Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4

Exercici 1.- Eliminar els identificadors del conjunt de dades (0.5p)

Responeu a les següent qüestions:

a. Quins són els atributs identificadors en el conjunt de dades?
Els atributs que identifiquen una persona de manera única i inequívocament són el DNI i el número de la SS.

b. Descriu les comandes que has emprat, indicant breument la funció de cadascun dels paràmetres que has utilitzat.
dades_subset = dades[,c(3,4,5)]
dades_subset_comparar = dades_subset
dades_subset_comparar$CP = as.numeric(dades_subset_comparar$CP). Amb aquesta comanda creem una nova variable "*dades_subset*" on hi guardarem el contingut de les columnes 3, 4 i 5 que corresponen a CP, Edat i Salari i a mes creem "*dades_subset_comparar"* que es el mateix però amb el CP com a *num* en comptes de *char* per poder fer comparacions numèriques.

Exercici 2.- Re-identificació (0.5p)

Responeu a les següent qüestions:

a. Quins d'aquests atributs o combinacions poden conduir a la identificació única d'un registre del conjunt de dades?
Hem trobat que tant per Edat, CP i Salari hi ha persones que tenen valors únics i llavors qualsevol dels 3 atributs poden identificar de manera única a una persona concreta. A més a més també hem trobat que les combinacions CP-Edat, CP-Salari, Edat-Salari i CP-Edat-Salari identifiquen a una persona concretament.

b. Indica quines comandes has fet servir per obtenir aquesta informació, explicitant el significat de les comandes i dels seus paràmetres.
Amb la comanda "*dades_CP_freq = table(dades_subset$CP)*", creem una nova taula a partir del subset que hem treballat en l'apartat anterior per a la columna CP. Com a sortida d'aquesta comanda rebrem una taula amb els diferents valors que pren el CP i la freqüència amb la que aquests valors surten, és a dir, la quantitat de persones amb cadascun dels CP.

Per fer una combinació, per exemple per saber quantes persones tenen el mateix CP i edat, la comanda és molt similar:
"*dades_CP_Edat_freq = table(dades_subset$CP,dades_subset$Edat)*"
Ara tenim una taula que ens mesura per a la variable1(CP) i per a la variable2(Edat) les combinacions possibles i quants cops hi apareixen.

Un cop hem provat totes les combinacions i tots els camps que considerem que poden ser quasi-identificadors utilitzarem:
"'1' %in% TAULA"
on *"TAULA"* és el nom de la taula generada anteriorment com per exemple "*dades_CP_freq*" i ens mostrarà "*TRUE*" si hi ha cap individu que es pot identificar de manera única.

Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4

Exercici 3.- Aplicació de soroll additiu (additive noise) (0.5p)

Responeu a les següent qüestions:

a. Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.

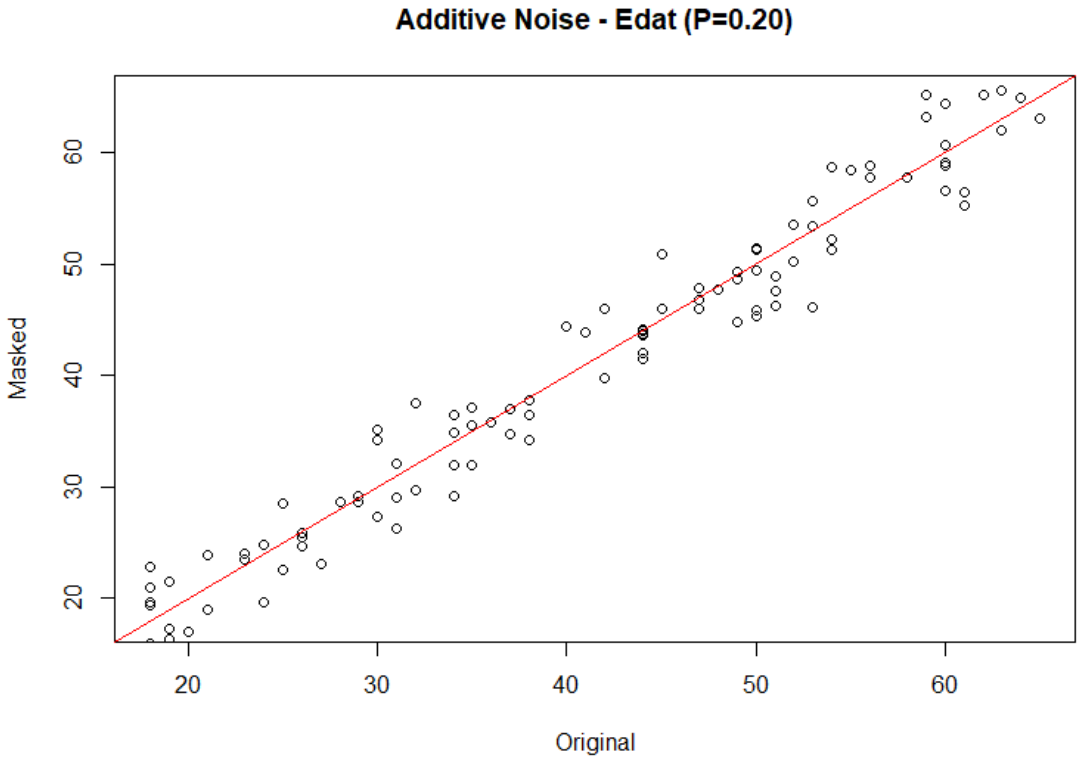
Nosaltres ho hem fet en 3 passos:

1: *"edat_soroll <- addNoise(dades_subset_comparar,'Edat',20)"*. Amb això creem un objecte de tipus *sdcMicroObj-class* en el qual guardarem el soroll que hem afegit a la variable edat, el primer paràmetre és el *data frame* a partir del qual es treballa, el segon paràmetre és l'atribut o atributs als quals es vol afegir soroll. El tercer paràmetre és la quantitat de soroll que afegim (20%).

2: *"dades.an = dades_subset_comparar"*. Creem un *data frame "dades.an"* que sigui igual que *"dades_subset_comparar"*, aquest pas el fem perquè és important l'ordre de les columnes en el *data frame* i així en assegurem que tenen el mateix ordre.

3: *"dades.an$Edat <- round(edat_soroll$xm)"*. Amb aquesta comanda guardem al camp Edat de dades.an el contingut arrodonit(ja que no té sentit parlar de decimals a l'edat) de l'atribut pertorbat a *"edat_soroll"*.

b. Executa el següent codi adjunt i explica els resultats obtinguts, justificant la resposta.



Obtenim aquesta gràfica, en l'eix d'abscisses tenim l'edat real de la persona i en l'eix d'ordenades l'edat després d'afegir el soroll. La línia vermella és per on haurien d'estar els punts en cas de no haver-hi soroll, llavors tots aquells punts que es troben per sobre

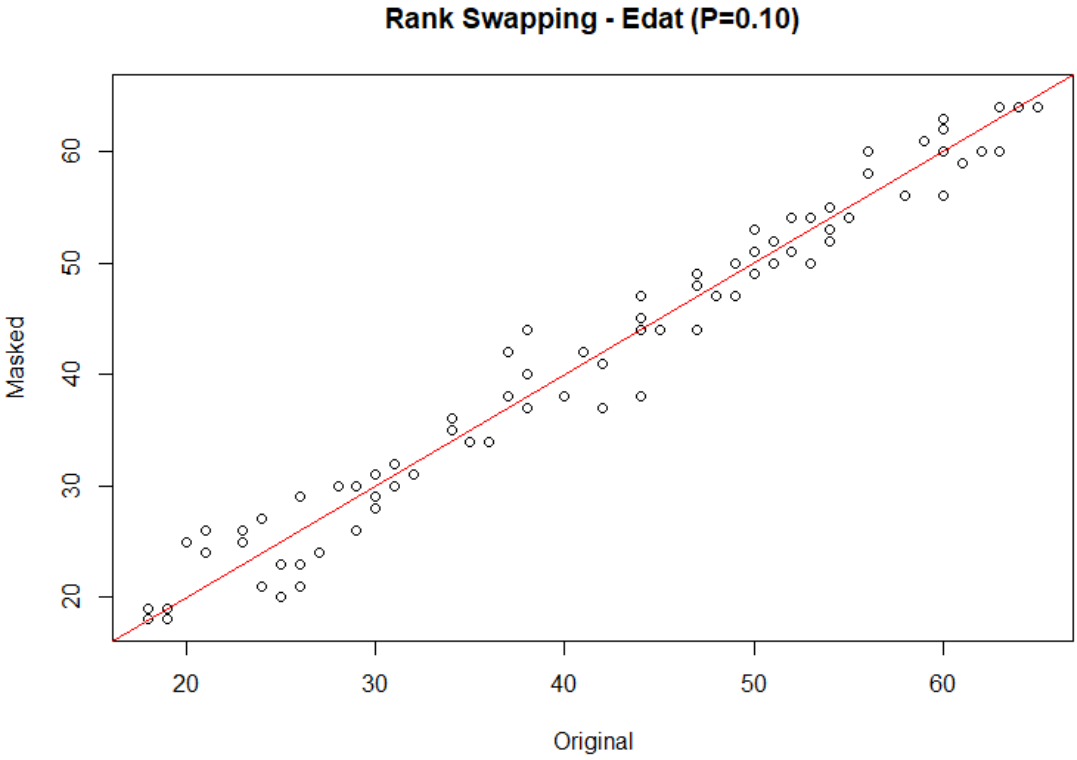
de la línia vermella són persones a les quals se'ls ha afegit edat (fins un 20%) i aquells que es troben per sota son persones a les quals se'ls ha tret edat (fins un 20%).

Exercici 4.- Aplicació del mètode d'intercanvi de rang additiu (rang swap) (0.5p)

Responeu a les següent qüestions:

a. Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.
"dades.rs <- rankSwap(dades_subset_comparar, 'Edat', P=10)". El primer paràmetre és el *data frame* a partir del qual es treballa. El segon paràmetre és l'atribut o atributs a les quals es vol fer el *rank swapping*. Finalment, el tercer paràmetre (P) és el percentatge amb el qual es farà el *rank swap.*

b. Genera una gràfica similar a la que hem emprat en l'exercici anterior – apartat b), on es pugui veure de forma visual i ràpida la dispersió o alteració de l'atribut Edat després d'aplicar el mètode d'intercanvi de rang.



Igual que en el cas anterior, en l'eix d'abscisses tenim l'edat real de la persona i, en l'eix d'ordenades, l'edat després de fer el *rank swapping*. També, com en el cas anterior, la línia vermella és on haurien d'estar els punts en el cas original. Llavors, tots aquells punts que es troben per sobre de la línia vermella són persones a les quals se'ls ha afegit edat i aquells que es troben per sota són persones a les quals se'ls ha tret edat.

Exercici 5.- Càlcul de la utilitat (0.5p)

Responeu a les següent qüestions:

a. Utilitza la comanda dUtility de la llibreria sdcMicro per a estimar la pèrdua de informació que s'ha produït en generar la versió pertorbada de l'atribut Edat emprant els mètodes de soroll additiu i intercanvi de rang (generats en els dos exercicis anteriors). Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.
Per *additive noise*:
"dUtility(obj=dades_subset_comparar, xm=dades.an)". El que fem és comparar les dades originals *"dades_subset_comparar"* amb les dades pertorbades amb *additive noise "dades.an"*.

Per *rank swapping*:
"dUtility(obj = dades_subset_comparar, xm=dades.rs)". Igual que abans, comparem les dades originals *"dades_subset_comparar"* amb les dades modificades (dades.rs), en aquest cas però, amb *rank swapping*.

b. Comenta els resultats obtinguts i descriu el significat dels valors obtingut i com es calculen.
Els resultats són molt similars però, com era d'esperar, el *rank swapping* ens dona menys pèrdua d'informació perquè hem afegit menys pertorbació.

Pèrdua de informació per *additive noise*: 11.01887

Pèrdua de informació per *rank swapping*: 9.03852

Aquests valors es calculen mesurant la distància entre les dades originals i les pertorbades, escalades per la desviació estàndard. Per a valors més alts significa que hi ha una major dispersió.

Exercici 6.- Càlcul del nivell de privacitat (0.5p)

Responeu a les següent qüestions:

a. Utilitza la comanda dRisk de la llibreria sdcMicro per a estimar el risc o nivell de privacitat en la versió pertorbada de l'atribut Edat, emprant els mètodes de soroll additiu i intercanvi de rang (generats en els dos exercicis anteriors). Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.
En ambdós casos la comanda és molt similar a l'anterior.

Per *additive noise*:
"dRisk(obj = dades_subset_comparar, xm=dades.an)". Els paràmetres són els mateixos que a l'exercici anterior, les dades originals *"dades_subset_comparar"* i les dades alterades amb *additive noise"dades.an"*.

Per *rank swapping*:
"*dRisk(obj = dades_subset_comparar, xm=dades.rs)*". De la mateixa manera, amb aquesta comanda estem estimant el risc o nivell de privacitat entre les dades originals *"dades_subset_comparar"* i les dades alterades amb *rank swapping "dades.rs".*

Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4

b. Comenta els resultats obtinguts i descriu el significat dels valors obtingut i com es calculen.

En ambdós casos obtenim *dRisk* = 1. *dRisk* és el valor en tant per 1 de la quantitat de registres que poden ser re-identificats. Això ho fa definint un interval per cada valor pertorbat. Aquest interval el calcula segons la desviació estàndard de l'atribut. Si el valor original està dins d'aquest interval, considera que pot ser re-identificat. Per tant, un valor 1 és el més dolent perquè significa que el 100% dels registres es podrien arribar a re-identificar.

Exercici 7.- Micro-agregació univariant i multivariant (1p)

Responeu a les següent qüestions:

a. Aplica el mètode de micro-agregació univariant amb un nivell d'agregació igual a 3 als atributs Edat i Salari, de forma independent, sobre conjunt de dades original. Indica les comandes i paràmetres que has fet servir i la comenta la seva funció.

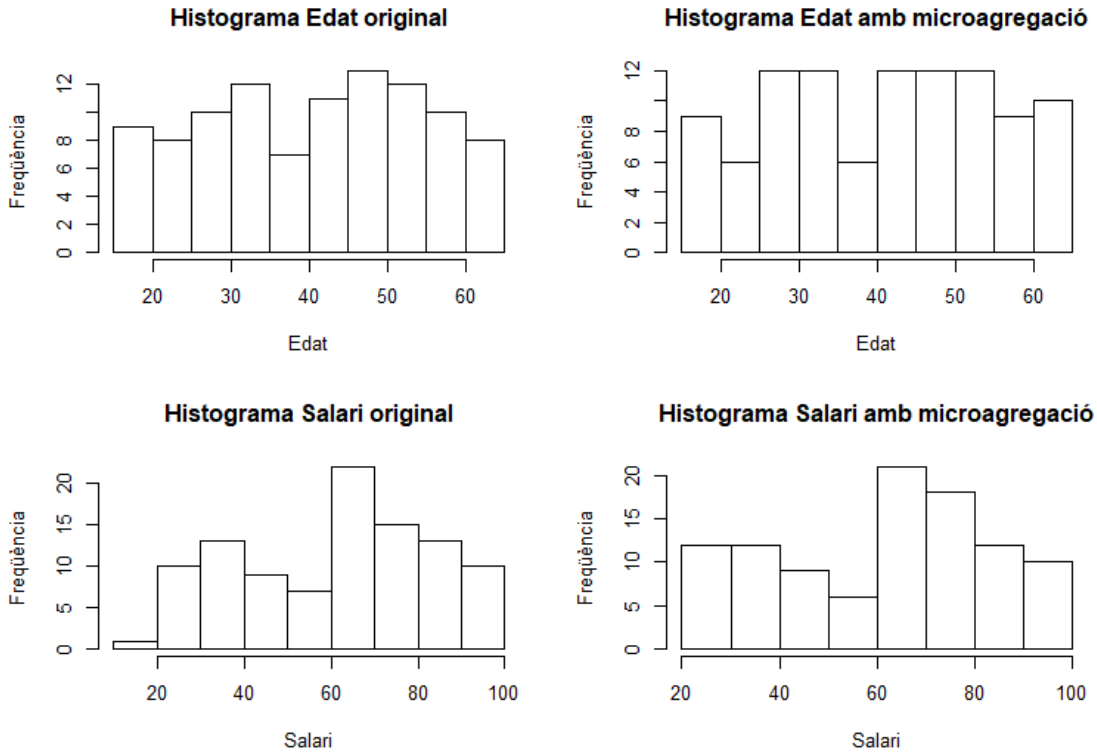
"microa <- microaggregation(obj=dades_subset_comparar, variables = c("Edat","Salari") , aggr = 3, method = "onedims")"
"dades.microa = dades_subset_comparar"
"dades.microa$Edat <- round(microa$mx$Edat)"
"dades.microa$Salari <- round(microa$mx$Salari)"

Hem utilitzat aquestes comandes per guardar en *"microa"* la micro-agregació univariant, el primer paràmetre és l'atribut original al qual es vol fer la micro-agregació, el segon paràmetre es el nivell d'agregació que, tal com se'ns diu, ha de ser de 3. I, per últim, el mètode que, en aquest cas, és univariant. Després es crea *"dades_microa"* on hi guardarem els atributs Edat i Salari modificats. Com en el cas de l'edat arrodonim per obtenir nombres enters.

b. Els histogrames són gràfiques que presenten la freqüència d'aparició segons els valors. Per tant, són molt interessants per veure quins són els valors més freqüents i quins són únics. Crea quatre gràfiques que presentin la informació de l'atribut Edat i Salari abans i després del procés de micro-agregació, i que permetin comparar de forma visual i senzilla, com s'han modificat el valors d'aquests dos atributs.

Indica el codi que has emprat, detallant la funció dels diferents Nota: la funció par(mfrow=c(2,2)) us pot ajudar a posar les 4 gràfiques en una sola imatge, de manera que simplifica la visualització.

Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4



"par(mfrow=c(2,2))"
"hist(microa$x$Edat, main = "Histograma Edat original", xlab = "Edat", ylab = "Freqüència")"
"hist(microamxEdat, main = "Histograma Edat amb microagregació", xlab = "Edat", ylab = "Freqüència")"
"hist(microa$x$Salari, main = "Histograma Salari original" , xlab = "Salari", ylab = "Freqüència")"
"hist(microamxSalari, main = "Histograma Salari amb microagregació", xlab = "Salari", ylab = "Freqüència")"

Hem creat 4 histogrames:

1. Histograma Edat original
2. Histograma Edat amb microagregació
3. Histograma Salari original
4. Histograma Salari amb microagregació

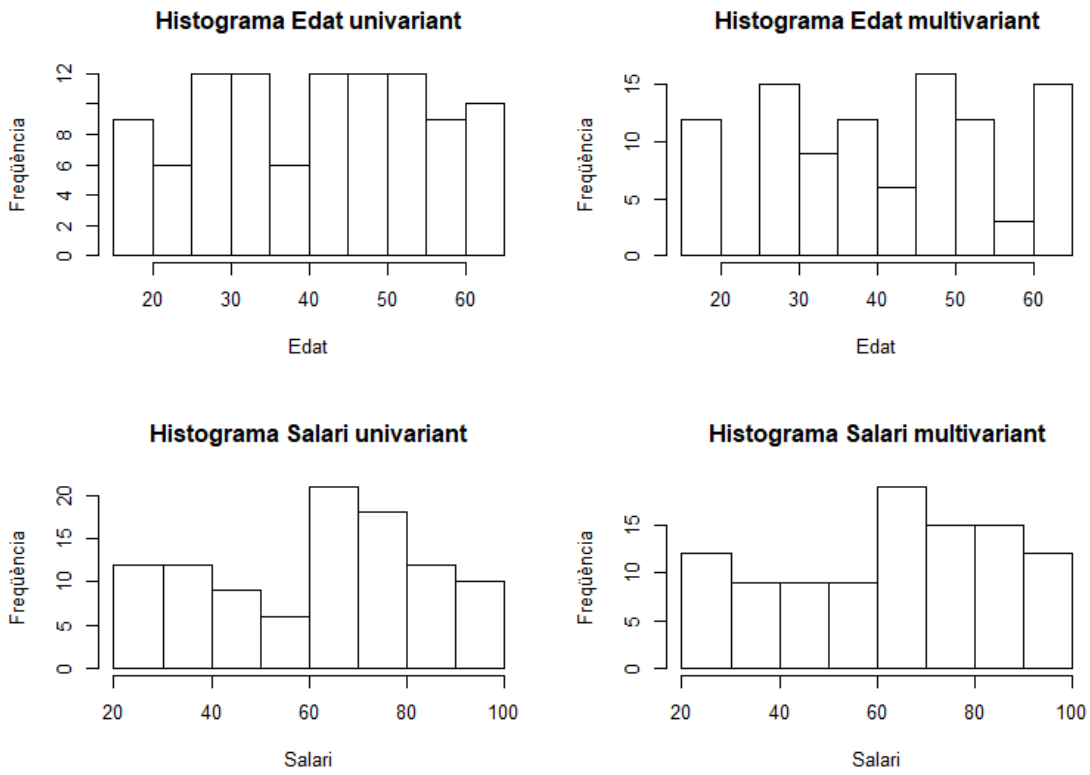
Per a tots els histogrames hem utilitzat com a primer paràmetre els valors els quals volem que siguin a l'eix d'abscisses, en aquest cas Edat o Salari. El segon paràmetre és el títol de l'histograma. El tercer paràmetre és el nom de l'eix horitzontal i, el quart paràmetre, el nom de l'eix vertical.

c. Aplica el mètode de micro-agregació multivariant MDAV amb un nivell d'agregació igual a 3 als atributs Edat i Salari sobre conjunt de dades original. Indica les comandes i paràmetres que has fet servir i la comenta la seva funció.
"microa_mv <- microaggregation(obj=dades_subset_comparar, variables = c('Edat','Salari'),aggr = 3, method = "mdav")"
"dades.microamv = dades_subset_comparar"

Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4

"dades.microamv$Edat <- round(microa_mv$mx$Edat)"
"dades.microamv$Salari <- round(microa_mvmxSalari)"
Em utilitzat aquestes comandes per guardar, en "microa_mv", la micro-agregació multivariant. El primer paràmetre és l'atribut original al qual es vol fer la micro-agregació. El segon paràmetre és el nivell d'agregació que, tal com se'ns diu, ha de ser de 3. Per últim el mètode, que en aquest cas és multivariant. Després i com en el cas anterior guardem en altre *data frame "dades.microamv"* el resultat.

d. Crea les gràfiques que presentin la informació dels atributs Edat i Salari anonimitzats en la seva versió multivariant, i compara els resultats obtinguts amb els obtinguts en el cas univariant.



"hist(microamxEdat, main="Histograma Edat univariant",xlab="Edat",ylab="Freqüència")"
"hist(microa_mvmxEdat, main="Histograma Edat multivariant",xlab="Edat",ylab="Freqüència")"
"hist(microamxSalari, main="Histograma Salari univariant",xlab="Salari",ylab="Freqüència")"
"hist(microa_mvmxSalari, main="Histograma Salari multivariant",xlab="Salari",ylab="Freqüència")"

Veiem que, per al cas de l'Edat amb el mètode univariant, les edats estan més repartides. Al cas multivariant, hi ha grups que concentren molts usuaris.

En canvi, en el cas del Salari, veiem que ens dona uns valors més repartits per al cas del mètode multivariant que no pas per el mètode univariant.

e. Realitza una estimació del risc i de la utilitat en els dos casos (univariant i multivariant). Comenta i justifica els resultats obtinguts.

Per a la utilitat:

- Càlcul de la utilitat per a la micro-agregació univariant: 3.548021
- Càlcul de la utilitat per a la micro-agregació multivariant: 19.8998

Com veiem, els resultats son els esperats, es pertorben més les dades en el cas del mètode multivariant que no pas en el mètode univariant.

Per al risc:

- Càlcul del risc per a la micro-agregació univariant: 0.97
- Càlcul del risc per a la micro-agregació multivariant: 0.41

Exercici 8.- Generalització d'atributs (1p)

Responeu a les següent qüestions:

a. Crea una funció pròpia que permeti generalitzar els codis postals de població, convertintlos en codis de província. Detalle el codi de la funció, així com alguns exemples del seu funcionament i valors de retorn.

```
"ProvinciaCP <- function(CP) {
        value = substr(CP,1,2)
        padding = "000"
        value = paste(value,padding,sep="")
        return (value)
}"
```

Amb aquesta funció, a partir d'un String d'entrada CP, agafem els 2 primers dígits que són els que ens indiquen la província i declarem un *padding* de 000 perquè els codis postals tenen una longitud de 5 dígits.
Després els concatenem i retornem el nou String.

Per exemple, algú amb el codi postal 08018 sabem que és de Barcelona ciutat i ens retorna 08000. Ara només sabem que és de Barcelona província.
Altre exemple seria 25610 (només surt un cop i és fàcil d'identificar) que sabem que és d'Os de Balaguer, ens retorna 25000 i ara només sabem que és de Lleida província.

b. Aplica la funció que has creat a l'atribut CP del conjunt de dades, per a tots els individus (registres) existents. Comenta el resultat obtingut.

"dades.CPGeneralitzat = dades_subset_comparar"
"dades.CPGeneralitzat$CP=as.numeric(ProvinciaCP(dades_subset$CP))". Amb aquestes comandes generalitzem el CP i el transformem a num per poder comparar-lo amb l'original.
Ara comparem amb l'original i obtenim:

- Utilitat: 2.037828
- Risc: 1

Com vegem, els mètodes no pertorbatius no perden gairebé gens informació cosa que també fa que sigui fàcil re-identificar un individu.

Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4

c. Aplica la funció que has creat a l'atribut CP del conjunt de dades, però només a aquells registres que siguin únics, és a dir, que la seva freqüència d'aparició sigui igual a 1 en tot el conjunt de dades. Comenta el resultat obtingut i compara-ho amb els resultats de l'apartat anterior.

Hem de modificar una mica el codi anterior i ens queda així:

```
"ProvinciaCPValorsUnics <- function(CP) {
        counter = 1
        for(CPindv in CP){
                if('1' %in% dades_CP_freq[CPindv]){
                        CP[counter] = ProvinciaCP(CPindv)
                }
                counter= counter + 1
        }
        return (CP)
}"
```

Ara comparem amb l'original i obtenim:

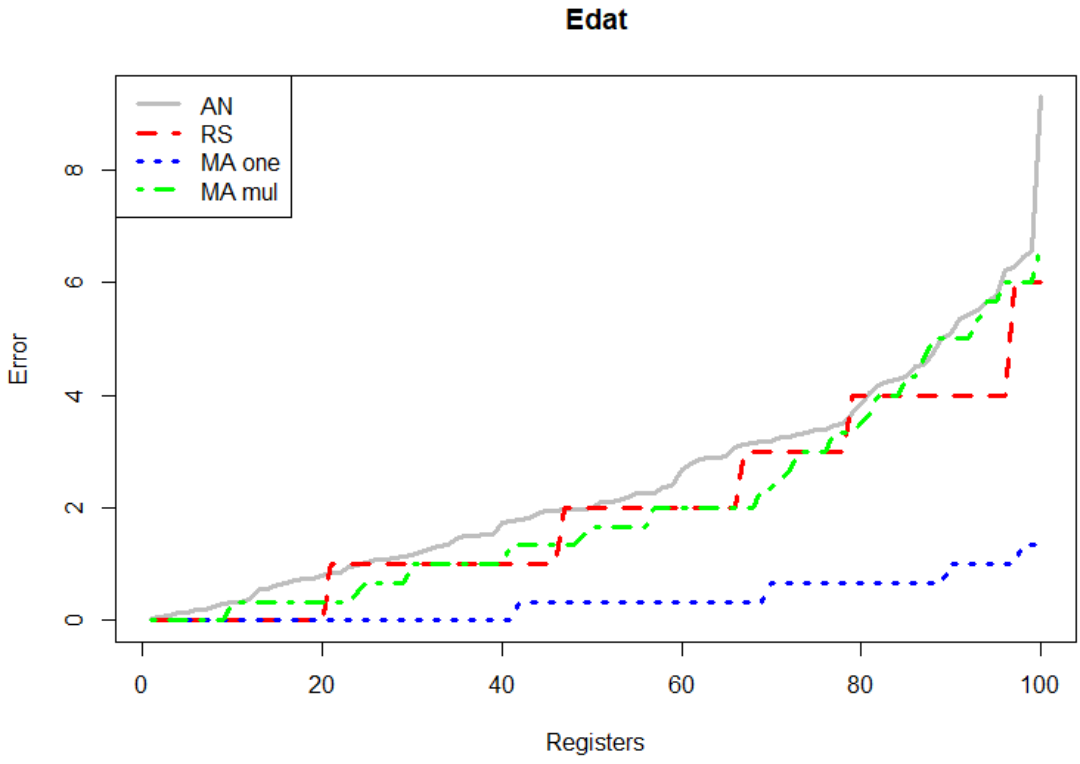
- Utilitat: 0.06534963
- Risc: 1

Gairebé no s'ha modificat el *"data frame"* original i com vegem la utilitat és molt baixa

Exercici 9.- Alternatives per mesurar la pèrdua de informació (1p)

Responeu a les següent qüestions:

a. Mostra la gràfica obtinguda.



Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4

b. Comenta i justifica els resultats obtinguts.

En aquesta gràfica es mostra l'error en els registres segons les pertorbacions que hem utilitzat. La que introdueix un valor major d'error és *Additive Noise,* en part perquè hem posat un valor més elevat (20%). *Rank Swapping* es veu clarament quan introdueix error i és quan efectua un canvi de rang, per això és veu tan esglaonat. Per micro-agregació univariant veiem que és la que menor error introdueix i que també està esglaonat perquè també fa intercanvis en les dades. Per micro-agregació multivariant es veu que esta esglaonada, però que puja més ràpidament ja que hi introdueix més pertorbació.

Exercici 10.- Alternatives per mesurar la pèrdua de informació (1p)

Responeu a les següent qüestions:

a. Expliqueu, atribut per atribut, quin tipus d'atribut és (numèric, categòric nominal, etc.) i a quina classe correspon segons la classificació de privacitat (identificador, quasi-identificador, etc.).

b. Expliqueu, atribut per atribut, quin mètode d'emmascarament heu utilitzar en cas de què calgués alterar l'atribut. Justifiqueu les decisions presses en cada cas.
DNI: L'hem eliminat perquè es tracta d'un identificador i volem garantir privacitat.

Sexe: L'hem deixat perquè considerem que és un atribut interessant i que no cal alterar per garantir privacitat als usuaris.

Tel: L'hem eliminat perquè l'única informació que podem obtenir del telèfon de manera significativa és la província cosa que també la podem obtenir a partir del CP i a més a més si coneixem el número de telèfon d'una persona la podem identificar cosa que és un problema per a la privacitat.

CP: Hem aplicat mètode no pertorbatiu, l'hem generalitzat perquè considerem que la província és una informació geogràfica força interessant, però que no cal ser tan precís com identificar el CP exacte.

Edat: Hem aplicat diversos mètodes pertorbatius. Hem fet rank swapping, com hem vist en el conjunt de dades anterior és un bon mètode per modificar l'atribut original. L'hem fet amb p = 10 perquè considerem que és un rang prou ampli com poder afegir privacitat i prou petit com per obtenir resultats realistes. A més, després de fer *rank swapping* afegim *additive noise* amb p = 10 per garantir més privacitat a l'edat dels usuaris. També hem considerat una bona idea fer micro-agregació multivariant amb el Salari l'hem fet amb 5 nivells perquè volem alterar més el resultat que no com passava en data set anterior.

NumFills: Hem aplicat un mètode pertorbatiu. Hem fet micro-agregació multinivell. Fem micro-agregació multinivell amb el nivell de estudis perquè tots dos tenen valors que estàn entre el 0 i el 4 i s'alteren les dades però com es un rang petit en les dues variables si altera l'atribut original no el desvia molt.

Juan Manuel Vallecillos Calzado – 1401596
José Antonio Cegarra Alonso – 1461305
GIS-b4

NivellEstudis: Mateixa explicació que al NumFills

Ocupacio: Hem aplicat metode no pertorbatiu. El que hem fet ha estat reduir el nombre de profesions a els camps professionals.

Salari: Hem aplicat mètode pertorbatiu. Hem afegit un 10% de soroll.

Prob: Creiem que es l'atribut fonamental d'aquest data set i que ha de ser verídic. Per tant, no l'hem modificat.

c. Avalueu el nivell de privacitat que heu assolit en el conjunt final, una vegada anonimitzats tots els atributs necessaris.
Obtenim un drisk que varia entre 0.1 i 0.24. Normalment està entre 0.12 i 0.16 pero hi ha casos molts bons en el que està al voltant de 0.1 i casos molt dolents en el que està al voltant de 0.2. Considerem que un nivell de privacitat de 0.1 es prou bo i no hem volgut baixar per no perdre informació.

d. Avalueu el nivell de utilitat de les dades que heu assolit en el conjunt final, una vegada anonimitzats tots els atributs necessaris. Us podeu ajudar de gràfiques i altres recursos que considereu rellevants.
Obtenim un dutility que varia entre 0.48 i 0.51. Son valors molt consistents i gairebé sempre s'obtenen valors dintre d'aquest rang. Considerem que per a un drisk de 0.1 aquest dutility està prou bé perquè si baixem el drisk el dutility puja dràsticament.

```
> dUtility(obj=dades_subset_nopertorb_nonumeric, xm=dades_subset_pertorb_nonumeric)
[1] 50.83904
> dRisk(obj=dades_subset_nopertorb_nonumeric, xm=dades_subset_pertorb_nonumeric)
[1] 0.1
```