TALL SOM FORTELLER

GaussSuppression: An R package for Tabular Data Suppression

Øyvind Langsrud, Daniel Lupp and Hege Marie Bøvelstad

https://cran.r-project.org/package=GaussSuppression
https://github.com/statisticsnorway/GaussSuppression

STATISTICS NORWAY

USER GROUP SDC WORKSHOP IN PARIS

20TH SEPTEMBER 2022

Statistisk sentralbyrå
Statistics Norway

TALL
SOM FORTELLER

GaussSuppression: An R package for Tabular Data Suppression

Contents

- History – Why a new suppression package?
- Introduction to the GaussSuppression package by examples
- Real application example
- K-disclosure suppression

Statistisk sentralbyrå
Statistics Norway

History – interface

- R package easySdcTable in 2016
  - To meet requirements of an IT-system.
- R package SmallCountRounding in 2018
  - Formula interface to define tables → tables created via a model matrix
- Hierarchical computations (not SDC) in 2018
  - To solve municipal accounts calculation problems
  - A spin-off is that model matrices can be made from usual sdc hierarchies
- General function `ModelMatrix` in the R package SSBtools in 2021
  - Formula interface interface and hierarchy interface combined into a single function
  - Should not belong to a specific purpose package

History – secondary suppression

- Curious to investigate secondary suppression based on classical linear algebra
  - Due to related work
  - Gaussian elimination seemed a very promising method
  - A fast method and better results than "SIMPLEHEURISTIC“ in sdcTable
  - But not a quality competitor to optimality-based methods
- Gauss suppression extra feature in easySdcTable in 2020
  - Made default in 2021
  - All use of “SIMPLEHEURISTIC” within the IT-system was changed to “Gauss“

History → New suppression package

- Main ingredients of a suppression package
  1) Interface to define and build tables from input data and hierarchies
  2) Interface to specify primary suppression
  3) A secondary suppression algorithm

- 1) and 3) already available
  - `ModelMatrix` and `GaussSuppression` fit perfectly together

- 2) can be specified by a user-defined function
  - That is, the primary suppression function is an input parameter

Ready made primary functions

- `PrimaryDefault` – frequency rule
- `DominanceRule` – application later in this talk
- `NcontributorsRule`
- `KDisclosurePrimary` – details later in this talk

```
PrimaryDefault <- function(freq, maxN = 3, protectZeros = TRUE, ...) {

  if(is.null(maxN))         stop("A non-NULL value of maxN is required.")
  if(is.null(protectZeros)) stop("A non-NULL value of protectZeros is required.")

  primary <- freq <= maxN
  if (!protectZeros)
    primary[freq == 0] <- FALSE

  primary
}
```

```
> z
    age      geo    eu freq
1 young    Spain    EU    7
2 young  Iceland nonEU    0
3 young Portugal    EU    2
4   old    Spain    EU   10
5   old  Iceland nonEU    1
6   old Portugal    EU    4

> GaussSuppressionFromData(data = z,
      dimVar = c("age", "geo", "eu"),
      freqVar = "freq", maxN = 2)
     age      geo freq primary suppressed
1  Total    Total   24       -          -
2  Total       EU   23       -       TRUE
3  Total    nonEU    1    TRUE       TRUE
4  Total  Iceland    1    TRUE       TRUE
5  Total Portugal    6       -       TRUE
6  Total    Spain   17       -          -
7    old    Total   15       -          -
8    old       EU   14       -       TRUE
9    old    nonEU    1    TRUE       TRUE
10   old  Iceland    1    TRUE       TRUE
11   old Portugal    4       -       TRUE
12   old    Spain   10       -          -
13 young    Total    9       -          -
14 young       EU    9       -       TRUE
15 young    nonEU    0    TRUE       TRUE
16 young  Iceland    0    TRUE       TRUE
17 young Portugal    2    TRUE       TRUE
18 young    Spain    7       -          -
```

Frequency rule suppression using `dimVar`

- `GaussSuppressionFromData`
  - The main function in the package
  - `ModelMatrix` used inside
- `ModelMatrix`: Three possibilities
  - `hierarchies`
  - `formula`
  - `dimVar` – automatic
- Frequency rule here: `freq` ≤ 2

"FALSE" changed to "-" in this presentation

Statistisk sentralbyrå
Statistics Norway

Hierarchies generated automatically

- in the background when `dimVar`

```
> z
    age      geo    eu freq
1 young    Spain    EU    7
2 young  Iceland nonEU    0
3 young Portugal    EU    2
4   old    Spain    EU   10
5   old  Iceland nonEU    1
6   old Portugal    EU    4

> GaussSuppressionFromData(data = z,
    dimVar = c("age", "geo", "eu"),
    freqVar = "freq", maxN = 2)
     age      geo freq primary suppressed
1  Total    Total   24       -          -
2  Total       EU   23       -       TRUE
3  Total    nonEU    1    TRUE       TRUE
4  Total  Iceland    1    TRUE       TRUE
5  Total Portugal    6       -       TRUE
6  Total    Spain   17       -          -
7    old    Total   15       -          -
8    old       EU   14       -       TRUE
9    old    nonEU    1    TRUE       TRUE
10   old  Iceland    1    TRUE       TRUE
11   old Portugal    4       -       TRUE
12   old    Spain   10       -          -
13 young    Total    9       -          -
14 young       EU    9       -       TRUE
15 young    nonEU    0    TRUE       TRUE
16 young  Iceland    0    TRUE       TRUE
17 young Portugal    2    TRUE       TRUE
18 young    Spain    7       -          -
```

```
> dimlists <- FindDimLists(z[c("age", "geo", "eu")])
> dimlists
$age
  levels codes
1      @ Total
2     @@   old
3     @@ young

$geo
  levels    codes
1      @    Total
2     @@       EU
3    @@@ Portugal
4    @@@    Spain
5     @@    nonEU
6    @@@  Iceland
```

Function in `SSBtools` originally made for `easySdcTable`

Statistisk sentralbyrå
Statistics Norway

Hierarchies can be input

```
> z
    age      geo    eu freq
1 young    Spain    EU    7
2 young  Iceland nonEU    0
3 young Portugal    EU    2
4   old    Spain    EU   10
5   old  Iceland nonEU    1
6   old Portugal    EU    4
```

```
> GaussSuppressionFromData(data = z,
    hierarchies = dimlists,
    freqVar = "freq", maxN = 2)
     age      geo freq primary suppressed
1  Total    Total   24       -          -
2  Total       EU   23       -       TRUE
3  Total    nonEU    1    TRUE       TRUE
4  Total  Iceland    1    TRUE       TRUE
5  Total Portugal    6       -       TRUE
6  Total    Spain   17       -          -
7    old    Total   15       -          -
8    old       EU   14       -       TRUE
9    old    nonEU    1    TRUE       TRUE
10   old  Iceland    1    TRUE       TRUE
11   old Portugal    4       -       TRUE
12   old    Spain   10       -          -
13 young    Total    9       -          -
14 young       EU    9       -       TRUE
15 young    nonEU    0    TRUE       TRUE
16 young  Iceland    0    TRUE       TRUE
17 young Portugal    2    TRUE       TRUE
18 young    Spain    7       -          -
```

```
> dimlists <- FindDimLists(z[c("age", "geo", "eu")])
> dimlists
$age
  levels codes
1      @ Total
2     @@   old
3     @@ young

$geo
  levels    codes
1      @    Total
2     @@       EU
3    @@@ Portugal
4    @@@    Spain
5     @@    nonEU
6    @@@  Iceland
```

Bigger data set

```
> x
     age       geo    eu year freq
1  young     Spain    EU 2014    7
2  young   Iceland nonEU 2014    0
3  young Portugal    EU 2014    2
4    old     Spain    EU 2014   10
5    old   Iceland nonEU 2014    1
6    old Portugal    EU 2014    4
7  young     Spain    EU 2015    9
8  young   Iceland nonEU 2015    2
9  young Portugal    EU 2015    5
10   old     Spain    EU 2015   12
11   old   Iceland nonEU 2015    3
12   old Portugal    EU 2015    7
13 young     Spain    EU 2016   11
14 young   Iceland nonEU 2016    5
15 young Portugal    EU 2016    7
16   old     Spain    EU 2016   15
17   old   Iceland nonEU 2016    5
18   old Portugal    EU 2016    9
```

- `year` is extra dimensional variable

Formula interface

```
> GaussSuppressionFromData(data = x,
    formula = ~eu*year + age:geo,
    freqVar = "freq", maxN = 2)
```

	year	age	geo	freq	primary	suppressed
1	Total	Total	Total	114	-	-
2	Total	Total	EU	98	-	-
3	Total	Total	nonEU	16	-	-
4	2014	Total	Total	24	-	-
5	2015	Total	Total	38	-	-
6	2016	Total	Total	52	-	-
7	2014	Total	EU	23	-	TRUE
8	2015	Total	EU	33	-	TRUE
9	2016	Total	EU	42	-	-
10	2014	Total	nonEU	1	TRUE	TRUE
11	2015	Total	nonEU	5	-	TRUE
12	2016	Total	nonEU	10	-	-
13	Total	old	Iceland	9	-	-
14	Total	old	Portugal	20	-	-
15	Total	old	Spain	37	-	-
16	Total	young	Iceland	7	-	-
17	Total	young	Portugal	14	-	-
18	Total	young	Spain	27	-	-

- Output table defined by a formula
- Hierarchical relations treated automatically
  - so that **eu** and **geo** are in the same column in the output

A small example again

```
> z
    age      geo    eu freq
1 young    Spain    EU    7
2 young  Iceland nonEU    0
3 young Portugal    EU    2
4   old    Spain    EU   10
5   old  Iceland nonEU    1
6   old Portugal    EU    4

> GaussSuppressionFromData(data = z,
      formula = ~geo + age,
      freqVar = "freq", maxN = 2)

       geo   age freq primary suppressed
1    Total Total   24       -          -
2  Iceland Total    1    TRUE       TRUE
3 Portugal Total    6       -       TRUE
4    Spain Total   17       -          -
5    Total   old   15       -          -
6    Total young    9       -          -
```

- The model matrix made internally

```
Total-Total Iceland-Total Portugal-Total Spain-Total Total-old Total-young
          1             .              .           1         .           1
          1             1              .           .         .           1
          1             .              .           .         1           .
          1             .              1           1         1           .
          1             .              .           1         1           .
          1             1              .           .         .           .
```

Statistisk sentralbyrå
Statistics Norway

Total-Total	Iceland-Total	Portugal-Total	Spain-Total	Total-old	Total-young
1	.	.	1	.	1
1	1	.	.	.	1
1	.	1	.	.	1
1	.	.	1	1	.
1	1	.	.	1	.
1	.	1	.	1	.

Candidates function

- determines priority order in the sequential algorithm

Primary function

Suppressed

Total-Total	Spain-Total	Total-old	Total-young	Portugal-Total
1	1	.	1	.
1	.	.	1	.
1	.	.	1	1
1	1	1	.	.
1	.	1	.	.
1	.	1	.	1

Iceland-Total
.
1
.
.
1
.

Statistisk sentralbyrå
Statistics Norway

Functions as parameters
- User defined or ready made

- `candidates` – determines priority order
- `primary` – primary suppression
- `forced` – cells forced to be not suppressed
- `hidden` – cells not to be published
- `singleton` – to handle problem of singletons or zeros

Specifying hierarchies

- can be done in several ways

These are more general.

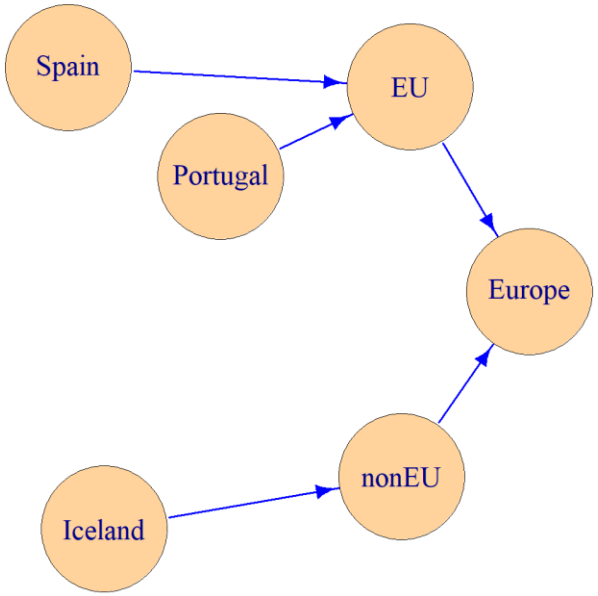
- A tree structure not needed.
- Sign can be negative

formulas

Europe = EU + nonEU

EU = Portugal + Spain

nonEU = Iceland



sdcTable

levels	codes
@	Europe
@@	EU
@@@	Portugal
@@@	Spain
@@	nonEU
@@@	Iceland

tauArgus

EU

@Portugal

@Spain

nonEU

@Iceland

This is the internal standard

mapsFrom	mapsTo	sign
EU	Europe	1
Portugal	EU	1
Spain	EU	1
nonEU	Europe	1
Iceland	nonEU	1

Benefits when there are linked tables

- Several tree-shaped hierarchies can be combined as a single hierarchy
- Formula interface is an easy way to specify several tables
- Either way, a single model matrix is created and the algorithm is the same

Microdata can be input

- Automatic aggregation to the appropriate level

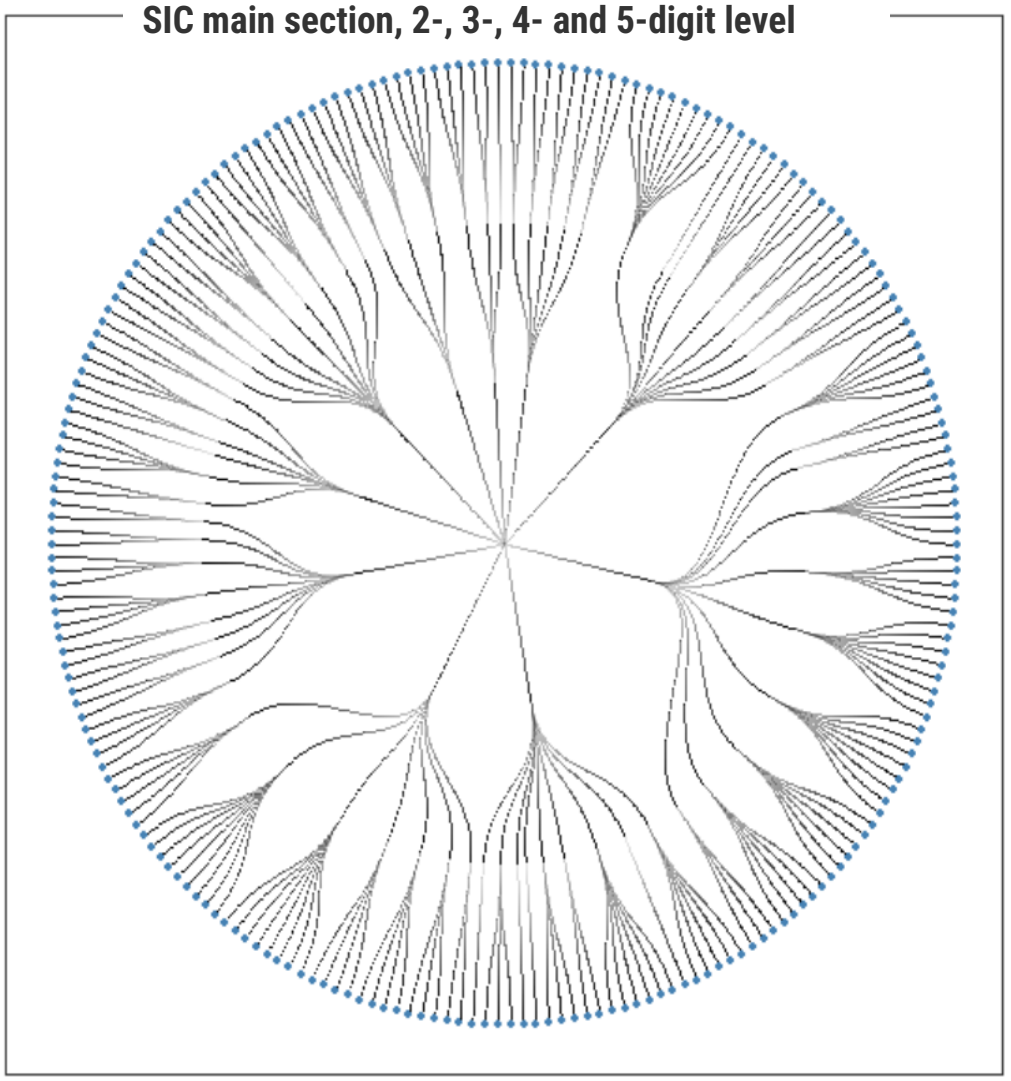
Real application example

Application to business statistics

- Information on the activity in the Norwegian business sector
- Register data and data from surveys
- Approximately 320.000 enterprises and 350.000 establishments

Problems to be handeled

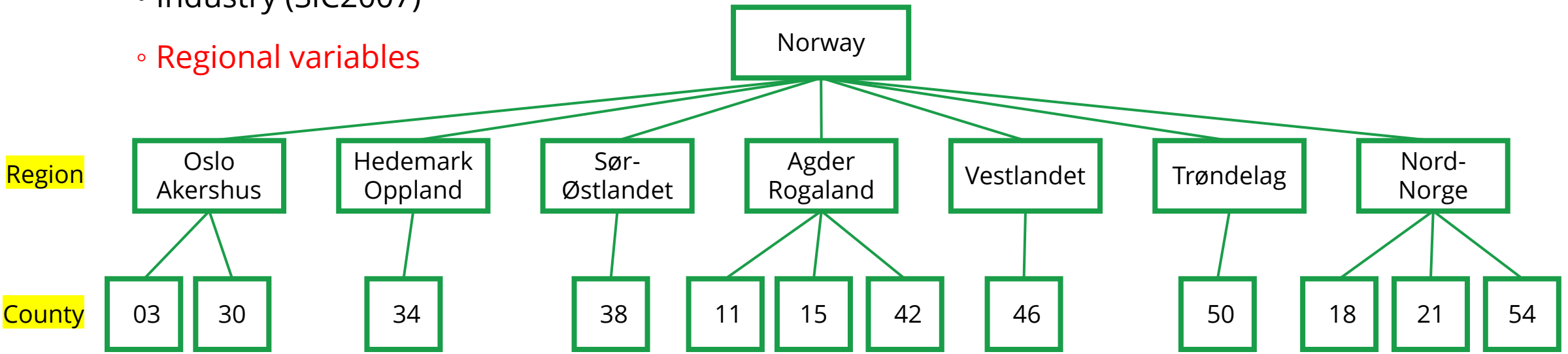
- 3 linked tables
- Hierarchies
  - Industry (SIC2007)



k sentralbyrå
lorway

Problems to be handeled

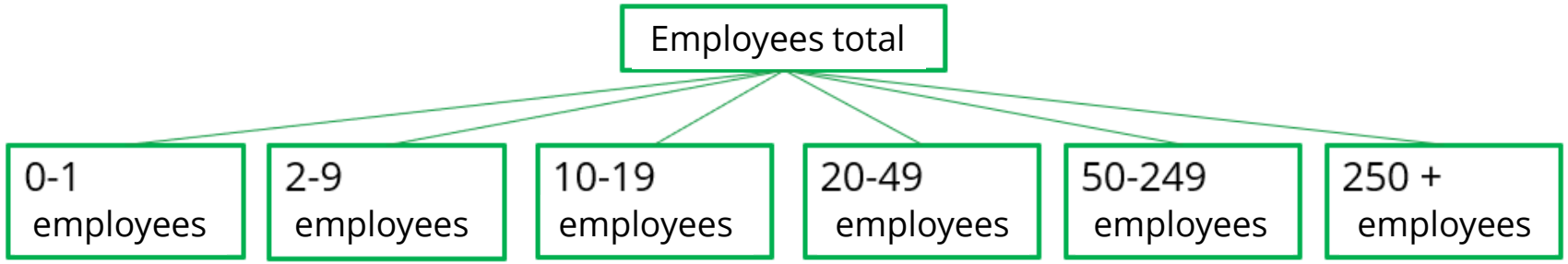
- 3 linked tables
- Hierarchies
  - Industry (SIC2007)
  - Regional variables



Statistisk sentralbyrå
Statistics Norway

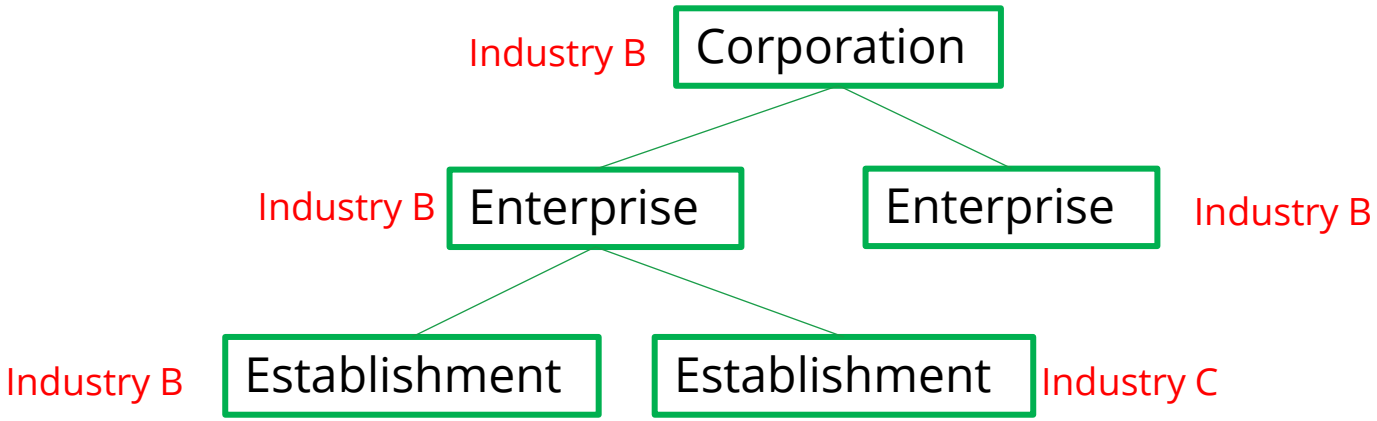
Problems to be handeled

- 3 linked tables
- Hierarchies
  - Industry (SIC2007)
  - Regional variables
  - Employment groups



Problems to be handeled

- 3 linked tables
- Hierarchies
  - Industry (SIC2007)
  - Regional variables
  - Employment groups
  - Holding information



Publication

Table 1:

Turnover by industry

(SIC2007 main section, 2-, 3-, 4- and 5-digit level)

Variables:

Turnover

Wages and salaries

Value added

Gross investments

⋮

Publication

Table 1:

Turnover by industry

(SIC2007 main section, 2-, 3-, 4- and 5-digit level)

```
formula =
        ~ (sic.main + sic.2digits + sic.3digits + sic.4digits + sic.5digits)
```

Table 2:

Turnover by number of persons employed and industry
(SIC2007 main section, 2- and 3-digit level)

```
formula = ~ (sic.main + sic.2digits + sic.3digits)
```

Publication

LINKED TABLES

Table 1:

Turnover by industry

(SIC2007 main section, 2-, 3-, 4- and 5-digit level)

```
formula =
        ~ (sic.main + sic.2digits + sic.3digits + sic.4digits + sic.5digits)
```

Table 2:

Turnover by number of persons employed and industry (SIC2007 main section, 2- and 3-digit level)

```
formula = ~(sic.main + sic.2digits + sic.3digits) * employment.group
```

Table 3:

Turnover by county, region and industry

(SIC2007 main section, 2- and 3-digit level)

```
formula = ~(sic.main + sic.2digits + sic.3digits) * (region + county)
```

Output for turnover

```
GaussSuppressionFromData(
  data = businessDat2020,
formula = ~ (sic.main + sic.2digits + sic.3digits + sic.4digits + sic.5digits) +
            (sic.main + sic.2digits + sic.3digits) * employment.group +
            (sic.main + sic.2digits + sic.3digits) * (region + county),
numVar = "turnover",
primary = DominanceRule,
n = c(1,2),
p = c(80,90),
charVar = "corporation",
protectZeros = TRUE,
…)
```

May also be a vector of variables

Keeps track of holding information

Specify primary suppression rule and parameters

Hierarchies are automatically generated from data

	naring	sys.gruppe	fylke	freq	omsetning	primary	suppressed
Total-Total-Total	Total	Total	Total	344450	5024582343	FALSE	FALSE
B-Total-Total	B	Total	Total	1357	139065668	FALSE	FALSE
C-Total-Total	C	Total	Total	19012	898165256	FALSE	FALSE
E-Total-Total	E	Total	Total	1541	32476263	FALSE	FALSE
F-Total-Total	F	Total	Total	58826	631619635	FALSE	FALSE
G-Total-Total	G	Total	Total	64965	1846090026	FALSE	FALSE
H-Total-Total	H	Total	Total	21303	423343703	FALSE	FALSE
I-Total-Total	I	Total	Total	15674	96402326	FALSE	FALSE
J-Total-Total	J	Total	Total	19185	272120976	FALSE	FALSE
L-Total-Total	L	Total	Total	52341	175952307	FALSE	FALSE
M-Total-Total	M	Total	Total	52184	283475855	FALSE	FALSE
N-Total-Total	N	Total	Total	22830	205280241	FALSE	FALSE
S-Total-Total	S	Total	Total	15232	20590087	FALSE	FALSE

Subset of output (approximately 9400 cells in total)

Run time

- Data: 350.000 establishments
- Suppressions

Table		Total suppressions	Primary	Secondary
1	SIC (1-5 digits)	205	140	65
2	SIC (1-3 digits) x employment groups	448	256	192
3	SIC (1-3 digits) x region and county	2101	1435	666

- ~ 30 minutes run time

TALL | SOM FORTELLER

K-Disclosure Suppression

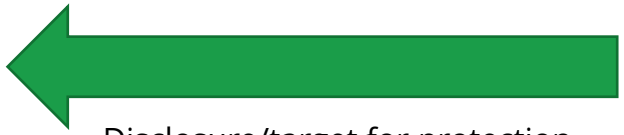
`GaussSuppression::SuppressKDisclosure`

Disclosure in frequency tables

Premise:

Protect disclosure of unit's cell membership, not disclosure of cell value

«Bob was seriously injured»



Disclosure/target for protection

«3 people were seriously injured»

Not disclosure

Small count primary

Level of injury in traffic accidents by city

	None	Light	Serious	Total
Paris	0	0	250	250
Oslo	0	0	5	5
Bergen	0	0	2	2

- With this notion of disclosure, these rows are virtually indistinguishable with respect to cell membership disclosure
- All units in Bergen/Oslo/Paris were seriously injured.
  - Why is Bergen the only one to be protected? ....and is it protected?
- Protection levels and intervals in CSP do not address the issue!
- Has been proposed that zeros must be protected as well...
- This is a heuristic focused on cell values, not targeted at disclosure

K-Disclosure

- Frequency table parallel to p% rule for volume tables:

Assume an attacker has knowledge of certain statistical units, can they disclose another unit's contribution to the table?

- Here: a unit's contribution is membership in (group of) cells
- We assume an attacker has knowledge of up to k statistical units

K-Disclosure

Assume an attacker has knowledge of k statistical units, can they disclose another unit's membership in cells of the table?

- One can show: answer is YES if and only if certain *differences* are less than k
- It is not *cell values* that are sensitive, but certain *cell differences*
- Table cells must be suppressed to prevent accurate recalculation of these differences

K-Disclosure

- k=0: attacker has no knowledge about any statistical units

	None	Light	Serious	Total
Paris	0	0	250	250
Oslo	0	0	5	5
Bergen	0	0	2	2

- k=1: knowledge of up to one statistical unit

	None	Light	Serious	Total
Paris	0	1	250	251
Oslo	1	0	50	51

- k=2: knowledge of up to two statistical units

	None	Light	Serious	Total
Paris	1	1	250	252
Oslo	2	0	50	52

- ...

K-Disclosure Suppression (k = 0)

	None	Light	Serious	Total
Paris	X	0	X	250
Prague	24	10	15	49
Berlin	0	2	5	7
Oslo	X	0	X	5
Bergen	X	0	X	2

Small difference

Small difference

Small difference

Suppress these...

...to protect these

Model matrix

```
 [1,] 1 . . . 1 . . . . . . . . . . . 1 . . . .
 [2,] 1 . . . 1 . . . . . . . . . . . 1 . . . .
 [3,] 1 . . . 1 . . . . . . . . . . . . 1 . . .
 [4,] 1 . . . . 1 . . . . . . . . . . . . . 1 .
 [5,] 1 . . . . 1 . . . . . . . . . . . . . 1 . .
 [6,] 1 . . . . 1 . . . . . . . . . . . . . . 1
 [7,] 1 . 1 . . . . . . . 1 . . . . . . . . . .
 [8,] 1 . 1 . . . . . . . 1 . . . . . . . . . .
 [9,] 1 . 1 . . . . . . . . 1 . . . . . . . . .
[10,] 1 . . 1 . . . . . . . . . 1 . . . . . . .
[11,] 1 . . 1 . . . . . . . . . 1 . . . . . . .
[12,] 1 . . 1 . . . . . . . . . . 1 . . . . . .
[13,] 1 1 . . . . . 1 . . . . . . . . . . . . .
[14,] 1 1 . . . . 1 . . . . . . . . . . . . . .
[15,] 1 1 . . . . . . 1 . . . . . . . . . . . .
```

Primary difference cells

```
 [1,] .
 [2,] .
 [3,] .
 [4,] .
 [5,] .
 [6,] .
 [7,] .
 [8,] .
 [9,] .
[10,] .
[11,] .
[12,] .
[13,] 1
[14,] 1
[15,] .
```

Statistisk sentralbyrå
Statistics Norway

…vs small count primary

Wrapper of `GaussSuppressionFromData` with `primary = KDisclosurePrimary`

```
SuppressKDisclosure(d,
        formula = ~ city + city:inj,
        freqVar = "freq",
        k = 0)
```

	None	Light	Serious	Total
Paris	X	0	X	250
Prague	24	10	15	49
Berlin	0	2	5	7
Oslo	X	0	X	5
Bergen	X	0	X	2

```
GaussSuppressionFromData(d,
                formula = ~ city + city:inj,
                freqVar = "freq",
                maxN =3, protectZeros = FALSE)
```

	None	Light	Serious	Total
Paris	0	0	250	250
Prague	24	10	15	49
Berlin	0	X	X	7
Oslo	0	0	X	X
Bergen	0	0	X	X

```
GaussSuppressionFromData(d,
                formula = ~ city + city:inj,
                freqVar = "freq",
                maxN =3, protectZeros = TRUE)
```

	None	Light	Serious	Total
Paris	X	X	X	250
Prague	24	10	15	49
Berlin	X	X	5	7
Oslo	X	X	5	X
Bergen	X	X	X	X

K-Disclosure

Assume an attacker has knowledge of k statistical units, can they disclose another unit's membership in cells of the table?

- Can adjust parameter k for more or less protection
- Flexible framework: one can protect not only against disclosure of single cell membership, but also *groups of cells*

All in Berlin are injured

	None	Light	Serious	Total
Paris	0	0	250	250
Prague	24	10	15	49
Berlin	0	2	5	7
Oslo	0	0	5	5
Bergen	0	2	0	2

Meaningful Combinations

- Can define any combination of categories that should be protected and include as parameter `mc_hierarchies`. Used in primary function, but not published.

```
 levels   codes
1      @ injured
2     @@ Serious
3     @@   Light
```

- Same as `GaussSuppression hierarchies` interface: can define any combination of categories without restriction
- Also a means of protecting against negative disclosure:

Protect against “unit is not a member of category A”

⇔

Protect against “unit is a member of combination of all other categories”

SuppressKDisclosure

- Package GaussSuppression well suited for this approach
  - Can define custom column vectors used as primary cells (need not be actual table cells)
  - Candidates for secondary suppression are customizable: only table cells can be suppressed, not difference cells
- Method under active development, working on implementation of more flexibility/features
  - Unknowns
  - More customizability with respect to what is considered sensitive/known

TALL SOM FORTELLER

Takk!

ssb.no

Statistisk sentralbyrå
Statistics Norway