2008-05

Using the Repeated Two-Sample Rank Procedure for Detecting Anomalies in Space and Time

Fricker, Ronald D.

NAVAL
POSTGRADUATE
SCHOOL

Using the Repeated Two-Sample Rank Procedure for Detecting Anomalies in Space and Time

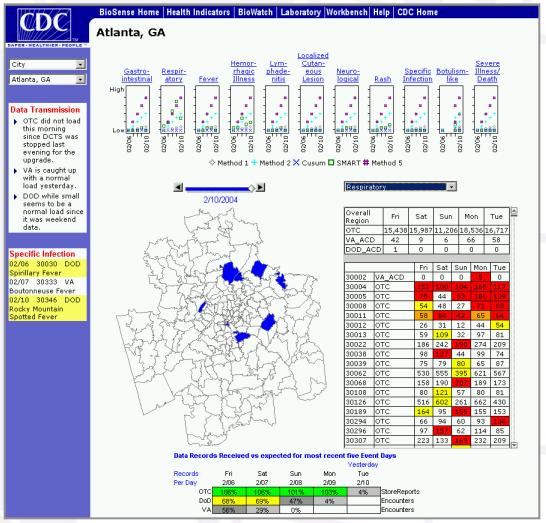
Ronald D. Fricker, Jr.

Interfaces Conference

May 31, 2008

Motivating Problem: Biosurveillance

- "…surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response." [1]
- On-going discussion in public health community about use of biosurveillance for "early event detection" vs. "situational awareness"



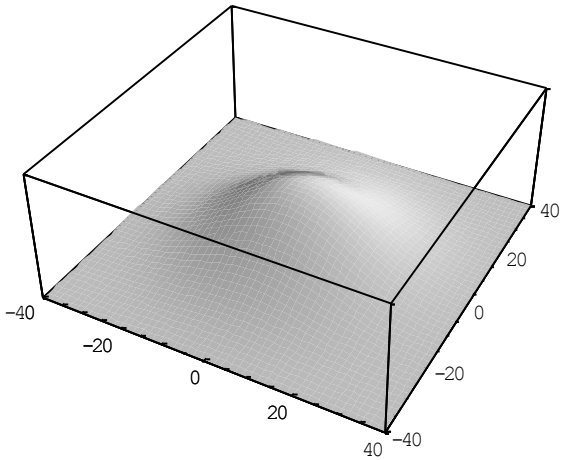
[1] CDC (www.cdc.gov/epo/dphsi/syndromic.htm, accessed 5/29/07)

Definitions

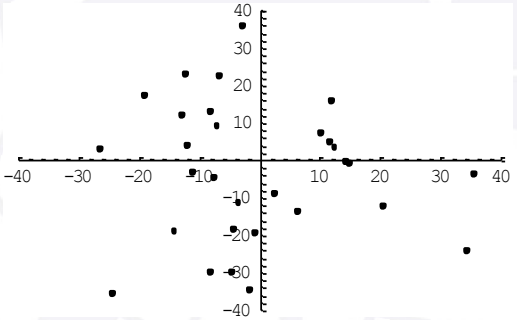
- *Early event detection*: gathering and analyzing data in advance of diagnostic case confirmation to give early warning of a possible outbreak
- *Situational awareness*: the real-time analysis and display of health data to monitor the location, magnitude, and spread of an outbreak

Illustrative Example

- ER patients come from surrounding area
  - On average, 30 per day
    - More likely from closer distances
  - Outbreak occurs at (20,20)
    - Number of patients increase linearly by day after outbreak



(Unobservable) distribution of ER patients' home addresses



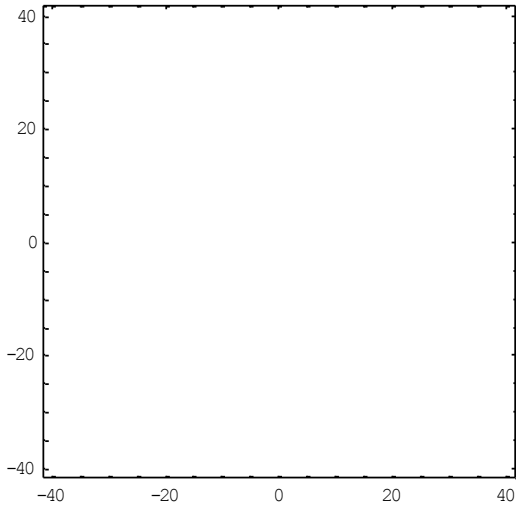
Observed distribution of ER patients' home addresses

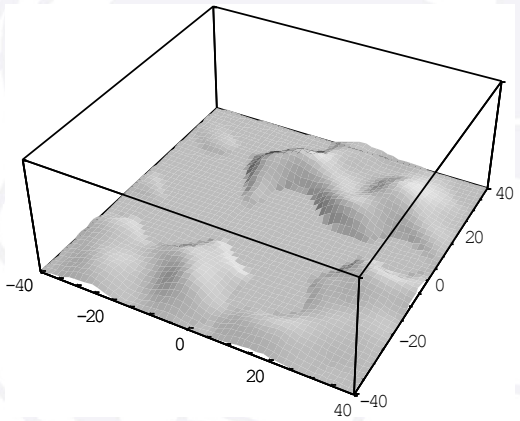
A Couple of Major Assumptions

- Can geographically locate individuals *in a medically meaningful way*
  - Non-trivial problem
  - Data not currently available
- Data is reported in a consistent and timely way
  - Public health community working this problem, but not solved yet
  - Assuming the above problems away…

Idea: Look at Differences in Kernel Density Estimates

- Construct kernel density estimate (KDE) of "normal" disease incidence using N historical observations
- Compare to KDE of most recent w observations





But how to know when to signal?

Solution: Repeated Two-Sample Rank (RTR) Procedure

- Sequential hypothesis test of estimated density heights
- Compare estimated density heights of recent data against heights of set of historical data
  - Single density estimated via KDE on *combined* data
- If no change, heights uniformly distributed
  - Use nonparametric test to assess

Data & Notation (1)

- Let $\mathbf{X}_i = \{X_{1i}, X_{2i}\}$ be a sequence of bivariate observations
  - E.g., latitude and longitude of a case
- Assume $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{\tau-1}$ ~ iid according to f_0
  - I.e., natural state of disease incidence
- At time τ, $\mathbf{X}_\tau, \mathbf{X}_{\tau+1}, ...$ ~ iid according to f_1
  - Corresponds to an increase in disease incidence
- Densities f_0 and f_1 unknown

Data & Notation (2)

- Assume a historical sequence $\mathbf{Y}_1, ..., \mathbf{Y}_N$ is available
  - Distributed iid according to f_0
- Followed by $\mathbf{X}_1, \mathbf{X}_2, ...$ which may change from f_0 to f_1 at any time
- For notational convenience, define $\mathbf{X}_i = \mathbf{Y}_{N+i}$ for $i \leq 0$

Estimating the Density

- Consider the $w+1$ most recent data points
- At each time period estimate the density

$$\hat{f}_n(\mathbf{x}) = \begin{cases} \dfrac{1}{N+n} \displaystyle\sum_{i=1-N}^{n} k_h(\mathbf{x}, \mathbf{X}_i), & n < w+1 \\ \dfrac{1}{N+w+1} \displaystyle\sum_{i=n-w-N-1}^{n} k_h(\mathbf{x}, \mathbf{X}_i), & n \geq w+1 \end{cases}$$

where k is a kernel function on $\mathbb{R}^2$ with bandwidth set to $h_i = \sigma_i \left(1/\left(N+w+1\right)\right)^{1/6}$

Calculating Density Heights

- The density estimate is evaluated at each historical and new point
  - For $n < w+1$

$$\underbrace{\hat{f}_n(\mathbf{X}_{1-N}), \ldots, \hat{f}_n(\mathbf{X}_0)}_{\text{historical observations}}, \underbrace{\hat{f}_n(\mathbf{X}_1), \ldots, \hat{f}_n(\mathbf{X}_n)}_{\text{new observations}}$$

  - For $n \geq w+1$

$$\underbrace{\hat{f}_n(\mathbf{X}_{n-w-N-1}), \ldots, \hat{f}_n(\mathbf{X}_{n-w-1})}_{\text{historical observations}}, \underbrace{\hat{f}_n(\mathbf{X}_{n-w}), \ldots, \hat{f}_n(\mathbf{X}_n)}_{\text{new observations}}$$

Under the Null, Estimated Density Heights are Exchangeable

- *Theorem*: The RTR procedure is asymptotically distribution free
  - I.e., the estimated density heights are exchangeable, so all rankings are equally likely
  - Proof: See Fricker and Chang (2008)
- Means can do a hypothesis test on the ranks each time an observation arrives
  - Signal change in distribution first time test rejects

Comparing Distributions of Heights

- Compute empirical distributions of the two sets of estimated heights:

$$\hat{J}_n(z) = \frac{1}{w+1} \sum_{i=n-w}^{n} I\left\{\hat{f}_n(\mathbf{X}_i) \le z\right\},$$

$$\hat{H}_n(z) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} I\left\{\hat{f}_n(\mathbf{X}_i) \le z\right\}$$

- Use Kolmogorov-Smirnov test to assess:

$$S_n = \max_z \left|\hat{J}_n(z) - \hat{H}_n(z)\right|$$

  - Signal at time $t = \min\left\{n : S_n > c\right\}$

Comparison Metrics

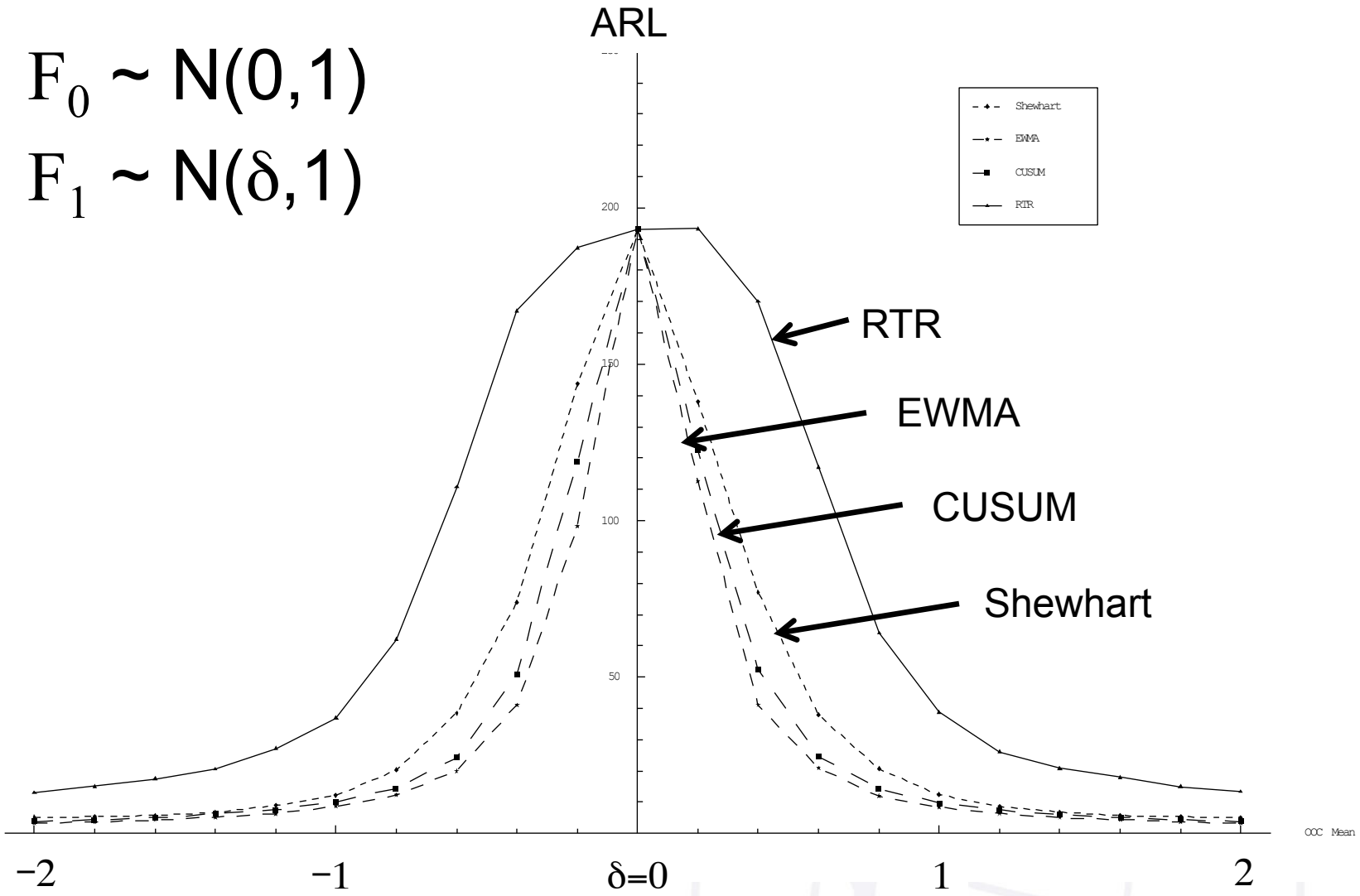
- How to find c?
  - Use ARL approximation based on Poisson clumping heuristic:

$$A \approx \left[\left(\frac{6.16c\,[c + 0.5/(w+1)]}{1 + (w+1)/N} \right) \exp\left\{ -2\left(c + \frac{1}{2(w+1)} \right)^2 \left(\frac{1}{w+1} + \frac{1}{N} \right)^{-1} \right\} \right]^{-1}$$

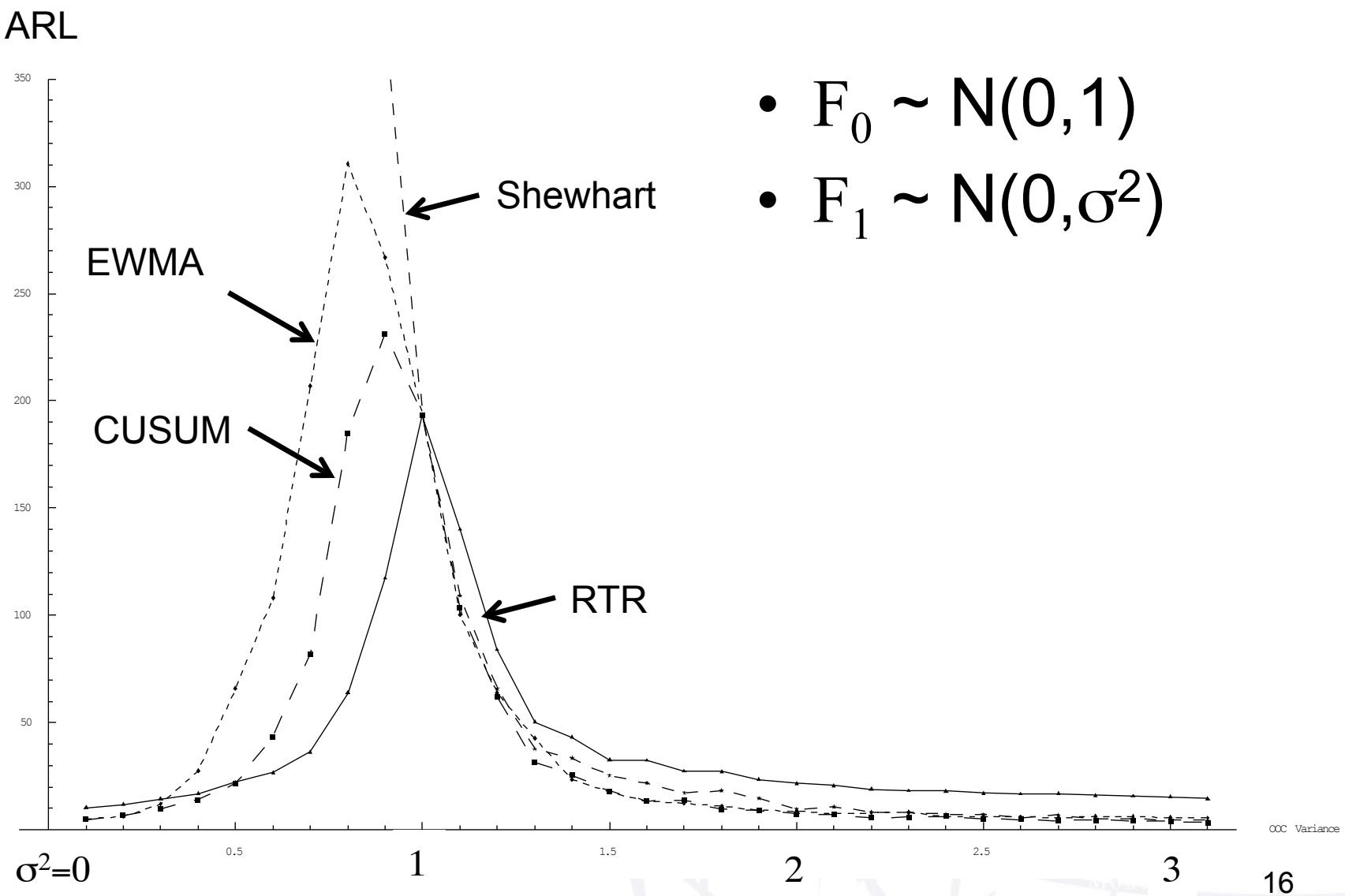
- Example: c=0.07754 with N=1,350 and $w+1$=250 gives A=900
  - If 30 observations per day, gives average time between (false) signals of 30 days

Performance Comparison #1

- $F_0 \sim N(0,1)$
- $F_1 \sim N(\delta,1)$



Performance Comparison #2



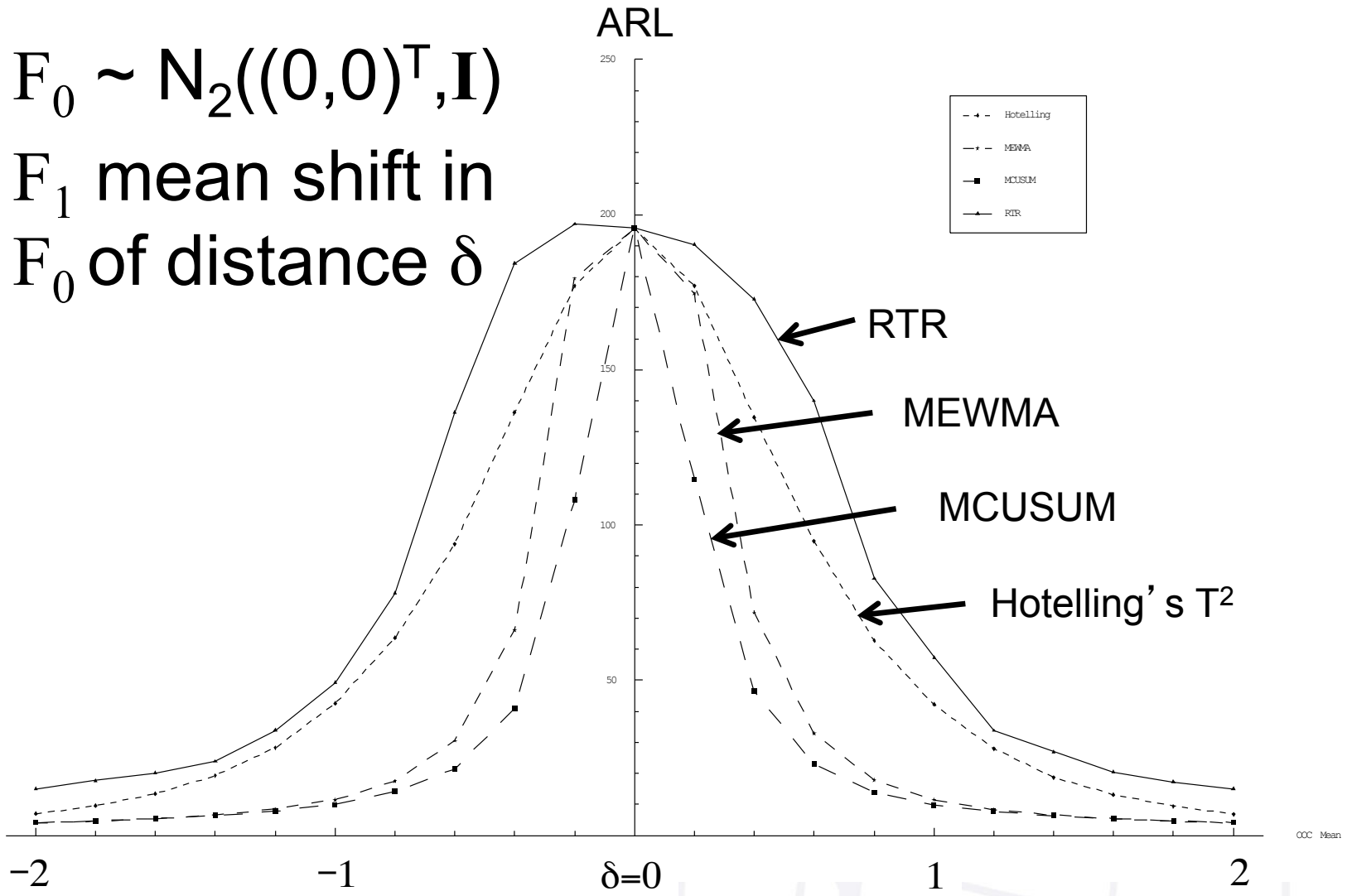
- $F_0 \sim N(0,1)$
- $F_1 \sim N(0,\sigma^2)$

Performance Comparison #3

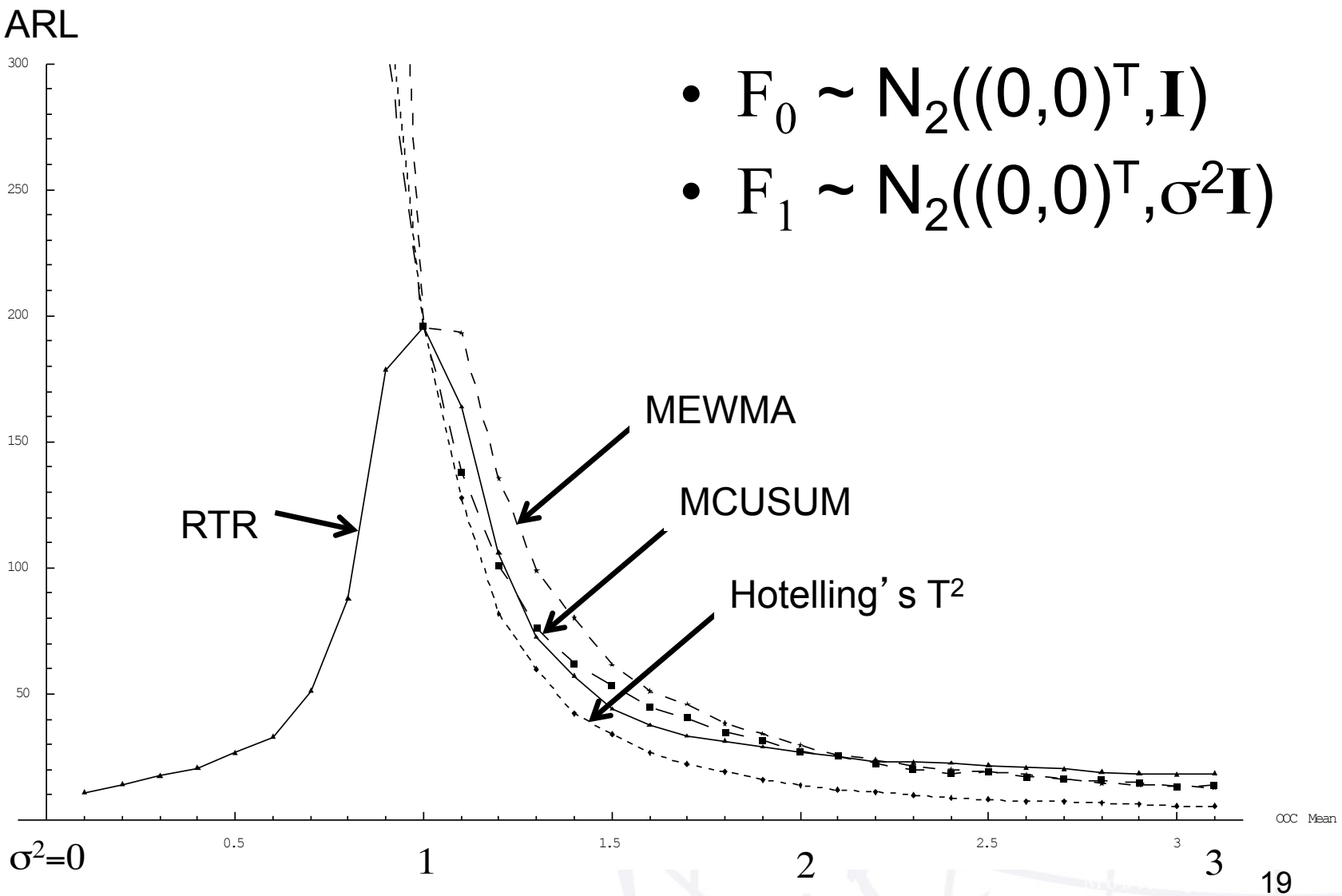
- $F_0 \sim N(0,1)$
- $F_1 \sim$ 

Performance Comparison #4

- $F_0 \sim N_2((0,0)^T, \mathbf{I})$
- F_1 mean shift in F_0 of distance δ



Performance Comparison #5



- $F_0 \sim N_2((0,0)^T, \mathbf{I})$
- $F_1 \sim N_2((0,0)^T, \sigma^2\mathbf{I})$

Plotting the Outbreak

- At signal, calculate optimal kernel density estimates and plot pointwise differences

$$\Delta_n(\mathbf{x}) = \max\left(\delta, \hat{h}_n(\mathbf{x}) - \hat{g}_n(\mathbf{x})\right)$$

where

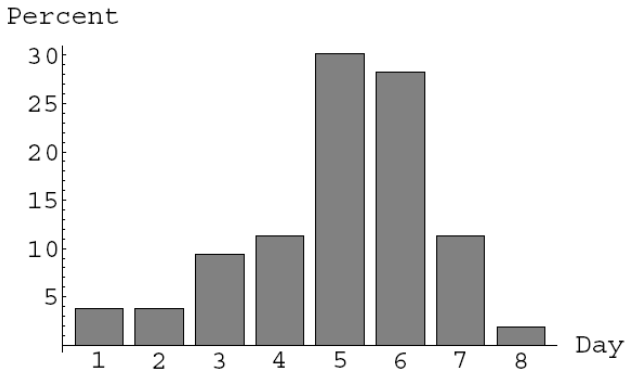
$$\hat{h}_n(\mathbf{x}) = \frac{1}{w+1}\sum_{i=n-w}^{n} k_h\left(\mathbf{x}, \mathbf{X}_i\right)$$

$$\hat{g}_n(\mathbf{x}) = \frac{1}{N}\sum_{i=n-w-N-1}^{n-w-1} k_h\left(\mathbf{x}, \mathbf{X}_i\right)$$

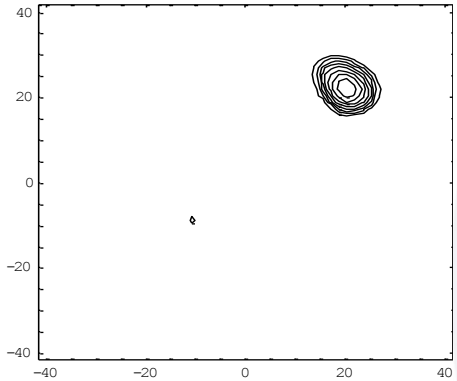
and $h_i = \sigma_i\left(\frac{1}{w+1}\right)^{1/6}$ or $h_i = \sigma_i\left(\frac{1}{N}\right)^{1/6}$

Example Results

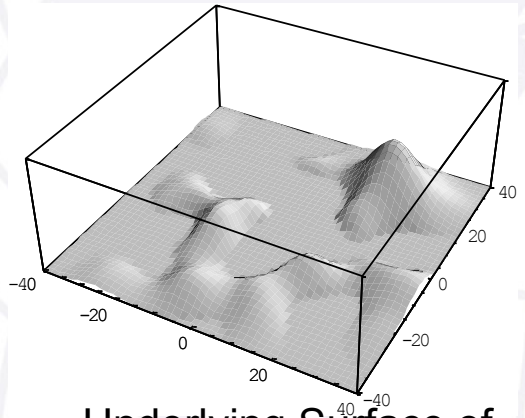
- Assess performance by simulating outbreak multiple times, record when RTR signals
  - Signaled middle of day 5 on average
  - By end of 5th day, 15 outbreak and 150 non-outbreak observations
  - From previous example:



Distribution of Signal Day

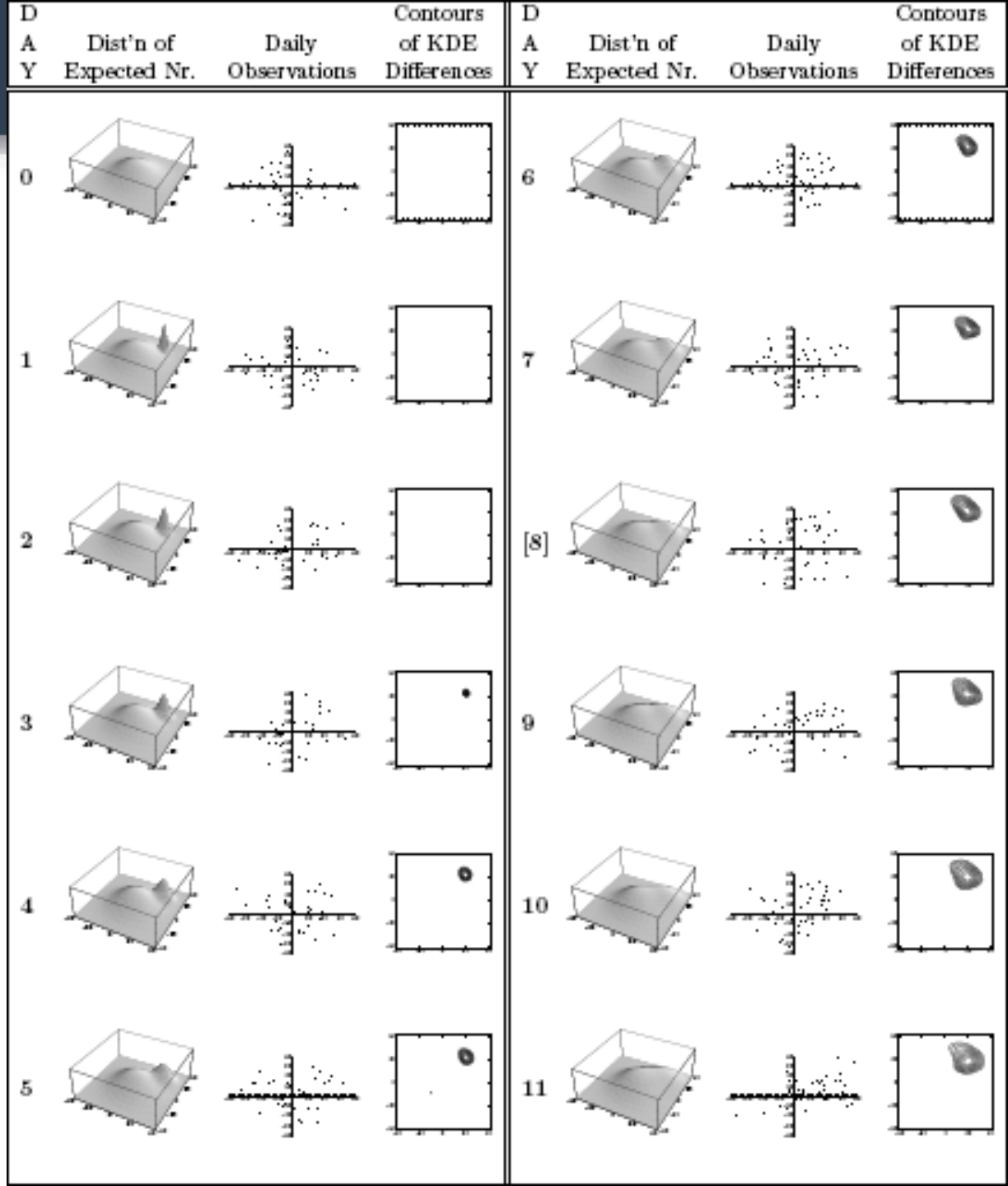


Estimate of Outbreak Location on Day 5



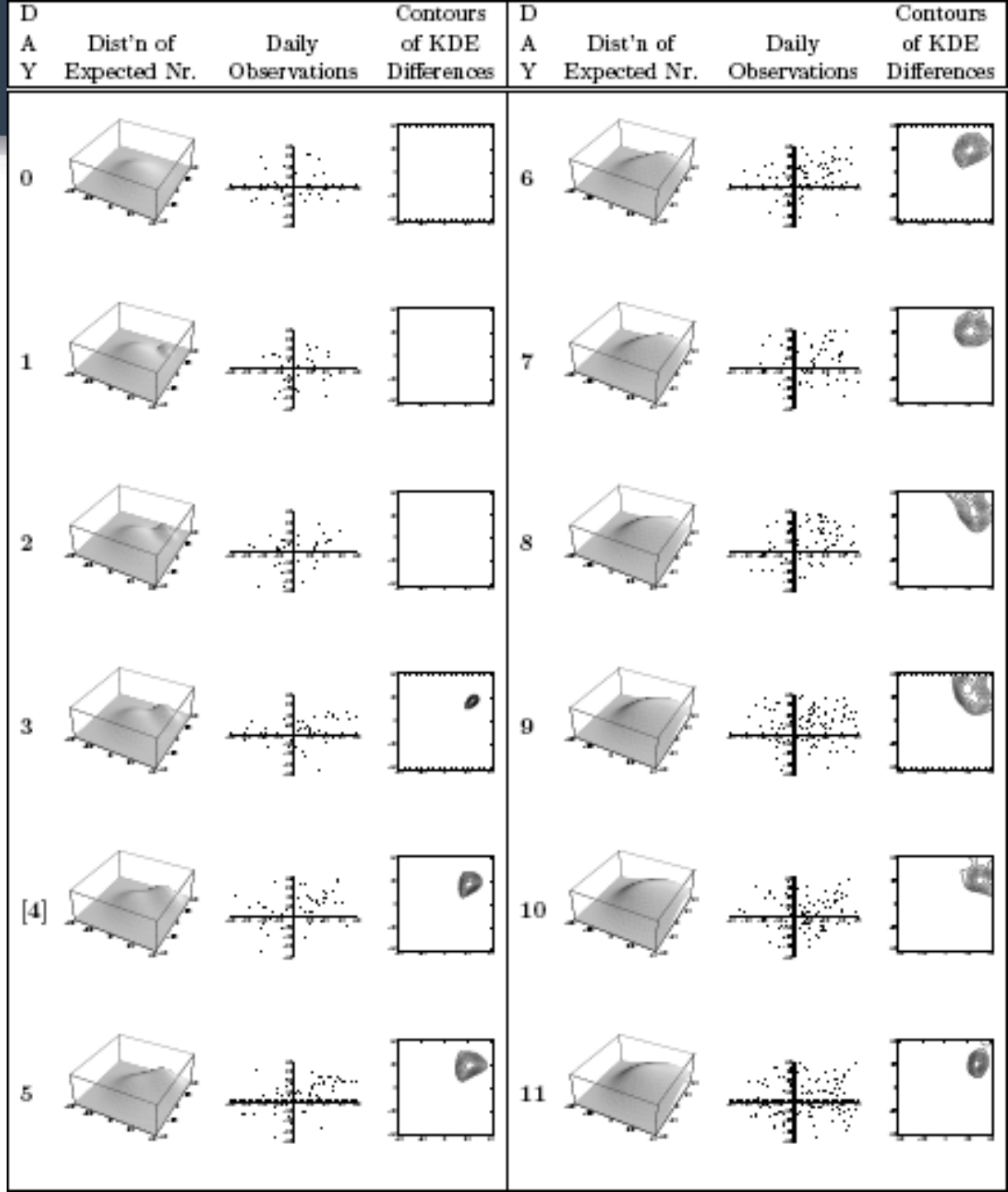
Underlying Surface of Density Height Differences

- Normal disease incidence ~ N ($\{0,0\}^t, \sigma^2\mathbf{I}$) with σ=15
  - Expected count of 30 per day
- Outbreak incidence ~ N ($\{20,20\}^t, d^2\mathbf{I}$), where d is the day of outbreak
  - Expected count is 30+d per day



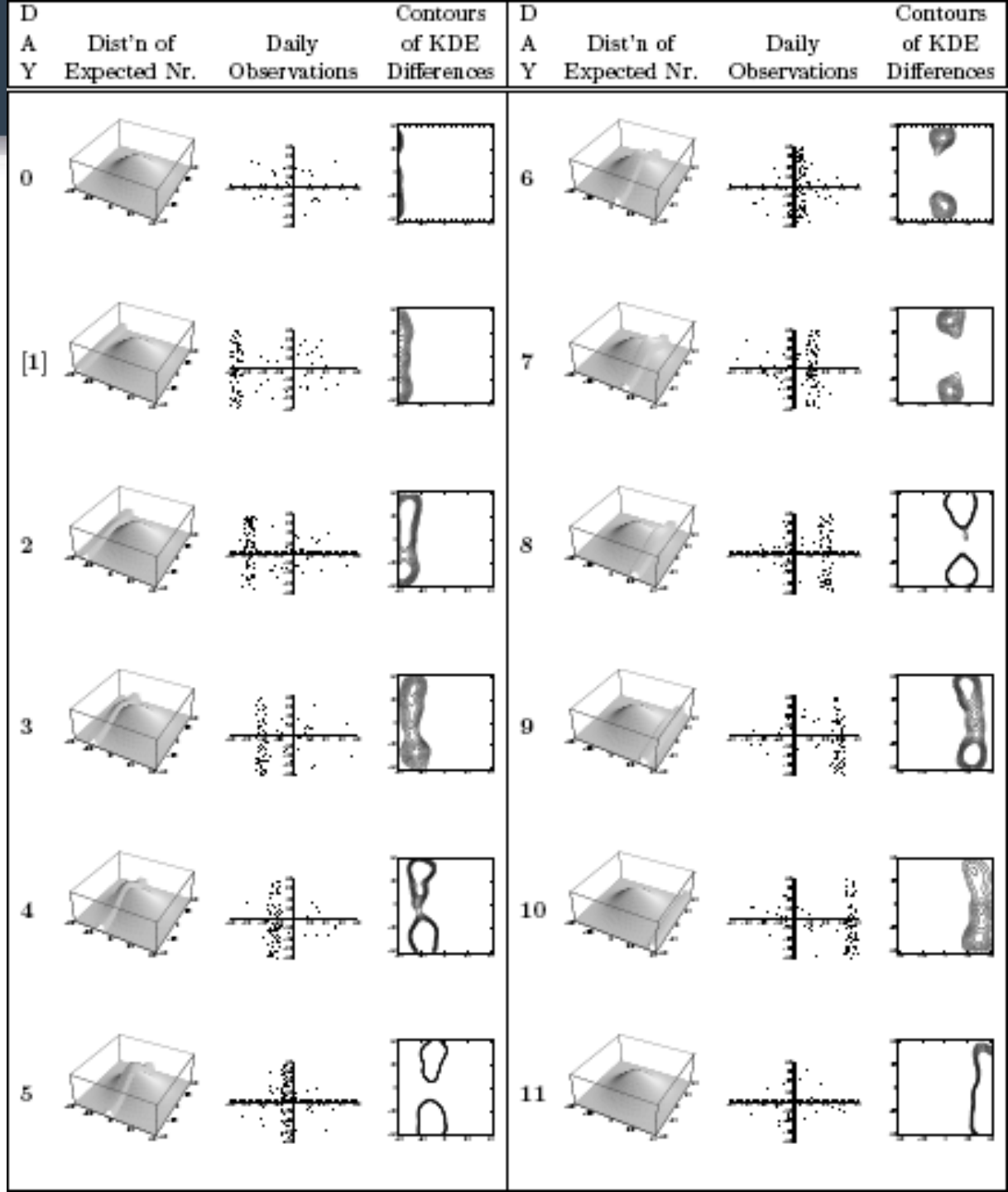
NAVAL POSTGRADUATE SCHOOL

- Normal disease incidence ~ N ($\{0,0\}^t, \sigma^2\mathbf{I}$) with σ=15
  - Expected count of 30 per day
- Outbreak incidence ~ N ($\{20,20\}^t, 2.2d^2\mathbf{I}$), where d is the day of outbreak
  - Expected count is 30+d^2 per day



NAVAL POSTGRADUATE SCHOOL

- Normal disease incidence ~ N ($\{0,0\}^t, \sigma^2\mathbf{I}$) with σ=15
  - Expected count of 30 per day
- Outbreak incidence sweeps across region from left to right
  - Expected count is 30+64 per day



Advantages and Disadvantages

- Advantages
  - Methodology supports both biosurveillance goals: early event detection and situational awareness
  - Incorporates observations sequentially (singly)
    - Most other methods use aggregated data
  - Can be used for more than two dimensions
- Disadvantage?
  - Can't distinguish increase distributed according to f_0
    - Unlikely for bioterrorism attack?
    - Won't detect an general increase in background disease incidence rate
      - E.g., Perhaps caused by an increase in population
      - In this case, advantage not to detect

Selected References

Selected Research:

- Fricker, R.D., Jr., and J.T. Chang, The Repeated Two-Sample Rank Procedure: A Multivariate Nonparametric Individuals Control Chart (in draft).
- Fricker, R.D., Jr., and J.T. Chang, A Spatio-temporal Method for Real-time Biosurveillance, *Quality Engineering* (to appear).
- Fricker, R.D., Jr., and D. Banschbach, Optimizing a System of Threshold Detection Sensors, in submission to *Operations Research*.
- Fricker, R.D., Jr., Knitt, M.C., and C.X. Hu, Comparing Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance, *Quality Engineering* (to appear).
- Joner, M.D., Jr., Woodall, W.H., Reynolds, M.R., Jr., and R.D. Fricker, Jr., A One-Sided MEWMA Chart for Health Surveillance, *Quality and Reliability Engineering International* (to appear).
- Fricker, R.D., Jr., Hegler, B.L., and D.A Dunfee, Assessing the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms, *Statistics in Medicine,* 2008.
- Fricker, R.D., Jr., Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance, *Advances in Disease Surveillance,* **3**:1, 2007.

Background Information:

- Fricker, R.D., Jr., and H. Rolka, Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, **91**, pp. 4-13, 2006
- Fricker, R.D., Jr., Syndromic Surveillance, *Encyclopedia of Quantitative Risk Assessment* (to appear).