente podrán hacerse con un manual que les ahorre mucho tiempo y quebraderos de cabeza.

No obstante, si bien dicho documento se considera un buen punto de comienzo para comprender los fundamentos, la elaboración de un indicador completo requiere de una mayor profundización en cada una de las etapas descritas en el propio manual. Por ello, todo buen marco teórico que se precie debe abordar en detalle cada una de las diferentes etapas de la construcción del indicador. Estas son, en orden, la recolección y selección de datos, el análisis multivariante, la imputación de datos ausentes, la normalización, la ponderación y, finalmente, la agregación... Por supuesto, esto debe comenzar por un sólido marco teórico que justifique las decisiones tomadas y debe finalizarse con un análisis de robustez y sensibilidad que evalúe el impacto de dichas decisiones del autor a lo largo de la construcción del indicador (Saisana et al., 2005). No obstante, en interés del presente estado de la cuestión, en primer lugar habrá de comprenderse bien la definición de indicador sintético y el porqué de su importancia.

Los indicadores sintéticos, definidos por autores como Freudenberg (2003) como aquellos construidos a partir de múltiples indicadores individuales (de ahora en adelante, *indicadores base*), son útiles para medir variables multidimensionales difíciles de acotar con un solo indicador. La mayoría de autores coinciden en la agregación de indicadores simples como elemento vertebrador de cualquier sintético, aunque para algunos resulta indispensable la existencia de un sistema complejo que, a partir de un modelo, permita comprender una realidad multidimensional (Nardo, et al., 2005). Por ejemplo, ¿cómo se podría medir el bienestar? A partir de indicadores base sobre la capacidad económica, los servicios a los que se tiene acceso, el nivel de educación, la salud... se podría atinar a dar un número que, por sí solo, no significaría nada concreto; pero que serviría para establecer comparativas con otros países y consigo mismo.

Los indicadores sintéticos gozan de una creciente importancia en el mundo académico y profesional, ya que aportan cierta monitorización a aspectos típicamente cualitativos y generalmente complejos. Su incipiente uso está documentado por numerosos autores, especialmente en la esfera pública y mediática (Saltelli, 2007). Sin embargo, pese a que el uso de estos medidores está muy extendido – como el Índice de Desarrollo Humano (IDH) –, también ha habido mucha discusión académica al respecto. Los detractores los critican por carecer de significado real y tangible, en tanto que muchas veces el foco del sintético radica en reflejar la variabilidad estadística subyacente en lugar de representar de forma fidedigna las interacciones de las realidades que agrupa. Además, también es muy criticado por las arbitrariedades presentes en su elaboración a través de la toma de decisiones de los autores (Grupp & Moge, 2004). De hecho, también suponen un riesgo notable para las instituciones allí donde son construidos negligentemente (Cherchye et al., 2007).

En cambio, los partidarios destacan cómo los indicadores permiten poner el foco del debate en ciertas problemáticas sociales, a la par que facilitan la comprensión de realidades complejas. De hecho, estos argumentos lograron que Amartya San, Premio Nobel de Economía en 1998, cambiara su postura al respecto. El economista comenzó a apoyarlos debido al impacto social propiciado por el debate en torno al IDH, tanto por atraer la atención pública como por dinamizar la toma de decisiones de entidades públicas y privadas (Greco et al., 2019). Así pues, los indicadores sintéticos también son especialmente útiles al reducir un gran número de medidores iniciales, lo cual además los convierte en una herramienta especialmente didáctica para aquellos

poco versados en el entorno observado por el indicador. Al fin y al cabo, resulta más sencillo comparar los mercados productivos de dos países a partir de un indicador sintético *Producción* que con los medidores IPRIX, IPRIM, IPRI... los cuales resultan complejos fuera del entorno industrial. Sin embargo, Para evitarlo, el proceso de elaboración ha de ser lo más transparente y justificado posible, de forma que el análisis del entorno proporcionado por el indicador no induzca a errores o manipulaciones. A tal efecto, la trazabilidad de sus medidores componentes es esencial.

Para que esta construcción sea adecuada a los máximos estándares propuestos por la OCDE (2008) y el mundo académico, cabe resaltar la importancia de los pasos que se estudiarán desde el punto de vista estadístico y metodológico en la construcción de los indicadores. En primer lugar, la imputación de información ausente que busca tratar un conjunto de datos completo donde, dependiendo de la naturaleza MCAR ("Missing completely at random"), MAR ("Missing at random") o NMAR ("Not missing at random") de los datos, debe valorarse un conjunto diferente de métodos. Para los casos MCAR, la imputación múltiple es descrita como típicamente la más robusta (OCDE, 2008), aunque debe estudiarse la aplicación de técnicas como la imputación simple EM que dispone de buena reputación en el mundo académico (Gold y Bentler, 2000). En este sentido, el TFG sobre el que se basa este TFM empleó un mecanismo empírico basado en el RMSE (error cuadrático medio) normalizado para dilucidar cuál de los métodos estudiados resulta más apto para el conjunto de datos estudiado.

En este sentido, cabe resaltar que, durante la realización del TFG, se halló una divergencia en cuanto a qué técnica era más apta para la imputación de datos según la estructura estudiada. Por ejemplo, emplear el método EM resultaba idóneo si solo se consideraban los indicadores base de cada una de las áreas del mercado laboral por separado, mientras que la imputación múltiple salía beneficiada si se consideraba el conjunto total de datos sin atender a dicho primer nivel de agregación. A esta capa de discrepancia, se debe añadir la consideración de los datos de panel, que incorporarían una dimensión adicional y, por tanto, aún mayores desavenencias entre los resultados proporcionados por las técnicas de imputación. Por ello, este TFM pretende realizar un acercamiento metodológico que arroje luz sobre qué técnica resulta más apta dependiendo de la estructura de datos estudiada.

Retomando la construcción de indicadores, otro de los aspectos más críticos es la asignación de pesos a los diferentes componentes. La ponderación determina la influencia que cada indicador tendrá en el resultado final, lo que puede alterar las interpretaciones y decisiones basadas en estos índices. Existen diversas técnicas para la asignación de pesos, y muchas de ellas se basan en la importancia "aparente" de cada indicador. Por ejemplo, "participatory methods" (Lee y Chan, 2008) como el BAP (*Budget Allocation Process*) o AHP (*Analytic Hierarchy Process*) suelen asignar pesos de manera subjetiva (Nardo *et al.*, 2005), basándose en criterio experto. Si bien en ciertas ocasiones estos pueden no reflejar adecuadamente las relaciones subyacentes entre los indicadores, su aplicación resulta idónea bajo ciertas premisas como una baja correlación. No obstante, en tanto que no valora la correlación subyacente entre las variables, puede sobreponderar determinadas realidades manifestadas por indicadores base de índole similar. En resumen, esta práctica puede llevar a una representación distorsionada de la importancia real de los componentes en el indicador sintético, en caso de que no se realice adecuadamente.

Para paliar esto, técnicas como el Principal Components Analysis (PCA) o el Factor Analysis (FA) realizan un análisis de la variabilidad explicada de cada uno de los indicadores base y los agrupan en torno a factores comunes ortogonales entre sí (Greco *et al.*, 2019). De esta manera, se optimiza la variabilidad explicada del conjunto de datos mientras se minimiza la cantidad de variables a estudiar. Si bien el enfoque es útil y puede resolver grandes problemas de dimensionalidad y correlación, también entorpece la trazabilidad del sintético final respecto de los indicadores base que lo componen.

Por otro lado, los trabajos de Becker et al. (2017) han abordado la brecha existente en la asignación de pesos a través del concepto de “importancia” en los indicadores compuestos. Este enfoque se centra en el cálculo de la importancia de cada componente mediante un análisis estadístico riguroso. En lugar de depender de técnicas que ignoren las correlaciones entre componentes, Becker y sus colaboradores proponen una metodología que asigna pesos estadísticos basados en la importancia final de los indicadores. Utilizando técnicas como la regresión penalizada, es posible identificar cómo cada indicador contribuye al resultado final, teniendo en cuenta no solo su variabilidad individual, sino también cómo se relaciona con otros indicadores.

El análisis de correlación juega un papel crucial en esta metodología, permitiendo descomponer la influencia de cada componente en el índice compuesto. Esta descomposición revela qué indicadores tienen un impacto significativo en el resultado y cuáles pueden ser redundantes o tener un efecto negativo. Los resultados de este análisis se pueden expresar en términos de ratios de correlación, que indican la relación entre los indicadores y el índice global. Esto permite combinar adecuadamente la subjetividad entendida como “importancia” que suele ser asignada por la técnica BAP con el análisis estadístico y robusto del PCA, que como se menciona anteriormente atiende a la variabilidad explicada de cada indicador base sobre el conjunto total. Por ello, este TFM propone profundizar en el concepto de la “importancia” y analizar el equilibrio entre el análisis subjetivo de las realidades subyacentes a los indicadores sintéticos a la par que proporcionar el enfoque estadístico de los métodos más matemáticos. De esta manera, se puede dotar a la comunidad de un marco metodológico que arroje luz sobre la interacción entre peso e “importancia”, cristalizando en una técnica de ponderación que permita al autor del indicador sintético encontrar una solución intermedia a este problema.

En el caso de la agregación, resulta capital comprender los dos principales métodos y su efecto sobre la sustituibilidad de los indicadores base, entendida como la capacidad de un indicador de reemplazar con buen desempeño el pobre rendimiento de otro sobre el resultado final (OCDE, 2008). En primer lugar, se dispone de la agregación aritmética, en la que los indicadores se suman ponderados según el peso asignado anteriormente, de manera que un buen rendimiento de uno puede compensar totalmente el pobre desempeño de otro. En contraste, la agregación geométrica multiplica cada indicador elevado al peso mencionado, de manera que el sintético final se ve muy penalizado cuando alguno de los indicadores base toma valores cercanos a 0. De esta manera, la sustituibilidad en la agregación geométrica solo se da forma parcial.

Para solucionar esto, recientemente nuevos autores (Biggeri et al., 2019) proponen métodos intermedios o “híbridos”. Este TFM busca profundizar en este nuevo abanico de técnicas y estudiar el impacto de su inclusión sobre los indicadores estudiados a través de estudios de robustez y sensibilidad sobre el resultado final del mismo.

En conclusión, la mejora metodológica en la construcción de indicadores compuestos, especialmente en las etapas de imputación, ponderación y agregación, es esencial para garantizar que estos índices representen fielmente la realidad compleja de los fenómenos que buscan medir. Al aplicar técnicas metodológicas para discernir la importancia de los indicadores base frente a su peso y al proporcionar directrices claras sobre cómo aplicar estas técnicas, se busca no solo elevar la calidad de los indicadores, sino también promover su uso efectivo en la toma de decisiones informadas. Este enfoque tiene el potencial de transformar la manera en que se interpretan y utilizan los indicadores compuestos en contextos políticos, económicos y sociales, asegurando que sean herramientas útiles y robustas para la evaluación de nuestro progreso como sociedad.

b. Motivación

La motivación detrás de este trabajo de fin de máster (TFM) radica en la necesidad de avanzar en la construcción de indicadores compuestos que reflejen de manera precisa y significativa realidades complejas, como la situación del mercado laboral. A lo largo de mi investigación previa, en la que desarrollé el Synthetic Labor Market Index (SLMI), he podido observar la importancia de una metodología robusta para la construcción del indicador. La experiencia adquirida en el desarrollo del mismo me ha llevado a reconocer que la imputación, asignación de pesos y la agregación de los mismos son etapas críticas que pueden influir significativamente en la validez y utilidad del indicador. Si bien la literatura sobre la construcción de indicadores compuestos está bien desarrollada y ofrece técnicas que pueden aplicarse en diversos contextos, cada indicador presenta particularidades que requieren un enfoque individualizado y fundamentado en el juicio del investigador. Al establecer un enfoque que contemple métodos estadísticos más precisos y una comprensión más profunda de la importancia relativa de las tres etapas de construcción consideradas, este trabajo permitirá mejorar la calidad y la interpretabilidad de los indicadores sintéticos en el ámbito del mercado laboral y más allá.

Mi interés por este tema surge de una profunda curiosidad por entender las dinámicas complejas que rigen el comportamiento del mercado laboral y su relación con el bienestar social. Me apasiona la idea de utilizar herramientas cuantitativas para resolver problemas reales y ofrecer soluciones que impacten positivamente en la sociedad. Además, creo firmemente que la investigación en esta área puede contribuir a un desarrollo más sostenible y equitativo. Este TFM representa una oportunidad para profundizar en un campo que me motiva y que considero esencial para abordar los desafíos contemporáneos.

De hecho, profundizar en el impacto social del trabajo realizado siempre me ha resultado esencial para dotar de propósito a mi análisis. En este sentido, un hecho que da fe de ello es la investigación sobre la relación entre el SLMI y la ley de Okun, plasmado en un artículo académico derivado del TFG donde se construye el SLMI. Esta investigación me ha proporcionado información muy interesante sobre cómo los cambios en el desempleo pueden influir en la producción económica, estableciendo conexiones significativas que pueden ser útiles de cara a futuros trabajos y líneas de investigación. Este impacto me ha motivado a seguir explorando esta línea de investigación de los indicadores sintéticos, con el objetivo de desarrollar indicadores que no solo sean estadísticamente válidos, sino que también ofrezcan soluciones prácticas a problemas actuales.

Así, esta investigación no solo busca fortalecer la metodología aplicada en el SLMI, sino que también me permitirá adquirir habilidades valiosas en la construcción y análisis de indicadores. Estoy convencido de que el trabajo que realice en este TFM será un paso crucial para seguir avanzando en esta línea de investigación, contribuyendo a la creación de indicadores que sean tanto precisos como útiles para la toma de decisiones en el ámbito político y económico. En definitiva, este proyecto representa no solo un reto académico, sino también una oportunidad de crecimiento personal y profesional en un campo que me apasiona profundamente.

c. Objetivos del proyecto

El principal objetivo de este trabajo será el análisis metodológico de propuestas orientadas a dotar a los indicadores sintéticos de mayor robustez en las etapas de imputación, ponderación y agregación. Para lograr este objetivo principal, el proyecto dispondrá de los siguientes objetivos de segundo orden:

1. *Revisión metodológica de las etapas críticas en la construcción de los indicadores sintéticos: imputación y ponderación.*

Resulta clave establecer un marco teórico sólido que permita construir y profundizar en esta labor de mejora metodológica propuesta en el TFM. Para ello se deberá analizar la literatura en profundidad para comprender en mayor profundidad las técnicas empleadas en el TFG y aquellas que, ya sea por su complejidad y no aplicabilidad al indicador estudiado, fueron descartadas. En conclusión, este objetivo permitirá sentar las bases de un análisis metodológico bien documentado.

2. *Análisis metodológico de las técnicas de imputación de información ausente según la estructura del conjunto de datos sobre la que se aplica*

Este objetivo expandirá el alcance y la comprensión de las técnicas de imputación empleadas en el TFG, tales como el método de imputación múltiple, EM o regresión lineal. Al aplicar dichas técnicas a conjuntos de datos diferentes, se busca dotar a la comunidad científica de un marco más exhaustivo y comprehensivo en la utilización de los métodos de imputación de información ausente. Por supuesto, esto se deberá ver dotado de un marco teórico detallado en la labor de mejora metodológica propuesta.

3. *Evaluación metodológica de la fase de ponderación con un análisis de la "importancia" proporcionada por el criterio experto y la asignación basada en la variabilidad estadística del método FA*

Se desarrollará un enfoque que integre la importancia real de cada variable en función de su impacto en el indicador final, evitando el uso de técnicas tradicionales que pueden no reflejar adecuadamente las correlaciones subyacentes. Se aplicarán métodos estadísticos, como la regresión penalizada en conjunción con el análisis de componentes principales, para determinar la asignación óptima de pesos a cada variable. Además, se creará un marco que permita la actualización de los pesos a medida que se disponga de nueva información o se produzcan cambios en el entorno del mercado laboral.

4. *Evaluación final del indicador en su replicabilidad en otros indicadores y divulgación de resultados*

Por último, se analizará el potencial impacto del indicador a través de una evaluación de su replicabilidad en otros indicadores relevantes de diferentes campos, como la salud pública, la educación y el desarrollo económico, fomentando un enfoque interdisciplinario y enriqueciendo el ámbito de la investigación, así como la interacción entre las técnicas empleadas. Además, se enfatizará la divulgación de este trabajo a través del correspondiente TFM y se explorarán otras vías de impacto social junto a futuras líneas de investigación.

d. Metodología de trabajo

En línea con lo anterior, la metodología del trabajo busca asentar una serie de tareas ordenadas cronológicamente que permita la consecución de los objetivos descritos. Dichas tareas son las siguientes:

Objetivo 1: *Revisión metodológica de las etapas críticas en la construcción de los indicadores sintéticos: imputación y ponderación:*

- *Análisis general de la bibliografía y elaboración del marco teórico general (actividad 1.1):* Para comprender el fenómeno e identificar exactamente el vacío que se pretende llenar con el análisis, un estudio general sobre los recientes desarrollos en el mundo de los indicadores

permitirá partir de unas bases actualizadas y debidamente documentadas. Este marco asegurará que todas las mejoras propuestas en la metodología estén justificadas teóricamente y tengan el potencial de aportar valor al índice sintético final.

- *Elección de las técnicas e indicadores a estudiar (tarea 1.2)*: Una vez la literatura ha sido debidamente estudiada, se deberá determinar qué propuestas se someterán a análisis y qué estructuras de datos se evaluarán a través de dichas técnicas. Esto permitirá ampliar los contextos en los que dichas técnicas pueden ser metodológicamente útiles en la construcción de indicadores.
- *Recolección y estructuración de datos en MATLAB para todos los indicadores a considerar (tarea 1.3)*: En paralelo a la revisión de literatura y una vez se ha determinado qué indicadores se estudiarán, los indicadores a estudiar deben ser recolectados y tratados en MATLAB para su análisis posterior. Este paso garantiza que todos los datos relevantes (SLMI y otros indicadores a estudiar) estén organizados de forma coherente, permitiendo que los análisis posteriores se apliquen sin obstáculos técnicos y asegurando la consistencia de los datos durante las fases de imputación y ponderación. MATLAB no solo permite el procesamiento eficiente de los datos, sino que también proporciona herramientas avanzadas para facilitar cualquier labor posterior de análisis.

Objetivo 2: *Análisis metodológico de las técnicas de imputación de información ausente según la estructura del conjunto de datos sobre la que se aplica:*

- *Revisión de literatura sobre los métodos de imputación y el impacto de diferentes estructuras de datos (tarea 2.1)*: En este caso resulta crítico comprender cómo diferentes conjuntos de datos (series temporales, paneles...) afectan a los resultados de la imputación, por lo que una revisión exhaustiva de la bibliografía disponible es capital para asegurar que todo análisis posterior está debidamente documentado. De igual modo, esta tarea contribuye a profundizar en el marco teórico del trabajo.

Análisis metodológico de los diferentes métodos de imputación sobre los resultados en MATLAB e IBM SPSS (tarea 2.2): En este caso, varios métodos de imputación (EM, múltiple, regresión lineal) se compararán sobre diferentes bases de datos de estructuras distintas con el objetivo de medir cuál puede resultar más aplicable o robusta según ciertas premisas. La evaluación de resultados garantiza que las propuestas metodológicas mejoran la robustez del índice.

Objetivo 3: *Evaluación metodológica de la fase de ponderación con un análisis de la "importancia" proporcionada por el criterio experto y la asignación basada en la variabilidad estadística del método FA:*

- *Revisión de literatura sobre la ponderación y sus potenciales mejoras metodológicas (tarea 3.1)*: En esta tarea se incluye la profundización en el mundo académico de la ponderación, para así asentar bajo sólidos pilares teóricos cualquier propuesta metodológica. Esta línea de trabajo contribuye a documentar un marco teórico que, posteriormente, será debidamente reflejado en la redacción del proyecto.
- *Análisis metodológico de la ponderación en MATLAB y evaluación de resultados (tarea 3.2)*: Aplicación y evaluación de las técnicas metodológicas propuestas en la fase de asignación de pesos. A través de este análisis se debe medir el impacto de la significancia y el impacto de correlaciones "escondidas" sobre la robustez y sensibilidad de los indicadores sintéticos estudiados.

Objetivo 4: Evaluación final del indicador y divulgación de resultados:

- Documento de seguimiento del proyecto: Anexo B (tarea 4.1): Esta etapa permite establecer el marco sobre el que se asientan los objetivos y el estado de la cuestión del proyecto en cuestión.
- Evaluación final de resultados (tarea 4.2): Esta etapa permite evaluar la consecución de los objetivos del proyecto en base a los resultados obtenidos. De igual modo, busca alinear dichos resultados con el marco teórico descrito originalmente y analizar el conjunto propuesto de manera holística.
- Redacción final del proyecto (tarea 4.3): En esta última etapa se detalla el análisis exhaustivo de las potenciales mejoras metodológicas introducidas así como el marco teórico sobre el que se asienta. De igual modo, se incluyen las decisiones clave y los hallazgos de cada etapa del proyecto, ofreciendo una descripción completa del trabajo realizado y proporcionando a otros investigadores una guía para la replicación del proceso en diferentes contextos.

El diagrama de Gantt que resume los pasos y actividades anteriormente descritas se encuentra a continuación:

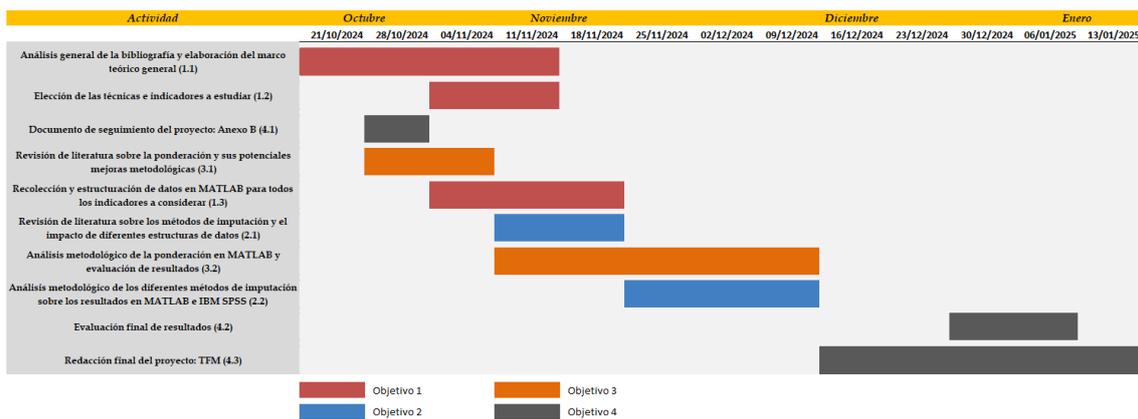


Ilustración 7. Diagrama de Gantt del proyecto

2. Marco conceptual del análisis

a. Introducción al SLMI

El Índice Sintético del Mercado Laboral (SLMI, por sus siglas en inglés) es una herramienta diseñada para evaluar el desempeño del mercado laboral desde una perspectiva integral, tomando en cuenta cuatro categorías fundamentales: Desempleo, Empleo, Protección a los Desempleados y Salarios. Estas categorías cubren aspectos clave del mercado laboral y proporcionan una visión holística sobre su funcionamiento y evolución en el tiempo. En el Apéndice del estudio se proporcionan los detalles específicos de los indicadores base utilizados en cada una de estas categorías, incluyendo tanto sus nombres como los códigos alfanuméricos asociados, así como las fuentes de datos empleadas para su recopilación.

El proceso de selección de los indicadores a incluir en el índice se basa en varios criterios rigurosos que garantizan la calidad y la utilidad de los datos. Uno de los criterios primordiales es la fiabilidad, ya que es esencial contar con datos confiables provenientes de fuentes públicas y reconocidas por su alta calidad. En este sentido, se priorizan las fuentes de datos oficiales como el Instituto Nacional de Estadística (INE) y el Servicio Público de Empleo Estatal (SEPE), instituciones que proporcionan información consistente y verificada, lo que permite asegurar la solidez de los resultados obtenidos a partir de los indicadores.

Otro criterio clave en la selección de los indicadores es la disponibilidad de datos, específicamente en lo que respecta al período comprendido entre 2008 y 2022. Este rango temporal permite capturar las variaciones y tendencias a largo plazo en el mercado laboral, abarcando un período suficiente para observar el impacto de diversos factores económicos y políticos. Además, se ha priorizado la inclusión de indicadores que se actualicen con regularidad, en particular de manera mensual. Las actualizaciones mensuales son cruciales para un análisis detallado de las tendencias y fluctuaciones en el mercado laboral, lo que permite identificar patrones más precisos y actualizados. Aunque se consideran datos trimestrales o anuales en algunos casos, se opta por no utilizar estos datos de manera frecuente debido a que su interpolación puede reducir la precisión de los resultados. Sin embargo, se incluyen dos variables clave, las tasas de empleo y desempleo trimestrales, dado su relevancia para el seguimiento continuo del mercado laboral.

Finalmente, la relevancia de los indicadores es un aspecto esencial para asegurar que cada uno de ellos se asigne correctamente a su categoría correspondiente dentro del índice. Esto garantiza que todos los indicadores estén alineados con el marco conceptual del estudio, lo que a su vez asegura que el SLMI brinde una representación precisa y significativa del mercado laboral en su conjunto. La correcta asignación de los indicadores a las categorías de Desempleo, Empleo, Protección a los Desempleados y Salarios es fundamental para que el índice sea capaz de reflejar de manera adecuada la complejidad y las dinámicas del mercado laboral, proporcionando una herramienta de análisis robusta y útil para la toma de decisiones.

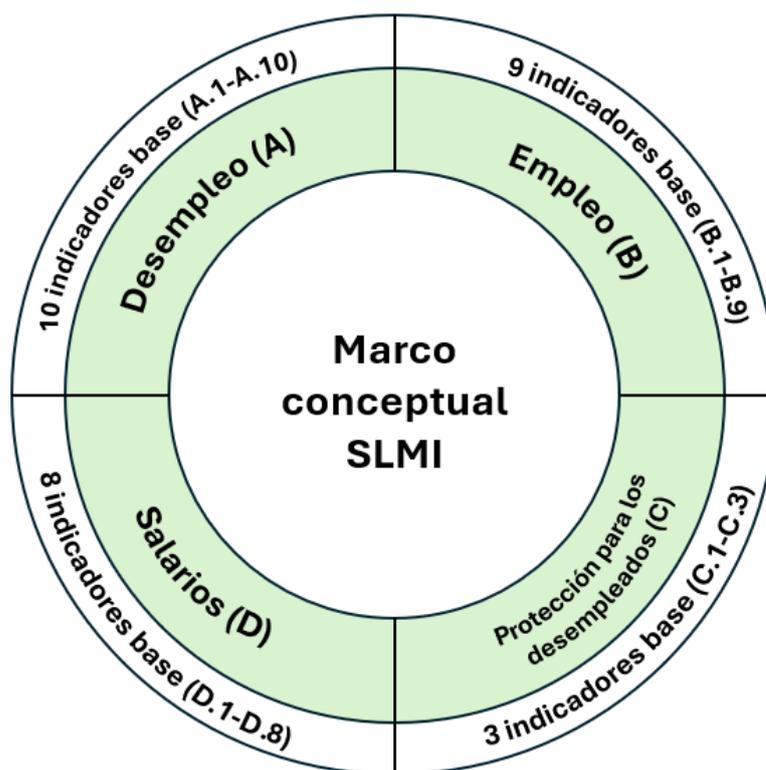


Ilustración 8. Marco conceptual del SLMI (Monitor Deloitte, 2022)

A continuación se explican en detalle de qué está conformada cada una de las categorías:

- **Categoría Desempleo:** Incluye 10 indicadores base, entre las cuales se encuentran:
 - **Desempleo registrado menores de 25 años (A.1):** Esta categoría hace referencia al desempleo registrado entre los jóvenes, un grupo que suele enfrentar mayores dificultades para ingresar al mercado laboral.
 - **Desempleo registrado en agricultura y pesca (A.2):** Involucra el desempleo registrado en sectores primarios como la agricultura y la pesca, que son fundamentales en ciertas economías, especialmente en áreas rurales.
 - **Desempleo registrado en construcción (A.3):** Refleja el desempleo registrado en la industria de la construcción, que suele ser cíclico y vulnerable a fluctuaciones económicas.
 - **Desempleo registrado en colectivo de difícil inserción (A.4):** Incluye a personas que enfrentan mayores barreras para acceder al empleo registrado debido a factores como discapacidad, larga duración de desempleo o problemas de formación.
 - **Desempleo registrado en industria (A.5):** Abarca el desempleo registrado dentro del sector industrial, que puede estar vinculado a cambios en la demanda de productos manufacturados o a la automatización.
 - **Desempleo registrado - Mujeres (A.6):** Considera específicamente el desempleo registrado entre las mujeres, quienes en muchos casos enfrentan mayores obstáculos laborales debido a la discriminación de género o la carga de trabajo doméstico no remunerado.
 - **Desempleo registrado en servicios (A.7):** Refleja el desempleo registrado en el sector servicios, que incluye una amplia gama de trabajos en áreas como comercio, educación, salud y hostelería.
 - **Desempleo registrado en personas sin empleo anterior (A.8):** Se refiere a aquellos individuos que no han tenido empleo registrado previamente y ahora buscan su primer trabajo.

- **Desempleo registrado en hombres (A.9):** Indica el desempleo registrado entre los hombres, un grupo en el que también se pueden observar disparidades dependiendo del sector o la región.
- **Desempleo registrado en tasa de desempleo trimestral (A.10):** Mide el desempleo registrado de forma puntual durante un trimestre específico, lo que permite evaluar las fluctuaciones en el corto plazo.
- **Categoría Empleo:** Incluye 9 indicadores sobre productividad laboral y tipos de contratos mediante:
 - **Contratos registrados (B.1):** Refleja la cantidad total de contratos registrados, independientemente de su tipo o duración.
 - **Contratos registrados a tiempo parcial (B.2):** Involucra los contratos registrados de tipo a tiempo parcial, es decir, aquellos con una jornada laboral inferior a la estándar.
 - **Contratos registrados CVE (B.3):** Hace referencia a los contratos registrados bajo el código CVE (que podría representar un tipo específico de contrato o sector).
 - **Contratos registrados indefinidos (B.4):** Se refiere a los contratos registrados que no tienen una fecha de finalización establecida, es decir, contratos indefinidos.
 - **Contratos registrados interinos (B.5):** Incluye los contratos registrados de tipo temporal o interino, generalmente vinculados a la sustitución de personal o trabajos por un tiempo determinado.
 - **Contratos registrados temporales (B.6):** Hace referencia a los contratos registrados de tipo temporal, con una duración específica y limitada en el tiempo.
 - **Productividad por hora efectivamente trabajada (B.7):** Mide la productividad registrada por hora trabajada efectivamente, reflejando la eficiencia laboral en función del tiempo de trabajo.
 - **Productividad por empleo a tiempo completo¹ (B.8):** Mide la productividad registrada por empleo a tiempo completo, que refleja la eficiencia laboral por trabajo completo.
 - **Tasa de empleo trimestral registrada (B.9):** Mide la tasa de empleo registrada de forma puntual durante un trimestre específico, lo que permite evaluar las fluctuaciones en el corto plazo.
- **Protección a los Desempleados:** Mide los incentivos públicos para transformar el estatus inactivo en activo mediante 3 indicadores base que reflejan el número de beneficiarios por categoría de subsidio:
 - **Beneficiarios de bienestar (C.1)**
 - **Beneficios contributivos por desempleo parcial (C.2)**
 - **Beneficios contributivos por desempleo total (C.3)**
- **Salarios:** Incluyen 8 indicadores sectoriales para captar el desempeño del mercado laboral y la evolución de la principal fuente de ingresos de los hogares. Las variaciones salariales incluyen las negociadas colectivamente (D.1) y los cambios salariales reales

(D.2-D.8), influenciados por condiciones de mercado, desempeño individual y factores económicos.

- **Incremento Salarial Pactado en Negociación Colectiva (D.1):** Este indicador mide el porcentaje de aumento salarial acordado entre los sindicatos y los empleadores a través de la negociación colectiva en un periodo determinado. Refleja los acuerdos alcanzados en los convenios colectivos sobre salarios, lo cual es fundamental para entender cómo se ajustan los salarios a nivel sectorial y regional.
- **Variación Interanual de los Rendimientos del Trabajo - Agricultura y Ganadería (D.2):** Este indicador mide el cambio porcentual en los rendimientos laborales (salarios, ingresos, etc.) en el sector agrícola y ganadero de un año a otro. Proporciona información sobre cómo ha evolucionado la remuneración de los trabajadores en este sector específico, el cual puede verse afectado por diversos factores como cambios en las políticas agrícolas, condiciones climáticas y la demanda de productos agrícolas.
- **Variación Interanual de los Rendimientos del Trabajo - Comercio (D.3):** Este indicador evalúa el cambio porcentual en los ingresos laborales de los trabajadores del sector comercio de un año a otro. Ayuda a medir el comportamiento del mercado laboral en términos de salarios en actividades comerciales, que incluyen tanto el comercio mayorista como minorista.
- **Variación Interanual de los Rendimientos del Trabajo - Construcción (D.4):** Este indicador refleja cómo han variado los salarios o rendimientos laborales de los trabajadores en el sector de la construcción en comparación con el año anterior. Los rendimientos del trabajo en construcción están influenciados por factores como la demanda de vivienda, infraestructuras y las políticas gubernamentales relacionadas con la construcción.
- **Variación Interanual de los Rendimientos del Trabajo - Hostelería (D.5):** Este indicador mide el cambio porcentual en los salarios de los trabajadores en la industria de la hostelería (restaurantes, bares, hoteles, etc.) en el transcurso de un año. Este sector es crucial para la economía, especialmente en áreas turísticas, y puede estar sujeto a variaciones estacionales y cambios en la demanda de servicios.
- **Variación Interanual de los Rendimientos del Trabajo - Industria (D.6):** Se mide el cambio porcentual en los rendimientos laborales dentro del sector industrial en comparación con el año anterior. Este sector engloba una gran variedad de actividades, como la manufactura, la producción de bienes y otros procesos industriales, que pueden verse influenciados por factores económicos globales, la innovación tecnológica y la política industrial.
- **Variación Interanual de los Rendimientos del Trabajo - Servicios (D.7):** Este indicador refleja la variación en los rendimientos laborales del sector de servicios de un año a otro. El sector servicios incluye una amplia gama de actividades, como el transporte, la educación, la salud, las finanzas, entre otros. La evolución de los rendimientos en este sector puede depender de factores como la demanda de servicios y las condiciones del mercado laboral.
- **Variación Interanual de los Rendimientos del Trabajo - Total (D.8):** Este indicador resume la variación porcentual en los rendimientos laborales a nivel general, sumando todos los sectores mencionados anteriormente (agricultura,

comercio, construcción, hostelería, industria y servicios). Proporciona una visión global de cómo han cambiado los ingresos laborales en la economía en su conjunto, lo que es útil para evaluar la tendencia general del mercado laboral.

En total, el SLMÍ utiliza 30 indicadores base públicos. Como se ha mencionado, se han priorizado variables provenientes de información pública para garantizar replicabilidad y fiabilidad, aunque esto pueda limitar la granularidad en ciertos contextos. Nótese que los datos han sido obtenidos del INE y del SEPE, ambas fuentes de gran prestigio y fiabilidad en lo referente a la estadística y el análisis socioeconómico en España.

En lo referente a la construcción del indicador, en primer lugar, se debe definir el marco teórico, que establece los conceptos a estudiar, determina los subgrupos del índice sintético y, además, define los criterios de calidad que deben mantenerse a lo largo del proyecto. En segundo lugar, se lleva a cabo la recolección de datos implica la estructuración, depuración y selección de variables, de acuerdo con principios como la pertinencia, oportunidad, eficiencia y completitud temporal. Dado que algunas variables no cuentan con todos los datos, la imputación de la información faltante debe ser la tercera etapa del proceso. En cuarto lugar, el análisis multivariante proporciona conclusiones clave sobre el comportamiento de los datos y las correlaciones entre las variables, de forma que estas puedan ser consideradas en las siguientes fases de ponderación y agregación.

Una vez comprendido el carácter de los datos, la normalización ofrece una escala común para todas las variables originales. Después de este paso, los datos estarán listos para ser ponderados y agregados, tomando especialmente en cuenta la estructura definida por el marco teórico y la información obtenida del análisis multivariante. Así, la ponderación asigna pesos a cada variable según su importancia relativa, aunque esto debe ser considerado en conjunto con la técnica de agregación utilizada. En algunos métodos, es posible una compensación total entre los indicadores componentes, por lo que la ponderación debe tener en cuenta esta posibilidad, particularmente si no es adecuada para el índice sintético. Además, los pesos deben considerar la correlación entre variables, ya que puede surgir el problema del doble conteo (Greco et al., 2019).

La fase de agregación reúne las diferentes variables en categorías y luego las agrega para formar el índice sintético final. Para esto, se utilizan técnicas aritméticas y geométricas. Finalmente, el análisis de robustez evalúa el impacto de las decisiones tomadas a lo largo del proceso de construcción, mediante la obtención de 16 resultados distintos generados por decisiones alternativas. Esto resalta la importancia de las técnicas utilizadas en el resultado final: a mayor variabilidad en el índice obtenido, menor robustez tendrá el índice sintético. Tras este paso, la divulgación de los resultados concluye el proceso de elaboración.

Estas etapas están ampliamente documentadas en la bibliografía, ya que cada una tiene una aplicación significativa en otros campos de la Estadística. Por ello, uno de los objetivos del proyecto será analizar y actualizar los manuales existentes de elaboración a fecha de 2022, para que futuros autores dispongan de una guía renovada y contrastada que les permita tomar decisiones adecuadas según el tipo de índice sintético que deseen construir. El otro objetivo, tal vez aún más importante, es la implementación efectiva del indicador sintético en el mercado laboral de España, llenando así el vacío actual en este ámbito. De este modo, el proyecto promoverá no solo la transformación social a través del citado indicador, sino también el establecimiento de bases que permitan a futuros autores desarrollar sus propios índices sintéticos. Para ello, la divulgación de los resultados debe ser igualmente promovida de manera adecuada.

3. Análisis de imputación con diferentes estructuras de datos

a. Estado del arte

El análisis metodológico de las técnicas de imputación de información ausente se centra en entender cómo los diferentes tipos de estructuras de datos, como las series temporales y los datos de panel, influyen en los resultados de la imputación. Para ello, es fundamental realizar una revisión detallada de la literatura existente que aborde estos métodos y sus aplicaciones en diversas situaciones. Este proceso de revisión no solo asegura que el análisis posterior esté bien fundamentado y documentado, sino que también permite ampliar el marco teórico del trabajo, proporcionando un contexto más amplio y actualizado sobre las mejores prácticas en la imputación de datos. Además, entender cómo cada tipo de estructura de datos impacta en los resultados de la imputación es esencial para elegir y aplicar las técnicas más adecuadas, lo que permitirá obtener resultados más precisos y fiables en la construcción de indicadores.

Es fundamental identificar el método de imputación más fiable para un conjunto de datos específico, aunque, según la OCDE (2008), no existe una solución universal. El primero de los enfoques estudiados es el uso de métodos de regresión, donde cada indicador con datos faltantes se modela como una variable dependiente, mientras que los indicadores restantes, especialmente aquellos altamente correlacionados, funcionan como variables independientes (Donders et al., 2006). Este método asume que los datos siguen una distribución MAR y, para mejorar su precisión, suele incorporar residuos con distribución normal en la regresión (Musil et al., 2000)

Otro método es la imputación mediante Expectation-Maximization (EM), que se sitúa entre la imputación simple y la múltiple, ya que emplea un algoritmo iterativo de dos fases. En la primera etapa, llamada esperanza, se estiman valores para los datos ausentes basándose en la información disponible. En la segunda, denominada maximización, se recalculan los parámetros de la distribución, como medias y covarianzas, incorporando las imputaciones previas. Este ciclo se repite hasta que el proceso converge en estimadores de máxima verosimilitud (OCDE, 2008). Su aplicación ha demostrado generar resultados precisos en estructuras de datos MCAR y MAR, siendo especialmente útil en comparación con otros métodos de imputación simple (Gold & Bentler, 2000). También se ha evaluado su desempeño en análisis de componentes principales, obteniendo buenos resultados (Malan et al., 2020). Sin embargo, su principal desventaja es que, en algunos casos, el proceso de convergencia puede ser lento.

Por último, la imputación múltiple (IM) consiste en un procedimiento aleatorio que genera N conjuntos de datos para representar la incertidumbre asociada a los valores faltantes. Cada conjunto se obtiene mediante imputaciones subyacentes, y posteriormente los parámetros se combinan para formar un conjunto final (Bennett et al., 2001). En este proceso, cada valor ausente recibe N valores imputados que luego se fusionan. Las técnicas subyacentes pueden ser diversas, incluyendo métodos regresivos, pero una de las más robustas se basa en métodos MCMC (Markov-Chain Monte Carlo), los cuales utilizan procesos estocásticos para modelar la dependencia de cada estado con respecto al anterior (Medel Esquivel et al., 2019). Aunque la justificación matemática de este enfoque es compleja, se ha demostrado que proporciona resultados altamente representativos y robustos en términos de calidad de los datos imputados (OCDE, 2008).

Aunque la imputación múltiple y el método EM presentan características destacadas, su implementación puede generar resultados imprevistos. Para minimizar esta incertidumbre, se recomienda realizar un análisis empírico, similar al propuesto por Gold y Bentler (2000). Este procedimiento consiste en eliminar aleatoriamente algunos datos de un conjunto disponible y aplicar diferentes técnicas de imputación. Posteriormente, se evalúa la diferencia entre los valores

imputados y los valores originales mediante un indicador, como el error cuadrático medio normalizado (NRMSE, por sus siglas en inglés).

Además del NRMSE, otros criterios útiles para medir la precisión de una técnica de imputación incluyen el coeficiente de correlación y el índice de concordancia de Willmot (1985), así como métricas tradicionales como el error cuadrático medio (RMSE) y el error absoluto medio (MAE) (OCDE, 2008). Sin embargo, estos métodos pueden no ser completamente adecuados en el contexto de indicadores sintéticos, especialmente cuando las escalas de los indicadores aún no han sido estandarizadas. En este sentido, un método de imputación con un rendimiento global superior podría verse afectado si su desempeño es pobre en una variable con una escala significativamente mayor que el resto, lo que incrementaría el error cuadrático.

El RMSE es útil para evaluar la precisión de la imputación. Sin embargo, en el caso de los indicadores sintéticos, si las escalas no están estandarizadas, un método de imputación con buen desempeño global puede verse afectado por una variable con una escala mayor que las demás, lo que aumenta el error.

Para contrarrestar este sesgo, es preferible utilizar un indicador normalizado basado en la media de la distribución, como el NRMSE, que permite comparar diversas técnicas de imputación de manera objetiva y confiable (Shcherbakov et al., 2013). Si se requiere una comparación entre subgrupos o categorías, la normalización facilita la agregación de los NRMSE de los diferentes indicadores, evitando la dificultad de un análisis individualizado, que podría ser poco viable debido a la cantidad de datos y el tiempo requerido. De esta manera, la ecuación del NRMSE para n indicadores se expresa de la siguiente forma:

$$NMRSE = \frac{\sqrt{\frac{\sum_i^N (X_i - \hat{X}_i)^2}{N}}}{\max(X_i) - \min(X_i)}$$

Donde X_i es el valor conocido y real de las variables i sin normalizar, \hat{X}_i es el valor imputado de cada variable i sin normalizar.

Cabe destacarse, por tanto, que a la hora de realizar la imputación una parte de los valores se deberán ocultar para obtener el NRMSE. La cantidad debe ser proporcional al porcentaje de datos ausentes de la estructura de datos original para así realizar imputaciones cuyos resultados e impactos sean replicables y fiables sobre la estructura original. Además, conocer el número de datos faltantes es capital para que se imputen el mismo número de datos sobre la muestra y que la suma de NRMSE de las variables pueda realmente ser comparable.

Nótese como la normalización min-máx. (presente en el denominador de la fórmula mostrada) habilita la comparación entre NRMSE de diferentes variables en tanto que lo reescala de 0 a 1, evitando complicaciones derivadas de un análisis individualizado. En este sentido, las estrategias de normalización más comunes para el cálculo del NRMSE son la normalización por la media (empleada en la obtención del NRMSE en el trabajo de Álvarez [2022]) y la Min-Max, que se utilizó en la construcción del propio SLMI. La normalización por la media puede ser útil cuando los datos tienen una distribución estable, pero podría sesgar la interpretación si existen valores extremos o, por ejemplo, si toma valores a 0. Estos problemas, normalmente analizados en el proceso de construcción de los sintéticos, no se ponen bajo la lupa cuando se realiza la imputación, puesto que el análisis teórico del NRMSE no necesariamente se suele incluir dentro del alcance metodológico de los trabajos.

Al contrario, en este trabajo y para mantener la consistencia con la construcción del SLMI, se aplicará la normalización Min-Max sobre el RMSE, la cual ajusta el error según la variabilidad

total de los datos, siendo más adecuada para datos heterogéneos o con valores atípicos a la par que no produce resultados extremos por negativos o situarse el promedio cercano al 0.

De esta manera, una vez descrito el mecanismo que permitirá comparar métodos de imputación, cabe destacar el verdadero valor añadido del proyecto, versado sobre la comparación de NRMSE para las diferentes estructuras de datos. A la hora de imputar los datos que pertenecen a variables que a su vez están enmarcadas en áreas que definen sintéticos disponibles en varios países, debe plantearse sobre qué estructura de datos ha de realizarse la imputación: ¿todo el conjunto de datos, incluyendo todas las variables al mismo tiempo y para todos los países o variable por variable para capturar mejor las variaciones específicas de cada indicador base? ¿Y si, al contrario, se imputa considerando todos los indicadores pero país a país?

A nivel teórico, hay más motivos para pensar que cuánta más información se disponga para realizar la imputación, en tanto que se dan las siguientes circunstancias:

1. **Mayor cantidad de información para la imputación:** Cuando se considera el set completo de datos, el modelo tiene acceso a una mayor cantidad de información para realizar la imputación (Young & Johnson, 2015). Esto puede ser crucial, especialmente si los datos están relacionados de manera no lineal o si existen patrones complejos que solo pueden ser capturados cuando se consideran todas las variables conjuntamente. Al evaluar sobre el set completo, el modelo puede aprovechar estas relaciones globales para hacer una imputación más precisa que si solo considerara información local (como áreas o subáreas específicas).
2. **Distribución dispar en la cantidad de datos ausentes:** En muchos casos, los datos faltantes no son distribuidos de manera uniforme entre las diferentes áreas o subáreas, lo que puede generar sesgos si se imputan de forma local. Si la imputación se realiza a nivel global, el modelo puede suavizar el % de datos faltantes y evitar imputaciones extremas debido a una posible preponderancia de información ausente en una cierta área o categoría.
3. **Riesgo de sobreajuste con imputación local:** Imputar por área o subárea puede llevar a un sobreajuste local si los modelos no son suficientemente robustos o si las áreas/subáreas tienen muy pocos datos. Los modelos podrían ajustarse excesivamente a las características locales de esas áreas, lo que puede resultar en imputaciones menos precisas cuando se generaliza a otras áreas o al conjunto completo de datos. Al trabajar con el set completo, se evita este riesgo y se obtiene una imputación que, si bien podría no ser perfectamente precisa en cada área, tiene una mayor capacidad de generalización.

En resumen, aunque la imputación por área o subárea tiene sus ventajas, como la posibilidad de capturar características locales específicas, la imputación global sobre el set completo de datos puede ofrecer mayor precisión en muchos casos, especialmente cuando se dispone de suficientes datos para reflejar regresiones más completas. En cualquier caso, esto deberá ser ratificado experimentalmente en la sección de resultados, aunque se recomienda estudiar la aplicación de una imputación u otra caso por caso según la estructura de datos estudiada

A continuación, se pretende analizar empíricamente qué nivel de agregación es más efectivo para la imputación sobre el indicador descrito y sobre las técnicas de imputación descritas en el presente apartado.

b. Resultados

Para aterrizar este análisis metodológico, se estudiará el impacto de imputar a diferentes niveles de agregación del SLMi y de un indicador con estructura de panel como se muestra a continuación:

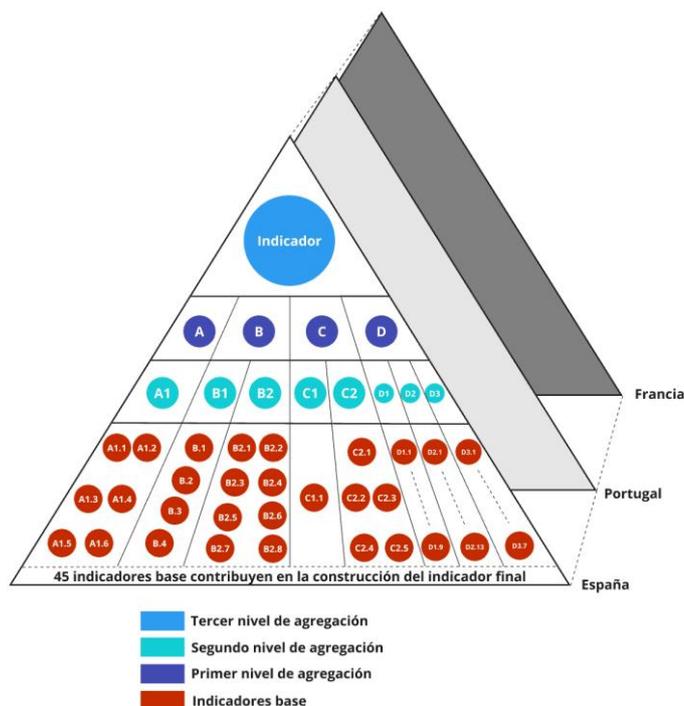


Ilustración 9. Estructura de panel del indicador estudiado en la imputación

La estructura mostrada pertenece al Senior Economy Tracker (Roch-Dupré et al., 2023), que mide el progreso de 27 economías europeas en su adaptación a un contexto de transición demográfica a lo largo del tiempo ($N = 18$ años). Su estructura de datos en forma de panel lo hace idóneo para la labor que se procederá a realizar (ver anexo 4).

De igual manera, el indicador muestra una tercera dimensión para conformar el panel: la distribución geográfica. Cada país tiene sus respectivos indicadores en cada subárea, siguiendo la misma estructura. Hay información disponible para los siguientes países: España (ESP), Portugal (PRT), Francia (FRA), Austria (AUT), Bélgica (BEL), República Checa (CZE), Dinamarca (DNK), Finlandia (FIN), Alemania (DEU), Grecia (GRC), Hungría (HUN), Irlanda (IRL), Italia (ITA), Luxemburgo (LUX), Países Bajos (NLD), Noruega (NOR), Polonia (POL), Eslovaquia (SVK), Suecia (SWE), Reino Unido (GBR), Estonia (EST), Eslovenia (SVN), Letonia (LVA), Lituania (LTU), Bulgaria (BGR), Croacia (HRV), Rumania (ROU).

A la hora de analizar el NRMSE y compararlo para las diferentes estructuras de datos, cabe estudiar qué cantidad de datos faltantes presenta este indicador. Sobre el total de indicadores para toda la serie temporal y países, el SET goza de un c.17% de datos faltantes, lo que aproximadamente se traducirá en que 3 de los 18 datos temporales deben “ocultarse” en cada variable para analizar el NRMSE que, se recuerda, es una muestra normalizada del error que presenta una estimación sobre el valor real que, previamente, se ha ocultado.

De esta manera, sobre el indicador estudiado, se han obtenido los siguientes valores de NRMSE incorporando la información cruzada de los países:

	Todas juntas	Por áreas		Por subáreas			
EM		A	0.339	A	0.339	A1	0.339
		B	0.585	B	0.632	B1	0.191
	1.930	C	0.304	C	0.333	B2	0.441
						C1	0.073
						C2	0.261
						D1	0.448
	D	0.867	D	1.333	D2	0.631	
					D3	0.254	
	1.930		2.095		2.638		2.638
IM		A	0.279	A	0.279	A1	0.279
		B	0.710	B	0.817	B1	0.203
	2.136	C	0.304	C	0.367	B2	0.614
						C1	0.073
						C2	0.294
						D1	0.509
	D	0.867	D	1.544	D2	0.772	
					D3	0.263	
	2.136		2.160		3.007		3.007
Regresión lineal		A	0.567	A	0.567	A1	0.567
		B	0.909	B	1.136	B1	0.333
	3.024	C	0.388	C	0.404	B2	0.803
						C1	0.104
						C2	0.300
						D1	0.588
	D	1.256	D	1.979	D2	0.952	
					D3	0.439	
	3.024		3.121		4.086		4.086

Tabla 5. Resultados del análisis de imputación considerando todos los países europeos

De esto se deduce que imputar todos los datos juntos, incorporando la información de panel en la imputación el NRMSE se reduce. De hecho, este comportamiento se ratifica también en cuando se imputa por áreas o por subáreas, en tanto que el NRMSE siempre aumenta cuanto menor información se impute de forma conjunta. Los datos mostrados incorporan la información de panel de los 27 países europeos que disponen del indicador.

Posteriormente, el análisis se ha reproducido obviando la característica de panel sobre el indicador, es decir, imputando las variables de cada país por separado y sumando los NRMSE

por comparabilidad con la tabla 4 en la que la imputación se realiza a lo largo del panel. Los resultados sin considerar los países europeos se muestran a continuación:

	Todas juntas	Por áreas		Por subáreas			
EM		A	1.251	A	1.251	A1	1.251
		B	1.499	B	2.194	B1	0.962
	5.213	C	0.565	C	0.755	C1	0.432
						C2	0.323
						D1	0.933
						D2	1.232
	D	2.814	D	3.347	D3	1.182	
5.213		6.129		7.547		7.547	
	Todas juntas	Por áreas		Por subáreas			
IM		A	1.431	A	1.493	A1	1.493
		B	1.849	B	2.387	B1	1.122
	5.732	C	0.612	C	0.835	C1	0.510
						C2	0.325
						D1	0.921
						D2	1.321
	D	3.127	D	3.429	D3	1.187	
5.732		7.019		8.144		8.144	
	Todas juntas	Por áreas		Por subáreas			
Regresión lineal		A	4.505	A	4.505	A1	4.505
		B	6.190	B	8.602	B1	3.407
	18.232	C	4.722	C	4.732	C1	2.251
						C2	2.481
						D1	3.795
						D2	5.891
	D	9.964	D	12.594	D3	2.908	
18.232		25.380		30.430		30.430	

Tabla 6. Resultados del análisis de imputación obviando la dimensionalidad regional

Se percibe una mejora consistente a todos los niveles cuando, en lugar de realizar imputaciones de $N = 18$, se realiza con un conjunto de datos de $N = 18 \cdot 27 = 486$. Tal y como arguyen Young y Johnson (2015), la imputación con datos de panel aumenta la cantidad de información disponible para el análisis, lo que mejora la eficiencia y reduce los errores estándar.

Esto queda patente en el estudio, en tanto que además el ajuste tiene un mayor impacto dependiendo del método empleado, dándose para la regresión la mayor reducción en el NRMSE. Después de todo, aunque el conjunto de datos original dispone de 18 datos, la imputación solo se puede aplicar sobre un set de datos completo. Esto reduce la mayoría de las variables a aproximadamente 12 datos (ya que la serie temporal debe estar completa para considerarse válida a la hora de imputar), a las que posteriormente han de eliminarse 3 más para evaluar el NRMSE. Por ello, la cantidad de información disponible para el indicador SET estudiado es realmente baja si no se imputa considerando los países.

De igual manera, emplear más variables, incluso aquellas que pertenecen a otras áreas, también facilita la labor de imputación y mejora los resultados para todas las técnicas de imputación estudiadas.

Este mismo análisis se ha realizado sobre el SLMI, con la salvedad de que este no se puede tratar como un panel de datos debido a la falta de datos comparables de otros países. Sin embargo, a pesar de esta limitación, sí se puede proceder a comparar la imputación a primer nivel de agregación (áreas), que es la que se realizó en la construcción del SLMI, comparándola frente a la imputación del segundo nivel de agregación (SLMI)

- Primer Nivel de Agregación (Áreas): Aquí, los datos se imputan sobre las áreas descritas de *Desempleo, Empleo, Protección a los Desempleados y Salarios y costes laborales*.
- Segundo Nivel de Agregación (SLMI): En este nivel, los datos faltantes se imputarán considerando las treinta variables que componen el indicador al mismo tiempo.

La comparación entre estos dos niveles de agregación tiene como objetivo evaluar la precisión y efectividad de la imputación en cada uno. En el primer nivel (áreas), es posible contar con una mayor granularidad de los datos, lo que podría facilitar una imputación más ajustada a la las desviaciones de índole más específica. En resumen, aunque no sea posible realizar una imputación a nivel de panel debido a la falta de datos internacionales comparables, la comparación entre las imputaciones a primer y segundo nivel de agregación puede ofrecer valiosa información sobre la robustez de las imputaciones realizadas.

Para llegar a esta conclusión, en primer lugar, se debe generar una matriz de datos completa, lo que implica eliminar los casos con información faltante tal y como se describió en el trabajo de Álvarez (2022). A continuación, se procederá a suprimir un porcentaje representativo de casos conocidos de la muestra, conforme a la tasa de datos ausentes de cada estructura de datos. En este proceso, se eliminan los datos correspondientes a tres meses, lo que conlleva una tasa de pérdida cercana al 2% por categoría. Sin embargo, esta eliminación no se aplica a dos variables específicas de cada subgrupo, ya que ciertos métodos de imputación requieren la presencia de datos parciales para funcionar correctamente. La selección de las dos variables cuyos datos completos se mantendrán en el conjunto de datos se realiza mediante un número generado con una distribución uniforme, asegurando que las matrices utilizadas para comparar técnicas sean aleatorias.

Posteriormente, se imputan los datos utilizando IBM SPSS, aplicando imputaciones simples mediante el método EM (Expectación-Maximización) y regresión, así como imputación múltiple (IM) utilizando el método MCMC.

Realizando los pasos mencionados, los resultados son consistentes con lo mostrado para el Senior Economy Tracker, obteniendo un NRMSE de 0.9747 para la técnica EM aplicada sobre todo el conjunto de datos sobre la serie temporal que se empleó en el trabajo (Álvarez, 2024) que además compara muy positivamente frente al 2.564 obtenido al construir el SLMI. En el caso del Sin embargo, debido a la robustez del indicador, esto apenas goza de impacto sobre el indicador final.

El resultado de aplicar la forma revisada sobre el SLMI se muestra a continuación en la ilustración:

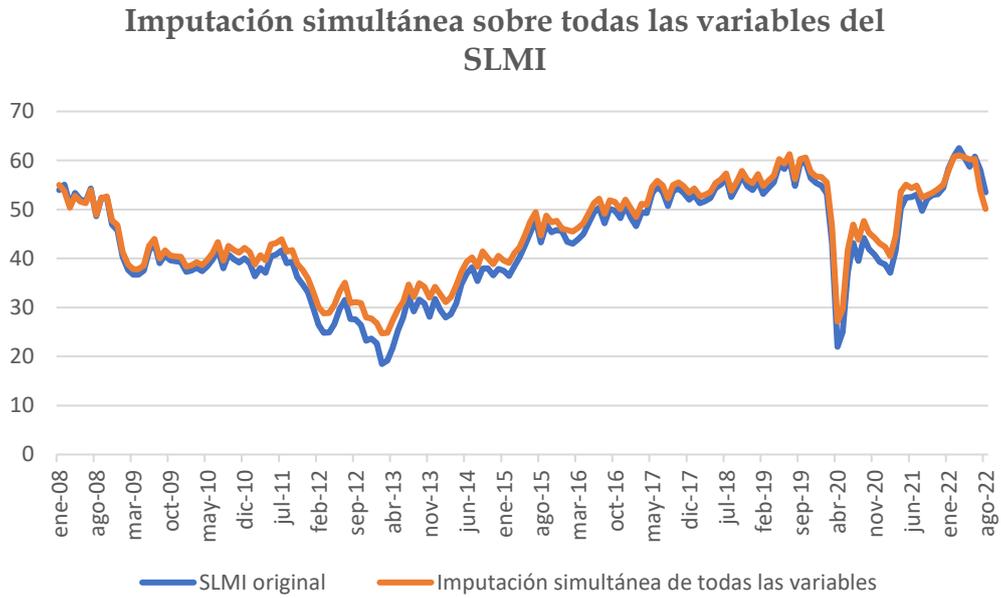


Ilustración 10. Impacto de la imputación simultánea sobre el SLMI

De esta forma, si bien el SLMI ha sido construido de una manera en la que el impacto de la imputación apenas altera el resultado, especialmente considerando el bajo número de datos faltantes (c.2%), la mejora en el NRMSE frente al TFG denota como una mayor cantidad de datos mejora la imputación.

4. Métodos avanzados de ponderación: FA vs análisis de importancia

a. Estado del arte

La evaluación metodológica de la fase de ponderación se enfoca en analizar la "importancia" asignada a cada variable, tanto a través del criterio experto como de métodos estadísticos basados en la variabilidad, como el análisis de componentes principales (PCA) y el análisis factorial (FA). En este sentido, la revisión de la literatura sobre la ponderación y sus posibles mejoras metodológicas se convierte en una tarea clave. Este proceso implica una profundización en los enfoques académicos existentes, lo cual permite fundamentar cualquier propuesta metodológica en bases teóricas sólidas. Esta revisión no solo proporciona una comprensión más profunda de cómo se asignan los pesos a las diferentes variables, sino que también contribuye a la construcción de un marco teórico robusto que servirá como soporte para las decisiones metodológicas que se adoptarán en el proyecto. El conocimiento adquirido en esta etapa será esencial para desarrollar propuestas que mejoren la precisión y la relevancia de la ponderación en la construcción de indicadores sintéticos.

El análisis factorial es una técnica estadística utilizada para reducir la dimensionalidad de los datos y encontrar estructuras subyacentes en un conjunto de variables observadas. Para asignar los pesos a las variables, sigue varios criterios fundamentales que permiten determinar la influencia de cada variable en los factores extraídos. Estos criterios se basan en la correlación entre variables, la varianza explicada, la rotación factorial, la carga factorial y la interpretación teórica (OCDE, 2008).

En primer lugar, el análisis factorial asigna pesos a las variables con base en la correlación entre ellas. Se parte de una matriz de correlaciones, donde se identifican variables que tienen relaciones significativas entre sí. Las variables altamente correlacionadas tienden a agruparse bajo un mismo factor, y el peso de cada una dependerá de qué tan fuerte sea su relación con dicho factor. Cuanto mayor sea la correlación entre una variable y un factor, mayor será su peso en la construcción del factor, en tanto que su varianza explicada sea preponderante sobre el mismo.

Cabe resaltar que en el análisis factorial, se busca reducir el número de variables manteniendo la mayor cantidad posible de información, por lo que la varianza explicada se convierte en el elemento decisor a la hora de elegir el número de factores que definirán dicha realidad. En este sentido, los factores se ordenan según la proporción de varianza que explican del conjunto de datos original. En consecuencia, las variables que contribuyen más a explicar la varianza del conjunto recibirán mayores pesos en los factores correspondientes. Esto se logra a través del cálculo de los valores propios (eigenvalues) de la matriz de correlaciones.

La rotación factorial también influye en la asignación de pesos, puesto que se utilizan técnicas como la rotación varimax para redistribuir las cargas factoriales y lograr una estructura más interpretable. La rotación busca maximizar las diferencias entre factores, de manera que cada variable se asocie de manera más clara con un solo factor en lugar de dispersarse en varios. Esto permite obtener factores más definidos y mejorar la asignación de pesos a las variables.

Finalmente, la interpretación teórica de los factores juega un papel crucial. Aunque el análisis factorial es un método matemático-estadístico, su éxito depende de la capacidad de los investigadores para interpretar los factores de manera coherente con el fenómeno estudiado. Después de todo, el propio método asigna pesos a las variables según su carga factorial y su varianza explicada sobre el conjunto de datos. En ciertos casos, la ponderación y agregación en torno al PCA o FA proporciona unos factores intermedios que proporcionan mayor granularidad y facilita la labor de los investigadores en lo referente a la trazabilidad e interpretabilidad de los

datos. Esto se conoce como agregación jerárquica (Szopik-Depczynska et al., 2018). Sin embargo, este enfoque no se ha realizado en el SLMI, donde el FA es empleado exclusivamente para obtener unos pesos que minimicen las correlaciones del conjunto.

Por otro lado, el análisis de importancia se basa en comprender las correlaciones entre las variables y evaluar la sensibilidad de cada una en relación con el modelo. A diferencia del análisis factorial, que se enfoca en identificar estructuras subyacentes, el análisis de importancia compara el índice de sensibilidad de primer orden de cada variable, considerando tanto correlaciones lineales como no lineales. Esto permite ajustar los pesos de manera que se minimice la desviación entre la importancia real de cada variable y su importancia deseada dentro del modelo.

Este enfoque es particularmente útil para garantizar que las variables que tienen un papel fundamental en la interpretación del modelo reciban un peso acorde a su impacto real. Así, al integrar el análisis de importancia con los criterios tradicionales del análisis factorial (correlación, varianza explicada, rotación factorial e interpretación teórica), se podría obtener un modelo más robusto, preciso y representativo de la realidad analizada.

En este sentido, el análisis de importancia representa una extensión del análisis factorial al incorporar un criterio adicional para la asignación de pesos. Mientras que el FA prioriza la identificación de factores latentes, el análisis de importancia busca mejorar la interpretación y fiabilidad del modelo ajustando los pesos de manera más precisa. Este contraste subraya la necesidad de combinar ambos enfoques para optimizar el cálculo de los indicadores sintéticos en aplicaciones reales.

Sin embargo, llegados a este punto, la pregunta que surge es, ¿bajo qué circunstancias debe aplicarse FA y en qué momento resulta más apropiado realizar un análisis de importancia?

El análisis factorial ofrece ventajas importantes en diversas situaciones. En primer lugar, es útil para reducir la dimensionalidad cuando el objetivo es condensar un gran conjunto de variables en unas pocas dimensiones subyacentes. Además, permite identificar realidades latentes al descubrir estructuras ocultas en los datos que explican las correlaciones entre variables observadas. En un enfoque exploratorio, el análisis factorial facilita la comprensión de cómo se agrupan las variables y genera hipótesis sobre relaciones subyacentes. También se puede aplicar de forma confirmatoria cuando ya existen hipótesis teóricas sobre la estructura de los datos, con el fin de validarlas. Finalmente, el análisis factorial es eficaz cuando hay grupos de variables altamente correlacionadas que representan factores comunes.

Por otro lado, el análisis de importancia también presenta varios puntos fuertes. Si el objetivo es asignar pesos a las variables minimizando la desviación entre la importancia real y la deseada, el análisis de importancia resulta adecuado. Asimismo, es valioso cuando se necesita evaluar el impacto de cada variable en la predicción o estabilidad del modelo, a través de un análisis de sensibilidad. El análisis de importancia es útil para capturar tanto correlaciones lineales como no lineales, algo que el análisis factorial podría no detectar completamente. Además, permite mejorar la interpretabilidad del modelo al ajustar los pesos de las variables para reflejar mejor su relevancia práctica o conceptual.

Aplicar ambas técnicas sobre el mismo conjunto de datos para comprender su estructura estadística puede ser muy beneficioso. En primer lugar, el análisis factorial (FA) revela estructuras subyacentes al identificar factores comunes que subyacen a un conjunto de variables correlacionadas. Este método reduce la dimensionalidad de los datos y puede descubrir factores latentes que son difíciles de identificar a simple vista, proporcionando una comprensión más profunda de cómo se relacionan las variables entre sí. En este sentido, el FA responde preguntas clave sobre qué realidades describen las dimensiones subyacentes que explican la variabilidad observada en los datos. Este proceso es especialmente útil cuando los datos son complejos o

cuando las relaciones entre las variables no son evidentes de inmediato, ya que el FA facilita la comprensión de los patrones subyacentes.

Mientras tanto, el análisis de importancia es útil para encontrar correlaciones no lineales, un área en la que el FA tiene limitaciones. Aunque el FA es eficaz para identificar correlaciones lineales y factores comunes, el análisis de importancia se especializa en entender cómo distintas variables impactan en un modelo predictivo, incluyendo relaciones más complejas y no lineales. A través del análisis de importancia, los pesos asignados a las variables pueden ajustarse para reflejar más precisamente su relevancia en un contexto específico, como la predicción o la toma de decisiones.

Otro aspecto clave a considerar en la combinación de ambos métodos es la trazabilidad e interpretabilidad de los resultados. Mientras que el análisis factorial puede introducir nuevas dimensiones mediante la extracción de factores, el análisis de importancia trabaja directamente con los indicadores originales, sin crear nuevas dimensiones. Esto permite ajustar los pesos de las variables sin perder la trazabilidad de los resultados, lo que facilita su interpretación práctica.

Finalmente, la aplicación de ambos métodos, sin necesidad de combinarlos, mejora la comprensión de la estructura de datos y puede facilitar el discernimiento en lo que a selección de técnica de ponderación se refiere. En resumen, el análisis factorial simplifica los datos al reducir su dimensionalidad, pero no siempre captura toda la complejidad de las relaciones. Por su parte, el análisis de importancia refina los pesos asignados a cada indicador, asegurando que la relevancia de cada variable se represente de manera precisa y clara.

En resumen, al aplicar el análisis factorial y el análisis de importancia sobre el set de datos, los autores pueden beneficiarse de la capacidad del FA para revelar estructuras subyacentes y las correlaciones entre variables, mientras que el análisis de importancia permite ajustar los pesos de manera más precisa para mejorar la interpretación y la aplicabilidad práctica de los resultados en base unas importancias deseadas que el autor podría haber discurrido al evaluar los resultados derivados del FA. De este modo, se puede obtener una visión más completa de los datos, logrando un equilibrio entre la comprensión profunda y la eficiencia de los modelos.

Una vez explicadas las implicaciones que podría tener el análisis de importancia, cabe plantearse cómo se materializa dicho análisis sobre las variables estudiadas del SLMI. Para ello, debe comprenderse brevemente la naturaleza de los indicadores sintéticos. Un indicador sintético y , calculado a partir de n variables de entrada normalizadas (x_i), puede resultar de combinar estas variables mediante un promedio aritmético ponderado (OCDE, 2008), representado como:

$$y = \sum_{i=1}^n w_i x_{ij}, \quad j = 1, 2, \dots, n$$

donde x_{ij} es el valor normalizado de un indicador base y w_i es el peso asignado a x_{ij} cumpliendo que $\sum_{i=1}^n w_i = 1$ y se cumple que cada peso toma un valor positivo.

Esto también puede tomar la forma de agregación geométrica:

$$y = \sum_{i=1}^n x_{ij}^{w_i}, \quad j = 1, 2, \dots, n$$

Un punto clave a comprender es que la influencia efectiva de cada variable x_i sobre el sintético y no siempre coincide con su peso w_i , ni guarda necesariamente una relación lineal con este. Esto se debe a que la influencia de x_i sobre el indicador sintético depende también de las correlaciones (que pueden no ser lineales) entre x_i y las demás variables del conjunto. Por lo tanto, aislar y cuantificar el impacto específico de cada variable resulta capital para comprender realmente qué peso debe asignarse a cada variable y, de hecho, no es algo necesariamente sencillo.

Para analizar la contribución de cada variable x_i al indicador sintético, se pueden definir medidas de importancia basadas en la relación entre ambos. Una forma común de hacerlo es medir cómo el sintético depende de x_i . Esto puede expresarse mediante la ecuación:

$$y = f_i(x_{ij}) + \varepsilon_j$$

donde $f_i(x_{ij})$ representa la parte del sintético explicada directamente por x_i , y ε_j es un término de error que recoge la influencia de las demás variables. Si la relación entre y y x_i es lineal, la dependencia puede cuantificarse usando el coeficiente de determinación R^2 , definido como:

$$R_i^2 = \frac{SS_{reg,i}}{SS_{tot}}$$

donde $SS_{reg,i}$ es la suma de cuadrados explicada por la regresión lineal del sintético sobre x_i , y SS_{tot} es la suma total de cuadrados de y . Sin embargo, este método no capta relaciones no lineales, ya que se basa exclusivamente en la regresión lineal cuando $f_i(x_{ij})$ se puede estimar mediante un único regresor (Chicco et al., 2021).

En el contexto de indicadores compuestos, aunque la fórmula de agregación suele ser lineal, las correlaciones entre variables pueden introducir no linealidades en la relación entre y y x_i . Para capturar estas dinámicas, se emplea una medida más general S_i , conocida como índice de sensibilidad de primer orden. Este enfoque mide la influencia potencialmente no lineal de cada variable en el indicador compuesto y puede interpretarse como la reducción esperada en la varianza de y al fijar x_i (Becker et al., 2017).

El índice de sensibilidad de primer orden se definiría por tanto como:

$$S_i = \frac{Var(E_{x_{\sim i}}[y|x_i])}{Var(y)}$$

donde $E_{x_{\sim i}}[y|x_i]$ es el valor esperado de y para x_i , asumiendo que el resto de variables i son evaluadas constantes. Para aterrizar el concepto, esto implica que cada valor de y puede explicarse a partir del valor de x_i (ajustado por el parámetro de regresión) cuando el resto de variables toman el valor esperado, i.e. su media. Esto puede interpretarse como una regresión no lineal de y sobre x_i , donde se observa la varianza explicada por la curva ajustada. De forma terrenal, S_i puede estimarse como:

$$\hat{S}_i = \frac{\sum_{j=1}^n (\hat{m}_{ij} - \bar{m}_i)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

donde \hat{m}_{ij} representa el valor ajustado de la regresión de y sobre x_i para la observación j , y \bar{m}_i es el promedio de esos valores ajustados. El denominador simplemente muestra la diferencia entre la variable y respecto de su media. Esta medida generaliza el R_i^2 , coincidiendo con él en caso de que la relación sea lineal, pero permitiendo analizar dependencias más complejas (Becker et al., 2017).

En resumen, el índice de sensibilidad de primer orden (S_i) proporciona una forma efectiva de evaluar la importancia de las variables de entrada en un indicador compuesto, considerando tanto las relaciones lineales como las no lineales. De hecho, la literatura (Hastie et al., 2008) ya describe los principales métodos de regresión que se pueden aplicar para ejecutar la regresión no lineal de y sobre x : procesos gaussianos y splines penalizados. Este último, que viene a basarse en una regresión con componentes elevados a una potencia, viene definido por:

$$f(x_i) = \beta_0 + \beta_1 \cdot x_i + \sum_i^N \beta_i (x_i - \alpha_i)_+^k$$

Donde la β_0 y β_1 son los coeficientes de la parte lineal y α_i representa los diferentes *knots*, puntos que representan discontinuidades en la regresión del modelo. Así, si se modela una regresión con 2 *knots* en 3 y 6, el spline ajustará el polinomio separadamente entre $[0,3]$, $[3,6]$ y $[6, \infty)$. Esto está reflejado en el símbolo $_+$, que implica que la diferencia $x - \alpha_i$ debe considerarse nula si resulta negativa, lo que implica que $x < \alpha_i$. La k se aplica según el tipo de ajuste deseado, aunque típicamente los splines cúbicos ($k = 3$) gozan de amplio respaldo en el mundo académico por proporcionar el equilibrio entre ajuste y suavidad en los modelos no lineales (Wood, 2017), por lo que estos se usarán en el estudio. El impacto de la posible variación de los *knots* se deja abierto para futuras líneas de investigación.

De la fórmula antes mencionada, cabe mencionar que se aplica una penalización (de ahí splines “penalizados”) para garantizar la continuidad y suavidad en los *knots*, es decir, que las derivadas de los polinomios obtenidos sean iguales en dichos puntos. Tal y como refleja Hastie et al. (2008), la penalización sobre la regresión viene dada por la inclusión del término en los mínimos cuadrados:

$$MCO = \sum (y_i - f(x_i))^2 + \lambda \int (f''(x_i))^2 dt$$

Donde $f''(x_i)$ es la segunda derivada de $f(x_i)$ y λ se considera el parámetro de suavizado que busca un compromiso adecuado entre ajuste y suavidad del modelo obtenido. En este caso, se han obtenido los *knots* y el parámetro de suavizado a partir de validación cruzada siguiendo el criterio establecido en la literatura (Ruppert et al., 2003).

De esta manera, queda resumido cómo se obtienen las correlaciones no lineales. Nótese que, de evaluarse S_i de forma meramente lineal, su valor coincidiría con R^2 , tal y como se ha mencionado originalmente.

En el ámbito del análisis de sensibilidad, se ha observado ampliamente (Xu y Gertner, 2008; Li et al., 2010; Mara y Tarantola, 2012) que, cuando las variables de entrada están correlacionadas, S_i (también conocida como índice de sensibilidad de primer orden) incluye tanto el efecto directo de la variable x_i como el impacto derivado de su correlación con otras variables. Esto se expresa como:

$$S_i = S_i^{\text{incorrelacionada}} + S_i^{\text{correlacionada}}$$

donde S_i^i representa la parte no correlacionada y S_i^c la parte correlacionada. Este desglose proporciona información adicional sobre la influencia de cada variable.

La distinción entre parte correlacionada e incorrelacionada permite una mayor comprensión de la naturaleza de la influencia de los indicadores base sobre el sintético, en este caso a nivel de área. Por ejemplo, si $S_i^c = S_i$, significa que el peso de la variable tiene poco impacto en el resultado, ya que su influencia proviene principalmente de su correlación con otras variables. Si S_i^c resulta negativo, indicaría un problema conceptual según Becker et al. (2017), generalmente derivado de correlaciones negativas entre variables, lo cual puede ser inapropiado en indicadores compuestos y de hecho se ha intentado evitar en la construcción del SLMi a la hora de normalizar las variables. Después de todo, en su construcción se aplicó la normalización min-máx y posteriormente se reorientaron aquellas variables para facilitar la asignación de los pesos en la ponderación y agregación.

Resulta muy importante “reorientar” la escala de los indicadores para que, a la hora de agregarlos, no se produzcan problemas conceptuales. Por ejemplo, al medir la situación del

mercado laboral de un país, mezclar variables positivas (tasa de empleo) con negativas (cierre de empresas) puede ser problemático para asignar pesos. De hecho, muchas técnicas de ponderación emplean solo pesos positivos, asumiendo esta condición como requisito. Por lo tanto, la propia OCDE (2008) recomienda orientar todos los indicadores de manera que tengan el mismo “sentido”: un valor mayor del indicador implicaría un mejor desempeño. Considerando que en el SLMI se ha seguido esta lógica y reorientado los indicadores de la variable “Desempleo” por describir realidades negativas en cuanto al desempeño del mercado laboral, resultaría extraño obtener valores de $S_i^{correlada}$, lo que implicaría problemas metodológicos

En el análisis de sensibilidad, la descomposición del coeficiente de correlación proporciona información adicional sobre cómo podría reducirse la incertidumbre, ya que correlaciones fuertes implican que las variables deben tratarse en grupos correlacionados en lugar de individualmente. Cabe esperarse, por tanto, que en conjuntos de datos con una gran número de variables, la importancia este dominada por la parte correlacionada.

De esta forma, resulta capital descomponer S_i , para lo que se sigue el enfoque de Xu y Gertner (2008), basado en regresión. Este método permite generalizar las relaciones entre y y x_i (o entre las x_i entre sí), tanto lineales como no lineales. Dicho índice se estima inicialmente mediante una regresión no lineal entre y y x_i , en una línea similar a la explicación previamente mostrada para S_i .

Para obtener la parte no correlacionada $S_i^{incorrelada}$, primero se realiza una regresión de x_i sobre el resto de variables, de manera que se obtenga una regresión multivariante para modelar x_i como función de las demás variables:

$$x_i = \beta_0 + \sum_{k \neq i} \beta_k x_k$$

donde los coeficientes β_k se obtienen mediante mínimos cuadrados. De esta manera, se define la varianza de x_i que puede ser explicada a través del resto de variables. Posteriormente, se calculan los residuos \hat{z}_i como:

$$\hat{z}_i = x_i - \hat{x}_i = x_i - (\beta_0 + \sum_{j \neq i} \beta_j x_j)$$

Al restar esta estimación \hat{x}_i del valor real de x_i , se aísla únicamente la parte de x_i que no puede ser explicada por las demás variables. Esto elimina la dependencia lineal entre x_i y x_k , puesto que toda la varianza de x_i que era explicada a través del resto de indicadores ha sido restada de su componente. Por tanto, el residuo \hat{z}_i representa la variabilidad que únicamente x_i puede explicar de y , que sigue siendo el indicador sintético.

Finalmente, se realiza la regresión de y sobre \hat{z}_i para obtener $S_i^{incorrelada}$ acometiendo la siguiente fórmula:

$$S_i^{incorrelada} = \frac{Var(E_{\hat{z}_i}[y|\hat{z}_i])}{Var(y)} = \frac{\sum_{j=1}^n (\hat{y}_j^i - \bar{y}^i)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

De esta manera, habiendo calculado $S_i^{incorrelada}$ y S_i , la obtención de $S_i^{correlada}$ resulta sencilla si recordamos la fórmula antes mencionada:

$$S_i = S_i^{incorrelada} + S_i^{correlada}$$

Una vez comprendido el proceso de cálculo de importancias, una cuestión que surge naturalmente es cómo ajustar los pesos de las variables para lograr una cierta importancia deseada. Esto se puede abordar mediante un algoritmo de optimización que minimice la

diferencia cuadrática entre una importancia deseada normalizada (S_i^*) y la importancia obtenida para los pesos optimizados (w_{opt}). El problema de optimización se formularía como:

$$w_{opt} = \underset{w}{\operatorname{argmin}} \sum_i^d (S_i^* - S_i(w))^2$$

donde S_i^* es el índice de sensibilidad deseado normalizado, y $S_i(w)$ se calcula para cada conjunto de pesos w como el índice de sensibilidad de primer orden (“importancia”) mencionado anteriormente (Becker et al., 2017). Nótese que, típicamente, el criterio experto asigna pesos en un intento de reflejar la “importancia” que se desea tengan los indicadores, por lo que dicho índice de sensibilidad objetivo realmente serían los pesos proporcionados por el panel de investigadores o expertos.

En resumen, este enfoque permite no solo analizar la influencia real de las variables en un indicador compuesto, sino también ajustar los pesos para alcanzar una configuración óptima. A continuación se analizarán los resultados de aplicar las fórmulas y teorías descritas sobre el SLMI.

b. Resultados

i. Desempleo

Como parte del análisis de importancia, se debe primero comprender la estructura de los datos y la potencial existencia de correlaciones no lineales que puedan no ser captadas por otros métodos de agregación como el Factor Analysis (FA).

Para ello, en primer lugar, se calculan las regresiones lineales y no lineales a través de splines de segundo grado que reflejen dichas no linealidades. Como ejemplo, se muestran las regresiones realizadas para la primera área del SLMI (“Desempleo”), donde la y es el indicador de área obtenido a partir de la agregación equiponderada de las variables base presentes en la tabla A2. Los indicadores base son las variables independientes de cada gráfica:

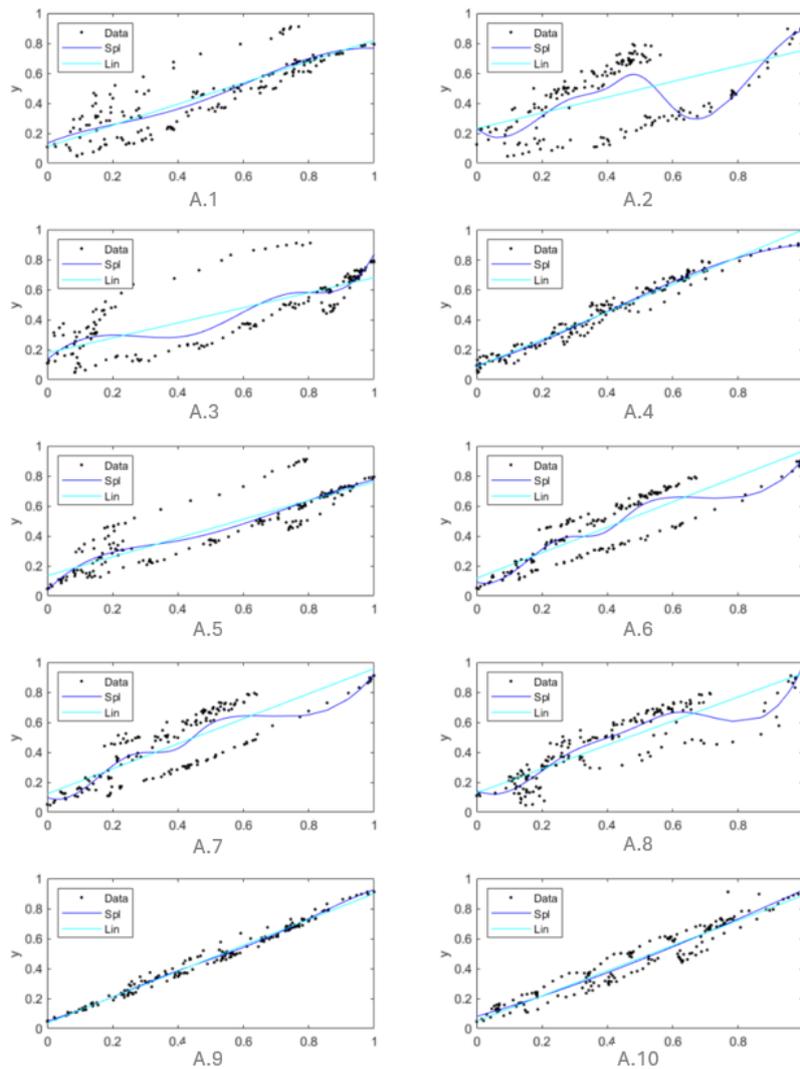


Ilustración 11. Distribuciones de los indicadores base de Desempleo frente al SLMi y sus ajustes lineal y por spline

Se observa claramente cómo la mayoría de las variables muestran un comportamiento razonablemente lineal, con la salvedad de la variable A.2 (Paro registrado agricultura y pesca) donde el spline se desvía en gran medida de la recta de correlación lineal. Si bien otras variables muestran una ligera variación, no se consideran suficientemente reseñables.

Nótese que, si bien la variable “Desempleo” ha sido recalculada de forma equiponderada (véase la ilustración A1 del anexo 2), el índice no variará en gran medida en tanto que los pesos proporcionados por el FA no distan en gran medida de aquellos otorgados a través de la equiponderación. Cabe mencionarse que, conceptualmente, resultaría extraño y difícilmente argumentable intentar asignar pesos óptimos a unos datos sobre los que se ha realizado un FA previamente, en tanto que las variables han perdido parte de su trazabilidad y los expertos encontrarán complicado proporcionar unos pesos deseados.

Retomando el análisis de importancia, a continuación se presenta el índice de primera sensibilidad para todas las variables de la categoría *Desempleo*, que es un reflejo de la correlación lineal y no lineal respecto del área estudiada, cuantificando el impacto mostrado en las gráficas mostradas anteriormente:

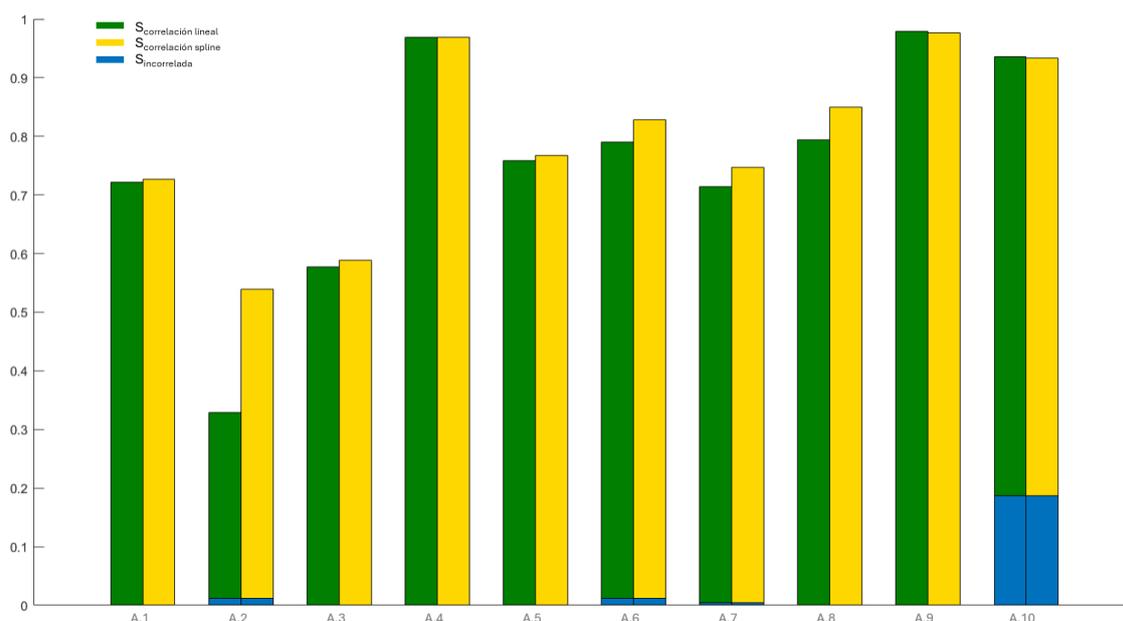


Ilustración 12. Índices de sensibilidad de primer orden sobre la categoría Desempleo

La variable A.2 es la que presenta la mayor desviación. No obstante esto se muestra con mayor detalle en la tabla que proporciona los valores precisos de cada hito en la ilustración 5:

Desempleo	S_{i_lineal}	S_{i_spline}	$S_{i_lineal}^{incorrelada}$	$S_{i_spline}^{incorrelada}$	$S_{i_lineal}^{correlada}$	$S_{i_spline}^{correlada}$
Paro Registrado (<25 años) A.1	0.7426	0.7464	0.0000	0.0000	0.7426	0.7464
Paro Registrado Agricultura y Pesca A.2	0.3089	0.5290	0.0119	0.0119	0.2970	0.5171
Paro Registrado Construcción A.3	0.5980	0.6327	0.0000	0.0000	0.5980	0.6327
Paro Registrado CVE A.4	0.9620	0.9630	0.0000	0.0000	0.9620	0.9630
Paro Registrado Industria A.5	0.7762	0.7838	0.0000	0.0000	0.7762	0.7838
Paro Registrado Mujeres A.6	0.7744	0.8151	0.0000	0.0000	0.7744	0.8151
Paro Registrado Servicios A.7	0.6977	0.7331	0.0038	0.0038	0.6939	0.7293
Paro Registrado Sin Empleo Previo A.8	0.7791	0.8396	0.0000	0.0000	0.7791	0.8396
Paro Registrado Varones A.9	0.9850	0.9810	0.0000	0.0000	0.9850	0.9810
Tasa de Paro Trimestral A.10	0.9281	0.9258	0.1898	0.1898	0.7383	0.7360

Tabla 7. Valores del índice de primera sensibilidad ("importancia") en la categoría Desempleo

De la gráfica y su correspondiente tabla de valores, se observa cómo la mayor variación entre la importancia obtenida linealmente y de spline se da precisamente en la variable A.2 ya mencionada, lo que refuerza el análisis inicial a través del análisis de correlación. Dada esta

variación, se considerará de ahora en adelante esta estimación por splines penalizados para el área “Desempleo”.

En segundo lugar, lo primordial que ha de valorarse es la existencia de una importancia incorrelada que apenas se muestra en el resto de las variables (0.1898). Nótese que la importancia, en tanto que es definida como el índice de sensibilidad de primer orden de x (indicador base) sobre y (indicador sintético), goza siempre de una componente correlada de x_i explicada a través del resto de variables $x_{\sim i}$ y una componente incorrelada de la variabilidad de y explicada a través de x_i . Esto está basado en el propio concepto de importancia: una variable no solo afecta sobre el conjunto según el peso recibido, sino también a través de aquellos otros indicadores con los que comparta comportamiento y , por tanto, correlación. Esto ha sido debidamente explicado en la sección de estado del arte.

Una vez comprendido el proceso a través del cual se han obtenido los valores de la tabla, se deduce que la variable A.10 puede ser explicada a partir de las demás de forma tan significativa como el resto de las variables, pero goza de una componente de incorrelación entre x_i e y que no aparece en el resto de indicadores. Por ello, de acuerdo con la teoría estudiada (Becker et al., 2017), sobre esta variable tendrá el mayor impacto la variación del peso que se pueda realizar a través de una optimización. Nótese que si $S_i = S_i^{correlada}$ el peso del indicador específico apenas tendrá impacto ya que los cambios en la importancia vendrán determinados por los movimientos en el resto de variables.

Cabe resaltar cómo el FA en cierta manera también revela esta tendencia, en tanto que le asigna un peso relativamente más bajo para minimizar el impacto de la variable. Si bien podría argüirse que esto se debe a una alta correlación entre este indicador y el resto de indicadores (aludiendo a $S_{10}^{correlada} = 0.7360$), parece que la existencia de una importancia incorrelada podría ser la clave de por qué el FA le asigna un peso menor que al resto, es decir, no porque el resto de indicadores no deban tomar otros valores, sino por la indiferencia estadística que supone cualquier tipo de variación sobre los mismos.

Después de todo, aunque el FA no considera el impacto sobre el sintético, sí que asienta sus resultados sobre la variabilidad del conjunto de datos que, *a posteriori*, dan forma al sintético. Por ello, el FA otorga un peso relativamente alto a A.2 y A.3, variables con menor correlación de x_i sobre $x_{\sim i}$ ($S_i^{correlada}$) a la par que minimiza A.10 por tratarse de la variable con una correlación mayor.

En cualquier caso, podría argüirse que el análisis de sensibilidad revela una capa más de información en el análisis multivariante de la estructura de datos estudiada.

A continuación, se debe comparar dicho índice de sensibilidad con el peso, no tanto en valores absolutos sino relativos para comparar cuán importante es cada variable sobre el área final. Para ello, se lleva a cabo un reescalado a 1 de la importancia obtenida por spline y se compara frente al peso original del SLMI, la varianza explicada por cada indicador base en el SLMI y también frente a la equiponderación con que ha sido construido el nuevo indicador *Desempleo* construido de forma equiponderada (figura A2 del anexo 2):

Desempleo	Peso asignado con FA	Índice de sensibilidad normalizado	Equiponderación
<i>Paro Registrado (<25 años) A.1</i>	0.1121	0.0939	0.1000
<i>Paro Registrado Agricultura y Pesca A.2</i>	0.1192	0.0665	0.1000
<i>Paro Registrado Construcción A.3</i>	0.1235	0.0796	0.1000

Paro Registrado CVE A.4	0.0717	0.1211	0.1000
Paro Registrado Industria A.5	0.1159	0.0986	0.1000
Paro Registrado Mujeres A.6	0.1084	0.1025	0.1000
Paro Registrado Servicios A.7	0.1124	0.0922	0.1000
Paro Registrado Sin Empleo Previo A.8	0.0908	0.1056	0.1000
Paro Registrado Varones A.9	0.0813	0.1234	0.1000
Tasa de Paro Trimestral A.10	0.0648	0.1165	0.1000

Tabla 8. Valores de importancia frente a los valores clave en la construcción de la categoría Desempleo

Queda por tanto patentes ciertas desviaciones en las aportaciones de cada variable al resultado final del área “Desempleo”. La mayor desviación se produce en las variables A.2, A.3, A.4, A.9 y A.10, todas con valores que se desvían más del 30% e incluso del 50% en los tres últimos casos. Como se ha mencionado, la importancia viene dominada en mayor medida por la correlación.

Este problema probablemente surge porque la parte correlacionada de S_i prevalece sobre la parte no correlacionada en todos los indicadores. Esto implica que las correlaciones, más que los pesos, son las que determinan la importancia, como se ha mencionado anteriormente. Considerando esta desviación, cabe preguntarse qué valores proporciona la matriz de correlaciones, en tanto que debería haber una conexión entre las variables altamente correladas y aquellas que gozan de una mayor importancia sobre el conjunto de datos:

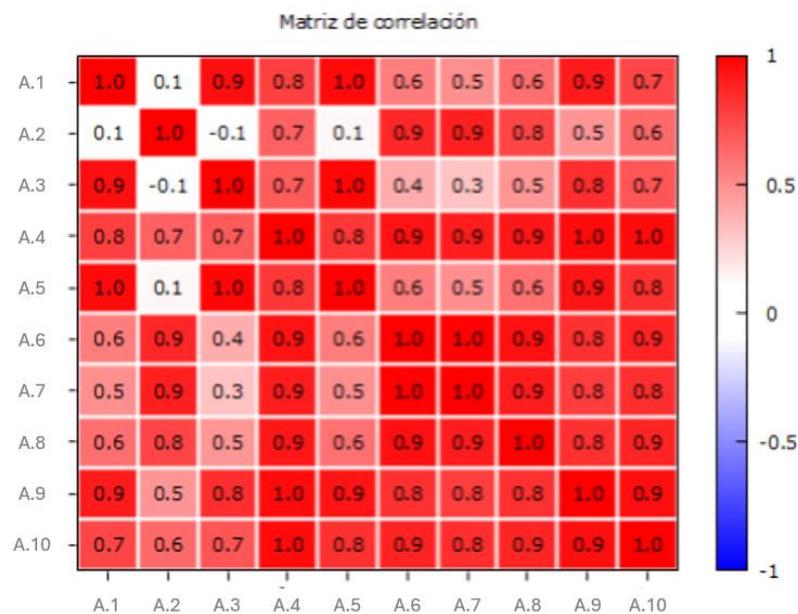


Ilustración 13. Matriz de correlaciones de la categoría Desempleo

Nótese cómo el área Desempleo está compuesta de indicadores altamente correlados entre ellos. No obstante, parece que la variable A.2 es la que menor correlación lineal muestra, lo cual fue también capturado por la ilustración y la tabla mostradas. Esto deja aún más patente la necesidad de considerar las correlaciones no lineales en tanto que, si bien en la tabla 6 A.2 muestra una S_i^{lineal} notablemente inferior al resto de variables, esto se reduce para el caso de los splines. Considerando además que la práctica totalidad de esta importancia procede de la importancia correlacionada de x_i frente a $x_{\sim i}$, se deduce que la matriz de correlaciones no capta adecuadamente las relaciones en este conjunto de datos.

Parece que, a mayor correlación, el FA busca asignar un menor peso para compensar aquella variabilidad que ya es explicada por otras variables. No obstante, en cuanto al impacto sobre la

importancia de cada variable, la variación del peso solo tendrá un efecto significativo en aquellos casos en los que la importancia no correlacionada goce de entidad, puesto que si no toda la importancia se obtendría a partir de las regresión lineal entre indicadores base y, por tanto, el resultado de alterar uno de sus pesos será marginal sobre el sintético si el impacto está mayormente definido a través de la variabilidad que es explicada a través del resto de variables.

En cualquier caso, esto ratifica la robustez del análisis llevado a cabo y permite proseguir con el análisis en torno a la búsqueda de unos pesos optimizados que minimicen la desviación entre la importancia deseada y la obtenida. Esto, en contraste con el FA donde los pesos son obtenidos estadísticamente, permite alinear los resultados con los deseos y preferencias del investigador, de manera que el criterio experto se mantiene vigente y se facilita la trazabilidad total de las variables y los sintéticos en su construcción.

A continuación, se muestra la ilustración 17 donde se observan los valores que podrían tomar los pesos para minimizar dicha desviación entre las importancias deseada y obtenida:

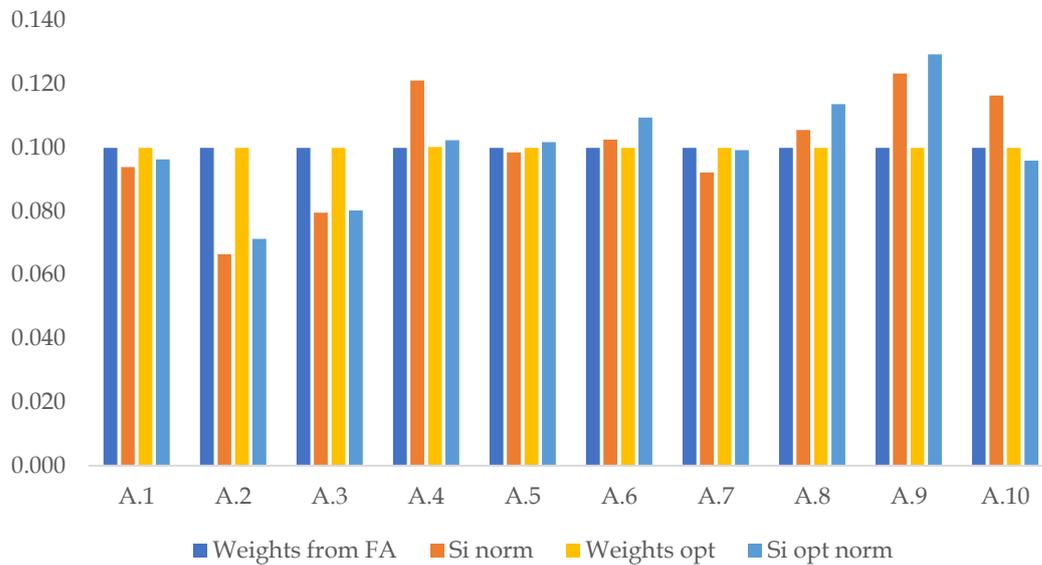


Ilustración 14. Resultados de la optimización para la categoría Desempleo

A partir de la ilustración mostrada cabe resaltar que el impacto de la optimización sobre la categoría es pequeño, en tanto que los pesos varían ligeramente y logran reducir las desviaciones en las variables A.2, A.3, A.4, A.7 y A.10. De hecho, la variación más notable se da en esta última variable, lo cual resulta consistente con lo explicado anteriormente. En cualquier caso, el impacto neto es positivo sobre el conjunto de datos y la optimización resulta exitosa calculando la desviación total de las importancia obtenida frente a la deseada como:

$$\sum_i^d (S_i^* - S_i(w))^2$$

donde S_i^* es la importancia deseada, que típicamente debería tomar los valores del peso asignados por los expertos, y $S_i(w)$ es la importancia obtenida para el set de pesos que minimizan dicho sumatorio.

De esta manera, el conjunto original mostraba una desviación de 0.054 frente a la nueva de 0.049.

ii. Empleo

A continuación, se presenta la distribución y ajustes lineal y spline del área “Empleo”:

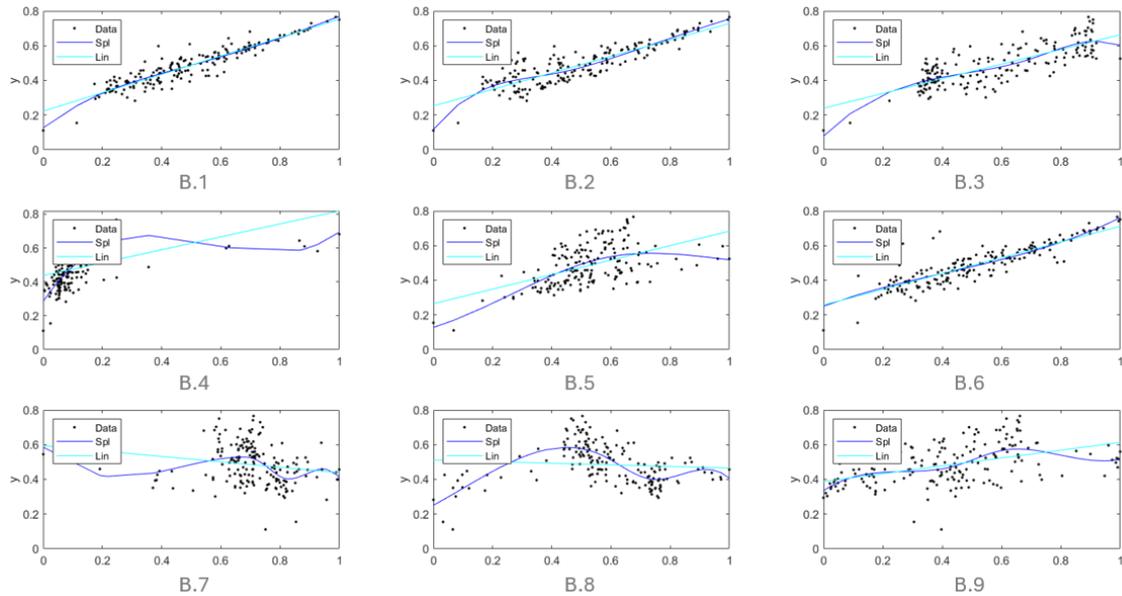


Ilustración 15. Distribuciones de los indicadores base de Empleo frente al SLMi y sus ajustes lineal y por spline

En este caso, si bien la mayoría de las variables se ajustan linealmente, también se da el caso de B.4 y B.8 donde el ajuste por spline resultará en un reflejo más fiel de la correlación entre las demás variables. Esto queda patente en la siguiente ilustración, donde la mayor variación entre ambos ajustes se observa precisamente sobre dichos indicadores:

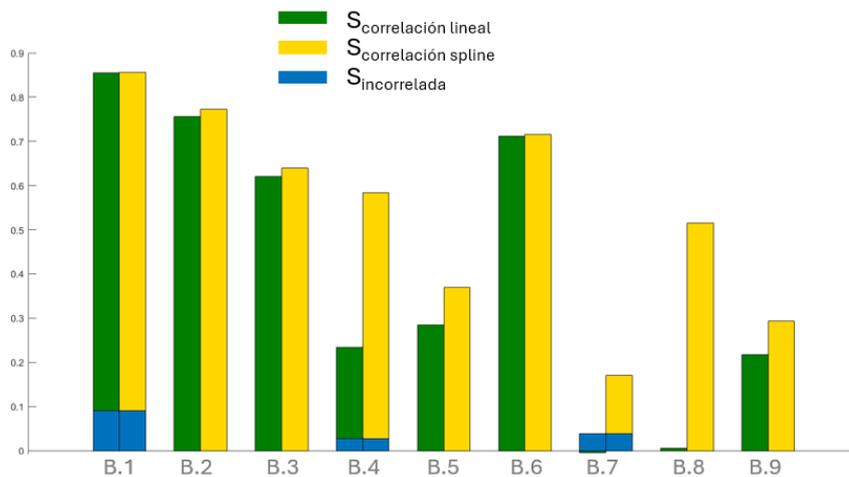


Ilustración 16. Índices de sensibilidad de primer orden sobre la categoría Empleo

De esta manera, los datos mostrados en la ilustración 19 se muestran desglosados en la siguiente tabla:

Empleo	S_{lineal}	S_{Spline}	$S_{\text{lineal incorrelada}}$	$S_{\text{Spline incorrelada}}$	$S_{\text{lineal correlada}}$	$S_{\text{Spline correlada}}$
Contratos Registrados B.1	0.8545	0.8561	0.0912	0.0912	0.7633	0.7649

<i>Contratos Registrados a Tiempo Parcial B.2</i>	0.7556	0.7729	0.0000	0.0000	0.7556	0.7729
<i>Contratos Registrados CVE B.3</i>	0.6202	0.6389	0.0000	0.0000	0.6202	0.6389
<i>Contratos Registrados Indefinidos B.4</i>	0.2339	0.5837	0.0275	0.0275	0.2064	0.5561
<i>Contratos Registrados Interinidades B.5</i>	0.2851	0.3700	0.0000	0.0000	0.2851	0.3700
<i>Contratos Registrados Temporales B.6</i>	0.7124	0.7161	0.0000	0.0000	0.7124	0.7161
<i>Productividad por hora efectivamente trabajada B.7</i>	0.0347	0.1712	0.0391	0.0391	(0.0044)	0.1321
<i>Productividad por puesto de trabajo a tiempo completo B.8</i>	0.0064	0.5148	0.0000	0.0000	0.0064	0.5148
<i>Tasa de empleo B.9</i>	0.2173	0.2940	0.0000	0.0000	0.2173	0.2940

Tabla 9. Valores de importancia frente a los valores clave en la construcción de la categoría Empleo

En el caso de esta categoría, cabe destacar cómo B.1, B.4 y B.7 gozan de la mayor parte incorrelada del conjunto, aunque también resalta cómo los indicadores B.5, B.7, B.8 y B.9 presentan importancias pequeñas, aunque notoriamente más altas cuando son evaluadas mediante splines. Esto revela correlaciones no lineales entre las variables y el área de *Empleo*. Por ello, también en este caso se evaluará la importancia como aquellos valores derivados de la regresión por splines penalizados. De hecho, al compararlas con la matriz de correlaciones, se observa cómo estas variables presentan de hecho ligeras variaciones:

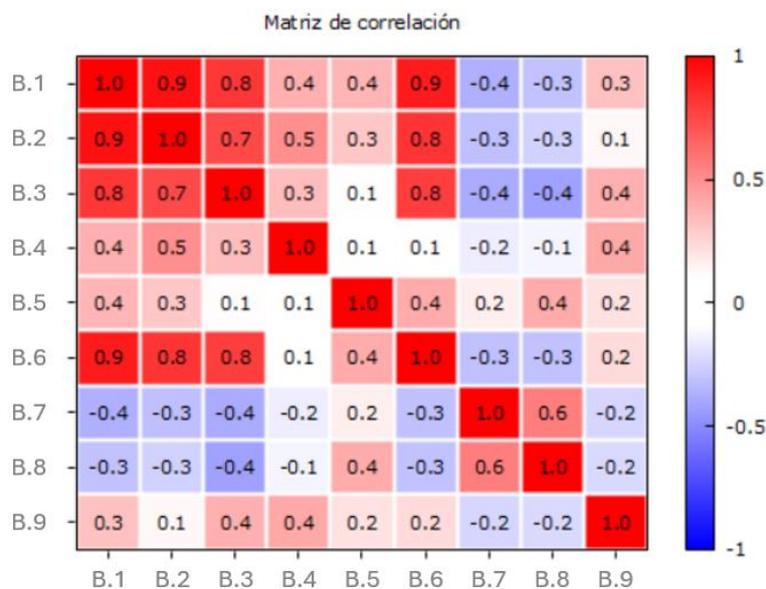


Ilustración 17. Matriz de correlaciones de la categoría Empleo

Una vez más, las importancias y las correlaciones lineales de la matriz de correlaciones están vinculadas, en tanto que las últimas variables de la categoría muestran importancias y correlaciones reducidas.

Empleo	Peso asignado con FA	Índice de sensibilidad normalizado	Equiponderación
<i>Contratos Registrados B.1</i>	0.1410	0.1741	0.1111
<i>Contratos Registrados a Tiempo Parcial B.2</i>	0.1291	0.1572	0.1111
<i>Contratos Registrados CVE B.3</i>	0.1005	0.1299	0.1111
<i>Contratos Registrados Indefinidos B.4</i>	0.1055	0.1187	0.1111
<i>Contratos Registrados Interinidades B.5</i>	0.0837	0.0752	0.1111
<i>Contratos Registrados Temporales B.6</i>	0.1457	0.1456	0.1111
<i>Productividad por hora efectivamente trabajada B.7</i>	0.0743	0.0348	0.1111
<i>Productividad por puesto de trabajo a tiempo completo B.8</i>	0.1139	0.1047	0.1111
<i>Tasa de empleo B.9</i>	0.1062	0.0598	0.1111

Tabla 10. Valores de importancia frente a los valores clave en la construcción de la categoría Empleo

Queda abierto para futuras líneas de investigación por qué FA en este pondera en demasía la variable B.1 siendo esta de las que gozan más importancia. Posiblemente, el análisis de importancia no resulte consistente con un alto número de variables.

A continuación, se presenta el ejercicio de optimización para la categoría de empleo:

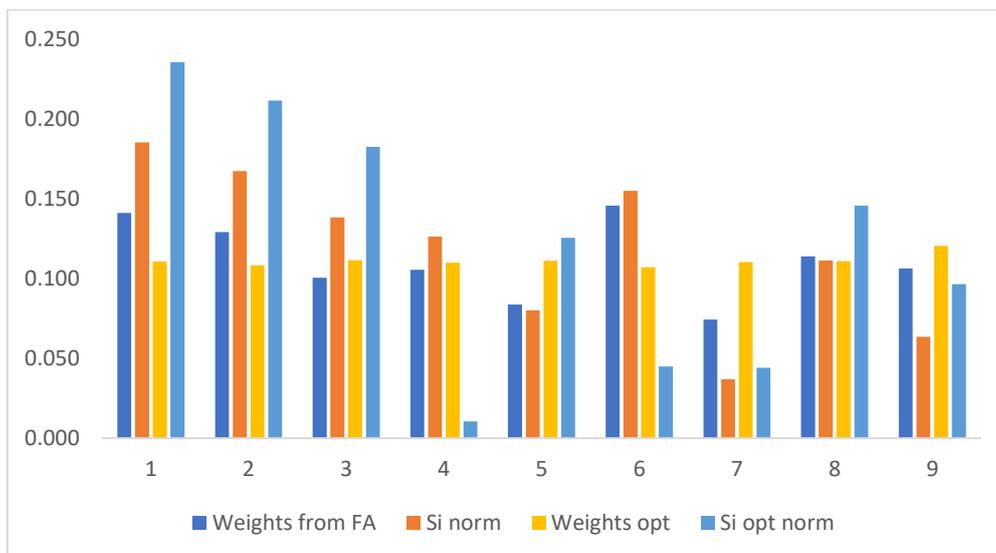


Ilustración 18. Resultados de la optimización para la categoría Empleo

Los datos son presentados cuantitativamente en la tabla resumen al final de la sección, donde se presentan las conclusiones y se abren nuevas líneas de investigación para futuros trabajos de análisis metodológico.

iii. Protección a los desempleados

A continuación, se presenta la distribución y ajustes lineal y spline del área *Protección a los desempleados*, que no fue ponderada mediante FA sino a través de equiponderación por criterio experto:

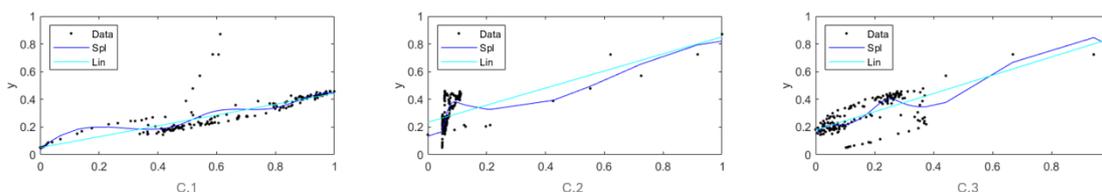


Ilustración 19. Distribuciones de los indicadores base de Protección a los desempleados frente al SLMI

Queda patente que C.2 debe ser ajustado por splines para capturar su impacto sobre la categoría, lo cual también es ratificado por la ilustración que muestra los valores de importancia para uno y otro método:

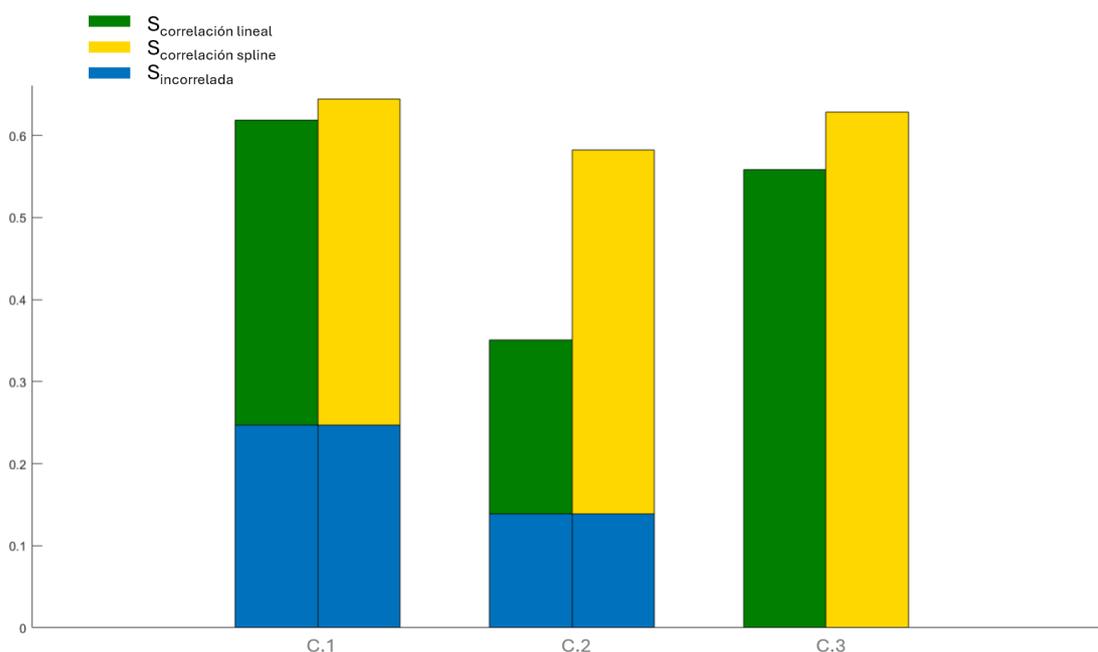


Ilustración 20. Índices de sensibilidad de primer orden sobre la categoría Protección a los desempleados

Cabe resaltar esta categoría, en la que de forma notable queda patente que este mecanismo habría sido metodológicamente más robusto que la simple equiponderación que se aplicó en la construcción original del SLMI. Por supuesto, si bien el impacto sobre el SLMI será reducido tanto por la escasa desviación frente a la importancia deseada como porque dicha área es la que menos pondera en el segundo nivel de agregación (10% frente al 30% del resto de áreas), desde el punto de vista teórico esto permite extraer ciertas conclusiones al respecto.

Los datos cuantitativos se presentan en la siguiente tabla:

Salarios y costes laborales	S_lineal	S_Spline	S_lineal incorrelada	S_Spline incorrelada	S_lineal correlada	S_Spline correlada
Beneficiarios Asistenciales	0.6192	0.6448	0.2472	0.2472	0.3720	0.3977
Prestaciones Contributivas Desempleo Parcial	0.3506	0.5822	0.1385	0.1385	0.2121	0.4437
Prestaciones Contributivas Desempleo Total	0.5587	0.6284	0.0000	0.0000	0.5587	0.6284

Tabla 11. Valores de importancia frente a los valores clave en la construcción de Protección a los desempleados

Nótese cómo en esta categoría la importancia incorrelada goza de una importancia mucho mayor en las dos primeras variables, lo cual se ve además ratificado por su matriz de correlaciones:

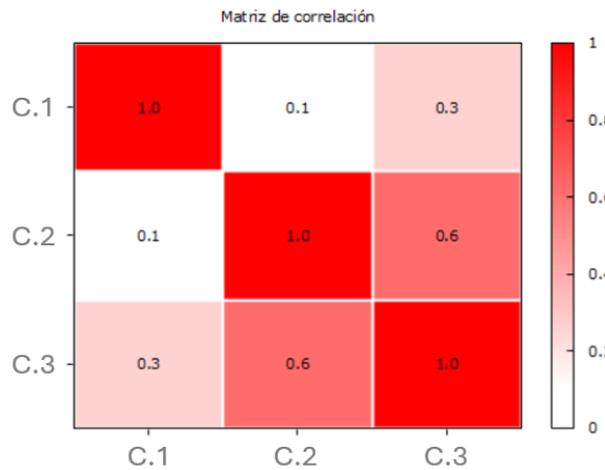


Ilustración 21. Matriz de correlaciones de la categoría de Protección a los desempleados

En este caso donde no se ha aplicado FA sobre el set de datos, parece que las dos primeras variables, visiblemente menos correladas tanto a través de la matriz de correlaciones como del análisis de importancia, serán aquellas donde la optimización podrá mejorar de forma más notable los resultados. Así, se podría proporcionar unos pesos que, sin ser equiponderados, proporcionen una importancia "equiponderada" a los tres indicadores.

De esta manera, se realiza el análisis mencionado y, efectivamente, se logra exitosamente obtener una importancia exactamente igual a la deseada:

	Equiponderación	Si	Si norm	Pesos optimizados	Si optimizado	Si optimizado norm	
Protección a los desempleados - C	C.1	0.333	0.645	0.348	0.332	0.630	0.333
	C.2	0.333	0.582	0.314	0.392	0.630	0.333
	C.3	0.333	0.628	0.339	0.275	0.630	0.333

Tabla 12. Resultados del análisis de sensibilidad de primer orden y su optimización para la categoría C

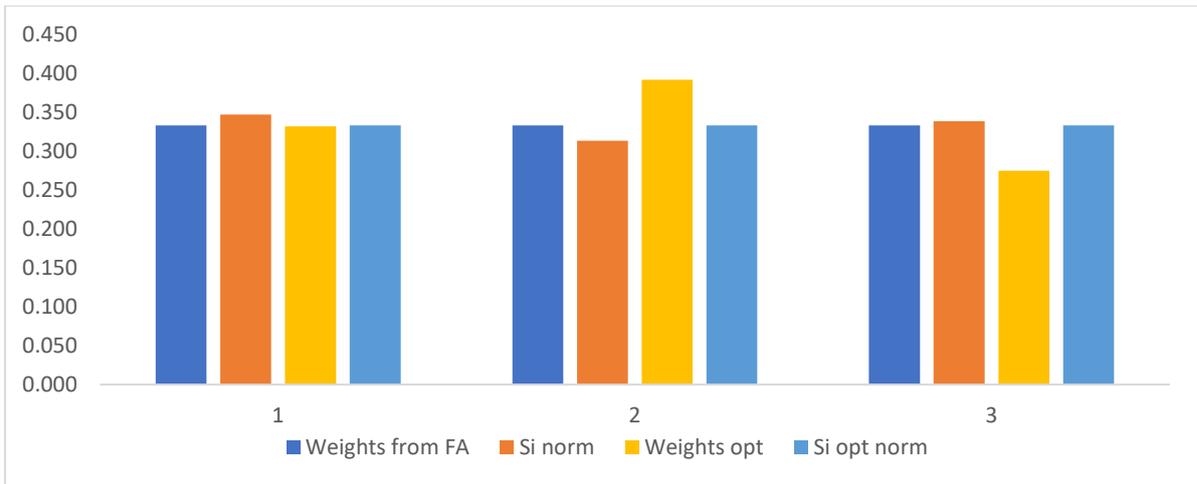


Ilustración 22. Resultados de la optimización para la categoría Protección a los desempleados

Queda por tanto patente que el análisis de importancia es muy exitoso allí donde el criterio experto puede ser usado y la dimensionalidad de los datos. En este caso, el panel de expertos presente en la elaboración del indicador habría de haber otorgado los pesos (0.332, 0.392, 0.275) para lograr la misma importancia (0.333) en los tres indicadores.

iv. Salarios y costes laborales

Los resultados de la categoría se presentan a continuación, siguiendo la línea de lo presentado anteriormente:

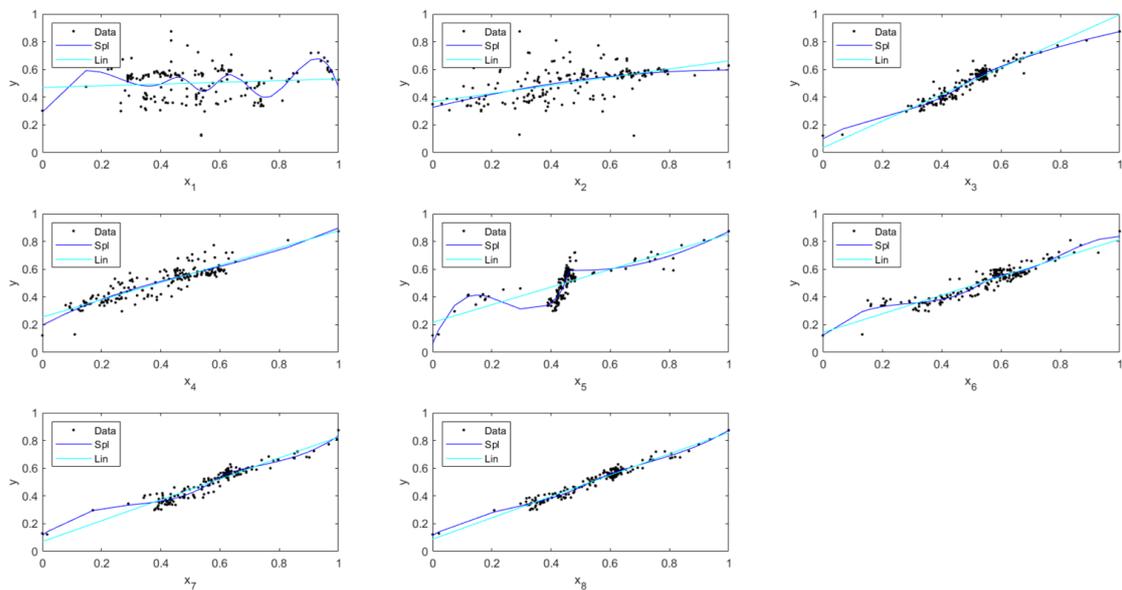


Ilustración 23. Distribuciones de los indicadores base de Salarios y costes laborales frente al SLMI

En este caso, nótese que las variables D.1 y D.5 presentan la mayor diferencia entre las regresiones lineal y por splines, lo cual es ratificado cuantitativamente por la siguiente gráfica:

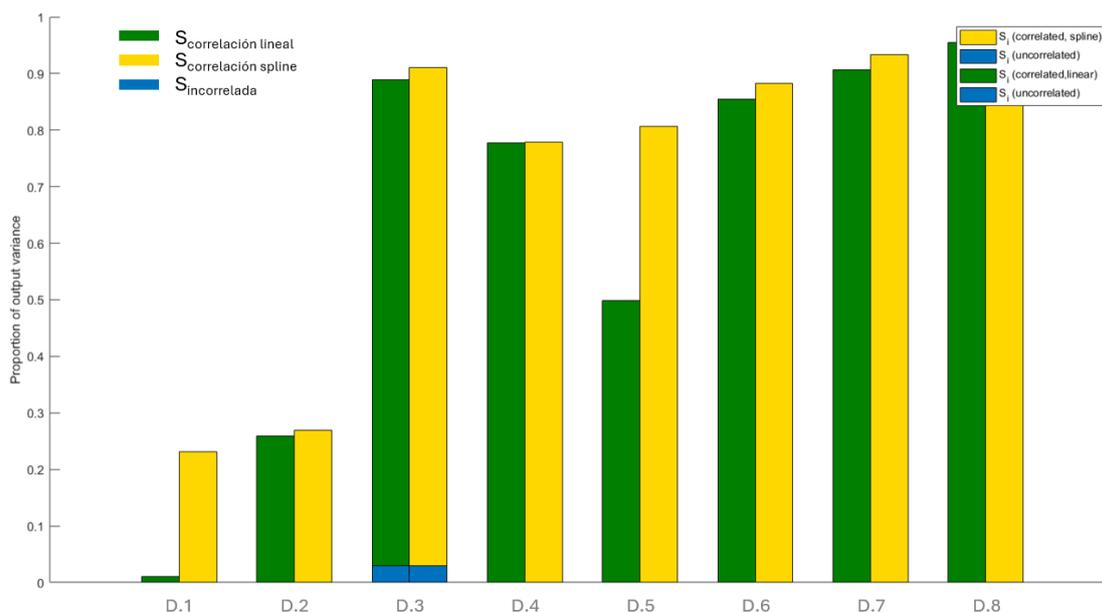


Ilustración 24. Índices de sensibilidad de primer orden sobre la categoría Salarios y costes laborales

En este caso, dada la diferencia visible tanto en D.1 como D.5 entre los métodos regresivos, resulta necesario una vez más asumir como la importancia de las variables la regresión por spline penalizados, que queda definida en la siguiente tabla:

Salarios y costes laborales	S_lineal	S_Spline	S_lineal incorrelada	S_Spline incorrelada	S_lineal correlada	S_Spline correlada
<i>Incremento Salarial Pactado en Negociación Colectiva D.1</i>	0.0105	0.2313	0.0000	0.0000	0.0105	0.2313
<i>Variación Interanual de los Rendimientos del Trabajo Agricultura y Ganadería D.2</i>	0.2598	0.2697	0.0000	0.0000	0.2598	0.2697
<i>Variación Interanual de los Rendimientos del Trabajo Comercio D.3</i>	0.8890	0.9112	0.0295	0.0295	0.8596	0.8817
<i>Variación Interanual de los Rendimientos del Trabajo Construcción D.4</i>	0.7779	0.7791	0.0000	0.0000	0.7779	0.7791
<i>Variación Interanual de los Rendimientos del Trabajo Hostelería D.5</i>	0.4986	0.8070	0.0000	0.0000	0.4986	0.8070
<i>Variación Interanual de los Rendimientos del Trabajo Industria D.6</i>	0.8553	0.8822	0.0000	0.0000	0.8553	0.8822
<i>Variación Interanual de los Rendimientos del Trabajo Servicios D.7</i>	0.9067	0.9333	0.0000	0.0000	0.9067	0.9333

<i>Variación Interanual de los Rendimientos del Trabajo Total D.8</i>	0.9549	0.9563	0.0000	0.0000	0.9549	0.9563
---	--------	--------	--------	--------	--------	--------

Tabla 13. Valores de importancia frente a los valores clave en la construcción de Salarios y costes laborales

De esta manera, comparando los valores con la matriz de correlaciones, se observa que la primera variable, si bien no goza de importancia alguna con el método lineal, sí se materializa su impacto en cierto modo a través de los splines, acercándose su importancia a la de la variable B.2. En cualquier caso, ambos indicadores gozan de una importancia total reducida y correlada.

Esto queda patente en la matriz de correlaciones:

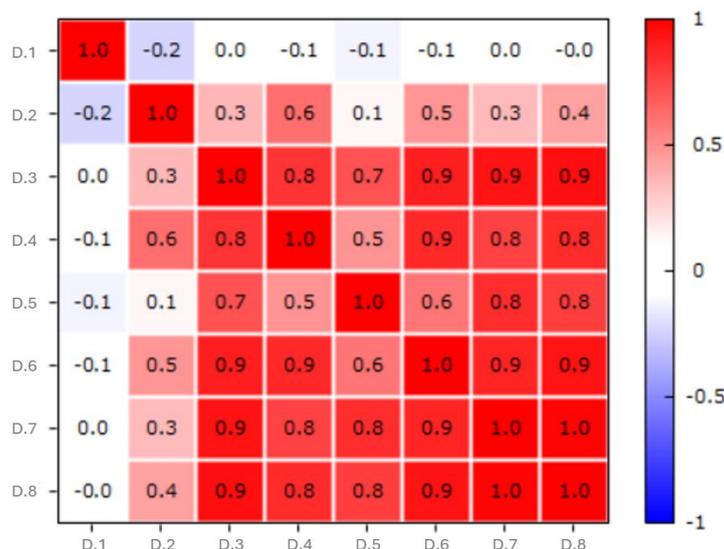


Ilustración 25. Matriz de correlaciones de la categoría Salarios y costes laborales

En este caso, al igual que en el área de *Empleo*, el FA no parece verse influenciado en gran medida por las correlaciones mostradas en la matriz de correlaciones. Realmente, el hecho de que no proporcione un mayor peso a la variable D.1 y D.2 deriva en que su importancia sobre el indicador final es despreciable, como se muestra a continuación:

Salarios y costes laborales	Peso asignado con FA	Indice de sensibilidad reescalada	Equiponderación
<i>Incremento Salarial Pactado en Negociación Colectiva D.1</i>	0.1081	0.0401	0.125
<i>Variación Interanual de los Rendimientos del Trabajo Agricultura y Ganadería D.2</i>	0.0916	0.0467	0.125
<i>Variación Interanual de los Rendimientos del Trabajo Comercio D.3</i>	0.1514	0.1579	0.125
<i>Variación Interanual de los Rendimientos del Trabajo Construcción D.4</i>	0.1097	0.1350	0.125
<i>Variación Interanual de los Rendimientos del Trabajo Hostelería D.5</i>	0.0968	0.1399	0.125
<i>Variación Interanual de los Rendimientos del Trabajo Industria D.6</i>	0.1327	0.1529	0.125

Variación Interanual de los Rendimientos del Trabajo Servicios D.7	0.1543	0.1618	0.125
Variación Interanual de los Rendimientos del Trabajo Total D.8	0.1556	0.1657	0.125

Tabla 14. Valores de importancia frente a los valores clave en la construcción de la categoría Salarios y costes laborales

De esta manera, se alcanzan los resultados de la optimización para esta categoría:

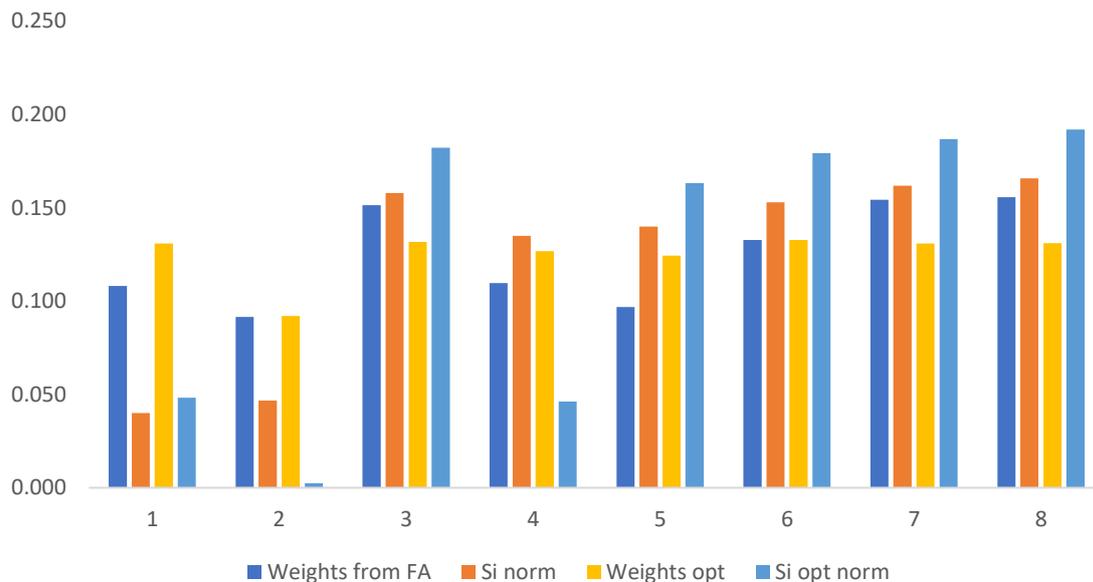


Ilustración 26. Resultados de la optimización para la categoría Salarios y costes laborales

Cabe resaltar el comportamiento anómalo de la importancia optimizada en este caso, denotando una vez más la complejidad de la optimización con un conjunto de variables elevado, lo que desaconseja el uso del análisis de importancia. En cualquier caso, esto queda abierto para futuros trabajos.

Para concluir se presentan los resultados de todas las categorías presentes en el SLMi, donde se integran además los resultados de la categoría *Salarios y costes laborales*:

		Pesos equiponderados sobre las áreas del SLMi					
		Equiponderación	Si	Si norm	Pesos optimizados	Si optimizado	Si optimizado norm
Desempleo – A	A.1	0.100	0.746	0.094	0.100	0.726	0.096
	A.2	0.100	0.529	0.067	0.100	0.538	0.071
	A.3	0.100	0.633	0.080	0.100	0.605	0.080
	A.4	0.100	0.963	0.121	0.098	0.772	0.102
	A.5	0.100	0.784	0.099	0.100	0.768	0.102
	A.6	0.100	0.815	0.103	0.101	0.826	0.110
	A.7	0.100	0.733	0.092	0.102	0.748	0.099
	A.8	0.100	0.840	0.106	0.102	0.858	0.114
	A.9	0.100	0.981	0.123	0.101	0.976	0.129
	A.10	0.100	0.926	0.116	0.097	0.724	0.096
Empleo – B	B.1	0.111	0.856	0.185	0.111	0.784	0.216
	B.2	0.111	0.773	0.167	0.111	0.710	0.196
	B.3	0.111	0.639	0.138	0.111	0.603	0.166
	B.4	0.111	0.584	0.126	0.114	0.034	0.009
	B.5	0.111	0.370	0.080	0.111	0.414	0.114
	B.6	0.111	0.716	0.155	0.110	0.156	0.043

	B.7	0.111	0.171	0.037	0.111	0.145	0.040
	B.8	0.111	0.515	0.111	0.111	0.483	0.133
	B.9	0.111	0.294	0.064	0.111	0.302	0.083
Protección a los desempleados - C	C.1	0.333	0.645	0.348	0.332	0.630	0.333
	C.2	0.333	0.582	0.314	0.392	0.630	0.333
	C.3	0.333	0.628	0.339	0.275	0.630	0.333
Costes laborales - D	D.1	0.125	0.231	0.040	0.124	0.225	0.046
	D.2	0.125	0.270	0.047	0.124	0.022	0.004
	D.3	0.125	0.911	0.158	0.124	0.886	0.180
	D.4	0.125	0.779	0.135	0.135	0.263	0.054
	D.5	0.125	0.807	0.140	0.124	0.795	0.162
	D.6	0.125	0.882	0.153	0.124	0.878	0.179
	D.7	0.125	0.933	0.162	0.124	0.908	0.185
	D.8	0.125	0.956	0.166	0.124	0.937	0.191

Tabla 15. Resultados para todas las categorías

v. Análisis de importancia sobre el último nivel de agregación

En cualquier caso, también cabe plantearse el escenario en que todas las variables sean incorporadas directamente sobre el SLMI, obviando el paso intermedio de subagregación en torno a las áreas mencionadas. Asumiendo que todos los indicadores base gozan del mismo peso en un contexto de equiponderación, cada uno se vería asignado con 0.033 en tanto que el SLMI está conformado por 30 indicadores base y agregado aritméticamente.

En pos de agilizar el análisis, se presenta a continuación la ilustración que muestra los pesos originales (0.033 para todos los indicadores), las importancias por splines de dichos 30 indicadores frente al SLMI calculado de forma equiponderada directamente y el resultado de su optimización:

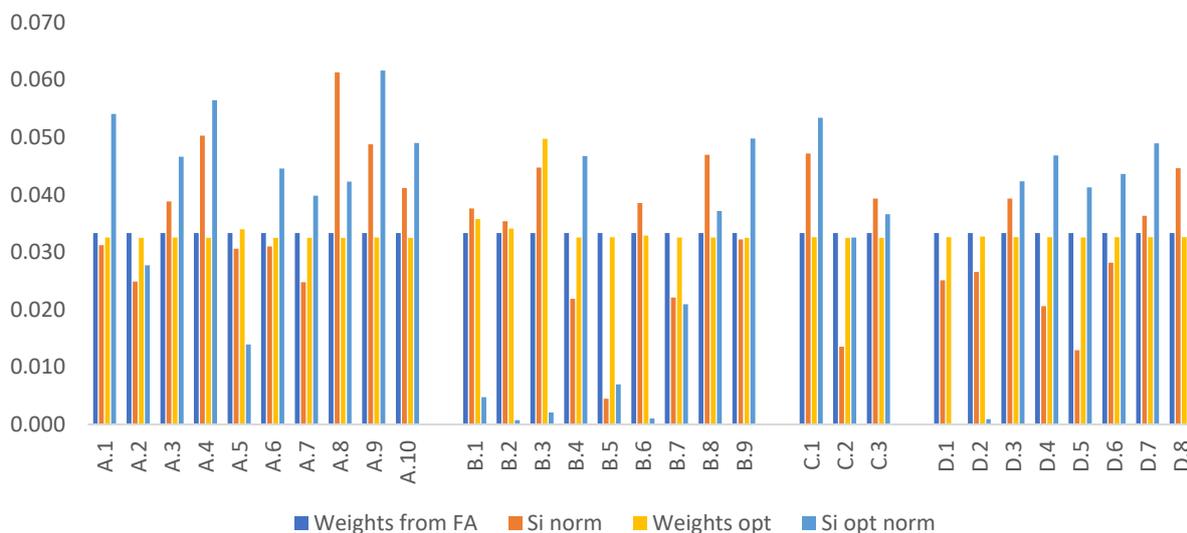


Ilustración 27. Resultados de la optimización para el SLMI equiponderado a partir de todos los indicadores base

De esta manera, se obtendría la siguiente tabla en la que se comparan, una vez más, los pesos proporcionados junto a su importancia frente a los pesos optimizados y su correspondiente importancia optimizada.

Pesos equiponderados agregando todos los indicadores base en torno al SLMI

	Equiponderación	Si	Si norm	Pesos optimizados	Si optimizado	Si optimizado norm	
Desempleo – A	A.1	0.033	0.326	0.031	0.033	0.811	0.054
	A.2	0.033	0.260	0.025	0.032	0.415	0.028
	A.3	0.033	0.406	0.039	0.033	0.699	0.047
	A.4	0.033	0.526	0.050	0.033	0.846	0.056
	A.5	0.033	0.320	0.031	0.034	0.209	0.014
	A.6	0.033	0.324	0.031	0.033	0.668	0.045
	A.7	0.033	0.259	0.025	0.033	0.597	0.040
	A.8	0.033	0.641	0.061	0.032	0.634	0.042
	A.9	0.033	0.510	0.049	0.033	0.924	0.062
	A.10	0.033	0.431	0.041	0.032	0.734	0.049
Empleo – B	B.1	0.033	0.394	0.038	0.036	0.071	0.005
	B.2	0.033	0.370	0.035	0.034	0.011	0.001
	B.3	0.033	0.468	0.045	0.050	0.031	0.002
	B.4	0.033	0.228	0.022	0.033	0.700	0.047
	B.5	0.033	0.047	0.004	0.033	0.104	0.007
	B.6	0.033	0.403	0.039	0.033	0.015	0.001
	B.7	0.033	0.231	0.022	0.033	0.313	0.021
	B.8	0.033	0.491	0.047	0.033	0.557	0.037
	B.9	0.033	0.337	0.032	0.033	0.746	0.050
Protección a los desempleados – C	C.1	0.033	0.493	0.047	0.033	0.800	0.053
	C.2	0.033	0.142	0.014	0.033	0.488	0.033
	C.3	0.033	0.411	0.039	0.032	0.549	0.037
Costes laborales – D	D.1	0.033	0.262	0.025	0.033	0.000	0.000
	D.2	0.033	0.278	0.027	0.033	0.014	0.001
	D.3	0.033	0.411	0.039	0.033	0.635	0.042
	D.4	0.033	0.215	0.021	0.033	0.702	0.047
	D.5	0.033	0.135	0.013	0.033	0.619	0.041
	D.6	0.033	0.294	0.028	0.033	0.654	0.044
	D.7	0.033	0.380	0.036	0.033	0.734	0.049
	D.8	0.033	0.466	0.045	0.033	0.713	0.048

Tabla 16. Resultados derivados de aplicar la optimización sobre todas las variables juntas

Al igual que en el caso anterior, la mejora más notable se da en el caso de la categoría de *Protección a los desempleados* puesto que, si bien aleja ligeramente la variable C.1, tanto C.2 y C.3 se acercan al objetivo de 0.033, redundando en unos residuos menores que ratifican por tanto la minimización realizada. No obstante, la mejora no es tan notoria como en el caso estudiado anteriormente, de lo que se deduce que, efectivamente, el análisis de importancia no resulta adecuado para lidiar con una gran cantidad de datos donde la misma está altamente influenciada por la correlación. Parece destacarse que, aquellas categorías con un grado alto de correlación en su mayoría lineal, el FA proporciona mayores ventajas que el análisis de importancia, en tanto que puede reducir la dimensionalidad a partir de la varianza explicada (aunque en el SLMI no se aplica esta reducción) y proporciona al autor unos pesos cuando se lidia con un número de variables demasiado elevado. El análisis de importancia no es capaz de proporcionar unos pesos que de forma notable alteren la estructura del sintético final, a menos que se den sobre un número reducido de variables como ocurrió en el caso de la categoría *Protección a los desempleados*.

Esto, además, puede verse ratificado por la otra agregación que en el caso del SLMI se realiza con criterio experto: la agregación de las áreas (agregación intermedia) en torno al SLMI, donde se asignan pesos del 30% a *Desempleo*, *Empleo* y *Costes laborales* y un 10% a *Protección a los desempleados*. Evaluar el impacto del análisis de importancia sobre esta agregación podría ratificar las conclusiones extraídas a lo largo del estudio.

A continuación, se proporciona la tabla con la optimización de las áreas frente al SLMI:

	Pesos	Si	Si norm	Pesos opt	Si opt	Si opt norm
Desempleo – A	0.300	0.528	0.344	0.156	0.686	0.256
Empleo – B	0.300	0.284	0.185	0.301	0.741	0.276
Protección a los desempleados – C	0.100	0.383	0.250	0.304	0.565	0.211
Costes laborales – D	0.300	0.339	0.221	0.239	0.689	0.257

Tabla 17. Resultados de la optimización para las áreas del SLMI

Excepto por el área *Desempleo*, cuya distancia frente a la importancia deseada no varía, el resto de las categorías presentan una importancia optimizada mucho más fiel a los pesos otorgados por el criterio experto. De esta forma, parece que allí donde el criterio experto impera, el análisis de importancia es una herramienta altamente valiosa para la ponderación.

4. Conclusiones

En lo relativo a la imputación, trabajar con el conjunto completo de datos para la imputación ofrece varias ventajas clave. En primer lugar, el modelo tiene acceso a una mayor cantidad de información, lo que permite capturar relaciones no lineales y patrones complejos que podrían perderse si solo se considera información local. Esto facilita una imputación más precisa al aprovechar relaciones globales entre variables. Además, la variabilidad entre diferentes áreas o subáreas se reduce, ya que los errores individuales tienden a balancearse, lo que genera resultados más estables y menos influenciados por la escasez de datos en grupos pequeños.

Otra ventaja importante es que la imputación global maneja mejor la distribución desigual de datos faltantes, reduciendo el riesgo de sesgos que pueden surgir si se imputan de forma local. También permite capturar mejor las correlaciones entre áreas o subáreas, lo que resulta en imputaciones más precisas cuando existen patrones comunes. Finalmente, la imputación a nivel local puede generar sobreajuste si las áreas o subáreas tienen pocos datos, haciendo que el modelo se adapte demasiado a características específicas y comprometa su capacidad de generalización. Al utilizar el set completo, se evita este riesgo y se obtiene una imputación más robusta y representativa del conjunto de datos en su totalidad.

Retomando la ponderación, ambos análisis estudiados proporcionan información valiosa sobre la estructura de datos estudiada, en particular si se considera que el FA no suele ir acompañado de un estudio de correlación no lineal. La combinación de ambos enfoques permite obtener una visión más completa de la relación entre variables y de las estructuras subyacentes que estas conforman.

Por otro lado, el análisis de importancia no sustituye al FA en aquellos casos donde las correlaciones entre variables son muy altas y el número de estas en la estructura de datos es considerablemente grande. En tales situaciones, la capacidad de reducción de variables que proporciona el FA es difícilmente compensable por las ventajas que ofrece el análisis de importancia. En particular, el FA es especialmente útil cuando se busca identificar la realidad subyacente que es descrita por varios indicadores. Un ejemplo de esto se puede observar en el caso del SLMi, donde el criterio experto aplicado en el último nivel de agregación no se basa en los indicadores individuales que componen cada área, sino en la realidad subyacente identificada a través del FA. Esto sugiere que las áreas dentro del SLMi no deben entenderse como una simple suma de sus indicadores, sino como una representación más profunda de una estructura latente.

En los casos en los que comúnmente se aplica el criterio experto, ya sea por la gran cantidad de variables presentes o por el interés en garantizar la trazabilidad de los datos, el análisis de importancia resulta particularmente adecuado. Este enfoque logra capturar con éxito los objetivos de los autores y proporciona resultados más fieles al criterio experto buscado.

Debido a estas características, la aplicación del análisis de importancia puede resultar especialmente útil en aquellos casos en los que los indicadores presentan múltiples niveles de agregación y emplean FA en el primer nivel, pero recurren al criterio experto en niveles posteriores. En este contexto, el análisis de importancia contribuye a evaluar las realidades subyacentes capturadas por el FA, permitiendo que el criterio experto se aplique sobre un conjunto de áreas más reducido y manejable.

Si bien en el SLMi no se redujo el número de indicadores mediante el cálculo de factores siguiendo una estructura de agregación jerárquica, como proponen Szopik-Depczynska et al. (2018), esta es una posibilidad a considerar en futuras repeticiones del análisis. En cualquier caso, dentro del SLMi se optó por reducir el número de variables mediante una agregación intermedia en torno a áreas específicas que, posteriormente, recibirían el criterio experto para su integración final dentro del SLMi. Es importante destacar que aplicar el análisis de importancia

directamente sobre 30 indicadores implica, de manera implícita, asignarles un peso relativo dentro del indicador final. Durante el estudio del SLMI, se evidenciaron las limitaciones del criterio experto en este tipo de escenarios, ya que su aplicación en un gran número de indicadores puede desvirtuar la capacidad de los expertos para asignar pesos de manera confiable, tal como lo advierten Saisana y Saltelli (2011).

El análisis de importancia ha sido replicado en distintos niveles dentro del SLMI. En primer lugar, se aplicó al primer nivel de agregación dentro de cada área, luego al segundo nivel de agregación en torno al SLMI y, finalmente, considerando todas las variables de manera directa sobre el SLMI. Para cada una de estas iteraciones, se obtuvo un conjunto de pesos optimizados que, a través de un análisis de sensibilidad, permitió generar diferentes configuraciones del SLMI. Esto facilitó la evaluación del impacto que tendrían diversas decisiones metodológicas sobre el indicador final.

A lo largo del estudio, se realizaron múltiples iteraciones con el propósito de obtener pesos óptimos bajo diversas configuraciones, aplicando también la imputación de datos sobre el conjunto total de variables. Este enfoque permitió analizar la robustez del SLMI en función de diferentes técnicas de imputación y ponderación. En este sentido, el análisis de importancia se replicó en varios niveles dentro del SLMI: en el primer nivel de agregación para cada área, en el segundo nivel de agregación en torno al SLMI y, finalmente, considerando todas las variables de manera directa sobre el SLMI. Para cada iteración, se obtuvo un grupo de pesos optimizados que, mediante un análisis de sensibilidad, permitió generar diferentes configuraciones del SLMI y evaluar el impacto de distintas decisiones metodológicas sobre el indicador final:



Ilustración 28. Diagrama de caja y bigotes sobre las diferentes iteraciones de SLMI obtenidas

De la gráfica, donde la variación del SLMI es menor en sus muchas iteraciones, se deduce la robustez del propio indicador. Si bien siempre existe margen para la mejora metodológica, los criterios de la OCDE (2008) proporcionan una base sólida sobre la que se asientan los sintéticos llamados a ser herramientas de transformación y análisis socioeconómico.

Para concluir cabe mencionar ciertas líneas de investigación que se dejan abiertas, tales como el análisis estadístico-matemático del método de los splines, especialmente en lo referente a los *knots* y el término de suavizado; el estudio metodológico de asignación de pesos por parte del FA frente a varianza explicada e importancia; así como el análisis de más técnicas de imputación.

De esta manera, se da por concluido el trabajo.

5. Referencias

- Álvarez, G. (2022). *Desarrollo de un KPI sintético para medir la recuperación española frente al COVID-19*. Madrid: Universidad Pontificia Comillas.
- Becker, W., Saisana, M., Paruolo, P., & Vandecasteele. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological indicators*, 80, 12-22.
- Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25 (5).
- Cherchye, L., Moesen, W., & Rogge, N. (2007). An Introduction to 'Benefit of the Doubt' Composite Indicators. *Social Indicators Research* 82, 111-145.
- Chicco, D., Warrens, M., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Computer Science*.
- Comillas - ICADE. (13 de Mayo de 2021). *Observatorio Criterias: un análisis económico y social*. Obtenido de <https://www.comillas.edu/noticias/63-comillas-icade/icade-cee/cee-investigacion/2351-observatorio-criterias-analisis-de-la-situacion-en-espana>
- Donders, A. R., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 59, 10, 1087-1091.
- Freudenberg, M. (2003). *Composite indicators of country performance: A critical assessment*. París: OECD Publishing.
- Gold, M. S., & Bentler, P. M. (2000). Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(3):319–355.
- Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2018). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Social Indicators Research* 141, 61–94.
- Grupp, H., & Mogege, M. E. (2004). Indicators for national science and technology policy: how robust are composite indicators? *Research Policy* 33, 9, 1373-1384.
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. Springer Series.
- Li, G. R. (2010). Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs. *The Journal of Physical Chemistry*, 114, 6022-6032.
- Malan, L., Smuts, C. M., Baumgartner, J., & Ricci, C. (2020). Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. *Nutrition Research*, 75:67-76.
- Mara, T. T. (2012). Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety*, 107, 115-121.
- Medel Esquivel, R., Gomez-Vargas, I., Vazquez, J., & García-Salcedo, R. (2019). *An introduction to Markov Chain Monte Carlo*. Mexico DF.

- Monitor Deloitte. (2022). *360 Smart Vision: Indicadores en tiempo real de la realidad socioeconómica de España*. Obtenido de <https://www2.deloitte.com/es/es/pages/strategy/articles/360-smart-vision.html/#dashboard>
- Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, 24 (7) 815-829.
- Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). *Handbook on constructing composite indicators*. París: OECD Publishing.
- OCDE. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. París: OECD Publications.
- Roch-Dupré, D., Aracil, E., & Calvo-Bascones, P. (2023). Monitoring socioeconomic readiness for the demographic transition: Introducing the Senior Economy Tracker. *Structural Change and Economic Dynamics*, Vol. 71, 430-443.
- Ruppert, D., Wand, M., & Carroll, R. (2003). *Semiparametric regression*. Cambridge: Cambridge University.
- Saisana, M., & Saltelli, A. (2011). Rankings and Ratings: Instructions for Use. *Hague Journal*, 3 (2): 247-268.
- Saltelli, A. (2007). Composite indicators between analysis and advocacy. *Social Indicators Research*, 81 (1), 65-77.
- Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. (2013). *A Survey of Forecast Error Measures*.
- Szopik-Depczynska, K., Cheba, K., B., I., Stakniak, M., Simboli, A., & Ioppolo, G. (2018). The study of relationship in a hierarchical structure of EU sustainable development indicators. *Ecological Indicators*, 120-131.
- Willmott, C., Ackleson, S., Davis, R., Feddema, J., Klink, K., Legates, D., . . . Rowe, C. (1985). Statistics for the Evaluation and Comparison of Models. *Journal of Geophysical Research*.
- Wood, S. (2017). *Generalized Additive Models*. New York: Chapman and Hall / CRC.
- Xu, C., & Gertner, G. (2008). *Reliability Engineering Systems*, 93, 1563-1573.
- Young, R., & Johnson, D. (2015). Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. *J Marriage Fam*, 77 (1) 277-294.

Anexo 1. Alineación con los ODS

La alineación de este proyecto con los Objetivos de Desarrollo Sostenible (ODS) establecidos por la ONU para 2030 es crucial, ya que busca contribuir al desarrollo inclusivo y sostenible mediante la mejora en la construcción de indicadores sintéticos. En este contexto, la mejora metodológica en la imputación y ponderación de los indicadores puede tener un impacto significativo en varios ODS, promoviendo un enfoque integral y multidimensional que abarca tanto la dimensión económica como la social. Al optimizar estos procesos, este proyecto se convierte en una herramienta clave para evaluar y mejorar las políticas públicas, particularmente en áreas relacionadas con el mercado laboral, la equidad de género y la reducción de las desigualdades.

Uno de los ODS más relevantes para este proyecto es el ODS 8, que se enfoca en promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo, y el trabajo decente para todos. Al mejorar la metodología de construcción de indicadores, se proporcionan herramientas más precisas y efectivas para evaluar el estado del mercado laboral (en el caso del SLMI) y otros aspectos económicos medidos por otros indicadores. La visibilidad de los problemas transitorios y estructurales que afectan al mercado laboral permite una mejor comprensión de los desafíos que enfrenta el sistema productivo. Con este conocimiento, los responsables de la formulación de políticas pueden tomar decisiones informadas, identificar áreas de mejora y desarrollar políticas públicas que impulsen tanto la productividad como las condiciones laborales, siguiendo la línea de los ODS 8.2 y 8.3, que buscan la mejora de la productividad y la innovación. La capacidad de monitorear realidades socioeconómicas con mayor precisión también facilita el desarrollo de políticas que apoyen un crecimiento económico más equitativo y sostenible.

Además, este proyecto también se alinea con el ODS 10, que busca reducir las desigualdades a nivel nacional e internacional. La mejora en la construcción de indicadores ayuda a capturar mejor las múltiples dimensiones de la desigualdad, proporcionando un marco robusto para evaluar y abordar las disparidades en el acceso a oportunidades y recursos.

Por otro lado, la mejora en las técnicas de imputación y ponderación de indicadores también está relacionada con el ODS 17, que promueve el fortalecimiento de los medios de implementación y la revitalización de la alianza mundial para el desarrollo sostenible. Un enfoque metodológico más riguroso y preciso facilita la colaboración entre académicos, responsables políticos y otros actores relevantes, permitiendo un intercambio de conocimientos y experiencias que enriquezcan la construcción de indicadores. Este proceso de colaboración y aprendizaje mutuo contribuye al fortalecimiento de los mecanismos de implementación de políticas públicas y al fortalecimiento de alianzas que apoyen el desarrollo sostenible, permitiendo que los ODS se logren de manera más efectiva a nivel global, nacional y local.

En resumen, la alineación de este proyecto con los ODS garantiza que la mejora en la construcción de indicadores no solo contribuya a los objetivos globales de desarrollo sostenible, sino que también refuerce la relevancia e impacto social de estos instrumentos en la búsqueda de un futuro más equitativo y sostenible. Al centrarse en la robustez y transparencia de la construcción de indicadores, el proyecto tiene como objetivo generar un cambio positivo que beneficie a la sociedad en su conjunto, asegurando que el desarrollo económico y social sea inclusivo y sostenible a largo plazo. Este enfoque integral y metodológico garantiza que las políticas públicas puedan ser evaluadas y ajustadas de manera más efectiva, impulsando un desarrollo que promueva tanto el bienestar económico como la justicia social.

Anexo 2. Composición del SLMI

A continuación, se muestran los materiales complementarios relativos al SLMI empleado en el presente trabajo (Comillas - ICADE, 2021):

Desempleo (A) 30%	Indicadores base
	Desempleo registrado (<25 years) ¹ (A.1)
	Desempleo registrado – Agricultura y pesca (A.2)
	Desempleo registrado – Construcción (A.3)
	Desempleo registrado – CVE (A.4)
	Desempleo registrado – Industria (A.5)
	Desempleo registrado – Mujeres (A.6)
	Desempleo registrado – Servicios (A.7)
	Desempleo registrado – Sin empleo previo (A.8)
	Desempleo registrado – Varones (A.9)
	Tasa de desempleo trimestral (A.10)
Employment (B) 30%	Indicadores base
	Contratos registrados (B.1)
	Contratos registrados a tiempo parcial (B.2)
	Contratos registrados – CVE (B.3)
	Contratos registrados – Indefinidos (B.4)
	Contratos registrados – Interinos (B.5)
	Contratos registrados – Temporales (B.6)
	Productividad por hora realmente trabajada (B.7)
	Productividad por trabajo a tiempo completo (B.8)
	Tasa de empleo trimestral (B.9)
Protection for the Unemployed (C) 10%	Indicadores base
	Beneficiarios Asistenciales (C.1)
	Prestaciones Contributivas Desempleo Parcial (C.2)
	Prestaciones Contributivas Desempleo Total (C.3)
Salarios (D) 30%	Indicadores base
	Incremento salarial pactado en negociación colectiva (D.1)
	Variación interanual de los rendimientos del trabajo – Agricultura (D.2)
	Variación interanual de los rendimientos del trabajo – Comercio (D.3)
	Variación interanual de los rendimientos del trabajo – Construcción (D.4)
	Variación interanual de los rendimientos del trabajo – Hostelería (D.5)
	Variación interanual de los rendimientos del trabajo – Industria (D.6)
	Variación interanual de los rendimientos del trabajo – Servicios (D.7)
Variación interanual de los rendimientos del trabajo – Total (D.8)	

Tabla A1. Composición detallada del SLMI

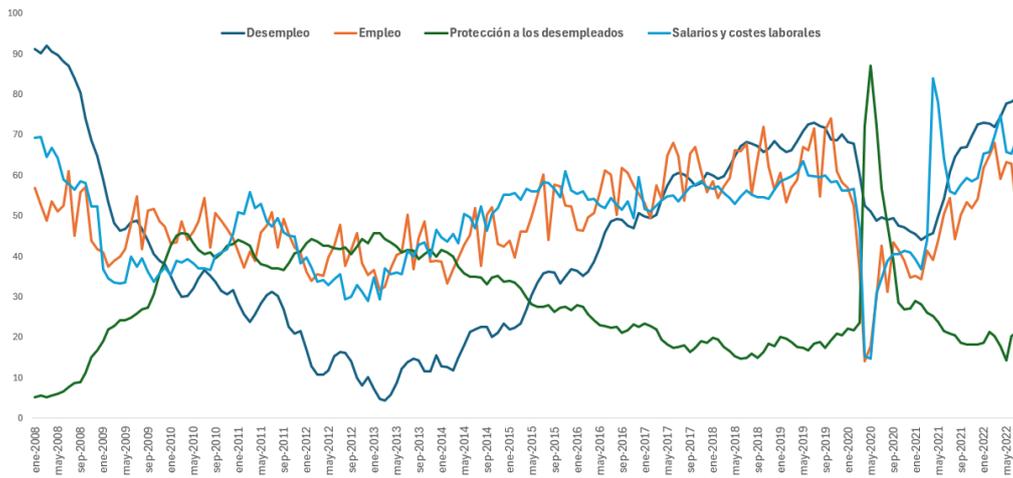


Ilustración A1. Áreas del SLMI a partir de la agregación equiponderada aritmética de sus componentes



Ilustración A2. SLMI con agregación equiponderada aritmética de sus áreas manteniendo criterio experto en el segundo nivel de agregación



Ilustración A3. SLMI con agregación equiponderada de todos los indicadores base sin agregación intermedia

Anexo 3. Estructura del Senior Economy Tracker

Los indicadores base del Senior Economy Tracker se presentan a continuación (Roch-Dupré et al., 2023):

Dimensión Demográfica (A)

Edad y esperanza de vida

- A1.1 Edad promedio de la población *
- A1.2 Población de adultos mayores (% del total) *
- A1.3 Esperanza de vida **
- A1.4 Calidad de la vida en la vejez *
- A1.5 Esperanza de vida después de los 55 años *
- A1.6 Ratio de dependencia *

Dimensión Institucional (B)

Salud y protección social (B1)

- B1.1 Gasto en salud per cápita **
- B1.2 Gasto en salud por parte del gobierno **
- B1.3 Densidad de trabajadores de la salud y protección social **
- B1.4 Proporción del gasto en salud financiado por el gobierno **

Pensiones y protección laboral (B2)

- B2.1 Pensión anual promedio *
- B2.2 Número de pensiones *
- B2.3 Tasa de reemplazo de pensiones *
- B2.4 Edad promedio de jubilación **
- B2.5 Duración de la jubilación **
- B2.6 Gasto individual en planes de pensión privados **
- B2.7 Gasto empresarial en planes de pensión privados **
- B2.8 Riesgo de pobreza en pensionados *

Dimensión Macroeconómica (C)

Mercado de bienes y servicios para adultos mayores (C1)

- C1.1 Gasto promedio en consumo *

Mercado laboral para adultos mayores (C2)

- C2.1 Duración de la vida laboral *
- C2.2 Ratio de transición laboral *
- C2.3 Trabajo remoto en personas mayores *
- C2.4 Tasa de empleo en adultos mayores *
- C2.5 Autoempleo en adultos mayores *

Dimensión Individual (D)

Participación en la sociedad (D1)

- D1.1 Falta de recursos para pequeños gastos semanales *
- D1.2 Imposibilidad de participar en actividades de ocio regularmente *
- D1.3 Imposibilidad de reunirse con familiares o amigos en restaurantes o hoteles al menos una vez al mes *
- D1.4 Desigualdad de ingresos (Ratio 80-20) *
- D1.5 Falta de conexión a internet para uso personal *
- D1.6 Vida independiente y autónoma (una persona mayor) *

- D1.7 Vida independiente y autónoma (dos o más personas mayores) *
- D1.8 Aprendizaje a lo largo de la vida *
- D1.9 Habilidades tecnológicas *

Seguridad financiera (D2)

- D2.1 Ingreso promedio relativo *
- D2.2 Ausencia de riesgo de pobreza *
- D2.3 Ausencia de privación material severa *
- D2.4 Ausencia de pobreza energética (hogares con 1 persona mayor) *
- D2.5 Ausencia de pobreza energética (hogares con 2 adultos y al menos 1 persona mayor) *
- D2.6 Trabajadores mayores en riesgo de pobreza *
- D2.7 Imposibilidad de afrontar gastos imprevistos (hogares con 1 persona mayor) *
- D2.8 Imposibilidad de afrontar gastos imprevistos (hogares con 2 adultos y al menos 1 persona mayor) *
- D2.9 Relación entre el costo de la vivienda y el ingreso disponible (hogares con 1 persona mayor) *
- D2.10 Relación entre el costo de la vivienda y el ingreso disponible (hogares con 2 adultos y al menos 1 persona mayor) *
- D2.11 Endeudamiento por vivienda principal *
- D2.12 Propiedad de vivienda (hogares con 1 persona mayor) *
- D2.13 Propiedad de vivienda (hogares con 2 adultos y al menos 1 persona mayor) *

Envejecimiento saludable y activo (D3)

- D3.1 Acceso a servicios de salud *
- D3.2 Autopercepción del estado de salud como bueno o muy bueno *
- D3.3 Personas con enfermedades o problemas de salud crónicos *
- D3.4 Autopercepción de ausencia de limitaciones en actividades diarias debido a problemas de salud *
- D3.5 Autopercepción de ausencia de falta de atención médica *
- D3.6 Esperanza de vida saludable a los 65 años *
- D3.7 Muertes prematuras debido a la exposición a partículas finas *

Donde los indicadores representados por * han sido obtenidos a partir de Eurostat y los que presentan ** de la OCDE.

Anexo 4. Código fuente

Nótese que las funciones empleadas en el código fuente descrito son proporcionadas por Becker et al. (2017) con ligeras variaciones para adecuarlo al conjunto de datos estudiado:

```
clear all
clear peso_var
load peso_var_norm.mat
n_cat=5;

%% Cargar archivos
for i=1:n_cat
    [norm_matrix{1,i},q{1,i},norm_tabla{1,i}]=xlsread([strcat('Tabla',
num2str(i)),'_Norm Mercado de Trabajo.xlsx']);
end
clear q

areas=xlsread('Areas Mercado de Trabajo.xlsx');
indice=xlsread('Indice.xlsx');
variables=xlsread('Variables normalizadas juntas.xlsx');
for i=1:n_cat
    if i~=3
        x=norm_matrix{1,i};
        y=areas(:,i);
        Sis=spline_SA(y,x);
        outSi=CI_Si_decomposer(x,y,1,i);
    end
end

Sis=spline_SA(indice,variables);
disp(Sis)
outSi=CI_Si_decomposer(variables,indice,1,1);

clear peso_var
for i=1:n_cat
    if i~=3
        for j=1:size(peso_var_norm{1,i},1)
            %peso_var{1,i}(j,1)=1/size(peso_var_norm{1,i},1);
            if i~=4
                %peso_var{1,i}(j,1)=max(peso_var_norm{1,i}(j,:));
                peso_var{1,i}(j,1)=1/size(peso_var_norm{1,i},1);
            else
                peso_var{1,i}(j,1)=1/size(peso_var_norm{1,i},1);
            end
        end
    end

    %% Normal
    outOpt=CIoptimiser(norm_matrix{1,i},peso_var{1,i},2,'ArAv')
    Optimised_weights{1,i}=outOpt.wopt
    Sis_at_optimised_weights{1,i}=outOpt.Sopt;
    for j=1:size(Sis_at_optimised_weights{1,i},1)

a{1,i}(j,1)=Sis_at_optimised_weights{1,i}(j,1)/sum(Sis_at_optimised_weights{1,i});
```

```
        end
    end
end
%% Equiponderación todas juntas
% clear peso_var
% for x=1:size(variables,2)
%     peso_var(x,1)=1/size(variables,2);
% end
%outOpt=CIoptimiser(variables,peso_var,2,'ArAv')
```