



Near-Real Time Surveillance against Bioterror Attack using Space-Time Clustering

March 16, 2001

Mark W. Koch

Signal and Image Processing Systems Department

Sean A. McKenna

Geohydrology Department





Problem, Approach, and Benefits

- **Problem**

- Use on-line electronic medical information to detect and characterize a biological weapons attack after only a *relatively small number* of infected victims begin to present symptoms.

- **Approach**

- Space-time clustering
 - » Sift through the massive amounts of medical records
 - » Focus attention on potentially suspicious times and places
 - » Indicate exposure to infectious disease or localized exposure to toxins

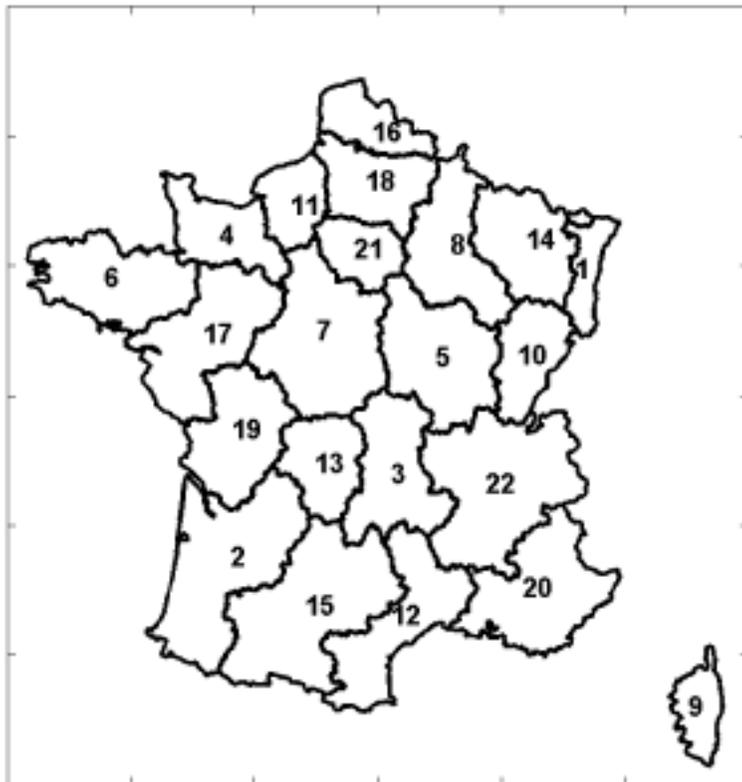
- **Associated Benefits**

- Early detection of emerging infectious diseases
- Accumulate a vast warehouse of health data for use in epidemiology.

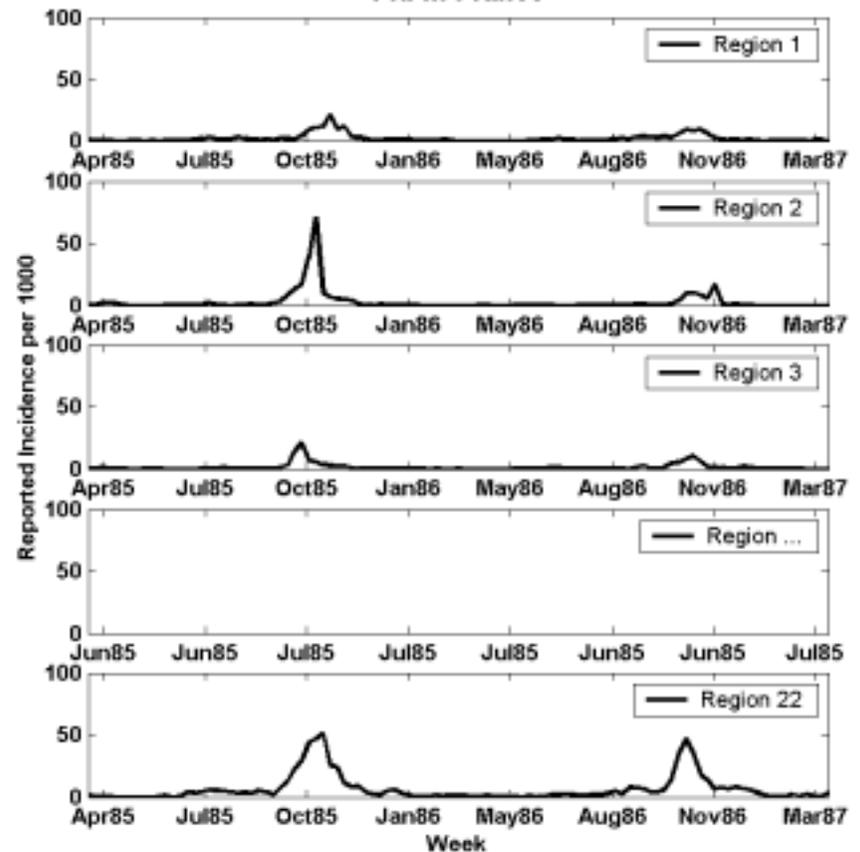
French Flu Data

- Weekly time series data for 22 regions in France, since 1984 to present
- Data from Sentinel Physicians on
 - Flu, Measles, Mumps, Chicken Pox, ...

Regions in France



Flu in France





Rapid Syndrome eValuation Project (RSVP) Data (Zelicoff)

- **Fast and easy data entry**
- **Return of information to the medical doctor**
- **Continuously monitoring public health**

- **Entered by physicians, in hospital ER's, 1 min.**
- **Syndrome information**
 - **Influenza Like Illness (ILI)**
 - **Adult respiratory distress syndrome**
 - **Bloody diarrhea and fever**
 - **Watery diarrhea**
 - **Rash with fever**
 - **Diffuse central nervous system dysfunction with fever**
- **6 Symptoms**
 - **Varying with syndrome**
- **6-10 Signs**
- **Zip code (150 possible)**
 - **Home & work**
- **High risk occupation**
 - **military, police, fire, child care, health care, food handler**
- **Contact with previous person? Travel?**



Space-Time Data Types

- **Detective or point data**
 - Investigation of a public health concern
 - Interview people
 - » Home, work or school address
 - » Time symptoms appeared
- **Research or aggregated data**
 - Aggregation of space information (zip-code, county, state, etc.)
 - Aggregation of time information (week, month, quarter, etc.)
 - Reduces database size
 - Insures privacy
 - Can lose information



Space-Time Clustering

- **Find clusters of adverse health events (disease) in space and time.**
 - Space-time clustering
 - » Incidence of a disease is temporarily higher at one place and time than others
 - Rare adverse health events
 - Sudden increase in frequency
- **Point data statistical tests (Space-Time)**
 - Knox, Mantel, k-NN.
- **Aggregated data tests**
 - Space or time not both.
- **Combine a aggregated-time test with a space clustering test!**

Aggregated Time-Clustering in a Region

- Assume cases are uniformly distributed in time.
- Does the aggregation of cases at the current time represent a significant cluster of cases?

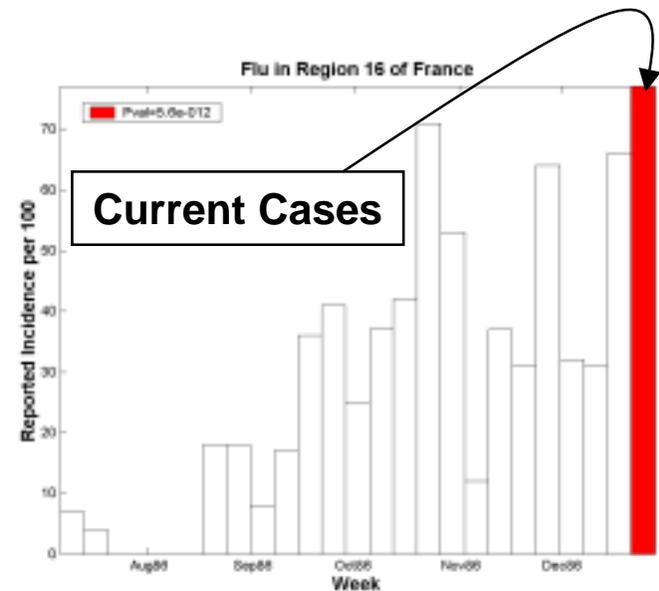
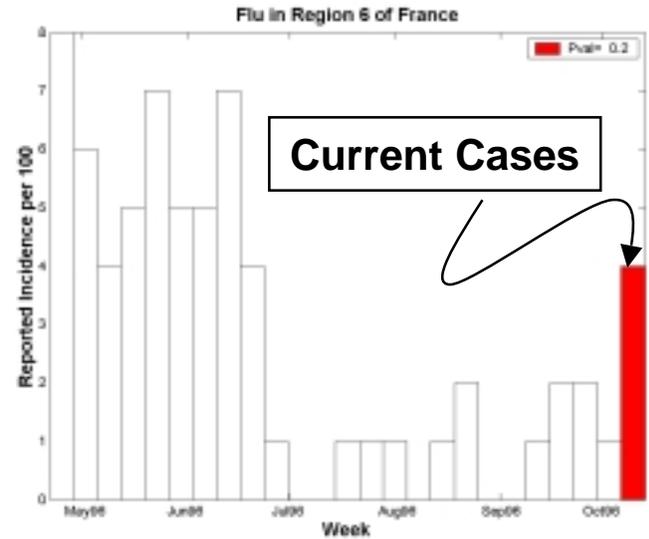
For an aggregation, y , of uniformly distributed events (U):

Y - is approximately
Poisson distributed

$$\begin{aligned}P(y \in U) &= P(\mathbf{Y} \geq y) \\ &= 1 - F_Y(y) \\ &= Pval(y)\end{aligned}$$

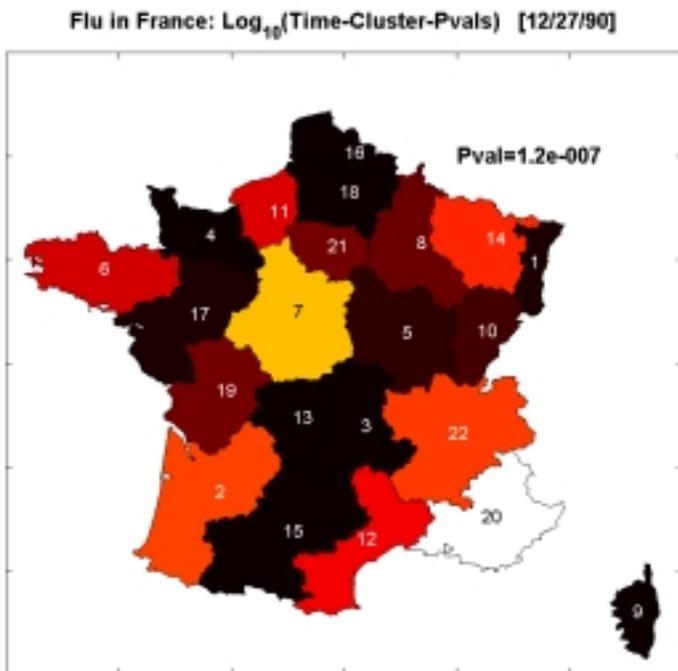
$F_Y(y)$ – is the Poisson CDF

Small $Pval$'s indicate significant
time - clusters

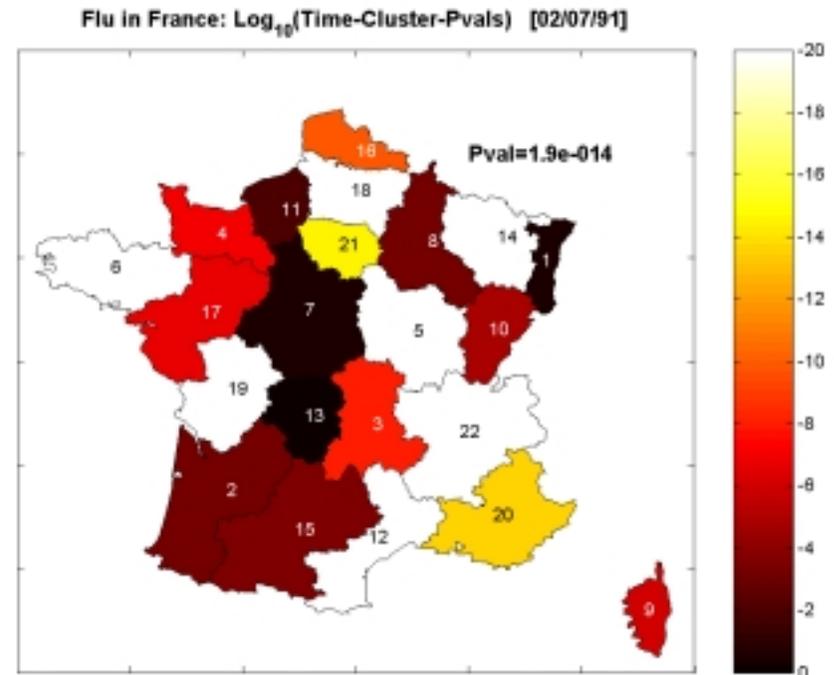


Space-Clustering

- Find a cluster of regions with large time clusters (low time-cluster-pvals)



7 Weeks before epidemic



2 Weeks before epidemic

"Min-Max" Space Clustering

- **Most methods use spatial auto-correlation**
 - These methods find clusters of non-significant events!
- **Combine Pvals in region cluster**
- **Identify region-clusters where every region has significant time-clusters.**

For a region r

$R(r)$ = "a region - cluster"
= $\{r, r\text{'s adjacent neighbors}\}$

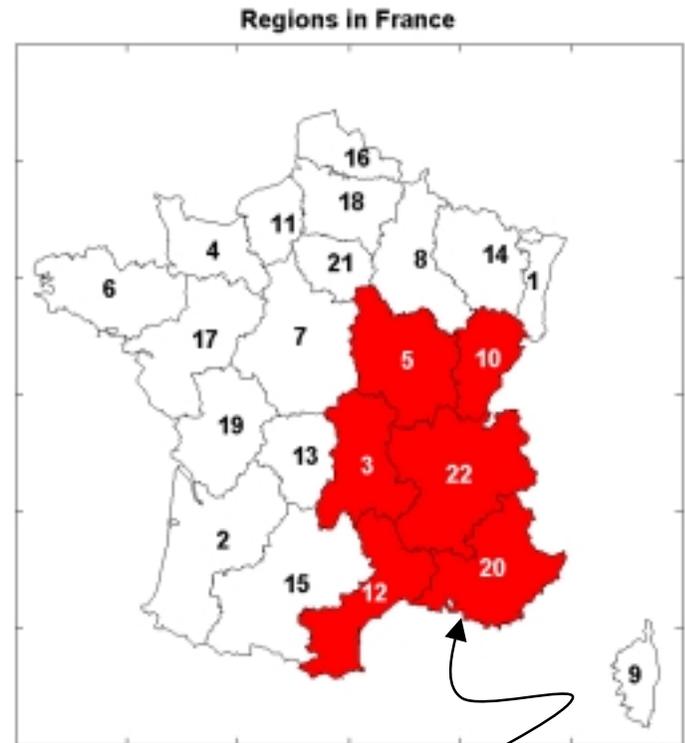
$$Pval(R(r)) = P(\bigvee_{i \in R(r)} \{y_i \in U\})$$

$$Pval(R(r)) = P(\bigvee_{i \in R(r)} \{\mathbf{Y}_i \geq y_i\})$$

$$\approx \max_{i \in R(r)} Pval(y_i)$$

For the set of regions, Φ

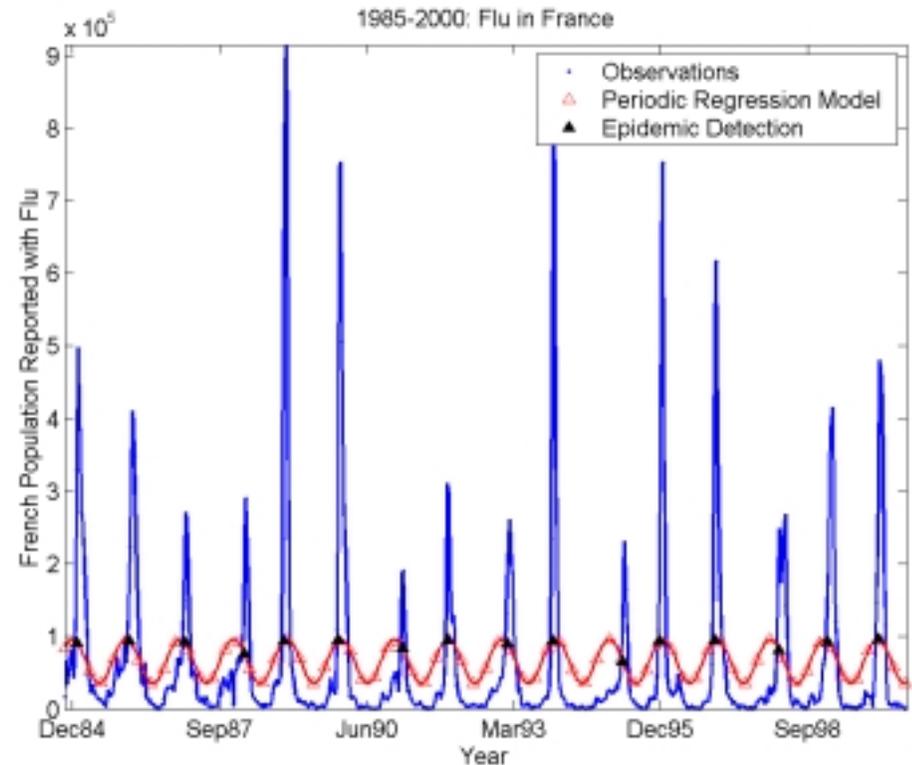
$$Pval(\Phi) = \min_{r \in \Phi} (Pval(R(r))) = \min_{r \in \Phi} (\max_{i \in R(r)} Pval(y_i))$$



Region-Cluster for region 22

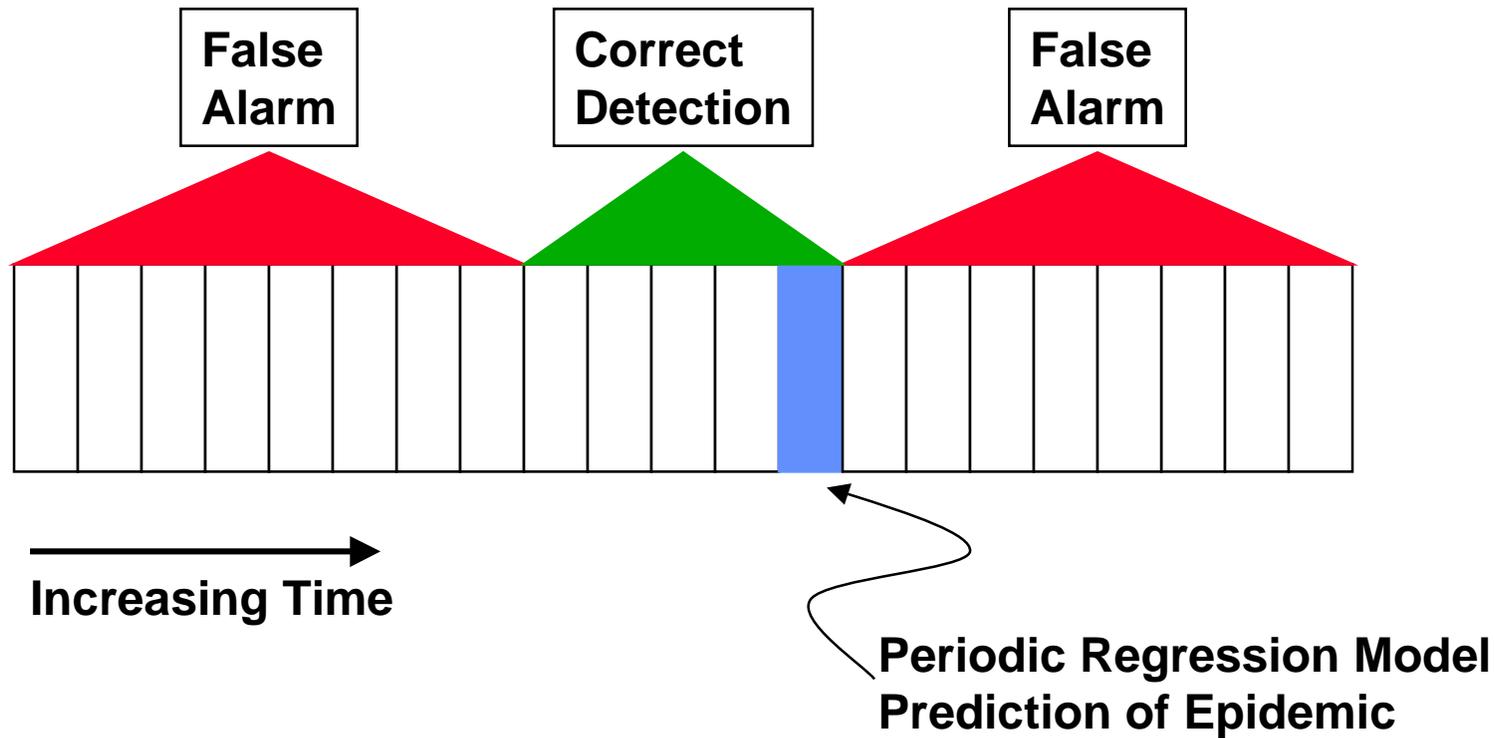
Determination of the Start of Flu Epidemics (Serfling)

- **Epidemic - *Observed value of incidence is above the 95% confidence threshold of a periodic regression model for two consecutive weeks.***
- **Doesn't incorporate a space criteria.**



Measuring Flu Epidemic Prediction Performance

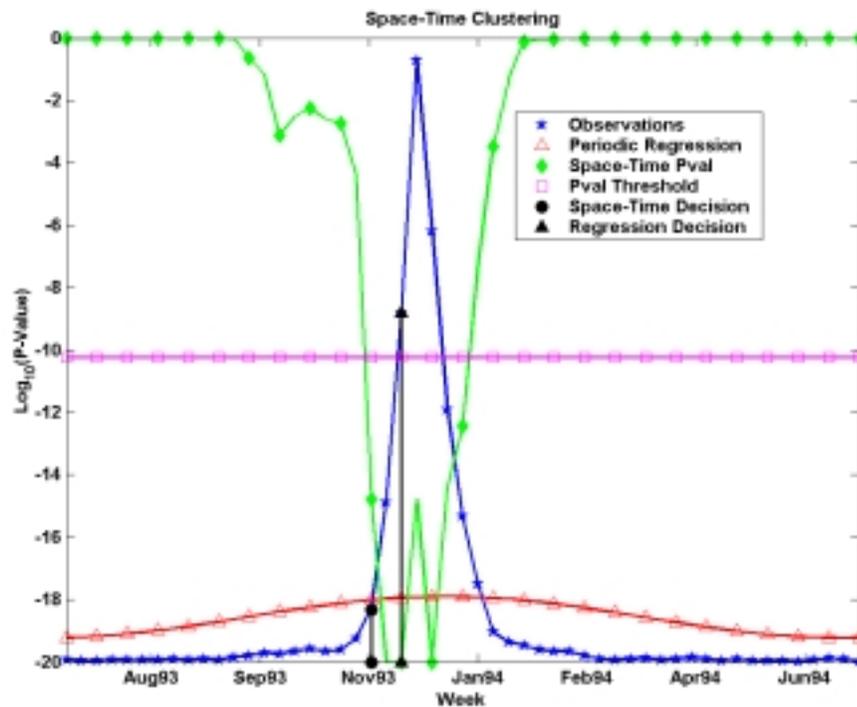
- Assume can't predict an flu epidemic more than 4 weeks earlier than method by Serfling.



Space-Time Prediction of Epidemics

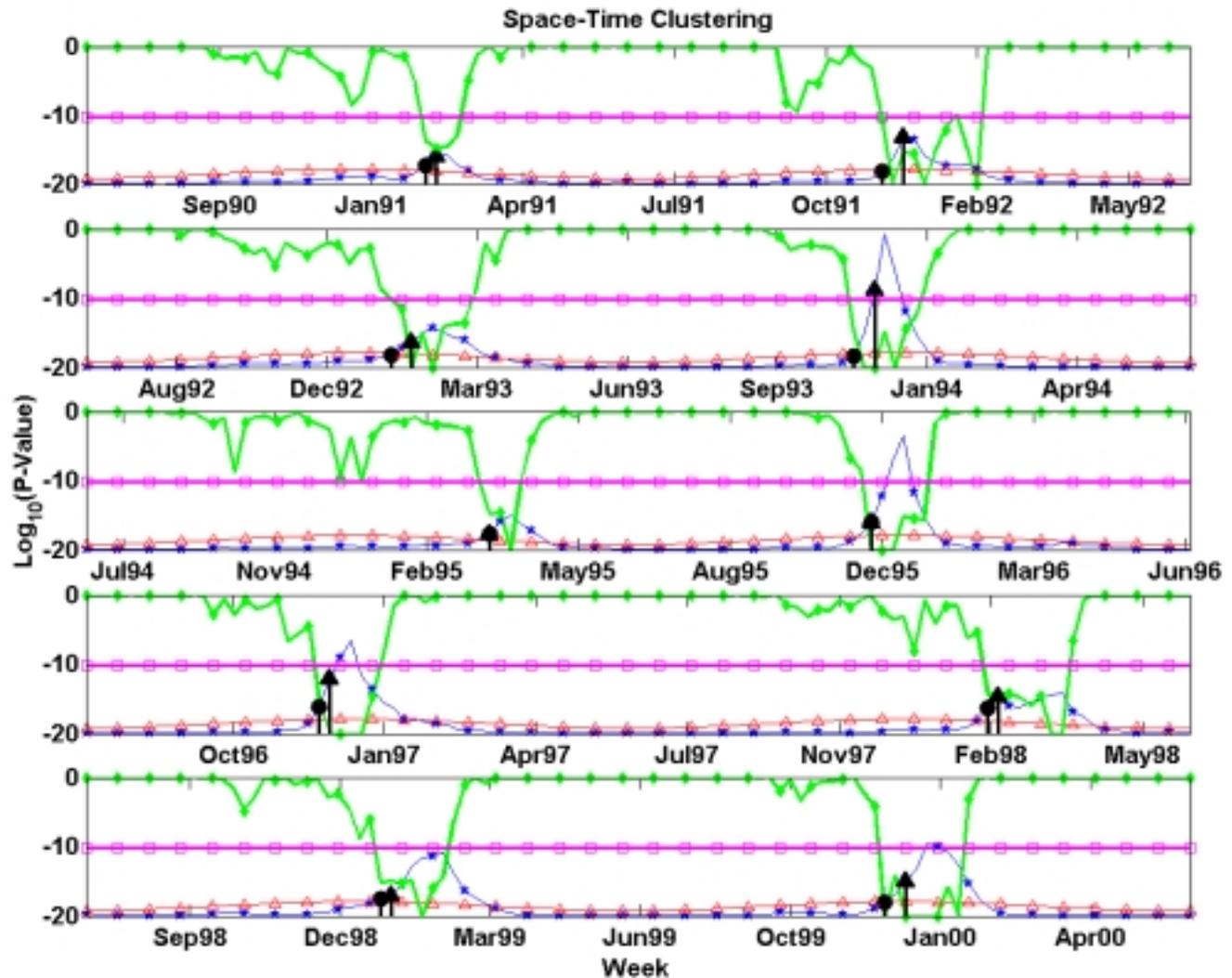
- Epidemic - *Space-Time Pval* falls below a *threshold*.
- Sensitivity
 - 13/14=93%
- Specificity
 - 223/224=99%

Weeks Before Serfling	# of Epidemics
0	3
1	4
2	4
3	0
4	2
5	1



Prediction of Flu 4 weeks before it's start.

Space-Time Prediction of Epidemics



Clustering RSVP Records

- **RSVP data contain both numerical and categorical information.**
 - **Numerical: has an ordering**
 - » Temperature, age
 - **Categorical: has no natural ordering**
 - » Abnormal lung sounds (unilateral, bilateral, wheezing, other)
- **Most clustering algorithms fail with categorical info.**
 - Require distance metrics to measure similarity and centroids to determine cluster center.
- **Link-based clustering (Gupta 2000)**
 - $\text{link}(p_i, p_j)$ - number of common neighbors between records p_i and p_j
 - neighbor - similarity(p_i, p_k) > threshold
 - similarity
 - » Expert-based
 - » Jaccard coefficient for two sets T_1 and T_2
 - $\text{similarity}(T_1, T_2) = |T_1 \cap T_2| / |T_1 \cup T_2|$



Conclusions

- **Mine electronic medical databases to detect and characterize bioterror attack**
- **Space-time clustering**
 - Focus attention on a potentially suspicious time and place
- **Extended an aggregated time clustering approach to space and time**
- **Predict a flu epidemic 3-4 weeks before it starts**
 - Sensitivity - 93%
 - Specificity - 99%