Dades sintètiques i privadesa

Josep Domingo-Ferrer

Universitat Rovira i Virgili, Tarragona
CYBERCAT-Centre de Recerca en Ciberseguretat de Catalunya

CYBERCAT

Càtedra de Privades@ de dades

josep.domingo@urv.cat

Barcelona, 15 de març de 2023

1. Introducció
2. Usos de les dades sintètiques
3. Mètodes de generació de dades sintètiques
4. Mesures d'utilitat i de confidencialitat en dades sintètiques
5. Estudi de cas: mètode de permutació
6. Conclusions

Introducció

- Idea: Generar ("inventar") dades de manera aleatòria però preservant alguns estadístics o relacions que existien en les dades originals[1].
- Les dades sintètiques són una oportunitat perquè els instituts d'estadística oficial (IEOs) puguin difondre microdades prou realistes per als usuaris, amb una bona utilitat, tot minimitzant el risc de revelar dades originals.

[1] K. Burnett-Isaacs, C. Girard, K. Sallier, G. Raab, A. Ramsden, M. Slokom, C. Task, M. Salamh, A. Wardlaw, E. Ghashim, R. Ostic, P. Belanger *et al.* 2022. *Synthetic Data for Official Statistics: A Starter Guide*. United Nations Economic Commission for Europe, Ginebra.

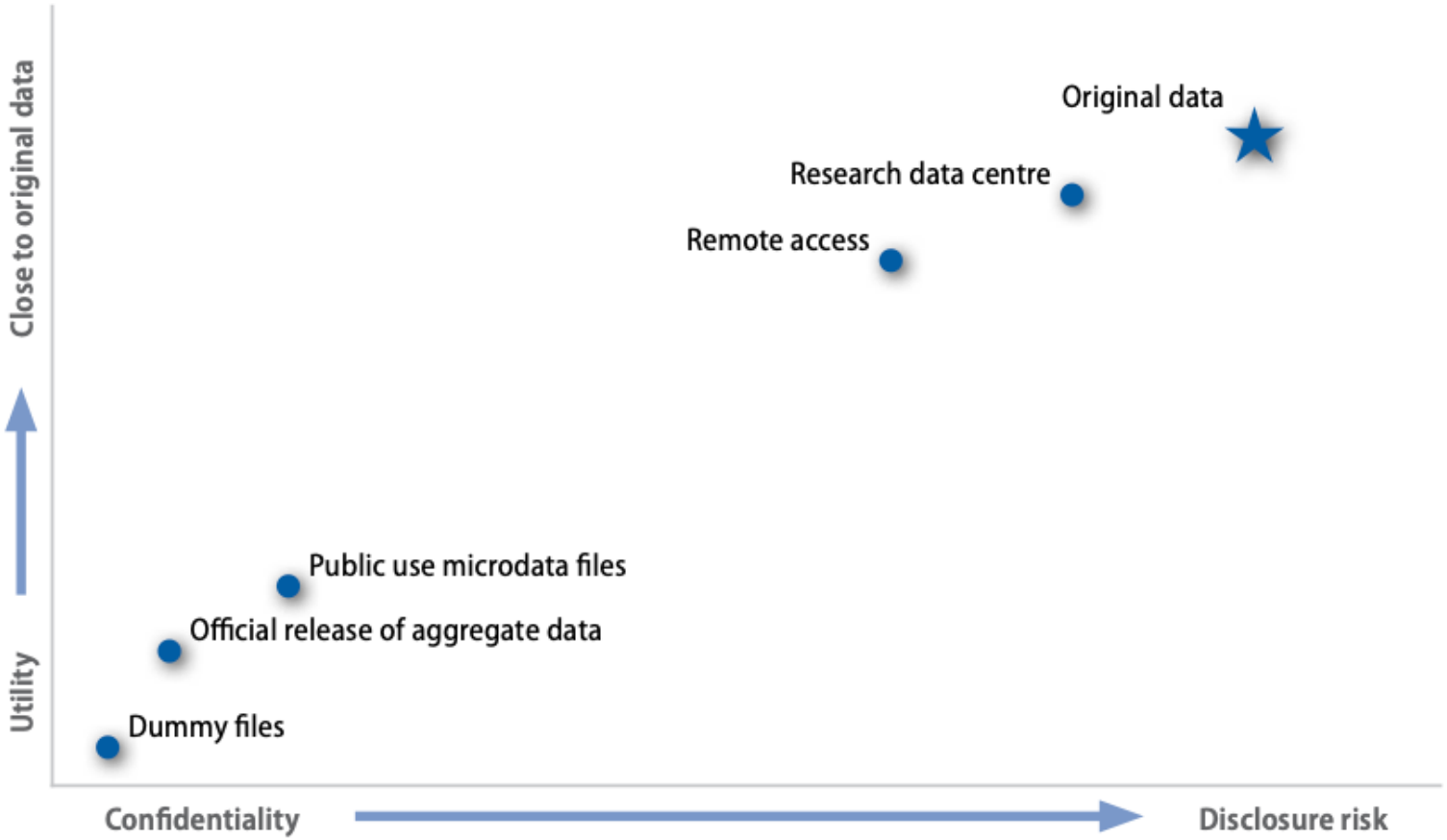
Alternatives tradicionals

- Les taules estadístiques no satisfan molts usuaris, que demanen més desagregació.
- L'accés a les microdades originals només es pot permetre en condicions molt controlades (centres d'accés segur).
- Les microdades anonimitzades (PUMF) no garanteixen la preservació exacta de les estadístiques i/o poden revelar informació sobre les microdades originals.

Per a més detalls vegeu[2].

[2] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer i P.-P. De Wolf. 2012. *Statistical Disclosure Control.* Wiley.

Confidencialitat i utilitat de les alternatives



Pros i contres de les dades sintètiques

- Pros: Sembla que protegeixen contra la reidentificació dels responents, perquè les dades són sintètiques.
- Contres:
  - Si un registre sintètic coincideix per atzar amb els atributs d'un cert responent, la reidentificació és probable i el responent no quedarà gaire consolat per molt que li diguem que les dades eren sintètiques.
  - La utilitat de les dades sintètiques es limita als estadístics i relacions preseleccionats de bon principi.
  - Les anàlisis sobre subconjunts aleatoris de dades no es preserven.
  - Les dades parcialment sintètiques o híbrides són més flexibles.

Tipus de dades sintètiques

- Fitxers ficticis. No preserven el valor analític, només l'estructura i les regles lògiques del fitxer original. Serveixen per depurar o per provar el programari. No hi ha risc de revelació.
- Fitxers totalment sintètics. Tots els atributs se sintetitzen. Poden preservar distribucions univariants, algunes de multivariants o alguns estadístics. Són semblants als PUMFs. Poden tenir més utilitat, però no estan exempts de risc.

Tipus de dades sintètiques (II)

- Fitxers parcialment sintètics. Només se sintetitzen alguns atributs (els més sensibles) o alguns valors d'alguns atributs. La resta són dades originals.
- Fitxers híbrids. Un fitxer híbrid $\mathbf{Y}$ s'obté combinant un fitxer original $\mathbf{X}$ i un de totalment sintètic $\mathbf{X}'$, per exemple via microagregació[3] o via combinació lineal:

$$\mathbf{Y} = \alpha\mathbf{X} + (1-\alpha)\mathbf{X}', \quad 0 < \alpha < 1.$$

[3] J. Domingo-Ferrer i Ú. González-Nicolás. 2010. Hybrid data using microaggregation. *Information Sciences*, 180(15):2834-2844.

Usos de les dades sintètiques

- Publicació de dades
- Desenvolupament d'anàlisis estadístiques
- Educació
- Proves de tecnologia

Publicació de dades

- El sintetitzador no sap què es farà de les dades. Cal molta confidencialitat i molta utilitat. Exemple:
  - Statistics New Zealand ha publicat fitxers de microdades sintètics (SURFs), generats amb un model matemàtic basat en les dades originals. Preserven algunes distribucions, atributs i relacions.
  - El SURF de 2007 contenia dades de renda: 11.000 registres, amb atributs Edat, Sexe, Ètnia, Nivell educatiu, Hores setmanals treballades i Ingressos.
  - El SURF de 2019 contenia dades escolars.

Desenvolupament d'anàlisis estadístiques

- Els investigadors o usuaris desenvolupen llur codi d'anàlisi estadística sobre dades sintètiques i, quan el tenen llest, l'executaran sobre les dades originals en un entorn controlat.
- El sintetitzador pot saber quines anàlisis o distribucions cal preservar. Cal molta confidencialitat i molta utilitat.
- Exemple: Dades sintètiques del *Scottish Longitudinal Study* (SLS). El SLS és una mostra del 5% de la població escocesa enllaçada al llarg del temps entre censos.
- Exemple: Prova d'algorismes d'aprenentatge automàtic a l'Australian Bureau of Statistics.

Educació

- Acadèmics, estudiants i usuaris necessiten dades per aprendre que donin resultats realistes.
- El sintetitzador pot saber les anàlisis que es faran i els usuaris poden tenir o no autorització de seguretat o un acord amb l'IEO. Cal molta utilitat i nivells variats de confidencialitat.
- Exemple: L'enquesta canadenca de mesures de salut (CHMS) inclou dades recollides amb un qüestionari, i també dades de laboratori. S'hi pot accedir en centres d'accés segur. S'han creat SUFs per a accés obert dels estudiants universitats, en els quals alguns dels atributs i dels pesos són sintètics.

Proves de tecnologia

- Per provar nou programari i nova tecnologia, calen dades fictícies que preservin l'estructura del fitxer i la taxa d'error real.
- En principi, no cal preservar utilitat, tot i que per provar mètodes d'aprenentatge automàtic pot caldre'n.
- Exemple: Prova dels programes de processament del Cens 2021 a l'*Office for National Statistics* (ONS, Regne Unit). ONS va fer servir dades sintètiques, en la majoria de casos dades fictícies, però en alguns casos preservant algunes distribucions.

Mètodes de generació de dades sintètiques

- Modelatge seqüencial
- Dades simulades
- Aprenentatge profund
- Permutació

Modelatge seqüencial

- La distribució multivariant conjunta de les dades originals no se sol conèixer.
- Modelar-la en un sol pas és massa difícil.
- Se sol descompondre el problema en el modelatge d'una seqüència de distributions condicionals i univariants:

$$f_{X_1,X_2,\ldots,X_P} = f_{X_1} \times f_{X_2|X_1} \times \ldots \times f_{X_p|X_1,X_2,\ldots,X_{P-1}}$$

- En un primer pas, hom estima en seqüència cada distribució condicional a partir de les dades originals.
- En un segon pas, hom genera valors sintètics per a cada atribut al seu torn, fent servir el model estimat per a la distribució condicional de l'atribut, a partir dels valors sintètics produïts per als atributs precedents.

Limitacions del modelatge seqüencial

- Cal decidir l'*ordre* en què els atributs se sintetitzaran, i els *models* que es faran servir per a cada atribut.
- No hi ha cap guia per triar l'ordre.
- No se sap tampoc quins models són millors, tot i que sovint es fa servir el model CART (*Classification and Regression Tree*).

El modelatge seqüencial a la pràctica

- Es pot fer servir el paquet *synthpop*[4], escrit en llenguatge R.
- Statistics Canada ha publicat dos fitxers generats amb aquest mètode, per fer-los servir en hackathons:
  - Un per a la 5th International Population Data Linkage Conference (2018).
  - L'altre per a la 2019 Canadian Cancer Research Conference.

[4] B. Nowok, C. Dibben i G. Raab. 2015. *synthpop: Bespoke creation of synthetic data in R*. `https://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf`

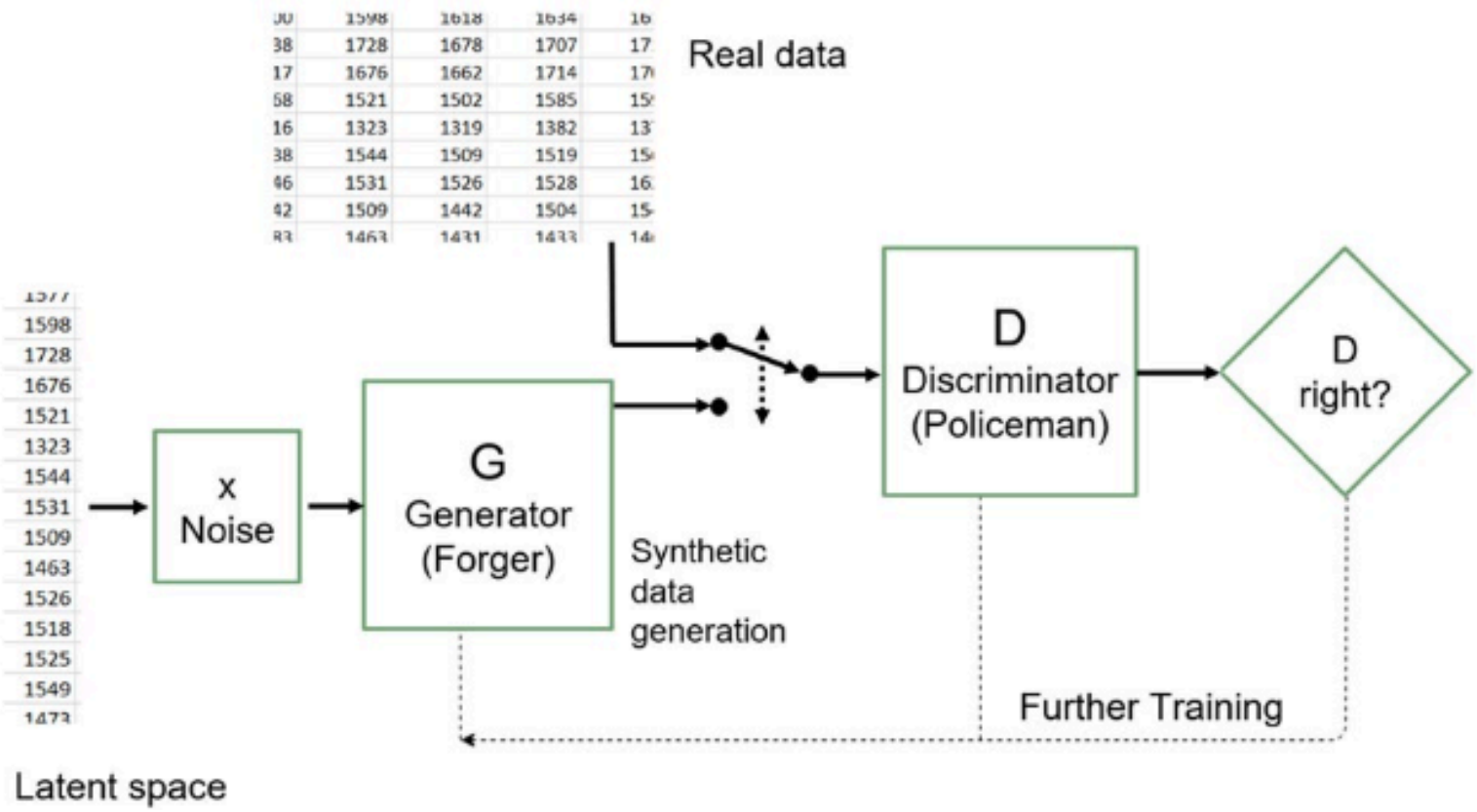
Dades simulades

- Hom pot crear fitxers sintètics mostrejant una distribució de probabilitat, sense que calguin dades originals.
- Per exemple, es pot generar un fitxer amb p atributs i n registres mostrejant p vectors independents $X_1, X_2, \ldots, X_p$ de mida n d'una distribució normal amb una certa mitjana i una certa variància.
- Els IEOs es fan servir sovint dades simulades com a dades fictícies per desenvolupar anàlisis.

Aprenentatge profund

- L'aprenentatge profund, i més concretament les xarxes adversarials generatives (GAN), es fan servir cada cop més per sintetitzar grans fitxers amb dependències complexes.
- Una GAN es basa en dues xarxes neuronals:
  1. Una xarxa (*generador*) intenta imitar l'estructura de les dades originals generant noves dades amb la mateixa distribució.
  2. L'altra xarxa (*discriminador*) rep mostres del generador i de les dades originals i intenta distingir si són sintètiques o originals.

GAN per generar dades sintètiques[5]



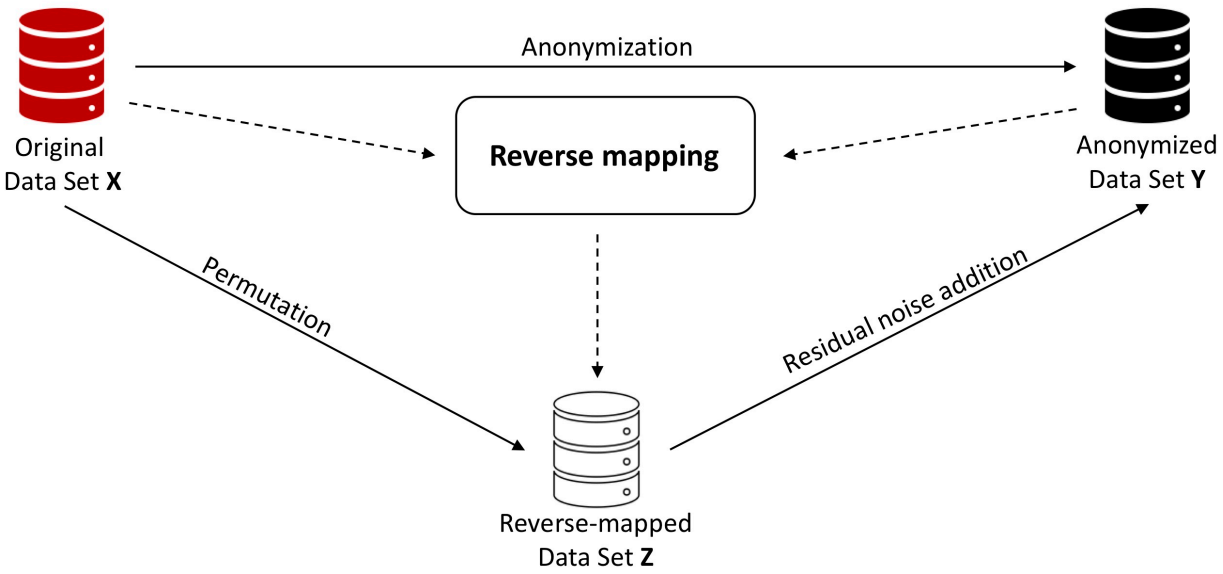
[5] I. Kaloskampis *et al.* 2020. Synthetic data in the civil service. *Significance*, 17:18-23.

Pros i contres de les GANs

- Pro: Forneixen dades sintètiques que preserven molta utilitat.
- Contres:
  - Les dades sintètiques poden coincidir amb les originals i per tant revelar-les.
  - L'aprenentatge profund requereix molt de càlcul.
  - No se sap com operen les GANs (capsa negra).

Dades sintètiques via permutació

El model de permutació[6] mostra que qualsevol mètode d'anonimització és *funcionalment equivalent* a permutar les dades originals i afegir-los un soroll residual (que no canvia els rangs resultants de la permutació).



[6] J. Domingo-Ferrer i K. Muralidhar. 2016. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337-338:11-24.

Dades sintètiques via permutació controlada[7]

1. Crear un fitxer de dades sintètiques mostrejant independentment les distribucions marginals estimades per a cada atribut original.
2. Substituir cada valor original pel corresponent valor sintètic amb el mateix rang.
3. Crear intervals de k rangs consecutius.
4. Permutar aleatòriament els valors dins de cada interval de rangs.

Amb $k = 1$ no hi ha permutació, només dades sintètiques que preserven rangs (força utilitat, poca confidencialitat). Amb $k = n$ la permutació és màxima (poca utilitat, força confidencialitat).

[7] K. Muralidhar i J. Domingo-Ferrer. 2023. Synthetic data generation via the permutation paradigm. Manuscrit sotmès.

Mesures d'utilitat

- Si el sintetitzador sap els usos que hom donarà a les dades sintètiques, pot emprar mesures d'utilitat que avaluïn l'impacte sobre aquests usos.
- Si no sap els usos, li caldran mesures d'utilitat generals:
  - Comparació de matrius de correlació.
  - Comparació de diagrames de dispersió.
  - Puntuacions de propensió (*propensity scores*[8]).
  - Mesura *UM* basada en matrius de covariància[9].

[8] M.-J. Woo, J. P. Reiter, A. Oganian i A. Karr. 2009. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1:111-124.

[9] J. Domingo-Ferrer, K. Muralidhar i M. Bras-Amorós. 2021. General confidentiality and utility metrics for privacy-preserving data publishing based on the permutation model. *IEEE Transactions on Dependable and Secure Computing* 18(5):2506-2517.

Puntuacions de propensió

1. Etiquetar els registres originals amb un atribut $I = 0$ i els sintètics amb $I = 1$.
2. Barrejar els registres originals i sintètics etiquetats.
3. Entrenar un classificador per predir I a partir dels altres atributs.
4. La utilitat U és més bona com més propera a $\hat{I} = 1/2$ és la predicció per a cada registre (és a dir, quan el classificador no pot distingir els registres originals dels sintètics.)
5. Es fa servir una fórmula per a U de manera que, com més bona és la utilitat, més s'acosta U a 0.

Mesura *UM* basada en matrius de covariància

- *UM* mesura la similitud entre la matriu de covariances de les dades originals $\mathbf{X}$ i la matriu de covariances de les dades sintètiques $\mathbf{Y}$.
- $0 \leq UM(\mathbf{X}, \mathbf{Y}) \leq 1$.
- Com més gran *UM*, més utilitat.

Mesures de confidencialitat

- Es poden fer experiments de reidentificació via enllaç de registres, però és costós i requereix fer hipòtesis sobre el que sap l'atacant.
- Es pot fer servir una mesura analítica *CM3* basada en correlacions canòniques[10].
- $0 \leq CM3(\mathbf{X}, \mathbf{Y}) \leq 1$.
- Com més gran *CM3*, més confidencialitat.
- *CM3* mesura la confidentialitat com la mínima informació mútua entre dades originals i sintètiques sobre totes les ordenacions per un sol atribut després de suprimir aquest atribut.

[10] J. Domingo-Ferrer, K. Muralidhar i M. Bras-Amorós. 2021.

Estudi de cas: mètode de permutació

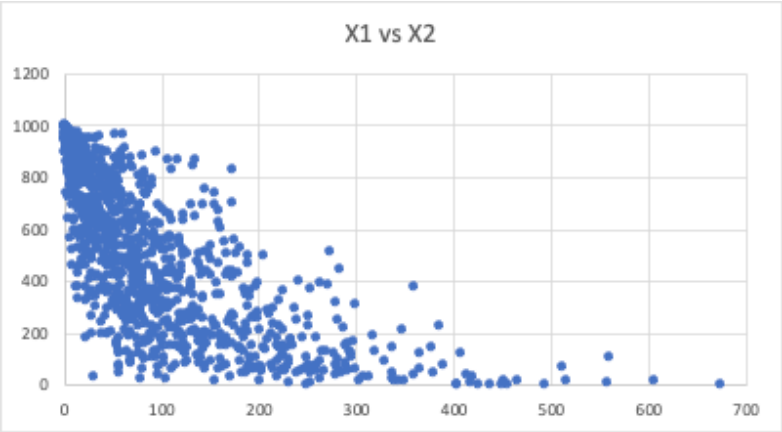
Fitxer original $\mathbf{X}$ amb $n = 1000$ registres i $p = 5$ atributs X_1, X_2, X_3, X_4 i X_5:

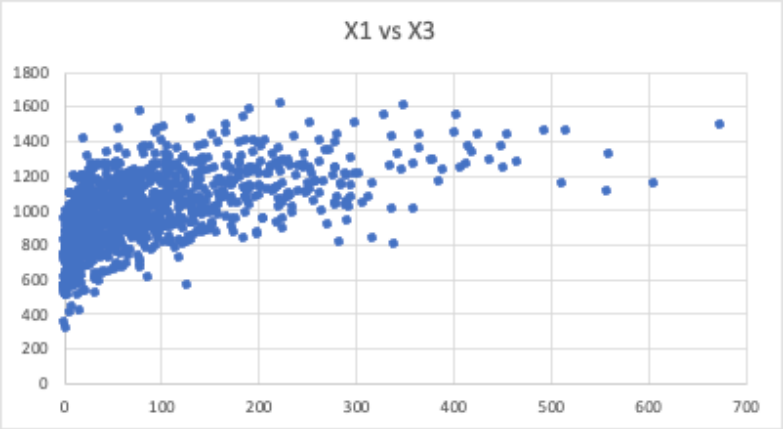
- $X_1 \sim Gamma(1, 100)$
- $X_2 \sim U[0, 1000]$
- $X_3 \sim N(1000, 200)$
- $X_4 \sim Discreta(\{1, 2, 3, 4, 5\})$
- $X_5 \sim Bernoulli(\{0, 1\})$

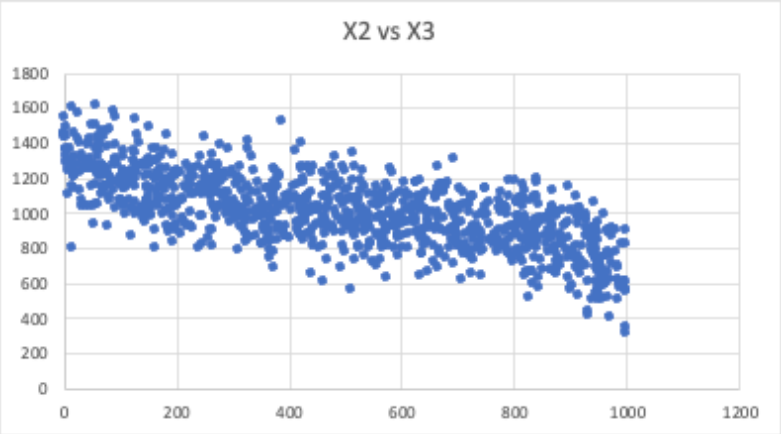
Mesures d'utilitat i de confidencialitat

k	U	UM	$CM3$
1	0.00268143	0.99990723	0.00000002
50	0.00262754	0.99997886	0.46724593
100	0.00317246	0.99974492	0.52022657
250	0.00940395	0.98950540	0.65528064
500	0.01443220	0.92641923	0.84282660
1000	0.01443220	0.66266165	0.97426504

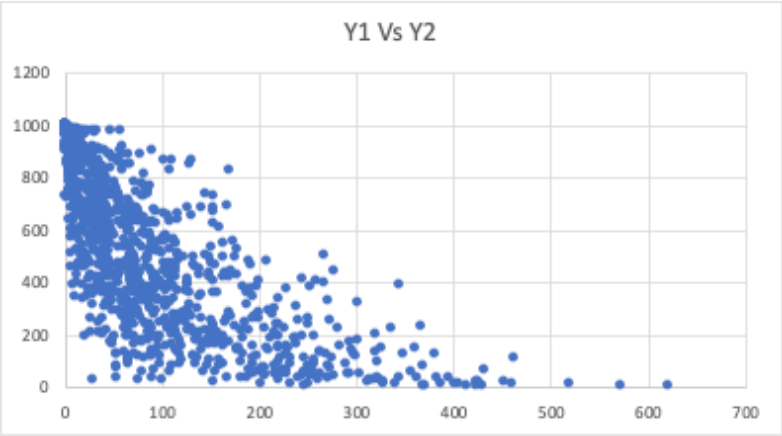
Diagrames de dispersió X_1, X_2, X_3

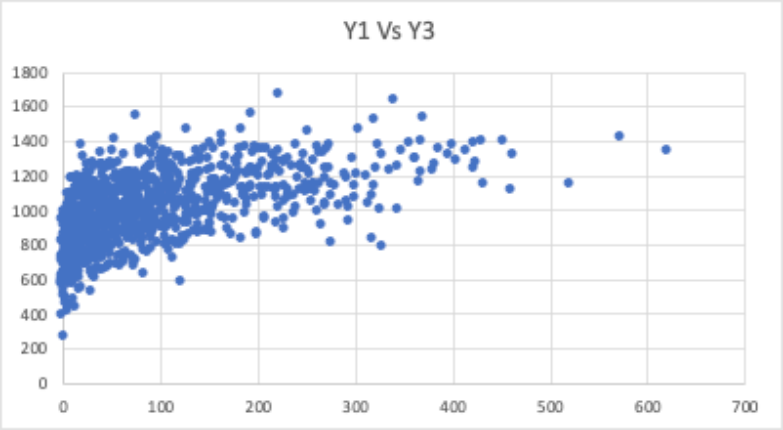


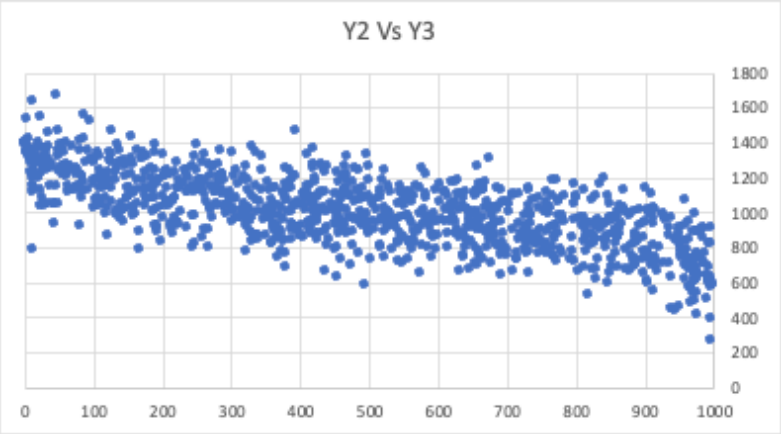




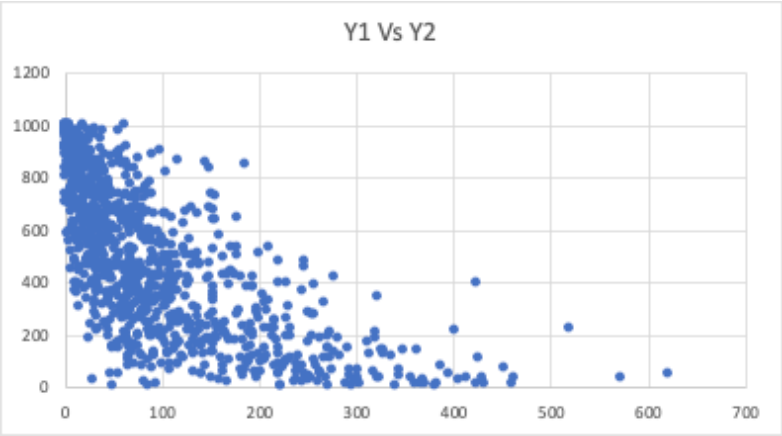
Diagrames de dispersió Y_1, Y_2, Y_3 per a $k = 1$

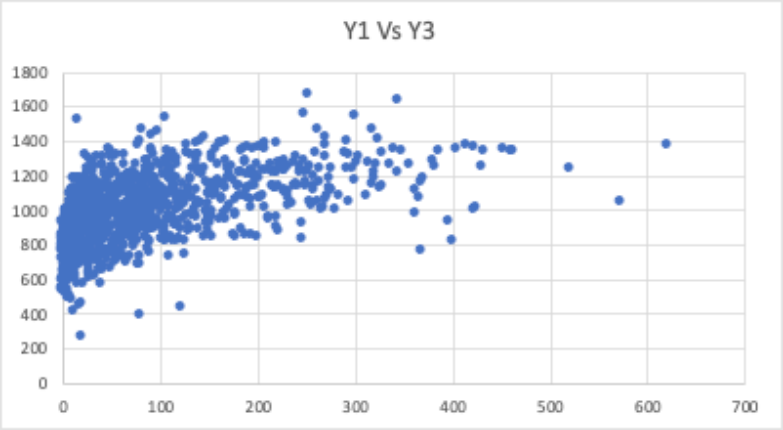


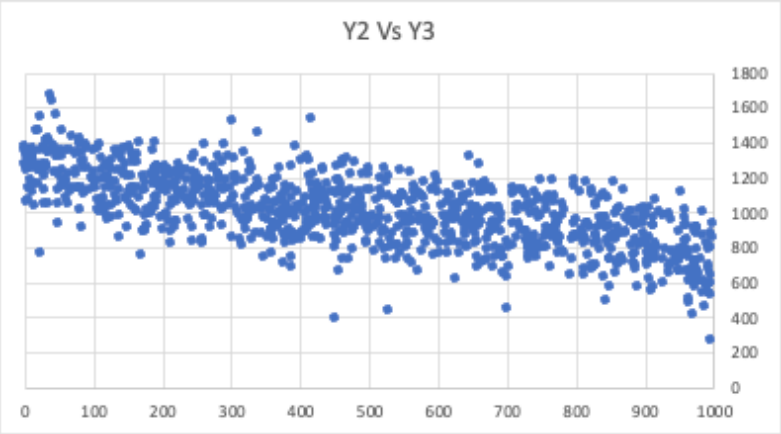




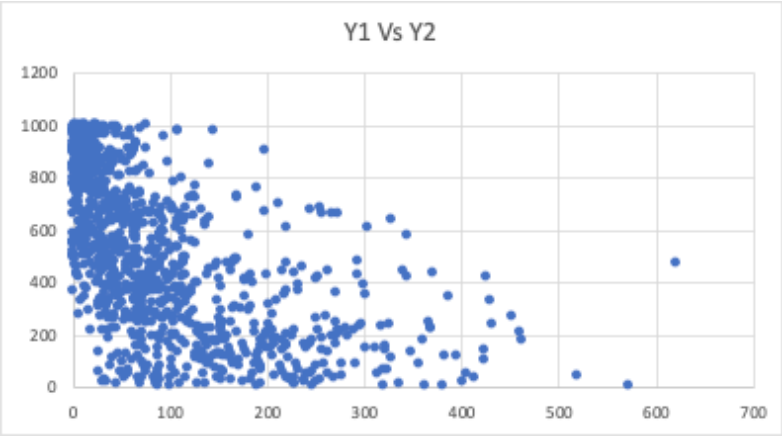
Diagrames de dispersió Y_1, Y_2, Y_3 per a $k = 50$

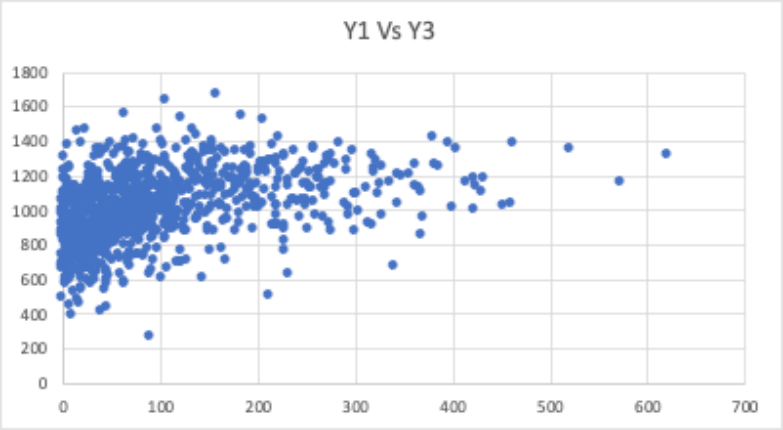


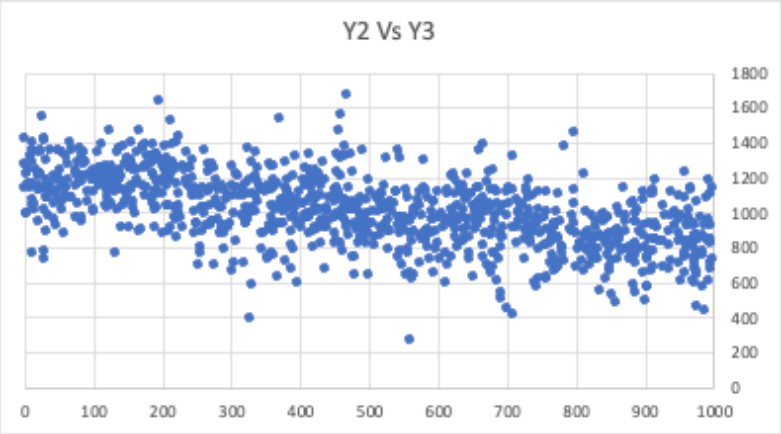




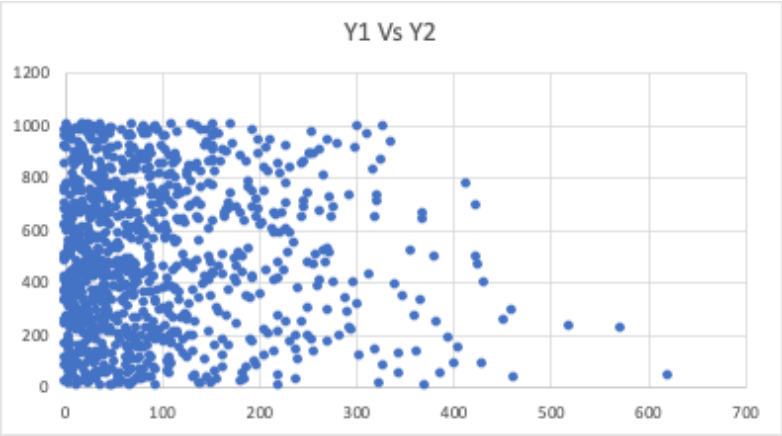
Diagrames de dispersió Y_1, Y_2, Y_3 per a $k = 250$

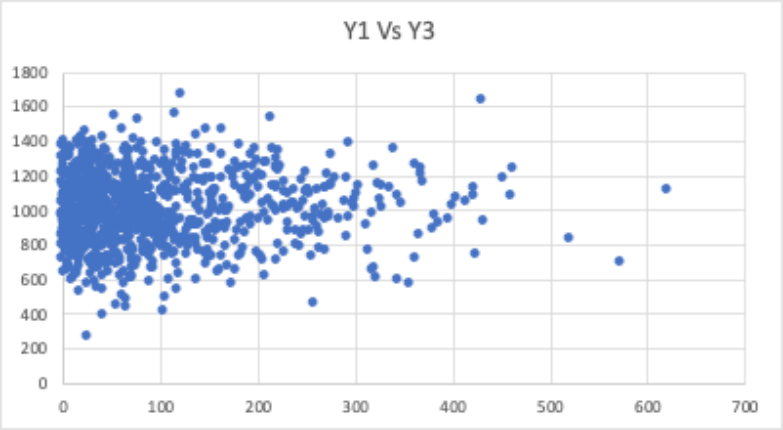


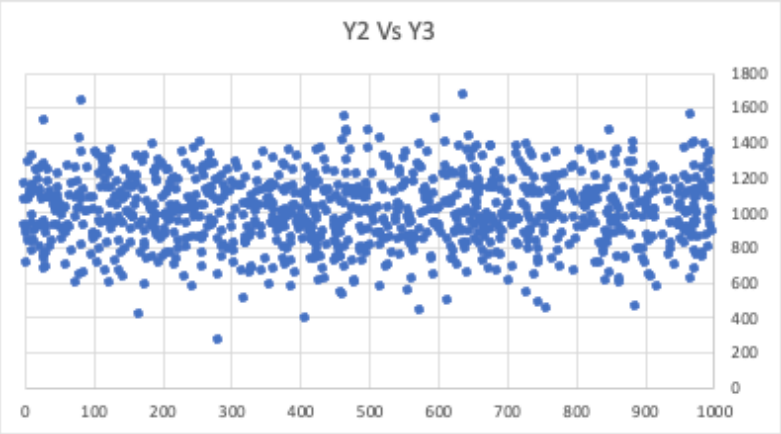




Diagrames de dispersió Y_1, Y_2, Y_3 per a $k = 1000$







Conclusions

- Les dades sintètiques són una alternativa atractiva per publicar informació estadística sense revelar informació privada.
- Cal que les dades sintètiques preservin utilitat estadística.
- Cal anar amb compte que les dades sintètiques no siguin massa semblants a les originals registre a registre.
- Hem vist diferents mètodes de generació i diferents mesures d'utilitat i de confidencialitat.
- L'informe Burnett-Isaacs *et al.* (2022) és un bon document per introduir-se en aquest tema.

GRÀCIES PER LA VOSTRA ATENCIÓ!!!!!