Análisis Regional España

Análisis provincial de la contracción del empleo en España durante el confinamiento nacional mediante Árboles de decisión

Giancarlo Carta / Rodolfo Méndez / Pep Ruíz / Angie Suárez

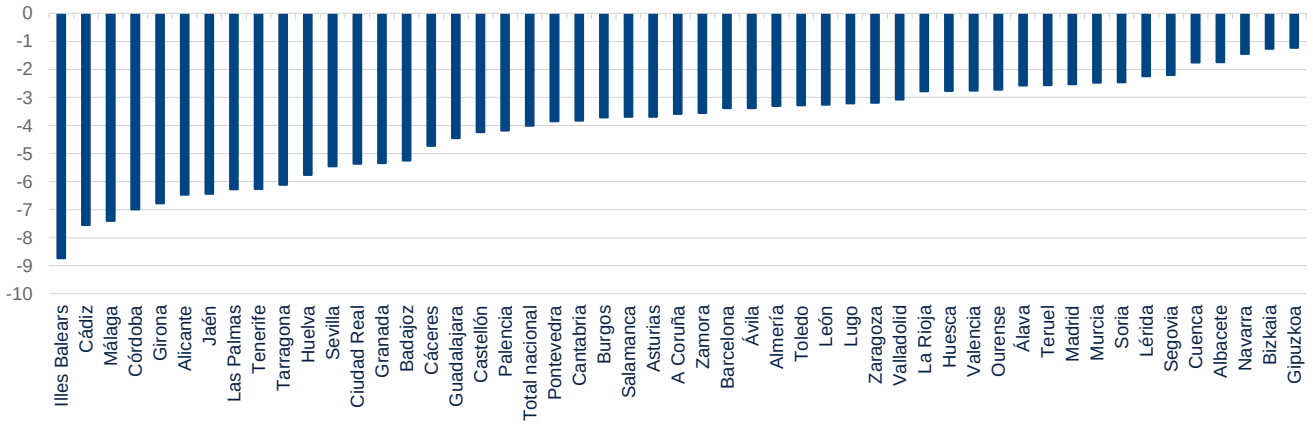
9 de diciembre de 2020

- **La diferencia de estructura productiva entre las provincias españolas constituye el determinante fundamental de la heterogeneidad en la magnitud de la fuerte contracción inicial del empleo provincial** en respuesta al confinamiento nacional impuesto por el Gobierno durante el primer cuatrimestre de 2020.
- Durante el período estudiado, **las provincias más adversamente afectadas,** en términos de empleo, por el confinamiento nacional, **fueron las de economía más dependiente de la actividad turística** y, aunque con excepciones relevantes, también de la actividad agrícola, pecuaria y pesquera.
- **Entre las menos impactadas, destacan las** provincias predominantemente **industriales** y aquellas con una **economía de servicios altamente diversificada** (principalmente aquellas que incluyen grandes ciudades).
- **Estos rasgos fueron menos relevantes** en la contracción inicial del empleo **durante la crisis inmobiliaria y financiera de 2008-2009,** siendo los más destacados en aquel entonces, el peso provincial del sector público (con implicaciones favorables) y de la banca (con implicaciones desfavorables).

1. Introducción

En esta nota se analiza e intenta explicar la heterogeneidad en la magnitud de la fuerte contracción experimentada por el empleo provincial al comienzo del período de severo confinamiento nacional impuesto por el Gobierno español a principios de 2020 con el objetivo de contener la propagación de la epidemia de COVID-19.

Gráfico 1. **ESPAÑA: TASA DE VARIACIÓN DE AFILIADOS A LA SEGURIDAD SOCIAL** (ABRIL-2020, % A/A)



Fuente: BBVA Research a partir de datos de la Seguridad Social

Adicionalmente, se contrastan los resultados obtenidos con los del análisis de la contracción del empleo provincial durante los primeros meses de la **crisis inmobiliaria y financiera de 2008**[1].

El confinamiento afectó negativamente en mayor o menor medida a la producción de bienes y servicios de casi todos los sectores productivos, lo que naturalmente se tradujo en una contracción del empleo. El efecto más directo fue el generado por la paralización impuesta sobre las actividades productivas consideradas no esenciales; pero luego están los efectos indirectos sobre la demanda de aquellos bienes y servicios cuya venta requiere la presencia física de los consumidores y sobre la oferta de bienes y servicios cuya producción requiere la presencia física de los trabajadores (muy limitada por las medidas de distanciamiento social). A todo ello se suma el efecto de la caída de rentas de los consumidores por la pérdida de empleo, contribuyendo también a la contracción de la demanda.

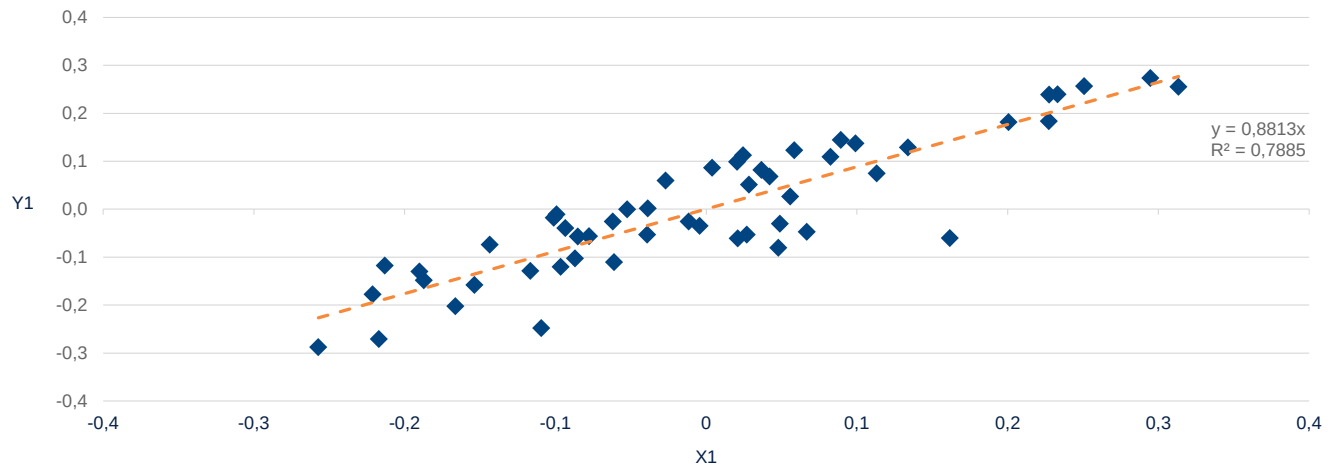
Entre los diversos factores que condicionan directamente la magnitud de estos efectos desde la perspectiva de una empresa en particular destacan los de tipo institucional, como el tipo de contratos laborales o modalidades de empleo, y de tipo tecnológico, como la mayor o menor posibilidad de vender virtualmente sus bienes y servicios (comercio electrónico) y/o de producirlo por trabajadores físicamente ausentes (teletrabajo o trabajo remoto). Pero en último término, las diferencias entre empresas en cuanto a la incidencia de estos factores están estrechamente asociadas al sector productivo o rama de actividad a la que pertenece cada una, lo que justifica la hipótesis analizada en el presente estudio de que **la heterogeneidad provincial en la respuesta del empleo a la COVID-19 podría explicarse fundamentalmente por diferencias de la estructura productiva entre provincias**.

Dejando de lado el caso de las medidas de paralización impuestas temporalmente sobre las actividades productivas consideradas “no esenciales” (que excluían principalmente a las actividades de los sectores sanitario y agroalimentario), el caso más significativo es el del sector servicios. Dentro de este, dadas las dificultades que plantea el contacto personal en el actual contexto, sería esperable un mayor impacto en los servicios personales (estética, mantenimiento físico y similares), la hostelería y la restauración. En contraste, las empresas de servicios financieros, gracias al incremento de la automatización y digitalización que han experimentado en los últimos años, han reducido de modo apreciable su dependencia tanto de la presencia física de clientes como de trabajadores en comparación al resto de servicios.

Los dos gráficos siguientes permiten ilustrar los puntos anteriores. **El Gráfico 2 deja ver la elevada correlación entre un indicador de “flexibilidad del mercado laboral” provincial** (que denominamos Y1[2]) y otro que sintetiza los rasgos de la “estructura productiva” de cada provincia (que denominamos X1), mostrando que al colocar el foco en el análisis de la estructura productiva no se obvia el papel de la flexibilidad del mercado laboral sino que tan solo permanece implícito. Por su parte, el Gráfico 3 muestra la clara correlación negativa entre el peso del sector de la hostelería (rasgo de la estructura productiva que mejor define la dependencia del sector turístico de una provincia) y la respuesta del empleo provincial durante el confinamiento inicial.

1: En el Apéndice B, también se contrastan los resultados con la heterogénea evolución del empleo provincial durante el conjunto de la fase de recuperación de la economía, previa al estallido de la pandemia.
2: Y1 y X1 se calculan mediante la técnica estadística de “correlación canónica” y representan las combinaciones lineales con mayor correlación de los dos grupos de variables siguientes: en el caso de Y1, el grupo de las variables disponibles con mayor incidencia directa en la sensibilidad del empleo provincial al confinamiento (las proporciones de, respectivamente, trabajadores con contratos temporales, trabajadores autoempleados o autónomos y grandes empresas con respecto al número total de empresas) y, en el caso de X1, el grupo de los pesos de los distintos sectores productivos en el empleo total provincial.

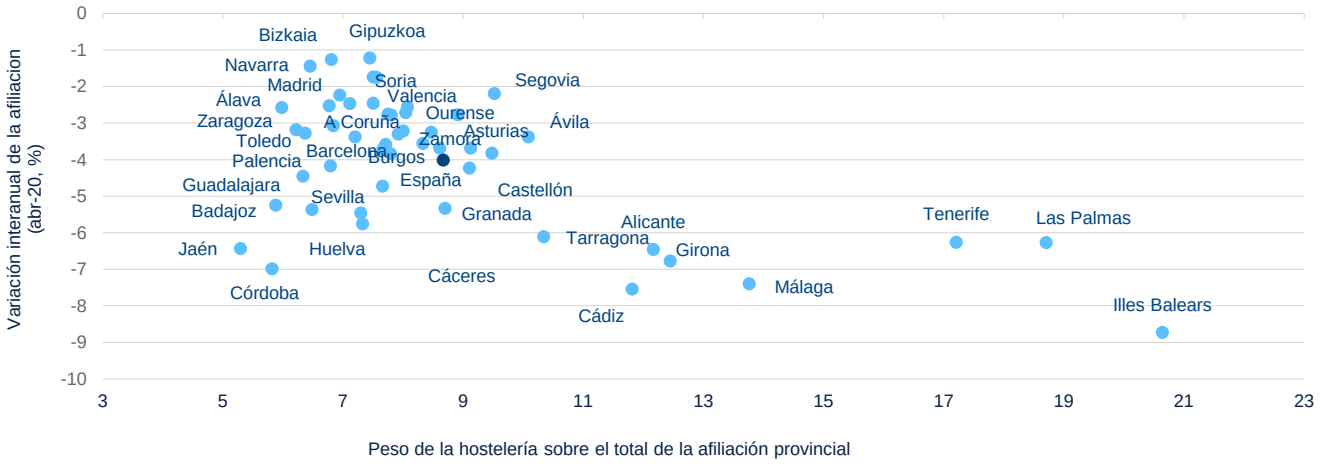
Gráfico 2. **ESTRUCTURA PRODUCTIVA (X1) Y FLEXIBILIDAD LABORAL (Y1)**



Fuente: BBVA Research a partir de datos de la Seguridad Social

Lo anterior intenta clarificar por qué el resto de este trabajo coloca el foco del análisis en el rol de la estructura productiva provincial en la explicación del fenómeno de nuestro interés, es decir, de la heterogeneidad en la magnitud de la fuerte contracción inicial del empleo provincial español en respuesta al confinamiento nacional del primer cuatrimestre de 2020. Se trata de una elección consistente con la adoptada por el Banco de España en su Informe Trimestral de la Economía Española[3] (específicamente en su Recuadro 7), que puede considerarse el antecedente más relevante en la literatura para el presente estudio.

Gráfico 3. **PESO DE LA HOSTELERÍA EN LA AFILIACIÓN A LA SEGURIDAD SOCIAL Y RESPUESTA DEL EMPLEO EN ABRIL**



Fuente: BBVA Research a partir de datos de la Seguridad Social

3: Para mayor detalle, véase https://repositorio.bde.es/bitstream/123456789/13709/1/be2003-it-Rec7.pdf

Sin embargo, el alcance y la aproximación metodológica de ambos trabajos difieren. El recuadro citado se concentra en ilustrar mediante gráficos similares al Gráfico 3, la potencial importancia del peso provincial de ciertos sectores productivos en las diferencias interprovinciales de la evolución del empleo durante la pandemia. En contraste, el presente estudio intenta precisar cuantitativamente dicha importancia, para lo cual es necesario ir más allá de la exploración de la relación del empleo provincial con el peso de cada sector productivo por separado, ya que la alta correlación entre el peso de los distintos sectores hace necesario considerarlos a todos en conjunto, y requiere asimismo considerar la posibilidad de que la relación del empleo con dichos pesos no sea simple o lineal. Es en función de estas consideraciones, que el presente estudio ha adoptado la aproximación metodológica de los "Árboles de Decisión" (*Decision Trees*), cuyos detalles técnicos se ofrecen en el apéndice final del trabajo.

El resto del observatorio se estructura como sigue: en la sección 2 se realiza una breve descripción de los datos empleados en el estudio; en la 3 se sintetizan los principales resultados obtenidos y en la 4 se extraen conclusiones.

2. Descripción de datos y variables

En los ejercicios realizados se emplearon los datos de las afiliaciones a la Seguridad Social, a nivel sectorial y provincial, como aproximación del empleo provincial[4]. Las series de afiliación son proporcionadas por el Ministerio de Inclusión, Seguridad Social y Migraciones, tienen una periodicidad mensual y están disponibles desde enero de 2009, desglosadas según la clasificación CNAE-2009 y para períodos anteriores, según la clasificación CNAE-93.

Los sectores productivos de las series de afiliación fueron reagrupados con el objetivo de aproximar el desglose sectorial del PIB en la Contabilidad Regional de España. Así, en las pruebas realizadas se evaluaron 11 sectores a partir de un total de 21 de acuerdo a la clasificación CNAE 2009 (ver Cuadro 1).

4: La alternativa a esta serie son los datos de la Encuesta de Población Activa (EPA) elaborada por el Instituto Nacional de Estadística. Sin embargo, la frecuencia trimestral de la serie impide realizar el análisis del impacto del confinamiento español en abril de 2020, viéndose este difuminado a lo largo del trimestre. Por otro lado, la afiliación a la Seguridad Social es un registro del número de trabajadores en alta laboral en los distintos regímenes, mientras que el empleo EPA se obtiene de una encuesta referida al empleo que reside en hogares familiares, lo que puede perder representatividad cuando se cruza un elevado desglose geográfico y sectorial.

Cuadro1. **AGRUPACIÓN SECTORIAL DE LA AFILIACIÓN**

Clasificación sectorial CNAE 2009	Agrupación sectorial utilizada en el análisis
(A) Agric., Gana. Silv. Y Pesca	(A) Agric., Gana. Silv. Y Pesca
(B) Ind. Extractivas	(B+D+E) Industria y suministros
(D) Suminis. Energía	
(E)Suminis. agua, resid.	
(C)Ind. Manufacturera	(C) Ind. Manufacturera
(F) Construcción	(F) Construcción
(G) Comercio. Rep. Vehícul.	(G+H) Comercio y transporte
(H) Transptes. Almacena.	
(I) Hostelería	(I) Hostelería
(J) Informac. Comunicac.	(J+K+L) Actividades Financ e Inmob
(K) Act. Financ. y Seguros	
(L) Act. Inmobiliarias	
(M) Actv .Prof.Cient. Téc.	(M+N) Actividades profesionales
(N) Actv .Admt. Serv.Auxil.	
(O) Admón Púb. Defen., S.S.	(O+P+Q) Administración Pública, Defensa y Sanidad
(P) Educación	
(Q) Actv .Sanit. Serv.Sociales.	
(R)Actv .Artis. Rec.y Entr.	(R+S) Actividades artisticas
(S) Otros Servicios	
(T) Hogares P. Domést.	(S+T+U) Resto
(U) Org. Extraterritoriales	

Fuente: BBVA Research a partir de datos de la Seguridad Social

Definición de variables o atributos objetivo y descriptivos (explicativos)

En las estimaciones realizadas, la variable dependiente se construye a partir de la tasa de variación de la afiliación provincial cuyo periodo se define según el evento a analizar.

La metodología exige que el problema a estudiar se formule como uno de clasificación o de predicción de clase: esto es, decidir a qué clase o grupo pertenece cada individuo en función de ciertas características o atributos individuales. Para lo cual es necesario que la variable a explicar (en nuestro caso, la tasa de variación de la afiliación provincial) sea discretizada y transformada en una variable categórica.

En este caso, se ha optado por la opción más simple, utilizar la mediana de la distribución de la variable objetivo o a explicar, para reexpresarla como una variable categórica que indique simplemente a qué grupo pertenece cada provincia: al grupo con un valor de la variable objetivo original mayor que la mediana, o al grupo con un valor menor o igual que la mediana.

Por ejemplo en el análisis del impacto inicial del confinamiento domiciliario en respuesta a la epidemia de COVID-19, en el que la variable objetivo se refiere a la tasa de variación del total de afiliados a la Seguridad Social entre abril 2019 y abril 2020, esta se reexpresó en la forma de una variable categórica binaria indicando, para cada provincia, a cuál de estos dos grupos o clases pertenece: el Grupo A formado por las provincias que experimentaron una caída del empleo menor o igual a la mediana (tasa de variación de las afiliaciones mayor o igual a la mediana) y el Grupo D formado por las provincias que experimentaron una caída del empleo mayor a la mediana (una tasa de variación de las afiliaciones menor a la mediana).

En el caso de los dos eventos disruptivos o shocks considerados (COVID-19 y crisis inmobiliaria y financiera) el criterio para establecer el periodo específico de dicho evento se desprende de las tasas de variación intermensual e interanual que experimentó la afiliación del conjunto de España[5].

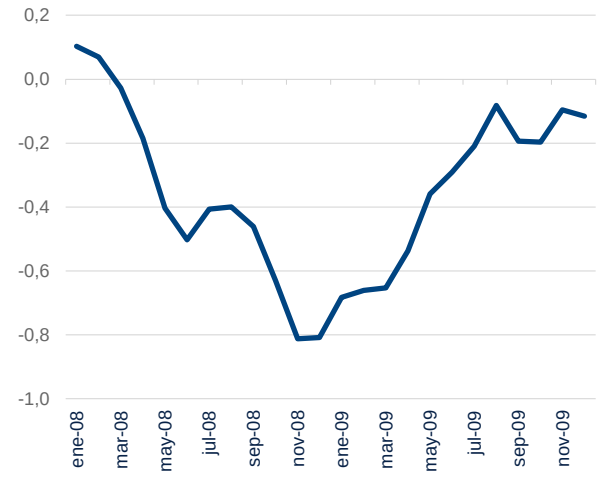
Así, para el análisis de la COVID-19, se selecciona como variable objetivo principal la tasa de variación interanual del número total de afiliados provinciales en Abril 2020 (es decir, su variación acumulada porcentual entre Abril 2019 y Abril 2020), pues como muestra el Gráfico 5, este es el mes donde la tasa intermensual de variación de la serie de afiliados nacionales alcanza su mínimo en 2020.

En lo que respecta al análisis, complementario, de la crisis inmobiliaria y financiera, la variable objetivo principal es la tasa de variación interanual de Abril de 2009 (es decir, su variación acumulada porcentual entre Abril de 2008 y Abril de 2009) que, como muestra el Gráfico 6, es la mínima tasa de variación interanual del período 2008-2009 (mientras que, como muestra el Gráfico 4, la mínima tasa intermensual se alcanzó en Diciembre 2008).

El último ejercicio considerado ya no trata de un evento disruptivo, sino del desempeño promedio del empleo durante la fase de recuperación económica previa a la COVID-19, que se ha definido como el período que abarca los años que van de 2014 a 2019. Por tanto, en este caso, la variable objetivo no se refiere a la tasa de variación interanual medida en un mes particular, sino al promedio de su valor para todo el período.

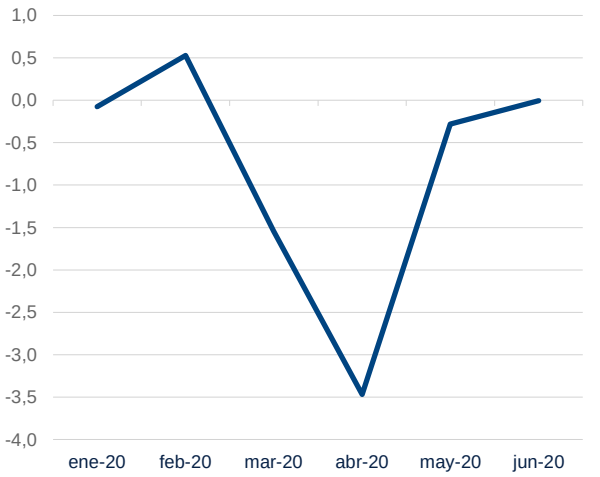
5: En el caso del cálculo de la tasa de variación intermensual se utilizan las series de afiliación corregidas de estacionalidad y efecto calendario ya que se están comparando meses con factores estacionales distintos en las dos crisis, no así en el caso de las tasas de variación interanuales.

Gráfico 4. **ESPAÑA: VARIACIÓN DE LA AFILIACIÓN TOTAL A LA SEGURIDAD SOCIAL** (% m/m, CVEC)



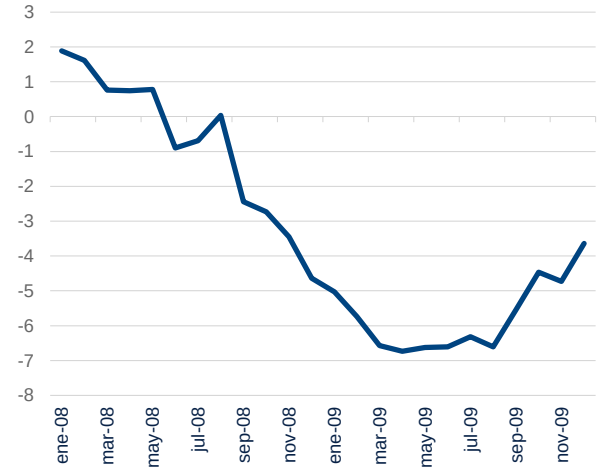
Fuente: BBVA Research a partir de datos de la Seguridad Social

Gráfico 5. **ESPAÑA: VARIACIÓN DE LA AFILIACIÓN TOTAL A LA SEGURIDAD SOCIAL** (% m/m, CVEC)



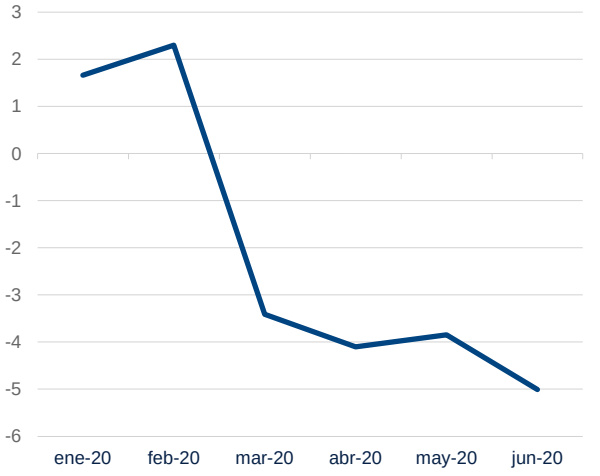
Fuente: BBVA Research a partir de datos de la Seguridad Social

Gráfico 6. **ESPAÑA: VARIACIÓN DE LA AFILIACIÓN TOTAL A LA SEGURIDAD SOCIAL** (% a/a, CVEC)



Fuente: BBVA Research a partir de datos de la Seguridad Social

Gráfico 7. **ESPAÑA: VARIACIÓN DE LA AFILIACIÓN TOTAL A LA SEGURIDAD SOCIAL** (% a/a, CVEC)



Fuente: BBVA Research a partir de datos de la Seguridad Social

En cada caso, las variables explicativas se refieren a la estructura de pesos del empleo sectorial (estructura productiva) previa al evento o período para el que se define la variable objetivo. En el caso de la COVID-19, se refiere a la estructura productiva promedio del período 2009-2019; en el de la crisis financiera, a la del período año 2006-2007 y en el caso de la fase de recuperación, a la del año 2013 (los años inmediatamente anteriores se consideran aún atípicos, por los efectos de la crisis de deuda europea).

El Cuadro 2 muestra en detalle las variables (objetivo y explicativas) utilizadas por los cuatro árboles que fundamentan nuestros resultados. El apéndice A contiene una muestra de ejercicios adicionales realizados a efectos de evaluar la sensibilidad y robustez de los resultados de estos cuatro árboles (básicamente se trató de utilizar variaciones en la manera de construir las variables objetivo y explicativas).

Cuadro 2. **VARIABLE OBJETIVO Y DESCRIPTIVA DE LAS ESTIMACIONES**

	Variable Objetivo	**Variable Descriptiva**
Crisis Sanitaria COVID-19 (2020)	*Variable categórica binaria* que recoge a cuál de las siguientes dos clases pertenece la provincia en función de la *tasa de variación interanual del total de afiliados entre Abril 2019 y Abril 2020* (ÁRBOL 1): **Clase A:** caída de la afiliación igual o menor a la mediana* **Clase D:** caída de la afiliación mayor a la mediana	Peso promedio de cada uno de los 11 sectores productivos sobre el total de afiliados para el período 2009-2019 (ver Cuadro 1)
Crisis Inmobiliaria y Financiera (2008-2009)	*Variable categórica binaria* que recoge a cuál de las dos clases (A y D) pertenece la provincia en función de la *tasa de variación interanual del total de afiliados entre Abril 2008 y Abril 2009* (ÁRBOL 2)	Peso promedio de cada sector productivo sobre el total de afiliados del período 2006-2007
Fase de expansión (2014-2019)	1) *Variable categórica binaria* indicando a cuál de los siguientes dos clases pertenece la provincia en función del *coeficiente de variación (volatilidad) de la tasa de variación interanual del total de afiliados para el período 2014-2019* (ÁRBOL 3): **Clase A:** coeficiente de variación menor que la mediana **Clase D:** coeficiente de variación mayor que la mediana 2) *Variable categórica binaria* indicando a cuál de los siguientes dos clases pertenece en función del *promedio de la tasa de variación interanual del número de afiliados durante el período 2014-2019* (ÁRBOL 4)*:* **Clase A:** tasa de variación menor que la mediana **Clase D**: tasa de variación mayor que la mediana	Peso promedio de cada sector productivo sobre el total de afiliados del año 2013

* Dado que una caída corresponde al valor absoluto de la tasa de variación cuando esta última es negativa, entonces una tasa de variación igual o mayor a la mediana equivale a una caída menor que la mediana, mientras que una tasa de variación menor que la mediana equivale a una caída mayor que la mediana.
Fuente: BBVA Research

3. Resultados

A continuación se presentan y analizan los principales resultados de la aplicación de la metodología de árboles de decisión a los cuatro problemas planteados (Cuadro 2). Es decir, tanto al problema central (determinar el papel de la estructura productiva provincial en la heterogénea respuesta inicial del empleo a la COVID-19), como al ejercicio complementario de determinar el papel de la estructura productiva provincial en la heterogénea respuesta inicial del empleo a la crisis financiera e inmobiliaria de 2008[6].

Crisis de la COVID-19

En líneas generales, el árbol estimado (ver Árbol 1) evidencia que las consecuencias económicas asociadas a la pandemia habrían impactado principalmente a las provincias más dependientes de los sectores turístico o agrario y a las enfocadas a servicios de menor valor añadido. Mientras que la mayor diversificación de las áreas urbanas o el mayor peso de la industria habrían permitido una mayor fortaleza. Se recuerda que la estimación de este árbol utiliza como variable objetivo la tasa de variación del total de afiliados entre Abril 2019 y Abril 2020 y como atributo descriptivo el peso porcentual promedio de cada sector productivo en el total de las afiliaciones de cada provincia durante el período 2009-2019 (véase el Cuadro 2). En su conjunto, el árbol muestra un elevado poder predictivo, con una precisión *in-sample* del 90% y *out-of-sample* del 84%[7].

En detalle, se nota como **la hostelería sería el sector con mayor poder discriminante a la hora de separar las provincias según el mayor o menor grado de ajuste observado en su mercado laboral. En particular, las provincias con un peso de los afiliados en este sector sobre el total mayor que un 9,4% experimentaron, en casi todos los casos, una variación de la afiliación más negativa que la mediana en el periodo considerado**. Así, el primer grupo en formarse (nodo 11), dominado por la clase D, puede considerarse como el de los **principales destinos turísticos de España** ya que incluye las islas y las provincias de la costa mediterránea. El mayor grado de afectación de estas zonas es bastante evidente si se considera que las restricciones a los movimientos introducidas con el confinamiento han impactado de forma diferencial a los sectores de la economía más dependientes de los flujos de personas, como el turismo y el consumo social[8].

En segundo lugar, **dentro de las provincias con un porcentaje de la hostelería inferior al umbral crítico (9,4%), se observa que un peso del sector agrario y pesquero mayor que el 18,1% se asoció a una respuesta más negativa** del mercado laboral. Este segundo grupo (nodo 10) recoge, con algunas excepciones, las provincias agrarias situadas en el centro-sur del país (Andalucía y Extremadura). La debilidad de este sector puede estar asociada tanto a una mayor temporalidad del mercado laboral como a la mayor dependencia de jornaleros provenientes de otros países, cuya llegada se vio limitada por el cierre de fronteras debido a la pandemia. Sin embargo, cabe destacar que las provincias agrarias a pesar de haberse visto más afectadas en el arranque de la pandemia, habrían mostrado en los meses siguientes una fortaleza y una recuperación más rápida. Esto, probablemente, como consecuencia de su naturaleza como actividad económica esencial.

El tercer sector relevante habría sido el **comercio**, cuyo umbral crítico se sitúa en el 18,9%, por debajo del cual la afiliación experimentó una caída menos acentuada que la mediana. En este grupo (nodo 4) se incluyen aquellas

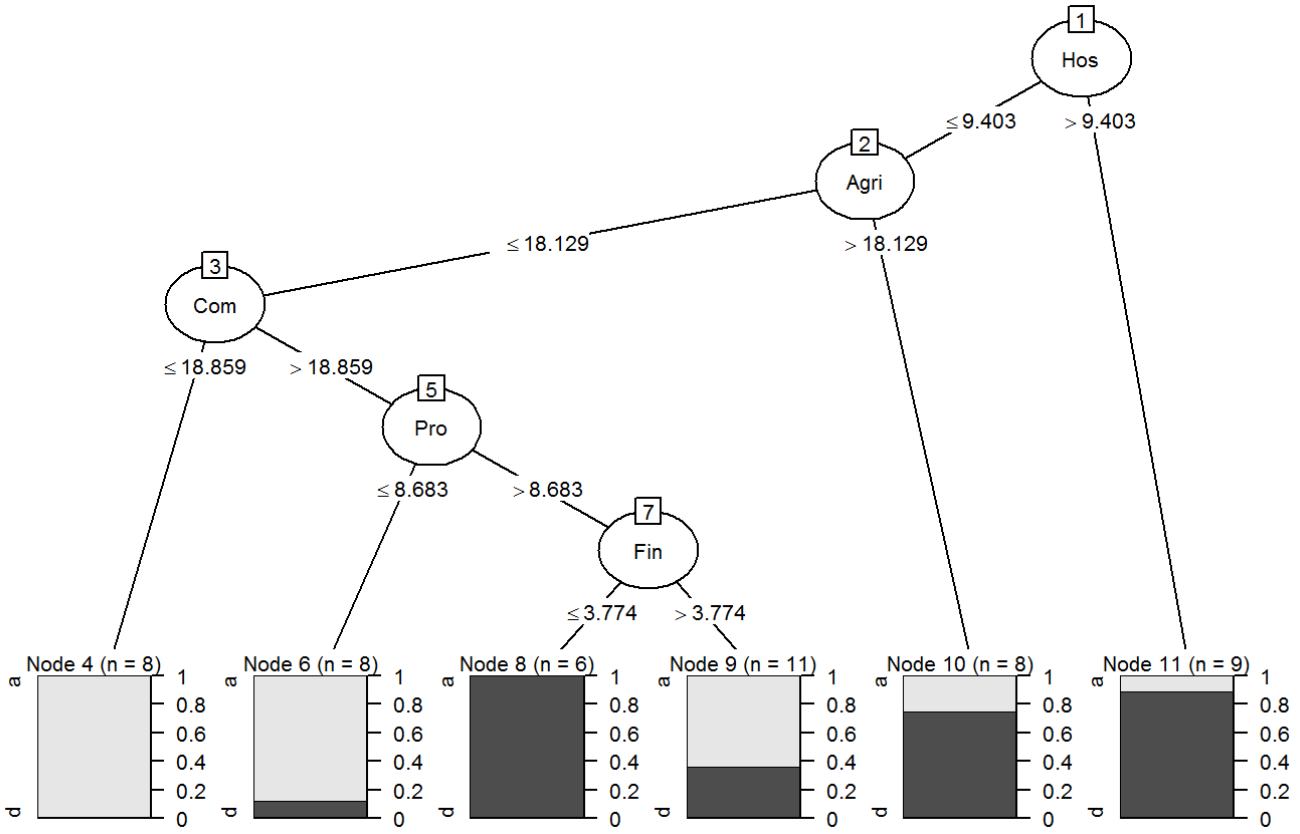
6: En el Apéndice B se presentan los resultados del ejercicio adicional: determinar el papel de la estructura productiva provincial en el comportamiento del empleo durante la fase de recuperación previa a la COVID-19.
7: Con precisión *in-sample* se entiende el porcentaje de provincias correctamente clasificadas por el árbol según la variable objetivo sobre el total de provincias en la muestra de estimación (*entrenamiento*), mientras que la precisión *out-of sample* se refiere al promedio de los aciertos en la clasificación de las provincias en un proceso de *cross-validation leave-one-out*, esto es, estimando el árbol repetidas veces en cada caso con una muestra que deja fuera una provincia distinta, y utilizando esta última para evaluar la precisión predictiva del árbol estimado sin su participación.
8: Servicios que normalmente se consumen de forma grupal: como bares, restauración, cine, discotecas, teatro, etc.

provincias que presentan un peso de la hostelería, de la agricultura y del comercio menor que el 9,4%, 18,1% y 18,9%, respectivamente. Dicho nodo engloba a las **provincias más industriales** y, en mayor medida, a las situadas en el noreste del país. La menor temporalidad, la mayor resiliencia y el mayor recurso a instrumentos de protección laboral de este sector podría explicar, en parte, el mejor comportamiento relativo de estas provincias.

Dentro del conjunto de provincias con un peso del comercio superior al 18,9% se identifican tres subgrupos diferentes:

- Nodo 6: **provincias de servicios "tradicionales"**, es decir las que presentan un peso de actividades profesionales y actividades administrativas inferior al 8,7%. Se trata en gran parte de zonas del interior del país, que mostraron una caída de la afiliación inferior a la mediana.
- Nodo 9: **provincias de servicios muy diversificadas y de alto valor añadido**. Incluye zonas con un peso de las actividades profesionales mayor que el 8,7% y de las actividades financieras, inmobiliarias, de información y comunicación mayor que el 3,8%. En su mayoría, este nodo incorpora a las grandes áreas urbanas y capitales del país, cuya economía más diversificada y enfocada a servicios que posibilitan realizar un mayor porcentaje de tareas en remoto les permitió mostrar una menor corrección.
- Nodo 8: **provincias con peso elevado de servicios profesionales (mayor que 8,7%) pero de menor valor añadido** (peso de actividades financieras menor que el 3,8%). Se trata de provincias con áreas urbanas medianas, cuyo menor enfoque hacia los servicios más avanzados y dependencia de las profesiones liberales y de empresas de menor tamaño podrían estar detrás de una caída más profunda de la afiliación.

Cuadro 2. **ÁRBOL 1: CRISIS SANITARIA** (COVID-19) - **VARIACIÓN INTERANUAL DE LA AFILIACIÓN, ABRIL 2020** (A/A, %) **- ESTRUCTURA PRODUCTIVA** (2009-2019) **- PRECISIÓN PREDICTIVA: 90%** (*OUT-OF-SAMPLE* 84%)



Nodo 4	Nodo 6	Nodo 8	Nodo 9	Nodo 10	Nodo 11
Álava	Lugo	Burgos	Madrid	A Coruña	Alicante
Guipúzcoa	Albacete	Cantabria	Asturias*	Almería*	Ávila*
Huesca	Ciudad Real*	Castellón	Barcelona	Badajoz	Baleares
La Rioja	Lérida	Guadalajara	Córdoba*	Cáceres	Cádiz
Navarra	Murcia	Palencia	León	Cuenca*	Girona
Segovia	Ourense	Salamanca	Pontevedra*	Granada	Las Palmas
Soria	Toledo		Sevilla*	Huelva	Málaga
Teruel	Zamora		Valencia	Jaén	S. C. Tenerife
			Valladolid		Tarragona
			Vizcaya		
			Zaragoza		

* Provincias que han tenido una evolución de la afiliación distinta a la de la mayoría del mismo nodo.
Fuente: BBVA Research

Crisis inmobiliaria y financiera

Para la evaluación del rol de la estructura productiva en el impacto de la crisis inmobiliaria y financiera en el empleo provincial, se realizan dos ejercicios. Por un lado, se evalúa la capacidad predictiva del Árbol 1 (COVID-19) para entender la crisis de 2008, y por otro, se estima un nuevo árbol[9].

Evaluación de la capacidad predictiva del Árbol 1 en el impacto de la crisis inmobiliaria y financiera

Para la realización de este ejercicio, se sustituyó la variable dependiente en el Árbol 1 (COVID-19) con la variable escogida para la evaluación de la crisis inmobiliaria y financiera (la tasa de variación interanual del total de afiliados entre abril 2008 y abril 2009), sin cambiar la estructura del árbol ni la composición de los nodos finales. Como resultado, se produjo un cambio en la pureza[10] de los nodos y también en la precisión predictiva del árbol, que bajó hasta el entorno del 60-65%, lo que evidencia que las dinámicas y los factores a nivel sectorial que han actuado detrás de la respuesta de las provincias a la crisis financiera habrían sido parcialmente diferentes a las de la crisis de la COVID-19.

En particular, **el grupo de las provincias turísticas (nodo 11 del Árbol 1) queda bien identificado**, es decir la respuesta a la variable objetivo es la misma en todas las que pertenecen a dicho grupo. Este comportamiento se nota en parte también en las **provincias industriales (nodo 4)**, mientras que en los otros nodos se observa una respuesta distinta según va cambiando la variable objetivo, lo que baja la capacidad predictiva y la pureza en estos grupos.

Estimación de un Árbol de Decisión para la crisis inmobiliaria y financiera

Dado la inadecuación de la estructura del Árbol 1 (COVID-19) para dar cuenta del comportamiento del empleo provincial durante la crisis de 2008, en esta subsección se presenta el resultado de la estimación de un nuevo árbol para identificar los patrones sectoriales que han sido relevantes durante la crisis de 2008. Con este propósito, a partir de la variable *objetivo tasa de variación interanual de las afiliaciones de Abril 2009* y la composición sectorial del mercado laboral de cada provincia en el periodo de 2006-2007 como variable explicativa se estima un nuevo árbol de decisión.

Los resultados muestran que las provincias más afectadas por la crisis de 2008 fueron las que tenían poco peso del sector público (que coinciden con las provincias que tiene un elevado peso del sector turístico), mientras que las industriales y las provincias de interior se vieron menos perjudicadas. Cabe destacar además un comportamiento menos negativo que la mediana también en las provincias caracterizadas por un peso menos elevado de las actividades financieras e inmobiliarias. Además la precisión predictiva tanto *in-sample* como *out-of-sample* se sitúa en niveles bastante altos.

En detalle, **el Árbol 2 identifica que el factor principal es el peso del sector público en la afiliación**. En particular, un porcentaje de afiliados a la **administración pública inferior al 14,5%**, se asoció a un retroceso del empleo más intenso que la mediana, para el periodo considerado. Lo anterior podría estar ligado a la menor

9: Estos dos ejercicios se realizaron también utilizando como variable objetivo la tasa de variación interanual de la afiliación total entre diciembre 2007 y diciembre 2008, al ser este último aquel mes donde se produjo la mayor caída de la afiliación en términos intermensuales en la pasada crisis. Los resultados de la evaluación de la capacidad predictiva del Árbol 1 son similares para las dos variables objetivo (abril 09 y diciembre 08). En el apéndice se incluye el árbol de decisión resultado de la estimación de diciembre 08 (Árbol 9).
10: Por pureza de nodo se entiende el porcentaje de provincias en un nodo que muestran la misma respuesta con respecto a la variable objetivo.

volatilidad del empleo público en momentos de crisis y a su papel suavizador de los ciclos económicos. Así, en este primer grupo (nodo 2) se incluye básicamente a los **destinos turísticos del Levante** (Balears y toda la costa mediterránea, desde Málaga hasta Girona).

En segundo lugar, **junto a un peso de la afiliación pública mayor que el 14,5%, un porcentaje de actividades financieras e inmobiliarias inferior al 9,0%** permitió un comportamiento del mercado laboral algo más favorable que la mediana (nodo 4). Dicho resultado está bastante en línea con la naturaleza de la crisis de 2008, que golpeó con mayor fuerza en los sectores ligados a la vivienda y las organizaciones financieras. Desde un punto de vista geográfico, el grupo del nodo 4 incluye en gran parte **provincias de interior poco pobladas y alejadas de los centros económicos y poblacionales del país**, cuya exposición a los sectores mencionados anteriormente es bastante marginal.

Por su parte, el nodo 6 recoge las provincias con un peso de la administración pública y de las actividades financieras superior a los umbrales especificados arriba y con un porcentaje de otras actividades sociales (incluye sectores como arte y entretenimiento) inferior al 3,7%. Las cuatro provincias (también en este caso zonas de interior) que forman parte de este grupo presentaron un retroceso del empleo más acentuado que la mediana.

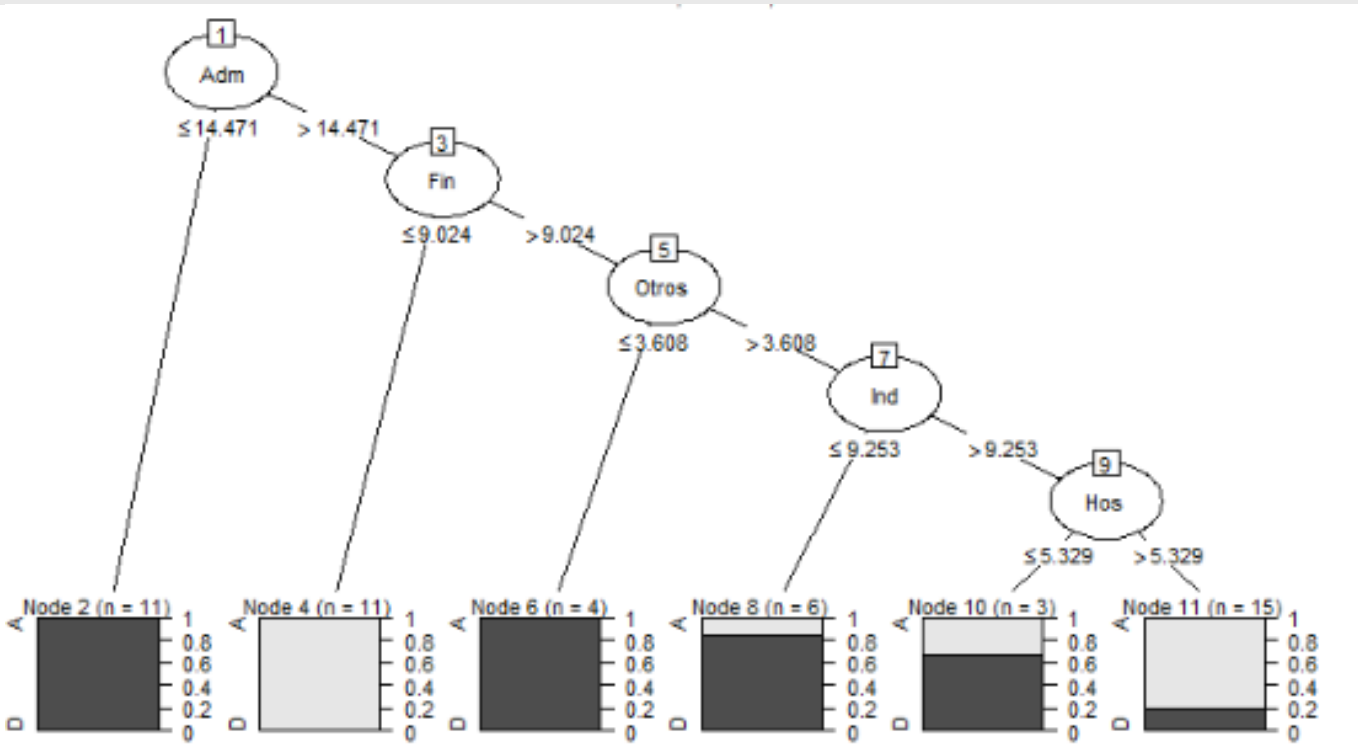
Finalmente, el **sector industrial** identifica a tres subgrupos diferentes:

- provincias con un **peso de la industria inferior al 9,3%** (nodo 8). Incluye el resto de destinos turísticos españoles no situados en el Levante.
- provincias industriales (peso > 9,3%) y con **baja relevancia de la hostelería** (<5,3%): noreste del país.
- provincias industriales pero con un **peso de la hostelería reseñable**.

Mientras que los primeros dos grupos experimentaron en su mayoría una caída de la afiliación más profunda que la mediana, el tercero se vio algo menos afectado.

Puede resultar extraño que por la naturaleza muy peculiar de la crisis de 2008, el sector de la construcción no aparezca como factor relevante. Este hecho podría en parte estar recogido en el primer grupo grupo (nodo 2) que identifica a la costa mediterránea y en las comunidades insulares, donde destacó el papel del *boom* inmobiliario.

Cuadro 3. **ÁRBOL 2: CRISIS INMOBILIARIA Y FINANCIERA DE 2008 - VARIACIÓN DE LA AFILIACIÓN, ABRIL 2009** (A/A, %) **- ESTRUCTURA PRODUCTIVA** (2006-2007) **- PRECISIÓN PREDICTIVA: 80%** (*OUT-OF-SAMPLE* 96%)



Nodo 2	Nodo 4	Nodo 6	Nodo 8	Nodo 10	Nodo 11
Alicante	A Coruña	Albacete	Tenerife	Álava*	Asturias
Almería	Badajoz	Ávila	Cádiz	Navarra	Barcelona*
Baleares	Cáceres	Burgos	Granada	Zaragoza	Cantabria
Castellón	Ciudad Real	Toledo	Madrid*		Córdoba
Girona	Cuenca		Las Palmas		Guadalajara
La Rioja	Huelva		Sevilla		Guipúzcoa
Málaga	Jaén				Huesca
Murcia	Lugo				León
Pontevedra	Soria				Lérida*
Tarragona	Teruel				Ourense
Valencia	Zamora				Palencia*
					Salamanca
					Segovia
					Valladolid
					Vizcaya

* Provincias que han tenido una evolución de la afiliación distinta a la de la mayoría del mismo nodo.
Fuente: BBVA Research

4. Conclusiones

En consistencia con el análisis económico, la aplicación de la metodología de Árboles de Decisión corrobora la importancia fundamental de la estructura productiva en la explicación de la heterogeneidad en la respuesta del empleo provincial a las medidas gubernamentales iniciales (y las más severas y generalizadas) de contención de la epidemia de la COVID-19.

Los resultados muestran que los rasgos más característicos de la estructura productiva del conjunto de provincias más adversamente afectadas por dichas medidas son una elevada dependencia del sector turístico (identificable por un peso del sector de la hostelería en el empleo provincial superior al 9,4%) o, alternativamente, y con importantes excepciones, una elevada dependencia del sector agrícola[11] (identificable por un peso de este sector superior al 18,1%). Ello, a pesar del trato favorable dado al sector por las medidas (que no habría podido contrarrestar sus consecuencias adversas sobre la disponibilidad de mano de obra en el período considerado).

Por el contrario, los rasgos adicionales más característicos de la estructura productiva del conjunto de provincias menos afectadas (obviamente, todas ellas caracterizadas por un peso de sus sectores hostelero y agrícola inferior a los umbrales citados en el párrafo anterior) son un elevado peso del sector industrial (identificable por un peso del sector comercial inferior al 18,8%) o, alternativamente, poseer una economía de servicios altamente diversificada (peso de los servicios profesionales y de los servicios financieros superiores al 8,7% y 3,7% respectivamente), como es el caso de las provincias en las que se ubican las grandes áreas urbanas.

El análisis realizado, a efectos comparativos, del rol de la estructura productiva en la respuesta del empleo provincial a la crisis inmobiliaria y financiera de 2008, solo comparte con los anteriores la homogeneidad en la intensidad de sus respuestas entre las provincias muy dependientes del turismo y, por otra parte, entre las muy dependientes de la actividad industrial. Pero los rasgos más característicos del grupo de provincias más adversamente impactadas en este caso son un peso reducido de la administración pública (inferior al 14,5%, que coincide con las provincias donde se observó el mayor boom inmobiliario) y, alternativamente, un peso muy elevado del sector bancario (identificable por un peso de los servicios financieros superior al 9%).

11: Incluyendo ganadería, silvicultura y pesca.

Apéndice A: Alternativas, Sensibilidad y Robustez

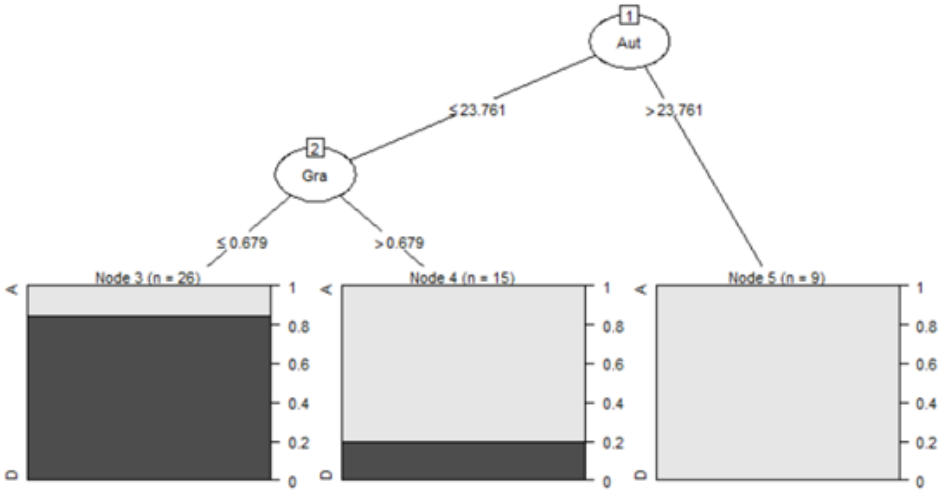
Crisis Sanitaria (COVID-19)

En el análisis de la crisis sanitaria se realizaron estimaciones adicionales a partir de variables objetivo alternativas, sin embargo fueron descartadas porque presentaban una capacidad predictiva inferior comparada con la obtenida a partir del Árbol 1, así como un menor nivel de desagregación. A continuación se presentan cuatro de estas estimaciones relacionadas con variantes de los atributos explicativos y con el componente de los ERTE.

Estimación del árbol de decisión a partir de atributos alternativos (excluyendo la estructura productiva)

Esta prueba excluye la estructura productiva empleada en el Árbol 1 presentado en la sección de resultados y en su lugar se emplean en la estimación tres atributos explicativos alternativos: el peso de los autónomos sobre el total de afiliados provinciales, la proporción de empresas con 50 o más empleados y el grado de temporalidad. El resultado que se obtiene se limita a dos nodos y la capacidad predictiva disminuye.

CUADRO 3. **ÁRBOL 5: ATRIBUTOS ALTERNATIVOS** (EXCLUYE ESTRUCTURA PRODUCTIVA) **- VARIACIÓN INTERANUAL DE LA AFILIACIÓN, ABRIL 2020** (A/A, %) **- PRECISIÓN PREDICTIVA: 88%** (*OUT-OF-SAMPLE* 78%)



Fuente: BBVA Research

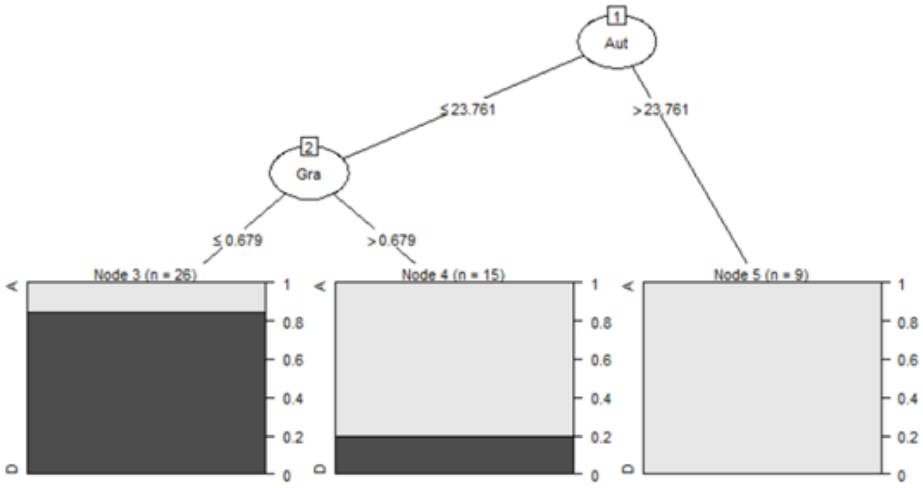
Estimación del árbol de decisión a partir de todos los atributos

A diferencia del Árbol 5, en esta estimación se combinan todos los atributos explicativos, es decir se incluye el peso de los autónomos sobre la afiliación total, la proporción de empresas con 50 o más empleados y el grado de temporalidad, además de la estructura productiva provincial.

El resultado de esta prueba es idéntico al obtenido en la estimación en la que se descarta la estructura productiva (Árbol 5) por lo que se realizó un análisis de correlación canónica: el mismo se basa en encontrar una combinación lineal de la estructura productiva, por una parte, y del trío (temporalidad, grandes empresas, autónomos), por otra, para que la correlación sea máxima. Los resultados evidencian que las dos muestran una

correlación de 88%, por lo que la estructura productiva de alguna forma está vinculada a la temporalidad o al peso de los autónomos y grandes empresas. Se concluye que es desaconsejable combinar ambos grupos en un mismo árbol de decisión.

Cuadro 4. **ÁRBOL 6: ATRIBUTOS ALTERNATIVOS JUNTO A LA ESTRUCTURA PRODUCTIVA - VARIACIÓN INTERANUAL DE LA AFILIACIÓN, ABRIL 2020** (A/A, %) **- PRECISIÓN PREDICTIVA: 88%** (*OUT-OF-SAMPLE* 80%)



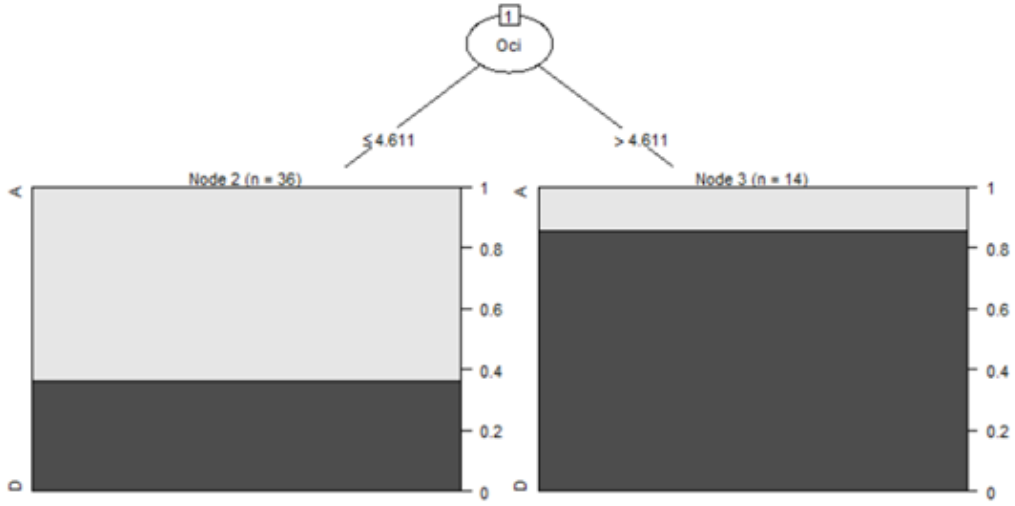
Fuente: BBVA Research

Estimación del árbol de decisión a partir de los datos de afiliación descontando los ERTE como variable objetivo y alternativamente la proporción de afiliados en ERTE

El análisis a partir de la afiliación descontando los ERTE como variable objetivo empobrece la interpretabilidad y se limita a dibujar un único nudo (Árbol 7), por lo que se estima un nuevo árbol incluyendo simplemente la proporción de afiliados en ERTE (Árbol 8) y se obtiene una mayor desagregación que el Árbol 7, pero con una capacidad predictiva fuera de muestra menor.

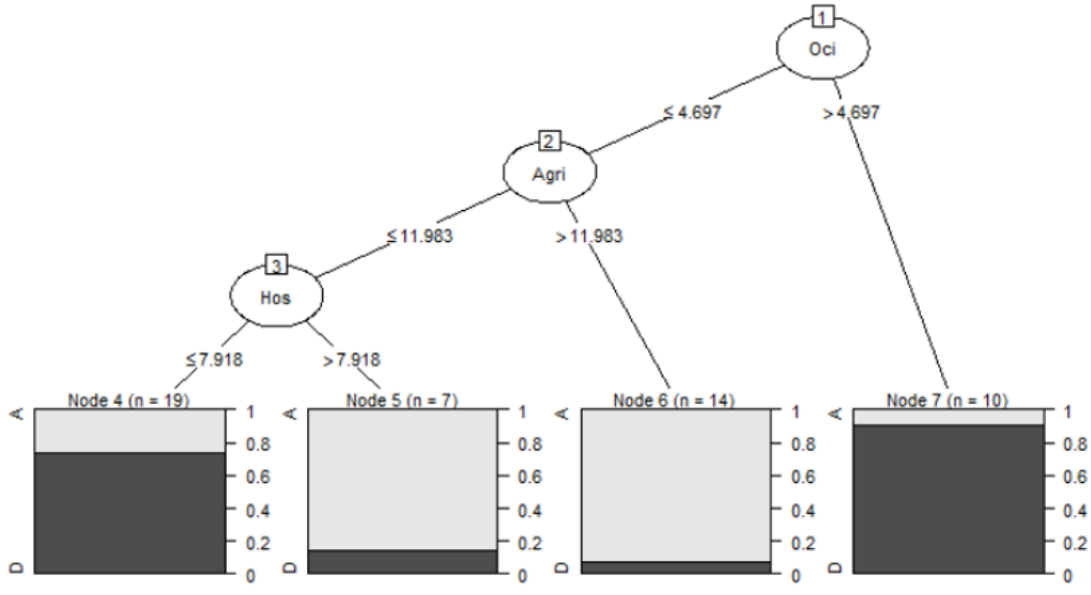
Así, se concluye que la extracción del componente de ERTE en las afiliaciones de los datos de afiliación deteriora tanto la capacidad predictiva como la interpretabilidad económica del árbol, sugiriendo una débil asociación de la intensidad del recurso a los ERTE con la estructura productiva provincial.

Cuadro 5. **ÁRBOL 7: DATOS DE AFILIACIÓN DESCONTANDO LOS ERTE COMO VARIABLE OBJETIVO - VARIACIÓN INTERANUAL DE LA AFILIACIÓN, ABRIL 2020** (A/A, %) **- ESTRUCTURA PRODUCTIVA** (2009-2019) **- PRECISIÓN PREDICTIVA: 72%** (*OUT-OF- SAMPLE* 70%)



Fuente: BBVA Research

Cuadro 6. **ÁRBOL 8: A PARTIR DE LA PROPORCIÓN DE AFILIADOS EN ERTE COMO VARIABLE OBJETIVO - VARIACIÓN INTERANUAL DE LA AFILIACIÓN, ABRIL 2020** (A/A, %) **- ESTRUCTURA PRODUCTIVA** (2009-2019) **- PRECISIÓN PREDICTIVA: 72%** (*OUT-OF- SAMPLE* 70%)



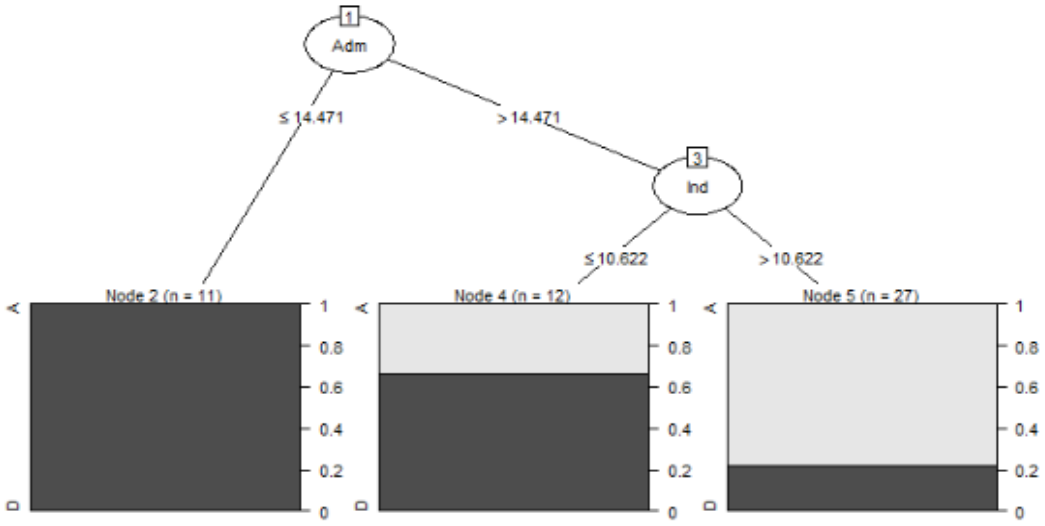
Fuente: BBVA Research

Crisis inmobiliaria y financiera

Para la evaluación del rol de la estructura productiva en el impacto de la crisis de 2008, se emplearon tres variables objetivo alternativas a la comentada en la sección de resultados (Árbol 2). Así, se estimaron tres árboles de decisión adicionales a partir de la variación interanual de la afiliación entre diciembre 2007 y diciembre 2008 (Árbol 9), del crecimiento del promedio del total de afiliados provincial del período abr08-dic08[12] respecto al promedio de los 9 meses precedentes (Árbol 10) y del crecimiento del promedio del total de afiliados provincial del período abr08-abr09 respecto al promedio de los 13 meses precedentes (Árbol 11).

En líneas generales los árboles estimados a partir de estas tres nuevas variables objetivo reflejan que los resultados se mantienen similares a los obtenidos en el Árbol 2 aunque con menor desagregación. La administración pública continúa siendo el principal factor discriminante a la hora de diferenciar la respuesta de las provincias a la crisis de 2008. También el hecho de tener un peso más elevado del sector industrial se asocia a una mayor fortaleza de los mercados laborales de cara al shock analizado. Los umbrales críticos de estos árboles también son similares a los obtenidos en el Árbol 2. Por el contrario, la precisión predictiva obtenida a partir de la estimación del Árbol 2 es claramente superior, especialmente la capacidad predictiva *out-of-sample.*

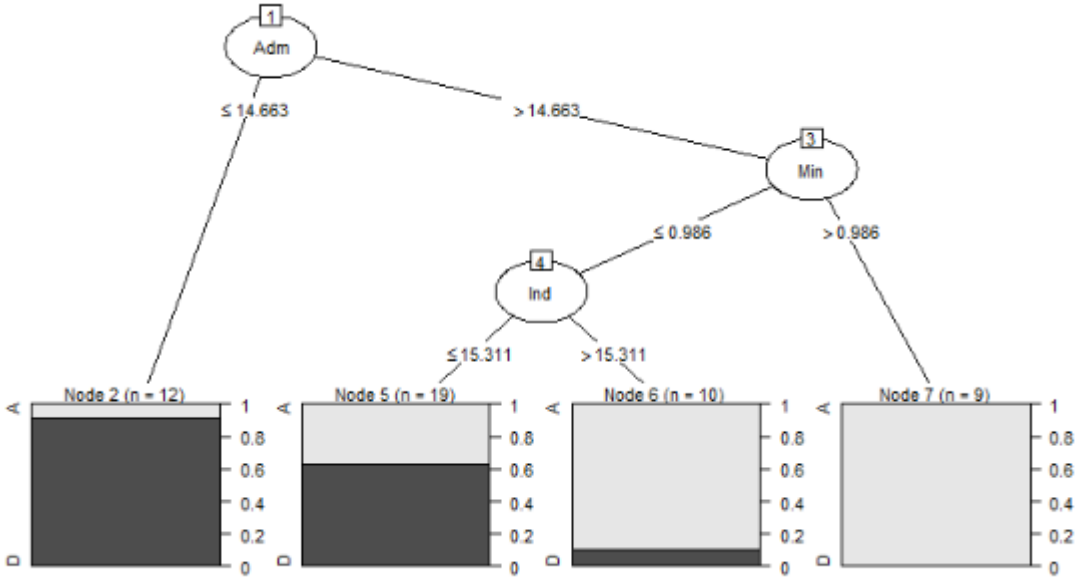
Cuadro 7. **ÁRBOL 9: CRISIS INMOBILIARIA Y FINANCIERA DE 2008 - VARIACIÓN INTERANUAL DE LA AFILIACIÓN, DICIEMBRE 2008** (%) **- ESTRUCTURA PRODUCTIVA** (2006-2007) **- PRECISIÓN PREDICTIVA: 82%** (*OUT-OF- SAMPLE* 76%)



Fuente: BBVA Research

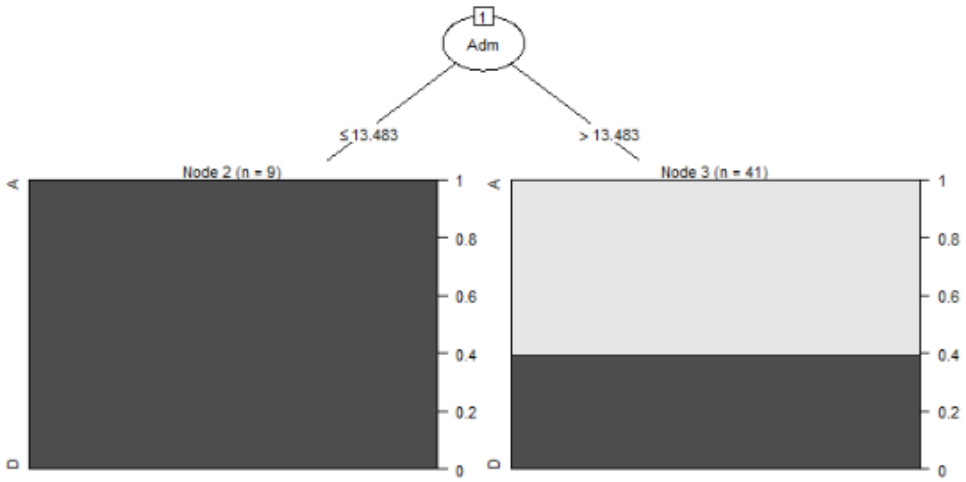
12: Se corresponde con el intervalo temporal donde las tasas intermensuales tienden a ser cada vez menores (el ritmo de caída tiende a acelerarse cada mes).

Cuadro 8. **ÁRBOL 10: CRISIS INMOBILIARIA Y FINANCIERA DE 2008 - VARIACIÓN DEL CRECIMIENTO DEL PROMEDIO DEL TOTAL DE AFILIADOS PROVINCIAL DEL PERÍODO ABR08-DIC08 RESPECTO AL PROMEDIO DE LOS 9 MESES PRECEDENTES - ESTRUCTURA PRODUCTIVA** (2006-2007) **- PRECISIÓN PREDICTIVA: 82%** (*OUT-OF- SAMPLE* 70%)



Fuente: BBVA Research

Cuadro 9. **ÁRBOL 11: CRISIS INMOBILIARIA Y FINANCIERA DE 2008 - VARIACIÓN DEL CRECIMIENTO DEL PROMEDIO DEL TOTAL DE AFILIADOS PROVINCIAL DEL PERÍODO ABR08-DIC08 RESPECTO AL PROMEDIO DE LOS 13 MESES PRECEDENTES - ESTRUCTURA PRODUCTIVA** (2006-2007) **- PRECISIÓN PREDICTIVA: 62%** (*OUT-OF- SAMPLE* 50%)



Fuente: BBVA Research

Apéndice B: Crecimiento y volatilidad en 2014-2019

Este trabajo pretende enfocarse no solo en el análisis de la respuesta provincial de cara a dos shocks muy específicos (crisis sanitaria y crisis de 2008), sino también trata de examinar el comportamiento de los mercados laborales locales a lo largo de periodos de relativa normalidad y en ausencia de *shocks* relevantes de grandes dimensiones. A este respecto, se toman en cuenta dos variables objetivo adicionales (véase el Cuadro 2). Por un lado, el crecimiento promedio de la afiliación[13] durante 2014-2019 con la finalidad de aislar los aspectos de la estructura productiva que fueron motores de la creación de empleo durante la última expansión. Por otro lado, se toma en cuenta la volatilidad de este crecimiento con el objetivo de analizar la estabilidad[14] en la generación de puestos de trabajo de cada provincia y su respuesta a *shocks* menores.

De la misma forma que en el caso de la crisis financiera, la simple sustitución de las nuevas variables dependientes en el Árbol 1, sin cambiar la estructura del árbol ni la composición de los nodos finales, redujo la capacidad predictiva hasta el entorno del 60%. Además, la pureza de los nodos finales se mantuvo solo en el caso de las provincias turísticas y, en menor medida, de las industriales. Lo anterior justifica la estimación de dos nuevos árboles para poder evidenciar las diferencias entre los períodos considerados. En este nuevo ejercicio se aplicará la estructura productiva de 2013[15].

Los resultados de estas nuevas estimaciones evidencian que tanto las provincias turísticas como las grandes áreas urbanas fueron los motores de la creación de empleo en el quinquenio anterior, y además lo fueron de forma estable en el tiempo. Por el contrario, las provincias menos diversificadas y de interior mostraron un menor dinamismo y una mayor volatilidad.

A continuación se desarrollan los resultados principales de los dos ejercicios. **En primer lugar, se analiza la volatilidad en el crecimiento de la afiliación en el último quinquenio.** En este caso, el sector de ocio y entretenimiento sustituye y complementa el papel de la hostelería (véase el Árbol 3). En detalle, las provincias con un peso de este sector superior al 4,4% habrían presentado un crecimiento menos volátil durante la última expansión. **El nodo 7 recoge tanto las típicas provincias turísticas como las grandes áreas urbanas.** Así, estas fueron no solo las que presentaron el mayor dinamismo en la creación del empleo entre 2014 y 2019, sino también lo hicieron de forma estable durante todo el periodo.

Por el contrario, las provincias con la mayor volatilidad habrían sido:

- provincias con un peso del ocio y entretenimiento inferior al umbral crítico y de la agricultura inferior al 15.6%. **Se trata de provincias industriales, de interior y poco pobladas**. El hecho de depender fuertemente de algunos sectores clave y de presentar mercados laborales de tamaño más reducido puede haber provocado que la evolución de la afiliación fuera menos estable y que shocks muy pequeños (locales, empresariales o sectoriales) fueran poco compensados y absorbidos con menor éxito.
- provincias con un peso del ocio y entretenimiento inferior al umbral crítico, con un peso de la agricultura superior al 15.6% y del comercio inferior al 19,8%. Se trata de **provincias muy dependientes del sector agrario y pesquero**. La naturaleza de este sector muy ligada a las temporadas o a los eventos climáticos podría justificar esta mayor volatilidad.

13: Con crecimiento promedio se entiende la media de las tasas interanuales de cada trimestre desde el 1T14 hasta el 4T19 en cada provincia.
14: Se entiende el coeficiente de variación de las tasas interanuales de cada trimestre desde el 1T14 hasta el 4T19 en cada provincia.
15: De la misma forma que en los otros dos árboles, se aplica la estructura productiva anterior al periodo considerado para la variable objetivo. El cambio estructural en esta estructura registrado entre 2011 y 2012 como consecuencia de la crisis de 2008 justifica además restringir el análisis de 2013.

Esta mayor volatilidad del sector agrario podría haberse visto compensada en algunas zonas por la presencia del comercio. Así, **el nodo 6, que recoge las provincias con un peso de la agricultura y del comercio superiores al umbral crítico (y del ocio y entretenimiento inferior)**, presentaron una mayor estabilidad.

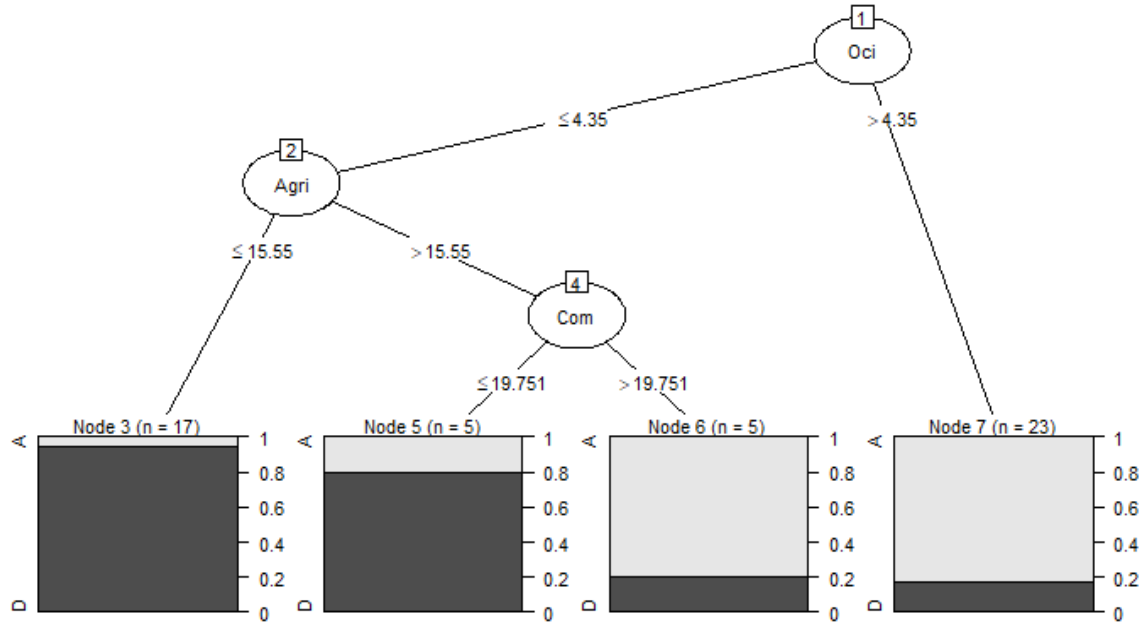
Por otro lado, **el Árbol 4 analiza el crecimiento promedio a lo largo de 2014-2019**. La hostelería resulta ser el factor relevante a la hora de explicar cómo evolucionó la afiliación provincial en el periodo considerado. En este caso, **un peso de la hostelería superior al 9,2% permitió un crecimiento de la afiliación superior a la mediana**. Así, **el nodo 7 recoge las provincias más turísticas**. Cabe destacar además que el umbral crítico de la hostelería es casi igual al encontrado en el árbol 1, lo que da aún más robustez al hecho de que el nodo turístico queda siempre bien identificado y las provincias que lo componen presentan un comportamiento similar entre ellas a lo largo del tiempo.

En segundo lugar, **entre las provincias con un peso de la hostelería inferior al umbral crítico, las que presentan un porcentaje de empleo en el sector público superior al 20,5% registraron una evolución de la afiliación algo inferior a la mediana**. Se puede identificar el **nodo 6, como el grupo de las provincias poco diversificadas y situadas en el centro peninsular**. El sector público en estas zonas ha compensado en algunos casos la poca creación de empleo por la falta de los sectores más dinámicos y/o de población, pero no con la intensidad suficiente como para generar un aumento sostenido de la afiliación.

Finalmente, la hostelería vuelve a aparecer como factor explicativo en la tercera división del árbol. En particular, se identifican dos grupos:

- provincias con un peso de la hostelería inferior al 7,2%, a las que se asocia un comportamiento del mercado laboral más favorable que la mediana. Se trata de un grupo formado en su mayoría por las **provincias con las mayores áreas urbanas del país o de las que están situadas cerca de estas**. La mayor diversificación económica y la mayor presencia de servicios de alto valor añadido podría haber provocado el fuerte crecimiento registrado en el periodo
- provincias con un peso de la hostelería superior al 7,2%, que experimentaron una mayor debilidad del mercado laboral. Se trata de **provincias con un sector turístico reseñable pero no tan masificado** y lejos de los niveles mostrados por las incluidas en el nodo 7. Esto podría haber llevado a un empuje inferior del sector, lo que, junto a la falta de otros sectores dinámicos, se habría traducido en una evolución más débil del mercado laboral.

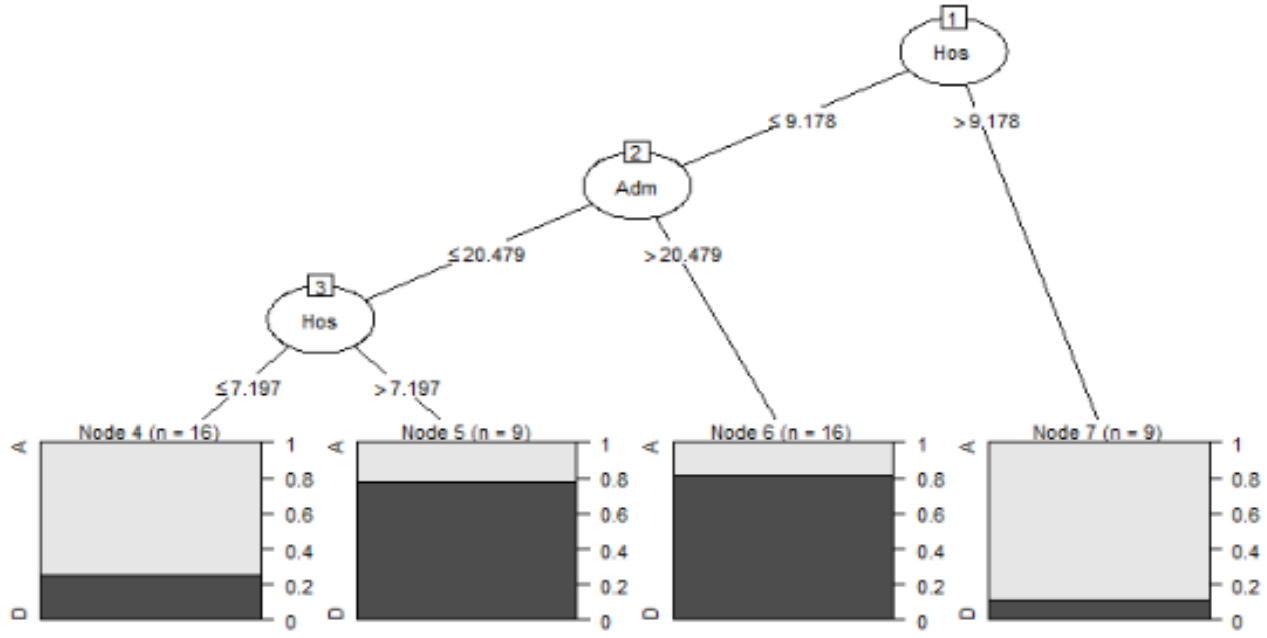
Cuadro 10. **ÁRBOL 3: VOLATILIDAD EN LA FASE EXPANSIVA - COEFICIENTE DE DISPERSIÓN DE LA VARIACIÓN INTERANUAL DE LA AFILIACIÓN, PROMEDIO 2014-2019** (%) **- ESTRUCTURA PRODUCTIVA** (2013) **- PRECISIÓN PREDICTIVA: 86%** (*OUT-OF- SAMPLE* 58%)



Nodo 3		Nodo 5	Nodo 6	Nodo 7	
Álava	Ourense	Huelva	Almería	Alicante	Huesca
Albacete*	Palencia	A Coruña*	Cuenca	Asturias*	La Rioja
Ávila	Pontevedra	Badajoz	Granada	Baleares	Madrid
Burgos	Salamanca	Cáceres	Murcia	Barcelona	Málaga
Ciudad Real	Segovia	Jaén	Sevilla*	Cádiz*	Navarra
León	Soria			Cantabria*	Las Palmas
Lérida	Teruel			Castellón	Tarragona
Lugo	Toledo			Córdoba	Tenerife
	Zamora			Girona	Valencia
				Guadalajara	Valladolid
				Guipúzcoa	Vizcaya*
					Zaragoza

* Provincias que han tenido una evolución de la afiliación distinta a la de la mayoría del mismo nodo.
Fuente: BBVA Research

Cuadro 11. **ÁRBOL 4: CRECIMIENTO PROMEDIO EN LA FASE EXPANSIVA - TASA DE VARIACIÓN INTERANUAL DE LA AFILIACIÓN,** PROMEDIO 2014-2019 **- ESTRUCTURA PRODUCTIVA** (2013) **- PRECISIÓN PREDICTIVA: 84%** (*OUT-OF- SAMPLE* 58%)



Nodo 4		Nodo 5	Nodo 6		Nodo 7
Guadalajara	A Coruña*	Burgos	Álava	Albacete*	Alicante
Huelva	Almería	Asturias	Ávila	Ciudad Real*	Baleares
Madrid	Barcelona	Cantabria	Badajoz	Granada*	Cádiz
Murcia	Cuenca	Castellón*	Cáceres	León	Girona
Navarra	Jaén*	Córdoba	Guipúzcoa	Ourense	Las Palmas
Sevilla	Lérida*	Huesca*	Salamanca	Palencia	Málaga
Toledo	Valencia	La Rioja	Soria	Teruel	S. C. Tenerife
Zaragoza	Vizcaya*	Lugo	Zamora	Valladolid	Segovia*
		Pontevedra			Tarragona

* Provincias que han tenido una evolución de la afiliación distinta a la de la mayoría del mismo nodo.
Fuente: BBVA Research

Apéndice C: Metodología de estimación

La metodología empleada en el estudio es la de **Árboles de Decisión** (*Decision Trees*), una de las más populares en el campo del **Aprendizaje Automático Supervisado** (*Supervised Machine Learning*) en general y en el de los **Datos Masivos** (*Big Data*) en particular.

Se escoge esta metodología porque el problema a tratar puede formularse naturalmente como uno de clasificación con potencial presencia de efectos no-lineales[16] y esta técnica permite abordar este tipo de problemas ofreciendo al mismo tiempo resultados susceptibles de una fácil y clara interpretación económica (en contraste con el resto de técnicas de aprendizaje automático supervisado).

Y es que la interpretabilidad de los árboles de decisión no tiene parangón en el mundo de las metodologías de aprendizaje automático supervisado, lo que obedece, en buena medida, a que replica (si bien en forma muy simplificada) formas típicas del razonamiento humano **consciente**, en contraste con los procesos de aprendizaje humano **inconscientes** que replican técnicas como las redes neuronales (*Neural Networks*).

La metodología exige que el problema a estudiar se formule como un problema de clasificación o de predicción de clase: esto es, decidir a qué clase o grupo pertenece cada individuo en función de ciertas características o atributos individuales. Para lo cual es necesario que la variable a explicar sea discretizada y transformada en una variable categórica.

En este caso, se ha optado por la opción más simple, utilizar los cuartiles de la distribución de la variable objetivo o a explicar, para reexpresarla como una variable categórica que indique simplemente a qué grupo pertenece cada provincia en función de los cuartiles que delimitan el intervalo en el que se encuentra el valor de la variable objetivo para la misma. En nuestros principales ejercicios se escoge de hecho sólo el cuartil central (la mediana), por lo cual la variable categórica resultante es binaria y simplemente indica si una provincia pertenece al grupo de aquellas para la que el valor de la variable objetivo es igual o mayor que la mediana ó al grupo complementario[17].

Por ejemplo en el análisis del impacto inicial del confinamiento domiciliario en respuesta a la epidemia de COVID-19, en el que la variable objetivo se refiere a la tasa de variación del total de afiliados a la Seguridad Social entre abril 2019 y abril 2020, esta se reexpresó en la forma de una variable categórica binaria indicando, para cada provincia, a cuál de estos dos grupos o clases pertenece: el Grupo A, formado por las provincias que experimentaron una caída del empleo menor o igual a la mediana (tasa de variación de las afiliaciones mayor o igual a la mediana) o Grupo D, formado por las provincias que experimentaron una caída del empleo mayor a la mediana (una tasa de variación de las afiliaciones menor a la mediana). [18]

A continuación se describen los rasgos fundamentales del algoritmo utilizado en la estimación de los árboles de decisión analizados en este trabajo. Se trata del **algoritmo C5.0 de Ross Quinlan** (que hemos empleado en el entorno del **lenguaje de programación estadística R**, gracias al paquete-interface C50), que a su vez no es más que una versión mejorada (generalización para incorporar atributos descriptivos continuos e incremento sustancial

16: De entrada, es esperable que haya umbrales críticos en el peso de un sector productivo para que su influencia en la respuesta del empleo total de una provincia a determinados shocks resulte determinante.
17: En ejercicios alternativos, se utiliza la variable categórica cuaternaria resultante de utilizar todos los cuartiles, pero en ningún caso se encontró que las variables explicativas contribuyeran a predecir la pertenencia a los grupos adicionales, por lo cual los resultados de estos ejercicios no resultaron realmente de utilidad y se omiten.
18: La alternativa es considerar los siguientes cuatro grupos o clases: Grupo A, con las provincias que mostraron una tasa de variación de las afiliaciones superior al tercer cuartil (percentil 75%); Grupo B, las que mostraron una tasa de variación mayor o igual a la mediana (segundo cuartil) pero menor que el tercer cuartil; Grupo C, con una tasa de variación menor o igual a la mediana pero mayor primer cuartil (percentil 25%) y Grupo D, el resto (variación menor o igual al primer cuartil).

de la eficiencia computacional) del algoritmo ID3 del mismo autor, un algoritmo de "dicotomización iterativa" (*"iterative dichotomizer"*) basado en la métrica de la entropía.

Recordemos que nuestros ejercicios siempre parten de una categorización de las provincias españolas en dos clases, que llamaremos en lo que sigue A y B (por ejemplo, en el Árbol 1, cada provincia pertenecía a una de dos clases: la clase de las que experimentaron una caída de sus afiliados igual o menor que la mediana, llamada A, o la clase contraria, denominada en aquel caso D).

Por otra parte, dado un grupo de provincias en el que la mayoría son de clase X (de modo que X = A si A≥B y X = B si B≥A) denominaremos **grado de homogeneidad (o pureza)** del grupo a la proporción de provincias de clase X dentro del grupo (llamado en adelante $\boldsymbol{P_x}$, tal que 1.0 > $\boldsymbol{P_x}$ ≥ 0.50) de tal modo que se considerará un grupo perfectamente homogéneo o puro si $\boldsymbol{P_x}$=100% y perfectamente heterogéneo o impuro si $\boldsymbol{P_x}$=50% .

A continuación se introduce la noción central del algoritmo, la **entropía**, una medida inversamente asociada al grado de homogeneidad del grupo y acotada al intervalo cerrado [0,1], de modo que una entropía igual a 0 indica que el grupo es perfectamente homogéneo o puro y una entropía igual a 1 que el grupo es perfectamente heterogéneo o impuro. **La entropía de un grupo se calcula aplicando la siguiente fórmula:**

$$\boldsymbol{E = P_x log_2(P_x) - (1 - P_x) log_2(1 - P_x)}$$

Dadas estas nociones, se puede entonces describir el objetivo último del algoritmo como el de dividir al conjunto total de provincias en dos grupos con la mínima entropía posible (máxima homogeneidad o pureza), tal que en uno predomine A y en el otro B, y que ello se realice sólo en función de las características de la estructura productiva de cada una de ellas: más concretamente, según que el peso provincial de uno o más sectores productivos exceda o no un cierto umbral (distinto por sector pero igual para todos las provincias) estimado por el mismo algoritmo.

Para lograr este objetivo el algoritmo en esencia sigue el procedimiento secuencial cuyas primeras etapas se esbozan a continuación:

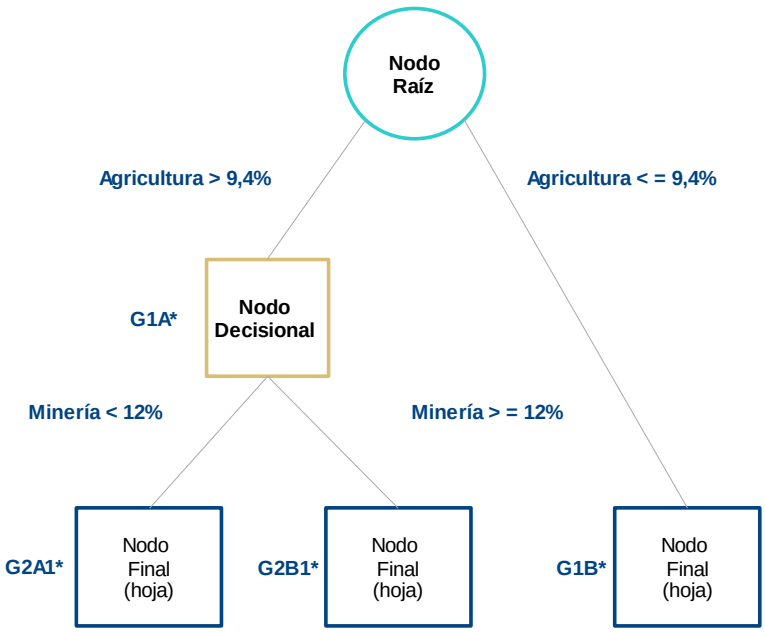
- Nodo raíz, dividir G0: en este punto inicial se parte del grupo formado por todas las provincias (G0) y su entropía asociada (E0) y se procede del siguiente modo:
  - Se escoge un sector productivo y se busca el umbral óptimo de su peso: aquel que permita dividir al conjunto total de provincias en los dos grupos G1A y G1B, cada uno dominada por una clase distinta, tal que el valor mínimo de sus entropías (E1 = min {E1A, E1B}) sea menor que el obtenido con cualquier otro umbral.
  - Se repite el procedimiento anterior para cada uno de los sectores productivos, y finalmente se selecciona el sector (y umbral óptimo asociado) que logró el valor mínimo de la entropía (E1*)
  - La división de G0 así obtenida se da por válida sólo si se obtiene una "ganancia de información", esto es, si E1* < E0, de otro modo el nodo superior o raíz se convertiría en el nodo final del árbol (sería imposible dividir G0 en grupos más homogéneos en función de la estructura productiva provincial)
- Nodo decisional, dividir G1A* (G1B*): en este punto se parte del grupo de provincias G1A* (G1B*), subgrupo de G0, con entropía asociada E1A* (E1B*) y se procede del siguiente modo:

- Se escoge un sector productivo, y se busca el umbral óptimo de su peso: aquel que permita dividir al grupo G1A* (G1B*) en los dos grupos G2A1 y G2B1 (G2A2, G2B2), cada uno dominada por una clase distinta, tal que el valor mínimo de sus entropías, E21 = min {E2A1, E2B1} (E22 = min {E2A2, E2B2}), sea menor que el obtenido con cualquier otro umbral.

- Se repite el procedimiento anterior para cada uno de los sectores productivos, y finalmente se selecciona el sector (y umbral óptimo asociado) que logró el valor mínimo de la entropía E21* (E22*).

- La división de G1A* (G1B*) así obtenida se da por válida sólo si se obtiene una "ganancia de información", esto es, si E21* < E1A* (E22* < E1B*), **y si además el número de provincias de los subgrupos resultantes de la división excede el valor mínimo seleccionado por el analista**, de otro modo el nodo inmediatamente superior se convertiría en el nodo final de esa rama del árbol, es decir, sería imposible dividir G1A* (G1B*) en grupos más homogéneos en función de la estructura productiva provincial (y cuyo tamaño excede el número mínimo de provincias impuesto por el analista).

La figura siguiente ilustra un resultado hipotético de los pasos anteriores: sólo hay dos divisiones que aportan reducción de la entropía (ganancia de información): (1) la del grupo original, G0, que se divide en los grupos G1A* y G1B* en función del sector agrícola (umbral óptimo de 9,4%), y (2) la del subgrupo G1A*, que puede volverse a dividir con alguna ganancia de entropía (en función del sector minería y un umbral óptimo de 12%). Por tanto, el nodo de G1B* se convierte en nodo final u hoja, al igual que los dos nodos que surgen de la división de G1A*.



Fuente: BBVA Research

AVISO LEGAL

El presente documento, elaborado por el Departamento de BBVA Research, tiene carácter divulgativo y contiene datos, opiniones o estimaciones referidas a la fecha del mismo, de elaboración propia o procedentes o basadas en fuentes que consideramos fiables, sin que hayan sido objeto de verificación independiente por BBVA. BBVA, por tanto, no ofrece garantía, expresa o implícita, en cuanto a su precisión, integridad o corrección.

Las estimaciones que este documento puede contener han sido realizadas conforme a metodologías generalmente aceptadas y deben tomarse como tales, es decir, como previsiones o proyecciones. La evolución histórica de las variables económicas (positiva o negativa) no garantiza una evolución equivalente en el futuro.

El contenido de este documento está sujeto a cambios sin previo aviso en función, por ejemplo, del contexto económico o las fluctuaciones del mercado. BBVA no asume compromiso alguno de actualizar dicho contenido o comunicar esos cambios.

BBVA no asume responsabilidad alguna por cualquier pérdida, directa o indirecta, que pudiera resultar del uso de este documento o de su contenido.

Ni el presente documento, ni su contenido, constituyen una oferta, invitación o solicitud para adquirir, desinvertir u obtener interés alguno en activos o instrumentos financieros, ni pueden servir de base para ningún contrato, compromiso o decisión de ningún tipo.

Especialmente en lo que se refiere a la inversión en activos financieros que pudieran estar relacionados con las variables económicas que este documento puede desarrollar, los lectores deben ser conscientes de que en ningún caso deben tomar este documento como base para tomar sus decisiones de inversión y que las personas o entidades que potencialmente les puedan ofrecer productos de inversión serán las obligadas legalmente a proporcionarles toda la información que necesiten para esta toma de decisión.

INTERESADOS DIRIGIRSE A:
BBVA Research: Calle Azul, 4. Edificio La Vela – 4ª y 5ª planta. 28050 Madrid (España).
Tel.:+34 91 374 60 00 y +34 91 537 70 00 / Fax:+34 91 374 30 25
bbvaresearch@bbva.com www.bbvaresearch.com