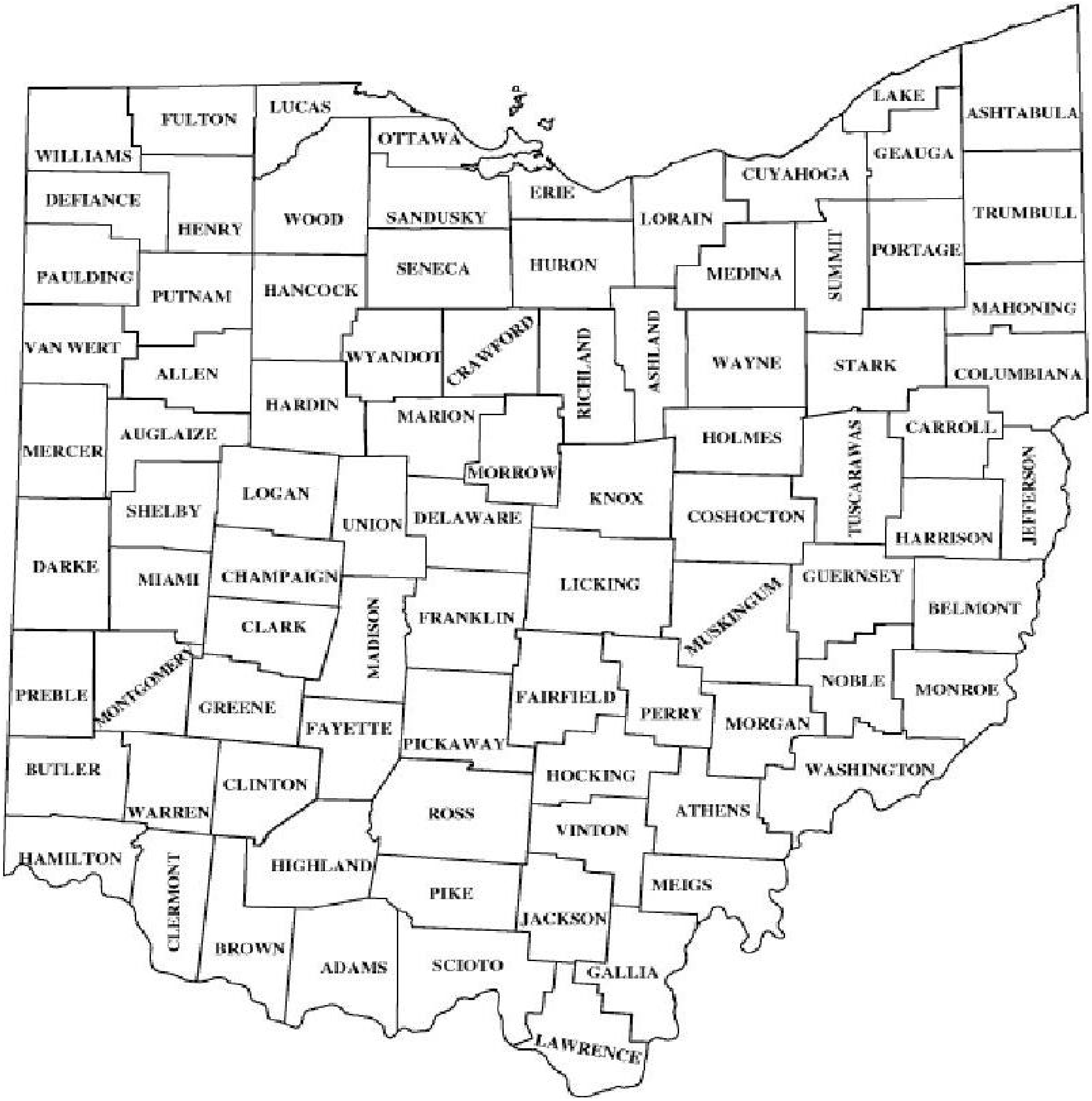
CROP DISEASE AND HOTSPOT GEOINFORMATICS: FUSARIUM HEAD BLIGHT

Murali Haran

The Pennsylvania State University

Ecometrics, Oct.13,2005

Ohio’s 53 Counties



Information available: weather data at sites (cells) of size 20km × 20km (over 300 cells total)

Goal: Using these weather covariates, we would like to predict risk of disease for that site.

Available Data

I. Experimental Plots:

Severity	Wheat Type	Susceptibility	Corn	Weather Cov 1-5

Susceptibility denotes how susceptible the wheat variety is to FHB disease on a scale of 0-3.

Corn (1 or -1) indicates whether there is corn residue present, typically due to corn being planted the previous year.

The 5 weather covariates include information on relative humidity, temperature, precipitation.

Data available from 1990-2005: 12 plots total in 7 states.

II. True risks/severity (after the fact):

Locn	Severity	Wheat Type	Suscept	Corn	Weather 1-5

Locn=County for Ohio, or Field's Latitude and Longitude for North Dakota.

Data for 53 counties for Ohio and 347 fields for ND.

III. RUC (satellite) information:

Latitude	Longitude	Weather 1-5

We want to use this data (available at 20km×20km sites for the entire state) to forecast disease risks.

Current Approach for Predicting Risk

Experimental plots set up by plant pathologists where weather covariates measured and crop (Fusarium Blight) disease studied.

Based on these plots: Researchers have crop disease rates and relevant covariates (including weather).

They fit models of the form:

$$r_{1s} = f(w_{1s1}, w_{1s2}, \ldots, w_{1s7}) + \epsilon_{1s}$$

where r_{1s} is the risk (this year, t=1) for a particular site (s), w_{1si} is covariate i for site s (for this year).

f is a non linear function.

ϵ_{1s} is the error term for site s.

Based on this model we can predict risk of disease for a new site, using covariates observed for that site:

$$\hat{r}_{1s} = f(w_{1s1}, w_{1s2}, \ldots, w_{1s7})$$

Goal 1: Can We Do Better?

There are several other sources of useful information:

1. Observed weather covariates at each site from last year w_{0si} (weather covariate i at site s at time '0').
2. Predicted risks for each site from last year $\hat{r}_{0s}$ based on the same model.
3. *True* observed risks (disease rates) last year R_{0c}.
   - These true risks are based on survey information collected county-wide.
   - Note that R_{0c} is the observed risk at the *county* level, while the covariates and risks of interest are at the site level.
4. The location of sites.

(1)-(3) provides past 'truth' to improve our estimates.
(4) provides information about adjacencies, helping us account for spatial relationships.
All of this information should be utilized to improve our estimates of risk. How ?

Reminder: Types of Spatial Data

Point level data: data is observed at several fixed locations (points) on a map. We may be interested in finding out information about other locations based on the spatial 'surface' obtained from this point-level information. E.g.: can only test soil for contaminants at a few fixed locations.

Areal (spatially aggregated) data: The map is partitioned into several regions with well-defined boundaries and data is observed for an entire region (it is aggregated over that region). For instance: disease rates for an entire county.

Point pattern data: The data is observed at random locations where the locations themselves are of interest. E.g.: is distribution of a plant species uniform in a region?

County level observed risks: areal data.

Our data: weather covariates observed at fixed locations (point-level) but if the weather information is fairly uniform within a 10km^2 region, we should think of it as areal data as well.

The 'Misalignment' Problem

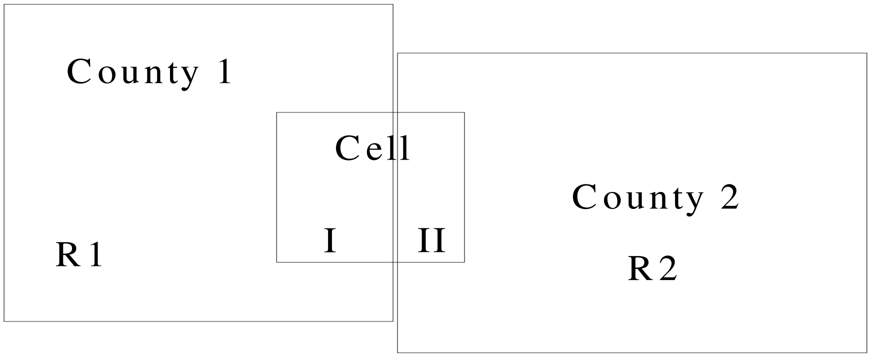
Ideal situation: we have true observed risks (disease rates) from last year at the site level, i.e. r_{0s}.
This would allow us to use observed discrepancies between the estimated risk values $\hat{r}_{0s}$ at each site and the true risk values r_{0s} at each site to directly improve our model.
Since r_{0s} is not available, we need to use R_{0c} to estimate r_{0s} before we work on improving our model.

This is generally called the "change of support" problem and is common. For example:
Spatial epidemiology problems: often health outcomes only available at county levels while socio-economic data may be available at zipcode level or pollutant monitoring sites are at the point-level.

Misalignment: Naive Approach

Use simple 'areal interpolation' to obtain an estimate of the disease risk in each cell.

For instance, if a cell is in more than one county:



Risk in cell = (R1 Area(I) + R2 Area(II))/ Area(Cell)

Misalignment (Contd.)

Unfortunately, this assumes that disease risk is evenly spread through the county (if this were true, we would not need to assess risk site by site).

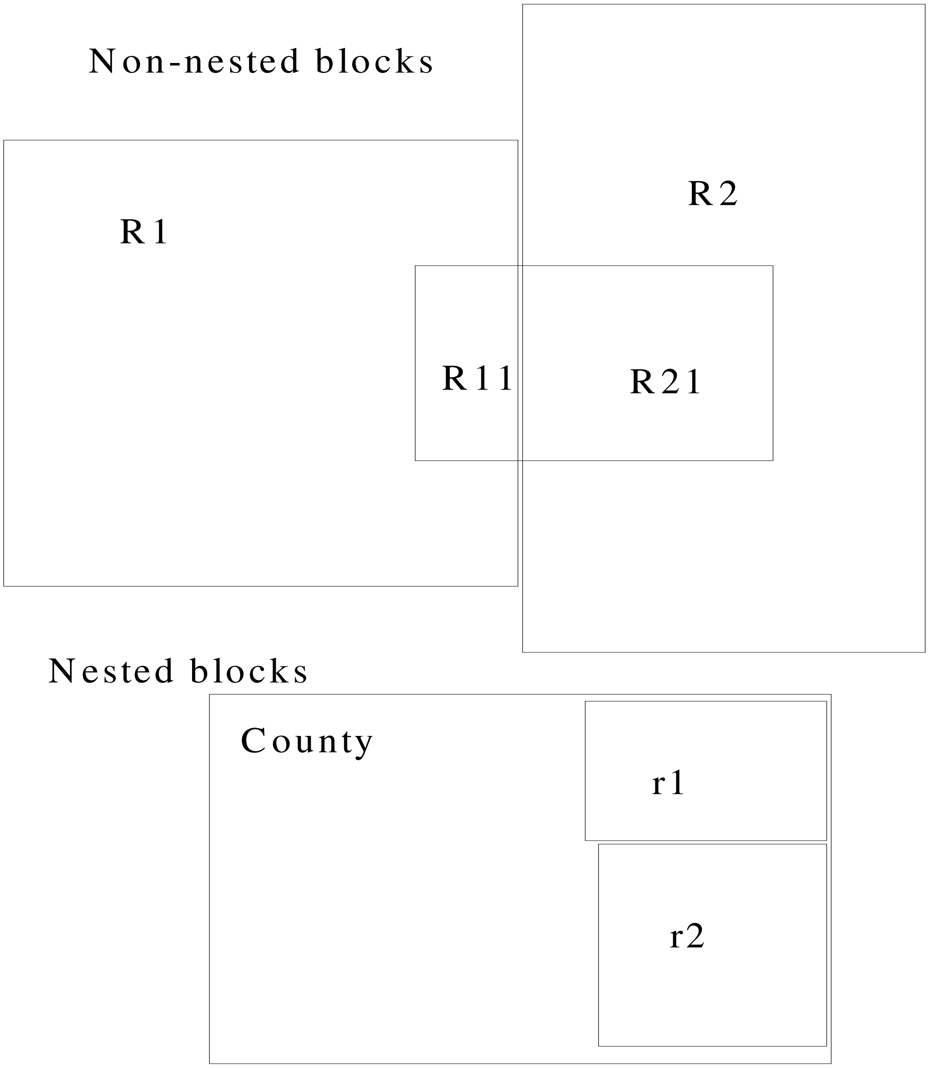
Also, it does not account for the variability of the associated estimate.

County level data: Need to be careful about ‘ecological fallacy’: relationships observed at county level may be inaccurate in judging the relationships between these same variables measured at the site (cell) level (cf. Gotway and Young, 2002).

For other states (e.g. North Dakota): true values of risk are available at the point level (specific farms). This presents another statistical challenge: using point level data for inference about aggregated data.

Should use model based (Bayesian) methods (cf. Banerjee, Carlin and Gelfand, 2005).

Areal Misalignment



There may be county level covariates, township level covariates etc. Natural to build model ‘hierarchically’: add up (or average) township predictions to get county predictions, but still account for covariates at all levels.

Bayesian Modeling: A 2-Slide Introduction

Instead of assuming parameters (θ) are fixed, think of parameters as random (they have distributions).

Assume *prior* distributions for parameters, $f(\theta)$.

As in frequentist modeling, we have a likelihood for the data $L(Y|\theta)$.

Once data are obtained, condition on data to obtain updated *posterior* distributions for the parameters, i.e. find $\pi(\theta|Y)$.

We use Bayes' Theorem:

$$\pi(\theta|Y) = \frac{f(Y,\theta)}{f(Y)} = \frac{L(Y|\theta)f(\theta)}{\int L(Y|\theta)f(\theta)d\theta}.$$

All statistical inference is based on properties of the posterior $\pi(\theta|Y)$, usually obtained by Monte Carlo simulation.

Some advantages: larger class of candidate models, allows for intuitive hierarchical specification of model, can account for uncertainty due to various sources, and does not rely on large sample theory.

Bayesian Modeling: A 2-Slide Introduction (Contd)

A short example:

Specify model $L(\mathbf{Y}|\theta)$ for the data where $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

Specify a prior distribution for the parameters $f(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

The prior can be specified either based on prior knowledge about the parameter θ or by using a prior that is as 'uninformative' as possible, i.e. it is dominated by the information from the likelihood (data).

$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ are the 'hyperparameters and may themselves be unknown. We can then place prior distributions on $\boldsymbol{\lambda}$, say $f(\boldsymbol{\lambda})$.

The posterior distribution is then

$$\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{Y}) = \frac{L(\mathbf{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})}{\int L(\mathbf{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})d\boldsymbol{\theta}d\boldsymbol{\lambda}}$$

Monte Carlo: If we simulate $(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_M), \ldots, (\boldsymbol{\theta}_1, \boldsymbol{\lambda}_M) \sim \pi$ can estimate all relevant properties of $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{Y})$ by simply using averages based on these samples.

Utilizing Spatial Information

We expect neighboring sites to have highly correlated disease rates/risks.

Important to note that the relationships between the risks of adjacent sites will most likely not be completely captured by available covariates.

For instance: while the weather covariates for neighbors are likely to be similar and would therefore account for some of the common risks, there are several other factors that are unaccounted for: Clearly since the disease is airborn, if a site is diseased the neighboring site is likely to also get the same disease.

We can 'borrow' information about a site from its neighbors: if a site's neighbors have a high risk of disease, that site is also more likely to have a high risk of disease.

Accounting for spatial dependence can help deal with spatial relationships and thereby act as a surrogate for other unobserved but important covariates.

Conditional Modeling: Markov Random Fields

One method for modeling spatial relationships is by using Markov Random Fields.

The disease counts (or risk) can be modeled conditionally: The true disease risk for a region i, say r_i, is assumed to be a random variable with mean=the average of all its neighbors and variance inversely proportional to the number of neighbors.

For instance: $r_i | r_{k \neq i} \sim N\left(\frac{\sum_{j \sim i} r_j}{n_i}, \frac{1}{\tau_c n_i}\right)$ where r_i is the risk for the ith region, n_i is the number of neighbors for the ith region $i \sim j$ implies i and j are neighbors and τ_c is a precision parameter.

Here we denote the distribution by: $(r_1, \ldots, r_n) \sim CAR(\tau_c)$

If all the risks are specified conditionally, how do we get the distribution of a single risk (marginal distribution) ?

By specifying prior distributions on the parameters (Bayesian model) and using Monte Carlo simulation methods.

Example of an MRF hierarchical model

Model due to Besag, York and Mollie (1991):

$$Y_i|\mu_i \sim \text{ Poisson}(E_i \exp(\mu_i)),\ i = 1, ..., N,$$

Y_i, E_i: observed count, estimate of expected disease events in region i respectively.

μ_i is our parameter of interest, $\mu_i = \theta_i + \phi_i$.

θ_i's vary in a completely unstructured way

$$\theta_i|\tau_h \overset{iid}{\sim} N(0, 1/\tau_h)$$

ϕ_i's account for spatial relationships: CAR (conditional autoregression) prior.

Add priors for the precision parameters τ_h, τ_c.

Specifying distributions one level at a time (conditionally) in this manner is refered to as hierarchical modeling.

Distribution of interest here: $\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \tau_h, \tau_c|Y)$, of $2N + 2$ dimensions. We use simulation methods for inference.

For example can find: $E(\mu_i|Y)$ for any region i and estimate any other posterior expectation.

Problem 2: Detecting and tracking hotspots

Once we have very good estimates of the risk for each cell we can study the 'risk' map.

Note that we now have (posterior) distributions for risk estimates, i.e. $(\pi(r_{1s}|Y))$ so we can automatically produce (Bayesian) confidence intervals for the risk estimates.

Using this map, we can:

- Look for hotspots, i.e. clusters of sites (cells) where the risk is high. All sites in this region can then take collective action based on this forecast.
- Hotspots can be found based on the risk surface at each time point. Studying hotspots over time allows us to learn about hotspot trajectories, i.e. how do the hotspots move over time?

  These trajectories can be helpful for forecasting high risk clusters.

Some References

Hierarchical Modeling and Analysis for Spatial Data, Banerjee, Carlin and Gelfand (2005).

R and geoR available from: www.cran.r-project.org

WINBUGS statistical software.