sdcmicro-exercise

Michelle Lam and Alex Reed

2023-05-25

Whale Entanglement sdcMicro Exercise

Your team acquired a dataset* from researchers working with whale entanglement data on the West Coast. The dataset contains both direct and indirect identifiers. Your task is to assess the risk of re-identification of the fisheries associated with the cases before considering public release. Then, you should test one technique and apply k-anonymization to help lower the disclosure risk as well as compute the information loss.

Please complete this exercise in pairs or groups of three. Each group should download the dataset and complete the rmd file, including the code and answering the questions. Remember to include your names in the YAML.

**This dataset was purposefully adapted exclusively for instruction use.*

Setup

```
library(sdcMicro)
whale_data <- read.csv("whale-sdc.csv")
```

Package & Data

```
head(whale_data)
```

Inspect the Dataset

```
##      case_id year month       type        county state    lat     long
## 1 20000201Er 2000     2 Gray Whale     San Diego    CA 32.670 -117.229
## 2 20000316Er 2000     3 Gray Whale        Orange    CA 33.383 -117.617
## 3 20000327Er 2000     3 Gray Whale   Los Angeles    CA 33.992 -118.804
## 4 20000330Er 2000     3 Gray Whale   Los Angeles    CA 33.710 -118.224
## 5 20000404Er 2000     4 Gray Whale Santa Barbara    CA 33.720 -118.080
## 6 20000610Er 2000     6 Gray Whale    Santa Cruz    CA 36.953 -121.910
##   inj_level condition     origin    gear fishery_license fine infraction_type
## 1         8     alive commercial gillnet      4649644859    1               1
## 2         8     alive commercial gillnet      7918308514    1               1
## 3         7     alive commercial gillnet      6621060947    0               0
## 4        10      dead commercial gillnet       3702613383    1               1
## 5        10      dead commercial     trap      7084197273    1               1
## 6         5     alive commercial gillnet       1321152653    0               0
```

```
str(whale_data)
```

```
## ’data.frame’:    348 obs. of  15 variables:
##  $ case_id        : chr  "20000201Er" "20000316Er" "20000327Er" "20000330Er" ...
##  $ year           : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2001 ...
##  $ month          : int  2 3 3 3 4 6 7 8 11 9 ...
##  $ type           : chr  "Gray Whale" "Gray Whale" "Gray Whale" "Gray Whale" ...
##  $ county         : chr  "San Diego" "Orange" "Los Angeles" "Los Angeles" ...
##  $ state          : chr  "CA" "CA" "CA" "CA" ...
##  $ lat            : num  32.7 33.4 34 33.7 33.7 ...
##  $ long           : num  -117 -118 -119 -118 -118 ...
##  $ inj_level      : int  8 8 7 10 10 5 3 3 10 3 ...
##  $ condition      : chr  "alive" "alive" "alive" "dead" ...
##  $ origin         : chr  "commercial" "commercial" "commercial" "commercial" ...
##  $ gear           : chr  "gillnet" "gillnet" "gillnet" "gillnet" ...
##  $ fishery_license: num  4.65e+09 7.92e+09 6.62e+09 3.70e+09 7.08e+09 ...
##  $ fine           : int  1 1 0 1 1 0 0 0 1 0 ...
##  $ infraction_type: int  1 1 0 1 1 0 0 0 1 0 ...
```

Q1. How many direct identifiers are present in this dataset? What are they? A: There is 1 direct identifier for fisheries: fishery_license

Q2. What attributes would you consider quasi-identifiers? Why? A: The rest of the attributes could be quasi-identifiers because they can be used in combination to identify a fishery.

Q3. What types of variables are they? Define them. (numeric, integer, factor or string) A: For the purposes of this exercise, the year, month, type, county, state, inj_level, condition, origin, gear, fine, infraction_type, lat, and long can be considered factor variables.

Make sure to have them set correctly.

```
fname_whale = "whale-sdc.csv"
file_whale <- read.csv(fname_whale)
# assign variables in df as factors (variables we want to define as categorical)
file_whale <- varToFactor(obj=file_whale, var=c("year","month", "type",
                                                "county","state", "inj_level",
                                                "condition", "origin", "gear",
                                                "fine", "infraction_type", "lat",
                                                "long"))
```

4 Considering your answers to questions 1, 2 and 3 create a SDC problem.

```
sdcInitial <- createSdcObj(dat=file_whale,
                       keyVars=c("year","month", "type", "county","state",
                                 "inj_level", "condition", "origin", "gear",
                                 "fine", "infraction_type", "lat", "long"),
                       numVars=NULL,
                       weightVar=NULL,
                       hhId=NULL,
                       strataVar=NULL,
                       pramVars=NULL,
                       excludeVars=c("fishery_license"),
                       seed=0,
                       randomizeRecords=FALSE,
                       alpha=c(1))

sdcInitial@risk$global$risk
```

Q4.1 What is the risk of re-identification for this dataset?

```
## [1] 1
```

A: The risk of re-identification for this dataset is 1 or 100%.

```
# look at sdc object
sdcInitial
```

Q4.2 To what extent does this dataset violate k-anonymity?

```
## The input dataset consists of 348 rows and 14 variables.
##
## The following variables have been deleted are not available in the output dataset:
##   --> fishery_license
##
##
##   --> Categorical key variables: year, month, type, county, state, inj_level, condition, origin, gea
## ----------------------------------------------------------------------
```

```
## Information on categorical key variables:
##
## Reported is the number, mean size and size of the smallest category >0 for recoded variables.
## In parenthesis, the same statistics are shown for the unmodified data.
## Note: NA (missings) are counted as seperate categories!
```

```
##     Key Variable Number of categories       Mean size
##             year                   20  (20)    17.400  (17.400)
##            month                   12  (12)    29.000  (29.000)
##             type                    8   (8)    43.500  (43.500)
##           county                   31  (31)    11.226  (11.226)
##            state                    3   (3)   116.000 (116.000)
##        inj_level                   11  (11)    31.636  (31.636)
##        condition                    2   (2)   174.000 (174.000)
##           origin                    3   (3)   116.000 (116.000)
##             gear                    8   (8)    43.500  (43.500)
##             fine                    2   (2)   174.000 (174.000)
##  infraction_type                    5   (5)    69.600  (69.600)
##              lat                  322 (322)     1.081   (1.081)
##             long                  331 (331)     1.051   (1.051)
##  Size of smallest (>0)
##                       1   (1)
##                      11  (11)
##                       2   (2)
##                       1   (1)
##                      32  (32)
##                       1   (1)
##                      32  (32)
##                       3   (3)
##                       4   (4)
##                     109 (109)
##                      22  (22)
##                       1   (1)
##                       1   (1)
```

```
## ----------------------------------------------------------------------
```

```
## Infos on 2/3-Anonymity:
##
## Number of observations violating
##   - 2-anonymity: 348 (100.000%)
##   - 3-anonymity: 348 (100.000%)
##   - 5-anonymity: 348 (100.000%)
##
## ----------------------------------------------------------------------
```

```
# look at which observations have a higher risk of being re-identified
sdcInitial@risk$individual
```

```
##        risk fk Fk
##   [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
##  [16,]    1  1  1
##  [17,]    1  1  1
##  [18,]    1  1  1
##  [19,]    1  1  1
##  [20,]    1  1  1
##  [21,]    1  1  1
##  [22,]    1  1  1
##  [23,]    1  1  1
##  [24,]    1  1  1
##  [25,]    1  1  1
##  [26,]    1  1  1
##  [27,]    1  1  1
##  [28,]    1  1  1
##  [29,]    1  1  1
##  [30,]    1  1  1
##  [31,]    1  1  1
##  [32,]    1  1  1
##  [33,]    1  1  1
##  [34,]    1  1  1
##  [35,]    1  1  1
##  [36,]    1  1  1
##  [37,]    1  1  1
##  [38,]    1  1  1
##  [39,]    1  1  1
##  [40,]    1  1  1
##  [41,]    1  1  1
##  [42,]    1  1  1
##  [43,]    1  1  1
##  [44,]    1  1  1
##  [45,]    1  1  1
##  [46,]    1  1  1
##  [47,]    1  1  1
##  [48,]    1  1  1
##  [49,]    1  1  1
##  [50,]    1  1  1
##  [51,]    1  1  1
##  [52,]    1  1  1
##  [53,]    1  1  1
##  [54,]    1  1  1
##  [55,]    1  1  1
##  [56,]    1  1  1
##  [57,]    1  1  1
##  [58,]    1  1  1
##  [59,]    1  1  1
##  [60,]    1  1  1
##  [61,]    1  1  1
##  [62,]    1  1  1
##  [63,]    1  1  1
##  [64,]    1  1  1
##  [65,]    1  1  1
##  [66,]    1  1  1
##  [67,]    1  1  1
##  [68,]    1  1  1
##  [69,]    1  1  1
##  [70,]    1  1  1
##  [71,]    1  1  1
##  [72,]    1  1  1
##  [73,]    1  1  1
##  [74,]    1  1  1
##  [75,]    1  1  1
##  [76,]    1  1  1
##  [77,]    1  1  1
##  [78,]    1  1  1
##  [79,]    1  1  1
##  [80,]    1  1  1
##  [81,]    1  1  1
##  [82,]    1  1  1
##  [83,]    1  1  1
##  [84,]    1  1  1
##  [85,]    1  1  1
##  [86,]    1  1  1
##  [87,]    1  1  1
##  [88,]    1  1  1
##  [89,]    1  1  1
##  [90,]    1  1  1
##  [91,]    1  1  1
##  [92,]    1  1  1
##  [93,]    1  1  1
##  [94,]    1  1  1
##  [95,]    1  1  1
##  [96,]    1  1  1
##  [97,]    1  1  1
##  [98,]    1  1  1
##  [99,]    1  1  1
## [100,]    1  1  1
## [101,]    1  1  1
## [102,]    1  1  1
## [103,]    1  1  1
## [104,]    1  1  1
## [105,]    1  1  1
## [106,]    1  1  1
## [107,]    1  1  1
## [108,]    1  1  1
## [109,]    1  1  1
## [110,]    1  1  1
## [111,]    1  1  1
## [112,]    1  1  1
## [113,]    1  1  1
## [114,]    1  1  1
## [115,]    1  1  1
## [116,]    1  1  1
## [117,]    1  1  1
## [118,]    1  1  1
## [119,]    1  1  1
## [120,]    1  1  1
## [121,]    1  1  1
## [122,]    1  1  1
## [123,]    1  1  1
## [124,]    1  1  1
## [125,]    1  1  1
## [126,]    1  1  1
## [127,]    1  1  1
## [128,]    1  1  1
## [129,]    1  1  1
## [130,]    1  1  1
## [131,]    1  1  1
## [132,]    1  1  1
## [133,]    1  1  1
## [134,]    1  1  1
## [135,]    1  1  1
## [136,]    1  1  1
## [137,]    1  1  1
## [138,]    1  1  1
## [139,]    1  1  1
## [140,]    1  1  1
## [141,]    1  1  1
## [142,]    1  1  1
## [143,]    1  1  1
## [144,]    1  1  1
## [145,]    1  1  1
## [146,]    1  1  1
## [147,]    1  1  1
## [148,]    1  1  1
## [149,]    1  1  1
## [150,]    1  1  1
## [151,]    1  1  1
## [152,]    1  1  1
## [153,]    1  1  1
## [154,]    1  1  1
## [155,]    1  1  1
## [156,]    1  1  1
## [157,]    1  1  1
## [158,]    1  1  1
## [159,]    1  1  1
## [160,]    1  1  1
## [161,]    1  1  1
## [162,]    1  1  1
## [163,]    1  1  1
## [164,]    1  1  1
## [165,]    1  1  1
## [166,]    1  1  1
## [167,]    1  1  1
## [168,]    1  1  1
## [169,]    1  1  1
## [170,]    1  1  1
## [171,]    1  1  1
## [172,]    1  1  1
## [173,]    1  1  1
## [174,]    1  1  1
## [175,]    1  1  1
## [176,]    1  1  1
## [177,]    1  1  1
## [178,]    1  1  1
## [179,]    1  1  1
## [180,]    1  1  1
## [181,]    1  1  1
## [182,]    1  1  1
## [183,]    1  1  1
## [184,]    1  1  1
## [185,]    1  1  1
## [186,]    1  1  1
## [187,]    1  1  1
## [188,]    1  1  1
## [189,]    1  1  1
## [190,]    1  1  1
## [191,]    1  1  1
## [192,]    1  1  1
## [193,]    1  1  1
## [194,]    1  1  1
## [195,]    1  1  1
## [196,]    1  1  1
## [197,]    1  1  1
## [198,]    1  1  1
## [199,]    1  1  1
## [200,]    1  1  1
## [201,]    1  1  1
## [202,]    1  1  1
## [203,]    1  1  1
## [204,]    1  1  1
## [205,]    1  1  1
## [206,]    1  1  1
## [207,]    1  1  1
## [208,]    1  1  1
## [209,]    1  1  1
## [210,]    1  1  1
## [211,]    1  1  1
## [212,]    1  1  1
## [213,]    1  1  1
## [214,]    1  1  1
## [215,]    1  1  1
## [216,]    1  1  1
## [217,]    1  1  1
## [218,]    1  1  1
## [219,]    1  1  1
## [220,]    1  1  1
## [221,]    1  1  1
## [222,]    1  1  1
## [223,]    1  1  1
## [224,]    1  1  1
## [225,]    1  1  1
## [226,]    1  1  1
## [227,]    1  1  1
## [228,]    1  1  1
## [229,]    1  1  1
## [230,]    1  1  1
## [231,]    1  1  1
## [232,]    1  1  1
## [233,]    1  1  1
## [234,]    1  1  1
## [235,]    1  1  1
## [236,]    1  1  1
## [237,]    1  1  1
## [238,]    1  1  1
## [239,]    1  1  1
## [240,]    1  1  1
## [241,]    1  1  1
## [242,]    1  1  1
## [243,]    1  1  1
## [244,]    1  1  1
## [245,]    1  1  1
## [246,]    1  1  1
## [247,]    1  1  1
## [248,]    1  1  1
## [249,]    1  1  1
## [250,]    1  1  1
## [251,]    1  1  1
## [252,]    1  1  1
## [253,]    1  1  1
## [254,]    1  1  1
## [255,]    1  1  1
## [256,]    1  1  1
## [257,]    1  1  1
## [258,]    1  1  1
## [259,]    1  1  1
## [260,]    1  1  1
## [261,]    1  1  1
## [262,]    1  1  1
## [263,]    1  1  1
## [264,]    1  1  1
## [265,]    1  1  1
## [266,]    1  1  1
## [267,]    1  1  1
## [268,]    1  1  1
## [269,]    1  1  1
## [270,]    1  1  1
## [271,]    1  1  1
## [272,]    1  1  1
## [273,]    1  1  1
## [274,]    1  1  1
## [275,]    1  1  1
## [276,]    1  1  1
## [277,]    1  1  1
## [278,]    1  1  1
## [279,]    1  1  1
## [280,]    1  1  1
## [281,]    1  1  1
## [282,]    1  1  1
## [283,]    1  1  1
## [284,]    1  1  1
## [285,]    1  1  1
## [286,]    1  1  1
## [287,]    1  1  1
## [288,]    1  1  1
## [289,]    1  1  1
## [290,]    1  1  1
## [291,]    1  1  1
## [292,]    1  1  1
## [293,]    1  1  1
## [294,]    1  1  1
## [295,]    1  1  1
## [296,]    1  1  1
## [297,]    1  1  1
## [298,]    1  1  1
## [299,]    1  1  1
## [300,]    1  1  1
## [301,]    1  1  1
## [302,]    1  1  1
## [303,]    1  1  1
## [304,]    1  1  1
## [305,]    1  1  1
## [306,]    1  1  1
## [307,]    1  1  1
## [308,]    1  1  1
## [309,]    1  1  1
## [310,]    1  1  1
## [311,]    1  1  1
## [312,]    1  1  1
## [313,]    1  1  1
## [314,]    1  1  1
## [315,]    1  1  1
## [316,]    1  1  1
## [317,]    1  1  1
## [318,]    1  1  1
## [319,]    1  1  1
## [320,]    1  1  1
## [321,]    1  1  1
## [322,]    1  1  1
## [323,]    1  1  1
## [324,]    1  1  1
## [325,]    1  1  1
## [326,]    1  1  1
## [327,]    1  1  1
## [328,]    1  1  1
## [329,]    1  1  1
## [330,]    1  1  1
## [331,]    1  1  1
## [332,]    1  1  1
## [333,]    1  1  1
## [334,]    1  1  1
## [335,]    1  1  1
## [336,]    1  1  1
## [337,]    1  1  1
## [338,]    1  1  1
## [339,]    1  1  1
## [340,]    1  1  1
## [341,]    1  1  1
## [342,]    1  1  1
## [343,]    1   1   1
## [344,]    1   1   1
## [345,]    1   1   1
## [346,]    1   1   1
## [347,]    1   1   1
## [348,]    1   1   1
```

```
# how many combinations of key variables does each record have
freq(sdcInitial, type = 'fk')
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

A: Looking at the sdc object, 100% of the observations in the data set violate 2, 3, and 5 anonymity.

5. Consider techniques that could reduce the risk of re-identification.

```
# Frequencies of year before recoding
table(sdcInitial@manipKeyVars$year)
```

Q5.1 Apply one non-perturbative method to a variable of your choice. How effective was it in lowering the disclosure risk?

```
## 
## 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 
##    9    1    3   10    9   10    8   11    5    8   15    9   14   11   23   51 
## 2016 2017 2018 2019 
##   53   30   45   23 
```

```
## Recode variable year (top coding)
sdcInitial <- groupAndRename(obj= sdcInitial, var= c("year"),
                             before=c("2000", "2001", "2002", "2003", "2004",
                                      "2005", "2006", "2007", "2008", "2009"),
                             after=c("2000-2009"))

## Recode variable year (bottom coding)
sdcInitial <- groupAndRename(obj= sdcInitial, var= c("year"),
                             before=c("2010", "2011", "2012", "2013", "2014",
                                      "2015", "2016", "2017", "2018", "2019"),
                             after=c("2010-2019"))
```

```
sdcInitial@risk$global$risk
```

```
## [1] 1
```

```
print(sdcInitial, 'kAnon')
```

```
## Infos on 2/3-Anonymity:
##
## Number of observations violating
##   - 2-anonymity: 348 (100.000%)
##   - 3-anonymity: 348 (100.000%)
##   - 5-anonymity: 348 (100.000%)
##
## ----------------------------------------------------------------------
```

A: When applying top and bottom coding to the sdc object for year, it was not effective at all for lowering disclosure risk. The risk is still 1 or 100%.

```
# apply k-3 anonymization
sdcInitial <- kAnon(sdcInitial, k = c(3))
sdcInitial@risk$global$risk
```

Q5.2 Apply (k-3) anonymization to this dataset.

```
## [1] 0.2408785
```

A: After k-3 anonymization, risk for this dataset decreased to about 0.24.

```
# show suppression rates
print(sdcInitial, 'ls')
```

Q6. Compute the information loss for the de-identified version of the dataset.

```
## Local suppression:
```

```
##           KeyVar | Suppressions (#) | Suppressions (%)
##             year |                7 |            2.011
##            month |              140 |           40.230
##             type |               31 |            8.908
##           county |              149 |           42.816
##            state |               20 |            5.747
##        inj_level |               89 |           25.575
##        condition |                2 |            0.575
##           origin |                 5 |            1.437
##             gear |                60 |           17.241
##             fine |                 1 |            0.287
##  infraction_type |                22 |            6.322
##              lat |               223 |           64.080
##             long |               245 |           70.402
```

```
## ----------------------------------------------------------------------
```

```r
#We can also compare the number of NAs before and after our interventions
# Store the names of all categorical key variables in a vector
namesKeyVars <- names(sdcInitial@manipKeyVars)

# Matrix to store the number of missing values (NA) before and after anonymization
NAcount <- matrix(NA, nrow = 2, ncol = length(namesKeyVars))
colnames(NAcount) <- c(paste0('NA', namesKeyVars)) # column names
rownames(NAcount) <- c('initial', 'treated') # row names

# NA count in all key variables (NOTE: only those coded NA are counted)
for(i in 1:length(namesKeyVars)) {
  NAcount[1, i] <- sum(is.na(sdcInitial@origData[,namesKeyVars[i]]))
  NAcount[2, i] <- sum(is.na(sdcInitial@manipKeyVars[,i]))}

# Show results
NAcount
```

```
##         NAyear NAmonth NAtype NAcounty NAstate NAinj_level NAcondition NAorigin
## initial      0       0      0        0       0           0           0        0
## treated      7     140     31      149      20          89           2        5
##         NAgear NAfine NAinfraction_type NAlat NAlong
## initial      0      0                 0     0      0
## treated     60      1                22   223    245
```

A: When looking at which variables had the highest suppression rates, the lat and long were the top 2 at about 64% and 70.4% respectively. The county variable was next highest at 42.8% and then month at about 40.2%. Additionally, when examining the number of NAs that were added in to each variable, the lat and long had the most with Lat going from 0 to 223 NAs and long going from 0 NAs to 245.