k-Anonimat i Privacitat diferencial en smart meter data

Arnau Coll Gramunt

Resum—Aquest projecte se centra en la comparació entre dos models d'anonimització de dades, *k*-anonimat i privacitat diferencial, en el context de dades de comptador intel·ligent. L'objectiu del treball és, per una banda, veure com afecte l'aplicació de cada model d'anonimització a la precisió i utilitat de les dades, usant una mètrica específica per la pèrdua d'informació, i per l'altra banda avaluar el nivell de privacitat proporcionat mitjançant record linkage. En *k*-anonimat s'utilitza l'algorisme Mondrian, el qual es fonamenta en la microagregació i és de tipus multivariable. Per privacitat diferencial es fa servir el mecanisme de Laplace per afegir soroll. El projecte comprèn des de l'estudi dels models d'anonimització, a la codificació dels algorismes els quals estan públics, fins a la fase final de determinar quins paràmetres d'ajustament aporten millors resultats en les dues tècniques, minimitzant la pèrdua d'informació i proporcionant el nivell de privacitat més alt possible.

Paraules clau—privacitat , Mondrian, Privacitat diferencial, dades de comptador intel·ligent, anonimització de dades, *k*-anonimat, microagregació.

Abstract—This project focuses on comparing two data anonymization models, *k*-anonymity and differential privacy, in the context of smart meter data. The aim of the study is twofold: firstly, to observe how the application of each anonymization model affects the precision and utility of the data, using a specific metric for information loss, and secondly, to evaluate the level of privacy provided through record linkage. The Mondrian algorithm, based on microaggregation and of a multivariate type, is employed for *k*-anonymity. For differential privacy, the Laplace distribution is used to add noise. The project encompasses everything from studying anonymization models to encoding the algorithms, which are publicly available, and the final phase of determining which adjustment parameters provide the best results in both techniques, minimizing information loss and providing the highest possible level of privacy.

Index Terms— privacy, Mondrian, Differential privacy, smart meter data, data anonymization, *k*-anonymity, microaggregation.

1 Introducció - Context del treball

Els smart meters o comptadors intel·ligents són eines molt útils les quals canvien la forma de funcionament de les xarxes elèctriques. En essència són aparells connectats a la xarxa que recullen dades exactes del consum elèctric en curts intervals de temps, per enviar-los després a l'empresa subministradora. Permeten, a part de fer una facturació exacta de l'energia elèctrica consumida sense la necessitat de fer estimacions, les quals poden generar errors, també possibiliten gestionar millor la demanda i optimitzar la distribució de l'energia.

A finals del 2018, segons la Comissió Nacional dels Mercats i Competència [1], prop del 98% dels comptadors de tipus 5 (domèstics) ja eren de tipus intel·ligent, fet que fa que el dia d'avui al 2024 puguem dir que ja disposem tots d'aquests comptadors i que s'ha acabat l'era dels electromecànics.

Preocupacions sobre la utilització i tractament de les dades generades a partir dels smart meters han sorgit, ja que no només són utilitzades per les empreses subministradores per millorar el servei, sinó que poden ser compartides amb proveïdors de serveis de tercers, o amb investigadors amb l'objectiu de proporcionar més informació sobre el consum d'electricitat.

Mentre que els smart meters no emmagatzemen dades identificadores com el nom, DNI o compte del banc del client, les soles dades del consum elèctric es consideren sensibles a la privacitat, ja que poden revelar hàbits generals i estils de vida de les persones.

Podríem pensar que la utilització d'un pseudònim per fer referència a cert usuari seria una solució per garantir la privacitat, no sabent a l'individu a qui representa la informació. Aquest pseudònim no és suficient per mitigar la violació de la privacitat de l'usuari, ja que està demostrat que es pot realitzar una de-pseudonimització de les dades fent servir un simple algorisme d'aparellament entre les dades de consum i les dades de facturació amb un identificador real [2].

Existeixen molts tipus d'atacs que es poden fer a les dades de consum, com atacs de re-identificació o el NILMA [3] el qual busca deduir els tipus d'electrodomèstics utilitzats en un domicili a partir del seu consum total

- *E-mail de contacte: arnaucg02@gmail.com*
- *Menció realitzada: Tecnologies de la Informació*
- *Treball tutoritzat per: Guillem Navarro Arribas (DEIC)*
- *Curs 2023/24*

d'energia.

La conclusió que podem extreure és que compartir les dades de consum elèctric en la seva forma original vulnera la seguretat i privacitat dels consumidors, per tant, hem d'aplicar alguna mesura a les dades per remeiar aquest problema.

Hi ha hagut diferents solucions que han estat proposades al llarg del temps, però en aquest treball desenvoluparem i avaluarem els models de privacitat diferencial i de *k*-anonimat.

En relació amb l'estat de l'art, el camp d'estudi sobre l'anonimització de dades smart meter no està gaire desenvolupat. És un camp relativament nou i s'està estudiant en l'actualitat, sorgint estudis en els darrers anys. Aquests treballs recents consideren algun tipus d'agregació pel model de *k*-anonimat [4], [5], [6]. Mentre que s'ha estudiat també privacitat diferencial en dades smart meter, algunes propostes actuals es basen en l'ús d'eines criptogràfiques les quals queden fora de l'abast d'aquest treball. Aquest projecte resulta interessant pel fet de comparar *k*-anonimat amb privacitat diferencial.

2 OBJECTIUS

Els objectius del projecte són múltiples i es llisten tenint en compte tant l'ordre de progrés temporal com la prioritat d'assoliment.

1. Estudiar, implementar i avaluar els models d'anonimització de *k*-anonimat i privacitat diferencial amb dades de smart meters. El nucli i l'objectiu principal del treball és fer una comparació entre els dos models d'anonimització observant quins paràmetres aporten millors resultats i quins paràmetres entre els dos models són equivalents.

2. Analitzar les bases teòriques de Mondrian i privacitat diferencial. Conèixer com funcionen els models de privacitat i els mètodes associats. Comprendre el perquè del seu ús i resultats.

3. Avaluar el nivell de privacitat proporcionat per cada tècnica d'anonimització. Fer ús de mètriques com record linkage per determinar la privacitat assolida.

4. Mesurar l'impacte en la utilitat de les dades després d'aplicar els models d'anonimització, avaluant la pèrdua d'informació assolida.

5. Implementació de l'algorisme Mondrian i privacitat diferencial amb mecanisme de Laplace.

6. Concloure sobre quin mètode i amb quins paràmetres és el que aporta millors resultats. Un cop havent fet l'anàlisi de resultats es buscarà donar recomanacions fonamentades sobre quin mètode és millor respecte a les mètriques considerades en el cas de voler anonimitzar dades de smart meter.

3 METODOLOGIA

Per l'assoliment de les diferents fites s'ha seguit una metodologia simple però efectiva. Tenint clar l'objectiu, realitzar recerca d'informació en recursos audiovisuals, articles científics i llibres educatius per obtenir el coneixement necessari per desenvolupar el que es vol. En el cas de la implementació de les tècniques d'anonimització s'han consultat recursos en línia per conèixer els fluxos dels algorismes i també veure maneres de posar-los en pràctica.

4 SMART METERS

4.1 Infraestructura

Els smart meters formen part d'una xarxa intel·ligent. Aquesta xarxa intel·ligent inclou sistemes de gestió de comunicacions amb l'estructura de flux de potència elèctrica de subministrament. Permeten a part de flux d'electricitat el transport d'informació. Aquesta xarxa intel·ligent fa ús de l'AMI (Advanced Metering Infraestrcure), permetent processar i recol·lectar la informació dels smart meters a través de protocols de comunicació, canals específics i també d'un sistema de gestió de dades (MDMS). EL MDMS emmagatzema les dades després de rebre-les dels concentradors de dades, els quals són nodes dintre de l'AMI i concentren informació rebuda de varis smart meters, i l'envien als servidors del proveïdor de serveis (MDMS). Aquests concentradors de dades estan a barris o zones reduïdes.

Amb aquesta explicació superficial del funcionament de l'ecosistema dels smart meters veiem que les dades de consum elèctric passen per molts nodes i cal preguntar-se en quin moment és l'adequat per aplicar els algorismes d'anonimització. Per donar una resposta a aquesta pregunta hem de determinar si considerem que l'empresa de serveis és un agent fiable, si confiem en el fet que faci anònimes les nostres dades abans d'emmagatzemar-les o processar-les. En cas que confiem en l'empresa subministradora es podrien aplicar les tècniques d'anonimització en els concentradors de dades, abans que arribin al MDMS, aprofitant que als concentradors s'ajunten dades de diferents usuaris es pot utilitzar *k*-anonimat i anonimitzar les dades abans que arribin al centre de gestió.

Per altra banda, si es determina que l'empresa subministradora no és confiable es poden aplicar algorismes d'anonimització en el propi smart meter abans d'enviar les dades. Se sol utilitzar privacitat diferencial local en models de zero confiança.

4.2 Dades

Com ja s'ha deixat entreveure en la introducció les dades amb les quals treballen els smarts meters i que poden suposar un risc a la privacitat són els consums d'energia elèctrica en diferents intervals de temps. En l'àmbit d'operació local guarden més dades, però no són d'interès per aquest treball, ja que no considerem atacs interns.

En la figura 1 es pot veure una taula de dades com les que emmagatzemen i comuniquen els smart meters.

CUSTOMER_KEY	Time of Reading	General Supply KWH
8150103	05:59:59	0.042
8150103	06:29:59	0.088
8150103	06:59:59	0.107
8150103	07:29:59	0.04
8150103	07:59:59	0.042
8150103	08:29:59	0.041
8150103	08:59:59	0.049
8150103	09:29:59	0.189
8150103	09:59:59	0.051
8150103	10:29:59	0.05
8150103	10:59:59	0.05

Fig. 1. Representació de les dades generades i emmagatzemades per un smart meter.

Les dades amb les quals es treballa en aquest projecte són reals i pertanyen a mesures de cases de Londres, ofertes pel govern Britànic amb col·laboració de l'empresa UK Power Networks [7]. Per avaluar els algorismes s'han formatat les dades originals de mode que han quedat una fila per punt de consum on cada columna representa un consum cada mitja hora en KWH. A l'utilitzar dades reals en vers a dades simulades ens permet cenyir-nos a un escenari realista en el moment d'obtenir conclusions.

5 *K*-ANONIMAT

El *k*-anonimat és un model de privacitat i alhora una propietat que poden complir conjunts de dades, en la que per cada registre existent n'hi ha $k-1$ més els quals són indistingibles del primer, tenint tots els mateixos atributs. Vist d'una altra forma és també una mesura de risc en la que agents externs puguin reidentificar persones o obtenir informació la qual no s'hauria de revelar. El *k*-anonimat s'aplica en les dades quan els atributs identificadors han estat eliminats i, per tant, només queden quasi-identificadors, els quals estan associats als atributs sensibles que són els interessants per donar a conèixer, també anomenats atributs confidencials.

La definició formal de *k*-anonimat (1)

Diem que el conjunt de dades D satisfà k-anonimat per un valor de k quan:

Per cada fila $f_1 \in D$, existeixen almenys unes altres $k-1$ files $f_2 \dots f_k \in D$ tal que

$$C_{qi(D)^r 1}=C_{qi(D)^r 2},\dots,C_{qi(D)^r 3}=C_{qi(D)^r k} \quad (1)$$

on $qi(D)$ són els quasi-identificadors de D, i $C_{qi(D)^r}$ representa les columnes de f que contenen quasi-identificadors.

En aquest model s'assumeix que no hi ha relacions entre entrades en el conjunt de dades ni que tampoc hi ha registres duplicats, sent tots els registres independents. S'ha de tenir en compte si aquesta independència no es dona i prendre mesures, ja que sinó es pot perdre efectivitat.

En el cas que es compleixin els requisits de zero relació esmentats anteriorment, *k*-anonimitat protegeix molt efectivament contra atacs de re-identificació amb record linkage (vegeu secció 7.2). Els atacants no poden reidentificar cap registre en específic, ja que n'existeixen *k* d'indistingibles en la base de dades protegida. Per altra banda, tenen $1/k$ possibilitats de fer una re-identificació, però no poden saber de cap manera si aquesta correspon al registre que estan intentant reidentificar o no (2).

$$KR.\mathrm{Re}id(B,A) \leq (0,1/k) \quad (1)$$

KR.Reid representa les dues maneres per mesurar el risc de re-identificació. *B* fa referència al conjunt de registres dels quals disposa l'atacant i *A* el conjunt de dades original. El valor de 0 indica que no hi ha registres a *B* que l'atacant hagi reidentificat amb certesa. Hi ha *1/k* registres que l'atacant ha re identificat però no té certesa sobre la correctesa dels enllaços.

En el moment de voler aplicar la *k*-anonimat a un algorisme o sistema és clau determinar un bon valor de *k*. Existeix una correlació positiva entre el nivell de privacitat i *k*, sent més alt el valor de *k* més privacitat que s'obté. Passa al revés amb la utilitat de les dades, ja que disminueix com de més *k* registres són els grups de dades. Trobar el valor de *k* no és trivial perquè s'han de tenir en compte molts factors. S'ha de tenir en compte la heterogeneïtat o homogeneïtat de les dades, en conjunts de dades heterogenis, on les dades provenen de diferents fonts, es pot assolir en línies generals més anonimat amb menys pèrdua d'informació. El volum de dades és un altre aspecte a tenir en compte, en conjunts de dades grans hi ha més oportunitats de trobar grups de registres similars sense la necessitat d'agrupar dades amb valors distants i, per tant, perdre informació de forma considerable. Per veure com afecte realment cert valor de *k* a les dades s'hauria d'avaluar de forma experimental, provant diferents valors i veure com afecte a la privadesa o utilitat de les dades. Utilitzar mètriques de pèrdua d'informació o simulacions d'utilització de dades anonimitzades pot ajudar a determinar el millor valor de *k*.

5.1 Mondrian

Mondrian és un algorisme basat en el model de *k*-anonimat el qual és de tipus gregari, multivariable, i implementa de forma heurística la microagregació. Al ser un algorisme de tipus multivariable, és a dir, que es treballa amb tots els atributs alhora, la complexitat de trobar la partició òptima que minimitza la pèrdua d'informació o maximitza la utilitat de les dades fa que aquest problema sigui NP-difícil. Mondrian, per tant, no garanteix trobar la solució òptima i recorrem a mètodes heurístics per aproximar la solució. Va ser proposat el 2006 [8] i s'utilitza en aplicacions o sistemes de protecció de dades des d'aleshores.

L'algorisme, primer de tot escull l'atribut el qual és particionera. S'ha de deixar clar que estem anonimitzant i que treballem amb els atributs quasi-identificadors, i no els atributs confidencials o identificadors. El criteri que s'utilitza normalment per escollir aquest atribut és la varietat de valors que pren. Si escollim un atribut amb valors que no variïn i l'utilitzem per dividir el conjunt de

dades, crearem particions poc heterogènies. Seguidament, a partir d'un valor del domini de valors de l'atribut, partim a la meitat el conjunt de dades creant dos subconjunts de la mateixa mida. Aquest procés de partiment continua de forma recursiva per cada sub-conjunt fins que hi hagi entre *k* i *2k -1* registres en cada un. Un cop ja tenim els grups de registres calculem la mitjana dels valors per cada atribut, en el cas que siguin numèrics, i substituïm aquesta mitjana pel valor anterior.

6 PRIVACITAT DIFERENCIAL

La privacitat diferencial és una propietat d'un algorisme i no d'un conjunt de dades. Per saber que un conjunt de dades compleix amb la privacitat diferencial hem de veure que l'algorisme que les ha generat la satisfaci. La definició formal de privacitat diferencial és la següent.

Un mecanisme *M* satisfà privacitat diferencial si per dos bases de dades veïnes (només difereixen en un únic registre) *D* i *D'* i tots els possibles conjunts de resultats *S*:

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S] \qquad (2)$$

Un algorisme o funció compleix amb la privacitat diferencial si la presència o absència d'un registre no canvia significadament el resultat. La tolerància que es té en el canvi la marca el valor d'*Èpsilon*. Petits valors d'*Èpsilon* requereixen que la funció produeixi resultats molt semblants i, per tant, aporta alts nivells de privacitat. La privacitat diferencial està definida en termes de distribucions de probabilitat, i seran aquestes les que utilitzarem per afegir soroll aleatori a les dades. Una de les característiques d'aquest model és que permet implementar la negació plausible, és a dir, els individus poden afirmar que les seves dades no han estat utilitzades per fer el càlcul de la funció, aquesta afirmació tindrà més credibilitat o menys depenent del valor d'*Èpsilon.*

6.1 Mecanisme de Laplace

El mecanisme de Laplace és una tècnica desenvolupada per aconseguir privacitat diferencial. Afegim soroll amb la distribució de Laplace a les dades.

En la equació 4, *F* satisfà privacitat diferencial, on *f* és una funció que retorna un valor numèric, la *s* és la sensibilitat de *f* i ε el paràmetre de privacitat.

$$F(x) = f(x) + Lap\left(\mu = 0, b = \frac{s}{\varepsilon}\right) \qquad (3)$$

Utilitzem un paràmetre de localització de 0 en la distribució de Laplace amb l'objectiu de no desviar sistemàticament el resultat en cap direcció. El paràmetre d'escala que s'utilitza es calcula com la divisió entre la sensibilitat i el valor d'*Èpsilon*. La sensibilitat d'una funció és la màxima diferència en l'output d'aquesta quan s'aplica a dos conjunts de dades que difereixen en un element. La sensibilitat s'ha de calcular per cada funció i conjunt de dades, ja que depèn d'aquests, si no es determina correctament la sensibilitat no estarem complint amb privacitat diferencial.

Pel càlcul de la sensibilitat en el projecte s'ha considerat que s'està treballant amb una funció identitat. S'ha determinat la sensibilitat com la diferència entre el valor de consum més gran i el més petit per cada franja de temps (cada columna). D'aquesta manera millorem la utilitat de les dades, ja que no agreguem un nivell uniforme de soroll als consums.

En considerar la funció identitat de cada columna de valors estem aplicant un enfocament de privacitat diferencial local, pel fet que anonimitzem les dades en grups de forma sistemàtica.

El valor de ε ens serveix per regular el nivell de privacitat, com més alt aquest valor menys privacitat que obtenim.

En el context del nostre treball i de les dades amb les quals treballem, on els consums elèctrics són decimals i de valors molt petits, la majoria no arriben a 1 kwh, hem d'utilitzar aquest valor *Èpsilon* per reduir tant com podem la pèrdua d'informació, ja que cap mínim canvi pot fer perdre la utilitat de les dades.

7 MESURES D'AVALUACIÓ DE LA UTILITAT I PRIVACITAT DE LES DADES

L'objectiu perseguit és trobar el balanç òptim entre privacitat i pèrdua d'informació. S'utilitza una mètrica de pèrdua d'informació adaptada a dades del context de smart meters per avaluar la pèrdua d'informació de les tècniques d'emmascarament. Per altra banda, es fa ús de record linkage per avaluar la privacitat.

7.1 Pèrdua d'informació

La pèrdua d'informació és una mètrica molt important a considerar a l'avaluar i comparar els models d'anonimització. A través d'aquesta mètrica identifiquem la pèrdua d'utilitat de les dades, aspecte el qual volem minimitzar tant com sigui possible.

$$PI(D, D') = \frac{1}{TN} \sum_{j=1}^{T} \sum_{i=1}^{N} \frac{|d_{ij} - d'_{ij}|}{\sqrt{2\sigma_j}} \qquad (4)$$

D és el conjunt de dades original; *D'* és el conjunt de dades anonimitzat; *T* és el nombre d'intervals de temps; *N* és el nombre de perfils o usuaris de consum; d_{ij} i d'_{ij} són els valors abans i després de ser anonimitzats, per l'interval de temps *j* i el perfil *i*; σ_j és la desviació estàndard de l'interval de temps *j* en *D*. [9].

Al normalitzar la diferència entre el valor original i el valor anonimitzat ens assegurem que tenim en compte la variabilitat inherent en cada columna, evitant que columnes amb gran variabilitat dominin la mètrica.

Com més baix el valor de *PI* menys pèrdua d'informació.

7.2 Record linkage

Record linkage és el procés de trobar registres en diferents bases de dades que corresponguin a la mateixa entitat, sense que en la major part de les vegades hi hagi atributs compartits i iguals entre els conjunts.

El record linkage ens permet avaluar el nivell de re-identificació de registres després d'aplicar una tècnica d'anonimat. En aquest treball utilitzem el record linkage per avaluar el nivell de privacitat ofert pels models d'emmascarament. Com més registres podem reidentificar entre el conjunt de dades original i el protegit, considerem que menys privacitat obtenim.

En la implementació de record linkage en aquest projecte, amb l'objectiu de reduir considerablement el temps d'execució, ja que es fan moltes execucions, s'han comparat els deu primers valors de consum de cada fila. Considerem que calcular la distància entre els deu primers valors i no amb els 48 valors totals no és significatiu alhora de trobar els enllaços correctes.

El funcionament del model de record linkage usat en el treball és diferent l'estàndard. En comptes de calcular la distància entre els primers deu valors de les columnes, s'utilitza un sistema de puntuació que permet ajustar petites diferències perquè afectin de manera considerable a la similitud. Això és especialment important perquè els valors dels consums de dades són molt petits, i diferències mínimes poden tenir un impacte significatiu en la identificació correcta dels registres. Per implementar aquest sistema, utilitzem una funció de similitud gaussiana, la qual és sensible a canvis en la proximitat de valors. Aquesta funció assigna una puntuació més alta quan els valors són més propers entre si i penalitza més quan hi ha diferències, encara que siguin petites. D'aquesta manera, es pot obtenir una mesura de similitud que reflecteixi millor la relació real entre els registres, facilitant així la tasca d'enllaçar-los correctament. Aquest enfocament pot arribar a tenir més sentit tenint en compte que una manera d'afegir soroll a les dades ha estat a partir del mecanisme de Laplace. El mecanisme de Laplace és utilitzat per garantir la privacitat diferencial afegint soroll als valors originals, la qual cosa pot introduir variacions menors però significatives en els registres. La funció de similitud gaussiana és especialment adequada per manejar aquestes petites variacions, ja que és capaç de detectar i quantificar les similituds malgrat la presència de soroll afegit, millorant així la precisió del record linkage. L'ús de distribucions en ambdós models facilita la coherència i l'eficàcia del procés de record linkage, perquè tots dos mètodes tracten les dades dins d'un marc estadístic similar.

Per altra banda, és important tenir en compte que un percentatge de re-identificació obtingut a través del record linkage no sempre reflecteix de manera directa el risc real per a la privacitat en el món real. Per exemple, un 68% de re-identificació indica que una proporció significativa dels registres pot ser enllaçada amb les dades originals, suggerint un alt risc de privacitat. Tanmateix, el context en què s'utilitzen les dades anonimitzades i la capacitat d'un adversari per accedir a informació addicional poden influir significativament en el risc real. A més, la presència d'altres mètodes d'atac o d'inferència que no són capturats pel record linkage pot afectar també la seguretat de les dades. Per tant, és crucial complementar el record linkage amb altres tècniques d'avaluació de privacitat, en el cas que es volgués testejar un mètode d'anonimització per una aplicació en el món real, a mode d'obtenir una visió completa i precisa del nivell de protecció real assolit.

8 RESULTATS

Un cop s'han posat a prova les tècniques d'anonimització obtenim els valors de pèrdua d'informació i el percentatge de re-identificació associat a cada valor de ε.

8.1 Pèrdua d'informació

Com més gran el valor de pèrdua d'informació més alta és la pèrdua d'informació en les dades. Per altra banda, com menor sigui la pèrdua d'informació més útils seran les dades per obtenir informació sobre patrons de consum.

8.1.1 Mondrian

Per l'avaluació de Mondrian s'han provat els valors de k des del dos al quinze. No s'han considerat valors més alts de k degut al nombre de valors de consum existents en les dades. La base de dades que anonimitzem compte amb 934 perfils de consum. Tot i que en termes específics de pèrdua d'informació valors bastant superiors a quinze pel valor de k no generarien valors absurds de pèrdua d'informació, per la poca variància de les dades, sí que seria absurd que un gran percentatge dels consums fossin els mateixos.

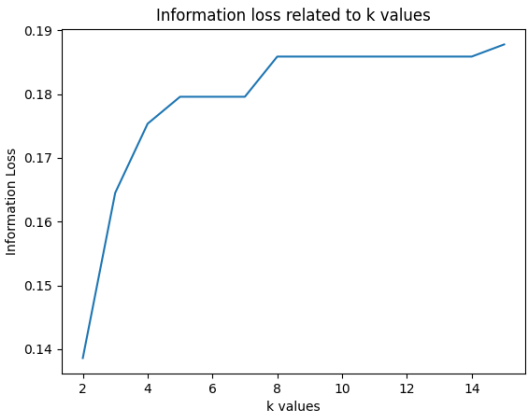

Fig. 2. Pèrdua d'informació a les dades anonimitzades per cada valor de *k*.

Com es pot observar a la figura 2, la pèrdua d'informació va augmentant a mesura que s'incrementa k. La diferència més gran en el canvi de pèrdua d'informació l'observem quan k és més gran de dos, sent aquest valor el que aporta menys pèrdua d'informació. A mesura que anem avançant la progressió en el valor de k,

l'augment de la pèrdua d'informació va disminuint, i es van formant planures que indiquen que diferents valors de *k* a partir d'un llindar no augmenten considerablement la pèrdua d'informació. El fet que s'estabilitzi la pèrdua d'informació és deu a diverses raons. En un cert punt, l'impacte addicional de fusionar més grups es redueix, ja que la variabilitat entre els grups es disminueix i la informació agregada es manté més consistent, especialment si les dades originals ja tenen una certa homogeneïtat en les seves distribucions.

8.1.2 Mecanisme de Laplace

Per l'avaluació del mecanisme de Laplace s'han provat els valors de ε del 10 fins al 370, incrementant de 10 en 10. Amb aquests valors de prova s'aconsegueix veure l'impacte quan s'utilitzen valors alts i baixos de ε. No he considerat representar valors més petits de deu, ja que per les mètriques posteriorment s'ha vist que no eren gens adequats. Al no provar una quantitat molt alta de valors ens permet poder presentar els resultats d'una forma més clara i còmode i el fet de provar cada número no canvia la tendència, sent l'interessant a mostrar. En relació a ε no existeix un valor que minimitzi la pèrdua d'informació, perquè sempre es pot anar més avall i usar una distribució la qual sumi menys soroll a les dades.

Un aspecte a tenir en compte quan avaluem el mecanisme de Laplace i que no hem de tenir en compte en avaluar Mondrian és el fet d'acabar obtenint consums negatius en les dades anonimitzades. Els valors de consum negatius no són naturals i no existeixen en les dades originals, per tant, s'han de considerar i estudiar la correlació que tenen aquests amb ε. Com menys valors de consum negatius existents a les dades més satisfactòria es considera que ha sigut la anonimització.

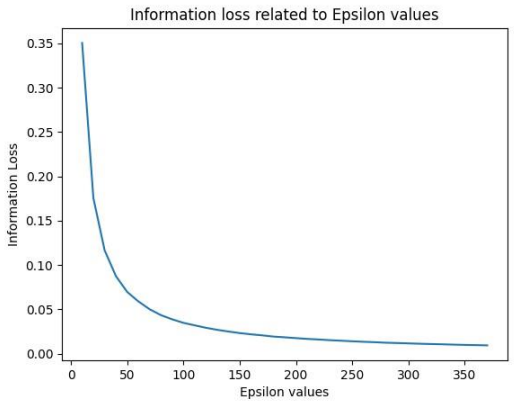

Fig. 3. Pèrdua d'informació a les dades anonimitzades per cada valor de ε.

Podem observar com la pèrdua d'informació disminueix a mesura que anem augmentant al valor de ε. Això és degut al fet de com més alt el valor de ε més baix és el paràmetre d'escala de la distribució de Laplace.

Mentre que quan ε val deu la pèrdua d'informació és molt considerable, quan tractem amb valors grans la disminució de la pèrdua d'informació és lineal i cada vegada disminueix també la diferència.

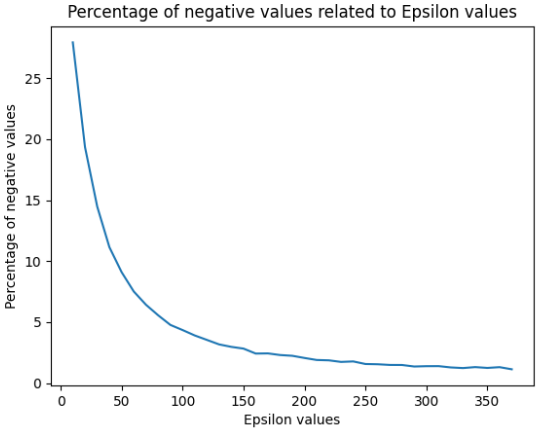

Fig. 4. Percentatge de nombres negatius amb relació als valors de ε associats.

S'observa que quan ε és inferior o igual a 50 apareixen més del 10% sobre el total de nombres negatius.

A partir de 210 el percentatge de nombres negatius ja s'estabilitza i es situa al 1%, disminuint més lentament a mesura que augmentem ε. És difícil dictaminar el límit de nombres negatius permesos i els que són acceptables.

8.2 Privadesa

Considerem el percentatge d'error en la re-identificació de registres com un indicador del nivell de privacitat assolit a les dades. Com més alt sigui el percentatge d'error d'enllaç a les dades més privacitat considerem que ha assolit el conjunt de dades anonimitzat.

8.2.1 Mondrian

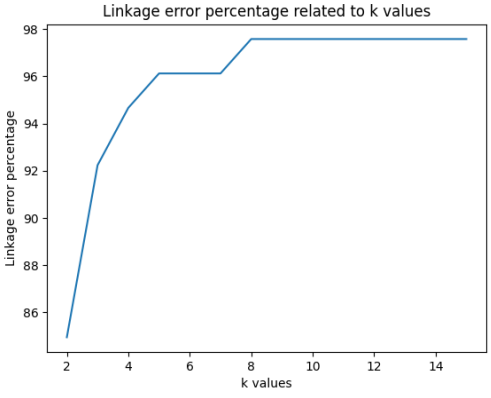

Fig. 5. Percentatge d'error de linkatge amb relació als valors de *k* associats.

És interessant comentar els resultats de les proves de record linkage per Mondrian, perquè són diferents de les prèviament esperades. Com es pot observar a la figura 5, el percentatge d'error ja és molt alt fins i tot quan *k* és igual a 2, sent del 85%. Aquest fet és en part sorprenent, ja

que en un principi podríem pensar que hauria d'estar entorn del 50% d'error, però és que la gran similitud en els valors de les dades i la poca variància d'aquestes fa que quan es calcula la distància entre possibles registres candidats en el moment de fer l'enllaç n'hi hagi molts amb una distància molt petita, fet que a més del 50% d'encertar per *k* igual a 2, dificulta molt trobar l'enllaç correcte.

8.2.2 Mecanisme de Laplace

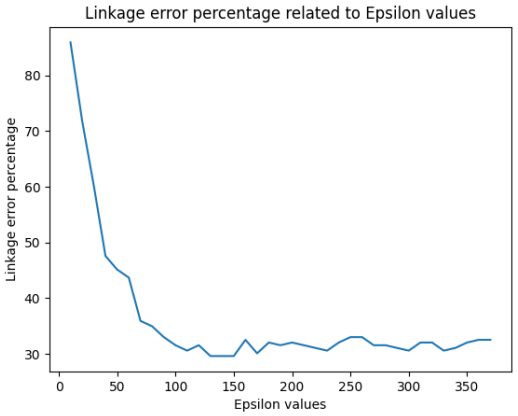

Fig. 6. Percentatge d'error de linkatge amb relació als valors de ε associats.

Els percentatges d'error de record linkage, quan els valors d'*Èpsilon* són baixos, són molt alts, ja que l'escala de la distribució de La Place que s'utilitza per afegir soroll és molt gran, sumant valors enters a nombres decimals. Per altra banda, quan els valors d'*Èpsilon* són molt baixos fan que es pugui identificar molt fàcilment quins registres són els corresponents en el data set anonimitzat.

9 COMPARATIVA

Amb l'objectiu de dictaminar quin model d'anonimització funciona millor en el context de dades smart meter hem d'analitzar si algun valor de *k* o de ε obté el mínim percentatge de re-identificació mentre que alhora genera la mínima pèrdua d'informació, tenint en compte els resultats dels dos models.

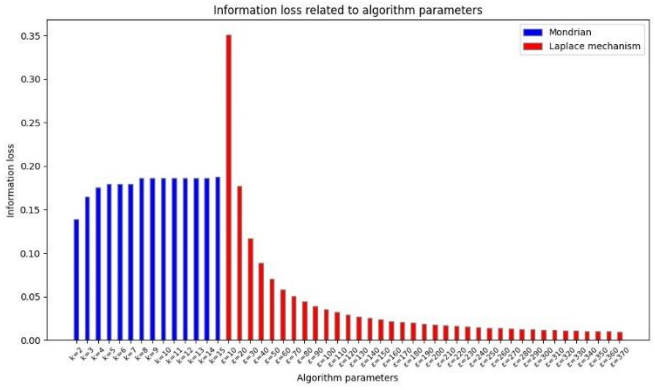

Fig. 7. Gràfica comparativa de la pèrdua d'informació a Mondrian i al mecanisme de Laplace.

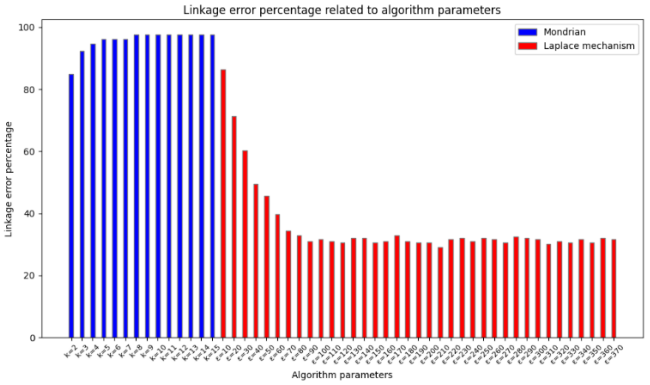

Fig. 8. Gràfica comparativa del percentatge d'error en la re-identificació a Mondrian i al mecanisme de Laplace.

Com es pot observar a la figura 7 i 8, no hi ha cap mètode d'anonimització que ofereixi alhora la mínima pèrdua d'informació i el mínim percentatge de re-identificació. Degut aquest fet, augmenta molt la complexitat de determinar quin model d'anonimització és més beneficiós. La decisió depèn de forma clau en el percentatge de re-identificació i amb el màxim o mínim que una entitat o individu consideri acceptables. Com s'ha comentat en la secció 7.2, la utilització de record linkage no és un terme absolut per jutjar la privacitat de les dades i no s'ha determinat de forma general un límit de percentatge d'error en la re-identificació de les dades a partir del qual és vulneri la privacitat o per altra banda asseguri que les dades siguin privades. També s'han de tenir en compte els valors negatius que apareixen en les dades quan les anonimitzem amb el mecanisme de Laplace. Quan ε és petita pot generar un percentatge molt considerable de nombres negatius com s'observa en la figura 4.

L'equivalència entre els paràmetres dels dos mètodes d'anonimització en termes de resultats en les mètriques són interessants d'assenyalar. En termes de pèrdua d'informació s'observa que una ε de 20 correspon a un valor de *k* de 4. Una ε de 20 genera 0,1760 de pèrdua d'informació, molt semblant al que causa $k = 2$, sent de 0,1753. Els valors de *k* de 5, 6 o 7 provoquen pèrdues d'informació similars, del 0,1795 per ser exactes. Per altra banda, en la mètrica de percentatge d'error en la re-identificació, els valors de $k = 2$ i $\varepsilon = 10$ proporcionen resultats molt semblants, un percentatge de 84,95% en ambdós casos.

En el cas de considerar només *k*-anonimat és sensat declarar $k = 2$ com la millor opció, ja que considero que ofereix una molt bona taxa de re-identificació i és el que aporta una pèrdua d'informació menor. Un percentatge de re-identificació del 85% es pot considerar molt bo, tot i que els altres valors més grans de *k* aporten percentatges més alts també aporten més pèrdua d'informació.

En termes de privacitat diferencial i el mecanisme de Laplace, si es considera com acceptable un percentatge d'error entorn del 30%, el qual té 1% de nombres nega-

tius, com més alt el valor de ε que s'esculli millor. En el cas de l' avaluació pròpia seria el valor 370, ja que ofereix un percentatge d'error de re-identificació igual que els valors més petits però amb menor número de nombres negatius i menys pèrdua d'informació. La pèrdua d'informació quan ε és de 370 és molt petita, molt menor que en el cas de $k = 2$, la millor opció de *k*-anonimat.

10 Conclusions

Partint des del supòsit de que compartir les dades de consum elèctric en la seva forma original vulnera la seguretat i privacitat dels consumidors, en aquest treball s'han examinat les implicacions en la utilitat i privacitat associades a l'anonimització de dades de comptadors intel·ligents.

S'han explorat i comparat dos models d'anonimització: el *k*-anonimat amb Mondrian i la privacitat diferencial mitjançant el mecanisme de Laplace. Els resultats mostren que, tot i que no hi ha un model que ofereixi simultàniament la mínima pèrdua d'informació i el mínim percentatge de re-identificació, ambdós mètodes tenen els seus avantatges i inconvenients.

El *k*-anonimat amb un valor de k=2 proporciona una bona taxa de re-identificació amb una pèrdua d'informació moderada, mentre que la privacitat diferencial, amb un valor òptim d'ε=370, redueix significativament la pèrdua d'informació respecte a altres valors de *k*, tot i que pot generar nombres negatius en les dades anonimitzades.

Una extensió possible del treball per aprofundir en com un escenari real impactaria l'anonimització en la utilitat de les dades smart meter, seria comparar els resultats de l'aplicació de models de predicció entre les dades originals i les anonimitzades. Observant com realment canviarien les prediccions de consum futur ens aproparia més a observar com perden verdaderament utilitat les dades.

En conclusió, és essencial seleccionar el model d'anonimització adequat segons les necessitats específiques de privacitat i la tolerància a la pèrdua d'informació acceptada en cada cas d'ús. L'estudi sobre anonimització de dades de comptadors intel·ligents continua sent un camp en desenvolupament, i les solucions bones poden variar a mesura que sorgeixin noves investigacions i tècniques.

Tot el codi relacionat amb la realització del treball està al GitHub, es poden consultar les dues implementacions de les tècniques d'anonimització avaluades al següent enllaç: https://github.com/ArnauCollGramunt/TFG-PrivacitatSmartMeterData.

Agraïments

Vull expressar el meu profund agraïment al meu tutor de Treball de Final de Grau, Guillem Navarro Arribas, per la seva guia i ajut prestat al llarg de tot el procés.

Bibliografia

[1] Informe sobre la efectiva integración de los contadores con telemedida y telegestión de consumidores eléctricos con potencia contratada inferior a 15 kw (equipos de medida tipo 5) en el año 2018 (Informe de la CNMC nº IS/DE/002/19). (2019). https://www.cnmc.es/sites/default/files/2623616_3.pdf

[2] S. Cleemput, M. A. Mustafa, E. Marin and B. Preneel, "Depseudonymization of Smart Metering Data: Analysis and Countermeasures," 2018 Global Internet of Things Summit (GIoTS), Bilbao, Espanya, 2018, pp. 1-6, doi: 10.1109/GIOTS.2018.8534430.

[3] M. Devlin and B. P. Hayes, "Non-Intrusive Load Monitoring using Electricity Smart Meter Data: A Deep Learning Approach," 2019 IEEE Power & Energy Society General Meeting (PESGM), Atlanta, GA, USA, 2019, pp. 1-5, doi: 10.1109/PESGM40551.2019.8973732.

[4] [1] Adewole, K.S., Torra, V.: DFTMicroagg: a dual-level anonymization algorithm for smart grid data. Int. J. Inf. Secur. (2022). https://doi.org/10.1007/s10207-022-00612-8.

[5] Alsaid, M., Slay, T., Bulusu, N., Bass, R.B.: K-anonymity applied to the energy grid of things distributed energy resource management system. In: Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services. pp. 581–582. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3498361.3538794.

[6] Gerlitz, C., Eriksson, A., Hansson, C.: Anonymisation score for time series consumption data. In: 27th International Conference on Electricity Distribution (CIRED 2023). pp. 428–432. Institution of Engineering and Technology, Rome,Italy (2023). https://doi.org/10.1049/icp.2023.0338

[7] SmartMeter Energy Consumption Data in London Households – London Datastore. (s.f.). London Datastore – Greater London Authority. https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[8] LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 25-25, doi: 10.1109/ICDE.2006.101.

[9] Yancey, W. & Winkler, William & Creecy, Robert. (2002). Disclosure risk assessment in perturbative microdata.