Introducció a la privadesa en la publicació de dades

PID_00274682

Guillermo Navarro-Arribas

Temps mínim de dedicació recomanat: 5 hores

Universitat Oberta de Catalunya

Guillermo Navarro-Arribas

Professor agregat al Departament d'Enginyeria de la Informació i de les Comunicacions de la Universitat Autònoma de Barcelona. El seu àmbit de recerca se centra en la seguretat i la privadesa informàtica i, més concretament, en la privadesa de dades, la privacitat en sistemes de comunicació i la tecnologia blockchain. Participa activament en projectes de recerca i és autor de diversos articles en el camp de la privacitat. Actualment imparteix docència sobre seguretat, privacitat i criptografia.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Cristina Pérez Solà

Com citar aquest recurs d'aprenentatge amb l'estil Harvard:
Navarro-Arribas, G. (2020) *Introducció a la privadesa en la publicació de dades*. [Recurs d'aprenentatge textual]. 1a. ed. Barcelona: Fundació Universitat Oberta de Catalunya (FUOC).

Primera edició: setembre 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Guillermo Navarro-Arribas
Producció: FUOC

Índex

Introducció

En moltes ocasions ens trobem davant la necessitat o possibilitat de fer públiques dades que poden contenir informació privada. Un exemple pot ser la publicació de dades censals, dades sobre l'activitat econòmica d'una població, o dades de pacients de centres sanitaris o estudis mèdics. Aquestes dades poden ser molt útils per fer estudis i recerca en molts camps, des d'estudis de mercat fins a recerca mèdica o social. No obstant això, les dades poden contenir informació privada relativa als individus o entitats que apareixen en els estudis que no volem o no podem revelar. Això impedeix o dificulta la publicació o cessió de dades en general. Aquesta publicació de dades es pot fer en diversos graus: les dades simplement es poden fer públiques o es poden cedir a terceres parts de manera exclusiva. Fins i tot en aquest últim cas, en què pot haver-hi contractes de confidencialitat en la cessió de dades, és possible que es necessiti igualment aplicar algun tipus d'anonimització. Avui en dia també són comunes iniciatives d'administracions públiques per publicar dades en obert com a part de programes de transparència. En aquest sentit volem poder publicar aquestes dades per a investigadors o el públic en general sense posar en risc per això la privadesa de les entitats o individus que hi apareixen.

Aquestes dades que es publiquen poden contenir informació privada de molts tipus i formes per als individus o entitats als quals fan referència. La informació privada pot estar exposada de manera molt clara i explícita, per exemple, amb la publicació de la identitat dels pacients d'un estudi clínic. En aquest cas la seva protecció pot ser relativament senzilla, atès que és fàcil identificar quina informació i quines dades cal protegir. Hi ha, però, la possibilitat que la informació privada es pugui obtenir de manera més tangencial o no tan explícita. Sol ser el cas de dades que aparentment no revelen informació privada per si soles però poden filtrar de manera total o parcial identitats o informació privada mitjançant la unió d'altres dades o mitjançant algun coneixement previ del possible atacant.

Tradicionalment, aquest problema s'ha tractat des de diversos àmbits. Investigadors d'agències d'estadística o mineria de dades i aprenentatge automàtic han treballat aquest problema i han donat lloc a disciplines com *statistical disclosure control* (SDC) o *privacy preserving data mining* (PPDM). Més recentment, s'ha acostumat a utilitzar el terme *tecnologies de millora de la privadesa* o *privacy enhancing technologies* (PET) per denominar totes les tecnologies encaminades a proporcionar privadesa, entre les quals hi ha les destinades a publicar dades.

En aquest mòdul estudiarem el problema associat a la privadesa en la publicació de dades. Veurem com s'aborda el problema, quines tècniques s'utilitzen per garantir certs nivells de privadesa i quines conseqüències té l'aplicació d'a-

quests mètodes de protecció. Ens centrarem en un tipus de dades relativament senzilles, les microdades, deixant l'estudi de privadesa en dades més complexes per a altres mòduls.

Objectius

Els objectius que l'estudiant ha d'assolir una vegada treballats els continguts d'aquest material didàctic són els següents:

1. Entendre els problemes associats a la publicació de dades privades.
2. Estudiar com es mesura i avalua la privadesa i pèrdua d'informació en la publicació de dades.
3. Entendre models de privadesa com la *k*-anonimitat i la privadesa diferencial.
4. Conèixer els mètodes principals d'anonimització de dades de propòsit general.

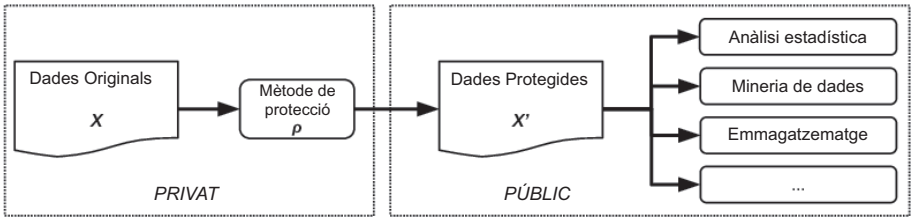
1. Introducció a la privadesa en la publicació de dades

En aquest mòdul donarem una visió general del problema de la privadesa en la publicació de dades simples. Introduirem el problema, el tipus de dades amb què treballarem, i plantejarem qüestions introductòries. Com veurem, un aspecte important és precisament com mesurar i avaluar la privadesa associada a unes dades. En aquest sentit veurem tècniques i mètodes utilitzats en l'avaluació de mètodes de protecció per revisar després alguns dels mètodes més populars en aquest camp.

1.1. El problema de la privadesa de dades

L'escenari que preveiem en la publicació de dades pot variar depenent del context, tipus de dades i ús que es vulgui donar a aquestes dades. De manera simplificada, considerem una situació com la mostrada en la figura 1. En aquest cas tenim un conjunt de dades X a les quals apliquem un mètode de protecció ρ, i obtenim una versió protegida de X denotada com a X', de manera que $X' = \rho(X)$.

Figura 1. Escenari simple de publicació de dades privades



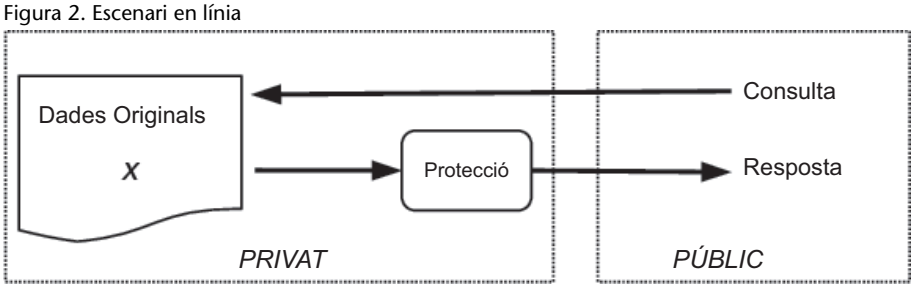
Les dades, una vegada protegides, poden tenir diferents usos o aplicacions. Per posar un exemple, es poden usar dades en estudis estadístics, aplicar tècniques de mineria de dades o aprenentatge automàtic, o simplement es poden distribuir o emmagatzemar dades per al seu ús futur.

La protecció de les dades originals la sol fer el propietari d'aquestes dades o l'entitat que en té la custòdia. L'objectiu que es planteja és que aquesta entitat pugui fer públiques les dades garantint que es preserva un nivell de privadesa o anonimat concret. Podem fer X' públic mantenint les dades originals X privades. Per a això, el mètode de protecció ρ aplicat haurà de garantir certes propietats o aplicar mesures concretes destinades a evitar la publicació de dades sensibles o considerades com a privades.

Encara que aquest és l'escenari principal que estudiarem, hi ha variacions interessants a tenir en compte. Una d'elles distingeix una publicació fora de línia

d'una en línia. El cas de la figura 1 correspondria al cas fora de línia, en què es publica una versió protegida de les dades i l'analista treballa directament sobre les dades publicades. En un escenari en línia, l'analista fa peticions a la base de dades i es protegeix la resposta que es retorna (figura 2).

Figura 2. Escenari en línia



1.1.1. Tipus de dades: microdades

En aquest mòdul ens centrarem en l'anonimització d'un tipus de dades conegudes com a **microdades** (en anglès, *microdata*).

Un fitxer de **microdades** es pot veure com una taula o matriu on cada fila correspon a un individu o entitat, que també denotem com a **registre**, i cada columna és un **atribut o variable** sobre els individus o entitats.

Denotarem com a X un conjunt de microdades, cada fila com a X_i per a $i = 1,\ldots,N$, i cada atribut o variable com a V_j, on $j = 1,\ldots,M$ (vegeu la taula 1), de manera que x_{ij} denota el valor de la variable V_j per a l'individu X_i.

Fitxer de microdades

En la definició de microdades parlem de *fitxer* perquè tradicionalment la manera més comuna de publicar microdades ha estat en fitxers CSV (*comma-separated values*). No obstant això, es poden publicar microdades en qualsevol altre suport, com taules en una base de dades, documents amb altres formats (JSON, Excel, ODS...) o fluxos de dades, etc.

Taula 1. Notació de microdades

	V_1	V_2	...		...	V_M
X_1				$\vdots$		
X_2				$\vdots$		
$\vdots$				$\vdots$		
	...	...	...	x_{ij}	...	...
$\vdots$				$\vdots$		
X_N				$\vdots$		

És important remarcar que en molts casos, com el de les agències d'estadística, és comuna la publicació de **dades tabulars** (*tabular data* en anglès), que contenen dades agregades a partir de microdades. Per exemple, es poden publicar les mitjanes d'edat d'una població per sexe.

Encara que aquest tipus de dades tabulars són útils, la publicació de microdades proporciona major flexibilitat a l'analista a l'hora de fer el seu estudi o a l'hora d'aplicar tècniques de mineria de dades o aprenentatge automàtic. Les dades tabulars no estan exemptes de riscos de privadesa, però en publicar dades no agregades les microdades solen ser més sensibles a la revelació d'informació privada.

Dades tabulars

Per obtenir informació sobre protecció de dades tabulars, podeu consultar: **J. Castro** (2012). «Recent advances in optimization techniques for statistical tabular data protection». *European Journal of Operational Research* (vol. 216, núm. 2, pàg. 257-269). https://bit.ly/2odvqy4

En aquest mòdul ens centrarem en microdades deixant una mica de costat la privadesa en dades tabulars. Aquestes solen ser estudiades en el camp de la recerca operativa i l'optimització, i el lector interessat en pot trobar més informació en la bibliografia del mòdul.

Exemple

Com a exemple senzill de microdades, tenim la taula 2, on es recullen diversos atributs o variables (columnes, V_j) respecte a individus concrets (files, X_i). Per exemple, el valor X_{42} denota el valor *Professor*, és a dir, la variable V_2 (en aquest cas *Occupation*) per a l'individu X_4 (en aquest cas *Sebastião Rodrigues*).

Taula 2. Exemple de microdades

Name	*Occupation*	*ZIP*	*Age*	*Sex*	*Income*
Tuco Benedicto	Runner	80222	42	M	25050
Jill McBain	School teacher	40831	28	F	18098
Ramon Miguel Vargas	Police det.	97206	51	M	28760
Sebastião Rodrigues	Professor	40505	23	M	26062
Mathilda Lando	Student	40831	12	F	790
Hank Quinlan	Police capt.	97206	68	M	30033
Margaret Fitzgerald	Boxer	80237	30	F	55093
Tom Doniphon	Horse rider	80911	46	M	17330
Ellen Ripley	Military lt.	97201	55	F	15700

Les dades contingudes en un fitxer de microdades poden ser, al seu torn, de diversos tipus. Podem tenir dades categòriques o numèriques de diferents formats o tipus. Entre les categòriques pot haver-hi dades nominals o ordinals. Podem tenir dades que facin referència a dates, URL, o que puguin tenir o no una interpretació semàntica intrínseca o mitjançant una ontologia. Com veurem, això condiciona el mètode de protecció que es pot aplicar o la manera en què s'aplica.

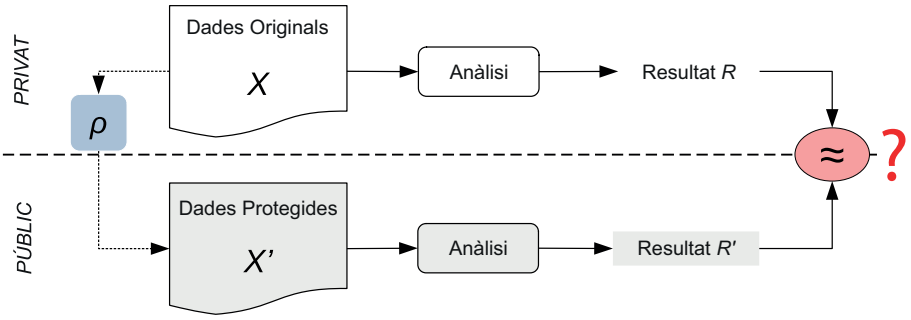
Algunes de les idees i principis de les tècniques i mètodes que veiem en aquest mòdul són aplicables també a dades més complexes que les microdades. Per exemple, en el mòdul següent veurem la protecció d'un altre tipus de dades com els grafs.

1.1.2. Privadesa i pèrdua d'informació

Aplicar un mètode de protecció a unes dades comporta alterar en certa manera aquestes dades. Aquesta alteració o distorsió de les dades ajuda a reduir el risc de revelació d'informació privada de les dades, però al seu torn comporta una pèrdua d'informació. En alterar les dades, perdem informació original que pot fer que aquestes dades perdin utilitat.

Per exemple, podem tenir unes dades mèdiques X sobre alguna malaltia que s'utilitzaran per fer estudis estadístics sobre la incidència de factors de risc en aquesta malaltia. Aquestes dades són distorsionades utilitzant un mètode de protecció ρ per mantenir la privadesa de les persones que apareixen, però aquesta distorsió pot fer que l'estudi estadístic sigui afectat. Podria ser que els resultats obtinguts a partir de X' (recordem que $X' = \rho(X)$) siguin diferents dels que obtindríem en fer el mateix estudi directament sobre X, com il·lustra la figura 3. Aquesta diferència podria ser menyspreable o molt petita i delimitada, o per contra ens podríem trobar amb una gran diferència en els resultats de manera que no es pogués utilitzar X' per fer aquest estudi perquè incorreríem en un error massa gran.

Figura 3. Resultats obtinguts de dades originals X i dades protegides



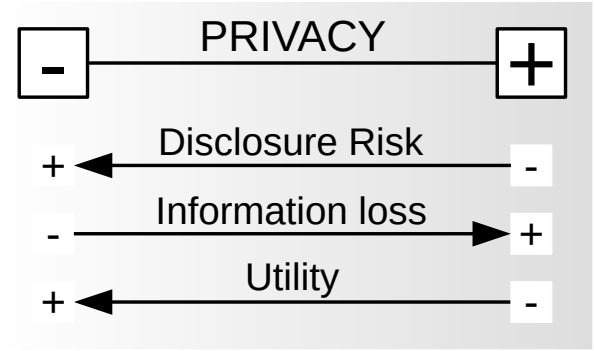
Un dels principals problemes que es tracta en privadesa de dades és precisament aconseguir un bon equilibri entre la privadesa i la utilitat. En general parlarem dels conceptes següents associats a un conjunt de dades:

- **Risc de revelació** (*disclosure risk*). És el risc que les dades proporcionin informació sensible o considerada com a privada. L'objectiu serà sempre intentar minimitzar el risc de revelació.
- **Pèrdua d'informació** (*information loss*). És la quantificació d'informació que es perd en aplicar un mètode de protecció. Això ens determina l'error que podem cometre en utilitzar les dades protegides en lloc de les dades originals a l'hora de fer una anàlisi estadística o aplicar un model d'aprenentatge automàtic. En aquest cas sempre buscarem minimitzar la pèrdua d'informació.
- **Utilitat** (*utility*). És la mesura de la utilitat de les dades per fer estudis sobre elles. Generalment, la utilitat d'unes dades protegides es considera respecte a la utilitat de les originals. Aquesta mesura es pot veure com a inversament proporcional a la pèrdua d'informació, fins al punt que moltes vegades

s'usen les dues mesures de manera indistinta. Se solen utilitzar les mateixes mesures simplement canviant-ne la interpretació. En el cas de la utilitat, voldrem maximitzar-la.

Com veiem en la figura 4, aquestes mesures estan molt relacionades. Si es minimitza el risc de revelació, s'aconsegueix més privadesa però també augmenta la pèrdua d'informació o, el que és equivalent, es redueix la utilitat. En general, no és possible, per exemple, reduir el risc de revelació i reduir al mateix temps la pèrdua d'informació.

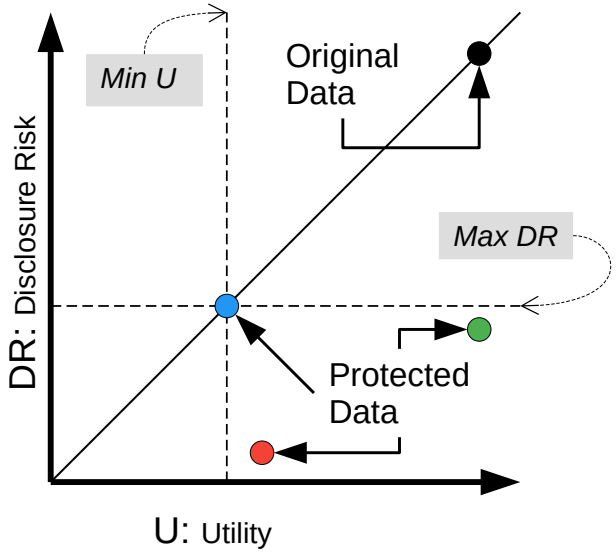
Figura 4. Risc de revelació, pèrdua d'informació i utilitat respecte a la privadesa



L'objectiu que se'ns planteja en la privadesa de dades és trobar un bon equilibri entre aquestes mesures. Volem reduir el risc de revelació fins a un nivell tolerable per mantenir un nivell de privadesa concret, però al mateix temps no volem perdre molta informació, volem mantenir la utilitat de les dades en un nivell acceptable.

En la figura 5 veiem el risc de revelació respecte a la utilitat. En general, podem situar les dades originals en el quadrant superior dret. És a dir, les dades originals tenen un alt risc de revelació i una gran utilitat. Per contra, les dades protegides se solen trobar en el quadrant inferior esquerre (baix risc de revelació i baixa utilitat).

Figura 5. Risc de revelació enfront d'utilitat



Un bon mètode de protecció és el que aconsegueix un bon equilibri entre risc de revelació i utilitat. En alguns casos es parteix d'un nivell de privadesa mínim tolerable (línia discontínua horitzontal), és a dir, un nivell màxim de risc de revelació, i l'objectiu seria aconseguir un mètode que, respectant aquesta quota màxima de revelació, proporcioni la major utilitat possible (és el cas del punt verd en la figura 5). També es pot plantejar el problema de manera inversa: partir d'un nivell d'utilitat mínim tolerable (línia discontínua vertical) i buscar el mètode que aconsegueixi minimitzar el risc de revelació en aquesta utilitat mínima (és el cas del punt vermell en la figura 5).

Com era d'esperar, en la pràctica, buscar aquest equilibri en la protecció de dades no és fàcil. Encara més, casos com els descrits anteriorment amb els punts vermell i verd no solen ser possibles. En general, donada una quota màxima de risc de revelació, no podrem augmentar la utilitat tant com vulguem. En un món ideal ens agradaria poder situar les nostres dades protegides en el quadrant inferior dret de la figura 5, però en la realitat ens trobarem en l'inferior esquerre. Això és així perquè, com hem comentat, aquestes mesures solen ser intrínsecament inverses. En el cas general, no podrem reduir el risc de revelació sense reduir la utilitat de les dades.

El punt blau de la figura 5 presentaria un bon balanç entre risc de revelació i utilitat. Si considerem les línies discontínues com els límits de risc de revelació i utilitat, és a dir, el màxim risc de revelació i el mínim d'utilitat tolerables, el punt blau assoleix el compromís entre les dues. En alguns casos pot ser que aquestes línies no se situïn de manera tan simètrica, imposant més restriccions en una mesura que en una altra.

Una altra dificultat important a què ens enfrontem a l'hora d'abordar el problema és com mesurar el risc de revelació, i la utilitat o pèrdua d'informació. Com veurem en els subapartats següents, hi ha diverses alternatives per a aquest tipus de mesures, i cadascuna té alguna particularitat, avantatge o inconvenient. Per exemple, no hi ha una única mesura que capturi de manera absoluta el risc de revelació associat a unes dades. El que tenim són diferents maneres de mesurar el risc de revelació o, més ben dit, diferents característiques que ens serveixen per estimar el risc de revelació en la mesura del possible. Això vol dir que podem tenir unes dades que donin un resultat molt bo amb una mesura de risc de revelació concreta, però no per això podem assumir que són totalment segures. Es podria donar el cas que es desenvolupés algun atac que aprofités alguna característica no prevista per aquesta mesura de risc de revelació concreta.

Podem dir que en el cas de la privadesa de dades no hi ha una solució perfecta. Sempre hem de jugar amb l'equilibri entre risc de revelació i utilitat, i pot resultar difícil fer afirmacions categòriques sobre la seguretat associada a unes dades concretes. Com diu una màxima molt aplicada en seguretat informàtica, la seguretat absoluta no existeix en la pràctica. Pel que respecta a

la privadesa de dades, podem dir que en el cas general no existeix la solució perfecta a la publicació de dades segures sense risc de revelació i sense pèrdua d'informació.

1.1.3. Classificacions de mètodes de protecció

A l'hora de considerar els mètodes de protecció, podem plantejar diverses classificacions o perspectives des dels quals abordar-los. Una classificació genèrica i que resulta interessant com a introducció consisteix a veure els mètodes de privadesa de dades en funció de l'ús que es vol fer de les dades protegides. Podem considerar la següent classificació de mètodes de protecció:

- De **propòsit específic** o orientats a computació (*computation-driven*). En aquest cas es volen protegir dades per fer una anàlisi (o modelatge) concreta i coneguda per endavant. Sabem la *computació* que es vol fer amb les dades protegides. Això fa que es puguin aplicar mètodes de protecció específicament dissenyats i parametritzats per a aquesta anàlisi. Es poden centrar els esforços a aconseguir minimitzar la pèrdua d'informació precisament en aquells aspectes més rellevants de cara a l'anàlisi que sabem que s'aplicarà. En aquesta categoria es descriuen també els mètodes **orientats a resultat**, en els quals considerem que les dades s'utilitzaran en un model concret de mineria de dades o aprenentatge automàtic i ens interessa la privadesa associada als resultats obtinguts a partir d'aquest model. En aquest últim cas la privadesa s'estudia en relació amb el coneixement que es pot inferir a partir de les dades i no tant en relació amb les dades per si mateixes.

- De **propòsit general** o orientats a dades (*data-driven*). És el cas en què no hi ha un ús específic per a les dades. Les dades protegides no seran usades en una sola anàlisi concreta o no coneixem l'estudi, model o ús que es farà de les dades protegides. En aquest sentit es busca obtenir el millor equilibri entre pèrdua d'informació i risc de revelació de manera general, sense afavorir una o altra característica de manera especial.

Els mètodes de protecció poden utilitzar diverses tècniques. Ens interessa destacar entre elles l'ús de mecanismes criptogràfics en mètodes de protecció. En aquest sentit podem distingir aquests mètodes:

- **Criptogràfics**. Fan ús de criptografia. Un cas interessant és poder publicar dades xifrades sobre les quals es pugui operar i fer anàlisis concretes sense necessitat de desxifrar-les. Aquest tipus de mètodes sol ser usat en mecanismes de propòsit específic o orientats a computació, en què sabem quina anàlisi (o operació) es farà sobre les dades. En aquest cas es pot definir un protocol criptogràfic per calcular una funció concreta sobre un conjunt de dades xifrades. És comú en aquests casos usar eines criptogràfiques com a

computació segura multipartita (*secure multiparty computation*) o criptografia homomòrfica.

- De ***masking***. Amb aquest terme ens referim als mètodes que utilitzen algun tipus d'alteració de les dades, com afegir soroll. Les dades no es xifren i es pot operar amb elles sense necessitat de protocols o eines criptogràfiques. Aquestes tècniques són les utilitzades en els mètodes de propòsit general o orientats a dades. En aquest cas es pot operar directament amb les dades i aplicar qualsevol tipus d'anàlisi.

En aquest mòdul ens centrarem en els mètodes de propòsit general i, per tant, en tècniques de *masking*. En el subapartat 1.4. veurem que n'hi ha de diferents tipus i repassarem algunes de les més conegudes.

1.2. Mesures de privadesa

En aquest subapartat revisem com estimar el nivell de privadesa que s'aconsegueix en aplicar un mètode de protecció, o simplement el risc de revelació (*disclosure*) associat a un conjunt de dades concret. De manera general, considerem els fets següents.

> Es produeix **revelació (*disclosure*)** d'informació privada quan es pot obtenir coneixement nou sobre un individu o registre del conjunt de microdades.

L'atacant pot tenir un coneixement previ, que veurà augmentat en analitzar les microdades publicades.

L'objectiu d'un atacant és obtenir la major quantitat d'informació a partir de les dades publicades. Quan aconsegueix identificar un individu o entitat en les dades protegides, diem que s'ha produït una **reidentificació**. En aquest cas l'atacant pot establir la correspondència exacta entre un registre de les dades protegides i un individu o entitat.

No tots els atributs o variables d'un mateix conjunt de dades presenten el mateix risc de revelació. És comú classificar atributs en les categories següents:

- **Identificadors**: atributs que identifiquen un individu per si sols i de manera única. Exemples típics són: nom, número de la Seguretat Social, DNI, etc.

- **Quasiidentificadors**: atributs que no identifiquen un individu per si sols, però en combinació entre ells i/o amb informació externa poden identi-

ficar algun individu. Per exemple: edat, codi postal, ciutat de residència, sexe, professió, data de naixement, etc.

- **Confidencials**: atributs sensibles respecte a l'individu. Alguns exemples són: malaltia, salari, afiliació política, etc.

- **No confidencials**: atributs que no tenen informació sensible sobre l'individu.

Aquestes categories no són disjuntes, i en alguns casos la classificació pot ser complicada, ja que depèn del context associat a les dades i la seva publicació.

A l'hora de protegir un fitxer de microdades, sembla clar que no podem publicar els atributs identificadors. Encara més, tampoc són susceptibles de ser protegides mitjançant algun mètode de protecció.

Exemple

No té molt sentit distorsionar amb soroll, per exemple, el nom o el DNI d'un individu en unes dades protegides. En distorsionar-lo perden la seva única utilitat, que és precisament identificar el registre, i no tenen interès.

És per això que els identificadors s'eliminen o es xifren, i és als quasiidentificadors que s'aplica el mètode de protecció. En alguns casos, podem tenir algun atribut confidencial al qual no s'aplica el mètode de protecció i, per tant, no es modifica. Això és així perquè poden ser l'objecte d'estudi en les microdades i es vol minimitzar l'error associat a aquest atribut. A més, com que és confidencial, s'assumeix que un possible atacant no en tindrà coneixement.

Exemple

En la taula 2 podem veure diversos atributs. Entre ells, considerem *Name* com a identificador; *Occupation*, *ZIP*, *Age* i *Sex* com quasiidentificadors no confidencials, i *Income* com a atribut confidencial. Depenent del context, l'atribut *Income* podria ser considerat també com a quasiidentificador. Més endavant veurem que això pot tenir implicacions de cara a com es protegeix tota la taula, ja que aquesta protecció s'aplica als atributs quasiidentificadors.

És important remarcar que no és suficient protegir dades eliminat simplement els atributs identificadors. Encara que unes dades sense atributs identificadors podrien semblar anònimes, no ho són. Un atacant podria obtenir informació sobre els individus a partir dels atributs quasiidentificadors.

Exemple

En la taula 3 es mostren les microdades de l'exemple anterior (vegeu la taula 2) sense el *nom*, que hem considerat com a identificador. Com a exemple, podem considerar un atacant que vol reidentificar *Ellen Ripley* per saber quant cobra, i l'única cosa que en sap és que és militar. Simplement amb aquest coneixement previ, l'atacant pot identificar de manera inequívoca quin registre correspon a *Ellen Ripley* i saber que té uns ingressos de *15.700*. L'atacant no ha necessitat atributs identificadors, ja que ha pogut reidentificar un registre simplement amb el coneixement d'un quasiidentificador.

Taula 3. Microdades sense atributs identificadors

Occupation	*ZIP*	*Age*	*Sex*	*Income*
Runner	80222	42	M	25050
School teacher	40831	28	F	18098
Police det.	97206	51	M	28760
Professor	40831	23	M	26062
Student	40831	12	F	790
Police capt.	97206	68	M	30033
Boxer	80237	30	F	55093
Horse rider	80911	46	M	17330
Military lt.	97201	55	F	15700

El problema de l'exemple anterior és que hi ha atributs que poden identificar de manera única algun registre. És a dir, hi ha atributs que tenen valors **únics**, que al costat del coneixement previ de l'atacant poden ajudar a reidentificar registres concrets. Això mateix pot passar amb combinacions d'atributs.

Exemple

Seguint amb l'exemple anterior, suposem que volem saber el que cobra *Jill McBain*. L'atacant sap que viu en el codi postal 40831 i que és dona. Amb aquest coneixement previ, l'atacant no pot reidentificar el registre que correspon a *Jill McBain*, però sí sabrà que guanya 18.098 o 790. En aquest cas podem dir que l'atacant té una probabilitat de reidentificació de 0,5 (identifica dos possibles registres que poden correspondre a *Jill MacBain*). Fixeu-vos que si l'atacant coneix solament el codi postal aquesta probabilitat serà del 0,33, i si coneix solament que és dona, serà del 0,2.

Identificació única mitjançant quasiidentificadors

Per veure el detall de l'estudi comentat en el text, podeu consultar L. Sweeney (2000). «Simple demographics often identify people uniquely?». *Data Privacy Working Paper* 3. Pittsburgh: Carnegie Mellon University. https://bit.ly/3fGYq1J

La unicitat de valors quasiidentificadors és sens dubte un problema de privadesa i és més comuna del que pot semblar. Com a curiositat, investigadors van determinar que es podia identificar de manera única el 87,1% de la població dels Estats Units amb tres senzills quasiidentificadors: sexe, data de naixement i codi postal.

De manera més concreta, distingim dos tipus de revelació:

- **Revelació d'atribut**. Es produeix quan un atacant pot adquirir informació nova respecte al valor d'un atribut d'un individu o registre. Aquesta informació nova pot ser el valor concret o simplement l'augment en la precisió de la informació sobre el valor (per exemple, un interval en el cas d'una variable numèrica).

- **Revelació d'identitat**. Es produeix quan l'atacant pot reidentificar de manera inequívoca el registre que correspon a un individu o entitat concrets, és a dir, quan es produeix una reidentificació.

A continuació veiem l'exemple anterior i el tipus de revelació que té lloc en cada cas.

Exemple: revelació d'atribut i identitat

Suposem que l'atacant intenta obtenir informació de la taula 3 basant-se en un coneixement previ concret:

- **Revelació d'atribut**. Si l'atacant sap que l'individu que busca és dona i viu en el codi postal 40831, pot saber que el seu salari és 18.098 o 790. Encara que no pot saber quin registre correspon a l'individu que busca, sí que sap que pot ser el segon o el cinquè, cosa que li permet millorar la precisió del coneixement que té sobre aquest subjecte. No solament el coneixement es refereix a l'atribut confidencial, també l'atacant pot obtenir informació d'altres atributs quasiidentificadors, en aquest cas que pot tenir una edat propera a 28 o 12.

- **Revelació d'identitat**. Si l'atacant sap que l'individu que busca és militar de professió, sabrà que es tracta de l'últim registre i, per tant, que el seu salari és de 15.700. En aquest cas l'atacant ha pogut reidentificar sense ambigüitat el registre que correspon a l'individu que cerca.

A continuació veurem com es pot mesurar la privadesa d'un conjunt de dades de manera més precisa.

1.2.1. Tipus de mesures de privadesa

Mesurar el nivell de privadesa o anonimat d'un conjunt de dades no és senzill i depèn moltes vegades de factors externs com el coneixement previ que tingui l'atacant sobre les dades. És important, no obstant això, poder establir algun indicador sobre la privadesa que s'aconsegueix en aplicar un mètode de protecció o senzillament alguna mesura del risc de revelació associat a unes dades concretes.

Aquí veurem diverses possibilitats. D'una banda, podem veure la revelació com una mesura booleana en què, donada una propietat o condició concreta, podem establir si les dades o el mètode de protecció la compleixen o no. És el cas de les propietats de *k*-anonimitat i privadesa diferencial. Per altra banda, també podem determinar la revelació basant-nos en alguna mesura quantificable. Veurem el cas d'establir mesures de revelació d'identitat quantificables basades en risc de reidentificació i unicitat.

1.2.2. *k*-anonimitat

La *k*-anonimitat (o *k*-anonimat) és un model de privadesa molt popular que podem definir de la manera següent.

Un conjunt de microdades X compleix ***k*-anonimitat** respecte als quasiidentificadors de X si qualsevol combinació de valors dels quasiidentificadors apareix com a mínim k vegades.

Podem dir que el conjunt X compleix k-anonimitat si sempre hi ha k registres com a mínim indistingibles entre ells respecte als seus quasiidentificadors. El conjunt de registres indistingibles rep també el nom de **classe d'equivalència** i es pot veure com un **conjunt d'anonimat**.

Originàriament, es defineix en el context de la privadesa en microdades, i per això la definició fa referència als quasiidentificadors. Avui dia, però, s'aplica en molts tipus de dades i requereix que els valors observables de les dades compleixin k-anonimitat.

Una definició més genèrica i aplicable més enllà de les microdades és que un conjunt de dades compleix k-anonimitat si sempre hi ha k elements com a mínim indistingibles entre ells. Això permet estendre aquest concepte a tot tipus de dades o successos observables o mesurables.

Exemple

En la taula 4 es mostra un exemple de microdades 3-anònims respecte als quasiidentificadors *occupation*, *ZIP* i *sex*. Com s'observa, hi ha sempre tres registres indistingibles respecte a aquests quasiidentificadors. En general, podem dir que la probabilitat de reidentificació o revelació d'identitat és, com a màxim, de 1/3. Un atacant amb algun coneixement previ mai podrà distingir quin registre d'entre els tres correspon a l'individu que busca (assumint que no coneix o té informació sobre l'atribut confidencial *income*).

Taula 4. Microdades k-anònims respecte als quasiidentificadors *Occupation*, *ZIP*, *Sex*

Occupation	ZIP	Sex	Income
Teacher	80122	M	10000
Teacher	80122	M	18098
Writer	97234	F	28000
Teacher	80122	M	20100
Writer	97234	F	30000
Writer	97234	F	9210

Cal destacar que en la taula 4 de l'exemple anterior veiem tres quasiidentificadors i un atribut confidencial al qual no s'ha aplicat el mètode de protecció. Com es comenta en el subapartat 1.1.1., això és comú de vegades i parteix de l'assumpció que aquest atribut, com que és confidencial, no pot ser conegut per l'atacant i, per tant, no comporta un risc per a la reidentificació. Això permet reduir la pèrdua d'informació respecte a aquest atribut, que sol ser l'objecte d'estudi. Per aquest motiu, la taula 4 es considera k-anònima respecte als quasiidentificadors. Podríem protegir també l'atribut confidencial i considerar la taula com a k-anònima respecte a tots els atributs.

La k-anonimitat es pot veure com una mesura de risc de revelació d'identitat, ja que prevé la reidentificació de registres; no obstant això, aquesta pot ser que no sigui suficient per protegir la revelació d'atribut.

Per exemple, en la taula 5 veiem unes dades que compleixen 3-anonimitat. Es tracta d'unes microdades similars a l'exemple anterior de la taula 4 però en les quals la distribució dels valors de l'atribut confidencial és una mica més curiosa. Si suposem que l'atacant sap que la persona que busca viu en el codi postal 97222, pot saber que els seus ingressos estan en l'interval [23.000,28.000]. De la mateixa manera, si sap que viu en el 80100, sabrà que els seus ingressos són de 10.000. En tots dos casos es produeix revelació d'atribut en major o menor mesura malgrat que les dades compleixen *k*-anonimitat per a $k = 3$. Això és el que se sol denominar un **atac d'homogeneïtat**, ja que aprofita l'homogeneïtat de l'atribut confidencial.

Taula 5. Microdades *k*-anònims i revelació d'atribut

Occupation	ZIP	Sex	Income
Teacher	80100	M	10000
Teacher	80100	M	10000
Teacher	80100	M	10000
Writer	97222	F	28000
Writer	97222	F	25000
Writer	97222	F	23000

Per tractar aquests problemes, s'han definit diversos models addicionals a la *k*-anonimitat. En els subapartats següents comentem els més rellevants.

l-diversitat

Per evitar el problema que veiem en la taula 5, *l*-diversitat (en anglès, *l-diversity*) busca garantir un nivell de diversitat mínim en l'atribut confidencial de cada classe d'equivalència.

> Un conjunt de microdades *k*-anònim compleix ***l*-diversitat** si cada classe d'equivalència o conjunt d'anonimat conté almenys *l* valors *ben representats* per l'atribut confidencial.

En aquesta definició un punt clau és el concepte de *ben representat*, ja que es pot interpretar de maneres diferents. Entre les possibles interpretacions, destaquem:

- **Valors diferents**. Requerim *l* valors diferents de l'atribut confidencial en cada classe d'equivalència.

- **Entropia**. Requerim que els valors confidencials en cada classe d'equivalència presentin una entropia major que $\log_2(l)$. Si denotem S_Q com el conjunt de valors de l'atribut confidencial en la classe d'equivalència Q, l'entropia de Q, $H(Q)$ és:

$$H(Q) = -\sum_{\in S_Q} p(Q,s)\log_2(p(Q,s))$$

on $p(Q,s)$ és la fracció de registres amb valors confidencials iguals a s en Q. Noteu que $\log_2$ és el logaritme en base 2*.

*** https://en.wikipedia.org/wiki)/Entropy_(information_theory)**

El conjunt de dades compleix l-diversitat d'entropia si $H(Q) \geq \log_2(l)$ per a tota Q en el conjunt de dades X.

Exemple (càlcul de l'entropia d'una classe d'equivalència)

En la taula 6a veiem unes microdades que compleixen 3-anonimitat i consten de dues classes d'equivalència amb tres elements cadascuna. L'atribut confidencial *Income* no s'ha protegit. Podem veure com calcular l'entropia per a les dues classes d'equivalència.

Taula 6. Exemple de l-diversitat en dades 3-anònimes

Occup.	ZIP	Sex	Income
Teacher	80100	M	10000
Teacher	80100	M	20000
Teacher	80100	M	10000
Writer	97222	F	28000
Writer	97222	F	25000
Writer	97222	F	23000

a. Exemple de microdades 3-anònimes

Valors diferents	2-diversitat
Entropia	1,8-diversitat
(c,l)-diversitat rec.	(3,2)-diversitat

b. Diferents tipus de l-diversitat

Considerem la primera classe Q_1 amb els valors de l'atribut confidencial: 10000, 20000 i 10000. En aquest cas tenim dos valors s: s_1 = 10000 i s_2 = 20000. La fracció d'aquests valors en Q_1 és de 2/3 i 1/3 respectivament. És a dir, $p(Q_1,s_1) = p(Q_1,10000) = 2/3$, i $p(Q_1,s_2) = p(Q_1,20000) = 1/3$.

L'entropia d'aquesta primera classe d'equivalència es pot calcular així:

$$H(Q_1) = -\sum_{i=1}^{2} p(Q_1,s_i)\log_2(p(Q_1,s_i))$$

$$= -\left(p(Q_1,s_1)\log_2 p(Q_1,s_1) + p(Q_1,s_2)\log_2 p(Q_1,s_2)\right)$$

$$= -\left((2/3\log_2 2/3) + (1/3\log_2 1/3)\right)$$

$$= -(-0{,}39 - 0{,}53) = 0{,}92$$

De manera anàloga, podem fer el mateix per a Q_2, amb la diferència que aquí tenim tres valors diferents per a s:

$$H(Q_2) = -\sum_{i=1}^{3} p(Q_2,s_i)\log_2(p(Q_2,s_i))$$

$$= -(-0{,}53 - 0{,}53 - 0{,}53) = 1{,}59$$

Com veiem, Q_2 té més entropia que Q_1, la qual cosa era d'esperar, ja que té major diversitat en el seu atribut confidencial.

- (c,l)-**diversitat recursiva**. En una classe d'equivalència Q, denotem com a r_i la freqüència del valor confidencial s_i en Q, i ordenem de manera decreixent la freqüència de tots els valors s_i de Q: $r_1 \geq r_2 \geq \ldots \geq r_m$.

  El conjunt de dades compleix (c,l)-diversitat si per a tota Q, $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$.

Exemple (diferents tipus de *l*-diversitat)

Seguim amb l'exemple de la taula 6a amb unes dades que compleixen 3-anonimitat i en què l'atribut confidencial és *income*. En la taula 6b podem veure quin nivell de *l*-diversitat compleixen en funció de la interpretació concreta de valors *ben representats*: valors diferents, entropia o diversitat recursiva. Les dades presenten dues classes d'equivalència de tres registres cadascuna. La primera Q_1 (*Teacher*, 80100, M) i la segona Q_2 (*Writer*, 97222, F).

Els diferents valors de *l*-diversitat es poden determinar de la manera següent:

- **Valors diferents**. Podem veure que el mínim nombre de valors diferents correspon a Q_1 amb 2 valors. Per tant, la taula compleix 2-diversitat respecte a valors diferents.

- **Entropia**. Com hem vist en l'exemple anterior, $H(Q_1) = 0{,}9183$, $H(Q_2) = 1{,}585$. Necessitem determinar un valor l tal que $H(Q_i) \geq \log_2(l)$ per a tot Q_i. En aquest cas ens fixem en la classe d'equivalència amb menor entropia, $H(Q_1)$, i veiem que per a $l = 1{,}8$ es compleix la desigualtat. El lector pot veure que $2^{H(Q_1)} = 1{,}88988$. En aquest cas direm que la taula compleix 1,8-diversitat respecte a l'entropia.

- **Diversitat recursiva**. Per cada Q_i, ordenem els valors de l'atribut confidencial per la seva freqüència (nombre de vegades que apareix en la classe):

  Q_1 :2 (valor 10000) $\geq$ 1 (valor 20000) $\rightarrow 2 \geq 1$

  Q_2 :1 (valor 28000) $\geq$ 1 (valor 25000) $\geq$ 1 (valor 23000) $\rightarrow 1 \geq 1 \geq 1$

  Ara busquem els valors c i l que compleixin la desigualtat $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$ tant per a Q_1 com a per a Q_2. En aquest cas podem veure que es compleix (3,2)-diversitat, ja que

  $$Q_1 : 2 < 3 * (1)$$

  $$Q_2 : 1 < 3 * (1 + 1)$$

t-proximitat

En alguns casos l'aplicació de *l*-diversitat pot resultar difícil o contraproduent. En certa mesura, *l*-diversitat assumeix una distribució uniforme dels valors de l'atribut confidencial. Suposem que tenim un atribut confidencial que determina tres rangs de salari o valors concrets. El 99% dels registres té el rang 1, el 0,9% el rang 2, i el 0,1% restant el rang 3. En aquest cas l'aplicació de *l*-diversitat és complicada i fins i tot podria induir a augmentar la revelació d'atribut.

La idea de la *t*-proximitat (o *t-closeness* en anglès) és fer que la distribució dels valors confidencials en una classe d'equivalència sigui la mateixa (o molt semblant) a la distribució d'aquests valors en tot el conjunt de dades.

Un conjunt de microdades compleix ***t*-proximitat** si per a totes les classes d'equivalència la distància entre la distribució de valors dels atributs confidencials a la classe i la distribució de l'atribut en tot el conjunt de dades és menor o igual que un llindar *t*.

Com a distància entre distribucions, se sol utilitzar la distància coneguda com a EMD (*earth mover's distance*), però es podria utilitzar qualsevol altra distància entre distribucions de probabilitats. No entrarem en detalls sobre com calcular la distància i assegurar, per tant, *t*-proximitat, però sí que és important remarcar el problema que pot tenir no respectar la distribució de l'atribut confidencial en microdades *k*-anònimes quan aquest no està protegit. Es pot trobar més informació en els articles detallats del quadre al marge o en la bibliografia del mòdul.

Earth mover's distance

Més informació sobre EMD i el seu ús per al càlcul de *t*-proximitat es pot trobar a **N. Li; T. Li; S. Venkatasubramanian** (2007). *T-closeness: privacy beyond k-anonymity and l-diversity*. Proceedings of the IEEE ICDE 2007, https://bit.ly/2oeric5

1.2.3. Privadesa diferencial

La privadesa diferencial és un model de privadesa que ha guanyat popularitat als últims anys. Inicialment es planteja en el context d'un escenari en línia (vegeu el subapartat 1.1.), però és extensible a la publicació de dades en general (escenari fora de línia).

Exemple (ús actual de privadesa diferencial)

Als últims anys, empreses com Google o Apple han anunciat l'ús de privadesa diferencial per protegir dades d'usuaris:

https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
https://research.google/pubs/pub42852/

El Census Bureau dels Estats Units també utilitza privadesa diferencial en l'elaboració del cens del 2020:

https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

En privadesa diferencial es considera que un conjunt de dades és privat si en afegir, eliminar o modificar un sol registre del conjunt no s'afecta el resultat de l'anàlisi que es fa sobre aquestes dades. Per a això es considera que una consulta a una base de dades hauria de donar el mateix resultat (o un resultat molt semblant) encara que s'esborrés o afegís un registre. És a dir, l'absència o presència d'un registre qualsevol no hauria de condicionar la resposta, o no hauria de ser perceptible a partir de les possibles consultes i respostes.

En la definició original considerarem dues bases de dades D_1 i D_2 que difereixen en un sol element o registre. Per exemple, una té un sol registre (o fila en el cas de microdades) més que l'altra. Considerem que es pot fer una consulta f a una base de dades D i obtenir un resultat $f(D)$. L'objectiu és trobar una funció K_f que, substituint f, garanteixi privadesa diferencial sense alterar molt el resultat. És a dir, K_f és una versió segura (que preserva la privadesa) de f. La idea més senzilla és pensar intuïtivament en K_f com a f més un soroll afegit: $K_f(D) = f(D) + soroll$.

Una funció K_f per a una consulta f proporciona ϵ-**privadesa diferencial** si per a totes les bases de dades D_1 i D_2 que difereixen en un sol element, té lloc per a tot $S \subseteq Rang(K_f)$:

$$[K_f(D_1) \in] \leq e^{\epsilon} \times [K_f(D_2) \in]$$

Moltes vegades la desigualtat de la definició anterior s'expressa també així:

$$\frac{P[K_f(D_1) \in]}{P[K_f(D_2) \in]} \leq e^{\epsilon}$$

Rang d'una funció

Recordem que el **rang** o **imatge** d'una funció f, que denotem com a $Rang(f)$, és el conjunt de tots els valors que pot prendre aquesta funció f.

Encara que la definició pot semblar una mica confusa, la seva interpretació és molt senzilla. La diferència entre la consulta $K_f(D_1)$ i $K_f(D_2)$ ha de ser el menys perceptible possible i queda limitada pel llindar ϵ. En aquest sentit ϵ s'utilitza com a indicador del nivell de privadesa proporcionat. Quant menor sigui ϵ major serà la privadesa, ja que més semblants seran les dues respostes. Podem dir que si tenim $\epsilon = 0$ la privadesa és màxima, les dues respostes $K_f(D_1)$ i $K_f(D_2)$ retornen el mateix resultat per a qualsevol consulta f i fan indistingible la presència o absència de qualsevol registre. De la mateixa manera, amb $\epsilon = 0{,}01$ ambdues respostes difereixen com a molt en el 1%.

Exemple

Suposem que tenim una base de dades amb un sol atribut V_1, *age*, com el que mostra la taula 7. En aquest cas tenim tres versions de la base de dades: D_1, D_2, i D_3. Si ens fixem podem veure que D_1 és la base de dades completa i que D_2 és igual a D_1 menys el registre amb el valor 42, i D_3 igual a D_1 sense el registre 12.

Considerem una possible consulta f, que pot ser la mitjana aritmètica $f(D_1) = \frac{1}{N}\sum_{i=1}^{N} x_{i1}$.

Taula 7. Exemple de dades per a privadesa diferencial

V_1 (*Age*)
42
12
28
51
23
68
30
46
55
62

a. Base de dades D_1

V_1 (*Age*)
12
28
51
23
68
30
46
55
62

b. Base de dades D_2

V_1 (*Age*)
42
28
51
23
68
30
46
55
62

c. Base de dades D_3

Podem fer la mateixa consulta a D_1 i D_2:

$$f(D_1) = \frac{42 + 12 + 28 + 51 + \ldots}{10} = 41{,}7$$

$$f(D_2) = \frac{12 + 28 + 51 + \ldots}{9} = 41{,}67$$

$$f(D_3) = \frac{42 + 28 + 51 + \ldots}{9} = 45{,}0$$

Comparant el resultat de les consultes, podem obtenir informació sobre el registre que s'ha eliminat en cada cas. Si comparem $f(D_1)$ i $f(D_2)$, la diferència és menor que si comparem $f(D_1)$ i $f(D_3)$, però en tots dos casos podem inferir informació sobre el valor eliminat. En el primer cas (D_2) aquest valor és proper a la mitjana i en el segon (D_3) és considerablement menor a la mitjana.

L'objectiu de la privadesa diferencial és fer que les diferències entre les consultes de l'exemple anterior siguin el menys uniforme possible o, millor dit, que a partir d'aquestes diferències no es pugui inferir informació sobre el registre eliminat. Per a això se sol aplicar algun tipus d'aleatorietat a la resposta. En aquest cas podríem afegir soroll a la resposta. D'aquesta manera, en observar les diferències entre les consultes no podrem destriar si aquestes diferències són degudes al soroll o a la mateixa consulta. Com més soroll més privadesa. Es busca que resulti més difícil obtenir informació sobre el valor eliminat fent que la incertesa sobre el resultat sigui major. D'altra banda, i com era presumible, un soroll molt elevat produirà major pèrdua d'informació.

En el subapartat 1.4.2. veurem com s'aplica soroll per aconseguir privadesa diferencial.

Exemple

Per mirar d'aclarir més la idea en què es basa la privadesa diferencial, aquí plantegem un altre exemple que s'utilitza comunament i està basat en un mètode conegut com a resposta aleatoritzada (*randomized response*).

Suposem que volem fer, a un grup de persones, una consulta sensible que admeti com a resposta *sí* o *no*. Per exemple: «Cobra més de 30.000 EUR anuals?» o «Pateix la malaltia *E*?». A més, volem preservar la privadesa de l'individu que respon. Per a això, cada individu utilitza l'estratègia següent per contestar: llança una moneda i, en funció de si surt cara o creu, fa el que es descriu a continuació.

$$\text{Llança una moneda} \begin{cases} \text{cara} & \rightarrow \text{respon amb honestedat} \\ \text{creu} & \rightarrow \text{llança la moneda} \begin{cases} \text{cara} & \rightarrow \text{respon } sí \\ \text{creu} & \rightarrow \text{respon } no \end{cases} \end{cases}$$

Com veiem, l'individu contesta la pregunta amb sinceritat amb el 50% de probabilitat i dona una resposta aleatòria amb la mateixa probabilitat. Podem veure la probabilitat que la resposta obtinguda sigui certa o no. Denotem com a A l'estratègia definida anteriorment, de manera que $A(sí)$ és el resultat d'aplicar aquesta estratègia quan la resposta original (veritable) és *sí*.

Per simplificar, considerem que la pregunta és: «Ha llegit el *Quixot*?». Les respostes obtingudes depenen del que s'ha obtingut en els llançaments de la moneda. Una persona que sí que ha llegit el *Quixot* respondrà que *sí* amb una probabilitat del 75% i que *no* amb una del 25%. De manera anàloga, les mateixes probabilitats les tenim per a la persona que no ha llegit el *Quixot*.

Podem veure aquestes probabilitats de manera més detallada desglossades per a cada cas en la taula 8. Sumant les probabilitats corresponents, tenim que $P[A(sí) = sí] = 0{,}75$, és a dir, la probabilitat que la resposta obtinguda amb l'estratègia A sigui *sí* quan la persona enquestada ha llegit realment el *Quixot* (la resposta veritable és *sí*) és del 75%. D'altra banda, tenim que $P[A(sí) = no] = 0{,}25$, la probabilitat que per als que sí han llegit el *Quixot* (resposta veritable *sí*) l'estratègia A retorni *no*. I, de manera anàloga, $P[A(no) = no] = 0{,}75$, i $P[A(no) = sí] = 0{,}25$.

Taula 8. Probabilitats de respostes per a l'estratègia A

Resposta original	1^a moneda	2^a moneda	Resposta protegida	Probabilitat
sí	cara	-	$A(sí) = sí$	0,5
sí	creu	cara	$A(sí) = sí$	0,25
sí	creu	creu	$A(sí) = no$	0,25
no	cara	-	$A(no) = no$	0,5
no	creu	cara	$A(no) = sí$	0,25
no	creu	creu	$A(no) = no$	0,25

Si ens fixem en les possibles respostes, la diferència entre cada possible cas sempre segueix una proporcionalitat màxima de 3. És a dir, entre els que han respost *sí*, n'hi ha 3/4 que sí que han llegit el *Quixot* i 1/4 que no:

$$\frac{P[A(sí) = sí]}{P[A(no) = sí]} = \frac{P[A(no) = no]}{P[A(sí) = no]} = 3$$

Per això es diu que el procés descrit compleix 1,1-privadesa diferencial o ($\ln 3$)-privadesa diferencial (recordem que $\ln e \simeq 1{,}1$).

Podem pensar en el procés descrit anteriorment com una base de dades que conté les respostes per a cada individu, i en la funció K_f com el procediment que hem denotat com a A.

D'una banda, s'aconsegueix un cert grau de privadesa o anonimitat. En obtenir una resposta *sí*, no podem saber si aquesta persona ha llegit realment el *Quixot*. D'altra banda, si el volum de població enquestada és suficientment elevat, els resultats globals seran significatius, i podem estimar el nombre de persones que han llegit realment el *Quixot*. Suposem que Y és el nombre de respostes positives que hem rebut, i assumim que p és la proporció real (veritable) de persones que han llegit el *Quixot*. Llavors tenim:

$$Y = \frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + p2$$

És a dir, de totes les persones enquestades, respondran que *sí* 3/4 de les que sí han llegit el *Quixot* (p) i 1/4 de les que no han llegit el *Quixot* ($1-p$). Llavors podem estimar que $p = 2Y - \frac{1}{2}$.

Una propietat interessant de la privadesa diferencial és el **teorema de composició seqüencial**.

La **composició seqüencial** en privadesa diferencial determina que, donada una funció aleatòria K_1 que proporciona ϵ_1-privadesa diferencial, i una altra funció K_2 que proporciona ϵ_2-privadesa diferencial, l'aplicació de les dues funcions al mateix conjunt de dades (K_1,K_2) proporciona ($\epsilon_1 + \epsilon_2$)-privadesa diferencial.

Sense entrar en detalls excessius, aquest teorema ens diu que cada vegada que es fa una consulta a la base de dades es produeix una pèrdua de privadesa acumulativa. En fer la consulta K_1 i després la K_2 passem a tenir privadesa diferencial amb $\epsilon = \epsilon_1 + \epsilon_2$. De fet, encara que tinguem solament una funció K_1, si fem la consulta t vegades, assumint que l'aleatorietat de cada resposta és independent, passem a tenir privadesa diferencial amb $\epsilon = t\epsilon_1$.

Aquest teorema té implicacions importants, ja que en certa manera limita de manera clara i precisa el nombre de consultes que podrem fer a un conjunt de dades, o conjunts de dades semblants (no disjuntes), al valor de ϵ que hàgim fixat. És el que se sol denominar **pressupost de privadesa**. Tenim un pressupost de privadesa delimitat per ϵ que anirem gastant cada vegada que es fa una consulta. Una vegada hàgim consumit aquest pressupost, no podrem fer més consultes si volem mantenir el nivell de privadesa delimitat per ϵ.

En el cas que les consultes K_1 i K_2 es facin a dos subconjunts disjunts de la base de dades, s'obté $\max(\epsilon_1,\epsilon_2)$-privadesa diferencial. Aquesta propietat es coneix com a **teorema de composició paral·lela**.

En la pràctica, la privadesa diferencial pot presentar uns requeriments de privadesa molt estrictes que fan que la seva aplicació comporti generalment molta pèrdua d'informació.

Actualment se sol utilitzar una definició de privadesa diferencial una mica més laxa, i això dona lloc al que es coneix com a (ϵ,δ)-privadesa diferencial.

Una funció K_f per a una consulta f, proporciona (ϵ,δ)-**privadesa diferencial** si per a totes les bases de dades D_1 i D_2 que difereixen per un sol element, per a tot $S \subseteq Rang(K_f)$:

$$[K_f(D_1) \in] \leq e^{\epsilon} \times [K_f(D_2) \in] + \delta$$

L'ús del paràmetre δ es pot veure com una generalització de ϵ-privadesa diferencial per a $\delta = 0$. La motivació per introduir aquest paràmetre és que la diferència entre la resposta a dues bases de dades que difereixen en un sol registre pot estar molt condicionada per registres que presentin valors atípics. Per exemple, en afegir un valor atípic, la diferència de les consultes pot ser molt gran. Això condiciona la sensibilitat de la consulta i, per tant, el paràmetre ϵ, encara que la probabilitat que aquests valors atípics apareguin sigui molt petita. Per això, moltes vegades s'eliminen aquests casos *rars* amb valors atípics. L'ús del paràmetre δ permet acceptar aquests casos *rars* sempre que ocorrin amb una probabilitat $\leq \delta$.

Tot el que hem explicat aquí és extensible a (ϵ,δ)-privadesa diferencial, inclosos els teoremes de composició. No obstant això, hem optat per centrar-nos en la definició original de ϵ-privadesa diferencial per simplificar l'exposició. Com sempre, referim al lector a la bibliografia del mòdul per aprofundir i ampliar els conceptes exposats aquí.

1.2.4. Risc de reidentificació i unicitat

Tant *k*-anonimitat com privadesa diferencial són models de privadesa que ens proporcionen en certa manera una avaluació booleana del risc de revelació associat a unes dades. Les dades compleixen o no la propietat amb tot el que això comporta. És cert que tots dos models són parametritzables i poden expressar diversos graus de privadesa o risc de revelació, però sempre determinen si les dades compleixen o no aquest nivell de privadesa.

En aquest subapartat veurem alguna mesura de risc de revelació i, per tant, de privadesa que no té aquesta naturalesa booleana. Es tracta de mesures que aporten una quantificació sobre algun aspecte relatiu al risc de revelació. Con-

cretament, veurem mesures de reidentificació basades en enllaç de registres. Comentem també el concepte d'unicitat, encara que no entrarem a calcular-lo detalladament, i referim el lector a la bibliografia del mòdul per aprofundir més en aquest tipus de mesures.

Unicitat

En aquest cas podem definir el risc associat a unes dades en funció de la probabilitat que el valor d'un atribut o combinació d'atributs puguin aparèixer amb una freqüència molt baixa en les dades. En la taula 3 del subapartat 1.2. ja vam veure que l'existència de valors únics és problemàtica.

Hi ha diverses maneres de considerar i mesurar la unicitat en un conjunt de dades, i una de les més comunes és la que es coneix com a **mesura de risc individual** (o *record-level risk uniqueness*). La idea és intentar establir la probabilitat que un registre sigui reidentificat en funció de la freqüència de la combinació dels valors dels seus atributs en les dades publicades, és a dir, la probabilitat que una combinació d'atributs sigui poc freqüent en les dades publicades i en la població de la qual s'obtenen les dades.

Mesures de reidentificació

Podem definir mesures de privadesa basant-nos en una estimació de la facilitat o dificultat que pot tenir reidentificar un atribut en les dades protegides. Donat un conjunt de dades originals X i les corresponents dades protegides $X' = \rho(X)$, podem mirar el percentatge de registres de X' que poden ser reidentificats per un atacant. El que sol dificultar aquesta estimació és que aquest percentatge dependrà del coneixement previ sobre les dades X que tingui l'atacant.

Podem denotar com a $B \subseteq X$ les dades que coneix l'atacant. Aquests dades poden incloure registres complets, encara que el més habitual és que tinguin registres amb un subconjunt d'atributs. L'objectiu de l'atacant és determinar, per a cada registre $b \in B$, el registre de X' que fa referència al mateix individu o entitat, utilitzant, per exemple, els atributs comuns, és a dir, establir un mapatge entre els registres de B i els de X'. Per establir aquest mapatge, es pot utilitzar una tècnica coneguda com a **enllaç de registres** (en anglès, *record linkage*).

L'enllaç de registres consisteix a enllaçar registres entre dos conjunts de dades que fan referència al mateix individu o entitat. En el nostre cas, per a cada registre $b_i \in B$ s'intenta enllaçar amb el registre $x'_j \in X'$ de manera que b_i i x'_j facin referència al mateix individu o entitat. Hi ha diversos mètodes i tècniques d'enllaç de registres. Un dels més utilitzats com a reidentificació és l'enllaç de registres basat en distància.

En l'**enllaç de registres basat en distància** entre dos conjunts de dades B i X, cada registre b_i de B s'enllaça amb el registre x'_j de X més proper. El registre més proper es defineix segons una funció de distància entre registres de B i X'.

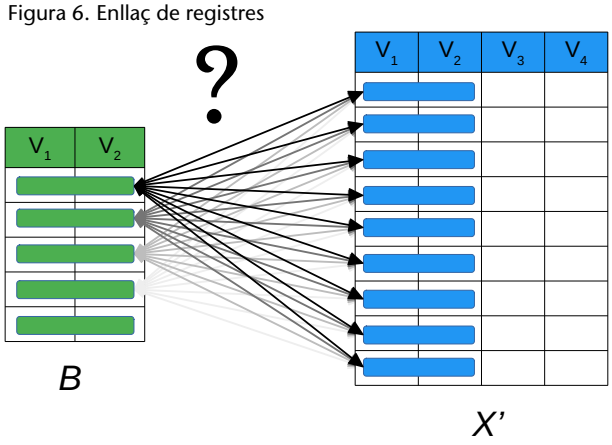
Depenent del tipus de dades i registres, s'utilitzarà una funció de distància o una altra. Per exemple, en el cas de registres amb atributs numèrics es pot utilitzar la de la distància euclidiana.

Una vegada fet l'enllaç de registres, es verifica quins registres s'han enllaçat correctament. Cal tenir en compte que en tot moment coneixem B, X, X' i els registres de B que corresponen als registres de X', cosa que ens permet determinar quins registres s'han enllaçat de manera correcta. Aquest percentatge de registres enllaçats de manera correcta es pot utilitzar com a mesura de risc de revelació.

La figura 6 il·lustra aquesta idea d'enllaç de registres basat en distància. En aquest cas, B té dos atributs V_1 i V_2, mentre que X' té aquests mateixos atributs a part de dos més. L'enllaç de registres basat en distància busca enllaçar cada registre de B amb el més proper de X'.

Si el primer registre de B és (b_{11},b_{12}), buscarà el registre de X', (x'_{i1},x'_{i2}) més proper utilitzant els mateixos atributs, és a dir, aquell registre $X'_i = (x'_{i1},x'_{i2},x'_{i3},x'_{i4})$ tal que $d((b_{11},b_{12}),(x'_{i1},x'_{i2})) \leq d((b_{11},b_{12}),(x'_{j1},x'_{j2}))$ per a tot $j \neq i$.

Figura 6. Enllaç de registres



Utilitzant aquesta estratègia, podem quantificar el risc de revelació d'un conjunt de dades protegit X' en funció de les dades conegudes per l'atacant, B. Però com podem determinar el coneixement que té l'atacant? Això no és sempre factible, i en aquest sentit es poden adoptar diverses estratègies. Hi haurà casos en què podrem utilitzar com a B dades públiques (publicades en un altre conjunt de dades) o podrem intentar intuir el coneixement que pot tenir algun tipus d'atacant.

Una possibilitat interessant és considerar el pitjor cas, això és, quan $B = X$. Aquesta situació pot semblar que no té sentit des d'un punt de vista pràctic, ja que si l'atacant coneix X no necessita fer cap reidentificació. No obstant això, sí que ens serveix com una mesura de revelació. La idea és intentar utilitzar l'enllaç de registres per buscar enllaços entre X i X'. Intuïtivament, si podem determinar l'enllaç de manera correcta, la protecció és pitjor que en el cas que no es pugui.

Un percentatge del 100% de reidentificacions correctes per enllaç de registres entre X i X' representa un risc molt alt (es pot establir l'enllaç correcte entre tots els registres de X i X'), i un de l'1% un risc molt baix. Això es pot veure com una **quota màxima del risc de reidentificació** o revelació d'identitat, ja que estem considerant el pitjor cas.

Exemple

Com a exemple, considerem les dades originals de la taula 9a, i dues versions diferents d'aquestes dades protegides (distorsionades utilitzant algun tipus de protecció), X' i X''.

Taula 9. Microdades anonimitzades de dues maneres diferents

V_1	V_2	V_3
10	33,4	1000
0	28,9	1010
30	10,3	922
20	80,0	20200
30	59,0	15320

a. Dades originals X

V_1	V_2	V_3
0	30,2	1000
0	31,2	1000
20	12,0	1000
30	82,1	20000
20	55,2	15000

b. Dades protegides X'

V_1	V_2	V_3
0	20,0	900
0	20,0	900
0	20,0	900
20	70,0	20000
20	70,0	20000

c. Dades protegides X''

Volem mesurar el risc màxim de reidentificació utilitzant enllaç de registres basat en distància. Per a això, buscarem, per a cada registre de X, quin és el registre més proper de X' i de X''. És a dir, apliquem enllaç de registres entre X i X''. La distància la calcularem com la distància euclidiana. Per exemple, la distància entre el primer registre de X i el primer registre de X' es pot obtenir així:

$$d(X_1,X'_1) = d\left((x_{11},x_{12},x_{13}),(x'_{11},x'_{12},x'_{13})\right) = d\left((10;33{,}4;1000),(0;30{,}2;1000)\right)$$

$$= \sqrt{\sum_{j=1}^{3}(x_{1j}-x'_{1j})^2} = \sqrt{(10-0)^2+(33{,}4-30{,}2)^2+(1000-1000)^2} \simeq 10{,}5$$

Podem calcular la distància de X_1 en tots els registres de X' i quedar-nos amb el registre de X' que té la menor distància. Farem el mateix per a la resta de registres de X. D'aquesta manera, obtenim la correspondència entre registres de X i X', i podem verificar si aquesta correspondència és correcta. És a dir, si l'enllaç de registres basat en distància realment ha enllaçat bé els registres.

En la taula 10 veiem com queda l'enllaç de registres entre X i X' (taula 10a), i entre X i X'' (taula 10b). En cada cas s'identifiquen els registres pel seu índex. Per exemple, en la taula 10a, $1 \to 2$ vol dir que s'ha enllaçat el primer registre de X amb el segon de X', i per tant aquest enllaç és incorrecte. Veiem que en el primer cas hi ha tres enllaços correctes de cinc, és a dir, s'aconsegueix reidentificar el 60% dels registres. En el segon cas tenim 2 de 5, és a dir el 40%.

Taula 10. Resultat de l'enllaç de registres

X		X'	Correcte?
1	→	2	NO
2	→	1	NO
3	→	3	SÍ
4	→	4	SÍ
5	→	5	SÍ

a. Enllaç entre X i X'

X		X'	Correcte?
1	→	1	SÍ
2	→	1	NO
3	→	1	NO
4	→	4	SÍ
5	→	4	NO

b. Enllaç entre X i X''

Aquests percentatges de reidentificació són una mesura que ens dona una estimació d'una possible quota màxima de reidentificació per enllaç de registres basat en distància. Es poden utilitzar com a estimació de risc de revelació. Com s'aprecia, els resultats són coherents. En el segon cas, per a X'', tenim un risc menor, ja que la protecció aplicada és major. A simple vista s'aprecia que la distorsió introduïda per generar X'' és major que la introduïda per generar X'.

En la pràctica, l'enllaç de registres basat en distància és molt costós de calcular tal com ho hem fet en aquest exemple. S'acostuma a plantejar com un problema d'optimització en què es busca maximitzar el nombre d'enllaços enllaçats correctament. Es poden utilitzar diferents funcions de distància i assignar pesos diferents a cada atribut en funció de la seva rellevància de cara a la reidentificació.

1.3. Mesures de pèrdua d'informació

Com hem comentat, en plantejar l'anonimització de dades hi ha sempre un conflicte entre el grau de privadesa i la pèrdua d'informació. En el subapartat previ hem vist algunes maneres de mesurar la privadesa, i de la mateixa manera podem utilitzar diverses mesures de pèrdua d'informació.

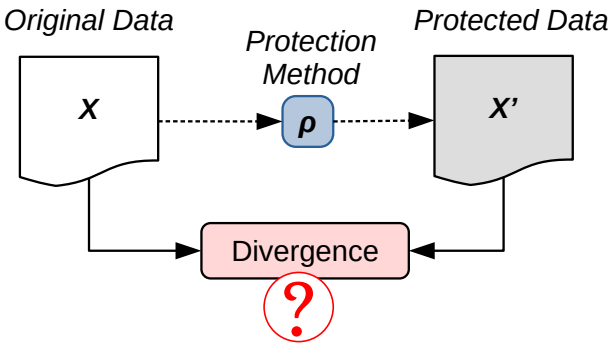
Poder estimar la informació que es perd en aplicar un mètode de protecció concret és important, ja que ens ajuda a estimar fins a quin punt les dades resultants seran útils en estudis posteriors. És comú referir-se a mesures de **pèrdua d'informació** i mesures d'**utilitat** de manera equivalent, encara que són conceptes inversament proporcionals (vegeu el subapartat 1.1.2.). Com més pèrdua d'informació menys utilitat. Malgrat ser mesures inverses, se solen utilitzar les mateixes mètriques diferenciant-ne únicament la interpretació.

De manera general, podem mesurar la pèrdua d'informació comparant les dades protegides amb les dades originals. Com més gran sigui la diferència entre les dues, més gran serà la pèrdua d'informació que s'ha produït. Aquesta diferència se sol denotar de manera genèrica com a **divergència** (figura 7).

En general, sembla assenyat assumir que si denotem la divergència entre dades originals i protegides com a $div(X,X')$, tenim que $div(X,X) = 0$ (la divergència, si els dos conjunts de dades són iguals, és 0) i $div(X,X') \geq 0$ (si els conjunts són diferents, serà major que 0). En alguns casos es requereix que la divergència

compleixi més propietats, com estar normalitzada en un interval (per exemple, [0,1]) o que sigui una mètrica, però aquí considerem el cas més general. Podem utilitzar diverses mesures diferents com a divergència i, depenent de la mesura utilitzada, estarem mesurant quin tipus d'informació es perd i en quin grau.

Figura 7. Pèrdua d'informació com a divergència entre dades protegides i originals



Per exemple, podríem definir una mesura de divergència senzilla per a microdades numèriques com la mitjana aritmètica de valor a valor entre les dades originals i les protegides. És a dir, donats $X' = \rho(X)$:

$$div_{avg}(X,X') = \frac{1}{MN}\sum_{i}^{N}\sum_{j}^{M}\frac{x_{ij}+x'_{ij}}{2}$$

Exemple

En la taula 11 es mostra un exemple de microdades amb tres atributs numèrics V_1, V_2, i V_3, on veiem les dades originals X (taula 11a) i dues versions protegides, X' (taula 11b) i X'' (taula 11c). Veiem que s'ha introduït algun tipus de distorsió diferent en X' i X''. Tant X' com X'' són versions protegides de X, però el nivell de protecció és diferent en un i altre cas.

Taula 11. Exemple de microdades anonimitzades per calcular la pèrdua d'informació

V_1	V_2	V_3
1	33,4	1000
0	28,9	1010
3	10,3	922
2	80,0	20200
3	59,0	15320

a. Dades originals X

V_1	V_2	V_3
0	30,2	1000
0	31,2	1000
2	12,0	1000
3	82,1	20000
2	55,2	15000

b. Dades protegides X'

V_1	V_2	V_3
0	40,0	900
0	20,0	1000
2	20,0	900
2	70,0	20000
2	50,0	20000

c. Dades protegides X''

Si apliquem la mesura de pèrdua d'informació que hem definit anteriorment com a div_{avg}, tenim:

$$div_{avg}(X,X') = \frac{1}{15}\sum_{i=1}^{5}\sum_{j=1}^{3}\frac{x_{ij}+x'_{ij}}{2} = \frac{1}{15}\left(\frac{1+0}{2}+\frac{33{,}4+30{,}2}{2}+\frac{1000+1000}{2}+\ldots\right)$$

$$= 2563{,}01$$

$$div_{avg}(X,X'') = \frac{1}{15}\sum_{i=1}^{5}\sum_{j=1}^{3}\frac{x_{ij}+x''_{ij}}{2} = \frac{1}{15}\left(\frac{1+0}{2}+\frac{33{,}4+40{,}0}{2}+\frac{1000+900}{2}+\ldots\right)$$

$$= 2722{,}62$$

Com veiem, la diferència entre la mitjana de les dades originals respecte a X'' és major que respecte a X'. En les taules veiem que la distorsió o alteració de les dades és major en X'' i, per tant, sembla raonable el resultat obtingut on observem major pèrdua d'informació en X''.

Es poden utilitzar diverses mesures de divergència per avaluar la pèrdua d'informació. En utilitzar div_{avg} veiem que es perd la informació capturada per la mitjana valor a valor, però tal vegada no és una mesura adequada per mesurar la pèrdua d'informació que es produeix, per exemple, respecte a la dispersió dels valors (variància o desviació estàndard). Per això és comú utilitzar diverses mesures en funció de què es vol mesurar o quina característica ens interessa analitzar respecte a la pèrdua d'informació. Aquestes mesures de pèrdua d'informació es poden combinar en una (agregant-les) o simplement donar per separat. Així mateix, es poden aplicar a atributs concrets o, com en el cas de div_{avg}, a tots.

El problema que tenim amb l'estudi de mesures de pèrdua d'informació és que no hi ha una o més mesures que predominin sobre la resta, i moltes vegades es desenvolupen *ad hoc* respecte a les dades o escenaris concrets. Com a exemple, en els subapartats següents discutirem algunes mesures de pèrdua d'informació genèriques per a dades numèriques i comentarem una mica alguna mesura específica. Per a més informació, referim al lector a la bibliografia del mòdul.

1.3.1. Exemple: pèrdua d'informació genèrica en dades numèriques

Com a exemple, veurem mesures de pèrdua d'informació genèrica que es poden utilitzar en dades numèriques. Per a això, usarem tres tipus de mesures d'error utilitzades correntment en l'anàlisi de dades. En aquest cas compararem dos conjunts de dades element a element; una altra opció que pot ser interessant és fer-ho registre a registre (considerant cada registre com un vector).

Donats dos conjunts de microdades A i B, amb N registres i M variables, podem considerar cada conjunt com una matriu i definir les funcions de divergència següents. Totes són mesures d'error en què valors propers a 0 signifiquen un error menor i, per tant, menor pèrdua d'informació.

- ***Mean squared error* (MSE)**. Mesura la mitjana de la diferència element a element al quadrat. Una característica d'aquesta mesura és que dona més pes a valors atípics (*outliers*) en elevar al quadrat la diferència.

$$MSE(A,B) = \frac{1}{NM}\sum_{ij}(a_{ij} - b_{ij})^2$$

- ***Mean absolute error (MAE)***. Mesura la mitjana de la diferència element a element en valor absolut. Com a mesura d'error absolut (al igual que l'MSE anterior), no té en compte la magnitud o escala de la dada que mesura. És a dir, un mateix error absolut de 2 respecte al número 10 és pitjor que respecte al número 1000 (tant MSE com MAE no capturen aquesta característica).

$$MAE(A,B) = \frac{1}{NM}\sum_{ij}|a_{ij} - b_{ij}|$$

- ***Mean relative error (MRE)***. Mesura la mitjana de la diferència element a element en valor relatiu respecte a l'element original. En aquest cas l'error és relatiu al valor original, per la qual cosa, a diferència de l'error absolut, sí que es té en compte la magnitud o escala de la dada.

$$MRE(A,B) = \frac{1}{NM}\sum_{ij}\frac{|a_{ij} - b_{ij}|}{|a_{ij}|}$$

Ara podem utilitzar aquestes funcions per avaluar diverses característiques entre dades originals X i dades protegides X', que considerem com a matrius. Denotem com a $cov(A)$ la matriu de covariància de A, i $corr(A)$ com la matriu de correlació de A. Llavors podem definir les mesures de pèrdua d'informació de la taula 12.

En comparar les matrius de covariància i correlació entre dades protegides i originals, mesurem com s'altera la relació entre les variables (columnes) dels conjunts de microdades, és a dir, en quin mesura la protecció aplicada a les dades originals modifica aquesta relació. Per relació entre variables entenem, per exemple, que si en les dades originals els valors alts d'una variable coincideixen amb els valors alts d'una altra, això hauria de mantenir-se en protegir les dades.

Matrius de covariància i correlació

La **matriu de covariància** captura la variància entre cada atribut (columna) i dona una matriu simètrica $M \times M$ en què cada element $cov(A)_{ij} = cov(V_i, V_j)$. Recordem que per a dos atributs V_i, V_j del conjunt de dades X, $cov(V_i, V_j) = \frac{1}{N}\sum_{r=0}^{N}(x_{ri} - \mu(V_i))(x_{rj} - \mu(V_j))$, on $\mu(V_i)$ és la mitjana de l'atribut V_i i N el nombre de registres.
De manera similar, la **matriu de correlació** captura la correlació entre atributs (columnes) i dona una matriu $M \times M$ en què cada element $corr(A)_{ij} = corr(V_i, V_j)$. Com a mesura de correlació, se sol utilitzar el coeficient de correlació de Pearson, que es calcula com la covariància dividida per la desviació estàndard de les dues variables (σ_{V_i} i σ_{V_j} respectivament): $corr(V_i, V_j) = \frac{cov(V_i, V_j)}{\sigma_{V_i}\sigma_{V_j}}$.

Taula 12. Exemple de mesures de pèrdua d'informació per a dades numèriques

$IL_{Id\_MSE}(X,X') = MSE(X,X')$
$IL_{Id\_MAE}(X,X') = MAE(X,X')$
$IL_{Id\_MRE}(X,X') = MRE(X,X')$
$IL_{Cov\_MSE}(X,X') = MSE(cov(X),cov(X'))$
$IL_{Cov\_MAE}(X,X') = MAE(cov(X),cov(X'))$
$IL_{Cov\_MRE}(X,X') = MRE(cov(X),cov(X'))$
$IL_{Corr\_MSE}(X,X') = MSE(corr(X),corr(X'))$
$IL_{Corr\_MAE}(X,X') = MAE(corr(X),corr(X'))$
$IL_{Corr\_MRE}(X,X') = MRE(corr(X),corr(X'))$

Exemple

Aquest tipus de mesures que utilitzen la matriu de covariància o correlació poden ser útils si s'aplicaran mètodes d'anàlisis en què aquesta correlació entre variables és important. Un exemple senzill pot ser el d'estudis en què es busca confirmar l'absència de causalitat en epidemiologia. En aquests casos, es vol confirmar que no hi ha correlació entre dos indicadors (variables), com viure en una certa zona i patir una certa malaltia. Perquè l'estudi sigui vàlid, hem d'assegurar que la correlació (o falta de correlació) entre les variables es manté en les dades protegides.

Podem definir mesures similars a les vistes aquí basades en uns altres indicadors estadístics que puguin capturar altres característiques de les dades.

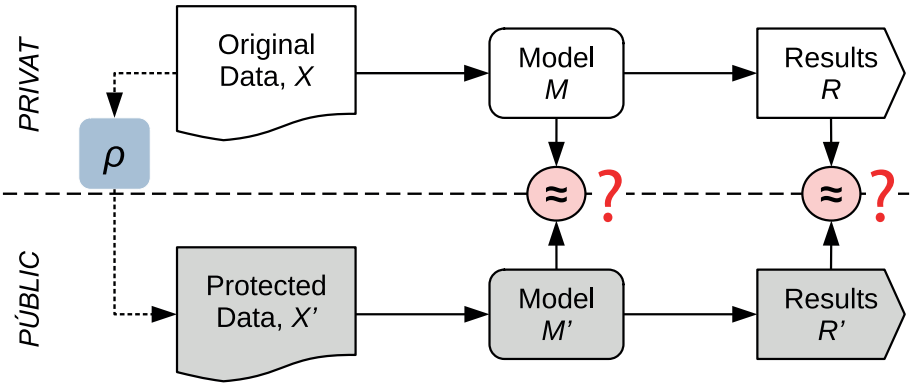
Aquest tipus de mesures solen ser una bona estimació genèrica sobre la pèrdua d'informació. Però pot ser que no siguin precises per a usos específics de les dades que vagin una mica més enllà de l'anàlisi estadística.

1.3.2. Exemple: pèrdua d'informació específica

Hi ha mesures de pèrdua d'informació específiques per a aplicacions concretes o per a l'ús concret que es donarà a aquestes dades. Això permet ajustar millor la quantificació de la pèrdua d'informació per a una anàlisi concreta. En aquests casos solem referir-nos a l'ús de dades protegides en models d'aprenentatge automàtic o mineria de dades.

A partir de les dades es genera un model, que s'utilitza per obtenir uns resultats. En el subapartat anterior quantificàvem la pèrdua d'informació comparant les dades originals amb les dades protegides. Ara la pèrdua d'informació la mesurarem comparant els models directament o comparant els resultats obtinguts a partir dels models de les dades originals i dades protegides (figura 8).

Figura 8. Mesures específiques de pèrdua d'informació



No hi ha una manera general o mètrica per mesurar la pèrdua d'informació en aquests escenaris. La majoria de les mesures es desenvolupen comparant directament els models (no sempre és fàcil o possible) o comparant els resultats per a un conjunt de tests concret, la qual cosa fa que siguin molt dependents del tipus de model que s'utilitzi.

Alguns exemples en què s'apliquen aquest tipus de mesures són els tipus de models següents:

- **Classificació**. Es compara generalment el resultat que poden donar models obtinguts a partir de les dades originals i protegides, utilitzant mesures típiques d'aquests models, com la precisió (en anglès, *precision*). Encara que és menys comú, també es pot comparar el model directament.

- **Regressió**. Es comparen de la mateixa manera els resultats amb models de regressió.

- **Clusterització**. Com a cas particular dels classificadors, hi ha mesures específiques per a clusterització. Aquestes mesures busquen comparar el model: la clusterització o particions obtingudes.

Models

Classificació. Problema que consisteix a determinar a quina categoria (o grup) pertany un element.
Regressió. Es tracta d'un problema típic en estadística i aprenentatge automàtic que busca estimar la relació entre valors dependents. Per exemple, la regressió lineal busca determinar la línia (o combinació lineal) que s'aproxima millor a un conjunt de valors.
Clusterització: Tipus de classificació en què s'agrupen les dades en clústers. Els elements d'un grup són més semblats entre ells que respecte a elements d'altres grups.

Exemple

Prenem un conjunt de dades conegut com a *Iris*. Es tracta d'un conjunt de dades molt popular i utilitzat per demostrar models de classificació. S'hi recullen dades morfològiques de l'iris de la flor de tres espècies diferents: *setosa*, *versicolour* i *virginica*. Les dades recollides són quatre atributs numèrics (en centímetres) (*Sepal length*, *Sepal width*, *Petal length*, *Petal width*) i la classe (espècie) a la qual pertany cada mostra. L'objectiu és generar un model amb aquestes dades que sigui capaç de classificar l'iris d'una flor segons aquests atributs morfològics.

Per a aquest exemple, comparem els models obtinguts de dades originals i protegides, és a dir, el model M i el model M' de la figura 8. Com a model, usarem un arbre de decisió.

Encara que pugui semblar que no té sentit aplicar protecció en aquest tipus de dades, les hem triades a causa de la seva popularitat en la comunitat d'aprenentatge automàtic i estadístic. Es tracta simplement d'un exemple en què podem veure l'efecte d'aplicar un mètode de protecció a uns atributs numèrics (el significat d'aquests atributs no és rellevant per al nostre objectiu didàctic).

Si al lector no el satisfà aquesta justificació, pot veure el conjunt de dades com un conjunt que fa referència a alguna malaltia, en el qual la classe (l'espècie de la flor) és una

Precisió en sistemes de classificació

La **precisió** en sistemes de classificació mesura la fracció d'instàncies rellevants entre totes les classificades, és a dir, la fracció que representen els positius veritables entre el total de positius veritables i falsos. Hi ha unes altres mesures que s'utilitzen en l'avaluació d'aquest tipus de sistemes, com l'exhaustivitat (*recall*), que mesura els positius veritables entre la suma de positius veritables i falsos negatius.

malaltia i els atributs numèrics són característiques físiques del pacient. O, més ben dit, pot pensar que es tracta d'informació sobre dades d'assajos per cultivar flors modificades genèticament per una empresa que vol mantenir els detalls dels assajos (mesures específiques de cada iris) privats de cara a futures patents o com a secret industrial.

Aclarit aquest punt, prenem els conjunts de dades següents:

- X: conjunt de dades *Iris* original.
- X': resultat de protegir les dades *Iris* amb soroll additiu en els atributs numèrics amb un paràmetre $p = 0{,}1\%$ de soroll.
- X'': resultat de protegir les dades *Iris* afegint soroll, aquesta vegada amb $p = 1\%$.

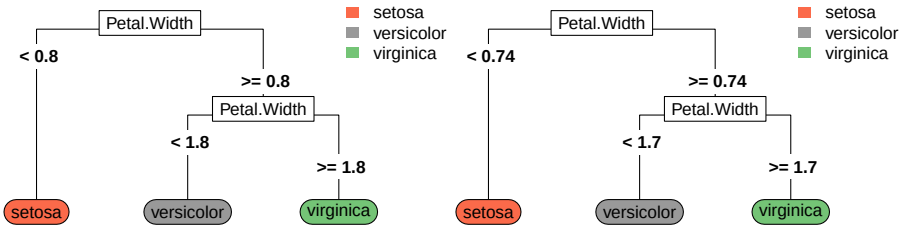
Sense entrar detalladament en el tipus de protecció (que veurem en el subapartat 1.4.2.), podem dir que el conjunt X'' ha estat protegit amb un nivell de soroll molt major. El normal és que X'' presenti major pèrdua d'informació que X' respecte a X.

A partir d'aquestes dades s'obté un model. En el nostre cas hem generat un **arbre de decisió** mitjançant partició recursiva. Concretament, hem entrenat el model per classificar noves dades segons els atributs *Petal.Width* i *Sepal.Width*. A partir de X, X', i X'' generem respectivament els models M, M', M''. En la figura 9 podem veure els tres arbres de decisió generats.

Soroll additiu

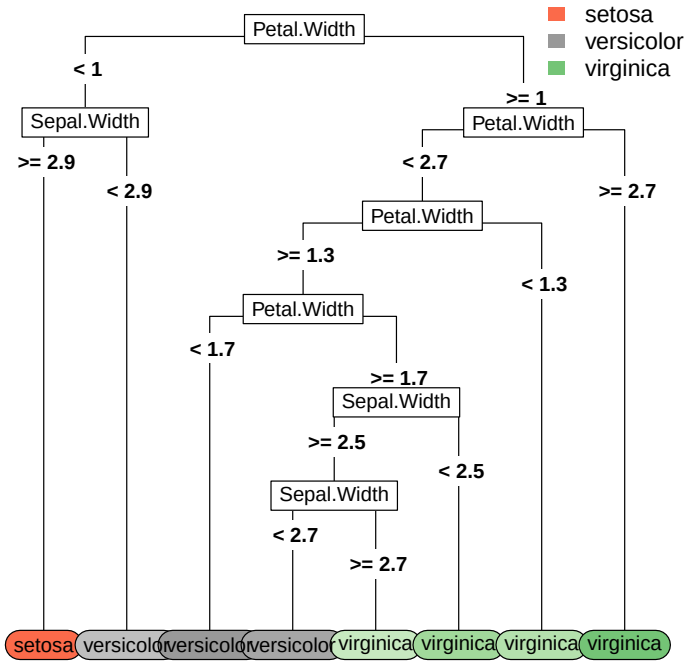
El soroll additiu, que veurem amb més detalls en el subapartat 1.4.2., consisteix a sumar soroll a les dades originals. Aquest soroll utilitza un paràmetre p. Com més petit sigui p, menys soroll es produirà i, per tant, menys distorsió (presumiblement, menor pèrdua d'informació i major risc de relevació).

Figura 9. Exemple de models generats a partir de dades protegides



9a. Model M obtingut a partir de les dades originals X

9b. Model M' a partir de les dades protegides X'



9c. Model M'' a partir de les dades protegides X''

Una possibilitat per mesurar la pèrdua d'informació és comparar aquests models. A simple vista sembla clar que hi ha major similitud entre M' i M que entre M'' i M.

La comparació dels models no és sempre senzilla, ja que necessitem alguna manera formal de poder fer la comparació (generalment no podem dir que «a simple vista» s'assemblen o no). Això dependrà molt del tipus de model i de la seva complexitat.

L'altra possibilitat és comparar el resultat d'aplicar el model a un conjunt de dades de test. En la taula 13 es mostra la classificació d'un conjunt de tests (definit pels atributs *Petal.Width* i *Sepal.Width*) utilitzant els tres arbres de decisió M, M', i M''. En general, veiem que la classificació és bastant bona i es produeixen més errors de classificació utilitzant M'' que utilitzant M'.

Taula 13. Classificació de dades de test com a mesura de pèrdua d'informació amb els models M, M' i M''

Petal.Width	*Sepal.Width*	M	M'	M''
0,5	3,0	*setosa*	*setosa*	*setosa*
0,2	3,5	*setosa*	*setosa*	*setosa*
2,7	1,3	*versicolor*	*versicolor*	*versicolor*
1,8	3,2	*virginica*	*virginica*	*virginica*
2,9	1,3	*virginica*	*virginica*	*virginica*
0,5	2,1	*setosa*	*setosa*	*versicolor*
2,1	2,6	*virginica*	*virginica*	*versicolor*
1,7	3,2	*versicolor*	*virginica*	*virginica*

És molt important remarcar que aquí entenem com a error la diferència de classificació en la qual s'incorre utilitzant M' o M'' respecte a M, no el fet que el model funcioni millor o pitjor. És a dir, no ens interessa veure si el model classifica bé o malament, ens interessa veure la diferència que hi ha entre utilitzar un model o un altre.

1.4. Mètodes de protecció

En aquest subapartat veurem diversos mètodes de protecció o *masking* de dades. Vegem quins mètodes ρ podem utilitzar per obtenir les dades protegides a partir de les dades originals, $X' = \rho(X)$. L'ús d'un o altre mètode pot dependre de diversos factors, com l'ús que es vol donar a aquestes dades, si es vol complir algun model concret de privadesa, el nivell de protecció o pèrdua d'informació que es vol obtenir, o el tipus de dades amb què treballem.

Com hem comentat a l'inici del mòdul, ens centrem en mètodes de propòsit general (o *data-driven*), anomenats de *masking*, que són els que alteren d'alguna manera les dades per protegir-les però sense utilitzar criptografia. Les dades protegides no estan xifrades, i es pot operar amb elles com es faria amb les dades originals.

Els mètodes de protecció que veurem es poden classificar segons com proporcionen aquesta protecció o alteració. En general hi ha dos tipus de mètodes de *masking*: pertorbatius i no pertorbatius. A més, considerarem també els mètodes coneguts com a generació sintètica de dades. Encara que no entrarem en els detalls sobre aquests mètodes, sí que els descriurem breument i veurem en

què consisteixen. D'aquesta manera, tenim mètodes de protecció dels tipus següents:

- **Mètodes pertorbatius**. Modifiquen les dades originals amb l'objectiu de prevenir la presència d'informació sensible. En certa manera, es poden veure com a mètodes que distorsionen les dades originals. Aquest tipus de mètodes introdueix errors en les dades protegides. És a dir, en les dades protegides hi ha informació incorrecta.

  Exemple

  Un individu que en les dades originals té l'atribut edat amb el valor 23, en les dades protegides podria tenir un altre valor com 25.

  Aquest error busca dificultar la revelació d'informació privada. Idealment, l'error introduït hauria de ser el mínim possible per no introduir massa pèrdua d'informació.

- **Mètodes no pertorbatius**. No alteren o modifiquen les dades introduint informació errònia. En canvi, el que fan és reduir la precisió o el detall de la informació, o fins i tot eliminar alguns valors. No introdueixen errors en el sentit que la informació de les dades protegides no és incorrecta encara que sí pot tenir menys detalls.

  Exemple

  En l'exemple anterior de l'individu amb edat de vint-i-tres anys, un mètode no pertorbatiu pot generalitzar el valor en les dades protegides indicant un interval com [20,39]. Un altre exemple seria substituir l'atribut que indica el lloc de residència amb el nom del barri, com *Los Royales*, *San Pedro* o *Los Pajaritos*, pel nom de la ciutat, en aquest cas *Sòria*. Aquest nou valor no aporta informació errònia sobre l'atribut, però sí pot aportar protecció en dificultar la reidentificació.

  Òbviament, aquesta generalització introdueix pèrdua d'informació deguda a la pèrdua de precisió en el valor protegit.

- **Generació de dades sintètiques**. Aquest mètode és una alternativa a la modificació de la informació original, ja que el que fa és generar les dades protegides de manera aleatòria preservant algunes característiques (per exemple, alguna propietat estadística). Típicament, s'obté un model o distribució a partir de les dades originals, i després es generen les dades protegides de manera aleatòria a partir d'aquest model. En aquest cas no hi ha una relació directa entre un registre de X i un altre de X'.

En aquest mòdul no pretenem proporcionar una llista exhaustiva de mètodes de protecció, sinó donar un esbós de la base dels principals mètodes de protecció que podem trobar en l'actualitat. La taula 14 resumeix els mètodes que discutirem en el mòdul. Com es veu, donem més rellevància als mètodes pertorbatius i pràcticament passem de puntetes pels de generació de dades sintètiques. Ho fem amb l'objectiu de simplificar aquesta exposició sense ometre

les bases i poder donar una visió global dels principals mètodes de protecció utilitzats avui dia. Cal destacar que els mètodes actuals moltes vegades combinen diverses estratègies o tipus de mètodes que són una extensió dels que veurem aquí.

Taula 14. Mètodes de protecció en aquest mòdul

Pertorbatius	Soroll additiu	Subapartat 1.4.2.
	Soroll multiplicatiu	Subapartat 1.4.2.
	Rank swapping	Subapartat 1.4.3.
	Microagregació	Subapartat 1.4.4.
No pertorbatius	Generalització	Subapartat 1.4.1.
	Supressió	Subapartat 1.4.1.
	Top/bottom coding	Subapartat 1.4.1.
Generació de dades sintètiques		Subapartat 1.4.5.

1.4.1. Generalització i supressió

A grans trets, la **generalització** (en anglès, *generalization* o *recoding*) consisteix a reemplaçar un valor per un altre més general. El cas més típic és la seva aplicació en dades categòriques mitjançant la combinació de diverses categories en una de nova. La categoria nova és menys específica que les anteriors.

La generalització es pot fer de diverses maneres depenent del tipus d'atributs que tractem i de la informació que en tinguem. Per il·lustrar-ho, podem veure un exemple de dades amb diferents atributs en la taula 15 i una possible protecció per generalització de cadascuna en la taula 16.

Taula 15. Exemple de generalització: dades originals

Occupation	ZIP	Age	Marital status
Runner	80222	42	Married
School teacher	40831	28	Divorced
Police det.	97206	51	Widow/er
Professor	40831	23	Married
Student	40831	12	Single
Police capt.	97206	68	Single
Boxer	80237	30	Divorced
Horse rider	80911	46	Married
Military lt.	97201	55	Single

Taula 16. Exemple de generalització: dades protegides

Occupation	ZIP	Age	Marital status
Sports	80***	[30, 49]	Married
Education	40***	[10, 29]	Divorced or Widow/er
Law enforcement	97***	[50, 69]	Divorced or Widow/er
Education	40***	[10, 29]	Married
Education	40***	[10, 29]	Single
Law enforcement	97***	[50, 69]	Single
Sports	80***	[30, 49]	Divorced or Widow/er
Sports	80***	[30, 49]	Married
Law enforcement	97***	[50, 69]	Single

Exemple

Vegem com s'han generalitzat les dades de la taula 16. Amb aquestes dades podem mostrar les diferents estratègies que es poden utilitzar per generalitzar diferents tipus de dades.

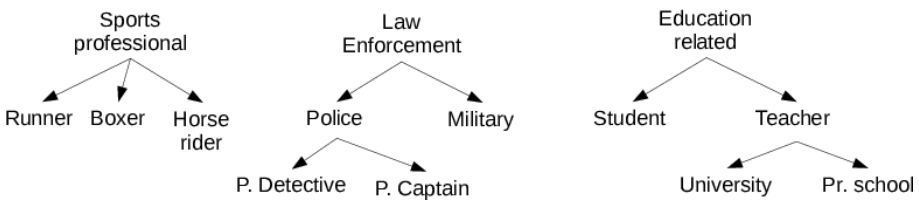
- *Occupation*. És un atribut categòric que s'ha generalitzat utilitzant una ontologia com la que es mostra en la figura 10. Aquest tipus d'ontologies poden venir donades o existir en entorns concrets, o en alguns casos es poden inferir a partir de la semàntica de l'atribut.

- *ZIP*. És un atribut categòric que per la seva pròpia naturalesa presenta un esquema jeràrquic. En la majoria de casos el codi postal d'un país es defineix per zones geogràfiques o de població, de manera que les primeres xifres són més generals.

- *Age*. Com que és un atribut numèric, una pràctica comuna és quantitzar l'atribut en classes més genèriques. En aquest cas, l'edat es generalitza en intervals de deu anys.

- *Marital status*. Es tracta d'un atribut categòric pel qual no disposem d'una ontologia o jerarquia. El que es pot fer en aquests casos és agrupar alguns atributs per crear categories. En l'exemple, s'agrupen *Divorced* i *Widow/er* perquè són els que tenen menor representació.

Figura 10. Ontologia per a l'atribut *Occupation*



En resum, i de manera general, podem aplicar diferents tipus de generalització depenent del tipus d'atribut:

- Atribut categòric: ús d'ontologies o jerarquies, o agrupació de valors.
- Atribut numèric: quantització en intervals.

La generalització és un mètode que s'aplica amb freqüència per aconseguir *k*-anonimitat. Com veiem, la taula 16 compleix la propietat de 3-anonimitat. Això ha estat deliberat clarament, i en la pràctica aconseguir una bona generalització és un problema difícil. Una generalització òptima és la que aconsegueix un nivell de privadesa desitjat (per exemple, complir un model de privadesa concret com *k*-anonimitat) reduint al màxim la pèrdua d'informació, és a dir, generalitzar al mínim possible per aconseguir un objectiu de privadesa concret. La generalització òptima se sol formalitzar com un problema d'optimització, i la seva resolució pot ser que no resulti fàcil. Per això, de vegades es busquen solucions aproximades aplicant generalitzacions genèriques.

Respecte a la uniformitat de la generalització, hi ha dues alternatives:

- ***Global recoding***. S'aplica la mateixa generalització a tots els valors. És el cas dels atributs *ZIP* o *age* en l'exemple anterior.

- ***Local recoding***. La generalització no és uniforme per al mateix atribut. Generalment s'intenta minimitzar la pèrdua d'informació generalitzant solament els valors necessaris. En l'exemple anterior, per a l'atribut *marital status* es generalitzen únicament alguns valors, i per a l'atribut *occupation* la generalització es fa en diversos nivells depenent del valor.

En alguns casos també s'aplica el que es coneix com a ***top and bottom coding***, que consisteix a aplicar la generalització únicament als valors més baixos i/o alts. Això té sentit en atributs que es puguin ordenar, com numèrics o ordinals, i sol afectar valors atípics (*outliers*).

La generalització es pot combinar amb **supressió** (en anglès, *supression*), que és una tècnica de protecció que consisteix a suprimir certs valors. Generalment, s'aplica a valors atípics o a aquells amb baixa representació i la generalització de la qual incorri en una gran pèrdua d'informació de la resta de valors de la mateixa categoria.

1.4.2. Soroll

L'ús de soroll per pertorbar dades és una pràctica molt estesa. Hi ha diferents estratègies a l'hora d'utilitzar soroll en privadesa de dades que es poden utilitzar en diversos escenaris. Generalment, s'intenta modelar el soroll de manera que el resultat final preservi algunes propietats respecte a les dades originals. Aquí veurem dos casos senzills i bastant comuns, que són el soroll additiu i el soroll multiplicatiu.

Soroll additiu

Naturalment, es tracta de sumar soroll (ε) a les dades originals X per obtenir unes dades protegides (X'):

$$X' = X + \varepsilon$$

Un cas senzill és utilitzar una distribució normal $N(\mu,\sigma^2)$ per modelar el soroll ε. Per protegir la variable V_j, utilitzem $N(\mu_j,\sigma_j^2)$, de manera que $\mu_j = 0$ i $\sigma_j^2 = pVar(V_j)$, i p és una constant que determina el nivell de pertorbació. És a dir, el soroll presenta valors al voltant de 0 (ja que la mitjana de la distribució μ_j és 0) i amb una variància proporcional als valors originals. Apliquem el mateix procediment a cada variable de manera independent.

Distribució normal

Recordem que una distribució normal o de Gauss es defineix com $N(\mu,\sigma^2)$, la mitjana és μ i la variància σ^2, amb la funció de densitat de probabilitat:

$$\frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

Aquesta tècnica es coneix com a **soroll no correlacionat**. És a dir, per a dues variables V_i i V_j, a les quals hem sumat soroll ε_i i ε_j respectivament, $Cov(\varepsilon_i,\varepsilon_j) = 0$. Es preserven les mitjanes i covariàncies en les dades protegides.

Exemple

En la taula 17 es mostra la protecció de dos atributs numèrics (un de sencer i un altre de real) utilitzant soroll no correlacionat amb una distribució normal i un paràmetre $p = 0{,}5$. Encara que es tracta d'un exemple amb molt poques dades que el fan molt poc representatiu, podem veure que les mitjanes de les variables es preserven bastant. Això és perquè el soroll que hem afegit té una mitjana de 0.

Taula 17. Exemple de soroll additiu no correlacionat

	Age	*Income*
	42	20000,00
	28	21000,00
	51	19003,20
	23	40000,55
	12	10000,00
	68	9500,33
	30	40400,00
	46	30300,00
	55	10100,00
	62	80700,20
Mean:	41,7	28100

a. Dades originals

	Age	*Income*
	56	6995,740
	41	25430,243
	35	3755,508
	13	39011,350
	5	1161,667
	57	23244,377
	24	43739,953
	40	9228,542
	49	10513,215
	75	77855,080
Mean:	39,5	24094

b. Dades protegides

Per generar el soroll, partim de la distribució normal $N(0,\sigma_j^2)$, on $\sigma_j^2 = pVar(V_j)$, per a $p = 0{,}5$. Vegem, com a exemple, com es genera el soroll per a l'atribut *age*, que denotem com a V_1. Per a això necessitem calcular la distribució normal $N(0,\sigma_j^2)$. Primer calculem la variància de V_1 com a $var(V_1) = \frac{1}{N}\sum_{i=1}^{N}(x_{i1} - \mu_1)^2$, on μ_1 és la mitjana aritmètica de l'atribut, $\mu_1 = \frac{1}{N}\sum_{i=1}^{N} x_{i1}$ (recordem que N és el nombre de registres):

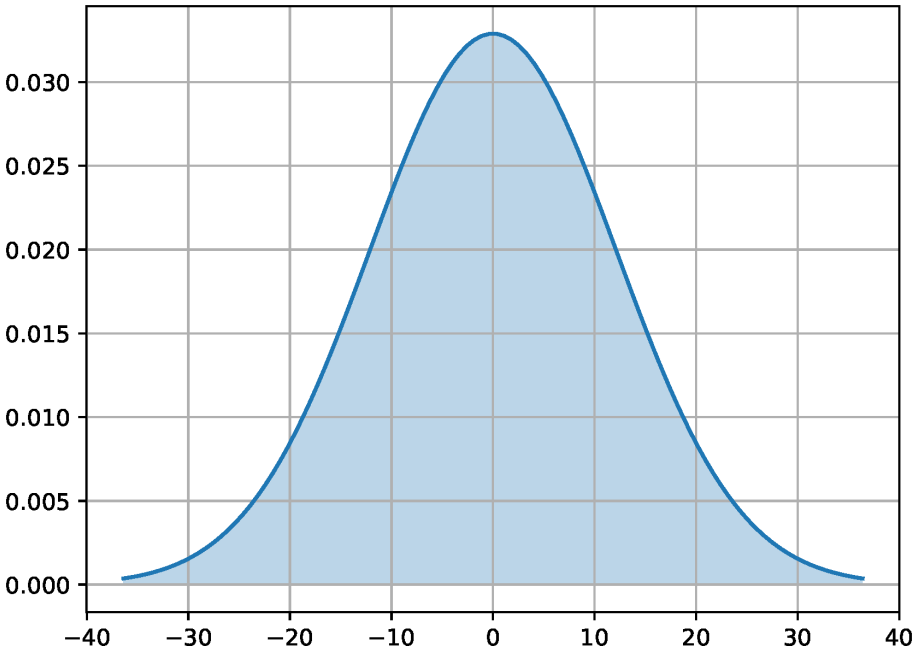
$$\mu_1 = \frac{1}{10}\sum_{i=1}^{10} x_{i1} = \frac{1}{10}(42 + 28 + 51 + \ldots) = 41{,}7$$

Ara podem calcular la variància de V_1:

$$var(V_1) = \frac{1}{10}\sum_{i=1}^{10}(x_{i1}-41{,}7)^2 = \frac{1}{10}\left((42-41{,}7)^2 + (28-41{,}7)^2 + (51-41{,}7)^2 + \ldots\right) = 294{,}21$$

Amb això definim la distribució $N(0{,}0{,}5 * 284{,}21) = N(0{,}142{,}105)$, la funció de densitat de probabilitats del qual es mostra en la figura 11.

Figura 11. Distribució normal $N(0{,}142{,}105)$



El soroll, que sumarem a les dades originals, s'obté com a mostres aleatòries d'aquesta distribució.

Alternativament, es pot aplicar un **soroll correlacionat** a V_i i V_j o X en general. Vegem aquest últim cas. El soroll ε segueix una distribució normal $N(0,p\Sigma)$, on Σ és la matriu de covariància de X, $\Sigma = cov(X)$. A diferència del soroll no correlacionat, aquí utilitzem una distribució normal multivariable, ja que treballem amb vectors de més d'una dimensió que corresponen als registres (files de la taula 18a).

La particularitat del soroll correlacionat és que es preserven els coeficients de correlació i les mitjanes.

Distribució normal multivariable

La distribució normal multivariable és una generalització de la distribució normal a vectors de més d'una dimensió. Es denota com a $N(\mu,\Sigma)$, de manera que μ és la mitjana i Σ la matriu de covariància. Referim al lector algun material introductori de la probabilitat i estadística per a més detalls sobre la seva definició i càlcul.

Exemple

En la taula 18 es mostra el mateix exemple anterior (taula 17), però aquesta vegada utilitzant soroll additiu correlacionat.

Taula 18. Exemple de soroll additiu correlacionat

	Age	*Income*		*Age*	*Income*
	42	20000,00		40	21713,166
	28	21000,00		27	20104,999
	51	19003,20		52	23539,095
	23	40000,55		21	37902,947
	12	10000,00		10	9986,307
	68	9500,33		68	8741,154
	30	40400,00		31	41709,833
	46	30300,00		45	31064,646
	55	10100,00		55	11442,116
	62	80700,20		60	78903,840
Mean:	41,7	28100	*Mean*:	40,9	28511
	a. Dades originals			b. Dades protegides	

En aquest cas la distribució utilitzada és la distribució normal multivariable $N(0,\Sigma)$, on Σ és la matriu de covariància de X (en el subapartat 1.3.1. recordem com es calcula aquesta matriu).

Una diferència important dels dos casos és la correlació entre variables. En general, el soroll correlacionat manté la correlació entre variables i el no correlacionat no.

Exemple

En la taula 19 es mostra el coeficient de correlació de Pearson (PCC) entre les dues variables per a les dades originals i els dos casos de protecció que hem vist en els dos exemples anteriors. Malgrat que es tracta d'un exemple massa senzill, es pot apreciar que el soroll correlacionat preserva millor aquesta correlació.

Taula 19. Correlació entre variables

	Originals (taula 17a, taula 17b)	No correlacionat (taula 17b)	Correlacionat (taula 18b)
PCC	0,149701	0,3623301	0,1483453

Soroll multiplicatiu

En aquest cas es calcula X' com el producte del soroll per les dades originals:

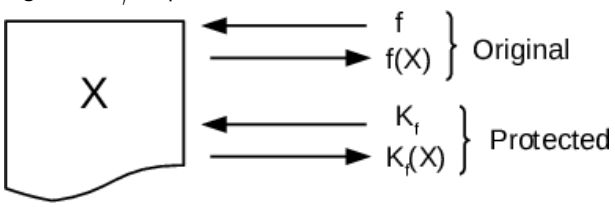
$$X' = X * \varepsilon$$

Un avantatge del soroll multiplicatiu respecte a l'additiu és que la pertorbació introduïda pel soroll (l'error en les dades protegides) és proporcional al valor a què s'aplica. Una variable amb valors petits tindrà una pertorbació menor que una amb valors majors.

Ús de soroll en privadesa diferencial

Un cas interessant d'aplicació de soroll com a mesura de protecció és el seu ús per garantir privadesa diferencial (vegeu el subapartat 1.2.3.). Una de les tècniques més comunes en privadesa diferencial consisteix precisament a afegir soroll a la resposta de la consulta.

Figura 12. K_f en privadesa diferencial



Podem simplificar el problema d'afegir soroll per aconseguir privadesa diferencial considerant una consulta f sobre una base de dades X que retornarà el resultat $f(X)$. En aquest cas considerem que les dades són numèriques i veiem cada registre com un vector de M elements (on M és el nombre d'atributs), de manera que $f : \mathcal{D} \rightarrow \mathbb{R}^M$, on $\mathcal{D}$ és el domini de tots els possibles conjunts de dades o bases de dades.

L'objectiu és definir una funció K_f com a $K_f(X) = f(X) + Noise$ que satisfaci privadesa diferencial. Partim de dues versions de X, que denotem com a X_1 i X_2 i que es diferencien en un únic registre. Recordem que la diferència entre X_1 i X_2 és que una té un registre més que l'altra, i això ho expressem denotant que la distància entre l'una i l'altra és 1 (un únic registre diferent), $d(X_1,X_2) = 1$.

Perquè K_f compleixi privadesa diferencial, necessitem definir el soroll *Noise* com una distribució independent de les dades X que depengui de com es comporta la consulta f amb conjunts de dades properes (amb distància 1). Per a això, en privadesa diferencial es defineix el concepte de sensibilitat global (*global sensitivity*).

La **sensibilitat global** d'una funció $f : \mathcal{D} \to \mathbb{R}^M$ és:

$$GS_f = \max_{X_1,X_2 \in \mathcal{D},\ d(X_1,X_2)=1} \|f(X_1) - f(X_2)\|_1$$

En aquesta definició, $\|\cdot\|_1$ és la norma L_1. Donat un vector $x \in \mathbb{R}^M$, $\|x\|_1 = \sum_{i=1}^{M} |x_i|$ (és a dir, la suma del valor absolut de tots els elements del vector). En certa manera, GS_f determina la diferència màxima que pot haver-hi entre el resultat d'aplicar f a dues versions qualssevol de la base de dades que difereixen en un sol registre. En el cas general és necessari que les dades tinguin unes quotes màxima i mínima perquè aquesta definició tingui sentit.

Exemple

Considerem l'exemple d'una base de dades amb informació sobre quins pacients pateixen esquizofrènia. En la taula 20 podem veure diverses versions de la base de dades, on un 1 indica que l'individu pateix esquizofrènia i un 0 que no la pateix.

Taula 20. Exemple de dades: pacients amb esquizofrènia

Name	*Schizophrenia*
Caracalla	1
Nerva	0
Commodus	1
Trajan	0
Hadrian	0

a. D_1

Name	*Schizophrenia*
Tiberius	1
Caracalla	1
Nerva	0
Commodus	1
Trajan	0
Hadrian	0

b. D_2

Name	*Schizophrenia*
Nerva	0
Commodus	1
Trajan	0
Hadrian	0

c. D_3

Name	*Schizophrenia*
Vespasian	0
Caracalla	1
Nerva	0
Commodus	1
Trajan	0
Hadrian	0

d. D_4

Si partim de la primera base de dades D_1, veiem que $d(D_1,D_2) = 1$, és dir, D_1 i D_2 difereixen en un sol registre. Tenim el mateix amb la resta de bases de dades respecte a D_1,

$d(D_1,D_3) = d(D_1,D_4) = 1$. No obstant això, podem veure que $d(D_3,D_4) = 2$, i $d(D_2,D_4) = 2$ (ja que difereixen en dos registres o individus).

Per il·lustrar la idea de sensibilitat, considerem la funció f_{count}, que fa una consulta que consisteix a sumar l'atribut *schizophrenia* (retorna el nombre de pacients que pateixen esquizofrènia). En aquest cas, $f_{count}(D_1) = 2$, $f_{count}(D_2) = 3$, $f_{count}(D_3) = 1$, $f_{count}(D_4) = 2$.

En aquest exemple senzill podem veure que la sensibilitat global de f_{count} és 1. Si apliquem la funció a qualsevol parell de versions de la base de dades D_i, D_j que difereixen en un sol registre, això és $d(D_i,D_j) = 1$, la diferència màxima que podem tenir entre $f_{count}(D_i)$ i $f_{count}(D_j)$ és 1. Per ser més concrets, aquesta diferència solament pot ser 0 o 1, és a dir, $GS_{f_{count}} = 1$, ja que $\|0\|_1 = 0$, i $\|1\|_1 = 1$.

Una vegada tenim la sensibilitat de la funció f, podem definir el soroll *Noise* necessari per afegir a aquesta funció. És a dir, definir $K_f(X) = f(X) + Noise$. La manera més típica de definir aquest soroll és la coneguda com a **mecanisme de Laplace**.

El **mecanisme de Laplace** per a la funció f és:

$$M_L(X,f,\epsilon) = f(X) + (N_1, \ldots, N_M)$$

on N_i són variables aleatòries independents que s'obtenen a partir d'una distribució de Laplace, concretament $N_i \sim Lap(0,GS_f/\epsilon)$, i M és el nombre de variables o atributs.

La particularitat d'aquesta definició és que $M_L(X,f,\epsilon)$ **satisfà ϵ-privadesa diferencial**, per la qual cosa podem utilitzar-la com la funció $K_f(X)$ que buscàvem.

$Lap(0,GS_f/\epsilon)$ és una distribució de Laplace. Recordem que una distribució de Laplace presenta la funció de densitat de probabilitats següent:

$$Lap(\mu,b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

En aquest cas tenim que $\mu = 0$ i $b = GS_f/\epsilon$.

Exemple

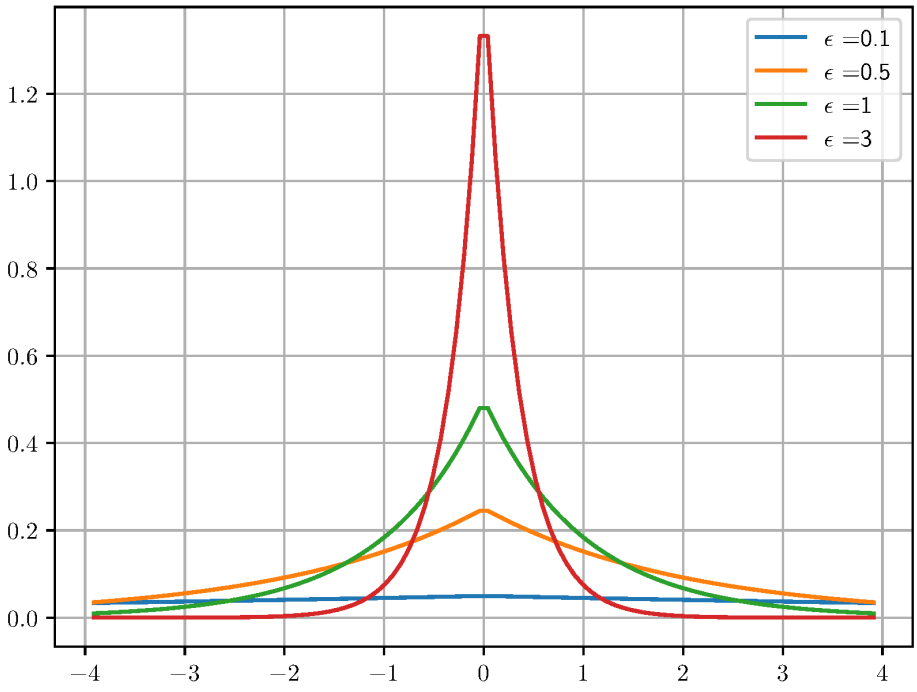
Seguint l'exemple anterior, on teníem la funció f_{count} amb una sensibilitat global $GS_{f_{count}} = 1$, apliquem el mecanisme de Laplace. La funció

$$M_L(X,f_{count},\epsilon) = f_{count} + N_1$$

on N_1 (en aquest cas solament tenim una variable) s'obté de manera aleatòria a partir de la distribució $Lap(0,1/\epsilon)$.

En la figura 13 podem veure aquesta distribució per a diversos valors de ϵ. Com es pot veure, valors de ϵ més grans generaran una distribució més precisa, la qual cosa fa que s'introdueixi menys soroll (la probabilitat d'obtenir un valor proper a 0 a partir de la distribució és major), i s'aconsegueix, per tant, un nivell de privadesa menor. Els valors de ϵ petits produeixen més soroll respecte a f_{count}, de manera que introdueixen major incertesa (i per tant privadesa) i, presumiblement, major pèrdua d'informació.

Figura 13. Funció de densitat de probabilitat de $Lap(0,1/\epsilon)$ per a ϵ = 0,1,0,5,1,3



Privadesa diferencial per a la publicació de dades fora de línia

Hi ha diverses estratègies per aplicar privadesa diferencial en escenaris fora de línia de publicació de dades. En la seva majoria es basen en consultes a l'histograma, en les quals es particiona el domini de les dades i es compta el nombre de registres en cada partició. Per obtenir un resum més detallat i referències, el lector pot consultar, entre altres fonts, el capítol 8 de l'obra següent: Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database anonymization privacy models, data utility, and microaggregation-based inter-model connections*. Califòrnia: Morgan & Claypool.

L'aplicació del mecanisme de Laplace és una solució molt interessant des del punt de vista teòric. En la pràctica l'ús de privadesa diferencial en aquests termes presenta algunes limitacions. Una és que la seva aplicació pot comportar una pèrdua d'informació molt elevada. Això, sumat a la limitació imposada pel teorema de composició seqüencial (vegeu el subapartat 1.2.3.), pot limitar molt el nombre de possibles consultes (funcions f) que es poden fer.

Hi ha millors aproximacions al mecanisme de Laplace, variants per a dades categòriques, i versions més laxes de privadesa diferencial. Referim el lector a la bibliografia del mòdul per obtenir-ne més informació i referències. També es pot aplicar privadesa diferencial en la publicació de dades en escenaris fora de línia.

1.4.3. *rank swapping*

El *rank swapping* és una variant d'un mètode de protecció més general conegut com a *data swapping*. La idea de *data swapping* és intercanviar valors d'un atribut entre ells. La justificació del mètode és que molta de la informació es-

tadística d'un atribut s'obté a partir de la freqüència dels seus valors, per la qual cosa l'intercanvi de valors entre registres no n'altera la freqüència final.

Algorithm 1: Rank swapping

Data: All values of variable V, and p

Result: Swapped values of variable V

1 Sort values of V as $(a_1,a_2,\ldots,a_n)$ so $a_i \leq a_j$ for all $1 \leq i < j \leq n$;
2 Mark all $a_i \in V$ as *unswapped*;
3 **for** $i = 1$ **to** n **do**
4 **if** a_i *is* unswapped **then**
5 Select l randomly from the interval $[i+1,M]$, where $M = \min(n, i + p * n/100)$;
6 Swap a_i with a_l;
7 Unmark a_i,a_l as *unswapped*;
8 Undo the sorting step from line 1 ;

Aquí veiem una variant que s'utilitza en variables ordinals o numèriques, en què l'intercanvi de valors es fa entre valors propers. En l'algorisme 1 es mostra l'aplicació de *rank swapping* a una variable o atribut. Quan es vol aplicar, a més, variables, s'aplica el mateix algorisme de manera individual a cada variable.

L'algorisme segueix els passos següents:

1) S'ordenen els valors de la variable.

2) Els valors s'intercanvien uns amb altres de manera aleatòria dins d'un interval delimitat per un paràmetre p.

3) Es desfà l'ordenació inicial.

Els valors intercanviats estan limitats a un interval del p% del total. Com més gran sigui p, menor serà el risc de revelació associat i major la pèrdua d'informació.

Exemple

En la taula 21 es mostra un exemple senzill de l'aplicació de *rank swapping* a un atribut numèric. Hem inclòs l'índex de cada valor entre claudàtors en la primera columna de cada taula per il·lustrar millor tots els passos.

En la primera taula (taula 21a) veiem les dades originals amb la seva ordenació original. El primer pas, en la taula 21b, és ordenar els valors. Una vegada ordenats, es fa el *swapping* tal com es descriu en l'algorisme 1 amb el paràmetre $p = 30$ resultant en la taula 21c. Finalment, en la taula 21d es mostra com queden les dades després de desfer l'ordenació inicial.

Taula 21. Exemple de *rank swapping*

	Age
[1]	42
[2]	28
[3]	51
[4]	23
[5]	12
[6]	68
[7]	30
[8]	46
[9]	55
[10]	61

a. Dades originals

	Age
[5]	12
[4]	23
[2]	28
[7]	30
[1]	42
[8]	46
[3]	51
[9]	55
[10]	61
[6]	68

b. Ordenació

	Age
[5]	23
[4]	12
[2]	42
[7]	51
[1]	28
[8]	55
[3]	30
[9]	46
[10]	68
[6]	61

c. Intercanvi (*swap*) amb $p = 30$

	Age
[1]	28
[2]	42
[3]	30
[4]	12
[5]	23
[6]	61
[7]	51
[8]	55
[9]	46
[10]	68

d. Desfer l'ordenació

En aplicar *rank swapping* s'intercanvien valors entre registres però no es modifiquen. D'aquesta manera, es manté la freqüència i distribució marginal de la variable. Per contra, en aplicar el mètode a cada variable de manera independent, si l'apliquem a més d'una variable, la correlació entre variables no es manté necessàriament en les dades protegides.

1.4.4. Microagregació

La **microagregació** és un mètode que proporciona privadesa uniformitzant conjunts de dades semblants. La idea és crear petits clústers o microclústers i substituir totes les dades de cada microclúster pel seu representant o centroide. Tots els registres del mateix microclúster passen a tenir el mateix valor i poden arribar a ser indistingibles entre ells. El grau de privadesa el determina el nombre de registres mínim que contenen els microclústers.

Exemple

De manera simplificada, en la taula 22 podem veure un exemple molt senzill de microagregació d'una variable numèrica sencera.

A partir de les dades originals de la taula 22a es creen microclústers amb un mínim de tres elements seguint aquests passos:

1) **Creació dels microclústers**. S'agrupen valors propers o semblats entre ells, i per a això, seguint la línia de molts algorismes de clusterització, se sol recórrer a una funció de distància o similitud (en aquest cas, la distància euclidiana). La taula 22b mostra els clústers de tres elements com a mínim utilitzant un color diferent per a cadascun.
2) **Substitució de valors**. Cada element de cada microclúster és substituït pel representant del microclúster (el centroide o centre del clúster). En aquest cas, aquest representant o centre s'ha calculat com la mitjana aritmètica dels valors del clúster, arrodonida per forçar que sigui un sencer. El resultat es mostra en la taula 22c.

Taula 22. Exemple de microagregació

Age	Age	Age
42	42	42
28	28	21
50	50	42
23	23	21
12	12	21
68	68	61
30	30	42
46	46	42
55	55	61
61	61	61
a. Dades originals	b. Partició	c. Substitució

El nombre mínim d'elements de cada microclúster, que generalment es denota com a k, determina el nivell de privadesa. Com més k més privadesa i més pèrdua d'informació.

A l'hora de formar els microclústers és important fer-ho de manera que la pèrdua d'informació posterior en substituir els valors pel representant o centre del microclúster sigui la menor possible. Aquest error és determinat per la distància de cada element al representant del clúster, que se sol mesurar com la suma dels quadrats entre grups (*within-groups sum of squares*) **SSE**. Podem veure SSE com la suma de les distàncies al quadrat de cada registre al seu representant.

De manera més formal, a partir d'unes dades originals X, apliquem microagregació per obtenir les dades protegides X'. Suposem que aquestes dades protegides contenen p microclústers en total. Representem cada clúster i (on $1 \leq i \leq p$) amb una funció característica χ_i, de manera que per a cada registre x $\chi_i(x) = 1$ si x pertany al clúster, i i $\chi_i(x) = 0$ en cas contrari. Denotem com a c_i el representant o centroide del clúster i, i el paràmetre k denota el nombre mínim d'elements. L'SSE es defineix de la manera següent utilitzant una funció de distància entre registres d:

$$SSE = \sum_{i=1}^{p} \sum_{x \in X} \chi_i(x)(d(x,c_i))^2$$

SSE és una mesura de pèrdua d'informació

En microagregació l'SSE és un cas de mesura de pèrdua d'informació específica en un mètode de protecció concret (vegeu el subapartat 1.3.2.).

La microagregació se sol formalitzar com un problema d'optimització que consisteix a minimitzar l'SSE subjecte a la restricció que cada microclúster ha de tenir un nombre d'elements entre k i $2k$.

Hi ha dos tipus de microagregació molt diferenciats:

- **Microagregació univariable**. És el cas en què la microagregació s'aplica a una sola variable. Aquest cas és interessant perquè hi ha algorismes que poden donar una solució òptima en temps polinomial.

- **Microagregació multivariable**. És el cas d'aplicar la microagregació a més d'una variable alhora. Cada element a microagregar es pren com un vector que correspon a tot el registre (o als atributs del registre que es microagregaran). A diferència del cas anterior, el problema d'optimització passa a ser *NP-hard* (o NP-complex), per la qual cosa s'utilitzen mètodes heurístics. Un aspecte molt interessant és que si s'aplica microagregació multivariable utilitzant tots els atributs s'aconsegueix *k*-anonimitat, amb *k* igual al paràmetre *k* de microagregació.

Mètodes heurístics per a la microagregació

Els mètodes heurístics utilitzats per aplicar microagregació multivariable consten dels passos següents:

- **Partició**. Es divideixen les dades en microclústers de grandària superior o igual a *k*. Aquesta divisió es basa en una **funció de distància** entre els vectors que representen els registres. Depenent del tipus de dades, s'utilitzarà una o altra funció de distància. En el cas de vectors numèrics és habitual utilitzar la distància euclidiana.

- **Agregació**. Es calcula el representant o centre de cada microclúster. El càlcul es fa utilitzant una **funció d'agregació**, amb la qual a partir del conjunt de registres que formen el microclúster s'obté el representant. Igual que en el cas anterior, aquesta funció dependrà del tipus de dades. En el cas numèric se sol utilitzar la mitjana aritmètica.

- **Substitució**. Cada registre se substitueix pel representant del microclúster al qual pertany en les dades protegides.

Hi ha diverses estratègies i algorismes per a la microagregació multivariable, entre els quals mostrem el conegut com a MDAV (*maximum distance to average vector*) en l'algorisme 2. En aquest cas l'algorisme està pensat per aplicar microagregació a totes les variables, amb la qual cosa es tracta tot el registre com un vector. S'assumeix també que tractem amb dades numèriques, i la partició es fa utilitzant la distància euclidiana i l'agregació utilitzant la mitjana aritmètica.

Algorithm 2: Algorisme MDAV.

Data: X: original data, k: minimum number of record for each cluster.

Result: Partition of X in clusters C_i.

while $|X| \geq 3k$ **do**
  $\bar{x}$ = average of all records in X;
  x_r = most distant record to $\bar{x}$ from X;
  C_r = cluster around x_r: x_r and the $k-1$ closest records to x_r;
  Remove records in C_r from X;
  x_s = most distant record to x_r from X;
  C_s = cluster around x_s: x_s and the $k-1$ closest records to x_s;
  Remove records in C_s from X;
if $|X| \geq 2k$ **then**
  $\bar{x}$ = average of all records in X;
  x_r = most distant record to $\bar{x}$ from X;
  C_r = cluster around x_r: x_r and the $k-1$ closest records to x_r;
  Remove records in C_r from X;
Form a cluster with the remaining records in X;

El funcionament d'MDAV és relativament senzill. La idea principal és partir de tot el conjunt de dades (en aquest cas vectors) i calcular el vector mitjà de totes (amb la distància euclidiana) $\bar{x}$. A partir d'aquest registre seguim el procediment següent:

1) Buscar el registre x_r que sigui més distant a $\bar{x}$ i crear un microclúster de k registres en total. El clúster es crea amb x_r i els $k-1$ registres més propers a x_r.

2) Buscar el registre x_s que sigui més distant a x_r i crear un microclúster de k registres de la mateixa manera que abans.

Aquest procés es va repetint fins que no queden suficients registres per crear un microclúster de k elements (l'últim microclúster tindrà entre k i $2k$ registres).

1.4.5. Generació de dades sintètiques

Una tècnica que difereix bastant de les vistes fins ara és la generació de dades sintètiques. Com el seu nom indica, la idea és generar un conjunt de dades artificials que reemplacen les dades originals. En certa manera es poden veure com a dades simulades. Solem distingir entre dues alternatives:

- **Dades totalment sintètiques**. Totes les dades protegides són generades de manera sintètica.

- **Dades parcialment sintètiques**. Es generen de manera sintètica algunes variables d'alguns registres (les que comporten un major risc de revelació).

Les dades es generen de manera aleatòria mantenint certs indicadors estadístics, models o relacions de les dades originals. La precisió amb què s'aconsegueixi construir el model de les dades originals determinarà la pèrdua d'informació en les dades generades. Hi ha una dificultat clara en la pràctica per aconseguir models precisos per a qualsevol tipus de dades. Això fa que moltes vegades el model utilitzat depengui de l'ús que es farà de les dades. Si es farà un estudi utilitzant algun tipus de mesura estadística concreta, s'utilitzarà un model que preservi aquesta mesura. De la mateixa manera, si s'aplicarà algun mètode d'aprenentatge automàtic es buscarà preservar les característiques particulars d'aquest mètode.

Amb relació al risc de revelació, les dades totalment sintètiques es consideren molt segures. *A priori* no és possible la reidentificació de registres, ja que totes les dades són generades i, per tant, no hi ha presència de dades originals. No hi ha atacs a aquest tipus de dades coneguts en la literatura. No obstant això, cal tenir present que això podria no ser sempre així. Un bon model podria generar els mateixos valors que els originals per a alguns atributs d'algun registre, cosa que faria possible algun tipus de revelació.

A diferència de les dades totalment sintètiques, les parcialment sintètiques sí que presenten risc de reidentificació i revelació de manera similar a la resta de mètodes de protecció pertorbatius o no pertorbatius.

Resum

En aquest mòdul hem presentat el problema de la privadesa en la publicació de dades senzilles. Concretament, ens hem centrat en la publicació de microdades. Com hem vist, un problema important a tractar en aquests casos és poder trobar un bon equilibri entre privadesa o risc de revelació i pèrdua d'informació o utilitat.

A l'hora de poder establir mesures o models de privadesa, hem vist la *k*-anonimitat i privadesa de diferencial i hem comentat algunes mesures de reidentificació com l'ús d'enllaç de registres. També hem vist que mesurar la pèrdua d'informació es pot fer de diverses maneres, algunes genèriques i altres més específiques de l'ús concret que es dona a les dades.

Finalment, hem revisat alguns dels principals mètodes pertorbatius i no pertorbatius de protecció de microdades.

Exercicis d'autoavaluació

1. Considerem un conjunt de microdades que ha estat protegit per generalització en la taula 23. S'han protegit els quasiidentificadors *Occupation*, *ZIP*, *Age*, i *Marital status*. L'atribut *Income* es considera confidencial i no ha estat protegit. Sabem també que les persones que apareixen en les dades són dels Estats Units.

Taula 23. Exercici: taula protegida mitjançant generalització

Idx	*Occupation*	*ZIP*	*Age*	*Marital status*	*Income*
[1]	Sports	80***	[30, 49]	Married	25050
[2]	Education	40***	[10, 29]	Divorced or widow/er	18098
[3]	Law enforcement	97***	[50, 69]	Divorced or widow/er	28760
[4]	Education	40***	[10, 29]	Married	26062
[5]	Education	97***	[10, 29]	Single	790
[6]	Law enforcement	97***	[50, 69]	Single	30033
[7]	Sports	80***	[30, 49]	Divorced or widow/er	55093
[8]	Sports	80***	[30, 49]	Married	17330
[9]	Law enforcement	97***	[50, 69]	Single	15700

Casualment, a internet trobem la publicació dels premiats a millor professor o professora del curs actual de la Societat d'Amics del Professor, que podem veure en la taula 24.

Taula 24. Exercici: publicació llista de professors premiats

Name	*Institution*	*City*
John Keating	Welton School	Portland (Oregon)
Erin Gruwell	Woodrow Wilson High School	Long Beach (Califòrnia)
Mark Thackeray	Seneca High School	Louisville (Kentucky)
Jaime Escalante	Garfield High School	East Los Angeles (Califòrnia)
Edna Krabbapel	Springfield Elementary School	Springfield (Illinois)
Rosemary Cross	East High School	Anchorage (Alaska)

a) Compleix la taula 23 la propietat de k-anonimitat? En cas afirmatiu determina el valor de k, i en cas negatiu justifica per què.

b) Quina informació podem obtenir sobre *John Keating* de la taula 23? Es produeix algun tipus de revelació d'informació?

c) I sobre *Mark Thackeray*?

2. En la taula 25a X tenim un conjunt de microdades, i en les taules 25b i 25c dues versions diferents protegides X' i X''. Sabem que el mètode de protecció aplicat ha estat el soroll additiu.

a) Volem mesurar la pèrdua d'informació que es produeix en X' i X''. Per a això utilitzarem algunes de les mesures de pèrdua d'informació de caràcter general que hem vist en el mòdul. Calcula les mesures de pèrdua d'informació denotades així en la taula 12 del subapartat 1.3.1.:

- IL_{Id_MSE}, IL_{Id_MAE}, IL_{Id_MRE}.
- IL_{Corr_MSE}, IL_{Corr_MAE}, IL_{Corr_MRE}.

Volem veure aquestes mesures de pèrdua d'informació per a X' i X''.

Taula 25. Exercici: taules numèriques amb diverses proteccions

V_1	V_2
10	90
9	80
8	70
7	60
6	50
5	40
4	30
3	20
2	10
1	9

a. Dades originals X

V_1	V_2
10,70	83,42
10,45	81,07
8,50	77,46
7,58	58,56
5,49	54,46
5,62	36,86
4,99	32,88
3,47	18,15
1,06	16,55
1,92	5,23

b. Dades protegides X'

V_1	V_2
16,79	146,61
14,66	126,57
13,37	114,44
13,68	118,22
10,67	88,79
11,55	95,05
9,82	78,29
8,24	61,17
7,44	55,94
6,49	57,07

c. Dades protegides X''

b) Quines conclusions podem treure comparant les mesures de pèrdua d'informació de X' i X'' sobre com s'han protegit les dades en cada cas? És important prestar especial atenció a la comparació de les mesures entre cada conjunt de dades protegides i entre els tipus de mesures, és a dir, entre les mesures que comparen l'error en els valors de les microdades directament i les que comparen la matriu de correlació.

3. En el subapartat 1.4.4. hem vist l'algorisme MDAV per a la microagregació multivariable. Concretament, hem vist com s'aplica aquest algorisme a dades numèriques. És a dir, cada registre s'interpreta com un vector numèric. Com es pot estendre aquest algorisme a dades categòriques? És possible? Si és que sí, què cal modificar de l'algorisme? I si és que no, per què?

4. Com a exercici general, recomanem provar mètodes de protecció i d'avaluació de microdades amb alguna eina de programari dissenyada per a això. Un exemple és el paquet *sdcMicro**. Es tracta d'una llibreria de *R* senzilla i que incorpora molts dels mètodes de protecció i avaluació que hem vist en el mòdul. També inclou conjunts de dades de diferents tipus sobre les quals fer proves, incorpora una interfície gràfica, en forma d'aplicació *shiny* (aplicació web feta en *R*), que permet fer proves sense haver de programar en *R*.

* https://bit.ly/32vPGI0

Recomanem revisar la documentació, on hi ha nombrosos exemples. Per a més informació, podeu consultar el web següent: http://sdctools.github.io/sdcmicro/index.html.

Solucionari

1. a) La taula 23 no compleix *k*-anonimitat per a $k > 1$. Per exemple, la combinació de valors (*Education*, *40****, [10,29], *Married*) és única. Sí que és cert que podríem dir que compleix 1-anonimitat o *k*-anonimitat per a $k = 1$. Ara bé, té sentit parlar de 1-anonimitat?

b) Si mirem la informació pública sobre *John Keating* en la taula 24, veiem que viu a Portland. Podem buscar el codi postal de Portland, que resulta que comença per 970, 971 o 972.

Tornant a la taula 23, podem veure que, com que es tracta d'un professor (professional de l'educació) i viu en un codi postal que comença per 97, es correspon al registre 5. Podem saber quant cobra, que és solter i que té una edat compresa entre 10 i 29 anys.

Hem pogut fer una reidentificació del registre que correspon a *John Keating*; és a dir, s'ha produït **revelació d'identitat**.

c) Seguint el mateix raonament anterior, podem identificar que *Mark Thackeray* correspon al registre 2 o 4, ja que el codi postal de Louisville (Kentucky) comença per 400 o 401.

En aquest cas no podem reidentificar el registre però sí obtenir informació nova sobre *Mark Thackeray*. Sabem que té entre 10 i 29 anys, que és casat o divorciat i que cobra 18.098 o 26.062. En aquest cas es produeix **revelació d'atribut**.

2. a) El càlcul de les mesures és immediat seguint les indicacions i fórmules del subapartat 1.3.1. Recomanem fer algun càlcul de manera manual per familiaritzar-nos amb les mesures. De totes maneres, aquests càlculs es poden fer de manera ràpida amb algun llenguatge de programació com *R* o el paquet Pandas* de Python.

*** https://pandas.pydata.org/**

Mostrem els resultats en la taula 26. Pot haver-hi algun petit error a causa de l'arrodoniment d'algunes xifres.

Taula 26. Exercici: resultats de pèrdua d'informació

IL_{Id_MSE}	$MSE(X,X')$	10,38
IL_{Id_MAE}	$MAE(X,X')$	2,34
IL_{Id_MRE}	$MRE(X,X')$	0,20
IL_{Corr_MSE}	$MSE(corr(X),corr(X'))$	0,00062
IL_{Corr_MAE}	$MAE(corr(X),corr(X'))$	0,01767
IL_{Corr_MRE}	$MRE(corr(X),corr(X'))$	0,01773

a. Resultats per a X'

IL_{Id_MSE}	$MSE(X,X'')$	1202,92
IL_{Id_MAE}	$MAE(X,X'')$	27,04
IL_{Id_MRE}	$MRE(X,X'')$	1,75
IL_{Corr_MSE}	$MSE(corr(X),corr(X''))$	$1{,}802e-6$
IL_{Corr_MAE}	$MAE(corr(X),corr(X''))$	0,0009493
IL_{Corr_MRE}	$MRE(corr(X),corr(X''))$	0,0009526

b. Resultats per a X''

b) Observant les dades protegides i les mesures de pèrdua d'informació, veiem un cas curiós. A simple vista, si observem els conjunts de microdades X' i X'' en les taules 25b i 25c, podria semblar que la taula X'' té major distorsió i, per tant, major pèrdua d'informació.

Podem confirmar les nostres sospites si mirem les mesures IL_{Id_MSE}, IL_{Id_MAE} i IL_{Id_MRE}. En els tres casos, l'error és major per a X'' que per a X'. Això vol dir que la diferència entre els valors de les taules és major en X'' que en X'.

No obstant això, veiem que la sospita inicial és qüestionada si mirem les mesures relatives a la matriu de correlació: IL_{Corr_MSE}, IL_{Corr_MAE} i IL_{Corr_MRE}. Ens trobem que l'error és molt

menor en X'' que en X'. És a dir, la correlació entre variables es manté molt més en X'' que en X'.

No podem dir simplement quin conjunt presenta major pèrdua d'informació que un altre perquè, com hem vist, aquesta pèrdua depèn de què observem. Si utilitzem les dades en una anàlisi per a la qual la correlació entre variables és molt important, el conjunt X'' presenta menor pèrdua. Per a altres casos, per exemple, si volem calcular mitjanes parcials, el conjunt X' serà més adequat.

A la vista dels resultats, sembla clar que, encara que en tots dos casos s'ha utilitzat soroll additiu com a mètode de protecció, en el cas de X' aquest soroll és **no correlacionat** i en X'' és **correlacionat**.

3. MDAV es pot aplicar *a priori* a qualsevol tipus de dades. De fet, podem veure un registre com un vector en què cada element pot ser d'un tipus de dades diferent. Per poder estendre l'algorisme, hem de veure quines operacions es fan sobre els vectors numèrics i reemplaçar-les per operacions que puguin acceptar altres tipus de dades com a paràmetre.

Si ens hi fixem bé, les operacions necessàries sobre vectors són de dos tipus:

- **Distància**. Necessitem calcular la distància entre vectors per poder determinar quins registres són els més distants (línies 3, 6 i 11 de l'algorisme 2). En l'algorisme s'utilitza la distància euclidiana. Per tant, és necessari canviar aquesta distància amb la distància apropiada per a altres tipus de dades. Alguns exemples poden ser l'ús de la distància cosinus per a vectors binaris, la distància de Jaccard per a conjunts de valors, l'*edit distance* per a cadenes de caràcters o símbols, etc.

- **Agregació**. L'agregació fa referència al càlcul de la mitjana de diversos vectors. Aquesta operació es fa en calcular el vector mitjà del conjunt de dades (línies 2 i 10) i en calcular el centre de cada microclúster (línies 4, 7, 12 i 14). De la mateixa manera, abans hem de definir una funció d'agregació que s'adapti al tipus de dades que tenim. Alguns exemples poden ser la mitjana o mitjana convexa per a dades ordinals, o la regla de pluralitat (o mode) per a dades nominals. Hi haurà tipus de dades per a les quals es podran definir funcions d'agregació específiques, i en general es pot compondre diferents agregadors per a dades heterogènies.

Glossari

classe d'equivalència *f* Cadascun dels conjunts de registres indistingibles entre ells en *k*-anonimitat. També es denomina *conjunt d'anonimat*.

enllaç de registres basat en distància *m* Tècnica que s'utilitza per determinar quins registres entre dos conjunts de dades fan referència a la mateixa entitat fent ús d'alguna mesura de distància entre registres.

generalització *f* Mètode de protecció no pertorbatiu.

identificador *m* Atribut que per si sol identifica de manera inequívoca un individu o entitat.

***k*-anonimitat** *f* Model de privadesa que determina que ha de haver-hi com a mínim *k* registres indistingibles en les dades protegides.

***l*-diversitat** *f* Propietat que requereix un cert nivell de diversitat en els atributs confidencials d'una classe d'equivalència.

mètode no pertorbatiu *m* Mètode de protecció que redueix la precisió o detall de la informació sense introduir informació errònia.

mètode pertorbatiu *m* Mètode de protecció que altera les dades generant informació errònia.

microagregació *f* Mètode de protecció pertorbatiu.

microdades *f* Conjunt de dades que es pot veure com una matriu on cada fila correspon a un individu o entitat i cada columna a un atribut o variable sobre l'individu o entitat.

pèrdua d'informació *f* Mesura de la quantitat d'informació que es perd en aplicar un mètode de protecció.

privadesa diferencial *f* Model de privadesa que requereix que la mateixa consulta a dues bases de dades que difereixen en un sol element retorni el mateix resultat o dos resultats molt semblants.

quasiidentificador *m* Atribut que per si sol no identifica de manera inequívoca un individu o entitat, però sí que pot arribar a fer-ho en combinació amb altres quasiidentificadors.

rank swapping *m* Mètode de protecció pertorbatiu.

reidentificació *f* Possibilitat d'identificar una entitat o individu en unes dades protegides.

revelació *f* Informació privada que es pot obtenir sense autorització. En el nostre cas, és la informació privada que s'obté a partir de dades protegides. Idealment, no hauria d'haver-hi revelació en dades protegides.

revelació d'atribut *f* Revelació d'informació sobre un individu o entitat que no arriba a permetre la seva reidentificació.

revelació d'identitat *f* Reidentificació d'un individu o entitat en unes dades protegides.

risc de revelació *m* Associat a unes dades, determina el risc que aquestes dades proporcionin informació sensible o considerada com a privada i que no es vol revelar.

soroll additiu o multiplicatiu *m* Mètode de protecció pertorbatiu.

***t*-proximitat** *f* Propietat aplicable a dades que compleixen *k*-anonimitat que requereix que la distribució dels atributs confidencials a cada classe d'equivalència segueixi la distribució del total de dades.

utilitat *f* Mesura de la utilitat de les dades per a una anàlisi. En dades protegides es mesura respecte a les dades originals i constitueix una estimació de la pèrdua d'informació.

Bibliografia

Aggarwal, C. C.; Yu, P. S. (editors) (2008). *Privacy-preserving data mining: models and algorithms*. Berlín: Springer.

Brand, R. (2002). «Microdata protection through noise addition. Inference control in statistical databases». *Lecture Notes in Computer Science* (vol. 2316). Berlín: Springer.

Casas Roma, J.; Romero Tris, C. (2017). *Privacidad y anonimización de datos*. Barcelona: Editorial UOC.

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database anonymization privacy models, data utility, and microaggregation-based inter-model connections*. Califòrnia: Morgan & Claypool.

Dwork, C.; Roth, A. (2014). «The algorithmic foundations of differential privacy». *Foundations and Trends in Theoretical Computer Science* (vol. 9, núm. 3-4, pàg. 211-407). <https://www.cis.upenn.edu/ aaroth/Papers/privacybook.pdf>

Torra, V. (2017). «Data privacy: foundations, new developtments and the big data challenge». *Studies in Big Data* (vol. 28). Berlín: Springer.

Willenborg, L.; de Waal, T. (2001). «Elements of statistical disclosure control». *Lecture Notes in Statistics*. Berlín: Springer.