<http://franklincenterhq.org/2536/analysis-of-missouri-sex-offender-registry-data/>

\FranklinCenter\Projects\2011\KSHB-Sex-Offenders

**Analysis of Missouri Sex Offender Registry Data**

By Earl Glynn Monday, July 18th, 2011

This is the first of four articles about analyzing distances between sex offenders and child daycare centers in Missouri as part of a [*joint project with KSHB NBC Action News in Kansas City*](http://www.nbcactionnews.com/dpp/news/local_news/investigations/loophole-in-law-allows-hundreds-of-missouri-sex-offenders-to-live-near-church-daycares).

The [Missouri State Highway Patrol](http://www.mshp.dps.missouri.gov/) maintains the online [Missouri Sex Offender Registry](http://www.mshp.dps.mo.gov/MSHPWeb/PatrolDivisions/CRID/SOR/SORPage.html) to provide a way for residents to find the location of sex offenders through searches or by maps.



*The Missouri State Highway Patrol maintains the online Missouri Sex Offender Registry*

In addition, the Patrol provides a single file that can be downloaded for analysis purposes.  Find the [link to download the Missouri offender file near the bottom of this page](http://www.mshp.dps.mo.gov/MSHPWeb/PatrolDivisions/CRID/SOR/SORPage.html). The **msor.xls** Excel file can be extracted from the **msor.zip**.

When opened in Excel the summary at the top indicated the file from June 20, 2011 had information for 15,420 offenders.  But the file had over 18,000 lines?

Inspection of the names and addresses showed the file was a list of offenses and there are multiple lines in the file for some offenders with one line per offense.  For example, there were separate lines for Roma J. Abel in Sullivan, MO for two offenses:  “Child Molestation-1st” and “Indecent Acts with Child.”

**Offender List from Offenses List**

We only wanted to analyze or plot a single point per offender, so the “extra” few thousand duplicate lines were a bit of a problem.  Deleting these extra lines manually in Excel would be quite tedious, and combining the individual offenses into a single string with the offenses for each offender with more than one would be a pain.

These manipulations on the data from the Excel file would be fairly easy using the [R statistical analysis language](http://www.r-project.org/). See the file [Missouri-Sex-Offenders.R](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/Missouri-Sex-Offenders.R) for details about these manipulations.

**R Script**

R can “read” an Excel file and manipulate the data programatically instead of manually with Excel.

The [R package RODBC](http://cran.r-project.org/web/packages/RODBC/index.html) can be used to read Excel files (or other files including various kinds of databases), but my first attempt to read the original msor.xls file failed.  After some inspection of the file, I discovered it was internally an XML file and not really the older Excel “XLS” format.

Since we needed to delete the top 13 lines of summary information anyway, after that deletion in Excel we saved the data to a new Excel file named **Missouri-Sex-Offenders.xls**.  R would read this file.

Three R statements read the entire Excel file into an R object called msor:

connection <- odbcConnectExcel("Missouri-Sex-Offenders.xls")  
msor <- sqlFetch(connection, "Sheet1", as.is=TRUE)  
odbcClose(connection)

[Note:  With an Excel 2007 or 2010 .xlsx file, change “odbcConnectExcel” to “odbcConnectExcel2007″.]

We can get a quick idea of the number of rows and columns, column names and structure of the data with these R statements:

**> dim(msor)**  
[1] 18234 9  
> # 18234 9

**> colnames(msor)**  
[1] "Name" "Address" "City" "St" "Zip" "County" [7] "Offense" "Count" "Compliant"

**> str(msor)**  
'data.frame': 18234 obs. of 9 variables:  
$ Name : chr "AARON, JEFFERY W" "ABBEY, DONALD S" "ABBOTT, JAMES E" "ABBOTT, RODNEY L" ...  
$ Address : chr "371 YEARGAN LN" "ADDRESS UNKNOWN" "940 HWY 5" "25466 HIGHWAY 5" ...  
$ City : chr "PINEVILLE" "ST CHARLES" "MANSFIELD" "LEBANON" ...  
$ St : chr "MO" "MO" "MO" "MO" ...  
$ Zip : chr "64854" "63301" "65704" "65536" ...  
$ County : chr "MCDONALD" "ST. CHARLES" "WRIGHT" "LACLEDE" ...  
$ Offense : chr "SEXUAL ABUSE IN THE SECOND" "SEXUAL MISCONDUCT-1ST" "RAPE 1ST DEGREE" "STAT SODOMY-1ST DEG-PERS UND 14" ...  
$ Count : num 1 1 1 2 1 1 1 3 1 1 ...  
$ Compliant: chr "Y" "Y" "Y" "Y" ...

The Excel data can be recognized in R.

After studying the summary information and the data, we discovered a number of “problem” address fields.

This R statement creates a table of the counts of all unique addresses, sorts that table in descending order, and displays the top 10 by frequency:

**> sort(table(msor$Address), decreasing=TRUE)[1:10]**

Incarcerated in MO  
2403  
Previously Exempt/Registration Required  
749  
Moved Out of State  
593  
Incarcerated Out of State  
255  
1621 N 1ST ST  
190  
651 MULBERRY ST  
154  
HOMELESS  
106  
UNKNOWN  
94  
1108 E 10TH ST  
52  
651 MULBERRY  
43

Obviously some of these addresses could not be geocoded, like “Homeless” or “Unknown”.

A delete.list was defined in R to list addresses that were not of interest.  A select boolean vector, with only True/False values, was defined for each msor row.  The value was “True” if the row was an invalid address, and “False” otherwise.

After defining select, we create an R data.frame with the 4203 problem addresses from 3543 offenders and write them to a file, [msor-problem-addresses.csv](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/msor-problem-addresses.csv), for further study.

delete.list <- c("Incarcerated in MO",  
"Previously Exempt/Registration Required",  
"Moved Out of State",  
"Incarcerated Out of State",  
"HOMELESS",  
"UNKNOWN",  
"UNKNOWN ADDRESS")

select <- msor$Address %in% delete.list

problem.addresses <- msor[select,]

write.csv(problem.addresses, "msor-problem-addresses.csv",  
row.names=FALSE)

Since select is a boolean vector, we can use !select (not select) to make a new object offense with the remaining offenses.

Next, the 198 out-of-state offender offense records can be written to another file [msor-out-of-state.csv](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/msor-out-of-state.csv) for analysis and removed from the offense object:

offense <- msor[!select,]

table(offense$St)

AR IA IL IN  KS    MO MS OK PA TN  
7  5 63  1 102 13833  1 15  1  3

select <- offense$St != "MO"

out.of.state <- offense[select,]

write.csv(out.of.state, "msor-out-of-state.csv", row.names=FALSE)  
offense <- offense[!select,]

After removing the out-of-state offenders, we create a file of how many offenses of what type are remaining in the analysis and sort it into descending order by frequency count:

counts <- sort(table(offense$Offense, useNA="ifany"), decreasing=TRUE)  
counts <- data.frame(Offense=names(counts), Count=as.numeric(counts))  
write.csv(counts, "msor-offender-offenses-counts.csv", row.names=FALSE)

The file [msor-offender-offenses-counts.csv](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/msor-offender-offenses-counts.csv) shows over 1200 different offenses — the number would be reduced a bit by standardizing the entries and spellings.

The top 20 most common offenses are:

|  |  |
| --- | --- |
| **Offense** | **Count** |
| SODOMY | 1188 |
| STATUTORY RAPE-2ND DEGRE | 932 |
| RAPE | 888 |
| SEXUAL ASSAULT | 739 |
| CHILD MOLEST-1ST DEGREE | 733 |
| STATUTORY SODOMY-2ND DEG | 726 |
| SEXUAL MISCONDUCT-1ST | 524 |
| SEXUAL ABUSE 1ST DEGREE | 523 |
| STATUTORY SODOMY-1ST DEG | 491 |
| SEXUAL ABUSE | 480 |
| DEVIATE SEXUAL ASSAULT | 400 |
| STATUTORY RAPE-1ST DEGRE | 374 |
| CHILD MOLEST-2ND DEGREE | 365 |
| POSSESSION OF CHILD PORNOGRAPHY | 212 |
| SEX MISCON-CHILD-1ST OFF | 154 |
| STAT SODOMY-1ST DEG-PERS UND 14 | 151 |
| INCEST | 150 |
| SEXUAL MISCONDUCT-2ND | 141 |
| STAT RAPE-1ST DEG-PERS UNDER 14 | 136 |
| CHILD MOLESTATION-1ST | 134 |

Now after deleting all the addresses that are out-of-state, or are invalid for geocoding,we use R’s unique function to eliminate all the duplicate records.  This results in a unique list of offenders.

The geocode object is a copy of the offender object with some changes needed for the geocoding process:

offender <- offense[,c("Name", "Address", "City", "St", "Zip", "County", "Compliant")]

offender <- unique(offender)

offender <- data.frame(N=1:nrow(offender), offender)

geocode <- offender  
geocode$status <- ""  
geocode$zip <- ""  
geocode$county <- ""  
geocode$state <- ""  
geocode$lat <- ""  
geocode$lng <- ""  
geocode$location.type <- ""  
geocode$formatted.address <- ""  
geocode$result.count <- ""

write.table(geocode, file="geocode-MO-offender-in-MASTER.txt",  
sep="\t", quote=FALSE, row.names=FALSE)

The 11,702 offender records in the tab-delimited [geocode-MO-offender-in-MASTER.txt](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/geocode-MO-offender-in-MASTER.txt) file will be the starting point for geocoding in the next article.

The R script shows the details of creating a file with one line per offender with a “Comments” column showing all the offenses for the offender: [msor-offender-master-file.csv](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/msor-offender-master-file.csv).

The Missouri-Sex-Offenders.R script created some additional files of summary information, including:

* table of sex offender registration compliance by county,  
  [msor-offenders-compliance-by-county.csv](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/msor-offenders-compliance-by-county.csv)
* list of addresses with 3 or more sex offenders (at least some are jails),  
  [msor-more-than-2-offenders-at-address.csv](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/msor-more-than-2-offenders-at-address.csv)

**Geocoding**

Next:  [Geocoding addresses from Missouri Sex Offender Registry](http://www.franklincenterhq.org/2541/geocoding-addresses-from-missouri-sex-offender-registry/)

**Program and Files**

R program:  [Missouri-Sex-Offenders.R](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/Missouri-Sex-Offenders.R).  Script to process sex offenses file from Missouri Highway Patrol and reduce it to the unique list of offender addresses that are likely to geocode to the State of Missouri. (you will want to modify the working directory in line 25 to your working directory)

Input File:  Missouri-Sex-Offenders.xls ([original file from Missouri State Highway Patrol web site](http://www.mshp.dps.mo.gov/MSHPWeb/PatrolDivisions/CRID/SOR/SORPage.html) can be downloaded  — see bottom of that page —  and modified as described above to be the input file for the script).

Output File:  [geocode-MO-offender-in-MASTER.txt](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/geocode-MO-offender-in-MASTER.txt).  This file is the starting point for the geocoding process that will add latitude/longitude coordinates to an address.

**Related**

* [Loophole in law allows hundreds of Missouri sex offenders to live near church daycares](http://www.nbcactionnews.com/dpp/news/local_news/investigations/loophole-in-law-allows-hundreds-of-missouri-sex-offenders-to-live-near-church-daycares), KSHB NBC Action News, Kansas City, July 18, 2011.
* [Geocoding addresses from Missouri Sex Offender Registry](http://www.franklincenterhq.org/2541/geocoding-addresses-from-missouri-sex-offender-registry/), Franklin Center CAR Blog, July 19, 2011.
* [Computing distance matrix between Missouri sex offenders and child daycare facilities](http://www.franklincenterhq.org/2544/computing-distance-matrix-between-missouri-sex-offenders-and-child-daycare-facilities/), Franklin Center CAR Blog, July 28,  2011.
* [Displaying Missouri sex offender/child day care facility proximity map using batchgeo.com](http://www.franklincenterhq.org/2563/displaying-missouri-sex-offenderchild-day-care-facility-proximity-map-using-batchgeo-com/), Franklin Center CAR Blog, July 28,  2011.

**Earl F Glynn • KansasWatchdog.org • Franklin Center for Government and Public Integrity**