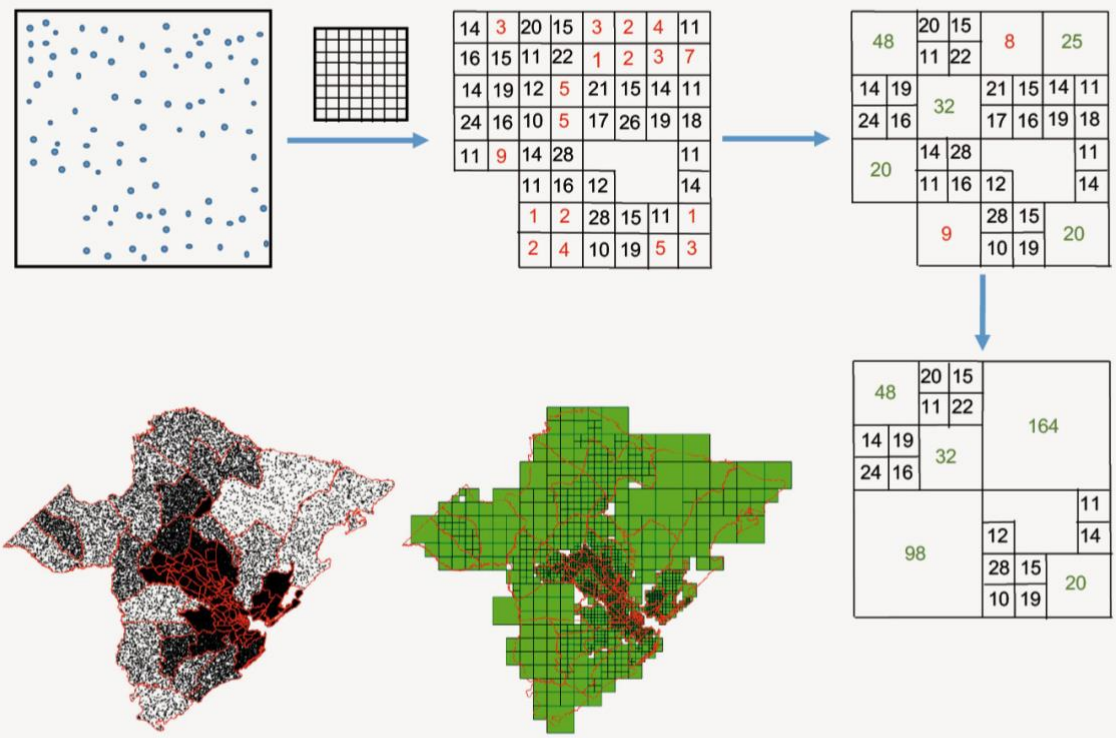
TESI DOCTORAL

Aplicacions de visualització d'informació georeferenciada

Raymond Lagonigro Bertran



ESCOLA
DE DOCTORAT
UVIC·UCC

TESI DOCTORAL

Aplicacions de visualització d'informació georeferenciada

Raymond Lagonigro Bertran

Directors de tesi: Ramon Oller Piqué i Joan Carles Martori Cañas

Programa de doctorat: Dret, Economia i Empresa

2020

ESCOLA DE DOCTORAT
UVIC | UVIC·UCC

Agraïments

Conscient que no seré capaç d'esmentar tots els que, d'una forma o altra, m'han acompanyat i ajudat durant l'elaboració de la tesi, vull, almenys, destacar algunes de les persones o entitats sense les quals aquest treball no hagués estat possible. Ha sigut un llarg procés en el que he après molt de tots els familiars, companys i amics que han participat d'aquest camí i que de forma directa o indirecta m'han ajudat a recorre'l.

Vull començar agraint a en Joan Carles Martori i en Ramon Oller, directors de la tesi, la paciència que han tingut, la confiança que m'han fet sentir en tot moment i, sobretot, el suport que m'han donat per desenvolupar els articles que conformen la tesi, i realitzar-ne l'elaboració final. Gràcies als dos per les reflexions que m'han ajudat a encaminar i anar millorant contínuament aquesta investigació.

Tinc molt presents també els companys del grup de recerca *Data Analysis and Modelling* i a la resta de companys dels dos departaments on realitzo docència i investigació, als quals agraeixo la col·laboració en aquest treball. De les seves aportacions i dels debats i xerrades que hem mantingut en innumerables ocasions n'he après i me n'he enriquit molt. Espero haver estat capaç de reflectir-ho en el meu treball.

Agraeixo a l'Institut d'Estadística de Catalunya l'interès en acceptar el projecte que els vam proposar i la signatura del conveni de col·laboració amb la Universitat de Vic – Universitat Central de Catalunya, "Preservació del secret estadístic en la difusió de dades geocodificades de població", que va ser l'origen de les investigacions que van donar lloc a la tesi.

Vull agrair també l'acollida que em van donar a *Le Laboratoire d'équité environnementale (LAEQ – INRS)* de Montreal amb qui vaig poder col·laborar durant un mes gràcies a un ajut per a estades de recerca fora de Catalunya, de la Universitat de Vic – Universitat Central de Catalunya. Va ser un període curt, però molt enriquidor, que ens va permetre definir la base per a dos dels articles que formen part d'aquesta tesi.

A la UVic-UCC són molts els amics que m'han acompanyat. Més enllà de les qüestions professionals, la seva companyia i escalf han estat també indispensables, sobretot en els moments de dubte.

Però especialment, la més profunda gratitud a la meva família. No hauria arribat fins aquí sense el suport personal que m'han donat tots ells per poder dur a terme aquesta tesi. Als meus germans, que sempre s'han interessat i m'han motivat en aquest procés i, sobretot, als meus pares i a les meves filles, que m'han recolzat tant i m'han transmès la força per tirar endavant, que han estat al meu costat en tot moment i de qui sempre he sentit l'empenta necessària. A ells els dec la culminació d'aquest camí.

A tots, el meu immens reconeixement i gratitud.

Taula de continguts

Resum

En un context on les eines de geolocalització són cada vegada més habituals, els sistemes de recopilació i difusió de dades estadístiques poden integrar noves eines per publicar informació georeferenciada amb una resolució espacial molt precisa. Al mateix temps, aquesta precisió pot generar problemes de confidencialitat de les persones, llars o empreses a qui fan referència aquestes dades, perquè en pot facilitar la identificació. Ens trobem amb un conflicte entre la capacitat per publicar dades amb un nivell de precisió geogràfica molt acurat i, al mateix temps, complir les diferents normatives de confidencialitat en la difusió de dades estadístiques.

Habitualment, les agències d'estadística utilitzen sistemes de recopilació basats en divisions espacials administratives que els permeten fer particions del territori orientades a una millor organització de les dades. Aquestes particions també permeten assegurar confidencialitat, però no sempre són la forma de difusió més adequada quan les dades han de ser utilitzades per realitzar determinades anàlisis espacials.

En aquesta tesi proposem una metodologia alternativa per la distribució de dades que, a partir d'una quadrícula inicial de mida constant, divideix jeràrquicament l'espai per obtenir una quadrícula irregular que agrupa la informació balancejant criteris de confidencialitat i maximització de la precisió espacial. Les quadrícules generades amb aquesta metodologia es basen en un sistema de codificació de les caselles que segueix la nomenclatura proposada per la Oficina d'Estadística de la Unió Europea, per la creació d'un conjunt de dades de població únic per la UE en forma de quadrícula.

Mesurar la importància dels efectes espacials requereix disposar de dades amb la precisió més acurada possible i, si és possible, a una escala espacial propera a la del fenomen estudiat. En aquest sentit, aquesta tesi també presenta dos estudis de dos fenòmens espacials utilitzant la informació disponible amb els sistemes actuals basats en seccions censals. En ambdós estudis els fenòmens investigats estan caracteritzats a una resolució espacial molt precisa, mentre

que les dades disponibles per detectar-ne possibles correlacions espacials estan agregades en unitats espacials a molta menys resolució. Per tal de comparar els diferents processos cal transformar les dades per tenir-les a una mateixa escala. Aquestes transformacions poden esbiaixar els resultats de les anàlisis. La metodologia proposada està orientada a permetre realitzar aquests tipus d'estudis amb dades més precises.

Aquesta metodologia està desenvolupada en una llibreria per l'entorn estadístic R, per tal de produir dades per a aquest entorn, però també exportables a formats estàndard per a qualsevol altre sistema d'informació geogràfica. Aquesta llibreria està publicada al repositori habitual de programari R i pot ser descarregada i instal·lada per publicar dades basades en quadrícula a partir de qualsevol conjunt de dades espacials. En el futur, aquesta llibreria permetrà que els instituts d'estadística publiquin informació amb una millor precisió geogràfica per la realització d'anàlisis espacials.

Abstract

In a context where geolocation tools are becoming more common, statistical data collection and dissemination systems should include new tools to release georeferenced information with precise spatial resolution. At the same time, this accuracy may trigger confidentiality issues for individuals, households or companies to whom this data refers, because it can enable their re-identification. There is a conflict between the dissemination of data at an accurate level of geographical precision and, at the same time, fulfill the different confidentiality regulations on the distribution of statistical data.

Statistical offices typically use collection systems based on administrative boundaries to divide the territory and assist data organization. These partitions are also useful to ensure confidentiality but may not be the most appropriate form of dissemination when data should be used to study spatial effects of different phenomena.

In this thesis we propose an alternative methodology for data distribution which takes an initial constant size grid, and successively divides the space to obtain an irregular grid that groups the information balancing confidentiality and spatial resolution. Grids produced using this methodology are based on a cell coding system following the indications proposed by the Statistical Office of the European Union, to represent the main characteristics of the population on a unique constant grid.

Measuring the importance of spatial effects requires having data at an accurate spatial resolution and, when possible, on a scale close to that of the phenomenon studied. In this regard, this thesis also presents two studies of spatial phenomena conducted with the information available using the current systems based on census tracts. In both studies the investigated phenomena are characterized at a very precise spatial resolution, while the available data to detect possible spatial correlations is aggregated in spatial units with much less resolution. In order to compare the different processes, data must be transformed to the same geographic scales. These transformations may have undesired effects on the

results of the analyses. The proposed methodology undertakes the production of more accurate spatial datasets to avoid those possible skewing effects.

The methodology is fully developed on a library in the statistical software R, in order to produce data for this environment, but it can also be exported to standard data formats for any other geographic information system. The library is published in the usual R software repository and can be downloaded and installed to publish grid-based datasets from any spatial point data. In the future, this library will allow statistical offices to provide more accurate spatial information to perform spatial analysis.

Llistat de figures i taules

Llistat Abreviatures

GDPR: General Data Protection Regulation

GIS: Geographic Information System

GWR: Geographically Weighted Regression

LISA: Local Indicators of Spatial Autocorrelation

MAUP: Modifiable Areal Unit Problem

OGC: Open Geospatial Consortium Inc.

OLS: Ordinary Least Squares

SAC: Spatial Autoregressive Confused

SAR: Spatial Auto Regression

SDC: Statistical Disclosure Control

SDM: Spatial Durbin Model

SDEM: Spatial Durbin Error Model

SE: Symbology Encoding

SEM: Spatial Error Model

SLD: Styled Layer Descriptor

SLX: Spatial Lag X

Compendi de publicacions

Lagonigro, R., Oller, R., Martori J.C. (2017). A Quadtree approach based on European geographic grids: reconciling data privacy and accuracy. SORT Statistics and Operations Research Transactions, 41(1), 139-158. doi: 10.2436/20.8080.02.55.

Web of Science Journal Citation Reports (JCR) Impact factor (2017): 1.344 – Q2 Statistics & Probability

Lagonigro, R., Martori, J.C., Apparicio P. (2018). Environmental noise inequity in the city of Barcelona. Transportation Research Part D: Transport and Environment, 63, 309-319. doi: 10.1016/j.trd.2018.06.007.

Web of Science Journal Citation Reports (JCR) Impact factor (2018): 4.051 – Q1 Transportation Science and Technology

Lagonigro, R., Martori, J.C., Apparicio P. (2020). Understanding Airbnb spatial distribution in a southern European city: The case of Barcelona. Applied geography, 115, doi: 10.1016/j.apgeog.2019.102136.

Web of Science Journal Citation Reports (JCR) Impact factor (2019): 3.508 – Q1 Geography

Acceptat, pendent de data de publicació:

Lagonigro, R., Oller, R., Martori J.C. (2020). AQuadtree: an R package for quadtree anonymization of point data. The R Journal.

Web of Science Journal Citation Reports (JCR) Impact factor (2019): 3.312 – Q1 Statistics and Probability; Q2 Computer Science, Interdisciplinary Applications

Altres

Lagonigro, R., Oller, R., Martori J.C. Preservació del secret estadístic en la difusió de dades geocodificades. Accèssit al premi de protecció de Dades en el Disseny 2016. Autoritat Catalana de Protecció de Dades.

Lagonigro, R., Martori, J.C., Apparicio P. (2018). AirBnB como mecanismo de gentrificación: el caso de Barcelona. International Conference on Regional Science - XLIV Reunión de Estudios Regionales. Valencia (Spain)

Presentació

Aquesta tesi com a compendi d'articles s'emmarca en l'anàlisi i la visualització de dades espacials reflexionant sobre la privacitat de la informació i, alhora, la necessitat de disposar d'informació acurada per tal de realitzar inferències estadístiques en els estudis espacials. Aquesta reflexió s'inicia a partir d'un conveni de col·laboració entre l'Institut d'Estadística de Catalunya i el grup de recerca *Data Modelling and Analys*is de la Universitat de Vic – Universitat Central de Catalunya. Concretament es centra en els processos de difusió de dades estadístiques de les entitats que realitzen la recollida, producció i distribució d'aquestes dades per tal que els investigadors puguin disposar d'informació precisa i, si convé, adequada a cada tipus d'estudi. Les dades localitzades geogràficament tenen una gran riquesa analítica i, en molts casos, permeten detectar i explicar determinats fenòmens espacials subjacents. Sovint, els investigadors depenen de les agències nacionals d'estadística, o altres organismes amb capacitat per recollir i difondre dades demogràfiques, econòmiques, sanitàries, o d'altres àmbits per a un determinat territori. La fiabilitat de les inferències que puguin derivar-se d'aquestes dades dependrà en gran mesura de que siguin prou ajustades a la realitat territorial i que l'escala espacial sigui el màxim d'específica.

Actualment les tècniques i mètodes de processament espacial permeten obtenir cada vegada més dades geolocalitzades i, per tant, que es disposi d'informació estadística a un nivell geogràfic molt detallat, cosa que dóna als investigadors la possibilitat d'estudiar fenòmens i interaccions espacials amb més precisió. Al mateix temps, la difusió de dades referenciades geogràficament genera problemes de confidencialitat de les persones, llars, empreses o qualsevol altre element particular a qui pertanyen aquestes dades perquè en pot facilitar la identificació. Hi ha doncs un conflicte entre la necessitat de disposar de dades amb el nivell de precisió més acurat possible i, al mateix temps, garantir la confidencialitat dels individus als que fan referència les dades.

La confidencialitat en la difusió d'informació estadística és una qüestió que ha estat molt analitzada i actualment és objecte de regulacions tant nacionals com supranacionals i, per tant, es disposa d'un marc de referència molt clar. El dret a la privacitat està reconegut a l'Article 12 de la Declaració Universal dels Drets Humans (1948) i l'Article 8 de la Convenció Europea dels Drets Humans (1950), i està protegit per les constitucions de la gran majoria de països. Així, la Comissió Europea, defineix la regulació que afecta la protecció de dades individuals a totes les institucions i agències d'estadística de la Unió Europea. En base a aquests criteris, els diversos països poden, a més, definir les seves pròpies regulacions. En l'àmbit de les dades estadístiques espacials, les restriccions de confidencialitat generalment s'apliquen de forma que les dades d'observacions individuals s'agrupen creant àrees geogràfiques definides a partir d'uns llindars en el nombre mínim d'observacions en aquestes àrees. Hi haurà doncs una relació directe entre els llindars que es defineixin i la mida de les zones geogràfiques creades per difondre informació; és a dir, si els llindars són prou alts, les àrees resultants seran suficientment grans per garantir que la informació difosa no permet identificar les observacions individuals.

Així com la privacitat és un aspecte bàsic a tenir en compte abans de difondre dades estadístiques, la precisió de les dades en l'àmbit de la recerca acadèmica és una qüestió no tant clarament definida quedant normalment en un segon terme. Sovint, moltes variables econòmiques, per exemple, només estan disponibles per determinades divisions administratives del territori, com poden ser les seccions censals. Per tant, les dades disponibles representen els valors mitjos per les variables d'interès a cada zona geogràfica. L'ús de dades agregades per estudiar les interaccions espacials entre determinats fenòmens pot comportar el que es coneix com a errors de biaix ecològic: els resultats de les interaccions investigades i dels models obtinguts poden dependre de la forma de les àrees geogràfiques tant a nivell de la mida com de les zones administratives que representen. Per tant, estudis realitzats a escales diferents o bé amb unes altres divisions administratives podrien haver donat lloc a resultats diferents.

En el terreny de l'econometria espacial, es poden destacar dues particularitats en els fenòmens que evidencien interacció espacial. D'una banda, les diverses observacions poden definir una similitud o dissimilitud en el fenomen estudiat en funció de la distància que hi ha entre elles (autocorrelació espacial). D'altra banda, tot i observar una dependència espacial entre les observacions, la influència dels factors explicatius en el model economètric, pot variar entre les diferents zones geogràfiques (heterogeneïtat espacial). En aquest sentit, la tesi aporta dos estudis d'anàlisi espacial, de dos tipus de fenòmens diferents, a la ciutat de Barcelona; en el primer cas en el domini de l'autocorrelació espacial i en el segon en el de la heterogeneïtat espacial. Ambdós estudis s'han realitzat a partir de les dades socioeconòmiques disponibles, actualment només a escala de secció censal. Per tant, en tots dos fenòmens pot haver tingut especial importància la definició de les àrees geogràfiques. Aquesta qüestió es planteja a l'apartat de discussió d'aquesta tesi.

Així doncs, els quatre articles que conformen aquest compendi tracten, d'una banda, la producció de dades estadístiques espacials i, de l'altra, les investigacions espacials amb els sistemes actuals de producció d'aquestes dades. En el primer article, *A Quadtree approach based on European geographic grids: reconciling data privacy and accuracy*, es revisen els criteris de privacitat per la producció de dades estadístiques espacials i es proposa un marc als agents responsables de custodiar i promoure l'ús d'aquestes dades, amb un doble objectiu. D'una banda que permeti assegurar la confidencialitat de la informació distribuïda i, de l'altra, que aquesta informació sigui el més acurada possible per tal que els resultats de les anàlisis espacials que se'n derivin estiguin menys esbiaixats. Els dos articles següents presenten dos exemples d'investigacions espacials a la ciutat de Barcelona, amb l'objectiu d'explicar les interrelacions entre dos fenòmens dels que es disposa d'informació geogràfica molt acurada, i les dades socioeconòmiques i demogràfiques distribuïdes amb els sistemes de producció actuals. En concret, l'article *Environmental noise inequity in the city of Barcelona*, analitza l'exposició al soroll ambiental de diversos grups de població considerats vulnerables i determina si aquests grups poden estar afectats per desigualtats en quant a molèsties de contaminació

acústica. En el tercer article, *Understanding Airbnb spatial distribution in a southern European city: The case of Barcelona*, s'explora la heterogeneïtat espacial en les interrelacions entre la densitat d'allotjaments turístics privats *Airbnb* i les característiques socioeconòmiques i demogràfiques de la població, així com l'atractiu turístic de les diverses zones de la ciutat. Finalment, el darrer article, *AQuadtree: an R Package for Quadtree Anonymization of Point Data*, descriu la implementació de la metodologia de producció de dades en una llibreria de programació en llenguatge R, publicada en codi obert a través d'un repositori públic. En la posterior discussió dels resultats aportats s'analitza, mitjançant l'ús d'aquesta llibreria la influència de l'escala utilitzada en la producció de les dades estadístiques i s'introdueixen algunes qüestions que planteja el possible biaix ecològic dels resultats obtinguts.

La tesi s'ha estructurat de la següent forma: en el capítol 1 es fa una revisió del marc teòric en els àmbits del tractament i visualització de la informació geogràfica des de la perspectiva de la confidencialitat estadística i, posteriorment, es contextualitzen els articles presentats a la tesi; a continuació el capítol 2 inclou els quatre articles publicats; en el capítol 3 es discuteixen els resultats obtinguts en els 4 articles, de forma transversal; finalment, el capítol 4 extreu conclusions i suggereix algunes línies de recerca futures.

Introducció

1 Introducció

Cada vegada més, les dades recollides per dur a terme estudis estadístics contenen informació geogràfica que permet situar-les en l'espai de forma molt precisa. Trobar sistemes adequats per visualitzar les dades espacials pot ajudar a interpretar determinats fenòmens que, en alguns casos, revelen interrelacions o processos lligats a l'origen d'aquests fenòmens. Situar la informació en el seu context espacial de forma adequada, per exemple, ofereix una visió molt més sintètica dels esdeveniments i pot donar indicis de quines són les eines estadístiques més apropiades per aprofundir en l'estudi dels fenòmens espacials. Més enllà de la visualització de les dades, l'anàlisi espacial explora com s'entrellacen les dades referenciades geogràficament, amb les seves característiques o amb els processos que les han produït. Per iniciar una investigació espacial, una primera anàlisi espacial descriptiva permet detectar problemes o mancances en les dades, o invalidar algunes hipòtesis necessàries per aplicar determinats mètodes estadístics. Així mateix, la visualització i l'anàlisi descriptiva de les dades permet definir l'estructura geogràfica més adequada a l'àrea d'estudi per tal de modelar els efectes espacials entre les observacions. Posteriorment, mitjançant tècniques d'econometria espacial es pot examinar i modelar la interacció entre les observacions analitzades a les diverses zones (autocorrelació espacial) i l'estructura espacial que defineixen aquestes observacions (heterogeneïtat espacial).

1.1 Visualització d'informació geogràfica

Actualment és molt fàcil visualitzar mapes digitals utilitzant ordinadors o dispositius mòbils. Sistemes com *Google Maps* i *Apple Maps* o el projecte de codi obert *OpenStreetMap* permeten l'ús de cartografia digital de forma molt detallada i precisa. Visvalingam (1994) defineix la Geovisualització com "l'ús de tecnologia informàtica per explorar dades espacials de forma visual... i l'ús de gràfics d'ordinador per aprofundir en el coneixement de dades". La geovisualització defineix un nou paradigma en el tractament de les dades

geogràfiques, que permet detectar fenòmens i interrelacions que inicialment poden quedar encobertes en volums de dades cada vegada grans i complexes (Orford, 2005).

Els sistemes d'informació geogràfica (GIS) proveeixen eines cartogràfiques per tal de crear mapes per visualitzar dades geogràfiques. Aquestes eines de cartografia digital permeten afegir capes d'informació sobreposant-les en mapes digitals per tal de mostrar contextos geogràfics. Els GIS es basen principalment en dues especificacions del *Open Geospatial Consortium* (OGC) que estableixen els reglaments per crear mapes que permetin visualitzar dades: el descriptor de capa estilitzada (SLD) (Lupp, 2007) i el codificador de símbols (SE) (Müller, 2006). Aquestes dues especificacions defineixen les regles bàsiques per renderitzar les dades geogràfiques i estructurar-les en diverses capes visuals damunt del mapa.

La semiologia cartogràfica engloba les teories dels símbols cartogràfics i la seva utilització, i defineix el conjunt de regles per tal transmetre la informació de la forma més clara possible mitjançant imatges cartogràfiques (Kraak & Ormeling, 2011). La representació cartogràfica és la principal forma gràfica utilitzada en l'àmbit de la geografia per representar de forma clara i objectiva les relacions entre els fenòmens espacials. L'ús de les diverses variables visuals ha de permetre que la informació es transmeti de forma clara i inambigua facilitant-ne la comprensió.

Així doncs, els mapes permeten integrar la dimensió espacial per facilitar la comprensió de determinats fenòmens i les seves interrelacions. La finalitat del mapa ha de ser representar les dades geogràfiques de la forma més adequada. Les formes, orientacions, textures, colors o mides, per exemple, permetran destacar diferents propietats dels objectes representats. Així, sintetitzar la informació en mapes és un dels primers passos en la investigació espacial, no només per facilitar-ne la comprensió a l'hora de fer-ne difusió, sinó també com part de la pròpia anàlisi.

En l'àmbit de les agències oficials d'estadística, els mapes jugaran un paper bàsic en totes les fases de la recol·lecció i difusió de les dades geolocalitzades

(United Nations, 2009). En la fase preparatòria, els mapes permetran definir les zones de forma que s'asseguri la cobertura del territori de la manera més adequada tenint en compte la posterior obtenció i emmagatzematge. També permetran monitoritzar la fase de recollida de dades i definir les millors estratègies en la planificació i control de recol·lecció. Però on més impacte i utilitat mostren els mapes és en la presentació, anàlisi i difusió de les dades, permetent, per exemple, la detecció de patrons locals de determinats indicadors demogràfics i socioeconòmics. Els mapes jugaran també un paper molt important en l'anàlisi de polítiques de planificació del territori tant en el sector públic com privat.

1.2 Dades espacials

Les tècniques i mètodes en l'àmbit de l'anàlisi estadística espacial deriven directament de la naturalesa de les dades a tractar. Cressie (1993) proposa una classificació de les dades espacials on destaquen principalment tres tipologies: dades de punts geogràfics, on per cada observació es disposa de les coordenades on es produeix, i es caracteritzen per la distribució espacial d'aquestes observacions; dades contínues, on es disposa d'informació per determinades variables d'interès a qualsevol punt del territori estudiat; dades per àrees geogràfiques, on tot i que les observacions puguin produir-se en punts geogràfics concrets, els valors associats es generen a partir de processos estadístics, basats en determinades particions de la zona en estudi. Per dur a terme estudis amb dades econòmiques o sanitàries, per exemple, tot i que la informació es recull a nivell individual, normalment les dades només es distribueixen de forma agregada. Les unitats d'agregació de les dades solen anar associades a particions preexistents del territori com poden ser seccions censals, barris, municipis, àrees metropolitanes o altres.

La definició de les unitats en el procés de generació de dades geogràfiques agregades és molt determinant, ja que poden esbiaixar els resultats de l'anàlisi espacial (Openshaw, 1977). L'impacte que té la definició de les àrees geogràfiques és conseqüència de dos efectes interrelacionats, l'escala i la forma, que conjuntament conformen un problema àmpliament estudiat i conegut

com a "*modifiable areal unit problem*" (MAUP) (Openshaw, 1984). L'escala es refereix a la dimensió espacial a la qual els fenòmens, entitats, patrons o processos són observats i caracteritzats (O'Neill & King, 1998). En el cas de l'escala de les àrees geogràfiques, el problema del MAUP deriva del fet que la dimensió d'aquestes àrees determina la quantitat d'observacions que s'agruparan en cadascuna i el nombre de zones que hi haurà disponibles per l'estudi. A part de la mida de les àrees, un segon problema és com se'n fa la delimitació perquè determina com s'agrupen les diverses observacions entre elles. Sovint, la definició de les zones es fa en base a delimitacions administratives o polítiques que poden amagar determinades interaccions espacials entres les observacions que s'hi agrupen. En aquest sentit diversos projectes, tant a nivell nacional com global, proposen sistemes de dades estadístiques en graelles quadriculades per representar la realitat socioeconòmica sense els efectes de les delimitacions administratives (Deichmann et al., 2001; Backer & Bloch Holst, 2011; Dmowska & Stepinski, 2017). Les dades en quadrícules, a més de ser independents de divisions administratives, permeten també que les dades siguin comparables entre territoris diferents o bé en moments diferents en el temps. A més, són una forma d'estandardització en la recol·lecció que de manera senzilla pugui integrar dades de diverses fonts.

1.3 Anàlisi de dades espacials

Les tècniques d'anàlisi espacial, quan les dades amb les que treballem presenten una clara vinculació geogràfica, permeten detectar i mesurar patrons espacials i seleccionar els mètodes més apropiats per tal de modelar i interpretar el fenomen estudiat.

Normalment els processos d'anàlisi de dades solen iniciar-se calculant descriptors estadístics bàsics. Aquesta informació ofereix una visió resumida de les dades, però no aporta informació de com els valors de les dades es relacionen amb l'espai i de com varien en els diferents llocs de la zona geogràfica a explorar. Com hem comentat anteriorment, representar les dades en mapes aporta una nova capacitat exploratòria i afegeix un component geogràfic a

l'anàlisi. Així mateix també pot ajudar a escollir les tècniques d'anàlisi espacial més adequades per continuar aprofundint en el coneixement de les dades.

Un dels primers instruments per l'exploració espacial quantitativa és la detecció d'autocorrelació espacial que indica si els valors de la variable explorada per uns determinats elements o observacions estan correlacionats amb els de les observacions properes geogràficament. Una autocorrelació espacial positiva evidencia que les observacions properes tenen valors semblants en les dades, mentre que una autocorrelació espacial negativa denota que les observacions properes tenen valors diferenciats. Un dels estadístics més utilitzats per detectar l'autocorrelació espacial és l'índex *Moran's I* que dóna una mesura global d'autocorrelació de la variable observada (Cliff, 1973). Per al càlcul del *Moran's I* s'utilitzen els valors que té la variable en cadascuna de les àrees, junt amb una estructura de pesos que defineix la relació espacial d'aquestes àrees entre si. Aquesta relació de veïnatge espacial pot definir-se de diverses formes: si les observacions són punts, es pot representar la distància entre ells; en canvi, si es treballa amb àrees, es pot contemplar el nivell de contigüitat entre les fronteres de les àrees; també es poden considerar els veïns més propers a cada àrea o punt o bé calcular les distàncies entre els centroides de les àrees o la llista d'àrees dins un llindar de distància; es pot mesurar com decau la distància o, fins i tot pot contemplar les trajectòries mínimes que permeten comunicar les àrees o punts entre si. L'índex *Moran's I* calcula un valor d'autocorrelació espacial global únic per tota l'àrea geogràfica estudiada. Per quantificar l'autocorrelació espacial de forma local es pot calcular el *Moran's I* i la seva significació per cada observació obtenint el que s'anomenen indicadors locals d'autocorrelació espacial (*LISA*) (Anselin, 1995). El valors del *LISA* es poden representar en un mapa sobre la zona geogràfica d'estudi per tal d'evidenciar visualment clústers de valors molt similars o molt diferents dels valors de la variable d'estudi. Els mapes *LISA* permeten destacar les àrees on la variable té valors alts, que estan envoltades per altres àrees on també són alts (*High-High*) o, contràriament, valors on la variable té valors baixos envoltades per altres àrees amb valors també baixos (*Low-Low*).

A partir d'aquí, per aprofundir en l'anàlisi de les interrelacions entre les variables habitualment s'aplicaran mètodes i models d'estadística espacial. El plantejament més habitual per quantificar les interrelacions entre les variables és començar aplicant tècniques estàndard de regressió lineal. Per fer una primera aproximació exploratòria, sense tenir en compte el component espacial, es pot realitzar una regressió de mínims quadrats ordinaris (*OLS*) que donarà un resum de valors estadístics per tota la zona d'estudi. Aquesta visió però, tot i que pot ser útil per ajudar a interpretar les interrelacions entre les dades, no incorpora el component geogràfic de les diverses observacions i, a més, trenca la hipòtesi de partida de les tècniques clàssiques de regressió d'independència entre les observacions. La presència d'autocorrelació espacial indica que les dades en un determinat lloc no són independents de les dades de les àrees veïnes.

Per continuar amb l'anàlisi serà útil visualitzar en un mapa alguns dels resultats obtinguts en els resultats del OLS, principalment els residus, que donaran una visió geogràfica i ajudaran en la comprensió d'aquests resultats. Per tal d'incorporar el component geogràfic, a partir de l'estructura espacial de les observacions, es podran modelar les interrelacions entre elles i els seus efectes espacials aplicant tècniques de regressió espacial.

Elhorst (2010) planteja una classificació dels principals models economètrics espacials, basats en els tres tipus de interaccions espacials introduïts per Manski (1993):

- Interrelació espacial endògena, on el fenomen estudiat per a un determinat element està influenciat pel mateix fenomen en els elements propers. En aquest cas s'aplica el model economètric *Spatial Lag* o *SAR* (*Spatial Auto Regression*) que utilitza l'estructura espacial de la zona d'estudi per mesurar la influència de la variable observada en cada àrea sobre la mateixa variable a les àrees properes.
- Interrelació espacial exògena, on el fenomen estudiat per a un determinat element està influenciat per altres característiques observables dels elements propers. El model economètric *SLX* (*Spatial Lag X*) aplica l'estructura espacial de les diverses àrees i les variables explicatives en cadascuna d'elles per mesurar-ne la influència a les àrees properes.

- Correlació espacial en els efectes provocats per factors desconeguts. En aquest cas el model economètric *SEM* (*Spatial Error Model*) planteja una correlació espacial, en allò que les característiques conegudes de les àrees no poden explicar totalment del fenomen estudiat.

Aquests tres models, de fet, són una simplificació de models més complexes i podrien generalitzar-se per obtenir-ne altres variacions. El model *SDM* (*Spatial Durbin Model*), per exemple, té en compte les interrelacions espacials endògenes i exògenes simultàniament (LeSage, 2008; Elhorst, 2010). El model *SAC* (*Spatial Autoregressive Confused*) considera que no hi ha una interacció endògena i situa les interrelacions espacials en els factors exògens i característiques desconegudes (Kelejian & Prucha, 2010).

Les tècniques de regressió espacial permeten incorporar l'autocorrelació o dependència espacial per mesurar les interrelacions entre les dades considerant que aquestes interrelacions són homogènies en tota la zona d'estudi. En alguns casos, tot i que les variables considerades permeten explicar el fenomen estudiat, les interrelacions no són de la mateixa forma a tota la zona i per tant ens trobem amb heterogeneïtat espacial o no-estacionarietat espacial (Anselin & Griffith, 1988). La tècnica de regressió ponderada geogràficament (*Geographically weighted regression – GWR*) (Brunsdon et al., 1996) permet avaluar el model tenint en compte que les variables explicatives no tenen el mateix efecte a tots els punts. La *GWR* realitza diverses regressions locals, cadascuna d'elles influenciada per les dades veïnes, permetent que els paràmetres del model variïn a través de la zona d'estudi. El resultat serà un conjunt d'indicadors estadístics per cada punt de regressió amb variacions locals en les interrelacions entre les variables.

Les diverses tècniques d'anàlisi espacial defineixen un marc consistent per tal d'avaluar i modelitzar fenòmens espacials i la seva interacció amb les característiques socioeconòmiques i demogràfiques de les diverses àrees geogràfiques. Permeten incorporar el component geogràfic de les dades que traduirà l'espai en un valor matemàtic més a tenir en compte en l'avaluació de les anàlisis estadístiques. La tècnica a aplicar en cada cas no serà única, sinó que els diversos indicadors estadístics de cada mètode ens permetran

aprofundir en el coneixement de les interrelacions entre les dades i també entendre els límits del model.

1.4 Confidencialitat estadística

La producció de dades estadístiques agregades és una forma de publicar informació resumida sobre un determinat territori, però, al mateix temps, és també una forma de garantir la confidencialitat dels individus a qui fan referència les dades. Les Nacions Unides determinen la confidencialitat estadística com una qüestió bàsica en l'administració de les dades personals dins els principis fonamentals sobre Estadística Oficial, i estableixen que les dades individuals recopilades per agències d'estadística han de tenir un tractament confidencial i han de ser utilitzades exclusivament amb finalitats estadístiques (Duncan et al., 2011). La European Comission (2009) en la regulació 223/2009 defineix les dades confidencials com aquelles "dades que permeten identificar les unitats estadístiques, ja sigui de forma directa o indirecta, donant a conèixer informació individual". S'entén per identificació directa la difusió d'informació que identifica de forma única una unitat estadística, com pot ser, per exemple, l'adreça postal. En canvi parlarem d'identificació indirecta quan es produeix a partir de la combinació d'un conjunt de variables. En aquest mateixa regulació, la Comissió Europea defineix que: "en les seves respectives esferes de competència, les Agències Nacionals d'Estadística i altres autoritats nacionals, així com la pròpia Comissió, han de prendre totes les mesures de regulació, administratives, tècniques i d'organització, per assegurar la protecció física i lògica de les dades confidencials".

La protecció i processament de les dades personals també ha estat recentment objectiu de debat i és objecte del nou marc legal de protecció de dades definit per la Regulació General de Protecció de Dades (GDPR) i la Regulació 2018/1725 del Parlament Europeu. Aquest nou marc delimita l'ús de les dades de les persones físiques per part de les institucions, organismes, oficines i agències de la Unió Europea així com el lliure intercanvi d'aquestes dades (Katulic & Protrka, 2019). La gran majoria de països (veure Figura 1) han adoptat lleis per protegir la privacitat de les dades personals a càrrec d'organismes tant

privats com públics (Banisar, 2019). Aquestes mesures han enfortit els drets fonamentals dels ciutadans en l'àmbit de les dades digitals, però també beneficia les empreses simplificant les normatives aplicables en el mercat únic digital.

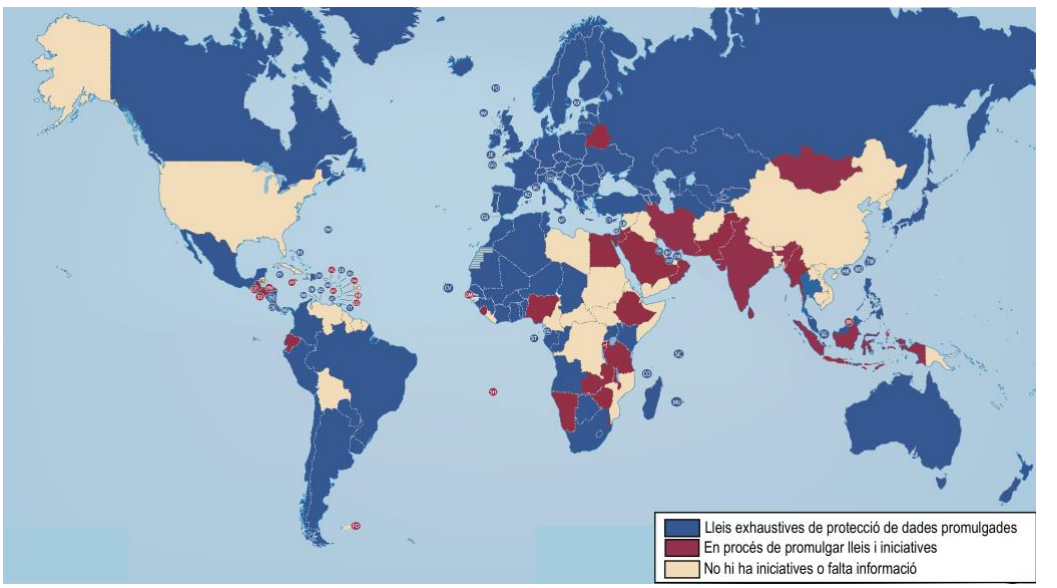


Figura 1. Mapa de països amb lleis per la protecció de dades personals(Banisar, 2019)

En l'àmbit de la confidencialitat estadística, les tècniques de control de la revelació estadística (SDC) tenen com a objectiu reduir el risc de revelació del secret estadístic en la difusió de dades i, al mateix temps, difondre tanta informació com sigui possible (Willenborg & de Waal, 2000). S'entén que es produeix revelació del secret estadístic quan a partir de les dades publicades per les agències o instituts d'estadística, es pot obtenir informació, no directament difosa i prèviament desconeguda sobre els titulars de les dades. Lambert (1993) defineix principalment dos tipus de revelació d'informació: la revelació d'identitat, que es produeix quan es pot vincular el titular d'una informació concreta amb les seves dades publicades; la revelació d'atributs, que consisteix en deduir una nova propietat o atribut a partir d'altres propietats publicades, encara que no necessàriament se'n dedueixi el titular. En aquest context, la localització geogràfica és una informació altament confidencial (Hundepool et al., 2010); és una dada que permet situar la informació en l'espai de forma precisa i, mitjançant

processos de geolocalització inversa, podria permetre revelar la identitat dels titulars de la informació (Armstrong & Ruggles, 1999).

Les tècniques de control de revelació estadística aplicades a informació geogràfica s'agrupen en el que Armstrong et al. (1999) defineixen com a emmascarament geogràfic. Els principals mètodes d'emmascarament geogràfic serien: transformacions afins, pertorbacions aleatòries, intercanvi de punts, agregació espacial i agregació de punts. Les transformacions afins desplacen els punts respecte la seva localització original, ja sigui movent-los una distància fixa, escalant la distància respecte un origen, o bé aplicant una rotació respecte un punt pivot. Les pertorbacions aleatòries apliquen a cada punt un desplaçament aleatori tant en distància com en direcció, normalment amb una distància màxima respecte la seva posició original; algunes variants apliquen també una distància mínima. En l'intercanvi de punts es desplacen punts a zones properes on hi ha característiques semblants. L'agregació espacial agrupa la informació dels punts individuals en àrees geogràfiques, ja siguin quadrícules o bé zones dissenyades ad hoc; en canvi, en l'agregació de punts s'agrupen uns quants punts propers geogràficament i es mostren de forma conjunta en una localització concreta.

En el cas dels estudis científics, les metodologies més aplicades per mantenir la confidencialitat són l'agregació espacial i la pertorbació aleatòria de les coordenades geogràfiques per ressituar la informació en el territori (Kounadi & Leitner, 2014). Les agències d'estadística apliquen metodologies d'agregació espacial en la publicació de la informació tant d'individus com de llars, empreses o entitats. Així, la protecció de la confidencialitat passa per agregar les dades individuals de manera que la informació es publica resumida per àrees per evitar que es puguin inferir les dades individuals. Evidentment, dependrà de com de grans i quants elements incloguin aquestes àrees per assegurar que la informació sigui realment confidencial (Zandbergen, 2014). El cens, per exemple, és un dels processos més acurats de recopilació de dades estadístiques utilitzats a la majoria de països, on es recopila informació detallada i individual sobre composició familiar, salut, ocupació, estudis, i altres característiques socio-econòmiques. Tant en la planificació del cens, o altres

processos de recopilació d'informació, com en la seva posterior divulgació les agències responsables utilitzen diversos processos per tal de garantir la confidencialitat de les dades (Gauthier, 2002; Longhurst et al., 2007; Zayatz, 2007; Forbes et al., 2009; Jansson, 2012).

1.4.1 Mesurar la confidencialitat

A partir dels mètodes d'emmascarament geogràfic existents una de les qüestions importants és poder-ne avaluar l'eficàcia per assegurar la confidencialitat. Alguns estudis plantegen que aplicar una distorsió prou gran a la localització dels punts inicials és suficient per garantir la privacitat de la informació (Kwan et al., 2004; Leitner & Curtis, 2006), en alguns casos es plantegen també determinades mesures o llindars mínims de privacitat (Cassa et al., 2006; Allshouse et al., 2010). Un dels marcs més àmpliament acceptats per avaluar el grau de confidencialitat de les dades espacials es basa en aplicar el concepte de "*k-anonymity*", prèviament introduït per Sweeney (2002) com a mecanisme per assegurar l'anonimat en les dades tabulars. En el cas genèric, el concepte de *k-anonymity* és una mesura probabilística del risc d'identificació dels titulars dels registres d'una taula de dades. Per considerar que un conjunt de dades compleix *k-anonymity* cal assegurar que per cada registre o informació individual dins la taula, hi ha un mínim de k-1 registres amb la mateixa combinació de valors en els atributs identificadors.

Aplicat a dades espacials, la mesura de *k-anonymity* dóna una estimació de la probabilitat d'identificar la posició geogràfica associada a unes dades concretes. Aquesta probabilitat vindrà determinada tant per la magnitud de la distorsió aplicada a les coordenades inicials de cada localització concreta, com per la densitat d'elements en la zona on es troba (Cassa et al., 2006; Allshouse et al., 2010). En el cas de les metodologies d'agregació espacial, el concepte de *k-anonymity* anirà lligat directament a la quantitat de punts agrupats en cada àrea. Un conjunt de dades espacials agregades assoleixen *k-anonymity* si en cada àrea s'agrupen les dades d'un mínim de k elements (Vu et al., 2012).

Alguns autors van demostrar algunes febleses del model de *k-anonymity* aplicat als mètodes d'agregació de dades, proposant millores al model o creant noves

mesures de privacitat. Machanavajjhala et al. (2007) demostren que quan les dades són molt semblants dins els subgrups, o bé es disposa d'informació complementària, es podria arribar a inferir informació específica individual. Utilitzant mesures d'entropia defineixen la propietat *ℓ-diversity* per mesurar com la informació confidencial dels elements és prou diferent entre els del mateix grup o àrea geogràfica. Ninghui et al. (2007) mostren algunes limitacions de la propietat *ℓ-diversity* i proposen el model *t-closeness* per mesurar la diferència entre les distribucions de la informació confidencial dels diversos grups respecte la distribució de la mateixa informació per al conjunt global de dades. Alguns autors han proposat també altres models per complementar el *k-anonymity* com *p-sensitive k-anonymity* (Truta & Vinay, 2006), *(α,k)-anonymity* (R. C.-W. Wong et al., 2006), *m-invariance* (Xiao & Tao, 2007), or *δ-presence* (Nergiz et al., 2007). En tots els casos, més enllà d'assegurar que no es puguin identificar els individus a qui pertany la informació, els models de privacitat pretenen que les dades agrupades siguin prou diferents entre si per tal que no es pugui inferir informació considerada confidencial.

1.5 Efectes de l'agregació espacial de dades en les anàlisis estadístiques

La distribució de dades de forma agregada permet garantir la confidencialitat si s'assegura que les agrupacions que es fan contenen un volum suficientment gran de dades individuals. Contràriament, quan més grans siguin les agrupacions menys utilitat tindran per l'aplicació de determinades tècniques analítiques espacials com per exemple anàlisis de patrons de punts o detecció de clústers (Armstrong et al., 1999). De fet, l'escala d'agregació defineix un punt de vista particular de la realitat geogràfica a través de la que s'investigarà un determinat fenomen (Levin, 1993). La dependència respecte l'escala d'agregació de les dades és, per tant, una propietat inherent als fenòmens geogràfics ja que la majoria de patrons geogràfics sota observació varien amb l'escala (Kolaczyk & Huang, 2010). Alguns patrons, de fet, només s'observen quan l'escala de l'anàlisi s'aproxima a l'escala operacional del fenomen estudiat (Allen et al., 1984).

El possible biaix que pot haver-hi en els resultats estadístics de les anàlisis espacials, degut a la mida de àrees utilitzades per agrupar les dades és un fenomen àmpliament estudiat i conegut. Gehlke & Biehl (1934) és un dels primers estudis que destaca com alguns valors estadístics podien variar significativament segons l'escala de les àrees geogràfiques utilitzades per agregar les dades individuals a estudiar. En concret, els autors van mostrar com els coeficients de correlació entre la delinqüència juvenil masculina i els preus mitjans de lloguer mensual, incrementaven a mesura que s'agrupaven les dades en àrees més grans. Robinson (1950) va introduir el terme "*Ecological fallacy*" per descriure l'error resultant d'inferir estadísticament relacions individuals a partir de les mateixes interrelacions en les dades agregades. S'entén per *correlació ecològica* la correlació entre variables considerades de forma agrupada, mentre que la *correlació individual* considera objectes estadístics indivisibles. L'autor va demostrar com la correlació ecològica no podia ser utilitzada per validar conclusions sobre correlacions individuals. McCarty et al. (1956) van generalitzar aquesta idea demostrant que les conclusions obtingudes a una determinada escala geogràfica no podien ser considerades vàlides a cap altra escala. O'Neill (1979) planteja l'efecte de l'escala geogràfica a partir de diverses organitzacions jeràrquiques de les àrees a tractar a diferents escales. L'article demostra els canvis en els models estadístics d'un determinat procés observat a diferents nivells d'escala d'agregació i denomina aquest efecte "*spatial transmutation*". Openshaw & Taylor (1979) continuen estudiant la variació de les correlacions entre diverses variables d'estudi a diferents escales definint el concepte "*modifiable areal unit problem*" (MAUP).

En el plantejament d'una anàlisi espacial, l'àrea geogràfica d'estudi pot dividir-se de moltes formes per obtenir un conjunt d'unitats no sobreposades. La definició d'aquestes unitats pot venir donada per requeriments propis de l'estudi o bé per la disponibilitat de les dades. Per tant, com que aquesta definició pot ser arbitraria i modificable, el resultat de l'estudi tindrà validesa només per aquella zonificació concreta i pot no ser extrapolable a altres possibles subdivisions. Com s'ha comentat abans, el MAUP inclou dos aspectes amb influència sobre el resultat de les anàlisis espacials: la mida de les àrees

geogràfiques en que s'agrupen les dades individuals ("*scale problem*") i la forma d'aquestes àrees ("*aggregation problem*"). De fet, aquests dos factors solen anar associats quan es defineix la partició o zonificació de la zona geogràfica a estudiar, i molts autors han estudiat la seva influència en els resultats dels estudis estadístics espacials de forma conjunta (Arbia, 1989; Fotheringham & Wong, 1991; Amrhein & Reynolds, 1996; D. W. S. Wong et al., 1999; Paelinck, 2000; D. W. S. Wong, 2004; Wu, 2004; Briant et al., 2010).

Alguns autors han plantejat tècniques per evitar els efectes del MAUP, però tal com planteja Openshaw (1977, 1978, 1984), la majoria de tècniques queden invalidades per la manca d'objectivitat en la definició de les zones, i, per tant, és inevitable la dependència dels resultats de les anàlisis respecte la unitat d'àrea utilitzada. En lloc d'evitar els efectes del MAUP, l'autor suggereix que cal incloure la informació de la zonificació en la valoració del resultats obtinguts en l'estudi i les possibles conseqüències que té aquesta partició concreta sobre les entitats espacials estudiades. De fet, alguns autors plantegen que els efectes de l'escala en els estudis espacials poden aportar informació que ajudi a entendre les característiques de determinats processos a diferents escales en els territoris (Jelinski & Wu, 1996; Wu, 2004). Tot i això, en alguns casos, l'escala pot amagar determinats patrons espacials o esbiaixar els resultats dels anàlisis (Rushton et al., 2006). En el cas d'anàlisi de dades de salut, per exemple, l'efecte de l'escala, redueix la capacitat de detectar clústers (Zimmerman et al., 2008; Jacquez & Rommel, 2009) o de detectar interrelacions entre algunes malalties i factors de risc que varien geogràficament (Diez Roux, 2001; Mazumdar et al., 2008).

A banda de l'escala, la definició de les fronteres o la forma de les àrees d'agregació és un segon component del MAUP i també podrà afectar els resultats de les anàlisis. Normalment la partició geogràfica no serà específica per al fenomen a estudiar sinó que les àrees estaran definides prèviament i dependran de qüestions administratives (seccions censals, barris, districtes, municipis, i altres) que poden presentar particularitats diferents a les que requereix l'estudi. Aquest biaix també pot influir els resultats de les anàlisis estadístiques i provocar errors d'interpretació en els models espacials (Openshaw & Alvanides, 1999). Aquest problema pot ser particularment evident

en els estudis que persegueixen capturar les característiques de l'entorn urbà més proper als llocs de residència dels individus (Mitra & Buliung, 2012). El comportament dels individus depèn de la seva pròpia percepció de l'entorn urbà, i no de la definició administrativa utilitzada per agrupar les dades individuals. En alguns casos, el biaix pot afectar, fins i tot, el signe de l'associació entre les característiques de la població i els factors estudiats (Riva et al., 2008; Coffee et al., 2013; Amini et al., 2016). Aquests efectes sobre els resultats dels estudis s'han demostrat especialment importants, per exemple, en les interrelacions entre les característiques demogràfiques o socioeconòmiques i la mobilitat (Viegas et al., 2009; Abdel-Aty et al., 2013) o l'activitat física (Boone-Heinonen et al., 2010; Houston, 2014). Per reduir els efectes de la zonificació Zhang & Kukadia (2005) van comparar els efectes de diversos sistemes d'agrupació administratius i també en quadrícules a diverses escales i van comprovar que l'agregació en quadrícules reduïa l'impacte del MAUP. En aquest sentit diversos autors han mostrat també com els sistemes d'agregació de les dades en quadrícules són una bona alternativa ja que la independència de fronteres administratives ofereix més robustesa per a l'anàlisi de dades espacials (Kim et al., 2006; Giuliani et al., 2011; Tammilehto-Luode, 2011).

1.6 Objectius de la tesi

Actualment, les agències d'estadística i altres organismes encarregats de recollir, tractar i publicar dades estadístiques disposen d'eines de geolocalització que els permeten etiquetar geogràficament les dades. A més, l'evolució dels sistemes d'informació geogràfica fan que es disposi d'informació individual amb un precisió geogràfica cada vegada més gran. Les tècniques i mètodes d'anàlisi espacial donen als investigadors la oportunitat de realitzar estudis geogràfics molt específics; però, per tal que els resultats de les anàlisis espacials siguin acurats cal que les dades que es difonen siguin el màxim de precises. D'altra banda, els identificadors geogràfics trenquen els principis de confidencialitat de la informació perquè permeten identificar els individus a qui fan referència. Així doncs, hi ha un conflicte evident entre la difusió de dades geogràfiques precises i la privacitat.

L'objectiu principal d'aquesta tesi és definir una metodologia per distribuir dades que tenen un component geogràfic, amb el màxim detall possible i, al mateix temps, mantenir la confidencialitat dels individus, empreses o entitats als quals fan referència aquestes dades. La tesi proposa un sistema d'agregació de dades geogràfiques per tal de mantenir la confidencialitat de la informació, de forma que qualsevol conjunt de dades produït tindrà com a restricció bàsica la no re-identificació de les dades individuals. Aquesta metodologia, proposada a nivell teòric, ha de poder ser implementada per tal de disposar d'un sistema informàtic que pugui generar un conjunt de dades geogràfiques de forma automàtica. Un objectiu complementari de la tesi serà desenvolupar una llibreria que implementi la metodologia proposada en un entorn de programari estadístic utilitzant formats de dades estàndard. Els sistemes basats en àrees administratives utilitzen descriptors numèrics o alfanumèrics per identificar l'àrea a la qual pertanyen les dades, de manera que l'agregació és automàtica. La metodologia proposada, a banda d'agregar els punts utilitza també un sistema d'identificació basat un estàndards europeus per tal de construir sistemes de distribució de dades transversals. El temps de computació és un punt crític perquè els volums de dades a gestionar normalment serà gran i per tant la llibreria haurà d'optimitzar el procés d'agregació geogràfica per tal de generar les dades agregades en un temps de computació raonable.

Els sistemes de difusió actuals utilitzen, en la majoria de casos, les seccions censals com a unitat d'àrea de màxima precisió. Amb l'objectiu de demostrar el potencial de la metodologia implementada, la tesi proposa dos casos d'estudi reals realitzats amb dades actualment disponibles només a nivell de secció censal. Així, a partir de les dades socioeconòmiques i demogràfiques disponibles per la ciutat de Barcelona, s'estudien des d'una perspectiva espacial les seves interrelacions amb dos fenòmens amb un clar component geogràfic. Ambdós fenòmens poden caracteritzar-se espacialment de forma molt precisa, i s'han d'aplicar transformacions espacials per equiparar-los a les dades censals. Aquestes transformacions poden introduir biaixos en els resultats de les anàlisis espacials.

A continuació es presenta una visió general dels 4 articles que conformen la tesi:

- El primer article és fruit d'un conveni de col·laboració entre l'Institut d'Estadística de Catalunya i el grup de recerca *Data Modelling and Analys*is de la Universitat de Vic – Universitat Central de Catalunya. L'article defineix una nova metodologia per construir una graella espacial amb cel·les de mida variable per difondre dades estadístiques geolocalitzades. A partir d'un conjunt de punts geogràfics el sistema proposat construeix una estructura espacial que cobreix tots els punts amb un quadricula de cel·les. La quadrícula es construeix a partir d'una organització jeràrquica, on cada cel·la es va dividint recursivament en quatre noves cel·les (*Quadtree*) per tal d'aconseguir cel·les de la menor mida possible. L'objectiu principal de la metodologia proposada és assegurar que l'estructura creada manté la confidencialitat de les dades, aplicant un criteri de *k-anonymity* per tal que tota cel·la contingui un mínim de punts. A més, per evitar identificar propietats concretes dels individus, el mecanisme també aplica criteris de *l-diversity* per assegurar un mínim d'individus amb igual combinació de valors per a propietats considerades crítiques per a la possible re-identificació. D'altra banda, tenint en compte els criteris de privacitat, es produeix una quadrícula amb la màxima precisió, creant cel·les tan petites com sigui possible. L'estructura creada, a més, es basa en estàndards europeus per la producció de dades estadístiques en quadrícules, i per tant, les dades poden acoblar-se a altres sistemes europeus, garantint també estabilitat tant transnacional com en el temps. L'Institut d'Estadística de Catalunya va aplicar la metodologia proposada en aquest article per publicar un subconjunt de les dades del Registre de població de Catalunya en una quadrícula multiresolució amb cel·les de fins a 62,5 metres (Institut d'Estadística de Catalunya, 2016). Posteriorment, aquestes mateixes dades es van difondre mitjançant una eina cartogràfica digital i interactiva de visualització i anàlisi, a la web de l'Institut Cartogràfic i Geològic de Catalunya (Institut Cartogràfic i Geològic de Catalunya, 2017).

En els següents dos articles es plantegen estudis espacials amb dos tipus de dades disponibles a un nivell de precisió molt alt, comparades a dades socioeconòmiques i demogràfiques obtingudes a través dels sistemes de difusió actuals, a nivell de secció censal. Aquesta diferència en l'escala de les dades pot donar lloc a errors o biaixos en els resultats de les anàlisis espacials. Aquesta qüestió es plantejarà en l'apartat de discussió de la tesi.

- El segon article avalua l'exposició al soroll d'alguns grups de població considerats vulnerables, a la ciutat de Barcelona. En concret es consideren dos grups de població segons l'edat, ja que, tal com planteja la bibliografia existent, són grups més sensibles a les conseqüències que pot provocar el soroll en la seva salut. Els altres grups de població considerats són els immigrats de fora de la Unió Europea, la població en atur, sense estudis secundaris o amb baixos ingressos econòmics i, per tant, amb poca capacitat per canviar de zona de residència i evitar les zones d'alts nivells de soroll. Així, la idea de l'estudi és comprovar la desigualtat respecte la molèstia provocada per alts nivells de soroll. A partir dels resultats dels diversos tests estadístics i les anàlisis de regressió espacial realitzades, els nens i els individus amb baixos ingressos no mostren afectacions respecte alts nivells de soroll, mentre que si que es detecta una interrelació positiva entre els nivells de soroll i la població a l'atur i les persones grans.
  Les dades dels nivells acústics a la ciutat de Barcelona estan disponibles a escala de tram de carrer mentre que les dades utilitzades per determinar els grups de població estan disponibles només a nivell de secció censal. Per tal de fer les dades comparables, les dades acústiques van ser agregades a escala de illes a partir de les dades urbanístiques de l'Ajuntament de Barcelona. Les dades de les seccions censals van ser baixades a les illes suposant una distribució dels grups de població totalment uniforme a totes les illes de cada secció censal. Aquesta suposició pot esbiaixar els resultats obtinguts en l'article, com es discutirà en l'apartat corresponent de la Tesi.

- El següent estudi es proposa interpretar la distribució espacial dels allotjaments *Airbnb* a la ciutat de Barcelona, i la interrelació entre aquest

tipus d'allotjaments turístics i els factors socioeconòmics i demogràfics de la població, així com l'atractiu turístic de les zones on estan ubicats. L'estudi proposa analitzar la no estacionarietat espacial de les interrelacions entre la densitat d'allotjaments i els factors explicatius a partir d'una regressió ponderada geogràficament. La ciutat de Barcelona ha experimentat diverses transformacions urbanístiques en els darrers anys i alguns barris han patit processos de gentrificació. Des del punt de vista turístic, Barcelona és la cinquena ciutat europea més visitada tenint en compte les pernoctacions dels turistes. La transformació d'habitatges en establiments turístics pot provocar alguns problemes de convivència i pot impactar en la qualitat de vida dels veïns, sobretot, quan la densitat d'allotjaments turístics és alta. L'estudi realitza una anàlisi espacial, tenint en compte la variabilitat de les correlacions espacials en les diverses àrees de la ciutat, per tal d'entendre com els allotjaments de curta durada s'han estès a la ciutat. Els principals factors que expliquen la variació de les proporcions de lloguers *Airbnb* respecte el total de llars són els ingressos i el nivell d'estudis de la població, així com la mida dels habitatges. A nivell local però, les interrelacions mostren patrons espacials de signe invers a les diverses zones de la ciutat.

La informació sobre els allotjaments turístics és molt precisa, fins al punt que en la majoria de casos cada allotjament proporciona les coordenades geogràfiques d'on està ubicat. Les dades sobre els punts d'atracció turística també estan disponibles a nivell de coordenades específiques. En canvi, tal com passava en l'estudi anterior, les dades referents a les característiques sòcio-econòmiques de població només estan disponibles a nivell de seccions censals. Per uniformitzar l'escala geogràfica, es van agregar les dades dels allotjaments i les dades turístiques i així poder comparar-les geogràficament i dur a terme anàlisis espacials.

- Finalment, el darrer article presenta la llibreria en R amb la metodologia descrita anteriorment, i aporta detalls sobre la implementació i l'ús dels diversos mètodes associats. La llibreria s'ha basat en objectes geogràfics

estàndard de l'entorn de programació estadístic R, però utilitzant un sistema d'agregació propi, a partir de les definicions del projecte GEOSTAT. El programari implementa diverses funcions per l'anàlisi i anonimització de punts espacials amb dades associades, utilitzant tècniques d'agregació i supressió. Així mateix, també proveeix una classe S4 de R, per crear, manipular i exportar quadrícules espacials. L'objectiu de la implementació és oferir mètodes automàtics per l'agregació de les dades, minimitzant el temps de computació mitjançant l'ús de tècniques de codificació geogràfica. A més, a llibreria desenvolupada aplica mesures d'entropia per tal de balancejar la resolució de la quadrícula i la possible pèrdua d'informació, de forma automàtica i parametritzada per l'usuari.

Publicacions

2 Publicacions

2.1 Publicació 1: A quadtree approach based on European geographic grids: reconciling data privacy and accuracy

2.2 Publicació 2: Environmental noise inequity in the city of Barcelona

2.3 Publicació 3: Understanding Airbnb spatial distribution in a southern European city: The case of Barcelona

2.4 Publicació 4: AQuadtree: an R Package for Quadtree Anonymization of Point Data

Discussió

3 Discussió

Una de les qüestions importants a plantejar-se per analitzar fenòmens espacials és l'escala a la que s'estudiaran (Johnston et al., 2019). Per tal de captar de la millor forma possible l'impacte dels indicadors analitzats, s'hauria d'utilitzar la resolució més acurada possible, utilitzant les unitats d'agregació espacial més petites disponibles o més properes a la resolució del fenomen estudiat. Però, en la majoria de casos, hi haurà limitacions en la disponibilitat de la informació, que vindran imposades pels sistemes preexistents de recol·lecció i distribució de dades estadístiques. La metodologia descrita en el primer article de la tesi, i implementada en el quart article, proposa un sistema per produir dades de forma més precisa que ofereixi millors possibilitats tant per la visualització com per l'anàlisi.

El segon i tercer articles inclosos a la tesi, són dos exemples d'anàlisi espacial que requeririen dades més precises. Aquests dos estudis realitzen anàlisis de dos fenòmens a la ciutat de Barcelona aplicant models espacials. En ambdós casos les dades que els descriuen tenen una precisió geogràfica molt acurada. Per altra banda, per tal de fer-los comparables a la informació disponible sobre la població o llars del territori, les dades s'han d'agregar en resolucions menys precises. Per exemple, els nivells acústics amb els que s'ha treballat en el segon article estan disponibles per trams de carrer, amb una mitjana de la longitud dels trams de 90 metres. Els allotjaments turístics *Airbnb*, estudiats en el tercer article, estan catalogats, en la majoria dels casos, amb les coordenades geogràfiques de la ubicació concreta de cadascun d'ells. Aquestes dades permeten realitzar anàlisis de distribució espacial molt acurats, però, en canvi, per aplicar models estadístics espacials, la informació socioeconòmica i demogràfica de la població no està disponible de forma tant precisa. En els models aplicats en els estudis, les unitats espacials utilitzades van venir determinades per la disponibilitat de les variables explicatives seleccionades i, per tant, les dades que descriuen els fenòmens es van haver d'agregar perdent

precisió. Aquestes transformacions poden haver provocat biaixos en els resultats.

Evidentment, quan cal tractar dades sobre individus, empreses, o llars, les mesures de privacitat limiten la forma com aquestes poden ser distribuïdes, però els sistemes actuals no estan dissenyats tenint en compte només finalitats de confidencialitat sinó que solen basar-se en decisions administratives. Per això, el *European Statistical System (ESSnet)* junt amb el *European Forum for Geography and Statistics (EFGS)* van engegar conjuntament el projecte *GEOSTAT* per tal de disposar d'un marc uniforme per representar les característiques de la població i les llars a partir de l'any 2011 per tota la unió europea en una quadrícula de cel·les de 1km^2. Aquest projecte va néixer per tal de donar solució a les mancances dels mètodes tradicionals d'estadístiques oficials que recopilaven les dades en sistemes jeràrquics d'unitats administratives. Aquests sistemes tradicionals, tot i que poden ser útils amb finalitats de comptabilització, són poc adequats per estudiar causes i efectes de molts fenòmens socioeconòmics i ambientals. L'objectiu del projecte no és substituir els sistemes basats en unitats administratives sinó complementar-los on aquets tenen limitacions. En aquest sentit, la legislació de la Unió Europea per al cens de població i habitatge de l'any 2021 inclou una llei específica per tal que els estats membres publiquin algunes dades del cens en una quadrícula amb cel·les de 1km^2, comuna per tota la UE (Eurostat, 2019). Aquesta nova forma de distribuir dades censals està dissenyada específicament amb finalitats de recerca però també, per exemple, per la definició de normatives amb una dimensió geogràfica. A més, l'ús d'una graella única ha de permetre realitzar anàlisis més flexibles a nivell transnacional. La metodologia d'agregació de dades proposada en aquesta tesi es basa en les directives del projecte *GEOSTAT* i implementa un sistema per produir dades agregades en quadrícules a partir de dades en punts geogràfics. Aquesta metodologia assoleix un doble objectiu: produeix dades amb la resolució espacial més acurada possible i, al mateix temps, manté la confidencialitat dels elements als que fan referència les dades. Mitjançant l'ús de la metodologia descrita es poden generar dades als nivells requerits per les noves lleis de cens de la UE, però també es poden

generar dades més precises a les zones amb altes densitats de població, més adaptades per tant, a les dades específiques a distribuir.

La llibreria s'ha implementat i publicat en el repositori públic de R perquè és el programari estadístic més complert per l'estimació de models economètrics espacials. Tanmateix, també proveeix mètodes per exportar les dades produïdes en formats de vectors geo-espacials estàndard i, per tant, poden ser tractades amb altres programaris estadístics i *GIS*. La llibreria implementa un tipus de dades específic que representa una graella de cel·les quadrades però, a diferència del projecte *GEOSTAT,* que defineix àrees de mida fixa, les cel·les creades amb la llibreria són de mida variable per tal d'adaptar-se a les diferents densitat de punts i aconseguir tenir millor resolució espacial. Tot i això, la definició de les cel·les permet també agrupar-les en una graella de mida constant per fer-les compatibles amb el projecte *GEOSTAT* perquè la codificació utilitzada coincideix amb les directrius del projecte europeu. La llibreria inclou mètodes per crear i gestionar quadrícules de mida constant, a qualsevol resolució, que cobreixin un territori donat, de manera equivalent a com ho fa el programari *Gridmaker* distribuït per l'Eurostat (Eurostat, 2020), però també permet crear quadrícules adaptades a un conjunt concret de punts geogràfics.

La llibreria també optimitza el rendiment en temps d'execució per tal de poder gestionar grans volums de dades. Per això s'utilitza el propi sistema de codificació de les cel·les per optimitzar les agregacions. Els testos realitzats en un ordinador amb processador Intel© 4 Core i7 de 2,5 GHz han permès agregar els 7.566.464 punts del registre de població de Catalunya de l'any 2014 en una graella amb 85.408 cel·les i una resolució de fins a 31,25 metres en 46 segons de temps d'execució.

Com a mecanisme de privacitat la metodologia es basa en el principis de *k-anonymity* i *ℓ-diversity*, descrits en la introducció, com a mesures del risc d'identificació dels individus a qui pertanyen les dades. En el cas de les metodologies d'agregació espacial, el concepte de *k-anonymity* implica que en cada àrea s'agrupen les dades d'un nombre mínim d'elements. La mesura *ℓ-diversity* persegueix la variabilitat en els valors de determinades propietats dins de les àrees, per evitar que es puguin re-identificar els individus a partir

d'encreuaments d'informació, o bé utilitzant coneixements previs de la zona en qüestió. En la construcció de la graella un dels paràmetres del procediment permet definir el llindar mínim d'elements per crear cada cel·la, de manera que es generaran les cel·les tant petites com sigui possible en funció d'aquest llindar, i tenint en compte també un dimensió de casella mínima que podrà ser parametritzada. Aquest mateix llindar es pot estendre a altres propietats considerades crítiques des del punt de vista de la re-identificació. Per tant, la confidencialitat dels conjunts de dades produïts es podrà controlar a partir dels paràmetres que s'utilitzin en la creació de la graella.

La metodologia proposada agrupa les dades de tots els punts geogràfics creant cel·les de la menor dimensió possible, per tant, permet produir dades amb millor resolució espacial que els mètodes habituals, basats en seccions censals. En les dues seccions següents utilitzem la llibreria AQuadtree, basada en aquesta metodologia, per tal de comparar algunes característiques espacials de les dades associades als dos fenòmens estudiats en el segon i tercer articles presentats en la tesi.

3.1 Allotjaments turístics de la ciutat de Barcelona

Per tal de comparar les dades agrupades en quadrícula i en seccions censals hem realitzat una anàlisi de la distribució espacial dels punts a partir de les dades dels allotjaments *Airbnb* de la ciutat de Barcelona utilitzades per l'estudi espacial del tercer article de la tesi. Els punts geogràfics corresponents a les ubicacions dels allotjaments s'han agregat per seccions censals i, utilitzant la llibreria *AQuadtree*, s'han agregat també en una quadrícula amb caselles d'entre 1 quilometre i 31,25 metres. Podem fer una anàlisi comparativa, amb algunes mesures habituals de distribució espacial de punts.

Taula 1. Diferències del centre mig i la desviació estàndard de la distribució dels nivells acústics

	Punts	Quadrícula	Seccions Censals
diferència al centre mig	(ref.)	18 metres	275 metres
diferència de desviació estàndard de la distància	(ref.)	21 metres (0.96%)	197 metres (8,79%)

La Taula 1 mostra la diferència de les coordenades del centre mig dels dos models d'agregació, respecte les coordenades del centre mig de les localitzacions dels allotjaments turístics de *Airbnb*; la taula també mostra la diferència de les desviacions estàndard de les distàncies. El centre mig és la mitjana de coordenades de tots els punts geogràfics considerats. Per als dos models amb dades agregades s'han utilitzat els centroides de cada àrea per calcular les mitjanes de coordenades ponderades segons el nombre d'allotjaments en cada àrea.

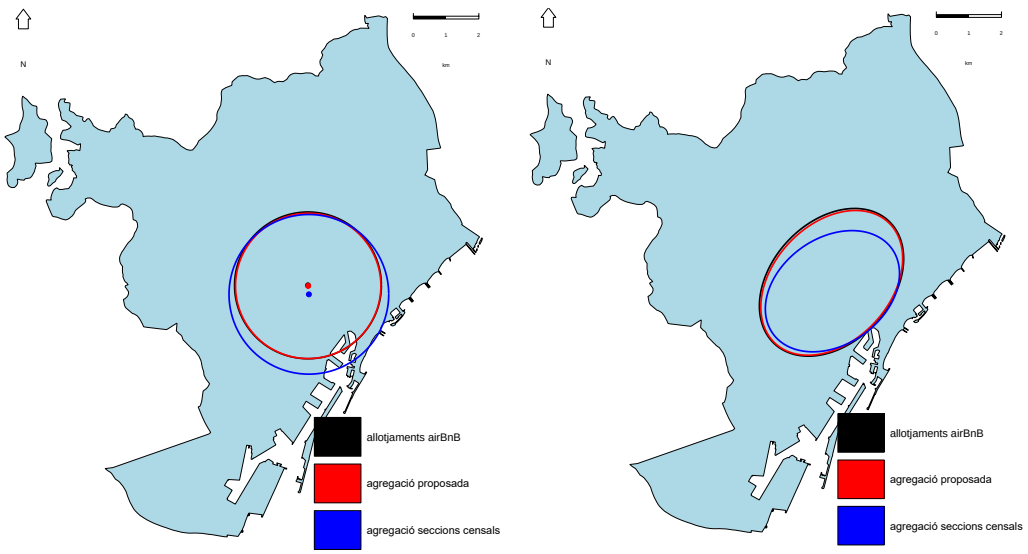


Figura 2. Centre mig i distància estàndard d'allotjaments Airbnb (esquerra); El·lipse de desviació estàndard (dreta)

La primera imatge de la Figura 2 mostra les dades de la Taula 1 de forma gràfica sobre el mapa de la ciutat; el punt central representa el centre mig i la circumferència mostra la desviació estàndard de les distàncies de cada allotjament respecte aquest centre. Com es pot veure en la imatge, les dues mesures, considerant els punts agregats en quadrícula o prenent les localitzacions individuals, són pràcticament idèntiques. En canvi, si considerem l'agregació en seccions censals, les coordenades del centre mig estan a una distància 15 vegades més gran i la desviació estàndard de la distància és gairebé un 8% més gran. La segona imatge de la Figura 2 mostra les el·lipsis de desviació estàndard dels tres casos que caracteritzen la dispersió dels punts al llarg de dos eixos ortogonals i mostren la direcció de la màxima dispersió dels

punts. Igual que abans, aquesta mesura de distribució espacial, considerant els punts agregats en quadrícula és pràcticament igual que si s'analitzen les localitzacions de forma individual.

Podem comparar gràficament les densitats locals dels allotjaments *Airbnb*, a partir de les superfícies de densitat calculades mitjançant la tècnica d'estimació de densitat *kernel* (Gatrell et al., 1996; Kwan et al., 2004). La Figura 3 mostra les tres estimacions de densitat a la zona central de la ciutat, on la densitat d'allotjaments turístics és més alta, pels tres casos: localitzacions individuals, agregació utilitzant la llibreria *AQuadtree* i agregació en seccions censals. El *bandwith* òptim en els tres casos està al voltant dels 300 metres. De forma semblant al que hem pogut comprovar amb les anàlisis de distribucions de punts, el patró que mostra l'estimació de la densitat en l'agregació en *AQuadtree* reprodueix millor a la que obtenim amb les localitzacions individuals que en l'agregació per seccions censals.

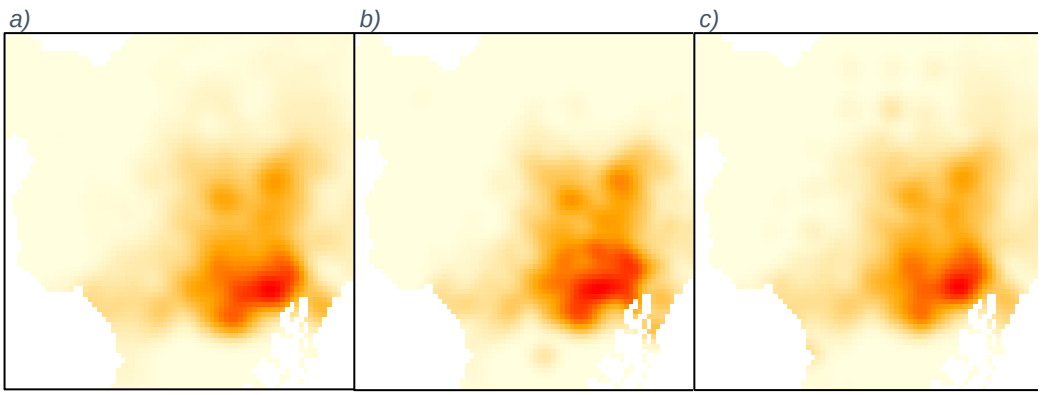


Figura 3. Estimació de la densitat kernel a la zona central de la ciutat: a) Localitzacions individuals. b) Agregació per seccions censals. c) Agregació en quadrícula AQuadtree.

Els exemples anteriors ens permeten observar gràficament les diferències entre algunes mesures d'anàlisi de distribució espacial segons el nivell d'agregació de les dades. Les agregacions s'han realitzat amb les dades dels allotjaments *Airbnb*, utilitzades en el tercer article presentat a la tesi, utilitzant, per una banda, les divisions administratives habituals en seccions censals i, per l'altra, una quadrícula en *AQuadtree* obtinguda amb la llibreria presentada. En tots els casos, les distribucions obtingudes utilitzant la quadrícula mostren patrons més propers als que es poden observar si es tracten les dades desagregades. Per

tant, en el cas concret de les dades utilitzades en el tercer article de la tesi, es pot observar com el nivell d'agregació té un efecte sobre la distribució espacial de la variable d'interès.

Els models estadístics espacials aplicats en els dos articles no poden reproduir-se perquè la informació necessària només estava disponible per seccions censals, raó per la qual són les unitats d'agregació que es van utilitzar. No obstant, podem veure algunes diferències més entre l'ús de la quadrícula i les seccions censals per als dos fenòmens estudiats.

Continuant amb les dades dels allotjaments turístics *Airbnb*, i tenint en compte que l'objectiu de la quadrícula ha de ser mantenir la privacitat de la informació, hem generat una nova quadrícula tenint en compte un criteri de confidencialitat més estricte, que asseguri un mínim de 100 observacions en cada casella.

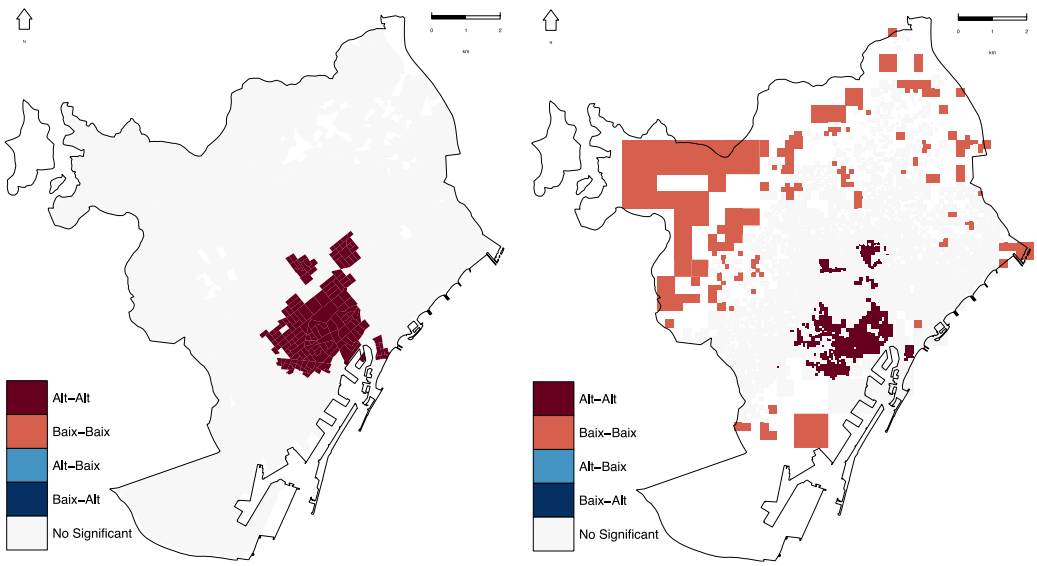


Figura 4. Clústers significants de densitats d'allotjaments Airbnb (Indicadors locals d'autocorrelació espacial). Seccions censals (esquerra) i agregació en quadrícula AQuadtree (dreta)

A partir d'aquesta quadrícula hem comparat l'anàlisi de clústers significants (*LISA*) amb el que es va obtenir a partir de les seccions censals. En l'estudi que es va realitzar a l'article no es van detectar clústers de baixa densitat d'allotjaments. És a dir, el gràfic LISA no va mostrar zones de baixa densitat d'allotjaments envoltades d'altres zones semblants. En canvi, com es pot veure a la Figura 4, la mateixa anàlisi tractant les dades en quadrícula si que mostra

diverses àrees de baixes densitats a les zones perifèriques de la ciutat. En el cas dels clústers amb alta densitat d'allotjaments, també podem observar certes diferències. Si bé en ambdós casos els clústers es troben situats al centre i zona antiga de la ciutat, el gràfic a partir de la quadrícula ens mostra aquests clústers de forma molt més precisa. Per tant, tot i no ser un factor que afecti les conclusions que es van obtenir en l'article, si que podem observar com l'anàlisi de clústers es veu afectat per la transformació que es va aplicar a les dades al pujar-les a nivell de seccions censals.

3.2 Nivells acústics de la ciutat de Barcelona

Si passem a les dades utilitzades en l'estudi presentat al segon article, referent a la interrelació espacial entre els nivells de soroll i els diversos grups de població, podem veure també algunes diferències derivades del canvi en la resolució espacial. Els nivells acústics estan caracteritzats per trams de carrer, amb una mitjana de la longitud dels trams de 90 metres. En canvi, l'escala mínima a la que estan disponibles les dades socioeconòmiques i demogràfiques, les seccions censals, tenen una àrea mitjana de gairebé 100.000 m^2 i un perímetre mig de 1.182 metres. Per realitzar els càlculs desenvolupats en l'estudi les dades dels nivells acústics de cada tram es van agregar per illes urbanístiques, i les dades socioeconòmiques i demogràfiques de la població es van baixar també a les illes urbanístiques suposant una distribució dels grups de població uniforme a totes les illes d'una mateixa secció censal. Per tant, per analitzar les dades acústiques segons les seccions censals es va introduir una pèrdua en la resolució disponible.

Utilitzant la mateixa quadrícula anterior, amb un mínim de 100 persones per casella per assegurar confidencialitat, compararem els patrons espacials de les zones amb alts nivells de soroll. Per calcular les dades acústiques de cada quadrícula s'ha realitzat una intersecció geogràfica entre els trams i la quadrícula per determinar els trossos dels trams dins de cada casella. El nivell acústic de cada àrea es calcula com una mitjana ponderada, segons la longitud del tram, dels nivells acústics de cada part de tram. El mateix procés permet obtenir els nivells acústics de les seccions censals.

La Figura 5 compara els patrons espacials que descriuen els mapes de soroll a la ciutat de Barcelona a les zones amb una mitjana de soroll[1] per sobre els 70dB(A). Les dues imatges ens mostren els patrons espacials segons si considerem els nivells de soroll a escala de secció censal o de quadrícula *AQuadtree*. En ambdós mapes destaquen les zones amb alts nivells de soroll del centre de la ciutat, tal com ja s'havia destacat en l'estudi; però si mirem l'agregació en quadrícula podem veure alguns patrons que no es detecten utilitzant seccions censals. S'evidencia de forma molt més clara, la relació entre les zones amb alts nivells de soroll i algunes vies principals de comunicació, de fet, el mapa utilitzant la quadrícula ressegueix de forma molt precisa algunes d'aquestes vies.

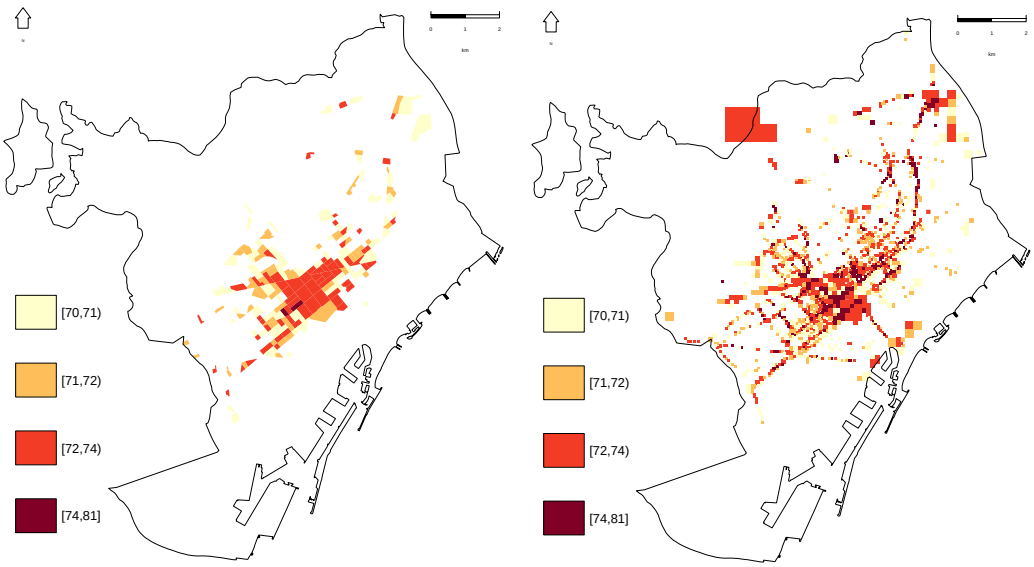


Figura 5. Nivells de soroll en zones per sobre els 70dB(A). Seccions censals (esquerra) i AQuadtree (dreta).

És interessant veure com la representació gràfica dels nivells de soroll de la Figura 5, utilitzant la quadrícula mostra de forma molt clara el que va concloure l'article respecte la relació entre els alts nivells de soroll i les vies urbanes. Tot i que, en l'article, aquesta conclusió va ser més focalitzada a la part central de la ciutat, mentre que aquí podem observar com la relació s'estén a altres zones. En aquest sentit coincidim amb el que demostren alguns autors respecte a com

[1] Mitjana ponderada dia-vespre-nit

els sistemes d'agregació de les dades en quadrícules són una millor alternativa ja que la independència de fronteres administratives permet que les anàlisis espacials es focalitzin en el fenomen estudiat (Kim et al., 2006; Giuliani et al., 2011; Tammilehto-Luode, 2011).

Taula 2. Proporcions de població resident respecte el total, segons el nivell de soroll mig.

	< 55 dB(A)	55 - 59 dB(A)	60 - 64 dB(A)	65 - 69 dB(A)	≥ 70 dB(A)
Proporció de població per seccions censals	0,55%	4,32%	32,72%	43,91%	18,50%
Proporció de població per quadrícula	1,63%	7,68%	30,26%	36,89%	23,54%

Podem comparar també el biaix que hi ha en la proporció de població segons el nivell de soroll mig de les diverses zones. La Taula 2 mostra la proporció de població a les diverses seccions censals i a les quadrícules, classificades segons el nivell de soroll de cada unitat espacial. Podem veure com, considerant les quadrícules, la proporció de població vivint en zones amb alts nivells de soroll és 5 punts més alta; paral·lelament, les zones amb els nivells de soroll més baixos també tenen proporcions més altes de població, en aquest cas, la proporció és el triple per les dades en quadrícula. Una vegada més, la precisió de la quadrícula ens permet detectar algunes particularitats del fenomen de forma més acurada.

En els exemples anteriors s'ha mantingut un llindar mínim de 100 persones per unitat espacial, per assegurar la confidencialitat. Tot i aquesta restricció, la mitjana de les àrees de la quadrícula és aproximadament 8 vegades menor al de les seccions censals, amb un coeficient de variació de gairebé la meitat. Per a aquells fenòmens que depenguin directament de l'àrea de les zones a tractar, els càlculs fets a partir de la quadrícula mostraran millor precisió que utilitzant les àrees més grans de les seccions censals. Com hem vist, la millora de la precisió en els resultats es fa bastant evident en la detecció de clústers. Tal com plantegen alguns autors, els patrons geogràfics varien amb l'escala o poden

passar desapercebuts si no s'utilitza l'escala d'anàlisi adequada (Allen et al., 1984; Kolaczyk & Huang, 2010). Per exemple, en el cas dels nivells acústics, si l'escala s'allunya de la xarxa vial, els patrons espacials lligats a aquesta xarxa poden passar inadvertits, com mostrava la Figura 5.

Conclusions

4 Conclusions

Les seccions censals han sigut una forma de distribució d'informació de la població, llars i altres elements, que s'ha mostrat útil, sobretot, en els processos de recollida i classificació. Però els sistemes actuals de posicionament geogràfic permeten referenciar la informació de manera molt més precisa. L'agregació de dades a nivell de secció censal limita la resolució a la que estarà disponible la informació a partir de criteris administratius, En alguns casos, aquestes agrupacions poden no ser del tot adequades als possibles tractaments que se'n voldrà fer.

A partir de la metodologia presentada i implementada a la tesi es poden agregar les dades amb millor resolució espacial, creant unitats d'informació més petites i mantenint els criteris de confidencialitat. Al igual que les seccions censals, les unitats creades tenen àrees que depenen de la densitat de població de cada zona, de forma que les àrees amb més densitat donen lloc a unitats d'agregació més petites. A diferència de les seccions censals, però, les unitats creades es basen en una quadrícula de base, que es manté fixa. Això afavoreix que qualsevol parell de quadrícules creades amb aquesta metodologia poden ser comparades entre si, ja que tenen una base comú. De fet la mateixa llibreria permet, a partir de dues quadrícules diferents, obtenir-ne la unió, amb la resolució compartida per les dues. Per tant, ofereix estabilitat en cas de, per exemple, haver de comparar dades de períodes diferents, ja que assegura una base comú per qualsevol quadrícula creada.

Al mateix temps, la quadrícula utilitzada es basa en els sistemes de referència de coordenades estàndard, i utilitza la nomenclatura proposada en els sistemes geogràfics i estadístics europeus per tal de disposar de formats transnacionals comuns. Per tant, a més de ser estables en el temps, les quadrícules generades són també comparables entre diferents territoris. D'altra banda, en cas que calguin dades agregades a nivells més alts, les quadrícules generades parteixen d'un sistema jeràrquic que permet agrupar les quadrícules però sense tenir en

compte fronteres administratives com fan els sistemes clàssics (seccions censals, barris, districtes, municipis, i altres).

Disposar de totes les dades necessàries per a una determinada anàlisi a una resolució espacial molt detallada pot ser complicat en algunes ocasions. La llibreria desenvolupada ofereix un sistema dinàmic per tal de poder generar les dades *ad hoc*, ja que els temps d'execució són baixos, i permet parametritzar tant la resolució com el nivell de confidencialitat. Les quadrícules, utilitzen una codificació basada en sistemes de referència de coordenades fixes, per tant són reproduïbles en cas que hi hagi variació en les dades, per exemple, per dur a terme estudis longitudinals.

Hem pogut comprovar com, per als dos fenòmens estudiats, agregar les dades segons la metodologia presentada produeix resultats més precisos en l'anàlisi de patrons espacials. D'altra banda, al no disposar de dades per calcular les variables explicatives a resolucions més precises, no s'han pogut reproduir les anàlisis espacials realitzades en els dos articles per detectar possibles biaixos en els resultats deguts a l'agregació de les dades. En una futura línia d'investigació, seria interessant aprofundir en els efectes sobre els models estadístics espacials, agregant les dades en quadrícula i en seccions censals. La manca de dades a un nivell espacial precís implica que cal generar dades simulades que presentin interrelacions espacials predeterminades per tal de comparar la capacitat de detecció de correlacions espacials amb cada sistema d'agregació. Aquesta mateixa comparació es pot ampliar fàcilment utilitzant quadrícules a diverses escales, per tal d'avaluar la influència de l'escala amb independència de la definició de les unitats d'agregació. Així, la llibreria pot ser una bona eina per continuar investigant els efectes del *MAUP*.

Caldrà tenir en compte que la quadrícula limita la representació de les estructures de veïnatge, ja que hi ha discontinuïtat espacial entre les caselles. A diferència de les seccions censals, que cobreixen de manera exhaustiva tot el territori, això no passa amb la quadrícula. La quadrícula no és una divisió del territori si no una agregació de punts i, per tant, moltes caselles poden no estar connectades. Així doncs, la relació de veïnatge no pot utilitzar matrius de pesos

basades en el contacte de les àrees si no que cal representar les distàncies entre elles.

Serà interessant també poder treballar amb la informació del cens de població i habitatge de l'any 2021, principalment, amb les dades que es publicaran en format de quadrícula amb cel·les de 1km2, comuna per tota la UE. Les quadrícules generades amb la metodologia proposada a la tesi estan basades en aquesta mateixa definició, per tant, comptar amb aquesta informació pot aportar idees per incorporar a la llibreria, però, sobretot, proporcionarà noves dades per aprofundir en els estudis basats en quadrícules.

Com hem vist, la metodologia implementa un sistema de privacitat basat en el nombre mínim d'elements agregats en les unitats creades. D'aquesta forma s'assegura que un conjunt de dades no trenca els criteris de privacitat establerts. Seria interessant també, utilitzant dades simulades realitzar algun estudi per avaluar possibles riscos de re-identificació en cas de disposar de diverses quadrícules, per un mateix territori, a resolucions diferents. Per exemple, per a un determinat estudi, que requereixi dades amb un nivell de confidencialitat alt (dades sanitàries o econòmiques, per exemple) es generarà una quadrícula que segurament contindrà caselles més grans que una altra on només s'utilitzin dades molt genèriques. Aquesta segona quadrícula, junt amb el possible coneixement del territori, podria debilitar la confidencialitat de la quadrícula inicial.

Finalment, cal tenir en compte que la llibreria està implementada en R, i depèn directament de les classes i mètodes per dades espacials de la llibreria *sp* i de l'abstracció per als sistemes de projecció espacial de la llibreria *rgdal*. Aquestes llibreries s'han actualitzat recentment per incorporar canvis en els sistemes de transformació de coordenades. Totes les llibreries amb dependències directes, entre elles la llibreria *AQuadtree*, han hagut d'actualitzar-se a una nova versió. Tot i això, el procés d'actualització s'ha dut a terme amb molt pocs canvis. La llibreria *sf*, creada recentment, engloba moltes funcionalitats de les llibreries *sp*, *rgdal* i *rgeos*. De moment encara és força recent i no és compatible amb algunes llibreries d'anàlisi espacial que es basen en els objectes de la llibreria *sp*. Els objectes de la classe *sf* poden transformar-se en objectes *sp*, i al revés també,

per tant, la llibreria *AQuadtree*, realitzant les transformacions oportunes, és també compatible amb *sf*. Tanmateix, en una futura versió de la llibreria *AQuadtree* es preveu afegir els mètodes per tal de tractar directament amb objectes *sf* sense haver de realitzar transformacions.

Referències

5 Referències

Abdel-Aty, Mohamed, Lee, Jaeyoung, Siddiqui, Chowdhury, & Choi, Keechoo. (2013). Geographical unit based analysis in the context of transportation safety planning. *Transportation Research Part A: Policy and Practice*, *49*, 62-75.

Allen, T. F. H., O'Neill, R. V., & Hoekstra, T. W. (1984). *Interlevel relations in ecological research and management: some working principles from hierarchy theory. General Technical Report - US Department of Agriculture, Forest Service* (Vol. 110). US Department of Agriculture, Forest Service, Rocky Mountain Forest and~….

Allshouse, William B., Fitch, Molly K., Hampton, Kristen H., Gesink, Dionne C., Doherty, Irene A., Leone, Peter A., … Miller, William C. (2010). Geomasking sensitive health data and privacy protection: An evaluation using an E911 database. *Geocarto International*, *25*(6), 443-452.

Amini, Nima, Rashidi, Taha, Gardner, Lauren, & Waller, S. Travis. (2016). Spatial Aggregation Method for Anonymous Surveys: Case Study for Associations Between Urban Environment and Obesity. *Transportation Research Record: Journal of the Transportation Research Board*, *2598*(1), 27-36.

Amrhein, Carl G., & Reynolds, Harold. (1996). Using spatial statistics to assess aggregation effects. *Geographical Systems*, *3*(2-3), 143-158.

Anselin, Luc. (1995). Local indicators of spatial association—LISA. *Geographical analysis*, *27*(2), 93-115.

Anselin, Luc, & Griffith, Daniel A. (1988). Do spatial effecfs really matter in regression analysis? *Papers in Regional Science*, *65*(1), 11-34.

Arbia, Giuseppe. (1989). The configuration of spatial data in regional economics. En *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems. Advanced Studies in Theoretical and Applied Econometrics*. Dordrecht, Netherlands: Springer Science & Business Media.

Armstrong, Marc P., & Ruggles, Amy J. (1999). Map hacking: On the use of inverse address-matching to discover individual identities from point-mapped information sources. En *Geographic Information and Society Conference*. Minneapolis, Minnesota: University of Minnesota.

Armstrong, Marc P., Rushton, Gerard, & Zimmerman, Dale L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, *18*(5), 497-525.

Backer, Lars, & Bloch Holst, VV. (2011). *GEOSTAT 1A – Representing Census data in a European population grid*. *The European Forum for GeoStatistics*.

Banisar, David. (2019). Data Protection Laws Around the World Map. *SSRN Electronic Journal*.

Boone-Heinonen, Janne, Popkin, Barry M., Song, Yan, & Gordon-Larsen, Penny. (2010). What neighborhood area captures built environment features related to adolescent physical activity? *Health & Place*, *16*(6), 1280-1286.

Briant, A., Combes, P. P., & Lafourcade, M. (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, *67*(3), 287-302.

Brunsdon, Chris, Fotheringham, A. Stewart, & Charlton, Martin E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4), 281-298.

Cassa, Christopher A., Grannis, Shaun J., Overhage, J. Marc, & Mandl, Kenneth D. (2006). A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *Journal of the American Medical Informatics Association*, *13*(2), 160-165.

Cliff, Andrew David. (1973). *Spatial autocorrelation*. London: Pion.

Coffee, Neil T., Howard, Natasha, Paquet, Catherine, Hugo, Graeme, & Daniel, Mark. (2013). Is walkability associated with a lower cardiometabolic risk? *Health & Place*, *21*, 163-169.

Cressie, Noel. (1993). *Statistics for spatial data*. New York, NY, USA: John Wiley & Sons.

Deichmann, Uwe, Balk, Deborah, & Yetman, Greg. (2001). Transforming population data for interdisciplinary usages: From census to grid. *CIESIN, Washington (DC): Center for International Earth Science Information Network University Working Paper*, *200*(1). Recuperat de http://sedac.ciesin.org/gpw-v2/GPWdocumentation.pdf

Diez Roux, Ana V. (2001). Investigating Neighborhood and Area Effects on Health. *American Journal of Public Health*, *91*(11), 1783-1789.

Dmowska, Anna, & Stepinski, Tomasz F. (2017). A high resolution population grid for the conterminous United States: The 2010 edition. *Computers, Environment and Urban Systems*, *61*, 13-23.

Duncan, George T., Elliot, Mark, & Salazar-González, Juan-José. (2011). Why Statistical Confidentiality? En *Statistical Confidentiality* (p. 1-26). Springer.

Elhorst, J. Paul. (2010). Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis*, *5*(1), 9-28.

European Comission. (2009). Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics. *Official Journal of the European Union*, *L 87*, 164. Recuperat de http://data.europa.eu/eli/reg/2009/223/2015-06-08

Eurostat. (2019). *EU legislation on the 2021 population and housing censuses*.

Eurostat. (2020). Eurostat GridMaker. Recuperat de https://github.com/eurostat/GridMaker

Forbes, Angela, Naylor, Jane, Leaver, Victoria, Gare, Melissa, Hawkes, Tim, & Camden, Mike. (2009). Confidentiality plans for the 2011 censuses in the United Kingdom, Australia and New Zealand: a comparison. *Joint UNECE/Eurostat work session on Statistical Data Confidentiality, Bilbao*, 2-4.

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, *23*(7), 1025-1044. Recuperat de http://ideas.repec.org/a/pio/envira/v23y1991i7p1025-1044.html

Gatrell, Anthony C., Bailey, Trevor C., Diggle, Peter J., & Rowlingson, Barry S. (1996). Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology. *Transactions of the Institute of British Geographers*, *21*(1), 256.

Gauthier, Pierre A. (2002). Balancing the need for detail and confidentiality in the Canadian Census. En *Population Census Conference in Ulaan Baatar, Mongolia. Online at*.

Gehlke, Charles E., & Biehl, Katherine. (1934). Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association*, *29*(185A), 169-170.

Giuliani, Gregory, Ray, Nicolas, & Lehmann, Anthony. (2011). Grid-enabled Spatial Data Infrastructure for environmental sciences: Challenges and opportunities. *Future Generation Computer Systems*, *27*(3), 292-303.

Houston, Douglas. (2014). Implications of the modifiable areal unit problem for assessing built environment correlates of moderate and vigorous physical activity. *Applied Geography*, *50*, 40-47.

Hundepool, Anco, Domingo-Ferrer, Josep, Franconi, Luisa, Giessing, Sarah, Lenz, Rainer, Longhurst, Jane, … Wolf, P. (2010). Handbook on statistical disclosure control. *ESSnet on Statistical Disclosure Control*.

Institut Cartogràfic i Geològic de Catalunya. (2017). Població de Catalunya. Visualització i anàlisi. Recuperat de https://betaportal.icgc.cat/poblacio.

Institut d'Estadística de Catalunya. (2016). Població de Catalunya georeferenciada a 1 de gener de ... Barcelona: Generalitat de Catalunya.

Jacquez, Geoffrey M., & Rommel, Robert. (2009). Local indicators of geocoding accuracy (LIGA): theory and application. *International Journal of Health Geographics*, *8*(1), 60.

Jansson, Ingegerd. (2012). Issues and plans for the disclosure control of the Swedish Census 2011. En *Workshop on Statistical Disclosure Control of Census Data, Luxembourg*.

Jelinski, Dennis E., & Wu, Jianguo. (1996). The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, *11*(3), 129-140.

Johnston, Ron, Jones, Kelvyn, & Manley, David. (2019). Why geography matters. *Significance*, *16*(1), 32-37.

Katulic, Tihomir, & Protrka, Nikola. (2019). Information security in principles and provisions of the EU data protection law. En *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings* (p. 1219-1225).

Kelejian, Harry H., & Prucha, Ingmar R. (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of Geographical Systems*, *12*(3), 241-257.

Kim, Karl, Brunner, I. Made, & Yamashita, Eric Y. (2006). Influence of Land Use, Population, Employment, and Economic Activity on Accidents. *Transportation Research Record: Journal of the Transportation Research Board*, *1953*(1), 56-64.

Kolaczyk, Eric D., & Huang, Haiying. (2010). Multiscale Statistical Models for Hierarchical Spatial Aggregation. *Geographical Analysis*, *33*(2), 95-118.

Kounadi, Ourania, & Leitner, Michael. (2014). Why Does Geoprivacy Matter? The Scientific Publication of Confidential Data Presented on Maps. *Journal of Empirical Research on Human Research Ethics*, *9*(4), 34-45.

Kraak, Menno-Jan, & Ormeling, Ferjan. (2011). Cartography: visualization of spatial data. *Choice Reviews Online*, *49*(01), 49-0290-49-0290.

Kwan, Mei-Po, Casas, Irene, & Schmitz, Ben. (2004). Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, *39*(2), 15-28.

Lambert, Diane. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, *9*(2), 313–331.

Leitner, M., & Curtis, A. (2006). A first step towards a framework for presenting the location of confidential point data on maps—results of an empirical perceptual study. *International Journal of Geographical Information Science*, *20*(7), 813-822.

LeSage, James P. (2008). An Introduction to Spatial Econometrics. *Revue d'économie industrielle*, (123), 19-44.

Levin, Simon A. (1993). Concepts of Scale at the Local Level. *Scaling Physiological Processes: Leaf to Globe*, 7-19.

Longhurst, Jane, Tromans, Nicola, Young, Caroline, & Miller, Caroline. (2007). Statistical disclosure control for the 2011 UK census. En *Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester* (p. 17-19).

Lupp, Markus. (2007). *Styled Layer Descriptor profile of the Web Map Service* (Vol. 0). Open Geospatial Consortium. Recuperat de http://www.opengeospatial.org/standards/symbol

Machanavajjhala, Ashwin, Kifer, Daniel, Gehrke, Johannes, & Venkitasubramaniam, Muthuramakrishnan. (2007). ℓ-diversity: Privacy beyond k-anonymity. En *ACM Transactions on Knowledge Discovery from Data* (Vol. 1, p. 24).

Manski, Charles F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, *60*(3), 531.

Mazumdar, Soumya, Rushton, Gerard, Smith, Brian J., Zimmerman, Dale L., & Donham, Kelley J. (2008). Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics*, *7*(1), 13.

McCarty, Harold H., Hook, John C., & Knos, Duane S. (1956). *The measurement of association in industrial geography*.

Mitra, Raktim, & Buliung, Ron N. (2012). Built environment correlates of active school transportation: neighborhood and the modifiable areal unit problem. *Journal of Transport Geography*, *20*(1), 51-61.

Müller, M. (2006). *Styled Layer Descriptor Implementation Specification (1.0.0)*.

Nergiz, Mehmet Ercan, Atzori, Maurizio, & Clifton, Chris. (2007). Hiding the presence of individuals from shared databases. En *Proceedings of the ACM SIGMOD International Conference on Management of Data* (p. 665-676). New York, New York, USA: ACM Press.

Ninghui, Li, Tiancheng, Li, & Venkatasubramanian, Suresh. (2007). t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. En *Proceedings - International Conference on Data Engineering* (p. 106-115). IEEE.

O'Neill, R. V. (1979). Transmutations across hierarchical levels. En G. S. Innis & R. V. O'Neill (Ed.), *Systems Analysis of Ecosystems* (p. 59-78). Fairland, Maryland, USA: International Cooperative Publishing House.

O'Neill, R. V., & King, A. W. (1998). Homage to ST. Michael; Or, Why are there so many books on scale? *Ecological scale: Theory and applications*, 3-15.

Openshaw, Stan. (1977). A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling. *Transactions of the Institute of British Geographers*, *2*(4), 459.

Openshaw, Stan. (1978). An Empirical Study of Some Zone-Design Criteria. *Environment and Planning A: Economy and Space*, *10*(7), 781-794.

Openshaw, Stan. (1984). The modifiable areal unit problem. *CATMOG (Concepts & Techniques in Modern Geography)*, *38*.

Openshaw, Stan, & Alvanides, Seraphim. (1999). Applying geocomputation to the analysis of spatial distributions. *Geographical Information Systems: Principles and Technical issues*, *1*, 267-282.

Openshaw, Stan, & Taylor, P. J. (1979). A million or so correlation coefficients, three experiments on the modifiable areal unit problem. *Statistical Methods in the Spatial Sciences*, 127-144.

Orford, Scott. (2005). Cartography and visualization. *Questioning Geography: Fundamental Debates*, 189-205.

Paelinck, Jean H. P. (2000). On aggregation in spatial econometric modelling. *Journal of Geographical Systems*, *2*(2), 157-165.

Riva, Mylène, Apparicio, Philippe, Gauvin, Lise, & Brodeur, Jean-Marc. (2008). Establishing the soundness of administrative spatial units for operationalising the active living potential of residential environments: an exemplar for designing optimal zones. *International Journal of Health Geographics*, *7*(1), 43.

Robinson, W. S. (1950). Ecological correlations and the behavior of Individuals. *American Sociological Review*, *15*(3), 351-357.

Rushton, Gerard, Armstrong, Marc P., Gittler, Josephine, Greene, Barry R., Pavlik, Claire E., West, Michele M., & Zimmerman, Dale L. (2006). Geocoding in Cancer Research. *American Journal of Preventive Medicine*, *30*(2), S16-S24.

Sweeney, Latanya. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, *10*(5), 557-570.

Tammilehto-Luode, Marja. (2011). Opportunities and challenges of grid-based statistics. En *World Statistics Congress of the International Statistical Institute* (Vol. 58, p. 2451-2457). Dublin.

Truta, Traian Marius, & Vinay, Bindu. (2006). Privacy protection: P-Sensitive k-Anonymity property. En *ICDEW 2006 - Proceedings of the 22nd International Conference on Data Engineering Workshops* (p. 94).

United Nations. (2009). *Handbook on geospatial infrastructure in support of census activities*. *Studies in Methods*. New York.

Viegas, José Manuel, Martinez, L. Miguel, & Silva, Elisabete A. (2009). Effects of the Modifiable Areal Unit Problem on the Delineation of Traffic Analysis Zones. *Environment and Planning B: Planning and Design*, *36*(4), 625-643.

Visvalingam, Mahes. (1994). Visualisation in GIS. En H. M. Hearnshaw & D. J. Unwin (Ed.), *Cartography and ViSC* (p. 18-25). New York, NY, USA: Wiley.

Vu, Khuong, Zheng, Rong, & Gao, Jie. (2012). Efficient algorithms for K-anonymous location privacy in participatory sensing. En *Proceedings - IEEE INFOCOM* (p. 2399-2407).

Willenborg, L., & de Waal, T. (2000). *Elements of Statistical Disclosure Control*. *Lecture Notes in Statistics, Springer Verlag, New York* (Vol. 1). Springer Science & Business Media.

Wong, David W. S. (2004). The Modifiable Areal Unit Problem (MAUP). En *WorldMinds: Geographical Perspectives on 100 Problems* (p. 571-575). Dordrecht: Springer Netherlands.

Wong, David W. S., Lasus, H., & Falk, R. F. (1999). Exploring the Variability of Segregation Index D with Scale and Zonal Systems: An Analysis of Thirty US Cities. *Environment and Planning A: Economy and Space*, *31*(3), 507-522.

Wong, Raymond Chi-Wing, Li, Jiuyong, Fu, Ada Wai-Chee, & Wang, Ke. (2006). (α, k)-anonymity. En *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06* (p. 754). New York, New York, USA: ACM Press.

Wu, Jianguo. (2004). Effects of changing scale on landscape pattern analysis: scaling relations. *Landscape Ecology*, *19*(2), 125-138.

Xiao, Xiaokui, & Tao, Yufei. (2007). M-invariance: Towards privacy preserving re-publication of dynamic datasets. En *Proceedings of the ACM SIGMOD International Conference on Management of Data* (p. 689-700). New York, New York, USA: ACM Press.

Zandbergen, Paul A. (2014). Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Advances in Medicine*, *2014*, 1-14.

Zayatz, Laura. (2007). Disclosure avoidance practices and research at the US Census Bureau: An update. *Journal of Official Statistics*, *23*(2), 253-265.

Zhang, Ming, & Kukadia, Nishant. (2005). Metrics of Urban Form and the Modifiable Areal Unit Problem. *Transportation Research Record: Journal of the Transportation Research Board*, *1902*(1), 71-79.

Zimmerman, Dale L., Fang, Xiangming, & Mazumdar, Soumya. (2008). Spatial clustering of the failure to geocode and its implications for the detection of disease clustering. *Statistics in Medicine*, *27*(21), 4254-4266.