# ESTIMACIÓN DE UN INDICADOR SINTÉTICO DE ALTA FRECUENCIA PARA LA ECONOMÍA ANDALUZA (ISAEA)

Carmen María Rubio Castaño

Manuel Alejandro Hidalgo Pérez

Universidad Pablo de Olavide

## Introducción

Los indicadores sintéticos son herramientas que combinan múltiples variables o indicadores individuales en una sola medida para proporcionar una visión más amplia y simplificada de una situación específica. Estos indicadores suelen utilizarse en diversos campos, como economía, finanzas, medio ambiente, salud pública y más. Así, por ejemplo, el Producto Interno Bruto (PIB) tal y como es calculado para las series trimestrales no deja de ser un indicador sintético que resume la actividad económica de un país, considerando la producción de bienes y servicios en un solo número.

La creación de estos indicadores puede requerir métodos estadísticos o matemáticos para ponderar y combinar las diferentes variables de manera que reflejen con precisión la situación que se desea medir. Sin embargo, la selección de las variables y la forma en que se combinan pueden generar debates, ya que la interpretación de estos indicadores sintéticos, a veces, puede simplificar en exceso realidades complejas.

Al iniciar un proyecto de indicadores sintéticos, es importante primero definir claramente el objetivo que se busca y el problema que se intenta resolver. En este caso, el objetivo es elaborar un indicador sintético de actividad para la economía andaluza (en adelante ISAEA) que sea capaz de informar en frecuencia mensual la evolución de la actividad andaluza. No es este un caso en el que se quiera sustituir al PIB como referencia de la evolución de la actividad (ni podría), sino de informar en una frecuencia mayor sobre la evolución de dicha actividad que, a su vez, puede ser también medido con mayor precisión por el PIB. Una vez definido el objetivo, el siguiente paso clave es identificar las principales variables o indicadores individuales que capturan los aspectos más relevantes del fenómeno que se busca medir, y que se desean incorporar en el indicador sintético final.

Cuando se tienen identificadas las variables individuales, es necesario recolectar los datos correspondientes a esos indicadores para los casos (países, regiones, etc.) y años que se incluirán en el análisis. Es muy importante revisar la integridad y calidad de los datos, realizando una limpieza en caso de ser necesario, para asegurarse que la información esté completa, al que le seguirá un tratamiento y estandarización de dicha información. Así, mediante dicha estandarización de indicadores, se eliminará el ruido posible que pueda transmitir las distintas métricas y escalas que no sean comparables y medidos en una escala común, así como la eliminación de la información que no resulte relevante para el objetivo, como son los factores estacionales o el ruido generado por eventos repetitivos más allá de la estacionalidad como es el calendario.

Sin embargo, la tarea más compleja de la realización de un indicador sintético, además de la elección de los indicadores parciales que lo componen, recae en la definición de las reglas de agregación de la información. Esta agregación tradicionalmente se lleva a cabo mediante la asignación de ponderaciones o pesos a cada una de las variables, pesos que se suelen asignar de acuerdo con su relevancia en la conformación del indicador sintético final. Con estos pesos se da más o menos importancia a ciertos indicadores en función del fenómeno que se estudia. Las ponderaciones suelen obtenerse mediante diversos métodos, pero los más habituales son los derivados de juicios de expertos, a partir de análisis estadísticos como son la regresión, el uso de simples correlaciones o el análisis factorial. La elección de cada uno de los métodos debe atenderse en función de los objetivos propuestos y de las necesidades de cada caso. No obstante, en la medida de ser lo más objetivos posibles, dichas ponderaciones deben atender, principalmente, a la pura información extraída de los datos.

Es por esto por lo que es absolutamente necesario justificar dicha calibración y qué se pretende con ello. Una opción es recurrir a fuentes estadísticas externas, como son los pesos de los sectores o de las ramas de actividad en el PIB de la economía, en el caso de un indicador como el que se quiere realizar para Andalucía. Otra opción es el de recurrir a estadísticos endógenos a la información ofrecida por los indicadores a usar. En este segundo caso, se usarían correlaciones, coeficientes de regresiones u otras métricas que permitan definir la relevancia de cada uno de los indicadores a la hora de incorporarse al indicador sintético.

En nuestra consideración, la segunda opción representa la alternativa más favorable. En este escenario, los datos mismos, sin intervención directa del analista, determinan su propio peso en la creación del indicador. Sin embargo, este peso estará, en última instancia, sujeto a las decisiones sobre qué métrica se empleará para definirlos. Tal como se detalla en este informe, siempre surgirán decisiones sobre la metodología a utilizar para identificar qué indicadores se incluyen y con qué nivel de relevancia en el indicador sintético. Por tanto, la justificación detrás de estas decisiones debe estar enfocada en maximizar la objetividad y la calidad del indicador construido, alineado con los objetivos planteados.

En el resto del informe se describe una primera aproximación para la construcción del indicador sintético. Para ello se van a utilizar los indicadores suministrados por el IECA y se va a plantear una metodología que, dados los condicionantes existentes y que serán detallados a lo largo del texto, podría llevarnos a obtener un indicador objetivo y que maximice el uso de la información suministrada por todos y cada uno de los indicadores individuales. No obstante, consideramos que este ejercicio que se presenta en este informe es solo una primera aproximación y que existen metodologías de reducción de variables que podrían ofrecer indicadores que mejoren el cumplimiento de los objetivos propuestos.

## Un indicador de actividad para la economía Andaluza

En primer lugar, antes de construir un indicador es necesario llevar a cabo una serie de etapas que permitan eliminar de los indicadores individuales aquella información que no es relevante para el seguimiento de la actividad cíclica andaluza. Así, para construir un indicador sintético de actividad económica y que minimice las diferencias con el perfil cíclico mostrado por el PIB de un país o, en este caso, de una región, se deben realizar los siguientes pasos.

#### 2.1 Selección y tratamiento de series

En primer lugar, se debe seleccionar una batería de indicadores potenciales y, entre los cuáles, poder escoger aquellos que de forma conjunta puedan aproximar de la mejor forma posible la actividad de la economía andaluza. Para ello deben seleccionarse previamente aquellos que muestren una estrecha relación contemporánea con el PIB y que cubran los principales sectores de la economía. Entre ellos, y a propuesta del IECA, se han seleccionado una batería de 14 indicadores, mensuales y trimestrales que mejor podrían capturar la información cíclica del PIB andaluz.

Estos indicadores son los siguientes:

* Consumo aparente de cemento
* Consumo de gasolina y gasóleo
* Exportaciones de bienes
* Importaciones de bienes
* IPIAN Div 23
* Índice general de producción industrial
* ICMA General sin estaciones de servicios
* Matriculación de turismos
* Índice ventas grandes superficies
* Pernoctaciones
* Cifra negocios del sector servicios
* Liquidación de presupuestos de la Junta de Andalucía. Capítulo 1
* Total de afiliados SS Total
* Índice de producción agrícola.

Los requisitos para que estas series puedan constituirse como candidatas a incluirse en el ISAEA son, al menos, tres.

1. En primer lugar, como se ha avanzado, deben tener o bien una vinculación directa con la evolución de la actividad productiva andaluza general, o bien deben reflejar la evolución en sectores parciales de la actividad productiva.
2. En segundo lugar, se debe exigir de estas series que se extiendan para un periodo de tiempo suficientemente largo.
3. En tercer lugar, deben aportar una garantía mínima de calidad.

Una vez hecha la selección y aplicado un filtro de calidad y suficiencia, se procede a eliminar aquellos componentes de las series que no son necesarios, y que incluso ocultan los componentes más relevantes. Junto con la rigurosa revisión y limpieza de los datos faltantes o atípicos es necesario proceder a la tradicional desestacionalización y eliminación de los efectos del calendario en la evolución de los datos de frecuencia superior a la anual[[1]](#footnote-1).

#### 2.2 Estacionariedad de los indicadores

Cualquier análisis que implique la búsqueda del mejor modelo posible para relacionar estadísticamente series temporales debe incorporar, en su inicio, además de un tratamiento como los indicados en el apartado anterior, el de transformar dichas series en series estacionarias. Es necesario que, para establecer relaciones econométricas entre series temporales, estas sean ergódicas. Generalmente dicha ergodicidad suele alcanzarse mediante la diferenciación de las series que la necesiten.

Para conocer el grado de diferenciación necesaria para transformar en estacionarios los indicadores individuales se aplica el *test* de Dickey-Fuller al logaritmo neperiano de cada serie. La implementación, para agilizar el procedimiento, se realizó automáticamente mediante la aplicación de un algoritmo que determinaría el grado de integración de cada serie en función de si cumple o no la hipótesis nula que marca el *test* aplicado[[2]](#footnote-2).

Así, siendo el indicador, el algoritmo en el primer paso aplica el *test*. Si se rechaza la hipótesis nula de estacionariedad, se procede a calcular su diferenciación, obteniendo una nueva serie . La diferenciación en tres períodos responde a la idea de que la aproximación de la actividad se va a establecer respecto a una serie trimestral (el PIB) y a su crecimiento intertrimestral.

A se le vuelve a aplicar el *test* de estacionariedad y si se acepta, el proceso finaliza. Si se rechaza, se vuelve a calcular la diferenciación hasta lograr que la hipótesis nula sea aceptada. El mismo procedimiento se replica para los datos trimestrales, aunque en este caso con diferenciaciones de un trimestre. Una vez finalizado se registra la cantidad de transformaciones necesarias para cada columna, lo que permite identificar y adecuar las series que requieren ajustes más significativos.

Sin embargo, la diferenciación de algunas series no es adecuadamente asignada por el proceso automático. Es el caso de 'Pernoctaciones', y cuya particular estructura derivada de los meses de COVID dificultan la aplicación de los tests. Es por ello por lo que a esta serie se le asigna manualmente una integración de orden 1 tras una tradicional comprobación visual de su comportamiento.

El resultado después del tratamiento y de la diferenciación hasta lograr la estacionariedad de las series puede observarse en las figuras 1 y 2, para las series mensuales y trimestrales respectivamente.

**Figura 1. Series mensuales CVEC y estacionarias.** **Imagen que contiene Calendario

Descripción generada automáticamente**

**Figura 2. Series trimestrales CVEC y estacionarias.**

Gráfico

Descripción generada automáticamente

## 3. Optimización de la selección de series para el ISAEA.

#### El uso de la correlación como identificación

El objetivo de esta sección es el de identificar cuáles son los indicadores preferidos a la hora de construir el ISAEA mediante el uso de una metodología objetiva. En este caso concreto, se propone el uso de regresiones por mínimos cuadrados (OLS) para justificar el cálculo del indicador sintético mediante el uso de estadísticos endógenos (generados por los propios datos). No obstante, y como se adelantó, este ejercicio solo corresponde a una aproximación previa que se concretará, finalmente, por el uso de métodos de reducción de la dimensionalidad para elaborar un indicador sintético de actividad para la economía andaluza.

Así pues, y a pesar de no corresponder a un ejercicio intermedio, este nos permite valorar qué grado de aproximación podemos logar, así como fijar algunos elementos de la metodología final que se aplicará para construir el ISAEA definitivo.

En primer lugar, y dado que el indicador objeto de aproximación tiene frecuencia trimestral, lo primero que debemos hacer es transformar los indicadores mensuales a esta frecuencia, aunque el indicador final tendrá frecuencia mensual. Una vez agregado a una frecuencia menor, mediante el uso de la media, se puede realizar una primera aproximación a la validez de los indicadores como potenciales fuentes de información para el seguimiento de la actividad económica andaluza. Para ello, el primer análisis natural corresponde a la visualización de la correlación contemporánea entre indicador y PIB trimestral de Andalucía. La figura 3 muestra este ejercicio, donde puede observarse que la correlación entre los diversos indicadores en tasa de crecimiento intertrimestral es elevada respecto al PIB, con excepción de los indicadores de producción agrícola, la liquidación de presupuestos y las exportaciones de bienes y servicios.

**Figura 3. Correlación entre el crecimiento de los indicadores y PIB trimestral**

Gráfico, Gráfico de barras

Descripción generada automáticamente

Una vez conocida la correlación, la primera idea por la que estaríamos tentados a proceder sería utilizar aquellos indicadores con un mayor valor de esta. Sin embargo, esto podría suponer un error por varias razones. En primer lugar, la falta de correlación simplemente indica la incapacidad de un indicador para aproximar por sí solo la evolución del PIB a lo largo del tiempo a partir de una especificación lineal. No obstante, esto no indica su capacidad para transmitir información relevante que no sea aportada por el resto de los indicadores. Es posible que su utilidad no sea nula a pesar de una baja correlación.

En segundo lugar, es importante considerar el período comprendido desde el primer trimestre de 2020 hasta mediados de 2021, el cual representa una fase que introduce variabilidad adicional en las series temporales. Esto dificulta la evaluación de las relaciones entre estas mediante medidas simples como el coeficiente de correlación u otras similares. Durante este corto período de tiempo, la variabilidad experimentada fue considerablemente alta en comparación con la tendencia histórica. Esta situación implica que, aunque las series temporales sean extensas, la correlación en estos meses recientes tendría un peso significativo en la correlación total entre las variables, potencialmente sesgando los resultados a favor de indicadores cuyo comportamiento durante la pandemia se asemejó más al del PIB. Las figuras 1 y 2 ilustran cómo cada serie temporal absorbió el impacto de la pandemia. A pesar de mostrar perfiles similares, algunas de estas series lo hicieron con una variabilidad más o menos intensa en relación con la del PIB. Por lo tanto, el uso exclusivo de este estadístico para seleccionar los indicadores que mejor reflejen la actividad económica de la región definida por el PIB puede no ser la opción más adecuada.

La idea, por ende, es emplear otras aproximaciones que nos permitan, en primer lugar, seleccionar indicadores por su relevancia para replicar la evolución de la actividad económica definida por el PIB y, en segundo lugar, extraer de ellos la información necesaria con la garantía de hacerlo correctamente.

Por todas estas razones, la estrategia previa para identificar los indicadores preferidos en la elaboración del indicador es la de encontrar la combinación de estos que mejor replique la evolución del PIB, condicionado a la particularidad que supone la existencia de fuertes "outliers" en torno a los meses de la pandemia. La variación de las series en dichos meses y la fuerte correlación “artificial” con la serie objetivo que es el PIB implica que debamos considerar estrategias de identificación que sean sensibles a este hecho y que, por lo tanto, descarten estadísticos que, o bien, incorporen dichos meses en su cálculo o bien nos permita elegir entre aquellos que mejor aproximen la evolución “a pesar” de dichos meses. Si esto no se hiciera, se podría estar cometiendo el error de asignar a una serie una fuerte capacidad de información cuando, en realidad, lo que sucede es que habrá una clara dominancia de la variación de la actividad recogida por dicho indicador y el PIB entre marzo y septiembre de 2020.

Por ejemplo, si realizamos una regresión entre el PIB y todos los indicadores, como se muestra en la siguiente tabla, el coeficiente de determinación será excepcionalmente alto, lo que implica una alta precisión de ese modelo. Sin embargo, este resultado se ve afectado, además de una evidente multicolinealidad, por los datos extremadamente atípicos de la pandemia, y que dominarán al resto por el mero hecho de su valor extremo.

Tabla 1. Regresión con indicadores del PIB andaluz

Tabla

Descripción generada automáticamente

Como se puede apreciar en la figura 4, la capacidad predictiva del modelo es alta (salvo, paradójicamente, su incapacidad para aproximar los datos extremos de la pandemia), aunque no hay garantía de que el mejor modelo sea aquél que utiliza todos los indicadores. Además de lo mencionado anteriormente, debemos considerar el riesgo tradicional de sobreajuste en un modelo de regresión, así como el problema de multicolinealidad, que se manifiesta en signos inesperados en indicadores con correlaciones positivas respecto a la variable dependiente.

**Figura 4. Serie de PIB observada y predicha por la regresión (tasa de crecimiento intertrimestral)**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Una prueba de lo que se está comentando, del posible sesgo en la estimación del modelo que supone considerar los meses de la pandemia, se pueden observar en la regresión que se obtendría si esta se realizara con los datos hasta diciembre de 2019. El coeficiente de determinación del modelo decae hasta el 88,4 %, lo que muestra que tratar de aproximar las variables que mejor aproximan al PIB usando las fechas que corresponden a la pandemia no es la mejor estrategia.

Tabla

Descripción generada automáticamente

La figura 5 muestra la evolución de la serie de PIB predicho y observada cuando se usa el modelo con datos hasta diciembre de 2019. Es obvio que nos muestra una aproximación bastante aceptable.

**Figura 5. Serie observada y predicha del PIB hasta diciembre de 2019**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Para finalizar con este análisis, podemos mirar la figura 6 como una muestra más del sesgo que puede suponer tomar en consideración los meses de la pandemia. Lo que nos dice esta figura, como la anterior figura 3, es que los indicadores, en general, mantienen una elevada correlación respecto al PIB. Sin embargo, esta es diferente para algunos de ellos. Nada mejor que para poder valorar esta diferencia que mostrar en una misma figura no tanto las correlaciones sino los coeficientes de ambas regresiones lineales mostradas en un solo gráfico. Esto es lo que nos enseña la figura 7.

**Figura 6. Coeficiente de correlación hasta diciembre de 2019**

Gráfico, Gráfico de barras

Descripción generada automáticamente

Así, y por ejemplo, para las ventas en grandes superficies, el índice de producción agrícola, las exportaciones, la liquidación de presupuestos y alguno otro indicador, los coeficientes en la regresión son significativamente diferentes dependiendo de si se incluyen o no los meses posteriores a la pandemia. Esto, obviamente, como se ha adelantado, obliga a tener en consideración este período antes de seleccionar o elegir cuáles son los mejores indicadores para aproximar la evolución de la actividad andaluza.

**Figura 7. Coeficientes de las regresiones dependiendo si se incluyen meses posteriores a enero de 2020**

Gráfico

Descripción generada automáticamente

Una prueba de que debe tenerse en cuenta estos meses como distorsionadores está en la proyección del PIB predicho a partir del modelo estimado hasta 2019. Al analizar únicamente la información disponible hasta diciembre de 2019, se observa en la figura 8 que, además de ser una buena aproximación para períodos pasados, el nuevo modelo proporciona estimaciones muy precisas durante los meses críticos de la pandemia, así como los posteriores.

**Figura 8. Comparación del PIB observado y el PIB predicho por el modelo teniendo en cuenta solo los meses hasta diciembre de 2019.**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

#### 3.2 Búsqueda iterativa del mejor modelo

Para solucionar los problemas planteados mientras se escogen los indicadores que optimizan la capacidad de reproducir la evolución del PIB se propone una estrategia doble basada en la idea de “cross validation” usada en la selección de modelos de aprendizaje automático.

1. En primer lugar, y como se ha adelantado, la estimación del modelo va a realizarse hasta diciembre de 2019 (grupo de entrenamiento), dejando los datos de 2020 y posteriores fuera del cálculo de este.
2. En segundo lugar, para seleccionar los mejores indicadores no se utilizarán estadísticos tradicionales de bondad de ajuste para el conjunto de los datos, sino métricas aplicadas para las predicciones posteriores a diciembre de 2019 (grupo de prueba). Este proceso de selección sigue, así, los pasos iterativos que se describen a continuación:
   1. En primer lugar, el proceso se caracteriza por buscar entre los modelos predictivos aquél que optimiza una métrica definida. Concretamente, de los indicadores disponibles se estiman regresiones con diferentes combinaciones posibles entre ellos, con un mínimo de 3 indicadores y un máximo de 11. Este enfoque de búsqueda exhaustiva de combinaciones de regresores trata de encontrar la mejor combinación posible de variables para predecir el PIB. Esto puede mejorar la precisión de los modelos predictivos al identificar las variables más influyentes en la economía de Andalucía.
   2. Se define una evaluación exhaustiva de las diferentes combinaciones. Al considerar una amplia gama de combinaciones de regresores (un total de 8086 combinaciones evaluadas), el proceso permite analizar sistemáticamente cómo diferentes conjuntos de variables afectan la capacidad del modelo para predecir el PIB. Esto proporciona una visión integral y detallada de la importancia relativa de cada variable en la predicción económica.
   3. Este proceso permitiría, aunque parcialmente, la identificación de relaciones ocultas. Explorar múltiples combinaciones de variables puede revelar relaciones no evidentes entre diversos aspectos económicos y el PIB. Esto puede desvelar conexiones subyacentes que, de otro modo, podrían pasarse por alto al analizar individualmente cada variable. Por ejemplo, variables con escasa correlación con el PIB, pero que podrían explicar una parte no evidente de su evolución y que quedaría oculta por la capacidad predictiva de otros indicadores con mayor correlación con la variable objetivo.

Como se ha indicado, el proceso de evaluación propuesto en este trabajo tiene similitudes con un enfoque de aprendizaje automático, específicamente con técnicas de selección de características y ajuste de modelos. Esto es así porque, en primer lugar, explora de forma sistemática las características que mejor predicen la evolución del PIB. Esto es similar a la selección de las características que se desarrollan en técnicas de aprendizaje automático. Así, este proceso busca determinar las mejores variables (regresores) que influencian la variable objetivo (PIB). Explora múltiples combinaciones para identificar las más relevantes en la predicción.

En segundo lugar, evalúa los modelos de forma similar a como se llevan a cabo en los algoritmos de aprendizaje automático. En este caso, con modelos de regresión lineal, se utilizan métricas de desempeño como el Error Cuadrático Medio (MSE). El objetivo es encontrar el modelo con el mejor desempeño predictivo en función de las características seleccionadas.

En tercer lugar, la optimización del modelo se diseña mediante la búsqueda de la combinación óptima de características que minimiza el error en la predicción del PIB. Esto se asemeja a la optimización de hiperparámetros en el aprendizaje automático, donde se busca maximizar el rendimiento del modelo. Finalmente, el enfoque iterativo prueba múltiples configuraciones y se realiza un seguimiento de los resultados para mejorar la precisión del modelo.

Así pues, aunque no emplea algoritmos de aprendizaje automático avanzados, este proceso comparte aspectos fundamentales con las metodologías utilizadas en el aprendizaje automático para seleccionar y evaluar variables predictoras en función de su capacidad para explicar o predecir un resultado específico.

Para ello se ha desarrollado un código que lleva a cabo los siguientes pasos:

1. Primero se definen las variables y conjuntos de fechas: Se inicia definiendo una lista de 13 regresores que podrían influir en el PIB y que se corresponde con la batería de indicadores seleccionada por el IECA. Luego, se establecen fechas para dividir los datos en conjuntos de entrenamiento y prueba. El entrenamiento es hasta diciembre de 2019. La prueba, todos los meses posteriores.

2. En segundo lugar, se procede a buscar la mejor combinación de regresores. Así, el código utiliza bucles anidados para iterar a través de todas las combinaciones posibles de regresores, desde combinaciones de 3 hasta 11 regresores.

3. Se entrena y evalúan los modelos. Para cada combinación de regresores, se entrena (estima) un modelo de regresión lineal con los datos de entrenamiento y se evalúa su desempeño utilizando los datos de prueba. Se calcula el Error Cuadrático Medio (MSE) para medir la precisión de las predicciones del PIB.

4. En cuarto lugar, se selecciona el mejor modelo. Durante el proceso de iteración, se realiza un seguimiento del modelo con el MSE más bajo, actualizando el "mejor modelo" y la combinación de regresores que proporcionan la mejor predicción.

5. Pero, además del MSE, se calcula una métrica adicional que evalúa la concordancia en el cambio de dirección de las predicciones del PIB en comparación con los datos reales. No solo nos interesa si se aproxima bien, sino también si la proyección genera predicciones que captan adecuadamente los cambios en los signos de las tasas de crecimiento del PIB a lo largo de los trimestres que se evalúan como período de prueba.

El proceso busca encontrar la combinación óptima de variables que minimiza el MSE y maximiza la precisión en la predicción de la dirección del PIB, contribuyendo así a identificar los factores más relevantes que impactan en la economía de Andalucía.

Siguiendo este proceso, la combinación de indicadores que minimiza el error cuadrático medio para los meses posteriores a diciembre de 2019 exportaciones de bienes, consumo de gasolina y gasóleo, matriculación de turismos y ICMA sin estaciones de servicios. Con solo cuatro indicadores es posible maximizar la capacidad de información para describir la evolución de la actividad económica andaluza similar al PIB.

Si realizamos la regresión entre el PIB y estos cuatro indicadores se obtiene el siguiente resultado:

Tabla

Descripción generada automáticamente

En la regresión se puede observar dos cuestiones. En primer lugar, los signos de los coeficientes son los esperados, salvo el de exportaciones, lo que indica posible multicolinealidad. En segundo lugar, la capacidad explicativa del modelo es elevada, pudiendo explicar el 90% de la variabilidad del PIB a lo largo de su serie completa.

Finalmente, la serie predicha del PIB por este modelo se muestra bastante coincidente con la observada del PIB. Incluso su capacidad de explicar ya no solo los meses posteriores al inicio del PIB sino los siguientes, es elevada.

**Figura 9 Comparación del PIB observado y el PIB predicho por el modelo teniendo en cuenta solo los trimestres hasta IV de 2019 y utilizando como prueba los posteriores. Mejor modelo**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Pero podríamos pensar que en este ejercicio los meses de 2020 influyen en la definición del modelo. Para ello se repite el ejercicio estimando el modelo hasta 2019 y minimizando el error entre el primer trimestre de 2021 y el tercero de 2023.

Para este ejercicio, los indicadores seleccionados serían cinco: Índice de producción agrícola – Andalucía, Consumo de gasolina y gasóleo, exportaciones e importaciones y Cifra negocios del sector servicios – Andalucía. La capacidad de predicción del modelo con estos cinco indicadores es muy elevada, del 91,2% para el conjunto del período.

Tabla

Descripción generada automáticamente

Finalmente, la representación gráfica de las series predichas y observadas de PIB muestran que solo estos tres indicadores reproducen el perfil cíclico marcado por el último, con un ajuste muy certero ya no solo para los trimestres de la pandemia sino para los posteriores.

**Figura 10 Comparación del PIB observado y el PIB predicho por el modelo teniendo en cuenta solo los trimestres hasta IV de 2019 y utilizando como prueba los posteriores a IV de 2020. Mejor modelo II.**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

No obstante, esta aproximación que por sí sola nos podría ofrecer pesos que permitieran construir un indicador muy aproximado a la evolución cíclica de la economía andaluza se antoja insuficiente.

En primer lugar, y a pesar de su capacidad de reproducir el ciclo andaluz, el uso de tan escaso número de indicadores nos ofrece la intuición de que hay información relevante que no se está incluyendo. En segundo lugar, cuanto menor sea el número de indicadores usados, más sensible será el indicador sintético a comportamientos erráticos o outliers de los indicadores usados, por muy bien que estos aproximen la evolución histórica o para los trimestres usados como prueba de la serie PIB.

Por estas razones, lo que se propone, como segundo paso en este trabajo, es el de aplicar metodologías de reducción de la dimensión del número de variables o indicadores que pueden ser usados para construir el indicador sintético. La idea, muy simple, es poder extraer de todos los indicadores presentes en la batería aportada por el IECA, toda la información relevante que permita aproximar la evolución del PIB, y con ello de la actividad económica de Andalucía.

Sin embargo, el problema del uso de esta metodología es que dejaría de lado indicadores con distinta frecuencia. Por esta razón, se propone como segundo ejercicio la aplicación de un modelo dinámico factorial desarrollado por Bańbura y Modugno (2014) y que permite, no solo rellenar en el proceso los datos “missings” sino también, emplear en un mismo proceso homogéneos, indicadores tanto mensuales como trimestrales. Esto es lo que se propone para una segunda fase del trabajo.

## 4. Definición de un análisis factorial para la estimación del ISEA

El análisis factorial dinámico (AFD) es una técnica estadística avanzada utilizada para analizar la estructura de relaciones entre variables a lo largo del tiempo, así como la estimación de componentes comunes que son usados para la reducción de la dimensionalidad. A diferencia del análisis factorial convencional, que se centra en la estructura de las variables en un único momento, el AFD examina cómo cambian las relaciones entre las variables a lo largo de múltiples períodos temporales.

La principal cualidad del AFD es que éste permite identificar patrones subyacentes, tendencias y cambios en la estructura de datos a lo largo del tiempo. Utilizando métodos matemáticos y estadísticos, descompone la variabilidad observada en los datos en factores subyacentes o latentes que explican las relaciones entre las variables en cada período de tiempo. Estos factores pueden representar tendencias, ciclos, estacionalidad u otros patrones dinámicos en los datos.

Una aplicación común del AFD es en el análisis de series temporales económicas o financieras, donde se examina la evolución de múltiples variables a lo largo del tiempo, como el crecimiento del PIB, las tasas de interés, el desempleo y otros indicadores económicos. El AFD puede ayudar a identificar factores subyacentes que contribuyen a los cambios observados en estas variables a lo largo del tiempo, lo que puede ser útil para predecir futuros movimientos económicos o para entender mejor las relaciones causales entre diferentes variables. Esta cualidad la hacen una metodología perfecta a la hora de estimar indicadores que traten de aproximar la evolución de agregados como el PIB.

Los modelos dinámicos factoriales ofrecen varias ventajas sobre los simples indicadores sintéticos obtenidos por la agregación ponderada de indicadores. Entre dichas ventajas se pueden enumerar las siguientes:

1. En primer lugar, capturan la complejidad de los datos al modelar las relaciones dinámicas entre las variables observadas. Esto es crucial en entornos económicos donde las relaciones entre las variables pueden ser no lineales o cambiar con el tiempo.
2. En segundo lugar, tienen en consideración la incertidumbre, ya que estos modelos tienen en cuenta la incertidumbre en la estimación de los parámetros del modelo, lo que proporciona además del indicador una medida de la precisión de las estimaciones.
3. Suelen mejorar la calidad de las predicciones, pues modelan las relaciones dinámicas entre las variables. Así, los modelos dinámicos factoriales tienden a ofrecer predicciones más precisas y robustas que los simples indicadores sintéticos.
4. En cuarto lugar, identifican factores latentes que subyacen a las variables observadas, lo que puede proporcionar información adicional sobre la estructura subyacente de la economía y ayudar a identificar tendencias ocultas.

Así pues, los modelos dinámicos factoriales son más sofisticados y flexibles que los simples indicadores sintéticos, lo que los hace más adecuados para capturar la complejidad y la dinámica de los datos económicos y financieros.

Dicho esto, el análisis factorial dinámico es un proceso meticuloso que implica una serie de etapas fundamentales para obtener resultados precisos y significativos. Estas etapas incluyen:

1. Especificación del modelo: En esta fase inicial, se define la estructura del modelo dinámico factorial, que incluye la identificación de las variables observadas relevantes y la formulación de las relaciones dinámicas entre ellas. Es crucial diseñar un modelo que capture de manera efectiva la complejidad de los datos y las interacciones temporales entre las variables. Además, en la especificación del modelo es necesario indicar la estructura estocástica del mismo (con la especificación de la estructura regresiva de los errores o el número de factores existentes).
2. Estimación de parámetros: Una vez especificado el modelo, se procede a estimar los parámetros de este. Esta tarea se lleva a cabo utilizando diversas técnicas estadísticas, como la máxima verosimilitud, que permiten encontrar los valores óptimos de los parámetros que mejor se ajustan a los datos observados. Es importante tener en cuenta la incertidumbre asociada a estas estimaciones y realizar pruebas de sensibilidad para evaluar su robustez.
3. Evaluación de la bondad de ajuste: Una vez estimado el modelo, se evalúa su bondad de ajuste mediante diversas métricas y pruebas estadísticas. Se comparan los valores predichos por el modelo con los datos observados para determinar si el modelo proporciona una representación precisa de la realidad. Se pueden realizar pruebas de hipótesis para verificar la significancia estadística de las relaciones estimadas y realizar diagnósticos de residuos para identificar posibles problemas en el modelo
4. Finalmente, la interpretación de los resultados: Se interpreta los resultados del análisis factorial dinámico para extraer conclusiones significativas sobre la estructura y dinámica de los datos. Esto implica identificar los factores latentes que subyacen a las variables observadas y analizar cómo estas relaciones evolucionan a lo largo del tiempo. La interpretación de los resultados puede requerir un conocimiento profundo del contexto económico y financiero, así como la aplicación de técnicas de visualización de datos para comunicar de manera efectiva los hallazgos.

Por lo tanto, el análisis factorial dinámico, al combinar la teoría estadística con el conocimiento del dominio para proporcionar información valiosa sobre la dinámica de los datos económicos y financieros, permite cumplir con los objetivos de obtener un indicador sintético de forma más adecuada que metodologías que dependen más de la subjetividad del analista. Son los datos los que hablan y son estos los que determinan la mejor aproximación que se quiere concretar en un indicador sintético.

### Especificación del modelo

En general, el modelo DFM propuesto para este análisis tiene la siguiente estructura:

(1)

Donde:

* son las variables observadas, en nuestro caso los indicadores, en el momento t, entre los que se encuentra el PIB trimestral.
* es un error idiosincrásico en el momento t que puede estar o no correlado serialmente.
* son los factores no observados en el momento t.
* ∼𝑁(0,𝑄) es un error asociado a la estructura estocástica de los factores.
* son las matrices de los pesos factoriales
* son las matrices con los coeficientes autorregresivos.

La especificación del modelo en el análisis factorial dinámico se fundamenta en la formulación de relaciones matemáticas que describan la dinámica de las variables observadas y los factores no observados a lo largo del tiempo. Así, bajo la primera ecuación del modelo (1) se describe la relación entre las variables observadas , que en nuestro caso serán los indicadores parciales suministrados por el IECA y usados en el ejercicio anteriormente planteado, en el momento y los factores no observados , ponderados por las matrices de pesos factoriales , junto con el término de error .

Bajo la segunda ecuación se define el llamado modelo de factores y que describe la evolución de los factores a lo largo del tiempo en función de sus valores anteriores, representados por los coeficientes autoregresivos , junto con el término de error .

En conjunto, estas ecuaciones proporcionan un marco teórico y matemático sólido para analizar la relación entre las variables observadas y los factores latentes, permitiendo capturar la complejidad y la dinámica de los datos a lo largo del tiempo. La especificación detallada del modelo es fundamental para garantizar la validez de los resultados obtenidos a partir del análisis factorial dinámico.

### 4.2 Optimización

El proceso de especificación del modelo en el análisis factorial dinámico sienta las bases para la posterior optimización y ajuste del modelo. Al definir las relaciones matemáticas entre las variables observadas, los factores no observados y los errores asociados, se establece un marco teórico riguroso que guía la selección de parámetros y la evaluación del ajuste del modelo a los datos empíricos. Este enfoque permite identificar la estructura subyacente de los datos y capturar de manera eficiente la variabilidad y la dinámica temporal presentes en el fenómeno estudiado. La optimización del modelo implica la búsqueda de los parámetros que maximizan la verosimilitud de los datos observados, garantizando así una representación precisa y fiel del proceso generador de los datos. De esta manera, la especificación detallada del modelo sirve como punto de partida crucial para el refinamiento y la mejora continua del modelo, contribuyendo a obtener resultados más precisos y significativos en el análisis factorial dinámico.

En este ejercicio se ha elegido, entre diversos métodos posibles para la optimización de los modelos dinámicos, el llamado algoritmo de Maximización de la Esperanza (EM). El EM es un método iterativo utilizado para estimar los parámetros de modelos estadísticos cuando hay variables no observadas o datos no presentes (missing data). Propuesto por Arthur Dempster, Nan Laird y Donald Rubin en 1977 (Dempster et al., 1977), el EM se ha convertido en una herramienta fundamental en el campo de la estadística y el aprendizaje automático.

El algoritmo EM consta de dos pasos alternativos: el paso de Expectativa (E-step) y el paso de Maximización (M-step). En el paso de Expectativa, se calculan las estimaciones de los valores esperados de las variables no observadas dados los datos observados y los parámetros actuales del modelo. En el paso de Maximización, se calculan los parámetros del modelo que maximizan la función de verosimilitud utilizando las estimaciones obtenidas en el paso anterior. Estos dos pasos se repiten iterativamente hasta que los parámetros convergen a un valor óptimo.

El algoritmo EM es especialmente útil en situaciones donde hay datos no presentes o variables latentes que afectan el modelo, pero no se pueden observar directamente. Al estimar tanto los parámetros del modelo como los valores faltantes simultáneamente, el EM proporciona una forma efectiva de abordar la incertidumbre en los datos y mejorar la precisión de las estimaciones.

Además de su aplicación en la imputación de datos no presentes, el algoritmo EM se utiliza en una amplia gama de problemas estadísticos y de aprendizaje automático, incluida la estimación de parámetros en modelos de mezcla, el ajuste de modelos de regresión con datos censurados y truncados, y la inferencia en modelos de redes Bayesianas, entre otros. Su versatilidad y eficacia lo convierten en una herramienta de elevado valor para analizar datos complejos y realizar inferencias precisas en una variedad de contextos.

Es, precisamente su cualidad para estimar y optimizar los resultados cuando algunos indicadores se presentan con datos no presentes por lo que este algoritmo ha sido elegido. Concretamente, esta ventaja lo coloca como técnica de optimación preferida en el caso de disponerse de indicadores de frecuencia mixta.

Así, y en primer lugar, debido a que es posible la existencia de datos no observados (o missings) en los diferentes vectores asociados a cada uno de los indicadores incluidos en la estimación, se pueden incluir variables de menor frecuencia como observadas en un período de mayor frecuencia específico, y para la cual los meses no observados simplemente serían tratados como missings. Por ejemplo, en un modelo mensual, se podrían incluir datos trimestrales que solo sucederían, o se observarían, en el tercer mes de cada trimestre. Para usar este método, simplemente se necesita combinar los datos en un solo conjunto de datos a la frecuencia más alta que se puede pasar a este modelo.

Por ejemplo, supongamos que estamos interesados ​​en la tasa de crecimiento del PIB real, que se mide a una frecuencia trimestral. Si el modelo factorial básico se especifica a una frecuencia mensual, que es nuestro caso, entonces la tasa de crecimiento trimestral en el tercer mes de cada trimestre, que es lo que realmente observamos, se aproxima por un promedio ponderado particular de tasas de crecimiento mensuales no observadas. Así pues, es necesario tener en cuenta este promedio móvil ponderado en particular al construir nuestro modelo, y esto es lo que hacen Bańbura et al. (2011) y Mariano y Murasawa (2010). El segundo enfoque sigue a estos autores en la construcción de un formulario de espacio de estados para modelar explícitamente las tasas de crecimiento trimestrales en términos de las tasas de crecimiento mensuales no observadas.

### 4.3 Evaluación del modelo y selección

En el proceso de estimación mediante el uso de DFM, se emplea un enfoque sistemático para evaluar y seleccionar el mejor modelo posible. Esto se logra a través de un procedimiento similar al descrito en la sección 3, pero implicando este tipo de modelos. El algoritmo, y que se asimilaría a un procedimiento denominado de “grid search” empleado en la modelización de modelos de aprendizaje automático, busca entre las múltiples combinaciones posibles de los indicadores disponibles, cuál aproxima mejor a la serie de PIB además de buscar el orden óptimo de los factores, así como el número de estos.

Así, entre todas las combinaciones posibles de indicadores, se procede a realizar un grid search para determinar las combinaciones óptimas de estos, el orden de los factores y el número de factores en el modelo DFM, y que implica la generación sistemática de diferentes configuraciones de modelos, evaluando su desempeño utilizando una métrica de evaluación, en este caso, el error cuadrático medio de la previsión del PIB. Esta métrica proporciona una medida de la precisión del modelo en predecir el valor real del PIB.

Así, se selecciona el modelo que minimiza el error cuadrático medio de la previsión del PIB como el mejor modelo entre todas las combinaciones evaluadas en el grid search. Este modelo óptimo se utiliza entonces para realizar pronósticos futuros del PIB y proporcionar información útil para la toma de decisiones económicas y políticas. Como se explicó en la sección anterior, el error cuadrático medio se estima para el período inmediatamente posterior a la Pandemia, con el objeto de evitar los sesgos en la estimación de parámetros que supone incorporar dichas fechas al cómputo de la optimización.

### 4.3 Resultados

Después de probar 8.086 posibles combinaciones diferentes, el error cuadrático medio se encuentra para un modelo DFM con dos factores y con una correlación serial de orden 2 para los residuos . Los indicadores que, bajo las anteriores premisas, minimizan esta métrica son los siguientes:

* + Importaciones de bienes,
  + Liquidación de presupuestos de la Junta de Andalucía. Capítulo 1,
  + Total afiliados SS Total,
  + Matriculación de turismos,
  + Consumo de gasolina y gasóleo,
  + Cifra negocios del sector servicios.

Estos indicadores conforman dos factores que, estimados, representarían la fuerza subyacente en el comportamiento de la economía andaluza. Una representación gráfica de los factores (denominados 0 y 1) puede comprobarse en la siguiente figura.

Interfaz de usuario gráfica, Gráfico, Aplicación, Gráfico de líneas

Descripción generada automáticamente

La siguiente tabla muestra la correlación entre cada uno de los dos factores y el conjunto de indicadores que participan en el modelo. La primera columna se corresponde con el coeficiente de determinación (lo que el factor uno explicaría de cada indicador) del factor 0 mientras que la columna segunda haría lo propio con el factor 1.

Tabla

Descripción generada automáticamente con confianza media

Como se puede comprobar, el factor 0 tendría mucha capacidad explicativa para el agregado que nos interesa aproximar con el modelo, el PIB. Este, por sí solo, explicaría un 74 %. Sin embargo, el factor 1, más asociado a la dinámica de un indicador como es la afiliación, explicaría una cuarta parte de su evolución. Sin embargo, dado que ambos factores conforman información que es ortogonal entre ellos, el uso de los dos puede resultar relevante a la hora de construir un indicador sintético del PIB y, por lo tanto, de la actividad andaluza.

La figura siguiente, que representa un mapa de calor donde los tonos y colores están identificados por la correlación de cada factor (primeras dos filas y columnas de la matriz) con el resto de los indicadores, nos ofrece una panorámica de cómo dichos factores logran explicar la dinámica de cada uno de los indicadores.

Así, los factores son ortogonales entre sí, por mera construcción matemática, mientras logran mostrar correlación elevada con una batería de indicadores. Por ejemplo, el indicador 0, que como hemos visto logra aproximar de forma cercana la evolución del PIB, estaría muy correlacionado con indicadores como el consumo de gasolina y gasóleo, el índice de producción industrial, la cifra de negocios del sector servicios y la matriculación de turismos. El factor 1, sin embargo, parece capturar mejor la información subyacente implícita en el comportamiento del mercado de trabajo.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

El siguiente paso implica transformar ambos factores en un indicador sintético., Sin embargo, esto no necesariamente se tendría que realizar de esta manera. Una posible acción hubiera sido pedir al modelo que realizara una estimación dentro de la muestra del PIB en frecuencia mensual, por lo que directamente el modelo ofrecería un resultado unívoco que podríamos considerar como un índice sintético de la actividad económica andaluza.

Sin embargo, en esta ocasión se ha procedido a realizar un paso diferente con el objeto de una mayor claridad en el resultado. Una vez se han obtenido ambos factores, estos se han usado como variables en una regresión donde la variable dependiente es el PIB trimestral. Dicha regresión ofrece ponderaciones en sus coeficientes con los que poder agregar ambos factores en uno solo. Dichas ponderaciones se aplican a los factores en frecuencia mensual con lo que se obtendría un indicador de actividad mensual que aproximaría a la evolución de la actividad económica andaluza. Lo que en último término denominamos ISEA.

Así, el resultado de e la regresión se observa en la tabla siguiente. Como se puede observar, la capacidad explicativa del modelo es elevada, un 87,9 %, mientras los dos factores aportan de forma significativa al indicador. Lo relevante de este ejercicio es que sabemos que dicha aportación no se solapa, pues ambos factores son ortogonales.

Tabla

Descripción generada automáticamente

Finalmente, el indicador se puede representar tanto en frecuencia mensual como en frecuencia trimestral. Para esta segunda se muestra la figura en el siguiente gráfico, desde el primer trimestre de 2021.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

## Implementación

La estimación e implementación del ISEA se ha llevado a cabo mediante el uso de lenguaje de programación de Python. Concretamente, para esto se han necesitado no solo la instalación de este lenguaje de programación, en su versión 3.9.7, sino que ha sido necesaria la instalación de librerías y su importación que son necesarias para llevar a cabo los pasos comentados.

Concretamente se han utilizado las librerías de numpy y pandas para la gestión de los indicadores y las bases de datos generadas a raíz de estos. Se ha utilizado la librería de statsmodels para la estimación de los modelos DFM, en concreto la clase DynamicFactorMQ. Además, se han usado librerías para obtener las representaciones gráficas necesarias.

Por lo tanto, para reproducir y calcular el ISEA sería necesario modificar y actualizar los códigos de los ficheros júpiter que han sido programados para tal fin. Sin embargo, para simplificar este procedimiento y elevar la autonomía que los equipos del IECA puedan tener en la elaboración del indicador sintético, se ha procedido a diseñar una aplicación con la que no solo pueden reproducir estos cálculos sin necesidad de modificar los códigos, sino que demás puedan establecer diseños diferentes para poder analizar posibles modelos que no se correspondan con el elegido por los autores de este informe.

Así, la aplicación diseñada para visualizar y analizar datos del ISEA lleva a cabo los siguientes pasos:

1. En primer lugar, se importan las bibliotecas de Python necesarias para el análisis de datos, como Pandas para la manipulación de datos, Streamlit para la creación de la aplicación web, y otras bibliotecas como Matplotlib, NumPy, y Statsmodels para análisis estadístico y visualización.

2. En segundo lugar, se definen las funciones necesarias de preprocesamiento de datos: Se definen varias funciones que realizan tareas de preprocesamiento de datos, como la estacionariedad de series temporales y la normalización de datos.

3. Se le pide al usuario que cargue los datos y, si lo estima oportuno, que los visualice: La aplicación permite al usuario cargar el archivo Excel que contiene los datos económicos mensuales y trimestrales de los indicadores usados para la generación del ISEA. Luego, se muestran gráficos interactivos para visualizar los datos seleccionados por el usuario.

4. Una vez realizado este paso, el programa transforma automáticamente estos datos: Se aplican técnicas de preprocesamiento a los datos cargados, como la media móvil trimestral y la estacionariedad de series temporales.

5. Se estima el modelo: Se estima un modelo de factores dinámicos utilizando la biblioteca Statsmodels. El usuario puede seleccionar los regresores (variables explicativas), el número de factores y el orden de autoregresión de los factores. El modelo se ajusta a los datos y se realiza una predicción del PIB utilizando los factores estimados. Sin embargo, recomendamos el uso de aquellos indicadores que, por defecto, minimizan el error cuadrático medio según se ha explicado en la sección 4 de este informe

6. se visualizan los resultados: Se muestra una comparación entre el crecimiento del PIB original y la predicción del modelo en un gráfico interactivo. El usuario puede observar cómo se comporta la predicción del modelo en comparación con los datos reales del PIB.

7. Se descarga el ISEA. En la localización dispuesta por el usuario se descarga un fichero Excel con la serie del indicador ISEA para el período completo para el que ha sido calculado.

## Bibliografía

Bańbura, Marta, and Michele Modugno. “Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data.” Journal of Applied Econometrics 29, no. 1 (2014): 133-160.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological), 39(1), 1-22.

Mariano, R. S., & Murasawa, Y. (2010). A coincident index, common factors, and monthly real GDP. Oxford Bulletin of economics and statistics, 72(1), 27-46.

1. En este sentido, de momento, se han procedido a usar directamente las series aportadas por el IECA con los procesos de desestacionalización y ajuste de atípicos. [↑](#footnote-ref-1)
2. Ver anexo 1 [↑](#footnote-ref-2)