Estimating the risk of a crop epidemic from coincident spatiotemporal processes

Murali Haran
Department of Statistics
The Pennsylvania State University
mharan@stat.psu.edu

K.Sham Bhat
Department of Statistics
The Pennsylvania State University
kgb130@psu.edu

Julio Molineros
Department of Biostatistics
Case Western Reserve University
jmolineros@darwin.EPBI.CWRU.edu

Erick DeWolf
Department of Plant Pathology
Kansas State University
dewolf1@ksu.edu

Abstract

Fusarium Head Blight (FHB) or 'scab' is a very destructive disease that affects wheat crops. Recent research has resulted in accurate weather-driven models that estimate the probability of an FHB epidemic based on experiments. However, these predictions ignore two crucial aspects of FHB epidemics: (1) An epidemic is very unlikely to occur unless the plants are flowering, and (2) FHB spreads by its spores, resulting in spatial and temporal dependence in risk. We develop a new approach that combines existing weather-based probabilities with information on flowering dates from survey data, while simultaneously accounting for spatial and temporal dependence. Our model combines two space-time processes, one associated with pure weather-based FHB risks and the other associated with flowering date probabilities. To allow for scalability, we model spatiotemporal dependence via a process convolutions approach. Our sample based approach produces a realistic assessment of areas that are persistently at high risk (the probability of an epidemic is elevated for extended time periods), along with associated estimates of uncertainty. We conclude with the application of our approach to a case study from North Dakota.

1 Introduction

Fusarium Head Blight (FHB), also called "Scab", is one of the most important diseases affecting wheat crops worldwide. In the United States, FHB is most commonly caused by the residue-borne fungus *Fusarium graminearum* (Parry et al., 1995). In the last few years, FHB has resulted in massive losses in the midwestern United States alone. Losses due to FHB have been to the tune of hundreds of millions of dollars in quality and yield; overall losses are higher when associated crop management and insurance costs are taken into account (McMullen et al. (1997), Nganje et al. (2004)). However, if an FHB epidemic is anticipated, farmers can mitigate effects and greatly reduce losses by using appropriate crop management practices. In addition, future decisions about where to plant wheat crops can be made in the most cost-effective manner. Thus there is a great incentive for the development of accurate methods for identifying regions with a high risk of epidemics.

Research in plant pathology over the past decade has produced models for risk prediction based on climate information (Rossi et al. (2003), Hooker et al. (2002), De Wolf et al. (2003)). Recent work by Molineros (2007) has resulted in models that have since been deployed as part of a widely used online Fusarium Head Blight Risk Assessment Tool (`http://wheatscab.psu.edu`). The model uses the national weather service network to predict the risk of disease for any location chosen. The weather information used by such models is available across the United States as "Rapid Update Cycle" (RUC) data, from a national operational weather prediction system where predictions are obtained from multiple sources including devices on commercial aircraft and surface reporting stations. Its primary purpose is to serve users needing frequently updated short-range weather forecasts. The weather-driven models of disease risk have been developed based on experimental plots set up across several states in the United States, where the severity of the disease for plants within these plots was measured. Several weather covariates were also observed, thereby providing a means for constructing a model for predicting risk based on weather information.

An FHB epidemic can only occur when weather conditions conducive to the epidemic coincide with the flowering period for the wheat crop. Therefore, risk predictions are only meaningful when also accounting for flowering dates. Fortunately, survey information on flowering dates is often available and can be used to estimate the distribution of flowering dates at sites of interest. Combining information from weather-driven risk predictions and flowering dates is at the core of assessment of the risk of an FHB epidemic. The probability that crops are flowering at a particular location on a given date is a space-time process that we are able to infer from the observed spatial process involving just the estimated flowering dates at several locations.

FHB is primarily spread to adjacent regions when spores of the fungi are windblown or splashed onto the wheat heads. It is most likely to spread under particular weather conditions, typically when the levels of relative humidity and daily temperatures are adequately high (De Wolf et al., 2003). The weather factors are taken into account by the plant pathologists' models. Strong spatial dependence in FHB risks are largely due to the movement of spores from crop residues where the fungus survives. The spatial distribution of risk depends on the proximity to a common source of inoculum (spores that infect plants), which can be either local say within the same field (Dill-Macky and Jones, 2000), or at some distance (Schmale et al., 2005). Relevant details on the epidemiology of FHB can be found in Sutton (1982), Fernando et al. (2000), Champeil et al. (2004), and Dufault et al. (2006). Thus spatial proximity is a strong predictor of FHB risk; if a farm has an FHB epidemic, an adjacent farm is very likely to also have an epidemic. Similarly, if a location has a high (or low) probability of an FHB epidemic on a particular day, it is likely to have a high (or low) probability of an epidemic the following day as well. It is therefore vital to account for spatio-temporal dependence to produce accurate risk estimates. Current models do not account for the spatial or temporal proximity; however, space-time dependence is an important feature of the model we propose here. Our sample based approach allows us to easily identify regions where the risk is high for extended periods of time, which is of particular interest to farmers.

Our goal in this paper is to develop a systematic statistical framework that combines information from state of the art weather based models of epidemic risk with information from survey data on flowering dates to identify regions that are persistently at high risk. Our model also accounts for critical space-time dependence and our sample based approach naturally allows us to express uncertainties about our estimates. Since we envision applications of this approach to large regions and data sets, computational considerations play an important role in how we develop our model. The rest of the paper is organized as follows: In Section 2, we describe our approach for combining raw weather model based maps with flowering data, and provide details about the separate space-time and space models we fit to the two data sets respectively. In Section 3, we describe in detail the application of our methods to data sets for the state of North Dakota, a region of particular economic interest due to its long history of severe FHB epidemics. These new risk maps can potentially provide state officials and farmers with an assessment of the areas that are truly at high risk and the periods when they are at high risk. This can be very helpful when making decisions about where and when to plant wheat crops and whether or not aggressive (and expensive) management practices should be implemented. We conclude with a discussion in Section 4.

2 Crop Epidemic Model

We consider the weather-driven plant pathology model for assessing the risk of an FHB epidemic described in Molineros et al. (2006). The risk model is a logistic regression model with predictors that include weather variables such as mean ambient relative humidity, and crop susceptibility (an indicator of how susceptible a particular variety of wheat is to FHB). For a particular plot or field, the response variable (whether or not the field has an FHB epidemic) is binary, with an epidemic defined as a disease severity of at least 10% for that field. The model was developed on the basis of data gathered from multiple experimental plots across the United States. The weather information gathered in the experiments and later utilized for prediction was based on knowledge of the disease biology for FHB. A critical observation is that the model was developed with specific knowledge of flowering dates of the research plots; the plant pathologists' model is therefore based only on data collected when the crops are flowering. Hence, implicitly, the model developed is one that relates weather information to the risk of an FHB epidemic, *given* that the crop is flowering. Hence, when using this model for predictions about risk, it is vital that we consider the probability that the crop is flowering at a given time.

Flowering information about wheat crops is obtained from surveys taken at several locations across the region. Since the survey locations do not coincide with the locations where weather information is available, there is a need for spatial interpolation of flowering information before it can be combined with the weather based probabilities above. We now describe our model for combining these two sources of information.

2.1 Spatio-temporal risk mapping

Denote the plant pathology model based probability of FHB epidemic for site $\mathbf{s}$ at time t, by $p(\mathbf{s},t) \in [0,1]$ for all $\mathbf{s} \in D$ and $t \in T$, where $D \in \mathbb{R}^2$ is the study region and T is the time interval of interest. Define the logit transformation, $r(\mathbf{s},t) = \log\left(\frac{p(\mathbf{s},t)}{1-p(\mathbf{s},t)}\right)$. Our model for $r(\mathbf{s},t)$ allows for spatio-temporal dependence and an independent error term,

$$r(\mathbf{s},t) = w(\mathbf{s},t) + \epsilon(\mathbf{s},t) \quad \text{for } \mathbf{s} \in D,\ t \in T, \tag{1}$$

where $w(\mathbf{s},t)$ is a space-time process, and $\epsilon(\mathbf{s},t) \overset{iid}{\sim} N(0,\lambda_r^{-1})$, is an independent error process with precision λ_r. Since the motivation of this research is to model risks across a much larger portion of the U.S. (potentially several states and several thousand locations overall), we consider the computationally convenient process convolutions approach due to Higdon (1998) for modeling the spatiotemporal dependence. This approach has the added benefit of easily allowing for non-stationary processes (cf. Higdon et al. (1999) and Calder et al. (2002)), which may be impor-

tant for modeling a spatial process across regions with vastly different climactic and geographic characteristics. The formulation is an alternative to the usual Gaussian process based model and relies on the fact that a continuous process can be created by convolving a continuous white noise process with a convolution kernel, k (Higdon (1998)). We first describe a spatial process. Following Higdon et al. (1999), denote a continuous white noise process z with precision λ_z as

$$w_A = \int_A z(u)du \sim N(0, \lambda_z^{-1}\mathsf{area}(A)) \;\; \forall A \subset D,$$

where for any subregions A, B contained in D,

$$\mathsf{Cor}(w_A, w_B) = \mathsf{area}(A \cap B),$$

thereby creating a continuous spatial process over the region D defined at any point $\mathbf{s}$ by

$$w(\mathbf{s}) = \int_D k(\mathbf{u} - \mathbf{s})z(\mathbf{u})d\mathbf{u}. \tag{2}$$

Here, the convolution kernel k is taken to be the circular normal,

$$k(\mathbf{u}) = (2\pi\sigma^2)^{-1} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{u}'\mathbf{u}\right\}.$$

In practice, the continuous white noise process z is replaced by a more convenient finite sum approximation $\mathbf{z}$ defined on a lattice $\mathbf{u}_1, \ldots, \mathbf{u}_L$ that covers the study region. The kernel used here corresponds to a Gaussian covariance function, a member of the Matérn family of covariance functions (Stein, 1999). This kernel has the advantage that it can be written in closed form, while this is not generally true for kernels corresponding to Matérn covariance functions. In spite of its popularity, Stein (1999) cautions against the use of the overly smooth Gaussian covariance function. However, in our work, since the weather information used to derive crop disease risks is available on an equally spaced set of grid points, there is little information about the behavior of the spatial covariance near the origin, which is critical for determining smoothness of the spatial process. We are also typically not interested in interpolation at a finer spatial or temporal scale than that provided by the data; hence, the smoothness implied by the Gaussian covariance function assumption appears to be a reasonable and convenient assumption.

In addition to the finite sum approximation, if we also allow for a non-zero mean $\mu(\mathbf{s})$ to capture spatial trends, we replace (2) by the following spatial process

$$w(\mathbf{s}) = \sum_{j=1}^{L} k(\mathbf{u}_j - s)z(\mathbf{u}_j) + \mu(\mathbf{s}),$$

where $z(\mathbf{u}_j)$ is the value of the white noise process at location $\mathbf{u}_j$. The 'knot' locations $(\mathbf{u}_j)$ are defined on a lattice covering the region of interest, for instance a 10×9 lattice. Following

our general discussion about process convolution models above, we can now define a space-time version of the model we will use for studying FHB epidemic risks.

$$w(\mathbf{s},t) = \sum_{j=1}^{L} k(\mathbf{u}_j - \mathbf{s}; v_j - t) z(\mathbf{u}_j, t) + \mu(\mathbf{s}),$$

where $(\mathbf{u}_j, v_j)$ defines our jth space-time knot. So our new set of knots are $((\mathbf{u}_1, v_1), \ldots, (\mathbf{u}_J, v_J))^T$, with $\mathbf{u}_j \in D$ and $v_j \in T$ for $j = 1, \ldots, J$, which defines a lattice over the entire region and time period of interest. $w(\mathbf{u}_j, v_j)$ is the white noise process ('knot process') at the jth knot. The model (1) for the random field can therefore be written as

$$r(\mathbf{s}_i, t_i) \mid \mathbf{z}, \lambda_z, \lambda_r, \boldsymbol{\beta}, \phi, \tau \sim N\left(\mathbf{X}(\mathbf{s}_i)\boldsymbol{\beta} + \sum_{j=1}^{J} K_{ij}(\phi, \tau) z(\mathbf{u}_j, v_j), \lambda_r^{-1}\right), \tag{3}$$

where $\mathbf{z}$ is the set of knot processes, λ_z, λ_r are precision parameters, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ are regression coefficients and ϕ, τ are space and time dependence parameters respectively. We assume a first order mean trend in the locations in space, so $\mathbf{X}(\mathbf{s}_i) = (1, s_{i1}, s_{i2})$. Typically, the locations will be (longitude,latitude), transformed to account for the curvature of the earth's surface. The kernel function, $K_{ij}(\phi, \tau) = \exp(-\|\mathbf{s}_i - \mathbf{u}_j\|/\phi - \|t_i - v_j\|/\tau)$, is space-time separable.

Let $\Theta = (\lambda_r, \lambda_z, \boldsymbol{\beta}, \phi, \tau)^T$. Details regarding prior selection are provided in Section 3. Once we specify a prior for Θ, we can obtain the posterior distribution of the risks and model parameters given the data using MCMC methods. Our inference regarding the underlying risks of an FHB epidemic at any site $\mathbf{s} \in D$ is then based on the posterior distribution,

$$\pi(\mathbf{z}, \Theta|\mathbf{r}) \propto L(\mathbf{r}|\mathbf{z}, \Theta) f_{\mathbf{z}}(\mathbf{z}|\Theta) f_{\Theta}(\Theta),$$

where $L(\mathbf{r}|\mathbf{z}, \Theta)$ is the likelihood from (3), with $\mathbf{r}$ the set of observed $r(\mathbf{s}, t)$, $f_{\mathbf{z}}$ is the joint distribution of the knot processes, and f_{Θ} is the prior on the model parameters. Risk estimates at *any* site $\mathbf{s}^* \in D$, and time $t^* \in T$ can be obtained from the posterior predictive distribution

$$\pi(r(\mathbf{s}^*, t^*)|\mathbf{r}) = \int \pi(r(\mathbf{s}^*, t^*), \Theta, \mathbf{z}|\mathbf{r}) d\Theta d\mathbf{z} = \int \pi(r(\mathbf{s}^*, t^*)|\Theta, \mathbf{z}, \mathbf{r}) \pi(\Theta, \mathbf{z}|\mathbf{r}) d\Theta d\mathbf{z}. \tag{4}$$

This is easily accomplished by standard MCMC based sampling in two stages: first using MCMC to draw samples from $\pi(\mathbf{z}, \Theta|\mathbf{r})$, then using the first stage samples to draw from $\pi(r(\mathbf{s}^*, t^*) \mid \mathbf{z}, \Theta)$. The kernel convolution approach provides large computational benefits over the usual Gaussian process models, which involve expensive matrix operations on large matrices. We discuss computation for this model in Section 3.

2.2 Incorporating the flowering process

We describe here the spatial model for flowering dates and how it can be used to adjust the epidemic probabilities to produce more accurate estimates of FHB risk. Let the observed flowering date at site $\mathbf{s} \in D$ be $\Psi(\mathbf{s})$ for a site $\mathbf{s} \in D$. We model the joint distribution of flowering dates via a Gaussian process so that for $\mathbf{s} \in D$,

$$\Psi(\mathbf{s}) = \gamma \mathbf{Y}(\mathbf{s}) + \xi(\mathbf{s}),$$

where $\mathbf{Y}(\mathbf{s})$ is the vector $(1, Y_1(\mathbf{s}), Y_2(\mathbf{s}))^T$, where Y_1 and Y_2 are the coordinates (longitude, latitude, transformed appropriately) for site $\mathbf{s}$, and γ is a vector of regression parameters. Since flowering dates are generally later in the year as we move north, there is a significant linear relationship between flowering dates and latitude. Note that the survey information does not typically contain the exact flowering date for a site; flowering dates are often estimated based on the physical characteristics of the crop before or after the actual flowering date. It is therefore important to account for non-spatial modeling error. We assume $\boldsymbol{\xi} = (\xi(\mathbf{s}_1), \ldots, \xi(\mathbf{s}_n))^T$ is distributed according to a zero mean Gaussian process with spatial dependence and measurement error,

$$\boldsymbol{\xi} \sim N(0, \lambda_d^{-1} I + H(\theta)),$$

where λ_d is a precision parameter, I is an $n \times n$ identity matrix, $H(\theta)$ is an $n \times n$ matrix specified by a Matérn covariance function (Matèrn, 1986), with parameter vector θ that controls the smoothness, scale and range of the dependence in the process (this is a standard specification, cf. Schabenberger and Gotway (2005).) We assume that if the date (t) is not within some c days of the true flowering date, the chance of an FHB epidemic is negligible (Parry et al., 1995). Hence, the risk of an FHB epidemic at site $\mathbf{s}$ and time t is obtained from (4) when $t \in (\Psi(\mathbf{s}) - c, \Psi(\mathbf{s}) + c)$ but is negligible (practically 0) if $t \notin (\Psi(\mathbf{s}) - c, \Psi(\mathbf{s}) + c)$.

We use survey data on flowering dates to estimate the distribution of flowering dates, which then allows us to estimate the probability that any given date t is a flowering date at any site $\mathbf{s}$. Note that the survey sites generally do not coincide with the sites where risk predictions are available. We complete the specification of the model by placing standard priors on $(\theta, \lambda_d, \gamma)$, with details provided in Section 3. This allows us to obtain, via MCMC (in similar fashion to Section 2.1), the posterior predictive distribution of flowering dates at any site $\mathbf{s}^* \in D$, $\pi(\Psi(\mathbf{s}^*) \mid \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi(\mathbf{s}_1), \ldots, \Psi(\mathbf{s}_n))^T$ are the survey flowering dates. $\delta(\mathbf{s}^*, t)$, the probability that t is within c days of a flowering date at a particular site $\mathbf{s}^* \in D$, is given by

$$\delta(\mathbf{s}^*, t) = P(t \in (\Psi(\mathbf{s}^*) - c, \Psi(\mathbf{s}^*) + c) \mid \boldsymbol{\Psi}) = \int I(t \in (\Psi(\mathbf{s}^*) - c, \Psi(\mathbf{s}^*) + c)) \pi(\Psi(\mathbf{s}^*) \mid \boldsymbol{\Psi})) d\Psi.$$

It is therefore easy to obtain a Monte Carlo estimate $\delta(\mathbf{s}^*, t)$, say $\hat{\delta}(\mathbf{s}^*, t)$, once we have samples $\Psi^{(1)}(\mathbf{s}^*), \ldots, \Psi^{(m)}(\mathbf{s}^*)$ from the posterior predictive distribution $\pi(\Psi(\mathbf{s}^*) \mid \mathbf{\Psi})$ for each $\mathbf{s}^*$:

$$\hat{\delta}(\mathbf{s}^*, t) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}(t \in (\Psi^{(j)}(\mathbf{s}^*) - c, \Psi^{(j)}(\mathbf{s}^*) + c)) \tag{5}$$

We can now produce risk maps that reflect the requirement that for an FHB epidemic to be likely, high risk weather conditions must coincide with flowering dates. These maps can then be used to assess the areas that were truly at the highest risk. Since we have access to the entire posterior distribution of the risk parameters it is easy to obtain a measure of uncertainty to go along with the risk predictions. The posterior predictive distributions of the risks for a given site $\mathbf{s}^*$ and time t, $\pi(r(\mathbf{s}^*, t)|\mathbf{r})$ quantify the predisposition that each site has towards an FHB epidemic. The posterior distributions of the flowering dates at each site, $\pi(\Psi(\mathbf{s})|\mathbf{d})$, provide a mechanism for finding the probability that a given date t is within c days of a flowering date. By combining these two sources of information, we can get a realistic assessment of the true risk, that is, the chance that an underlying predisposition due to weather conditions actually turns into an epidemic due to coincident flowering dates. The posterior distribution of the risk of an FHB epidemic for a site $\mathbf{s}^*$ and time t is estimated by using the models for risk and flowering and considering the product, $\pi(p(\mathbf{s}^*, t)|\mathbf{r})\hat{\delta}(\mathbf{s}^*, t)$, where $\pi(p(\mathbf{s}^*, t)|\mathbf{r})$ is obtained by transforming the posterior predictive samples from the (logit-transformed) distribution of risks $\pi(r(\mathbf{s}^*, t)|\mathbf{r})$. Furthermore, regions that are persistently at risk can be found by looking at locations $\mathbf{s}$ such that 95% credible regions for $\pi(p(\mathbf{s}^*, t)|\mathbf{r})\hat{\delta}(\mathbf{s}^*, t)$ are above a certain threshold for multiple times t. This is, in fact, an approach pursued in our application in Section 3.

3 Case Study: An Application to FHB in North Dakota

3.1 Data and model for FHB in North Dakota

We apply our methods to a data set on FHB risks in the U.S. state of North Dakota for a 51 day period between June 8th and July 28th, 2005. The data available includes risks associated with 545 sites corresponding to centroids of 20km $\times$ 20km sized cells arranged on a grid across the state. A plant pathology model (Molineros et al., 2006) is used to predict these risks using the weather information available from Rapid Update Cycle (RUC) data at each of these locations. In addition, survey data providing information on flowering dates are available at 365 locations across the state. The locations of the surveyed information do not coincide with those at which RUC data are available. However, it is possible to estimate the distribution for flowering dates at the RUC data sites using this survey information using the model in Section 2.2.

Figure 1 shows the weather-driven model risks at the grid locations for two dates, July 9th and 10th, in 2005. We have information on risks at the same locations for the remaining 44 days, with 1 week of missing data between July 19 and July 25th, for a total of 23,980 risk values from the weather-driven model. The spatio-temporal model in Section 2.1 is used to obtain estimates of the true underlying risk of an FHB epidemic. Figure 2 shows the flowering dates at surveyed locations around the state. This map is spatially smoothed as discussed in Section 2.2 to obtain distributions of flowering dates at the RUC sites. Information from the two sets of estimates is then combined to obtain estimates of the risk of FHB epidemics across North Dakota.

Here we describe some of the details of our approach for obtaining spatiotemporally smoothed weather driven risks in the context of our North Dakota data sets. We use a flat prior for $\boldsymbol{\beta}$, uniform priors for ϕ, τ and inverse gamma priors for λ_r, λ_z, with parameters for the priors obtained by using a combination of domain knowledge and exploratory data analysis. We also ensured that the distributions of the parameters are roughly centered on estimates obtained from the variogram of points which have small differences in space and time (Higdon, 1998), but that the priors were wide enough to allow for reasonable values for each parameter.

$$\boldsymbol{\beta} \propto 1, \quad \lambda_r \sim \mathsf{IG}(5, 14), \quad \lambda_z \sim \mathsf{IG}(5, 2.5)$$
$$\phi \sim \mathsf{Unif}(10, 600), \quad \tau \sim \mathsf{Unif}(0.1, 30),$$

For knot locations, we selected 25 spatial locations which were approximately equally spaced, and for each spatial location, we selected 7 equally spaced time points between the first and last observed dates, yielding a total of 175 knots. We selected the number of knots in space and time according to the heuristics described in Short et al. (2007). For instance, we note that the posterior distribution for the spatial dependence parameter stayed well above the knot-to-knot distance. We account for the curvature of the earth by transforming the latitudes and longitudes into approximate two-dimensional geodesic distances (see Banerjee (2005) for details.)

The posterior distribution for weather model based risks is obtained from $\pi(\mathbf{z}, \boldsymbol{\beta}, \lambda_r, \lambda_z, \phi, \tau | \mathbf{r})$, and we obtained a sample based version of the distribution of each spatiotemporal parameter using MCMC. For prediction at a new location $(\mathbf{s}^*, t)$, we used the posterior predictive distributions

$$r(\mathbf{s}^*, t) = \mathbf{X}(s^*, t)\boldsymbol{\beta} + \sum_{j=1}^{J} \mathbf{K_{*j}}^T \mathbf{z}(u_j) + \epsilon(\mathbf{s}^*),$$

where $\mathbf{K_{*j}}$ is the jth column of the kernel matrix $\mathbf{K}$. For the North Dakota data set, we used a sample consisting of 1000 draws from the posterior predictive distribution of $r(\mathbf{s}, t)$ for each of 545 locations for each of the 44 days in the study, which is obtained by retaining every tenth sample from the posterior distribution. Given the large number of parameters, retaining the subsamples is a convenient and computationally convenient approach for producing estimates

based on samples with reduced autocorrelations. The posterior predictive mean risk and 95% confidence intervals can easily be computed for any spatiotemporal location, and can easily be transformed into smoothed posterior probabilities $p(\mathbf{s}, t)$.

Modeled flowering date data is available at 365 survey sites, while we would like to predict the distribution for flowering dates at each of the 545 RUC locations. We fit the model discussed in Section 2.2, using standard inverse gamma priors for λ_d, λ_k, uniform priors for ν, ϕ and a flat prior for $\boldsymbol{\beta}$, with reasonable hyperparameters selected according to Finley et al. (2007). We used MCMC to obtain the posterior distribution for the parameters, which then allows us to simulate from the posterior predictive distribution for flowering dates at each RUC location. For each location, we have 1000 prediction samples at each of the 545 locations, giving us the posterior predictive distribution of the flowering date at each RUC location. The posterior mean flowering dates are shown in Figure 2. As discussed in Section 2.2, we use a mean function that is linear in the transformed locations. We assume a flowering window of c=7 days, which is the period of time in which an outbreak can occur (Parry et al., 1995). We then estimate the probability that an RUC site $\mathbf{s}$ is flowering at a given time t,

$$\delta(\mathbf{s}, t) = \mathsf{Pr}(t \in (d(\mathbf{s}) - 7, d(\mathbf{s}) + 7)).$$

Since we have a have a sample based distribution at each location of flowering dates, for each time t we can compute a Monte Carlo estimate for $\delta(\mathbf{s}, t)$ as in (5).

We then compute P(epidemic)=$p(\mathbf{s}, t|$flowering at time $t) * \hat{\delta}(\mathbf{s}, t)$ for a location $\mathbf{s}$ and time t, where $p(\mathbf{s}, t|$ flowering at time t) and $\hat{\delta}$(s, t) are the estimated spatiotemporal weather based risks and flowering date probabilities respectively. Thus, P(epidemic) is an estimate for the probability of an epidemic which has been adjusted for spatial and temporal dependence and the probability of flowering. For example, Figure 1 shows these adjusted FHB epidemic probabilities for July 9th and 10th, 2005.

In order to determine regions where crops are under long-term threat to FHB, we find areas of elevated risk of epidemic, which we define as any area where the probability of an epidemic is unusually high for a length of time. We consider a probability of 0.5 to be elevated, and three days as the definition of an extended period of elevated risk. We compute a 95% credible region for P(epidemic) at each spatiotemporal location $(\mathbf{s},\ t)$. To do this we compute a set of samples for P(epidemic) by multiplying each sample for the smoothed probability conditioned at $(\mathbf{s}, t)$ by the point estimate $\hat{\delta}(\mathbf{s}, t)$. From these samples, we calculate a 95% credible set for each location and time, as shown for six locations and times in Table 1, to ensure that the area has a statistically significant elevated risk. From this information, we can deduce the spatial locations $\mathbf{s}$ where there is potential for an epidemic over an extended period of time. The map of such persistent high risk regions is shown in Figure 3.

We utilized posterior predictive checks (Gelman et al., 2003) as a heuristic for assessing model fit for both the spatiotemporal risk and spatial flowering date models and found that our models fit well. For instance, we computed the coverage probabilities for the 95% credible interval by using the raw epidemic probabilities at each location. We found that for 91.8% of the locations, the 95% credible region contained the raw epidemic probability. Further investigation suggested that most of the observations that were not included in the credible sets generally had very low raw epidemic probabilities and very low posterior predicted epidemic probabilities and hence were inconsequential in risk assessment (as final estimates of risk were practically zero for all these regions.)

3.2 Computational details

The posterior distribution from Subsection 2.1 on risk and the distribution of flowering dates as described in Subsection 2.2 are analytically intractable so we naturally turn to sample based inference via MCMC. Since we make heavy use of MCMC, it is worth noting a few computational details. The process convolutions approach greatly reduces the dimensionality of the posterior distributions since the dimensions of the distribution depends on the number of knot locations selected, not on the number of sites. The model for spatiotemporal risks was fit in `C++`. We used block updating rather than univariate updating of the parameters in the Metropolis-Hastings algorithm. Block updating schemes generally result in Markov chains that are more efficient ('better mixing') than univariate schemes (cf. Liu et al. (1994).) The computation was manageable since matrix operations were performed on matrices of dimension 175×175, resulting in a significant computational gain when compared to fitting a standard (non process-convolution) Gaussian process. Of course, for much larger data sets, we would likely revert to some compromise between univariate and multivariate updates since the matrix computations would become prohibitive. MCMC for the distribution of flowering dates was relatively fast and convenient. We utilized spBayes (Finley et al., 2007), an `R package` (Ihaka and Gentleman, 1996) for fitting the model for flowering dates and obtaining samples from the posterior predictive distribution.

To ensure that our MCMC based estimates were reliable, we used standard heuristics such as starting the chain from different initial values and comparing resulting estimates. To determine how long to run the Markov chains, we used a stopping rule based on Monte Carlo standard errors for the posterior mean estimates computed by consistent batch means (Jones et al., 2006): when the standard errors for all parameter estimates were low enough, the chain was stopped. For instance, the posterior mean estimates of ϕ, τ, λ_r had standard errors of under 0.2, 0.01 and 10^{-4} respectively, while $\beta_0, \beta_1, \beta_2$ had standard errors under 5×10^{-4}. MCMC was run for 100,000 iterations for the model for flowering dates, and all relevant standard errors were very small, below

10^{-4} in every case.

4 Discussion

We find that our approach successfully combines two different factors, weather conditions and flowering dates, along with space-time dependence in each, to estimate the risk of an FHB epidemic. Current approaches do not consider flowering and ignore space-time dependence, which are both key factors in determining risk. Our sample based methodology automatically provides error estimates which we, in turn, use to produce risk maps that identify regions that have significantly elevated risks for extended periods of time. Our process convolutions based modeling strategy also scales well, which is critical if these methods are to be useful for data sets involving many more observations in space and time. The results from our North Dakota case study suggest to us (the plant pathologists JM and ED), based on our domain expertise, that our methods produce good estimates. The sample based approach provides the additional flexibility of easily assessing regions that are persistently at significantly high risk, as described in Section 3. Of course, what may be of great practical use will be assessing the areas of high or low risk under several different scenarios. This can be achieved using our approach with multiple historical data sets along with weather forecasts. An interesting approach to consider would also involve modeling the two space-time processes jointly, though our current data sets and domain expertise do not suggest obvious approaches for doing this.

We note that the methodology developed here is potentially applicable to a large number of important related problems, including models for other crop epidemics. We believe our sample based approach for combining inference from two processes, the first corresponding to a predisposition (in our case, 'optimal' weather for FHB spread and establishment) and the second corresponding to a necessary triggering event (in our case, crop flowering), has applications to other research problems in ecology. For example, the authors' ongoing research on the spread of invasive plant species can be divided into two processes: the first related to the presence of nutrients in the soil, which describes the predisposition of a location towards hosting invasive plant species, while the second process involves a triggering event such as optimal weather conditions.

Acknowledgments: The authors thank Brad Carlin and Andrew Finley for helpful discussions. We also thank Marsha McMullen for providing us with spatially explicit data for North Dakota.

Location (Lat, Long)	Date	$p(\mathbf{s},t)$	$\hat{\delta}(\mathbf{s},t)$	Probability of Epidemic (95% region)
(46.55,-96.72)	7/13	0.3616	0.968	0.3510 (0.2460,0.5299)
(47.71,-97.74)	7/7	0.1390	0.618	0.1137 (0.0650,0.2535)
(47.39,-96.99)	7/5	0.1086	0.665	0.1454 (0.0878,0.2175)
(47.39,-96.99)	7/8	0.5504	1.000	0.4252 (0.2621,0.6791)
(48.72,-97.27)	7/18	0.4192	1.000	0.2947 (0.1666,0.5075)
(48.72,-97.27)	7/19	0.4900	0.966	0.3272 (0.2227,0.5291)

Table 1: Posterior Epidemic Probabilities with $p(\mathbf{s},t)$=weather based, $\hat{\delta}(\mathbf{s},t)$=estimated flowering probability

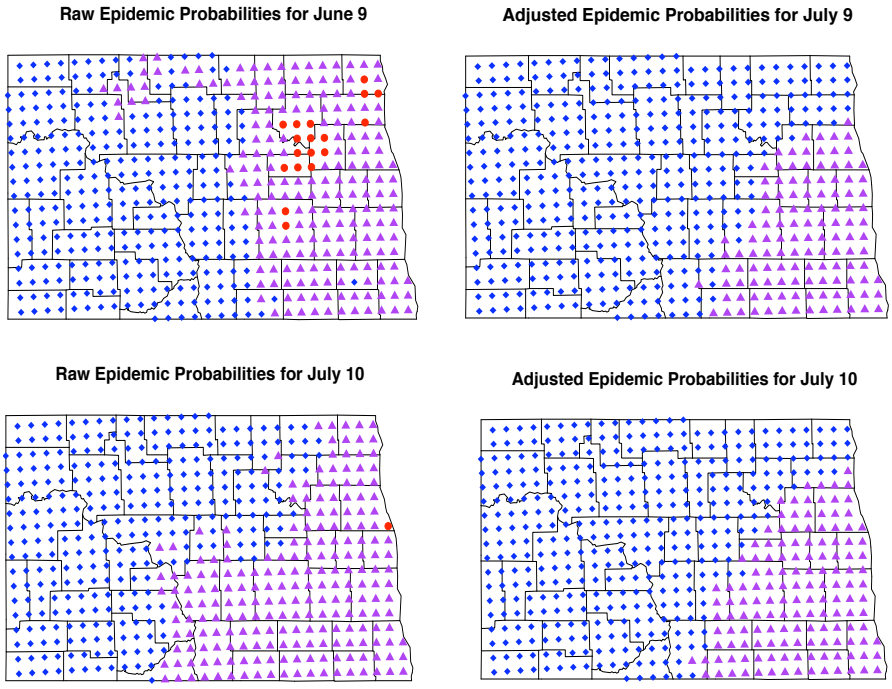

Figure 1: North Dakota raw and adjusted risk maps, July 9-10, 2005, ●=high risk,▲=medium risk, ♦=no risk), Left column:Raw risk maps, Right column: Adjusted risk maps

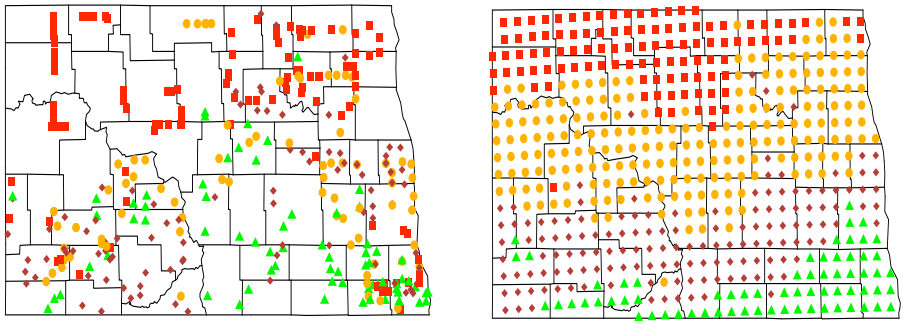

Figure 2: North Dakota flowering dates, 2005: ▲: <July 6, ♦: July 6 to July 12, ●:July 13 to July 18, ■: after July 20. Left: flowering survey data, Right: kriged map (means from posterior predictive distribution)

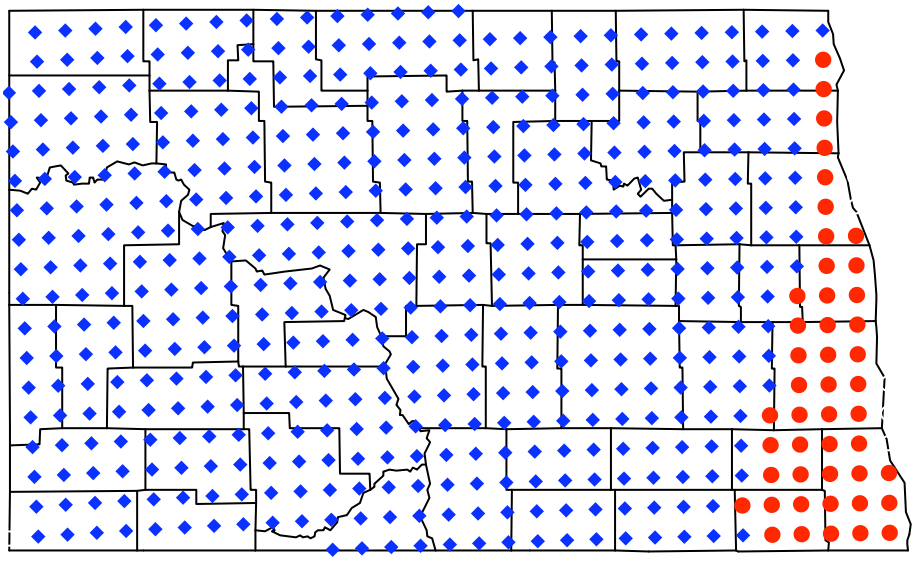

Figure 3: North Dakota elevated risk region ●=prolonged elevated risk, ♦=no prolonged elevated risk)

References

Banerjee, S. (2005). On geodetic distance computations in spatial modeling. *Biometrics*, 61(2):617–625.

Calder, C. A., Holloman, C. H., and Higdon, D. M. (2002). Exploring space time structure in ozone concentration using a dynamic process convolution model. In *Case Studies in Bayesian Statistics Volume VI*, pages 165–177. Springer-Verlag Inc.

Champeil, A., Doré, T., and Fourbet, J. (2004). Fusarium head blight: epidemiological origin of the effects of cultural practices on head blight attacks and the production of mycotoxins by Fusarium in wheat grains. *Plant Science*, 166(6):1389–1415.

De Wolf, E., Molineros, J., Wei, C., Lipps, P., Madden, L., and FRANCL, L. (2003). Development and deployment of the next generation prediction models for Fusarium head blight. *National Fusarium Head Blight Forum*.

Dill-Macky, R. and Jones, R. (2000). The Effect of Previous Crop Residues and Tillage on Fusarium Head Blight of Wheat. *Plant Disease*, 84(1):71–76.

Dufault, N., De Wolf, E., Lipps, P., and Madden, L. (2006). Role of Temperature and Moisture in the Production and Maturation of Gibberella zeae Perithecia. *Plant Disease*, 90(5):637–644.

Fernando, W., Miller, J., Seaman, W., Seifert, K., and Paulitz, T. (2000). Daily and seasonal dynamics of airborne spores of Fusarium graminearum and other Fusarium species sampled over wheat plots. *Can. J. Bot*, 78(4):497–505.

Finley, A., Banerjee, S., and Carlin, B. (2007). spBayes:an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19:1–20.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman and Hall.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean (Disc: p191-192). *Environmental and Ecological Statistics*, 5:173–190.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A., editors, *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, pages 761–768. Clarendon Press [Oxford University Press].

Hooker, D., Schaafsma, A., and Tamburic-Ilincic, L. (2002). Using Weather Variables Pre- and Post-heading to Predict Deoxynivalenol Content in Winter Wheat. *Plant Disease*, 86(6):611–619.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40.

Matèrn, B. (1986). Spatial variation. *Lecture Notes in Statistics, Number 36. Springer Verlag, New York (2nd edition, first edition published in 1960)*.

McMullen, M., Jones, R., and Gallenberg, D. (1997). Scab of wheat and barley: A re-emerging disease of devastating impact. *Plant Disease*, 81(12):1340–1348.

Molineros, J., De Wolf, E., Madden, L., , and Paul, P. (2006). Incorporation of variety resistance to spring wheat Fusarium Head Blight modeling. Technical report, Fargo, ND, North Central Division: American Phytopathological Society Meeting.

Molineros, J. E. (2007). Understanding the challenges of Fusarium Head Blight forecasting, Ph.D. dissertation. Technical report, Department of Plant Pathology, Pennsylvania State University.

Nganje, W., Bangsund, D., Leistritz, F., Wilson, W., and Tiapo, N. (2004). Regional Economic Impacts of Fusarium Head Blight in Wheat and Barley. *Review of Agricultural Economics*, 26(3):332–347.

Parry, D., Jenkinson, P., and McLeod, L. (1995). Fusarium ear blight(scab) in small grain cereals: a review. *Plant pathology*, 44(2):207–238.

Rossi, V., Giousue, S., Pattori, E., Spanna, F., and Del Vechio, A. (2003). A model estimating the risk of Fusarium head blight on wheat. *EPPO/OEPP Bulletin*, 33(3):421–425.

Schabenberger, O. and Gotway, C. (2005). *Statistical Methods For Spatial Data Analysis*. CRC Press.

Schmale, D., Shah, D., and Bergstrom, G. (2005). Spatial Patterns of Viable Spore Deposition of Gibberella zeae in Wheat Fields. *Phytopathology*, 95(5):472–479.

Short, M. B., Higdon, D. M., and Kronberg, P. P. (2007). Estimation of Faraday Rotation Measures of the Near Galactic Sky Using Gaussian Process Models. *Bayesian Analysis*, 2(4):665–680.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag Inc.

Sutton, J. (1982). Epidemiology of wheat head blight and maize ear rot caused by Fusarium graminearum. *Can. J. Plant Pathol*, 4:195–209.