2016 SISMID Module 16
Lecture 9: Infectious Disease Data

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Why space-time modeling?

We will concentrate on infectious diseases in humans.

Infectious diseases are disorders caused by organisms, such as bacteria, viruses, fungi or parasites.

With comprehensive data (e.g. contact information for communicable diseases), spatial information would be (largely) unneeded.

But without such information, space is acting as a surrogate, or summary of sources of transmission that are geographically close to a susceptible.

Infectious vs. chronic diseases

Table 1 : Comparison of analysis and modeling issues for cancer and infectious diseases, with respect to space-time modeling.

	Cancer	Infectious
Incidence	Rare	Can be non-rare
Associations	Carcinogens	Transmission
Time scale	Long	Often short
Space-time interactions	Difficult to detect	Default
Age	Mostly increasing with age	Often in children
Objectives	Description	Description
	Etiology	Etiology
	Interventions (e.g. screening)	Interventions (e.g. vaccination)
		Understand dynamics

Overview

Keeling and Rohan (2008, Chapter 7) give an overview of spatial modeling.

For directly transmitted diseases, individuals have to be in the same geographical location, and spread will occur when individuals move in space.

The type of model used will depend on the host organism (human, animal, plant), what is known about the organism's behavior, and the geographical scale that is considered.

A big distinction concerns the form of the data we receive; do we see individuals, with point locations, or aggregated counts with respect to some administrative regions?

If infection can only be passed to a small number of individuals (as is the case for sexually transmitted diseases) then network models are advantageous.

Infectious Disease Data

The aims of infectious disease modeling include:

- understanding the mechanisms of spread,
- estimating the durations of the latent and infectious periods, and
- the size of the epidemic, often with the aim of determining strategies for disease control.

The modeling of infectious disease data has a huge literature, though relatively little has been written on spatial analysis; analyzing as a function of time is the norm.

Infectious Disease Data

A starting dichotomy is in terms of deterministic versus stochastic models.

References:

- The classic text on deterministic models is Anderson and May (1991).
- Books that consider both include Daley and Gani (1999) and Bailey (1975).
- Books on stochastic modeling include Becker (1989), Andersson and Britton (2000) and Halloran *et al.* (2010).

Deterministic Models

The law of mass action is central to deterministic modeling.

In a population context, if the individuals in a population mix homogenously, the rate of interaction between two different subsets of the population is proportional to the product of the numbers in each of the subsets concerned.

Groups of individuals defined by their disease status are described in continuous time (usually) via differential equations and there is no randomness — may be thought of as producing the mean of a random process.

Let X_t represent the number of susceptibles at time t, Y_t the number of infectives and Z_t the number of immune individuals with $X_t + Y_t + Z_t = N$.

Deterministic Models

Kermack and McKendrick (1927) proposed the following classic equations to describe the dynamics of the general epidemic (where we assume frequency dependent transmission):

$$\begin{aligned} \frac{dX}{dt} &= -\frac{\beta Y}{N} X, \\ \frac{dY}{dt} &= \frac{\beta Y}{N} X - \gamma Y, \\ \frac{dZ}{dt} &= \gamma Y, \end{aligned}$$

subject to initial conditions (X_0, Y_0, Z_0) with $Z_0 = 0$.

β is the infection parameter and γ the removal parameter.

Deterministic models can be embedded within a statistical framework for inference, or a stochastic approach can be taken from the onset.

A meta-population model

There is not a great deal of cross-referencing between the infectious disease and statistical research communities.

Meta-population models are discussed in depth in Keeling and Rohan (2008, Chapter 7).

The idea is to divide the population into distinct subpopulations each of which has its own dynamics, but which may depend on interactions between the subpopulations.

A meta-population model

In anticipation of spatial modeling we shall refer to the subpopulations as areas.

Recall the simple deterministic SIR model within a single area:

$$\begin{aligned} \frac{dX}{dt} &= -\lambda X \\ \frac{dY}{dt} &= \lambda X - \gamma Y. \end{aligned}$$

With frequency dependent transmission the force of infection $\lambda = \beta Y/N$[1].

[1] Under this form the rate of contact between a susceptible and a member of the population does not depend on the population size, β is the product of the contact rate and the per-contact probability of infection, given contact between an infective and a susceptible, Y/N is the prevalence, i.e. the probability that a random contact is with an infective

A meta-population model

We let X_i, Y_i represent the number of susceptibles and infectives in area i, $i = 1, \ldots, n$.

The deterministic equations for the n areas are:

$$\begin{aligned} \frac{dX_i}{dt} &= -\lambda_i X_i \\ \frac{dY_i}{dt} &= \lambda_i X_i - \gamma_i Y_i. \end{aligned}$$

One form for the force of infection in area i is

$$\begin{aligned} \lambda_i &= \beta_i \sum_j \rho_{ij} \frac{Y_j}{N_j} \\ &= \beta_i \frac{Y_i}{N_i} + \beta_i \sum_{j:j \neq i} \rho_{ij} \frac{Y_j}{N_j} \end{aligned} \tag{1}$$

so we have contributions from the area that we are modeling, and potentially all other areas.

A metapopulation model

With respect to the model

$$\lambda_i = \beta_i \sum_j \rho_{ij} \frac{Y_j}{N_j},$$

the ρ_{ij} parameters are a measure of the strength of interaction between areas i and j.

More precisely, ρ_{ij} measures the relative strength of transmission from area j to area i, so that $\rho_{ij} = 0$ means no direct transmission from area j to area i.

Recall that λ is the product of the rate of contact ($= c_0$ for frequency dependent), the probability a host is infectious Y/N and the probability of transmission p.

Hence, the above form shows that in this model transmission takes place in area j, since the probability a host is infectious is Y_j/N_j.

Transmission kernels

The strength of interaction is referred to the level of coupling between the areas.

In many cases the level of interaction between two areas will depend on the distance d between them, and this can be captured by a transmission kernel, K.

Common examples:

1. Exponential: $K \propto \exp(-Ad)$.
2. Gaussian: $K \propto \exp(-Ad^2)$.
3. Power law: $K \propto d^{-A}$.

For individual-based models (Keeling and Ross, 2007, p. 268) the rate of transmission to a susceptible individual i is

$$\lambda_i = \beta \sum_{j \in \text{ infectious}} K(d_{ij}),$$

where d_{ij} is the distance between susceptible individual i and infective j.

SIR Models

In the Susceptible-Infected-Removed (SIR) model the deterministic differential equations are replaced by probabilistic descriptions of the transitions.

Suppose we have x_t susceptibles and y_t infectives at time t.

Then consider Markov transitions with the probability of infection during a time interval $[t, t + \delta t]$ being

$$\beta x_t y_t + o(\delta t)$$

and the probability of removal being

$$\gamma y_t + o(\delta t).$$

In these statements, δt is a small time interval and $o(\delta t)$ is a small term.

Computation for compartmental models

To address the inference problem, various approaches have been suggested:

- For small populations, auxiliary variable approaches are tractable (Gibson and Renshaw, 1998; O'Neill and Roberts, 1999; O'Neill and Becker, 2001; Neal and Kypraios, 2015).
- Discrete approximations (Lekone and Finkenstädt, 2006).
- Diffusion process approximation (Cauchemez and Ferguson, 2008).
- Particle filtering, for likelihood or Bayesian inference (He *et al.*, 2010; Koepke *et al.*, 2015).
- Gaussian process approximate Bayesian inference (Jandarov *et al.*, 2014).
- Approximate Bayesian Computation (ABC) (McKinley *et al.*, 2009; Toni *et al.*, 2010; Neal, 2012).

The last three require simulation from the model, which is straightforward.

Disease mapping type models ignore the infectious aspect (Mugglin *et al.*, 2002; Knorr-Held and Richardson, 2003; Bauer *et al.*, 2015).

Forms of Data

We may have count or point data from a survey (at a single time point), with the response being infected/not infected. In this case the models we have used throughout the course may be applied.

Counts from a region may available at regular intervals (e.g. weekly) and these may have sub-regional information.

We might fit the Type IV space-time interaction disease mapping model to data of this type though this model has little direct connection with the underlying biology.

This lecture will concentrate on count data, see Riley (2007) for a review of four approaches to spatial modeling.

Motivating example: hand, foot and mouth disease

Hand, Foot and Mouth Disease (HFMD) is caused by an acute contagious viral infection and there have been large-scale outbreaks in Asia during the past 20 years (Tong and Bible, 2009).

Mostly in children, and transmitted primarily via the fecal-oral route.

Cases are most infectious during the first week of acute illness but may continue to shed virus in the stool for weeks. Incubation period is 3–5 days.

For EV71 HFMD the basic reproduction number[2] R_0 is estimated as 5.5 and for CoxA16 HFMD R_0 is estimated as 2.5 (Ma *et al.*, 2011).

Very little is known about the etiology of the viral strains causing HFMD or the factors associated with its outbreak and spread.

[2] the average number of secondary cases arising from an average primary case in an entirely susceptible population

Motivating example: hand, foot and mouth disease

In 2003, the Chinese Center for Disease Control and Prevention (CCDC) established a disease surveillance system which regulates the reporting of 39 notifiable infectious diseases including HFMD.

The purpose of the surveillance system is to monitor epidemics of infectious diseases, identify high case occurrence areas, predict and control epidemics, and provide information for formulating policy.

Each reported case of HFMD from the CCDC infectious disease surveillance system consists of the patient's geographical location, gender and age and the symptom onset date.

We analyze data from 59 regions in the Central North region of China over the period 2009–2011.

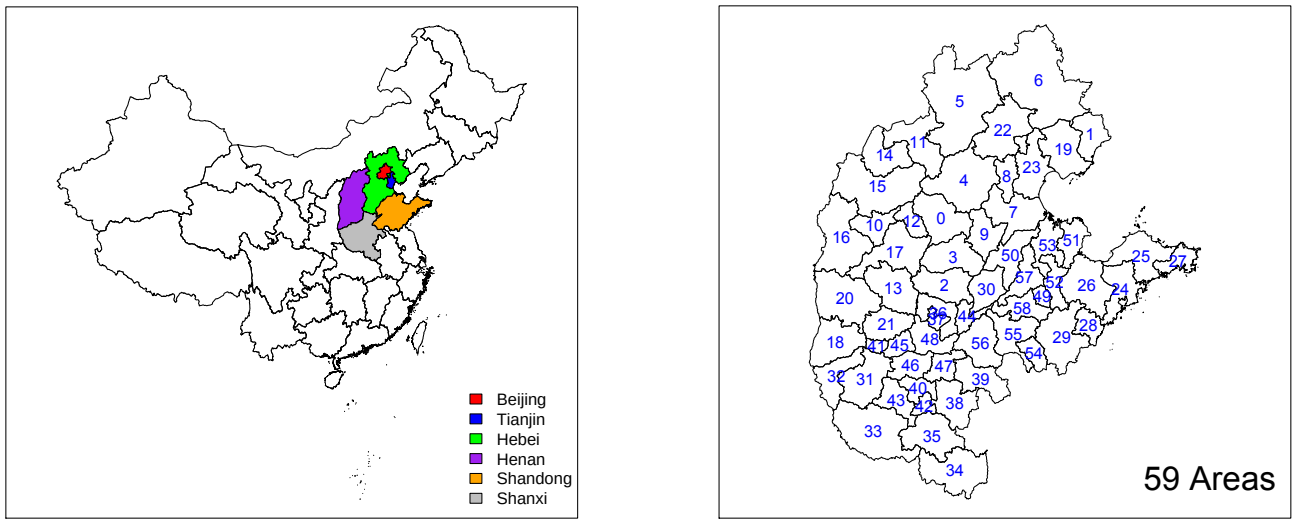

Figure 1 : Map of the central north region in relation to China as a whole (on the left) and map of the 59 prefecture (on the right).

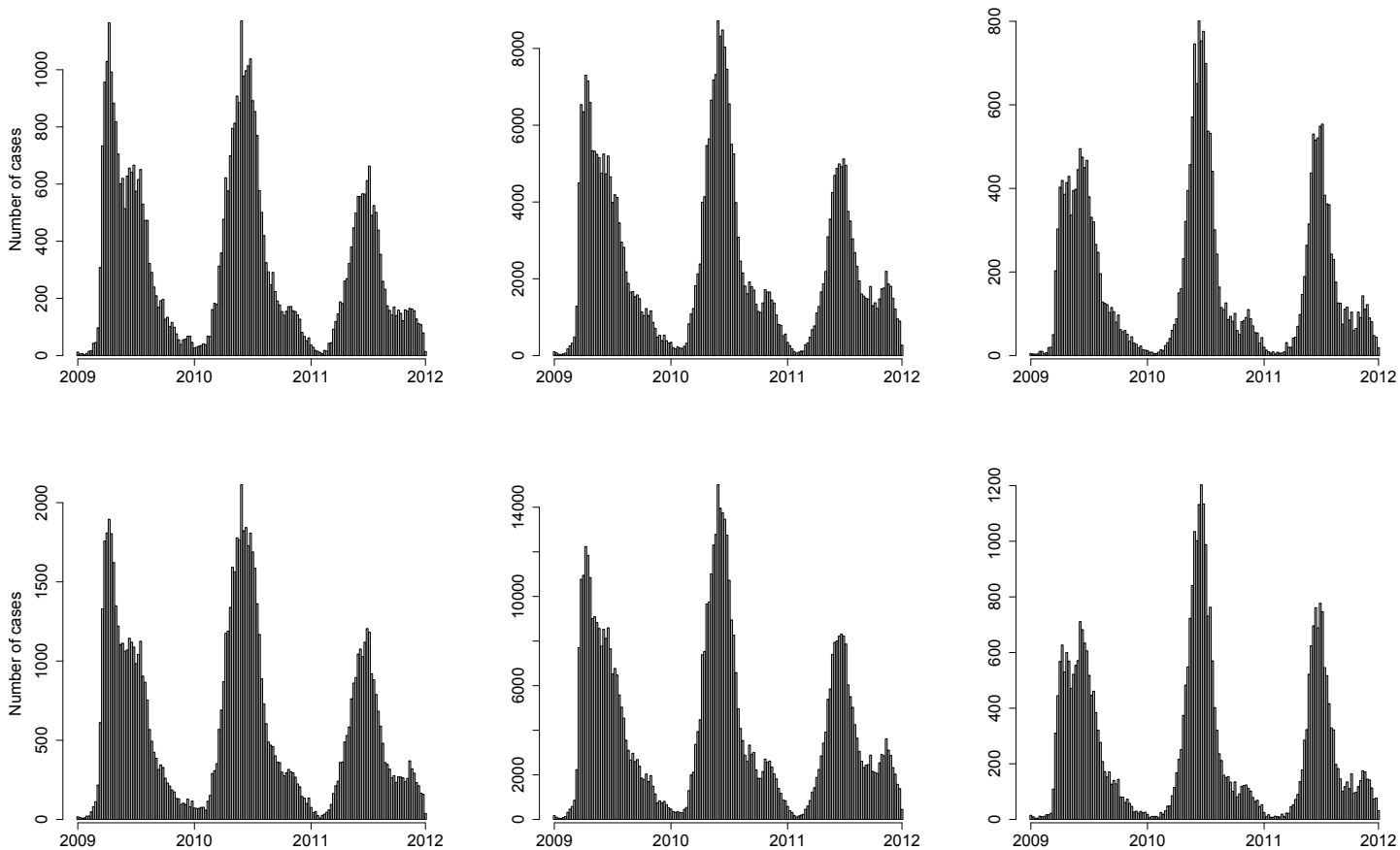

Figure 2 : Weekly epidemic curves of HFMD cases in the central north region of China: Top row females (0–1, 1–6, > 6), bottom row males (0–1, 1–6, > 6).

An epidemic/endemic framework

A statistical framework for analyzing spatio-temporal, aggregated infectious disease data was proposed by Held *et al.* (2005).

The framework was extended by Paul *et al.* (2008), Paul and Held (2011), Held and Paul (2012), Meyer and Held (2014) and Geilhufe *et al.* (2014).

These models are implemented within the `surveillance` package in `R` (Meyer *et al.*, 2015) and have been applied to a variety of diseases; see, for example, Höhle *et al.* (2011) and Herzog *et al.* (2011).

Notably the implementation does not provide a straightforward way to allow age/gender and space to be in the model, though Meyer and Held (2015) use survey information on contact rates in the epidemic/endemic model.

An epidemic/endemic model

We let Y_{it} denote the observed count of HFMD cases in area i and in week t for a generic age-gender group, $i = 1, \ldots, n$, $t = 1, \ldots, T$.

The population in area i is denoted N_i, and assumed constant during the study period.

The estimated incubation period of HFMD is between 3 and 7 days and individuals are most infectious for one week, and so we take a weekly time scale.

Under the model of Held *et al.* (2005) it is assumed that

$$Y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it})$$

with the mean μ_{it} being decomposed into three terms, which we refer to as autoregressive (AR), neighborhood (NE) and endemic (EN).

An epidemic/endemic model

Specifically,

$$\mu_{it} = \lambda_{it}^{\text{AR}} y_{i,t-1} + \lambda_i^{\text{NE}} \sum_{i'=1}^{n} w_{i'i} y_{i',t-1} + N_i \lambda_{it}^{\text{EN}}, \tag{2}$$

where:

- The autoregressive rate λ_i^{AR} dictates the contribution to the risk from the cases in area i in the previous time period.
- The neighborhood rate λ_i^{NE} determines the contribution from the neighboring areas, with the weights taken as $w_{i'i} = 1/|\text{ne}(i')|$, for $i' \in \text{ne}(i)$, where $\text{ne}(i)$ is the set of neighbors of area i, and $w_{i'i} = 0$ for $i' \notin \text{ne}(i)$ (Paul *et al.*, 2008).
- The endemic component λ_{it}^{EN} is a catch all term for contributions not catered for by the autoregressive and neighborhood components and, for example, includes seasonality. Note the N_i population multipliers[3].

[3] But no populations for the AR and NE components, with what justification?

An epidemic/endemic model

For the autoregressive self area rate:

$$\log \lambda_{it}^{\text{AR}} = \alpha_0^{\text{AR}} + \mathbf{z}_{it}^{\top} \boldsymbol{\beta}^{\text{AR}} + b_i^{\text{AR}},$$

where

- $\mathbf{z}_{it}$ represent a $q \times 1$ vector of area-time specific bases,
- $\boldsymbol{\beta}^{\text{AR}}$ is a $q \times 1$ vector of association parameters, and
- $b_i^{\text{AR}} \sim_{iid} \text{N}(0, \sigma^2_{\text{AR}})$ is an area-level autoregressive random effect.

For the neighborhood area rate:

$$\log \lambda_i^{\text{NE}} = \alpha_0^{\text{NE}} + b_i^{\text{NE}},$$

with

- $b_i^{\text{NE}} \sim_{iid} \text{N}(0, \sigma^2_{\text{NE}})$ an area-level neighborhood random effect.

An epidemic/endemic model

For the endemic component:

$$\log \lambda_{it}^{\text{EN}} = \alpha_0^{\text{EN}} + b_i^{\text{EN}} + \beta_{\text{sin}} \sin\left(\frac{t}{52} 2\pi\right) + \beta_{\text{cos}} \cos\left(\frac{t}{52} 2\pi\right),$$

where

- $b_i^{\text{EN}} \sim_{iid} \text{N}(0, \sigma^2_{\text{EN}})$ is the area-level endemic random effect.
- Seasonality is modeled via the sin/cos terms.

Modeling the weights

More recently (Meyer and Held, 2014) the weights are assumed to follow a power law,

$$w_{i'i} = \frac{o_{i'i}^{-A}}{\sum_{k=1}^{n} o_{ki}^{-A}}$$

where $o_{i'i}$ is the number of areas that must be crossed when moving between areas i and i', and A is a power which may be estimated.

The limit $A \to \infty$ corresponds to first-order dependency, and $A = 0$ gives equal weight to all areas.

The normalization ensures that $\sum_{k=1}^{n} w_{ki} = 1$ for all rows of the weight matrix (infecteds are being allocated to neighbors).

The power law allows "contact" between areas that are a large distance apart since it is "heavy-tailed".

Model Comparison

The above model is very flexible and so in practice many models may be fitted.

In terms of interpretation it is difficult to assess what is "big"

Interval estimates on regression coefficients β may be examined to assess significance (do the intervals contain zero?).

Seeing if random effects are needed is more difficult.

Using AIC or BIC is not straightforward in a mixed model framework.

Paul and Held (2011) compare models by comparing one-step ahead predictions (with the point to be predicted removed) with the observed value.

Covariate model for HFMD data

We include in the AR model:

- An indicator variable of whether school was in session ($z = 1$) or not ($z = 0$).
- The rest of the covariate vector $\mathbf{z}_{it}$ consists of a set of meteorological variables, specifically temperature, precipitation, relative humidity and wind speed.
- The meteorological data were obtained from the National Climate Data Center of the Department of Commerce of the United States, which includes over 300 weather stations from mainland China.
- We applied a tensor product cubic regression spline model to the averaged daily weather data from the monitoring stations, to obtain area-level summaries such as area averages.
- We assumed a change-point model for temperature and linear terms for the remaining variable,
- The area-level meteorological covariates entered in the model at a one week lag based on previous epidemiological studies of HFMD.

Computation

Bayesian inference for the parameters is made using an MCMC algorithm (not possible with INLA).

The proposed model was implemented using `WinBUGS` (Spiegelhalter *et al.*, 1998).

The advantage of a Bayesian framework is that any quantity that is a function of the model parameters, such as the local effective reproductive numbers, can be inferred via MCMC with the relevant posterior distributions, along with the associated uncertainty estimates.

A General Model

Recall: i, j and t represent area, age-gender stratum and week, respectively.

We assume

$$y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it})$$

with

$$\mu_{it} = \underbrace{\sum_{j=1}^{J} N_{ij} \sum_{j'=1}^{J} \lambda_{itjj'}^{\text{AR}} \frac{y_{i,t-1,j'}}{N_{ij'}}}_{\text{Self-Area Autoregressive}} + \underbrace{\sum_{i'=1}^{n} \lambda_{i}^{\text{NE}} w_{i'i} y_{i',t-1}}_{\text{Neighboring Area}} + \underbrace{\sum_{j=1}^{J} N_{ij} \lambda_{itj}^{\text{EN}}}_{\text{Endemic}}.$$

HFMD Model

We assume frequency dependent transmission and self-area infections from neighbors

$$
\begin{aligned}
\log \lambda^{\mathsf{AR}}_{itjj'} &= \alpha^{\mathsf{AR}}_0 + \alpha_{jj'} + \mathbf{z}^{\mathsf{T}}_{it}\boldsymbol{\beta}^{\mathsf{AR}} + b^{\mathsf{AR}}_i, \\
\log \lambda^{\mathsf{NE}}_{i} &= \alpha^{\mathsf{NE}}_0 + b^{\mathsf{NE}}_i, \\
\log \lambda^{\mathsf{EN}}_{itj} &= \alpha^{\mathsf{EN}}_j + b^{\mathsf{EN}}_i + \beta_{\mathsf{sin}} \sin\left(\frac{t}{52}2\pi\right) + \beta_{\mathsf{cos}} \cos\left(\frac{t}{52}2\pi\right).
\end{aligned}
$$

HFMD models

We consider the following models:

- Model 1 has fixed autoregressive and neighborhood components, i.e., $b_i^{\text{AR}} = 0$ and $b_i^{\text{NE}} = 0$, and the endemic component includes the random effects, $b_i^{\text{EN}} \sim_{iid} \text{N}(0, \sigma^2_{\text{EN}})$.
- Model 2 adds an autoregressive random effect, $b_i^{\text{AR}} \sim_{iid} \text{N}(0, \sigma^2_{\text{AR}})$, to M1.
- Model 3 adds a neighborhood random effect, $b_i^{\text{NE}} \sim_{iid} \text{N}(0, \sigma^2_{\text{NE}})$, to M2.
- Model 4 adds an endemic ICAR spatial random effect, $b^{\text{EN}}_{i,\text{ICAR}} \sim \text{ICAR}(\sigma^{\text{ICAR}}_{\text{EN}})$, to M3.

Models can be compared with DIC, but this measure is known to under penalize complex models (Plummer, 2008).

Contributions from different components

Recall $y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it})$ with

$$\mu_{it} = \underbrace{\sum_{j=1}^{J} N_{ij} \sum_{j'=1}^{J} \lambda_{itjj'}^{\text{AR,FD}} \frac{y_{i,t-1,j'}}{N_{ij'}}}_{\mu_{it}^{\text{AR}}} + \underbrace{\sum_{i'=1}^{n} \lambda_{i}^{\text{NE,FD}} w_{i'i} y_{i',t-1}}_{\mu_{it}^{\text{NE}}} + \underbrace{\sum_{j=1}^{J} N_{ij} \lambda_{itj}^{\text{EN}}}_{\mu_{it}^{\text{EN}}}.$$

To quantify the contribution from various components over time, we can define

$$\mu_t^{\text{C}} = \sum_{i=1}^{n} \mu_{it}^{\text{C}},$$

The proportion from component C is then

$$p_t^{\text{C}} = \frac{\mu_t^{\text{C}}}{\mu_t^{\text{AR}} + \mu_t^{\text{NE}} + \mu_t^{\text{EN}}}.$$

	M1: fixed AR, fixed NE iid END	M2: iid AR, fixed NE iid END	M3: iid AR, iid NE iid END	M4: iid AR, iid NE ICAR + iid END
α_0^{AR}	-4.76 (-4.90, -4.63)	-4.69 (-4.83, -4.55)	-4.86 (-5.01, -4.73)	-4.86 (-5.01, -4.72)
α_0^{NE}	-5.82 (-5.87, -5.767)	-5.80 (-5.86, -5.75)	-5.30 (-5.52, -5.09)	-5.30 (-5.53, -5.08)
α_1^{EN}	-20.77 (-23.73, -18.48)	-20.24 (-22.48, -18.52)	-19.86 (-22.22, -18.11)	-20.02 (-21.84, -18.16)
α_2^{EN}	-21.11 (-22.12, -20.27)	-21.35 (-22.65, -20.29)	-21.86 (-22.82, -20.8)	-22.09 (-23.16, -21.08)
α_3^{EN}	-22.39 (-25.83, -20.39)	-23.33 (-27.23, -20.84)	-25.18 (-31.41, -21.73)	-24.58 (-29.73, -21.37)
α_4^{EN}	-11.63 (-11.89, -11.37)	-11.48 (-11.76, -11.2)	-11.49 (-11.75, -11.16)	-11.53 (-11.73, -11.34)
α_5^{EN}	-20.22 (-20.83, -19.68)	-20.05 (-20.81, -19.38)	-20.63 (-21.49, -20.01)	-20.86 (-21.46, -20.11)
α_6^{EN}	-21.85 (-25.46, -20.22)	-22.36 (-25.83, -20.52)	-23.37 (-27.33, -21.06)	-23.39 (-27.35, -21.12)
β_{school}	0.066 (0.06, 0.072)	0.073 (0.067, 0.080)	0.072 (0.066, 0.078)	0.072 (0.066, 0.079)
β_{tempr}	0.022 (0.021, 0.023)	0.023 (0.023, 0.024)	0.024 (0.023, 0.024)	0.023 (0.023, 0.024)
β_{tempr2}	-0.0263 (-0.0267, -0.0259)	-0.029 (-0.029, -0.028)	-0.029 (-0.029, -0.028)	-0.029 (-0.029, -0.028)
β_{humid}	-0.0014 (-0.0016, -0.0012)	-0.0023 (-0.0025, -0.0020)	-0.0023 (-0.0025, -0.0021)	-0.0023 (-0.0025, -0.0020)
β_{windsp}	0.051 (0.049, 0.053)	0.042 (0.039, 0.045)	0.042 (0.040, 0.045)	0.042 (0.040, 0.045)
β_{precip}	0.010 (-0.01, 0.032)	0.050 (0.029, 0.071)	0.052 (0.031, 0.073)	0.052 (0.031, 0.073)
$\beta_{sin,1}$	9.74 (7.5, 12.71)	9.29 (7.555, 11.59)	8.87 (7.22, 11.28)	9.04 (7.08, 10.87)
$\beta_{sin,2}$	10.6 (9.72, 11.55)	10.78 (9.68, 12.02)	11.24 (10.29, 12.14)	11.42 (10.47, 12.47)
$\beta_{sin,3}$	4.39 (2.33, 7.73)	5.007 (2.52, 8.44)	5.52 (2.56, 9.92)	5.02 (2.2, 9)
$\beta_{sin,4}$	0.71 (0.60, 0.83)	0.682 (0.59, 0.78)	0.67 (0.58, 0.76)	0.67 (0.58, 0.77)
$\beta_{sin,5}$	10.24 (9.76, 10.87)	10.09 (9.50, 10.70)	10.6 (10.01, 11.4)	10.8 (10.07, 11.39)
$\beta_{sin,6}$	4.12 (2.61, 7.31)	4.483 (2.72, 7.45)	4.74 (2.73, 7.96)	4.69 (2.74, 7.89)
$\beta_{cos,1}$	1.07 (0.48, 1.79)	0.945 (0.43, 1.57)	0.83 (0.34, 1.46)	0.85 (0.33, 1.41)
$\beta_{cos,2}$	0.98 (0.74, 1.24)	1.105 (0.82, 1.41)	1.23 (0.98, 1.49)	1.27 (1.01, 1.56)
$\beta_{cos,3}$	1.89 (0.80, 3.72)	2.437 (1.03, 4.81)	4.17 (1.761, 9.07)	3.81 (1.50, 7.80)
$\beta_{cos,4}$	-0.84 (-0.92, -0.75)	-0.866 (-0.94, -0.79)	-0.85 (-0.925, -0.78)	-0.85 (-0.92, -0.78)
$\beta_{cos,5}$	0.64 (0.47, 0.81)	0.656 (0.49, 0.83)	0.83 (0.65, 1.03)	0.86 (0.66, 1.05)
$\beta_{cos,6}$	2.05 (1.03, 4.07)	2.36 (1.19, 4.51)	3.16 (1.58, 5.86)	3.16 (1.5, 5.88)
σ_{AR}	-	0.060 (0.050, 0.074)	0.056 (0.047, 0.069)	0.057 (0.047, 0.056)
σ_{NE}	-	-	0.82 (0.69, 1.01)	0.83 (0.69, 1.01)
σ_{EN}	0.97 (0.81, 1.20)	1.01 (0.84, 1.23)	1.05 (0.87, 1.29)	0.66 (0.20, 0.91)
σ_{EN}^{ICAR}	-	-	-	1.08 (0.58, 1.78)
p_D	103.1	162.1	217.9	215.0
DIC ($\times 10^5$)	3.011	2.996	2.980	**2.980**

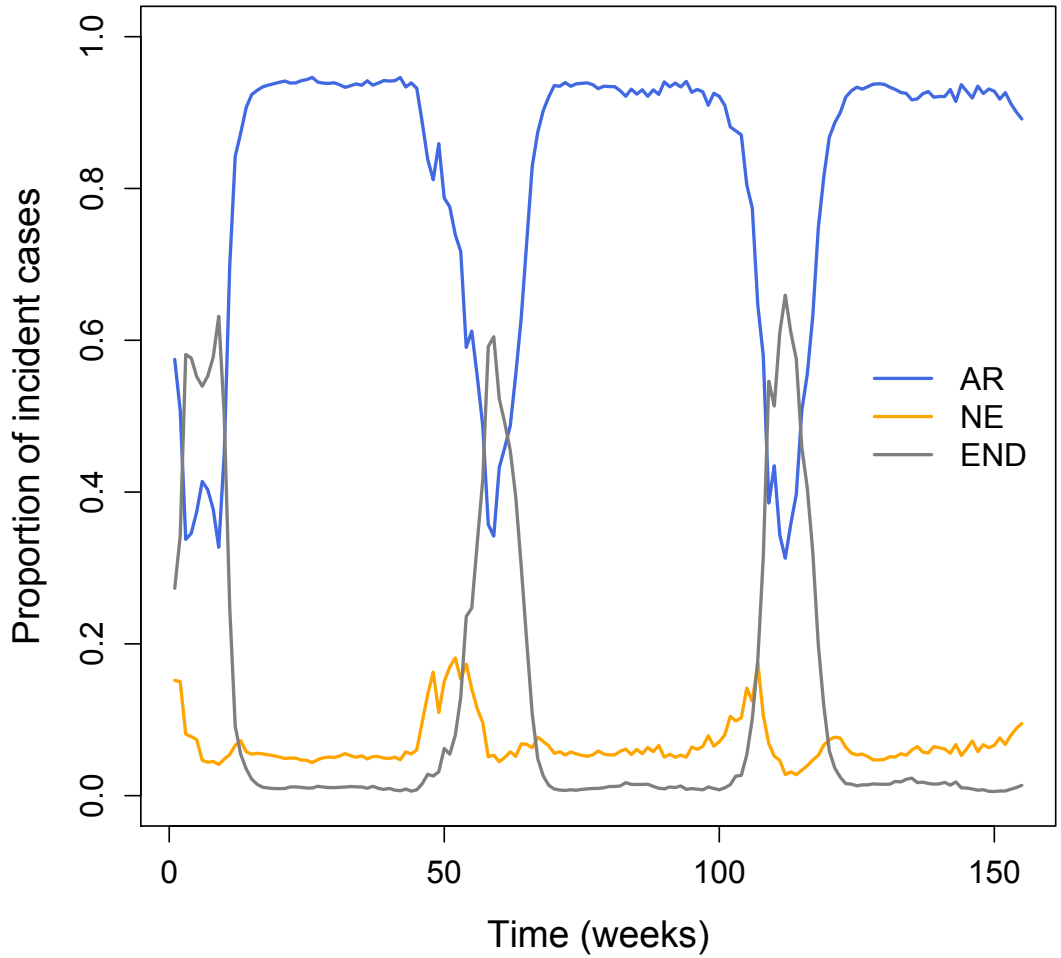

Figure 3 : Estimated proportion of infection risk contributed from AR, NE, EN components.

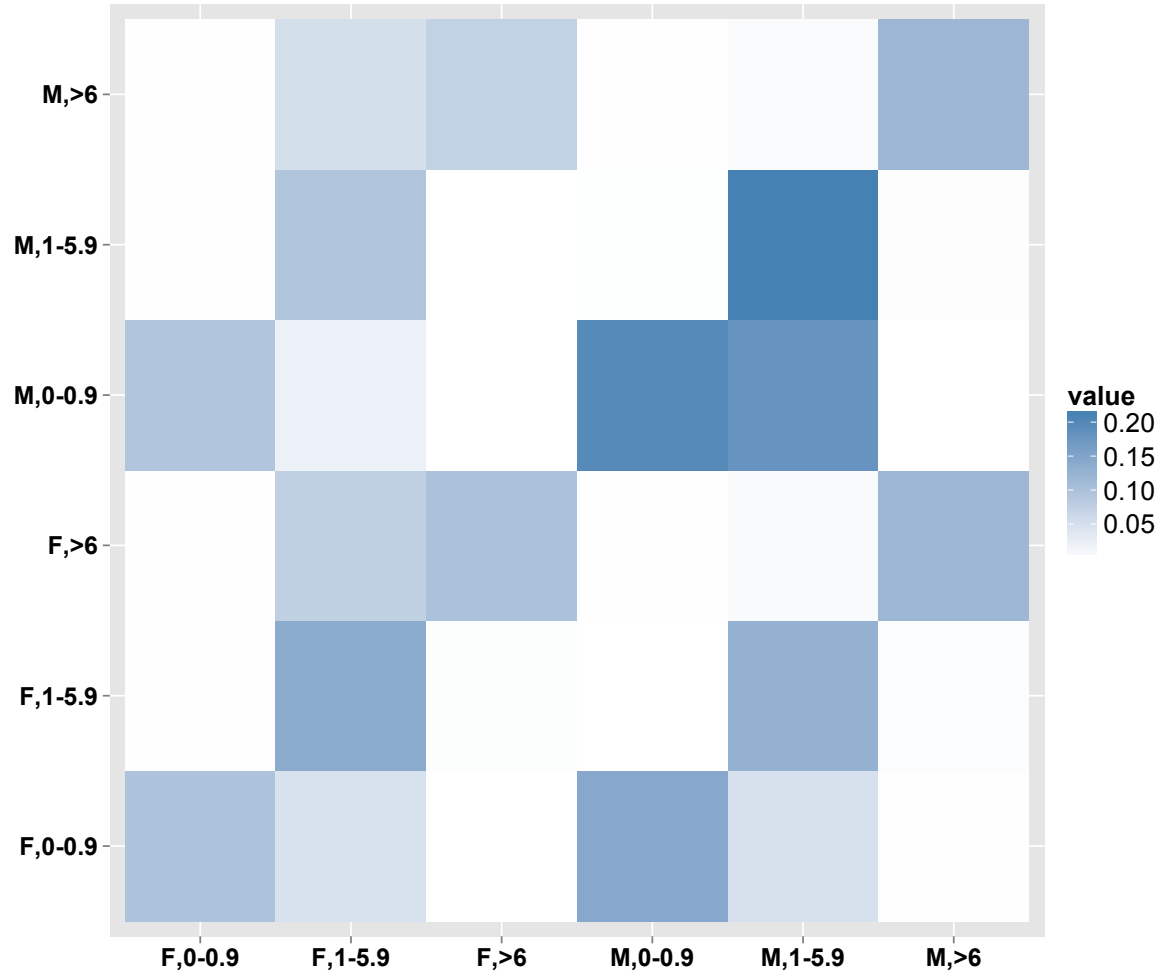

Figure 4 : Estimated transmission rate between age-gender strata, Model 4.

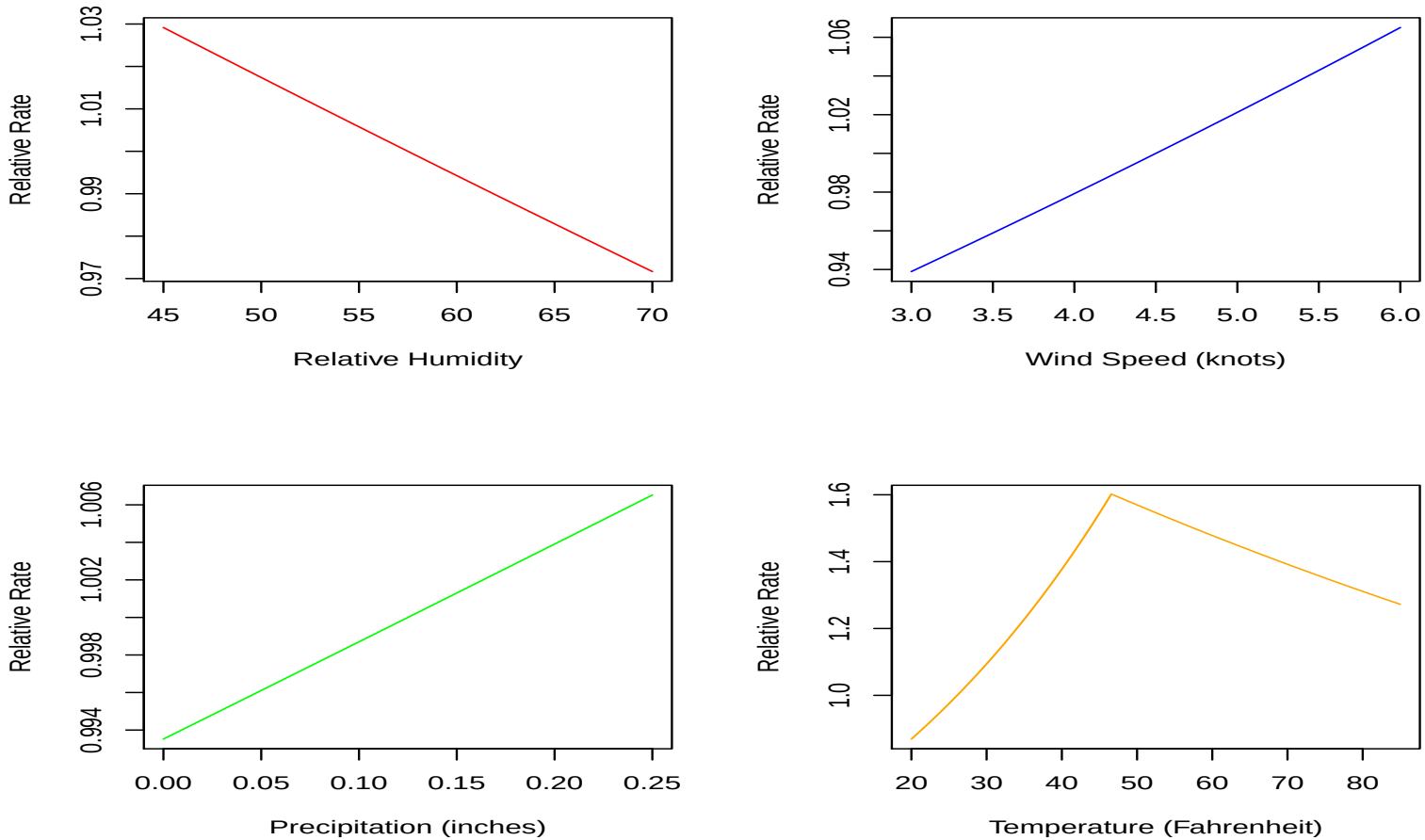

Figure 5 : Summary of meteorological associations.

Between strata transmission

Table 2 : The probabilities of a strata j susceptible being infected by a strata j' infective (given the same values of covariates).

	Infective Strata					
Susc Strata	F, 0–1	F, 1–6	F, >6	M, 0–1	M, 1–6	M, > 6
F, 0–1	0.50	0.00	0.00	0.48	0.00	0.00
F, 1–6	0.09	0.33	0.18	0.04	0.23	0.13
F, > 6	0.00	0.02	0.57	0.00	0.00	0.41
M, 0–1	0.42	0.00	0.00	0.58	0.00	0.00
M, 1–6	0.06	0.22	0.01	0.30	0.38	0.01
M, >6	0.00	0.02	0.48	0.00	0.01	0.49

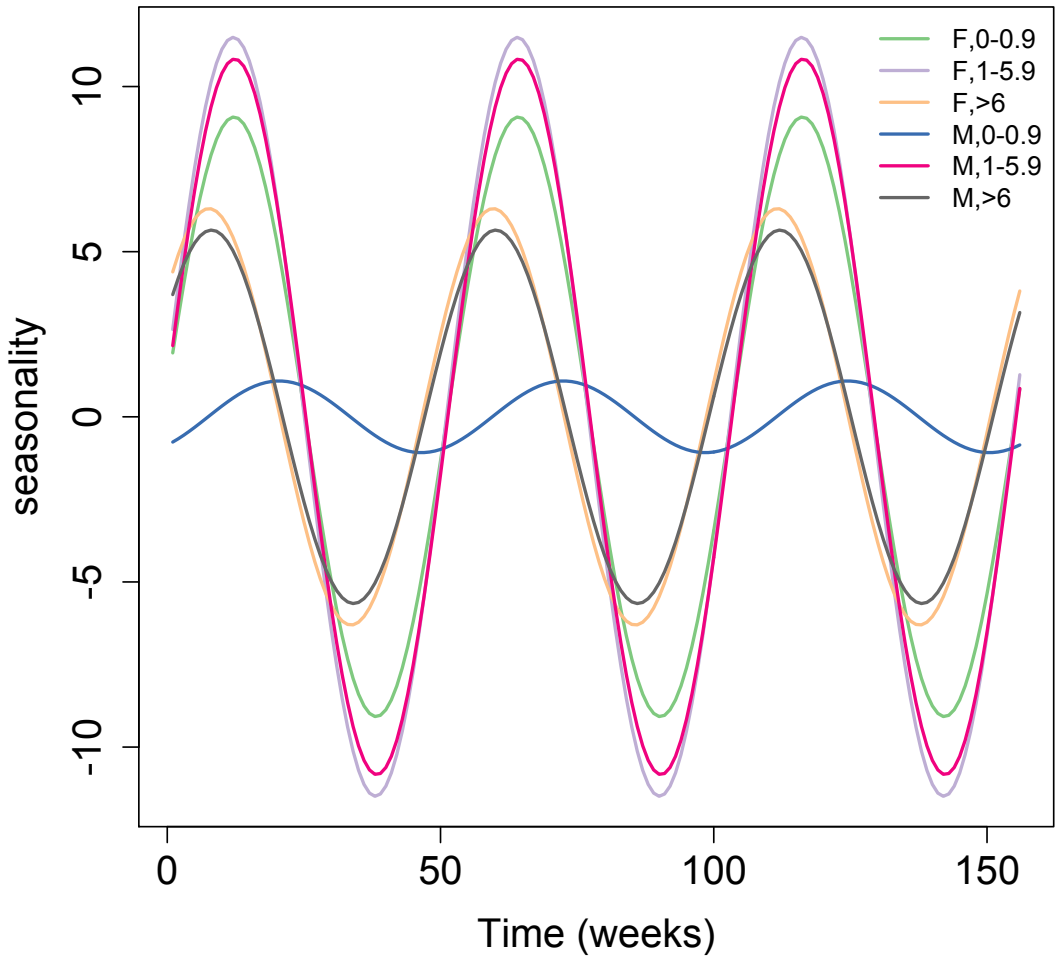

Figure 6 : Estimated seasonality for age-gender subgroups from Model 4.

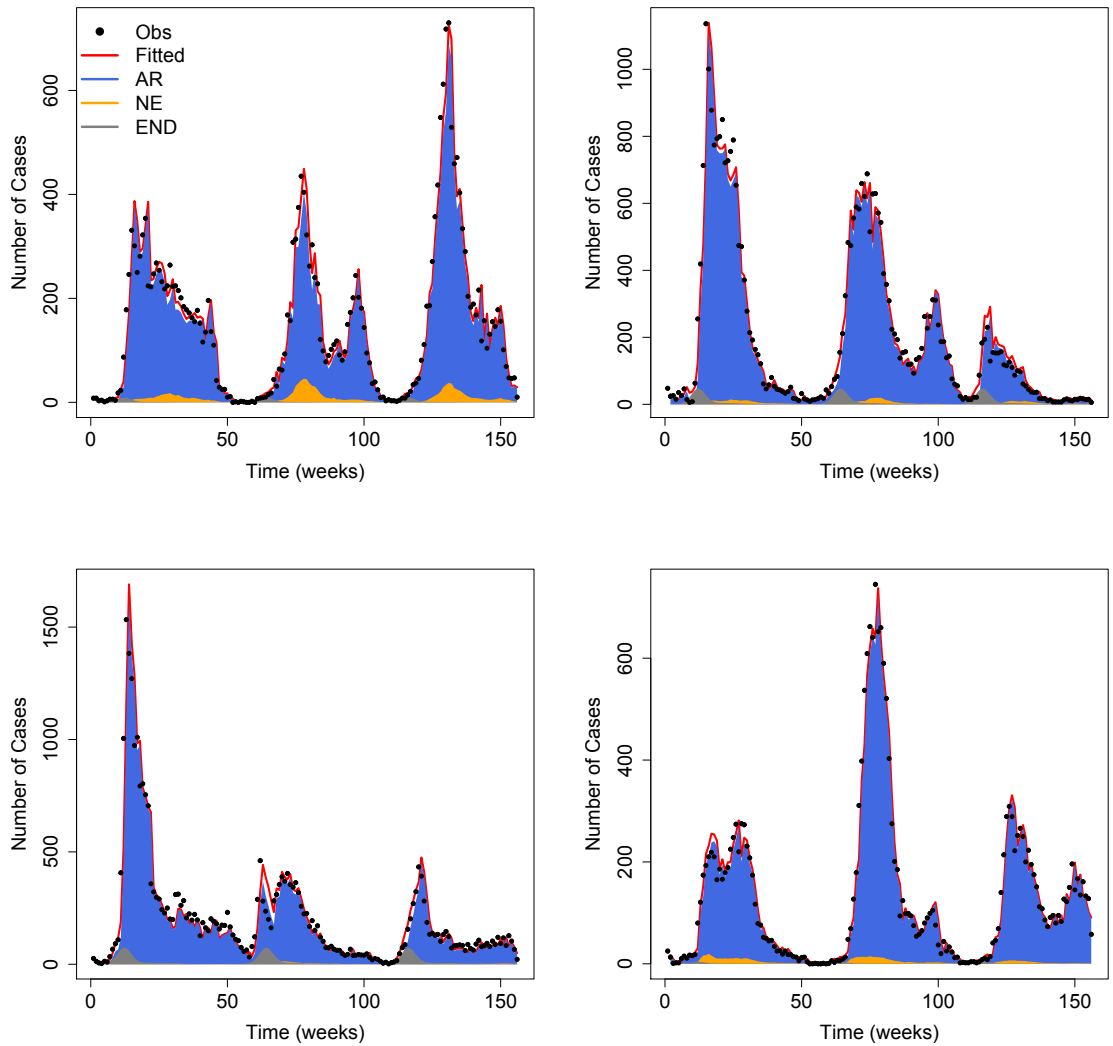

Figure 7 : Fitted number of cases for selected areas, with contribution by different components.

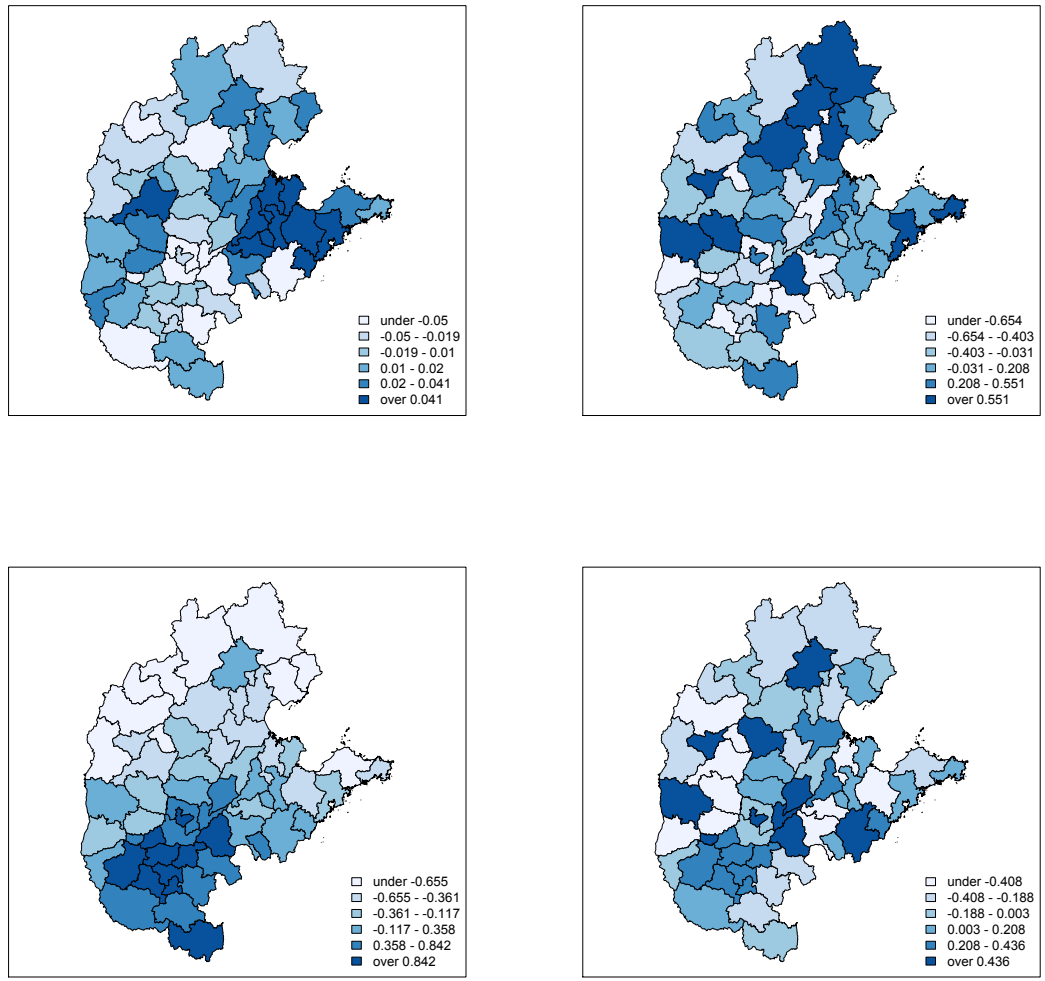

Figure 8 : Maps of the estimated random effects from AR, NE and EN components using model 4.

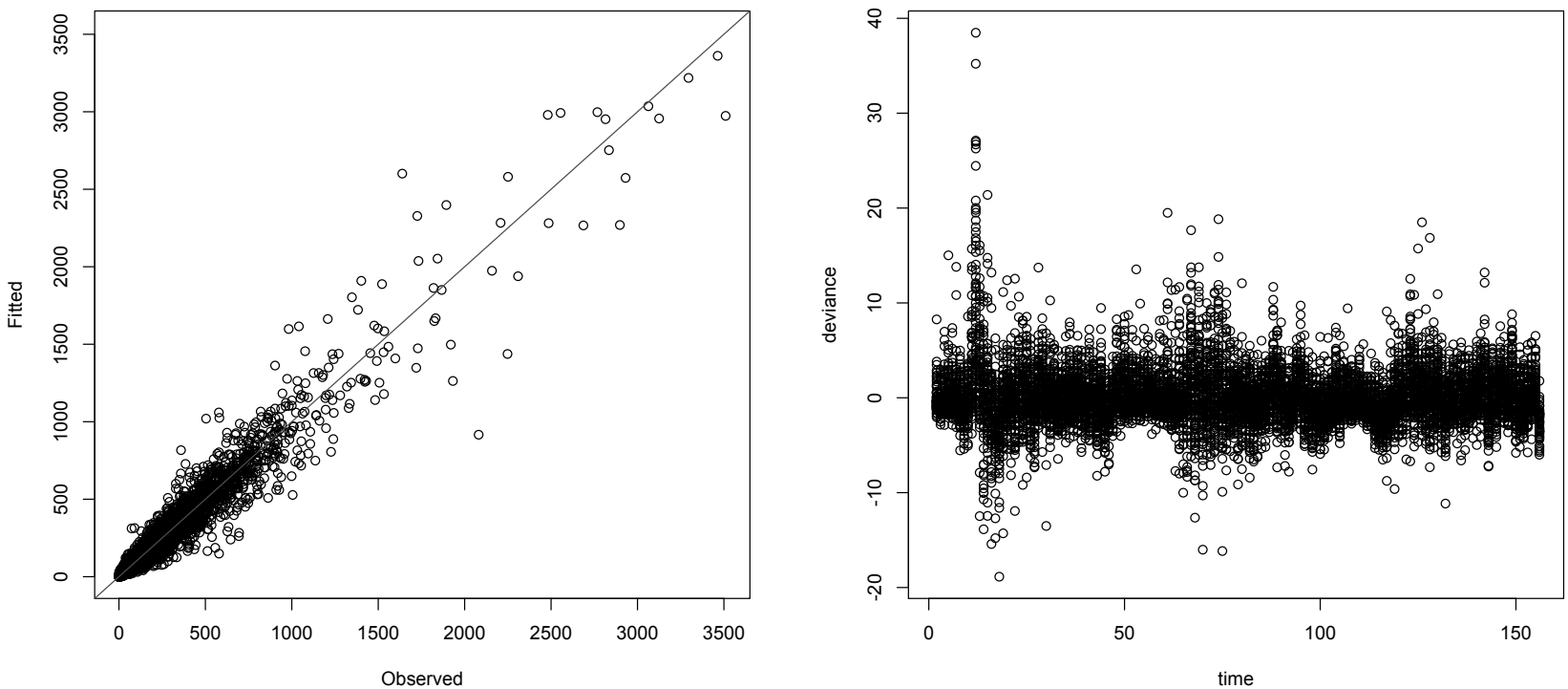

Figure 9 : Comparison of fitted and observed counts, and the plot of deviance residuals from Model 4.

Spatial TSIR Models

Xia *et al.* (2004) describe a model for modeling measles infections in time and space.

Measles epidemics may be characterized into:

1. Type I behavior with endemic cycles in large cities.
2. Type II dynamics predictable epidemics with local extinction during epidemic troughs in medium-sized communities.
3. Type III dynamics with irregular epidemics interpersed with prolonged periods of local disease extinction (so-called epidemic fade-outs) in small communities.

Regional movement of hosts is important.

Relationship to TSIR Model

For the Xia *et al.* (2004) TSIR framework with a 'gravity model' for movement between areas:

$$\begin{aligned}
\mathrm{E}[Y_{it}|y_{i,t-1}] &= \lambda_{it} y_{i,t-1} \\
\lambda_{it} &= \frac{\beta_t S_{it} (y_{i,t-1} + \delta_{it})^{\alpha}}{N_i} \\
\delta_{it} &\sim \mathrm{Gamma}(m_{it}, 1) \\
\mathrm{E}[\delta_{it}] &= m_{it} \\
&= \theta N_{it}^{\tau_1} \sum_{i'=1}^{n} \frac{y_{i',t-1}}{d_{i'i}^{\rho}}
\end{aligned}$$

So with $\alpha = \tau_1 = \tau_2 = 1$ and $S_{it} \approx N_i$, we could write

$$\lambda_{it} = \lambda_t^{\mathrm{AR}} y_{i,t-1} + \lambda^{\mathrm{NE}} N_{it} \sum_{i'=1}^{n} \frac{y_{i',t-1}}{d_{i'i}^{\rho}}$$

where $\lambda_t^{\mathrm{AR}} = \beta_t$, $\lambda^{\mathrm{NE}} = \theta$ and we have a distance-based weighting scheme.

Spatial TSIR Models

Inference is ad-hoc (as self-confessed by the authors!) with some parameters fixed using previously obtained estimates.

Estimates for the remaining parameters obtained by minimizing short-term prediction error and matching aspects of the long-term behavior.

More rigorous approach described in Jandarov *et al.* (2014).

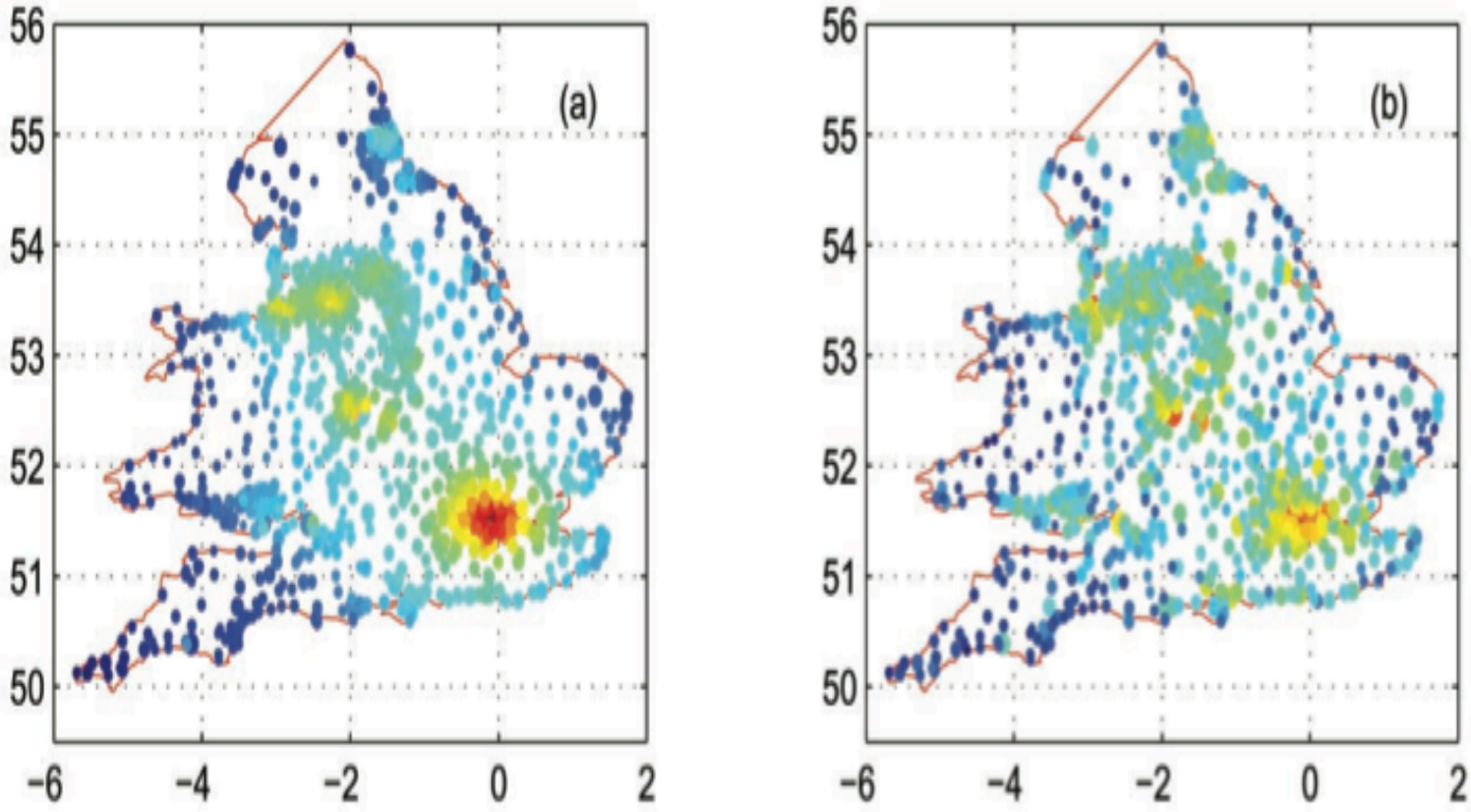

Figure 5: Epidemical isolation in the model measles population. *a*, Per capita influx of infections from all the other areas. *b*, Per capita emigrants from the area. The values of the per capita influx or emigrants (from low to high) are indicated by color from blue to cyan to yellow to red.

Figure 10 : Figure from Xia *et al.* (2004).

A model for plantation data

We now turn our attention to the situation in which point data are available.

Brown *et al.* (2014) describe a spatial susceptible-infectious (SI) model in which the intensity at time t and location $\mathbf{x}_i$ is

$$\lambda(\mathbf{x}_i, t) = \mu + \sum_{j:\tau_j<t} \theta f(\mathbf{x}_i - \mathbf{x}_j; \sigma)$$

where τ_j is the infection time of individual j and

$$\theta f(\mathbf{x}_i - \mathbf{x}_j; \sigma)$$

is the transmission rate from individual j to individual i, and μ is the environmental contribution.

A model for plantation data

The data concern plant infections transmitted by aphids and the Gaussian function is chosen to represent spatial connection:

$$f(d;\sigma) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{d^2}{2\sigma^2}\right).$$

As usual with models such as these, the likelihood is not straightforward to calculate, and an auxiliary variable method is used.

A model for plantation data

Point data:

- A Bayesian approach to inference is taken.
- The likelihood, given known infection times $\tau_1, \ldots, \tau_n$ (total observation period is $[0, T]$), is

$$
\begin{aligned}
L(\mu, \theta, \sigma^2) &= \left[\prod_{i:\tau_i \leq T} \exp\left\{ - \int_0^{\tau_i} \lambda(\mathbf{x}_i, t) dt \right\} \lambda(\mathbf{x}_i, \tau_i) \right] \\
&\times \prod_{i:\tau_i > T} \exp\left\{ - \int_0^{\tau_i} \lambda(\mathbf{x}_i, t) dt \right\}
\end{aligned}
$$

- The data are interval censored (plants are surveyed on six occasions) and so the unobserved times are imputed via an auxiliary variable scheme.
- Much of Brown *et al.* (2014) concerns computation.

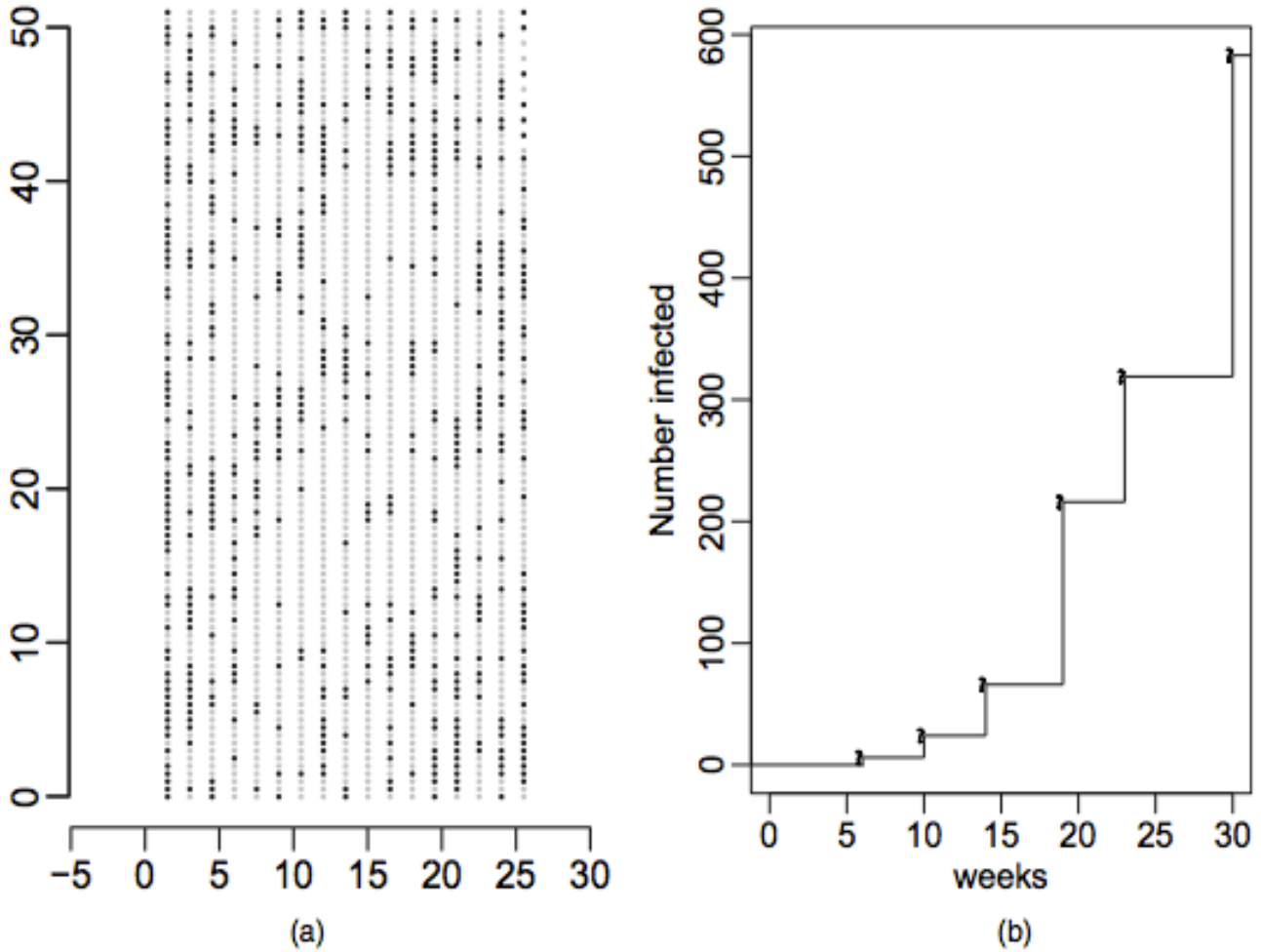

Fig. 1. (a) Location of 1742 sugar canes which are infected (•) or uninfected (•) at the end of the study period, along with (b) the cumulative number of infections over time

Figure 11 : Raw data, from Brown *et al.* (2014).

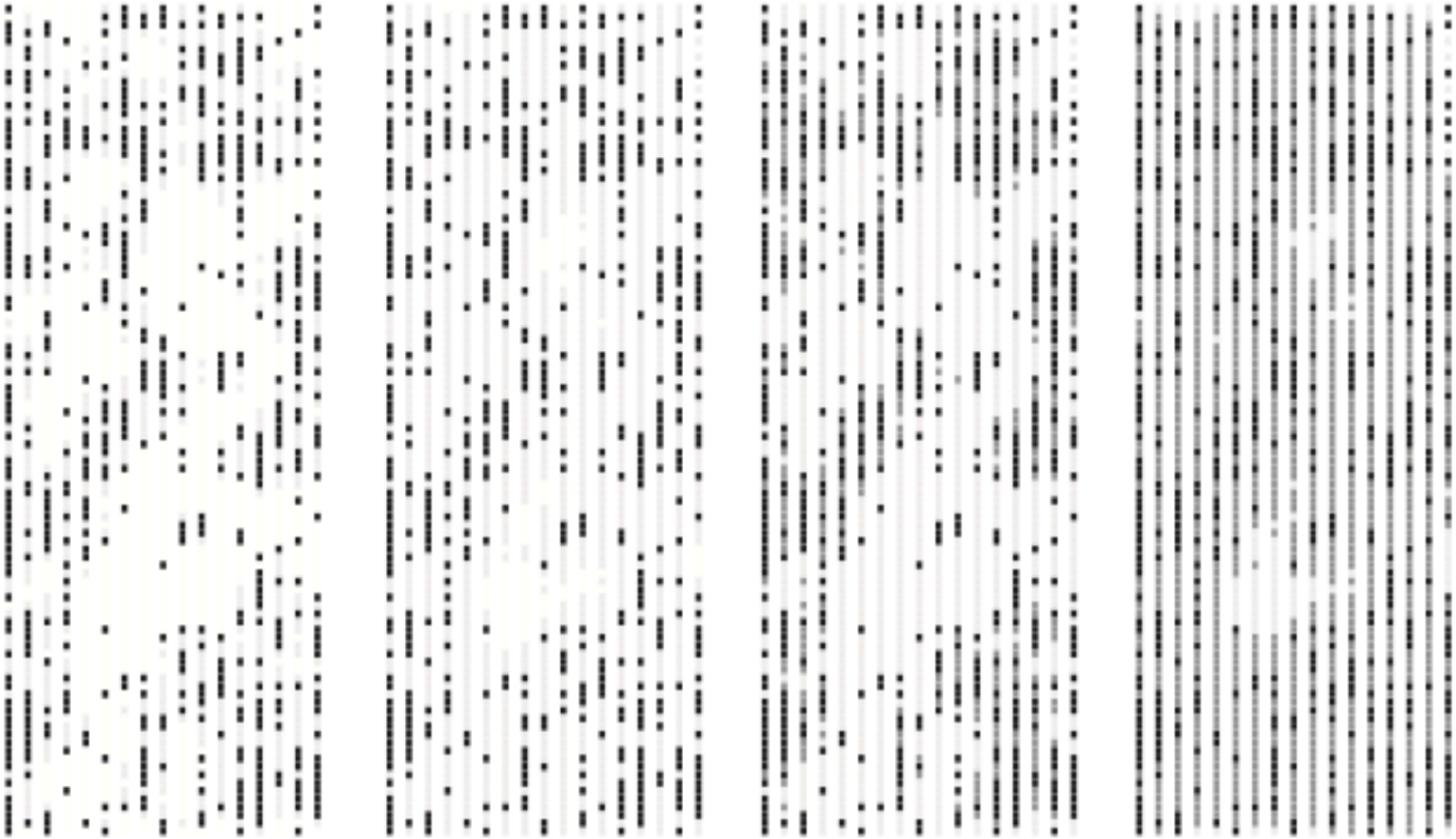

Fig. 5. Probabilities forecast at week 30 of each plant being infected by (a) week 35, (b) week 40, (c) week 50 and (d) week 60: the probabilities are 0–19% (□), 20–79% (▒), 80–99% (▓) and 100% (■) (colour versions are available in the on-line supplement)

Figure 12 : Predictions, from Brown *et al.* (2014).

Animal applications

Animal disease epidemics: A number of authors have considered data in the form of the infectious status of farms.

For example, data on foot and mouth disease (FMD) have been analyzed by a number of authors including Keeling *et al.* (2001), Lawson and Zhou (2005), Diggle (2006) and Jewell *et al.* (2009).

In the latter, a Susceptible-Infected-Notified-Removed model is assumed.

Likelihood is constructed from a time inhomogenous Poisson point process with rates that depend on the states of each farm over time.

Spatial transmission is modeled using a Cauchy-type kernel and computation is via RJMCMC, again with auxiliary variables.

Conclusions

This lecture has largely concentrated on spatio-temporal models for aggregated count data (though we touched on point data at the end) – with such data much fine detail is lost and so biologically motivated models are difficult to fit.

The full SIR formulation (and its spin offs, such as SEIR) are computationally hard to fit since the likelihood is analytically intractable.

Spatio-temporal methods may be used to assess the effect of intervention programs, see for example Azman *et al.* (2012).

The space-time models within the `surveillance` package do not currently allow adjustment for age and gender, i.e. different transmission dynamics for different stratum, which is problematic.

Much work to be done!!!

References

Anderson, R. and May, R. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*. Springer Lecture Notes.

Azman, A., Luquero, F., Rodrigues, A., Palma, P., Grais, R., Banga, C., Grenfell, B., and Lessler, J. (2012). Urban cholera transmission hotspots and their implications for reactive vaccination: evidence from Bissau City, Guinea Bissau. *PLoS Neglected Tropical Diseases*, **6**, 1–11.

Bailey, N. (1975). *The Mathematical Theory of Infectious Diseases*. Oxford University Press.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., and Wang, Y. (2015). Space-time spline smoothing of infectious disease counts. *Statistics in Medicine*. To appear. DOI: 10.1002/sim.6785.

Becker, N. (1989). *Analysis of Infectious Disease Data*. Chapman and Hall.

Brown, P., Chimard, F., Remorov, A., Rosenthal, J., and Wang, X. (2014). Statistical inference and computational efficiency for spatial infectious disease models with plantation data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 467–482.

Cauchemez, S. and Ferguson, N. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society Interface*, **5**, 885–897.

Daley, D. and Gani, J. (1999). *Epidemic Modelling: An Introduction*. Cambridge University Press.

Diggle, P. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical Methods in Medical Research*, **15**, 325–336.

Geilhufe, M., Held, L., Skrøvseth, S., Simonsen, G., and Godtliebsen, F. (2014). Power law approximations of movement network data for modeling infectious disease spread. *Biometrical Journal*, **56**, 363–382.

Gibson, G. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, **15**, 19–40.

Halloran, M., I.M. Longini, J., and Struchiner, C. (2010). *Design and Analysis of Vaccine Studies*. Springer, New York.

He, D., Ionides, E., and King, A. (2010). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*, **7**, 271–283.

Held, L. and Paul, M. (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal*, **54**, 824–843.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, **5**, 187–199.

Herzog, S., Paul, M., and Held, L. (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in german surveillance data. *Epidemiology and Infection*, **139**, 505–515.

Höhle, M., Siedler, A., Bader, H.-M., Ludwig, M., Heininger, U., and von Kries, R. (2011). Assessment of varicella vaccine effectiveness in Germany: a time-series approach. *Epidemiology and Infection*, **139**, 1710–1719.

Jandarov, R., Haran, M., Bjornstad, O., and Grenfell, B. (2014). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of the Royal Statistical Society, Series C*, **63**, 423–444.

Jewell, C., Kypraios, T., Neal, P., and Roberts, G. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, **4**, 465–496.

Keeling, M. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton and Oxford.

Keeling, M., Woolhouse, M., Shaw, D., Matthews, L., Chase-Topping, M., Haydon, D., Cornell, S., Kappey, J., Wilesmith, J., and Grenfell, B. (2001). Dynamics of the 2001 uk foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, **294**, 813–817.

Kermack, W. and McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character*, **115**, 700–721.

Knorr-Held, L. and Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society: Series C*, **52**, 169–183.

Koepke, A., Longini Jr, I., Halloran, M., Wakefield, J., and Minin, V. (2015). Predictive modeling of cholera outbreaks in Bangladesh. *Annals of Applied Statistics*.

Lawson, A. and Zhou, H. (2005). Spatial statistical modeling of disease outbreaks with particular reference to the uk foot and mouth disease (fmd) epidemic of 2001. *Preventive Veterinary Medicine*, **71**(3), 141–156.

Lekone, P. and Finkenstädt, B. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, **62**, 1170–1177.

Ma, E., Fung, C., Yip, S., Wong, C., Chunag, S., and Tsang, T. (2011). Estimation of the basic reproduction number of enterovirus 71 and coxsackievirus a16 in hand, foot and mouth disease outbreaks. *Pediatric Infectious Disease Journal*, **30**, 675–679.

McKinley, T., Cook, A., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, **5**, 24.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics*, **8**, 1612–1639.

Meyer, S. and Held, L. (2015). Incorporating social contact data in spatio-temporal models for infectious disease spread. *Submitted for publication*.

Meyer, S., Held, L., and Höhle, M. (2015). Spatio-temporal analysis of epidemic phenomena using the R package `surveillance`. *Preprint available from http://arxiv.org/abs/1411.0416*. Revised for Journal of Statistical Software.

Mugglin, A., Cressie, N., and Gemmell, I. (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine*, **21**, 2703–2721.

Neal, P. (2012). Efficient likelihood-free bayesian computation for household epidemics. *Statistics and Computing*, **22**, 1239–1256.

Neal, P. and Kypraios, T. (2015). Exact bayesian inference via data augmentation. *Statistics and Computing*, **25**, 333–347.

O'Neill, P. and Becker, N. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics*, **2**, 99–108.

O'Neill, P. and Roberts, G. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series A*, **162**, 121–129.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, **30**, 1118–1136.

Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, **27**, 6250–6267.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539.

Riley, S. (2007). Large-scale spatial-transmission models of infectious disease. *Science*, **316**, 1298–1301.

Spiegelhalter, D., Thomas, A., and Best, N. (1998). *WinBUGS User Manual*. Cambridge, UK, 1.1.1 edition.

Tong, C. and Bible, J. (2009). Global epidemiology of enterovirus 71. *Future Virology*, **4**, 501–510.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. (2010). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202.

Xia, Y., Bjørnstad, O. N., and Grenfell, B. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, **164**, 267–281.