Advancements in Automated Disease Outbreak Detection: A Comparative Simulation Study of a Novel Method against State-of-the-Art Approaches

Kasper Schou Telkamp[a,1,*], Lasse Engbo Christiansen[a,1], Jan Kloppenborg Møller[b,2]

[a]*Epidemiology Research, Statens Serum Institut, Artillerivej 5, Copenhagen S, 2300, Denmark*
[b]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Asmussens Allé, Building 303B, Lyngby, 2800, Denmark*

Abstract

Background: Here is a long text that tells me about the background of this article. I want to make this text long, so I can illustrate how the abstract is formatted. *Methods*: These methods have been crucial in the development of this outbreak detection algorithm. *Results*: These results are outstanding and will forever change the way we employ statistical outbreak detection. *Conclusion*: Please use my method, as it will result in significant advancements within disease outbreak detection.

Keywords: generalized mixed effects models, hiearchical generalized lionear models, outlier, outbreak, statistical surveillance

1. Introduction

The fight against infectious disease not only requires proper treatment of patients and implementation of preventive measures but also demands early detection of emerging disease outbreaks. Today epidemiologists and public health professionals utilize a range of tools and methodologies to effectively tackle disease outbreaks. Here, laboratory-based approaches play a crucial role in outbreak investigations and may involve techniques such as molecular epidemiology (Honardoost et al., 2018, Struelens and Brisse (2013)) and, more recently, Whole Genome Sequencing (WGS) (Koeser et al., 2012, Baldry (2010)).

However, in recent years, there has been a growing interest in statistical methods for automated and early detection of disease outbreak. Statistical outbreak detection begins with the identification of an aberrant number of cases of a particular disease within a specific time and space. When an increase in the number of cases is detected, a signal or alarm is raised by the detection method. Subsequently, an epidemiologist assesses the public health relevance of the aberration to determine if further investigation is warranted. These methodologies encompass various statistical techniques, including regression analysis, time series methodology, methods inspired by statistical process control, approaches incorporating spatial information, and multivariate outbreak detection. A comprehensive review of these methods can be found in studies by Buckeridge (2007) and Unkel et al. (2012).

To establish a golden standard, this article will focus on the method initially proposed by Farrington et al. (1996) and the subsequent improvements proposed by Noufaily et al. (2013). These methods offer advanced statistical tools for detecting and monitoring disease outbreaks and are currently *the* methods of choice at European public health institutes (Hulth et al., 2010). They can be accessed through the R package called `surveillance` developed by Salmon et al. (2016). It is a well known fact, that one limitation

*Corresponding author
Email addresses: `ksst@ssi.dk` (Kasper Schou Telkamp), `lsec@ssi.dk` (Lasse Engbo Christiansen), `jkmo@dtu.dk` (Jan Kloppenborg Møller)
[1]This is the first author footnote.
[2]Another author footnote.

of these detection algorithms is an occasional lack of specificity, leading to false alarms that can overwhelm the epidemiologists with verification tasks (Bédubourg and Strat, 2017). Therefore, in this article, these established methods will be compared to a novel outbreak detection algorithm based on hierarchical models. This article introduces this new algorithm as an innovative approach to outbreak detection and aims to assess its performance in comparison to already existing methods.

While hierarchical models have earned a reputation within ecology (Bolker et al., 2009, Zuur et al. (2009)), urban energy modeling (Real et al., 2021, Palmer Real et al. (2022)), and other fields, their application in the automatic detection of disease outbreaks is relatively unproved. However, there is a promising paper by Heisterkamp et al. (2006) that applied a hierarchical time series model to detect infectious disease outbreaks in empirical data from *Rubella* and *Salmonella*. The authors concluded that the method is a powerful and versatile way of analyzing time series of routinely recorded laboratory data. In the context of this article, the focus will be on the prospective detection of disease outbreaks, considering the potential of hierarchical models to effectively identify and respond to emerging outbreaks in a timely manner.

2. Novel outbreak detection algorithm

The novel algorithm utilizes a generalized mixed effects model or a hierarchical generalized linear model as a modeling framework to model the count observations y and assess the unobserved random effects u. These random effects are used directly in the detection algorithm to characterize an outbreak. The theoretical foundations of these models will be further discussed in Sections 2.1 and 2.2.

The first step involves fitting either a hierarchical Poisson Normal or Poisson Gamma model with a log link to the reference data. Here, it is possible to include an arbitrary number of covariates by supplying a model formula. In order to account for structural changes in the time series, e.g. an improved and more sensitive diagnostic method or a new screening strategy at hospitals, a rolling window with width k is used to estimate the time-varying model parameters. Also, it is assumed that the count is proportional to the population size n. Hence in terms of the canonical link the model for the fixed effects is

$$\log(\lambda_{it}) = x_{it}\beta + \log(n_{it}), \quad i = 1, \ldots, m, \quad t = 1, \ldots, T \tag{1}$$

Here x_{it} and β are p-dimensional vectors of covariates and fixed effects parameters respectively, where p denotes the number of covariates or fixed effects parameters, m denotes the number of groups, and T denotes the length of the period.

In the second step of the algorithm, as a new observation becomes available, the algorithm infers the one-step ahead random effect u_{it_1} for each group using the obtained model estimates θ_{t_0}. Here, t_0 represents the current time point, and t_1 represent the one-step ahead time points. The threshold U_{t_0} for detecting outbreak signals is defined as a quantile of the distribution of random effects in the second stage model. This threshold can be calculated based on either a Gaussian distribution using the plug-in estimate $\hat{\sigma}_{t_0}$ or a Gamma distribution using the plug-in estimate $\hat{\phi}_{t_0}$. The choice of distribution depends on the specific modeling framework and assumptions used in the analysis.

In the final step, the inferred random effect $\hat{u}_{it_1}$ is compared to the upper bound U_{t_0}, and an alarm is raised if $\hat{u}_{it_1} > U_{t_0}$. If an outbreak is detected, the related observation y_{it_1} is omitted from the parameter estimation in the future. Thus, resulting in a smaller sample size for the rolling window until that specific observation is discarded.

2.1. Generalized mixed effects models

The generalized mixed effects model can be represented by its likelihood function

$$L_M(\theta; y) = \int_{\mathbb{R}^q} L(\theta; u, y) du \tag{2}$$

where y is the observed random variable, θ is the model parameters to be estimated and U is the q unobserved random variables. The likelihood function L is the joint likelihood of both the observed and the unobserved random variables. The likelihood function for estimating θ is the marginal likelihood L_M

obtained by integrating out the unobserved random variables. In general it is difficult to solve the integral in (2) if the number of unobserved random variables is more than a few and hence numerical methods must be used. Thus, an outline of the Laplace approximation is included in this section.

2.1.1. Hiearchical models

It is useful to formulate the model as a hierarchical model containing a *first stage model*

$$f_{Y|u}(y; u, \beta) \tag{3}$$

which is a model for the observed random variables given the unobserved random variables, and a *second stage model*

$$f_U(u; \Psi) \tag{4}$$

which is a model for the unobserved random variables. Here β represent the fixed effects parameters and Ψ is a model parameter. The total set of parameters is $\theta = (\beta, \Psi)$. Hence the joint likelihood is given as

$$L(\beta, \Psi; u, y) = f_{Y|u}(y; u, \beta) f_U(u; \Psi) \tag{5}$$

To obtain the likelihood for the model parameters (β, Ψ) the unobserved random variables are integrated out. The likelihood function for estimating (β, Ψ) is as in (2) the marginal likelihood

$$L_M(\beta, \Psi; y) = \int_{\mathbb{R}^q} L(\beta, \Psi; u, y) du \tag{6}$$

where q is the number of unobserved random variables, and β and Ψ are the parameters to be estimated.

2.1.2. Laplace approximation

The Laplace approximation will be outlined in the following. A thorough description of the Laplace approximation in nonlinear mixed effects models is found in Wolfinger and Lin (1997).

For a given set of model parameters θ the joint log-likelihood $\ell(\theta, u, y) = \log\big(L(\theta, u, y)\big)$ is approximated using a second order Taylor approximation around the optimum $\tilde{u} = \hat{u}_\theta$ of the log-likelihood function w.r.t. the unobserved random variables u, i.e.,

$$\ell(\theta, u, y) \approx \ell(\theta, \tilde{u}, y) - \frac{1}{2}(u - \tilde{u})^T H(\tilde{u})(u - \tilde{u}) \tag{7}$$

where the first-order term of the Taylor expansion disappears since the expansion is done around the optimum $\tilde{u}$ and $H(\tilde{u}) = -\ell''_{uu}(\theta, u, y)|_{u=\tilde{u}}$ is the negative Hessian of the joint log-likelihood evaluated at $\tilde{u}$.

It is readily seen that the joint log-likelihood for the hierarchical model specified in 3 and 4 is

$$\ell(\theta, u, y) = \ell(\beta, \Psi, u, y) = \log f_{Y|u}(y; u, \beta) + \log f_U(u; \Psi) \tag{8}$$

which implies that the Laplace approximation becomes

$$\ell_{M,LA}(\theta, y) = \log f_{Y|u}(y; \tilde{u}, \beta) + \log f_U(\tilde{u}, \Psi) - \frac{1}{2} \log \left| \frac{H(\tilde{u})}{2\pi} \right| \tag{9}$$

2.1.3. Formulation of the generalized mixed effects model

The generalized mixed effects model utilized in the novel outbreak detection algorithm is formulated as a hierarchical Poisson Normal model. This section presents the joint likelihood function for the first and second stage models.

In order to simplify the notation, the probability density functions are presented for a specific observation. Hence, the subscripts indicating the group and time are omitted. The conditional distribution of the count observations is assumed to be a Poisson distribution with intensities λ

$$f_{Y|u}(y;u,\beta) = \frac{\lambda\exp(u)^y}{y!}\exp\big(-\lambda\exp(u)\big) \tag{10}$$

Also, it is assumed that the count is proportional to the population size n. Hence, in terms of the canonical link for the Poisson distribution the model for the fixed effects is

$$\log(\lambda_{it}) = x_{it}\beta + \log(n_{it}), \quad i = 1,\ldots,m, \quad t = 1,\ldots,T \tag{11}$$

The probability density function for the distribution of the random effects is assumed to follow a zero mean Gaussian distribution, $u \sim \mathrm{N}(0, I\sigma^2)$, i.e.

$$f_U(u;\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{u^2}{2\sigma^2}\right) \tag{12}$$

where σ is a model parameter.

Henceforth, the total set of parameters are $\theta = (\beta, \sigma)$ and the model can be formulated as a two-level hierarchical model

$$Y|u \sim \mathrm{Pois}\big(\lambda\exp(u)\big) \tag{13a}$$

$$u \sim \mathrm{N}(0, I\sigma^2) \tag{13b}$$

The joint likelihood for the count observations y and the random effects u becomes

$$L(\beta,\sigma;u_{it},y_{it}) = \prod_{t=1}^{T}\prod_{i=1}^{m}\frac{\big(\lambda_{it}\exp(u_{it})\big)^{y_{it}}}{y_{it}!}\exp\big(-\lambda_{it}\exp(u_{it})\big)\prod_{t=1}^{T}\prod_{i=1}^{m}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{u_{it}^2}{2\sigma^2}\right) \tag{14}$$

2.2. Hiearchical generalized linear models

In this section selected theory related to hierarchical generalized linear models is presented. The model class was initially formulated by Lee and Nelder (1996) as a natural generalization of the generalized linear model to also incorporate random effects. A starting point in hierarchical modelling is an assumption that the distribution of random effects may be modeled by an exponential dispersion family. This family of models were first introduced by Fisher and Russell (1922), and has proven to play an important role in mathematical statistics because of their simple inferential properties. The exponential dispersion family considers a family of distributions, which can be written on the form

$$f_Y(y;\theta) = c(y,\phi)\exp\big(\phi\{\theta y - \kappa(\theta)\}\big) \tag{15}$$

Here the parameter $\phi > 0$ is called the *precision parameter*, which in some cases represents a shape parameter as for the Gamma distribution. In other cases the precision parameter represents an over-dispersion that is not related to the mean. These distributions combine with the so-called *standard conjugate distributions* in a simple way, and lead to marginal distributions that may be expressed in a closed form suited for likelihood calculations. For an introduction to the concept of *standard conjugate distributions* and the definition of a hierarchical generalized linear model, refer to Section 6.3 and Section 6.5 of Madsen and Thyregod (2011), respectively.

2.2.1. Formulation of the hiearchical generalzied linear model

The hierarchical generalized linear model used by the novel outbreak detection algorithm is formulated as a hierarchical Poisson Gamma model. This section present the derivation of the marginal distribution of Y along with the joint likelihood function for the first and second stage models.

In the compound Poisson Gamma model the conditional distribution of the count observations are assumed to be a Poisson distribution with intensities λ

$$f_{Y|u}(y;u,\beta) = \frac{(\lambda u)^y}{y!}\exp(-\lambda u) \tag{16}$$

The probability density function for the random effects u are assumed to follow a reparametrized Gamma distribution with mean 1, $u \sim \mathrm{G}(1/\phi, \phi)$ that is

$$f_u(u;\phi) = \frac{1}{\phi\Gamma(1/\phi)}\left(\frac{u}{\phi}\right)^{1/\phi-1}\exp(-u/\phi) \tag{17}$$

Subsequently, the model can be formulated as a two-level hierarchical model

$$Y|u \sim \mathrm{Pois}(\lambda u) \tag{18a}$$
$$u \sim \mathrm{G}(1/\phi, \phi) \tag{18b}$$

Given 16 and 17, the probability function for the marginal distribution of Y is determined from

$$\begin{aligned} g_Y(y;\beta,\phi) &= \int_{u=0}^{\infty} f_{Y|u}(y;u,\beta) f_u(u;\phi)\, du \\ &= \int_{u=0}^{\infty} \frac{(\lambda u)^y}{y!}\exp(-\lambda u)\frac{1}{\phi\Gamma(1/\phi)}\left(\frac{u}{\phi}\right)^{1/\phi-1}\exp(-u/\phi)\, du \\ &= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}}\int_{u=0}^{\infty} u^{y+1/\phi-1}\exp\big(-u(\lambda\phi+1)/\phi\big)\, du \end{aligned} \tag{19}$$

In 19 it is noted that the integrand is the kernel in the probability density function for a Gamma distribution, $\mathrm{G}\big(y+1/\phi, \phi/(\lambda\phi+1)\big)$. As the integral of the density shall equal one, it is found by adjusting the norming constant that

$$\int_{u=0}^{\infty} u^{y+1/\phi-1}\exp\Big(-u/\big(\phi/(\lambda\phi+1)\big)\Big)\, du = \frac{\phi^{y+1/\phi}\Gamma(y+1/\phi)}{(\lambda\phi+1)^{y+1/\phi}} \tag{20}$$

Therefore, it is shown that the marginal distribution of Y is a Negative Binomial distribution, $Y \sim \mathrm{NB}\big(1/\phi, 1/(\lambda\phi+1)\big)$. The probability function for Y is

$$\begin{aligned} P[Y=y] &= g_Y(y;\beta,\phi) \\ &= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}}\frac{\phi^{y+1/\phi}\Gamma(y+1/\phi)}{(\lambda\phi+1)^{y+1/\phi}} \\ &= \frac{\Gamma(y+1/\phi)}{\Gamma(1/\phi)y!}\frac{1}{(\lambda\phi+1)^{1/\phi}}\left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y \\ &= \binom{y+1/\phi-1}{y}\frac{1}{(\lambda\phi+1)^{1/\phi}}\left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y, \quad \text{for } y = 0,1,2,\ldots \end{aligned} \tag{21}$$

where the following convention is used

$$\binom{z}{y} = \frac{\Gamma(z+1)}{\Gamma(z+1-y)y!} \tag{22}$$

for z real and y integer values. Consequently, the mean and variance of Y are given by

$$\mathrm{E}[Y] = \lambda \qquad \mathrm{V}[Y] = \lambda(\lambda\phi+1) \tag{23}$$

The joint likelihood function for estimating (β, ϕ) is

$$L(\beta,\phi;y_{it}) = \prod_{t=1}^{T}\prod_{i=1}^{m}\binom{y_{it}+1/\phi-1}{y_{it}}\frac{1}{(\lambda_{it}\phi+1)^{1/\phi}}\left(\frac{\lambda_{it}\phi}{\lambda_{it}\phi+1}\right)^{y_{it}} \tag{24}$$

2.2.2. Inference on individual groups

Consider the compound Poisson Gamma model in 18, and assume that a value $Y = y$ has been observed.

The conditional distribution of u for given $Y = y$ is found using Bayes Theorem. In order to simplify the notation, the subscript indicating the group and time are omitted.

$$\begin{aligned} g_u(u|Y=y) &= \frac{f_{y,u}(y,u)}{g_Y(y;\lambda,\phi)} \\ &= \frac{f_{y|u}(y;u)g_u(u)}{g_Y(y;\lambda,\phi)} \\ &= \frac{1}{g_Y(y;\lambda,\phi)}\left(\frac{(\lambda u)^y}{y!}\exp(-\lambda u)\frac{1}{\phi\Gamma(1/\phi)}\left(\frac{u}{\phi}\right)^{1/\phi-1}\exp(-u/\phi)\right) \\ &\propto u^{y+1/\phi-1}\exp\big(-u(\lambda\phi+1)/\phi\big) \end{aligned} \tag{25}$$

Here, the *kernel* of the probability density function is identified

$$u^{y+1/\phi-1}\exp(-u(\lambda\phi+1)/\phi) \tag{26}$$

as the kernel of a Gamma distribution, $\mathrm{G}(y+1/\phi, \phi/(\lambda\phi+1))$, i.e. the conditional distribution of u for given $Y = y$ can be written as

$$u|Y=y \sim \mathrm{G}\big(y+1/\phi, \phi/(\lambda\phi+1)\big) \tag{27}$$

The mean of the conditional distribution is given by:

$$\mathrm{E}[u|Y=y] = \frac{y\phi+1}{\lambda\phi+1} \tag{28}$$

And the variance of the conditional distribution is:

$$\mathrm{V}[u|Y=y] = \frac{(\phi^2+\phi)}{(\lambda\phi+1)^2} \tag{29}$$

These formulas provide the mean and variance of the conditional distribution of u given the observed value $Y = y$.

2.2.3. The rationale for employing the Gamma distribution as a second stage model

The choice of the Gamma distribution for modeling the random effects has been motivated by several reasons. Firstly, the support of the Gamma distribution, which ranges from 0 to infinity, aligns with the mean-value space, denoted as $\mathcal{M}$, for the Poisson distribution. This ensures that the random effects are constrained within a meaningful range for the underlying Poisson process.

Secondly, the two-parameter family of Gamma distributions offers considerable flexibility, encompassing a wide range of shapes and distributions that can span from exponential-like distributions to fairly symmetrical distributions on the positive real line. This flexibility allows the model to capture various patterns and characteristics observed in the data.

Additionally, the choice of the Gamma distribution has benefits in terms of the derivation of the marginal distribution of the response variable Y. The kernel $u^{\alpha-1}\exp(-u/\beta)$ of the Gamma distribution used for modeling the random effects exhibits a similar structure to the kernel $u^y\exp(-u)$ of the likelihood function corresponding to the sampling distribution of Y. This similarity facilitates the analytical computation of the integral involved in deriving the marginal distribution, as it can be expressed in terms of known functions.

Overall, the Gamma distribution is selected due to its alignment with the mean-value space of the Poisson distribution, its flexibility in capturing diverse distributions, and its analytical convenience in computing the marginal distribution of the response variable.

3. Parameter estimation

The generalized mixed effect and hierarchical generalized linear models are implemented in R using the open-source R package `TMB` (Template Model Builder) developed by Kristensen et al. (2016). This package facilitates efficient maximum likelihood estimation and uncertainty calculations for the parameter set $\theta = (\beta, \Psi)$ and random effects u. The presentation of the parameter estimation conducted in `TMB` is strongly inspired by Chapter 2 in Kristensen et al. (2016) and Section 5.10 in Madsen and Thyregod (2011).

`TMB` maximizes a user-provided objective function in the form of a C++ template, to estimate the maximum likelihood for the parameter set $\theta = (\beta, \Psi)$. The following code chunk illustrates the C++ template for the hierarchical Poisson Normal model specified in 13:

```
#include <TMB.hpp>
template<class Type>
Type objective_function<Type>::operator() ()
{
  // R input data
  DATA_VECTOR(y);                              // Count data
  DATA_VECTOR(x);                              // Population size
  DATA_MATRIX(X);                              // Design matrix
  PARAMETER_VECTOR(u);                         // Random effects
  // Parameters
  PARAMETER_VECTOR(beta);                      // Fixed effects parameters
  PARAMETER(log_sigma_u);                      // Model parameter
  vector<Type> lambda  = exp(X*beta-log(x)+u); // Construct 'lambda'
  Type sigma_u = exp(log_sigma_u);             // And the model parameters
  Type mean_ran = Type(0);
  // Objective function
  Type f = 0;                                  // Declare the objective
  f -= sum(dnorm(u,mean_ran,sigma_u,true));    // Calculate the objective
  f -= sum(dpois(y,lambda,true));              // Calculate the objective
  return f;
}
```

The objective function maximizes the marginal log-likelihood function, which integrates out the random effects u

$$\ell_M(\theta; y) = \int_{\mathbb{R}^q} \ell(\theta; u, y) du \tag{30}$$

where $\ell(\theta, u, y)$ is the joint log-likelihood function of the data given the parameters and random effects. The maximizer $\hat{u}_\theta$ of the joint log-likelihood $\ell(\theta; u, y)$ with respect to the random effects u is defined as:

$$\hat{u}_\theta = \arg\max_u \ell(\theta; u, y) \tag{31}$$

Using $H(\hat{u}_\theta)$ to denote the negative Hessian of the joint log-likelihood evaluated at $\hat{u}_\theta$; i.e,

$$H(\hat{u}_\theta) = -\ell''_{uu}(\theta, u, y)|_{u=\hat{u}_\theta} \tag{32}$$

The Laplace approximation for the marginal log-likelihood $\ell_M(\theta)$ is

$$\ell_{M,LA}(\theta, y) = \ell(\theta, u, y) - \frac{1}{2} \log \left| \frac{H(\hat{u}_\theta)}{2\pi} \right| \tag{33}$$

Our estimate of θ minimizes the negative log of the Laplace approximation, i.e.,

$$\hat{\theta} = \arg\min_{\theta} -\ell_{M,LA}(\theta, y) \tag{34}$$

The minimization of the Laplace approximation for the marginal likelihood is then performed using conventional R optimization routines (e.g., BFGS) to optimize the objective and obtain our estimate $\hat{\theta}$. Uncertainty of the estimate $\hat{\theta}$, or any differentiable function of the estimate $\phi(\hat{\theta})$, is obtained by the δ-method:

$$\mathrm{V}\big(\phi(\hat{\theta})\big) = -\phi'_\theta(\hat{\theta})\Big(\Delta^2 \ell_{M,LA}(\hat{\theta}, y)\Big)^{-1} \phi'_\theta(\hat{\theta})^T \tag{35}$$

Additionally, `TMB` utilizes Automatic Differentiation (AD) techniques (Griewank and Walther, 2008) to evaluate first, second, and potentially third-order derivatives. This approach enhances the computational efficiency and accuracy of the parameter estimation process in the implemented models. Therefore, even though the random effects are analytically integrated out in the hierarchical Poisson Gamma model, and the Laplace approximation is not needed, implementing the joint log-likelihood function in `TMB` can still result in more efficient computations. For a comprehensive introduction to the concept of AD, it is recommended to read Section 2.1 and Section 2.2 of Fournier et al. (2012).

4. Simulation study

This subsection includes a thorough description of the simulation study conducted to evaluate the performance of the novel outbreak detection algorithm compared to state-of-the-art algorithms. The simulations cover various scenarios, adapted from the study by Noufaily et al. (2013).

The subsection begins by describing the method used to simulate the baseline data. These data are generated using a Negative Binomial model with a time-dependent mean $\mu(t)$. Next, the assumptions regarding the simulated outbreaks are outlined, including the outbreak size and distribution in time.

The evaluation measures used to assess the performance of the outbreak detection algorithms are then presented. These measures are designed to capture relevant quantities in the context of outbreak detection.

4.1. The simulated baseline data

The simulated baseline data is generated using a Negative Binomial model with a mean parameter μ and a variance parameter $\phi\mu$. The dispersion parameter ϕ is assumed to be greater than or equal to 1. The mean $\mu(t)$ is defined by a linear predictor that includes a trend component and a seasonality component represented by Fourier terms.

The equation for $\mu(t)$ is given as:

$$\mu(t) = \exp\left(\theta + \beta_t + \sum_{j=1}^{m}\left(\gamma_1 \cos\Big(\frac{2\pi j t}{52}\Big) + \gamma_2 \sin\Big(\frac{2\pi j t}{52}\Big)\right)\right) \tag{36}$$

In this equation, m represents the number of Fourier terms used to model seasonality. When $m = 0$, it indicates the absence of seasonality, while $m = 1$ corresponds to annual seasonality.

To cover a wide range of data sets encountered in practical applications, 28 different parameter combinations are generated. These combinations vary in terms of trends (represented by different values of β), seasonalities (represented by different values of γ_1 and γ_2), baseline frequencies of reports (represented by different values of θ), and dispersion (represented by different values of ϕ). The specific parameter values for the 28 scenarios are provided in Table 1.

Table 1: Parameters and criteria utilized to generate the 28 scenarios.

Scenario	θ	ϕ	β	γ_1	γ_2	m	Trend
1	0.10	1.5	0.0000	0.00	0.00	0	0
2	0.10	1.5	0.0000	0.60	0.60	1	0
3	0.10	1.5	0.0025	0.00	0.00	0	1
4	0.10	1.5	0.0025	0.60	0.60	1	1
5	-2.00	2.0	0.0000	0.00	0.00	0	0
6	-2.00	2.0	0.0000	0.10	0.30	1	0
7	-2.00	2.0	0.0050	0.00	0.00	0	1
8	-2.00	2.0	0.0050	0.10	0.30	1	1
9	1.50	1.0	0.0000	0.00	0.00	0	0
10	1.50	1.0	0.0000	0.20	-0.40	1	0
11	1.50	1.0	0.0030	0.00	0.00	0	1
12	1.50	1.0	0.0030	0.20	-0.40	1	1
13	0.50	5.0	0.0000	0.00	0.00	0	0
14	0.50	5.0	0.0000	0.50	0.50	1	0
15	0.50	5.0	0.0020	0.00	0.00	0	1
16	0.50	5.0	0.0020	0.50	0.50	1	1
17	2.50	3.0	0.0000	0.00	0.00	0	0
18	2.50	3.0	0.0000	1.00	0.10	1	0
19	2.50	3.0	0.0010	0.00	0.00	0	1
20	2.50	3.0	0.0010	1.00	0.10	1	1
21	3.75	1.1	0.0000	0.00	0.00	0	0
22	3.75	1.1	0.0000	0.10	-0.10	1	0
23	3.75	1.1	0.0010	0.00	0.00	0	1
24	3.75	1.1	0.0010	0.10	-0.10	1	1
25	5.00	1.2	0.0000	0.00	0.00	0	0
26	5.00	1.2	0.0000	0.05	0.01	1	0
27	5.00	1.2	0.0001	0.00	0.00	0	1
28	5.00	1.2	0.0001	0.05	0.01	1	1

To simulate the baseline data without outbreaks, 100 replicates are generated for each of the 28 parameter scenarios. Each replicate consist of a time series of size $T = 624$ weeks.

The 624 weeks are divided into three periods: weeks 1-313 are used for training, weeks 313-575 are considered as baseline weeks, and weeks 576-624 are designated as the test weeks for evaluation.

The simulation results are based on the test weeks of all the replicates, totaling $100 \times 49 = 4900$ weeks, for each of the 28 data scenarios and each method investigated.

4.2. The simulated outbreaks

The outbreaks starting in week t_i are simulated using the following procedure. First, a constant value k is chosen at random. The size of the outbreak, denoted as v, is then generated randomly from a Poisson distribution with a mean equal to k times the standard deviation of the baseline count in that scenario.

Next, the outbreak is distributed randomly in time according to a discretized log-normal distribution with a mean of 0 and a standard deviation of 0.5, represented as $Z \sim \lfloor \mathrm{LN}(0, 0.5^2) \rfloor$. This is achieved by drawing v random numbers, which correspond to the outbreak size, from the specified log-normal distribution and then rounding down these numbers to the nearest integer.

The probability mass function for the discretized log-normal distribution is visualized in Figure 1.

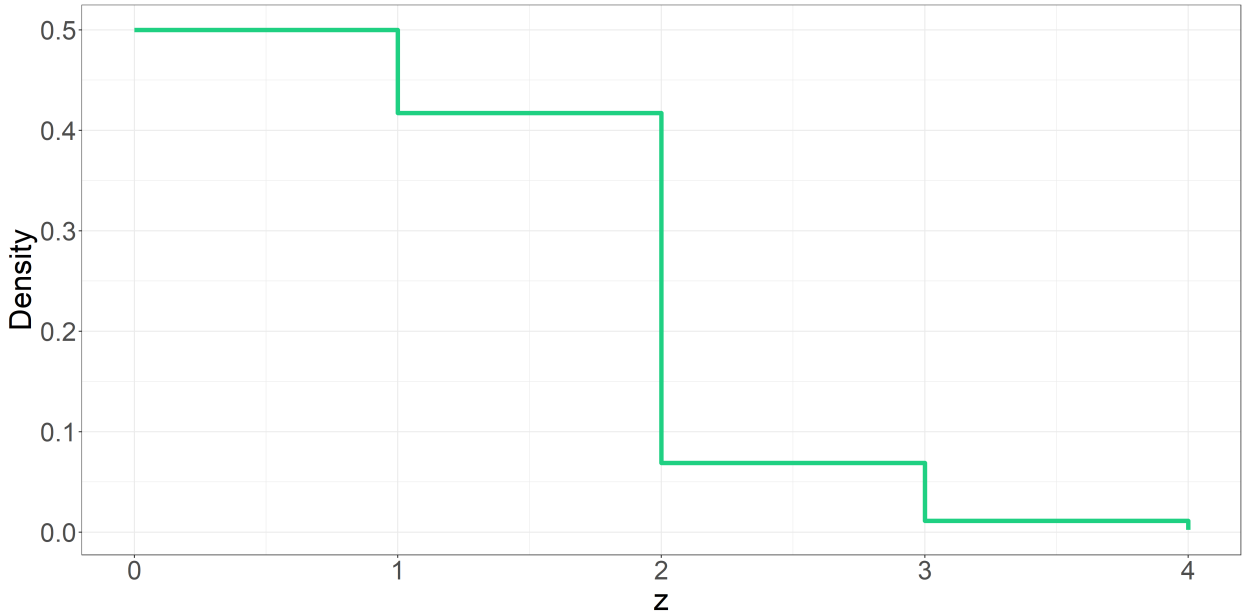


Figure 1: Stairstep plot of the probability mass function for the discretized log-normal distribution with a mean of 0 and a standard deviation of 0.5, i.e. $Z \sim \lfloor \text{LN}(0, 0.5^2) \rfloor$.

Typically, outbreak durations of 2-4 weeks are observed when values of k are in the range of 2-10. To simulate the outbreaks, the outbreak cases are added to the baseline count in week $t_i + z_i$, where t_i represents the start time of the outbreak and z_i represents the number of weeks after the start of the outbreak. The start and end times of the outbreaks are recorded for evaluating the performance of the methods.

To simulate outbreaks, the following procedure is followed:

- **Outbreaks in baseline weeks:** For each data series, four outbreaks are generated. The start time of each outbreak is randomly selected from the baseline weeks (weeks 313-575). The value of k is sampled randomly with replacement from the set $\{2, 3, 5, 10\}$. It should be noted that different outbreaks are generated for each of the 2800 runs.
- **Outbreaks in current weeks:** For each data series, one outbreak is generated. The start time of the outbreak is randomly chosen from the last 49 weeks (weeks 576-624). The value of k is sampled randomly in the range of 1 to 10. Similar to the previous case, different outbreaks are generated for each of the 2800 runs.

One randomly chosen realization for scenario 8, 12, 13, and 20 are visualized in Figure 2.

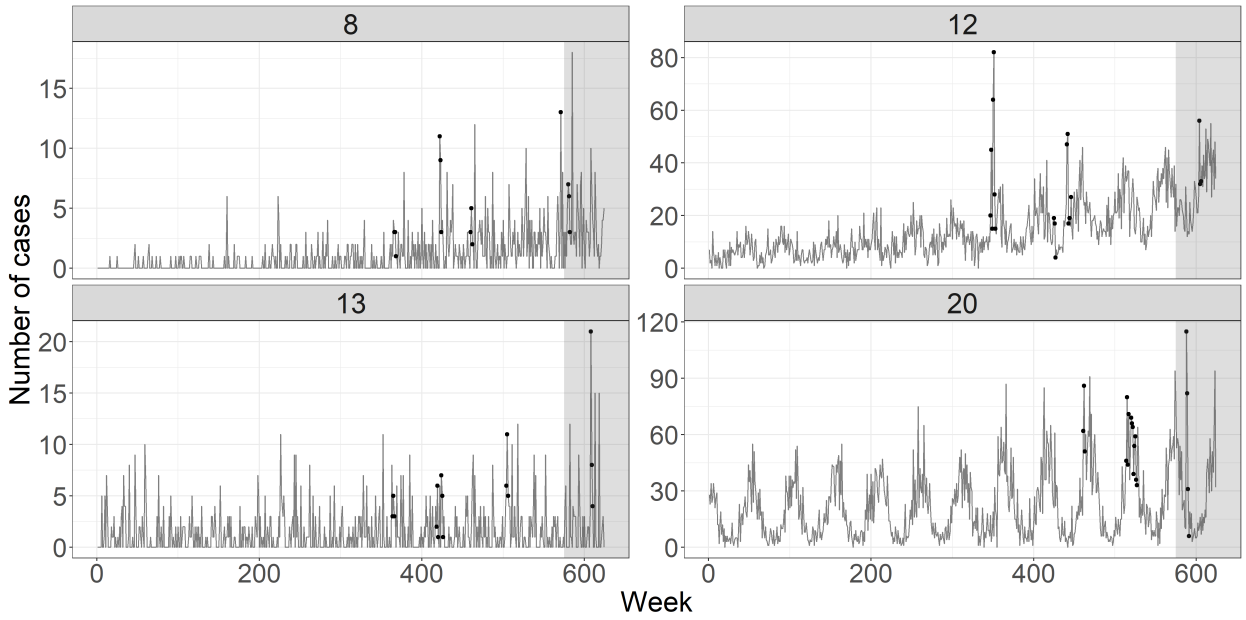


Figure 2: Plots of one randomly chosen realization for scenario 8, 12, 13, and 20 (see Table 1). During outbreaks (circles), outbreak cases are added to the baseline data. Four outbreaks are added during the baseline weeks and 1 outbreak is added during the test weeks. The results are based on the data obtained in the test weeks (grey area).

Evidently, the scenarios vary in their epidemiological characteristics, such as seasonality, trend, and incidence.

4.3. Evaluation measures

To evaluate the performance of the outbreak detection system in the absence and presence of outbreaks, several measures are employed to assess its effectiveness. These measures are specifically designed to capture relevant quantities in the given context.

In the absence of outbreaks in the data, one of the primary measures used is the FPR. This is calculated for each of the 28 scenarios, before the addition of the simulated outbreaks to the baseline data. The FPR is determined by calculating the proportion of the 49 weeks and 100 replicates in which the observed value exceeds the threshold in the absence of any simulated outbreaks.

Another measure is the POD of an outbreak. Likewise, this is calculated for each of the 28 scenarios, but this time it is in the presence of the simulated outbreaks. The algorithm is applied iteratively for the 49 current weeks, and an outbreak is considered detected if the observed value exceeds the threshold at least once within the start and end times of the outbreak. The POD of an outbreak is then determined as the proportion of outbreaks detected out of the 100 runs.

It is important to note that the FPR is a rate per week, while the POD is a rate per realization. These evaluation measures are chosen because they provide insights into the performance of the detection system on individual time series.

5. Results of the simulation study

5.1. False positive rate

In general, the method introduced by Farrington et al. (1996) tends to have relatively higher FPR compared to the other methods. This observation is consistent with the results presented in Table 2. However, this outcome is not surprising since the Farrington method is known to be overly sensitive, making

it more prone to producing false alarms. On the other hand, the improved method by Noufaily et al. (2013) outperforms the other methods by minimizing the FPR.

Additionally, it is noted that the novel algorithms, using the two different modeling frameworks, perform somewhere in between the two state-of-the-art algorithms in terms of minimizing the FPR.

Table 2: Summary statstics of the FPR obtained in the 28 scenarios using the four different methods.

Method	median(FPR)	mean(FPR)	sd(FPR)	min(FPR)	max(FPR)
Farrington	0.022	0.031	0.031	0	0.217
Noufaily	0.000	0.009	0.015	0	0.125
Poisson Normal	0.021	0.017	0.020	0	0.128
Poisson Gamma	0.021	0.017	0.020	0	0.128

Upon examining Figure 3, it becomes even more apparent that the Noufaily method consistently outperforms the other methods. Furthermore, it is interesting to note that scenario 8 and 15 consistently pose challenges for all methods, while scenarios 13-20 prove to be particularly problematic for the novel method.

A closer look reveals that scenarios 13-20 have the highest overdispersion parameters, indicating increased variability in the data. On the other hand, scenario 8 incorporates both a steep trend and a seasonality component, which can complicate the detection process for all methods.

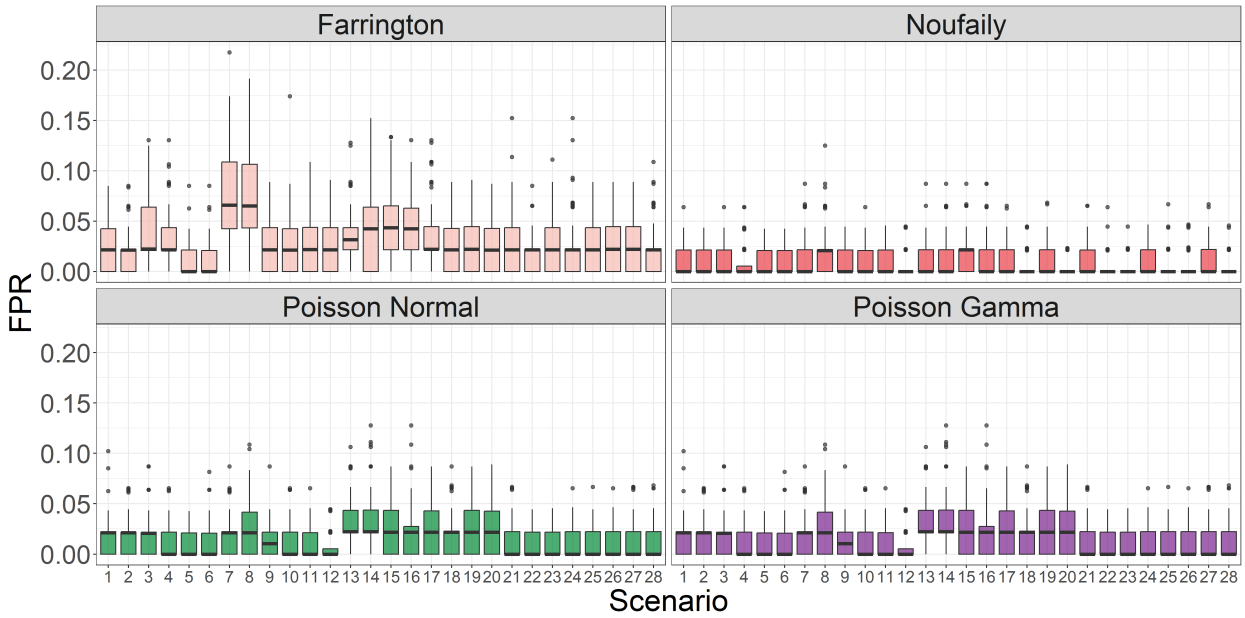


Figure 3: FPR obtained in each of the 28 scenarios for each of the methods applied.

5.2. Probability of detection

As expected, the POD of an outbreak increases with the size of k. Intuitively, when the outbreak size v is larger, it becomes more likely to be detected by the outbreak detection algorithms. In Table 3, it is evident that the Farrington method performs very well in terms of POD, closely followed by the novel method using either modeling framework. Moreover, it can be seen that the Noufaily method is outperformed by the other methods, w.r.t. POD.

The high performance of the Farrington method can be attributed to its sensitivity in detecting outbreaks. Similarly, the novel method utilizing both modeling frameworks demonstrates its effectiveness in detecting outbreaks of varying sizes.

Table 3: Summary statistics of the POD of an outbreak of size k times the standard deviations of the baseline data for each of the methods applied.

Method	k	median(POD)	mean(POD)	sd(POD)	min(POD)	max(POD)
Farrington	2	0.261	0.285	0.179	0.000	0.625
	4	0.462	0.493	0.173	0.091	0.778
	6	0.845	0.800	0.174	0.375	1.000
	8	0.971	0.914	0.147	0.375	1.000
	10	1.000	0.932	0.096	0.714	1.000
Noufaily	2	0.111	0.127	0.108	0.000	0.400
	4	0.300	0.328	0.182	0.091	0.750
	6	0.721	0.682	0.249	0.125	1.000
	8	0.845	0.822	0.193	0.250	1.000
	10	1.000	0.913	0.109	0.700	1.000
Poisson Normal	2	0.244	0.267	0.174	0.000	0.714
	4	0.441	0.460	0.190	0.100	0.818
	6	0.826	0.758	0.222	0.235	1.000
	8	1.000	0.903	0.162	0.375	1.000
	10	1.000	0.945	0.108	0.571	1.000
Poisson Gamma	2	0.244	0.267	0.174	0.000	0.714
	4	0.441	0.460	0.190	0.100	0.818
	6	0.826	0.758	0.222	0.235	1.000
	8	1.000	0.903	0.162	0.375	1.000
	10	1.000	0.945	0.108	0.571	1.000

In Figure 4, the variability in POD of outbreaks can be observed across the 28 scenarios. The level of variability in POD is generally low when the outbreak size factor k is set to 1, indicating that only a few outbreaks are detected in these scenarios. Similarly, when k is set to 10, indicating that almost all outbreaks are detected, the variability in POD is also low.

On the other hand, the variability in POD across the scenarios is highest when k is set to 5, and around 75% of the outbreaks are detected.

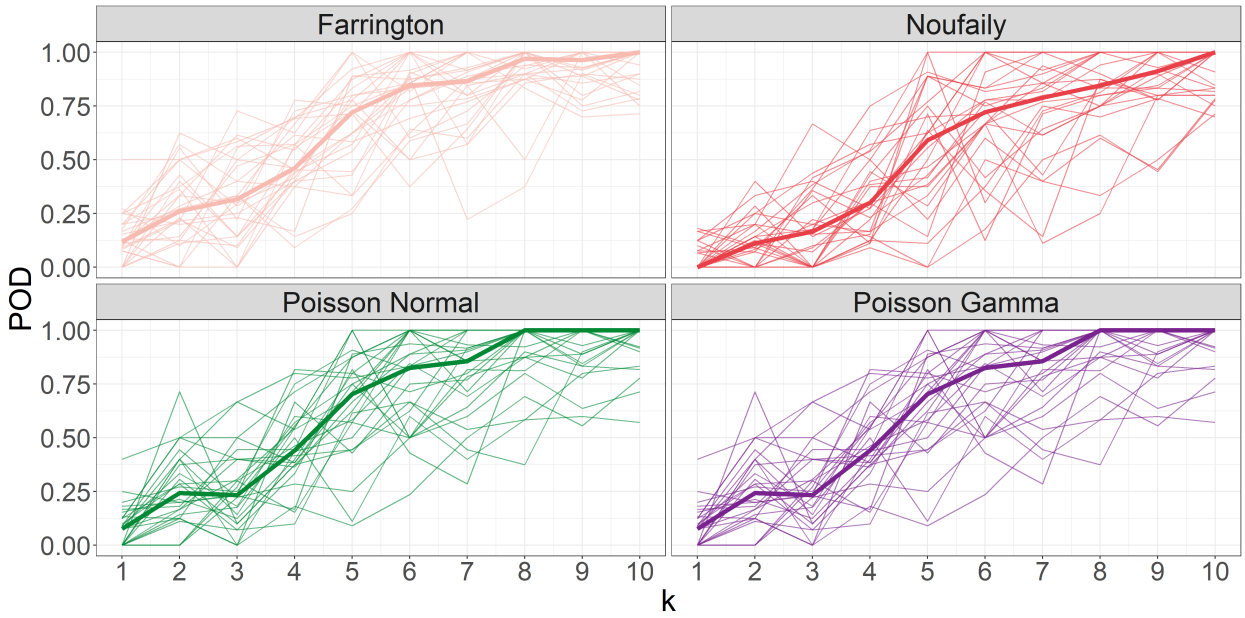
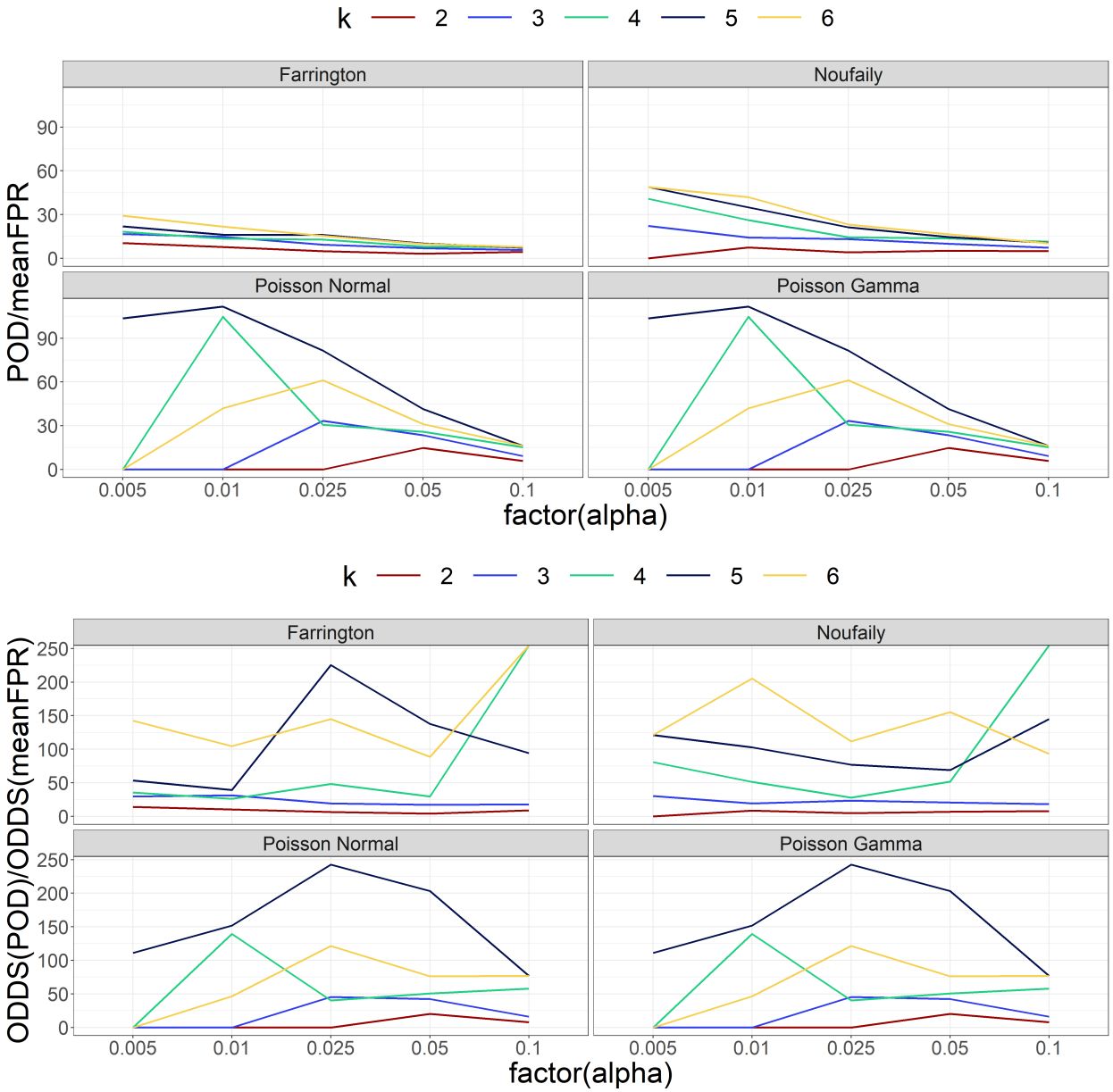


Figure 4: POD of an outbreak of a random size v drawn from a Poisson distribution with mean equal to k times the standard deviations of the baseline data. The x-axis shows increasing values of k. The POD for each scenario is plotted along with the median curves (bold) across all 28 scenarios.

It is important to bear in mind that an outbreak of size v is randomly distributed in time according to a discretized log-normal distribution with a mean of 0 and a standard deviation of 0.5, denoted as $Z \sim \lfloor \mathrm{LN}(0, 0.5^2) \rfloor$. The probability mass function of Z is shown in Figure 1. From the figure, it can be observed that 50% of the outbreak cases are added to the same week as the outbreak starts, 42% are added to the following week, and only 7% are added two weeks after the start. Therefore, the simulated outbreak cases are not observed in a single week only but rather in several concurrent weeks.

Consequently, an outbreak of size v generated from a Poisson distribution with a mean equal to k times the standard deviation of the baseline series is perceived to be relatively smaller than initially perceived in the simulation setup. For example, an outbreak of size $k = 4$ times the standard deviation may only be perceived as an outbreak signal of size 2 times the standard deviation in an individual week.

5.3. Varying significance levels



6. Evaluations on data

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7. Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

8. Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

Susannah Baldry. Attack of the clones. *Nature Reviews Microbiology*, 8(6):390, 2010. doi: 10.1038/nrmicro2369. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-77952857458&doi=10.1038%2fnrmicro2369&partnerID=40&md5=228603d5ee11cdcdf04bea3f5de7c9a8. Cited by: 9; All Open Access, Bronze Open Access.

Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135, 2009. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2008.10.008. URL https://www.sciencedirect.com/science/article/pii/S0169534709000196.

David L. Buckeridge. Outbreak detection through automated surveillance: A review of the determinants of detection. *Journal of Biomedical Informatics*, 40(4):370–379, 2007. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2006.09.003. URL https://www.sciencedirect.com/science/article/pii/S1532046406000980. Public Health Informatics.

Gabriel Bédubourg and Yann Le Strat. Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study. *PLoS ONE*, 12(7), 2017. doi: 10.1371/journal.pone.0181227. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85024904807&doi=10.1371%2fjournal.pone.0181227&partnerID=40&md5=0a18530ff75e9efd8643ca4a743097f7. Cited by: 27; All Open Access, Gold Open Access, Green Open Access.

C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):547–563, 1996. ISSN 09641998, 1467985X. URL http://www.jstor.org/stable/2983331.

R. A. Fisher and Edward John Russell. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922. doi: 10.1098/rsta.1922.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1922.0009.

David A. Fournier, Hans J. Skaug, Johnoel Ancheta, James Ianelli, Arni Magnusson, Mark N. Maunder, Anders Nielsen, and John Sibert. Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249, 2012. doi: 10.1080/10556788.2011.597854. URL https://doi.org/10.1080/10556788.2011.597854.

Andreas Griewank and Andrea Walther. *Evaluating Derivatives*. Society for Industrial and Applied Mathematics, second edition, 2008. doi: 10.1137/1.9780898717761. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898717761.

Simon H. Heisterkamp, Arnold L. M. Dekkers, and Janneke C. M. Heijne. Automated detection of infectious disease outbreaks: hierarchical time series models. *Statistics in Medicine*, 25(24):4179–4196, 2006. doi: https://doi.org/10.1002/sim.2674. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2674.

Maryam Honardoost, Azam Rajabpour, and Ladan Vakil. Molecular epidemiology; new but impressive. *Medical Journal of the Islamic Republic of Iran*, 32(1), 2018. doi: 10.14196/mjiri.32.53. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065203514&doi=10.14196%2fmjiri.32.53&partnerID=40&md5=31ba202c2698dce55f73f5bfb6efb1a7. Cited by: 4; All Open Access, Gold Open Access, Green Open Access.

A. Hulth, N. Andrews, S. Ethelberg, J. Dreesman, D. Faensen, W. van Pelt, and J. Schnitzler. Practical usage of computer-supported outbreak detection in five european countries. *Eurosurveillance*, 15(36):1 – 6, 2010. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-77957607700&partnerID=40&md5=ff916d06bcf38218b2388d0874eed9f8. Cited by: 31.

Claudio U. Koeser, Matthew T. G. Holden, Matthew J. Ellington, Edward J. P. Cartwright, Nicholas M. Brown, Amanda L. Ogilvy-Stuart, Li Yang Hsu, Claire Chewapreecha, Nicholas J. Croucher, Simon R. Harris, Mandy Sanders, Mark C. Enright, Gordon Dougan, Stephen D. Bentley, Julian Parkhill, Louise J. Fraser, Jason R. Betley, Ole B. Schulz-Trieglaff, Geoffrey P. Smith, and Sharon J. Peacock. Rapid whole-genome sequencing for investigation of a neonatal mrsa outbreak. *NEW ENGLAND JOURNAL OF MEDICINE*, 366(24):2267–2275, JUN 14 2012. ISSN 0028-4793.

Kasper Kristensen, Anders Nielsen, Casper W. Berg, Hans Skaug, and Bradley M. Bell. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21, 2016. doi: 10.18637/jss.v070.i05.

Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):619–678, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346105.

Henrik Madsen and Poul Thyregod. *Introduction to general and generalized linear models*. Texts in statistical science. CRC Press, 2011. ISBN 9781420091557.

Angela Noufaily, Doyo G Enki, Paddy Farrington, Paul Garthwaite, Nick Andrews, and André Charlett. An improved algorithm for outbreak detection in multiple surveillance systems. *Online Journal of Public Health Informatics*, 32(7):1206–1222, mar 2013.

Jaume Palmer Real, Jan Kloppenborg Møller, Rongling Li, and Henrik Madsen. A data-driven framework for characterising building archetypes: A mixed effects modelling approach. *Energy*, 254, 2022. ISSN 0360-5442. doi: 10.1016/j.energy.2022.124278.

J. Palmer Real, J. Kloppenborg Møller, C. Rasmussen, K. B. Lindberg, I. Sartori, and H. Madsen. Simulating heat load profiles in buildings using mixed effects models. volume 2069. IOP Publishing, 2021. doi: 10.1088/1742-6596/2069/1/012138.

Maëlle Salmon, Dirk Schumacher, and Michael Höhle. Monitoring count time series in r: Aberration detection in public health surveillance. *Journal of Statistical Software*, 70(10):1—–35, 2016. doi: 10.18637/jss.v070.i10. URL https://www.jstatsoft.org/index.php/jss/article/view/v070i10.

M J Struelens and S Brisse. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. *Eurosurveillance*, 18(4):20386, 2013. doi: https://doi.org/10.2807/ese.18.04.20386-en. URL https://www.eurosurveillance.org/content/10.2807/ese.18.04.20386-en.

Steffen Unkel, C. Paddy Farrington, Paul H. Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 175(1):49–82, 2012. ISSN 09641998, 1467985X. URL http://www.jstor.org/stable/41409708.

Russell D. Wolfinger and Xihong Lin. Two taylor-series approximation methods for nonlinear mixed models. *Computational Statistics and Data Analysis*, 25(4):465—490, 1997. doi: 10.1016/S0167-9473(97)00012-1. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0031222254&doi=10.1016%2fS0167-9473%2897%2900012-1&partnerID=40&md5=dbf7ca1e47d836029af4044ff9880463. Cited by: 87.

Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, and Graham M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health. Springer New York, NY, 2009. ISBN 9780387874586.