



---

## Center for **S**tatistical **E**cology and **E**nvironmental **S**tatistics

---

Large Scale Plant Disease Forecasting: Case Study of Fusarium Head Blight

By Murali Haran,<sup>1</sup> Julio Molineros,<sup>2</sup> and G.P. Patil<sup>1</sup>

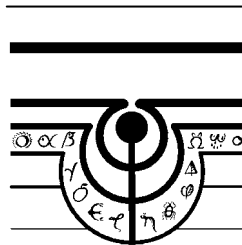
<sup>1</sup>Department of Statistics, <sup>2</sup>Department of Plant Pathology,  
The Pennsylvania State University,  
University Park, PA 16802, USA

This material is based upon work partially supported by (i) the National Science Foundation under Grant No. 0307010, (ii) The United States Environmental Protection Agency under Grant No. CR-83059301 and (iii) The Pennsylvania Department of Health using Tobacco Settlement Funds and Grant No. ME 01324.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

[Presented at the 7<sup>th</sup> Annual International Conference on Digital Government Research]

Technical Report Number 2006-0530  
TECHNICAL REPORTS AND REPRINTS SERIES  
May 2006



---

Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802

G. P. Patil  
Distinguished Professor and Director  
Tel: (814)865-9442 Fax: (814)865-1278  
Email: [gpp@stat.psu.edu](mailto:gpp@stat.psu.edu)

<http://www.stat.psu.edu/~gpp>  
<http://www.stat.psu.edu/hotspots>

DGOnline News  
[Environmental and Ecological Statistics-Springer](http://www.stat.psu.edu/hotspots)

# Large scale plant disease forecasting: Case study of Fusarium Head Blight

Murali Haran  
Statistics Department  
Pennsylvania State University  
University Park, PA  
mharan@stat.psu.edu

Julio Molineros  
Department of Plant Pathology  
Pennsylvania State University  
University Park, PA  
jem320@psu.edu

G.P.Patil  
Statistics Department  
Pennsylvania State University  
University Park, PA  
gpp@stat.psu.edu

## ABSTRACT

Hotspots are locations or regions that have consistently high levels of disease and may have characteristics unlike those of the surrounding areas. We study disease mapping and hotspot detection for Fusarium Head Blight (FHB), a disease which affects wheat crops. The data available include geo-referenced observations of disease intensity for past years, radar-based weather data and information on crop/disease management practices. Risk predictions are available from weather-driven models of disease biology based on experimental data. Also, observed disease rates based on surveys done after each season provide information about the ground truth. We describe some methods for utilizing these sources of information for validation and for improving prediction in future years. The varying sources of information available to us are often sparse and spatially and temporally misaligned which leads to several interesting and challenging statistical and computational issues in order to help develop a geoinformatic disease management system for an advisory to the farmer on whether to spray or not.

## Categories and Subject Descriptors

H.4 [Information Systems Applications]: types of systems

## General Terms

Algorithms, Plant Disease, Management

## Keywords

wheat crop disease, spatial data, hotspot detection, space-time hotspots

## 1. INTRODUCTION

Hotspots are locations or regions that have consistently high levels of disease and may have characteristics unlike those of the surrounding areas. We study disease mapping and

hotspot detection for Fusarium Head Blight (FHB), a disease which affects wheat crops. Risk predictions for FHB are available from weather-driven models of disease biology based on experimental data. We study data for the year 2005 in North Dakota. In addition, observed disease rates based on surveys done in 2005 provide information about the ground truth both during and after the flowering period (the time when wheat is susceptible to disease). These observations are therefore potentially useful for providing a way to assess the accuracy of the model predictions and for improving risk predictions in the future.

## 2. RESEARCH

One main challenge in using different sources of information for the purpose of model predictions lies in the spatial and temporal misalignment between the varying sources. The model predictions are based on Rapid Update Cycle (RUC) radar-based information about the weather conditions are available for a grid with cells of size  $20\text{km}^2 \times 20\text{km}^2$ . However, the observations of the ground truth are based on surveys of multiple farms within each of several counties. To further complicate matters, the observations are relatively sparse, with only a total of 370 farm-level observations between June 21st and August 8th, 2005, fewer than 7 observations per day. Also, on several days, multiple observations were taken in a small localized survey involving fewer than 5 counties. Thus, for any given date, there is little or no information about the true disease rates at most locations. On the other hand, the model predictions are available at every location on the RUC grid consisting of 545 cells over the entire state for each of 51 days between June 8th and July 28th, 2005. These model predictions at each time are for disease risk in a particular cell 3 weeks later. This leads to two main approaches: One based on developing fast algorithms for processing these predictions to identify FHB 'hotspots' for immediate action, and the other model-based approach that incorporates other sources of information such as survey data and spatial dependencies to optimally predict the true risks at each location.

### Upper level scan statistic

One approach we investigate is using a fast algorithmic technique for detecting significant hotspots based on the model risk predictions. The algorithmic technique we propose to use is an extension of the Upper Level Scan Statistic (ULS) proposed by [2]. ULS scan statistic was developed for regions naturally partitioned into subdivisions, typically counties,

postal zip codes, or subregions based on other methods of forming boundaries. At each cell  $i$  we have available a count of some event (typically a disease),  $Y_i$ , and a known ‘size’  $A_i$  which is proportional to the expected number of events for that cell.  $A_i$  is often taken to be the population of the cell or the area of the cell, perhaps corrected for covariates that determine disease counts. [2] develop a fast algorithm that searches through a subset of possible hotspots to find a few candidate hotspots that are most likely to be significant hotspots. Significant hotspots are determined via a classical likelihood ratio test where the p-values are obtained through simulation.

While [2] discuss this approach in the context of disease mapping based on count data, we develop methodology for its use in the current situation, where the risk surface is a continuous variable on the (0,1) interval. The new versions of the ULS developed in our crop disease context will employ gamma and beta models rather than the binomial and poisson models of [2].

In the original ULS scan statistic the cell count  $Y_i$  was modeled as either a Binomial or Poisson random variable with expectation given by  $A_i p_i$  and  $A_i \lambda_i$  respectively, where  $p_i$  is the Binomial probability of an event and  $\lambda_i$  is the intensity parameter for that cell. Now we consider two methods for modeling the risk for cell  $i$ ,  $r_i$ , which is a continuous random variable: the Gamma density and the Beta density. The Gamma density has support  $(0, \infty)$  so we transform  $r_i$  as follows:  $y_i = -\log(1 - r_i)$ . This is clearly a monotone transformation and  $y_i \in (0, \infty)$ . We can develop the model for the  $y_i$  by making reasonable assumptions. For instance we assume that expected risk is proportional to some known ‘size’ value  $A_i$  (provided by the domain experts). We omit details here. Alternatively, a Beta model places a distributional assumption directly on the risk parameter  $r_i$ , since the Beta distribution has support on  $(0, 1)$ . The Beta density can be reparametrized so that it can easily take into account assumptions about the relative size  $A_i$  of each cell and how it affects the expected risk. Our work also compares and contrasts the two modeling approaches.

## Model based approaches

In addition to the fast algorithmic approaches discussed above, we will also consider more model based approaches for such data. There has been some promising recent work in the area of spatial misalignment (see the discussion and references in [1] for instance). A hierarchical Bayesian approach is an appropriate framework for integrating various sources of information. To fix ideas: let the true risk for site  $s$  at any time  $t$  be  $\rho_{st}$ . These true risks are, of course, unobservable, but the true parameters of interest. The true risks are available for farms scattered across the state, denoted by  $R_{ft}$  for farm  $f$  and time  $t$ . We treat the model developed by the experimentalists as a black box that takes weather covariates and other relevant information as input and produces an estimate of risk based on the plant pathologists’ model. We denote this predicted risk by  $r_{st}$  respectively and assume it is an imperfect surrogate for the true risk  $\rho_{st}$ . Obviously, since the survey data are sparse, we will rarely have  $R_{ft}$  for a farm overlapping a given site even though we will have  $r_{st}$  at all times for all sites. We also know that nearby sites and farms are likely to have similar risks due to similar

conditions and presence of the spores that cause FHB. Our approach is to formulate a model that relates  $r_{st}$ ,  $R_{ft}$ ,  $\rho_{st}$  for all sites and available farms, thus allowing us to predict risk for any site based on all available information.

## Broader Applicability

The methods we propose will be of use not only in crop disease forecasting but also in other important problem where data are available in similar form, namely some combination of easily available but possibly less accurate model predictions and more accurate but very sparse survey data. Also, situations where there are compelling reasons to use information based on local conditions (such as weather) and experimental models, along with sparse survey information, could potentially benefit from methods and algorithms that are able to combine them in an optimal fashion.

## 3. ACKNOWLEDGMENTS

This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010 and (ii) the United States Environmental Protection Agency under Grant No. CR- 83059301. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

## 4. REFERENCES

- [1] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Models and Analysis for Spatial Data*. Chapman & Hall CRC, 2004.
- [2] G. P. Patil and C. Taillie. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11(2):183–197, 2004.