

Grau en Estadística

Títol: Anàlisi estadística del número de trucades a la Línia d'Atenció contra la Violència Masclista de la Generalitat de Catalunya

Autor: Mireia Bolívar Rubia

Director: David Moríña Soler

Departament: Departament d'Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2024



UNIVERSITAT DE
BARCELONA



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística

Dedicatòria

A totes les víctimes de la violència de gènere, amb l'esperança que un dia puguem viure en un món lliure d'aquesta xacra.

Agraïments

Al meu tutor, per la seva guia i dedicació, sense els quals aquest projecte no hauria estat possible.

A la meva família i amics, pel seu suport incondicional i per estar sempre al meu costat a cada pas d'aquest camí.

Aquest treball forma part del projecte “Data science against gender-based violence” finançat per la Fundació Obra Social La Caixa (codi LCF/PR/SR22/52570005).

Resum

La violència de gènere s'ha mantingut gairebé invisible durant dècades, afectant greument dones i nenes a escala mundial. Aquest estudi se centra en la violència física o sexual per raó de gènere contra les dones adultes a Catalunya, abordant l'important subnotificació dels casos, sovint conegut com a l' 'Iceberg de la violència de gènere'.

Utilitzant dades de la *Línia d'Atenció Contra la Violència Masclista*, comparem les estimacions d'un model de subnotificació extern amb els registres reals de les trucades per validar-ne la precisió. A més, examinem l'impacte de la pandèmia de la COVID-19 en la freqüència de trucades de la línia d'atenció, reconeixent que les situacions de confinament agreugen la violència masclista. Els models ARIMA i la Xarxa Neural Artificial (ANN) s'utilitzen per generar prediccions de trucades mensuals.

Les troballes confirmen la precisió del model de subnotificació i quantifiquen l'augment significatiu de la violència durant el confinament de la COVID-19. A més, s'identifica el model ANN com el millor model predictiu.

Paraules clau: Violència de gènere, subnotificació, Línia d'Atenció Contra la Violència Masclista, impacte COVID-19, Sèries Temporals, ARIMA, Xarxes Neuronals Artificials.

Gender-based violence has remained nearly invisible for decades, severely affecting women and girls worldwide. This study focuses on gender-based physical or sexual violence against adult women in Catalonia, addressing the significant underreporting of cases, often known as the 'Gender Violence Iceberg'.

Using data from the '*Línia d'Atenció Contra la Violència Masclista*', we compare the estimates of an external underreporting model with actual call records to validate its accuracy. Additionally, we examine the impact of the COVID-19 pandemic on the frequency of hotline calls, recognizing that confinement situations exacerbate gender-based violence. ARIMA and Artificial Neural Network (ANN) models are used to generate monthly call predictions.

The findings confirm the underreporting model's accuracy and quantify the significant increase in violence during the COVID-19 lockdown. Moreover, the ANN model is identified as the best predictive model.

Key words: Gender-based violence, underreporting, Línia d'Atenció Contra la Violència Masclista, COVID-19 impact, Time Series, ARIMA, Artificial Neural Networks.

Clasificació AMS

- 62P25 Applications to social sciences
- 37M10 Time series analysis
- 62M10 Time series, auto-correlation, regression, etc.
- 91D10 Models of societies, social and urban evolution
- 60G25 Prediction theory (aspects of stochastic processes)
- 68T01 General topics in artificial intelligence
- 68T05 Learning and adaptive systems

Índex

Agraïments	ii
Resum	iii
Clasificació AMS	iv
Índex de Figures	vii
Índex de Taules	xi
1 Introducció	1
1.1 Context i motivació	1
1.2 Objectius principals	3
2 Metodologia	5
2.1 Dades	5
2.1.1 Estructura del Sistema Sanitari Català	5
2.1.2 Fonts utilitzades	6
2.2 Models	7
2.2.1 Validació del model per estimar l'evolució real de casos de Violència de Gènere	7
2.3 Mètriques per a la quantificació de l'impacte de la COVID-19 al número de trucades	9
2.3.1 Test d'arrels unitàries de Dickey-Fuller	9
2.3.2 Causal Impact	11
2.4 Predicció de Sèries Temporals	13
2.4.1 Introducció a les Sèries Temporals	13
2.4.2 Processos estocàstics	14
2.4.3 Metodologia Box-Jenkins	16
2.4.4 Xarxes Neuronals Artificials (ANN)	18
2.5 Mètriques d'error	21

3	Resultats	23
3.1	Validació externa del model de subnotificació	23
3.1.1	Evolució mensual del nombre de trucades a la Línia d'Atenció contra la Violència Masclista	23
3.1.2	Anàlisi del Model Extern de Subnotificació	25
3.1.3	Comparació i validació del model	26
3.2	Quantificació de l'impacte de la COVID-19	27
3.2.1	Test de Dickey-Fuller	28
3.2.2	Model Causal Impact	29
3.3	Predicció del número mensual esperat de trucades a la Línia d'Atenció Contra la Violència Masclista	31
3.3.1	Modelització de Sèries Temporals	31
3.3.2	Modelització d'una Xarxa Neuronal Artificial	41
3.3.3	Millor model	43
4	Discussió i Conclusions	45
4.1	Validació externa del Model de subnotificació	45
4.1.1	Evolució del número de trucades	45
4.1.2	Evolució dels registres del model de subnotificació	46
4.1.3	Validació	46
4.2	Quantificació de l'impacte de la COVID-19	47
4.3	Predicció i comparació de models predictius	48
4.4	Conclusions	50
	Bibliografia	51
A	Codi R	55

Índex de figures

2.1	Sistema Sanitari Català	5
2.2	Valors estimats dels paràmetres (mediana posterior i interval de credibilitat del 95%) per a cada subàrea	9
2.3	Exemple de l'output del plot de la funció CausalImpact	12
2.4	Exemple de l'output del summary de la funció CausalImpact	12
2.5	Esquema de la metodologia Box-Jenkins	16
2.6	Estructura d'una xarxa neuronal	19
2.7	Exemple d'una xarxa neuronal amb quatre entrades i una capa oculta amb tres neurones ocultes.	19
3.1	Número de Trucades per mes a l'Àrea Metropolitana Nord de Barcelona	24
3.2	Casos Estimats i Registrats per mes a l'Àrea Metropolitana Nord de Barcelona	25
3.3	Validació del model amb el número de trucades	26
3.4	Número de trucades mensuals a la Línia d'Atenció contra la Violència Masclista	27
3.5	Test de Dickey-Fuller per al període pre estat d'alarma	28
3.6	Test de Dickey-Fuller per al període post estat d'alarma	29
3.7	Resultats Summary Anàlisi Casual Impact	30
3.8	Resultats Plot Anàlisi Casual Impact	30
3.9	Gràfics per analitzar la variància de la sèrie	32
3.10	<i>Monthplot</i> de la sèrie de trucades	32
3.11	Gràfic temporal on s'observa la mitjana abans d'aplicar la diferenciació regular	33
3.12	ACF de la sèrie	33
3.13	ACF i PACF de la sèrie	34
3.14	Validació de la normalitat	36
3.15	Validació de la variància	37
3.16	Contrast de Breusch-Pagan per validar la variància	37
3.17	Gràfics per validar la independència dels residus	38
3.18	Comprovació de la volatilitat del model	38
3.19	Predicció Out-of-Sample model SARIMA	40
3.20	Predicció dels 12 mesos futurs per al model SARIMA	40

3.21	Arquitectura de la xarxa neuronal utilitzada per la predicció del número mensual del número de trucades	41
3.22	Predicció Out-of-Sample model ANN	42
3.23	Predicció dels 12 mesos futurs per al model ANN	43

Índex de taules

3.1	Descriptives del Número de Trucades	24
3.2	Descriptives dels Casos Registrats i Estimats	26
3.3	Resum Descriptives	27
3.4	Resum Descriptives de diferents franges temporals	28
3.5	Comparació de la variància de la sèrie al aplicar diferents diferenciacions	33
3.6	Estimació del model 1	35
3.7	Estimació del model 1.1	35
3.8	Estimació del model 1.2	35
3.9	Estimació del model 1 amb paràmetres significatius	35
3.10	Estimació del model 2 amb paràmetres significatius	36
3.11	Comparació dels models estimats a partir del criteri AIC	36
3.12	Anàlisi de l'estabilitat del model per a les dades d'entrenament	39
3.13	Anàlisi de l'estabilitat del model per a la sèrie sencera	39
3.14	Mètriques d'Error per al Conjunt d'Entrenament (SARIMA).	39
3.15	Mètriques d'Error per al Conjunt de Prova (SARIMA).	40
3.16	Mètriques d'Error per al Conjunt d'Entrenament (ANN).	42
3.17	Mètriques d'Error per al Conjunt de Prova (ANN).	43
3.18	Mètriques d'Error per comparar els diferents models	44

Capítol 1

Introducció

1.1 Context i motivació

Durant dècades la violència masclista ha estat un fenomen invisibilitzat. Aquest fet ha constituït un gran problema social que afecta dones i nenes de tot el món, generant un impacte devastador en la vida de les víctimes i en la societat en el seu conjunt.

Segons l'Organització Mundial de la Salut (OMS), aproximadament el 30% de les dones arreu del món han experimentat violència física i/o sexual en algun moment de les seves vides ([World Health Organization, 2021](#)), sent això un objectiu específic de desenvolupament en l'agenda per a l'any 2030. Tot i que existeix una àmplia varietat de tipus de violència contra la dona, aquest estudi se centra específicament en la violència física o sexual dirigida a dones adultes.

D'acord amb la Macroenquesta de Violència contra la Dona ([Ministerio de Igualdad, 2020](#)), elaborada per la Delegació del Govern espanyol contra la Violència de Gènere, el 57,3% de les dones de 16 o més anys, residents a Espanya, van declarar haver estat exposades a algun tipus de violència masclista (física o sexual) al llarg de la seva vida. A escala autonòmic, segons l'Enquesta sobre Violència Masclista a Catalunya 2021 ([Generalitat de Catalunya, 2021](#)), el 79,3% de les dones també han patit violència masclista almenys una vegada en algun moment.

Aquestes dades posen de manifest que la violència de gènere és una xacra social que no entén de raça, edat, sexe, cultura, país o lloc. Constitueix una vulneració dels drets i llibertats de les dones que s'ha d'erradicar i, per això, és fonamental aturar-la des de tots els àmbits. Les conseqüències de la violència de gènere són devastadores, i els impactes en aquelles que aconseguixen sobreviure poden ser permanents.

Des del començament del 2024 fins al març del mateix any, les víctimes mortals a causa de la violència masclista a Espanya han ascendit a sis, segons l'últim balanç de la Delegació del

Govern contra la Violència de Gènere ([Ministerio de Igualdad, nda](#)). Així i tot, el titular que ens sobta recau en el fet que cap d'aquestes víctimes no havia presentat denúncia. Aquesta xifra posa de manifest una altra dimensió del problema que cal abordar amb urgència: la subnotificació dels episodis de violència masclista, coneguda com l' 'Iceberg de la Violència de Gènere'. Entre el 55 i el 95% de les dones supervivents de violència de gènere no revelen ni utilitzen serveis de salut, legals o policials ([World Health Organization, 2007](#)). Es considera, llavors, que aquesta subnotificació d'episodis de violència de gènere, mina la qualitat de les dades associades, donant lloc a resultats pobres i esbiaixats, aconseguint que la societat mai entengui l'abast del problema real.

És crucial que la societat i les autoritats prenguin mesures per abordar aquesta qüestió amb urgència, proporcionant recursos adequats per a les víctimes, millorant els protocols de resposta i treballant per canviar les actituds i les normes socials que perpetuen la violència masclista.

A escala nacional, el Ministeri d'Igualtat, mitjançant la Delegació del Govern contra la Violència de Gènere, ofereix el Servei telefònic d'informació, assessorament jurídic i atenció psicosocial immediata per personal especialitzat a totes les formes de violència contra les dones a través del número 016 ([Ministerio de Igualdad, ndb](#)).

A Catalunya, l'Institut Català de les Dones, adjunt al Departament d'Igualtat i Feminismes, dissenya, promou, coordina i avalua les polítiques d'igualtat de gènere desenvolupades pel Govern Regional (Generalitat de Catalunya). L'institut ofereix la Línia d'Atenció a les dones en situació de violència (900 900 120), servei que ofereix suport i assessorament a les víctimes de violència de gènere ([Generalitat de Catalunya, nda](#)).

Sota aquest context, per abordar el problema de la violència masclista de manera més efectiva, és primordial quantificar la magnitud dels casos de violència de gènere no denunciats i intentar reconstruir l'iceberg complet de la violència de gènere. Amb aquest objectiu neix un estudi titulat "*Exploring What Lies Beneath the Tip of the Gender Violence Iceberg*" de [Moriña et al. \(2024\)](#). Aquesta investigació analitza la violència de gènere a Catalunya, amb l'objectiu d'estimar el nombre real de casos de violència de gènere considerant que molts casos no són reportats.

Tenint en compte que la Línia d'Atenció de la Generalitat de Catalunya és un servei anònim i accessible per a tothom, plantegem la hipòtesi que els seus registres proporcionen una quantificació més acurada del fenomen que els registres policials o clínics. Per tant, un dels objectius d'aquest treball és comprovar si l'estimació que produeix el model extern de subnotificació és similar a l'evolució del nombre de trucades. Considerarem, llavors, que el model estima adequadament si ambdues evolucions són similars.

La subnotificació no és l'únic front que hem d'afrontar. Segons Acosta ([Lorente-Acosta, 2020](#)), qualsevol circumstància de confinament potencia els factors de risc de violència de gènere, tant a nivell individual com social. L'autor assenyala que en situacions de confinament es creen les condicions idònies perquè els elements de la violència de gènere es magnifiquin, ja que les dones que pateixen aquesta violència es troben en una situació d'aïllament i debilitació de la xarxa de suports, sotmeses a un major control a través de la clausura forçada a casa seva. Sota aquesta premissa, és evident que les dades indiquen que sota les circumstàncies creades pel confinament obligatori instigat per la pandèmia de la COVID-19, hi ha hagut un augment de la violència de gènere, reflectit en un important increment de les trucades al 016 ([Lorente-Acosta et al., 2022](#)).

Tenint en compte aquest altre supòsit, els dos objectius addicionals del treball es centren en tractar la sèrie temporal obtinguda a partir de les dades dels registres de la Línia d'Atenció Contra la Violència Masclista de tota la regió catalana. Primer, visualitzarem i quantificarem l'impacte de la COVID-19 en les trucades, verificant així les paraules d'Acosta. Finalment, generarem prediccions precises del nombre mensual de trucades, modelitzant la nostra sèrie temporal amb un ARIMA i una Xarxa Neuronal Artificial.

1.2 Objectius principals

Al llarg d'aquest estudi, ens endinsarem en les profunditats d'aquest 'iceberg' a partir de diferents fronts. Els objectius principals de la nostra investigació són:

- Comprovar si l'estimació de l'evolució que produeix el model extern de subnotificació és similar a l'evolució del nombre de trucades a la línia d'atenció contra la Violència Masclista, pensant que aquest servei proporciona una quantificació més acurada del fenomen que els registres policials o clínics.
- Quantificar l'impacte de la COVID-19 sobre el nombre de trucades a la Línia d'Atenció a la Violència Masclista, considerant que situacions de confinament augmenten considerablement els casos de maltractament contra la dona.
- Generar prediccions acurades de l'evolució futura del nombre mensual de trucades a la línia d'atenció contra la violència masclista basades en models de sèries temporals. Els principals models predictius utilitzats inclouen models convencionals ARIMA i algorismes d'aprenentatge automàtic com les Xarxes Neuronals Artificials (ANN).

L'estudi comença amb una revisió exhaustiva de la literatura existent sobre la Violència Masclista. A continuació, es descriuran les dades i la metodologia utilitzada per assolir els objectius, incloent-hi l'explicació del model de subnotificació, les mètriques per a quantificar

l'impacte de la COVID-19, l'explicació dels models predictius per a sèries temporals i les mètriques d'error per a la seva comparació. Seguidament, es validarà el model de subnotificació, s'analitzarà l'impacte de la COVID-19 en el nombre de trucades i es generaran les seves prediccions. Finalment, es presentaran els resultats de l'anàlisi i el poder predictiu dels models, que s'avaluaran mitjançant diverses mètriques de rendiment. L'informe conclourà amb discussions dels resultats i conclusions acadèmiques i personals.

Aquest treball aspira a anar més enllà de la mera descripció i comprensió del fenomen, cercant generar prediccions precises sobre la seva evolució futura i intentant trobar una solució a la gran subnotificació. Basant-nos en models estadístics, pretenem oferir una eina valuosa per a la planificació de polítiques i estratègies d'intervenció, amb l'objectiu últim de combatre la violència masclista i construir un futur més segur i equitatiu per a tothom.

Capítol 2

Metodologia

2.1 Dades

2.1.1 Estructura del Sistema Sanitari Català

Per a poder contextualitzar adequadament les dades utilitzades en aquest estudi, és essencial comprendre l'estructura del Sistema Sanitari Català.

El sistema sanitari català està dividit en set regions sanitàries: la Regió Sanitària de l'Alt Pirineu, de Lleida, del Camp de Tarragona, de les Terres de l'Ebre, de la Catalunya Central, la Regió Sanitària de Girona i la Regió Sanitària de Barcelona.



Figura 2.1: Sistema Sanitari Català

La Regió Sanitària de Barcelona, a més, està dividida en tres àrees metropolitanes o urbanes:

- Àmbit Metropolità Sud
- Àmbit Metropolità Nord

- Barcelona Ciutat

A la Figura 2.1 s'observa la il·lustració de la estructura sencera del Sistema Sanitari Català. En color rosa trobem identificada la Barcelona Metropolitana Nord, regió utilitzada per a respondre un dels objectius del present estudi.

2.1.2 Fonts utilitzades

Les dades utilitzades provenen de dues fonts principals del sistema públic de salut a Catalunya: les trucades a la Línia d'Atenció Contra la Violència Masclista i els diagnòstics de violència de gènere produïts a la regió sanitària metropolitana nord de Barcelona.

D'una banda, els diagnòstics mensuals de casos de violència de gènere registrats a l'àrea bàsica de salut Metropolitana Nord van ser facilitats pel Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària (SIDIAP). Aquestes dades, concretament, les farem servir a la validació del model de subnotificació, donat que el model extern utilitza els registres en qüestió per a la seva estimació.

D'altra banda, les trucades a la Línia d'Atenció a les Dones en Situació de Violència (900 900 120), han estat extretes de la següent font de dades: **Línia d'Atenció Contra la Violència Masclista**. Gestionades pel Departament d'Igualtat i Feminismes de la Generalitat de Catalunya, són un servei que ofereix suport i assessorament a les víctimes de violència de gènere.

Concretament, aquesta línia, gratuïta, confidencial i disponible 24 hores del dia i els 365 dies de l'any, tal i com el seu nom indica, no només està adreçada a les dones que pateixen aquesta violència, sinó que també a les persones del seu entorn, els professionals que realitzen atenció a les dones en situació de violència masclista o a qualsevol persona que tingui alguna demanda relacionada amb les violències masclistes ([Generalitat de Catalunya, nda](#)).

El servei en qüestió, a part d'estar destinat a donar suport a les dones que s'enfronten a aquest tipus de violència, també ofereix diferents tipus de serveis, com poden ser atenció directa d'equips psicològics especialitzats, seguiment dels casos o bé activació dels serveis d'emergència, entre d'altres ([Generalitat de Catalunya, nda](#)). A continuació veiem el seu preprocessament.

2.1.2.1 Tractament de les dades

El preprocessament de les dades de la Línia d'Atenció Contra la Violència Masclista ha implicat diverses etapes per assegurar la qualitat i la coherència de les dades. La variable "Situació.Laboral" es va renombrar a "Situació.Personal.Econòmica" per reflectir millor el seu contingut sobre la independència econòmica, mentre que "Detall.situació.laboral" es va mantenir per descriure la situació laboral específica. Es van unificar camps repetits i excloure variables

amb valors uniformes. Els camps buits i "No consta"es van substituir per NA, i en la columna "Nombre.fills.filles", els valors buits o "No consta"es van canviar per Çap fill/a"si no es tenia fills.

Es van corregir errors en la variable "Dataï ajustar la variable "Any"perquè coincidís amb l'any indicat en "Data". Les hores en la variable "Hora"es van unificar per tenir un format consistent, i la variable "Edat" es va ajustar per incloure totes les franges d'edat majors de 60 anys en una sola categoria. Es van crear dues noves variables: "Setmanai "Setmana.Freq", que indiquen respectivament en quina setmana es troben des del començament de les trucades i la quantitat de trucades rebudes en aquella setmana.

Es va filtrar per "Motiu.Trucada"per incloure només casos de violència masclista (violència física o sexual) o camps buits, i es van excloure els casos amb menors d'edat. També es van filtrar les dades per comarques de l'Àrea Metropolitana Nord de Barcelona, excloent Barcelona i L'Hospitalet de Llobregat mitjançant una proporció de casos basada en dades de l'INE. Mentre que per a les prediccions i per la quantificació de la COVID-19 es fan servir les trucades registrades a tota Catalunya.

Finalment, encara que les dades estan registrades diàriament, s'han agregat per a tractar-les mensualment. Mentre que per a la validació externa del model utilitzem la franja temporal des del gener del 2010 fins al desembre del 2021, la quantificació de l'impacte de la COVID-19 i les prediccions, en canvi, tenen en compte els registres des de gener del 2013 fins al desembre del 2023.

Per a més informació, consultar ([Generalitat de Catalunya, ndb](#)), on trobem un PowerBI i la descripció de les variables originals.

2.2 Models

2.2.1 Validació del model per estimar l'evolució real de casos de Violència de Gènere

El model que es busca validar es basa en l'estudi titulat "*Exploring What Lies Beneath the Tip of the Gender Violence Iceberg*" de [Moriña et al. \(2024\)](#). Aquest treball s'enfoca en l'anàlisi de la violència de gènere a Catalunya, tenint com a objectiu principal estimar el nombre real de casos, pensant que molts no són denunciats.

Per a contextualitzar, aquest estudi es basa en dades del sistema públic de salut a Catalunya, específicament de la regió sanitària Barcelona i les seves subàrees. Concretament, les dades en

les quals ens centrarem, com ja s'ha dit, són: el número mensual de casos de violència de gènere registrades en l'Àmbit Metropolità Nord de la Regió Sanitària de Barcelona entre el gener de 2010 i desembre de 2021, on les víctimes eren dones de 16 anys o més (Moriña et al., 2024).

Per estimar el nombre esperat de casos setmanals de violència de gènere reportats al sistema de salut, Moriña et al. van desenvolupar un model matemàtic basat en la suposició que el 32% de les dones han experimentat violència de gènere almenys una vegada a la seva vida, i només el 16% cerca atenció mèdica primària. Utilitzant la mitjana dels casos setmanals segons la macroenquesta de violència de gènere (2019), aquest escenari ofereix una estimació bàsica sense ajustaments per subregistre (Moriña et al., 2024).

2.2.1.1 Detalls del model

Pel que fa a la definició del model, s'assumeix que el nombre real setmanal de casos de violència de gènere X_t segueix una distribució de Poisson amb una mitjana λ , que augmenta per un factor de β durant el període de confinament obligatori (del 14 de març al 24 de juny de 2020), és a dir, $E(X_t) = \lambda + I(t) \cdot \beta$, on $I(t)$ pren el valor 1 durant el confinament i 0 en altres períodes (Moriña et al., 2024).

El nombre de casos diagnosticats al sistema d'atenció primària de salut pública, Y_t , és només una fracció del procés real X_t :

$$Y_t = \begin{cases} q_0 \circ X_t, & t \leq t' \\ q_t \circ X_t, & t > t' \end{cases}$$

on $\circ$ és l'operador d'aprimament binomial¹, definit com $q_t \circ X_t = \sum_{i=1}^{X_t} Z_i$, amb Z_i variables aleatòries de Bernoulli independents i idènticament distribuïdes amb probabilitat d'èxit q_t , com $q_t = 1 - 1(1 - q_0) \cdot e^{-\alpha \cdot (t - t')}$. t' es el punt de canvi desconegut en què la formació en consciència per als professionals de l'atenció primària comença a impactar al nombre setmanal de diagnòstics reportats (Moriña et al., 2024).

També cal destacar que per a la modelització lineal de q_t , α és el moment en què $q_\alpha = 1$, és a dir, els processos registrats i observats coincideixen. En el cas de la modelització exponencial de q_t , α és proporcional a la velocitat a la qual desapareix el problema de la subnotificació com a conseqüència de l'activitat de formació, així que també està relacionat amb el moment en el temps quan tots els casos estan registrats. Tots els paràmetres (q_t, α, β, t') s'estimen mitjançant mostreig de Gibbs, destinada a construir una cadena de Markov amb valors que convergeixen

¹L'operador d'aprimament binomial és una tècnica utilitzada en processos estocàstics per modelar la reducció d'esdeveniments en una seqüència temporal, en funció d'una probabilitat d'èxit determinada, es pot utilitzar un enfocament com ara el model de Bernoulli.

2.3. Mètriques per a la quantificació de l'impacte de la COVID-19 al número de trucades

a una distribució objectiu, utilitzant un enfocament de modelatge bayesià jeràrquic per evitar la no identificabilitat del model (Moriña et al., 2024).

S'assumeixen distribucions prèvies normals per a λ amb una desviació estàndard de 1, la qual cosa condueix al promig setmanal esperat de casos (Moriña et al., 2024).

A la Taula 2.2 presentem un resum dels valors dels paràmetres estimats per a cada subàrea, extreptes de l'estudi (Moriña et al., 2024).

	$\hat{q}_0$	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\mu}$
A	0.024 (0.020, 0.030)	7.0 (6.8, 7.2)	1.3 (0.2, 3.7)	0.01 (0.008, 0.012)	520 (518, 522)
B	0.026 (0.021, 0.031)	7.0 (6.8, 7.2)	0.8 (0.1, 2.4)	0.01 (0.011, 0.017)	505 (500, 506)
C	0.025 (0.019, 0.033)	4.0 (3.8, 4.2)	1.4 (0.2, 3.9)	0.004 (0.003, 0.006)	508 (496, 517)
D	0.041 (0.036, 0.047)	13.0 (12.8, 13.2)	1.8 (0.3, 2.7)	0.006 (0.005, 0.007)	503 (497, 507)
E	0.103 (0.079, 0.134)	1.4 (1.2, 1.5)	0.9 (0.2, 2.0)	0.598 (0.154, 1.706)	508 (506, 509)
F	0.018 (0.015, 0.022)	14.0 (13.8, 14.1)	1.3 (0.2, 3.8)	0.007 (0.006, 0.008)	509 (507, 511)

Figura 2.2: Valors estimats dels paràmetres (mediana posterior i interval de credibilitat del 95%) per a cada subàrea

Finalment, després d'estimar els paràmetres, es va reconstruir el procés tenint en compte que $Y_i|X_i \sim \text{Binom}(x_i, q_t)$.

L'anàlisi va mostrar que les activitats de formació i sensibilització dirigides als professionals de la salut primària tenen un impacte significatiu en la reducció del subregistre de casos de violència de gènere (Moriña et al., 2024). A més, el model va revelar que es necessitarien més de vint anys per a que es revelés la totalitat del fenomen de la violència de gènere a Catalunya, sense la implementació d'activitats addicionals de sensibilització (Moriña et al., 2024).

L'estudi també va examinar l'impacte del confinament obligatori durant la pandèmia de la COVID-19, trobant un augment de mitjana de 1.25 casos per setmana, el que suggereix un efecte significatiu de les mesures de confinament en la incidència de violència de gènere (Moriña et al., 2024).

2.3 Mètriques per a la quantificació de l'impacte de la COVID-19 al número de trucades

2.3.1 Test d'arrels unitàries de Dickey-Fuller

Amb l'objectiu de veure l'impacte del confinament per la COVID-19, analitzarem si les propietats de la sèrie temporal del nombre de trucades a la línia d'atenció a dones en situació de violència masclista canvien pre i postconfinament mitjançant el test de Dickey-Fuller.

Propòsit: El Test de Dickey-Fuller Augmentat és un test que permet determinar la presència d'una arrel unitària en les sèries de temps i, per tant, permet detectar si hi ha o no *estacionarietat* (Cuéllar, 2021).

En altres paraules, s'utilitza per a detectar estadísticament la presència de conducta tendencial en les sèries temporals mitjançant un contrast d'hipòtesis.

Considerem un procés estocàstic de la forma:

$$Y_t = \phi Y_{t-1} + u_t, -1 \leq \phi \leq 1$$

On Y_t és la sèrie de temps i u_t és un terme d'error de soroll blanc.

Hipòtesis:

$$H_0 : \beta = 0 \quad (\text{és equivalent a } \phi = 1)$$

$$H_1 : \beta < 0 \quad (\text{és equivalent a } \phi < 1)$$

La hipòtesi nul·la del test d'ADF estableix que la sèrie de temps té una arrel unitària, és a dir, és no estacionària. La hipòtesi alternativa, d'altra banda, suggereix que la sèrie és estacionària.

La idea darrere del test és avaluar si el coeficient de la variable endarrerida és significativament diferent de zero després d'eliminar l'efecte dels termes endarrerits addicionals.

El procediment del test de ADF consisteix a estimar un model de regressió lineal en què la variable dependent és la sèrie temporal en qüestió i les variables independents són la mateixa sèrie retardada i els seus retards. Per estimar els coeficients de la regressió, s'utilitza un mètode de mínims quadrats ordinaris, i es calcula un estadístic de prova basat en la diferència entre el valor estimat del coeficient de la variable retardada i la seva desviació estàndard.

Aquest estadístic de prova es compara amb els valors crítics de la distribució de Dickey-Fuller per decidir si es rebutja o no la hipòtesi nul·la. Si l'estadístic de prova és inferior al valor crític, es pot rebutjar la hipòtesi nul·la i concloure que la sèrie temporal és estacionària.

També podem rebutjar la hipòtesi nul·la si el p-valor de la prova és inferior a algun nivell de significació. En aquest cas, utilitzarem un nivell del 95%.

La sintaxi utilitzada per a aquest test en el llenguatge R ('adf.test') ([RDocumentation](#), nd), utilitza els paràmetres 'x', on s'indiquen les dades, 'nlag' i 'output', que per defecte és TRUE.

El paràmetre 'nlag' es refereix al nombre de retards (lags) de les diferències de la sèrie

2.3. Mètriques per a la quantificació de l'impacte de la COVID-19 al número de trucades

temporal que s'inclouen en el model per a corregir l'autocorrelació dels errors. L'elecció del nombre de retards depèn de la naturalesa de les dades. Es fa ús de 'nlags=0' quan la sèrie és molt simple i curta on no s'espera autocorrelació en els errors. Mentre que s'utilitza 'nlags>0' quan és necessari incloure retards addicionals per a corregir l'autocorrelació. Els retards addicionals ajuden a millorar l'especificació del model i evitar problemes d'autocorrelació residual.

2.3.2 Causal Impact

Per quantificar l'impacte del confinament per COVID-19 en el nombre de trucades al telèfon d'atenció contra la violència masclista, també utilitzem el paquet CausalImpact a R. Desenvolupat per Google, CausalImpact està dissenyat per estimar l'efecte causal d'una intervenció en un context de sèries temporals (Brodersen et al., 2015).

El paquet construeix un model de sèrie temporal estructural bayesià donat una sèrie de temps de resposta i un conjunt de sèries temporals de control. Amb aquest model s'intenta predir el contrafactual, o com hauria canviat la mètrica de resposta després de la intervenció si no hagués succeït. Vegeu Brodersen et al., *Annals of Applied Statistics* (2015) per obtenir més informació.

Pel que fa a les hipòtesis del model, suposem que existeix un conjunt de sèries temporals de control que no es van veure afectades per la intervenció. Si ho fossin, podríem subestimar o exagerar sense voler l'impacte real. Alternativament, podríem creure erròniament que es va produir un efecte quan, de fet, no hi havia. A més, el model suposa que la relació preestablerta entre les covariables i les sèries temporals tractades es mantindrà durant el període posterior.

Si es compleixen les condicions esmentades anteriorment, el paquet pretén fer la inferència contrafactual tan senzilla com ajustar un model de regressió, però molt més potent. La funció CausalImpact() és l'únic punt d'entrada del paquet. La funció crea un model de sèrie temporal, executa una inferència posterior sobre el contrafactual i retorna un objecte CausalImpact donat una sèrie de temps de resposta i un conjunt de sèries temporals de control.

La sintaxi utilitzada en el llenguatge R es troba al següent listing:

```
1 library(CausalImpact)
2 impact <- CausalImpact(data = NULL, pre.period = NULL, post.period = NULL)
```

Listing 2.1: Sintaxi aplicada en R pel model CausalImpact

Els resultats es poden resumir en una taula, una descripció verbal o un gràfic.

A la següent figura 2.3 observem un exemple de l'output que et retorna la funció 'plot()'.

```
1 plot(impact)
```

Listing 2.2: Sintaxi aplicada en R pel plot de la funció CausalImpact

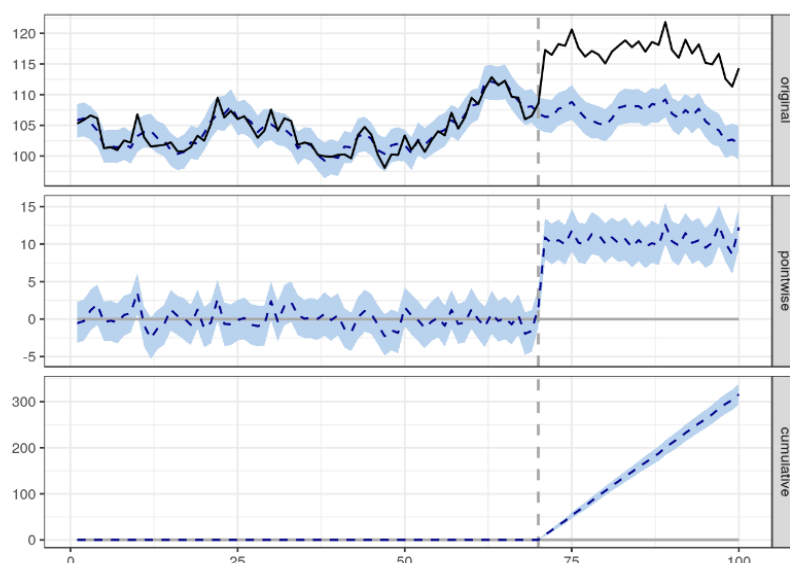


Figura 2.3: Exemple de l'output del plot de la funció CausalImpact

El "plot" té tres panells per defecte 2.3. Les dades i una predicció contrafactual per al període posterior al tractament es mostren al primer plafó. La discrepància entre les dades observades i les prediccions contrafactuals es mostra al segon plafó. Aquest és l'efecte causal puntual estimat pel model. El tercer plafó suma les contribucions puntuals del segon plafó, donant com a resultat un gràfic de l'efecte acumulat de la intervenció.

Per a obtenir un resum numèric dels resultats, s'utilitza el següent listing:

```
summary(impact)
```

Listing 2.3: Sintaxi aplicada en R pel summary de la funció CausalImpact

```
## Posterior inference {CausalImpact}
##
##               Average      Cumulative
## Actual         117         3511
## Prediction (s.d.) 107 (0.37) 3196 (11.03)
## 95% CI          [106, 107] [3174, 3217]
##
## Absolute effect (s.d.) 11 (0.37) 316 (11.03)
## 95% CI          [9.8, 11] [294.9, 337]
##
## Relative effect (s.d.) 9.9% (0.35%) 9.9% (0.35%)
## 95% CI          [9.2%, 11%] [9.2%, 11%]
##
## Posterior tail-area probability p: 0.001
## Posterior prob. of a causal effect: 99.9%
##
## For more details, type: summary(impact, "report")
```

Figura 2.4: Exemple de l'output del summary de la funció CausalImpact

Com observem a l'exemple de la Figura 2.4, la mitjana (al llarg del temps) durant la fase

posterior a la intervenció es discuteix a la columna 'Average'. Quan es considera una variable resposta que representa una quantitat de flux (e.g. nombre de trucades) la columna 'Cumulative' és útil ja que agrega punts de temps individuals.

Per obtenir orientació addicional sobre la interpretació correcta de la taula resum ("summary"), el paquet proporciona una interpretació verbal, que es pot imprimir mitjançant:

```
summary(impact, "report")
```

Listing 2.4: Sintaxi aplicada en R pel "report" de la funció CausalImpact

Encara que l'opció més simple és deixar que el paquet decideixi com construir un model de sèrie temporal per a les dades disponibles, tanmateix, hi ha diverses opcions que ens permeten tenir una mica més de control sobre aquest procés. Aquestes opcions es passen a 'model.args' com a elements de llista individuals. Una altra opció, en lloc d'utilitzar el model predeterminat, es pot utilitzar el paquet *bsts* per especificar el nostre propi model, proporcionant major grau de flexibilitat. Veure la següent font per a més informació: [An R package for causal inference using Bayesian structural time-series models](#).

2.4 Predicció de Sèries Temporals

Les nostres dades, tot i estar recollides i registrades diàriament, hem decidit treballar-les mensualment. Per aquest motiu, a l'hora de fer l'anàlisi s'ha de tenir en compte el factor temps i, com a conseqüència, s'apliquen diferents metodologies estadístiques que s'adeqüen a les sèries temporals.

En aquest apartat, expliquem les diferents metodologies aplicades a la sèrie del número mensual de trucades a la Línia d'Atenció contra la violència de gènere, del 2013 al 2024. El supòsit d'aplicar dos models recau en la necessitat d'identificar quin d'ells s'adequa millor a les nostres dades, per tal d'optimitzar la precisió i eficiència de les prediccions.

A continuació ens endinsem en els models ARIMA i les Xarxes Neuronals Artificials.

2.4.1 Introducció a les Sèries Temporals

Abans de començar amb l'explicació teòrica dels mètodes convencionals emprats a l'estudi, és essencial contextualitzar les característiques de les sèries temporals.

Definició: Una Sèrie Temporal és una seqüència de N observacions (dades) sobre una variable ordenades cronològicament i recollides de manera equiespaiada en el temps.

La seva expressió matemàtica és:

$$\{X_t\}_{t=1,\dots,N} = \{X_1, X_2, \dots, X_N\}$$

on X_t és l'observació nº t ($1 \leq t \leq N$) de la sèrie, N és el nombre d'observacions que té la sèrie sencera i t ens indica el moment temporal.

Així, les sèries temporals estan constituïdes per diferents components, els quals faciliten la identificació del tractament necessari en funció de la seva presència o absència.

2.4.1.1 Components de les sèries temporals

Les components de variació que es consideren habitualment són les següents:

1. **Component de Tendència:** Es denomina tendència, creixent o decreixent, a un comportament o moviment durant un llarg període de temps. La tendència s'identifica com un moviment suau de la sèrie a llarg termini.
2. **Component Estacional:** Registra les fluctuacions que es produeixen en períodes iguals o inferiors a un any, repetits de forma regular als diferents anys.
3. **Component cíclica:** Aquest component reflecteix comportaments recurrents, amb un període superior a un any.
4. **Component irregular:** també anomenat "soroll" o component aleatòria, recull les alteracions de la sèrie que no són regulars i són purament aleatòries o irregulars. En conseqüència, recull la incapacitat del model per explicar a la perfecció el comportament de la sèrie temporal.

2.4.2 Processos estocàstics

Abans d'iniciar amb l'explicació de la metodologia Box-Jenkins, és necessari introduir els conceptes previs que aquesta metodologia té en compte.

Principalment, per poder aplicar l'enfocament de Box-Jenkins, les sèries temporals s'han de processar com a processos estocàstics, és a dir, s'han de satisfer una sèrie de propietats.

Definició: Un procés estocàstic és un conjunt de variables aleatòries, ordenades cronològicament i que, en general, s'identifiquen amb el temps.

Definició: Donat un conjunt d'observacions $(X_{t1}, \dots, X_{t2})$, un procés és estacionari si la seva distribució conjunta és la mateixa al llarg del procés:

$$F_{(X_{t_1}, \dots, X_{t_2})}(x_{t_1}, \dots, x_{t_2}) = F_{(X_{t_1+h}, \dots, X_{t_2+h})}(x_{t_1+h}, \dots, x_{t_2+h}) \quad \forall t, h$$

Concretament, en sentit dèbil, un procés estocàstic és estacionari si les seves propietats estadístiques no canvien amb el temps, és a dir, no depenen de l'origen de la sèrie. Es caracteritza per una mitjana, una variància i una covariància constants en el temps.

Així, si la sèrie temporal estudiada compleix les propietats anteriors, serà necessari, només, que segueixi una distribució normal. En aquest sentit, al ser el procés estacionari, podrem aplicar la metodologia Box-Jenkins.

2.4.2.1 Propietats processos estacionaris

Introduïm les propietats i característiques que presenten els processos estacionaris, necessaris per a identificar i validar els models.

Primerament, definim la funció d'autocovariància pels processos estocàstics:

$$\gamma(h) = \text{Cov}(X_t, X_{t+h}) = E((X_t - \mu)(X_{t+h} - \mu))$$

I té les següents propietats:

- $\gamma(0) = \sigma^2 > 0$
- $\gamma(k) = \gamma(-k) \quad \forall k$
- $|\gamma(k)| \leq \gamma(0)$

D'altra banda, la funció d'autocorrelació simple (FAS), o en anglès *Auto-correlation Function* (ACF) es defineix com:

$$\rho(h) = \rho(X_t, X_{t+h}) = \frac{E((X_t - \mu)(X_{t+h} - \mu))}{\sqrt{E(X_t - \mu)^2 E(X_{t+h} - \mu)^2}} = \frac{\gamma(h)}{\gamma(0)}$$

Que presenta les següents propietats:

- $\rho(0) = 1$
- $\rho(k) = \rho(-k) \quad \forall k$
- $|\rho(k)| \leq 1 \quad \forall k$

Per comprovar que una sèrie és estacionària, hem d'estudiar el *Correlogram* (gràfic de l'ACF $\rho(k)$ en funció del retard k) i observar que l'ACF decaigui ràpidament cap al 0.

Després, cal estudiar la funció d'autocorrelació parcial (FAP), o en anglès *Partial Auto-Correlation Function (PACF)*, necessària per a definir la resta de paràmetres del model SARIMA.

$$\text{corr}(X_{t+h}, X_t \mid X_{t+h-1}, \dots, X_{t+1})$$

En darrer lloc, definim el terme de soroll blanc, necessari per facilitar la lectura dels pròxims apartats:

- **Soroll blanc:** cas particular del processos estacionaris, on les variables són incorrelacionades $\rho(h) = 0$, $\forall h$ distribuïdes al llarg del temps, amb la mitjana centrada en zero.

2.4.3 Metodologia Box-Jenkins

Una vegada hem explicat la terminologia necessària, introduïm la metodologia Box-Jenkins.

Box Jenkins és un enfocament sistemàtic per identificar, estimar i verificar models $ARMA(p, q)$, $ARIMA(p, d, q)$ o $SARIMA(p, d, q)(P, D, Q)_s$ per a sèries temporals. Se segueix un procés iteratiu fins a trobar el model òptim (mirar [Brockwell and Davis \(2002\)](#)). Es segueix l'estructura de l'esquema de la Figura 2.5.

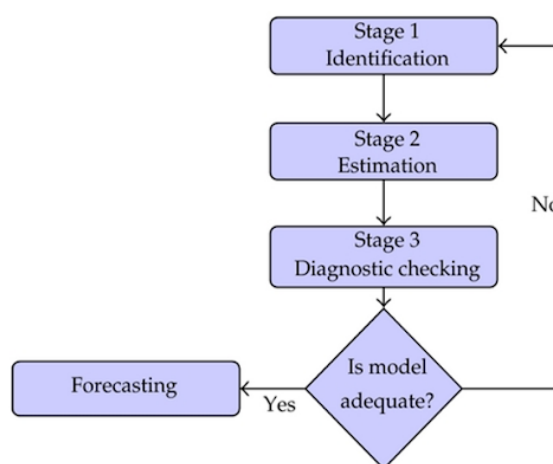


Figura 2.5: Esquema de la metodologia Box-Jenkins

2.4.3.1 Identificació

Primer, analitzem la sèrie temporal per identificar si és estacionària o no, de manera que s'avalua una variància constant i una mitjana centrada en zero. Si la variància no és constant, transformarem la sèrie amb logaritme, mentre que si la mitjana no està centrada en zero, aplicarem tantes diferenciacions regulars com calguin, per a aconseguir aquesta estacionarietat.

D'altra banda, observarem si la sèrie presenta patrons estacionals per a què es pugui recollir al model. En cas que s'observi estacionalitat, caldrà l'aplicació d'una diferenciació estacional a la sèrie.

Una vegada aplicades aquestes transformacions, avaluem la funció d'autocorrelació (ACF) i autocorrelació parcial (PACF). En aquest sentit, si la funció decau ràpida cap a 0, es verifica que estem davant d'una sèrie temporal estacionària.

Al tenir preparada la sèrie, procedim a identificar un model SARIMA (*Seasonal Autorregressive Integrated Moving Average*). La seva expressió general és:

$$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)Z_t$$

On $Z_t \sim WN$, que indica que es distribueix com un soroll blanc.

Als models SARIMA, la primera part segueix una estructura $ARIMA(p,d,q)$ i ens indica la part regular del model. La segona part, segueix un $ARIMA(P,D,Q)$ i fa referència a la part estacional.

D'una banda, el paràmetre d defineix l'ordre de diferenciacions regulars, mentre que el paràmetre D ens indica l'ordre de diferenciacions estacionals. D'altra banda, els paràmetres p i q defineixen l'ordre d'autoregressió o mitjana mòbil regular, obtinguts observant els primers retards (lags) significatius de l'ACF i el PACF. Quant als paràmetres P i Q , que es refereixen a la part estacional, s'obtenen identificant els retards múltiples a s (el paràmetre que representa l'estacionalitat).

2.4.3.2 Estimació

Una vegada hem identificat el model i els seus paràmetres, estimem els seus coeficients a partir de mètodes no lineals per trobar la funció de màxima versemblança ([Gardner et al., 1980](#)).

A partir de l'estimació, es verifica si aquests coeficients són significatius, de manera que es pugui simplificar el model final. Així, un coeficient es considera significatiu si el seu estadístic (*T-Ratio*) és major a 2 en valor absolut.

En cas que s'hagi plantejat més d'un model, triarem el model òptim mitjançant el criteri de l'AIC (Criteri d'Informació d'Akaike).

$$AIC = -2\log L(\phi, \theta, \sigma^2|X) + 2p$$

On $L(\phi, \theta, \sigma^2|X)$ és la funció de màxima versemblança i p és el nombre de paràmetres. Es tria el model que presenta un menor AIC.

2.4.3.3 Validació

A la tercera etapa, ja simplificat el possible model, analitzem els seus residus per verificar que siguin soroll blanc: independents, seguint una distribució normal amb mitjana i variància constant.

Aquesta validació la farem tant visualment mitjançant gràfics, com a partir de tests estadístics. De manera que si els residus compleixen aquesta sèrie de propietats, acceptarem el model com adequat i l'utilitzarem per a la previsió. Si els residus, en canvi, no són validats, repetirem les etapes anteriors fins a trobar el model òptim.

2.4.3.4 Model adequat i predicció

A la quarta i última etapa, farem les prediccions necessàries amb el model que s'acaba de validar. Es calculen diferents mètriques d'error per a avaluar el rendiment d'aquestes prediccions (vegeu 2.5).

2.4.4 Xarxes Neuronals Artificials (ANN)

Les Xarxes Neuronals Artificials, (ANN) per les seves sigles en anglès, són un tipus sofisticat d'algorismes d'aprenentatge automàtic inspirats en les xarxes neuronals biològiques del cervell humà. Són molt eficaços per modelar relacions complexes i no lineals, el que les fa adequades per a tasques com la previsió de sèries temporals on els mètodes tradicionals, com ARIMA, poden ser inadequats a causa dels canvis en l'estructura de dades i el gradient.

Una xarxa neuronal es pot pensar com una xarxa de "neurones" organitzades en tres capes connectades per enllaços. Els predictors (o entrades) formen la capa inferior i les previsions (o sortides) formen la capa superior. També pot haver-hi capes intermèdies que continguin "neurones amagades". Aquesta capa oculta, permet la relació no lineal entre les entrades i les sortides, permetent al model més graus de llibertat (Salguero, 2021).

La relació entre la sortida (y_t) i les entrades ($y_{t-1}, \dots, y_{t-p}$) té la següent representació matemàtica:

$$y_t = w_0 + \sum_{j=1}^q w_j \cdot g \left(w_{0,j} + \sum_{i=1}^p w_{i,j} \cdot y_{t-i} \right) + \epsilon_t$$

on, $w_{i,j}$ ($i=0, 1, 2, \dots, p$, $j=1, 2, \dots, q$) i w_j ($j=0, 1, 2, \dots, q$) són paràmetres del model que es denominen pesos de connexió; p és el nombre de nodes d'entrada; i q és el nombre de nodes ocults. A la Figura 2.6 observem l'estructura d'una xarxa neuronal amb els seus paràmetres.

Si només s'utilitzen les entrades i les sortides d'una xarxa neuronal, la funció del model és equivalent a una regressió lineal. Per això, una xarxa neuronal necessitarà funcions d'activació no lineals i unitats o capes ocultes per diferenciar-se.

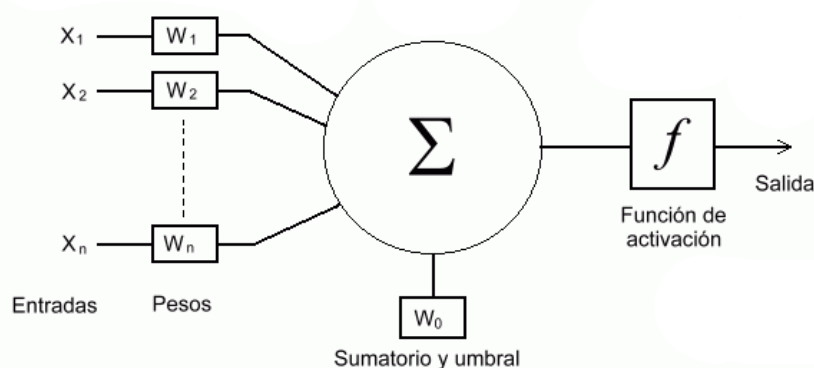


Figura 2.6: Estructura d'una xarxa neuronal

Ara, a la Figura 2.7, il·lustrem una xarxa neuronal petita, que posseeix una capa oculta i en ella neurones ocultes, que permetran distribuir els pesos dels paràmetres a la vegada que s'utilitzen funcions d'activació no lineals.

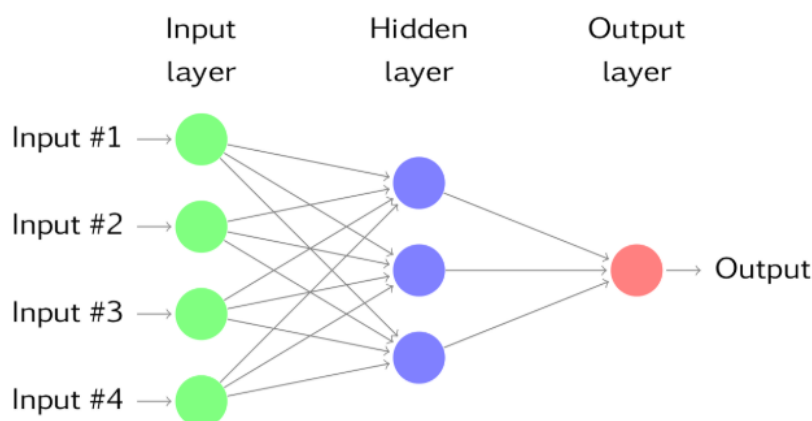


Figura 2.7: Exemple d'una xarxa neuronal amb quatre entrades i una capa oculta amb tres neurones ocultes.

A les xarxes neuronals, els nodes de cada capa estan connectats amb altres nodes o fins i tot amb ells mateixos (cada connexió és diferent a més perquè també tenen un pes associat), depenent de l'arquitectura i la interconnexió de totes les neurones d'una xarxa, poden classificar-se en diferents categories: Xarxes Neuronals Feed-Forward, Xarxes Neuronals Convolucionals (CNN) i Xarxes Neuronals Recurrents (RNN).

En termes de sèries temporals, les característiques d'entrada solen consistir en observacions anteriors de la sèrie, mentre que la capa de sortida emet el valor previst per al següent pas de temps. La capa oculta treballa igual, cada neurona d'una capa oculta calcula una suma ponderada de les seves entrades, afegeix un terme de biaix i aplica una funció d'activació per

introduir la no-linealitat.

Entre els avantatges de les xarxes neuronals per a la previsió de sèries temporals, trobem: segueixen un modelatge no lineal, és a dir, poden capturar relacions no lineals complexes a les dades; tenen flexibilitat, i però això poden gestionar una varietat de característiques d'entrada i no estan limitats pels supòsits dels models estadístics tradicionals; a més, les ANN es poden escalar per adaptar-se a grans conjunts de dades i arquitectures complexes.

A continuació descrivim el procés d'implementació que s'ha utilitzat a partir del paquet 'neuralnet' a R.

2.4.4.1 Implementació de xarxes neuronals artificials a R: paquet NeuralNet

El paquet 'neuralnet' a R està dissenyat per crear i entrenar xarxes neuronals de feed-forward. En una xarxa neuronal de feed-forward, la informació es mou en una direcció: des dels nodes d'entrada, a través dels nodes ocults (si n'hi ha) i finalment als nodes de sortida. No hi ha cicles ni bucles a la xarxa.

La funció `neuralnet` a R presenta els següents paràmetres: "*formula*" és la fórmula que ha d'aprendre la xarxa neuronal, a "*data*", s'especifiquen les dades que contenen les variables especificades a la fórmula, "*hidden*" és un vector d'enters que especifica el nombre de neurones ocultes (vèrtexs) a cada capa, "*rep*" és el nombre de repeticions per a l'entrenament de la xarxa neural i "*threshold*" és un valor numèric que especifica el llindar per als derivats parcials de la funció d'error com a criteri de parada (Coneo, 2019).

La funció d'activació que es troba definida dins de la funció 'neuralnet' és la sigmoide, o funció logística, que es defineix matemàticament com:

$$\sigma(x) = \frac{1}{1 + \exp[-x]}$$

No obstant però, abans d'ajustar el model ANN, les dades de la sèrie temporal s'han de preprocessar per estructurar-les adequadament a la xarxa neuronal.

1. Normalització de les dades:

Les dades s'escalen a un interval de [0, 1] per garantir que totes les característiques d'entrada contribueixen per igual al model. Això s'aconsegueix calculant els valors mínim (mi) i màxim (ma) de la sèrie temporal i aplicant una transformació.

Aquest pas de normalització ajuda a accelerar el procés de formació i a millorar la convergència del model.

2. Variables retardades:

Per capturar dependències temporals, es creen funcions retardades de la sèrie temporal. Aquí, el nombre d'observacions retardades s'estableix pel paràmetre m . Per a cada pas de temps, les m observacions anteriors s'inclouen com a característiques d'entrada.

Finalment, la sèrie temporal intersecada amb els retards, es converteix en una base de dades.

Després, les dades es divideixen en conjunts d'entrenament (train) i proves (test) per avaluar el rendiment del model. Normalment, els punts de dades més recents es reserven per fer proves per avaluar el rendiment del model amb dades no vistes.

Posteriorment, es construeix una fórmula per definir la relació entre la variable resposta (valors futurs de la sèrie temporal) i les variables explicatives (característiques retardades). Així, consegüentment, s'utilitza la funció '*neuralnet*' per entrenar la xarxa neuronal amb les dades d'entrenament.

Finalment, utilitzem la xarxa neuronal entrenada per a fer prediccions tant sobre els conjunts de dades d'entrenament com de proves. Es tornen a transformar les prediccions a l'escala original utilitzant la inversa de la transformació de normalització.

L'avaluació del rendiment del model es fa mitjançant diferents mètriques d'error, les quals es veuran a la següent secció.

2.5 Mètriques d'error

Una vegada hem explicat les metodologies aplicades per modelitzar la sèrie temporal del número de trucades a línia d'atenció a la Violència de Gènere, es decideix quin mètode s'ajusta millor segons diferents mètriques d'error o rendiment.

És fonamental obtenir precisió a les dades d'entrenament, però també és crucial que els resultats siguin fiables i aproximats per a dades desconegudes. Això és important perquè els algoritmes no poden ser 100% eficients amb aquestes dades, ja que podrien estar esbiaixats i presentar un sobre ajustament.

Tenint en compte que y_i es refereix a les observacions i $\hat{y}_i$ a les prediccions:

- **L'error quadràtic mitjà (RMSE):** aquest indicador mostra la diferència dels valors pronosticats pel model i els valors reals registrats, és a dir, s'ocupa de les desviacions del valor real. Es pot interpretar com una mesura de la capacitat de predicció del model. Encara que és una mesura imperfecta per a l'avaluació, és molt comú, essent una molt bona opció quan s'entrena una xarxa neuronal profunda.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- **Mitjana de l'error absolut (MAE):** aquesta mesura ens ajuda a identificar la magnitud de la diferència entre la predicció i el valor real, i es calcula com la mitjana de la suma dels valors absoluts de les diferències entre les prediccions i els valors reals. En altres termes, calcula la mitjana absoluta dels errors en les prediccions, sense considerar el seu signe.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- **Mitjana de l'error absolut en percentatge (MAPE):** és una mètrica que mesura el valor promig de l'error en percentatge. A l'expressar-se en percentatge, és independent de l'escala i fàcil d'interpretar.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100$$

- **L'error quadràtic mig percentual (RMSPE):** Aquesta mesura té les mateixes característiques que l'RMSE però expressada com a valor percentual, cosa que simplifica la interpretació dels resultats.

$$RMSPE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \cdot 100$$

- **Mitjana de l'error absolut normalitzat (nMAE):** amb aquest indicador es poden comparar dades amb diferents escales segons el MAE.

$$nMAE = \frac{MAE}{\hat{y}} = \frac{\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|}{\frac{1}{n} \sum_{i=1}^n |y_i|}$$

- **L'error quadràtic mig percentual normalitzat (nRMSE):** amb aquest indicador es poden comparar dades amb diferents escales segons el RMSE.

$$nRMSE = \frac{RMSE}{\hat{y}} = \frac{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \right)^2}}{\frac{1}{n} \sum_{i=1}^n |y_i|}$$

Així, per a poder determinar quin model ofereix les prediccions més precises, compararem les mètriques de regressió i seleccionarem el model que produeixi els resultats més baixos per a la mètrica estudiada.

Capítol 3

Resultats

En aquest capítol ens centrarem únicament en els resultats obtinguts per l'estudi realitzat, la finalitat del qual és donar resposta als objectius plantejats. L'anàlisi següent ha estat fet mitjançant les bases de dades mencionades al capítol 2.

3.1 Validació externa del model de subnotificació

L'objectiu principal d'aquest primer apartat de resultats, és comprovar si l'estimació de l'evolució que produeix el model de subnotificació és similar a l'evolució del número de trucades a la línia d'atenció contra la Violència Masclista. La premissa subjacent és que aquest servei proporciona una quantificació més acurada del fenomen que no pas els registres policials o clínics.

L'anàlisi abasta el període de 2013 a 2022, al context específic de l'Àrea Metropolitana Nord de Barcelona, permetent avaluar la precisió i l'eficàcia del model extern, així com el seu potencial per capturar aquests casos no registrats que viuen sota l'«iceberg» de la Violència de Gènere.

Abans de començar a analitzar la mera validació gràfica del model a partir de la comparació de les dues bases de dades, presentarem les dades individualment.

3.1.1 Evolució mensual del nombre de trucades a la Línia d'Atenció contra la Violència Masclista

El gràfic següent 3.1 presenta l'evolució mensual del nombre de trucades a la línia d'atenció contra la violència de gènere a l'Àrea Metropolitana Nord de Barcelona, des del 2013 fins a l'any 2024.

Observem com entre l'any 2013 i el 2020, el nombre de trucades mensuals va fluctuar entre 150 i 300. Durant aquest període, s'observa la presència d'alguns pics ocasionals, però en

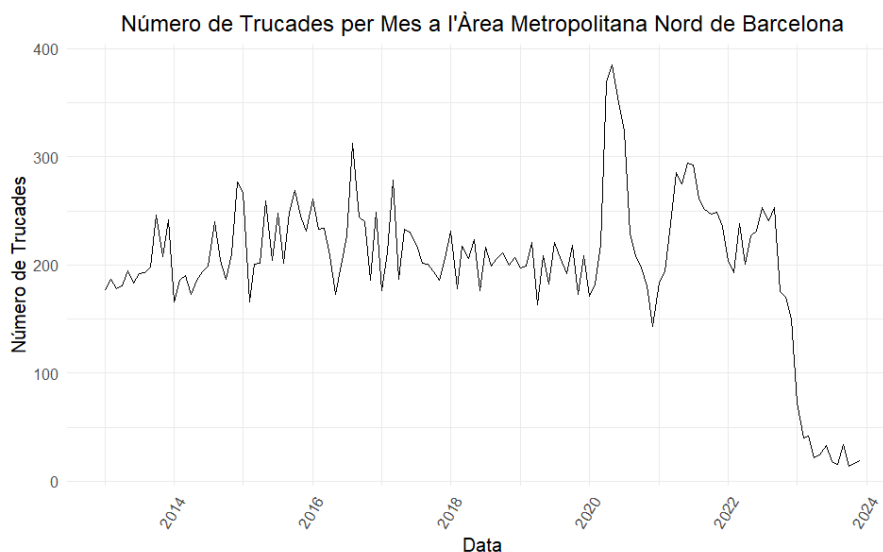


Figura 3.1: Número de Trucades per mes a l'Àrea Metropolitana Nord de Barcelona

general la tendència d'aquest període és força constant. D'altra banda, és notable que al 2016 i al 2017 es van registrar diversos pics significatius, superant les 300 trucades en alguns mesos.

A l'inici del 2020, coincidint amb el començament de la pandèmia de la COVID-19 i els confinaments, és notable un augment significatiu al número de trucades. Així, el gràfic mostra un pic notablement alt al número de trucades durant el primer trimestre de 2020, assolint el seu màxim al mes d'abril. Aquest pic coincideix amb els primers mesos de confinament estricte a Espanya. Després del pic inicial, el número de trucades comença a disminuir, encara que segueix essent més alt que als anys anteriors a la pandèmia. Finalment, a partir del 2022 intuïm una gran disminució en el número de trucades, que continua fins al 2024.

A la taula següent, Taula 3.1, observem un resum de les seves descriptives més destacables. El número mínim de trucades mensuals va ser de 14, mentre que el màxim va ser de 385, que coincideix amb els mesos de confinament estricte a Espanya. La mediana i la mitjana, ambdues al voltant de 200 trucades, mostren que la distribució està centrada en aquest valor. La desviació estàndard de 67,621 reflecteix una variabilitat considerable al número de trucades mensuals.

Mínim	Mediana	Mitjana	Màxim	Desviació
14	204	201.05	385	67.62

Taula 3.1: Descriptives del Número de Trucades

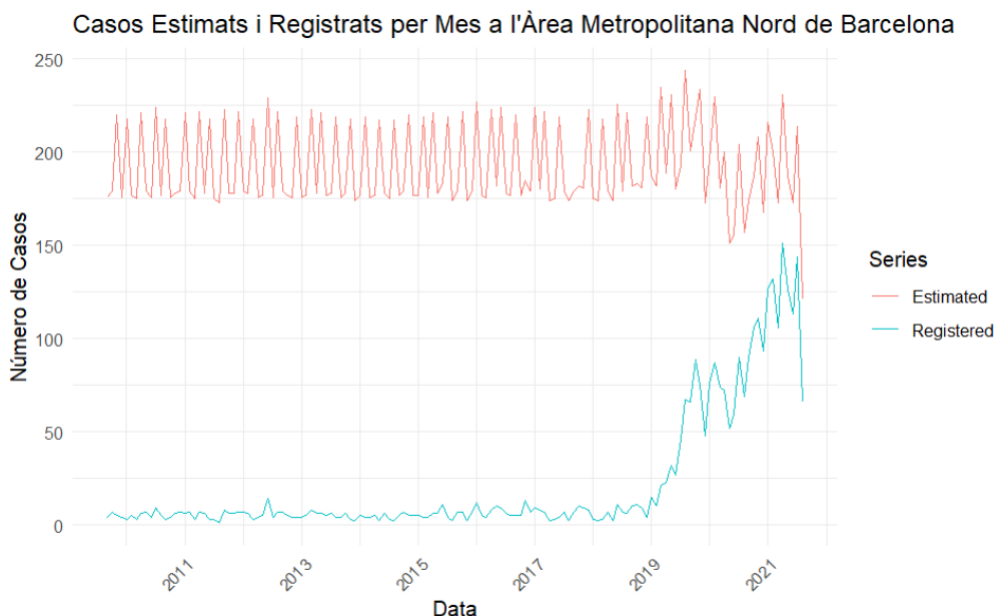


Figura 3.2: Casos Estimats i Registrats per mes a l'Àrea Metropolitana Nord de Barcelona

3.1.2 Anàlisi del Model Extern de Subnotificació

Així mateix, a continuació analitzem el model que es pretén validar, on s'observen tant els casos reals "registrats" de VG, com els casos estimats pel model.

En aquest cas, a la Figura 3.2, veiem com la franja temporal d'aquest model s'estén des del gener de 2010 fins al desembre de 2021.

Els casos registrats, representats en blau, mostren una tendència bastant estable, des del 2010 fins a aproximadament el 2019, on el número de casos es manté baix, amb menys de 50 casos mensuals en promig. A partir de 2018, veiem un augment significatiu en el número de casos registrats, que continua incrementant-se fins a finals de 2021.

D'altra banda, els casos estimats pel model de subnotificació, representats en vermell, mostren una tendència diferent. Veiem com la tendència és estable al llarg del temps. Contràriament als 50 casos mensuals dels casos registrats, als casos estimats s'observa com el número de casos fluctua al voltant d'un promig de 200 casos mensuals, a tota la llargada de la sèrie històrica. A més, observem una petita variació estacional, amb pics i valls que es repeteixen de manera consistent cada any.

Amb tot, aquesta diferència entre casos reals i estimats es redueix al final del període analitzat, de manera que hi ha una convergència entre sèries.

A continuació, a la Taula 3.2, podem veure aquesta diferència numèricament. Els casos registrats presenten valors més petits i menys variables en comparació amb les estimacions.

	Mínim	Mediana	Mitjana	Màxim	Desviació
Casos registrats	1	6	21.49	151	34.64
Casos estimats	121	179	192.5	244	22.6

Taula 3.2: Descriptives dels Casos Registrats i Estimats

3.1.3 Comparació i validació del model

Finalment, per tal de validar el model, a continuació, comparem el model de subnotificació amb el número de trucades a la línia d'atenció contra la violència masclista a un mateix gràfic.

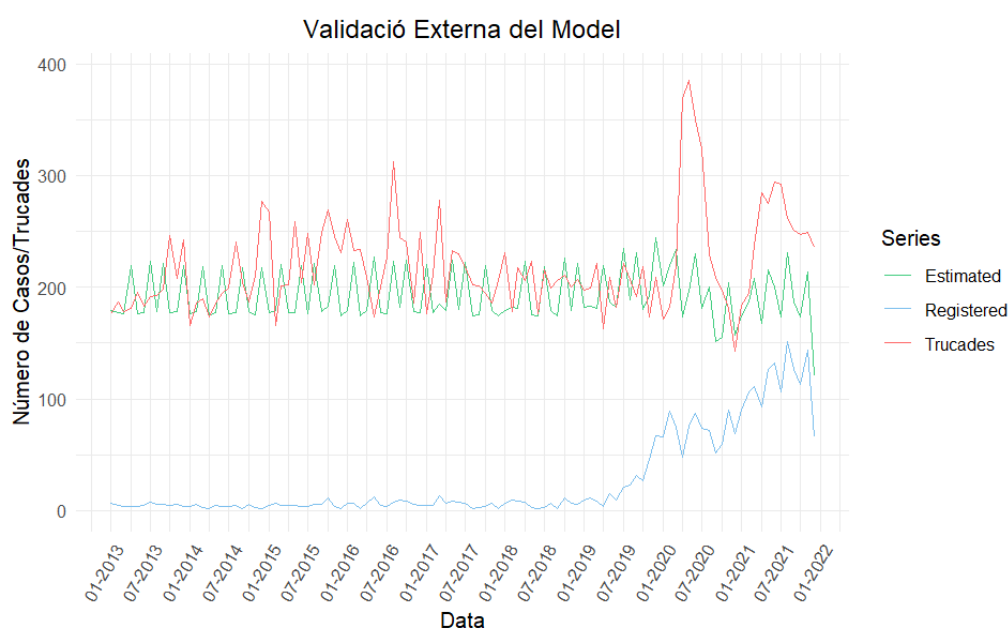


Figura 3.3: Validació del model amb el número de trucades

Tal com s'observa a la Figura 3.3, hem ajustat una finestra temporal on coincideixen les dades de les dues fonts. Així, el període d'estudi compren des de gener de 2013 fins a gener de 2022, on trobem: casos estimats del model de subnotificació (línia verd), casos detectats (línia blava) i número de trucades a la línia d'atenció a les víctimes (línia vermella).

Principalment, a partir del gràfic, veiem com durant tot el període els casos estimats coincideixen en promig amb el número de trucades, essent 192.4 i 219 el número mitjà de casos estimats i el número mitjà de trucades, respectivament. Així i tot, hi ha una gran diferència amb el nombre de casos registrats, que no superen els 21 casos de mitjana.

D'altra banda, als inicis del 2020, on s'observa el pic de 385 trucades a causa del confinament, veiem com el model no és capaç de registrar aquest augment, en comparació amb el nombre de trucades. Després, des de mitjans del 2020 fins a finals del 2021, tant les trucades com els casos estimats mostren una tendència descendent. No obstant això, els casos estimats presenten una major estabilitat i una reducció més moderada en comparació amb les trucades. Així mateix,

contràriament a la tendència descendent, els casos registrats des de finals del 2019 mostren una tendència alcista.

Finalment, respecte a la variabilitat associada, observem que tant els casos estimats com el número de trucades mostren pics i valls; no obstant però, els casos estimats es mantenen bastant estables, a diferència de la variabilitat que s'observa a les trucades, que és bastant major.

A la Taula 3.3, presentem aquesta comparació amb més detall, on es mostra un resum de les descriptives de les tres sèries analitzades, concretament dins la finestra temporal ajustada.

	Mínim	Mediana	Mitjana	Màxim	Desviació
Número de trucades	143	207.5	219	385	42.44
Casos estimats	121	181	192.4	244	22.99
Casos registrats	2	7	26.84	151	34.64

Taula 3.3: Resum Descriptives

3.2 Quantificació de l'impacte de la COVID-19

L'objectiu principal d'aquest següent apartat és analitzar i quantificar l'impacte de la COVID-19 al registre del número de trucades a la Línia d'Atenció contra la violència de gènere. L'anàlisi abasta el número mensual de trucades, a la línia d'atenció contra la Violència Masclista, a tota Catalunya.

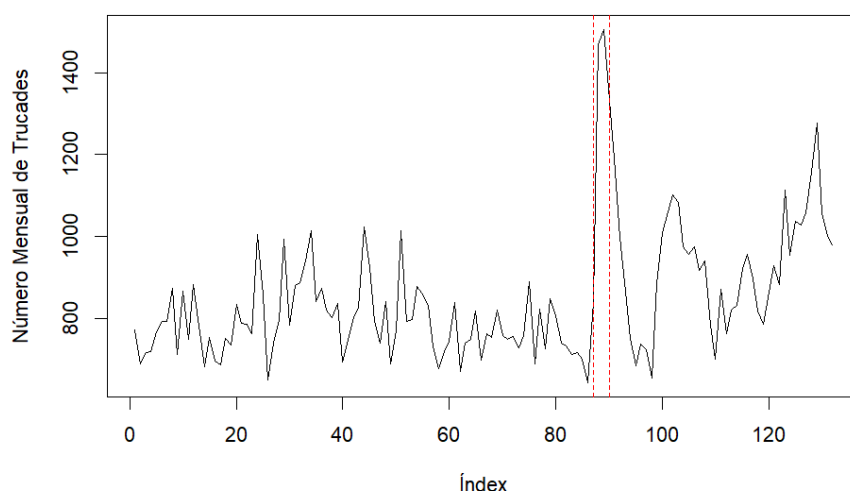


Figura 3.4: Número de trucades mensuals a la Línia d'Atenció contra la Violència Masclista

La Figura 3.4, mostra l'evolució del número mensual d'aquestes trucades al llarg del temps, des del gener del 2013 fins al desembre del 2023. Les línies vermelles representen la franja

temporal del període en què es va declarar, a Espanya, l'estat d'alarma per la COVID-19: del 14 de març fins al 21 de juny del 2020.

En aquesta gràfica, podem observar que abans de l'impacte de la pandèmia, el nombre de trucades es mantenia dins d'un rang relativament estable, amb oscil·lacions típiques al llarg dels mesos. Tanmateix, a partir del començament de les restriccions per la pandèmia, representat per la primera línia vermella, detectem un increment significatiu en el nombre de trucades, arribant a un pic màxim de 1506 trucades al mes de maig.

Després d'aquest pic, intuïm una petita disminució gradual en el nombre de trucades, seguit d'un gran augment mantingut fins al final de la sèrie.

A la Taula 3.4 comparem estadístics per als diferents períodes. S'observa una gran diferència en valors de mitjana i desviació tipus, abans i després del confinament.

	Mediana	Mitjana	Desviació
Període sencer	811	847.8	153.61
Període pre-confinament	771	788.6	84.75
Període post-confinament	940	937.8	154.52

Taula 3.4: Resum Descriptives de diferents franges temporals

3.2.1 Test de Dickey-Fuller

Per tal d'analitzar si les propietats d'estacionarietat de la sèrie canvien pre i post-confinament, duem a terme el test de Dickey-Fuller augmentat (ADF) en dos segments diferents de les dades:

1. Des de gener de 2013 fins a febrer del 2020 (abans de l'estat d'alarma)
2. Des de juliol del 2020 fins al desembre de 2023 (després de l'estat d'alarma)

Com hem vist visualment, abans de la pandèmia, es pot observar que el número de trucades no presentava una tendència clara, mantenint-se relativament estable amb fluctuacions moderades. Aquesta observació es corrobora amb el test de Dickey-Fuller augmentat (ADF), que indica que la sèrie era estacionària:

```
Augmented Dickey-Fuller Test
data: datesMens$Trucades[0:87]
Dickey-Fuller = -7.0288, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary
```

Figura 3.5: Test de Dickey-Fuller per al període pre estat d'alarma

L'estadístic de prova 3.5 és -7.0288, amb un p-valor menor a 0.01, el que suggereix que hi ha suficient evidència per rebutjar la hipòtesi nul·la de la presència d'una arrel unitària. Això suggereix que la sèrie és estacionària, i que les seves propietats estadístiques (com la mitjana i la variància) són constants en aquest període de temps.

Donats aquest resultat, confirmem que el número de trucades es manté en un rang relativament estable amb oscil·lacions típiques al llarg dels mesos abans de l'impacte de la pandèmia.

D'altra banda, després de l'impacte de l'estat d'alarma, a partir del punt de ruptura situat des del març de 2020 fins al juny (la línia vermella), es nota un augment significatiu en el número de trucades. Aquest increment és avaluat de nou amb el test de Dickey-Fuller per a la segona part de la sèrie temporal (sense tenir en compte els mesos de confinament), la qual cosa suggereix una tendència creixent:

```
Augmented Dickey-Fuller Test
data:  dadesMens$Trucades[95:132]
Dickey-Fuller = -2.6885, Lag order = 0, p-value = 0.3043
alternative hypothesis: stationary
```

Figura 3.6: Test de Dickey-Fuller per al període post estat d'alarma

L'estadístic de Dickey-Fuller 3.6 és -2.6885 amb un p-valor de 0.3043. Aquest p-valor és major al nivell de significança de 0.05, el que ens indica que no tenim suficient evidència com a rebutjar la hipòtesi nul·la d'una arrel unitària. Això suggereix que la sèrie no és estacionària en aquest període de temps, és a dir, les seves propietats estadístiques de la sèrie canvien amb el temps.

El resultat és coherent amb l'observació visual, on després del confinament, la sèrie mostra un comportament més variable amb tendència a l'alça i fluctuacions significatives, reflectint un canvi al patró de les trucades.

3.2.2 Model Causal Impact

Finalment, per quantificar de manera més precisa l'impacte de la COVID-19 al número de trucades, utilitzem l'anàlisi de "*Casual Impact*", que permet identificar i quantificar canvis en una sèrie temporal atribuïbles a un esdeveniment específic (veure Capítol 3 per a més informació). En aquest cas, l'esdeveniment en concret és l'estat d'alarma per la COVID-19. Els resultats d'aquesta anàlisi els presentem a la Figura 3.7 següent:

L'anàlisi d'impacte causal revela que la pandèmia de la COVID-19 va tenir un efecte significatiu en el nombre de registres a la línia d'atenció contra la violència masclista. En termes

Posterior inference {CausalImpact}		
Actual	Average	Cumulative
Prediction (s.d.)	938	40324
95% CI	786 (17)	33785 (711)
	[753, 817]	[32369, 35150]
Absolute effect (s.d.)	152 (17)	6539 (711)
95% CI	[120, 185]	[5174, 7955]
Relative effect (s.d.)	19% (2.5%)	19% (2.5%)
95% CI	[15%, 25%]	[15%, 25%]
Posterior tail-area probability p:	0.00102	
Posterior prob. of a causal effect:	99.89806%	

Figura 3.7: Resultats Summary Anàlisi Casual Impact

absoluts, es van observar 6.539 trucades addicionals durant el període analitzat, cosa que representa un augment mitjà del 19% en comparació amb les prediccions basades en la tendència històrica prèvia a la pandèmia. L'interval posterior del 95% de l'efecte mitjà és [120, 185]. Com que això n'exclou 0, concloem (correctament) que la intervenció va tenir un efecte causal en la variable resposta. Finalment, l'alta provabilitat posterior d'un efecte causal (99.898%) i el baix p-valor (0.00102) subratllen la robustesa d'aquestes troballes.

Per complementar els resultats numèrics de l'anàlisi d'impacte causal, presentem a continuació el gràfic que il·lustra els canvis en el nombre de crides a la línia d'atenció contra la violència masclista abans i després de l'inici del confinament causat per la pandèmia de la COVID-19.

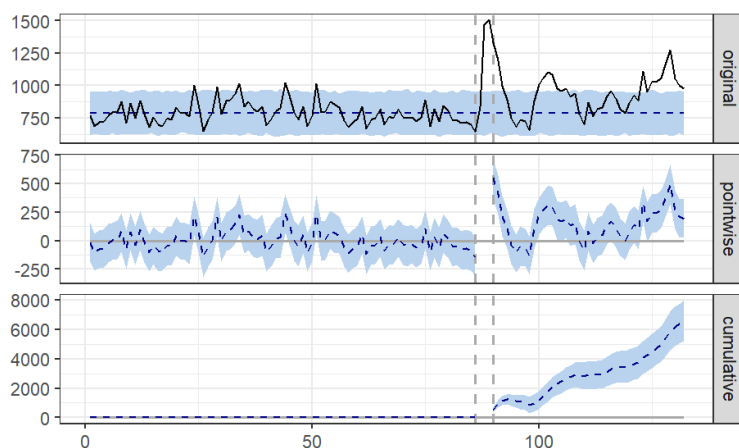


Figura 3.8: Resultats Plot Anàlisi Casual Impact

Al gràfic de la Figura 3.8 es visualitza clarament l'impacte de la COVID-19 al nombre de trucades a la línia d'atenció contra la violència masclista:

- **Abans de l'estat d'alarma:** La sèrie temporal original es manté dins dels límits de

predicció.

- **Després de l'estat d'alarma:** S'observa un augment pronunciat al nombre de trucades, que excedeix significativament els intervals de confiança predits, corroborant l'existència d'un efecte causal de la pandèmia.
- **Increment Puntual i Acumulat:** Tant les diferències puntuals com les acumulades destaquen l'augment en la demanda de serveis d'atenció, amb increments sostinguts i acumulats en el temps.

En resum, el gràfic complementa els resultats numèrics de l'anàlisi d'impacte causal i proporciona una visualització clara i contundent de l'increment en el nombre de trucades a la línia d'atenció contra la violència masclista a causa del confinament per la COVID-19.

3.3 Predicció del número mensual esperat de trucades a la Línia d'Atenció Contra la Violència Masclista

Aquest darrer apartat pretén donar resposta al tercer objectiu d'aquest estudi: fer prediccions precises sobre l'evolució futura del nombre mensual de trucades a la línia d'atenció contra la violència masclista, mitjançant l'ús de models de sèries temporals.

Per dur a terme aquesta anàlisi, hem fet servir dos enfocaments principals: models de sèries temporals i xarxes neuronals. Ambdós resultats, es compararan per avaluar-ne la precisió i fiabilitat en la predicció del nombre futur de trucades.

3.3.1 Modelització de Sèries Temporals

Utilitzem el model ARIMA (Autoregressive Integrated Moving Average) i la seva variant estacional, SARIMA (Seasonal ARIMA), per capturar l'estructura temporal de la sèrie de trucades.

3.3.1.1 Preparació prèvia al model

Per començar amb la metodologia Box-Jenkins, és a dir, proposar un model estocàstic capaç d'explicació i predicció, les sèries temporals utilitzades han de complir uns requisits (explicats a la Metodologia [2](#) per a més detalls).

En primer lloc, determinem si la variància de la sèrie és constant a partir d'un gràfic de caixes (boxplot) per a cada any i un gràfic de mitjana-variància.

Observant les Figures [3.9](#), es veu clarament com la variància no es mostra constant. Amb els rangs interquartílics del boxplot [3.9\(a\)](#), es pot veure com la variància abans del període del

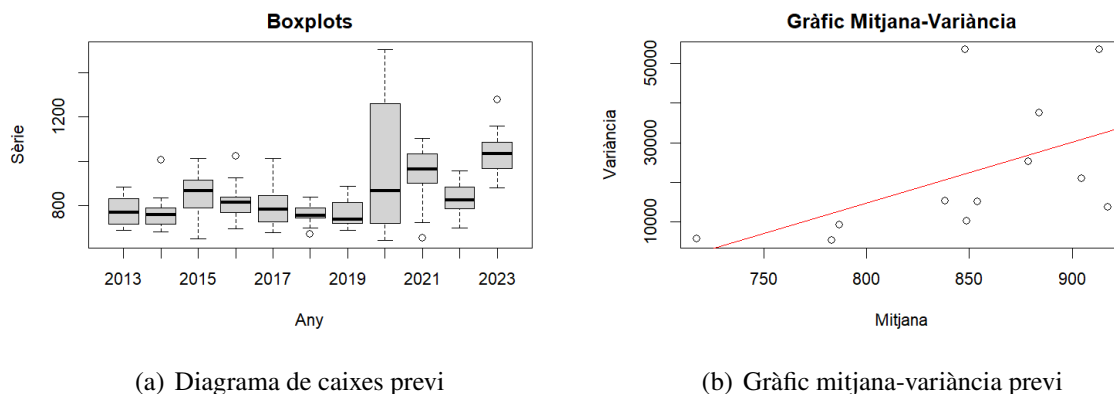


Figura 3.9: Gràfics per analitzar la variància de la sèrie

confinament es mantenia constant, mentre que posteriorment s'observa aquesta no constància. Això es contrasta amb el gràfic de mitjana-variància 3.9(b), on es veu que a major nivell de la sèrie, major variància.

Al no presentar-se una variància constant al llarg del temps, apliquem una transformació logarítmica a la nostra sèrie temporal.

Una vegada aplicada la transformació i, per tant, acabat l'estudi de la variància, analitzem l'estacionalitat de la sèrie temporal.

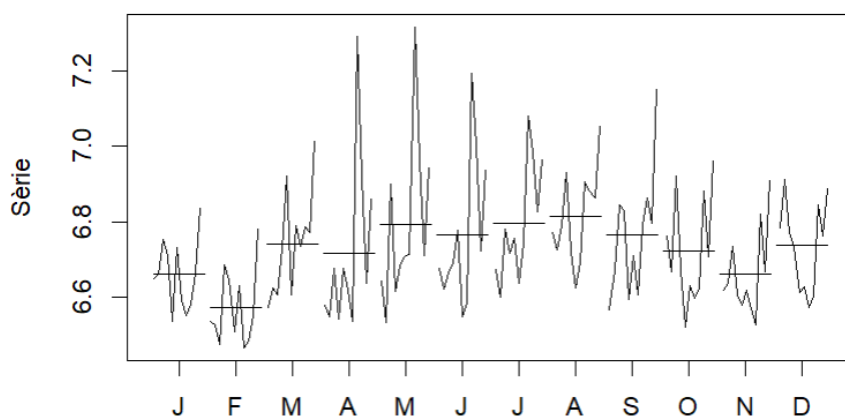


Figura 3.10: *Monthplot* de la sèrie de trucades

A partir del *monthplot* de la Figura 3.10, veiem que hi ha un lleuger patró estacional al llarg dels mesos. Eliminem aquesta estacionalitat mitjançant una diferenciació estacional ($D=1$) d'ordre $S=12$.

Una vegada aplicada aquesta diferenciació estacional, s'observa al gràfic 3.11 com la mitjana està centrada al zero, concretament 0.029, però no arriba a ser constant a tota la sèrie temporal.

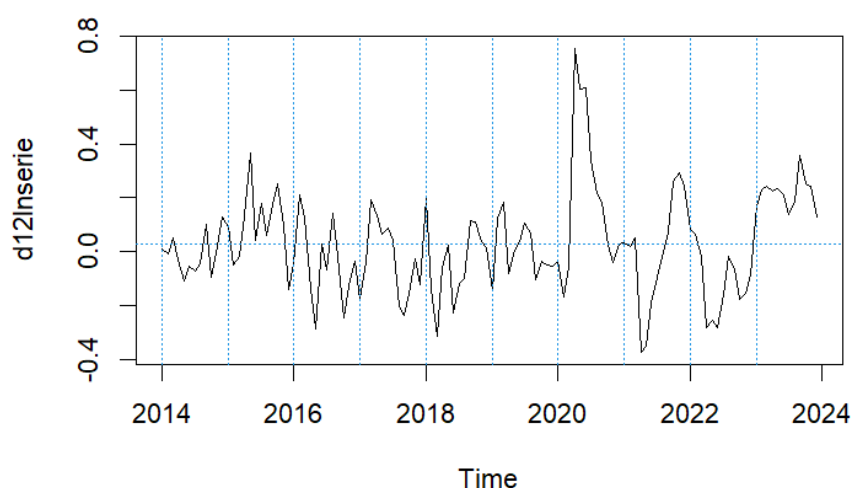


Figura 3.11: Gràfic temporal on s'observa la mitjana abans d'aplicar la diferenciació regular

Com a conseqüència, s'apliquen tantes diferenciacions regulars com es considerin necessàries.

Amb aquest supòsit, hem decidit aplicar només una diferenciació regular, donat que la variància augmentava al aplicar una diferenciació més, com es pot observar a la darrera columna de la taula 3.5.

lnserie	d12lnserie	d1d12lnserie	d1d1d12lnserie
0.027	0.035	0.026	0.059

Taula 3.5: Comparació de la variància de la sèrie al aplicar diferents diferenciacions

Finalment, s'analitza l'ACF per a determinar si la sèrie triada és realment estacionària.

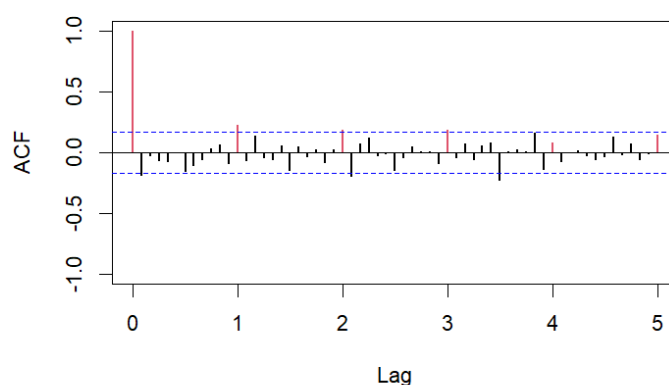


Figura 3.12: ACF de la sèrie

Com observem a la figura 3.12, el gràfic decau relativament ràpid cap a zero, el que ens indica que la sèrie transformada és, després de tot, estacionària.

3.3.1.2 Identificació i estimació del model

Una vegada hem corroborat que la sèrie transformada és finalment estacionària, passem al següent bloc: identificar els possibles models basant-nos en l'ACF i el PACF, seleccionant els components autoregressius (AR), integrats (I) i mitjana mòbil (MA) que millor s'ajustin. Se seguirà el principi de parsimònia. En altres paraules, es proposarà agafar un model senzill que s'adapti a la sèrie temporal estudiada sense la necessitat de complicar el model amb paràmetres addicionals.

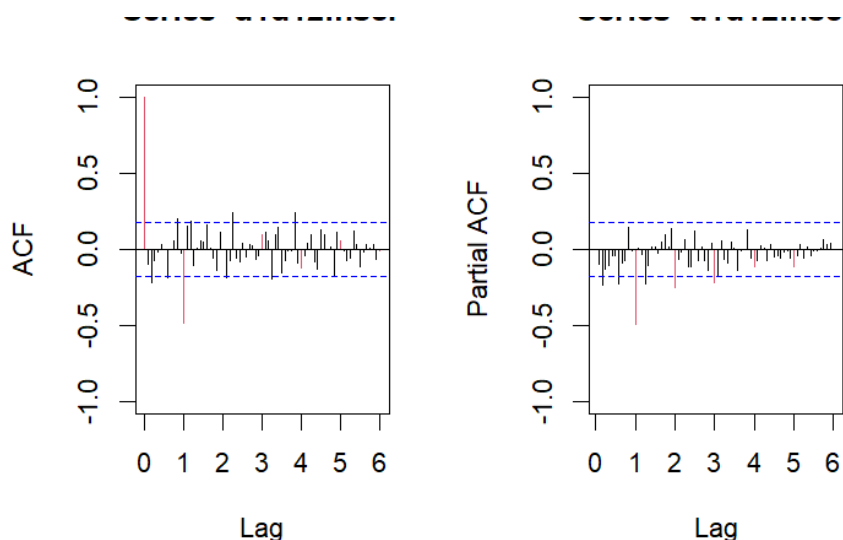


Figura 3.13: ACF i PACF de la sèrie

Amb l'ACF i PACF de la figura 3.13, per la part regular del model, identifiquem un AR(2) i un MA(2), degut a que tant al PACF com a l'ACF (respectivament) es veuen dos retards regulars fora de les bandes de confiança. En quant a la part estacional, es consideren els retards de color vermell.

Així, s'ha identificat un SAR(3) i un SMA(1), amb $S = 12$. Les diferenciacions a aplicar són: $d = 1$, $D = 1$ i $S = 12$.

Amb això, es proposen els models següents:

- **Model 1:** $ARMA(2, 2)SMA(1)_{12}$ per $W_t = \ln(X_t)$

$$(1 - \phi_1 B - \phi_2 B^2)W_t = (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta_1 B^{12})Z_t, \quad Z_t \sim N(0; \sigma_Z^2) \quad (3.1)$$

- **Model 2:** $AR(2)SAR(3)_{12}$ per $W_t = \ln(X_t)$

$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12} - \Phi_2 B^{24} - \Phi_3 B^{36})W_t = Z_t, \quad Z_t \sim N(0; \sigma_Z^2) \quad (3.2)$$

Per a cadascun dels dos models, estimarem els seus coeficients i es tornaran a ajustar eliminant els no significatius, seguint el principi de parsimònia.

	ar1	ar2	ma1	ma2	sma1	intercept
Coef	1.4157	-0.5685	-1.6987	0.6987	-0.8998	0.0006
s.e.	0.1848	0.1236	0.2060	0.2017	0.1736	0.0003
T-ratios	7.6628	-4.6006	-8.2454	3.4630	-5.1825	1.6342

Taula 3.6: Estimació del model 1

Per al Model 1, com s'observa la taula 3.6, l'*intercept* no ha resultat significatiu en quant que el seu *T-Ratio*, en valor absolut, és menor a 2. En conseqüència, es torna a ajustar el model SARIMA eliminant aquest coeficient, ara fent ús de la sèrie no-estacionària $\ln X_t$.

	ar1	ar2	ma1	ma2	sma1
Coef	-0.3034	0.6960	-0.0193	-0.9764	-0.9366
s.e.	0.0729	0.0728	0.3642	0.3546	0.2166
T-ratios	-4.1602	9.5543	-0.0530	-2.7537	-4.3231

Taula 3.7: Estimació del model 1.1

En aquest cas, a la taula 3.7, es veu com la part regular MA presenta un *T-Ratio* no significatiu. Tornem a ajustar el model però eliminant aquest paràmetre.

	ar1	ar2	ma1	sma1
Coef	0.7294	-0.0835	-0.9782	-0.8444
s.e.	0.0946	0.0934	0.0612	0.1238
T-ratios	7.7089	-0.8940	-15.9743	-6.8190

Taula 3.8: Estimació del model 1.2

Al tornar a ajustar el model, ara torna a sortir un coeficient de la part regular AR no significatiu (3.8). Es re-ajusta un cop més.

	ar1	ma1	sma1
Coef	0.6811	-0.9917	-0.8531
s.e.	0.0772	0.1310	0.1288
T-ratios	8.8200	-7.5726	-6.6248

Taula 3.9: Estimació del model 1 amb paràmetres significatius

A partir de la taula 3.9 veiem com finalment tots els coeficients són significatius. El model ajustat final es un $ARIMA(1, 1, 1)(0, 1, 1)_{12}$.

De la mateixa manera que hem estimat el model 1, estimem el model 2. Es presenta el model estimat final a la taula 3.10, reajustat mitjançant la sèrie $\ln X_t$, eliminant tota la part regular AR, el SAR(3) i l'*intercept*.

	sar1	sar2
Coef	-0.6485	-0.3344
s.e.	0.0861	0.0867
T-ratios	-7.5327	-3.8564

Taula 3.10: Estimació del model 2 amb paràmetres significatius

Una vegada tenim aquests dos models estimats, comparem els seus AIC i es decideix amb quin model farem l'anàlisi als següents apartats (vegeu la taula 3.11):

	df	AIC
Model 1	4	-157.3550
Model 2	3	-133.6732

Taula 3.11: Comparació dels models estimats a partir del criteri AIC

D'aquesta manera, el model que presenta un millor ajust és el *Model 1*, ja que és el que presenta un AIC més petit. Així, el model que finalment es validarà és el model següent:

Model 1: $ARIMA(1, 1, 1)(0, 1, 1)_{12}$

$$(1 - \phi_1 B)(1 - B)(1 - B^{12})X_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})Z_t \quad (3.3)$$

3.3.1.3 Validació del model

En aquesta secció, validarem el model SARIMA triat, amb l'objectiu de verificar si els supòsits subjacents del model es compleixen i si el model és capaç de capturar de manera adequada l'estructura de les dades temporals. Així, s'estudiaran les propietats dels termes de pertorbació.

Primerament, s'analitza que el terme de pertorbació del model segueixi una distribució normal i una variància constant.

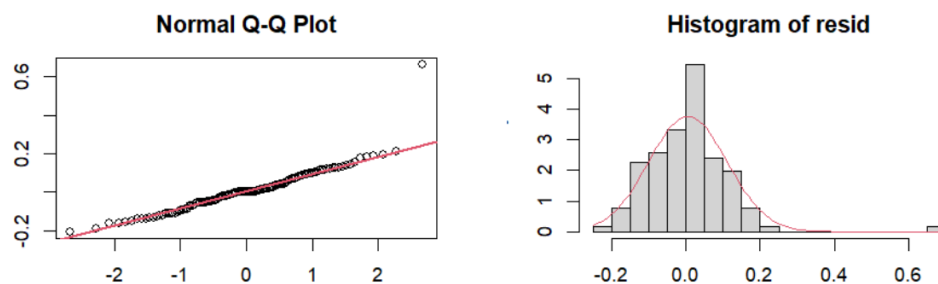


Figura 3.14: Validació de la normalitat

Amb la figura 3.14 podem dir que el model sembla ajustar-se bé amb el "Q-Q Plot", de manera que els residus es troben perfectament alineats a la recta diagonal. L'histograma ens confirma que els residus s'ajusten a una distribució Normal. No obstant això, visualitzem un petit outlier, que no acaba sent preocupant.

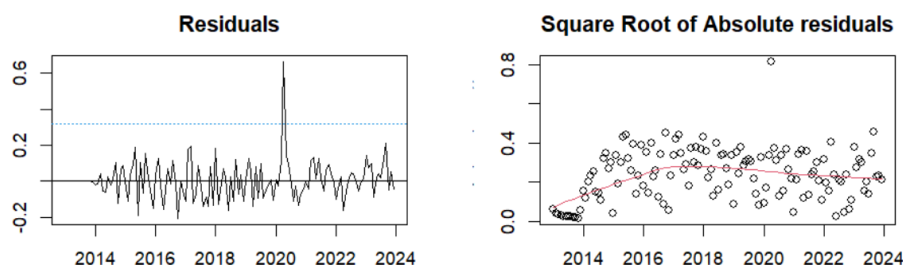


Figura 3.15: Validació de la variància

Per a l'anàlisi de la variància, ens fixem en els gràfics de residus (3.15). En el gràfic de l'esquerra veiem que la majoria dels residus es troben dins de les bandes de confiança al 95%, indicant que la variància és constant al llarg del temps. D'altra banda, en el gràfic de la dreta, es corrobora aquesta conclusió, ja que la línia vermella es presenta bastant anivellada. Finalment, comprovem a partir del contrast de Breusch-Pagan que, amb un p-valor major a 0.05, es verifica que els residus són homocedàstics (3.16).

```
Homoscedasticity Test
-----

studentized Breusch-Pagan test

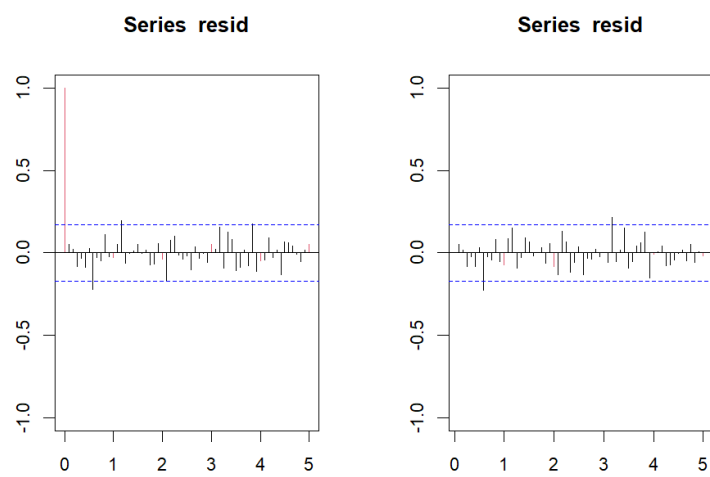
data: resid(model) ~ I(obs - resid(model))
BP = 0.00847, df = 1, p-value = 0.9267
```

Figura 3.16: Contrast de Breusch-Pagan per validar la variància

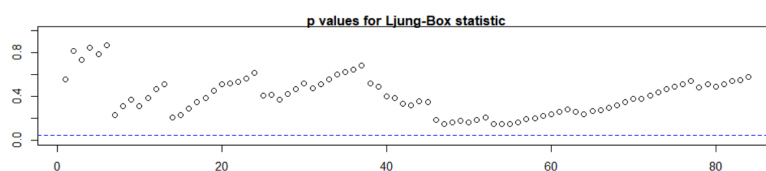
Seguint amb la validació del model, analitzem que el terme de pertorbació sigui independent entre els retards. En altres paraules, que els residus no estiguin auto-correlacionats.

Observant l'ACF i el PACF del terme de pertorbació de la figura 3.17(a), no sembla que els retards presentin autocorrelació, donat que presenten la mateixa distribució que el soroll blanc. D'altra banda, a partir del contrast Ljung-Box, a la figura 3.17(b), es veu com tots els residus es troben per sobre del nivell de significació de 0.05, concloent que els residus no estan auto-correlacionats i que, per tant, són independents entre ells.

Tant el test de Durbin-Watson, amb un p-valor de 0.279, com els resultats numèrics de Ljung-Box, presenten p-valors majors al nivell de significança de 0.05, verificant que els residus són independents.



(a) ACF i PACF del terme de pertorbació



(b) P-valors del Contrast Ljung-Box

Figura 3.17: Gràfics per validar la independència dels residus

Finalment, analitzem el ACF i el PACF al quadrat per a analitzar la presència de volatilitat.

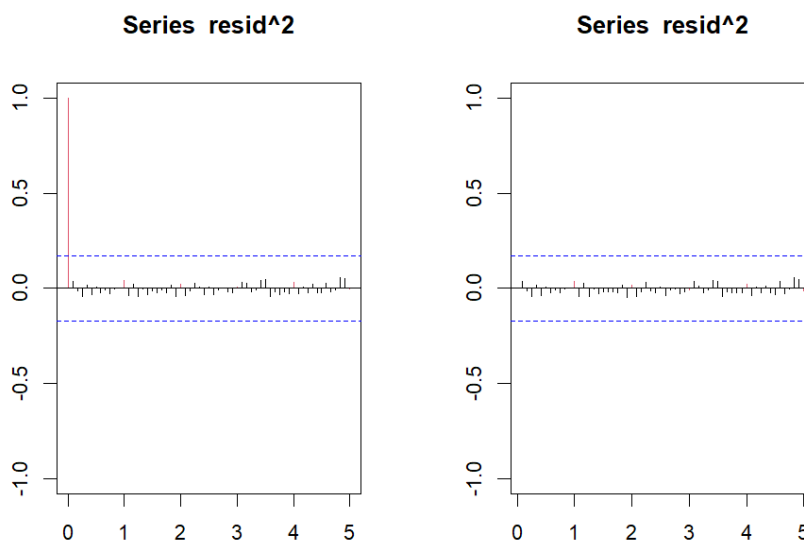


Figura 3.18: Comprovació de la volatilitat del model

Observant la figura 3.18, veiem que no hi ha cap retard fora de les bandes de confiança, indicant que els gràfics presenten soroll blanc i que, per tant, el model recull bé la volatilitat.

3.3.1.4 Predicció

Per acabar, comprovem l'estabilitat del model validat per complir amb l'objectiu: generar prediccions de l'evolució de les trucades a la línia d'atenció contra la violència masclista.

Ajustem el model tant per a les dades d'entrenament (tota la sèrie menys l'últim any), com per a la sèrie completa.

	ar1	ma1	sma1
Coef	0.6811	-0.9917	-0.8531
s.e.	0.0772	0.1310	0.1288
T-ratios	8.8200	-7.5726	-6.6248
Sigma2	0.01235		
Versemblança	82.68		
AIC	-157.35		

Taula 3.12: Anàlisi de l'estabilitat del model per a les dades d'entrenament

	ar1	ma1	sma1
Coef	0.6636	-1.0000	-0.8278
s.e.	0.0743	0.0484	0.1269
T-ratios	8.9315	-20.6762	-6.5252
Sigma2	0.01275		
Versemblança	72.13		
AIC	-136.26		

Taula 3.13: Anàlisi de l'estabilitat del model per a la sèrie sencera

Comparant les taules 3.12 i 3.13, veiem com els paràmetres presenten les mateixes magnituds en quant als coeficients estimats i els signes. Així, concloem que els paràmetres compleixen les condicions d'estabilitat indicant, per tant, que el model és estable al llarg del temps.

Ara, realitzem la predicció *Out-of-sample* per observar si el model prediu correctament les trucades registrades durant tot l'any 2023.

El gràfic de la figura 3.19 deixa veure com el model sembla capturar la tendència general i l'estacionalitat de les dades. Les mètriques d'error pel conjunt d'entrenament s'observen a la taula 3.14:

	RMSE	MAE	nRMSE	nMAE	RMSPE	MAPE
Conjunt d'entrenament SARIMA	178.117	156.515	0.174	0.153	0.508	0.147

Taula 3.14: Mètriques d'Error per al Conjunt d'Entrenament (SARIMA).

Finalment, fem les prediccions a llarg termini tenint en compte totes les dades registrades.

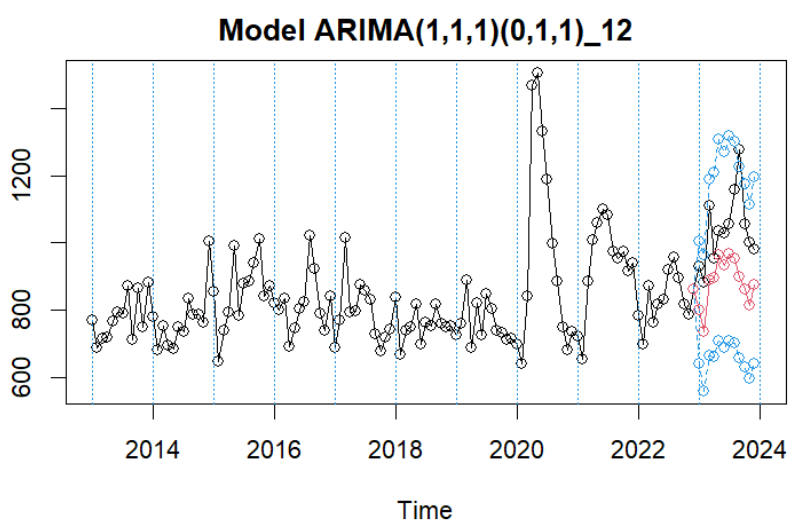


Figura 3.19: Predicció Out-of-Sample model SARIMA

Al gràfic de la figura 3.20, observem la predicció de les trucades per a l'any 2024, segons el model SARIMA estimat i validat.

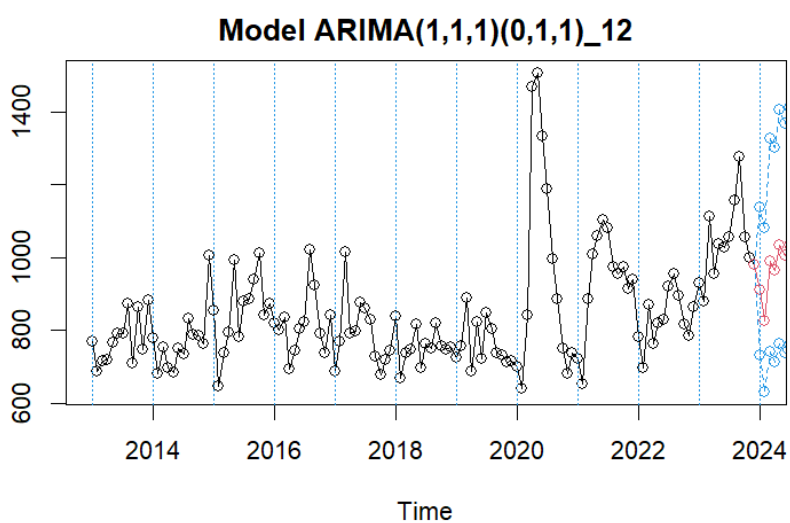


Figura 3.20: Predicció dels 12 mesos futurs per al model SARIMA

Les mètriques d'error pel conjunt de prova es troben a la taula següent (3.15):

	RMSE	MAE	nRMSE	nMAE	RMSPE	MAPE
Conjunt de prova SARIMA	280.913	81.092	0.27	0.079	0.287	0.083

Taula 3.15: Mètriques d'Error per al Conjunt de Prova (SARIMA).

3.3.2 Modelització d'una Xarxa Neuronal Artificial

Una vegada aplicada la modelització d'un SARIMA convencional a la nostra sèrie temporal, en aquesta secció desenvolupem un model de Xarxes Neuronals (ANN), aprofitant la seva capacitat per capturar relacions no lineals i patrons complexos a les dades.

Primerament, per a l'entrenament de les dades, hem utilitzat la sèrie temporal sencera, escalant les dades entre 0,9 vegades el valor mínim i 1,1 vegades el valor màxim de la sèrie, utilitzant $m = 5$ observacions anteriors ('lags') per a capturar les dependències temporals.

Seguidament, s'ha dividit la sèrie en conjunts d'entrenament i prova, amb els darrers 3 anys de dades destinades a la seva validació.

Després, per a l'entrenament del model, la xarxa neuronal es dissenya amb dues capes ocultes, cadascuna amb 6 neurones, i s'entrena amb una fórmula que relaciona les observacions futures amb les passades. El llindar ('threshold') utilitzat com a criteri de parada ha estat de 0.00125.

La figura 3.21 mostra l'estructura de la xarxa neuronal entrenada. El gràfic il·lustra les connexions entre les entrades, les capes ocultes i la sortida.

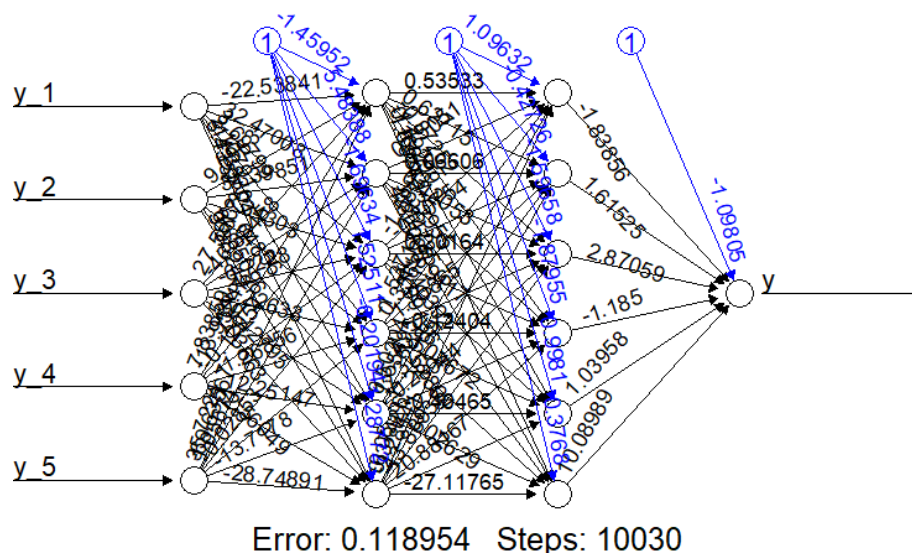


Figura 3.21: Arquitectura de la xarxa neuronal utilitzada per la predicció del número mensual del número de trucades

Les prediccions per al conjunt d'entrenament i prova es reescalen als valors originals i es comparen amb les dades reals.

El següent gràfic 3.22 mostra les prediccions del conjunt d'entrenament comparades amb la

sèrie real. Les línies negres i vermelles representen els valors observats i predits respectivament.

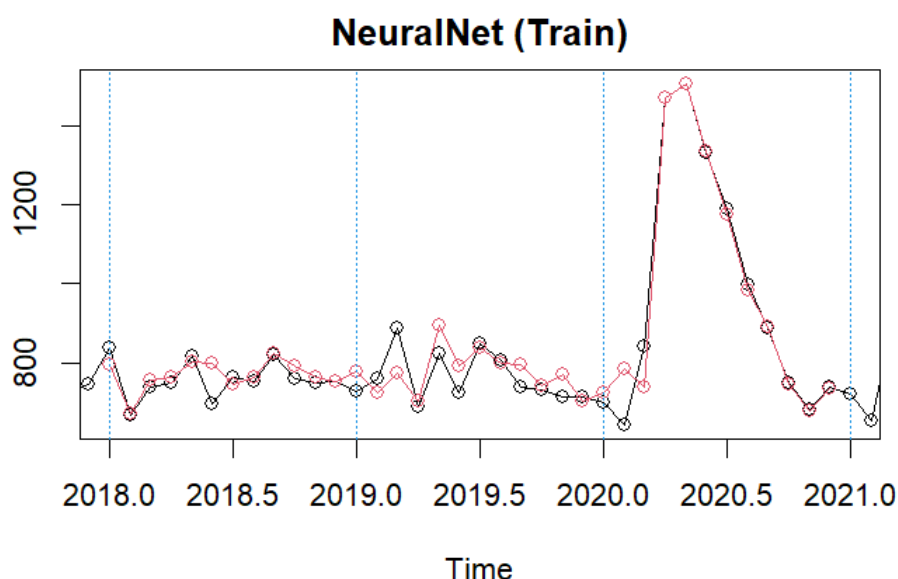


Figura 3.22: Predicció Out-of-Sample model ANN

Observem que el model ANN proporciona prediccions raonablement precises per a la major part de la sèrie temporal, de manera que capta adequadament les tendències generals, però no aconsegueix capturar del tot l'estacionalitat de les dades.

Les mètriques d'error per a les prediccions en el conjunt d'entrenament les trobem a la taula següent 3.16:

	RMSE	MAE	nRMSE	nMAE	RMSPE	MAPE
Conjunt d'entrenament ANN	46.999	30.331	0.056	0.0364	0.060	0.038

Taula 3.16: Mètriques d'Error per al Conjunt d'Entrenament (ANN).

Finalment, al gràfic 3.23 es mostren les prediccions del conjunt de prova comparades amb les dades reals.

Veiem com el model de xarxa neuronal capta adequadament les tendències generals de la sèrie temporal, exceptuant les prediccions en dades no observades, on sembla que no acaba de seguir la tendència creixent que es començava a observar a partir del 2022.

D'altra banda, l'estacionalitat és parcialment capturada pel model. Les prediccions mostren patrons recurrents similars als observats a les dades històriques, encara que amb algunes discrepàncies en l'amplitud i el temps exacte dels pics i valls.

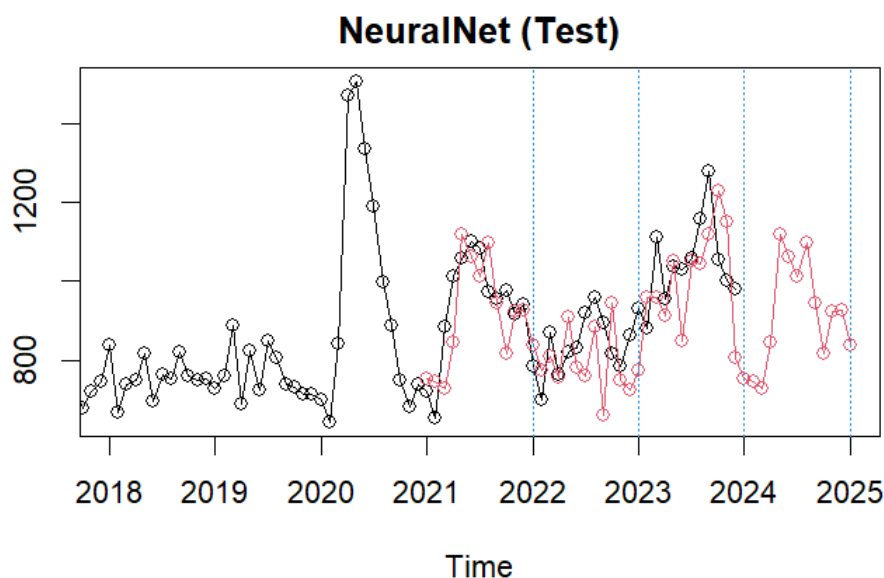


Figura 3.23: Predicció dels 12 mesos futurs per al model ANN

A continuació presentem les mètriques d'error calculades per a avaluar la precisió de les prediccions del conjunt de prova:

	RMSE	MAE	nRMSE	nMAE	RMSPE	MAPE
Conjunt de prova ANN	113.656	95.008	0.121	0.101	0.136	0.109

Taula 3.17: Mètriques d'Error per al Conjunt de Prova (ANN).

3.3.3 Millor model

Després d'aplicar els diferents models per a modelar la nostra sèrie temporal i fer les prediccions per a cada una d'elles, s'ha d'avaluar quin és el model òptim per a fer les prediccions sobre el número mensual de trucades a la línia d'atenció contra la violència masclista.

Calculem i resumim les diferents mètriques d'error per a determinar quin model és el més adequat. En altres paraules, comparem aquestes mètriques per a concloure quina metodologia explica i prediu millor el número de trucades.

La taula següent (3.18) resumeix aquestes mètriques per als models SARIMA i ANN tant en els conjunts de dades d'entrenament com de proves:

	RMSE	MAE	nRMSE	nMAE	RMSPE	MAPE
Conjunt d'entrenament SARIMA	178.117	156.515	0.174	0.153	0.508	0.147
Conjunt d'entrenament ANN	46.999	30.331	0.056	0.0364	0.060	0.038
Conjunt de prova SARIMA	280.913	81.092	0.27	0.079	0.287	0.083
Conjunt de prova ANN	113.656	95.008	0.121	0.101	0.136	0.109

Taula 3.18: Mètriques d'Error per comparar els diferents models

D'una banda, el model ANN supera significativament el model SARIMA en el conjunt d'entrenament en totes les mètriques d'error, cosa que indica un millor ajust i prediccions més precises durant l'entrenament.

D'altra banda, al conjunt de proves, el model ANN encara mostra un RMSE, RMSPE i nRMSE inferiors en comparació amb el model SARIMA, la qual cosa indica una millor precisió predictiva global. Tanmateix, el model SARIMA presenta un millor rendiment en termes de MAE, nMAE, i MAPE.

Tenint en compte tant el rendiment del conjunt d'entrenament com de prova, el model ANN generalment demostra una precisió predictiva superior amb valors RMSE i nRMSE més baixos en ambdós conjunts de dades. Encara que hi ha bastant marge d'error donat que la precisió d'aquestes mètriques no és gaire satisfactòria, concloem que el model ANN és més precís per predir el nombre de trucades mensuals a la línia d'atenció contra la violència masclista de tota la regió catalana.

Capítol 4

Discussió i Conclusions

4.1 Validació externa del Model de subnotificació

L'objectiu principal d'aquest primer estudi ha consistit en comprovar si l'estimació de l'evolució dels registres produït pel model extern de subnotificació és similar a l'evolució del número de trucades a la Línia d'Atenció contra la Violència Masclista. Considerem que aquest servei proporciona una quantificació més acurada del fenomen, que no pas els registres clínics o policials. Amb això, els resultats d'aquest estudi revelen la capacitat del model per a capturar la incidència real de la violència de gènere.

A continuació, es discutirà la validació del model i es relacionarà l'estructura canviant de totes les dades observades amb diversos fenòmens externs. Aquests fenòmens proporcionen un context per interpretar els augments i disminucions dels registres, cosa que resulta fonamental per entendre la dinàmica subjacent de la violència de gènere i el seu report al llarg del temps.

4.1.1 Evolució del número de trucades

D'una banda, els pics significatius que s'observen el 2016 i el 2017, amb més de 300 trucades en alguns mesos, poden estar estretament relacionats amb esdeveniments particulars que van augmentar la consciència pública i la disposició de les víctimes a demanar ajut.

D'altra banda, l'increment del número de trucades coincidint amb l'inici de la pandèmia de la COVID-19 i els seus confinaments el 2020, verifica les observacions de [Lorente-Acosta \(2020\)](#), que assenyalava que qualsevol circumstància de confinament actua com a factor de risc. Llavors, aquest increment és atribuïble a l'augment de l'estrès i les tensions a la llar, així com a la dificultat per escapar-se de situacions de violència.

Posteriorment, el gran descens en el nombre de trucades observat a partir del 2022, podria estar explicat tant per la relaxació de les mesures de confinament, que podrien haver reduït les

tensions a les llars, com per millores en la gestió i la prevenció de casos de violència de gènere.

Aquestes observacions principalment suggereixen que la situació excepcional del confinament va exacerbar les condicions que condueixen a la violència de gènere, fent que més víctimes busquessin ajuda a través de les línies telefòniques de suport. Més endavant veurem com numèricament el confinament va impactar notablement en aquest fenomen.

4.1.2 Evolució dels registres del model de subnotificació

A partir de l'observació dels casos registrats de violència de gènere, pensem com l'augment significatiu de denúncies a partir del 2019 està estretament relacionat amb esdeveniments específics que van augmentar la consciència pública i la disposició de les víctimes a demanar ajuda.

Concretament, durant el 2019, es van dur a terme nombroses campanyes de sensibilització a nivell estatal i local. '*Respeto los límites sí o sí*' i altres iniciatives de la Delegació del Govern contra la Violència de Gènere, que han ajudat a incrementar la visibilitat del problema de la violència de gènere i fomentar a més víctimes a buscar ajuda.

A més, el Pacte d'Estat contra la Violència de Gènere aprovat el 2017 ([Ministerio de Igualdad, 2017](#)), que inclou diverses mesures de protecció i suport a les víctimes, també sembla que ha contribuït a aquest augment en les denúncies.

D'altra banda, respecte als casos estimats pel model de subnotificació, la seva tendència estable al llarg del temps amb una mitjana molt major als casos registrats, ens suggereix la gran diferència que hi ha entre la incidència real de la violència de gènere amb els registres oficials.

Adicionalment, cap al final del període analitzat, s'observa una convergència entre les sèries de casos registrats i estimats, suggerint una possible reducció a la bretxa de subnotificació. No obstant això, aquesta convergència no és completa, cosa que indica que encara hi ha una quantitat considerable de casos que no es registren i que el model estima. Això subratlla la importància de seguir millorant els sistemes de registre i les polítiques de suport a les víctimes per reflectir amb més precisió la magnitud del problema i garantir que totes les víctimes rebin l'ajuda necessària.

4.1.3 Validació

Finalment, tenim en compte que les crides a la línia d'atenció reflecteixen una aproximació més raonable al nombre real d'incidents de violència de gènere que no pas els registres policials. Així, a partir de l'anàlisi comparativa, els resultats indiquen que, de mitjana, els casos estimats pel model coincideixen amb el nombre de trucades, suggerint que el model capta la tendència general dels casos reals de violència de gènere.

Tot i això, s'observen algunes limitacions. Als inicis del 2020, coincidint amb el confinament a causa de la pandèmia de COVID-19, es registra un pic en el nombre de trucades que el model no aconsegueix captar. Aquesta incapacitat per adaptar-se a canvis bruscos a la realitat social reflecteix una limitació significativa del model.

D'altra banda, encara que el model aconsegueix captar la tendència general, el fet que els casos estimats presentin una major estabilitat i una reducció més moderada en comparació amb les trucades, ens suggereix que el model no reflecteix completament les fluctuacions ràpides observades a les trucades. En altres paraules, concloem que el model no captura correctament la variabilitat dels casos reals.

Tal com s'ha comentat a la metodologia d'aquest estudi, la distribució que s'utilitza per a les estimacions és la Poisson, que, encara que adequada per a certs tipus d'anàlisi, pot provocar sobredispersió o variància extra-Poisson en el context de fenòmens recurrents ([Navarro et al., 2001](#)). Això condueix a la subestimació dels errors estàndards dels coeficients, podent derivar en la significació estadística de factors que realment no estiguin associats amb el fenomen.

L'estudi de [Navarro et al. \(2001\)](#), indica que per abordar aquestes limitacions, es proposa l'ús de la distribució Binomial Negativa com una alternativa vàlida per analitzar aquest tipus de fenòmens. Aquesta distribució és capaç de capturar part de la variància no identificada per la regressió de Poisson, encara que presenta alguns inconvenients, com ara la impossibilitat de controlar variables que fluctuen en el temps i possibles intervals de temps amb diferents nivells de risc. Això no obstant, la Binomial Negativa permet estimar, almenys en part, la variabilitat que la Poisson no pot identificar, oferint una alternativa per capturar la variabilitat dels casos reals de violència de gènere.

En definitiva, concloem que encara que el model reflecteix bé la tendència general, és incapaç d'adaptar-se a canvis bruscos i de capturar completament la variabilitat i volatilitat observades al nombre de trucades. Com a conseqüència, suggerim l'ús de la Binomial Negativa per millorar l'estimació i adaptar-se millor a canvis bruscs.

Tot i aquestes limitacions, el model es considera adequat per reflectir la realitat dels casos no registrats de violència de gènere a l'àrea metropolitana nord de Barcelona.

4.2 Quantificació de l'impacte de la COVID-19

El segon objectiu al qual ens enfrontàvem, era analitzar i quantificar l'impacte que va causar el confinament de la COVID-19 al número de trucades i, per tant, a la violència de gènere.

Principalment, l'anàlisi de l'impacte de la pandèmia revela canvis significatius en la dinàmica

d'aquestes dades, aportant una comprensió més profunda de com les crisis poden afectar la demanda de serveis de suport per a víctimes de violència de gènere.

D'una banda, el test de Dickey-Fuller Augmentat confirma que la sèrie temporal del nombre de trucades es va tornar no estacionària després del confinament, reflectint un canvi significatiu en la dinàmica de les trucades. Aquesta troballa suggereix que la pandèmia i el confinament van alterar l'estructura temporal de la sèrie i van provocar canvis en els patrons de comportament observats a les dades.

D'altra banda, la metodologia de CausalImpact va permetre quantificar l'impacte específic de la COVID-19 al nombre de trucades. Els resultats mostren un increment significatiu, amb 6,539 trucades addicionals durant el període analitzat, que representen un augment mitjà del 19% en comparació amb les prediccions basades en la tendència històrica prèvia a la pandèmia. Aquest increment no sols és estadísticament significatiu, sinó també socialment rellevant, destacant la creixent necessitat de suport i recursos per a les víctimes de violència masclista en situacions de crisi.

Aquests resultats tenen una clara implicació. Subratllen la necessitat d'enfortir els serveis d'atenció contra la violència masclista en temps de crisi i postcrisi. La COVID-19 no només va crear un escenari crític durant el confinament, sinó que arrel d'aquest fenomen les trucades s'han intensificat. L'augment sostingut en el nombre de trucades durant la pandèmia indica que les crisis poden exacerbar la violència de gènere i augmentar la demanda de suport, cosa que corrobora les paraules de [Lorente-Acosta et al. \(2022\)](#). Finalment, l'augment del registre de trucades postconfinament, ens podria estar indicant que les campanyes i les polítiques aplicades durant l'any 2019 han començat a donar efecte, acompanyant a les dones a buscar ajuda.

En resum, la pandèmia de COVID-19 va tenir un impacte profund i sostingut a la dinàmica de les crides a la línia d'atenció contra la violència masclista. Aquests resultats destaquen la importància d'enfortir els serveis de suport en temps de crisi i d'adaptar-se als canvis a la demanda per proporcionar el suport necessari a les víctimes de violència de gènere.

4.3 Predicció i comparació de models predictius

Finalment, la predicció del nombre de crides a la línia d'atenció contra la violència masclista utilitzant models de predicció ARIMA i ANN revela diferències significatives en la precisió i el rendiment dels dos enfocaments.

Tal com hem vist a l'apartat de resultats, considerant el rendiment tant al conjunt d'entrenament com al de prova, el model ANN generalment demostra una precisió predictiva superior.

Això indica que el model ANN és més precís per predir el nombre de trucades mensuals a la línia d'atenció contra la violència masclista a la regió catalana. La capacitat de l'ANN per controlar no-linealitats i captar patrons complexos en les dades ho fa una eina poderosa per a la predicció en sèries temporals complexes com la que s'analitza en aquest estudi.

No obstant això, és important assenyalar que els models ANN tenen certes limitacions en comparació dels models ARIMA. Els procediments de Machine Learning, com ara les xarxes neuronals, no es recolzen en un model estadístic tradicional, ja que es basen en iteracions i no proporcionen intervals de confiança. Aquesta manca d'intervals de confiança pot ser un desavantatge significatiu en contextos en què és crucial tenir una mesura de la incertesa associada a les prediccions.

D'altra banda, un dels avantatges significatius dels models ANN és la seva capacitat per adaptar-se millor a canvis estructurals a la sèrie temporal. Els models convencionals com els SARIMA sovint tenen dificultats per capturar aquests canvis abruptes, mentre que les xarxes neuronals poden ajustar-se més fàcilment a noves dinàmiques a les dades. Aquesta adaptabilitat fa que els models ANN siguin particularment útils en situacions en què s'anticipen canvis als patrons subjacents, com els observats durant la pandèmia de COVID-19.

Els gràfics de predicció proporcionen una clara visualització de la capacitat predictiva dels dos models. El model SARIMA mostra prediccions que, encara que gaire precises, té problemes a capturar les fluctuacions i pics observats a les dades reals. En canvi, el model ANN mostra una capacitat superior per seguir les fluctuacions i capturar millor la variabilitat a les dades, com s'observa en els períodes posteriors al confinament. No obstant això, ambdós models no acaben de capturar la tendència alcista de després del confinament per la COVID-19, especialment el model SARIMA.

En resum, el model ANN ofereix una millor precisió predictiva general en comparació del model SARIMA, especialment en mètriques clau com RMSE i nRMSE. Això suggereix que les xarxes neuronals artificials són una eina més efectiva per a la predicció del nombre de crides a la línia d'atenció contra la violència masclista, donat el seu rendiment superior en capturar patrons complexos i no lineals en les dades. Tot i això, el model SARIMA encara té mèrits en termes de minimització d'errors absoluts, cosa que podria ser útil en contextos específics. L'elecció del model pot dependre del balanç entre la necessitat de precisió global i la minimització d'errors absoluts, però els resultats indiquen que, en general, el model ANN és més adequat per a aquest propòsit. A més, és crucial considerar les limitacions inherents de cada enfocament, especialment pel que fa a la provisió d'intervals de confiança i la capacitat per manejar canvis estructurals a la sèrie temporal.

4.4 Conclusions

Finalitzant aquest extens projecte d'investigació, podem dir que s'han complert tots els objectius proposats a l'inici de l'estudi. D'una banda, hem validat i comprovat la bona precisió del model de subnotificació, en quant que l'evolució de les seves estimacions segueixen la mateixa tendència que els registres a la Línia d'Atenció Contra la Violència Masclista. D'altra banda, hem pogut quantificar tant gràficament com numèricament l'impacte que ha tingut la COVID-19 en el número de trucades i, per tant, quantificar el seu impacte al fenomen de la violència masclista. Finalment, podem dir que hem estat capaços de contribuir en la quantificació d'aquest fenomen, proporcionant prediccions de les trucades a la Línia d'Atenció contra la Violència Masclista a tota la regió catalana.

Recordant la hipòtesi formulada a l'inici del treball, que sosté que la capacitat de les trucades per proporcionar una imatge més precisa del fenomen supera la dels registres policials o clínics, s'ha comprovat que el model de subnotificació validat és eficaç en capturar la tendència general de la violència de gènere a la Regió Sanitària Metropolitana Nord de Barcelona. Considerem llavors que el model reflecteix la incidència real amb més precisió que els registres policials.

Tot i això, el model presenta limitacions per adaptar-se a canvis bruscos en la realitat social i incapacitat en capturar la variabilitat del fenomen. Aquests resultats ens donen peu a proposar una millora per a aquest model i proposar la implementació de la distribució Binomial Negativa com a alternativa. Aquest canvi milloraria l'estimació del model de subnotificació aconseguint una major captura de la variabilitat que la distribució de Poisson no acaba d'identificar.

Addicionalment, l'augment significatiu de les trucades i de les denúncies observades post-confinament, ens han fet reflexionar sobre la importància de les campanyes de sensibilització i les polítiques públiques, crucials per a augmentar la consciència i la disposició de les víctimes a cercar ajuda.

La pandèmia, en particular, que va tenir un impacte profund i perllongat en la dinàmica de les trucades durant i postconfinament, valida la hipòtesi que situacions de crisi poden exacerbar la violència de gènere. A més, ens destaca la necessitat de reforçar els serveis de suport durant les crisis. Es conclou que la implementació d'aquestes iniciatives demostra la importància de mantenir i reforçar les mesures de sensibilització i suport a les víctimes per continuar avançant en la lluita contra la violència de gènere.

Finalment, els resultats subjacents al darrer objectiu, ens han revelat que per a generar prediccions precises, l'elecció del model predictiu ha de considerar el balanç entre precisió global i minimització d'errors absoluts, avaluant les limitacions inherents de cada enfocament.

És cert que hem observat com principalment el model ANN demostra una precisió predictiva superior en comparació amb el model SARIMA, especialment en la captura de patrons complexos i no lineals a les sèries temporals. No obstant això, els models ANN com que no tenen intervals de confiança, ens impliquen una limitació en contextos on és crucial mesurar la incertesa associada a les prediccions. Per tant, encara que el model ANN sembli capturar millor la variabilitat del fenomen, com cap dels dos models aconsegueixen capturar la tendència alcista postconfinament, suggerim la necessitat de continuar explorant i trobar la metodologia predictiva que millor s'adapti a les nostres dades.

Encara que el model de subnotificació i els models predictius utilitzats en aquest estudi han mostrat ser eines útils per comprendre i preveure la incidència de la violència de gènere, és essencial continuar millorant aquests mètodes i adaptar-se a les dinàmiques canviants de la societat. Les crisis, com la pandèmia de COVID-19, ressalten la importància de la flexibilitat i capacitat d'adaptació d'aquests models per proporcionar un suport adequat i efectiu a les víctimes de violència de gènere.

Una de les principals limitacions trobades en aquest estudi ha estat que les dades utilitzades per a la validació del model extern pertanyen únicament a l'Àrea Metropolitana Nord de Barcelona. Seria d'interès disposar de dades per a tota la regió catalana, ja que això permetria una validació més robusta, una aplicació més àmplia del model de subnotificació i contribuiria a una comprensió més exhaustiva del fenomen de la violència de gènere a Catalunya.

En conclusió, aquesta investigació és un complement valuós als estudis sobre Violència de Gènere, perquè presenta diversos models per predir els casos no registrats, així com el model de subnotificació i les prediccions de les trucades a la Línia d'Atenció contra la Violència Masclista a Catalunya. Alhora, les troballes d'aquest estudi ressalten la importància de les campanyes de sensibilització i les polítiques públiques a augmentar la visibilitat i la denúncia de la violència de gènere. Factors polítics, activistes i professionals poden utilitzar aquests coneixements per unir forces, lluitar contra la violència masclista i crear una societat més segura i justa per a tothom.

Personalment, no només ha estat un procés enriquidor pel fet d'investigar sobre un dels fenòmens més invisibilitats de la nostra societat, com és la Violència de Gènere, sinó que també s'han millorat els meus coneixements en estadística i se n'han ampliat uns de nous amb la implementació de les Xarxes Neuronals Artificials. Addicionalment, l'informe en qüestió s'ha redactat utilitzant el llenguatge *LaTeX*, fet que m'ha permès adquirir noves habilitats en l'ús d'eines avançades per a la presentació de treballs acadèmics.

Encara que sempre basat en la sòlida formació teòrica adquirida durant el grau d'estadística, la realització d'aquest treball ha estat un recorregut enriquidor, d'aprenentatge continu i ple de novetats.

Bibliografia

- Alice, M. (2015). Fitting a neural network in r; neuralnet package. <https://www.r-bloggers.com/2015/09/fitting-a-neural-network-in-r-neuralnet-package/>. Accessed: 2024-06-26.
- Ara (2020). Més de la meitat de les dones de l'estat han patit algun tipus de violència masclista al llarg de la seva vida. https://www.ara.cat/societat/mes-meitat-dones-violencia-masclista_1_1018188.html. Accessed: 2024-06-26.
- Brockwell, P. and Davis, R. (2002). *An Introduction to Time Series and Forecasting*, volume 39.
- Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247 – 274.
- Casas González, M. (2020). Forecasting de series temporales con redes neuronales. <https://www.linkedin.com/pulse/forecasting-de-series-temporales-con-redes-neuronales-casas-gonzalez/>. Accessed: 2024-06-26.
- Charte, F. (2016). Redes neuronales con r. <https://fcharte.com/tutoriales/20160203-R-RedesNeuronales/>. Accessed: 2024-05-15.
- Coneo, Y. A. S. (2019). Redes neuronales con r. https://yuasaavedraco.github.io/Docs/Redes_Neuronales_con_R.html. Accessed: 2024-05-15.
- Cuéllar, J. L. G. (2021). Analysis of time series using r. <https://rpubs.com/JessicaCuellar/843574>. Accessed: 2024-05-22.
- García, P. (2021). Una de cada tres dones al món pateix violència masclista, segons l'onu. *À Punt*. Accessed: 2024-04-01.
- Gardner, G., Harvey, A. C., and Phillips, G. D. (1980). Algorithm as 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of kalman filtering. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):311–322.

Generalitat de Catalunya (2021). Enquesta sobre les violències masclistes a catalunya 2021. https://interior.gencat.cat/ca/el_departament/publicacions/seguretat/estudis-i-enquestes/enquesta_de_violencia_masclista/enquesta-sobre-les-violencias-masclistes-a-catalunya-2021/index.html. Accessed: 2024-06-26.

Generalitat de Catalunya (2024). Igualtat i feminismes posa en marxa el nou servei d'intervenció en crisi que garanteix atenció psicològica urgent i presencial especialitzada en violències masclistes. <https://govern.cat/salaprensa/notes-premsa/580662/igualtat-i-feminismes-posa-en-marxa-el-nou-servei-d-intervencio-en-cri-si-que-gar>. Accessed: 2024-04-01.

Generalitat de Catalunya (n.d.a). Línia d'atenció contra la violència masclista 900 900 120. <https://igualtat.gencat.cat/ca/ambits-dactuacio/violencias-masclistes/linia-datencio-contra-la-violencia-masclista-900-900-120->. Accessed: 2024-04-01.

Generalitat de Catalunya (n.d.b). Línia d'atenció contra la violència masclista 900 900 120. https://analisi.transparenciacatalunya.cat/Societat-benestar/Linia-d-atenci-contra-la-violencia-masclista-900-9/q2sg-894k/about_data. Accessed: 2024-04-21.

Jim (2022). Augmented dickey-fuller test in r. <https://www.r-bloggers.com/2022/06/augmented-dickey-fuller-test-in-r/>. Accessed: 2024-05-22.

Lorente-Acosta, M. (2020). Violencia de género en tiempos de pandemia y confinamiento. *Revista Española de Medicina Legal*, 46(3):139–145.

Lorente-Acosta, M., Lorente-Martínez, M., and Lorente-Martínez, M. (2022). Impacto de la pandemia por covid-19 y el confinamiento en los homicidios por violencia de género en España. *Revista Española de Medicina Legal*, 48(1):36–43.

Ministerio de Igualdad (2017). Plan estratégico para la violencia de género. <https://violenciagenero.igualdad.gob.es/wp-content/uploads/FolletoPEVGcastweb.pdf>. Accessed: 2024-06-26.

Ministerio de Igualdad (2020). Macroencuesta de violencia contra la mujer 2019. <https://violenciagenero.igualdad.gob.es/macroencuesta2015/macroencuesta2019/>. Accessed: 2024-04-21.

Ministerio de Igualdad (n.d.a). Delegación del gobierno contra la violencia de género. <https://violenciagenero.igualdad.gob.es/>. Accessed: 2024-04-21.

- Ministerio de Igualdad (n.d.b). Teléfono 016. <https://violenciagenero.igualdad.gob.es/informacion-3/recursos/telefono016/>. Accessed: 2024-04-21.
- Moriña, D., Millán, I., Fernández-Fontelo, A., Puig, P., Toran, P., Gómez-Maldonado, M., and Falguera, G. (2024). Exploring what lies beneath the tip of the gender-based violence iceberg. *medRxiv*, pages 2024–02.
- Navarro, A., Utzet, F., Puig, P., Caminal, J., and Martín, M. (2001). La distribución binomial negativa frente a la de poisson en el análisis de fenómenos recurrentes. *Gaceta Sanitaria*, 15(5):447–452.
- RDocumentation (n.d.). `adf.test`: Augmented dickey-fuller test. <https://www.rdocumentation.org/packages/aTSA/versions/3.1.2.1/topics/adf.test>. Accessed: 2024-05-22.
- Rodriguez-Jimenez, R., Fares-Otero, N. E., and García-Fernández, L. (2023). Gender-based violence during covid-19 outbreak in spain. *Psychological Medicine*, 53(1):299–300.
- Salguero, R. A. F. (2021). Series temporales avanzadas: Aplicación de redes neuronales para el pronóstico de series de tiempo. Master's thesis, Universidad de Granada. Accessed: 2024-06-12.
- World Health Organization (2007). Global plan of action, to strengthen the role of the health system within a national multisectoral response, to address interpersonal violence, in particular against women and girls, and against children.
- World Health Organization (2021). Violence against women prevalence estimates, 2018 – executive summary. <https://www.who.int/publications/i/item/9789240026681>. Accessed: 2024-04-01.

Apèndix A

Codi R

```
1
2 # Llibreries utilitzades
3 library(dplyr)
4 library(lubridate)
5 library(stringr)
6 library(ggplot2)
7 library(CausalImpact)
8 library(tseries)
9 library(tidyverse)
10 library(neuralnet)
11 library(knitr)
12 library(kableExtra)
13 library(urca)
14
15
16 # Dades
17 ### LINEA D'ATENCIÓ CONTRA LA VIOLENCIA MASCLISTA A TOTA LA REGIÓ DE
    CATALUNYA
18 dades <- read.csv("L_nia_d_atenci__contra_la_viol_ncia_masclista_
19 900_900_120_20240225.csv", header=TRUE, sep=",")
20
21 ## Lectura de les dades
22 dades<-dades[dades$Sexe=="Dona",]
23 dades$Edat[dades$Edat=="Menors de 18 anys"]<-"Menys de 18 anys"
24
25 ## Agafem casos de VGS i VGF
26 dades <- dades[dades$Motiu.Trucada == "Violencia masclista" |
    dades$Violencia.fisica=="Si" | dades$Violencia.sexual=="Si",]
27
28 ## Dades mensuals
29 dades$Data <- as.Date(paste0(substr(dades$Data, 1, 6), dades$Any),
    format="%d/%m/%Y")
30 dades <- dades[order(dades$Data), ]
```

```

31 dades$Month <- month(dades$Data)
32 dadesMens <- dades %>% group_by(Any, Month) %>% summarise(Trucades=n())
33
34 ## Serie temporal
35 dades_ts <- ts(dadesMens$Trucades, freq=12, start=c(2013, 1))
36 plot(dades_ts)
37
38 # 1) VALIDACIO EXTERNA DEL MODEL DE SUBNOTIFICACIO
39
40 ### Plot de trucades + casos estimats pel model
41 ## Area Metropolitana Nord de Barcelona
42
43 ## Lectura de les dades del model a validar
44 casos_estimats_model <- read.csv("CasosVG_ABS_MN.csv", sep="")
45
46 ## Lectura de les dades del numero de trucades
47 trucades_ABS_MN <- read.csv("dades_trucades_Area_Metropolitana_Nord.csv")
48
49 ## Convertir la columna 'Data' a format de data i agregar la columna
   'Month'
50 trucades_ABS_MN$Data <- as.Date(trucades_ABS_MN$Data, format="%d/%m/%Y")
51 trucades_ABS_MN$Month <- month(trucades_ABS_MN$Data)
52
53 ## Agrupar les dades per any ('Any') i mes ('Month') i contar el numero
   de trucades
54 dadesMonth_ABS_MN <- trucades_ABS_MN %>% group_by(Any, Month) %>%
   summarise(Trucades=n())
55 dadesMonth_ABS_MN$date <- as.Date(paste0(dadesMonth_ABS_MN$Any, "-",
   dadesMonth_ABS_MN$Month, "-01"))
56
57 ## Grafiquem el numero de trucades
58 ggplot(dadesMonth_ABS_MN, aes(x = date, y = Trucades)) +
59 geom_line() +
60 labs(title = "Numero de Trucades per Mes a l'Area Metropolitana Nord de
   Barcelona",
61       x = "Data",
62       y = "Numero de Trucades") +
63 theme_minimal() +
64 theme(plot.title = element_text(hjust = 0.5), axis.text.x =
   element_text(angle = 60, hjust = 1))
65
66 ##Estadistics basics
67 summary(dadesMonth_ABS_MN$Trucades)
68 sd(dadesMonth_ABS_MN$Trucades)
69
70 ### Monthly aggregation estimated cases
71 reg <- casos_estimats_model[casos_estimats_model$Series=="Registered", ]

```

```

72 est <- casos_estimats_model[casos_estimats_model$Series=="Estimated", ]
73 reg$date <- as.Date(paste(reg$Year, reg$Week, 1, sep="-"),
74   format="%Y-%U-%u")
75 est$date <- reg$date
76 reg <- reg[!is.na(reg$date), ]
77 est <- est[!is.na(est$date), ]
78 casos_estimats_model <- rbind(reg, est)
79 casos_estimats_model$Month <- month(casos_estimats_model$date)
80 casos_estimats_model$Year <- year(casos_estimats_model$date)
81 casos_estimats_model <- casos_estimats_model %>% group_by(Series, Year,
82   Month) %>%
83   summarise(Cases=sum(Cases))
84
85 ## Graficar les dades
86 ggplot(casos_estimats_model, aes(x = date, y = Cases, color = Series)) +
87   geom_line() +
88   labs(title = "Casos Estimats i Registrats per Mes a l'Area Metropolitana
89     Nord de Barcelona",
90     x = "Data",
91     y = "Numero de Casos") +
92   theme_minimal() +
93   scale_x_date(date_labels = "%Y", date_breaks = "24 months") +
94   theme(axis.text.x = element_text(angle = 45, hjust = 1))
95
96 ## Estadistics basics
97 casos_registrats <- casos_estimats_model[casos_estimats_model$Series
98   == "Registered",]$Cases
99
100 summary(casos_registrats)
101 sd(casos_registrats)
102
103 casos_estimats <- casos_estimats_model[casos_estimats_model$Series
104   == "Estimated",]$Cases
105
106 summary(casos_estimats)
107 sd(casos_estimats)
108
109 ### Finestra temporal on coincideixen les dades de les dues fonts
110 casos_estimats_model <- casos_estimats_model[casos_estimats_model$Year
111   >= 2013, ]
112 vdadesMonth_ABS_MN <- dadesMonth_ABS_MN[dadesMonth_ABS_MN$Any <= 2021, ]
113 casos_estimats_model$date <- as.Date(paste0(casos_estimats_model$Year,
114   "-", casos_estimats_model$Month, "-01"))

```

```

111 ### Visualitzacio evolucio casos detectats en AP + trucades linia
      d'atencio víctimes + estimacio model underreporting GBV
112
113 ## Etiquetes adequades per a la llegenda
114 casos_estimats_model$Series <- as.factor(casos_estimats_model$Series)
115 dadesMonth_ABS_MN$Series <- "Trucades"
116
117 ## Combinem els dataframes
118 combined_data <- rbind(casos_estimats_model,
119                        data.frame(date = dadesMonth_ABS_MN$date,
120                                Cases = dadesMonth_ABS_MN$Trucades,
121                                Series = "Trucades"))
122
123 ## Graquifem
124 global <- combined_data %>%
125 ggplot(aes(x = date, y = Cases, col = Series)) +
126 geom_line() +
127 labs(title = "Validacio Externa del Model",
128      x = "Data",
129      y = "Numero de Casos/Trucades",
130      color = "Series") +
131 theme_minimal() +
132 scale_x_date(date_labels = "%m-%Y", date_breaks = "6 month") +
133 theme(plot.title = element_text(hjust = 0.5),
134       axis.text.x=element_text(angle=60, hjust=1)
135 ) +
136 ylim(0, 390) +
137 scale_color_manual(values = c("Estimated" = "#43CD80", "Registered" =
138                               "#7EC0EE", "Trucades" = "#FF6A6A"))
139
140 print(global)
141
142 ## Estadistics basics
143 summary(dadesMonth_ABS_MN$Trucades)
144 sd(dadesMonth_ABS_MN$Trucades)
145
146 casos_registrats1 <- casos_estimats_model[casos_estimats_model$Series
147 == "Registered",]$Cases
148 summary(casos_registrats1)
149 sd(casos_registrats1)
150
151 casos_estimats1 <- casos_estimats_model[casos_estimats_model$Series
152 == "Estimated",]$Cases
153 summary(casos_estimats1)
154 sd(casos_estimats1)
155
156 # 2) QUANTIFICACIO DE L'IMPACTE DE LA COVID

```

```

153
154 ## Grafiquem la franja temporal del covid (confinament)
155 conf_ini <- which(dadesMens$Any == 2020 & dadesMens$Month == 3)
156 conf_fin <- which(dadesMens$Any == 2020 & dadesMens$Month == 6)
157
158 plot(dadesMens$Trucades, type="l", ylab="Numero Mensual de Trucades",
159       xlab="Index")
159 abline(v = conf_ini, col="red", lwd=1, lty=2)
160 abline(v = conf_fin, col="red", lwd=1, lty=2)
161
162 ## Estadistics basics
163 summary(dadesMens$Trucades)
164 sd(dadesMens$Trucades)
165 summary(dadesMens$Trucades[0:87])
166 sd(dadesMens$Trucades[0:87])
167 summary(dadesMens$Trucades[90:132])
168 sd(dadesMens$Trucades[90:132])
169
170 ## ADF a la serie temporal pre-confinament
171 resultado_adf <- adf.test(dadesMens$Trucades[0:87], k=0) #type 2
172     constante + tend
173
174 ## Imprimir el resultat
175 print(resultado_adf)
176
177 ## ADF post-confinament
178 resultado_adf <- adf.test(dadesMens$Trucades[95:132], alternative =
179     "stationary", k=0)
180
181 ## Imprimir el resultado
182 print(resultado_adf)
183
184 ##CAUSAL IMPACT
185 set.seed(1234)
186 impacto_causal <- CausalImpact(dadesMens$Trucades, pre.period = c(1,
187     86), post.period = c(90,132))
188
189 # Graficar els resultats
190 plot(impacto_causal)
191 summary(impacto_causal)
192
193 # 3) PREDICCIONS
194
195 ## ARIMA
196
197 dades <- read.csv("L_nia_d_atenci__contra_

```

```

196 la_viol_ncia_masclista_900_900_120_20240225.csv", header=TRUE, sep=",")
197 dades<-dades[dades$Sexe=="Dona",]
198 dades$Edat[dades$Edat=="Menors de 18 anys"]<-"Menys de 18 anys"
199
200 ## Agafem casos de VGS i VGF
201 dades <- dades[dades$Motiu.Trucada == "Violencia masclista" |
    dades$Violencia.fisica=="Si" | dades$Violencia.sexual=="Si",]
202
203 ## Dades mensuals
204 dades$Data <- as.Date(paste0(substr(dades$Data, 1, 6), dades$Any),
    format="%d/%m/%Y")
205 dades <- dades[order(dades$Data), ]
206 dades$Month <- month(dades$Data)
207 dadesMens <- dades %>% group_by(Any, Month) %>% summarise(Trucades=n())
208
209 ## Serie temporal
210 ser <- ts(dadesMens$Trucades,
211     freq=12,
212     start=c(2013, 1))
213
214 plot(ser, ylab = "Trucades", xlab="Temps")
215
216 ## Identificacio del model
217 ## Variancia
218 boxplot(ser~floor(time(ser)), ylab = "Serie", xlab="Any",
    main="Boxplots")
219
220 m = apply(matrix(ser, ncol=11), 1, mean)
221 v = apply(matrix(ser, ncol=11), 1, var)
222 fit <- lm(v ~ m)
223 plot(v~m, main="Grafic Mitjana-Variancia", xlab="Mitjana",
    ylab="Variancia")
224 abline(fit, col = "red")
225
226 ## Apliquem logaritme
227 lnser=log(ser)
228 plot(lnser)
229
230 boxplot(lnser~floor(time(lnser)), ylab = "Serie", xlab="Any",
    main="Boxplots")
231
232 m = apply(matrix(lnser, ncol=11), 1, mean)
233 v = apply(matrix(lnser, ncol=11), 1, var)
234 fit <- lm(v ~ m)
235 plot(v~m, main="Grafic Mitjana-Variancia", xlab="Mitjana",
    ylab="Variancia")
236 abline(fit, col = "red")

```

```

237
238 ## Estacionalitat
239 plot(decompose(lnser))
240
241 monthplot(lnser, ylab = "Serie")
242
243 ## Diferenciacio estacional
244 d12lnser = diff(lnser,12)
245 plot(d12lnser, ylab="d12lnserie")
246 abline(v=2013:2023, h=mean(d12lnser), lty=3, col=4)
247
248 mean(d12lnser)
249
250 ## Mitjana constant?
251 d1d12lnser <- diff(d12lnser)
252 plot(d1d12lnser, main="d1d12ser")
253 abline(h=0)
254
255 abline(h=mean(d1d12lnser), col=2)
256
257 mean(d1d12lnser)
258
259 var(lnser)
260 var(d12lnser)
261 var(d1d12lnser)
262 var(diff(d1d12lnser))
263
264 vari <- data.frame(lnSerie=var(lnser), d12lnserie=var(d12lnser),
265 d1d12lnserie=var(d1d12lnser), d1d1d12lnserie=var(diff(d1d12lnser)) )
266
267 colnames(vari) <- c("lnSerie", "d12lnserie", "d1d12lnserie",
268 "d1d1d12lnserie")
269
270 kable(vari, caption="Comparacio de variancies", align = "c", booktabs=T)
271 %>%
272 kable_styling(position = "center",
273 latex_options = "HOLD_position", font_size = 9)
274
275 #####Validation#####
276 validation=function(model,dades){
277   s=frequency(get(model$series))
278   resid=model$residuals
279   par(mfrow=c(2,2), mar=c(3,3,3,3))
280   #Residuals plot
281   plot(resid, main="Residuals")
282   abline(h=0)

```

```

282 abline(h=c(-3*sd(resid),3*sd(resid)),lty=3,col=4)
283 #Square Root of absolute values of residuals (Homocedasticity)
284 scatter.smooth(sqrt(abs(resid)),
285 main="Square Root of Absolute residuals",
286 lpars=list(col=2))
287
288 #Normal plot of residuals
289 qqnorm(resid)
290 qqline(resid,col=2,lwd=2)
291
292 ##Histogram of residuals with normal curve
293 hist(resid,breaks=20,freq=FALSE)
294 curve(dnorm(x,mean=mean(resid),sd=sd(resid)),col=2,add=T)
295
296 ## Individual Correlation Tests
297 #ACF & PACF of residuals
298 par(mfrow=c(1,2))
299 acf(resid,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,s-1)),lwd=1)
300 pacf(resid,ylim=c(-1,1),lag.max=60,col=c(rep(1,s-1),2),lwd=1)
301 par(mfrow=c(1,1))
302
303 #ACF & PACF of square residuals
304 dev.new(width=12, height=6)
305
306 par(mfrow=c(1,2))
307 acf(resid^2,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,s-1)),lwd=1)
308 pacf(resid^2,ylim=c(-1,1),lag.max=60,col=c(rep(1,s-1),2),lwd=1)
309 par(mfrow=c(1,1))
310
311 #Global Correlation Test
312 #Ljung-Box p-values
313 par(mar=c(2,2,1,1))
314 tsdiag(model,gof.lag=7*s)
315 cat("\n-----\n")
316 print(model)
317
318 #Stationary and Invertible
319 cat("\nModul of AR Characteristic polynomial Roots: ",
320 Mod(polyroot(c(1,-model$model$phi))),"\n")
321 cat("\nModul of MA Characteristic polynomial Roots: ",
322 Mod(polyroot(c(1,model$model$theta))),"\n")
323
324 #Model expressed as an MA infinity (psi-weights)
325 psis=ARMAtoMA(ar=model$model$phi,ma=model$model$theta,lag.max=72)
326 names(psis)=paste("psi",1:72)
327 cat("\nPsi-weights (MA(inf))\n")
328 cat("\n-----\n")

```

```

329 print(psis[1:20])
330
331 plot(psis,type="h",main="Pesos Psis - MA infinito")
332
333 #Model expressed as an AR infinity (pi-weights)
334 pis=-ARMAtoMA(ar=-model$model$theta,ma=-model$model$phi,lag.max=72)
335 names(pis)=paste("pi",1:72)
336 cat("\nPi-weights (AR(inf))\n")
337 cat("\n-----\n")
338 print(pis[1:20])
339
340 plot(pis,type="h",main="Pesos Pis - AR infinito")
341
342 # #Some Complementary Tests
343 # cat("\nNormality tests\n")
344 # cat("\n-----\n")
345 # ##Shapiro-Wilks Normality test
346 # print(shapiro.test(resid(model)))
347 #
348 # suppressMessages(require(nortest,quietly=TRUE,warn.conflicts=FALSE))
349 # ##Anderson-Darling test: Normality
350 # print(ad.test(resid(model)))
351 #
352 # suppressMessages(require(tseries,quietly=TRUE,warn.conflicts=FALSE))
353 # ##Jarque-Bera test: Normality
354 # print(jarque.bera.test(resid(model)))
355 #
356 # cat("\nHomoscedasticity Test\n")
357 # cat("\n-----\n")
358 # suppressMessages(require(lmtest,quietly=TRUE,warn.conflicts=FALSE))
359 # ##Breusch-Pagan test
360 # obs=get(model$series)
361 # print(bptest(resid(model)~I(obs-resid(model))))
362 #
363 # cat("\nIndependence Tests\n")
364 # cat("\n-----\n")
365 #
366 # ##Durbin-Watson test
367 # print(dwtest(resid(model)~I(1:length(resid(model)))))
368 #
369 ##Ljung-Box test
370 cat("\nLjung-Box test\n")
371 print(t(apply(matrix(c(1:4,(1:4)*s)),1,function(el) {
372   te=Box.test(resid(model),type="Ljung-Box",lag=el)
373   c(lag=(te$parameter),statistic=te$statistic[[1]],
374     p.value=te$p.value)})))
375

```

```

376
377 #Sample ACF vs. Teoric ACF: similar?
378 par(mfrow=c(2,2),mar=c(3,3,3,3))
379 acf(dades, ylim=c(-1,1), lag.max=36, main="Sample ACF")
380
381 plot(ARMAacf(model$model$phi, model$model$theta, lag.max=36), ylim=c(-1,1),
382      type="h", xlab="Lag", ylab="", main="ACF Teoric")
383 abline(h=0)
384
385 #Sample PACF vs. Teoric PACF
386 pacf(dades, ylim=c(-1,1), lag.max=36, main="Sample PACF")
387
388 plot(ARMAacf(model$model$phi, model$model$theta, lag.max=36,
389             pacf=T), ylim=c(-1,1),
390      type="h", xlab="Lag", ylab="", main="PACF Teoric")
391 abline(h=0)
392 par(mfrow=c(1,1))
393 }
394 ##### End Validation #####
395
396 ## Model fitting i validacio
397 ## Analitzem el ACF
398
399 acf(d1d12lnser, ylim=c(-1,1), lag.max=60, col=c(2, rep(1,11)), lwd=2)
400
401 ## Identificacio del model
402 par(mfrow=c(1,2))
403 acf(d1d12lnser, lag.max=72, ylim=c(-1,1), col=c(2, rep(1,11)))
404 pacf(d1d12lnser, lag.max=72, ylim=c(-1,1), col=c(rep(1,11), 2))
405
406 ### Primer model
407 par(mfrow=c(1,1))
408 (mod1=arima(d1d12lnser, order=c(2,0,2),
409            seasonal=list(order=c(0,0,1), period=12)))
410
411 coef1<- data.frame(round(rbind(mod1$coef, sqrt(diag(mod1$var.coef))),
412                      mod1$coef/sqrt(diag(mod1$var.coef))), 4))
413 row.names(coef1) <- c("Coef", "s.e.", "T-ratios")
414
415 kable(coef1) %>%
416   kable_styling(position = "center", latex_options = c("HOLD_position"))
417
418 (mod1=arima(lnser, order=c(2,1,2), seasonal=list(order=c(0,1,1), period=12)))
419
420 coef1<- data.frame(round(rbind(mod1$coef, sqrt(diag(mod1$var.coef))),
421                      mod1$coef/sqrt(diag(mod1$var.coef))), 4))

```

```

422 row.names(coef1) <- c("Coef","s.e.,"T-ratios")
423
424 kable(coef1) %>%
425   kable_styling(position = "center", latex_options = c("HOLD_position"))
426
427 (mod1=arima(lnser,order=c(2,1,1),seasonal=list(order=c(0,1,1),period=12)))
428
429 coef1<- data.frame(round(rbind(mod1$coef,sqrt(diag(mod1$var.coef)),
430   mod1$coef/sqrt(diag(mod1$var.coef))),4))
431 row.names(coef1) <- c("Coef","s.e.,"T-ratios")
432
433 kable(coef1) %>%
434   kable_styling(position = "center", latex_options = c("HOLD_position"))
435
436 (mod1=arima(lnser,order=c(1,1,1),seasonal=list(order=c(0,1,1),period=12)))
437
438 coef1<- data.frame(round(rbind(mod1$coef,sqrt(diag(mod1$var.coef)),
439   mod1$coef/sqrt(diag(mod1$var.coef))),4))
440 row.names(coef1) <- c("Coef","s.e.,"T-ratios")
441
442 kable(coef1) %>%
443   kable_styling(position = "center", latex_options = c("HOLD_position"))
444
445
446 ### Segon model
447 (mod4=arima(d1d12lnser,order=c(2,0,0),
448   seasonal=list(order=c(3,0,0),period=12)))
449
450 coef4<- data.frame(round(rbind(mod4$coef,sqrt(diag(mod4$var.coef)),
451   mod4$coef/sqrt(diag(mod4$var.coef))),4))
452 row.names(coef4) <- c("Coef","s.e.,"T-ratios")
453
454 kable(coef4) %>%
455   kable_styling(position = "center", latex_options = c("HOLD_position"))
456
457 (mod4=arima(lnser,order=c(0,1,0),seasonal=list(order=c(2,1,0),period=12)))
458
459 coef4<- data.frame(round(rbind(mod4$coef,sqrt(diag(mod4$var.coef)),
460   mod4$coef/sqrt(diag(mod4$var.coef))),4))
461 row.names(coef4) <- c("Coef","s.e.,"T-ratios")
462
463 kable(coef4) %>%
464   kable_styling(position = "center", latex_options = c("HOLD_position"))
465
466 AIC(mod1, mod4)
467
468 ## Validacio del model

```

```

469 dades = d1d12l1nser
470 model=mod1
471 validation(mod1)
472
473 ## Predicció
474
475 ##### Estabilitat del Model (SENSE CONSTANT!!!!) #####
476 ultim=c(2022,12)
477 pdq=c(1,1,1)
478 PDQ=c(0,1,1)
479
480 serie2=window(ser,end=ultim)          #Incomplete series: leave last year
      out!
481 lnserie2=log(serie2)
482 serie1=window(ser,end=ultim+c(1,0)) #complete series
483 lnserie1=log(serie1)
484
485 (modA=arima(lnserie1,order=pdq,seasonal=list(order=PDQ,period=12)))
486 coefA<- data.frame(round(rbind(modA$coef,sqrt(diag(modA$var.coef)),
487 modA$coef/sqrt(diag(modA$var.coef))),4))
488 row.names(coefA) <- c("Coef","s.e.,"T-ratios")
489 kable(coefA) %>%
490   kable_styling(position = "center", latex_options = c("HOLD_position"))
491
492 (modB=arima(lnserie2,order=pdq,seasonal=list(order=PDQ,period=12)))
493 coefB<- data.frame(round(rbind(modB$coef,sqrt(diag(modB$var.coef)),
494 modB$coef/sqrt(diag(modB$var.coef))),4))
495 row.names(coefB) <- c("Coef","s.e.,"T-ratios")
496 kable(coefB) %>%
497   kable_styling(position = "center", latex_options = c("HOLD_position"))
498
499 ## Predicció
500 pred=predict(modB,n.ahead=12) #using all but last year observations
501 pr<-ts(c(tail(lnserie2,1),pred$pred),start=ultim,freq=12)
502 se<-ts(c(0,pred$se),start=ultim,freq=12)
503
504 #Intervals
505 tl<-ts(exp(pr-1.96*se),start=ultim,freq=12)
506 tu<-ts(exp(pr+1.96*se),start=ultim,freq=12)
507 pr<-ts(exp(pr),start=ultim,freq=12)
508
509
510 ts.plot(ser,tl,tu,pr,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=c(2013,2024),
511 type="o",main=paste("Model ARIMA(",paste(pdq,collapse=""),")",
512 ("",paste(PDQ,collapse=""),")_12",sep=""))
513 abline(v=2013:2024,lty=3,col=4)
514

```

```

515 (previs=window(cbind(tl,pr,tu,ser,error=round(ser-pr,3)),start=ultim))
516
517 obs=window(ser,start=ultim)
518 (mod.EQM1=sqrt(sum(((obs-pr)/obs)^2)/12))
519
520 (mod.EAM1=sum(abs(obs-pr)/obs)/12)
521
522 ## Metriques d'error:
523
524 RMSPE<- function(obs,pr){
525   r <- sqrt(sum(((obs-pr)/obs)^2)/12)
526   return(r)
527 }
528
529 MAPE <- function(obs,pr){
530   r <- sum(abs(obs-pr)/obs)/12
531   return(r)
532 }
533
534 MAE <- function(obs,pr){
535   r <- sum(abs(obs-pr))/12
536   return(r)
537 }
538
539 RMSE <- function(obs,pr){
540   r <- sqrt(sum(((obs-pr))^2)/12)
541   return(r)
542 }
543
544 nRMSE <- function(obs,pr){
545   r <- RMSE(obs,pr)/abs(mean(obs))
546   return(r)
547 }
548
549 nMAE <- function(obs,pr){
550   r <- MAE(obs,pr)/abs(mean(obs))
551   return(r)
552 }
553
554 ## Dades entrenament
555 RMSPE(obs,pr)
556 MAPE(obs,pr)
557 RMSE(obs,pr)
558 MAE(obs,pr)
559 nRMSE(obs,pr)
560 nMAE(obs,pr)
561

```

```

562 ## Predicció futura
563 pred=predict(modA,n.ahead=12)
564 pr<-ts(c(tail(lnserie1,1),pred$pred),start=ultim+c(1,0),freq=12)
565 se<-ts(c(0,pred$se),start=ultim+c(1,0),freq=12)
566
567 #Intervals
568 tl1<-ts(exp(pr-1.96*se),start=ultim+c(1,0),freq=12)
569 tu1<-ts(exp(pr+1.96*se),start=ultim+c(1,0),freq=12)
570 pr1<-ts(exp(pr),start=ultim+c(1,0),freq=12)
571
572 ts.plot(ser,tl1,tu1,pr1,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=c(2013,2024),
573 type="o", main=paste("Model ARIMA(",paste(pdq,collapse=","),")
574 (" ,paste(PDQ,collapse=","),")_12",sep=""))
575 abline(v=2013:2024,lty=3,col=4)
576
577 (previs1=window(cbind(tl1,pr1,tu1),start=ultim+c(1,0)))
578
579 ## Dades prova
580 RMSPE(obs,pr)
581 MAPE(obs,pr)
582 RMSE(obs,pr)
583 MAE(obs,pr)
584 nRMSE(obs,pr)
585 nMAE(obs,pr)
586
587 # ANN
588
589 measures=function(pred,obs){
590
591     err=pred-obs
592     perr=err/pred
593
594     c(RMSE=sqrt(mean(err^2)),
595       MAE=mean(abs(err)),
596       nRMSE=sqrt(mean(err^2))/mean(obs),
597       nMAE=mean(abs(err))/mean(obs),
598       RMSPE=sqrt(mean(perr^2)),
599       MAPE=mean(abs(perr)))
600 }
601
602 m=5
603 ny=1
604 serieTot = dades_ts
605 mi=min(serieTot)*0.9
606 ma=max(serieTot)*1.1
607
608 dades=serieTot

```

```

609 for (i in 1:m){
610   dades=ts.intersect(dades, stats::lag(serieTot,-i))
611 }
612 dades=as.data.frame(dades)
613 names(dades)=paste0("y",c("",paste0("_",1:m)))
614 dades=(dades-mi)/(ma-mi)
615
616 set.seed(123456)
617 kf=1
618 years=3
619 npred=12*years
620 idx=c(rep(1,nrow(dades)-npred),rep(2,npred))
621
622 train=as.matrix(dades[idx!=2,])
623 test=as.matrix(dades[idx==2,])
624
625 resp="y"
626 if(ny>1) resp=paste0(resp,c("",paste0("_",1:(ny-1))),collapse="+")
627 expl=paste0("y_",ny:m, collapse = "+")
628 f <- as.formula(paste0(resp, "~",expl))
629
630 net <- neuralnet(f, train, hidden = c(6, 6), rep = 1, threshold =
        0.00125)
631
632 plot(net)
633
634 ## Predicció
635 pred.train=mi+
636 predict(net,train[nrow(train)-(npred-1):0,-c(1:ny)])*(ma-mi)
637 pred.test=mi+predict(net,test[, -c(1:ny)])*(ma-mi)
638 obs.train=mi+train[nrow(train)-(npred-1):0,1:ny]*(ma-mi)
639 obs.test=mi+test[,1:ny]*(ma-mi)
640
641 pr=ts(pred.train[, 1], start=2024-2*years,freq=12)
642 ts.plot(serieTot, pr, lty=c(1,1), col=c(1,2),
        xlim=2024-c(2*years,years), type="o", main="NeuralNet (Train)")
643 abline(v=2024-years-years:0,lty=3,col=4)
644
645 (me3=measures(pred.train, obs.train))
646
647 ## Predicció futura
648 pr2=ts(pred.test,start=2024-years, end=2025, freq=12)
649 ts.plot(serieTot,pr2,lty=c(1,1),col=c(1,2),xlim=c(2018,2025),type="o",
        main="NeuralNet (Test)")
650 abline(v=2025-years:0,lty=3,col=4)
651
652
653 (me4=measures(pred.test,obs.test))

```