UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

MÁSTER EN ESTADÍSTICAS OFICIALES E INDICADORES SOCIALES Y ECONÓMICOS

TRABAJO FIN DE MÁSTER

2021

Imputación de datos mediante Random Forest

Autor:
Lasai Alai BARREÑADA TALEB

Tutores:
David SALGADO
Elena ROSA
Rosa ALONSO

18 de octubre de 2021

Resumen/ Abstract

Facultad de Ciencias Matematicas

Máster Oficial de Estadísticas Oficiales

Imputación de datos mediante Random Forest

by Lasai Alai BARREÑADA TALEB

La información disponible es cada vez mayor y los institutos de estadística oficiales deben hacer uso de esta información para crear procesos innovadores y eficaces. El *statistical lear ning* es el conjunto de técnicas usadas para la mejor comprensión de los datos. Los *random forests*, basados en un *ensemble* de arboles de decisión, son una de las técnicas mas utilizadas de aprendizaje supervisado. En este trabajo se han usado *random forests* para la imputación de datos en encuestas económicas coyunturales y mas concretamente en los Índices de Cifras de Negocios de la Industria. La imputación se trata del proceso mediante el cual se asigna un valor a un ítem para el que previamente no se tenia información. En este estudio se elabora la metodología para la imputación después de analizar los criterios de calidad necesarios para la producción de una estadística oficial. En primer lugar se realiza la selección de variables o *feature selection* más interesante para el cálculo de las cifras de negocios. Posteriormente, se aborda el proceso de selección de parámetros para la obtención del modelo óptimo de bos ques aleatorios para el conjunto de datos seleccionado. Finalmente se realiza una aplicación práctica del bosque aleatorio para las imputaciones y se evalúan obteniendo un resultado satisfactorio.

The amount of available information in National Statistical Institutes is increasing ra pidly and they shall make use of it to develop innovative and effective processes. Statistical learning is the set of techniques used for better understanding of data. Random Forests, ba sed on decision tree ensembles, are one of the most used techniques of supervised learning. In this thesis Random Forest have been used to impute data in short term business statistics. Imputation is defined as the method to give value to an item that previously was missing.In this study a new methodology is developed after analysing the quality requirements for of ficial statistics. Firstly, the feature selection is carried out in order to get the set of variables that will be included in the model. After this, the tuning of the forests is carried out to get the optimum forest. Finally, this model is used to impute the missing values and the assessment of the accuracy of the estimation is carried out having satisfactory results.

Índice general

Índice de figuras

[1]Todas las figuras en las que no se menciona la fuente son de elaboración propia con datos del INE.

Índice de cuadros

Capítulo 1

Introducción

1.1. Statistical learning

Llamamos *statistical learning* o aprendizaje estadístico a un conjunto de herra mientas utilizadas para la mejor comprensión de los datos. Este aprendizaje puede ser supervisado o no supervisado. El aprendizaje supervisado consiste en la crea ción de modelos para predecir o estimar un *output* en función del input provisto mientras que en el aprendizaje no supervisado no existe *output* con lo que el objetivo es entender las relaciones entre los datos (James y col., 2013). Dentro del *statistical learning* encontramos diferentes técnicas como los algoritmos de regresión logística, arboles de decisión o clasificación, *k* vecinos más cercanos o bosques aleatorios. Es esta última técnica de aprendizaje supervisado, los bosques aleatorios o *random fo rests* (RF), la que será utilizada en este trabajo. Este algoritmo introducido en 2001 por el estadístico estadounidense Leo Breiman consiste en la combinación de mu chos árboles de clasificación o regresión para mejorar los resultados que arrojaría un único árbol. Se explicará con mayor detalle el funcionamiento y modalidades de los bosques aleatorios en el Capítulo 4.

1.2. Estadísticas Oficiales

Estas técnicas, aunque no son muy recientes, todavía no se han consolidado en las metodologías de producción de estadísticas oficiales [1]. La estadística oficial cum ple un rol esencial en la sociedad ya que es la base para la toma de decisiones de entidades políticas, económicas o sociales y por ende tiene unos estándares de ca lidad muy altos donde no es sencillo introducir técnicas innovadoras debido a la gran carga de trabajo de los profesionales dedicados a la producción que no cuentan con el tiempo suficiente para continuar produciendo los

productos estadísticos y a su vez invertir tiempo en innovación. Sin embargo, la expansión del concepto de calidad, centrado anteriormente en el sesgo y la varianza, hacia otras dimensiones como la puntualidad o el *timeliness* [2] hace difícil a los institutos nacionales de esta dística competir con productores de estadísticas privados que gracias a estas y otras técnicas innovadoras son capaces de, por ejemplo, calcular el tráfico en tiempo real

[1]Las estadísticas oficiales son aquellas publicadas por el sistema nacional de estadística como bien público.

[2]Tiempo desde el evento de referencia hasta la publicación de la estadística. Traducido como opor tunidad

como nos ofrece Google. Además de cumplir con estos altos estándares de calidad los institutos de estadística oficial deben ser capaces de cuantificar la precisión de sus datos (MacFeely, 2016) para proveer a los usuarios con indicadores de calidad. El uso del RF permitiría al Instituto Nacional de Estadística (INE) tanto optimizar los procesos de imputación como calcular el error de imputación y la precisión de las imputaciones.

En los últimos años el *statistical learning* y el big data se están empezando a usar en la producción de estadísticas oficiales. En el año 2018 la oficina de estadística ale mana (Destatis) llevó a cabo un profundo análisis del uso del *machine learning* en la estadística oficial. La mayoría de los países indicó el uso del *machine learning*, in cluido el INE en España. En ese momento el INE contaba con tres proyectos donde se usaba *machine learning*, muy por debajo de Statistics Canada que es la institu ción donde mayor penetración tiene el *machine learning* con 36 proyectos. De estos tres proyectos del INE, dos se centran en el uso del aprendizaje automático para la selección de unidades influyentes para la depuración selectiva. La mayoría de los proyectos donde se usa el aprendizaje automático son experimentales y cabe desta car que la segunda aplicación más común es la imputación de datos y el método más usado dentro del *machine learning* son los RF (Beck y col., 2018).

Con el fin de crear una guía para la producción de estadísticas oficiales de ca lidad Eurostat creó el Código de Buenas Prácticas de la Unión Europea (ECoP)[3]en el año 2004 y lo actualizó en 2011 y 2017 (Eurostat, 2017). Dentro de los principios del ECoP este trabajo busca fortalecer la rentabilidad y la solidez metodológica de los procesos estadísticos así como mejorar las dimensiones 13 y 14 asociadas a la precisión y fiabilidad y puntualidad y *timeliness*.

Además la estadística debe hacer hincapié en el *timeliness* a la hora de tener en cuenta al usuario moderno final. Estos usuarios están acostumbrados, gracias a la inmediatez de internet y las redes sociales, a que el periodo entre un fenómeno y te ner información acerca del mismo sea mínimo y por lo tanto cuando los institutos de estadística publican información con *timeliness* de 1 año (estadísticas estructurales) hay algunas dimensiones de calidad que se ven resentidas como puede ser la rele vancia. En el caso de las estadísticas económicas coyunturales [4]este periodo es por definición inferior, pero mantenerlo mínimo, siempre sin perder en otras dimensio nes de calidad, debe ser un objetivo de cualquier entidad productora de estadísticas oficiales. Por otro lado, el *statistical learning* al ser una técnica que usa la inteligencia artificial permitirá un uso de los recursos disponibles en las unidades productoras mucho más eficiente con lo que los profesionales podrán invertir este tiempo en cual quier otro proceso necesario y así seguir el

principio 10 de rentabilidad del ECoP.

Este trabajo se ha realizado en colaboración con el INE, donde trabajo actualmen te en el Gabinete de la Presidencia, gracias a una beca de postgrado. Gracias a estar trabajando dentro del instituto y después de firmar el acuerdo de confidencialidad tengo acceso a los microdatos de la encuesta de Índices de Cifra de Negocios en la

[3]Código de 15 principios elaborado por la Comisión Europea con el objetivo de establecer pautas para preservar la calidad de las estadísticas europeas.

[4]Estadísticas económicas que buscan reflejar las tendencias más cercanas en el tiempo con periodi cidad de recogida y difusión inferior al trimestre.

Industria (ICN) que serán los utilizados para la construcción del modelo y la reali zación de las imputaciones. Los resultados, sin embargo, siempre serán agregados y verificando el secreto estadístico, cumpliendo así con el principio de confidenciali dad y las leyes vigentes.

El INE es un organismo autónomo público dependiente de la secretaria de estado de Economía y Apoyo a la empresa. El INE se encarga de la producción y difusión de la mayoría de las estadísticas oficiales españolas y de la coordinación con Eurostat. Es la institución más importante de producción estadística oficial del Estado. El INE cuenta con un presupuesto de 189,78 millones de euros (BOE, 2020) en el año 2021 de los cuales más del 60 % son destinados a gastos de personal y un 23 % a inversiones en mejoras de material y sistemas informáticos.

Dentro de las tareas del Instituto se encuentran, entre otras, las siguientes (BOE, 2001):

- Coordinar las unidades estadísticas en el gobierno central.
- Escribir el borrador del Plan Estadístico Nacional (PEN).
- Crear y proponer estándares dentro del proceso estadístico.
- Mejorar e investigar sobre metodología estadística.
- Aplicar y evaluar la aplicación de la confidencialidad estadística.
- Crear y mantener directorios de empresas, edificios etc.
- Llevar a cabo los censos de población decenales.
- Crear y proponer estándares dentro del proceso estadístico.
- Crear estadísticas e indicadores sociales y económicos.
- Crear el Inventario de Operaciones Estadísticas (IOE).
- Proponer nuevas regulaciones en relación con la estadística.
- Difundir de manera clara y accesible las estadísticas producidas.
- Coordinar las relaciones internacionales con otros institutos de estadística.

El INE forma parte del Sistema Estadístico Europeo (ESS del inglés *European Sta tistical System*) que nace de la relación de Eurostat, todos los institutos de

estadística oficiales de los diferentes países de la Unión Europea y el espacio EFTA así como di ferentes autoridades nacionales (ONAs) como pueden ser las unidades estadísticas de los ministerios. El objetivo de esta unión es asegurar que las estadísticas que se producen a nivel europeo se produzcan usando las mismas definiciones, estándares y clasificaciones para asegurar la comparabilidad de las estadísticas. Además, la ESS coordina su trabajo con otros organismos internacionales como la OCDE, la red de bancos centrales europeos (SEBC), la Organización de las Naciones Unidas (NNUU) o el Fondo Monetario Internacional (FMI).

1.3. Software

Todo el tratamiento de los datos y obtención de los resultados se ha realizado utilizando el lenguaje de programación R. El lenguaje de programación R es un len guaje orientado a objetos, diseñado para la estadística computacional y la visualiza ción de datos (Team R, 2000). El lenguaje está diseñado de tal forma que los propios usuarios son los que crean paquetes para cada proceso concreto y estos se alojan en *"The Comprehensive R Archive Network "* (CRAN) o cualquier otra plataforma co mo GitHub [5] de donde cualquiera puede descargarlos y usarlos en su código. Esto permite tener herramientas muy actualizadas e innovadoras con el beneficio de ser gratuitas y además con la calidad que exigen los millones de usuarios que comprue ban la precisión de las técnicas cuando se publican. Cabe destacar que la mayoría de los paquetes más utilizados suelen ir acompañados de artículos científicos pu blicados en revistas de renombre que explican el funcionamiento y el resultado del paquete.

Los paquetes principales usados en este trabajo son: para la gestión de datos **da ta.table** (Dowle y col., 2019) , para la creación de bosques aleatorios **ranger**[6](Wright y Ziegler, 2017) y para la visualización de datos **ggplot** (Wickham, 2007). Además de estos paquetes se han utilizado otros creados específicamente por el departamento de metodología del INE como *FastReadFtw*,*StQ* y *fastReadfwf*. A pesar de no ser este lenguaje el oficial del INE se ha escogido debido a su carácter de código abierto y la cantidad de paquetes creados por otros estadísticos oficiales que lo convierte en ideal para la creación y divulgación científica . Por otro lado, es uno de los lenguajes más usados en el desarrollo de metodologías de producción de estadísticas oficiales en el ESS con iniciativas como la plataforma *"Awesome official Statistics"* impulsada por el instituto nacional de estadística neerlandés (CBS) y presentada tanto en la Comisión Económica de las Naciones Unidas para Europa (UNECE) como en diferentes confe rencias internacionales sobre estadísticas oficiales. Esta iniciativa busca fomentar la colaboración internacional mediante la creación de una lista de todos los softwares de código abierto y acceso libre desarrollados o usados por las diferentes organiza ciones de estadística oficial y además están agrupadas en función de las fases del Generic Statistical Business Process Model[7](GSBPM) correspondiente en el que son usados. En esta lista se observa que R es el lenguaje preferido de la mayoría de esta dísticos ya que permite la creación de paquetes que son descargados e instalados de manera sencilla por los usuarios que los deseen lo que genera una portabilidad muy interesante para la reusabilidad de la herramienta. Dentro de los diferentes paquetes destinados a la creación de *random forest* se ha escogido ranger debido a su rapidez para la creación de árboles de regresión donde supera con creces a todos los demás teniendo resultados de predicción similares (Wright

y Ziegler, 2017).

[5]Plataforma para alojar proyectos utilizando el sistema de *Git* de control de versiones. [6]RANdom forest GEneRator.

[7]Estándar para describir el proceso de producción estadístico desarrollado por la UNECE y adop tado por la mayoría de los países.

1.4. Objetivos y estructura

El objetivo de este trabajo de fin de máster es desarrollar una nueva metodología de imputación de datos que cumpla con todos los requisitos de calidad necesarios para la estadística oficial. El objetivo principal del estudio es crear una metodología utilizando RF para la imputación de valores faltantes en encuestas económicas co yunturales. El trabajo se basará en la encuesta de ICN pero el objetivo es desarrollar una técnica modular que pueda ser usada en otras encuestas económicas coyuntura les como el Índice de Precios Industriales (IPRI) o Índices de Producción Industrial (IPI). A la hora de imputar los datos mediante bosques aleatorios se considerará la precisión de esta técnica teniendo en cuenta las revisiones en periodos posteriores donde el dato faltante fue imputado y después enviado el real por la unidad infor mante así como la desviación con respecto a la imputación llevada a cabo actualmen te por el instituto. Por otro lado, los bosques aleatorios permiten obtener indicadores de precisión basándose en las observaciones *out of bag* que se explicaran más tarde en el Capítulo 4.

Los objetivos principales del Trabajo de Fin de Máster son:

Introducción teórica a los diferentes métodos de imputación haciendo hincapié en los usados en la actualidad en la red Eurostat.

Revisión de los estándares de calidad del sistema estadístico europeo.

Introducción teórica a la encuesta de Índices de Cifras de Negocios en la In dustria del INE.

Estudio teórico de los diferentes modelos de bosques aleatorios y su aplicación en la estadística oficial.

Preparación de los ficheros a imputar de la encuesta.

Aplicación de *random forest* a los periodos escogidos para la imputación de valores faltantes.

Análisis de la calidad de las imputaciones y los resultados obtenidos.

Para abordar estos objetivos el trabajo se estructura de la siguiente manera: el Capítulo 2 se centra en el marco teórico de la imputación en estadísticas oficiales. En este capítulo se plantea el porqué de la imputación así como diferentes métodos con sus ventajas y desventajas usados en la estadística oficial para después explicar las diferentes dimensiones de calidad necesarias en estas estadísticas. El Capítulo 3 se centra en el análisis y breve explicación de todo el proceso de la encuesta de ICN y las características del conjunto de datos disponible haciendo hincapié en las variables que se usarán en el siguiente capítulo para construir los bosques aleatorios. En este capítulo cabe destacar la selección de regresores para la construcción del modelo así como la creación de variables derivadas a partir de las existentes. El Capítulo 4 comienza por una revisión y explicación de los bosques aleatorios así como ejemplos de uso en

estadísticas oficiales para después acabar explicando cómo

ha sido la metodología usada en este trabajo. Por último, el Capítulo 5 presentara los pasos seguidos para obtener los resultados y analizará la calidad de estos para que en el Capítulo 6 se añadan las conclusiones y los posibles trabajos futuros.

Todo el código utilizado para la elaboración de los modelos y las figuras que aparecen en este trabajo está disponible en repositorio GitHub en el siguiente en lace y en el anexo A.6 se puede observar parte del código para la obtención de los resultados del envío 1.

Capítulo 2

Marco Teórico

En este capítulo se abordará el problema de la falta de respuesta en los procesos estadísticos y el concepto de calidad en la estadística oficial. Para esto se revisarán los conceptos teóricos de la falta de respuesta y se presentarán diferentes metodo logías para la imputación de datos. El último apartado del capítulo se centra en la explicación de las diferentes dimensiones de la calidad de las estadísticas haciendo hincapié en aquellas afectadas por los valores *missing* y la imputación.

2.1. Falta de respuesta: Imputación

La recogida de datos es la fase anterior al procesamiento de los datos según el GSBPM. La recogida es *el proceso sistemático de obtención de datos para las estadísticas oficiales* (SDMX [1], 2009) y se puede realizar utilizando diferentes métodos. En pri mer lugar, dependiendo de si la información se recoge *ad hoc* para la encuesta o se reutiliza información ya existente se dividen las fuentes en encuestas o registros administrativos. Los **registros administrativos** son aquellas bases de datos de la ad ministración pública (datos ya recogidos) a las que los productores de estadística tienen, o deberían tener, acceso. El uso de esta información se debe maximizar para reducir la carga de trabajo a las personas encuestadas (Principio 9 ECoP). A veces el registro administrativo no se recoge con fines estadísticos por lo que los institutos nacionales de estadística construyen un registro estadístico a partir del administra tivo. Otro método de recogida de datos son las **encuestas** que son documentos con una serie de ítems o preguntas que se construyen para dar respuesta al fenómeno de estudio. Las encuestas se pueden realizar de diferentes formas haciendo más o menos participe al entrevistador, siendo directas o indirectas, utilizando medios di gitales o analógicos, etc... Es usual que se combinen los registros administrativos con las encuestas como ocurre en ICN donde se utiliza el Directorio Central de Empresas

(DIRCE)[2] para la construcción del marco muestral y después sobre esta población se realiza la encuesta de cifras de negocios que se observa en el anexo A.1. Cuando el método de recogida de información son encuestas, como es el caso de ICN, es co mún que las unidades no respondan al cuestionario por diversas razones como no comprender la pregunta, olvidar responderla o simplemente no querer hacerlo. Un

[1]Statistical Data and Metadata eXchange (SDMX) es una iniciativa que busca estandarizar y moder nizar el intercambio de datos y metadatos

[2]Sistema de información único donde están todas las empresas españolas.

Tipología	Definición
MCAR	La probabilidad de que el valor sea *missing* no depende del valor real dela respuesta o el valor de las variables auxiliares. Ejemplo de esto sería cuando el encuestado se olvida de responder un ítem o cuando se pier de información en el procesamiento. Éste es el caso más favorable para la imputación ya que el subgrupo de falta de respuesta tendrá las mismas características que aquellos que han respondido. Sin embargo, es una si tuación poco común en la realidad.
MAR	En este caso la probabilidad de no responder depende del valor de las variables auxiliares pero no de la variable de estudio. Este caso ocurre cuando diferentes subgrupos de población comparten patrones a la hora de no querer responder a la cuestión. El objetivo será identificar estos grupos para así convertir los valores *missing* MAR en MCAR para cada grupo. La mayoría de las técnicas de imputación asumen que la falta de respuesta es de este tipo.
NMAR	En el último caso la falta de respuesta depende tanto de las variables au xiliares como de la variable a imputar. Esto ocurre por ejemplo en la de moscopia donde el voto dependiendo del partido al que vaya dirigido es más o menos probable de ser respondido. Este es el caso menos favorable pero en este estudio no será necesario profundizar ya que se asume que las empresas tienen obligación de proporcionar datos correctos al INE y por lo tanto ninguno de los valores *missing* debería entrar en esta catego ría.

CUADRO 2.1: Tipologías de falta de respuesta

valor *missing* es un marcador de posición para un dato cuyo tipo es conocido pero su valor no lo es. La clasificación más común para la falta de respuesta contempla tres tipologías: **MCAR** (*Completely missing at random*), **MAR** (*Missing at random*) y **NMAR** (*Not missing at random*) (Acock, 2005; Bennett, 2001; Donders y col., 2006). Algunos autores también sugieren incluir la categoría NI (*Non ignorable*) que corresponde a aquellos valores ausentes que no son ni MAR ni MCAR, sin embargo su estudio no se abordará en este trabajo por la complejidad que conlleva pero está ampliamente explicado en Muthen & Muthen (2004). Una mayor explicación sobre estos concep tos se encuentra en el cuadro 2.1.

Esta falta de respuesta puede ser únicamente en uno de los ítems de la encues ta lo que causaría falta de respuesta parcial. Por ejemplo, un problema común en encuestas socioeconómicas es que las unidades responden a muchas preguntas ex ceptuando su renta debido a que creen que esta información es demasiado privada para compartirla. El otro caso es el de falta de respuesta total o de unidad que ocurre cuando una unidad que está dentro de la muestra no da ninguna información para el periodo correspondiente. Es el experto en la materia el que debe determinar si la cantidad de ítems sin respuesta es suficiente para considerar

si la falta de respues ta es total o parcial (De Waal y col., 2007). En el caso de ICN la falta de respuesta siempre es total ya que la empresa debe enviar el cuestionario completo.

La imputación consiste en asignar un valor a un ítem o un grupo de ítems que previamente no tenía valor o ese valor se consideraba erróneo o no ajustado a la realidad. La imputación es por lo tanto un proceso por el que se generan valores artificiales y por lo tanto introduce un error de imputación. Sin embargo, este error

cuenta con la ventaja de ser medible ya que el especialista puede analizar la preci sión de las imputaciones y de esta forma estimar el error de imputación. El objetivo siempre será la generación de valores que se asemejen lo máximo posible al valor real, pero al ser esto difícil se buscará que tengan sesgo mínimo en primer lugar y varianza mínima en segundo lugar. La imputación es preferible en encuestas donde la distribución de la población está altamente sesgada como es el caso de las en cuestas económicas (Särndal y Lundström, 2005) donde una pequeña cantidad de empresas muy grandes generan la mayor parte de la actividad y la gran mayoría son empresas pequeñas con apenas impacto. Las buenas prácticas indican que para tener buenas imputaciones es necesario revisar las técnicas de imputación y mejorar las continuamente y es por eso por lo que aplicar el aprendizaje automático para la realización de las imputaciones puede ser una solución apropiada ya que la propia técnica se revisa y mejora continuamente. Sin embargo, esto no significa que el ex perto encargado de la operación no deba depurar las imputaciones para asegurarse de que no ocurren errores en el algoritmo.

Cabe destacar que la imputación no es una técnica usada exclusivamente para los ítems donde la respuesta esta *missing*. Si en el proceso de validación y depuración uno de los datos se presupone erróneo por no cumplir con los *edits* [3]este dato se de be corregir y esta corrección no es otra cosa que convertir este ítem en valor *missing* y después imputar un valor nuevo, teniendo o no en cuenta la respuesta considerada errónea. Sin embargo, este trabajo se centrará en la imputación para aquellas unida des que no hayan respondido al cuestionario o que el dato que hayan proporcionado (presumiblemente erróneo) no se pueda utilizar para obtener el valor real.

2.2. Tipos de imputación

La metodología para las estadísticas económicas modernas (**MEMOBUST**) (Schol tus y col., 2014) distingue tres grandes enfoques para la imputación. El primero con siste en la **imputación deductiva o lógica**, que consiste en usar reglas de derivación con la información disponible para estimar el valor faltante. El segundo consiste en usar reglas de predicción estadísticas para obtener **modelos** donde calcular impu taciones. Ejemplo de esto son la imputación por regresión, razón o media que más tarde se explicarán y también los random forest que se usarán en este trabajo. El ter cer grupo consiste en **utilizar unidades similares** (*Donor imputation*) para imputar con este valores a las unidades donde no hay respuesta. Dentro de este grupo se en cuentran las técnicas *hot deck* y de vecino más cercano. Además de estos tres existe también en enfoque más manual de la imputación que consiste en la imputación por el experto en la materia.

[3]Se denomina *edit* a el conjunto de requisitos que deben cumplir las respuestas para ser validas.

2.2.1. Imputación deductiva

Es el método que tiene preferencia sobre todos los demás (Scholtus y col., 2014) pero en muchos de los casos no puede ser usado. Este método es especialmente in teresante cuando tenemos falta de respuesta parcial ya que utilizando el valor de otros ítems se puede deducir el valor faltante. Por ejemplo, teniendo la cifra de ne gocios de España y del extranjero y no teniendo el total podríamos imputar el valor total como la suma de ambos.

No obstante, este método no se estudiará con profundidad debido a que este tipo de correcciones se llevan a cabo previamente durante la recogida de datos por la Delegación Provincial correspondiente y no son objeto de estudio. Para finalizar cabe destacar el paquete *deducorrect* (Van Der Loo y col., 2011) de R para la realización de estas imputaciones.

2.2.2. Imputación basada en modelos

Como su nombre indica este enfoque de imputación consiste en encontrar el mo delo predictivo adecuado para la obtención de la imputación. El modelo toma por lo tanto la información disponible y puede ser más o menos complejo. Dentro de estas técnicas se explicarán a continuación la imputación por media, razón y regresión.

Imputación por media

Como su nombre indica cada valor faltante es reemplazado por la media de todos los valores disponibles en su versión más simple. Este modelo tiene el problema de no representar la distribución real del fenómeno ya que existirán muchos casos del valor de la media que no se ajustan a la realidad (Särndal y Lundström, 2005). Para reducir esto se pueden realizar imputaciones por media para grupos concretos lo que reduciría este problema. La fórmula para calcular los valores imputados es la siguiente:

$$\tilde{y}_i = \frac{\sum_{k \in obs} y_k}{n_{obs}}.$$

Siendo $\tilde{y}_i$ los valores imputados y_k el valor de la variable de estudio en la obser vación k y n el número total de observaciones.

Este método tiene la ventaja de ser muy sencillo y no necesitar información au xiliar pero solamente arrojaría resultados satisfactorios a la hora de calcular medias y totales poblacionales pero en ningún caso obtendríamos microdatos ajustados a la realidad por lo que no es un método interesante para el problema planteado en este estudio. Por otro lado, la existencia de *outliers* afecta de manera muy negativa a esta técnica ya que las imputaciones se alejarán más todavía de la distribución real (Scholtus y col., 2014).

Imputación por razón

La imputación por razón tiene en cuenta una sola variable auxiliar y asume que

esta variable es proporcional a la variable de estudio. El proceso de estimación con siste en calcular la razón de la variable auxiliar y la variable de estudio sobre el con junto de datos sin valores *missing*. Una vez calculada esta razón se multiplica por el valor de la variable auxiliar de los valores *missing* y así se obtiene la imputación. La fórmula que explica este proceso es la siguiente:

$$\tilde{y_i} = R\hat{x_i} = \frac{\sum_{k \in obs} y_k}{\sum_{k \in obs} x_k x_i}.$$

Siendo $\tilde{y_i}$ el valor imputado, y_k el valor de la variable que se desea imputar de la observación k, x_k el valor de la variable auxiliar de la observación k y x_i es el valor de la variable auxiliar en la observación i donde no hay respuesta para y_i.

Esta estimación será mejor cuanta mayor linealidad exista entre la variable de estudio y la auxiliar. En este caso al igual que en la imputación por la media se pueden calcular razones para subgrupos dentro de la población si se dispone de más información. El inconveniente es que se necesita información de buena calidad sobre una variable auxiliar para el total poblacional. Para la obtención de esta información es habitual hacer uso de los registros estadísticos disponibles o otras operaciones estadísticas sobre la misma muestra.

Imputación por regresión

Esta técnica es una generalización de las dos anteriores para un conjunto de va riables auxiliares $x_1, \ldots, x_n$. El modelo más simple es el de la regresión lineal que tiene la siguiente expresión:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n + (e).$$

Donde α es el parámetro de la constante y $\beta_1 ... \beta_n$ serán los parámetros desco nocidos de cada una de las variables auxiliares, e será el error y y es la variable de estudio. Normalmente la estimación de estos parámetros se realiza mediante míni mos cuadrados ordinarios (Scholtus y col., 2014) y por lo tanto se obtiene un modelo desde el que fácilmente se puede predecir el valor de la variable de estudio. El valor estimado se puede obtener tanto sumando el error como no haciéndolo. Sumar el error no es necesario cuando el objetivo de la imputación es obtener medias o tota les poblacionales pero si se quiere observar la variabilidad es conveniente añadir el error (De Waal y col., 2007, cap. 7.3). El modelo sin error siempre tendrá el mismo resultado y por lo tanto es determinista pero la regresión añadiendo el componente de error será determinista si la forma de escoger el error lo es, de lo contrario las imputaciones serán estocásticas.

De esta fórmula se derivan las dos imputaciones explicadas anteriormente. Si la regresión no cuenta con regresores y únicamente utiliza la media de la variable de estudio obtendremos una imputación donde todos los valores faltantes serán la media de los valores respondidos que se corresponde a la imputación por media.

Por otro lado, eliminando la constante y teniendo únicamente un regresor numérico se obtiene la imputación por razón.

Las imputaciones realizadas mediante regresión tienen la ventaja de ser fácil mente calculables pero son demasiado suaves (*smooth*) lo que permite obtener resul tados adecuados para las estadísticas poblacionales como la media o el total pero no son adecuados para tener el fichero final microdatos. Esta forma de

imputación tiende a centrar los valores en la media por lo que será más adecuado cuanto menor sea la variabilidad de la variable de estudio.

Estas técnicas serán más precisas cuanto mayor sea la relación entre las variables y normalmente se especifica una jerarquía en las técnicas para usar la más adecuada en cada unidad. Por ejemplo, si se ha comprobado que la técnica más precisa es la imputación por regresión y la siguiente más precisa la imputación por la media, se imputaría mediante regresión todas aquellas unidades que tengan información auxiliar disponible y para las que no lo tuvieran se usaría la imputación por la media (Särndal y Lundström, 2005, 161).

La regresión es una técnica muy extendida en la estadística y prácticamente todos los softwares permiten su cálculo sin embargo tiene el problema de que para que el resultado sea correcto se deben cumplir los supuestos de linealidad, normalidad y no colinealidad de las variables e independencia y homocedasticidad de los errores del modelo.

***Donor imputation* (*Hot deck*)**

Esta técnica de imputación consiste en seleccionar a un *donor* (donante) para la asignación del valor al *recipient* (receptor con la observación ausente) (Andrid ge y Little, 2010). La selección del donante se puede hacer de diversos métodos pero el objetivo es obtener un donante lo más similar posible para que la imputación sea más precisa. De esta forma, una vez seleccionado el donante el proceso consistirá en imputar el valor del ítem del donante en el receptor. Estas técnicas tienen la des ventaja de que todos los donantes son parte de las observaciones y esto conlleva a asumir que no existen diferencias entre las personas que responden y las que no (Véase por ejemplo, Särndal y Lundström, 2005, página 161).

Andridge y Little (2010) definen el termino *hot deck* como el uso de un donante disponible en el mismo conjunto de datos que los valores ausentes y se contrapo ne a la imputación *cold deck* que consiste en el uso de conjuntos de datos diferentes para la selección del donante como por ejemplo periodos anteriores. En este apar tado se analizarán los diferentes métodos de imputación *hot deck* más usados en la estadística oficial.

Imputación *hot deck* aleatoria y secuencial: Es el método más simple de impu tación usando un donante. El proceso consiste en seleccionar un donante al azar de las observaciones y usarlo para imputar un valor al receptor. Este método tiene la ventaja de no requerir información auxiliar. La imputación *hot deck* aleatoria será útil si el objetivo es obtener una matriz de datos rectangular completa (sin valores

missing) pero las imputaciones no serán muy precisas si existe variabilidad en el fe nómeno.

En el caso de tener información adicional, como por ejemplo, la pertenencia a subgrupos de población se podría restringir la selección aleatoria del donante a ese grupo concreto. Si en vez de realizar una asignación aleatoria del donante se asigna la unidad más próxima en el registro con las características deseadas será imputación ***hot deck* secuencial**. Para este método es recomendable aleatorizar el conjunto de datos (Scholtus y col., 2014).

En ambos métodos la desviación típica de los totales y la media aumentará ya

que siempre existe la posibilidad de que un *outlier* sea el donante. Las estimaciones *hot deck* serán insesgadas únicamente para cuando los valores faltantes son MCAR que es poco probable que ocurra por lo que se recomienda reducirlo utilizando la información auxiliar (Little y Rubin, 2002)

Imputación por Vecino más cercano: Esta imputación se diferencia de las ante riores en que en vez de seleccionar una unidad con las mismas características se selecciona una unidad que minimice una función de distancia previamente defini da. La función general de distancia usada es la distancia de Minkowski (De Waal y col., 2007,251).

En la versión más simple donde solamente tenemos una variable auxiliar el do nante será aquel para el que la diferencia en esa variable sea mínima. En el caso de tener varias variables auxiliares el donante sería aquel que menor suma de todas las distancias tuviera pero nótese que primero las variables se deben estandarizar o usar distancias relativas como la de Mahalanobis (Scholtus y col., 2014). A cada una de estas variables se le puede aplicar una ponderación para así dar más o menos importancia a las variables que se deseen.

Este método obtiene imputaciones deterministas ya que únicamente habrá una observación que minimice la distancia (podría haber empates pero es poco probable) pero también tiene una variante estocástica en la que se seleccionan *k* vecinos más cercanos (Batista, Monard y col., 2002) y después se selecciona al azar o asignando una probabilidad acorde a la distancia, un donante.

La desventaja de esta técnica es que la distancia no puede calcularse de la misma forma para variables categóricas y numéricas pero esto puede solucionarse usando las variables categóricas para crear subgrupos homogéneos donde después extraer el donante. Esto daría más importancia a las variables categóricas que a las numéri cas. Por otro lado, se pueden definir dos funciones de distancia, una para variables numéricas y otra para variables categóricas, y aplicando la ponderación adecuada definir la distancia como la suma de ambas (De Waal y col., 2007).

2.2.3. Imputación de experto

Este método de imputación es el menos estadístico ya que consiste en que el o la responsable del producto estadístico determine el valor *missing* usando su capacidad analítica y toda la información de la que disponga.

2.3. Calidad en las estadísticas oficiales

La calidad según la norma ISO 9000 es: *"el grado en el que un conjunto de carac terísticas inherentes a un objeto (producto, servicio, proceso, persona, organización, sistema o recurso) cumple con los requisitos"*. Para los procesos estadísticos las características principales son relevancia, precisión, fiabilidad, *timeliness*, puntualidad, coherencia, comparabilidad, accesibilidad y claridad (Eurostat, 2017; OECD, 2011).

2.3.1. Relevancia

La relevancia es el nivel en el que los outputs estadísticos cumplen las necesida des de los usuarios. Esta característica está directamente ligada a los intereses de los usuarios por el tema del que se esté desarrollando la estadística y por lo tanto no es una característica de la estadística en sí misma sino del tema en cuestión sobre el que se investiga (Elvers y Lindén, 2015). Para asegurar que la estadística es relevante pa ra la mayor parte de los usuarios el INE (y Eurostat) realiza cada tres años encuestas de satisfacción al usuario donde se encuentran representados los diferentes perfiles de usuario. De estas encuestas se obtienen aquellos puntos en los que hacer hincapié para mejorar la relevancia de los productos. Por ejemplo, en la última encuesta de 2019 la mayoría de usuarios apuntaron que lo más destacable a mejorar sería una mayor desagregación de la información (INE, 2020).

2.3.2. Precisión y fiabilidad

Precisión

La precisión es la cercanía de las estimaciones al valor real. La **diferencia entre el valor real y el estimado se considera error** (OECD) y cuanto menor sea este mayor será la precisión. La precisión es considerada una de las características de calidad más importantes y por eso el estudio de las fuentes de error durante el proceso de producción estadística es amplio.

Fuentes de error: Uno de los enfoques más utilizado para analizar las fuentes de error es el “Total Survey Error” (TSE) (Groves y Lyberg, 2010) pero no permite ana lizar correctamente los diseños de encuesta que tenga múltiples fuentes, muestras o modos. Para solucionar este problema el estadístico Li-Chung Zhang presenta el “Two-phase life-cycle model” (Zhang, 2012) que pretende solucionar los problemas del anterior modelo cuando existe integración de fuentes o métodos de recogida. Ambos modelos se centran en analizar los errores surgidos de la medición del con cepto por un lado y de la representación de la población por otro lado. El modelo de dos fases incluye además el análisis de la integración de dos poblaciones que es algo común cuando se combinan registros administrativos y encuestas. Como se observa en las figuras 2.1 y 2.2 el error surgido de la medida de las variables se analiza en el lado izquierdo y el derecho analiza el error surgido en la representación de las unidades. La figura 2.2 corresponde a la segunda fase donde dos fuentes de datos se combinan para la obtención de un conjunto de datos integrado como ocurre cuando

el marco muestral se selecciona usando registros administrativos pero la informa ción se obtiene usando encuestas.

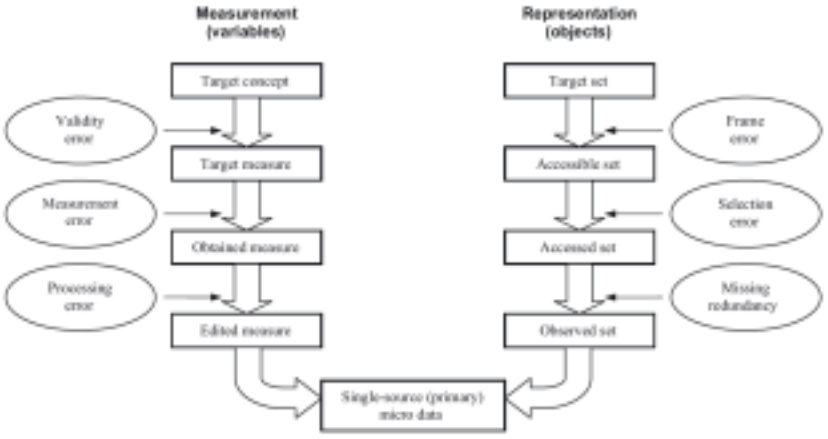


FIGURA 2.1: Fase 1 del Marco de evaluación de la calidad. Fuente: Zhang, 2012.

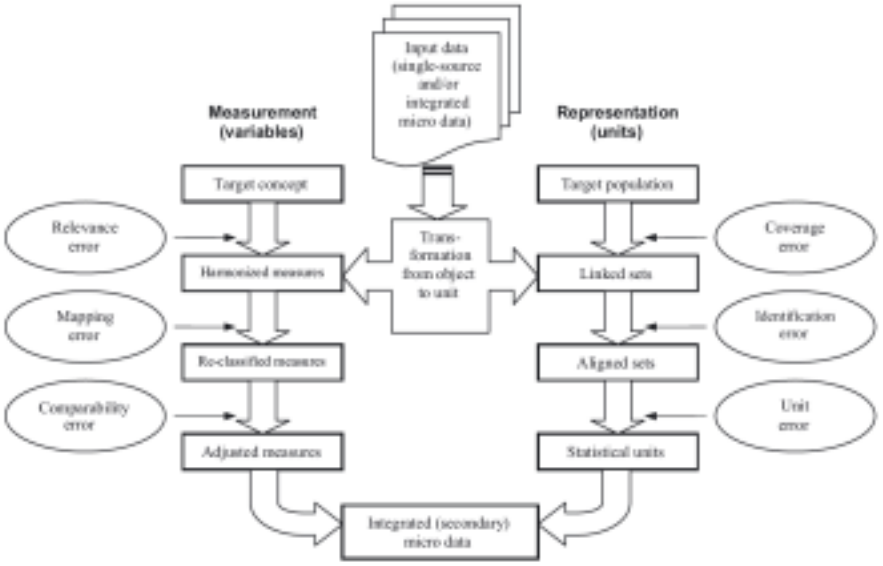


FIGURA 2.2: Fase 2 del Marco de evaluación de la calidad. Fuente: Zhang, 2012.

Error de cobertura: Es el error surgido por no cubrir a toda la población ob jetivo en la muestra. Este error puede ser sobrecobertura o subcobertura. En el primer caso se selecciona una unidad en el marco muestral que no pertene ce a la población objetivo y en el segundo caso serán unidades que no estén en la marco muestral pero que si pertenecen a la población objetivo. Existe un último caso en el que una unidad es clasificada erróneamente en un grupo o categoría a la que no pertenece.

Error de muestreo Es uno de los errores clásicos y más importantes en las estadísticas oficiales y se trata de un "Quality and performance indicators" (QAF) del sistema estadístico europeo y siempre que sea posible deberá ser calculado. Los *QAF* son una serie de indicadores de calidad aprobados por Eurostat en 2010 con el objetivo de representar la calidad de las estadísticas de manera estandarizada.

Este error se puede medir en coeficiente de variación para la estimación o

inter valos de confianza. Cuanto menor sea el CV o el intervalo menor error existirá.

El problema del error de muestreo es que únicamente podrá ser calculado en aquellos casos donde el diseño muestral sea probabilístico. De esta forma, al gunas encuestas económicas del INE no pueden cuantificar este error debido al muestreo por ***cut off***. Esta estrategia de muestreo establece un límite por en cima del cual se seleccionan todas las unidades y por debajo no se selecciona ninguna. Por ejemplo, en ICN se seleccionan todas las empresas por encima de una determinada cifra de negocio en cada comunidad autónoma y actividad económica.

Error por falta de respuesta: La falta de respuesta definida previamente afec ta a las estadísticas producidas incrementando el sesgo y la varianza. El sesgo aumenta debido a la falta de respuesta de la tipología NMAR que considera que la falta de respuesta en las variables de estudio depende del valor de estas variables y por lo tanto las estimaciones no las tendrán en cuenta. La varianza aumenta por el hecho de reducir la muestra efectiva en el caso de reponde ración. En el caso de solucionar el problema mediante imputación, el método para el cálculo de la varianza debería cambiar para que este valor fuera consis tente (Kim, 2001). [4]

Eurostat define 2 indicadores de calidad para este tipo de error:

- **Tasa de falta de respuesta por unidad**: Número de unidades sin informa ción o información no utilizable dividido por el total de unidades objetivo en la muestra (sujeto o no a ponderaciones). Eurostat hace hincapié en la distinción entre esta tasa y la infracobertura definida previamente. Cuan to menor sea este valor mejor será la calidad.
- **Tasa de falta respuesta por ítem**: Para una variable concreta es la razón entre el número de unidades no respondidas y el total de unidades en la muestra (sujeto o no a ponderaciones). Cuanto menor sea este indicador mejor será la calidad.

Error de medida: Es la diferencia entre el dato recogido y el dato real. Este error puede ser sistemático, por ejemplo responder un valor en euros en vez de miles de euros, o puede ser aleatorio. En el caso aleatorio y si se está usando un diseño muestral probabilístico este error estará contemplado en el error de muestreo.

[4]Para más información acerca del cálculo de la varianza después de las imputaciones véase Kim 2000

Error de procesamiento: En todo el proceso de recopilación de datos, codifica ción, edición, estimación etc... pueden ocurrir error de procesamiento por fallos humanos o computacionales. Estos errores suelen ser difíciles de detectar y se debe tener especial cuidado con ellos.

2.3.3. Fiabilidad

La fiabilidad indica la diferencia entre el primer valor estimado y los siguientes valores estimados con lo que nos indica como ha variado el dato en las revisiones. Esta dimensión a menudo no se tiene en cuenta debido a que algunas estadísticas

no están sujetas a revisión pero será un indicador clave en este estudio ya que las revi siones permiten analizar la precisión de las imputaciones y de esta forma determinar como de bueno es el modelo de imputación.

Se utilizan tres indicadores clave para medir la fiabilidad. **MAR** (*Mean Absolute Revision*) indica el tamaño medio de las revisiones en valor absoluto, **RMAR** (*Relative Mean Absolute Revisión*) indica el tamaño relativo al estimador de las revisiones y **MR** (*Mean Revision*) nos indicara si las estimaciones iniciales estaban infraestimadas (MR Positivo) o sobrestimadas (MR Negativo). La revisión de las estadísticas permite al usuario comprender la robustez de las estadísticas analizando la diferencia entre la primera y ultima publicación (McKenzie y Gamba, 2008). El Instituto Nacional de Estadística define su política de revisión como un trabajo llevado a cabo para mejorar la calidad estadística y contemplan diferentes tipologías de revisión además de la inclusión de información otorgada en periodos posteriores. Entre estos motivos de revisión se encuentran un cambio metodológico o la corrección de errores con las fuentes de datos o procesamiento (INE, 2015).

2.3.4. Puntualidad y *timeliness*

El ***timeliness*** se refiere al periodo de tiempo entre el fenómeno de estudio y la publicación de este. Para encuestas económicas coyunturales este periodo es rela tivamente corto por ejemplo ICN tiene un *timeliness* de 51 días. Sin embargo este valor se debe reducir al máximo para preservar la relevancia temporal de las esta dísticas pero siempre sin perder calidad en alguna de las otras dimensiones como la precisión.

La **puntualidad** es la diferencia entre el momento en el que se anunció la pu blicación de las estadísticas y la publicación real. En el INE todas las publicaciones anunciadas se publican exactamente en la fecha y hora indicadas pero aunque ac tualmente no suponga un problema no se debe pasar por alto que es una dimensión primordial para mantener la confianza y seriedad de la institución de cara al público.

2.3.5. Coherencia y comparabilidad

Esta dimensión se refiere al nivel en el que los diferentes procesos estadísticos uti lizan las mismas clasificaciones, definiciones, poblaciones y metodología. Esto per mite el uso conjunto de diferentes estadísticas y la validación cruzada de ambas. Por

ejemplo, el ICN puede validarse conjuntamente con otras estadísticas económicas coyunturales como el IPI o IPRI. Mantener altos niveles de coherencia es primordial para Eurostat ya que la **comparabilidad** entre regiones/países, dominios o periodos temporales es un caso específico de coherencia.

2.3.6. Accesibilidad y claridad

La **accesibilidad** según la OCDE es el nivel de facilidad que existe para obtener información estadística. Como indica el principio 15 del ECoP y el principio primero del los principios fundamentales de las Estadísticas Oficiales de las Naciones Unidas (*UN Fundamental Principles*), los productores de estadísticas

oficiales deben facilitar de forma imparcial el acceso a esta información de manera equitativa. Aunque las TIC hayan permitido una mayor accesibilidad al público general no se debe olvidar la brecha tecnológica y por lo tanto se deben seguir teniendo en cuenta otros medios de difusión para el público no usuario de internet. Por otro lado, se debe hacer hinca pié en la creación de productos estadísticos accesibles para el público con diversidad funcional ya que se debe preservar la equidad en la accesibilidad.

La **claridad** es la dimensión asociada a la fácil comprensión de las estadísticas y está directamente relacionada con los metadatos que las acompañan. Las publicacio nes deben ser comprensibles para todo el público al que van dirigidas y por lo tanto se deben acompañar de metadatos, gráficos, mapas o explicaciones que faciliten la comprensión. El ESS está continuamente intentando mejorar la claridad de sus pu blicaciones y un proyecto con este objetivo en el que pude participar fue "Coding Lab Project" donde se buscaba replicar, utilizando los lenguajes de programación R o Python, los gráficos de los artículos de *Statistics Explained* y que de esta forma el usuario pudiese de manera sencilla comprobar de donde salían los resultados pre sentados (Grazzini, 2021).

2.3.7. Rentabilidad como dimensión de calidad

Aunque el coste de las estadísticas no sea considerado como una dimensión de calidad (Eurostat, 2017; OECD, 2011) si se indica que debe tenerse muy en cuenta. El ECoP menciona en sus principio 9 y 10 la necesidad de reducir la carga de respuesta a los encuestados (Coste del usuario) y mejorar la rentabilidad de las estadísticas (Coste para los productores). Estos costes, sean o no considerados dimensiones de calidad, afectan de manera directa a los productos estadísticos y el desarrollo de nuevas metodologías como la que se explicará en este trabajo pueden ser soluciones interesantes a ambos problemas. Por un lado, el *machine learning* permite hacer una explotación de los datos existentes y de esta forma reducir el coste para el usuario y por otro lado al tratarse de un proceso automatizado la rentabilidad se verá in crementada, aunque el coste de implantación pueda ser relativamente alto, ya que reducirá el coste mensual de producir las estadísticas coyunturales.

2.3.8. Comunicar la incertidumbre:

Las estadísticas oficiales aun teniendo estándares muy altos de calidad son siem pre aproximaciones del valor real. El problema es que a menudo esta incertidumbre no se comunica con la misma claridad que la estimación puntual (Manski, 2014). Es to genera en el público una sensación de certeza que no se ajusta a la realidad y por lo tanto la política de comunicación de los institutos nacionales de estadística debe ría hacer hincapié no solo en comunicar la estimación puntual sino que también se deberían publicar claramente (en notas de prensa etc) indicadores como el intervalo de confianza del indicador o el error con la correspondiente explicación clara de su significado.

2.3.9. Dimensiones más importantes:

Aunque todas las dimensiones se deben tener en cuenta para producir estadísti cas con la calidad óptima, un estudio realizado en España por Costa y

col. (2014) ase gura que las consideradas como más importantes por los usuarios son la precisión y fiabilidad y la menos importante la relevancia (Costa y col., 2014). En el estudio se encuesta a personas de diferentes ámbitos (universidades, administración central y regional, periodistas, etc) y se les pregunta sobre la importancia para ellos de las diferentes dimensiones de calidad. Anteriormente en 2001, Gordon Brackstone, es tadístico de *Statistics Canada* desarrollo una teoría sobre la importancia jerárquica de las dimensiones de calidad. Según su estudio, la relevancia es la dimensión más importante sin la cual el resto carecen de sentido y por eso entiende las dimensiones como una pirámide que se puede observar en la figura 2.3. Una vez satisfecha esta dimensión las siguientes en la jerarquía de importancia son aquellas que permiten a la estadística estar disponible (Accesibilidad y *timeliness*) (Brackstone, 2002). Según este enfoque una vez obtenidas estas tres dimensiones es cuando la interpretabili dad, precisión y coherencia cobran sentido. El estudio de Costa utiliza las respuestas brindadas por los encuestados para construir un modelo que determine la percep ción de calidad total en función de las diferentes dimensiones. Es este modelo el que otorga mayores coeficientes a la fiabilidad y precisión y una menor importancia a la relevancia. Según los autores, esto podría deberse al hecho de que los encuestados no le dan importancia a la relevancia porque la asumen como parte de la estadística y por lo tanto el modelo no le da la misma importancia que al resto de dimensiones.

Relevancia

Accesibilidad Timeliness

Precision Interpretabilidad Coherencia

FIGURA 2.3: Jerarquía de las dimensiones de calidad de G. Brackstone (2001)

Capítulo 3

Encuesta de Cifras de Negocios en la Industria (ICN)

Los índices de Cifras de Negocios de la Industria (ICN) son una estadística co yuntural económica llevadas a cabo por el Instituto Nacional de Estadística. Las esta dísticas económicas coyunturales tienen como objetivo medir el ciclo económico de la manera más inmediata posible. Estas encuestas suelen contar con pocas variables en comparación con las encuestas realizadas en los hogares

o con las económicas estructurales. Debido a la necesidad de representar las tendencias económicas la pu blicación debe ser poco tiempo después del periodo de referencia. Para el caso con creto del ICN el tiempo entre el estudio del fenómeno y la publicación es de 51 días. En estas encuestas al ser una de las prioridades la publicación más pronta posible se reduce el tiempo y recursos disponibles para los trabajos de depuración e impu tación y es por esto por lo que una estrategia de imputación automática mediante *random forest* podría reducir el *timeliness* sin afectar a la calidad. En este capítulo se abordará el proceso para la elaboración de los ICN, definiendo sus características, su proceso de recogida, el cálculo de los índices y su posterior corrección. Por otro lado, se presentarán las variables de la encuesta usadas en el RF del Capítulo 5.

3.1. Índice de Cifras de Negocio en la Industria

ICN tiene como objetivo medir la evolución de la actividad de las empresas que forman parte del sector industrial en España, a partir de sus cifras de negocios. Los resultados se presentan en forma de índices ya que el objetivo es medir la variación de las cifras de negocios tomando como referencia un año base, que actualmente es el 2015 y se actualiza cada 5 años. La unidad estadística es el establecimiento que no tiene por qué coincidir con la empresa por lo que existen variables de identifi cación tanto de empresa como de establecimiento para cada una de las unidades informantes.

La encuesta tiene una publicación mensual e incluye la tasa de variación anual, mensual y media del año en curso a diferentes niveles de desagregación. Estos nive les son:

Según la clasificación nacional de actividades económicas (CNAE-2009) por secciones, divisiones y subdivisiones.

Según el destino económico de los bienes, MIGs.

Según Comunidades autónomas.

Según el mercado donde se ha facturado.

Además los datos se presentan corregidos de efectos de calendario y corregidos de efectos de calendario y estacionales para permitir la realización de comparaciones certeras con periodos previos.

3.1.1. CNAE-2009 en ICN

La CNAE-2009 es la clasificación de actividades económicas que asigna un có digo de hasta 4 dígitos a las empresas en función de su actividad económica. La versión internacional de la clasificación fue aprobada por el Parlamento Europeo con el nombre NACE Rev.2 en 2006 indicando además que esta debe ser la clasifica ción usada en las estadísticas comunitarias (Comisión Europea, 2006). Dependiendo del número de dígitos la desagregación será mayor o menor. El nivel más general se representa con una letra y se refiere a la actividad que en el caso de ICN se co rresponde a la industria extractiva (letra B) e industria manufacturera (letra C). El siguiente nivel consiste en 2 dígitos y representa la división de la

actividad econó mica como puede ser por ejemplo la industria de la alimentación (código 10) dentro de la industria manufacturera. Añadiendo un dígito más a este código se obtiene el grupo de actividades económicas al que pertenece la empresa que siguiendo el mis mo ejemplo es "Procesado y conservación de frutas y hortalizas"(código 103). Por último, añadiendo otro dígito se obtiene la clase que es el nivel más desagregado en esta clasificación y corresponde por ejemplo a "Procesado y conservación de pata tas"(código 1031)

Además de esta desagregación existe otra alternativa que representa un nivel intermedio entre las actividades y las divisiones denominada **MIGs** (Main industrial groupings). Existen 6 MIGs correspondientes a bienes intermedios, bienes de capital o equipo, bienes de consumo, bienes de consumo duradero, bienes de consumo no duradero y energía (Unión Europea, 2020).

3.2. Recogida de información

La recogida de datos en el ICN se realiza a través de las Delegaciones Provincia les del INE. Los datos se recogen en un cuestionario mensual que puede ser enviado por diversas vías como la plataforma IRIA[1], por correo electrónico, postal o fax. Los cuestionarios pueden recibirse a partir del día 1 del mes siguiente al de referencia y van grabándose y depurándose a medida que lleguen. Las Delegaciones Provincia les son las encargadas de enviar esta información parcialmente depurada a Servicios Centrales donde se realiza el resto de depuración, la imputación y el cálculo de los índices. Servicios centrales recibe 3 envíos diferentes a lo largo del mes siguiente al

[1]Portal para el cumplimiento de encuestas online del INE

de referencia, el primero hacia el día 17 del mes siguiente de referencia, el segundo al día 27 y el tercero y último hacia el día 8 de dos meses después del mes de referencia. Con cada envío de información las Delegaciones Provinciales proveen de dos archivos diferentes, PID y PGR que se explican a continuación.

PID: El PID contiene los paradatos o metadatos de procesamiento. Estos paradatos se definen como la información auxiliar que describen al proceso de recogida de los datos (West, 2011). Ejemplos de esto son la fecha de realización de la encuesta, el tiempo que se tarda en completarla o el entrevistador que la realiza. Las variables incluidas en el estudio y extraídas del PID se indican en la tabla 3.1.

PGR: El PGR contiene los datos proporcionados por los establecimientos. Esta es la información del fenómeno que se busca representar por lo que en el ICN serán las 5 diferentes cifras de negocios dependiendo el mercado donde se hayan generado. Las variables incluidas en el estudio y extraídas del PGR se indican en la tabla 3.1.

Por último, una vez depurada la información, Servicios Centrales del INE crea un archivo denominado **FDE** donde aparecen todas las cifras de negocios depura das para el cálculo de los índices finales además de variables de validación para saber por ejemplo si es la primera vez que la unidad aporta información, si el dato fue imputado manual o automáticamente o si la unidad se ha dado de baja de la

encuesta.

3.3. Cálculo de los índices

Para calcular los índices que se publican primero se deben calcular los índices elementales. Estos son los índices por agregado elemental o nivel de desagregación más bajo posible. Para el cálculo de estos índices no existen ponderaciones y se cal culan para cada cruce entre comunidad autónoma y divisiones (dos dígitos) o sub división (tres dígitos) de la CNAE-09 [2]. Una vez calculados estos se obtiene el índice elemental de la siguiente forma:

$$I^{mt}_{i} = {}_{2015}I^{m-1t}_{i} \cdot \frac{\sum_j f^{mt}_{j,i}}{\sum_j f^{m-1t}_{j,i}} \quad \text{. (3.1)}$$

Donde:

I^{mt}_{i} es el índice con base 2015 del agregado elemental *i* en el mes *m* del año t. $f^{mt}_{j,i}$ es el valor en términos monetarios de la facturación del establecimiento *j* que pertenece al agregado elemental *i* (Cruce de actividad y CCAA) en el periodo *mt*.

[2] Actualmente existen 37 índices elementales

Como se puede observar en la fórmula para el calculo del índice sera necesario que el establecimiento haya proporcionado datos por lo menos en el mes *m* y *m-1* con lo que al menos debe llevar 2 meses en la muestra.

Para preservar el secreto estadístico los índices elementales se usan únicamen te para el cálculo de los índices agregados y esta información no se hace pública. Para obtener los índices agregados en primer lugar se obtienen las ponderaciones por agregado que se calculan usando la Estadística Estructural de Empresas: Sector Industrial del año base de los índices de Cifra de Negocios. De esta forma la ponde ración del agregado elemental *i* seria la siguiente:

$$W_i = \frac{\text{Cifra de negocios del agregado } i \text{ en el 2015}}{\text{CN total de la industria Secciones B y C en el 2015}\cdot}$$

Partiendo de estas ponderaciones elementales se obtienen las ponderaciones de agregados como la suma de todas las ponderaciones que se incluyen en ese agrega do.

De esta forma los índices de cualquier agregación funcional se calculan de ma nera intuitiva sumando los índices elementales multiplicados por su ponderaciones de la agregación funcional que se esté calculando:

${}_{2015}I^{mt}_{i} \times W_i$. Donde:

$$_{2015}I^{mt}_{A} = \sum_{i \in A}$$

$_{2015}I^{m}{}_{A}t$ es el índice con base 2015 del agregado *A*.

$_{2015}I^{m}{}_{i}t$ es el índice del agregado elemental *i* que pertenece a la agregación

A. W_i es la ponderación de *i* en el agregado *A*.

3.3.1. Correcciones de los índices

El ajuste de calendario se realiza en base al *Estándar del INE para la corrección de efectos estacionales y de efectos de calendarios de las series coyunturales* que sigue las re comendaciones de la ESS. Tanto las series corregidas por calendario como las corre gidas de efectos estacionales y calendario se obtienen usando el software Jdemetra+ reconocido por Eurostat en 2015 (Grudkowska y col., 2013).

Correcciones de calendario: Los efectos de calendario son aquellas variaciones en las series temporales causadas por el cambio del número de días laborables o festivos o el número de días de vacaciones del mes (Scholtus y col., 2014).La aplicación de estas técnicas es necesaria ya que el reglamento Europeo exige que las estadísticas coyunturales estén corregidas de efectos de calendario (Comisión Europea, 1998).

Los modelos usados para la corrección son del tipo regARIMA que buscan corregir los efectos de días hábiles (número de días hábiles en el mes), festivos y hábiles de Semana Santa y año bisiesto.

Correcciones de efectos estacionales y de calendario Además de corregir los efec tos de calendario mencionados anteriormente se corrigen también los efectos esta cionales. Las fluctuaciones estacionales son movimientos que ocurren con intensidad similar en cada mes, cada trimestre o cada estación del año y que se espera que sigan ocurriendo. Una vez corregidos estos efectos la serie resultante reflejará los cambios en la tendencia, ciclo y componente irregular es decir aquello nuevo de este periodo.

3.4. Imputación en ICN

Actualmente la imputación se realiza mediante imputación por media [3] para ca da uno de los cruces de la Comunidad Autónoma y División/Subdivisión que se publican. La fórmula usada en el INE para estas imputaciones es la siguiente:

$$\hat{y}_k = y^{t-1}_k \times TV^{i}_{\text{intermensual}},$$

Donde:

y^{t-1}_k será el valor de la variable de estudio en el mes anterior al periodo de refe rencia de la unidad *k*.

$TV^{i}_{\text{intermensual}}$ será la tasa de variación intermensual del estrato *i* al que pertenece la unidad *k* (Cruce CCAA y División/Subdivisión CNAE-2009).

Aquí se pueden encontrar dos problemas: el primero que el dato del mes anterior sea demasiado grande o pequeño y el segundo que dentro del estrato haya unidades muy dispares en tamaño. Si el dato del mes anterior es grande porque se ha realizado un pedido importante por ejemplo, el mes siguiente debería reducirse drásticamente su cifra de negocios pero con este método de imputación esto no se contemplará ya que puede que el estrato tenga una tasa de variación positiva. Por otro lado al existir unidades de tamaños diferentes si se da la casuística en la que una de las unidades más importantes del estrato no responde el cuestionario y se imputa con la tasa de variación de las demás unidades pequeñas se puede cometer un error importante ya que para las unidades muy grandes la tasa de variación es menos sensible. Para las unidades de mayor tamaño que no verifiquen los *edits* se suele aplicar imputación por experto debido al gran efecto que puede tener una imputación errónea en estas unidades.

[3] Véanse 2.2.2

3.5. Variables *Random Forests*

Para la construcción del RF se han escogido una serie de variables que actuaran de regresores. Se han clasificado las variables dependiendo el archivo de origen de las mismas y por lo tanto existen variables provenientes del PID, PGR y FDE.

En el cuadro 3.1 se muestran las variables extraídas de los diferentes archivos de recogida del ICN [4]. Entre las variables a destacar se encuentran *cn01.anterior* que co mo es lógico será una de las variables que mejor explique la cifra de negocios ya que lo más habitual será no encontrar variaciones excesivamente bruscas. Otras variables a destacar son aquellas relacionadas con la actividad como *cnaeest3* o *acti.anterior* que indicarán el grupo de actividad al que pertenece la unidad. Por último, la variable que indica el tamaño de la empresa *codtame* es también una de las más importantes a la hora de imputar la cifra de negocios ya que la cifra de negocios esta correlacionada con el tamaño.

Con lo que respecta al cuadro 3.2 en este se muestran las variables derivadas de las anteriores que se ha creído podrían ser interesantes para el RF. Como se observa en la figura 3.1 entre estas variables destacan las variaciones intermensuales e in teranual de la cifra de negocios. Por otro lado también son importantes los grupos MIG y División/Subdivisión al que pertenecen. Otra variable que ha resultado tener importancia para calcular la cifra de negocios es *threshold*

que es una variable dico tómica que indica si la cifra de negocios de la unidad está por encima del percentil 95 en su grupo o no. Esta variable busca corregir los errores de subestimación que se generaban en las unidades muy grandes de cada CNAE.

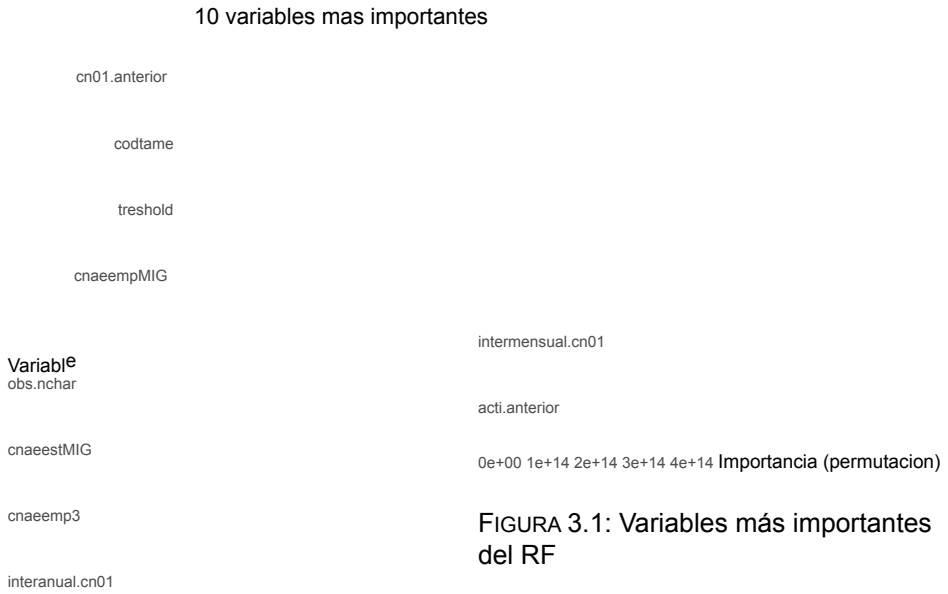


FIGURA 3.1: Variables más importantes del RF

[4]Para el significado de las variables véase las tablas 3.1 y 3.2

Nombre Variable Descripción

Variables PID

numidest Código de referencia de cada unidad sirve de identificador único cnaeemp CNAE de la empresa a 4 dígitos
provem Código de provincia de la empresa
codddpp Código de la delegación provincial que recoge la unidad codtame TAME (código que indica tamaño) de la empresa ntrab número de trabajadores de la empresa
actual A para alta y B para baja
obsanual1 Observaciones anuales
obsanual2 Observaciones anuales
prioridp Prioridad de la unidad, indica la importancia de la unidad de cara a la recogida. Sirve para que las delegaciones sepan con quién es más importante contactar
resulta Información sobre la situación de la recogida
cnaeest CNAE del establecimiento a 4 dígitos
proves Código de provincia del establecimiento

Variables PGR

NUMIDEST Código de referencia de cada unidad sirve de identificador único CN01 Cifra de Negocios total
CN01A Código de actualización de la variable CN01
CN02 Cifra de Negocios del mercado interior
CN02A Código de actualización de la variable CN02
CN03 Cifra de Negocios de la zona euro
CN03A Código de actualización de la variable CN03
CN04 Cifra de Negocios de la zona no euro dentro de la Unión Europea CN04A Código de actualización de la variable CN04
CN05 Cifra de Negocios del Resto del Mundo excluyendo la Unión Euro pea
CN05A Código de actualización de la variable CN05

Variables FDE

CCAA Comunidad autónoma del establecimiento depurada acti Actividad

CNAE-09 a 4 dígitos depurado
CN01.anterior Cifra de Negocios total del mes anterior
CN02.anterior Cifra de Negocios del mercado interior del mes anterior CN03.anterior Cifra de Negocios de la zona euro del mes anterior CN04.anterior Cifra de Negocios de la zona no euro dentro de la UE del mes ante rior
CN05.anterior Cifra de Negocios del resto del mundo excluyendo UE del mes an terior

CUADRO 3.1: Variables usadas como regresores (PGR Y PID)

Nombre Variable Descripción

Variables Derivadas

envioPID Número de envio del PID (1,2 o 3)
cnaemp(3,2,1) Actividad CNAE-09 a 3,2 y 1 dígitos respectivamente de la empresa
cnaeest(3,2,1) Actividad CNAE-09 a 3,2 y 1 dígitos respectivamente del es tablecimiento
match.cnae(4,3,2,1) Coincidencia entre el código CNAE de la empresa y el esta blecimiento a 4,3,2 y 1 digitos
cnaemp**MIG** Main industrial grouping del CNAE correspondiente a la empresa cnaemp**Sub** Sub del CNAE correspondiente a la empresa cnaest**MIG** Main industrial grouping del CNAE correspondiente al esta blecimiento
cnaesest**Sub** Sub del CNAE correspondiente al establecimiento match.cnaeMIG Coincidencia entre el MIG del establecimiento y empresa match.cnaeSub Coincidencia entre el Sub del establecimiento y empresa match.prov Coincidencia entre la provincia del establecimiento y empre sa
match.CCAA Coincidencia entre la CCAA del establecimiento y empresa obs Si el establecimiento tiene observaciones obs.nchar El número de caracteres en el establecimiento **treshold Si la CN01 está por encima del percentil 95 para el cruce de CNAE a dos digitos y CCAA**
interanual.cn01 Variación interanual de la CN01
interanual.cn02 Variación interanual de la CN02
interanual.cn03 Variación interanual de la CN03
interanual.cn04 Variación interanual de la CN04
interanual.cn05 Variación interanual de la CN05
intermensual.cn01 Variación intermensual de la CN01
intermensual.cn02 Variación intermensual de la CN02
intermensual.cn03 Variación intermensual de la CN03
intermensual.cn04 Variación intermensual de la CN04
intermensual.cn05 Variación intermensual de la CN05
intermensual.CCAA Variación intermensual de la CN01 agrupada por CCAA

CUADRO 3.2: Variables derivadas

Capítulo 4

Random forest

Este capítulo comienza explicando los orígenes del random forest en los árboles de decisión para después explicar el funcionamiento de los mismos. A continuación se explica el funcionamiento del algoritmo usado en RF y la metodología seguida para la selección de los hiperparámetros y el *tuning* o elección de los parámetros óptimos.

4.1. Del árbol de decisión al random forest

Los árboles de decisión son métodos de *statistical learning* donde la función apren dida con los datos de entrenamiento [1]toma la forma de un árbol. Los árboles se van dividiendo en nodos que podrían ser entendidos como reglas de "si ocurre esto en tonces este es el resultado"(*if-then rules*). Los árboles de decisión son usados en múl tiples campos debido a su sencillez y la interpretabilidad de los resultados (Mitchell, 1997).

Los árboles de decisión son interpretados de manera muy sencilla. Todos los ár boles parten de una raíz o *root* y acaban en diferentes nodos hoja o *leaf node*. En cada decisión el árbol se divide en dos nodos en función de la variable regresora o clasi ficadora más importante [2]. Dependiendo de si la variable de estudio es categórica o continua el tipo de árbol será distinto siendo los primeros árboles de clasificación y los segundos de regresión. También existen árboles de supervivencia (Bou-Hamad y col., 2011) para analizar conjuntos de datos de supervivencia pero no serán objeto de este estudio ya que los datos usados no son de esta tipología. Una vez construido el árbol completo se obtiene la lista de reglas de decisión que serán las usadas para clasificar o predecir una nueva observación. Cuando se utilizan árboles de decisión para predecir la variable de estudio en una observación simplemente se debe ir eva luando las reglas hasta alcanzar el *leaf node* correspondiente. Finalmente el valor que se le asignará a esa predicción será la categoría más probable entre el conjunto de datos de entrenamiento en ese *leaf node* o la media de la variable de estudio de las observaciones que estén dentro del nodo final.

Llamamos **partición binaria recursiva**, al conjunto de algoritmos mediante el cual se obtienen los resultados de los árboles de decisión. Entre los algoritmos más

[1]Los datos de entrenamiento son aquellos usados en aprendizaje automático para construir los modelos.

[2]La importancia de la variable dependerá del algoritmo escogido.

utilizados encontramos **ID3** , **CH4.5** (Hssina y col., 2014) o el más utilizado por la comunidad estadística **CART** (*Classification regression tree*) (Alsagheer y col., 2017). Este método se utiliza ampliamente debido a sus propiedades que son las siguientes:

Técnica no paramétrica y libre de tests de contraste.

Se pueden usar variables predictoras y de estudio de todo tipo (Continuas, dicotómicas, categóricas etc).

No es necesario imputar los datos *missing*.

Es una técnica robusta a la que no afectan los *outliers*.

El algoritmo de **CART** comienza con todas las observaciones en un nodo inicial. Realiza la mejor partición binaria entre todos los predictores y se divide el nodo según esta partición. La selección de la mejor partición posible puede basarse en diferentes criterios pero los más comunes son el índice de Gini para clasificación y la suma de los residuos de los cuadrados para árboles de regresión (Cutler y col., 2012). Este paso se repite tantas veces hasta que se cumpla el criterio de parada [3](Merkle y Shaffer, 2011). Una vez obtenidos todos los nodos la predicción de la variable de estudio sera la media o moda del nodo final donde la nueva unidad acabe. En la figura 4.1 se observa un ejemplo ilustrativo de partición binaria realizado con la librería *rpart* de R. En este caso el árbol cuenta con dos particiones únicamente y 3 nodos finales. Ambas particiones se realizan utilizando la variable edad ya que es la más relevante en este caso. De esta forma, si una nueva unidad con 50 años llegara y tuviese que ser clasificada acabaría en el nodo verde de la derecha clasificada como persona ocupada ya que es la categoría más común entre las observaciones con esas mismas características.

FIGURA 4.1: Ejemplo Árbol de decisión

El principal problema de estos algoritmos es la determinación del tamaño del ár bol ya que un árbol demasiado grande estará sobreajustado pero un árbol demasiado pequeño puede no predecir bien la variable de estudio. T. Mitchell (1997) indica que un tamaño del árbol superior a 25 nodos hace que la precisión del árbol disminu ya progresivamente a medida que se aumenta este tamaño. Es por esto por lo que

[3]Se denomina así al criterio escogido por el investigador para dejar de generar nuevos nodos.

lo usual es hacer crecer a un árbol y después "podarlo"(*pruning*). Dependiendo del algoritmo se utilizan diferentes métodos de poda pero uno de los más habituales es el de complejidad-error (Breiman y col., 1984) que como su nombre indica tiene en cuenta tanto la complejidad (tamaño) del árbol como el error de predicción [4].El *pru ning* no es necesario en los RF pero el tema será discutido con mayor profundidad en el apartado 4.3.

Para mejorar el rendimiento predictivo de los árboles se han desarrollado méto dos de combinación de árboles que obtienen mucho mejores resultados predictivos con el inconveniente de perder interpretabilidad. Dentro de estos métodos destacan el *bagging, boosting y random forest* (James y col., 2013).

Bagging (Breiman, 1996): En lugar de ajustar un único árbol, se ajustan muchos de ellos en paralelo formando un ”bosque”. En cada nueva predicción, todos los árboles que forman el ”bosque” participan aportando su predicción. Como valor final, se toma la media de todas las predicciones (variables continuas) o la clase más frecuente (variables cualitativas). Este método surge debido a la gran variabilidad que tienen los árboles al depender del conjunto de entrenamiento. Para reducir esta varianza se hace uso del bootstrap[5] y después se calcula un árbol para cada una de las muestras. El resultado final será el promedio de todos los árboles generados. RF es uno de los métodos más conocidos de *bagging*.

Boosting (Drucker y Cortes, 1996): Los árboles se generan de forma parecida al *bagging* pero en este caso de manera secuencial, el segundo árbol tendrá en cuenta al primero y sucesivamente. Cada árbol se generará usando la información de los re siduos de los árboles anteriores denominados *weak learners* hasta obtener el óptimo. Como valor final, al igual que en *bagging*, se toma la media de todas las predicciones (variables continuas) o la clase más frecuente (variables categóricas). Los algoritmos más utilizados son *AdaBoost* , *Gradient Boosting* y *Stochastic Gradient Boosting* (Versión estocástica del anterior).

Random Forests (Breiman, 2001): Es una técnica similar al *bagging* pero en este caso se añade un parámetro *mtry* que indica el número de regresores que tendrá cada división. Estos regresores se escogen aleatoriamente en cada división para que si una variable es muy importante no resulten todos los árboles similares, esto es, decorrelaciona los árboles. El funcionamiento de los RF se explicará detalladamente en el apartado 4.2.

4.2. Metodología *Random Forest*

Leo Breiman definió los random forests en 2001 como clasificadores consistentes en un conjunto de árboles de decisión donde cada uno de los árboles usa un vector aleatorio e independiente θ_k resultante en un clasificador $h(x, \theta_k)$ siendo x el vector

[4]Para más métodos de poda véase Mingers, J. 1989.

[5]Generar *b* muestras con remplazamiento de la muestra observada

input. De esta forma, la clasificación será aquella más popular en el conjunto de los *k* árboles. Este es por lo tanto un método de **aprendizaje supervisado** lo que significa que tanto el input (regresores) como el output (variable de estudio) deben ser indicados al usarlos. Al igual que con los árboles de decisión dependiendo de la variable respuesta que se estudie existen RF de clasificación o regresión.

Los random forest podrían considerarse como una evolución del *bagging* y una técnica competidora al *boosting* (Cutler y col., 2012). Se considera una evolución ya que ambos métodos buscan reducir la varianza pero el RF decorrelaciona los dife rentes árboles generados para así obtener una mayor reducción de la varianza[6]. En el *bagging* esta correlación ocurre ya que muchas veces una variable tiene una impor tancia muy superior al resto. Por ejemplo, si quisiéramos determinar si una persona esta jubilada o no es coherente pensar que la variable

que generaría el primer split sea la edad y por lo tanto todos los árboles generados serían similares. Para elimi nar la correlación los RF eligen un número aleatorio m de los p predictores en cada partición. De esta forma al menos $(p - m)/p$ de las particiones no cuentan con es ta variable importante y los árboles serán menos similares (James y col., 2013). Con $m = p$ RF será lo mismo que el *bagging* por lo que es importante la elección de m. El número de predictores aleatorios que se usa por defecto para clasificación es $m = \sqrt{p}$ y para regresión $m = p/3$ (Breiman, 2001; Cutler y col., 2012; James y col., 2013) pero este parámetro sera analizado con mayor profundidad en el apartado 4.3.

4.2.1. Construcción del RF y algoritmo

Como se ha mencionado anteriormente los RF son una técnica de *ensemble* de los árboles de decisión por lo que su construcción está basada en la combinación de múltiples árboles de decisión. El algoritmo utilizado para la construcción de estos árboles es CART ya que tiene la ventaja de poder utilizar tanto variables categóricas como numéricas como regresoras y acepta variables de estudio de ambas tipologías también (Lewis, 2000). Los parámetros principales a indicar para la construcción del modelo son pocos y se indican en el cuadro 4.1. Teniendo en cuenta estos parámetros el RF se construye siguiendo el algoritmo 1.

[6]La reducción de la varianza de la media de árboles será proporcional a la correlación

Parámetro	**Definición**	**Valor por defecto**
mtry o *m*	Número de variables aleatorias usadas en cada partición	$\sqrt{p}$, $p/3$ en regresión
n	Número de observaciones en cada árbol	*N*
replace	Usar o no reemplazamiento en la muestra	SI
min.node.size	Número mínimo de unidades en un nodo final	1 clasificación y 5 regre sión
num.trees	Número de árboles en el RF	500 o 1000
splitrule	Criterio usado para hacer las divi siones	*Gini impurity*, *Variance*

CUADRO 4.1: Parámetros principales del RF

Algoritmo 1: Algoritmo de Breiman (Cutler y col., 2012; Liaw, Wiener y col., 2002)

Seleccionar el conjunto de entrenamiento S para cada uno de los j árboles que comprendan el RF.

1. Obtener una muestra *bootstrap* de tamaño n de S, obteniendo así S_j
2. Usar S_j como conjunto de entrenamiento y aplicar partición binaria recursiva (CART) pero con las siguientes especificaciones:
   - *a*) Empezar con todas las observaciones en el nodo raíz
   - *b*) Para cada partición hasta que se cumpla el criterio de parada realizar: 1) Seleccionar *mtry* predictores aleatorios.

2) Hacer la partición óptima (según el *splitrule*) sobre esos *mtry* predictores.

Las predicciones para regresión serán la media de las observaciones que haya en el nodo terminal correspondiente y la moda para los árboles de clasificación.

4.2.2. Propiedades y características

Las principales características de los RF son la clave de su extendido uso en la investigación. En general la técnica se usa debido a su facilidad para trabajar con conjuntos de datos de alta dimensionalidad que combinen variables de diferentes tipos sin la necesidad de crear un modelo paramétrico. En este apartado se men cionarán algunas de las propiedades y características de los RF que demuestran su verdadera utilidad.

Medida de proximidad: Los RF proporcionan una medida de proximidad entre dos observaciones como la proporción de árboles en los que ambas terminan o no en el mismo nodo final. Esta proximidad será útil para la imputación y visualización.

En el caso de que en todos los árboles dos unidades acaben en el mismo nodo esta medida será de 1 y en caso contrario 0. De esta forma, se puede obtener una medida de distancia ajustada que otorgará mayor peso a aquellos predictores más relevan tes para la predicción de la respuesta y menor peso a aquellos irrelevantes para el estudio (Cutler y col., 2012).

Validación Out Of Bag: (OOB) Al usarse una muestra *bootstrap* de tamaño *N* en cada construcción de un árbol no se utilizan todas las observaciones posibles si no que algunas se repetirán y otras unidades no participarán en la creación del mismo. En el caso de no usar *bootstrap* la muestra de cada árbol será de *N*/2 y la mitad de las observaciones serán OOB en cada ejecución. Para cada árbol, las unidades que no participen en su construcción se denominan *Out of Bag observations* y son especialmente útiles para determinar el error de la estimación y la importancia de los predictores. Calcular el error de las predicción sobre el conjunto de entrenamiento sería demasiado optimista y es por esto que el error de generalización se calcula para cada unidad con los árboles en los que la unidad no ha participado o sea en aquellos que era OOB. Esta forma de validación es computacionalmente mucho más eficiente que el *bootstrap* o la validación cruzada ya que el propio método calcula el error de generalización OOB automáticamente. Este error se usa comúnmente para determinar el número de árboles adecuado para cometer un error de generalización pequeño sin incrementar drásticamente el tiempo de procesamiento y así establece el *mtry* óptimo (Friedman y col., 2001).

Importancia de las variables (VIM): Los random forest son capaces de calcular la importancia de las variables predictoras utilizando diferentes métodos donde desta can los presentados por Breiman y Cutler (Breiman, 2015). El primer método o me dida de importancia de las variables (*VIM*) es la **permutación**. Esta técnica consiste en construir un árbol de decisión. Una vez obtenido el árbol usarlo para

predecir el valor de las observaciones OOB y calcular el error del árbol. Una vez obtenido esto el valor de la variable de la que se está calculando la importancia es permutado en todas las observaciones OOB. Se repite el mismo proceso que con los valores reales y se computan los errores de clasificación o el error cuadrático medio (regresión) y se obtendrá la importancia de la variable como la diferencia entre el error del con junto original y el del conjunto permutado (Cutler y col., 2012). Es intuitivo pensar que si el cambio de valores en una variable afecta poco a la predicción esa variable tendrá poca importancia y viceversa. Mientras que el primer método mide por lo tanto el grado de perdida de precisión, el segundo método que es la ***Gini importan ce o impurity importance*** mide el grado de disminución de la impureza (*impurity*)[7] (Nembrini y col., 2018). Este método es mucho más simple ya que únicamente cal cula la perdida de importancia de Gini al dividir un nodo en la variable objetivo y lo agrega a lo largo de todo el RF. Este método es mucho más rápido que el anterior y tiene resultados similares. Según Cutler (2012), aunque el RF no se vea afectado de

[7]Para más información acerca de los diferentes VIMs véase Nembrini y col., 2018.

manera brusca por la inclusión de variables irrelevantes lo común es eliminar aque llas en las que la importancia sea cercana a cero y hacer crecer un nuevo bosque con únicamente estas variables .

4.3. Hiperparámetros y *tuning*

Los hiperparámetros son aquellos parámetros que controlan el proceso de apren dizaje en las diferentes técnicas de *statistical o machine learning*. El RF es conocido por no requerir demasiada atención en el ajuste de estos parámetros ya que el efecto que tienen sobre el algoritmo no tiene gran influencia (Kuhn, Johnson y col., 2013). En esta sección se explicará cuáles son los valores óptimos según diversos expertos:

4.3.1. Número de predictores (*m* o *mtry*)

El número de predictores aleatorios que seleccionar en cada división. Este pa rámetro es clave para añadir aleatoriedad al algoritmo ya que con un *mtry* = *p* el random forest sería determinista salvo por el *subsampling* (Scornet, 2017). Según Breiman lo óptimo sería seleccionar el número de variables por defecto explicado en el apartado anterior y la mitad y el doble de este número para finalmente escoger de entre esos tres aquel que mejores resultados obtenga (Breiman, 2001). Usar un *mtry* pequeño resultará en árboles menos correlacionados lo que aumentará la estabili dad. Sin embargo también existirá la posibilidad de que variables no importantes sean seleccionadas y por lo tanto la precisión será menor. Para conjunto de datos de gran dimensionalidad será importante la selección de un *mtry* grande ya que se debe asegurar que al menos una variable importante aparezca en las divisiones. Sin em bargo, en el caso de conjuntos de datos que tienen muchas variables relevantes este parámetro debe ser pequeño para que se seleccionen tanto las variables influyentes como las no influyentes en las divisiones (Probst y col., 2019).

El tiempo de procesamiento computacional es proporcional al número de *mtry*

ya que la mayoría de la memoria en el RF es usada para determinar la división óptima y esto será más costoso cuantas más variables haya (Wright y Ziegler, 2017).

4.3.2. Tamaño del bosque (*num.trees)*

El número de árboles a escoger se entiende sencillamente ya que corresponde al número de replicas en una simulación de Monte Carlo con lo que cuanto mayor sea el número de árboles mejores serán los resultados. Sin embargo, el número de árboles afecta negativamente al tiempo de computación con lo que la técnica usada comúnmente es escoger todos los demás parámetros y determinar el tamaño del bosque como el momento en el que el error alcanza su límite de reducción (Cutler y col., 2012; Probst y Boulesteix, 2017; Probst y col., 2019; Scornet, 2017). Los valores por defecto suelen ser 500 o 1000 pero la elección de un mayor número de árboles nunca afectara negativamente a los resultados únicamente a la memoria utilizada para su cálculo y por lo tanto al tiempo de clasificación o regresión.

4.3.3. Tamaño de la muestra (*n*) y reemplazamiento

El tamaño de la muestra determina el número de observaciones a escoger para el entrenamiento de cada uno de los árboles. Su efecto es similar a *mtry* ya que un tamaño de muestra menor creará árboles más diversos con lo que la precisión agre gada aumenta pero la precisión individual se ve reducida por el menor tamaño de muestra de entrenamiento. Encontrar el equilibrio entre precisión y estabilidad será la clave para la selección de este parámetro según Probst (2019).

El análisis de Martinez-Muñoz y Suárez (2010) sobre el valor óptimo de este pa rámetro determino que depende del fenómeno de estudio en cuestión pero que se puede optimizar usando la estimación OOB del error del RF. Por otro lado, los inves tigadores determinaron que el uso de un número de muestra inferior al establecido por defecto es beneficioso ya que reduce el coste computacional construyendo árbo les más simples.

Mientras algunos autores coinciden en que el uso del remplazamiento no afecta al rendimiento (Martínez-Muñoz y Suárez, 2010) otros estudios concluyen que al usar boostrap con reemplazamiento se incluye un pequeño sesgo con las variables categóricas de diferente número de categorías (Janitza y col., 2016).

4.3.4. Número de unidades en nodo final (*min.node.size*)

Este parámetro indica el número mínimo de unidades en un nodo final y cuanto más pequeño sea, más grande (mayor número de divisiones) será el árbol. Al igual que con otros parámetros el problema de situarlo en los valores por defecto de 1 o 5 dependiendo el tipo de árbol es el coste computacional. Un aumento del tamaño de los nodos hará que el tiempo de computación se reduzca exponencialmente y su rendimiento predictivo no se ve afectado de manera drástica (Probst y col., 2019)

4.3.5. Regla de partición (*splitrule*)

La regla de división es la fórmula por la cual el algoritmo determina cual es la

división óptima en cada una de las particiones del árbol. Existen diferentes métodos pero los definidos por Breiman (2001) fueron para clasificación la **impureza de Gini** y la **varianza ajustada** para la regresión. Ambos métodos favorecen la división en variables que tienen una gran cantidad de posibilidades de división como son las variables continuas o las categóricas con muchas categorías (Probst y col., 2019).

4.3.6. Poda o *pruning*

La poda consiste en la reducción del tamaño del árbol de decisión para no so breajustar el modelo. Sin embargo, este problema que es común en la metodología **CART** no ocurre en los RF. Como ya justificó Breiman (2001) no es necesaria la poda debido principalmente a dos razones. Por un lado, los árboles se construyen usando una muestra boostrap y por otro lado cada árbol se construye usando unos predic tores distintos. Esto consigue que el RF tenga árboles con alta capacidad predictiva y decorrelacionados entre ellos.

Capítulo 5

Resultados y calidad de las imputaciones

En este capítulo se explicará la metodología seguida para la construcción de los RF así como el posterior análisis de los resultados obtenidos con estos. Cabe destacar que el estudio se ha realizado para los ICN de noviembre de 2020 pero el mismo pro cedimiento podría realizarse para la imputación de datos en cualquier otro periodo.

5.1. Preprocesamiento de los datos

Esta sección explica el proceso llevado a cabo para la obtención del conjunto de datos necesario para la aplicación del RF. Como se explicó en el el Capítulo 3, ICN recoge datos continuamente pero estos son enviados a servicios centrales únicamen te en 3 momentos para cada mes de referencia. A lo largo de este capítulo se usará el termino *envío* para referirse a los 3 diferentes momentos en los que la información llega a Servicios Centrales del INE y puede usarse para la imputación. Este capítulo solamente abordará el análisis del envío 1 pero los resultados del 2 y el 3 pueden encontrarse en A.3 y A.4 respectivamente.

5.1.1. Conjunto de datos

Como se ha mencionado anteriormente en el Capítulo 3 ICN cuenta con 2 ar

chivos de recogida de datos (PID y PGR) y 1 archivo de datos depurados (FDE). La muestra total *n* es de 12402 unidades o establecimientos en noviembre de 2020. De las cuales 11963 son válidas. Con estas cifras el ICN tuvo una tasa de respuesta en este periodo del 93 %. Sin embargo, el objetivo de este estudio es obtener imputación no únicamente para el resultado final del mes de referencia sino para cada uno de los envíos que se realizan. Como es lógico, la cantidad de unidades que responden en el primer envío es inferior al 80 % con lo que la tasa de imputación será superior en el primer envío e ira disminuyendo hasta el 3 envío de información por las Delega ciones Provinciales. El número de unidades informantes e imputaciones por envíos aparecen en la tabla 5.1. Para la distribución relativa de las unidades por MIGs y división/subdivisión véase anexo A.2.2.

Envío	Recibidos Validos Imputados Tasa de imputación
1	11963 2699 22,56 %
2	10465 1178 11,11 %
3	11801 741 6,27 %

CUADRO 5.1: Número de unidades informantes e imputaciones por envío

5.1.2. Preparación ICN

Los datos enviados por las Delegaciones Provinciales consisten debido al diseño en tres envíos de información en 6 archivos de datos, tres correspondientes al PGR y los mismos correspondientes al PID. El primer paso para trabajar con los datos es por lo tanto obtener estos 6 ficheros y determinar las unidades que serán impu tadas y las que se usarán de entrenamiento. Para este estudio se han usado todas aquellas unidades con información completa (envío de cuestionario válido) como conjunto de entrenamiento y las unidades donde no se tiene información acerca de la cifra de negocios serán el conjunto sobre el que realizar las imputaciones. No se ha usado grupo de control debido a la naturaleza de los RF que permiten usar las observaciones OOB como validación cruzada.

Una vez realizado este trabajo se extrajeron las variables regresoras que se cre yeron interesantes que están disponibles en la tabla 3.1 y se calcularon las variables derivadas disponibles en la tabla 3.2. Sin embargo el algoritmo del RF de la libre ría **ranger** no puede trabajar con valores *missing* con lo que en el próximo apartado se explicará cómo se redujo la cantidad de unidades con valores *missing* sin perder información útil en el modelo.

5.1.3. Missing values

El tratamiento de los valores *missing* es clave en los RF ya que las observaciones que presenten algún ítem sin respuesta de entre todos los regresores no podrán ser usadas para el entrenamiento del método [1]. En este tratamiento se busca eliminar todos aquellos valores ausentes de la tipología **MCAR** [2] que se corresponden a falta de respuesta aleatoria.

Comunidad autónoma, provincia y actividad: Estas variables presentaban falta

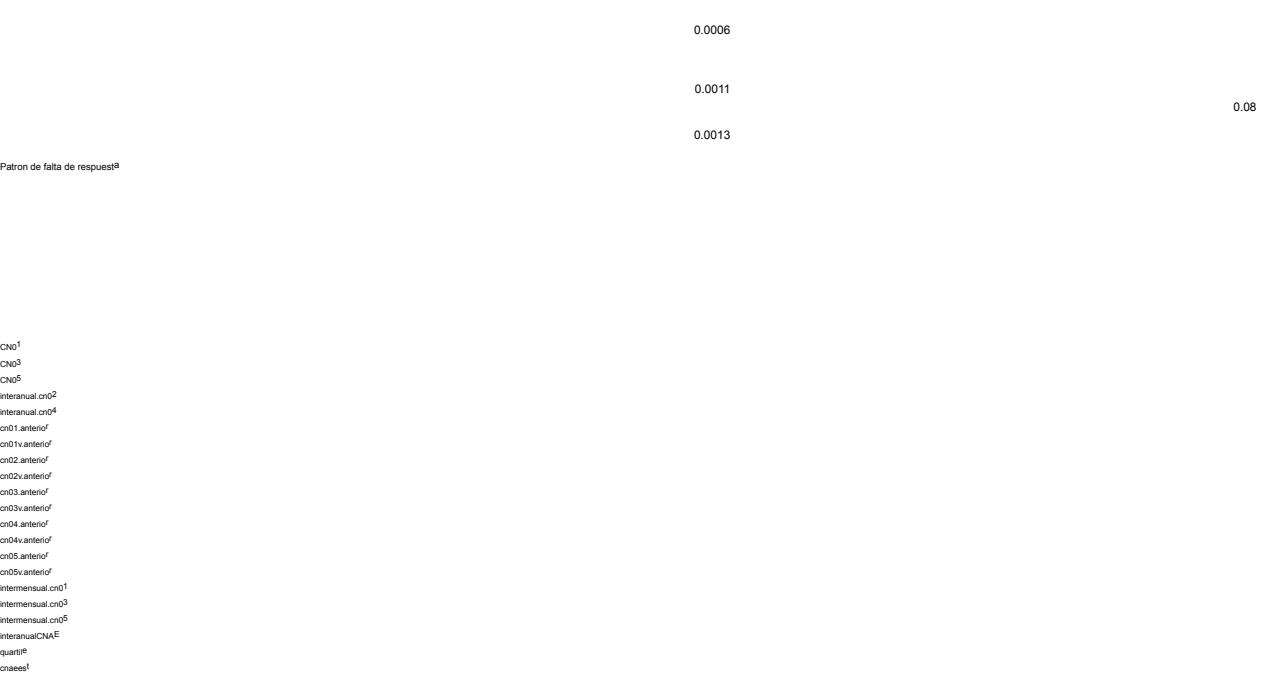
de respuesta en algunas observaciones por lo que la forma de solucionarlo fue; por un lado el uso del archivo FDE depurado para las observaciones que ya habían in formado en periodos anteriores y para aquellas nuevas observaciones el PID. La realización de esto podría conllevar un pequeño error debido a que si una empresa ha cambiado de localización o de actividad económica en el último mes este cambio no estaría contemplado pero esto sería poco habitual en ICN debido a la dificultad de cambiar la localización o el proceso productivo en la industria.

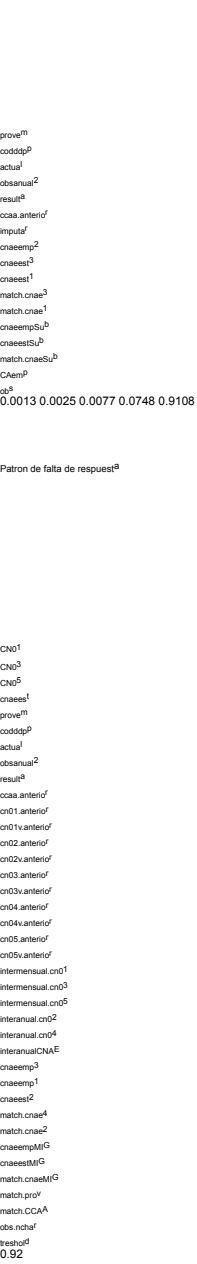
[1]Para variables categóricas donde no se tenga información se puede crear una nueva categoría para los valores *missing* y así solventar el problema.

[2]Véase tabla 2.1.

Sin embargo, esta falta de respuesta no se puede solucionar en todos los casos. Existen 218 observaciones de las cuales no se puede calcular la variación interanual debido a que llevan menos de 1 año en la encuesta lo que hace que se deban eliminar del estudio. De estas además, 86 son de nuevo ingreso en el mes de referencia lo que hace que no se tenga información en el FDE y por lo tanto no sean adecuadas para su uso. En el caso de las observaciones que llevan más de un periodo en la muestra pero menos de un año una solución para no perder estas observaciones podría ser la imputación de las tasas de variación interanuales. No obstante, debido a la pequeña cantidad de unidades se ha optado por eliminarlas del estudio con la intención de incorporarlas en periodos posteriores cuando se pueda calcular esta variación.

Como se observa en el gráfico 5.1 (a) creado con el paquete *VIM* (Kowarik y Templ, 2016) el patrón de falta de respuesta más común es el interesante para el estudio ya que se trata de aquellas observaciones donde no se dispone de la información rela tiva al PGR (Cifra de negocios en sus 5 variantes) que será imputado mediante RF. Sin embargo tras la depuración se obtiene el patrón (b) que es el necesario para la construcción del modelo donde únicamente falta información acerca de las cifras de negocios. En total se realizarán 4618 imputaciones siendo la mayoría (2699) corres pondientes al primer envío.





(b) Después

(a) Antes

FIGURA 5.1: Patrón de respuesta antes y después de tratamiento de *missing*

5.2. Resultados

En esta sección se construirán los bosques aleatorios para después analizar en el apartado siguiente cuál de ellos ha tenido una mejor *performance*.

El paquete **ranger** (Wright y Ziegler, 2017) con el que se construyen los árboles utiliza el criterio de reducción de la impureza de los nodos midiéndose con la varian za de respuesta estimada que se calcula según la fórmula 5.1 donde y es la variable de estudio, μ la media de esta variable en el nodo correspondiente y n el número de unidades que hay en el nuevo nodo (Steinberg, 2009). Siguiendo esta fórmula la división buscará que los dos nodos sean lo más homogéneos posibles para que la varianza sea menor y por lo tanto la reducción de la varianza mayor.

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2. \quad (5.1)$$

Para cada uno de los envíos se construirá un RF y se imputarán los valores *mis sing* para posteriormente comprobar su calidad con diferentes indicadores. Antes de realizar el modelo se comprueba que la información provista por las unidades es aceptable gracias a la variable de paradatos *resulta*. Con esta variable algunas uni dades que habían otorgado información pasan a ser valores *missing* y por lo tanto susceptibles de imputación y otras unidades pasan a tener cifra de negocios 0 por que la unidad no haya operado durante ese periodo o cualquier otra situación que lo requiera indicada por la delegación provincial que recoge los datos.

5.2.1. Selección de los parámetros óptimos

Este apartado se centrará en la selección de los parámetros óptimos para la cons trucción del RF correspondiente al envío 1.

Eliminar variables no importantes

Como sugiere Cutler (2012) es interesante realizar un primer RF para determinar aquellas variables con importancia cercana a 0 y eliminarlas del modelo. Siguiendo este consejo se ha realizado un RF con los parámetros predeterminados mostrados en la tabla 5.2. Como se aprecia en la figura A.2 encontramos varias variables con importancia negativa lo que quiere decir que la permutación de los valores en la submuestra OOB ha conllevado en una mejora del error del modelo con lo que estas variables no son buenas para predecir el valor de la cifra de negocios. Sin embargo, al ser esta una primera aproximación se ha optado por mantener todas las variables y considerar la eliminación de algunas en trabajos futuros donde se analicen más periodos y se obtenga una lista de variables poco importantes consistentes a lo largo de diferentes periodos.

Número de árboles

Para determinar el número de árboles óptimo del RF se construyeron modelos con las características predeterminadas que se observan en la tabla 5.2 y con diferen tes valores en el parámetro *num.trees*. Se han escogido 2 errores para este análisis. El error más importante es el error OOB que es aquel que se mide sobre cada árbol con

las observaciones que no han sido usadas para su construcción. El error *RMSE* (véa se formula 5.2) OOB sirve de validación cruzada ya que usa las observaciones del conjunto de entrenamiento no usadas en cada árbol. Además, la estimación de error OOB convergerá hacia la validación cruzada *leave-one-out* con el tamaño de árboles suficiente (Friedman y col., 2001). El otro error evaluado se corresponde al *RMSE* del conjunto de datos de entrenamiento. Como es lógico el error más bajo es aquel del conjunto de entrenamiento y por eso se usa el OOB error que es menos optimista. Los resultados de la figura 5.2 muestran que el número de árboles óptimo es 221 donde el error OOB alcanza un mínimo. Aunque el mínimo del gráfico se encuentre en 1 árbol esto se debe a que la muestra a

OOB de este único árbol aleatoriamente ha tenido muy buenos resultados para esa semilla pero este error estará sesgado y por eso en el siguiente número de árboles comprobado tiene un aumento pronunciado del error. Una vez seleccionado este parámetro el resto se debe seleccionar probando diferentes combinaciones o lo que es lo mismo usando un *grid*.

mtry n replace num.trees splitrule Var.Imp. seed p/3 9618 TRUE
500 Varianza Permutación 123

CUADRO 5.2: Parámetros RF inicial

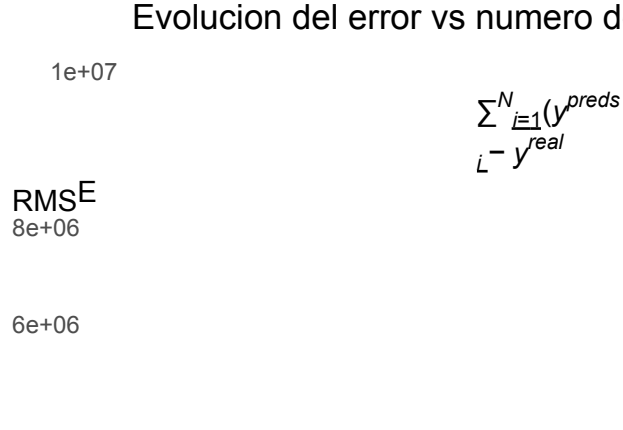


FIGURA 5.2: Evolución del error en función del número de árboles

s

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i^{preds} - y_i^{real})^2}{N}} \quad (5.2)$$

Número de predictores (*mtry*) y tamaño de los nodos (*min.node.size*)

Como se ha mencionado en la sección anterior estos parámetros se escogen uti lizando un *grid*[3]. De esta forma se evalúan los diferentes RF posibles para escoger aquel que menor error OOB y mayor R^2tenga. El R^2 o coeficiente de determinación

[3]Conjunto de diferentes combinaciones de parámetros.

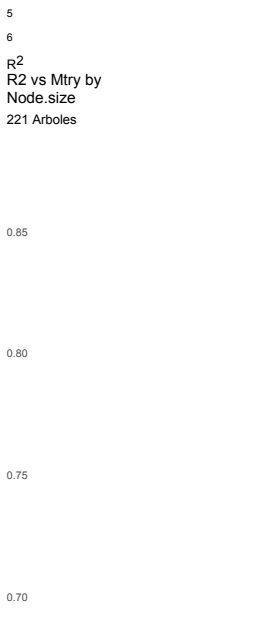
es otra medida de calidad que representa el porcentaje de variabilidad explicada por las variables independientes (Conjunto de predictores) sobre la variable dependiente (Cifra de negocios) (véase la fórmula 5.3).

$$R^2 = (\sum_i y_i - \hat{y}_i^{preds})^2 \quad (\sum_i y_i - \bar{y}_i)^2 . \quad (5.3)$$

En el análisis de *grid* ilustrado en la figura 5.3 se observa que el OOB RMSE depende de la cantidad de variables predictoras que se usen en cada división de los nodos obteniendo un mínimo error con 33 variables predictoras de las 50 existentes en el conjunto de datos. Este valor es mucho mayor al recomendado por Breiman 50/3 y puede deberse a la gran importancia que tienen pocas variables del modelo lo que hace necesario que alguna de estas esté incluida para que el error individual de ese árbol no sea demasiado grande. Como era de esperar el R^2sigue un patrón similar pero alcanzando su máximo en el mismo modelo que el anterior.

Con lo que respecta al tamaño de los nodos, este no afecta de manera sustancial a los resultados pero si se obtiene un mínimo para un *node size* de 1 con lo que este será el valor utilizado en el modelo óptimo aunque podría aumentarse si se quiere optimizar el coste computacional ya que nodos finales de 6 unidades no cambiarían significativamente el modelo y reducirían el tiempo de procesamiento.

(a) OOB RMSE



(b) R^2

FIGURA 5.3: OOB Error y R^2 en función de *mtry* y *nodesize*

Modelo final

El modelo final cuenta con 221 árboles, *mtry* de 33 y un tamaño de los nodos finales de mínimo 1. Con estos parámetros se comete un error cuadrático medio de

7641207 y se obtiene un R^2 de más del 87 %. Viendo estos resultados se podría decir que el modelo ajusta de manera muy satisfactoria en el conjunto de entrenamiento pero se debe probar su capacidad predictiva.

RF *mtry Trees node size seed OOB RMSE* R^2 *MAE Time*

17 500 5 123 33 221 1
123

Default 9519087 0.812 427679 6,46 **Óptimo** 7641207 0.878 278352.9 6.36

CUADRO 5.3: Parámetros y resultados RF óptimo y predeterminado

En la tabla 5.3 se puede observar como una elección de los parámetros óptima puede mejorar sustancialmente tanto el ajuste del modelo como el tiempo de compu tación necesario.

5.2.2. Capacidad predictiva

Para comprobar la capacidad predictiva del modelo en primer lugar se analizará el conjunto de datos de entrenamiento para posteriormente realizar las imputaciones y compararlas con aquellas realizadas por el INE.

El primer análisis consistirá en analizar el valor predicho con respecto al real. Para esto se han separado las unidades en grandes (CN >100 millones) y pequeñas (CN <100 millones). Como se puede observar en el gráfico 5.4 (a) en el caso de las unidades pequeñas el modelo predice muy bien exceptuando algunas unidades para las que la estimación es muy superior al valor real. Sin embargo, lo contrario ocurre con las unidades grandes como se aprecia en el gráfico 5.4 (b), que tienden a estar infraestimadas y en general a ajustar peor que las pequeñas. Se podría decir que, por un lado, ajusta peor en las unidades de mayor tamaño y por otro lado, ajusta peor en las unidades que tienen gran tamaño pero que en periodos anteriores no superaban el *threshold* de situarse en el percentil 95 (véase anexo A.3). Las empresas de mayor tamaño son aquellas correspondientes al MIG

BC (Bienes de consumo) (Véase anexo A.4) por lo que teniendo en cuenta que el modelo ajusta peor en estas unidades podría ser interesante una mayor atención a este estrato.

(a) Pequeñas (b) Grandes

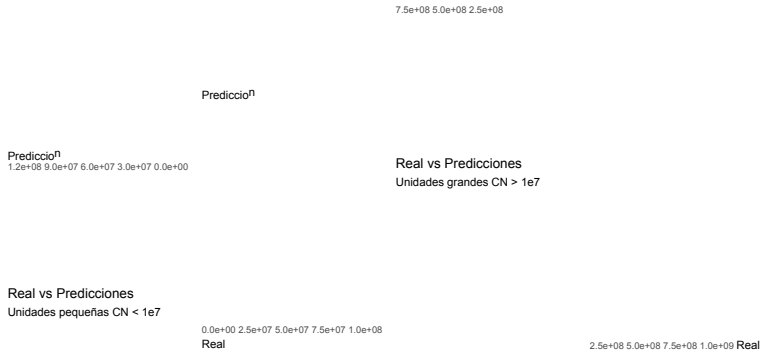
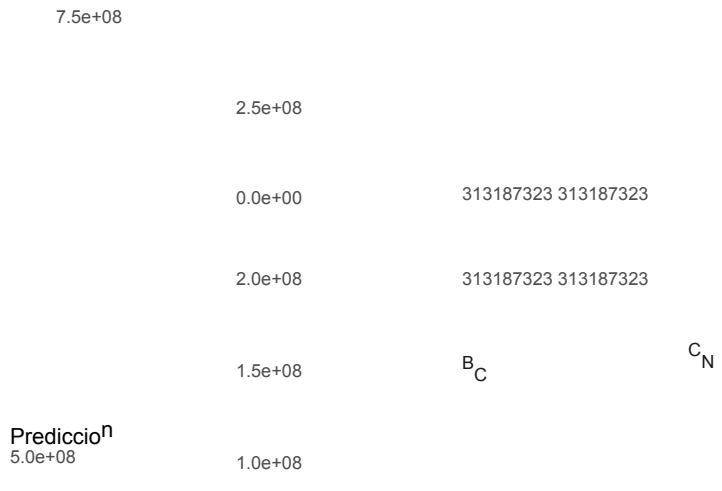


FIGURA 5.4: Ajuste de las predicciones en el conjunto de entrena miento

Realizando un análisis más profundo acerca de las unidades que más error gene ran se obtienen los resultados del gráfico 5.5. En este caso se presentan únicamente aquellas observaciones que han generado al menos un 1 % al error cuadrático total. Además se han desagregado por MIGs y los colores representan la división de esos establecimientos. En primer lugar, destaca lo ya observado anteriormente y es la gran cantidad de unidades pertenecientes al grupo de los bienes de consumo. Ade más, dentro de este grupo se observa como las unidades grandes donde se comete error son aquellas de la división 29 y las unidades pequeñas las de la división 30. Es tos estratos se corresponden a la industria automovilística y la industria aeronáutica, ferroviaria y naval respectivamente que son divisiones de enorme tamaño y peso en la cifra de negocios total y deberán tenerse en cuenta más detenidamente. Si se ob serva que el modelo falla sistemáticamente en unas divisiones o MIGs concretas se podrían crear variables que indiquen la pertenencia a estos grupos para que así el modelo pueda ajustarse mejor.

 Real vs Predicciones



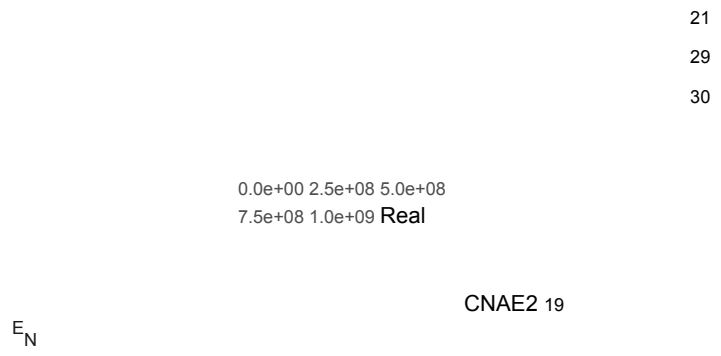


FIGURA 5.5: Diferencia entre el valor real de las unidades más influ
yentes (1 % Error)

Una vez visto que el modelo funciona de manera satisfactoria con el conjunto de test se usará para predecir los valores del conjunto donde la cifra de negocios era un valor *missing* y así realizar las imputaciones.

5.3. Imputaciones

Usando el modelo con los parámetros de la tabla 5.3 y la función *predict* del pa quete *ranger* se obtienen las imputaciones que se analizaran a continuación.

5.3.1. Error de las imputaciones

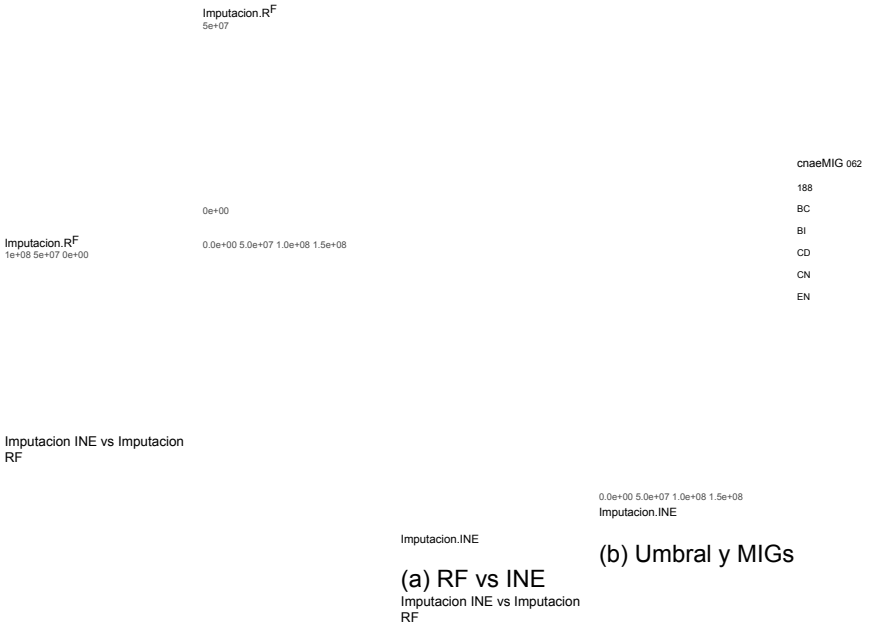
El error se define como la diferencia entre el valor real y el estimado o imputado en este caso. Una vez aplicado el modelo al conjunto de datos a imputar se utiliza

el archivo de datos FDE de noviembre de 2020 para comprobar la cercanía de las imputaciones a la real [4].

En primer lugar se calcula el error cuadrático medio (*RMSE*) de las imputaciones que es de 2,05 millones de €. Este valor es inferior al obtenido en el conjunto de en trenamiento que es de 2,77 millones de €. El error medio absoluto (*MAE*), en cambio, es superior al obtenido en el conjunto de entrenamiento con un valor de 521177 €. La distribución de los errores que se puede observar en el histograma A.5 muestra claramente como la mayoría de las imputaciones se sitúan en torno al valor 0 lo que indica el buen funcionamiento del modelo.

Si realizamos la comparación de los valores imputados por el modelo contra los

imputados por el INE que se aprecia en el gráfico 5.6 (a) podemos observar que de nuevo el modelo tiende a sobreestimar la cifra de negocios de aquellas empresas de tamaño pequeño/mediano y subestima las más grandes. Esto es lo mismo que ocurría con el conjunto de entrenamiento pero en este caso el problema se acentúa. Un análisis más preciso es el del gráfico 5.6 (b) que analiza la precisión con respecto al MIG al que pertenecen. Aquí se puede observar una sobreestimación en aquellas unidades pertenecientes al MIG de consumo no duradero (CN) y lo contrario con el grupo de bienes intermedios (BI).



(a) RF vs INE

(b) Umbral y MIGs

FIGURA 5.6: Comparación del valor imputado y el real

Si analizamos detenidamente el comportamiento de las imputaciones por divi sión/subdivisión (véase tabla A.1) se observa como las divisiones de mayor tamaño que son la 10A y 24 generan un error relativo inferior a su peso. Esto probablemente se deba a que en estos estratos existan una mayor cantidad de unidades y por lo tanto el modelo se entrene mejor. Sin embargo, en estratos más pequeños el error

[4]Nótese que se asumirá una imputación como valor correcto por el INE pero si el dato no es reen viado no se puede comprobar con certeza si este valor es real.

es muy superior al peso del estrato probablemente por el motivo contrario. Al igual que con el conjunto de entrenamiento es interesante analizar las unidades que gene ran más error. Es por esto que se han escogido aquellas que generan más de un 1 % del error total cómo podemos apreciar en el gráfico 5.7. Al igual que en el conjunto de entrenamiento encontramos dentro del MIG de bienes de consumo

una amplia mayoría de unidades pertenecientes a la división 29. Con lo que respecta al MIG de bienes intermedios (BI) la división que destaca es la 24 correspondiente a la metalur gia mientras que en el MIG de consumo no duradero encontramos una mayoría de unidades pertenecientes a la división 11 de fabricación de bebidas.

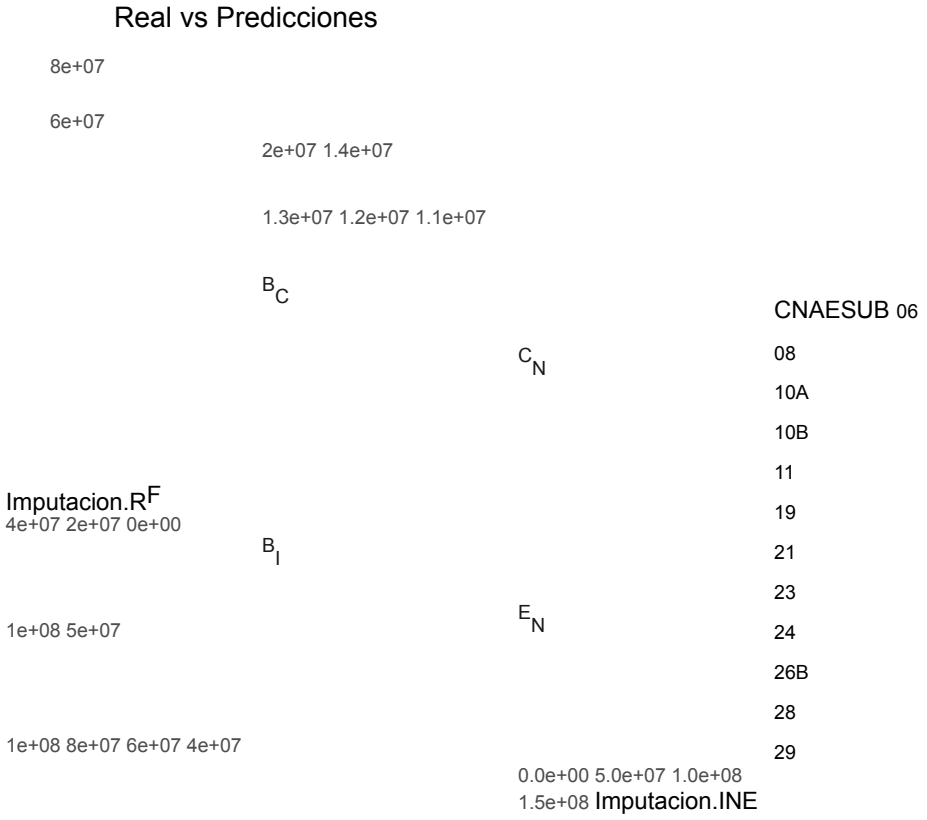


FIGURA 5.7: Diferencia entre el valor imputado de las unidades más influyentes (1 % Error)

5.3.2. Envíos

Como se ha comentado anteriormente el ICN, por el momento, recoge datos en 3 diferentes envíos con lo que el proceso llevado a cabo hasta el momento se puede re petir con 3 conjuntos de datos diferentes donde presumiblemente el modelo debería mejorar debido a la incorporación de nuevas unidades.

Al igual que para el envío 1 es crucial la selección de hiperparámetros y se ha se guido la misma estrategia que se puede observar en el anexo A.3 y A.4. El resultado final son 3 modelos resumidos en la tabla 5.4. Observando los gráficos de ajuste del envío 2 (A.3.3) y envío 3 (A.4.3) se percibe claramente una mejora de las predicciones realizadas por el modelo gracias a el mayor numero de unidades en el conjunto de entrenamiento y un menor numero de unidades a imputar.

RF *mtry num.trees node size seed OOB RMSE* R^2 *MAE Time*
33 221 1 123 38 601 2 123 45 141 4 123

Envío 1 7641207 0.878 278352.9 6.36 **Envío 2** 8011386 0.853 295373.1 22.83 **Envío 3** 7392445 0.876 326074.7 6,53

CUADRO 5.4: Parámetros y resultados RF diferentes envíos

Capítulo 6

Conclusiones y trabajo futuro

El objetivo de este trabajo de fin de máster era crear un proceso de imputación de valores *missing* para la encuesta de Índices de Cifras de Negocios de la Industria basado en aprendizaje estadístico y más concretamente en *random forests*. Antes de desarrollarla se han explicado las diferentes técnicas existentes así como los crite rios de calidad que se deben seguir en la estadística oficial para la creación de un proceso en este ámbito. De esta forma después de analizar los diferentes métodos de imputación existentes se ha explicado el funcionamiento de los RF para después aplicarlos a la encuesta de noviembre de 2020. Se ha construido el modelo haciendo uso de la librería *ranger* con la regla de división de reducción de la varianza y los pa rámetros óptimos basándose en el error *out of bag*. Los modelos finales pueden verse en 5.4 y tienen todos una capacidad predictiva muy satisfactoria.

La principal conclusión que se saca de este trabajo es la viabilidad que tienen los RF como herramienta de imputación en conjuntos de datos de estadísticas económi cas coyunturales. Además, cabe destacar la velocidad de la creación de modelos así como la versatilidad para añadir o eliminar variables sean del tipo que sean.

Más concretamente, se ha observado que el método tiende a subestimar aquellas observaciones que tienen un gran tamaño y lo contrario ocurre con las observacio nes de menor tamaño. Esto indica la necesidad de añadir variables que contemplen esto como podría ser la inclusión de un *threshold* desagregado a diferentes niveles por ejemplo. En general la inclusión de más variables podría ser una forma de en riquecer el modelo ya que los RF no se ven penalizados por la inclusión de muchos predictores.

Para la inclusión de nuevas variables, una de las ventajas de los RF es que permi ten el cálculo de la importancia de las variables (haciendo uso de las observaciones OOB) y analizando los 3 modelos de los 3 envíos diferentes todos coinciden en que la variable más importante con amplia diferencia es la cifra de

negocios del mes an terior al de referencia y después se encuentran dos variables relativas al tamaño de la empresa. Esto indica que por un lado la cifra de negocios en periodos anteriores es un factor importante a tener en cuenta pero también tiene gran importancia el ta maño de la empresa y el tamaño relativo en su estrato. Esto nos indica que variables como la cifra de negocios en periodos anteriores como un año antes del periodo de referencia o un trimestre antes podrían ser variables interesantes para posteriores análisis. Lo mismo ocurre con las tasas de variación que podrían calcularse las tasas trimestrales, bimestrales, semestrales etc... para ser de utilidad en aquellos procesos

industriales con gran estacionalidad. Otro lugar donde encontrar información que podría enriquecer el análisis son el resto de encuestas económicas coyunturales. Por ejemplo, se podrían usar variables de IPRI o IPI ya que los precios industriales y la producciones industrial tienen relación directa[1]con la cifra de negocios y además se podría incluir como regresora la tasas de variación de precios en un estrato concreto o la tasas intermensuales de la unidad en cuestión.

En cuanto a la posible extensión del trabajo por un lado se trataría de la impu tación del resto de cifras de negocio (no solo el total) y en segundo lugar el cálculo de los índices. En este trabajo, se ha realizado únicamente la imputación de la cifra de negocios total debido a la necesidad de clasificar primero las empresas por mercados antes de poder saber en qué mercado imputar la cifra de negocios con lo que para la realización de estas imputaciones sería necesario primero diseñar un clasificador (p. ej. RF) y esto sería tema de otro trabajo de fin de máster. Con lo que respecta a los índices, aunque este trabajo se ha centrado únicamente en la imputación de los datos y el análisis de los resultados en comparación con los microdatos, el INE no publica microdatos sino que tiene el deber de publicar la lista de índices actualizada men sualmente y es para esto para lo que se realizan las imputaciones. Por esta razón, el trabajo futuro consiste en calcular los índices a través de las imputaciones realiza das mediante RF. En caso de obtener índices similares al INE esto nos indicaría un buen desempeño de la metodología para el objetivo con el que se ha desarrollado. En este sentido la idea sería ser capaces de realizar la estimación del índice en cada uno de los periodos y compararla con el índice calculado en cada envío por el INE. Además, a partir de noviembre de 2021 esta encuesta pasará a realizarse mediante la plataforma IRIA de cumplimiento de cuestionarios online. Esto no solo permitirá la implementación de *edits* al cumplimentar el cuestionario consiguiendo así elimi nar todos los ítem *missing* si no que permitirá el envío de la información a servicios centrales del INE automáticamente en cuanto la unidad complete el cuestionario. Además, se contarán con más paradatos e información acerca de la recogida lo que permitirá añadir regresores al modelo. Con todo esto, el INE podría ser capaz de estimar en tiempo real o realizar un *nowcasting* de la cifra de negocios de todas las unidades que falten por enviar información y calcular los índices con la cantidad de información disponible. Además, en el momento en el que se realice el *nowcasting*, la muestra de datos estará completa durante todo el periodo de recogida y esto permi tirá acelerar la macrodepuración. Como es lógico, cuando existan pocas unidades, las imputaciones serán malas pero la ventaja es que se podrá analizar lo bien que funciona y la convergencia hacia el índice real.

[1]Aproximadamente *ICN* = *IPRI* × *IPI*

Bibliografía

Acock, Alan C. (nov. de 2005). «Working with missing values». En: *Journal of Marriage and Family* 67.4, págs. 1012-1028. ISSN: 00222445. DOI: 10.1111/j.1741- 3737. 2005.00191.x. URL: http://doi.wiley.com/10.1111/j.1741-3737.2005. 00191.x.

Alsagheer, Radhwan HA, Abbas FH Alharan y Ali SA Al-Haboobi (2017). «Popular decision tree algorithms of data mining techniques: a review». En: *International Journal of Computer Science and Mobile Computing* 6.6, págs. 133-142.

Andridge, Rebecca R. y Roderick J. A. Little (abr. de 2010). «A Review of Hot Deck Imputation for Survey Non-response». En: *International Statistical Review* 78.1, págs. 40-64. ISSN: 03067734. DOI: 10.1111/j.1751- 5823.2010.00103.x. URL: http://doi.wiley.com/10.1111/j.1751-5823.2010.00103.x.

Batista, Gustavo EAPA, Maria Carolina Monard y col. (2002). «A study of K-nearest neighbour as an imputation method.» En: *His* 87.251-260, pág. 48.

Beck, Martin, Florian Dumpert y Joerg Feuerhake (dic. de 2018). «Machine Learning in Official Statistics». En: arXiv: 1812 . 10422. URL: http : / / arxiv . org / abs / 1812.10422.

Bennett, Derrick A. (oct. de 2001). «How can I deal with missing data in my study?» En: *Australian and New Zealand Journal of Public Health* 25.5, págs. 464-469. ISSN: 13260200. DOI: 10.1111/j.1467- 842X.2001.tb00294.x. URL: http://doi. wiley.com/10.1111/j.1467-842X.2001.tb00294.x.

BOE (2001). *Real Decreto 508/2001, de 11 de mayo, por el que se aprueba el Estatuto del Instituto Nacional de Estadística.* https://www.boe.es/eli/es/rd/2001/05/11/508/con.

— (2020). *Boletin oficial del estado: Ley 11/2020, de 30 de diciembre, de Presupuestos Ge nerales del Estado para el año 2021.* https://boe.es/boe/dias/2020/12/31/pdfs/BOE-A-2020-17339.pdf. Bou-Hamad, Imad, Denis Larocque, Hatem Ben-Ameur y col. (2011). «A review of survival trees». En: *Statistics surveys* 5, págs. 44-71.

Brackstone, Gordon J (2002). *How important is accuracy?* Citeseer.

Breiman, Leo (1996). «Bagging predictors». En: *Machine learning* 24.2, págs. 123-140. — (2001). «Random forests». En: *Machine learning* 45.1, págs. 5-32.

— (2015). «Random forests leo breiman and adele cutler». En: *Random*

Forests-Classification Description. URL: https://www.stat.berkeley.edu/~breiman/RandomForests/ cc_home.htm.

Breiman, Leo y col. (1984). *Classification and regression trees*. CRC press.

Comisión Europea (1998). *Reglamento (CE) NO 1165/98 del Consejo*. https : / / eur - lex . europa . eu / legal - content / ES / TXT / PDF / ?uri = CELEX : 31998R1165&from=ES.

— (2006). *Reglamento (CE) n o 1893/2006 del Parlamento Europeo y del Consejo*. https : / / eur - lex . europa . eu / legal - content / ES / TXT / PDF / ?uri = CELEX : 02006R1893-20190726&from=ES.

Costa, Alex, Jaume Garcıá y Josep Lluis Raymond (sep. de 2014). «Are All Qua lity Dimensions of Equal Importance when Measuring the Perceived Quality of Official Statistics? Evidence from Spain». En: *Journal of Official Statistics* 30.3, págs. 547-562. ISSN: 2001-7367. DOI: 10 . 2478 / jos - 2014 - 0034. URL: https : //www.sciendo.com/article/10.2478/jos-2014-0034.

Cutler, Adele, D. Richard Cutler y John R. Stevens (2012). «Random Forests». En: *Ensemble Machine Learning*. Boston, MA: Springer US, págs. 157-175. DOI: 10 . 1007/978-1-4419-9326-7_5. URL: http://link.springer.com/10.1007/978-1-4419-9326-7_5.

De Waal, Ton, Jorden Pannekoek y Sander Scholtus (2007). *Statistical data editing and imputation*. Vol. 29. 29, pág. 51. ISBN: 9780470542804.

Donders, A. Rogier T. y col. (2006). «Review: A gentle introduction to imputation of missing values». En: *Journal of Clinical Epidemiology* 59.10, págs. 1087-1091. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2006.01.014.

Dowle, Matt y col. (2019). «Package 'data. table'». En: *Extension of 'data. frame*.

Drucker, Harris y Corinna Cortes (1996). «Boosting decision trees». En: *Advances in neural information processing systems*, págs. 479-485.

Elvers, Eva y Håkan Lindén (sep. de 2015). «Quality Concept for Official Statistics». En: *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: John Wiley y Sons, Ltd, págs. 1-13. DOI: 10 . 1002 / 9781118445112 . stat03101 . pub2. URL: http : //doi.wiley.com/10.1002/9781118445112.stat03101.pub2.

Eurostat (2017). *European Statistics Code of Practice*. Eurostat, Luxembourg. DOI: 10. 2785/798269. URL: https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani y col. (2001). *The elements of sta tistical learning*. Vol. 1. 10. Springer series in statistics New York. Grazzini, J (2021). «Statistics Coded – Storytelling through literate programming and runnable computing». En:

Groves, R. M. y L. Lyberg (ene. de 2010). «Total Survey Error: Past, Present, and Future». En: *Public Opinion Quarterly* 74.5, págs. 849-879. ISSN: 0033-362X. DOI: 10.1093/poq/nfq065. URL: https://academic.oup.com/poq/article-lookup/doi/10.1093/poq/nfq065.

Grudkowska, Sylwia y col. (2013). «Advanced Tools for Time Series Analysis and Seasonal Adjustment in the New JDemetra+». En: *JSM Proceedings Paper*.

Hssina, Badr y col. (2014). «A comparative study of decision tree ID3 and C4. 5». En: *International Journal of Advanced Computer Science and Applications* 4.2, págs. 13-19. INE (mar. de 2015). *Política de revisión del Instituto Nacional de Estadística*. — (mar. de 2020). *Encuesta de satisfacción de los usuarios de*

James, Gareth y col. (2013). *An introduction to statistical learning*. Vol. 112. Springer. Janitza, Silke, Harald Binder y Anne-Laure Boulesteix (2016). «Pitfalls of hypothe sis tests and model selection on bootstrap samples: causes and consequences in biometrical applications». En: *Biometrical Journal* 58.3, págs. 447-473. Kim, Jae Kwang (2001). «Variance estimation after imputation». En: 27.1, pág. 173. URL: http://projecteuclid.org/euclid.aos/1083178946.

Kowarik, Alexander y Matthias Templ (2016). «Imputation with the R Package VIM». En: *Journal of Statistical Software* 74.7, págs. 1-16.

Kuhn, Max, Kjell Johnson y col. (2013). *Applied predictive modeling*. Vol. 26. Springer. Lewis, Roger J (2000). «An introduction to classification and regression tree (CART) analysis». En: *Annual meeting of the society for academic emergency medicine in San Francisco, California*. Vol. 14.

Liaw, Andy, Matthew Wiener y col. (2002). «Classification and regression by ran domForest». En: *R news* 2.3, págs. 18-22.

Little, Roderick J. A. y Donald B. Rubin (ago. de 2002). *Statistical Analysis with Mis sing Data*. Hoboken, NJ, USA: John Wiley y Sons, Inc. ISBN: 9781119013563. DOI: 10.1002/9781119013563. URL: http://doi.wiley.com/10.1002/9781119013563.

MacFeely, Steve (dic. de 2016). «The Continuing Evolution of Official Statistics: Some Challenges and Opportunities». En: *Journal of Official Statistics* 32.4, págs. 789-810. ISSN: 2001-7367. DOI: 10.1515/jos- 2016- 0041. URL: https://www.sciendo. com/article/10.1515/jos-2016-0041.

Manski, Charles (mayo de 2014). *Communicating Uncertainty in Official Economic Sta tistics*. Inf. téc. Cambridge, MA: National Bureau of Economic Research. DOI: 10.3386/w20098. URL: http://www.nber.org/papers/w20098.pdf.

Martínez-Muñoz, Gonzalo y Alberto Suárez (2010). «Out-of-bag estimation of the optimal sample size in bagging». En: *Pattern Recognition* 43.1, págs. 143-152.

McKenzie, Richard y Michela Gamba (2008). «Interpreting the results of Revision Analyses: Recommended Summary Statistics». En: *Contribution to OECD/Euros tat Task Force on "Performing Revisions Analysis for Sub-Annual Economic Statistics*. URL: https://www.oecd.org/sdd/40315546.pdf.

Merkle, Edgar C. y Victoria A. Shaffer (2011). «Binary recursive partitioning: Back ground, methods, and application to psychology». En: *British Journal of Mathe matical and Statistical Psychology* 64.1, págs. 161-181. DOI: https : / / doi . org / 10 . 1348 / 000711010X503129. eprint: https : / / bpspsychub . onlinelibrary . wiley.com/doi/pdf/10.1348/000711010X503129. URL: https://bpspsychub. onlinelibrary.wiley.com/doi/abs/10.1348/000711010X503129.

Mitchell, Tom M (1997). *Machine learning*. McGraw-hill New York. Nembrini, Stefano, Inke R König y Marvin N Wright (mayo de 2018). «The revival of the Gini importance?» En: *Bioinformatics* 34.21, págs. 3711-3718. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty373. eprint: https://academic.oup.com/ bioinformatics/article-pdf/34/21/3711/26146979/bty373_supplement\ _nembrini.pdf. URL: https://doi.org/10.1093/bioinformatics/bty373. OECD (2011). *Quality dimensions, core values for oecd statistics and procedures for plan ning and evaluating statistical activities*.

Probst, Philipp y Anne-Laure Boulesteix (2017). «To Tune or Not to Tune the Number of Trees in Random Forest.» En: *J. Mach. Learn. Res.* 18.1, págs. 6673-6690. Probst, Philipp, Marvin N Wright y Anne-Laure Boulesteix (2019). «Hyperparame ters and tuning strategies for random forest». En: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3, e1301.

Särndal, Carl-Erik y Sixten Lundström (ene. de 2005). *Estimation in Surveys with Non response*. Chichester, UK: John Wiley y Sons, Ltd. ISBN: 9780470011355. DOI: 10. 1002/0470011351. URL: http://doi.wiley.com/10.1002/0470011351.

Scholtus, Sander, Rob van de Laar y Leon Willenborg (2014). *The memobust handbook on methodology for modern business statistics (MEMOBUST Handbook)*. Scornet, Erwan (2017). «Tuning parameters in random forests». En: *ESAIM: Procee dings and Surveys* 60, págs. 144-162.

Steinberg, Dan (2009). «CART: classification and regression trees». En: *The top ten algorithms in data mining*. Chapman y Hall/CRC, págs. 193-216. Team R, Core (2000). «R language definition». En: *Vienna, Austria: R foundation for statistical computing*.

Unión Europea (2020). *Reglamento de ejecución (UE) 2020/1197 de la Comisión*. https : / / eur - lex . europa . eu / legal - content / ES / TXT / PDF / ?uri = CELEX : 32020R1197&from=EN.

Van Der Loo, Mark, Edwin De Jonge y Sander Scholtus (2011). *Correction of rounding, typing, and sign errors with the deducorrect package*. Citeseer.

West, Brady T. (ago. de 2011). «Paradata in Survey Research». En: *Survey Practice* 4.4, págs. 1-8. ISSN: 2168-0094. DOI: 10.29115/SP- 2011- 0018. URL: https:// surveypractice.scholasticahq.com/article/3036- paradata- in- survey research.

Wickham, Hadley (2007). *The ggplot package*.

Wright, Marvin N. y Andreas Ziegler (2017). «ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R». En: *Journal of Statis tical Software* 77.1. ISSN: 1548-7660. DOI: 10.18637/jss.v077.i01. URL: http: //www.jstatsoft.org/v77/i01/.

Zhang, Li-Chun (2012). «Topics of statistical theory for register-based statistics and data integration». En: *Statistica Neerlandica* 66.1, págs. 41-63.

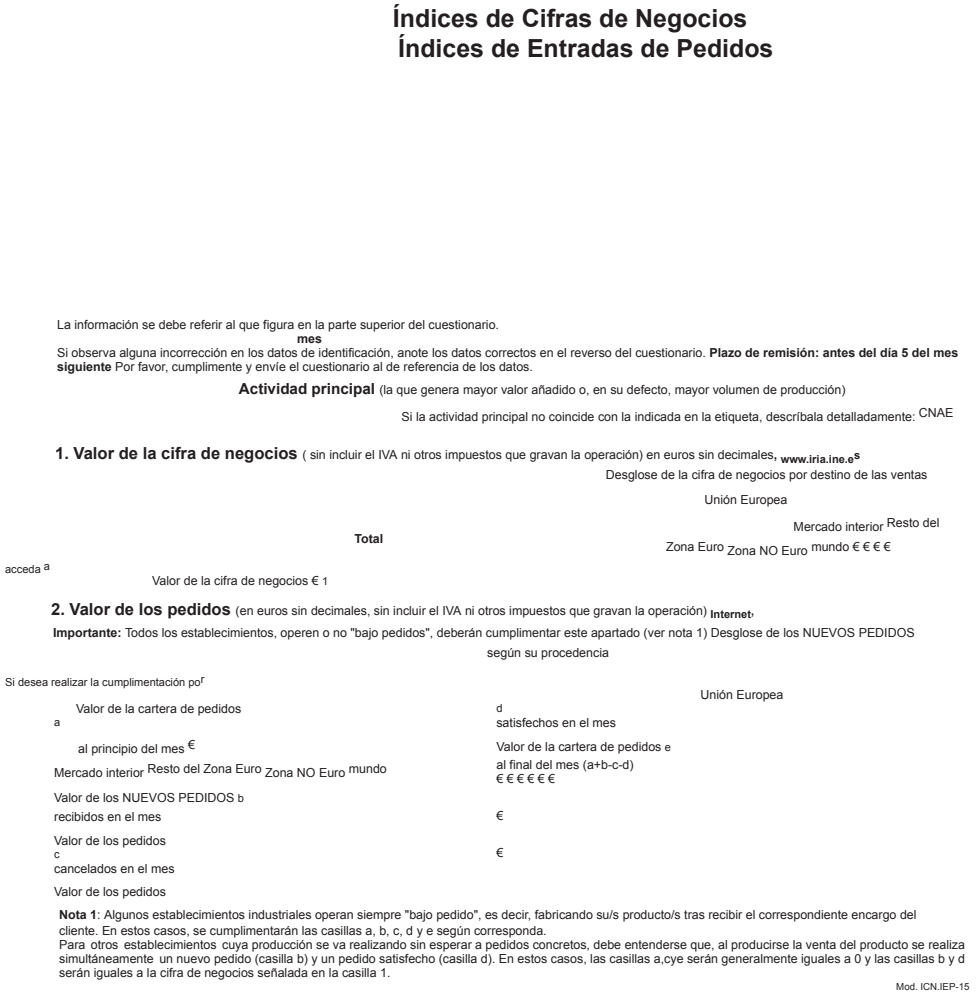
Apéndice A

Anexo

A.1. Encuesta ICN

Este cuestionario es el enviado a todas las unidades en la muestra de ICN y que debe ser cumplimentado. Aquellos cuestionarios no recibidos son los

susceptibles de imputación.



Índices de Cifras de Negocios
Índices de Entradas de Pedidos

La información se debe referir al que figura en la parte superior del cuestionario.
mes
Si observa alguna incorrección en los datos de identificación, anote los datos correctos en el reverso del cuestionario. **Plazo de remisión: antes del día 5 del mes siguiente** Por favor, cumplimente y envíe el cuestionario al de referencia de los datos.

Actividad principal (la que genera mayor valor añadido o, en su defecto, mayor volumen de producción)

Si la actividad principal no coincide con la indicada en la etiqueta, descríbala detalladamente: CNAE

1. Valor de la cifra de negocios (sin incluir el IVA ni otros impuestos que gravan la operación) en euros sin decimales, **www.iria.ine.es**
Desglose de la cifra de negocios por destino de las ventas

	Total	Unión Europea		Mercado interior	Resto del mundo
		Zona Euro	Zona NO Euro		
acceda a					
Valor de la cifra de negocios € 1					

2. Valor de los pedidos (en euros sin decimales, sin incluir el IVA ni otros impuestos que gravan la operación) **Internet,**
Importante: Todos los establecimientos, operen o no "bajo pedidos", deberán cumplimentar este apartado (ver nota 1) Desglose de los NUEVOS PEDIDOS según su procedencia

Si desea realizar la cumplimentación por

		Unión Europea			
	Valor de la cartera de pedidos a al principio del mes €	d satisfechos en el mes	Valor de la cartera de pedidos e al final del mes (a+b-c-d)		
		Mercado interior	Resto del Zona Euro	Zona NO Euro	mundo
Valor de los NUEVOS PEDIDOS b recibidos en el mes		€	€	€	€
Valor de los pedidos c cancelados en el mes		€	€	€	€
Valor de los pedidos					

Nota 1: Algunos establecimientos industriales operan siempre "bajo pedido", es decir, fabricando su/s producto/s tras recibir el correspondiente encargo del cliente. En estos casos, se cumplimentarán las casillas a, b, c, d y e según corresponda.
Para otros establecimientos cuya producción se va realizando sin esperar a pedidos concretos, debe entenderse que, al producirse la venta del producto se realiza simultáneamente un nuevo pedido (casilla b) y un pedido satisfecho (casilla d). En estos casos, las casillas a,cye serán generalmente iguales a 0 y las casillas b y d serán iguales a la cifra de negocios señalada en la casilla 1.

Mod. ICN.IEP-15

FIGURA A.1: Encuesta ICN base 2015. Fuente *INE*

A.2. Envío 1

A.2.1. Variables menos importantes

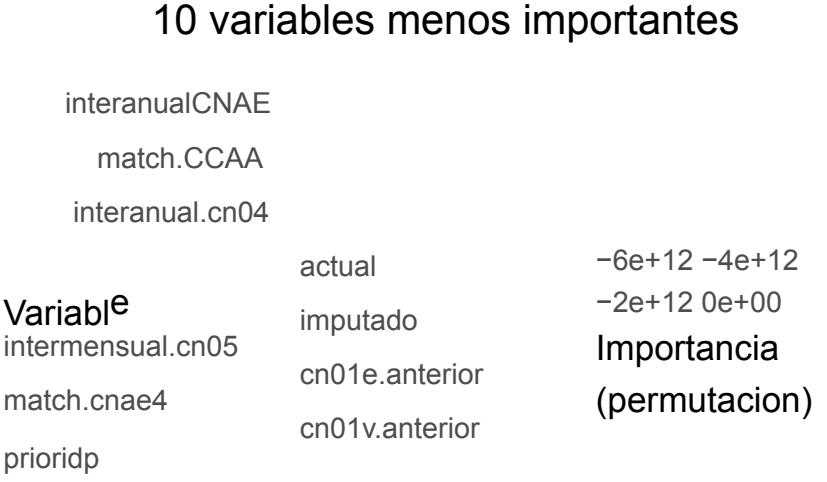


FIGURA A.2: Variables menos importantes usando método permutación

A.2.2. Tabla de frecuencias de la muestra según MIGs y división/subdivisión

Estas tablas muestran las frecuencias de cada uno de los grupos a los que se hará alusión en los análisis para ver la importancia relativa de cada una de las categorías.

Grupo MIG	Unidades	Porcentaje			
BI	5427	45.36	CN	3620	30.26
BC	2293	19.17	CD	590	4.93
EN	31	0.26			

Grupo División/Subdivisión	Unidades	Porcentaje			
10A	1621	13.55	23	1056	8.83
25B	889	7.43	28	719	6.01
22	622	5.20	20B	528	4.41
25A	476	3.98	08	432	3.61
33	428	3.58	31	402	3.36
162	366	3.06	11	357	2.98
16	340	2.84	17	337	2.82
24	318	2.66	29	302	2.52
15	290	2.42	14	287	2.40
10B	278	2.32	27B	274	2.29
13B	263	2.20	13A	175	1.46
32B	160	1.34	21	148	1.24
20A	147	1.23	32C	129	1.08
26B	127	1.06	30A	112	0.94
161	94	0.79	32A	87	0.73
27A	51	0.43	26A	47	0.39
30B	33	0.28	19	21	0.18
26C	17	0.14	07	11	0.09
06	9	0.08	12	7	0.06
05	1	0.01	062	1	0.01
188	1	0.01			

A.2.3. Gráfico *Scatter* Umbral

Este gráfico muestra el ajuste del modelo teniendo en cuenta aquellas unidades que están por encima del percentil 95 en su estrato (*TRUE*) y las que no (*FALSE*)

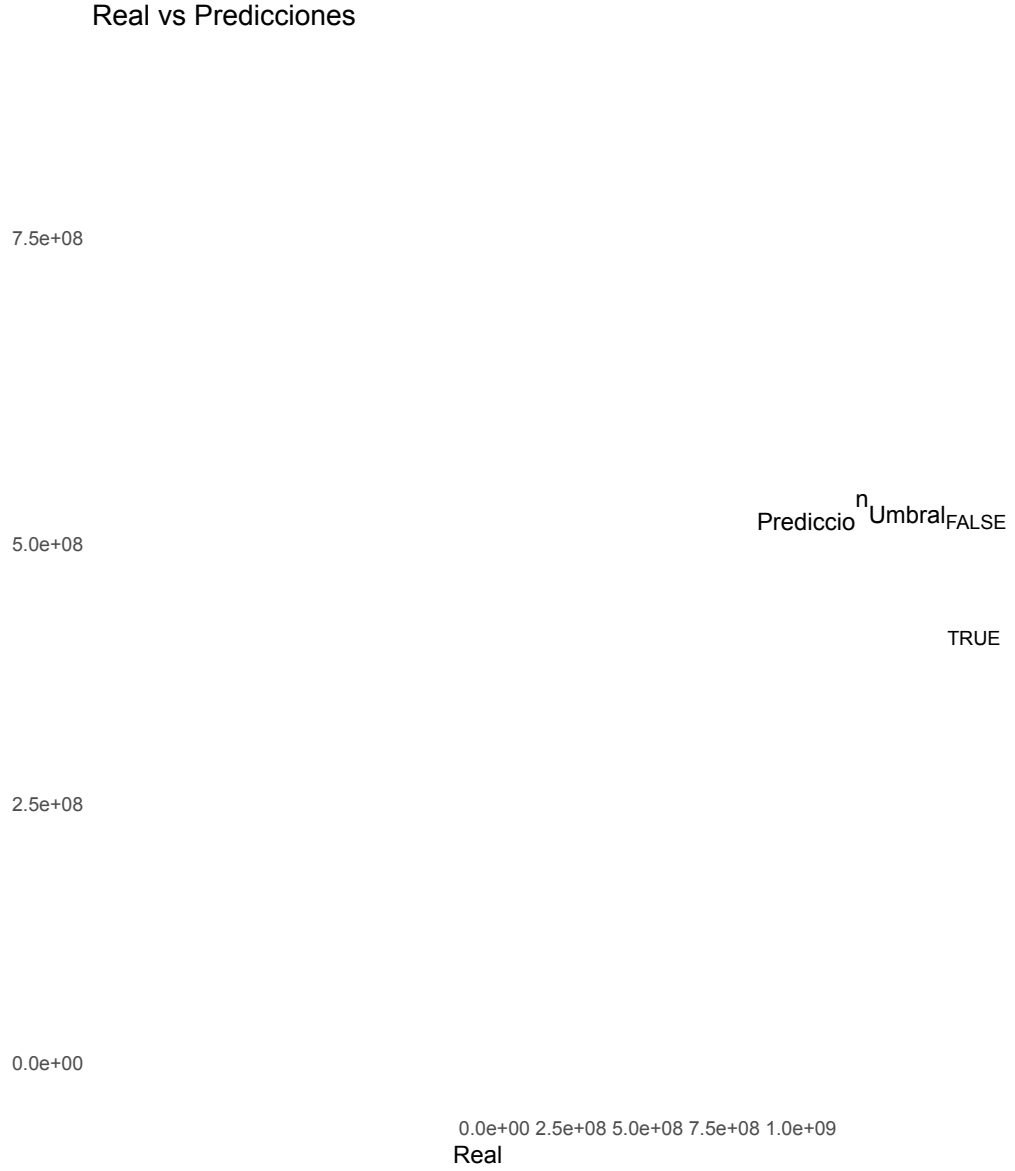
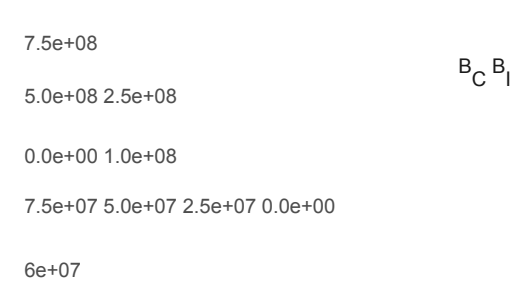


FIGURA A.3: *Scatterplot* del ajuste del modelo por *threshold*

A.2.4. *Scatterplot* MIGs y *threshold*

Este gráfico muestra el ajuste del modelo en el conjunto de datos de entrena miento dividido por los principales grupos industriales (MIGs)



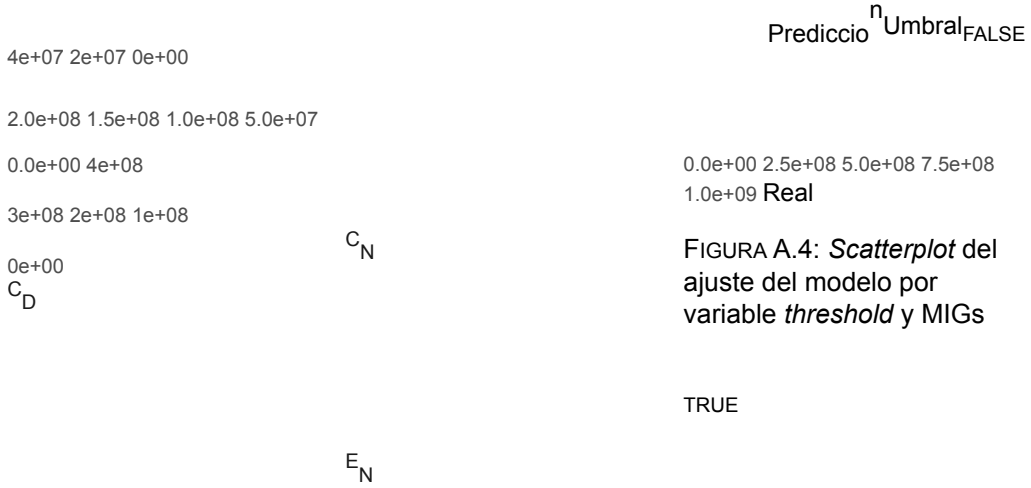


FIGURA A.4: *Scatterplot* del ajuste del modelo por variable *threshold* y MIGs

A.2.5. Importancia relativa de las subdivisiones y error (ENVÍO 1)

Error generado en las imputaciones del modelo con respecto a las imputaciones del INE en términos relativos y por División-Subdivisión CNAE-2009 en el envío 1. Estas tablas buscan ilustrar como el error generado por las subdivisiones grandes es inferior, en términos relativos, al error generado por las pequeñas.

División/Subdivisión	*RF INE*	*Error Absoluto Relativo*

10A	23.173	18.187
24	22.559	11.492
11	12.424	8.452
28	14.497	7.569
29	5.703 4.828	7.412
33	5.193 5.628	4.346
25B	5.481 4.845	4.110
20B	2.367 2.281	4.070
23	4.343 4.371	3.606
21	5.479 5.817	3.557
25A	5.447 5.383	3.485
10B	3.350 3.108	2.556
08	1.864 1.711	2.196
26B	4.861 5.307	2.097
22	0.744 0.361	1.712
17	0.905 0.533	1.664
30A	3.314 3.533	1.653
14	3.728 3.993	1.325
27B	1.928 2.047	1.298
31	0.549 0.471	1.220
06	1.597 1.768	1.022
15	1.157 1.128	0.985
16	0.214 0.000	0.942
19	0.452 0.376	0.864
162	0.873 0.888	0.776
05	0.209 0.030	0.578
13B	0.875 0.923	0.508
20A	0.140 0.020	0.363
32B	0.657 0.679	0.339
13A	0.457 0.416	0.337
32C	0.463 0.497	0.221
27A	0.470 0.439	0.218
30B	0.191 0.159	0.211
161	0.352 0.350	0.196
32A	0.157 0.128	0.134
12	0.367 0.409	0.110
188	0.074 0.069	0.101
26A	0.123 0.151	0.044
07	0.156 0.141	0.027
26C	0.110 0.111	0.013
062	0.007 0.001	0.004
	0.044 0.043	
	0.002 0.001	

CUADRO A.1: Error relativo en % de las imputaciones en compara ción al peso relativo por División/Subdivisión (Envío 1)

 A.2.6. Histograma de los errores de las imputaciones

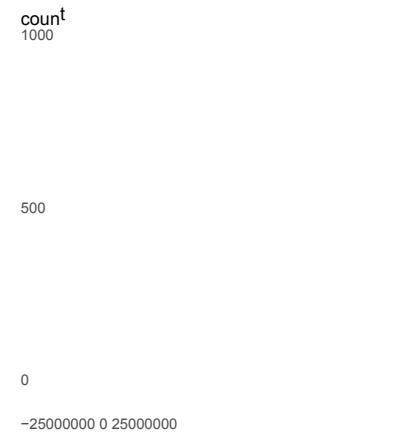


FIGURA A.5: Histograma del error de las imputaciones

A.3. Envío 2

En este apartado se presentan los resultados del RF correspondiente al envío 2 con parámetros disponibles en la tabla 5.4

A.3.1. Hiperparámetros del bosque 2

Evolución del error de entrenamiento y el error OOB en función del número de arboles. Número óptimo **601**

Apéndice A. Anexo 61 Evolucion del error vs numero de arboles

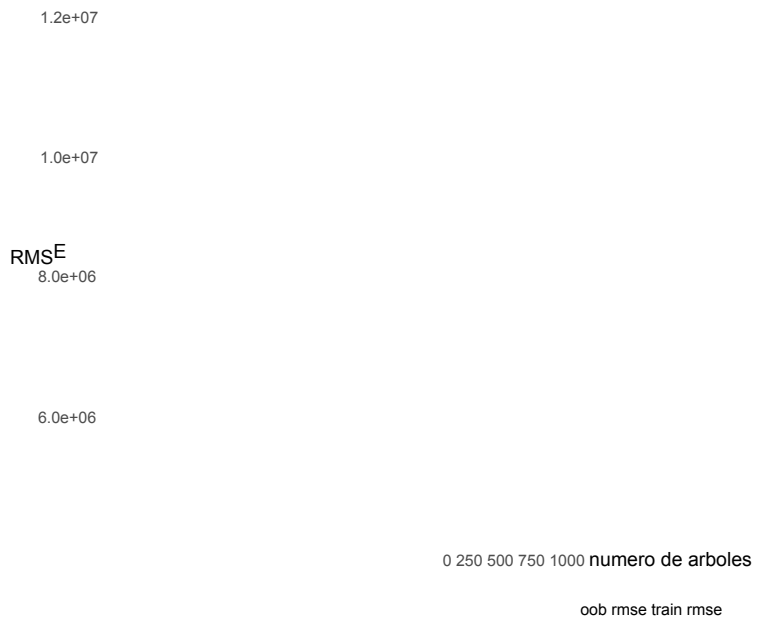


FIGURA A.6: Error en función del número de arboles

Análisis del ajuste del modelo en función de diferentes hiperparámetros :

Error OOB vs Mtry by Node.size

601 Arboles

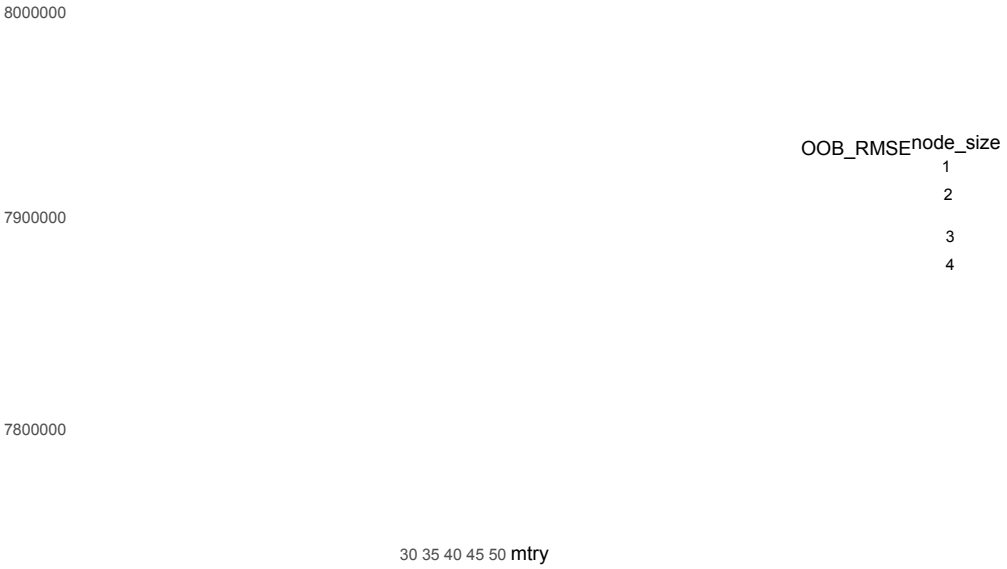


FIGURA A.7: Error RMSE OOB en función de *mtry*

FIGURA A.8: R^2 en función

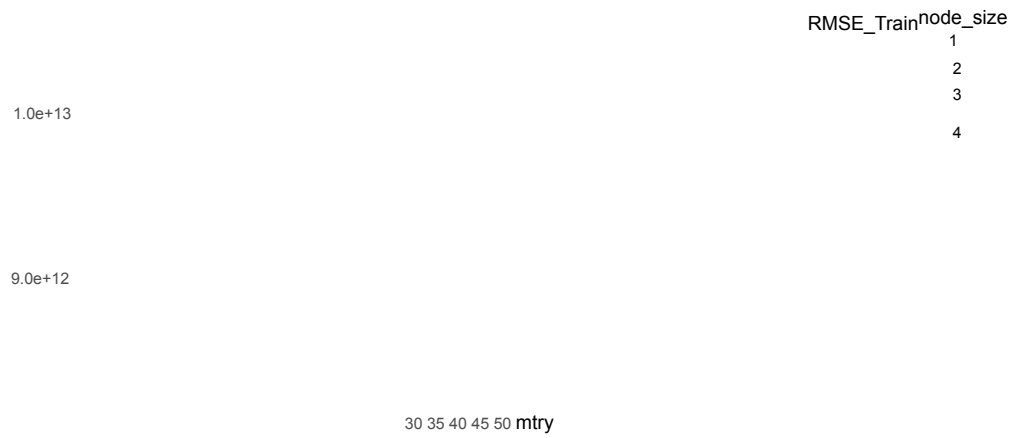
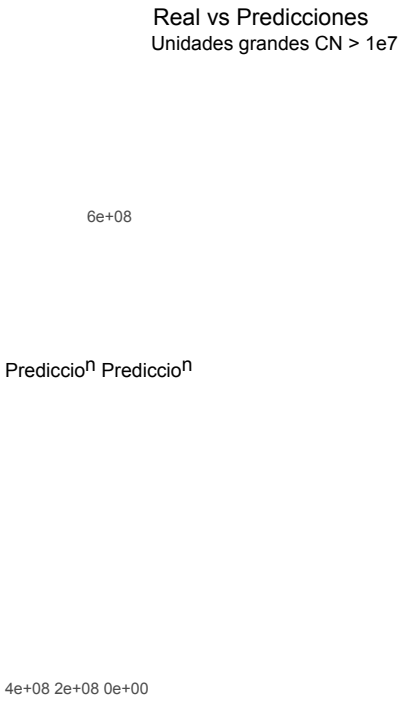


FIGURA A.9: Error RMSE de entrenamiento en función de *mtry*

A.3.2. Predicciones en el conjunto de entrenamiento



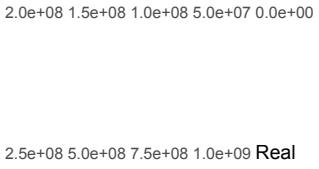


FIGURA A.10: *Scatterplot* de las predicciones y los valores reales (Uni dades grandes)

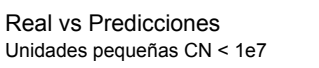


FIGURA A.11: *Scatterplot* de las predicciones y los valores reales (Uni dades pequeñas)

 Real vs Predicciones

FIGURA A.12: *Scatterplot* de las predicciones y los valores reales por *threshold*

A.3.3. Imputaciones RF (ENVÍO 2)

En este apartado se mostraran los diferentes gráficos y tablas usados para evaluar la calidad de las imputaciones.

Importancia relativa de las subdivisiones y error

Error generado en las imputaciones del modelo con respecto a las imputaciones del INE en términos relativos y por División-Subdivisión CNAE-2009 en el envío 2.

División/Subdivisión	*RF INE*	*Error Absoluto Relativo*
10A	22.26	20.31
24	20.73	16.64
28	21.46	8.04
29	24.02	7.91
11	4.81 5.42	6.83
25B	7.40 6.74	4.38
20B	4.82 4.43	3.91
21	4.34 4.05	3.17
08	5.07 5.44	2.88
23	2.72 2.24	2.80
25A	0.78 0.24	2.56
33	3.46 3.32	2.40
22	1.73 1.83	2.13
14	1.28 1.35	1.98
17	3.87	1.63
06	4.09	1.57
27B	0.69	1.46
10B	0.67	1.42
16	4.36	1.32
30A	4.68	1.16
162	0.33	0.83
31	0.00	0.68
32B	1.49	0.56
15	1.59	0.46
05	2.47	0.44
13B	2.61	0.40
13A	0.53	0.31
26B	0.54	0.26
20A	0.43	0.25
32C	0.62	0.23
30B	0.90	0.22
27A	0.92	0.22
188	0.89	0.18
161	0.90	0.17
32A	0.60	0.16
26A	0.58	0.06
07	0.20	0.04
062	0.12	0.03
26C	0.13	0.02
	0.05	
	0.49	
	0.45	
	0.31	
	0.31	
	0.16	
	0.15	
	0.26	

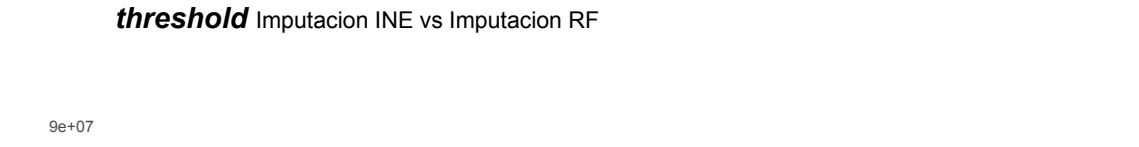
	0.23 0.18 0.14 0.19 0.17 0.29 0.28 0.34 0.32 0.52 0.54 0.09 0.08 0.07 0.08 0.01 0.00 0.01 0.00 0.06 0.06	

CUADRO A.2: Error relativo en % de las imputaciones en comparación al peso relativo por División/Subdivisión (Envío 2)

Imputación RF en comparación con la imputación del INE

FIGURA A.13: Precisión de las imputaciones

Imputación RF en comparación con la imputación del INE por *threshold*



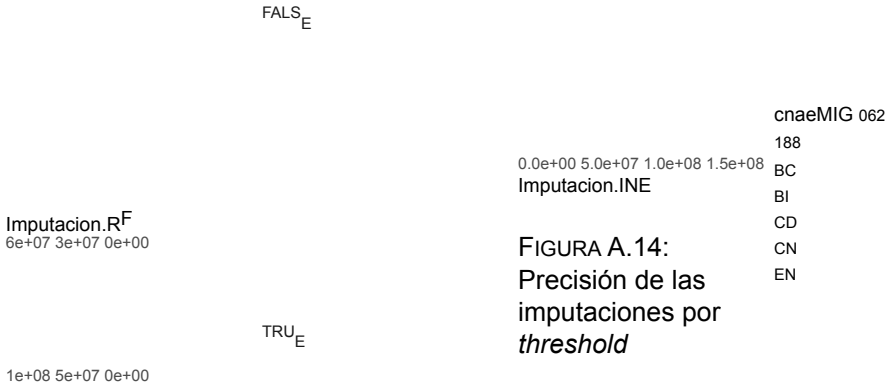


FIGURA A.14: Precisión de las imputaciones por *threshold*

Imputación RF en comparación con la imputación del INE por MIGs

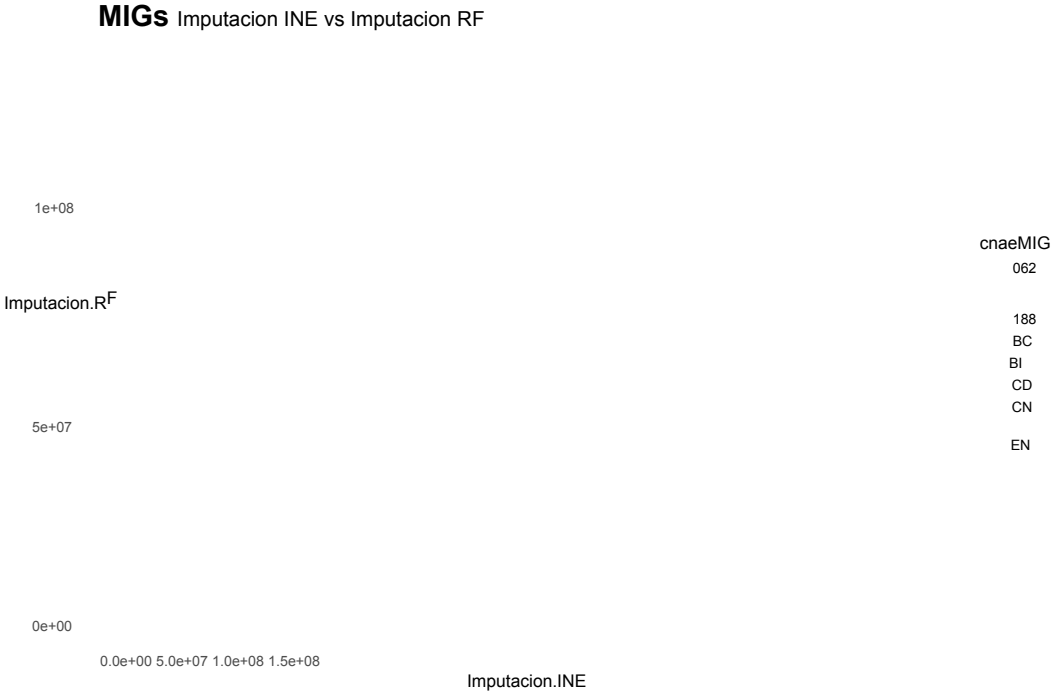
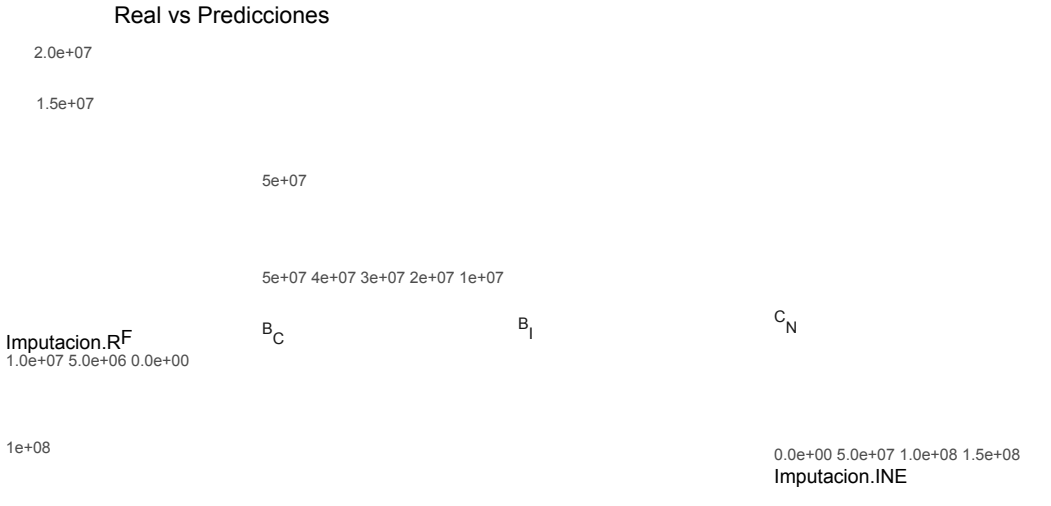


FIGURA A.15: Precisión de las imputaciones por MIGs

Imputación RF en comparación con la imputación del INE en unidades erróneas

Comparación del valor imputado por el modelo y el INE de las unidades que generaron mas de un 1 % del error total dividido por MIGs y División/Subdivisión.



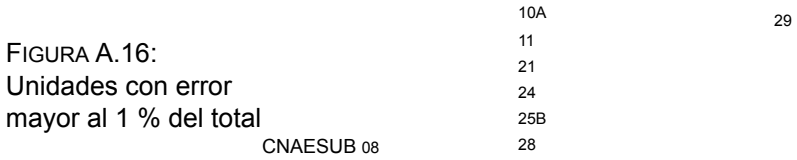


FIGURA A.16:
Unidades con error
mayor al 1 % del total

A.4. Envío 3

En este apartado se presentan los resultados del RF correspondiente al envío 3 con parámetros disponibles en la tabla 5.4

A.4.1. Hiperparámetros del bosque 3

Evolución del error de entrenamiento y el error OOB en función del número de arboles. Número óptimo **141**

FIGURA A.17: Error en función del número de arboles

Análisis del ajuste del modelo en función de diferentes hiperparámetros:

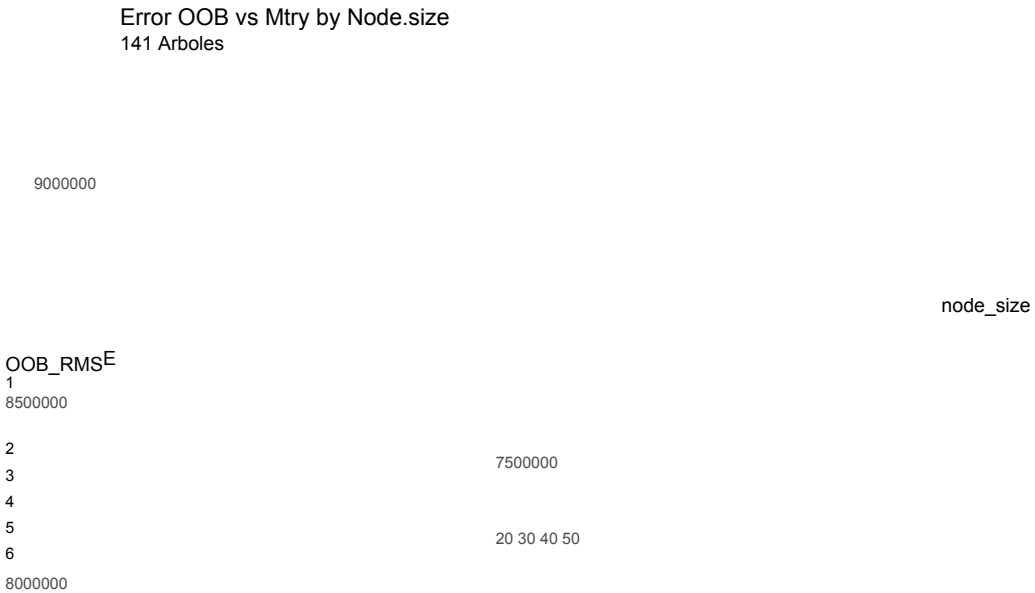


FIGURA A.18: Error RMSE OOB en función de *mtry*

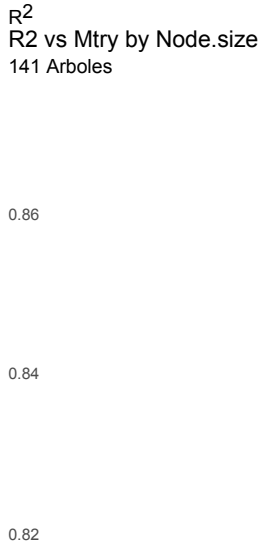


FIGURA A.19: R^2 en función de *mtry*

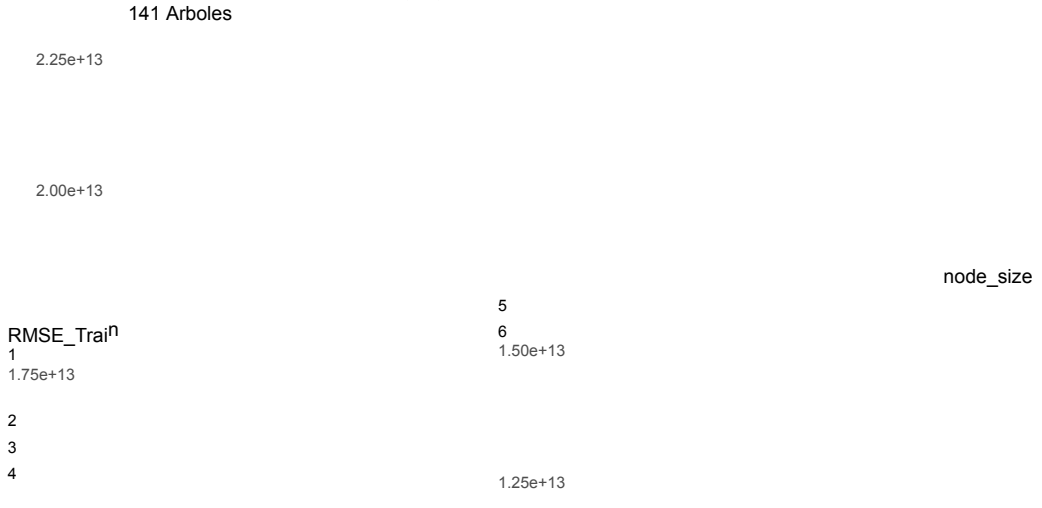


FIGURA A.20: Error RMSE de entrenamiento en función de *mtry*

mtry

A.4.2. Predicciones en el conjunto de entrenamiento

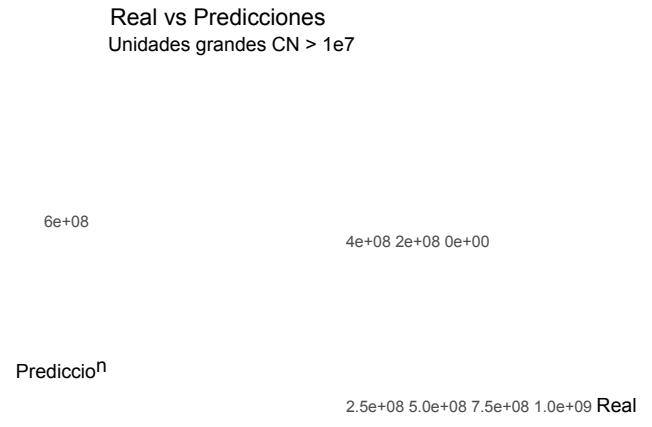


FIGURA A.21: *Scatterplot* de las predicciones y los valores reales (Uni dades grandes)

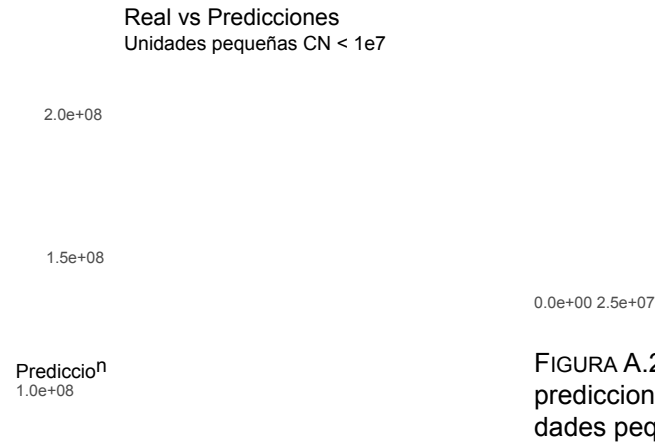


FIGURA A.22: *Scatterplot* de las predicciones y los valores reales (Uni dades pequeñas)

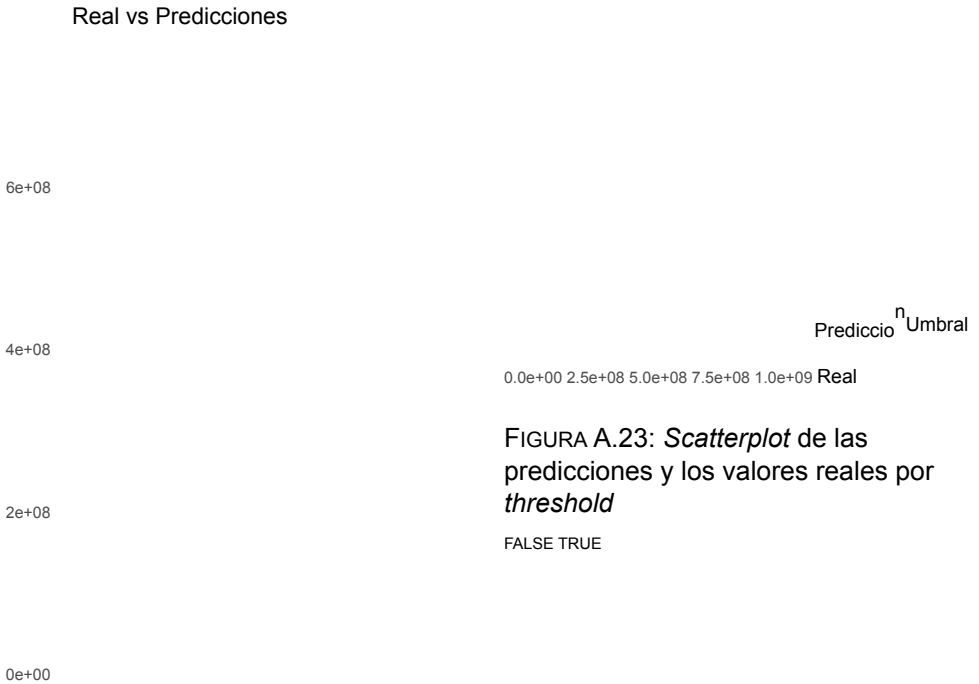


FIGURA A.23: *Scatterplot* de las predicciones y los valores reales por *threshold*

A.4.3. Imputaciones RF (ENVÍO 3)

En este apartado se mostraran los diferentes gráficos y tablas usados para evaluar la calidad de las imputaciones.

Importancia relativa de las subdivisiones y error

Error generado en las imputaciones del modelo con respecto a las imputaciones del INE en términos relativos y por División-Subdivisión CNAE-2009 en el envío 3.

División/Subdivisión	*RF INE*	*Error Absoluto Relativo*

10A	17.55	15.58
24	15.33	14.01
28	13.65	13.99
20B	16.00	5.92
29	7.14 7.88	4.87
11	7.96 8.20	4.77
08	6.56 6.63	4.12
25A	4.05 4.63	3.63
17	1.04 0.09	3.46
23	2.62 2.63	3.25
25B	9.25 9.58	3.01
16	4.43 4.13	2.69
14	4.20 4.26	2.57
30A	0.89 0.73	2.47
22	0.78 0.99	2.37
06	1.06 1.40	1.84
27B	3.67 3.68	1.56
33	0.43 0.00	1.53
32B	1.05 1.40	1.18
10B	1.53 1.28	1.07
15	1.74 1.64	0.88
13B	2.87 2.75	0.84
31	0.39 0.21	0.71
162	0.94 0.75	0.61
27A	1.32 1.29	0.60
21	0.86 0.80	0.58
13A	0.70 0.63	0.40
32C	0.63 0.61	0.30
30B	0.27 0.32	0.27
26B	0.23 0.17	0.26
05	0.33 0.28	0.25
32A	0.33 0.29	0.12
161	0.16 0.10	0.11
20A	0.06 0.04	0.07
07	0.96 0.95	0.05
26A	0.27 0.26	0.04
	0.02 0.01	
	0.07 0.06	

CUADRO A.3: Error relativo en % de las imputaciones en comparación al peso relativo por División/Subdivisión (Envío 3)