ANÀLISI I PREDICCIÓ TEMPORAL I ESPACIAL DE LA CRIMINALITAT A GIRONA

Sergi Payarol Calzado

Directora: Isabel Salamaña

Treball Final de Grau

Facultat de Lletres – UdG

9 de juny del 2023

Taula de contingut

Agraïments

A la Policia Municipal de Girona, Mossos d'Esquadra i a la Unitat Municipal d'Anàlisi Territorial per donar-me accés a les dades per realitzar aquest treball. Sobretot a l'Arnau Pararol, per la seva constant ajuda al llarg d'aquests mesos i a la Isabel Salamaña, pels seus consells i ajuda que tant han calgut.

Agraeixo a en David Pavón el seu gran suport a l'hora de gestionar les tramitacions i convenis pertinents per poder accedir les dades de la policia. A en Xavi Garcia Acosta, per haver-me ajudat i motivat en tot moment.

I sobretot, agrair a l'Andrea Garcia la seva paciència i suport, ja que sense ella aquest treball no hagués estat possible.

RESUM

En aquest treball es realitza una anàlisi i una predicció de la criminalitat a la ciutat de Girona dels últims anys. Les dades que s'han utilitzat procedeixen de la Policia Municipal i dels Mossos d'Esquadra de Girona. Aquestes engloben tots els registres sobre els delictes ocorreguts a la ciutat, com també les persones detingudes pel propi fet delictiu. Les dades sobre els delictes tenen una franja temporal que va del 2000 al 2023, mentre que les dades sobre les persones detingudes van del 2018 fins el 2022.

Les dades contenen informació confidencial. És per aquest motiu que no es recullen en aquest TFG els apartats on es mostren dades sensibles. Per aquesta raó, les agrupacions espacials sempre s'han realitzat a una escala de sectors i subsectors policials.

El projecte comença fent una explicació sobre els instruments, proves i components utilitzats a l'etapa d'anàlisi predictiva. S'explica el procediment realitzat a l'etapa de preprocessament de les dades, per així poder analitzar més còmodament les dades a l'etapa d'anàlisi exploratòria de les dades. És en aquest apartat on s'han respost diferents preguntes: Què? Quan? On? Qui? I Per què? Sent aquestes les preguntes més típiques en una anàlisi d'intel·ligència policial i que es troben documentades a la guia de l'Organització per a la Seguretat i la Cooperació a Europa (OSCE, 2017).

En l'etapa d'anàlisi predictiva s'han realitzat diferents tipus de models per sèries temporals corresponents a cada sector policial. La finalitat d'aplicar diferents models és el de discernir sobre quins poden ser més fiables o adequats per predir la mitjana de delictes per dies de cada mes que poden ocorre en els mesos següents a la data actual. El propòsit d'aquests models és veure si, a partir de les dades registrades, pot ser viable fer prediccions amb aquests models.

Paraules clau: Criminalitat, espai geogràfic, temps, predicció, sèrie temporal, Girona, Policia predictiva, delictes, detinguts.

1. INTRODUCCIÓ

Aquest TFG s'ha realitzat amb la col·laboració de la Policia Municipal de Girona, els Mossos d'Esquadra i la Unitat Municipal d'Anàlisi Territorial (UMAT) de l'Ajuntament de Girona. El principal objectiu és realitzar anàlisis avançades amb caràcter prospectiu, per predir la criminalitat de Girona a partir de dades sobre delictes ocorreguts en els últims anys. Així doncs, el que es proposa en aquest projecte és dissenyar una eina composta per un seguit de models predictius per sèries temporals pels diferents sectors policials de la ciutat, i així poder aportar informació de valor a la policia perquè els ajudi a millorar la gestió dels seus recursos. La predicció anirà enfocada a saber la mitjana de delictes per dies de cada mes en una franja de tres a sis mesos a futur des de la data actual. Per tal d'arribar a aquesta fita, s'ha utilitzat una metodologia estructurada en diferents fases i que es documenta en l'apartat metodològic d'aquest treball.

És important explicar que a part de la predicció sobre la criminalitat de Girona, s'ha realitzat una anàlisi exploratòria de les dades on s'ha pogut treure força informació. No obstant, l'objectiu és obtenir una visió general dels delictes històrics de la ciutat. En algunes ocasions, aquesta anàlisi pot generar preguntes que no s'han respost. Cal remarcar que un dels propòsits del treball és el de crear una base de partida sobre la criminalitat de Girona i que no s'ha aprofundit en temes concrets ni s'han donat resposta a tots els dubtes que han sorgit. La idea és que, a partir d'aquesta informació surtin nous projectes que afinin encara més en certes tipologies del delicte o analitzar a fons, per exemple, el perquè de la baixada de delictes en certs moments de la sèrie temporal.

Feta aquesta primera introducció formal, passarem a posar en context el tema principal. Els fets delictius són ben presents en la nostra societat, i representen la cara fosca de la vida en els llocs, del territori. Però, què en sabem d'aquests delictes? Quines tipologies hi ha? A on succeeixen? Quan solen ocórrer? Qui els comet? Són preguntes que segurament ens hem fet vàries vegades a la nostra vida i possiblement la resposta ha variat al llarg del temps. És més, tota la informació que podem obtenir procedeix directament dels mitjans de comunicació. Si utilitzem la prospecció en aquest tema sorgeixen més preguntes com: es poden prevenir els delictes? Es poden predir?

Es tracta de preguntes molt agosarades de fer, però tot això ja es porta analitzant des de fa anys per la policia predictiva.

Si observem l'acte delictiu com un punt que s'engloba dins de l'espai-temps, el conjunt d'aquests punts formen una sèrie temporal, on queda recollida de manera seqüencial i ordenada la informació. A partir d'aquí es pot analitzar la sèrie amb diferents tècniques per intentar detectar certs patrons. Aquestes anàlisis avançades són els que duen a terme la policia predictiva.

Als Estats Units s'han implementat diferents mecanismes que ajuden a combatre el crim. Alguns d'aquests projectes són l'"Operació LASER, que va començar al 2011" (Lapowsky,

I., 2018) i que identifica àrees on es creu que es pot produir violència armada; o "PredPol, que ajuda a localitzar els *hot spots* amb una alta probabilitat de delictes relacionats amb la propietat" (Lau, T., 2020). A Espanya també s'han fet diferents projectes. També, des del món acadèmic hi ha diversos autors que han fet anàlisis i prediccions molt completes i interessants i que analitzarem breument en el marc teòric.

Així doncs, el que es pretén en aquest Treball Final de Grau és analitzar i predir la criminalitat de Girona.

Les dades corresponen als registres sobre els delictes registrats en els últims 23 anys (2000 – 2023), com també als registres de les persones detingudes dels últims 4 anys (desembre 2018 – 2022). A partir d'aquestes dades, s'ha realitzat un projecte amb caràcter prospectiu, amb el propòsit d'implementar una sèrie de models que ajudin a predir la mitjana de delictes per dies de cada mes en els diferents sectors policials de Girona. Aquests models serviran per ajudar a optimitzar les zones de patrullatge de la policia.

Per tant, aquest TFG constarà de diferents apartats on es recull tot el procés seguit per assolir els objectius i les necessitats de la policia. A més, s'acompanya, de dos informes que s'ha generat a petició de la policia, els quals complementen el treball i poden ser d'interès pels lectors d'aquest treball.

2. METODOLOGIA

En aquest apartat veurem la metodologia que s'ha seguit per a realitzar aquest projecte.

El treball consta de cinc fases, l'última de les quals queda fora de l'abast del TFG.

Primera fase

En primer lloc, es va realitzar una primera reunió amb la Policia Municipal de Girona i la Unitat Municipal d'Anàlisi Territorial (UMAT) per tal d'explicar la importància d'aplicar anàlisis avançades sobre els delictes que s'han registrat a la ciutat, al llarg dels darrers 23 anys; així com, de les persones detingudes dels últims 4 anys. Tanmateix, es va concretar els objectius principals que es volien assolir amb aquest treball. Aquests objectius es detallaran al següent apartat.

Segona fase

Un cop obtingudes les dades, es va fer un preprocessament d'aquestes. A l'apartat "5.1 Preprocessament de les dades" es podrà veure el procés que s'ha seguit en aquesta fase. Tanmateix, en aquesta fase es va fer una segona reunió amb la Policia Municipal i la UMAT on es va analitzar detalladament tota la informació de les dades. És a dir, es va explicar totes les variables susceptibles de ser d'estudiades i també es va donar algunes pinzellades sobre el procediment que segueix la policia a l'hora de registrar un delicte.

Tercera fase

Una vegada realitzat un primer processament de les dades, es començar a elaborar la part d'Anàlisi Exploratòria de les Dades (EDA, per les seves sigles en anglès). En aquesta etapa del projecte es dona resposta a les següents preguntes: Quina tipologia de delictes tenim enregistrats? Quan han ocorregut els delictes? On han ocorregut els delictes? Qui ha realitzat els delictes? I per què ocorren els delictes?

Per a analitzar els motius pels quals tenen lloc els delictes, s'ha fet servir variables externes al nostre joc de dades corresponents a la densitat i al total de població per barris de Girona i altres variables meteorològiques.

Quarta fase

En aquesta fase es realitza una tercera reunió on es fa seguiment del treball, i es resolen dubtes que han sorgit al llarg del treball.

Després de realitzar l'apartat d'EDA, es procedeix a l'etapa d'anàlisi predictiva de les dades. A partir del preprocessament, de l'EDA de les dades i de la recerca, s'ha arribat a la conclusió que els models més adequats per aquest projecte són algoritmes de sèries temporals. Així doncs, en veure que les sèries presenten certa estacionalitat, s'han prioritzat models que tinguin en compte aquests components, com per exemple

SARIMA, Holt-Winters i Prophet. També s'ha utilitzat ARIMA, que no té en compte la part estacional, però s'ha considerat interessant d'aplicar-lo al nostre projecte.

La part d'anàlisi predictiva s'ha dividit en dos grans blocs: en el primer, s'han realitzat un conjunt de proves amb els diferents models on s'ha dut a terme una predicció univariada per cadascun dels sectors policials. És a dir, s'ha tingut en compte una única variable; en el segon bloc, s'ha realitzat una predicció multivariada a escala municipal. És a dir, s'han utilitzat més variables predictores per enriquir el model.

Cinquena fase

Aquesta darrera fase ha quedat fora de l'abast del TFG per la seva extensió. Però dins del projecte com a tal, és fonamental per a la implementació i obtenció de resultats dels models predictius que farà servir la Policia Municipal.

En aquesta fase s'implementaran els algoritmes codificats en fitxers en format PY per tal que es puguin executar al sistema operatiu de la policia i així poder obtenir les dades predites de manera automàtica.

En conclusió, el que veurem en aquest treball és un recull de les quatre primeres fases. Primerament, en el marc teòric exposarem en quines referències ens hem basat per a realitzar totes les proves i sobre l'ús dels models. Seguidament, en l'apartat de preprocessament, es resumiran els aspectes més rellevants d'aquest apartat. A continuació, es veuran tots els resultats de l'anàlisi exploratòria de les dades i finalment veurem els resultats de l'etapa d'anàlisi predictiva. Tot el projecte s'ha realitzat a partir de dues taules: una taula amb registres sobre delictes i una segona taula amb el registre de persones detingudes. Ambdues taules tenen diferents franges temporals que explicarem més endavant.

3. OBJECTIUS

L'objectiu principal del treball és realitzar una anàlisi i una predicció temporal i espacial de la criminalitat a Girona. Per tant, és un estudi amb caràcter prospectiu, amb interès d'ajudar a la policia a obtenir informació predictiva sobre possibles delictes als diferents sectors policials.

Es desglossen aquest objectiu en diferents subobjectius:

- Generar tres informes on es reculli tota la informació que permeti tenir una visió general de la criminalitat a Girona i respondre les preguntes sobre el què, el quan, l'on, el qui i el perquè dels delictes.
  Al primer informe es farà el preprocessament de les dades; al segon s'exposarà tota la informació corresponent a l'exploració de les dades; i al tercer i últim informe es recolliran els resultats dels models predictius.
- Utilitzar models predictius per sèries temporals pels sectors policials de Girona (que es recolliran al tercer informe del primer subobjectiu). L'objecte de predicció correspon a la mitjana de delictes per dies de cada mes de l'any.
  Per aquest projecte, no es busca obtenir el model perfecte, sinó investigar i generar resultats primerencs que ajudin a futurs investigadors a obtenir informació base de partida. Tot i això, si els models donen bons resultats, es podran implementar al sistema de la policia per fer-se servir com a una eina complementària en la gestió de patrulles.

4. MARC TEÒRIC

Els Mossos d'Esquadra i la Policia Municipal de Girona han registrat durant aquests últims anys tots els delictes que han ocorregut a la ciutat. Des de la UMAT s'han realitzat diferents anàlisis sobre tots aquests registres, per així poder fer una gestió més efectiva de la seguretat de la ciutat. De fet, és necessari que es facin anàlisis sobre els delictes. Com bé diuen González-Álvarez, et al.:

> *"Els estudis sobre el crim han mostrat de manera consistent que es poden identificar regularitats en la manera d'operar dels criminals en diferents delictes. Així, per exemple, des de la psicologia i la sociologia ambiental, s'ha demostrat que els criminals tendeixen a actuar a llocs que coneixen (zona de confort) i no fan grans desplaçaments. D'altra banda, s'ha demostrat que la distribució dels crims no és homogènia, sinó que tendeix a concentrar-se en llocs i moments determinats en allò que es coneix com a hot spots"* (González-Álvarez, et al., 2020).

Així doncs, la policia predictiva suposa un canvi en el que respecta a la gestió sobre la reactivitat vers els delictes. La idea és que la prevenció i la predicció siguin les eines cabdals en aquest procés de transformació policial. Des de l'Organització per a la Seguretat i la Cooperació a Europa (OSCE) s'ha creat una guia on s'informa sobre com s'ha de realitzar l'Activitat Policial basada en la Intel·ligència (ILP, per les seves sigles en anglès). "L'enfocament proactiu i prospectiu de la ILP facilita la prevenció, la reducció, la repressió i la desarticulació de la delinqüència. Un element clau d'aquest enfocament és la recopilació i l'anàlisi sistemàtica de la informació i les dades pertinents per a la prevenció i la reducció de la delinqüència, seguida per l'elaboració d'informes d'intel·ligència" (OSCE, 2017).

S'ha de tenir en compte, però, que la policia predictiva només es pot aplicar a delictes que segueixen un patró. Aquells delictes que siguin situacionals o esporàdics no podran predir-se amb cert grau de precisió, tal com ens comenta Kaufman et al., a *Predictive policing and the polítics of paterns* el 2018.

Això últim és important, ja que les anomalies que trobem en les sèries temporals s'hauran de tenir en compte pels models predictius. De no fer-ho, la variabilitat de les dades pot ser massa gran perquè el model pugui tenir la capacitat d'entendre la distribució de les observacions.

Si ens fixem en el cas espanyol, la Secretaria de l'Estat de Seguretat (SES) ha desenvolupat el protocol de valoració policial sobre el risc que un autor dels fets sobre violència de gènere pugui ser reincident. Aquest protocol utilitza VioGen, el sistema de seguiment integral dels casos de violència de gènere, el qual permet pronosticar el risc de patir una nova agressió que pot tenir una dona que denuncia un fet. Si es continua analitzant el cas espanyol, ens podem adonar que els SIG juguen un paper rellevant en

aquesta lluita contra el crim. L'espai geogràfic i el temps són dues variables indispensables per ajudar a identificar patrons i concentracions de fets delictius.

Però, sobre quins models estem parlant? Quins són els utilitzats en un context més tècnic?

Per tal de concloure quins models poden ser els més adequats per aquest treball, s'ha realitzat una recerca de documents on s'hagin implementat models predictius corresponents a la criminalitat.

És interessant veure com en gairebé tots els treballs que s'han realitzat s'utilitzen models per sèries temporals. De fet, en un treball final integrador realitzat per Zambrano que es va realitzar a Buenos Aires, el 2020, es realitza un model de sèrie temporal de tipus *single step*. En aquest s'intenta predir el nombre de delictes per mes, tenint en compte cada cantonada de certs carrers d'un espai geogràfic. Pel que respecta al model utilitzat per sèries de temps, "s'utilitza el model *Croton*, ja que és un model especial per regressions de recompte, que té en compte únicament l'històric de l'ocurrència dels delictes" (Zambrano, R., 2020).

En altres treballs s'han utilitzat mètodes seqüencials de Monte Carlo, el qual "permet estimar una distribució a posteriori que implícitament incorpori la variable temporal" (Viviana, P., 2014). En aquest treball, per exemple, el que es vol aconseguir és una estimació de la distribució espacial del risc en una àrea en concret a partir d'aplicar "un model per barreges gaussianes (GMM), per així generar la caracterització probabilística del risc final" (Viviana, P., 2014).

Com es pot veure, la predicció del crim es pot treballar a partir de diferents models, però el que sempre es té present en aquests projectes és l'espai i el temps com a components indispensables per l'anàlisi i posterior predicció. De fet, tots els models que s'han utilitzat en aquest projecte contemplen aquests dos elements. A més a més, els models escollits són força utilitzats al món acadèmic. Per exemple l'estudi de Divya Sindhuri "*Time series analysis and forecasting of crime dat"* del 2019, on analitza i prediu els crims ocorreguts a Chicago i Los Ángeles, ho fa a partir de models de sèries temporals com ARIMA, Auto ARIMA i Facebook Prophet. ARIMA i Facebook Prophet (actualment conegut com a Prophet) són dos dels models que s'han implementat en aquest TFG.

Així com els models ARIMA i Auto ARIMA no contemplen l'estacionalitat, el model SARIMA sí que té en compte la part estacional de la sèrie temporal. De fet, a *Seasonal Autoregressive Integrated Moving Average Model for Crime Analysis in Daudi Arabia,* escrit per Noor, T., et al*.,* s'utilitza SARIMA per predir els crims a Aràbia Saurí. D'altra banda, a *A statistical study on the impact of the weather on cryme* de Robert Kane, es fa servir el model Holt Winters.

Com es veurà en aquest treball, les dades de la Policia Municipal de Girona mostren certa estacionalitat en els diferents sectors policials, i és per aquest fet que cal utilitzar

models que tinguin en compte l'estacionalitat de la sèrie, com per exemple SARIMA, Prohpet o Holt Winters. Tot i això, també s'utilitzarà ARIMA, que no contempla l'estacionalitat de la sèrie, per tal de tenir en compte aquelles sèries on no es vegi una estacionalitat molt marcada.

A continuació, es farà una introducció a les sèries temporals i es descriuran les proves estadístiques i els components d'una sèrie abans de procedir a veure els resultats del treball.

4.1 Introducció a les sèries temporals

"Les sèries temporals són col·leccions d'observacions que segueixen un ordre i tenen una forma seqüencial en el temps" (VIllavicencio, J., 2010). L'explotació d'aquestes sèries permet a l'investigador observar l'evolució d'un determinat valor en el temps i, d'aquesta manera, poder extraure un coneixement sobre la variable en qüestió de tipus històric i evolutiu. Aquesta seria la fórmula d'una sèrie temporal additiva:

$$X(t) = T(t) + S(t) + C(t) + E(t)$$

On:

- $T(t)$: Seria la tendència. Representa el patró general de creixement o decreixement de la sèrie a llarg termini.

- $S(t)$: Representa l'estacionalitat. És a dir, els patrons regulars i repetitius que ocorren en períodes fixos, com l'estacionalitat anual, trimestral, mensual, diària, etc.

- $C(t)$: Seria la component cíclica. Representa les fluctuacions que no són regulars que ocorren en períodes més llargs que l'estacionalitat.

- $E(t)$: Correspon a l'error o residu. Per tant, es refereix a la variació aleatòria o que no pot ser explicada pels altres components. En tractar-se d'un comportament sistemàtic o regular, aquest fet pot fer que empitjori la precisió en la predicció.

Pel que fa als components de la sèrie temporal que acabem de veure, els analitzarem amb més profunditat a l'apartat 4.2.

Així doncs, com que les dades són observacions que s'han agafat de manera seqüencial i tenen un ordre en el temps, s'ha cregut oportú fer ús dels algorismes per sèries temporals. És per aquest fet, que abans de seguir amb els diferents apartats, contextualitzarem aquest tema i seguirem veient què és una sèrie temporal, quins requisits previs ha d'haver-hi per treballar amb ella i quins models es poden fer servir per realitzar aquest estudi. En aquest cas, analitzarem els models que s'han utilitzat per aquest projecte.

Ja hem vist que una sèrie temporal és una evolució en el temps d'un valor o d'una sèrie de valors que s'han recollit de forma seqüencial i que es troben ordenats en el temps. Aquestes sèries poden tenir una periodicitat anual, semestral, trimestral, mensual, diària, etc. Com es podrà apreciar més endavant en la fase d'anàlisi exploratòria de les dades i en la fase predictiva, les sèries seran desglossades en diferents escales temporals, per així poder veure els patrons que s'hi donen. Tot i que les nostres dades estan recollides de manera diària, per l'etapa d'anàlisi predictiva s'utilitzarà una periodicitat mensual.

Les sèries temporals es poden agrupar segons si són univariants o multivariants:

- **Univariants**: Són aquelles on només hi ha una variable depenent en el temps. Això significa que només es recopila informació i es fa un seguiment d'una sola variable al llarg del temps. Un exemple podria ser la sèrie temporal de delictes mensuals de la ciutat de Girona. És a dir, tenim una variable, que en aquest cas correspon al nombre de delictes mensuals, que han estat registrats de manera seqüencial i mantenen un ordre en el temps.
- **Multivariants**: Són aquelles on es recopila i s'enregistra més d'una variable en el temps. Això implica que es prenen múltiples variables i s'observen i enregistren simultàniament en cada moment temporal. Un exemple en aquest cas seria un conjunt de variables, on tindríem una variable dependent que en aquest cas serien els delictes recollits mensualment; i unes variables predictores (o exògenes) que ajudarien a enriquir el procés predictiu i d'anàlisi de la sèrie temporal de delictes. Aquestes variables predictores podrien ser factors meteorològics, de densitat de població, nombre de població, etc. Dit d'una altra manera, la variable dependent és aquella que volem analitzar o predir i les predictores són les variables que s'utilitzen per explicar o predir les variacions en les variables dependents.

Hem de tenir en compte que les sèries multivariants ofereixen la possibilitat de capturar interaccions i relacions entre les variables en el context temporal. Això permet una anàlisi més completa i pot proporcionar una millor comprensió del comportament de les variables i les seves interrelacions. En canvi, les sèries univariants se centren exclusivament en una única variable, oferint una visió més estreta i específica de la seva evolució temporal.

Per aquest projecte ens hem centrat a fer prediccions per sèries temporals univariants amb els sectors policials. D'altra banda, pel conjunt del municipi de Girona s'han utilitzat altres variables predictores amb les quals s'ha pogut fer una predicció multivariant.

Dins del ventall de sèries temporals que ens podem trobar, hi ha dos grans grups que hem de tenir en compte prèviament abans de començar l'anàlisi. Aquests dos tipus de sèries els podem catalogar com:

- **Sèries estacionàries:** Són aquelles en què les seves propietats estadístiques no varien en el temps. Això significa que la mitjana, la variància i l'estructura de correlació de la sèrie temporal són constants al llarg del temps. Per tant, els patrons, tendències i cicles es mantenen constants i previsibles al llarg del temps.

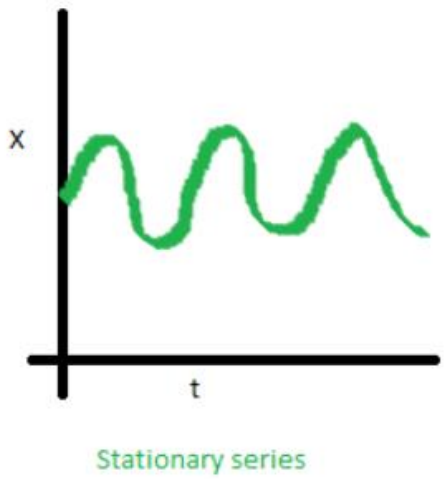

Figura 4.1: Exemple de sèrie estacionària. Citat per Duk2.

- **Sèries no estacionàries**: Són aquelles en què les propietats estadístiques varien en el temps. Això pot incloure canvis en la mitjana, la variància i l'estructura de correlació. Així doncs, els patrons, les tendències i els cicles poden canviar al llarg del temps i no es poden predir de manera consistent.

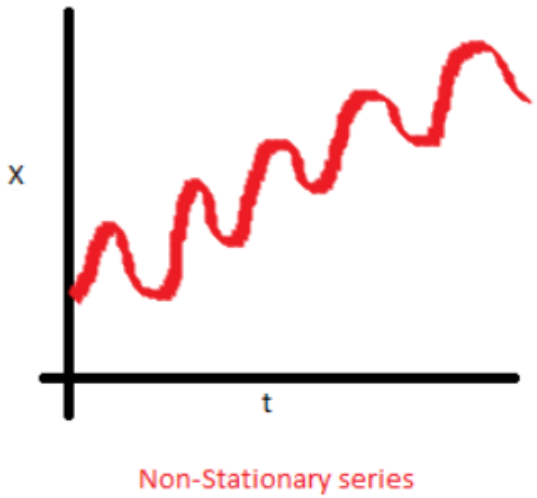

Figura 4.2: Exemple de sèrie no estacionària. Citat per Duk2.

Un aspecte que s'ha de tenir en compte a l'hora de realitzar un model per sèries temporals és la necessitat que aquesta disposi de mostres suficients per obtenir resultats robustos. Així doncs, veurem que per la realització dels models en la fase d'anàlisi predictiva, s'utilitzaran els sectors policials i l'àmbit municipal de la ciutat. No s'ha pogut implementar una predicció pels subsectors policials, ja que sinó el projecte s'estendria excessivament. Tampoc s'han utilitzat els barris oficials de Girona perquè no hi ha suficients observacions com per obtenir un model consistent. "De fet, la llargada de la sèrie temporal ha de ser, generalment, d'almenys 20 observacions" (McCleary et al., 1980, p. 20).

Per últim, s'ha de tenir en compte les diferents estratègies de validació que es poden dur a terme a l'hora de valorar la capacitat predictiva del model. Aquestes han estat extretes d'un article publicat per l'expert en ciència de dades Mario Filho a *How to do time series cross-validation in Python*. Aquests són els tipus de validació que existeixen:

- **Validació simple de fracció temporal**: Aquesta estratègia consisteix a dividir les dades en un conjunt d'entrenament i un conjunt de prova, deixant com a mínim el 50% de les dades per a l'entrenament. És un dels enfocaments més senzills, i en la majoria de casos pot funcionar amb bons resultats, però pot ser el menys robust en comparació amb altres estratègies.

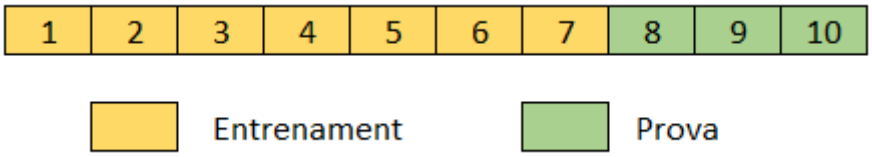

Figura 4.3: Exemple de validació simple de fracció temporal. Autor: Filho, M.

- **Validació de finestra corredissa (*Rolling window*)**: En aquesta validació, es mou una finestra temporal al llarg de la sèrie i es realitzen múltiples validacions. En cada iteració es recopilen dades històriques fins a un punt determinat i es realitza la predicció per al període següent. Aquest enfocament permet avaluacions repetides i successives de la precisió del model en diferents segments de la sèrie temporal.

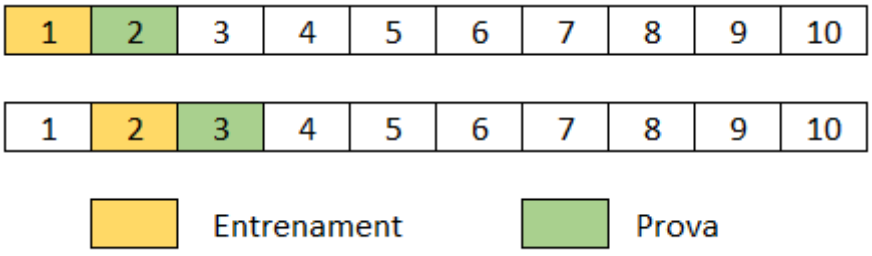

Figura 4.4: Exemple de validació de finestra corredissa. Autor: Filho, M.

- **Validació de la finestra d'ampliació (*Expanding window*)**: A diferència de la finestra corredissa, en aquesta validació es comença amb un petit conjunt

d'entrenament inicial i a mesura que s'avança en el temps, es va augmentant la mida del conjunt d'entrenament. Aquest enfocament permet avaluar el rendiment del model a mesura que es disposa de més dades disponibles.

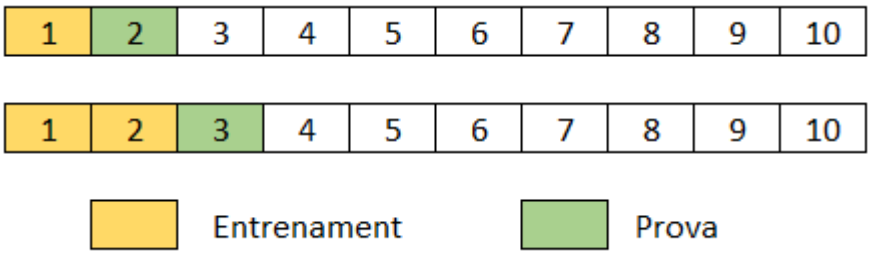

Figura 4.5: Exemple de validació de la finestra d'ampliació. Autor: Filho, M.

- **Validació per establir una bretxa entre formació i validació (*Gap walk-forward)***: Per aquesta estratègia, el que es fa és moure una finestra temporal al llarg de la sèrie, però deixant un espai buit o una bretxa entre la part d'entrenament i la part de validació. Aquest fet permet avaluar la capacitat del model per predir valors en períodes on no ha estat entrenat. Aquesta validació és de les més robustes que hi ha.

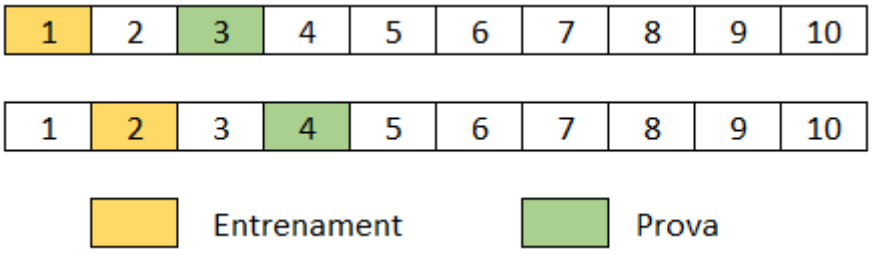

Figura 4.6: Exemple de validació per establir una bretxa entre formació i validació. Autor: Filho, M.

Per a la realització del projecte, s'ha utilitzat el primer mètode com a validació. Per tant, seria interessant que en posteriors projectes s'utilitzessin altres tipus de validacions i així poder ampliar l'anàlisi.

A continuació, s'explicarà quin és el principal objectiu d'una sèrie temporal.

4.1.1 Objectiu de la sèrie temporal

Un dels principals objectius de les sèries temporals és el de descriure els canvis en una variable analitzada en el temps, o de predir els seus valors futurs. Quan es tracta de fer prediccions, "podem seguir dues vies: el *single-step* i el *multi-step"* (Aragon, F., 2017). Aquests conceptes fan referència a dues formes diferents de realitzar prediccions per al futur.

En el *single-step* es realitza una predicció pel següent valor futur de la sèrie. Per exemple, volem predir quina és la mitjana de delictes diaris que es pot donar el mes següent

(futur). En aquest cas, s'agafaria tot el registre històric de mitjana de delictes i s'intentaria predir el mes següent.

Per la seva part, el *multi-step* es refereix a fer una predicció de múltiples punts temporals en el futur. Per tant, dividiríem el nostre registre històric de casos delictius, depenent del nombre de passos que volguéssim predir. S'ha de tenir en compte que en l'ús del *multi-step* es pot donar el cas que, com més lluny es vulgui predir, més incert i menys precisos seran els resultats.

Per últim, per tal de poder assolir l'objectiu que ens proposem amb la sèrie temporal, s'ha de complir una sèrie de requisits que influiran en la manera com tractem la nostra sèrie temporal. Així doncs, per tal de poder realitzar una predicció, la sèrie necessita certa correlació serial. És a dir, que hi hagi correlació entre el valor actual i el valor anterior. Aquesta correlació es pot extreure a partir de les gràfiques d'autocorrelació ACF i d'autocorrelació parcial PACF, que es veuran més endavant. A part de tenir en compte l'autocorrelació serial, també s'ha de capturar la tendència, el comportament estacional observat i l'heteroscedasticitat, entre d'altres. A mode de resum sintètic del que s'ha realitzat en aquest treball, aquests són tots aquests aspectes que s'han de treballar:

- La variabilitat de les dades
- La tendència
- Si la sèrie és additiva o multiplicativa
- L'estacionalitat, cicles
- L'autocorrelació
- Si és estacionària
- Si hi ha presència d'heteroscedasticitat
- La irregularitat, presència d'anomalies

A més a més, per l'apartat de l'anàlisi multivariant s'ha fet una selecció de variables que explicarem en l'apartat "6.3.2.1 Selecció de variables".

A continuació, explicarem cadascun dels elements que hem comentat anteriorment.

4.2 Components, proves estadístiques i indicadors

En aquest apartat detallarem els components i les proves que s'han realitzat prèviament a la fase d'anàlisi predictiva i els indicadors que ens han servit per veure la capacitat predictora del model. Així doncs, el que es pretén és que el lector es familiaritzi amb els processos que s'han dut a terme en aquest projecte i, més endavant, veure els resultats de cada component, de les proves estadístiques i dels indicadors.

Com s'ha comentat a l'apartat dels objectius, abans de la implementació dels models es realitzarà una anàlisi prèvia de les dades per tenir en compte els diferents components i altres valoracions. Comencem doncs, per detallar què entenem per variabilitat de les dades.

4.2.1 Variabilitat de les dades

Per veure la variabilitat de les dades utilitzarem el coeficient de variació (CV).

El CV en una sèrie temporal proporciona una mesura relativa de la dispersió o variabilitat en relació amb la mitjana de la sèrie. O, dit d'una altra manera, "la CV és una mesura que ens diu com d'agrupat és un conjunt de dades" (Sujit, S., 2015).

Aquest coeficient s'utilitza per avaluar la relativa estabilitat o consistència de les observacions en una sèrie temporal. El CV es calcula com la desviació estàndard dividida per la mitjana de la sèrie temporal, tot multiplicat per 100 per expressar el resultat en percentatge.

La fórmula per calcular el coeficient de variació és la següent:

$$CV = \left(\frac{\sigma}{\mu}\right) * 100$$

On:

- σ: Mesura la dispersió de les dades respecte a la mitjana. Indica la quantitat de variabilitat en les observacions.
- μ: És el valor mitjà de la sèrie temporal.

El CV proporciona una perspectiva relativa de la variabilitat en una sèrie temporal. Valors més baixos de CV indiquen una menor variabilitat relativa, mentre que valors més alts de CV indiquen una major variabilitat relativa.

En general, si el CV és baix (per exemple, inferior al 10%), indica que es tracta d'una sèrie temporal relativament estable i consistent, amb poca variabilitat entre les observacions. D'altra banda, si el CV és alt (per exemple, superior al 10%), indica una sèrie temporal més volàtil, amb una major variabilitat entre les observacions.

Hem de tenir en compte que aquest coeficient no té en compte específicament els pics estacionals en la sèrie temporal. El CV és una mesura de la variabilitat relativa en relació amb la mitjana de la sèrie i considera tota la variabilitat present en les dades, incloent-hi els pics estacionals si hi estan presents.

Tot i això, ens permet comparar la variabilitat relativa entre diferents sèries temporals.

4.2.2 Naturalesa additiva o multiplicativa

En aquest subapartat veurem la diferència entre una sèrie temporal additiva i una de multiplicativa, ja que aquestes fan referència a la forma com s'interpreten els components de la descomposició d'una sèrie temporal.

Un model additiu assumeix que els components de la sèrie (tendència, estacionalitat, residuals) s'ajunten de manera additiva per formar la sèrie total. Això significa que els efectes de la tendència, l'estacionalitat i els residuals s'agreguen de manera lineal per obtenir els valors observats de la sèrie temporal. En un model additiu, el canvi en un component no afecta la magnitud de l'altre component. Per exemple, si la tendència està augmentant, l'estacionalitat i els residuals es mantindran constants en termes absoluts. En altres paraules, a mesura que s'avança en el temps i el valor mig de la sèrie fluctua, la desviació típica es manté aproximadament en la mateixa distància.

Un model multiplicatiu, d'altra banda, assumeix que els components de la sèrie s'ajunten de manera multiplicativa. Això significa que els efectes de la tendència, l'estacionalitat i els residuals es multipliquen per formar la sèrie total. En un model multiplicatiu, els canvis en un component afecten la magnitud de l'altre component. Per exemple, si la tendència està augmentant, l'estacionalitat i els residuals també augmentaran proporcionalment. També ho podem entendre com la fluctuació de la mitjana en la sèrie ve acompanyada pel desplaçament de la desviació típica.

"Per poder saber si el model és additiu o multiplicatiu es pot fer la comparació dels coeficients de variació de les sèries diferència i quocient" (Aragon, F., 2017). Els passos per poder calcular-ho són els següents:

Primer es calcula la sèrie diferenciada (Dt):

$$Dt = y_t - y_t - 1$$

On:

y_t: És el valor de la sèrie actual.

A continuació, es calcula la sèrie quocient (Ct):

$$Ct = y_t / y_t - 1$$

Calculem el coeficient de variació per la sèrie diferenciada i del quocient:

$$CVDt\ o\ CVCt = \left(\frac{\sigma}{\mu}\right) * 100$$

Finalment, es conclou que:

- Si CVC < CVD, ens trobem davant d'un model multiplicatiu
- Si CVC > CVD, ens troben davant d'un model additiu

4.2.3 L'estacionalitat

L'estacionalitat en una sèrie temporal fa referència a un patró o cicle recurrent que es repeteix a intervals regulars. Sol ser d'una durada concreta i està associat a factors estacionals com ara les estacions de l'any, els dies de la setmana, els períodes de vacances, etc.

Quan una sèrie temporal té un component estacional, significa que hi ha canvis regulars i previsibles en les dades al llarg del temps. És a dir, hi ha "oscil·lacions que es produeixen amb un període inferior o igual a l'any" (UCM, 2013). Aquesta variabilitat pot ser anual, mensual, setmanal o diària, segons el patró temporal observat.

Aquesta estacionalitat pot manifestar-se com a pics i valls recurrents en les dades, amb valors més alts o més baixos en certs moments de cada cicle. Per tal d'analitzar l'estacionalitat en una sèrie es faran servir dues aproximacions: en una es farà una descomposició estacional i en l'altre es calcularà l'índex de variància estacional. Vegem doncs per veure què entenem per descomposició estacional.

4.2.3.1 Descomposició estacional per sectors

La descomposició estacional no és res més que descompondre la sèrie temporal per components. És a dir, separar pel component de tendència, d'estacionalitat i variabilitat cíclica i els residus. Com s'apreciarà més endavant en el subapartat "6.3.1.1 Resultats dels components i proves estadístiques", s'utilitzarà una funció anomenada "*seasonal_decompose()*" (Perktold, J., et al., 2023) que ajudarà a desglossar la sèrie pels diferents components. En ella podrem veure:

- **Gràfica de la sèrie original**: Mostra la sèrie temporal original sense cap modificació. Dóna una visió general del comportament general de la sèrie, incloent-hi les tendències, la variabilitat estacional i els residus.
- **Gràfica de la tendència**: Representa el component de tendència de la sèrie temporal. La tendència indica la direcció general dels canvis en la sèrie al llarg del temps. Aquesta gràfica ens permet identificar si hi ha una tendència creixent, decreixent o constant en la sèrie.
- **Gràfica del component estacional**: Mostra la variabilitat estacional en la sèrie. Aquesta gràfica ens permet identificar els patrons repetitius o cíclics que es repeteixen al llarg del temps. Ajuda a comprendre les fluctuacions sistemàtiques que es produeixen en períodes específics de l'any.
- **Gràfica dels residus**: Representa els errors residuals, és a dir, les diferències entre els valors observats i els valors predits pels components de tendència i

estacionalitat. Aquesta gràfica ens permet identificar si hi ha alguna estructura no capturada pels components anteriors, com soroll o irregularitats aleatòries.

4.2.3.2 Índex de Variància Estacional

Una altra tècnica per veure l'estacionalitat de la sèrie temporal és utilitzar l'Índex de Variància Estacional (IVE).

L'IVE és una mesura utilitzada per quantificar la magnitud de la variabilitat estacional en una sèrie temporal. S'utilitza per avaluar quina proporció de la variabilitat total de la sèrie temporal està relacionada amb l'estacionalitat. Si l'índex és alt, indica que una part significativa de la variabilitat total és atribuïble al component estacional. D'altra banda, si l'índex és baix, indica que la variabilitat estacional té un impacte menor en la sèrie temporal.

Per calcular l'IVE s'han de realitzar aquests passos:

En primer lloc, s'elimina la tendència. Per fer això, primer es calcula la tendència mitjançant mitjanes mòbils centrades. En aquest treball, s'ha utilitzat una finestra de mida dos (o d'ordre dos) per calcular la mitjana mòbil centrada, tal com s'ha realitzat en els apunts de temari de "sèries temporals" (Martínez, R., s.d.).

$$t\text{endencia}_{\text{t}} = \frac{1}{2} \times (\text{mean}DelicXmontht - 1 + \text{mean}DelicXmontht)$$

I després s'elimina dividint les dades originals per la tendència.

$$serie\_no\_t\text{endencia}_{\text{t}} = \frac{meanDelicXmonth_{\text{t}}}{tendencia_{\text{t}}}$$

A continuació, eliminem els valors irregulars fent la mitjana dels valors calculats anteriorment.

$$\%\ avg\_mensual_{\text{m}} = \frac{(\sum serie\_no\_tendencia_{\text{t}})}{N_{\text{m}}}$$

$(\sum serie\_no\_tendencia_{\text{t}})$: és la suma dels valors de la sèrie sense tendència per cada temps $_{\text{t}}$ en el mes $_{\text{m}}$, i N_{m} és el nombre de mostres en el mes $_{\text{m}}$.

Per acabar d'entendre l'IVE, s'inclou un exemple a continuació:

Suposem que pel mes de febrer tenim un índex de 0,30. Aquest valor suggereix que la variació estacional en aquest mes és relativament baixa. Indica que la influència dels factors estacionals és menor en comparació amb altres mesos. Els canvis en les condicions estacionals tenen un impacte limitat en aquest mes en particular pel sector analitzat.

Suposem ara que en el mes d'octubre tenim un índex d'1,30. Aquest valor indica que la variació estacional en aquest mes és relativament alta. És a dir, ens diu que la influència dels factors estacionals en aquest mes és significativa i que pot tenir un impacte considerable en el comportament o rendiment del sector. Aquest mes en particular mostra una presència destacada de patrons estacionals que influeixen en els valors observats. Si multipliquéssim el resultat per 100 tindríem que hi ha hagut un 130% dels delictes en el mes d'octubre en relació amb un mes estàndard (que seria igual a 100%).

4.2.4 Autocorrelació ACF i PACF

La correlació serial en una sèrie temporal fa referència a la relació entre les observacions successives en el temps. Indica si hi ha alguna relació o dependència entre els valors passats i els valors futurs de la sèrie temporal.

Quan una sèrie temporal té una correlació serial significativa, vol dir que els valors successius de la sèrie estan relacionats entre si i que les observacions anteriors poden proporcionar informació útil per a predir els valors futurs.

Això implica que hi ha una estructura de dependència temporal a la sèrie. La correlació serial es pot mesurar utilitzant diferents mètriques com l'autocorrelació o la funció d'autocorrelació parcial. L'autocorrelació mesura la correlació entre una observació i una o diverses observacions anteriors, mentre que la funció d'autocorrelació parcial mesura la correlació entre una observació i les observacions anteriors tenint en compte les correlacions parcials amb les observacions intermèdies.

En aquesta part introduirem el concepte de *lag*, que es refereix al nombre de períodes de temps que s'endarrereix una sèrie respecte a una altra sèrie. És a dir, es refereix al nombre de punts anteriors que s'utilitzen per predir el valor actual de la sèrie. Per tal d'analitzar l'autocorrelació es poden utilitzar les gràfiques de les dues funcions més comunes: l'*Autocorrelation Function* (ACF) i la *Partial Autocorrelation Function (*PACF).

- **ACF**: En la gràfica ACF s'ha de prestar atenció en els valors que sobresurten de les línies horitzontals que marquen l'interval de confiança. Per tant, aquestes "barreres" signifiquen que hi ha una correlació significativa en aquell *lag*. I, per tant, es pot tenir en compte aquell *lag* pel model. En general, com més lluny en el temps, menys significança hi ha en els *lags* i per aquest fet se sol agafar els primers *lags* ja que acostumen a ser els més importants. Per entendre millor què significa cada *lag* vegem el següent:

$corr(y_t, y_{t-1})$ = Correspondria a la correlació entre el primer *lag* i el segon.

$corr(y_t, y_{t-2})$ = Correspondria a la correlació entre el primer *lag* i el tercer.

> "Els valors poden anar de -1 a 1. Un valor pròxim a 1 indica una correlació forta entre intervals i, per tant, els valors del dia en qüestió pugen seguint la tendència del dia anterior. Mentre que si el valor és negatiu la correlació és a la inversa. És a dir, els valors d'avui pugen quan els d'ahir anaven a la baixa." (Villalba, R., 2020).

Hem de tenir en compte que a partir de l'observació d'ACF podrem saber *p*, la part autoregressiva no estacional (*p*) o estacional (*P*) dels models ARIMA i SARIMA. On per ARIMA, *p*=1 indicarà que es té en compte el valor de la sèrie actual i que es basa en el valor anterior. Si *p*=2, el valor actual es basa en els dos valors anteriors, i així successivament. En el cas del model SARIMA, *P*=1 indicarà que es té en compte el valor actual, que es basa en el valor del mateix mes (en cas que la sèrie sigui mensual) de l'any anterior. Si *P*=2, es tindrà en compte el valor actual, que es basa en el valor dels dos mesos anteriors de l'any anterior.

- **PACF**: Els resultats de PACF indiquen el mateix que ACF però sense tenir en compte la influència dels intervals intermedis. És a dir:

$corr(y_t, y_{t-2})$ = Correspondria a la correlació entre el primer *lag* i el tercer, sense tenir en compte y_{t-1}.

Amb PACF podem esbrinar la mitjana mòbil (MA) dels models ARIMA i SARIMA. En el model ARIMA, *q*=1 indicarà que es té en compte el valor actual i la influència de l'error del valor anterior. En canvi, si s'utilitza SARIMA i tenim *Q*=1, es tindrà en compte el valor actual i la influència de l'error del mes de l'any anterior.

4.2.5 Estacionarietat: Prova estadística de Dickey Fuller augmentat

L'algoritme de Dickey-Fuller és una prova estadística que es basa en un model autoregressiu (AR) "per determinar si una sèrie temporal amb correlació serial té o no una arrel unitària" (Perktold, J., et al., 2023), la qual cosa indica si la sèrie és estacionària o no. És a dir, si les propietats estadístiques com la mitjana, la variància i l'estructura de correlació, no varien en el temps

Per tal d'aplicar aquest algorisme podem utilitzar la funció *adfuller()*. La informació d'aquest test es pot estructurar amb els següents indicadors:

- **ValorP**: És el p-valor associat a la prova d'hipòtesi que la sèrie no és estacionària. Si aquest valor és menor que un nivell de significança predeterminat (generalment 0,05), es rebutja la hipòtesi nul·la de no estacionarietat i es conclou que la sèrie és estacionària. D'altra banda, si el valor *p* és més gran que el nivell de significança, no es pot rebutjar la hipòtesi nul·la de no estacionarietat i es conclou que la sèrie és no estacionària.

- **lags**: Nombre de retards utilitzats en la regressió de la prova de Dickey-Fuller augmentada. Aquesta regressió es fa servir per determinar si hi ha una relació de dependència entre els valors de la sèrie en diferents moments en el temps.
- **N_observacions**: Nombre d'observacions (punts de dades) utilitzats a la prova.
- **1%, 5% i 10%**: Valors crítics associats amb diferents nivells de significança (99%, 95% i 90% respectivament) per a la prova d'hipòtesis. Si l'estadística de prova és menor que aquests valors crítics, es pot rebutjar la hipòtesi nul·la de no estacionarietat. En altres paraules, si l'estadística de prova és més baixa que el valor crític, la sèrie es considera estacionària amb un nivell de confiança determinat.
- **CoefRM**: Coeficient de regressió que s'utilitza a la regressió de la prova de Dickey-Fuller augmentada. Aquesta regressió es fa servir per determinar si hi ha una relació de dependència entre els valors de la sèrie en diferents moments en el temps. El coeficient de regressió indica la taxa en què els valors de la sèrie s'ajusten a un valor d'equilibri a llarg termini. Si aquest coeficient és negatiu i significatiu, indica que la sèrie està convergint un valor d'equilibri i, per tant, es considera estacionària.

4.2.6 Heteroscedasticitat: Resultats de Breush Pagan i White

La prova de Breusch-Pagan i la prova de White són dues proves estadístiques utilitzades per avaluar l'heteroscedasticitat en un model. Hem d'entendre "l'heteroscedasticitat com a l'existència de variància no constant en els residuals d'un model" (Pedrosa, J., 2020). Indica que la variabilitat dels errors residuals canvia en funció dels valors de les variables independents. Dit d'una altra manera, els residuals tendeixen a tenir una variabilitat més gran en certes regions de l'espai de les variables independents i una variabilitat més petita en altres regions. Això pot afectar la precisió de les estimacions i les inferències realitzades a partir del model. Aquest fet pot distorsionar els resultats i, per això, és important saber si la nostra sèrie té heteroscedasticitat. Vegem dues imatges.

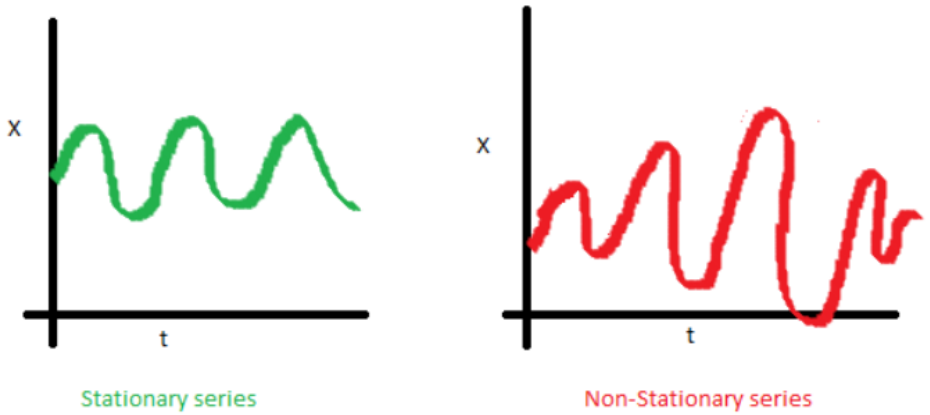

Figura 4.7: Exemple d'homoscedasticitat a la gràfica de l'esquerra. I exemple d'heteroscedasticitat a la gràfica de la dreta. Citat per Duck2.

A la figura 4.7 podem veure com en la sèrie de l'esquerra no hi ha presència d'heteroscedasticitat, per tant, podem dir que hi ha presència d'homoscedasticitat. D'altra banda, la sèrie de l'esquerra presenta clarament heteroscedasticitat, atès que la variància va augmentant i disminuint al llarg de la sèrie temporal. A continuació, parlarem sobre les proves més comunes per detectar la presència d'heteroscedasticitat.

La prova de Breusch-Pagan és una prova que avalua si la variància és constant i si hi ha una relació sistemàtica entre els residus d'un model i una o més variables explicatives.

El procediment general de la prova de Breusch-Pagan és ajustar un model de regressió, obtenir els residus, realitzar una regressió dels residus al quadrat en les variables explicatives i obtenir un estadístic de prova. "Si el valor de la prova és significativament diferent de zero i menor que 0,05, indica que hi ha evidència d'heteroscedasticitat" (Statalogos, 2021).

D'altra banda, **la prova de White**, també coneguda com a prova d'homoscedasticitat generalitzada, és una extensió de la prova de Breusch-Pagan. Aquesta prova té en compte la possibilitat de correlació entre els residuals d'un model, amb la qual estima una regressió auxiliar dels quadrats dels residuals en totes les variables explicatives i les seves interaccions.

A continuació, s'utilitzen els quadrats dels residus ajustats com a variables dependents en un model de regressió auxiliar. Els coeficients d'aquesta regressió auxiliar es fan servir per calcular un estadístic de prova, i si aquest és significativament diferent de zero i menor que 0,05, indica la presència d'heteroscedasticitat.

Hem de considerar que la prova de White pot ser més robusta en casos en què hi ha heteroscedasticitat i correlació entre els residuals. És a dir, "s'utilitza per provar els errors heteroscedàstics en l'anàlisi de regressió" (Benites, L., 2021) i, per això, s'ha cregut pertinent utilitzar les dues proves.

En general, si el valor *p* associat a la prova de Breusch-Pagan és menor que el nivell de significació (per exemple, 0,05), aleshores hi ha evidència d'heteroscedasticitat a la mostra. Si el valor *p* és més gran que el nivell de significació, llavors no hi ha prou evidència per concloure que hi ha heteroscedasticitat a la mostra.

Per últim, cal puntualitzar que aquestes proves d'hipòtesis estadístiques no poden concloure si una hipòtesi és certa o falsa de manera definitiva. Els resultats d'una prova d'hipòtesi només indiquen si hi ha prou evidència per rebutjar la hipòtesi nul·la o no.

En el nostre cas tenim que:

- **Hipòtesi nul·la (H 0)**: L'homoscedasticitat és present. És a dir, els residus es distribueixen amb la mateixa variància.

- **Hipòtesi alternativa (H A)**: L'heteroscedasticitat és present. Per tant, els residus no es distribueixen amb la mateixa variància.

En cas que es necessiti més endavant reduir la presència d'heteroscedasticitat, es poden provar aquestes opcions:

- **Transformació logarítmica**

```
np.log(sectorx['meanDelicXmonth'])
```

- **Transformació quadràtica**

```
np.power(sectorx['meanDelicXmonth'], 2)
```

Convertim les dades a l'escala quadràtica x^2 .

- **Transformació de boxcox**

```
boxcox(sectorx['meanDelicXmonth'])[0]
```

S'utilitza per aconseguir una distribució més semblant a la normalitat en les dades. Les convertim amb $x^{(\lambda)}$ on:

- λ: És el paràmetre de la transformació que cal determinar. Es tracta d'un valor que pot prendre diferents valors i s'ajusta per obtenir la millor aproximació a una distribució normal. En aquest cas l'algorisme de boxcox ja ho determina automàticament.

4.2.7 Anomalies en les dades

Les anomalies en les sèries temporals poden indicar dues coses: o bé que hi ha hagut un error en registrar la dada i que, per tant, s'ha de solucionar; o que es tracta d'un esdeveniment caòtic que trenca la sèrie com, per exemple, la COVID.

Per tal de localitzar els valors atípics, hem utilitzat *IsolationForest*. A continuació, s'exposen els passos més importants per utilitzar aquest model.

- Estandarditzem les dades del sector utilitzant el ***StandardScaler***. Les dades es transformen de manera que tinguin una mitjana de 0 i una desviació estàndard d'1.
- Creem un model d'***IsolationForest*** per detectar anomalies en les dades. El paràmetre ***contamination*** s'estableix amb la fracció d'anomalies que es vol detectar. Després ajustem el model d'*IsolationForest* utilitzant les dades estandarditzades i fem una predicció d'aquells valors atípics.

És important entendre que l'algoritme *IsolationForest* és un mètode popular per a la detecció d'anomalies en dades, incloent-hi sèries temporals. S'utilitza per identificar valors atípics o anòmals en un conjunt de dades "basant-se en el principi que les anomalies són menys freqüents i se separen més fàcilment que les instàncies normals" (Meng, S., 2022).

4.2.8 Indicadors estadístics dels models

En aquest darrer apartat parlarem dels indicadors que s'han utilitzat per veure com d'adient és un model. Cal remarcar, que d'indicadors n'hi ha molts més, però hem seleccionat aquells més populars que hem anat veient en tots els treballs relacionats amb prediccions de sèries temporals.

AIC = $-2 * log(L) + 2 * k$

On:

- $log(L)$: És el logaritme de la funció de versemblança màxima del model (likelihood). La funció de versemblança (likelihood function) és una funció que descriu la relació entre els paràmetres d'un model estadístic i les dades observades.
- k: És el nombre de paràmetres lliures del model.

Akaike's Information Criterion és un criteri estadístic utilitzat per comparar i seleccionar models estadístics. Mesura la qualitat relativa d'un model en funció de la seva capacitat de fer un bon ajustament de les dades i de la seva simplicitat. L'AIC té en compte tant la capacitat predictiva del model com la seva capacitat de descriure les dades amb el menor nombre de paràmetres possible. L'objectiu és trobar el model amb l'AIC més baix, ja que aquest paràmetre indica un bon equilibri entre l'ajustament de les dades i la simplicitat del model.

MSE = $\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$
On:

- n: És el nombre d'observacions o casos.
- y_i: Representa el valor observat de la variable de resposta per a l'observació i.
- $\bar{y}$: És el valor mitjà observat de la variable de resposta.

L'Error Quadràtic Mitjà (o MSE, per les seves sigles en anglès, *Mean Squared Error*) és una mesura de l'error mitjà al quadrat entre els valors pronosticats pel model i els valors reals observats.

Calcula la diferència al quadrat entre cada valor pronosticat y_i i el valor real $\bar{y}$ per a cada observació, suma totes aquestes diferències al quadrat i les divideix pel nombre d'observacions n.

Aquesta mesura de l'error ens dóna una idea de quant s'allunya el model dels valors reals, on valors més alts de l'MSE indiquen un pitjor ajustament del model.

MAE = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \overline{y}|$
On:

- n: És el nombre d'observacions o casos.
- y_i: Representa el valor observat de la variable de resposta per a l'observació i.
- $\bar{y}$: És el valor mitjà observat de la variable de resposta.
- $||$: Indica el valor absolut.

L'Error Absolut Mitjà (o MAE, per les sigles en anglès, *Mean Absolute Error*) calcula la diferència absoluta entre cada valor pronosticat y_i i el valor real $\bar{y}$ per a cada observació, suma totes aquestes diferències absolutes i les divideix pel nombre d'observacions n. Aquesta mesura de l'error ens dóna una idea de la magnitud de l'error entre el model i les dades observades, sense tenir en compte la direcció de l'error.

L'ús del MAE és similar al MSE, ja que proporciona una mesura quantitativa de l'error entre el model i les dades observades. No obstant, el MAE té avantatges en casos on és important evitar l'amplificació de grans errors, ja que no es calculen els quadrats de les diferències. El MAE és més robust a valors atípics o valors extrems que podrien influir en el MSE. Tot i això, el MSE pot ajudar en el sentit que és més sensible als errors grans. És a dir, si tenim especial preocupació per la influència dels errors grans a l'anàlisi, el MSE pot ser més útil. Com que el MSE eleva els errors al quadrat, els errors més grans tindran un impacte més significatiu en la mesura global de l'error. Això pot ser desitjable en certes circumstàncies, com ara en problemes on els grans errors tenen conseqüències més importants o són més crítics.

RMSE = $sqrt(MSE)$

És una versió modificada de l'Error Quadràtic Mitjà (MSE) que té l'avantatge d'estar en la mateixa escala que la variable de la resposta original. Per tant, podem entendre l'RMSE com la dispersió entre les prediccions i els valors reals o, dit d'una altra manera, "l'RMSE com la desviació estàndard de la variància inexplicada" (sitiobigdata, 2018).

MAPE = $\text{mean}\left(\left|\frac{y-\hat{y}}{y}\right|\right) \times 100$
On:

- y: Representa els valors reals de la variable de resposta.
- $\hat{y}$: Representa els valors pronosticats pel model.
- $|x|$: Representa el valor absolut de x.

El *Mean Absolute Percentage Error* (MAPE) calcula l'error mitjà com a percentatge de la variable de la resposta original. En aquest projecte, considerarem que un MAPE és positiu quan el resultat sigui igual o inferior a 0,10, ja que indicarà que en el que respecta la mitjana, els pronòstics difereixen en un 10% o menys dels valors reals.

CRMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$

On:

- n: És el nombre de mostres.
- y_i: És el valor real de la mostra i.
- $\overline{y}$: És la mitjana dels valors reals.

Centered Root Mean Squared Error (CRMSE) és una mesura de l'error quadràtic mitjà que considera la diferència entre les prediccions del model i els valors reals, centrant les dades al voltant de la mitjana.

En primer lloc, es resta la mitjana dels valors reals (*y_bar*) a cada valor real (*y_i*). Aquesta diferència centra les dades respecte a la mitjana. A continuació, s'eleven al quadrat les diferències centrades, es calcula la mitjana d'aquests quadrats i finalment se'n fa l'arrel quadrada per obtenir el CRMSE.

El CRMSE és útil per avaluar l'error quadràtic mitjà tenint en compte la desviació de les prediccions respecte a la mitjana dels valors reals. Això pot ser útil per mesurar com de bé s'ajusta el model a les dades en relació amb la mitjana i per identificar possibles resultats esbiaixats en les prediccions. “Si el resultat del CRMSE es troba per sota de 100, vol dir que el rendiment de l’equació és acceptable” (Ermeydan, I., et al., 2022).

R2 = $1 - \left(\frac{SSR}{SST}\right)$

On:

- SSR: és la suma dels quadrats dels residus.
- SST: és la suma total dels quadrats.

El coeficient de determinació (*R2*) és una mesura estadística que indica la proporció de la variància de la variable dependent que és explicada pel model de regressió. Prendrà valors entre 0 i 1, on un valor més proper a 1 indica un millor ajust del model. És important entendre que el coeficient de determinació és rarament utilitzat en sèries temporals. Això és degut a que normalment es treballa amb sèries no estacionàries i pot

provocar que *R2* es distorsioni i no mostri un resultat fiable. És per aquest fet "que *R2* per sèries estacionàries és preferible a l'ordinari *R2*" (Hassouna, F., 2020) quan la sèrie no és estacionària.

Valor_real: És el valor real de la mostra.

Valor_predit: És el valor predit pel model.

Error_relatiu= $\left|\frac{y-\hat{y}}{y}\right|$

On:

- y: És el valor real.
- $\hat{y}$: És el valor predit per al model.
- $||$: Indica el valor absolut.

L'error relatiu és una mesura que permet quantificar la diferència relativa entre el valor real i el valor predit pel model. Un valor d'error relatiu proper a zero indica un bon ajustament entre el valor real i el valor predit, mentre que un valor més alt indica una diferència més gran entre aquests valors.

Error Relatiu Acumulat = $\sum_{i=1}^{n}\left|\frac{y_i-\hat{y_i}}{y_i}\right|$

On:

- n: És el nombre de mostres.
- y_i: És el valor real de la mostra i.
- $\hat{y_i}$: És el valor predit per al model per a la mostra i.
- $\sum$: Indica la suma acumulada dels valors absoluts de l'error relatiu.
- $||$: Indica el valor absolut.

L'error relatiu acumulat permet quantificar l'error relatiu per a totes les mostres en conjunt. Un valor d'error relatiu acumulat proper a zero indica un bon ajustament global entre els valors reals i els valors predits pel model, mentre que un valor més alt indica una diferència més gran en conjunt.

4.3 Models implementats

En aquest subapartat presentarem els models utilitzats en el projecte. Tots els models són exclusius per realitzar prediccions per sèries temporals. Així doncs, el primer que veurem serà el model SARIMA, i continuarem amb ARIMA, Holt Winter i finalitzarem amb Prophet.

4.3.1 SARIMA

SARIMA (*Seasonal Autoregressive Integrated Moving Average*) és una extensió del model ARIMA que incorpora components estacionals i variables exògenes per capturar patrons més complexos presents a les dades. Tota la informació corresponent als paràmetres que es veuran a continuació, s'han extret de la seva documentació oficial "*statsmodels.tsa.statespace.sarimax.SARIMAX*" (Perktold, J., et al., 2023).

Una característica crucial del model SARIMA és que permet predir la part estacional de les dades. Aquesta estacionalitat es capta per *P, D, Q* i *m*, on:

- ***P***: És el valor que mostra les transformacions que necessiten les dades per convertir-les en estacionàries. Aquest valor s'aconsegueix a partir del valor màxim de PACF que ja hem comentat anteriorment, a l'apartat "4.2.4 Autocorrelació ACF i PACF".

- ***D***: És el grau d'integració estacional. Indica el nombre de vegades que s'ha de diferenciar la sèrie temporal de forma estacional per aconseguir l'estacionarietat.

- ***Q***: És l'ordre de la mitjana mòbil estacional. Determina el nombre de valors retardats dels errors de pronòstic en el component de mitjana mòbil estacional del model SARIMA. Aquest valor es pot extreure dels resultats de la gràfica ACF.

- ***m***: És la periodicitat del component estacional. Representa el nombre de períodes en una temporada. Per exemple, si les dades són trimestrals, *m* s'establiria en 4. Si les dades són mensuals, *m* s'establiria en 12.

A part d'aquests paràmetres també tenim l'ordre no estacional:

- ***p***: És l'ordre autoregressiu. Indica el nombre de valors retardats de la variable dependent en el component autoregressiu del model SARIMA. El valor de *p* es pot extreure a partir de la interpretació de la gràfica ACF.

- ***d***: És el grau d'integració. Determina el nombre de vegades que s'ha de diferenciar la sèrie temporal per obtenir l'estacionarietat.

- ***q***: És l'ordre de la mitjana mòbil. Especifica el nombre de valors retardats dels errors de pronòstic en el component de mitjana mòbil del model SARIMA. El valor de *q* el podem extreure a partir de la informació que ens aporta PACF.

Aquests paràmetres són components importants del model SARIMA per predir les dependències de la sèrie temporal. El component autoregressiu (*p*) captura les relacions

entre els valors passats de la sèrie temporal i els valors actuals, mentre que la mitjana mòbil (*q*) captura les dependències dels errors de pronòstic.

Alguns dels paràmetres que més utilitzaren en el model SARIMA són els següents:

- **endog**: Representa el procés de la sèrie temporal observada, anomenat *y*. Aquest paràmetre conté les dades històriques que es volen predir.
- **exog** (opcional): És una matriu de regressors exògens. Les variables exògenes són factors addicionals que poden influir en la sèrie temporal i formen part dels components SARIMA, a diferència d'altres algorismes que són univariants.
- **order(p, d, q)**: Especifica els components estacionaris del model SARIMA. Es pot definir com a iterador o com a iterador d'iteradors. El paràmetre d'ordre consta de tres valors: (*p, d, q*).
- **seasonal_order(P, D, Q, m)**: Especifica els components estacionals del model SARIMA.
- **freq**: La freqüència temporal de les dades.
- **maxiter**: Màxim d'iteracions.
- **enforce_stationarity**: És un booleà que indica si cal transformar els paràmetres autoregressius per garantir l'estacionarietat en el component autoregressiu del model. Per defecte, s'estableix en *True*, realitzant la transformació per garantir l'estacionarietat.
- **enforce_invertibility**: És un booleà que indica si cal transformar els paràmetres de mitjana mòbil per garantir la invertibilitat en el component de mitjana mòbil del model. Per defecte, s'estableix en *True*, realitzant la transformació per garantir la invertibilitat. Per garantir la invertibilitat, ens referim a la capacitat del model per aconseguir que els errors residuals tinguin una estructura aleatòria i no hi hagi cap dependència sistemàtica. És necessari tenir en compte aquest element, ja que en cas contrari ens pot portar a prediccions poc fiables.

4.3.2 ARIMA

ARIMA (*Autoregressive Integrated Moving Average*) és una extensió del model ARMA (*Autoregressive Moving Average*), que incorpora un component d'integració per tractar amb sèries no estacionàries. Hem de tenir en compte que el model ARIMA no contempla la part estacional i tampoc se li poden afegir variables predictores. Tota la documentació

del model ARIMA l'hem extret de "*statsmodels.tsa.arima.model.ARIMA*" (Perktold, J., et al., 2023).

El model ARIMA combina tres components principals, que ja hem vist anteriorment amb SARIMA:

Component autoregressiu (AR): Aquest component utilitza valors retardats de la sèrie temporal per predir valors futurs. Cada valor de la sèrie temporal s'expressa com una combinació lineal dels valors retardats anteriors, on els coeficients són els paràmetres del model. L'ordre autoregressiu **(p)** determina el nombre de valors retardats que s'han d'incloure al model. En aquest cas, *p* es pot extraure a partir de la gràfica PACF.

Component d'integració (I): Aquest component s'utilitza per transformar una sèrie temporal no estacionària en una sèrie estacionària. Es realitza una diferenciació de la sèrie temporal per eliminar les tendències i els patrons no desitjats. L'ordre d'integració **(d)** indica el nombre de vegades que cal diferenciar la sèrie per aconseguir l'estacionarietat.

Component de mitjana mòbil (MA): Aquest component utilitza els errors de pronòstic anteriors per predir valors futurs. S'assumeix que els errors estan correlacionats i s'utilitzen en combinació amb els coeficients del model per calcular les prediccions. L'ordre de la mitjana mòbil **(q)** especifica el nombre d'errors de pronòstic retardats que s'han d'incloure en el model. Aquest valor es pot aconseguir a partir de la gràfica PACF.

Així doncs, a través d'anàlisis exploratòries, com ara gràfics de la sèrie, funcions d'autocorrelació i funcions d'autocorrelació parcial, s'intenta comprendre les correlacions i dependències temporals presents a les dades per seleccionar els paràmetres òptims.

Pel que fa als paràmetres d'aquest model, són força semblants a SARIMA, tret del *seasonal_order()*, que no hi és.

4.3.3 Model de Holt Winters

El model de Holt Winters es coneix com a triple suavització exponencial i és una extensió del model de suavització exponencial simple que afegeix components de tendència i estacionalitat. Tota la documentació referent a Holt Winters la podem trobar a "*statsmodels.tsa.holtwinters.ExponentialSmoothingk"* (Perktold, J., et al., 2023).

Aquest model s'utilitza principalment per a pronòstics a curt termini i és adequat per a sèries temporals amb patrons clars de tendència i estacionalitat. Per tant, si veiem que tenim sectors amb tendència i estacionalitat, pot ser adequat.

El mètode es basa en tres components:

- **Component de nivell (*nivel*)**: Representa el valor mitjà de la sèrie temporal sense tenir en compte la tendència o els efectes estacionals. Aquest component s'actualitza a cada pas de temps amb una funció de suavització exponencial.
- **Component de tendència (*trend*)**: Captura la direcció i l'evolució de la tendència en la sèrie temporal. Aquest component es calcula mitjançant una funció de suavització exponencial per estimar el canvi de nivell al llarg del temps.
- **Component estacional (*seasonal*)**: Representa els efectes estacionals o cíclics presents a la sèrie temporal. Aquest component té en compte els patrons repetitius en un període determinat i s'actualitza amb una funció de suavització exponencial.

Alguns dels paràmetres que més s'utilitzaran són els següents:

- **error**: Indica el tipus d'error.
- **seasonal_periods**: Fa referència al nombre de períodes en una temporada. En cas que posem 12, per exemple, significa que les dades tenen un patró estacional anual.
- **trend**: Especifica el tipus de model de tendència.
- **seasonal**: Indica el tipus de model estacional.

4.3.4 Prophet

Prophet és una llibreria que ha estat desenvolupada per Facebook. Com les anteriors llibreries que hem presentat, Prophet serveix també per l'anàlisi i predicció de dades temporals. Tota la documentació de Prophet, l'hem extret de la seva pàgina oficial tant en GitHub com a la web "Prophet" (Coleman, A., 2023).

La gran diferència entre Prophet i la resta de models és la seva simplicitat i facilitat d'ús. Aquest es basa en un model additiu en el qual la sèrie temporal s'expressa com la suma de components comuns com la tendència, les oscil·lacions periòdiques i els efectes de les vacances. Aquest model permet capturar de manera més senzilla i precisa les característiques complexes de les sèries temporals, com ara els canvis de tendència, els patrons estacionaris i les fluctuacions a curt termini.

Prophet ja porta incorporades funcions que permeten detectar canvis de tendència i estacionalitat en les dades, i permet als usuaris especificar els esdeveniments importants que poden afectar la sèrie temporal, com ara vacances, per a millorar la precisió de les prediccions. Això últim és molt important, ja que en el cas d'aquest projecte, la COVID marca un abans i un després en les sèries temporals i serà interessant utilitzar aquesta funcionalitat del model per capturar aquesta anomalia.

Alguns dels paràmetres que més utilitzarem són:

- **growth**: Permet especificar el tipus de creixement de la sèrie temporal. Pot ser *linear* per al creixement lineal o *logistic* per al creixement logístic.
- **changepoints**: Són els punts en el temps on es considera que hi ha un canvi en la tendència de la sèrie temporal. Aquest paràmetre permet especificar explícitament aquests punts.
- **holidays**: Permet afegir dates específiques de vacances o esdeveniments especials que poden afectar la sèrie temporal. Es pot proporcionar un DataFrame amb les dates i les etiquetes corresponents.
- **n_changepoints**: És el nombre total de punts de canvi que es consideren en el model. Si no s'especifica, el valor per defecte és 25.
- **changepoint_range**: És un valor entre 0 i 1 que especifica la proporció de dades a utilitzar per a trobar els punts de canvi. Per exemple, si s'estableix en 0,8, s'utilitzarà el 80% de les dades per a detectar els punts de canvi.
- **yearly_seasonality**, **weekly_seasonality**, **daily_seasonality**: Aquests paràmetres permeten habilitar o deshabilitar la presència de components estacionaris anuals, setmanals i diaris a la sèrie temporal, respectivament.
- **seasonality_mode**: Determina el tipus de model de temporada a utilitzar. Pot ser *additive* per a una estacionalitat additiva o *multiplicative* per a una estacionalitat multiplicativa.
- **seasonality_prior_scale**, **holidays_prior_scale**, **changepoint_prior_scale**: Aquests paràmetres permeten ajustar la força de les regularitzacions per a la temporada, les vacances i els punts de canvi, respectivament. Un valor més gran indica una regularització més feble.
- **mcmc_samples**: És el nombre de mostres MCMC (Cadenes de Markov de Monte Carlo) que cal utilitzar per a l'estimació dels paràmetres del model. Un valor més gran pot millorar la precisió, però augmentarà el temps de càlcul.
- **interval_width**: És l'amplada de l'interval de confiança per a les prediccions. Un valor més gran indicarà intervals més amples i, per tant, més incertesa en les prediccions.
- **uncertainty_samples**: És el nombre de mostres que cal utilitzar per a estimar la incertesa en les prediccions. Un valor més gran pot proporcionar una millor estimació de la incertesa però augmentarà el temps de càlcul.
- **stan_backend**: És una opció que permet canviar el motor *backend* utilitzat per a l'ajust del model. Per defecte, utilitza el motor *Stan*, però es pot canviar a *pymc3* o altres opcions compatibles.

5. CAS D'ESTUDI: ANÀLISI I PREDICCIÓ DE LES DADES

Aquest capítol és la part central del projecte en el qual, com s'ha explicat, es procedirà a fer una anàlisi i una predicció temporal i espacial de la criminalitat a Girona. L'objectiu és fer un estudi de caràcter prospectiu amb l'interès d'ajudar a la Policia Municipal de Girona a obtenir informació predictiva (de futur) en relació amb possibles delictes als diferents sectors policials de la ciutat.

El treball s'organitza amb distints apartats i s'ha englobat algunes de les fases que normalment s'utilitzen a "l'extracció de coneixement de les dades"(Fayyad, U. 1996). Aquestes fases s'han englobat en: Preprocessament de les dades, Anàlisi Exploratòria de les dades (EDA) i Anàlisi predictiva.

És important entendre que les fases de preprocessament i EDA no tenen una delimitació ben marcada que les separi. És habitual que es faci un preprocessament inicial i, a mesura que a l'EDA es vagi observant els resultats de les anàlisis, es pot tornar a preprocessar les dades. Per tant, són fases dependents l'una de l'altra.

A continuació, es presenta cada fase i es detalla els resultats més rellevants de cadascuna. Ens ha semblat d'interès acompanyar com a documents annexos l'informe tècnic policial on es troben totes les gràfiques i resultats i a on es poden consultar els informes corresponents. S'ha de tenir en compte que l'informe de preprocessament no és públic i, per tant, no s'hi tindrà accés. El motiu és que en aquest informe hi ha dades confidencials i, per qüestions de seguretat i privacitat, s'ha exclòs del TFG.

5.1 Preprocessament de les dades

L'etapa de preprocessament de les dades és una de les que més temps requereixen. Consisteix en detectar inconsistències, dades esbiaixades, valors buits o mal introduïts, etc. i netejar-los, transformar-los, i seleccionar aquelles característiques o variables que es creu que poden ser més interessants pel projecte. Com que en aquesta fase és habitual visualitzar les variables que composen el joc de dades, a continuació es presenta cadascuna d'elles amb una breu descripció:

- **NumerDocu:** Número de document on queda registrat el fet delictiu. Pot estar repetit.
- **K_NUMER_ACCI:** Representa el mateix que NumerDocu.
- **Data_instruccio:** Data instrucció
- **Data_inici_fets:** dia/mes/any dels fets
- **Any_fets:** Any dels fets
- **Mes_fets:** Mes dels fets
- **Dia_fets:** Dia dels fets
- **Temps_inici_fets:** Hora dels fets
- **Hora_fets_ini:** Hora dels fets simplificat (0-23h)

- **IdCarrer:** Identificador del carrer
- **K_ID_CARRER:** Representa el mateix que IdCarrer
- **NomCarrer:** Nom del carrer
- **Numer:** No hi ha dada
- **CNum:** No hi ha dada
- **Escala:** No hi ha dada
- **Pis:** No hi ha dada
- **Porta:** No hi ha dada
- **IdLocal:** No hi ha dada
- **IdEdifici:** Identificador de l'edifici
- **K_ID_EDIFICI:** Representa el mateix que IdEdifici
- **LOCALITZAT:** Autor dels fets localitzat
- **LugarConcreto1:** Lloc concret dels fets
- **LugarConcreto2:** Representa el mateix que Lugar Concreto1
- **AmbitoEspacial:** Àmbit espacial dels fets
- **CodiSubsector:** Codi dels subsectors policials
- **K_CODI_SUBSE_PM:** Representa el mateix que CodiSubsector
- **K_CODI_SUBSEPM_2019:** Representa el mateix que CodiSubsector
- **Subsector:** Noms dels subsectors policials
- **Sector:** Nom dels sectors policials
- **EjeX:** Coordenades geogràfiques de longitud
- **EjeY:** Coordenades geogràfiques de latitud
- **Coordenades:** Booleà sobre si hi ha coordenades
- **XEtrs89:** Coordenades UTM de longitud
- **YEtrs89:** Coordenades UTM de latitud
- **Xwsg:** Coordenades geogràfiques de longitud
- **Ywsg:** Coordenades geogràfiques de latitud
- **Gravedad:** Gravetat del delicte
- **CodiServei:** El codi del servei.
- **Grupo_siagin:** Grup delictiu (Delictes)
- **Subgrupo_siagi:** Subgrup delictiu
- **NomSiagi:** Nom del subgrup de tipologia de delictes.
- **OrigenExtern:** Origen de la font d'informació
- **Origen:** Cos policial que ha realitzat el registre
- **CodiFet:** Codi de NomSiagi
- **TipusOrigen:** Representa el mateix que Grupo_siagi
- **GrupOrigen:** Grup del delicte.
- **NomOrigen:** Descripció del delicte
- **Documento:** Tipus de document registrat.
- **NumerComplet:** Número complet del document del fet delictiu
- **Ordre:** Ordre de prioritat del delicte.

En el registre d'informació de la Policia Municipal i Mossos d'Esquadra s'hi recullen un total de 22 variables. La franja temporal va del 2000 al 2023, un total de 124.291 registres de fets delictius. Cal recordar que les dades sobre les persones detingudes van del 2018 fins al 2022.

Una feina a fer abans d'iniciar les distintes fases del treball, és entendre bé cada variable, analitzar-la una per una per veure si hi ha presència de valors nuls, errors de sintaxi, valors duplicats, etc.

Un aspecte a remarcar i que girarà entorn al projecte és la duplicitat de la variable NumerDocu. Aquesta variable correspon al número de registre d'un fet delictiu. Com s'ha comentat anteriorment, aquestes dades poden estar duplicades. Aquest fet es dóna perquè en un mateix delicte, es poden imputar diferents càrrecs. Per exemple, un robatori amb violència o intimidació pot comportar que al delinqüent se l'imputi per lesions també. Així doncs, el número de document (NumerDocu) d'aquest delicte estarà duplicat i farà referència a cadascun dels càrrecs imputats del mateix delicte.

Per tant, en posteriors anàlisis s'haurà de tenir en compte aquest detall, ja que del total de 124.291 delictes enregistrats, 113.469 corresponen a fets delictius únics.

Pel que fa a l'espai temporal, com s'ha explicat, les dades corresponen a la franja temporal que va del 2000 al 2023. No obstant, com es veurà més endavant, només es treballarà amb les dades corresponents a la franja de temps entre el 2009 i el 2022.

En el cas de les localitzacions dels delictes, en tot el conjunt de dades registrades, hi ha un total de 9.849 adreces úniques i de 822 carrers en tot l'àmbit de Girona.

Una de les feines que ha comportat més temps ha estat la de geolocalitzar els delictes que no tenien coordenades. De 124.291 registres que hi ha al joc de dades, gairebé 99.000 registres estaven sense coordenades. Tot i ser una suma elevada, s'han agafat els registres amb adreces úniques. D'aquesta manera, s'ha baixat de 99.000 a gairebé 10.000 registres. A continuació, s'han seleccionat aquells registres corresponents a adreces úniques que no tenen coordenades. Aquest procés ha ajudat a reduir el total a 3.836 registres sense coordenades. Per tal de trobar aquestes últimes coordenades s'ha codificat un mètode *getLatLon()* amb Python per tal de fer peticions a l'API de Nominatim perquè a partir del nom del carrer es retornin les coordenades.

Una vegada s'han comprovat totes les variables i seleccionat aquelles més interessants pel treball, s'han guardat diferents Data Frames (df) en format CSV per utilitzar-los més endavant. En un primer CSV s'han guardat els delictes totals, en un altre els delictes únics i a l'últim CSV s'ha guardat el df que s'utilitzarà per l'etapa d'anàlisi predictiva. Aquest s'ha incorporat a l'etapa de preprocessament després d'haver realitzat la fase d'EDA. En aquest CSV s'han agregat els sectors policials calculant la mitjana de delictes per dies del mes per cadascun d'ells. La franja temporal engloba del 2009-01 fins el 2022-12.

5.2 Anàlisi Exploratòria de les Dades (EDA)

En aquest apartat realitzarem una anàlisi exploratòria de les dades en profunditat. Se seguirà un ordre d'escala pel que respecta el temps i l'espai. S'ha de tenir present, que les conclusions i els arguments que s'utilitzen en aquest apartat d'EDA s'han extret a partir de les pròpies dades. En alguns casos s'ha realitzat recerca externa a les dades, però, en la majoria dels casos, la informació extreta de les dades ha estat la principal font d'informació.

En cas que es vulgui tenir accés a totes les gràfiques i informació extreta de les dades, es pot consultar l'informe tècnic "Anàlisi exploratòria de les dades".

Abans de passar al següent subapartat, s'ha de tenir present que les gràfiques mostraran els delictes per tipologia i els delictes únics per separat. És a dir, els delictes que tinguin en compte la tipologia, seran aquells on s'englobin els diferents càrrecs imputats en un mateix delicte, per tant, amb un número de document duplicat. D'altra banda, els delictes únics faran referència als fets únics, sense tenir en compte els seus càrrecs i que s'agruparan per número de document (NumerDocu).

S'ha de tenir present, que del total de delictes, s'ha procedit a eliminar per l'EDA, tots aquells delictes enregistrats al C/Bacià. Aquest carrer és on es troba la comissaria central de la Policia Municipal a Girona. Tots aquells delictes que no s'han pogut georeferenciar, s'han geolocalitzat en aquest carrer. Per aquesta etapa d'anàlisi exploratòria, s'ha cregut pertinent, eliminar-los. Així doncs, no es comptabilitzaran 6.434 delictes totals enregistrats al carrer Bacià, que corresponen a un 5,18 % del total de delictes enregistrats a Girona. Pel que fa als delictes únics enregistrats, no se'n comptabilitzaran 6.173, que corresponen al 4,97 % del total de delictes únics enregistrats a la ciutat.

A continuació, analitzarem la tipologia de delictes que s'han registrat entre el 2000 i el 2023.

5.2.1 Tipologia dels delictes

Per començar, s'ha d'observar quines tipologies de delictes hi ha enregistrades. Hi ha quatre classes de delictes: "Altres fets penals", "Patrimoni", "Persones" i "Seguretat vial". D'entre aquestes quatre tipologies, destaquen els delictes contra el patrimoni. Hi ha un total de 94.560 delictes contra el patrimoni; 14.293 delictes contra les persones, 11.292 delictes d'altres fets penals i 4.146 delictes contra la seguretat vial. A continuació s'observarà quina distribució hi ha en cadascun d'ells.

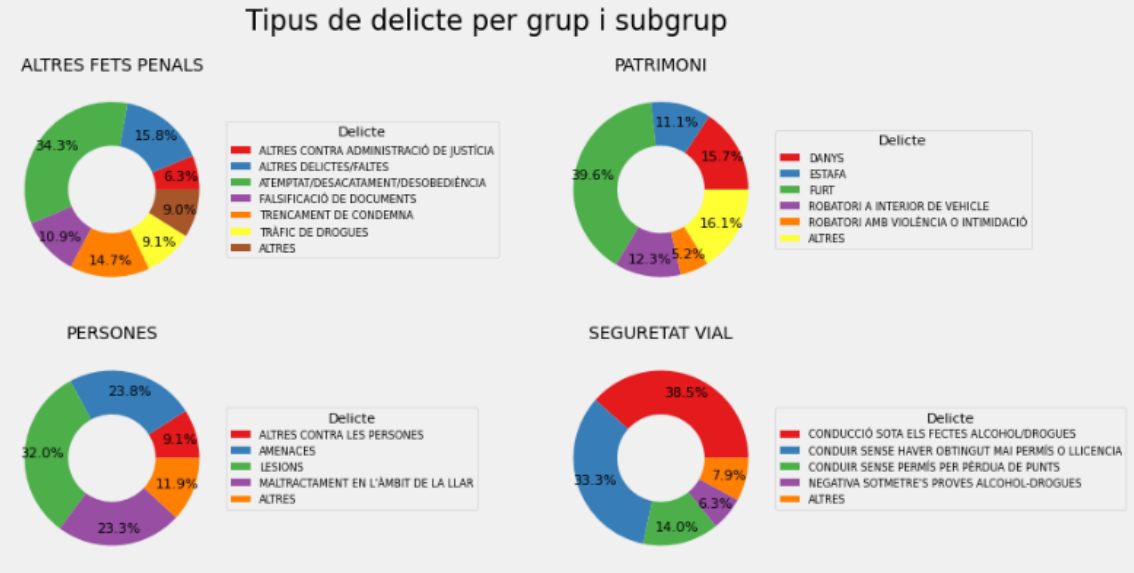

Figura 5.1:Tipologia dels delictes per grup i subgrup. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes amb tots els càrrecs i tipus de delicte. Autor: Sergi Payarol.

Altres fets penals

Els delictes englobats dins de la categoria "Altres fets penals" es distribueixen d'una manera força equitativa. S'observa com els delictes d'atemptat", desacatament o desobediència són els més presents, amb un 34,3%. D'altra banda, els "Altres delictes contra l'administració de justícia" són els que presenten un percentatge més baix, amb un 6,3%.

Patrimoni

Els delictes contra el patrimoni són els que han tingut més presència al llarg de tots aquests anys (2000 - 2023). El furt és el més present, amb gairebé un 40%. Mentre que el robatori amb violència o intimidació ha estat el menys present, amb un 5,2%. Tot i això, aquest 5,2% es tradueix en 4.917 delictes sobre el total d'aquest grup.

Persones

Els delictes de lesions són els més presents, amb un 32%. Destaca el 23,3% de delictes enregistrats com a maltractament en l'àmbit de la llar. Aquests últims superen els 3.000 delictes sobre el total d'aquest grup. La categoria d'amenaces també rep una porció força elevada amb un 23,8%.

Seguretat vial

Els delictes enregistrats sobre conducció sota els efectes de l'alcohol o drogues són els més presents, amb un 38,5%, que representa 1.596 delictes. El delicte de conduir sense permís per pèrdua de punts també és força elevat, i arriba al 33,3%.

Els subtipus de delictes més ocorreguts al llarg d'aquest temps són: el furt, amb més de 35.000 casos; els danys, amb gairebé 15.000 casos; el robatori a interior de vehicle, amb gairebé 12.000 casos; i l'estafa, amb poc més de 10.000 casos. Així doncs, el furt és el delicte que més es dóna. La principal característica d'aquest tipus de delicte és que no es produeix cap tipus de violència sobre les pertinences ni sobre les persones. En canvi, els danys i el robatori a interiors de vehicles implica un cert tipus de violència sobre béns materials.

Després d'aquest primer grup de delictes més recurrents, el següent grup està format pel robatori amb violència o intimidació, lesions, robatori amb força a un domicili, amenaces i maltractament en l'àmbit de la llar. Aquest segon grup ofereix unes característiques més sèries pel que respecta la violència física i/o mental de les persones i de les pertinences. Cadascun dels delictes d'aquest segon grup té un total de menys de 5.000 casos.

Tenint en compte la distribució del delicte segons la seva tipologia seria interessant, en futurs projectes, analitzar més a fons aquests primers grups, o inclús focalitzar-se en un tipus de delicte en concret, com per exemple el furt que és el més present a Girona.

Una vegada s'ha observat la distribució dels delictes tenint en compte la seva tipologia, es passa a analitzar la seva distribució temporal. Per tant, es formula una pregunta d'entrada: Quan han ocorregut els delictes?

5.2.2 Anàlisi temporal: Quan han ocorregut els delictes?

En aquest subapartat s'analitza la sèrie temporal dels fets delictius. Consulteu la primera gràfica de l'apartat "4. Anàlisi temporal: Quan han ocorregut els delictes?" de l'informe EDA. Com es pot observar a la gràfica, es poden descartar els delictes enregistrats del 2000 al 2008 i els del 2023 degut a que no hi ha dades suficients com per extreure'n conclusions. Així doncs, eliminant els anys 2000 - 2008 i 2023, es descarten 689 delictes únics, que representen un 0,64% del total de delictes únics; i 803 delictes totals, que representen un 0,68% de tots els delictes totals.

Una vegada s'ha eliminat aquests anys es pot fer una primera exploració de la sèrie. A continuació, observem la distribució dels delictes únics agrupats per dies.

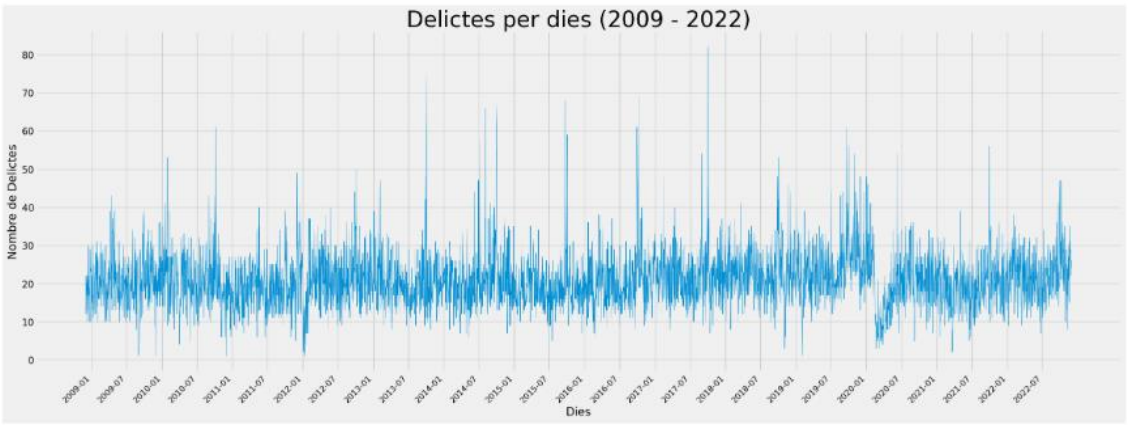

Figura 5.2: Sèrie temporal dels delictes per dies, del 2009 al 2022. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

Com es pot detectar a la gràfica, la distribució dels delictes per dies segueix una certa estacionalitat. Si s'observa, a simple vista, la gràfica es poden remarcar tres fets. D'una banda, el 2012 va haver-hi una baixada de delictes important. Tanmateix, el 2020, amb el confinament produït per la COVID, es percep una disminució rellevant dels delictes. Per últim, hi ha certs dies que despunten exageradament respecta la resta. Més endavant s'analitzarà quins són aquests dies.

A continuació, es presenta la sèrie amb els delictes agrupats per mesos de l'any.

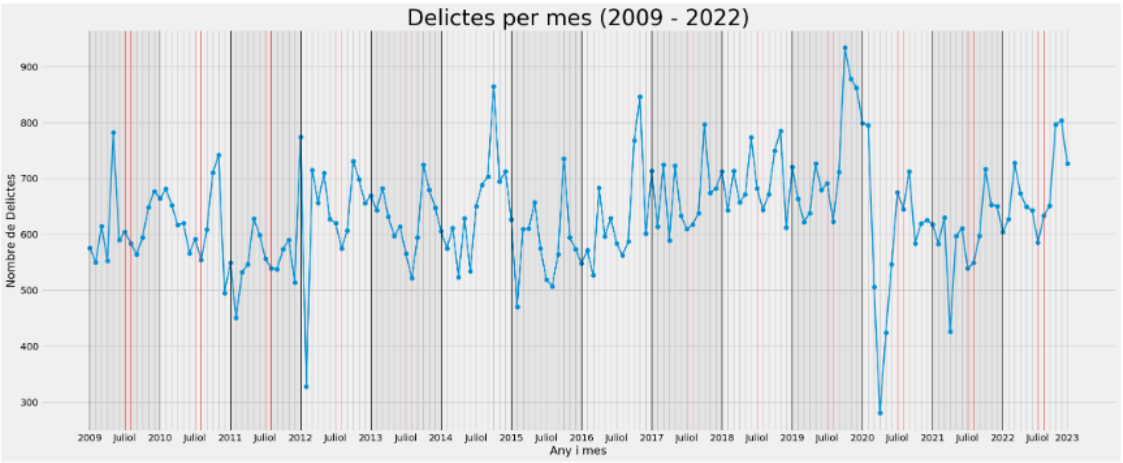

Figura 5.3: Sèrie temporal dels delictes per mesos, del 2009 al 2022. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

En aquesta gràfica es pot contemplar l'estacionalitat dels delictes més detalladament. Durant els mesos d'abril, maig, octubre i novembre es produeixen els pics, mentre que a l'estiu baixen els casos.

Es pot observar més clarament la baixada de delictes durant el 2012 i la COVID. És important veure com a partir de la pandèmia, s'ha trencat la tendència de delictes, tot i que a partir de l'octubre de 2022 la línia de regressió torna a enfilar-se cap a dalt. La tendència puja una altra vegada per anivellar-se a les dades del 2017 al 2019. També

pot ser que en ser el període de temps on acostumen a ocórrer més delictes, hagi despuntat el 2022 per baixar més endavant.

Si s'observen les anomalies de la sèrie, hi ha certs pics que es troben molt per sobre de la resta: l'octubre de 2014, l'octubre de 2016 i, sobretot, l'octubre de 2019. És interessant veure que tots els pics es donen a l'octubre. És probable que en ser les Fires de Sant Narcís de Girona a l'octubre, els casos delictius augmentin. Més endavant s'analitzarà quins tipus de delictes es produeixen.

Una vegada vista la distribució per mesos, a continuació observarem el nombre de delictes segons el dia de la setmana.

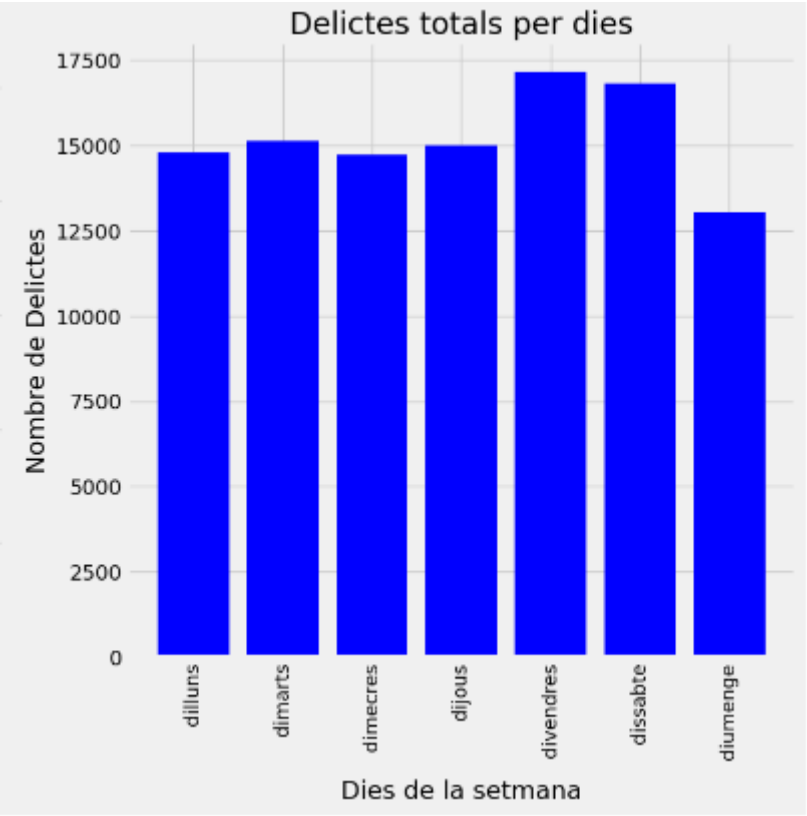

Figura 5.4: Delictes totals agregats per dies de la setmana. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

La distribució dels delictes per dies de la setmana mostra una distribució força equitativa entre el dilluns i el dijous. No obstant això, el divendres i el dissabte mostren un repunt de casos, per després caure el diumenge, sent aquest el dia que menys delictes hi ha. S'ha d'entendre que es mostren només els delictes únics totals, sense tenir en compte cap tipologia. Per tant, no es poden extreure conclusions sòlides sobre els motius d'aquesta distribució. Sí que es pot pensar, però, que els divendres i els dissabtes és quan les persones sortim més, i hi ha més concurrència de persones que visiten la ciutat per compra o oci. Com que el furt és el delicte que més es comet, podria haver-hi una certa correlació entre aquests dos fets.

Si es baixa encara més l'escala temporal i s'analitza la distribució per hores, s'obtenen els següents resultats:

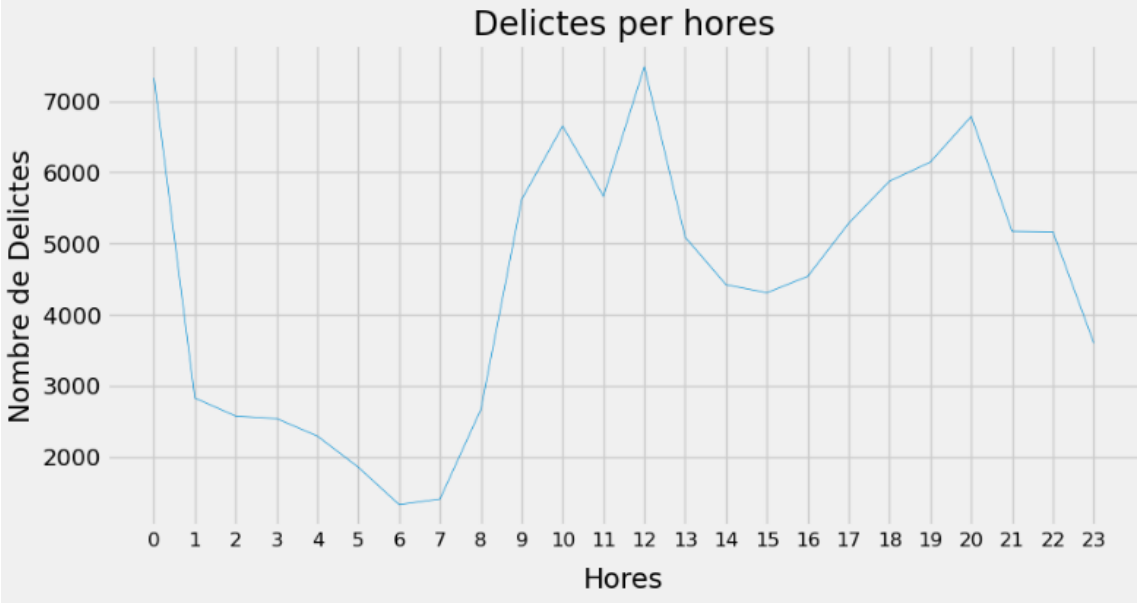

Figura 5.5: Sèrie temporal dels delictes per hores del dia. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

Es pot veure clarament com els repunts de delictes es produeixen a les 24:00h de la nit, de les 10h del matí a les 12h del migdia i als voltants de les 20:00 del vespre. A l'altre extrem de la balança, des de l'1h de la matinada a les 7h del matí és quan menys fets delictius es cometen. També es detecta una lleugera disminució durant les 13h i les 16h. Analitzant-ho per franges horàries, la distribució varia lleugerament.

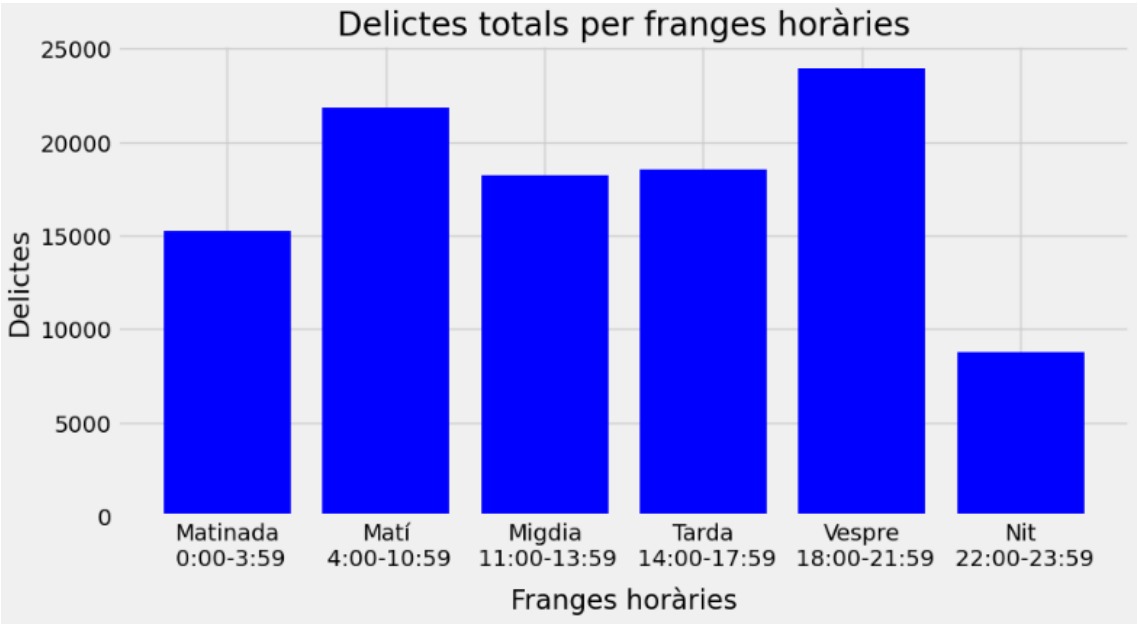

Figura 5.6: Delictes totals agregats per franges horàries. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

La discretització per franges horàries s'ha fet segons la distribució oficial que s'estipula al web del Parlament de Catalunya.

El vespre encapçala la primera posició amb més nombres de delictes, mentre que el matí és el segon moment del dia quan es produeixen més delictes. S'ha de tenir en compte, que a la gràfica anterior, les 12h del migdia i de la nit, són les hores més problemàtiques. A l'haver-ho discretitzat per franges horàries, la distribució de delictes ha pogut quedar descompensada, ja que algunes franges tenen més hores que d'altres. Per aquest fet, s'ha de pensar, que quan es parla d'hores i de franges horàries, la visió general de quan es cometen els delictes pot canviar.

Una vegada s'ha fet una aproximació temporal a diferents escales sobre els delictes, procedirem a fer l'anàlisi tenint en compte la tipologia de delictes.

5.2.2.1 Tipologia dels delictes en el temps

Abans de veure les tipologies en el temps, s'ha cregut oportú mostrar la distribució dels delictes segons la seva gravetat per dies de la setmana.

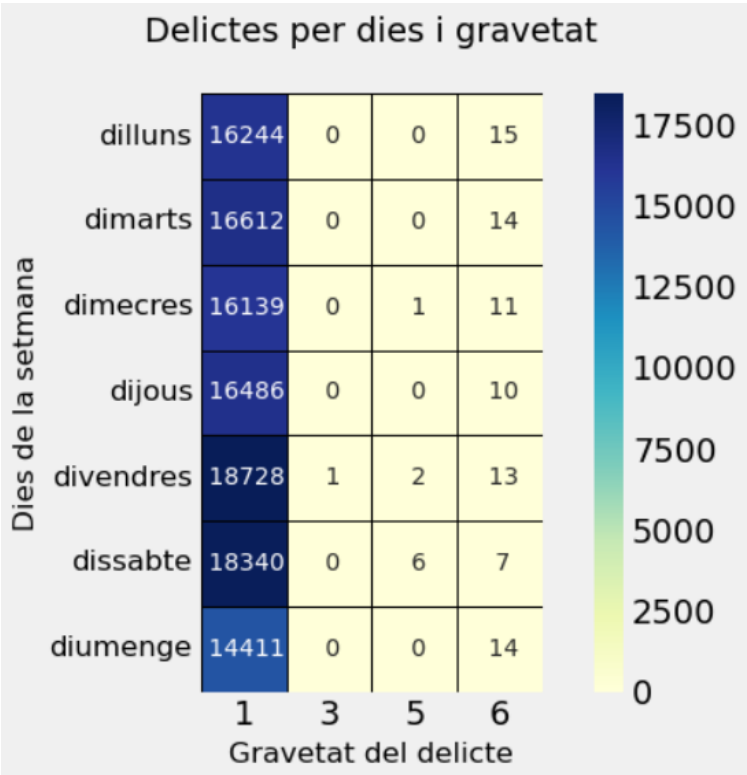

Figura 5.7: Taula de contingència sobre els delictes agrupats per gravetat del delicte i dies de la setmana. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

Clarament, es pot veure que tots els delictes tenen gravetat 1. S'ha de tenir en compte que un delicte de gravetat 1 és el més greu que pot haver-hi, mentre que un delicte 6, correspondrà a un delicte lleu. Així doncs, tots els delictes enregistrats entre el 2009 i el

2022 són molt greus. Aquest fet pot ser degut a que la població no es molesta en fer una denúncia a no ser que sigui un delicte força greu.

A continuació, analitzarem la distribució temporal dels delictes per tipologia a una escala mensual.

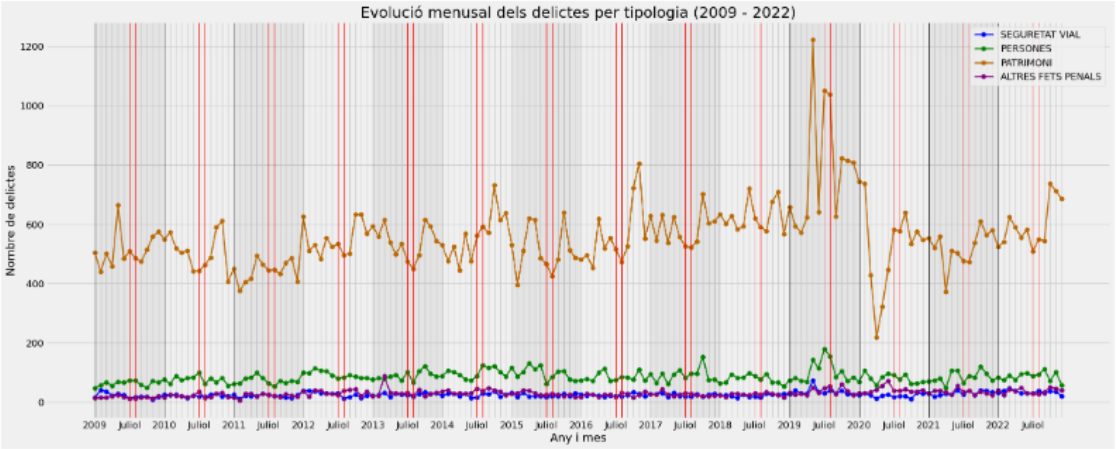

Figura 5.8: Sèries temporals dels delictes per tipologia de delicte. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

Si s'observa l'evolució dels delictes, tenint en compte la seva tipologia, es veu que en el cas del patrimoni se segueix una certa estacionalitat. Es pot veure un repunt important de delictes contra el patrimoni en el 2019, i la baixada en picat a causa del confinament el 2020.

És interessant veure com els delictes contra les persones segueixen en certa manera l'estacionalitat dels de patrimoni. S'ha de tenir present que un delicte contra el patrimoni també pot tenir associat un càrrec contra les persones.

En el cas dels delictes contra les persones és difícil veure la seva distribució temporal, atès a l'ombra que produeixen els delictes contra el patrimoni. Així doncs, es procedeix a veure la seva distribució per separat.

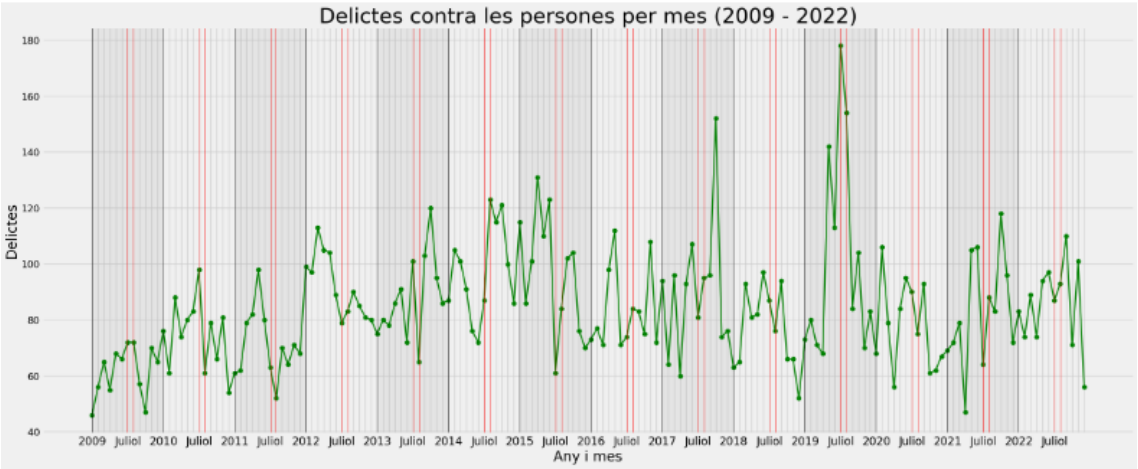

Figura 5.9: Sèrie temporal dels delictes contra les persones per mes. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

En aquesta gràfica, es veu l'evolució temporal dels delictes contra les persones, que segueix certa estacionalitat, amb una certa tendència positiva a partir del 2012. Mentre que el juliol del 2019 hi ha el pic més important, que segurament està relacionat amb l'augment de casos de delictes contra el patrimoni.

Posant el focus en els cinc dies amb més delictes enregistrats, es detecta el següent:

Data inici dels fets	Total de delictes únics
01/11/2017	82
01/11/2013	75
06/11/2016	69
24/10/2015	68
02/11/2014	67

Figura 5.10: Taula amb els cinc dies amb més delictes registrats. Per realitzar aquesta taula s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

- El dia 1 de novembre del 2017, el 60% dels casos van ser delictes per furt.
- El dia 1 de novembre del 2013, el 57% dels casos van ser delictes per furt.
- El dia 6 de novembre del 2016, el 71% dels casos van ser delictes per furt.
- El dia 24 d'octubre del 2015, el 71% dels casos van ser delictes per furt.
- El dia 2 de novembre del 2014, el 67% dels casos van ser delictes per furt.

Així doncs, els cinc pitjors dies enregistrats, els delictes que més es van produir van ser furts. És interessant veure que no ha sortit l'octubre del 2014, 2016 o del 2019, que són els mesos amb un repunt important de delictes. Aquest fet es pot donar perquè potser en aquests mesos la distribució de delictes diària contempla una mitjana més elevada sense tenir casos atípics molt extrems. És important veure quins dies són els que han estat més problemàtics, ja que poden ajudar a entendre si hi ha un factor a tenir en compte en una franja de temps molt reduïda (24 h). Per tant, són diferents aproximacions, que poden aportar informació diferent.

En el cas de l'octubre del 2019, el que més va haver-hi són furts, representant un 22% sobre el total. A continuació, hi ha un 15% de robatori a interior de vehicle i danys en ambdós casos. Com es pot veure, són la classe de delictes que ja s'ha analitzat prèviament.

Un fenomen interessant de veure és la distribució dels delictes per setmanes de cada mes. En aquest cas, s'ha estructurat un mes en cinc setmanes, sent la cinquena setmana la que menys dies té (4/5 dies). Es veu primer l'agrupació dels delictes per aquestes setmanes.

Setmana	Total de delictes únics
Primera	21890
Segona	20433
Tercera	21270
Quarta	20622
Cinquena	8545

Figura 5.11: Taula amb el total de delictes distribuïts per setmanes. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

La primera i la tercera setmana és on hi ha més presència delictiva. Tot i això, la cinquena setmana té una suma ben alta, tenint en compte que només es compten els delictes durant 4/5 dies. A més a més, també s'ha observat que el dia 1 és el dia amb més delictes del mes, segons el còmput total, amb 4.533 delictes únics.

Tot seguit, es farà una distribució de les setmanes per tots els anys registrats. Es veurà per cada mes, quina és la setmana amb més delictes únics registrats.

Any	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2009	Quarta	Quarta	Tercera	Tercera	Tercera	Primera	Quarta	Segona	Quarta	Quarta	Tercera	Segona
2010	Quarta	Quarta	Quarta	Segona	Tercera	Tercera	Quarta	Tercera	Tercera	Tercera	Segona	Tercera
2011	Segona	Tercera	Tercera	Segona	Tercera	Segona	Tercera	Tercera	Tercera	Primera	Primera	Segona
2012	Segona	Quarta	Primera	Quarta	Tercera	Quarta	Tercera	Quarta	Tercera	Quarta	Primera	Tercera
2013	Tercera	Tercera	Primera	Segona	Segona	Segona	Tercera	Tercera	Tercera	Quarta	Primera	Tercera
2014	Quarta	Tercera	Tercera	Tercera	Segona	Primera	Segona	Primera	Primera	Quarta	Primera	Tercera
2015	Segona	Quarta	Segona	Tercera	Segona	Primera	Tercera	Quarta	Tercera	Quarta	Primera	Tercera
2016	Tercera	Quarta	Primera	Tercera	Primera	Quarta	Tercera	Tercera	Quarta	Cinquena	Primera	Tercera
2017	Quarta	Tercera	Tercera	Segona	Segona	Segona	Tercera	Segona	Quarta	Segona	Primera	Quarta
2018	Segona	Tercera	Tercera	Tercera	Tercera	Tercera	Segona	Segona	Tercera	Quarta	Primera	Quarta
2019	Primera	Primera	Tercera	Primera	Tercera	Tercera	Tercera	Segona	Tercera	Tercera	Primera	Tercera
2020	Tercera	Segona	Segona	Segona	Quarta	Tercera	Segona	Quarta	Tercera	Tercera	Tercera	Tercera
2021	Tercera	Tercera	Segona	Segona	Tercera	Tercera	Segona	Quarta	Tercera	Tercera	Segona	Tercera
2022	Quarta	Quarta	Primera	Segona	Tercera	Tercera	Tercera	Segona	Quarta	Segona	Primera	Quarta

Figura 5.12: Taula on es mostra la distribució de setmanes al llarg dels anys. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

En la taula 5.12 s'ha generat una taula creuada. Es pot veure clarament, com la tercera setmana té més presència respecta les demès. Però, s'ha de tenir en compte que la primera setmana és la que té més delictes enregistrats. Això fa pensar, que els delictes que es cometen a la primera setmana estan ben concentrats. Com es pot veure a la taula creuada, les primeres setmanes del mes de novembre mostren un clúster força evident. També s'ha de tenir en compte la quarta i cinquena setmana que es concentren en el mes d'octubre.

Per tant, les festes de Girona, com les Fires de Sant Narcís, mostren possiblement una correlació alta amb el nombre de delictes.

Una vegada s'ha fet la visió temporal dels fets delictius, es pot tenir una visió general de com es distribueixen al llarg del temps a diferents escales. A continuació, es procedirà a l'anàlisi espacial. Per tant, es respondrà la gran pregunta: On es cometen els delictes?

5.2.3 Anàlisi espacial: On es cometen els delictes?

En aquest apartat es localitzaran els delictes en el territori. A més a més, també s'observarà la seva evolució temporal tot tenint en compte la tipologia de cada delicte.

És important entendre que, per qüestions de privacitat i de seguretat, l'escala que s'utilitzarà per contextualitzar geogràficament els delictes és de sectors i subsectors policials. En cap moment es mostrarà la localització exacta del delicte.

S'aconsella seguir aquest apartat amb l'ajuda dels mapes interactius que s'han generat en el *Jupyter notebook* de l'informe tècnic. En concret, l'apartat "5. Anàlisi espacial: On es cometen els delictes?" de l'EDA. Els mapes interactius aporten informació addicional si es clica a sobre de la geometria.

L'aportació gràfica que es farà en aquest apartat serà referent als mapes estàtics. És per aquest fet, que es referenciarà en tot moment els mapes interactius de l'informe policial que s'estiguin comentant.

Al primer mapa de "Delictes totals per subsectors policials", es pot identificar clarament un clúster que es concentra al sector oest del municipi. Santa Eugènia, Can Gibert del Pla, la Devesa, i l'Eixample Sud, entre d'altres, és on n'hi ha hagut més delictes; mentre que la cara est del municipi és on n'hi ha hagut menys. Pot ser que el nombre d'habitants per subsectors i la densitat de població siguin variables que tinguin un pes rellevant en aquests resultats.

Ara bé, el mapa de "Delictes totals per sectors policials" mostra clarament que els sectors amb major presència delictiva són el sector 1 i el sector 2.

Al sector 1 s'engloben els subsectors de la Devesa, el Güell, el Mercadal, Sant Feliu-Catedral, Ajuntament-Rambles i Sant Domènec-Fora muralles; mentre que en el sector 2 hi ha l'Eixample Nord, l'Eixample Sud, Casernes i Migdia. A continuació, hi ha el sector 8 (Mas Xirgu, Palau, etc.), el sector 4 (Santa Eugènia) i el sector 3 (Sant Narcís, Parc Central) amb bastants fets delictius enregistrats; seguidament hi ha el sector 6 (Domeny, Taialà, Fontajau, etc.); i, per últim, estan els sectors amb menys casos delictius: el sector 7 (Sant Daniel, Montjuïc, Campdorà, Pont Major, etc.) i el sector 5 (Font de la Pólvora, Vila-roja, Pedreres, Mas Ramada, etc.).

Mapa de calor dels delictes únics agregats per dies i subsectors policials

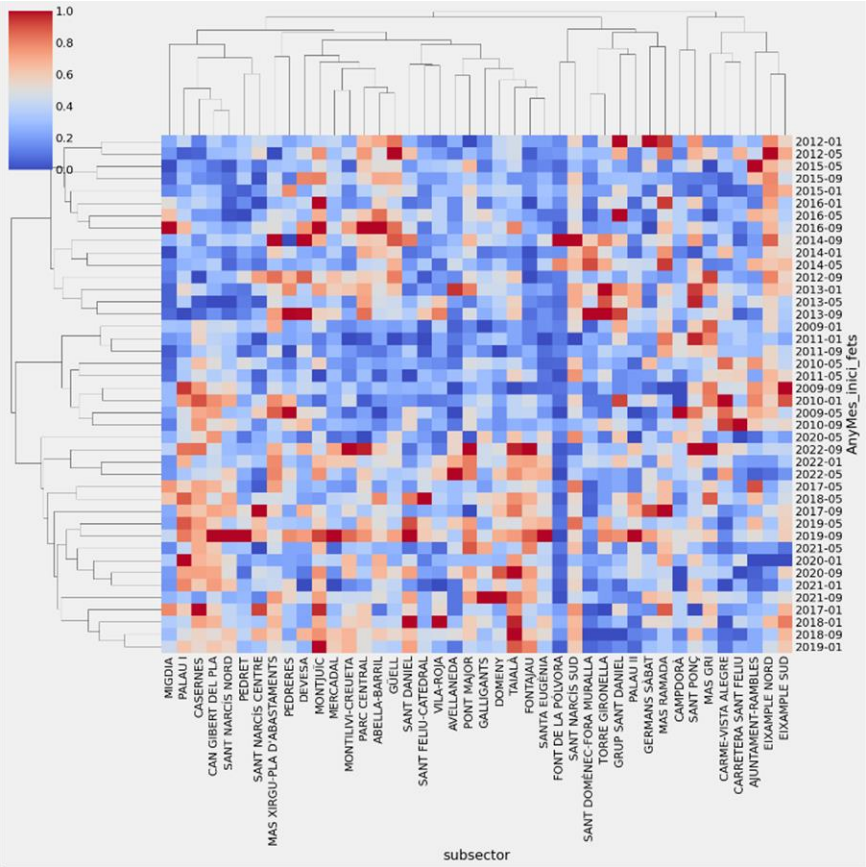

Figura 5.13: Mapa de calor on es mostra la distribució dels delictes agrupats per dies i subsectors. Les dades han sigut estandarditzades per tal que es mostri la desviació de les dades respecte la mitjana. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

Al mapa de clústers 5.13 s'ha aplicat una transformació a les dades perquè tinguin una distribució amb mitjana zero i una dispersió relativa a la desviació estàndard original. En escalar les dades amb una desviació estàndard d'1, s'obté una nova distribució de les dades en què la mitjana és 0 i la dispersió (variabilitat) es mesura en termes de desviacions estàndard.

Quan hi ha subsectors amb una desviació estàndard propera a 1, indica que aquests subsectors tenen una dispersió relativament alta respecta a la seva mitjana. Això significa que els valors estan més dispersos al voltant de la mitjana i que hi ha una variabilitat considerable entre ells.

D'altra banda, els subsectors amb una desviació estàndard propera a 0, indiquen que tenen una dispersió molt baixa respecta a la seva mitjana. Això significa que els valors estan més concentrats al voltant de la mitjana, i que hi ha poca variabilitat entre ells.

Per tant, es pot veure com hi ha sectors on la distribució dels delictes és més homogènia en el temps i no mostra una tendència molt abrupta en cap moment. Un exemple podria

ser Font de la Pólvora, que mostra una distribució dels delictes homogènia sense tenir molta variabilitat. També es pot dir el mateix de Santa Eugènia, ja que no es percep molta variabilitat en les dades. Això no treu que no hi hagi delictes. Al contrari, Santa Eugènia mostra una continuïtat sense pics importants (tret del 2019) i això indica que el nombre de delictes és constant en el temps.

D'altra banda, hi ha subsectors on la seva desviació és més propera a 1 en alguns moments, la qual cosa indica que la variabilitat de les dades és més dispersa i que se solen donar moments on la presència de delictes és més elevada. L'Eixample Sud podria ser un exemple, o Taialà. Per tant, aquesta informació ens indica quins subsectors tenen casos atípics, esporàdics de fets delictius; mentre que una variabilitat més propera a 0, és a dir, més propera a la mitjana significa que són subsectors on la presència de delictes és més homogènia. Els dos casos són importants tenir-los en compte ja que mostren el propi tarannà del subsector respecta els delictes que s'hi produeixen.

A continuació, veurem una composició de mapes on es mostren l'evolució espacial i temporal dels delictes per subsectors policials.

Delictes únics per anys i subsectors i distribuïts en diferents franges temporals

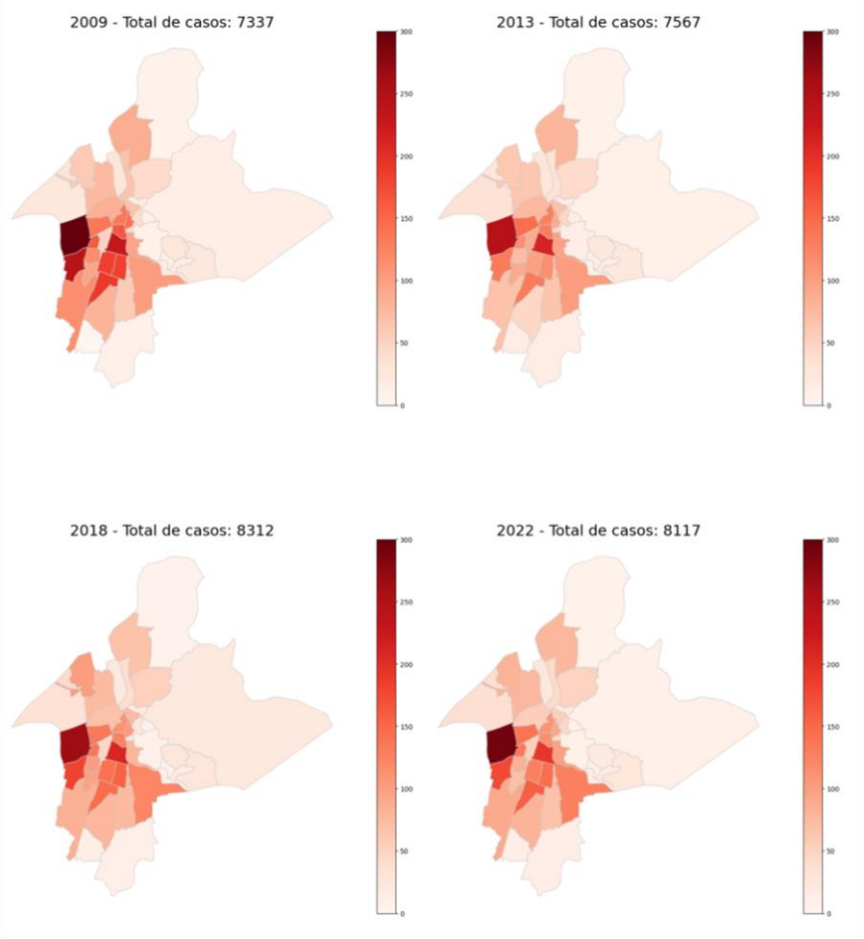

Figura 5.14: Composició de mapes on es mostren els delictes anuals per subsectors en diferents franges temporals. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.

Clarament, Santa Eugènia mostra un major nombre de delictes al llarg del temps. La resta de subsectors mostren una continuïtat força estable. En certs moments, els subsectors del centre mostren un canvi lleuger de delictes, però no es percep res inusual.

Si ara observem el GIF de l'informe tècnic on es mostra una evolució temporal del mapa de calor per anys, veurem com es dibuixa un *hot spot* força present al centre de la ciutat, concretament als subsectors policials del Mercadal, Eixample Nord i Eixample Sud. S'ha de tenir en compte que el mapa de calor mostra la distribució de la densitat de punts (delictes) al municipi. Així doncs, es pot veure com el *hot spot* més present es troba al Mercadal. De totes maneres, també és interessant veure com al 2014 es reparteixen més *hot spots*, sobretot a la Font de la Pólvora i als subsectors limítrofs al Mercadal.

El que potser és més interessant de veure és el següent mapa de calor de l'informe tècnic, on es mostren el total de casos per tots els mesos de la sèrie temporal. A continuació, es mostra una composició de mapa on s'ha representat el *hot spot* principal, la seva perifèria més propera i llunyana i els "satèl·lits" que mostren *hot spots* intermitents al llarg de la sèrie temporal.

Distribució dels *hot spots* i perifèries més propera i allunyada dels delictes

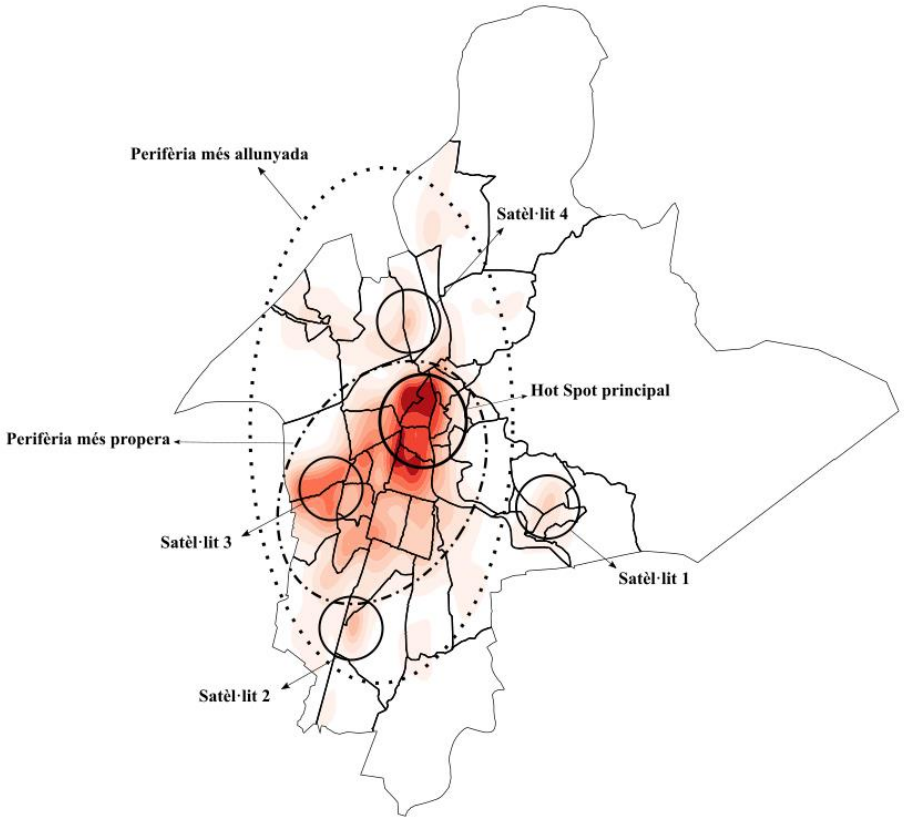

Figura 5.15: Mapa de calor on es mostra la densitat de delictes totals. Les delimitacions corresponen als subsectors policials. Es pot apreciar un seguit de satèl·lits que es localitzen en diferents parts del mapa. A més a més, el hot spot *principal se situa ben bé en el centre del municipi, on es generen un seguit de perifèries al voltant d'aquest amb una projecció dels delictes en direcció sud-oest. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes únics. Autor: Sergi Payarol.*

Veient l'evolució mensual es pot percebre més el moviment del *hot spot* i de la seva projecció en l'espai. En aquest sentit, la projecció del *hot spot* és generalment cap al sector sud-oest.

El punt central del *hot spot* va pivotant entre el Mercadal i l'Eixample Nord al llarg de tota la sèrie temporal. La seva perifèria més propera (Santa Eugènia, Sant Narcís, Eixample Sud, Casernes, Migdia, Parc Central, el Güell i la Devesa van mostrant-se intermitentment, però amb una presència notòria de delictes.

En certs moments de l'any, més pròxims a finals i a principis d'any, la projecció arriba encara més lluny, cap a la perifèria més allunyada (Domeny, Taialà, Fontajau, Montjuïc, Palau, Mas Xirgu, Mas Gri, etc.).

És interessant veure com al llarg de la sèrie certs "satèl·lits" es mostren feblement en alguns subsectors. En aquest cas, s'han distingit quatre satèl·lits.

El primer satèl·lit correspon als subsectors de Font de la Pólvora, Mas Ramada i Grup Sant Daniel. El 2014 és quan es pot distingir amb més força el *hot spot*. Aquest *hot spot* es localitza a Font de la Pólvora, i correspon a una sèrie de delictes que es van registrar, i que corresponen a la defraudació de fluid elèctric on es van denunciar múltiples veïns (x84) que punxaven la llum.

Un segon satèl·lit correspon als subsectors de Mas Xirgu, Palau i Mas Gri. Aquest *hot spot* es mostra amb major freqüència que l'anterior. En aquesta zona són freqüents els delictes contra l'oci i la restauració, el comerç, etc. De fet, és una àrea on es concentren força discoteques, supermercats i botigues.

Un tercer satèl·lit correspon als subsectors de Santa Eugènia, Sant Narcís Nord/Centre i Can Gibert del Pla. Aquest *hot spot* és molt més freqüent i mostra més intensitat. És lògic, atès que es localitza a la perifèria més propera del principal *hot spot*. Els delictes que es cometen aquí són molt més diversos. S'ha de remarcar que Santa Eugènia és el subsector policial amb més delictes registrats.

Un últim satèl·lit correspon als subsectors de Fontajau i Pedret. Aquest es troba ben a prop de la perifèria més propera, i sobretot del *hot spot* principal, tot i que a l'hora d'englobar-lo en una àrea s'ha optat per ficar-lo dins de la perifèria més allunyada, ja que la projecció dels delictes més forta és cap al sud-oest. És interessant veure que els delictes que es cometen en aquesta zona és en gran part contra la seguretat vial. De fet, aquest *hot spot* se situa ben bé a la sortida i entrada de Girona cap a l'N-II direcció Sarrià de Ter.

Així doncs, a partir d'aquest mapa de calor, i de la composició que s'ha fet per resumir les perifèries i els diferents *hot spots* es mostra la distribució temporal i espacial per subsectors en el temps. Aquesta informació pot enriquir en gran mesura aquesta anàlisi

que estem fent, i pot fer aflorar noves preguntes que es puguin resoldre en projectes futurs.

A continuació, s'analitzarà la tipologia dels delictes en l'espai, per subsectors policials.

5.2.3.1 Tipologia dels delictes en l'espai

En aquest subapartat analitzarem la tipologia de delictes en els subsectors policials. Per tal d'observar la taula de contingència, adreceu-vos a l'apartat "5.1 Tipologia dels delictes en l'espai" de l'EDA, i a la taula "Tipologia de delictes distribuïts per subsectors".

A continuació, s'analitzarà la taula de contingència. S'ha de tenir en compte que la taula mostra la distribució de cada tipus de delicte per tots els subsectors.

Altres fets penals

Tenen una distribució força concentrada en certs subsectors: Can Gibert del Pla, Eixample Nord, Mercadal, Parc Central i Santa Eugènia.

Patrimoni

Els subsectors amb més presència de delictes contra el patrimoni són: Can Gibert del Pla, Casernes, Devesa, Eixample Sud, Güell, Mercadal, Parc Central i Santa Eugènia.

Persones

Els delictes contra les persones es concentren majoritàriament a Can Gibert del Pla, Eixample Sud, Mercadal, Parc Central, Sant Narcís Nord i Santa Eugènia.

Seguretat Vial

Els subsectors amb més presència de delictes contra la seguretat vial són: Can Gibert del Pla, Casernes, Devesa, Eixample Nord, Güell, Mas Xirgu, Sant Narcís Nord i Santa Eugènia.

Es pot veure com en certs subsectors hi ha una distribució força elevada en tots els tipus de delicte. En aquest cas es pot destacar Santa Eugènia, Can Gibert del Pla, Casernes, Eixample Nord i Eixample Sud, el Güell i el Mercadal. És interessant veure com els delictes contra les persones es concentren més a Santa Eugènia i Can Gibert del Pla.

D'altra banda, si s'analitza la distribució dels delictes segons la seva tipologia en els diferents subsectors s'aprecia el següent:

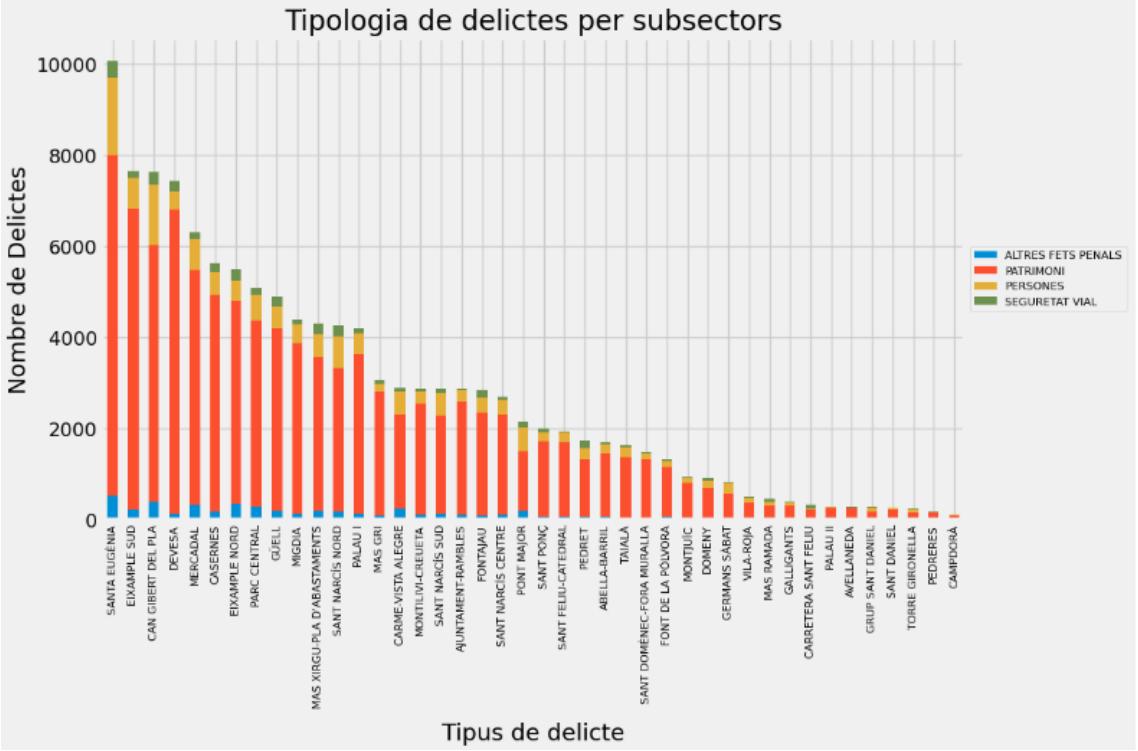

Figura 5.16: Tipologia dels delictes per subsectors policials. Per realitzar aquesta gràfica s'ha utilitzat la taula de delictes amb tots els càrrecs i tipus de delicte. Autor: Sergi Payarol.

En la majoria de subsectors, la presència de delictes contra el patrimoni és força present. El subsector de Santa Eugènia és on més delictes contra el patrimoni i les persones hi ha. A continuació, l'Eixample Sud, Can Gibert del Pla i la Devesa són els subsectors amb major presència delictiva. És interessant veure com Can Gibert del Pla és el segon subsector on es donen més delictes contra les persones. D'altra banda, els subsectors que tenen una presència més baixa de delictes (menys de 1.000) es localitzen als sectors 5 i 7. En aquests subsectors, la distribució dels delictes sembla estar més balancejada entre dues tipologies. Es pot posar com a exemple Grup Sant Daniel o Torre Gironella, on la proporció de delictes contra el patrimoni i les persones està força equilibrada.

Com bé s'ha pogut veure al llarg d'aquests apartats, la distribució espacial i temporal té associats uns patrons que se solen repetir al llarg de la sèrie. En aquesta anàlisi falta, però, saber qui comet els delictes. Al llarg del següent apartat farem una aproximació de les persones detingudes que s'han enregistrat des del desembre del 2018 fins el 2022 i les relacionarem amb la taula de delictes que s'ha estat estudiant.

5.2.4 Anàlisi de les persones detingudes: Qui comet els delictes?

La taula de persones detingudes està formada per una sèrie de registres sobre els actors del delicte que han estat arrestats. Les dades enregistrades són confidencials, i per raons de seguretat, no es mostrarà cap dada sensible. Les variables que s'analitzaran seran el sexe de les persones detingudes, la seva edat i la nacionalitat.

El nombre total de persones detingudes que hi ha enregistrats segons el delicte és de 1.095. De persones detingudes úniques, n'hi ha un total de 854. Després de fusionar la taula de persones detingudes amb la de delictes per codi ID del fet, veiem que el nombre total puja a 1.252 registres. Aquest fet és degut a que els delictes poden tenir més d'una tipologia amb un mateix ID del fet. És a dir, un detingut pot haver comès un delicte que afecti el patrimoni i a les persones a la vegada. Per aquest motiu, sortirà dues vegades a la taula de delictes.

Per tant, es pot veure com en 157 delictes el detingut haurà estat imputat amb múltiples càrrecs.

Tanmateix, quan s'ha fusionat la taula de persones detingudes amb la taula de delictes únics, han sortit un total de 727 registres. Per tant, s'ha perdut registres de la taula de persones detingudes. Aquest fet és degut a que la clau forana de la taula de persones detingudes no correspon amb la seva clau forana de la taula de delictes. Per tant, en la primera fusió entre la taula de delictes i la taula de persones detingudes, la suma total podria haver estat lleugerament més elevada.

Al llarg d'aquest apartat s'ha de tenir present que s'observen les característiques de l'autor dels fets per delicte. És a dir, no s'estan analitzant les persones detingudes amb identificador únic, sinó que estem focalitzant-nos primer en el delicte, i després en el detingut. Això voldrà dir que hi haurà delictes on l'autor dels fets és el mateix.

Com que s'analitzen les dades amb cert grau de confidencialitat, no s'ha afinat en analitzar els tipus de delictes, ni el nombre total d'aquests sobre cada detingut. Tot i que seria un estudi interessant de fer, i convidem que en futurs projectes s'estudiï aquest tema amb més profunditat.

D'aquestes persones detingudes enregistrades per delicte, la distribució per sexe és la següent:

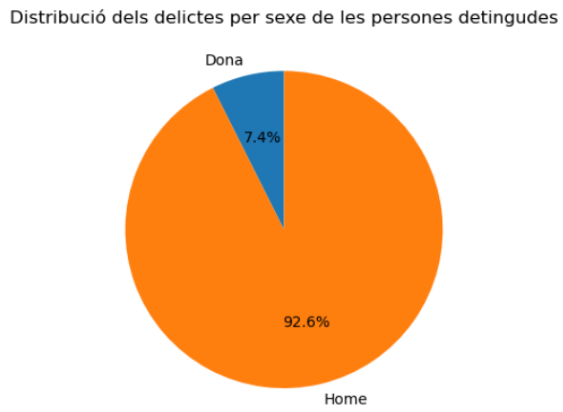

Figura 5.17: Distribució de les persones detingudes per sexe. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes. Autor: Sergi Payarol.

Com es pot veure, el 92,6% dels individus arrestats són homes, mentre que el 7,4% restant són dones. S'ha de tenir present que es parla de persones detingudes enregistrades entre el desembre de 2018 i el 2022. És a dir, quatre anys d'informació. En el cas de les dones, hi ha 106 delictes on l'autor dels fets és una dona. Per tant, és una mostra massa petita com per extraure conclusions sòlides.

A partir d'aquí, analitzarem dues taules: Una taula de persones detingudes úniques, que s'ha fusionat amb la taula de delictes únics; i una segona taula que s'ha fusionat amb la taula de delictes totals tenint en compte les diferents tipologies de delicte.

Si observem la distribució dels delictes, segons la seva tipologia i sexe dels individus arrestats es pot veure el següent:

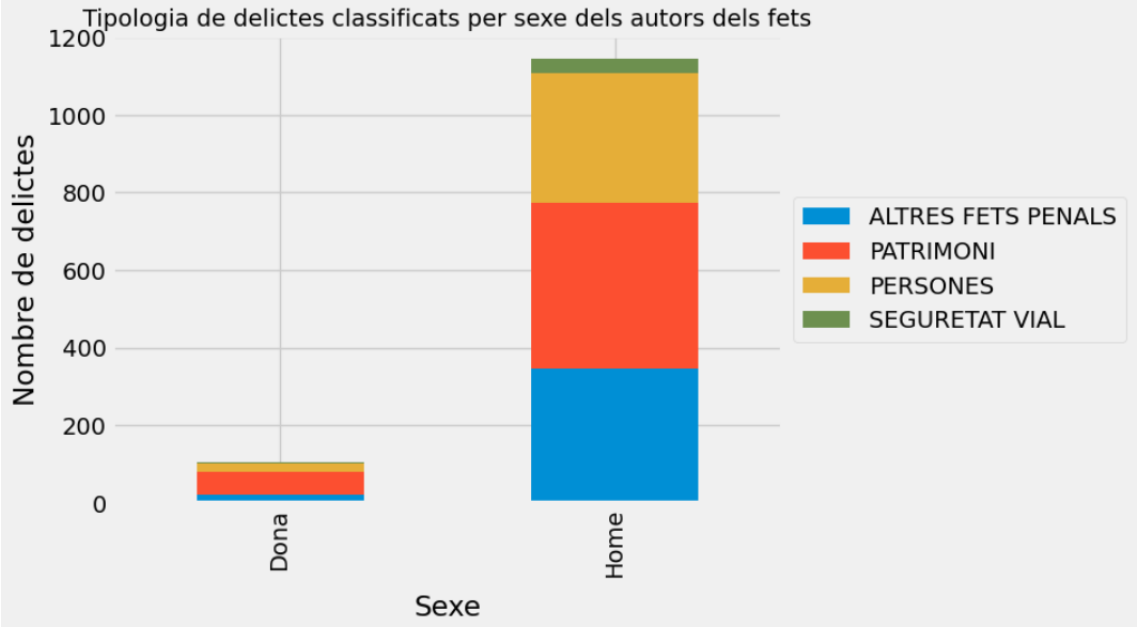

Figura 5.18: Tipologia dels delictes per sexe dels autors dels fets. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes amb tots els càrrecs i tipus de delicte. Autor: Sergi Payarol.

En primer lloc, veiem com els delictes contra la seguretat vial són força escassos. En els homes, la distribució entre els delictes contra el patrimoni, les persones i altres fets penals és força equilibrada. Mentre que en les dones, els delictes contra el patrimoni són una mica més elevats que la resta.

Si s'afina un grau més en la tipologia dels delictes per sexe s'obtenen les següents gràfiques.

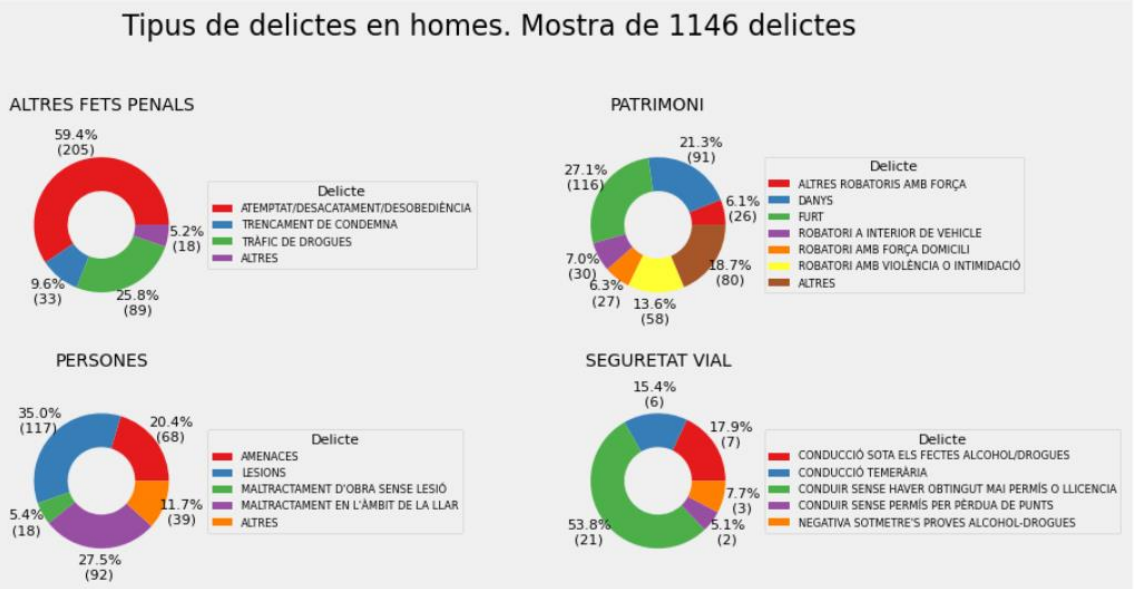

Figura 5.19: Tipologia dels delictes en homes. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes amb tots els càrrecs i tipus de delicte. Autor: Sergi Payarol.

En la composició de gràfiques circulars de sobre, hi ha la distribució dels delictes únics on l'autor dels fets ha estat un home. Es pot veure els tipus de delicte segons el seu grup en valors proporcionals i absoluts.

Altres fets penals

Amb un 59,4% l'atemptat, desacatament o desobediència són els més presents. Tot i que destaca també el 25,8% de tràfic de drogues.

Patrimoni

Al que respecta els delictes contra el patrimoni, hi ha un equilibri entre les categories. No obstant, es pot diferenciar els furts i els danys com els més presents.

Persones

Les lesions és la categoria més present en els delictes contra les persones amb un 35%. El maltractament en l'àmbit de la llar és la següent categoria amb un 27,5%.

Seguretat vial

Prop del 54% dels delinqüents han estat persones detingudes per conduir sense haver obtingut mai permís o llicència de conduir.

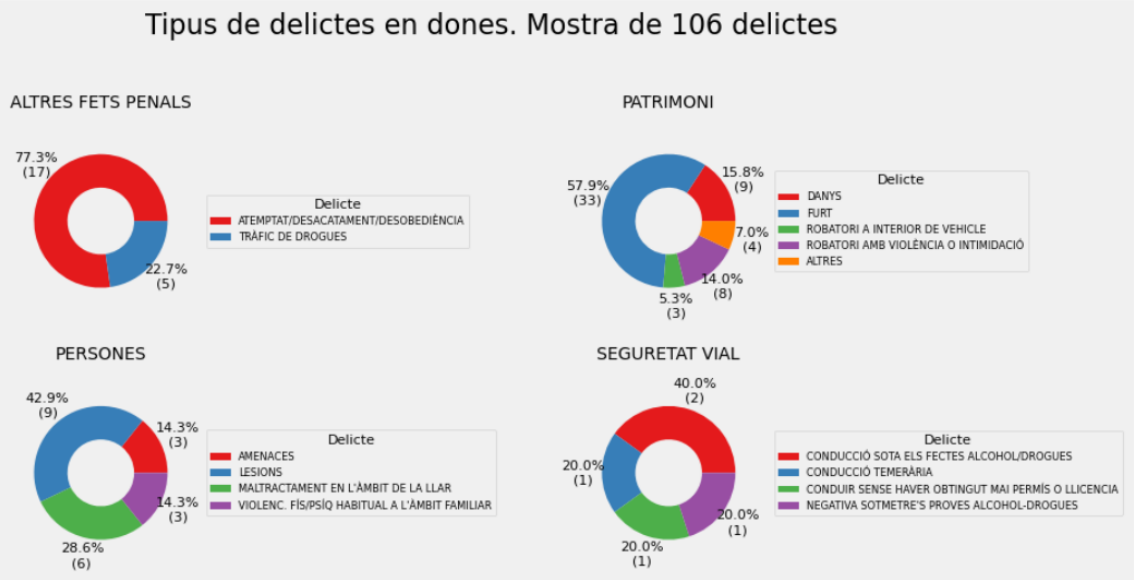

Figura 5.20: Tipologia dels delictes en dones. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes amb tots els càrrecs i tipus de delicte. Autor: Sergi Payarol.

Pel que respecta a les dones, la mostra és molt petita (106 registres). La seva distribució és la següent:

Altres fets penals

Amb un 77,3% l'atemptat, desacatament o desobediència és el tipus de delicte més present.

Patrimoni

Al que respecta els delictes contra el patrimoni, el 58% dels registres són per furts.

Persones

Les lesions conformen la categoria amb més registres, amb un 42,9%.

Seguretat vial

El 40% dels registres s'han realitzat per una conducció sota els efectes de l'alcohol o les drogues.

És important remarcar que aquesta mostra és massa petita per extreure conclusions que generalitzin la tipologia de delictes en dones.

A continuació, es mostra com es distribueixen els delictes tenint en compte l'edat de les persones detingudes.

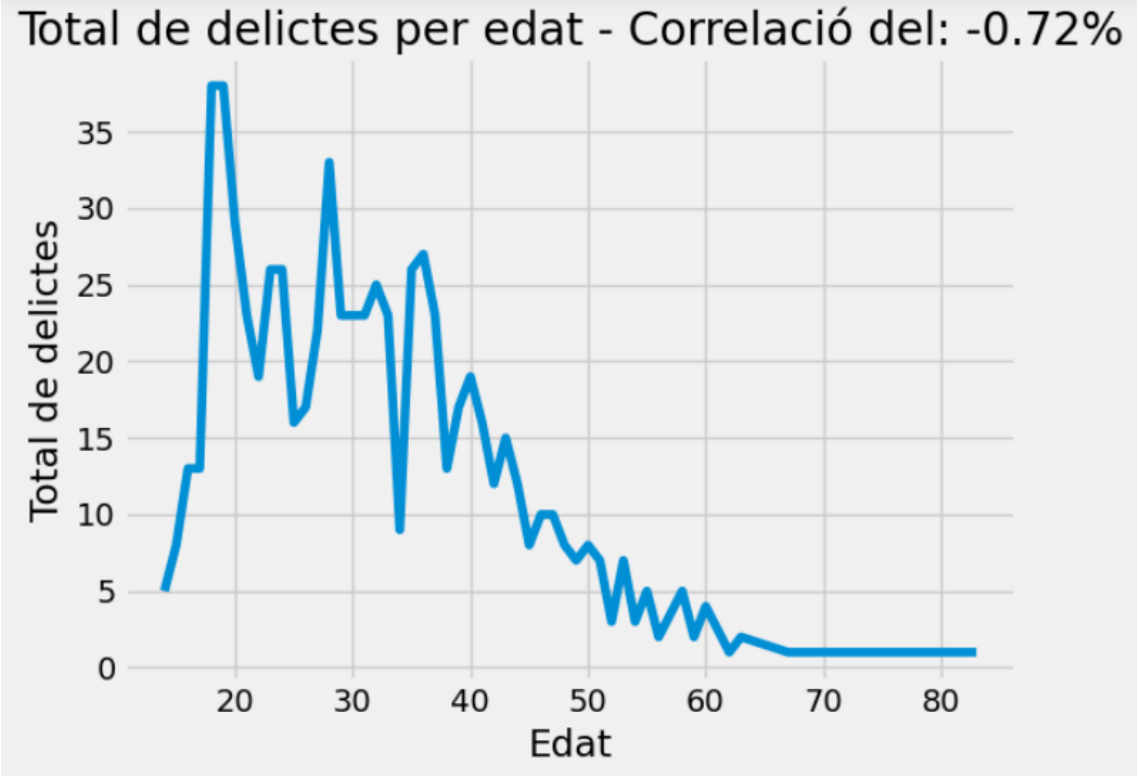

Figura 5.21: Correlació entre l'edat de les persones detingudes i el total de delictes. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes. Autor: Sergi Payarol.

La majoria de les persones detingudes solen tenir entre 20 i 30 anys. Mentre que els individus arrestats d'edat més avançada són més rars. Per tant, es pot veure que els joves són els que més delinqueixen i, a mesura que avancen en edat, la proporció de delinqüents és menor. De fet, hi ha una correlació negativa del -0.72% que ho corrobora.

A continuació, es mostra com és la distribució de l'edat, segons el sexe dels autors dels fets.

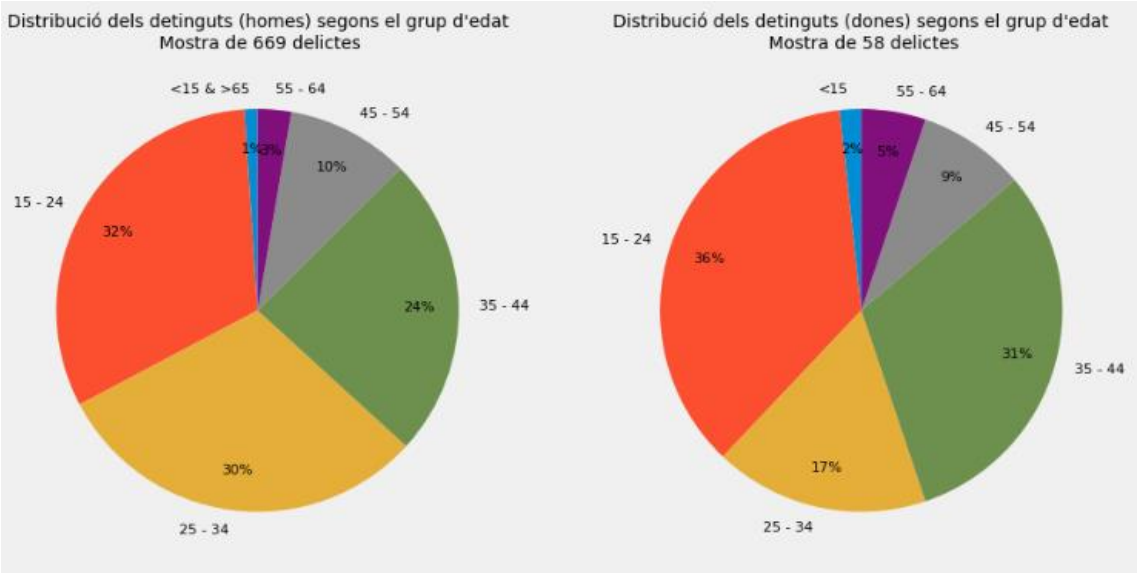

Figura 5.22: Distribució del nombre de persones detingudes per rangs d'edat i sexe. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes úniques. Autor: Sergi Payarol.

Es pot observar com la distribució sol ser força igualada entre l'edat de les persones detingudes segons el seu sexe. La franja d'homes d'entre 25 i 34 anys té una proporció més elevada que la de les dones, mentre que aquestes tenen una proporció més gran de delinqüents entre els 15 i 24 anys i els 35 i 44 anys.

A continuació, s'analitza com es distribueix l'edat dels individus arrestats segons la tipologia de delicte.

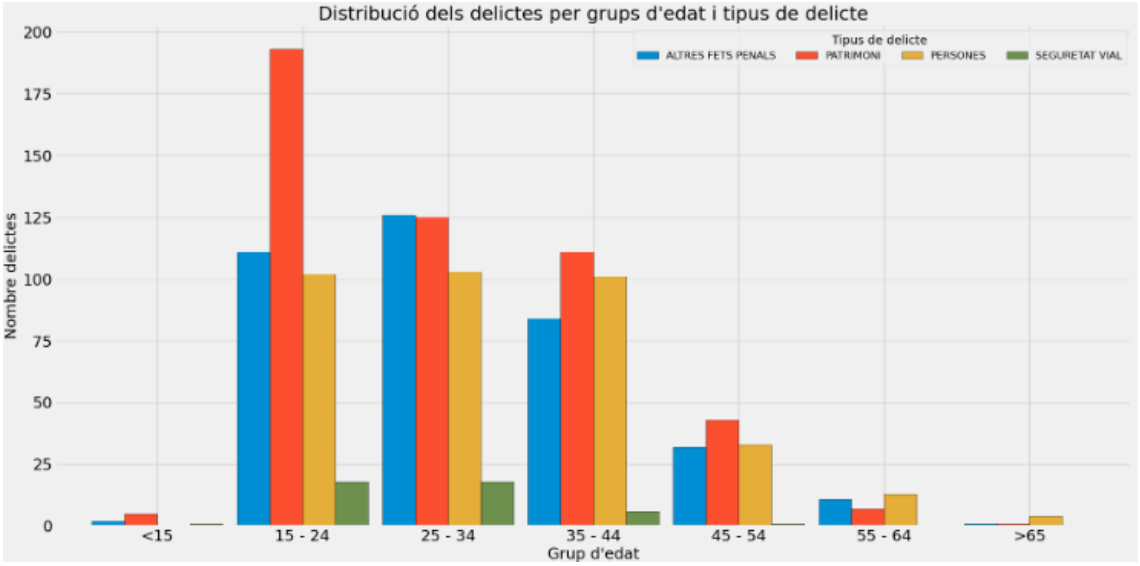

Figura 5.23: Distribució del nombre de persones detingudes per rangs d'edat i tipus de delicte. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes amb tots els càrrecs i tipus de delicte. Autor: Sergi Payarol.

És interessant veure com els delictes contra el patrimoni són força més elevats entre els individus arrestats de 15 a 24 anys. I amb l'edat, aquests baixen. També s'observa com el delicte contra les persones es manté estable en persones detingudes que tenen entre 15 i 44 anys, fins que a partir dels 45 cau en picat. Pel que respecta els altres fets penals, es veu un repunt en individus arrestats d'entre 25 i 34 anys. Per últim, els delictes contra la seguretat vial també es donen més entre persones detingudes de la mateixa franja d'edat.

A continuació, s'analitzarà la taula de contingència "Subtipus de delictes per franges d'edat dels autors dels fets" que es troba a l'apartat "6. Anàlisi de les persones detingudes: Qui comet els delictes?" de l'informe tècnic "Anàlisi Exploratòria de les dades".

A tall de conclusió, es pot destacar el següent:

Aquesta taula de contingència mostra la distribució dels subtipus de delictes entre els diferents rangs d'edat.

<15

Entre els menors de 15 anys, els delictes més presents són per tràfic de drogues, robatori de vehicle, robatori amb violència o intimidació, danys, conduir sense permís i atemptat, desacatament o desobediència.

15-24

En aquest grup d'edat, els delictes més presents són per tràfic de drogues, robatori amb violència o intimidació, maltractament en l'àmbit de la llar, lesions, furt, danys, conduir sense permís, atemptat, desacatament o desobediència i amenaces.

25-34

En aquest grup d'edat, els delictes més presents són per tràfic de drogues, danys, maltractament en l'àmbit de la llar, lesions, furt, danys, atemptat, desacatament o desobediència i amenaces.

35-44

En aquest grup d'edat, els delictes més presents són per tràfic de drogues, maltractament en l'àmbit de la llar, lesions, furt, danys, atemptat, desacatament o desobediència i amenaces.

45-54

En aquest grup d'edat, els delictes més presents són per tràfic de drogues, maltractament en l'àmbit de la llar, lesions, furt, danys, atemptat, desacatament o desobediència i altres robatoris amb força.

55-64

En aquest grup d'edat, els delictes més presents són per tràfic de drogues, maltractament en l'àmbit de la llar, lesions, furt i atemptat, desacatament o desobediència.

>65

Dins d'aquest grup d'edat, els delictes més presents són per violència física i/o psicològica habitual a l'àmbit familiar, maltractament en l'àmbit de la llar, lesions, danys i amenaces.

Així doncs, es pot veure que a mesura que s'avança en edat els següents delictes augmenten:

- Violència física i/o psicològica habitual a l'àmbit familiar (Sobretot els majors de 65 anys)
- Maltractament en l'àmbit de la llar
- Lesions
- Furt (No hi ha registres de persones detingudes majors de 65 anys), per aquest tipus de delicte, veiem que els individus arrestats entre 35 i 44 anys són on tenen una proporció més elevada.

Una vegada s'ha analitzat el sexe i l'edat de les persones detingudes, només queda per analitzar la seva procedència. Així doncs, veiem com es distribueixen les nacionalitats dels autors dels fets.

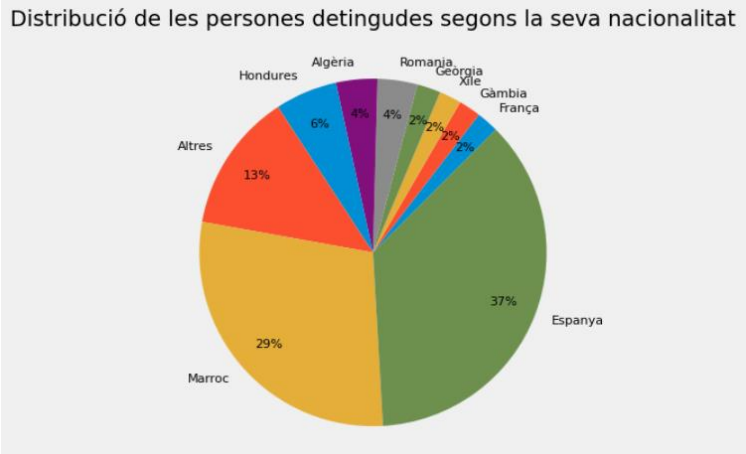

Figura 5.24: Proporció de les persones detingudes per nacionalitat. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes úniques. Autor: Sergi Payarol.

Es pot veure clarament com els individus arrestats de procedència espanyola són els majoritaris, tot i que els de procedència marroquina se situen en segona posició, amb un 29% del total de delictes. A partir d'aquesta gràfica sorgeix una pregunta: quin pes té cada nacionalitat tenint en compte la proporció de persones segons origen que resideixen a Girona? Seria interessant de veure aquesta distribució i es convida a que en posteriors anàlisis s'aprofundeixi en aquest aspecte.

Si ara observem aquesta distribució tenint en compte el sexe de les persones detingudes, s'observa el següent:

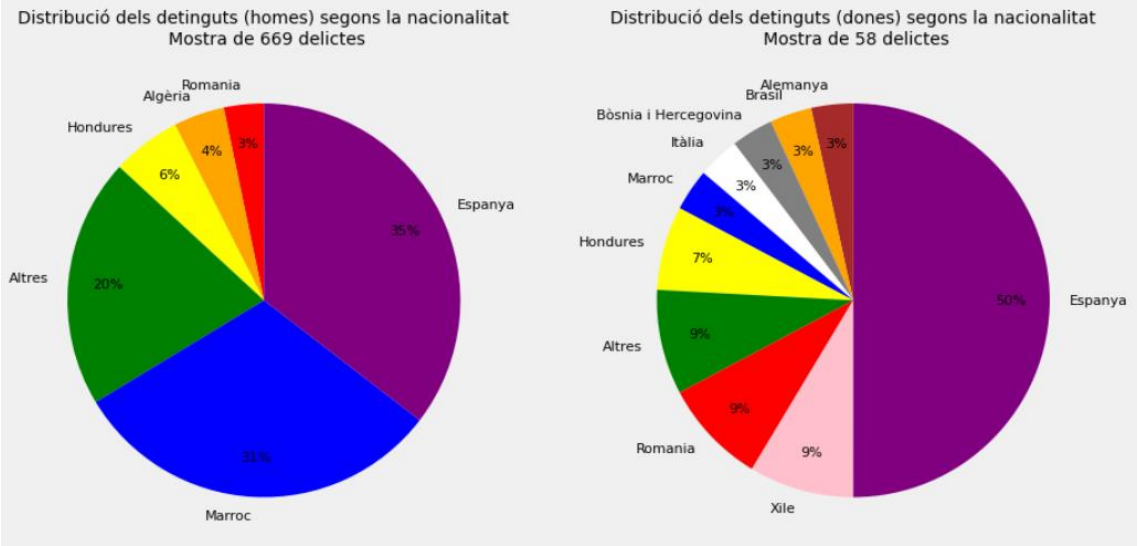

Figura 5.25: Proporció de persones detingudes per nacionalitat i sexe. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes úniques. Autor: Sergi Payarol.

La proporció d'homes d'origen espanyol i marroquí engloba pràcticament el 66%. Mentre que entre les dones, el 50% és d'origen espanyol, tot i que també es pot destacar la presència de xilenes i romaneses que engloben un 18% de la mostra. A continuació, es mostra la distribució dels delictes segons rang d'edat i nacionalitat dels autors dels fets.

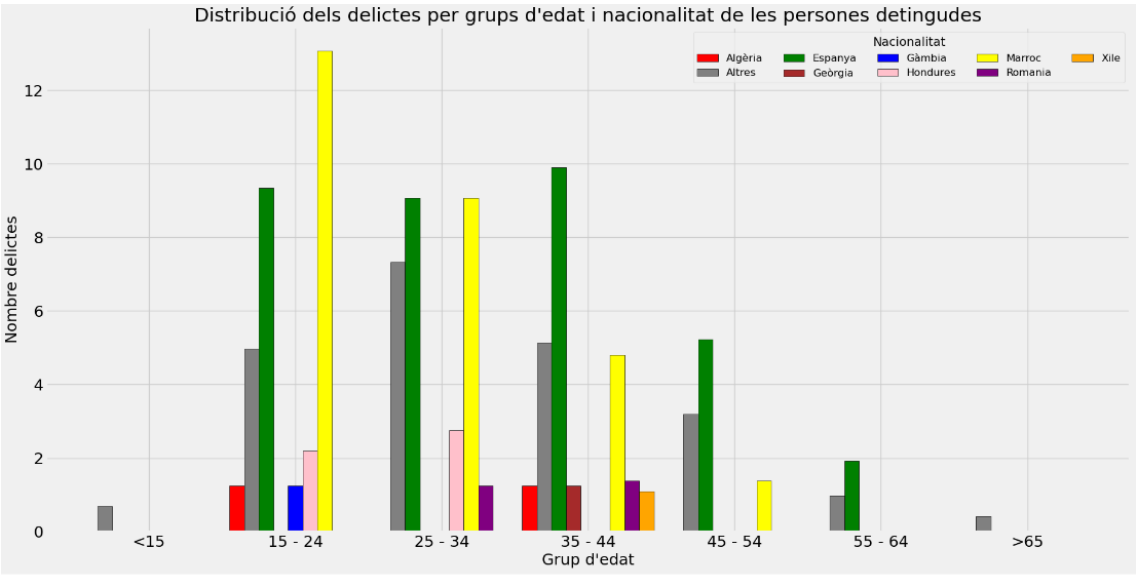

Figura 5.26: Proporció de persones detingudes per rangs d'edat i nacionalitat dels individus arrestats. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes úniques. Autor: Sergi Payarol.

Observem clarament, com la majoria dels individus arrestats d'origen marroquí es concentren entre els 15 i 24 anys. Mentre que els espanyols estan més balancejats entre els rangs d'edat d'entre 15 – 24 i 35 - 44 anys.

Una vegada s'ha fet una anàlisi general sobre les persones detingudes, es procedirà a fer una aproximació temporal.

5.2.4.1 Aproximació temporal

En aquest subapartat analitzarem l'evolució temporal dels fets delictius, tenint en compte els autors dels fets. A continuació, es mostren dues sèries temporals dividides per sexe dels arrestats.

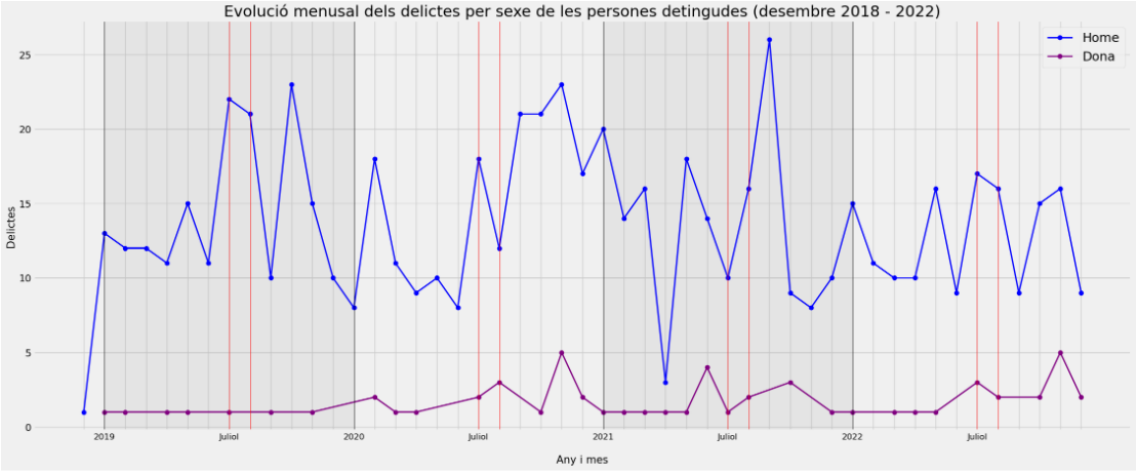

Figura 5.27: Sèries temporals del nombre de delictes segons les persones detingudes (homes i dones) per mesos. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes. Autor: Sergi Payarol.

En la gràfica de sobre es pot veure l'evolució dels delictes comesos per homes i dones. A primera vista, s'aprecia que la sèrie de les dones queda força silenciada per la dels homes. S'ha de pensar que la mostra de les dones és de 106 observacions, i per tant, no és gaire representativa.

En canvi, entre els homes, es pot veure com segueix una certa estacionalitat, on els pics es concentren al setembre i a l'octubre. S'ha de pensar que aquesta sèrie temporal va del desembre del 2018 fins al 2022, amb el qual la pandèmia es troba ben enmig. És important de tenir en compte aquest fet, ja que fa que la mostra pugui no ser gaire representativa d'una sèrie que englobi més anys. Tanmateix, són les úniques dades proporcionades per la Policia Municipal.

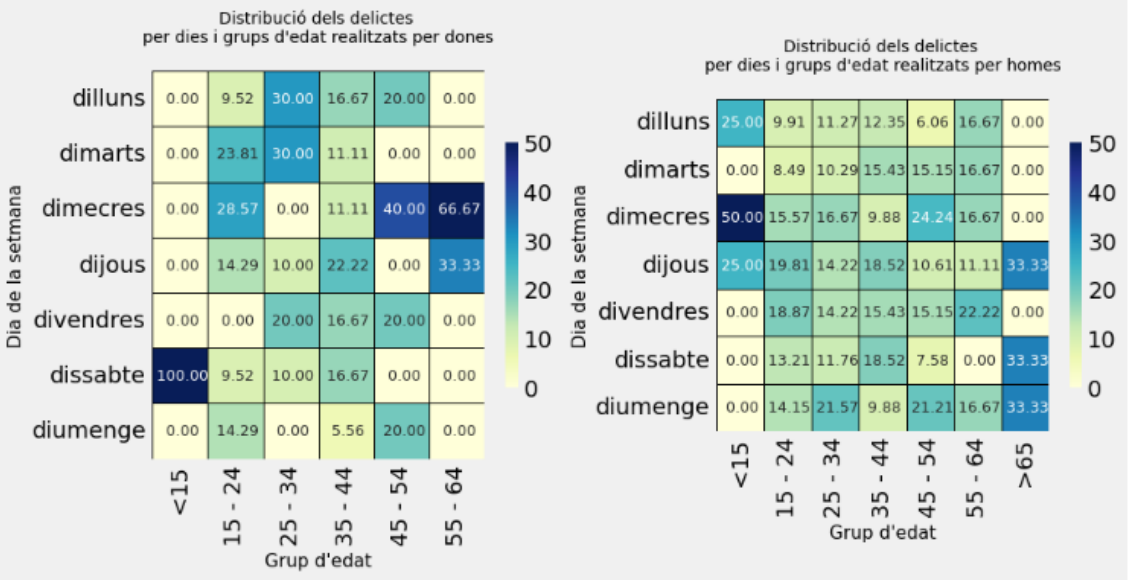

Figura 5.28: Taules de contingència sobre la distribució dels delictes per dies i grups d'edat segons el sexe. Per realitzar aquestes taules s'ha utilitzat la taula de persones detingudes. Autor: Sergi Payarol.

A les taules de contingència de sobre es pot veure la distribució de persones detingudes per grups d'edat, sexe i dies de la setmana.

Pel que respecta les dones, veiem una agrupació important de delictes en les de 45-54 anys els dimecres. Els delictes per menors de 15 anys es concentren el dissabte. Els delictes entre les franges de 15-24 i 25-34 es distribueixen entre el dilluns i el dimecres. Mentre que en la franja de 55-64 s'agrupen en el dimecres i el dijous.

Pel que fa als homes, destaquen dos grups: un primer grup on els delictes es distribueixen més homogèniament entre el dimecres i el diumenge per persones entre 15-24, 25-34 i 35-44. I un segon grup de delictes més concentrats, com per exemple els delictes dels dilluns, dimecres i dijous per menors de 15 anys; els delictes dels dimecres per persones de 45-54 anys; els delictes del divendres per persones de 55-64 anys; els delictes ben concentrats el dijous, dissabte i diumenge per persones majors de 65 anys. D'altra banda, observem que el diumenge és un dels dies que té unes proporcions més elevades en diferents franges d'edat.

Una pregunta que sorgeix en veure aquestes taules de contingència, és si la mostra que s'ha agafat de persones detingudes té les mateixes característiques que s'ha extret de l'anàlisi de la taula de delictes. A primera vista, es poden veure algunes característiques semblants, però febles. És a dir, els divendres i dissabtes són els dies de la setmana amb major presència delictiva. També, que els joves d'entre 15 i 34 anys són els que més delinqueixen. Pel que respecta a les dones del rang d'edat 15-24 es concentren més els dimarts i dimecres, mentre que els homes es localitzen més de dimecres a divendres. D'altra banda, les dones que tenen entre 25 i 34 es concentren el dilluns i el dimarts.

Mentre que els homes de 25-34 anys es distribueixen més entre el dimecres i el diumenge.

A continuació, es mostra la distribució dels delictes per hores dels dies segons el sexe de les persones detingudes.

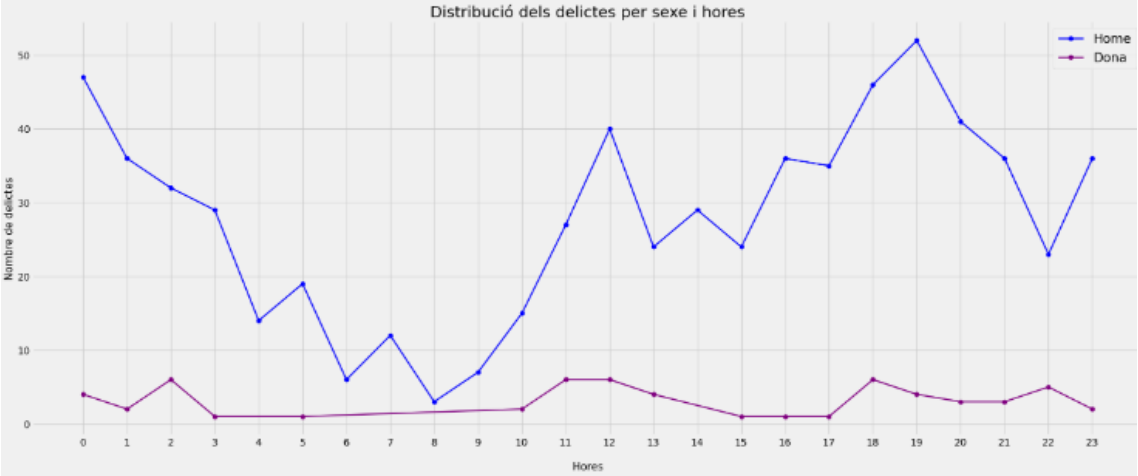

Figura 5.29: Sèries temporals del nombre de delictes segons les persones detingudes (homes i dones) per hores. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes. Autor: Sergi Payarol.

La distribució dels delictes per hores i sexe dels autors dels fets té força semblança a les gràfiques d'anàlisi temporal que s'han fet sobre els delictes. Els homes acostumen a delinquir més cap a la tarda, vespre i nit, mentre que les dones es concentren més durant el migdia o la tarda.

Si es discretitzen les hores per franges horàries s'obté el següent:

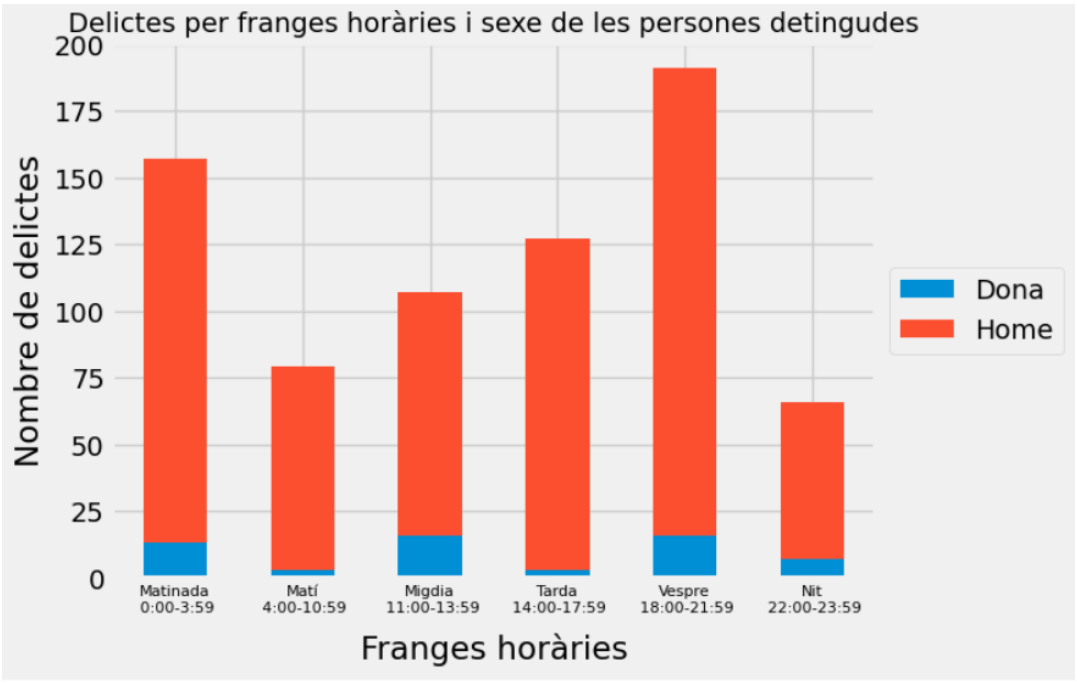

Figura 5.30: Delictes de persones detingudes per franges horàries i sexe. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes úniques. Autor: Sergi Payarol.

El vespre continua sent quan més delictes es cometen. La matinada se situa en segona posició. A l'anàlisi temporal dels delictes que s'ha realitzat anteriorment, la matinada estava força per sota i el matí es posicionava en segona posició. Es podria concloure, que en aquests darrers anys, hi ha hagut una tendència a la baixa en els delictes que es cometen al matí, amb una pujada dels que es cometen durant la matinada.

A continuació, es farà la distribució per franges horàries i per nacionalitat de les persones detingudes.

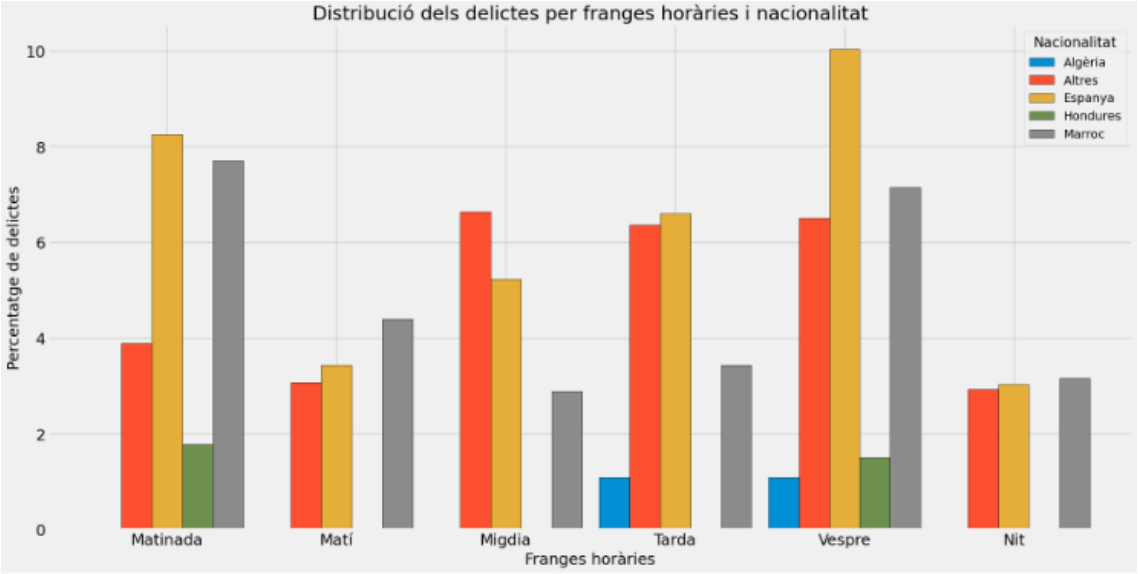

Figura 5.31: Distribució dels delictes de persones detingudes per franges horàries i nacionalitat. Per realitzar aquesta gràfica s'ha utilitzat la taula de persones detingudes úniques. Autor: Sergi Payarol.

En la gràfica de sobre es mostra la proporció de delictes tenint en compte la nacionalitat de persones arrestades en les diferents franges horàries. Es pot observar com els espanyols i els marroquins tendeixen a delinquir més durant el vespre i la matinada. Si afinem encara més en l'anàlisi, es pot veure com la distribució dels espanyols presenta un biaix negatiu (sense tenir en compte els delictes de la matinada, ni la nit). És a dir, des del matí cap al vespre hi ha una tendència positiva clara. D'altra banda, la distribució dels delictes comesos per persones d'origen marroquí presenta una distribució bimodal (si no tenim en compte la nit). És a dir, hi ha dos pics ben marcats en els extrems de la gràfica, mentre que la part central cau formant una "u". Aquest fet ens indica dues coses: la primera és que els espanyols delinqueixen més durant el dia, i amb més freqüència al vespre. La segona és que els marroquins delinqueixen més quan sol ser més fosc (matinada i vespre).

A continuació, s'analitzarà la distribució espacial dels individus arrestats.

5.2.4.2 Aproximació espacial

En aquest subapartat es farà una aproximació espacial sobre els delictes a nivell de subsectors tenint en compte les persones detingudes i la seva nacionalitat.

Tot seguit s'analitzarà la taula de contingència "Distribució dels delictes per subsectors policials i nacionalitat de la persona detinguda" de l'apartat "6.2 Aproximació espacial" de l'informe tècnic "Anàlisi exploratòria de les dades".

En aquesta taula de contingència es mostren les nacionalitats que més presència delictiva tenen distribuïdes per subsectors. Dins la categoria "Altres" s'han englobat totes aquelles persones detingudes on la seva proporció és menor d'1. A continuació, es mostra un resum del que es pot extreure a la taula de contingència, tenint en compte el país d'origen dels individus arrestats.

Algèria

Són presents a Santa Eugènia, Sant Feliu catedral, Parc central, Eixample Sud, Montilivi - Creueta i Ajuntament - Rambles

Espanya

Es localitzen a Santa Eugènia, Sant Narcís Nord, Parc Central, Mercadal, Mas Xirgu, Eixample Sud, Devesa i Can Gibert del Pla.

Hondures

Delinqueixen més a Santa Eugènia, Sant Narcís Nord i Sud, Palau I i Can Gibert del Pla.

Marroc

Són present a Santa Eugènia, Sant Feliu Catedral, Parc Central, Eixample Sud, Devesa i Can Gibert del Pla.

Xile

Delinqueixen més a mas Gri, Fontajau, Eixample Sud i Casernes.

Com s'ha pogut veure, els subsectors que més han sortit, com per exemple Santa Eugènia, Sant Narcís i el Mercadal, corresponen als subsectors on més presència delictiva hi ha, i que s'ha analitzat en els primers apartats del projecte.

Una vegada s'ha analitzat amb cert grau de profunditat els individus arrestats, passarem a veure altres variables externes a aquests jocs de dades que s'ha vist. Per tant, s'intentarà respondre a la pregunta: per què es donen els delictes?

5.2.5 Més enllà de les dades analitzades: Per què es donen els delictes?

En aquest apartat s'intentarà trobar factors externs que ajudin a entendre millor la dinàmica delictiva d'aquests últims anys.

Un apunt que s'ha de tenir en compte és que en aquest projecte es treballa amb sectors i subsectors policials. Trobar variables que s'ajustin a les delimitacions sectorials que ha creat la policia no ha estat possible. És per aquest motiu que s'utilitzaran els barris oficials de Girona i l'àmbit municipal per fer l'anàlisi multivariant.

Les variables utilitzades per aquest apartat han estat la densitat i el total de població per barri. L'escala temporal d'aquestes variables és anual. També s'han utilitzat variables meteorològiques, per veure si es dóna algun tipus de relació amb els fets delictius. Per aquestes últimes, l'escala que s'ha fet servir és a escala municipal de Girona i són dades recollides mensualment. Es recomana que es mirin els mapes interactius de l'apartat "7. Més enllà de les dades analitzades: Per què es donen els delictes?" de l'informe tècnic d'EDA, per així poder veure amb més deteniment el perquè de l'elecció dels barris.

Així doncs, es procedeix a analitzar primer la densitat i el total de població per barri.

5.2.5.1 Densitat i total de població per barris

El primer que s'ha realitzat és una agrupació de nombre de delictes per barri. Una vegada s'ha realitzat aquest pas, s'han afegit les variables de "Densitat neta de població", "Distribució de la població en valors absoluts i de percentatge" i s'ha calculat l'índex de criminalitat per cada 1.000 habitants.

La densitat neta i la distribució de població provenen de fonts oficials. En el cas de la densitat neta, s'ha calculat en base a la superfície urbana construïda i no sobre el total de superfície del barri.

Per l'índex de criminalitat s'ha seguit la fórmula oficial que fa servir el Ministeri de l'Interior:

$$\frac{Delictes\ totals}{Població\ total} * 1000$$

A continuació, s'analitzarà la distribució d'aquestes variables per barris.

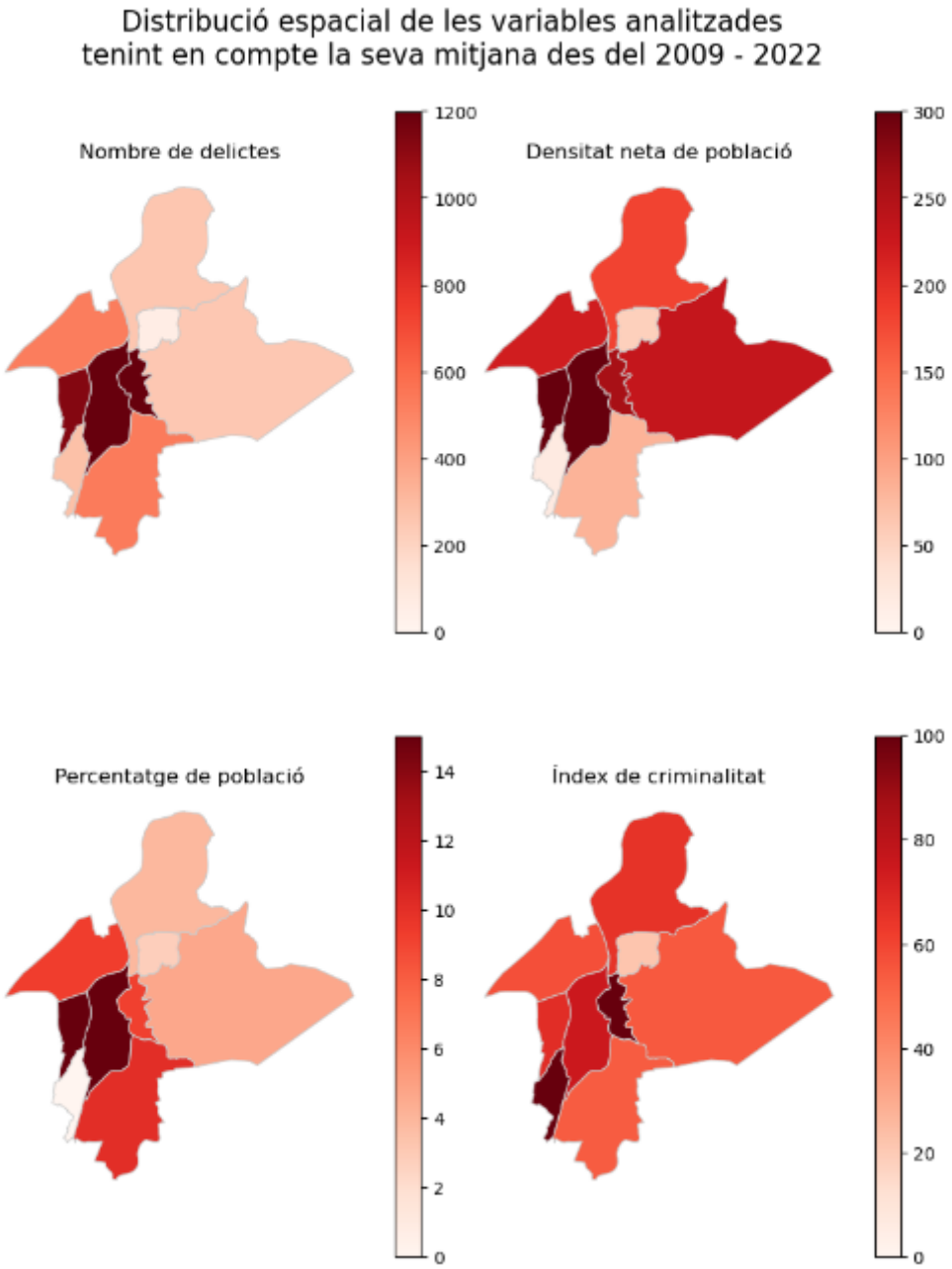

Figura 5.32: Composició de mapes on es mostra la distribució de les diferents variables analitzades en els barris de Girona. Autor: Sergi Payarol.

En la composició de mapes 5.32 tenim representades la distribució de les variables analitzades.

D'una banda, la mitjana de delictes se situa força per sobre dels 1.000 delictes en els barris de l'Eixample, Santa Eugènia i el Centre. També s'aprecia com el barri Oest i Sud no es queden enrere. Com a recordatori, el *hot spot* es concentra en aquests barris que s'han mencionat, sent l'Eixample i el Centre els barris nucli del *hot spot*. Per la seva part, Santa Eugènia se situa a la perifèria més propera i els dos restants en la perifèria més allunyada.

La densitat neta de població ha estat calculada en relació al nombre d'habitants i la superfície de sòl urbà destinada a ús residencial. Si ens fixem en el barri de l'Est, la seva superfície total de l'àrea és una de les més grans. No obstant, si analitzem la seva densitat neta de població veiem que està pràcticament igualada a la del barri del Centre. Per tant, tot i que la superfície en Km^2 sigui diferent, la superfície de sòl urbà destinada a ús residencial és força semblant (Centre: 0,35 Km^2, Est: 0,2 Km^2 el 2022).

Així doncs, es pot veure com Santa Eugènia és el barri amb una densitat més elevada. L'Eixample, el barri de l'Oest i el del Centre van a la cua. Per la seva part, el percentatge de població segueix una dinàmica força semblant. Tot i que s'ha de remarcar que el barri de l'Eixample engloba pràcticament el 44% dels gironins i Santa Eugènia el 17% al 2022. Si s'observa la seva mitjana durant aquests últims anys, es pot veure que no ha variat gaire.

Per últim, hi ha l'índex de criminalitat. En primer lloc, s'ha d'excloure Mas Xirgu, atès que no és representatiu en ser una zona de serveis i equipaments on viuen només 24 habitants censats. El problema rau en el fet que en haver-hi tan poques persones residint, el nombre de delictes d'aproximadament 320 de mitjana queda molt per sobre del total de residents. És per aquest fet que l'índex de criminalitat queda esbiaixat. S'ha de tenir en compte que en aquest barri és on es concentren algunes discoteques, molts supermercats i botigues i activitats d'oci que podrien explicar el fet d'aquest biaix.

En general, l'índex de criminalitat és força elevat als barris del Centre, l'Eixample i Santa Eugènia, sent el barri del Centre el que té l'índex més elevat amb 150 delictes de mitjana al mes per cada 1.000 habitants.

S'ha de remarcar que el barri del Centre és on hi ha més activitat comercial i d'oci, com també és on es poden trobar la majoria dels monuments i llocs històrics, ja que engloba els sectors oficials (no policials) del Barri Vell, el Mercadal i el Carme. Aquest fet pot derivar que aquest barri sigui el que tingui més afluència de persones, i de retruc fer augmentar el nombre de delictes, per la seva correlació tan alta, com bé podem veure a la taula 5.33.

	Total delictes
Densitat neta	0.53 %
Total població	0.95 %

Figura 5.33: Correlació de la densitat i total de població per barris amb el total de delictes. Autor: Sergi Payarol.

Si s'observen les correlacions entre les variables es destaca que el total de població té una correlació molt elevada amb el que respecta el total de delictes. Aquest fet ens indica que, a major nombre de població, més elevat serà el total de delictes. També hi ha una certa correlació entre la densitat neta i el total de delictes.

Si ens fixem ara en les taules de "Correlació de les diferents variables per barris" de l'apartat "7.1 Densitat i total de població per barris" que hi ha a l'informe "EDA", es poden extreure les següents conclusions, tot i que s'ha de tenir en compte que cada mostra de barri conté només 14 observacions i, per tant, no són gaire representatives:

- **Centre**: Existeix una correlació negativa moderada entre la densitat neta i el total de delictes. És a dir, a mesura que augmenta la densitat, el nombre de casos baixa. En el cas de l'índex de criminalitat passa el mateix. No obstant, hem de tenir en compte que estem fent l'anàlisi amb 14 observacions. No es poden extreure conclusions molt sòlides al respecte. D'altra banda, sabem que el barri del Centre és on es localitza el *hot spot* principal de delictes. Aquí entren en joc altres factors que expliquen el perquè dels delictes. Un exemple podria ser que en aquest barri és on hi ha més zona comercial, activitats d'oci, com per exemple bars i restaurants. Això pot fer que el nombre de delictes sigui elevat, però també fa que la densitat de població, o el total de població quedin silenciats.
- **Eixample**: En aquest cas, hi ha correlacions positives. Pel que fa a la distribució de població amb el total de delictes, hi ha una correlació moderada del 40%. És a dir, a mesura que augmenta el nombre de població, augmenta el nombre de delictes. El mateix passa amb la densitat neta. L'índex de criminalitat és força baix.
- **Est**: No s'observen correlacions gaire destacables.
- **Montjuïc**: Hi ha una correlació moderada entre la densitat de població, el total de delictes i l'índex de criminalitat. És a dir, a mesura que la densitat augmenta, el total de delictes i l'índex de criminalitat augmenten també.
- **Nord**: No s'observen correlacions gaire destacables.
- **Oest**: En primer lloc, s'aprecia una correlació moderada entre la distribució de població i el total de delictes. És a dir, a mesura que augmenta el nombre de persones, augmenten els delictes. D'altra banda, a mesura que la densitat de població augmenta, disminueix el nombre de delictes. D'altra banda, hi ha un possible error en les dades, ja que hi ha una correlació significativa entre la distribució de població i la densitat neta. És a dir, a mesura que augmenta el nombre de població disminueix la densitat. Això no hauria de ser així. Per tant, s'ha de vigilar amb les interpretacions.
- **Santa Eugènia**: Existeix una correlació moderada entre la densitat neta i el total de delictes i entre l'índex de criminalitat i la distribució de població.
- **Sud**: Hi ha una correlació moderada entre la distribució de població absoluta i l'índex de criminalitat i el total de delictes.

En la composició de gràfiques amb títol “Distribució temporal de les variables analitzades del 2009 al 2022” es pot veure la sèrie temporal anual de les variables analitzades, distribuïdes per barris.

D’aquestes gràfiques es poden extreure les següents observacions:

- **Centre**: Les diferents variables presenten una correlació serial. El confinament per la COVID va fer que augmentés el total de població i la densitat neta de població, mentre que els delictes van caure estrepitosament.
- **Eixample**: En els darrers anys de la COVID s’observa una pujada de delictes que va en relació amb el total de població i la densitat.
- **Est**: Hi ha una baixada contínua de la densitat i la distribució de població. No es pot observar una relació directa amb el total de delictes. Sí que es pot veure, que sobre el 2014 es produeix un repunt important de delictes.
- **Montjuïc**: Es percep una tendència lleugerament a l’alça del total de delictes que va amb relació amb la densitat de població. Però, després de la COVID, la densitat ha caigut en picat.
- **Nord**: Destaca un repunt important de casos el 2019 que pot estar relacionat amb la pujada de densitat.
- **Oest**: S’observa una tendència a l’alça sobre el nombre de casos. També, que la densitat cau estrepitosament, possiblement degut a un error de càlcul.
- **Santa Eugènia**: Hi ha un repunt de casos a partir del 2016 que pot estar en relació amb el total de població i la densitat.
- **Sud**: No es percep una correlació serial molt significativa entre el total de delictes i el total de població i densitat.

En el següent apartat s’analitzaran les variables meteorològiques, per així veure si hi ha algun tipus de relació amb el nombre de delictes.

5.2.5.2 Anàlisi municipal: Agrupació de variables meteorològiques

En aquest apartat es recullen les anàlisis correlatives sobre les variables meteorològiques i el total de delictes mensuals a escala de municipi. Aquestes variables s’han seleccionat a petició expressa de la Policia Municipal.

Les variables meteorològiques que s’han fet servir són les següents:

- Humitat relativa màxima mitjana
- Humitat relativa mitjana
- Precipitació acumulada
- Precipitació número dies
- Pressió atm mitjana
- Temperatura màxima mitjana

- Temperatura mitjana

Les correlacions entre les variables són les següents:

	Total de delictes
Humitat relativa màxima mitjana	0.25 %
Humitat relativa mitjana	0.35 %
Precipitació acumulada	0.10 %
Precipitació número dies	0.07 %
Pressió atm mitjana	-0.04 %
Temperatura màxima mitjana	-0.08 %
Temperatura mitjana	-0.07 %

Figura 5.34: Correlació entre les diferents variables meteorològiques i el total de delictes en l'àmbit municipal de Girona. Autor: Sergi Payarol.

En general, no es mostren unes correlacions molt significatives amb el total de delictes. Per tant, serà interessant utilitzar altres mecanismes per veure la relació entre les variables en l'etapa d'anàlisi predictiva.

5.2.6 Resum de l'anàlisi exploratòria

Una vegada acabada l'Anàlisi Exploratòria de les Dades, es poden extraure vàries conclusions:

- Els delictes contra el patrimoni són els que es donen més, sobretot el furt.
- Hi ha un increment important dels delictes en els mesos d'octubre, novembre, maig i abril.
- El divendres i el dissabte són els dies més problemàtics.
- El vespre i el matí són les franges amb més presència delictiva.
- Els subsectors del Mercadal, Eixample Nord, Devesa, Santa Eugènia, Sant Narcís, entre d'altres que es localitzen al centre del municipi, són on hi ha més presència de delictes.
- El sector 1 i el sector 2 són els que tenen un total de delictes més elevat respecte a la resta.
- Les persones detingudes d'origen espanyol i marroquí, sobretot joves d'entre 15 i 24 anys i 25 i 34 anys són els que més delinqueixen.
- La densitat de població i el total de població per barris, són variables importants que s'haurien d'analitzar amb major profunditat en posteriors anàlisis. No obstant, no podem obviar el fet que hi ha algunes zones de la ciutat més freqüentada per persones que ens visiten (compra i oci). Ben segur que aquest distorsiona la qüestió de població resident i densitats de població.

- Les variables meteorològiques presenten unes correlacions força febles respecte el total de delictes a escala de municipi. Tot i així, se seguiran analitzant aquestes variables en l'etapa d'anàlisi predictiva multivariant.

5.3 Anàlisi predictiva dels sectors policials i de l'àmbit municipal de Girona

Per aquest apartat ens centrarem en resumir tot el procés que s'ha realitzat en l'etapa d'anàlisi predictiva. A tall de resum, primer s'ha realitzat una anàlisi dels components i la valoració de les proves estadístiques. A continuació, s'ha realitzat la predicció dels delictes per sectors policials i per l'àmbit municipal de Girona. Hem de tenir en compte que pels sectors policials, s'han utilitzat models univariants. És a dir, s'ha emprat una sola variable per realitzar la predicció. Els delictes s'han agrupat en la mitjana de delictes per dies de cada mes de l'any. Per tant, l'objectiu serà predir la mitjana de delictes per dies dels següents sis mesos.

D'altra banda, s'ha realitzat una predicció multivariant. És a dir, s'han utilitzat variables predictores, en aquest cas factors meteorològics, per enriquir més el procés de predicció.

Així doncs, el que primer veurem són els resultats dels components i de les proves estadístiques i seguidament veurem la predicció univariant. Per tal de veure totes les gràfiques resultants de cada predicció, es pot consultar l'informe "Anàlisi predictiva", on surten els resultats per cada sector. A continuació, es recolliran els models que millors resultats han donat i es valoraran els seus indicadors.

5.3.1 Anàlisi per sèries temporals univariants per sectors policials

En aquest apartat es resumiran els resultats dels components i de les proves estadístiques i es mostraran els resultats dels indicadors dels models.

5.3.1.1 Resultats dels components i proves estadístiques

En aquest subapartat veurem els diferents resultats dels components per sectors policials i també sobre les proves estadístiques. Començarem per veure la variabilitat de les dades.

Variabilitat de les dades

Per tal d'observar la variabilitat dels delictes temporalment, s'ha calculat el coeficient de variació. Si ens fixem en les gràfiques que hi ha en l'informe tècnic "Anàlisi predictiva", i anem a l'apartat "4.2 Variabilitat de les dades per sectors" (a partir de la mitjana i sd) veurem el següent:

- **Sector 1**: El CV és 28,15%, per tant, un coeficient de variació força elevat, la qual cosa ens indica una variabilitat important. Tot i així, la sèrie es pot considerar homogènia perquè no supera el 30%.
- **Sector 2**: El CV és 17,25%. Es tracta d'un coeficient inferior a l'anterior i, per tant, hi ha menys variació i es pot afirmar que la sèrie és homogènia.
- **Sector 3**: El CV és 17,68%, de manera que podem afirmar el mateix que l'anterior.
- **Sector 4**: El CV és 22,27%, la qual cosa indica una dispersió més elevada, sobretot en els darrers anys de la sèrie. Tot i així, podem dir que és homogènia.
- **Sector 5**: El CV és 34,07%, que representa un coeficient de variació molt elevat. Aquest fet és degut al valor anòmal del 2014. Si no tinguéssim en compte aquest valor, el coeficient seria relativament baix. No obstant, hem d'acabar afirmant que la sèrie no és homogènia perquè supera el 30%.
- **Sector 6**: El CV és 26,17% i, per tant, és de volatilitat elevada, sobretot en els darrers anys de la sèrie. No obstant, hi ha homogeneïtat.
- **Sector 7**: El CV és 25,29%, és a dir, també de volatilitat elevada, sobretot en els últims anys de la sèrie. Tot i així, la sèrie és homogènia.
- **Sector 8**: El CV és 19,85%. El coeficient de variació mostra una dispersió constant i elevada al llarg de la sèrie. No obstant, la sèrie és homogènia.

La tendència

Pel que fa a la tendència per sectors policials s'obtenen els següents resultats:

Sector	Pendent	Valor p	Presència de tendència
Sector 1	-0.001	0.742	False
Sector 2	-0.002	0.117	False
Sector 3	0.002	0.025	True
Sector 4	0.007	0.000	True
Sector 5	-0.002	0.001	True
Sector 6	0.004	0.000	True
Sector 7	0.002	0.000	True
Sector 8	0.003	0.001	True

Figura 5.35: Resultats de tendència en els diferents sectors policials. Autor: Sergi Payarol.

Els resultats de la tendència s'han realitzat a partir de la mitjana de delictes únics per dies de cada mes.

A partir dels resultats anteriors podem concloure vàries afirmacions.

D'una banda, hi ha evidència estadística significativa en tots els sectors menys en el sector 1 i en el sector 2. Per tant, per aquests dos últims, no podem estar segurs que hi hagi presència de tendència i que, per tant, la tendència és plana.

D'altra banda, la resta de sectors sí que presenten una tendència significativa, estadísticament parlant. Hem de pensar que un valor de *p* inferior a 0,05 ens permet acceptar la hipòtesi nul·la i afirmar que hi ha tendència. Així doncs, s'observa

primerament que el sector 5 té una tendència negativa. Això significa que la sèrie presenta una disminució gradual en el temps. Aquest fet pot ser degut al valor atípic del 2014, que pot crear una lleugera elevació del pendent en negatiu.

La resta de sectors tenen una tendència positiva, la qual cosa implica que els delictes poden estar en augment.

Sèrie additiva o multiplicativa

Els resultats sobre si les sèries per sectors són additives o multiplicatives són els següents:

Sector	CVC	CVD	Additiu
Sector 1	7.21	0.28	True
Sector 2	9.36	0.17	True
Sector 3	9.60	0.18	True
Sector 4	9.71	0.22	True
Sector 5	6.82	0.34	True
Sector 6	6.78	0.26	True
Sector 7	6.54	0.25	True
Sector 8	9.06	0.20	True

Figura 5.36: Resultats de si la sèrie és additiva en els diferents sectors policials. Autor: Sergi Payarol.

Així doncs, es pot concloure que totes les sèries són additives. Aquest fet ens indica que els components de la sèrie, com la tendència, l'estacionalitat o els valors residuals s'hauran d'ajustar de manera additiva. És a dir, sumant els valors i no multiplicant com es faria en un model multiplicatiu.

Estacionalitat, cicles

Per aquesta secció es presentaran els resultats sobre la presència d'estacionalitat en les diferents sèries dels sectors policials. Per establir si hi ha estacionalitat a les sèries hem descompost cada sèrie pels seus components a partir de la funció *seasonal_decompose()*. Per tal de poder veure les gràfiques de cada sector, podeu dirigir-vos a l'apartat "4.5.1 Descomposició estacional per sectors" de l'informe "Anàlisi predictiva". Tot seguit es presenta un resum del que es veu a les gràfiques:

- **Sector 1**: No es percep una tendència clara, però sí que es veu l'estacionalitat cada 12 mesos. Pel que fa als residus, observem alguns valors atípics que es tindran en compte per l'etapa d'anàlisi predictiva.
- **Sector 2**: Hi ha una tendència negativa en els darrers anys. Els patrons estacionals anuals se centren enmig de l'any i observem una altra vegada una mica de soroll en les dades.
- **Sector 3**: No es percep una tendència clara. Hi ha una estacionalitat marcada anualment on es poden donar en períodes de cada 6 mesos aproximadament. S'observen residus força dispars.

- **Sector 4**: Hi ha una lleugera tendència positiva en aquests darrers anys. Una estacionalitat marcada, anual i amb períodes de 6 mesos aproximadament. Pel que fa als residus, es mostren força pròxims a la mitjana, tret del 2019 i el 2020.
- **Sector 5**: No hi ha una tendència gaire clara. L'anomalia del 2014 distorsiona en gran mesura els resultats.
- **Sector 6**: S'observa una tendència positiva en aquests darrers anys, una estacionalitat anual clara i uns residus que mostren força soroll en els dos extrems.
- **Sector 7**: No es percep una tendència clara. Hi ha una estacionalitat anual i uns residus força estables, tret del 2019.
- **Sector 8**: No hi ha una tendència clara, l'estacionalitat és anual i se centra sobretot al final de cada any.

A continuació, es mostren els resultats de l'Índex de Variància Estacional (IVE), per així poder identificar la variabilitat estacional de cada sèrie temporal.

Mesos	sector1	sector2	sector3	sector4	sector5	sector6	sector7	sector8
1	0.95	1.04	1.02	1.04	1.12	0.99	0.99	1.03
2	0.99	1.0	0.95	1.01	0.93	1.03	1.04	0.96
3	1.03	0.94	1.05	0.99	1.05	0.99	0.92	0.99
4	0.96	0.96	0.96	0.96	0.93	1.0	1.01	0.96
5	1.07	1.06	1.05	1.03	1.03	1.01	1.05	1.02
6	1.02	0.98	0.97	0.99	1.04	1.02	1.03	0.99
7	0.98	0.99	0.99	0.97	0.98	0.95	0.89	0.97
8	0.97	0.97	0.99	1.0	0.94	1.02	1.02	0.99
9	1.04	1.03	1.04	1.06	1.05	0.99	1.07	1.08
10	1.18	1.03	1.02	0.99	1.02	1.07	1.05	1.01
11	0.99	1.01	1.0	0.99	1.02	0.98	0.97	1.04
12	0.84	0.99	0.97	0.97	0.88	0.97	0.97	0.96

Figura 5.37: Resultats dels IVE en els diferents sectors policials per mesos de l'any. Autor: Sergi Payarol.

Tot seguit es mostren els resultats en diferents gràfiques de barres.

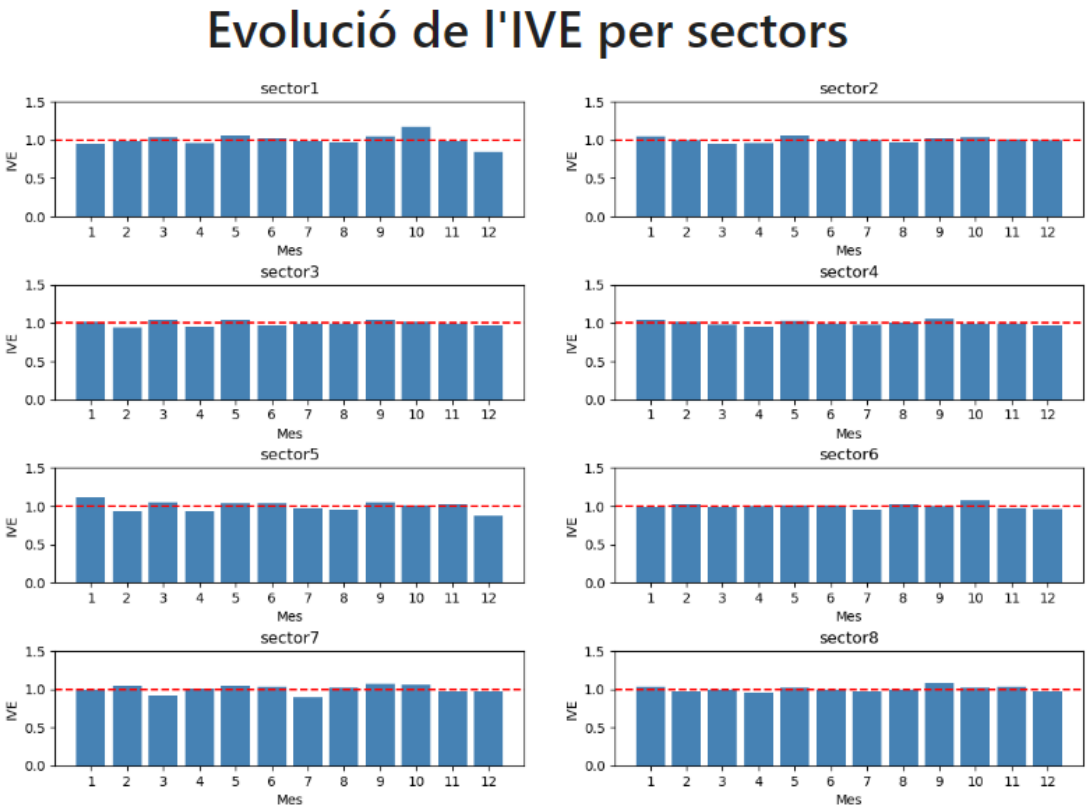

Figura 5.38: Composició de gràfiques on es mostra l'IVE per sectors policials. Autor: Sergi Payarol.

- **Sector 1**: Presenta un IVE lleugerament superior a 1 a gairebé tots els mesos, indicant una influència estacional significativa en aquest sector.
- **Sector 2**: Mostra un IVE proper a 1 en la majoria dels mesos, suggerint una influència estacional moderada.
- **Sector 3**: Exhibeix un IVE lleugerament superior a 1 en alguns mesos, reflectint una influència estacional rellevant en aquests períodes concrets.
- **Sector 4**: Té un IVE proper a 1 en tots els mesos, indicant una influència estacional consistent, però moderada.
- **Sector 5**: Mostra un IVE superior a 1 en alguns dels mesos, assenyalant una influència estacional moderada.
- **Sector 6**: Presenta un IVE proper a 1 en la majoria dels mesos, indicant una influència estacional moderada.
- **Sector 7**: Té un IVE superior a 1 en alguns dels mesos, amb el qual es demostra una influència estacional moderada.
- **Sector 8**: Exhibeix un IVE proper a 1 en alguns dels mesos, expressant una influència estacional moderada.

Autocorrelació

La correlació serial és una de les parts més importants dins del conjunt de proves prèvies a la predicció de les sèries temporals. L'autocorrelació ens indica si els valors successius de la sèrie estan relacionats entre si. Totes les observacions que estiguin per darrere dels *lags* aportaran informació útil per a predir els valors futurs.

Per aquest projecte s'han utilitzat les gràfiques ACF (*Autocorrelation Function*) i el PACF (*Partial Autocorrelation Function*). S'ha de tenir en compte que alguns dels resultats de les primeres gràfiques no correspondran amb els valors *p,d,q* i *P,D,Q* dels models. Això és degut a que en alguns dels models s'ha hagut d'iterar amb diferents valors per la part autoregressiva i de mitjana mòbil fins aconseguir un resultat acceptable. També, en alguns casos s'ha fet ús de l'"auto arima de pmdarima" (Smith, T., 2023), per veure si de manera automàtica es podien obtenir els millors paràmetres. No obstant, en cap cas, *auto_arima()* ens ha donat bons resultats i, per tant, l'hem extret de l'informe tècnic.

Sector	SARIMA		ARIMA	
	ACF/PACF	Final	ACF/PACF	Final
Sector 1	(1,1,1)(1,1,1,12)	(0,1,1)(2,1,2,12)	(1,1,1)	(0,1,1)
Sector 2	(1,0,1)(1,0,1,12)	(0,0,2)(1,1,1,12)	(1,0,1)	(1,1,1)
Sector 3	(1,0,1)(1,0,1,12)	(1,2,0)(2,2,2,12)	(1,0,1)	(2,2,1)
Sector 4	(1,0,1)(1,0,1,12)	(1,0,0)(0,1,1,12)	(1,0,1)	(2,2,1)
Sector 5	(1,0,1)(1,1,1,12)	(0,1,5)(0,1,1,12)	(1,0,1)	(1,1,1)
Sector 6	(1,0,1)(1,1,1,12)	(0,0,2)(3,1,1,12)	(1,0,1)	(1,2,1)
Sector 7	(1,1,1)(1,1,1,12)	(1,1,1)(2,1,1,12)	(1,1,1)	(2,1,4)
Sector 8	(1,1,1)(1,1,1,12)	(2,1,0)(2,1,0,12)	(1,1,1)	(1,1,1)

Figura 5.39: Paràmetres d'order() i seasonal_order() en els models SARIMA i ARIMA pels diferents sectors policials. Autor: Sergi Payarol.

A la taula 5.39 tenim el resum de tots els paràmetres *p,d,q* i *P,D,Q* dels models implementats amb SARIMA i *p,d,q* pels models ARIMA. Com es pot observar, la majoria de paràmetres inicials no corresponen als finals. Això és perquè s'han hagut de realitzar diverses iteracions per cada sector. A més a més, dins de cada comprovació s'ha verificat que els residus dels resultats del model no eren significants. És a dir, es trobaven per dins de l'interval de significança de la funció d'autocorrelació. Aquestes gràfiques de comprovació no s'han inclòs en l'informe ni es comentaran aquí, ja que si no s'allargaria massa aquest apartat. D'altra banda, hem cregut que no calia explicar la majoria dels sectors, si cap d'ells menys el sector 7 i el sector 8 han tingut bons resultats amb ARIMA i SARIMA.

A continuació, mostrarem els resultats de la gràfica ACF del model ARIMA del sector 7 i del model SARIMA del sector 8.

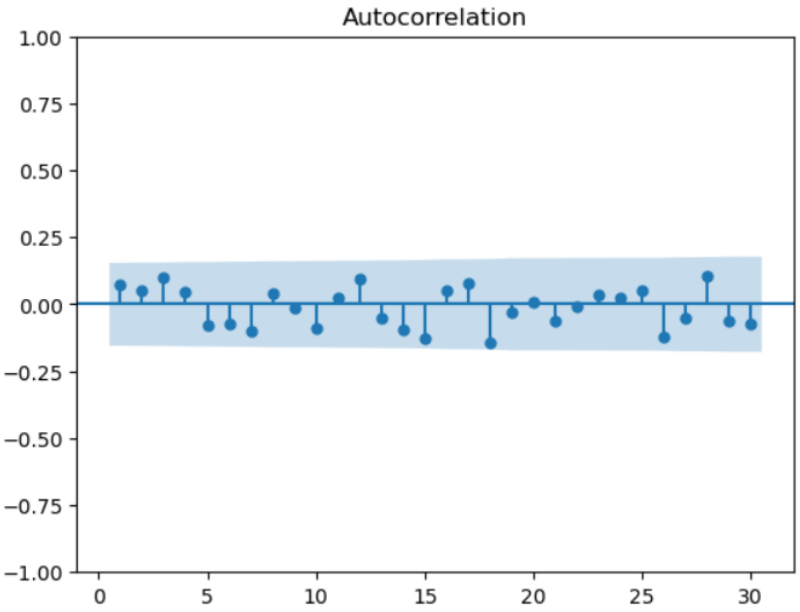

Figura 5.40: Resultats de l'autocorrelació dels residus del model ARIMA en el sector 7. Podem observar com els residus es troben per dins de l'interval de significança. Autor: Sergi Payarol.

A la primera gràfica 5.40 tenim l'autocorrelació dels residus dels resultats del model ARIMA del sector 7. Com podem veure, tots els *lags* es troben per dins de la banda de significança. Això vol dir que el model s'ha adaptat bé i que no s'ha d'extraure més informació útil dels errors.

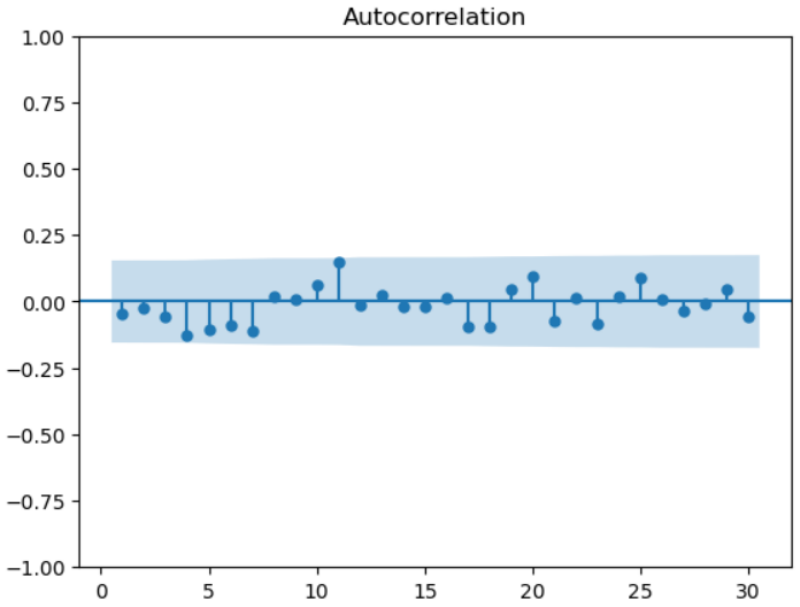

Figura 5.41: Resultats de l'autocorrelació dels residus del model SARIMA en el sector 8. Podem observar com els residus es troben per dins de l'interval de significança. Autor: Sergi Payarol.

En aquesta segona gràfica 5.41, es mostra l'autocorrelació dels residus dels resultats del model SARIMA del sector 8. Es veu com tots els *lags* tornen a estar dins de la banda de

significança, la qual cosa indica que els residus no són estadísticament significatius i que ja s'ha capturat tota la informació necessària.

Tot seguit mostrarem els resultats de l'estacionarietat.

Estacionarietat

Comprovar l'estacionarietat d'una sèrie també és una de les proves més importants abans de passar a l'etapa d'Anàlisi predictiva. L'estacionarietat d'una sèrie es dóna quan aquesta presenta unes propietats estadístiques que no canvien al llarg del temps. Com per exemple, que no hi ha una tendència molt marcada de creixement o decreixement, o que la variabilitat de les dades es manté constant. A continuació, s'analitzaran els resultats de la prova de Dickey Fuller augmentat.

Sector	Valor p	Estacionarietat
Sector 1	0.14	False
Sector 2	0.00	True
Sector 3	0.00	True
Sector 4	0.00	True
Sector 5	0.00	True
Sector 6	0.91	False
Sector 7	0.00	True
Sector 8	0.00	True

Figura 5.42: Resultats de l'estacionarietat pels diferents sectors policials. Autor: Sergi Payarol.

Tots els sectors, menys el sector 1 i el sector 6 tenen sèries estacionàries. Ens hem guiat pel valor de *p*, on aquest mostra un nivell de significança de 0,05. Si aquest és inferior a aquest llindar, podem rebutjar la hipòtesi nul·la i acceptar que la sèrie és estacionària. Així doncs, per l'etapa d'anàlisi predictiva, haurem de tenir en compte la no estacionarietat del sector 1 i del sector 6.

Heteroscedasticitat

Per la presència d'heteroscedasticitat, s'han utilitzat les proves de Breush-Pagan i White. Cal recordar que la presència d'heteroscedasticitat implica que la variància no és constant en els residus del model. És important tenir clar el grau d'heteroscedasticitat de les sèries, perquè, si es dóna el fet que els resultats de les proves són positives, haurem de treballar la sèrie fins que es pugui homogeneïtzar.

Sector	Valor p Pagan	Valor p White	Heteroscedasticitat
Sector 1	0.23	0.45	False
Sector 2	0.54	0.83	False
Sector 3	0.37	0.59	False
Sector 4	0.49	0.74	False
Sector 5	0.53	0.17	False
Sector 6	0.14	0.33	False
Sector 7	0.36	0.62	False
Sector 8	0.09	0.18	False

Figura 5.43: Resultats de la presència d'heteroscedasticitat pels diferents sectors policials. Autor: Sergi Payarol.

Els resultats ens indiquen que no hi ha evidència suficient com per afirmar la presència d'heteroscedasticitat a les sèries. Així doncs, podem concloure que les sèries temporals sobre els sectors policials tenen homoscedasticitat.

Presència d'anomalies

Per últim, hem localitzat les anomalies en les dades que poden ser problemàtiques a l'hora d'implementar els models. És per aquest fet que s'ha utilitzat el model *IsolationForest*. Aquest algorisme proporciona una puntuació d'anomalia per a cada observació, on una puntuació baixa indica presència d'anomalia, i una puntuació alta mostra que l'observació analitzada és normal. El seu funcionament és a partir dels arbres de decisió, utilitzant particions generades. El que es fa és "aïllar els valors atípics seleccionant aleatòriament una característica del conjunt de funcions donat i després seleccionant aleatòriament un valor dividit entre els valors màxim i mínim d'aquesta característica" (Dhiraj, K., 2018).

Es recomana consultar les gràfiques de l'apartat "4.9 Anomalies en les dades" de l'informe "Anàlisi predictiva". A continuació, mostrarem els anys on s'han detectat valors anòmals per sectors policials.

- **Sector 1**: El 2013 i el 2020.
- **Sector 2**: El 2016 i el 2020.
- **Sector 3**: El 2019 i el 2020.
- **Sector 4**: El 2019.
- **Sector 5**: El 2014.
- **Sector 6**: 2016 i el 2018.
- **Sector 7**: 2012 i el 2019.
- **Sector 8**: 2019 i el 2020.

5.3.1.2 Resultats dels models implementats en els sectors policials

En aquest subapartat presentarem els resultats dels indicadors dels millors models per cada sector. Hem de tenir en compte que (P) = Prophet, (A) = ARIMA i (S) = SARIMA.

Indicadors	S1(P)	S2(P)	S3(P)	S4(P)	S5(P)	S6(P)	S7(A)	S8(S)
MSE	0.28	0.33	0.07	0.14	0.01	0.11	0.04	0.12
MAE	0.46	0.48	0.22	0.32	0.09	0.27	0.16	0.25
RMSE	0.53	0.57	0.27	0.38	0.11	0.33	0.2	0.35
MAPE	0.09	0.13	0.09	0.1	0.09	0.15	0.16	0.06
CRMSE	0.45	0.57	0.21	0.38	0.11	0.33	0.2	0.33
R2	0.858	0.383	0.161	0.168	0.296	0.273	0.211	0.633
Valor_real1	3.58	2.9	2.26	3.9	1.03	1.23	0.94	3.06
Valor_predit1	4.12	3.77	2.7	3.6	1.03	1.68	0.99	3.27
Error_relatiu1	0.15	0.3	0.19	0.08	0.0	0.36	0.05	0.07
Valor_real2	3.9	3.71	2.71	2.97	1.06	2.1	1.19	2.81
Valor_predit2	3.77	3.89	2.66	3.29	0.95	2.05	1.14	2.87
Error_relatiu2	0.18	0.35	0.21	0.18	0.11	0.39	0.09	0.09
Valor_real3	4.3	3.63	2.67	3.2	1.2	2.43	0.77	3.6
Valor_predit3	4.22	4.13	2.74	3.26	1.07	1.89	1.06	3.69
Error_relatiu3	0.2	0.48	0.24	0.2	0.22	0.61	0.48	0.12
Valor_real4	7.06	4.26	3.23	2.84	1.06	2.19	0.87	4.26
Valor_predit4	6.44	4.23	3.11	3.38	1.05	1.92	1.07	3.53
Error_relatiu4	0.29	0.49	0.28	0.39	0.23	0.73	0.7	0.29
Valor_real5	7.0	4.7	2.57	3.87	1.07	2.1	1.47	4.07
Valor_predit5	6.18	4.21	2.84	3.26	1.22	2.11	1.12	4.05
Error_relatiu5	0.41	0.59	0.38	0.55	0.37	0.74	0.94	0.29
Valor_real6	4.87	5.1	2.48	3.13	0.77	1.77	1.03	4.29
Valor_predit6	4.31	4.27	2.88	3.23	0.91	2.04	1.03	3.91
Error_relatiu6	0.52	0.76	0.54	0.58	0.54	0.89	0.94	0.38

Figura 5.44: Resultats dels millors models pels diferents sectors policials. Autor: Sergi Payarol.

A la taula 5.44 hi ha representats els resultats dels indicadors del model per cadascun dels sectors policials. Per tal de tenir una visió general dels resultats dels models per sector, s'ha realitzat una sèrie de mapes on es recullen els resultats de la bondat d'ajust (R2), del percentatge d'error absolut mitjà (MAPE), de l'error absolut mitjà (MAE) i de l'error de l'arrel quadrada mitjana (RMSE). Per últim, es mostra una gràfica amb tots els errors relatius acumulats per sectors.

Resultats d'alguns dels indicadors dels models per sectors policials

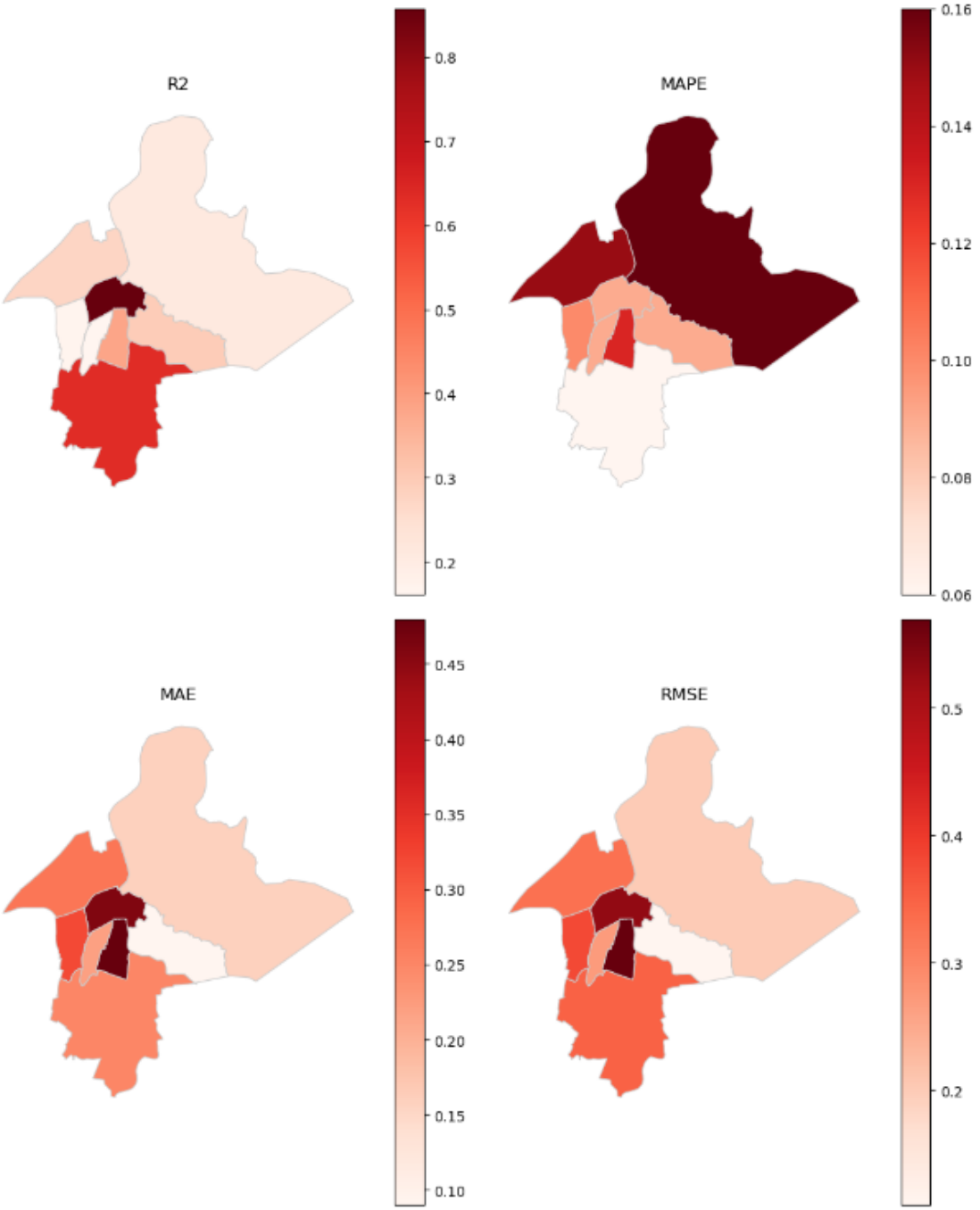

Figura 5.45: Composició de mapes on es veuen representats alguns dels indicadors extrets dels models pels diferents sectors policials. Autor: Sergi Payarol.

Si ens fixem en el coeficient de determinació, gairebé tots els sectors mostren un resultat força baix. El sector 1 i el sector 8 mostren un R2 força bo, mentre que el sector 2 arriba gairebé al 40%. D'altra banda el sector 5, el sector 6 i el sector 7 no arriben ni al 30%. Per últim, el sector 3 i el sector 4 mostren una bondat d'ajust per sota del 20%. Hem de tenir en compte que aquesta mesura estadística ens indica la proporció de la variància de la variable dependent que és explicada pel model. És a dir, els models de la majoria de sectors tenen dificultat en explicar la variabilitat de les dades.

En canvi, el sector 1 i el sector 8 mostren un coeficient força alt, amb el que podem interpretar que els models expliquen bé la variabilitat de la variable dependent.

Pel que fa a MAPE, el sector 1, el sector 3, el sector 4, el sector 5 i el sector 8 mostren un percentatge de l'error mitjà absolut per sota o igual del 10%. Això vol dir que, de mitjana, els pronòstics difereixen dels valors reals en un 10% o menys, mentre que la resta de sectors superen el 15%, però se situen per sota del 20%.

El MAE en tots els sectors és força acceptable. La mitjana de l'error absolut més alta es troba en el sector 2 amb 0,48, valor que indica que, de mitjana, hi ha un error del 0,48 entre els valors predits i els valors originals. Aquest resultat no és excessivament elevat, donat el rang de valors que tenim i, per tant, podem afirmar que és bo.

Per últim, l'RMSE sol ser força estable en la majoria de sectors. Podem destacar el sector 1 i el sector 2 amb un RMSE de més del 0,50. Però, donat el rang de valors amb els quals estem treballant, no sembla un resultat molt elevat. Hem d'entendre que l'RMSE ens diu l'error quadràtic mitjà entre els valors predits i els originals, en la mateixa escala que la variable resposta.

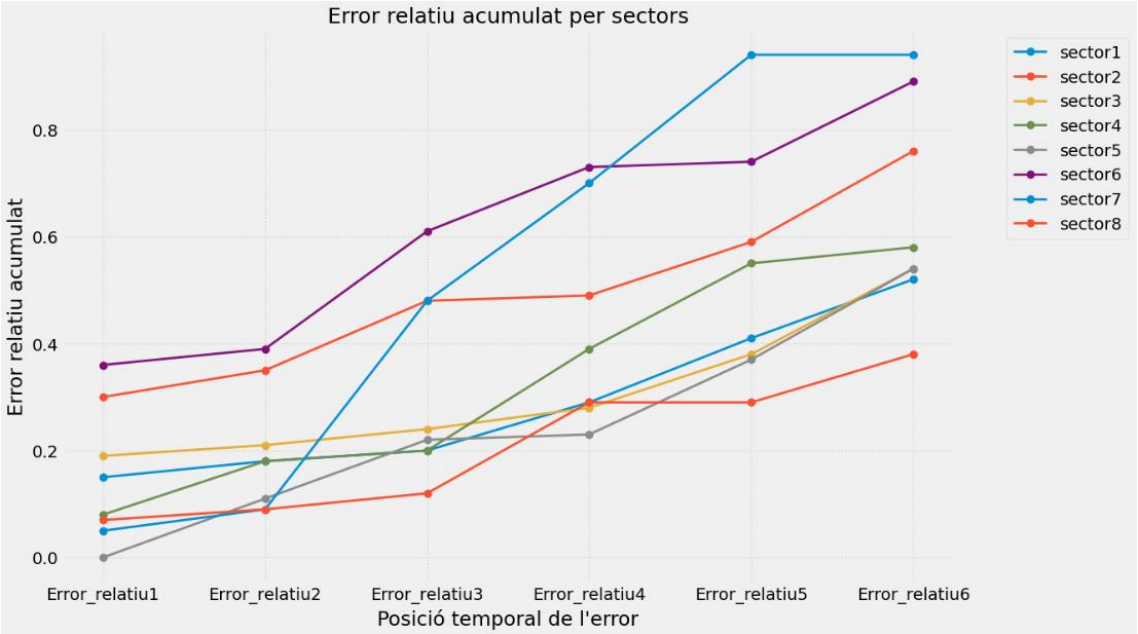

Figura 5.46: Resultats dels errors relatius acumulats pels diferents sectors policials. Autor: Sergi Payarol.

A la gràfica 5.46 se'ns mostra l'error relatiu acumulat per mesos. A simple vista podem veure que, a mesura que la predicció és més a futur, l'error s'eleva. Hem de pensar que si "l'interval de l'error es troba entre el 5% i el 10%, el resultat és acceptable, mentre que si l'error relatiu és superior al 10%, el resultat és poc confiable" (Graciela, M., 2005). Si seguim aquest criteri, el sector 4, el sector 5, el sector 7 i el sector 8 tenen un error relatiu per sota del 10% en el primer mes predit. Per tant, aquí tenim un altre indici que els models es poden millorar, buscant altres algorismes o altres variables predictores que ajudin a enriquir el model i obtenir més bons resultats. Ara bé, si optem per

considerar un error acceptable per sota del 20%, la majoria de sectors entren en aquest interval, exceptuant el sector 2 i el sector 6. El segon mes encara és acceptable en aquests últims sectors que es posicionen per sota del 20%. El tercer mes, es comença a apreciar que tenim menys sectors sota l'interval del 20%. Així doncs, podem concloure que la predicció pot ser acceptable fins els dos mesos per a tots els sectors menys el sector 2 i el sector 6; i fins a tres mesos pel sector 1, el sector 4 i el sector 8.

5.3.2 Anàlisi per sèries temporals multivariants: Girona

En aquest apartat es realitzarà l'anàlisi per sèries temporals multivariants de l'àmbit municipal de Girona. Al tractar-se d'una aproximació multivariant s'utilitzaran altres variables predictores corresponents a factors meteorològics.

Com que el model Prophet ens ha donat bons resultats a l'apartat anterior, ara s'utilitzarà només aquest model per l'àmbit de Girona. Així doncs, s'ha de tenir en compte que en utilitzar aquest model no és necessari analitzar separadament els components de tendència, estacionalitat i altres components d'una sèrie temporal, ja que aquest model els detecta automàticament i es poden ajustar a partir dels hiperparàmetres. Tot i això, hem realitzat la mateixa anàlisi per cada component per així tenir una comprensió més robusta, tal com es pot consultar a l'informe tècnic d'Anàlisi predictiva.

Així doncs, primer se seleccionen les millors variables pel nostre model. Una vegada s'han seleccionat les característiques, es procedirà a veure els resultats del model.

5.3.2.1 Selecció de variables

En aquest apartat seleccionarem les millors variables pel nostre model.

Utilitzarem *RandomForestRegressor* per a seleccionar les millors variables pel nostre model i RFE.

RandomForestRegressor

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(X)$$

On:

$\hat{y}$: És la predicció final de la variable objectiu.

N: És el nombre d'arbres en el conjunt.

$f_i(X)$: És la predicció realitzada per l'arbre i en el conjunt per a les característiques X.

El càlcul de la importància de les variables es basa en dos factors principals:

- **Mida de la reducció de la impuresa**: Es mesura quanta reducció de la impuresa (com ara l'error quadràtic mitjà o la desviació estàndard) es produeix en cada arbre de decisió de cada partició basada en una variable determinada. Aquesta mesura indica la importància de la variable en la separació i la predicció precisa de les dades.
- **Freqüència d'ús de la variable**: Es mesura quants cops una variable és seleccionada per dividir els nodes en tots els arbres de decisió. Això reflecteix la freqüència amb la qual la variable s'utilitza per prendre decisions importants en la construcció dels arbres.

A continuació, mostrarem la gràfica amb l'ordre descendent de les variables més importants.

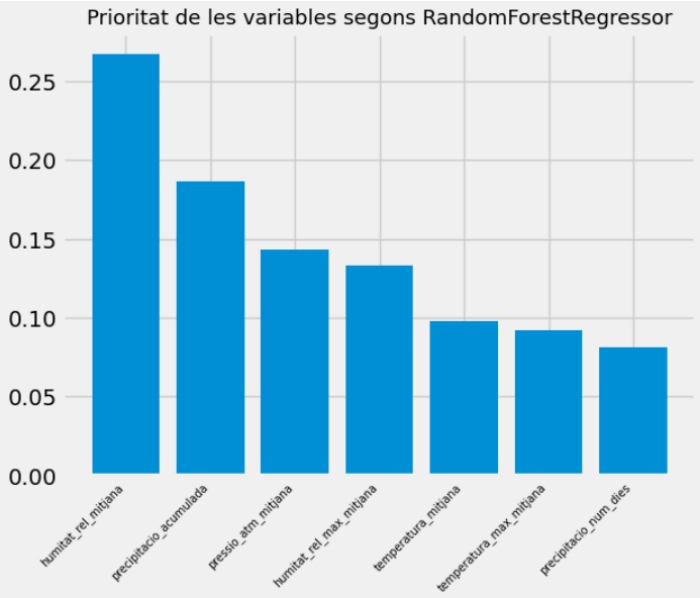

Figura 5.47: Resultat de RandomForestRegressor. Autor: Sergi Payarol.

RFE

L'algorisme RFE (*Recursive Feature Elimination*) és una tècnica de selecció de variables que s'utilitza per reduir la dimensionalitat d'un conjunt de dades i seleccionar les característiques més importants per a un model de regressió. Així doncs, aquest algorisme aplica una recursivitat on s'eliminen les característiques menys importants del conjunt de dades original. Això ho fa a partir de l'ajust successiu del model, en aquest cas *RandomForestRegressor*, i avalua la seva precisió.

Tot seguit, mostrarem el resultat de l'algorisme RFE.

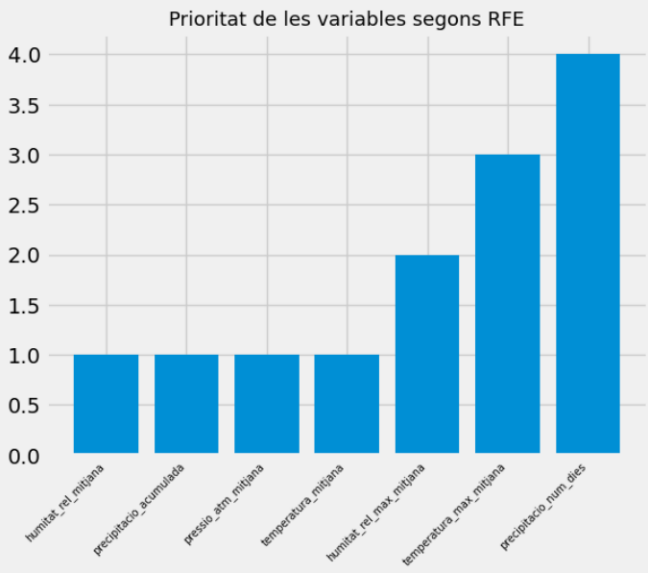

Figura 5.48: Resultat d'RFE. Autor: Sergi Payarol.

Així doncs, ens quedarem amb humitat_rel_mitjana, precipitació_acumulada, pressió_atm_mitjana, temperatura_mitjana i humitat_rel_max_mitjana.

5.3.2.2 Resultats del model

En aquest apartat detallarem tots els indicadors que hem extret del model. Abans, però, mostrarem la predicció que ha realitzat Prophet.

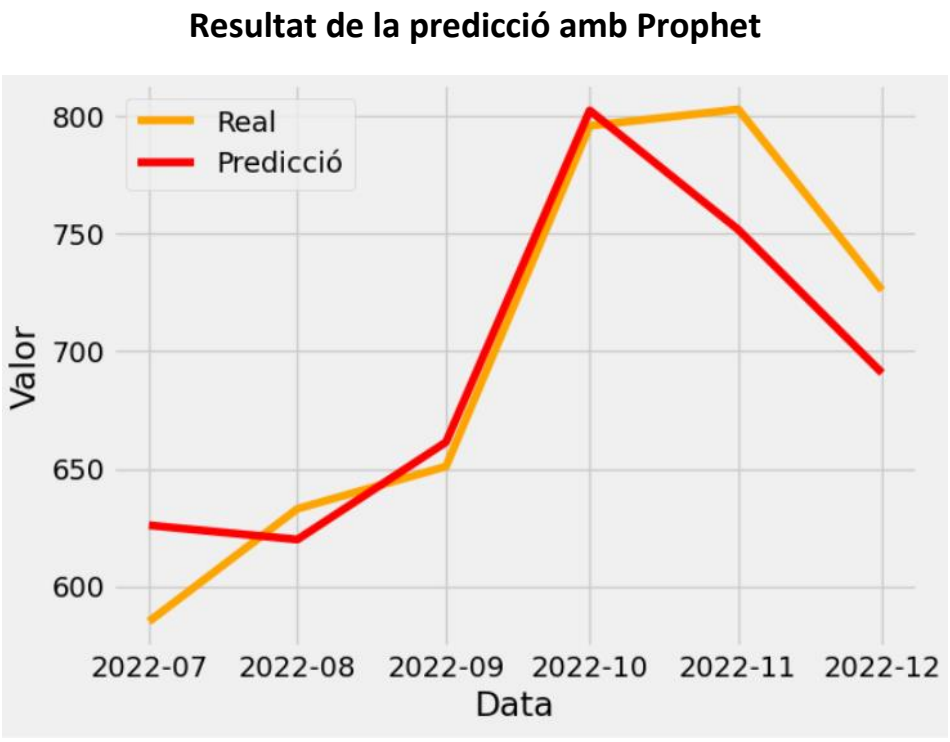

Figura 5.49: Resultat de la predicció amb Prophet. Autor: Sergi Payarol.

A la gràfica de sobre mostrem la predicció feta pel model. Podem veure a simple vista que la predicció és força acceptable. Tot seguit mostrem la taula amb els resultats dels indicadors.

Indicadors	Resultats
MSE	977.03
MAE	26.22
RMSE	31.26
MAPE	0.04
CRMSE	30.477
R2	0.86
Valor real 1	585
Valor predit 1	625.98
Error relatiu 1	0.07
Valor real 2	633
Valor predit 2	619.92
Error relatiu 2	0.09
Valor real 3	651
Valor predit 3	661.35
Error relatiu 3	0.11
Valor real 4	796
Valor predit 4	802.52
Error relatiu 4	0.12
Valor real 5	803
Valor predit 5	751.79
Error relatiu 5	0.18
Valor real 6	726
Valor predit 6	690.79
Error relatiu 6	0.23

Figura 5.50: Resultats dels indicadors del model Prophet per l'àmbit municipal de Girona. Autor: Sergi Payarol.

En general, els resultats són força bons. Tenim un MSE força elevat, el qual indica una certa dispersió entre les prediccions i els valors reals amb un error quadràtic mitjà relativament alt. Hem de tenir en compte que l'MSE és sensible als errors quadràtics i penalitza molt més els errors grans. Això significa que els errors més grans tindran un impacte molt més elevat en el càlcul de l'MSE. Així doncs, si tenim valors atípics, aquests poden fer que l'error en la predicció sigui més elevat. És per aquest fet, que l'MSE és interessant, però s'han de considerar altres indicadors.

Pel que fa al MAE, aquest ens indica l'error mitjà absolut entre els valors reals i els valors predits. A diferència de l'MSE, el MAE no és tan sensible als valors atípics. Tenim un MAE de 26,22. Si valorem els valors originals, que tenen tres dígits, el MAE és força baix.

L'RMSE ens indica l'arrel quadrada de l'error mitjà quadràtic. És a dir, la dispersió entre les prediccions i els valors reals. Així doncs, podem considerar el resultat de l'RMSE baix.

També podem veure que tenim un MAPE força reduït. Aquest indicador mesura l'error mitjà percentual absolut entre els valors reals i els predits. Per tant, el que veiem és que els pronòstics difereixen un 4% o menys dels valors reals.

Pel que fa al CRMSE, el resultat és de 30,477. Aquest calcula l'error quadràtic mitjà tenint en compte la desviació de les prediccions respecte a la mitjana dels valors reals. Si tenim un CRMSE que s'aproxima a 0, podem dir que el model s'ajusta molt bé a les dades en relació amb la mitjana.

D'altra banda, l'R2 és força bo. El coeficient de determinació indica la bondat de l'ajust del model i ens diu quin percentatge de variabilitat és explicada. Així doncs, Prophet explica aproximadament el 86% de la variabilitat observada en els valors reals.

Per últim, els errors relatius acumulats mostren un error en el temps força acceptable, sent el sisè mes el que té l'error més alt (26%). A continuació, mostrem una gràfica on es mostra l'error relatiu acumulat.

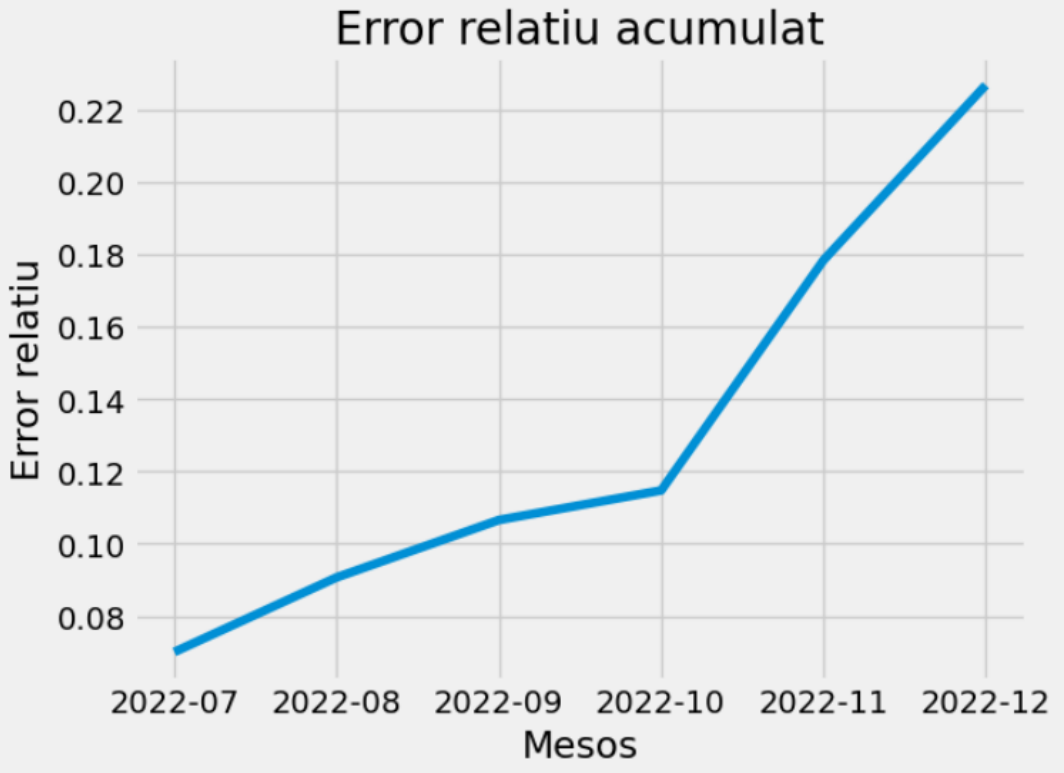

Figura 5.51: Resultats dels errors relatius acumulats per l'àmbit municipal de Girona. Autor: Sergi Payarol.

Podem observar com passat el quart mes, l'error s'eleva bastant, la qual cosa indica que l'error fins al tercer o quart mes pot ser força acceptable. A partir d'aquesta informació es pot saber fins a quin punt serà segur acceptar un valor predit. Com ja s'ha comentat, és normal veure que, com més lluny es vulgui predir, més incerta serà la predicció i més error tindrà.

6. CONCLUSIONS

Els algorismes per fer prediccions són cada vegada més sofisticats amb resultats molt propers a la realitat. No obstant, si en aquesta predicció hi englobem variables com els delictes ocorreguts en una ciutat, la capacitat de predicció és molt més incerta i difícil d'aconseguir. Aquest fet es deu a que el delicte en si pot englobar multitud de factors que, de cap manera, s'ha volgut analitzar en aquest projecte. De fet, el delicte en si pot mostrar una certa estacionalitat, si els agrupem per diferents franges temporals. És a partir d'aquesta estacionalitat i d'altres components que hem analitzat en aquest treball que es pot obtenir una capacitat predictiva amb cert grau d'error. Com a eina complementària que ajudi a la policia a gestionar el territori en temes de seguretat pot estar bé i ser útil. Ara bé, s'ha d'entendre que en aquest projecte no s'ha tingut com a objectiu treure els millors models que tinguin una capacitat predictiva molt bona. Sí que és veritat, que l'objectiu inicial radica en elaborar diferents models per sectors policials, per així poder-se implementar. No obstant, tot i que els resultats dels models són correctes, no s'ha de descartar l'opció de seguir endavant amb altres projectes que tinguin com a objectiu millorar els models que hem presentat en aquest treball.

Malgrat que aquest projecte ha tingut com a principal objectiu la predicció, l'extracció de coneixement de les dades ha estat molt enriquidora. A partir de l'anàlisi exploratòria de les dades s'ha pogut entendre la distribució dels delictes segons la seva tipologia. El furt, els danys, el robatori a interiors de vehicles i l'estafa són els delictes que més s'han registrat en aquests últims anys. Seria interessant en futurs projectes, analitzar amb més profunditat aquests delictes en concret. Per exemple, a partir del delicte per furt, que és el que es dóna més, amb diferència, sorgeixen diferents preguntes: On es cometen aquests furts? Qui els comet? Quins elements configuren l'entorn que rodeja el delicte? Són preguntes molt interessants de respondre i que podria englobar-se perfectament en un segon TFG.

També s'han analitzat temporalment els delictes. Aquest estudi s'ha realitzat a diferents escales temporals, on s'han entès les dinàmiques a gran escala (anual); s'ha vist que els mesos més problemàtics són l'octubre, el novembre, el maig i l'abril; que els divendres i dissabtes són els dies amb més registres de delictes; i les franges del vespre i del matí són les que més despunten de la resta, però si observem les hores del dia, les dotze de la nit i del migdia i les vuit del vespre són les hores més crítiques.

Al següent apartat hem realitzat una anàlisi espacial dels delictes. Ha estat molt interessant veure el desplaçament dels delictes, per densitat, en el mapa de calor. Amb aquesta informació s'ha pogut teoritzar les diferents zones delictives al municipi de Girona. El *hot spot* principal es localitza al Mercadal i l'Eixample Nord. També s'ha detectat la perifèria més propera al *hot spot*, que és on es projecten més els delictes cap al sud-oest, sent Santa Eugènia i Can Gibert del Pla, els subsectors que més despunten. Tanmateix, a la perifèria més allunyada, la part nord i sud del municipi, es produeixen

forces casos, però amb menor proporció que els del centre. Finalment, s'han pogut localitzar altres *hot spots* o "satèl·lits", que de manera intermitent es mostren feblement en diferents moments de la sèrie temporal.

Una vegada s'ha obtingut una visió general dels delictes a nivell temporal i espacial, s'ha passat a analitzar les persones detingudes. És important remarcar que per aquest projecte s'ha descartat aprofundir en el perfil dels individus arrestats. És un estudi que s'escapa de l'abast del projecte, però que seria molt interessant d'emprendre en futurs estudis. Per aquest projecte, doncs, ens hem focalitzat en entendre primer el delicte i, en segon lloc, el detingut. Hem obtingut dades força interessants i que expliquen a una escala temporal i espacial la distribució dels delictes i els autors dels fets. Els homes han representat la mostra més gran i que per tant, les conclusions que hem pogut extreure són més fermes que no pas amb la mostra de les dones, que ha estat força reduïda.

Als últims apartats de la fase exploratòria s'han realitzat diferents anàlisis correlatives sobre variables externes al nostre joc de dades. Hem pogut comprovar que la densitat i el total de població per barris té una forta correlació amb el nombre de delictes. Si recordem bé, aquestes variables s'han hagut d'analitzar a partir dels barris oficials, i no dels sectors policials. El principal motiu és que les delimitacions dels sectors policials no casen bé amb la dels barris oficials. Per aquest fet, s'ha agafat la informació a partir dels subsectors policials on hem pogut fusionar-ho amb la capa de barris. Per últim, les variables meteorològiques no han mostrat una certa correlació, tot i que no hem descartat utilitzar-les a l'etapa d'anàlisi predictiva.

Al llarg del treball, hem aprofundit en certs aspectes, però una de les aspiracions d'aquest treball és la de proporcionar una visió general dels delictes a Girona en aquests últims anys. És a partir d'aquesta primera exploració que es pot obtenir una base de partida. I és que aquesta és una de les finalitats del propi projecte, aportar un coneixement que convidi i motivi a investigadors a anar més enllà i aporti nou coneixement que ens ajudi a entendre com es pot gestionar el territori en termes de seguretat.

En general, tot el procés de treball ha estat molt enriquidor a nivell personal i espero que aquesta informació generada pugui servir per la gestió de les patrulles, però també per l'ordenació del territori. I és que, de fet, tot i ser un treball tècnic i instrumental, els fonaments han estat pròpiament geogràfics. S'ha analitzat el territori, l'espai distribuït per sectors policials aplicant la variable temporal; s'ha vist com les persones detingudes actuen en aquests sectors; s'han utilitzat variables com la densitat, el nombre de població per barris i factors meteorològics. La geografia ha estat present al llarg de tot el treball, aportant aquesta visió interdisciplinària tan pròpia d'aquesta bella disciplina.

7. BIBLIOGRAFIA I WEBGRAFIA

Acharya, S., (2021). *What are RMSE and MAE?* Recuperat de: https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383 [Consulta: 02/04/2023]

Aragon, F., (2017). *Series temporales*. Recuperat de: https://github.com/FrancisArgnR/SeriesTemporalesEnCastellano [Consulta: 01/04/2023]

Benites, L., (2021). *Prueba blanca: definición, ejemplos*. Recuperat de: https://statologos.com/prueba-blanca/ [Consulta: 28/03/2023]

Coleman, A., (2023). *Prophet*. Recuperat de: https://github.com/facebook/prophet [Consulta: 12/05/2023]

Dhiraj, K., (2018). *Anomaly detection using Isolation Forest in Python.* Recuperat de: https://blog.paperspace.com/anomaly-detection-isolation-forest/#:~:t. [Consulta: 01/03/2023]

Duk2, (2016). *Series estacionarias: Por qué son importantes para trabajar con modelos*. Recuperat de: https://estrategiastrading.com/series-estacionarias/ [Consulta: 16/05/2023]

Ermeydan, İ., & Akgönen, A. İ. (2022). Investigation on behavior and seismic performance of reduced beam sections. *Revista de la construcción*, *21*(2), 427-446. Recuperat de: https://www.researchgate.net/publication/364566571_Investigation_on_behavior_and_seismic_performance_of_reduced_beam_sections [Consulta: 02/05/2023]

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34. Recuperat de: http://140.118.5.28/DataMining_Notes/fayyad.pdf [Consulta: 08/05/2023]

Filho, M., (2022). *How to do time series cross-validation in Python.* Recuperat de: https://forecastegy.com/posts/time-series-cross-validation-python/ [Consulta: 07/05/2023]

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016. Recuperat de: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.295.3501&rep=rep1&type=pdf [Consulta: 03/05/2023]

González-Álvarez, J. L., Hermoso, J. S., & Camacho-Collados, M. (2020). Policía predictiva en España. Aplicación y retos futuros. *Behavior & Law Journal*, *6*(1), 26-41. Recuperat

de: https://www.behaviorandlawjournal.com/BLJ/article/view/75/90 [Consulta: 15/04/2023]

Graciela, M., (2005). *El proceso de medición: Análisis y comunicación de datos experimentales.* Recuperat de: https://www.unrc.edu.ar/unrc/digtal/El_proceso_de_med.pdf [Consulta: 04/05/2023]

Hassouna, F., (2020). *What is the problema with using R-squared in time series models?* Recuperat de: https://www.researchgate.net/post/What-is-the-problem-with-using-R-squared-in-time-series-models [Consulta: 02/05/2023]

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in psychology*, *6*, 727. Recuperat de: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00727/full [Consulta: 05/05/2023]

Kane, R. (2018). *A Statistical Study on the Impact of Weather on Crime: Technical Report* (Doctoral dissertation, Dublin, National College of Ireland). Recuperat de: https://norma.ncirl.ie/3449/1/robertkane.pdf [Consulta: 01/04/2023]

Kaufmann, M., Egbert, S., & Leese, M. (2018). Predictive policing and the politics of patterns. *The British Journal of Criminology*, *59*(3), 674-692. Recuperat de: https://academic.oup.com/bjc/article/59/3/674/5233371 [Consulta: 12/05/2023]

KNIMETV, (2022). *Data Science Pronto! – Why is R^2 not used to measure time series analysis performance*. Recuperat de: https://www.youtube.com/watch?v=MN5AYymKDhc [Consulta: 07/05/2023]

Lapowsky, i., (2018). *How the LAPD uses data to predict crime.* Recuperat de: https://www.wired.com/story/los-angeles-police-department-predictive-policing/ [Consulta: 03/04/2023]

Lau, T., (2020). *Predictive policing explained*. Recuperat de: https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained [Consulta: 04/04/2023]

Martínez, R., (s.d.). *Tema 6: Series temporales.* Recuperat de: https://www.uv.es/ceaces/pdf/tur/tema6.pdf [Consulta: 07/05/2023]

Meng, S., (2022). *Isolation forest anomaly detection for telemetry time series data.* Recuperat de: https://medium.com/@juniper.cto.aiml.2021/isolation-forest-anomaly-detection-for-telemetry-time-series-data-bbd71adacfaf [Consulta: 05/05/2023]

Ministeri de l'Interior, (2022). *Estadística de seguridad: actuaciones policiales.* Recuperat de: https://estadisticasdecriminalidad.ses.mir.es/publico/portalestadistico/dam/jcr:bae4fa

e5-2acf-4153-a8a7-f06edd0b7bb0/Informe-metodologico-Actuaciones-Policiales.pdf [Consulta: 01/05/2023]

Noor, T. H., Almars, A. M., Alwateer, M., Almaliki, M., Gad, I., & Atlam, E. S. (2022). SARIMA: A Seasonal Autoregressive Integrated Moving Average Model for Crime Analysis in Saudi Arabia. *Electronics*, *11*(23), 3986. Recuperat de: https://www.mdpi.com/2079-9292/11/23/3986 [Consulta: 02/05/2023]

OSCE, (2017). Guía de la OSCE sobre actividad policial basada en la intel·ligència. Recuperat de: https://www.osce.org/files/f/documents/6/4/455536.pdf [Consulta: 16/04/2023]

Parlament de Catalunya, (2018). *Parts del dia i expressió de les hores.* Recuperat de: https://www.parlament.cat/document/intrade/6698 [Consulta: 11/05/2023]

Pedrosa, J., (2020). *Heterocedasticidad*. Recuperat de: https://economipedia.com/definiciones/heterocedasticidad.html [Consulta: 19/03/2023]

Perktold, J., *et al.,* 2023). *statsmodels.tsa.arima.model.ARIMA*. Recuperat de: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima.model.ARIMA.html [Consulta: 03/04/2023]

Perktold, J., *et al.,* 2023). *statsmodels.tsa.holtwinters.ExponentialSmoothing*. Recuperat de: https://www.statsmodels.org/devel/generated/statsmodels.tsa.holtwinters.ExponentialSmoothing.html [Consulta: 01/04/2023]

Perktold, J., et al., 2023). *Statsmodels.tsa.seasonal.seasonal_decompose*. Recuperat de: https://www.statsmodels.org/devel/generated/statsmodels.tsa.seasonal.seasonal_decompose.html [Consulta: 03/04/2023]

Perktold, J., *et al.,* 2023). *statsmodels.tsa.statespace.sarimax.SARIMAX*. Recuperat de: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html [Consulta: 05/05/2023]

Perktold, J., *et al.,* 2023). *Statsmodels.tsa.stattools.adfuller*. Recuperat de: https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html [Consulta: 01/03/2023]

Sindhuri , D., (2019). *Time series analysis and forecasting of crime data.* Recuperat de: https://scholarworks.calstate.edu/downloads/wp988k17c [Consulta: 10/05/2023]

Smith, T., (2023). *Pmdarima.arima.ARIMA()*. Recuperat de: https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html [Consulta: 09/05/2023]

Statalogos, (2021). *La prueba de breush-Pagan: definición y ejemplo*. Recuperat de: https://statologos.com/prueba-pagana-breusch/ [Consulta: 01/05/2023]

Sujit, S., (2015). *Do you use coeficient of variation to determine forecastability?* Recuperat de: https://blog.arkieva.com/do-you-use-coefficient-of-. [Consulta: 03/04/2023]

UCM, (2013). *Tema 6: Modelización con datos de series temporales*. Recuperat de: https://www.ucm.es/data/cont/docs/518-2013-10-25-Tema_6_EctrGrado.pdf [Consulta: 08/05/2023]

Villalba, R., (2020). *Series temporales con ARIMA I.* Recuperat de: http://enrdados.net/post/series-temporales-con-arima-i/ [Consulta: 08/05/2023]

Villavicencio, J. (2010). Introducción a series de tiempo. *Puerto Rico*. Recuperat de: https://d1wqtxts1xzle7.cloudfront.net/38458362/manual_intro_series_tiempo-libre. [Consulta: 10/05/2023]

Viviana, P., (2014). *Modelación y predicción de focos de criminalidad basado en modelos probabilísticos.* Recuperat de: https://repositorio.uchile.cl/handle/2250/129832#:~:text. [Consulta: 01/05/2023]

Zambrano, R., (2020). *Predicción de ocurrència de delitós: implementación de modelos predictivos en base a los casos registrados en CABA durante el periodo 2017-2019.* Recuperat de: http://bibliotecadigital.econ.uba.ar/econ/collection/tpos/document/1502-1919_ZambranoMedinaRA [Consulta: 01/05/2023]