Rapid detection of spatiotemporal clusters

Markus Loecher, Berlin School of Economics and Law

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

July 2nd, 2015

Table of contents

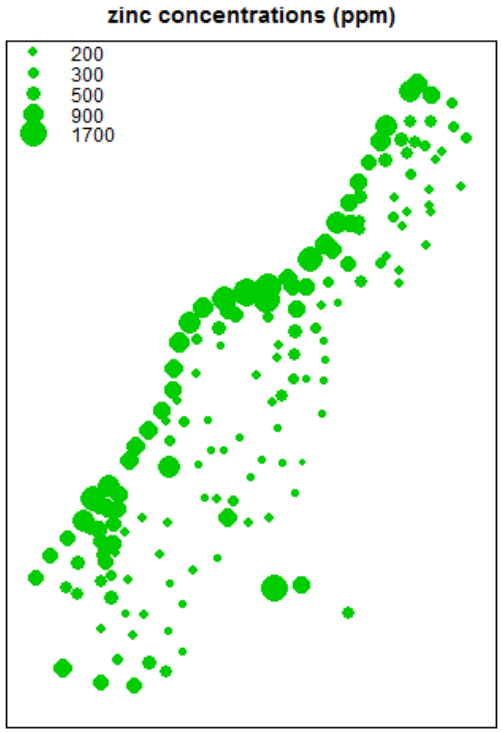
Motivation
- Spatial Plots in R
- RgoogleMaps

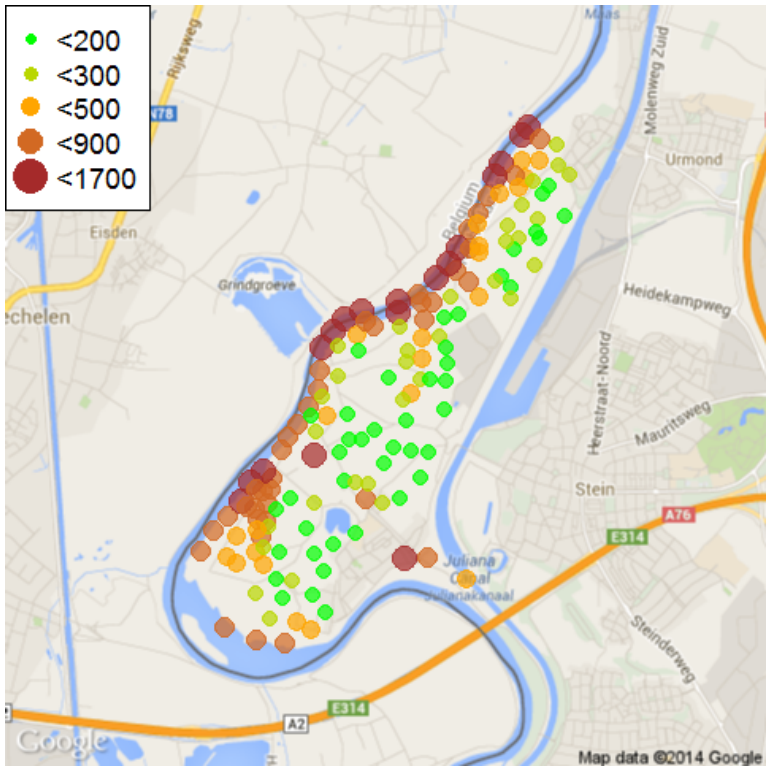
Spatial cluster detection
- Spatiotemporal Clusters
- Hot spot analysis
- Unsupervised as Supervised Learning

TreeHotspots
- Data Rotation
- Leaves as HotSpots
- Likelihood Ratio

Outlook

Motivation I





The meuse data set gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Heavy metal concentrations are from composite samples of an area of approximately 15 m x 15 m.

San Francisco Crime Data

https://data.sfgov.org/Public-Safety/Map-Crime-Incidents-from-1-Jan-2003/gxxq-x39z

SF OpenData

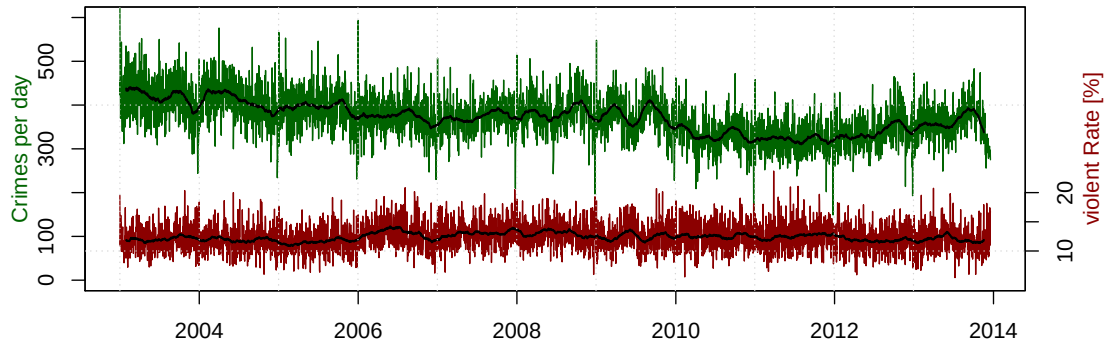
Map: Crime Incidents - from 1 Jan 2003

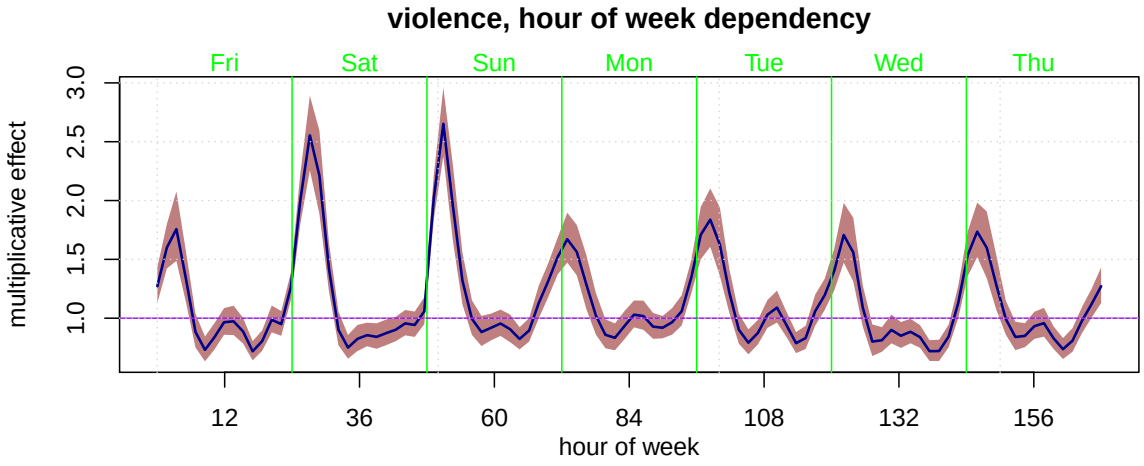
Based on SFPD Incidents - from 1 January 2003

Incidents are derived from SFPD Crime Incident Reporting system. Updated on a daily basis;data available from 1 Jan 2003 up until two

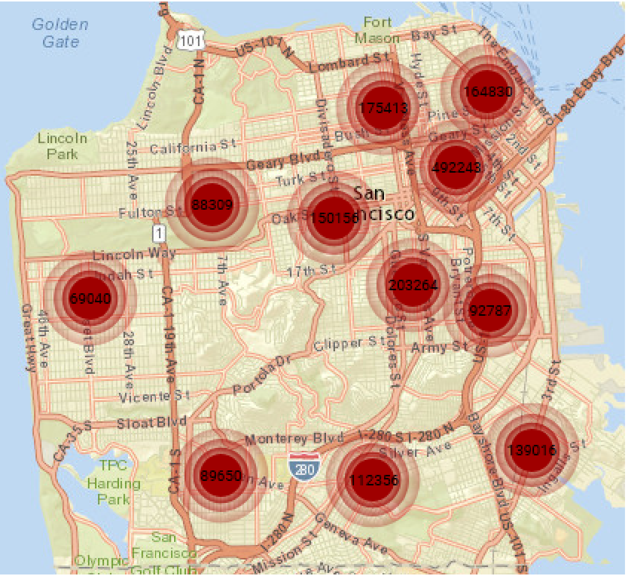
	Category	Descript	DayOfWeek	Date	Time	PdDis
1	Click to sort	BATTERY	Wednesday	04/20/2005 12:00:00 AM	04:00	MISSIO
2	EFT	GRAND THEFT FROM A BUILDING	Sunday	01/13/2008 12:00:00 AM	18:00	PARK
3	ASSAULT	AGGRAVATED ASSAULT WITH A KNIF	Sunday	05/05/2013 12:00:00 AM	04:10	INGLES
4	DRIVING UNDER THE INFLUENCE	DRIVING WHILE UNDER THE INFLUEN	Tuesday	07/08/2003 12:00:00 AM	01:00	SOUTHE
5	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Friday	10/04/2013 12:00:00 AM	20:53	TENDER
6	BURGLARY	BURGLARY OF APARTMENT HOUSE, U	Tuesday	08/14/2007 12:00:00 AM	07:00	NORTHI
7	DRUG/NARCOTIC	POSSESSION OF MARIJUANA	Tuesday	03/04/2008 12:00:00 AM	14:23	INGLES
8	OTHER OFFENSES	DRIVERS LICENSE, SUSPENDED OR R	Wednesday	07/05/2006 12:00:00 AM	15:50	INGLES
9	LARCENY/THEFT	GRAND THEFT FROM A BUILDING	Wednesday	12/10/2003 12:00:00 AM	09:30	INGLES
10	NON-CRIMINAL	STAY AWAY OR COURT ORDER, NON-I	Monday	01/17/2011 12:00:00 AM	15:35	INGLES
11	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Saturday	01/07/2006 12:00:00 AM	22:00	NORTHI
12	LARCENY/THEFT	PETTY THEFT BICYCLE	Sunday	11/13/2011 12:00:00 AM	18:00	MISSIO
13	SEX OFFENSES, FORCIBLE	ASSAULT TO RAPE WITH BODILY FOR	Monday	02/17/2014 12:00:00 AM	14:30	INGLES
14	SUSPICIOUS OCC	INVESTIGATIVE DETENTION	Wednesday	04/11/2012 12:00:00 AM	15:10	INGLES
15	VEHICLE THEFT	STOLEN TRUCK	Saturday	08/30/2003 12:00:00 AM	11:00	TARAVA
16	NON-CRIMINAL	LOST PROPERTY	Monday	04/08/2013 12:00:00 AM	15:15	NORTHI
17	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Tuesday	06/16/2009 12:00:00 AM	22:00	TARAVA

Temporal Analysis

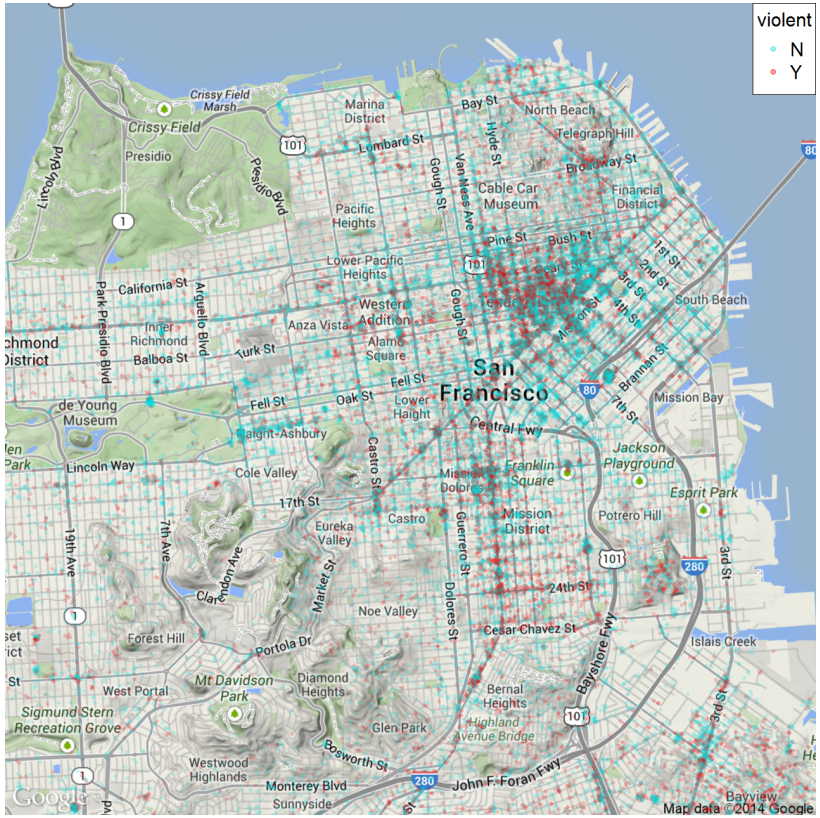


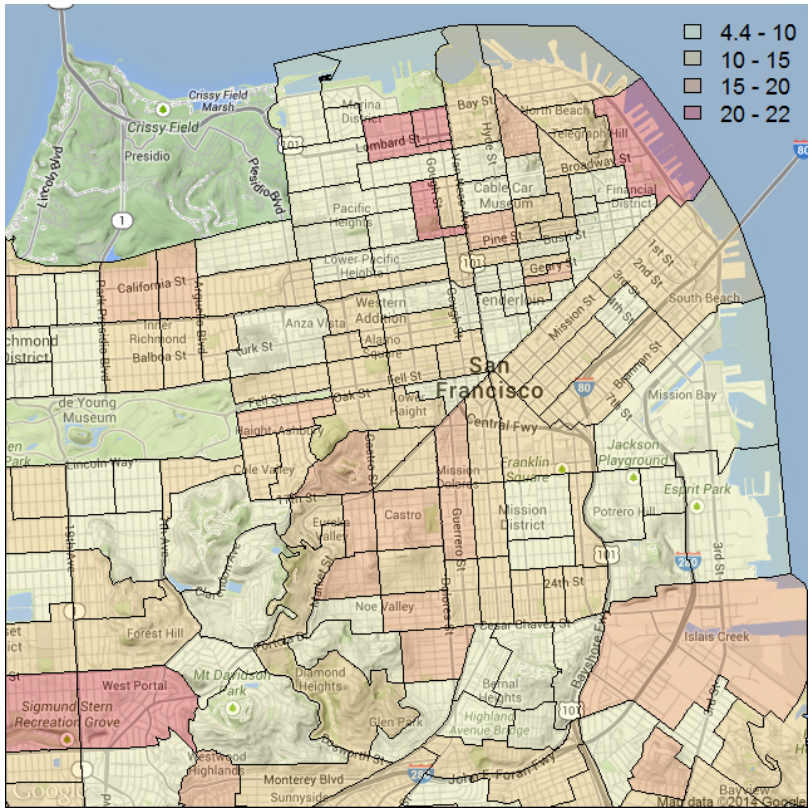


SF Open Data


Golden Gate
Fort Mason
Bay St
101
US-101 N
Lombard St
Lincoln Blvd
CA-1 N
Hyde St
175413
164830
Pine St
Bush St
Divisadero St
Lincoln Park
California St
25th Ave
Geary Blvd
Geary St
492243
Turk St
San
88309
Fulton St
Oak St
150156
1
Lincoln Way
7th Ave
17th St
7th St
2nd St
Bryant St
69040
203264
92787
Great Hwy
40th Ave
28th Ave
CA-1 19th Ave
Clipper St
Army St
Dolores St
Portola Dr
Vicente St
3rd St
Sloat Blvd
CA-35 S
Monterey Blvd
I-280 S
I-280 N
Silver Ave
139016
TPC Harding Park
89650
280
112356
CA-1 S
Geneva Ave
Mission St
Bayshore Blvd
US-101 S
I-80 E Bay Brg
San Francisco Golf Club
Olympic

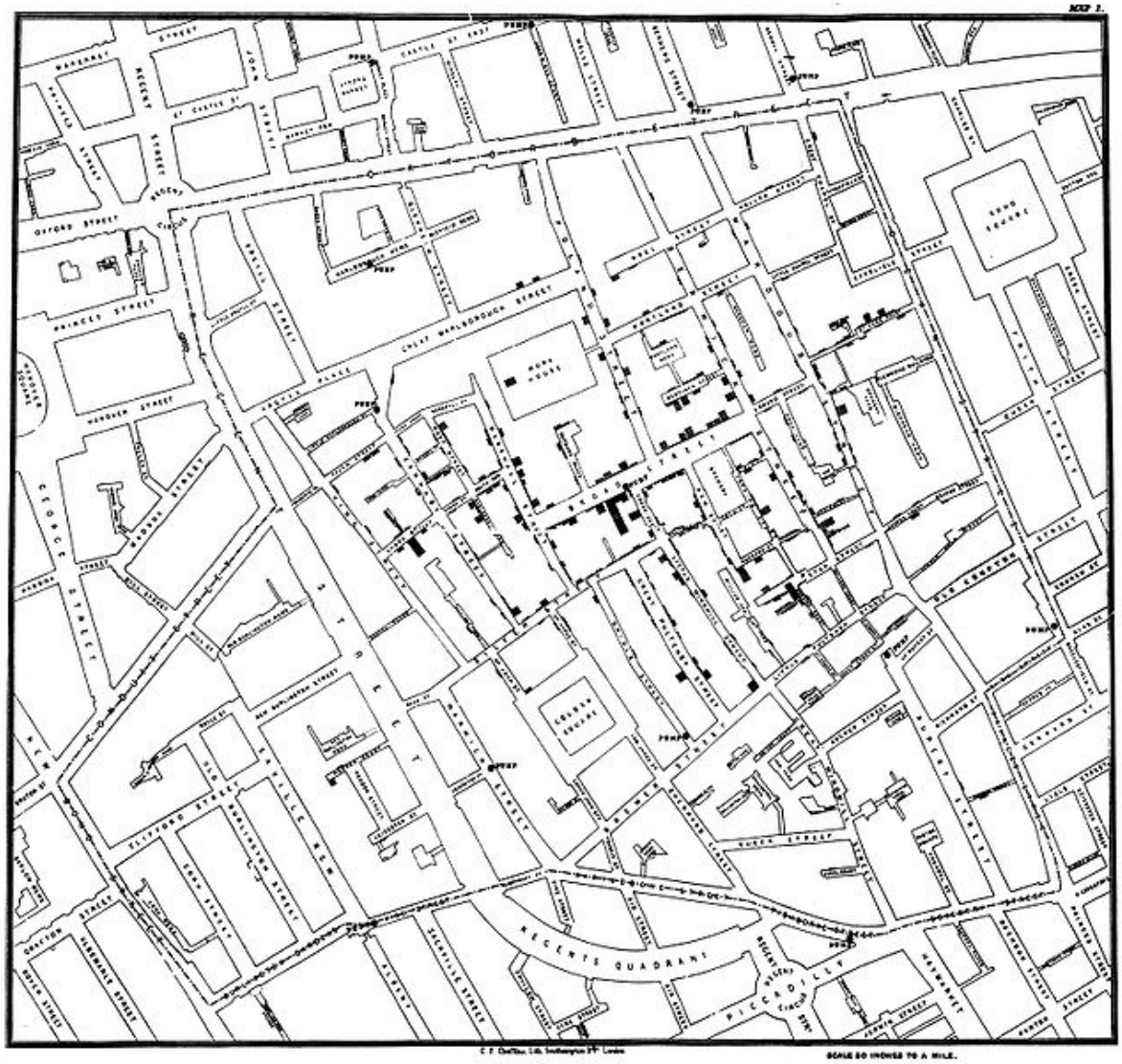
One Environment





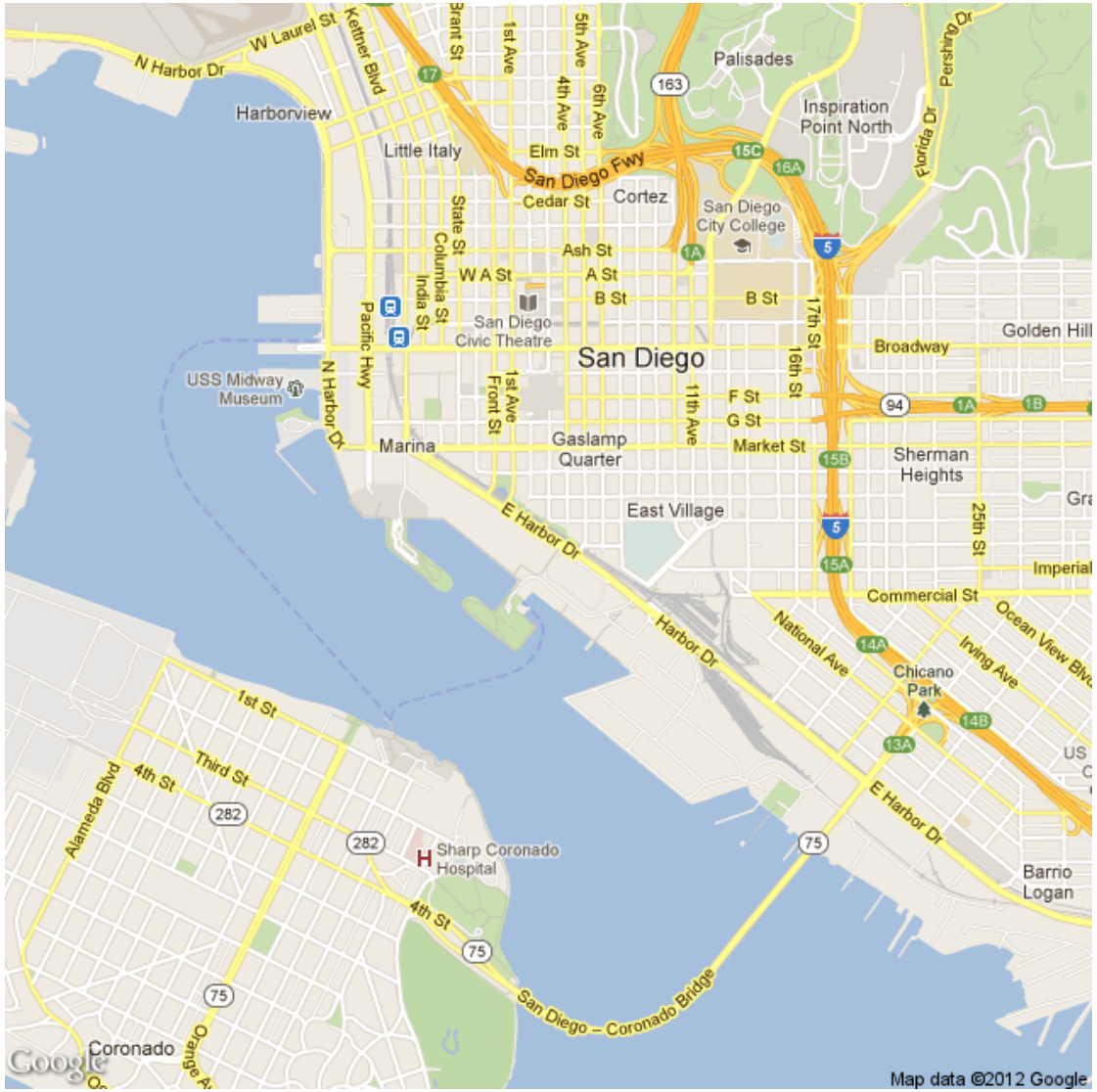
Not having to leave R is **priceless**

Back to 1854


BROAD STREET
REGENT STREET
PICCADILLY
GEORGE STREET
OXFORD STREET
GOLDEN SQUARE
SOHO SQUARE
LONG ACRE
CHARING CROSS
REGENT'S QUADRANT
SCALE 80 INCHES TO A MILE.

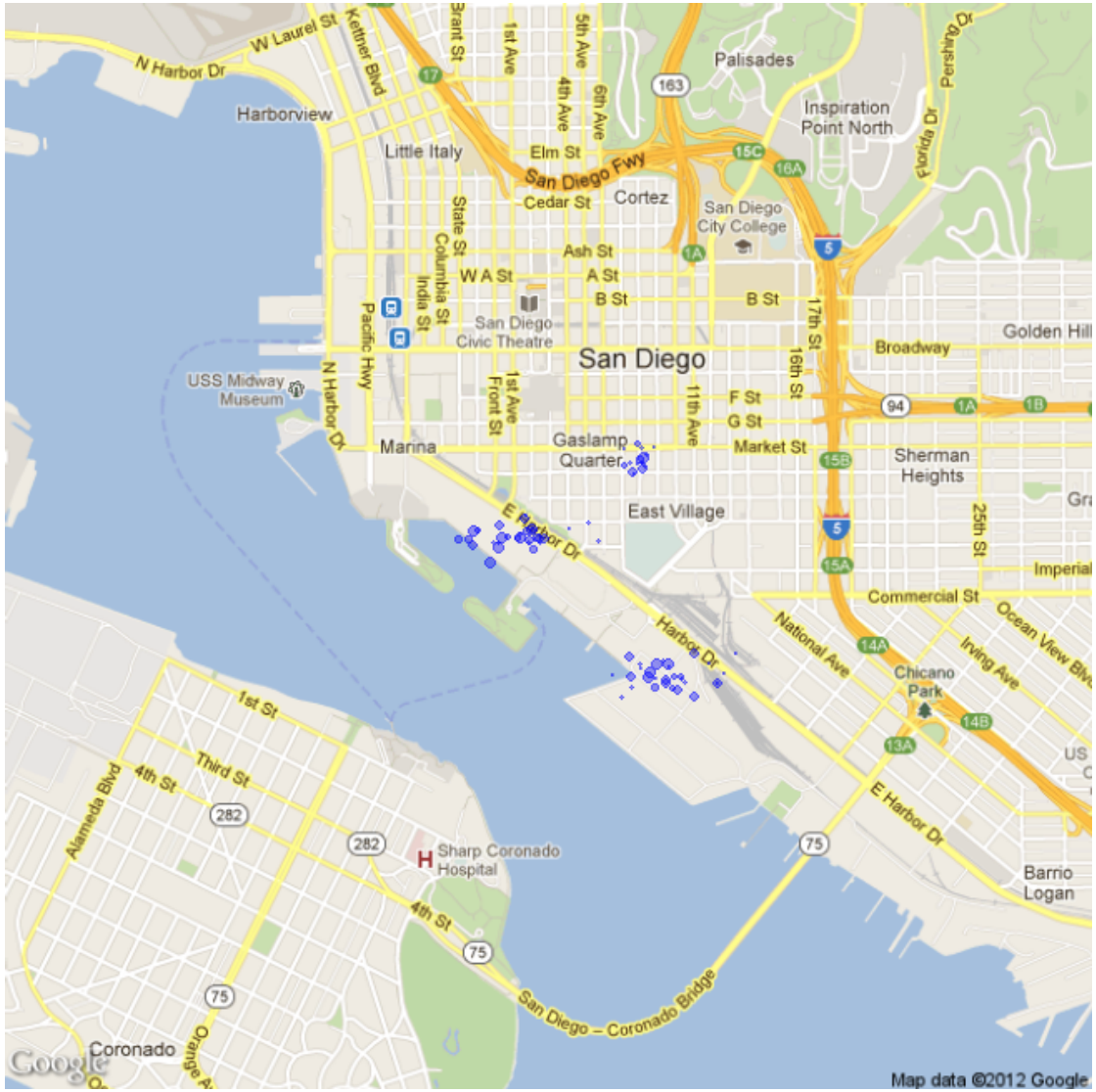
RgoogleMaps

mapSD = **GetMap**(center=c(32.7073, -117.162), zoom=10, destfile='SDconv.png')



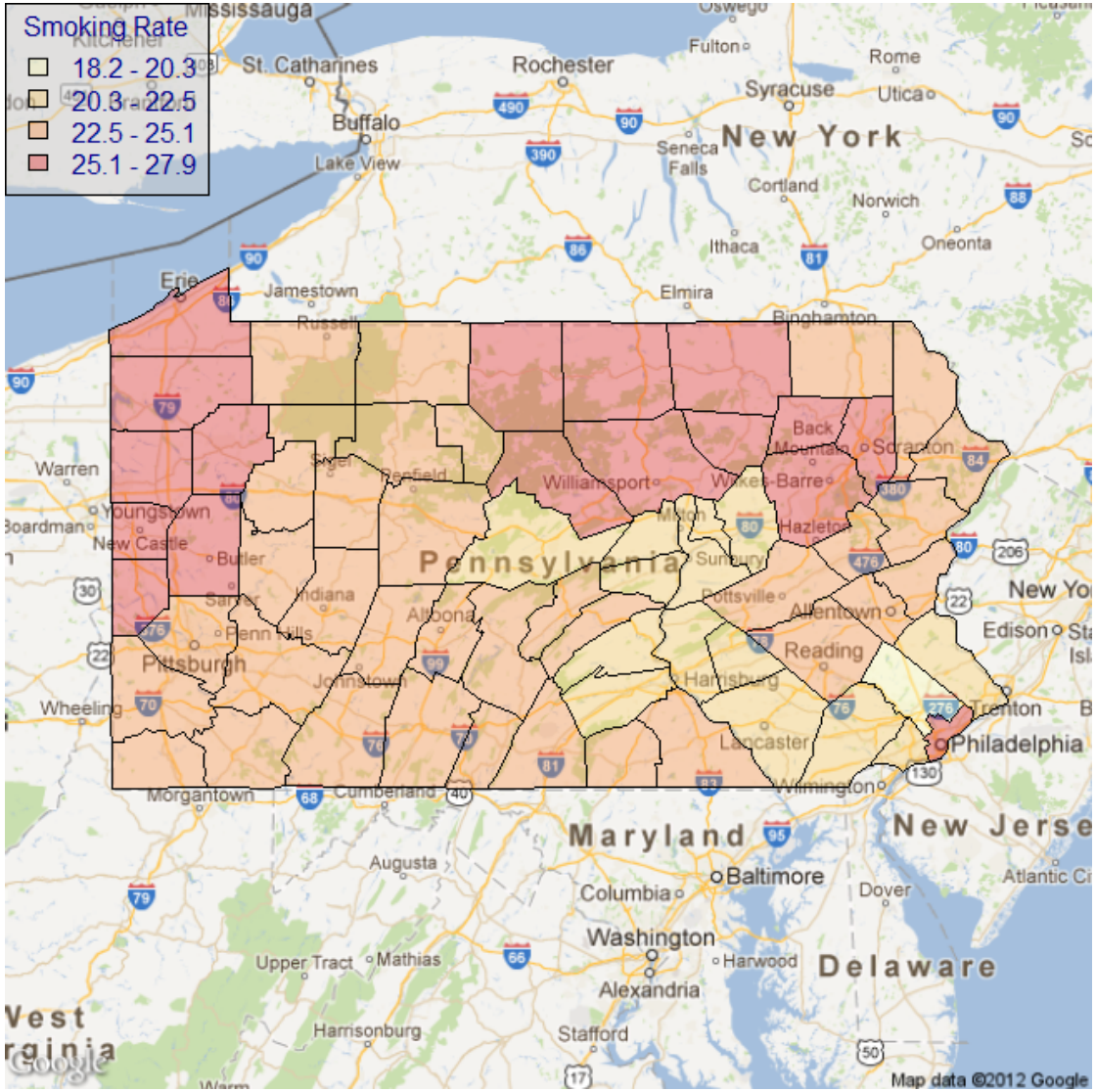
PlotOnStaticMap

PlotOnStaticMap(mapSD, lat=myTrails$lon, lon=myTrails$lon, col=myTrails$col, cex = myTrails$cex)



PlotPolysOnStaticMap

PlotPolysOnStaticMap(map, shp, lwd=.5, col = shp[,'col']);
shp=importShapefile(shpFile,projection='LL');



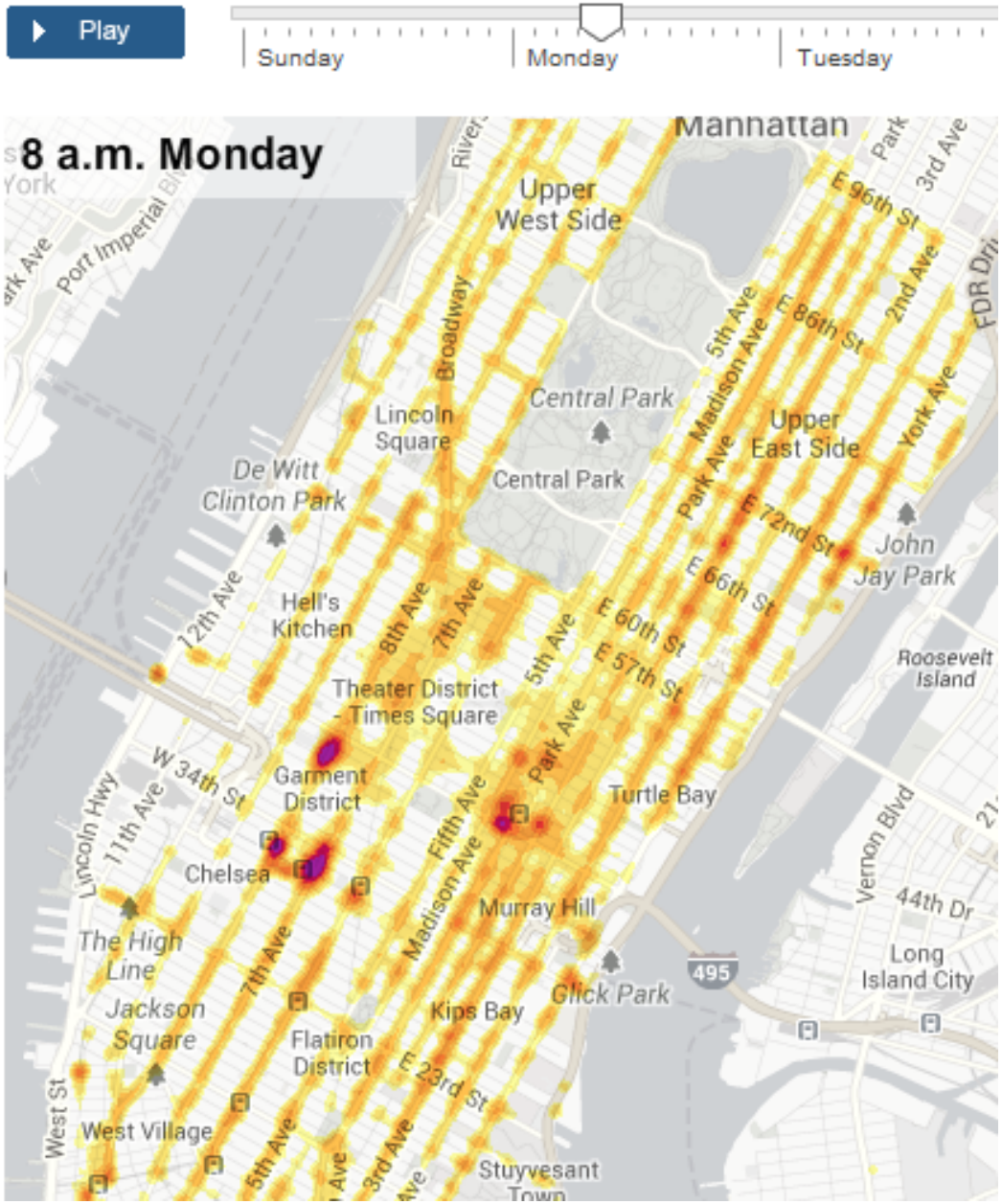
Spatiotemporal Clusters

Scoring unusual events in space and time has been an active and important field of research for decades: How do we

- distinguish normal fluctuations in a stochastic count process from real additive events ?
- identify spatiotemporal clusters where the event is most strongly pronounced ?
- efficiently graph these clusters in a map overlay ?

Supervised learning algorithms are proposed as an alternative to the computationally expensive scan statistic.
The task can be reduced to detecting over-densities in space relative to a background density.

NYC cab data


Play
Sunday
Monday
Tuesday
8 a.m. Monday
Manhattan
Upper West Side
Central Park
Lincoln Square
Central Park
Upper East Side
De Witt Clinton Park
Hell's Kitchen
Theater District - Times Square
Garment District
Chelsea
Turtle Bay
Murray Hill
The High Line
Jackson Square
Flatiron District
Kips Bay
Glick Park
West Village
Stuyvesant Town
John Jay Park
Roosevelt Island
Long Island City
Broadway
5th Ave
Madison Ave
Park Ave
2nd Ave
York Ave
FDR Dri
E 96th St
E 86th St
E 72nd St
E 66th St
E 60th St
E 57th St
E 23rd St
12th Ave
8th Ave
7th Ave
Fifth Ave
W 34th St
11th Ave
Lincoln Hwy
West St
Vernon Blvd
44th Dr
495

Hot spots

- Relatively compact areas of “high intensity”
- What is baseline ?

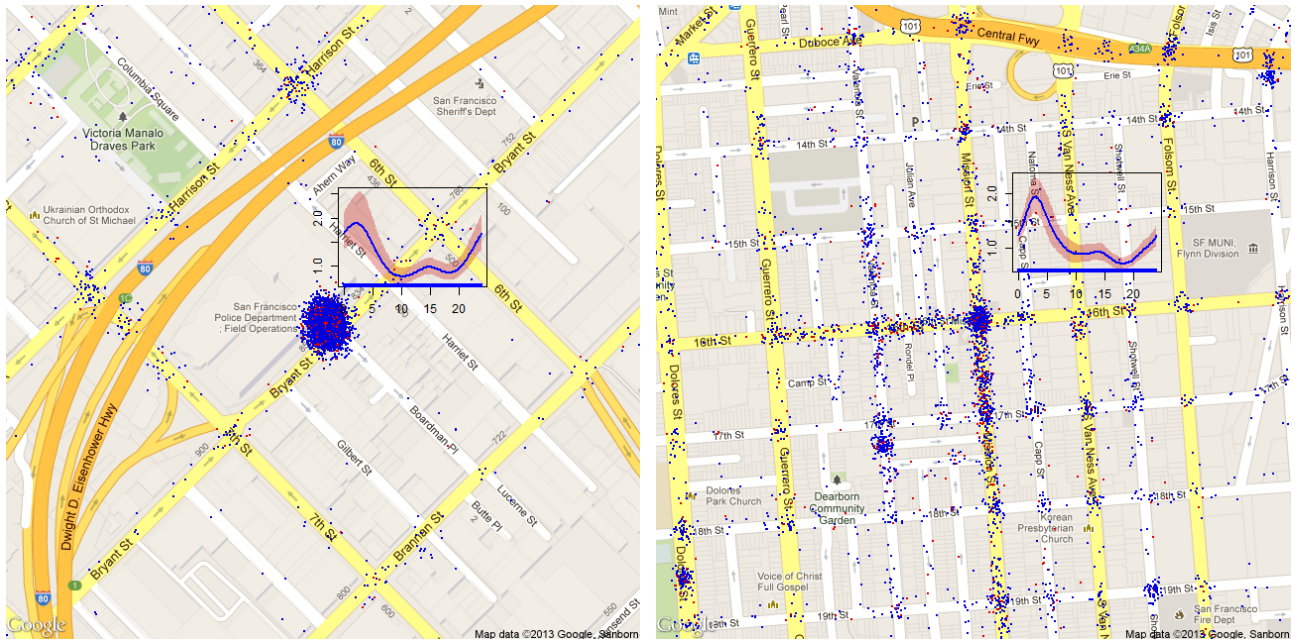
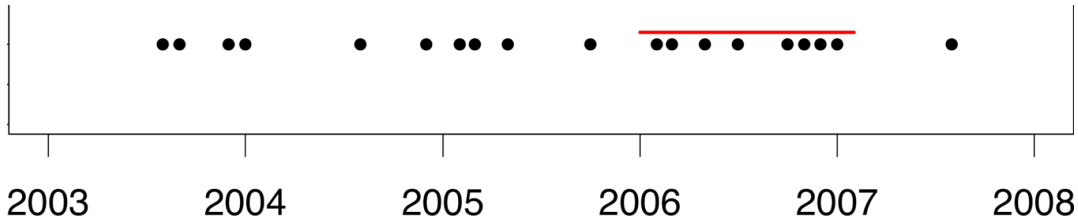


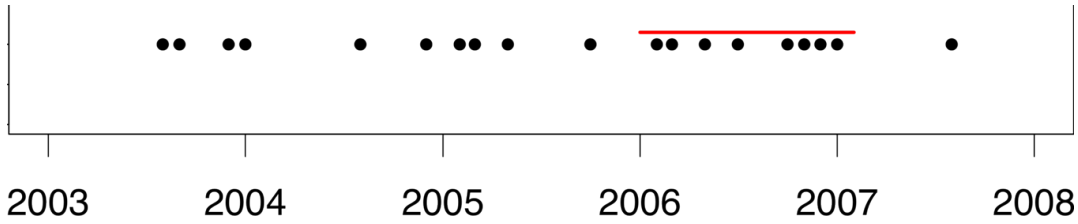
Figure: (left) One “hot spot” found where the contextual information provided by the map is invaluable. (right) Another cluster of crime activity spread along the street grid.

Unusual Clusters



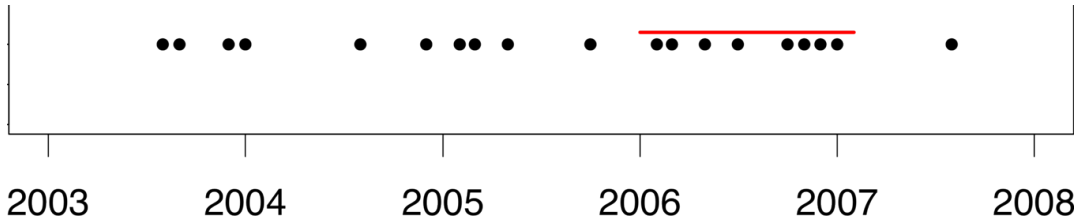
- "Over a 5 year period there were 19 cases of a particular type of cancer reported in a town. A physician notes that there is a 1 year period that contains eight cases. "
- "On Aug 17, **the U.S. Army suspended all operations of the Black Hawk helicopter** after the third crash in 25 days. The 3 crashes were about seven times the expected rate based on the previous 5 years. ($S_{25} = 3$)"

Unusual Clusters



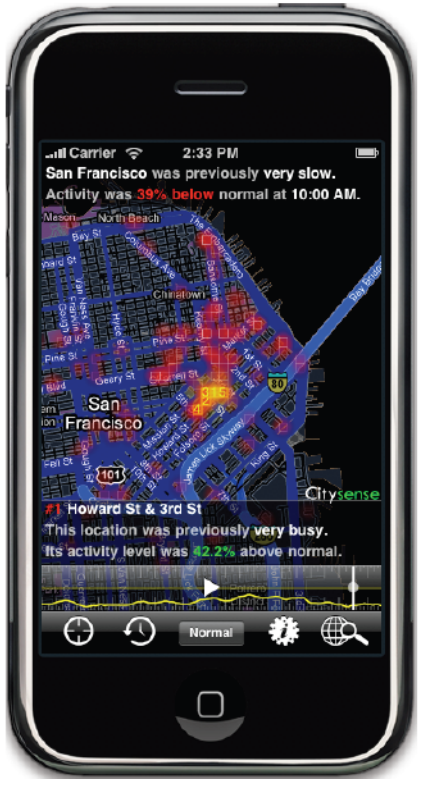
- "Over a 5 year period there were 19 cases of a particular type of cancer reported in a town. A physician notes that there is a 1 year period that contains eight cases. "
- "On Aug 17, **the U.S. Army suspended all operations of the Black Hawk helicopter** after the third crash in 25 days. The 3 crashes were about seven times the expected rate based on the previous 5 years. ($S_{25} = 3$)"
- " **Alarming number of inmate deaths in Harris County**: In a 10 month period, 11 inmates died at the troubled Harris County Jail, which is about twice the expected rate. The U.S. Department of Justice ordered the city of Houston to pay a fine of $1000 a day until the cause was found. ($S_{10} = 11$)"

Unusual Clusters



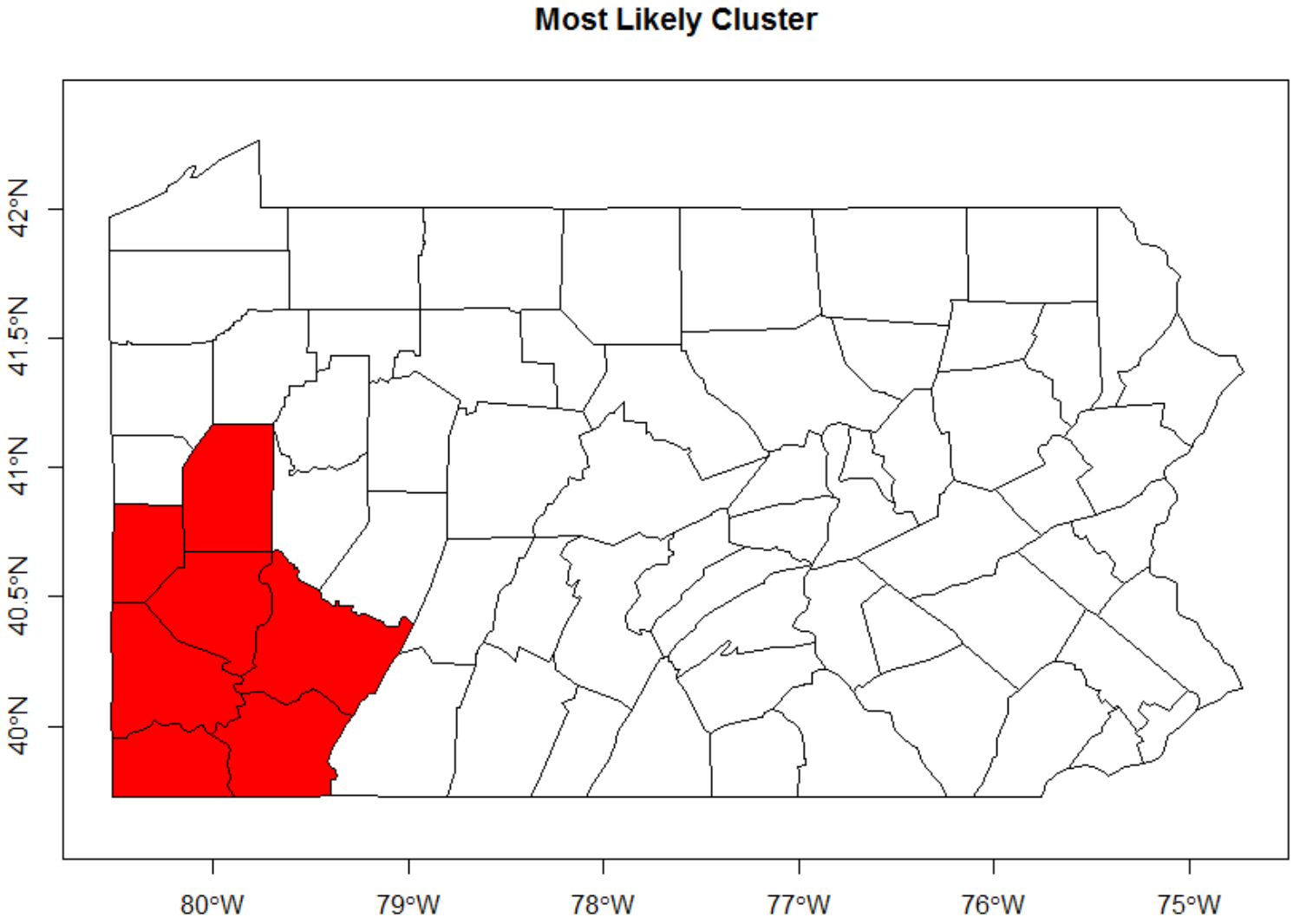
- "Over a 5 year period there were 19 cases of a particular type of cancer reported in a town. A physician notes that there is a 1 year period that contains eight cases. "
- "On Aug 17, **the U.S. Army suspended all operations of the Black Hawk helicopter** after the third crash in 25 days. The 3 crashes were about seven times the expected rate based on the previous 5 years. ($S_{25} = 3$)"
- " **Alarming number of inmate deaths in Harris County**: In a 10 month period, 11 inmates died at the troubled Harris County Jail, which is about twice the expected rate. The U.S. Department of Justice ordered the city of Houston to pay a fine of $1000 a day until the cause was found. ($S_{10} = 11$)"

Scan Statistic



This type of spatial surveillance is computationally expensive: $O(R \cdot N^4)$

Pennsylvania Lung Cancer



Package *SpatialEpi: kulldorff()* function

Unsupervised as Supervised Learning

Introduced in Hastie *et al* for density estimation or association rule generalizations. Problem must be enlarged with a simulated data set generated by Monte Carlo techniques

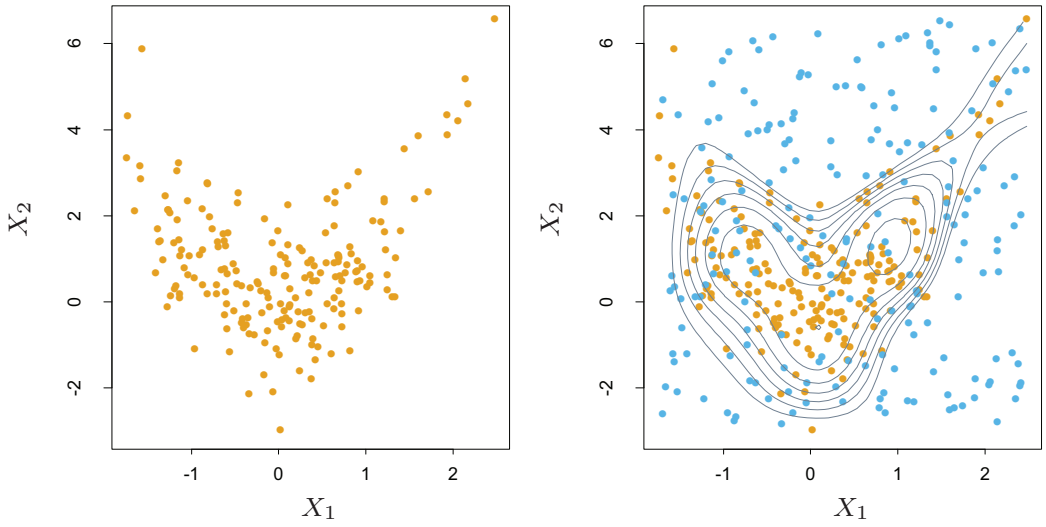


FIGURE 14.3. *Density estimation via classification. (Left panel:) Training set of* 200 *data points. (Right panel:) Training set plus* 200 *reference data points, generated uniformly over the rectangle containing the training data. The training sample was labeled as class* 1*, and the reference sample class* 0*, and a semiparametric logistic regression model was fit to the data. Some contours for* $\hat{g}(x)$ *are shown.*

In epidemiology cases and population naturally provide two classes, for anomaly detection the background "population" is taken to be some sort of average.

CART

A cluster found by a classification tree visualized on a Google map tile. The numeric labels indicate the fraction of the positive class labels.

TreeHotspots

R package with new functionalities

- Rotation of Data
- Visualization of selected leaves

TreeHotspots

R package with new functionalities

- Rotation of Data
- Visualization of selected leaves
- Overlay on maps (Google and OSM)

TreeHotspots

R package with new functionalities

- Rotation of Data
- Visualization of selected leaves
- Overlay on maps (Google and OSM)
- User written splitting functions for *rpart*:

TreeHotspots

R package with new functionalities

- Rotation of Data
- Visualization of selected leaves
- Overlay on maps (Google and OSM)
- User written splitting functions for *rpart*:
  - Baseline Distributions eliminate need for point augmentation
  - SatScan Poisson and Binomial Likelihood, e.g.

$$\left(\frac{c}{E[c]}\right)^{c} \cdot \left(\frac{C-c}{C-E[c]}\right)^{C-c}$$

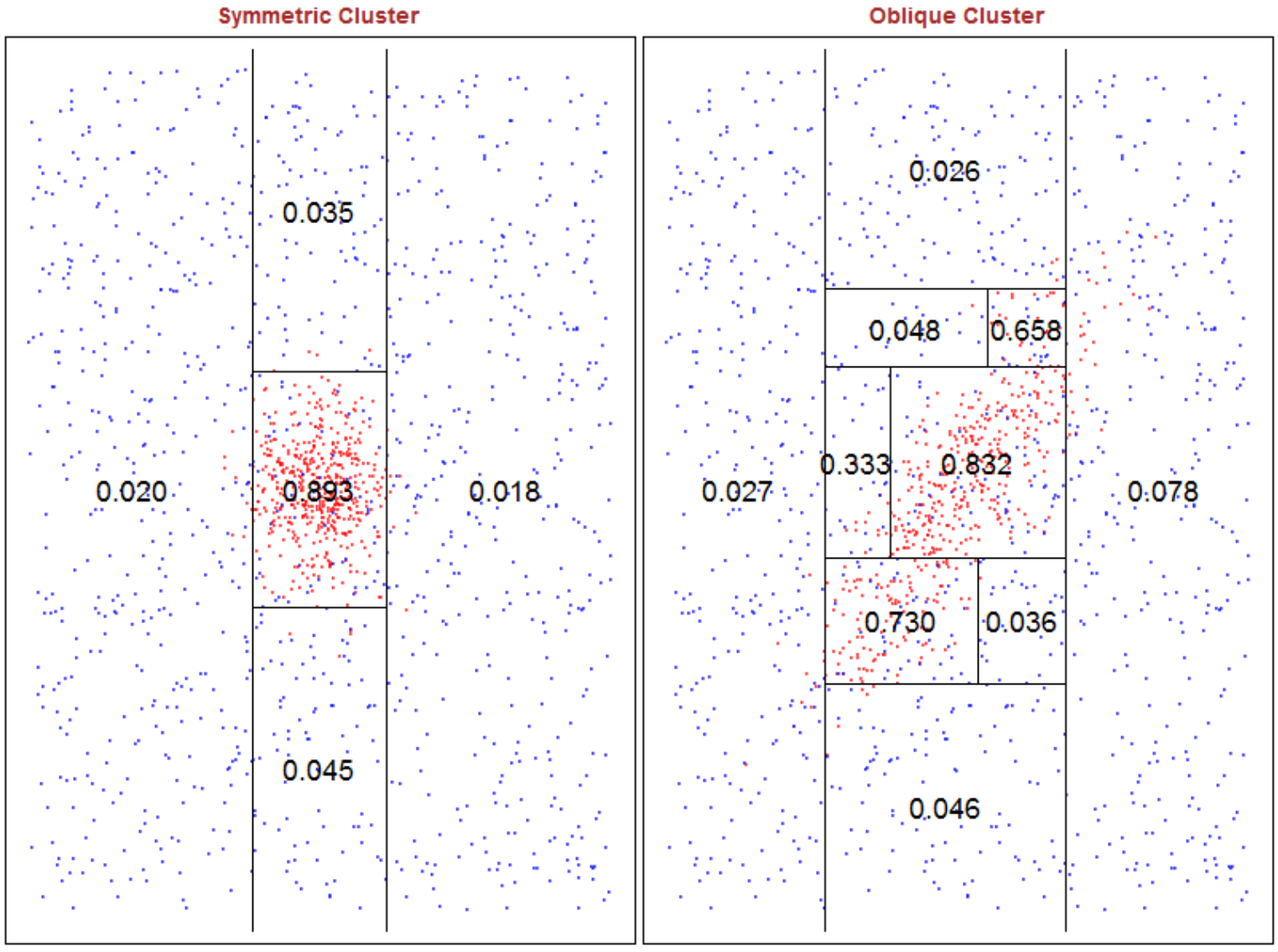
TreeHotspots

R package with new functionalities

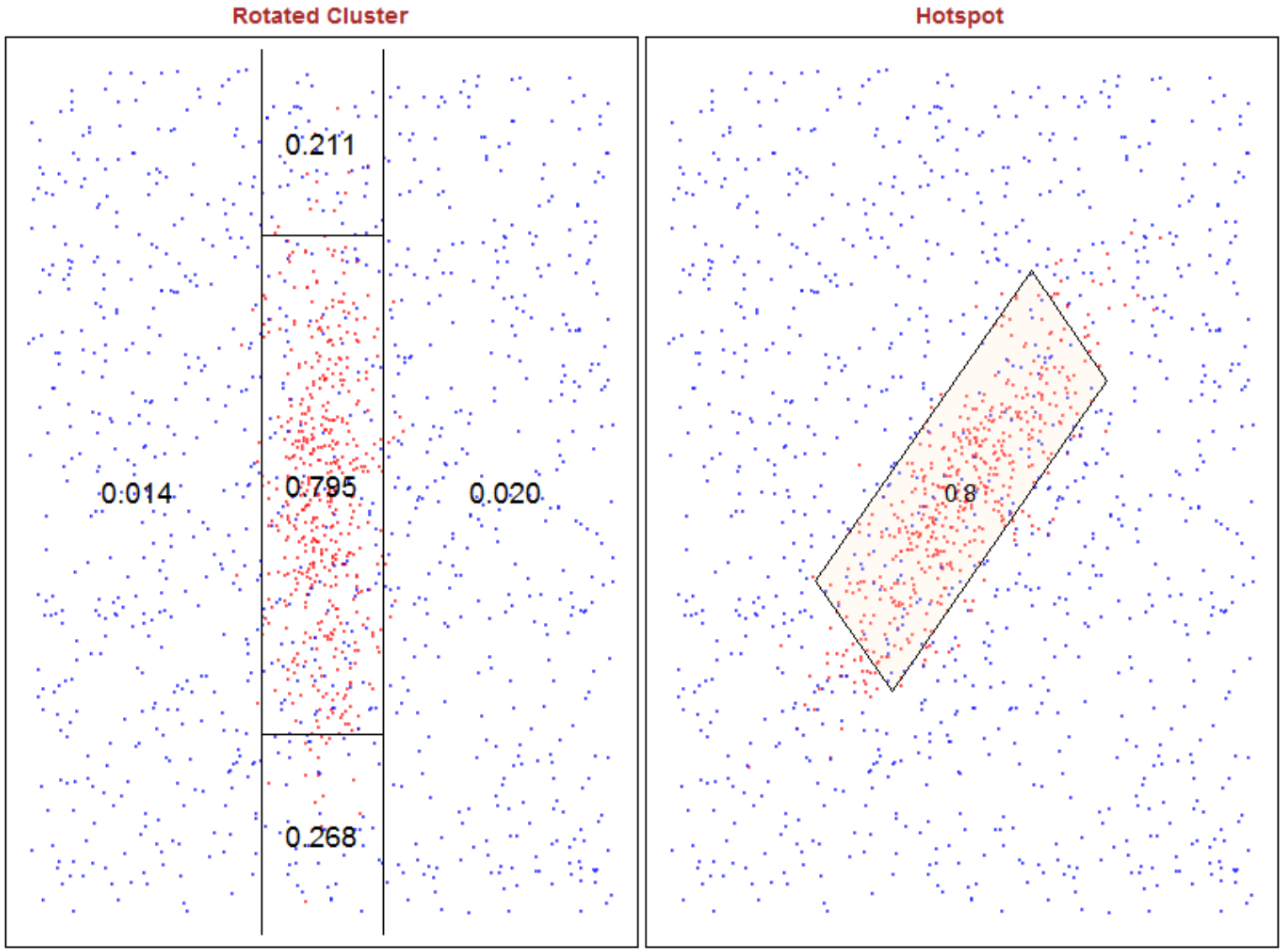
- Rotation of Data
- Visualization of selected leaves
- Overlay on maps (Google and OSM)
- User written splitting functions for *rpart*:
  - Baseline Distributions eliminate need for point augmentation
  - SatScan Poisson and Binomial Likelihood, e.g.

$$\left(\frac{c}{E[c]}\right)^{c} \cdot \left(\frac{C-c}{C-E[c]}\right)^{C-c}$$

Simulations



Simulations



SF crime data, on map

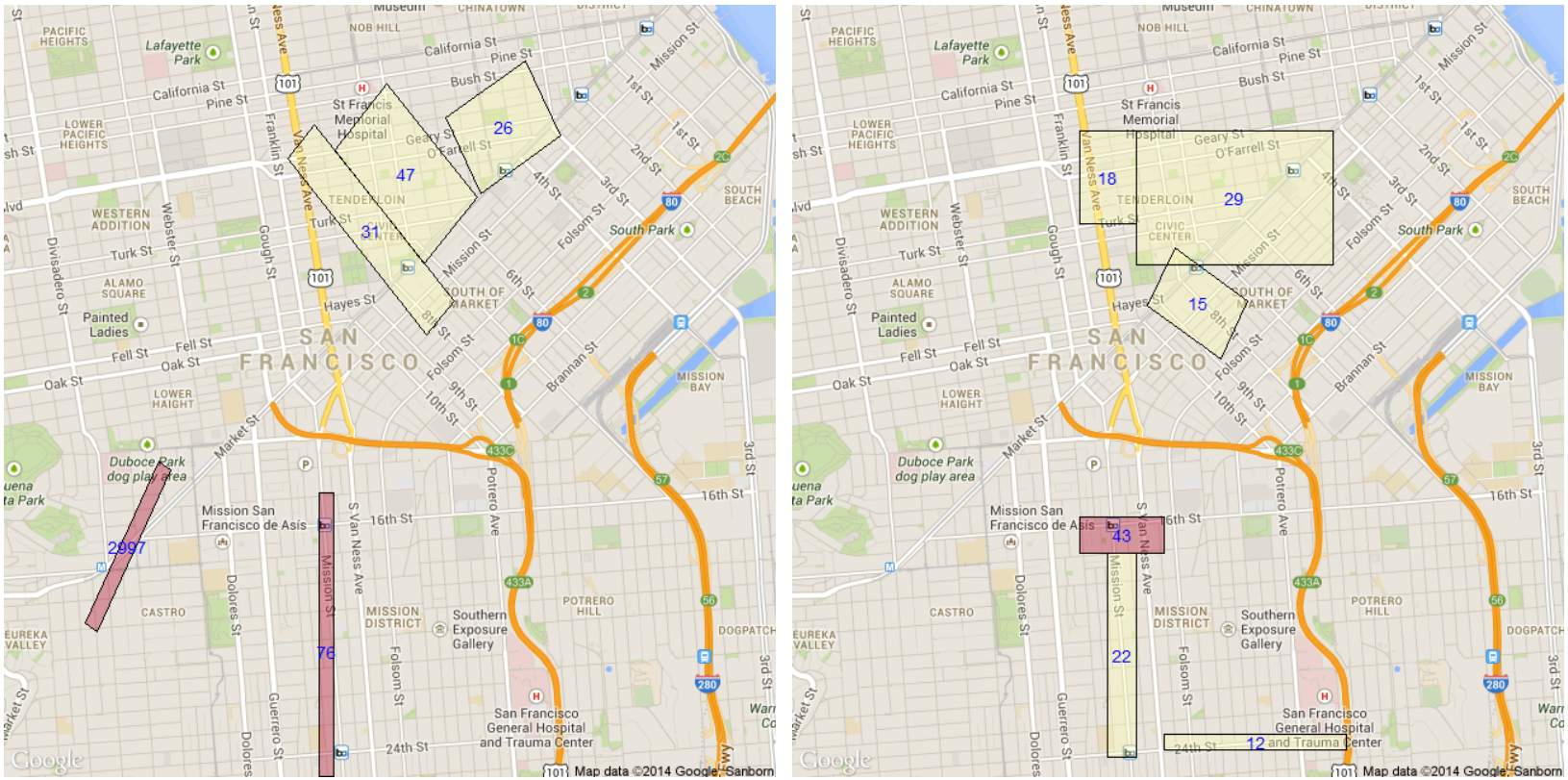


Figure: positive class is given by (left) drug crimes and (right) robbery related incidents.

Outlook

- Power Study
- R library *partykit*

Outlook

- Power Study
- R library *partykit*
- lat lon distortion

Outlook

- Power Study
- R library *partykit*
- lat lon distortion
- CRAN or github release

Outlook

- Power Study
- R library *partykit*
- lat lon distortion
- CRAN or github release