Arrests 2010 Census Tracts analysis

```
##           used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 1978527 105.7    2929328 156.5  2929328 156.5
## Vcells 3431523  26.2    8388608  64.0  6544799  50.0
```

Preprocessing

Remove date and NTA variables, convert to factor location, month and KY_CD.

```
## 'data.frame':    419420 obs. of  14 variables:
##  $ KY_CD     : Factor w/ 70 levels "101","102","103",..: 70 4 70 15 70 70 70 4 70 4 ...
##  $ LAW_CAT_CD: Factor w/ 5 levels "","F","I","M",..: 1 2 4 2 4 2 4 2 4 2 ...
##  $ AGE_GROUP : Factor w/ 5 levels "<18","18-24",..: 3 2 3 1 4 4 3 2 2 3 ...
##  $ PERP_SEX  : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PERP_RACE : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 6 3 3 3 3 7 6 7 7 3 ...
##  $ geoid     : Factor w/ 2308 levels "36005000100",..: 236 636 1273 369 43 2187 983 1422 1592 1880 .
##  $ MONTH     : Factor w/ 12 levels "1","2","3","4",..: 1 11 3 12 12 12 11 11 10 9 ...
##  $ Pop1      : num  751 4544 183 2394 5337 ...
##  $ Male.P    : num  26.9 41.8 44.8 41.6 50.1 46.1 46.6 47.7 49.6 48.4 ...
##  $ MdAge     : num  20.4 31.4 30.8 32.2 30.4 33.8 36.8 34.2 35.1 37.3 ...
##  $ Hsp1P     : num  19.2 14 2.7 25 64.4 26.4 14.1 42.2 21.1 30.1 ...
##  $ WNHP      : num  61.5 1.2 39.9 19.8 3.2 20.5 3.8 8.8 53.3 17.7 ...
##  $ BNHP      : num  9.9 82.6 3.3 42.4 29 44.7 71.8 44.6 2.8 12.9 ...
##  $ ANHP      : num  7.1 0.2 49.7 10.3 1.8 4 7.6 1.5 19.1 33.1 ...
```

Check for NA

```
##      KY_CD LAW_CAT_CD  AGE_GROUP   PERP_SEX  PERP_RACE      geoid      MONTH
##          0          0          0          0          0          0          0
##       Pop1     Male.P      MdAge      Hsp1P       WNHP       BNHP       ANHP
##          0       3860       3796       4790       4410       4969       9686
```

A possibility is to get rid of all NAs rows, the portion of deleted rows would be relatively small (of course we're introducing some bias here).

```
## [1] 0.9752515
```

Other possibilities would be to impute values for numerical variables (using median, mean or more sofisticated methods). For simplicity we just delete missing values rows.

Description

Ideally 2010 data are our training set and 2011 data are the test set. The goal of the analysis is to identify is some covariates are correlated with the arrests rate: more specifically if the response is well explained by some non spatial covariates alone, some spatial alone or interaction between the two.

A reasonable response variable would be the count of arrests divided by the local (space zone) population, also grouping by any other covariates value.

To get an idea of the dataset used on which models are tested a

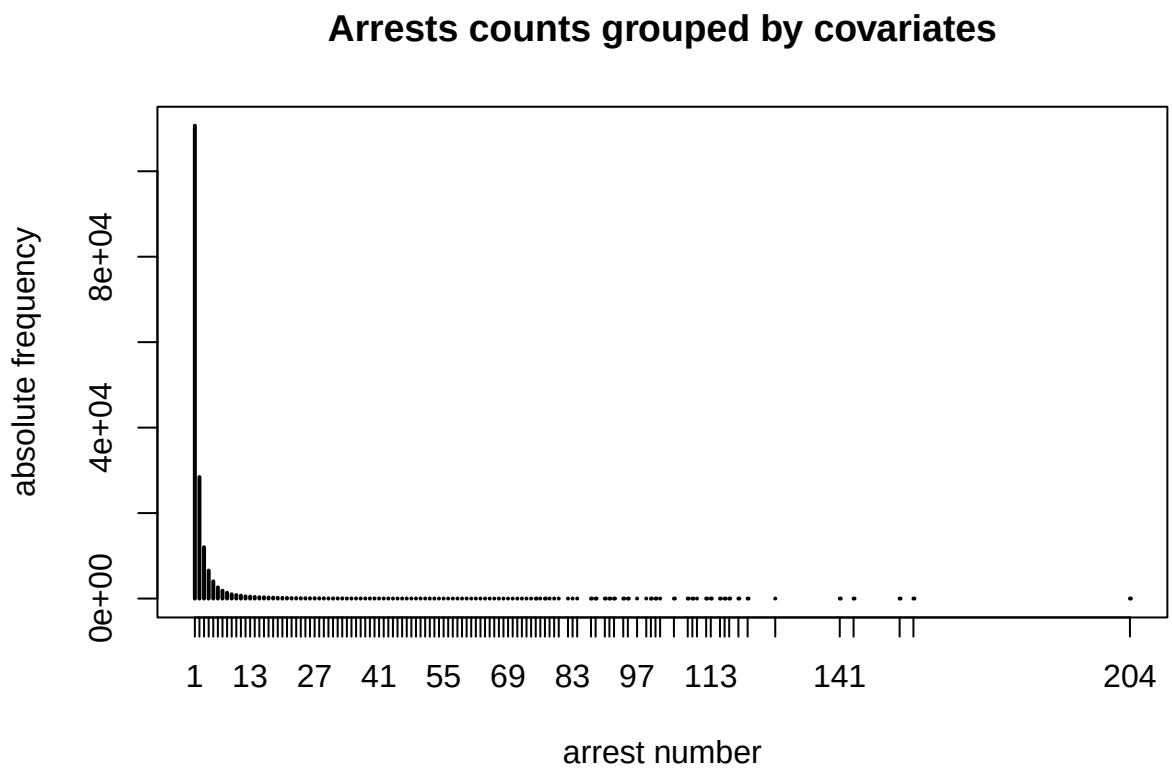
```
## `summarise()` has grouped output by 'KY_CD', 'LAW_CAT_CD', 'AGE_GROUP',
## 'PERP_SEX', 'PERP_RACE', 'geoid', 'Pop1', 'Male.P', 'MdAge', 'Hsp1P', 'WNHP',
## 'BNHP'. You can override using the `.groups` argument.
```

```
## [1] 174289     15
```

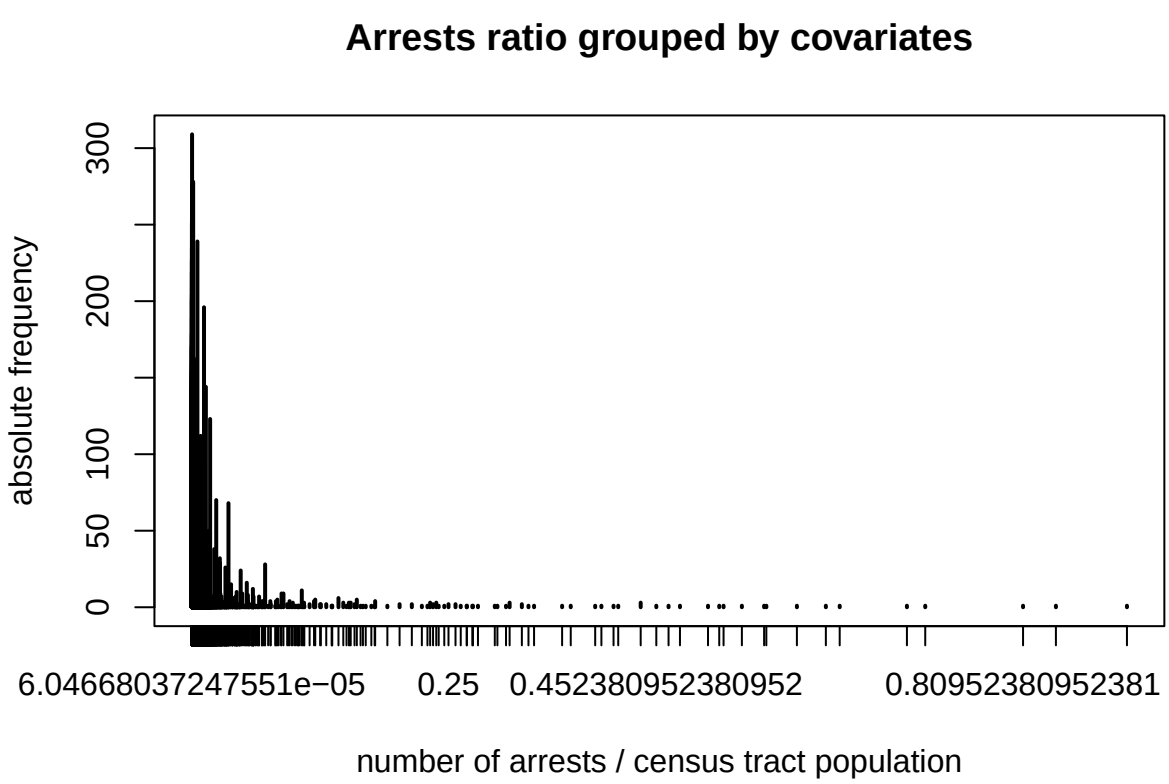
```
##  [1] "KY_CD"      "LAW_CAT_CD" "AGE_GROUP"  "PERP_SEX"   "PERP_RACE"
##  [6] "geoid"      "Pop1"       "Male.P"     "MdAge"      "Hsp1P"
## [11] "WNHP"       "BNHP"       "ANHP"       "count"      "y"
```

Still a huge number of observations compared to the number of variables, but what if we add interactions?

Let's look at the distribution of the counts



We can see an inflation of ones. The ratios present a similar table.



Let's count the hipothetical number of interaction terms if ones considers only interactions between spatial zones and selected arrests covariates along with the obervations / number of parameter ratio (underestimate since there are other variables):

```
##      KY_CD LAW_CAT_CD  AGE_GROUP   PERP_SEX  PERP_RACE      geoid       Pop1
##         70          5          5          2          7       2213       1835
##     Male.P      MdAge      Hsp1P       WNHP       BNHP       ANHP      count
##        204        305        679        762        672        492        112
##          y
##      12630
```

Not including KY_CD:

```
## geoid
## 42047
```

```
##  geoid
## 4.1451
```

Including KY_CD

```
##  geoid
## 196957
```

```
##     geoid
## 0.8849089
```

The ratio is already less than one. Given the inflation of ones it seems clear that some compromise has to be adopted one possibility is to not include interactions between KY_CD variable hoping that the less detailed LAW_CAT_CD variable would still give some insights. We decide to not employ the MONTH time variable as a covariate but use it for a model selection method.

Variables description

Original dataset selected variables:

Census stratification variables:

Explorative analysis

Models

Model selection method

Given the previously described constraints, in order to be able to apply a cross validation (CV) selection method we choose to ignore the time (MONTH) factor using MONTH as index to create the CV folds as described below. Choose k: the number of validation sets (example k = 4) each validation set is made by grouped observations of 12 / k (3) months and the months left are used to fit the model. To try to compensate and average for seasonal fluctuations the validation months are chosen as spaced as possible, for example, in the case k = 4 the first validation set is (january, may, september), the second set is (february, june, october), the third is (march, july, november) and the forth is (april, august, december); in order to make each response comparable having used a different number of months a new response is defined as the arrests ratio divided by the number of months used in the grouping.

Note on quantitative covariates

The simplest assumption is to assume a linear (monotone) trend of the response as a function of quantitative covariates.

LASSO

```
## Loaded glmnet 4.1-8
```

Elasticnet

Scad MCP

Grouped LASSO