SaTScan performance analysis - Executive summary

SaTScan detects ~8 space-time clusters on October 2011 data

Main limit is that reported clusters cannot overlap geographically, whatever the time of the event → Small number of clusters

When parameters vary, only the most significant results are very consistent

- May be different without the non-overlapping constraint

Adding a cluster size bound allows detection of more clusters with more even significance

Ellipse scan performs better than circular but is more expensive

- Computation: Ellipse ~ Circles x 100 on data
- Ellipses fit events better than circles – elongated events on road network
- Results are more significant when ellipse can elongate more – no constraint

0. Experiment design

SaTScan experiment parameters
Time analysis
Time bounds: ***October***
Year / Month / Day / Generic: ***Day***
Aggregation period: ***1 Day (minimum)***
Cluster
max/min bounds: ***all time clusters***
Adjust day-of-week: ***No***
Spatial analysis
Window shape
Circular
Ellipse
Compactness constraint: none / mid / strong
Cluster size bound / cluster count bound: *none / 1km / 0.5km for circular – none / 25k cart.units / 10k for ellipses*
Type of analysis
Retrospective/Prospective: ***Retrospective***
Proba model: ***Space-Time permutation – only one that fits the data***
High/Low rates: ***Low rates***
Cluster selection: ***No geographical overlap whatever the time***
Coord system
Lat/Long or Cartesian:
Replications: ***0***
Output
Temporal graph for clusters
kml file

SaTScan experiments summary

Name of experiment	Run time PC	Clusters detected	Bound cluster radius (radius km/semi-minor axis units)	Ellipse constraint	Max/Med/Min cluster size rad km / semi-minor k cart. units	Max/med/min semi-major (k cart.units)	Stat score: max/nedian/min	Replicas
graph	1’09’’	6	NA		2.4/0.8k/0.2		7500/1200/500	0
graph05	24’’	10	0.5		0.5/0.48/0.2		4700/2300/1100	0
graph1	19’’	10	1.0		0.94/0.6/0.08		6700/1500/750	0
graph_elp	2h10’	3	NA	mid	50/24/10	78/71/20	9300/7600/600	0
graph_elp_10k	7’	10	10k	mid	10/9.5/6	50/45/8	6300/1400/800	0
graph_elp_25k	25’	6	25k	Mid	25/23/10	98/72/20	9400/900/75	0
graph_elp_10k_none	3’	10	10k	none	10/9.7/6	50/49/30	8500/1950/750	0
graph_elp_10k_strong	7’	10	10k	strong	10/9.7/6.5	50/20/8	5000/1300/750	0

Main limit of executable SaTScan: small number of reported clusters

Small number of reported clusters

→ Maximum = 10, down to 3 sometimes

→ Important parts of Manhattan where no event is detected over the whole month

→ Due to *"Non-geographical overlap"* option only?

→ Actually, test statistic are computed for every point of the grid. It would be possible to report all of them and choose a significance threshold

→ ***Makes the results difficult to interpret since for each area***

Currently, impossible to use SaTScan as an exploratory tool

This version is better to detect the most significant event on a given period

1.1. Cluster size bound - Circular

None 1km 0.5km

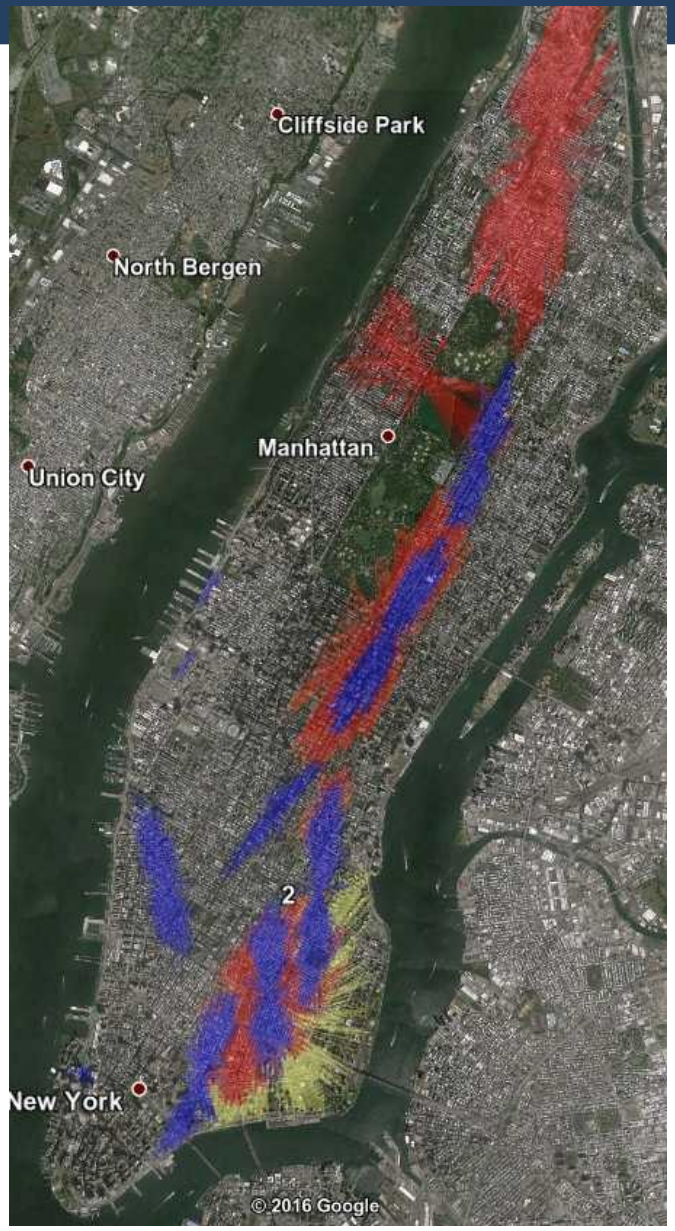
Significance
→ Increases with cluster size
Max test statistic: 7500 / 6700 / 4500
→ More equally distributed with small size bound
Median test statistic: 1200 / 1500 / 2300

Number of clusters
→ High when small bound (overlap option)

Consistency
→ Overlapping clusters always have same dates
→ Strong consistency 1km / 0.5km

Event Analysis
→ Halloween parade

1.2. Cluster size bound - Ellipses



Significance → same as circular
→ Increases with cluster size
→ More equally distributed with small size bound

Number of clusters → same as circular
→ High when small bound (overlap option)

Consistency
→ Overlapping clusters always have same dates
→ Strong consistency for high test significances

Small bound allows to detect a greater number of clusters of more even significance

Bound in cartesian units, arbitrary unit

None 25k 10k

2. Circles vs Ellipses

2. Circles vs Ellipses

Significance
→ Ellipse clusters have ~25% better test statistic
→ More equally distributed with small size bound
Median test statistic: 1200 / 1500 / 2300

Number of clusters
→ Similar

Shape of clusters
→ Flat ellipses are favored, possibly because they fit better the road network

Consistency
→ Quite bad for none or large size bounds
→ Very strong with small size bounds
→ Most overlapping events have same dates

Computation for circles is much cheaper
→ No bound: x120
→ Mid bound: x50
→ Small bound: x30

3. Ellipse constraints

None Medium Strong

Significance

→ No constraint => Most significant

Max test stat 8500 / 6300 / 5000

Number of clusters

→ Similar

Consistency

→ On most significant

→ Two geographical neighbors may have different dates

Possible improvements

Increase # reported cluster with overlap options

→ Report clusters which are non overlapping in geography OR time

→ Select manually clusters

Search for more ellipses

→ Add ellipse shapes (flatter)

→ Add ellipse angles

Adjust for day-of-week effect

→ TBD: check if many events are if the week-end

Possible next steps:

1) Get SaTScan source code and adapt it to the problem

2) Look for other techniques → Seem more reasonable at first

No easy event interpretability

Scan	# Cluster	Main st 1	Main st 2	Time	Significance	Zone	Research	Day-of-Week
Elp	1	34th → 75th st	Lexington / Park / Madison	29 → 30	9000	E.Mid / UES	**Not relevant – Central Park Pumpkin Festival**	Sat29
	2	E.Bway → 14th st	Bowery → Pitt st	3 → 6	7500	LES	**Not found**	Mon3
	(3)				500			
Elp25k	1	34th → 75th st	Lexington – Park - Madison	29 → 30	9300	E.Mid / UES	**Not found**	
	2	Canal St → 14th st	Bowery – Allen – Essex	3 –> 6	7300	LES	**Not found**	Mon3
	3	90th → 100th	Madison - 5th – Central Park - Columbus	8 → 9	1100	UWS		Sat8
	4-5-6				600 / 200 / 70			
Elp10k	1	40th → 70th	Lexington – Park - Madison	29 → 30	6300	E.Mid / UES	**Not found**	Sat29
	2	Grand – Houston – Delancey - 10th	Essex – Allen - 2nd Ave	24 → 27	3500	LES/ EV	**24: Grub Street Food - not significant / not found**	Mon24
	3	79th → 96th	Park - Madison	1 → 2	2600	E.Mid / UES	**Not found**	Sat1
	4	Bleeker – Christopher – W 14th	6th → 8th	24 → 27	1800	Greenwhich/ West Village	**Not found**	Mon24
	5 → 10							

October 2011 NY events

Holiday

- 10: Columbus Day
- 31: Halloween (off work?)

10: Columbus parade, 5thAve&44th st → 5thAve&79th → NO

31: Halloween Parade, Greenwich village → NO

29: snow storm → NO, *a priori not localized in particular neighborhood*

- 3 inches of snow in 1 day
- https://www.washingtonpost.com/blogs/capital-weather-gang/post/historic-october-northeast-storm-epic-incredible-downright-ridiculous/2011/10/31/gIQApy7LZM_blog.html
- Urgency state in New Jersey, two rail service closed in NY area

Occupy Wall Street → NO

- Sat 8 & Sat 1 in Union Square
  - http://live.nydailynews.com/Event/Occupy_Wall_Street_Protests_Rock_New_York_City?Page=2
- 5: 15k demonstrators from Foley Square to Zuccotti Park *Wikipedia* → ***NO, different location***
  https://en.wikipedia.org/wiki/Timeline_of_Occupy_Wall_Street#October_2011
- 15*:* thousands protestors from downtown Manhattan to Times Square Armed Forces recruiting station, Broadway? → ***NO, different location***
  - *http://abcnews.go.com/Business/occupy-wall-street-movement-worldwide/story?id=14743648*

Event research protocol & conclusions

Google

- “october DD” “2011” nyc lower east side
- “october DD” “2011” nyc park avenue, etc
- Check 1st page of results
- Nothing found

Research from events

- Occupy Wall Street
- Columbus Parade
- Halloween Parade

  → No match found

Conclusions

- Detected events are not easily interpretable
  - Traffic perturbation do not always coincide with events
- It is strange that Halloween Parade and Columbus Day Parade are not detected

Possible reasons

- Need to double check the dates: shift?
  - Raw file: taxi1110.csv → format dates → Urbane aggregation → Writecounts python = hourly + daily → SaTScan Wizard = import file + aggregation window
  - Error in reading dates
  - Error in aggregation
    - Urbane code
    - Python hourly to daily aggregation
    - SaTScan count file importation Wizard
    - SaTScan hourly to daily aggregation
- Too big spatial window? → New experiments with smaller window