Detection Algorithms for Biosurveillance: A tutorial

*Tutorial slides by Andrew Moore*

RODS: <http://www.health.pitt.edu/rods> Auton Lab: [http://www.autonlab.org](http://www.autonlab.org/)

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 1

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew’s tutorials: [http://www.cs.cmu.edu/~awm/tutorials](http://www.cs.cmu.edu/%7Eawm/tutorials) .

Comments and corrections gratefully received.

**Many Methods!**

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 2

*Details of these methods and bibliography available from “Summary of Biosurveillance-relevant statistical and data mining technologies” by Moore, Cooper, Tsui and Wagner. Downloadable (PDF format) from* [www.cs.cmu.edu/~awm/biosurv-methods.pdf](http://www.cs.cmu.edu/%7Eawm/biosurv-methods.pdf)

|  |  |
| --- | --- |
| Greg Cooper | Professor Computer Science and [gfc@cbmi.upmc.edu](mailto:gfc@cbmi.upmc.edu) RODS lab, U. Pitt |
| Bill Hogan | Assistant Professor RODS lab, U. Pitt [wrh@cbmi.pitt.edu](mailto:wrh@cbmi.pitt.edu) |
| Andrew Moore | Professor and co-director of Computer Science, [awm@cs.cmu.edu](mailto:awm@cs.cmu.edu) Auton Lab Carnegie Mellon |
| Robin Sabhnani | Senior Programmer, Auton Robotics Institute, [sabhnani@cs.cmu.edu](mailto:sabhnani@cs.cmu.edu) Lab Carnegie Mellon |
| Rich Tsui | Research Professor and RODS lab, U. Pitt [tsui@cbmi.pitt.edu](mailto:tsui@cbmi.pitt.edu) associate Director of RODS lab |
| Mike Wagner | Professor and Director of RODS lab, U. Pitt [mmw@cbmi.pitt.edu](mailto:mmw@cbmi.pitt.edu) RODS lab |
| Weng-Keen Wong | Graduate Student, Auton Lab Computer Science, [wkw@cs.cmu.edu](mailto:wkw@cs.cmu.edu)  Carnegie Mellon |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Has  Pitt/CMU tried it? | Tried  but little used | Tried  and used | Under development | Multivariate  signal tracking? | Spatial  ? |
| Time-weighted averaging | Yes | Yes |  |  |  |  |
| Serfling | Yes |  | Yes |  |  |  |
| ARIMA | Yes | Yes |  |  |  |  |
| SARIMA + External Factors | Yes |  | Yes |  |  |  |
| Univariate HMM | Yes |  | Yes |  |  |  |
| Kalman Filter | Yes | Yes |  |  |  |  |
| Recursive Least Squares | Yes |  | Yes |  |  |  |
| Support Vector Machine | Yes | Yes |  |  |  |  |
| Neural Nets | Yes | Yes |  |  |  |  |
| Randomization | Yes |  | Yes | Yes |  |  |
| Spatial Scan Statistics | Yes |  |  | (w/ Howard Burkom) | Yes | Yes |
| Bayesian Networks | Yes |  |  | Yes | Yes |  |
| Contingency Tables | Yes |  | Yes |  |  |  |
| Scalar Outlier (SQC) | Yes | Yes |  |  |  |  |
| Multivariate Anomalies | Yes |  | Yes |  | Yes |  |
| Change-point statistics | Yes |  |  | Yes |  |  |
| FDR Tests | Yes |  | Yes |  | Yes |  |
| WSARE (Recent patterns) | Yes |  | Yes | Yes | Yes | Yes |
| PANDA (Causal Model) | Yes |  |  | Yes | Yes | Yes |
| FLUMOD (space/Time HMM) |  |  |  | Yes | Yes | Yes |

Biosurveillance Detection Algorithms: Slide 3

Copyright © 2002, 2003, Andrew Moore

What you’ll learn about

* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new patterns

Biosurveillance Detection Algorithms: Slide 4

Copyright © 2002, 2003, Andrew Moore

What you’ll learn about

* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…

Biosurveillance Detection Algorithms: Slide 5

Copyright © 2002, 2003, Andrew Moore

What you’ll learn about

* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…

*Boring Names.*

Biosurveillance Detection Algorithms: Slide 6

Copyright © 2002, 2003, Andrew Moore

Spatial Scan Statistics

WSARE

Multivariate Anomaly Detection

patterns

Univariate Anomaly Detection

These are all powerful statistical methods, which means they all have to have one thing in common…

*Boring Names.*

What you’ll learn about

* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new

Biosurveillance Detection Algorithms: Slide 7

Copyright © 2002, 2003, Andrew Moore

Spatial Scan Statistics

WSARE

Multivariate Anomaly Detection

patterns

Univariate Anomaly Detection

What you’ll learn about

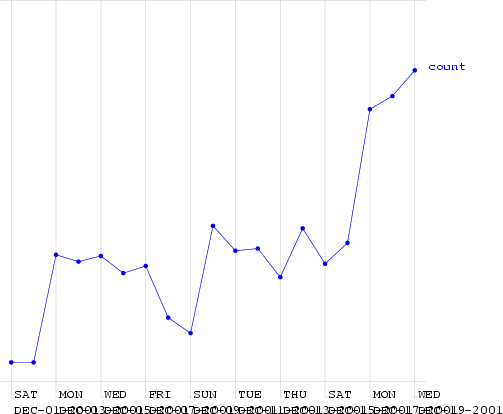
* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new

|  |  |  |
| --- | --- | --- |
| Signal | Univariate Time Series | |
|  |  |
|  | |  |
| Time  Example Signals:   * Number of ED visits today * Number of ED visits this hour * Number of Respiratory Cases Today * School absenteeism today * Nyquil Sales today   Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 8 | | |

Biosurveillance Detection Algorithms: Slide 9

Copyright © 2002, 2003, Andrew Moore

(When) is there an anomaly?

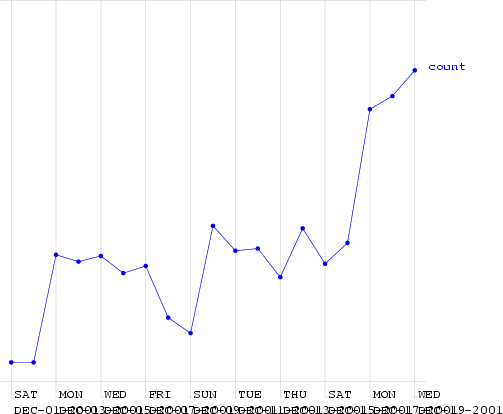


Biosurveillance Detection Algorithms: Slide 10

Copyright © 2002, 2003, Andrew Moore

(When) is there an anomaly?

This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?



Biosurveillance Detection Algorithms: Slide 11

Copyright © 2002, 2003, Andrew Moore

(Ramp attack)

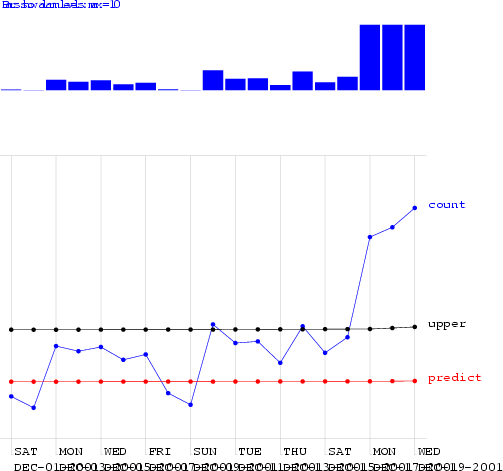
(When) is there an anomaly?

Here (much too high for a Friday)

This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?

|  |  |  |
| --- | --- | --- |
| Signal | An easy case | |
|  |  |
|  | |  |
| Time  Dealt with by Statistical Quality Control  Record the mean and standard deviation up the the current time.  Signal an alarm if we go outside 3 sigmas  Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 12 | | |

|  |  |  |
| --- | --- | --- |
| Signal | An easy case: Control Charts | |
|  | Upper Safe Range  Mean |
|  | |  |
| Time  Dealt with by Statistical Quality Control  Record the mean and standard deviation up the the current time.  Signal an alarm if we go outside 3 sigmas  Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 13 | | |

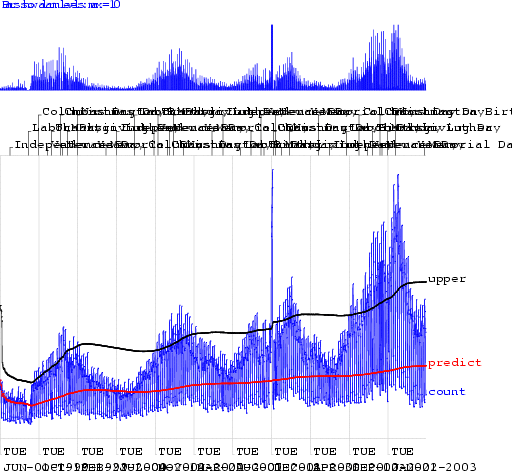


Biosurveillance Detection Algorithms: Slide 14

Copyright © 2002, 2003, Andrew Moore

Control Charts on the Norfolk Data

Alarm Level



Biosurveillance Detection Algorithms: Slide 15

Copyright © 2002, 2003, Andrew Moore

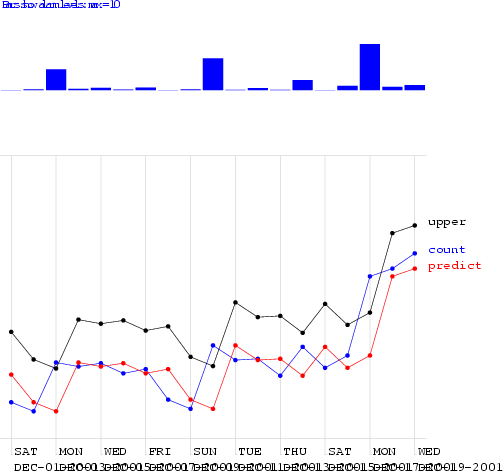
Control Charts on the Norfolk Data

Alarm Level

Biosurveillance Detection Algorithms: Slide 16

Copyright © 2002, 2003, Andrew Moore

Looking at changes from yesterday

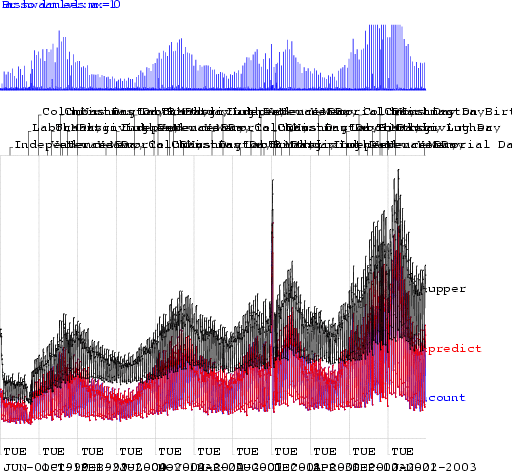


Biosurveillance Detection Algorithms: Slide 17

Copyright © 2002, 2003, Andrew Moore

Looking at changes from yesterday

Alarm Level

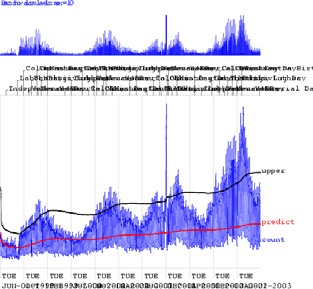
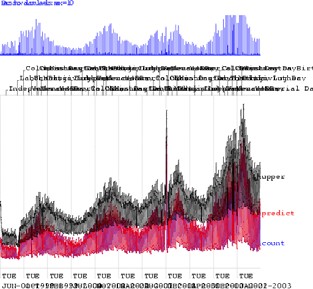


Biosurveillance Detection Algorithms: Slide 18

Copyright © 2002, 2003, Andrew Moore

Looking at changes from yesterday

Alarm Level



Biosurveillance Detection Algorithms: Slide 19

Copyright © 2002, 2003, Andrew Moore

Change from yesterday: Too sensitive to recent changes

Control Chart:

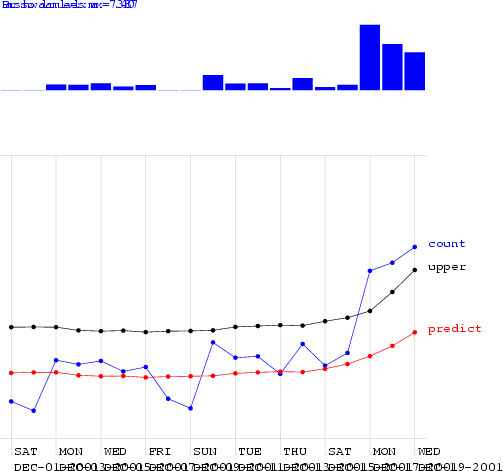
Too insensitive to recent changes

We need a happy medium:

Biosurveillance Detection Algorithms: Slide 20

Copyright © 2002, 2003, Andrew Moore

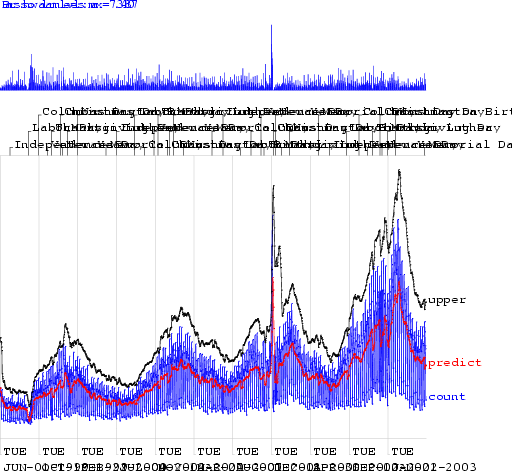
Moving Average



Biosurveillance Detection Algorithms: Slide 21

Copyright © 2002, 2003, Andrew Moore

Moving Average



Biosurveillance Detection Algorithms: Slide 22

Copyright © 2002, 2003, Andrew Moore

Moving Average

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

Moving Average 3 0.36 3.45 0.33 3.79

Moving Average 7 0.58 2.79 0.51 3.31

Moving Average 56 0.54 2.72 0.44 3.54

hours\_of\_daylight 0.58 2.73 0.43 3.9

hours\_of\_daylight is\_mon 0.7 2.25 0.57 3.12

hours\_of\_daylight is\_mon ... is\_tue 0.72 1.83 0.57 3.16

hours\_of\_daylight is\_mon ... is\_sat 0.77 2.11 0.59 3.26

CUSUM 0.45 2.03 0.15 3.55

sa-mav-1 0.86 1.88 0.74 2.73

sa-mav-7 0.87 1.28 0.83 1.87

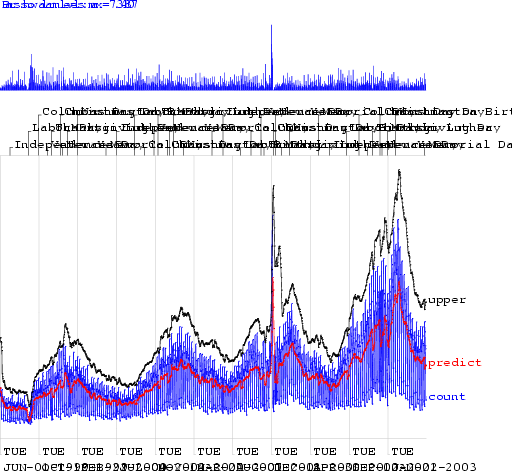
sa-mav-14 0.86 1.27 0.82 1.62

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 24



Biosurveillance Detection Algorithms: Slide 23

Copyright © 2002, 2003, Andrew Moore

Moving Average



Algorithm Performance

i

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

Moving Average 56 0.54 2.72 0.44 3.54

hours\_of\_daylight 0.58 2.73 0.43 3.9

hours\_of\_daylight is\_mon 0.7 2.25 0.57 3.12

hours\_of\_daylight is\_mon ... is\_tue 0.72 1.83 0.57 3.16

hours\_of\_daylight is\_mon ... is\_sat 0.77 2.11 0.59 3.26

CUSUM 0.45 2.03 0.15 3.55

sa-mav-1 0.86 1.88 0.74 2.73

sa-mav-7 0.87 1.28 0.83 1.87

sa-mav-14 0.86 1.27 0.82 1.62

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 25

Moving Average 3 0.36 3.45 0.33 3.79

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

hours\_of\_daylight 0.58 2.73 0.43 3.9

hours\_of\_daylight is\_mon 0.7 2.25 0.57 3.12

hours\_of\_daylight is\_mon ... is\_tue 0.72 1.83 0.57 3.16

hours\_of\_daylight is\_mon ... is\_sat 0.77 2.11 0.59 3.26

CUSUM 0.45 2.03 0.15 3.55

sa-mav-1 0.86 1.88 0.74 2.73

sa-mav-7 0.87 1.28 0.83 1.87

sa-mav-14 0.86 1.27 0.82 1.62

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 26



Algorithm Performance

i



Algorithm Performance

i

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

hours\_of\_daylight is\_mon 0.7 2.25 0.57 3.12

hours\_of\_daylight is\_mon ... is\_tue 0.72 1.83 0.57 3.16

hours\_of\_daylight is\_mon ... is\_sat 0.77 2.11 0.59 3.26

CUSUM 0.45 2.03 0.15 3.55

sa-mav-1 0.86 1.88 0.74 2.73

sa-mav-7 0.87 1.28 0.83 1.87

sa-mav-14 0.86 1.27 0.82 1.62

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 28



Algorithm Performance

i

|  |  |  |
| --- | --- | --- |
| Signal | Seasonal Effects | |
|  |  |
|  | |  |
| Time  Fit a periodic function (e.g. sine wave) to previous data. Predict today’s signal and 3-sigma confidence intervals. Signal an alarm if we’re off.  Reduces False alarms from Natural outbreaks. Different times of year deserve different thresholds.  Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 27 | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |
| hours\_of\_daylight | 0.58 | 2.73 | 0.43 | 3.9 |

Biosurveillance Detection Algorithms: Slide 29

A simple form of ANOVA

Copyright © 2002, 2003, Andrew Moore

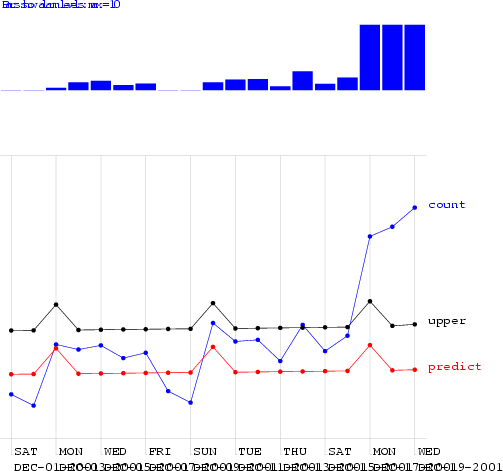
Day-of-week effects

Fit a day-of-week component E[Signal] = a + deltaday

E.G: deltamon= +5.42, deltatue= +2.20, deltawed=

+3.33, deltathu= +3.10, deltafri= +4.02,

deltasat= -12.2, deltasun= -23.42



Biosurveillance Detection Algorithms: Slide 30

Copyright © 2002, 2003, Andrew Moore

Regression using Hours-in-day & IsMonday

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

hours\_of\_daylight is\_mon ... is\_tue 0.72 1.83 0.57 3.16

hours\_of\_daylight is\_mon ... is\_sat 0.77 2.11 0.59 3.26

CUSUM 0.45 2.03 0.15 3.55

sa-mav-1 0.86 1.88 0.74 2.73

sa-mav-7 0.87 1.28 0.83 1.87

sa-mav-14 0.86 1.27 0.82 1.62

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

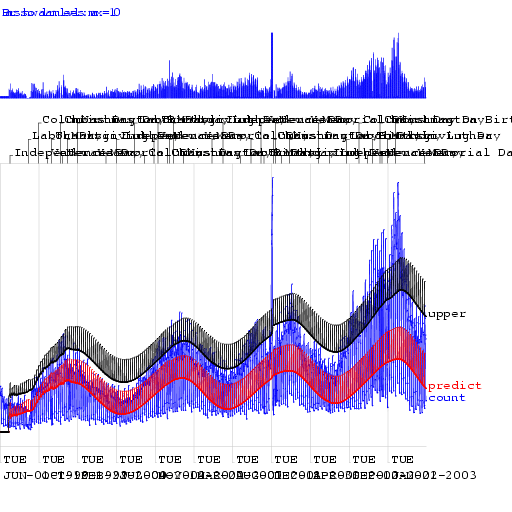
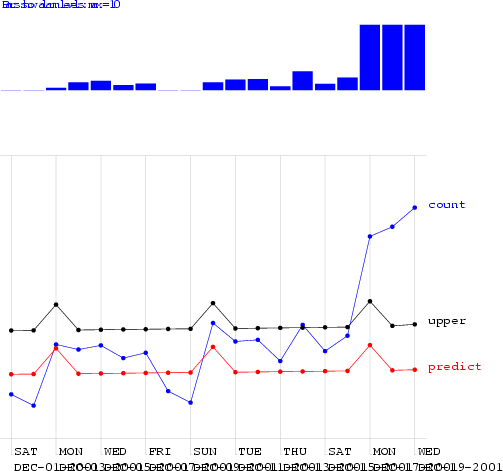
Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 32



Algorithm Performance

i



Biosurveillance Detection Algorithms: Slide 31

Copyright © 2002, 2003, Andrew Moore

Regression using Hours-in-day & IsMonday

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |
| hours\_of\_daylight | 0.58 | 2.73 | 0.43 | 3.9 |
| hours\_of\_daylight is\_mon | 0.7 | 2.25 | 0.57 | 3.12 |

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

CUSUM 0.45 2.03 0.15 3.55

sa-mav-1 0.86 1.88 0.74 2.73

sa-mav-7 0.87 1.28 0.83 1.87

sa-mav-14 0.86 1.27 0.82 1.62

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

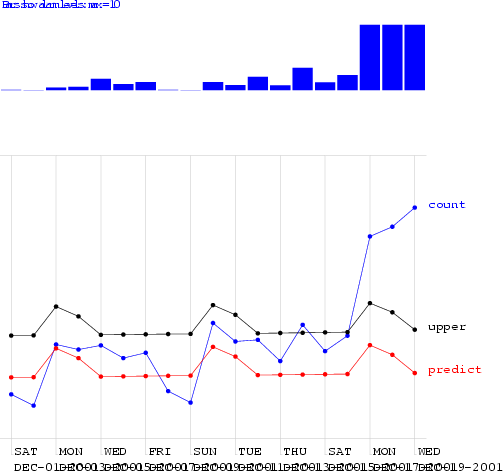
Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 34



Algorithm Performance

i



Biosurveillance Detection Algorithms: Slide 33

Copyright © 2002, 2003, Andrew Moore

Regression using Mon-Tue

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |
| hours\_of\_daylight | 0.58 | 2.73 | 0.43 | 3.9 |
| hours\_of\_daylight is\_mon | 0.7 | 2.25 | 0.57 | 3.12 |
| hours\_of\_daylight is\_mon ... is\_tue | 0.72 | 1.83 | 0.57 | 3.16 |
| hours\_of\_daylight is\_mon ... is\_sat | 0.77 | 2.11 | 0.59 | 3.26 |

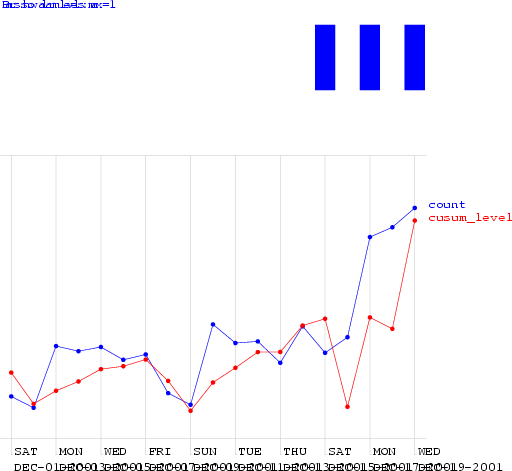
Biosurveillance Detection Algorithms: Slide 35

Copyright © 2002, 2003, Andrew Moore

* Keep a running sum of “surprises”: a sum of excesses each day over the prediction
* When this sum exceeds threshold, signal alarm and reset sum

CUSUM

* CUmulative SUM Statistics



Biosurveillance Detection Algorithms: Slide 36

Copyright © 2002, 2003, Andrew Moore

CUSUM

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

sa-mav-1 0.86 1.88 0.74 2.73

sa-mav-7 0.87 1.28 0.83 1.87

sa-mav-14 0.86 1.27 0.82 1.62

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

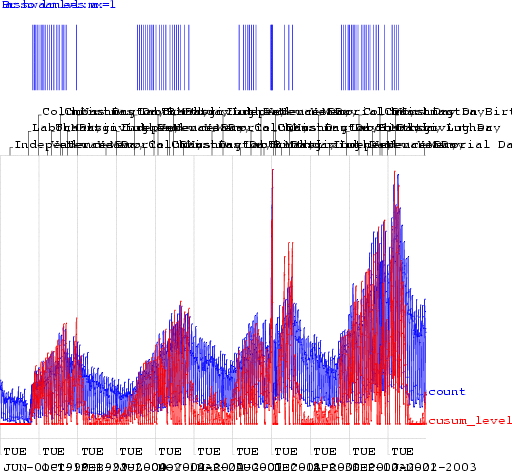
Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 38



Algorithm Performance

i



Biosurveillance Detection Algorithms: Slide 37

Copyright © 2002, 2003, Andrew Moore

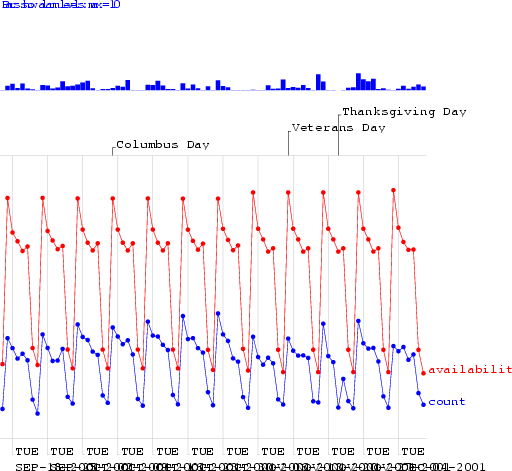
CUSUM

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |
| hours\_of\_daylight | 0.58 | 2.73 | 0.43 | 3.9 |
| hours\_of\_daylight is\_mon | 0.7 | 2.25 | 0.57 | 3.12 |
| hours\_of\_daylight is\_mon ... is\_tue | 0.72 | 1.83 | 0.57 | 3.16 |
| hours\_of\_daylight is\_mon ... is\_sat | 0.77 | 2.11 | 0.59 | 3.26 |
| CUSUM | 0.45 | 2.03 | 0.15 | 3.55 |

Biosurveillance Detection Algorithms: Slide 39

Copyright © 2002, 2003, Andrew Moore

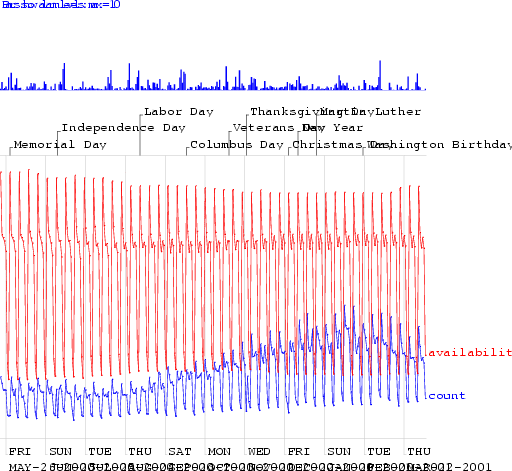
The Sickness/Availability Model



Biosurveillance Detection Algorithms: Slide 40

Copyright © 2002, 2003, Andrew Moore

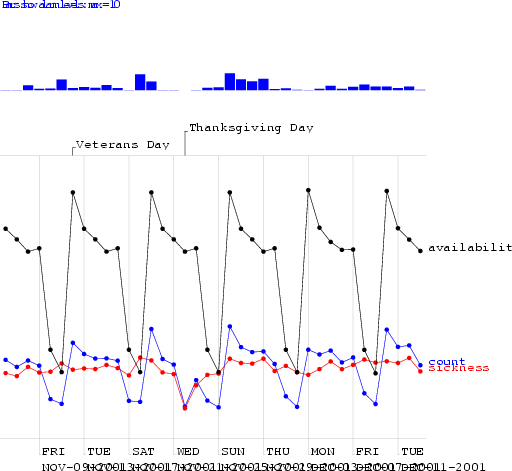
The Sickness/Availability Model



Biosurveillance Detection Algorithms: Slide 41

Copyright © 2002, 2003, Andrew Moore

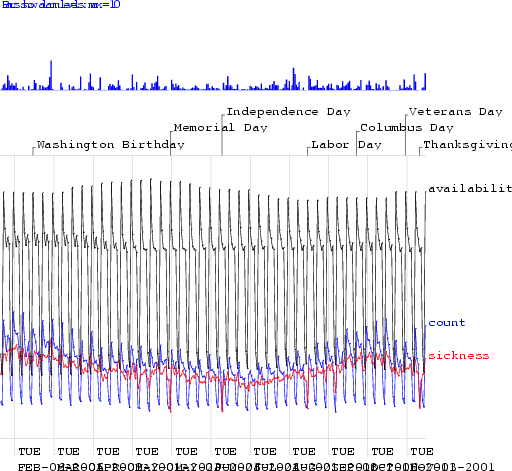
The Sickness/Availability Model



Biosurveillance Detection Algorithms: Slide 42

Copyright © 2002, 2003, Andrew Moore

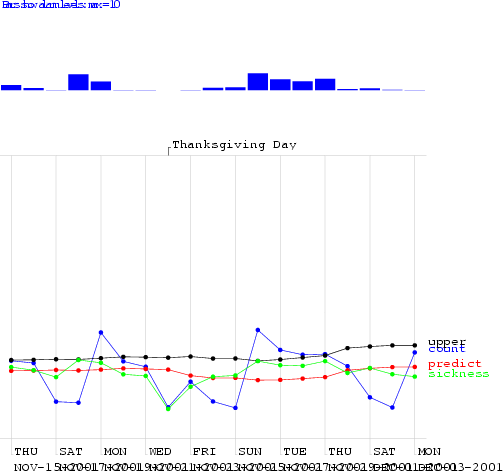
The Sickness/Availability Model



Biosurveillance Detection Algorithms: Slide 43

Copyright © 2002, 2003, Andrew Moore

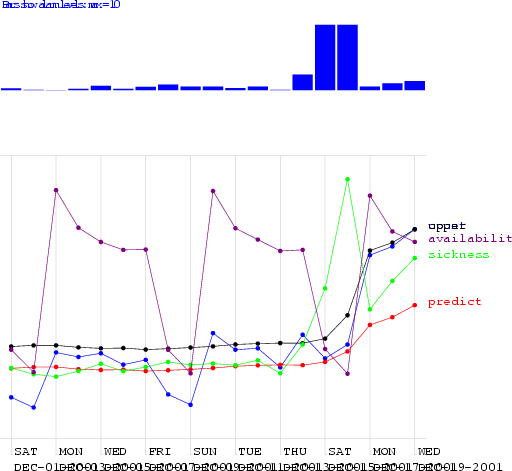
The Sickness/Availability Model



Biosurveillance Detection Algorithms: Slide 44

Copyright © 2002, 2003, Andrew Moore

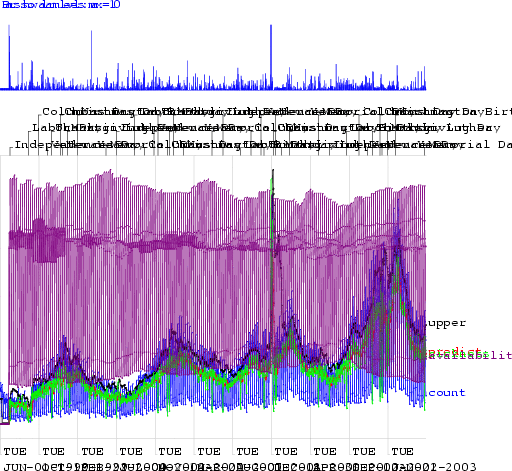
The Sickness/Availability Model



Biosurveillance Detection Algorithms: Slide 45

Copyright © 2002, 2003, Andrew Moore

The Sickness/Availability Model



Biosurveillance Detection Algorithms: Slide 46

Copyright © 2002, 2003, Andrew Moore

The Sickness/Availability Model

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

sa-regress 0.73 1.76 0.67 2.21

Cough with denominator 0.78 2.15 0.59 2.41

Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 47

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d

Cough with denominator 0.78 2.15 0.59 2.41

Cough with MA 0.65 2.78 0.57 3.24

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 48



Algorithm Performance

i

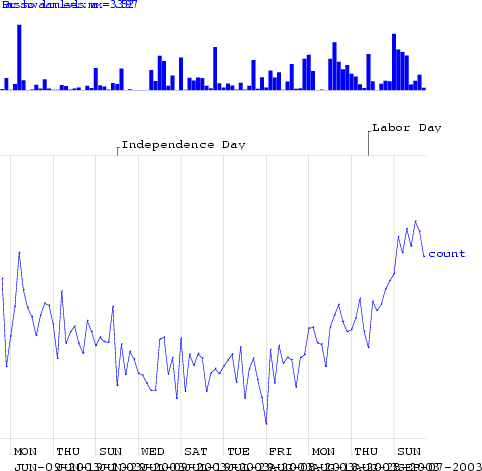


Algorithm Performance

i

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |
| hours\_of\_daylight | 0.58 | 2.73 | 0.43 | 3.9 |
| hours\_of\_daylight is\_mon | 0.7 | 2.25 | 0.57 | 3.12 |
| hours\_of\_daylight is\_mon ... is\_tue | 0.72 | 1.83 | 0.57 | 3.16 |
| hours\_of\_daylight is\_mon ... is\_sat | 0.77 | 2.11 | 0.59 | 3.26 |
| CUSUM | 0.45 | 2.03 | 0.15 | 3.55 |
| sa-mav-1 | 0.86 | 1.88 | 0.74 | 2.73 |
| sa-mav-7 | 0.87 | 1.28 | 0.83 | 1.87 |
| sa-mav-14 | 0.86 | 1.27 | 0.82 | 1.62 |

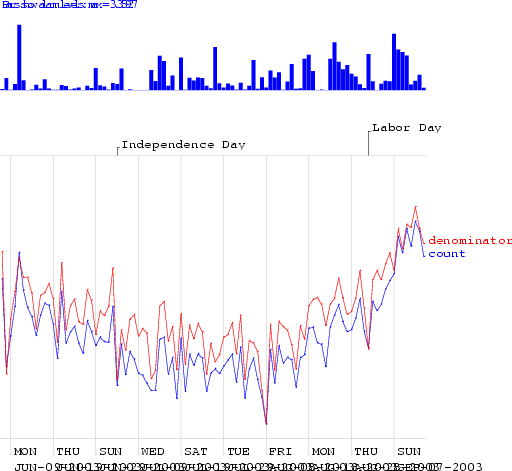
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |
| hours\_of\_daylight | 0.58 | 2.73 | 0.43 | 3.9 |
| hours\_of\_daylight is\_mon | 0.7 | 2.25 | 0.57 | 3.12 |
| hours\_of\_daylight is\_mon ... is\_tue | 0.72 | 1.83 | 0.57 | 3.16 |
| hours\_of\_daylight is\_mon ... is\_sat | 0.77 | 2.11 | 0.59 | 3.26 |
| CUSUM | 0.45 | 2.03 | 0.15 | 3.55 |
| sa-mav-1 | 0.86 | 1.88 | 0.74 | 2.73 |
| sa-mav-7 | 0.87 | 1.28 | 0.83 | 1.87 |
| sa-mav-14 | 0.86 | 1.27 | 0.82 | 1.62 |
| sa-regress | 0.73 | 1.76 | 0.67 | 2.21 |



Biosurveillance Detection Algorithms: Slide 49

Copyright © 2002, 2003, Andrew Moore

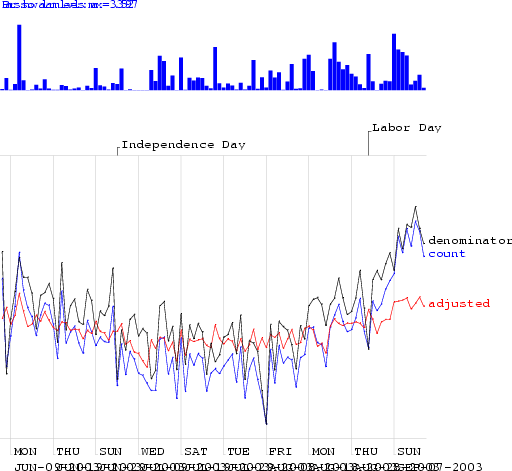
Exploiting Denominator Data



Biosurveillance Detection Algorithms: Slide 50

Copyright © 2002, 2003, Andrew Moore

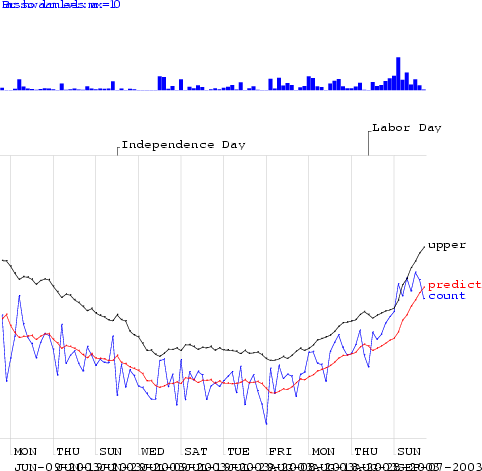
Exploiting Denominator Data



Biosurveillance Detection Algorithms: Slide 51

Copyright © 2002, 2003, Andrew Moore

Exploiting Denominator Data



Biosurveillance Detection Algorithms: Slide 52

Copyright © 2002, 2003, Andrew Moore

Exploiting Denominator Data

Al

low

i

n

g

o

ne

Fa

lse

Al

ar

m

p

er

TW

O

w

eeks…

s

a

a

ay

pi

k

cti

r

s

e

on

am

to

s

o

p

d

e

d

t

f

at

t

e

t

a

e

c

ct

c

k

t

ed

A

ll

o

w

n

g

o

n

e

F

al

s

e

Al

a

r

m

p

er

SI

X

w

eeks…

ra

D

spi

c

a

r

a

ys

k

tio

am

t

e

s

n

p

o

of

d

e

de

t

at

t

te

ec

a

c

ct

te

k

d



Algorithm Performance

i

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 53

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| standard control chart | 0.39 | 3.47 | 0.22 | 4.13 |
| using yesterday | 0.14 | 3.83 | 0.1 | 4.7 |
| Moving Average 3 | 0.36 | 3.45 | 0.33 | 3.79 |
| Moving Average 7 | 0.58 | 2.79 | 0.51 | 3.31 |
| Moving Average 56 | 0.54 | 2.72 | 0.44 | 3.54 |
| hours\_of\_daylight | 0.58 | 2.73 | 0.43 | 3.9 |
| hours\_of\_daylight is\_mon | 0.7 | 2.25 | 0.57 | 3.12 |
| hours\_of\_daylight is\_mon ... is\_tue | 0.72 | 1.83 | 0.57 | 3.16 |
| hours\_of\_daylight is\_mon ... is\_sat | 0.77 | 2.11 | 0.59 | 3.26 |
| CUSUM | 0.45 | 2.03 | 0.15 | 3.55 |
| sa-mav-1 | 0.86 | 1.88 | 0.74 | 2.73 |
| sa-mav-7 | 0.87 | 1.28 | 0.83 | 1.87 |
| sa-mav-14 | 0.86 | 1.27 | 0.82 | 1.62 |
| sa-regress | 0.73 | 1.76 | 0.67 | 2.21 |
| Cough with denominator | 0.78 | 2.15 | 0.59 | 2.41 |
| Cough with MA | 0.65 | 2.78 | 0.57 | 3.24 |

Biosurveillance Detection Algorithms: Slide 54

Copyright © 2002, 2003, Andrew Moore

Other state-of-the-art methods

* Wavelets
* Change-point detection
* Kalman filters
* Hidden Markov Models

Biosurveillance Detection Algorithms: Slide 55

Copyright © 2002, 2003, Andrew Moore

Spatial Scan Statistics

WSARE

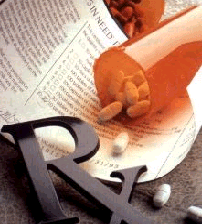
Multivariate Anomaly Detection

patterns

Univariate Anomaly Detection

What you’ll learn about

* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new



Biosurveillance Detection Algorithms: Slide 56

Copyright © 2002, 2003, Andrew Moore

Multiple Signals

**L a b**

**F l u**

**W e b M D**

**S c h o o l**

**C o u g h & C o l d**

**C o u g h S y r u p**

**R e s p**

**V i r a l**

**D e a t h**

0

7/1/99 10/1/ 99 1/1/00 4/1/00 7/ 1/00 10/1/00 1/1/01

Copyright © 2002, 2003, Andrew Moore date Biosurveillance Detection Algorithms: Slide 57

500

1000

1500

Multivariate Signals

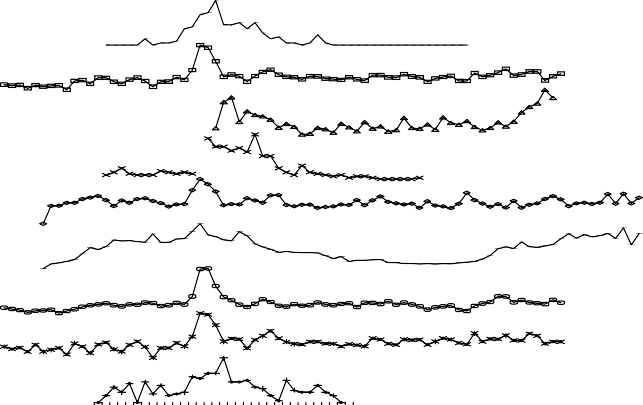
(relevant to inhalational diseases)

c ough.s y r.liq.dec

tabs .c aps

throat.c ough nas al

2000



Multi Source Signals

**F o o t p r i n t o f I n f l u e n z a i n R o u t i n e l y C o l l e c t e d D a t a**

Lab

Flu WebMD School

Cough&

Cold Throat Resp Viral

Death

2 7 3 1 3 5 3 9 4 3 4 7 5 1 3 7

weeks

1 1 1 5 1 9 2 3 2 7 3 1 3 5 3 9 4 3 4 7 5 1 3

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 58

daily sales

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

Another View

Red: Cough Sales

Blue: ED Respiratory Visits

Question: why might that not be the

best we can do?

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 60

ED Respiratory Visits

Signal

Cough Sales

|  |  |  |
| --- | --- | --- |
| Signal | What if you’ve got multiple signals? | |
|  | Red: Cough Sales  Blue: ED Respiratory Visits |
|  | |  |
| Time  Idea One:  Simply treat it as two separate alarm-from- signal problems.  …Question: why might that not be the best we can do?  Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 59 | | |

Another View

Red: Cough Sales

Blue: ED Respiratory Visits

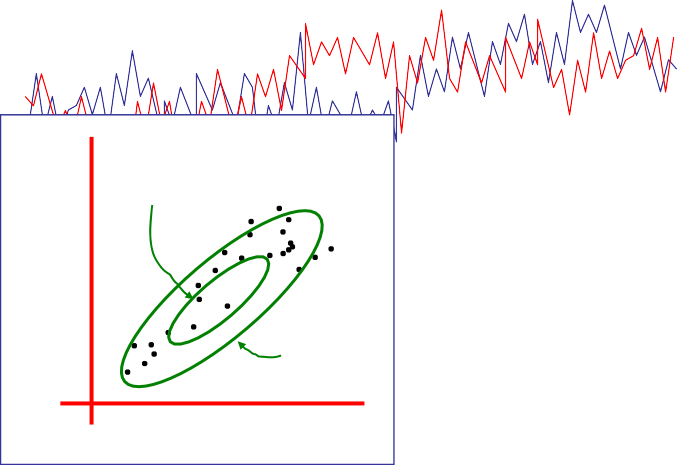
Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 61

ED Respiratory Visits

This should be an anomaly

Question: why might that not be the best we can do?



One Sigma

2 Sigma

ED Respiratory Visits

Biosurveillance Detection Algorithms: Slide 62

Copyright © 2002, 2003, Andrew Moore

Good Practical Idea:

Model the joint with a Gaussian This is a sensible N-dimensional

SQC

…But you can also do N- dimensional modeling of dynamics (leads to the idea of Kalman Filter model)

N-dimensional Gaussian

Red: Cough Sales

Blue: ED Respiratory Visits

Signal

Signal

Cough Sales

Cough Sales

But what if joint N-dimensional distribution is highly non-Gaussian?

Red: Cough Sales

Blue: ED Respiratory Visits

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 63

ED Respiratory Visits

Signal

Cough Sales

Biosurveillance Detection Algorithms: Slide 64

Copyright © 2002, 2003, Andrew Moore

Spatial Scan Statistics

WSARE

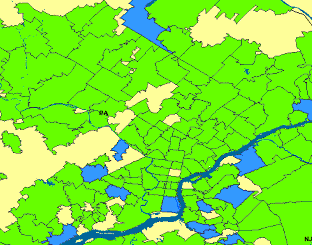
Multivariate Anomaly Detection

patterns

Univariate Anomaly Detection

What you’ll learn about

* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new

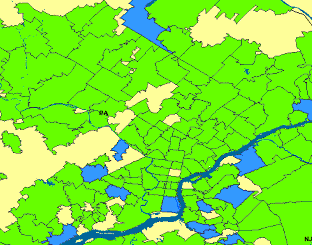


Biosurveillance Detection Algorithms: Slide 65

Copyright © 2002, 2003, Andrew Moore

One Step of Spatial Scan

Entire area being scanned



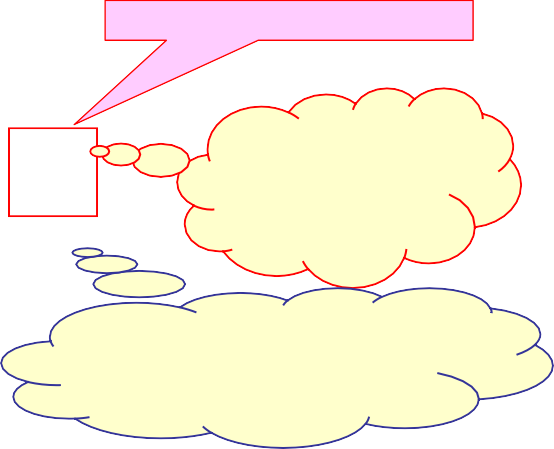
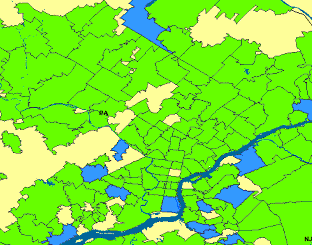
Biosurveillance Detection Algorithms: Slide 66

Copyright © 2002, 2003, Andrew Moore

Current region being considered

One Step of Spatial Scan

Entire area being scanned



Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

Biosurveillance Detection Algorithms: Slide 67

Copyright © 2002, 2003, Andrew Moore

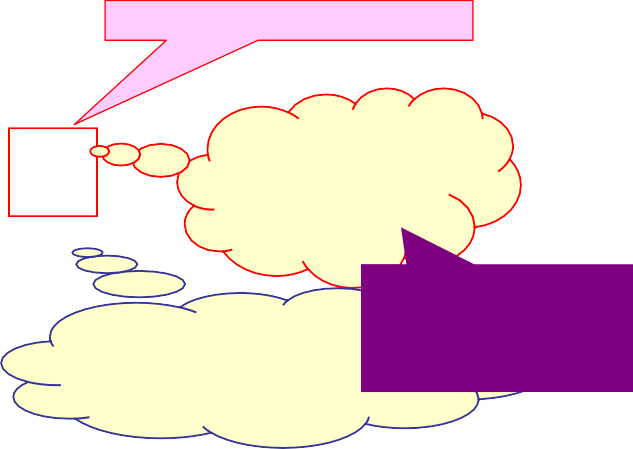
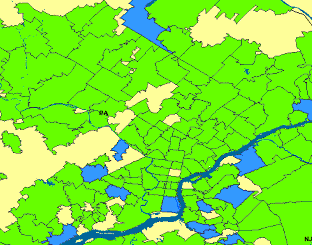
I have a population of 5300 of whom 53

are sick (1%)

Current region being considered

One Step of Spatial Scan

Entire area being scanned



Biosurveillance Detection Algorithms: Slide 68

Copyright © 2002, 2003, Andrew Moore

*function (e.g. Kulldorf’s*

population of 2,200,000 of *score)*

whom 20,000 are sick (0.9%)

***So...*** *is that a big deal? Evaluated with Score*

Everywhere else has a

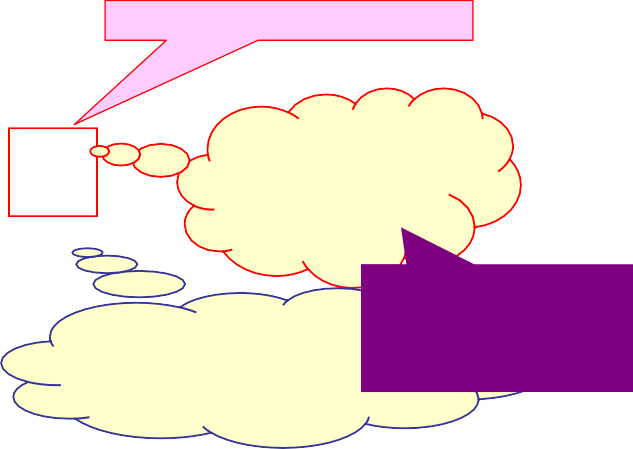
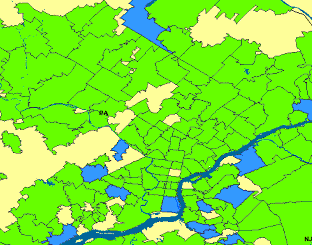
I have a population of 5300 of whom 53

are sick (1%)

Current region being considered

One Step of Spatial Scan

Entire area being scanned



Biosurveillance Detection Algorithms: Slide 69

Copyright © 2002, 2003, Andrew Moore

*function (e.g. Kulldorf’s*

population of 2,200,000 of *score)*

whom 20,000 are sick (0.9%)

***So...*** *is that a big deal? Evaluated with Score*

Everywhere else has a

I have a population of 5300 of whom 53

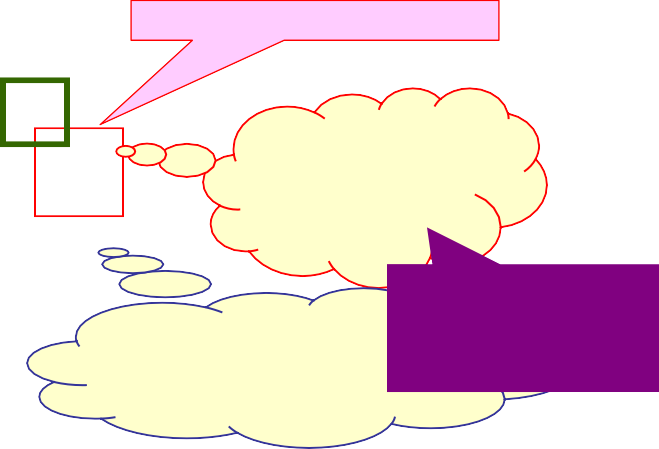
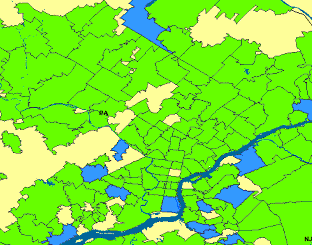
are sick (1%)

[Score = 1.4]

Current region being considered

One Step of Spatial Scan

Entire area being scanned



Biosurveillance Detection Algorithms: Slide 70

Copyright © 2002, 2003, Andrew Moore

*function (e.g. Kulldorf’s*

population of 2,200,000 of *score)*

whom 20,000 are sick (0.9%)

***So...*** *is that a big deal? Evaluated with Score*

Everywhere else has a

I have a population of 5300 of whom 53

are sick (1%)

[Score = 1.4]

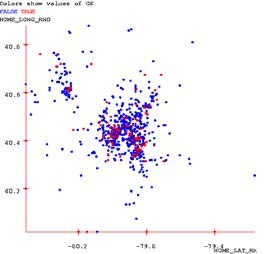
[Score = 9.3]

Many Steps of Spatial Scan

Entire area being scanned

Highest scoring region in search so far

Current region being considered



the max, S\* W

See [Glaz and Balakrishnan, 99] for details

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 71

Compute the likelihood of the data given the hypothesis that the rate of occurrence is uniform everywhere, L0

For some geographical region, W, compute the likelihood that the rate of occurrence is uniform at one level inside the region and uniform at another level outside the region, L(W).

Compute the likelihood ratio, L(W)/L0

Repeat for all regions, and find the largest likelihood ratio. This is the scan statistic, S\*W

Report the region, W, which yielded

5.

3.

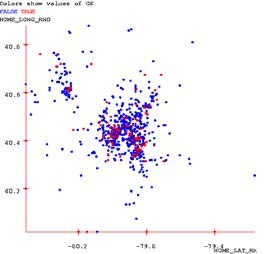
4.

Standard scan statistic question: Given the geographical locations of occurrences of a phenomenon, is there a region with an unusually high (low) rate of these occurrences?

Scan Statistics

Standard approach: 1.

2.



Biosurveillance Detection Algorithms: Slide 72

Copyright © 2002, 2003, Andrew Moore

Generate many randomized versions of the data set by shuffling the labels (positive instance of the phenomenon or not).

Compute S\*W for each randomized data set. This forms a baseline distribution for S\*W if the null hypothesis holds.

Compare the observed value of S\*W against the baseline distribution to determine a p-value.

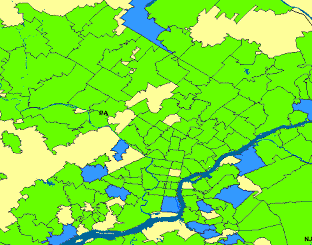
Given that region W is the most likely to be abnormal, is it significantly abnormal?

Significance testing

Standard approach: 1.

2.

3.



N

Fast squares speedup

N

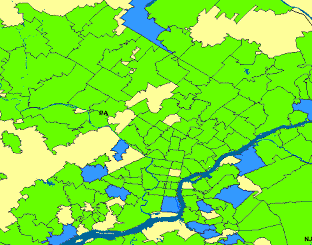
* Theoretical complexity of fast squares: O(N2) (as opposed to naïve N3), if maximum density region sufficiently dense.

*If not, we can use several other speedup tricks.*

* In practice: 10-200x speedups on real and artificially generated datasets.

*Emergency Dept. dataset (600K records): 20 minutes, versus 66 hours with naïve approach.*

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 73



N

Fast rectangles speedup

N

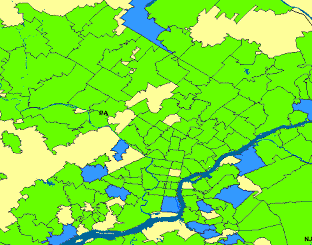
* Theoretical complexity of fast rectangles: O(N2log N) (as opposed to naïve N4)

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 74

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  | | | | | | | | |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  | | | | | | | | |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |



N

Fast oriented rectangles speedup

N

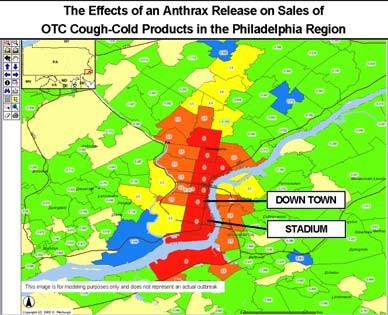
* Theoretical complexity of fast rectangles: 18N2log N (as opposed to naïve 18N4)

Copyright © 2002, 2003, Andrew Moore

*(Angles discretized to 5 degree buckets)*

Biosurveillance Detection Algorithms: Slide 75

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  | | | | | | | | |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

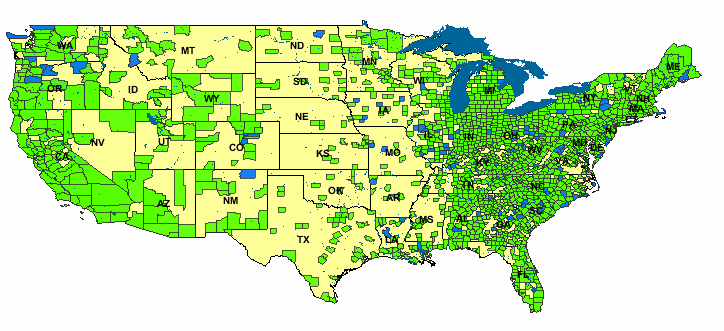
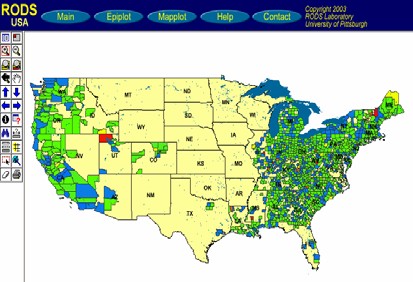


Biosurveillance Detection Algorithms: Slide 76

* Traditional Scan Statistics very expensive, especially with Randomization tests
* New “Historical Model” Scan Statistics
* Proposed new WSARE/Scan Statistic hybrid

Copyright © 2002, 2003, Andrew Moore

Why the Scan Statistic speed obsession?



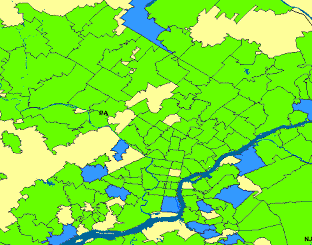
Biosurveillance Detection Algorithms: Slide 77

* Traditional Scan Statistics very expensive, especially with Randomization tests
* New “Historical Model” Scan Statistics
* Proposed new WSARE/Scan Statistic hybrid

Copyright © 2002, 2003, Andrew Moore

Why the Scan Statistic speed obsession?

Why the Scan Statistic speed obsession?



# Traditional Scan Statistics very expensive, especially with Randomization tests

* New “Historical Model” Scan Statistics
* Proposed new WSARE/Scan Statistic hybrid

This is the strangest region because the age distribution of respiratory cases has changed dramatically for no reason that can be explained by known background changes

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 78

Biosurveillance Detection Algorithms: Slide 79

Copyright © 2002, 2003, Andrew Moore

Spatial Scan Statistics

WSARE

Multivariate Anomaly Detection

patterns

Univariate Anomaly Detection

What you’ll learn about

* Noticing events in bio- event time series
* Tracking many series at once
* Detecting geographic hotspots
* Finding emerging new

But there’s potentially more data than aggregates

Suppose we know that today in the ED we had…

* 421 Cases
* 78 Respiratory Cases
* 190 Males
* 32 Children
* 21 from North Suburbs
* 2 Postal workers

(etc etc etc)

Have we made best use of all possible information?

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 80



Biosurveillance Detection Algorithms: Slide 81

**Absenteeism by zipcode Farm Workers Recent month**

Copyright © 2002, 2003, Andrew Moore

**Nyquil Sales by state**

**Recent 30 mins**

**Collapse by county**

**Among Men Recent week**

**Diarrhea by Neighborhood Among Elderly Recent 24 hrs**

**Diarrhea By Street**

**Among Children Recent 3 hours**

There are so many things to look at

Massive Computer Analysis

Human Analysts

Biosurveillance Detection Algorithms: Slide 82

Copyright © 2002, 2003, Andrew Moore

WSARE v2.0

* What’s Strange About Recent Events?
* Designed to be easily applicable to any date/time-indexed biosurveillance-relevant data stream.

Biosurveillance Detection Algorithms: Slide 83

Copyright © 2002, 2003, Andrew Moore

* Inputs:

WSARE v2.0

3. Which attributes to use?

2. Time Window Length

1. Date/time-indexed biosurveillance- relevant data stream

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| WSARE v2.0   * Input 1. Date/time-indexed 2. Time Window 3. Which   s: biosurveillance- Length attributes to use? relevant data stream  **Example** “last 24 hours” “ignore key and  weather” | | | | | | | | | | | | | | | | |
| Primary Key | Date | Time | Hospital | ICD9 | Prodrome | Gender | Age | Home | | | Work | | | Recent Flu Levels | Recent Weather | (Many more…) |
| Large Scale | Medium Scale | Fine Scale | Large Scale | Medium Scale | Fine Scale |
| h6r32 | 6/2/2 | 14:12 | Down- town | 781 | Fever | M | 20s | NE | 15217 | A5 | NW | 15213 | B8 | 2% | 70R | … |
| t3q15 | 6/2/2 | 14:15 | River- side | 717 | Respirat ory | M | 60s | NE | 15222 | J3 | NE | 15222 | J3 | 2% | 70R | … |
| t5hh5 | 6/2/2 | 14:15 | Smith- field | 622 | Respirat ory | F | 80s | SE | 15210 | K9 | SE | 15210 | K9 | 2% | 70R | … |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 84 | | | | | | | | | | | | | | | | |

Simple WSARE

* Given 500 day’s worth of ER cases at 15 hospitals…

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 86

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| WSARE v2.0   * Inputs: 1. Date/time-indexed 2. Time Window 3. Which biosurveillance- Length attributes to use? relevant data stream   3. And here’s   * Outputs: 1. Here are the 2. Here’s why how seriously you records that most should take it   surprise me | | | | | | | | | | | | | | | | |
| Primary Key | Date | Time | Hospital | ICD9 | Prodrome | Gender | Age | Home | | | Work | | | Recent Flu Levels | Recent Weather | (Many more…) |
| Large Scale | Medium Scale | Fine Scale | Large Scale | Medium Scale | Fine Scale |
| h6r32 | 6/2/2 | 14:12 | Down- town | 781 | Fever | M | 20s | NE | 15217 | A5 | NW | 15213 | B8 | 2% | 70R | … |
| t3q15 | 6/2/2 | 14:15 | River- side | 717 | Respirat ory | M | 60s | NE | 15222 | J3 | NE | 15222 | J3 | 2% | 70R | … |
| t5hh5 | 6/2/2 | 14:15 | Smith- field | 622 | Respirat ory | F | 80s | SE | 15210 | K9 | SE | 15210 | K9 | 2% | 70R | … |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 85 | | | | | | | | | | | | | | | | |

|  |  |
| --- | --- |
| Date | Cases |
| Thu 5/22/2000 | C1, C2, C3, C4 … |
| Fri 5/23/2000 | C1, C2, C3, C4 … |
| : | : |
| : | : |
| Sat 12/9/2000 | C1, C2, C3, C4 … |
| Sun 12/10/2000 | C1, C2, C3, C4 … |
| : | : |
| Sat 12/16/2000 | C1, C2, C3, C4 … |
| : | : |
| Sat 12/23/2000 | C1, C2, C3, C4 … |
| : | : |
| : | : |
| Fri 9/14/2001 | C1, C2, C3, C4 … |

Simple WSARE

* Given 500 day’s worth of ER cases at 15 hospitals…
* For each day…
  + Take today’s cases

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 87

Simple WSARE

* Given 500 day’s worth of ER cases at 15 hospitals…
* For each day…
  + Take today’s cases
  + The cases one week ago
  + The cases two weeks ago

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 88

|  |  |
| --- | --- |
| Date | Cases |
| Thu 5/22/2000 | C1, C2, C3, C4 … |
| Fri 5/23/2000 | C1, C2, C3, C4 … |
| : | : |
| : | : |
| Sat 12/9/2000 | C1, C2, C3, C4 … |
| Sun 12/10/2000 | C1, C2, C3, C4 … |
| : | : |
| Sat 12/16/2000 | C1, C2, C3, C4 … |
| : | : |
| Sat 12/23/2000 | C1, C2, C3, C4 … |
| : | : |
| : | : |
| Fri 9/14/2001 | C1, C2, C3, C4 … |

|  |  |
| --- | --- |
| Date | Cases |
| Thu 5/22/2000 | C1, C2, C3, C4 … |
| Fri 5/23/2000 | C1, C2, C3, C4 … |
| : | : |
| : | : |
| Sat 12/9/2000 | C1, C2, C3, C4 … |
| Sun 12/10/2000 | C1, C2, C3, C4 … |
| : | : |
| Sat 12/16/2000 | C1, C2, C3, C4 … |
| : | : |
| Sat 12/23/2000 | C1, C2, C3, C4 … |
| : | : |
| : | : |
| Fri 9/14/2001 | C1, C2, C3, C4 … |

rth

go

rth

go

Simple WSARE

* Given 500 day’s wo of ER cases at 15 hospitals…
* For each day…
  + Take today’s cases
  + The cases one week ago
  + The cases two weeks a
* Ask: “What’s different about today?”

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 89

Simple WSARE

* Given 500 day’s wo of ER cases at 15 hospitals…
* For each day…
  + Take todayF’iselcdassewse use:
  + The cases one week ago

***Sym***T***p***h***to***e***m***ca***s***s, eAsgtew, oGwenedeekrs, aCoarse Location,

* AFsinke:L“oWcathioant, ’***I***s***CD***d***9***if***D***fe***e***r***r***e***iv***n***ed***t ***Features,***
* Date, Time of Day, Prodr

***C***a***e***b***n***o***s***u***u***t***s*** t***B***o***lo***d***c***a***k***y***D***?***e***”***rived Features, Work***

***Details, Colocation Details***

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 90

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| DATE\_AD | ICD9 | PRODRO | GENDER | place2 | … | … |
|  |  |  |  |  | … | … |
| 12/9/00 | 786.05 | 3 | F | s-e | … | … |
| 12/9/00 | 789 | 1 | F | s-e | … | … |
| 12/9/00 | 789 | 1 | M | n-w | … | … |
| 12/9/00 | 786.05 | 3 | M | s-e | … | … |
| : | : | : | : | : | … | … |
| 12/16/00 | 787.02 | 2 | M | n-e | … | … |
| 12/16/00 | 782.1 | 4 | F | s-w | … | … |
| 12/16/00 | 789 | 1 | M | s-e | … | … |
| 12/16/00 | 786.09 | 3 | M | n-w | … | … |
| 12/23/00 | 789.09 | 1 | M | s-w | … | … |
| 12/23/00 | 789.09 | 1 | F | s-w | … | … |
| 12/23/00 | 782.1 | 4 | M | n-w | … | … |
| : | : | : | : | : | … | … |
| 12/23/00 | 786.09 | 3 | M | s-e | … | … |
| 12/23/00 | 786.09 | 3 | M | s-e | … | … |
| 12/23/00 | 780.9 | 2 | F | n-w | … | … |
| 12/23/00 | V40.9 | 7 | M | s-w | … | … |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| DATE\_AD | ICD9 | PRODRO | GENDER | place2 | … | … |
|  |  |  |  |  | … | … |
| 12/9/00 | 786.05 | 3 | F | s-e | … | … |
| 12/9/00 | 789 | 1 | F | s-e | … | … |
| 12/9/00 | 789 | 1 | M | n-w | … | … |
| 12/9/00 | 786.05 | 3 | M | s-e | … | … |
| : | : | : | : | : | … | … |
| 12/16/00 | 787.02 | 2 | M | n-e | … | … |
| 12/16/00 | 782.1 | 4 | F | s-w | … | … |
| 12/16/00 | 789 | 1 | M | s-e | … | … |
| 12/16/00 | 786.09 | 3 | M | n-w | … | … |
| 12/23/00 | 789.09 | 1 | M | s-w | … | … |
| 12/23/00 | 789.09 | 1 | F | s-w | … | … |
| 12/23/00  : | 782.1 | 4 | M | n-w | … | … |
|  | : | : | : | : | … | … |
| 12/23/00 | 786.09 | 3 | M | s-e | … | … |
| 12/23/00 | 786.09 | 3 | M | s-e | … | … |
| o1m2/23e/00, | IC780D.9 | 9, 2 | F | n-w | … | … |
| 12/23/00 | V40.9 | 7 | M | s-w | … | … |

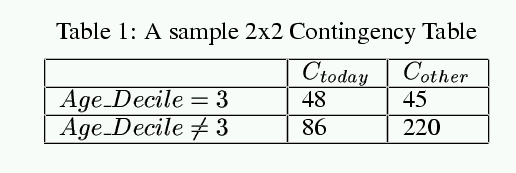
Biosurveillance Detection Algorithms: Slide 91

Copyright © 2002, 2003, Andrew Moore

Example

**Sat 12-23-2001 (daynum 36882, dayindex 239)**

**35.8% ( 48/134) of today's cases have 30 <= age < 40 17.0% ( 45/265) of other cases have 30 <= age < 40**



Biosurveillance Detection Algorithms: Slide 92

Copyright © 2002, 2003, Andrew Moore

Example

**Sat 12-23-2001 (daynum 36882, dayindex 239)**

**FISHER\_PVALUE = 0.000051**

**35.8% ( 48/134) of today's cases have 30 <= age < 40 17.0% ( 45/265) of other cases have 30 <= age < 40**

Biosurveillance Detection Algorithms: Slide 93

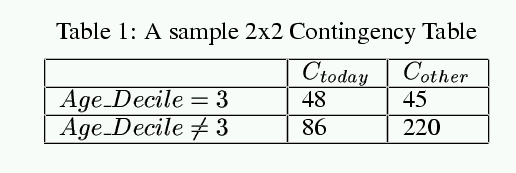
Copyright © 2002, 2003, Andrew Moore

* Try ICD9 = x for each value of x
* Try Gender=M, Gender=F
* Try CoarseRegion=NE, =NW, SE, SW..
* Try FineRegion=AA,AB,AC, … DD (4x4 Grid)
* Try Hospital=x, TimeofDay=x, Prodrome=X,

…

* [In future… features of census blocks]

Searching for the best score…



Biosurveillance Detection Algorithms: Slide 94

Copyright © 2002, 2003, Andrew Moore

Example

**Sat 12-23-2001 (daynum 36882, dayindex 239)**

**FISHER\_PVALUE = 0.000051 RANDOMIZATION\_PVALUE = 0.031**

**35.8% ( 48/134) of today's cases have 30 <= age < 40 17.0% ( 45/265) of other cases have 30 <= age < 40**

Multiple component rules

* We would like to be able to find rules like:

There are a surprisingly large number of children with respiratory problems today

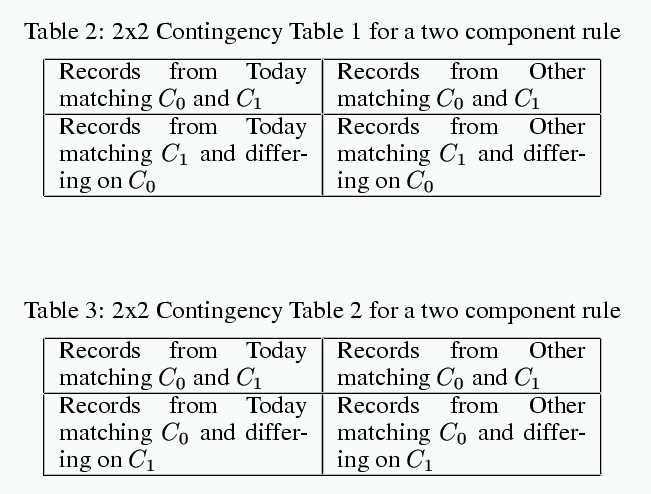
or

There are too many skin complaints among people from the

affluent neighborhoods

* These are things that would be missed by casual screening
* BUT
  + The danger of overfitting could be much worse
  + It’s very computationally demanding
  + How can we be sure the entire rule is meaningful?

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 95



Biosurveillance Detection Algorithms: Slide 96

Copyright © 2002, 2003, Andrew Moore

Checking two component rules

* Must pass both tests to be allowed to live.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| WSARE v2.0   * Inputs: 1. Date/time-indexed 2. Time Window 3. Which biosurveillance- Length attributes to use? relevant data stream   3. And here’s   * Outputs: 1. Here are the 2. Here’s why how seriously you records that most should take it   surprise me | | | | | | | | | | | | | | | | |
| Primary Key | Date | Time | Hospital | ICD9 | Prodrome | Gender | Age | Home | | | Work | | | Recent Flu Levels | Recent Weather | (Many more…) |
| Large Scale | Medium Scale | Fine Scale | Large Scale | Medium Scale | Fine Scale |
| h6r32 | 6/2/2 | 14:12 | Down- town | 781 | Fever | M | 20s | NE | 15217 | A5 | NW | 15213 | B8 | 2% | 70R | … |
| t3q15 | 6/2/2 | 14:15 | River- side | 717 | Respirat ory | M | 60s | NE | 15222 | J3 | NE | 15222 | J3 | 2% | 70R | … |
| t5hh5 | 6/2/2 | 14:15 | Smith- field | 622 | Respirat ory | F | 80s | SE | 15210 | K9 | SE | 15210 | K9 | 2% | 70R | … |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 97 | | | | | | | | | | | | | | | | |

Time

mally,

re ov

s

b

i

70R

n

patter

7e0rRe’s d a rul

Biosurveillance Detection Algorithms: Slide 98

Copyright © 2002, 2003, Andrew Moore

:

an: ce

y: ch

u:st b

m:atic j

d: ra

:this

:

:

:

:

:

:

:

:

:

:

a…

e

2e%r, th

foun

Kg9 ov

have

1a5r2c1h0i

e I’d

SnE se

hanc

K9bee

0% c

1521I’0ve

2

SE

80

s

F

Respira tory

622

Smith- field

14:15

6/2/2

t5hh5

…

ns

the

2%

J3

nt all

15222

accou

into

NE

king

J3

Ta

15222

NE

60

s

M%

Reeesnpir1a5 tory

7t’1s7b

Roidvaery- side

1B4:u1t5t

6/2/2

t3q15

…

d!

70R

2%

resse

imp

B8

S1c5a2le13

be too

on’t

NW

A5 D

S1c5a2le17

NE

20

s

M

Fse.ver

7l8e1m

Dopwrno- town

14:12

6/2/2

h6r32

Fine Scale

Mediu m

Large Scale

Fine Scale

Mediu m

Large Scale

(Many more…

)

Recent Weathe r

Recen t Flu Levels

Work

Home

Ag est

Gende

hr e Ea

atory

Prodrom ees in t h respir

ICD

9ca

wit

Hospita l 8% of er-50s

Date

Nor

a

Primary Key

2. Here’s why

1. Here are the records that most surprise me

* Output s:
* Input s:

WSARE v2.0

3. And here’s how seriously you should take it

3. Which attributes to use?

2. Time Window Length

1. Date/time-indexed biosurveillance- relevant data stream

Biosurveillance Detection Algorithms: Slide 100

Biosurveillance Detection Algorithms: Slide 99

Copyright © 2002, 2003, Andrew Moore

Normally:

0.8% of records (50/6205) have time before 2pm and prodrome = Hemorrhagic But recently:

2.1% of records (19/907) have time before 2pm and prodrome = Hemorrhagic Pvalue = 0.0484042

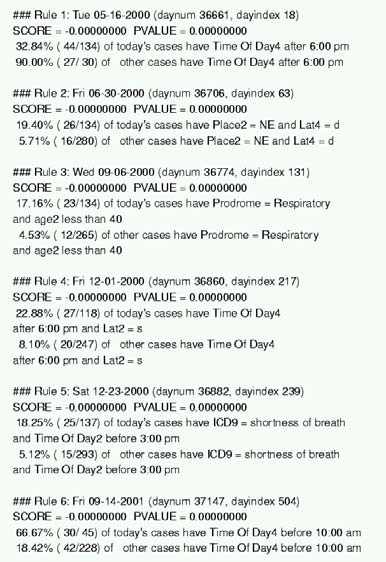
Which means that in a world where nothing changes we'd expect to have a result this significant about once

every 20 times we ran the program

WSARE on recent Utah Data

Saturday June 1st in Utah:

The most surprising thing about recent records is:



Results on Emergency Dept Data

Copyright © 2002, 2003, Andrew Moore

|  |  |  |
| --- | --- | --- |
| WSARE | | 3.0 |
|  | * “Taking into account recent flu levels…” * “Taking into account that today is a public holday…” * “Taking into account that this is Spring…” * “Taking into account recent heatwave…” * “Taking into account that there’s a known natural Food-borne outbreak in progress…” | |
|  | Copyright © 2002, 2003, Andrew Moore | Bonus: More efficient use of historical data  Biosurveillance Detection Algorithms: Slide 101 |

Biosurveillance Detection Algorithms: Slide 102

Copyright © 2002, 2003, Andrew Moore

Analysis of variance

* Good news:

If you’re tracking a daily aggregate (e.g. number of flu cases in your ED, or Nyquil Sales)…then ANOVA can take care of many of these effects.

* But…

What if you’re tracking a whole joint distribution of transactional events?



“On the day after a major holiday, expect a boost in the morning followed by a lull in the afternoon”

Biosurveillance Detection Algorithms: Slide 103

“The Viral prodrome is more likely to co-occur with a Rash prodrome than Botulinic”

Copyright © 2002, 2003, Andrew Moore

“On Cold Tuesday Mornings the folks coming in from the North part of the city are more likely to have respiratory problems”

Idea: Bayesian Networks

“Patients from West Park Hospital are less likely to be young”

Biosurveillance Detection Algorithms: Slide 104

Copyright © 2002, 2003, Andrew Moore

All historical data

WSARE 3.0



Biosurveillance Detection Algorithms: Slide 105

Copyright © 2002, 2003, Andrew Moore

All historical data

WSARE 3.0



Biosurveillance Detection Algorithms: Slide 106

Copyright © 2002, 2003, Andrew Moore

What *should* be happening today?

WSARE 3.0

Today’s Environment

All historical data



Biosurveillance Detection Algorithms: Slide 107

Copyright © 2002, 2003, Andrew Moore

What’s strange about today, considering its environment?

What *should* be happening today?

Today’s Cases

Environment

WSARE 3.0

Today’s

All historical data



And how big a deal is this, considering how

Biosurmveiullacnche DseeteactriocnhAlgIo’rvitehmds:oSlnidee1?08

Copyright © 2002, 2003, Andrew Moore

What’s strange about today, considering its environment?

What *should* be happening today?

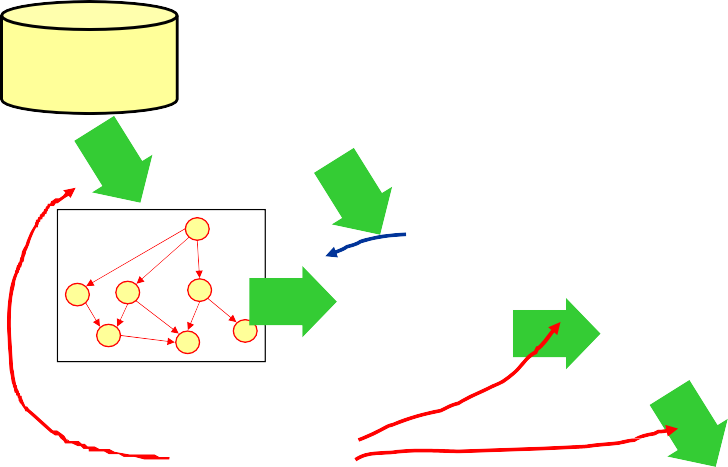
Today’s Cases

Environment

WSARE 3.0

Today’s

All historical data



And how big a deal is this, considering how

Biosurmveiullacnche DseeteactriocnhAlgIo’rvitehmds:oSlnidee1?09

**Expensive**

Copyright © 2002, 2003, Andrew Moore

What’s strange about today, considering its environment?

**Cheap**

What *should* be happening today?

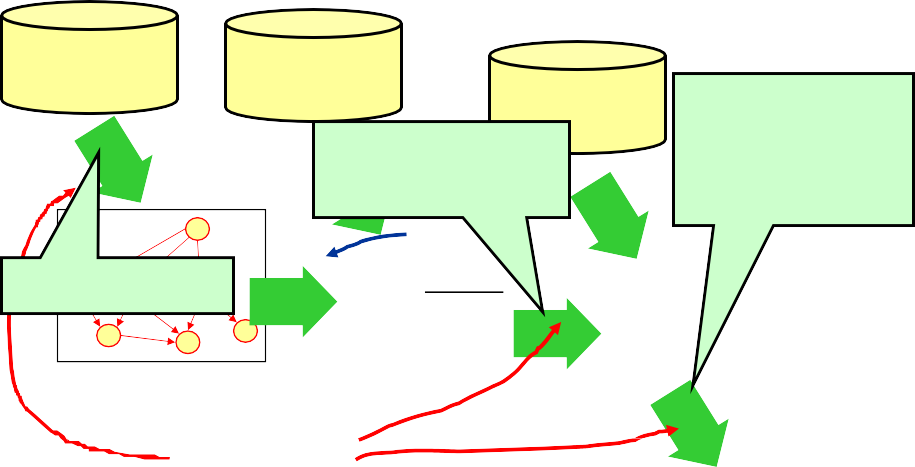
Today’s Cases

Environment

WSARE 3.0

Today’s

All historical data



And how big a deal is this, considering how

Biosurmveiullacnche DseeteactriocnhAlgIo’rvitehmds:oSlnidee1?10

**Expensive**

Copyright © 2002, 2003, Andrew Moore

What’s strange about today, considering its environment?

* RADSEARCH
* Racing Randomization
* Differential Randomization

Today’s Cases

Environment

* All-dimensions Trees

**Cheap**

What *should* be happening today?

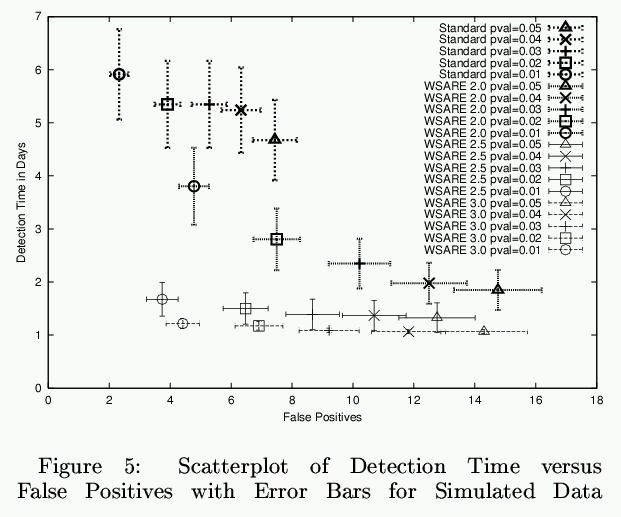
WSARE 3.0

Today’s

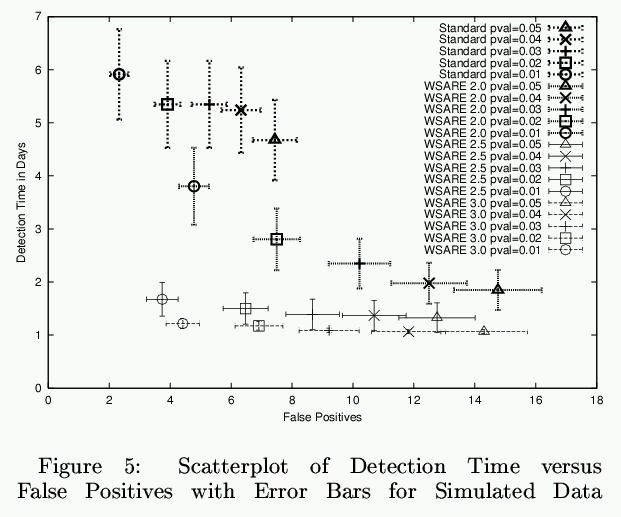
All historical data

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 111

Copyright © 2002, 2003, Andrew Moore Biosurveillance Detection Algorithms: Slide 112



Results on Simulation



**WSARE3.0**

Results on Simulation

**WSARE2.5**

**WSARE2.0**

**Standard**

Biosurveillance Detection Algorithms: Slide 113

Copyright © 2002, 2003, Andrew Moore

Conclusion

* One approach to biosurveillance: one algorithm monitoring millions of signals derived from multivariate data

instead of

Hundreds of univariate detectors

* Modeling historical data with Bayesian Networks to allow conditioning on unique features of today
* Computationally intense unless we’re tricksy!

Biosurveillance Detection Algorithms: Slide 114

...and keep relearning them as data arrives online...

Copyright © 2002, 2003, Andrew Moore

* m..u.wlteivaalrsioatneeeddattoalearn Bayes Nets from databases with

millions of records...

* instead of
* H...uinndthreedesndowf euntyivpaicrailalytesedaercthecatboorust a billion alternative Bayes net structures for modeling 800,000 records in 10
* Mmoidnuetliensg historical data with Bayesian Networks to allow conditioning on unique features of today
* Computationally intense unless we’re tricksy!

mdounrinitgorrianngdommiilzliaotinosn of signals derived from

•• O.n..oenalypwperohaacvhe ttoo dboioitsu10rv,0e0i0llatnimcees: oonntehealrgeoplriictahsm

large database...

* Searching over thoCusoanndcs louf csoinotinngency tables on a

yright © 2002, 20

03, Andrew Moore

veillance Detection Algorithms: Sli



Bayesian Network Spatial Scan

[Neill, Moore, Schneider, Cooper

Biosur Wagner, Wong] de 116

Cop

Historical Model Scan Statistic

[Hogan, Moore, Neill, Tsui, Wagner]

Fast Scan for Oriented Regions

[Neill, Moore et al.]

Fast Scan Statistic

[Neill, Moore]

BARD: Airborne Attack Detection

[Hogan, Cooper]

***General Detectors***

WSARE meets Scan Statistics

***Specific Detectors***

PANDA2: Patient-based Bayesian Network

[Cooper, Levander et. al]

Other New Algorithmic Developments

Biosurveillance Detection Algorithms: Slide 115

Copyright © 2002, 2003, Andrew Moore

Conclusion

* One approach to biosurveillance: one algorithm monitoring millions of signals derived from multivariate data

instead of

Hundreds of univariate detectors

* Modeling historical data with Bayesian Networks to allow conditioning on unique features of today
* Computationally intense unless we’re tricksy!
* WSARE 2.0 Deployed during the past year
* WSARE 3.0 about to go online
* WSARE now being extended to additionally exploit over the counter medicine sales

yright © 2002, 20

03, Andrew Moore

veillance Detection Algorithms: Sli



de 117

[Neill, Moore, Schneider, Cooper

Biosur Wagner, Wong]

Cop

Bayesian Network

informationSpatial Scan

[wrh@cbmi.pitt.edu](mailto:wrh@cbmi.pitt.edu) for

Historical Model Scan Statistic

[Hogan, Moore, Neill,

Please conTtsaui,cWt aBgniellr] Hogan

Please conFtaasct tSGcarneSgtaCtisoticoper gfc@cbmi.[uNepilml, Mcoo.ree]du for

information

Fast Scan for Oriented Regions

[Neill, Moore et al.]

BARD: Airborne Attack Detection

[Hogan, Cooper]

***General Detectors***

WSARE meets Scan Statistics

***Specific Detectors***

PANDA2: Patient-based Bayesian Network

[Cooper, Levander et. al]

Other New Algorithmic Developments

Biosurveillance Detection Algorithms: Slide 118

Copyright © 2002, 2003, Andrew Moore

Information Gain, Decision Trees

Probabilistic Reasoning, Bayes Classifiers, Density Estimation

Probability Densities in Data Mining Gaussians in Data Mining Maximum Likelihood Estimation Gaussian Bayes Classifiers Regression, Neural Nets Overfitting: detection and avoidance

The many approaches to cross-validation Locally Weighted Learning

Bayes Net, Bayes Net Structure Learning, Anomaly Detection

Andrew's Top 8 Favorite Regression Algorithms (Regression Trees, Cascade Correlation, Group Method Data Handling (GMDH), Multivariate Adaptive Regression Splines (MARS), Multilinear Interpolation, Radial Basis Functions, Robust Regression, Cascade Correlation + Projection Pursuit

Clustering, Mixture Models, Model Selection K-means clustering and hierarchical clustering

Vapnik-Chervonenkis (VC) Dimensionality and Structural Risk Minimization

PAC Learning

Support Vector Machines

Time Series Analysis with Hidden Markov Models

* Papers on these and other anti-terror applications: [www.cs.cmu.edu/~awm/antiterror](http://www.cs.cmu.edu/%7Eawm/antiterror)
* Papers on scaling up many of these analysis methods: [www.cs.cmu.edu/~awm/papers.html](http://www.cs.cmu.edu/%7Eawm/papers.html)
* Software implementing the above: [www.autonlab.org](http://www.autonlab.org/)
* Copies of 18 lectures on 25 statistical data mining topics: [www.cs.cmu.edu/~awm/781](http://www.cs.cmu.edu/%7Eawm/781)
* CD-ROM, powerpoint-synchronized video/audio recordings of the above lectures: [awm@cs.cmu.edu](mailto:awm@cs.cmu.edu)

For further info

Biosurveillance Detection Algorithms: Slide 119

Copyright © 2002, 2003, Andrew Moore

These and other Biosurveillance algorithms papers and free software available from

<http://www.autonlab.org/>

See also: <http://www.health.pitt.edu/rods>

1. WSARE 3.0 : Bayesian Network based Anomaly Pattern Detection

Wong, Moore, Cooper and Wagner [ICML/KDD 2003]

1. Fast Grid Based Computation of Spatial Scan Statistics Neill and Moore [NIPS 2003]

References