Universitat de Girona

MODELS ECONOMÈTRICS EN ENTORNS BIG DATA ENFOCATS A DESIGUALTATS

Xavier Perafita Basart

MODELS ECONOMÈTRICS EN ENTORNS BIG DATA ENFOCATS A DESIGUALTATS SOCIALS · **XAVIER PERAFITA BASART** · 2023

Tesi Doctoral

Xavier Perafita Basart

2023

Universitat de Girona

MODELS ECONOMÈTRICS EN ENTORNS BIG DATA ENFOCATS A DESIGUALTATS

MODELS ECONOMÈTRICS EN ENTORNS BIG DATA ENFOCATS A DESIGUALTATS

Tesi Doctoral

MODELS ECONOMÈTRICS EN ENTORNS BIG DATA ENFOCATS A DESIGUALTATS

Xavier Perafita Basart

2023

Programa de Doctorat en Dret, Economia i Empresa

Dirigit per:

Dr. Marc Saez Zafra

Memòria presentada per optar al títol de doctor per la Universitat de Girona

El **Prof. Dr. Marc Saez Zafra**, de la Universitat de Girona i membre del Grup de Recerca en Estadística, Econometria i Salut (GRECS),

CERTIFICO:

Que el treball titulat *Models economètrics en entorns Big Data enfocats a desigualtats*, que presenta en **Xavier Perafita Basart** per a l'obtenció del títol de doctor, ha estat realitzat sota la seva direcció.

Que la tesi es presenta com un compendi d'articles, indicant la idoneïtat d'aquest formati demostrant la rellevància de la contribució específica del doctorand a les publicacions presentades.

Tanmateix, la tesi reflecteix fidelment el treball realitzat pel doctorand, que ha estatelaborada d'acord amb el codi de bones pràctiques de l'Escola de Doctorat i que no conté cap element plagiat.

I, perquè així consti i tingui els efectes oportuns, signo aquest document.

Prof. Dr. Marc Saez Zafra

Universitat de Girona, Girona

Girona, 12 de setembre de 2023

AGRAÏMENTS

A l'**Albert**, el meu pare, per ensenyar-me com, tot i les dificultats a la vida, hi ha valors que mai s'han de perdre i que al final són aquests els que ens defineixen com a persona. Gràcies per insistir quan tots els professors ens deien que no "servia per això d'estudiar". Per ser primer de tot pare i ara amic i per no acabar de perdre la paciència amb el meu caos en el que tant m'agrada viure. A la **Marta**, la meva Gogo, que, tot i que ens matem, saps que en el fons t'estimo, però no tornis en una temporada no sigui que canviï d'opinió. A la **Tresa**, la meva mare, que, tot i que no hi és, a vegades tinc la sensació que no ha marxat, realment em vas ensenyar més del que ningú altre hauria pogut fer.

Gràcies al meu director de tesis, del qual he après i continuo aprenent. **Marc**, crec que no hauria pogut trobar un director que em cuidés i m'ajudés tant!!. Gràcies per creure tant en mi, en què jo podia i que puc, en dir-me doctor abans de ni tan sols haver començar el doctorat i, no pateixis, que ja no me'n recordo de quan em canviaves el cognom. No crec que mai tingui suficients paraules per agrair-te tot el que has fet per mi i tot el temps que em dediques.

Als meus companys de l'Observatori, crec que no havia coincidit amb un grup de persones que eren tan diferents de mi i amb el temps m'he anat nodrint de les seves essències. Gràcies **Laura**, per ensenyar-me de la teva passió i el teu sentit de justícia. A les meves dues companyes de cafès: a tu **Alba**, gràcies per ajudar-me a ser més pacient i a la vegada rebel. A tu **Angi**, gràcies per ensenyar-me de la teva passió i la vida sana, encara que ja saps que tinc debilitats. **Berta**, gràcies per ensenyar-me la solució més pràctica als meus problemes i a tu **Nerea** per veure la senzillesa de les coses. Gràcies **Pau**, per ensenyar-me que s'atrapen més mosques amb una gota de mel que amb un litre de vinagre.

Als meus amics, en especial a en **Martí** i les seves frases filosòfiques que no tenen fi i a la **Sílvia**, per fer-me sentir com un més de la seva família. També us agreixo infinit permetre'm ser part de la vida de la coseta més maca del món que es diu **Irina**. A en **Claudi**, per ser una persona amb la qual puc estar igual que aquell primer dia d'institut que ens vam asseure junts als seients de l'autobús. A

la **Laia**, per sempre veure la part positiva de les coses. A l'**Iria**, la meva gallegiña, gràcies per com diries tu "fer-me" tot i que ja saps que no és veritat.

Tinc la sort, de poder compartir moments amb moltes persones especials, intentaré sintetitzar. **Pau**, per ser el meu torturador preferit i jo sé que soc el teu pupes preferit; **Núria** per la teva creativitat i enfadar-te cada 5 frases meves; **Pol** gràcies per ajudar-me amb les coses més bàsiques; **Núria** gràcies per les xerrades a l'oficina arreglant al món; **Maria** gràcies per ensenyar-me que es pot estimar a tots els éssers vius; **Jess** gràcies per ensenyar-me que tinc un rerefons espiritual; **Cesc** i l'**Edu** per ajudar-me a treure el competidor que porto dins. A l'equip (**Silvia**, **Carla** i **Maribel**) de la Càtedra de Promoció de la Salut, per ajudar-me i assessorar-me. Als demés, no me n'he oblidat, us ho agrairé en persona i així tenim una excusa per veure'ns.

També vull agrair a la infinitat de companys i entrenadors que he tingut al llarg de la meva vida, que m'han ajudat de fer d'un esport la meva passió i durant molt de temps la meva obsessió. En especial aquells, amb els quals he compartit moments que m'han ajudat a ser una persona constant i lluitadora.

Tinc una amiga, que és la meva companya de sopars, juemingos i que m'acaba enredant amb qualsevol excusa, que sempre em diu que les persones arriben a la teva vida per una finalitat, les que s'han de quedar es queden i les que han de marxar, és perquè ja han fet la seva funció a la teva vida. Aquestes persones que han marxat, que segurament mai llegiran això, també m'agradaria agrair-los-hi que hi hagin sigut.

Per acabar, hi ha una frase que fem servir en totes les celebracions amb la meva família i que també voldria fer en aquesta tesi:

"Pels presents i pels absents"

Xevi

LLISTAT DE PUBLICACIONS

Aquesta tesi es presenta com a un compendi de dues publicacions, les quals es presenten a continuació:

ARTICLE 1

Títol: Clustering of Small Territories Based on Axes of Inequality

Autors: Perafita X, Saez M

Resum: Antecedents: en el present article, hem realitzat un estudi abans de crear una e-cohort per al disseny de la mostra. Aquesta e-cohort ha de possibilitar la representació efectiva de la província de Girona per facilitar-ne l'estudi segons els eixos de desigualtat. Mètodes: El territori objecte d'estudi es divideix per municipis, tenint en compte els diferents eixos. L'estudi consisteix en una comparació de 14 algoritmes d'agrupació, juntament amb 3 conjunts de dades d'informació municipal per detectar l'agrupació que era més consistent. Abans de realitzar el clustering, es va realitzar un procés de selecció de variables per descartar aquelles que no fossin útils. La comparació es va dur a terme seguint dos eixos: resultats i representació gràfica. Resultats: Es van analitzar els resultats intra-clúster per observar la coherència de l'agrupació. Finalment, hem estudiat la probabilitat de pertànyer a un clúster com el de la capital. Conclusions: Aquesta agrupació pot ser la base per treballar amb una mostra significativa i representativa del territori.

Revista: International Journal of Environmental Research and Public Health. 2022; 19(6):3359. **DOI**: 10.3390/ijerph19063359

Impact Factor (2021): 4.614 (Q1 Public Health, Environmental and Occupational Health, posició 139 de 562).

ARTICLE 2

Títol: Housing Supply and How It Is Related to Social Inequalities—Air Pollution, Green Spaces, Crime Levels, and Poor Areas—In Catalonia

Autors: Perafita X, Saez M

Resum: Hem fet una recerca de més de 12.000 habitatges que s'ofereixen al mercat de lloguer a Catalunya i hem valorat la possibilitat que les famílies per sota del llindar de pobresa puguin llogar aquests habitatges. En aquest sentit, s'ha volgut avaluar si la situació econòmica de les famílies és capaç d'incidir en el seu entorn social, ambiental i de seguretat. Hem observat com la seva situació econòmica pot permetre a les famílies la possibilitat de desenvolupar una vida sense exposició a riscos per a la salut, i com les limitacions econòmiques generen desavantatges en diversos àmbits de la vida. Els resultats mostren com les famílies en risc de pobresa viuen en condicions menys favorables i experimenten una ampliació de diferents bretxes, amb els preus actuals que condueixen a una possible trampa de pobresa per als col·lectius més desfavorits. Com més gran és el percentatge de població per sota del llindar, menor és la possibilitat de no poder llogar un habitatge en comparació amb les zones amb menor prevalença de població per sota del llindar. Aquesta associació es va observar tant en considerar el risc de manera lineal com no lineal. Linealment, la probabilitat de no llogar un habitatge es va reduir un 8,36% per cada 1% d'augment de la taxa de població en risc de pobresa extrema. En el segon, tercer i quart quartils, la probabilitat de no poder llogar un habitatge va disminuir un 21,13%, un 48,61% i un 57,79%, respectivament. A més, l'efecte va ser diferent dins i fora de les àrees metropolitanes, en la primera mostra una disminució del 19,05% en la probabilitat de lloguer d'un habitatge, mentre que fora de les àrees metropolitanes la probabilitat va augmentar un 5,70%.

Revista International Journal of Environmental Research and Public Health 2023, 20(8), 5578. **DOI**: https://doi.org/10.3390/ijerph20085578

Impact Factor (2021): 4.614 (Q1 Public Health, Environmental and Occupational Health, posició 139 de 562).

Altres publicacions d'interès que s'han realitzat durant la tesi es mostren en la següent caixa i es detallen en l'**Annex I**: Altres publicacions relacionades durant el període de la tesi.

Caixa 1. Altres publicacions d'interès

2020

1. Promoting equity through monitoring inequalities in the semi-rural region of Girona.
2. Plan municipal de salud, bienestar y desarrollo sostenible en una comunidad rural: Práctica a nivel micro de promoción de la salud y la equidad a través de los datos del Observatorio en desigualdades sociales y de la salud.
3. Promoviendo la equidad a partir de la monitorización de las desigualdades en la región de Girona: Diagnóstico participado de necesidad de datos sobre desigualdades para el fomento de la equidad en áreas rurales.

2021

4. Metodologia de treball de les dades sobre cobertures del sòl per la realització d'indicadors del àmbit Medi i entorn de l'[O]bservatori.
5. Procés participatiu per a la identificació de necessitats als municipis en matèria d'indicadors de salut i desigualtat social.
6. Metodologia de treball de les bases de dades Mortalitat de l'Institut Nacional d'Estadística (INE). Unificació per la creació dels indicadors de salut de l'[O]bservatori de Desigualtats Socials i de Salut.
7. Promoting Equity through Monitoring Inequalities in the Semi-rural Region of Girona. Participatory Process to Identify Municipalities' Needs for Data and Information.

2022

8. Atlas de los determinantes sociales de la salud en España. Evolución y variabilidad entre Comunidades Autónomas. Comunidades Autónomas.

9. Observatorio sobre determinantes sociales y desigualdades en salud y bienestar.
10. Atlas de determinantes sociales de la salud en España: evolución y variabilidad entre las comunidades autónomas.

2023

11. Contaminants i desigualtats - C[O]NTAMINANTS
12. IndiMuniDem (IMD)
13. GirTrans (GiT)
14. Spatiotemporal variability in socioeconomic inequalities in vaccination against COVID-19 in Catalonia.
15. El atlas de los determinantes sociales de la salud en España y sus CCAA 2022 .
16. Spatiotemporal variability in socioeconomic and environmental inequalities in vaccination against COVID-19 in Catalonia, Spain.
17. Spatiotemporal variability in socioeconomic inequalities in vaccination against COVID-19 in Catalonia, Spain

LLISTAT D'ABREVIATURES

AGNES: Agglomerative Nesting

AMB: Àrea Metropolitana de Barcelona

AMS: Alternative Model Selection

BICO: Balanced Iterative Reducing and Clustering using Hierarchies meets coresets for k-means clustering

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies

BR: Bridge Regressio

CART: Classification and regression trees

CLARA: Clustering Large Applications

CLARANS: Clustering Large Applications based on Randomized Search

DBN: Deep Belief Networks

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

DENCLUE: Density-Based Clustering

DIANA: Divisie Analysis

Dipsalut: Organisme de Salut Pública de la Diputació de Girona

DSS: Determinants socials de la salut

ETL: Extraure, transformar i carregar (Extract, Transform, Load)

GAN: Generative Adverse Networks

IA: Intel·ligència artificial

K-NN: K-nearest neighbour classification

LARS: Least angle regression

LASSO: Least Absolute Shrinkage and Selection Operator

LOF: Local outlier factor

MAP: Màxims a posteriori

MCMC: Cadenes de Markov Monte Carlo

MCP: Minimax Concave Penalty

NNG: Non-negative Garrote

OPTICS: Ordering Points To Identify the Clustering Structure

PAM: Partitioning Around Medoids

QBLL: Quasi-maximum likelihood estimator

QMLE: Quasi-Bayesian local likelihood

QR: Quantile regressions

RBM: Restricted Boltzmann machine

RR: Ridge Regressio

RSS: Suma Residual de Quadrats

SCAD: Smoothly Clipped Absolute Deviation

SNN: Shared Nearest Neighbour

ZB: Zettabyte

LLISTAT DE FIGURES, TAULES I CAIXES

Llistat de figures

Llistat de taules

Llistat de caixes

ABSTRACT

In the last twenty years, the study of social inequalities has been a topic of special interest, revealing how differences between social classes have increased. During the same period, data generation has exponentially grown, leading to the normalization of the term "Big Data," which is widely recognized by the general population. This thesis examines the econometric mechanisms used in the management of massive data and applies them to social inequalities.

To achieve this, existing mechanisms have been identified through a systematic review, which revealed five types of algorithms: classifiers, penalized classifiers, quasi-Bayesian penalized classifiers, clusters, and neural networks.

Based on this review, the results of different algorithms have been utilized and compared to perform the clustering of municipalities in the province of Girona. The best grouping has served as the foundation for designing the sampling of the cohort driven by the Social, Environmental, and Health Observatory of Dipsalut.

To study the inequality of vulnerable families in Catalonia, data on the typology of households and their environments have been obtained to analyze the living conditions of these families. The market they have access to is more limited, where the characteristics of the homes do not play a significant role in the possibility of renting housing; instead, the environment determines where they can live. These environments are more insecure, unhealthy, and impoverished, highlighting how low-income families are forced to reside in settings that exacerbate the poverty trap. Additionally, the study demonstrates that different logics regarding inequalities exist in urban and rural areas, which should be considered when designing and implementing any policies addressing inequality.

RESUM

En els últims vint anys, l'estudi de les desigualtats socials han estat un tema d'especial interès, mostrant com les diferències entre les classes socials han augmentat. En el mateix període, la generació de dades ha augmentat de manera exponencial, provocant la normalització del terme Big Data, el qual és reconegut per la població general. Aquesta tesi estudia quins són els mecanismes economètrics que s'utilitzen en la gestió de dades massives i ho aplica a les desigualtats socials.

Per a fer-ho, s'han detectat els mecanismes existents a través d'una revisió sistemàtica, en la qual, s'han detectat cinc tipus d'algorismes: classificadors, classificadors amb penalització, classificadors gairebé bayesians amb penalització, agrupadors i les xarxes neuronals.

A partir d'aquesta revisió, s'han utilitzat i comparat els resultats de diferents algorismes per a realitzar la clusterització dels municipis de la província de Girona. La millor agrupació s'ha usat com a base pel disseny del mostreig de la cohort que impulsa l'Observatori Social, Ambiental i de Salut del Dipsalut.

Per a estudiar la desigualtat a Catalunya de les famílies vulnerables, s'ha obtingut dades sobre la tipologia de llars i els seus entorns per a estudiar en quina situació poden viure aquestes famílies. El mercat al qual tenen accés és més reduït, on les característiques de les llars no juguen un paper significatiu en la possibilitat de llogar o no l'habitatge i acaba sent l'entorn el que determina on poden viure. Aquests entorns són més insegurs, insalubres i pobres. Mostrant com les famílies amb baixos recursos, han de viure en entorns que acreixen el parany de la pobresa. A més, es demostra que existeixen diferents lògiques sobre les desigualtats en àrees urbanes i en àrees rurals, les quals s'haurien de tenir en compte en l'hora de dissenyar i aplicar qualsevol política sobre desigualtats.

RESUMEN

En los últimos veinte años, el estudio de las desigualdades sociales ha sido un tema de especial interés, mostrando cómo las diferencias entre las clases sociales han aumentado. En el mismo período, la generación de datos ha aumentado de manera exponencial, provocando la normalización del término "Big Data", el cual es reconocido por la población en general. Esta tesis estudia cuáles son los mecanismos econométricos que se utilizan en la gestión de datos masivos y los aplica a las desigualdades sociales.

Para ello, se han identificado los mecanismos existentes a través de una revisión sistemática, en la cual se han detectado cinco tipos de algoritmos: clasificadores, clasificadores con penalización, clasificadores casi bayesianos con penalización, agrupadores y redes neuronales.

A partir de esta revisión, se han utilizado y comparado los resultados de diferentes algoritmos para realizar la clusterización de los municipios de la provincia de Girona. La mejor agrupación se ha utilizado como base para el diseño del muestreo de la cohorte que impulsa el Observatorio Social, Ambiental y de Salud del Dipsalut.

Para estudiar la desigualdad en Cataluña de las familias vulnerables, se han obtenido datos sobre la tipología de hogares y sus entornos para estudiar en qué situación pueden vivir estas familias. El mercado al que tienen acceso es más reducido, donde las características de los hogares no juegan un papel significativo en la posibilidad de alquilar o no la vivienda, siendo el entorno el que determina dónde pueden vivir. Estos entornos son más inseguros, insalubres y pobres, lo que muestra cómo las familias de bajos recursos deben vivir en entornos que aumentan la trampa de la pobreza. Además, se demuestra que existen diferentes lógicas sobre las desigualdades en áreas urbanas y áreas rurales, las cuales deben tenerse en cuenta a la hora de diseñar y aplicar cualquier política sobre desigualdades.

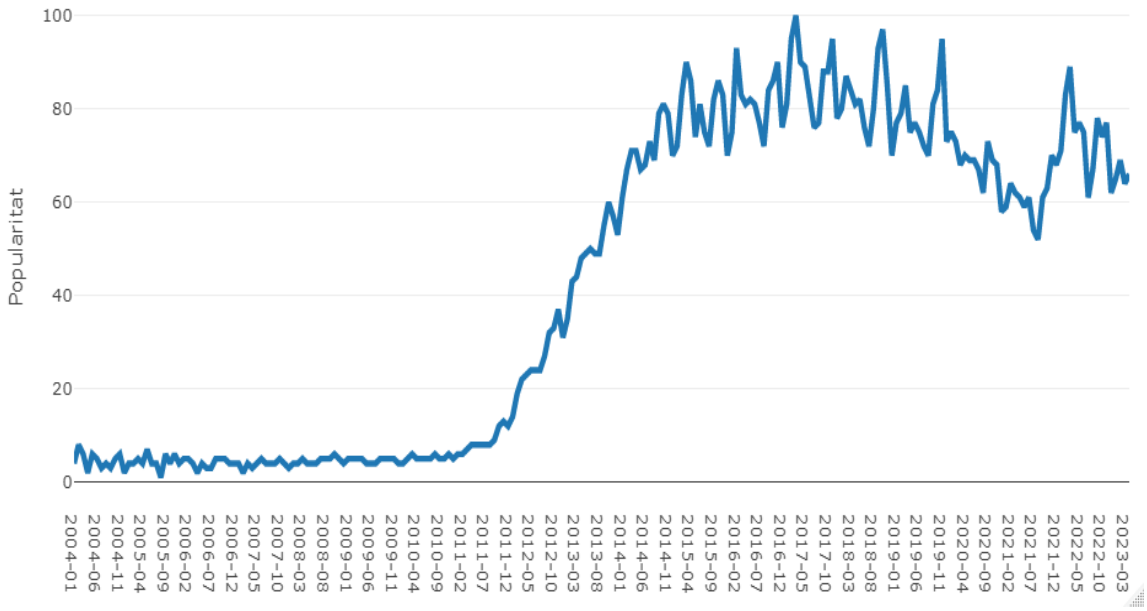
TAULA DE CONTINGUTS

1 INTRODUCCIÓ

1.1 Big Data

El terme *Big Data* ha augmentat la seva popularitat de forma dràstica en les últimes dècades. Sobretot en l'última, des de la implementació en l'àmbit industrial[1]. La **Figura 1** mostra el grau de popularitat del terme a escala mundial. La popularitat es mesura respecte a totes les cerques realitzades en el buscador de Google.

Figura 1. Popularitat del terme "Big Data" en les cerques del buscador de Google a escala mundial.

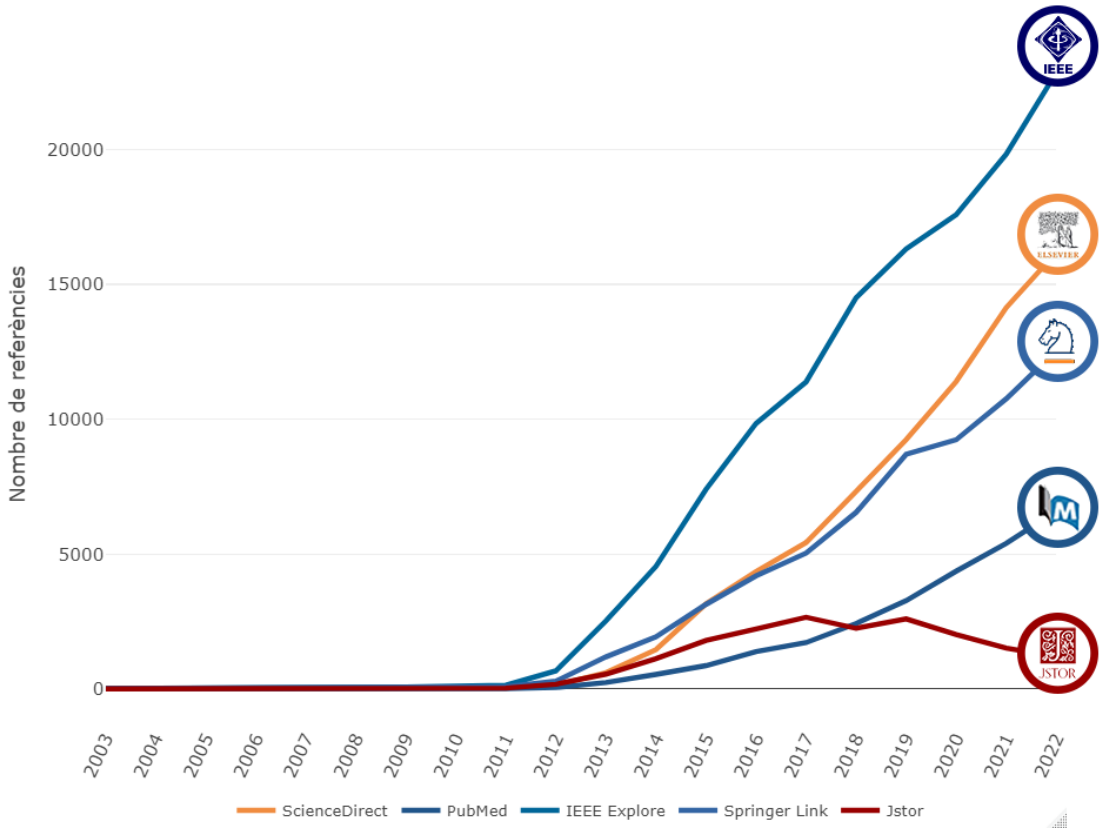


Font: Gràfic d'elaboració pròpia a partir de les dades extretes del software Google Trends2. La màxima popularitat d'un terme, es mostra respecte al total de cerques que es realitzen en la mateixa regió i període, es valora com a 100. Una popularitat residual o insuficient es valora com a 0.

En l'àmbit acadèmic, quan s'utilitza el terme "Big Data" com a paraula clau de cerca, s'observa un augment similar al vist en la **Figura 1**. En la **Figura 2** es mostra l'evolució del terme en les bases de dades literàries: PubMed, ScienceDirect, Jstor, IEE Explore i SpringerLink.

No sorprèn que el concepte cada cop estigui més universalitzat, ja que el volum de dades amb què es treballa, en tots els àmbits, augmenta constantment.

Figura 2. Resultats de terme "Big Data" en les cerques a bases de dades literàries.

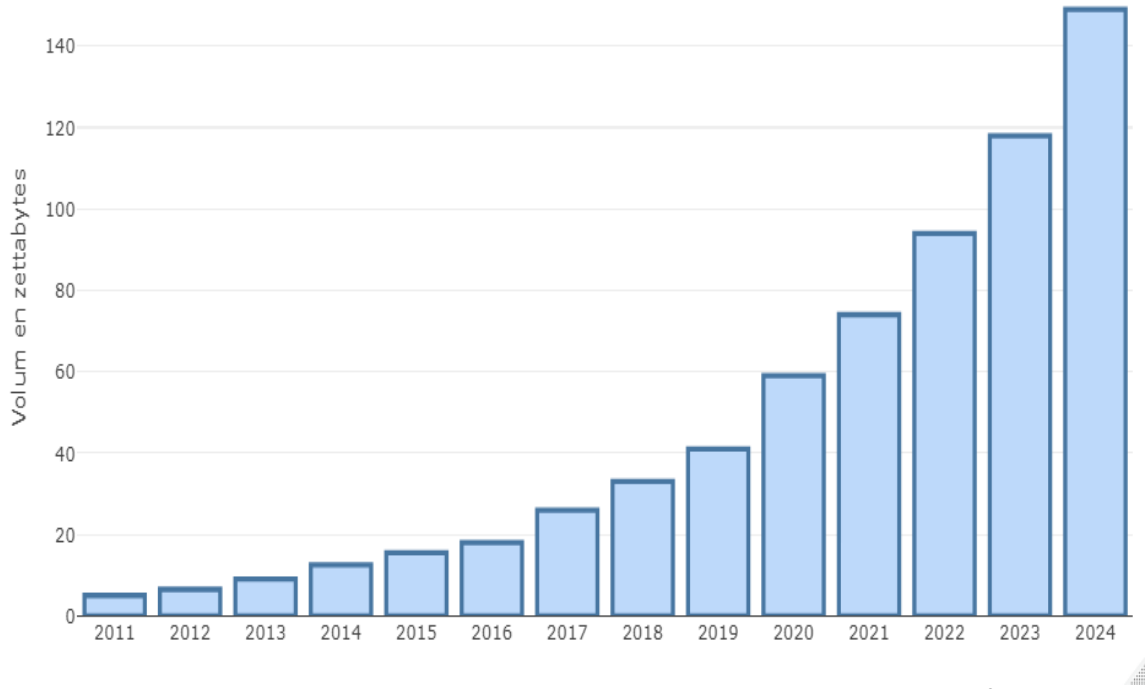


Font: Gràfic d'elaboració pròpia a partir del recull del nombre de publicacions que mostren les bases de dades: PubMed, ScienceDirect, Jstor, IEE Explore i SpringerLink.

D'acord amb les estimacions del International Data Group Forecasts[3], veure **Figura 3**, el volum estimat de dades i informació creades, capturades, copiades i consumides a tot el món en el 2010 era de 2 zettabytes. El 2015 aquest volum estimat va augmentar fins els 18 zetabytes. El 2020, el volum gairebé es va triplicar respecte al 2015, estimant-se el volum total en 59 zetabytes. El 2024 es preveu que el volum generat sigui gairebé el triple que del 2020. Amb un volum total de 149 zetabytes.

Aquest augment exponencial del volum de dades, presenta un escenari que requereix l'ús de tècniques i tecnologies enfocades a l'optimització dels processos. La generació de dades de forma massiva depèn que els algoritmes siguin més adaptables a diferents situacions. Aquesta adaptabilitat condueix a l'aparició d'estructures més flexibles que permeten la incorporació de noves i diferents fonts de dades[4].

Figura 3. Volum estimat de dades creades, capturades, copiades i consumides a tot el món del 2010 al 2020 i estimat fins al 2025



Font: Gràfic d'elaboració pròpia a partir de les estimacions del International Data Group Forecasts, Statista[3].

1.1.1 Les bases prèvies al Big Data

En l'últim segle, sobretot a partir dels anys seixanta, l'ús i processament de dades ha anat prenent cada cop més pes fins a arribar als entorns Big Data actuals. De fet, la importància d'emmagatzemar, llegir i aprendre de les dades és inherent a l'ésser humà. Els antics egipcis van intentar emmagatzemar el coneixement més gran, guardat en format físic, a la biblioteca d'Alexandria l'any 300[1]. Els romans utilitzaven les estadístiques per analitzar les seves estratègies militars i determinar quina era la millor distribució de la seva força militar per obtenir millors resultats en les guerres.

El 1662, John Graunt va crear un dels primers sistemes d'emmagatzematge d'informació estructurada[5]. El sistema consistia en unes taules de vida sobre les diferents causes de mortalitat i altres fenòmens vinculats la ciutat de Londres. Aquest sistema va permetre generar un mecanisme d'alerta per la pesta bubònica que estava devastant Londres[6]. Més tard, el 1854, John Snow va desenvolupar un sistema per detectar casos de còlera[6,7], també a la ciutat de Londres. Snow va representar sobre un mapa de la ciutat totes les morts de

còlera. Així va poder determinar els focus d'infecció i va obtenir un patró sobre la propagació de la mortalitat. Utilitzant la visualització de dades com a eina de detecció de patrons.

El primer projecte a gran escala va ser ordenat l'any 1937 per l'administració de Franklin D. Roosevelt als Estats Units. El projecte consistia a realitzar per primera vegada el seguiment de les nòmines a 26 milions de nord-americans així com un identificador per més de 3.500 empreses[8]. Uns anys més tard es va inventar el primer ordinador.

Una de les primeres eines de processament de dades es va crear enmig de la Segona Guerra Mundial i va ser desenvolupada pels anglesos per desxifrar el codi enemic. Aquesta eina es va anomenar Colossus[9,10] i va ser capaç de cercar patrons a una velocitat de 5.000 caràcters per segon. En aquest punt es va crear una eina computacional primitiva que va ja podia crear una gran base de dades.

En les dues últimes dècades, les noves tecnologies han permès augmentar la velocitat, la varietat i el volum de dades que es poden crear i gestionar. Això ha permès als humans superar les seves limitacions en la manipulació de dades i augmentar la seva comprensió de l'entorn. Tot i l'evolució dels mecanismes, aquestes noves tecnologies es basen en teoremes o eines clàssiques. Cada cop hi ha més camps on s'apliquen aquests nous mecanismes, i per tant el desenvolupament, adaptació i variació d'eines i tècniques és constant. Aquests mecanismes han de poder-se adaptar a: noves necessitats, noves estructures relacionals de dades, noves tipologies de dades i indicadors.

1.1.2 Definició

En l'àmbit acadèmic s'atribueix l'encunyació del terme *Big Data* a Magoulas l'any 2005[11]. Tot i que prèviament es troben documents que ja analitzaven el fenomen i les problemàtiques associades a aquest[12]. Fora de l'àmbit acadèmic, també podem trobar publicacions anteriors a Magoulas, que utilitzaven el concepte *Big Data* per definir macroconjunts de dades[13,14].

El terme *Big Data* defineix aquells entorns on els conjunts de dades a computar són tan grans o complexes de treballar que els programes tradicionals de gestió de dades ho fan de forma ineficient o directament no els poden processar. Aquests conjunts de dades es poden presentar amb diferents formats: estructurats, semiestructurats o no estructurats i amb diverses tipologies d'informació.

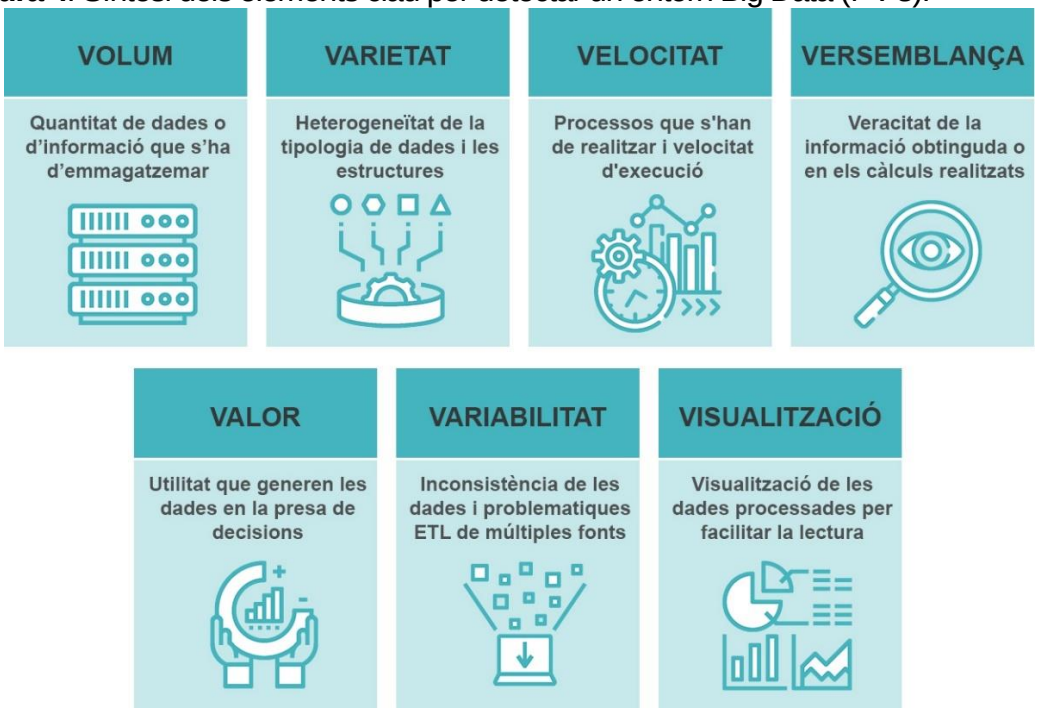
1.1.3 Característiques dels entorns Big Data

Per detectar un entorn Big Data, aquest necessita complir diferents requisits, veure **Figura 4**. El 2001, Lanley va definir les tres característiques principals que havia de tenir un entorn de dades per poder-se definir com un entorn Big Data: Volum, Varietat i Velocitat[15]. Aquests tres elements, coneguts com les tres V[16], representen un repte comú en la majoria dels projectes que es basen en la gestió de dades.

El volum es refereix a la quantitat de dades o informació que haurà de processar el sistema. La varietat consisteix a analitzar l'heterogeneïtat de la tipologia de dades i l'estructura en què aquestes es troben. La velocitat observa el nombre de processos que s'han de realitzar amb les dades i la velocitat a la qual s'han d'executar. En implementar les primeres infraestructures Big Data en l'àmbit industrial apareixen dos elements nous: la versemblança i el valor[17–21]. La versemblança es refereix a la veracitat que té aquella informació que s'està capturant i quin grau d'incertesa incorpora. El valor és el coneixement o la presa de decisions que se'n deriven d'aquesta informació. Actualment, estan acceptats dos elements més: la variabilitat i la visualització[22]. La variabilitat fa referència a la inconsistència de les dades per la presència de valors erronis o alterats i als problemes derivats de les càrregues de dades de múltiples fonts. L'últim element analitza com es poden visualitzar les dades una vegada processades.

Actualment, existeixen altres característiques alternatives que es poden tenir en compte pels entorns Big Data: Volatilitat, Virtual, Visualització, Complexitat, Viscositat o Gerro[16,23].

Figura 4. Síntesi dels elements clau per detectar un entorn Big Data (*7 V's*).



Font: Gràfic d'elaboració pròpia.

1.1.4 La mineria de dades i el Machine Learning

Una de les parts més atractives i claus dels entorns Big Data és la gestió i processament de grans quantitats de dades. Aquesta funció es coneix com a mineria de dades i consisteix en el processament de dades per detectar patrons, tendències, interaccions o aprofundir en les relacions de les dades.

El concepte de la mineria de dades està estretament lligat a l'estadística i a les ciències de la computació, en especial a les tècniques de Machine Learning. Aquesta connexió dual és especialment evident en l'estadística inferencial, on les tècniques de Machine Learning permeten incorporar les complexitats computacionals als models d'intel·ligència artificial supervisats i no supervisats, generant solucions que permeten una gran flexibilitat a l'hora d'incorporar noves fonts de dades.

Arthur Samuel, va ser l'autor que va encunyar el terme Machine Learning a partir d'un exemple on connectava un sistema primitiu que aprenia d'una base de dades senzilla[24]. Amb aquest exemple, es va poder veure el potencial que tindria un sistema que aprengués d'una gran base de dades amb informació de valor.

Tota aquesta infraestructura necessita ser acompanyada d'eines estadístiques que permetin generar un aprenentatge eficient, de valor i significatiu[25,26].

1.2 Econometria

Històricament, els economistes han tractat amb dades per realitzar els seus estudis. Aquests estudis requereixen elaborar anàlisis d'informació a través de l'anàlisi de dades. Actualment, la majoria d'eines enfocadades al tractament de grans volums d'informació, treballen en quatre categories: resum, estimació, prova d'hipòtesis i predicció[27], veure **Figura 5**.

Figura 5. Categories de l'anàlisi d'informació a través del processament de dades.

Font: Gràfic d'elaboració pròpia a partir de l'article de Varian[27]

1.2.1 Definició

L'econometria és la branca de l'economia que aplica mètodes estadístics i matemàtics per estudiar i quantificar fenòmens basats en teoremes econòmics amb la finalitat de verificar-los o rebutjar-los.

La primera aparició del terme "Econometria", dins de l'àmbit acadèmic, l'utilitza Ciompa el 1910[28,29]. El terme, com es va definir en aquell moment, no té res a veure amb el que es coneix en l'actualitat. Ciompa utilitzava al terme per referir-se a les regles matemàtiques de la comptabilitat. El 1936, Ragnar Frisch encunya i desenvolupa el concepte com es coneix avui en dia[30]. Frisch, juntament amb Tinbergen, van desenvolupar diverses formulacions matemàtiques en l'àmbit de l'economia. Tots dos van ser reconeguts amb el Premi Nobel d'Economia, per haver desenvolupat i aplicat models dinàmics a l'anàlisi dels processos econòmics.

1.2.2 Interpretació de la probabilitat

En estadística, la probabilitat té un gran pes. Tanmateix, no existeix un consens sobre com s'interpreta. D'aquesta forma podem trobar dos corrents interpretatius. Per un costat, tenim les interpretacions freqüentistes que interpreten la probabilitat com la freqüència de resultats seqüencials que s'obtenen d'uns successos, assaigs o fenòmens idèntics al llarg del temps. Per l'altre, trobem les interpretacions bayesianes. Els bayesians incorporen la incertesa, interpretant la probabilitat com una probabilitat condicionada a un esdeveniment aleatori en funció d'un segon esdeveniment en cadascuna de les seqüències.

Tot i que l'estadística Bayesiana es va desenvolupar un segle abans que la freqüentista, no és fins a mitjans del segle XX que va començar a prendre protagonisme. Al segle XIX, amb Laplace[31–33] al capdavant es van desenvolupar les bases de l'estadística bayesiana, prèviament teoritzada per Thomas Bayes[34]. Durant el següent segle, va sorgir el paradigma de l'estadística freqüentista on Fisher[35–37] hi va prendre un paper destacat. El desenvolupament de la ciència de la computació i el descobriment de nous mètodes matemàtics van permetre recuperar pes a l'estadística Bayesiana. A partir de la dècada dels cinquanta es va desenvolupar el mètode Markov Chain Monte–Carlo (MCMC)[38,39]. Aquestes cadenes aleatòries permeten generar mostreigs aleatoris a partir d'una distribució de probabilitat i van permetre el desenvolupament dels models jeràrquics bayesians i posteriors mètodes basats en l'enfocament Bayesià.

Els dos corrents tenen avantatges i desavantatges. L'estadística freqüentista no calcula la probabilitat d'hipòtesis mentre que la bayesiana utilitza les probabilitats de les dades i de les hipòtesis. Els mètodes freqüentistes no requereixen una distribució a priori, en canvi, els models bayesians necessiten especificar-la.

Les eines de Big Data, sobretot les enfocades a l'aprenentatge automàtic, utilitzen aproximacions bayesianes, ja que tenen avantatges tant teòrics com pràctics. En primer lloc, respon a l'excés d'ajustaments i permet automatitzar la selecció dels hiperparàmetres. Els estimadors màxims a posteriori (MAP),

permeten a les eines treballar de forma eficient amb dades sobreparametritzades. Aquesta sobreparametrització apareix per la falta de bases de dades de bona qualitat. També permet capturar la incertesa dels paràmetres, generar prediccions i realitzar comparacions quantitatives entre models alternatius.

1.2.3 Inferència de prediccions estadístiques i heteroscedasticitat

La funció principal de l'estadística, en les eines Machine Learning, és la generació d'estimacions en funció d'unes dades que s'estan constantment actualitzant per generar prediccions el més acurades possibles[27,40]. Aquestes estimacions permeten generar un valor resum el qual facilita la interpretació del fenomen i genera un coneixement d'interès. A més, pot aportar noves perspectives o patrons de fenòmens de qualsevol àmbit d'investigació.

A partir d'aquestes inferències es poden generar els models predictius que és on les metodologies Big Data i Machine Learning aprofiten l'eficiència computacional[27]. Les dades són recollides i sistemàticament reutilitzades per validar els valors predits pels models, buscant un model, que pugui predir amb molta exactitud la informació que entrarà al sistema[41].

Aquests models, tot i l'estabilitat que busquen per generar prediccions, es poden veure afectats per múltiples factors[42,43]. Destaquen tres: soroll en les dades, addició de noves tipologies de dades no considerades prèviament i canvi de patró del fenomen predit. Qualsevol d'aquests tres elements pot afectar a les estimacions i prediccions generades per la IA[44].

L'estabilitat i la reproductibilitat de les modelitzacions requereixen homogeneïtat i independència[41,42]. Les dades a gran escala tenen major facilitat de violar el supòsit d'homogeneïtat[45]. Aquest supòsit es viola generalment per la integració de múltiples fonts de dades en diferents formats. Això es coneix com a heteroscedasticitat sintàctica[45,46]. També es pot veure violat per l'heteroscedasticitat semàntica[47,48], que pot aparèixer amb la unió de diferents bases de dades, quan presenten variables amb diferents interpretacions o significats pel mateix fenomen.

1.3 Machine Learning

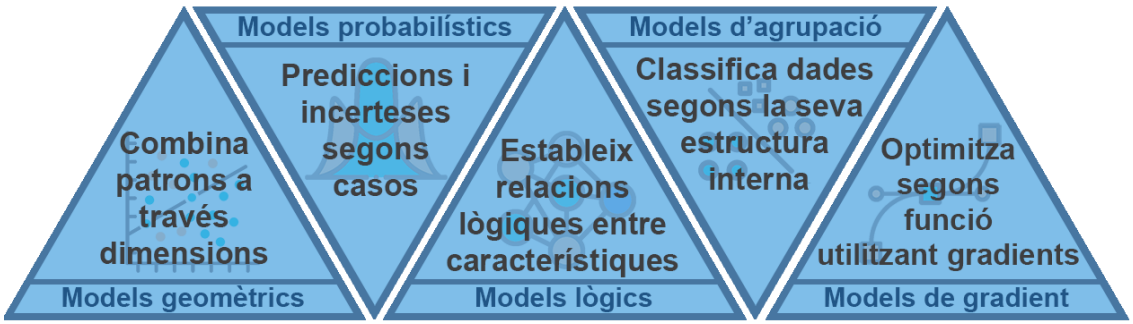
1.3.1 Definició

La primera persona que va encunyar el terme Machine Learning va ser l'autor Artur Samuel[24], l'any 1959. Samuel va definir en aquell moment el concepte Machine Learning com "el camp d'estudi que dota als ordinadors de la capacitat d'aprendre sense ser programats explícitament". De forma sintètica expressava la capacitat d'una màquina a aprendre a partir de l'experiència.

1.3.2 Els models i l'aprenentatge

Actualment, el Machine Learning constitueix una de les branques principals de les ciències d'intel·ligència artificial, centrant-se en la creació d'algoritmes i models per resoldre tasques o detectar patrons. Segons la tasca a resoldre el model resultant diferirà, sent aquest l'element nuclear del Machine Leanring. Segons el tipus de model la consolidació de l'aprenentatge és diferent[49].

Figura 6. Models d'aprenentatge utilitzat en Machine Learning



Font: Gràfic d'elaboració pròpia.

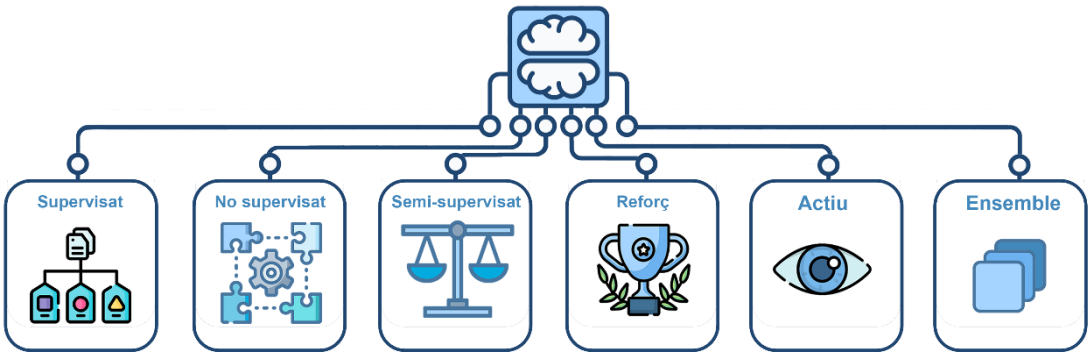
La **Figura 6** mostra els principals models que s'usen en Machine Learning[50]. Si la tasca a resoldre consisteix a combinar característiques o patrons a través de diverses dimensions, el model d'aprenentatge utilitza és el model geomètric. Si aquestes característiques es poden separar a través d'un hiperplà el model geomètric és linealment separable. En cas que no sigui possible el model serà linealment no separable. Si la tasca a realitzar és predir i estimar la incertesa d'un fenomen el model usat és el probabilístic. Aquest es basa a detectar la distribució dels casos per obtenir-ne la millor estimació possible on l'estadística

bayesiana i pren protagonisme. Quan la tasca consisteix a relacionar característiques o resultats es fa ús dels models lògics. Si la tasca a dur a terme consisteix a agrupar o dividir un conjunt de dades s'usen els models d'agrupació. Aquests tenen en compte l'estructura de dades per agrupar els objectes. Finalment, si la tasca a fer consisteix a identificar i portar a terme optimitzacions de resultats, el model emprat serà el de gradients.

1.3.3 Tipus d'algoritmes d'aprenentatge

Per a qualsevol eina, el tipus d'estratègia d'aprenentatge és clau per la potencialitat de l'eina. De forma genèrica es troben dos enfocaments a l'hora de construir els algoritmes sobre els quals l'entorn obtindrà informació i generarà un aprenentatge: mètodes supervisats i no supervisats. Els mètodes supervisats són aquells en el que l'eina és entrenada amb una estructura de dades prèviament treballada. En el conjunt de dades els valors d'entrada estan associats amb uns valors de sortida. Així el model pot aprendre en funció de les noves dades introduïdes. En l'estructura de dades els mètodes no supervisats no s'associa cap valor entrada i sortida. D'aquesta forma, l'eina troba les seves pròpies agrupacions o patrons.

Figura 7. Principals mètodes d'aprenentatge de les eines enfocades al Machine Learning



Font: Gràfic d'elaboració pròpia.

La **Per a** qualsevol eina, el tipus d'estratègia d'aprenentatge és clau per la potencialitat de l'eina. De forma genèrica es troben dos enfocaments a l'hora de construir els algoritmes sobre els quals l'entorn obtindrà informació i generarà un aprenentatge: mètodes supervisats i no supervisats. Els mètodes supervisats són aquells en el que l'eina és entrenada amb una estructura de dades

prèviament treballada. En el conjunt de dades els valors d'entrada estan associats amb uns valors de sortida. Així el model pot aprendre en funció de les noves dades introduïdes. En l'estructura de dades els mètodes no supervisats no s'associa cap valor entrada i sortida. D'aquesta forma, l'eina troba les seves pròpies agrupacions o patrons.

Figura 7 mostren múltiples mecanismes d'aprenentatge[50,51]. Més enllà dels mètodes supervisats i no supervisats existeixen tipologies d'algoritmes d'aprenentatges que són hibridacions o variacions dels mètodes anteriors, com per exemple el mètode semi supervisat. Aquest mètode consisteix a utilitzar conjunt de dades estructurades i processades amb conjunts de dades sense depurar. El mètode de recompensa construeix un entorn que aprèn en funció de l'experiència generada, on cada resultat està associat a un premi o una penalització. El mètode actiu consisteix a seleccionar els casos que vol fer servir el model per aprendre, en lloc de fer ús de tot un conjunt de dades estàtic i predefinit. El model Ensemble combina múltiples models individuals per generar un aprenentatge més sòlid i consistent.

1.4 Revisió sistemàtica

1.4.1 Cerca

S'ha creat una revisió sistemàtica que consisteix en una cerca en les bases de dades literàries: EconLit, Springer Link, ScienceDirect, Google Scholar i IEEE Explore fins al març de 2023. La cerca s'ha realitzat utilitzant les paraules clau "*Econometrics*" i "*Big Data*" connectades a través del booleà "*AND*" i aplicant el filtre de llengua "*English*".

1.4.2 Depuració

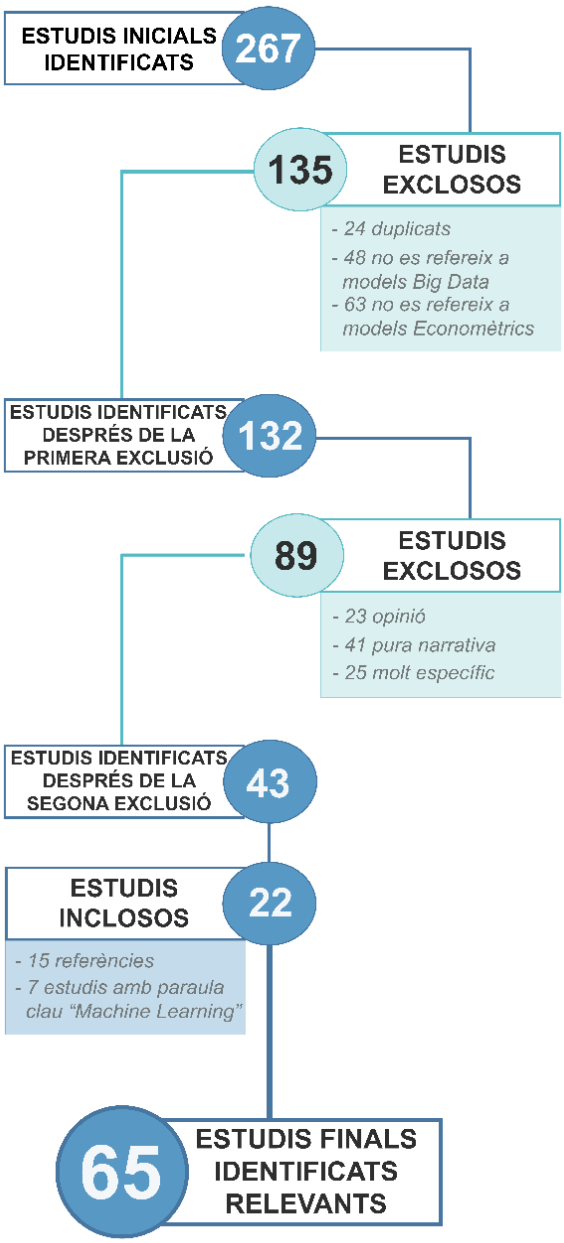
El recompte total de documents obtinguts amb les cerques a les diferents bases de dades literàries ha estat de 267 documents. Aquests han passat per uns filtres que es detallen a continuació, veure **Figura 8**.

Un cop recopilades totes les referències, s'han revisat els títols i els resums. Aquesta primera revisió permet actuar com a filtre, on s'eliminen: les referències duplicades, les que no són articles o estudis, les que són purament narratives o d'opinió i les que no estan vinculades ni fan referència a models economètrics i a entorns Big Data.

Un cop depurats els articles, s'han llegit i discutit. En aquesta segona fase, s'han comprovat totes les referències i cites. Les que han sigut d'interès s'han revisat i les que aportaven informació rellevant s'han incorporat a la revisió. En aquesta segona fase, s'ha observat com el terme "Machine Learning" es repeteix en la majoria d'estudis. En aquest punt, s'ha decidit afegir el terme a la cerca. La síntesi dels articles es pot trobar en les taules: **Taula 1**, **Taula 2**, **Taula 3**, **Taula 4**, **Taula 5** i **Taula 6**.

Per tant, s'ha afegit el terme "Machine Learning" a una nova cerca connectada al terme "Econometrics" pel booleà "AND", juntament amb tota la literatura que havia superat els filtres anteriors. Finalment, tots els estudis que han superat tots els filtres s'han afegit a la revisió sistemàtica.

Figura 8. Diagrama de flux dels articles seleccionats i procés de depuració utilitzant les paraules clau "*Big Data*", "*Econometrics*" i "*Machine Learning*"



Font: Gràfic d'elaboració pròpia.

1.4.3 Resultats

La mineria de dades és un concepte estretament relacionat amb l'estadística i la informàtica i, en particular, amb el *Machine Learning*. Les dues branques estan enfocades a manipular dades, encara que amb algunes diferències. La mineria de dades se centra a generar coneixement a través d'algoritmes, models i prediccions de valor. Totes aquestes accions són supervisades de forma intensiva per almenys un humà. En canvi, el *Machine Learning* se centra en

l'anàlisi de resultats i implica menys interacció humana. Consisteix en un algoritme inicial que interactua amb noves dades de forma iterativa.

1.4.3.1 Mètodes supervisats: Classificadors

Un dels algoritmes més utilitzats en l'agrupació de casos és el K-Nearest Neighbors (K-NN)[52]. Aquest model d'agrupació ordena segons característiques específiques, els objectes veïns (*k*) més propers entre ells. En l'àmbit computacional és senzill i eficient, també en grans bases de dades. Basat en una regressió no paramètrica, aquest mètode divideix la mostra en *k* característiques, independentment del tipus de característica. Aquestes submostres comprenen *m* grups de la mostra, cadascun amb les seves pròpies característiques individuals i globals. Aquest mètode va ser introduït el 1951 per Fix i Hodges[53].

Diversos autors han estudiat, criticat i optimitzat el mètode, incloent-hi Cover i Hart[54], que van establir les propietats del K-NN; Hellman[55] que ho va ajustar respecte a la taxa d'error de Bayes; així com Dudani[56] i Bailey i Jain[57], que van reduir els problemes computacionals associats. L'atractiu d'aquest mètode es troba en la precisió de la informació proporcionada per cada grup de la mostra i la generació de prediccions segmentades per a una o totes les submostres. Tots aquests aspectes, combinats amb la facilitat d'ús i el creixement de la intel·ligència artificial basats en models predictius, expliquen l'augment de la seva popularitat durant l'última dècada.

Una altra eina clàssica de classificació de mostres és el mètode Classification And Regression Trees (CART), conegut popularment com a mètode de l'arbre de decisió. És una tècnica molt popular entre els econometristes en la classificació de conjunts de dades. El 1984, Breiman et al.[58] va recuperar aquesta tècnica que s'havia introduït a la dècada del seixanta[59], però que no havia rebut molta atenció per part de la comunitat[60]. Aquesta tècnica consisteix a generar un arbre de creixement. L'arbre classifica diferents característiques de múltiples casos a partir d'un conjunt de dades, que actuen com les seves arrels. Aquest mètode permet subdividir el conjunt de dades en nodes i fer una predicció eficient tant dins com fora d'una mostra controlada, fins i tot amb valors perduts.

L'aleatorietat en les mostres afecta els mecanismes de classificació. En el seu assaig, Varian va presentar múltiples opcions per fer front a l'aleatoriat[27]. També va realitzar una crítica dirigida a la comunitat econometrista, ja que considerava que sempre s'utilitzava la mateixa tècnica: el Bootstrap, tot i no ser sempre la mes adequada. El Bootstrap és un procediment inspirat en la tècnica de remostreig Jackknife[61,62] que va presentar Efron el 1979[63]. Consisteix en un remostreig de les dades originals per obtenir inferències dels paràmetres i aplicar-los a la mostra original. L'aleatorietat de la mostra original es resol mitjançant el remostreig de parts de les mostres[64]. El 1981, Rubin[65] va presentar un Natural Bayesian Analogue del Bootstrap. Aquest actua com una extensió bayesiana d'aquest mètode. Uns anys més tard, Efron va presentar una variació que corregeix el biaix i accelera del procés del Bootstrap[66] així com el càlcul aproximat bayesià[67]. Varian també va proposar dos mètodes més, tot i que no van ser gaire utilitzats[27]. El primer s'anomena Bagging[68] i consisteix en la modelització de valors mitjans. El mecanisme crea múltiples mostres d'entrenament de diferents dimensions i utilitza els resultats com a una mitja combinada. Les mostres que contenen valors repetits s'ajusten durant el procés. Per tant, aquest mètode redueix la variància i és molt eficient en models individuals inestables, però produeix diversos problemes quan es combina amb altres mètodes.

El segon mètode és el Boosting i és una altra alternativa als mètodes Bootstrap i Bagging. Va ser presentat el 1990 per Schapire[69]. Aquesta tècnica es basa en un marc d'anàlisi matemàtica anomenat probabilitat aproximadament correcte. Al mateix temps, actuava com a resposta a una proposta de Kearns[70] i Valiant[71]. Consisteix a generar un indicador potent ordenant els resultats dels classificadors de més feble a menys i fusionant-los per crear un classificador més robust[69]. Freund va proposar una alternativa que és més eficient, però amb certs problemes pràctics amb els classificadors més febles[72]. L'any 1997, Schapire i Freund van presentar una altra variant del mètode Boosting que combina els valors més febles en un sol resultat[73]. Aquesta combinació correspon a la suma de pesos d'aquests valors. Després els converteixen en un nou classificador. La principal variació d'aquest mètode consisteix en com es pesen les dades.

Utilitzant ambdues tècniques de classificació, Bagging i Boosting, es pot crear un conjunt de submostres que generin un grup d'arbres de decisió amb la mateixa longitud[74]. Aquest bosc està format per un conjunt d'arbres de decisió ponderats que no estan correlacionats. L'aleatorietat corregeix el sobreajust de la formació de dades inherent als arbres de decisió. Aquest algoritme va ser proposat per Ho[75,76] els anys 1995 i 1998 fent ús d'un model aleatori subespacial que fa servir una selecció de característiques d'una submostra. Aquest mètode es coneix com a Random Decision Forest. El 2001, Breiman[77] va desenvolupar un sistema de boscos aleatoris que va simplifica el mètode, restringint els arbres binaris, per aconseguir una major simplicitat. Aquesta opció redueix la tendència de la superposició de les dades. L'algoritme genera bons resultats i és eficient fins i tot amb bases de dades grans. Si s'observen arbres de forma aïllada poden presentar problemes analítics.

Taula 1. Síntesi dels principals mecanismes dels mètodes supervisats: Classificadors

Any	Algoritme	Autor	Descripció
1951	K-nearest neighbour classification (K-NN)	Fix & Hodges	Utilitzat com a mètode de classificació o regressió. Consisteix a agrupar la mostra en *m* mostres mitjançant *k* característiques per obtenir *m* valors per a cadascun dels *m* grups.
1979	Bootstrap	Efron	Utilitzat com a model inferencial, consisteix a remostrejar les dades originals i obtenir algunes inferències pels paràmetres a aplicar a la mostra original.

1984	Classification and regression trees (CART)	Breiman *et al.*	Utilitzat com a model predictiu, basat en l'observació d'un conjunt de dades i la subdivisió dels ítems amb *n* nivells o característiques que es poden utilitzar com a base per a la presa de decisions.
1990	Boosting	Schapire	Utilitzat com a model de mitjanes, consisteix a generar un indicador potent fusionant els resultats dels indicadors de més febles a menys per crear un classificador més robust.
1994	Bagging	Breiman	Utilitzat com a model de mitjanes, consisteix a crear diferents mostres d'entrenament i utilitzar els resultats com a resultat mitjà combinat.
1995	Random forest (Multiple CART)	Ho	Utilitzat com a mètode de classificació, consisteix a produir un nombre molt gran de CART per a l'entrenament, obtenint-ne els resultats, i crear un classificador o un predictor de tots ells.

Font: Taula d'elaboració pròpia.

1.4.3.2 Mètodes supervisats: Classificadors amb penalització als predictors

Un model estadístic clàssic molt estès en la literatura acadèmica de la comunitat economista és la regressió lineal. Consisteix en un model bàsic que mesura la relació entre dues variables. Aquest mètode i les seves variants serveixen per interpretar un o diferents esdeveniments, estimar-los i predir-los. Una anàlisi en profunditat de les regressions lineals presenta certes limitacions amb les classificacions. El principal problema sorgeix quan s'intenta crear un model que permeti modelitzar a partir d'un conjunt de característiques d'un grup de casos o individus[78–81]. A més, és difícil seleccionar un subconjunt d'individus a partir d'una base de dades quan s'apliquen algoritmes diferents.

Els algoritmes basats en la selecció de variables són molt comuns, ja que aquesta és una tasca clau per a la modelització estadística en una gran base de dades. Són útils per als investigadors per simplificar els models reduint les dades d'entrenament i la variabilitat. Aquests algoritmes es poden utilitzar en diferents tipus de models: autoregressius[82], autoregressius vectorials[83] i predictius[84]. Les característiques s'escullen afegint i eliminant algunes variables predictibles, en

un grup de patrons, utilitzant un conjunt de diferents bondats d'ajust que permet ajustar els millors models[85].

Quan el conjunt de dades és gran i multivariant o quan el nombre de variables és més gran que la dimensió de la mostra, el model presenta diferents limitacions. L'enfocament per obtenir submostres i subconjunts de dades per etapes és naturalment inestable. A més, dificulta l'extracció de valors a causa de la seva complexitat subòptima[86]. Els mètodes basats en penalitzacions es presenten com una alternativa a aquest enfocament clàssic. La penalització consisteix a ponderar els coeficients a partir de les variables menys contributives. Aquestes tècniques estan augmentant en popularitat gràcies a la facilitat de detectar interaccions. Hi ha diversos mètodes basats en la penalització per evitar el sobreajustament durant el tractament de les dades.

El primer mètode que va aparèixer basat en la penalització dels coeficients va ser la Ridge Regression (RR). Aquest mètode és útil per reduir el problema de la multicol·linealitat. Va ser proposat el 1943 per Tikhonov[87], però van ser Hoerl i Kennard 1970[88,89], els que van ampliar el mètode aplicant els casos de dimensions finites i el van desenvolupar seguint un enfocament estadístic. L'algoritme funciona a partir d'un estimador de regressió, que permet reduir els errors de predicció, disminuint la mida dels coeficients grans per tal de reduir el sobreajustament. A més, penalitza els coeficients, però no funciona com a selector de variables, dificultant així la interpretació dels models.

Anys més tard, Frank i Friedman van proposar la Bridge Regression (BR)[90], que funciona com una generalització de la Ridge Regression. El mecanisme resumeix les estimacions dels coeficients beta penalitzant la suma residual de quadrats (RSS) en un punt. Compta amb moltes bondats com ara la dispersió i la imparcialitat. Tanmateix, presenta diversos problemes a l'hora de sistematitzar el procés per generar inferències. Aquest procediment es fa més complicat en aplicacions pràctiques a causa de la seva rigidesa. El 2011, Polson la va estendre a una versió de mètodes bayesians que simplifica i facilita l'aplicabilitat del mètode[91].

El mateix any que es va presentar la Bridge Regression, es va reformular un mètode en forma de mostreig de Gibbs desenvolupat per George i McCullochin[92] el 1993. Abans havia estat introduït per Geman i Geman[93], però en la versió proposada per George i McCulloch, l'algoritme funciona com un selector de variables. Es basa en el mètode Markov Chain Monte Carlo (MCMC)[38] que funciona com un algoritme aleatori, a més a més, de ser una alternativa clau als algoritmes deterministes en inferències estadístiques. El mostreig de Gibbs s'utilitza per fer inferències a partir d'un gran conjunt de dades amb moltes variables. Aquest, pot dividir qualsevol variable, grup de variables o subgrup d'una mostra. Consisteix a crear unes cadenes de valors de Markov correlacionant-les amb les mostres i eliminant les cadenes que no estiguin correlacionades. Com més gran sigui la cadena, millors seran els resultats obtinguts. Aquest mètode és útil quan la distribució és desconeguda o quan la distribució d'una variable és coneguda, però és difícil mostrejar-la directament.

El 1995, Breiman va proposar l'estimador de Non-Negative Garrote (NNG), que és una versió a escala de l'estimació de mínims quadrats[94]. L'algoritme diferencia entre els coeficients que obtenen un mínim quadrat gran o petit. En el cas dels coeficients amb un quadrat mínim gran, els factors de constrenyiment són propers a 1, mentre que els que tenen un quadrat mínim petit són identificats com a redundants per a una predicció i el factor de constrenyiment és 0 o proper a 0. En particular, aquest mètode no treballa bé amb un conjunt de dades petit o amb un nombre més gran de prediccions que la mida de la mostra, ja que els mínims quadrats funcionen de manera ineficient.

Un dels mètodes de penalització més populars s'anomena Least Absolute Shrinkage and Selection Operator (LASSO). Va ser desenvolupat el 1996 per Tibishirani[95] per facilitar la selecció i regularització de variables i obtenir una major precisió en la predicció del model. Anteriorment, Santosa i Symes[96], havien presentat un mètode similar. La principal diferència entre aquests dos mètodes és que el proposat per Santosa i Symes es basava en l'ús del ℓ^1 per ajustar els coeficients i la seva aplicació es basava en la geofísica. En canvi, el mètode de Tibishirani es basa en el model Non-Negative de Breiman. En aquest mètode, els coeficients β tendeixen a 0, donant lloc a un millor ajust i anul·lant o

minimitzant determinades variables febles que afecten els valors de predicció. A més permet la seva extensió a altres models. El mètode LASSO sovint es complementa amb altres paràmetres de regularització[97].

Un altre mètode popular que es destaca és la Smoothly Clipped Absolute Deviation (SCAD) proposada per Fan i Li l'any 2001. Funciona com un mètode de penalització no còncau[98]. L'algoritme redueix els coeficients petits a zero i tendeix a reduir-ne els altres menys febles a zero, mantenint el valor dels coeficients grans. La dificultat d'aquest mètode rau en triar els coeficients que s'han d'arrodonir a zero. Aquest procés genera un coeficient imparcial per als coeficients grans. A diferència d'altres mètodes, l'estimador SCAD presenta totes les propietats desitjables; és a dir, imparcial, dispersió i continuïtat[99]. El 2005, Zou i Hastie van presentar una adaptació de LASSO que també satisfà les tres propietats, conegut com a Elastic Net[100]. Aquesta combina els mètodes LASSO i Ridge per trobar els coeficients de Ridge i després aplicar la penalització LASSO. Aquest procés genera un doble constrenyiment que pot provocar un augment del biaix. Per resoldre-ho l'algoritme reescala els coeficients.

Hi ha altres alternatives interessants: la selecció de Bayes, la selecció de models adaptatius, Least-Angle Regression (LARS) o la Minimax Concave Penalty (MCP). La selecció de Bayes es basa en el Alternative Model Selection (AMS) proposada per Shen i Ye l'any 2002[101] i funciona aplicant una penalització. Tanmateix, en aquest cas, la penalització es basa en un concepte de graus de llibertat generalitzats i depèn de la mida del model. Aquest procediment redueix el biaix de selecció i condueix a més precisió. Shen i Ye critiquen els models de selecció, ja que només funcionen bé en un tipus de situació i no permeten l'adaptació en diferents entorns.

LARS, desenvolupat per Efron et al.[102], és un algoritme que ajusta els models de regressió lineal a un gran nombre de variables. Aquest model de regressió permet determinar les variables de resposta i obtenir els seus coeficients. Les variables s'estimen en una direcció equiangular a la de les seves correlacions i residus. Els valors s'afegeixen un a un per aconseguir el millor model. El model funciona quan el nombre de variables és superior a la mida de la mostra i

presenta una fàcil adaptació a diferents algoritmes. A més, és eficient computacionalment, però és especialment sensible als efectes del soroll. Per la seva banda, les regles i factors de probabilitat bayesians quantifiquen el valor del model ajustant el tipus de model. Aquest mètode beneficia als models senzills que busquen controlar el sobreajustament d'estructures complexes. En els models complexos l'ajust acostuma a ser més petit. S'aplica el mètode empíric de Bayes per crear un criteri de selecció adaptatiu i millorar-ne els criteris de selecció fixos[103].

En models dispersos, les penalitzacions MCP són una altra alternativa a SCAD i LASSO. Les penalitzacions MCP s'obtenen de forma més eficient i els coeficients presenten un menor biaix. Aquest mètode va ser proposat per Zhang al 2010[104]. El mecanisme té un funcionament molt similar al SCAD. Inicialment, aplica el mateix constrenyiment que SCAD, però a diferència d'aquest el constrenyiment es comença a reduir molt abans del que ho fa SCAD.

Els models bayesians han augmentat en popularitat al món de l'econometria des de l'augment de les metodologies Big Data[83]. Un altre mètode utilitzat per seleccionar i penalitzar les variables és el model de regressió Spike-and-Slab. Aquest, consisteix a escollir les característiques a incloure en l'anàlisi i realitzar una regressió lineal simultàniament. La selecció de variables es realitza com una fusió del component Spike-and-Slab sobre els efectes dels coeficients. La primera part calcula la probabilitat spike dels coeficients per tal que la selecció sigui 0. La segona part calcula la distribució slab per als coeficients de regressió. Les diferents característiques funcionen simultàniament per generar una regressió lineal i arrodonir les característiques més febles a 0. Aquest mètode va ser ajustat el 2005 per Ishwaran i Rao[105] però inicialment presentat per Mitchell i Beauchamp[106] i tractat per Madigan i Raftery[107]. És un mètode amb una complexitat matemàtica elevada que dificulta la seva implementació.

Taula 2. Síntesi dels principals mecanismes dels mètodes supervisats: Classificadors amb penalització als predictors

Any	Algoritme	Autor	Descripció
1970	Ridge regression (RR)	Hoerl & Kennard	Utilitzat com a mètode de penalització, consisteix a reduir els errors de predicció mitjançant la disminució de la mida de grans coeficients amb l'objectiu de reduir el sobreajustament i obtenir millors resultats amb presencia de multicolinearitat.
1993	Bridge regression (BR)	Frank & Friedman	Utilitzat com a mètode de classificació, penalització i predicció, consisteix en una generalització d'una regressió de cresta que resumeix les estimacions del coeficient beta i penalitza la suma residual de quadrats.
1993	Gibbs sampling	George & McCulloch	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a vincular un MCMC a un subconjunt de dades i, a partir de la seva correlació fer prediccions amb els valors obtinguts.
1995	Non-negative Garrote (NNG)	Breiman	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a detectar els coeficients principals i controlar els coeficients redundants tendint-los a zero per tal d'obtenir un millor predictor.
1996	Least Absolute Shrinkage and Selection Operator (LASSO)	Robert Tibishirani	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a realitzar la selecció i regularització de la variable per tal d'aconseguir una major precisió de la predicció reduint l'impacte de les variables més febles.
2001	Smoothly Clipped Absolute Deviation (SCAD)	Fan & Li	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a reduir a zero els coeficients més febles, reduir a zero els altres menys febles i mantenir el valor dels coeficients forts. Aquest procés genera un valor imparcial amb dispersió i continuïtat.

2002	Adaptative model selection (AMS)	Shen & Ye	Utilitzat com a mètode de classificació, penalització i predicció penalitzada en funció de la mida del model, permet adaptar-se a múltiples situacions, així com obtenir un baix biaix i una major precisió.
2004	Least angle regression (LARS)	Efron *et al.*	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a determinar quines variables són variables de resposta.
2005	Elastic Net	Zou & Hastie	Utilitzat com a mètode de classificació, penalització i predicció, combina els mètodes LASSO i Ridge. Aquest procés genera una doble contracció i reescala els coeficients.
	Spike-and-slab	Ishwaran & Rao	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a escollir les característiques i aplicar simultàniament una regressió lineal basada en una teoria bayesiana.
2010	Penalitzacions concoves minimax (MCP)	Zhang	Utilitzat com a mètode de penalització i predicció, consisteix a realitzar la selecció i regularització de la variable per tal d'aconseguir una major precisió de la predicció reduint l'impacte de les variables més febles.
2014	Bayesian bridge regression	Polson *et al.*	Utilitzat com a una extensió de la clàssica regressió de Bridge, facilita l'aplicació del mètode a la pràctica utilitzant els principis bayesians.

Font: Taula d'elaboració pròpia.

1.4.3.3 Mètodes supervisats: Classificador quasi bayesians i altres penalitzadors dels predictors

Si a l'anàlisi de dades se li incorpora un patró temporal, aquest dificulta els estudis amb mètodes bayesians clàssics. Scott i Varian van presentar un mètode Bayesià per a aquests tipus de sèries temporals quan el conjunt de dades és massa gran. Proposen buscar múltiples submostres d'un conjunt de dades que facin aflorar les variables més rellevants. Aquest mètode aconsegueix reduir el temps de processament i també en millora l'eficiència computacional[40]. Si les dades contenen ruptures, això es resol mitjançant un enfocament Bayesià basat en models de panells que estimen els valors que falten[108].

La teoria quasi bayesiana va ser presentada per McCray[109] el 1984 com una alternativa quan la distribució de les dades és desconeguda. Es basa en un model multivariant amb una parametrització variable en el temps. Aquest mètode assumeix l'existència d'un gran nombre d'errors tenint en compte que la distribució de les dades és desconeguda. Tanmateix, permet estimar els valors sense conèixer la distribució utilitzant una probabilitat màxima. El procés exigeix que els coeficients no es ponderin de forma prèvia, fins i tot quan desconeixem la seva distribució. Així redueix alguns problemes computacionals en l'obtenció de les probabilitats[110]. Els resultats de la selecció de variables per aquest mètode són equivalents a la selecció de variables obtinguda aplicant una penalització a la suma de quadrats i a la suma de quadrats residuals. Hi ha diferents estimacions basades en mètodes quasi bayesians com l'estimació de quasi-maximum likelihood estimate (QMLE) o el quasi-Bayesian local likelihood (QBLL).

El QMLE va ser presentat per Weiis[111] com una estimació d'un θ obtingut de maximitzar una funció vinculada al logaritme de la funció de probabilitat. Aquesta funció és més simple que la funció de quasi probabilitat, però és més consistent i asimptòticament normal. La pèrdua d'eficiència es compensa amb una minimització de la pèrdua d'informació real. S'han desenvolupat algunes extensions del model basades en distribucions gaussianes[112] i de Poisson[113]. Un enfocament de la QBLL va ser exposada per Petrova et al.[114]. A diferència del mètode de QMLE, augmenta els estimadors més freqüents per permetre un tractament bayesià dels paràmetres de deriva. Consisteix a enfocar els paràmetres de manera no paramètrica per a l'estimació, aplicant una ponderació basada en una densitat prèvia. Proporciona un valor que és asimptòticament vàlid i permet generar inferències i prediccions.

Una altra alternativa és el quantile regression (QR)[115]. Aquesta és una eina que produeix estimacions robustes entre intervals analitzant múltiples variables[116]. Aquest mètode estima la mitjana condicional o altres quantils de la variable resposta. Aquest mètode va perdre popularitat en detriment del mètode de mínims quadrats. Analitzar un gran conjunt de dades sense ordinadors pel mètode QR es converteix en una tasca difícil de realitzar.

Taula 3. Síntesi dels principals mecanismes dels mètodes supervisats: Classificadors quasi bayesians i altres penalitzadors dels predictors

Any	Algoritme	Autor	Descripció
1978	Quantile regressions (QR)	Koenker & Bassett	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a estimar els mitjans condicionals o altres quantils de la variable resposta.
1984	Quasi-Bayesian	McCray	Utilitzat com a mètode de classificació, penalització i predicció, permet que una regressió bayesiana funcioni sense conèixer la distribució de dades i obtingui resultats equivalents o similars.
1986	Quasi-maximum likelihood estimator (QMLE)	Weiss	Utilitzat com a mètode de classificació, penalització i predicció, permet que una regressió bayesiana funcioni sense conèixer la distribució de dades i aconsegueixi resultats equivalents o similars.
2019	Quasi-Bayesian local likelihood (QBLL)	Petrova et al.	Utilitzat com a mètode de classificació, penalització i predicció, consisteix a maximitzar una funció vinculada al logaritme de la funció de probabilitat per tal d'aconseguir un valor.

Font: Taula d'elaboració pròpia.

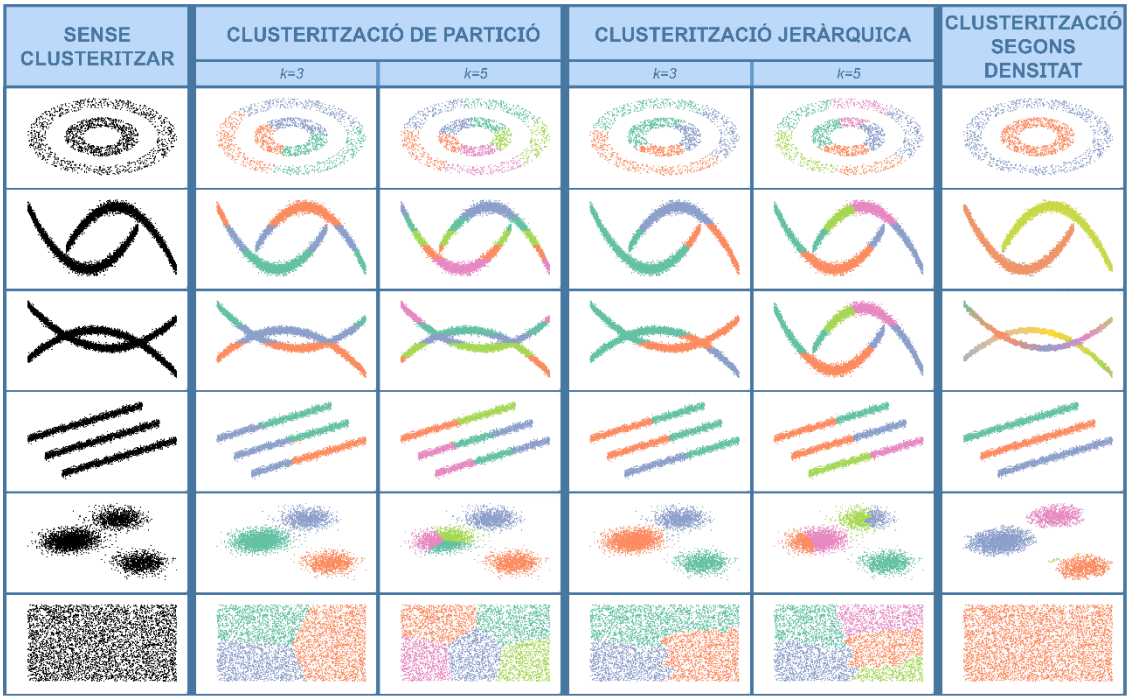
1.4.3.4 Mètodes no supervisats: Clustering

Tots els models, mètodes i algorismes presentats fins ara requereixen una interacció supervisada que modifiqui determinades variables per tal de generar una major precisió en la selecció de variables i predictors. Els mètodes que es presenten a continuació, a diferència dels descrits fins ara, es basen en la interacció no supervisada i permeten que l'algoritme s'adapti a l'entorn que genera les dades, així com a canvis en els patrons o ajustament dels valors.

El clúster és una anàlisi basada en un grup d'objectes d'un conjunt de dades. Es busquen patrons en les dades i s'agrupen segons les similituds i diferencien dels objectes. En l'anàlisi de clúster, s'utilitzen diferents algoritmes per agrupar els objectes, i la diferència entre ells rau en la forma en com s'agrupen, veure **Figura 9**. El potencial i popularització de l'agrupació de dades es troba en l'automatització dels processos.

Un dels mètodes d'agrupament més antics va ser establert el 1894 per Karl Pearson[117] i el va anomenar model mixta. En aquest model, es representa un conjunt *k* de subgrups, on cadascun dels *k* subgrups que té la seva pròpia distribució dins d'un conjunt de dades *K* sense requerir la identificació individual. Tot i que s'atribueixen identitats o pesos a subgrups basats en casos en entorns no supervisats, un cop subagrupats, es porta a terme un procés de normalització de les dades a 1. Els models mixts s'utilitzen per fer prediccions i inferències sobre les propietats de cada subgrup a partir dels casos agrupats. Algunes extensions basades en la teoria bayesiana i la distribució gaussiana estan disponibles per a aquests mètodes.

Figura 9. Visualització de diferents agrupacions segons el tipus de clusterització: mètode de partició, mètode jeràrquic i basats en densitat, utilitzant distribucions complexes



Font: Gràfic d'elaboració pròpia.

Un dels algoritmes més utilitzats va ser desenvolupat l'any 1957 per Stuart Lloyd[118] i s'anomena K-means, tot i que no va ser publicat fins al 1982[119]. Consisteix en un procés que consta de dues etapes. En la primera, s'escull un nombre *k* de centroides i s'assigna aleatòriament el seu inici. Els casos individuals s'assignen a cada *k* centroide en funció d'una minimització de les distàncies respecte a les mitjanes de cada centroide. En segon lloc, els centroides es tornen a calcular amb els punts assignats i s'actualitzen

iterativament fins que no hi hagi variació en cap cas. L'algoritme pot variar en funció de la mesura de proximitat. Aquest concepte va ser introduït per Hugo Steinhaus el 1956[83] i presentat el 1967 per James MacQueen[120]. L'any 1965 es va presentar una variant de l'algoritme amb un mètode gairebé idèntic al de Lloyd[121]. Hartigan i Wong[122] van presentar més tard una versió més eficient del mètode inicial. També ho va fer Dunn el 1973[123,124], que va proposar un mètode computacional més complex que considera que un cas podia pertànyer a més d'un clúster. Aquest mètode va rebre el nom de FUZZY per la forma en què generava les agrupacions.

Les tècniques de clusterització per partició tenen limitacions amb la sensibilitat als valors atípics. Alternativament, es pot prendre com a valor central d'un clúster el valor real del conjunt de dades i no un valor mitjà. Aquests punts centrals s'anomenen medoids o K-medoids. El primer mecanisme basat en K-medoids, va ser presentat l'any 1986 per Kaufman i Rousseeuw, anomenat Partitioning Around Medoids (PAM)[125,126]. A diferència dels algoritmes K-Means, aporta major robustesa amb el soroll de les dades i els valors outliers. No obstant això, és computacionalment més complexa. El 1990, Kaufman i Rousseeuw proposen una extensió de PAM que millora les limitacions computacionals, permeten la lectura de més observacions anomenades Clustering Large Applications (CLARA)[127]. Amb el temps, Ng i Han, van proposar una variació de CLARA més eficient, sobretot amb dades espacials, anomenada Clustering Large Applications based on Randomized Search (CLARANS)[128].

Tot i que els medoids treballaven el soroll de les dades, de forma alternativa, es va desenvolupar un algoritme que es va popularitzar dins la comunitat científica. El 1996, Martin Ester, Hans-Peter Kriegel, Jörg Sander i Xiaowei Xu van introduir el Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[129]. El seu funcionament és similar als algoritmes de partició i es basa en la distribució de les densitats de la base de dades. A diferència del mètode anterior, permet que l'algoritme detecti el nombre de clústers de forma autònoma. Aquesta agrupació es realitza segons la concentració de densitat dins d'un radi, la mida del qual és l'últim que ha de decidir l'algoritme. Es distingeixen tres tipus de casos: els que es troben dins del radi (nucli), els que es troben a la vora del radi

(la vora) i els que estan lluny d'aquest (soroll). Aquest mètode permet una major resistència al soroll i funciona amb bases de dades de diferents mides i formes. Sovint el mètode K-means té dificultat o no es pot adaptar a aquestes situacions.

L'any 1998[130] es va proposar un mètode alternatiu sota el nom de Density-Based Clustering (DENCLUE), que es va desenvolupar per classificar un conjunt de dades amb un nivell de soroll augmentat i es va convertir en la solució de referència amb conjunts de dades amb soroll. Es basa en la densitat dels grups i la influència dels casos entre ells per la generació d'aquests grups. La suma de les funcions d'influència representa la densitat de cada grup. És més ràpid que el DBSCAN i permet categoritzar amb diferents mides de densitat. Un any més tard, Mihael Ankerst et. al van presentar un altre algoritme també basat en l'agrupació de densitats[131]. Aquest mètode ordena les dades en clústers segons la densitat de cadascuna de les estructures de dades, de manera que els casos espacialment propers es consideren veïns i s'agrupen. Aquest ordenament és equivalent als clústers basats en densitats. Millora el mètode DBSCAN, ja que és més eficient detectant clústers i subclústers de diferents densitats.

Quan el conjunt de dades inclou valors atípics, es fa difícil per a molts mètodes abordar aquests valors, motiu pel qual es va proposar el Factor Local Outlier (LOF)[132] l'any 2000. Aquest mecanisme es basa en la detecció de valors atípics segons la seva densitat i calcula la distància de cada cas respecte als seus veïns. El 2003, es va crear un mètode anomenat Shared Nearest Neighbour (SNN) amb l'objectiu de detectar diferents clústers amb diferents formes, mides i densitats en un gran conjunt de dades multidimensional[133]. A diferència del mètode DBSCAN, aquest algorisme detecta clústers amb diferents densitats.

L'agrupació jeràrquica també es pot utilitzar per analitzar les dades sense tenir en compte la densitat de les dades i per generar alguns clústers. En els mètodes jeràrquics, els casos no es divideixen en clústers només una vegada, sinó que es fan particions successives de manera iterativa segons la jerarquia existent. El seu ús ha augmentat sobretot gràcies a l'anàlisi de dades de les xarxes socials. Consisteix a construir una jerarquia de grups amb diferents nivells aplicant dues tècniques diferents segons com s'hagin d'agrupar els objectes: aglomerativa o

divisiva. Aquest mètode va ser proposat inicialment de forma aglomerativa el 1951[134]. La tècnica aglomerativa crea jerarquies ascendents, en les quals cada objecte té un grup, i els grups es generen durant el procés de jerarquia creixent. La tècnica Agglomerative Nesting (AGNES)[135] és dels mètodes de jerarquies ascendents més representatius.

Amb el mètode divisiu, l'objecte s'inclou inicialment en un sol grup i després els grups es divideixen en ordre descendent. Una de les principals tècniques del mètode divisiu s'anomena Divisie Analysis (DIANA)[136]. Tot i que tots dos tenen limitacions computacionals quan es treballen amb grans conjunts de dades, permeten fer inferències. Altres mètode de clusterització utilitzant jerarquitzacions són BIRCH[137] i BICO[138].

Taula 4. Síntesi dels principals mecanismes dels mètodes no supervisats: Clustering

Any	Algoritme	Autor	Descripció
1894	Model mixta	Person	Utilitzat com a agrupador i ponderador, consisteix a dividir el conjunt de dades en k subgrups per obtenir una inferència sobre les propietats de cada subgrup a partir dels casos agrupats.
1967	K-means	MacQueen	Utilitzat com a classificador i predictor, consisteix a dividir un conjunt de dades amb un nombre de grup de k assignant els casos més propers a partir de la minimització de les distàncies entre cada cas i els grups.
1973	Fuzzy	Dunn	Utilitzat com a classificador i predictor, consisteix a dividir un conjunt de dades amb un nombre de grup de k assignant els casos en diferents conjunts de grups en una partició difusa, variant entre 0 i 1, fins a trobar una la agrupació òptima.
1986	Partitioning Around Medoids (PAM)	Kaufman & Rousseeuw	Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades en k medoids per obtenir una inferència sobre les propietats de cada subgrup a partir dels casos agrupats.

1990	Clustering Large Applications (CLARA)	Kaufman & Rousseeuw	Utilitzat com a classificador i predictor, consisteix a dividir de forma aleatòria el conjunt de dades en diferents submostres. Després clusteritzar cada subconjunt de dades per extreure els k medoids i trobar la clusterització amb major similitud entre cada medoid i les seves observacions associades per obtenir la clusterització final.
	Divisie Analysis (DIANA)	Kaufmann & Rousseeuw	Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades de forma jerarquia de forma aglomerativa (damunt a envall) per generar una classificació dels casos.
1990	Agglomerative Nesting (AGNES)	Kaufmann & Rousseeuw	Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades de forma jerarquia de forma divisiva (damunt a avall) per generar una classificació dels casos.
1994	Clustering Large Applications based on Randomized Search (CLARANS)	Ng & Han	Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades en k medoids de forma iterativa fins a trobar la distribució més òptima de k.
1996	Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	Ester et al.	Utilitzat com a agrupador i predictor, consisteix a dividir un conjunt de dades amb un nombre de grup de k en funció de la densitat dels casos. Això permet adaptar-lo a diferents tipus de bases de dades.
1997	Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)	Zhang et al.	Utilitzat com a agrupador, enfocat a grans volums de dades, desglossa les dades en petits resums que s'agrupen en lloc dels punts de dades originals. Els resultats capturen la informació de la distribució.
1998	Density-based Clustering (DENCLUE)	Hinneburg & Keim	Utilitzat com a agrupador i predictor en un conjunt de dades amb un gran nivell de soroll, consisteix en identificar un clúster a partir de la influència dels casos entre ells en la generació de grups.

1999	Ordering Points To Identify the Clustering Structure (OPTICS)	Ankerst et al.	Utilitzat com a agrupador i predictor, consisteix a emmagatzemar les dades en clústers segons la densitat de cadascun dels clústers corresponents a un ampli conjunt de paràmetres d'un algorisme de clúster basat en densitat.
2000	Local outlier factor (LOF)	Ertöz et al.	Utilitzat com a classificador, consisteix en una identificació a partir de la densitat dels casos allunyats del seu veí.
2003	Shared nearest neighbour (SNN)	Ertöz, Steinbach & Kumar	Utilitzat com a agrupador i predictor, consisteix a identificar diferents clústers amb diferents formes, mides i densitats en un gran conjunt de dades multidimensional.
2013	BIRCH meets coresets for k-means clustering (BICO)	Fichtenberger et al.	Utilitzat com a agrupador ràpid de grans volums de dades, calcula un nucli S, per cada centre C, el cost d'entrada P es pot aproximar calculant el cost de generar S.

Font: Taula d'elaboració pròpia.

1.4.3.5 Mètodes no supervisats: Xarxes neuronals

Un altre mètode per llegir i combinar dades consisteix a crear una xarxa neuronal de dades. Aquestes es van proposar a mitjan segle XX. Tot i que inicialment va sorgir en aplicacions més enfocades a Robòtica i Neurobiologia, actualment s'apliquen a diverses disciplines. També es pot aplicar en regressions utilitzades per fer prediccions i classificar casos.

Les xarxes neuronals van ser introduïdes el 1943 per McCulloch i Pitts[139]. Es basen en la connexió existent entre els nodes que representen els casos i l'anàlisi de les dades. Aquesta xarxa està formada per casos connectats entre si al llarg del conjunt de dades en què, com en el cas d'una xarxa neuronal biològica, les connexions entre casos transmeten iterativament informació a tota la xarxa a través dels nodes.

Una de les primeres limitacions d'aquestes xarxes senzilles és que no tenien capacitat d'aprenentatge. El 1949, Hebb va desenvolupar un principi de ponderació actualitzat per als nodes[140] que va permetre el desenvolupament de l'aprenentatge basat en dades. El pes calculat durant el procés de ponderació

s'emmagatzema entre les connexions de cada cas, de manera que el canvi de pes depèn de la interacció entre aquests nodes. Si els dos casos estan actius al mateix temps, la relació serà més forta i el pes serà més gran, en canvi, si s'activen en diferents moments, el pes serà menor. Així, fins i tot les relacions febles tenen un patró, de més a menys pes. Algunes de les limitacions d'aquest mètode inclouen la temporalitat de les dades, els retards, el control motor o l'estimulació incondicional, que requereixen solucions externes. Malgrat aquestes limitacions, aquest mètode va proporcionar la primera base per a l'aprenentatge de xarxes neuronals.

El 1986, Smolensky va proposar el mètode de la Restricted Boltzmann Machine (RBM)[141]. La RBM és un model generatiu basat en la distribució de probabilitat apresa per casos. A diferència de la majoria de les eines de xarxa, aquest mètode no té sortida adaptativa. La generació d'aprenentatge es realitza mitjançant la distribució dels nodes en dues capes: visible i invisible, per la qual cosa a cada node de la capa visible se li assigna un node a la capa invisible. Atès que el mètode determinarà la distribució en funció prèvia de la determinació dels nodes invisibles, el model predictiu varia d'una manera o altra. El seu ús es basa en la reducció multidimensional, la classificació, la regressió i el modelatge predictiu.

Un any després es va introduir un mètode que permet classificar les dades sense tenir en compte la presència de soroll a la xarxa[142]. Consisteix a recrear un conjunt de dades per forçar l'algoritme a generar dades de sortida a partir de les dades originals mitjançant un procés comprimit. Si les dades no tenen característiques o no estan correlacionades, la sortida serà similar o igual a les dades originals. En canvi, si les dades contenen característiques o patrons que les correlacionen, la sortida es modificarà i generarà un nou valor a partir del patró basat en l'aprenentatge. Un dels principals atractius d'aquest mètode és que ignora el soroll de les dades.

A principis de segle, Hilton va utilitzar conjunts de RBM per desenvolupar un tipus diferent d'aprenentatge a la xarxa neuronal[143] anomenada Deep Belief Network (DBN). Aquesta xarxa actua de manera diferent a altres mètodes d'aprenentatge,

ja que es basa en l'aprenentatge de tots els casos alhora. Les xarxes neuronals usen l'aprenentatge de processos basats en les vores, els exteriors i característiques senzilles que generen una primera segregació. A continuació, procedeixen a analitzar les dades profundes per generar un valor final. En canvi, DBN fa servir dades profundes i superficials al mateix nivell per calcular els pesos i les probabilitats de si un cas està implicat en un esdeveniment específic. Això permet desenvolupar models predictius.

En l'última dècada, Goodfellow et al. va proposar un altre mètode per interactuar amb xarxes neuronals, les Generative Adverse Networks (GAN). Aquests són un enfocament al modelatge generatiu mitjançant xarxes neuronals convolucionals[144] basat en dos submodels. El primer model entrena les dades per generar nous exemples, i el segon model és un model discriminador que intenta classificar si el cas és real (a partir del conjunt de dades) o fals (generat en el primer submodel). Els dos models estan entrenats per comprovar constantment si els casos generats simulen els casos reals. La propagació posterior s'aplica als dos models per garantir que el generador produeixi millors exemples. Un cop el model discriminant es pot enganyar diverses vegades, això vol dir que els casos generats poden simular la realitat. En aquest punt, tenim dues subxarxes: la xarxa associada al generador de casos és una xarxa neuronal desconvulsional i la xarxa associada al discriminador és una xarxa neuronal convolusional. GAN té moltes aplicacions en traduccions, generació d'imatges i validació de prediccions.

Taula 5. Síntesi dels principals mecanismes dels mètodes no supervisats: Xarxes neuronals

Any	Algoritme	Autor	Descripció
1949	Hebbian Learning	Hebb	Utilitzat com a predictor i ponderador, consisteix a ponderar les interaccions entre tots els casos per tal de generar un patró d'aquests pesos.
1986	Restricted Boltzmann machine (RBM)	Smolensky	Utilitzat com a classificador de pes i predictor, consisteix a detectar les relacions entre els nodes en capes (visibles i invisibles) per tal de generar un valor de probabilitat sobre si un cas es troba dins d'una característica.

1987	Autoencoders	LeCun	Utilitzat com a classificador, predictor i ponderador, consisteix a detectar les relacions entre les dades i generar un nou valor a partir del patró d'aquestes relacions.
2006	Deep belief networks (DBN)	Hilton	Utilitzat com a generador de pes classificador i de models predictius, consisteix a detectar el valor ponderat de cada característica i generar prediccions segons les relacions entre els nodes en funció de cada característica.
2014	Generative Adversarial Networks (GAN)	Goodfellow et al.	Utilitzat com a corrector i generador automàtic, consisteix a generar un conjunt de dades falses (a partir de les dades originals) per discriminar-les fins que aquestes dades s'assemblen a les reals.

Font: Taula d'elaboració pròpia.

1.4.3.6 Resum

La **Taula 6** i la **Figura 10**, mostren de forma sintètica els resultats de la revisió sistemàtica detallada anteriorment. En el **Annex III**: Recull de principals softwares per es poden trobar els principals softwares per desenvolupar cadascun dels mètodes presentats.

Taula 6. Resum sintètic dels mètodes supervisats (Classificadors, Classificadors amb penalització als predictors i Classificadors quasi bayesians i altres penalitzadors dels predictors) i mètodes no supervisats (Clustering i Xarxes neuronals)

MÈTODE		**ANY**	**ALGORITME**	**AUTOR**
MÈTODES SUPERVISATS	**CLASSIFICADORS**	1951	K-nearest neighbour classification (K-NN)	Fix & Hodges
			Utilitzat com a mètode de classificació o regressió. Consisteix a agrupar la mostra en m mostres mitjançant k característiques per obtenir m valors per a cadascun dels m grups.	
		1979	Bootstrap	Efron
			Utilitzat com a model inferencial, consisteix a tornar a mostrejar les dades originals i obtenir algunes inferències per als paràmetres a aplicar a la mostra original.	

MÈTODES SUPERVISATS	**CLASSIFICADORS**	1984	Classification and regression trees (CART)	Breiman et al.
			Utilitzat com a model predictiu basat en l'observació d'un conjunt de dades i la subdivisió dels ítems amb n nivells o característiques que es poden utilitzar com a base per a la presa de decisions.	
		1990	Boosting	Schapire
			Utilitzat com a model de mitjana, consisteix a generar un indicador potent fusionant els resultats dels indicadors de més febles a menys per crear un classificador més robust.	
		1994	Bagging	Breiman
			Utilitzat com a model de mitjana, consisteix a crear diferents mostres d'entrenament i utilitzar els resultats com a resultat mitjà combinat.	
MÈTODES SUPERVISATS	CLASSIFICADORS	1995	Random forest (multiple CART)	Ho
			Utilitzat com a mètode de classificació, consisteix a produir un nombre molt gran de CART per a l'entrenament, obtenint-ne els resultats, i crear un classificador o un predictor de tots ells.	
	CLASSIFICADORS AMB PENALITZACIÓ	1970	Ridge regression (RR)	Hoerl & Kennard
			Utilitzat com a mètode de penalització, consisteix a reduir els errors de predicció mitjançant la disminució de la mida de grans coeficients amb l'objectiu de reduir el sobreajustament i obtenir millors resultats amb presencia de multicol·linealitat.	
		1993	Bridge regression (BR)	Frank & Friedman
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix en una generalització d'una regressió de cresta que resumeix les estimacions del coeficient beta i penalitza la suma residual de quadrats.	
			Gibbs sampling	George & McCulloch
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix en una generalització d'una regressió de cresta que resumeix les estimacions del coeficient beta i penalitza la suma residual de quadrats.	
		1995	Non-negative Garrote (NNG)	Breiman
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix a detectar els coeficients principals i controlar els redundants tendint-los a zero per tal d'obtenir un millor predictor.	

MÈTODES SUPERVISATS	**CLASSIFICADORS AMB PENALITZACIÓ**	1996	Least Absolute Shrinkage and Selection Operator (LASSO)	Tibishirani
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix a realitzar la selecció i regularització de la variable per tal d'aconseguir una major precisió de la predicció reduint l'impacte de les variables més febles.	
		2001	Smoothly Clipped Absolute Deviation (SCAD)	Fan & Li
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix a reduir a zero els coeficients més febles, reduir a zero els altres menys febles i mantenir el valor dels coeficients forts. Aquest procés genera un valor imparcial amb dispersió i continuïtat.	
		2002	Adaptative model selection (AMS)	Shen & Ye
			Utilitzat com a mètode de classificació, penalització i predicció penalitzada en funció de la mida del model, permet adaptar-se a múltiples situacions, així com obtenir un baix biaix i una major precisió.	
		2004	Least angle regression (LARS)	Efron *et al.*
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix a determinar quines variables són variables de resposta.	
		2005	Spike-and-slab	Ishwaran & Rao
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix a escollir les característiques i aplicar simultàniament una regressió lineal basada en una teoria bayesiana.	
			Elastic Net	Zou & Hastie
			Utilitzat com a mètode de classificació, penalització i predicció, combina els mètodes LASSO i Ridge. Aquest procés genera una doble contracció i reescala els coeficients.	
		2010	Penalitzacions concoves minimax (MCP)	Zhang
			Utilitzat com a mètode de penalització i predicció, consisteix a realitzar la selecció i regularització de la variable per tal d'aconseguir una major precisió de la predicció reduint l'impacte de les variables més febles.	
		2014	*Bayesian bridge regression*	Zou & Hastie
			Utilitzat com a una extensió de la clàssica regressió de Bridge, facilita l'aplicació del mètode a la pràctica utilitzant els principis bayesians.	

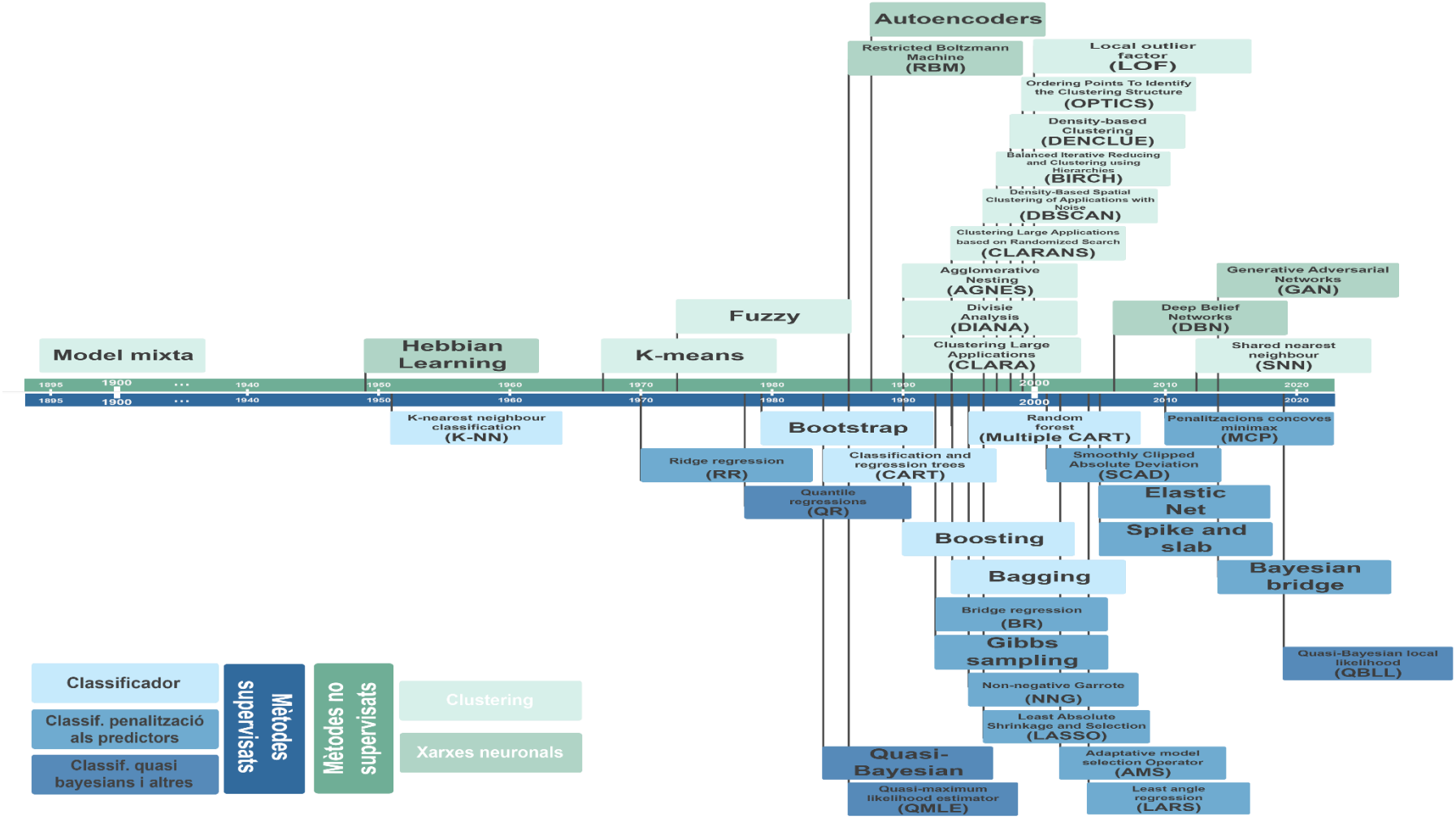
MÈTODES SUPERVISATS	**CLASSIFICADORS QUASI-BAYESIANS I ALTRES**	1984	Quasi-Bayesian	McCray
			Utilitzat com a mètode de classificació, penalització i predicció, permet que una regressió bayesiana funcioni sense conèixer la distribució de dades i obtingui resultats equivalents o similars.	
		1986	Quasi-maximum likelihood estimator (QMLE)	Weiss
			Utilitzat com a mètode de classificació, penalització i predicció, permet que una regressió bayesiana funcioni sense conèixer la distribució de dades i obtingui resultats equivalents o similars.	
		2019	Quasi-Bayesian local likelihood (QBLL)	Petrova *et al.*
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix a maximitzar una funció vinculada al logaritme de la funció de probabilitat per tal d'obtenir un valor.	
	1978		Quantile regressions (QR)	Koenker & Bassett
			Utilitzat com a mètode de classificació, penalització i predicció, consisteix a estimar els mitjans condicionals o altres quantils de la variable resposta.	
MÈTODES NO SUPERVISATS	**CLUSTERING**	1894	Model mixta	Person
			Utilitzat com a agrupador i ponderador, consisteix a dividir el conjunt de dades en k subgrups per obtenir una inferència sobre les propietats de cada subgrup a partir dels casos agrupats.	
		1967	K-means	MacQueen
			Utilitzat com a classificador i predictor, consisteix a dividir un conjunt de dades d'un nombre de grup de k, assignant els casos més propers a partir de la minimització de les distàncies entre cada cas i grup.	
		1973	Fuzzy	Dunn
			Utilitzat com a classificador i predictor, consisteix a dividir un conjunt de dades amb un nombre de grup de k assignant els casos en diferents conjunts de grups en una partició difusa, variant entre 0 i 1, fins a trobar una la agrupació òptima.	
		1986	Partitioning Around Medoids (PAM)	Kaufman & Rousseeuw
			Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades en k medoids per obtenir una inferència sobre les propietats de cada subgrup a partir dels casos agrupats.	

MÈTODES NO SUPERVISATS	**CLUSTERING**	1990	Clustering Large Applications (CLARA)	Kaufman & Rousseeuw
			Utilitzat com a classificador i predictor, consisteix a dividir de forma aleatòria el conjunt de dades en diferents submostres. Després clusteritzar cada subconjunt de dades per extreure els k medoids i trobar la clusterització amb major similitud entre cada medoid i les seves observacions associades per obtenir la clusterització final.	
			Divisie Analysis (DIANA)	Kaufmann & Rousseeuw
			Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades de forma jerarquia de forma aglomerativa (damunt a envall) per generar una classificació dels casos.	
		1990	Agglomerative Nesting (AGNES)	Kaufmann & Rousseeuw
			Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades de forma jerarquia de forma divisisva (damunt a avall) per generar una classificació dels casos.	
		1994	Clustering Large Applications based on Randomized Search (CLARANS)	Ng & Han
			Utilitzat com a classificador i predictor, consisteix a dividir el conjunt de dades en k medoids de forma iterativa fins a trobar la distribució més òptima de k.	
		1996	Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	Ester *et al.*
			Utilitzat com a agrupador i predictor, consisteix a dividir un conjunt de dades amb un nombre de grup de k en funció de la densitat dels casos. Això permet adaptar-lo a diferents tipus de bases de dades.	
		1997	Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)	Zhang et al.
			Utilitzat com a agrupador, enfocat a grans volums de dades, desglossa les dades en petits resums que s'agrupen en lloc dels punts de dades originals. Els resultats capturen la informació de la distribució.	
		1998	Density-based Clustering (DENCLUE)	Hinneburg & Keim
			Utilitzat com a agrupador i predictor en un conjunt de dades amb un gran nivell de soroll, consisteix en identificar un clúster a partir de la influència dels casos entre ells en la generació de grups.	
		1999	Ordering Points To Identify the Clustering Structure (OPTICS)	Ankerst *et al.*
			Utilitzat com a agrupador i predictor, consisteix a emmagatzemar les dades en clústers segons la densitat de cadascun dels clústers corresponents a un ampli conjunt de paràmetres d'un algorisme de clúster basat en densitat.	

MÈTODES NO SUPERVISATS	**CLUSTERING**	2000	Local outlier factor (LOF)	Ertöz *et al.*
			Utilitzat com a classificador, consisteix en una identificació a partir de la densitat dels casos allunyats del seu veí.	
		2003	Shared nearest neighbour (SNN)	Ertöz, Steinbach & Kumar
			Utilitzat com a agrupador i predictor, consisteix a identificar diferents clústers amb diferents formes, mides i densitats en un gran conjunt de dades multidimensional.	
		2013	BIRCH meets coresets for k-means clustering (BICO)	Fichtenberger et al.
			Utilitzat com a agrupador ràpid de grans volums de dades, calcula un nucli S, per cada centre C, el cost d'entrada P es pot aproximar calculant el cost de generar S.	
	XARXES NEURONALS	1949	Hebbian Learning	Hebb
			Utilitzat com a predictor i ponderador, consisteix a ponderar les interaccions entre tots els casos per tal de generar un patró d'aquests pesos.	
		1986	Restricted Boltzmann Machine (RBM)	Smolensky
			Utilitzat com a classificador, ponderador i predictor, consisteix a detectar les relacions entre els nodes en capes (visibles i invisibles) per tal de generar un valor de probabilitat sobre si un cas es troba dins d'una característica.	
		1987	Autoencoders	LeCun
			Utilitzat com a classificador, predictor i ponderador, consisteix a detectar les relacions entre les dades i generar un nou valor a partir del patró d'aquestes relacions.	
		2006	Deep Belief Networks (DBN)	Hilton
			Utilitzat com a ponderador i classificador de models predictius, consisteix a detectar el valor ponderat de cada característica i generar prediccions segons les relacions entre els nodes en funció de cada característica.	
		2014	Generative Adversarial Networks (GAN)	Goodfellow et al.
			Utilitzat com a corrector i generador automàtic, consisteix a generar un conjunt de dades falses (a partir de les dades originals) per discriminar-les fins que aquestes dades s'assemblen a les reals.	

Font: Taula d'elaboració pròpia.

Figura 10.Resum sintètic dels mètodes supervisats (Classificadors, Classificadors amb penalització als predictors i Mètodes supervisats: Classificador quasi bayesians i altres penalitzadors dels predictors) i no supervisats (Clustering i Xarxes neuronals)



Font: Gràfic d'elaboració pròpia

2 JUSTFICACIÓ

Les eines i metodologies que es poden utilitzar en un entorn Big Data són variades. Aquest ventall d'eines dificulta la possibilitat d'escollir-ne una que permeti fer una bona anàlisi sobre un fenomen i que no quedi mitigada o distorsionada per una mala elecció de l'instrument d'anàlisis. En funció de la tipologia de dades, volum d'informació, l'estudi estadístic a realitzar o futures innovacions o perspectives de l'entorn poden condicionar quin mecanisme s'ha d'escollir. Per focalitzar l'àmbit d'estudi de la tesi, s'ha usat el marc conceptual dels determinants socials de la salut.

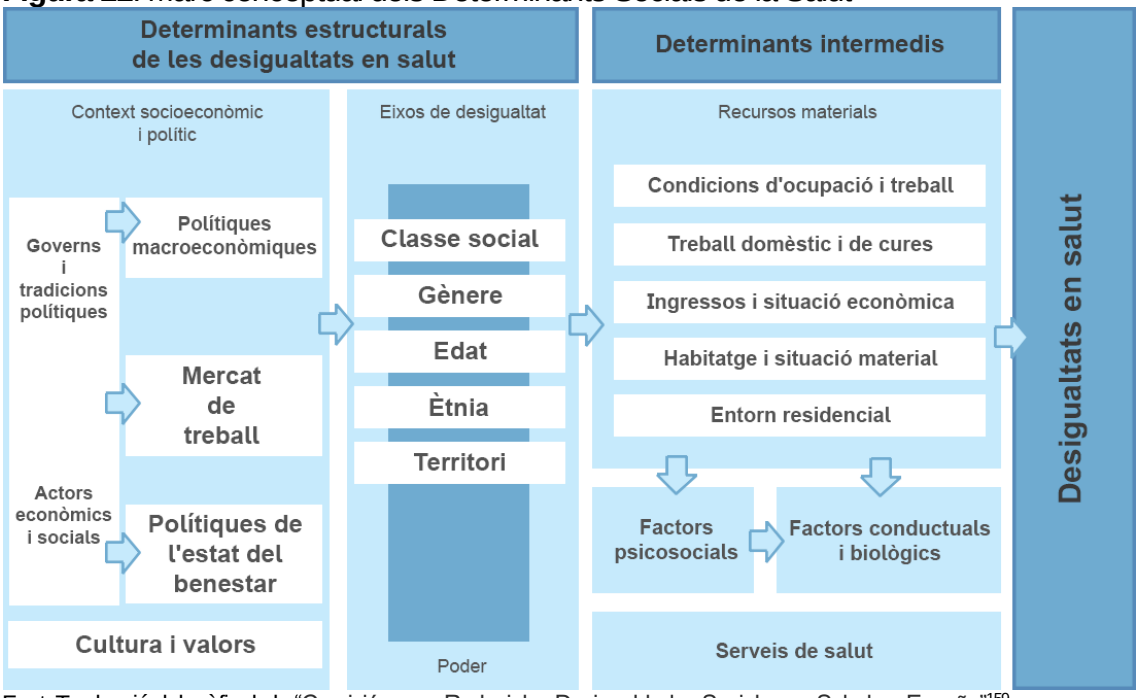
Els determinants socials de la salut (DSS) són aquelles circumstàncies en què les persones neixen, creixen, treballen, viuen i envelleixen. També ho són algunes circumstàncies més estructurals com: els sistemes de salut o les polítiques econòmiques i socials[145]. Les diferències evitables de la salut entre els individus d'una societat teixeixen un entramat de diferències que es coneix com a desigualtats.

El marc conceptual de les desigualtats socials en salut, veure **Figura 11**, té tres components principals: determinants estructurals, eixos de desigualtat i recursos materials. Els determinants estructurals es refereixen a les circumstàncies macrosocials que estan implantades en una societat, com ara: les polítiques econòmiques, l'estratificació social, la distribució del poder entre els estaments o bé la cultura i els seus valors. Els eixos de desigualtat, són aquells elements que jerarquitzen la societat: la classe social, el gènere, l'edat, el territori o l'ètnia. Els recursos materials, condicionats pels determinats estructurals, es refereixen a l'accés dels individus a uns béns o serveis de qualitat que tenen un impacte sobre les seves condicions de vida.

La versió moderna dels determinants socials, va ser proposada per Lalonde el 1974[146] i ampliada per Dahlgren & Whitehead el 1991[147]. L'OMS el 2008, va crear la comissió de determinants socials de la salut que va articular com els determinants de la salut generaven les desigualtats. No obstant, el marc està està en constant evolució[148–153] i crítica[154–157].

Amb la incorporació dels Objectius de Desenvolupament Sostenible, de l'agenda 2030 de l'ONU[158], la salut ambiental ha pres especial rellevància a l'hora d'estudiar les desigualtats en salut. També ha tingut repercussió en l'estudi dels seus determinants, ja que alguns estan estretament relacionats amb diversos factors ambientals.

Figura 11. Marc conceptual dels Determinants Socials de la Salut



Font: Traducció del gràfic de la "Comisión para Reducir las Desigualdades Sociales en Salud en España"[159].

Més enllà del marc teòric, és necessari identificar i generar noves fonts d'informació que permetin mesurar l'evolució de la desigualtat dels territoris. Així, es pot realitzar un seguiment i dotar de coneixement als diferents actors perquè puguin actuar per reduir les situacions de vulnerabilitat.

També és necessari desenvolupar models estadístics i economètrics que permetin diferenciar entre quins elements generen diferències o desigualtats dins d'una societat. A més, aporten informació sobre els elements que tenen un impacte directe, tant negatiu com positiu, sobre la desigualtat, per ajudar a orientar polítiques en favor de les persones en situació de vulnerabilitat..

3 OBJECTIUS

Aquesta tesi pretén avaluar quins són els mecanismes economètrics que s'utilitzen per tractar dades en entorns Big Data. Per abordar aquest objectiu general, es van definir els següents objectius específics:

1. Realitzar una revisió sistemàtica de la literatura acadèmica sobre els mecanismes economètrics més rellevants i àmpliament utilitzats en l'anàlisi de dades en entorns Big Data.
2. Recopilar conjunts de dades relacionats amb les distribucions demogràfiques, socioeconòmiques, geogràfiques i ambientals de la població de Catalunya.
3. Utilitzar mètodes estadístics adequats per analitzar i interpretar les dades recopilades d'àrees petites a Catalunya i estudiar-ne les desigualtats resultants on les poblacions estan repartides de forma heterogènia al llarg del territori.

4 RESULTATS

Els resultats de la tesi es presenten en els següents articles originals:

Caixa 2. Article I

CITACIÓ: Perafita X, Saez M. Clustering of Small Territories Based on Axes of Inequality. *International Journal of Environmental Research and Public Health 2022, Vol 19, Page 3359*. 2022;19(6):3359. doi:10.3390/IJERPH19063359

[**Impact Factor (2021)**: 4.614 (Q1 Public Health, Environmental and Occupational Health, posició 139 de 562)]

Caixa 3. Article II

CITACIÓ: Perafita X, Saez M. Housing Supply and How It Is Related to Social Inequalities—Air Pollution, Green Spaces, Crime Levels, and Poor Areas—In Catalonia. *International Journal of Environmental Research and Public Health*. *2023, Vol 20, Page 5578. 202320(*8):5578. https://doi.org/10.3390/ijerph20085578

[**Impact Factor (2021)**: 4.614 (Q1 Public Health, Environmental and Occupational Health, posició 139 de 562)]

4.1 Article I

Clustering of Small Territories Based on Axes of Inequality

Perafita X, Saez M

International Journal of Environmental Research and Public Health: 2022 March 12. doi:10.3390/IJERPH19063359

Caixa 4. Síntesis del article I

Context de l'article

- S'està dissenyant una cohort electrònica per mostrejar els municipis de la província de Girona.
- La província de Girona presenta una heterogeneïtat geogràfica i de distribució poblacional.
- Menys d'un 10% dels municipis supera els 10.000 habitants.

Què aporta l'article?

- Els resultats obtinguts mostren com la base de dades suavitzada presenta una menor variabilitat a l'hora d'agrupar els municipis.
- S'observa com en totes les representacions existeixen patrons geogràfics.
- Els resultats mostren com la capital de la província és un valor atípic, en totes les agrupacions i per totes les bases de dades.
- La metodologia aplicada mostra com els algoritmes enfocats a treballar amb soroll: OPTICS, DBSCAN i SUBCLU no agrupen de forma òptima, ja que els casos no es poden ignorar.

International Journal of
Environmental Research and Public Health

MDPI

Article

Clustering of Small Territories Based on Axes of Inequality

Xavier Perafita [1,2] and Marc Saez [2,3,*]

[1] Observatori—Organisme Autònom de Salut Pública de la Diputació de Girona (Dipsalut), 17003 Girona, Spain; xperafita@dipsalut.cat
[2] Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, 17003 Girona, Spain
[3] CIBER of Epidemiology and Public Health (CIBERESP), 28029 Madrid, Spain
* Correspondence: marc.saez@udg.edu; Tel.: +34-972-418338; Fax: +34-972-418032

Abstract: Background: In the present paper, we conduct a study before creating an e-cohort for the design of the sample. This e-cohort had to enable the effective representation of the province of Girona to facilitate its study according to the axes of inequality. Methods: The territory under study is divided by municipalities, considering these different axes. The study consists of a comparison of 14 clustering algorithms, together with 3 data sets of municipal information to detect the grouping that was the most consistent. Prior to carrying out the clustering, a variable selection process was performed to discard those that were not useful. The comparison was carried out following two axes: results and graphical representation. Results: The intra-cluster results were also analyzed to observe the coherence of the grouping. Finally, we study the probability of belonging to a cluster, such as the one containing the county capital. Conclusions: This clustering can be the basis for working with a sample that is significant and representative of the territory.

Keywords: big data; clustering; hierarchical k-means; e-cohort; classifiers; machine learning; inequalities

Citation: Perafita, X.; Saez, M. Clustering of Small Territories Based on Axes of Inequality. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3359. https://doi.org/10.3390/ijerph19063359

Academic Editor: Paul B. Tchounwou

Received: 7 February 2022
Accepted: 10 March 2022
Published: 12 March 2022

1. Background

Currently, the concept of "health inequalities" refers to the impact that factors, such as wealth; education; employment; racial or ethnic group; exposure to environmental factors, including air pollution or weather variables; urban or rural residences; and/or the social conditions of an individual's workplace or dwelling, have on the distribution of health and disease among the population. The study of the characteristics of the population and the geographical area of residence is the methodological support that allows for intervention points focused on the prevention and the disappearance of existing health inequalities to be identified.

Initially, socioeconomic inequalities were identified with health inequality [1]. Health inequality can be defined as an inequity in the spread of a disease. In other words, health inequality is the systematic and potentially avoidable differences in one or more health aspects across socially, economically, demographically, or geographically defined populations or population groups. Two conditions must be met for a difference in health to be considered as an inequality: (1) it must be considered socially unjust and (2) potentially avoidable (i.e., there are instruments available that could be used to avoid it) [1].

There is evidence that inequalities in health exist. While the Ladonde [2] and Black [3] Reports pointed this out, it was the Acheson Report [1] that firmly concluded that inequalities in health have a socioeconomic explanation. To date, twenty years later, most of these relationships have been demonstrated, and not an insignificant proportion is caused by environmental problems [1]. These factors are generally, but not exclusively, linked to gender, social and economic conditions [1,4,5].

In general, the living environment, and thus environmental conditions, can contribute to socioeconomic inequalities in health, either independently or, more likely, jointly [1,5]. The first is differential exposure: the most economically disadvantaged groups has a

greater exposure to environmental problems, including, air pollution. The second is differential susceptibility to exposure (i.e., the main adverse health effects) resulting from environmental problems, which occur among the most economically disadvantaged people due to their greater vulnerability.

When we think about a longitudinal study to observe how health inequalities, individuals' health, income, or another specific characteristic evolve over time, our thoughts very quickly turn to creating a cohort. This is immediately followed by considerations of the high cost and logistical difficulties of managing a cohort in terms of obtaining users, processing the sample, managing the information, and even handling and looking after the sample.

There are many cohorts in which the number of individuals easily surpasses 100,000 marks, including the Framingham Heart Study [6] the Current Management of Secondary Hyperparathyroidism: A Multicenter Observational Study (COSMOS) [7], and the NutriNet-Santé Study [8]. When the sample is large, the governance of the user and their data become extremely costly. The sample is acquired in the traditional way, via a letter explaining to the individual concerned that they have been selected to take part in a project and what it consists of involves some costs that are sufficiently high as to consider alternatives to the cohort [9–12]. Another point of consideration is that the cost of increasing, improving, or simply demonstrating the significance for a group or subgroup that was not initially contemplated can be so high that many researchers decide not to incorporate any more individuals into the cohort beyond a theoretical framework. Financial constraints and a lack of logistical resources are factors that generally mean that traditional cohorts have limits.

This is where digital considerations come into play. An *electronic-cohort* or an *e-cohort* is a traditional but digitally managed cohort [13]. This management can be entirely digital via user interactions with websites, platforms, apps, or by post [9]. It can also be of a hybrid nature, depending on the type of information needed to be previously collected and the level of difficulty of obtaining the information automatically. Some traditional cohorts, some of them *novel cohorts* with a high number of individuals, are starting to test the transformation of traditional cohorts into electronic cohorts, seeking their improvement. These improvements basically focus on optimizing the cost/efficiency of the project and obtaining and managing data.

The marginal cost of the sample in an e-cohort is practically zero [11], although some costs inherent to longitudinal studies and linked to maintaining and managing the sample remain. They are, nonetheless, significantly lower than the cost of traditional acquisition. This cost reduction not only signifies monetary savings, but also logistical ease in terms of the human factor. Currently, the e-cohorts that have published results focus on using a webapp as the working platform, sometimes including external elements, such as smartwatches [14] or diaries that must be kept up [9], with the user being able to choose different format. These external elements end up not being used by the individuals, causing sample mortality and making this a weakness of e-cohorts that needs to be addressed [10,11,14] to be able to obtain data without the user having to directly intervene with the app or the mobile phone.

The e-cohort also reduces the costs linked to data collection, minimizing the logistical costs of obtaining, cleaning, homogenizing, processing, and automating all the information concerning the sample. In a cohort, the time spent purging everyone's information quickly adds up to many hours, while digitally doing so allows for "interviewing" the sample, thus eliminating the time spent on this task. We must also consider that the information is obtained in this way just once or twice a year, especially if the sample is large. This lack of information about the user during certain periods causes a data lag, generating an information gap that the traditional cohort cannot resolve. The e-cohort enables different and several surveys to be carried out at no extra economic cost, although consideration must be given to ensure that the sample is not saturated with activity.

In e-cohorts, the data can be obtained in different ways, which, for the sake of simplification, can be separated into two groups: the first where the user interacts, and the second where the user is "passive". In the first, the user interacts directly with the website, app, or mobile device, and consciously responds to the information requested, such as answering a survey or a question about their perceived state of health. Although users' fatigue thresholds have not yet been established, the e-cohort is an attractive option, thanks to the possibility of asking more users more questions at a lower cost. In addition, all the answers enter a digital process where they are easily automated, further reducing the cost and increasing the efficiency of the process. The same logic can be applied to the use of external elements, for example, a smartwatch that can supply minute-by-minute information about the evolution of an individual's heart rate. The results obtained using these tools are unbiased compared to the data obtained using traditional tools, and they also provide information that is consistent over time.

It has been demonstrated that the most effective way to gather users for a sample is by offering a monetary incentive [9,12,13], which the user receives once they have responded to the questions.

There has been a case in which the sample was opened up by applying citizen science. In these cases, the e-cohorts have to buy their sample with a census, or via a similar means, to validate whether the sample obtained is representative of the study population [11,13]. The sample must be validated by separating the different demographic characteristics. In various cases, it has been observed that there are groups that do not tend to take part in these experiences, so additional efforts are required to sample these groups correctly. Conversely, young women with a higher educational level tend to participate most in this type of initiative, leading to their oversampling [14]. This can cause biases, which must be controlled when performing the inferences. It has also been shown that a population with little or no digital skills find responding to the questions problematic. Despite this limitation, very few individuals emerge to complicate the sampling of specific groups [11].

One common limitation of the cohorts that is not resolved by the e-cohort emerges when seeking a way to use a sample to represent a set of territories. If we want to significantly represent the population of Catalonia, it is sufficient that it is random throughout the territory. Meanwhile, if we want to work with a specific axis, such as age, it is sufficient to make a small adjustment and increase the size of the sample.

The Public Health Observatory of Girona Province (Dipsalut) is designing an e-cohort to carry out a longitudinal study to simultaneously examine the health of the population and its socioeconomic situation. The province of Girona is defined as a semi-rural territory [15], with 221 municipalities and a population of approximately 770,000 people. Less than 10% of the municipalities have more than 10,000 inhabitants, substantially limiting statistical significance and causing us to encounter the limitations of the statistical secret.

This e-cohort must not only allow us to obtain a significant representation of all the municipalities in the territory, but it must also optimize the resources and the sample. A municipality codified as LAU level 2 by Eurostat is the smallest existing territorial division at the national level in Spain, where there is a decision-making power over local policies. The present paper explains the process of carrying out clustering in the province of Girona. The clustering must allow similar municipalities to be clustered for the purpose of constructing a representative sample of the different territories. This sample must enable the generation of a set of indicators that present the inequalities that exist in the territories [16]. Furthermore, its design must revolve around the five major axes of inequality: sex, age, social class, migratory process, and territory. This sample was controlled and had to be regulated, so working with an open sample was not a consideration.

This paper explains the process used to cluster the municipalities into 6 groups according to their similarities, and how 14 clustering algorithms were tested to find the ones were the most effective and representative of the province. Finally, statistical modeling was used to observe if there were significant differences between the clusters to draw the final conclusions.

Int. J. Environ. Res. Public Health 2022, 19, 3359

2. Methods

2.1. Methods Prior to Carrying out the Study, the Data Set, and the Data Sources

As explained earlier, the diversity of the territory of Girona requires a large number of variables to determine the differences and similarities between its municipalities. These differences can range from an economic point of view, where the main cities in the province have a larger number of specific companies and sectors, to the migratory processes that the areas experience or the number of elderly people who live there. As can be seen in Figure 1, we carried out a review of all the indicators that exist in the main databases that provide information on the municipalities in the province of Girona. From this, we obtained 541 variables. These were then processed based on the availability of data for the study period, data availability for most municipalities throughout the study period, as well as the elimination of variables that we considered to be duplicates or redundant, and those that did not contribute any relevant information to the study.

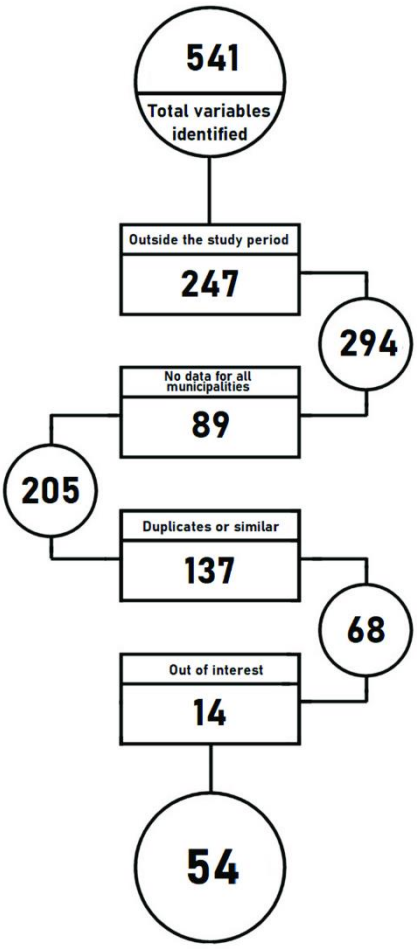


Figure 1. Debugging the process of all the detected variables up to the final data model. Source: authors' own elaboration.

Prior to the clustering, a final set of 54 potential variables encompassing the areas of demography, economy, job market, public spending, health, and populational and geographical incidences and emergencies were identified.

2.1.1. Data Sources

The data used were supplied by different official sources. They were the Statistical Institute of Catalonia (IDESCAT) [17–23], Xifra [24–30], Open Data Generalitat [31–34], the Department of Territory and Sustainability of the Government of Catalonia [35,36], and the National Statistics Institute (INE) [37]. Obtaining information for 199 of the 221 municipalities was difficult because many of them are bound by the obligation of the statistical secret due to the small population figures of below 10,000.

2.1.2. Demographic Area

Ethnic and cultural diversity and populational polarization have positive repercussions on the economy and generate cultural and social combinations [38]. Migratory movements also have an effect on the socioeconomic levels of the population [39], causing modifications to the diseases and states of health linked to the populational pyramid that can lead to changes in health policies.

The following indicators were used to evaluate the demographic situation of each municipality: the average age of the population, total population, population resident abroad [23], net migration and population [22], immigration rate and the native population index [27], and population density [30].

2.1.3. Economic Area

Economic capacities can determine the significant differences between the inhabitants of a municipality. As described in the literature [40], poverty does not solely consist of the economic capacity of a person to meet minimum expenses, but it also has implications in terms of health, education, and the chance to save money to have a better quality of life. The standard of living can also be determined by access to basic goods, such as housing.

The following indicators were used to evaluate the economic status of each municipality: personal income tax [20], the result of the tax return per declarant [28], and gross income per person [37]. The following types of indicators were collected to evaluate the degree of poverty in each municipality: the distribution of the sources of income and the Gini index [24].

The state of housing was also included as an economic indicator, because a direct relation between the state of housing and the economy of a municipality is considered to exist, including the number of residences, average rental price [40], cadastral value and number of urban plots, and number of immovable properties and their cadastral value [24].

2.1.4. The Job Market Area

A municipality's job market shows the type of employment that exists in that area and the predominant sector. Depending on the sector, the industry and the working conditions linked to the different sectors of a municipality, the lifestyle of the people that live there, are positively affected to varying degrees [41].

The following indicators were used to evaluate the job market of each municipality: social security affiliations, according to the registered home address of the affiliated person and the activity sector; social security affiliations according to the percentage of the active foreign-born population [17]; unemployment [26]; unemployment among foreign-born persons [26]; and the temporary employment rate [25].

2.1.5. Area of Public Spending

Public spending shows the amount of money spent by the local government of a municipality to cover the needs of its inhabitants. There is discussion in the literature as to whether an increase in public spending has a direct impact on citizens and their levels of poverty [42–44], health [45], and education [46].

The following indicators were used to evaluate the public spending of each municipality: the number of libraries [18] and sports facilities [32].

2.1.6. Area of Health

This area considers the state of health of the inhabitants of a municipality. Given that the territories were generally very small, we had access to data that were more purely biological. Traffic accidents were also observed as they impact the health of a territory and its preventive strategies, focusing on pedestrians, cyclists, cars, and motorcycles [47]. Aging must also be considered in this area, since it is one of the most predominant demographic phenomena in Europe in the twenty-first century. There are indexes that show how aging has different effects on the population in terms of fertility, age, and birth rate [48]. This phenomenon involves some specific public policies that have a direct impact on the population and their state of health.

The following indicators were used to evaluate the state of health of each municipality: the number of births and deaths and the gross mortality rate [19] and birth rate [29]; the number of traffic victims to evaluate the possible impacts on the inhabitants of a municipality [49]; the variables of the aging and global dependency indexes [30]; and the Synthetic Fertility Index and the natural population growth [29].

2.1.7. Area of Population Incidences and Emergences

The incidences and emergencies of the inhabitants of each municipality show the population's one-off and recurrent needs in terms of the emergency services. There are social factors that generally contribute to the use of these services [50]. The following indicator was used to evaluate the incidences in each municipality: the number of emergency phone calls [31].

2.1.8. Geographic Area

The geography of the province of Girona is diverse and varied. There are coastal, mountainous, and flat areas, and the geographical characteristics of each area is instrumental in the development of a type of commerce and populational structure. The following indicators were used to evaluate the geography of each municipality: the extension of herbaceous crops [34] and woody cultivation [34], land extension in km^2, and the singular entities in each municipality [21]. The altitude, latitude, and longitude of each municipality were added later [21], in addition to whether it was a county capital. Additionally, included was whether these municipalities were in a mountainous area [36] or coastal [35]. These variables show the different types of environments and their geographical positions.

2.1.9. Alternative Data Sets

Two databases parallel to the working one were developed: a nominal data set and a smoothed data set. These had to enable the observation of whether the smoothing of data or the transformation of the indicators from a percentual to a nominal value improved the cluster forming. In the nominal data set, the data was obtained from the sources mentioned above. A z-score transformation was performed for the smoothed data set [51]. The same number of variables was maintained in both datasets.

2.2. Control of Missing Value or Statistical Confidentiality

There was a set of data that was lost because they are bound by the obligation of the statistical secret, so they could not be collected. In these cases, an estimated value was assigned to each of those lost sets.

2.3. Variable Selection

We carried out a variable selection process, spike and slab, according to the population [52]. The aim was to eliminate the redundant variables and excessive noise. Other methods for selecting variables were also employed: Ridge Regression [53,54], LASSO [55], Elastic Net [56], SCAD [57], MCP [58] and LARS [59].

2.4. Cluster Analysis

Once the variables were selected, a clustering process was carried out to detect the municipalities that were similar among them. Given that the data set represented such different types of municipalities, it was decided to carry out a preliminary task with 14 different algorithms. This process was required to observe the algorithms that adapted best to the type of data, which is why they were of different types: partitional, hierarchical, one-pass, density-based, and big data clustering.

Among the partitional clustering methods, the following were used: k-means [60], Partitioning Around Method (PAM) [61], Clustering Large Applications (CLARA) [61], fuzzy clustering [61], CLARANS [62], and EA [63]. The hierarchical methods employed were Divisive Analysis (DIANA) [61], Agglomerative Nesting (AGNES) [61], and hierarchical k-means [64]. Additionally, the density-based methods used were SUBCLU [65], DBSCAN [66], and OPTICS [66]. The big data and One Pass methods were BICO [67] and BIRCH [68].

The cluster analyzed responds to a grouping based on a measure of distance where each observation initially acts as a cluster.

$$X = \{x_i | i = 1, \ldots, r\} \text{ com a dades base } A$$

$$A = \{a_i | i = 1, \ldots, n\} \text{ for } n = 53$$

These clusters fuse iteratively together, depending on their proximity until no more of them can be fused.

$$Ci = \{c_{ij} | j = 1, \ldots, K\} \text{ for } K = 6$$

Each new fusion can generate a new centroid in each cluster.

$$D = \{d_i | i = 1, \ldots, K\} \text{ for } K = 6$$

Mapping of the Clustering

The clusterings created using the hierarchical k-means algorithm were represented to evaluate whether they followed a geographical pattern on the map of the region under study (i.e., Girona). The map was created for three points in time, 2015, 2016, and 2017. The maps of the municipalities were obtained from the Cartographic and Geographic Institute of Catalonia [69]. The mapping was also used to observe whether there was a variation in the municipalities over the years.

2.5. Data Analysis

A multinominal logistic regression was carried out, for which the dependent variable (π_j) is the cluster generated, where $j = 1, 2, 3, 4, 5$. The variable of the reference group was 6, modeled in the following way:

$$logit(\pi_j) = \log\left(\frac{\pi_j}{\pi_6}\right) = x^T \beta_j, \text{ for } j = 1, \ldots, 5$$

It was adjusted as follows to find the estimated probability ($\hat{\pi}_j$) of the events:

$$\hat{\pi}_6 = \frac{1}{1 + \sum_{j=2}^{j=6} exp(x^T \hat{\beta}_j)}, \text{ for } j = 6$$

$$\hat{\pi}_j = \frac{exp(x^T \hat{\beta}_j)}{1 + \sum_{j=2}^{j} exp(x^T \hat{\beta}_j)}, \text{ for } j = 1, \ldots, 5$$

The final result enables the clusters to be compared with the municipality of Girona.

2.6. Software

All the analyses were carried out using the free R software. The packages used were *glmnet*, *ncvreg*, *lars*, *spikeslab*, and *data sets* for the variable selection method; *data sets*, *stats*, *factoextra*, *cluster*, *dbscan*, *subspace*, *stream*, *clv*, *stream*, and *fpc* for the clustering and validation of the clusters; *nlme*, *tidyverse*, *moments*, and *nnet* for mining the data; and *factoextra*, *ggplot2*, *gridExtra*, *cowplot*, *rgdal*, and *tmap* for the graphic representation.

3. Results

3.1. Area and Period of Study

A process of clustering small areas of Catalonia using a set of 54 variables was carried out. A prior task was performed to select the variables that were most relevant to the different areas, as explained in the following sections.

The study period was initially 2010 to 2018. However, given the small dimensions of both the territory and population, the data are bound by the obligations of the statistical secret, presenting limitations regarding accessing the available information. Consequently, the study period was changed to 2015–2017, when the data are more consistent and relatively unproblematic regarding lost values. All the municipalities were therefore represented by a high level of consistency.

In this study, we considered 221 of the 948 municipalities belonging to the region of Catalonia. The number of inhabitants varied between 83 and 99,013 (average inhabitants: 3412, standard deviation: 9081.349 inhabitants, median inhabitants: 746, Q1 298 inhabitants, and Q3 2290 inhabitants). The population density varied between 1 and 4493 inhabitants per km^2 (average: 45 inhabitants/km^2, standard deviation: 464.216 inhabitants/km^2, median inhabitants: 45/km^2, Q1 20 inhabitants/km^2, and Q3 130 population/km^2).

3.2. Variable Selection

To eliminate the redundant variables and excessive noise, we carried out a variable selection process, spike and slab, according to the population [52]. The models were based on the relationship with respect to the number of inhabitants in a municipality. The mean squared error of the predictions was used as a method comparison criterion [70]. The spike and slab method presents the smallest mean squared error (MSE) (see Table 1).

Table 1. Study of the number of optimal variables from different variable selection methods, according to the MSE.

Method	MSE	Number of Variables	
		Selected	Non-Selected
Ridge Regression	25,981.73	54	0
Lasso	55,404.96	12	42
Elastic Net	70,199.54	53	1
SCAD	50,711.94	14	40
MCP	50,711.94	16	38
LARS	41,167.40	34	17
Spike and Slab	25,302.36	53	1

Source: author's own elaboration.

The dimensions of the final dataset are defined in 54 variables for 221 municipalities over 3 years, thus obtaining a final sample of 35,802 cases.

3.3. Clustering

The number of clusters obtained from the supervised methods was six (Figure 2). This number was validated based on the application of the Elbow method in a task carried out

prior to the process of clustering. The number of optimized clusters does not change in any of the three data sets.

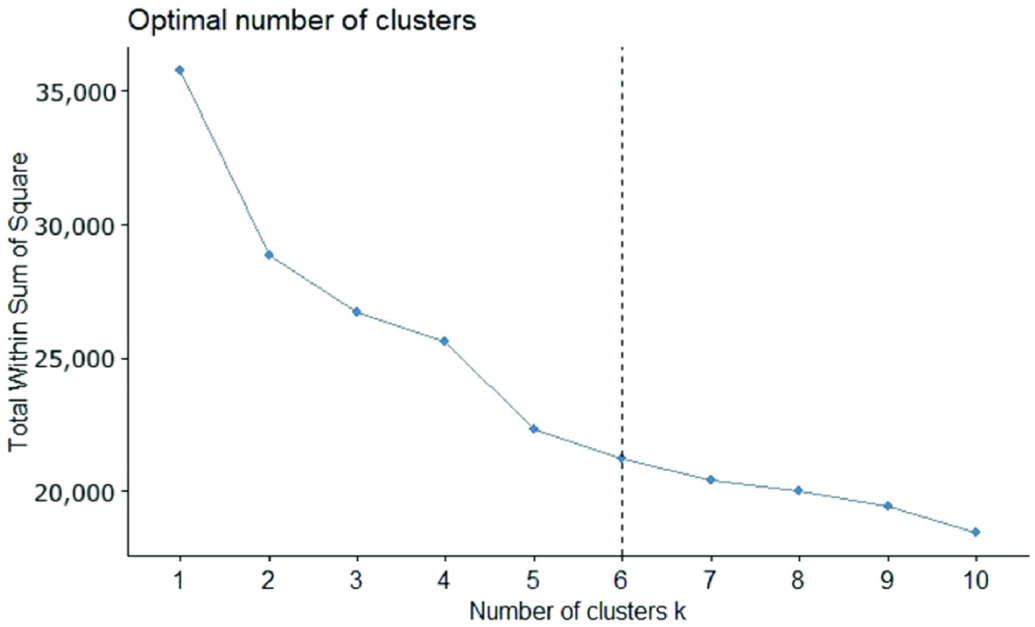


Figure 2. Process of obtaining the optimal number of clusters from the Elbow method. Source: authors' own elaboration.

The results of the clustering process are presented in Table 2 (external and internal validation of clustering), Table 3 (number of observations for each cluster and data set), and Figures 3 and 4 (results of clustering).

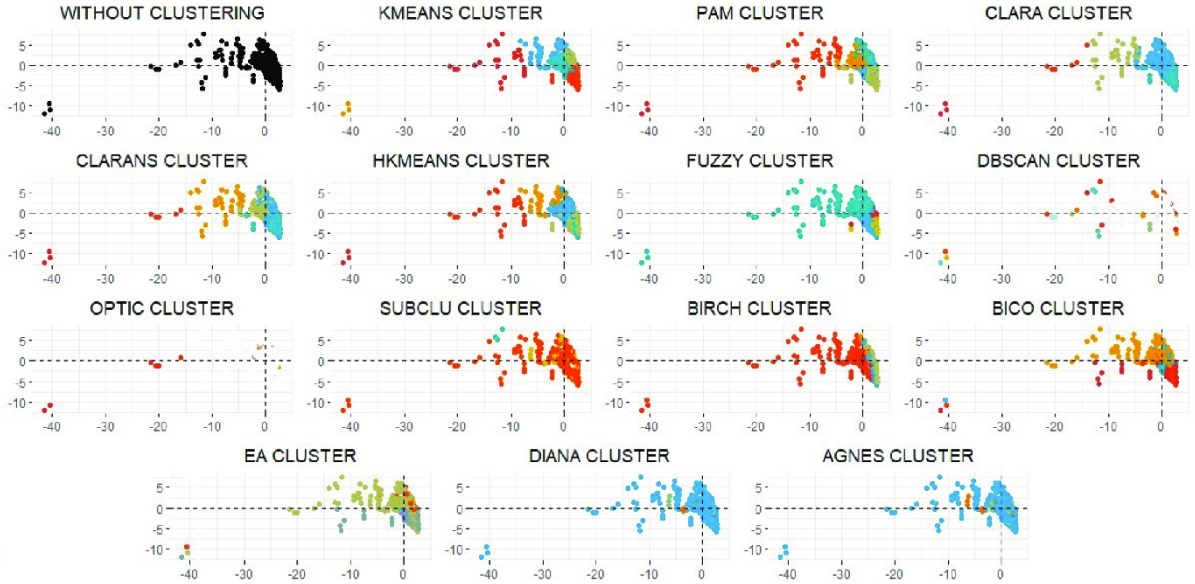


Figure 3. Representation of the different algorithms performed to study the clustering of municipalities. Source: authors' own elaboration.

Table 2. External and internal validation of clustering.

Name	Nº Clusters	Noise Point	Avg Between	Avg Within	Avg Silhouette	DUNN Index	Entropy	WB Ratio	CH Index	Separation Index
Data Set: Original										
K-MEANS	6	0	9.962	7.569	0.084	0.087	1.407	0.760	91.998	2.877
PAM	6	0	9.836	7.639	0.065	0.065	1.509	0.777	85.240	2.567
CLARA	6	0	10.499	8.070	0.074	0.038	0.961	0.769	59.973	2.488
CLARANS	6	0	10.064	7.766	0.070	0.068	1.206	0.772	83.459	2.739
HKMEANS	6	0	10.407	7.639	0.120	0.078	1.217	0.734	89.323	3.174
FUZZY	3	0	9.928	9.000	0.067	0.025	0.580	0.907	27.103	1.716
BIRCH	6	0	9.232	8.614	−0.073	0.029	1.671	0.933	18.810	2.437
BICO	6	0	9.560	8.539	−0.030	0.024	1.343	0.893	16.487	2.304
EA	6	0	9.560	8.539	−0.030	0.024	1.343	0.893	16.487	2.304
DIANA	4	0	12.017	9.128	0.024	0.044	0.256	0.760	6.536	3.020
AGNES	4	0	10.363	9.130	−0.072	0.044	0.422	0.881	5.193	2.886
Data set: Nominal										
K-MEANS	6	0	9.191	5.266	0.196	0.065	1.241	0.573	278.179	2.232
PAM	6	0	7.641	7.379	−0.101	0.011	1.509	0.966	3.012	1.006
CLARA	6	0	8.728	5.459	0.123	0.037	1.228	0.625	244.045	1.403
CLARANS	6	0	8.694	5.358	0.137	0.037	1.293	0.616	256.041	1.499
HKMEANS	6	0	9.195	5.268	0.195	0.065	1.240	0.573	278.081	2.241
FUZZY	4	0	9.109	6.255	0.077	0.015	0.862	0.687	141.029	1.567
BIRCH	6	0	7.609	6.744	−0.137	0.008	1.563	0.886	25.802	1.031
BICO	6	0	7.808	6.465	−0.008	0.012	1.517	0.828	26.727	1.259
EA	6	0	7.808	6.465	−0.008	0.012	1.517	0.828	26.727	1.259
DIANA	4	0	7.158	7.470	−0.089	0.012	0.243	1.044	0.789	1.440
AGNES	4	0	6.461	7.451	−0.233	0.008	0.422	1.153	1.680	1.165
Data set: Z-score										
K-MEANS	6	0	10.149	7.759	0.061	0.104	1.241	0.765	83.072	2.789
PAM	6	0	9.352	9.103	−0.039	0.036	1.509	0.973	3.456	2.374
CLARA	6	0	10.013	7.846	0.060	0.093	1.228	0.784	78.518	2.382
CLARANS	6	0	10.014	7.766	0.079	0.099	1.293	0.775	83.033	2.422
HKMEANS	6	0	10.15	7.758	0.061	0.104	1.240	0.764	83.097	2.771
FUZZY	4	0	10.263	8.413	0.038	0.049	0.862	0.820	65.039	2.406
BIRCH	6	0	9.266	8.711	−0.116	0.040	1.563	0.940	16.763	2.478
BICO	6	0	9.434	8.515	−0.028	0.039	1.517	0.903	20.023	2.400
EA	6	0	9.434	8.515	−0.028	0.039	1.517	0.903	20.023	2.400
DIANA	4	0	8.796	9.185	−0.087	0.051	0.243	1.044	1.643	3.160
AGNES	4	0	8.491	9.194	−0.156	0.039	0.422	1.083	1.554	2.841

Source: authors' own elaboration.

Table 3. Distribution of the number of cases according to the data set and the type of grouping.

Name	Cluster 1 (C1)			Cluster 2 (C2)			Cluster 3 (C3)			Cluster 4 (C4)			Cluster 5 (C5)			Cluster 6 (C6)		
	O [1]	N [2]	Z [3]	O [1]	N [2]	Z [3]	O [1]	N [2]	Z [3]	O [1]	N [2]	Z [3]	O [1]	N [2]	Z [3]	O [1]	N [2]	Z [3]
K-MEANS	25	44	127	258	127	15	62	15	3	121	360	360	3	114	114	194	3	4
PAM	235	235	235	165	117	117	122	97	97	92	30	30	46	3	3	3	181	181
CLARA	425	239	328	163	95	115	58	284	25	5	38	6	10	4	2	2	3	187
CLARANS	347	277	235	166	122	117	101	49	97	41	5	30	5	3	3	3	207	181
HKMEANS	355	360	360	41	132	132	183	109	109	57	44	44	24	15	15	3	3	3
FUZZY	170	345	597	492	33	34	1	1	1	0	284	28	0	0	3	0	0	0
BIRCH	44	32	32	96	43	44	161	219	170	50	132	53	126	50	128	186	187	236
BICO	95	198	193	35	48	101	6	3	3	331	179	122	151	74	153	45	161	91
EA	95	3	3	45	161	91	331	198	193	151	179	122	35	74	153	6	48	101
DIANA	627	630	612	24	18	33	9	12	15	3	3	3	0	0	0	0	0	0
AGNES	594	594	594	33	33	33	33	33	33	3	3	3	0	0	0	0	0	0

Source: authors' own elaboration. [1]: cluster based on the original data set. [2]: cluster based on the nominal data set. [3]: cluster based on the z-score data set. Source: authors' own elaboration.

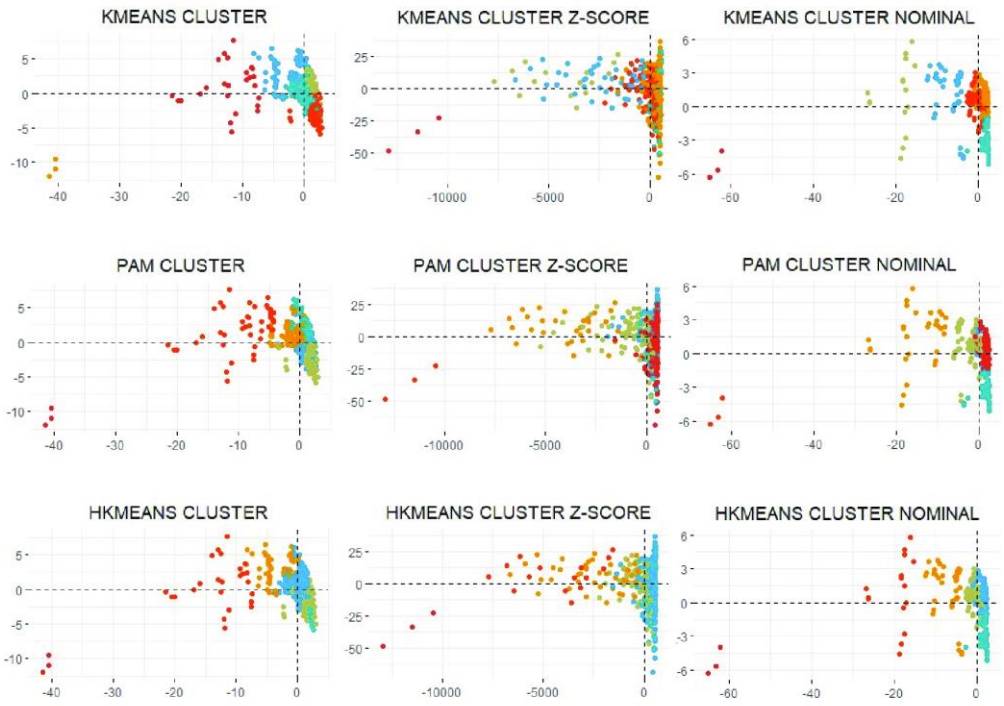


Figure 4. Representation of the results of the k-means, PAM, and hierarchical k-means algorithms for the different data sets. Source: authors' own elaboration.

The diversity of the municipalities in Girona presents a well-recognized heterogeneity. The capital has a little over 100,000 inhabitants (103,369 inhabitants), while there are less than 50,000 (47,235 inhabitants) in the next largest municipality. There is also important diversity in a geographical sense, with a set of municipalities located in mountainous areas and others located on the Mediterranean coast. This heterogeneity across the entire area generates some obvious socioeconomic and health differences. The density-based clustering algorithms do not work this heterogeneity optimally. Many municipalities, including the capital of the province, are detected as outliers. This type of algorithm does

not allow all the municipalities to be classified, and so they were ruled out. However, the rest of the models classified all the municipalities (see Figure 3).

An external and internal validation study was carried out to choose between the rest of the algorithms. A graphic validation was later designed using a cloud of points and the mapping of the clusters. The clustering produced by the hierarchical k-means method was consequently chosen.

As shown in Table 2, the internal validation values [71] of the algorithms, k-means, hierarchical k-means, PAM, and CLARANS, present the optimum values in the original database. In the nominal and smoothed data set, we observe how the PAM algorithm obtains some internal validation results that are inferior to the rest of the previously mentioned algorithms.

The external validation shows how PAM is the algorithm presenting a difference between inferior clusters in all the data sets. However, the intra-cluster difference varies depending on the data set. The three algorithms that present the relation of the most optimum intra-between cluster differences can be highlighted: k-means, PAM, and hierarchical k-means. The entropy value [71] that shows the best clustering is presented in the fuzzy, DIANA, and AGNES algorithms for the different data sets. The CH index [72] shows how the k-means, PAM, CLARANS, and hierarchical k-means algorithms are the ones that present the best construction of the clusters.

Table 3, which shows the distribution of the clusters, helps with the conceptualization of the dimensions of the clusters. It can be observed how the different clustering has a main cluster in the original data set, which has a greater number of cases than the rest. This main cluster varies from 186 to 627 in the different algorithms. There are two types of clustering: those in which the main cluster captures most cases, and those in which the cases are distributed more homogeneously between the clusters. In most of the groupings, there is a second cluster with a weight greater than 20% for all the observations. The groupings in which the main cluster retains at least 50% of the sample are CLARA, CLARANS, hierarchical k-means, fuzzy, BICO, EA, DIANA, and AGNES. Meanwhile, k-means, PAM, and BIRCH are the algorithms that distribute the individuals in the most balanced way. The nominal and smoother data sets present a more uniform distribution of the clusters in the municipalities.

Once the validations of the clusters and their dimensions have been analyzed, a graphic representation of them must be produced. This representation must allow the algorithms that generate a visually intuitive clustering to be detected to facilitate choosing the final clustering (Figure 3).

Figure 4 shows how the k-means, PAM, and hierarchical k-means algorithms are the dimensions that generate a more visually intuitive clustering for the different data sets. The representations based on the nominal data set show how the distribution is reduced. In the smoothed data set, the cases are smoothed in a more obvious manner.

The graphic representation using the clouds of points does not allow a pattern that is significantly better than the rest to be detected. Therefore, Figure 5 shows the groupings of the k-means, PAM, and hierarchical k-means algorithms on the study map (province of Girona).

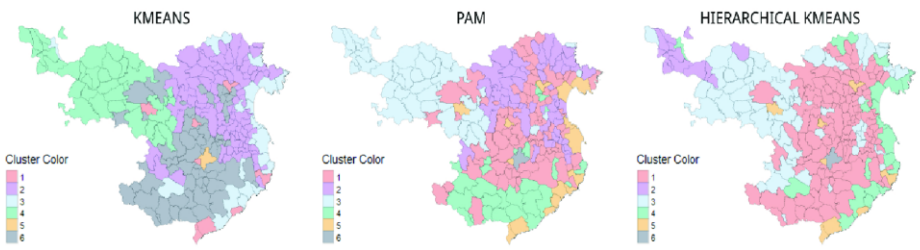


Figure 5. Representation of the cluster map: k-means, PAM and hierarchical k-means, according to the normal dataset, to observe their spatial distribution. Source: authors' own elaboration.

3.4. Mapping of the Clustering

The maps illustrate how the clustering carried out using the original data set enables us to detect that the k-means and hierarchical k-means algorithms differentiate between the set of coastal municipalities and some county capitals together. They also cluster the set of inland municipalities that link Barcelona and France. They do not detect a differentiation between the mountain municipalities, although they do differentiate between a subregion of them. A small cluster for some of the municipalities with a high population is generated. Regarding PAM, the mountain and coastal municipalities are clearly differentiated. Some county capitals are also added to these last clusterings. A set of municipalities very close to Barcelona and the municipalities nearest the French border can be identified, as can the inland municipalities dispersed in a first and second ring around the county capitals. In all three clusterings, Girona is grouped independently.

The clusters generated by the k-means, PAM, and hierarchical k-means algorithms, based on the nominal and smoothed data sets, are very similar. The k-means and hierarchical k-means algorithms detect the first grouping of the municipalities located in the mountainous areas. K-means detects a subset of these municipalities since they belong to the inland municipalities. Both algorithms also detect a set of municipalities that belong to the coast, together with some county capitals. The municipalities nearest the French border and those closest to Barcelona are detected. Meanwhile, PAM detects a pattern among the municipalities next to France (Figures 6 and 7).

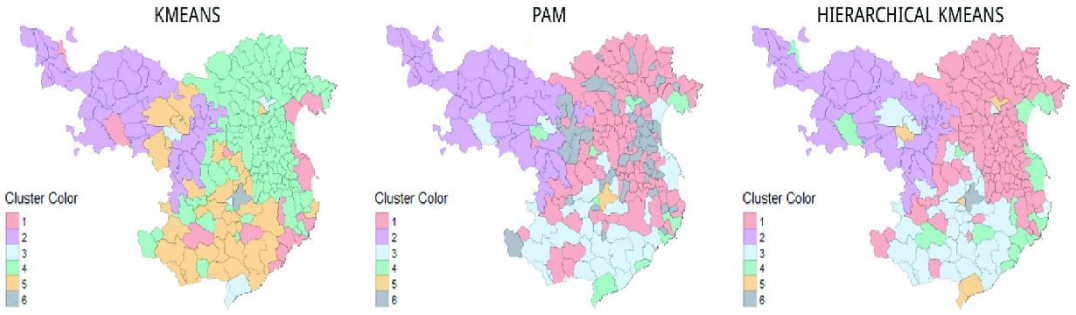


Figure 6. Representation of the cluster map: k-means, PAM, and hierarchical k-means, according to the nominal dataset, to observe their spatial distribution. Source: authors' own elaboration.

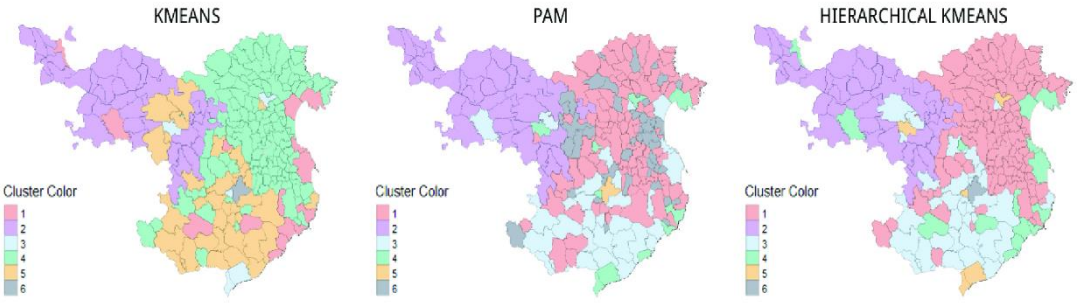


Figure 7. Representation of the cluster map: k-means, PAM, and hierarchical k-means, according to the z-score dataset, to observe their spatial distribution. Source: authors' own elaboration.

3.5. Descriptive Study of the Clustering

Table 4 shows the variability of the clusterings. Notably, the k-means and hierarchical k-means algorithms are the data sets with the least variability in all three data sets, indicating that these clusterings do not undergo changes and are stable over time.

Table 4. Measurement of the number of cases that vary between clusters to study the variability of results.

	0 Changes	*1 Changes*	*2 Changes*	*0 Changes*	*1 Changes*	*2 Changes*	*0 Changes*	*1 Changes*	*2 Changes*
	Data Set: Original			*Data Set: Nominal*			*Data Set: Z-Score*		
K-MEANS	202	19	0	217	4	0	217	4	0
PAM	120	98	3	74	142	5	74	142	5
CLARA	155	65	1	56	162	3	56	162	3
CLARANS	181	40	0	56	162	3	56	162	3
HKMEANS	196	25	0	217	4	0	217	4	0
FUZZY	54	167	0	10	210	1	10	210	1
BIRCH	172	46	3	197	24	0	197	24	0
BICO	172	46	3	196	25	0	173	48	0
EA	172	46	3	197	24	0	173	48	0
DIANA	172	46	3	197	24	0	221	0	0
AGNES	172	46	3	197	24	0	221	0	0

Source: authors' own elaboration.

The algorithm chosen is hierarchical k-means, because it presents the optimum and secure properties to generate a sample that endures over the years. Six clusters can be detected in this algorithm. The first cluster contains the municipalities near the French border (Empordà), and the second contains the municipalities located in mountainous areas. The third group focuses on the inland municipalities of the territory. The fourth group is made up of the coastal municipalities and some provinces in the county. The fifth group detects the territory's important municipalities, be it economically or in terms of population. The sixth and last group separates the capital from the rest of the municipalities.

The results of the descriptive study of the clustering are shown in Table 5 (descriptive analysis by conglomerates, robust values). As can be observed, the size of the population is very different among the six groups. There is an obvious contrast between the high number of people that live in the capital (98,255) and the median population of the municipalities located in the other county capitals (37,042) and close to the coast (10,709), with lower population numbers than the rest of the cluster. The population density is also higher in these groups, and especially in the capital (2512). It can be observed how the native population figures are quite similar for all the clusters, except the capital, where this figure is higher (40.22). Meanwhile, the ratios of immigrants in the inland municipalities (0.082) and the mountainous areas (0.061) are lower than in the rest of the clusters, with the highest ratios in the coastal municipalities (0.217) and the other county capitals (0.225).

The internal and external flow of movements is greatest in the capitals of the county (28) and in the capital of the province (555). The migratory balance is also higher in the capital than in the rest of the clusters. The different weights in the distribution of jobs in the sectors in each cluster can also be observed. The mountain and border clusters (7.23 and 6.61, respectively) have the highest percentage of the population employed in agriculture. Meanwhile, the inland municipalities (20.98) have a higher percentage of the population employed in the industrial sector. The weight of the construction sector is similar in all the clusters, except for the capital, which has a lower percentage (4.67). The services sector predominates in all the clusters, with the greatest weight (81.94) in the capital. The unemployment rate increases in line with the weight of the population of each cluster. Likewise, the clusters with the highest population densities are where the Gini index is highest.

Inequality is greatest in the capital (36), followed by the coastal municipalities (34.60) and the main county capitals (34.10).

Income from salaries is highest in the capital (10,277) and lower in the coastal municipalities (7393) and the county capitals (7218). Income derived from unemployment benefits is lower in the coastal municipalities (2488) and the county capitals (2221).

Table 5. Descriptive analysis by conglomerates, only robust values (median (1st quartile–3rd quartile)).

FRENCH BORDER (C1)	*MOUNTAIN (C2)*	*INLAND (C3)*	*COASTAL (C4)*	*OTHERS (C5)*	*CAPITAL (C6)*	*FRENCH BORDER (C1)*	*MOUNTAIN (C2)*	*INLAND (C3)*	*COASTAL (C4)*	*OTHERS (C5)*	*CAPITAL (C6)*
n = 360	***n = 132***	***n = 109***	***n = 44***	***n = 15***	***n = 3***	***n = 360***	***n = 132***	***n = 109***	***n = 44***	***n = 15***	***n = 3***
pob_res_alestranger						cadastre_parcel_u					
12 (6–25)	10 (4–16)	35 (14–65.25)	324 (276–456)	855 (685–1263)	4160 (3941–4339.5)	407.5 (235.25–698.5)	407.5 (194.75–608.75)	1330.5 (957.5–2527.5)	4540 (2292–7232)	6541 (5850.5–9234)	10,649 (10,645–10,677)
saldo_migratori_intern						cadastre_inmo_u					
1 ((−7)–8.25)	1 ((−5)–6)	4.5 ((−8)–30.25)	−8 ((−42)–33)	−2 ((−30.5)–55)	9 ((−39.5)–51)	463.5 (247.25–878.5)	579.5 (242–1139)	2372.5 (1166–4260.75)	15,982 (8132–21,720)	32,691 (26,451.5–38,173.5)	79,579 (79,242–79,713.5)
saldo_migratori_extern						cadastre_valor					
2 (0–6)	1 (0–4)	8 (0–18)	48 (2–85)	3 ((−16.5)–212.5)	546 (310–699)	22,797.5 (13,956.5–44,019.5)	24,778.5 (11,902–70,339.75)	144,625.5 (64,595–250,763.25)	686,228 (312,990–12,862,52)	1,374,697 (1,291,656–2,110,633.5)	4005,166 (3,806,792.5–4,036,354.5)
saldo_migratori_total						atur_mig					
2 ((−5)–11)	3 ((−2.25)–7)	14.5 ((−4.5)–40.25)	22 ((−11)–107)	28 ((−6.5)–202.5)	555 (361–659.5)	22.71 (9.79–43.46)	10.5 (4.83–36.605)	124.915 (56.603–279.955)	625.17 (477.92–948.83)	2190.42 (1582.955–3230.75)	5730.42 (5447.835–6093.585)
irpf_base_imp						atur_mig_estranger					
20,129 (18,708.25–21,803.5)	19,582.5 (17,367–21,283.5)	20,578.5 (19,228.75–21,582.5)	18,577 (17,700–19,991)	18,736 (17,123–19,331.5)	24,800 (24,443–25,100)	2.96 (1.08–7.123)	0.96 (0.06–3.455)	15.54 (3.958–36.293)	184.33 (127.67–346.58)	619.08 (432.5–820.5)	1644.83 (1552.415–1774.29)
irpf_couta_auto						inde_env					
5129.5 (4538.25–5815)	4831.5 (4158–5676)	4745 (4381.25–5296)	4749 (4540–5065)	4513 (4197–4740.5)	6647 (6615.5–6733)	130.255 (101.812–157.438)	154.23 (119.182–192.27)	93.6 (82.613–117.955)	97.55 (88.81–120.07)	81.53 (77.885–116.02)	81.95 (81.38–85.05)
nascuts_vius						tax_bruta_mort					
4 (2–9)	3 (1–7)	27.5 (14–55.25)	100 (84–143)	302 (290–342.5)	1048 (1041.5–1074.5)	9.16 (6.455–12.795)	9.05 (5.695–13.413)	7.805 (6.412–9.773)	8.42 (7.85–9.32)	7.68 (6.265–8.95)	7.22 (7.215–7.42)
morts_num						index_rec					
2 (1–4)	1 (0.75–3)	13.5 (7–23.25)	42 (29–61)	131 (106–137.5)	344 (338.5–357)	147.22 (110–196.243)	158.57 (124.52–217.957)	109.7 (99.032–129.367)	107.47 (100.38–146.15)	108.23 (101.18–112.64)	96.05 (95.695–96.14)
saldo_pobl						index_dep_glob					
2 (0–5)	1 (0–4)	16 (6–29.25)	65 (40–88)	193 (161.5–208.5)	704 (684.5–736)	60.595 (54.788–64.713)	56.185 (49.905–62.543)	54.47 (52.33–56.37)	54.04 (52.64–54.87)	51.12 (42.165–52.085)	50.06 (49.785–50.24)
mobilitat_estudiants_uni_foramun						edat_mitja					
33.511 (5–20)	57.586 (0–11.25)	44.974 (40–105)	111.312 (135–300)	272.496 (597.5–660)	58.381 (1120–1177.5)	44.15 (42.2–45.725)	45.40 (43.6–47.2)	41.50 (40.275–43.225)	41.50 (40.8–43)	41.30 (39.9–42.5)	40.0 (39.9–40.1)

Table 5. *Cont.*

FRENCH BORDER (C1)	*MOUNTAIN (C2)*	*INLAND (C3)*	*COASTAL (C4)*	*OTHERS (C5)*	*CAPITAL (C6)*	*FRENCH BORDER (C1)*	*MOUNTAIN (C2)*	*INLAND (C3)*	*COASTAL (C4)*	*OTHERS (C5)*	*CAPITAL (C6)*
n = 360	***n = 132***	***n = 109***	***n = 44***	***n = 15***	***n = 3***	***n = 360***	***n = 132***	***n = 109***	***n = 44***	***n = 15***	***n = 3***
mobilitat_estudiants_uni_mun						creix_natu					
0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)	10,785 (10,737.5–10,890)	−1 ((−3)–1)	−1 ((−3)–1)	2 ((−4)–11.25)	12 (0–28)	3 ((−22.5)–120.5)	326 (313–336)
renda_mitja						index_sint_fecund					
12,195 (11,375.5–13,261.5)	13,084 (12,102.75–14,568.25)	12,304 (11,315.25–13,375.25)	10,629 (9818–11,665)	10,104 (9600.5–10,938.5)	13,183 (12,930.5–13,355)	1.28 (0.838–1.74)	1.3 (0.768–1.74)	1.415 (1.175–1.675)	1.45 (1.3–1.61)	1.350 (1.2–1.74)	1.45 (1.435–1.49)
total_pobl						taxa_estreng					
579 (284.75–1035.75)	340.5 (181.75–829)	3525.5 (1713.5–5474.5)	10,709 (10,231–17,677)	37,042 (33,972–39,096)	98,255 (97,920.5–98,634)	0.101 (0.066–0.138)	0.061 (0.049–0.112)	0.082 (0.044–0.117)	0.217 (0.164–0.298)	0.225 (0.161–0.258)	0.18 (0.179–0.182)
biblio						index_autoc					
0 (0–0)	0 (0–0)	1 (0–1)	1 (1–3)	3 (1–5)	18 (18–18)	30.685 (24.788–34.858)	35.03 (24.377–40.385)	28.525 (23.395–35.197)	33.11 (21.49–38.24)	32.88 (20.445–37.505)	40.22 (40.175–40.22)
ss_total_mig						densitat_pob					
228 (113–405)	145 (83–326)	1539 (793–2293)	4109 (3671–6547)	13,633 (12,460–14,595)	39,427 (38,747–40,147)	41 (20–76)	16 (5–34)	135 (60.75–190)	423 (175–630)	1171 (758.5–2214)	2512 (2503.5–2521.5)
ss_ext_mig						contract_tempo					
9.83 (6.332–14.315)	5.42 (2.015–9.15)	7.535 (3.947–11.123)	17.08 (14.14–22.09)	17.37 (13.51–24.36)	16.77 (16.405–17.125)	0.812 (0.5–1)	0.883 (0.702–1)	0.834 (0.75–0.906)	0.838 (0.794–0.886)	0.861 (0.819–0.902)	0.898 (0.893–0.9)
ss_agricultura_per						gini					
6.606 (3.541–12.228)	7.23 (2.91–12.821)	2.61 (1.487–5.697)	2.572 (1.55–4.059)	1.277 (0.384–2.425)	0.627 (0.596–0.633)	31.3 (28.8–33.6)	31.9 (28.975–34.8)	28.5 (27.4–30.6)	34.6 (32.7–36.1)	34.1 (31.7–36.6)	36 (35.45–36.1)
ss_industria_per						renda_bruta_mitja					
11.765 (8.747–16.981)	15.155 (6.744–23.149)	20.977 (15.936–28.685)	9.818 (7.502–14.758)	11.207 (10.869–19.345)	12.768 (12.647–12.857)	14,791 (13,581.5–16,171.5)	15,874 (14,383.75–17,778.25)	14,926.5 (13,472.5–16,295.75)	12,626 (11,634–13,970)	12,011 (11,342.5–12,968)	16,303 (16,006.5–16,559)
ss_construccio_per						renda_salari					
8.889 (6.589–10.714)	7.833 (5.66–10.086)	7.93 (6.58–9.378)	9.756 (7.456–10.343)	6.298 (5.141–6.65)	4.661 (4.655–4.77)	8258.5 (7528–9185)	8732.5 (7775–10,044.5)	9430 (8399–10,639)	7393 (6793–8117)	7218 (6956–7662.5)	10,277 (10,067.5–10,454.5)
ss_serveis_per						renda_pensions					
70.588 (64.057–74.803)	67.458 (60.34–75.506)	65.896 (61.232–72.396)	75.795 (65.677–78.832)	79.983 (69.048–80.24)	81.937 (81.779–82.07)	2861 (2546–3379.75)	3209 (2814.75–3747.25)	2717 (2463.75–2959.75)	2488 (2174–2744)	2221 (1795–2749.5)	2963 (2920.5–3007)
equipament						renda_atur					
0 (0–3.978)	0 (0–6.082)	2.61 (1.768–4.425)	2.06 (1.32–2.78)	0.81 (0.44–2.05)	1.83 (1.825–1.835)	237.5 (189.75–294)	234.5 (184.75–287)	242.5 (209–283.5)	326 (282–358)	305 (255.5–401.5)	245 (235–263.5)

Table 5. *Cont.*

FRENCH BORDER (C1)	*MOUNTAIN (C2)*	*INLAND (C3)*	*COASTAL (C4)*	*OTHERS (C5)*	*CAPITAL (C6)*	*FRENCH BORDER (C1)*	*MOUNTAIN (C2)*	*INLAND (C3)*	*COASTAL (C4)*	*OTHERS (C5)*	*CAPITAL (C6)*
n = 360	***n = 132***	***n = 109***	***n = 44***	***n = 15***	***n = 3***	***n = 360***	***n = 132***	***n = 109***	***n = 44***	***n = 15***	***n = 3***
preu_mig_lloguer						capitalcomarca					
487.73 (432.805–522.745)	472.56 (387.272–514.478)	498.545 (435.03–545.448)	454.62 (408.96–480.18)	422.2 (378.66–434.205)	515.46 (500.545–538.245)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–1)	0 (0–1)	1 (1–1)
num_habitatges						geo_altitud					
6 (3–13)	4 (1.75–11.25)	34.5 (16.75–78.25)	190 (141–296)	842 (712.5–917)	3267 (3199–3291.5)	82 (33.75–161)	953.5 (362–1180.5)	111 (89.75–172)	31 (12–148)	39 (13–260)	70 (70–70)
transit_victim						munt					
111.5 (1–280.25)	121 (1–298.5)	132 (2–260.5)	111 (1–263)	92 (2–263.5)	76 (38.5–186.5)	0 (0–0)	0 (0–0)	0 (0–0)	1 (0–1)	0 (0–1)	0 (0–0)
trucades_emer						costa					
2 (1–6)	2 (1–5)	3 (1–10)	5 (2–23)	5 (1.5–12.5)	1 (1–253.5)	0 (0–0)	1 (1–1)	0 (0–0)	0 (0–0)	0 (0–1)	0 (0–0)
super_conreu_herb						latitud					
15 (4–67.5)	10.5 (3–23.75)	10 (3–21)	36 (9–102)	4 (2–30.5)	3 (3–9.5)	42.175 (42.038–42.298)	42.257 (42.144–42.35)	41.935 (41.827–42.03)	42.125 (41.917–42.219)	42.182 (41.699–42.237)	41.982 (41.982–41.982)
super_conreu_lleny						longitud					
110.5 (0–279.25)	120 (0–297.5)	131 (0–259.5)	110 (0–262)	91 (0–262.5)	75 (37.5–185.5)	2.946 (2.812–3.04)	2.327 (2.072–2.612)	2.76 (2.638–2.883)	3.073 (2.662–3.129)	2.792 (2.657–2.848)	2.824 (2.824–2.824)

The cost of renting housing is similar among the clusters. However, the cadastral value is not, with the highest values in the capital (4,005,166) and the lowest around the French border (22,797.5).

On observing the breakdown of the population balance, it can be observed how this balance is lower in the mountainous areas (1) and the border areas (2), than in the capital (704) and some other municipalities (193). A similar dynamic appears in relation to the natural growth of the population and the dependency index. The border and mountain municipalities have the same negative natural growth rate (−1) and the highest dependency indexes (60.60 and 56.19). Conversely, there is a higher natural growth rate in the capital and county capitals and a lower dependency index (50 and 48.04). The number of traffic accident victims is similar in all the clusters, except in the capital (76). However, more phone calls are registered in the coastal municipalities (5) and the other county capitals (5) than in the rest of the clusters.

Geographically, it can be observed how the highest municipalities are found in the mountain municipalities (953.5).

3.6. Inference

The clusters represent the variability of the territory, which, as we have shown, is very varied, and therefore these different realities are so different that they do not follow a normal distribution. The Kruskal–Wallis [73] and Mann–Whitney tests [74] show that there are significant differences among the clusters. To observe these differences from the clusters, we assume that we do not have the presence of multiculturalism or outliers. A multinominal logistic regression was performed to observe these differences [75]. The odds ratios of the regression are presented in Table 6.

Table 6. Probability of a municipality belonging to each of the clusters (odds ratio).

	French Border	Mountain	Inland	Coastal	Others
Intercept	0.9992802 (***)	1.0002685 (***)	0.9998384 (***)	1.0004043 (***)	1.0006086
Population residing abroad	1.0083055 (***)	1.0002468 (*)	0.9887537 (***)	1.0011125 (***)	0.9988869
Internal migratory balance	0.9902796 (***)	0.9960504 (***)	1.0013632 (***)	0.9811316 (***)	0.9980404
External migratory balance	1.0023456 (***)	0.9995994 (***)	0.995847 (***)	0.9994775 (***)	0.999062
Taxable base of personal income tax	0.999858	0.999797	1.0000104	1.0009097 (**)	0.9995569
Self-employed income tax contributions	1.000437	1.00045	1.000327	1.000503	1.000133
Number of births	0.9866577 (***)	1.0149668 (***)	0.9957985 (***)	1.0328247 (***)	0.9798185
Number of deaths	0.9873149 (***)	1.0627083 (***)	1.0011853 (***)	1.0125107 (***)	0.9523698
Population balance	0.9993343 (***)	0.9550756 (***)	0.9946195 (***)	1.020063 (***)	1.0288215
Mobility of university students outside the municipality	0.9887809 (***)	0.9835925 (***)	1.0066694 (***)	1.0053018 (***)	1.0113427
Mobility of university students within the municipality	1.0010445 (·)	0.9985847 (***)	1.0020055 (***)	0.9969711 (***)	0.9982406
Average income	0.9991602 (*)	0.9999274	0.9992583 (·)	0.9990069 (·)	0.9996053
Total population	0.9979256 (***)	1.0011634 (***)	0.9986322 (***)	0.9928825 (***)	0.9990906
Library count	0.9920155 (***)	0.9969202 (***)	1.0082681 (***)	1.0105826 (***)	0.9944124
Average number of people registered with social security	1.0036874 (***)	0.9979587 (**)	1.0018299 (**)	1.0142032 (***)	1.0004368
Average number of foreigners registered with social security	0.9832582 (***)	0.9783802 (***)	1.0052646 (***)	1.0045112 (***)	1.0420857
Percentage of workers engaged in the agricultural sector registered with social security	0.9919438 (***)	0.9796335 (***)	0.981117 (***)	1.0864516 (***)	0.9754141

Table 6. *Cont.*

	French Border	Mountain	Inland	Coastal	Others
Percentage of workers in industry registered with social security	0.9765902 (***)	1.0136796 (***)	1.0162364 (***)	0.9900073 (***)	1.0222547
Percentage of workers in the construction sector registered with social security	1.0616384 (***)	1.0078621 (***)	0.9817759 (***)	0.9884191 (***)	0.9674953
Percentage of workers in the services sector registered with social security	1.0120162 (***)	0.9976513 (***)	0.9992377 (***)	0.9597444 (***)	1.0410581
Sports facilities count	0.9866832 (***)	1.0077482 (***)	0.9896743 (***)	1.0081863 (***)	1.0040921
Average rental price	0.9997722	1.0005119	1.0012895 (·)	0.9906604 (***)	0.9999714
Count of homes available for rent	0.9943651 (***)	0.9992177 (***)	0.9999074 (·)	0.9828655 (***)	1.0073002
Emergency calls count	0.9970014 (***)	1.0067798 (***)	1.0035279 (***)	1.015406 (***)	1.0054075
Area of woody cultivation	1.0163713 (***)	0.9976557 (***)	0.9898313 (***)	1.0010048 (·)	0.9971145
Number of properties according to land register	1.0001438	0.9999955	1.0003972 (·)	1.0017136 (***)	1.0002099
Average unemployment	1.007197 (***)	1.001434 (***)	1.006958 (***)	1.005407 (***)	1.004737
Aging ratio	0.9945135 (***)	0.9960614 (***)	0.9962998 (***)	0.9685533 (***)	0.9920995
Active population replacement rate	1.0006615	0.9971274 (**)	0.9963596 (***)	0.9999958	1.0030382
Middle age	1.0023732 (***)	1.0010506 (***)	0.9681195 (***)	1.0175038 (***)	1.0253728
Synthetic fertility rate	0.9998299 (***)	0.9925758 (***)	0.999746 (***)	1.0019436 (***)	1.0064768
Proportion of native born population	0.9526581 (***)	0.993467 (***)	1.0901165 (***)	1.0573994 (***)	0.9445746
Percentage of the number of temporary contracts	1.004969 (***)	0.9980301 (***)	0.9977705 (***)	1.002199 (***)	0.9973226
Average gross income	1.0006414 (*)	1.0004035	1.0006079 (·)	0.9984342 (**)	1.0003484
Average income from pensions	1.0004604	1.000368	0.9998378	1.0016645 (**)	1.0009924
County capital (no)	0.9981343 (***)	0.9978039 (***)	0.9988672 (***)	1.007165 (***)	0.9986435
Municipality located in the mountains (no)	1.0027483 (***)	1.0010293 (***)	0.9967967 (***)	1.0005993 (***)	0.9989383
Latitude	0.9739104 (***)	1.011893 (***)	0.9863013 (***)	1.0184511 (***)	1.0271034
Number of traffic deaths	0.9848847 (***)	1.0027088 (***)	1.0106366 (***)	0.9989089 (·)	1.0035241
Area for herbal cultivation	1.0016808 (*)	0.9997817	0.9872208 (***)	1.0102234 (***)	1.0029894
Number of plots according to land register	1.0003769	1.0006836	1.0015421 (**)	0.9990165 (*)	1.000732
Total cadastral value	0.999998	0.9999899 (*)	0.9999944	0.9999917 (*)	1.0000028
Average foreign born unemployment	1.0109255 (***)	0.9937618 (***)	0.9975241 (***)	1.0235579 (***)	0.9959258
Gross mortality rate	1.0067221 (***)	1.0322658 (***)	0.9621324 (***)	1.0172427 (***)	0.9872505
Overall dependency ratio	1.0779853 (***)	0.9850141 (***)	0.9621172 (***)	1.0158033 (***)	0.9834363
Natural population growth	0.9792645 (***)	1.0495036 (***)	0.9959008 (***)	0.9769135 (***)	0.9939988
Immigration rate	0.9994612 (***)	0.999584 (***)	0.9999851 (***)	1.0002735 (***)	1.0007975
Population density	0.9987993 (*)	0.9982249 (**)	0.9990685 (·)	0.9983053 (*)	0.9988891
Gini index	0.9536537 (***)	1.0278178 (***)	0.9691616 (***)	1.013044 (***)	1.048768
Average income from salary	1.0003585	0.9998022	1.0004226	1.000985 (·)	1.0003853
Average income from unemployment benefits	1.0049536 (***)	1.0021693 (**)	1.0039937 (***)	1.0138551 (***)	0.9964328
Altitude	0.9969976 (***)	1.0022741 (***)	0.9995159	1.0034163 (***)	0.9988485
Municipality located on the coast (no)	0.9786072 (***)	1.0144727 (***)	0.9987252 (***)	1.0047179 (***)	1.0041124
Length	1.002225 (***)	0.9997354 (***)	0.9968903 (***)	1.0015902 (***)	1.0005632

(***) = $p \leq 0.001$. (**) = $p \leq 0.01$. (*) = $p \leq 0.05$. (·) = $p \leq 0.1$. Source: authors' own elaboration.

4. Discussion

The execution of the algorithms and data sets show how the validation improves when working with more stable data, such as the nominal values smoothed by the z score. This stability is translated into less variability in the construction of the clusters in the three periods. The variability improves when working with the smoothed data set. This is a relevant point when considering the design of a longitudinal study to find the individuals that are representative of the same type of municipality.

Of the clustering presented, three of the maps can be identified as the most representative of the territory. The first was the map created with the original data set using the PAM algorithm, which managed to determine six clusters: the French border, the mountainous area, the outskirts of Barcelona, the coast, the area inland, and the capital of the province. The other two were the maps generated with the nominal and smoothed data sets, using the hierarchical k-means algorithm, which showed five clusters: the French border, the mountainous area, the coast, the area inland, and the capital of the province, in addition to a sub-cluster of the main county capitals. The more solid algorithm was chosen at the expense of the loss of the cluster adjoining Barcelona.

The multinominal logistic regression shows that there are differences among the clusters and the capital. There are no significant differences demographically between the municipalities grouped as county capitals and the capital of the province. The clusters of the mountainous areas, the French border, and the coast have the probability of having lower population balances and lower population densities than the capital. Consequently, the probability of having a higher global dependency index than the capital is higher.

Economically, there are less differences between the clusters. Any differences stem from salaries with respect to the capital, giving the coastal areas a lower probability. However, they have a higher probability of obtaining a gross average income and a pension than the capital. On the coast, both the gross average income and income from salaries have a higher probability of being the same as those of the capital. However, pensions have a lower probability.

The probability of having a rental housing offer equal to that of the capital is less in the mountain, border, and coastal clusters. However, the probability of owning property is higher with respect to the capital. Nonetheless, there is less probability that they are valued the same as the capital.

The job market presents significant differences, except for the municipalities in the county capitals. In the rest of the clusters, there is a greater probability of being unemployed than in the capital. Notably, the probability of having an immigrant unemployment rate equal to that of the capital is lower on the coast and in the mountains. The probability of having workers who are employed in the agricultural sector with respect to the probability of the same in the capital is greater in the coastal municipalities, and lower in the other municipalities. The probability of having workers employed in the services sector works inversely.

The probability of having sports facilities and libraries with respect to the capital is higher in the coastal municipalities. We find the inverse in the border municipalities, which have a negative probability. There are no significant differences in the municipalities of the county capitals.

In terms of interpreting the health variables, there are no significant differences with respect to the municipalities of the country capitals. For the rest of the clusters, the probability of having an aging index, similar to that of the capital, is negative. The probability of having the same death rate as the capital is higher in the border municipalities and on the coast, and negative in the others. The recovery index also has a negative probability. In the inland municipalities, the probability of having a mean age equal to that of the capital is lower than in the rest of the clusters. Traffic incidences and victims are more probable in the mountain and coastal municipalities than in those of the capital.

Clear differences were observed between the clusters and the capital. However, few significant differences were observed in the subgroup of the municipalities in the county capitals.

In conclusion, working with microdata is complicated, in terms of both making comparisons and modeling and clustering, especially if they are socioeconomic data. The difficulties of working with indicators, indexes, and rates complicate the data mining process and, later, the reading of the results. A smoothing or standardization process is necessary to work effectively. It must be considered that using percentages with such small data sets mean that these can drastically change from year to year. These possible irregularities accentuate the variations and generate an elevated volatility. This volatility affects the clustering and models, making their classification difficult. These factors end up translating into a high variability of the observations in the groups. However, this way of working can end up impeding the detection of new emerging clusters.

The functions based on density do not work optimally with variables that have such different realties as these. Figure 3 shows how they do not manage to classify all the municipalities. It should be tested whether re-clustering the outliers results in being able to classify all the municipalities, even though this means generating a final clustering superior to the k-number of the chosen clusters. The hierarchical k-means and k-means algorithms generate a cluster that does not present large significant differences with respect to the capital, so we can therefore work with five clusters rather than six. This helps us to design the simplest sample with the possibility of generating the most segregations. Another point for further study is whether the subgroup detected by PAM presents significant differences to the other groups to maintain the six clusters. A priority when designing a clustering to be able to extract a set of individuals to carry out a longitudinal study using digital tools is that these groupings endure for as long as possible.

Another point to bear in mind is that the number of years studied should always be higher than the number of clusters we want to create. This way, we can know in which cluster the municipalities are classified, most times, to be able to find a cluster–territory relationship and a trend. This was not possible in this study due to the lack of data.

5. Conclusions

This article aims to help researchers and other decision-making institutions facilitate a comparison of the structuring and grouping of small areas, especially in those cases where the differences between them are so large. It also endeavors to show an optimal way of transforming and working on datasets to facilitate the resulting groupings. Two of the main limitations in grouping such diverse and small populations is, on the one hand, the lack of data and, on the other, the lack of experiences that endured over time, where we can observe their evolution.

If we want to analyze the impacts of spatial variables such, as NDVI or the pollutants $PM_{2.5}$, PM_{10}, NO_2 or CO_2, it is advisable to generate data at a lower level than the municipality, as municipalities, while not the smallest administrative division, are the smallest division that has political decision-making power. This would allow us to segment a cohort from census tracts or districts in the future and reduce the potential ecological fallacies that cohort data may generate. In addition, it would capture the inequality that can be observed between the rich and poor areas in cities better. The lack of experience working in small areas, along with the nature of most indicators, makes these processes difficult.

Currently, it is essential to start generating data at the scale of small areas, even smaller than those of a municipality, because otherwise we will not always be masking the inequalities through averages and aggregate values of population subsets, in which wealth has blurred the levels of poverty. On the other hand, the microdata permits the creation and adaptation of new indicators that allow the inequalities and the phenomena that occur in the territorial field to be captured more efficiently.

Data protection policies, although necessary, often prevent the study of the reality of territories. They also make it difficult to study individuals in a particular way. These

mechanisms end up making it difficult to observe inequalities as well as study the sensitivity that each individual has, with respect to their social conditions and how these affect them.

To facilitate the best clustering process, it would be useful to carry out trend studies and predictive modeling to observe the subsequent years and to be able to forecast where each municipality will be classified, to help create a clustering that endures over time.

Author Contributions: X.P.: Conceptualization, Methodology, Investigation, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Project Administration. M.S.: Methodology, Investigation, Data Curation, Writing—Review and Editing, Supervision, Visualization, Project Administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data, including the code to produce the figures, can be requested from the first author (xperafita@dipsalut.cat).

Acknowledgments: This study was carried out within the "Cohort-Real World Data" subprogram of CIBER of Epidemiology and Public Health (CIBERESP). We appreciate the comments of two anonymous reviewers of a previous version of this work who, without doubt, helped us to improve our work. The usual disclaimer applies.

Conflicts of Interest: The manuscript is an original contribution that has not been published previously, whole or in part, in any format, including electronically. All authors will disclose any actual or potential conflicts of interest, including any financial, personal, or other relationships with other people or organizations that could inappropriately influence or be perceived to influence their work within three years of beginning the submitted work.

Abbreviations

AGNES	Agglomerative nesting
CLARA	Clustering large applications
DIANA	Divisive analysis
Dipsalut	Public Health Observatory of Girona Province
IDESCAT	Statistical Institute of Catalonia
MSE	Mean squared error
PAM	Partitioning around methods

References

1. Acheson, D. *Independent Inquiry into Inequalities in Health Report*; The Stationary Office: London, UK, 1998.
2. Lalonde, M. *A New Perspective on the Health of Canadians. A Working Document*; Government of Canada: Ottawa, ON, Canada, 1974.
3. Department of Health and Social Security. *Inequalities in Health: Report of a Research Working Group*; Department of Health and Social Security: London, UK, 1980.
4. Deguen, S.; Zmirou-Navier, D. Social inequalities resulting from health risks related to ambient air quality—A European review. *Eur. J. Public Health* **2010**, *28*, 27–35. [CrossRef] [PubMed]
5. Bowen, W. An analytical review of environmental justice research: What do we really know? *Environ. Manag.* **2002**, *29*, 3–15. [CrossRef] [PubMed]
6. Long, M.T.; Fox, C.S. The framingham heart study-67 years of discovery in metabolic disease. *Nat. Rev. Endocrinol.* **2016**, *12*, 177–183. [CrossRef] [PubMed]
7. Cannata-Andía, J.B.; Fernández-Martín, J.L.; Cannata-Andia, J.B.; Fernandez-Martin, J.L.; Zoccali, C.; London, G.M.; Locatelli, F.; Ketteler, M.; Ferreira, A.; Covic, A.; et al. Current management of secondary hyperparathyroidism: A multicenter observational study (COSMOS). *J. Nephrol.* **2008**, *21*, 290–298. [PubMed]
8. Hercberg, S.; Castetbon, K.; Czernichow, S.; Malon, A.; Mejean, C.; Kesse, E.; Touvier, M.; Galan, P. The nutrinet-santé study: A web-based prospective study on the relationship between nutrition and health and determinants of dietary patterns and nutritional status. *BMC Public Health* **2002**, *10*, 142. [CrossRef] [PubMed]

9. Chatzitheochari, S.; Fisher, K.; Gilbert, E.; Calderwood, L.; Huskinson, T.; Cleary, A.; Gershuny, J. Using new technologies for time diary data collection: Instrument design and data quality findings from a mixed-mode pilot survey. *Soc. Indic. Res.* **2017**, *137*, 379–390. [CrossRef] [PubMed]
10. McManus, D.D.; Trinquart, L.; Benjamin, E.J.; Manders, E.S.; Fusco, K.; Jung, L.S.; Spartano, N.L.; Kheterpal, V.; Nowak, C.; Sardana, M.; et al. Design and preliminary findings from a new electronic cohort embedded in the framingham heart study. *J. Med. Internet Res.* **2019**, *21*, e12143. [CrossRef] [PubMed]
11. Pouchieu, C.; Méjean, C.; Andreeva, V.A.; Kesse-Guyot, E.; Fassier, P.; Galán, P.; Hercberg, S.; Touvier, M.; Paolotti, D.; Kaliraman, V. How Computer literacy and socioeconomic status affect attitudes toward a web-based cohort: Results from the nutrinet-santé study. *J. Med. Internet Res.* **2015**, *17*, e34. [CrossRef]
12. Toledano, M.B.; Smith, R.B.; Brook, J.P.; Douglass, M.; Elliott, P. How to establish and follow up a large prospective cohort study in the 21st century—Lessons from UK COSMOS. *PLoS ONE* **2015**, *10*, e0131521. [CrossRef] [PubMed]
13. Kesse-Guyot, E.; Assmann, K.; Andreeva, V.; Castetbon, K.; Méjean, C.; Touvier, M.; Salanave, B.; Deschamps, V.; Péneau, S.; Fezeu, L.; et al. Lessons learned from methodological validation research in e-epidemiology. *JMIR Public Health Surveill.* **2016**, *2*, e160. [CrossRef]
14. Spartano, N.L.; Lin, H.; Sun, F.; Lunetta, K.L.; Trinquart, L.; Valentino, M.; Manders, E.S.; Pletcher, M.J.; Marcus, G.M.; McManus, D.D.; et al. Comparison of on-site versus remote mobile device support in the framingham heart study using the health eheart study for digital follow-up: Randomized pilot study set within an observational study design. *JMIR mHealth uHealth* **2019**, *7*, e13238. [CrossRef] [PubMed]
15. Methodology—Rural Development—Eurostat. Available online: https://ec.europa.eu/eurostat/web/rural-development/methodology (accessed on 11 June 2021).
16. Amat, P.B.; Lazaro-Lasheras, L.; Oliveras, S.; Perafita, X.; Tarrés, A.; Vilà, A. Promoting equity through monitoring inequalities in the semi-rural region of Girona. *Eur. J. Public Health* **2020**, *30* (Suppl. 5), ckaa166.306. [CrossRef]
17. IDESCAT. Afiliats I Afiliacions a la Seguretat Social Segons Residència Padronal de L'afiliat. Available online: https://www.idescat.cat/pub/?id=afi (accessed on 10 March 2021).
18. IDESCAT. Enquesta de Biblioteques. Available online: https://www.idescat.cat/pub/?id=bib (accessed on 31 December 2021).
19. IDESCAT. Estadística de Naixements. Available online: https://www.idescat.cat/pub/?id=naix (accessed on 31 December 2021).
20. IDESCAT. Impost Sobre la Renda de les Persones Físiques. Available online: http://www.idescat.cat/pub/?id=irpf (accessed on 31 December 2021).
21. IDESCAT. Indicadors Demogràfics I de Territori. Available online: http://www.idescat.cat/pub/?id=inddt&n=215 (accessed on 31 December 2021).
22. IDESCAT. Moviments Migratoris. Available online: https://www.idescat.cat/pub/?id=mm (accessed on 31 December 2021).
23. IDESCAT. Padró D'inhabitants Residents a L'estranger. Available online: https://www.idescat.cat/pub/?id=phre (accessed on 31 December 2021).
24. XIFRA. Cadastre. Available online: http://xifra16.ddgi.cat/qualitat/cadastre2.asp?opCad=A&IdMenu=03031201 (accessed on 31 December 2021).
25. XIFRA. Atur Registrat. Available online: https://www.ddgi.cat/xifra/atur/aturPeriodes.asp?IdMenu=03060602&agrupat=7 (accessed on 31 December 2021).
26. XIFRA. Atur Registrat Estrangers. Available online: https://www.ddgi.cat/xifra/atur/aturEstPeriodes.asp?IdMenu=03060802&agrupat=7 (accessed on 31 December 2021).
27. XIFRA. Característiques de la Població. Available online: https://www.ddgi.cat/xifra/Indicadors/demografia/dpt_TEG.asp?IdMenu=04020303 (accessed on 31 December 2021).
28. XIFRA. Impost Sobre la Renda de Les Persones Físiques (IRPF). Available online: https://www.ddgi.cat/xifra/indicadors/ActivEcon/irpf2.asp?IdMenu=04051002 (accessed on 31 December 2021).
29. XIFRA. Moviment Natural de la Població. Available online: https://www.ddgi.cat/xifra/Indicadors/demografia/dnd_TBM.asp?IdMenu=04020402 (accessed on 31 December 2021).
30. XIFRA. Població. *Recomptes.* Available online: https://www.ddgi.cat/xifra/Indicadors/demografia/dpt_km2.asp?IdMenu=04020103 (accessed on 31 December 2021).
31. Government of Catalonia. Dades de Trucades Operatives Gestionades Pel CAT112 | Dades Obertes de Catalunya. Available online: https://analisi.transparenciacatalunya.cat/Seguretat/Dades-de-trucades-operatives-gestionades-pel-CAT11/mfqb-sbx4 (accessed on 31 December 2021).
32. Government of Catalonia. Espais Esportius I Complementaris Censats Per Municipality | Dades Obertes de Catalunya. Available online: https://analisi.transparenciacatalunya.cat/Esport/Espais-esportius-i-complementaris-censats-per-muni/v99k-i424 (accessed on 31 December 2021).
33. Government of Catalonia. Preu Mitjà del Lloguer D'habitatges Per Municipality | Dades Obertes de Catalunya. Available online: https://analisi.transparenciacatalunya.cat/Habitatge/Preu-mitj-del-lloguer-d-habitatges-per-municipality/qww9-bvhh (accessed on 31 December 2021).
34. Government of Catalonia. Superfícies Municipals Dels Conreus Herbacis A Catalunya | Dades Obertes de Catalunya. Available online: https://analisi.transparenciacatalunya.cat/Medi-Rural-Pesca/Superf-cies-municipals-dels-conreus-herbacis-a-Cat/nuvr-btxv (accessed on 31 December 2021).

35. Departament de la Vicepresidència i de Polítiques Digitals i Territori. Per Municipality. Available online: https://territori.gencat.cat/ca/06_territori_i_urbanisme/observatori_territori/litoral/regim_sol_litoral/per_municipality (accessed on 31 December 2021).
36. Departament de la Vicepresidència i de Polítiques Digitals i Territori. Territoris de Muntanya. Available online: https://territori.gencat.cat/ca/06_territori_i_urbanisme/politica_de_muntanya/territoris_de_muntanya/ (accessed on 31 December 2021).
37. INE. Atlas de Distribución de Renta de Los Hogares. Available online: https://www.ine.es/dynt3/inebase/es/index.htm?padre=7132 (accessed on 31 December 2021).
38. Bove, V.; Elia, L. Migration, diversity, and economic growth. *World Dev.* **2017**, *89*, 227–239. [CrossRef]
39. Foulkes, M.; Schafft, K.A. The Impact of migration on poverty concentrations in the United States, 1995–2000. *Rural Sociol.* **2010**, *75*, 90–110. [CrossRef]
40. Banerjee, A.; Duflo, E. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*; Public Affairs: New York, NY, USA, 2011; p. 320.
41. Lozano, M.; Rentería, E. Work in Transition: Labour Market Life Expectancy and Years Spent in Precarious Employment in Spain 1986–2016. *Soc. Indic. Res.* **2019**, *145*, 185–200. [CrossRef]
42. Anderson, E.; d'Orey, M.A.J.; Duvendack, M.; Esposito, L. Does government spending affect income poverty? A meta-regression analysis. *World Dev.* **2018**, *103*, 60–71. [CrossRef]
43. Ravallion, M. Growth, inequality and poverty: Looking beyond averages. *World Dev.* **2001**, *29*, 1803–1815. [CrossRef]
44. Son, H.H.; Kakwani, N. Global estimates of pro-poor growth. *World Dev.* **2008**, *36*, 1048–1066. [CrossRef]
45. Filmer, D.; Pritchett, L. The impact of public spending on health: Does money matter? *Soc. Sci. Med.* **1999**, *49*, 1309–1323. [CrossRef]
46. Birdsall, N. Public spending on higher education in developing countries: Too much or too little? *Econ. Educ. Rev.* **1996**, *15*, 407–419. [CrossRef]
47. Ameratunga, S.; Hijar, M.; Norton, R. Road-traffic injuries: Confronting disparities to address a global-health problem. *Lancet* **2006**, *367*, 1533–1540. [CrossRef]
48. Bloom, D.E.; Luca, D.L. *The Global Demography of Aging*; Elsevier: Amsterdam, The Netherlands, 2016; Volume 1, pp. 3–56.
49. Observatori de la Seguretat Viària. Accidents de Trànsit Amb Morts o Ferits Greus a Catalunya. Available online: http://transit.gencat.cat/ca/observatori/dades_obertes/ (accessed on 31 December 2021).
50. Agarwal, G.; Lee, J.; McLeod, B.; Mahmuda, S.; Howard, M.; Cockrell, K.; Angeles, R. Social factors in frequent callers: A description of isolation, poverty and quality of life in those calling emergency medical services frequently. *BMC Public Health* **2019**, *19*, 684. [CrossRef] [PubMed]
51. Barbaree, H.; Mewhort, D. The effects of the z-score transformation on measures of relative erectile response strength: A re-appraisal. *Behav. Res. Ther.* **1994**, *32*, 547–558. [CrossRef]
52. Ishwaran, H.; Rao, J.S. Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Stat.* **2005**, *33*, 730–773. [CrossRef]
53. Hoerl, A.E. Application of ridge analysis to regression problems. *Chem. Eng. Prog.* **1962**, *58*, 54–59.
54. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
55. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
56. Antoniadis, A.; Fan, J.; Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2001**, *67*, 301–320.
57. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
58. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
59. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–451. [CrossRef]
60. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108. [CrossRef]
61. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data*; Ltd, ch2, ch3, ch4, ch5, ch6; John Wiley & Sons: New York, NY, USA, 1990; pp. 68–279.
62. Ng, R.; Han, J. CLARANS: A method for clustering objects for spatial data mining. *Knowl. Data Eng. IEEE Trans.* **2002**, *14*, 1003–1016. [CrossRef]
63. Carnein, M.; Trautmann, H. EvoStream—Evolutionary stream clustering utilizing idle times. *Big Data Res.* **2018**, *14*, 101–111. [CrossRef]
64. Arai, K.; Barakbah, A.R. Hierarchical K-means: An algorithm for centroids initialization for K-means. *Rep. Fac. Sci. Eng.* **2007**, *36*, 25–31.
65. Kröger, P.; Kriegel, H.P.; Kailing, K. Density-connected subspace clustering for high-dimensional data. In Proceedings of the 2004 SIAM International Conference on Data Mining (SDM), Lake Buena Vista, FL, USA, 22–24 April 2004; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2004; pp. 246–257.
66. Hahsler, M.; Piekenbrock, M.; Doran, D. Dbscan: Fast density-based clustering with R. *J. Stat. Softw.* **2019**, *91*, 1–30. [CrossRef]

67. Fichtenberger, H.; Gillé, M.; Schmidt, M.; Schwiegelshohn, C.; Sohler, C. BICO: BIRCH meets coresets for k-means clustering. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8125, pp. 481–492.
68. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: A new data clustering algorithm and its applications. *Data Min. Knowl. Discov.* **1997**, *1*, 141–182. [CrossRef]
69. ICGC. Base Municipal. Available online: https://www.icgc.cat/Administracio-i-empresa/Descarregues/Capes-de-geoinformacio/Base-municipal (accessed on 31 December 2021).
70. Allen, D.M. Mean square error of prediction as a criterion for selecting variables. *Technometrics* **1971**, *13*, 469. [CrossRef]
71. Celeux, G.; Soromenho, G. An entropy criterion for assessing the number of clusters in a mixture model. *J. Classif.* **1996**, *13*, 195–212. [CrossRef]
72. Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.—Simul. Comput.* **1974**, *3*, 1–27. [CrossRef]
73. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*; John Wiley & Sons: New York, NY, USA, 1973.
74. Neuhäuser, M. Wilcoxon—Mann—Whitney test. In *International Encyclopedia of Statistical Science*; Miodrag, L., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 1656–1658.
75. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S-PLUS*; Springer: New York, NY, USA, 2003.

4.2 Article II

Housing Supply and How It Is Related to Social Inequalities—Air Pollution, Green Spaces, Crime Levels, and Poor Areas—In Catalonia

Perafita X, Saez M

International Journal of Environmental Research and Public Health: 2023 April 19. doi:10.3390/IJERPH19063359

Caixa 5. Síntesis del article II

Context de l'article

- Una família que gasti més del 40% dels ingressos en pagar l'habitatge està en situació d'entrar en la trampa de la pobresa.
- A Europa el 2018, el 10,20% de llars en risc de pobresa superava aquest llindar.
- La llar actua com un element clau en el desenvolupament vital dels membres d'una unitat família.

Què aporta l'article?

- L'estudi mostra com el preu de les propietats a Catalunya no afavoreix que les famílies en situació de vulnerabilitat puguin sortir de la trampa de la pobresa. En més de 12.000 llars només 9, tenen un cost on una família per sota el llindar de la pobresa podria pagar un lloguer inferior al 40% de la seva renda.
- Es mostren patrons diferents en l'Àrea Metropolitana de Barcelona (AMB) i fora d'aquesta. Dins l'AMB l'augment del NDVI augmenta la probabilitat de no llogar una llar en un 125%. En canvi, fora de l'AMB disminueix un 78,21%. Dins l'AMB, per cada increment en 1 $\mu g/m^3$ de PM10 la probabilitat de no poder llogar una llar disminueix en 7,03%. Fora l'AMB la probabilitat augmenta en 8,08%.
- Existeix una pobresa latent a tot el territori català que constitueix una pobresa estructural.

International Journal of
Environmental Research and Public Health

MDPI

Article

Housing Supply and How It Is Related to Social Inequalities—Air Pollution, Green Spaces, Crime Levels, and Poor Areas—In Catalonia

Xavier Perafita [1,2] and Marc Saez [2,3,*]

[1] Observatori—Organisme Autònom de Salut Pública de la Diputació de Girona (Dipsalut), 17003 Girona, Spain
[2] Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, 17003 Girona, Spain
[3] CIBER of Epidemiology and Public Health (CIBERESP), 28029 Madrid, Spain
* Correspondence: marc.saez@udg.edu

Abstract: We carried out a search of over 12,000 houses offered on the rental market in Catalonia and assessed the possibility of families below the poverty threshold being able to rent these homes. In this regard, we wanted to evaluate whether the economic situation of families is able to influence their social environment, surroundings, and safety. We observed how their economic situation can allow families the possibility of developing a life without exposure to health risks, and how economic constraints result in disadvantages in several areas of life. The results show how families at risk of poverty live in less favourable conditions and experience a widening of different gaps, with current prices leading to a possible poverty trap for the most disadvantaged groups. The higher the percentage of the population below the threshold, the lower the possibility of not being able to rent a house compared to areas with a lower prevalence of population below the threshold. This association was observed both when considering the risk linearly and non-linearly. Linearly, the probability of not renting a house was reduced by 8.36% for each 1% increase in the prevalence of population at risk of extreme poverty. In the second, third and fourth percentage quartiles, the probability of not being able to rent a house decreased by 21.13%, 48.61%, and 57.79%, respectively. In addition, the effect was different inside and outside of metropolitan areas, with the former showing a decrease of 19.05% in the probability of renting a house, whereas outside metropolitan areas the probability increased by 5.70%.

Keywords: inequality; housing poverty; NDVI; pollutants; poverty trap; life limitations

Citation: Perafita, X.; Saez, M. Housing Supply and How It Is Related to Social Inequalities—Air Pollution, Green Spaces, Crime Levels, and Poor Areas—In Catalonia. *Int. J. Environ. Res. Public Health* **2023**, *20*, 5578. https://doi.org/10.3390/ijerph20085578

Academic Editors: Joanna Mazur and Paul B. Tchounwou

Received: 1 November 2022
Revised: 4 April 2023
Accepted: 11 April 2023
Published: 19 April 2023

1. Introduction

Social inequality can be viewed from different perspectives. These can be political, economic, social, or health-related, among others, and they can be interrelated [1–5]. Consequently, inequality culminates in the generation of situations of marginalisation and social exclusion. This problem is highlighted by the policies pursued by governments and institutions to try to minimalize these inequalities.

If we focus on income inequalities, there are several policies pursued in Europe to minimise the gap including minimum wages, taxes for redistributive purposes, and the establishment of welfare that guarantees minimum services for the whole population. As Piketty shows, the growth of economies alone does not reduce inequalities stemming from income and wealth [6]. These inequalities originate from the poor distribution of wealth among the population. However, the discourse around this line of thought does not revolve around an element that is common in the lives of everyone: housing.

Housing is one of the inequalities generated within the economic sphere and encompasses under-housing, housing deficiencies, unhealthy households, sub-renting, employment, unhealthy environments, and the cost of housing; all of which are elements that lead to social exclusion and other inequalities. Housing was declared a universal right in 1948

in the Universal Declaration of Human Rights (article 25), and later in the International Covenant on Economic, Social and Cultural Rights. In fact, it is one of the economic, social, and cultural rights with the greatest impact. Irrespective of the debate as to whether housing is a right or an obligation of governments [7], the fact is that there is a part of the population in situations of vulnerability with respect to homes [8].

Not all homes are optimal for developing a dignified and healthy lifestyle [9,10]. While most countries have their own regulations that determine what these minimums should be, there is still a significant gap between what is understood as a decent home [10] and each country's domestic market prices.

This difference in prices and housing features causes a mismatch between the possibilities of accessing a decent home for different social classes. Although homes are one of the few financial assets most likely to be distributed among the middle classes, not all families can access one [11]. Those who can access a home must be alert to what percentage of their income is allocated to paying for it. If this percentage is very high, the pressure on the household may increase, having a negative effect on the social and economic life and health of its members. This is the reason why rent subsidies are provided in various countries [12,13]. The fact that house prices are not within the means of all families means that those with fewer financial resources cannot buy a home and must rent a property instead [14]. It is recommended that, whether for purchase or rental, the percentage of household income allocated to a home does not exceed 40% of the total income. If this percentage is not adhered to, one can more easily fall into the poverty trap. In 2015, 11.3% of the population living in the European Union lived in households that spent more than 40% of their family income on a home [15]. In 2018, 10.2% of households [16] at risk of poverty spent more than this threshold figure. Although this ceiling is generally 40%, EU countries adopt different thresholds ranging from 30% to 40% [16].

These features are also directly related to price. The type of housing and installations and the greater the number of square metres, toilets, and rooms it has, the higher its price will be. The area where the home is located must also be considered. If it is in an unsafe neighbourhood with a higher crime rate or there are fewer facilities nearby, the price drops significantly [17–19].

Another factor to consider is whether or not the area surrounding the home has green zones, including parks, forests, gardens, and trees, which are all linked to better health [20–24] and the acquisition of healthy habits [25]. Proximity to green spaces also affects the final price of a house, and so the further a house is from a green area, the lower its price. That said, the size of the green areas within the vicinity of the house does not seem to affect its final price [26]. Likewise, living in an area with high-density traffic or high levels of pollution will also have a direct impact on people's quality of life [27]. This is also reflected in the price of a home whereby the greater the levels of pollution are (whether due to noise or air quality), the lower the price of the home will be [28–30].

Therefore, a low home price could be linked to a poorer quality home. This trend can become a cyclical problem, where lower income families are not able to access a large variety of homes. In addition, the properties to which they will have access will be those with a worse set of features, implying that they cannot access quality neighbourhoods in safer areas or those which provide a better quality of life.

Consequently, we can interpret the home as a key element in the development of people, right from childhood itself, and a key factor in the growth and reduction of inequalities in the medium and long terms.

In Spain, economic growth has been closely linked to the construction of private homes [31]. This growth has transformed households, considered a first-necessity good, into a speculative good. This speculation has resulted in a decrease in the number of households that can be accessed by low-income families.

On a demographic level, Catalonia has a unique structure in that more than 21% of its population lives in Barcelona; the second most populous city in Spain. Furthermore, more than 41% of its population lives either in Barcelona or one of its 36 adjacent towns,

meaning that 41% of the population occupies just 1.97% of the region. By the same token, 66% reside in cities with more than 20,000 inhabitants which, in turn, is equivalent to just 6.61% of the total territory [32].

This rural/urban dichotomy has direct implications for inequality, and in those that affect living conditions above all. Urbanization has been present in many areas of the Catalan territory at the same time as inequalities have been accentuated and residential segregation manifested [33,34]. Recently, house prices have generally increased, especially in the metropolises [35–37]. In 2020, rental prices had increased so dramatically in Catalonia that a law was passed to regulate them. This law allows rents to be capped based on the Catalan Housing Agency's Average Price Index, which sets an average reference price for specific areas where abusive price increases have been identified. In addition, with the appearance of homes destined for short-term rental platforms such as Airbnb, the number of homes available for rent has been reduced. This reduction is even greater in popular tourist areas and has resulted in a 7% increase in rental prices [38,39].

These circumstances end up affecting those with fewer economic resources as they must allocate a larger part of their income to cover housing, often at the expense of living in a healthy environment [40]. This paper aims to study the conditions under which families who are below the poverty threshold can rent a home. Our hypothesis is that the socially excluded cannot rent a property with conditions that ensure an optimal standard of living. We also assume that to have a home to live in, low-income families must interact with environments of high social inequality.

Our objective in this work is to study the ability of people at risk of social exclusion to access the rental housing market in Catalonia at the small scale (either municipality or district). Specifically, we evaluated the association between the possibility that a family at risk of exclusion does not rent a home and the type of socio-economic environment, the nearby urban greenery (green areas per inhabitant in square meters, the Normalized Difference Vegetation Index (NDVI) in a range of 500 m), the level of exposure to air pollution (PM_{10}, NO_2 and O_3), and the level of known crimes in the area.

2. Materials and Methods

2.1. Area and Period of Study

We used a cross-sectional observational design, and the data were extracted from multiple databases in September 2021 to study accessibility. The area of study consisted of the geolocated information of each home combined with the information from the small areas (either municipality or district), to enable access to the most granular information possible. The study is based on data from households for rent in September 2021.

In Catalonia, as in the rest of Spain, only municipalities with 75,000 inhabitants or more are divided into districts. The districts, although they are an administrative division, group neighbourhoods with noticeable homogeneity in terms of socioeconomic variables (similar population density, similar disposable income, similar inequality index, etc.), see Figures 1–4. In addition, the districts are the smallest administrative unit for which information is obtained for the variables of interest in this study (i.e., housing).

Int. J. Environ. Res. Public Health **2023**, *20*, 5578

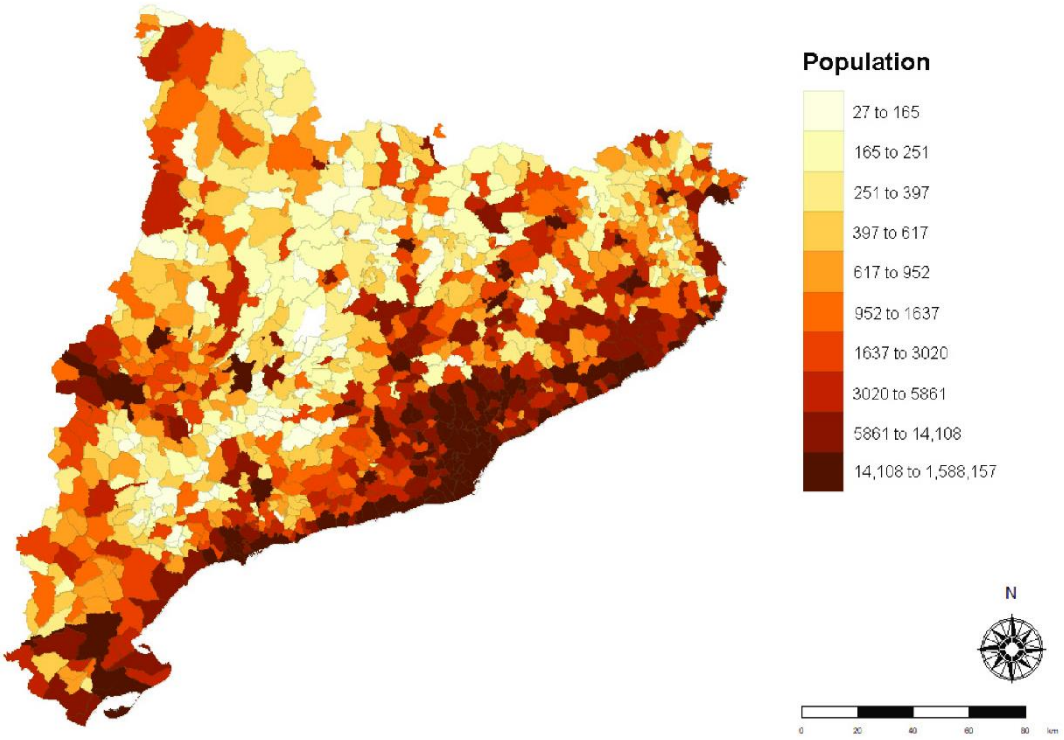


Figure 1. Map of the population of Catalonia by municipalities.

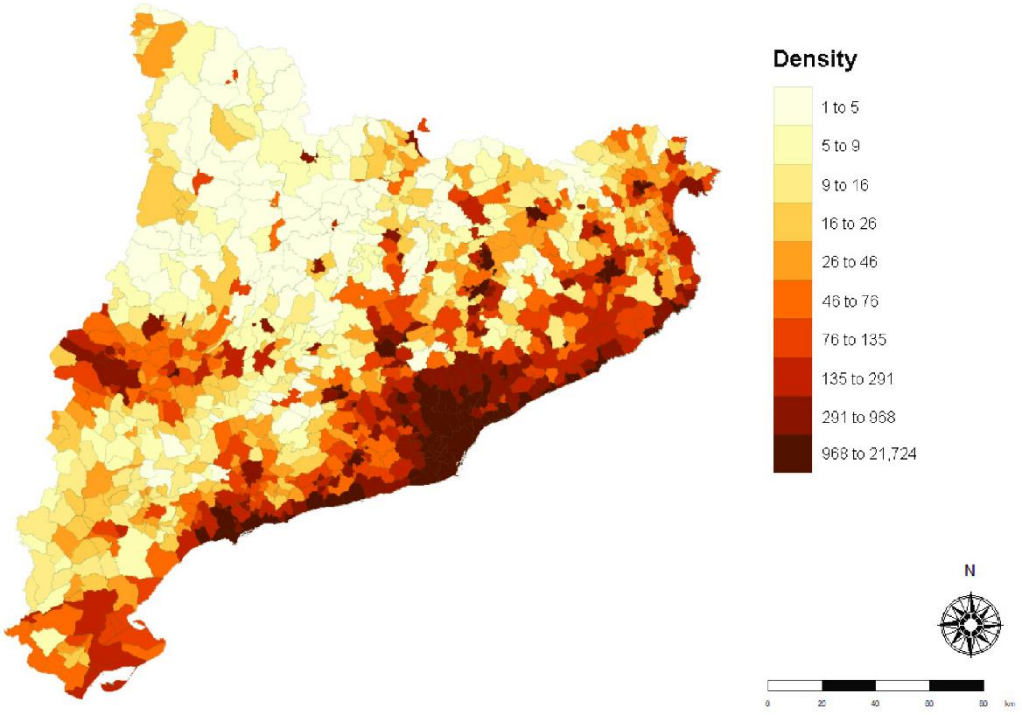


Figure 2. Map of the density of Catalonia by municipalities.

Int. J. Environ. Res. Public Health **2023**, *20*, 5578

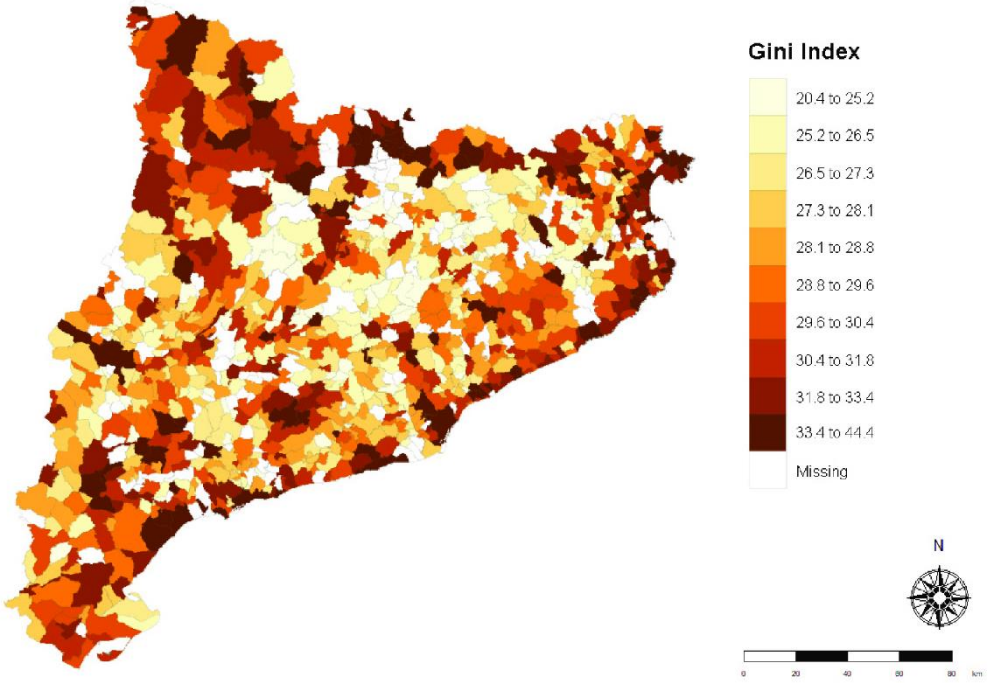


Figure 3. Map of the Gini index of Catalonia by municipalities.

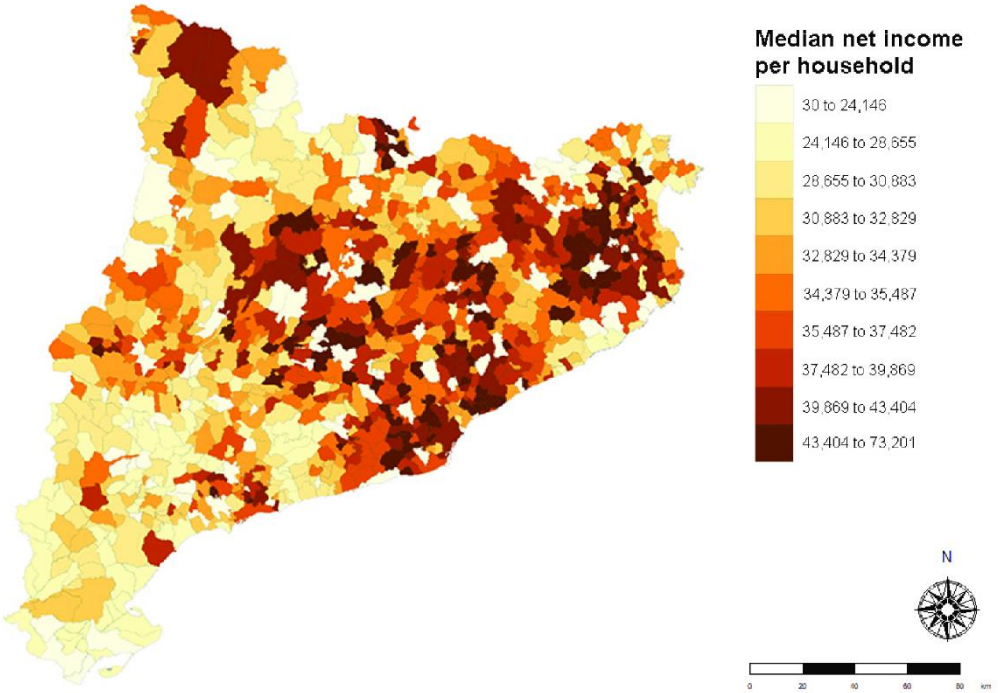


Figure 4. Map of the median net income per household of Catalonia by municipalities.

2.2. Methods Prior to Executing the Study, the Dataset, and Data Sources

2.2.1. Data Sources

The data used were extracted from different sources, all of them official, and translated into a dataset of more than 50 variables (see Supplementary Table S1). The sources used to carry out the study were: Statistical Institute of Catalonia (IDESCAT), Cartographic and Geological Institute of Catalonia (ICGC), Government of Catalonia Open Data, the Mossos d'Esquadra police force, the National Statistics Institute (INE), Habitaclia estate agency and the Research Group on Statistics, Econometrics and Health (GRECS), University of Girona.

2.2.2. Homes

A web scraping of the Habitaclia website was carried out to determine the number of rental homes available in Catalonia. After reviewing the main rental platforms in Spain the Habitaclia portal was chosen because it is not only one of the platforms with the largest number of homes registered on it, but its web structure also allowed greater ease of data extraction for subsequent processing and geolocation.

Web Scraping

Web scraping (a technique using software to extract public information from websites in an automated way) allowed the following features of each home to be obtained: price, m^2, number of toilets and rooms, province, municipality, district, and street (where the house is located). The total number of rental homes in Catalonia was 12,796, comprising 11,320 homes in the province of Barcelona (mean: 58.35 homes, standard deviation: 566.23 homes, median: 4 homes, first quartile -Q1-: 2 homes, and third quartile -Q3-: 15.75 homes); 582 in the province of Girona (mean: 4.89 homes, standard deviation: 10.48 homes, median: 2 homes, first quartile -Q1-: 1 home, and third quartile -Q3-: 4.5 homes); 234 in the province of Lleida (mean: 4.03 homes, standard deviation: 17.82 homes, median: 1 home, first quartile -Q1-: 1 home, and third quartile -Q3-: 1 home); and 660 in the province of Tarragona (mean: 10.82 homes, standard deviation: 28.37 homes, median: 3 homes, first quartile -Q1-: 1 home, and third quartile -Q3-: 3 homes). Figures 5 and 6 shows the result for geolocalised homes in the area of study.

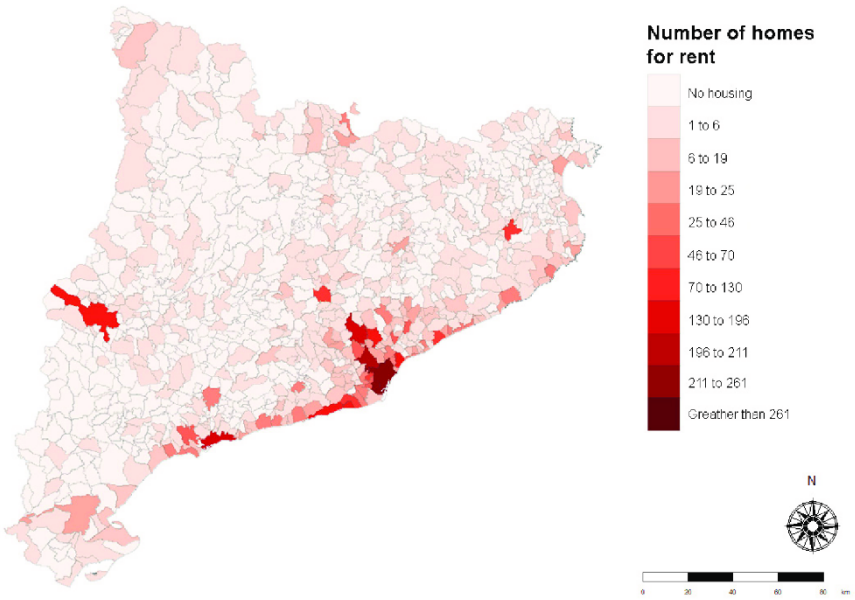


Figure 5. Map of number of homes for rent.

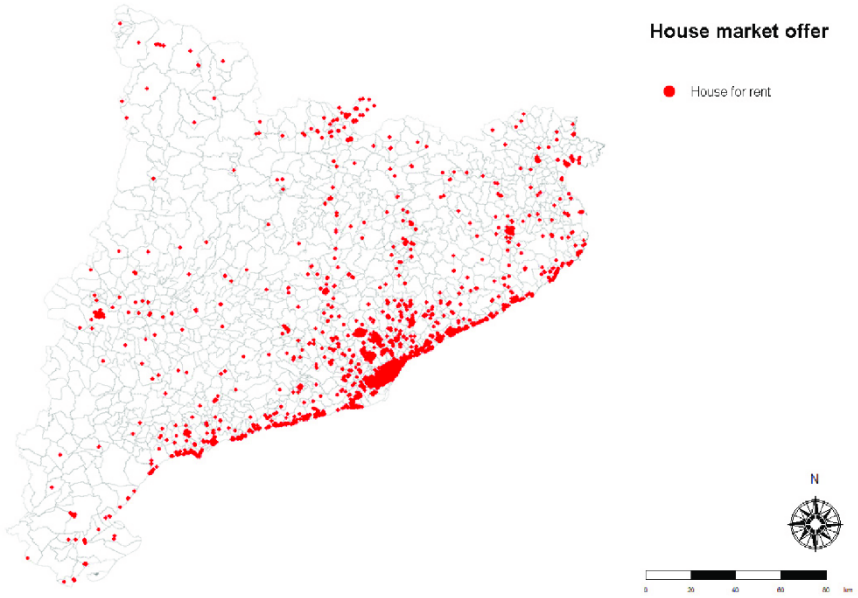


Figure 6. Map of real estate rental offer (each point is a house).

Geolocalisation, Atypical Values, Errors, and Debugging the Web Scraping

The data obtained from the web scraping were used to localise each home in its town and district of reference. Of the total number of homes obtained (n = 12,796), n = 11,289 were successfully geolocated. This debugging reported the following errors: platform errors (n = 3); homes with some missing information (n = 42); and homes where the geolocated municipality and province do not match (n = 142). All homes that were atypical due to their high rental price and could affect the objective of the study were also discarded (n = 1320).

2.2.3. Variables

Dependent Variables

We considered, as a dependent variable, a variable indicating that, in September 2021, the home was not rented (value 1 and 0 if it was rented).

Explanatory Variables

— *Environmental and demographic data*

The cultural diversity [41] of the people living in a neighbourhood or area determines the socioeconomic level of its population. This socioeconomic level directly affects the area or neighbourhood, generating either favourable economic growth or reversing this growth [42].

The data were obtained taking the district in each town where the home was located as the reference. If they were subject to statistical secrecy, the value for the town itself was used. The variables used to determine the socioeconomic fabric of the territory were: average age, average size of homes, number of single-person households, and the Gini index [43].

— *Vegetation and the presence of greenery*

The relationship between the presence of parks, gardens, trees, and other green elements has been scientifically proven [44–46] to be linked to health [47]. The NDVI [48] and the surface area of green areas per inhabitant in square metres [49] were used to study the relationship between the homes and the surrounding greenery. The WHO

recommendation regarding the latter is that every city should have at least 9 m^2 of greenery per inhabitant [21].

In addition, the relationship between the presence of urban vegetation and the quality of life and health is known. There is also a relationship between urban vegetation and physical activity and mental health [50].

To observe the nearby NDVI of each house, a 500-metre buffer was constructed [51], where each house acts as a centroid of its own buffer. In this way, the vegetation close to each house can be analysed.

– *Air pollutants*

In recent years, there has been an ever expanding amount of scientific evidence concerning increased levels of pollution [52–54]. Consequently, pollution is highly relevant when considering the development of a healthy lifestyle.

The following variables regarding air pollution [55,56] were used: particulate matter equal to or less than 10 microns (PM_{10}), nitrogen dioxide (NO_2), and ozone (O_3). In 2021, the WHO recommended new levels of contamination [57]. Families living in areas below these pollution levels will see their cardiovascular and respiratory health improve, in addition to reducing their burden of morbidity and mortality. We created two dummy variables for each pollutant and looked at whether homes were in locations where air pollution was above or below the old and current WHO limits.

– *Safety*

The safety of a town or area is an important element in determining its socio-economic status [58]. Low-income areas are directly related to different types of crime [59]. Although Europe in general does not have such a high or significant ratio as other countries [60], it is still an element to consider, especially for families who rent [18]. The number of crimes is also related to the movement of drugs, gangs, and other elements that can impact an area and condition the lives of the families living there [61].

The total number of crimes known to the Catalan police force (Mossos d'Esquadra) in the Basic Police Area (ABP) of each home was used [62]. The Mossos d'Esquadra are responsible for citizen safety, public order, investigation, and traffic control in Catalonia.

– *Ability to pay*

The rental price of homes determines the type of families who will be able to live in an area [63]. Each small area (either municipality or district) has similar prices for a home, which vary depending on the features, the socioeconomic level of the neighbourhood, the presence of nearby urban greenery, and how safe the area is. Migratory movements also affect the socioeconomic levels of the population [64].

The threshold for Catalonia (EUR 11,365.60) [65] was used to determine which flats could be rented by families at risk of poverty. The social structure of the territory was also obtained through the variables presented above. We have controlled the neighbouring areas of each house to observe their closest socioeconomic environments. We do so to observe the interaction of each dwelling with the areas close to them.

The variables linked to income to determine the ability to pay were net income per person per household, median per unit of consumption, and average salary [43]. The percentages of people at risk of poverty (60% of the median) and extreme poverty (40% of the median) were also obtained.

Control Variables

We controlled for different types of variables: features of the home, social characteristics, population size, and typologies of areas. More specifically, the area in square meters, and the number of rooms and toilets were used to consider the different features. In terms of population characteristics, we used the number of inhabitants, the average size and age of the home, and the population density in the district where each home was located.

The differences presented in the small areas (either municipality or district) forced the creation of multiple control variables to capture the heterogeneity of the territory: size of each district in square kilometres (source: extraction of the area through the municipal raster [66]); the variable density_urban, which measures the density of the urban areas of each town (source: data from the urban map of Catalonia, [49]); the variable area_metro, which captures whether or not the district in question belong to the metropolitan area of Barcelona (source: authors' own); the variable capitals (source: authors' own) to capture whether or not the district belongs to a provincial capital; and the variable density_ (source: authors' own), which measures whether the district is in a high- (High) or low- (Low) density area, depending on the reference population. Apart from the variable density_urban, the variables were observed at the district level.

2.2.4. Geospatial Groupings

The data were placed into four groups. The first set of data was georeferenced and applied to the characteristics derived from the dwellings as well as the NDVI and the dummy control variables. The second set was grouped by district or town depending on the availability of the information. The data in this second group correspond to the socioeconomic data and the dichotomous variables of rural/urban areas and belonging to the metropolitan area of Barcelona. The inhabitants by districts are distributed as follows: mean: 6167 inhabitants, standard deviation: 16,556.16 inhabitants, median: 1730 inhabitants, first quartile -Q1-: 433 inhabitants, and third quartile -Q3-: 6251 inhabitants. By municipality they are distributed as follows: mean: 16,378 inhabitants, standard deviation: 258,190.7 inhabitants, median: 961 inhabitants, first quartile -Q1-: 322 inhabitants, and third quartile -Q3-: 3903 inhabitants.

The third type of grouping included the Basic Health Areas (ABS), which are defined as the territorial units through which the primary health care services are organized [67]. These can have different dimensions depending on the accessibility of the population to be served. The variables in this third grouping consist of the air pollution data. The population distribution of the ABS is: mean: 20,636 inhabitants, standard deviation: 9192.051 inhabitants, median: 20,638 inhabitants, first quartile -Q1-: 14,018 inhabitants, and third quartile -Q3-: 26,558 inhabitants.

The last grouping includes the Basic Police Areas (ABP), which are defined as the basic territorial implementation units, themselves defined using territorial and police criteria [64]. The data in this last group are those related to the number of crimes. The distribution of people of in the ABP is: mean: 381,153 inhabitants, standard deviation: 577,201.9 inhabitants, median: 129,374 inhabitants, first quartile -Q1-: 65,844 inhabitants and third quartile -Q3-: 258,179 inhabitants.

2.3. Data Analysis

We specified generalised linear models (GLM) with variable response of a Bernoulli distribution, according to the possibility of renting a home, i.e., a variable indicating that, in September 2021, a specific home (of the 11.289 homes analysed) was not rented.

We included as the following explanatory variables in the GLM: the percentage of the population at risk of social exclusion in the district of each home; the vegetation index (also in the district); the number of m^2 of green area per inhabitant in the town of each home; dichotomous variables indicating if the values of the air pollutants (PM_{10}, NO_2 and O_3) predicted in the district exceed the limits defined by the WHO; and the number of known crimes in the district of each home. We also controlled for the control variables defined above.

For the effect of the metropolitan area on the possibility of a family at risk of social inclusion renting a home, the model was re-estimated, including the interaction of the metropolitan area.

Details are shown in the Supplementary Material.

The values of the coefficients and the Odds Ratios generated through the model were used to study which variables affect the event of households at risk of poverty renting a home.

We included the variables linearly (that is, the response of the dependent variable is the same to an increase in the variable, regardless of the level of the variable) and non-linearly (that is, the response of the dependent variable to an increase in the variable will depend on the level of the level of the variable) in the model, testing different reference categories to see how they interacted with the event of households at risk of poverty renting a home.

The models were compared using the Akaike Information Criterion (AIC), which reviews and penalises the flexibility of the model [68] and the control of multicollinearity.

2.4. Software

The processes of web scraping and geolocalisation were carried out using Spyder software (version 5.1.5) [69] based on the Python language. The libraries used were random, time [70], requests [71], IPython [72], bs4 [73], datetime [74], csv [75], pandas [76], os [77], re [78], math [79], msvcrt [80], tabulate [81], tkinter [82], tqdm [83], and googlemaps [84].

Obtaining and debugging the data and the statistical modelling were carried out with the software RStudio (version 4.1.3) [85], based on the R language. The libraries used were rgdal [86], rgeos [87], raster [88], tmap [89], BBmisc [90], haven [91], dplyr [92], and ggplot2 [93].

3. Results

3.1. Bivariate Analyses

The results are presented in Tables 1 and 2 (bivariate analysis). Table 1 considers whether the home could be rented by a household at risk of social exclusion. Table 2 also considers the area where the home is located, either within the Metropolitan Area (AMB) or Outside the AMB (OAMB). The variables show quite asymmetric distributions, implying that only robust statistics (median and first and third quartile) should be used when interpreting the data.

Table 1. Bivariate analysis according to the possibility of renting.

	Housing to Rent	
Variables	**Can Rent (n = 5174)**	**Cannot Rent (n = 5923)**
Percentage of people with an equivalized disposable income below the risk of extreme poverty threshold (%)		
Mean (sd)	8.967 (3.979)	8.465 (3.941)
Median (Q1–Q3)	7.7 (6.5–10.6)	7.7 (5.6–8.2)
Min–Max	2–34.3	2–24.6
Percentage of people with an equivalized disposable income below the risk of poverty threshold (%)		
Mean (sd)	17.246 (7.096)	15.32 (7.116)
Median (Q1–Q3)	14.7 (12.3–20.9)	13.3 (9.5–15.5)
Min–Max	4.8–57.1	4.8–42.2
Gini Index (%)		
Mean (sd)	33.251 (3.905)	35.752 (3.759)
Median (Q1–Q3)	32.9 (30.5–36.1)	36.1 (33.1–38.3)
Min–Max	22.6–43.9	23.5–43.9
PM_{10} ($\mu m/m^3$)		
Mean (sd)	23.476 (2.945)	23.989 (2.815)
Median (Q1–Q3)	23.567 (21.664–25.444)	23.893 (22.126–26.201)
Min–Max	12.172–34.057	12.172–34.057

Table 1. *Cont.*

Housing to Rent		
Variables	**Can Rent (n = 5174)**	**Cannot Rent (n = 5923)**
NO_2 ($\mu m/m^3$)		
Mean (sd)	27.416 (7.546)	27.969 (7.517)
Median (Q1–Q3)	26.358 (23.053–30.19)	27.922 (23.186–30.19)
Min–Max	4.994–52.524	4.994–52.48
O_3 ($\mu m/m^3$)		
Mean (sd)	52.385 (8.096)	51.604 (7.542)
Median (Q1–Q3)	51.111 (46.078–57.988)	49.978 (45.469–57.929)
Min–Max	22.592–80.507	22.592–75.387
Vegetation vigour (NDVI) within 500 metres of the house (index between −1 and 1)		
Mean (sd)	0.252 (0.098)	0.235 (0.079)
Median (Q1–Q3)	0.226 (0.189–0.304)	0.221 (0.192–0.261)
Min–Max	−0.028–0.69	−0.019–0.69
Green areas per inhabitant (m^2)		
Mean (sd)	14.738 (17.768)	11.913 (14.809)
Median (Q1–Q3)	7.847 (7.847–15.752)	7.847 (7.847–7.847)
Min–Max	1.955–346.709	0–273.822
Number of known crimes		
Mean (sd)	11,952,441 (6986.762)	15,356.411 (8702.922)
Median (Q1–Q3)	9780 (6498–15,510)	14,303 (8056–22,791)
Min–Max	341–27,890	467–27,890
Gross household income (EUR)		
Mean (sd)	46,423.656 (16,329.446)	57,562.09 (22,441.094)
Median (Q1–Q3)	41,990 (36,195.75–50,566)	54,708 (41,990–60,502)
Min–Max	20,489–105,681	26,291–132,268
Individual gross income (EUR)		
Mean (sd)	18,491.176 (6438.809)	23,198.498 (8409.087)
Median (Q1–Q3)	17,004 (13,966–20,606)	22,484 (17,004–24,532)
Min–Max	7218–39,760	8853–44,928
Can money be saved? (n—%-)		
Can save money	9—0.15%-	0—0%-
Cannot save money	5165—99.85%-	5923—100%-
Is it above the former WHO PM_{10} limit? (n—%-)		
Under limit WHO	617—11.05%-	338—6.05%-
Over limit WHO	4557—88.95%-	5585—93.95%-
Is it above the new WHO PM_{10} limit? (n—%-)		
Under limit WHO	4—0.07%-	1—0.02%-
Over limit WHO	5170—99.93%-	5922—99.98%-
Is it above the former WHO NO_2 limit? (n—%-)		
Under limit WHO	4845—93.64%-	5512—93.06%-
Over limit WHO	329—6.36%-	411—6.94%-
Is it above the new WHO NO_2 limit? (n—%-)		
Under limit WHO	55—1.06%-	10—0.17%-
Over limit WHO	5119—98.94%-	5913—99.83%-

Table 2. Bivariate analysis according to the possibility of renting inside or outside the metropolitan area.

Variables	Housing to Rent			
	Outside the Metropolitan Area (n = 2912)		Metropolitan Area (n = 8185)	
	Can Rent (n = 2182)	Cannot Rent (n = 730)	Can Rent (n = 2992)	Cannot Rent (n = 5193)
Percentage of people with an equivalised disposable income below the risk of extreme poverty threshold (%)				
Mean (sd)	8.733 (3.654)	8.206 (3.552)	9.138 (4.193)	8.502 (3.991)
Median (Q1–Q3)	7.9 (5.825–11)	7.6 (5.5–10.3)	7.7 (6.6–9.4)	7.7 (5.6–7.8)
Min–Max	2–34.3	2–20.4	2.9–19.4	2.2–24.6
Percentage of people with an equivalised disposable income below the risk of poverty threshold (%)				
Mean (sd)	17.513 (6.491)	15.983 (6.129)	17.05 (7.501)	15.227 (7.24)
Median (Q1–Q3)	16.3 (12.3–21.6)	15.4 (11.3–19.9)	14.6 (12.3–17.7)	13.3 (9.5–14.7)
Min–Max	4.8–57.1	4.8–34.3	6.6–35.9	5.6–42.2
Gini Index (%)				
Mean (sd)	31.705 (3.39)	33.142 (4.459)	34.379 (3.87)	36.119 (3.497)
Median (Q1–Q3)	31.5 (29.4–33.7)	32.45 (30.4–35.5)	34.1 (31.7–38.3)	36.1 (33.1–38.3)
Min–Max	22.6–43.9	23.5–43.9	24.4–41.5	25.5–41.5
PM_{10} ($\mu m/m^3$)				
Mean (sd)	23.007 (2.987)	23.207 (2.744)	23.818 (2.866)	24.099 (2.808)
Median (Q1–Q3)	23.081 (21.285–24.769)	22.914 (22.007–24.734)	24.091 (22.126–25.897)	23.893 (22.126–26.446)
Min–Max	12.172–34.057	12.172–34.057	15.16–30.262	15.16–30.262
NO_2 ($\mu m/m^3$)				
Mean (sd)	27.17 (8.05)	26.346 (8.663)	27.595 (7.152)	28.197 (7.314)
Median (Q1–Q3)	26.397 (23.004–30.142)	26.45 (20.37–30.378)	26.35 (23.363–30.19)	27.922 (23.469–30.19)
Min–Max	4.994–52.524	4.994–52.48	13.52–52.419	13.52–52.419
O_3 ($\mu m/m^3$)				
Mean (sd)	54.702 (8.598)	55.679 (9.466)	50.694 (7.259)	51.032 (7.045)
Median (Q1–Q3)	55.031 (49.298–61.035)	56.67 (49.949–61.215)	49.743 (45.469–55.981)	49.966 (44.8–56.394)
Min–Max	22.592–80.507	22.592–75.387	37.651–71.425	37.651–71.425
Vegetation vigour (NDVI) within 500 metres of the house (index between −1 and 1)				
Mean (sd)	0.294 (0.119)	0.308 (0.131)	0.222 (0.064)	0.225 (0.062)
Median (Q1–Q3)	0.285 (0.207–0.367)	0.312 (0.207–0.385)	0.213 (0.179–0.245)	0.218 (0.191–0.242)
Min–Max	(−0.028)–0.69	0–0.69	0.003–0.578	(−0.019)–0.609
Green areas per inhabitant (m^2)				
Mean (sd)	22.124 (22.272)	29.638 (30.838)	9.352 (10.747)	9.422 (8.138)
Median (Q1–Q3)	16.223 (12.041–25.786)	20.774 (12.041–31.746)	7.847 (7.847–7.847)	7.847 (7.847–7.847)
Min–Max	1.955–346.709	0–273.822	3.953–192.599	3.953–192.599
Number of known crimes				
Mean (sd)	8857.759 (4311.136)	8543.019 (3690.514)	14,209.326 (7667.583)	16,314.196 (8776.855)
Median (Q1–Q3)	9121 (5000–12,940)	8727 (5321–12,163)	14,303 (8056–22,791)	15,074 (8056–27,890)
Min–Max	341–15,535	467–15,535	3522–27,890	3522–27,890
Gross household income (EUR)				
Mean (sd)	42,140.024 (8797.947)	47,905.192 (11,046.034)	49,547.614 (19,534.271)	58,919.597 (23,287.825)
Median (Q1–Q3)	40,907 (36,194–47,221)	47,075 (39,387–54,708)	42,731 (36,574–54,962)	54,962 (41,990–71,287)
Min–Max	20,489–81,141	27,325–81,141	28,500–105,681	26,291–132,268

Table 2. *Cont.*

Variables	Housing to Rent			
	Outside the Metropolitan Area (n = 2912)		Metropolitan Area (n = 8185)	
	Can Rent (n = 2182)	Cannot Rent (n = 730)	Can Rent (n = 2992)	Cannot Rent (n = 5193)
Individual gross income (EUR)				
Mean (sd)	16,252.919 (3319.883)	18,292.252 (4042.72)	20,123.489 (7572.718)	23,888.187 (8631.341)
Median (Q1–Q3)	15,900 (13,966–17,644)	17,490 (15,233–21,415)	17,659 (13,812–23,771)	23,771 (17,004–24,950)
Min–Max	7218–28,575	10,731–28,575	8987–39,760	8853–44,928
Can money be saved? (n—%-)				
Can save money	9—0.41%-	0—0.0%-	0—0.0%-	0—0.0%-
Cannot save money	2173—99.59%-	730—100%-	2992—100%-	5193—100%-
Is it above the former WHO PM_{10} limit? (n—%-)				
Over limit WHO	1859—85.20%-	663—90.82%-	2698—90.17%-	4922—94.78%-
Under limit WHO	323—14.80%-	67—9.18%-	294—9.83%-	271—5.22%-
Is it above the new WHO PM_{10} limit? (n—%-)				
Over limit WHO	2178—99.82%-	729—99.86%-	2992—100%-	5193—100%-
Under limit WHO	4—0.18%-	1—0.12%-	0—0.0%-	0—0.0%-
Is it above the old WHO NO_2 limit? (n—%-)				
Over limit WHO	179—8.20%-	50—6.85%-	150—5.01%-	361—6.95%-
Under limit WHO	2003—91.80%-	680—93.15%-	2842—94.99%-	4832—93.05%-
Is it above the new WHO NO_2 limit? (n—%-)				
Over limit WHO	2127—97.48%-	720—98.63%-	2992—100%-	5193—100%-
Under limit WHO	55—2.52%-	10—1.37%-	0—0.0%-	0—0.0%-

The results of the bivariate analysis (Table 1) show that generally the properties that cannot be rented by families at risk of social exclusion tend to be found in areas with a high Gini index (Can rent: 32.9, Cannot rent: 36.1). It is also observed that the percentage of people at risk of social exclusion is lower (Can rent: 14.7, Cannot rent: 13.3). However, it should be noted that the percentage of families living in extreme conditions does not vary depending on whether the houses can or cannot be rented by households with adverse social conditions (Can rent: 7.7, Cannot rent: 7.7). Table 2 shows how, in both areas, the homes that cannot be rented by families at risk of poverty are in the areas with a higher Gini score (AMB—Can rent: 31.5; Cannot rent: 32.45. OAMB—Can rent: 34.1; Cannot rent: 36.1). It can also be seen how these homes are found to be in areas where there is a lower percentage of people at risk of poverty (AMB—Can rent: 16.3; Cannot rent: 15.4. OAMB—Can rent: 14.6; Cannot rent: 13.3). However, it can also be seen that the percentage of people below the poverty line is higher outside the metropolitan area than inside it. There is also a higher percentage of people in extreme poverty outside of the metropolitan areas. It can be observed that outside of the metropolitan area there are variations among the homes to which families with social difficulties and those that do not have access (AMB—Can rent: 7.9; Cannot rent: 7.6. OAMB—Can rent: 7.7; Cannot rent: 7.7).

The number of rental homes available for a family at risk of social exclusion that would allow them to save at least 30% of their income is just 9 out of the total of 11,097 homes (all of them outside the AMB). Income also shows how families with social limitations can live in areas where gross income per capita income is lower (Can rent: 41,990; Cannot rent: 54,708). If we look at the situation by area, the phenomenon is the same but the value for gross income per person and per household is higher within the AMB.

Green areas present no differences in terms of the homes that can be rented and those that cannot (Can rent: 7.847, Cannot rent: 7.847). There is, however, a difference regarding the NDVI near to the homes, with those that cannot be rented presenting a lower value (Can

rent: 0.226, Cannot rent: 0.221). If we analyse the green areas, it is outside the metropolitan area that there are higher values of NDVI and green areas. Notably, when differentiating the areas it was observed that the homes that cannot be rented have higher NDVI values (AMB—Can rent: 0.285, Cannot rent: 0.312. OAMB—Can rent: 0.213, Cannot rent: 0.218).

In general, the homes that can be rented have lower levels of PM_{10} (Can rent: 23.567; Cannot rent: 23.893) and NO_2 (Can rent: 26.358; Cannot rent: 27.922), and higher levels of O_3 (Can rent: 51.111; Cannot rent: 49.978). However, when we looked at the different areas, it was detected that the levels of PM_{10} (AMB—Can rent: 23.081; Cannot rent: 22.914. OAMB—Can rent: 24.091; Cannot rent: 23.893) are reversed and are higher in the houses that can be rented by low-income families. Another way to analyse pollutant levels is by using the limits proposed by the WHO. If we take the former level recommended by this organisation, we can see that most properties are above the recommended PM_{10} levels. However, with the new WHO-recommended levels, the number of households with an optimal level of pollution is reduced, especially regarding NO_2.

If we analyse the number of crimes, the analysis shows that this number is lower for properties that can be rented by families with financial difficulties (Can rent: 9780; Cannot rent: 14,303). When observed by area, it is outside the metropolitan area that families at risk of poverty live in homes where there are more crimes (OAMB—Can rent: 9121; Cannot rent: 8727). In contrast, in the metropolitan area, crime is higher in areas where homes can only be rented by wealthy families (AMB—Can rent: 14,303; Cannot rent: 15,074).

3.2. Results of the Estimation of the Generalised Linear Models (GLM)

Tables 3 and 4 show the Odds Ratio of the GLM estimation models, with which we looked for the association between air pollutants, socioeconomic status, and the vegetation near a home and the possibility of not renting/renting by a family at risk of social exclusion. We controlled for socioeconomic and demographic variables and non-observed confounders for all these factors. Tables 3 and 4 also show the 95% confidence intervals (95% ICr, as of now) and their *p*-value. The variables on income (per person and per household) were not significant in terms of the possibility of renting. Similarly, the new limits recommended by the WHO and the possibility to save were not significant.

Table 3. Association between air pollutants and socioeconomic variables with the possibility of not renting a home for a family at risk of social exclusion.

Variable	UNADJUSTED			ADJUSTED [1]		
	OR (95% CI)	Pr (>|z|)		OR (95% CI)	Pr (>|z|)	
Percentage of people at risk of poverty threshold [Quartile 1]						
Risk of poverty threshold Q2	0.6464 (0.5496–0.7601)	1.33×10^{-7}	(***)	0.7887 (0.6481–0.9593)	0.017621	(*)
Risk of poverty threshold Q3	0.3884 (0.3426–0.4401)	$<2 \times 10^{-16}$	(***)	0.5139 (0.4362–0.6051)	1.55×10^{-15}	(***)
Risk of poverty threshold Q4	0.2855 (0.2492–0.3268)	$<2 \times 10^{-16}$	(***)	0.4221 (0.3502–0.5082)	$<2 \times 10^{-16}$	(***)
Average value NDVI range 500 metres [Quartile 1]						
NDVI Q2	0.7982 (0.7013–0.9082)	0.00063	(***)	1.1946 (1.0044–1.4209)	0.044492	(*)
NDVI Q3	0.8224 (0.7243–0.9336)	0.00253	(**)	1.3942 (1.1679–1.6649)	0.000237	(***)
NDVI Q4	0.6191 (0.5437–0.7048)	4.32×10^{-13}	(***)	1.3155 (1.0616–1.6312)	0.012317	(*)

Table 3. *Cont.*

Variable	UNADJUSTED			ADJUSTED [1]		
	OR (95% CI)	Pr (>|z|)		OR (95% CI)	Pr (>|z|)	
Average PM_{10} [Quartile 1]						
PM_{10} Q2	0.7859 (0.6899–0.8953)	0.00029	(***)	0.6841 (0.5368–0.8716)	0.002139	(**)
PM_{10} Q3	1.1249 (0.9769–1.2955)	0.10209		1.0607 (0.8110–1.3867)	0.666568	
PM_{10} Q4	1.1275 (0.9615–1.3222)	0.13971		0.6379 (0.4802–0.8464)	0.001873	(**)
Green areas per inhabitant (m^2)	0.9976 (0.9947–1.0005)	0.11222		2.9040 (1.7695–4.7033)	1.81×10^{-5}	(***)
Is the average PM_{10} over or under the former WHO limits? [over limit]						
pm10_expounder limit WHO	0.4358 (0.3656–0.5189)	$<2 \times 10^{-16}$	(***)	0.5315 (0.4282–0.6589)	8.97×10^{-9}	(***)
Average NO_2 [Quartile 1]						
NO_2 Q2	0.3887 (0.3418–0.4418)	$<2 \times 10^{-16}$	(***)	0.4399 (0.3770–0.5128)	$<2 \times 10^{-16}$	(***)
NO_2 Q3	0.5584 (0.4788–0.6510)	1.05×10^{-13}	(***)	0.8091 (0.6689–0.9783)	0.028884	(*)
NO_2 Q4	0.5103 (0.4338–0.6001)	4.53×10^{-16}	(***)	0.8134 (0.6729–0.9831)	0.032684	(*)
Is the average NO_2 over or under the former WHO limits? [over limit]						
no2_expounder limit WHO	1.0207 (0.8359–1.2461)	0.84066		0.7123 (0.5630–0.9008)	0.004653	(**)
Average O_3 [Quartile 1]						
O_3 Q2	0.9915 (0.8670–1.1339)	0.90113		1.0416 (0.8929–1.2152)	0.604212	
O_3 Q3	0.9834 (0.8533–1.1335)	0.81705		1.5826 (1.3398–1.8704)	6.88×10^{-8}	(***)
O_3 Q4	0.8432 (0.7188–0.9892)	0.03627	(*)	1.6228 (1.3386–1.9691)	8.75×10^{-7}	(***)
Number of crimes [Quartile 1]						
Number of crimes Q2	1.4464 (1.2765–1.6393)	7.32×10^{-9}	(***)	1.1324 (0.9578–1.3391)	0.145864	
Number of crimes Q3	1.4613 (1.2880–1.6586)	4.08×10^{-9}	(***)	1.4342 (1.1794–1.7444)	0.000304	(***)
Number of crimes Q4	4.3761 (3.5951–5.3335)	$<2 \times 10^{-16}$	(***)	4.2064 (2.9383–6.0237)	4.31×10^{-15}	(***)

(***) = $p \leq 0.001$; (**) = $p \leq 0.01$; (*) = $p \leq 0.05$. Source: authors' own elaboration. [1] Additionally, adjusted for the control variables.

Table 4. Association between air pollutants and socioeconomic and health variables and the possibility of not renting a home for a family at risk of social exclusion by area.

Variable	Adjusted			Adjusted 1			Adjusted 1		
	Non Filtred			Metropolitan Area			Out Of Metropolitan Area		
	OR (95% CI)	Pr (>|z|)		OR (95% CI)	Pr (>|z|)		OR (95% CI)	Pr (>|z|)	
Percentage of people at risk of extreme poverty threshold	0.916365 (0.904883–0.927941)	$<2 \times 10^{-16}$	(***)	0.809471 (0.786280–0.833023)	$<2 \times 10^{-16}$	(***)	1.056953 (1.007717–1.106804)	0.02044	(*)
Average value NDVI range 500 metres	0.217895 (0.120654–0.392645)	4.13×10^{-7}	(***)	2.251505 (0.870287–5.848026)	0.094835	(·)	0.396411 (0.142164–1.099841)	0.07613	(·)
Average PM_{10}	0.970661 (0.950686–0.991048)	0.004992	(**)	0.929724 (0.891240–0.969703)	0.000708	(***)	1.080775 (1.031068–1.133166)	0.00125	(**)
Square metres of green area per inhabitant	1.006783 (1.003523–1.010092)	4.64×10^{-5}	(***)	1.041741 (1.024869–1.059444)	1.30×10^{-6}	(***)	1.007544 (1.002478–1.012570)	0.00309	(**)
Average NO_2	1.003829 (0.995226–1.012518)	0.384383		1.023022 (1.011949–1.034271)	4.31×10^{-5}	(***)	0.976399 (0.957402–0.995677)	0.01688	(*)
Average O_3	1.013291 (1.005693–1.020963)	0.000593	(***)	1.019605 (1.009848–1.029496)	7.81×10^{-5}	(***)	1.004532 (0.989968–1.019289)	0.54348	
Number of crimes	1.000049 (1.000042–1.000056)	$<2 \times 10^{-16}$	(***)	1.000142 (1.000126–1.000159)	$<2 \times 10^{-16}$	(***)	1.000037 (1.000005–1.000069)	0.02233	(*)

(***) = $p \leq 0.001$; (**) = $p \leq 0.01$; (*) = $p \leq 0.05$; (·) = $p \leq 0.1$. Source: authors' own elaboration. [1] Additionally, adjusted for the control variables.

There is a relationship between areas with a high percentage of people below the poverty line and the ease of renting a home for families at risk of vulnerability. The higher the percentage of population below the threshold, the lower the possibility of not being able to rent a home compared to areas with a lower percentage of population below the threshold. In the second quartile, the probability of a home not being rented decreases by 21.13% and by 48.61% and 57.79% in the third and fourth quartiles. With crimes, the phenomenon is reversed, the greater the number of crimes, it's harder not to rent compared to homes in areas where there are fewer crimes. In the third quartile, the probability of not renting increases by 43.42% and in the quartile with the highest number of crimes, the probability increases by 320.64%.

Relationships were also observed between the possibility of not renting a home and pollutants. For the PM_{10} pollutant, the second and last quartiles are less likely not to be rented than the homes in the first pollution quartile. The odds decrease by 31.59% and 36.21%, respectively. For the NO_2 pollutant, the probability of not being rented decreases to a lesser extent compared to the least polluted quartile. In the second quartile the probability decreases by 56.01%, and in the third and fourth quartiles the probability decreases by 19.09% and 18.66%, respectively. For O_3, the higher the level of contamination, the greater the probability of not being able to be rented compared to less contaminated homes. In the third quartile the probability of not being rented increases by 58.26% and in the fourth quartile the probability is 62.28%.

In addition, if the homes are in areas above the old limits indicated by the WHO, the probability of not renting decreases by 46.85% if they exceed PM_{10} levels and by 28.77% if they exceed NO_2 levels.

Regarding green spaces, for each increase of one m^2 of green area per inhabitant, the probability of not renting the home will increase by 190%. If we look at the NDVI close to the home, the higher the vegetation index, the more difficult it is not to rent compared to homes in areas where there is a lower index. In the second quartile the probability increases by 19.46%, in the third quartile by 39.42%, and in the quartile with the highest number of crimes the percentage is a little lower at 31.55%.

If we carry out the study considering the areas where the homes are located, we can detect divergences between them. In this case, the study was performed with linear variables to facilitate interpretation and comparison.

The probability of not renting a home is reduced by 8.36% for every 1% increase in the population at risk of extreme poverty. The effect is different inside and outside the metropolitan area. Within the metropolitan area, the probability reduction is 19.05%. On the other hand, outside the metropolitan area the probability increases by 5.70%. Crimes have a similar effect in the two zones. For each crime that occurs, the probability increases by 0.005%. Within the metropolitan area it increases by 0.014%, while outside the metropolitan area the probability increases by 0.004.

If we look at pollutants, the effects are different for each 1 µg/m3 increase. For each increase of 1 $\mu g/m^3$ in the PM_{10} pollutant, the probability of not renting decreases by 2.93%. We observed a different pattern between areas. In the metropolitan area, the probability decreases by 7.03%. Outside the area the probability increases by 8.08%. The NO_2 pollutant had the opposite effect between areas to PM_{10}. For each increase of 1 $\mu g/m^3$, the probability of not renting increases by 0.38%. Within the metropolitan area the probability increases by 2.30%, outside the probability it decreases by 2.36%. For the pollutant O_3, for each increase of 1 $\mu g/m^3$ the probability of not renting decreased by 1.32%. Inside and outside the metropolitan area, the probability of not renting increases by 1.96% and 0.45%, respectively.

As for green areas, for each increase of 1 m^2 of green area per inhabitant, the probability of not renting will also increase by 0.68%. Within the metropolitan area, the probability will increase by a greater magnitude, 4.17%. On the other hand, outside the metropolitan area the probability will increase by 0.75%. The nearby NDVI showed a different pattern to the m^2 of green area. For every 1% increase in the index, the probability of not renting decreases by 78.21%. Within the metropolitan area the probability increases by 125.15% and, on the other hand, outside the metropolitan area the probability decreases by 60.35%.

4. Discussion

The reduction of poverty and related inequalities is one of the great struggles repeated cyclically throughout history. In recent years, it has been shown that increased productivity in not reflected in society in terms of salary increases or a greater ability to cover basic material needs [94]. We can also find in the literature multiple studies on where and how poverty is distributed [95–97] among the different social classes [98].

As has been shown, the lower and low-income social classes have higher mortality rates than the wealthy classes [99–101]. Although the aim of this study was not to determine the relative risk of mortality of families below the poverty line, we find significant relationships between the areas with the highest levels of pollutants, PM_{10}, NO_2, and O_3, and the tendency for these families to live there. As is well known, the most disadvantaged social classes tend to have higher exposure to environmental pollutants than the rest of the population [102], causing a higher probability of developing diseases [103–106].

There are various articles that study the relationship between exposure to environmental pollutants and social class [107–110]. Most of them are based on the individual or the socio-economic status (SES) of the areas, and the authors relate these to the levels of environmental pollution recorded. As in some other studies [111,112], our work was carried out with reference to housing, subsequently linking this to the area where the homes are located. In Oslo [111], it was found that socially deprived neighbourhoods have a greater exposure to air pollution. Wheeler and Ben-Shlomo [112], analysing a survey from England, found that socially deprived neighbourhoods have higher exposure to air pollution. However, our consideration is the possibility of the home being rented to low-income families, based on the real market supply. Notably too, our study includes different types of cities, while previous studies have generally considered metropolises and major cities. Yet, despite these differences, the results are similar, only differing when we analyse our results by area type. In this regard, the similarities with previous studies are accentuated in the more urbanised areas and differ in the more rural areas. These differences may be due to variations in the type of territory, the vegetation levels, and the socioeconomic and demographic status of the territories studied. We must also consider the possibility that the new generations in

Catalonia have changed their housing consumption patterns and prefer to rent rather than buy. In fact, the most recent studies [14] show that young people want to buy but cannot.

This difference between zones is repeated among various key variables, but with a reversed trend. Urban greenery per capita, for example, is a green variable of high interest to most major cities worldwide [113,114]. These cities stand out for their high population density [115,116], accentuating problems of access to vegetation, while the direct relationship between vegetation and health is well known. Even in large cities, policies and regulations are in force to increase and protect the number of square kilometres of green areas per capita and their vegetation [117,118]. There are several studies linking inequality and green areas [119,120]. As with previous studies, we found that an increase in square metres of green space per capita acts as a barrier to access for low-income families. The same phenomenon occurs if we look at studies of NDVI close to the home. When we look at the phenomenon in rural areas, the patterns of urban green space per capita and NDVI are reversed. It is the areas with less urban green space that generate access difficulties for families at risk of poverty. This phenomenon can be explained by the fact that there is a large volume of greenery in the rural areas of Catalonia. The evidence shows that historically there has been residential segregation, with poverty concentrated in zones or neighbourhoods [121,122]. These areas end up becoming communities of vulnerable people, perpetuating the poverty trap [123,124]. In our study, we observed this very trend. The higher the percentage of the population at risk of poverty, the more likely they are to rent these homes. Results show [125] that approximately 18% of European households and 10% of Spanish households have difficulty meeting their monthly payments. In our study, we found that there is a minimum population of 7.7% at risk of social exclusion who may present these difficulties, even in affluent neighbourhoods. These differences can be explained by the fact that in the present study we only considered data from our territory. If we differentiate the data by areas, we can see that in rural areas there is a low segregation of the population with low resources in the most precarious areas, which is not the case in urban areas.

The international literature shows that poverty is one of the most stable predictors of crime [126,127]. Moreover, the degree of urbanisation and safety are also directly related to the crimes committed [128]. Areas with greater social deprivation are associated with a higher level of crime [129,130]. Our results show that the most depressed areas have the lowest crime rates. If we look at behaviour in rural areas, it does reproduce the logic pointed out in previous studies. However, in urban areas, housing to which low-income families do not have access is found in the neighbourhoods with the highest crime rates. There are two possible reasons for these differences. They could be explained by the type of data identified as crimes, since there may be limitations when it comes to capturing data through administrative record channels. It could also be possible that crimes are not recorded in the place where they occurred, but rather in the place where the people who commit them live. We find it plausible to think that people who commit crimes do so more in rich areas than in poor ones.

Our study, unlike most of those mentioned above, performs an analysis of housing, allowing data to be obtained inside and outside large cities to see the similarities and differences. We also focused on small areas (either municipality or district), although this made the study more difficult because it meant creating different control variables. To this effect, we obtained data of multiple types that have an impact on inequalities in large cities and the most depopulated areas.

5. Conclusions

Out study may have some limitations stemming from its design. Firstly, the ecological inference fallacy must be considered when working with an ecological study. This fallacy complicates inferences at the individual level, given that there can be confounding elements inherent in the design typology. However, we tried to control the bias in the models as far as possible by including social, environmental, demographic, geographical, crime, and

confounding variables, all of which were addressed at a small territorial level. It would be interesting to be able to carry out the study through individuals rather than through the possibility of renting to be able to verify whether these data are correct or, conversely, to show new variations in the inequalities.

Secondly, the processing of data on different scales may be hiding inequalities and altering the study. This could be especially pertinent with reference to rural areas where it is more difficult to obtain the same level of data granularity as in urban areas. Furthermore, a family living in one area does not necessarily share the same average value as another family in the area. This fact leads to a possibly random error of measurement. This explains why the data have been measured with error, because otherwise we would obtain inconsistent estimators [131]. Lastly, we observe that there is a greater volume of data in the Barcelona area than in other areas. Notably, this province is home to more than 73% of the total population of the territory.

We believe these limitations are offset by the strengths of the study. Firstly, we conducted a study using small areas that allowed district data to be analysed. Although we are not the only authors to conduct a study of small areas, the model responds well to both small and rural areas. Secondly, the models obtained used multiple observed and unobserved confounders. Thirdly, we combined an interesting set of variables that are closely linked to inequalities. All of them stem from poverty and although studied in a combined way, our model can respond to multiple intersectionalities. Fourthly, the data obtained study inequality through the market supply curve, allowing the reality of the housing market at any given time to be captured and its limitations to be observed, which would be more difficult if worked from the demand curve or the balance point.

Lastly, we obtained a model that encompasses different key aspects caused by poverty: environmental pollution, the poverty cycle and trap, crime, and vegetation in relation to small areas and of different types.

Although the present study only considers outdoor pollutants, it would be interesting to obtain a more accurate reading of the exposure of each home to indoor pollution. It would also be interesting to be able to cross reference the different variables in the model at the individual level and see how they develop in the different districts.It would be advisable to carry out a study at the Catalonia level to determine if there are areas that have already become or are in the process of becoming places where people with high or low purchasing power are concentrated in order to deconstruct these areas of poverty or wealth.

We think that this article can shed light on the housing problem in Catalonia and that our results can be used to adjust the different policies implemented in the fight against poverty. In the first place, inequality is not generated solely by a lack of economic resources. Policies on environmental contamination should be accelerated to reduce future morbidity problems in the population, which will culminate in affecting future public health policies. In addition, a way must be found to encourage and help low-income families to save so as not to fall into the poverty trap while ensuring a healthy environment to reduce the latent inequality present throughout the territory. These social, economic, and environmental policies must be resolved today so as to have a smaller, less significant impact on future policies and on the Catalan economy.

Finally, the model is easily reproducible at other scales, which makes it possible to reduce the ecological fallacy and also to be used in other countries.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijerph20085578/s1, Suppelementary File S1. Methods. Supplementary Table S1. Variables.

Author Contributions: Conceptualization, X.P.; Data curation, X.P. and M.S.; Investigation, X.P. and M.S.; Methodology, X.P. and M.S.; Project administration, X.P. and M.S.; Supervision, M.S.; Visualization, M.S.; Writing—original draft, X.P. and M.S.; Writing—review and editing, X.P. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data, including the code to produce the figures, can be requested from the first author (xperafita@dipsalut.cat).

Acknowledgments: This study was carried out within the 'Cohort-Real World Data' subprogram of CIBER of Epidemiology and Public Health (CIBERESP). We appreciate the comments of two anonymous reviewers and of the academic editors of a previous version of this work who, without doubt, helped us to improve our work. The usual disclaimer applies.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wilderink, L.; Bakker, I.; Schuit, A.J.; Seidell, J.C.; Pop, I.A.; Renders, C.M. A Theoretical Perspective on Why Socioeconomic Health Inequalities Are Persistent: Building the Case for an Effective Approach. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8384. [CrossRef] [PubMed]
2. Navarro, V.; Shi, L. The Political Context of Social Inequalities and Health. *Int. J. Health Serv.* **2001**, *31*, 1–21. [CrossRef] [PubMed]
3. Power, A. Social Inequality, Disadvantaged Neighbourhoods and Transport Deprivation: An Assessment of the Historical Influence of Housing Policies. *J. Transp. Geogr.* **2012**, *21*, 39–48. [CrossRef]
4. Yan, Y.; Gai, X. High Achievers from Low Family Socioeconomic Status Families: Protective Factors for Academically Resilient Students. *Int. J. Envron. Res. Public Health* **2022**, *19*, 15882. [CrossRef]
5. Thomson, K.; Hillier-Brown, F.; Todd, A.; McNamara, C.; Huijts, T.; Bambra, C. The Effects of Public Health Policies on Health Inequalities in High-Income Countries: An Umbrella Review. *BMC Public Health* **2018**, *18*, 869. [CrossRef]
6. Piketty, T. *Capital in the Twenty-First Century*; Harvard University Press: Cambridge, MA, USA, 2015.
7. Henkin, L. *How Nations Behave: Law and Foreign Policy*, 2nd ed.; Published for the Council on Foreign Relations by Columbia University Press; Columbia University Press: New York, NY, USA, 1979; ISBN 0231047568.
8. Frazer, H.; Marlier, E. Homelessness and housing exclusion across EU member states, analysis and suggestions on the way forward. In *EU Network of Independent Experts on Social Inclusion*; CEPS/INSTEAD: Luxembourg, 2009.
9. Sandel, M.; Wright, R. When Home Is Where the Stress Is: Expanding the Dimensions of Housing That Influence Asthma Morbidity. *Arch. Dis. Child.* **2006**, *91*, 942–948. [CrossRef]
10. World Health Organitzation. *Environmental Health Inequalities in Europe: Assessment Report*; World Health Organitzation: Geneva, Switzerland, 2012.
11. Andrews, D.; Caldera Sánchez, A. The Evolution of Homeownership Rates in Selected OECD Countries: Demographic and Public Policy Influences. *OECD J. Econ. Stud.* **2011**, *2011*, 1–37. [CrossRef]
12. Baker, E.; Lester, L.; Mason, K.; Bentley, R. Mental Health and Prolonged Exposure to Unaffordable Housing: A Longitudinal Analysis. *Soc. Psychiatry Psychiatr. Epidemiol.* **2020**, *55*, 715–721. [CrossRef]
13. Denary, W.; Fenelon, A.; Schlesinger, P.; Purtle, J.; Blankenship, K.M.; Keene, D.E. Does Rental Assistance Improve Mental Health? Insights from a Longitudinal Cohort Study. *Soc. Sci. Med.* **2021**, *282*, 114100. [CrossRef]
14. Fuster, N.; Arundel, R.; Susino, J. From a Culture of Homeownership to Generation Rent: Housing Discourses of Young Adults in Spain. *J. Youth Stud.* **2019**, *22*, 585–603. [CrossRef]
15. Czischke, D.; van Bortel, G. An Exploration of Concepts and Polices on 'Affordable Housing' in England, Italy, Poland and The Netherlands. *J. Hous. Built Environ.* **2018**, *2018*, 1–21. [CrossRef]
16. Caturianas, D.; Lewandowski, P.; Sokołowski, J.; Kowalik, Z.; Barcevičius, E. *Policies to Ensure Access to Affordable Housing EN STUDY*; European Parliament: Luxembourg, 2020.
17. Kisiala, W.; Rącka, I. Spatial and Statistical Analysis of Urban Poverty for Sustainable City Development. *Sustainability* **2021**, *13*, 858. [CrossRef]
18. Buonanno, P.; Montolio, D.; Raya-Vílchez, J.M. Housing Prices and Crime Perception. *Empir. Econ.* **2012**, *45*, 305–321. [CrossRef]
19. Gu, Y.; Wang, Q.; Yi, G. Stationary Patterns and Their Selection Mechanism of Urban Crime Models with Heterogeneous Near-Repeat Victimization Effect. *Eur. J. Appl. Math.* **2017**, *28*, 141–178. [CrossRef]
20. Campagna, G. Linking Crowding, Housing Inadequacy, and Perceived Housing Stress. *J. Environ. Psychol.* **2016**, *45*, 252–266. [CrossRef]
21. Maryanti, M.; Khadijah, H.; Uzair, A.; Rahman, M. The Urban Green Space Provision Using the Standards Approach: Issues and Challenges of Its Implementation in Malaysia. *WIT Trans. Ecol. Environ.* **2017**, *210*, 369–379.
22. Roe, J.J.; Aspinall, P.A.; Ward Thompson, C. Coping with Stress in Deprived Urban Neighborhoods: What Is the Role of Green Space According to Life Stage? *Front. Psychol.* **2017**, *8*, 1760. [CrossRef]
23. Ulmer, J.M.; Wolf, K.L.; Backman, D.R.; Tretheway, R.L.; Blain, C.J.; O'Neil-Dunne, J.P.; Frank, L.D. Multiple Health Benefits of Urban Tree Canopy: The Mounting Evidence for a Green Prescription. *Health Place* **2016**, *42*, 54–62. [CrossRef]

24. Wolch, J.R.; Byrne, J.; Newell, J.P. Urban Green Space, Public Health, and Environmental Justice: The Challenge of Making Cities 'Just Green Enough'. *Landsc. Urban Plan.* **2014**, *125*, 234–244. [CrossRef]
25. Villeneuve, P.J.; Jerrett, M.; Su, J.G.; Weichenthal, S.; Sandler, D.P. Association of Residential Greenness with Obesity and Physical Activity in a US Cohort of Women. *Environ. Res.* **2018**, *160*, 372–384. [CrossRef]
26. Morancho, A.B. A Hedonic Valuation of Urban Green Areas. *Landsc. Urban Plan.* **2003**, *66*, 35–41. [CrossRef]
27. Dadvand, P.; Rivas, I.; Basagaña, X.; Alvarez-Pedrerol, M.; Su, J.; de Castro Pascual, M.; Amato, F.; Jerret, M.; Querol, X.; Sunyer, J.; et al. The Association between Greenness and Traffic-Related Air Pollution at Schools. *Sci. Total Environ.* **2015**, *523*, 59–63. [CrossRef] [PubMed]
28. Beimer, W.; Maennig, W. Noise Effects and Real Estate Prices: A Simultaneous Analysis of Different Noise Sources. *Transp. Res. D Transp. Environ.* **2017**, *54*, 282–286. [CrossRef]
29. Tang, M.; Niemeier, D. How Does Air Pollution Influence Housing Prices in the Bay Area? *Int. J. Environ. Res. Public Health* **2021**, *18*, 12195. [CrossRef]
30. Das, R.C.; Chatterjee, T.; Ivaldi, E. Nexus between Housing Price and Magnitude of Pollution: Evidence from the Panel of Some High- and-Low Polluting Cities of the World. *Sustainability* **2022**, *14*, 9283. [CrossRef]
31. Novoa, A.M.; Bosch, J.; Díaz, F.; Malmusi, D.; Darnell, M.; Trilla, C. El Impacto de La Crisis En La Relación Entre Vivienda y Salud. Políticas de Buenas Prácticas Para Reducir Las Desigualdades En Salud Asociadas Con Las Condiciones de Vivienda. *Gac. Sanit.* **2014**, *28*, 44–50. [CrossRef]
32. IDESCAT. Statistical Yearbook of Catalonia. Altitude, Surface Area and Population. Municipalities. Available online: https://www.idescat.cat/pub/?id=aec&n=925&lang=es (accessed on 1 September 2021).
33. Gutiérrez, A.; Domènech, A. Identifying the Socio-Spatial Logics of Foreclosed Housing Accumulated by Large Private Landlords in Post-Crisis Catalan Cities. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 313. [CrossRef]
34. Checa, J.; Nel·lo, O.; Manuel, J.; Castellano, P.; Piñeira-Mantiñán, J.; Manuel González Pérez, J.; Plane, D.; Springer, S. Residential Segregation and Living Conditions. An Analysis of Social Inequalities in Catalonia from Four Spatial Perspectives. *Urban Sci.* **2021**, *5*, 45. [CrossRef]
35. López-Rodríguez, D.; Matea, M.; de los Lanos, M. Public Intervention in the Housing Rental Market: A Review of the International Experience (La intervención pública en el mercado del alquiler de vivienda: Una revisión de la experiencia internacional). *Occas. Pap.* **2020**, 2002. [CrossRef]
36. Torres-Pruñonosa, J.; García-Estévez, P.; Raya, J.M.; Prado-Román, C. How on Earth Did Spanish Banking Sell the Housing Stock? *Sage Open* **2022**, *12*. [CrossRef]
37. Álvarez-Román, L.; García-Posada, M. Are House Prices Overvalued in Spain? A Regional Approach. *Econ. Model.* **2021**, *99*, 105499. [CrossRef]
38. Blanco-Romero, A.; Blázquez-Salom, M.; Cànoves, G. Barcelona, Housing Rent Bubble in a Tourist City. Social Responses and Local Policies. *Sustainability* **2018**, *10*, 2043. [CrossRef]
39. Garcia-López, M.À.; Jofre-Monseny, J.; Martínez-Mazza, R.; Segú, M. Do Short-Term Rental Platforms Affect Housing Markets? Evidence from Airbnb in Barcelona. *J. Urban Econ.* **2020**, *119*, 103278. [CrossRef]
40. Garrido-Yserte, R.; Mañas-Alcón, E.; Gallo-Rivera, M.T. Housing and Cost of Living: Application to the Spanish Regions. *J. Hous. Econ.* **2012**, *21*, 246–255. [CrossRef]
41. UNESCO. *Basic Texts of the 2005 Convention on the Protection and Promotion of the Diversity of Cultural Expressions*, 2019th ed.; UNESCO: Paris, France, 2019.
42. Piasek, G.; Fernández Aragón, I.; Shershneva, J.; Garcia-Almirall, P. Assessment of Urban Neighbourhoods’ Vulnerability through an Integrated Vulnerability Index (IVI): Evidence from Barcelona, Spain. *Soc. Sci.* **2022**, *11*, 476. [CrossRef]
43. Instituto Nacional de Estadística (INE). Household Income Distribution Atlas. Available online: https://www.ine.es/dynt3/inebase/es/index.htm?padre=7132 (accessed on 3 September 2021).
44. Donovan, G.H.; Butry, D.T.; Michael, Y.L.; Prestemon, J.P.; Liebhold, A.M.; Gatziolis, D.; Mao, M.Y. The Relationship between Trees and Human Health: Evidence from the Spread of the Emerald Ash Borer. *Am. J. Prev. Med.* **2013**, *44*, 139–145. [CrossRef]
45. Gascon, M.; Triguero-Mas, M.; Martínez, D.; Dadvand, P.; Rojas-Rueda, D.; Plasència, A.; Nieuwenhuijsen, M.J. Residential Green Spaces and Mortality: A Systematic Review. *Environ. Int.* **2016**, *86*, 60–67. [CrossRef]
46. Kondo, M.C.; Fluehr, J.M.; McKeon, T.; Branas, C.C. Urban Green Space and Its Impact on Human Health. *Int. J. Environ. Res. Public Health* **2018**, *15*, 445. [CrossRef]
47. Rojas-Rueda, D.; Nieuwenhuijsen, M.J.; Gascon, M.; Perez-Leon, D.; Mudu, P. Green Spaces and Mortality: A Systematic Review and Meta-Analysis of Cohort Studies. *Lancet Planet Health* **2019**, *3*, e469–e477. [CrossRef]
48. Institut Cartogràfic i Geològic de Catalunya Vegetation Index of Normalized Difference. Available online: https://www.icgc.cat/en/Public-Administration-and-Enterprises/Downloads/Aerial-photos-and-orthophotos/NDVI (accessed on 3 September 2021).
49. Generalitat de Catalunya Dades Del Mapa Urbanístic de Catalunya | Dades Obertes de Catalunya. Available online: https://analisi.transparenciacatalunya.cat/en/Urbanisme-infraestructures/Dades-del-mapa-urban-stic-de-Catalunya/epsm-zskb (accessed on 12 December 2021).
50. Dadvand, P.; Bartoll, X.; Basagaña, X.; Dalmau-Bueno, A.; Martinez, D.; Ambros, A.; Cirach, M.; Triguero-Mas, M.; Gascon, M.; Borrell, C.; et al. Green Spaces and General Health: Roles of Mental Health Status, Social Support, and Physical Activity. *Environ. Int.* **2016**, *91*, 161–167. [CrossRef]

51. Su, J.G.; Dadvand, P.; Nieuwenhuijsen, M.J.; Bartoll, X.; Jerrett, M. Associations of Green Space Metrics with Health and Behavior Outcomes at Different Buffer Sizes and Remote Sensing Sensor Resolutions. *Environ. Int.* **2019**, *126*, 162–170. [CrossRef]
52. Dominski, F.H.; Lorenzetti Branco, J.H.; Buonanno, G.; Stabile, L.; Gameiro da Silva, M.; Andrade, A. Effects of Air Pollution on Health: A Mapping Review of Systematic Reviews and Meta-Analyses. *Environ. Res.* **2021**, *201*, 111487. [CrossRef] [PubMed]
53. Han, C.; Xu, R.; Zhang, Y.; Yu, W.; Zhang, Z.; Morawska, L.; Heyworth, J.; Jalaludin, B.; Morgan, G.; Marks, G.; et al. Air Pollution Control Efficacy and Health Impacts: A Global Observational Study from 2000 to 2016. *Environ. Pollut.* **2021**, *287*, 117211. [CrossRef] [PubMed]
54. Noël, C.; Vanroelen, C.; Gadeyne, S. Qualitative Research about Public Health Risk Perceptions on Ambient Air Pollution. A Review Study. *SSM Popul. Health* **2021**, *15*, 100879. [CrossRef] [PubMed]
55. Saez, M.; Barceló, M.A. Spatial Prediction of Air Pollution Levels Using a Hierarchical Bayesian Spatiotemporal Model in Catalonia, Spain. *medRxiv* **2021**, *6*, 21258419. [CrossRef]
56. Saez, M.; Tobias, A.; Barceló, M.A. Effects of Long-Term Exposure to Air Pollutants on the Spatial Spread of COVID-19 in Catalonia, Spain. *Environ. Res.* **2020**, *191*, 110177. [CrossRef]
57. World Health Organization. Ambient (Outdoor) Air Pollution. Available online: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed on 6 October 2021).
58. Choe, J. Income Inequality and Crime in the United States. *Econ. Lett.* **2008**, *101*, 31–33. [CrossRef]
59. Coccia, M. A Theory of General Causes of Violent Crime: Homicides, Income Inequality and Deficiencies of the Heat Hypothesis and of the Model of CLASH. *Aggress Violent Behav.* **2017**, *37*, 190–200. [CrossRef]
60. Kim, B.; Seo, C.; Hong, Y.-O. A Systematic Review and Meta-Analysis of Income Inequality and Crime in Europe: Do Places Matter? *Eur. J. Crim. Policy Res.* **2020**, *2020*, 1–24. [CrossRef]
61. Bennett, T.; Holloway, K.; Farrington, D. The Statistical Association between Drug Misuse and Crime: A Meta-Analysis. *Aggress Violent Behav.* **2008**, *13*, 107–118. [CrossRef]
62. Mossos d'Esquadra Catàleg de Dades Obertes. Available online: https://mossos.gencat.cat/en/els_mossos_desquadra/indicadors_i_qualitat/dades_obertes/cataleg_dades_obertes/index.html (accessed on 3 September 2021).
63. Bove, V.; Elia, L. Migration, Diversity, and Economic Growth. *World Dev.* **2017**, *89*, 227–239. [CrossRef]
64. Foulkes, M.; Schafft, K.A. The Impact of Migration on Poverty Concentrations in the United States, 1995–2000. *Rural Sociol.* **2010**, *75*, 90–110. [CrossRef]
65. IDESCAT. Llindar de Risc de Pobresa. Per Composició de La Llar. Available online: https://www.idescat.cat/indicadors/?id=anuals&n=10411&lang=es (accessed on 12 December 2021).
66. ICGC. Base Municipal. Available online: https://www.icgc.cat/Administracio-i-empresa/Descarregues/Capes-de-geoinformacio/Base-municipal (accessed on 12 June 2021).
67. Institut Català de la Salut Atenció Primària Girona Àrea Bàsica de Salut (ABS). Available online: http://www.icsgirona.cat/ca/contingut/primaria/370 (accessed on 3 September 2021).
68. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* **1974**, *19*, 716–723. [CrossRef]
69. Raybaut, P. Spyder IDE 2009. Available online: https://www.spyder-ide.org/ (accessed on 17 December 2022).
70. Time—Time Access and Conversions. 2021. Available online: https://docs.python.org/3/library/time.html (accessed on 17 December 2022).
71. Prewitt, N.; Larson, S.M. Requests: HTTP for HumansTM. 2021. Available online: https://requests.readthedocs.io/en/latest/ (accessed on 17 December 2022).
72. Pérez, F.; Granger, B.E. IPython: A System for Interactive Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 21–29. [CrossRef]
73. Richardson, L. Bs4 Dummy Package for Beautiful Soup 2020. Available online: https://pypi.org/project/bs4/ (accessed on 18 December 2022).
74. Datetime—Basic Date and Time Types. Available online: https://docs.python.org/3/library/datetime.html (accessed on 18 December 2022).
75. Csv—CSV File Reading and Writing 2021. Available online: https://docs.python.org/3/library/csv.html (accessed on 18 December 2022).
76. The pandas development team Pandas-Dev/Pandas: Pandas 2020. Available online: https://doi.org/10.5281/zenodo.3509134 (accessed on 18 December 2022).
77. Os—Miscellaneous Operating System Interfaces 2021. Available online: https://docs.python.org/3/library/os.html (accessed on 18 December 2022).
78. Re—Regular Expression Operations. Available online: https://docs.python.org/3/library/re.html (accessed on 18 December 2022).
79. Math—Mathematical Functions 2021. Available online: https://docs.python.org/3/library/math.html#constants (accessed on 18 December 2022).
80. Msvcrt—Rutinas Útiles Del Entorno de Ejecución MS VC++ 2021. Available online: https://docs.python.org/es/3/library/msvcrt.html (accessed on 18 December 2022).
81. Tabulate PyPI 2021. Available online: https://pypi.org/project/tabulate/ (accessed on 18 December 2022).
82. Tkinter—Python Interface to Tcl/Tk 2022. Available online: https://docs.python.org/3/library/tkinter.html (accessed on 18 December 2022).

83. Da Costa-Luis, C.; Larroque, S.K.; Altendorf, K.; Mary, H.; Richard, S.; Korobov, M.; Raphael, N.; Ivanov, I.; Bargull, M.; Rodrigues, N. Tqdm: A Fast, Extensible Progress Bar for Python and CLI 2022. Available online: https://zenodo.org/record/7046742 (accessed on 18 December 2022).
84. Googlemaps PyPI 2021. Available online: https://pypi.org/project/googlemaps/ (accessed on 18 December 2022).
85. RStudio Team RStudio: Integrated Development Environment for R 2022. Available online: https://www.rstudio.com/categories/integrated-development-environment/ (accessed on 18 December 2022).
86. Bivand, R.; Keitt, T.; Rowlingson, B. Rgdal: Bindings for the "Geospatial" Data Abstraction Library 2022. Abstraction Library. Available online: https://CRAN.R-project.org/package=rgdal (accessed on 17 December 2022).
87. Bivand, R.; Rundel, C. Rgeos: Interface to Geometry Engine-Open Source ('GEOS') 2021. Available online: https://CRAN.R-project.org/package=rgeos (accessed on 17 December 2022).
88. Hijmans, R.J. Raster: Geographic Data Analysis and Modeling 2022. Available online: https://CRAN.R-project.org/package=raster (accessed on 17 December 2022).
89. Tennekes, M. Tmap: Thematic Maps in R. *J. Stat. Softw.* **2018**, *84*, 1–39. [CrossRef]
90. Bischl, B.; Lang, M.; Bossek, J.; Horn, D.; Richter, J.; Surmann, D. BBmisc: Miscellaneous Helper Functions for B. Bischl, 2022. Available online: https://cran.r-project.org/web/packages/BBmisc/BBmisc.pdf (accessed on 17 December 2022).
91. Wickham, H.; Miller, E.; Smith, D. Haven: Import and Export "SPSS", "Stata" and "SAS" 2022. Available online: https://CRAN.R-project.org/package=haven (accessed on 17 December 2022).
92. Wickham, H.; François, R.; Henry, L.; Müller, K. Dplyr: A Grammar of Data Manipulation. 2022. Available online: https://www.researchgate.net/publication/275646200_dplyr_A_Grammar_of_Data_Manipulation (accessed on 17 December 2022).
93. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2016; Available online: https://ggplot2.tidyverse.org/ (accessed on 17 December 2022).
94. Isidro Luna, V.M. The Persistence of Poverty in Capitalist Countries. *Econ. Inf.* **2016**, *400*, 67–82. [CrossRef]
95. Aaberge, R.; Brandolini, A. Multidimensional Poverty and Inequality. *Handb. Income Distrib.* **2015**, *2*, 141–216. [CrossRef]
96. Sandmo, A. The Principal Problem in Political Economy: Income Distribution in the History of Economic Thought. *Handb. Income Distrib.* **2015**, *2*, 3–65. [CrossRef]
97. Chen, X.; Pei, Z.; Chen, A.L.; Wang, F.; Shen, K.; Zhou, Q.; Sun, L. Spatial Distribution Patterns and Influencing Factors of Poverty-A Case Study on Key Country From National Contiguous Special Poverty-Stricken Areas in China. *Procedia Environ. Sci.* **2015**, *26*, 82–90. [CrossRef]
98. Guio, A.-C.; Marlier, E.; Nolan, B. *Improving the Understanding of Poverty and Social Exclusion in Europe*; Publications Office of the European Union: Luxembourg, 2021; ISBN 978-92-76-34284-7.
99. Brønnum-Hansen, H.; Foverskov, E.; Andersen, I. Income Inequality in Life Expectancy and Disability-Free Life Expectancy in Denmark. *J. Epidemiol. Community Health* **2021**, *75*, 145–150. [CrossRef]
100. Rehnberg, J.; Fors, S.; Fritzell, J. Divergence and Convergence: How Do Income Inequalities in Mortality Change over the Life Course? *Gerontology* **2019**, *65*, 313–322. [CrossRef] [PubMed]
101. Shi, J.; Tarkiainen, L.; Martikainen, P.; van Raalte, A. The Impact of Income Definitions on Mortality Inequalities. *SSM Popul. Health* **2021**, *15*, 100915. [CrossRef] [PubMed]
102. Hajat, A.; Hsia, C.; O'Neill, M.S. Socioeconomic Disparities and Air Pollution Exposure: A Global Review. *Curr. Environ. Health Rep.* **2015**, *2*, 440–450. [CrossRef] [PubMed]
103. Dominici, F.; McDermott, A.; Daniels, M.; Zeger, S.L.; Samet, J.M. Revised Analyses of the National Morbidity, Mortality, and Air Pollution Study: Mortality Among Residents Of 90 Cities. *J. Toxicol. Environ. Health Part A.* **2006**, *68*, 1071–1092. [CrossRef]
104. Khomenko, S.; Cirach, M.; Pereira-Barboza, E.; Mueller, N.; Barrera-Gómez, J.; Rojas-Rueda, D.; de Hoogh, K.; Hoek, G.; Nieuwenhuijsen, M. Premature Mortality Due to Air Pollution in European Cities: A Health Impact Assessment. *Lancet Planet Health* **2021**, *5*, e121–e134. [CrossRef]
105. Liu, C.; Chen, R.; Sera, F.; Vicedo-Cabrera, A.M.; Guo, Y.; Tong, S.; Coelho, M.S.Z.S.; Saldiva, P.H.N.; Lavigne, E.; Matus, P.; et al. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *N. Engl. J. Med.* **2019**, *381*, 705–715. [CrossRef]
106. Strak, M.; Weinmayr, G.; Rodopoulou, S.; Chen, J.; de Hoogh, K.; Andersen, Z.J.; Atkinson, R.; Bauwelinck, M.; Bekkevold, T.; Bellander, T.; et al. Long Term Exposure to Low Level Air Pollution and Mortality in Eight European Cohorts within the ELAPSE Project: Pooled Analysis. *BMJ* **2021**, *374*, 1904. [CrossRef]
107. Loizeau, M.; Buteau, S.; Chaix, B.; McElroy, S.; Counil, E.; Benmarhnia, T. Does the Air Pollution Model Influence the Evidence of Socio-Economic Disparities in Exposure and Susceptibility? *Environ. Res.* **2018**, *167*, 650–661. [CrossRef]
108. Samoli, E.; Stergiopoulou, A.; Santana, P.; Rodopoulou, S.; Mitsakou, C.; Dimitroulopoulou, C.; Bauwelinck, M.; de Hoogh, K.; Costa, C.; Marí-Dell'Olmo, M.; et al. Spatial Variability in Air Pollution Exposure in Relation to Socioeconomic Indicators in Nine European Metropolitan Areas: A Study on Environmental Inequality. *Environ. Pollut.* **2019**, *249*, 345–353. [CrossRef]
109. Tonne, C.; Milà, C.; Fecht, D.; Alvarez, M.; Gulliver, J.; Smith, J.; Beevers, S.; Ross Anderson, H.; Kelly, F. Socioeconomic and Ethnic Inequalities in Exposure to Air and Noise Pollution in London. *Environ. Int.* **2018**, *115*, 170–179. [CrossRef]
110. Zhou, Q.; Zhang, X.; Chen, J.; Zhang, Y. Do Double-Edged Swords Cut Both Ways? Housing Inequality and Haze Pollution in Chinese Cities. *Sci. Total Environ.* **2020**, *719*, 137404. [CrossRef]
111. Naess, O.; Piro, F.N.; Nafstad, P.; Smith, G.D.; Leyland, A.H. Air Pollution, Social Deprivation, and Mortality: A Multilevel Cohort Study. *Epidemiology* **2007**, *18*, 686–694. [CrossRef]

112. Wheeler, B.W.; Ben-Shlomo, Y. Environmental Equity, Air Quality, Socioeconomic Status, and Respiratory Health: A Linkage Analysis of Routine Data from the Health Survey for England. *J. Epidemiol. Community Health* **2005**, *59*, 948–954. [CrossRef]
113. Park, J.; Kim, J.H.; Lee, D.K.; Park, C.Y.; Jeong, S.G. The Influence of Small Green Space Type and Structure at the Street Level on Urban Heat Island Mitigation. *Urban Urban Green* **2017**, *21*, 203–212. [CrossRef]
114. Schetke, S.; Qureshi, S.; Lautenbach, S.; Kabisch, N. What Determines the Use of Urban Green Spaces in Highly Urbanized Areas?—Examples from Two Fast Growing Asian Cities. *Urban Urban Green* **2016**, *16*, 150–159. [CrossRef]
115. Laan, C.M.; Piersma, N. Accessibility of Green Areas for Local Residents. *Environ. Sustain. Indic.* **2021**, *10*, 100114. [CrossRef]
116. Wang, M.; Qiu, M.; Chen, M.; Zhang, Y.; Zhang, S.; Wang, L. How Does Urban Green Space Feature Influence Physical Activity Diversity in High-Density Built Environment? An on-Site Observational Study. *Urban Urban Green* **2021**, *62*, 127129. [CrossRef]
117. Lin, Y.; Shui, W.; Li, Z.; Huang, S.; Wu, K.; Sun, X.; Liang, J. Green Space Optimization for Rural Vitality: Insights for Planning and Policy. *Land Use Policy* **2021**, *108*, 105545. [CrossRef]
118. Wang, Q.; Lan, Z. Park Green Spaces, Public Health and Social Inequalities: Understanding the Interrelationships for Policy Implications. *Land Use Policy* **2019**, *83*, 66–74. [CrossRef]
119. Moran, M.R.; Bilal, U.; Dronova, I.; Ju, Y.; Gouveia, N.; Caiaffa, W.T.; Friche, A.A.D.L.; Moore, K.; Miranda, J.J.; Rodríguez, D.A. The Equigenic Effect of Greenness on the Association between Education with Life Expectancy and Mortality in 28 Large Latin American Cities. *Health Place* **2021**, *72*, 102703. [CrossRef]
120. Schüle, S.A.; Gabriel, K.M.A.; Bolte, G. Relationship between Neighbourhood Socioeconomic Position and Neighbourhood Public Green Space Availability: An Environmental Inequality Analysis in a Large German City Applying Generalized Linear Models. *Int. J. Hyg. Environ. Health* **2017**, *220*, 711–718. [CrossRef]
121. French, D.; Vigne, S. The Causes and Consequences of Household Financial Strain: A Systematic Review. *Int. Rev. Financ. Anal.* **2019**, *62*, 150–156. [CrossRef]
122. Kleinhans, R.; van der Land, M.; Doff, W.; Kleinhans, R.; Doff, Á.W.; Doff, W.; van der Land, M. Dealing with Living in Poor Neighbourhoods. *J. Hous. Built Environ.* **2010**, *25*, 381–389. [CrossRef]
123. Radosavljevic, S.; Haider, L.J.; Lade, S.J.; Schlüter, M. Implications of Poverty Traps across Levels. *World Dev.* **2021**, *144*, 105437. [CrossRef]
124. Yang, T.; Pan, H.; Zhang, X.; Greenlee, A.; Deal, B. How Neighborhood Conditions and Policy Incentives Affect Relocation Outcomes of Households from Low-Income Neighborhoods—Evidence from Intra-City Movement Trajectories. *Cities* **2021**, *119*, 103415. [CrossRef]
125. Eurostat Inability to Make Ends Meet-EU-SILC Survey. Available online: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_mdes09&lang=en (accessed on 19 December 2021).
126. McCall, P.L.; Land, K.C.; Parker, K.F. An Empirical Assessment of What We Know About Structural Covariates of Homicide Rates: A Return to a Classic 20 Years Later. *Homicide Stud.* **2010**, *14*, 219–243. [CrossRef]
127. Wilson, W.J. *The Truly Disadvantaged*, 2nd ed.; University of Chicago Press: Chicago, IL, USA, 2012; ISBN 9780226901268.
128. Massey, D.S. Inheritance of Poverty or Inheritance of Place? The Emerging Consensus on Neighborhoods and Stratification. *Contemp. Sociol.* **2013**, *42*, 690–695. [CrossRef]
129. Graif, C. Delinquency and gender moderation in the moving to opportunity intervention: The role of extended neighborhoods. *Criminology* **2015**, *53*, 366–398. [CrossRef]
130. Alkire, S.; Foster, J.E.; Seth, S.; Santos, M.E.; Roche, J.M.; Ballon, P. *Multidimensional Poverty Measurement and Analysis*; Oxford University Press: Oxford, UK, 2015; ISBN 978-19-0719-471-9.
131. Greene, W.H. *Econometric Analysis*, 8th ed.; Pearson: London, UK, 2018; ISBN 9353061075.

5 DISCUSSIÓ

La investigació realitzada en aquesta tesi tenia com a objectiu: identificar els mecanismes economètrics que s'utilitzen per tractar dades en entorns Big Data. Un cop identificats, s'han aplicat els mecanismes oportuns per estudiar la desigualtat en zones heterogèniament diferents, en quant a distribució de població es refereix. Les següents seccions aporten una discussió global dels principals resultats obtinguts en l'**Article I** i l'**Article II**. També es presenten les possibles limitacions i fortaleses associades.

5.1 Cohort electrònica

Les cohorts electròniques són una derivació de les cohorts tradicionals, que busquen a través d'eines electròniques: reduir els costos econòmics i de gestió associats als participants i a la captació d'informació.

En la literatura científica, trobem diferents cohorts que són purament electròniques, veure **Taula 7**: Nutrinet-Santé[160], Communauté de Patients pour la Recherche (ComPaRe)[161], Wales Multimorbidity e-Cohort (WMC)[162], 1970 British Cohort Study (BCS70)[163], Avon Longitudinal Study of Parents and Children (ALSPAC)[164], Millennium Cohort Study (MCS)[165], SAIL dementia e-cohort (SAIL-DeC)[166], UK COSMOS cohort[167], Electronic Framingham Heart Study (eFHS)[168], COVIDiet[169,170] i eCardia[171].

Taula 7. Breu descripció, disseny i període de les e-cohorts electròniques identificades.

e-Cohort	Mostreig	Breu descripció
NutriNet-Santé Study *França* (2009-2029)	**No probabilístic**, mostra de conveniència	Cohort francesa que té com a objectiu estudiar les relacions entre la seva dieta i estils de vida dels participants i la seva salut. La base poblacional de l'estudi és la població francesa, de la qual té una mostra de més de **200.000 participants**.
ComPaRe *França* (2017-Actualitat)	**No probabilístic**, mostra de conveniència	Cohort de reclutament de pacients adults amb afeccions cròniques a França per saber-ne les experiències i necessitats de més de **1.600 participants**.
Wales Multimorbidity e-Cohort *Regne Unit* (2000-2019)	**No probabilístic**, mostra de conveniència	Cohort gal·lesa que té com a objectiu estudiar la multimorbiditat dels usuaris del sistema de salut a Gal·les, des del 1990. La base poblacional de l'estudi és la població gal·lesa, de la qual té una mostra de més de **100.000 participants**.

1970 British Cohort Study *Regne Unit* (2020-2021)	**No probabilístic**, mostra de conveniència	Cohort anglesa que té com a objectiu estudiar el desenvolupament des de la infància dels nadons per veure'n el desenvolupament cognitiu, emocional, educatiu, econòmic i social. La base poblacional de l'estudi és la població anglesa, de la qual se n'han obtingut més de **17.000 participants**.
Avon Longitudinal Study of Parents and Children (ALSPAC) *Regne Unit* (1991-Actualitat)	**No probabilístic**, mostra de conveniència	Cohort anglesa que buscar observar com els factors biològics, socials, econòmics i ambientals influeixen en la salut i el benestar a mesura que els individus creixen. Es van reclutar **14.000 dones embarassades** entre 1990 i 1991 a Avon, UK.
Millennium Cohort Study *Regne Unit* (2000-Actualitat)	**Probabilístic**, Estratificat per conglomerats	Cohort anglesa que té com a objectiu seguir el desenvolupament de les condicions de vida dels nadons al Regne Unit. L'estudi segueix aproximadament a **19.000 infants**, durant el període 2000-2002.
SAIL dementia e-cohort (SAIL-DeC) *Regne Unit* (2016-Actualitat)	**No probabilístic**, mostra de conveniència	Cohort gal·lesa que té com a objectiu l'estudi dels i les participants vinculats a la demència i els seus factors de risc. Combina múltiples fonts de dades tant administratives i de registre com de qüestionaris en línia. Inicialment es van capturar uns **1.250.000 participants**.
UK COSMOS cohort *Regne Unit* (2012-Actualitat)	**No probabilístic**, mostra de conveniència	Cohort anglesa que busca conèixer quins factors ambientals, comportamentals i genètics poden afectar a la salut de les persones, enfocat a malalties cròniques. Inicialment es van reclutar més de **100.000 homes i dones de més de 18 anys**.
Electronic Framingham Heart Study (eFHS) *Framingham* (2017-Actualitat)	**No probabilístic**, mostra de conveniència	Evolució de la cohort tradicional aplicant la tecnologia, buscant ampliar les causes i els factors que provoquen les morbiditats i els factors de risc. La cohort va reclutar a **790 participants**.
COVIDiet *Múltiples països* (2020-2021)	**No probabilístic**, mostra de conveniència	Cohort de diferents països que investiga l'associació entre el tipus de dieta, els factors socioeconòmics i el risc i gravetat de la COVID-19 dels i les participants. La cohort compta amb més de **500.000 participants**.
eCARDIA Estados Unidos (1980-Actualitat)	**No probabilístic**, mostra de conveniència	Cohort nord-americana que investiga la salut cardiovascular de més de **6.000 participants**. Reuneix informació detallada sobre la salut dels i les participants així com altres aspectes com els seus estils de vida.

Font: Taula d'elaboració pròpia, a partir de les cerca a la base literària PubMed amb la cerca: "Cohort Studies"[MeSH Major Topic] AND ("new technologies"[Title/Abstract] OR "smartphone app"[Title/Abstract] OR "mobile application"[Title/Abstract] OR "Web-based"[Title/Abstract] OR "mHealth"[Title/Abstract] OR "mobile devices"[Title/Abstract]). De tots els articles detectats, s'han seleccionat el més rellevants d'acord amb el contingut.

Altres incorporen elements o gadgets per digitalitzar part de la informació dels i les participants, veure **Taula 8**: Hertfordshire Cohort Study[172], 1946 National Study of Health and Development (NSHD)[173], Southampton Women's Study[174] i la Millennium Cohort Study (MSC6)[175].

Taula 8. Breu descripció, disseny, gadgets i període de les e-cohorts identificades.

e-Cohort	Mostreig	Gadgets	Breu descripció
Hertfordshire Cohort Study Vertical Impact *Comtat de Hertfordshire* (2009-2029)	**No probabilístic**, mostra de conveniència	Acceleròmetre USB i dispositiu GeneActiv	Cohort del comtat anglès de Hertfordshire que estudia com l'activitat física afecta a la salut òssia, la musculatura així com a la capacitat física i l'artrosi. En l'actualitat conta amb una cohort a **224 participants**, tot i que inicialment va arribar a tenir més de 40.000 registres.
1946 MRC National Study of Health and Development (NSDH) *Regne Unit* (1946-2020)	**No probabilístic**, mostra de conveniència	Acceleròmetre, CAPI i examen cognitiu d'Addenbrooke (Ipad – versió 3)	Cohort anglesa que segueix a una mateixa generació des de el 1946 fins a l'actualitat, per veure'n l'evolució de la infantesa fins la vida adulta. Recopilant informació sobre la salut física, mental i social. La base poblacional són persones nascudes en la mateixa setmana de Març de 1946, amb un total inicial de **5.362 participants**.
Southampton Women's Study (SWS) *Southampton* (1998-Actualitat)	**No probabilístic**, mostra de conveniència	Monitor de freqüència cardíaca i acceleròmetre	Cohort a la ciutat anglesa de Southampton, que té per objectiu estudiar el desenvolupament dels nadons des de la gestació i examinar com els factors materns i intrauterins estan relacionats amb els gens i la seva incidència en l'evolució del nadó. La base poblacional de la cohort consta de **12.500 dones**.
Millennium Cohort Study (MCS6) *Regne Unit* (2015-2016)	**Probabilístic**, Estratificat per conglomerats	Acceleròmetre i dietari a través de App/Web.	Cohort anglesa que busca trobar com afecten els factors biològics, socials, econòmics i ambientals al desenvolupament de salut i benestar. La base poblacional són els infants nascuts al 2000. La cohort pot accedir a més de **11.000 nuclis familiars**.

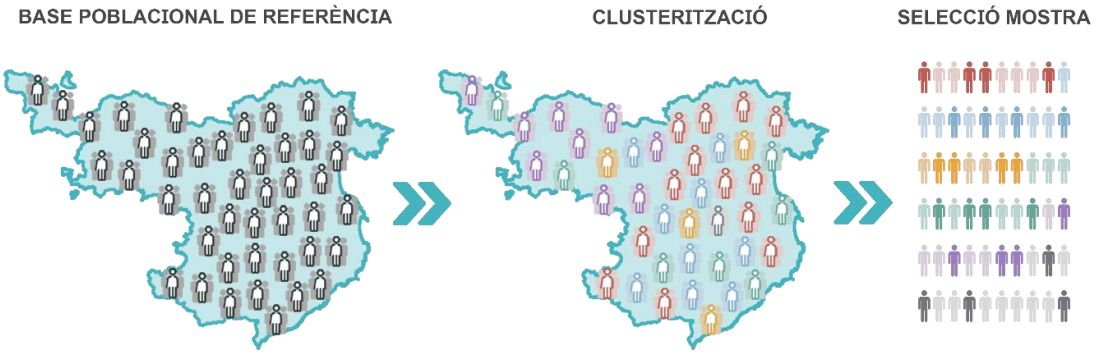
Font: Taula d'elaboració pròpia, a partir de les cerca a la base literària PubMed amb la cerca: "Cohort Studies"[MeSH Major Topic] AND ("new technologies"[Title/Abstract] OR "smartphone app"[Title/Abstract] OR "mobile application"[Title/Abstract] OR "Web-based"[Title/Abstract] OR "mHealth"[Title/Abstract] OR "mobile devices"[Title/Abstract]).

Per estudiar les desigualtats dels 221 municipis de la província de Girona s'ha dissenyat i s'està construint una cohort que començarà al primer trimestre del 2024. S'ha optat per una cohort electrònica pels avantatges de gestió dels i les participants i la facilitat d'obtenció de dades. A més el mecanisme d'obtenció de dades pot ser fàcilment modificat per obtenir nova informació. Tot i que, en la literatura no estan clares totes les limitacions de les e-cohorts, se'n poden

detectar algunes. La cohort de Framingham, demostra que la incorporació de gadgets als estudis és complexa. L'ús que en fan els i les participants decreix progressivament amb el pas de les setmanes[176]. Per altre costat, es coneix poc quina és la saturació dels participants de les cohorts electròniques. A més, les cohorts de participació voluntària com per exemple la Nutrinet-Sante, mostren com els resultats presenten biaixos de resposta en part generats per una participació més elevada de les dones joves amb estudis universitaris.

Un dels avantatges que tenen les cohorts electròniques és que els seus mecanismes augmenten la possibilitat d'arribar a moltes més persones. La majoria d'e-cohorts estan basades en mostres de participació voluntàries. Només l'anglesa Millennium Cohort Study realitza un mostreig probabilístic estratificat. Aquesta estratificació es basa principalment en les característiques clàssiques: edat, sexe i territori. Altres cohorts, que s'han portat a terme a la província, com el MESGI50[177], també realitzen una estratificació segons l'edat i nombre d'habitants de cada municipi.

Figura 12. Procés de clusterització de la província de Girona



Font: Gràfic d'elaboració pròpia

Les cohorts identificades que caracteritzen la mostra d'un territori gran, ho fan pels estratificadors clàssics i apliquen conglomerats de grans zones. La clusterització presentada en l'**Article I**, es basa en una caracterització de diferents elements socials, econòmics, geogràfics i de salut. Buscant una caracterització del territori més enllà de les variables socials tradicionals i així poder capturar la seva heterogeneïtat. Els resultats presentats mostren, una caracterització amb coherència geogràfica i també socioeconòmica que

permeten una captura de mostra estratificada representativa del territori, veure **Figura 12**.

5.2 Clusterització

La clusterització és un element clau, sobretot en aquest tipus de casos, on es busca la representativitat d'un conjunt de municipis tan heterogeni. Dels 221 municipis que conformen la província de Girona, només el 10% superen els 10.000 habitants. Aquesta distribució poblacional té interaccions en totes les variables que s'utilitzen per a la classificació.

Cap de les cohorts anteriorment identificades, veure **Taula 7** i **Taula 8**, utilitzen una caracterització dels seus conglomerats tan extensa. Quan s'usen variables, més enllà de les clàssiques socials o demogràfiques, aquestes presenten problemàtiques de lectura i comparació que acaben afectant l'agrupació. És per això que s'han creat tres bases de dades diferents per veure'n els resultats obtinguts i observar les diferències entre elles.

Els resultats de les clusteritzacions, mostren que és necessari una estandardització prèvia abans de començar a realitzar qualsevol classe de clusterització[178–180]. Treballar amb un conjunt de dades amb soroll, implica dur a terme processos de constrenyiment i posterior depuració.

Tampoc s'ha detectat que cap cohort utilitzi una clusterització de més d'un any. Les cohorts que realitzen clusteritzacions, prenen un any de referència i generen les agrupacions necessàries. La cohort anglesa Millennium Cohort Study, pren de referència les variables socioeconòmiques de les famílies del regne unit de l'any 2000. En l'àmbit local, MESGI50 pren de referència l'any les dades poblacionals de l'any 2012. En canvi, les clusteritzacions presentades en els resultats s'han fet per múltiples anys. Així es pot veure en quines agrupacions, els municipis canvien de clúster i escollir aquella clusterització que sigui més estable, veure **Taula 9**.

Els mecanismes de partició, són els més comuns per realitzar agrupacions de municipis o ciutats[181–185]. Els algoritmes com OPTICS o DBSCAN també

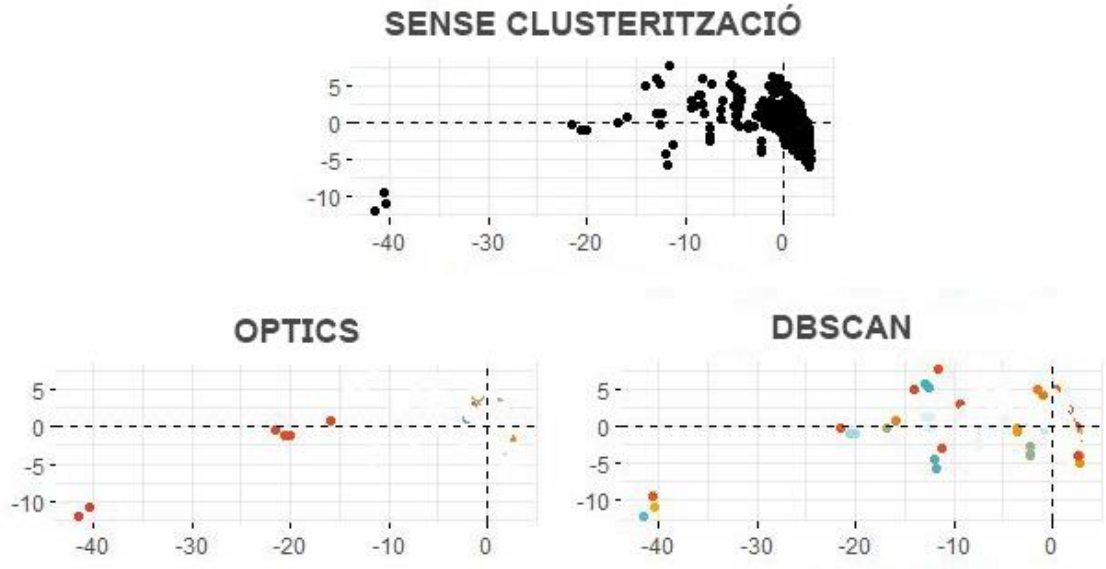
s'utilitzen per dur a terme agrupacions en àrees petites, tot i que no són tan comuns[186]. De forma similar, en l'**Article I** es mostra com els mecanismes de partició presenten validacions òptimes. Per contra, els algoritmes com DBSCAN i OPTICS eliminen els municipis detectats com a soroll i no els clusteritzen, veure **Figura 13**.

Taula 9. Mesurament del nombre de casos que varien de clúster per estudiar la variabilitat dels resultats.

	0 canvis	1 canvi	2 canvis	0 canvis	1 canvi	2 canvis	0 canvis	1 canvi	2 canvis
	Data set : Original			*Data set : Nominal*			*Data set : Z-score*		
K-MEANS	202	19	0	217	4	0	217	4	0
PAM	120	98	3	74	142	5	74	142	5
CLARA	155	65	1	56	162	3	56	162	3
CLARANS	181	40	0	56	162	3	56	162	3
HKMEANS	196	25	0	217	4	0	217	4	0
FUZZY	54	167	0	10	210	1	10	210	1
BIRCH	172	46	3	197	24	0	197	24	0
BICO	172	46	3	196	25	0	173	48	0
EA	172	46	3	197	24	0	173	48	0
DIANA	172	46	3	197	24	0	221	0	0
AGNES	172	46	3	197	24	0	221	0	0

Font: Taula d'elaboració pròpia a partir de l'article Perafita, X.; Saez, M, 2022[187]

Figura 13. Clusterització amb soroll utilitzant OPTICS i DBSCAN

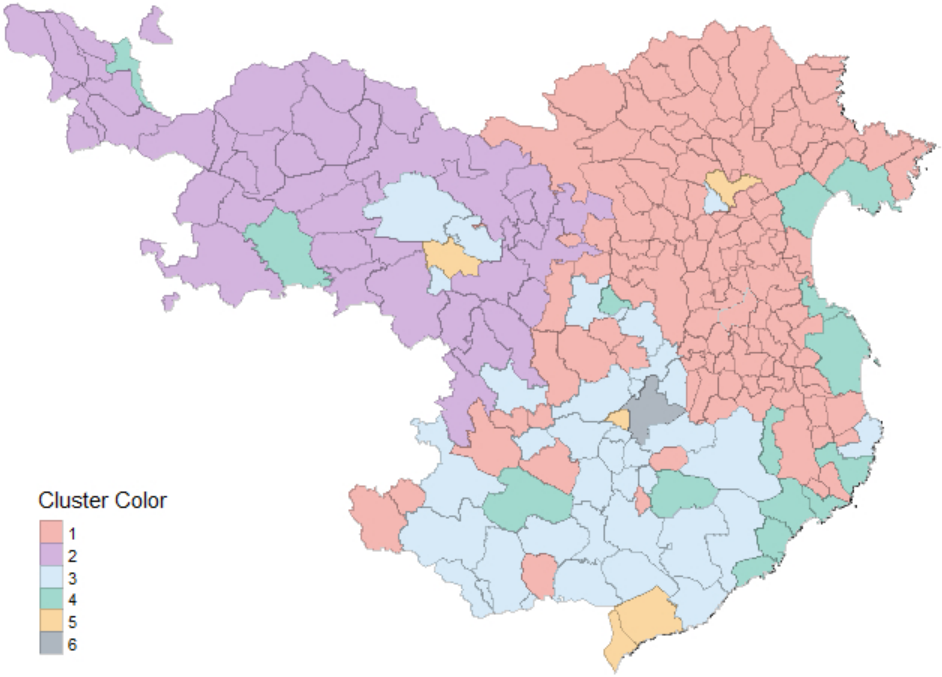


Font: Gràfic d'elaboració pròpia a partir de l'article Perafita, X.; Saez, M, 2022[187]

En conseqüència, s'han descartat per aquest projecte, ja que no es pot eliminar cap municipi. Encara que els algoritmes de partició presenten bons resultats i

són comuns en aquest tipus d'escenaris, el mecanisme que presenta els millors resultats per la seva agrupació és un algoritme jerarquitzat, veure **Figura 14**.

Figura 14. Clusterització final amb el mètode k-mean jeràrquic de la província de Girona



Font: Gràfic d'elaboració pròpia a partir de l'article Perafita, X.; Saez, M, 2022[187]

El resultat final d'aquesta clusterització es tradueix en l'agrupació dels municipis en 6 grups, on Girona que és la capital de la província queda aïllada en un clúster únic. La resta queden agrupats en tres grans clústers i dos restants que capturen realitats de municipis més específiques. En primer lloc trobem els clúster 1 i 2 que són els municipis fronteres, on el clúster 1 recull més municipis muntanyosos. El clúster 3 recull els municipis d'interior. El clúster 4 recull algunes capitals de comarca i els principals municipis de la costa brava. El municipi 5 recull els municipis més densos i amb major activitat econòmica amb la excepció de Girona.

Aquesta nova classificació de municipis és una proposta alternativa a les agrupacions clàssiques que segueixen els límits administratius. Així doncs, la mostra que s'obtindrà permetrà donar un pes diferent a la tipologia de municipis i equilibrar el pes dels municipis més urbanitzats amb els més rurals.

La elecció dels clústers estables pel disseny final del mostreig respon a les limitacions de la lectura de les dades en micro-realitats. La sensibilitat a petits

canvis en un municipi amb poca població fa que sigui necessari treballar amb un conjunt de períodes. Com es pot veure en els resultats del **Article I**, els municipis tendeixen a tenir canvis i no trobar-se en el mateix clúster cada any, més enllà del mecanisme utilitzat per fer l'agrupació.

Alternativament es presenten en l'**Annex IV**: Estudi intra-entre clústers per 10 clústers l'estudi per 10 clústers.

5.2.1 Limitacions i fortaleses

Tot i els múltiples processos per realitzar una agrupació: estandardització de la base de dades, constrenyiment de variables, comparació i validació de les diferents agrupacions, l'ús de la clusterització resultant pel disseny d'una mostra pot presentar certes limitacions.

Encara que s'ha aplicat el mètode Elbow per trobar quin és el nombre de clústers més òptim, pot ser que les divisions no siguin suficients. És plausible pensar que s'hagin d'augmentar el nombre d'agrupacions per trobar més tipologies de municipis i assegurar una heterogeneïtat mínima que representi millor les diversitats demogràfiques, socials, econòmiques i geogràfiques del territori.

Tot i les possibles limitacions, el procés detallat en l'**Article I**, és un mètode més complex que l'utilitzat per la resta de cohorts digitals identificades prèviament. Aquest procés, permet donar una guia de quins algoritmes són més eficaços en aquesta tipologia d'agrupacions. A més, l'agrupació resultant acaba sent més consistent en el temps, ja que considera les clusteritzacions de múltiples anys. Aquesta consistència és interessant sobretot en els estudis longitudinals perquè pot augmentar el soroll de les dades i reduir la validesa i fiabilitat de l'estudi.

També es fàcilment replicable en altres províncies o països. La clusterització serà tant bona com les dades permetin trobar la riquesa i realitat de cada territori, veure **Annex V**: Fonts d'informació identificades en l'Article I. A més, l'enfoc es pot generar a diferents nivell administratius. Si es volgués aplicar a tota una comunitat autònoma es podria seguir treballant per municipis, ja que el mostreig

continuaria sent representatiu per tots ells. Alternativament es podria aplicar el mateix model passant la unitat territorial de municipi a comarques. No obstant, s'hauria de vigilar de no emmascarar les realitats dels municipis petits amb les capitals de comarca o altres municipis, ja que existeixen zones molt polaritzades on la capital o algun municipi concret centralitzen el gran gruix de població i/o activitat econòmica.

5.3 La clusterització i la desigualtat

La clusterització és molt útil per realitzar una primera categorització o estructuració d'un conjunt de dades. Així i tot, les agrupacions per si soles, només poden jerarquitzar i ordenar la informació. Aquestes agrupacions, poden presentar limitacions a l'hora de detectar homogeneïtat en els casos o no poden identificar possibles noves agrupacions que estan presents en formes de subgrups. Aquests subgrups, poden ser possibles grups marginals que no solen ser fàcilment reconeguts pels algoritmes[188].

A més, les agrupacions per si soles, no permeten analitzar o comparar resultats. Requereixen d'un procés de modelització de dades que permeti obtenir uns resultats per fer-ne interpretacions i prediccions.

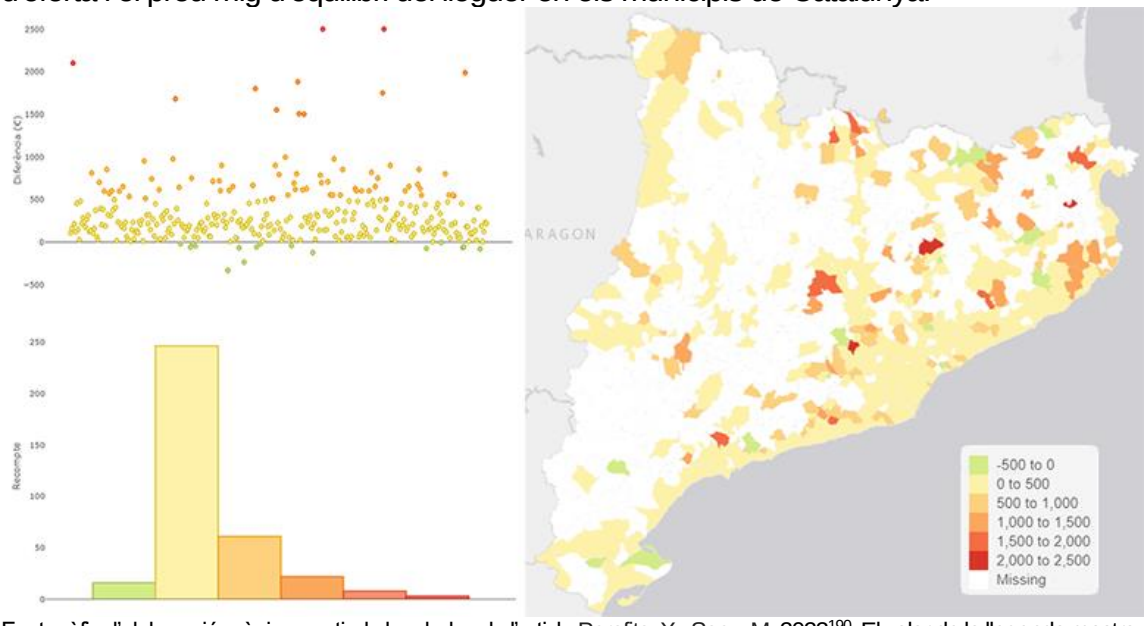
Així doncs, s'ha d'entendre la clusterització com una eina més d'un gran engranatge. És un mecanisme eficaç per categoritzar els individus i veure'n el seu estat i la seva evolució. Permetent realitzar lectures amb major riquesa i profunditat, però per si sola, una agrupació no genera resultats per estudiar la desigualtat.

5.4 L'habitatge com a clau de la desigualtat

L'habitatge és una necessitat bàsica a la qual tothom té dret. A més, acaba actuant com a determinant social, cultural, econòmic, ambiental i de la salut[152,188,189]. Les polítiques d'habitatges socials són extenses, tot i que no sempre acaben afavorint a les famílies destinatàries de l'ajuda[178,188,189]. En l'actualitat existeix evidència sobre les desigualtats vinculades al mercat immobiliari i els entorns de les persones que hi viuen.

És comú trobar que s'estudiï la desigualtat a través del mercat immobiliari, veure **Annex II**. De fet, la majoria d'investigacions es basen en dades de registre on queda reflectit el preu en què s'han llogat o comprat els habitatges. Es pot entendre aquest preu com el punt on l'oferta i la demanda del mercat es creuen, trobant així un punt d'equilibri. Però aquest preu d'equilibri només permet veure a quines llars l'individu o família hi té accés. En l'actualitat les tècniques d'obtenció de dades com el *web scrapping* i l'aparició de portals que venen les seves dades han permès poder estudiar el mercat immobiliari d'una forma més dinàmica. En l'**Article II** s'ha estudiat l'accés a l'habitatge a través de l'oferta. Basar-se en l'oferta del mercat permet detectar, en el moment de fer l'estudi, permet veure a quins entorns queden abocades les famílies amb pocs recursos econòmics.

Figura 15. Mapa, gràfic de dispersió i histograma de la comparació entre el preu mig d'oferta i el preu mig d'equilibri del lloguer en els municipis de Catalunya.



Font gràfic d'elaboració pròpia a partir de les dades de l'article Perafita, X.; Saez, M, 2023[190]. El valor de la llegenda mostra l'excés del preu mig de les llars ofertes al mercat respecte el preu mig de les llars llogades.

Si s'analitzen les dades oficials de lloguer per municipis a Catalunya[191] es pot observar la disparitat entre els preus oferts al mercat i els preus d'equilibri. De tots els municipis on s'ha identificat que existeix com a mínim un habitatge per llogar, el preu mig de les propietats està per sota el preu d'equilibri en setze. La resta estan per sobre, arribant a valors diferencials màxims de 2.500€, veure **Figura 15**. El gràfic de dispersió mostra com la majoria de diferències entre el

preu d'oferta i el preu d'equilibri es troben entre els zero euros i els cinc-cents euros. Reflectint que existeix un conjunt de llars que excloses temporalment del punt d'equilibri.

En gran part, els articles mostren com les característiques de les llars tenen afectes directes sobre les desigualtats en salut. També hi juga un paper important l'entorn de la llar. Tot i que de forma intuïtiva, es pot vincular ràpidament el preu d'un habitatge al nombre de metres quadrats, el nombre d'habitacions o el nombre de lavabos, els resultats presentats en aquesta tesi mostren com l'accés de les famílies per sota el llindar de la pobresa a un habitatge no està determinant per cap d'aquests elements.

En canvi, els nivells de contaminació on està situat l'habitatge, el nivell socioeconòmic o la delinqüència de la zona si que resulta significatiu a l'hora de que les famílies més vulnerables puguin llogar un habitatge. D'aquesta manera, podem entendre que el primer pas que construeix la desigualtat, provinent de les llars, són els entorns en els que poden viure les persones i no l'estat de les cases.

5.5 La perpetuació de la desigualtat

En l'actualitat existeix una àmplia evidència sobre la relació entre l'estat dels habitatges i les desigualtats, sobretot les vinculades a la salut[192–195]. La majoria dels estudis, es presenten o bé com a casos particulars de grans ciutats o bé com a resultats d'enquestes pel que fa a grans regions o de país.

Un dels principals estratificadors de la desigualtat, en totes les seves variants, és la classe social. Existeix una amplia evidència que les classes benestant presenten millors estats de salut que les classes més pobres[196–199]. També es troba relació entre les classes socials més desfavorides i l'exposició a nivells més elevats de contaminació. Els estudis enfocats a relacionar malalties amb nivells de contaminació es basen en l'exposició dels individus. En el **Article II**, s'han canviat els individus per propietats. Mostrant que com més augmenta la contaminació en nivells de PM10, major és la possibilitat de ser llogades.

Els estudis també mostren com les zones més empobrides tendeixen a presentar menor interès per part dels demandants d'habitatges [200–202]. Els resultats segueixen amb aquesta lògica, com major és el percentatge de pobresa extrema de la zona on hi ha un habitatge, major serà la probabilitat de llogar-lo per part de les famílies vulnerables. Històricament, existeix una segregació residencial que divideix la pobresa per zones o barris. Aquestes zones acaben esdevenint comunitats de persones vulnerables i perpetuant la trampa de la pobresa. A més, els resultats bivariants, mostren com existeix una pobresa latent, fins i tot en els barris benestants del territori català.

Figura 16. Mapeig de les ciutats amb estudis identificat que vinculen temàtiques socials, polítiques o ambientals relacionades amb la desigualtat a través de la llar.

Font: Gràfic d'elaboració pròpia a partir de la cerca realitzada en la base de dades acadèmica SCOPUS. S'ha realitzat la següent cerca: (TITLE ("rental hous*") AND TITLE-ABS-KEY ("inequal*")) AND (LIMIT-TO (SUBJAREA , "SOCI") OR LIMIT-TO (SUBJAREA , "ENVI") OR LIMIT-TO (SUBJAREA, "ECON")). De tots els articles detectats s'han seleccionat per els que per contingut aportaven al estudi. Posteriorment s'han afegit 14 articles vinculats.

L'accés a entorns verds també juga un paper clau en la salut de les persones[203–207]. També en la desigualtat. De fet, l'accés a la vegetació és d'alt interès per la majoria de les ciutats rellevants a escala mundial. Els resultats del **Article II** mostren que la proximitat a vegetació, es relaciona amb una reducció de la probabilitat de llogar l'habitatge per part de les famílies més vulnerables. Són pocs els estudis que estudien les desigualtats presentant diferències entre zones urbanes i zones rurals. La majoria se centren en metròpolis o ciutats principals,

veure **Figura 16**. Els que si ho fan, mostren que les tendències entre elles poden ser diferents[208]. Les probabilitats que té una família en risc d'exclusió social de poder llogar un habitatge dins i fora de l'àrea metropolitana són diferents, veure **Figura 17**. Això demostra que les accions o polítiques s'haurien d'aplicar en funció de les àrees on es vol reduir la desigualtat.

Un altre element a tenir en compte és l'impacte que tenen els elements relacionats amb la vegetació com el NDVI proper o verd urbà, ja que tots dos actuen com a barrera d'accés dins de l'àrea metropolitana. Tot i que és el NDVI proper dels habitatges el que té més pes. En canvi, fora de l'àrea metropolitana, el NDVI actua com a facilitador del lloguer. Fora de les zones urbanes, un menor NDVI o verd urbà pot expressar major grau d'urbanització.

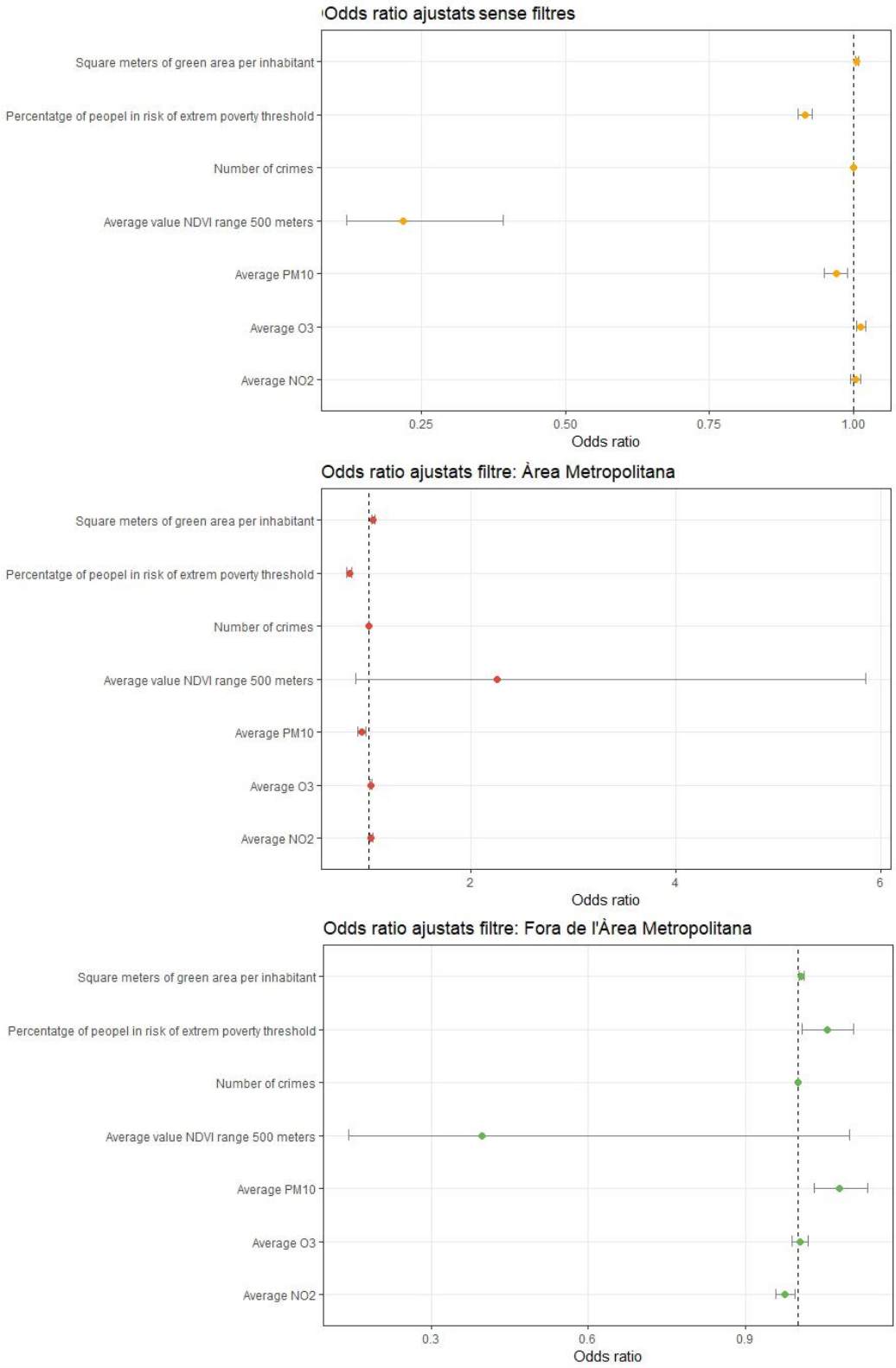
En les zones rurals existeix una segregació de la població de baixos recursos en les zones més precàries. Fet que no succeeix en les zones urbanes, on la probabilitat que una família per sota el llindar de la pobresa pugui llogar augmenta. El mateix fenomen es dona amb les zones més contaminades per PM10. Com més contaminada és la zona on es troba l'habitatge més probable serà llogar-lo, dins de l'àrea metropolitana, en canvi, fora es redueix la possibilitat.

5.5.1 Limitacions i fortaleses

Les desigualtats necessiten entendre l'estat de l'individu, propi i de la seva comunitat. Estudiar la desigualtat a través de les llars, és un mecanisme que posa més èmfasis en l'entorn que genera la desigualtat que en el mateix individu.

Si bé és cert que els resultats presenten tres àmbits a la vegada: socioeconòmic, ambiental i de seguretat enfocats a les famílies en risc d'exclusió social, els resultats que se'n deriven poden ser que no capturin la realitat de les famílies. Aquestes poden desenvolupar les seves activitats en entorns allunyats de les llars.

Figura 17. Associació entre els contaminants atmosfèrics i les variables socioeconòmiques i sanitàries i la possibilitat de no llogar un habitatge per a una família en risc d'exclusió social per zones



Ajustat pel tipus de llar (m2,habitacions,lavabos), variables de tipus social (densitat i població) i variables de la tipologia d'àrees (capital i àrea del districte).
Font: Gràfic d'elaboració pròpia a partir de les dades del article: Perafita, X.; Saez, M, 2023[190].

També és cert que la llar només presenta una foto de tot un collage que representa la desigualtat. La llar és un element més que pot contribuir a millorar la qualitat de vida. Però per si sol, no és un element que garanteixi millor qualitat de vida. Altres factors com l'accés a un sistema eficient de transport públic, una bona qualitat del servei sanitari, un bon sistema educatiu o altres factors poden ajudar a millorar la qualitat de vida i reduir-ne les desigualtats vinculades.

Això implica que, per entendre bé la desigualtat, calen dades de millor qualitat i complexitat que permetin modelar la desigualtat a nivell individual i capturar-ne l'evolució per entendre: què està passant, com està passant i perquè està passant.

No obstant, a l'**Article II** es presenta un model que permet estudiar la determinació social al moment. Al treballar sobre un bé de primera necessitat amb demanda constant, el preu d'oferta reflecteix aquells entorns als quals certs col·lectius no hi tenen accés. L'estudi mostra una forma eficient de detectar patrons que condicionen la vida i la salut de les persones. A més, permet planificar actuacions i dissenyar polítiques a mig i llarg termini que actuïn sobre els entorns que estratifiquen la societat i generen desigualtat.

6 CONCLUSIONS

Resultats sobre la revisió sistemàtica:

- Existeixen múltiples mecanismes economètrics que s'utilitzen en l'actualitat en els entorns Big Data. La majoria d'eines aplicades a intel·ligències artificials, aprenentatges automàtics i similars, es basen en eines d'estadística clàssica. Sovint, és necessari combinar diferents tipus d'algoritmes per poder processar la informació i obtenir-ne resultats rellevants.

Obtenció de dades de la població de Catalunya (**Article I** i **Article II**):

- Existeixen múltiples fonts de dades oficials que permeten descarregar informació. Tanmateix, moltes estan protegides pel secret estadístic. Alternativament, existeixen altres fonts que són generades per tercers. Aquestes fonts no estan controlades, però generen informació de forma activa per tenir impacte sobre algun tipus d'activitat o sector. Amb tècniques d'obtenció de dades tipus *webscrapping* es poden aconseguir i explotar noves fonts de dades que permeten estudiar certs fenòmens amb un altre perspectiva.
- La combinació de múltiples fonts de dades, permet estudiar els patrons de desigualtat amb una major riquesa, però s'ha de vigilar l'autocorrelació i aplicar mecanismes de constrenyiment per controlar la incidència de les variables sobre els estudis.

Ús de mètodes estadístics per interpretar dades i estudiar-ne la desigualtat de la població de Catalunya (**Article I** i **Article II**):

- Les clusteritzacions mostren com hi ha patrons socioeconòmics, culturals, ambientals i geogràfics que permeten dividir els 221 municipis de la província, sent Girona l'únic municipi que té una agrupació per ell sol.
- El procés de clusterització mostra com és necessari aplicar bones pràctiques per poder fer una agrupació consistent de les dades. Aquestes bones pràctiques impliquen: estandardització de les dades, estudi de la

variabilitat de les agrupacions i mecanismes de comparació intra-entre clústers.

- Les característiques de la llar no tenen impacte directe sobre la possibilitat de llogar per part de les famílies de baixos recursos. Aquest fet mostra que la desigualtat no està generada de manera directa per les característiques de la llar, i que per tant, no és un factor determinant per si sol, sinó que són les característiques de l'entorn residencial les que generen i conformen aquesta desigualtat.
- En tot Catalunya, al moment de fer l'estudi, només existeixen 9 llars en les quals una família per sota el llindar de la pobresa pot estalviar prou per a poder sortir d'aquesta situació de vulnerabilitat.
- Les desigualtats presents en les zones urbanes i les zones rurals són diferents, destacant l'impacte que tenen la vegetació i els contaminants en el lloguer en les zones urbanes respecte a les rurals.

El treball dut en la tesi mostra una forma alternativa de categoritzar la població d'un territori. Aquesta categorització permet crear noves agrupacions que representin la realitat d'un territori tenint en compte indicadors d'interès. Aquest enfocament pren especial rellevància a l'hora de realitzar estudis de comportament poblacional, ja que les agrupacions més clàssiques com per exemples les divisions administratives o el recompte poblacional, poden generar males inferències o resultats esbiaixats. També es mostra com es poden trobar fonts on la mateixa població sigui la generadora d'informació per fer seguiment d'un fenomen. L'estudi presentat en l'**Article II** n'és un exemple. L'ús de les xarxes neuronals, en especial les estructures de grafs com a base per l'estudi de les desigualtats és l'evolució natural la tesi. Aquesta tècnica permetria focalitzar els estudis en les connexions entre els casos i permetria millorar els estudis de les desigualtats i aprofundir en les seves causes.

7 BIBLIOGRAFIA

1. MacLeod RM. *The Library of Alexandria : Centre of Learning in the Ancient World*. I.B. Tauris ; In the U.S.A. and Canada distributed by St. Martin's Press; 2000.

2. Google Trends. Accessed May 16, 2023. https://trends.google.com/trends

3. Total data volume worldwide 2010-2025. Statista. Accessed April 20, 2023. https://www.statista.com/statistics/871513/worldwide-data-created/

4. Ojokoh BA, Samuel OW, Omisore OM, et al. Big data, analytics and artificial intelligence for sustainability. *Scientific African*. 2020;9:e00551. doi:10.1016/j.sciaf.2020.e00551

5. Glass DV. John Graunt and His Natural and Political Observations. *Notes and Records of the Royal Society of London*. 1964;19(1):63-100.

6. Snow J. THE MODE OF PROPAGATION OF CHOLERA. *Assoc Med J*. 1856;4(163):135.

7. Smith GD. Commentary: Behind the Broad Street pump: aetiology, epidemiology and prevention of cholera in mid-19th century Britain. *International Journal of Epidemiology*. 2002;31(5):920-932. doi:10.1093/ije/31.5.920

8. Social Security History. Accessed October 22, 2020. https://www.ssa.gov/history/ibm.html

9. Copeland BJ. Colossus: its origins and originators. *IEEE Annals of the History of Computing*. 2004;26(4):38-45. doi:10.1109/MAHC.2004.26

10. Copeland BJ. *Colossus: The Secrets of Bletchley Park's Code-Breaking Computers*. OUP Oxford; 2006.

11. Halevi G, Moed H. The evolution of big data as a research and scientific topic: Overview of the literature. *Research Trends*. 2012;30:3-6.

12. Diebold FX. On the Origin(s) and Development of the Term "Big Data." *SSRN Electronic Journal*. Published online October 3, 2012. doi:10.2139/ssrn.2152421

13. Mashey JR. Big Data and the Next Wave of InfraStress Problems, Solutions, Opportunities. In: *1999 USENIX Annual Technical Conference (USENIX ATC 99)*. USENIX Association; 1999. https://www.usenix.org/conference/1999-usenix-annual-technical-conference/big-data-and-next-wave-infrastress-problems

14. Cox M, Ellsworth D. Managing big data for scientific visualization. In: Vol 97. MRJ/NASA Ames Research Center; 1997:21-38.

15. Laney D. 3D data management: Controlling data volume, velocity and variety. *META group research note*. 2001;6(70):1.

16. Patgiri R, Ahmed A. Big Data: The V's of the Game Changer Paradigm. In: ; 2016. doi:10.1109/HPCC-SmartCity-DSS.2016.0014

17. Jin X, Wah BW, Cheng X, Wang Y. Significance and Challenges of Big Data Research. *Big Data Research*. 2015;2(2):59-64. doi:10.1016/j.bdr.2015.01.006

18. Chen H, Chiang RHL, Storey VC. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*. 2012;36(4):1165-1188. doi:10.2307/41703503

19. Demchenko Y, de Laat C, Membrey P. Defining architecture components of the Big Data Ecosystem. In: *2014 International Conference on Collaboration Technologies and Systems (CTS)*. ; 2014:104-112. doi:10.1109/CTS.2014.6867550

20. Ishwarappa, Anuradha J. A brief introduction on big data 5Vs characteristics and hadoop technology. *Procedia Computer Science*. 2015;48(C):319-324. doi:10.1016/j.procs.2015.04.188

21. Ishwarappa, Anuradha J. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. *Procedia Computer Science*. 2015;48:319-324. doi:https://doi.org/10.1016/j.procs.2015.04.188

22. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*. 2015;35(2):137-144. doi:10.1016/j.ijinfomgt.2014.10.007

23. Kitchin R, McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. 2016;3(1):2053951716631130. doi:10.1177/2053951716631130

24. Samuel AL. Some studies in machine learning using the game of checkers. II-Recent progress. *Annual Review in Automatic Programming*. 1969;6(PART 1):1-36. doi:10.1016/0066-4138(69)90004-4

25. Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*. 2021;76:243-297. doi:10.1016/j.inffus.2021.05.008

26. Junfei Qiu, Youming Sun. A Research on Machine Learning Methods for Big Data Processing. In: *Proceedings of the 4th International Conference on Information Technology and Management Innovation*. Atlantis Press; 2015:920-928. doi:10.2991/icitmi-15.2015.155

27. Varian HR. Big data: New tricks for econometrics. *Journal of Economic Perspectives*. 2014;28(2):3-28. doi:10.1257/jep.28.2.3

28. Ciompa P. *Grundrisse Einer Oekonometrie Und Die Auf Der Nationalökonomie Aufgebaute Natürliche Theorie Der Buchhaltung: Ein Auf Grund Neuer Ökonometrischer Gleichungen Erbrachter Beweis, Dass Alle Heutigen Bilanzen Falsch Dargestellt Werden*. Poeschel; 1910.

29. Ciompa P. *Zarys Ekonometryi i Teorya Buchalteryi*. wydawca nieznany; 1910.

30. Frisch R. Note on the Term “Econometrics.” *Econometrica*. 1936;4(1):95-95. doi:10.2307/1907124

31. Laplace PS. Mémoire sur les probabilités. *Mémoires de l’Académie Royale des sciences de Paris*. 1781;1778:227-332.

32. Laplace PS. *Théorie Analytique Des Probabilités*. Vol 7. Courcier; 1820.

33. Laplace PS. *Essai Philosophique Sur Les Probabilités*. Bachelier; 1825.

34. Bayes Mr, Price Mr. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*. 1763;53:370-418.

35. Fisher RA. *Statistical Methods for Research Workers*. Springer; 1992.

36. Fisher RA. Theory of statistical estimation. In: Vol 22. Cambridge University Press; 1925:700-725.

37. Fisher RA. Design of experiments. *British Medical Journal*. 1936;1(3923):554.

38. Chernozhukov V, Hong H. An MCMC approach to classical estimation. *Journal of Econometrics*. 2003;115(2):293-346. doi:10.1016/S0304-4076(03)00100-3

39. van Ravenzwaaij D, Cassey P, Brown SD. A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*. 2018;25(1):143-154. doi:10.3758/s13423-016-1015-8

40. Scott SL, Varian HR. Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*. 2014;5(1-2):4-23. doi:10.1504/IJMMNO.2014.059942

41. Bühlmann P, van de Geer S. Statistics for big data: A perspective. *Statistics & Probability Letters*. 2018;136:37-41. doi:10.1016/j.spl.2018.02.016

42. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010;72(4):417-473.

43. Shah RD, Samworth RJ. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2013;75(1):55-80.

44. L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM. Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*. 2017;5:7776-7797. doi:10.1109/ACCESS.2017.2696365

45. Barbella M, Tortora G. A semi-automatic data integration process of heterogeneous databases. *Pattern Recognition Letters*. 2023;166:134-142. doi:10.1016/j.patrec.2023.01.007

46. Stańczyk U, Baron G. On heterogeneity or sub-classes aspect in construction of stylometric input datasets. *Procedia Computer Science*. 2022;207:2526-2535. doi:10.1016/j.procs.2022.09.311

47. Iddianozie C, Palmes P. Towards smart sustainable cities: Addressing semantic heterogeneity in Building Management Systems using discriminative models. *Sustainable Cities and Society*. 2020;62:102367. doi:10.1016/j.scs.2020.102367

48. Ali S, Chong I. Semantic Mediation Model to Promote Improved Data Sharing Using Representation Learning in Heterogeneous Healthcare Service Environments. *Applied Sciences*. 2019;9(19):4175. doi:10.3390/app9194175

49. Pineda JM. Modelos predictivos en salud basados en aprendizaje de maquina (machine learning). *Revista Médica Clínica Las Condes*. 2022;33(6):583-590. doi:10.1016/j.rmclc.2022.11.002

50. Goodfellow I, Bengio, Yoshua, Courville, Aaron. *Deep Learning*. MIT Press; 2016.

51. Aggarwal CC. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing; 2018. doi:10.1007/978-3-319-94463-0

52. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*. 1992;46(3):175. doi:10.2307/2685209

53. Silverman BW, Jones MC. E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *International Statistical Review / Revue Internationale de Statistique*. 1989;57(3):233. doi:10.2307/1403796

54. Cover TM, Hart PE. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*. 1967;13(1):21-27. doi:10.1109/TIT.1967.1053964

55. Hellman ME. The Nearest Neighbor Classification Rule with a Reject Option. *IEEE Transactions on Systems Science and Cybernetics*. 1970;6(3):179-185. doi:10.1109/TSSC.1970.300339

56. Dudani SA. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man and Cybernetics*. 1976;SMC-6(4):325-327. doi:10.1109/TSMC.1976.5408784

57. Bailey T, Jain AK. NOTE ON DISTANCE-WEIGHTED k-NEAREST NEIGHBOR RULES. *IEEE Transactions on Systems, Man and Cybernetics*. 1978;SMC-8(4):311-313. doi:10.1109/tsmc.1978.4309958

58. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Taylor & Francis; 1984. https://books.google.es/books?id=JwQx-WOmSyQC

59. Morgan JN, Sonquist JA. Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*. 1963;58(302):415-434. doi:10.2307/2283276

60. Gordon AD, Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. *Biometrics*. 1984;40(3):874. doi:10.2307/2530946

61. Quenouille MH. Approximate Tests of Correlation in Time-Series. *Journal of the Royal Statistical Society Series B (Methodological)*. 1949;11(1):68-84.

62. Jaeckel LA. *The Infinitesimal Jackknife*. Bell Laboratories Memorandum; 1972.

63. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 1979;7(1):1-26.

64. Efron B, Tibshirani R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*. 1986;1(1):54-77. doi:10.1214/ss/1177013817

65. Rubin DB. The Bayesian Bootstrap. *The Annals of Statistics*. 1981;9(1):130-134. doi:10.1214/aos/1176345338

66. Efron B. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*. 1987;82(397):171-185. doi:10.2307/2289144

67. DiCiccio T, Efron B. More Accurate Confidence Intervals in Exponential Families. *Biometrika*. 1992;79(2):245. doi:10.2307/2336835

68. Breiman L. Bagging predictors. *Machine Learning*. 1996;24(2):123-140. doi:10.1007/bf00058655

69. Schapire RE. The strength of weak learnability. *Machine Learning*. 1990;5(2):197-227. doi:10.1007/bf00116037

70. Kearns M. Thoughts on hypothesis boosting. Published online 1988.

71. Kearns M, Valiant L. Cryptographic Limitations on Learning Boolean Formulae and Finite Automata. *Journal of the ACM (JACM)*. 1994;41(1):67-95. doi:10.1145/174644.174647

72. Freund Y. Boosting a weak learning algorithm by majority. *Information and Computation*. 1995;121(2):256-285. doi:10.1006/inco.1995.1136

73. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. 1997;55(1):119-139. doi:10.1006/jcss.1997.1504

74. Friedman JH, Hall P. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*. 2007;137(3):669-683. doi:10.1016/j.jspi.2006.06.002

75. Ho TK. Random decision forests. In: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. Vol 1. IEEE Computer Society; 1995:278-282. doi:10.1109/ICDAR.1995.598994

76. Ho TK. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998;20(8):832-844. doi:10.1109/34.709601

77. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

78. Leamer EE. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley; 1978.

79. Leamer EE. Let's Take the Con Out of Econometrics. *The American Economic Review*. 1983;73(1):31-43.

80. Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society Series A (Statistics in Society)*. 1995;158(3):419-466. doi:10.2307/2983440

81. Hansen HF. Choosing Evaluation Models: A Discussion on Evaluation Design. *Evaluation*. 2005;11(4):447-462. doi:10.1177/1356389005060265

82. Marcellino MG, Stock JH, Watson MW. A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series. *IGIER Working Paper No 285*. Published online April 13, 2005. doi:10.2139/ssrn.687782

83. Koop G, Korobilis D. Model uncertainty in Panel Vector Autoregressive models. *European Economic Review*. 2016;81:115-131. doi:10.1016/j.euroecorev.2015.09.006

84. Pesaran H, Timmermann A. A Recursive Modelling Approach to Predicting UK Stock Returns. *The Economic Journal*. 2000;110(460):159-191.

85. Castle J, Qin X, Reed W. How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms. *University of Canterbury, Department of Economics and Finance, Working Papers in Economics*. Published online 2009.

86. Breiman L. Heuristics of instability and stabilization in model selection. *Annals of Statistics*. 1996;24(6):2350-2383. doi:10.1214/aos/1032181158

87. Tikhonov AN. On the stability of inverse problems. *Proceedings of the USSR Academy of Sciences*. 1943;39:195-198.

88. Hoerl AE. Application of ridge analysis to regression problems. *Chemical Engineering Progress*. 1962;58:54-59.

89. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 1970;12(1):55-67. doi:10.1080/00401706.1970.10488634

90. Frank IE, Friedman JH. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*. 1993;35(2):135. doi:10.2307/1269656

91. Polson NG, Scott JG, Windle J. The Bayesian bridge. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2014;76(4):713-733. doi:10.1111/rssb.12042

92. George EI, McCulloch RE. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*. 1993;88(423):889. doi:10.2307/2290777

93. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1984;PAMI-6(6):721-741. doi:10.1109/TPAMI.1984.4767596

94. Breiman L. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*. 1995;37(4):384. doi:10.2307/1269730

95. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267-288.

96. Santosa F, Symes WW. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*. 1986;7(4):1307-1330. doi:10.1137/0907087

97. Knight K, Fu W. Asymptotics for Lasso-Type Estimators on JSTOR. *The Annals of Statistics*. 2000;28(5):1356-1378.

98. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*. 2001;96(456):1348-1360.

99. Antoniadis A, Fan J. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*. 2001;96(455):939-955.

100. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2005;67(2):301-320.

101. Shen X, Ye J. Adaptive Model Selection. *Journal of the American Statistical Association*. 2002;97(457):210-221.

102. Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *The Annals of Statistics*. 2004;32(2):407-451.

103. Cui W, George EI. Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*. 2008;138(4):888-900. doi:10.1016/j.jspi.2007.02.011

104. Zhang CH. Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *The Annals of Statistics*. 2010;38(2):894-942.

105. Ishwaran H, Rao JS. Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*. 2005;33(2):730-773.

106. Mitchell TJ, Beauchamp JJ. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*. 1988;83(404):1023-1032.

107. Madigan D, Raftery AE. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*. 1994;89(428):1535-1546.

108. Smith SC, Timmermann A, Zhu Y. Variable selection in panel models with breaks. *Journal of Econometrics*. 2019;212(1):323-344. doi:10.1016/j.jeconom.2019.04.033

109. McCray JH. A Quasi-Bayesian Audit Risk Model for Dollar Unit Sampling. *The Accounting Review*. 1984;59(1):35-51.

110. Giron FJ, Rios S. Quasi-Bayesian Behaviour: A more realistic approach to decision making? *Trabajos de Estadistica Y de Investigacion Operativa*. 1980;31(1):17-38. doi:10.1007/BF02888345

111. Weiss AA. Asymptotic Theory for ARCH Models: Estimation and Testing. *Econometric Theory*. 1986;2(1):107-131.

112. Hall P, Yao Q. Inference in ARCH and GARCH Models with Heavy-Tailed Errors. *Econometrica*. 2003;71(1):285-317.

113. Wooldridge J. Quasi-Likelihood Methods for Count Data. *Handbook of Applied Econometrics*. 1997;2.

114. Petrova K, Galvão A, Giraitis L, Kapetanios G. *A Quasi-Bayesian Local Likelihood Method for Modelling Parameter Time Variation in DSGE Models*.; 2005.

115. Koenker R, Bassett G. Regression Quantiles. *Econometrica*. 1978;46(1):33-50.

116. Portnoy S. Asymptotic behavior of regression quantiles in non-stationary, dependent cases. *Journal of Multivariate Analysis*. 1991;38(1):100-113. doi:10.1016/0047-259X(91)90034-Y

117. Pearson K. Contributions to the Mathematical Theory of Evolution. *Journal of the Royal Statistical Society*. 1893;56(4):675-679.

118. Lloyd SP. Least square quantization in PCM. *Bell Telephone Laboratories Paper*. 1957;18.

119. Lloyd SP. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. 1982;28(2):129-137. doi:10.1109/TIT.1982.1056489

120. MacQueen J. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967;Volume 1: Statistics:281-297.

121. Forgy E. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*. 1965;21:768-769.

122. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*. 1979;28(1):100. doi:10.2307/2346830

123. Dunn JC. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*. 1973;3(3):32-57. doi:10.1080/01969727308546046

124. Bezdek J. *Pattern Recognition With Fuzzy Objective Function Algorithms*.; 1981. doi:10.1007/978-1-4757-0450-1

125. Kaufmann L, Rousseeuw P. Clustering by Means of Medoids. *Data Analysis based on the L1-Norm and Related Methods*. Published online January 1, 1987:405-416.

126. Kaufman L, Rousseeuw PJ. Partitioning Around Medoids (Program PAM). In: *Finding Groups in Data*. John Wiley & Sons, Ltd; 1990:68-125. doi:10.1002/9780470316801.ch2

127. Kaufman L, Rousseeuw PJ. *Clustering Large Applications (Program CLARA)*.; 1990. doi:https://doi.org/10.1002/9780470316801.ch3

128. Ng R, Han J. CLARANS: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*. 2002;14:1003-1016. doi:10.1109/TKDE.2002.1033770

129. Sander J, Ester M, Kriegel HP, Xu X. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*. 1998;2(2):169-194. doi:10.1023/A:1009745219419

130. Hinneburg A, Keim DA. A General Approach to Clustering in Large Databases with Noise. *Knowledge and Information Systems*. 2003;5(4):387-415. doi:10.1007/s10115-003-0086-9

131. Ankerst M, Breunig M, Kriegel HP, Sander J. OPTICS: Ordering Points to Identify the Clustering Structure. In: *Sigmod Record*. Vol 28. ; 1999:49-60. doi:10.1145/304182.304187

132. Breunig M, Kriegel HP, Ng R, Sander J. LOF: Identifying Density-Based Local Outliers. In: *ACM Sigmod Record*. Vol 29. ; 2000:93-104. doi:10.1145/342009.335388

133. Ertöz L, Steinbach M, Kumar V. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In: *SIAM ICDM*. ; 2003. doi:10.1137/1.9781611972733.5

134. Florek K, Łukaszewicz J, Perkal J, Steinhaus H, Zubrzycki S. Sur la liaison et la division des points d'un ensemble fini. In: Vol 2. ; 1951:282-285.

135. Kaufman L, Rousseeuw PJ. *Agglomerative Nesting (Program AGNES)*.; 1990. doi:https://doi.org/10.1002/9780470316801.ch5

136. Kaufman L, Rousseeuw PJ. *Divisive Analysis (Program DIANA)*.; 1990. doi:https://doi.org/10.1002/9780470316801.ch6

137. Zhang T, Ramakrishnan R, Livny M. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*. 1997;1(2):141-182. doi:10.1023/A:1009783824328

138. Fichtenberger H, Gillé M, Schmidt M, Schwiegelshohn C, Sohler C. BICO: BIRCH meets coresets for k-means clustering. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 8125 LNCS. Springer, Berlin, Heidelberg; 2013:481-492. doi:10.1007/978-3-642-40450-4_41

139. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*. 1943;5(4):115-133. doi:10.1007/BF02478259

140. Hebb DO. *The Organization of Behavior: A Neuropsychological Theory*. J. Wiley; Chapman & Hall; 1949.

141. Rumelhart DE, McClelland JL. Information Processing in Dynamical Systems: Foundations of Harmony Theory. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press; 1987:194-281. https://ieeexplore.ieee.org/document/6302931

142. Le Cun Y. Learning Process in an Asymmetric Threshold Network. In: *Disordered Systems and Biological Organization*. Springer Berlin Heidelberg; 1986:233-240. doi:10.1007/978-3-642-82657-3_24

143. Hinton G. Deep belief networks. *Scholarpedia*. 2009;4(5):5947. doi:10.4249/scholarpedia.5947

144. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 27*. Published online 2014:2672-2680.

145. Social determinants of health: Key concepts. Accessed April 25, 2023. https://www.who.int/news-room/questions-and-answers/item/social-determinants-of-health-key-concepts

146. Lalonde M. *A New Perspective on the Health of Canadians*. Minister of Supply and Services Canada; 1974. http://www.phac-aspc.gc.ca/ph-sp/pdf/perspect-eng.pdf

147. Dahlgren G, Whitehead M. Policies and strategies to promote social equity in health. Background document to WHO - Strategy paper for Europe. *Arbetsrapport*. Published online December 1991. Accessed April 26, 2023. https://ideas.repec.org//p/hhs/ifswps/2007_014.html

148. Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. *International Review of Psychiatry*. 2014;26(4):392-407. doi:10.3109/09540261.2014.928270

149. Stiefel MC, Straszewski T, Taylor JC, et al. Using the County Health Rankings Framework to Create National Percentile Scores for Health Outcomes and Health Factors. *Perm J*. 2020;25:20.012. doi:10.7812/TPP/20.012

150. Bettencourt-Silva JH, Mulligan N, Sbodio M, et al. Discovering New Social Determinants of Health Concepts from Unstructured Data: Framework and Evaluation. *Stud Health Technol Inform*. 2020;270:173-177. doi:10.3233/SHTI200145

151. Duh-Leong C, Dreyer BP, Huang TTK, et al. Social Capital as a Positive Social Determinant of Health: A Narrative Review. *Acad Pediatr*. 2021;21(4):594-599. doi:10.1016/j.acap.2020.09.013

152. Marí-Dell'Olmo M, Oliveras L, Barón-Miras LE, et al. Climate Change and Health in Urban Areas with a Mediterranean Climate: A Conceptual Framework with a Social and Climate Justice Approach. *Int J Environ Res Public Health*. 2022;19(19):12764. doi:10.3390/ijerph191912764

153. Rhee TG, Marottoli RA, Cooney Jr LM, Fortinsky RH. Associations of Social and Behavioral Determinants of Health Index with Self-Rated Health, Functional Limitations, and Health Services Use in Older Adults. *Journal of the*

American Geriatrics Society. 2020;68(8):1731-1738. doi:https://doi.org/10.1111/jgs.16429

154. McQueen DV. Three challenges for the social determinants of health pursuit. *Int J Public Health*. 2009;54(1):1-2. doi:10.1007/s00038-008-8167-x

155. Fink DS, Keyes KM, Cerdá M. Social Determinants of Population Health: A Systems Sciences Approach. *Curr Epidemiol Rep*. 2016;3(1):98-105. doi:10.1007/s40471-016-0066-8

156. Frank J, Abel T, Campostrini S, Cook S, Lin VK, McQueen DV. The Social Determinants of Health: Time to Re-Think? *Int J Environ Res Public Health*. 2020;17(16):5856. doi:10.3390/ijerph17165856

157. Walton AL. The Limits of 'Social Determinants of Health' Language. *AJN The American Journal of Nursing*. 2023;123(1):11. doi:10.1097/01.NAJ.0000911484.04250.06

158. Weiland S, Hickmann T, Lederer M, Marquardt J, Schwindenhammer S. The 2030 Agenda for Sustainable Development: Transformative Change through the Sustainable Development Goals? *Politics and Governance*. 2021;9(1):90-95. doi:10.17645/pag.v9i1.4191

159. Comisión para Reducir las Desigualdades Sociales en Salud en España. Propuesta de políticas e intervenciones para reducir las desigualdades sociales en salud en España. *Gaceta Sanitaria*. 2012;26(2):182-189. doi:10.1016/j.gaceta.2011.07.024

160. Hercberg S, Castetbon K, Czernichow S, et al. The Nutrinet-Santé Study: a web-based prospective study on the relationship between nutrition and health and determinants of dietary patterns and nutritional status. *BMC Public Health*. 2010;10(1):242. doi:10.1186/1471-2458-10-242

161. Tran VT, Riveros C, Péan C, Czarnobroda A, Ravaud P. Patients' perspective on how to improve the care of people with chronic conditions in France: a citizen science study within the ComPaRe e-cohort. *BMJ Qual Saf*. 2019;28(11):875-886. doi:10.1136/bmjqs-2018-008593

162. Lyons J, Akbari A, Agrawal U, et al. Protocol for the development of the Wales Multimorbidity e-Cohort (WMC): data sources and methods to construct a population-based research platform to investigate multimorbidity. *BMJ Open*. 2021;11(1):e047101. doi:10.1136/bmjopen-2020-047101

163. Sullivan A, Brown M, Hamer M, Ploubidis GB. Cohort Profile Update: The 1970 British Cohort Study (BCS70). *International Journal of Epidemiology*. Published online July 18, 2022:dyac148. doi:10.1093/ije/dyac148

164. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol*. 2013;42(1):97-110. doi:10.1093/ije/dys066

165. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*. 2014;43(6):1719-1725. doi:10.1093/ije/dyu001

166. Schnier C, Wilkinson T, Akbari A, et al. The Secure Anonymised Information Linkage databank Dementia e-cohort (SAIL-DeC). *Int J Popul Data Sci*. 5(1):1121. doi:10.23889/ijpds.v5i1.1121

167. Toledano MB, Smith RB, Brook JP, Douglass M, Elliott P. How to Establish and Follow up a Large Prospective Cohort Study in the 21st Century--Lessons from UK COSMOS. *PLoS One*. 2015;10(7):e0131521. doi:10.1371/journal.pone.0131521

168. McManus DD, Trinquart L, Benjamin EJ, et al. Design and Preliminary Findings From a New Electronic Cohort Embedded in the Framingham Heart Study. *J Med Internet Res*. 2019;21(3):e12143. doi:10.2196/12143

169. Merino J, Joshi AD, Nguyen LH, et al. Diet quality and risk and severity of COVID-19: a prospective cohort study. *Gut*. 2021;70(11):2096-2104. doi:10.1136/gutjnl-2021-325353

170. Tessier AJ, Moyen A, Lawson C, et al. Lifestyle Behavior Changes and Associated Risk Factors During the COVID-19 Pandemic: Results from the Canadian COVIDiet Online Cohort Study. *JMIR Public Health Surveill*. 2023;9:e43786. doi:10.2196/43786

171. Kershaw KN, Liu K, Goff DC, et al. Description and initial evaluation of incorporating electronic follow-up of study participants in a longstanding multisite cohort study. *BMC Medical Research Methodology*. 2016;16(1):125. doi:10.1186/s12874-016-0226-z

172. Syddall HE, Simmonds SJ, Carter SA, Robinson SM, Dennison EM, Cooper C. The Hertfordshire Cohort Study: an overview. *F1000Res*. 2019;8:82. doi:10.12688/f1000research.17457.1

173. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology*. 2006;35(1):49-54. doi:10.1093/ije/dyi201

174. Inskip HM, Godfrey KM, Robinson SM, Law CM, Barker DJ, Cooper C. Cohort Profile: The Southampton Women's Survey. *Int J Epidemiol*. 2006;35(1):42-48. doi:10.1093/ije/dyi202

175. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*. 2014;43(6):1719-1725. doi:10.1093/ije/dyu001

176. Pathiravasan CH, Zhang Y, Wang X, et al. Factors associated with long-term use of digital devices in the electronic Framingham Heart Study. *npj Digit Med*. 2022;5(1):1-11. doi:10.1038/s41746-022-00735-1

177. Corominas Barnadas JM, López-Pousa S, Vilalta-Franch J, et al. Estudio MESGI50: descripción de una cohorte sobre la madurez y el envejecimiento satisfactorio. *Gaceta Sanitaria*. 2017;31(6):511-517. doi:10.1016/j.gaceta.2016.07.017

178. Luiz O do C, Heimann LS, Boaretto RC, et al. Differences in living conditions and health between cities: construction of a composite indicator. *Rev Saúde Pública*. 2009;43:115-122. doi:10.1590/S0034-89102009000100015

179. Pinheiro Junior RVB, Carneiro Junior N, Sala A, et al. Primary health care performance according to clusters of convergent municipalities in the state of São Paulo. *Rev bras epidemiol*. 2022;25:e220017. doi:10.1590/1980-549720220017

180. Baxter LK, Sacks JD. Clustering cities with similar fine particulate matter exposure characteristics based on residential infiltration and in-vehicle commuting factors. *Science of The Total Environment*. 2014;470-471:631-638. doi:10.1016/j.scitotenv.2013.10.019

181. Feng W, Lischko A, Martin EG, et al. Who Are the Local Policy Innovators? Cluster Analysis of Municipal Tobacco Control Policies in Massachusetts. *J Public Health Manag Pract*. 2023;29(2):151-161. doi:10.1097/PHH.0000000000001649

182. Baxter LK, Sacks JD. Clustering cities with similar fine particulate matter exposure characteristics based on residential infiltration and in-vehicle commuting factors. *Science of The Total Environment*. 2014;470-471:631-638. doi:10.1016/j.scitotenv.2013.10.019

183. Caparrós Martínez JL, Milán García J, Rueda López N, de Pablo Valenciano J. Mapping green infrastructure and socioeconomic indicators as a public management tool: the case of the municipalities of Andalusia (Spain). *Environ Sci Eur*. 2020;32(1):144. doi:10.1186/s12302-020-00418-2

184. Du X, Niu D, Chen Y, Wang X, Bi Z. City classification for municipal solid waste prediction in mainland China based on K-means clustering. *Waste Manag*. 2022;144:445-453. doi:10.1016/j.wasman.2022.04.024

185. YoshimiTanaka O, Drumond Júnior M, Cristo EB, Spedo SM, Pinto NR da S. Uso da análise de clusters como ferramenta de apoio à gestão no SUS. *Saude soc*. 2015;24:34-45. doi:10.1590/S0104-12902015000100003

186. de Domingo M, Ortigosa N, Sevilla J, Roger S. Cluster-Based Relocation of Stations for Efficient Forest Fire Management in the Province of Valencia (Spain). *Sensors (Basel)*. 2021;21(3):797. doi:10.3390/s21030797

187. Perafita X, Saez M. Clustering of Small Territories Based on Axes of Inequality. *International Journal of Environmental Research and Public Health*. 2022;19(6):3359.

188. Theories for social epidemiology in the 21st century: an ecosocial perspective | International Journal of Epidemiology | Oxford Academic. Accessed May 10, 2023. https://academic.oup.com/ije/article/30/4/668/705885

189. Financiarización de la vivienda para alquiler y la precarización de las familias de bajos ingresos en Medellín (Colombia) | Boletín de la Asociación de Geógrafos Españoles. Accessed May 10, 2023. https://bage.age-geografia.es/ojs/index.php/bage/article/view/3319

190. Perafita X, Saez M. Housing Supply and How It Is Related to Social Inequalities—Air Pollution, Green Spaces, Crime Levels, and Poor Areas—In Catalonia. *International Journal of Environmental Research and Public Health*. 2023;20(8):5578. doi:10.3390/ijerph20085578

191. Departament de Territori. Preu mitjà del lloguer d'habitatges per municipi | Dades obertes de Catalunya. Accessed March 22, 2023. https://analisi.transparenciacatalunya.cat/Habitatge/Preu-mitj-del-lloguer-d-habitatges-per-municipi/qww9-bvhh

192. Chen B. Coincided disparity between housing price and health outcome. *The Lancet Regional Health – Europe*. 2023;27. doi:10.1016/j.lanepe.2023.100593

193. Ayala L, Bárcena-Martín E, Cantó O, Navarro C. COVID-19 lockdown and housing deprivation across European countries. *Social Science & Medicine*. 2022;298:114839. doi:10.1016/j.socscimed.2022.114839

194. Otavova M, Faes C, Bouland C, et al. Inequalities in mortality associated with housing conditions in Belgium between 1991 and 2020. *BMC Public Health*. 2022;22(1):2397. doi:10.1186/s12889-022-14819-w

195. Marí-Dell'Olmo M, Novoa AM, Camprubí L, et al. Housing Policies and Health Inequalities. *Int J Health Serv*. 2017;47(2):207-232. doi:10.1177/0020731416684292

196. Eisenberg-Guyot J, Blaikie K, Andrea SB, et al. A tutorial on a marginal structural modeling approach to mediation analysis in occupational health research: Investigating education, employment quality, and mortality. *Am J Ind Med*. 2023;66(6):472-483. doi:10.1002/ajim.23471

197. European Parliament, Directorate-General for Parliamentary Research Services, Scholz N. *Addressing Health Inequalities in the European Union : Concepts, Action, State of Play : In-Depth Analysis*. Publications Office; 2020. doi:10.2861/567478

198. Lago S, Cantarero D, Rivera B, et al. Socioeconomic status, health inequalities and non-communicable diseases: a systematic review. *Z Gesundh Wiss*. 2018;26(1):1-14. doi:10.1007/s10389-017-0850-z

199. Moor I, Spallek J, Richter M. Explaining socioeconomic inequalities in self-rated health: a systematic review of the relative contribution of material, psychosocial and behavioural factors. *J Epidemiol Community Health*. 2017;71(6):565-575. doi:10.1136/jech-2016-207589

200. Multidimensional Discrimination in the Online Rental Housing Market: Implications for Families With Young Children: Housing Policy Debate: Vol 0, No 0. Accessed May 16, 2023. https://www.tandfonline.com/doi/abs/10.1080/10511482.2021.2010118?journalCode=rhpd20

201. Martiniello B, Verhaeghe PP. Does the neighbourhood of the dwelling and the real estate agency matter? Geographical differences in ethnic discrimination on the rental housing market. *Urban Studies*. 2022;59(15):3201-3221. doi:10.1177/00420980221086502

202. Searching for housing in the digital age: Neighborhood representation on internet rental housing platforms across space, platform, and metropolitan segregation - Chris Hess, Arthur Acolin, Rebecca Walter, Ian Kennedy, Sarah Chasins, Kyle Crowder, 2021. Accessed May 16, 2023. https://journals.sagepub.com/doi/abs/10.1177/0308518X211034177

203. Campagna G. Linking crowding, housing inadequacy, and perceived housing stress. *Journal of Environmental Psychology*. 2016;45:252-266. doi:10.1016/J.JENVP.2016.01.002

204. Maryanti M, Khadijah H, Uzair A, Rahman M. The urban green space provision using the standards approach: issues and challenges of its implementation in Malaysia. In: ; 2016:369-379. doi:10.2495/SDP160311

205. Roe JJ, Aspinall PA, Ward Thompson C. Coping with Stress in Deprived Urban Neighborhoods: What Is the Role of Green Space According to Life Stage? *Frontiers in Psychology*. 2017;0(OCT):1760. doi:10.3389/FPSYG.2017.01760

206. Ulmer JM, Wolf KL, Backman DR, et al. Multiple health benefits of urban tree canopy: The mounting evidence for a green prescription. *Health & place*. 2016;42:54-62. doi:10.1016/J.HEALTHPLACE.2016.08.011

207. Wolch JR, Byrne J, Newell JP. Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough.' *Landscape and Urban Planning*. 2014;125:234-244. doi:10.1016/J.LANDURBPLAN.2014.01.017

208. Kim SR, Yoo JW. A Study on the Design Direction of Public Rental House through 'Small-Scale Housing Improvement Project'-Focusing on the Works were Selected for the Final of the LH Housing Design Award-. *Journal of the Architectural Institute of Korea*. 2022;38(10):49-60. doi:10.5659/JAIK.2022.38.10.49

8 ANNEX

8.1 Annex I: Altres publicacions relacionades durant el període de la tesi

1. Batlle Amat P, Lazaro-Lasheras L, Oliveras S, Perafita X, Tarrés A, Vilà A. Promoting equity through monitoring inequalities in the semi-rural region of Girona. *European Journal of Public Health, 30 (Supplement_5)* 2020. doi: https://doi.org/10.1093/eurpub/ckaa166.306
2. Batlle Amat P, Vila A, Lazaro-Lasheras L, Tarrés A, Perafita X, Pou Marti N, Juvinyà D, Oliveras Casadella S, Pujol Fuster J, Siches T, Canales M, Blázquez J. Plan municipal de salud, bienestar y desarrollo sostenible en una comunidad rural: Práctica a nivel micro de promoción de la salud y la equidad a través de los datos del Observatorio en desigualdades sociales y de la salud. Presentat a: *X Congreso Internacional de Salud, Bienestar y Sociedad*; September 3, 2020; París, França. Internacional. https://cgscholar.com/cg_event/events/Wes20/proposal/50196
3. Lazaro-Lasheras L, Batlle Amat P, Oliveras Casadella S, Perafita X, Tarrés A, Vila A. Promoviendo la equidad a partir de la monitorización de las desigualdades en la región de Girona: Diagnóstico participado de necesidad de datos sobre desigualdades para el fomento de la equidad en áreas rurales. Presentat a: *X Congreso Internacional de Salud, Bienestar y Sociedad*; September 3, 2020; París, França. Internacional. https://cgscholar.com/cg_event/events/Wes20/proposal/ 50173
4. Perafita X, Tarrés A, Batlle Amat P, Lazaro-Lasheras L, Martin B, Ruiz N, Vilà A. Metodologia de treball de les dades sobre cobertures del sòl per la realització d'indicadors del àmbit Medi i entorn de l'[O]bservatori. Organisme de Salut Pública de la Diputació de Girona (Dipsalut); 2021. Disponible a: https://observatori.dipsalut.cat/graf_indicadors/medi_ambient_i_salut/medi_i_entorn/cobertures_usos_sol/metodologia-cobertura-sols.pdf
5. Sánchez J.M, Lazaro-Lasheras L, Tarrés A, Vilà A, Batlle Amat P, Perafita X, Ruiz N, Trias J. Procés participatiu per a la identificació de necessitats als municipis en matèria d'indicadors de salut i desigualtat social. Organisme de

Salut Pública de la Diputació de Girona (Dipsalut); 2021. Disponible a: https://observatori.dipsalut.cat/storage/755/Dipsalut_Informe_Final_Proces_Participatiu_Obs.pdf

6. Perafita X, Lazaro-Lasheras L, Vilà A, Batlle Amat P, Martin B, Ruiz N, Tarrés A. Metodologia de treball de les bases de dades Mortalitat de l'Institut Nacional d'Estadística (INE). Unificació per la creació dels indicadors de salut de l'[O]bservatori de Desigualtats Socials i de Salut. Organisme de Salut Pública de la Diputació de Girona (Dipsalut); 2021. Disponible a: https://observatori.dipsalut.cat/graf_indicadors/estat_de_salut/mortalitat/METODOLOGIA_MORTALITAT.pdf
7. Tarrés A, Lazaro-Lasheras L, Battle Amat P, Vila A, Perafita X, Oliveras S. Promoting Equity through Monitoring Inequalities in the Semi-rural Region of Girona. Participatory Process to Identify Municipalities' Needs for Data and Information. Presentat a: *11th IUHPE European Conference on Health Promotion*; Abril 21, 2021, Girona, Espanya.
8. Daponte-Codina A, Cabrera-León A, Mateo-Rodríguez I, Campoy F, Perafita X, Sáez M, Barceló MA. Atlas de los determinantes sociales de la salud en España. Evolución y variabilidad entre Comunidades Autónomas. Comunidades Autónomas. *Granada: Escuela Andaluza de Salud Pública*; 2022. Disponible a: http://www.easp.es/atlasdss/
9. Perafita X; Battle Amat P; Vila A; Tarrés A. Observatorio sobre determinantes socailes y desigualdades en salud y benestar. Presentat a: *24TH IUPHE World Conference on Health Promotion*; Maig 19, 2022 Montréal, Canada. Internacional.
10. Cabrera-León A, Saez M, Campoy F, Perafita X, Mateo I, Barceló MA, Daponte-Codina A. Atlas de determinantes sociales de la salud en España: evolución y variabilidad entre las comunidades autónomas. Presentat a: *XL Reunión Anual de la Sociedad Española de Epidemiología (SEE) y XVII Congresso da Associação Portuguesa de Epidemiologia (APE)*; September 2, 2022 Donostia-San Sebastián, Spain. Internacional.
11. Perafita X, Saez M, Barceló M.A, Tarrés A, Pou Marti N, Battle Amat P, Martin B, Ruiz N, Vilà A. Contaminants i desigualtats - C[O]NTAMINANTS. Girona:

Organisme Autònom de Salut Pública de la Diputació de Girona; 2023. Disponible a: https://observatori.shinyapps.io/ATMOSFERA_PM10/

12. Perafita X, Saez M, Barceló M.A, Tarrés A, Pou Marti N, Ruiz N, Vilà A, Battle Amat P. IndiMuniDem (IMD). Girona: Organisme Autònom de Salut Pública de la Diputació de Girona; 2023. Disponible a: https://observatori.shinyapps.io/APP_DEMOGRAFIA/
13. Perafita X, Saez M, Barceló M.A, Tarrés A, Pou Marti N, Battle Amat P, Ruiz N, Vilà A. GirTrans (GiT). Girona: Organisme Autònom de Salut Pública de la Diputació de Girona; 2023. Disponible a: https://observatori.shinyapps.io/APP_TRANSIT/
14. Moreno-Vásquez M, Perafita X, Saez M, Barceló MA. Spatiotemporal variability in socioeconomic inequalities in vaccination against COVID-19 in Catalonia. Presentat a: *XLI Congreso de la Sociedad Española de Epidemiología (SEE) y XVIII Congresso da Associaição Portuguesa de Epidemiologia (APE).* Oporto, Portugal, 5-8 de setembre de 2023. Internacional.
15. Daponte A, Perafita X, Campoy F, Saez M, Mateo I, Barceló MA, Cabrera A, Sánchez-Cantalejo C. El atlas de los determinantes sociales de la salud en España y sus CCAA 2022. Presentat a: *XLI Congreso de la Sociedad Española de Epidemiología (SEE) y XVIII Congresso da Associaição Portuguesa de Epidemiologia (APE)*. Oporto, Portugal, 5-8 de setembre de 2023. Internacional.
16. Barceló MA, Moreno MA, Perafita X, Saez M. Spatiotemporal variability in socioeconomic and environmental inequalities in vaccination against COVID-19 in Catalonia, Spain. Presentat a: *Spatial Statistics 2023. Climate and the Environment*. University of Colorado, Boulder, Estats Units, 18 a 21 de juliol de 2023. Internacional.
17. Moreno M, Barceló MA, Perafita X, Saez M. Spatiotemporal variability in socioeconomic inequalities in vaccination against COVID-19 in Catalonia, Spain. Presentat a: *METMA LATAM*. Quito, Equador, 26 a 28 de juny de 2023. Internacional.

8.2 Annex II: Recull d'articles vinculats al habitatge i les desigualtats.

Taula 10. Revisió literària sobre els articles basats en el mercat immobiliària i les desigualtats.

Article	Any	Tipus	Països	Font	Zona	Descricpcio
Financialization of rental housing and the precariousness of low-income families in Medellín (Colombia)	2023	Polític	**Colòmbia**, Medellín	-	Urbà	Analitzar la **financerització de l'habitatge a través del lloguer a Medellín** i les seves **conseqüències entre les famílies de renda baixa**. La financerització de l'habitatge mitjançant el lloguer s'evidencia amb la **compra d'Habitatge d'Interès Social**, promoguda per l'Estat, **per part de famílies amb excedents de capital**.
Does the neighbourhood of the dwelling and the real estate agency matter? Geographical differences in ethnic discrimination on the rental housing market	2022	Polític	**Bèlgica**, Anvers	-	Urbà	Investigar en quina mesura **la composició ètnica i socioeconòmica del veïnat està relacionada amb els nivells de discriminació al mercat de lloguer** d'habitatges i com això es **relaciona amb les teories de discriminació ètnica**. Una composició socioeconòmica més baixa es relaciona amb unes taxes d'invitació generals més baixes.
Searching for housing in the digital age: Neighborhood representation on internet rental housing platforms across space, platform, and metropolitan segregation	2021	Polític	**Estats Units**	On-line	Urbà	Aquest article estudia la **comprensió de com les plataformes en línia afecten la dinàmica de cerca d'habitatges a través dels seus biaixos i segmentació**, i destaca el potencial i els límits de l'ús de les dades disponibles en aquestes plataformes per produir estimacions de lloguer de zones petites.
Multidimensional Discrimination in the Online Rental Housing Market: Implications for Families With Young Children	2022	Polític	**Estats Units**	On-line	Urbà	Examinen la **discriminació en l'habitatge** que viuen les persones que pertanyen a **múltiples grups desfavorits**. Trobant un **patró dinàmic de discriminació multidimensional** i donen suport a arguments per a un **enfocament interseccional per comprendre i combatre la desigualtat**.
The unequal availability of rental housing information across neighborhoods	2021	Social	**Estats Units**	On-line i enquesta	Urbà	L'estudi estudia **com varia la informació disponible dels lloguers segons la tipologia de barris**. L'estudi es basa en la plataforma Craiglist. Els resultats mostren que **l'accés a informació varia segons la composició socioeconòmica i ètnica dels barris**.

Online rental housing market representation and the digital reproduction of urban inequality	2020	Polític	**Estats Units**	On-line	Urbà	Estudi de **com el lloguer d'habitatges pot reduir les desigualtats segons raça i classe social**. Basat en anuncis de **Craiglist**. L'estudi mostra com el **mercat d'habitatge està segregats digitalment generant mobilitzacions residencials**. Les tecnologies poden **reforça la segregació racial**.
Migration and inequality in rental housing: Affordability stress in the Chinese cities	2020	Social	**Xina**	Enquesta	Urbà	L'article analitza com **els treballadors immigrants a la Xina** estan sotmesos a la **desigualtat en accessibilitat i assequibilitat al lloguer**. L'estudi es basa en la tensió de la renda. Els resultats mostren una **bretxa creixent entre el nord i el sud del país**, on en les **zones més urbanitzades** hi ha **major presència d'immigrants**.
Analyzing the private rental housing market in Shanghai with open data	2019	Social	**Xina**, Xangai	On-line	Urbà	L'estudi estudia els **preus de lloguer a Xangai**. Els resultats mostren que la **concertació de lloguer més cars es troben al centre de la ciutat**, **mobilitzant a les famílies pobres**. El **preu del lloguer, està influenciat per oferta laboral, salaris, serveis públic entre d'altres**.
Closed doors everywhere? A meta-analysis of field experiments on ethnic discrimination in rental housing markets	2019	Polític	**Estats Units, Canadà** i **Europa**	-	Urbà	L'article consisteix en un **meta-anàlisis sobre la discriminació ètnica en el mercat de lloguer** d'habitatges. Els resultats mostren que **la discriminació disminueix en el temps**, relacionant la **falta d'informació de la classe social dels sol·licitants**.
India's residential rental housing	2017	Social	**Índia**	Enquesta	Urbà	L'article **estudia l'estat del lloguer a l'Índia**, a través del **cens i enquestes**. Els resultats mostren una **escassetat d'habitatges i falta d'accessibilitat econòmica**. També es destaquen l'augment del nombre de cases buides, trobant desigualtats entre les necessitats dels habitatges i el nombre de cases buides.
The financialisation of rental housing: A comparative analysis of New York City and Berlin	2016	Social	**Estats Units**, Nova York i **Europa**, Berlín	-	Urbà	L'article compara com **les inversions privades han impactat sobre el mercat de lloguer** a **Nova York** i **Berlín**. Es mostra com **la financerització** ha generat un **augment de la desigualtat d'accés al habitatge** i ha configurat **espais abandonats**. Proposa proteccions al lloguer.

Forced Displacement From Rental Housing: Prevalence and Neighborhood Consequences	2015	Polític	**Estats Units**, Milwaukee	Enquesta	Urbà	L'estudi es basa en una **enquesta de llogaters a Milwaukee** i n'estudia el seu **desplaçament forçós**. Els resultats mostren que **un de cada vuit inquilins** experiments una **mobilització forçosa a zones més precàries i amb majors índex de criminalitat**.
People like us? Social status, social inequality and perceptions of public rental housing	2014	Social	**Xina**, Hong Kong	Enquesta	Urbà	L'article estudia com les **polítiques neoliberals han afectat al mercat immobiliari**. Es basa en un enquesta que examina l'experiència de **3.000 llogaters** a **Hong Kong** i com els hi **afecta la percepció del estatus social i igualtat** a l'hora de pagar un lloguer.
Fair and affordable? racial and ethnic segregation and inequality in New York city rental housing	2011	Social	**Estats Units**, Nova York	Enquesta	Urbà	L'article **analitza l'equitat i desigualtat en les polítiques d'habitatge assequibles a Nova York**. Els resultats mostren que **existeix segregació racial i ètnica**. També mostren com els **habitatges mixtos funcionen millor que els habitatges públics**.
Air Pollution, Social Deprivation, and Mortality. A Multilevel Cohort Study	2007	Ambiental	**Noruega**, Oslo	Enquesta	Urbà	L'estudi examina **com afecta la contaminació del aire a diferents zones de Oslo**. L'estudi es condueix a partir de **registres demogràfics i de contaminació**. Els resultats mostren que les **zones més pobresa presenten nivells més alts de contaminació**.
Environmental equity, air quality, socioeconomic status, and respiratory health: a linkage analysis of routine data from the Health Survey for England	2005	Ambiental	**Regne Unit**	Enquesta	Rural i urbà	L'estudi examina **l'impacte de la contaminació** del **Regne unit** en **relació la salut i l'estatus socioeconòmic**. Els resultats mostra que les **persones en àrees mes desafavorides** tenen **més risc de mort prematura** relacionada amb la contaminació del aire. Es proposen millores polítiques tant socials com de qualitat del aire.
Which communities have better accessibility to green space? An investigation into environmental inequality using big data	2020	Ambiental	**Xina**, Xangai	On-line	Urbà	L'estudi analitza a **Xangai l'accessibilitat als espais verds** segons **estatus econòmic i preu de l'habitatge**. Els resultats mostren **desigualtats respecte els preus dels habitatges**. Proposen millores en les planificació per evitar desigualtats ambientals.
The role of informal green spaces in reducing inequalities in urban green space availability to children and seniors	2020	Ambiental	**Polònia**, Warsaw & Łódź	Enquesta	Urbà	L'estudi investiga **la desigualtat d'accés als espais verds "informals"**. Els resultats mostren que existeix una **desigualtat en la distribució del espais verds formals**. Proposen millores de gestió a l'accés per reduir-ne la desigualtat.

Measuring socio-economic disparities in green space availability in post-socialist cities	2021	Ambiental	**Hongria**, Debrecen, Kecskemét & Szeged	On-line	Urbà	L'estudi examinar la **facilitat d'espais verds en les zones urbanes** de les ciutats hongareses: **Debrecen**, **Kecskemét** i **Szeged**. Les zones s'estudien **en funció dels seus factors socioeconòmics**. Els resultats mostren **desigualtats entre disponibilitat d'espais verds i tipus de barris**.
Valuing urban green amenities with an inequality lens	2021	Ambiental	**Suècia**, Estocolm	On-line	Urbà	L'estudi analitza les **zones verdes afecten els preus del habitatges**. S'utilitzen **dades de venda dels habitatges de les afores** de **Suècia**, **Estocolm**. Els **resultats mostren que les zones forestals tenen impacte sobre els preus dels apartaments**.
Access to urban green space and environmental inequalities in Germany	2017	Ambiental	**Alemanya**, 53 ciutats	Enquesta	Urbà	Aquest estudi relaciona la **disponibilitat d'àrees verdes en zones urbanes** de **53 ciutats alemanyes** i els **seus nivells socioeconòmics**. Es realitza una combinació de llars amb dades de població i l'accessibilitat al verd en un radi de 500 metres. Els resultats mostren que hi ha **diferències entre ciutats i entre grups socioeconòmics**.
Exploring the equality of accessing urban green spaces: A comparative study of 341 Chinese cities	2021	Ambiental	**Xina**, 341 ciutats	On-line	Urbà i rural	Aquest estudi a**nalitza l'accés al verd de múltiples ciutats xineses**. S'han creat **dos índex que mesuren l'equitat d'accés al verd**. Els **resultats mostren disparitats d'igualtat d'accés en tot el país**. També destaquen la **relació PIB per càpita** i **la densitat** respecte aquests **indicadors**.
Do area-based intervention programs affect house prices? A quasi-experimental approach	2017	Social	**Noruega**, Oslo	-	Urbà	L'estudi es realitza a **Oslo**, per **relacionar l'atracció dels barris segons el seu nivell de preus**. Els resultats **mostren un augment significatiu en els preus en dues de les tres àrees**.
JUE Insight: City-wide effects of new housing supply: Evidence from moving chains	2022	Social	**Finlàndia**, Hèlsinki	On-line	Urbà	Aquest estudi analitza els **efecte de els habitatges de nova construcció en el mercat del centre de Hèlsinki**. Els resultats mostren que aquest **augment de la oferta beneficia** els **barris i les persones de renda mitja i baixa** ja que se'n millora l'accessibilitat a l'habitatge d'aquestes àrees.
The Urban Poor, Rental Accommodations, and Housing Policy in Korea	2002	Social	**Corea del Sud**, Seül	Enquesta	Urbà	L'article analitza els **problemes relacionats amb els habitatges situats en zones de baixos ingressos a Corea del Sud**. Destaca la **presencia de lloguers il·legals** i **habitatges deficients** que són utilitzats **per persones pobres**. Proposen **millores de les polítiques de habitatges socials**, centrant-se en el lloguer i no en la venta.

Urban poverty neighbourhoods: Typology and spatial concentration under China's market transition, a case study of Nanjing	2006	Social	**Xina**, Nanjing	Enquesta	Urbà	L'estudi investiga la **relació espacial de la pobresa urbana a la ciutat de Nanjing**. S'identifiquen tres tipus de barris: zones degradades del centre de la ciutat, comunitat de treballadors precàries i assentaments d'immigrants en zones rurals. **La pobresa va relacionada amb el desenvolupament urbà i la marginalitat** de les zones.
Accessibility of green areas for local residents	2021	Ambiental	**Països Baixos**, Amsterdam; **Àustria**, Viena; **Alemanya**, Berlin	Online	Urbà	L'estudi es basa en comprendre **com la població accedeix als espais verds** en diferents ciutats europees. Els resultats mostren com **els espais verds estan condicionats per la ubicació de la ciutat i la seva densitat**.
How does urban green space feature influence physical activity diversity in high-density built environment? An on-site observational study	2021	Ambiental	**Xina**, Xangai	Enquesta	Urbà	L'estudi relaciona els **espais verds urbans del centre de Xangai** amb **la salut pública de la seva població**. Els resultats mostren com la **proporció de cobertura verda** i la **densitat d'arbustos** estan **relacionades de manera positiva amb l'activitat física**. La **densitat d'arbres** presenta una **relació inversa i negativa**.

Font: Taula d'elaboració pròpia a partir de la cerca realitzada en la base de dades acadèmica SCOPUS. S'ha realitzat la següent cerca: (TITLE ("rental hous*") AND TITLE-ABS-KEY ("inequal*")) AND (LIMIT-TO (SUBJAREA , "SOCI") OR LIMIT-TO (SUBJAREA , "ENVI") OR LIMIT-TO (SUBJAREA, "ECON")). De tots els articles detectats s'han seleccionat per els que per contingut aportaven al estudi. Posteriorment s'han afegit 14 articles vinculats.

8.3 Annex III: Recull de principals softwares per a Machine Learning

Taula 11. Recull de principals softwares per desenvolupar els mètodes supervisats i no supervisats identificats en la revisió sistemàtica.

Algoritme	Software	
K-nearest neighbour classification (K-NN)	*Python*	*WEKA*
	R	*RapidMiner*
	TensorFlow	*Azure Machine Learning*
	Keras	*KNIME*
	MATLAB	*Orange*
Bootstrap	*Python*	*SPSS*
	R	*Stata*
	MATLAB	*Efron's Bootstrap Software*
	SAS	*GraphPad Prism*
	Julia	*JMP*
Classification and regression trees (CART)	*Python*	*Orange*
	R	*KNIME*
	MATLAB	*Azure Machine Learning*
	WEKA	*TensorFlow*
	RapidMiner	*Google Cloud AutoML*

Boosting	*Python*	*RapidMiner*
	R	*KNIME*
	Scikit-learn	*Azure Machine Learning*
	XGBoost	*AWS SageMaker*
	MATLAB	*Google Cloud AutoML*
Bagging	*Scikit-learn*	*RapidMiner*
	MATLAB	*Orange*
	R	*KNIME*
	Weka	*Azure Machine Learning*
	TensorFlow	*AWS SageMaker*
Random forest (multiple CART)	*Scikit-learn*	*RapidMiner*
	R	*Orange*
	MATLAB	*KNIME*
	Weka (Java)	*Azure Machine Learning*
	H2O	*AWS SageMaker*
Ridge regression (RR)	*Scikit-learn*	*RapidMiner*
	R	*Orange*
	MATLAB	*KNIME*
	Statsmodels	*Azure Machine Learning*
	TensorFlow	*AWS SageMaker*

Bridge regression (BR)	*R*	*RapidMiner*
	Scikit-learn	*Orange*
	MATLAB	*KNIME*
	TensorFlow	*Azure Machine Learning*
	Statsmodels	*AWS SageMaker*
Gibbs sampling	*Stan*	*WinBUGS/OpenBUGS*
	JAGS	*GeNIe Modeler*
	PyMC3	*KNIME*
	TensorFlow Probability	*Azure Machine Learning*
	R	*AWS SageMaker*
Non-negative Garrote (NNG)	*R*	*RapidMiner*
	Scikit-learn	*Orange*
	MATLAB	*KNIME*
	TensorFlow	*Azure Machine Learning*
	Statsmodels	*AWS SageMaker*
Least Absolute Shrinkage and Selection Operator (LASSO)	*Scikit-learn*	*RapidMiner*
	R	*Orange*
	MATLAB	*KNIME*
	Statsmodels	*Azure Machine Learning*
	TensorFlow	*AWS SageMaker*

Smoothly Clipped Absolute Deviation (SCAD)	*R* *Scikit-learn* *MATLAB* *TensorFlow* *Statsmodels*	*RapidMiner* *Orange* *KNIME* *Azure Machine Learning* *AWS SageMaker*
Adaptative model selection (AMS)	*Scikit-learn* *R* *XGBoost* *TensorFlow* *Weka*	*RapidMiner* *Orange* *KNIME* *Azure Machine Learning* *AWS SageMaker*
Least angle regression (LARS)	*Scikit-learn* *R* *MATLAB* *Statsmodels* *TensorFlow*	*RapidMiner* *Orange* *KNIME* *Azure Machine Learning* *AWS SageMaker*
Spike-and-slab	*R* *Stan* *JAGS* *PyMC3* *TensorFlow*	*WinBUGS/OpenBUGS* *GeNIe Modeler* *KNIME* *Azure Machine Learning* *AWS SageMaker*

Elastic Net	*Scikit-learn* *R* *MATLAB* *Statsmodels* *TensorFlow*	*RapidMiner* *Orange* *KNIME* *Azure Machine Learning* *AWS SageMaker*
Penalitzacions concoves minimax (MCP)	*R* *Scikit-learn* *MATLAB* *TensorFlow* *Statsmodels*	*RapidMiner* *Orange* *KNIME* *Azure Machine Learning* *AWS SageMaker*
Bayesian bridge regression	*Stan* *JAGS* *PyMC3* *TensorFlow* *R*	*WinBUGS/OpenBUGS* *GeNIe Modeler* *KNIME* *Azure Machine Learning* *AWS SageMaker*
Quasi-Bayesian	*Scikit-learn* *R* *TensorFlow* *MATLAB* *Statsmodels*	*RapidMiner* *Orange* *KNIME* *Azure Machine Learning* *AWS SageMaker*

Quasi-maximum likelihood estimator (QMLE)	*R* *Scikit-learn* *MATLAB* *Stata* *Statsmodels*	*EViews* *SAS* *SPSS* *Azure Machine Learning* *AWS SageMaker*
Quasi-Bayesian local likelihood (QBLL)	*R* *Python* *Stan* *MATLAB* *SAS/IML*	*KNIME* *GeNIe Modele* *RapidMiner* *Azure Machine Learning* *AWS SageMaker*
Quantile regressions (QR)	*R* *Scikit-learn* *Statsmodels* *MATLAB* *SAS*	*SPSS* *Stata* *EViews* *Azure Machine Learning* *AWS SageMaker*
Model mixta	*R* *Python* *MATLAB* *SAS* *Stata*	*SPSS* *GraphPad Prism* *JMP* *Azure Machine Learning* *AWS SageMaker*

K-means	*Scikit-learn*	*RapidMiner*
	R	*Orange*
	MATLAB	*KNIME*
	TensorFlow	*Azure Machine Learning*
	Weka	*AWS SageMaker*
Fuzzy	*Scikit-fuzzy*	*Fuzzy Logic Designer*
	R	*G fuzzy*
	MATLAB	*FuzzyCLIPS*
	Octave	*Azure Machine Learning*
	Java Fuzzy Logic Toolkit	*AWS SageMaker*
Partitioning Around Medoids (PAM)	*R*	*Orange*
	Python	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Weka	*AWS SageMaker*
Clustering Large Applications (CLARA)	*R*	*Orange*
	Python	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*

Divisie Analysis (DIANA)	*R*	*Orange*
	Python	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*
Agglomerative Nesting (AGNES)	*R*	*Orange*
	Python	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*
Clustering Large Applications based on Randomized Search (CLARANS)	*Python*	*Orange*
	R	*KNIME*
	Java	*RapidMiner*
	MATLAB	*Azure Machine Learning*
	Julia	*AWS SageMaker*
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	*Scikit-learn*	*Orange*
	R	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)	*Scikit-learn*	*Orange*
	R	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*
Density-based Clustering (DENCLUE)	*Python*	*Orange*
	R	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*
Ordering Points To Identify the Clustering Structure (OPTICS)	*Scikit-learn*	*Orange*
	R	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*
Local outlier factor (LOF)	*Scikit-learn*	*Orange*
	R	*KNIME*
	MATLAB	*RapidMiner*
	Julia	*Azure Machine Learning*
	Java	*AWS SageMaker*

Shared nearest neighbour (SNN)	*Scikit-learn* *R* *MATLAB* *Julia* *Java*	*Orange* *KNIME* *RapidMiner* *Azure Machine Learning* *AWS SageMaker*
BIRCH meets coresets for k-means clustering (BICO)	*Scikit-learn* *R* *MATLAB* *Julia* *Java*	*Orange* *KNIME* *RapidMiner* *Azure Machine Learning* *AWS SageMaker*
Hebbian Learning	*TensorFlow* *PyTorch* *NEST* *Brian2* *MATLAB*	*Neural Designer* *GeNN* *Emergent Neural Simulation Software* *Azure Machine Learning* *AWS SageMaker*
Restricted Boltzmann Machine (RBM)	*TensorFlow* *PyTorch* *scikit-learn* *MATLAB* *Theano*	*KNIME* *RapidMiner* *Neural Designer* *Azure Machine Learning* *AWS SageMaker*

Autoencoders	*TensorFlow* *PyTorch* *Keras* *scikit-learn* *MATLAB*	*KNIME* *RapidMiner* *Neural Designer* *Azure Machine Learning* *AWS SageMaker*
Deep Belief Networks (DBN)	*TensorFlow* *PyTorch* *Keras* *Theano* *MATLAB*	*KNIME* *RapidMiner* *Neural Designer* *Azure Machine Learning* *AWS SageMaker*
Generative Adversarial Networks (GAN)	*TensorFlow* *PyTorch* *Keras* *MATLAB*	*Neural Designer* *GAN Lab* *Azure Machine Learning* *AWS SageMaker*

Font: Taula d'elaboració pròpia

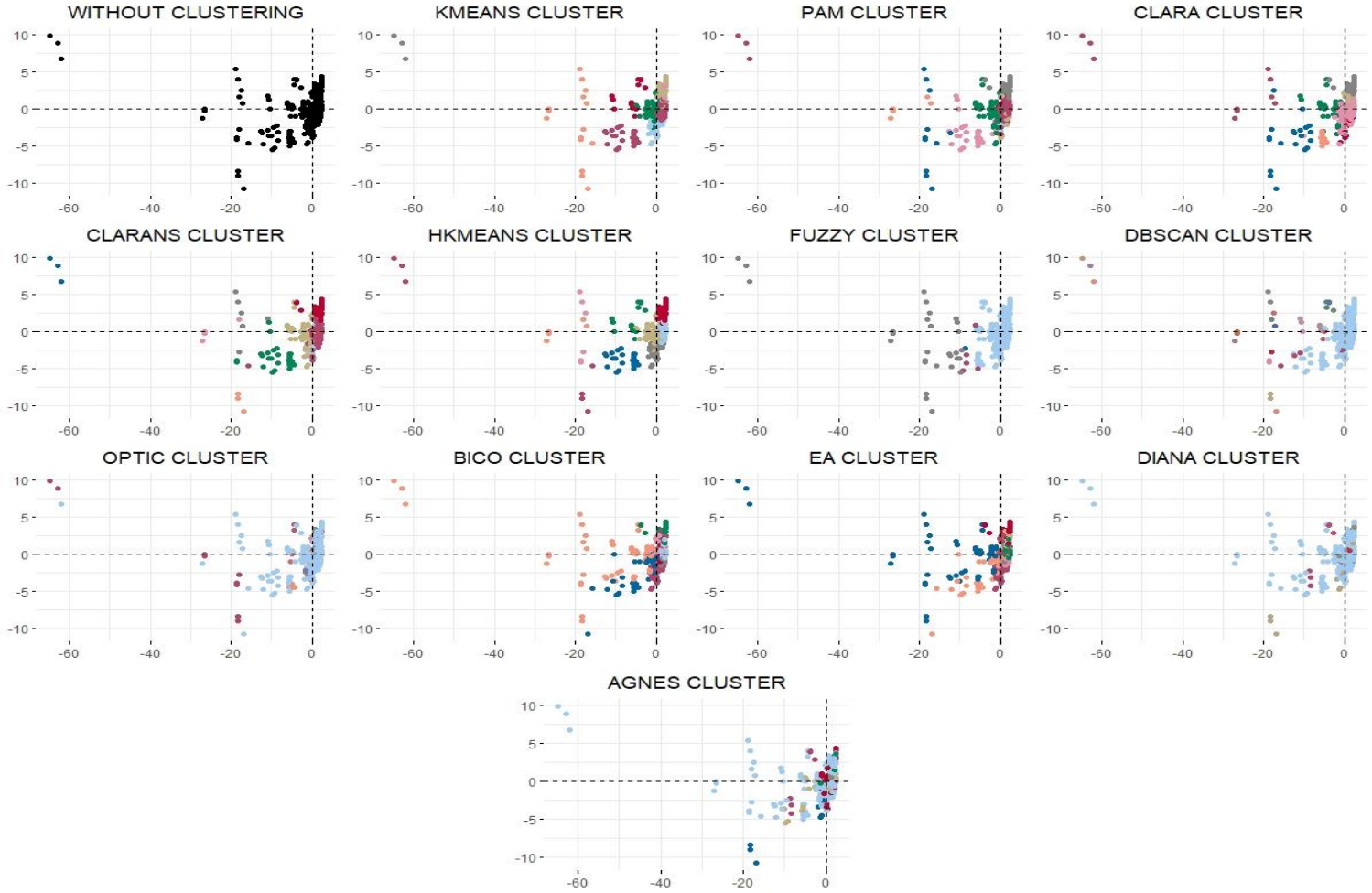
8.4 Annex IV: Estudi intra-entre clústers per 10 clústers.

Taula 12. Validació intra-entre clústers per a 10 clústers utilitzant la base de dades estandarditzada per z-score

Name	Nº Clusters	Noise Point	Avg Between	Avg Within	Avg Silhouette	DUNN Index	Entropy	WB Ratio	CH Index	Separation Index
Data set: Z-score										
K-MEANS	10	0	8,7562	5,2300	0,1298	0,0355	1,9118	0,5973	173,4225	1,7544
PAM	10	0	8,5385	5,4082	0,0913	0,0289	1,9805	0,6334	163,3552	1,4971
CLARA	10	0	8,6918	5,5430	0,0612	0,0396	1,8444	0,6377	153,4150	1,4097
CLARANS	10	0	9,0511	5,4920	0,1237	0,0515	1,5391	0,6068	165,6828	1,6965
HKMEANS	10	0	9,6981	5,4041	0,2028	0,0922	1,4347	0,5572	180,5587	2,6274
FUZZY	4	0	23,1511	6,9918	0,4351	0,0530	0,2349	0,3020	125,6927	7,0499
BICO	10	0	8,4412	6,1908	0,0172	0,0170	2,0855	0,7334	39,0796	1,5683
EA	10	0	8,4412	6,1908	0,0172	0,0170	2,0855	0,7334	39,0796	1,5683
DIANA	9	0	8,1173	7,8470	-0,2971	0,0161	0,5394	0,9667	1,9863	1,5233
AGNES	10	0	7,6427	7,8028	-0,2636	0,0154	1,2448	1,0209	2,7850	1,3453

Font: Taula d'elaboració pròpia

Figura 18. Representació dels diferents mecanismes d'agrupació aplicat a la base de dades estandarditzada amb z-score, on k=10



Font: Gràfic d'elaboració pròpia

8.5 Annex V: Fonts d'informació identificades en l'Article I

Variables	Tipus	Anys	Temporalitat	Tots els municipis	Font
Població a 1 de gener. Per sexe	Demografia	1998-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per sexe i edat any a any	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per sexe i edat quinquennal	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per sexe i edat en grans grups	Demografia	2000-2020	Anual	Si	IDESCAT
Poblacióa 1 de gener. Per sexe i generacions	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per lloc de naixement. Totals	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per lloc de naixement i sexe	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per lloc de naixement (CA i estranger)	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per lloc de naixement (CA i estranger) i sexe	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per nacionalitat i sexe	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per nacionalitat, sexe i edat quinquennal	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per nacionalitat, sexe i edat en grans grups	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per nacionalitat (continents)	Demografia	2000-2020	Anual	Si	IDESCAT
Població a 1 de gener. Per nacionalitat (continents) i sexe	Demografia	2000-2020	Anual	Si	IDESCAT
Població. Per sexe	Demografia	1975-1996	Anual	Si	IDESCAT
Población extranjera a 1 de enero. Por municipios	Demografia	2000-2020	Anual	Si	IDESCAT
Població per sexe	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons sexe i edat any a any	Demografia	2001-2011	10 anys	Si	IDESCAT

Població segons sexe i edat quinquennal	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons sexe i edat en grans grups	Demografia	2001-2011	10 anys	Si	IDESCAT
Població per sexe i generacions	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons sexe i edat quinquennal	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons sexe i edat en grans grups	Demografia	2001-2011	10 anys	Si	IDESCAT
Població. Per sexe, grans grups d'edat i estat civil	Demografia	2001-2011	10 anys	No	IDESCAT
Població. Per sexe i estat civil	Demografia	2001-2011	10 anys	Si	IDESCAT
Població. Per sexe i estat civil (agregat)	Demografia	2011	10 anys	Si	IDESCAT
Població. Per sexe, edat qinquennal i estat civil	Demografia	2001	10 anys	Si	IDESCAT
Població. Lloc de naixement per comunitats autònomes i estat civil	Demografia	2001	10 anys	Si	IDESCAT
Població. Nacionalitat per continents i estat civil	Demografia	2001-2011	10 anys	No	IDESCAT
Població. Per sexe, edat quinquennal, nacionalitat estrangera i estat civil	Demografia	2011	10 anys	No	IDESCAT
Població. Per nacionalitat, sexe, grans grups d'edat i estat civil	Demografia	2011	10 anys	No	IDESCAT
Població segons lloc de naixement. Totals	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons lloc de naixement, sexe i edat	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons lloc de naixement per comunitats autònomes	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons lloc de naixement per comunitats autònomes i sexe	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons lloc de naixement per comunitats autònomes i edat	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons país de naixement i sexe	Demografia	2011	10 anys	No	IDESCAT
Població segons país de naixement, sexe i nacionalitat	Demografia	2011	10 anys	No	IDESCAT
Població segons lloc de naixement per continents	Demografia	2001-2011	10 anys	Si	IDESCAT

Població segons lloc de naixement per continents i sexe	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons lloc de naixement per continents i nacionalitat	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons lloc de naixement i nivell d'instrucció. Població 10 anys i més	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons el lloc de naixement i relació amb l'activitat. Total població	Demografia	2001	10 anys	Si	IDESCAT
Població segons lloc de naixement i sectors d'activitat dels ocupats (CCAE93)	Demografia	2001	10 anys	Si	IDESCAT
Població segons el lloc de naixement i branques d'activitat dels ocupats	Demografia	1991	10 anys	Si	IDESCAT
Població segons lloc de naixement i branques d'activitat dels ocupats (CCAE93)	Demografia	2001	10 anys	Si	IDESCAT
Població segons el lloc de naixement i branques d'activitat dels desocupats	Demografia	1991	10 anys	Si	IDESCAT
Població segons el lloc de naixement i professió dels ocupats (CCO94)	Demografia	2001	10 anys	Si	IDESCAT
Població segons el lloc de naixement i professió dels ocupats	Demografia	1991	10 anys	Si	IDESCAT
Població segons el lloc de naixement i professió dels desocupats	Demografia	1991	10 anys	Si	IDESCAT
Població segons nacionalitat i sexe	Demografia	2001	10 anys	Si	IDESCAT
Població segons nacionalitat, sexe i grup d'edat	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons nacionalitat per continents	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons nacionalitat per continents i sexe	Demografia	2001-2011	10 anys	Si	IDESCAT
Població segons nacionalitat per continents, sexe i edat	Demografia	2001-2011	10 anys	Si	IDESCAT

Població segons nacionalitat per continents i lloc de naixement	Demografia	2001-2011	10 anys	Si	IDESCAT
Població de 2 anys i més segons coneixement del català	Demografia	2001-2011	10 anys	No	IDESCAT
Població de 2 anys i més segons coneixement del català (%)	Demografia	2001-2011	10 anys	No	IDESCAT
Població de 2 anys i més segons coneixement del català i sexe	Demografia	2001-2011	10 anys	No	IDESCAT
Població de 2 anys i més segons coneixement del català i edat.	Demografia	2001-2011	10 anys	Si	IDESCAT
Població de 2 anys i més segons coneixement del català per edat i sexe	Demografia	2001-2011	10 anys	Si	IDESCAT
Població de 2 anys i més segons coneixement del català i lloc de naixement.	Demografia	2001-2011	10 anys	Si	IDESCAT
Població de 2 anys i més segons coneixement del català i any d'arribada a Catalunya	Demografia	1986-1996	10 anys	Si	IDESCAT
Població de 10 anys i més segons coneixement del català i nivell d'instrucció.	Demografia	2001-2011	10 anys	Si	IDESCAT
Població de 2 anys i més segons coneixement del català i estudis en curs.	Demografia	2001-2011	10 anys	Si	IDESCAT
Coneixement del català dels ocupats per professió (CCO94)	Demografia	2001-2011	10 anys	Si	IDESCAT
Llars. Total	Demografia	2001-2011	10 anys	No	IDESCAT
Llars. Per nombre de famílies	Demografia	2001-2011	10 anys	No	IDESCAT
Llars. Per nombre de famílies (agregat)	Demografia	2001-2011	10 anys	No	IDESCAT
Llars. Per nombre i tipus de nucli	Demografia	2001-2011	10 anys	No	IDESCAT
Llars. Per nombre i tipus de nucli (agregat)	Demografia	2001-2011	10 anys	No	IDESCAT
Llars. Per nombre i tipus de nucli i, per nombre de fills	Demografia	2001-2011	10 anys	No	IDESCAT

Llars. Per nombre i tipus de nucli i, per nombre de fills menors de 16 anys	Demografia	2001-2011	10 anys	No	IDESCAT
Llars d'estrangers. Per grandària de la llar amb algun dels membres estranger	Demografia	2001	10 anys	No	IDESCAT
Llars d'estrangers. Per grandària de la llar amb tots els membres estrangers	Demografia	2001	10 anys	No	IDESCAT
Composició de les llars per generacions	Demografia	2001-2011	10 anys	No	IDESCAT
Nuclis de matrimonis. Per nombre de fills menors de 16 anys	Demografia	2001	10 anys	No	IDESCAT
Nuclis de parelles de fet. Per estat civil	Demografia	2001-2011	10 anys	No	IDESCAT
Nuclis de parelles de fet. Per nombre de fills menors de 16 anys	Demografia	2001-2011	10 anys	No	IDESCAT
Nuclis monoparentals. Per tipus	Demografia	2001-2011	10 anys	No	IDESCAT
Nuclis monoparentals. Per tipus i fills menors de 16 anys	Demografia	2001-2011	10 anys	No	IDESCAT
Nuclis monoparentals. Per estat civil. Pares	Demografia	2001-2011	10 anys	No	IDESCAT
Nuclis monoparentals. Per estat civil. Mares	Demografia	2001-2011	10 anys	No	IDESCAT
Nuclis monoparentals. Per estat civil. Total	Demografia	2001-2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per nombre de fills nascuts vius. Fins a 8 i més fills	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per nombre de fills nascuts vius. Fins a 4 i més fills	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per edat any a any i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per edat quinquennal i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Dones per nombre de fills biològics tinguts i edat	Salut	2011	10 anys	No	IDESCAT
Dones per nombre de fills biològics tinguts, nacionalitat i edat	Salut	2011	10 anys	No	IDESCAT

Dones per nombre de fills biològics tinguts i estat civil	Salut	2011	10 anys	No	IDESCAT
Dones per nombre de fills biològics tinguts, estat civil i edat	Salut	2011	10 anys	No	IDESCAT
Dones casades i no casades per nombre de fills biològics tinguts i edat	Salut	2011	10 anys	No	IDESCAT
Dones casades i no casades per nombre de fills biològics tinguts i edat	Salut	2011	10 anys	No	IDESCAT
Dones solteres i no solteres per nombre de fills biològics tinguts i edat	Salut	2011	10 anys	No	IDESCAT
Dones per nombre de fills biològics tinguts i relació amb l'activitat econòmica	Salut	2011	10 anys	No	IDESCAT
Dones per edat i convivència amb els pares	Salut	2011	10 anys	No	IDESCAT
Dones per edat i convivència amb els pares	Salut	2011	10 anys	No	IDESCAT
Dones segons si tenen fills o no per nacionalitat i per la de la parella	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per any de naixement, nacionalitat i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per estat civil i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per estat civil, edat i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per lloc de naixement, edat i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per nacionalitat, edat i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Dones de 16 anys i més. Per principals països de nacionalitat i nombre de fills nascuts vius	Salut	2011	10 anys	No	IDESCAT
Habitatges principals. Per règim de tinença	Mercat Laboral	2001	10 anys	Si	IDESCAT

Habitatges principals segons l'any de construcció de l'edifici	Mercat Laboral	2001	10 anys	Si	IDESCAT
Habitatges principals segons l'any de construcció i règim de tinença	Mercat Laboral	1991	10 anys	Si	IDESCAT
Habitatges principals segons l'any de construcció i nombre d'habitacions	Mercat Laboral	1991	10 anys	Si	IDESCAT
Habitatges principals. Per superfície útil	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Habitatges principals segons la superfície útil i règim de tinença	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Habitatges principals segons la superfície útil i any de construcció	Mercat Laboral	1991	10 anys	Si	IDESCAT
Habitatges principals. Per superfície útil i instal.lacions	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Habitatges principals. Per nombre d'habitacions	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Habitatges principals segons instal·lacions	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Habitatges principals segons instal·lacions i serveis (II)	Mercat Laboral	1991	10 anys	Si	IDESCAT
Habitatges principals segons instal·lacions i serveis (III)	Mercat Laboral	1991	10 anys	Si	IDESCAT
Habitatges principals amb calefacció segons el tipus de combustible	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Habitatges principals amb calefacció segons el combustible utilitzat	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Habitatges principals segons el nombre d'habitatges de l'edifici	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Establiments col·lectius. Per tipus.	Mercat Laboral	2011	10 anys	Si	IDESCAT

Habitatges principals per any d'arribada a l'habitatge	Mercat Laboral	1996-2001	10 anys	Si	IDESCAT
Habitatges principals per disponibilitat de segona residència	Mercat Laboral	2011	10 anys	Si	IDESCAT
Habitatges principals segons el temps dútilització de les segones residències	Mercat Laboral	2001	10 anys	Si	IDESCAT
Habitatges principals segons disponibilitat de vehicle	Mercat Laboral	2001	10 anys	Si	IDESCAT
Habitatges principals per problemes en l'habitatge	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població. Per any d'arribada a Catalunya	Demografia	1986-1991	10 anys	Si	IDESCAT
Població. Per any d'arribada a Catalunya i sexe	Demografia	1986-1991	10 anys	Si	IDESCAT
Població. Per any d'arribada a Catalunya, sexe i edat	Demografia	1986-1991	10 anys	Si	IDESCAT
Població. Per lloc de naixement i any d'arribada a Catalunya	Demografia	1986-1991	10 anys	Si	IDESCAT
Població. Per any d'arribada a Catalunya i professió dels ocupats	Demografia	1986-1991	10 anys	Si	IDESCAT
Població. Per any d'arribada a Catalunya i professió dels ocupats (CCO94)	Demografia	1986-1991	10 anys	Si	IDESCAT
Població. Per any d'arribada a Catalunya i professió dels desocupats	Demografia	1986-1991	10 anys	Si	IDESCAT
Població ocupada resident i llocs de treball localitzats per sexe	Demografia	1986-2001	10 anys	Si	IDESCAT
Població que estudia resident i llocs d'estudi localitzats per sexe	Demografia	1986-2001	10 anys	Si	IDESCAT
Població ocupada resident i llocs de treball localitzats per tipus de transport	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població que estudia resident i llocs d'estudi localitzats per tipus de transport	Mercat Laboral	2001	10 anys	Si	IDESCAT

Població ocupada resident segons desplaçament al lloc de treball per tipus de transport	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població que estudia resident segons desplaçament al lloc d' estudi per tipus de transport	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població ocupada resident segons desplaçament al lloc de treball per mitjans de transport. Resposta múltiple	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població que estudia resident segons desplaçament al lloc d'estudi per mitjans de transport. Resposta múltiple	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població ocupada segons nombre de viatges diaris per anar a treballar	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més que estudia resident segons nombre de viatges diaris per anar a estudiar	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població ocupada resident segons temps de desplaçament per anar a treballar	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més que estudia resident segons temps de desplaçament per anar a estudiar	Mercat Laboral	2001	10 anys	Si	IDESCAT
Localització de l'ocupació per branques d'activitat (CCAE-93)	Mercat Laboral	2001	10 anys	Si	IDESCAT
Localització de l'ocupació per branques d'activitat. Població de 16 anys i més	Mercat Laboral	1986-1991	10 anys	Si	IDESCAT
Localització de l'ocupació per professions (CCO-94)	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Localització de l'ocupació laboral per professions. Població de 16 anys i més	Mercat Laboral	1991	10 anys	Si	IDESCAT
Població de 16 anys i més segons nivell d'instrucció	Mercat Laboral	1991-2001	10 anys	Si	IDESCAT
Població de 16 anys i més segons nivell d'instrucció i sexe	Mercat Laboral	2001	10 anys	Si	IDESCAT

Població de 16 anys i més segons nivell d'instrucció, sexe i edat	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més segons estudis en curs. Resposta múltiple	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més segons estudis en curs i sexe. Resposta múltiple	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més segons estudis en curs, sexe i edat. Resposta múltiple	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més amb FP o 3er grau segons tipus d'estudis realitzats	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més amb FP o 3r grau segons tipus d'estudis realitzats i sexe	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població de 16 anys i més amb FP o 3er grau segons tipus d'estudis realitzats, sexe i edat	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població per relació amb l'activitat. Total població. Recomptes	Mercat Laboral	2001	10 anys	Si	IDESCAT
Població per relació amb l'activitat i sexe	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Població per relació amb l'activitat, sexe i edat	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Població per relació amb l'activitat, estat civil i sexe	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Població de 16 anys i més per relació amb l'activitat. Dades bàsiques	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Població per relació amb l'activitat, nivell d'instrucció i sexe. Població>=16 anys	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per grans sectors d'activitat. Població de 16 anys i més (CCAE93)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per grans sectors d'activitat i sexe. Població de 16 anys i més (CCAE93)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT

Ocupats per branques d'activitat. Població de 16 anys i més. Recomptes (CCAE93)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per branques d'activitat. Població de 16 anys i més. Recomptes	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per branques d'activitat i sexe. Població de 16 anys i més (CCAE93)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per branques d'activitat i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per branques d'activitat, sexe i edat. Població de 16 anys i més (CCAE93)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per branques d'activitat, sexe i edat. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per branques d'activitat, professió i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per branques d'activitat, professió i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per professió. Població de 16 anys i més. Recomptes (CCO94)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per professió. Població de 16 anys i més. Recomptes (CNO79)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per professió i sexe. Població de 16 anys i més (CCO94)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per professió i sexe. Població de 16 anys i més (CNO79)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per professió, sexe i edat. Població de 16 anys i més (CCO94)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per professió, sexe i edat. Població de 16 anys i més (CNO79)	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT

Ocupats per situació professional i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per situació professional, sexe i edat. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per condició socioeconòmica. Recomptes. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per condició socioeconòmica i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per hores treballades. Població de 16 anys i més. Recomptes	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per hores treballades i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per hores treballades i sectors d'activitat. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per hores treballades i professió. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats per hores treballades i situació professional. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Població per grans sectors d'activitat. Població de 16 anys i més. Recomptes	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Ocupats i desocupats per grans sectors d'activitat i sexe. Població >= 16 anys	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Desocupats per branques d'activitat. Població de 16 anys i més. Recomptes	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Desocupats per branques d'activitat i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Desocupats per branques d'activitat, sexe i edat. Població 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT

Desocupats per branques d'activitat, professió i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Desocupats per professió. Població de 16 anys i més. Recomptes	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Desocupats per professió i sexe. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Desocupats per professió, sexe i edat. Població de 16 anys i més	Mercat Laboral	1986-2001	5 anys	Si	IDESCAT
Creixement intercensal de la població. 2001-2011	Demografia	2001-2011	10 anys	Si	IDESCAT
Creixement intercensal de la població, per components. 2001-2011	Demografia	2001-2011	10 anys	Si	IDESCAT
Creixement intercensal de la població, per components i sexe. 2001-2011	Demografia	2001-2011	10 anys	Si	IDESCAT
Creixement intercensal de la població, per components en taxa. Mitjana anual 2001-2011	Demografia	2001-2011	10 anys	Si	IDESCAT
Creixement intercensal de la població, per components. 1996-2001	Demografia	1996-2001	5 anys	Si	IDESCAT
Creixement intercensal de la població, per components i sexe. 1996-2001	Demografia	1996-2001	5 anys	Si	IDESCAT
Creixement intercensal de la població, per components en taxa. Mitjana anual 1996-2001	Demografia	1996-2001	5 anys	Si	IDESCAT
Creixement intercensal de la població, per components. 1991-1996	Demografia	1991-1996	5 anys	Si	IDESCAT
Creixement intercensal de la població, per components i sexe. 1991-1996	Demografia	1991-1996	5 anys	Si	IDESCAT
Creixement intercensal de la població, per components en taxa. Mitjana anual 1991-1996	Demografia	1991-1996	5 anys	Si	IDESCAT
Creixement intercensal de la població, per components. 1986-1991	Demografia	1986-1991	5 anys	Si	IDESCAT

Creixement intercensal de la població, per components i sexe. 1986-1991	Demografia	1986-1991	5 anys	Si	IDESCAT
Creixement intercensal de la població, per components en taxa. Mitjana anual 1986-1991	Demografia	1986-1991	5 anys	Si	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013). Sexe	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013) amb intervals de confiança (95%). Sexe	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013). Grup d'edat. Total	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013) amb intervals de confiança (95%). Grup d'edat. Total	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013). Grup d'edat. Homes	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013) amb intervals de confiança (95%). Grup d'edat. Homes	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013). Grup d'edat. Dones	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 2013) amb intervals de confiança (95%). Grup d'edat. Dones	Salut	2013-2020	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 1976). Sexe	Salut	1988-2008	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE) amb intervals de confiança	Salut	1988-2008	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 1976). Grup d'edat. Ambdós sexes	Salut	1988-2008	Anual	No	IDESCAT
Taxa de mortalitat estandarditzada (PEE 1976). Grup d'edat. Homes	Salut	1988-2008	Anual	No	IDESCAT

Taxa de mortalitat estandarditzada (PEE 1976). Grups d'edat. Dones	Salut	1988-2008	Anual	No	IDESCAT
Població de 15 anys i més segons nivell de formació assolit. Per nivells agregats	Mercat Laboral	2019	Anual	No	IDESCAT
Indicadors dels estudis de la població de 25 a 64 anys. Per sexe	Mercat Laboral	2019	Anual	No	IDESCAT
Població de 15 anys i més segons nivell de formació assolit i sexe	Mercat Laboral	2019	Anual	No	IDESCAT
Població de 15 anys i més segons nivell de formació assolit i grups d'edat decennal	Mercat Laboral	2019	Anual	No	IDESCAT
Població de 15 anys i més segons nivell de formació assolit i nacionalitat	Mercat Laboral	2019	Anual	No	IDESCAT
Renda familiar disponible bruta (RFDB). Revisió estadística 2019. Índex	Economia	2010-2018	Anual	No	IDESCAT
Renda familiar disponible (RFDB). Revisió estadística 2019. Per principals components	Economia	2010-2018	Anual	No	IDESCAT
Renda familiar disponible bruta (RFDB). Revisió estadística 2019. Per principals recursos (%)	Economia	2010-2018	Anual	No	IDESCAT
Renda familiar disponible bruta (RFDB). Revisió estadística 2019. Per principals usos (%)	Economia	2010-2018	Anual	No	IDESCAT
Producte interior brut.	Economia	2010-2018	Anual	No	IDESCAT
Valor afegit brut. Per sectors	Economia	2010-2018	Anual	No	IDESCAT
Valor afegit brut. Per sectors (%)	Economia	2010-2018	Anual	No	IDESCAT
Parc de vehicles, per tipus	Economia	1991-2012	Anual	Si	IDESCAT
Parc de vehicles. Índex de motorització	Economia	1991-2012	Anual	Si	IDESCAT
Població resident a l'estranger. Per lloc d'inscripció i sexe	Demografia	2009-2021	Anual	Si	IDESCAT

Població resident a l'estranger. Per lloc de naixement, sexe i edat quinquennal	Demografia	2009-2021	Anual	Si	IDESCAT
Població resident a l'estranger. Per país de residència (més de 500 residents) i sexe	Demografia	2009-2021	Anual	Si	IDESCAT
Població resident a l'estranger. Per continent de residència, sexe i edat quinquennal	Demografia	2009-2021	Anual	Si	IDESCAT
Població resident a l'estranger. Per continent de residència i lloc de naixement	Demografia	2009-2021	Anual	Si	IDESCAT
Població resident a l'estranger. Per lloc de naixement i continent de residència (agregat)	Demografia	2009-2021	Anual	Si	IDESCAT
Població resident a l'estranger. Per país de residència (més de 500 residents) i lloc de naixement (agregat)	Demografia	2009-2021	Anual	Si	IDESCAT
Emigracions externes. Per continent de destinació.	Demografia	2005-2020	Anual	Si	IDESCAT
Altes i baixes de residència	Demografia	1981-1994	Anual	Si	IDESCAT
Migracions. Totals	Demografia	1998-2020	Anual	Si	IDESCAT
Migracions. Sexe i grups d'edat	Demografia	2002-2020	Anual	Si	IDESCAT
Immigracions. Per municipi de destinació i lloc de procedència.	Demografia	1998-2020	Anual	Si	IDESCAT
Emigracions. Per municipi de procedència i lloc de destinació.	Demografia	1998-2020	Anual	Si	IDESCAT
Saldos migratoris interns. Municipis	Demografia	1998-2020	Anual	Si	IDESCAT
Immigracions externes. Per continent de procedència.	Demografia	2004-2020	Anual	Si	IDESCAT
Impost sobre la renda de les persones físiques (IRPF)	Economia	2000-2019	Anual	Si	IDESCAT
Impost de béns immobles de naturalesa urbana (IBI)	Economia	2006-2020	Anual	No	IDESCAT

Impost de béns immobles de naturalesa rústica (IBI)	Economia	2006-2020	Anual	No	IDESCAT
Cadastre immobiliari urbà	Economia	2006-2020	Anual	No	IDESCAT
Cadastre immobiliari rústic	Economia	2006-2020	Anual	No	IDESCAT
Estimacions de població ETCA i de població estacional ETCA	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població estacional ETCA, per trimestre	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població vinculada ETCA i taxa de vinculació	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població vinculada no resident ETCA, per trimestre	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població vinculada	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població ETCA i de població estacional ETCA	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població estacional ETCA, per trimestre	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població vinculada ETCA i taxa de vinculació	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població vinculada no resident ETCA, per trimestre	Demografia	2015-2020	Anual	No	IDESCAT
Estimacions de població vinculada	Demografia	1986-2020	Anual	No	IDESCAT
Estimacions de poblacio per sexe	Demografia	1986-2020	Anual	No	IDESCAT
Estimacions de poblacio per sexe i edat any a any	Demografia	1986-2020	Anual	No	IDESCAT
Estimacions de poblacio per sexe i edat quinquennal	Demografia	1986-2020	Anual	No	IDESCAT
Moviment demogràfic de les estimacions postcensals de població. Homes.	Demografia	2012-2020	Anual	No	IDESCAT
Moviment demogràfic de les estimacions postcensals de població. Dones.	Demografia	2012-2020	Anual	No	IDESCAT

Moviment demogràfic de les estimacions postcensals de població.	Demografia	2012-2020	Anual	No	IDESCAT
Població segons hagin tingut alguna relació laboral d'ocupació, per sexe.	Mercat Laboral	2006-2018	Anual	No	IDESCAT
Població segons hagin tingut alguna relació laboral d'ocupació, per sexe. Percentatge	Mercat Laboral	2006-2018	Anual	No	IDESCAT
Efectius de les policies locals. Per graduació	Incidéncies i emergències	1993-2020	Anual	Si	IDESCAT
Eleccions al Parlament de Catalunya. Dades generals	Despesa pública	2010,2012,2015,2017	Aleatori	Si	IDESCAT
Resultats de les eleccions al Parlament de Catalunya. Vots a partits	Despesa pública	2010,2012,2015,2017	Aleatori	Si	IDESCAT
Residus municipals. Generació. Total registrat i generació per càpita	Mercat Laboral	2004-2020	Anual	No	IDESCAT
Residus municipals. Generació. Recollida selectiva registrada. Per tipus de residu	Mercat Laboral	2004-2020	Anual	No	IDESCAT
Residus municipals. Tractament. Recollida no selectiva registrada. Per tipus de tractament	Mercat Laboral	2004-2020	Anual	No	IDESCAT
Renda Garantida de Ciutadania	Economia	2018-2020	Anual	No	IDESCAT
Persones reconegudes legalment com a discapacitades per sexe	Demografia	1998-2020	Anual	No	IDESCAT
Persones reconegudes legalment com a discapacitades per edat	Demografia	1998-2020	Anual	No	IDESCAT
Persones reconegudes legalment com a discapacitades segons el tipus de discapacitat	Demografia	2018-2020	Anual	No	IDESCAT
Persones reconegudes legalment com a discapacitades segons el grau de discapacitat	Demografia	1998-2020	Anual	No	IDESCAT
Nascuts vius segons sexe	Salut	1975-2020	Anual	Si	IDESCAT

Nascuts vius segons sexe i edat de la mare	Salut	1975-2020	Anual	No	IDESCAT
Nascuts vius segons lloc de residència/inscripció i sexe	Salut	1975-2020	Anual	Si	IDESCAT
Nascuts vius segons el mes del part i el sexe	Salut	1975-2020	Anual	Si	IDESCAT
Nascuts vius segons sexe i pes al néixer	Salut	1975-2020	Anual	No	IDESCAT
Nascuts vius segons edat de la mare i pes al néixer	Salut	1975-2020	Anual	No	IDESCAT
Nascuts vius segons sexe i nacionalitat de la mare	Salut	1975-2020	Anual	Si	IDESCAT
Nascuts vius segons edat i nacionalitat de la mare	Salut	1975-2020	Anual	Si	IDESCAT
Nascuts vius segons mare estrangera per sexe i edat de la mare	Salut	1975-2020	Anual	Si	IDESCAT
Nascuts vius segons mare estrangera per sexe i mes de part	Salut	1975-2020	Anual	Si	IDESCAT
Parts segons edat de la mare i multiplicitat	Salut	1975-2020	Anual	No	IDESCAT
Parts segons edat de la mare i normalitat	Salut	1975-2020	Anual	No	IDESCAT
Parts segons edat de la mare i maturitat	Salut	1975-2020	Anual	No	IDESCAT
Parts segons edat de la mare, tipus de part i maturitat	Salut	1975-2020	Anual	No	IDESCAT
Parts segons edat de la mare i assistència sanitària	Salut	1975-2020	Anual	No	IDESCAT
Parts segons multiplicitat i normalitat	Salut	1975-2020	Anual	No	IDESCAT
Parts segons multiplicitat i maturitat	Salut	1975-2020	Anual	No	IDESCAT
Morts fetals tardanes	Salut	1975-2020	Anual	No	IDESCAT
Naixements, defuncions i matrimonis. Recomptes	Demografia	1975-2020	Anual	Si	IDESCAT
Matrimonis segons el lloc de residència / inscripció	Demografia	2012-2020	Anual	Si	IDESCAT
Matrimonis segons el tipus de celebració	Demografia	1976-2011	Anual	Si	IDESCAT
Matrimonis segons el mes de celebració	Demografia	2012-2020	Anual	Si	IDESCAT
Atur registrat (a 31 de març). Recomptes	Mercat Laboral	2005-2020	Anual	Si	IDESCAT
Atur registrat. Per sexe. Mitjanes anuals	Mercat Laboral	2005-2020	Anual	Si	IDESCAT

Atur registrat. Per sectors. Mitjanes anuals	Mercat Laboral	2009-2020	Anual	Si	IDESCAT
Atur registrat. Per branques d'activitat. Mitjanes anuals	Mercat Laboral	2009-2021	Anual	Si	IDESCAT
Alumnes residents i llocs d'estudi localitzats. Ensenyaments universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Alumnes residents per lloc d'estudi. Ensenyaments universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Llocs d'estudi localitzats per lloc de residència de l'alumne. Ensenyaments universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Alumnes residents i llocs d'estudi localitzats per sexe. Ensenyaments universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Alumnes residents i llocs d'estudi localitzats. Ensenyaments no universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Alumnes residents per lloc d'estudi. Ensenyaments no universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Llocs d'estudi localitzats per lloc de residència de l'alumne. Ensenyaments no universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Alumnes residents i llocs d'estudi localitzats per sexe. Ensenyaments no universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Alumnes residents i llocs d'estudi localitzats per edat. Ensenyaments no universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Alumnes residents i llocs d'estudi localitzats per nivell d'estudis. Ensenyaments no universitaris	Demografia	2012-2020	Anual	Si	IDESCAT
Biblioteques per tipus	Despesa pública	2002-2008	Anual	Si	IDESCAT
Biblioteques per titularitat	Despesa pública	2002-2008	Anual	Si	IDESCAT
Població segons llengua inicial	Demografia	2008-2018	Quinquenal	No	IDESCAT
Població segons llengua habitual	Demografia	2008-2018	Quinquenal	No	IDESCAT

Població segons usos lingüístics i àmbits d'ús	Demografia	2008-2018	Quinquenal	No	IDESCAT
Defuncions segons sexe	Salut	1975-2020	Anual	Si	IDESCAT
Defuncions segons el lloc de residència/inscripció i sexe	Salut	1975-2020	Anual	Si	IDESCAT
Defuncions segons el mes de la mort i sexe	Salut	1975-2020	Anual	Si	IDESCAT
Afiliats a la Seguretat Social segons residència padronal de l'afiliat. Per sexe. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliats a la Seguretat Social segons residència padronal de l'afiliat. Per grups d'edat. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliats a la Seguretat Social segons residència padronal de l'afiliat. Per nacionalitat. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliacions a la Seguretat Social segons residència padronal de l'afiliat. Per sexe. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliacions a la Seguretat Social segons residència padronal de l'afiliat. Per grups d'edat. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliacions a la Seguretat Social segons residència padronal de l'afiliat. Per nacionalitat. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliacions a la Seguretat Social segons residència padronal de l'afiliat. Per tipus de relació laboral. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliacions a la Seguretat Social segons residència padronal de l'afiliat. Per sector d'activitat. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Afiliacions a la Seguretat Social segons residència padronal de l'afiliat. Règim general de la Seguretat Social i Règim especial de treballadors autònoms. A últim dia del mes	Mercat Laboral	2012-2020	Anual	No	IDESCAT
Capital de provincia	Geografia	1975-2020	Anual	Si	PROPI
Municipi de muntanya	Geografia	1975-2020	Anual	Si	PROPI

Municipi de costa	Geografia	1975-2020	Anual	Si	PROPI
Altitud	Geografia	1975-2020	Anual	Si	IDESCAT
Latitud	Geografia	1975-2020	Anual	Si	IDESCAT
Longitud	Geografia	1975-2020	Anual	Si	IDESCAT
Capital de la comarca	Geografia	1975-2020	Anual	Si	PROPI
Habitant per km² de sòl urbà	Demografia	2012-2020	Anual	Si	XIFRA
Habitant per km² de sòl urbanitzable	Demografia	2012-2020	Anual	No	XIFRA
Habitant per km² de sòl no urbanitzable	Demografia	2012-2020	Anual	Si	XIFRA
Població. Variació percentual	Demografia	2000-2020	Anual	Si	XIFRA
Variació percentual de l'estructura	Demografia	2000-2020	Anual	Si	XIFRA
Habitants per km²	Demografia	2000-2020	Anual	Si	XIFRA
Índex de masculinitat	Demografia	2000-2020	Anual	Si	XIFRA
Edat mitjana de la població	Demografia	2000-2020	Anual	Si	XIFRA
Percentatge de joves	Demografia	2000-2020	Anual	Si	XIFRA
Percentatge d'adults	Salut	2000-2020	Anual	Si	XIFRA
Percentatge de població gran	Salut	2000-2020	Anual	Si	XIFRA
Taxa de joventut	Salut	2000-2020	Anual	Si	XIFRA
Índex d'envelliment	Salut	2000-2020	Anual	Si	XIFRA
Índex de sobreenvelliment	Salut	2000-2020	Anual	Si	XIFRA
Índex de dependència global	Salut	2000-2020	Anual	Si	XIFRA
Índex de dependència juvenil	Salut	2000-2020	Anual	Si	XIFRA
Índex de dependència senil	Salut	2000-2020	Anual	Si	XIFRA
Índex de recanvi de la població en edats actives	Salut	2000-2020	Anual	Si	XIFRA
Índex de potencialitat	Demografia	2000-2020	Anual	Si	XIFRA
Índex de tendència	Demografia	2000-2020	Anual	Si	XIFRA
Dimensió mitjana de la llar	Demografia	1986-2011	Quinquenal	No	XIFRA
Taxa de llars unipersonals	Demografia	1986-2011	Quinquenal	No	XIFRA
Taxa de llars amb família nombrosa	Demografia	1986-2011	Quinquenal	No	XIFRA

Variació percentual d'estrangers	Demografia	2000-2019	Anual	Si	XIFRA
Índex d'autoctonisme	Demografia	2000-2019	Anual	Si	XIFRA
Taxa d'estrangeria global	Demografia	2000-2019	Anual	Si	XIFRA
Taxa de coneixement del català	Demografia	1986-2011	Quinquenal	Si	XIFRA
Taxa de titulats universitaris	Demografia	1986-2011	Quinquenal	No	XIFRA
Taxa bruta de mortalitat	Salut	2000-2020	Anual	Si	XIFRA
Taxa bruta de natalitat	Salut	2000-2020	Anual	Si	XIFRA
Taxa general de fecunditat	Salut	2000-2020	Anual	Si	XIFRA
Índex sintètic de fecunditat	Salut	2000-2020	Anual	Si	XIFRA
Creixement natural	Salut	2000-2020	Anual	Si	XIFRA
Saldo migratori per 1.000 habitants	Demografia	2000-2019	Anual	Si	XIFRA
Creixement migratori	Demografia	2000-2019	Anual	Si	XIFRA
Ràtio d'immigració exterior	Demografia	2000-2019	Anual	Si	XIFRA
Densitat de biblioteques per 10.000 habitants	Despesa pública	2000-2018	Bianual	Si	XIFRA
Densitat d'arxius i museus per 10.000 habitants	Despesa pública	2000-2018	Bianual	Si	XIFRA
Densitat de sales de cinema i teatres per 10.000 habitants	Despesa pública	2000-2018	Bianual	Si	XIFRA
Poliesportiu cobert per 10.000 habitants	Despesa pública	2000-2018	Bianual	Si	XIFRA
Piscina coberta per 10.000 habitants	Despesa pública	2009-2020	Anual	Si	XIFRA
Camps de futbol, rugbi, etc. per 10.000 habitants	Despesa pública	2009-2020	Anual	Si	XIFRA
Pistes de tennis per 10.000 habitants	Despesa pública	2009-2020	Anual	Si	XIFRA
Residus Kg/hab./dia	Salut	2000-2020	Anual	Si	XIFRA
% recollida selectiva sobre total	Salut	2000-2020	Anual	Si	XIFRA
Emissions de CO2 d'edificis residencials per càpita	Salut	2005-2019	Anual	Si	XIFRA

Emissions de CO2 d'edificis del sector terciari per càpita	Salut	2005-2019	Anual	Si	XIFRA
Emissions de CO2 per tractament de residus sòlids i urbans per càpita	Salut	2005-2019	Anual	Si	XIFRA
Emissions de CO2 del transport per càpita	Salut	2005-2019	Anual	Si	XIFRA
Emissions de CO2 industrial per càpita	Salut	2005-2019	Anual	Si	XIFRA
Oficines de farmàcia per 1.000 habitants	Demografia	2000-2020	Anual	Si	XIFRA
Variació percentual de centres educatius	Despesa pública	1992-2020	Anual	No	XIFRA
Taxa d'ocupació de centres educatius	Despesa pública	1992-2020	Anual	No	XIFRA
Taxa d'ocupació de les aules	Despesa pública	1992-2020	Anual	No	XIFRA
Taxa d'atenció del professorat	Despesa pública	1992-2020	Anual	No	XIFRA
Places en residències per a gent gran	Salut	2001-2020	Anual	Si	XIFRA
Beneficiaris de pensions no contributives de la Seguretat Social	Economia	2001-2020	Anual	No	XIFRA
Policies locals per 1.000 habitants	Despesa pública	2000-2020	Anual	Si	XIFRA
Taxa d'habitatges secundaris	Economia	1986-2011	Quinquenal	No	XIFRA
Taxa d'habitatges vacants	Economia	1986-2011	Quinquenal	No	XIFRA
Taxa d'habitatges de lloguer	Economia	1986-2011	Quinquenal	No	XIFRA
Taxa d'habitatges hipotecats	Economia	1986-2011	Quinquenal	No	XIFRA
Taxa d'habitatges de menys de 50 m²	Economia	1986-2011	Quinquenal	No	XIFRA
Taxa d'habitatges de menys de 60 m²	Economia	1986-2011	Quinquenal	No	XIFRA
Taxa d'habitatges de més de 100 m²	Economia	1986-2011	Quinquenal	No	XIFRA
Taxa d'habitatges de més de 105 m²	Economia	1986-2011	Quinquenal	No	XIFRA
Cementiris per 10.000 habitants	Despesa pública	2017-2020	Anual	Si	XIFRA
Variació percentual de turisme	Economia	2015-2019	Anual	No	XIFRA

Places turístiques per 1.000 habitants	Economia	2015-2019	Anual	No	XIFRA
Oficines de turisme per 1.000 habitants	Despesa pública	2000-2020	Anual	Si	XIFRA
Centres de cotització. Variació percentual	Mercat Laboral	2008-2020	Anual	No	XIFRA
Índex d'especialització	Mercat Laboral	2008-2020	Anual	Si	XIFRA
Grandària mitjana	Mercat Laboral	2008-2020	Anual	Si	XIFRA
Variació percentual d'entitats financeres	Economia	1987-2001	Anual	No	XIFRA
Índex de motorització per 1.000 habitants	Economia	2000-2002	Anual	Si	XIFRA
Entitats per 1.000 habitants	Economia	1987-2001	Anual	No	XIFRA
Renda familiar disponible. Variació percentual	Economia	2000-2018	Anual	Si	XIFRA
RBFD per habitant	Economia	2000-2018	Anual	No	XIFRA
RBFD per habitant de 16 anys i més	Economia	2000-2018	Anual	No	XIFRA
PIB per habitant	Economia	2001-2019	Anual	No	XIFRA
PIB per habitant de 16 anys i més	Economia	2001-2019	Anual	No	XIFRA
Base imposable per declarant (euros)	Economia	2000-2019	Anual	Si	XIFRA
Quota resultant per declarant (euros)	Economia	2000-2019	Anual	Si	XIFRA
Variació percentual del IBI	Economia	2000-2019	Anual	Si	XIFRA
Percentatge d'assalariats per subsector econòmic	Mercat Laboral	2008-2020	Anual	Si	XIFRA
Assalariats per trams d'empresa	Mercat Laboral	2009-2020	Anual	Si	XIFRA
Taxa d'ocupació estimada per sexe	Mercat Laboral	2012-2020	Anual	Si	XIFRA
Taxa d'atur registral	Mercat Laboral	2005-2020	Anual	Si	XIFRA
Taxa d'atur registral, estimada per sexe	Mercat Laboral	2005-2020	Anual	Si	XIFRA
Taxa d'atur registral, estimada per sexe, evolució	Mercat Laboral	2008-2020	Anual	Si	XIFRA

Taxa d'atur registral, estimada per edats	Mercat Laboral	2005-2020	Anual	Si	XIFRA
Estructura	Mercat Laboral	2006-2020	Anual	Si	XIFRA
Contractació registrada. Variació percentual	Mercat Laboral	2006-2020	Anual	No	XIFRA
Taxa de contractació temporal	Mercat Laboral	2006-2020	Anual	No	XIFRA
Taxa de contractació a temps parcial	Mercat Laboral	2006-2020	Anual	No	XIFRA
Taxa d'activitat	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'inactivitat	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'ocupació	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'atur censal	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'activitat segons edat i sexe	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'inactivitat segons edat i sexe	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'ocupació segons edat i sexe	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'atur censal segons edat i sexe	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Índex de dependència econòmica	Mercat Laboral	1996-2001	Quinquenal	Si	XIFRA
Taxa d'autocontenció	Demografia	1996-2011	Quinquenal	No	XIFRA
Taxa d'autosuficiència	Demografia	1996-2011	Quinquenal	No	XIFRA
Codi municipi	Classificació	1975-2020	Anual	Si	Dades Obertes Catalunya
Codi comarca	Classificació	1975-2020	Anual	Si	Dades Obertes Catalunya

Codi provincia	Classificació	1975-2020	Anual	Si	Dades Obertes Catalunya
Vacunació per la COVID	Salut	2019-2021		Si	Dades Obertes Catalunya
UTM X	Geografia	1975-2020	Anual	Si	Dades Obertes Catalunya
UTM Y	Geografia	1975-2020	Anual	Si	Dades Obertes Catalunya
Longitud	Geografia	1975-2020	Anual	Si	Dades Obertes Catalunya
Latitud	Geografia	1975-2020	Anual	Si	Dades Obertes Catalunya
Agermanaments de municipis	Despesa pública	?	Sense anys	No	Dades Obertes Catalunya
Preu mitjà del lloguer d'habitatges per municipi	Economia	2007-2021	Anual	Si	Dades Obertes Catalunya
Caps de municipi de Catalunya georeferenciats	Geografia	1975-2020	Anual	Si	Dades Obertes Catalunya
Consum d'energia elèctrica per municipis i sectors de Catalunya	Economia	2013-2020	Anual	Si	Dades Obertes Catalunya
Llistat d'agrupacions de municipis per a funcionaris d'habilitació nacional	Despesa pública	?	Sense anys	No	Dades Obertes Catalunya
Codis postal per municipis de Catalunya	Classificació	1975-2020	Anual	Si	Dades Obertes Catalunya
Població de Catalunya per municipi, rang d'edat i sexe	Demografia	2019-2020	Anual	Si	Dades Obertes Catalunya
Dades de l'activitat administrativa a la DGOTU per municipis	Economia	2013-2020	Anual	Si	Dades Obertes Catalunya
Registre central de població del CatSalut: població per municipi	Demografia	2012-2020	Anual	Si	Dades Obertes Catalunya
Municipis compresos en les àrees bàsiques policials (ABP)	Classificació	1975-2020	Anual	Si	Dades Obertes Catalunya
Dades històriques de població dels municipis de Catalunya	Demografia	1990-2019	Anual	Si	Dades Obertes Catalunya

Consum de gas natural canalitzat per municipis i sectors de Catalunya	Economia	2013-2020	Anual	Si	Dades Obertes Catalunya
Relació Municipis i Partit Judicial de Catalunya	Classificació	1975-2020	Anual	Si	Dades Obertes Catalunya
Dades setmanals de COVID-19 per comarca	Salut	2019-2020	Diàri	No	Dades Obertes Catalunya
Registre d'entitats, serveis i establiments socials (serveis socials bàsics i especialitzats)	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Registre de casos de COVID-19 a Catalunya per municipi i sexe	Salut	2019-2020	Diàri	Si	Dades Obertes Catalunya
Nombre de persones amb discapacitat per tipologia. Municipis de més de 20.000 habitants	Demografia	2015-2020	Anual	No	Dades Obertes Catalunya
Nombre de persones amb discapacitat per grups d'edat. Municipis de més de 20.000 habitants	Demografia	2015-2020	Anual	No	Dades Obertes Catalunya
Nombre de persones amb discapacitat per grau. Municipis de més de 20.000 habitants	Demografia	2015-2020	Anual	No	Dades Obertes Catalunya
Registre de les associacions de voluntaris de protecció civil	Incidéncies i emergències	?	Sense anys	Si	Dades Obertes Catalunya
Ocupació comercial de ciutats catalanes (IATC)	Economia	?	Sense anys	No	Dades Obertes Catalunya
Dades generals dels ens locals de Catalunya	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Instal·lacions d'autoconsum elèctric	Economia	?	Sense anys	No	Dades Obertes Catalunya
Equitat digital als centres educatius de Catalunya	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Centres tècnics de tacògraf CTT	Despesa pública	?	Sense anys	No	Dades Obertes Catalunya
Actuacions dels Bombers de la Generalitat	Incidéncies i emergències	2010-2020	Anual	Si	Dades Obertes Catalunya
Registre d'establiments del sector de l'alimentació animal i de l'àmbit dels SANDACH	Economia	?	Sense anys	Si	Dades Obertes Catalunya

Personal docent en centres públics titularitat del Departament d'Educació	Despesa pública	2018-2020	Anual	Si	Dades Obertes Catalunya
Llistat de beneficiaris de les subvencions en espècie per a la contractació d'espectacles professionals inclosos en el catàleg d'espectacles Programa.cat	Despesa pública	2012-2020	Anual	Si	Dades Obertes Catalunya
Àrees i zones de referència en matèria de política d'habitatge a Catalunya	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Incendis forestals a Catalunya. Anys 2011-2020	Incidéncies i emergències	2011-2020	Anual	Si	Dades Obertes Catalunya
Espais esportius i complementaris censats al Cens d'Equipaments Esportius de Catalunya (CEEC)	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Oferta educativa dels estudis post-obligatoris	Despesa pública	2021	Anual	Si	Dades Obertes Catalunya
Òrgans judicials d'Espanya	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Partits judicials d'Espanya	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Participació catalana en programes finançats per la Unió Europea (2014-2020)	Despesa pública	?	Sense anys	Si	Dades Obertes Catalunya
Certificats d'eficiència energètica d'edificis: tancaments	Economia	?	Sense anys	Si	Dades Obertes Catalunya
Indicadores de renta media y mediana. Renta neta mitja per persona	Economia	2015-2019	Anual	Si	INE
Indicadores de renta media y mediana. Renda neta mijta per llar	Economia	2015-2019	Anual	Si	INE
Indicadores de renta media y mediana. Mitja de la renta per unitat de consum	Economia	2015-2019	Anual	Si	INE
Indicadores de renta media y mediana. Mediana de la renta per unitat de consum	Economia	2015-2019	Anual	Si	INE

Indicadores de renta media y mediana. Renda bruta mitja per persona	Economia	2015-2019	Anual	Si	INE
Indicadores de renta media y mediana. Renda bruta mitja per llar	Economia	2015-2019	Anual	Si	INE
Distribución por fuente de ingresos. Renda bruta mitja per persona	Economia	2015-2019	Anual	Si	INE
Distribución por fuente de ingresos. Fonts d'ingressos: salari	Economia	2015-2019	Anual	Si	INE
Distribución por fuente de ingresos. Fonts d'ingressos: pensions	Economia	2015-2019	Anual	Si	INE
Distribución por fuente de ingresos. Fonts d'ingressos: prestacions per atur	Economia	2015-2019	Anual	Si	INE
Distribución por fuente de ingresos. Fonts d'ingressos: altres prestacions	Economia	2015-2019	Anual	Si	INE
Distribución por fuente de ingresos. Fonts d'ingressos: altres ingressos	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo. Poblacion amb ingressos per unitat de consum per sota dels 5.000€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo. Poblacion amb ingressos per unitat de consum per sota dels 7.500€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo.	Economia	2015-2019	Anual	Si	INE

Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota dels 5.000€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota dels 7.500€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota dels 10.000€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota dels 5.000€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota dels 7.500€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota dels 10.000€	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo. Poblacion amb ingressos per unitat de consum per sota del 40% de la mediana	Economia	2015-2019	Anual	Si	INE

Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo. Poblacion amb ingressos per unitat de consum per sota del 50% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo. Poblacion amb ingressos per unitat de consum per sota del 60% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo. Poblacion amb ingressos per unitat de consum per sota del 140% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo. Poblacion amb ingressos per unitat de consum per sota del 160% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo. Poblacion amb ingressos per unitat de consum per sota del 200% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota del 40% de la mediana	Economia	2015-2019	Anual	Si	INE

Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota del 50% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota del 60% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota del 140% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota del 160% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y tramos de edad. Poblacion amb ingressos per unitat de consum per sota del 200% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota del 40% de la mediana	Economia	2015-2019	Anual	Si	INE

Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota del 50% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota del 60% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota del 140% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota del 160% de la mediana	Economia	2015-2019	Anual	Si	INE
Porcentaje de población con ingresos por unidad de consumo por debajo/encima de determinados umbrales relativos por sexo y nacionalidad. Poblacion amb ingressos per unitat de consum per sota del 200% de la mediana	Economia	2015-2019	Anual	Si	INE
Indice de Gini	Economia	2015-2019	Anual	Si	INE
Ratio 80/20	Economia	2015-2019	Anual	Si	INE
Edat mitja de la població	Economia	2015-2019	Anual	Si	INE
Percentatge de població menor de 18 anys	Economia	2015-2019	Anual	Si	INE
Percentatge de poblacio de 65 i més anys	Economia	2015-2019	Anual	Si	INE
Dimensió mitja de la llar	Economia	2015-2019	Anual	Si	INE
Percentatge de llars unipersonals	Economia	2015-2019	Anual	Si	INE
Recompte població	Economia	2015-2019	Anual	Si	INE

Font: Taula d'elaboració pròpia

8.6 Annex VI: Fonts d'informació utilitzades en l'Article II

ID	Variables	Description	Font de dades	Nivell
		CASES		
1	Housing ID	ID de cada vivenda	Habitaclia	Individual
38	Price	Precio de la vivienda mensual	Habitaclia	Individual
39	Annual price	Precio de la vivienda anual	Habitaclia	Individual
40	Square meters	Metros cuadrados de la vivienda	Habitaclia	Individual
41	Toilets	Numero total de baños	Habitaclia	Individual
42	Bedrooms	Numero total de habitacines	Habitaclia	Individual
43	Municipality	Municipio identificado de la vivienda	Habitaclia	Individual
44	Street	Calle identificada de la vivienda	Habitaclia	Individual
45	Latitude	Latitud de la vivienda	Own	Individual
46	Longitude	Longitud de la vivinda	Own	Individual
		DADES DEMOGRÀFIQUES I MEDI AMBIENTALS		
49	Total rents	Nombre de contractes que s'han donat d'alta al Registre de fiances en el període considerat	Dades obertes	Municipality
50	Mean rent house	Import del lloguer mensual mitjà. Dades anonimitzades per municipis amb menys de 6 habitatges registrats	Dades obertes	Municipality
51	Energy certificate ID	Identificador del tràmit del certificat d'eficiència energètica	Dades obertes	Municipality
52	Empty house tax	Municipis on s'aplica l'impost sobre els habitatge buits creat per la Llei 14/2015 i que han estat determinats per la Llei 4/2016, del 23 de desembre	Dades obertes	Municipality
3	Area	Metros cuadrados de cada sección censal	ICGC	District
5	Boundary 40 median income	Porcentaje de personas con ingressos por debajo del 40% de la mediana	INE	District
6	Boundary 50 median income	Porcentaje de personas con ingressos por debajo del 50% de la mediana	INE	District
7	Boundary 60 median income	Porcentaje de personas con ingressos por debajo del 60% de la mediana	INE	District

8	Boundary 140 median income	Porcentaje de personas con ingressos por debajo del 140% de la mediana	INE	District
9	Boundary 160 median income	Porcentaje de personas con ingressos por debajo del 160% de la mediana	INE	District
10	Boundary 200 median income	Porcentaje de personas con ingressos por debajo del 200% de la mediana	INE	District
11	Income: net per person	Ingresso de la renda neto por persona	INE	District
12	Income: gross per person	Ingresso de la renda bruta por persona	INE	District
13	Income: net household	Ingresso de la renda neta del hogar	INE	District
14	Income: household gross	Ingresso de la renda bruta del hogar	INE	District
15	Income: median unit of consumption	Mediana del ingresso por unidad de consumo	INE	District
16	Income: salary	Ingresso de la renda por salario	INE	District
17	Gini	Índice de Gini	INE	District
18	P80/20	Ratio 80/20	INE	District
19	Population	Población que vive en cada distrito	INE	District
20	Middle Ages	Edad media de cada distrito	INE	District
21	Mean home size	Dimensión de las viviendas de cada distrito	INE	District
22	Single-person homes	Porcentage de viviendas unipersonales de cada distrito	INE	District
23	Income: net per neighboring person	Media del ingresso de la renda neto por persona de los distritos vecinos	INE	District
24	Income: gross per neighboring person	Media del ingresso de la renda bruta por persona de los distritos vecinos	INE	District
25	Income: net neighboring household	Media del ingresso de la renda neta del hogar de los distritos vecinos	INE	District
26	Income: gross from neighboring household	Media del ingresso de la renda bruta del hogar de los distritos vecinos	INE	District
27	Income: median neighboring unit of consumption	Mediana del ingresso por unidad de consumo de los distritos vecinos	INE	District
28	Income: neighboring salary	Media del ingresso de la renda por salario de los distritos vecinos	INE	District
47	Urban (Y/N)	El distrito de la vivienda és en una zona urbana? #Metodologia OCDE	Own	District
48	Metropolitan Area (Y/N)	El distrito de la vivienda esta dentro de la Area Metropolitana? #Metodologia OCDE	Own	District
53	Can rent? (Y/N)	Can you rent the house?	Own	Individual

54	Can rent and save? (Y/N)	Can you rent and save the 30% of the rent?	Own	Individual
55	Capital (Y/N)	Is it a Capital of province?	Own	Individual
		VEGETACIÓ I PRESÈNCIA DE VERD		
2	NDVI	El Índice de Vegetación de Diferencia Normalizada (NDVI)en un radio de 500 metros de cada casa. Valores possibles: -1;-1.	ICGC	District
		CONTAMINANTS		
29	PM10	Media de las pequeñas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro aerodinámico es menor que 10 µm	GRECS	ABS
30	Old WHO limit PM10 (Y/N)	Sobrepasa los límites antiguos de la OMS?	Own	ABS
31	New WHO limit PM10 (Y/N)	Sobrepasa los límites actuales de la OMS?	Own	ABS
32	NO2	Media del compuesto químico formado por los elementos nitrógeno y oxígeno, uno de los principales contaminantes entre los varios óxidos de nitrógeno. El dióxido de nitrógeno es de color marrón-amarillento.	GRECS	ABS
33	Old WHO limit NO2 (Y/N)	Sobrepasa los límites antiguos de la OMS?	Own	ABS
34	New WHO limit NO2 (Y/N)	Sobrepasa los límites actuales de la OMS?	Own	ABS
35	O3	Media del gas altamente reactivo, es capaz de absorber luz infrarroja y ultravioleta, contribuyendo al efecto invernadero y proporcionando protección contra la luz ultravioleta del Sol.	GRECS	ABS
36	Old WHO limit O3 (Y/N)	Sobrepasa los límites antiguos de la OMS?	Own	ABS
37	New WHO limit O3 (Y/N)	Sobrepasa los límites actuales de la OMS?	Own	ABS
		SEGURETAT		
4	Known crimes	Nombre total de crims coneguts pels Mossos d'Esquadra que s'inclou:	Mossos d'Esquadra	ABP

Font: Taula d'elaboració pròpia.