



# ESTUDI d'EMISSIONS de CO<sub>2</sub>

Vehicles Generalitat de Catalunya

Preparat per



**Barcelona  
Supercomputing  
Center**

*Centro Nacional de Supercomputación*



# 1. Resum executiu

En aquest document es proposa una metodologia per estimar les emissions de CO<sub>2</sub> de cotxes del tipus M1 i N1 dels vehicles registrats a la Comunitat de Catalunya. Aquesta estimació es realitza a partir dels vehicles del padró proporcionat per la DGT que ja conté un valor d'emissions provinent del fabricant o similar. S'utilitza aquesta informació per crear algoritmes que estimen l'emissió de CO<sub>2</sub> a partir de les altres propietats tècniques dels vehicles.

En aquest informe es detalla tot el treball tècnic realitzat per poder avaluar les diferents possibilitats per a la creació d'aquests algoritmes, incloent un estudi de les propietats estadístiques i qualitatives de la base de dades utilitzada, i els mètodes desenvolupats per completar aquells registres amb informació mancanta a més del CO<sub>2</sub>, i que són necessàries per a poder estimar el nivell d'emissions.

Els resultats presentats en aquest informe proporcionen un mètode vàlid per estimar les emissions de CO<sub>2</sub> dels vehicles amb categoria d'homologació M1 i N1. A més a més, es presenten les estimacions de tots els errors, tant de les estimacions dels valors que falten com de les estimacions de CO<sub>2</sub>. Els models seleccionats són aquells denominats de regressió lineal, que han estat preferits sobre altres, per brindar bons rendiments sense sacrificar interpretabilitat.

Finalment, s'inclouen discussions sobre les possibles millores que es poden introduir en aquests models. Juntament amb aquest document, es lliura també la base de dades original amb tots els camps que falten imputats mitjançant els mètodes discutits.

|                                                               |           |
|---------------------------------------------------------------|-----------|
| <b>1. Descripció del projecte</b>                             | <b>3</b>  |
| 1.1. Objectius                                                | 3         |
| 1.2. Dades                                                    | 3         |
| <b>2. Processament de dades</b>                               | <b>5</b>  |
| 2.1. Anàlisi exploratòria de les dades                        | 5         |
| 2.1.1 Quantitat de files                                      | 5         |
| 2.1.2 Quantitat de files amb informació vàlida per a entrenar | 7         |
| 2.1.3. Valors típics de les variables                         | 9         |
| 2.1.4. Correlacions entre variables                           | 10        |
| 2.2. Neteja                                                   | 18        |
| 2.2.1. Valors que falten                                      | 18        |
| 2.2.2. Outliers                                               | 21        |
| 2.2.3. Errors en l'escriptura de dades                        | 27        |
| 2.2.4. Altres errors                                          | 30        |
| 2.3. Valors mancants                                          | 34        |
| <b>3. Models d'imputació</b>                                  | <b>37</b> |
| 3.1 Missing Forests                                           | 37        |
| 3.2. Imputació per any                                        | 38        |
| 3.3. Imputació per grups d'anys                               | 39        |
| 3.4. Imputació per finestra lliscant                          | 40        |
| 3.5. Imputació per finestra acumulativa                       | 42        |
| <b>4. Models</b>                                              | <b>46</b> |
| 4.1. Mètriques de rendiment                                   | 46        |
| 4.1.1. Model directe                                          | 47        |
| 4.1.2. Model amb filtrat d'outliers                           | 47        |
| 4.2. Extracció de colinealitat                                | 48        |
| 4.3. Regressió lineal regularitzada                           | 51        |
| 4.4. Enginyeria de variables                                  | 51        |
| 4.5. Significances                                            | 53        |
| 4.6. Optimització de models lineals                           | 57        |
| 4.7. Quadrats mínims no-negatius                              | 59        |
| 4.8. Comparació amb altres models                             | 63        |
| 4.8.1. Arbres de decisió                                      | 63        |
| 4.8.2. XGBoost                                                | 64        |
| 4.9. Validació contra les dades de l'IDAE                     | 67        |
| <b>5. Conclusions</b>                                         | <b>70</b> |
| 5.1. Model final                                              | 70        |
| 5.2. Recomanacions futures                                    | 72        |

# 1. Descripció del projecte

Les emissions del parc vehicular de la Comunitat de Catalunya són objecte d'interès de la Secretaria de Medi Ambient i Sostenibilitat de la Generalitat i també de la Secretaria d'Hisenda.

Existeixen en circulació molts cotxes per als quals no es té una estimació fiable del seu nivell d'emissions, en la seva major part perquè la normativa vigent al moment de la seva posada en circulació no exigia aquesta informació. Més en concret, la majoria dels cotxes fabricats abans de 2002, i molts dels fabricats entre 2002 i 2009, no registren informació oficial sobre el seu nivell d'emissions de CO<sub>2</sub>.

Per altra banda, el padró vehicular conté una gran quantitat de cotxes de l'última dècada per als quals *sí hi ha* registre i informació oficial. Es planteja llavors la possibilitat d'utilitzar aquestes dades i diverses tècniques estadístiques per inferir el nivell d'emissions que tenien aquells cotxes i per als quals no es disposa d'informació. S'entén, que aquesta estimació tindrà alguns biaixos inherents a causa de les dades facilitades per la DGT, amb la qual cosa és important estudiar també el conjunt de dades disponibles de manera crítica, per identificar aquests biaixos i determinar si causen una sobreestimació o una subestimació de les emissions.

## 1.1. Objectius

L'objectiu principal d'aquest estudi és assignar a tots els vehicles del tipus M1 i N1 del padró de vehicles de la Comunitat de Catalunya (proveïts per la Direcció General de Tráfico) una estimació del seu nivell d'emissions de CO<sub>2</sub> "a km. 0", és a dir, les emissions que tindria cada vehicle al moment de fabricació donades les seves característiques tècniques.

## 1.2. Dades

El padró de vehicles proveït per la Direcció General de Tráfico està completament anonimitzat i conté un total de 5.271.750 vehicles, dels quals 3.467.365 són de categoria d'homologació M1 i 218.905 són de categoria d'homologació N1.

Les columnes descriptives lliurades per a cada vehicle, separades per un caràcter '|', són:

| Nom camp                | Descripció                                                                                                                                                                                                                                                                                                 | Tipus i Longitud |
|-------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| FEC_PRIM_MAT            | Data Primera Matriculació.                                                                                                                                                                                                                                                                                 | 'YYYYMMDD'       |
| MARCA                   | Marca del vehicle                                                                                                                                                                                                                                                                                          | VARCHAR (60)     |
| MODEL                   | Model del vehicle                                                                                                                                                                                                                                                                                          | VARCHAR (50)     |
| DES_SERVICIO            | Descripció de serveis públics-particulars                                                                                                                                                                                                                                                                  | VARCHAR (30)     |
| DES_TIPO                | Descripció del tipus de vehicle                                                                                                                                                                                                                                                                            | VARCHAR (50)     |
| CLASIFICACION_INDUSTRIA | Identifica el tipus de vehicle segons les normes de indústria                                                                                                                                                                                                                                              | VARCHAR (4)      |
| NUM_HOMOLOGACION        | Conté la contrasenya d'homologació que apareix consignada en el requadre de certificació de la targeta d'ITV. Pot estar en blanc                                                                                                                                                                           | VARCHAR (35)     |
| CILINDRADA              | Cilindrada del vehicle                                                                                                                                                                                                                                                                                     | DECIMAL (5,0)    |
| POTENCIA_FISCAL         | Potència del vehicle                                                                                                                                                                                                                                                                                       | DECIMAL (5,2)    |
| POTENCIA_NETA           | (Potència neta màxima en kW). Obligatori, excepte en el cas de remolcs i semiremolcs. Per recollir potència real en KW de l'epígraf corresponent de la targeta d'ITV                                                                                                                                       | DECIMAL (5,2)    |
| MASA_MAX                | Pes màxim del vehicle                                                                                                                                                                                                                                                                                      | DECIMAL (7,0)    |
| MOM                     | Massa en ordre de marxa. Si la massa en servei és 0 o nul·la agafem la massa del vehicle en circulació (Aquesta dada no apareix consignat a la targeta d'ITV, sinó que es dedueix de la tara. En cas de motocicletes, ciclomotors, vehicles de tres rodes, quadricicles lleugers i no lleugers, i remolcs) | INTEGER          |
| TARA                    | Tara del vehicle (pes del vehicle)                                                                                                                                                                                                                                                                         | DECIMAL (7,0)    |
| NUM_PLAZAS              | Nombre de places d'un vehicle. Per a un vehicle de càrrega, aquest camp indicarà el nombre de places màxim permès quan el vehicle està habilitat per a càrrega de mercaderies (ex: seients posteriors abatuts). -                                                                                          | DECIMAL (3,0)    |
| COMBUSTIBLE             | Descripció del tipus de propulsió                                                                                                                                                                                                                                                                          | VARCHAR (50)     |
| TIPO_ALIMENTACION       | Valors: M = monocombustible; B = biocombustible; F = flexicombustible                                                                                                                                                                                                                                      | VARCHAR (1)      |
| CONSUM                  | Consum en wh / km                                                                                                                                                                                                                                                                                          | INTEGER          |
| EMISIONES_UE            | Nivell d'emissions del motor que apareix en l'homologació de tipus                                                                                                                                                                                                                                         | VARCHAR (8)      |
| CO2                     | Emissions CO2 en g/km                                                                                                                                                                                                                                                                                      | DECIMAL (6,3)    |
| CATEGORIA_HOMOLOGACION  | Conté la contrasenya d'homologació que apareix consignada en el requadre de certificació de la targeta d'ITV. Pot estar en blanc.                                                                                                                                                                          | VARCHAR (35)     |

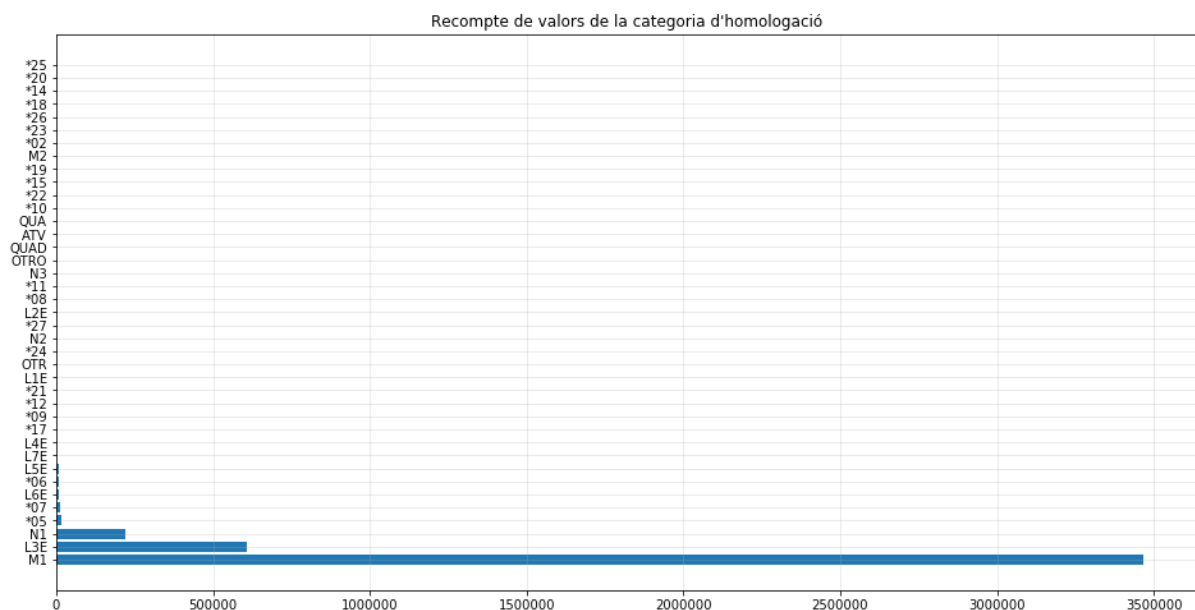
|                     |                                                                                                             |              |
|---------------------|-------------------------------------------------------------------------------------------------------------|--------------|
| COD_PROVINCIA       | Codi INE de la província del vehicle                                                                        | INTEGER (2)  |
| MUNICIPI            | Nom del municipi on està domiciliat el vehicle                                                              | VARCHAR (30) |
| CATEGORIA_ELECTRICO | Categoria de vehicle elèctric                                                                               | VARCHAR (4)  |
| AUTONOMIA_ELECTRICO | Autonomia del vehicle elèctric. Primeres 4 posicions del camp autoselect de les dades tècniques del vehicle | VARCHAR (4)  |
| CO2 IDAE            | Variable extreta de la pàgina web <a href="http://www.idae.es">www.idae.es</a> .                            |              |

## 2. Processament de dades

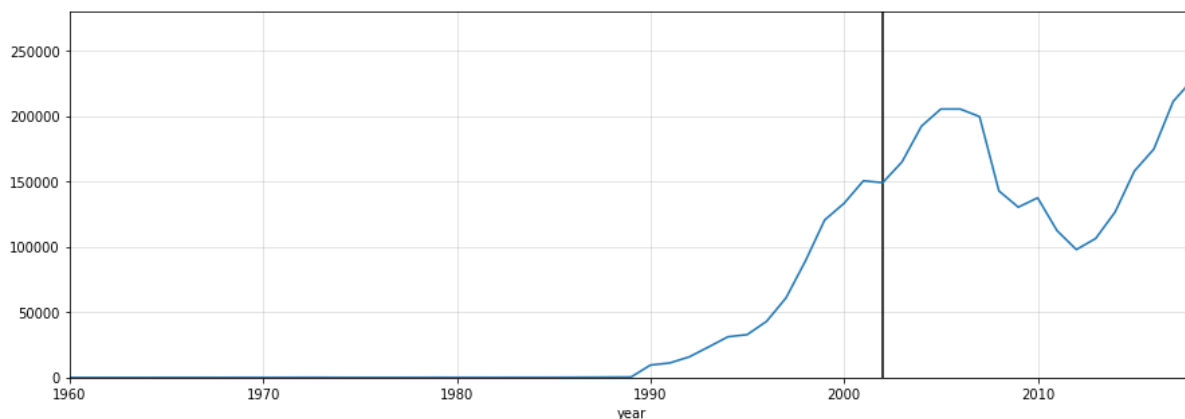
### 2.1. Anàlisi exploratòria de les dades

#### 2.1.1 Quantitat de files

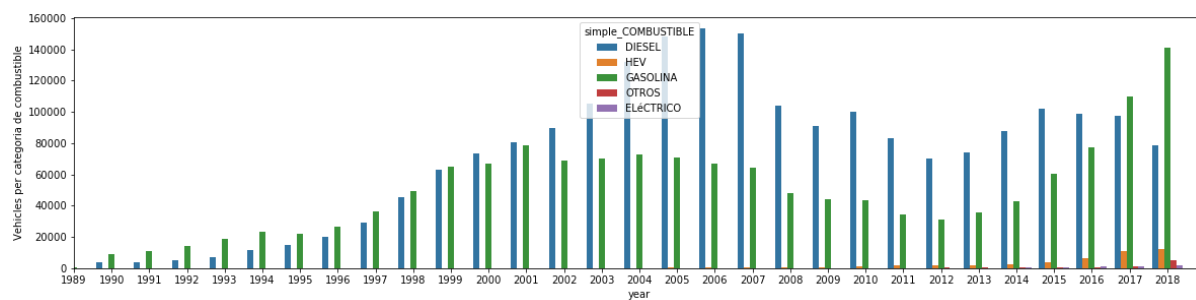
En els següents gràfics observem la quantitat de vehicles per categoria d'homologació, i la quantitat de vehicles M1 i N1 per any de registre:



|           |                    |
|-----------|--------------------|
| <b>M1</b> | 3.467.365 (79.59%) |
| <b>N1</b> | 218.905 (5.02%)    |

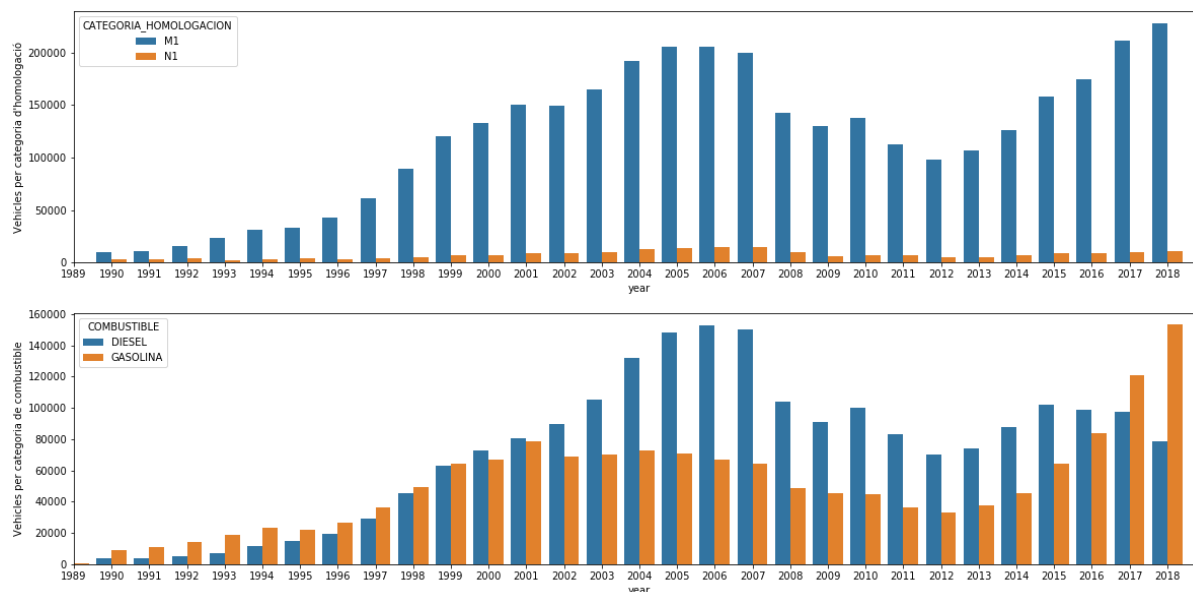


En les dades obtingudes no s'indica si els vehicles són considerats com *històrics*, cosa que es podria esbrinar simplement tenint accés a la matrícula del cotxe. D'altra banda, els cotxes d'anys previs a 1989 suposen un percentatge del ~ 0124% del total dels cotxes, de manera que no resulta en un biaix important si els descartem d'ara en endavant.



D'aquí en endavant també només ens referirem als vehicles de tipus M1 i N1, que són el focus d'interès expressat per la Secretaria de Medi Ambient i Sostenibilitat i la Secretaria d'Hisenda.

Veiem que la quantitat de vehicles M1 és molt més gran que la de vehicles N1. En el gràfic de sota podem veure l'impacte en la matriculació de vehicles de la crisi de 2009.



Separant els cotxes pels dos combustibles més populars, veiem també com en els anys més recents s'ha abandonat el cotxe dièsel pel de gasolina.

S'observen en els camps de marca i model (on es pot introduir text lliure), que hi han 526 marques úniques i 95.756 models únics. Els camps que identifiquen els models dels cotxes presenten moltes incongruències i errors creats a l'hora de la inserció d'aquests camps, com per exemple "NISSANQASHQAI" o "NISSANQASH". Discutirem aquests problemes més endavant.

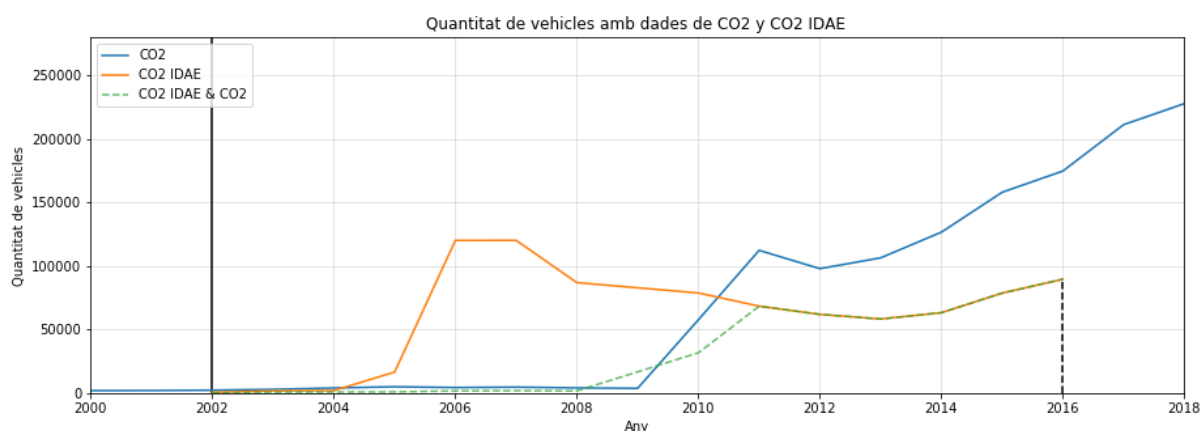
## 2.1.2 Quantitat de files amb informació vàlida per a entrenar

Dels 3.686.270 vehicles a la base de dades, 1.381.681 tenen informació de CO<sub>2</sub>, i 877.368 contenen informació en el camp CO<sub>2</sub> de IDAE. No obstant això, de vegades la informació apareix present, però és un zero. Si descartem els zeros considerant-los com a valors no vàlids, trobem que tenen informació només de CO<sub>2</sub> 1.352.655 vehicles i tenen informació de CO<sub>2</sub> IDAE 877.361 vehicles, dels quals no tenen informació de CO<sub>2</sub> 404.342 vehicles, mentre que tenen ambdós valors 473.019 vehicles.

|           | Té CO2 IDAE    | No té CO2 IDAE   | Total            |
|-----------|----------------|------------------|------------------|
| Té CO2    | 473.019        | 879.636          | <b>1.352.655</b> |
| No té CO2 | 404.342        | 1.929.273        | <b>2.333.615</b> |
| Total     | <b>877.361</b> | <b>2.808.909</b> | <b>3.686.270</b> |

En principi, hem d'intentar estimar les emissions de 1.929.273 vehicles. No obstant això, no només no estan totes les dades que necessitem per a fer l'estimació (com veurem més endavant), sinó que la distribució de vehicles amb l'objectiu que falta no és aleatòria.

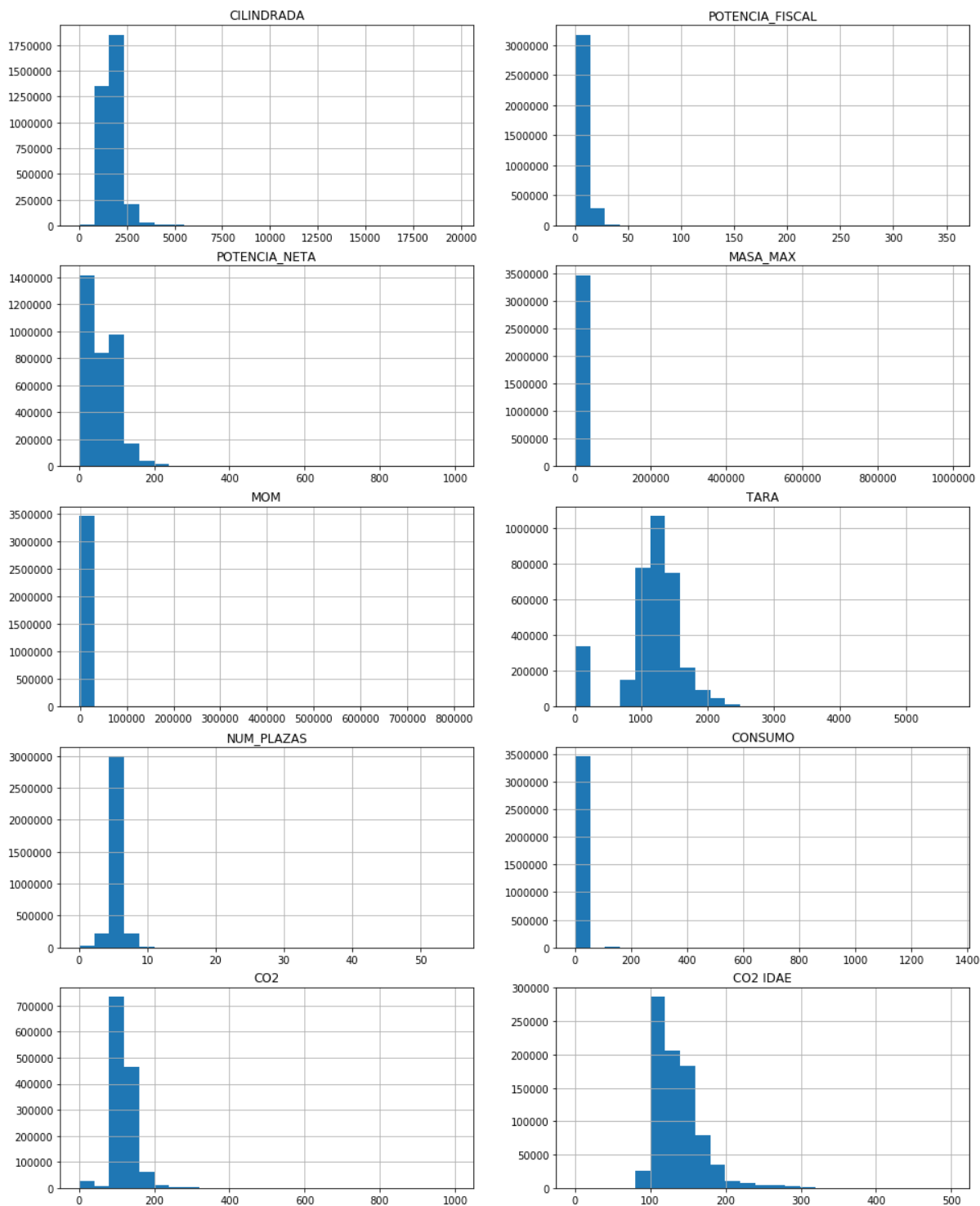
Per exemple, la quantitat de vehicles a predir es troba aclaparadorament en anys anteriors, com veiem per la quantitat de vehicles amb informació de CO2 i CO2 IDAE per any:



Observant l'edat del parc amb la quantitat de vehicles per any de matriculació (indicant 2008 com l'any a partir del qual és obligatori incloure informació d'emissions de CO2) veiem que el problema és tractable perquè, en proporció, hi ha més cotxes moderns que antics. Per altra banda, aquesta distribució de les dades que faltaven introdueix un biaix a la baixa en l'estimació d'emissió de cotxes més antics: els cotxes més recents amb què entrenarem els models estadístics tenen substancialment menys emissions que els cotxes més antics.

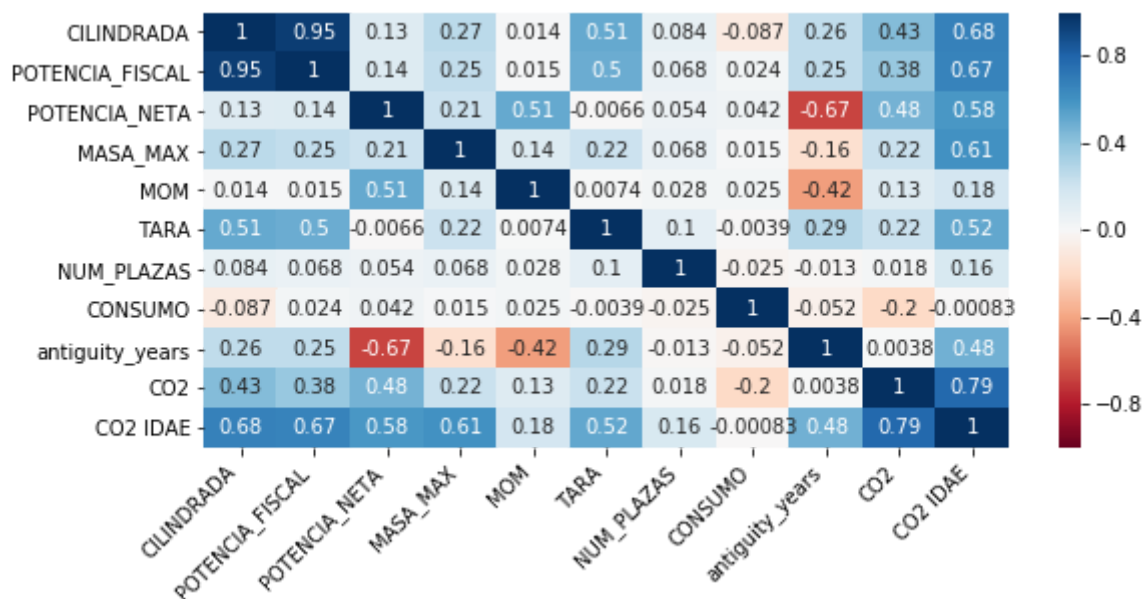
### 2.1.3. Valors típics de les variables

Aquí mostrem histogrames de les variables numèriques del conjunt de dades:



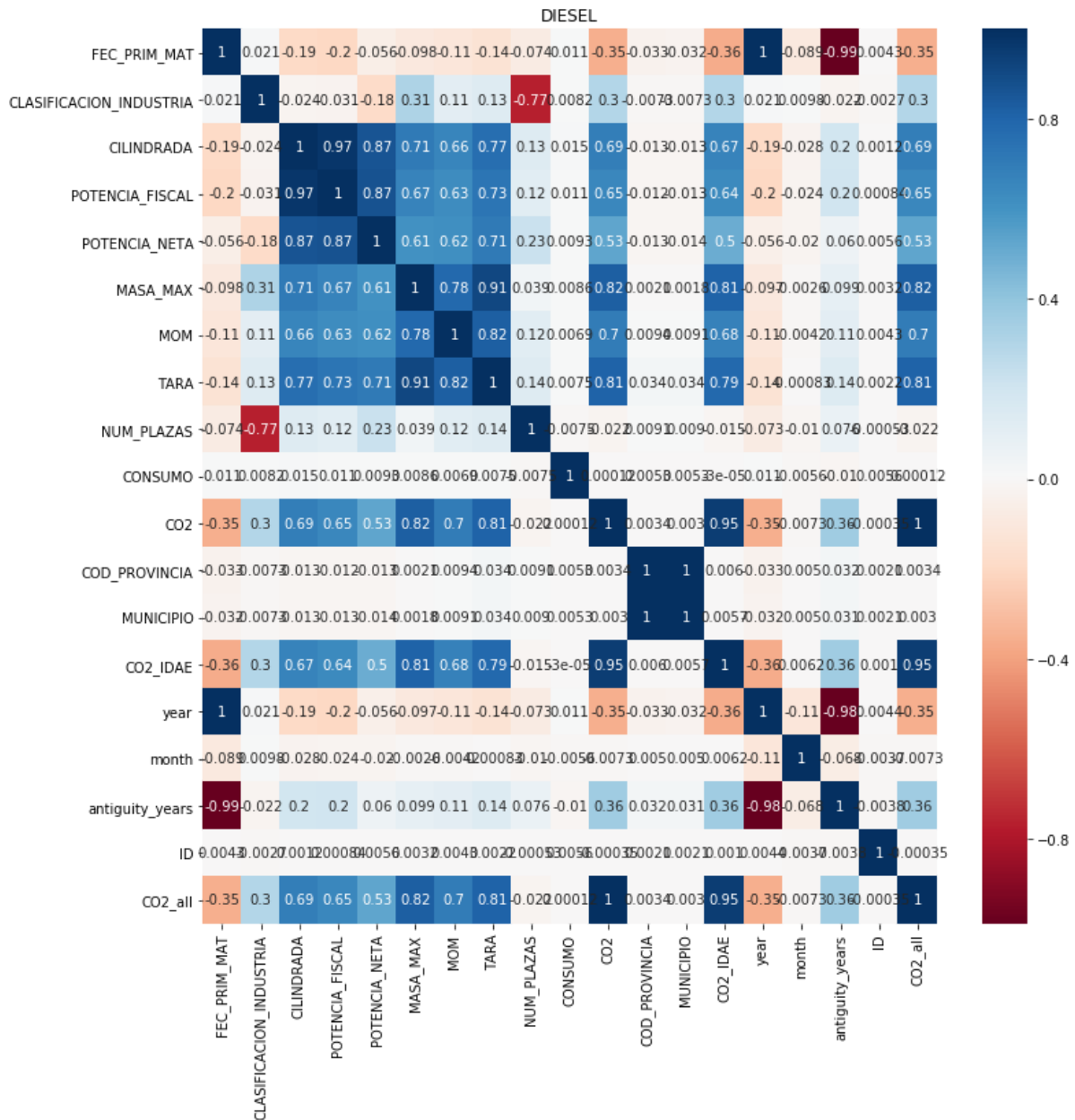
## 2.1.4. Correlacions entre variables

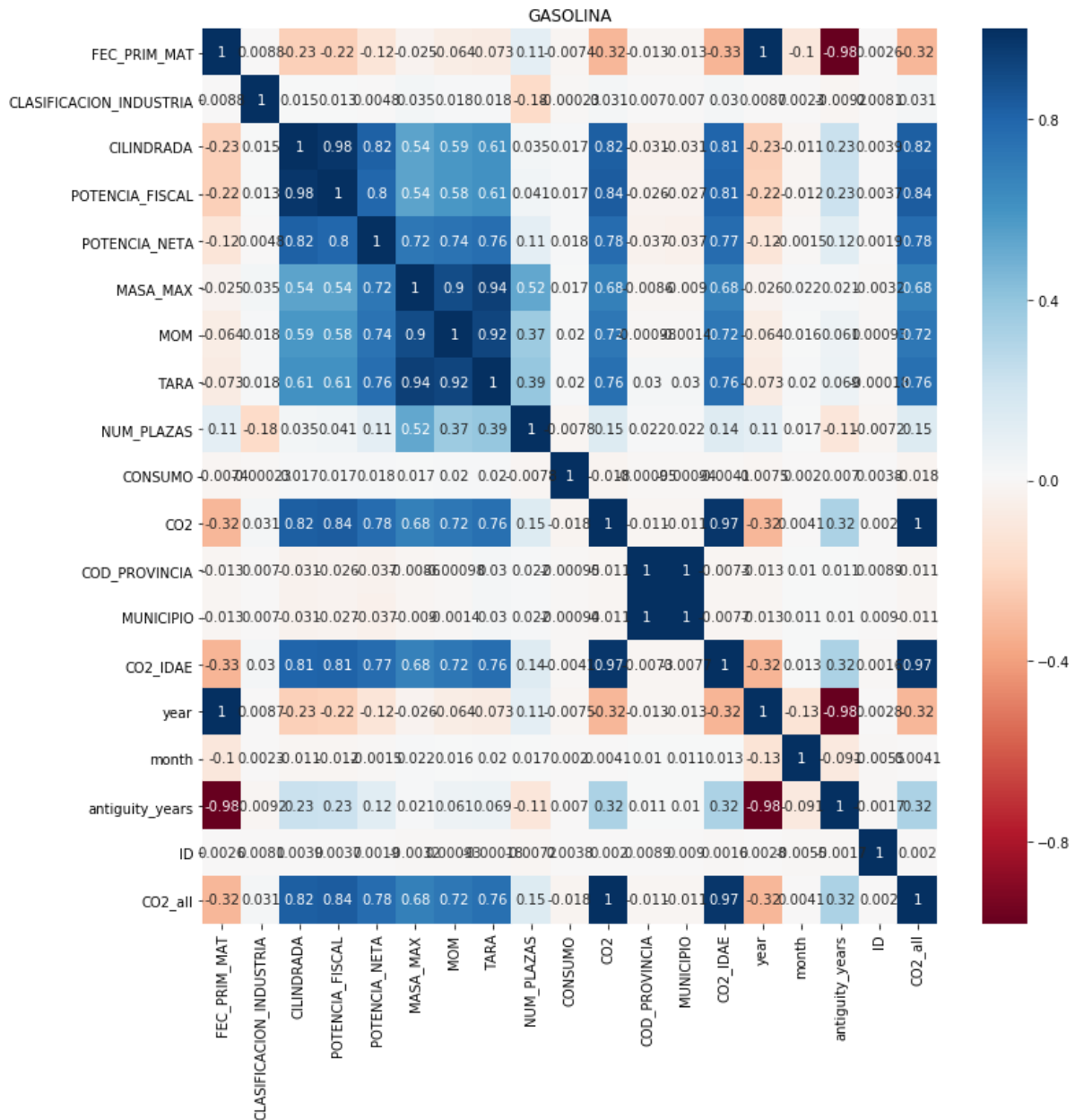
Les correlacions entre les variables numèriques mostren una saludable presència de diversos factors correlacionats que poden ajudar a la predicció:

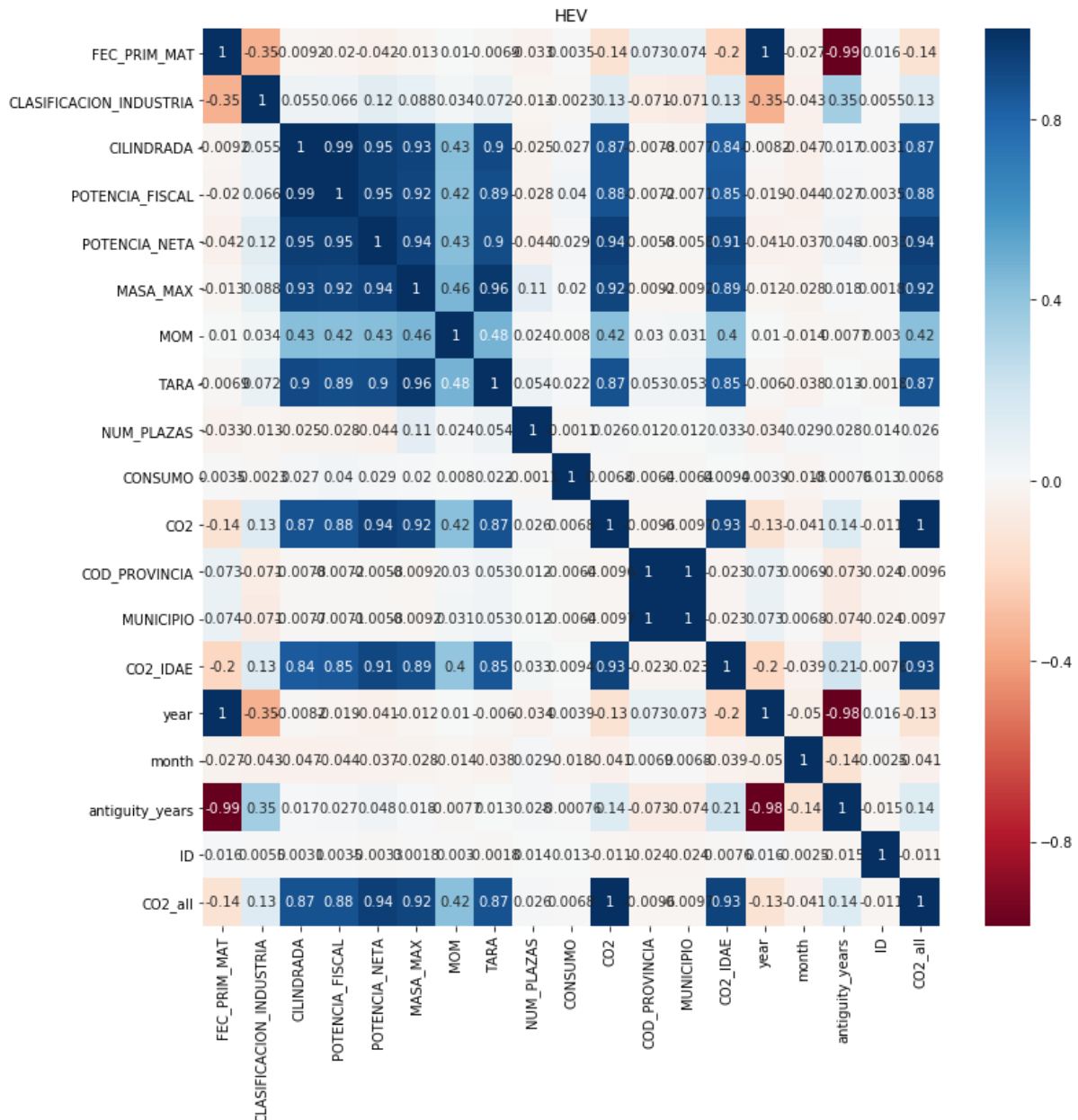


Cal notar que no totes les variables correlacionen positivament amb l'emissió de CO2, en particular el consum.

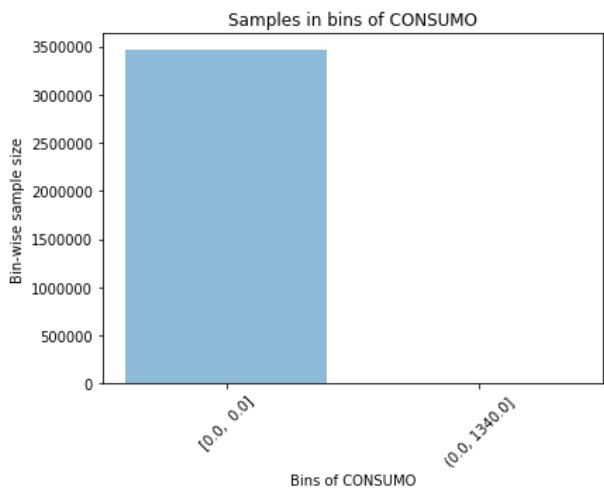
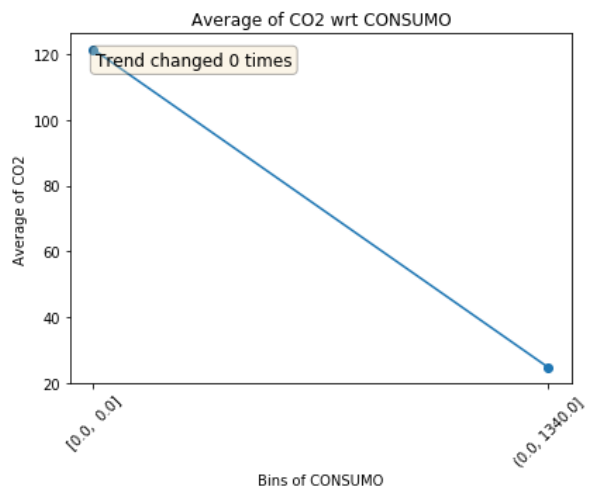
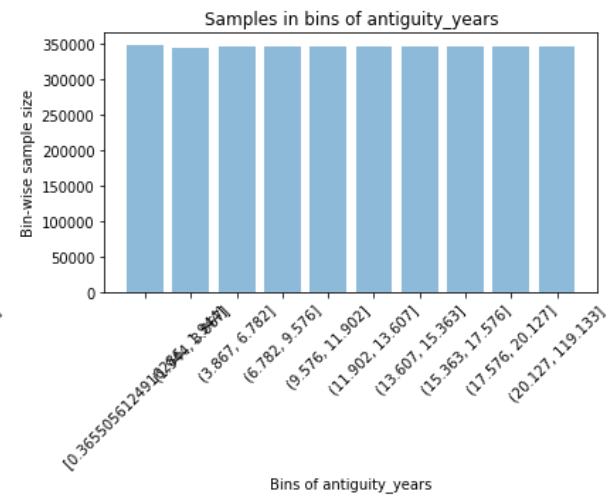
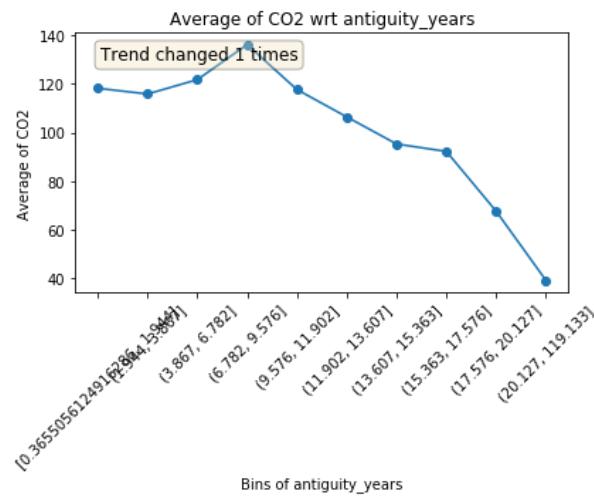
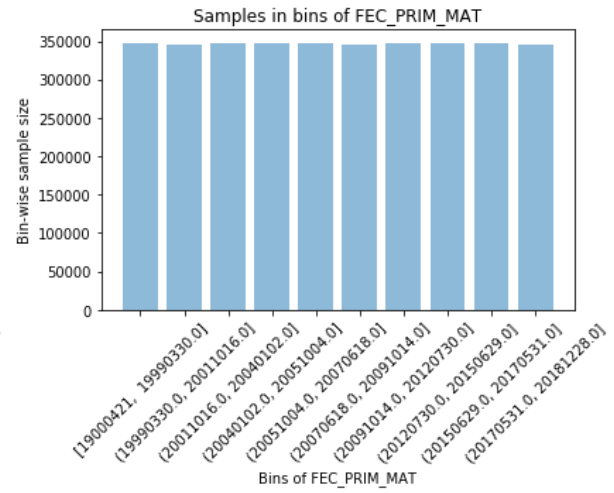
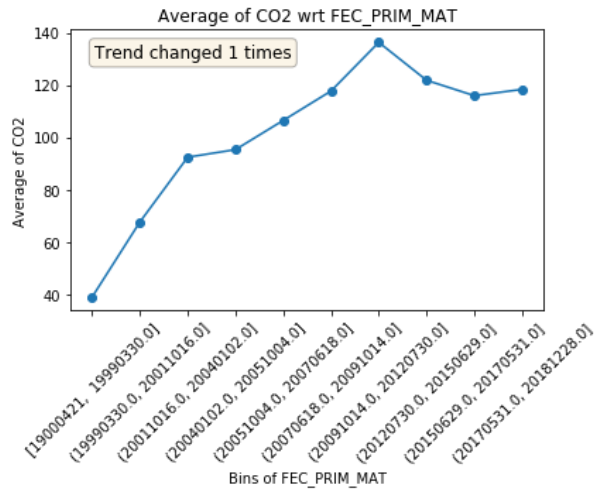
A continuació mostrem el mateix gràfic de dalt segregat pels tres tipus més comuns de combustible: vehicles de gasolina, de dièsel i els híbrids. S'observen característiques que són d'esperar, com ara el CO2 dels vehicles híbrids està menys negativament correlacionat amb l'antiguitat dels vehicles, ja que aquest tipus de combustible resulta en unes emissions molt més baixes de CO2.

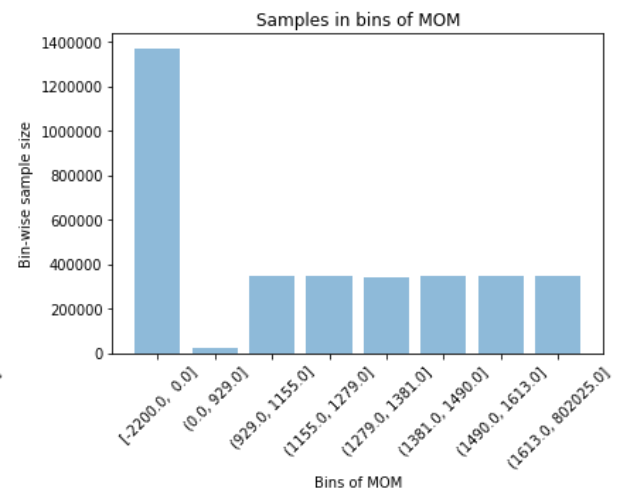
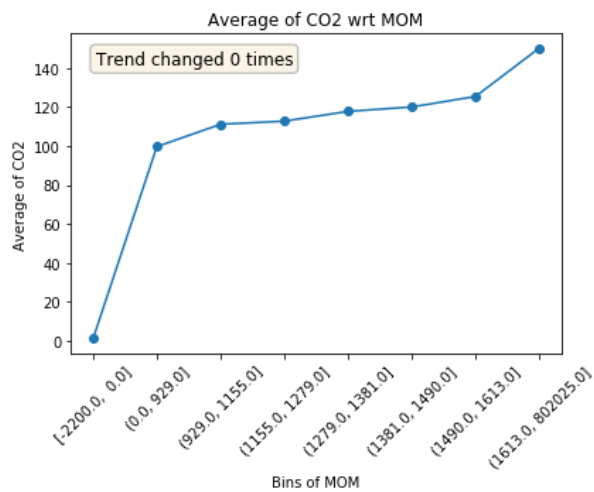
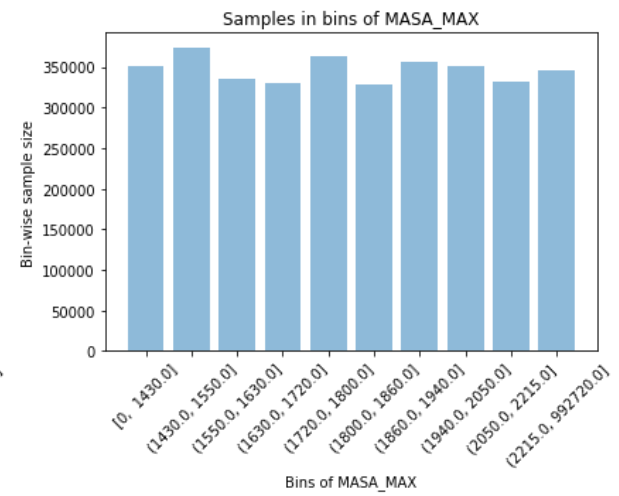
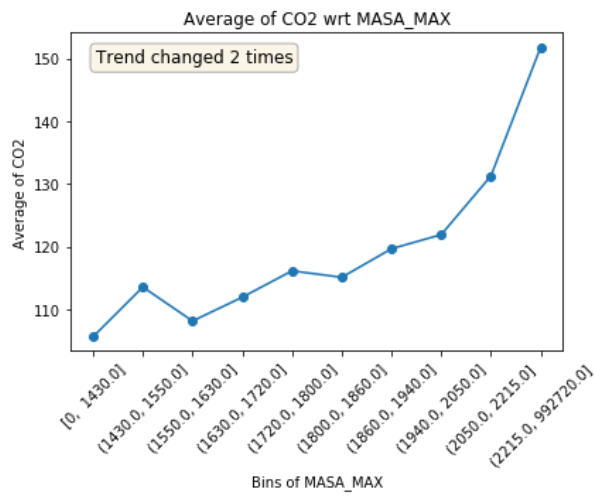
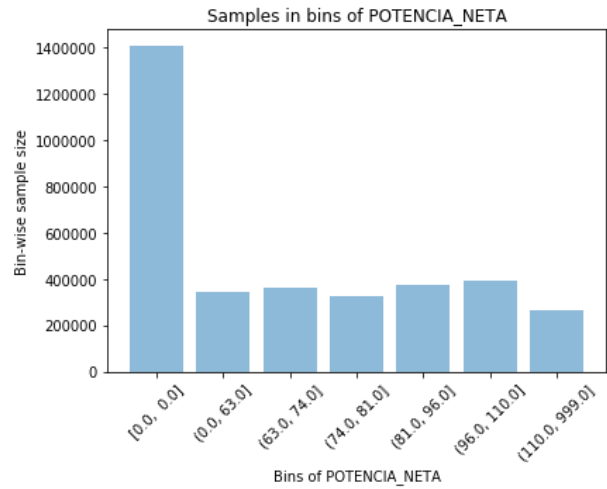
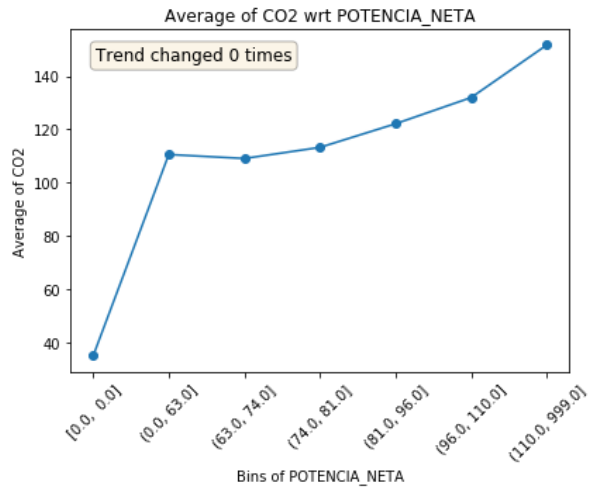


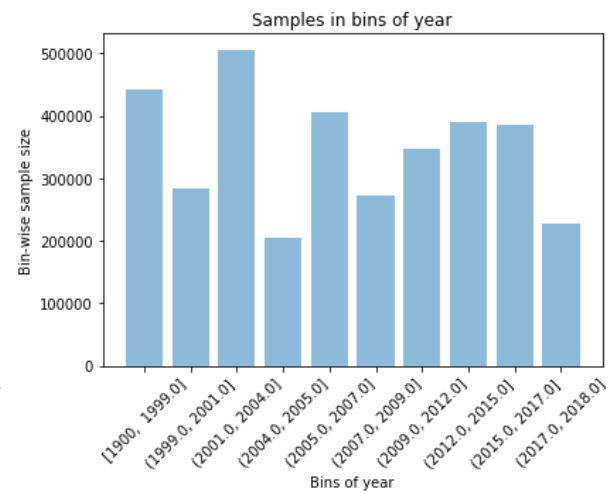
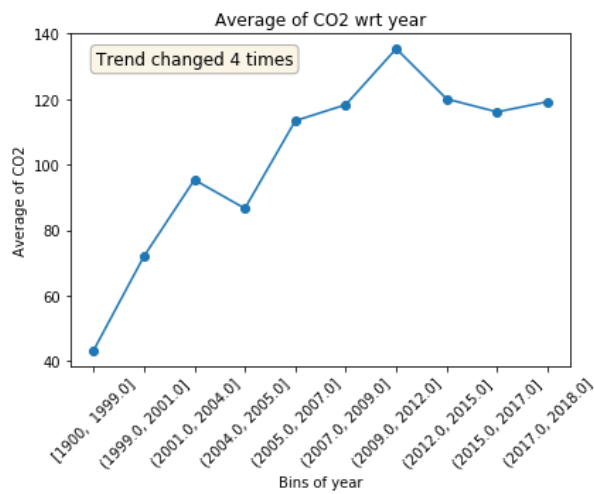
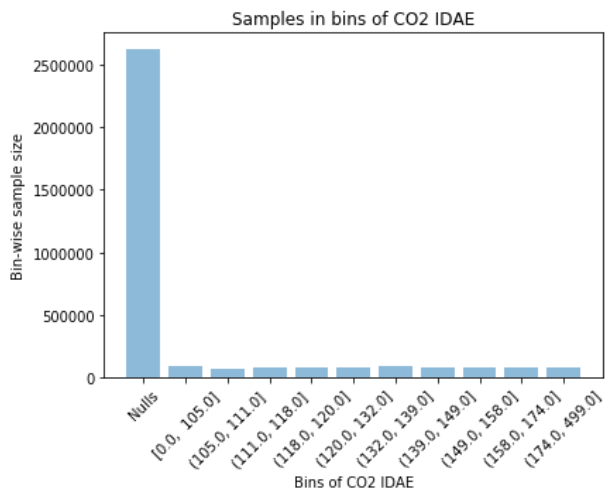
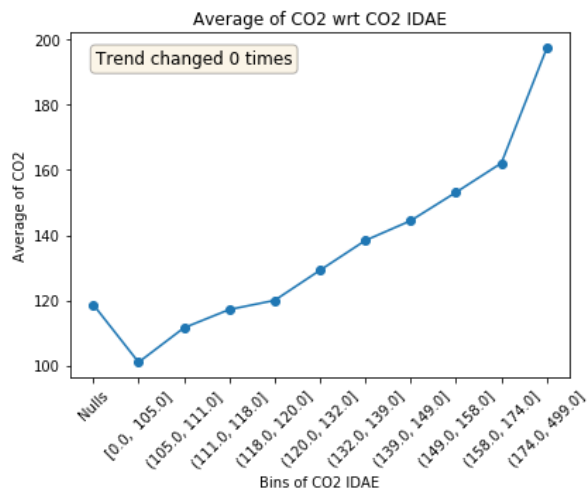
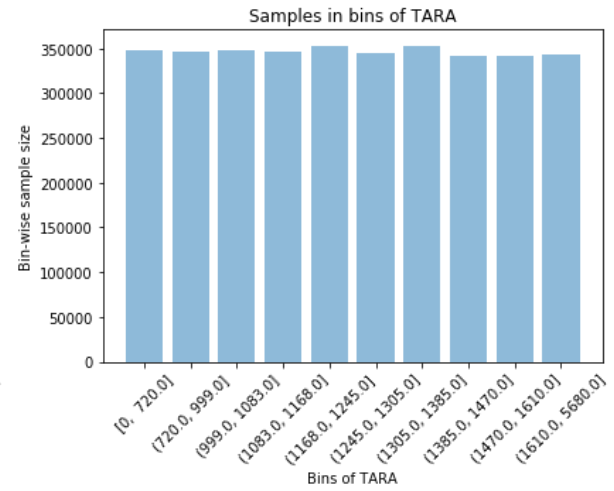
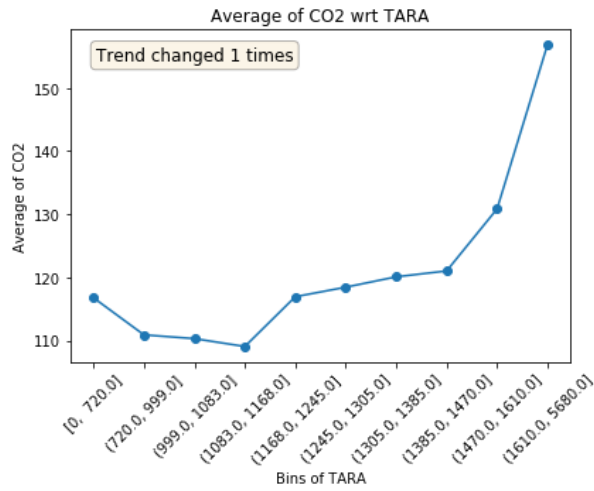


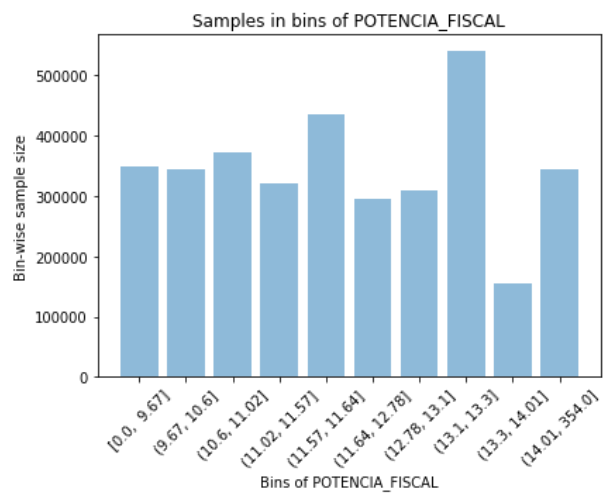
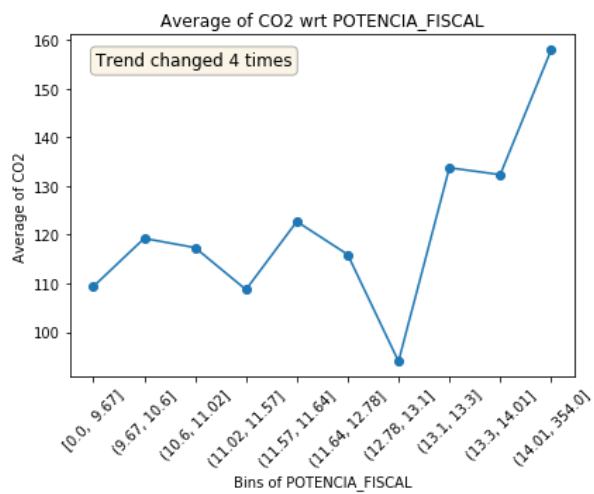
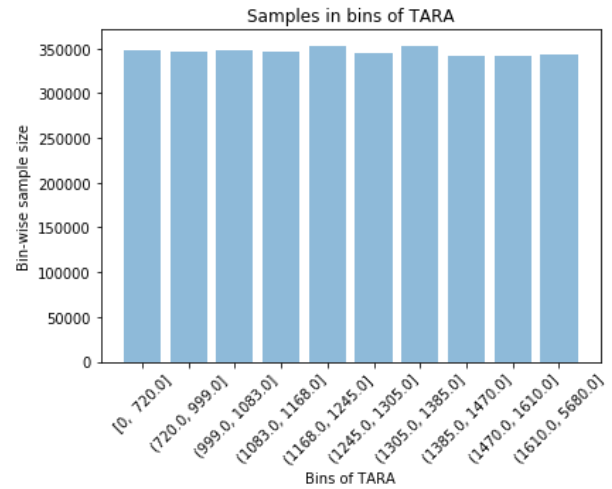
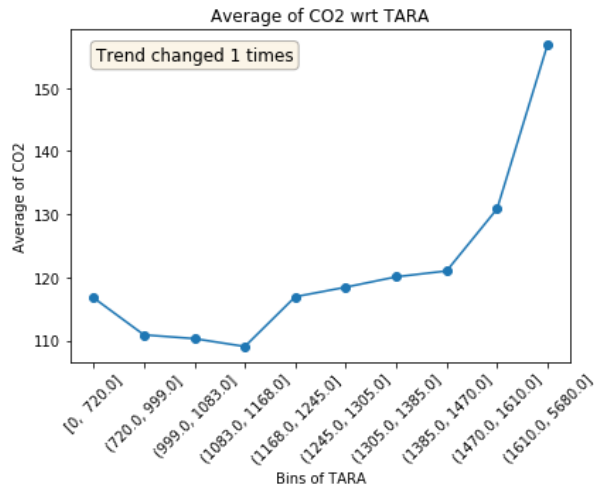


No obstant això, les correlacions amb la variable objectiu poden ser no lineals. Apliquem aquí una tècnica que ens ajuda a identificar-les visualment, i que consisteix en calcular per a cada variable rangs que continguin una quantitat similar de valors, i després per a cada un dels rangs calcular la mitjana de la variable objectiu. A continuació podem veure per a cada variable de la base de dades, el gràfic de la dreta mostra l'histograma de valors en cada rang (notar que els rangs no estan equiespaiats), i a l'esquerra la mitjana de valors de CO2 trobats per a cada rang, indicant quantes vegades canvia la tendència.







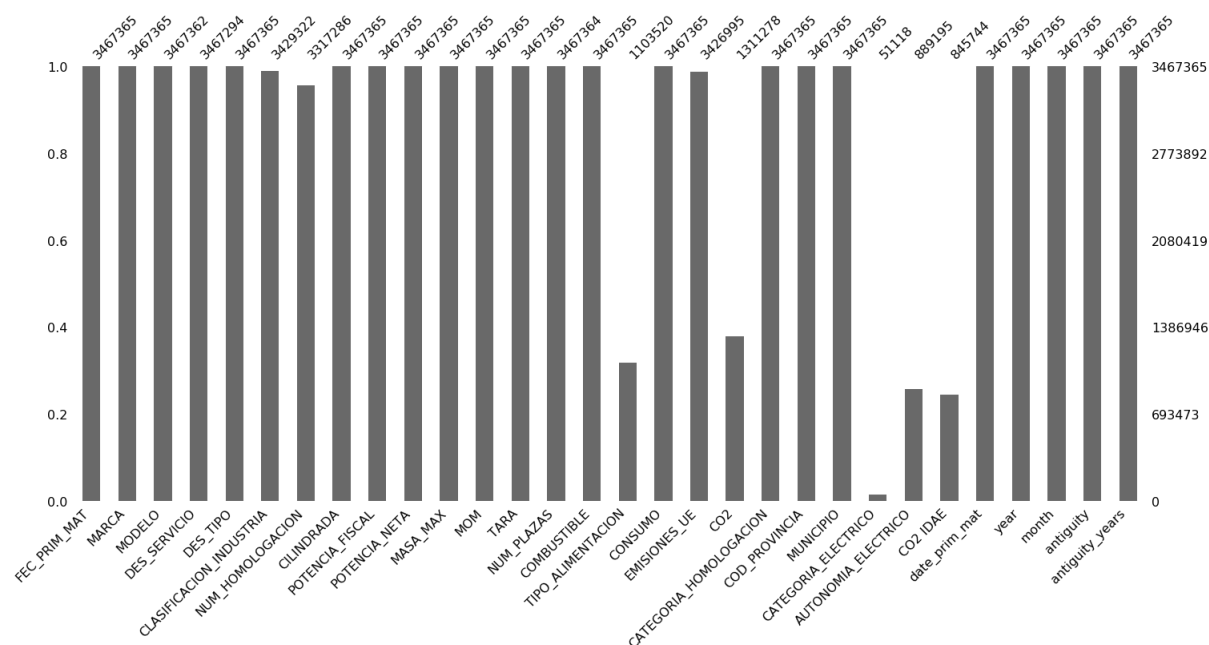


Aquí observem diversos punts interessants: el consum no és una variable molt rellevant perquè hi ha molts valors que falten (són zero). La potència fiscal té una dependència amb l'emissió de CO2 molt complexa o conté bastant soroll, ja que canvia de tendència fins a 4 vegades.

## 2.2. Neteja

### 2.2.1. Valors que falten

Es va trobar a més una quantitat considerable de valors que falten en algunes variables que a priori poden ser significatives. Aquí veiem una gràfica de la quantitat de valors trobats per columna o variable de la base de dades:



Aquests valors que falten són els valors que no van ser introduïts a la base de dades a l'hora de l'inserció dels vehicles a la base de dades, és a dir, les variables on no es va introduir cap valor. Els valors que falten, com podem observar més endavant, no són sistemàtics, és a dir, contenen informació addicional sobre el valor real que hem de tenir en compte.

Segons Rubin<sup>1</sup>(1976), la classificació de valors que falten normalment es divideix en tres grups. Primer de tot hi ha els valors que falten completament aleatoris, en anglès *missing completely at random*, on la probabilitat de ser un valor que falta per a cada observació és la mateixa. Això implica que les causes per la qual hi ha valors que falten no estan relacionades amb la col·lecció de les dades. En aquest cas es poden ignorar les complexitats que comporta l'existència de valors que falten, a part de la pèrdua d'informació intrínseca. Segon, hi ha els valors que falten aleatoris, en anglès *missing at random*, on la probabilitat de ser un valor faltant és la mateixa només dins dels grups definits. Aquest cas s'assembla molt més al nostre. Hem d'estudiar en més detall aquests, ja que ens ajudaria a omplir les dades de forma més precisa. Finalment, si cap de les suposicions anteriors es compleix, llavors estem parlant de valors que falten no aleatoris,

<sup>1</sup> 1976. "Inference and Missing Data." Biometrika 63 (3): 581–90.

en anglès *missing not at random*. Això significa que els valors que falten contenen informació addicional a tenir en compte a l'hora de modelitzar les dades. De forma matemàtica normalment s'expressen de la següent manera:

$$MCAR : P (R | Z_{comp}) = P (R)$$

$$MAR : P (R | Z_{comp}) = P (R | Z_{obstant})$$

$$MNAR : P (R | Z_{comp}) = P (R | Z_{obstant}, Z_{meus})$$

Per al nostre cas, volem trobar els grups en què la suposició *MAR* es compleixi, ja que ens permetria imputar els valors que falten dins dels grups com si fossin *MCAR*.

En el següent gràfic podem observar tots els patrons de valors per emplenar. Aquest gràfic representa totes les combinacions de patrons de valors mancants que s'observen en les dades, després de remoure els valors que considerem com a erronis, que van ser introduïts a l'hora d'omplir la base de dades. Més endavant explicarem en més detall com interpretar aquest gràfic.



## 2.2.2. Outliers

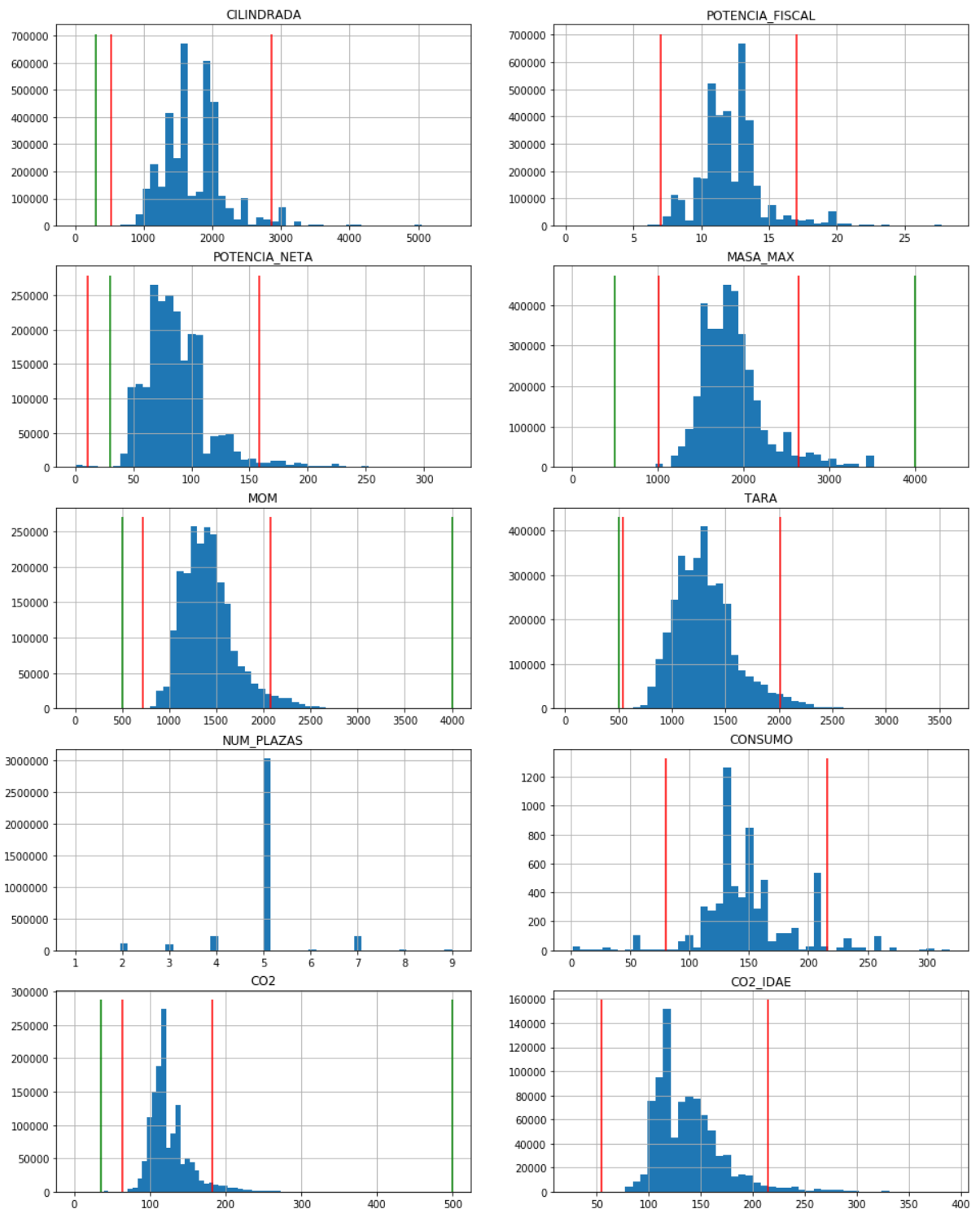
Explorant la distribució de valors de cada columna veiem també alguns patrons sospitosos, com que el mínim d'algunes columnes sigui zero (per exemple, cilindrada), o valors màxims (MOM, o MASA\_MAX) que estan fora d'escalles admissibles:

|                                | count     | mean        | std     | min        | 25%         | 50%         | 75%         | max         |
|--------------------------------|-----------|-------------|---------|------------|-------------|-------------|-------------|-------------|
| <b>FEC_PRIM_MAT</b>            | 3467365.0 | 20075267.79 | 69059.3 | 19000421.0 | 20021217.00 | 20070618.00 | 20140404.00 | 20181228.00 |
| <b>CLASIFICACION_INDUSTRIA</b> | 3429322.0 | 1006.57     | 98.6    | 10.0       | 1000.00     | 1000.00     | 1000.00     | 3300.00     |
| <b>CILINDRADA</b>              | 3467365.0 | 1724.39     | 493.4   | 0.0        | 1398.00     | 1598.00     | 1968.00     | 19680.00    |
| <b>POTENCIA_FISCAL</b>         | 3467365.0 | 12.17       | 2.7     | 0.0        | 10.74       | 11.64       | 13.19       | 354.00      |
| <b>POTENCIA_NETA</b>           | 3467365.0 | 52.68       | 50.6    | 0.0        | 0.00        | 63.00       | 88.00       | 999.00      |
| <b>MASA_MAX</b>                | 3467365.0 | 1767.87     | 739.8   | 0.0        | 1581.00     | 1800.00     | 2000.00     | 992720.00   |
| <b>MOM</b>                     | 3467365.0 | 852.60      | 1288.5  | -2200.0    | 0.00        | 1155.00     | 1435.00     | 802025.00   |
| <b>TARA</b>                    | 3467365.0 | 1171.41     | 468.2   | 0.0        | 1048.00     | 1245.00     | 1425.00     | 5680.00     |
| <b>NUM_PLAZAS</b>              | 3467364.0 | 5.05        | 0.7     | 0.0        | 5.00        | 5.00        | 5.00        | 55.00       |
| <b>CONSUMO</b>                 | 3467365.0 | 0.23        | 6.0     | 0.0        | 0.00        | 0.00        | 0.00        | 1340.00     |
| <b>CO2</b>                     | 1311278.0 | 121.06      | 30.9    | 0.0        | 107.00      | 118.00      | 135.00      | 999.00      |
| <b>COD_PROVINCIA</b>           | 3467365.0 | 14.28       | 11.5    | 8.0        | 8.00        | 8.00        | 17.00       | 43.00       |
| <b>MUNICIPIO</b>               | 3467365.0 | 14417.02    | 11551.0 | 8001.0     | 8077.00     | 8196.00     | 17079.00    | 48036.00    |
| <b>CO2 IDAE</b>                | 845744.0  | 137.19      | 31.8    | 0.0        | 115.00      | 132.00      | 154.00      | 499.00      |
| <b>year</b>                    | 3467365.0 | 2007.46     | 6.9     | 1900.0     | 2002.00     | 2007.00     | 2014.00     | 2018.00     |
| <b>month</b>                   | 3467365.0 | 6.47        | 3.4     | 1.0        | 4.00        | 6.00        | 10.00       | 12.00       |
| <b>antiguity_years</b>         | 3467365.0 | 11.40       | 6.9     | 0.0        | 5.10        | 11.90       | 16.41       | 119.13      |

Si considerem tots els valors iguals a zero com un outlier, la tasca que ens queda és identificar valors mínims i màxims raonables per a cada variable. Per això podem 1) definir els criteris arbitràriament, o 2) identificar dins la base de dades una definició de rangs típics.

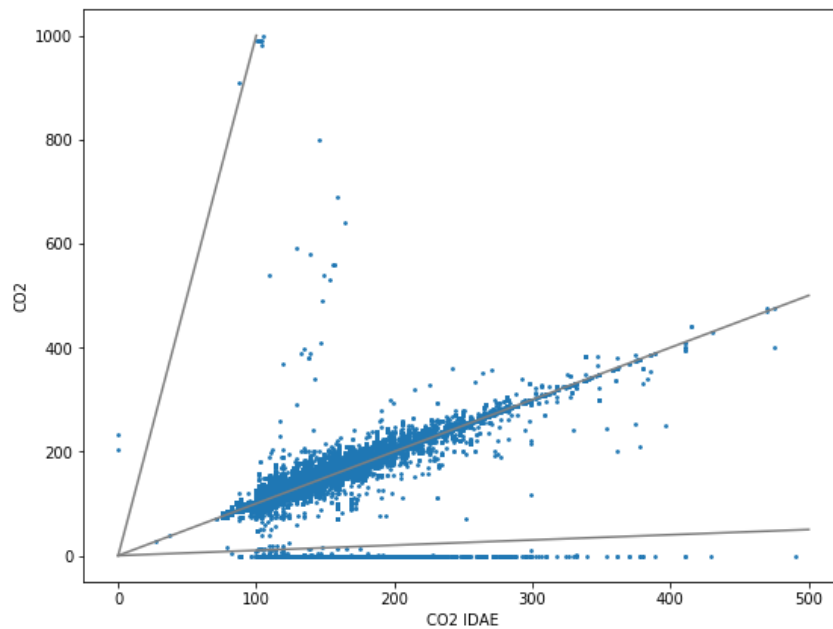
En el primer cas, fem servir com a definició dels criteris suggerits per la Secretaria de Medi Ambient i Sostenibilitat i la Secretaria d'Hisenda, especificant que els rangs acceptables de les variables seran entre 35 i 499 g/km de CO<sub>2</sub>, cilindrades superiors a 300 cm<sup>3</sup>, potència neta major a 30, totes les masses majors a 500 kg i menors a 4000 kg, i la resta que no hi ha condicions es demana que siguin majors de zero.

En el segon cas apliquem el criteri de Tukey, en el qual es consideren com outliers a qualsevol valor que es trobi més lluny de 1.5 vegades el rang interquartil. És a dir, si Q1 i Q3 són els valors del primer i l'últim quartil, el rang interquartil es defineix com  $IQR = Q3 - Q1$ , i estaríem descartant tots els valors que estan per sota de  $Q1 - 1.5 * IQR$  o per sobre de  $Q3 + 1.5 * IQR$ . En la següent gràfica mostrem els histogrames de cada variable (en barres blaves) al costat dels rangs de normalitat definits manualment (en verd), i pel criteri de Tukey (en vermell):



Com veiem, el rang proposat pel criteri de Tukey és més restrictiu que el del criteri manual. En particular, afecta en major quantitat els vehicles més grans emissors de CO2. És per això que ens vam quedar amb el criteri manual.

Hem d'explorar en més detall per entendre quines columnes estan correlacionades amb el CO2. La referència de CO2 proporcionada per IDAE hauria de ser gairebé idèntica (quan existeixen els dos valors), però no obstant això veiem que no és perfecte:

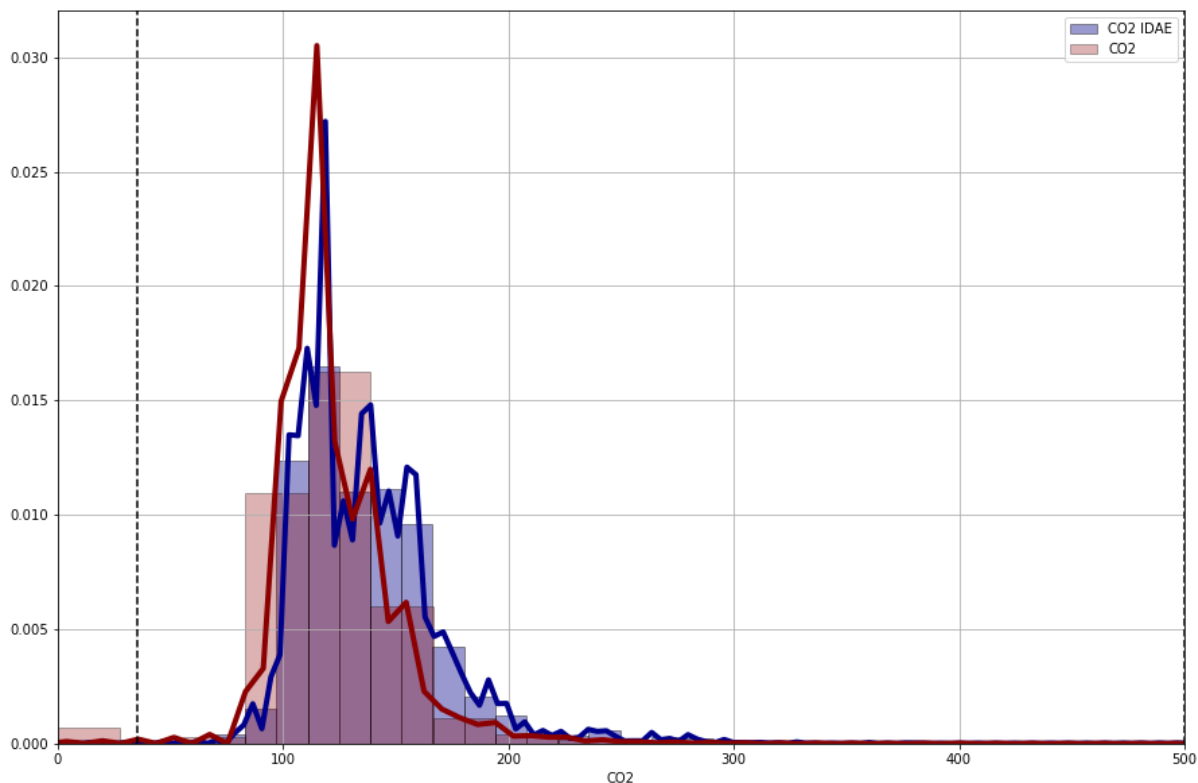


En el gràfic de dalt s'indiquen tres pendents: La identitat (central), que és quan el valor de CO2 coincideix amb el de CO2 IDAE, i les de baix i dalt que mostren quan el CO2 és un dècim o deu vegades el valor IDAE, que indiquen quan hi ha errors tipogràfics com el desplaçament del lloc de la coma cap a l'esquerra o dreta. A més a més, es veu que existeixen molts vehicles pels quals el valor de CO2 és igual a zero, mentre que sí tenen el valor indicat de l'IDAE.

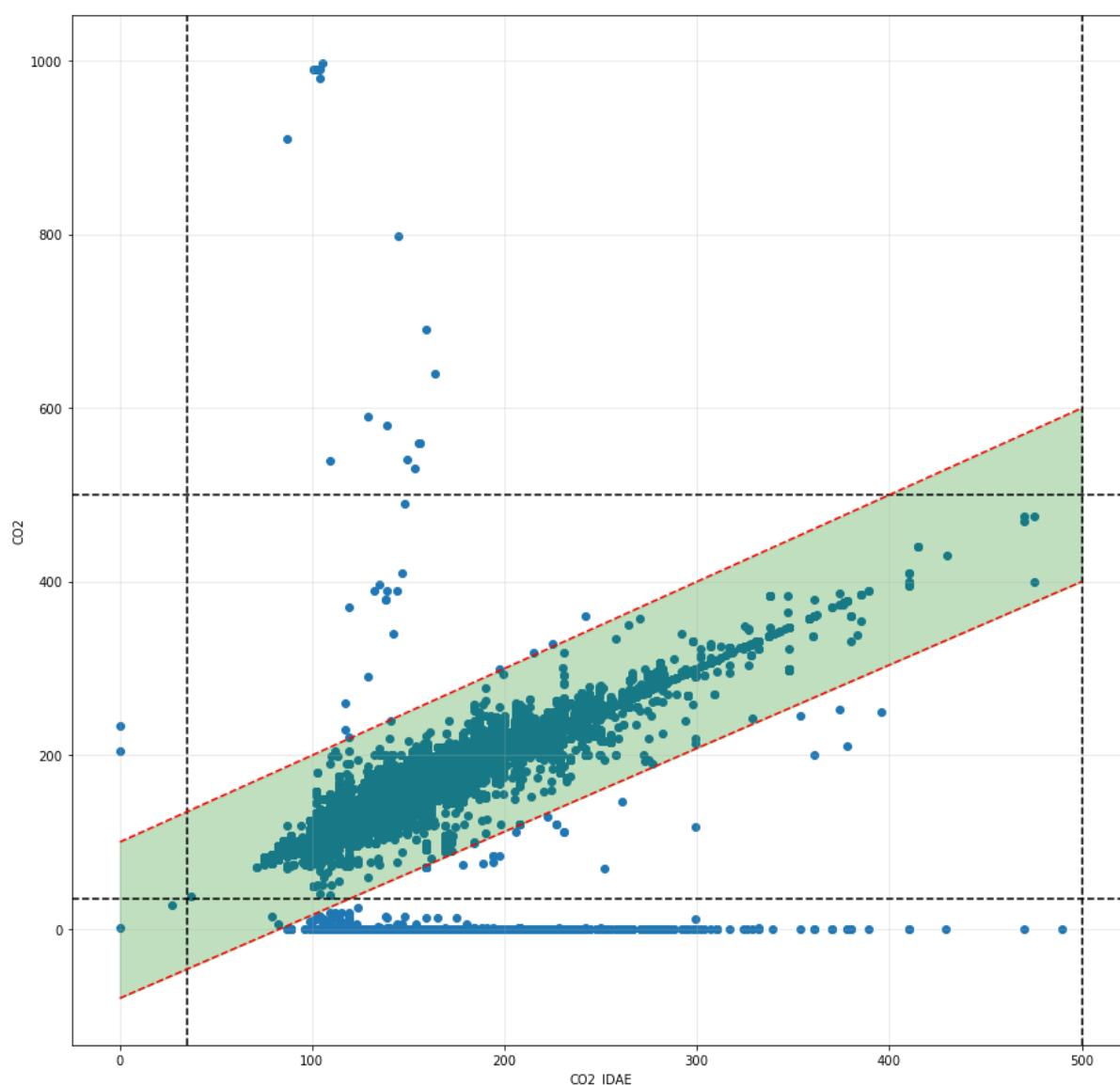
Els valors més petits de CO2 corresponen a alguns vehicles elèctrics, però hi ha moltes observacions que són indiscutiblement errors. A continuació ensenyem una mostra d'una taula filtrant per veure valors petits de CO2, amb errors obvis com un Peugeot 407 dièsel que emet només 1 g/km de CO2:

| MARCA        | MODELO                | CILINDRADA | POTENCIA_FISCAL | POTENCIA_NETA | MASA_MAX | MOM    | TARA | NUM_PLAZAS | COMBUSTIBLE | CO2 IDAE | CO2   |
|--------------|-----------------------|------------|-----------------|---------------|----------|--------|------|------------|-------------|----------|-------|
| DAIMLER BENZ | CHRYSLER              | 1796       | 12.48           | 120.00        | 1935     | 1495.0 | 0    | 5.0        | GASOLINA    | NaN      | 8.70  |
| HYUNDAI      | I30                   | 1396       | 10.74           | 73.19         | 1820     | 1260.0 | 1260 | 5.0        | GASOLINA    | NaN      | 10.74 |
| FORD         | FOCUS                 | 1753       | 12.31           | 0.00          | 1755     | 1465.0 | 1390 | 5.0        | DIESEL      | NaN      | 1.00  |
| BMW I        | I3                    | 647        | 14.55           | 125.00        | 1760     | 1440.0 | 1390 | 4.0        | ELÉCTRICO   | NaN      | 13.00 |
| OPEL         | CORSA E               | 1248       | 10.02           | 70.00         | 1720     | 1259.0 | 1259 | 5.0        | DIESEL      | 101.0    | 10.00 |
| BMW I        | I3                    | 647        | 14.55           | 125.00        | 1760     | 1440.0 | 1390 | 4.0        | ELÉCTRICO   | NaN      | 13.00 |
| BMW I        | I3                    | 647        | 14.55           | 125.00        | 1730     | 1390.0 | 1390 | 4.0        | ELÉCTRICO   | NaN      | 13.00 |
| VOLKSWAGEN   | TOURAN                | 1968       | 13.19           | 100.00        | 2260     | 1675.0 | 1675 | 5.0        | DIESEL      | NaN      | 9.75  |
| VOLKSWAGEN   | GOLF                  | 1598       | 11.64           | 0.00          | 1830     | 1337.0 | 1337 | 5.0        | DIESEL      | NaN      | 12.00 |
| PEUGEOT      | 407 SR SPORT HDI      | 1560       | 11.47           | 80.00         | 2020     | 1437.0 | 1437 | 5.0        | DIESEL      | 140.0    | 1.00  |
| BMW I        | I3                    | 647        | 14.55           | 125.00        | 1760     | 1440.0 | 1390 | 4.0        | ELÉCTRICO   | NaN      | 12.00 |
| VOLKSWAGEN   | TIGUAN                | 1968       | 13.19           | 103.00        | 2250     | 1654.0 | 1654 | 5.0        | DIESEL      | NaN      | 13.19 |
| BMW          | I3                    | 647        | 14.55           | 125.00        | 1730     | 1390.0 | 1315 | 4.0        | ELÉCTRICO   | NaN      | 13.00 |
| TOYOTA       | TOYOTA YARIS          | 1329       | 10.42           | 73.00         | 1490     | 1117.0 | 0    | 5.0        | GASOLINA    | 114.0    | 5.00  |
| VOLKSWAGEN   | TRANSPORTER CARAVELLE | 2461       | 16.48           | 0.00          | 3000     | 2395.0 | 2320 | 9.0        | DIESEL      | NaN      | 1.00  |
| BMW          | I3                    | 647        | 14.55           | 125.00        | 1730     | 1390.0 | 1315 | 4.0        | ELÉCTRICO   | NaN      | 13.00 |
| DS           | DS 4 1.2 PURETECH S&S | 1199       | 8.73            | 96.00         | 1790     | 1330.0 | 1330 | 5.0        | GASOLINA    | 119.0    | 14.00 |
| OPEL         | ZAFIRA-A              | 1598       | 11.63           | 74.00         | 1950     | 1265.0 | 1190 | 7.0        | GASOLINA    | NaN      | 1.00  |

En el gràfic de més avall s'observa que la densitat dels valors de CO2 provinents d'IDAE està una mica desplaçada a la dreta. Atès que els valors de CO2 d'IDAE solen ser d'anys més anteriors, coincideix amb l'observació que els valors antics de CO2 tenen valors grans.



La inclusió del dataset d'IDAE es pot fer de diverses maneres. Per exemple, es pot considerar part de la veritat objectiva, encara que això no pot ser directe perquè com ja vam veure els dos valors no coincideixen en molts casos. Una altra forma és que es pot utilitzar com a reemplaçament del valor vertader quan aquest no existeix, és a dir, prendre com a veritat absoluta els valors de CO<sub>2</sub> quan n'hi ha, i sinó, els d'IDAE si existeixen. També es podria considerar construir una combinació dels dos valors, a partir de criteris que permetin avaluar la probabilitat que algun sigui més correcte que l'altre. Finalment, podria no utilitzar-se directament per a l'entrenament, si no com una mesura de prova o validació dels models de més alt nivell, per tal de tenir una avaluació del rendiment independent del dataset i la partició original entre entrenament/prova. En qualsevol cas, cal tenir en compte la diferència estadística entre els dos datasets (tal com la que es mostra més amunt), que pot introduir biaixos complexos i perjudicar el model silenciosament.

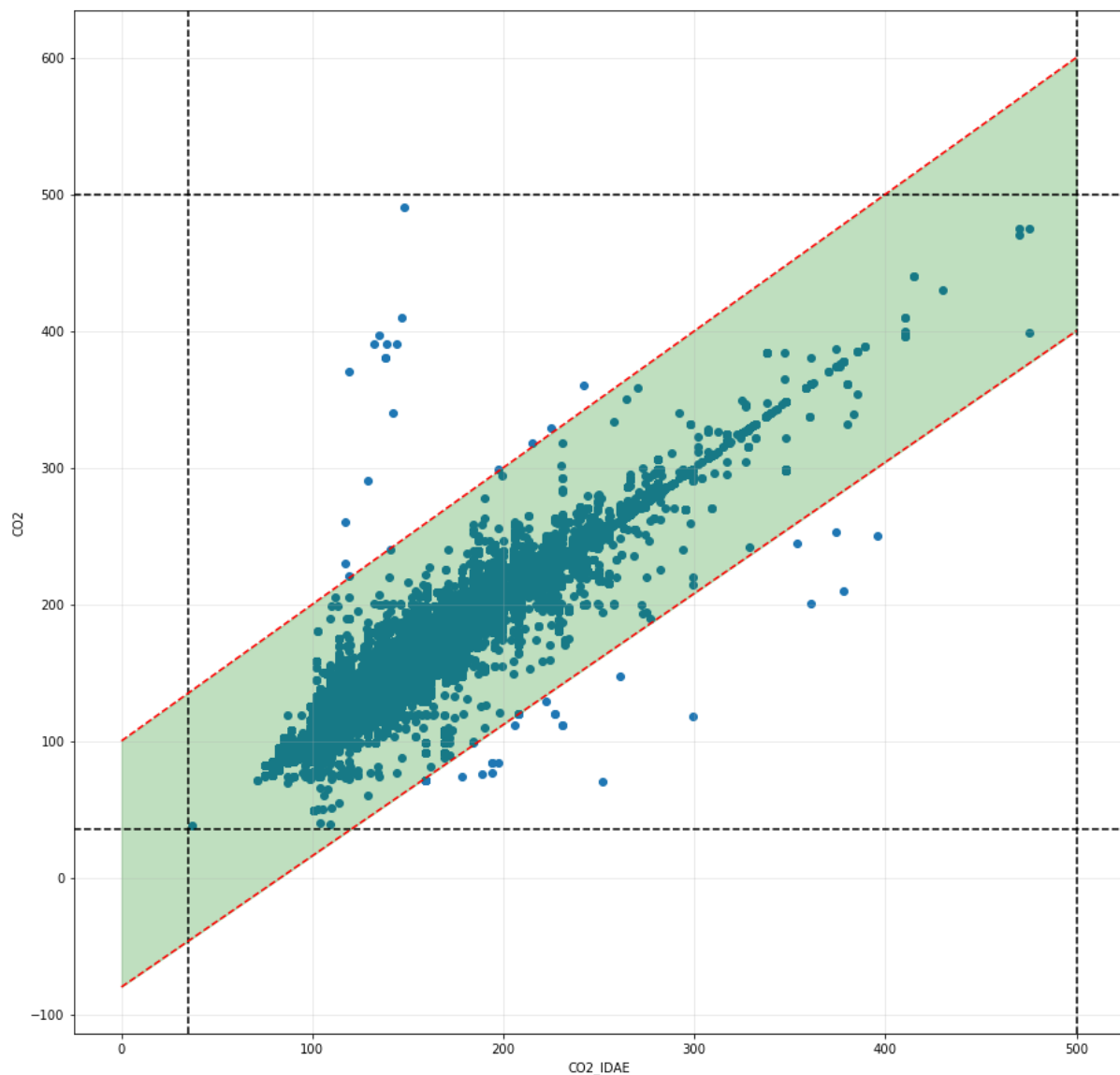


En el gràfic anterior es mostra un possible rang de valors que considerem vàlid per al CO2 d'IDAE. Aquest rang està compost entre les dues funcions lineals:

$$i_{low} = -80 + 0.96 * CO2$$

$$i_{high} = 100 + CO2$$

Per descomptat, això només podem fer-ho per als vehicles que podem observar ambdós valors de CO2. Per a la resta de vehicles no tenim cap referència de com estimar si el CO2 de l'IDAE és un valor vàlid o no. Les línies negres ratllades mostren els criteris establerts per als quals es consideren com a valors incorrectes. El següent gràfic mostra els valors que queden després de remoure els valors segons els límits establerts. Observem que encara hi ha valors que presenten incongruències entre les columnes de CO2 i CO2 d'IDAE.



El resultat dels anàlisis d'aquesta secció és que els valors considerats outliers han estat remoguts de la base de dades utilitzada per a fer les estimacions de CO2. Aquests vehicles que presenten valors erronis van ser etiquetats, per a la seva posterior imputació automàtica (veure més a baix).

### 2.2.3. Errors en l'escriptura de dades

En els camps de tipus caràcter, trobem molts valors introduïts amb errors ortogràfics o errors involuntaris introduïts a l'hora d'escriure les dades. Aquests valors els hem de considerar si volem utilitzar aquestes columnes.

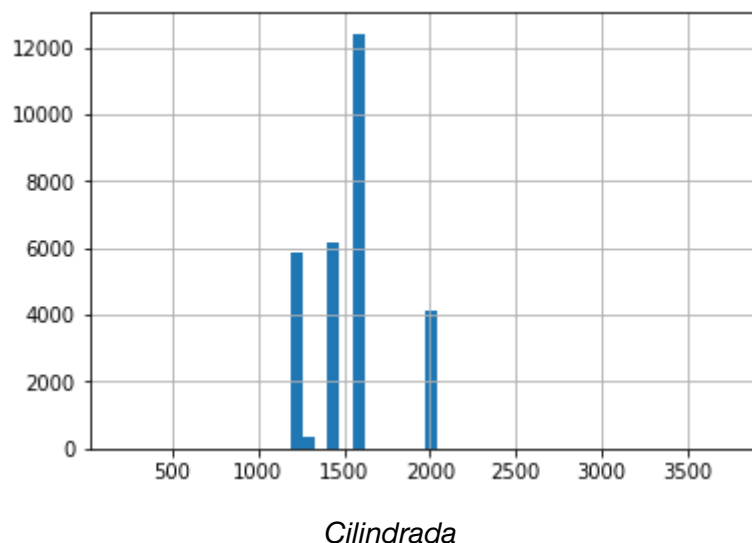
Per exemple, el Nissan Qashqai és un cotxe relativament popular. Si busquem els models que contenen les lletres "QASH" en l'exacte mateix ordre apareixen més de 40 diferents models únics, la majoria només un cop:

#### **MODEL Observacions**

NISSAN QASHQAI 30099  
NISSAN QASHQAI + febrer 2260  
NISSAN QASHQAI 37  
NISSAN QASHQAI febrer 20  
NISSAN QASHQAI, febrer 17  
NISSAN QASHQAI +2 15  
NISSAN QASHQAI2 14  
NISSAN QASHAQI 14  
NISSAN QASHQAI +2 14  
NISSAN QASHGAI 8  
Nissan QASHQAI 7  
NISSAN QASHAI 7  
NISSANQASHQAI 4  
NISSAN QASHQAI 4  
NISSAN QASHQI 3  
NISSAN QASHAQAI 3  
NISSN QASHQAI 2  
NISSAN QASHQAI'J10 2  
Nissa QASHQAI 2  
NISASN QASHQAI 1  
NSSAN QASHQAI 1  
Missan QASHQAI 1  
NISSAN QASHQAIç 1  
NISSAN QASHQAI \* gener 2  
NISSAN QASHQ2AI + 2 1  
NISSAN QASHOAI 1  
NISSAN QASHAAI + 2 1

Nissan QASHQAI + 2 1  
 NIDDSN QASHQAI 1  
 NISSAN QASHQ AI 1  
 NISSAN QASHQAI-2 1  
 NISSAN QASHQAU 1  
 NISSAN QASHQAI 4X4 1  
 NISSAN QASHQA1 1  
 NIOSSAN QASHQAI + 2 1  
 NISSAN QASHQAIÑJ10 1  
 NIISAN QASHQAI 1  
 NISAAN QASHQAI 1  
 BNISSAN QASHQAI + 2 1  
 NISSAN QASHDAI 1  
 NISSAN QASHQAI-J10 1  
 NISSAN QASHAQI + 2 1  
 NISSAN QASHQAI + 2 1  
 NISSAQN QASHQAI 1  
 NISSAB QASHQAI 1  
 Missan QASHQAI + 2 1

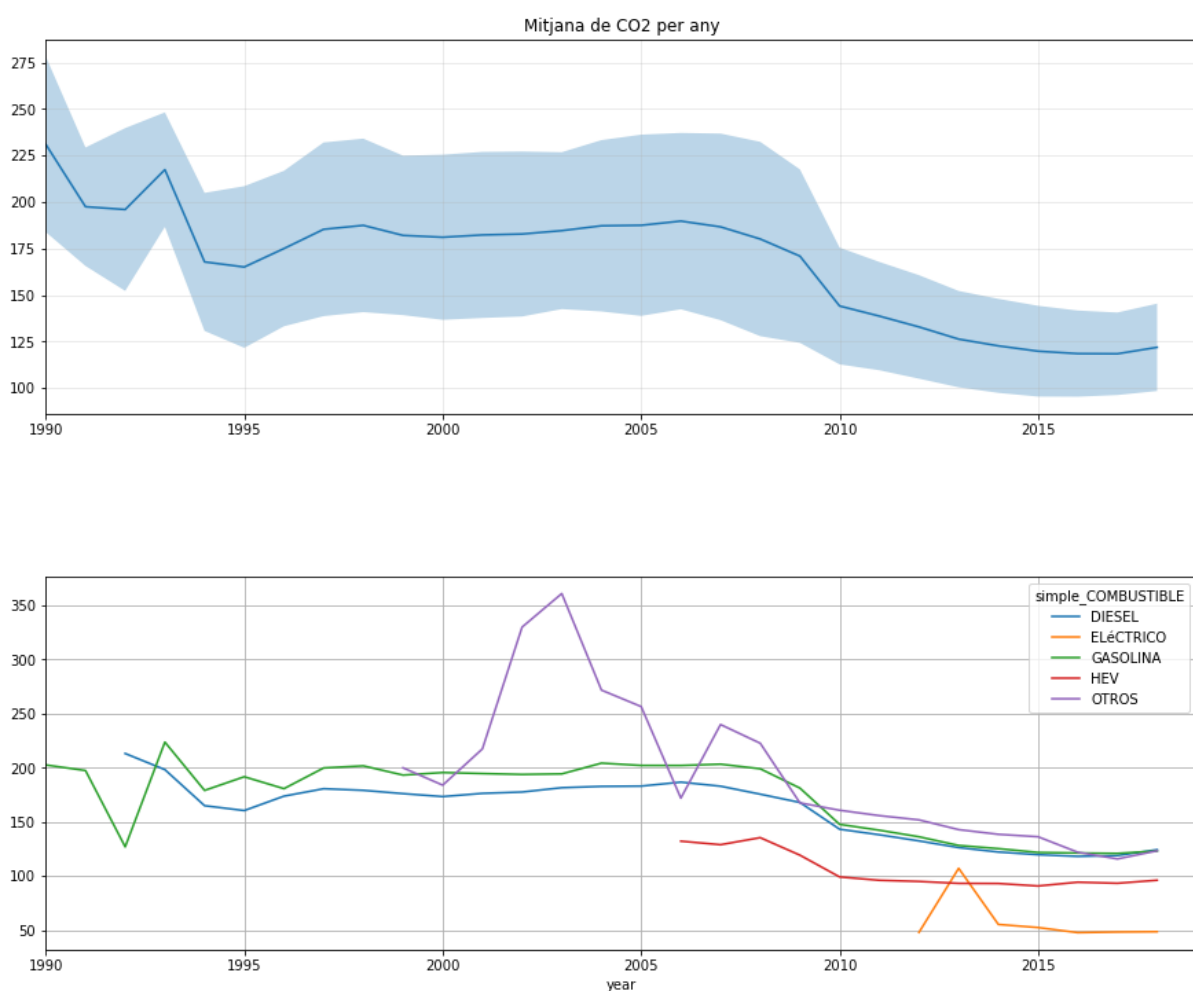
El nombre que apareix a la dreta dels noms de dalt mostra la quantitat de cops que un vehicle està registrat amb aquest model. Tot i això, si s'exploren altres variables sobre aquestes observacions que cauen dins dels més de 40 diferents models únics, s'observen que hi ha unes 3 o 4 tipologies generals i no 40. Per exemple, mirant l'histograma de valors trobats per a tots els models amb "QASH", veiem una distribució amb pocs valors diferents:



Una bona neteja de les columnes de marques i models serviria per a poder identificar i esmenar errors en les columnes de propietats tècniques dels vehicles, o millor encara en la mateixa estimació de CO2. *En els models proposats en aquest informe hi ha una molt alta*

possibilitat que dos vehicles idèntics (marca / model / any), que haurien de tenir les mateixes emissions, apareguin amb un càlcul de CO2 estimat diferent només perquè un d'ells té algun error en alguna de les seves característiques. Això no queda reflectit en un error en el model predictiu, sinó a la base de dades, que es podria corregir en un cas hipotètic, senzillament amb la neteja de les marques i models, i corregint els valors que falten/erronis basant-se no només en qualsevol vehicle similar, si no donant més pes a aquells que són el mateix model.

En les següents gràfiques podem observar com la mitjana d'emissions de CO2 per vehicle ha anat canviant amb el pas dels anys, tant com per al total de combustibles com per al desagregat:



Veient els gràfics de dalt, podem observar clarament que per als vehicles que anem a estimar les equacions, la distribució pot variar. Això pot resultar en subestimacions de les emissions a predir. Per tal de corregir aquest biaix, caldrà incloure una variable regressora que representi l'edat dels vehicles.

## 2.2.4. Altres errors

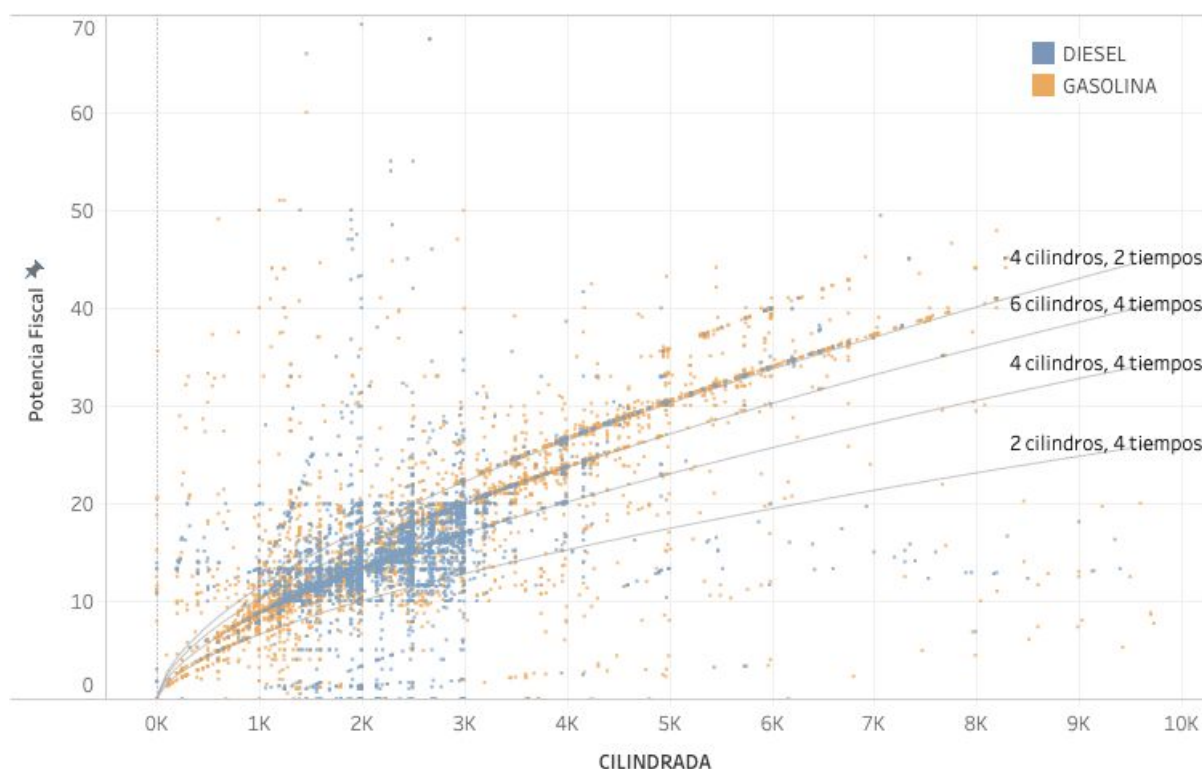
Ademés dels errors tipogràfics explicats en el subapartat anterior, també trobem errors similars però sobre les variables numèriques. Per exemple, variables numèriques que estan escrites amb comes (,) en comptes de (.). En general aquests errors són menors i són fàcils de netejar.

Com ja hem esmentat abans, un altre cas senzill són els valors anòmals. Situant tots els valors d'una variable en concret en ordre i observant quins són els valors més atípics (ja sigui massa petits o massa grans) podem descartar que aquests siguin valors adequats. Per exemple, una cilindrada menor a 10 el considerem com un valor mal introduït, o un nombre de places igual a 55, també es considera un valor erroni.

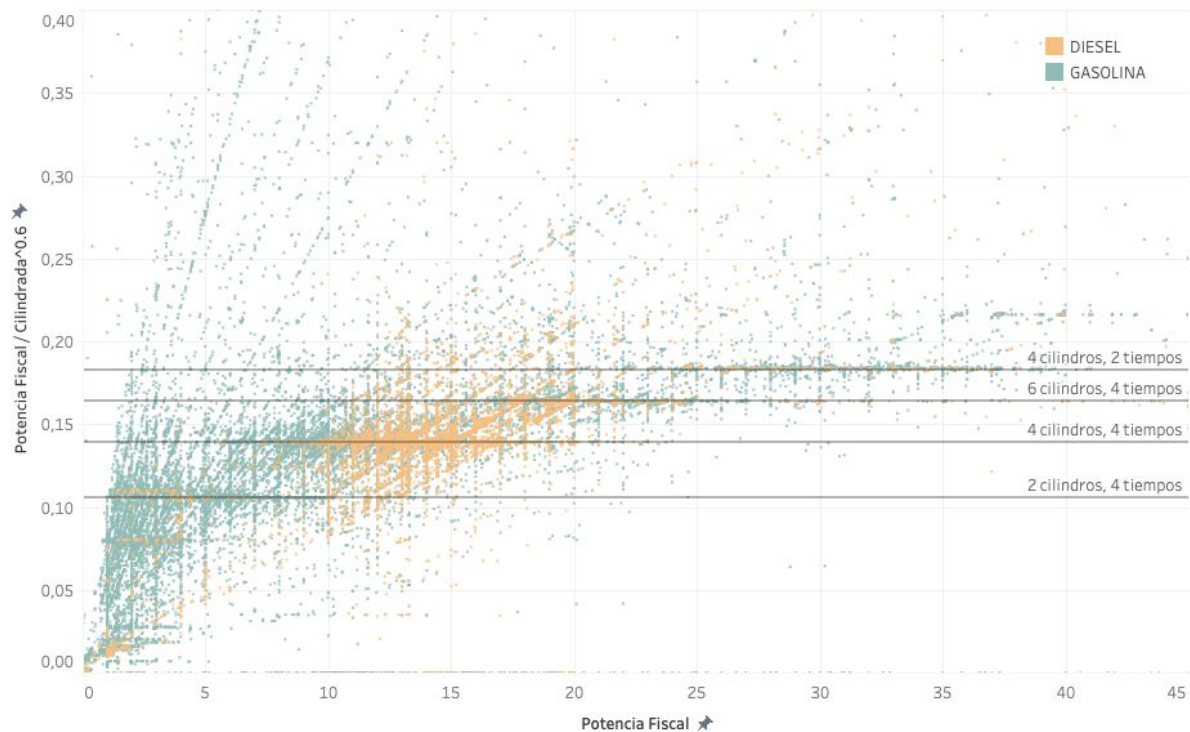
El problema més gran s'observa quan hi ha valors erronis i no tan diferents a la resta, de manera que no podem filtrar-los com outliers. Vam detectar aquests a través de la inspecció entre relacions entre variables. Per exemple, algunes variables són, en principi, un càlcul directe de les altres, com ara la potència fiscal que es calcula com:

$$\text{Potència fiscal} = t * (C / N)^{0,6} * N$$

On  $t$  és un factor numèric que depèn dels temps del motor,  $C$  és la cilindrada i  $N$  el nombre de cilindres.

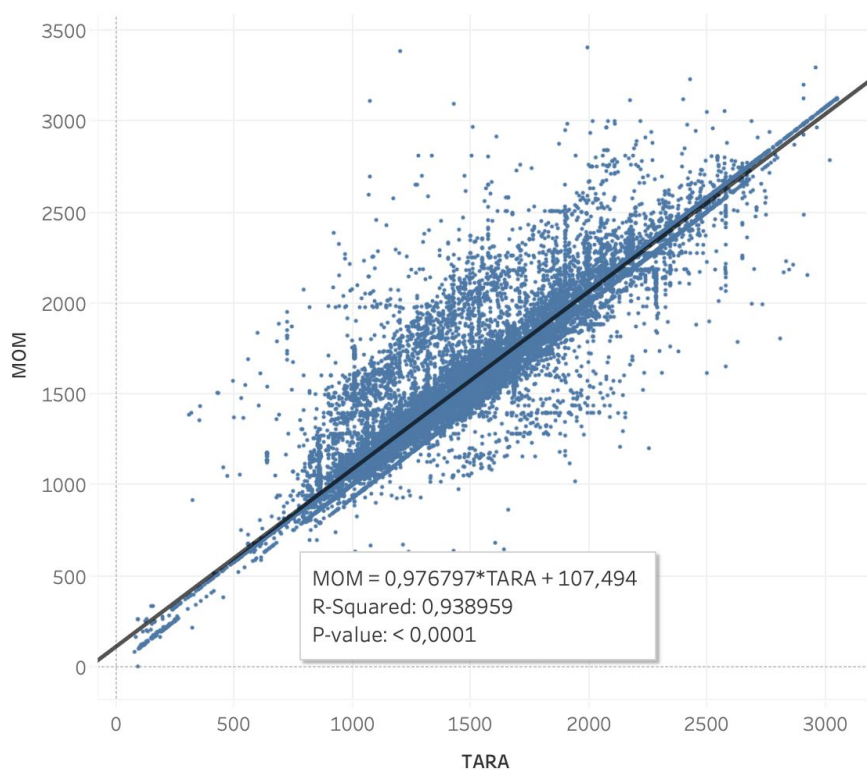


Si grafiquem el quocient entre la potència fiscal i la potència 0.6 de la cilindrada hauríem d'observar una sèrie de valors constants. No obstant això, observem el següent comportament (cal comparar amb les línies horitzontals que marquen alguns valors típics que es poden esperar):



A priori no podem saber si el valor correcte és la Cilindrada o la potència fiscal, però clarament hi ha molts vehicles per als quals existeix algun error.

Una altra exemple de relació lineal i senzilla és la de la de la massa en moviment, que en principi hauria de ser la tara més una constant típica (per exemple el pes d'un conductor). S'observa en canvi, una relació amb una dispersió relativament gran:



La tendència en el gràfic ens indica que de mitjana veiem la relació que esperem, però en molts casos la desviació és altíssima (més de 500 kg).

Finalment, hem notat que la variable DES\_TIPO presenta valors estranys. Es suposa que els vehicles que pertanyen a les categories d'homologació M1 i N1 tenen unes certes característiques a complir. Tot i això, observem errors. No tenim manera de corroborar si aquests errors pertanyen a la introducció de la variable DES\_TIPO o a la introducció de la categoria d'homologació (CATEGORIA\_HOMOLOGACION), tot i que sospitem que hi ha errors en els dos camps.

| CATEGORIA_HOMOLOGACION | DES_TIPO                              | CO2     | CO2_IDAE |
|------------------------|---------------------------------------|---------|----------|
| M1                     | CAMIÓN                                | 15      | 1        |
|                        | CAMIÓN ARTICULADO                     | 0       | 1        |
|                        | CAMIÓN CAJA                           | 21      | 0        |
|                        | CAMIÓN FRIGORÍFICO                    | 1       | 0        |
|                        | CAMIÓN FURGÓN                         | 51      | 0        |
|                        | CAMIÓN JAULA                          | 1       | 0        |
|                        | CAMIÓN MIXTO                          | 7       | 0        |
|                        | CAMIÓN PLATAFORMA                     | 1       | 0        |
|                        | CUATRICICLO PESADO                    | 6       | 0        |
|                        | FURGONETA                             | 253     | 33       |
|                        | FURGONETA MIXTA                       | 59      | 1        |
|                        | MOTOCICLETA DE 2 RUEDAS SIN SIDECAR   | 95      | 1        |
|                        | TODO TERRENO                          | 36816   | 26757    |
|                        | TRACTOCAMIÓN                          | 0       | 1        |
|                        | TURISMO                               | 1248161 | 818949   |
| N1                     | CAMIONETA                             | 12      | 4        |
|                        | CAMIÓN                                | 507     | 151      |
|                        | CAMIÓN ARTICULADO                     | 1       | 19       |
|                        | CAMIÓN ARTICULADO CAJA                | 0       | 1        |
|                        | CAMIÓN ARTICULADO CONTRA INCENDIOS    | 2       | 2        |
|                        | CAMIÓN ARTICULADO FRIGORÍFICO         | 0       | 0        |
|                        | CAMIÓN ARTICULADO FURGÓN              | 0       | 3        |
|                        | CAMIÓN ARTICULADO GRÚA                | 1       | 1        |
|                        | CAMIÓN ARTICULADO HORMIGONERA         | 0       | 0        |
|                        | CAMIÓN ARTICULADO JAULA               | 0       | 0        |
|                        | CAMIÓN ARTICULADO PLATAFORMA          | 0       | 0        |
|                        | CAMIÓN ARTICULADO TALLER              | 0       | 0        |
|                        | CAMIÓN ARTICULADO VIVIENDA O CARAVANA | 0       | 2        |
|                        | CAMIÓN ARTICULADO VOLQUETE            | 163     | 1        |
|                        | CAMIÓN BASURERO                       | 11      | 0        |
|                        | CAMIÓN BOTELLERO                      | 0       | 1        |
|                        | CAMIÓN CAJA                           | 5044    | 1420     |
|                        | CAMIÓN CISTERNA                       | 0       | 0        |
|                        | CAMIÓN FRIGORÍFICO                    | 531     | 180      |
|                        | CAMIÓN FURGÓN                         | 691     | 2774     |
|                        | CAMIÓN ISOTERMO                       | 28      | 40       |
|                        | CAMIÓN JAULA                          | 1       | 0        |
|                        | CAMIÓN MIXTO                          | 18      | 3        |
|                        | CAMIÓN PARA CANTERA                   | 0       | 1        |
|                        | CAMIÓN PLATAFORMA                     | 5       | 0        |
|                        | CAMIÓN PORTACONTENEDORES              | 1       | 2        |
|                        | CAMIÓN PORTAVEHÍCULOS                 | 39      | 15       |
|                        | CAMIÓN SILO                           | 0       | 2        |
|                        | CAMIÓN TALLER                         | 1       | 9        |
|                        | CUATRICICLO PESADO                    | 0       | 0        |
|                        | FURGONETA                             | 58402   | 24268    |
|                        | FURGONETA MIXTA                       | 349     | 931      |
|                        | MOTOCICLETA DE 2 RUEDAS SIN SIDECAR   | 2       | 0        |
|                        | TODO TERRENO                          | 35      | 105      |
|                        | TRACTOCAMIÓN                          | 3       | 2        |
|                        | TURISMO                               | 566     | 1687     |

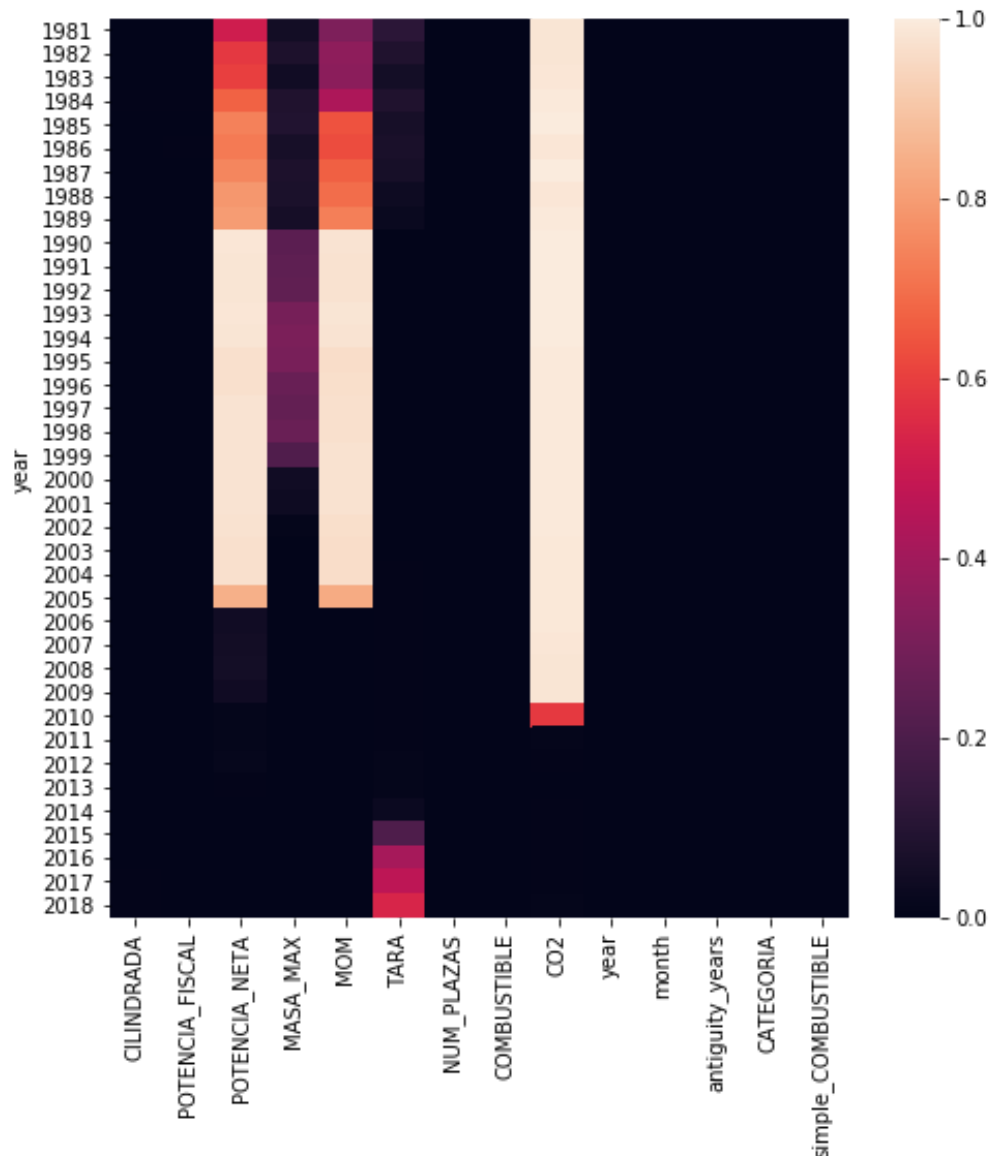
## 2.3. Valors mancants

Finalment discutim un altre tipus d'error observat que és el dels valors que falten a la base de dades. Solucionar aquest error és crucial per a l'objectiu del projecte, ja que per a poder calcular les emissions de tots els vehicles mitjançant regressions lineals hem de tenir totes les dades disponibles que siguin necessàries per als models. Aquests models no admeten valors mancants. Més endavant discutirem els mètodes d'imputació que van ser explorats i utilitzats.

Observem que només algunes variables presenten valors buits o valors mancants. Com ja hem indicat abans, molt probablement els valors que falten no siguin aleatoris. Això ens indica que els valors que falten contenen informació addicional per al dataset. Quan els valors que falten són aleatoris, la imputació resulta més senzilla, ja que es poden imputar amb valors estàndard com la mitjana o la mediana. En canvi, quan no és el cas, hem de fer servir tècniques més complexes.

Després de remoure les dades atípiques i reemplaçar-les per valors mancants, les variables que presenten valors que falten per al subconjunt d'observacions amb *CATEGORIA\_HOMOLOGACION* M1 i N1 són:

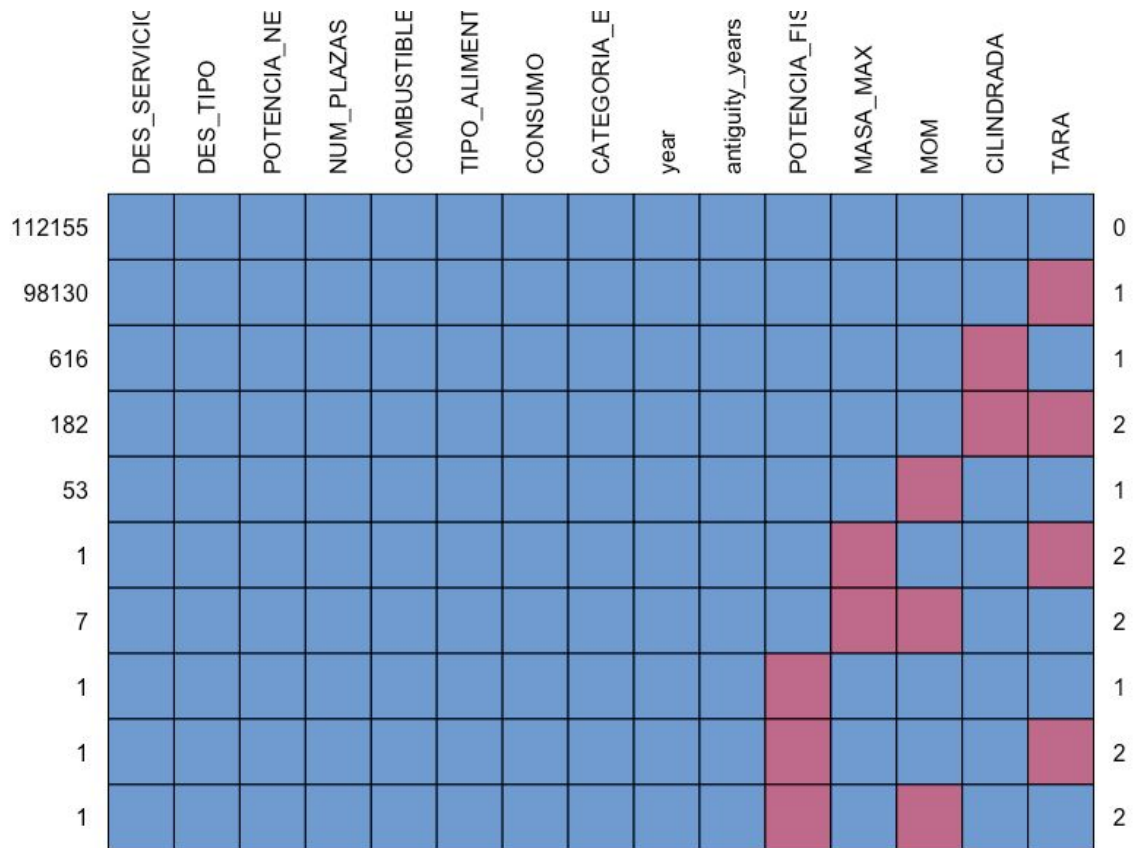
- *MASA\_MAX* per als anys anteriors a 1990
- *MOM* per als anys anteriors a 2005
- *TARA* a partir de 2014 en endavant
- *TIPO\_ALIMENTACION* per anys anteriors a 2012
- *EMISIONES\_UE* per anys anteriors a 1992
- *CO2* per anys anteriors a 2010
- *AUTONOMICO\_ELECTRICO* per anys anteriors a 2014



La diferència entre variables categòriques i numèriques també és molt important. Alguns mètodes d'imputació només serveixen per omplir variables numèriques i d'altres només per a variables categòriques. També hi han altres tècniques, com és el cas de l'ús de regressors més avançats emprats per a omplir aquests valors mancants, que serveixen per a ambdós casos.

Les observacions sense valors mancants o sense valors atípics les considerem vàlides, o que no han de ser estimades, sense tenir en compte si són correctes o no. Les variables que contenen valors que falten necessiten que aquests siguin estimats, i intentarem fer servir els valors que sí tenim per tal de fer-ho. Cada observació té un diferent patró de valors buits. Si agrupem aquests patrons entre els diferents grups de variables que tenen diferents patrons (considerem que tenen un mateix patró si contenen exactament les

mateixes variables amb valors buits) observem el següent gràfic dels patrons dels valors que falten per a l'any 2018:



A dalt veiem en l'eix Y dues escales, el valor de l'esquerra mostra la quantitat d'observacions que contenen aquest patró, mentre que el de la dreta mostra la quantitat d'observacions que contenen exclusivament aquests valors mancants. L'eix X conté els noms de les diferents variables. Llavors, les columnes indiquen les diferents variables avaluades i les files indiquen la quantitat d'observacions que tenen concretament aquests valors que falten, sent vermell quan contenen valors que falten i blau quan no. Podem observar que els dos grups amb major quantitat d'observacions són les observacions que no presenten valors que falten (112.155), seguit d'observacions que només tenen com a valor buit la TARA (98.130). Finalment, la resta de patrons de valors que falten representen un petit percentatge del nombre total d'observacions (~ 0.4%).

Un problema a tenir en compte és que mai podem observar els valors que falten reals. L'única forma d'avaluar el biaix o error de la imputació és mitjançant la imputació de valors mancants artificials que produïm només per a aquest propòsit, extraient valors el més similars possibles als valors que falten originals, imputar-los, i observar l'error que es produeix.

Després de l'exploració dels valors que falten, es va observar un clar empitjorament en la col·lecció de les dades per als anys més antics. És a dir, a banda de que les dades de CO2 pateixin un gran deteriorament per als anys previs a 2002 i entre els anys 2002 i 2008, també la col·lecció en general de dades resulta més errònia. Per exemple, dades que han estat considerades com outliers solen correspondre a anys més antics a la base de dades.

Hi ha diferents formes d'avaluar l'estimació, però la més precisa és la d'estimar exactament l'error d'imputació sobre cadascuna de les variables, és a dir, iterar sobre cadascuna de les variables a imputar i estimar quin és l'error sobre la imputació d'aquesta variable per separat de la resta.

L'objectiu de la imputació és llavors trobar grups on els valors que falten dins del grup siguin aleatoris. Sabent que és una suposició forta, hem de trobar la forma en què els errors estimats d'imputació siguin els més petits possibles. En la següent secció expliquem els mètodes utilitzats per tal de trobar els grups on els valors que falten siguin aleatoris.

## 3. Models d'imputació

### 3.1 Missing Forests

La tècnica Missing Forest és una tècnica d'imputació de valors que falten no paramètrica la qual permet imputar tant variables numèriques com variables categòriques. La tècnica consisteix en construir un bosc d'arbres de decisió per tal de predir les variables desitjades que contenen valors buits. D'aquesta manera, es genera un bosc per a cadascuna de les variables que es requereix imputar i llavors s'estima quin és l'error de cadascun d'aquests. Aquest model resulta adequat per al nostre objectiu gràcies a la gran quantitat d'observacions que conté la base de dades.

El paquet de *R* utilitzat per fer aquesta imputació es diu "*missForest*" i ens permet estimar els errors a l'hora de fer la imputació. Per una banda, el mètode utilitzat per a estimar els errors s'anomena *Out of Bag Error* (OOB error). *Stekhoven and Buhlmann* [2011] varen demostrar que aquesta tècnica representa una estimació vàlida dels errors de la imputació. Per altra banda, la mètrica utilitzada per a estimar els errors és el *MSE* per a variables numèriques i *PFC* (*proportion of falsely classified* en anglès o proporció de variables mal classificades) per a les variables categòriques. En aquest estudi només imputem variables numèriques.

Finalment, esmentar una frase de Joseph L. Schafer que diu: "*Failure of an imputation model does not damage the integrity of the entire dataset, but only the portion that is imputed.*"

## 3.2. Imputació per any

El primer mètode emprat és el de separar la base de dades a imputar entre els diferents anys en què les observacions van ser introduïdes a la base de dades, mitjançant la variable *FECHA\_PRIM\_MAT*. Aquesta variable ens permet separar les observacions en diferents grups, dins dels quals volem que els valors que falten siguin aleatoris. Sabent que els valors que falten varien molt amb la data de matriculació, separar entre diferents anys és un mètode que a priori sembla vàlid. És important remarcar que això no implica que els valors que falten dins dels diferents anys siguin aleatoris, però és definitivament una tècnica millor que agrupar-los tots dins del mateix grup, ja que com hem pogut observar anteriorment, els valors que falten estan molt lluny de ser aleatoris. La següent taula mostra els errors amb la mètrica RMSE estimats per a cada un dels anys imputats i cada una de les variables imputades:

|       | CILINDRADA | POTENCIA_FISCAL | MASA_MAX    | MOM         | TARA       |
|-------|------------|-----------------|-------------|-------------|------------|
| X2018 | 241.698348 | 1.604102        | 191.501044  | 135.062179  | 149.845839 |
| X2017 | 237.996675 | 1.436056        | 190.599848  | 132.410545  | 137.749160 |
| X2016 | 248.580743 | 1.464033        | 2111.889503 | 543.232712  | 161.132428 |
| X2015 | 260.122271 | 1.477731        | 201.313429  | 3609.659953 | 155.602844 |
| X2014 | 248.813188 | 1.467327        | 202.188837  | 1566.627831 | 145.970308 |
| X2013 | 244.378032 | 1.425251        | 214.219118  | 714.640159  | 145.084280 |
| X2012 | 265.999741 | 1.436584        | 226.935936  | 195.187704  | 151.387101 |
| X2011 | 264.940714 | 1.214794        | 224.242373  | 168.108935  | 150.870931 |
| X2010 | 271.290241 | 1.258630        | 218.987659  | 155.997790  | 151.168198 |
| X2009 | 276.985605 | 1.330651        | 209.479361  | 168.413495  | 143.853206 |
| X2008 | 315.770725 | 1.419083        | 205.949287  | 172.979932  | 153.874613 |
| X2007 | 295.664316 | 1.469645        | 198.562920  | 1734.113856 | 152.326320 |
| X2006 | 280.781006 | 1.472808        | 203.145553  | 1761.825925 | 128.156702 |
| X2005 | 287.751389 | 1.574038        | 414.303670  | 1061.320056 | 130.730875 |
| X2004 | 273.804973 | 1.656300        | 182.610914  | 246.060956  | 154.386501 |
| X2003 | 271.360362 | 1.608666        | 183.012171  | 555.091551  | 148.383546 |
| X2002 | 264.988248 | 1.507766        | 688.324811  | 216.858444  | 136.827423 |
| X2001 | 262.106268 | 1.421222        | 191.870436  | 277.799510  | 131.253889 |
| X2000 | 260.149969 | 0.000000        | 194.210767  | 285.345837  | 133.656452 |
| X1999 | 0.000000   | 1.479811        | 212.278398  | 247.280427  | 146.561256 |
| X1998 | 0.000000   | 0.000000        | 3783.634294 | 266.740831  | 144.539492 |
| X1997 | 326.612471 | 1.818902        | 206.076572  | 292.463648  | 151.768756 |
| X1996 | 335.867586 | 1.886786        | 222.518476  | 464.506874  | 156.482250 |
| X1995 | 0.000000   | 2.125307        | 365.853755  | 356.407069  | 168.162461 |
| X1994 | 0.000000   | 0.000000        | 245.069154  | 621.419443  | 140.709057 |
| X1993 | 0.000000   | 0.000000        | 211.911460  | 205.308294  | 153.831839 |
| X1992 | 359.850108 | 0.000000        | 207.522621  | 181.154128  | 152.543898 |
| X1991 | 0.000000   | 0.000000        | 231.220873  | 208.688568  | 194.145922 |
| X1990 | 0.000000   | 0.000000        | 215.749538  | 226.652721  | 170.386697 |

A la taula anterior observem que hi ha certs anys per als quals la imputació funciona molt pitjor, com podem observar per MASA\_MAX en l'any 1998 o la variable MOM l'any 2015. Els anys on apareix un 0 és que no hi ha valors a imputar.

### 3.3. Imputació per grups d'anys

El segon mètode és bastant similar a l'anterior, amb la diferència que els grups són d'anys consecutius i no incloents. La tècnica, per tant, és molt similar a l'anterior, només que canvien els grups d'anys per als quals s'imputen les variables alhora. És a dir, la imputació en comptes de ser any per any es fa en grups de dos anys. Primer el període 2017-2018, després 2015-2016, etc... Aquesta tècnica s'ha fet servir ja que no sabem els anys en què els cotxes són introduïts per primera vegada ni en quins anys quins cotxes són més comuns, per tant s'intenta crear subgrups per als quals els valors que falten puguin ser similars. A continuació ensenyem la taula d'errors d'aquesta metodologia:

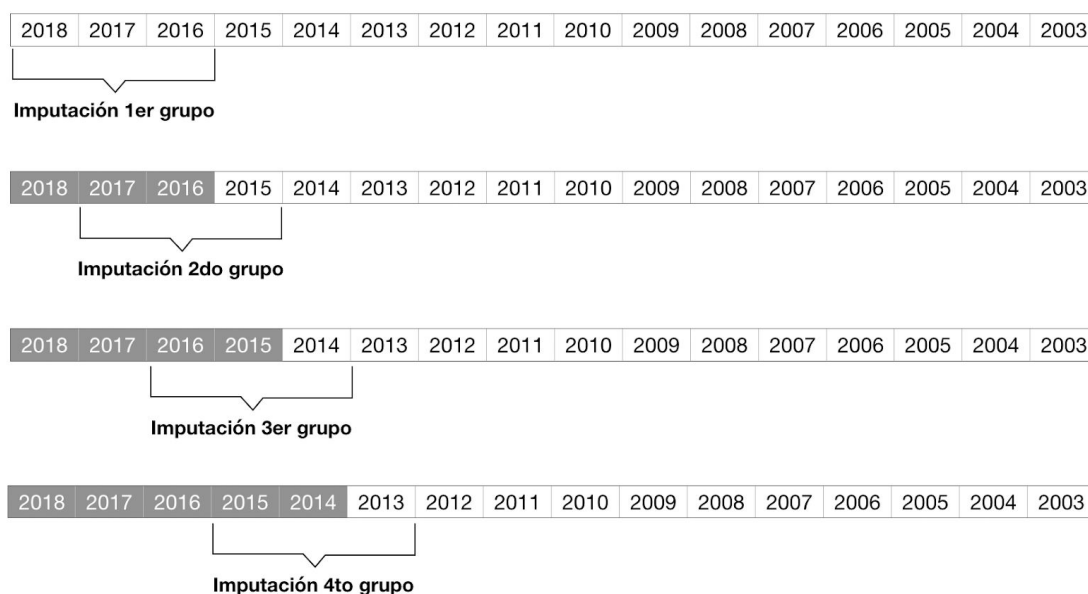
|       | CILINDRADA | POTENCIA_FISCAL | MASA_MAX    | MOM         | TARA       |
|-------|------------|-----------------|-------------|-------------|------------|
| X2018 | 237.693673 | 1.529061        | 191.588341  | 132.371012  | 141.388369 |
| X2017 | 237.693673 | 1.529061        | 191.588341  | 132.371012  | 141.388369 |
| X2016 | 248.951415 | 1.435457        | 1535.208666 | 2522.722591 | 156.985426 |
| X2015 | 248.951415 | 1.435457        | 1535.208666 | 2522.722591 | 156.985426 |
| X2014 | 246.868774 | 1.466107        | 203.175403  | 1252.251967 | 142.338074 |
| X2013 | 246.868774 | 1.466107        | 203.175403  | 1252.251967 | 142.338074 |
| X2012 | 259.591653 | 1.242335        | 226.308268  | 180.897301  | 155.251251 |
| X2011 | 259.591653 | 1.242335        | 226.308268  | 180.897301  | 155.251251 |
| X2010 | 281.360813 | 1.333931        | 214.301812  | 160.691227  | 147.252787 |
| X2009 | 281.360813 | 1.333931        | 214.301812  | 160.691227  | 147.252787 |
| X2008 | 292.432546 | 1.485622        | 210.942522  | 1331.917933 | 157.352037 |
| X2007 | 292.432546 | 1.485622        | 210.942522  | 1331.917933 | 157.352037 |
| X2006 | 263.973642 | 1.574808        | 327.224604  | 1681.049958 | 132.084354 |
| X2005 | 263.973642 | 1.574808        | 327.224604  | 1681.049958 | 132.084354 |
| X2004 | 267.423952 | 1.596777        | 188.946745  | 418.266444  | 146.228917 |
| X2003 | 267.423952 | 1.596777        | 188.946745  | 418.266444  | 146.228917 |
| X2002 | 260.492875 | 1.479194        | 508.596800  | 247.343396  | 143.159203 |
| X2001 | 260.492875 | 1.479194        | 508.596800  | 247.343396  | 143.159203 |
| X2000 | 264.667390 | 1.459147        | 203.663369  | 268.902667  | 142.343726 |
| X1999 | 264.667390 | 1.459147        | 203.663369  | 268.902667  | 142.343726 |
| X1998 | 314.908781 | 1.703090        | 2902.165715 | 280.444241  | 151.209933 |
| X1997 | 314.908781 | 1.703090        | 2902.165715 | 280.444241  | 151.209933 |
| X1996 | 354.264626 | 2.058439        | 291.543824  | 422.589990  | 169.011991 |
| X1995 | 354.264626 | 2.058439        | 291.543824  | 422.589990  | 169.011991 |
| X1994 | 0.000000   | 0.000000        | 228.528279  | 501.631572  | 150.214829 |
| X1993 | 0.000000   | 0.000000        | 228.528279  | 501.631572  | 150.214829 |
| X1992 | 389.501852 | 0.000000        | 210.217433  | 209.580075  | 170.544742 |
| X1991 | 389.501852 | 0.000000        | 210.217433  | 209.580075  | 170.544742 |
| X1990 | 0.000000   | 0.000000        | 231.015788  | 216.889761  | 167.198288 |

A la taula anterior s'observen errors bastant similars als del mètode anterior tenim valors repetits ja que s'han agrupat per grups de 2 anys. L'avantatge d'aquest mètode és que la variància d'aquestes imputacions és menor, i l'escala global que es pot observar a la dreta del gràfic és menor que en el mètode anterior. El desavantatge és que en utilitzar aquest mètode podem afegir soroll extra en els anys on hi ha pocs valors mancants. Per exemple, si comparem amb el gràfic anterior, la variable MOM l'any 2015 presenta un error d'imputació bastant alt, i això repercuteix l'error estimat en aquest model per a les observacions de l'any 2016 també.

### 3.4. Imputació per finestra lliscant

El tercer mètode proposat és el d'anar omplint les dades més noves primer i després anar afegint les observacions d'anys més anteriors, amb una finestra temporal lliscant (veure esquema més avall). L'objectiu d'això és que els valors que falten a emplenar siguin els més similars als valors observats a la base de dades. En el cas ideal, extraient les observacions amb valors buits la distribució dels altres valors hauria de ser la mateixa que quan s'inclouen les variables amb valors a emplenar.

Aquesta tècnica la fem servir a causa de la gran dependència que tenen els valors respecte a l'any que van ser introduïts a la base de dades. Llavors, la base de dades s'omple a partir del present, iterant any rere any cap enrere.



*Esquema d'imputació amb finestra lliscant*

|       | CILINDRADA | POTENCIA_FISCAL | MASA_MAX    | MOM         | TARA       |
|-------|------------|-----------------|-------------|-------------|------------|
| X2018 | 246.901302 | 1.552341        | 1139.074999 | 313.781534  | 147.348845 |
| X2017 | 246.901302 | 1.552341        | 1139.074999 | 313.781534  | 147.348845 |
| X2016 | 246.901302 | 1.552341        | 1139.074999 | 313.781534  | 147.348845 |
| X2015 | 263.394553 | 1.469371        | 196.500505  | 3611.070967 | 151.148738 |
| X2014 | 250.947947 | 1.454950        | 199.533465  | 1566.609734 | 144.873115 |
| X2013 | 248.981498 | 1.429456        | 215.928716  | 713.789929  | 142.270559 |
| X2012 | 248.031055 | 1.465921        | 227.560577  | 195.371640  | 154.247242 |
| X2011 | 259.825790 | 1.249677        | 225.145183  | 166.876170  | 149.763134 |
| X2010 | 275.381075 | 1.260078        | 221.736895  | 153.667586  | 151.931942 |
| X2009 | 276.911426 | 1.369976        | 207.059469  | 170.296168  | 143.736194 |
| X2008 | 312.474997 | 1.366856        | 206.119515  | 169.654136  | 156.637601 |
| X2007 | 275.425653 | 1.476091        | 199.120142  | 1733.997912 | 152.571732 |
| X2006 | 282.489829 | 1.482449        | 199.929887  | 1764.779491 | 123.792363 |
| X2005 | 281.970304 | 1.627215        | 411.429702  | 1061.318464 | 140.736814 |
| X2004 | 268.093818 | 1.661658        | 184.092173  | 245.420780  | 145.136796 |
| X2003 | 272.071806 | 1.575900        | 185.768222  | 557.321541  | 145.227536 |
| X2002 | 265.626610 | 1.542617        | 689.418272  | 219.066850  | 141.417549 |
| X2001 | 255.102010 | 1.460773        | 203.504634  | 278.567802  | 141.066577 |
| X2000 | 265.178395 | 0.000000        | 195.694114  | 287.160265  | 130.445243 |
| X1999 | 0.000000   | 1.536405        | 215.603369  | 246.250923  | 145.366049 |
| X1998 | 0.000000   | 0.000000        | 3784.095835 | 268.327716  | 150.096113 |
| X1997 | 319.346763 | 1.767376        | 207.663983  | 291.544966  | 148.458148 |
| X1996 | 336.914434 | 1.936018        | 217.769563  | 463.542842  | 159.350322 |
| X1995 | 0.000000   | 2.124016        | 366.115917  | 356.673155  | 168.432679 |
| X1994 | 0.000000   | 0.000000        | 250.669677  | 601.698434  | 145.705530 |
| X1993 | 0.000000   | 0.000000        | 217.610283  | 205.619117  | 162.649525 |
| X1992 | 373.794530 | 0.000000        | 207.752356  | 185.123609  | 158.206768 |
| X1991 | 0.000000   | 0.000000        | 219.271321  | 209.996239  | 197.594058 |
| X1990 | 0.000000   | 0.000000        | 228.069758  | 228.303874  | 176.999275 |

La taula anterior mostra els errors d'aquest mètode. Tret d'algunes variables i anys puntuals sembla obtenir uns resultats vàlids per a la majoria d'anys i variables.

### 3.5. Imputació per finestra acumulativa

Finalment, l'últim mètode proposat és molt similar a l'anterior. S'emplenen les dades més noves primer i després es van afegint les observacions d'anys més anteriors, per posteriorment anar omplint-les iterativament, de forma similar al mètode anterior (veure esquema més avall), amb la diferència que s'utilitzen tots els anys més nous per tal d'emplenar les observacions més velles. L'objectiu d'això és que els valors que falten a emplenar siguin el més similars als valors observats a la base de dades. Com els valors més nous tenen un percentatge molt menor de valors buits, aquest mètode sembla adequat. La diferència amb el mètode anterior, és que per aquest mètode la imputació dels valors d'anys més anteriors utilitza les dades de tots els anys posteriors ja imputats prèviament. És a dir, la imputació de l'any 2010 utilitza les observacions des del 2011 fins 2018 ja imputades sense valors buits.

De manera similar al mètode anterior, aquesta tècnica la fem servir a causa de la gran dependència que tenen els valors respecte a l'any que van ser introduïts a la base de dades. A la següent taula podem observar els errors respectius a les diferent variables i anys. El principal problema d'aquest mètode és que quan hi ha una variable que té un any amb molts valors que falten, tots els anys anteriors es veuen afectats a l'hora de fer la imputació.

|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 2018 | 2017 | 2016 | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|



**Imputación 1er grupo**

|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 2018 | 2017 | 2016 | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|



**Imputación 2do grupo**

|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 2018 | 2017 | 2016 | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|



**Imputación 3er grupo**

|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 2018 | 2017 | 2016 | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|



**Imputación 4to grupo**

|       | CILINDRADA | POTENCIA_FISCAL | MASA_MAX    | MOM         | TARA       |
|-------|------------|-----------------|-------------|-------------|------------|
| X2018 | 246.249283 | 1.583114        | 190.451051  | 133.319019  | 154.149953 |
| X2017 | 242.300866 | 1.570456        | 196.309872  | 137.789535  | 133.799292 |
| X2016 | 243.531203 | 1.543226        | 1140.460565 | 312.528846  | 134.083995 |
| X2015 | 237.345098 | 1.502417        | 1021.402959 | 1662.709036 | 141.415482 |
| X2014 | 253.822766 | 1.513180        | 948.616936  | 1649.640809 | 139.741578 |
| X2013 | 250.738483 | 1.458436        | 898.558507  | 1576.732090 | 143.236426 |
| X2012 | 254.393506 | 1.530022        | 860.324243  | 1506.216358 | 139.218385 |
| X2011 | 251.053446 | 1.538196        | 820.512351  | 1434.777022 | 140.553590 |
| X2010 | 259.986081 | 1.512451        | 783.903229  | 1360.829386 | 143.174695 |
| X2009 | 262.594213 | 1.532878        | 752.399474  | 1301.123780 | 142.068946 |
| X2008 | 262.208388 | 1.556128        | 721.551511  | 1243.039404 | 146.920316 |
| X2007 | 277.388865 | 1.653142        | 682.640594  | 1307.201761 | 144.265366 |
| X2006 | 283.850987 | 1.692548        | 650.047395  | 1362.046779 | 149.970460 |
| X2005 | 287.476531 | 1.644371        | 634.500580  | 1357.095025 | 152.368832 |
| X2004 | 293.152878 | 1.706273        | 612.937136  | 1302.298068 | 153.098130 |
| X2003 | 297.165904 | 1.748623        | 595.072725  | 1250.360927 | 154.797762 |
| X2002 | 306.202669 | 1.649795        | 602.594702  | 1210.150564 | 155.168542 |
| X2001 | 304.034260 | 1.617222        | 588.686172  | 1176.969145 | 154.764871 |
| X2000 | 309.510167 | 0.000000        | 577.255219  | 1146.501138 | 154.464464 |
| X1999 | 0.000000   | 1.622354        | 570.615309  | 1121.349361 | 156.359475 |
| X1998 | 0.000000   | 0.000000        | 779.170200  | 1100.110699 | 154.265915 |
| X1997 | 310.961718 | 1.692700        | 772.200928  | 1084.957124 | 158.775835 |
| X1996 | 310.329110 | 1.688771        | 765.677419  | 1074.992279 | 157.570195 |
| X1995 | 0.000000   | 1.680108        | 762.215889  | 1068.232368 | 163.163946 |
| X1994 | 0.000000   | 0.000000        | 757.957864  | 1062.749545 | 159.535482 |
| X1993 | 0.000000   | 0.000000        | 756.719319  | 1057.766228 | 159.845648 |
| X1992 | 318.964146 | 0.000000        | 756.130734  | 1054.348145 | 161.792410 |
| X1991 | 0.000000   | 0.000000        | 753.777807  | 1051.426884 | 160.661701 |
| X1990 | 0.000000   | 0.000000        | 752.925030  | 1049.250617 | 159.281055 |

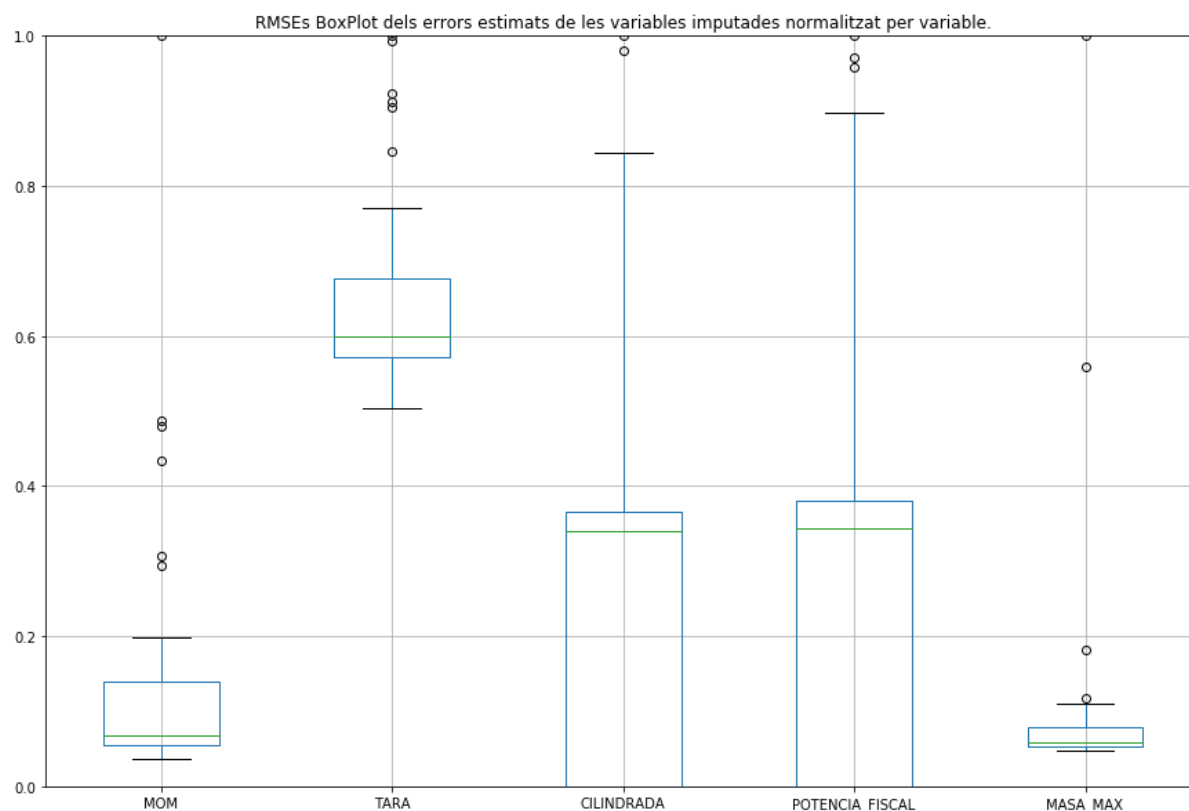
En la taula anterior observem els errors d'aquest mètode. Observem que el problema d'aquest mètode és que quan existeix un any amb una variable que té molts valors

mancants, aquesta pot tenir un error gran i aquest error s'arrossega cap als anys més anteriors per a la mateixa variable. Per tant, aquest mètode sembla afegir biaix innecessari a l'hora d'estimar els valors mancants de certes variables.

|                        | Imputation method | RMSE        |
|------------------------|-------------------|-------------|
| <b>CILINDRADA</b>      | Year              | 212         |
|                        | 2 Year groups     | 261         |
|                        | Rolling cumsum    | <b>209</b>  |
|                        | Rolling fixed     | 212         |
| <b>POTÈNCIA FISCAL</b> | Year              | 1.27        |
|                        | 2 Year groups     | 1.57        |
|                        | Rolling cumsum    | 1.22        |
|                        | Rolling fixed     | <b>1.14</b> |
| <b>MASSA MAX</b>       | Year              | <b>395</b>  |
|                        | 2 Year groups     | 465         |
|                        | Rolling cumsum    | 714         |
|                        | Rolling fixed     | 408         |
| <b>MOM</b>             | Year              | <b>512</b>  |
|                        | 2 Year groups     | 596         |
|                        | Rolling cumsum    | 1143        |
|                        | Rolling fixed     | 602         |
| <b>TARA</b>            | Year              | 166         |
|                        | 2 Year groups     | 165         |
|                        | Rolling cumsum    | <b>151</b>  |
|                        | Rolling fixed     | 151         |

La taula anterior mostra la mitjana dels errors imputats per mètode i per variable de tots els anys. S'observa que cap dels mètodes és superior a la resta de mètodes per a totes les variables. Cada variable funciona millor amb diferents mètodes. Això pot ésser per culpa de la particularitat dels valors buits de cada variable.

A continuació mostrem un gràfic final dels errors pel mètode d'imputació anual, on es poden veure els errors normalitzats per a cada variable. Aquest gràfic es correspon amb els errors de la taula de la pàgina 38.



Observem que hi ha variables on els errors tenen una major variància i d'altres on els errors estan més concentrats en certes regions.

A causa de què no hi ha cap mètode superior a la resta, hem decidit escollir el primer mètode per raó de la simplicitat d'aquest. Donat que els altres mètodes no suposen una gran millora de les imputacions però sí un increment de la dificultat, creiem adient quedar-nos amb el primer mètode d'agrupar per anys únics.

## 4. Models

En aquesta secció es discuteixen els models de predicció d'emissions de CO<sub>2</sub> considerats en aquest informe. Per a totes aquestes consideracions, en general:

- Es va dividir el conjunt de dades aleatòriament en un 80% per a l'entrenament dels models i un 20% com a validació. Es va prendre una llavor d'aleatorietat fixa per poder comparar justament el rendiment dels diferents models.
- Per a tots els models es va fer un escalat de variables normal, restant la seva mitjana i dividint per la seva desviació estàndard.
- Es van separar les dades per combustible, però de manera simplificada: només es va agrupar per dièsel, gasolina, vehicles híbrids HEV, altres híbrids i elèctrics (per als quals no es va fer cap estimació), i altres (butà, biodièsel, etc.).
- Es va avaluar el rendiment dels diferents criteris d'outliers.

Cal recordar que el dataset d'entrenament només conté aquelles observacions que tenen valors de CO<sub>2</sub>, i que aquestes només són presents per a anys posteriors a 2002 (i majoritàriament per a anys posteriors a 2008). Això introdueix un biaix potencial, com es discuteix més amunt, però que probablement sigui contrarestat per la inclusió de l'antiguitat dels cotxes.

### 4.1. Mètriques de rendiment

A l'hora de fer models, es van utilitzar tècniques de mínims quadrats parcials i regressions regularitzades amb norma L1 (Lasso) i L2 (Ridge) per a identificar aquelles variables que tenen la major rellevància per a predir el CO<sub>2</sub>.

Per a mesurar el rendiment dels models s'utilitzen les següents mètriques:

Coeficient  $R^2$ :

$$R^2 \equiv 1 - \frac{SSR}{SST}; \quad SST = \sum_i (i_i - \bar{i})^2; \quad SSR = \sum_i (i_i - f(i))^2 = \sum_i i_i^2 - 2 \sum_i i_i f(i) + \sum_i f(i)^2; \quad \bar{i} = \frac{1}{n} \sum_{i=1}^n i_i$$

Arrel quadrada de la mitjana dels errors quadrats (*Root-mean-square error* (RMSE)):

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (i_j - \hat{i}_j)^2}$$

Error mitjà absolut (Mean absolute error (MAE)):

$$MAE = \frac{1}{N} \sum_{t=1}^N abs(i_t - \hat{i}_t)$$

Error mitjà percentual (Mean average percentage error (MAPE)):

$$MAPE = \frac{1}{N} \sum_{t=1}^N \frac{abs(i_t - \hat{i}_t)}{i_t} * 100$$

#### 4.1.1. Model directe

En una primera instància, es va treballar amb les dades que estaven "completes", en el sentit que tenien informació en totes les seves columnes (més enllà de filtrar aquells valors que falten o iguals a zero, no es va efectuar cap revisió de si aquesta informació era correcta).

Es va comparar la possibilitat d'arribar a un model únic per a tots els tipus de combustibles, però es va descartar aquesta alternativa perquè atorgava una precisió un 10% inferior als diferents models per a cada tipus de combustible. Amb aquestes dades, es van aconseguir les següents mètriques:

| Combustible     | RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE   |
|-----------------|------|----------------------|---------------------|--------|
| <b>Dièsel</b>   | 17.2 | 0.637                | 0.613               | 10.94% |
| <b>Gasolina</b> | 14.4 | 0.683                | 0.677               | 9.42%  |
| <b>HEV</b>      | 36.2 | 0.357                | 0.13                | 5.486% |
| <b>Altres</b>   | 3.6  | 0.616                | 0.879               | 8.955% |

#### 4.1.2. Model amb filtrat d'outliers

Arreglant la base de dades i traient valors que cauen fora dels rangs determinats manualment (veure a dalt neteja d'outliers) es pot millorar sensiblement la predicció. Filtrant d'aquesta manera, queden únicament 994.018 observacions, dividides per combustible, de la següent forma:

DIÈSEL: 595.798 vehicles  
 GASOLINA: 366.131 vehicles  
 HEV: 26.037 vehicles  
 ALTRES: 5.485 vehicles

I s'aconsegueixen les següents mètriques, mostrant millores substancials en diverses mètriques:

| Combustible | RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE   |
|-------------|------|----------------------|---------------------|--------|
| Dièsel      | 14.2 | 0.765                | 0.767               | 8.171% |
| Gasolina    | 11.9 | 0.768                | 0.764               | 6.552% |
| HEV         | 6.8  | 0.803                | 0.810               | 4.994% |
| altres      | 15.9 | 0.627                | 0.558               | 9.749% |

Aquest exercici ens indica que hem d'intentar millorar la identificació de potencials errors o altres outliers en les mostres, i després imputar els valors correctes mitjançant les tècniques explicades anteriorment.

## 4.2. Extracció de colinealitat

Com vam veure a la secció de correlacions, hi ha diverses variables que (encara que tinguin soroll) estan fortament relacionades amb altres. Per millorar el rendiment dels ajustos, es va fer un extracció de la multicol·linealitat entre variables utilitzant la tècnica d'anàlisi de components principals.

Aquesta tècnica es pot entendre com una rotació de la matriu de covariància entre variables a un espai ortogonal, on les noves variables són totes ortonormals entre elles. Es realitza l'ajust a un model multilineal utilitzant aquestes noves variables, i després es pot aplicar la rotació inversa per poder expressar el model ajustat en termes de les variables originals. En termes matemàtics, escrivim totes les nostres  $N$  observacions com una matriu

$$\vec{X} = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \dots \\ \vec{x}_N \end{pmatrix}$$

On  $\vec{x}_n = (x_1, x_2, \dots, x_M)$  és el vehicle  $n$  amb valors  $x_1, \dots, x_M$  a les  $M$  propietats disponibles. Sempre podem escriure aquesta matriu com  $\vec{Z} = \vec{X}V$ , on  $V$  és una matriu de mida  $M \times M$  ortonormal,  $VV^T = I$  que quantifica la covariància entre les columnes

de  $X$ . Per això, en realitat, en lloc d'utilitzar directament les observacions, fem servir la seva versió normalitzada i amb les mitjanes restades:

$$\tilde{x}_j = \frac{x_j - \langle x_j \rangle}{\sigma_j}$$

Les columnes i files de  $V$  són els vectors base dels espais, i es pot entendre com una rotació.

$$z_i = \sum_j \tilde{x}_j V_{ji}$$

Com en la base rotada les dimensions són ortogonals entre elles (sense col·linealitat), fem la regressió multilinear per tal de trobar els coeficients:

$$y = \sum_i a_i z_i + C$$

Per a poder escriure la regressió en termes dels coeficients originals fem servir les equacions de més amunt per a reemplaçar:

$$\begin{aligned} y &= \sum_i a_i \left( \sum_j \tilde{x}_j V_{ji} \right) + C \\ &= \sum_{i,j} a_i V_{ji} \left( \frac{x_j - \langle x_j \rangle}{\sigma_j} \right) + C \\ &= \sum_j \frac{(\sum_i a_i V_{ji})}{\sigma_j} x_j - \sum_j \frac{(\sum_i a_i V_{ji}) \langle x_j \rangle}{\sigma_j} + C \\ &= \sum_j \tilde{a}_j x_j + \tilde{C} \end{aligned}$$

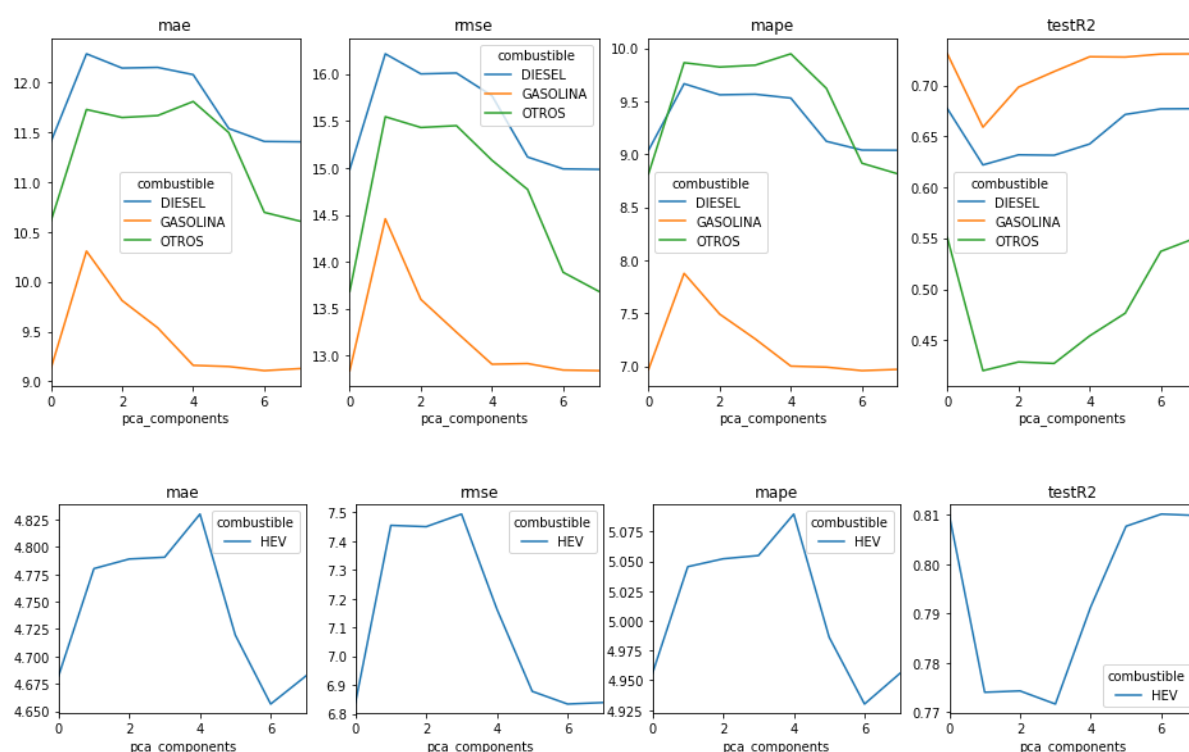
On

$$\begin{aligned} \tilde{a}_j &= \frac{(\sum_i a_i V_{ji})}{\sigma_j} \\ \tilde{C} &= C - \sum_j \frac{(\sum_i a_i V_{ji}) \langle x_j \rangle}{\sigma_j} \end{aligned}$$

En fer la rotació  $V$  és possible treure aquells components que aporten menys per descriure la variància de les dades, simplement ordenant-los pels seus autovalors (que són sempre positius), i traient els més petits. D'aquesta manera, extraïem els components que no aporten més que soroll, i reduïm la complexitat del model.

En les gràfiques següents, veiem les mètriques de cada ajust per als diferents combustibles (separant els HEV per permetre veure millor la diferència) en funció de la quantitat de components que conservem - definint zero components com *no realitzar la transformació de PCA*. Llavors, veiem que si només conservem un o dos components, les mètriques són dolentes, però milloren fins que bàsicament se saturen al voltant de quatre o cinc components, depenent del combustible.

D'aquestes gràfiques podem concloure que si bé hi ha una presència de col·linealitat en les dades (perquè és possible extreure components sense massa perjudici per a les mètriques), tampoc és extrema, ja que no es poden treure massa components exceptuant per a la gasolina.



### 4.3. Regressió lineal regularitzada

Pels requeriments del projecte, s'utilitzen models de regressió multilinear com a models predictius. La regressió lineal és una funció polinòmica de primer grau de les variables explicatives, minimitzant la suma dels quadrats de les diferències entre les variables dependents observades en el dataset i les prediccions de la funció lineal.

$$i_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \dots + \beta_p * x_{ip} + \varepsilon_i$$

On es busca minimitzar

$$\beta_{min}, \beta_0 \left\{ \frac{1}{N} \sum_{i=1}^N (i_i - \beta_0 - x_i \beta T)^2 \right\}$$

Aquests models són senzills de comprendre, però depenen dels nostres criteris per a saber quines variables convé incloure i quines no. Per això, podem utilitzar models de regressió lineal regularitzada, en els quals incloem en la funció a minimitzar un terme que penalitzi la quantitat de variables on els coeficients siguin diferents de zero.

El model LASSO és una d'aquestes variants. El paràmetre de regularització limita la quantitat i la mida dels regressors a fer servir per a aquesta regressió i així evitar construir equacions massa complicades que no extrapolen bé a casos no observats.

$$\beta_{min}, \beta_0 \left\{ \frac{1}{N} \sum_{i=1}^N (i_i - \beta_0 - x_i \beta T)^2 \right\} \text{ st } \sum_{j=1}^p |\beta_j| \leq t$$

En aquest projecte hem aplicat models LASSO, però no vam detectar cap component a descartar més enllà dels seleccionats a partir de la descomposició en components principals. Es va comprovar a més que el rendiment és bàsicament idèntic amb altres models de regularització, com Ridge o ElasticNet.

### 4.4. Enginyeria de variables

En l'aprenentatge automàtic és molt comú trobar que es pot guanyar precisió construint variables noves (normalment no lineals) en funció de les originals. Aquestes variables solen dissenyar-se a partir del coneixement específic sobre el domini amb el qual es treballa. Per exemple, un dels problemes d'aquest projecte, és que els vehicles de més antiguitat tenen més emissions. Per això, cal construir una variable que capturi l'antiguitat dels vehicles, mesurada com la quantitat de dies des de la data de primera matriculació fins al final de l'any de l'exercici que es vol avaluar. Per exemple, com les dades avaluades són fins a

2018 (l'última matriculació observada és el 28/12/18), la variable antiguitat per a un vehicle matriculat el 2018/12/20 serien 11 dies, o per al 2005/07/18 són 4.914 dies. Per conveniència, normalitzem aquesta variable per la quantitat mitjana de dies d'un any, 365.25 (que té en compte els anys de traspàs). Els valors dels exemples serien llavors  $11 / 365.25 = 0.03011636$ , i  $4914 / 365.25 = 13.453799$ . Aquesta normalització no afecta al model, i només serveix per identificar els valors dels coeficients resultants en una escala natural. Es va tenir en compte també la quantitat d'anys sencers, però no va donar el mateix rendiment.

Més en general, es poden provar automàticament combinacions de variables. Com vam realitzar després una descomposició en PCA, no vam provar combinacions lineals, i només vam explorar el guany de rendiment dels algorismes amb combinacions no lineals. En concret, es van explorar totes les combinacions de segon grau, és a dir, totes les multiplicacions entre variables possibles.

De les variables numèriques que disposem per a cada vehicle, seleccionem 8 d'elles: CILINDRADA, POTENCIA\_FISCAL, POTENCIA\_NETA, MASA\_MAX, MOM, TARA, NUM\_PLAZAS, antiguitat, i al combinar-les obtenim **44** ( $= 9 * 8/2 + 8$ ) variables del tipus TARA \* CILINDRADA, ANTIGUITAT \* POTÈNCIA FISCAL, etc.

El model lineal utilitzant *totes* aquestes variables aconsegueix les mètriques següents:

| Combustible     | RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE |
|-----------------|------|----------------------|---------------------|------|
| <b>Dièsel</b>   | 12.7 | 0.814                | 0.815               | 7.3% |
| <b>Gasolina</b> | 12.3 | 0.79                 | 0.75                | 6.1% |
| <b>HEV</b>      | 53.7 | 0.89                 | -10.7               | 4.5% |
| <b>Altres</b>   | 11.8 | 0.80                 | 0.76                | 6.9% |

Notem que alguns combustibles ara mostren una diferència entre el coeficient de determinació d'entrenament i de prova, indicant que tenim algun grau de sobre-determinació (overfit). En particular, els vehicles HEV no donen un bon coeficient. El guany pel que fa als models sense totes aquestes variables dissenyades dóna un guany aproximat d'un 5% extra al R<sup>2</sup> de prova. Si bé matemàticament aquest guany podria estar justificat, el model resultant i els seus coeficients perden la major part de la seva interpretabilitat, i és important llavors sopesar els avantatges i desavantatges d'utilitzar aquestes variables extres.

## 4.5. Significances

Els resultats d'una regressió lineal es poden interpretar amb més detall veient el nivell de significança estadística trobada per a la regressió, i per a cada un dels coeficients per separat. En particular, cada coeficient de la regressió té associat un error estàndard, i es pot calcular un valor-p per a cada un d'ells. Amb els valors-p podem establir algun valor límit sota del qual considerem què canvis en aquesta variable es tradueixen en canvis significatius en la variable objectiu (usualment s'utilitza el límit 0.05).

En els exemples calculats més amunt, traient les variables MOM i NUM\_PLAZAS, obtenim que totes les variables tenen significança alta per a tots els combustibles. És important notar que per aquest anàlisi no estem fent servir la transformació a l'espai ortogonal de PCA, la qual cosa proporcionaria una possible millora en els resultats però confondria la manera d'estudiar les significances de les variables originals.

A continuació detallem els resultats de les significances i nivells d'error per a cada coeficient de cada combustible:

**DIESEL**

R-squared: 0.753

Adj. R-squared: 0.753

| =====           |          |          |          |       |         |
|-----------------|----------|----------|----------|-------|---------|
| ===             |          |          |          |       |         |
|                 | coef     | std err  | t        | P> t  | [0.025  |
| 0.975]          |          |          |          |       |         |
| -----           |          |          |          |       |         |
| ---             |          |          |          |       |         |
| const           | -22.8785 | 0.210    | -109.094 | 0.000 | -23.290 |
| -22.467         |          |          |          |       |         |
| CILINDRADA      | 0.0164   | 0.000    | 73.007   | 0.000 | 0.016   |
| 0.017           |          |          |          |       |         |
| POTENCIA_FISCAL | 0.6686   | 0.043    | 15.407   | 0.000 | 0.584   |
| 0.754           |          |          |          |       |         |
| POTENCIA_NETA   | -0.1633  | 0.001    | -124.619 | 0.000 | -0.166  |
| -0.161          |          |          |          |       |         |
| MASA_MAX        | 0.0004   | 1.62e-05 | 27.005   | 0.000 | 0.000   |
| 0.000           |          |          |          |       |         |
| TARA            | 0.0786   | 0.000    | 649.828  | 0.000 | 0.078   |
| 0.079           |          |          |          |       |         |
| antiguity_years | 2.8631   | 0.008    | 347.724  | 0.000 | 2.847   |
| 2.879           |          |          |          |       |         |

**GASOLINA**

R-squared: 0.764

Adj. R-squared: 0.764

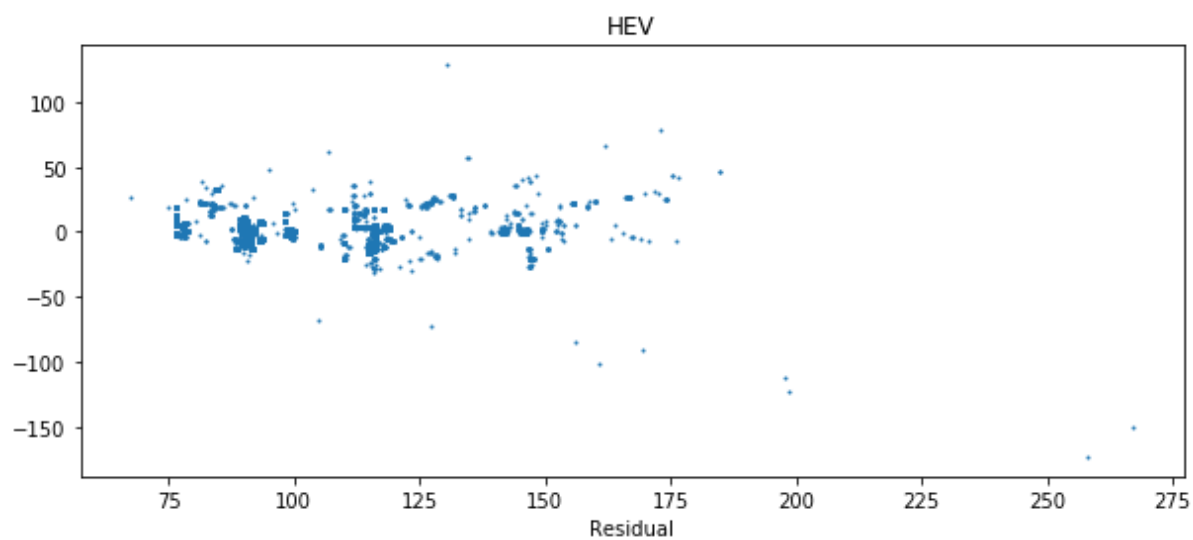
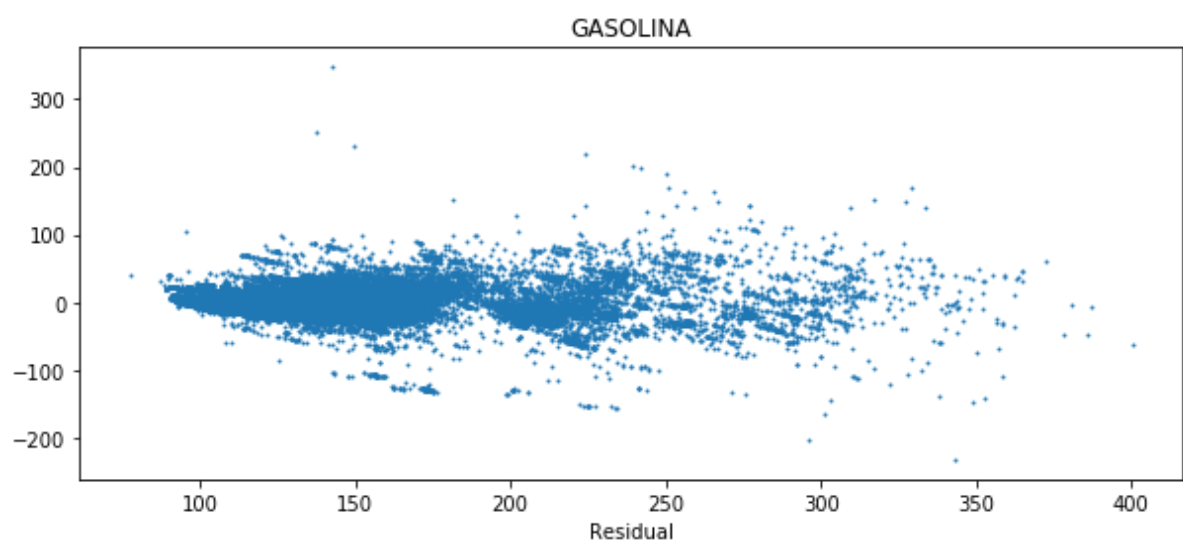
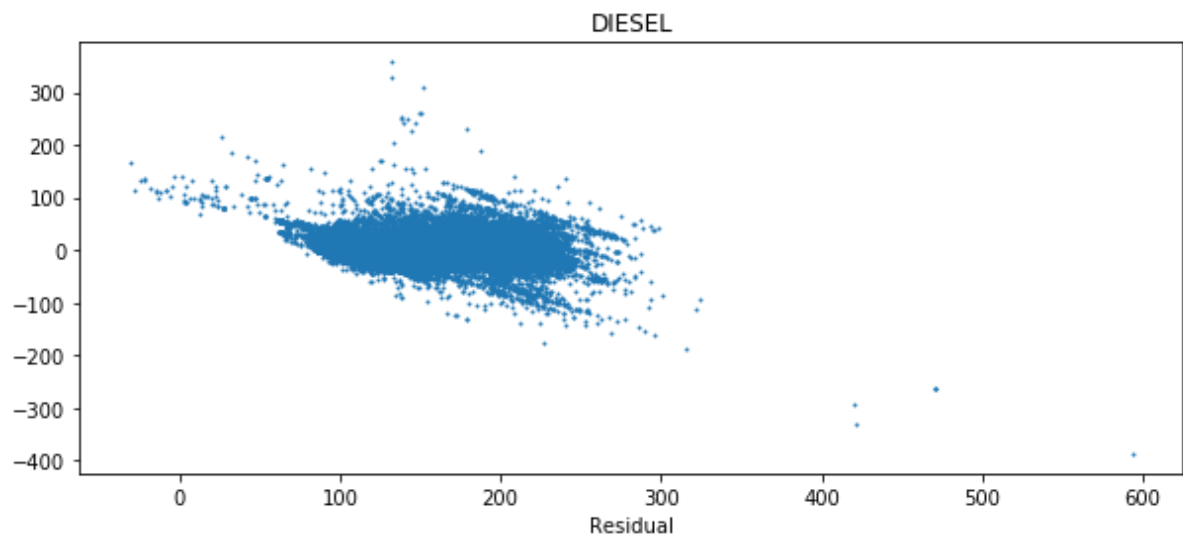
| =====           |         |         |         |       |        |
|-----------------|---------|---------|---------|-------|--------|
| ===             |         |         |         |       |        |
|                 | coef    | std err | t       | P> t  | [0.025 |
| 0.975]          |         |         |         |       |        |
| -----           |         |         |         |       |        |
| ---             |         |         |         |       |        |
| const           | 17.2020 | 0.208   | 82.853  | 0.000 | 16.795 |
| 17.609          |         |         |         |       |        |
| CILINDRADA      | 0.0074  | 0.000   | 31.909  | 0.000 | 0.007  |
| 0.008           |         |         |         |       |        |
| POTENCIA_FISCAL | 4.1785  | 0.039   | 106.029 | 0.000 | 4.101  |
| 4.256           |         |         |         |       |        |
| POTENCIA_NETA   | 0.0193  | 0.001   | 18.899  | 0.000 | 0.017  |
| 0.021           |         |         |         |       |        |
| MASA_MAX        | 0.0067  | 0.000   | 25.415  | 0.000 | 0.006  |
| 0.007           |         |         |         |       |        |
| TARA            | 0.0334  | 0.000   | 100.944 | 0.000 | 0.033  |
| 0.034           |         |         |         |       |        |
| antiguity_years | 1.9912  | 0.009   | 225.091 | 0.000 | 1.974  |
| 2.009           |         |         |         |       |        |

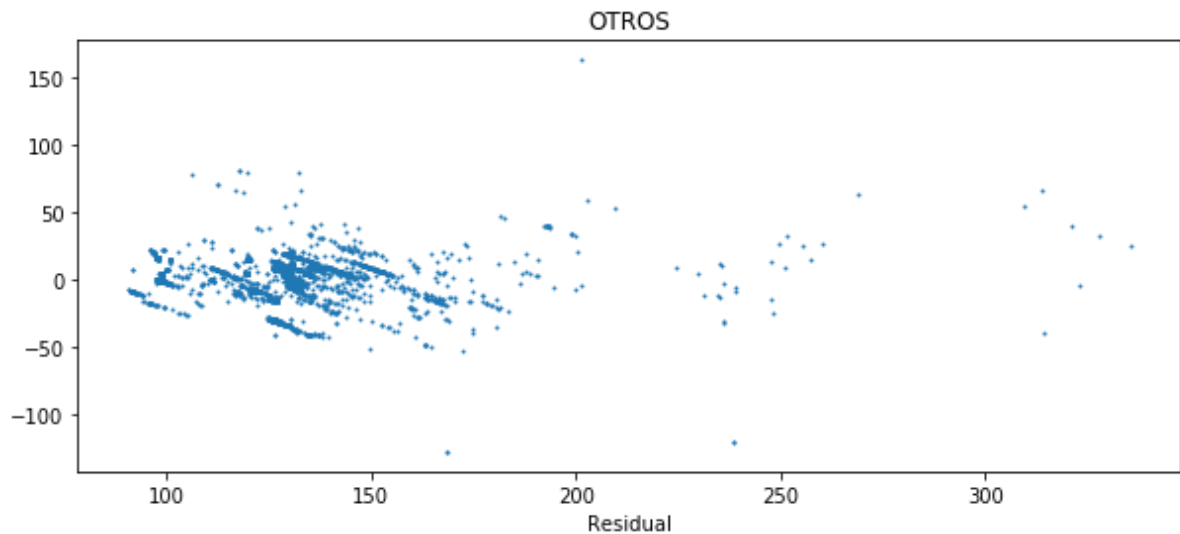
|                 |         |         |        |       |        |
|-----------------|---------|---------|--------|-------|--------|
| HEV             |         |         |        |       |        |
| R-squared:      | 0.824   |         |        |       |        |
| Adj. R-squared: | 0.824   |         |        |       |        |
| =====           |         |         |        |       |        |
| ===             |         |         |        |       |        |
|                 | coef    | std err | t      | P> t  | [0.025 |
| 0.975]          |         |         |        |       |        |
| -----           |         |         |        |       |        |
| ---             |         |         |        |       |        |
| const           | 16.1500 | 0.689   | 23.455 | 0.000 | 14.800 |
| 17.500          |         |         |        |       |        |
| CILINDRADA      | 0.0037  | 0.001   | 6.810  | 0.000 | 0.003  |
| 0.005           |         |         |        |       |        |
| POTENCIA_FISCAL | -0.4927 | 0.112   | -4.419 | 0.000 | -0.711 |
| -0.274          |         |         |        |       |        |
| POTENCIA_NETA   | 0.2418  | 0.003   | 77.615 | 0.000 | 0.236  |
| 0.248           |         |         |        |       |        |
| MASA_MAX        | 0.0302  | 0.000   | 87.485 | 0.000 | 0.029  |
| 0.031           |         |         |        |       |        |
| antiguity_years | 0.3044  | 0.017   | 17.476 | 0.000 | 0.270  |
| 0.339           |         |         |        |       |        |

|                 |         |         |        |       |        |
|-----------------|---------|---------|--------|-------|--------|
| OTROS           |         |         |        |       |        |
| R-squared:      | 0.603   |         |        |       |        |
| Adj. R-squared: | 0.603   |         |        |       |        |
| =====           |         |         |        |       |        |
| ===             |         |         |        |       |        |
|                 | coef    | std err | t      | P> t  | [0.025 |
| 0.975]          |         |         |        |       |        |
| -----           |         |         |        |       |        |
| ---             |         |         |        |       |        |
| const           | 11.4731 | 2.871   | 3.997  | 0.000 | 5.845  |
| 17.101          |         |         |        |       |        |
| CILINDRADA      | 0.0229  | 0.004   | 5.096  | 0.000 | 0.014  |
| 0.032           |         |         |        |       |        |
| POTENCIA_FISCAL | 2.2217  | 0.784   | 2.835  | 0.005 | 0.685  |
| 3.758           |         |         |        |       |        |
| POTENCIA_NETA   | 0.0775  | 0.018   | 4.302  | 0.000 | 0.042  |
| 0.113           |         |         |        |       |        |
| TARA            | 0.0403  | 0.002   | 21.280 | 0.000 | 0.037  |
| 0.044           |         |         |        |       |        |
| antiguity_years | 2.1215  | 0.119   | 17.801 | 0.000 | 1.888  |
| 2.355           |         |         |        |       |        |

Per a una estimació global del model hem d'inspeccionar els residus en cada punt, i verificar que no hi ha informació no capturada mitjançant patrons obvis - és a dir, que els residus estan dispersos al voltant del zero de manera homogènia per a tot el rang de predicció.

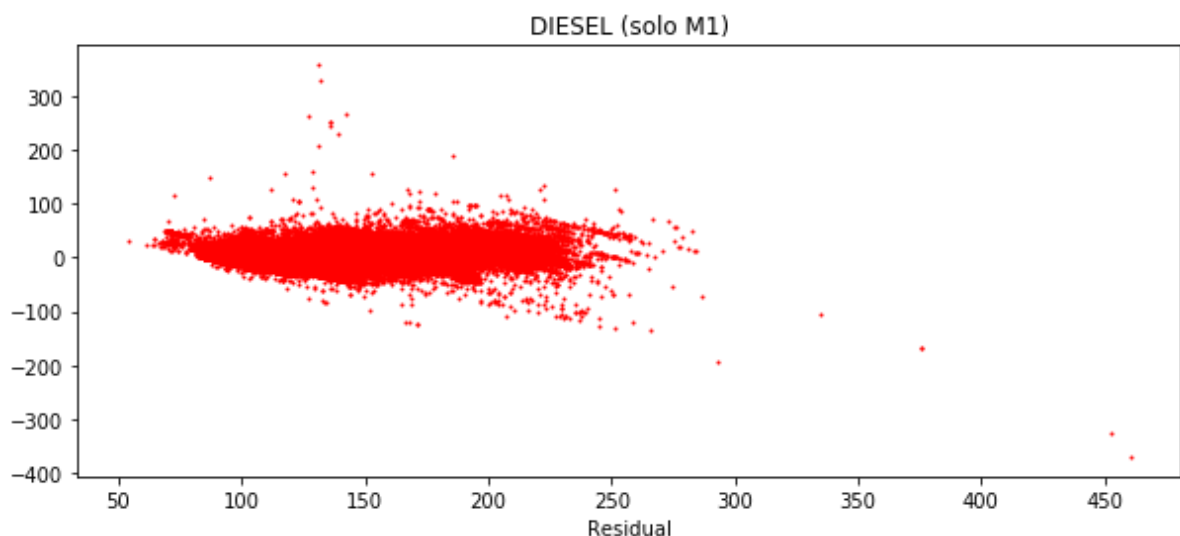
Observem els següents patrons en els residus:



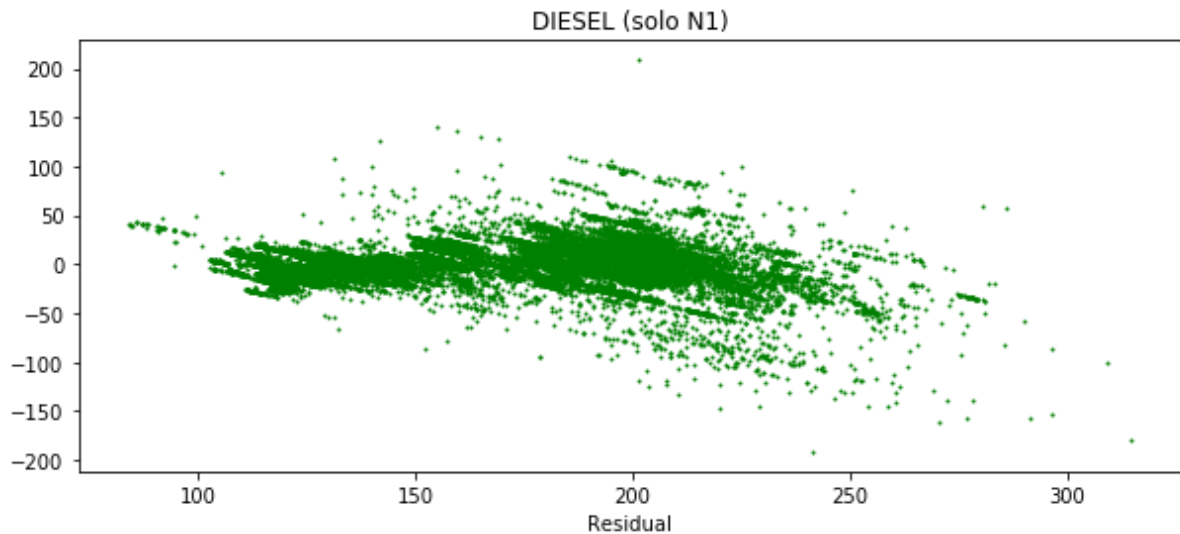


Veiem que per als motors de gasolina i dièsel, els residus mostren diversos outliers i una lleugera tendència a la heteroscedasticitat, és a dir, que la variància dels errors no és constant per les observacions realitzades. Això porta a concloure que els errors en les variàncies no són completament fiables, però els coeficients són no esbiaixats (és a dir, és probable que les variàncies trobades no siguin les mínimes).

És recomanable estudiar més en profunditat les suposicions de fons que porten a aquesta distribució dels residus. Per exemple, si reproduïm la mateixa configuració però tenint en compte *només els vehicles amb categoria M1*, veiem un patró de residus una mica més simètrica, amb menys heteroscedasticitat (només a valors baixos de CO<sub>2</sub>), tot i que encara hi ha outliers presents:



Els vehicles N1 per sí sols tenen menys observacions, amb una tendència a la baixa dels residus:



## 4.6. Optimització de models lineals

Utilitzant les regressions lineals descrites més amunt, es va realitzar una optimització dels paràmetres (variables, components PCA, categories separades o juntes) i es va escollir aquells que maximitzen les mètriques. Tenint en compte només les dades provinents de CO<sub>2</sub> de la DGT, es va arribar a un model lineal per a cada tipus de combustible:

$$\begin{aligned} \text{CO}_{2\text{-DIESEL}} = & 0.01637 * \text{CILINDRADA} + 0.6687 * \text{POTENCIA\_FISCAL} - 0.1633 \\ & * \text{POTENCIA\_NETA} + 0.0004367 * \text{MASA\_MAX} + 0.07857 * \text{TARA} + 2.863 \\ & * \text{ANTIGUITAT} - 22.88 \end{aligned}$$

$$\begin{aligned} \text{CO}_{2\text{-GASOLINA}} = & 0.01536 * \text{CILINDRADA} + 2.831 * \text{POTENCIA\_FISCAL} + 0.0146 \\ & * \text{POTENCIA\_NETA} + 0.006599 * \text{MASA\_MAX} + 0.03388 * \text{TARA} + 2.0 \\ & * \text{ANTIGUITAT} + 20.2 \end{aligned}$$

$$\begin{aligned} \text{CO}_{2\text{-HEV}} = & -0.007138 * \text{CILINDRADA} + 2.071 * \text{POTENCIA\_FISCAL} + 0.141 \\ & * \text{POTENCIA\_NETA} + 0.0316 * \text{MASA\_MAX} + 0.002087 * \text{TARA} + 0.2678 \\ & * \text{ANTIGUITAT} + 6.332 \end{aligned}$$

$$\begin{aligned} \text{CO}_{2\text{-ALTRES}} = & 0.02284 * \text{CILINDRADA} + 2.222 * \text{POTENCIA\_FISCAL} + 0.07852 \\ & * \text{POTENCIA\_NETA} + 0.0003962 * \text{MASA\_MAX} + 0.03985 * \text{TARA} + 2.121 \\ & * \text{ANTIGUITAT} + 11.36 \end{aligned}$$

El rendiment del model final per a cada combustible és:

| combustible       | RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE | Components |
|-------------------|------|----------------------|---------------------|------|------------|
| <b>Dièsel</b>     | 14.6 | 0.753                | 0.752               | 8.4% | 6          |
| <b>Gasolina</b>   | 11.8 | 0.763                | 0.771               | 6.6% | 5          |
| <b>HEV</b>        | 6.9  | 0.802                | 0.816               | 4.9% | 6          |
| <b>Uns altres</b> | 15.7 | 0.603                | 0.621               | 9.7% | 6          |

Si utilitzem la variable IDAE per a complementar els valors existents de CO2, el model resultant és el següent:

$$CO2_{DIÈSEL} = 0.02527 * \text{CILINDRADA} + 0.07.175 * \text{POTENCIA\_FISCAL} - 0.1924 * \text{POTENCIA\_NETA} + 0.004.972 * \text{MASA\_MAX} + 0.08.138 * \text{TARA} + 3.12 * \text{ANTIGUITAT} - 33.54$$

$$CO2_{GASOLINA} = 0.01696 * \text{CILINDRADA} + 3.1 * \text{POTENCIA\_FISCAL} + 0.002.618 * \text{POTENCIA\_NETA} + 0.007.507 * \text{MASA\_MAX} + 0.03.776 * \text{TARA} + 2.664 * \text{ANTIGUITAT} + 8.355$$

$$CO2_{HEV} = -0.008.084 * \text{CILINDRADA} + 2.457 * \text{POTENCIA\_FISCAL} + 0.1777 * \text{POTENCIA\_NETA} + 0.0295 * \text{MASA\_MAX} + 0.002.655 * \text{TARA} + 0.3988 * \text{ANTIGUITAT} + 6.421$$

$$CO2_{ALTRES} = 0,02284 * \text{CILINDRADA} + 2.222 * \text{POTENCIA\_FISCAL} + 0,07852 * \text{POTENCIA\_NETA} + 0.0003962 * \text{MASA\_MAX} + 0,03985 * \text{TARA} + 2.121 * \text{ANTIGUITAT} + 11.36$$

Amb aquestes mètriques:

| Combustible       | RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE  | Components |
|-------------------|------|----------------------|---------------------|-------|------------|
| <b>Dièsel</b>     | 14.5 | 0.806                | 0.806               | 7.9%  | 6          |
| <b>Gasolina</b>   | 12.1 | 0.823                | 0.834               | 6.4%  | 6          |
| <b>HEV</b>        | 7.5  | 0.799                | 0.800               | 5.2 % | 5          |
| <b>Uns altres</b> | 15.7 | 0.603                | 0.621               | 9.7%  | 6          |

## 4.7. Quadrats mínims no-negatius

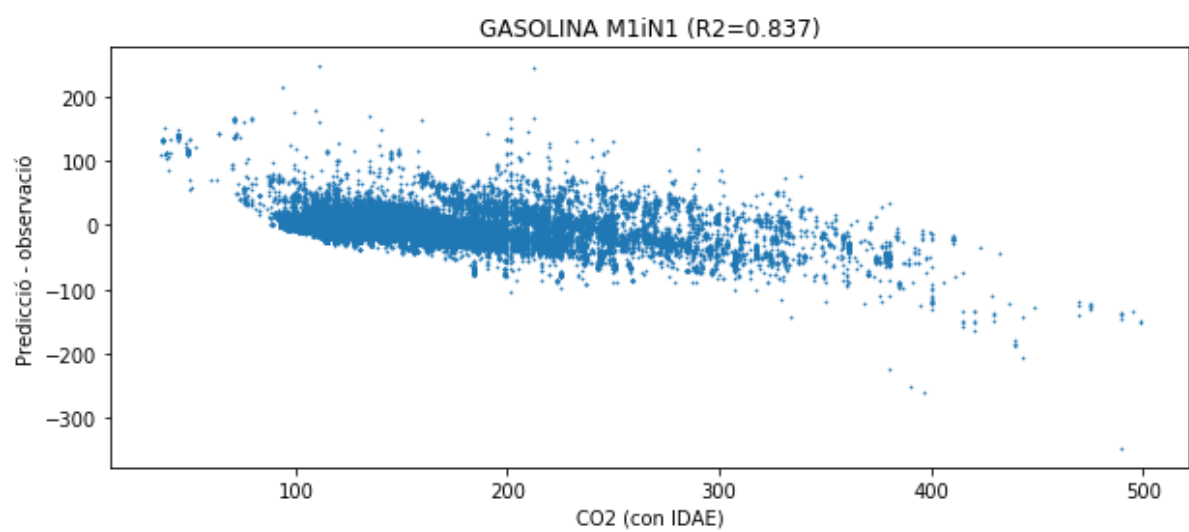
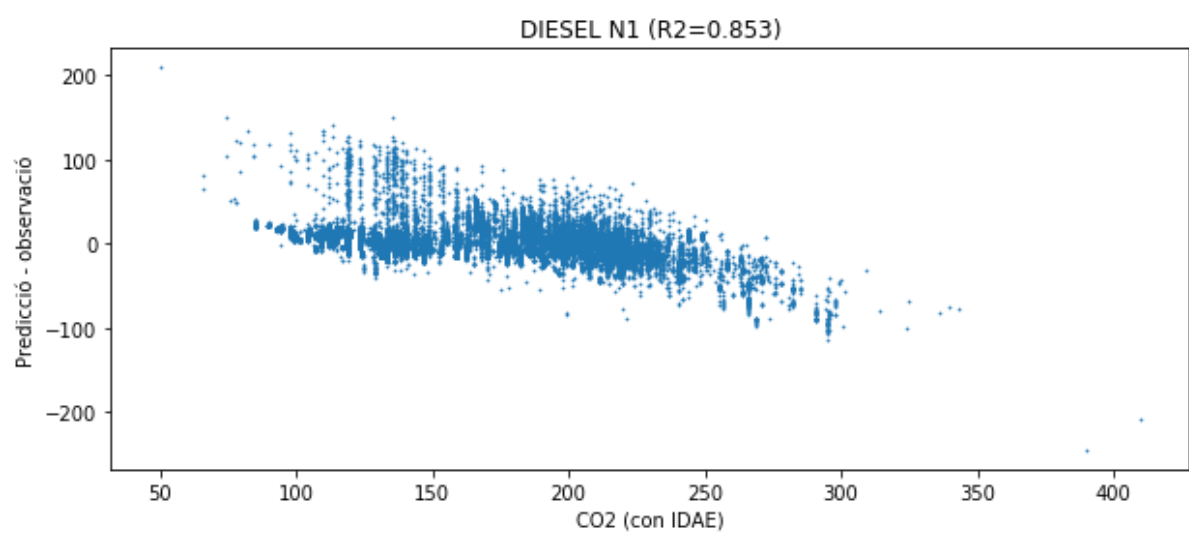
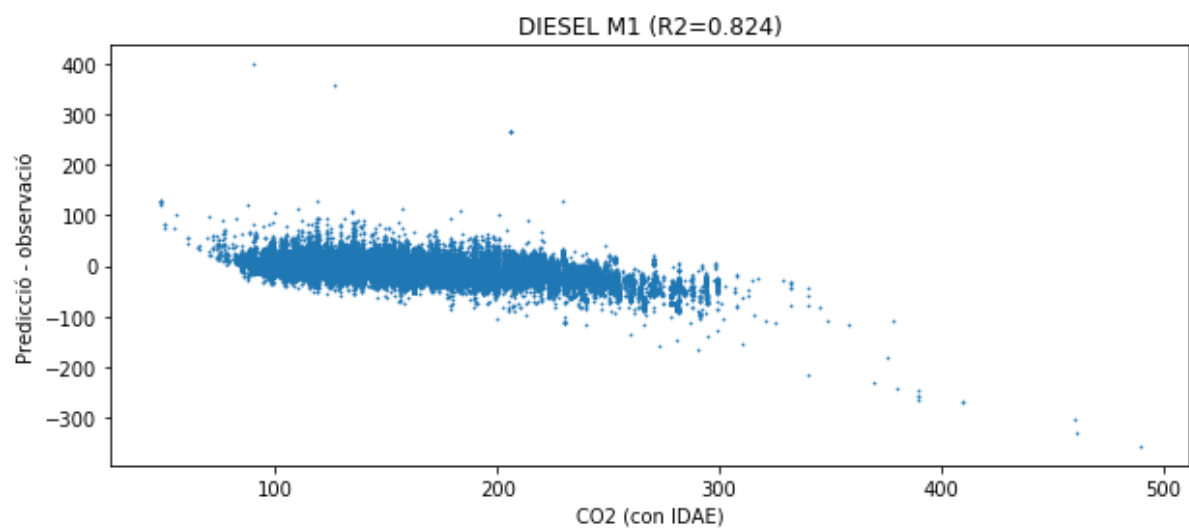
Notem que els paràmetres de la secció anterior, amb la seva transformació a components principals, resulten en alguns coeficients negatius que no són intuïtius, per exemple, potència neta per als motors dièsel. Això es deu al fet que es tracta d'un model estadístic, i no físic, i les variables seleccionades simplement minimitzen la fórmula indicada (en aquest cas la norma L2 de les variables dependents). Encara que en principi correctes, aquestes fórmules ens dificulten la interpretació, i per això se suggereix una exploració d'un model de quadrats mínims amb restriccions, en particular l'anomenat *Non-Negative Least Squares*. En aquest problema d'optimització, donada una matriu  $A$  i un vector  $i$ , es busca trobar el vector  $x$  que minimitza la distància  $|Ax - i|^2$  subjecta al fet que les components de  $x$  siguin positives. Donada la seva forma, és un problema convex, del tipus que es troba en programació quadràtica, i se soluciona de manera iterativa (en el nostre cas, simplement amb una llibreria del llenguatge *Python* que implementa aquesta solució, *scipy.optimize.nnls*).

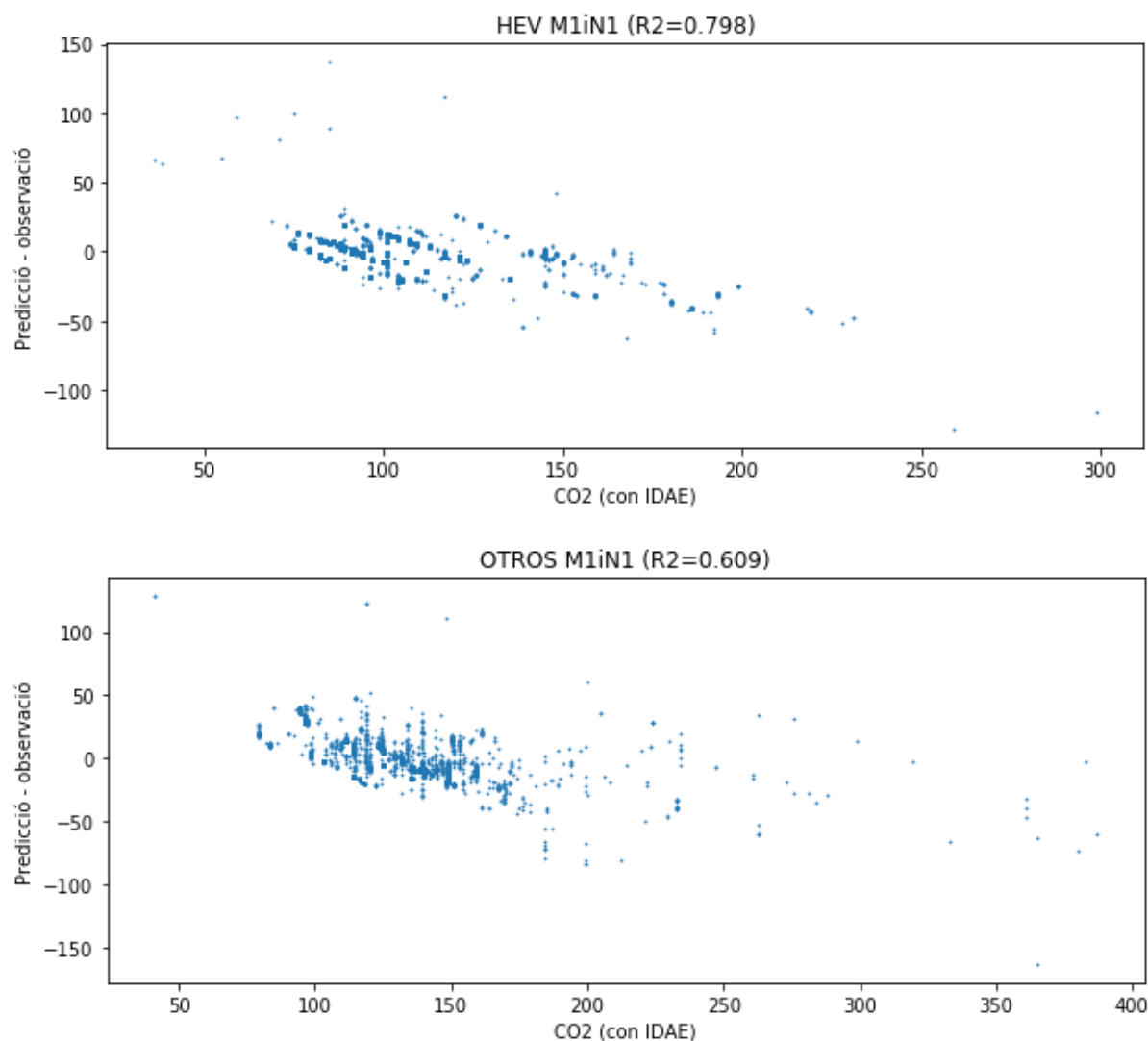
Aplicant aquesta solució hem de descartar la descomposició en components principals, tot i que ja vam veure en la secció anterior que no era òptim descartar massa components. D'altra banda, per a aquesta solució resulta més convenient separar els vehicles en categoria M1 i N1 del combustible dièsel.

Després d'optimitzar, utilitzant tant les dades de CO2 proveïdes per la DGT com les dades de de CO2 obtingudes de l'IDAE, es van aconseguir les mètriques següents:

| Combustible | Categoria | RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE |
|-------------|-----------|------|----------------------|---------------------|------|
| DIÈSEL      | M1        | 12.9 | 0.824                | 0.824               | 7.28 |
|             | N1        | 16.3 | 0.853                | 0.85                | 6.9  |
| GASOLINA    | M1 i N1   | 12.0 | 0.837                | 0.837               | 6.3  |
| HEV         | M1 i N1   | 7.5  | 0.798                | 0.79                | 5.1  |
| ALTRES      | M1 i N1   | 15.5 | 0.609                | 0.59                | 9.5  |

Els residus per a cada combustible guarden similitud amb els dels models de regressió lineal, amb una lleugera heteroscedasticitat i una tendència a subestimar les emissions de vehicles amb valors alts de CO2.





Finalment, cal preguntar-se si la tendència a subestimar les emissions de vehicles amb alts valor de CO2 és un problema dels models que hem avaluat o de les mateixes dades: hi ha menys vehicles amb alts valors de CO2, i com hem vist en l'exploració inicial hi ha molts valors que probablement siguin erronis per al CO2 (valors exageradament alts per a cotxes petits, o valors molt petits per a cotxes grans) que poden introduir el biaix observat.

## 4.8. Comparació amb altres models

### 4.8.1. Arbres de decisió

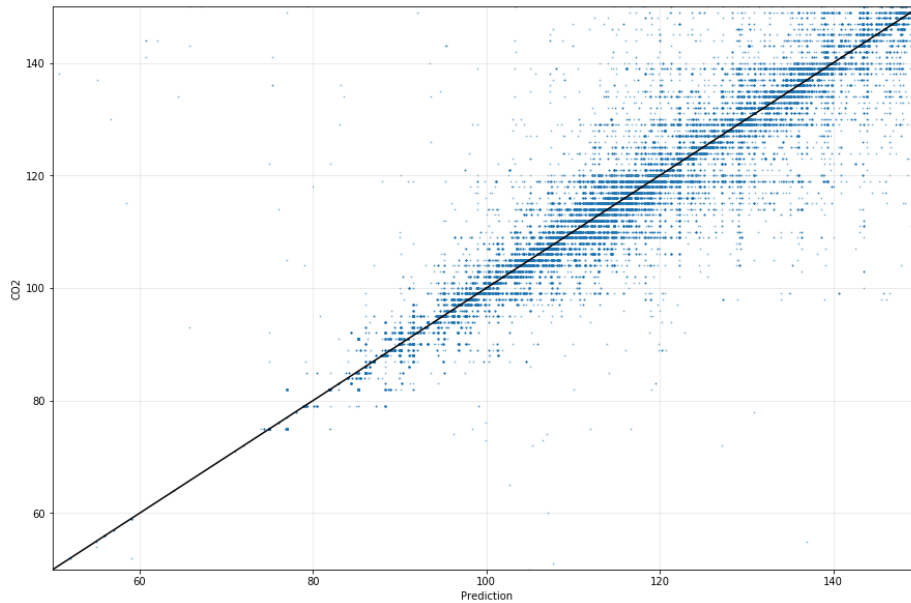
Els mètodes que fan servir arbres de decisió per a l'aprenentatge automàtic són considerats com alguns dels millors algoritmes per a l'aprenentatge supervisat. Un model típicament emprat per la seva senzillesa i gran interpretabilitat és la d'arbres de decisió. Els arbres de decisió presenten una solució al problema de classificació i regressió, i són creats a partir de la partició recursiva, tal que cada node o fulla de l'arbre representa un punt de deliberació i decisió.

Atès que només volem utilitzar aquest model per comparar els resultats amb el model de regressió lineal, només valorarem el paràmetre de màxima profunditat per a establir la mida òptima de l'arbre. Com més gran sigui l'arbre més gran és la quantitat de possibilitats o decisió, però al mateix temps s'augmenta la possibilitat d'establir massa normes que no serveixen per obtenir un millor resultat. És per això que és tan important validar la mida dels arbres.

A la taula següent es mostren les mètriques usades per a avaluar aquest model per als diferents combustibles:

| Combustible | RMSE  | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE  |
|-------------|-------|----------------------|---------------------|-------|
| Dièsel      | 16.07 | 0.922                | 0.916               | 9.01% |
| Gasolina    | 12.65 | 0.939                | 0.925               | 6.23% |
| HEV         | 5.03  | 0.97                 | 0.96                | 2.72% |
| Altres      | 13.03 | 0.984                | 0.876               | 6.23% |

Podem observar que aquests models no lineals obtenen una precisió molt més gran que els models lineals. En la gràfica següent veiem el valor correcte de CO<sub>2</sub> en funció de la predicció d'un únic model per a tots el combustibles:

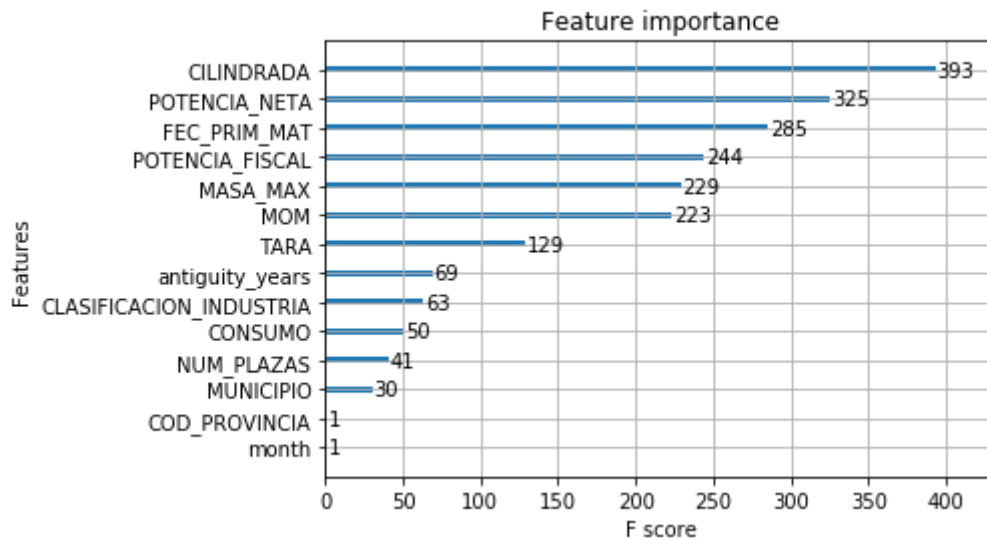


Podem observar que les prediccions són clarament no lineals, ja que cada observació cau a una fulla de l'arbre. A més a més, observem que tot i ser un únic model per a tots els diferents combustibles, s'observa una precisió molt bona.

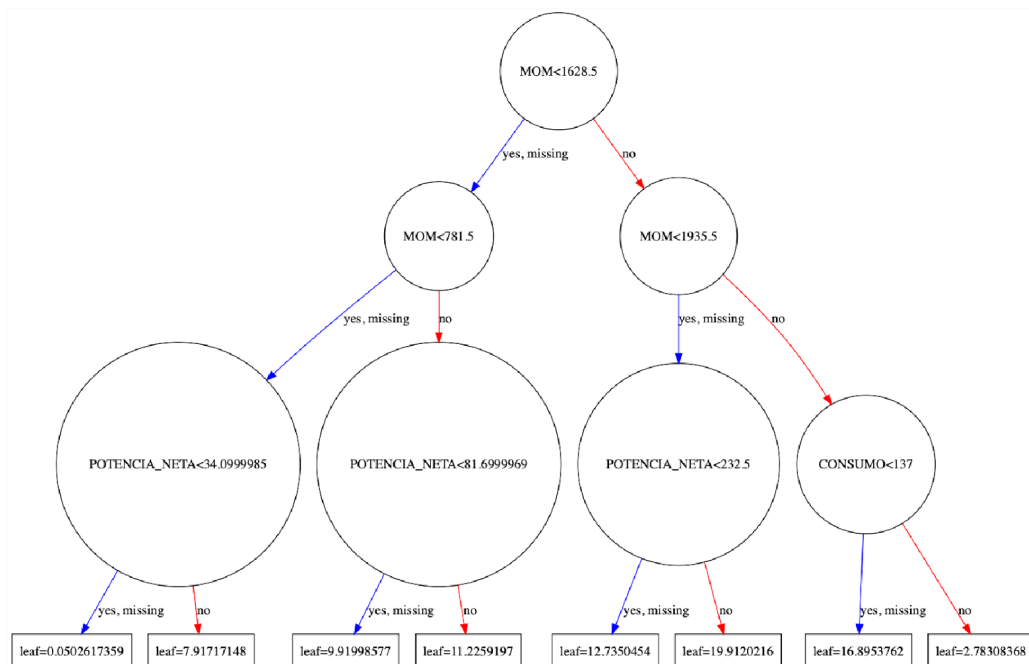
### 4.8.2. XGBoost

Extreme Gradient Boosting, més conegut com XGBoost, és un model no paramètric basat en la tècnica de *gradient boosting* mitjançant la creació d'arbres de decisió. Aquest mètode és conegut per ser un dels mètodes més eficaços a l'hora de predir taules estructurades. A més, aquest mètode té l'avantatge que pot tractar amb valors mancants sense la necessitat que siguin imputats prèviament. Això, juntament amb la seva alta flexibilitat i capacitat de paral·lelització el fan un candidat perfecte per al problema a solucionar.

El major inconvenient per a l'ús d'aquesta tècnica és la baixa interpretabilitat. En créixer arbres de decisió, ja siguin regressors o classificadors, un sobre l'altre, corregint els errors dels previs arbres, és difícil obtenir una avaluació de quines són les variables que més influeixen a l'hora de prendre les decisions. La forma més senzilla d'avaluar la importància de les variables és mesurar els pesos que donen aquests arbres a les variables utilitzades per predir el valor desitjat. La gràfica següent mostra un exemple:

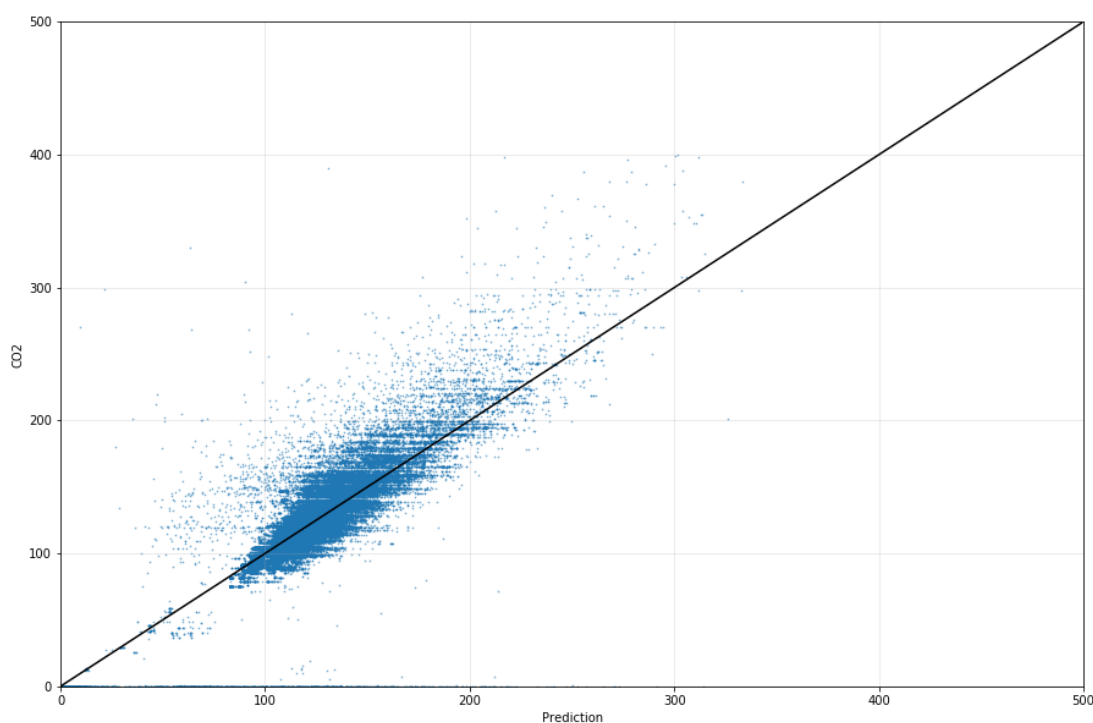


En el següent gràfic també podem observar la forma en què les particions dels arbres són creades mostrant l'exemple d'un dels arbres creats per predir el valor de CO2:



Els valors de les fulles llavors són ponderats i agrupats per tal d'obtenir una predicció final per a cada observació. Com ja hem esmentat, els valors que falten no han de ser imputats, ja que si hi ha un valor mancants, s'assumeix que la norma de l'arbre és que l'observació va per la partició a la qual pertanyen el major nombre d'observacions amb valors no mancants. D'aquesta manera també es poden tractar amb els valors que falten sense haver d'imputar-los. Això és una de les majors avantatges que presenta aquest model, ja que resulta innecessari tota la part d'imputació de valors buits.

Un altre cop, hem fet un únic regressor per a tots els combustibles. Com podem observar en el gràfic, la regressió sembla bastant encertada pel que fa a la predicció del CO2. En el següent gràfic observem que aquest mètode subestima en bastants casos els valors de CO2.



La taula següent mostra les mètriques obtingudes a partir de la creació de diferents models per a cada combustible. Observem que no comporta una millora substancial sobre els arbres de decisió, encara que probablement aquests resultats es puguin millorar probablement fent una validació adequada dels paràmetres d'aquest model. Aquesta validació no s'ha fet per culpa de la prioritització dels models lineals en aquest estudi.

| Combustible | RMSE  | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE  |
|-------------|-------|----------------------|---------------------|-------|
| Dièsel      | 16.46 | 0.871                | 0.870               | 9.43% |
| Gasolina    | 13.04 | 0.879                | 0.88                | 6.78% |
| HEV         | 4.96  | 0.960                | 0.948               | 3.08% |
| Altres      | 12.86 | 0.946                | 0.918               | 7.59% |

## 4.9. Validació contra les dades de l'IDAE

Com es va comentar anteriorment, la columna d'emissions de CO<sub>2</sub> d'IDAE proveeix una font de dades independent que pot ser utilitzada de diverses maneres. Un cas possible és fer servir el valor d'IDAE en els vehicles que no tenen CO<sub>2</sub>. Aquesta estratègia té l'avantatge que sembla enriquir el dataset substancialment, tant en nombre de vehicles com en la quantitat d'anys abastada. Per l'altre banda, les propietats estadístiques del CO<sub>2</sub> IDAE són lleugerament diferents que les del CO<sub>2</sub> dels fabricants, i no s'ha pogut explorar si això és un defecte de les dades o si realment contenen informació diferent.

En aquesta primera aproximació, el model lineal resultant utilitzant la combinació de dades de CO<sub>2</sub> de DGT i d'IDAE dona el rendiment següent:

| Combustible | RMSE | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE |
|-------------|------|----------------------|---------------------|------|
| Dièsel      | 14.5 | 0.806                | 0.806               | 7.9% |
| Gasolina    | 12.1 | 0.823                | 0.834               | 6.4% |
| HEV         | 7.5  | 0.799                | 0.800               | 5.2% |
| Altres      | 15.7 | 0.603                | 0.621               | 9.7% |

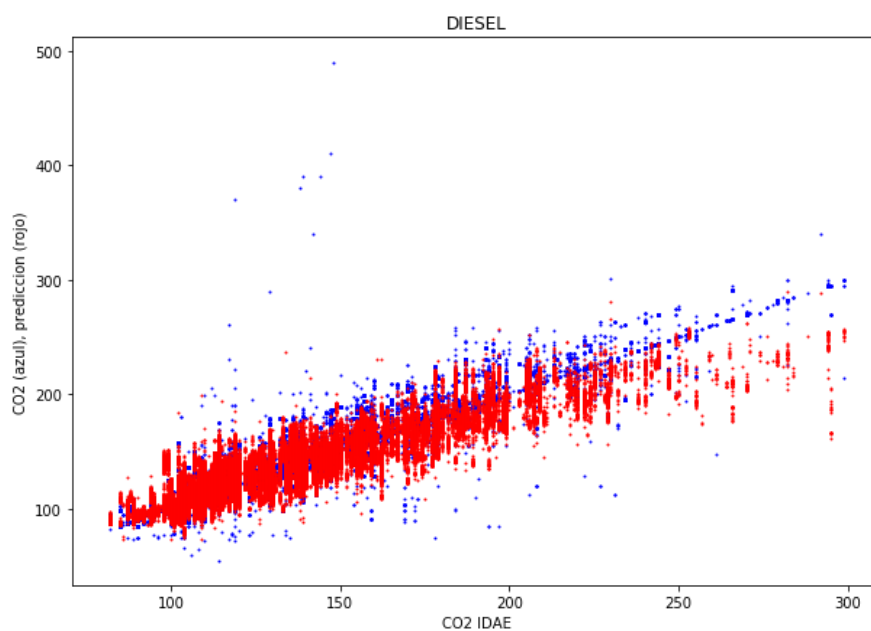
És a dir, una millora en les mètriques de R<sup>2</sup> i MAPE, i un petit empitjorament deper l'RMSE.

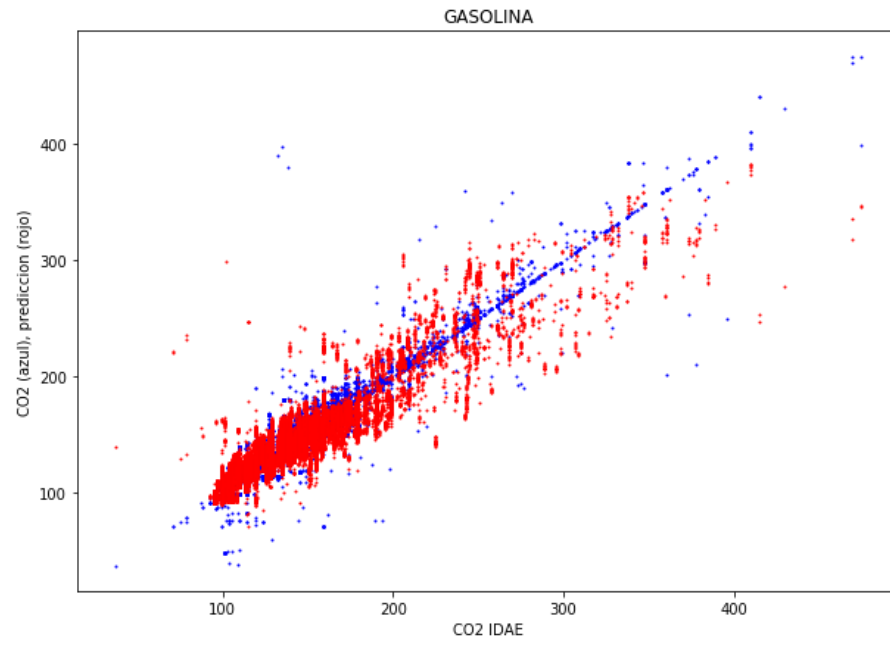
En canvi, podem fer servir el dataset d'IDAE com una validació dels resultats del model final que resulta d'entrenar només amb les dades de CO<sub>2</sub> de la DGT. En aquest cas, el que fem és prendre les dades de l'IDAE com els valors vertaders, i comparem el rendiment que té la nostra predicció contra el rendiment que té la columna de CO<sub>2</sub>.

Observem que la columna de CO<sub>2</sub> té millor correlació amb la de valors d'IDAE que els valors predits pel nostre model. En particular:

|                      | CO2 vs IDAE | Predicció vs IDAE |
|----------------------|-------------|-------------------|
| <b>Dièsel</b>        |             |                   |
| <b>RMSE</b>          | 9.2         | 13.3              |
| <b>MAPE</b>          | 4.3         | 8.2               |
| <b>R<sup>2</sup></b> | 0.875       | 0.740             |
| <b>Gasolina</b>      |             |                   |
| <b>RMSE</b>          | 6.1         | 10.8              |
| <b>MAPE</b>          | 2.3         | 6.2               |
| <b>R<sup>2</sup></b> | 0.938       | 0.802             |
| <b>HEV</b>           |             |                   |
| <b>RMSE</b>          | 7.5         | 7.2               |
| <b>MAPE</b>          | 4.7         | 5.4               |
| <b>R<sup>2</sup></b> | 0.804       | 0.786             |

Si grafiquem els valors d'ambdós enfront dels valors d'IDAE, identifiquem que el major problema es veu per als grans valors de CO2, on el model predictiu tendeix a subestimar les emissions (que ja vam veure parcialment en l'estudi dels residus).





## 5. Conclusions

### 5.1. Model final

Considerant totes les exposicions anteriors, i ponderant fortament la interpretabilitat dels models de predicció, es va decidir utilitzar el model de quadrats mínims no negatius, amb les categories M1 i N1 juntes, excepte per als motors dièsel, descartant el nombre de places, i prenent com a objectiu la informació de CO2 proveïda per la DGT quan era existent, i la d'IDAE en el seu reemplaç. Com a variable dissenyada addicional a les variables originals, només es va conservar l'antiguitat, calculada com:

$$\text{antiguitat} = (\text{Dies des de la matriculació fins al 31/12/2018}) / 365.25$$

S'entén que aquest càlcul es realitza una sola vegada per als vehicles en el parc actual, o quan un vehicle d'una altra comunitat es registra a Catalunya.

Per tal d'estimar els errors, es realitza una validació creuada aleatòria amb 20 separacions, i es calcula el model final utilitzant totes les dades disponibles.

Els coeficients finals per a cada combustible (amb l'error estimat en l'última xifra indicada entre parèntesis) són:

|                        | DIÈSEL N1 | DIÈSEL M1  | GASOLINA   | HEV       | OTROS     |
|------------------------|-----------|------------|------------|-----------|-----------|
| <b>CILINDRADA</b>      | 0.0114(3) | 0.01642(6) | 0.01149(1) | 0         | 0.034(3)  |
| <b>MASA_MAX</b>        | 0.0256(1) | 0.01140(8) | 0          | 0.0279(3) | 0         |
| <b>MOM</b>             | 0.0311(3) | 0.0575(3)  | 0.0400(1)  | 0         | 0         |
| <b>POTENCIA_FISCAL</b> | 2.70(7)   | 0          | 3.88(4)    | 0.85(9)   | 0.0(4)    |
| <b>POTENCIA_NETA</b>   | 0.0264(6) | 0          | 0          | 0.191(9)  | 0.069(7)  |
| <b>TARA</b>            | 0         | 0.0051(1)  | 0.0095(1)  | 0         | 0.0413(6) |
| <b>ANTIGÜEDAD</b>      | 2.922(9)  | 3.471(2)   | 2.605(2)   | 0.392(6)  | 1.99(4)   |
| <b>CONSTANTE</b>       | -25.6(4)  | -37.15(5)  | 4.35(9)    | 14.3(9)   | 19(1)     |

Les fórmules, escrites de manera complerta i a precisió de 4 dígitos uniforme, quedarien de la forma següent:

$$CO2_{DIÈSEL-N1} = 0.01144 * CILINDRADA + 2.699 * POTENCIA\_FISCAL + 0.02635 * POTENCIA\_NETA + 0.02562 * MASA\_MAX + 0.03115 * MOM + 0 * TARA + 2.922 * ANTIGÜEDAD - 25.64$$

$$CO2_{DIÈSEL-M1} = 0.01642 * CILINDRADA + 0 * POTENCIA\_FISCAL + 0 * POTENCIA\_NETA + 0.0114 * MASA\_MAX + 0.05745 * MOM + 0.005106 * TARA + 3.471 * ANTIGÜEDAD - 37.15$$

$$CO2_{GASOLINA} = 0.01149 * CILINDRADA + 3.879 * POTENCIA\_FISCAL + 0 * POTENCIA\_NETA + 0 * MASA\_MAX + 0.04008 * MOM + 0.009541 * TARA + 2.605 * ANTIGÜEDAD + 4.35$$

$$CO2_{HEV} = 0 * CILINDRADA + 0.8533 * POTENCIA\_FISCAL + 0.1909 * POTENCIA\_NETA + 0.02794 * MASA\_MAX + 0 * MOM + 0 * TARA + 0.3922 * ANTIGÜEDAD + 14.28$$

$$CO2_{OTROS} = 0.03399 * CILINDRADA + 0 * POTENCIA\_FISCAL + 0.06862 * POTENCIA\_NETA + 0 * MASA\_MAX + 0 * MOM + 0.04134 * TARA + 1.996 * ANTIGÜEDAD + 18.89$$

El rendiment del model final per a cada combustible és (amb l'error en la última xifra indicada en parèntesi):

| Combustible     | Categoria      | RMSE    | Train R <sup>2</sup> | Test R <sup>2</sup> | MAPE     | Observacions |
|-----------------|----------------|---------|----------------------|---------------------|----------|--------------|
| <b>DIÈSEL</b>   | <b>M1</b>      | 12.9(1) | 0.824(2)             | 0.824(3)            | 7.28(3)% | 803972       |
|                 | <b>N1</b>      | 16.3(6) | 0.853(1)             | 0.85(1)             | 6.9(2)%  | 62389        |
| <b>GASOLINA</b> | <b>M1 i N1</b> | 12.0(2) | 0.837(1)             | 0.837(5)            | 6.32(6)% | 498459       |
| <b>HEV</b>      | <b>M1 i N1</b> | 7.5(6)  | 0.798(1)             | 0.79(3)             | 5.1(2)%  | 26650        |
| <b>OTROS</b>    | <b>M1 i N1</b> | 16(2)   | 0.609(4)             | 0.6(1)              | 9.5(7)%  | 5484         |

## 5.2. Recomanacions futures

En aquesta secció analitzem les possibles passes futures a seguir en l'hipotètic cas que sigui oportú seguir millorant els models, tant d'imputació com d'estimació.

En primer lloc, s'ha d'utilitzar mètodes per a netejar el camp de models. Com ja hem demostrat anteriorment, existeixen una gran quantitat d'errors existents a la base de dades a causa de la introducció errònia de caràcters. Aquesta neteja no s'ha pogut realitzar per la curta duració del projecte, i la gran dificultat que suposa o presenta netejar aquests de forma automàtica. Recomanem tècniques i eines que poden ser útils com ara l'*OpenRefine* de *Google*.

Una possible forma de procedir seria comptar els models més comuns a la base de dades, i poc a poc netejar les categories menys comuns, intentant trobar un model que teòricament hauria de ser el mateix, però a l'hora d'introduir les dades s'han introduït de forma errònia.

Aquest mètode pot necessitar invertir molt de temps, però potencialment podria ajudar a mitigar l'observació que hem fet més a dalt: *En els models proposats en aquest informe hi ha una molt alta possibilitat que dos vehicles idèntics (marca/model/any), que haurien de tenir les mateixes emissions, apareguin amb un càlcul de CO2 estimat diferent* només perquè un d'ells té algun error en alguna de les seves característiques. Això no significa que hi hagi un error en el model predictiu, sino en la base de dades i parcialment en com s'imputen els valors buits.

En segon lloc, seria recomanable explorar altres models que molt probablement siguin més precisos a l'hora d'estimar els valors de CO2 desitjats. En aquest informe només hem treballat amb dos dels mètodes més comuns, que són els arbres de decisió i els Extreme Gradient Boosting, els quals donen uns resultats molt vàlids sense incidir en la validació dels paràmetres. És més, per a tots els combustibles, les mètriques mostrades superen la regressió lineal. El problema d'aquests models és l'alta complexitat i, en el cas de l'XGBoost, la difícil interpretabilitat. Aquest darrer model presenta la gran avantatge que suposa la capacitat de no haver d'imputar els valors mancants, encara que, en qualsevol cas, és recomanable valorar tant amb els mètodes proposats d'imputació com sense ells, i veure com es poden aconseguir millors resultats. El balanç entre precisió i interpretabilitat és un dels temes més discutits dins del món de la inferència estadística computacional. Degut als requeriments d'aquest estudi, s'han emfatitzat els models lineals, però seria interessant fer una exploració més detallada sobre altres models.

En tercer lloc, seria recomanable treballar més en profunditat la validació dels resultats comparant aquests amb les dades provinents d'IDAE, per entendre la lleugera heteroscedasticitat observada en els residus.

Finalment, caldria revisar la variable DES\_TIPO, on podem observar moltes incongruències. Vehicles que en teoria són M1 o N1, apareixen com a motocicletes en el camp de DES\_TIPO. El problema, com ja hem esmentat abans, és que no tenim forma addicional de corroborar quin dels dos camps és correcte.