Clàudia Barahona Tarragó

ESTIMACIÓ DE L'IMPACTE DE LES CONDICIONS AMBIENTALS I LA MOBILITAT SOBRE LA INCIDÈNCIA DE MALALTIES INFECCIOSES DE TRANSMISSIÓ AÈRIA A PARTIR DE GRANS VOLUMS DE DADES

TREBALL DE FI DE GRAU

Dirigit pel Dr. Alexandre Fabregat Tomàs

Grau d'Enginyeria Biomèdica

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2024

Agraïments

M’agradaria expressar el meu agraïment a tothom que ha contribuït a aquest treball. Especialment, vull donar les gràcies a:

El Dr. Alexandre Fabregat Tomàs, pel seu compromís i guiatge en la investigació. La seva orientació, recolzament i constància han estat crucials per al desenvolupament i èxit de l'estudi.

Els meus pares i la meva germana, pel seu suport incondicional, així com la seva confiança, paciència, ajuda i amor durant tot el procés d’aquest treball.

Les meves companyes de carrera, amb qui vam esdevenir amigues, per aquests quatre anys d’un acompanyament i ajut immillorables.

Finalment, a totes aquelles persones que en algun moment m’han animat i ajudat al llarg del procés que ha estat aquesta investigació. Moltes gràcies a tots.

Resum

En aquest treball s'investiga la relació entre les condicions ambientals i la mobilitat i la incidència de malalties infeccioses de transmissió aèria, amb un enfocament especial en la COVID-19 i la grip dins l'àmbit territorial de Catalunya. Mitjançant l'anàlisi de grans volums de dades públiques, i amb el llenguatge de programació Python, s'han recopilat i processat diverses fonts, incloent dades de vigilància sindròmica, meteorològiques, cartogràfiques i de mobilitat humana. L'estratificació de les dades originals ha permès investigar el rol de l'edat, la regió sanitària, el sexe i el nivell socioeconòmic sobre la incidència.

Per assolir l'objectiu, per a cada diagnòstic s'han ajustat models de regressió lineal múltiple (MLR) utilitzant valors retardats de les variables de temperatura, humitat, mobilitat interna i interior i exterior de Catalunya i incidència de l'any 2022. Posteriorment, els models de MLR s'han utilitzat per a predir nous valors del 2023 i se n'ha avaluat la precisió.

Els resultats mostren una relació inversa entre les condicions ambientals i la incidència, i, amb els viatges, una relació positiva amb els d'entrada i els interiors i una de negativa amb els de sortida. Per a la COVID-19 s'ha evidenciat un major impacte de la mobilitat davant de les condicions ambientals i, per a la grip, ha estat a l'inrevés, ja que la temperatura sembla tenir un major efecte sobre la incidència.

Paraules clau: malalties infeccioses, mobilitat, dades meteorològiques, MLR, cartografia

Resumen

En este trabajo se investiga la relación entre las condiciones ambientales y la movilidad y la incidencia de enfermedades infecciosas de transmisión aérea, con un enfoque especial en la COVID-19 y la gripe dentro del ámbito territorial de Cataluña. Mediante el análisis de grandes volúmenes de datos públicos, y con el lenguaje de programación Python, se han recopilado y procesado varias fuentes, incluyendo datos de vigilancia sindrómica, meteorológicos, cartográficos y de movilidad humana. La estratificación de los datos originales ha permitido investigar el rol de la edad, la región sanitaria, el sexo y el nivel socioeconómico sobre la incidencia.

Para lograr el objetivo, para cada diagnóstico se han ajustado modelos de regresión lineal múltiple (MLR) utilizando valores retardados de las variables de temperatura, humedad, movilidad interna e interior y exterior de Cataluña e incidencia del año 2022. Posteriormente, los modelos de MLR se han usado para predecir nuevos valores del 2023 y se ha evaluado su precisión.

Los resultados muestran una relación inversa entre las condiciones ambientales y la incidencia, y, en cuanto a los viajes, una relación positiva con los de entrada y los interiores y una de negativa con los de salida. Por lo que hace a la COVID-19, se ha evidenciado un mayor impacto de la movilidad ante las condiciones ambientales y, por lo que hace a la gripe, ha sido al revés, puesto que la temperatura parece tener un mayor efecto sobre la incidencia.

Palabras clave: enfermedades infecciosas, movilidad, datos meteorológicos, MLR, cartografía

Abstract

This work investigates the relationship between environmental conditions and the mobility and the incidence of airborne infectious diseases, with a special focus on COVID-19 and influenza within the territorial scope of Catalonia. Through the analysis of large volumes of public data, and using the Python programming language, several sources have been collected and processed, including syndromic surveillance, weather, cartographic and human mobility data. Stratification of the original data has allowed us to investigate the role of age, health region, gender, and socioeconomic status on incidence.

To achieve the objective, multiple linear regression models (MLR) have been adjusted for each diagnosis using delayed values of the variables of temperature, humidity, internal and external mobility of Catalonia and incidence of the year 2022. The MLR models have subsequently been used to predict new 2023 values and their accuracy has been assessed.

The results show an inverse relation between environmental conditions and incidence, and, regarding travel, a positive relation with incoming and indoor mobility and a negative relation with outgoing mobility has been revealed. For COVID-19, a greater impact of mobility has been shown against environmental conditions and, for the flu, it has been the other way around since the temperature seems to have a greater effect on the incidence.

Key words: infectious diseases, mobility, meteorological data, MLR, cartography

Índex

Llista de Figures

1 Introducció

1. 1 Context

Les malalties infeccioses han estat i es mantenen inextricablement lligades als humans, no només per factors sanitaris, sinó també per factors socials, culturals, econòmics i polítics, de manera que les epidèmies són esdeveniments socials i biològics des dels principis de la humanitat [1]. Amb una àmplia gamma de tipologies i classificacions es troben entre les cinc primeres causes de mort a escala global, i la seva transmissió roman un problema, donat que són causades per microorganismes que proliferen ràpidament, s'adapten fàcilment i muten freqüentment, de manera que contínuament n'hi ha d'emergents [2]. L'impacte i la rellevància que tenen és considerable, i la situació a Catalunya no n'és una excepció. És per les dramàtiques emergències que han causat en les últimes dècades, i també els sacsejos als sistemes mèdics, socials, econòmics i polítics, amb la profunda i duradora empremta que això deixa, que aquest estudi pretén estimar l'impacte de les condicions meteorològiques i la mobilitat en la incidència de dues d'elles, la COVID-19 i la grip. Per a arribar a tals resultats, la investigació ha aprofundit en la primera, i s'ha servit de la segona per a propòsits comparatius.

Així doncs, del conjunt de dades de malalties infeccioses de transmissió aèria, d'entre els 9 possibles diagnòstics que es registren al Sistema d'Informació per a la Vigilància d'Infeccions a Catalunya (SIVIC), s'ha optat per posar l'atenció en la malaltia causada pel SARS-CoV-2. Tot i l'abundància d'estudis realitzats sobre la COVID-19, no deixa de ser sorprenent la manera com un virus del qual abans del 2019 no es tenia constància ha pogut afectar en tants aspectes de la vida quotidiana de la gent en l'àmbit mundial.

El procés d'investigació científica és considerablement costós tant en termes econòmics com temporals. Especialment al començament de la pandèmia, el factor temporal va ser crític. La notable alta transmissibilitat del virus, combinada amb un coneixement inicial pràcticament inexistent, van provocar un ràpid augment de contagis a escala mundial en un curt període de temps. Aquesta situació va accelerar la urgència de la investigació sobre la COVID-19, amb els diferents països ajustant i implementant mesures de contenció del virus a mesura que es publicaven nous articles amb descobriments rellevants.

La pandèmia de la COVID-19, doncs, ha suposat un desafiament sanitari global en impactar profundament en tots els aspectes de la societat i en requerir una resposta científica rigorosa per comprendre la seva evolució i factors determinants. D'entre els estudis realitzats, n'hi ha diversos que han posat l'atenció en la influència dels factors geogràfics i climàtics en la incidència. Canvis en la temperatura, la humitat, la radiació ultraviolada o el vent han estat relacionats amb la transmissibilitat del virus, de manera que alguns estudis classifiquen la COVID-19 com una infecció estacional, una característica que es veu en altres malalties respiratòries [3]. Alguns exemples d'estudis, com el de M. S. Al-Khateeb et al. (2023), conclouen que algunes variables climàtiques són factors significants a l'hora de determinar la velocitat de transmissió de la COVID-19. En concret, aquest afirma que la velocitat de transmissió augmenta durant els mesos de fred a les regions de l'hemisferi nord i, a l'hemisferi sud, la humitat és el factor que mostra una relació més forta amb el nombre de casos de COVID-19 [4].

En la mateixa línia, però en un estudi enfocat a Catalunya, A. Tobías et al. (2021) suggereixen que els climes càlids i humits podrien haver causat una reducció de la incidència a Catalunya durant la primera onada pandèmica [5]. En un altre estudi, M. Moazeni et al. (2023) indiquen que la temperatura és el factor climàtic que més influeix en la pandèmia de

la COVID-19 en la majoria dels països [6]. I, com aquests, múltiples estudis més reafirmen la negativa correlació entre la temperatura i la taxa de creixement [7], [8], [9].

No obstant això, encara hi ha una manca d'evidència, ja que altres estudis han mostrat resultats heterogenis i contradictoris, sense arribar a trobar associació entre la transmissió de la COVID-19 i els factors meteorològics [10], [11]. O, d'altres, en diferents províncies d'un mateix país han obtingut correlacions positives i negatives entre el COVID-19 i, en aquest cas, la temperatura [12].

A banda dels factors ambientals, també s'ha suggerit que la mobilitat humana afecta la transmissió de la COVID-19. I és que, donat que el SARS-CoV-2 és un virus que es transmet per via aèria, algunes de les primeres mesures que es van implementar van ser per evitar el contacte interpersonal. Aquestes mesures incloïen estrictes quarantenes i, posteriorment, tocs de queda i restriccions de la mobilitat. Essencialment, la mobilitat es va veure com un indicador indirecte de les interaccions socials i, per tant, es va considerar un factor clau en l'evolució de la malaltia.

Per a les fases inicials de la pandèmia, es va demostrar que la mobilitat estava altament relacionada amb la velocitat de transmissió de la COVID-19 [13]. Tot i això, diversos estudis apunten a que aquesta forta relació varia significativament arreu de món i al llarg del temps, de manera que a partir de l'estiu del 2020 la mobilitat passa a ser menys important per la predicció de la transmissió [14]. De fet, N. K. Bergman et al. (2023) indiquen que això s'evidencia sobretot en zones on es van aplicar altres mesures com les de l'ús de la mascareta o la ventilació dels espais, entre d'altres [15].

En aquest estudi, s'han recopilat dades de cinc fonts diferents, totes elles d'accés públic. Les fonts en qüestió són les relatives a les dades meteorològiques i de localització de les estacions meteorològiques automàtiques, a les dades de mobilitat, a la informació cartogràfica de les regions sanitàries i de les àrees bàsiques de salut de Catalunya i a les dades de vigilància sindròmica de les malalties infeccioses. Partint d'aquesta base, per a començar la recerca s'han analitzat més exhaustivament les dades d'incidència de la COVID-19 del 2020, l'any que emergeix la malaltia, fins al 2023. Addicionalment, per analitzar si les diferents mostres, desagregades per edat, sexe, regió sanitària i índex socioeconòmic, tenen una mateixa distribució estadística, s'ha realitzat el test Kolmogórov-Smirnov amb les dades del 2022. S'ha seleccionat aquest any perquè el model de regressió lineal múltiple que s'utilitza per a determinar la relació entre els diferents factors i la incidència s'ajusta amb les dades del 2022.

En el marc contextual de les dades referents a les diferents recerques científiques exposades anteriorment, aquest treball s'ha centrat en investigar la relació entre les condicions ambientals, la mobilitat i la incidència dels diagnòstics de COVID-19 i grip. Per a assolir aquest objectiu, s'han ajustat models de regressió lineal múltiple per a cada diagnòstic, utilitzant valors retardats de les variables de temperatura, humitat, mobilitat interna, interior i exterior de Catalunya i incidència de l'any 2022. Els resultats dels models s'han avaluat mitjançant els coeficients de determinació (R^2), i el model amb el valor més alt ha estat emprat per predir nous valors del 2023, avaluant-ne posteriorment la precisió. A més, per a cada model s'han analitzat els resultats dels coeficients dels predictors que, principalment, són els que indiquen el pes de cada factor en la incidència. Les dades utilitzades corresponen a l'any 2022, i això es deu a que les dades de mobilitat humana de caràcter públic que ofereix el Ministeri de Transports i Mobilitat Sostenible espanyol comencen a registrar-se a partir d'aquest any. I és que per tal de construir un model de regressió lineal múltiple, és imprescindible que les mostres continguin dades vàlides de tots

els factors, incloent-hi les condicions ambientals, la mobilitat i la incidència. Per a aquesta part de l'estudi, a banda de la COVID-19, s'ha volgut afegir el diagnòstic d'una altra malaltia infecciosa, en aquest cas de caràcter estacional. Es tracta de la grip, i s'ha optat per afegir-la a l'estudi perquè d'aquesta manera, en primer lloc, es pot comparar el grau de precisió de l'ajust dels models de MLR per a diferents malalties, així com l'impacte dels factors ambientals i la mobilitat; i, en segon lloc, es poden analitzar els resultats de cada model per a identificar patrons comuns i diferències clau entre diagnòstics.

D'entre els diferents llenguatges de programació adequats per a la recopilació, processament i anàlisi estadística de dades, en aquest estudi s'ha implementat el de Python, juntament amb l'entorn de desenvolupament Jupiter Notebook. Python ofereix una àmplia gamma de biblioteques especialitzades que faciliten les tasques d'anàlisi estadística i visualització de dades. Algunes de les quals s'ha fet ús són, per a la manipulació de dades, la biblioteca Pandas, que permet treballar amb estructures de dades complexes d'una manera eficient; per a realitzar operacions matemàtiques complexes per a grans volums de dades, NumPy; per a realitzar proves estadístiques avançades i ajustament de models, SciPy i StatsModels; i, per a la visualització de dades, Matplotlib i Seabons, que han permès crear gràfics i visualitzacions detallades i personalitzades per a facilitar la interpretació dels resultats. L'ús combinat de totes aquestes eines ha facilitat el procés d'anàlisi al llarg dels diferents estadis de l'estudi.

1. 2 Objectius

L'objectiu principal d'aquest estudi ha estat estimar l'impacte de les condicions ambientals i la mobilitat en la incidència de malalties infeccioses de transmissió aèria —en concret, de la COVID-19 i de la grip. Els diferents objectius són:

1. Estudiar l'evolució de la incidència de la COVID-19 de manera general, però també desagregada per edat, regió sanitària, sexe i índex socioeconòmic.
2. Determinar, mitjançant l'ús del test de Kolmogórov-Smirnov, si les distribucions estadístiques de cada desagregació de COVID-19 coincideixen i, en conseqüència, identificar si les mostres provenen de la mateixa població o no.
3. Ajustar els models de regressió lineal múltiple per als diagnòstics de COVID-19 i grip de l'any 2022 amb les variables meteorològiques i de mobilitat per a diferents valors retardats d'aquestes.
4. Identificar quins models de MLR de cada diagnòstic són més precisos mitjançant l'anàlisi del coeficient de determinació.
5. Avaluar els valors dels coeficients dels predictors, els quals marquen el pes de cada factor en la incidència. Addicionalment, representar geogràficament els valors de cada ABS per a la posterior anàlisi.
6. Predir els valors d'incidència de les tres primeres setmanes de l'any 2023 per als diagnòstics de COVID-19 i grip.
7. Avaluar la precisió del model de regressió lineal múltiple, així com la qualitat de les prediccions calculades, amb la complementarietat de representacions geogràfiques.

2 Metodologies i Materials

2. 1 Recopilació i Preprocessament de les Dades

En aquest estudi s'ha dut a terme una recopilació exhaustiva de dades provinents de diverses fonts, publicades pels organismes oficials a Catalunya i l'Estat Espanyol, amb el propòsit d'analitzar i relacionar la incidència de malalties infeccioses respiratòries amb altres factors com ara els ambientals o de mobilitat.

2. 1. 1 Dades de la Vigilància Sindròmica d'Infeccions a l'Atenció Primària a Catalunya

Les dades relatives a la vigilància sindròmica d'infeccions a l'Atenció Primària són de tipus obert, de manera que qualsevol usuari d'Internet pot accedir-hi [16]. El conjunt en qüestió prové del sistema del Departament de Salut que monitora les malalties epidemiològiques de la grip i la COVID-19, i també altres virus respiratoris, per tal d'identificar tendències i evolucions, caracteritzar microbiològicament la natura dels virus circulants i desenvolupar futurs mètodes preventius. Es tracta del Sistema d'Informació per a la Vigilància d'Infeccions a Catalunya (SIVIC) [17].

El conjunt de dades del qual es parla té unes dimensions aproximades de 14M de files i 16 columnes, i es pot filtrar i exportar en format CSV. Per a cada entrada de la taula es recull el número de casos diagnosticats, juntament amb el valor de la població total corresponent, de manera que posteriorment es fa possible calcular-ne la incidència. D'entre els diferents virus respiratoris existents, són 9 els que es comprenen en la taula: altres IRA (Infeccions Respiratòries Agudes), bronquiolitis, COVID-19, escarlatina, faringoamigdalitis, faringoamigdalitis estreptocòccica, grip, impetigen i pneumònia.

Pel que fa a l'estructuració temporal de les dades, l'interval entre mostres és setmanal, de manera que es recull la data inicial i final de la setmana implicada; i, geogràficament, la classificació més extensiva és la de regió sanitària, mentre que la màxima desagregació que es pot fer és per ABS (Àrea Bàsica de Salut) —ambdós conceptes es definiran a l'apartat específic de les dades geogràfiques. A més a més, també hi ha la possibilitat de desagregar per sexe, edat i índex socioeconòmic (ISC). Per a la creació d'aquest últim, es tenen en compte factors com el percentatge de persones amb estudis insuficients, el percentatge de població desocupada, el percentatge de gent que viu sola, el nivell de renda, entre d'altres. L'índex s'interpreta com un indicador de privació, on valors més elevats indiquen nivells socioeconòmics més baixos (1 = nivell alt; 2 = nivell mitjà-alt; 3 = nivell mitjà-baix; 4 = nivell baix).

Per a aquest estudi, inicialment s'han extret les dades compreses en el període del 2019 al 2023, ambdós inclosos. La Figura 1 mostra el nombre de casos registrats en aquest període, desagregats per diagnòstic. A grans trets, es pot veure com, aproximadament a partir del gener de 2020, comencen a aparèixer casos de la COVID-19 i, durant els mesos que corresponen a la pandèmia de la COVID-19, hi ha una davallada general dels casos detectats de les altres malalties infeccioses que es transmeten per via aèria. El més probable és que la quarantena, juntament amb les posteriors mesures per prevenir la propagació del virus, com ara les mascaretes i les distàncies de seguretat o la restricció de viatges, van ajudar a reduir la incidència de la resta de malalties respiratòries. Aquestes dades, així com altres estudis com el de J. Reina et al. (2021) [18], evidencien l'efectivitat de les mesures que es van imposar i permeten veure com la pandèmia ha influït en els patrons epidemiològics generals.

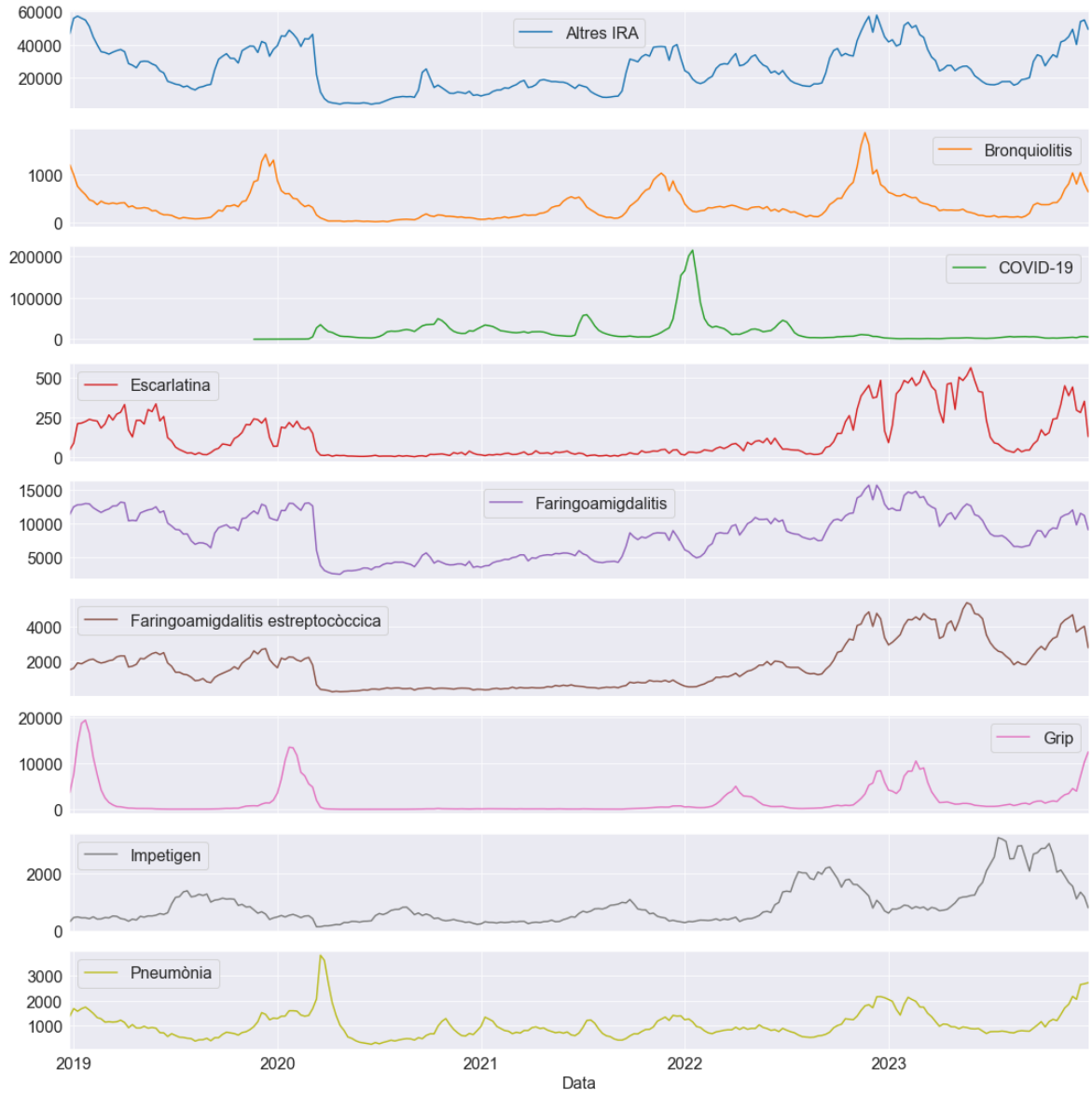


Figura 1. Nombre de casos registrats de les diferents malalties infeccioses respiratòries entre els anys 2019 i 2023.

Com a part del procés de preprocessament de les dades s'han exclòs aquelles línies de la taula que tenien una regió sanitària o una edat 'No disponible' i les que tenien un ISC igual a -1, per tal de centrar l'anàlisi en les mostres amb informació completa. Addicionalment, els grups d'edat s'han reagrupat en intervals més grans, reduint els 19 grups inicials a 7.

2. 1. 2 Dades Cartogràfiques de la Divisió Territorial Sanitària Oficial de Catalunya

En referència a la informació geogràfica de la divisió territorial sanitària oficial de Catalunya, el Departament de Salut presenta l'opció de descarregar-ne les dades obertament. Les dades són vectorials i estan en format ESRI Shapefile (SHP) [19].

Aquesta organització de les àrees geogràfiques sanitàries es desenvolupa de la següent manera: les àrees bàsiques de salut s'agrupen en sectors sanitaris, els quals, al seu torn, s'agrupen en regions sanitàries. Per a aquest estudi, les divisions més rellevants i de les quals s'ha fet ús són les de regió sanitària i d'àrea bàsica de salut. Seguint les definicions que proporciona CatSalut, els conceptes mencionats es corresponen a:

- Les regions sanitàries (RS). Se'n presenten deu en total i fan referència a les divisions del territori català delimitades segons factors geogràfics, socioeconòmics i demogràfics (Figura 2). Es regulen per decrets del Govern.
- Les àrees bàsiques de salut (ABS). Són les unitats territorials bàsiques del sistema sanitari a Catalunya, en les quals s'ofereix l'atenció primària a la població. Estan regulades a la Llei 15/1990 d'ordenació sanitària de Catalunya i a les diverses ordres del Departament de Salut. N'hi ha unes 380, que són actualitzades periòdicament en termes de designació i delimitació, i es conformen per un o més municipis en l'àmbit rural, o per districtes o barris a les àrees urbanes [20]. Es pot veure en la Figura 3 que a les zones amb més densitat de població, com ho és l'àrea metropolitana de Barcelona, hi ha un major nombre d'ABS, mentre que a les àrees menys poblades les ABS tenen una major extensió territorial.

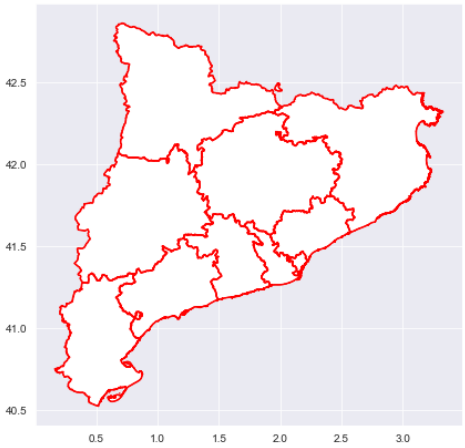


Figura 2. Distribució geogràfica de les regions sanitàries de Catalunya.

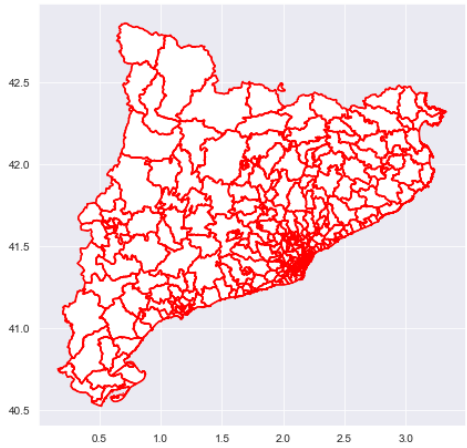


Figura 3. Distribució geogràfica de les àrees bàsiques de salut de Catalunya.

Per a treballar amb les dades geoespacials, s'han llegit i transformat els fitxers SHP corresponents a la informació de les RS i ABS en GeoDataFrames (unes estructures de dades de Geopandas). Aquests DataFrames amb què s'ha partit inclouen la columna *geometry*, la qual conté les coordenades que defineixen la forma i la ubicació de cada polígon. Posteriorment s'han transformat les coordenades al nou sistema de referència de coordenades, en anglès CRS, EPSG:4326. És important transformar la projecció, sobretot si s'ha de treballar amb conjunts de dades provinents de diferents fonts, ja que d'aquesta manera s'unifica el sistema de referència i així es permet una millor interoperabilitat [21].

2. 1. 3 Dades Meteorològiques de la XEMA: Temperatura i Humitat Relativa

El Departament d'Acció Climàtica, Alimentació i Agenda Rural proveeix les dades registrades a totes les estacions de la Xarxa d'Estacions Meteorològiques Automàtiques (XEMA) del Servei Meteorològic de Catalunya (METEOCAT) [22]. Són dades de domini públic i s'inclouen mesures horàries o semi-horàries de diferents variables meteorològiques, com ara la temperatura, la humitat relativa, la precipitació o la velocitat del vent, entre d'altres [23]. En aquest estudi són d'interès les dues primeres. Totes les dades es poden filtrar i exportar en un fitxer de tipus CSV.

Una de les columnes conté el codi identificador de l'Estació Meteorològica Automàtica, una informació que més endavant permetrà posicionar geogràficament el valor de les mesures.

Per a començar a treballar amb les dades, atès que hi ha una columna que recull el resultat del procés de validació, s'han filtrat únicament les dades vàlides. D'altra banda, donat que les dades corresponents a les infeccions a l'Atenció Primària tenen una periodicitat setmanal, el conjunt de mesures de temperatura i humitat relativa s'ha reagrupat mostrejant-se setmanalment, realitzant-ne la mitjana. Posteriorment, s'ha reduït el conjunt de dades per a mantenir només les dels anys 2022 i 2023, necessàries per a aquest estudi.

2. 1. 4 Dades de les Estacions Meteorològiques Automàtiques

La posició geogràfica de cadascuna de les estacions meteorològiques s'ha obtingut mitjançant les dades que es poden descarregar, en format CSV, des del portal de dades obertes de Catalunya [24]. La taula inclou les metadades associades a cadascuna de les estacions de la Xarxa d'Estacions Meteorològiques Automàtiques (XEMA), integrada a la Xarxa d'Equipaments Meteorològics de la Generalitat de Catalunya (Xemec), del Servei Meteorològic de Catalunya.

Del total de columnes que hi ha en el conjunt de dades, les que han estat d'interès són les relatives al codi de l'estació i a la georeferència, la qual conté la informació de les coordenades dels punts on es troben les estacions.

Tal com s'ha realitzat amb les dades de la divisió territorial sanitària de Catalunya, les coordenades dels punts de les estacions s'han transformat al mateix sistema de coordenades de referència, el EPSG:4326. En la Figura 4 es mostra la localització de les diferents estacions, les quals estan repartides de manera força uniforme per tot el territori, potser de manera més predominant al voltant de les grans ciutats.

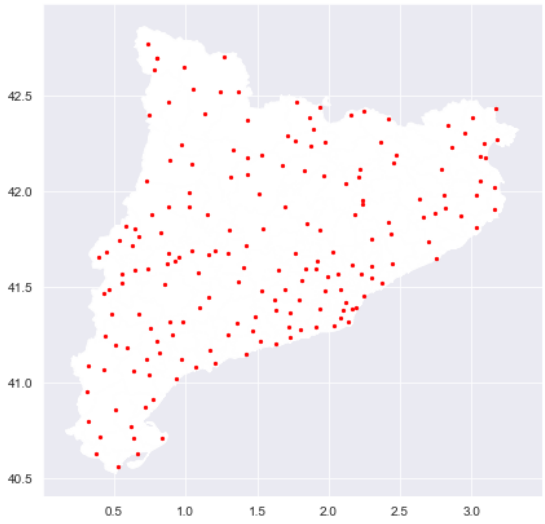


Figura 4. Distribució geogràfica de les estacions meteorològiques automàtiques de Catalunya.

2. 1. 5 Dades de Mobilitat Autonòmica

El Ministeri de Transports i Mobilitat Sostenible espanyol deixa a l'abast de tothom la informació relativa a la mobilitat nacional dels ciutadans [25]. El conjunt de dades conté registres diaris per hores, obtinguts a partir del dia 1 de gener del 2022, i contempla tots aquells viatges amb origen o destí en territori nacional (incloent els viatges a/de l'estranger).

Entre d'altres anàlisis, una de les quals han realitzat des del ministeri és per viatges, que és la d'interès per a l'estudi. Aquesta anàlisi ofereix la mobilitat horària per a cada parell d'origen i destí, agregant els viatges segons rangs de distància, tipus d'activitat i perfil sociodemogràfic del viatger (residència, sexe, rang d'edat i rang de renda), quantificant-la tant en número de viatges com en número de viatgers-km.

Les dades de viatges estan disponibles tant a nivell autonòmic com provincial i municipal (que inclou agregacions de municipis per protecció de dades). Per a aquest estudi s'han emprat les de tipus autonòmic, filtrant per la comunitat autònoma de Catalunya, i compreses entre l'1 de gener del 2022 i el 31 de desembre del 2023. Els viatges es divideixen en tres tipus de mobilitat: interior, exterior relativa a les entrades i exterior de sortides (Figura 5).

Figura 5. Viatges diaris a Catalunya en l'any 2022 diferenciats pel tipus de mobilitat.

En l'anterior figura s'observa que el número de viatges a l'interior de Catalunya és més elevat i força constant durant gran part de l'any, però disminueix cap allà al mes d'abril (a Setmana Santa), als mesos d'estiu, sobretot a l'agost, i a les dates de Nadal. Aquests períodes es corresponen amb etapes vacacionals, en què la gent no se desplaça tant interiorment per arribar als llocs de treball i d'estudi. En prestar atenció a la segona i tercera gràfica, és just en aquests períodes en què es troben els pics, especialment a l'abril. Això indicaria que hi ha més turistes que arriben de fora, però també més gent que surt de Catalunya.

Per a poder treballar amb les dades, donat que els registres son horaris, s'han resumit reagrupant-se en mostres de tipus setmanal, realitzant una suma dels valors del número de viatges.

Com a apunt, les dades en qüestió les han obtingut, principalment, a través dels registres anonimitzats de més de 13 milions de línies mòbils proporcionats per un operador mòbil. A diferència d'estudis anteriors, aquests registres inclouen tant esdeveniments actius (interacció del dispositiu amb xarxes mòbils) com passius (dades de sondes de xarxa que es registren en canviar d'antena). Això proporciona una alta granularitat temporal i una resolució espacial de cel·la de telefonia, amb una precisió de desenes o centenars de metres a les ciutats i de diversos quilòmetres a les zones rurals [26].

2. 2 Processament de les Dades

2. 2. 1 Càlcul de la Taxa d'Incidència

La incidència es defineix com el «nombre de casos nous d'una malaltia determinada que apareixen en un grup o una població durant un període de temps concret» [27]. Es tracta d'una mesura clau en epidemiologia perquè proporciona informació essencial sobre la propagació i el risc de la malaltia dins d'una població.

És preferible fer ús de la incidència enlloc del nombre de casos perquè és una manera d'estandarditzar les mesures i, en conseqüència, de poder realitzar comparacions significatives entre poblacions de diferents grandàries.

Per a calcular la taxa incidència, s'utilitza la fórmula de l'equació (1), on n és un factor de multiplicació que es fa servir per expressar la taxa en una forma més manejable.

$$Taxa\ d'incidència = \frac{\text{Nombre de casos durant un període}}{Població\ en\ risc\ durant\ el\ mateix\ periode} \times 10^n \quad (1)$$

En aquest estudi s'ha utilitzat un factor de multiplicació igual a 5, ja que és el que més s'adequa per a una millor comprensió de les dades i és el més comú per a treballar amb taxes d'incidència. Així doncs, la incidència s'expressa per a cada 100.000 habitants.

2. 2. 2 Estudi General de l'Evolució de la COVID-19 a Catalunya

Tot i que es parteix d'un conjunt de dades de malalties infeccioses respiratòries que van des del 2019 fins al 2023, el registre de casos de COVID-19 s'inicia aproximadament a finals del 2019. És per això que l'anàlisi i la interpretació dels resultats es realitzen del període comprés entre els anys 2020 i 2023.

Amb aquest estudi es pretén comprendre de manera completa el desenvolupament de la pandèmia de la COVID-19, amb una anàlisi desagregada per edat, regió sanitària, sexe i índex socioeconòmic, que ajuda a identificar-ne patrons en els diferents grups.

Per a dur a terme l'estudi, s'han realitzat gràfics de línies. Un primer gràfic amb la incidència total (per 100.00 habitants) al llarg del període especificat, i quatre gràfics més per mostrar la incidència desagregada per edat, regió sanitària, sexe i índex socioeconòmic.

2. 2. 3 Estudi de les Distribucions Estadístiques de les Mostres. Test Kolmogórov-Smirnov (K-S)

El test Kolmogórov-Smirnov és una prova estadística no paramètrica que permet determinar si dues mostres de probabilitat tenen la mateixa distribució estadística. S'utilitza per comparar o bé una mostra amb una distribució de probabilitat de referència (test K-S d'una mostra) o bé per comparar dues mostres (test K-S de dues mostres) [28]. En aquest estudi s'ha emprat el test K-S de dues mostres.

Els avantatges d'aquest test respecte a d'altres són que, en tractar-se d'una prova de tipus no paramètrica, no fa suposicions especifiques sobre la forma de la distribució de les dades, de manera que considera tota la distribució, incloent la forma i la posició.

El test K-S a Python es pot realitzar amb la funció 'ks_2samp()', la qual retorna el valor de l'estadístic K-S i el p-value. Internament, la funció realitza els següents passos:

1. Formulació de la hipòtesi. La hipòtesi nul·la és que ambdues mostres provenen de la mateixa distribució.
2. Càlcul de l'estadístic de la prova. Es calcula la funció de distribució acumulada empírica (ECDF) per a cadascuna de les mostres i es determina la distància màxima entre les ECDFs, anomenada estadístic K-S.
3. Determinació del valor crític o p-value. Mitjançant el càlcul del p-value, es determina la significació estadística de l'estadístic K-S.

En l'estudi s'han encarat les distribucions de les dades agregades per les diferents variables: regió sanitària, edat, sexe i ISC. Per a cada una de les quatre variables s'han realitzat múltiples tests K-S, comparant totes les combinacions possibles de valors que poden prendre. Així, s'ha fet possible deduir la probabilitat que cada parell de mostres provingués de la mateixa distribució.

Prèviament al test, s'han generat gràfics de densitat d'incidència per a cada variable. Es tracta d'una eina complementaria al test K-S, que proporciona una visualització intuïtiva i qualitativa de les distribucions, la qual cosa és de gran utilitat per a identificar diferències i similituds entre les distribucions i, així, començar a intuir els resultats del test K-S que es realitzarà posteriorment.

Per a preparar les dades per al test, s'han filtrat les mostres relatives al diagnòstic de COVID-19 i compreses solament l'any 2022 —s'ha optat per la selecció d'aquest únic any perquè és el que s'ha usat per a fer el model d'estimació de la incidència que es presentarà més endavant. Tot seguit, en primer lloc, s'ha treballat amb les dades agregades per regió sanitària, les quals s'han agrupat per data per obtenir per a cada setmana i cada regió el número de casos i de població. Posteriorment, a aquest DataFrame se li ha afegit una columna amb el valor del càlcul de la incidència (Figura 6) i, a continuació, s'han reorganitzat les dades en forma d'una taula pivot en què les files són dates, les columnes són regions i els valors són incidències (Figura 7).

	data_inici	nom_regio	diagnostic	casos	poblacio	incidencia
14084	2022-01-03	Alt Pirineu i Aran	COVID-19	1694	67200	2520.833333
14091	2022-01-03	Barcelona Ciutat	COVID-19	34767	1637358	2123.359705
14100	2022-01-03	Barcelona Metropolitana Nord	COVID-19	46370	2045466	2266.965083
14109	2022-01-03	Barcelona Metropolitana Sud	COVID-19	23320	1111700	2097.688225
14118	2022-01-03	Camp de Tarragona	COVID-19	10060	521398	1929.428191

Figura 6. Primeres 5 files del DataFrame amb les dades de COVID-19 del 2022 agrupades per data i regió sanitària.

nom_regio	Alt Pirineu i Aran	Barcelona Ciutat	Barcelona Metropolitana Nord	Barcelona Metropolitana Sud	Camp de Tarragona	Catalunya Central	Girona	Lleida	Penedès	Terres de l'Ebre
data_inici										
2022-01-03	2520.833333	2123.359705	2266.965083	2097.688225	1929.428191	2413.915255	2069.020907	2263.725237	1909.758929	2147.091453
2022-01-10	2783.835604	2425.949030	2773.668675	2518.141619	2434.972587	2820.720027	2598.948305	2800.651772	2370.909091	2521.741546
2022-01-17	2796.588727	2365.445241	3013.922139	2588.823904	2880.964283	3079.691935	2910.137856	3349.960476	2582.338135	2722.202916
2022-01-24	2023.991276	1649.133580	2074.219427	1822.605157	2435.796700	2193.472975	2022.446161	2772.082611	2004.761496	2149.625935
2022-01-31	1380.267632	960.370116	1204.230744	1071.189283	1452.316991	1242.608129	1116.658153	1542.712792	1177.462065	1245.630724

Figura 7. Primeres 5 files del DataFrame de la Fig. 6 reorganitzat en forma de taula pivot.

Amb les dades preparades, s'ha implementat un bucle per iterar sobre les combinacions de regions i així, per a cada parell, poder executar el test K-S. Cal mencionar que s'ha optat per realitzar la prova sense normalitzar prèviament les dades, ja que així es poden comparar les distribucions en la seva forma original i no s'altera la relació entre les mostres.

Els resultats de p-value i estadístic K-S obtinguts s'han emmagatzemat en un DataFrame. Això ha permès poder-los representar en forma de matriu, a través de la creació d'un objecte 'PairGrid()' de seaborn, en el qual en la diagonal superior s'han exposat els valors de p-value i en la diagonal inferior els valors estadístics. Així s'aconsegueix tenir una visió global dels resultats de cada test.

Tot aquest procediment efectuat per a la variable de regió sanitària s'ha dut a terme per a les altres tres variables d'edat, sexe i índex socioeconòmic.

2. 2. 4 Model de Regressió Lineal Múltiple (MLR)

La regressió lineal múltiple és «una tècnica estadística que utilitza varies variables explicatives per a predir el resultat d'una variable de resposta. L'objectiu de la MRL és modelitzar la relació lineal entre les variables explicatives (independents o predictors) i les variables de resposta (dependents)» [29].

Com s'ha mencionat, la regressió busca la relació entre variables i pretén determinar com un o múltiples fenòmens n'influencien un altre. Addicionalment, la regressió pot ser d'utilitat per a predir una resposta utilitzant un nou conjunt de predictors.

En aquest estudi, d'entre la varietat de mètodes de regressió disponibles, s'ha emprat la regressió lineal. Es tracta probablement d'una de les tècniques de regressió més importants i emprades, així com de les més senzilles, oferint una fàcil però eficient interpretació dels resultats. Es tracta d'una regressió lineal de tipus múltiple perquè fa servir més d'una variable independent.

L'equació 2 mostra la fórmula i càlcul de la regressió lineal múltiple:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \qquad (2)$$

$$on\ y = variable\ dependent, \quad x = variables\ independents, \quad p = nº\ de\ predictors,$$
$$\beta_0 = intercept\ (terme\ constant), \ \beta_p = coeficients\ dels\ predictors,$$
$$\epsilon = terme\ d'error\ (residu)$$

La regressió lineal calcula l'estimació dels coeficients de regressió, els quals indiquen el pes que té de cadascun dels predictors sobre el resultat de la variable dependent. Per altra

banda, un altre paràmetre resultant a tenir en compte és el coeficient de múltiple determinació (R^2), el qual és una mesura estadística que avalua quina proporció de la variació del resultat pot explicar-se amb la variació de les variables independents. En altres paraules, mesura la proximitat de les dades a la línia de regressió lineal ajustada [30], [31].

Per a realitzar el model de regressió lineal múltiple i la posterior predicció de valors futurs, s'han incorporat les dades de grip del 2022 juntament amb el diagnòstic de COVID-19. Aquesta integració permet comparar com cadascuna de les malalties s'ajusta al model de MLR, proporcionant una anàlisi detallada de com factors ambientals i de mobilitat influeixen en l'evolució de la incidència. A més, permet contraposar les prediccions de cada diagnòstic. Per a l'estudi, aquest enfocament amplia la comprensió de les dinàmiques epidemiològiques associades a cada patogen i aporta una visió més holística dels factors que influeixen en la transmissió de cada malaltia.

A continuació es detalla el procediment dut a terme per a la implementació de la regressió lineal múltiple, utilitzant totes les dades exposades en l'apartat 2.1.

2. 2. 4. 1 Selecció i Filtratge de les Dades Geogràfiques i d'Estacions Meteorològiques

Per tal de poder ajuntar els diferents conjunts de dades, primerament se va filtrar a través de les estacions meteorològiques i les dades ambientals de temperatura i humitat relativa. Del DataFrame inicial d'estacions de la XEMA, algunes d'elles han estat desmantellades, per la qual cosa només s'han conservat les que foren operatives durant el període d'interès, 2022-2023. Per altra banda, donada la possibilitat que d'algunes estacions no se'n tingueren dades meteorològiques, s'han mantingut únicament aquelles que coexistien en ambdós DataFrames. Com a resultat, el nombre d'estacions, que inicialment era de 239, es redueix a 187.

Considerant que cada ABS en la seva extensió territorial pot o bé no contenir cap estació, o bé contenir-ne una o múltiples, a continuació s'ha filtrat per a mantenir solament les ABS que en tenen almenys una. Per a aconseguir-ho, GeoPandas presenta la funció 'sjoin()' que realitza una unió espacial entre dos GeoDataFrames. És important tenir en compte que totes dues taules contenen una columna geometry amb coordenades, de manera que les ubicacions de les estacions són punts i les àrees geogràfiques de les ABS són polígons. En aquesta funció s'especifiquen els paràmetres 'inner' i 'within' per tal de buscar els punts que cauen dins d'algun dels polígons i mantenir únicament les ABS que tenen una o més estacions (Figura 8). Així, s'obté una llista de les ABS amb què es treballarà, en la qual s'ha passat de 379 a 116 ABS.

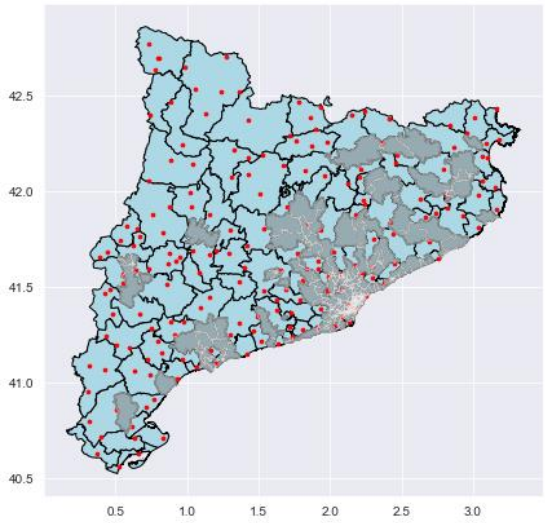


Figura 8. Representació geogràfica de les ABS i les estacions meteorològiques automàtiques. En color gris es troben marcades les ABS que no presenten cap estació compresa al seu territori.

En la figura presentada anteriorment s'observa que és sobretot a les zones amb una població més densa que hi ha més ABS i, com que les estacions estan repartides de manera més o menys uniforme per tot el territori català, és aquí on hi ha menys àrees bàsiques de salut que continguin una o més estacions.

2. 2. 4. 2 Gestió i Agregació de les Dades Meteorològiques per Àrea Bàsica De Salut

En aquest punt, el següent pas ha estat gestionar les ABS que contenen en el seu territori més d'una estació de la XEMA. Ha fet falta identificar quines estacions es troben en cadascuna de les ABS per tal de poder tenir un únic valor de temperatura i d'humitat relativa per a cada data, i per a fer-ho s'ha creat un DataFrame en què cada fila representa una ABS amb una llista de codis de les estacions associades.

Una vegada es va conèixer la relació entre les ABS i les estacions, a través d'un diccionari de mapeig entre els codis de les estacions i els noms de les ABS, es va afegir una nova columna al DataFrame de dades ambientals mapejant cada codi d'estació al seu corresponent nom d'ABS. D'aquesta manera, ja va ser possible agrupar les dades per ABS i data, guardant la mitjana dels valors de temperatura i d'humitat relativa per a cada registre.

Addicionalment, donat que per a realitzar el model d'estimació d'incidència s'usaran com a variables d'entrada o independents les corresponents a les dades ambientals i de mobilitat de les setmanes anteriors, al DataFrame s'han afegit noves columnes amb aquests valors. És a dir, amb l'ús d'un bucle, s'ha dut a terme el procés de creació de variables retardades o 'lagging' (amb la funció shift(i)), amb el que s'han agregat 10 noves columnes, amb els valors de temperatura i d'humitat de les cinc setmanes anteriors (Figura 9).

	NOMABS	DATA_LECTURA	temp_mitjana	hum_mitjana	temp_k-1	hum_k-1	temp_k-2	hum_k-2	temp_k-3	hum_k-3	temp_k-4	hum_k-4	temp_k-5	hum_k-5
0	Alcarràs	2021-12-27	7.872537	89.173134	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Alcarràs	2022-01-03	5.280952	79.709821	7.872537	89.173134	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Alcarràs	2022-01-10	8.722917	76.069444	5.280952	79.709821	7.872537	89.173134	NaN	NaN	NaN	NaN	NaN	NaN
3	Alcarràs	2022-01-17	0.432292	82.177083	8.722917	76.069444	5.280952	79.709821	7.872537	89.173134	NaN	NaN	NaN	NaN
4	Alcarràs	2022-01-24	1.983631	79.474702	0.432292	82.177083	8.722917	76.069444	5.280952	79.709821	7.872537	89.173134	NaN	NaN

Figura 9. Primeres 5 files del DataFrame que conté les dades de temperatura i humitat de la setmana actual i de les cinc anteriors per a cada ABS i data únics.

2. 2. 4. 3 Regressió Lineal Múltiple (MLR)

Dues de les implementacions de models de regressió lineal que més s'utilitzen a Python són les que s'importen de les llibreries de codi obert de 'scikit-learn' i 'statsmodels'. En aquest estudi s'ha optat per la segona opció, ja que scikit-learn està principalment orientada a la predicció, per la qual cosa pràcticament no disposa de funcionalitats que mostrin resultats més detallats. En canvi, statsmodels és més complet i mostra els paràmetres estadístics avançats del model.

El següent procediment s'ha dut a terme tant amb les dades de grip com amb les de COVID-19 i és d'aquestes últimes que se'n detallarà el procés. Per a acabar de preparar les dades per al model, s'ha partit de tres DataFrames. El primer és el que conté les dades de Vigilància Sindròmica d'Infeccions del diagnòstic de COVID-19 de l'any 2022, el segon és l'obtingut en l'apartat 2.2.4.2, que conté les dades ambientals per a cada data i ABS únics, i el tercer és el de les dades de mobilitat. Tots ells s'han combinat, de manera que s'han realitzat dues fusions ('merge()') on les columnes en comú dels dos primers DF eren les de data i ABS, i la columna en comú amb el tercer DF era únicament la columna de data, ja que les dades de mobilitat són de tot Catalunya i no estan desagregades.

S'han desenvolupat cinc models de MLR, cadascun amb unes variables independents diferents. Aquestes sis variables en cada model tenen un valor de retard diferent, que varia des de k-1 a k-5. Això vol dir que s'integren en el model mostres de fins a cinc setmanes anteriors per analitzar la seva influència en la setmana actual. Per a aconseguir-ho, s'ha anat iterant entre els diferents valors de retard mitjançant un bucle.

Abans de començar amb les iteracions, s'han creat dues variables, una llista i un DataFrame, per a anar emmagatzemant els valors de R^2 i els coeficients del model, respectivament.

Considerant que el bucle s'itera cinc vegades, a continuació es proporcionarà una explicació detallada de la iteració k-1.

Abans d'ajustar el model, es realitza un preprocessament de les dades combinades dels tres DataFrames. En aquesta fase, s'eliminen els valors nuls i infinits per assegurar la integritat de les dades i evitar interferències en les anàlisis posteriors. Les mostres restants, amb l'objectiu de calcular-ne la incidència, s'agrupen per data, ABS, temperatura, humitat i viatges amb un retard de k-1 setmanes, i se suma el nombre de casos i població per a cada grup. Posteriorment, es calcula la incidència per cada 100.000 habitants i s'ordenen les dades per nom d'ABS i data. Es crea una nova columna amb el càlcul de la incidència amb retard k-i setmanes, i es tornen a eliminar els registres nuls. Això és necessari perquè el primer valors d'incidència k-1 per a cada ABS seran NaN, ja que el DataFrame no conté registres anteriors a la primera data de cada ABS.

Inicialment, es va intentar realitzar el test amb aquestes dades preprocessades, però el model no s'ajustava correctament. A causa de la presència de valors extrems en les variables d'incidència i incidència k-1, es va optar per calcular el logaritme natural d'ambdues. Prèviament, es va comprovar que no hi hagués cap valor de zero en les dades, ja que pot causar problemes a l'hora de calcular el logaritme.

Una vegada realitzat aquest procediment, el resultat és un DataFrame de 5704 files (Figura 10), què és també el nombre d'observacions que s'introdueixen al model de MLR. Aquest nombre d'observacions permet realitzar una anàlisi robusta.

	data	nom_abs	temp_k-1	hum_k-1	viatges_k-1_Interior	viatges_k-1_Ext: entrades	viatges_k-1_Ext: sortides	incidencia	incidencia_k-1	log_incidencia_k-1	log_incidencia
115	2022-01-10	Alcarràs	5.280952	79.709821	1.315367e+08	586271.281	533750.440	2545.978589	2194.729853	7.693814	7.842270
230	2022-01-17	Alcarràs	8.722917	76.069444	1.454129e+08	496421.585	496897.738	3541.855125	2545.978589	7.842270	8.172406
345	2022-01-24	Alcarràs	0.432292	82.177083	1.455062e+08	509650.166	518216.285	3026.351908	3541.855125	8.172406	8.015113
460	2022-01-31	Alcarràs	1.983631	79.474702	1.491442e+08	548118.527	552347.127	1369.768722	3026.351908	8.015113	7.222397
575	2022-02-07	Alcarràs	7.572768	68.599702	1.527948e+08	572017.457	582991.939	678.457082	1369.768722	7.222397	6.519821
...	...	...	...	...	...	...	...	...	...	...	...
5333	2022-11-21	Vilassar de Dalt	15.446131	69.074405	1.560854e+08	645613.814	650717.264	345.493344	371.385622	5.917241	5.844973
5447	2022-11-28	Vilassar de Dalt	12.857500	71.608333	1.561366e+08	640896.540	645895.570	368.284229	345.493344	5.844973	5.908855
5594	2022-12-12	Vilassar de Dalt	11.350000	83.008333	1.364091e+08	723167.564	724518.440	216.149452	368.284229	5.908855	5.375970
5709	2022-12-19	Vilassar de Dalt	11.546131	87.404762	1.575892e+08	640073.150	643087.326	237.139730	216.149452	5.375970	5.468650
5818	2022-12-26	Vilassar de Dalt	14.339286	74.452381	1.549196e+08	629308.221	752148.416	203.252033	237.139730	5.468650	5.314447

Figura 10. Resum del DataFrame de les dades de COVID-19. Hi ha la variable dependent i variables independents que s'usen per ajustar el model de regressió lineal múltiple.

De l'anterior figura es pot puntualitzar que els valors de temperatura, humitat, log_incidència i log_incidència_k-1 son únics per a cada ABS i data, mentre que les tres categories de viatges, per a cada data, prenen el mateix valor per a totes les ABS.

Amb la finalitat d'ajustar el model, s'han definit les variables X i Y. El conjunt de variables X, o variables independents, està constituït per les columnes de temperatura, humitat relativa, log_incidencia, viatges interiors, viatges exteriors d'entrada i viatges exteriors de sortida amb un retard d'una setmana. Per altra banda, la variable Y, o variable dependent, es correspon al logaritme de la incidència de la setmana actual. Amb aquestes assignacions, l'equació de regressió lineal múltiple pren la següent forma:

$$y^k = \beta_0 + \beta_1 {x_1}^{(k-1)} + \beta_2 {x_2}^{(k-1)} + \beta_3 {x_3}^{(k-1)} + \beta_4 {x_4}^{(k-1)} + \beta_5 {x_5}^{(k-1)} + \beta_6 {x_6}^{(k-1)} + \epsilon$$

$$\begin{gathered} on\ y^k = variable\ dependent - log.de\ la\ incidència\ en\ el\ temps\ k \\ {x_1}^{(k-1)} = temperatura\ amb\ un\ retard\ d'una\ setmana \\ {x_2}^{(k-1)} = humitat\ relativa\ amb\ un\ retard\ d'una\ setmana \\ {x_3}^{(k-1)} = log.de\ la\ incidència\ amb\ un\ retard\ d'una\ setmana \\ {x_4}^{(k-1)} = viatges\ interiors\ amb\ un\ retard\ d'una\ setmana \\ {x_5}^{(k-1)} = viatges\ exteriors\ (entrades)\ amb\ un\ retard\ d'una\ setmana \\ {x_6}^{(k-1)} = viatges\ exteriors\ (sortides)\ amb\ un\ retard\ d'una\ setmana \\ \beta_0 = intercepte\ (terme\ constant) \\ \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 = coeficients\ per\ a\ cada\ predictor \\ \epsilon = terme\ d'error\ (residu) \end{gathered} \quad (3)$$

Posteriorment, les variables independents van ser estandarditzades. Aquest pas es va realitzar per contrarestar les diferències d'escala entre variables, ja que poden influir de manera desigual en el model de regressió. L'estandardització va ajustar les variables per tenir una mitjana de zero i una desviació estàndard d'una unitat.

El següent pas ha estat afegir una constant a les variables independents estandarditzades, necessari per a incloure l'intercept en els càlculs del model de MLR. Finalment, s'ha ajustat el model de regressió lineal múltiple utilitzant les variables independents estandarditzades i la variable dependent. El model en qüestió és el dels mínims quadrats ordinaris (OLS), de la llibreria statsmodels.

Dels resultats de cada iteració, se n'ha emmagatzemant el coeficient de determinació i els coeficients calculats per a cada variable independent, incloent la constant.

2. 2. 4. 4 Representació Geogràfica dels Resultats del Model de Regressió Lineal Múltiple

De manera anàloga a l'apartat anterior, aquesta secció se centra en el procediment per al diagnòstic de COVID-19, però s'ha realitzat també per a la grip. Per a representar els resultats del coeficient de determinació i dels coeficients de cada variable independent sobre el mapa, en primer lloc ha estat necessari calcular aquests valors per a cada ABS i setmana de retard úniques. És per aquest motiu, que s'ha alterat lleugerament el procediment que s'incorpora dins al bucle.

S'ha seguit la mateixa metodologia descrita en l'apartat 2.2.4.3, fins a obtenir el DataFrame representat en la Figura 10. A continuació, s'ha implementat un nou bucle per a cada àrea bàsica de salut, iterant entre les diferents ABS per a omplir un DataFrame amb les seves dades corresponents i, posteriorment, ajustar el model de regressió lineal múltiple. Per assegurar la robustesa dels resultats, totes les MLR realitzades comptaven, com a mínim, amb 30 mostres. En el cas de la grip, cal puntualitzar que dues de les ABS (La Granadella i Artesa de Segre) no presentaven prou mostres per a dur a terme el test i s'han descartat. Els resultats obtinguts s'han anat recopilant en un DataFrame (Figura 11), on figura el nom de l'ABS, el nombre de setmanes de retard temporal, el valor de R^2 i de R^2 ajustada, l'intercept i els coeficients de les sis variables independents.

	NOMABS	k	R2	R2_adj	Intercept	Coef_temp	Coef_hum	Coef_incidencia	Coef_viatges_interior	Coef_viatges_ext:entrades	Coef_viatges_ext:sortides
0	Alcarràs	1	0.793346	0.764511	6.089952	0.022741	-0.101897	0.473434	-0.094024	-0.026177	-0.112291
1	Alfarràs - Almenar	1	0.743596	0.707818	6.481635	-0.031255	-0.063940	0.379094	-0.052938	-0.043044	-0.014503
2	Almacelles	1	0.765322	0.731796	6.447587	0.062898	0.008337	0.412403	-0.037928	-0.130253	0.016756
3	Alt Berguedà	1	0.412064	0.330026	6.670290	-0.000321	-0.060575	0.133601	-0.073677	0.041569	-0.160964
4	Alt Camp Est	1	0.793658	0.765520	6.206111	0.005571	-0.134851	0.417762	-0.042662	-0.004656	-0.137869
...	...	...	...	...	...	...	...	...	...	...	...
570	Vic - 2 Sud	5	0.619272	0.562163	5.463852	-0.068375	-0.035369	0.234645	0.079082	-0.037330	0.071546
571	Viladecans - 1	5	0.743888	0.705471	5.534974	-0.049973	-0.006661	0.370772	0.040952	-0.088480	0.088726
572	Vilafant	5	0.377451	0.284068	5.852879	0.007465	-0.051783	0.142432	0.000447	0.021974	-0.138754
573	Vilafranca del Penedès - 2	5	0.634101	0.579216	5.634297	-0.161008	0.030408	0.323483	0.075344	0.070163	0.059681
574	Vilassar de Dalt	5	0.732103	0.691919	5.944638	-0.011746	-0.009447	0.340011	0.032725	-0.133037	0.103277

Figura 11. Resum del DataFrame dels resultats del model de regressió lineal múltiple per a cada ABS i setmana de retard per al diagnòstic de COVID-19.

Una vegada obtingut el DataFrame amb les dades (Figura 11), s'ha implementat un bucle per a iterar entre les diferents setmanes de retard temporal per a la representació geogràfica del coeficient de determinació. En cada iteració, les dades corresponents a la setmana en qüestió s'han guardat en un DataFrame, el qual s'ha combinat amb el DataFrame de les coordenades geogràfiques de les ABS. Així, cridant la funció 'plot()' sobre el DataFrame resultant, s'ha generat la figura.

Amb relació a la representació geogràfica dels coeficients, en lloc de generar 30 figures, corresponents als sis coeficients per a cadascuna de les cinc setmanes de retard temporal, s'ha trobat més oportú representar únicament aquells coeficients del model que ha obtingut un valor de R^2 més elevat en l'apartat 2.2.4.3. D'aquesta manera, es posa atenció en els resultats del model que millor s'ha ajustat als punts de dades observats.

2. 2. 4. 5 Predicció de valors a partir del Model de Regressió Lineal Múltiple

En aquesta secció, s'ha fet ús el model de regressió lineal múltiple per a realitzar i analitzar les prediccions sobre dades noves. En primer lloc, donat que les prediccions que s'han fet són de l'any 2023, s'ha dut a terme el procés de preprocessat de les dades. Aquest coincideix amb el descrit en l'apartat 2.2.4.3, on s'obté un DataFrame com el de la Figura 10, però en aquest cas amb les dades del 2023.

Per a realitzar una primera predicció, s'ha seleccionat una àrea bàsica de salut, concretament la de 'Tarragona - 2', i s'han predit els valors d'incidència per als diagnòstics de COVID-19 i grip per a les tres primeres setmanes del 2023.

Inicialment, per a obtenir les prediccions, s'ha filtrat el conjunt de dades preprocessat per seleccionar les files corresponents a l'ABS i dates desitjades. A continuació, s'han utilitzat les variables predictores o independents, les quals s'han estandarditzat per garantir la consistència amb el procés d'ajustament del model. S'ha afegit una constant al conjunt de dades estandarditzades i, mitjançant la funció 'predict()' associada al model de MLR, s'han calculat les prediccions.

És important destacar que les prediccions s'obtenen en escala logarítmica. Per facilitar-ne la interpretació, s'ha aplicat la funció exponencial als valors predits per retornar-los a l'escala original.

Aquest procediment s'ha realitzat tant per al diagnòstic de COVID-19 com de grip. Finalment, amb l'objectiu de comparar les prediccions amb les observacions reals, els valors s'han representat en un gràfic de línies.

A més a més, a banda de predir els valors de l'ABS de 'Tarragona - 2', s'han predit també les incidències de la resta d'ABS per a les tres primeres setmanes. En aquest cas, el procediment segueix el mateix format descrit anteriorment, però incorporat dins d'un bucle que itera per cada ABS. Addicionalment, per a avaluar l'efectivitat del model predictiu i veure en quina proporció s'acosten les prediccions als valors reals observats, s'ha utilitzat el càlcul d'error relatiu (equació 3). S'usa com a mesura de precisió i es defineix com a la relació entre l'error i el valor real observat, sense aplicar el valor absolut per mantenir el signe i identificar si la predicció és major o menor que l'observació. És important destacar que, tot i això, com cap valor d'incidència és negatiu, no es dona l'efecte de compensació d'errors de predicció positius i negatius, la qual cosa és crucial per obtenir resultats precisos. Tots els valors d'error s'han emmagatzemat en un DataFrame per una posterior anàlisi.

$$E_r = \frac{\left(y_{observació} - y_{predicció}\right)}{y_{observació}} \times 100 \quad (3)$$

Per a cada setmana, s'han representat gràficament amb punts les prediccions en comparació amb les observacions reals. Així, s'ha visualitzat la comparació directa entre ambdós valors de totes les ABS.

Amb els valors de l'error relatiu, s'han creat tres mapes, un per a cada data, que mostren el valor de l'error per a cada ABS. Això permet identificar per a quines ABS les prediccions han estat majors, menors o iguals als valors reals.

Tant els gràfics de punts com els mapes s'han realitzat per als diagnòstic de COVID-19 i de grip.

3 Resultats i discussió

3. 1 Estudi General de l'Evolució de la COVID-19 a Catalunya

Per a tenir una visió global de sobre la distribució i estructura de les dades, primerament s'ha realitzat un estudi general de l'evolució de la COVID-19 durant el període comprés entre els anys 2020-2023. A través d'aquesta anàlisi es pretén observar la progressió del virus des de diferents perspectives, examinant l'evolució general de la incidència, així la seva distribució per regions sanitàries, edat, sexe i ISC.

3. 1. 1 Evolució de la Incidència

En primer lloc, es va estudiar l'evolució temporal de la incidència de la COVID-19 per cada 100.000 habitants. Es poden observar les fluctuacions de la incidència al llarg del temps (Figura 12), amb un pic significatiu i fàcilment detectable a finals del 2021 – principis del 2022. En analitzar més detingudament aquest període amb alts nivells d'incidència, es troba que coincideix amb la 6a onada epidèmica a Espanya, la qual, oficialment, es va iniciar el 14 d'octubre del 2021 i es va donar per finalitzada el 27 de març del 2022, una informació que encaixa amb les dades mostrades. Els casos de COVID-19 es van incrementar de manera molt notable durant aquest interval, inicialment degut a la variant Delta, però fonamentalment sota l'efecte de la variant Ómicron que es caracteritza per la seva alta transmissibilitat [32].

Per altra banda, els diferents decreixements que es veuen en el transcurs del temps podrien ser a causa de les diferents mesures de confinament i distanciament social, així com de l'ús de mascaretes, la campanya massiva de vacunació, l'estacionalitat o el tancament de llocs públics i la cancel·lació d'esdeveniments.

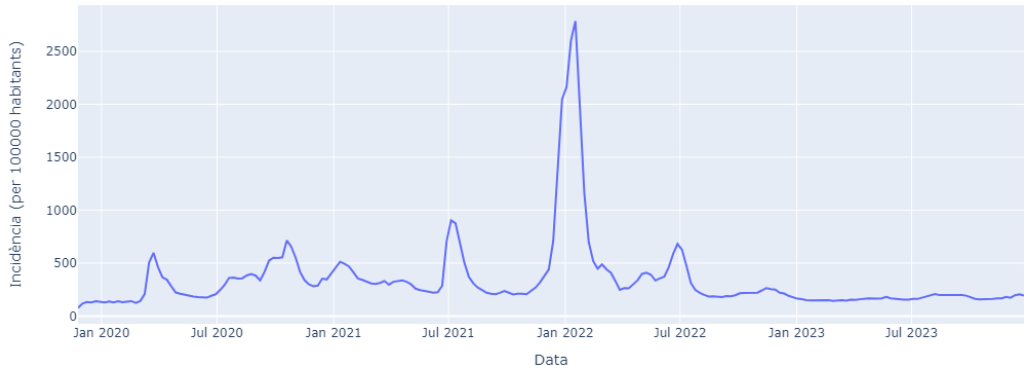


Figura 12. Taxa de diagnòstics clínics de COVID-19 per 100.000 habitants des de l'any 2020 fins al 2023.

3. 1. 2 Evolució de la Incidència per Regions Sanitàries

El territori català està dividit en 10 regions sanitàries oficials i la visualització de l'evolució de la incidència al llarg del temps per a cadascuna d'elles (Figura 13) permet entendre com la pandèmia de la COVID-19 ha afectat diferents regions de manera desigual.

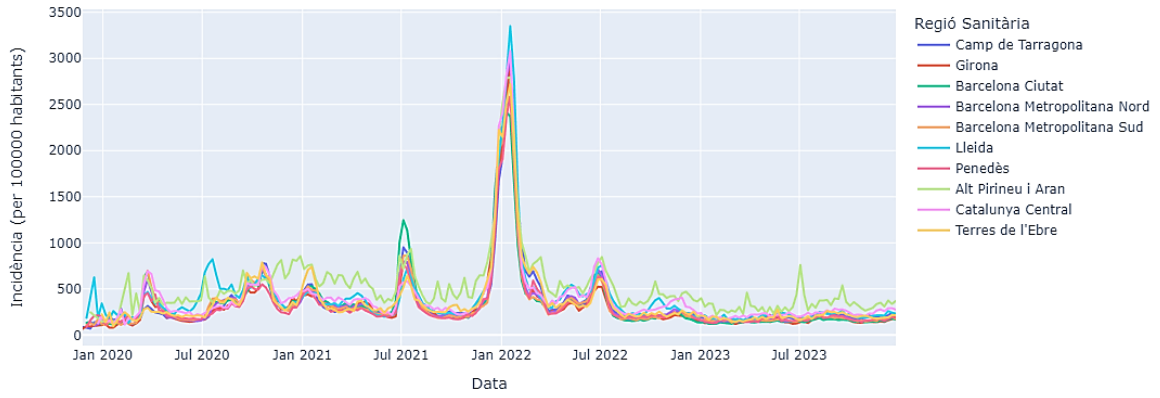


Figura 13. Taxa de diagnòstics clínics de COVID-19 per 100.000 habitants des de l'any 2020 fins al 2023 desagregats per regió sanitària catalana.

La Figura 13 mostra que, en general, la distribució de la incidència de les regions sanitàries és bastant similar entre elles. És a dir, els pics i les davallades tendeixen a coincidir, indicant que la major part de les vegades les regions han experimentat de manera sincronitzada les onades de la pandèmia. Cal mencionar que la regió representada en verd, que correspon a l'Alt Pirineu i Aran, és la que sobresurt i es desmarca una mica més de la resta. Això probablement es deu al fet que per a aquesta regió es contempla un nombre de població i de casos menor, i, una mínima variació en el nombre d'afectats per la malaltia, produeix un gran canvi en el valor de la incidència.

S'ha decidit representar la incidència de COVID-19 per regions sanitàries i no incloure la distribució per àrees bàsiques de salut perquè, donat que hi ha unes 380 ABS, el resultat il·lustrat al llarg del temps no seria clar ja que la gran quantitat d'àrees generaria un gràfic excessivament carregat i difícil d'interpretar.

3. 1. 3 Evolució de la Incidència per Grups d'Edat

En aquest apartat s'analitza com ha evolucionat la incidència de la COVID-19 en diferents grups d'edat a través de la visualització d'un gràfic de línies (Figura 14).

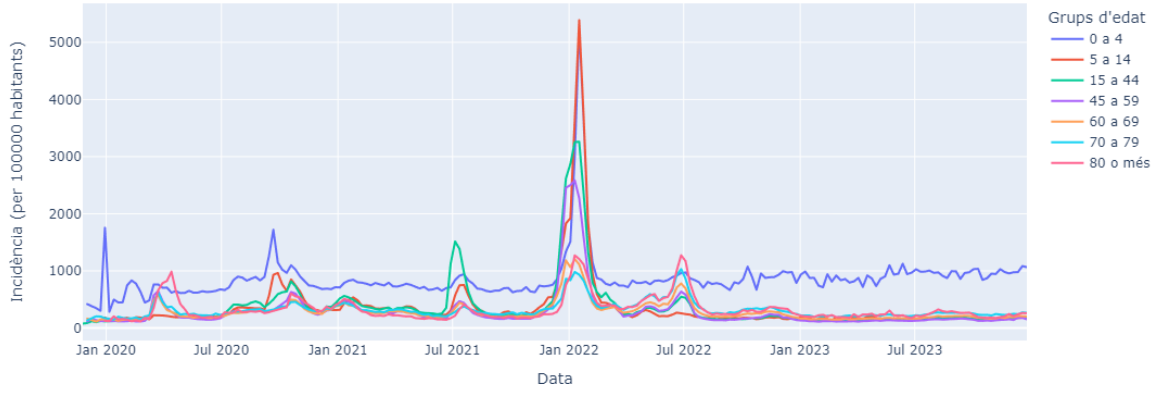


Figura 14. Taxa de diagnòstics clínics de COVID-19 per 100.000 habitants des de l'any 2020 fins al 2023 desagregats per grups d'edat.

Com en el cas de la incidència per regions sanitàries, els diferents grups d'edat també presenten, en línies generals, una distribució pareguda entre ells. Tot i això, és adient esmentar que hi ha un grup que pràcticament durant tota la línia temporal registrada té un

valor d'incidència superior a la resta de grups. Es tracta d'aquells diagnosticats de COVID-19 d'una edat entre 0 a 4 anys. Donat que l'aspecte d'aquesta línia en concret, en ressaltar per sobre de tota la resta de manera constant, no semblava que fos correcte, es va comparar el resultat obtingut amb el que s'ofereix a la pàgina web del SIVIC. Allà es pot visualitzar aquesta mateixa representació realitzada a partir de les dades que s'han usat per a aquest estudi. En equiparar ambdós resultats, en la gràfica del SIVIC la línia del grup d'edat de 0 a 4 anys seguia una distribució homogènia a la resta de grups. En prendre consciència d'aquesta divergència, va provar-se de contactar amb la Generalitat de Catalunya, ja que és qui facilita les dades, a través de la bústia de contacte del portal *Govern obert*. Després de descriure la situació, textualment es va demanar: «Exposat això, em preguntava si em podríeu ajudar a entendre a què es podrien deure aquestes diferències. Pot ser que hi hagi algun detall en les dades o en el procés que se m'hagi passat per alt, o que vosaltres realitzeu algun processat extra?». Malauradament, fins al dia d'escriptura d'aquest treball no s'ha rebut resposta a la consulta, de manera que no se n'ha pogut fer cap altra interpretació.

A banda del que ja s'ha exposat, de la figura es podria destacar que, durant el període representat, els altres tres grups que sobresurten en els pics a la resta de grups serien el de 5 a 14 anys —al setembre del 2020 i gener del 2022—, el de 15 a 44 —al juliol del 2021— i el de 80 o més anys —a l'abril del 2020 i al juliol del 2022—, però de manera més lleugera que el grup de 0 a 4 anys. Aquests resultats coincideixen amb els que mostra el portal del Centre Nacional d'Epidemiologia de l'Institut de Salut Carlos III d'Espanya [33].

3. 1. 4 Evolució de la Incidència per Sexe

En relació a l'evolució de la incidència de la COVID-19 segons el sexe, a Catalunya hi ha hagut lleugerament més infeccions en les dones que en els homes (Figura 15). Aquesta diferència es fa especialment evident en els diversos pics d'incidència que s'han donat al llarg del període representat.

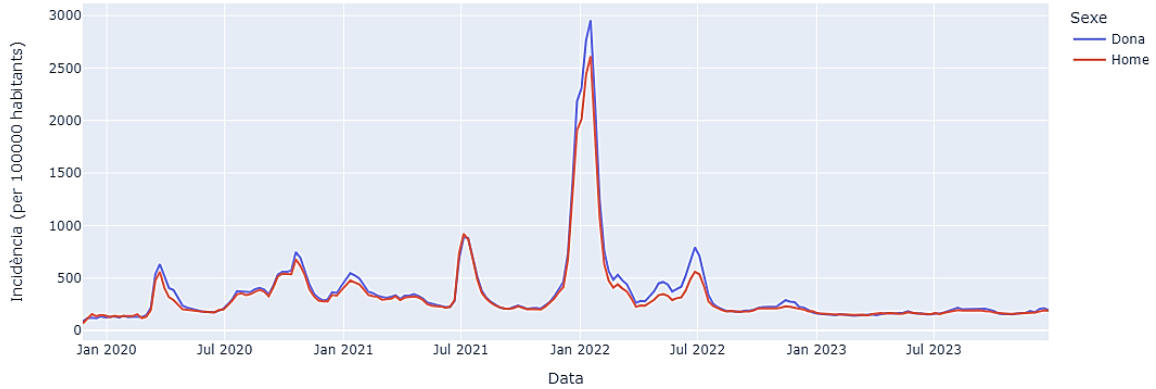


Figura 15. Taxa de diagnòstics clínics de COVID-19 per 100.000 habitants des de l'any 2020 fins al 2023 desagregats per sexe.

3. 1. 5 Evolució de la Incidència per Índex Socioeconòmic (ISC)

El gràfic de la Figura 16 mostra la incidència de COVID-19 desglossada per índex socioeconòmic (ISC). S'observa que, al llarg del temps, els pics d'incidència coincideixen en gran proporció entre els diferents nivells socioeconòmics, indicant una generalitzada

afectació de les onades de contagi. No obstant això, els resultats apunten diferències importants en funció de la onada en que es posa l'atenció.

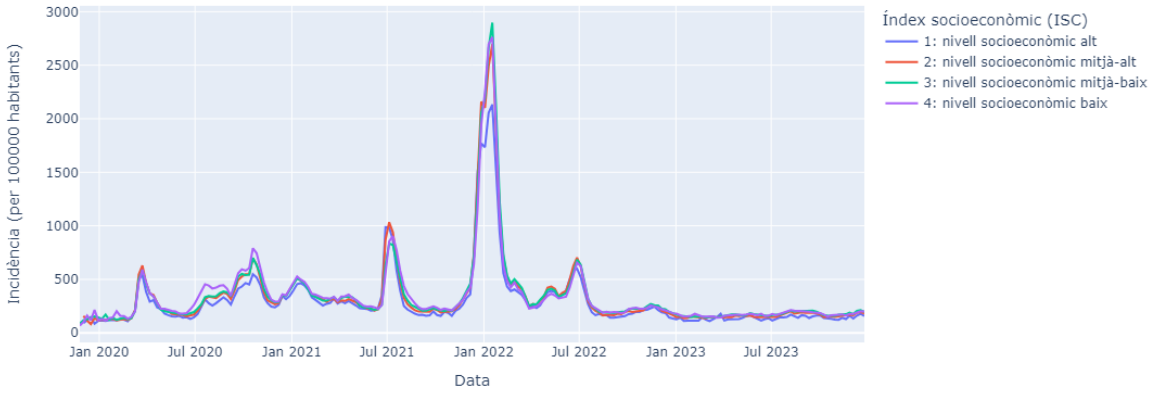


Figura 16. Taxa de diagnòstics clínics de COVID-19 per 100.000 habitants des de l'any 2020 fins al 2023 desagregats per índex socioeconòmic (1=nivell alt; 2=nivell mitjà-alt; 3=nivell mitjà-baix; 4=nivell baix)..

D'entre les diferents onades, en la segona (octubre-desembre 2020), en la cinquena (juny-setembre 2021) i en la sisena (octubre 2021 - abril 2022) és quan més s'evidencia la diferència entre els grups amb un nivell de major o de menor privació. Es mostra que en la segona i sisena onada la privació tenia una relació positiva amb la incidència, mentre que en la cinquena onada aquesta relació es torna negativa. Tals resultats coincideixen en gran part amb els presentats per E. López-Bazo (2024) [34], amb la discrepància que les dades d'incidència de la sisena onada mostrades en l'article tenen una relació negativa amb la privació. Posant èmfasi en els resultats de la segona onada, en línia amb E. Roel et al. (2022) [35], els valors pre-vacunació es podrien associar a què «those with low socioeconomic status are more likely to be exposed to infection because of poorer working and housing conditions and to develop severe disease because of poorer health status».

És alhora interessant posar atenció a treballs com el de M. A. Barceló et al. (2024) [36], que demostren que els individus amb un nivell socioeconòmic més baix tenien les taxes de vacunació més baixes, una manca en la cobertura vacunal que podria haver influït directament en el risc de contreure la malaltia. Malgrat això, tant E. López-Bazo com E. Roel et al. mencionen que la vacunació massiva a Catalunya[1] podria haver moderat l'efecte de la privació socioeconòmica sobre la incidència de la COVID-19, una afirmació que concordaria bastant amb els resultats de la Figura 16.

Les diferents discussions evidencien la complexitat de la relació entre l'índex socioeconòmic i la COVID-19.

3. 2 Estudi de les Distribucions Estadístiques de les Mostres. Test Kolmogórov-Smirnov (K-S)

El test Kolmogórov-Smirnov (K-S) s'ha realitzat per a les quatre variables presents en el conjunt de dades: regió sanitària, edat, sexe i índex socioeconòmic. L'objectiu del test és acabar deduint si les diferents distribucions són estadísticament iguals.

[1] La campanya de vacunació massiva a Catalunya es considera que va començar l'1 de gener del 2021.

3. 2. 1 Test K-S de les Mostres Agregades per Regió Sanitària

En la Figura 17 es mostra la distribució de densitat d'incidència, en forma de gràfic KDE (Kernel Density Estimation), de les diferents regions sanitàries. Totes les regions mostren una distribució de tipus asimètrica positiva, o també anomenada asimetria a la dreta, ja que la cua dreta és més extensa que l'esquerra. Això indica que la major part de les dades estan concentrades a l'esquerra de la mitjana, amb alguns valors dispersos a la dreta. També es pot apreciar que tenen dos pics, un amb un valor deu vegades major que l'altre, un aspecte que indica que s'està parlant de distribucions bimodals.

Així doncs, en general, a primera vista, no sembla que les distribucions de les diferents regions siguin prou similars per a concloure que totes provenen d'una mateixa població. Algunes localitats com Camp de Tarragona i Barcelona Met. Nord són més similars entre elles, però d'altres com Alt Pirineu i Aran s'allunyen bastant de la distribució de la resta.

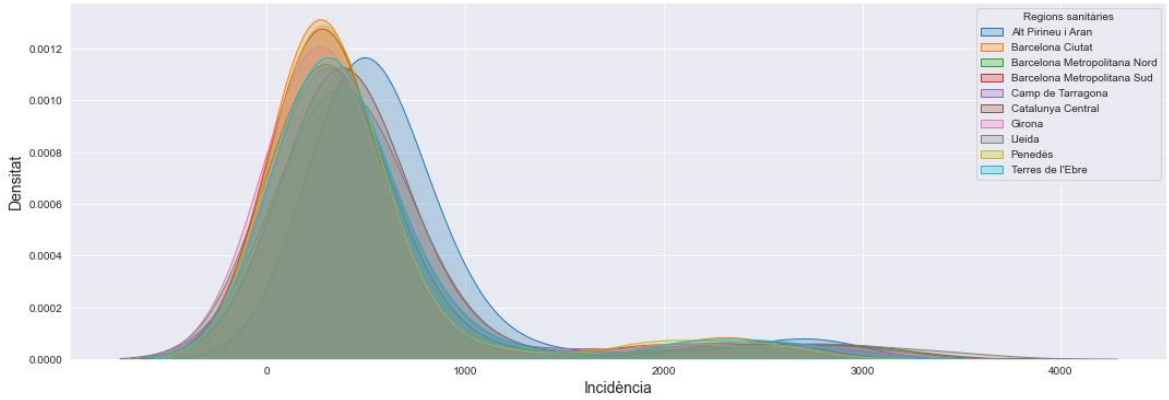


Figura 17. Densitat d'incidència de COVID-19 de l'any 2022 per regió sanitària.

Els resultats del test Kolmógorov-Smirnov (Figura 18) evidencien i reafirmen el que se suposava en veure les distribucions de densitat. De les 45 combinacions de regions a què se'ls ha realitzat el test, 18 d'elles tenen un p-value major a 0.05. Això indica que en aquests 18 tests es confirma la hipòtesi nul·la, la qual suggereix que les dues distribucions coincideixen. Per altra banda, en referència a l'estadístic, aquells tests als quals es demostra la hipòtesi nul·la, el valor de la distància K-S és menor en comparació a la resta. Una menor distància màxima entre les funcions de distribució acumulada de les mostres significa una major semblança entre ambdues.

Globalment, els resultats dels tests K-S realitzats indiquen que les dades de les deu regions no provindrien del mateix grup poblacional. Sí que és veritat que hi ha regions com Barcelona Met. Nord o el Penedès que tenen una distribució semblant a 6/9 regions, però, per contrapartida, la regió de l'Alt Pirineu i Aran té una distribució diferent a totes les altres regions.

Figura 18. Matriu de resultats del test Kolmogórov-Smirnov per a les diferents combinacions de regions sanitàries. A la diagonal superior es troben els valors de p-value, acolorits en verd quan són majors de 0.05, que és el valor de significança establert, i en taronja quan són iguals o menors a 0.05. A la diagonal inferior es mostren els valors de l'estadístic i les caselles estan acolorides de manera simètrica als valors de p-value.

3. 2. 2 Test K-S de les Mostres Agregades per Grup d'Edat

La Figura 19 presenta la densitat d'incidència per als diferents grups d'edat. Es pot observar que, en comparació amb el gràfic de les regions sanitàries, aquestes distribucions són més diferents entre elles. Es tracta de distribucions asimètriques positives i bimodals, de manera que la major part de les dades s'acaparen a l'esquerra de la mitjana mentre que a la dreta hi ha pocs valors però molt extrems. També cal destacar que hi ha distribucions, com les dels grups d'edat de 60 a 69 o de 70 a 79 anys, que tenen un pic més elevat i estret, un aspecte que indica que una major proporció de les dades es concentra al voltant de la moda, suggerint una menor variabilitat. Altres distribucions com la del grup de 45 a 59 anys presenten un pic més baix i ample, indicant una major dispersió o variabilitat.

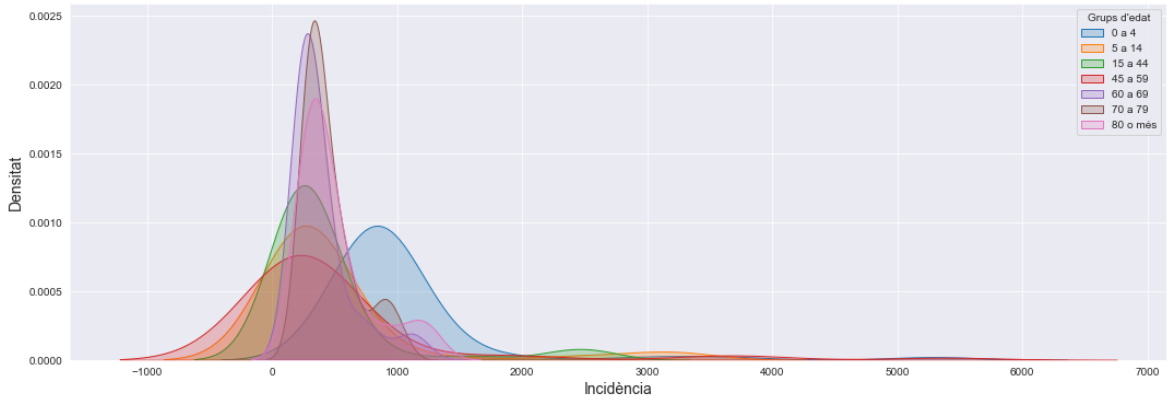


Figura 19. Densitat d'incidència de COVID-19 de l'any 2022 per grups d'edat.

Amb les observacions presentades, no sembla que en general els diferents grups d'edat provinguin d'una mateixa distribució, i això mateix exposen els resultats del test K-S (Figura 20). Només confirmen la hipòtesi nul·la, per una banda, els grups de 5 a 14, de 15 a 44 i de 45 a 59 i, per altra banda, els grups de 70 a 79 i de 80 o més, que entre ells mostren una distribució pareguda. Aquests amb un p-value major a 0.05 són també els que tenen un valor d'estadístic menor.

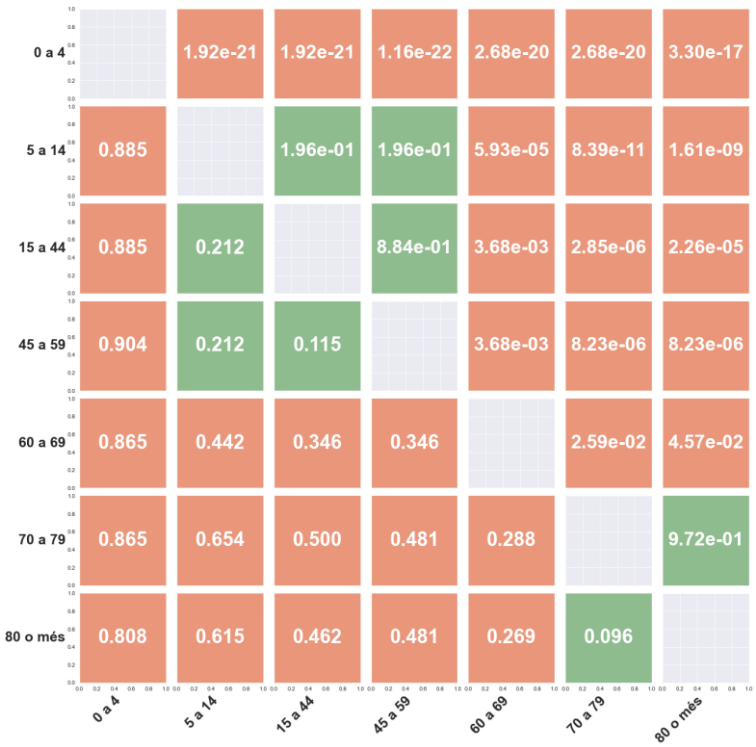


Figura 20. Matriu de resultats del test Kolmogórov-Smirnov per a les diferents combinacions de grups d'edat. A la diagonal superior es troben els valors de p-value, acolorits en verd quan són majors de 0.05, que és el valor de significança establert, i en taronja quan són iguals o menors a 0.05. A la diagonal inferior es mostren els valors de l'estadístic i les caselles estan acolorides de manera simètrica als valors de p-value.

3. 2. 3 Test K-S de les Mostres Agregades per Sexe

Les corbes de distribució de densitat d'incidència per sexe (Figura 21) revelen dues distribucions semblants, deixant entreveure que les mostres provenen d'una mateixa població. Les distribucions són asimètriques positives i bimodals, amb una concentració de les dades a l'esquerra de la mitjana i alguns valors extrems.

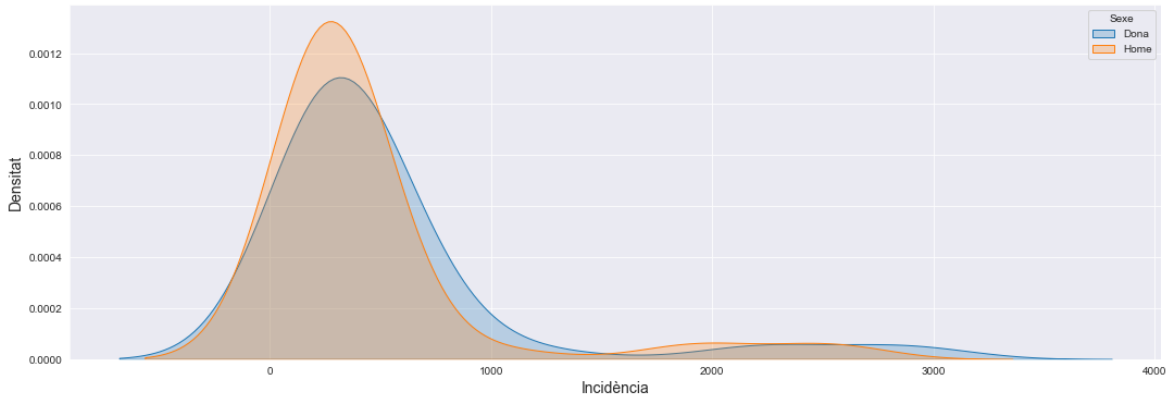


Figura 21. Densitat d'incidència de COVID-19 del 2022 per sexe.

Donat que només calia realitzar un test K-S, ja que hi ha dues mostres, els resultats no s'han expressat en forma de matriu sinó que es presenten a continuació: el valor del p-value és de 0.574 i la distància K-S és de 0.1538. Amb això exposat, es confirma la hipòtesi nul·la, perquè el p-value és major a 0.05 i, per tant, sembla que les mostres provenen de la mateixa distribució subjacent.

3. 2. 4 Test K-S de les Mostres Agregades per Índex Socioeconòmic (ISC)

En la Figura 22 s'il·lustren les distribucions de densitat per als diferents índexs socioeconòmics. Les distribucions presenten asimetria positiva i són bimodals, mostrant dos pics cadascuna. Tres dels índexs semblen tenir una distribució altament pareguda i, l'1, que es correspon al nivell socioeconòmic alt, sembla que es desmarca una mica més de la resta, però sense deixar de tenir pràcticament el mateix valor de moda en el primer pic. Dit això, es podria preveure que el test K-S no evidenciarà una gran diferència estadística entre les distribucions de les mostres.

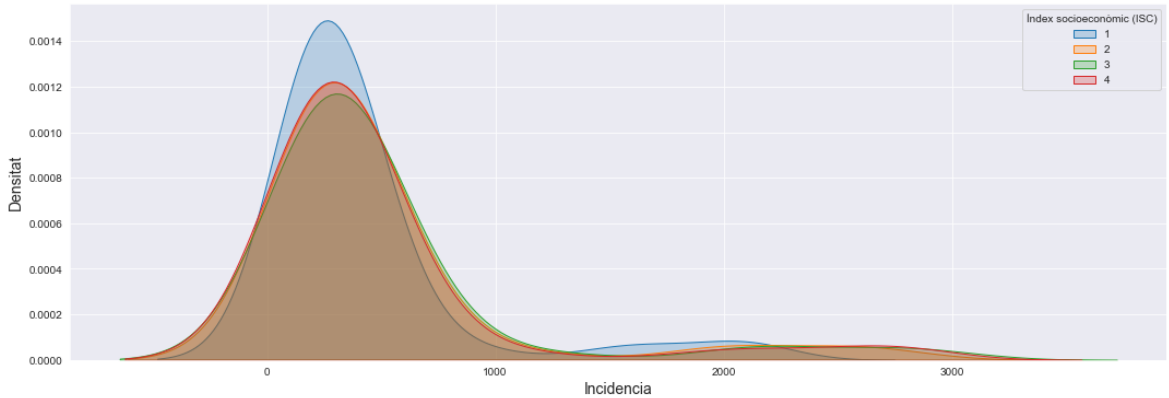


Figura 22. Densitat d'incidència de COVID-19 del 2022 per índex socioeconòmic (ISC).

Els resultats del test Kolmogórov-Smirnov posen de manifest el que s'havia predit: les mostres de tots els ISCs sembla que provenen d'una mateixa distribució (Figura 23). D'entre els valors de la matriu, es palpable com la distància K-S del nivell socioeconòmic alt amb la resta d'índexs és major que la distància dels altres índexs entre ells. Això és precisament el mateix que s'havia observat en les corbes de distribució de densitat, en les quals la distribució de l'índex 1 s'allunya una mica de les altres.

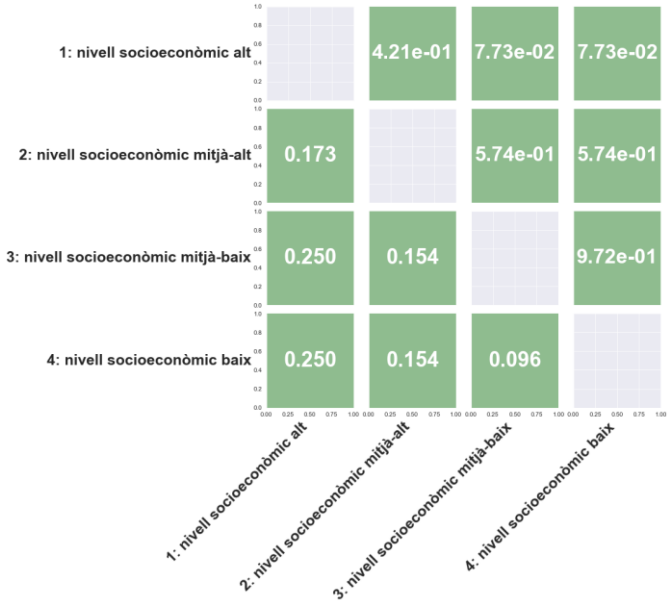


Figura 23. Matriu de resultats del test Kolmogórov-Smirnov per a les diferents combinacions de ISCs. A la diagonal superior es troben els valors de p-value, acolorits en verd quan són majors de 0.05, que és el valor de significança establert, i en taronja quan són iguals o menors a 0.05. A la diagonal inferior es mostren els valors de l'estadístic i les caselles estan acolorides de manera simètrica als valors de p-value.

3. 3 Anàlisi del Model de Regressió Lineal Múltiple (MLR)

El model de regressió lineal múltiple s'ha ajustat per als diagnòstics de COVID-19 i de grip. En les pròximes seccions es presenten els resultats obtinguts i es comparen els d'ambdues malalties.

Cal mencionar que per als models de MLR (els no desagregats per ABS) de la COVID-19 s'han utilitzat 5704 observacions, i per ajustar els models de la grip s'ha fet ús de 3703 observacions.

3. 3. 1 Coeficients de determinació (R^2)

La Figura 24 presenta la bondat d'ajust de les rectes de regressió als valors de les mostres. La R^2 pren valors de 0 a 1, en què 0 denota la inexistència de relació entre les variables i 1 indica que tots els punts es troben sobre la recta de regressió i, per tant, l'ajust és perfecte.

És important tenir en consideració que, com més gran és el nombre de predictors, més s'eleva el valor de R^2, de manera que cada predictor explica una part de la variabilitat observada en y. Per aquest motiu, la R^2 no serveix per a comparar models amb un nombre de predictors diferent.

Es pot observar que la tendència de totes dues rectes és decreixent, de manera que a mesura que augmenta el retard temporal, la capacitat explicativa del model disminueix. Per a la COVID-19, els valors inicials del coeficient de determinació són considerablement alts, indicant que el model presenta un bon ajust (arribant a explicar un 87% de la variància de les observacions). En el cas de la grip, els valors de R^2 són més constants, però inferiors als de la COVID-19. Aquesta estabilitat indica que l'ajust del model de la grip és menys sensible a la variació del retard temporal en comparació a la COVID-19.

Així doncs, el model de MLR, tant per a les dades de la COVID-19 com per a les de la grip, té el valor de R^2 més elevat quan les variables independents del model són les relatives a les dades de la setmana anterior (k-1). Tot i aquesta similitud, el model de COVID-19 presenta un ajust més proper a 1 i, per tant, un millor ajust.

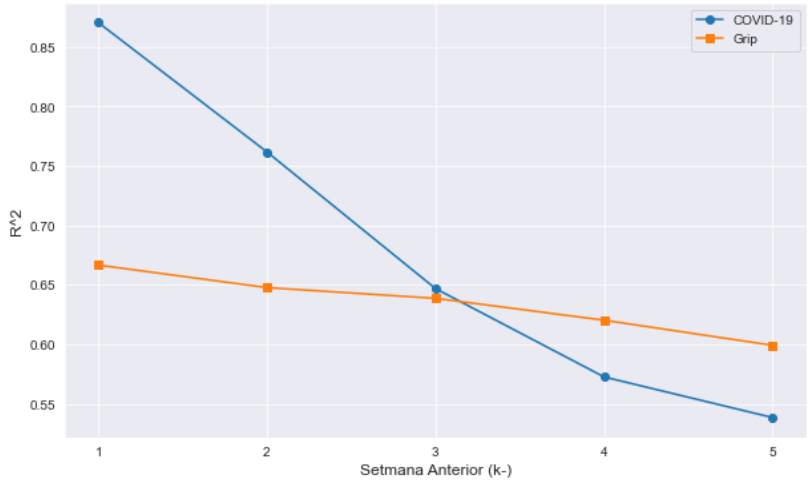


Figura 24. Gràfic de línies dels coeficients de determinació per a un retard de k setmanes, en blau per al diagnòstic de COVID-19 i en taronja per al de grip.

La Figura 25 i la Figura 26, per als diagnòstics de COVID-19 i grip, proporcionen una visió espacial de la capacitat predictiva dels models de MLR a través de les ABS, tot considerant diferents retards de setmanes (k-n).

Els resultats de la COVID-19 mostren, per a les dades de la setmana (k-1), un valor alt de R^2 en la majoria de les ABS. Es pot observar clarament que, a mesura que s'usen dades de setmanes més anteriors, la precisió del model decau notablement, de manera que en el primer plot pràcticament tot el mapa es mostra de color vermell i, a partir d'aquí, la predominança vermella es va atenuant i convertint-se en blau. Això suggereix que la capacitat de precisió del model no és homogènia a través de Catalunya. Algunes ABS mostren un ajust consistentment millor, mentre que d'altres, inclús per a (k-1), denoten una baixa precisió. Aquest aspecte podria indicar diferències locals pel que fa a la dinàmica de contagi de la COVID-19, a la recopilació de dades o a altres factors no capturats pel model.

Quant als resultats de la grip, la línia corresponent a aquest diagnòstic en la Figura 24 ja indicava que la tendència del valor de R^2 és decreixent i poc variable a mesura que augmenta el nombre de setmanes de retard, i així es veu també reflectit en la Figura 26. En general, els valors del coeficient de determinació en les diferents ABS són més aviat baixos. Tot i això, sembla que amb les dades de la setmana k-1 diverses ABS presenten valors relativament alts, amb colors més vermellosos, però tampoc es pot identificar un patró clar.

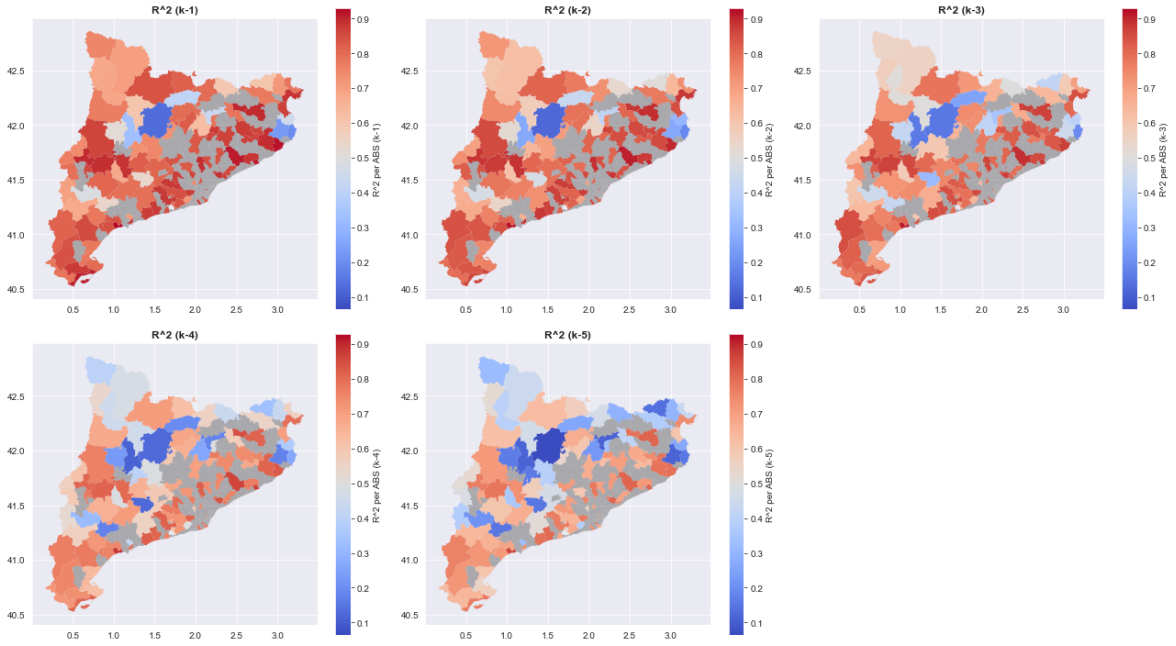


Figura 25. Representació geogràfica de les ABS i els valors del coeficient de determinació del model de MLR per a les dades de cada setmana de retard per al diagnòstic de COVID-19. L'escala de colors va dels blaus (que representen valors més baixos) als vermells (que representen valors més elevats), passant pel blanc. En color gris es troben marcades les ABS que no presenten cap estació compresa al seu territori i, per tant, no se'n registren dades meteorològiques per a poder introduir-les al model.

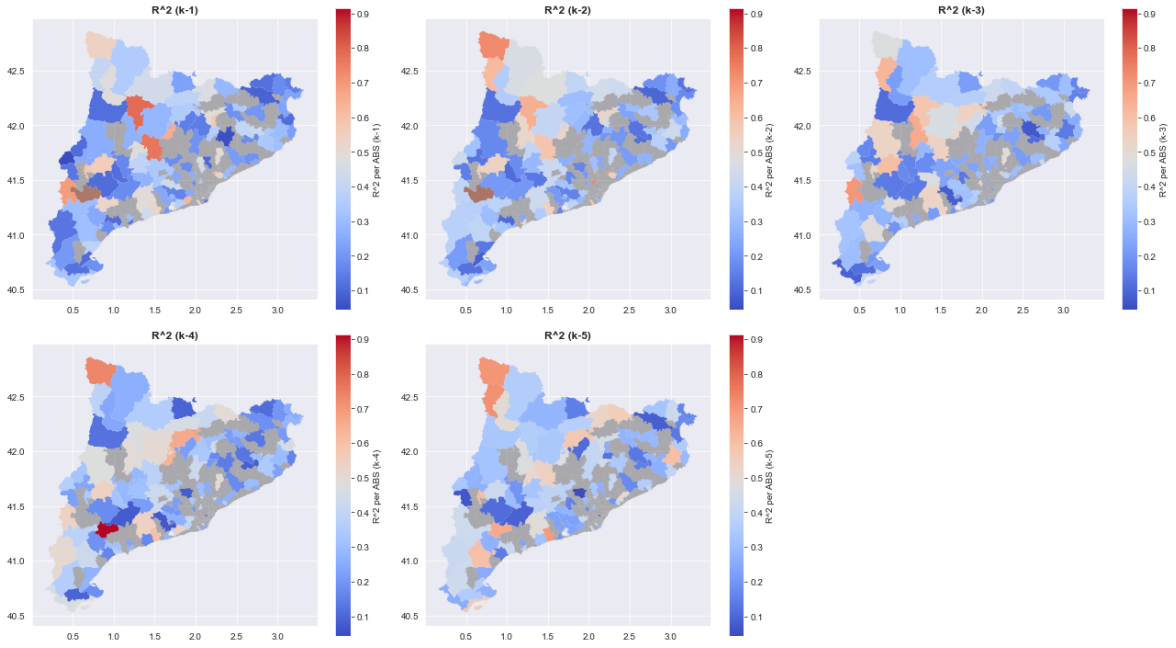


Figura 26. Representació geogràfica de les ABS i els valors del coeficient de determinació del model de MLR per a les dades de cada setmana de retard per al diagnòstic de grip. L'escala de colors va dels blaus (que representen valors més baixos) als vermells (que representen valors més elevats), passant pel blanc. En color gris es troben marcades les ABS que no presenten cap estació compresa al seu territori i, per tant, no se'n registren dades meteorològiques per a poder introduir-les al model. També s'inclouen en aquest grup les ABS que no disposaven de prou registres com per a realitzar el test.

Així doncs, tot i que ambdós diagnòstics comparteixen algunes tendències generals com la disminució progressiva a mesura que s'utilitzen dades de setmanes més antigues, hi ha diferències significatives en el nivell de precisió del model, en el qual la COVID-19 mostra resultats notablement més alts en comparació a la grip, i en la distribució geogràfica de la capacitat predictiva. Pel que fa a aquest últim punt, sí que sembla que les ABS amb valors més alts o més baixos difereixen entre ambdós diagnòstics, però no s'identifica cap patró evident en aquest sentit.

3. 3. 2 Coeficients dels predictors

Després d'observar que els models de COVID-19 i grip presenten un millor ajust amb un retard d'una setmana, l'anàlisi s'ha centrat en aquests models per a aprofundir en els resultats dels coeficients dels predictors com a clau per entendre les dinàmiques de contagi. Els coeficients són els que indiquen el pes que té el predictor sobre la variable dependent.

En la Taula 1 es mostren els valors dels coeficients, i del seu respectiu p-value, per als diagnòstics de COVID-19 i grip. Pel que fa al valor de la constant, aquesta significa que si la resta dels coeficients fossin zero, el valor de la y seria el de la constant. Amb les dades de la taula, tenint en compte que els valors d'incidència són logarítmics, en termes absoluts la incidència de COVID-19 seria de 436.24 i la de grip seria de 233.92.

Amb relació a la COVID-19, el predictor que mostra una influència significativa en la incidència és, fonamentalment, la incidència de la setmana anterior. El coeficient és positiu, de manera que si la incidència ha augmentat la setmana anterior, també ho farà aquesta. Tal resultat és coherent, ja que el virus es transmet principalment per via aèria i, per tant, un increment en el nombre de casos d'infectats augmenta la probabilitat de transmissió.

Els següents coeficients que més influeixen són els associats a viatges exteriors, tant d'entrada com de sortida. Els viatges d'entrada tenen una relació directa amb la incidència, en indicar que un major nombre de persones que arriben a Catalunya, com serien els turistes, poden contribuir a un augment dels casos. D'altra banda, els viatges de sortida mostren una relació inversa, suggerint que una disminució de les sortides del territori pot correlacionar-se amb un augment en la incidència. Uns altres coeficients menys significatius són els ambientals, de temperatura i humitat relativa, ambdós negatius, indicant que una disminució de la temperatura o de la humitat relativa pot provocar un augment de la incidència. Aquest resultat és congruent amb molts estudis epidemiològics prèviament mencionats.

Finalment, el coeficient relatiu als viatges interiors és relativament petit i mostra una relació inversa amb la incidència, palesant que una disminució de la mobilitat interna pot augmentar la incidència. No obstant això, és crucial considerar el valor del p-value, que indica la significança estadística dels coeficients. En el cas del coeficient de mobilitat interna, el valor del p-value és considerablement gran (> 0.05), tot suggerint que la variable independent no té un efecte significatiu sobre la variable dependent. En contrast, la resta de p-values dels altres coeficients són < 0.05, manifestant significança estadística.

Pel que fa a la grip, la variable que més influència té en la incidència és la incidència de la setmana anterior, amb un coeficient positiu, cosa que indica que un augment en la incidència de la setmana anterior es correlaciona amb un augment la setmana següent. El segon coeficient més influent és el relatiu a la temperatura, el qual té un valor negatiu, suggerint que una disminució de la temperatura pot provocar un augment de la incidència. Aquests resultats són congruents amb la naturalesa estacional de la grip, que generalment es presenta a l'hivern, quan les temperatures són més baixes.

Els següents coeficients que més influeixen són els dels viatges exteriors d'entrada i de sortida. El coeficient dels viatges d'entrada és positiu, mentre que el de viatges de sortida és negatiu, similar a la situació de la COVID-19. En el cas dels viatges de sortida, el p-value és superior a 0.05, indicant que el resultat no és significatiu. Un altre coeficient amb una influència menor és el relatiu a la mobilitat interna. Aquest coeficient és positiu, suggerint que un augment en els desplaçaments per Catalunya es correlaciona amb augment de la incidència, una dada lògica, ja que més moviments poden facilitar la transmissió del virus.

Per últim, el coeficient d'humitat relativa és positiu, és el que menys influeix en la incidència i té un p-value superior a 0.05, de manera que no és significatiu. Cal esmentar que la resta de p-values que no s'han comentat són inferiors a 0.05 i, en conseqüència, asseguren la significança estadística dels coeficients respectius.

Taula 1. Comparació dels resultats del model de regressió lineal múltiple de les dades de la setmana (k-1) per als diagnòstics de COVID-19 i grip. S'exposen els coeficients amb els seus respectius p-values.

	COVID-19		GRIP	
	Coef.	p-value	Coef.	p-value
Constant	6.0782	0.0000	5.4550	0.0000
β_1 = temperatura	-0.0198	0.0040	-0.0739	0.0000
β_2 = humitat	-0.0177	0.0001	0.0035	0.6369
β_3 = incidència	0.7191	0.0000	0.5619	0.0000
β_4 = Viatges interiors	-0.0015	0.7561	0.0199	0.0214
β_5 = Viatges ext. d'entrada	0.0636	0.0000	0.0492	0.0013
β_6 = Viatges ext. de sortida	-0.0571	0.0000	-0.0219	0.1226

Tal com s'ha fet amb els resultats del coeficient de determinació, per als coeficients dels predictors també s'han representat les dades geogràficament, desagregades per ABS (Figura 27).

En primer lloc, pel que fa al coeficient de temperatura, es mostra una distribució predominantment negativa en la major part de les ABS, tant per a la COVID-19 com per a la grip. En el cas de la COVID-19, algunes zones mostren coeficients més propers a zero, tot indicant una menor influència de la temperatura. En canvi, per a la grip, els valors tendeixen a ser més negatius, a excepció d'alguna ABS que presenta coeficients lleugerament positius.

Els coeficients d'humitat relativa, en el cas de la COVID-19, són majoritàriament negatius, presentat una distribució força homogènia per tot el territori català. Per a la grip, tot i tenir una distribució també majoritàriament negativa, hi ha més variabilitat, especialment en les ABS del sud-oest.

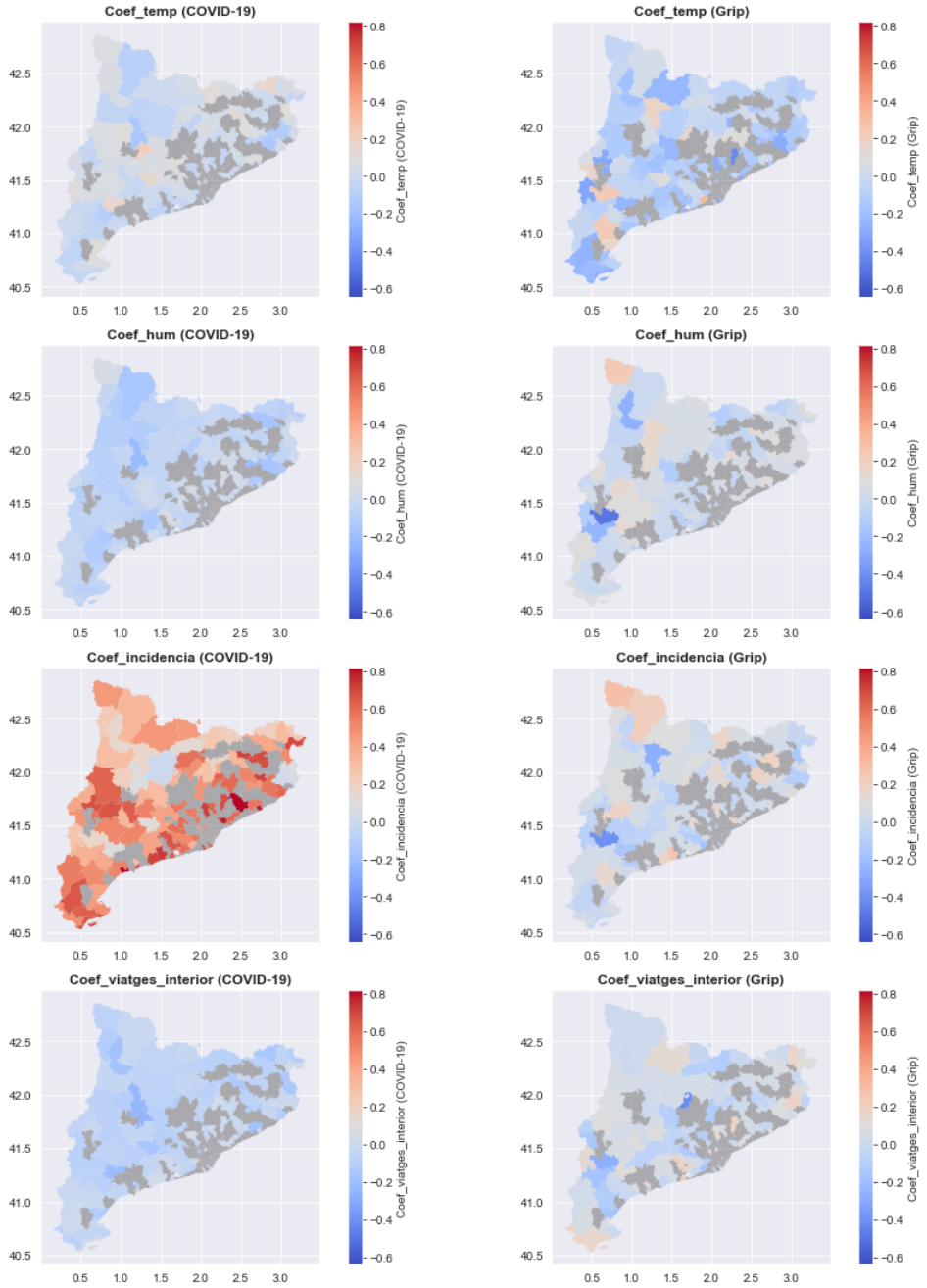
Envers la distribució geogràfica dels coeficients d'incidència, els de la COVID-19 són fortament positius a gairebé totes les ABS, sobretot a les zones de la costa. Algunes ABS del centre i del nord-oest presenten valors de coeficients més moderats, inclús lleugerament negatius. Per a la grip també s'identifiquen coeficients positius per a la majoria de les ABS, però amb una intensitat menor. No obstant això, algunes ABS tenen coeficients negatius que semblen arribar fins a -0.3.

La influència dels viatges interiors per a la COVID-19 és predominantment negativa i força homogènia per a les ABS de Catalunya. Respecte a la grip, els patrons són semblants però menys pronunciats, amb algunes ABS que mostren coeficients propers a zero o positius.

En referència als coeficients d'entrades de viatges exteriors a Catalunya, tant la grip com la COVID-19 revelen resultats força variats per a les diferents ABS. Per a la COVID-19, els coeficients es mouen entorn valors més propers al zero, mentre que per a la grip la variabilitat és major, amb predominança de valors entorn a zero o lleugerament positius.

Per últim, amb relació a la COVID-19, els coeficients de sortides de viatges exteriors són predominantment negatius i força semblants entre les diferents ABS. Els de la grip són majoritàriament negatius, però presenten una major variabilitat, sense arribar a identificar-se un patró geogràfic clar.

Essencialment, tot i que els patrons generals són similars entre ambdues malalties, les magnituds dels coeficients i les distribucions geogràfiques mostren variacions. En línies generals, les distribucions dels sis coeficients de la COVID-19 no presenten tanta variabilitat entre les diferents ABS, mentre que la grip mostra una diversitat de valors més pronunciada.



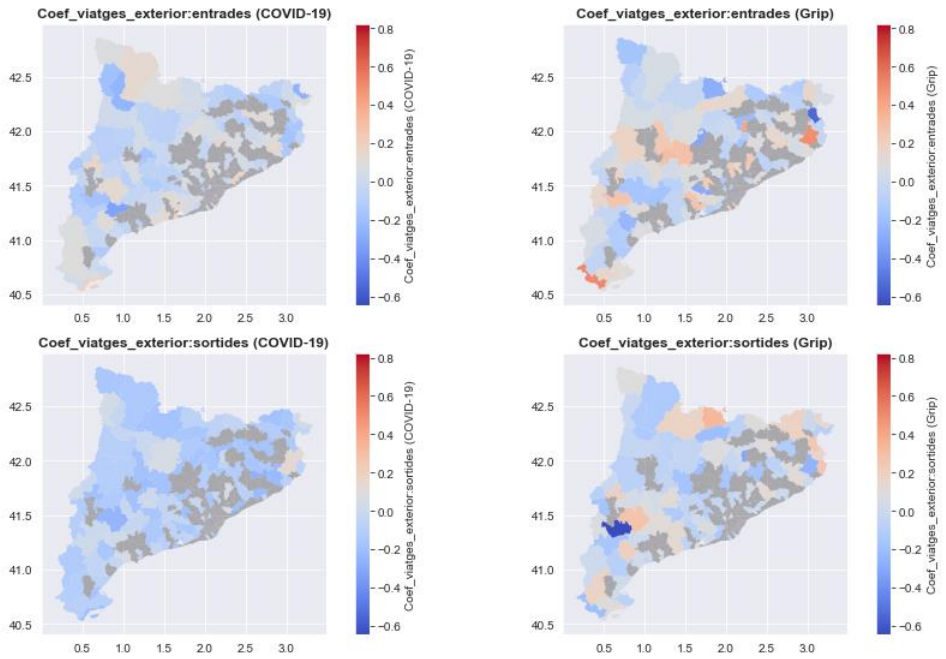


Figura 27. Representació geogràfica dels sis coeficients del model de MLR per a cada ABS per als diagnòstics de COVID-19 i de grip. L'escala de colors va dels blaus (que representen valors més baixos) als vermells (que representen valors més elevats), passant pel blanc. En color gris es troben marcades les ABS que no presenten cap estació compresa al seu territori i, per tant, no se'n registren dades meteorològiques per a poder introduir-les al model.

3. 3. 3 Predicció de Valors d'Incidència per a Una Àrea Bàsica de Salut

Una vegada s'han estimat els coeficients, el model de regressió lineal múltiple pot ser usat per a fer prediccions sobre dades noves, és a dir, per obtenir estimacions de la incidència futura donades les condicions presents.

Primerament, per a avaluar un cas concret, s'ha escollit l'àrea bàsica de salut de 'Tarragona - 2' i s'han calculat les prediccions de COVID-19 i de grip per a les tres primeres setmanes de 2023. Aquestes són les setmanes dels dies 2, 9 i 16 de gener.

Per una banda, la Figura 28 mostra la comparació de les prediccions amb les observacions reals per a la COVID-19. Per a la primera setmana, es pot observar una diferència significativa entre la predicció i la realitat, fent palesa d'una sobreestimació del nombre d'incidències esperades. A la segona setmana, la diferència entre ambdós valors és menor, amb la predicció que s'acosta més al valor observat. A la tercera setmana, malgrat que la predicció continua sent superior a l'observació, la diferència és més petita en comparació a la primera setmana.

Per una altra banda, en el cas de la grip (Figura 29), per a la primera setmana la predicció és lleugerament superior, però s'acosta força al valor observat. A la segona setmana, la diferència entre ambdós valors augmenta i es fa més notable. I, a la tercera setmana, aquesta diferència s'amplia encara més.

En resum, el model tendeix a preveure valors superiors d'incidències dels que realment es produeixen. És rellevant tenir en compte que, com s'ha destacat en l'apartat 3.3.2, la incidència de la setmana anterior és el factor que més influeix en la predicció de la incidència actual, tant per a la COVID-19 com per a la grip. Aquestes tendències es poden observar clarament en ambdues figures.

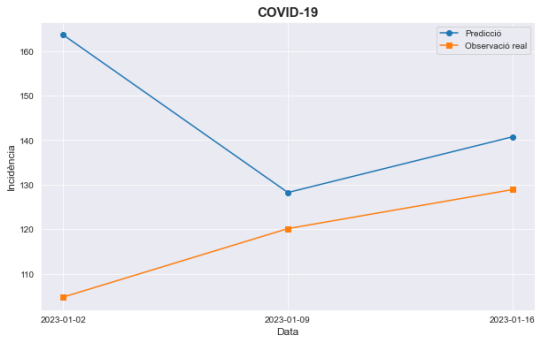


Figura 28. Representació gràfica de la comparació de les prediccions (línia blava) amb les observacions reals d'incidència (línia taronja) de les tres primeres setmanes del 2023 per a la COVID-19.

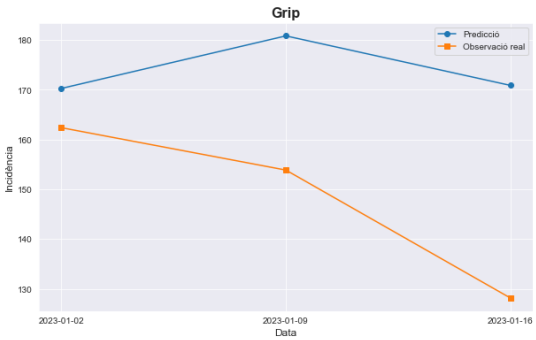


Figura 29. Representació gràfica de la comparació de les prediccions (línia blava) amb les observacions reals d'incidència (línia taronja) de les tres primeres setmanes de 2023 per a la grip.

3. 3. 4 Predicció de Valors d'Incidència per a Totes les Àrees Bàsiques de Salut i Avaluació dels Resultats

A banda de la predicció d'incidència de 'Tarragona - 2', s'han predit les de COVID-19 i grip de les tres primeres setmanes del 2023 per a totes les ABS. La Figura 30 i la Figura 31 mostren la comparació entre les prediccions del model de MLR i les observacions reals.

Pel que fa als resultats de la COVID-19 (Figura 30), s'observa que, en general, la major part dels punts es troben sobre la línia de referència, suggerint que les prediccions del model són raonablement properes a les observacions reals per a la majoria dels casos. Tot i això, s'identifiquen diversos punts dispersos lluny de la línia, la major part dels quals es troba per sobre de línia, és a dir, que correspon a observacions amb valors elevats, un fet que indica que el model tendeix a subestimar les incidències quan aquestes són més altes.

Amb relació a les mostres de cada data, les dos primeres setmanes semblen presentar una major dispersió en les prediccions. Això podria suggerir que, a mesura que avança el mes de gener i els valors es van estabilitzant després de les vacances de Nadal, les prediccions esdevenen més precises, sense pics inesperats.

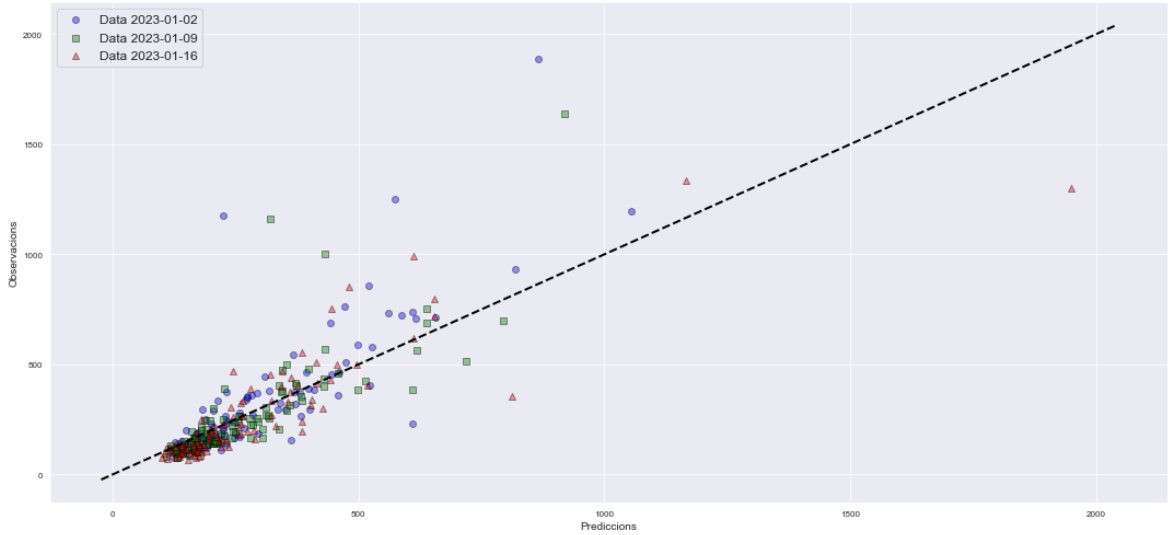


Figura 30. Representació gràfica per comparar les prediccions i les observacions reals de la incidència de COVID-19 per a tres setmanes diferents. Cada setmana està representada per punts d'un color i d'una forma diferents. La línia discontínua té una pendent igual a 1, i permet identificar la precisió de la predicció, en la qual els valors sobre la línia indiquen una predicció perfecta.

Per a la grip (Figura 31), en comparació a la COVID-19, la representació és força similar, amb la majoria dels punts concentrats prop de la línia, palesant que el model manté una capacitat predictiva raonable per a la majoria de les observacions. No obstant això, la dispersió és menor i es dona especialment per a valors alts d'incidència observada, amb alguns punts notablement allunyats de la línia de referència.

Com amb la COVID-19, el model tendeix a subestimar els valors alts d'incidència, i la desviació d'alguns d'aquests punts dispersos suggereix que el model lineal no capta completament les relacions entre les variables. Les observacions corresponents a les tres setmanes semblen mantenir un patró similar. Aquesta consistència en el temps indica que el model de la grip manté una precisió relativament constant en les seves prediccions.

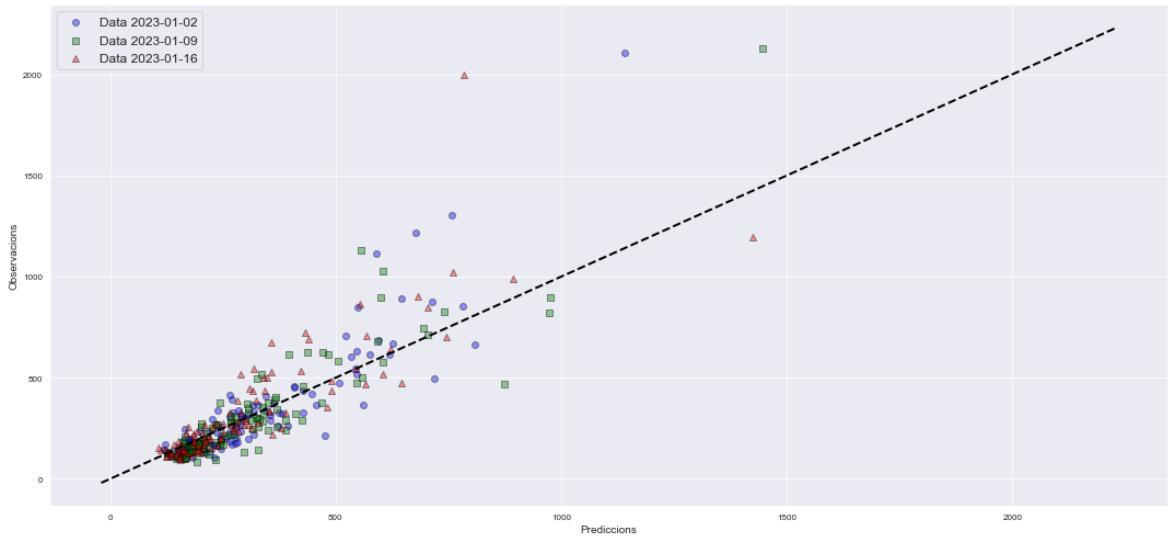


Figura 31. Representació gràfica per comparar les prediccions i les observacions reals de la incidència de grip per a tres setmanes diferents. Cada setmana està representada per punts d'un color i d'una forma diferents. La línia discontinua té una pendent igual a 1, i permet identificar la precisió de la predicció, en la qual els valors sobre la línia indiquen una predicció perfecta.

En conclusió, tot i que tots dos models presenten prediccions raonablement properes a la línia, el de la grip sembla tenir una precisió lleugerament superior en comparació al de la COVID-19. Pel que fa a les desviacions, ambdós models en presenten, però les de la grip són menys pronunciades. És complicat deduir a què es deuen els punts dispersos, ja que probablement hi ha més factors que no es tenen en compte, com ara aspectes de la naturalesa dels virus, com ho és la transmissibilitat. Aquest factor podria jugar un paper important, perquè la COVID-19 es transmet més fàcilment que la grip, un fet que pot provocar que d'una setmana a l'altra els valors augmentin més del que el model prediria.

A més, caldria estudiar perquè per a valors reals baixos el model tendeix a sobreestimar-los i per a valors reals alts a subestimar-los. Aquest patró suggereix que els valors extrems no són capturats adequadament pels models actuals i caldria o bé ajustar el model amb més predictors, o bé cercar un altre tipus de model que no fos lineal.

Per a quantificar en quina mesura s'han allunyat les prediccions dels valors reals, per a cada àrea bàsica de salut i data s'ha calculat l'error relatiu (ER) com a percentatge. Els resultats es mostren representats als mapes d'ABS (Figura 32 i Figura 33), en els quals l'escala de color es centra en el zero per a facilitar la identificació de les àrees en què el model té una major desviació de les observacions reals. Els resultats de l'error relatiu s'interpreten de la següent manera:

- Si ER = 0, el valor predit pel model coincideix exactament amb el valor real.
- Si ER > 0, la predicció és inferior al valor de l'observació real. Això implica que el model està subestimant la variable d'interès.
- Si ER < 0, la predicció és superior al valor de l'observació real. Això implica que el model està sobreestimant la variable d'interès.

Per al diagnòstic de COVID-19, a la primera setmana s'observen zones amb errors relatius tant positius com negatius. A primera vista no es detecta un patró clar; no obstant això, les ABS amb blaus més foscos, per a les quals el model sobreestima la incidència real, es concentren al nord-est de Catalunya. A la resta del territori, els colors són variats, sense una predominança.

La segona setmana mostra una distribució més equilibrada dels errors relatius. Es percep una tendència general cap a una millor precisió del model, ja que les àrees grises són més prevalents. Així i tot, algunes ABS al llarg del territori presenten errors relatius significatius, tant positius com negatius.

Els errors relatius de la tercera setmana revelen una distribució geogràfica més marcada. S'observa una concentració de valors negatius a les àrees del sud-oest del territori. A més, algunes ABS aïllades mostren valors extrems que arriben fins al -100%. Pel que fa a les zones amb errors relatius positius, hi ha una presència més notable, però no s'identifica un patró evident.

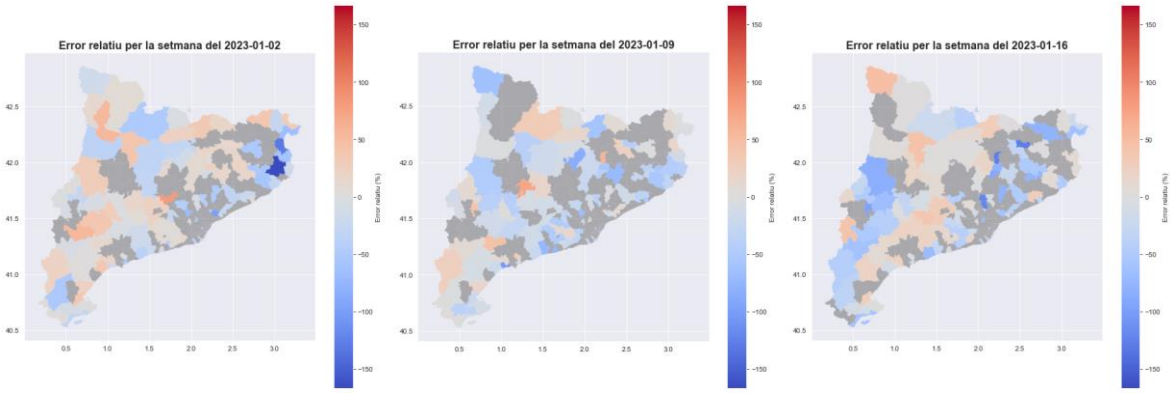


Figura 32. Representació geogràfica de l'error relatiu (%) entre les prediccions i les observacions reals de la COVID-19 per a cada ABS. Es mostren tres mapes, un per a cada setmana de prediccions. L'escala de color està centrada amb el zero en el color blanc, de manera que els colors vermellosos indiquen un percentatge d'error positiu (predicció < observació) i els colors blavosos indiquen un percentatge d'error negatiu (predicció > observació). En color gris fosc es troben marcades les ABS de les quals no es tenen prediccions i/o observacions reals per a aquella setmana o, en no presentar estacions meteorològiques, les quals no s'han introduït al model.

Per al diagnòstic de la grip, la primera setmana s'observen errors relatius tant negatius com positius distribuïts per tot el territori. Els errors més significatius es concentren principalment al nord-est, on el model tendeix a sobreestimar la incidència real (colors blaus més foscos). Per a la resta de zones hi ha una presència variada d'àrees grises, en les quals les prediccions s'apropen molt a l'observació real, i d'àrees vermelles, en les quals el valors de la incidència han estat subestimats.

La segona setmana presenta una distribució homogènia dels errors relatius, tant positius com negatius. Hi ha presència d'algunes àrees amb valors notablement extrems, especialment de color blau, en què se sobreestima la incidència real. Tot i això, les àrees grises són més prevalents, suggerint una millora en la precisió de les prediccions.

La tercera setmana presenta, en general, colors més atenuats en comparació amb la segona, un arranjament que indica una menor dispersió. Hi ha una major presència d'ABS que s'apropen al 0% d'error, i les zones vermelles són més prevalents que les de color blau. També hi ha algunes ABS amb errors relatius significatius, però no tan extrems com a les dues primeres setmanes.

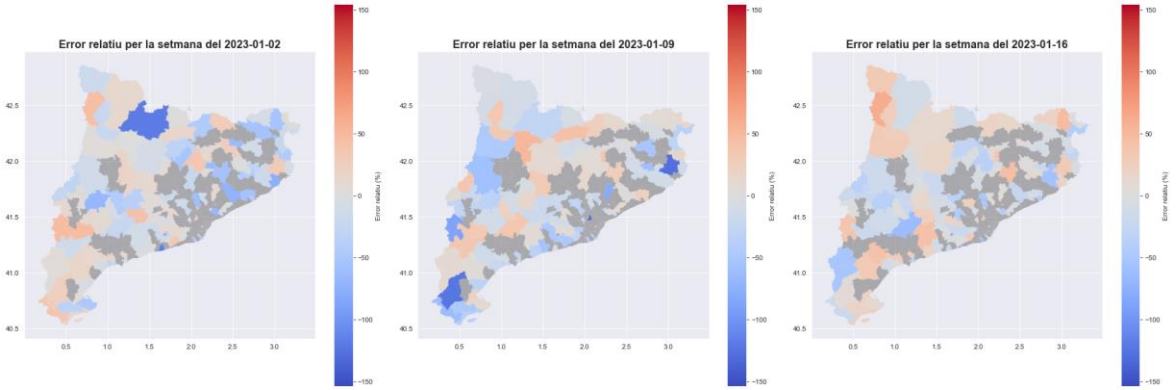


Figura 33. Representació geogràfica de l'error relatiu (%) entre les prediccions i les observacions reals de la grip per a cada ABS. Es mostren tres mapes, un per a cada setmana de prediccions. L'escala de color està centrada amb el zero en el color blanc, de manera que els colors vermellosos indiquen un percentatge d'error positiu (predicció < observació) i els colors blavosos indiquen un percentatge d'error negatiu (predicció > observació). En color gris fosc es troben marcades les ABS de les quals no es tenen prediccions i/o observacions reals per a aquella setmana o, en no presentar estacions meteorològiques, les quals no s'han introduït al model.

En conclusió, les diferents representacions geogràfiques, tant de COVID-19 com de grip, evidencien que l'ER no presenta uniformitat ni temporalment ni geogràficament, i en la major part dels contextos no s'hi identifica un patró clar. En ambdós casos, la precisió del model és raonablement alta en diverses àrees, però encara hi ha marge per a la millora, especialment en la reducció dels errors relatius significatius. Per a perfeccionar l'ajust del model, caldria identificar i analitzar les ABS que mostren una major desviació, ja que aquestes requereixen un ajustament més precís i podrien fer evidents les mancances del model.

4 Conclusions

En síntesi, aquest estudi ha aportat una anàlisi exhaustiva de la relació entre les condicions ambientals i la mobilitat amb la incidència de la COVID-19 i la grip a Catalunya. A partir de les dades i resultats obtinguts, es poden extreure conclusions significatives en relació als objectius plantejats a l'inici.

En primer lloc, s'ha pogut avaluar l'evolució de la incidència de COVID-19 entre els anys 2020 i 2023, tant de manera general com desagregant les dades per edat, regió sanitària, sexe i índex socioeconòmic. Els resultats dels diferents gràfics desagregats han mostrat una distribució relativament homogènia d'incidència entre els diferents grups, indicant que la major part de les vegades els grups han experimentat les onades de la pandèmia de manera sincronitzada. No obstant això, les majors diferències s'observen entre els diferents grups d'edat, on sobretot en els pics de les onades destaquen per sobre dels demés els grups que engloben edats dels 5 als 44 anys, i de 80 o més anys. Aquests resultats coincideixen amb els presentats per diversos organismes estadístics oficials.

S'ha emprat la prova Kolmogórov-Smirnov per determinar si les mostres dels grups de cada desagregació (regió sanitària, edat, sexe i índex socioeconòmic) provenien d'una mostra subjacent. Els resultats han revelat que els grups d'índex socioeconòmic i de sexe, respectivament, tenen distribucions estadístiques semblants. Pel que fa als grups d'edat, només 4 de les 21 combinacions han obtingut un valor de significança major a 0.05, i 18 de les 45 combinacions de regions semblen provenir d'una mateixa població. Això indica que aquests factors han provocat diferències significatives en la incidència. En el cas de l'edat, tals diferències es poden atribuir a variacions en l'exposició al virus, la susceptibilitat a la malaltia i les activitats socials. Quant a les regions, factors com la densitat de població, les condicions ambientals i la mobilitat poden haver contribuït a les diferències en la taxa de contagi entre els diferents grups.

Amb relació als models de regressió lineal múltiple de la COVID-19 i la grip, per a tots dos diagnòstics les mostres que millor s'ajusten són les corresponents al retard d'una setmana, amb una R^2 de 0.87 per la COVID-19 i de 0.66 per la grip. Amb aquests models s'han obtingut uns coeficients dels predictors de la COVID-19, de major a menor influència en la incidència, i o bé positius (+) o bé negatius (-), que són, tots ells amb un retard d'una setmana: incidència (+), viatges d'entrada a Catalunya (+), viatges de sortida de Catalunya (-), temperatura (-) i humitat (-). Per a la grip, de major a menor impacte en la incidència, i també amb un retard d'una setmana, són: incidència (+), temperatura (-), viatges d'entrada a Catalunya (+) i viatges interiors per Catalunya (+). Els signes dels factors coincideixen per als dos diagnòstics, i revelen una relació negativa dels factors ambientals amb la incidència i, per als viatges, una relació positiva amb els d'entrada i interiors, i amb els de sortida una de negativa. Aquests resultats són congruents i evidencien per a la COVID-19 un major impacte de la mobilitat davant de les condicions ambientals sobre la incidència i, per a la grip, és a l'inrevés, ja que la temperatura sembla tenir un major efecte sobre la incidència. Això demostra el que molts altres estudis epidemiològics han conclòs, i és que les temperatures més baixes poden afavorir la supervivència dels virus, augmentant la seva transmissibilitat, i uns nivells baixos d'humitat poden facilitar que les partícules que contenen virus estiguin més temps en suspensió, incrementant així la probabilitat de ser

inhalades. Cal mencionar que alguns dels coeficients mostraven resultats no significatius, ja que tenen un p-value major a 0.05, i per aquest motiu no s'han esmentat aquí.

Quant a l'abast del model ajustat, s'han pogut predir nous valors d'incidència per al 2023, obtenint prediccions raonablement properes a les observacions reals, amb una precisió lleugerament superior en el cas de la grip en comparació a la COVID-19. És probable que tenint en compte més factors per al model, com ara la transmissibilitat dels virus, la precisió augmentés. Amb el càlcul de l'error relatiu de les prediccions de cada ABS, s'ha avaluat l'efectivitat del model predictiu i s'ha observat que l'ER no presenta uniformitat ni temporalment ni geogràfica i, en la major part dels contextos, no s'hi identifica un patró clar.

Malgrat tot, aquest estudi compta amb algunes limitacions, com ara les que presenta el model de regressió lineal múltiple. Aquest, que ja és més complex que el de regressió simple, en moltes ocasions no és capaç de capturar la complexitat espaciotemporal de la transmissió del virus. Caldria explorar models no lineals o d'aprenentatge automàtic per veure si aconsegueixen captar les dinàmiques més complexes. Per altra banda, es podrien millorar alguns aspectes de les dades emprades, com és el cas de les dades de mobilitat, les quals seria millor tenir desagregades, per exemple, per municipis, per així obtenir resultats més precisos; o de les dades meteorològiques, per a les quals seria més adient aproximar els valors de temperatura i humitat per a les ABS que no presenten cap estació meteorològica a partir de registres d'estacions properes, i així no "perdre" les dades ambientals d'aquelles àrees bàsiques de salut.

Referències

[1] J. Sommerfeld, «Plagues and peoples revisited», *EMBO Rep*, vol. 4, núm. S1, 2003, doi: 10.1038/sj.embor.embor845.

[2] World Health Organization, «The top 10 causes of death». Consulta: 1 juliol 2024. [En línia]. Disponible a: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] A. Fontal, M. J. Bouma, A. San-José, L. López, M. Pascual, i X. Rodó, «Climatic signatures in the different COVID-19 pandemic waves across both hemispheres», *Nat Comput Sci*, vol. 1, núm. 10, 2021, doi: 10.1038/s43588-021-00136-6.

[4] M. S. Al-Khateeb, F. A. Abdulla, i W. K. Al-Delaimy, «Long-term spatiotemporal analysis of the climate related impact on the transmission rate of COVID-19.», *Environ Res*, vol. 236, p. 116741, nov. 2023, doi: 10.1016/j.envres.2023.116741.

[5] A. Tobías, T. Molina, M. Rodrigo, i M. Saez, «Meteorological factors and incidence of COVID-19 during the first wave of the pandemic in Catalonia (Spain): A multi-county study», *One Health*, vol. 12, p. 100239, juny 2021, doi: 10.1016/j.onehlt.2021.100239.

[6] M. Moazeni, M. Rahimi, i A. Ebrahimi, «What are the effects of climate variables on COVID-19 pandemic? A systematic review and current update», *Adv Biomed Res*, vol. 12, núm. 1, p. 33, 2023, doi: 10.4103/abr.abr_145_21.

[7] G. Donzelli, A. Biggeri, A. Tobias, L. N. Nottmeyer, i F. Sera, «Role of meteorological factors on SARS-CoV-2 infection incidence in Italy and Spain before the vaccination campaign. A multi-city time series study», *Environ Res*, vol. 211, p. 113134, ago. 2022, doi: 10.1016/j.envres.2022.113134.

[8] C. Valero, R. Barba, D. P. Marcos, N. Puente, J. A. Riancho, i A. Santurtún, «Influencia de los factores meteorológicos en la incidencia de COVID-19 en España», *Med Clin (Barc)*, vol. 159, núm. 6, p. 255-261, set. 2022, doi: 10.1016/j.medcli.2021.10.010.

[9] J. Wang, K. Tang, K. Feng, i W. Lv, «High Temperature and High Humidity Reduce the Transmission of COVID-19», *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3551767.

[10] Y. Yao *et al.*, «No association of COVID-19 transmission with temperature or UV radiation in Chinese cities», *European Respiratory Journal*, vol. 55, núm. 5, p. 2000517, maig 2020, doi: 10.1183/13993003.00517-2020.

[11] S. Xiao *et al.*, «Meteorological conditions are heterogeneous factors for COVID-19 risk in China», *Environ Res*, vol. 198, 2021, doi: 10.1016/j.envres.2021.111182.

[12] F. Shahzad, U. Shahzad, Z. Fareed, N. Iqbal, S. H. Hashmi, i F. Ahmad, «Asymmetric nexus between temperature and COVID-19 in the top ten affected provinces of China: A current application of quantile-on-quantile approach», *Science of the Total Environment*, vol. 736, 2020, doi: 10.1016/j.scitotenv.2020.139115.

[13] A. Lison, J. Persson, N. Banholzer, i S. Feuerriegel, «Estimating the effect of mobility on SARS-CoV-2 transmission during the first and second wave of the COVID-19 epidemic, Switzerland, March to December 2020», *Eurosurveillance*, vol. 27, núm. 10, 2022, doi: 10.2807/1560-7917.ES.2022.27.10.2100374.

[14] O. Gatalo, K. Tseng, A. Hamilton, G. Lin, i E. Klein, «Associations between phone mobility data and COVID-19 cases», *The Lancet Infectious Diseases*, vol. 21, núm. 5. 2021. doi: 10.1016/S1473-3099(20)30725-8.

[15] N. K. Bergman i R. Fishman, «Correlations of mobility and Covid-19 transmission in global data», *PLoS One*, vol. 18, núm. 7 July, 2023, doi: 10.1371/journal.pone.0279484.

[16] Govern Obert. Generalitat de Catalunya, «Vigilància sindròmica d'infeccions a Atenció Primària». Consulta: 12 març 2024. [En línia]. Disponible a: https://analisi.transparenciacatalunya.cat/Salut/Vigil-ncia-sindr-mica-d-infeccions-a-Atenci-Prim-r/44sy-txnv/about_data

[17] Govern Obert. Generalitat de Catalunya, «Sistema d'Informació per a la Vigilància d'Infeccions a Catalunya». Consulta: 5 març 2024. [En línia]. Disponible a: https://sivic.salut.gencat.cat/

[18] J. Reina, R. M. Arcay, M. Busquets, i H. Machado, «Impact of hygienic and social distancing measures against sars-cov-2 on respiratory infections caused by other viruses», *Revista Espanola de Quimioterapia*, vol. 34, núm. 4, 2021, doi: 10.37201/req/017.2021.

[19] Departament de Salut. Generalitat de Catalunya, «Cartografia. Capa amb les àrees bàsiques de salut, sectors sanitaris, àrees de gestió assistencial i regions sanitàries, any 2024». Consulta: 1 abril 2024. [En línia]. Disponible a: https://salutweb.gencat.cat/ca/departament/estadistiques-sanitaries/cartografia/index.html

[20] Generalitat de Catalunya. CatSalut., «Regions sanitàries», Consulta: 19 juny 2024. [En línia]. Disponible a: https://catsalut.gencat.cat/ca/coneix-catsalut/catsalut-territori/index.html

[21] C. Stal, L. De Sloover, J. Verbeurgt, i A. De Wulf, «On Finding a Projected Coordinate Reference System», *Geographies*, vol. 2, núm. 2, 2022, doi: 10.3390/geographies2020017.

[22] Govern Obert. Generalitat de Catalunya. METEOCAT, «Dades meteorològiques de la XEMA». Consulta: 19 maig 2024. [En línia]. Disponible a: https://analisi.transparenciacatalunya.cat/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee/about_data

[23] Govern Obert. Generalitat de Catalunya. METEOCAT, «Metadades variables meteorològiques». Consulta: 24 abril 2024. [En línia]. Disponible a: https://analisi.transparenciacatalunya.cat/Medi-Ambient/Metadades-variables-meteorol-giques/4fb2-n3yi/about_data

[24] Govern Obert. Generalitat de Catalunya. METEOCAT, «Metadades estacions meteorològiques automàtiques». Consulta: 1 maig 2024. [En línia]. Disponible a: https://analisi.transparenciacatalunya.cat/Medi-Ambient/Metadades-estacions-meteorol-giques-autom-tiques/yqwd-vj5e/about_data

[25] Ministerio de Transportes y Movilidad Sostenible, «Estudio de la movilidad nacional con tecnologías Big Data en formato abierto». Consulta: 9 juny 2024. [En línia]. Disponible a: https://data.mitma.gob.es/public/mov-autonomica

[26] Ministerio de Transportes y Movilidad Sostenible, «Metodología del Estudio de Movilidad con bigdata». Consulta: 20 juny 2024. [En línia]. Disponible a: https://www.transportes.gob.es/ministerio/proyectos-singulares/estudios-de-movilidad-con-big-data/metodologia-del-estudio-de-movilidad-con-bigdata

[27] TERMCAT. Centre de Terminologia, «Definició del terme incidència». Consulta: 23 juny 2024. [En línia]. Disponible a: https://www.termcat.cat/es/cercaterm/fitxa/NDQ1OTE3OQ%3D%3D

[28] Y. Dodge, «Kolmogorov–Smirnov Test», en *The Concise Encyclopedia of Statistics*, New York, NY: Springer New York, p. 283-287. doi: 10.1007/978-0-387-32833-1_214.

[29] A. Hayes, «Multiple Linear Regression (MLR) Definition», Investopedia.

[30] Minitab, «Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?», *Regression Analysis*, 2013.

[31] H. Coolican, «Regression and multiple regression», en *Research Methods and Statistics in Psychology*, 2018. doi: 10.4324/9781315201009-20.

[32] A. Serrano-Cumplido *et al.*, «Efecto de la variante Ómicron sobre la incidencia y la letalidad durante la 6.ª onda epidémica COVID-19 en España», *Medicina de Familia. SEMERGEN*, vol. 50, núm. 2, 2024, doi: 10.1016/j.semerg.2023.102073.

[33] Centro Nacional de Epidemiologia del Instituto de Salud Carlos III de España, «Situación y evolución de la pandemia de COVID-19 en España». Consulta: 20 juny 2024. [En línia]. Disponible a: https://cnecovid.isciii.es/covid19/

[34] E. López-Bazo, «The complex link between socioeconomic deprivation and COVID-19. Evidence from small areas of Catalonia», *Spat Spatiotemporal Epidemiol*, vol. 49, p. 100648, juny 2024, doi: 10.1016/j.sste.2024.100648.

[35] E. Roel, B. Raventós, E. Burn, A. Pistillo, D. Prieto-Alhambra, i T. Duarte-Salles, «Socioeconomic Inequalities in COVID-19 Vaccination and Infection in Adults, Catalonia, Spain», *Emerg Infect Dis*, vol. 28, núm. 11, 2022, doi: 10.3201/eid2811.220614.

[36] M. A. Barceló, X. Perafita, i M. Saez, «Spatiotemporal variability in socioeconomic inequalities in COVID-19 vaccination in Catalonia, Spain», *Public Health*, vol. 227, 2024, doi: 10.1016/j.puhe.2023.11.024.