Homework 3: Blocking Approaches Applied to the El Salvodoran Conflict

Rebecca C. Steorts, STA 490/690

Total points on assignment: 5 (reproducibility) + 10 points for the assignment.

El Salvador Civil War

We will continue with our exploration of the UNTC data set from El Salvador.

```r
library(knitr)
library(RecordLinkage)
```

```r
# read in data
df <- read.csv("../sv-mauricio/sv-mauricio.csv")
ent_id <- df$HandID
# Filter out records with ground truth, leaving dept 1 and 7
df <- df[!is.na(ent_id),]
ent_id <- ent_id[!is.na(ent_id)]
new_df <- df[,c(3:8,10)]
head(new_df)
```

```
##          lastname firstname day month year geocode dept
## 26   ALEMAN SOLIS   ALFREDO   2     5 1984   70000    7
## 64           CRUS    CARMEN  21    10 1981   10000    1
## 66        MONTOYA    CARMEN  NA     3 1982   70000    7
## 70  PAS SINGUENSA JUAN JOSE  22    10 1980   70000    7
## 112        GUIYEN   TEODORO  NA    NA 1983   70000    7
## 144      MANOQUIN     JULIA  NA     3 1982   70000    7
```

Recall that we are only considering two municipalities in El Salvador now, which is what was considered in Sadinle (2014).

Blocking applied to the El Salvadoran conflict

In this assignment, you should explore deterministic and probablistic blocking methods and how these work on the El Salvadoran data set. Find at least one deterministic and probabilistic blocking criterion that seems suitable for this data set. Illustrate its effectiveness on the data using the reduction ratio, precision, and recall. Utilize other vizualiations that might also help you in explaining your results. How would you use these blocking methods to remove duplicate records in the data set?