Statistinio atskleidimo kontrolės vadovo
3 priedas

RIZIKŲ APSKAIČIAVIMO IR STATISTINIO ATSKLEIDIMO METODŲ TAIKYMO PAVYZDŽIAI NAUDOJANT *R* PAKETĄ

1. Statistinio atskleidimo kontrolės Vadovo V skyriuje aprašytas rizikas galima apskaičiuoti naudojant *R* programos *sdcMicro* paketą. Pateiktame pavyzdyje naudojami socialinės apklausos, turinčios hierarchinę struktūrą, duomenys.

2. Pirmas žingsnis – duomenų analizė. Duomenys nuskaitomi, pavyzdžiui, iš *csv* formato failo, naudojant *R* paketą *utils (1 kodas).*

1 kodas. Reikalingų paketų paruošimas naudoti ir duomenų nuskaitymas, asmenų ir kintamųjų skaičius, kintamųjų pavadinimai

```
# reikalingi paketai
library(sdcMicro) # sdcMicro paketas rizikoms skaičiuoti
library(utils) #nuskaityti csv failą

setwd("C:/SDC") # darbinis katalogas, kuriame yra duomenys
fname = "duomenys.csv" # duomenų failo pavadinimas
file <- read.csv(fname, header = TRUE, sep = ",", dec = ".") #nuskaitomi duomenys

dim(file) # Failo dimensijos (stebėjimai, kintamieji)
#[1] 11360  40
colnames(file) # Kintamųjų pavadinimai
# [1] "HH070"     "HH071"     "HH081"     "HH091"     "HY040G"    "HY050G"
#[7] "HY060G"    "HY070G"    "HY090G"    "HY120G"    "HY130G"    "HS090"
#[13] "HS110"     "HH_SIZE"   "MIE_KAIM"  "REGION"    "NUID"      "ETHNICITY"
#[19] "LANGUAGE"  "HY020"     "WGTPOP"    "SAVIVAL"   "HY17_1"    "RB220"
#[25] "RB230"     "RB240"     "AGE"       "PB190"     "PE010"     "PH020"
#[31] "WGTHH"     "ASMID"     "LYTIS"     "PE020"     "PL040"     "PL060"
#[37] "PL111"     "ACT_ST3"   "PE040B"    "ACT_ST1"
```

Nuskaitytame duomenų rinkinyje yra 11 360 asmenų, 5 131 namų ūkis ir 40 kintamųjų.

Norint apžvelgti kintamųjų reikšmes – kategoriniams kintamiesiems naudojamos dažnių lentelės, tolydiems kintamiesiems – suvestinė statistika. Norint įtraukti praleistų reikšmių skaičių (*NA* ar „.“), naudojama *useNA = "ifany"* funkcijoje *table()* (*2 kodas*).

2 kodas. Dažnių lentelė kintamajam „lytis“ ir statistika kintamajam „Namų ūkio pajamos“

```
# Dažnio lentelė kintamajam LYTIS (kategorinis)
table(file$LYTIS, useNA = "ifany")
#1       2
#5060 6300

# Statistika kintamajam HY020P_K (Metinės namų ūkio pajamos, tolydus)
summary(file$HY020)
#Min.     1st Qu.   Median   Mean      3rd Qu.    Max.
#-658.2   8347.7  14091.5   17322.9  22267.1   130450.9
```

1 lentelėje pateikiami duomenų rinkinio kintamieji, jų surinkimo lygis (asmuo (IND), namų ūkis (HH)), kintamojo tipas (kategorinis, pusiau tolydus, tolygus) ir reikšmių diapazonai. Daugelio kategorinių kintamųjų skaitinės reikšmės yra kodai, nurodantys reikšmes, pavyzdžiui, kintamasis *MIE_KAIM* gali įgyti reikšmes: 1 – žymi miestą ir 2 – žymi kaimą.

1 lentelė. Kintamųjų apžvalga

Nr.	Pavadinimas	Aprašymas	Lygis	Tipas	Reikšmės
1	NUID	Namų ūkio ID	HH	.	10000400–40300700
2	ASMID	Asmens ID	IND	.	1–13
3	REGION	Regionas	HH	kategorinis	1–2
4	SAVIVAL	Savivaldybė	HH	kategorinis	11–94

Nr.	Pavadinimas	Aprašymas	Lygis	Tipas	Reikšmės
5	MIE_KAIM	Miestas / kaimas	HH	kategorinis	1–2
6	WGTHH	Asmens svoris	HH	svoris	9.857–2810.328
7	WGTHH	Populiacijos svoris	IND	svoris	19.37–12207.74
8	HH_SIZE	Namų ūkio dydis	HH	kategorinis	1–13
9	LYTIS	Lytis	IND	kategorinis	1–2
10	PB190	Santuokinė padėtis	IND	kategorinis	1–5
11	AGE	Amžius	IND	Pusiau tolydus	-1–99
12	ETHNICITY	Namų ūkio galvos tautybė	HH	kategorinis	Visos reikšmės tuščios
13	LANGUAGE	Kalba, kuria kalba namų ūkio galva	HH	kategorinis	Visos reikšmės tuščios
14	RB230	Tėvo (patėvio, globėjo) ID	IND	kategorinis	1–12
15	RB240	Motinos (pamotės, globėjos) ID	IND	kategorinis	1–11
16	RB250	Sutuoktinio / sugyventi-nio ID	IN	kategorinis	1–13
17	PH020	Ar serga kokia nors lėtine liga?	IND	kategorinis	1–2
18	PE010	Ar šiuo metu mokosi?	IND	kategorinis	1–2
19	PE040B	Aukščiausiais įgytas išsilavinimo lygis	IND	kategorinis	0–8
20	PE020	Kur dabar mokosi?	IND	kategorinis	2–8
21	ACT_ST1	Aktyvumo statusas	IND	kategorinis	1–6
22	ACT_ST3	Užimtumo statusas	IND	kategorinis	1–2
23	PL111	Įmonės veiklos kodas	IND	kategorinis	0–9
24	PL040	Užimtumo statusas pagrindiniame darbe	IND	kategorinis	1–4
25	PL06	Kiek valandų dirba per savaitę?	IND	tolydus	1–98
26	HH091	Būste yra tualetas su nutekamuoju vandeniu	HH	Kategorinis	1–2
27	HH081	Būste yra vonia ar dušas	HH	Kategorinis	1–2
28	HY17_1	Ar turi žemės ūkio paskirties žemės?	HH	Kategorinis	1–2
29	HS110	Ar turi automobilį?	HH	Kategorinis	1–2
30	HS090	Ar turi kompiuterį?	HH	Kategorinis	1–2
31	HY020	Disponuojamosios namų ūkio pajamos	HH	Tolydus	
32	HY040G	Pajamos iš turto ir žemės nuomos	HH	Tolydus	
33	HY050G	Socialinės išmokos šeimai	HH	Tolydus	
34	HY060G	Socialinės atskirties ir kitos socialinės išmokos	HH	Tolydus	
35	HY070G	Kompensacijos būstui išlaikyti	HH	Tolydus	
36	HY090G	Pajamos iš palūkanų, dividendų	HH	Tolydus	
37	HH070	Būsto išlaikymo išlaidos	HH	Tolydus	

Nr.	Pavadinimas	Aprašymas	Lygis	Tipas	Reikšmės
38	HH071	Būsto paskola	HH	Tolydus	
39	HY120G	Žemės ir nekilnojamo turto mokesčiai	HH	Tolydus	
40	HY130G	Piniginė parama kitiems asmenims	HH	Tolydus	

Individuali rizika (angl. *individual risk*)

3. Analizuojamame duomenų rinkinyje yra šie jautrūs kintamieji: *tautybė, kintamieji, susiję su asmens darbo jėgos būkle, ir kintamieji, kuriuose yra informacija apie namų ūkio pajamas ir išlaidas (1 lentelė)*. Tiesioginių identifikatorių duomenų rinkinyje nėra, jei jie būtų – juos reikėtų pašalinti. Tačiau yra keli ryšiai tarp kintamųjų, kuriuos reikia išsaugoti nuasmeninimo proceso metu. Namų ūkio kintamieji turi tas pačias reikšmes visiems to namų ūkio nariams. Namų ūkio dydžio reikšmė atitinka faktinį tame namų ūkyje esančių asmenų skaičių duomenų rinkinyje. Taip pat atkreipiame dėmesį į tai, kad kintamieji *LANGUAGE* ir *ETNICITY* neturi nurodytų reikšmių. Kintamieji, kuriuose yra tik trūkstamos reikšmės, šiame etape turėtų būti pašalinti iš duomenų rinkinio (*3 kodas*) ir pašalinti iš nuasmeninimo proceso.

3 kodas. Pašalinami kintamieji su tuščiomis reikšmėmis

```
# Pašalinami kintamieji, kurių visos reikšmės tuščios
file <- file[,!names(file) %in% c('LANGUAGE', 'ETHNICITY')]¶
```

4. Duomenų rinkinyje yra du svorių koeficientai: *WGTHH* ir *WGTPOP*. $WGTPOP = WGTHH \cdot HH\_SIZE$. Apskaičiuodami atskleidimo riziką naudosime namų ūkių svorius *WGTHH* ir asmenų svorius *WGTPOP*.

5. Toliau analizuojami galimi atskleidimo scenarijai, pagal kuriuos pasirenkami kvaziidentifikatoriai (raktiniai kintamieji).

Svarstomi dviejų rūšių atskleidimo scenarijai: 1) atpažinimas remiantis kitais viešai prieinamais duomenų rinkiniais; 2) spontaniškas atpažinimas.

Daroma prielaida, kad galima rasti demografinius gyventojų duomenis, kuriuose pateikiami tokie kintamieji: lytis, amžius, gyvenamoji vieta (savivaldybė, miestas / kaimas) ir kiti kintamieji, tokie kaip šeiminė padėtis, kintamieji, susiję su išsilavinimu ir profesine padėtimi, kurie taip pat yra šiame duomenų rinkinyje. Remiantis turimo duomenų rinkinio analize, pasirenkami kategoriniai kvaziidentifikatoriai *SAVIVAL, MIE_KAIM, HH_SIZE, LYTIS, PB190, AGE, PL111* ir kintamieji, susiję su išsilavinimu. Šie kintamieji gali leisti įsibrovėliui identifikuoti asmenį ar namų ūkį, remiantis kitais turimais duomenų rinkiniais.

Kadangi duomenų rinkinys bus viešoji rinkmena, tai identifikuojančių kintamųjų rinkinys yra per didelis (9 kintamieji), kas lemia didelį raktų derinių skaičių (visos galimos raktinių kintamųjų reikšmių kombinacijos). Dėl tokio pasirinkimo galima identifikuoti daug respondentų, todėl reikia sumažinti identifikuojančių kintamųjų skaičių, šiuo atveju iš duomenų rinkinio pašalinami *HY17_1, PE010*.

Namų ūkio rizika

6. Namų ūkio struktūra yra svarbi duomenų naudotojams, todėl reikia į ją atsižvelgti vertinant riziką. Kadangi kai kurie kintamieji skaičiuojami namų ūkio lygmeniu, tai kiekvieno namų ūkio nario reikšmės yra identiškos.

6.1. Pirmiausia yra nuasmeninami tik namų ūkio kintamieji. Po to jie sujungiami su kintamaisiais asmens lygiu, o tada nuasmeninami asmens ir namų ūkio lygio kintamieji kartu. Tokiu būdu užtikrinama, kad namų ūkio kintamųjų reikšmės išliktų tokios pat kiekvienam namų ūkio nariui ir namų ūkio struktūra negalėtų būti naudojama asmens (-ų) atpažinimui.

Pavyzdyje pateikiamas namų ūkio nuasmeninimo kintamųjų pasirinkimas bei jų išsaugojimas *fileHH* (*4 kodas*). Kiekvienas namų ūkis turės tiek pat įrašų, kiek turi narių (pvz., trijų asmenų namų

ūkis bus pakartotas tris kartus *fileHH*), o prieš analizuojant namų ūkio lygio kintamuosius bus paliekamas tik vienas įrašas kiekvienam namų ūkiui.

4 kodas. Namų ūkio nuasmeninimo kintamųjų pasirinkimas

```
### Kintamųjų pasirinkimas (namų ūkio lygis)
# Raktiniai kintamieji (namų ūkio lygis)
selectedKeyVarsHH = c('MIE_KAIM', 'SAVIVAL', 'HH_SIZE')

# Skaitiniai kintamieji¶
numVarsHH = c('HY020', 'HY040G', 'HY050G', 'HY060G', 'HY070G', 'HY090G', 'HH070',
'HH071', 'HY120G', 'HY130G')

# imties svoris (WGTPOP) (namų ūkio)
weightVarHH = c('WGTPOP')

# Kintamieji, kurie netinka viešajai rinkmenai (NU lygis)
varsNotToBeReleasedHH <- c("HY17_1")

# Visi namų ūkio lygio kintamieji
HHVars <- c('NUID', selectedKeyVarsHH, numVarsHH, weightVarHH)

# Namų ūkių poaibis
fileHH <- file[,HHVars]

# Pašalina pasikartojančias eilutes pagal NUID
fileHH <- fileHH[which(!duplicated(fileHH$NUID)),]
dim(fileHH)

## [1] 5131   15¶
```

Faile *fileHH* yra 5 131 namų ūkis ir 15 kintamųjų. Remiantis *fileHH* duomenimis, sukuriamas *sdsMicro* objektas ir pavadinamas *sdcHH (*5 kodas*)*.

5 kodas. *sdcMicro* objekto sukūrimas

```
# Sukurti pradinį SDC objektą namų ūkio kintamiesiems

sdcHH <- createSdcObj(dat = fileHH, keyVars = selectedKeyVarsHH,
                      weightVar = weightVarHH, numVars = numVarsHH)
```

Pirmiausia vertinamas *2-ų, 3-ų* ir *5-ių* vienetų anonimiškumą pažeidžiančių namų ūkių skaičius. *2 lentelėje* parodytas *2-ų, 3-ų* ir *5-ių* vienetų anonimiškumą pažeidžiančių namų ūkių skaičius ir procentinė dalis nuo visų namų ūkių. *6 kodas* parodo, kaip rasti šias reikšmes naudojant *sdcMicro*.

2 lentelė. Namų ūkiai, pažeidžiantys j vienetų anonimiškumą

j vienetų anonimiškumo lygis	Skaičius NŪ pažeidžiančių	Procentai nuo visų NŪ
2	134	2,61 %
3	262	5,11 %
5	564	10,99 %

6 kodas. j vienetų-anonimiškumą pažeidžiančių namų ūkių skaičius

```
# j vienetų anonimiškumą pažeidžiančių stebėjimų skaičius

print(sdcHH)
#Infos on 2/3-Anonymity:

#Number of observations violating
#- 2-anonymity: 134 (2.612%)
#- 3-anonymity: 262 (5.106%)
#- 5-anonymity: 564 (10.992%)
```

6.2. Naudinga peržiūrėti namų ūkio (-ų) reikšmes, pažeidžiančias *j* vienetų anonimiškumą, nes tai gali padėti išsiaiškinti, kurie kintamieji paverčia šiuos namų ūkius unikaliais. Ši analizė prisideda prie tinkamo statistinio atskleidimo kontrolės metodų pasirinkimo. *7 kode* pademonstruota, kaip surandami namų ūkiai, kurie pažeidžia 3-ų ir 5-ių vienetų anonimiškumą.

7 kodas. Namų ūkiai, pažeidžiantys j vienetų anonimiškumą

```
# Rodyti j vienetų anonimiškumą pažeidžiančių įrašų raktinių kintamųjų reikšmes
fileHH[sdcHH@risk$individual[,2] < 3, selectedKeyVarsHH] # 3-anonymity
fileHH[sdcHH@risk$individual[,2] < 5, selectedKeyVarsHH] # 5-anonymity
```

Bendroji rizika (angl. *global risk*)

7. Turint individualiąsias rizikas arba namų ūkio rizikas, visam duomenų rinkiniui yra skaičiuojama bendroji atskleidimo rizika (angl. *global risk*). *FileHH* kiekvienas įrašas reiškia namų ūkį. Todėl čia naudojama individualioji, o ne hierarchinė rizika, nes asmuo šiuo atveju nurodo namų ūkį. Individualiosios ir bendrosios rizikos priemonės automatiškai atsižvelgia į namų ūkio svorį, kuris buvo apibrėžtas *6 kode*. Analizuojamame duomenų faile bendrasis rizikos matas, apskaičiuotas naudojant pasirinktus raktinius kintamuosius, yra 0,06 % ir atitinka 2,83 numatomą atpažinimą. *8 kode* pavaizduota, kaip atspausdinti bendrosios rizikos matą.

8 kodas. Bendroji rizika

```
print(sdcHH, "risk")
#Risk measures:
#Number of observations with higher risk than the main part of the data: 1
#Expected number of re-identifications: 283 (0.06 %)
```

8. Bendroji rizika nepateikia informacijos apie individualiąsias rizikas. Gali būti, kad keleto namų ūkių individualioji rizika didelė, tačiau bendroji rizika – maža. Todėl reikia peržiūrėti individualiąsias rizikas.

9 kodo pavyzdyje peržiūrima individualioji namų ūkio rizika. Rezultatas rodo, kad didžiausia individualioji rizika – 11,18 %, o 21-o namų ūkio individualioji rizika – didesnė nei 2 %, t. y. šių namų ūkiai individualioji rizika viršija nustatytą 2 % ribą.

9 kodas. Stebėjimai, kai individuali rizika yra didesnė nei 2 %

```
# Stebėjimas su didesne rizika, nei nurodyta riba (0.02)
fileHH[sdcHH@risk$individual[, "risk"] > 0.02,]
#didžiausia rizika
max(sdcHH@risk$individual[, "risk"])
#[1] 0.118537
```

9. Kadangi yra namų ūkių, pažeidžiančių *2-ų* vienetų anonimiškumą, reikia imtis statistinio atskleidimo kontrolės metodų, kad unikalių namų ūkių skaičius būtų sumažintas.

Tikslinga peržiūrėti, kaip pasiskirstęs namų ūkio dydis. *10 kodo* rezultatai rodo, kad šiame duomenų rinkinyje yra labai mažai namų ūkių, kurių narių skaičius yra 8 arba daugiau. Šių reikšmių negalima pergrupuoti ar slėpti, nes iš turimos informacijos apie namų ūkio narius galima išskaičiuoti namų ūkio dydį, todėl vienintelė išeitis yra pašalinti tokius namų ūkius iš duomenų rinkinio.

10 kodas. Namų ūkių šalinimas

```
# Dažnio lentelė kintamajam HH_SIZE
table(sdcHH@manipKeyVars$HH_SIZE)
#1       2        3      4       5      6     7   8    9   10
#1595 1933  838  532  165   52   11 3    1   1

# Pašalinti didelius namų ūkius (8 ar daugiau nariai) iš file ir fileHH
file <- file[!file[,'HH_SIZE'] >= 8,]
fileHHnew <- fileHH[!fileHH[,'HH_SIZE'] >= 8,]

# Sukurti naują „sdcMicro" objektą pagal naują failą be pašalintų namų ūkių
sdcHH <- createSdcObj(dat=fileHHnew, keyVars=selectedKeyVarsHH,
                          weightVar=weightVarHH, numVars = numVarsHH)
```

10. Analizuodami, kodėl *2-ų* vienetų anonimiškumas nėra tenkinamas, matome, kad kintamasis SAVIVAL, turintis 60 reikšmių, gali būti pagrindinis kintamasis, dėl kurio peržengiamos nustatytos anonimiškumo ribos. Vienas iš sprendimų – perkoduoti savivaldybes į apskritis (*11 kodas*).

11 kodas. Savivaldybių perkodavimas į apskritis

```
sdcHH@manipKeyVars$SAVIVAL[sdcHH@manipKeyVars$SAVIVAL == 42 |
                sdcHH@manipKeyVars$SAVIVAL == 85 |
                        sdcHH@manipKeyVars$SAVIVAL == 89 |
                sdcHH@manipKeyVars$SAVIVAL == 86 |
                sdcHH@manipKeyVars$SAVIVAL == 79 |
                sdcHH@manipKeyVars$SAVIVAL == 81 |
                        sdcHH@manipKeyVars$SAVIVAL == 13 |
                sdcHH@manipKeyVars$SAVIVAL == 41 ] <- 10

table(sdcHH@manipKeyVars$SAVIVAL) # Patikrinami grupavimo rezultatai
```

11. Perskaičiavus atskleidimo rizikas (*12 kodas*), duomenų rinkinyje vis dar yra 12 unikalių namų ūkių, bet bendroji atskleidimo rizika sumažėjo nuo 0,05 % iki 0,01 %.

12 kodas. Rizikų perskaičiavimas

```
# perskaičiuoti riziką rankiniu būdu pakeitus reikšmes sdcMicro objekte
calcRisks(sdcHH)

#Infos on 2/3-Anonymity:

#  Number of observations violating
# - 2-anonymity: 12 (0.234%) | in original data: 129 (2.517%)
# - 3-anonymity: 22 (0.429%) | in original data: 257 (5.014%)
# - 5-anonymity: 50 (0.975%) | in original data: 559 (10.905%)
# perskaičiuoti riziką rankiniu būdu pakeitus reikšmes sdcMicro objekte
sdcHH <- calcRisks(sdcHH)

# Bendroji rizika
print(sdcHH, "risk")

#Risk measures:

#Number of observations with higher risk than the main part of the data:

#in modified data: 0

#in original data: 1

#Expected number of re-identifications:

#in modified data: 0.31 (0.01 %)

#in original data: 2.76 (0.05 %)
```

12. Kadangi atlikus reikšmių grupavimą ir didelių namų ūkių pašalinimą nebuvo pasiektas reikiamas *5-ių* vienetų anonimiškumo lygis, dar papildomai bus naudojamas reikšmių slėpimo (ang. *local suppression*) metodas. Reikšmių slėpimo rezultatai pateikti *13 kode*. Pasirinktas trečias variantas, nes *5-ių* vienetų anonimiškumas pasiektas paslėpus mažiau kintamojo *SAVIVAL* reikšmių ir neslepiant kintamojo *HH_SIZE* reikšmių.

13 kodas. Reikšmių slėpimas su svarbos vektoriumi ir be svarbos vektoriaus

```
# Reikšmių slėpimas, kad būtų pasiektas 5-ių vienetų anonimiškas

#1
sdcHH <- localSuppression(sdcHH, k = 5, importance = NULL) # be svarbos vektoriaus
print(sdcHH, "ls")
##Local Suppression:
##KeyVar | Suppressions (#) | Suppressions (%)
```

```
##MIE_KAIM |             →          0 |        →0.000
##SAVIVAL |               46 |         →         0.897
##HH_SIZE |            →6 |        →0.117
## ----------------------------------------------------------------------------
# Anuliuoti slėpimą, kad pamatytumėte svarbos vektoriaus poveikį
sdcHH <- undolast(sdcHH)

#2
#Pakartojamas reikšmių slėpimą, sumažindami slopinimų skaičių HH_SIZE
sdcHH <- localSuppression(sdcHH, k = 5, importance = c(2, 3, 1))
print(sdcHH, "ls")
##Local Suppression:
##KeyVar | Suppressions (#) | Suppressions (%)
##MIE_KAIM |        →     4 |        →          0.078
##SAVIVAL    |           →          48 |        →           0.936
##HH_SIZE    |            →           0 |         →          0.000
##----------------------------------------------------------------------------
# Anuliuoti slėpimą, kad pamatytumėte svarbos vektoriaus poveikį
sdcHH <- undolast(sdcHH)

#3
# Pakartojamas reikšmių slėpimas, pakeičiant svarbos vektoriaus reikšmes
sdcHH <- localSuppression(sdcHH, k = 5, importance = c(3, 2, 1))
print(sdcHH, "ls")
## Local Suppression:
##  KeyVar | Suppressions (#) | Suppressions (%)
## MIE_KAIM |          23 |        →0.449
## SAVIVAL |           27 |        →0.527
## HH_SIZE |            0 |        →0.000
```

13. *12 kode* apskaičiuota bendroji rizika yra beveik nulis, kaip ir tikėtinas atpažinimų skaičius, todėl daroma išvada, kad remiantis kategoriniais kintamaisiais duomenys buvo pakankamai nuasmeninti.

14. Kitas žingsnis – sujungti apdorotus namų ūkio kintamuosius su neapdorotais asmenų kintamaisiais, kad būtų galima nuasmeninti asmens lygio kintamuosius. *14 kode* parodyta, kaip sujungti šiuos failus.

14 kodas. Namų ūkio ir asmenų kintamųjų failų sujungimas ir *sdcMicro* objekto sukūrimas asmenų kintamųjų nuasmeninimui

```
### Kintamųjų pasirinkimas (asmenų lygio)
selectedKeyVarsIND = c('LYTIS', 'PB190', 'AGE', 'PE040B', 'PL111') # raktiniai kintamieji

# Imties svoris
selectedWeightVarIND = c('WGTHH')

# Namų ūkio ID
selectedHouseholdID = c('NUID')

# Kintamieji netinkami viešajai rinkmenai (IND level)
varsNotToBeReleasedIND <- c("PE010", 'RB220', 'RB230', 'RB240')
# Visi asmenų kintamieji
INDVars <- c(selectedKeyVarsIND)¶
HHmanip <- extractManipData(sdcHH) # pakeisti HH kintamieji

# Anonimizuotų HH duomenų rinkinių ir individualaus lygio kintamųjų sujungimas
indVars <- c("NUID", "ASMID", selectedKeyVarsIND, "WGTHH") # NUID ir ne NU kintamieji
fileInd <- file[indVars] # file poaibis be HHVars
```

```
fileCombined <- merge(HHmanip, fileInd, by.x = c('NUID'))
fileCombined <- fileCombined[order(fileCombined[,'NUID'],  fileCombined[,'ASMID']),]

dim(fileCombined)
##[1] 11317→22


# SDC objektas tik su IND kintamaisiais
sdcCombined <- createSdcObj(dat = fileCombined, keyVars = c(selectedKeyVarsIND),
                      weightVar = selectedWeightVarIND, hhId = selectedHouseholdID)
```

15. Sujungus duomenų failus, vertinamas įrašų, pažeidžiančių *2-ų, 3-ų* ir *5-ių* vienetų anonimiškumą (*15 kodas*), skaičius. Bendroji hierarchinė atskleidimo rizika yra 1,11 %, tai atitinka maždaug 125 numatomus atpažinimus. Čia naudojama hierarchinė rizika, nes identifikavus vieną namų ūkio narį, būtų galima identifikuoti ir kitus to paties namų ūkio narius. Net 2 661 įrašas turi individualią hierarchinę riziką, viršijančią 1 %, o didžiausią – 40,05 %.

15 kodas. Rizikos prieš nuasmeninimą

```
# Asmenų skaičius pažeidžiančių j vienetų anonimiškumą
print(sdcCombined)
## Infos on 2/3-Anonymity:
##
## Number of observations violating
## - 2-anonymity: 1063 (9.393%)
## - 3-anonymity: 2129 (18.812%)
## - 5-anonymity: 3876 (34.249%)

# Bendroji rizika
print(sdcCombined, 'risk')
## Risk measures:
##
##Number of observations with higher risk than the main part of the data: 20
##Expected number of re-identifications: 50.40 (0.45 %)
##Information on hierarchical risk:
## Expected number of re-identifications: 125.23 (1.11 %)


# Įrašų su gana didele rizika skaičius
dim(fileCombined[sdcCombined@risk$individual[, "hier_risk"] > 0.01,])
## [1] 2661   22

# Didžiausia individuali rizika
max(sdcCombined@risk$individual[, "hier_risk"])
## [1] 0.4005182
```

16. Iš *15 kode* pateiktos informacijos matyti, kad rizikos labai didelės, todėl reikia imtis duomenų nuasmeninimo priemonių.

17. Viena iš priemonių – kintamųjų perkodavimas. Pirmiausia *16 kode* atliekamas kintamojo *amžiaus* perkodavimas. Šis perkodavimas sumažino *2-ų* vienetų anonimiškumą pažeidžiančių asmenų skaičių iki 36, tačiau to neužtenka, todėl perkoduojamas įgytas išsilavinimo lygis į 3 grupes: *žemesnis nei vidurinis*, *vidurinis* ir *aukštasis*, taip pat perkoduojama ir įmonės veikla į 3 grupes (*16 kodas*).

16 kodas. Asmens kintamųjų perkodavimas

```
#perkoduojama į amžiaus grupes
sdcCombined@manipKeyVars$AGE[sdcCombined@manipKeyVars$AGE <= 15] <- 15
sdcCombined@manipKeyVars$AGE[sdcCombined@manipKeyVars$AGE >= 16 &
              sdcCombined@manipKeyVars$AGE < 25] <- 20
...
sdcCombined@manipKeyVars$AGE[sdcCombined@manipKeyVars$AGE >= 65] <- 65
```

```
table(sdcCombined@manipKeyVars$AGE) # patikrinti rezultatus
#1520   30    40    50      60      65
#1430   898  961  1147  1851  2151 2879
#0 = 'Ikimokyklinis ugdymas'  1 = 'Pradinis ugdymas' 2 = 'Pagrindinis ugdymas'
#3 = 'Vidurinis ugdymas' 4 = 'Profesinis mokymas turint vidurinį išsilavinimą'
#6 = 'Bakalauro studijos' 7 = 'Magistrantūros studijos' 8 = 'Doktorantūra'

sdcCombined@manipKeyVars$PE040B[sdcCombined@manipKeyVars$PE040B == 0 |
                  sdcCombined@manipKeyVars$PE040B == 1 |
                  sdcCombined@manipKeyVars$PE040B == 2 ] <- 1 # žemesnis nei vidurinis
sdcCombined@manipKeyVars$PE040B[sdcCombined@manipKeyVars$PE040B == 3 |
                  sdcCombined@manipKeyVars$PE040B == 4] <- 3 # vidurinis
sdcCombined@manipKeyVars$PE040B[sdcCombined@manipKeyVars$PE040B == 6 |
                  sdcCombined@manipKeyVars$PE040B == 7 |
                  sdcCombined@manipKeyVars$PE040B == 8 ] <- 6 # aukštasis
table(sdcCombined@manipKeyVars$PE040B, useNA = "ifany")

calcRisks(sdcCombined) # perskaičiuoja rizikas
```

Kadangi atlikus grupavimą nepavyko pasiekti *5-ių* vienetų anonimiškumo, papildomai naudojama reikšmių slėpimo funkcija (*17 kodas*).

17 kodas. Reikšmių slėpimas 5-ių vienetų anonimiškumui pasiekti

```
# Reikšmių slėpimas naudojant svarbos vektorių
sdcCombined <- localSuppression(sdcCombined, k = 5, importance = c(4, 5, 1, 2, 3))

print(sdcCombined, "ls")

#KeyVar | Suppressions (#) | Suppressions (%)
#  LYTIS   |            0 |        →0.000
#  PB190   |           69 |        →0.610
#  AGE     |            0 |        →0.000
#  PE040B|              0 |        →0.000
#  PL111   |             0 |        →0.000
```

18. Pritaikius reikšmių slėpimo funkciją perskaičiuojamos rizikos. Gautas rezultatas parodo, kad pritaikius šią funkciją buvo pasiektas *5-ių* vienetų anonimiškumas. Bendroji hierarchinė atskleidimo rizika sumažėjo iki 0,01 % (arba 1,5) tikėtino teisingo atpažinimo. Didžiausia individuali hierarchinė identifikavimo rizika yra 0,17 %. Šios rizikos yra pakankamos, kad šis duomenų rinkinys būtų skelbiamas.

Globalus perkodavimas (angl. *global recoding*)

19. Taikant globalųjį perkodavimą reikia nuspręsti, kokio dydžio turėtų būti naujos grupės ir kokios reikšmės turėtų būti sujungtos į vieną grupę. Naujos grupės turėtų būti parinktos atsižvelgiant į vartotojų poreikius ir minimizuojant perkodavimo metu prarandamos informacijos kiekį. *18 kode* pademonstruotas kategorinio kintamojo perkodavimas, taikant *sdcMicro* paketo funkciją *groupVars()*.

18 kodas. Kategorinio kintamojo perkodavimas naudojant *R* paketo *sdcMicro* funkciją *groupVars()*

```
sdcInitial <- groupVars(obj = sdcInitial, var = c("Regionas"),
                        before = c("Regionas 1", "Regionas 2", "Regionas 3", "Regionas
                        4", "Regionas 5"),
                        after = c("Šiaurės", "Šiaurės", "Centrinis", "Pietų", "Pietų"))
```

20. *19 kode* pateiktas pavyzdys, kaip tolydaus kintamojo „age“, nusakančio respondentų amžių, reikšmes galima sugrupuoti į vienodo dydžio intervalus kas dešimt metų.

19 kodas. Tolydaus kintamojo perkodavimas naudojant *R* paketo *sdcMicro* funkciją *GlobalRecode()*

```
sdcInitial <- globalRecode(sdcInitial, column = c('age'),
                           breaks = 10 * c(0:10)

# Frequencies of age after recoding
```

```
table(sdcInitial@manipKeyVars$age)
##    (0,10]   (10,20]   (20,30]   (30,40]   (40,50]   (50,60]   (60,70]   (70,80]   (80,90]
(90,100]
##       462       483       344       368       294       214       172        94        26
3
```

21. Galima pasirinkti ir nevienodo ilgio intervalus, pavyzdžiui, tokias amžiaus grupes: 1–5, 6–11, 12–17, 18–21, 22–25, 26–49, 50–64 ir 65+. Kaip jas sugrupuoti, parodyta *20 kode*.

20 kodas. Tolydaus kintamojo perkodavimas naudojant *R* paketo *sdcMicro* funkciją *GlobalRecode()*

```
sdcInitial <- globalRecode(sdcInitial, column = c('age'),
                           breaks = c(0, 5, 11, 17, 21, 25, 49, 65, 100))

# Frequencies of age after recoding
table(sdcInitial@manipKeyVars$age)
##    (0,5]   (5,11]  (11,17]  (17,21]  (21,25]  (25,49]  (49,65] (65,100]
##      192      317      332      134      142      808      350      185
```

22. Svarbu atkreipti dėmesį, kad funkcija *GlobalRecode()* leidžia sudaryti tik iš kairės atvirus intervalus, todėl į tai būtina atsižvelgti renkantis intervalų ribas. Reikšmės, nepatenkančios nė į vieną iš intervalų, pakeičiamos tuščiomis (*NA*). Pateiktame pavyzdyje amžius, lygus 0, taps tuščia reikšme, kas lems tam tikrą informacijos praradimą.

Viršaus ir apačios perkodavimas (angl. *top and bottom coding*)

23. Viršaus ir apačios perkodavimas gali būti atliekamas su *R* paketo *sdcMicro* funkcija *topBotCoding()*. Viršutinės ir apatinės skirstinio reikšmės negali būti perkoduojamos tuo pačiu metu, reikia įvykdyti atskiras komandas. Žemiau esantis *21 kodas* iliustruoja, kaip atliekamas didesnių nei 64 metai ir mažesnių nei 5 metai reikšmių perkodavimas.

21 kodas. Viršaus ir apačios perkodavimas

```
# Top coding at age 65
sdcInitial <- topBotCoding(obj = sdcInitial, value = 65, replacement = 65,
                           kind = 'top', column = 'age')

# Bottom coding at age 5¶
sdcInitial <- topBotCoding(obj = sdcInitial, value = 5, replacement = 5,
                           kind = 'bottom', column = 'age')
```

24. Norint sukurti keletą viršutinių reikšmių kategorijų, pavyzdžiui, 65–80 metų ir daugiau kaip 80 metų, galima naudoti anksčiau aprašytą funkciją *groupVars()* arba tai atlikti rankiniu būdu.

Duomenų panaudojamumo ir informacijos praradimo matai

25. Taikant bet kurį statistinio atskleidimo kontrolės metodą, atsiranda tam tikras informacijos praradimas. Žemiau pateikiama keletas pavyzdžių, kaip galima įvertinti prarastą informaciją.

25.1. Vienas iš siūlomų būdų galėtų būti ryšių lentelių palyginimas. *22 kode* pavaizduota, kaip skiriasi originalių ir nuasmenintų duomenų ryšių lentelės.

22 kodas. Ryšių lentelių palyginimas

```
# Contingency table (cross tabulation) of the variables region and urban/rural
table(sdcInitial@origData[, c('REGION', 'URBRUR')]) # before anonymization
##       URBRUR
## REGION   1   2
##      1 235  89
##      2 261  73
##      3 295  76
##      4 304  71
##      5 121 139
##      6 100 236

table(sdcInitial@manipKeyVars[,c('REGION', 'URBRUR')]) # after anonymization
```

```
##       URBRUR
## REGION   1   2
##      1 235  89
##      2 261  73
##      3 295  76
##      4 304  71
##      5 105 130
##      6  79 201¶
```

25.2. *Informacijos praradimo matas IL1S* įvertina atstumą tarp pradinio duomenų rinkinio X ir galutinio duomenų rinkinio Z. Naudojant *sdcMicro* paketą, šį matą galima apskaičiuoti su funkcija *dUtility()*, kaip parodyta pavyzdyje.

23 kodas. Informacijos praradimo matas IL1S

```
#Evaluating IL1s measure for all variables in the sdcMicro object sdcInitial
sdcInitial <- dUtility(sdcInitial)
#Calling the result of IL1s
sdcInitial@utility$il1
##[1] 0.2203791

#IL1s for a subset of the numerical quasi-identifiers
subset <- c('INCRMT', 'INCWAGE', 'INCFARMBSN')
dUtility(obj = sdcInitial@origData[,subset], xm = sdcInitial@manipNumVars[,subset],
method = 'IL1')
##[1] 0.5641103
```

Nors funkcijos pavadinimas susietas su duomenų panaudojamumu, ji grąžina informacijos praradimo matą. Jis naudingas lyginant kelis skirtingus metodus. Kuo šio mato reikšmė (t. y. atstumas tarp originalių ir pakeistų duomenų) didesnė, tuo mažesnis duomenų rinkinio panaudojamumas, tačiau tuo pat metu ir mažesnė atskleidimo rizika. Mažesnė mato reikšmė reiškia, kad duomenys pritaikius nuasmeninimo metodus nedaug skiriasi nuo originalių.

25.3. *Tikrinės reikšmės (*angl. *eigenvalues)* naudojamos skirtingiems metodams palyginti. Anksčiau paminėta funkcija *dUtility()* grąžina tikrinių reikšmių skirtumą tarp originalių duomenų rinkinio ir nuasmeninto duomenų rinkinio. Kuo šis skirtumas didesnis, tuo didesni duomenų rinkinio pokyčiai buvo atlikti ir tuo didesnį informacijos praradimą turime.

24 kodas. Tikrinės reikšmės (angl. *eigenvalues*)

```
#Comparison of eigenvalues of continuous variables
dUtility(obj = sdcInitial@origData[,contVars],
         xm = sdcInitial@manipNumVars[,contVars], method = 'eigen')
##[1] 2.482948

#Comparison of robust eigenvalues of continuous variables*
dUtility(obj = sdcInitial@origData[,contVars],
         xm = sdcInitial@manipNumVars[,contVars], method = 'robeigen')
##[1] -4.297621e+14
```

25.4. *Gini koeficientas* – vienas iš dispersijos matų, skaičiuojamas originaliam ir nuasmenintam duomenų rinkiniui ir analizuojama, kaip keičiasi Gini koeficiento reikšmės. Vienas iš būdų apskaičiuoti šį koeficientą – pritaikyti R paketo *laeken* funkciją *gini()*.

25 kodas. Gini koeficientas

```
#Gini coefficient before anonymization
gini(inc = sdcInitial@origData[selInc,'INC'],
weights = curW[selInc], na.rm = TRUE)$value # before
##[1] 34.05928

#Gini coefficient after anonymization
```

```
gini(inc = sdcInitial@manipNumVars[selInc,'INC'],
weights = curW[selInc], na.rm = TRUE)$value # after
##[1] 67.13218
```