Inference of latent event times and transmission networks in individual level infectious disease models

Justin Angevaare[a,*], Zeny Feng[a], Rob Deardon[b]

[a] *University of Guelph, Canada*
[b] *University of Calgary, Canada*

Abstract

Transmission networks indicate who-infected-whom in epidemics. Reconstruction of transmission networks is invaluable in applying and developing effective control strategies for infectious diseases. We introduce transmission network individual level models (TN-ILMs), a competing-risk, continuous time extension to individual level model framework for infectious diseases of Deardon et al. (2010). Through simulation study using a Julia language software package, Pathogen.jl, we explore the models with respect to their ability to jointly infer latent event times, latent disease transmission networks, and the TN-ILM parameters. We find good parameter, event time, and transmission network inference, with enhanced performance for inference of transmission networks in epidemic simulations that have higher spatial signals in their infectivity kernel. Finally, an application of a TN-ILM to data from a greenhouse experiment on the spread of tomato spotted wilt virus is presented.

Keywords: individual level infectious disease model, transmission network, epidemics, Julia language

1. Introduction

The propagation of disease through a population can be described by a disease transmission network. Transmission networks are directed graphs that indicate who-infected-whom in an epidemic. These transmission networks are different from contact networks, which are usually undirected graphs indicating potential interaction, but not necessarily disease transmission. The transmission network is often assumed to be a directed subgraph of an underlying contact network. The use of simulation methods to study transmission network topology and corresponding epidemic dynamics is a highly active area of research. Recent

*Corresponding author
Email addresses: jangevaa@uoguelph.ca (Justin Angevaare), zfeng@uoguelph.ca (Zeny Feng), robert.deardon@ucalgary.ca (Rob Deardon)
URL: https://jangevaare.github.io (Justin Angevaare), https://zfeng.uoguelph.ca (Zeny Feng), https://people.ucalgary.ca/~robert.deardon (Rob Deardon)

simulation work has largely focused on behavioural modulation of human contact networks during epidemics, such as Azizi et al. (2020), Abdulkareem et al. (2020), Sun et al. (2017), and Pipatsart et al. (2017). Through such work we better understand our capacity to quench socially transmitted diseases affecting humans.

Through reconstruction of transmission networks, we can learn how exactly particular diseases are spread from individual to individual. Transmission network reconstruction is often a resource intensive process, involving contact tracing and increasingly, phylogenetic analysis (Giardina et al., 2017). While resource intensive, the level of detail such investigations provide is invaluable in identifying and applying effective control strategies. As our methods and technology have improved, transmission network reconstruction has moved from a retrospective analysis, into an analysis we may perform in real time during ongoing epidemics. That being said, additional work is needed to understand the uncertainty involved in such transmission network reconstructions (Moshiri et al., 2019).

Often, spatial models without explicit contact networks are used in infectious disease research. Spatial data sets are easier to acquire in comparison to contact networks. Though there is some loss of information, spatial models that represent individuals through a static location are flexible in that they can be applied to diseases transmitted through direct or indirect contact as well as those transmitted by vectors.

In Section 2, we introduce transmission network individual level models (TN-ILMs), and a method for joint inference of latent event times, transmission networks, and TN-ILM parameters using a new open-source software package, Pathogen.jl. The inference capabilities of the methods implemented in Pathogen.jl are assessed with respect to spatially driven epidemics through simulation study in Section 3. We finally provide an application of a spatial TN-ILM to data from a greenhouse experiment on the spread of tomato spotted wilt virus in Section 4.

2. Transmission network individual level models

In the individual level model (ILM) framework of Deardon et al. (2010) for the spread of infectious diseases, the rates that describe transition between disease classes are individual specific. These transition rates are functions of relevant risk factors for each individual represented in the epidemic model. The ILM framework was described for discrete time, where transition rates are tied to the states of individuals in the population at the start of regular time periods. In the case of transmission of a disease to an individual (*i.e.*, transition of that individual into the infectious class in a susceptible-infectious-removed (SIR) model), that transition rate is not only specific to the susceptible individual, but also specific to the set of infectious individuals in the population at the beginning

of a time period, $I_{(t)}$. The transition of individual i from the susceptible state to the infectious state, $\lambda_{SI}(i,t)$, is the culmination of various sources of infectious pressure. $\lambda_{SI}(i,t)$ is calculated as:

$$\lambda_{SI}(i,t) = \left[\Omega_S(i) \sum_{k \in I_{(t)}} \Omega_T(k)\kappa(i,k)\right] + \epsilon(i,t) \text{ for } i \in S_{(t)}, \tag{1}$$

where,

- $I_{(t)}$ is the set of infectious individuals during the t^{th} time period,
- $S_{(t)}$ is the set of susceptible individuals during the t^{th} time period,
- $\Omega_S(i)$ is a function of risk factors associated with the risk of susceptible individual i contracting the disease (susceptibility),
- $\Omega_T(k)$ is a function of risk factors associated with the risk of infection transmission from the k^{th} individual (transmissibility),
- $\kappa(i,k)$ is an infection kernel, a function of risk factors involving a susceptible individul i and infectious individual k, which often describes the connectivity between these individuals, and,
- $\epsilon(i,t)$ is a function of risk factors associated with infection of the i^{th} individual during the t^{th} time period, which is not otherwise explained by the model. This may refer to an infection with a source external to the population. $\epsilon(i,t)$ is also referred to as the *sparks function*.

Transmission network individual level models (TN-ILMs) are an extension of the Deardon et al. (2010) ILM framework. TN-ILMs are described in continuous time, where time periods are variable in length. These time periods, represent the time between consecutive disease state transition events in the population, during which, membership in the susceptible, infected, and removed sets ($S_{(t)}$, $I_{(t)}$, and, $R_{(t)}$ respectively), are unchanged. The length of these time periods are denoted as Δ_t.

TN-ILMs differ from ILMs in their explicitness to sources of disease transmission. In TN-ILMs, competing transmission rates for every susceptible - infected pair of individuals (i and j respectively) are given as:

$$\lambda^*_{SI}(i,j,t) = \Omega_S(i)\Omega_T(j)\kappa(i,j) \text{ for } i \in S_{(t)}, \ j \in I_{(t)}, \tag{2}$$

and with,

$$\lambda^*_{SI}(i,t) = \epsilon^*(i,t) \text{ for } i \in S_{(t)} \tag{3}$$

as the transmission rate for infections originating from any other source (*e.g.* a transmission with an origin external to the population).

Once infected, individuals may transition into a removed state, which can refer to recovery with acquired immunity, death, quarantine, *etc.* This transition rate is given as:

$$\lambda_{IR}(j,t) = \Omega_R(j) \text{ for } j \in I_{(t)} \tag{4}$$

where $\Omega_R(k)$ is a function of risk factors associated with the transition of an infected individual j into the removed state during the t^{th} time period.

Using these transition rates, the likelihood of an SIR TN-ILM is calculated as:

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{T-1} \psi^*(t)\upsilon(t) \exp\left\{-\upsilon(t)\Delta_t\right\}, \tag{5}$$

where,

$$\upsilon(t) = \sum_{i \in S_{(t)}} \sum_{j \in I_{(t)}} \lambda^*_{SI}(i,j,t) + \sum_{i \in S_{(t)}} \lambda^*_{SI}(i,t) + \sum_{j \in I_{(t)}} \lambda_{IR}(j,t), \tag{6}$$

$$\psi^*(t) = \begin{cases} \frac{\lambda^*_{SI}(i,j,t)}{\upsilon(t)} \text{ if } i \in (S_{(t)} \cap I_{(t+1)}) \text{ by transmission from individual } j, \\ \frac{\lambda^*_{SI}(i,t)}{\upsilon(t)} \text{ if } i \in (S_{(t)} \cap I_{(t+1)}) \text{ by exogenous transmission}, \\ \frac{\lambda_{IR}(j,t)}{\upsilon(t)} \text{ if } j \in (I_{(t)} \cap R_{(t+1)}). \end{cases} \tag{7}$$

Simulation of epidemics following a TN-ILM can be accomplished using the Gillespie (1977) algorithm.

Neither exact event times, or the transmission network are likely to be known in an epidemic. These data are required for the application of TN-ILMs, so they must be imputed. This imputation can be done via a data augmentation process within a Bayesian framework. Data augmentation treats latent data as additional parameters to be inferred. The resulting joint posterior distribution is high dimensional, and analytically intractable, necessitating the use of Markov Chain Monte Carlo (MCMC) or other computational approaches to perform inference.

The TN-ILM approach shares similarities to EpiForest of Li et al. (2019), however TN-ILMs are generalized to allow for arbitrary functions of risk factors to describe source specific (*i.e.* parent infection specific) transmission rates with latent parameters. With TN-ILMs, use of such risk functions extends beyond transmission events, to all disease state transitions. TN-ILMs also enable inference of latent event times.

Pathogen.jl (Angevaare et al., 2020) is a package for the Julia language (Bezanson et al., 2018), which provides implementations for simulation and MCMC (with data augmentation) using TN-ILMs. We utilize this package exclusively to perform our simulation study.

3. Simulation Study

We designed a simulation study to evaluate transmission network and event time inference performance, using the TN-ILM simulation and MCMC methods included with Pathogen.jl (Angevaare et al., 2020) for a distance-based SIR model. The parameters of the simulation study were set according to a 2^3 factorial design. The three factors were:

1. two levels for the strength of the spatial mechanism in infection transmission,
2. two levels of population size, and,
3. two levels of population area.

Population density (individuals per unit area) was kept equal across all scenarios.

Seventy-two replicates were generated for each eight simulation scenarios. From these epidemic simulations, observational data were generated. Inclusion of an epidemic simulation was subject to the criteria of that simulation resulting in $\geq 50\%$ of the population contracting the disease. If this criteria was not met, a new epidemic was simulated for that scenario. This criteria ensured adequate observations to use in inference. MCMC was conducted in order to estimate the joint posterior distribution of the TN-ILM parameters, event times, and transmission network, using the same model structure as was used for the epidemic simulations.

3.1. Model structure

A spatial TN-ILM was constructed where the infectivity between two individuals was described by a Euclidean distance based power law function as shown in Equation 8. The susceptibility and transmissibility functions were set equal to 1. Thus,

$$\begin{aligned} \lambda^*_{SI}(i,j,t) =& \Omega_S(i)\Omega_I(k)\kappa(i,j) \text{ for } i \in S_{(t)}, j \in I_{(t)} \\ \Omega_S(i) =& 1.0 \\ \Omega_I(j) =& 1.0 \\ \kappa(i,j) =& \alpha d_{i,j}^{-\beta}, \end{aligned} \tag{8}$$

where,

- α and β are parameters describing the baseline strength and spatial decay of infectious pressure respectively, and,
- $d_{i,j}$ is the Euclidean distance between a susceptible individual i and infectious individual j.

Constant rates were selected for the infectious period with $\Omega_r = \eta$.

Scenario	1, 2	3, 4	5, 6	7, 8
Population size (individuals)	25	100	25	100
Area (units distance)	10×10	20×20	20×5	40×10
Population density	0.25			
Epidemic length (units time)	100.0			

Table 1: The conditions of the eight scenarios of the TN-ILM simulation study

Parameter	**Description**	**Scenario**	
		1, 3, 5, 7	**2, 4, 6, 8**
ϵ	Exogenous infection rate	0.0001	
α	Epidemic strength	0.5	1.0
β	Spatial decay	3.0	5.0
η	Removal rate	0.05	

Table 2: Parameter descriptions and values for the TN-ILM used in the simulation study

3.2. Simulation scenarios

Each simulated epidemic was generated for a population of either 25 or 100 individuals, and all simulations occurred over the same length of 100 time units, as detailed in Table 1. The simulations varied with respect to the TN-ILM parameters, specifically those of the infectivity kernel. The two parameter sets used in the simulation study were selected to give scenarios resulting in epidemics that were either strongly or weakly spatially driven while having comparable overall incidence. The remaining parameter describing the length of the infectious period, was constant across all scenarios. A complete description of all TN-ILM parameters and their values by scenario is found in Table 2.

3.3. Inference

When performing inference, the TN-ILM structure was assumed to be known. That is, inference was performed for an SIR TN-ILM, where the transition rates were described by the same functions as were used during simulation. Priors were set for each of the TN-ILM parameters, and were consistent across all simulation scenarios. The priors were selected to ensure all parameters were positive, relevant for the spatial scale in the simulations, but otherwise fairly uninformative. Flat priors are selected except for ϵ, which receives an exponential prior to favour parametrizations that rely less on exogenous exposures. The priors were:

$$\begin{aligned}
\epsilon &\sim \text{Exponential}(0.0001), \\
\alpha &\sim \text{Uniform}(0.0, 2.0), \\
\beta &\sim \text{Uniform}(1.0, 8.0), \text{ and,} \\
\eta &\sim \text{Uniform}(0.0, 1.0).
\end{aligned}$$

For the event times, a Uniform$(0.0, 5.0)$ prior was selected for the delay in time units between each infection observation and corresponding latent true onset of infectiousness, as well as for the delay between a removal observation and latent true removal time.

For each of the 576 overall simulations, a Markov chain was initialized by selecting the set of parameter values and event times that yielded the highest marginal posterior density out of 100k random samples from the indicated prior distributions, as implemented in Pathogen.jl. An initial transition kernel variance for the TN-ILM parameters was set as a diagonal matrix comprised of the variances of the corresponding prior distributions, which is the default option in Pathogen.jl. Adaptive MCMC (Roberts and Rosenthal, 2007) was employed, which automatically tunes the transition kernel variance and eventually converges to an optimal transition kernel variance for Metropolis-Hastings (Hastings, 1970) MCMC of the TN-ILM parameters.

Event times were updated with 5 block updates per iteration of Metropolis-Hastings MCMC for simulation scenarios with populations of 25, and with 20 block updates for scenarios with populations of 100. This ensured a similar number of events were updated in each block across all simulations. An event time transition kernel variance of 1.0 was selected for all event types across all scenarios. 100k MCMC iterations were performed, with the first 50K of those iterations treated as a burn-in period, and not used to inform the inference. The burn-in period was selected based on visual inspection of full trace plots, which appeared to indicate convergence to the posterior distribution for all model parameters across all simulations by this point. Of the retained 50k iterations, every 10^{th} iteration was assumed to be independent sample from the joint posterior distribution.

Posterior means were used as point estimates of TN-ILM parameters. Accuracy and precision of these point estimates is examined across the various epidemic scenarios. Further, the posterior mean event times and posterior mean transmission network are compared with the true event times and true transmission network with mean squared error.

3.4. Results and discussion

3.4.1. Model parameter inference

Posterior means are shown in Figures 1 through 4 for the various scenarios.

Posterior means of ϵ appear to be close to the true values across all scenarios, as seen in Figure 1. ϵ effectively captures rare events, whether those rare events are true exogenous transmissions, or if they are unlikely endogenous transmissions that the TN-ILM failed to identify. Due to the stochastic nature of simulations, there will be some simulations that are significantly higher in such events, and we expect that these represent the few apparent outliers in Figure 1.

Posterior mean estimates of α and β, the two parameters of our power law infectivity kernel, are shown in Figures 2 and 3, respectively. These parameters tend to be strongly positively correlated with one another, as a relatively low α and β pairing compared to relatively high α and β pairing may describe a similar level of infectious pressure to a susceptible individual. Underestimation of β is consistent with underestimation of, or inability to detect, the spatial dynamics of an epidemic. In Figures 2 and 3, we see a rather consistent underestimation of these parameters across all scenarios, though performance seems to be improved under scenarios 3, 4, 7, 8, where more individuals are being observed. There may be identifability issues with α and β due to our scenario specification. The removal rate in our simulated epidemics results in individuals that remain as possible sources of infection for a large portion of our epidemics. The more concurrent potential infection sources there are, the more difficult it is to determine the true source of infection, and the poorer the identifiability of spatial parameters. There is some indication that if this is an issue, it is lessened in scenarios $5-8$, where the population area is has an aspect ratio of $4:1$, rather than the $1:1$ aspect ratio of the population area in scenarios $1-4$.

Finally, η, described the transition rate between infectious and removed states, and was inferred accurately across all scenarios, as seen in Figure 4. With larger population sizes in scenarios 3, 4, 7 and 8, estimation is improved due to larger sample sizes.

If an epidemic simulation failed to meet the criteria described in Section 3, you may expect that through inference we would have found that these epidemics were more consistent with lower infectious pressure, and/or higher removal rates than what they were truly generated by, due to the stochastic nature of these epidemic simulations. However rare, it should be noted that such epidemic simulations are not represented in these results.

3.4.2. Event times

Under the conditions of our simulation study, the data augmentation process was able to accurately reconstruct epidemics. An example result from the data augmentation process can be seen in Figures 5. In some additional exploratory SEIR simulation and inference it was found that inclusion of an exposed state with weak priors describing the length of the latent period could introduce substantial uncertainty into the data augmentation process.

A 3 factor analysis of variance was also conducted to determine if performance of event time inference varied across population size, population area, or TN-ILM parametrization. Using the posterior means as our point estimates, mean squared error was calculated for each simulation. Using a 1% level of significance, no significant differences were found. Event time mean squared error by simulation scenario is shown in Figure 6.

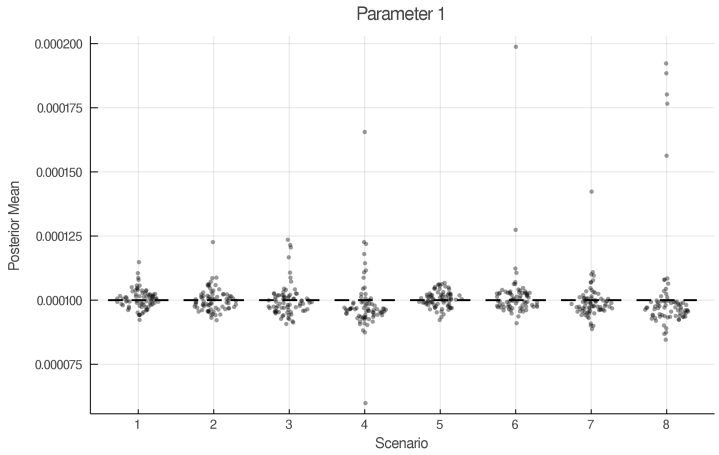

Figure 1: Point estimates for ϵ in comparison to the actual values used in epidemic simulation. Actual values are indicated with horizontal dashed line at 0.0001.

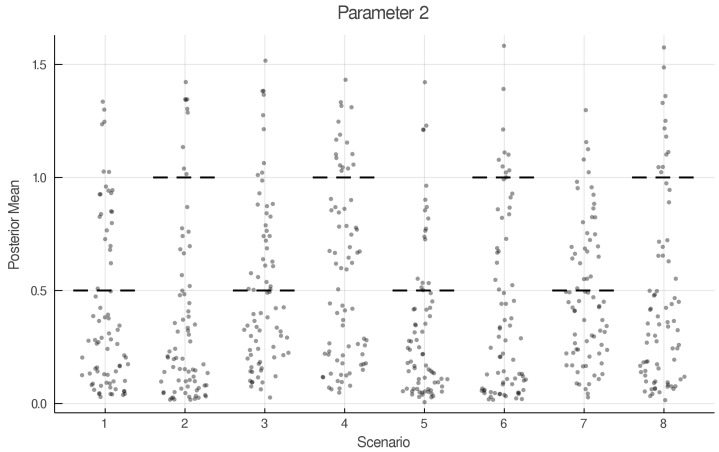

Figure 2: Point estimates for α are shown in comparison to the values set values under each simulation scenario indicated with horizontal dashed lines at 0.5 and 1.0.

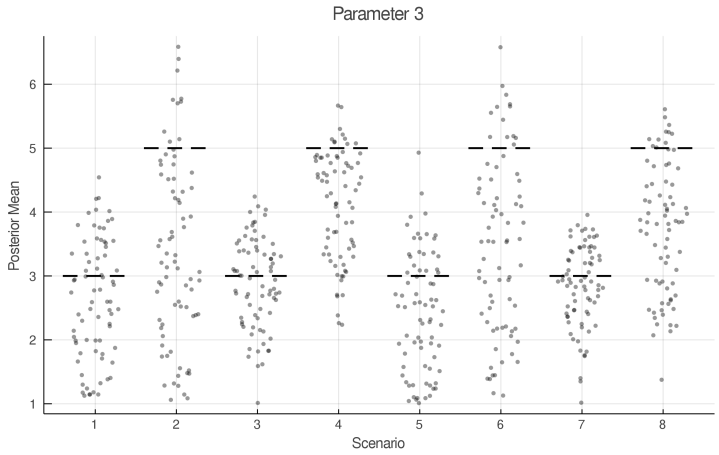

Figure 3: Point estimates for β are shown in comparison to the values set under each simulation scenario, indicated with horizontal dashed lines at 3.0 and 5.0.

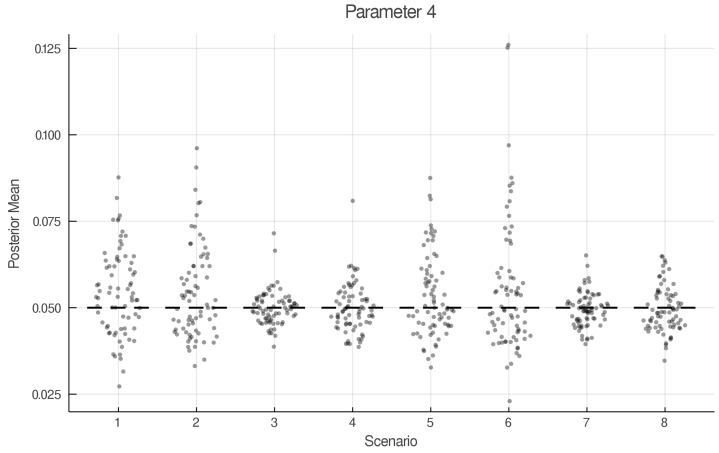

Figure 4: Point estimates for η are shown in comparison to the values set under each simulation scenario, indicated with horizontal dashed lines at 0.05.

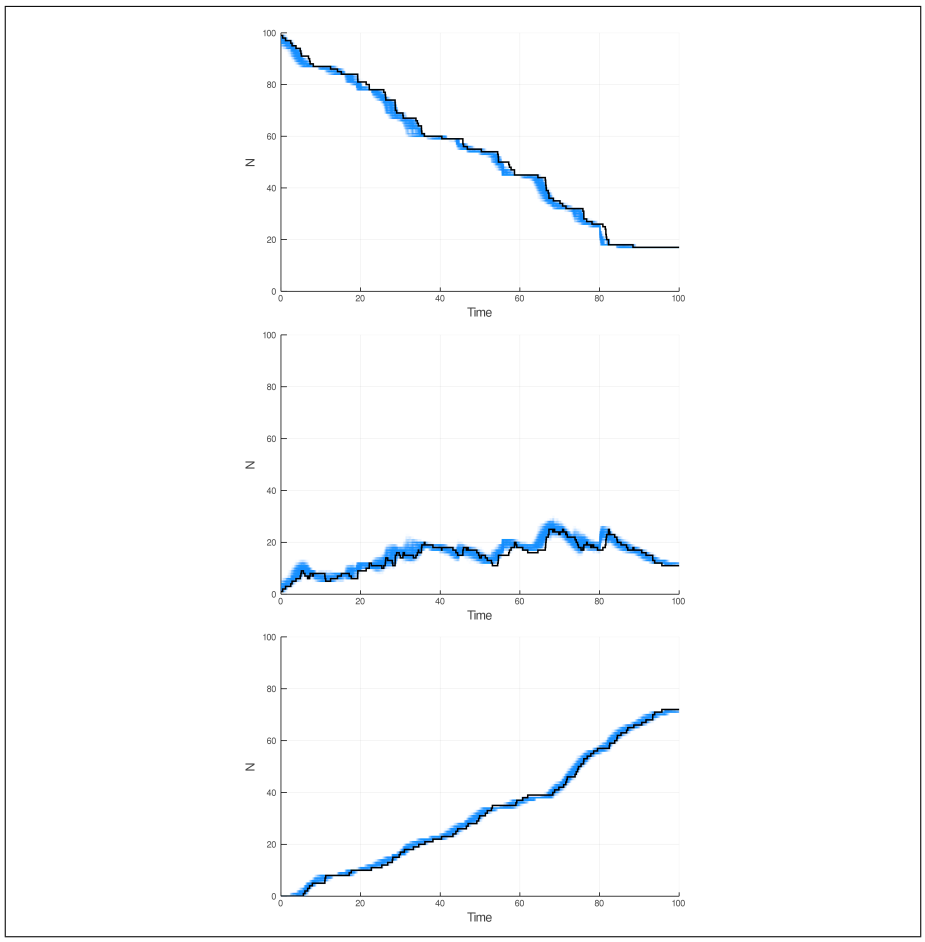

Figure 5: Posterior distribution (blue) vs. actual (black) number of susceptible (top), infectious (middle), and removed (bottom) individuals during one particular scenario 4 simulation.

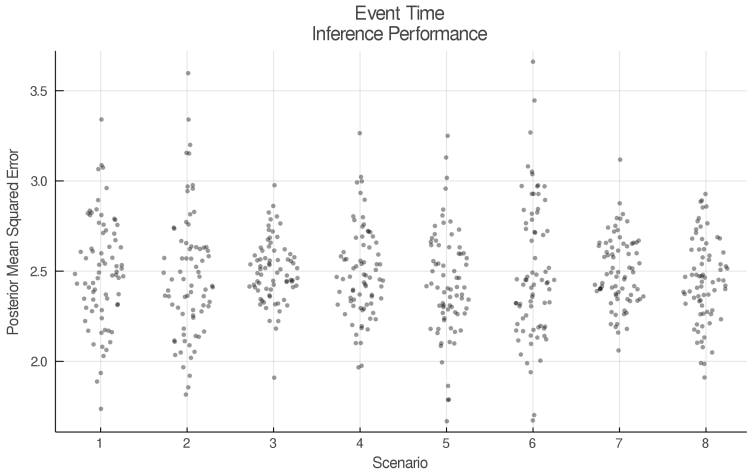

Figure 6: Mean squared error of the estimates of event times and the true event time for each simulation.

3.4.3. Network posterior mean squared error

The posterior mean transmission networks were calculated for each simulated epidemic. An example of such a posterior mean transmission network in shown for a single simulation in Figure 7, along with the true transmission network for that simulation. The posterior mean transmission network from each simulation was compared to the true transmission networks using mean squared error used to quantify their similarity. These mean squared errors are shown in Figure 8. Once again, a 3 factor ANOVA used to determine if performance differed significantly across the different scenarios. Spatial strength of the epidemic was found to be significantly associated with inference performance at the 1% level of significance. Specifically we found that transmission networks were better inferred on average when stronger spatial dynamics were present.

3.4.4. Relation of transmission network and event time performance

Accounting for simulation scenario, we examined whether there was correlation between inference performance of transmission network and of event times. It may be expected that with stronger inference of the transmission network, that inference of event times would be improved as the transmission network constrains event times. However, in our simulation study, such a relationship could not be detected (results not shown).

3.4.5. Computation time

To perform 100K iterations, it required approximately 8 minutes for simulation scenarios that had populations of size 25, and approximately 184 minutes for simulation scenarios that involved populations of size 100. In our simulation

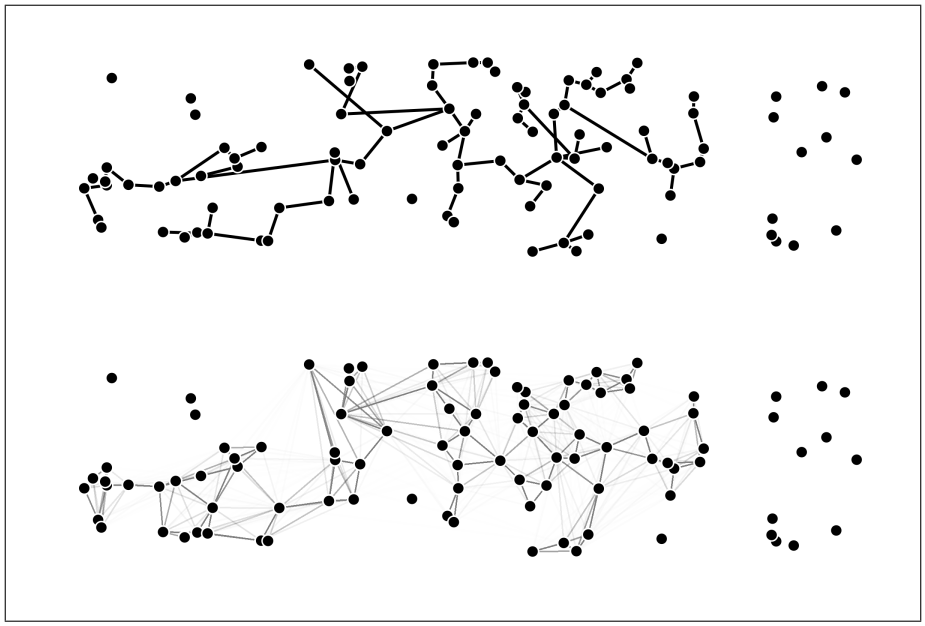

Figure 7: True transmission network (top) and transmission network posterior distribution (bottom) of the same scenario 4 simulation as in Figure 5.

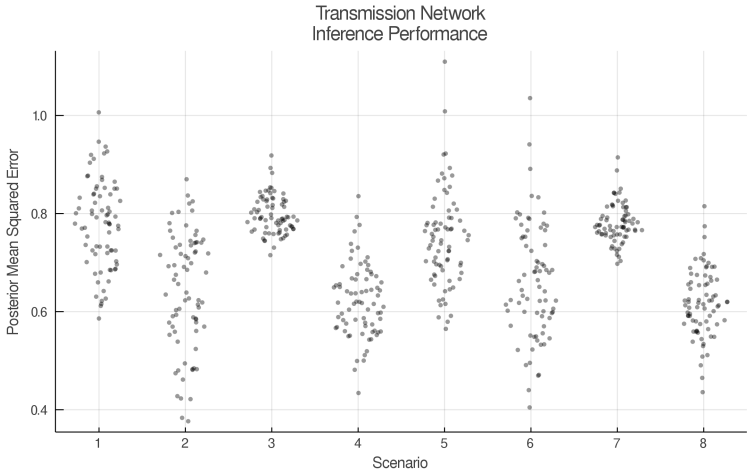

Figure 8: Mean squared error of the transmission network posterior mean for each simulation.

study, we ran ten epidemic simulations simultaneously on an Intel® Xeon® CPU E5-2670 (12 cores @ 2.30GHz), with 64GB 2133MHz RAM.

4. Application: Tomato Spotted Wilt Virus Experiment

4.1. Description

MCMC was performed to fit a TN-ILM to the spread of Tomato Spotted Wilt Virus (TSWV) in a greenhouse experiment presented by Hughes et al. (1997). TSWV is a vector borne plant disease, transmitted by thrips (Hughes et al., 1997). The disease affects over 1000 different plant species, and is responsible for significant economic losses (Parrella et al., 2003). Due to TSWV's host range, and the impact of infection, TSWV has been studied extensively, and specialized models incorporating plant and vector dynamics have been developed to describe TSWV epidemics (*e.g.* Ogada et al. (2016)).

In the TSWV experiment, 520 pepper plants regularly spaced within a 10m x 26m greenhouse were examined for the presence of TSWV once every two weeks. Plants were not removed after showing signs of infection by TSWV. The study concluded after 14 weeks, which saw a total of 327 individual plants infected. Data for this experiment has previously been processed for use as an example dataset in the EpiILM R package (Warriyar K. V. et al., 2020). We used this dataset for our application.

The first infected plant was detected in this experiment at the second time point (day 28). We considered infections detected at or before the third time point (day 42) to be initial infections in order to be sure that the TN-ILM was not applied to the initial seeding event in the experiment.

4.2. Inference

4.2.1. Model structure

We conducted inference for a spatial SI TN-ILM using the TSWV data. Exposed and removed states were not used in the interest of simplicity and computation time.

The only individual level risk data that we had available for the TSWV experimental epidemic was location. We used an SI TN-ILM, that describes infectivity using a Euclidean distance-based power law kernel. Susceptibility, $\Omega_S(i)$, and transmissibility, $\Omega_T(j)$, functions were set equal to 1.0. Thus, a susceptible plant, i, is infected by an infectious plant j, at rate $\lambda^*_{SI}(i,j,t)$, where

$$\begin{aligned}
\lambda^*_{SI}(i,j,t) =& \Omega_S(i)\Omega_I(j)\kappa(i,j) \text{ for } i \in S_{(t)}, j \in I_{(t)} \\
\Omega_S(i) =& 1.0 \\
\Omega_T(j) =& 1.0 \\
\kappa(i,j) =& \alpha d_{i,j}^{-\beta}, \qquad (9)
\end{aligned}$$

where $d_{i,j}$ is the Euclidean distance between plants i and j.

As the TSWV experiment epidemic was conducted in a closed greenhouse environment, the exogenous transmissions where assumed to occur with rate

$$\lambda^*_{SI}(i,t) = 0.0.$$

Marginal prior distributions were selected to provide the appropriate support to the model parameters, but be otherwise rather uninformative towards TN-ILM parametrizations that would describe epidemics ranging from having a strong spatial signal (*i.e.* relatively high values of α and β in Equation 9), to those with weaker spatial signals (*i.e.* relatively low values of α and β). Specifically the prior distributions were selected to be:

- $\alpha \sim \text{Gamma}(1.0, 1.0)$, and
- $\beta \sim \text{Gamma}(1.0, 0.5)$.

The augmented times of infection onset were assumed to occur within the 2 week period between plant inspections. Specifically, a $\text{Uniform}(0.0, 14.0)$ prior distribution was assumed for the observation delay in days for each record of infectiousness.

4.2.2. MCMC

Four Markov chains were initialized by again, selecting the parameter and augmented event time datasets that provided the highest marginal posterior densities from 50k generations from the prior distributions.

Each of the four Markov chains were run for 50k iterations. Event time data augmentation was completed in 20 batches of updates per iteration. The exact composition of these event time batches was randomly generated at each iteration. The Adaptive Metropolis-Hastings approach of Roberts and Rosenthal (2007) was utilized to tune the proposal distribution for the TN-ILM parameters. Convergence to the posterior distribution of the TN-ILM was determined by visual assessment of trace plots. Visual assessment was also used to determine an appropriate burn-in period.

4.3. Results and discussion

4.3.1. Assessing convergence

Trace plots for all four chains for both SI TN-ILM parameters, α and β, appeared to show rapid convergence, as seen in Figure 9. A burn-in period of 25k iterations was selected as a conservative choice that would still leave sufficient samples for inference. The posterior means, and variances, and effective sample sizes for each of the model parameters were calculated for each Markov chain and found to be very similar (not shown) - further evidencing convergence of TN-ILM parameters to their posterior distribution.

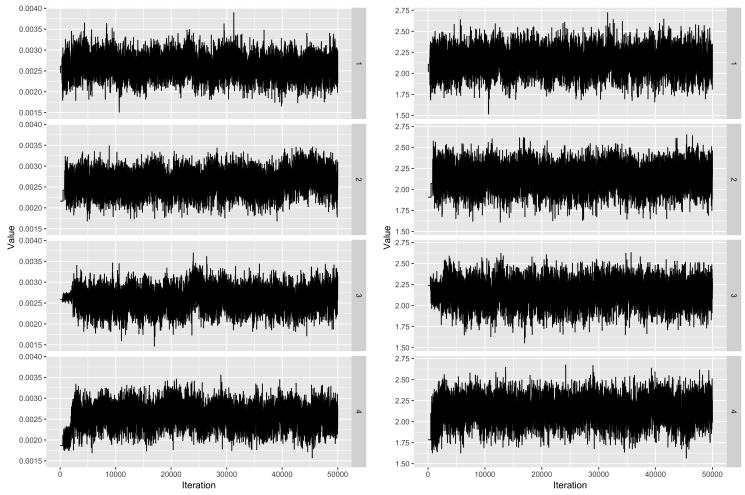

Figure 9: Trace plots from 4 independent Markov Chains for inference of α (left), and β (right) parameters of a power law infectivity kernel for the TSWV experimental epidemic. Trace plots consistently exhibit quick convergence to the posterior distribution. The posterior distributions of α and β appear distinct from their priors of Gamma$(1.0, 1.0)$ and Gamma$(1.0, 0.5)$, respectively.

It is difficult to visually assess convergence of some of the higher dimensional aspects of TN-ILMs involving data augmentation. For the transmission network posterior distribution, we considered the degree distributions of the mean transmission networks for each chain. For a transmission network, the out-degree represents the number of transmissions, or secondary cases from each infected individual. The out-degree distribution is one way of summarizing a transmission network. Similarity amongst independent chains with respect to degree distribution, as seen in Figure 10, supports that convergence had also occurred for the transmission network themselves.

4.3.2. Parameter estimation

By calculating the mean from the retained samples of each of four Markov chains, point estimates of 2.60×10^{-3} and 2.15 were found for α and β respectively. The spatial decay parameter, β, may be higher than had a removed class had been involved in the model - which may be a more realistic model for a TSWV epidemic. A higher spatial decay would effectively reduce the impact of earlier incidences on latter transmissions in a strongly spatially driven epidemic.

4.3.3. Event time data augmentation

The original TSWV experiment data was to a temporal resolution of 2-weeks. Using event time data augmentation, posterior distributions for exact time of onset of infectiousness were generated for each incidence. The posterior distribution for the epidemic curve can be visualized, as in Figure 11. Posterior

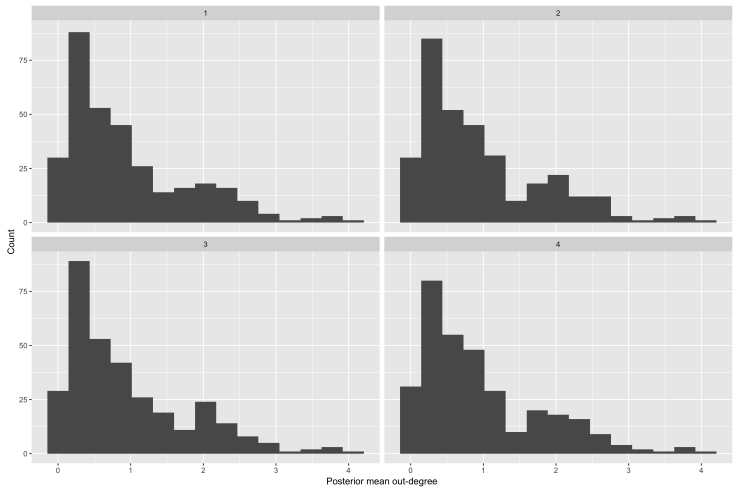

Figure 10: Posterior mean out-degrees for individual plants part of the TSWV experimental epidemic for each of four independent Markov chains. Each Markov chain displays a similar distribution, supporting that they have converged.

means can also be calculated in order to produce a posterior mean epidemic curve.

4.3.4. Computation time

To complete 50k iterations across 4 chains, for an epidemic involving a population of 520, Pathogen.jl required approximately 78.9 hrs. With using 20 event batches per iteration, this represents a total of 4.2 million TN-ILM likelihood calculations for the TSWV experimental epidemic.

5. Future work

We have introduced TN-ILMs, and have shown that inference can be successfully carried out using these models with data augmented MCMC, however, many avenues for future work are possible. First, our simulation study represents an investigation into inference of latent transmission network and event times in spatially driven TN-ILMs. However our metric for quantifying similarity of transmission network posterior estimates to true transmission networks did not consider closeness of incorrect transmission paths. For instance, spatial distance from true transmission source to other potential transmission sources weighted by the posterior density may allow us to say more about transmission network inference in spatially driven epidemics. We could have also considered degree-separation from true transmission sources in a similar manner. Other transmission network summary statistics may be considered as well; even if

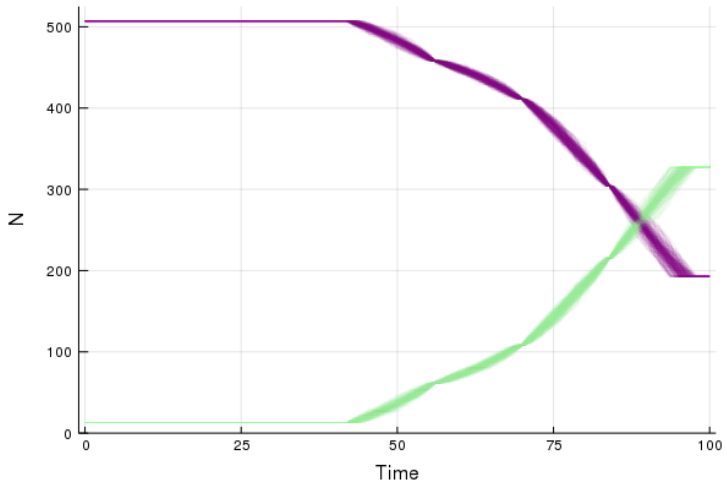

Figure 11: Visualization of the event time posterior distribution, where opacity represents posterior densities for the epidemic curve. The purple and green lines display the count of individuals in the susceptible, and infected class through time, respectively.

transmission sources were not consistently and reliably identified, it is of interest if global network statistics, such as degree distributions, were accurately reconstructed in the data augmentation process.

In the scenarios studied here, we found inference for parameters, transmission networks, and event times all worked well. However, it would be worthwhile to explore further TN-ILM structures and parametrizations in investigation into the relationship between the inference of transmission networks and event times.

With exploratory simulations and an understanding of implementation of TN-ILMs in Pathogen.jl, we can determine that there is a steep increase in computational requirements in applying TN-ILMs to larger populations. The impact of population size on inference, however, is less clear. Further simulation studies could be conducted with additional population sizes across a variety of TN-ILMs to develop general guidance for application of TN-ILMs with regards to population size and model complexity.

As our understanding of uncertainty involved in joint inference of event times, transmission networks, and TN-ILM parameters improves, our confidence in using these models to inform risk-based control measures in real time should increase. Performance of precision control measures possible through TN-ILMs should thus be compared to those derived from traditional ILMs in future work to ascertain the value of the extra complexity involved in a TN-ILM based analysis.

While the TN-ILMs explored here have been spatial, they have not represented any geography-dependent risk factors (*e.g.* environmental, or social determinants of health). Recent work by Mahsin et al. (2020) has demonstrated how latent, spatially structured covariates can be incorporated into ILMs. This work could be expanded to TN-ILMs, and implemented into a future version of Pathogen.jl.

Continued improvement to computational performance - with the Julia language, and the Pathogen.jl package will further motivate their use. Currently, data augmentation of transmission networks and event times for large populations can be time consuming, as seen with our TSWV experimental epidemic analysis. While Julia is a high level language, and writing functions following the signature expected of Pathogen.jl is not difficult, implementations of common risk functions into the Pathogen.jl package itself, such as power law infectivity kernel, or macros to assist in their construction, may also increase the package's accessibility.

6. Acknowledgements

We thank two anonymous reviewers for their constructive feedback. This research was funded via a Highly Qualified Personnel (HQP) scholarship from the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) / University of Guelph Partnership, as well as from Dr. Feng's and Dr. Deardon's Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants.

References

Abdulkareem, S.A., Augustijn, E.W., Filatova, T., Musial, K., Mustafa, Y.T., 2020. Risk perception and behavioral change during epidemics: Comparing models of individual and collective learning. PloS One URL: https://doi.org/10.1371/journal.pone.0226483.

Angevaare, J., Feng, Z., Deardon, R., 2020. jangevaare/Pathogen.jl: v0.4.5. URL: https://doi.org/10.5281/zenodo.3703648.

Azizi, A., Montalvo, C., Espinoza, B., Kang, Y., Castillo-Chavez, C., 2020. Epidemics on networks: Reducing disease transmission using health emergency declarations and peer communication. Infectious Disease Modelling 5, 12–22. URL: https://doi.org/10.1016/j.idm.2019.11.002.

Bezanson, J., Chen, J., Chung, B., Karpinski, S., Shah, V.B., Vitek, J., Zoubritzky, L., 2018. Julia: Dynamism and performance reconciled by design. Proceedings of the ACM on Programming Languages 2, 120:1–120:23. URL: https://doi.org/10.1145/3276490.

Deardon, R., Brooks, S.P., Grenfell, B.T., Keeling, M.J., Tildesley, M.J., Savill, N.J., Shaw, D.J., Woolhouse, M.E., 2010. Inference for individual-level models of infectious diseases in large populations. Statistica Sinica 20, 239–261.

Giardina, F., Romero-Severson, E.O., Albert, J., Britton, T., Leitner, T., 2017. Inference of transmission network structure from hiv phylogenetic trees. PLoS Computational Biology 13, e1005316. URL: https://doi.org/10.1371/journal.pcbi.1005316.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry 81, 2340–2361. URL: https://doi.org/10.1021/j100540a008.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109. URL: https://doi.org/10.1093/biomet/57.1.97.

Hughes, G., McRoberts, N., Madden, L.V., Nelson, S.C., 1997. Validating mathematical models of plant-disease progress in space and time. Mathematical Medicine and Biology: A Journal of the IMA 14, 85–112.

Li, M., Shi, X., Li, X., Ma, W., He, J., Liu, T., 2019. Epidemic Forest: A spatiotemporal model for communicable diseases. Annals of the American Association of Geographers 109, 812–836.

Mahsin, M., Deardon, R., Brown, P., 2020. Geographically dependent individual-level models for infectious diseases transmission. Biostatistics , 1–17URL: https://doi.org/10.1093/biostatistics/kxaa009.

Moshiri, N., Ragonnet-Cronin, M., Wertheim, J.O., Mirarab, S., 2019. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. Bioinformatics 35, 1852–1861. URL: https://doi.org/10.1093/bioinformatics/bty921.

Ogada, P.A., Moualeu, D.P., Poehling, H., 2016. Predictive models for tomato spotted wilt virus spread dynamics, considering *frankliniella occidentalis* specific life processes as influenced by the virus. PloS One 11, e0154533. URL: https://doi.org/10.1371/journal.pone.0154533.

Parrella, G., Gognalons, P., Gebre-Selassiè, K., Vovlas, C., Marchoux, G., 2003. An update of the host range of tomato spotted wilt virus. Journal of Plant Pathology 85, 227–264. URL: https://www.jstor.org/stable/41998156.

Pipatsart, N., Triampo, W., Modchang, C., 2017. Stochastic models of emerging infectious disease transmission on adaptive random networks. Computational and Mathematical Methods in Medicine 2017, 1–11. URL: https://doi.org/10.1155/2017/2403851.

Roberts, G.O., Rosenthal, J.S., 2007. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. Journal of Applied Probability 44, 458–475. URL: https://doi.org/10.1239/jap/1183667414.

Sun, M., Zhang, H., Kang, H., Zhu, G., Fu, X., 2017. Epidemic spreading on adaptively weighted scale-free networks. Journal of Mathematical Biology 74, 1263–1298. URL: https://doi.org/10.1007/s00285-016-1057-6.

Warriyar K. V., V., Almutiry, W., Deardon, R., 2020. Individual-level modelling of infectious disease data: EpiILM. The R Journal 12, 87–104. URL: https://doi.org/10.32614/RJ-2020-020.