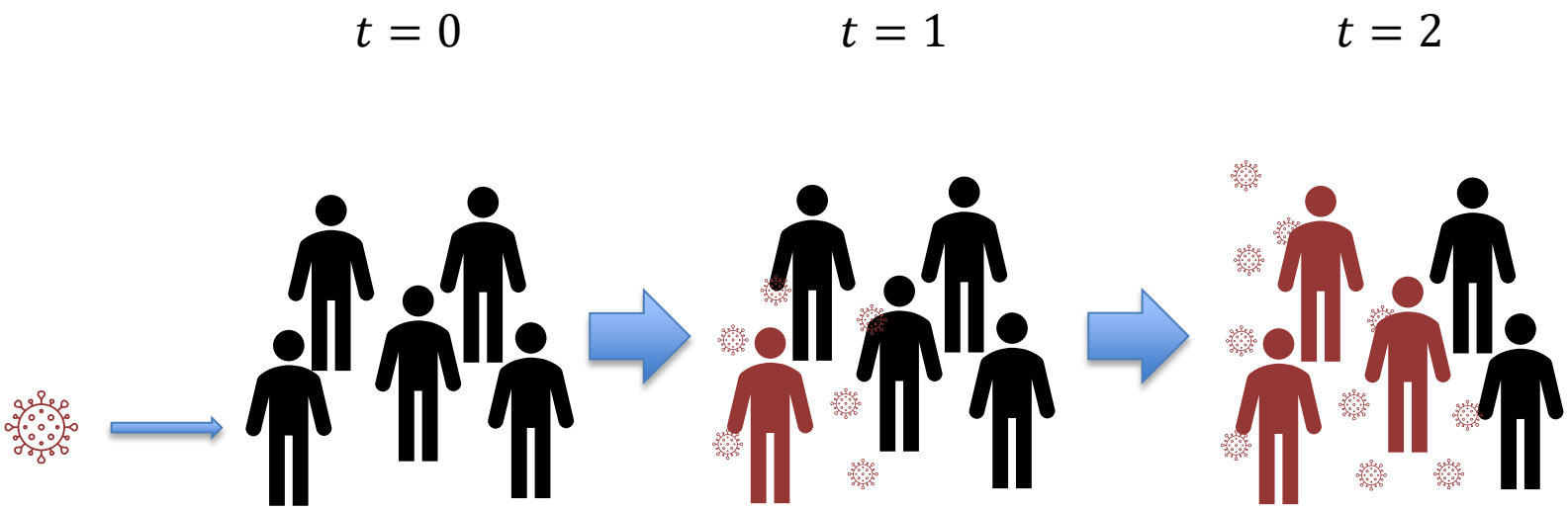
Predicting the Risk of Novel Pathogen Introductions from Disease Surveillance Data

Nelson Walker, Trevor Hefley, Daniel Walsh, Ian McGahan, Daniel Storm, Daniel Skinner

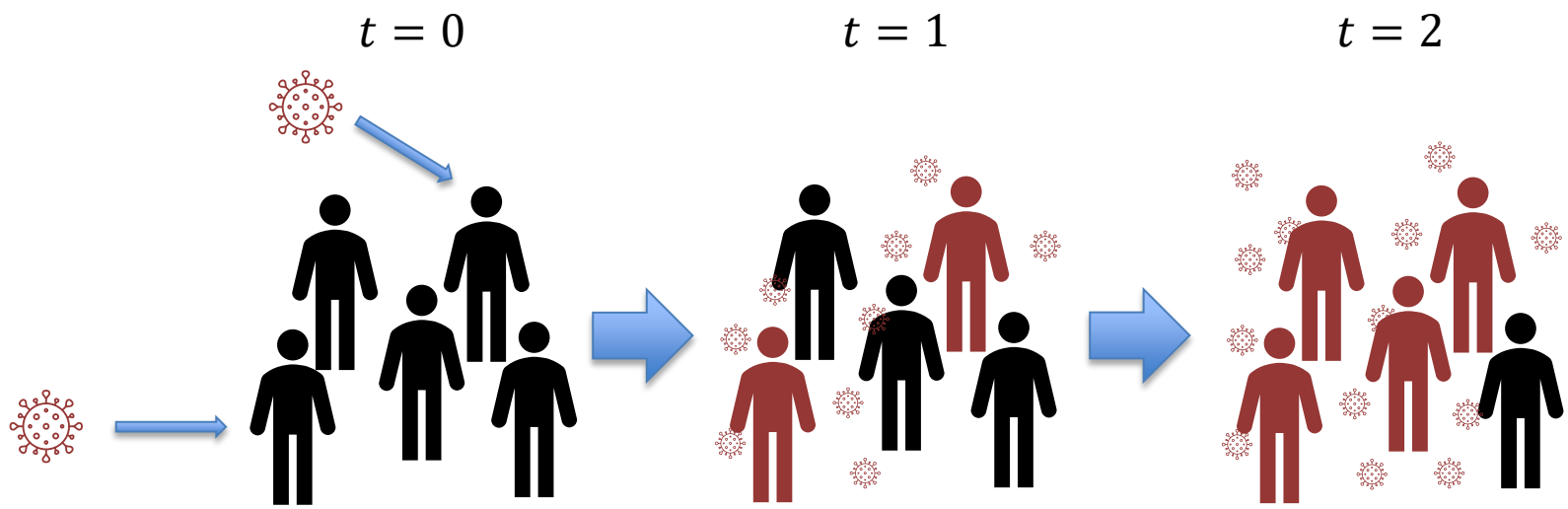
DISTRIBUTION STATEMENT A - Approved for public release; distribution is unlimited. 412TW-PA-22293

The Spread of Infectious Disease



- Could we identify where and when the pathogen was introduced?

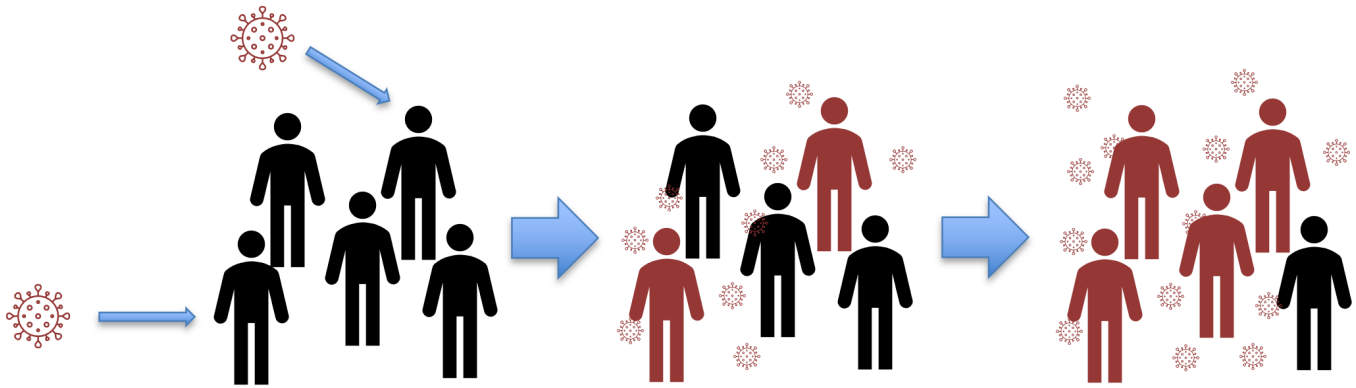
The Spread of Infectious Disease



- Could we identify where and when the pathogens were introduced?

What Statisticians Typically Do

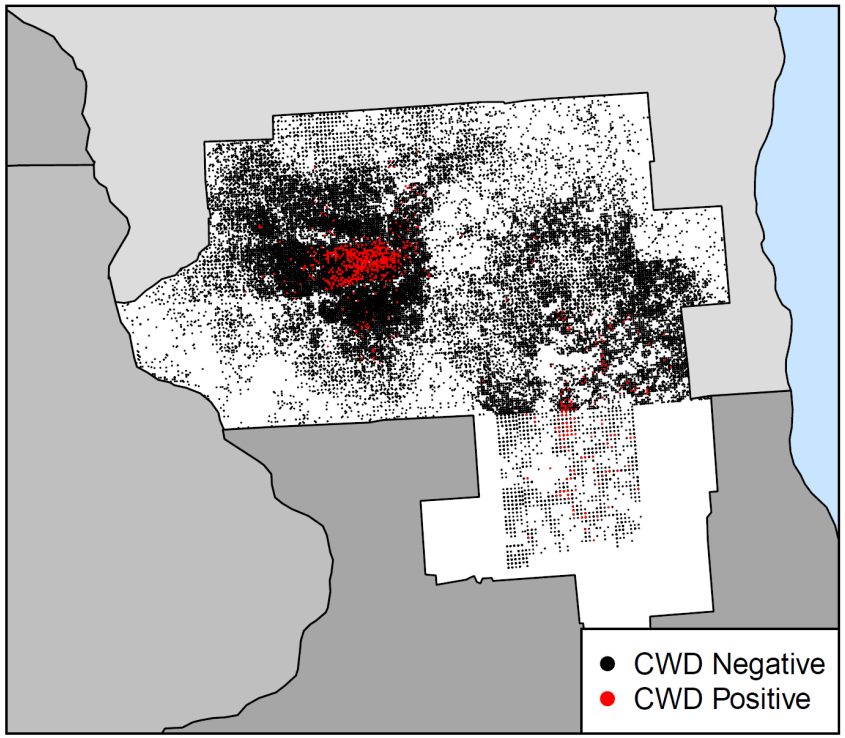
- Statisticians typically focus on estimating, predicting, and forecasting the spread of a pathogen.
  - Ordinary differential equations (Vynnycky and White, 2010).
  - Partial differential equations (PDEs; e.g., Hooten et al., 2013).
- The problem of estimating 'how many,' 'where,' and 'when' pathogen introductions occurred is typically ignored (Hefley et al., 2020).



- Vynnycky, E. and White, R. G. (2010). *An Introduction to Infectious Disease Modeling*. New York: Oxford University Press.
- Hooten, M. B., Garlick, M. J., and Powell, J. A. (2013). Computationally efficient statistical differential equation modeling using homogenization. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 405–428.
- Hefley, T.J., Russell, R.E., Ballmann, A.E., and Zhang, H. (2020). When and where: estimating the date and location of introduction for exotic pests and pathogens. Pre-print: https://arxiv.org/abs/2006.16982.

Example: Chronic Wasting Disease (CWD)

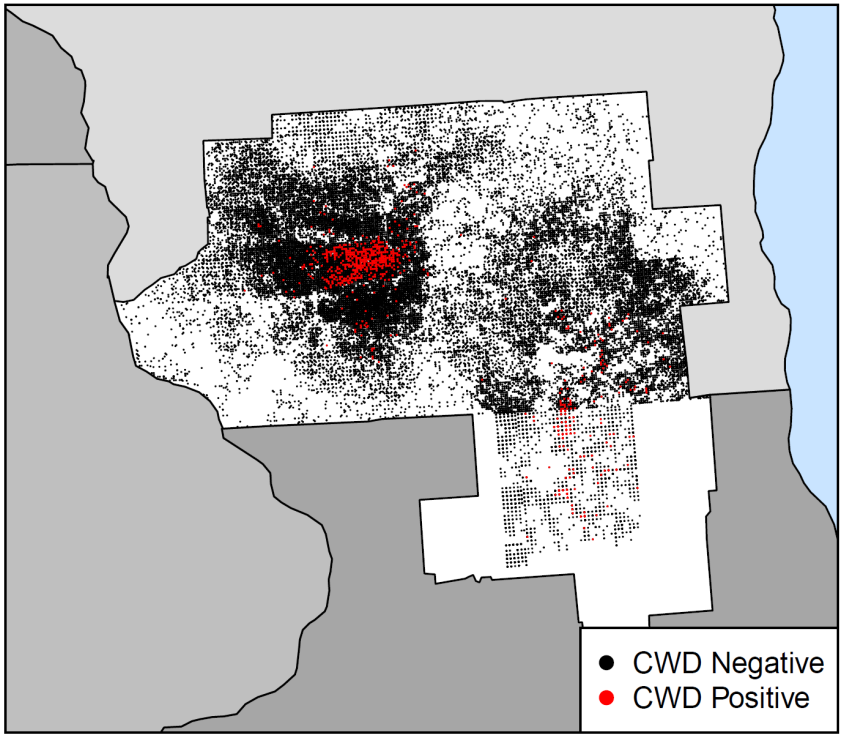
By Terry Kreeger, Wyoming Game and Fish and Chronic Wasting Disease Alliance. - https://www.usgs.gov/media/images/deer-chronic-wasting-disease, Public Domain, https://commons.wikimedia.org/w/index.php?curid=68352483



Our Questions

Can we estimate, predict, and forecast pathogen spread AND:

1. Estimate where, when, and how many pathogen introductions occurred?
2. Identify spatial covariates that are associated with locations of pathogen introduction?
3. Extrapolate and predict where pathogen introductions may occur in the broader region of Iowa, Minnesota, and Michigan, Illinois, and Wisconsin?



Several Components Required

- Mathematical model – a mixture of partial differential equations to represent the spread of the pathogen from numerous locations.
- Statistical model – link the latent pathogen intensity with spatio-temporal disease surveillance data.
  - Problem: Trans-dimensional estimation.
    - The number of pathogen introductions is a positive discrete random variable.
  - Problem: Find a principled way to obtain inference on where pathogen introductions occur.

Mathematical Modeling

- Ecological diffusion with exponential growth process expressed as a partial differential equation (PDE):

$$\frac{\partial}{\partial t} u_j(\mathbf{s}, t) = \left(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2} \right) \left[\mu(\mathbf{s}) u_j(\mathbf{s}, t) \right] + \lambda(\mathbf{s}) u_j(\mathbf{s}, t).$$

- A sum of ecological diffusion PDEs:

$$u(\mathbf{s}, t) = \sum_{j=1}^{J} u_j(\mathbf{s}, t)\,.$$

- Define a bivariate Gaussian kernel as the initial conditions for the jth PDE:

$$c_j\left(\mathbf{s}, t_{0j}\right) = \frac{\theta_j}{\sqrt{4\pi^2 \sigma_1^2 \sigma_2^2}} e^{\frac{-||s_{1,j} - \omega_{1,j}||^2}{2\sigma_1^2}} e^{\frac{-||s_{2,j} - \omega_{2,j}||^2}{2\sigma_2^2}}.$$

- Williams, P.J., M.B. Hooten, J.N. Womble, G.G. Esslinger, M.R. Bower, and T.J. Hefley. (2017). An integrated data model to estimate spatio-temporal occupancy, abundance, and colonization dynamics. *Ecology*, 98, 328-336.
- Hefley, T.J., Russell, R.E., Ballmann, A.E., and Zhang, H. (2020). When and where: estimating the date and location of introduction for exotic pests and pathogens. Pre-print: https://arxiv.org/abs/2006.16982.

Enter the Applied Mathematician

- Hooten et al. (2013) used a homogenization technique to more easily numerically solve a PDE.
  - Followed by Hefley et al. (2017), Williams et al. (2017) and others.
- The numerical method was inadequate for working with a mixture of PDEs.
- Ian McGahan:
  - Applied Mathematician Post-doc.
  - University of Wisconsin, Department of Statistics.
- Provided an approximate analytical solution to the PDE.

$$u_j(\mathbf{s},t) = \frac{\theta_j}{\mu(\mathbf{s})2\pi(\sigma^2 + 2\bar{\mu}(\mathbf{s})(t - t_{0j}))} e^{\bar{\lambda}(\mathbf{s})(\mathbf{s}-t_{0j})} e^{\frac{-||\mathbf{s}-\boldsymbol{\omega}_j||^2}{2(\sigma^2+2\bar{\mu}(\mathbf{s})(t-t_{0j}))}}$$

- Hooten, M. B., Garlick, M. J., and Powell, J. A. (2013). Computationally efficient statistical differential equation modeling using homogenization. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 405–428.
- Williams, P.J., M.B. Hooten, J.N. Womble, G.G. Esslinger, M.R. Bower, and T.J. Hefley. (2017). An integrated data model to estimate spatio-temporal occupancy, abundance, and colonization dynamics. *Ecology*, 98, 328-336.
- Hefley, T. J., Hooten, M. B., Russell, R. E.,Walsh, D. P., and Powell, J. A. (2017b). When mechanism matters: Bayesian forecasting using models of ecological diffusion. Ecology Letters, 20, 640–650.

Statistical Model: Bayesian Hierarchical Mixture of Ecological Diffusion Models (BHMEDM)

$$y_i \sim \text{Bern}(p(\mathbf{s}_i, t_i))$$

$$g(p(\mathbf{s}_i, t_i)) = u(\mathbf{s}_i, t_i) e^{\mathbf{b}_i' \boldsymbol{\tau}}$$

$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J} u_j(\mathbf{s}_i, t_i)$$

Question: This looks simple enough. How do we obtain inference on the spatial process for the pathogen introductions and predict new introductions?

Parameter Model and Prior Distributions

-Model component for the PDE initial condition parameters:

$$p(\boldsymbol{\Omega}, \mathbf{t}_0, J | \beta_0, \boldsymbol{\beta}) \equiv p\left(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_J, t_{01}, t_{02}, \dots, t_{0J}, J \middle| \beta_0, \boldsymbol{\beta}\right) = \mathrm{IPPP}\big(\Lambda(\mathbf{s}, t)\big),$$

$$\log\big(\Lambda(\mathbf{s}, t)\big) = \beta_0 + \boldsymbol{x}(\mathbf{s})'\boldsymbol{\beta},$$

$$p((\log(\theta), \boldsymbol{\tau})') = MVN\left(\begin{pmatrix} \mu_\theta \\ \boldsymbol{\mu}_{\boldsymbol{\tau}} \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \boldsymbol{\rho}'_{\tau,\theta} \\ \boldsymbol{\rho}_{\theta,\tau} & \boldsymbol{\Sigma}_{\boldsymbol{\tau}} \end{pmatrix}\right).$$

-Priors for other parameters are uninteresting.

Fitting the BHMEDM: Problems

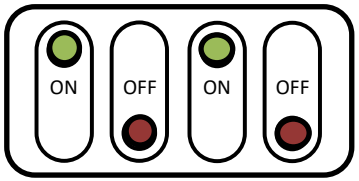
- J is unknown, positive discrete parameter.
  - Trans-dimensional estimation problem.
    - J+1 results in three more parameters to estimate $(t_{0,J+1}, \omega_{1,J+1}, \omega_{2,J+1})'$.

- Tuning the proposal distribution for $(\mathbf{\Omega}, \mathbf{t}_0, J)'$ is a major, if not impossible challenge.

Solutions

- Over parameterize BHMEDM with an arbitrarily large number of PDEs (Pseudo-introductions) where the time and locations of introduction were "picked."

$$u(s_i, t_i) = \sum_{j=1}^{J^*} v_j u_j(s_i, t_i)$$



- Posterior inference on v_j's leads to inference on the number, locations, and times of introduction $(\mathbf{\Omega}^*, \mathbf{t}_0^*, J^*)'$.

- Rousseau, J. and Mengersen, K. (2018). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* 73, 689–710.

Bayesian Imputation

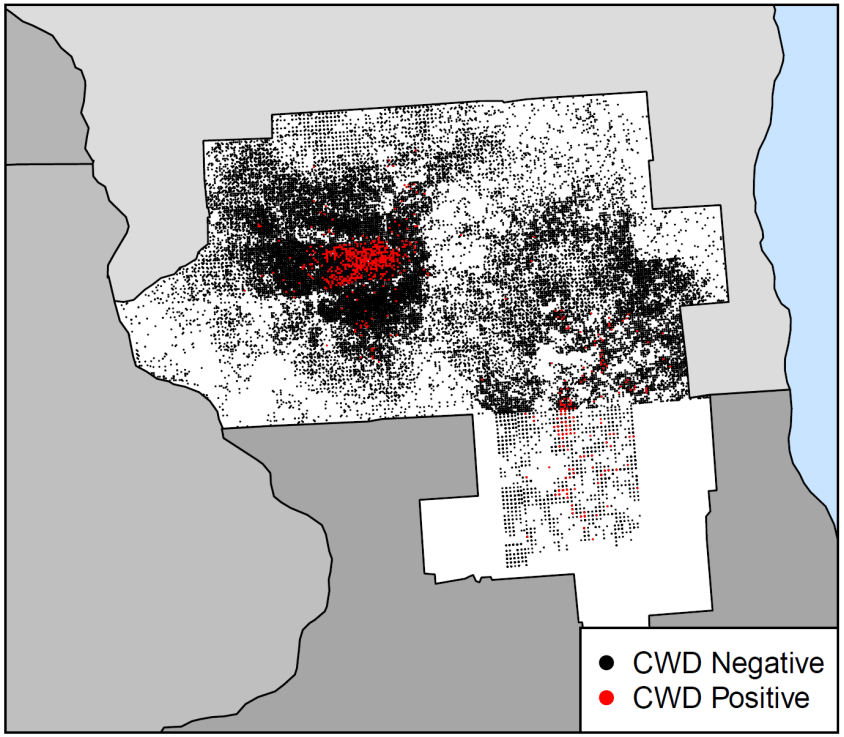
- Problem: If pseudo-introductions are mis-specified, the model is mis-specified.
  - The degree that the pseudo-introductions were mis-specified affects the performance of the BHMEDM.

- My approach: Employ Bayesian imputation by using draws from selected pseudo-introductions to obtain $p(\beta_0, \boldsymbol{\beta} \mid \boldsymbol{\Omega}^*, \mathbf{t}_0^*, J)$ (Scharf, Hooten, and Johnson, 2017).

Scharf, H.R., M.B. Hooten, and D.S. Johnson. (2017). Imputation approaches for animal movement modeling. *Journal of Agricultural, Biological, and Environmental Statistics*, 22, 335-352.

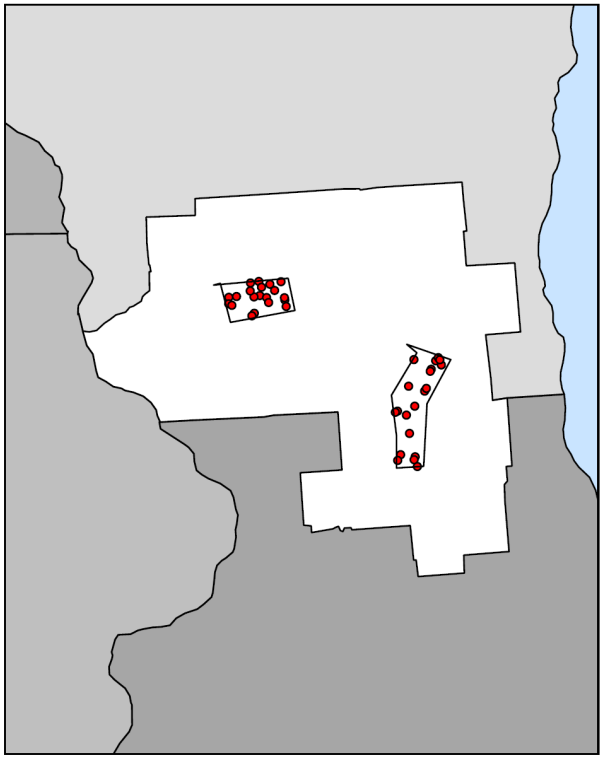
CWD Surveillance Data

- Disease surveillance data for CWD in white-tailed deer.
- Study area: southern Wisconsin and northern Illinois.
- Collected from 2001-2006.
- Estimate the number, locations, and times of pathogen introduction.
- Obtain inference on $\Lambda(\mathbf{s}, t)$ within the study area.
- Extrapolate $\Lambda(\mathbf{s}, t)$ to Iowa, Illinois, Michigan, Minnesota, and Wisconsin and predict where other introductions may occur.



Define Pseudo-Introductions

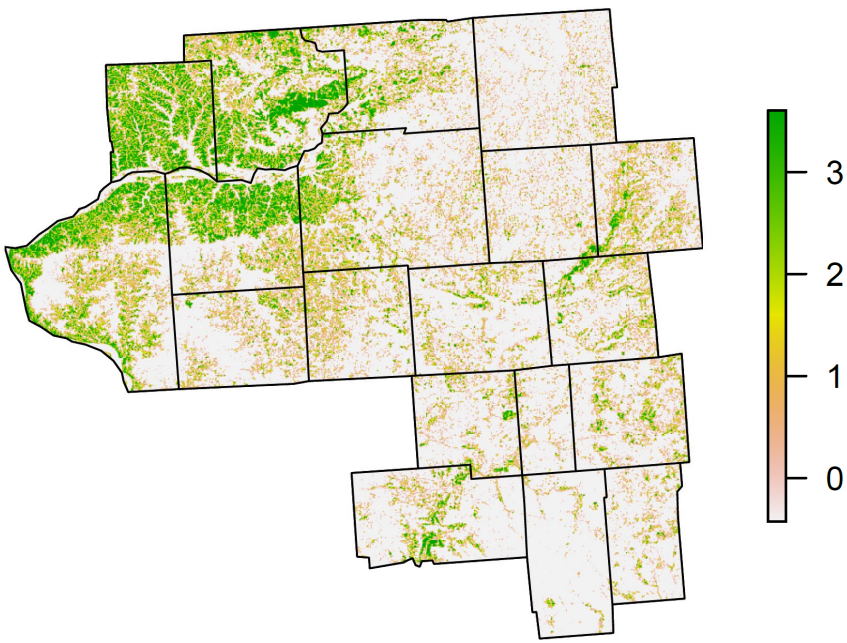
- Define J^*=40.
  - Computationally feasible.
- Define two regions within the study area that contain the majority of CWD cases.
- Draw twenty pseudo-introductions from each region in the study area using a bivariate uniform distribution.
- Draw times of introduction for pseudo-introductions as $\mathbf{t}_0^* \equiv (t_{01}^*, t_{02}^*, \dots, t_{0J^*}^*)' \sim \text{Unif}(1994, 2004)$.



Covariates for BHMEDM

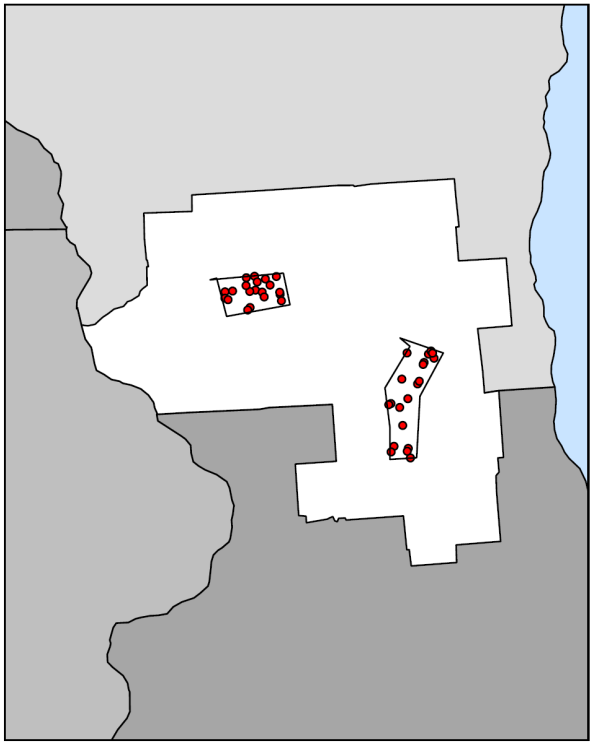
- Susceptibility parameters and covariate:
  - $(\log(\theta), \tau_{sex})'$
  - $b_{sex,i} = \begin{cases} 1 \; if \; sex = male \\ 0 \; if \; sex = female \end{cases}$
- Diffusion covariate:
  - Centered and scaled forest density (*).
  - Centered and scaled development density.
  - East vs. west indicator.
- Growth covariates:
  - Centered and scaled soil clay content.
  - Centered and scaled soil cation exchange capacity (*).
  - Centered and scaled soil organic carbon content.
  - Centered and scaled forest density (*).
  - Centered and scaled development density.
  - East vs. west indicator.
- Priors for $(\log(\theta), \tau)', \boldsymbol{\alpha}, \boldsymbol{\gamma}$ are all multivariate Gaussian.

Forest Density (CS)

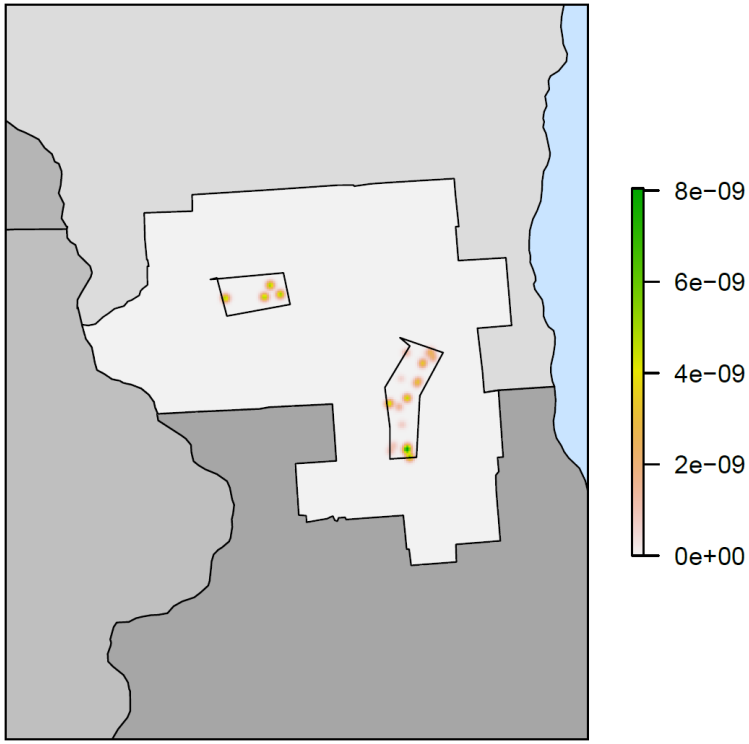


Rivera, N. A., Brandt, A. L., Novakofski, J. E., and Mateus-Pinilla, N. E. (2019). Chronic wasting disease in cervids: Prevalence, impact and management strategies. Veterinary Medicine: Research and Reports, 10:123-139.

Pre-Specified Pseudo-Introductions

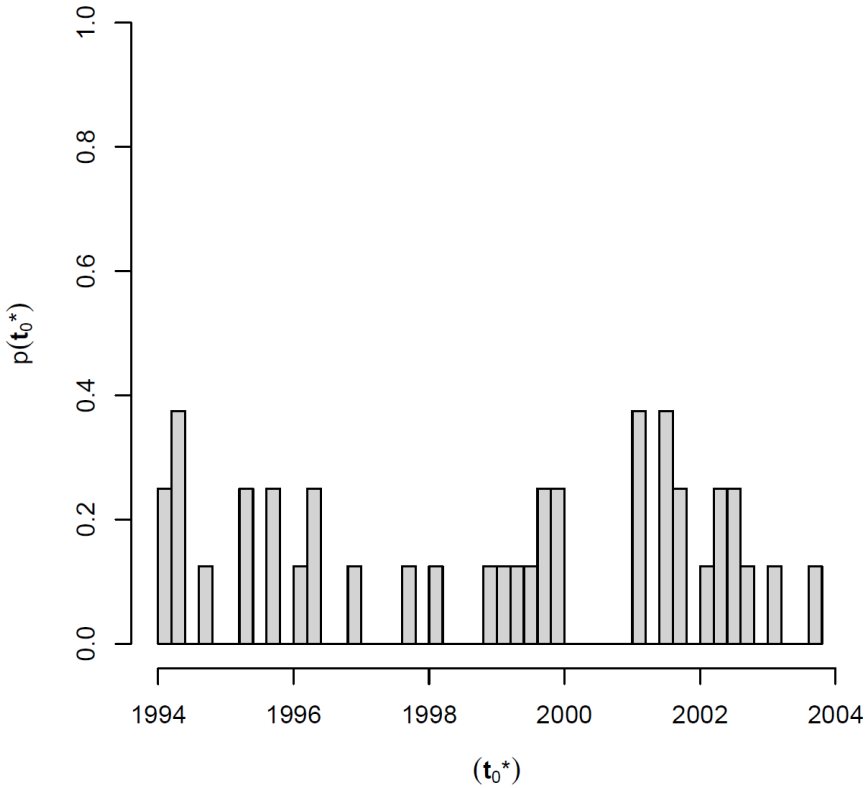


Selected Pseudo-Introductions

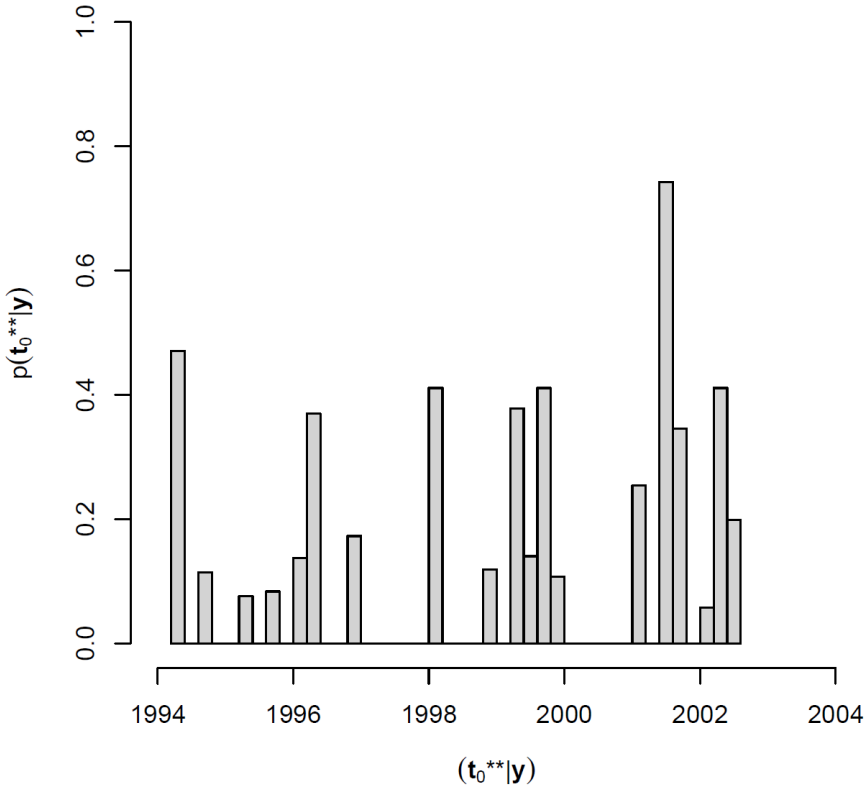


Parameter	Mean	0.025 Quantile	0.975 Quantile
$J^{**} = \sum v_j^k$	12.1	10	15

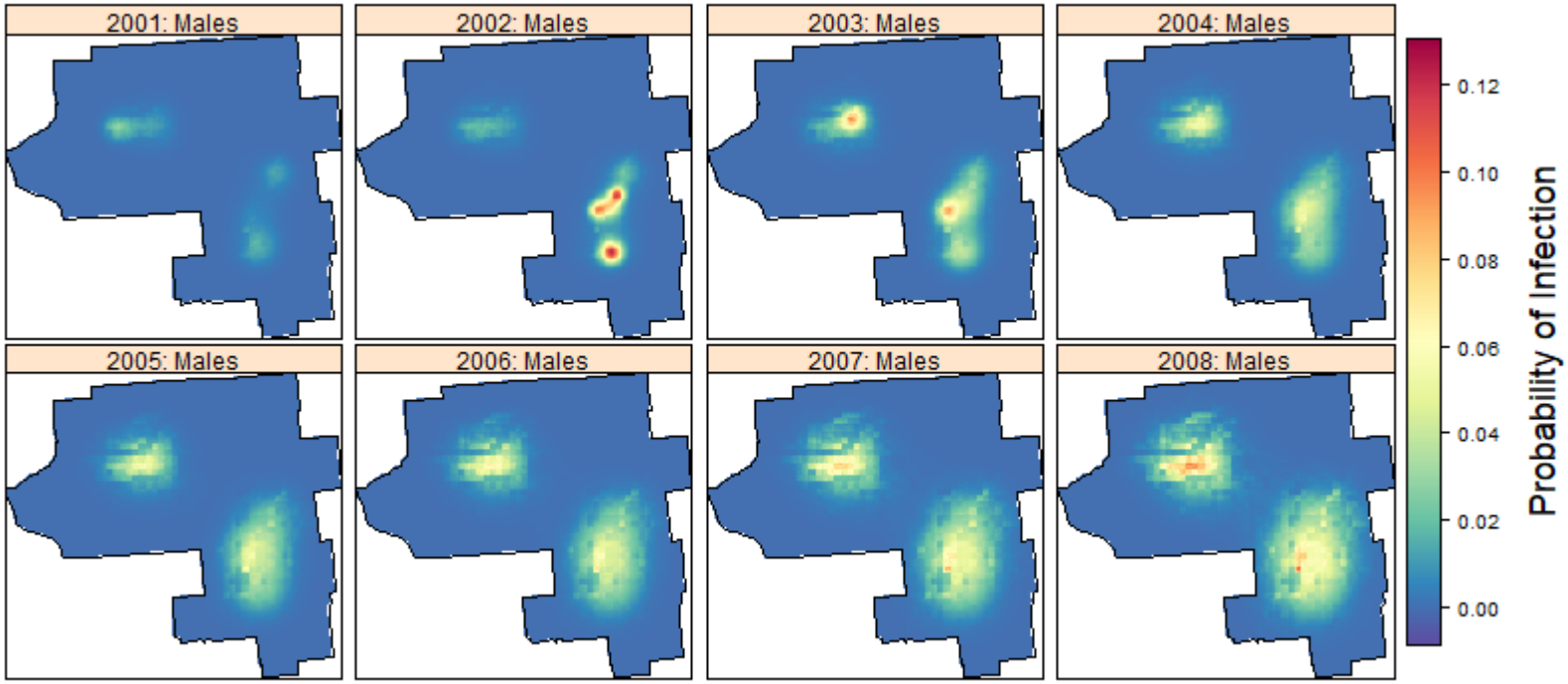
Pre-Specified Pseudo-Introduction Times



Selected Pseudo-Introduction Times



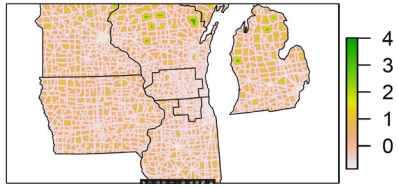
Predicted and Forecasted Probabilities



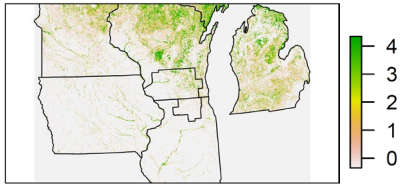
Bayesian Imputation with the IPPP

- Centered and scaled covariates:
  - $x_{\text{highway}}(\mathbf{s})$ - distance to nearest highway
  - $x_{\text{water}}(\mathbf{s})$ - water density
  - $x_{\text{crops}}(\mathbf{s})$ - crop density
  - $x_{\text{forest}}(\mathbf{s})$ - forest density
- Priors
  - $(\beta_0, \beta_{highway}, \beta_{water}, \beta_{crops}, \beta_{forest})' \sim MVN(\mathbf{0}, 10^6 \boldsymbol{I})$

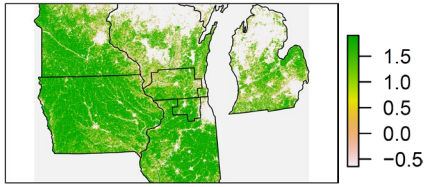
Distance to Nearest Highway



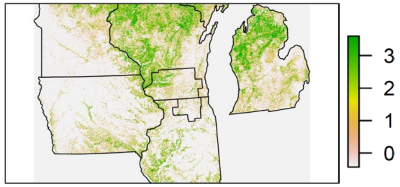
Water Density



Crop Density

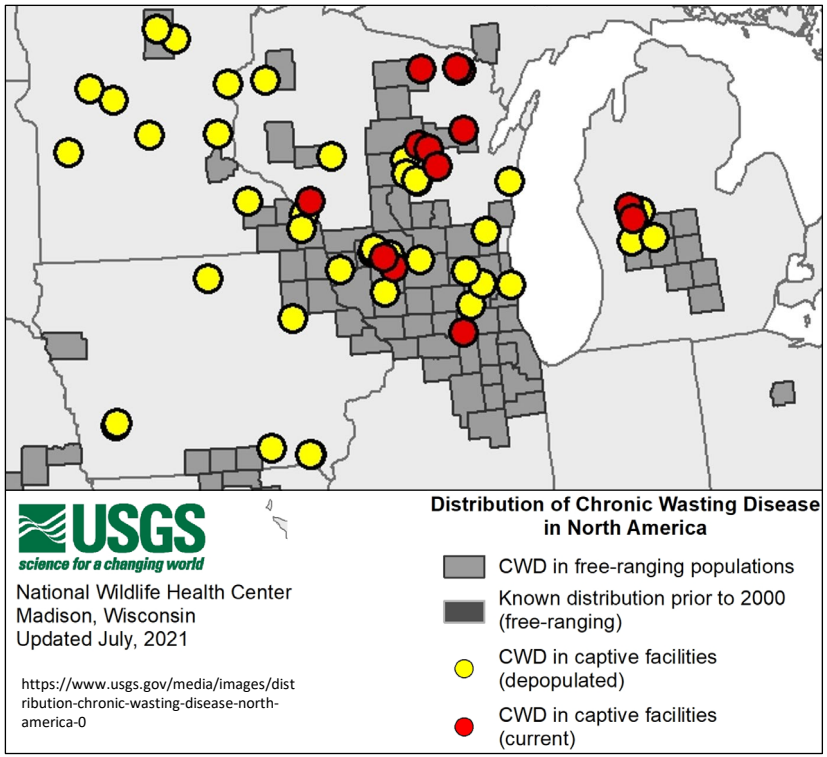


Forest Density

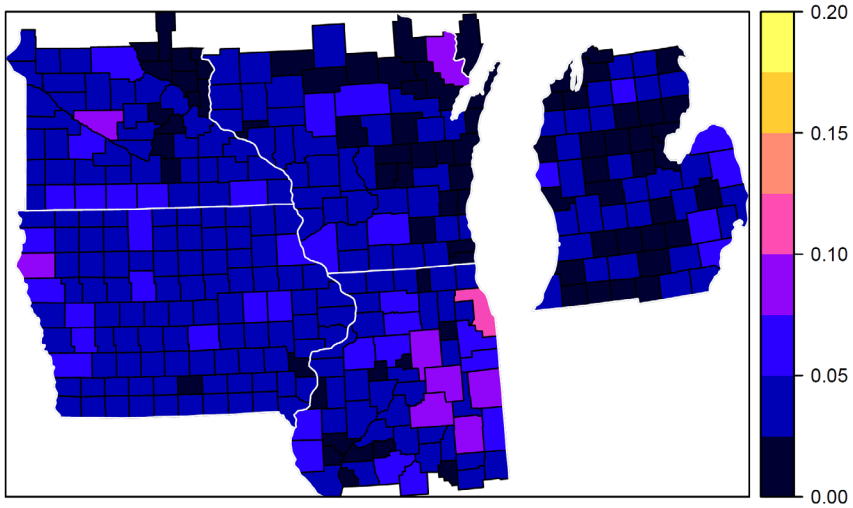


Results

Counties with CWD Present
(as of July 2021)



Expected Number of Introductions by County (2004-2013)



Discussion

- This is a very challenging and relevant problem in light of COVID-19.
  - Few researchers have attempted to develop solutions.
- We obtained inference on the number, locations, and times of pathogen introduction.
- We provide a form of inference on the spatio-temporal process for pathogen introductions that was impossible before.
- Framework is generalizable for other types of disease surveillance data besides binary data.
- If applied to COVID-19 surveillance data, it may be possible to better understand the impact of novel COVID-19 introductions in a region vs. community spread.
- Future work might account for temporal variability in the intensity function, $\Lambda(\mathbf{s}, t)$.

Questions

Choosing Pseudo-Introduction Times

- First CWD cases recorded in WI in 2001 and IL in 2002.
- Evidence suggests that CWD was first introduced sometime before 2001.
- Additional case hot-spots suggest multiple introductions. Some before 2001, some perhaps after.
- Drawing pseudo-introduction times from Uniform distribution is convenient for spatio-temporal extrapolation to a future time interval.
- Uniform(1994, 2004) seemed a reasonable exploratory distribution for pseudo-introductions.
  - Not too early, not too late.

- Chickering, D. M. and Heckerman, D. (1997). Efficient approximations of the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*. 29, 181-121.
- Goh, G. Spring 2020. Stat 950 Lecture Notes, Kansas State University.
- Rousseau, J. and Mengersen, K. (2018). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*. 73, 689-710.

More on Priors to Empty the BHMEDM of Superfluous Pseudo-introductions

- Rousseau and Mengersen (2018) focused on overfitting traditional mixture models (i.e., $J^* > J$):

$$y_i \sim \text{Bern}\big(p(\mathbf{s}_i, t_i)\big),$$
$$g\big(p(\mathbf{s}_i, t_i)\big) = u(\mathbf{s}_i, t_i),$$
$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J^*} \eta_j u_j(\mathbf{s}_i, t_i),$$

  where η_j is a non-negative weight (probability), such that $\sum_{j=1}^{J^*} \eta_j = 1$.

- J is the true number of introductions.
- The following prior will "empty" the over-parameterized model of superfluous pseudo-introductions as $n \to \infty$:

$$p(\boldsymbol{\eta}) = \text{Dirichlet}\left(\frac{1}{2}\right).$$

  - This is also the Jeffreys prior for a multinomial distribution.

- Less informative priors that favor weights close to zero will empty the model of superfluous pseudo-introductions.

- Rousseau, J. and Mengersen, K. (2018). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73, 689–710.
- Frühwirth-Schnatter, S.; Celeux, G. & Robert, C. P. (2020). *Handbook of Mixture Analysis*. Chapman and Hall/CRC, Florida.

More on Priors to Empty the BHMEDM of Superfluous Pseudo-introductions

- We apply this principle to a related situation. Since this:

$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J^*} \eta_j u_j(\mathbf{s}_i, t_i),$$

where η_j is a non-negative weight (probability), such that $\sum_{j=1}^{J^*} \eta_j = 1$

- Is similar to this:

$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J^*} v_j u_j(\mathbf{s}_i, t_i)\,,$$

where

$$v_j \sim \mathrm{Bern}(\tilde{q}_j)$$

and

$$\tilde{q}_j = \frac{q_j\, p(\mathbf{y}|\ldots, v_j=1)}{q_j\, p(\mathbf{y}|\ldots, v_j=1) + (1-q_j) p(\mathbf{y}|\ldots, v_j=0)},$$

and

$$q_j = P(v_j = 1) \sim \mathrm{Beta}(0.5, 0.5).$$

- Rousseau, J. and Mengersen, K. (2018). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73, 689–710.
- Frühwirth-Schnatter, S.; Celeux, G. & Robert, C. P. (2020). *Handbook of Mixture Analysis*. Chapman and Hall/CRC, Florida.

Applying BMA to Multiple BHMEDM Fits

- Recall that BMA provides a way to address uncertainty associated with model selection.
- Let $\varDelta$ be a quantity of interest.
- The posterior distribution of $\varDelta$ given data $\mathbf{y}$ is defined as:

$$p(\varDelta|\mathbf{y}) = \sum_{k=1}^{K} p(\varDelta|M_k, \mathbf{y}) p(M_k|\mathbf{y}),$$

where $M_1, M_2, \ldots, M_K$ are the candidate models and:

$$\mathrm{E}(\varDelta|\mathbf{y}) = \sum_{k=1}^{K} \widehat{\varDelta}_k p(M_k|\mathbf{y}),$$

$$\mathrm{Var}(\varDelta|\mathbf{y}) = \sum_{k=1}^{K} \left\{\mathrm{Var}(\varDelta|\mathbf{y}, M_k) + \widehat{\varDelta}_k^2\right\} p(M_k|\mathbf{y}) - \{\mathrm{E}(\varDelta|\mathbf{y})\}^2,$$

where $\widehat{\varDelta}_k = E(\varDelta|\mathbf{y}, M_k)$.

- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian modeling averaging: A tutorial. *Statistical Science*, 14(4): 382-401.
- Goh, G. Spring 2020. Stat 950 Lecture Notes, Kansas State University.

Applying BMA to Multiple BHMEDM Fits

- The posterior probability for model M_k is given by:

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{\sum_{k=1}^{K} p(\mathbf{y}|M_k)p(M_k)},$$

where the marginal likelihood of model M_k is:

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|\boldsymbol{\Theta}_k, M_k)p(\boldsymbol{\Theta}_k|M_k)d\boldsymbol{\Theta}_k\,,$$

  - $\boldsymbol{\Theta}_k$ is the vector of parameters of model M_k,
  - $p(\boldsymbol{\Theta}_k|M_k)$ is the prior density of Θ_k under M_k,
  - $p(\mathbf{y}|\boldsymbol{\Theta}_k, M_k)$ is the likelihood, and
  - $p(M_k)$ is the prior probability that M_k is the true model.
- The marginal likelihood may be calculated using Monte Carlo integration or another method.
- Remember that in the "introductions" are treated like missing data. I propose that for now they are ignored. BMA may be applied to any parameter in the BHMEDM, but I propose it can also include IPPP parameters.

- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian modeling averaging: A tutorial. *Statistical Science*, 14(4): 382-401.
- Goh, G. Spring 2020. Stat 950 Lecture Notes, Kansas State University.

Evaluating MCMC Convergence

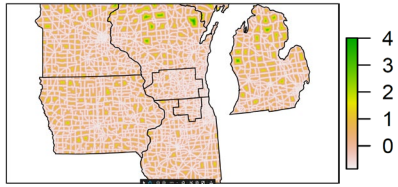
- What did I do? Check for stationarity for each parameter within a single chain.
- Three important aspects for evaluating MCMC convergence:
  1. Use multiple MCMC chains with different starting values.
  2. Ensure efficient simulation.
  3. Compare variation between and within chains. Between and within variation should become roughly equal.
     - Evaluating stationarity within chains and common distribution between chains.
     - Employ the Gelman-Rubin convergence diagnostic.

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. CRC Press, Boca Raton, Florida.

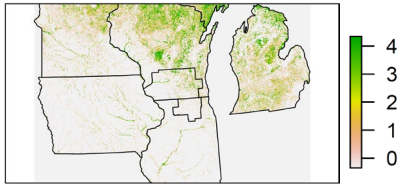
Bayesian Imputation with the IPPP

- Centered and scaled covariates:
  - $x_{\text{highway}}(\mathbf{s})$ - distance to nearest highway
  - $x_{\text{water}}(\mathbf{s})$ - water density
  - $x_{\text{crops}}(\mathbf{s})$ - crop density
  - $x_{\text{forest}}(\mathbf{s})$ - forest density
- Priors
  - $(\beta_0, \beta_{highway}, \beta_{water}, \beta_{crops}, \beta_{forest})' \sim MVN(0, 10^6 I)$

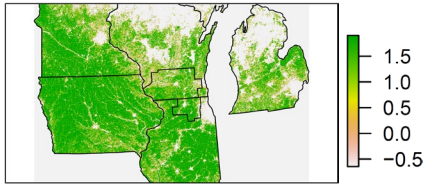
Distance to Nearest Highway



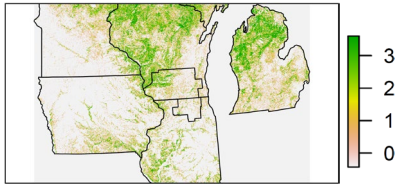
Water Density



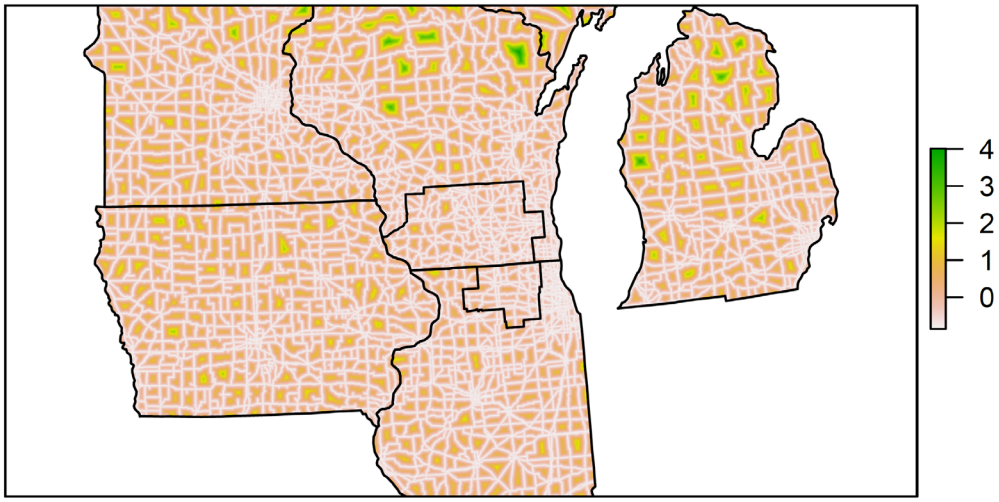
Crop Density



Forest Density



Bayesian Imputation with the IPPP

- Centered and scaled covariates:
  - Distance to nearest highway;
  - Water density;
  - Crop density;
  - Forest density.
- Priors
  - $MVN(\mathbf{0}, 10^6 \boldsymbol{I})$

Distance to Nearest Highway (Centered and scaled)