01: Preprocessament de les dades

Soulaiman el Hamri

2025-05-03

Contents

1. Introducció

Aquest document descriu el **preprocessament inicial de les dades meteorològiques i de qualitat de l'aire** registrades a la ciutat de Barcelona. Aquest pas és essencial per garantir la **coherència, qualitat i integritat** de les dades abans de realitzar qualsevol anàlisi estadística o modelització posterior.

1.1. Fonts de dades

- **Dades meteorològiques (1991–2025)**: registres diaris de les estacions meteorològiques de Barcelona.
- **Metadades meteorològiques**: descripcions de cada acrònim i unitats.
- **Fitxer d'estacions**: ubicació i municipi de cada estació meteorològica.
- **Dades de qualitat de l'aire**: registres històrics dels principals contaminants (NO2, PM10, PM2.5, O3, etc.).

2. Descripció de les dades

2.1. Dades meteorològiques (Meteocat)

Els conjunts de dades meteorològiques han estat descarregats del portal de dades obertes de la Generalitat de Catalunya. Cada fitxer conté **observacions diàries** mesurades per les estacions automàtiques de Barcelona (XEMA). A Barcelona hi ha 4 estacions operatives i tenim un fitxer csv per cada estació.

Les columnes principals són:

- **DATA:** Data de la mesura, en format `"YYYY-MM-DD"` o `"YYYY-MM-DD HH:MM:SS"`.
- **CODI_ESTACIO:** Codi identificador únic de l'estació meteorològica.
- **ACRÒNIM:** Sigla curta que identifica la variable mesurada (ex.: `TM`, `TX`, `HRM`, `PPT`, `PM`...)
- **VALOR:** Valor mesurat de la variable corresponent.

Aquesta informació es completa amb un fitxer de metadades, que inclou:

- **ACRÒNIM:** Coincideix amb el del dataset principal.
- **NOM_VARIABLE:** Nom complet de la variable (ex.: "Temperatura mitjana diària").
- **UNITAT:** Unitat de mesura (ex.: °C, %, mm, hPa).
- **CODI_VARIABLE:** Identificador intern de la variable.

A continuació es mostra una taula amb les variables meteorològiques rellevants:

ACRÒNIM	NOM VARIABLE	UNITAT
TM	Temperatura mitjana diària	°C
HRM	Humitat relativa mitjana	%
PPT	Precipitació acumulada	mm
PM	Pressió atmosfèrica mitjana	hPa
VVM10	Velocitat mitjana del vent a 10 m	m/s

Aquestes dades són fonamentals per contextualitzar els nivells de contaminació en funció de la meteorologia.

2.2. Dades de qualitat de l'aire

El conjunt de dades conté registres de **mesures horàries** de concentració de contaminants atmosfèrics des de l'any 1991 fins al 2025. L'origen de les dades és de portal Open Data BCN.

Les columnes principals inclouen:

- **CODI_ESTACIO**: Codi alfanumèric de l'estació de mesura.
- **NOM_ESTACIO:** Nom de l'estació (ex.: "Eixample", "Gràcia", "Zona Universitària"...).
- **DATA:** Data de mesura.
- **HORA:** Hora de mesura (de 1 a 24).

- **CONTAMINANT:** Substància mesurada (ex.: `NO2`, `PM10`, `PM2.5`, `O3`).
- **VALOR:** Valor de concentració horària (normalment en µg/m³).

Es descarten columnes complementàries com:

- magnitud, codi_ine, municipi, geocoded_column, codi_comarca, nom_comarca

Contaminants d'interès seleccionats:

- **NO2 (diòxid de nitrogen)**: Indica presència de trànsit intens.
- **PM10 i PM2.5 (partícules en suspensió)**: Afavoreixen problemes respiratoris.
- **O3 (ozó troposfèric)**: Forma contaminant secundari amb impactes en salut i vegetació.

3. Preprocessat

3.1. Dades meteorologiques

Aquest apartat descriu detalladament el procés de preparació inicial de les dades meteorològiques corresponents als anys 1995 a 2025. El flux de treball inclou la lectura dels fitxers, la unificació en un únic conjunt de dades, l'enriquiment amb metadades descriptives, la conversió correcta de dates i el filtratge per seleccionar només dades rellevants per a l'anàlisi.

Primerament, farem la lectura dels fitxers per estacions:

```
# Forcem que totes les columnes es llegeixin com a text per evitar errors en unir-les
col_types_forcats <- cols(.default = "c")

# Llegim els CSVs de les estacions de Barcelona des de les seves rutes
df_D5 <- read_csv("../data/raw/meteocat/dades_estacions/30597_D5.csv",
                  col_types = col_types_forcats, na = c("", "NA"))
df_X2 <- read_csv("../data/raw/meteocat/dades_estacions/30597_X2.csv",
                  col_types = col_types_forcats, na = c("", "NA"))
df_X4 <- read_csv("../data/raw/meteocat/dades_estacions/30597_X4.csv",
                  col_types = col_types_forcats, na = c("", "NA"))
df_X8 <- read_csv("../data/raw/meteocat/dades_estacions/30597_X8.csv",
                  col_types = col_types_forcats, na = c("", "NA"))
```

Un cop llegits, els quatre conjunts s'uneixen en un únic dataframe mitjançant `bind_rows()`. Aquesta operació genera un sol conjunt homogeni, amb totes les observacions agrupades.

```
# Unifiquem tots els conjunts de dades en un únic dataframe
df_meteo <- bind_rows(df_D5, df_X2, df_X4, df_X8)
```

A continuació, es genera un resum estadístic (`summary(df_meteo)`) per identificar possibles inconsistències o valors anòmals.

```
# Mostrem un resum del conjunt complet
summary(df_meteo)
```

```
##      EMA                DATA                TM                TX
##  Length:30287       Length:30287       Length:30287       Length:30287
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  Data Extrem TX           TN           Data Extrem TN          HRM
##  Length:30287       Length:30287       Length:30287       Length:30287
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##      HRX            Data Extrem HRX        HRN            Data Extrem HRN
##  Length:30287       Length:30287       Length:30287       Length:30287
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       PM                 PX           Data Extrem PX           PN
##  Length:30287       Length:30287       Length:30287       Length:30287
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  Data Extrem PN         PPT               RS24h              VVM10
##  Length:30287       Length:30287       Length:30287       Length:30287
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     DVM10              VVX10           Data Extrem VVX10     DVVX10
##  Length:30287       Length:30287       Length:30287       Length:30287
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Per entendre i etiquetar correctament les variables (ex. TX = Temperatura màxima, HRM = Humitat mitjana), aquestes metadades són clau per enriquir les observacions i preparar-les per visualització i interpretació.

```r
# Carreguem el fitxer de metadades
df_metadata <- read_csv("../data/raw/meteocat/MeteoCat_Metadades.csv", col_types = cols(
  CODI_VARIABLE = col_double(),
  NOM_VARIABLE = col_character(),
  UNITAT = col_character(),
  ACRÒNIM = col_character()
))
```

La columna de data s'estandarditza al format ISO (`%Y-%m-%dT00:00:00Z`) per garantir compatibilitat amb sistemes com ArcGIS Online i assegurar la correcta ordenació temporal.

```r
# Normalitzem la data de lectura
df_meteo <- df_meteo %>%
  mutate(DATA_LECTURA = dmy(DATA) %>% format("%Y-%m-%dT00:00:00Z"))
```

Els fitxers meteorològics tenen una estructura de "taula ampla", on cada variable meteorològica és una columna. Per facilitar les operacions d'anàlisi, es transforma a format "longitudinal" o **long format**, on cada fila representa un registre d'una sola variable per dia i estació. Això es fa mitjançant una iteració sobre cada acrònim (ex. TX, HRM, PPT) i es construeix un dataframe agregat.

A més, per a certes variables que tenen una **data extrema** associada (ex. temperatura màxima amb hora), també es processa i incorpora aquesta informació en format ISO 8601.

```r
# Variables que tenen columna de 'Data Extrem'
variables_amb_extrem <- c("TX", "TN", "HRX", "HRN", "PX", "PN", "VVX10")
# Transformem a format llarg
acronims <- df_metadata$ACRÒNIM

resultats <- map_dfr(acronims, function(acronim) {
  col_valor <- sym(acronim)

  # Si la variable té data extrem, l'afegim
  if (acronim %in% variables_amb_extrem) {
    col_extrem <- sym(paste0("Data Extrem ", acronim))
    dades <- df_meteo %>%
      select(DATA_LECTURA, CODI_ESTACIO = EMA, valor = !!col_valor,
             data_extrem = !!col_extrem)
  } else {
    dades <- df_meteo %>%
      select(DATA_LECTURA, CODI_ESTACIO = EMA, valor = !!col_valor) %>%
      mutate(data_extrem = NA)
  }

  # Tractament del valor i format final
  dades %>%
    filter(!is.na(valor)) %>%
    mutate(
      ACRÒNIM = acronim,
      VALOR = str_replace(valor, ",", ".") %>% as.numeric(),
      DATA_EXTREM = if_else(
        !is.na(data_extrem),
        format(ymd_hms(data_extrem), "%Y-%m-%dT%H:%M:%SZ"),
        NA_character_
      )
    ) %>%
    select(DATA_LECTURA, DATA_EXTREM, CODI_ESTACIO, ACRÒNIM, VALOR)
}) %>%
  left_join(df_metadata, by = "ACRÒNIM") %>%
  select(DATA_LECTURA, DATA_EXTREM, CODI_ESTACIO, ACRÒNIM, VALOR, CODI_VARIABLE,
         NOM_VARIABLE, UNITAT)
```

Es comprova la presència de `NA` per columna amb `colSums(is.na(...))` i es realitza una ordenació final per data, estació i variable (`arrange()`) per garantir una estructura ordenada i fàcilment interpretable.

```r
na_summary <- colSums(is.na(resultats))
print(na_summary)
```

```
##  DATA_LECTURA   DATA_EXTREM  CODI_ESTACIO       ACRÒNIM         VALOR
##             0        190773             0             0             0
## CODI_VARIABLE  NOM_VARIABLE        UNITAT
##             0             0             0
```

```r
#  Ordenem per data, estació i variable
resultats <- resultats %>%
  arrange(DATA_LECTURA, CODI_ESTACIO, ACRÒNIM)
```

Els valors NA (valors buits) a la columna DATA_EXTREM es deuen al fet que **no totes les variables meteorològiques tenen una data extrema associada**. Això és un comportament esperat i correcte segons la naturalesa de les dades.

Per últim, el resultat és un fitxer `.csv` preparat i net, que conté totes les observacions diàries de les quatre estacions meteorològiques, en format llarg i enriquit amb metadades.

```
# Guardem el resultat final en CSV
write_csv(resultats, "../data/processed/meteocat/meteocat_1995_2025_bcn_processed.csv")
```

3.2. Dades de contaminants

En aquest apartat es duu a terme el preprocessament de les dades de qualitat de l'aire, provinents del conjunt històric (1991–2025) que conté mesuraments horaris dels principals contaminants atmosfèrics a diferents estacions de Catalunya. Ens centrarem exclusivament en les estacions **ubicades dins del municipi de Barcelona**, i seleccionarem només aquells **contaminants d'interès** per a l'estudi, amb un filtratge, neteja i preparació estructurada per a l'anàlisi posterior.

```
# Carreguem les dades crues de qualitat de l'aire
contaminants_raw <- read_csv("../data/raw/contaminants/dades_qualitat_aire_1991_2025.csv")

head(contaminants_raw)
```

```
## # A tibble: 6 x 40
##   codi_eoi nom_estacio          data                magnitud contaminant unitats
##   <chr>    <chr>                <dttm>                 <dbl> <chr>       <chr>
## 1 08307012 Vilanova i la Geltrú 2025-04-07 00:00:00        6 CO          mg/m3
## 2 43171002 Vila-seca (IES Vila~ 2025-04-07 00:00:00       14 O3          µg/m3
## 3 43162005 Vandellòs (Barranc ~ 2025-04-07 00:00:00        7 NO          µg/m3
## 4 25120001 Lleida               2025-04-07 00:00:00        9 PM2.5       µg/m3
## 5 25196001 Montsec              2025-04-07 00:00:00       14 O3          µg/m3
## 6 43103001 Perafort (Puigdelfí) 2025-04-07 00:00:00       12 NOX         µg/m3
## # i 34 more variables: tipus_estacio <chr>, area_urbana <chr>, codi_ine <chr>,
## #   municipi <chr>, codi_comarca <chr>, nom_comarca <chr>, h01 <dbl>,
## #   h02 <dbl>, h03 <dbl>, h04 <dbl>, h05 <dbl>, h06 <dbl>, h07 <dbl>,
## #   h08 <dbl>, h09 <dbl>, h10 <dbl>, h11 <dbl>, h12 <dbl>, h13 <dbl>,
## #   h14 <dbl>, h15 <dbl>, h16 <dbl>, h17 <dbl>, h18 <dbl>, h19 <dbl>,
## #   h20 <dbl>, h21 <dbl>, h22 <dbl>, h23 <dbl>, h24 <dbl>, altitud <dbl>,
## #   latitud <dbl>, longitud <dbl>, geocoded_column <chr>
```

```
# Filtratge per municipi (municipi == "Barcelona")
df_bcn_contaminants <- contaminants_raw %>%
  filter(municipi == "Barcelona")

# Definim els contaminants rellevants per a l'estudi
contaminants_interessants <- c("NO2", "PM10", "PM2.5", "O3")

# Seleccionem només les observacions corresponents a aquests contaminants
df_bcn_contaminants <- df_bcn_contaminants %>%
  filter(contaminant %in% contaminants_interessants)

# Comptem quantes observacions hi ha per estació i contaminant
df_bcn_contaminants %>%
```

```
  count(nom_estacio, contaminant) %>%
  arrange(nom_estacio, contaminant)
```

```
## # A tibble: 34 x 3
##    nom_estacio                       contaminant     n
##    <chr>                             <chr>       <int>
##  1 Barcelona (Ciutadella)            NO2          7383
##  2 Barcelona (Ciutadella)            O3           7475
##  3 Barcelona (Eixample)              NO2          9231
##  4 Barcelona (Eixample)              O3           9525
##  5 Barcelona (Eixample)              PM10         6684
##  6 Barcelona (Eixample)              PM2.5         829
##  7 Barcelona (Gràcia - Sant Gervasi) NO2          9422
##  8 Barcelona (Gràcia - Sant Gervasi) O3           9528
##  9 Barcelona (Gràcia - Sant Gervasi) PM10         4557
## 10 Barcelona (Gràcia - Sant Gervasi) PM2.5         253
## # i 24 more rows
```

```
# Eliminem columnes que no aporten valor analític directe
df_bcn_contaminants <- df_bcn_contaminants %>%
  select(-magnitud, -codi_ine, -municipi, -codi_comarca, -nom_comarca, -geocoded_column)


# Comprovem la presència de valors nuls a les columnes principals
na_summary_contaminants <- colSums(is.na(df_bcn_contaminants))
print(na_summary_contaminants)
```

```
##      codi_eoi   nom_estacio          data   contaminant       unitats
##             0             0             0             0             0
## tipus_estacio   area_urbana           h01           h02           h03
##             0             0          4808          5175          4842
##           h04           h05           h06           h07           h08
##          4684          4803          4863          4827          4955
##           h09           h10           h11           h12           h13
##          5688          6428          8044          8604          8487
##           h14           h15           h16           h17           h18
##          7452          6829          6716          5241          4535
##           h19           h20           h21           h22           h23
##          4436          4322          4363          4547          4562
##           h24      altitud       latitud     longitud
##          4813             0             0             0
```

```
# Ordenem les dades per facilitar l'anàlisi temporal posterior
df_bcn_contaminants <- df_bcn_contaminants %>%
  arrange(nom_estacio, contaminant, data)

# Crear carpeta de sortida si no existeix
dir.create("../data/processed/contaminants", recursive = TRUE, showWarnings = FALSE)

# Guardar el dataset filtrat
write_csv(df_bcn_contaminants,
          "../data/processed/contaminants/contaminants_bcn_filtrat.csv")
```

El dataset resultant conté les observacions horàries dels contaminants **NO2, PM10, PM2.5 i O3** a les diferents estacions de Barcelona, estructurat cronològicament i amb les columnes essencials per a l'anàlisi temporal i espacial. En la següent fase es podrà integrar amb les dades meteorològiques per estudiar les relacions entre **qualitat de l'aire** i **condicions ambientals**.

4. Conclusions generals

Aquest document ha permès dur a terme una primera fase fonamental del projecte: la preparació, neteja i estructuració dels conjunts de dades que seran la base de les anàlisis posteriors sobre la qualitat de l'aire a la ciutat de Barcelona.

S'han abordat dues fonts de dades complementàries:

- **Dades meteorològiques (1995–2025)**, proporcionades pel Servei Meteorològic de Catalunya, que han estat unificades, enriquides amb metadades descriptives i filtrades geogràficament per a les estacions situades dins del municipi de Barcelona. El conjunt resultant conté variables ambientals clau (temperatura, humitat, precipitació, pressió atmosfèrica, etc.) en format net i estandarditzat.
- **Dades de contaminants atmosfèrics (1991–2025)**, provinents de l'administració ambiental catalana, que han estat transformades i filtrades per obtenir observacions horàries dels contaminants **NO2, PM10, PM2.5 i O3** a estacions urbanes de Barcelona. Les dades han estat reorganitzades cronològicament i simplificades per facilitar l'anàlisi temporal i espacial.

Gràcies a aquest treball de preprocessament:

- S'ha garantit la **coherència temporal i geogràfica** entre les diferents fonts.
- S'ha assegurat la **qualitat i integritat de les dades**, descartant camps no rellevants i detectant possibles valors nuls.
- S'han creat conjunts de dades **preparats per a la fusió** i anàlisi conjunta, amb l'objectiu d'estudiar **les relacions entre condicions meteorològiques i nivells de contaminació atmosfèrica**.

A partir d'aquesta base sòlida, en la següent fase del projecte es podrà dur a terme una anàlisi exploratòria, visualització de sèries temporals, estudi de correlacions i construcció de models explicatius o predictius.

Aquest preprocessament inicial constitueix, doncs, una etapa **clau per garantir la robustesa i el rigor analític** de tot el treball posterior.