UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

MÁSTER EN ESTADÍSTICAS OFICIALES E INDICADORES SOCIALES Y ECONÓMICOS

TRABAJO FIN DE MÁSTER

2020-2021

Estandarización de la Imputación en la Encuesta de Transporte de Viajeros

---

Autora:

Noelia Mos Pérez

Tutores:

Alba M. Franco Pereira

Elena Rosa Pérez

David Salgado Fernández

14 de Julio de 2021

Agradecimientos

A mis tutores Alba, Elena y David por su dedicación y paciencia.

A mis familias por su apoyo y cariño.

A la vida, por darme tanto cuando otros no tienen nada.

UNIVERSIDAD COMPLUTENSE DE MADRID

Resumen

Facultad de Ciencias Matemáticas

Máster en Estadísticas Oficiales e Indicadores Sociales y Económicos

Estandarización de la imputación en la Encuesta de Transporte de Viajeros

Por Noelia Mos Pérez

La imputación consiste en estimar los valores perdidos o *missings* recurriendo a otros datos aportados por la unidad o a datos de otras unidades semejantes. Su importancia radica en el hecho de que aumenta la calidad de las estimaciones y se confirma por ser una fase incluida en el estándar de estadística oficial *Generic Statistical Business Process Model* (*GSBPM)*. Este estándar nace con la necesidad de estandarizar los procesos estadísticos entre organismos del mismo país y de diferentes países. Teniendo esto en cuenta, el objetivo de este Trabajo de Fin de Máster (TFM) es usar la Encuesta de Transporte de Viajeros (publicada mensualmente por el Instituto Nacional de Estadística) para desarrollar un sistema estandarizado de imputación que sustituya al que se emplea actualmente y que pueda ser aplicable a otras operaciones estadísticas. Para ello, se ha recurrido a dos fases: una de clasificación de unidades en imputables y no imputables mediante el paquete *ranger* del software libre R y otra de imputación propiamente dicha con el paquete *simputation* del mismo software.

The imputation process consists on estimating missing values using other data provided by the unit or data by other similar units. Its importance lies in the fact that it improves the quality of the estimations and is confirmed due to the inclusion into the official statistical standard Generic Statistical Business Process Model (*GSBPM)*. This standard was born because of the need of standardize statistical processes among agencies in the same country and in different countries. Taking this into account, the goal of this project is to use the Traveler's Transport Survey (published monthly by INE Spain) to develop an imputation system to replace the current one and that can be applicable to other statistical operations. Two phases have been used for this purpose: a classification of imputable and non-imputable units by means of the *ranger* package of the free software R and another phase of imputation with the *simputation* package of the mentioned software.

Índice de contenidos

Índice de figuras y tablas

1. Introducción

En esta primera sección se expondrá la motivación del trabajo o su razón de ser, así como un primer esbozo de lo que significa imputar, los errores asociados a la imputación y su importancia en el proceso de producción estadística.

A pesar de los esfuerzos de la Estadística Oficial –producción estadística objetiva y práctica procedente de organismos públicos acerca de la situación económica, social, ambiental, etc. principalmente para el gobierno y los ciudadanos[1] (Naciones Unidas, 2014)– por reducir la carga de respuesta de los informantes [principio 9 del *European Statistical Code of Practice*[2] (Eurostat, 2018)] y por hacer los cuestionarios accesibles para todos los usuarios (cuestionarios con instrucciones, sencillez en las preguntas y aportación de distintas vías para completar la encuesta) la falta de respuesta total y parcial siguen siendo un problema que afecta a la calidad de las estimaciones y dificulta el proceso de producción en general.

Imputar, según *Handbook of Statistical Data Editing and Imputation*, es "crear un conjunto de datos completo antes de la fase de estimación, sustituyendo los valores que faltan por valores estimados a partir de los datos disponibles" (Pannekoek, Scholtus, & De Waal, 2011, p.7). Por tanto, podemos considerar la imputación como el método por el que se solventa la falta de respuesta parcial, es decir, cuando una unidad no ha respondido algunas preguntas –pero otras sí- se estiman los valores de esas celdas concretas procurando preservar la distribución estadística del conjunto de datos. Para ello se recurre, según el caso, a las respuestas de otras unidades con características similares, a los agregados del grupo al que pertenece o bien a las propias respuestas de esa unidad en otros ítems de la misma encuesta o en ocasiones previas. Por su parte, la falta de respuesta total –una unidad no responde ninguna de las preguntas del cuestionario- se recurre a la reponderación; ese caso no será objeto de este Trabajo de Fin de Máster (TFM).

[1] https://unstats.un.org/unsd/dnss/hb/S-fundamental%20principles_A4-WEB.pdf
[2] https://ec.europa.eu/eurostat/documents/4031688/9394048/KS-02-18-142-ES-N.pdf/e792b761-6f09-42a9-a1e0-3a3356a0de1c

Este TFM surgió durante la realización de mis 300 horas de prácticas curriculares en el departamento de Estadísticas Industriales y Servicios del Instituto Nacional de Estadística[3] (INE, desde ahora), que fueron dedicadas a comprender el carácter necesario de la imputación para evitar sesgos en aquellos casos en que las unidades que no responden lo hacen por un motivo que "las une" –los *missing* deben responder a cierta aleatoriedad– (Pannekoek, Scholtus, & De Waal, 2011). Además, estas prácticas me han servido para tener acceso a dos herramientas imprescindibles para este TFM: la base de datos de la Encuesta de Transporte de Viajeros[4] (ETV) y la nota metodológica que se utiliza internamente en el INE (con las correspondientes explicaciones detalladas de métodos, codificación, etc). La coordinación entre prácticas y TFM ha sido perfecta ya que han compartido no solo temática, sino tutores: siendo Elena Rosa Pérez la tutora en el INE y Alba Franco Pereira la tutora académica, en coordinación con David Salgado Fernández, se han mantenido reuniones regulares desde el comienzo de las prácticas para orientar y evaluar el progreso del TFM.

La importancia de llevar a cabo una fase de imputación viene confirmada por el hecho de estar presente en uno de los estándares más importantes de la estadística internacional: *Generic Statistical Business Process Model*[5] (GSBPM), que incluye la imputación en las fases 5.3 y 5.4, para las cuales un paquete estadístico muy apropiado es *simputation*, del software R, que permite emplear gran variedad de métodos de imputación y se ha diseñado de acuerdo a *Tidy tools manifesto*[6], documento que aboga por reutilizar estructuras de datos existentes, componer funciones simples, así como llevar a cabo una programación funcional y un diseño para humanos.

Concretamente, a lo largo del presente documento se expone cómo se ha aplicado la técnica de imputación a la ETV realizada por el INE, cuyas características se detallarán más adelante pero que ya podemos adelantar que en 2020 tuvo una tasa media anual de no respuesta por ítem del 17,4% según detallan los metadatos ofrecidos por el INE en su página web oficial[7]. En la misma fuente se encuentra la justificación de la relevancia de esta operación estadística: cubrir las necesidades de los usuarios de las estadísticas, concretamente de las empresas de transportes que emplean los datos en los procesos de

[3] https://ine.es/
[4] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176906&menu=ultiDatos&idp=1254735576820
[5] https://www.ine.es/clasifi/estandar_procesos.pdf
[6] https://cran.r-project.org/web/packages/tidyverse/vignettes/manifesto.html
[7] https://www.ine.es/dynt3/metadatos/es/RespuestaDatos.html?oe=30163

negociación de concesiones con las autoridades. También tiene utilidad de cara a la investigación (en materia turística, por ejemplo) y para políticas desarrolladas por el Ministerio de Transportes, Movilidad y Agenda Urbana, entre otros.

Pero el proceso de imputación es, al fin y al cabo, un mecanismo de estimación y, por tanto, tiene asociados errores y sesgos como el que se genera debido a que los modelos de regresión diseñados se ajustan mediante los parámetros de los ítems respondidos, lo cual no tiene por qué encajar perfectamente con los ítems no respondidos. Este sesgo es aceptable siempre que sea insignificante en comparación con otras fuentes de inexactitud en el resultado, como la infraestimación de la varianza de muestreo (Pannekoek, Scholtus, & De Waal, 2011).

Otro error inducido por la imputación ocurre de cara a la estimación de distribuciones y medidas de dispersión, dado que subestima la variación en las puntuaciones por el efecto de regresión a la media, lo cual lleva a distribuciones muy apuntadas con áreas de colas muy pequeñas. Esto empeora a medida que aumenta el número de valores perdidos y es más acentuado cuando se emplea el método de imputación por la media; para contrarrestarlo se aconseja añadir una perturbación aleatoria en la imputación. (Scholtus, 2013). Otra forma de abordar el problema es mediante remuestreo o con imputación múltiple –la variación entre imputaciones sirve para estimar el aumento de la varianza debido a la falta de respuesta y la correspondiente imputación; como resultado surgen múltiples conjuntos de datos imputados– (Pannekoek, Scholtus, & De Waal, 2011).

La imputación también presenta limitaciones tales como el hecho de que no sea aplicable a datos cualitativos o que no siempre se cumple la exigencia recogida por Fellegi & Holt de que las reglas de imputación deben derivarse directamente de los *edits* [requisitos predefinidos a los que los datos deben amoldarse (Comisión Europea, 2014)]. Pongamos como ejemplo que los *edits* relacionan numerosas veces dos variables, estas dos variables deberían ser regresores en los modelos de imputación. A esto se le suman los errores humanos asociados a la imputación tales como la imputación "creativa": no documentar todos los cambios realizados y/o no respetar los protocolos establecidos por organismos europeos de Estadística Oficial. Un ejemplo de estos organismos es la Comisión Europea, que destaca en su web *EU Science Hub*[8] el

[8] https://ec.europa.eu/jrc/en/coin/10-step-guide/step-4

problema en términos de comparabilidad: algunos países no cuentan con la suficiente información adicional para imputar de manera fiable, lo cual distorsiona la posición de todos los países en los diferentes rankings que se publican (O'Connor, 2015).

Esto debe ser especialmente considerado en la actualidad, dado que la era COVID-19 ha ocasionado el cierre temporal de muchas empresas, por lo que las encuestas económicas no podrán apoyarse en los datos de los meses previos del mismo modo que venían haciendo, la fiabilidad de la imputación se reducirá y las posiciones de los países en determinadas clasificaciones no serán exactas.

2. La Encuesta de Transporte de Viajeros

La Estadística de Transporte de Viajeros es una operación estadística coyuntural llevada a cabo por el INE, que proporciona información acerca del número de viajeros transportados mensualmente. Para esto se recurre tanto a una encuesta –a empresas cuya actividad principal o no principal esté dentro de la sección H clases 4931, 4939 de la Clasificación Nacional de Actividades Económicas (CNAE 2009[9]) y que son tanto la unidad informante como la unidad estadística– como a registros administrativos: Aviación Civil y Puertos del Estado.

Los resultados se difunden clasificados según transporte urbano (autobús y metro), interurbano (autobús, ferrocarril, avión y barco), especial y discrecional por autobús[10]. Cuando hablamos de Estadística de Transporte de Viajeros nos estamos refiriendo a la operación que abarca transporte por tierra, mar y aire y que usa tanto una encuesta como registros administrativos; mientras que la Encuesta de Transporte de Viajeros se encarga solo del transporte por carretera y es en esta última en la que se ha trabajado.

2.1 Muestra

El marco del que se parte es el Directorio Central de Empresas[11] (DIRCE: sistema de información único con actualización anual que reúne a todas las empresas españolas y a sus unidades locales ubicadas en el territorio nacional) y otras fuentes auxiliares como los directorios de empresas con concesiones de transporte de viajeros de las Comunidades Autónomas.

El muestreo es aleatorio estratificado –teniendo en cuenta Comunidad Autónoma, tamaño de la empresa en términos de empleados y actividad económica–, excepto para las empresas con más de 50 asalariados que se investigan exhaustivamente, así como aquellas con un gran volumen de viajeros o las que se dedican al transporte urbano regular. El reparto de tamaño muestral se hace de forma proporcional al número de empresas y se realiza una rotación anual del 25% para evitar el cansancio de los informantes.

[9] https://www.seg-social.es/wps/wcm/connect/wss/f8ba3b82-130e-47ea-a197-c27224fa38e1/T63-TABLA+DE+CNAE09+.pdf?MOD=AJPERES

[10] https://www.ine.es/daco/daco43/notatvir.pdf

[11] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550

2.2 Estimadores

El estimador del número de viajeros en el mes *m* para cualquier dominio *D* es:

$$\hat{V}_{D,m} = \sum_h \sum_{k \in D} F_h V_{hk,m}$$

Siendo:

$V_{hk,m}$ el número de viajeros consignados en la unidad *k* seleccionado en el estrato *h* el mes *m*.

$F_h = N_h/n_h$ h=1,...,H[12], factor de elevación para todos los cuestionarios comprendidos en el estrato h-ésimo. Los factores de elevación reflejan cuantas unidades de la población quedan representadas por una unidad de la muestra

El estimador del total de viajeros en el mes m, para el total nacional, viene dado como suma de los totales estimados en cada uno de los estratos:

$$\hat{V}_m = \sum_h \sum_k F_h V_{hk,m}$$

El factor de elevación (N_h/n_h) se mantiene fijo durante el año porque la falta de respuesta se combate con imputación, por lo que no son necesarias las labores de reponderación.

2.3 Métodos de imputación que aplica el INE en la ETV

En la ETV se imputa solamente la variable viajeros, a pesar de que recaba información acerca de los ingresos y del personal, como se muestra en el cuestionario incluido en Anexo I. Esto viene justificado por el hecho de que esos datos –ingresos y personal– no son difundidos, sin embargo, son variables que se utilizan en la fase de depuración y por ello sería conveniente que el INE comenzase a imputarlas. Además, existe otra operación estadística que sí explota las variables ingresos y personal y, por consiguiente, se imputan de acuerdo con las normas de imputación establecidas en esa operación estadística: los Indicadores de Actividad del Sector Servicios (IASS)[13]

[12] Siendo *H* el último estrato definido

[13] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176863&menu=ultiDatos&idp=1254735576778

Los métodos que emplea el INE para imputar la variable viajeros son:

Método 1: Cuando la empresa tiene datos anteriores al que es objeto de imputación el programa buscará, desde el mes anterior al mes de referencia, hasta el mismo mes del año anterior, la tasa de variación –cambio en porcentaje entre dos valores- de las empresas (más concretamente de todas las variables de esas empresas) ubicadas en el mismo estrato que la empresa a imputar, con relación al mes de referencia. Si para una variable concreta dicha tasa es menor que la tasa inicial se realizará la imputación de dicha variable, aplicando la variación de las empresas ubicadas en su mismo estrato. De superarse la tasa inicial, se pasará a hacer lo mismo con el mes siguiente.

Método 2: En caso de haber llegado hasta el mismo mes del año anterior y de quedar todavía variables pendientes de imputación, se realiza un promedio de los últimos 4 meses cuya información haya dado la empresa.

Método 3: El uso de este método se eliminó a partir de 2017 debido a que conduce a resultados insatisfactorios. Consistía en que si la empresa no había dado ningún dato anterior al mes que tratamos de imputar (es decir, la empresa es nueva, por ejemplo), sus datos se imputarían con la media de los valores de las variables de las empresas contenidas en su mismo estrato.[14]

En pocas palabras, el primer método se basa en tasas de variación de empresas del mismo estrato que la que se debe imputar, el segundo método recurre a un promedio de los datos que sí haya dado la empresa previamente y el último método se aplica en caso de que no sea posible implementar el segundo porque no hay datos previos de la empresa a imputar y consiste en una media de los valores de las empresas del mismo estrato.

2.4 Variables

A continuación se muestra una tabla resumen de las 11 variables numéricas –y de valores siempre positivos o iguales a 0- de la ETV con las que principalmente se ha trabajado en este TFM. Estas son las "variables objetivo" y que se pretenden imputar a partir de otras variables "de control" como la provincia, el CNAE, etc. Para todo esto ha sido necesario contar con una variable de identificación única que, en este caso, el INE denomina "NORDEN".

[14] Observaciones obtenidas de la nota metodológica interna del INE (he accedido durante mis prácticas)

Código	Variable
VMET	Viajeros de metro
VURBRG	Viajeros de urbano regular general
VURBRE	Viajeros de urbano regular especial escolar
VURBRL	Viajeros de urbano regular especial laboral
VURBDIS	Viajeros de urbano discrecional
VREGCR	Viajeros de interurbano regular cercanías
VREGMD	Viajeros de interurbano regular media distancia
VREGLD	Viajeros de interurbano regular larga distancia
VESC	Viajeros de interurbano regular especial escolar
VLAB	Viajeros de interurbano regular especial laboral
VDIS	Viajeros de interurbano discrecional

Tabla 1. Lista de variables de la ETV

De cara a la difusión de resultados, la agregación que se hace se basa en la unificación de los dos tipos de transporte escolar (urbano regular especial escolar e interurbano regular especial escolar), así como los transportes laborales (urbano regular especial laboral e interurbano regular especial laboral) y discrecional (urbano discrecional e interurbano discrecional). Además, el transporte urbano regular general es difundido a nivel de Comunidad Autónoma.

Llegados a este punto cabe destacar que la Estadística Oficial a menudo trabaja con variables "semicontinuas", que son aquellas en las que las respuestas se corresponden o bien con un único valor (típicamente 0) o bien con una distribución continua y sesgada de otros valores. La dificultad de este tipo de variables reside en que modelizarlas a menudo requiere más de un tipo de regresión, así como en que los 0 son valores válidos, es decir, no representan valores negativos o perdidos. Son bastante comunes en la encuestas económicas, especialmente si estas miden el fenómeno en varios momentos en el tiempo (Olsen & Schafer 2001), como es el caso de la ETV, encuesta económica continua que presenta valores que, según la lógica, deberían ser enteros (viajeros, en definitiva, personas) y que, sin embargo, son continuos porque son fruto de estimaciones.

3. R en la Estadística Oficial. GSIM, CSPA y GSBPM

R se ha convertido en el lenguaje común entre estadísticos, metodólogos y científicos de datos de todo el mundo. Las razones por las que la comunidad estadística oficial está adoptando rápidamente R son claras: tiene millones de usuarios –que son, a la vez, desarrolladores de R- por todo el planeta, hay un amplio apoyo de la industria y combina una gran cantidad de funcionalidades para la preparación de datos, metodología, visualización y construcción de aplicaciones. Además, los programas informáticos basados en R se intercambian mediante normas y técnicas estandarizadas. R es un software libre, es decir, gratuito (Kowarik, 2019).

La utilidad del software R para las Estadísticas Oficiales se confirma anualmente por la Comisión Europea con la celebración de la conferencia anual titulada "Use of R in Official Statistics"[15], iniciada en 2013. La última tuvo lugar en Austria, durante 4 días en diciembre del 2020.

El hecho de que muchos investigadores y la mayor parte del personal al servicio de las Estadísticas Oficiales utilicen el mismo software contribuye a las ideas de estandarización y reutilización del GSBPM ya mencionado y del *Common Statistical Production Architecture*[16] (CSPA). Esto se ve complementado con el *General Statistical Information Model*[17] (GSIM) que propone que todos los *building blocks* o partes del proceso productivo sean tratados con la lógica del siguiente esquema:

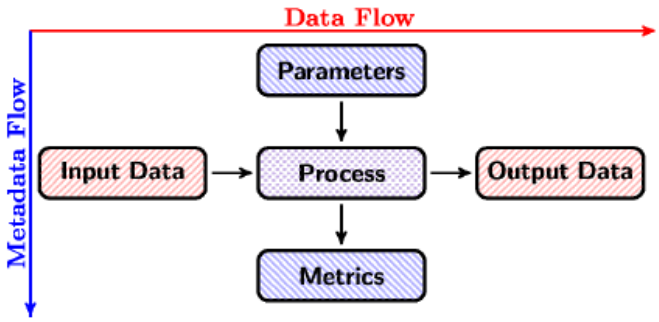


Figura 1. Esquema de flujo de datos y metadatos

[15] https://ec.europa.eu/eurostat/cros/content/use-r-official-statistics-uros2020_en
[16] https://statswiki.unece.org/display/CSPA/CSPA+v2.0
[17] https://unstats.un.org/unsd/classifications/expertgroup/egm2015/ac289-22.PDF

La lógica de trabajo pasa por diseñar cada uno de los *building blocks* o piezas del proceso a partir de su *Input* (los datos, documentos o cualquier "materia prima" que introducimos), esta sufre un *Process* o transformación que viene marcado por los parámetros que utilizamos (puede ser un método, una forma de hacer o cualquier indicación que le demos al programa que estemos utilizando). De ese proceso, siguiendo el flujo de los metadatos, se derivan las métricas: subconjuntos de resultados, por ejemplo, agregados o cada uno de los datos fruto del proceso y que configurará el *Output Data*, que no es más que el producto final, aquello que buscábamos al iniciar el proceso del *building block.* Por supuesto, el *Output* de un proceso a menudo es el *Inputs* del siguiente proceso.

En el caso de este Trabajo de Fin de Máster, el proceso general puede describirse diciendo que los datos Input son las bases de datos de la ETV con valores perdidos, el proceso sería la imputación en sí misma (para lo cual utilizaríamos como parámetros, por ejemplo, el método de imputación o el vector que nos dicte si se debe imputar o no, lo cual se explicará detalladamente en el capítulo Metodología y Resultados) y como datos Output sería la misma base de datos pero ya sin valores perdidos. Esta misma estrategia puede emplearse en cada paso concreto del proceso y ser diseñada de tal manera que pueda reutilizarse en otros contextos, bien para otra operación estadística, para la misma pero en otro país o para otro momento del tiempo, favoreciendo así el uso eficiente de los recursos [principio 10 del *Code of Practice* (Eurostat, 2018)].

Por su parte, el estándar de producción GSBPM -diseñado por la UNECE y adoptado por la mayoría de países con un sistema estadístico desarrollado- pretende proporcionar un listado estándar de tareas estadísticas de forma que en todas las oficinas europeas de Estadísticas Oficiales se tengan en cuenta todas las tareas y su respectivo orden (usamos la expresión "tener en cuenta" y no "realizar" porque el GSBPM no es lineal, es decir, en muchas partes del proceso estadístico es necesario retroceder algunos pasos y repetir tareas o, incluso, no realizarlas).

Si nos centramos en la parte del GSBPM que atañe a este proyecto (fases 5.3 y 5.4: *data editing and imputation*) esta incluye las subfases: 5.3.1 Ejecutar detección y tratamiento de errores (input), 5.3.2 Ejecutar detección y tratamiento de errores (output), 5.3.3 Elaborar informe de depuración e imputación, 5.3.99 Otras tareas.

En la versión española del GSBPM[18] estas fases están comprendidas como una sola, puesto que no tiene sentido tratarlas individualmente porque son absolutamente dependientes. Cada una de las tareas expuestas unas líneas atrás pueden ser realizadas mediante distintos paquetes del software R, que se recogen en la siguiente ilustración[19] -a la cual se ha aplicado zoom en la imagen de la derecha para mostrar mejor los paquetes asociados a las faes 5.3 y 5.4 del *GSBPM*-:

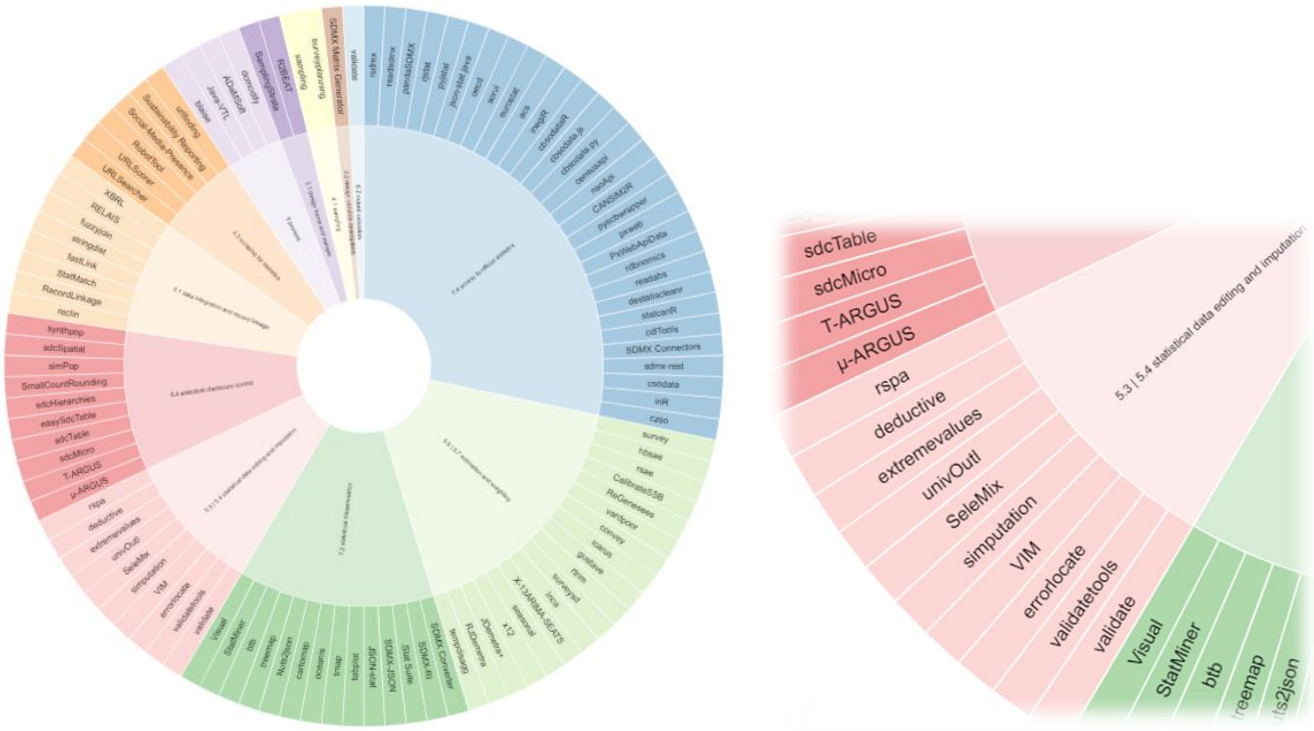


Figura 2. Rueda de paquetes del software R en consonancia con las fases del GSBPM

Como ya se ha expuesto en un apartado previo, en este trabajo se ha optado por recurrir al paquete de R s*imputation* –desarrollado por Mark van der Loo[20], trabajador del Instituto Nacional de Estadística de Holanda (CBS)[21]- porque la gran variedad de métodos que incluye y por ajustarse a los principios del *Tidy tools manifesto* (reutilizar estructuras de datos existentes, funciones simples, programación funcional y un diseño para humanos). Este paquete y una explicación más en profundidad del proceso de imputación serán abordados en el siguiente apartado.

[18] https://www.ine.es/clasifi/estandar_procesos.pdf
[19] https://github.com/SNStatComp/awesome-official-statistics-software
[20] https://github.com/markvanderloo
[21] https://www.cbs.nl/en-gb

4. Paquete *simputation*[22]

Sabiendo que la imputación consiste en estimar los valores no respondidos y rellenarlos preservando la distribución estadística del conjunto de datos para mejorar y simplificar la estimación de agregados, se procede a clasificar los valores perdidos de acuerdo a *Handbook of Statistical Data Editing and Imputation* (Pannekoek, Scholtus, & De Waal, 2011):

a) *Missing completely at random* (MCAR): la probabilidad de no respuesta no está asociada al valor, sino a la aleatoriedad: un informante que olvida responder, un dato que se pierde en el proceso de grabación de datos,...

b) *Missing at random* (MAR): la probabilidad de que un valor sea perdido depende de la variable auxiliar (pero no de del valor de la variable objetivo). Por ejemplo, observamos patrones de no respuesta asociados al tamaño de la empresa o a la provincia en la que se ubica.

c) *Not missing at random* (NMAR): el valor de la variable objetivo condiciona la probabilidad de respuesta.

Otra forma de clasificarlos es "ignorable" (si el conocimiento del parámetro asociado al mecanismo de no respuesta no ayuda a calcular los parámetros de interés) y "no ignorable" (asociado al mecanismo NMAR). Esto se deriva directamente del hecho de que el gran problema de la falta de respuesta es la introducción de sesgos (las unidades no responden por un motivo que está ligado directamente a la variable objetivo).

Los distintos modelos de imputación buscan aproximarse lo máximo posible al valor real, así como a la consistencia interna. Esos métodos pueden ser comparados –aplico uno y luego otro y decido cual es mejor– o combinados –un modelo de imputación para cada subpoblación–. Algunos de los principales métodos de imputación existentes, según *Handbook of Statistical Data Editing and Imputation* (Pannekoek, Scholtus, & De Waal, 2011) y Van der Loo, (2021), son:

4.1 Deductive imputation: se aplica en los casos en que el resto de respuestas de la unidad permiten deducir el valor perdido. Por ejemplo, la modalidad *proxy* (pongamos por caso que una unidad no ha contestado su nacionalidad pero sí la

[22] https://cran.r-project.org/web/packages/simputation/simputation.pdf

de su pareja, así que se le imputa esta última), *balance edits* (reglas matemáticas: si según mis *edits* A+B=C, imputar B contando con A y C no es un problema).

4.1.1 I*mputation by variable derivation* → funciones impute_proxy e impute_const en *simputation.*

Mientras que la función *proxy* sirve bien para una imputación por razón, por la media del grupo o para copiar un valor desde otra variable, la función *const* simplemente imputa una constante.

4.2 Model-based imputation: se emplea como información adicional variables fuertemente correlacionadas con la variable objetivo.

4.2.1 Decistion tree imputation → funciones impute_cart e impute_rf en *simputation.*

Sirven para cualquier tipo de datos. Mientras que la versión *cart* recurre a un árbol de clasificación y regresión, *rf* emplea el método *random forest* (conjunto de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales).

4.2.2 EM-based imputation → función impute_em en s*imputation.*

Para variables numéricas con distribución normal multivariante. Se estima la media y la matriz de varianzas y covarianzas en base al algoritmo Esperanza-Maximización de Dempster, Laird y Rubin (1977). Este algoritmo alterna pasos de esperanza (paso E), donde se computa la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables, y un paso de maximización (paso M), donde se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E.

4.3 <u>*Regression imputation*</u>: se emplea información adicional, cuidando el principio de parsimonia mediante la revisión de indicadores de calidad (BIC, R^2 ...). Los residuos pueden ser incluidos en la formulación (si estimamos medidas de dispersión) o no (si lo que interesa estimar son medias y totales).

4.3.1 Robust linear regression imputation → función impute_rlm en *simputation.*

Puede utilizarse para imputar variables numéricas empleando predictores numéricos y/o categóricos. En la estimación M, la minimización de los cuadrados de los residuos se sustituye por una función convexa alternativa de los residuos que disminuye la influencia de los valores atípicos.

4.3.2 Lasso/elastic net/ridge regression imputation → función impute_es en *simputation.*

Útil para imputar variables numéricas empleando predictores numéricos y/o categóricos. Para este método, los coeficientes de regresión se encuentran minimizando la suma de cuadrados de los residuos aumentados con un término de penalización dependiendo del tamaño de los coeficientes. Para la regresión de lazo (*lasso)* (Tibshirani, 1996), el término de penalización es la suma de cuadrados de los coeficientes. Para la regresión de cresta (*ridge)* (Hoerl y Kennard, 1970), el término de penalización es la suma de los valores absolutos de los coeficientes.

4.3.3 Median imputation → función impute_median en *simputation.*

Se emplea cuando no existe información auxiliar.

4.4 <u>*Hot deck imputation*</u>: es un procedimiento no paramétrico basado en utilizar los valores de un registro ("donante", unidades sin errores) para reemplazar los valores erróneos y perdidos en otro registro ("receptor", con características muy similares al donante). La imputación basada en donantes es más adecuada que la de modelos para casos en los que se deben imputar varias variables en un registro y sus relaciones se deben mantener lo máximo posible (multivariado).

4.4.1 *Random Hot Deck Imputation* → función impute_rhd en *simputation.*

Puede emplearse en cualquier tipo de datos (numéricos/categóricos/mixtos). Un valor faltante se copia de un registro muestreado. Opcionalmente se toman muestras dentro de un grupo, o con probabilidades de muestreo no uniformes.

4.4.2 *Sequential hot deck imputation* → function impute_shd en *simputation.*

El conjunto de datos se ordena utilizando las variables predictoras. Los valores faltantes o sus combinaciones se copian del registro anterior (o posterior si estamos imputando datasets "antiguos") donde el valor está disponible.

4.4.3 *Predictive mean matching* → función impute_pmm en *simputation.*

Solo para variables numéricas. Los valores que faltan o sus combinaciones se imputan primero utilizando un modelo predictivo. A continuación, estas predicciones se sustituyen por los valores observados más cercanos a la predicción (aquel con la menor desviación absoluta respecto a la predicción).

4.4.4 *K-nearest neighbour imputation* → función impute_knn en *simputation.*

Aplicable a cualquier tipo de dato. Para cada registro que contiene valores faltantes, los *k* registros completados más similares se determinan en base al coeficiente de similitud de Gower (Gower, 1971). Después, se realiza un muestreo de esas unidades para elegir al "donante" efectivo.

5. Metodología y resultados

La metodología inicialmente diseñada planteaba únicamente una fase de construcción y depuración de un fichero de datos de la ETV desde 2009 a 2020 y aplicar imputación sobre las distintas variables de viajeros. Sin embargo, al empezar a trabajar nos dimos cuenta de que el software estaba imputando unidades cuya actividad no se correspondía con la actividad imputada. Por ejemplo, una empresa de transporte escolar debe ser imputada solo en su categoría de viajeros, porque sabemos, sin necesidad de imputar, que tendrá 0 viajeros de otro tipo de transporte:

	Viajeros de metro	Viajeros de transporte escolar	Viajeros de transporte laboral
Empresa de transporte escolar (**Antes** de imputar)	NA	1000	NA
Empresa de transporte escolar (**Después** de imputar)	500	1000	300

Esto nos hizo darnos cuenta de que antes de imputar era necesario implantar un árbol de decisión (técnica de *machine learning* basada en la clasificación binaria) que entrenase al algoritmo para imputar con valor 0 cuando una empresa nunca se había dedicado a un tipo de transporte concreto o cuando el periodo no se correspondiese con el periodo activo de la empresa (por ejemplo, se debe asignar 0 al transporte escolar en el mes de agosto).

El siguiente paso en la metodología es la imputación de los valores de Enero de 2020 a partir de todo lo aprendido gracias al fichero de datos que data del 2009. Este mes se ha elegido arbitrariamente y debemos recordar que en ese entonces la crisis por COVID 19 aún no había estallado en España pero, aún así, se ha definido el regresor "Confinamiento" de cara a la imputación futura de datos "post COVID" y siempre teniendo en mente la intención de estandarizar y utilizarlo en el proceso de producción de la ETV; el TFM surgió a raíz de la necesidad acaecida durante el confinamiento de 2020 cuando se detectó que los métodos de imputación utilizados podrían no ser los

más adecuados porque el cambio en la serie temporal fue el más acusado de la última década.

La tercera etapa definida en esta metodología será la breve evaluación de la calidad de las imputaciones, comparando los datos imputados con los reales (de ahí el por qué se ha imputado Enero de 2020 y no una fecha más próxima al momento actual, Junio de 2021). A continuación se expone detalladamente la metodología y resultados de cada una de las fases de este TFM.

5.1 Árbol de decisión

5.1.1 Metodología

Se trata de un modelo predictivo formado por reglas binarias (si/no) de tipo no paramétrico que simplifica los predictores y se compone de dos fases: entrenamiento de un parte de los datos o fase de prueba y luego, test para cuantificar la capacidad predictiva del modelo.

Se ha hecho uso del paquete *ranger*[23] de R, que es un modelo *Random Forest* pero más rápido computacionalmente hablando. Los modelos *Random Forest* están formados por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante bootstrapping (remuestreo: para medir la calidad de la inferencia consideramos las primera muestra como si fuese la población –los datos de todas las unidades existentes– y la submuestra extraída de la primera muestra es la que se emplea en los cálculos).

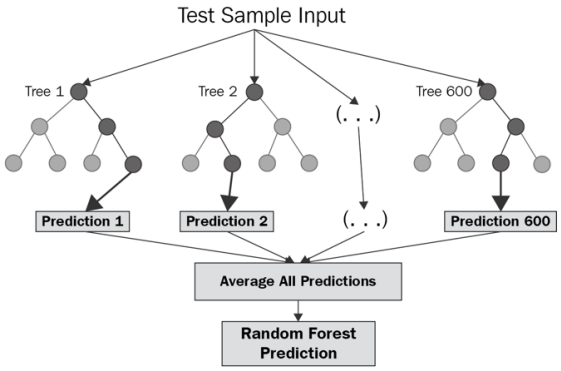


Figura 3. Funcionamiento de un árbol de decisión[24]

[23] https://cran.r-project.org/web/packages/ranger/ranger.pdf
[24] https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Algunas ventajas de los árboles de decisión son la posibilidad de utilizar como predictores tanto variables numéricas como categóricas, que no exige ninguna distribución específica de los datos porque son modelos no paramétricos, no se ven muy influenciados por *outliers* y el algoritmo subyacente no se corresponde con modelos lineales, por lo que no afectan los problemas de multicolinealidad en los regresores.

Por el contrario, como desventajas, son sensibles a datos de entrenamiento en los que un grupo (si dividimos en estratos por características) es notablemente más cuantioso que los demás y tienden a presentar una varianza elevada y sobreajuste de los datos –se da tanta importancia a los datos utilizados en la fase de entrenamiento que es incapaz de generalizar a los datos test–[25]. Para este último defecto se recurre a técnicas que combinan árboles individuales –siendo la predicción final la media de todos ellos–, como es el caso del *Random Forest* que se aplicará en este TFM con la finalidad de decidir a qué medios de transporte aplicar la imputación y a cuáles no por el motivo ya explicado al comienzo de esta sección. Por tanto, el *Random Forest* es, a la vez, una herramienta de clasificación y regresión. Para ello se han realizado las siguientes tareas:

1. Construcción del fichero de 500 empresas y más de 200 000 observaciones con los datos de 2009 a 2020 filtrando de forma que solo quede una fila por unidad y mes (en ocasiones, para el mismo mes hay más de un dato porque el INE detecta errores y se produce un recontacto o porque se hace una imputación y después se consigue el dato real, así que nos quedamos solo con el último dato aportado por la empresa para cada mes).
2. Regresores para cada tipo de transporte (los tipos de transporte fueron expuestos en la subsección variables):

- Variable target: a partir de la variable "Origen" se define como 0 y 1 si el dato no ha sido imputado o si sí lo ha sido, respectivamente. Además, se han creado una serie de regresores para cada tipo de transporte:

- n1: Número de meses con valor no nulo consecutivos.

- n2: Variable indicadora de la variable n1 que tomará el valor 0 si n1=0 y 1 en caso contrario.

[25] https://rpubs.com/Joaquin_AR/255596

- n3: Número de meses con valor *missing* consecutivos.
- n4: Variable indicadora de la variable n3 que tomará el valor 0 si n3=0 y 1 en caso contrario.
- n5: Número de meses con valor nulo consecutivos.
- n6: Variable indicadora de la variable n5 que tomará el valor 0 si n5=0 y 1 en caso contrario.
- n7: Fracción del total de meses no nulos consecutivos entre la longitud total de la serie (siendo esta el número de meses que cada empresa está en muestra).
- n8: Fracción del total de meses *missing* entre la longitud total de la serie.
- n9: Fracción del total de meses no nulos entre la longitud total de la serie.
- n10: Total de meses no nulos.
- n11: Variable indicadora de la variable n10 que tomará el valor 0 si n10=0 y 1 en caso contrario.
- n12: Fracción correspondiente a n10.
- n13: Total de meses nulos, excluyendo perdidos.
- n14: Variable indicadora de la variable n13 que tomará el valor 0 si n13=0 y 1 en caso contrario.
- n15: Fracción correspondiente a n14.
- n16: Total de meses con valor perdido.
- n17: Variable indicadora de la variable n15 que tomará el valor 0 si n13=0 y 1 en caso contrario.
- n18: Fracción correspondiente a n16.

El tipo de regresores que se corresponden con dicotomizaciones de sus predecesores (n2, n4, n6...) parecen estar recogiendo la misma información que otros. Sin embargo, esta información es ligeramente diferente y, matemáticamente, es posible incluir ambos porque desaparece el problema de la multicolinealidad al usar un árbol de decisión que, en definitiva, es un modelo no lineal cuya potencialidad puede ser mayor si se incluyen

estos regresores que en el caso contrario. En cualquier caso, se analizará la importancia de cada regresor mediante la función *importance* de R para contrastar esto.

También se emplean una serie de regresores que no son asociados al tipo de transporte y que, en ocasiones, no requieren transformación sino que ya son variables aportadas por la ETV pero que, en todo caso, podrían ser claves para el *Random Forest*:

- "Mes": Variable categórica que toma 12 valores.
- "Confinamiento": Variable indicadora que toma los siguientes valores: 0 para los meses anteriores a marzo de 2020, 1 para los meses de marzo a agosto de 2020 y 2 desde agosto de 2020 a diciembre 2020.
- "Vargestion": Se trata de una variable nominal que indica distintos tipos de incidencia y puede ser insertada en el modelo porque los random forests tienen la ventaja de que integran de modo muy natural regresores cualitativos y cuantitativos.
- "NTrabajadores": Número de trabajadores de la unidad en el mes correspondiente.
- "ObservacionesDepuracion": Variable indicadora que tomará el valor 0 si no existen observaciones escritas en el archivo para esa unidas y 1 en caso contrario.

3. *Random Forest* → ¿cómo funciona el paquete *ranger* de R?

Una de las peculiaridades del paquete *ranger,* creado por Marvin N. Wright, es que no puede trabajar con valores perdidos -lo cual es algo paradójico en este caso-, por lo que se ha imputado el regresor "NTrabajadores" con valores aleatorios de manera "provisional" para poder aplicar la función r*anger*. Parece algo ilógico tener que imputar para crear un árbol de decisión que indique cuándo se puede imputar o cuándo no, sobre todo si esa imputación es aleatoria y no fruto de un proceso refinado. Este paso viene apoyado por dos cuestiones: (1) el objetivo del trabajo es la imputación de viajeros, "NTrabajadores" es una de las variables explicativas entre muchas otras y, por tanto, no requiere una imputación sofisticada, (2) algunos autores recomiendan introducir cierta aleatoriedad en el proceso de imputación para corregir errores, como se ha expuesto en la introducción.

Uno de los añadidos de *ranger* es que permite definir el número de árboles que se elaboran. Ese número se ha definido de acuerdo a una gráfica de estabilización que indica a partir de qué número de árboles el error no se reduce, como esta:

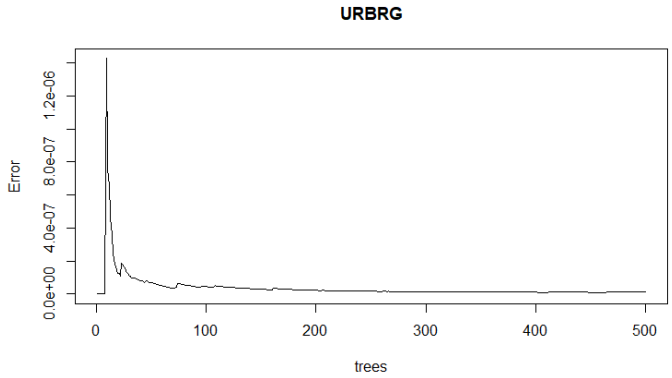


Figura 4. Gráfica de estabilización del error según el número de árboles para viajeros de urbano regular general

En la gráfica se observa que el resultado es igual si computamos 100 árboles que con 500. Por lo tanto, sería más eficiente computar solo 100. Este paso de control del número de árboles es necesario para reducir los tiempos de procesamiento del software y no por una cuestión de *overfitting* o sobreajuste –ya que los *random forest* son robustos ante este defecto[26]-.

Era de esperar que todas las gráficas siguieran una pauta similar a la Figura 5, por tanto para la base de datos con la que se trabaja en este TFM no sería imprescindible hacer estos gráficos porque no hay tantos datos como para que sea necesario reducir los tiempos de procesado. La finalidad de añadir esa línea de código era facilitar la estandarización y que utilizasen este recurso quienes tuviesen más datos o peores máquinas. Queremos incidir aquí en la importancia no solo en la forma de la función que nos indica el número de árboles a incluir, sino también de la magnitud del error. En este sentido queremos mostrar como algunas gráficas de estabilización del error en función del número de árboles pueden arrojar resultados realmente sorprendentes, por ejemplo:

[26] https://www.cs.us.es/cursos/rac-2018/temas/tema-05.pdf

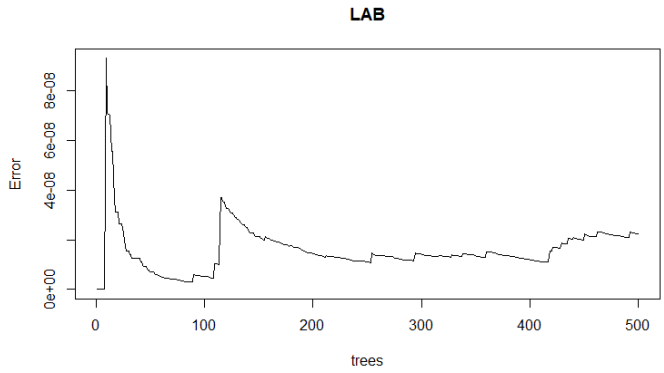


Figura 5. Gráfica de estabilización del error según el número de árboles para viajeros de interurbano regular especial laboral

Se observa que el error es menor si empleamos 100 árboles que si recurriésemos a 200, 300,... Esto debe ser comprendido en términos numéricos, es decir, aunque la gráfica parezca tan drástica y haya tanta diferencia entre picos debido a una cuestión de escala, numéricamente hablando, el error oscila entre el 0 y el 0,00000008. Es decir, apenas nada pero, aún así, se ha modificado el script para calcular cada *ranger* con el número de árboles más adecuado para su tipo de transporte. El resto de tipos de transporte cuyas gráficas vale la pena ver se encuentran en Anexo II.

Posteriormente, el procedimiento r*anger* aplica la partición del archivo para utilizar una parte de los datos como entrenamiento y aplicar lo aprendido al resto de datos (test). Llegados a este punto podemos considerar que ya se ha aplicado el algoritmo y, antes de utilizar las predicciones que ha producido, se calcula la importancia de cada uno de los regresores. Es decir, R nos indica cuántas veces ha utilizado el modelo cada uno de los regresores. Una vez realizado ese paso se procede a la evaluación de la capacidad de clasificación del árbol creado mediante una curva ROC (*Receiver Operating Characteristic*, "Característica Operativa del Receptor" en español).

Las curvas ROC miden la capacidad de clasificación normalmente en términos binarios –unidad imputable/no imputable, en este caso– mediante la "sensibilidad", siendo esta la probabilidad de clasificar correctamente a una unidad como "positiva" –imputable- y "especificidad", que es la probabilidad de clasificar correctamente a una unidad como "negativa" –no imputable- (Pepe, 2003). Respecto al gráfico, el eje x representa

1-especificidad que es la probabilidad de equivocación al clasificar como negativo o no imputable. Por ello, la curva será peor cuanto más a la derecha esté[27].

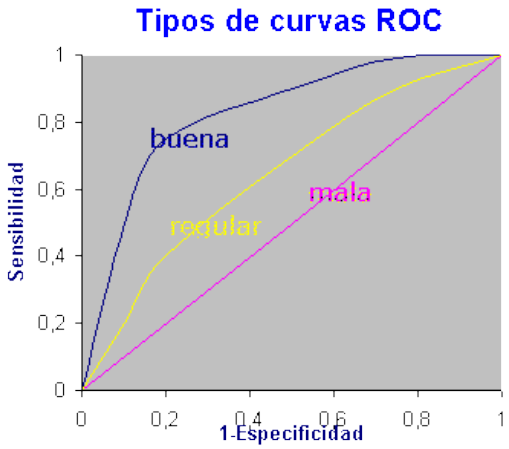


Figura 6. Tipos de curvas ROC[28]

Una vez dibujada la curva nos interesa calcular el área bajo ella (AUC), que nos dirá si la clasificación es de baja exactitud (0,5 – 0,7), útil para algunos propósitos (0,71 – 0,9) o de exactitud alta (0,91 – 1), según Muñoz Pichardo & del Valle Benavides (2020).

5.1.2 Resultados

Los resultados del árbol de decisión son una serie de predictores (1 y 0) que definen si la unidad es imputable o no, respectivamente, así como el peso o importancia de cada regresor para el algoritmo. Un ejemplo de este *output* relativo a las importancias se expone a continuación.

[27] http://ares.inf.um.es/00Rteam/pub/mamutCola/modulo2.html
[28] http://www.hrc.es/bioest/roc_1.html

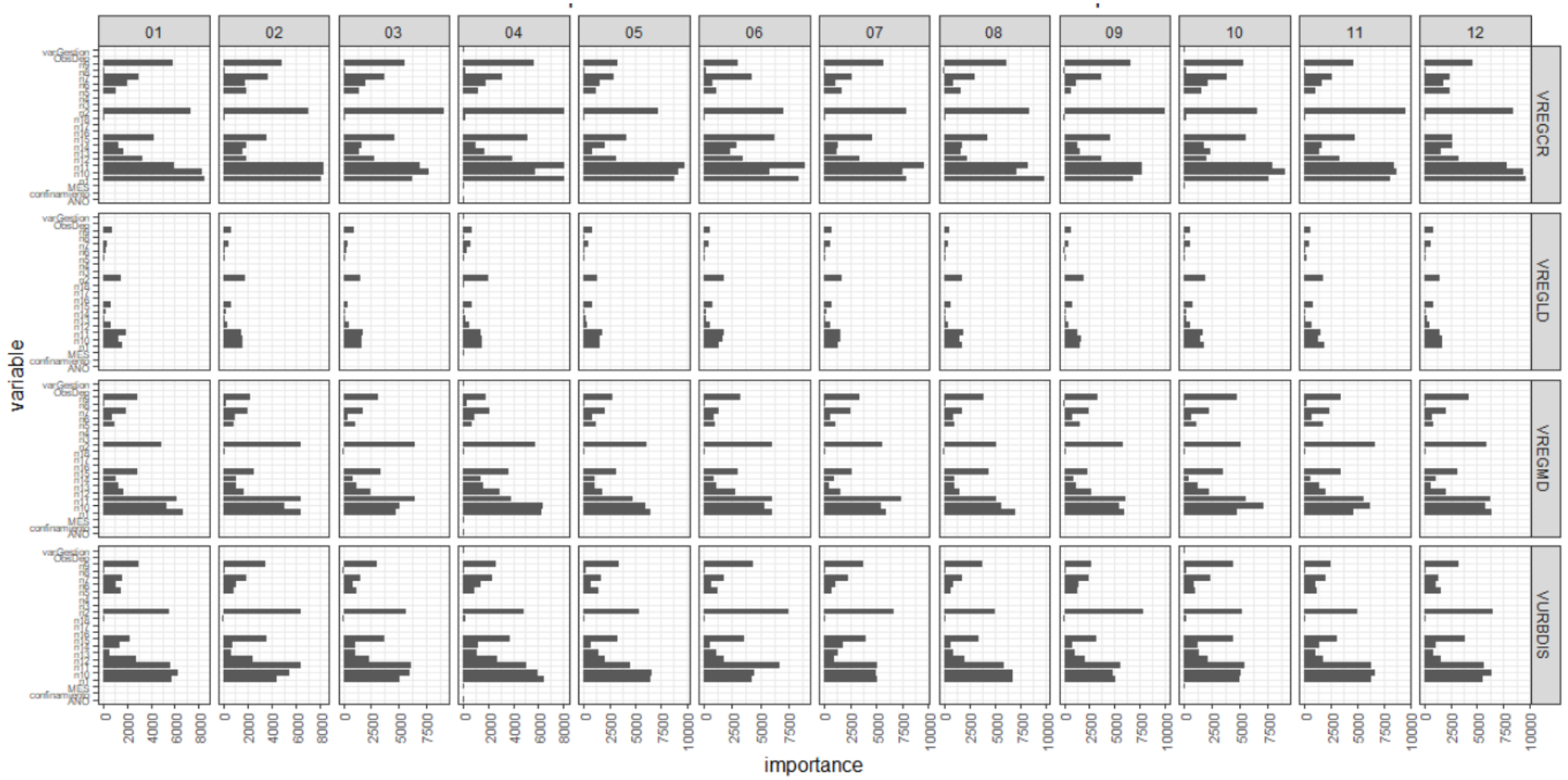


Figura 7. Gráfico de barras de la importancia de cada regresor para 4 tipos de transporte

El gráfico muestra que existen unas oscilaciones mensuales muy importantes y que cada tipo de transporte se apoya en regresores diferentes. Aún así, observamos que ciertos regresores no aportan prácticamente nada (confinamiento, MES, n8, ObservacionesDepuración, varGestion,...). Además, se confirma que los regresores que se derivan de otros mediantes dicotomización sí son útiles (n2, n11, n14,...). El resto de gráficos para otros tipos de transporte pueden consultarse en Anexo IV.

Que unos regresores aporten más información que otros al modelo no significa que el modelo en sí sea bueno. Para valorar esto se recurre a la curva ROC –cuyo funcionamiento se ha explicado en el subapartado Metodología-.

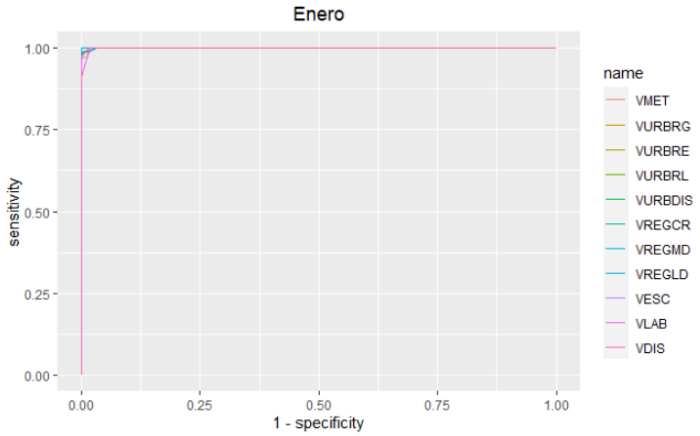


Figura 8. Curva ROC para cada tipo de transporte de la ETV en el mes de Enero

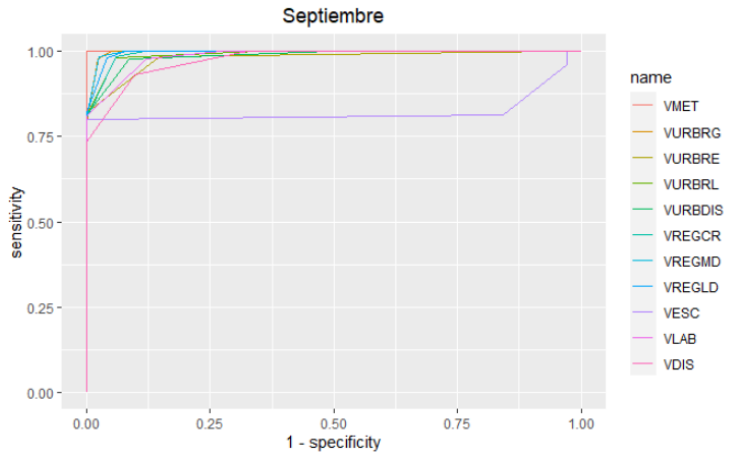


Figura 9. Curva ROC para cada tipo de transporte de la ETV en el mes de Septiembre

En general observamos que las curvas son bastante buenas, sobre todo la de Enero, en la que algunos tipos de transporte se "superponen" al 1 o extremo máximo de calidad de la curva y todos los demás se encuentran muy cerca. En el caso de Septiembre, existen predicciones muy buenas y otras un tanto alejadas pero que siguen por encima del 0.75.

Se ha expuesto el gráfico correspondiente al mes de Enero porque es el que se ha utilizado para poner a prueba el sistema de imputación diseñado y, como se mostrará más adelante, el que se ha utilizado para evaluar su capacidad predictiva. Por otra parte, se ha presentado el gráfico de Septiembre porque es el más "problemático" para algunos tipos de transporte (como se muestra en la Figura 10). El resto de curvas ROC correspondientes a los meses entre febrero y diciembre puede consultarse en Anexo III.

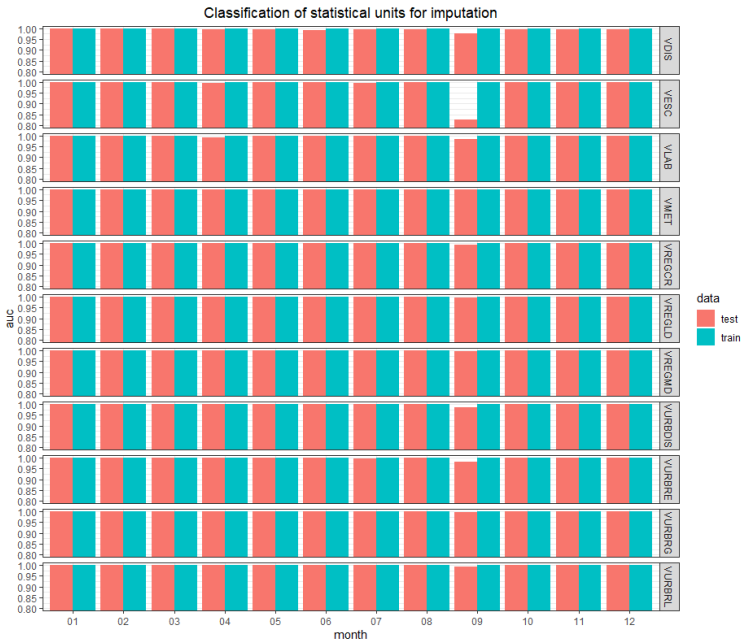


Figura 10. Áreas bajo la curva ROC de los datos de entrenamiento y los datos del test para cada tipo de transporte de la ETV para todos los meses

Se observa que el árbol de decisión ha sido capaz de clasificar entre imputables y no imputables las observaciones asociadas a todos los tipos de transporte, obteniéndose una clasificación excelente según las curvas ROC (con excepciones muy puntuales en el mes de Septiembre debido al comienzo del curso escolar y al descenso del número de excursiones turísticas). El producto final de esa clasificación es una serie de predicciones materializadas en 0 y 1 que significan "dato no imputable" y "dato imputable", respectivamente. Un valor perdido será "no imputable" cuando la

dedicación de la empresa no concuerde con lo que se pregunta. Por ejemplo, una empresa de transporte urbano discrecional que no ha respondido tendrá valor 0 (“no imputable”) para todas las variables (número de viajeros de metro, de escolar, etc) excepto para la variable “número de viajeros de urbano discrecional”, que tendrá valor 1 en las predicciones del árbol y otro valor fruto de la imputación que se expondrá unas páginas más adelante.

Se exponen a continuación, a modo de ejemplo, las primeras 1000 predicciones de los tipos de transporte urbano regular general e interurbano discrecional:

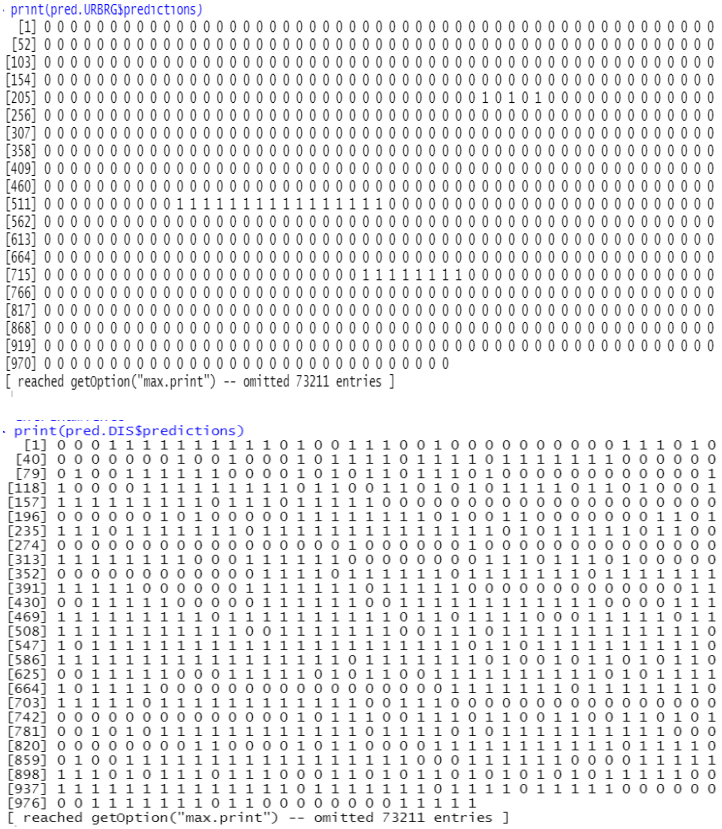


Figura 11. Output de R con las predicciones para dos tipos de transporte de la ETV

Cabe destacar que los 0 y 1 de esta imagen reflejan probabilidades. No se trata de una variable dicotómica, sino de la probabilidad de que la unidad sea imputable (1) o no imputable (0), de lo que se deduce que el modelo es bastante bueno porque prácticamente todas las probabilidades expuestas son “extremas” (0 y 1). Pero encontrar un 0,0002 no indica error, sino que para ese tipo de transporte la predicción no es perfecta –pero está muy cerca de serlo y podemos considerar esa unidad no imputable-. Eso es lo que ha ocurrido en los casos de VURBRL, VURBDIS, VREGMD, VLAB:

```
> print(pred.URBDIS$predictions) #Esta variable es la que me dice cuando imputar (1)
 y los 0 se imputan como 0
   [1] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002304683
   [7] 0.0002304683 0.0000000000 0.0000000000 0.0002304683 0.0000000000 0.0002304683
  [13] 0.0000000000 0.0000000000 0.0000000000 0.0002304683 0.0000000000 0.0000000000
  [19] 0.0002304683 0.0000000000 0.0000000000 0.0002304683 0.0002304683 0.0000000000
  [25] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000
  [31] 0.0000000000 0.0002304683 0.0002304683 0.0002304683 0.0000000000 0.0002304683
  [37] 0.0000000000 0.0000000000 0.0002304683 0.0000000000 0.0000000000 0.0000000000
  [43] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002304683
  [49] 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0002304683
  [55] 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0002304683
  [61] 0.0000000000 0.0000000000 0.0002304683 0.0000000000 0.0002304683 0.0000000000
  [67] 0.0000000000 0.0000000000 0.0000000000 0.0002304683 0.0002304683 0.0002304683
  [73] 0.0002304683 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002304683
  [79] 0.0002304683 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002304683
  [85] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002304683
  [91] 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0002304683
  [97] 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0000000000 0.0000000000
 [103] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002304683 0.0002304683
 [109] 0.0000000000 0.0002304683 0.0002304683 0.0002304683 0.0000000000 0.0000000000
 [115] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002304683 0.0000000000
 [121] 0.0000000000 0.0002304683 0.0000000000 0.0000000000 0.0002304683 0.0000000000
 [127] 0.0002304683 0.0002304683 0.0000000000 0.0000000000 0.0000000000 0.0002304683
 [133] 0.0000000000 0.0002304683 0.0002304683 0.0000000000 0.0002304683 0.0000000000
 [139] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.9982304683 0.9982304683
 [145] 0.9982304683 0.0000000000 0.9982304683 0.0002304683 0.0002304683 0.0000000000
 [151] 0.0002304683 0.0002304683 0.0000000000 0.0000000000 0.0000000000 0.0000000000
 [157] 0.0000000000 0.0000000000 0.0002304683 0.0000000000 0.0002304683 0.0002304683
 [163] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000
 [169] 0.0000000000 0.0002304683 0.0002304683 0.0002304683 0.0002304683 0.0002304683
```

Figura 12. Output de R con las predicciones para el transporte urbano discrecional

Este "producto final" de la primera fase del proyecto (árbol de decisión >> imputación por donantes >> imputación por mediana) es imprescindible para imputar, puesto que se trata de una corrección necesaria para que el paquete *simputation* haga los cálculos solo para aquellas observaciones (cruce de empresa por tipo de transporte y mes, es decir, cada observación no se corresponde con una unidad) en las que tenga sentido imputar y le otorgue automáticamente un 0 a las observaciones perdidas que no se correspondan con el tipo de transporte al que se dedica la empresa que ha generado el valor perdido.

5.2 Imputación por donantes de viajeros

De los múltiples modelos posibles de imputación se ha elegido el basado en donantes (*Hot Deck Imputation)* porque es el más favorable para casos en los que se deben imputar varias variables en un registro. Además, este método permite explotar la potencialidad de contar con un fichero de datos tan extenso en el tiempo y recurrir así a valores que la propia unidad declaró en el pasado.

5.2.1 Metodología

En primer lugar, se ha creado un nuevo archivo de datos solo con las unidades imputables según el árbol de decisión y que, a su vez, cuentan con una incidencia que permite imputar (por ejemplo, si la codificación de la incidencia dice que la empresa está cerrada definitivamente, CD, no será necesario aplicar imputación, solo asignarle un 0). Sobre este archivo se ha ido repitiendo la imputación oleada a oleada, siendo

estas los momentos del mes en el que el INE finalmente accede a los datos. Es decir, los cuestionarios se graban hasta en tres ocasiones durante el mismo mes -por ser una encuesta mensual, y en cada una se van añadiendo a la base de datos los cuestionarios que no pudieron ser recabados en las oleadas previas, así como las modificaciones por errores detectados en envíos previos.

Concretamente para la imputación se ha recurrido a *Sequential Hot Deck Imputation* (copia los valores faltantes del registro anterior) y a *Predictive Mean Matching* (calcula un modelo predictivo de los perdidos y luego esas predicciones son sustituidas por los valores observados más cercanos a la predicción). Estos métodos son sólidos al tratarse de la imputación de Enero de 2020, cuando los valores pasados de cada empresa –concretamente se ha usado Diciembre de 2019- aún eran un referente por no haber sucedido aún la pandemia y los profundos cambios asociados.

5.2.2 Resultados

	Perdidos			
	Oleada 1	Oleada 2	Oleada 3	Resultado final
VMET	4	3	1	0
VURBRG	40	40	27	0
VURBRE	124	66	48	0
VURBRL	39	18	14	0
VURBDIS	57	28	22	0
VREGCR	106	54	41	0
VREGMD	86	43	29	0
VREGLD	10	6	3	0
VESC	301	164	122	0
VLAB	117	58	44	0
VDIS	400	213	160	0

Tabla 2. Número de perdidos por oleada y tipo de transporte

Como resultado de esta fase del proyecto observamos en la Tabla 2 que cada vez son menos los *missings* hasta llegar a 0 tras la tercera oleada. Se pueden consultar algunas tablas en las que resumidamente se observan los valores reales y los imputados en la sección Evaluación de la calidad de la imputación.

5.3 Imputación por mediana de viajeros

Hasta ahora el INE ha recurrido a imputación por la media del grupo. En este trabajo se ha aplicado la imputación por mediana porque es bien sabido que esta medida es más robusta que la media. Esta es otra de las aportaciones de este TFM al INE junto con la de facilitar la estandarización del sistema de imputación en dos fases: decidir qué datos se deben imputar y después llevar a cabo la imputación por diversos métodos.

5.3.1 Metodología

La imputación por mediana aquí elaborada otorga a cada valor perdido el valor mediano del grupo de la unidad. Estos grupos se han formado a partir de 3 variables:

- Comunidades Autónomas (CCAA): se hipotetiza que la situación geográfica será decisiva para la variable objetivo de imputación, el número de viajeros que transporta cada empresa.
- TAME (CodTame): esta variable clasifica a las empresas según el número de asalariados con los que cuenta, siendo las categorías las mostradas en la Tabla 3:

TAME	Número de asalariados
00	sin asalariados
01	1 y 2 asalariados
12	3 a 5 asalariados
13	6 a 9 asalariados
14	10 a 19 asalariados
15	20 a 49 asalariados
16	50 a 99 asalariados
17	100 a 199 asalariados
18	200 a 499 asalariados
19	500 a 999 asalariados
20	1000 a 4999 asalariados
21	5000 o más asalariados

Tabla 3. Códigos de TAME de acuerdo al número de asalariados[29]

[29] Fuente: nota metodológica interna del INE (obtenida durante mis prácticas)

Tiene sentido pensar que este código será indicativo del tamaño de la empresa, es decir, cuantos más viajeros transporte (variable objetivo) más asalariados contratará (variable TAME).

- Clasificación Nacional de Actividades Económicas[30] (CodCNAEDirce): de acuerdo con el código que se otorgue a las empresas en base al CNAE, se crearán grupos para imputación por mediana.

5.3.2 Resultados

La Tabla 2 es aplicable también a los resultados de imputación por mediana ya que, aunque los valores imputados son ligeramente diferentes de un método a otro, el resultado final en ambos es que no quedan perdidos.

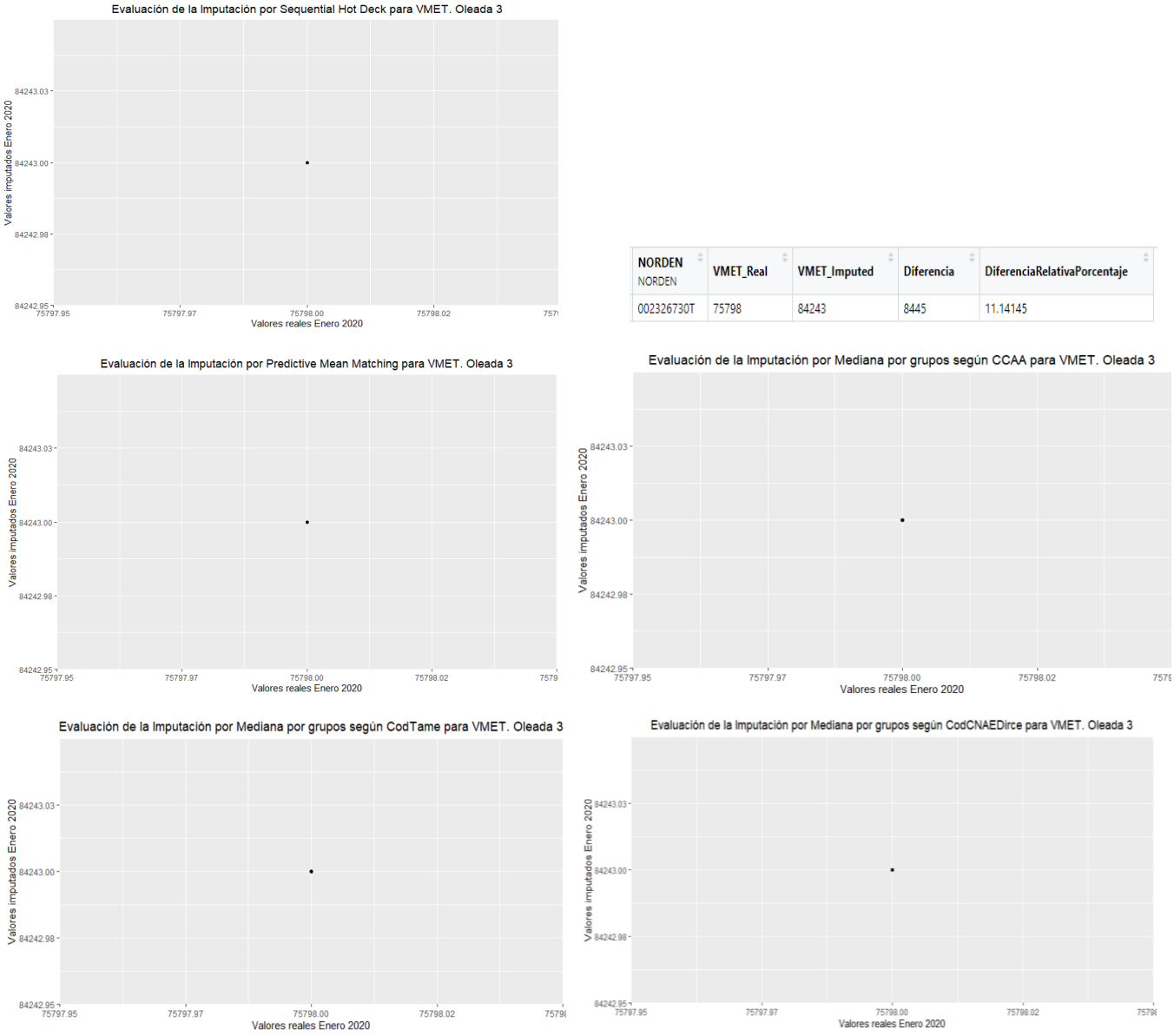
5.4 Evaluación de la calidad de la imputación

La finalidad de imputar datos pasados (Enero de 2020) es poder evaluar la calidad de la imputación realizada comparando estos valores con los reales, es decir, con los que el INE obtuvo para ese mes de ese año. Para ello, se han elaborado unas gráficas que plasman los valores reales (eje x) y los imputados (eje y) para cada tipo de imputación realizada. En otras palabras, se dibuja la unión poligonal de los puntos para su comparación con la recta y=x; cuanto más se parezca esta a una recta de pendiente 1, mejor habrá sido la imputación realizada. Además, se ha calculado una medida del error cometido al imputar: la diferencia relativa, es decir, (valor imputado-valor real)/valor real. El potencial de esta medida está en que el signo indica si se ha infraestimado o sobreestimado, de lo cual carecen las diferencias al cuadrado, por mencionar otra opción.

Se presentan a continuación los resultados por tipo de transporte calculados para la tercera oleada que es con la que se calculan los agregados que el INE publica en su página web. El resto de gráficos y tablas correspondientes a la primera y segunda oleada se pueden consultar en el Anexo V.

[30] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614

5.4.1 VMET



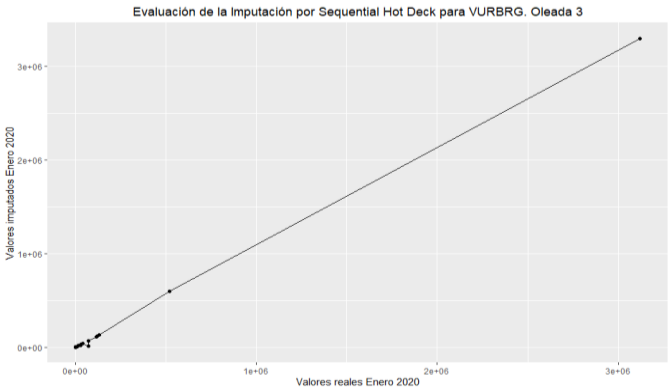
NORDEN NORDEN	VMET_Real	VMET_Imputed	Diferencia	DiferenciaRelativaPorcentaje
002326730T	75798	84243	8445	11.14145

Figura 13. Gráficos de valores reales frente a imputados por cinco métodos distintos para VMET en Enero de 2020

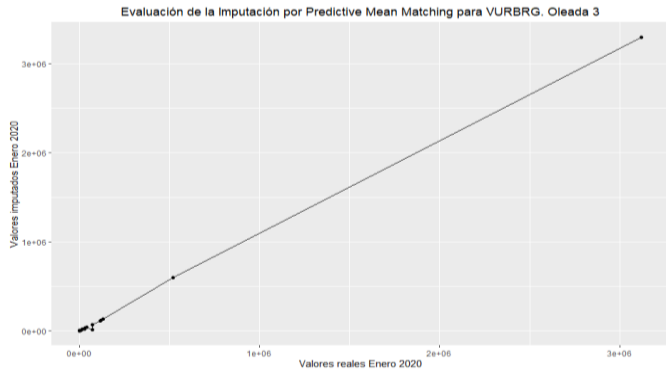
Para la variable VMET (Viajeros de Metro) no se pueden obtener conclusiones de la calidad de la imputación porque finalmente solo se imputó una unidad, por lo que no existe recta. Esto se debe a que en España solo existen 7 empresas de metro en las ciudades de Barcelona, Bilbao, Madrid, Málaga, Palma de Mallorca, Sevilla y Valencia.

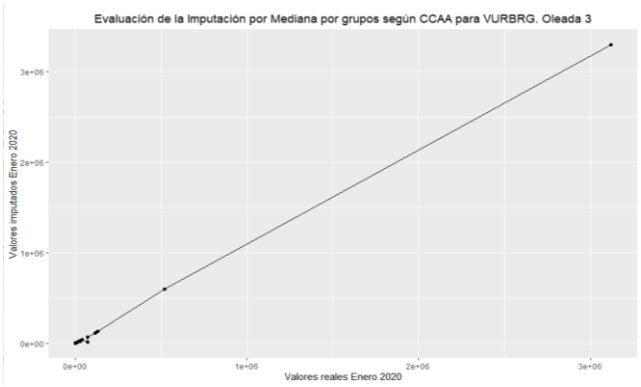
Sin embargo, numéricamente la tabla nos indica que hubo un error de 8445 viajeros (sobreestimación del 11,14%).

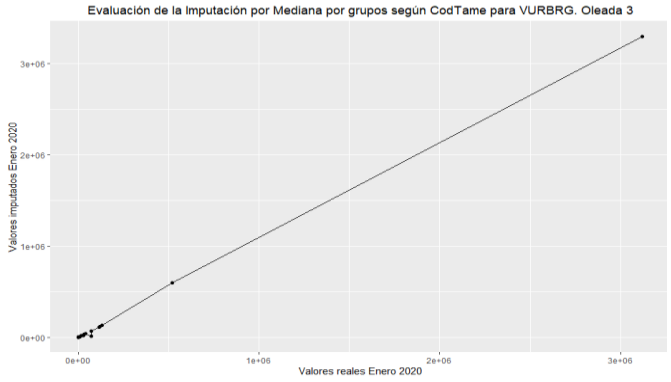
5.4.2 VURBRG



NORDEN NORDEN	VURBRG_Real	VURBRG_Imputed	Diferencia	DiferenciaRelativaPorcentaje
002010357T	17373	17179	-194	-1.1166753
002201533T	920	814	-106	-11.5217391
002630463T	*NA*	1600	*NA*	*NA*
0182290600	25648	21676	-3972	-15.4865876
0188663800	68849	10698	-58151	-84.4616480
0205689600	1391	1600	209	15.0251618
0208705900	2856	4291	1435	50.2450980
0223035500	282	270	-12	-4.2553191
0243488100	450	1870	1420	315.5555556
0306000400	14029	17548	3519	25.0837551
0904000100	39281	39015	-266	-0.6771722
096434511T	130451	129903	-548	-0.4200811
1104000100	119228	113936	-5292	-4.4385547
1104001100	10815	8499	-2316	-21.4147018
137472198T	186	262	76	40.8602151
1704000100	14959	15889	930	6.2169931







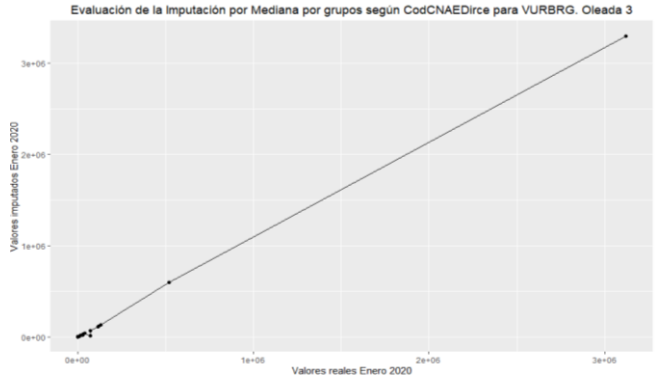
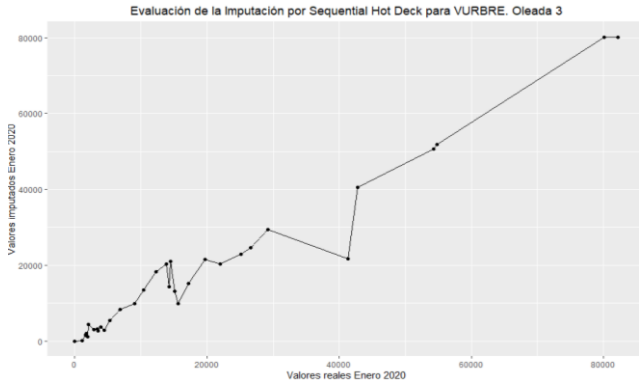


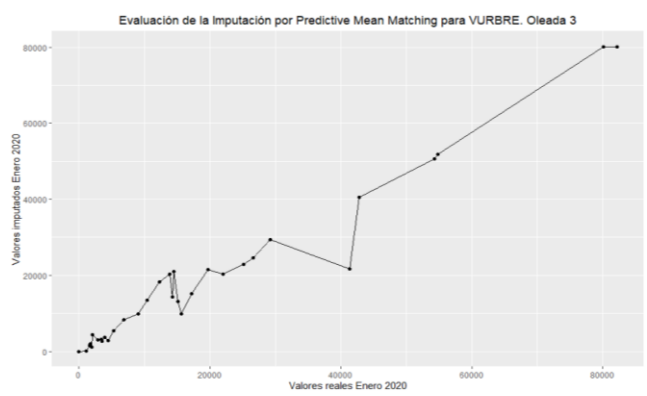
Figura 14. Gráficos de valores reales frente a imputados por cinco métodos distintos para VURBRG en Enero de 2020

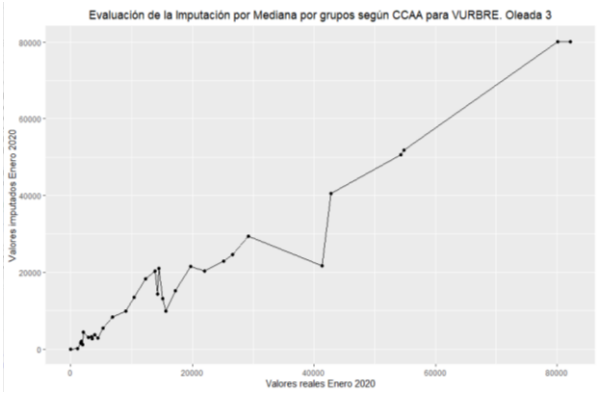
En los viajeros del urbano regular general observamos que todas las gráficas se aproximan bastante a la recta de pendiente 1, por lo que todos los métodos de imputación aplicados han sido igual de buenos. En la tabla de la Figura 14 se aprecia que casi todos los errores son por infraestimación (signo negativo en la columna DiferenciaRelativaPorcentaje), aunque de alguna manera los valores positivos han sido capaces de compensarlos; de otra forma las gráficas no arrojarían resultados tan buenos.

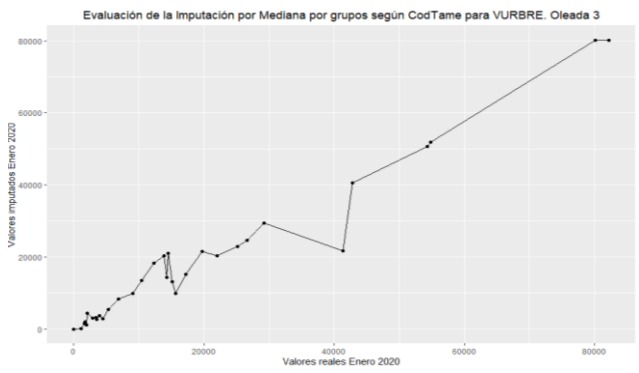
5.4.3 VURBRE



NORDEN NORDEN	VURBRE_Real	VURBRE_Imputed	Diferencia	DiferenciaRelativaPorcentaje
000889904T	19718	21555	1837	9.3163607
002010357T	14488	20936	6448	44.5057979
002030680T	2880	2996	116	4.0277778
002066901T	1584	1527	-57	-3.5984848
002125438T	26640	24660	-1980	-7.4324324
002228141T	1728	1987	259	14.9884259
002288718T	9000	9789	789	8.7666667
002573074T	2000	4420	2420	121.0000000
002630463T	NA	1987	NA	NA
0104272900	42845	40605	-2240	-5.2281480
0124252900	3461	3201	-260	-7.5122797
0156090400	0	0	0	NaN
0182290600	54793	51912	-2881	-5.2579709
0192206700	29220	29331	111	0.3798768
0201028100	14256	14256	0	0.0000000
0202090100	17208	15234	-1974	-11.4714086
0207319400	54288	50688	-3600	-6.6312997
0208705900	4495	2779	-1716	-38.1757508
0211459300	22040	20383	-1657	-7.5181488







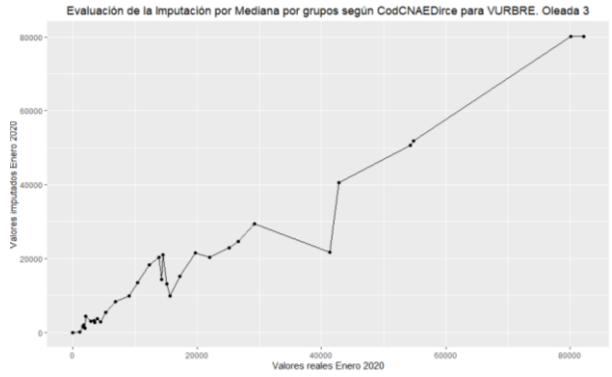
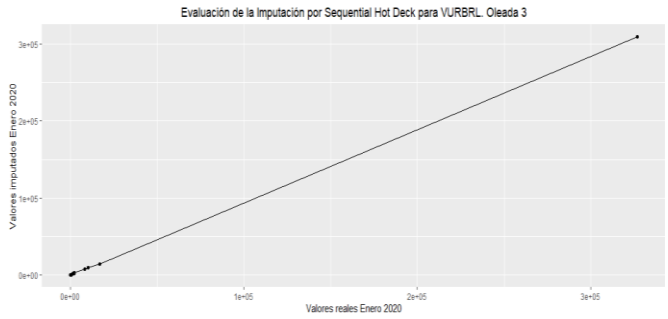


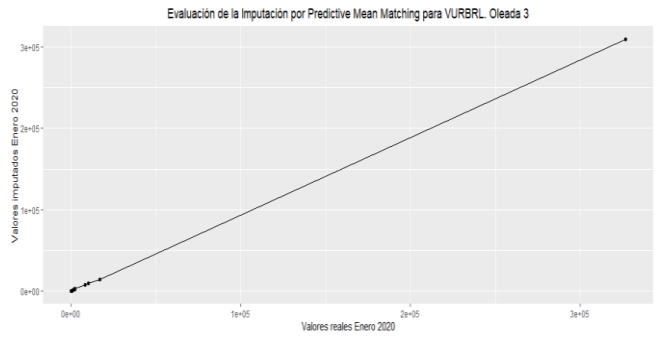
Figura 15. Gráficos de valores reales frente a imputados por cinco métodos distintos para VURBRE en Enero de 2020

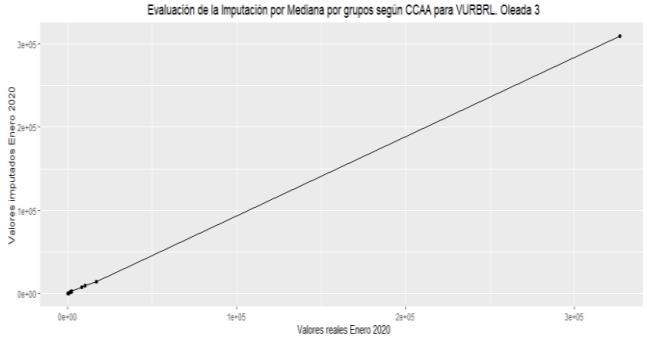
En el caso del transporte urbano regular especial escolar la imputación ha sido menos acertada. Las gráficas de la Figura 15 alcanzan cierta estabilidad a partir de, aproximadamente, 45 000 viajeros, lo cual podría indicar que la imputación es mejor en empresas con mayor volumen de viajeros, bien porque estas responden más, bien porque cuentan con más miembros en su grupo y la imputación por donantes y por mediana con grupos de CodTame se vuelve más robusta (aunque la primera hipótesis parece más probable porque las gráficas de todos los tipos de imputación son idénticas).

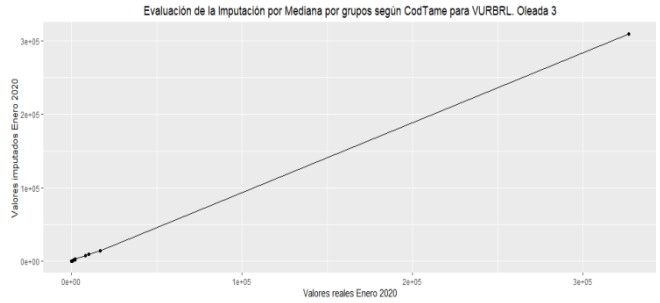
5.4.4 VURBRL



NORDEN NORDEN	VURBRL_Real	VURBRL_Imputed	Diferencia	DiferenciaRelativaPorcentaje
000889904T	2232	3258	1026	45.9677419
002010357T	612	730	118	19.2810458
002288718T	8140	8085	-55	-0.6756757
0124252900	2027	2053	26	1.2826838
0220390500	991	1473	482	48.6377397
0229410600	*NA*	21400	*NA*	*NA*
0808004800	10319	9891	-428	-4.1476887
2806003400	0	0	0	NaN
3004000800	1656	2139	483	29.1666667
3108001300	16800	14953	-1847	-10.9940476
5006001800	326700	309600	-17100	-5.2341598







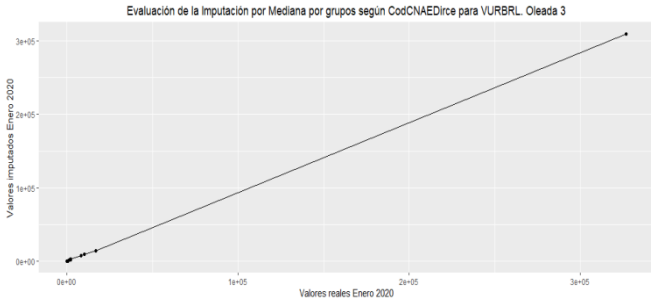
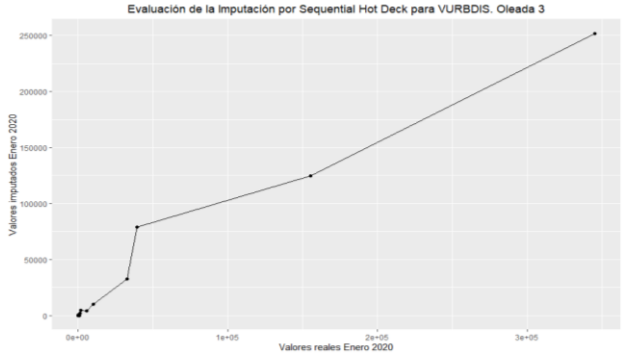


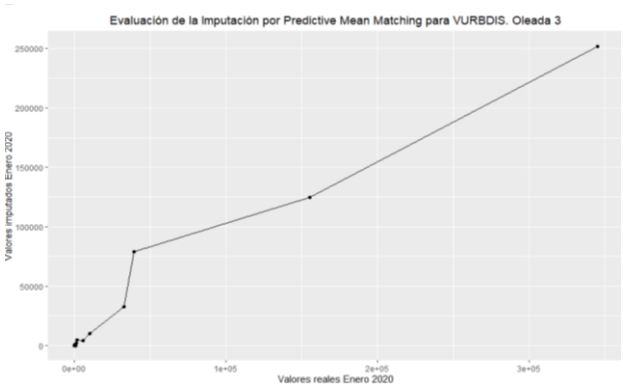
Figura 16. Gráficos de valores reales frente a imputados por cinco métodos distintos para VURBRL en Enero de 2020

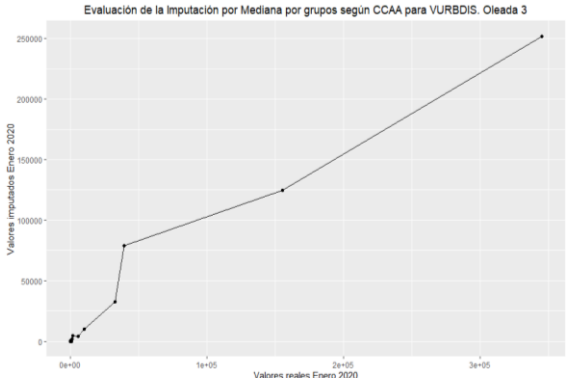
La variable Viajeros de urbano regular especial laboral ha sido perfectamente imputada según las gráficas de la Figura 16. Sin embargo, la tabla de esa misma figura nos muestra que se han dado tanto sobreestimaciones como subestimaciones. De ahí que se haya complementado con una tabla; la gráfica no siempre es totalmente ilustrativa, entre otras cosas porque nos movemos en cifras muy dispares (de 0 a 300 000 viajeros) y la mayoría de los datos se concentran en una parte de la gráfica.

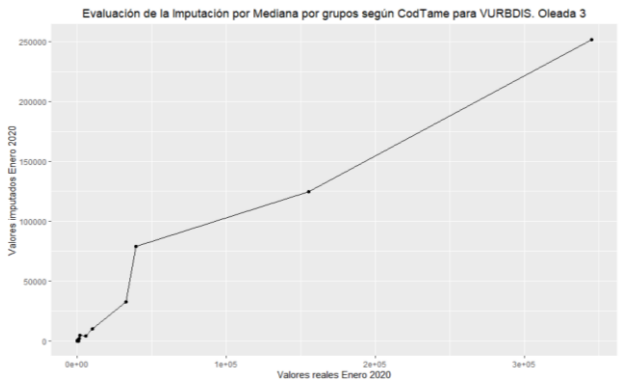
5.4.5 VURBDIS



NORDEN NORDEN	VURBDIS_Real	VURBDIS_Imputed	Diferencia	DiferenciaRelativaPorcentaje
002010357T	776	886	110	14.1752577
002125438T	990	1305	315	31.8181818
002279469T	0	0	0	NaN
0124252900	1957	4393	2436	124.4762391
0156090400	10320	10212	-108	-1.0465116
0211459300	919	0	-919	-100.0000000
0214556000	5542	4011	-1531	-27.6254060
0220390500	912	1356	444	48.6842105
0220862700	960	618	-342	-35.6250000
0225549100	62	25	-37	-59.6774194
0243773900	1338	2145	807	60.3139013
0808004800	39341	78902	39561	100.5592130
106506553T	32594	32701	107	0.3282813
116927237T	92	68	-24	-26.0869565
168483363T	840	578	-262	-31.1904762
2806003400	0	0	0	NaN
2908010000	154982	124789	-30193	-19.4816172







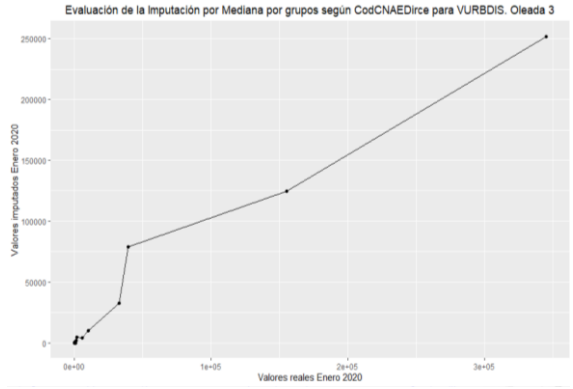
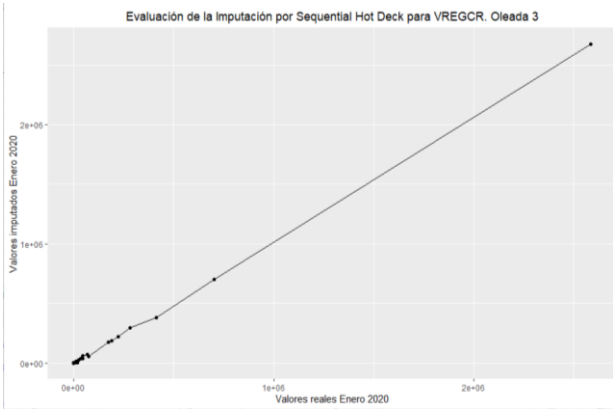


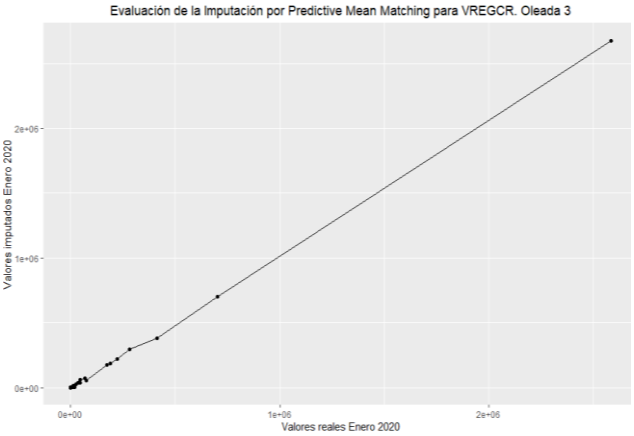
Figura 17. Gráficos de valores reales frente a imputados por cinco métodos distintos para VURBDIS en Enero de 2020

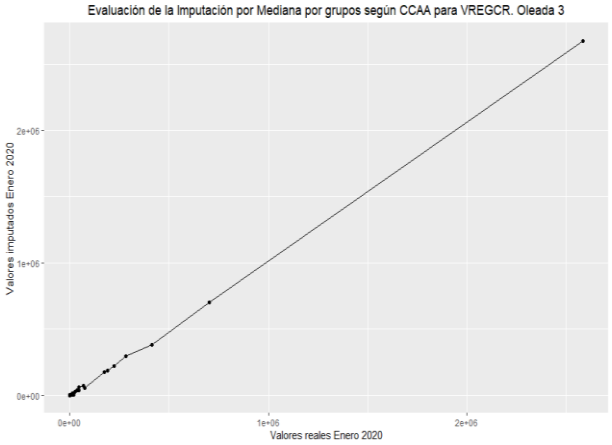
Para los viajeros de urbano discrecional la estimación ha sido bastante semejante a la realidad salvo por una unidad que evita que las gráficas de la Figura 17 sean una recta de pendiente 1. Como en casos anteriores observamos que la tabla de esa figura muestra errores relativos bastantes altos (columna DiferenciaRelativaPorcentaje) pero que se han compensado los unos a los otros y de nuevo todos los métodos de imputación muestran gráficas casi idénticas, asunto que se discutirá en las Conclusiones y Trabajo Futuro.

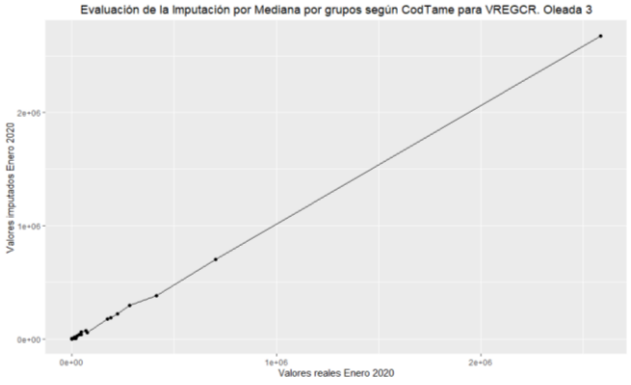
5.4.6 VREGCR



NORDEN NORDEN	VREGCR_Real	VREGCR_Imputed	Diferencia	DiferenciaRelativaPorcentaje
001799025T	1418	1563	145	10.22566996
001914276T	1214	1086	-128	-10.54365733
001915959T	345	386	41	11.88405797
001987423T	1445	852	-593	-41.03806228
002072173T	228	249	21	9.21052632
002201533T	2528	2488	-40	-1.58227848
002228141T	45	63	18	40.00000000
0182290600	18320	1194	-17126	-93.48253275
0194565600	45440	63278	17838	39.25616197
0201028100	1065	1176	111	10.42253521
0205450400	4162	3732	-430	-10.33157136
0205689600	2888	2988	100	3.46260388
0206721500	26015	25134	-881	-3.38650778
0207434100	352	162	-190	-53.97727273
0212281500	217	270	53	24.42396313
0215634600	11488	13328	1840	16.01671309
0223035500	44052	38277	-5775	-13.10950695







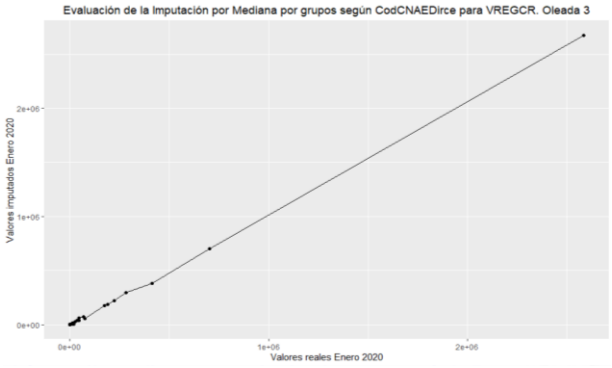
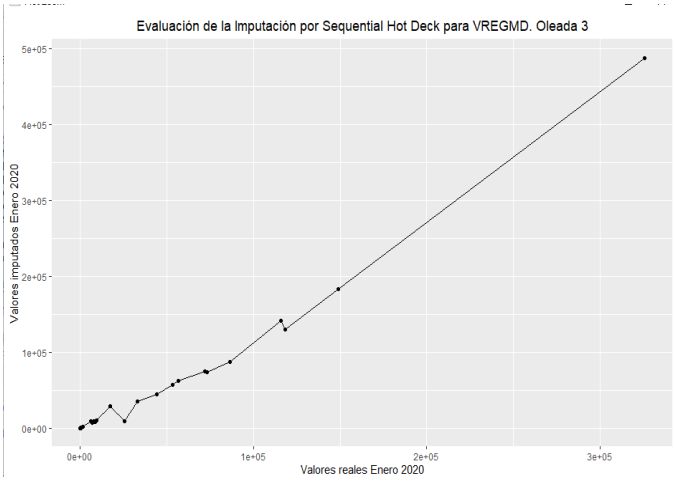


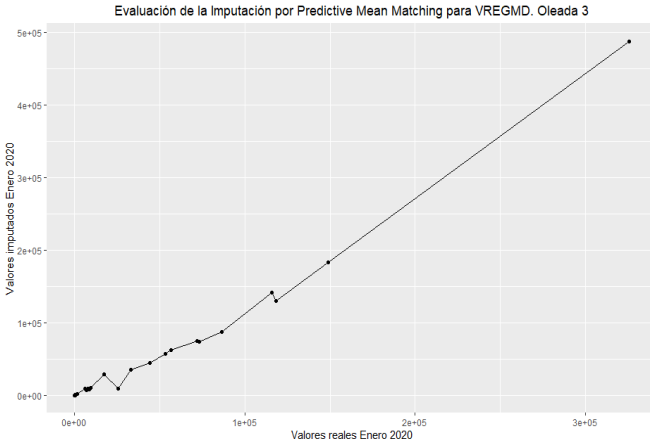
Figura 18. Gráficos de valores reales frente a imputados por cinco métodos distintos para VREGCR en Enero de 2020

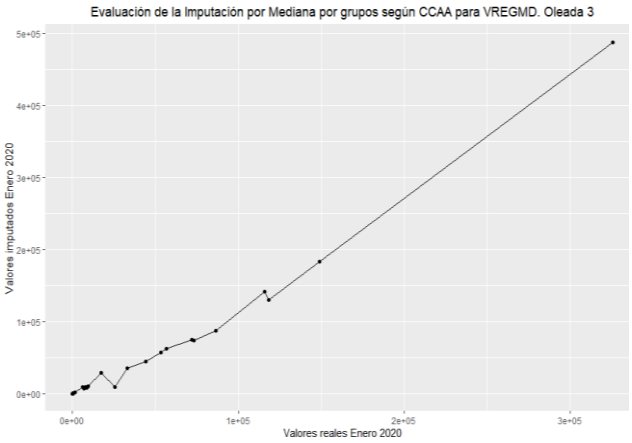
La variable Viajeros de interurbano regular cercanías no presenta, en cuanto a calidad de la imputación realizada, ninguna diferencia con casos mencionados anteriormente (ver Figura 18) y, por tanto, no requiere un amplio comentario.

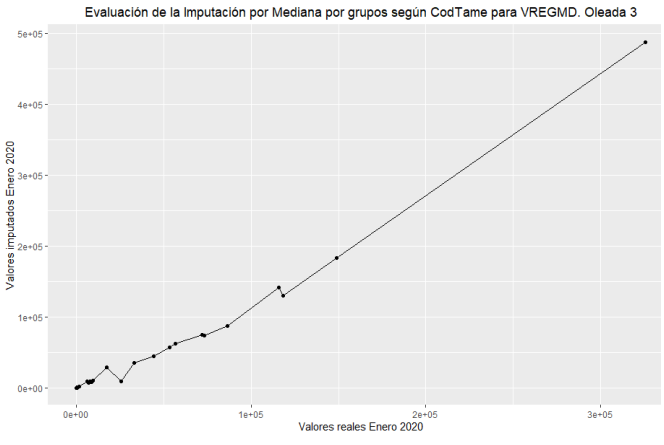
5.4.7 REGMD



NORDEN NORDEN	VREGMD_Real	VREGMD_Imputed	Diferencia	DiferenciaRelativaPorcentaje
000602528T	1560	1428	-132	-8.4615385
001799025T	238	352	114	47.8991597
001933243T	17415	28989	11574	66.4599483
001987423T	7773	8624	851	10.9481539
002191322T	33164	34820	1656	4.9933663
0191921100	223	262	39	17.4887892
0197800900	6756	6849	93	1.3765542
0208705900	6050	8637	2587	42.7603306
0209228700	388	420	32	8.2474227
0218629400	86293	86735	442	0.5122084
0223035500	94	82	-12	-12.7659574
023161413T	115637	141096	25459	22.0163097
0606000400	56750	61963	5213	9.1859031
0647205100	17	43	26	152.9411765
1271014100	118164	129975	11811	9.9954301
137472198T	8576	7723	-853	-9.9463619
1506001500	53257	56960	3703	6.9530766







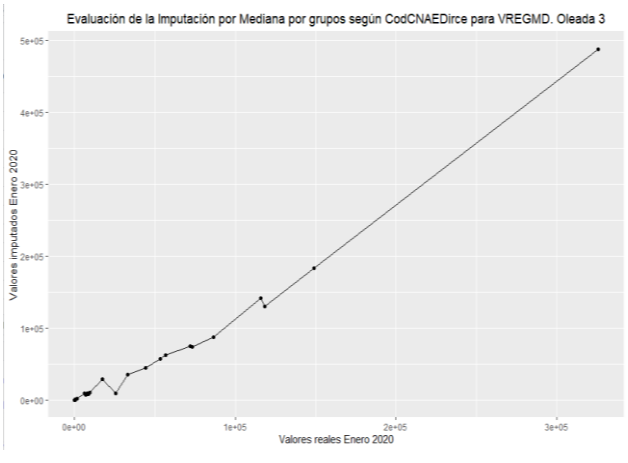
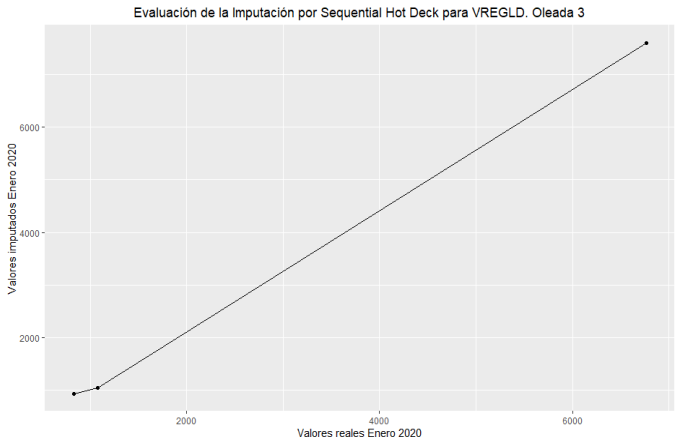
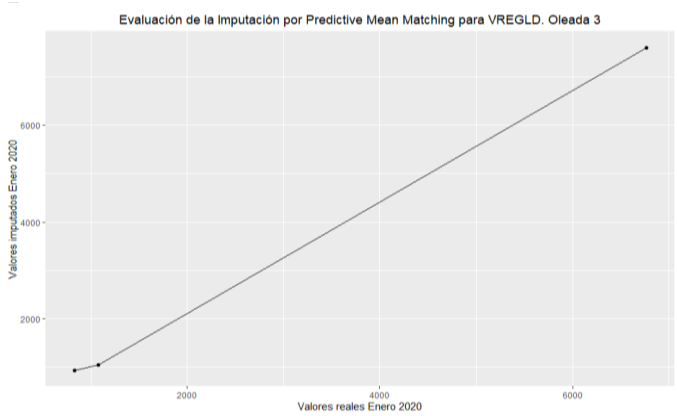


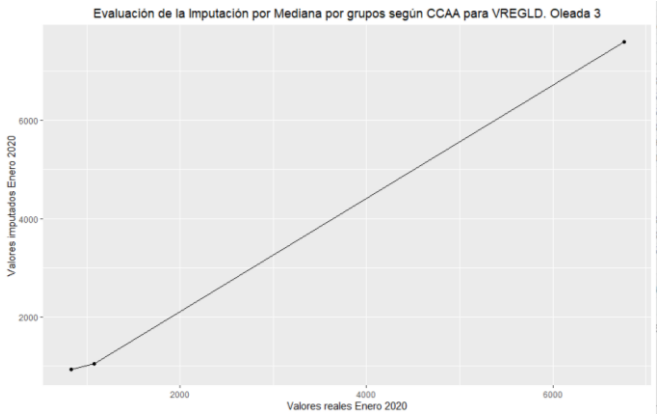
Figura 19. Gráficos de valores reales frente a imputados por cinco métodos distintos para VREGMD en Enero de 2020

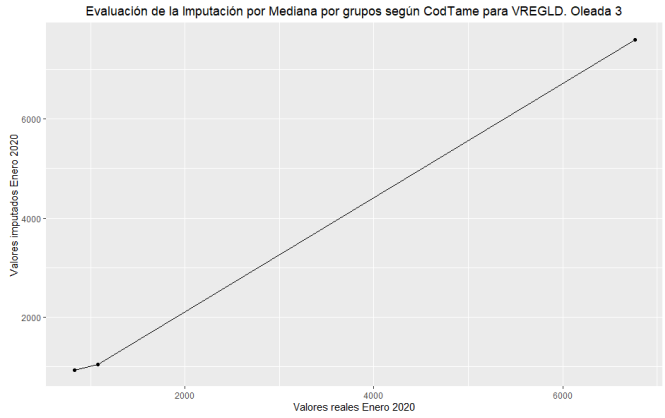
5.4.8 REGLD



NORDEN NORDEN	VREGLD_Real	VREGLD_Imputed	Diferencia	DiferenciaRelativaPorcentaje
023161413T	6763	7602	839	12.405737
2806004800	1078	1049	-29	-2.690167
4306001400	827	938	111	13.422007







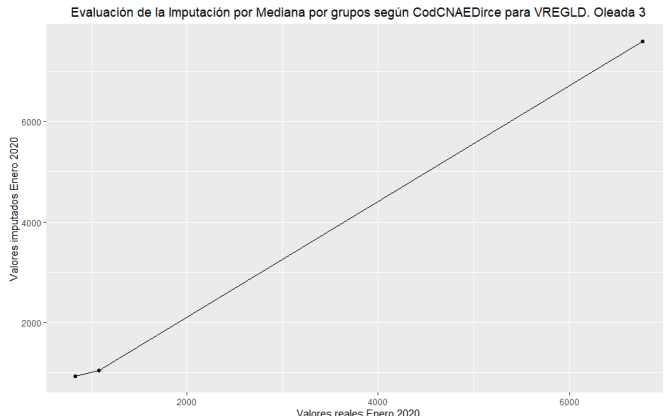
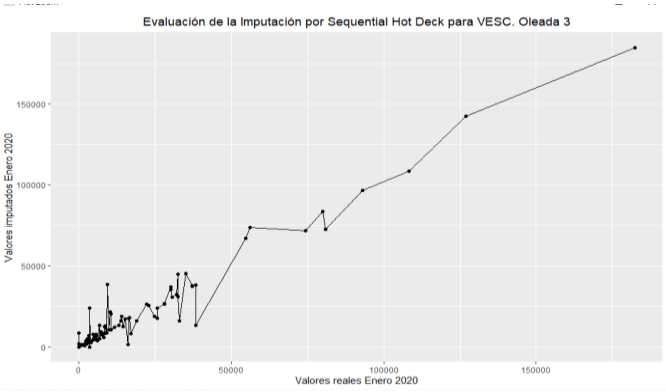
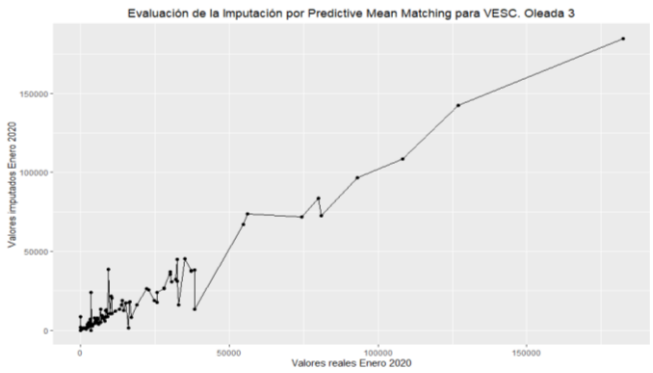


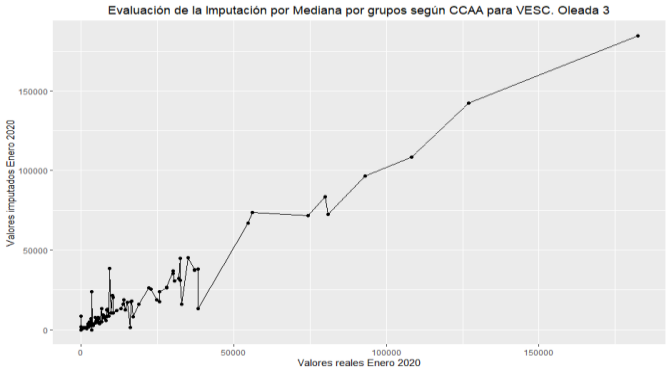
Figura 20. Gráficos de valores reales frente a imputados por cinco métodos distintos para VREGLD en Enero de 2020

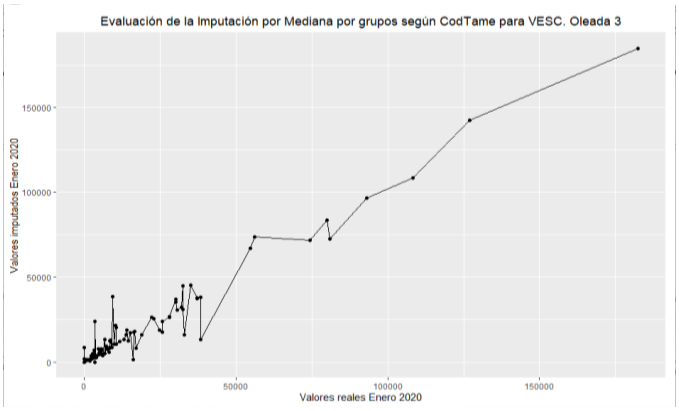
5.4.9 VESC



NORDEN NORDEN	VESC_Real	VESC_Imputed	Diferencia	DiferenciaRelativaPorcentaje
000602528T	5172	7056	1884	36.426914
000661698T	6840	5140	-1700	-24.853801
001131149T	8624	13002	4378	50.765306
001298991T	4658	4032	-626	-13.439244
001307830T	32480	44711	12231	37.657020
001345609T	3570	4082	512	14.341737
001799025T	8496	8676	180	2.118644
001808746T	6660	13320	6660	100.000000
001841744T	1872	798	-1074	-57.371795
001913539T	2376	2500	124	5.218855
001914276T	16434	17748	1314	7.995619
001915527T	2412	2339	-73	-3.026534
001915758T	3950	2700	-1250	-31.645570
001915959T	7416	8009	593	7.996224
001956980T	3640	3096	-544	-14.945055
001984634T	3150	2700	-450	-14.285714
001987423T	14571	12304	-2267	-15.558301







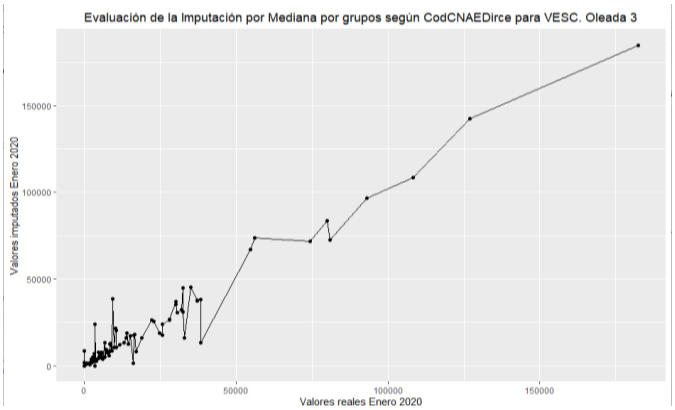
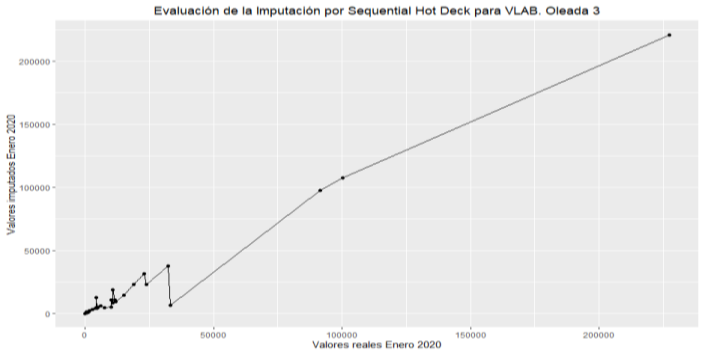


Figura 21. Gráficos de valores reales frente a imputados por cinco métodos distintos para VESC en Enero de 2020

Los viajeros de interurbano regular especial escolar han sido bastante bien imputados al principio de la serie, es decir, donde existen muchas observaciones con un volumen de viajeros relativamente pequeño (ver Figura 21). No son las gráficas más semejantes a una recta de pendiente 1 de todas las que se han mostrado, pero en general se puede decir que la imputación parece haber sido buena ya que este tipo de transporte tiene bastantes más observaciones que otros y, aún así, la gráfica se asemeja a una recta con la excepción de 3 observaciones en torno a los 50 000 viajeros.

5.4.10 VLAB



NORDEN NORDEN	VLAB_Real	VLAB_Imputed	Diferencia	DiferenciaRelativaPorcentaje
001307830T	10185	10907	722	7.0888562
001799025T	4357	4826	469	10.7642874
002201533T	1700	2228	528	31.0588235
002403779T	4978	4750	-228	-4.5801527
0108000400	23125	31702	8577	37.0897297
0108005900	91481	97940	6459	7.0604825
012697223T	570	1268	698	122.4561404
0191921100	11660	10816	-844	-7.2384220
0192206700	11840	9510	-2330	-19.6790541
0205450400	10740	8340	-2400	-22.3463687
0205689600	4110	4110	0	0.0000000
0206327500	3000	3360	360	12.0000000
0206721500	6042	6080	38	0.6289308
0213942500	5044	5044	0	0.0000000
0217481200	33133	6650	-26483	-79.9293755
0219371100	10116	5406	-4710	-46.5599051
0223035500	18870	23185	4315	22.8669846

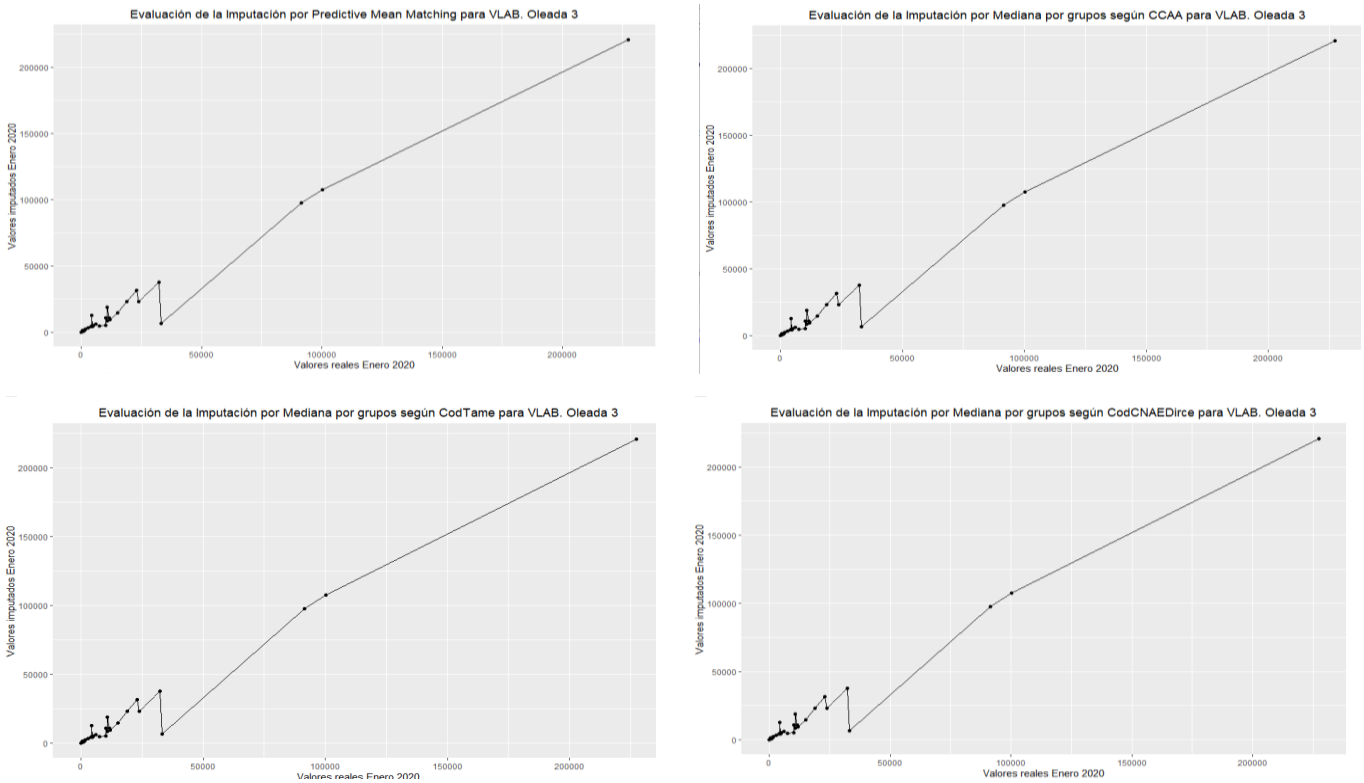
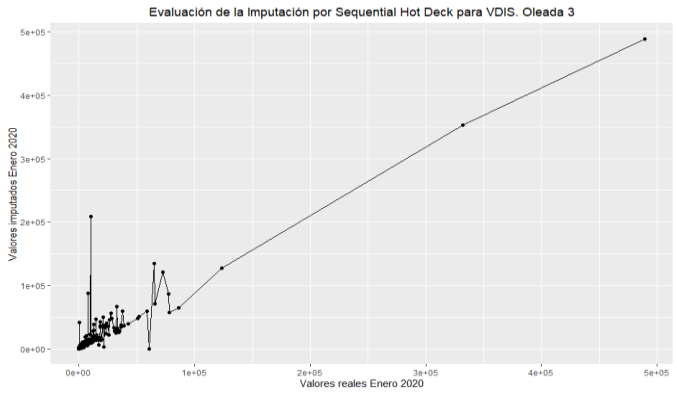


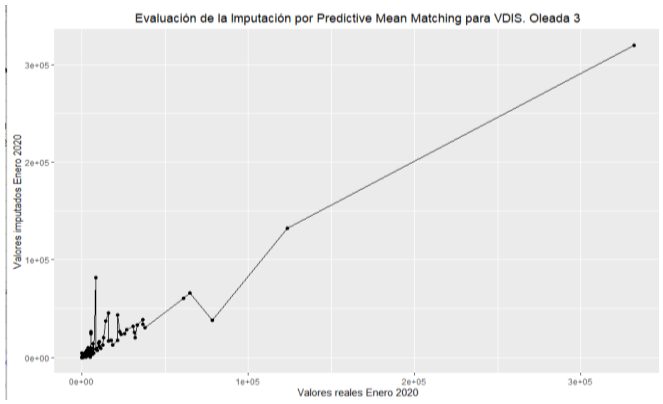
Figura 22. Gráficos de valores reales frente a imputados por cinco métodos distintos para VLAB en Enero de 2020

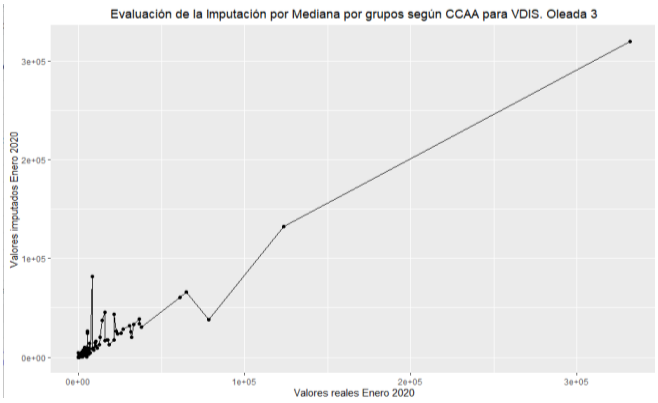
Las gráficas de la Figura 22, correspondientes a viajeros de interurbano regular especial laboral, no se ajustan perfectamente a una recta de pendiente 1 probablemente debido al efecto de las vacaciones de Navidad cuya duración es diferente según la empresa –por lo que la elección de donante es muy delicada en este caso.

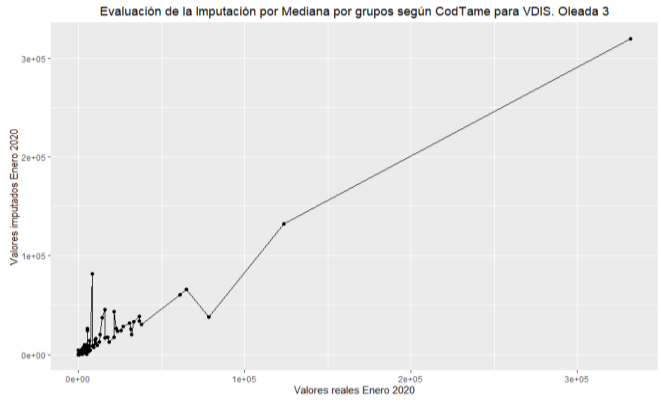
5.4.11 VDIS



NORDEN NORDEN	VDIS_Real	VDIS_Imputed	Diferencia	DiferenciaRelativaPorcentaje
000228349T	500	2146	1646	329.2000000
000321657T	864	288	-576	-66.6666667
000383820T	1286	2670	1384	107.6205288
000602528T	2094	6427	4333	206.9245463
000661698T	735	1230	495	67.3469388
000777816T	1600	4108	2508	156.7500000
000889904T	10078	208730	198652	1971.1450685
001131149T	21371	2420	-18951	-88.6762435
001190294T	3102	6302	3200	103.1592521
001253763T	4641	4682	41	0.8834303
001284716T	1368	2511	1143	83.5526316
001298991T	343	1266	923	269.0962099
001300563T	4229	10559	6330	149.6807756
001307830T	8100	87276	79176	977.4814815
001318084T	380	206	-174	-45.7894737
001345609T	1780	447	-1333	-74.8876404
001525260T	180	1177	997	553.8888889







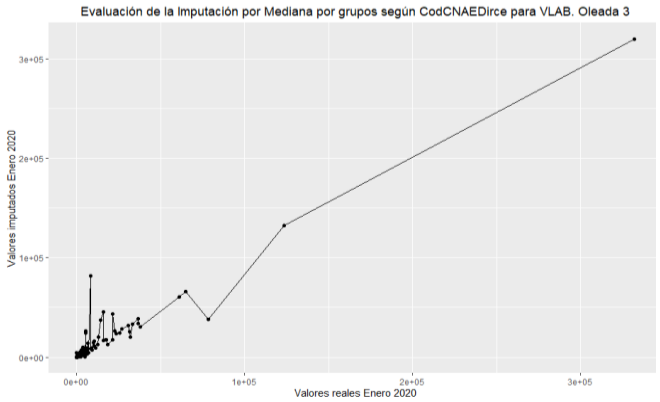


Figura 23. Gráficos de valores reales frente a imputados por cinco métodos distintos para VDIS en Enero de 2020

En el caso del transporte interurbano discrecional observamos en las gráficas que uno de los tipos de imputación arroja un resultado ligeramente distinto del resto de los métodos. En general la imputación en este caso es de buena calidad aunque observemos bastantes casos de sobreestimación en la tabla de la Figura 23 a causa del efecto de los días festivos en Enero.

6. Conclusiones y trabajo futuro

En un trabajo de producción estadística, lo escrito nunca reflejará suficientemente bien todo el trabajo de investigación que hay detrás. Con el fin de acercar un poco más al lector a la realidad de las tareas realizadas se adjunta a este archivo el código en R empleado.

En conclusión, se ha diseñado un sistema de imputación multietápico y con diferentes métodos de imputación en el que se recurre a árboles de decisión (con un complejo sistema de regresores detrás), donantes (utilización de registros anteriores al periodo a imputar) y medianas (calculadas para grupos formados en base a 3 variables distintas todas ellas estrechamente relacionadas con la variable objetivo). Como resultado se ha obtenido un archivo sin valores perdidos de calidad relativamente buena y con microdatos listos para el cálculo de agregados. Es decir, gracias a que se ha imputado un periodo pasado se han podido comparar los valores imputados con los reales y hemos observado que, en general, las gráficas y la medida de error calculada (diferencia relativa) nos indican que los resultados son bastantes buenos. No parece existir un patrón por "grupos", es decir, no podemos hablar de que la imputación es mejor en los transportes urbanos que en los interurbanos o en los discrecionales frente al resto, etc.

También en las gráficas de evaluación de la calidad de la imputación hemos observado que todos los métodos de imputación reflejan resultados iguales a nivel visual (lo cual no significa que los valores imputados sean idénticos se use el método que se use, aunque sí similares). Esto indica que no puedo concluir cual de los métodos es mejor, para ello sería necesario hacer el cálculo de agregados y repetir el proceso para todos los meses del años, lo cual va más allá del alcance de este TFM que tenía como objeto la imputación microdato y es lo que se ha hecho. Por el mismo motivo –carecer de agregados- no se puede ofrecer para cada tipo de transporte una medida global del error cometido (por ejemplo, una media de las diferencias relativas). Esto se debe a que la diferencia relativa se corresponde con cada unidad y tipo de transporte porque las unidades tienen un volumen muy distinto de viajeros. Es decir, un error de 5 000 viajeros en el metro de Palma cambia radicalmente los resultados, lo cual no ocurriría con el metro de Madrid porque tiene un volumen de viajeros muy superior.

Como cierre de conclusiones, cabría recalcar que en este TFM se ha planteado como objetivo desde mis prácticas en el INE que fuese un proyecto eminentemente práctico y que pudiese aportar algo valioso a las instituciones productoras de Estadísticas Oficiales. Se puede concluir que esta aportación ha sido, por una parte, los elementos conductores a la estandarización (en el plano teórico, la construcción de las fases teniendo en cuenta el gráfico de flujo de datos propuesto por el GSBPM y GSIM y, en el plano práctico, uso de software libre) y, por otra, el diseño de una metodología basada en dos fases: un árbol de decisión para definir las unidades imputables y una segunda fase de imputación por diversos métodos –siendo uno de ellos la versión mejorada de lo que se hace actualmente en el INE, al pasar de usar la imputación por media a imputación por mediana.

Como líneas de trabajo futuro cabría afinar el algoritmo del árbol de decisión eliminando los regresores con menos importancia (por una cuestión de parsimonia, ya que en este tipo de algoritmos no lineales la presencia de regresores "poco útiles" no genera problemas) y, sobre todo, diseñando otros alternativos –especialmente si se pretende estandarizar el procedimiento y aplicar este sistema de imputación a otras encuestas que necesariamente dependerán de otras variables y, por ende, de otros regresores. En la segunda etapa (siguiendo con la metodología diseñada en este trabajo) cabría mejorar la regresión aumentando la robustez, ya que los modelos siempre se pueden perfeccionar. Una posible vía para lograr esto sería incluir series temporales.

Por otra parte, sería conveniente aplicar todo el procedimiento (tanto la fase de árbol de decisión como la de imputación en sí misma) a otra serie de variables presentes en la encuesta y en las que también se detectan valores perdidos como son la variable ingresos y la variable personal; de esta manera, la propia imputación de viajeros mejoraría porque se podría tener en cuenta las variables ingresos y personal en el modelo de regresión. Asimismo, centrándonos en la fase 2 –de imputación propiamente dicha– sería interesante como trabajo futuro recurrir a otros métodos diferentes de la imputación por donantes y la mediana –y que han sido brevemente explicados en la subsección "paquete *simputation*"– y realizar el cálculo de agregados para tener una idea más precisa de la calidad de la imputación realizada.

Por otra parte, teniendo en cuenta la voluntad estandarizadora del proyecto y el propósito de que llegue a tantos trabajadores de las Estadísticas Oficiales como sea

posible, habría que redactar un documento equivalente al presente pero en inglés y redactado en un formato más conciso, como una especie de manual.

6.1 Limitaciones del proyecto

La situación sanitaria de COVID-19 ha afectado gravemente a las empresas, suponiendo el cierre temporal o definitivo de muchas ellas, lo cual ha implicado que la tasa media anual de no respuesta por ítem en la ETV aumentase del 8,96% en 2019 al 17,4% en 2020[31]. Debemos tener en cuenta que estas tasas ofrecidas por el INE son no ponderadas, por lo tanto, no se muestra que son las empresas pequeñas las que han dejado de responder, como se observa en los gráficos de la sección Evaluación de la calidad de la imputación, que muestran que las rectas dibujadas se separan más de la línea de pendiente 1 en las empresas con menor número de viajeros. Luego la falta de respuesta está sesgada porque depende del tamaño de la empresa, pero la imputación se realiza de igual manera en todas las unidades. Partiendo de la base de que la imputación recurre a valores previos al periodo de referencia en cuestión y/o a unidades similares en cierta característica a aquella que no ha respondido, el hecho de que la tasa de no respuesta haya aumentado tanto supone una limitación muy importante porque no podemos fijarnos en otras unidades para imputar un valor similar (porque muchas unidades han dejado de responder) y tampoco podemos imputar siguiendo la tendencia de respuestas ofrecidas hasta el momento (porque la situación económica ha cambiado radicalmente). Por tanto, aunque este trabajo no se ve afectado por la pandemia, su aplicación futura sí se verá limitada hasta que la situación se estabilice.

Una segunda limitación ha sido la complejidad de trabajar con un archivo tan extenso: 33 variables (7 de identificación y organización interna, 11 de tipos de transporte, 11 asociadas a los ingresos y 4 de personal) que llegaron a ser más de 200 cuando se empezaron a calcular los regresores (18 regresores para cada uno de los 11 tipos de transporte más otros añadidos) y más de 200 000 observaciones de 500 empresas de transporte a lo largo de 11 años –lo cual supone una limitación añadida puesto que la forma de nombrar las variables desde 2009 ha ido variando, lo cual pone de manifiesto la necesidad de estandarizar para sacar el máximo provecho de las series temporales-.

[31] https://www.ine.es/dynt3/metadatos/es/RespuestaDatos.html?oe=30163

Por último, cabe destacar un problema recurrente al que se enfrentan las Estadísticas Oficiales que también está presente en este TFM y es la modelización de variables "semicontinuas" que son aquellas que tienen valor 0 o un valor continuo, como se explicaba en el apartado variables. A priori esto parece ajeno en un trabajo dedicado a la imputación de viajeros (puesto que los viajeros son personas y, por ende, unidades enteras), pero lo cierto es que se trata de valores estimados que presentan la forma de números continuos.

7. Referencias bibliográficas

Comisión Europea (2014). *Automatic Editing*. Recuperado el 8 de junio de 2021 de: https://ec.europa.eu/eurostat/cros/content/automatic-editing-method_en

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22.

Eurostat. (2017). *Código de buenas prácticas de las estadísticas europeas*. Recuperado el 8 de junio de 2021 de: https://ec.europa.eu/eurostat/documents/4031688/9394048/KS-02-18-142-ES-N.pdf/e792b761-6f09-42a9-a1e0-3a3356a0de1c#:~:text=El%20C%C3%B3digo%20de%20Buenas%20Pr%C3%A1cticas%20de%20las%20Estad%C3%ADsticas%20Europeas%20es,estad%C3%ADsticos%20y%20la%20producci%C3%B3n%20estad%C3%ADstica.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Kowarik, A. (6 de noviembre de 2019). *Use of R in Official Statistics - uRos2020 - CROS - European commission*. Recuperado el 23 de mayo de 2021 de: https://ec.europa.eu/eurostat/cros/content/use-r-official-statistics-uros2020_en

Muñoz Pichardo, J. M. & del Valle Benvides, A. R. (junio de 2020). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones.* Recuperado el 1 de junio de https://idus.us.es/bitstream/handle/11441/63201/Valle%20Benavides%20Ana%20Roc%C3%ADo%20del%20TFG.pdf?sequence=1

Naciones Unidas. (2014). *Principios Fundamentales de las Estadísticas Oficiales*. Recuperado el 25 de abril de https://unstats.un.org/unsd/dnss/hb/S-fundamental%20principles_A4-WEB.pdf

O'connor, L. (20 de enero de 2015). *Step 4: Imputation of missing data*. Recuperado el 31 de mayo de 2021 de: https://ec.europa.eu/jrc/en/coin/10-step-guide/step-4

Olsen, M., & Schafer, J. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association, 96*(454), 730-745. Recuperado el 17 de junio de 2021 de: http://www.jstor.org/stable/2670310

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

Scholtus, S. (2013). *Imputation - Main Module (Theme) - CROS - European Commission*. Recuperado el 23 de mayo de 2021 de: https://ec.europa.eu/eurostat/cros/content/imputation-main-module-theme_en

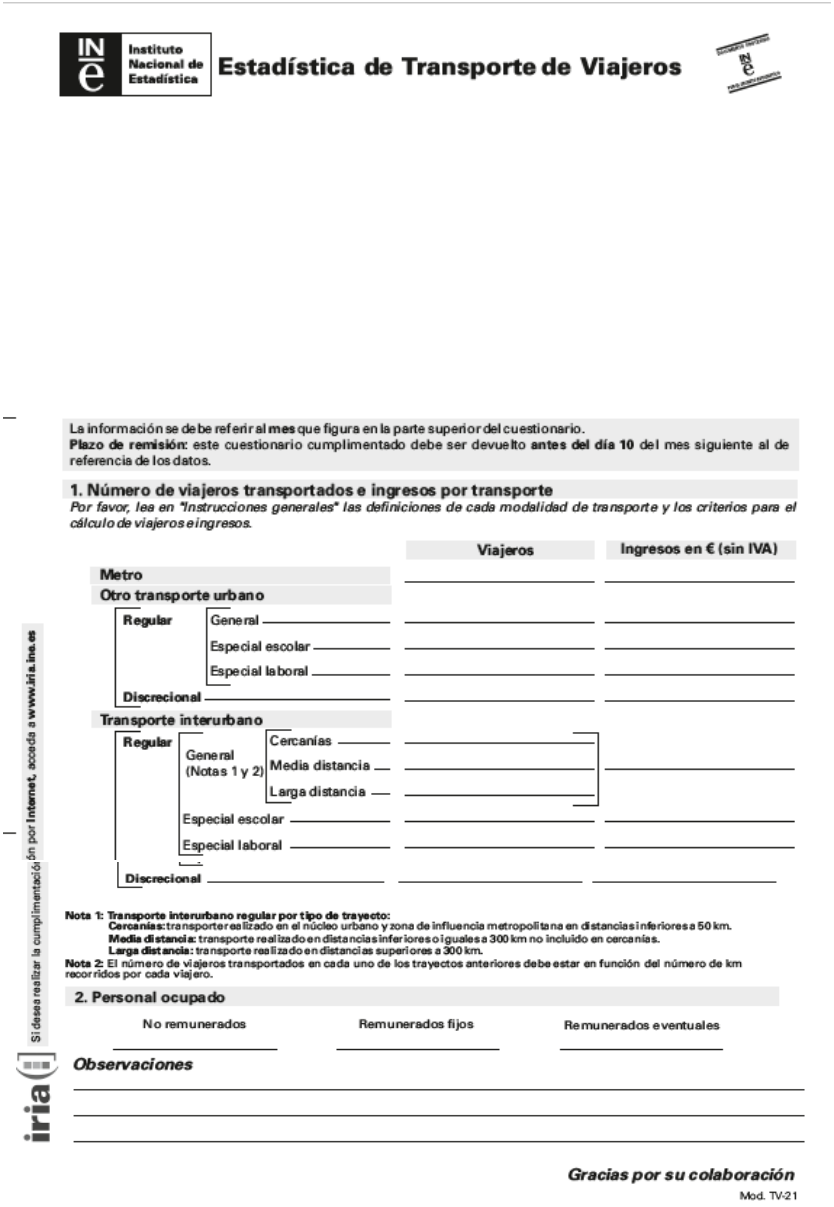
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Van der Loo, M. (2021). *Simple Imputation [R package simputation version 0.2.6]*. Recuperado el 28 de marzo de: https://cran.r-project.org/web/packages/simputation/

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (Vol. 563). John Wiley & Sons.

8. Anexos

8.1 Anexo I: Cuestionario íntegro de la ETV, año 2021



Instituto Nacional de Estadística

Estadística de Transporte de Viajeros

Si desea realizar la cumplimentación por **Internet**, acceda a **www.iria.ine.es**

La información se debe referir al **mes** que figura en la parte superior del cuestionario.
Plazo de remisión: este cuestionario cumplimentado debe ser devuelto **antes del día 10** del mes siguiente al de referencia de los datos.

1. Número de viajeros transportados e ingresos por transporte

Por favor, lea en "Instrucciones generales" las definiciones de cada modalidad de transporte y los criterios para el cálculo de viajeros e ingresos.

			Viajeros	Ingresos en € (sin IVA)
Metro				
Otro transporte urbano				
Regular	General			
	Especial escolar			
	Especial laboral			
Discrecional				
Transporte interurbano				
Regular	General (Notas 1 y 2)	Cercanías		
		Media distancia		
		Larga distancia		
	Especial escolar			
	Especial laboral			
Discrecional				

Nota 1: Transporte interurbano regular por tipo de trayecto:
Cercanías: transporte realizado en el núcleo urbano y zona de influencia metropolitana en distancias inferiores a 50 km.
Media distancia: transporte realizado en distancias inferiores o iguales a 300 km no incluido en cercanías.
Larga distancia: transporte realizado en distancias superiores a 300 km.
Nota 2: El número de viajeros transportados en cada uno de los trayectos anteriores debe estar en función del número de km recorridos por cada viajero.

2. Personal ocupado

No remunerados	Remunerados fijos	Remunerados eventuales

Observaciones

Gracias por su colaboración

Mod. TV-21

Modificaciones en la identificación *(Cumplimente sólo los apartados sujetos a variación)*

Nombre o razón social de la empresa — NIF

Domicilio Social (calle, plaza, paseo, avenida...)

Código Postal — Municipio — Provincia

Teléfono — Fax — Url

Persona de contacto a quien dirigirse para aclaraciones sobre este cuestionario

D./Dña. _____ Tfno. _____ E-mail _____

Instrucciones generales

Tipo de transporte

Metro: Recoge los viajeros transportados por las compañías metropolitanas.

Otro transporte urbano: El que discurre íntegramente por suelo urbano o urbanizable o se dedica a comunicar entre sí núcleos urbanos diferentes situados dentro del mismo municipio. En este apartado se deben incluir además los viajeros de tranvías y funiculares urbanos.

Transporte interurbano: El que se realiza entre núcleos urbanos pertenecientes a distintos términos municipales. Dentro de este tipo de transporte también se debe incluir a los viajeros que tomando un autobús interurbano se desplacen únicamente dentro de un mismo término municipal.

El transporte urbano y el interurbano se pueden clasificar en:

Transporte regular:

General: El destinado a transportar, en general, a todo tipo de pasajeros en autobuses que tienen un itinerario preestablecido sujeto a calendarios y horarios prefijados, tomando a los pasajeros en paradas fijas. Para realizar este tipo de transporte se requiere una concesión administrativa.

Especial escolar: El destinado a transportar en autobuses exclusivamente a escolares o estudiantes.

Especial laboral: El destinado a transportar en autobuses exclusivamente a colectivos laborales homogéneos (trabajadores de empresa, militares, líneas de servicio a aeropuertos o a estaciones de trenes para el transporte de tripulaciones...).

Transporte discrecional: El que realiza servicios de transporte no regular de viajeros, sin sujeción a itinerario u horario alguno (autocares contratados para visitas turísticas, excursiones, alquileres de autobuses con conductor...).

Cálculo del total de viajeros por mes

Para metro: Se contabilizarán todos los billetes expedidos tanto en taquillas como por máquinas expendedoras, más las validaciones de bonos y/o abonos de transporte en el mes de referencia.

Para el transporte urbano e interurbano:

Transporte regular:

General: Se contabilizarán todos los billetes expedidos tanto en taquillas como por máquinas expendedoras, más las validaciones de bonos y/o abonos de transporte en el mes de referencia.

Especial escolar: Si no puede calcular con precisión el número de escolares transportados en el mes, estímelos multiplicando el número de plazas de los autobuses escolares por el número de días lectivos del correspondiente mes y por el número de viajes diarios a los colegios (dos o cuatro, según sea la jornada continua o partida).

Especial laboral: Si no puede calcular con precisión el número de viajeros transportados en el mes, estímelos multiplicando el número de plazas de los autobuses de empresa por el número de días hábiles del correspondiente mes y por el número de viajes diarios de los autobuses (seis en el caso de que se cubran los tres turnos de trabajo habituales).

Transporte discrecional: Se contabilizarán los billetes vendidos en el mes de referencia. Si no puede hacerse este cálculo con precisión, estímelos multiplicando el número medio de viajeros por autobús en el mes de referencia por el número de autobuses contratados en dicho mes para realizar servicios de transporte discrecional.

[illegible]

Cálculo del total de viajeros por mes

Para metro: Se contabilizarán todos los billetes expedidos tanto en taquillas como por máquinas expendedoras, más las validaciones de bonos y/o abonos de transporte en el mes de referencia.

Para el transporte urbano e interurbano:

Transporte regular:

General: Se contabilizarán todos los billetes expedidos tanto en taquillas como por máquinas expendedoras, más las validaciones de bonos y/o abonos de transporte en el mes de referencia.

Especial escolar: Si no puede calcular con precisión el número de escolares transportados en el mes, estímelos multiplicando el número de plazas de los autobuses escolares por el número de días lectivos del correspondiente mes y por el número de viajes diarios a los colegios (dos o cuatro, según sea la jornada continua o partida).

Especial laboral: Si no puede calcular con precisión el número de viajeros transportados en el mes, estímelos multiplicando el número de plazas de los autobuses de empresa por el número de días hábiles del correspondiente mes y por el número de viajes diarios de los autobuses (seis en el caso de que se cubran los tres turnos de trabajo habituales).

Transporte discrecional: Se contabilizarán los billetes vendidos en el mes de referencia. Si no puede hacerse este cálculo con precisión, estímelos multiplicando el número medio de viajeros por autobús en el mes de referencia por el número de autobuses contratados en dicho mes para realizar servicios de transporte discrecional.

Cálculo del total de ingresos

Se contabilizarán los importes facturados por la empresa en el mes de referencia por la prestación de servicios de transporte (realizados por ella o subcontratados) sin incluir el IVA ni las subvenciones recibidas. Para el transporte en metro y en autobuses urbanos e interurbanos de línea regular general, se tendrá en cuenta el número de billetes sencillos vendidos, el número de validaciones contabilizadas de los bonos múltiples de viajes, el número de abonos de transporte mensuales y anuales vendidos, siempre contabilizando mensualmente los ingresos por la parte proporcional que corresponda a cada mes.

	Información básica sobre Protección de Datos
Responsable	Instituto Nacional de Estadística
Finalidad	Realización de esta Encuesta del Plan Estadístico Nacional
Legitimación	Cumplimiento de una misión realizada en interés público o en el ejercicio de poderes públicos
Destinatarios	No se cederán datos a terceros, salvo obligación legal
Derechos	De acuerdo con los artículos 89.2 del Reglamento 2016/679 relativo a la protección de datos de personas físicas y 25.3 de la Ley Orgánica 3/2018 de Protección de Datos Personales y Garantía de Derechos Digitales no podrán ejercerse los derechos de acceso, rectificación, oposición y limitación de tratamiento.
Información adicional	Puede consultar la información adicional y detallada sobre Protección de Datos en la página web del INE: **http://www.ine.es/proteccion_datos/oe/30163**

Naturaleza, características y finalidad

Esta encuesta se enmarca dentro del Plan Estadístico Nacional. La Estadística de Transporte de Viajeros tiene como objetivo proporcionar información sobre el número de viajeros transportados en cada uno de los medios de transporte y de otras variables relevantes para el sector.

Legislación — **Estadística de cumplimentación obligatoria**

Secreto Estadístico

Serán objeto de protección y quedarán amparados por el **secreto estadístico**, los datos personales que obtengan los servicios estadísticos tanto directamente de los informantes como a través de fuentes administrativas (art. 13.1 de la Ley de la Función Estadística Pública de 9 de mayo de 1989, LFEP). Todo el personal estadístico tendrá la obligación de preservar el secreto estadístico (art. 17.1 de la LFEP).

Obligación de facilitar los datos

Las Leyes 4/1990 y 13/1996 establecen la **obligación de facilitar los datos** que se soliciten para la elaboración de esta Estadística.

Los servicios estadísticos podrán solicitar datos de todas las personas físicas y jurídicas, nacionales y extranjeras, residentes en España (artículo 10.1 de la LFEP).

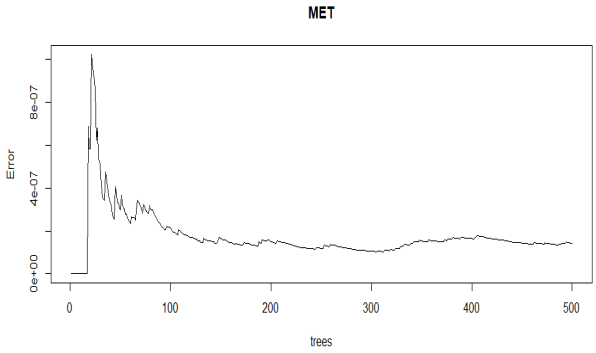
Todas las personas físicas y jurídicas que suministren datos, tanto si su colaboración es obligatoria como voluntaria, **deben contestar de forma veraz, exacta, completa y dentro del plazo** a las preguntas ordenadas en la debida forma por parte de los servicios estadísticos (art. 10.2 de la LFEP).

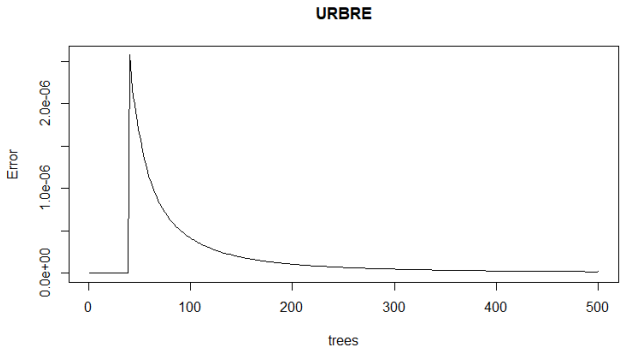
El incumplimiento de las obligaciones establecidas en esta Ley, en relación con las estadísticas para fines estatales, **será sancionado** de acuerdo con lo dispuesto en las normas contenidas en el presente Título (art. 48.1 de la LFEP).

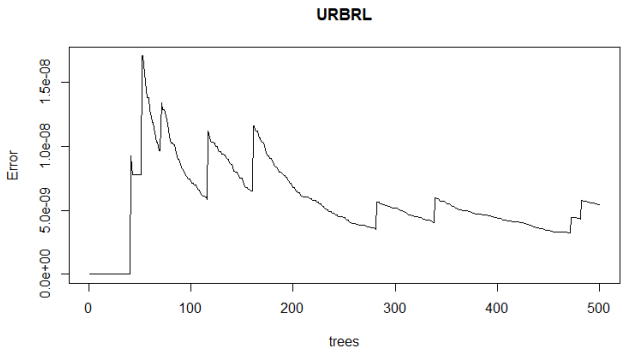
Las infracciones muy graves serán sancionadas con multas de **3.005,07 a 30.050,61 euros.** Las infracciones graves serán sancionadas con multas de **300,52 a 3.005,06 euros.** Las infracciones leves se sancionarán con multas de **60,10 a 300,51 euros** (art. 51.1, 51.2 y 51.3 de la LFEP).

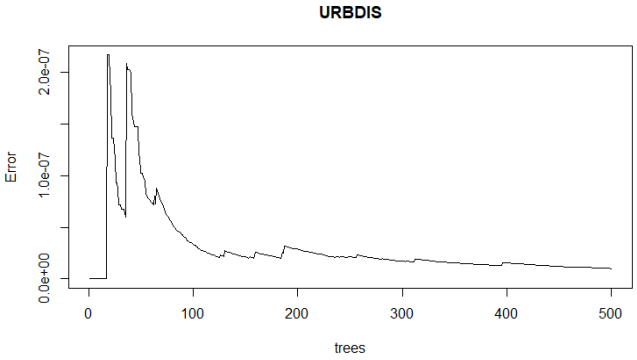
Nota: El cuestionario está disponible en las distintas lenguas cooficiales de las comunidades autónomas.

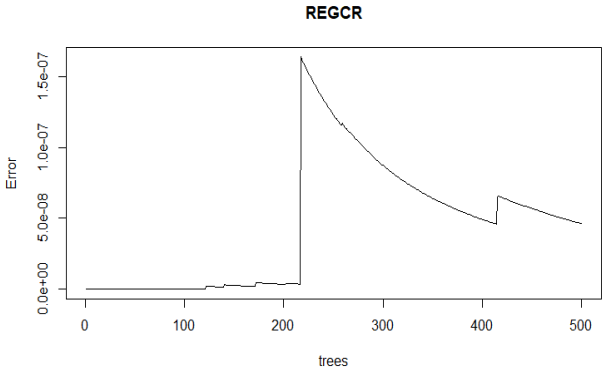
8.2 Anexo II: Gráficos de estabilización del error en *ranger* según el número de árboles de decisión empleados

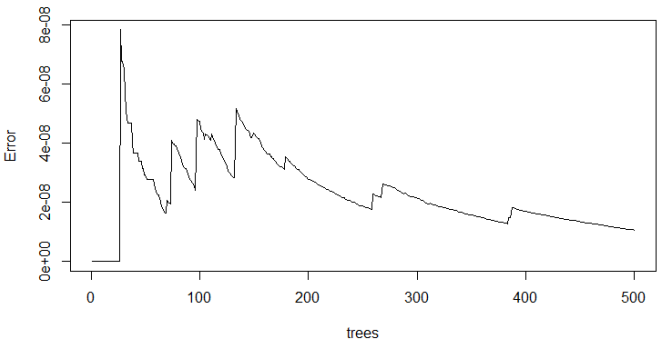


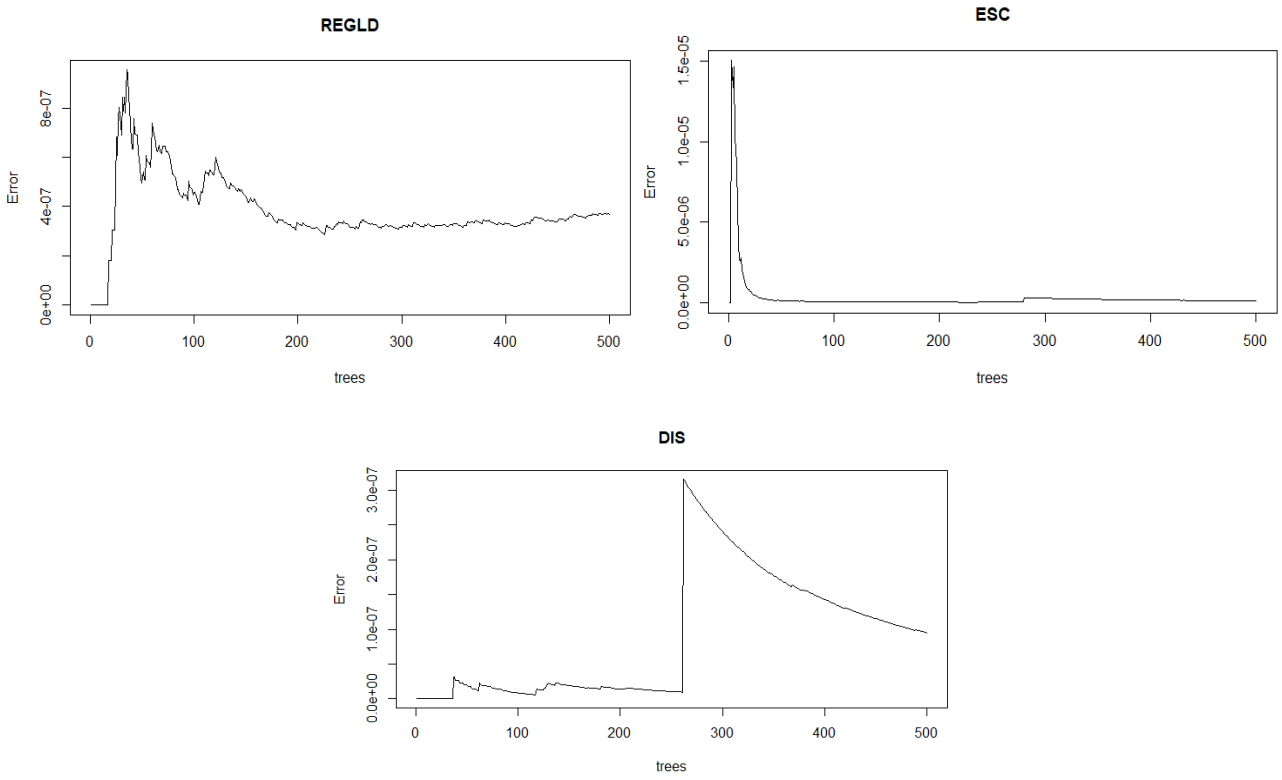




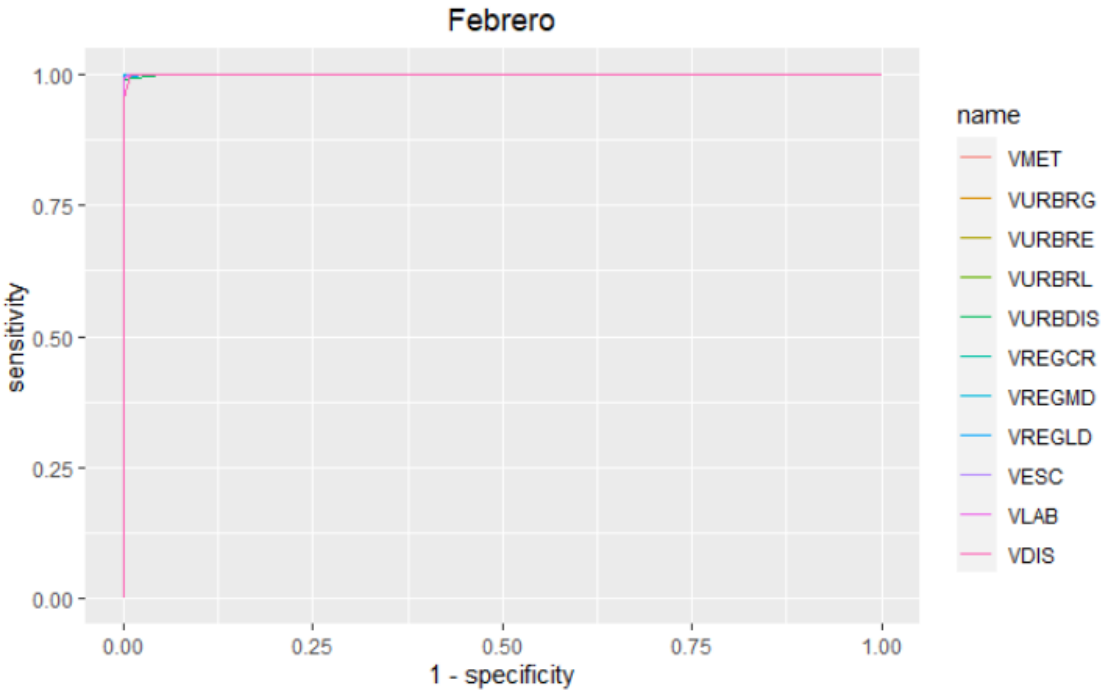


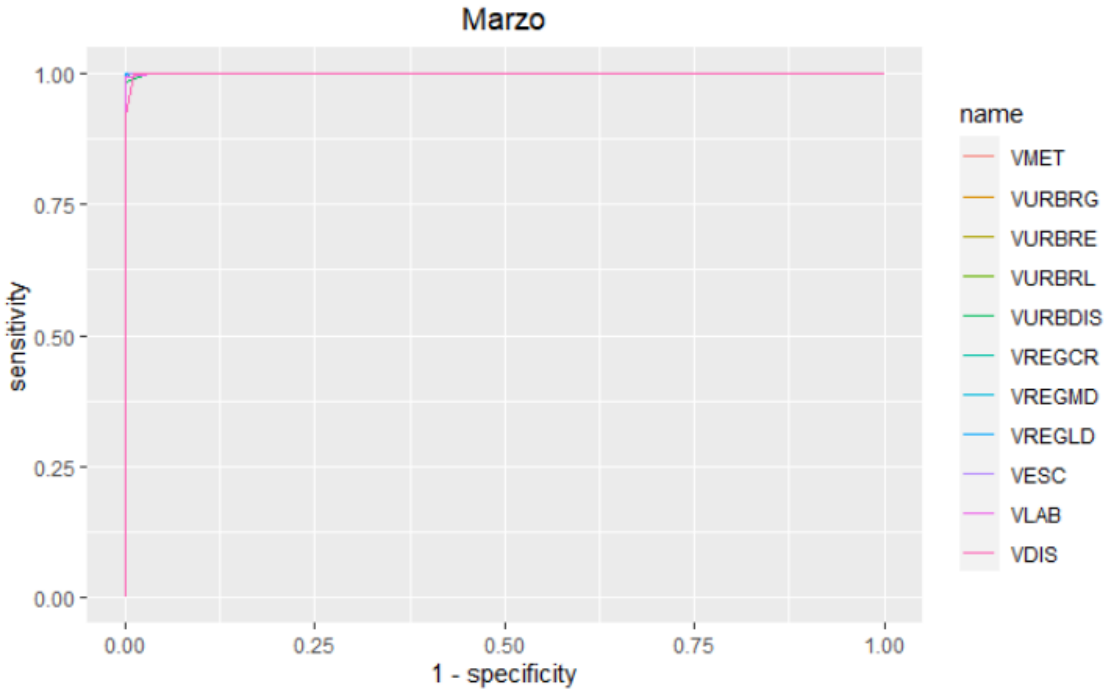


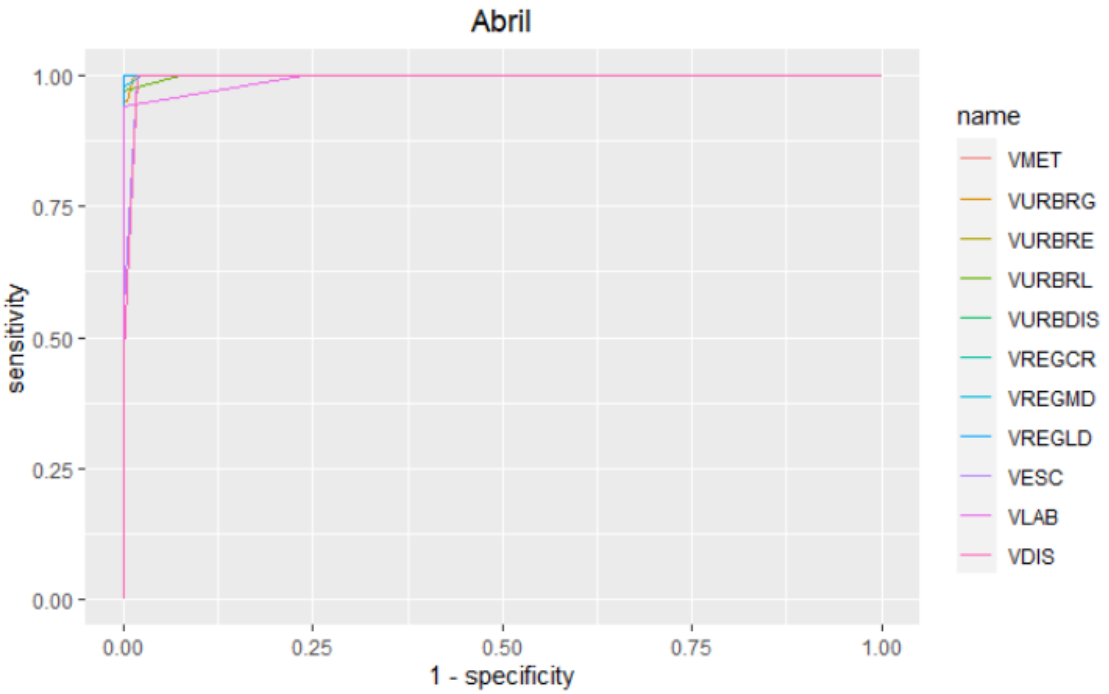


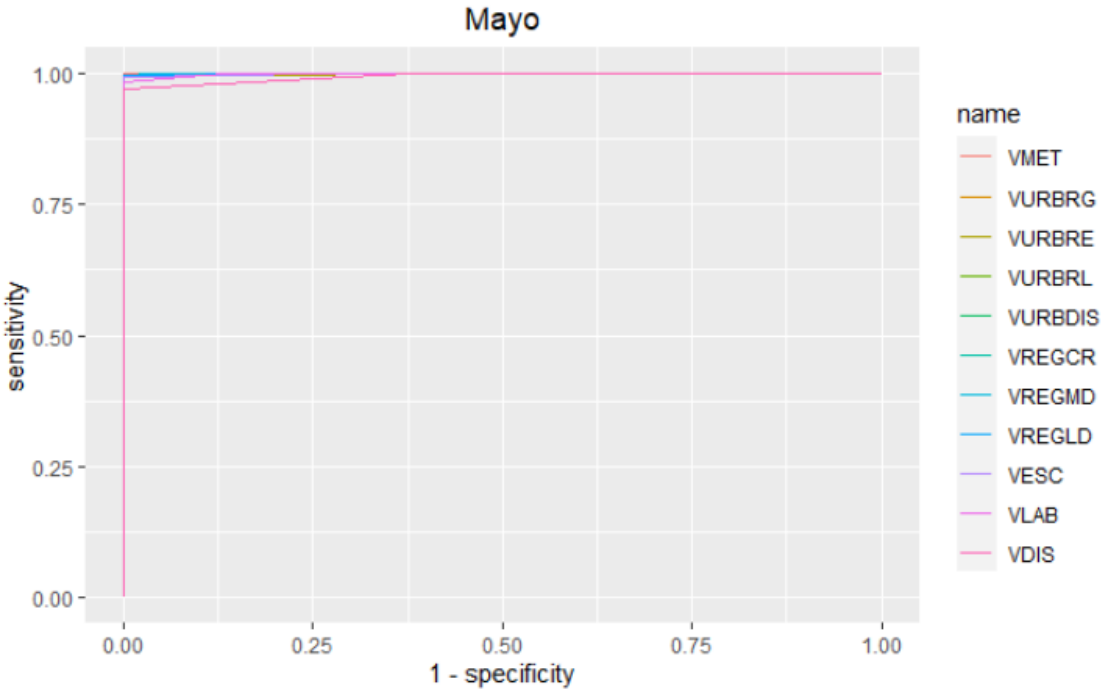


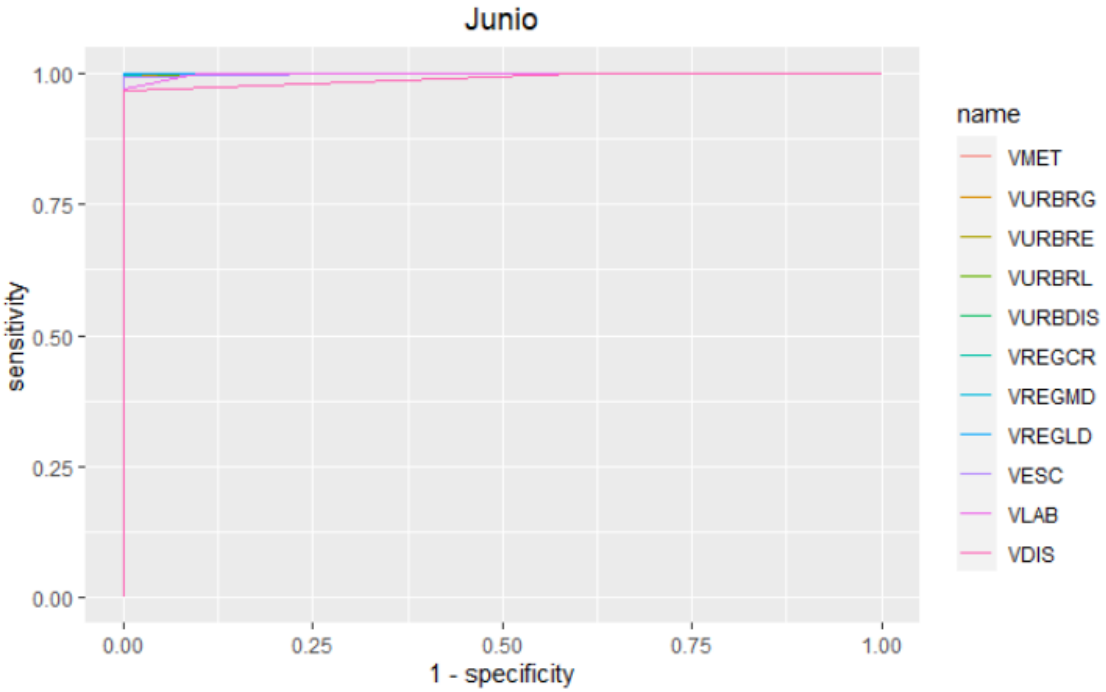
8.3 Anexo III: Curvas ROC para todos los tipos de transporte por mes

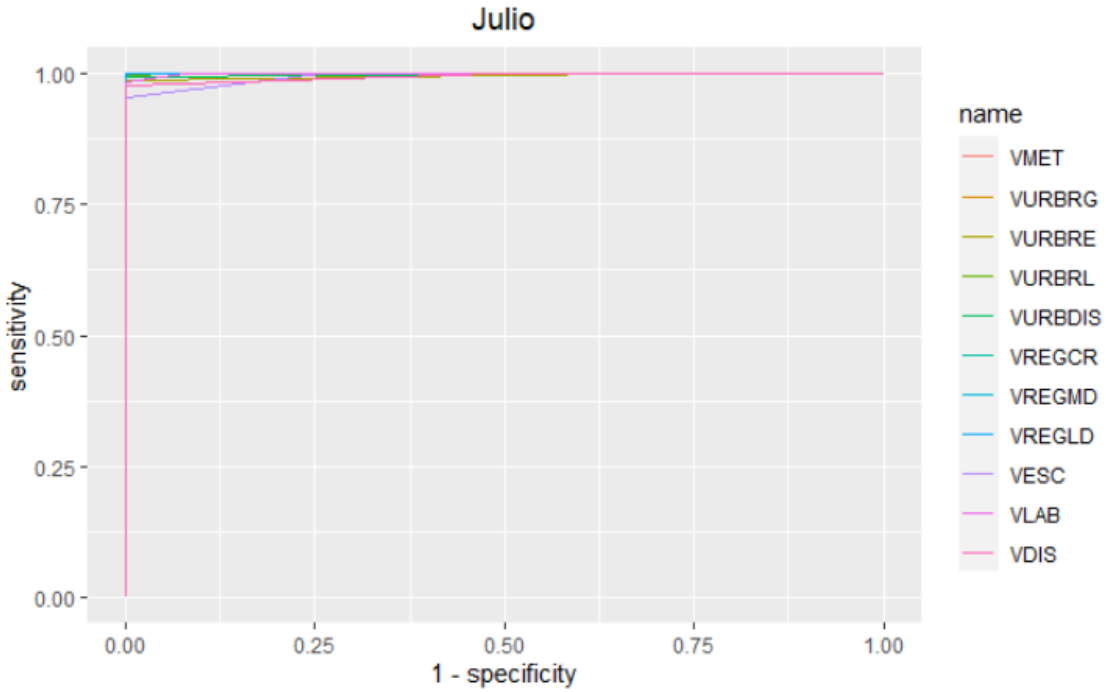


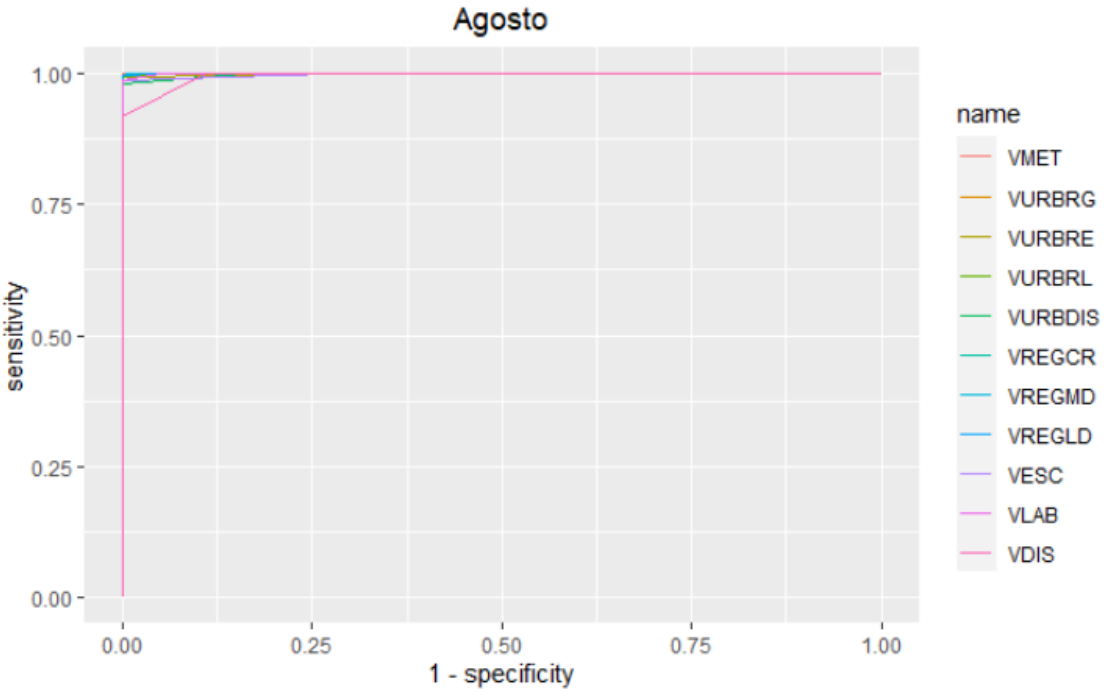

Marzo
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS

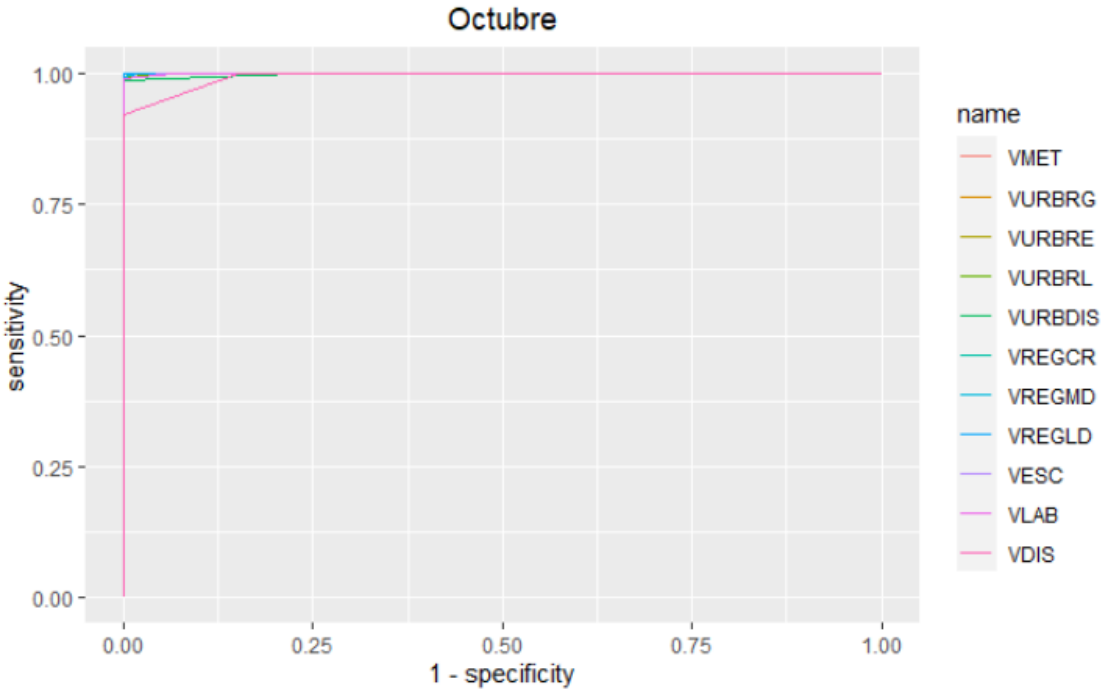

Abril
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS

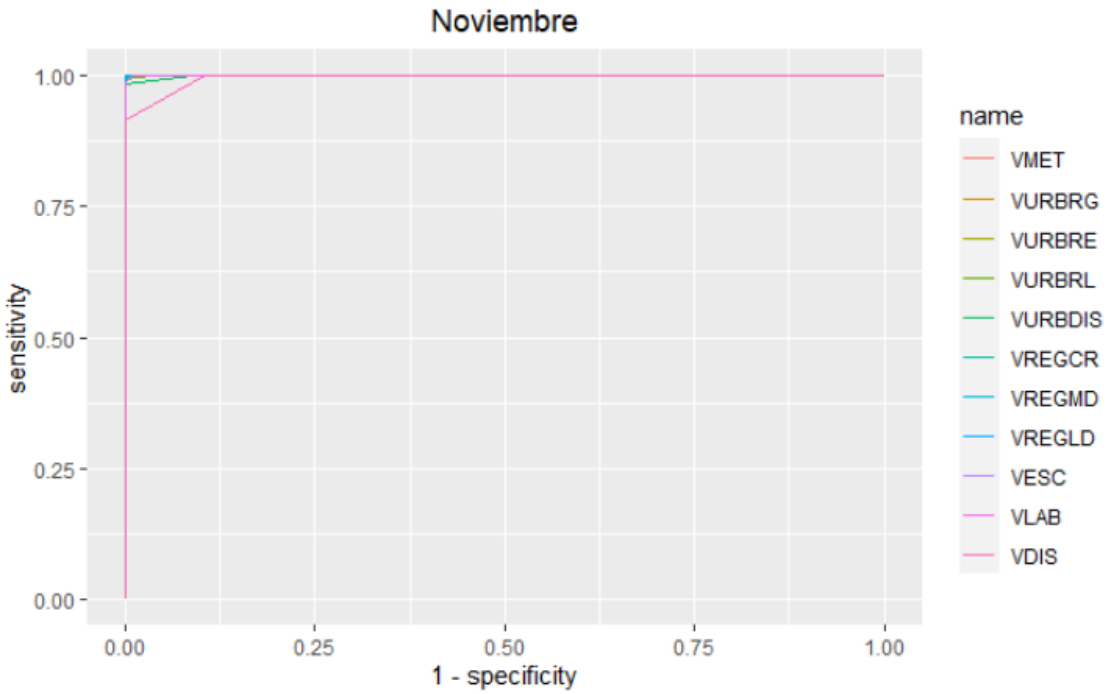

Mayo
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS

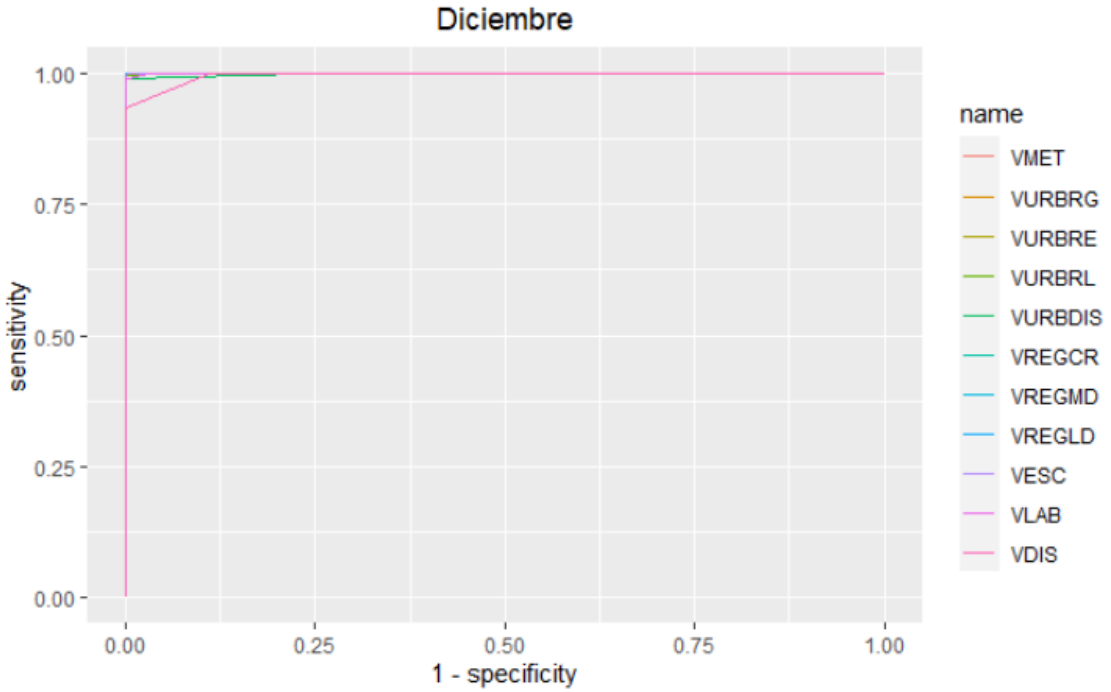

Junio
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS


Julio
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS


Agosto
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS


Octubre
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS


Noviembre
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS

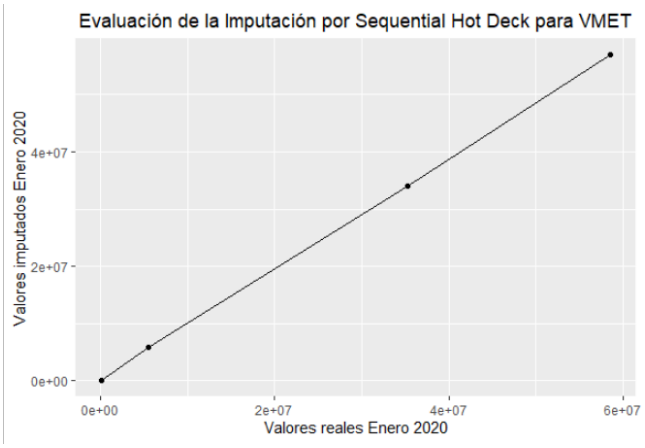

Diciembre
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity
name
VMET
VURBRG
VURBRE
VURBRL
VURBDIS
VREGCR
VREGMD
VREGLD
VESC
VLAB
VDIS

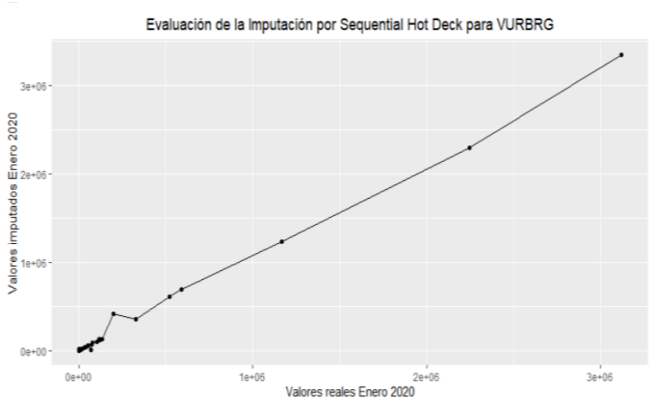
8.4 Anexo IV: Gráficos acerca de la importancia de cada regresor para cada tipo de transporte en el modelo del árbol de decisión por mes

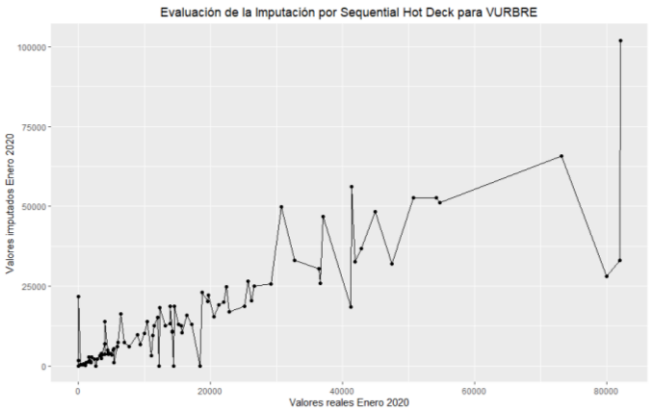

01
02
03
04
05
06
07
08
09
10
11
12
VDIS
VESC
VLAB
variable
varGestion
ObsDep
n9
n8
n7
n6
n5
n4
n3
n2
n18
n17
n16
n15
n14
n13
n12
n11
n10
n1
MES
confinamiento
ANO
0
5000
10000
15000
importance

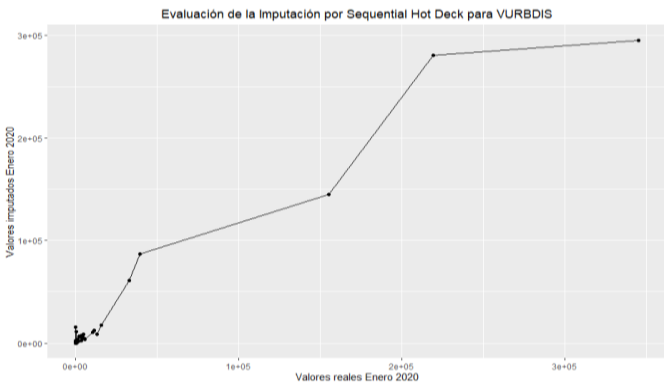
8.5 Anexo V: Evaluación de la calidad de la imputación en la primera y segunda oleada de recogida de datos en Enero de 2020

Oleada 1

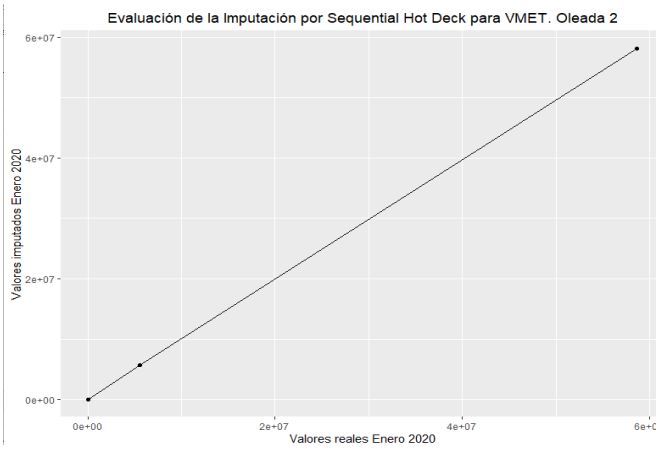


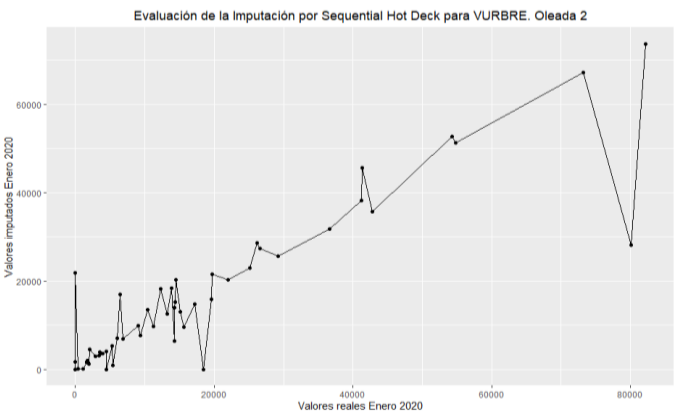


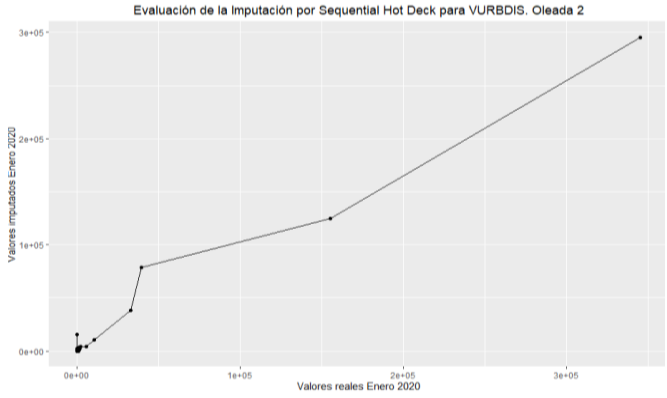


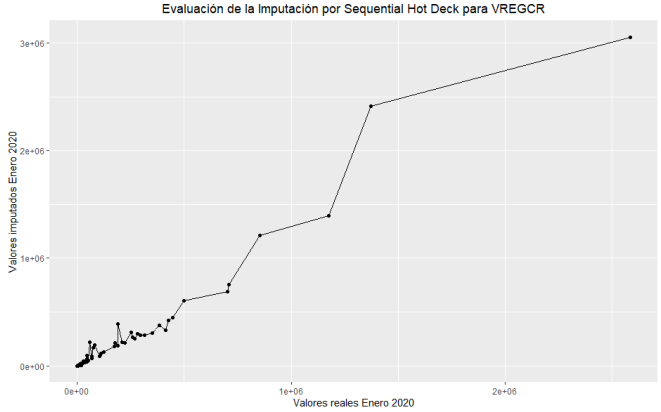


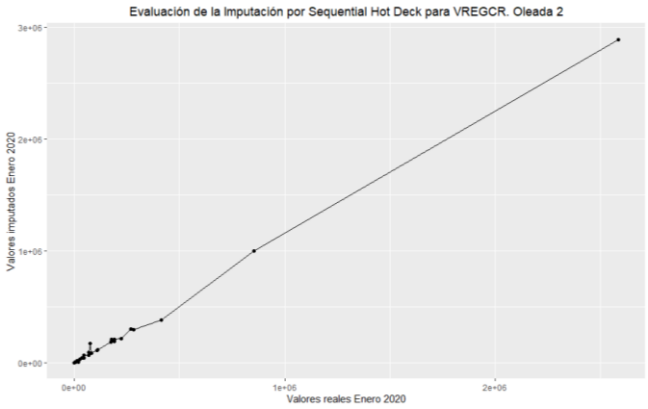
Oleada 2











Como ya se ha demostrado lo que se pretendía (las curvas mejoran notablemente de la oleada 1 a la 2) no se expondrán el resto de gráficas relativas a los 6 tipos de transporte restantes.