Oslo, 2018

Data privacy: sdcMicro and sdcTable

Vicenç Torra

January 15, 2018

Privacy, Information and Cyber-Security Center
SAIL, School of Informatics, University of Skövde, Sweden

Outline

1. sdcMicro
2. sdcTable

sdcMicro

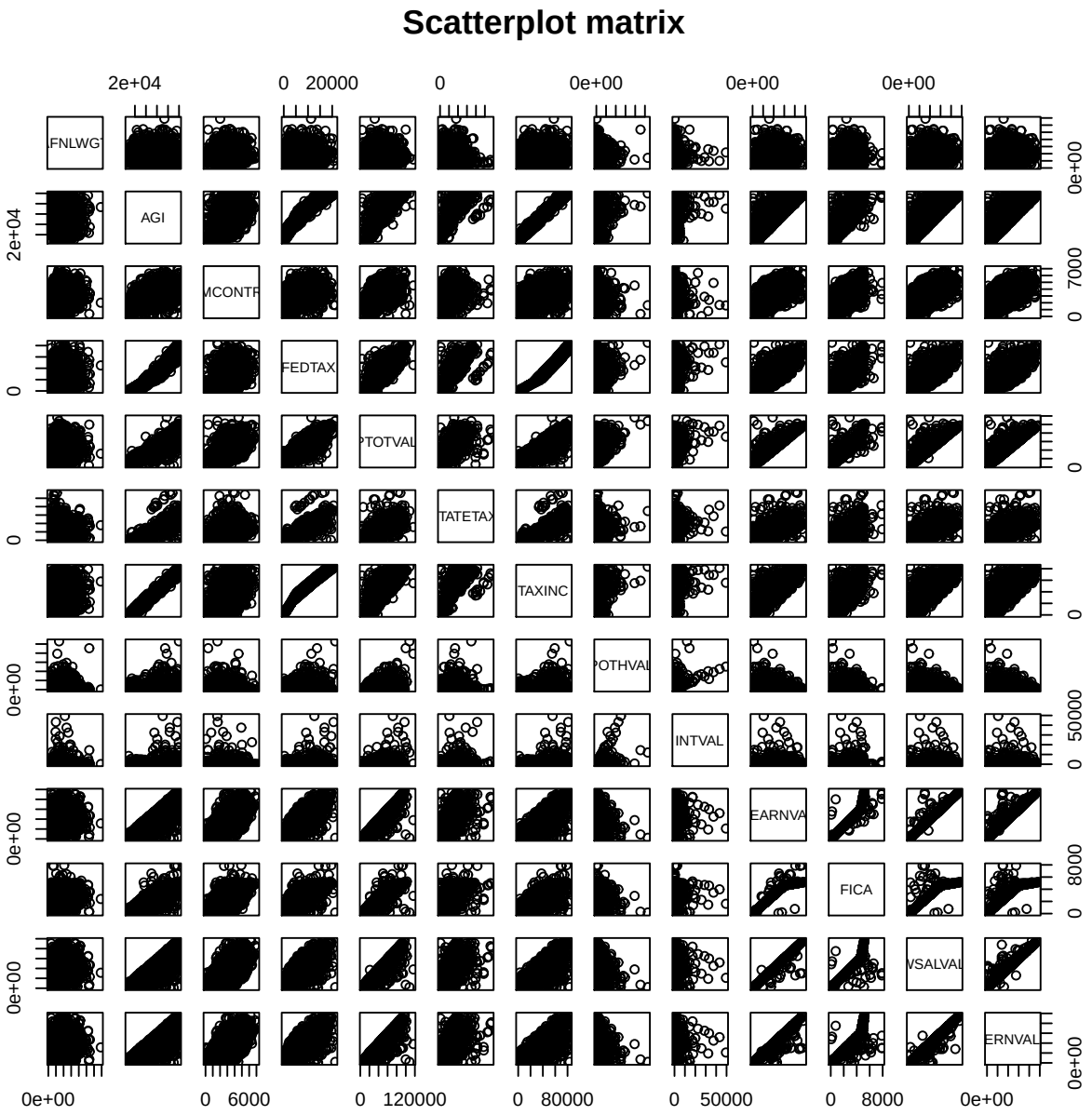
The file

- Read the file `CASCrefmicrodata` (from sdcMicro) and plot scatterplot

```
> data(CASCrefmicrodata)
> colnames(CASCrefmicrodata)
 [1] "AFNLWGT"  "AGI"      "EMCONTRB" "FEDTAX"   "PTOTVAL"  "STATETAX"
 [7] "TAXINC"   "POTHVAL"  "INTVAL"   "PEARNVAL" "FICA"     "WSALVAL"
[13] "ERNVAL"
> pairs(~AFNLWGT+AGI+EMCONTRB+FEDTAX+PTOTVAL+STATETAX+
        TAXINC+POTHVAL+INTVAL+PEARNVAL+FICA+WSALVAL+ERNVAL,
        data=CASCrefmicrodata,main="Scatterplot matrix")
```

- I will use AGI and TAXINC.

The file: Scatterplots (observe: AGI vs. TAXINC)



Masking: rank swapping

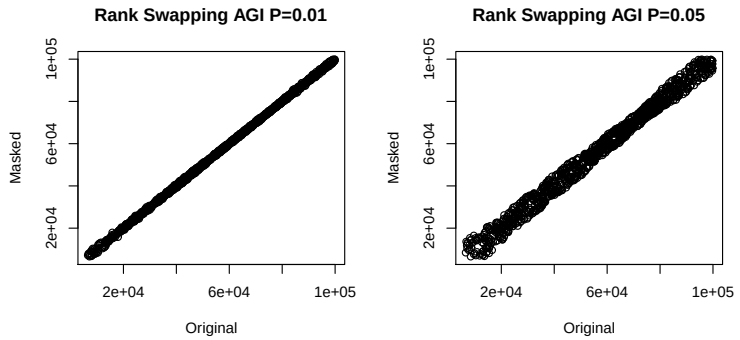
- We mask the data set using rank swapping with $p = 0.01, 0.05, 0.1, 0.2, 0.3$

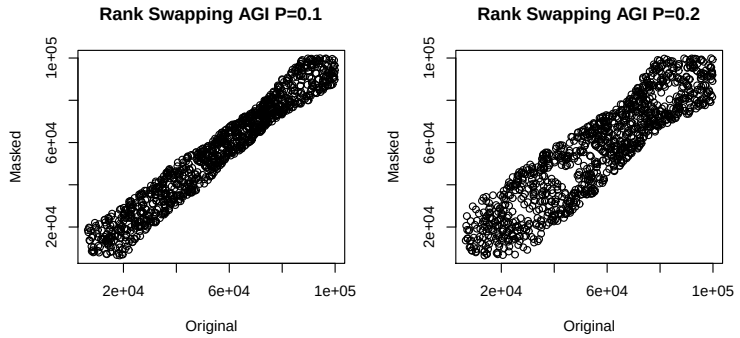
```
casc.rs0.01<-rankSwap(CASCrefmicrodata,
                      variables=colnames(CASCrefmicrodata),
                      TopPercent=0,BottomPercent=0,K0=-1,R0=-1,P=0.01)
```

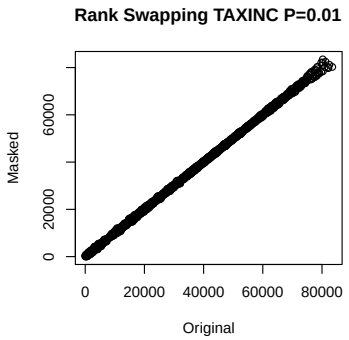
By default top and bottom coding, modification to increase correlation

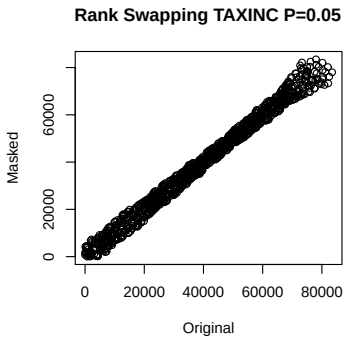
Masking: rank swapping. Scatterplots

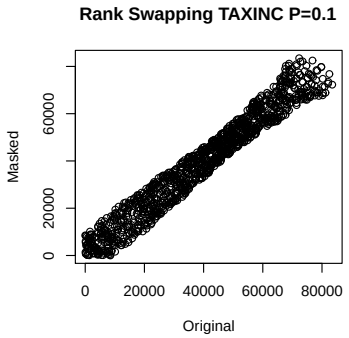
- Variables AGI and TAXINC original vs. masked ($p = 0.01, 0.05, 0.1, 0.2$)

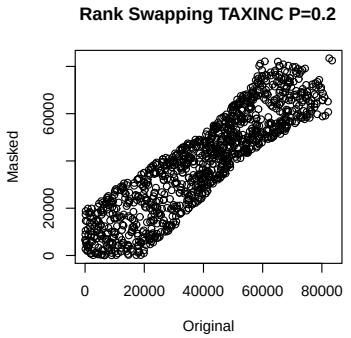






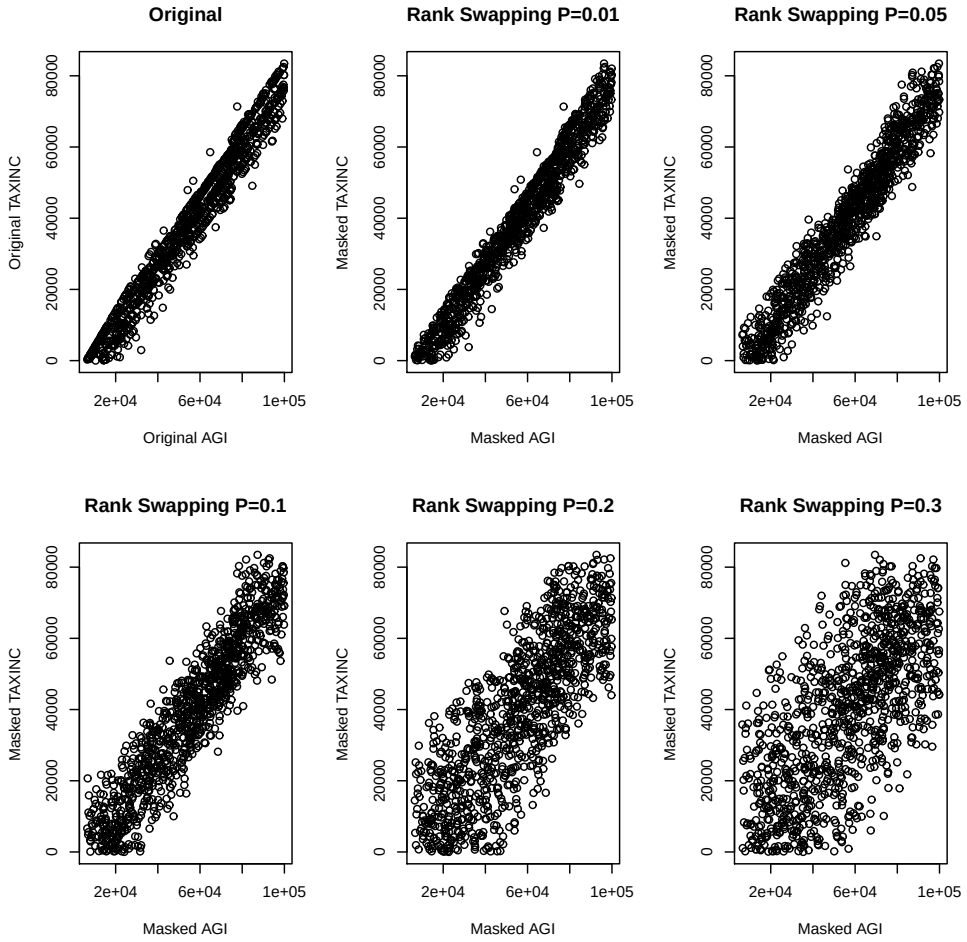






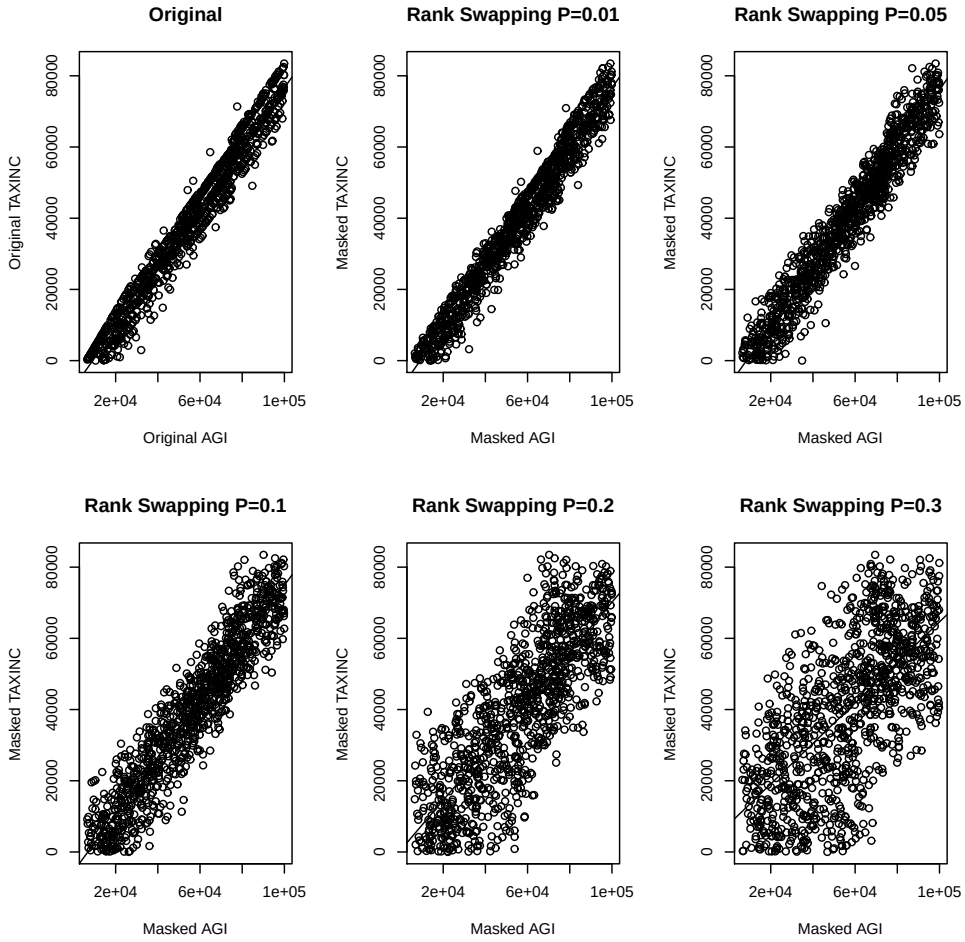
Masking: rank swapping. Scatterplots

- AGI vs. TAXINC ($p = 0$ (original) and $p = 0.01, 0.05, 0.1, 0.2, 0.3$)



Masking: rank swapping. Scatterplots

- Regression AGI/TAXINC (original and $p = 0.01, 0.05, 0.1, 0.2, 0.3$)



Masking: microaggregation.

- We mask the data set using microaggregation with $k = 3, 4, 5, 6, 7, 10, 15$ (all variables) with MDAV

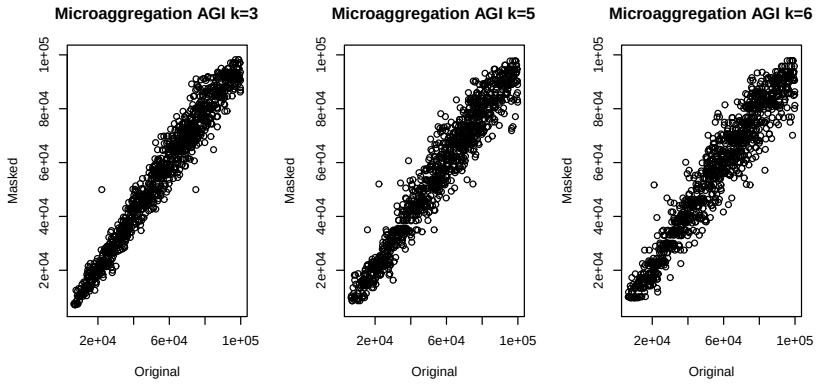
```
casc.ma.3<-microaggregation(casc, variables=cascV,
                            aggr=3, method="mdav")
```

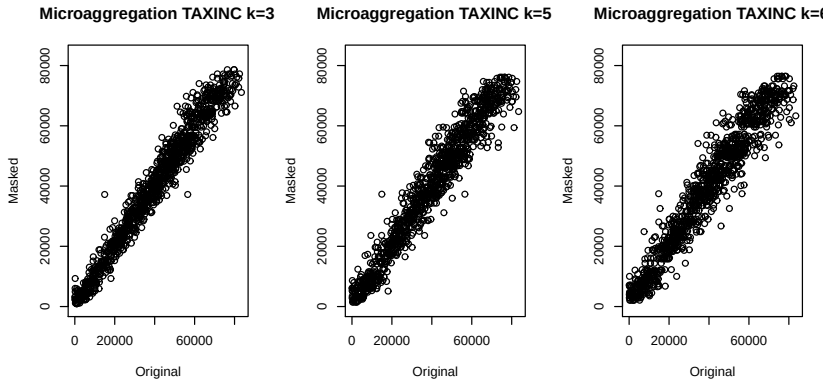
- Observe: when $k = 3$, we have sets of 3 indistinguishable records

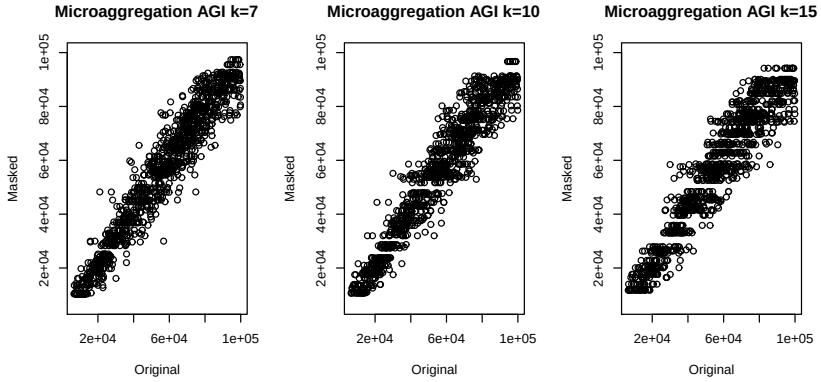
```
> casc.ma.3$mx[casc.ma.3$mx$AGI==10554.000,]
     AFNLWGT   AGI EMCONTRB   FEDTAX  PTOTVAL STATETAX TAXINC ...
196   195905 10554 2252.667 412.6667 12437.67      156   2756 ...
1054  195905 10554 2252.667 412.6667 12437.67      156   2756 ...
1065  195905 10554 2252.667 412.6667 12437.67      156   2756 ...
```

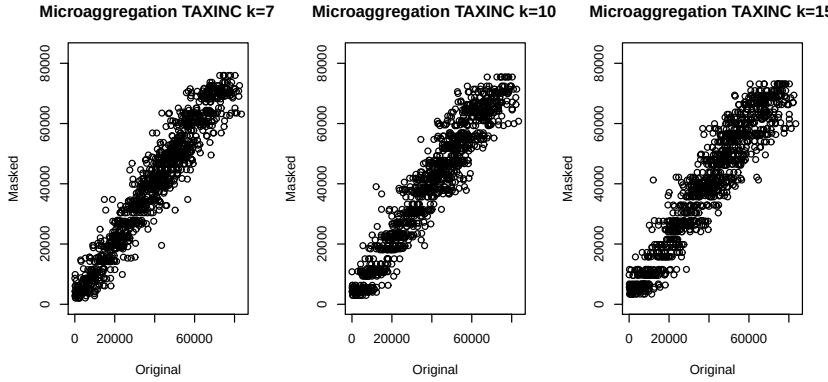
Masking: microaggregation. Scatterplots.

- Variables AGI and TAXINC original vs. masked ($p = 0.01, 0.05, 0.1, 0.2$)









Masking: microaggregation.

- We mask the data set using microaggregation with $k = 3, 4, 5, 6, 7, 10, 15$ (all variables) with individual ranking

```
casc.ma.od.3<-microaggregation(casc, variables=cascV,
                               aggr=3, method="onedims")
```

- Observe: when $k = 3$, no sets of 3 indistinguishable records

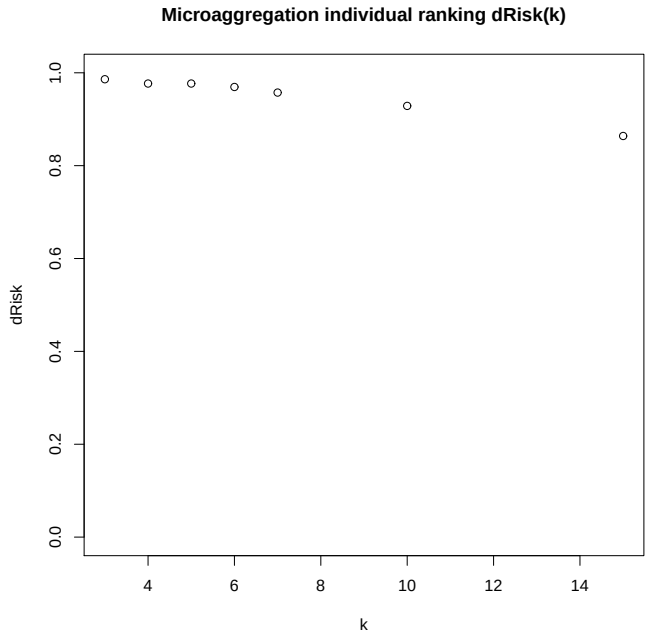
```
> casc.ma.od.3$mx[casc.ma.od.3$mx$AGI==57563.00,]
     AFNLWGT   AGI EMCONTRB   FEDTAX  PTOTVAL STATETAX   TAXINC ...
2   251340.0 57563 2633.333 6082.333 42080.00 1904.000 39263.67 ...
112 188836.7 57563 1563.667 5247.333 30871.33 3208.333 34969.67 ...
253 386669.3 57563 2403.000 7396.667 48840.33 3552.333 44052.00 ...
```

Masking: Disclosure risk.

- Risk using `dRisk` for rank swapping with $p = 0.01$.

```
> dRisk(obj=casc, xm=casc.rs0.01)
[1] 0.7601852
```

- Risk for microaggregation with MDAV is 0
- Risk for microaggregation with individual ranking:

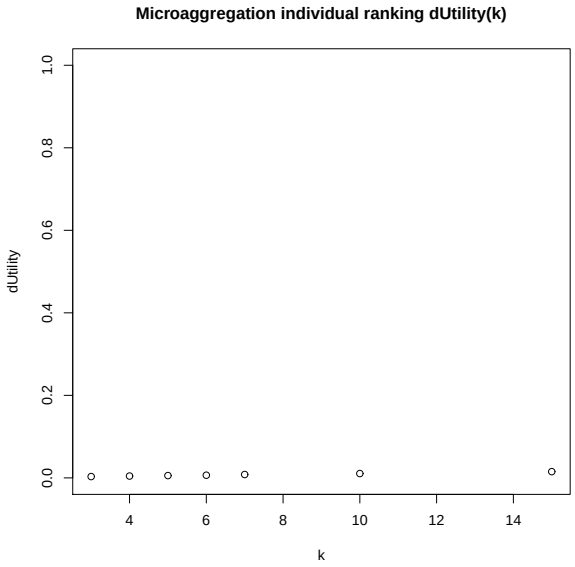


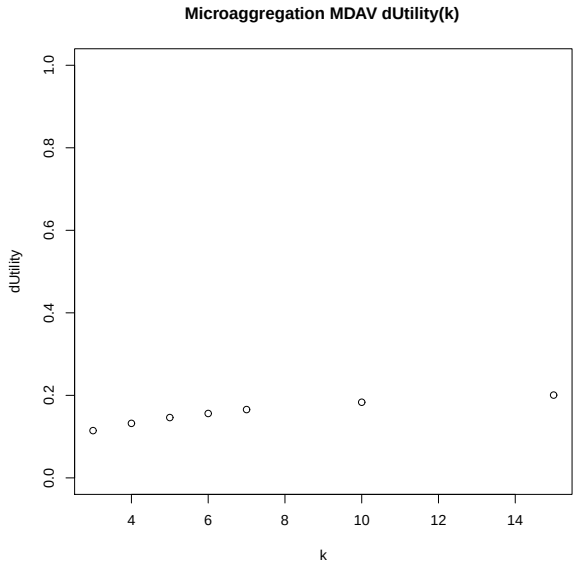
Masking: Information loss.

- Utility/information loss using `dUtility` for microaggregation. (utility in terms of the distance original vs. masked records)

```
> dUtility(obj=casc, xm=casc.rs0.01)
[1] 0.02068808
```

- Individual ranking (left) and MDAV (right)

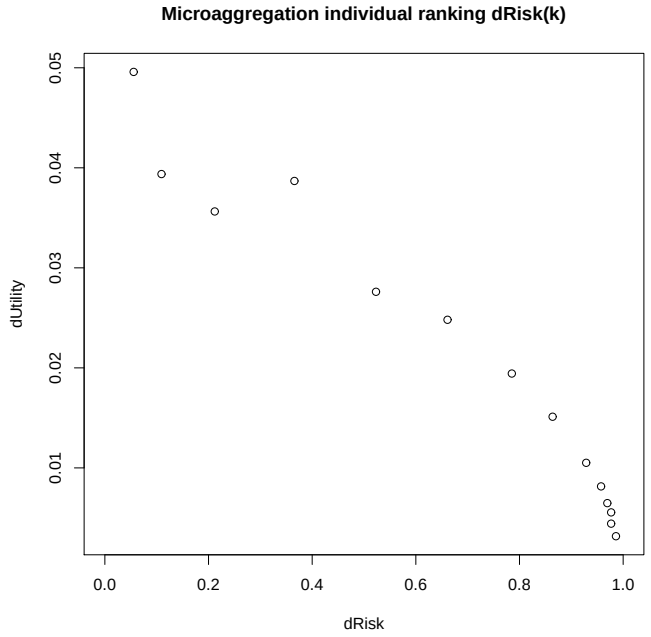




- Note: individual ranking less information loss, larger risk.

Masking: Information loss.

- R-U map (or IL vs. Risk)



(risk as abscissa, as in previous figures k and p as abscissa but in R-U map Utility in abscissa)

Masking: Methods in sdcMicro.

- Methods
  - noise addition (`addNoise`)
  - recoding (globalRecode, groupAndRename)
  - suppression
  - microaggregation
  - pram
  - rank swapping (`rankSwap`)
  - shuffle
  - top and bottom coding (`topBotCoding`)

sdcTable

The problem

- Aggregates of data with respect to a few variables. Ex. (Castro, 2012)

	P_1	P_2	P_3	P_4	P_5	Total
M_1	2	15	30	20	10	77
M_2	72	20	1	30	10	133
M_3	38	38	15	40	5	136
TOTAL	112	73	46	90	25	346

Cell (M_2, P_3): number of people with profession P_3 living in municipality M_2.

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

Cell (M_2, P_3): total salary received by people with profession P_3 living in M_2.

The problem

- File `ex1.rec` with records with professions, municipalities, and salaries (that lead to the tables).
- 1. Define level structure (hierarchical structure) for each variable.

```
bildLevels <- function (values) {
    rbind(c("@","Total"),
          t(mapply(function(v) { c("@@",v) }, values)))}
ex1.t.lprofs <- as.data.frame(bildLevels(ex1.profs),
                              stringsAsFactors=FALSE)
> ex1.t.lprofs
   V1    V2
    @ Total
P1 @@    P1
P2 @@    P2
P3 @@    P3
P4 @@    P4
P5 @@    P5
```

The problem

- 2. Creation of the problem

```
ex1.t.dim <- list(Munic=ex1.t.lmunic,Profs=ex1.t.lprofs)
ex1.t.vIn <- c(2,3)   ## in which column we find Munic and Profs in ex
ex1.t.prob <- makeProblem(
    data = ex1.rec,
    dimList = ex1.t.dim,
    dimVarInd = ex1.t.vIn,
    freqVarInd=NULL, ## if not NULL, column with cell counts
    numVarInd=c(4), ## if not NULL, columns with numerical variables:
    weightInd=NULL, ## if not NULL, column with with weights for secon
    sampWeightInd=NULL ## if not NULL, column with sampling weights fo
)
```

The problem

- 3. Mark the cells to be suppressed:

```
ex1.t.prob <- primarySuppression(ex1.t.prob, type='freq', maxN=3)
```

- 4. Suppression (primary and secondary suppression) (HITAS, OPT, HYPER) three methods

```
ex1.t.resHITAS <- protectTable(ex1.t.prob, method="HITAS")
(...)
> reshape2::dcast(tabHITAS, tabOPT$Munic ~ tabOPT$Profs,
                  value.var=c("newC"))
  tabOPT$Munic   P1   P2   P3   P4   P5 Total
1           M1   --  450  720  400   --  2290
2           M2 1140  540   --  570   --  2592
3           M3   -- 1178   --  800   --  3438
4        Total 2222 2168 1117 1770 1043  8320
```

Thank you

References

- T. Benschop, C. Machingauta, M. Welch, Statistical disclosure control for microdata: A practical guide, 2016.
- M. Templ, Statistical disclosure control for microdata: Methods and applications in R, Springer, 2017.
- V. Torra, Data Privacy: Foundations, New Developments and the Big Data Challenge, Springer, 2017.
- J. Castro, Recent advances in optimization techniques for statistical tabular data protection, European Journal of Operational Research 216 (2012) 257-269.