

*Generalized Linear Mixed Models
for Detecting 'Unusual' Incident
Spatial Clusters of Disease*

Ken Kleinman, 9/23/2002

Our team: Courtney Adams, Inna
Dashevsky, Al DeMaria, Pat Kludt, Ben
Kruskal, **Ross Lazarus, Richard Platt**

Outline

- Motivation
- Our setting and data
- Our approach to setting thresholds
- Important outstanding questions
- Summary

Motivation

- Based on surveillance, how do we decide when something unusual is going on? (Previous session.)
- When geographical data is available?
- My motivation: anthrax. Crucial to respond as quickly as possible: one day earlier could save lives, money, resources.
- Details apply in other cases as well

Motivation

- The goal is to decide when to break out the Cipro
- More realistically, when to suggest it's time to do some field epidemiology and/or gold-standard tests (x-rays, for anthrax).
- This decision should be made systematically, not based on subjective assessments.
- My professional bias: Statistically (not by rule-of-thumb)

Our setting and data

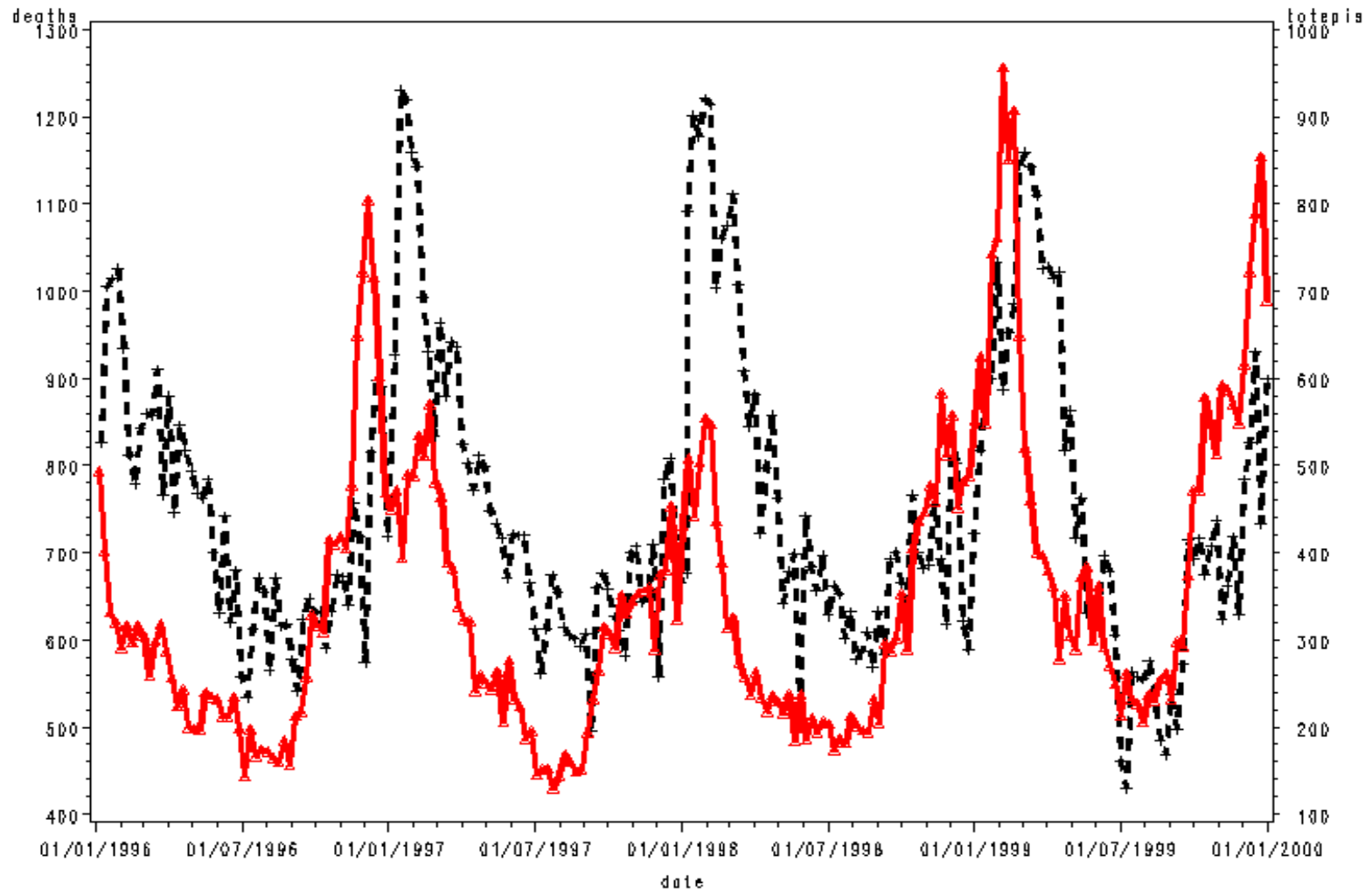
- The setting: An HMO and Provider group with ~250,000 members in eastern Mass.
- Automated ambulatory medical records:
 - **Live**, continuously updated records of each patient contact, including ICD-9 diagnosis
 - Part of standard practice: our system imposes **no** additional burden on busy care providers
 - A commercial system, implemented elsewhere; easy to implement our approach elsewhere

Syndrome information

- Define Lower Respiratory Illness (LRI) syndrome: contains 119 ICD-9 codes, including cough, pneumonia, bronchitis (adapted from ESSENCE project)
- Phase 1 anthrax would be diagnosed in LRI
- For historical data (96-99):
 - 118,557 LRI **visits**
 - 73,752 LRI **episodes** (any previous LRI complaint more than 6 weeks in past)

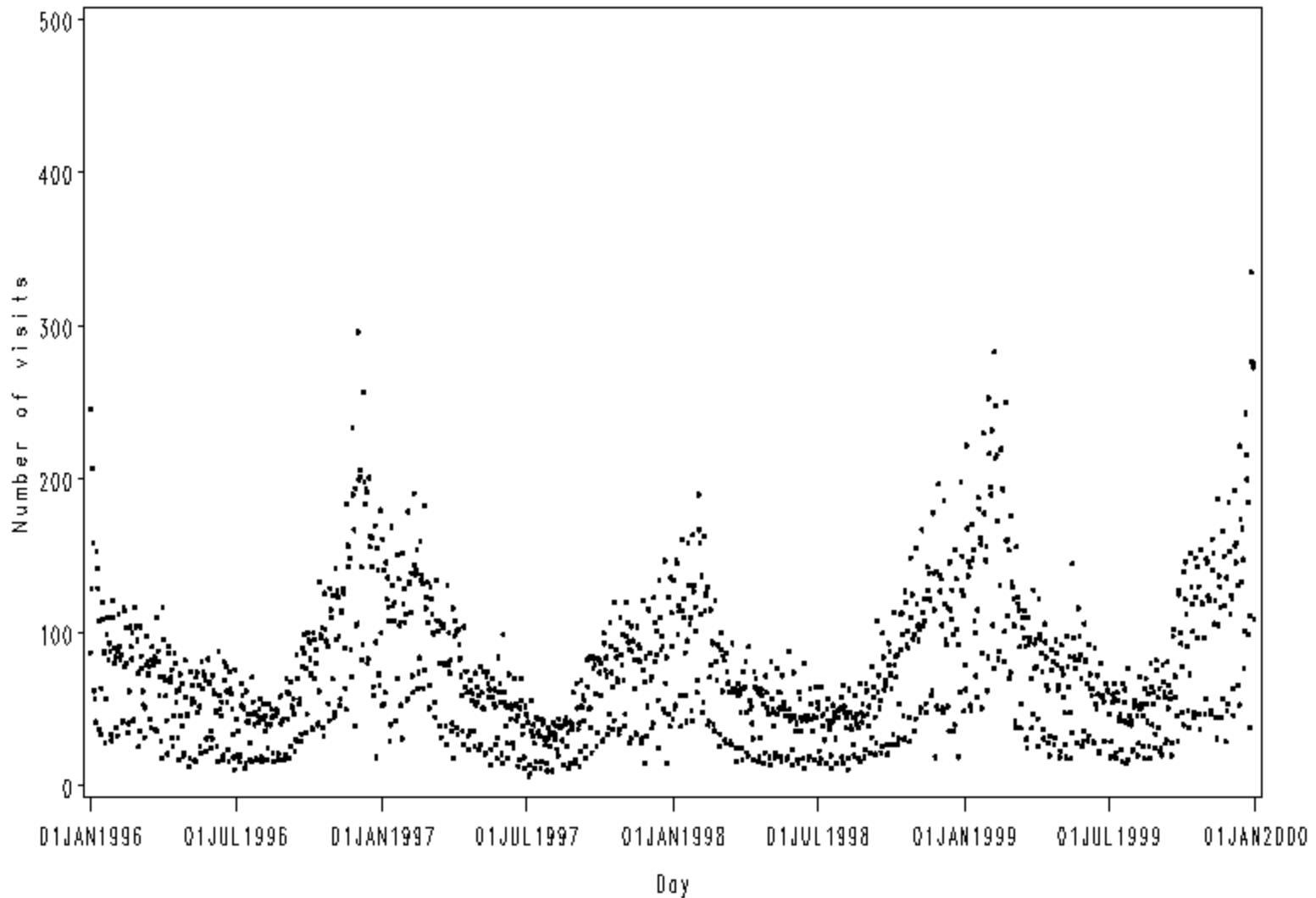
Syndromic approach: validity

Weekly US Pneumonia/Influenza deaths (CDC) and HPHC/Vanguard LoRI Episodes



A closer look at the data

Sum of LRI visits by day
previous 42 days w/a LRI complaint



Data analysis?

- With this data, you might use a classic time-series analysis
- This will get you a threshold for the whole area under surveillance
- This alone is not great for public health purposes:
 - Not very sensitive. Need at *least* 10 extra cases to pass threshold in applications I know of
 - Where will you send the Cipro (field staff)? Don't want to dose (x-ray) 2.5 million people!

Spatial approach

- Geocode everyone using HMO records
 - Not just cases, but all insurees
- We get the exact latitude and longitude as well as census tract of residence
 - Only billing address, though. And the coding changes as the coding database improves.

Setting, revisited

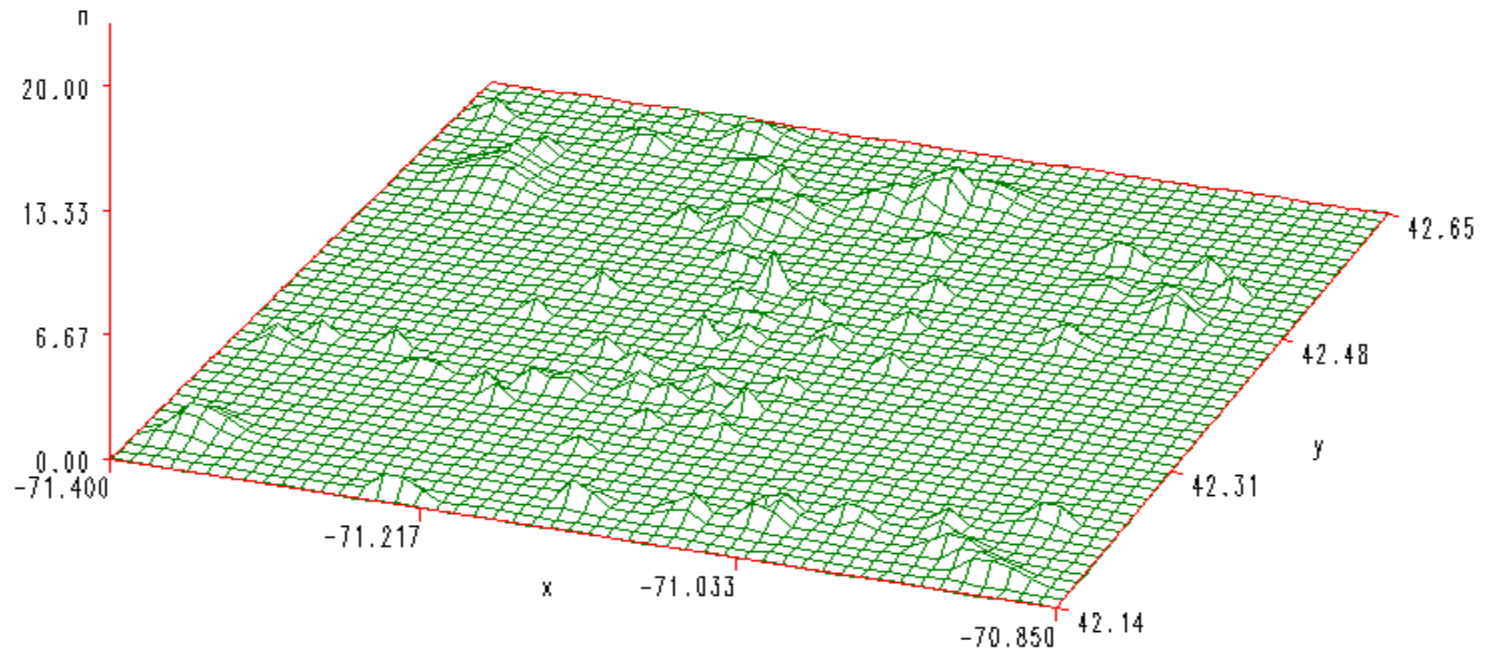
Coverage by census tract



As of October 31, 1999

Data, revisited

December 7 1998



Restating the main question: Is the bump too high?

- Many (most?) ‘spatial’ techniques tend to summarize over time and assume it to be static. Then they ask: Is there a bump?
- In contrast, time is as important a dimension for us as space.
- **We need to know whether there’s an event *now* just as much as *where* it happened.**

Is the bump too high?

- Some spatial techniques take time into account, and at least two are adapted for surveillance: they repeatedly incorporate more data and ask whether there is a bump ‘now’:
1) Kulldorf, 2) Rogerson: both 2001, JRSSA.

(The following complaints may refer only to the presentation in the above articles)

- Both assume time periods are equivalent (i.e. no adjusting for season or day of week)
- Both seemingly designed for smaller data streams

Is the bump too high?

- Both tools seem to have a uniform time-space distribution as the null
- Recall that phase 1 anthrax resembles flu. We *know* there will be clusters of flu symptoms
- **We need to know whether the bumps we're seeing are *bigger than we'd expect*, based on history, not bigger than random association.**
Not 'are there clusters?' but
'are the clusters (that we know exist) too big?'

A proposed analytic approach

- Treat each census tract as independent individuals (not true: closer in space ==> more highly correlated)
- Treat each day as a repeated observation on the census tract: count of syndrome visits each day is our outcome
- These are longitudinally repeated binomial observations: the denominator is the number of insureds living in the tract (updated daily)

Proposed approach

- We use the Generalized Linear Mixed Model approach to logistic regression
- This is essentially a logistic regression that takes into account the correlation between repeated days observed on a given tract, or, equivalently, different baseline risk in different tracts
- We could also use a GLMM Poisson regression

Proposed approach

- The model can be fit using the GLIMMIX macro that's distributed with SAS
- This allows relatively easy use of the system by other insurers, care providers, or analysts
- We could also have used GEEs, but GLMM allows a separate estimate for each tract, which is crucial to what we will do

Proposed approach

- The model looks just like a logistic regression, with some additional subscripts and one more parameter:

$$E(y_{it} | b_i) = n_{it} p_{it}; \quad \log \left(\frac{p_{it}}{1 - p_{it}} \right) = x_{it} \mathbf{b} + b_i$$

- where i is the census tract with repeated days t , y_{it} is the number of visits, n_{it} is the number of insured, and b_i is a random effect: $b_i \sim \mathbf{N}$

Proposed approach: face validity

- Fixed effects (x_{it}) in our current model: 11 months, 6 days of week, national holiday
- Odds by month highest in winter months, lowest in summer
- Odds by day highest Mondays, lowest on weekends
- OR for holidays less than 1

Proposed approach

- The random effect b_i models the unique features of each census tract: is there a little community of hypochondriacs somewhere? Are there more elderly or children?
- The estimated variance of the b_i is not 0: there really are differences between tracts.
- The estimated random effects (a.k.a. shrinkage or empirical Bayes estimators) are the odds of a case of LRI in tract i relative to the average tract.

Using the model for public health

- To use the model, we invert the estimated logit for each tract to get an estimated binomial p_{it} for each census tract i on day t
- Example: surveillance day t is a Monday in April

Suppose the estimated intercept = -8, April effect = -0.6, Monday effect = 0.3, that it is not a holiday, and that census tract i has estimated $b_i = 0.3$. Then

$$\hat{p}_{it} = \left(\frac{e^{-8-0.6+0.3+0.3}}{1 + e^{-8-0.6+0.3+0.3}} \right) = 0.000335$$

Using the model

- Then we calculate the probability of seeing as many cases as we saw, or more
- Based on Binomial distribution:

$$\Pr(X = x) = \binom{n_{it}}{x} \hat{p}_{it}^x (1 - \hat{p}_{it})^{n_{it} - x}$$

- $P(3+ \text{ cases}) = 0.0049$; $P(4+) = 0.000402$
- This is basically a p-value, for H_0 : the data come from a binomial distribution with the p_{it} estimated from the model

Using the model

- There are 529 census tracts in the relatively heavily-populated area we focus on
- We estimate a p-value for each tract each day
- A small multiple comparisons problem \implies
~180,000 tests/year
- We do a Bonferroni-like thing to adjust for this.

Using the model

- Instead, we report the number of years we would have to test every tract every day in order to (statistically) expect one p-value this small (or smaller).
- This is like when the weather service says it's a 'hundred year flood'; we can say 'it's a hundred year count'.
- One advantage to expressing it this way is that big is bad.

Using the model

- The inverse of the expected value of the number of p-values this size or smaller in one year, assuming independence of all tests is

$$(\text{nominal } p_{it} * 365 * \# \text{tracts})^{-1}$$

- Example, revisited:

count of 3: $p_{it} = .0049 \Rightarrow (.0049 * 193085)^{-1} = .001$ years,
i.e. we expect counts this ‘unusual’ 946 times per year

count of 4: $p_{it} = .0004 \Rightarrow (.0004 * 193805)^{-1} = .013$
years, i.e. expect counts this extreme 78 times per year

count of 6: $p_{it} = 0.00002 \Rightarrow 2.6$ times per year

Our current report

Wednesday, 18 September 2002*

Town	Census tract	Cases	Denominator	Years between [^]
Boston	250250706	2	299	0.003
Chelsea	250251603	1	69	0
Brookline	250214012	2	827	0
Boston	250250910	1	353	0
Medford	250173397	1	361	0

- * The 5 most extreme tracts are shown...

[^] Estimated number of years between daily counts this extreme...

Our current report

- Some days are more exciting looking than September 18th...

Thursday, 14 March 2002

Weston 250173672 4 477 0.153

- This event happens only once every 55 days, rare enough that the Mass. DPH was interested.
- Note that only 4 cases was enough to be interesting

Problems with this approach

- An event may take place over more than one census tract, at a low enough level that no single one of them looks terribly interesting
- Our model doesn't assess this possibility
- To help alleviate this problem, we currently rely on the visual cortex.

Neighboring areas: current approach

Iri syndrome

12/03/2001: 5 most extreme counts

Everett, 3424, 2 cases

Weymouth, 4228, 2 cases



Watertown, 3702, 2 cases

Quincy, 4182, 2 cases

Boston, 1001, 2 cases

With approximate place and last four digits of 1990 census tract number

Important outstanding problems

- ① What to do with cases in adjacent tracts. The present solution is just short of totally inadequate. Directions: incorporating additional test based on SaTscan?
- ② Incorporating spatial correlation into the model. This will smooth baseline risks across census tracts. This is relatively easy, but probably won't affect prediction much.

Important outstanding problems

- ③ Incorporating individual-level covariates. We should be more concerned when a 20 year-old is a case than when an 80 year-old is.
- ④ Simulating data so we can determine how well the model (or other models) will work under different conditions. Note that this means modeling flu, for the most part, not anthrax. I think this is the thorniest and most important task right now.

Summary

- There's no ideal tool already developed out there. The current technique addresses some drawbacks of available tools but has others of its own.
- We probably need a suite of tools to be powerful to different alternatives
- Assessing the sensitivity of any proposed tool requires a viable simulation of the background noise.

Why surveillance for anthrax?

Course of the disease

Timeline	Course (inhalation anthrax)
Day 0	Exposure
Day 2-7	Phase 1: Flu-like; fever, cough, fatigue
Day or two later	Phase 2: Severe breathing problems, shock
Day or two later	Death in 90% of untreated cases

Why surveillance for anthrax?

Victim's behavior

Timeline	Course	Action
Day 0	Exposure	
Day 2-7	Phase 1	Visit their Doctor?
Day or two later	Phase 2	Call 911? Go to the hospital, get Cipro
Day or two later	Death	

Why surveillance for anthrax?

Public health response

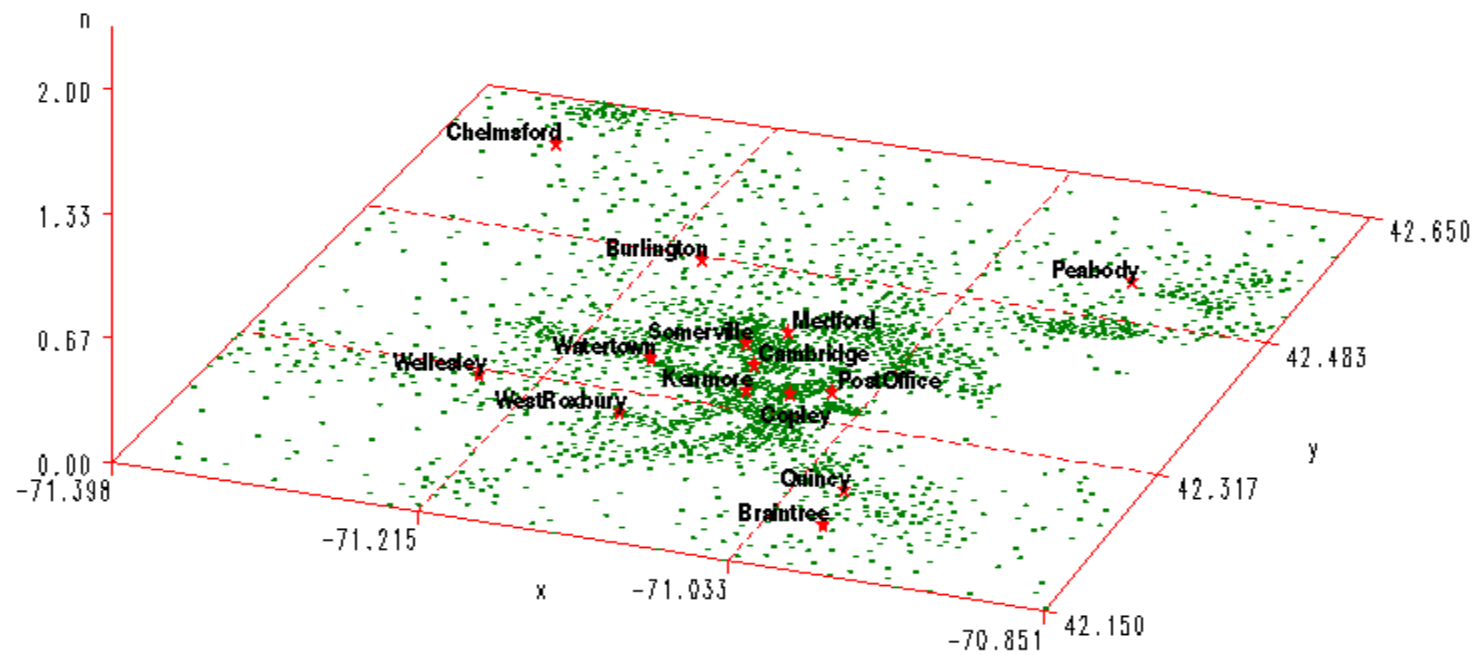
Timeline	Course	Action	Reaction
Day 0	Exposure		
Day 2-7	Phase 1	Visit Doctor?	
Day or two later	Phase 2	Hospital	Get ready for the flood; lots of people already in phase 2. Many beyond help.

Why surveillance for anthrax?

Public health response

Timeline	Course	Action	Reaction
Day 0	Exposure		
Day 2-7	Phase 1	Visit Doctor?	Break out the Cipro: most victims in Phase 1 or earlier

CBG centroids and HVMA clinics



Other approaches

- Another possibility to control for the impact of all those tests is to make connections with imaging problems
- They need to know, e.g., which voxel of a brain is activated in a particular time frame; many millions of tests for a single experiment
- A promising approach there is to control the ‘false discovery rate.’ Useful here?

Important outstanding problems

- ⑤ Including a ‘badness-of-season’ indicator to be more sensitive during years when the flu is not so bad, less when it is worse than usual. This requires fitting the model more often than we current do, but should not be too difficult. On the other hand, if the time between exposure and symptoms is more variable than currently thought, this might miss events.
- ⑥ Assess boundary definitions: CT vs. CBG...