# Burglaries in Chicago
# Near-repeat analysis

Alberto Garrido, Lin Xia, Louise Geraghty and Manuel Aragonés

## Presentation outline:

I. Descriptive statistics
II. Background on Spatial-Temporal analysis
III. Knox implementation
IV. Results
V. Conclusions

# Descriptive statistics

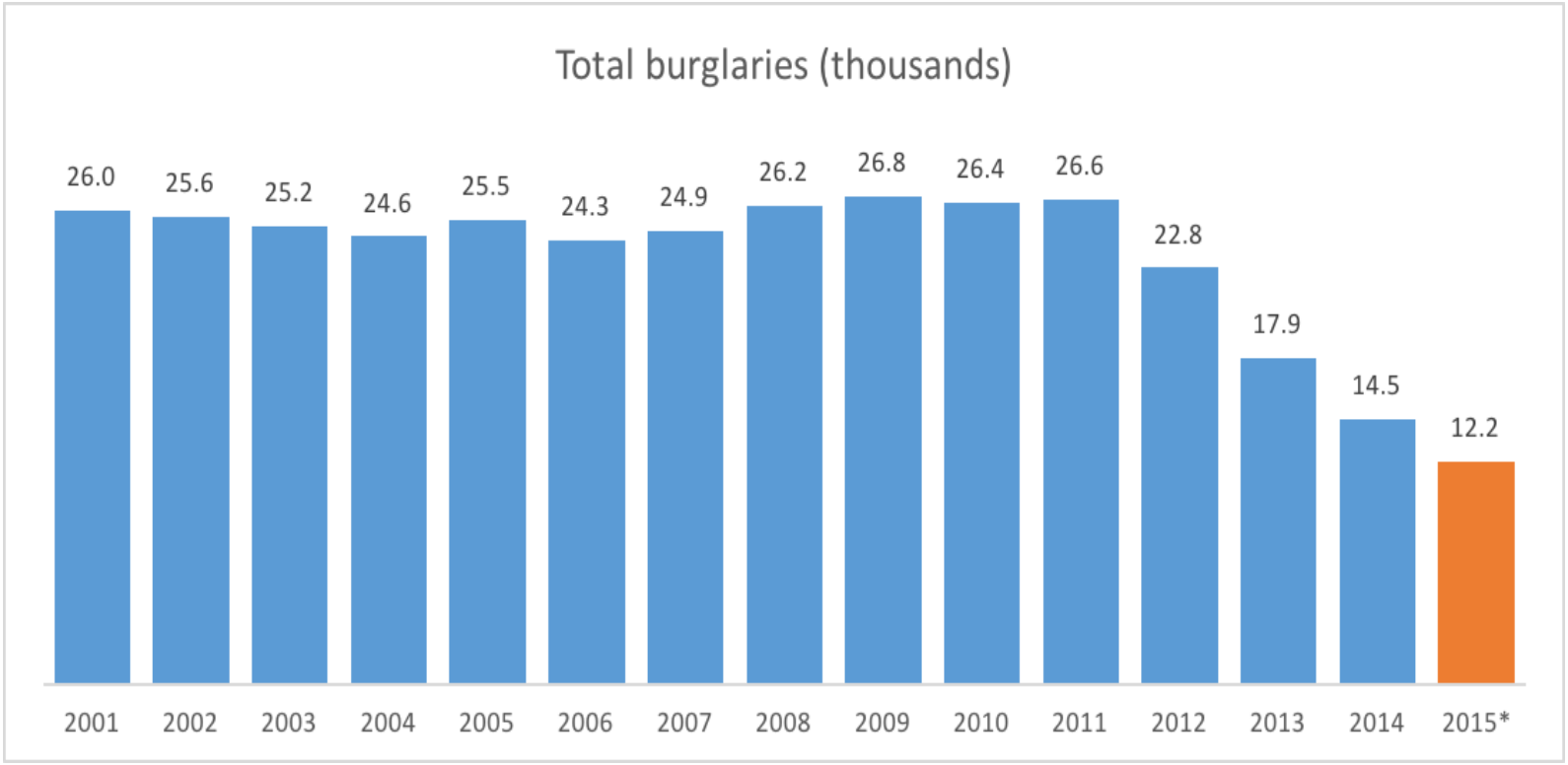
# Data in brief

- Data:
  - 339,778 observations
  - IUCR
  - Primary.Type
  - Description
  - Location

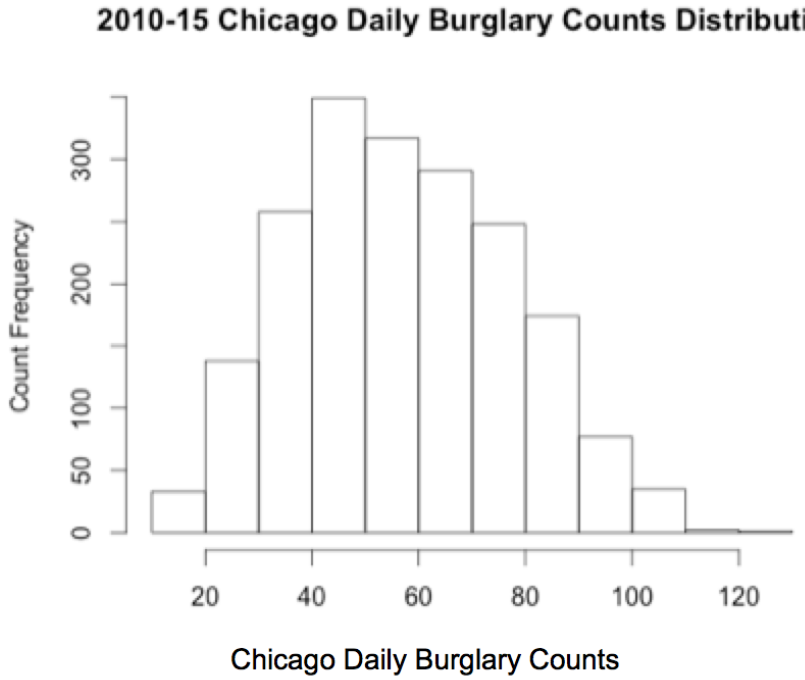
- Spatial
  - Latitude & longitude
  - # Blocks: 37,575
  - # Beats: 281
  - # CAs: 77
  - # Wards: 50
  - # Districts: 24

- Temporal:
  - 2001-to date (2001 without spatial)
  - dym_hm

# The number of burglaries has decreased



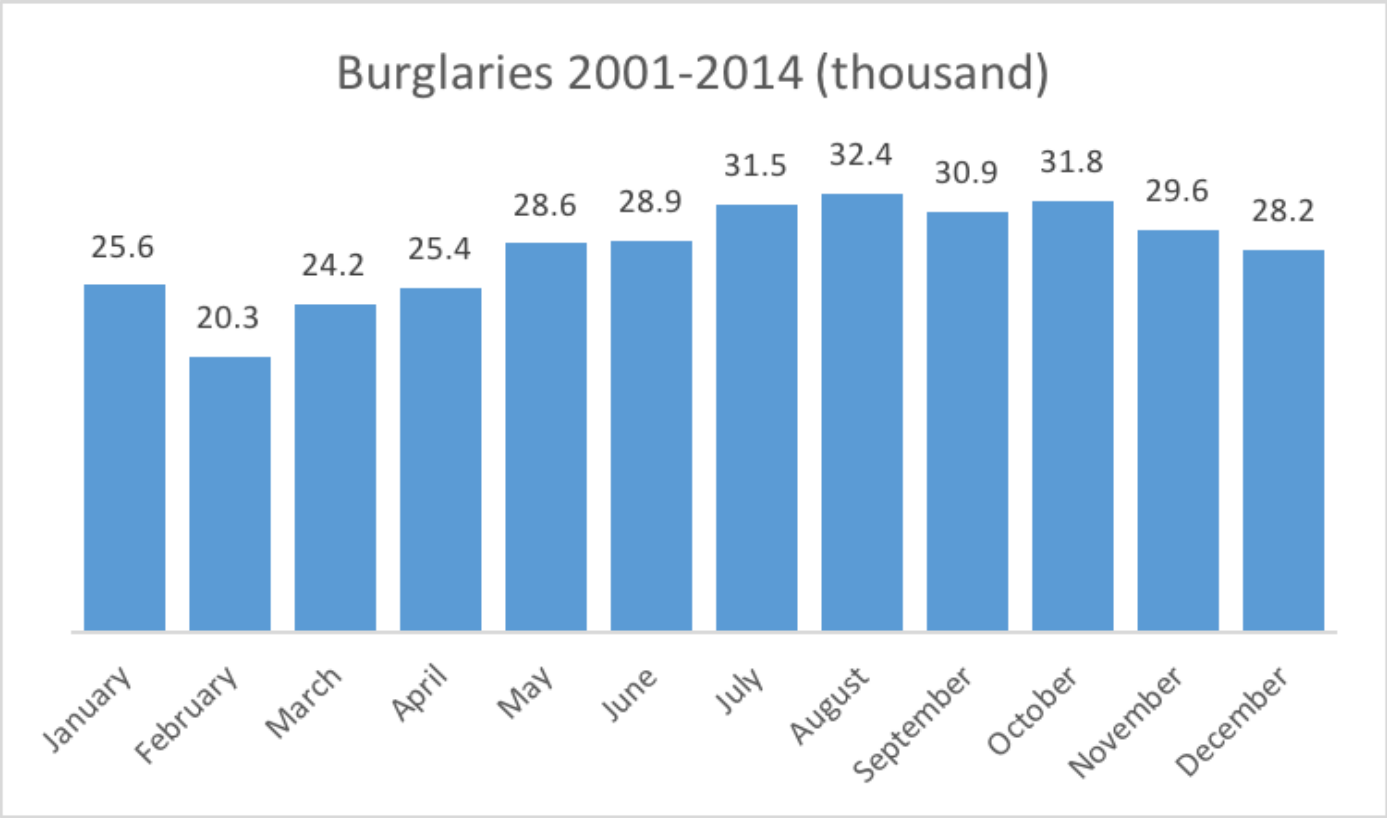
# More than 57 burglaries each day in Chicago



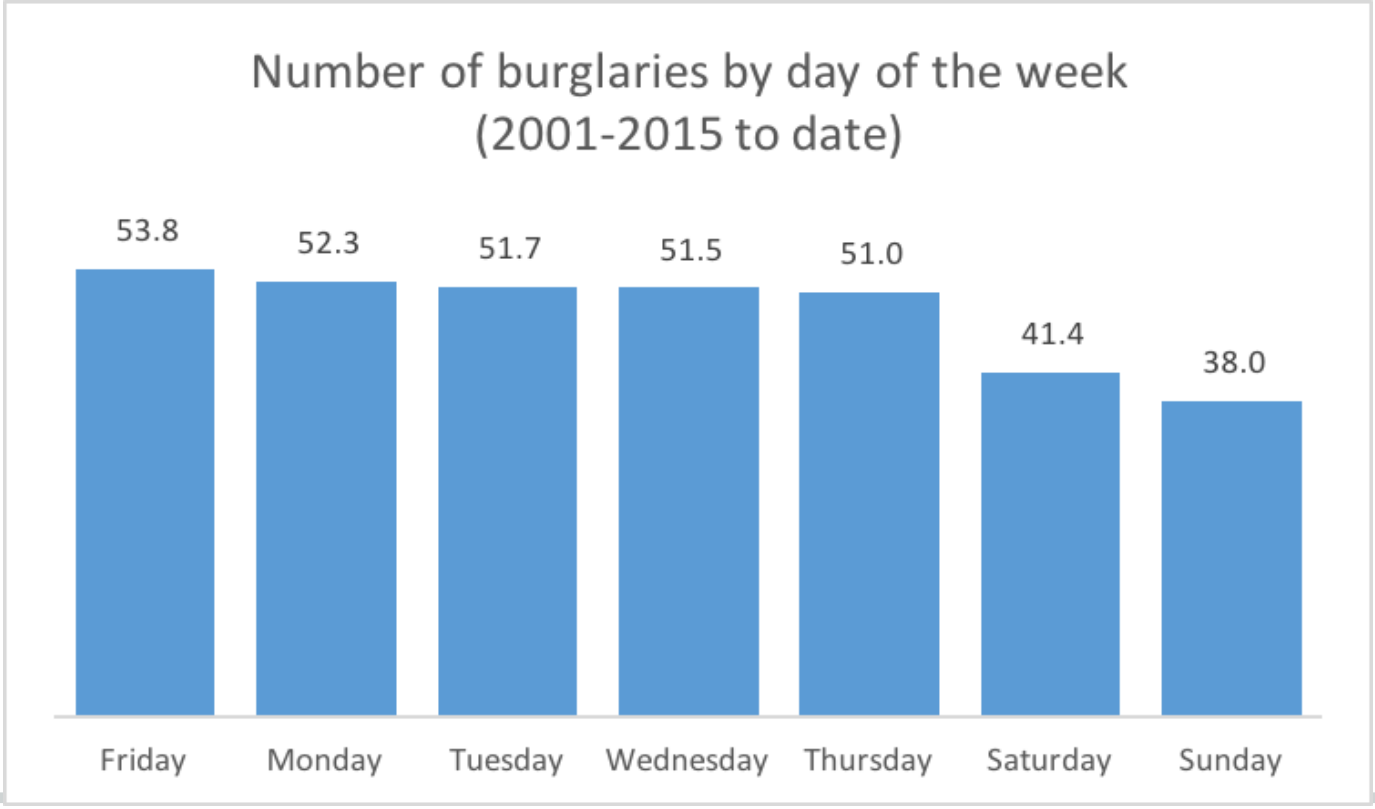

In Chicago from 1/01/2010-04/07/2015:

- Mean daily burglary counts in this time period is 57.5
- Standard deviation is 20.27
- Minimum number is 13 per day
- Maximum is 124 per day

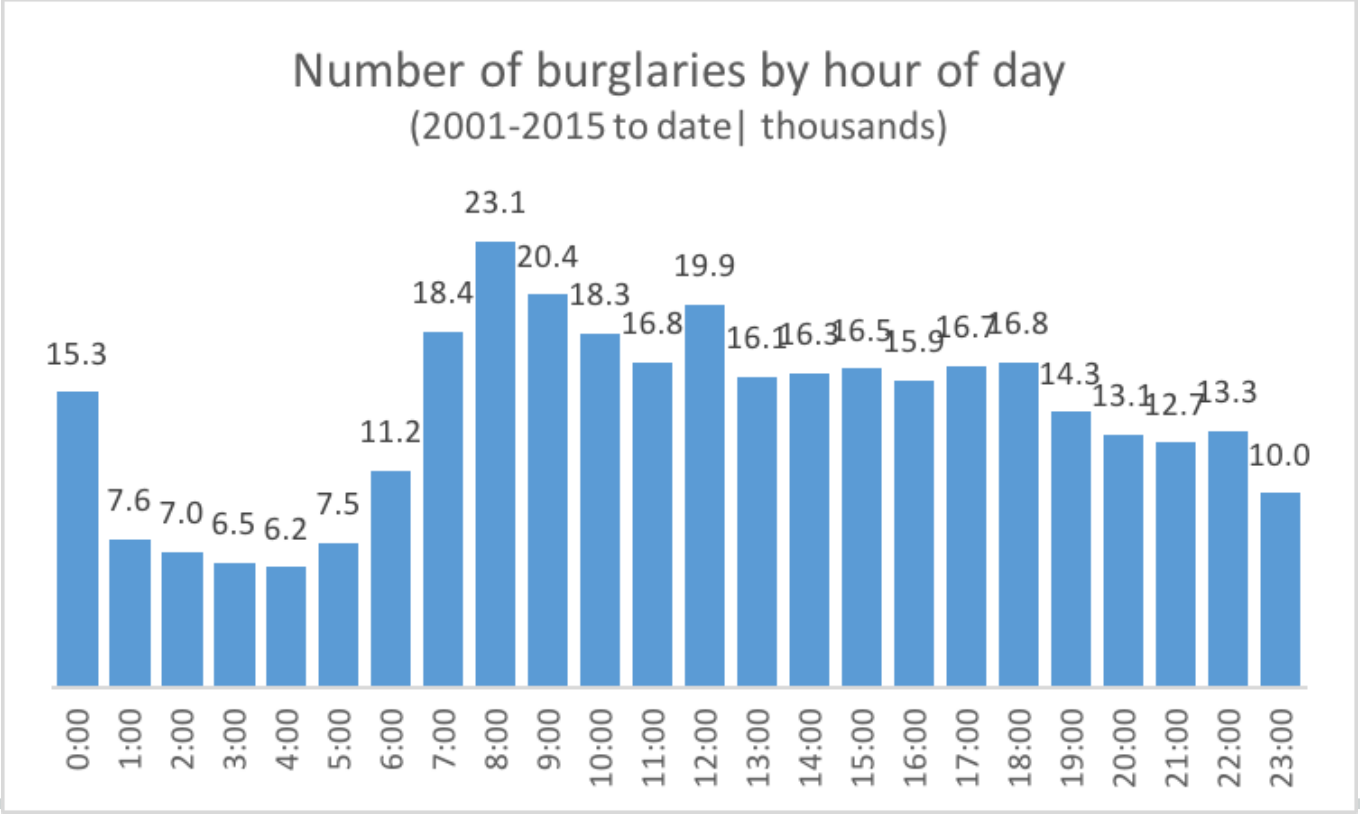
# July-October, the highest burglary rates



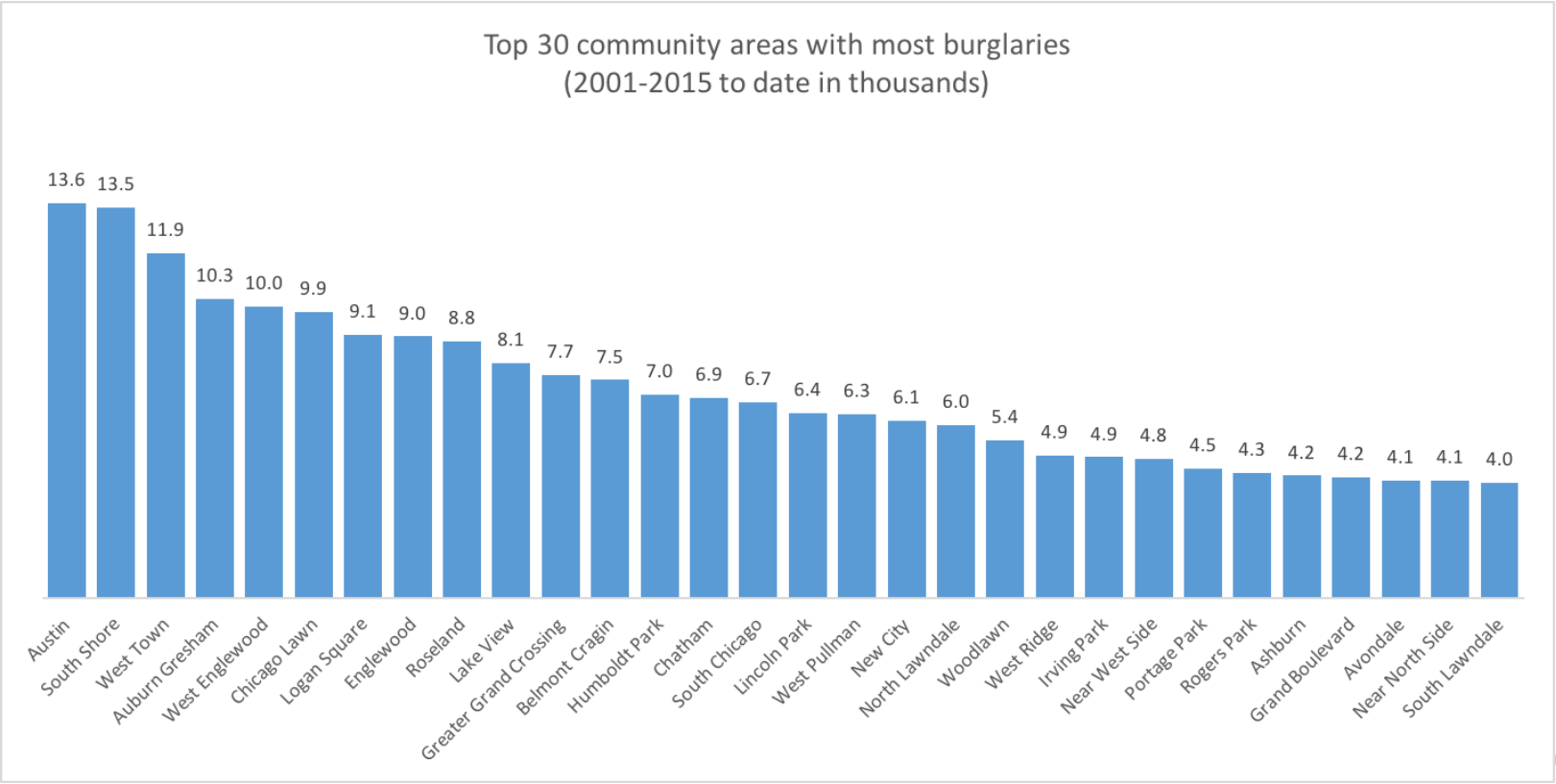
# Saturday and Sunday the days with less burglaries



# Most burglaries happen during the day



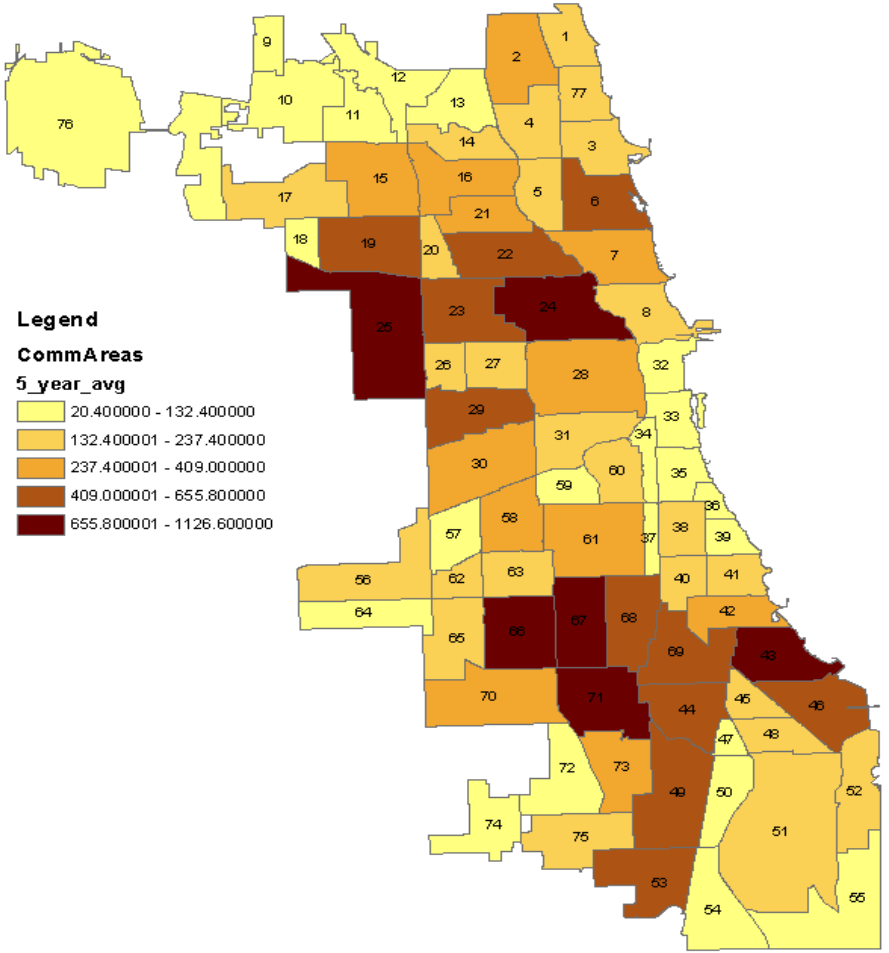
# The 30 community areas with most burglaries



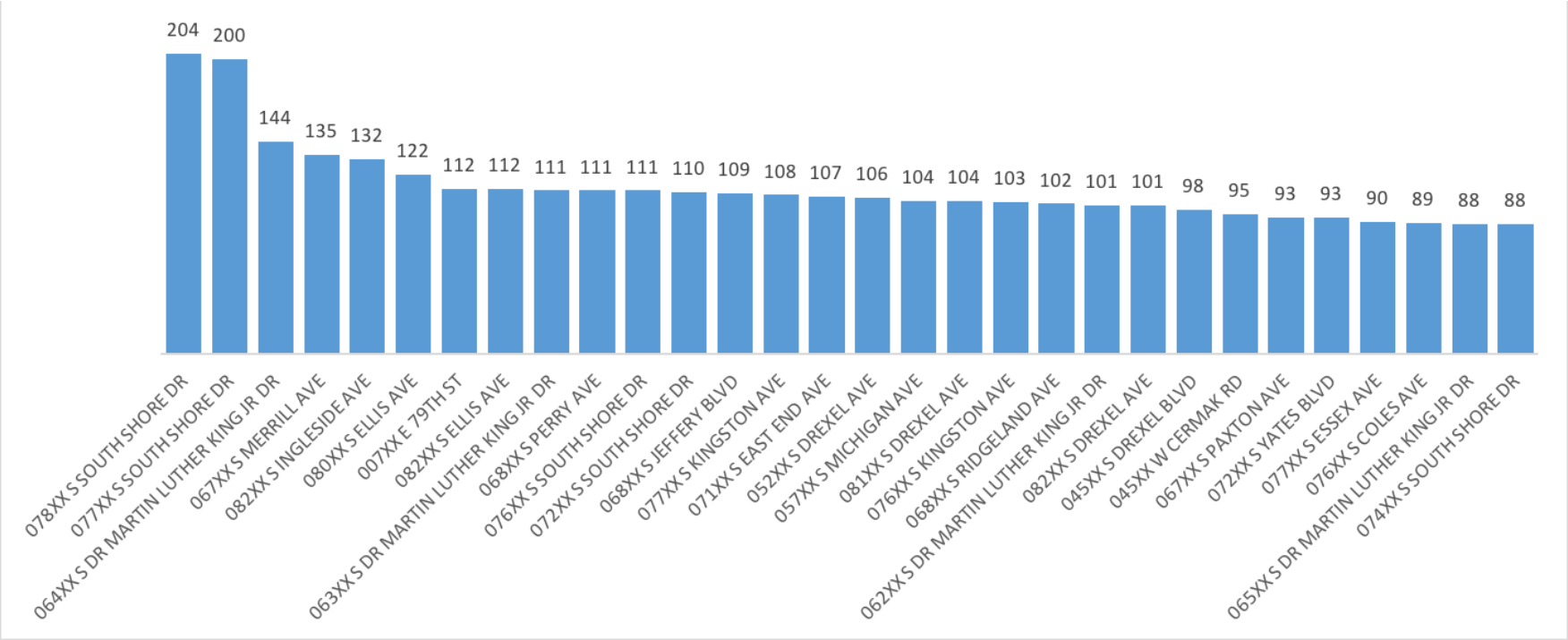
# Yearly Burglaries by CA

Top 6 communities with highest yearly average burglaries from 2010-15:

1. South shore: 1126.6
2. Austin: 1067.4
3. West Town: 816.2
4. Chicago Lawn: 738
5. West Englewood: 737.8
6. Auburn Gresham: 715.8



# The 30 blocks with most burglaries



# Background on Spatial-Temporal analysis

# Detection of Spatio-Temporal clusters

**A**utomatically detect regions of space that are “anomalous,” “unexpected,” or otherwise “interesting.” Finding regions of space where the values of some quantity (the “count”) are significantly higher than expected, given some other “baseline” information.

1. What is Spatial and Spatio-Temporal cluster analysis
    - **Spatial cluster detection ≠ cluster detection**: with spatial cluster detection we try to **find regions where** some quantity is significantly higher than expected, **adjusting for the underlying population** or baseline. Cluster detection just tries to find groups of data points.
    - **Spatial cluster detection ≠ anomaly detection**: with anomaly detection you focus on **single data points**, and asks whether it is normal or not. In spatial cluster detection you focus on **finding anomalous spatial groups or pattern**s.

2. Identify the locations, shapes, sizes, and other parameters of potential clusters

3. Determine whether each of these anomalous regions is due to a genuine and relevant cluster, or simply a chance occurrence:
    - Are they **significant**?
    - Testing: to tell which are likely to be “true” clusters and which are likely to have **occurred by chance**.

# Spatial scan statistic

1. Detect spatial regions (Knox-test):
   - Where underlying count **rate are significantly higher** inside the region than outside the region:
     - $F(S)=Pr(Data|H_1(s))/Pr(Data|H_0(s))$

2. Identify the most significant regions:
   - The *F(S)* score is calculated for each region.
   - The region with the highest score *F(S)* is the “most significant region”
   - Randomization Testing: Monte Carlo assuming $C_{all}/P_{all}$~Uniformly

3. Limitations:
   - Does not scale well
   - Low power to detect elongated regions
   - Limited to Binomial and Poisson

# Knox implementation

# Using Knox to forecast risk areas

## Objective

- Measure near repeat with Knox statistic.
- Obtain the $F(S)=Pr(Data|H_1(s))/Pr(Data|H_0(s))$ score
- We want to allow law enforcement authorities to allocate resources in a manner that allows them to prevent future burglaries, or apprehend burglars.

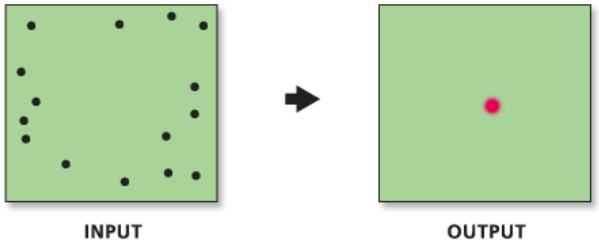
# Knox implementation

## Estimating Knox ratio:

- Pr(Data|$H_1$(s))/Pr(Data|$H_0$(s))
- Knox hat/Expected knox

## New method

1. The geographic centroid (in XY Coordinates) to the polygons of census tracts & blocks
2. From Shapefile to Dataframe
3. From Dataframe to a subset Dataframe
4. From Subset Dataframe to array (1 to 77 or each block)
5. Then run Knox for every array
6. Export results



# Using Knox to forecast risk areas

## The approach

1. Estimate Knox Matrix for each CA (Knoxhat/Knoxbar)
2. Prepare a time-distance matrix with burglaries information for each block
3. Assume the value from the CA is equal to the block
4. Multiply Knox Matrix by Time-distance Matrix
5. Obtain the score for each block
6. Rank the blocks by score

# Example of Knox ratio + score for a specific block

| | | | A) Information for the past 14 days in a specific block | | | | | B) Knox Statistic for CA | | | | | AxB | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CA | Block ID | days / ft | 600 | 1200 | 1800 | 2400 | 3000 | 600 | 1200 | 1800 | 2400 | 3000 | 600 | 1200 | 1800 | 2400 | 3000 |
| 5 | 514002000 | t-1 | 0 | 0 | 0 | 0 | 0 | 2.21 | 1.27 | 0.96 | 0.80 | 1.02 | - | - | - | - | - |
| 5 | 514002000 | t-2 | 0 | 0 | 0 | 0 | 0 | 0.58 | 1.20 | 1.04 | 0.99 | 1.08 | - | - | - | - | - |
| 5 | 514002000 | t-3 | 0 | 0 | 0 | 0 | 0 | 0.82 | 1.11 | 0.92 | 0.89 | 0.96 | - | - | - | - | - |
| 5 | 514002000 | t-4 | 0 | 0 | 0 | 0 | 0 | 0.52 | 1.00 | 0.95 | 0.95 | 0.99 | - | - | - | - | - |
| 5 | 514002000 | t-5 | 0 | 0 | 0 | 0 | 1 | 0.54 | 0.99 | 0.99 | 0.93 | 0.97 | - | - | - | - | 0.97 |
| 5 | 514002000 | t-6 | 0 | 0 | 0 | 0 | 1 | 1.19 | 0.98 | 1.05 | 0.96 | 1.00 | - | - | - | - | 1.00 |
| 5 | 514002000 | t-7 | 0 | 0 | 0 | 0 | 1 | 0.78 | 0.90 | 0.98 | 0.87 | 0.91 | - | - | - | - | 0.91 |
| 5 | 514002000 | t-8 | 0 | 0 | 0 | 0 | 1 | 0.60 | 0.94 | 1.04 | 0.95 | 1.00 | - | - | - | - | 1.00 |
| 5 | 514002000 | t-9 | 0 | 0 | 0 | 0 | 1 | 0.88 | 0.96 | 1.00 | 0.95 | 1.00 | - | - | - | - | 1.00 |
| 5 | 514002000 | t-10 | 0 | 0 | 0 | 0 | 1 | 0.76 | 0.94 | 0.98 | 0.95 | 0.98 | - | - | - | - | 0.98 |
| 5 | 514002000 | t-11 | 0 | 0 | 1 | 0 | 1 | 0.68 | 0.93 | 1.01 | 0.99 | 0.99 | - | - | 1.01 | - | 0.99 |
| 5 | 514002000 | t-12 | 0 | 0 | 1 | 0 | 1 | 0.57 | 0.92 | 0.99 | 0.97 | 0.95 | - | - | 0.99 | - | 0.95 |
| 5 | 514002000 | t-13 | 0 | 0 | 1 | 0 | 1 | 0.99 | 0.92 | 0.97 | 0.99 | 0.96 | - | - | 0.97 | - | 0.96 |
| | | | | | | | | | | | | | | | | | **11.73** |

# Using Knox to forecast risk areas

## Advantages

- Our method lets you use Knox statistic to estimate near repeat
- It incorporates the near repeat probability and ranks based on the most recent information
- Scalability
- “Speed”

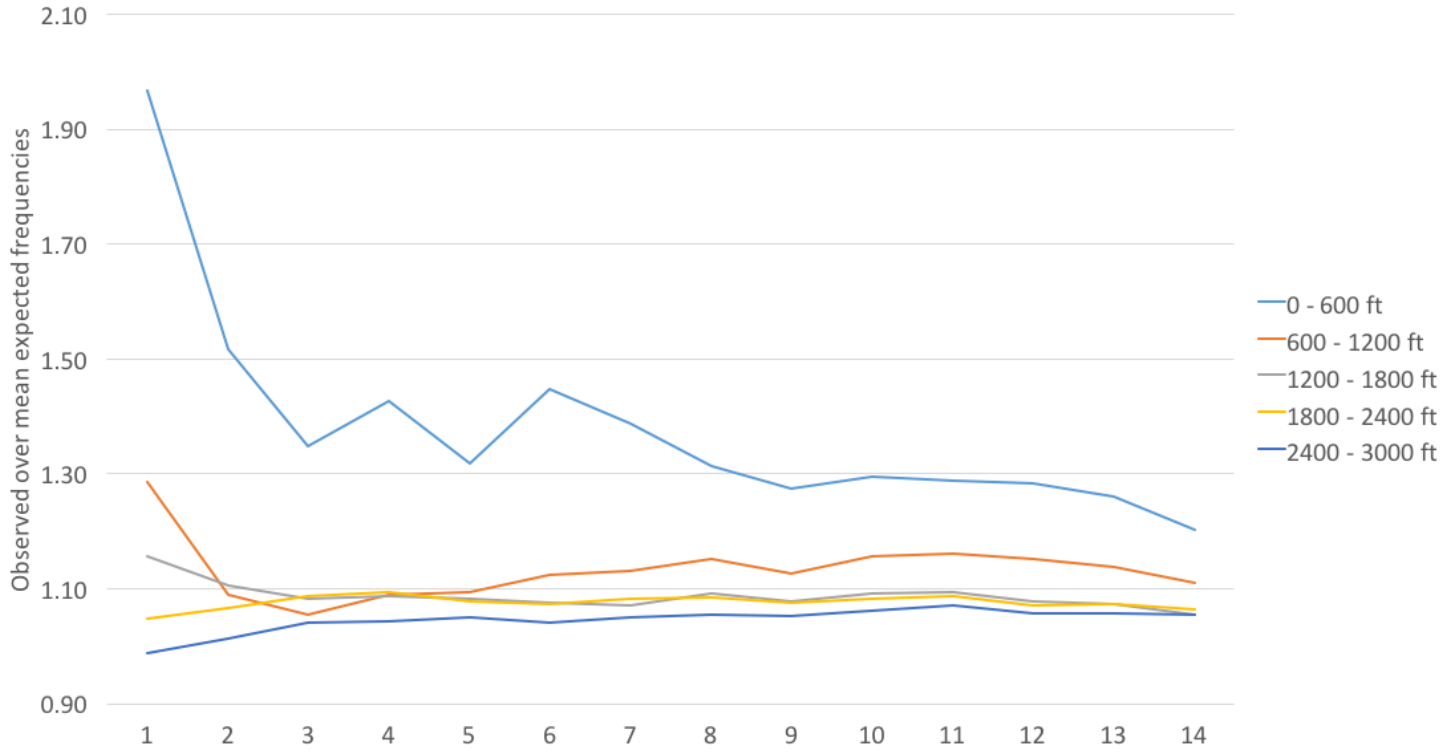
## Disadvantages

- Trade off (time spaces vs knox): we want precise estimates, but if we disaggregate time and/or space too much, we increase Knox’s standard error (or simply does not run).
- We have to measure out of sample deviance to see how good it predicts
- It is solely based on past burglaries information, we have not incorporated other crimes or socio economic variables

# Results

# Expected frequencies

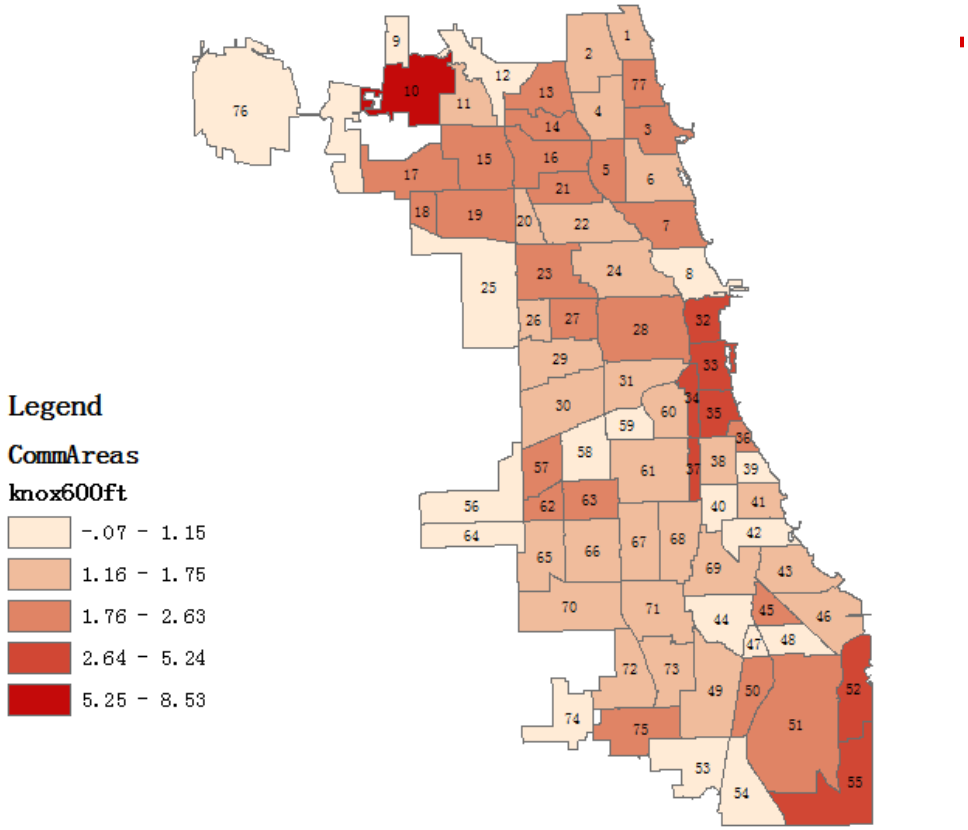
**(All CA's; 30 days)**



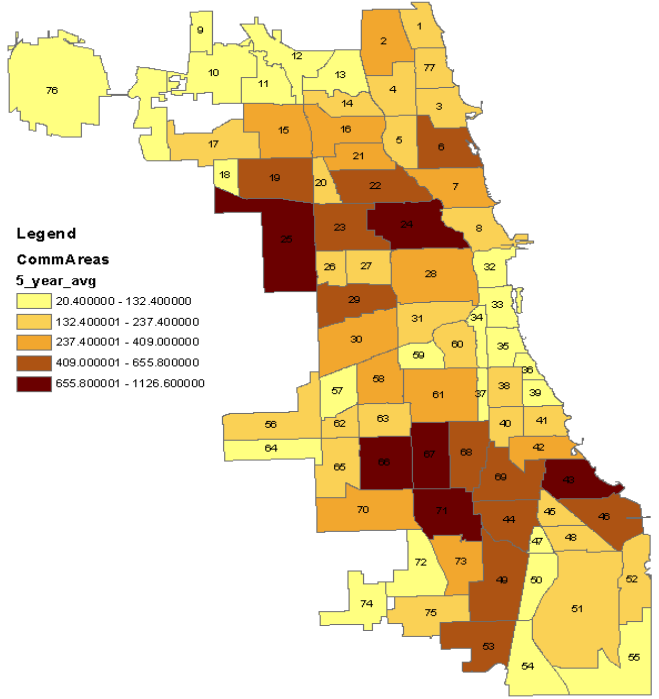
# Knox by CA

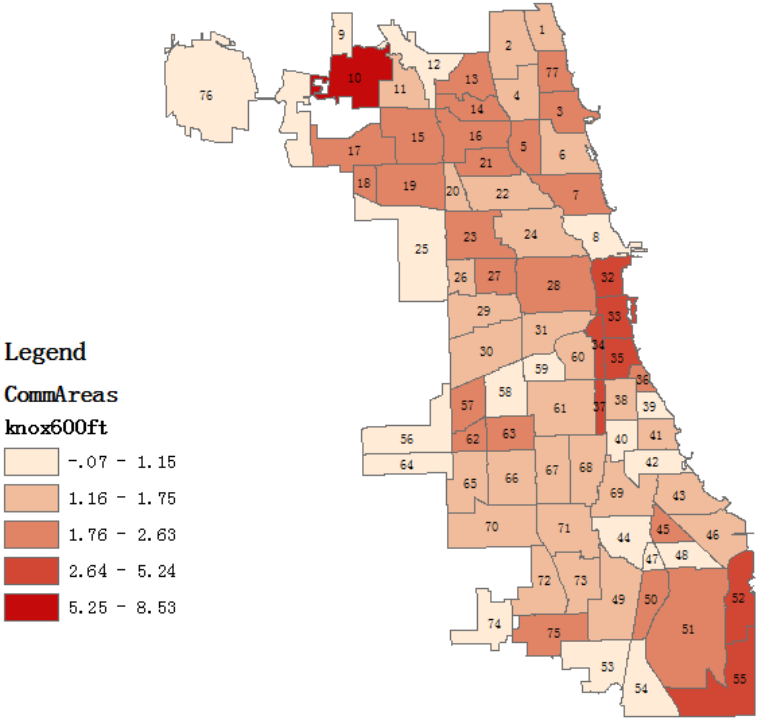
Communities with highest Knox index for 1 day and 600ft:

1. Norwood Park: 8.53
2. Near South Side: 5.24
3. Douglas: 4.49
4. Hegewisch: 3.74
5. Armour Square: 3.26
6. East Side: 3.17
7. Fuller Park: 3.16
8. Loop: 2.95



# Average burglary rate vs Knox ratio by CA



# Conclusions

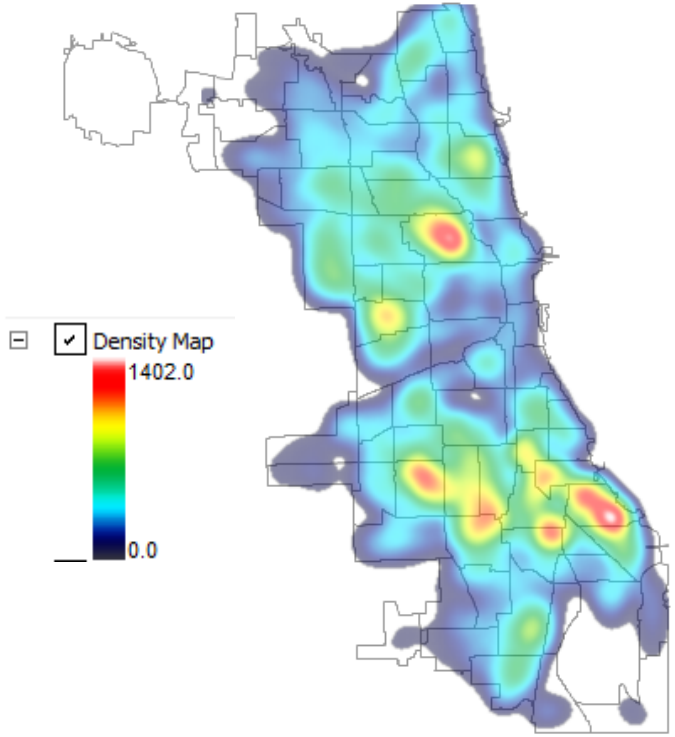
# Conclusions

1. Knox statistic allows you to identify community areas with near repeat patterns
2. Knox decay is faster in distance than by day (could be the units)
3. With the correct XY coordinates, the analysis can be more precise
4. The Knox ratio can be estimated each day with the latest information
5. Further analysis should take advantage of non-parametric methods for estimating spatial-temporal clusters (Neill)

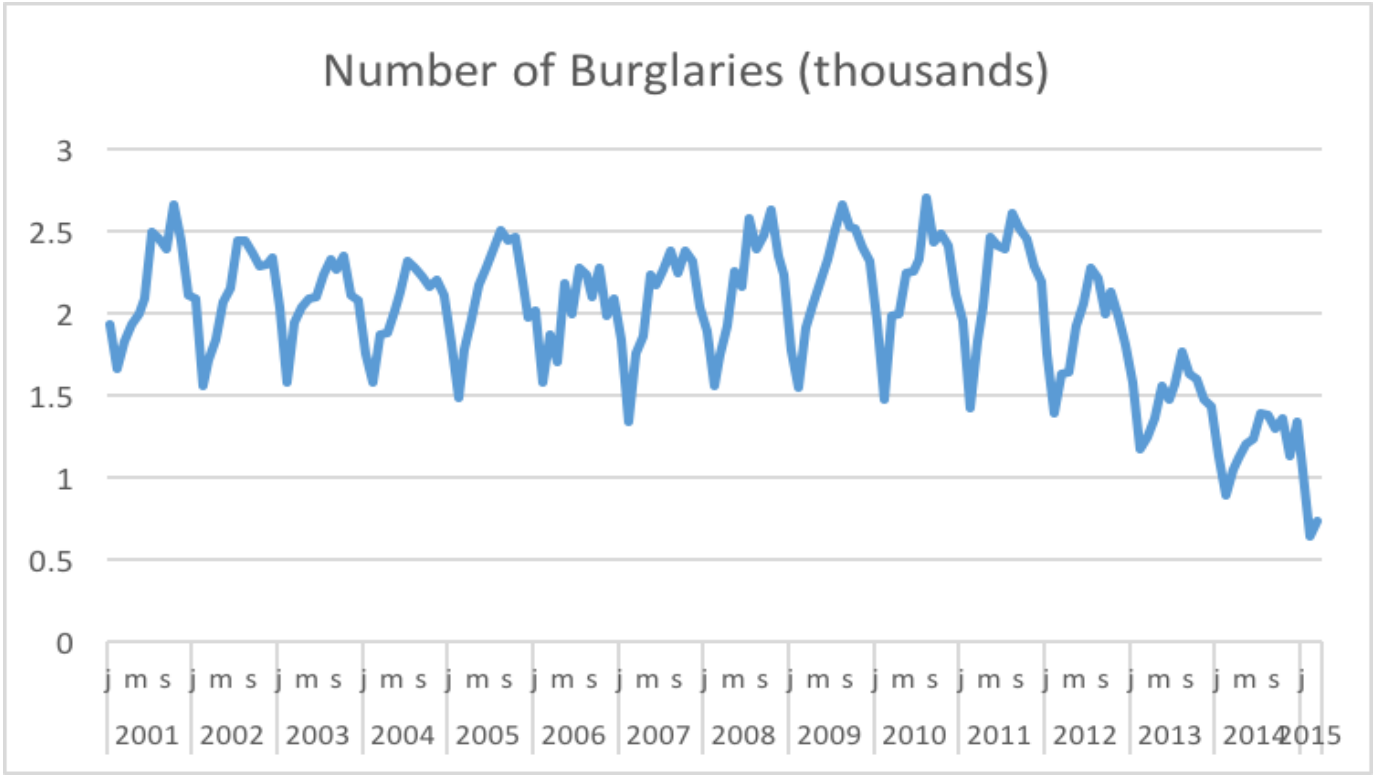
# Thank you.

# Annex

# Burglaries in Chicago with community area overlay (2014)



# Certain stationality in burglaries



# Average burglary values

--Geographic unit : 800 **census tracts** in Chicago (average 11 census tracts per community)

-- Time unit : 1923 **days** from 01/01/2010-- 04/07/2015

| | |
|---|---|
| **Minimum** Average Burglary counts per day per tract | 0.00052 ( 1 burglary in 2010-15 ) |
| **Maximum** Average Burglary counts per day per tract | 0.3686 ( 709 burglaries in 2010-15 ) |
| **Standard Deviation** of ave. burglary counts per day per tract | 0.0537 |
| **Mean** ave. burglary counts per day per tract | 0.0719 |

# Using Knox to forecast risk areas

Possible police approach

1. Obtain last days burglaries
2. Estimate Knox Matrix for each observation (their information will not be jittered)
3. Prepare a time-distance matrix with burglaries information for each data point
4. Multiply Knox Matrix by Time-distance Matrix (for each observation)
5. Obtain the score for each observation
6. Rank the observations by score
7. Allocate resources based on scores & available resources

# Problems we encountered

- Same location problem
  - Out of the 3 community areas with the highest knox for one day and same location (Archer Heights, North Center, and Portage Park), we have 132 burglaries that happened in the same day and block. Out of those 132, 58% were registered in the same location as well (76 burglaries). Of the 76 burglaries, 22 happened in apartment building (those 22 were registered in the 3 aforementioned community areas).
  - The latter can be solved by estimating the Knox from 0-600 feet.

# Suggestions

- Run analysis with Centroid
- See a map with the variation of Knox by CA…
- Do a simulation (with hold-out sample) to predict
  - random week
  - choose an area based on the model results (0-14days; parameter)
  - see how good is the prediction vs. baseline
- Other models (more complicated for the future)