Journal de la Société Française de Statistique
Vol. 157 No. 1 (2016)

Introduction to statistical inference for infectious diseases

Titre: Introduction à l'inférence statistique pour les maladies infectieuses

Tom Britton[1] and Federica Giardina[1]

Abstract: In this paper, we first introduce the general stochastic epidemic model for the spread of infectious diseases. Then, we give methods for inferring model parameters such as the basic reproduction number R_0 and vaccination coverage v_c assuming different types of data from an outbreak such as final outbreak details and temporal data or observations from an ongoing outbreak. Both individual heterogeneities and heterogeneous mixing are discussed. We also provide an overview of statistical methods to perform parameter estimation for other stochastic epidemic models. In the last section we describe the problem of early outbreak detection in infectious disease surveillance and statistical models used for this purpose.

Résumé : Dans cet article, nous introduisons le modèle stochastique épidémique général pour la propagation des maladies infectieuses. Nous décrivons ensuite des méthodes pour l'inférence des paramètres du modèle tels que le nombre de reproduction de base R_0 et la couverture vaccinale v_c à partir de différents types de données épidémiques telles que des informations sur l'état final de l'épidémie et des données temporelles ou des observations pour une épidémie en cours. La prise en compte d'hétérogénéités individuelles et des contacts hétérogènes est discutée. Nous fournissons également une vue d'ensemble des méthodes statistiques pour l'estimation des paramètres d'autres modèles épidémiques stochastiques. Dans la dernière section nous décrivons le problème de la détection précoce d'épidémies dans la surveillance des maladies infectieuses et les modèles statistiques utilisés dans ce contexte.

Keywords: stochastic epidemic models, basic reproduction number, vaccination coverage, MCMC, infectious disease surveillance, outbreak detection
Mots-clés : modèles épidémiques stochastiques, nombre de reproduction de base, couverture vaccinale, MCMC, surveillance des maladies infectieuses, détection d'épidémies
AMS 2000 subject classifications: 62M05, 65C40

1. Introduction

Infectious disease models aim at understanding the underlying mechanisms that influence the spread of diseases and predicting disease transmission. Mathematical models have been increasingly used to evaluate the potential impact of different control measures and to guide public health policy decisions.

Deterministic models for infectious diseases in humans and animals have a vast literature, (e.g. Anderson and May, 1991; Keeling and Rohani, 2008). Although these models can sometimes be sufficient to model the mean behaviour of the underlying stochastic system and guide towards parameter estimates, they do not allow the quantification of the uncertainty associated to model parameters estimates (Becker, 1989). Stochastic models (Andersson and Britton, 2000; Britton,

[1] Department of Mathematics, Stockholm University, Stockholm, Sweden
E-mail: tom.britton@math.su.se and E-mail: federica@math.su.se

 ISSN: 2102-6238

2004; Diekmann et al., 2013), can be used to infer relevant epidemic parameters and provide estimates of their variability.

Infectious disease data are commonly collected by surveillance systems at certain space and time resolutions. The main objectives of surveillance systems are early outbreak detection and the study of spatio-temporal patterns. Early outbreak detection commonly relies on statistical algorithms and regression models for (multivariate) time series of counts accounting for both time and space variations.

In this overview paper, we start by analysing the general stochastic epidemic model, which describes the spread of a Susceptible Infected Recovered (SIR) disease assuming a closed population with homogeneous mixing, and show how to make inference on important epidemiological parameters, namely the basic reproduction number R_0 and the critical vaccination coverage v_c. We then describe inference procedures for various extensions increasing model realism. Moreover, we describe statistical models used for the analysis and forecasting of time series of infectious disease data in surveillance settings.

Section 2 defines the general stochastic model, and describes inference procedures for R_0 and v_c depending on the available data (final size or temporal data). Section 3 presents extensions of the general stochastic models treating both individual and mixing heterogeneities and Section 4 discusses the main issues in statistical inference from ongoing outbreaks, relating estimates of the exponential growth rate r to R_0 using e.g. serial intervals and generation time estimation. The main challenge in parameter estimation for epidemic models is that the infection process is not observed. Section 5 presents an overview of statistical methods to estimate transmission model parameters dealing with the missing data and describes recent advances in statistical algorithms to improve computational performance. Section 6 shows how statistical models with space/time structures can be applied to infectious disease surveillance settings for early outbreak detection and forecasting. Section 7 mentions some further extensions and model generalizations as well as new approaches to perform statistical inference for infectious diseases.

2. Inference for a simple stochastic epidemic model

2.1. A simple stochastic epidemic model and its data

We start by defining a simple stochastic model known as the general stochastic epidemic model (e.g. Section 2.3 in Andersson and Britton, 2000). This model considers a so-called SIR-disease where individuals at first are Susceptible. If they get infected they immediately become Infectious (an infectious individual is called an infective) and remain so until they Recover assuming immunity during the rest of the outbreak. Individuals can hence get infected at most once. The general stochastic epidemic assumes a closed population in which individuals mix uniformly in the community, and all individuals are equally susceptible to the disease and equally infectious if they get infected.

Consider a closed population of size n. An individual who gets infected immediately becomes infectious and remains so for an exponentially distributed time with rate parameter γ. During the infectious period an individual has "close contact" with other individuals randomly in time at rate λ, each such contact is with a uniformly selected individual, and a close contact is a contact which results in infection if the contacted person is susceptible; otherwise the contact has no effect.

 ISSN: 2102-6238

Let $(S(t),I(t),R(t))$ denote the numbers of susceptible, infectious and recovered individuals at time t. Because the population is closed and of size n we have $S(t)+I(t)+R(t)=n$ for all t. At the start of the epidemic we assume that $(S(0),I(0),R(0))=(n-1,1,0)$, i.e. that there is one initially infective and no immune individuals. The model is Markovian implying that it may equivalently be defined by its jump rates. An infection occurs at t with rate $\lambda I(t)S(t)/n$ (since each infective has close contacts at rate λ and a close contact results in infection with probability $S(t)/n$). The other event, recovery, occurs at t with rate $\gamma I(t)$, since each infective recovers at rate γ.

The epidemic evolves until the first (random) time T when there are no infectives. Then both rates are 0 and the epidemic hence stops. The final size of the epidemic is denoted $Z=R(T)$, the number of individual that were infected during the outbreak, all others still being susceptible $(S(T)=n-Z)$.

The epidemic model has two parameters, λ and γ, plus the population size n. The perhaps most important quantity for any epidemic model is called the *basic reproduction number* and denoted R_0. The definition of R_0 is that it equals the average number of infections caused by a *typical* individual during the early stage of an outbreak (when nearly all individuals are still susceptible). It is often defined assuming that the population size n tends to infinity. For the general stochastic epidemic, the basic reproduction equals

$$R_0=\lambda/\gamma.$$

This is so because an individual infects others at rate λ (when all individuals are susceptible) while infectious, and the mean duration of the infectious period equals $1/\gamma$. The most important property of R_0 is that it has a threshold value at 1: if $R_0>1$, i.e. if infected individuals infect more than one individual on average, then the epidemic can take off thus producing a "major outbreak", whereas if $R_0\leq 1$ the disease will surely die out without affecting a large fraction of individuals. This has important consequences for vaccination. If, prior to the outbreak, a fraction v are vaccinated (or immunized in some other way), then the number of infections caused by a typical individual is reduced to $R_0(1-v)$ since only the fraction $1-v$ of all contacts result in infection. The new reproduction number is hence $R_v=(1-v)R_0$. For the same reason as above, a positive fraction of the community may get infected if and only if $R_v>1$. Using the expression for R_v this is seen to be equivalent to $v>1-1/R_0$. The value v_c where we have equality is denoted the *critical vaccination coverage* and given by

$$v_c=1-\frac{1}{R_0}.$$

The conclusion is hence that the fraction necessary to vaccinate (or isolate in some other way) to surely avoid a big epidemic outbreak is a simple function of R_0. This explains why R_0 and v_c are considered the perhaps two most important parameters in infectious disease epidemiology (cf. Anderson and May, 1991).

Now we study inference procedures for these parameters (and others) in the general stochastic model. What we can infer, and with what precision, depends on the available data. We mainly focus on the two extreme types of data. The first is where we only observe the final size $Z=R(T)$. The second situation is where we have detailed information about the state of all individuals throughout the outbreak, i.e. where we observe the complete process $\{(S(t),I(t),R(t));0\leq t\leq T\}$,

 ISSN: 2102-6238

called complete observation. In reality, it is often the case that some temporal information is available even if the exact state of all individuals is not known. For example, the onset of symptoms may sometimes be observed for infected individuals. How the onset of symptoms relate to the time of infection and time of recovery depends on the disease in question. Since we do not considering any specific disease, we treat the two extreme situations of final size and complete observation (Sections 2.2 and 2.3) , the precision of any estimator based on partial temporal observations will lie between these two situations.

There are many extensions of the model defined above. For example, it is sometimes assumed that the infectious period is different from the exponential distribution assumed above. The situation where it is assumed non-random is called the continuous time Reed-Frost epidemic model, but also other distributions may be relevant. Another extension is where the disease has a latent period, i.e. where there is a period between when an individual gets infected and until he or she becomes infectious. Such models are often referred to as SEIR epidemics, where the "E" stands for "exposed (but not yet infectious)". Some perhaps even more important extensions are where the community is considered heterogeneous with respect to disease spreading. For example, some individuals (like children and elderly) may be more susceptible to the disease but it is also possible that certain individuals are more infectious be shedding more virus during the infectious period. A different form of heterogeneity of high relevance is where the community has heterogeneous social structures, which all communities do. For example, individuals are more likely to spread the disease to members of the same household than to a random individual in the community (Section 3).

There are two main reasons why making inference in infectious disease outbreaks is harder than in many other situations. The first is that infection events are not independent: whether I get infected is not at all independent of whether my friends get infected. Most standard theory for statistical inference is based on independent events, but such methods are hence not applicable in our situation. The second complicating factor is that we rarely observe the most important events: when and by whom an individual is infected and when they stop being infectious. Instead we observe surrogate observations such as onset of symptoms and stop of symptoms or similar, and to infer the former from the latter is not straightforward. Statistical methodology to analyse such data imputing missing observations is reviewed in Section 5.

2.2. Final size data

Most disease outbreaks of concern, whether in human or animal populations, consist of many individuals getting infected, implying that by necessity the population size n is also large. However, in veterinary science it also happens that controlled experiments are performed, where disease spread is studied in detail in several small isolated units (e.g. Klinkenberg et al., 2002). We start by describing how to make inference in this situation, i.e. when observing disease spread in many small units. We do this for the somewhat simpler discrete time Reed-Frost model in which an infected individual in generation i infects other individuals independently with a probability p in generation $i+1$. If we start with k isolated *pairs* of individuals, one being initially infected and the other initially susceptible, then p is estimated by $\hat{p} = Z/k$, the observed fraction that were infected by the infected "partner" of the same isolated unit. This estimator is based on a binomial experiment and it is well-known that it is unbiased with a standard error of $s.e.(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/k}$.

 ISSN: 2102-6238

A confidence bound on the estimator is constructed using the normal distribution and it is observed that the uncertainty in the estimator decreases with the number of pairs in the experiment as expected. Having estimated the transmission probability p, the natural next step is to estimate R_0. However, this is not straightforward since the transmission probability p (for each specific individual) may vary depending on the animal being in one isolated pair or in some herd (it will be most likely smaller in this case). If the transmission probability is the same when the individual is in a herd, the basic reproduction number will equal $R_0 = mp$ if there are m individuals in the vicinity of any individual. This type of inference, for isolated units, can be extended to situations where there are more than two individuals out of which at least one is initially inoculated. However, the inference gets fairly involved even with very moderate unit sizes (e.g. size 4 units) due to the dependence between individuals getting infected. We refer the reader to e.g. Becker and Britton (1999), who also consider vaccinated and unvaccinated individuals with the aim to estimate vaccine efficacy, for further treatment of these aspects.

We now treat the situation when one large outbreak takes place in a large community (of uniformly mixing homogeneous individuals). As before we let n denote the population size and we assume data consists of the final size Z, i.e. the ultimate number of infected individuals during the course of the outbreak. Using results from probabilistic analyses of a class of epidemic models (containing the general stochastic epidemic model) it is known that in case a major outbreak occurs in a large community, then the outbreak size Z is approximately normally distributed with mean $n\tau$ and variance $n\sigma^2$ where τ and σ^2 are functions of the model parameters. These results, together with delta-method, can be used to obtain an explicit estimate $\hat{R}_0$ and standard error for the estimate (see Section 5.4 in Diekmann et al., 2013):

$$\hat{R}_0 = \frac{-\log(1-Z/n)}{Z/n} \qquad s.e.(\hat{R}_0) = \frac{1}{\sqrt{n}}\sqrt{\frac{1+c_v^2(1-Z/n)\hat{R}_0^2}{(Z/n)(1-Z/n)}}.$$

The point estimate is based on the so-called final size equation for the limiting fraction infected τ: $1-\tau = e^{-R_0\tau}$. The expression for the standard error contains one unknown parameter c_v which is the coefficient of variation of the duration of the infectious period T_I: $c_v^2 = V(T_I)/E(T_I)^2$. For the general stochastic epidemic the infectious period is exponential leading to that $c_v = 1$ whereas $c_v = 0$ for the Reed-Frost epidemic. Most infectious diseases have an infectious period with less variation than the exponential distribution, so replacing c_v by 1 usually gives a conservative (i.e. large) standard error.

In case the outbreak takes place in a large community it may be that the total number of infected Z is not observed, but instead the number of infected Z_m in a sample of size (say) m may be the data at hand. Therefore, there are two sources of error in the estimate: the uncertainty from the final outcome being random, and the uncertainty from observing only a sample of the community. The latter is bigger the smaller sample is taken. In this situation, the estimator of R_0 and its uncertainty can be shown to be

$$\hat{R}_0 = \frac{-\log(1-Z_m/m)}{Z_m/m}$$

$$s.e.(\hat{R}_0) = \sqrt{\frac{1+c_v^2(1-Z_m/m)\hat{R}_0^2}{n(Z_m/m)(1-Z_m/m)} + \frac{(1-m/n)(1-(1-Z_m/m)\hat{R}_0)^2}{m(Z_m/m)(1-Z_m/m)}}.$$

 ISSN: 2102-6238

The above approximation uses the delta-method together with the fact that $V(Z_m) = E(V(Z_m|Z)) + V(E(Z_m|Z))$. We see that the first term in the square root equals the standard error when observing the whole community and the second term vanishes if $m = n$ as expected. If on the other hand $m \ll n$ the second term under the square root dominates; then nearly all uncertainty comes from observing only a small sample.

Another fundamental parameter mentioned above is the critical vaccination coverage v_c: the necessary fraction to immunize in order to surely prevent a major outbreak. For our simple model we know that $v_c = 1 - 1/R_0$. The estimator for this quantity is obtained by plugging in the estimator for R_0 given above, and a standard error is obtained using the delta-method again. The result is (see Section 5.4 in Diekmann et al., 2013)

$$\hat{v}_c = 1 - \frac{1}{\hat{R}_0} = 1 - \frac{Z/n}{-\log(1-Z/n)} \qquad s.e.(\hat{v}_c) = \frac{1}{\sqrt{n}} \sqrt{\frac{1 + c_v^2(1-Z/n)\hat{R}_0^2}{\hat{R}_0^4 (Z/n)(1-Z/n)}}.$$

In case only a sample is observed the following estimator and standard error can be derived:

$$\hat{v}_c = 1 - \frac{Z_m/m}{-\log(1-Z_m/m)}$$

$$s.e.(\hat{v}_c) = \sqrt{\frac{1 + c_v^2(1-Z_m/m)\hat{R}_0^2}{n\hat{R}_0^4 (Z_m/m)(1-Z_m/m)} + \frac{(1-m/n)(1-(1-Z_m/m)\hat{R}_0)^2}{m\hat{R}_0^4 (Z_m/m)(1-Z_m/m)}}.$$

As when estimating R_0 the second term vanishes as $m \to n$ whereas it dominates if we have a small sample, i.e. $m \ll n$.

The above estimates were based on final size data from one outbreak assuming that all n individuals were initially susceptible. In many situations there are also initially immune individuals when an outbreak occurs. Suppose as above that there are n initially susceptible and Z/n denotes the fraction infected among the initially susceptible, but that there were additionally n_I initially immune individuals. Then the estimate $\hat{R}_0$ above is actually an estimate of the effective reproduction number $R_E = sR_0$, where $s = n/(n+n_I)$ denotes the fraction initially susceptible (just as if a fraction $1-s$ were vaccinated). The estimate of R_0 and v_c (the fraction necessary to vaccinate assuming everyone is susceptible) are then given by the expressions above replacing $\hat{R}_0$ by $\hat{R}_0/s$. The corresponding standard errors are as before but dividing by s for $\hat{R}_0$, and multiplying by s for $\hat{v}_c$.

2.3. *Temporal data*

The estimates of the previous section were based on observing the final outcome of an outbreak, denoted Z. Quite often some temporal data, such as weekly reported cases, are also observed. This will improve inference for R_0 and v_c as compared with final size data. However, for the simple scenario of the current section where there are no individual heterogeneities and where individuals mix uniformly, the gain from having temporal data is limited. In Andersson and Britton (2000), Exercise 10.3, the precision based on final size data is compared with the estimation precision from so-called complete data, meaning that the time of infection and time of recovery of all infected individuals are observed. Even with such very detailed data the gain in reduced standard

 ISSN: 2102-6238

error is only of the order 10-15% for some common parameter values. Since most temporal data is less detailed than complete data, but more detailed than final size data, the gain from such temporal data will be even smaller, say 5-10%. A disadvantage with using temporal data in the analysis is that the estimators and their uncertainties are quite involved, using for example martingale methods, as compared to the rather simple estimators for final size data given above. Further, for some partial temporal data types it might even be hard to specify what is observed in terms of model quantities and estimators may therefore be lacking. For this reason we do not present estimators for temporal data and refer the interested reader to e.g. Diekmann et al. (2013), Section 5.4.

Having temporal data is hence not so important for precision in estimation of R_0 and the critical vaccination coverage v_c when having a homogeneous community that mixes (approximately) uniformly. However, temporal data may be useful for many other reasons. Firstly, having temporal data enables estimation of the two model parameters λ and γ separately, and not only the ratio of the two $R_0 = \lambda/\gamma$. Another important reason is that it may be used as model validation. It can for example happen that the close contact parameter (λ) changes over time, for example due to increasing precautions of uninfected individuals. Without temporal data such deviation from the model above cannot be detected. Similarly, if the community actually is heterogeneous in some way this will typically lead to a quicker decrease of incidence as compared to a homogeneous community. Another reason to collect temporal data is of course that it is not necessary to wait until the end of the outbreak before making inference. This is particularly important for new emerging outbreaks (see Section 4 below). Moreover, infectious diseases surveillance systems rely on the availability of temporal data for early outbreak detection and forecasting, as explained in Section 6.

3. Heterogeneities

The model treated in the previous section assumed a community of homogeneous individuals that mix uniformly. Reality is of course not like that and various heterogeneities affect the spreading patterns of an infectious disease. The type of heterogeneities to consider will depend on both the type of community and the type of disease. Think for example of influenza and a sexually transmitted disease; for these two diseases the relevant contact patterns clearly differ. Roughly speaking, heterogeneities can be divided into two different sorts, individual heterogeneities and mixing heterogeneities. These will be discussed below in separate subsections as they quite often require different methods of both modelling and statistical analysis.

3.1. Individual heterogeneities

Individual heterogeneities are factors which affect the risk of getting infected or of spreading the disease onwards. This can for example be age and/or gender, (partial) immunity or vaccination status. Such factors can often be used to categorize individuals into different types, and outbreak data will then be reported as final size (or temporal) data separately for the different cohorts. This type of data is often called a multitype epidemic outbreak. Final size data would then be to observe the number, or fraction, infected in the different cohorts. If there are k groups we let the final fraction of infected in each group be denoted by $\tilde{\tau}_1, \ldots, \tilde{\tau}_k$, and the known community fractions

 ISSN: 2102-6238

of the different groups are given by $\pi_1, \ldots, \pi_k$ (so π_i is the community fraction of individuals being of type i). From this data we would like to estimate the model parameters $\{\lambda_{ij}, \gamma_i\}$; there is now a close contact rate between all pairs of groups (λ_{ij}/n is the rate at which an infectious i-individual infects a given susceptible type-j individual) and a type-specific recovery rate γ_i. In general we hence have $k^2 + k$ model parameters whereas the data vector has dimension k. Clearly it will hence not be possible to estimate all parameters from final size data. In fact, it will not even be possible to estimate the basic reproduction number R_0 consistently, where R_0 is now the largest positive eigenvalue of the so-called next generation matrix M with elements $m_{ij} = \lambda_{ij}\pi_j/\gamma_i$. An intuitive explanation to this result is easy to give for the situation where $\lambda_{ij} = \alpha_i\beta_j$, so the first factor is the infectivity of i-individual and the second factor the susceptibility of j-individuals. By observing the fraction of infected of the different types in a multitype epidemic it is possible to infer which of the types are more susceptible to the disease, but the data contains less information on which types are more infectious in case they get infected. However, and the latter affects the estimation of R_0 as well. The equations which to base parameter estimates on are the following (corresponding to the final size equations for the multitype epidemic model):

$$1 - \tilde{\tau}_j = e^{-\sum_i \lambda_{ij}\pi_i\tilde{\tau}_i/\gamma_i}, \qquad j = 1, \ldots, k.$$

If the number of parameters are reduced down to k, or if some parameters are known, the k equations above may be used to estimate the remaining parameters including R_0. Uncertainty estimates can also be obtained using probabilistic results of Ball and Clancy (1993), but to derive them explicitly remains an open problem.

An important common particular type of multitype setting is where there are asymptomatic cases. For many infectious diseases certain infected individuals have no symptoms but may still spread the disease onwards. This situation is slightly different from the description above in that there are not two distinguishable types of individuals; it is only upon infection that individuals react differently and either become symptomatic or asymptomatic. The most challenging statistical feature is that the asymptomatic cases are rarely observed. In order to make good inference in this situation it is necessary to obtain information about the fraction of asymptomatic cases, for example by testing for antibodies in a random sample in the community.

3.2. Heterogeneous mixing

Individuals are also heterogeneous in the way they mix with each other. In the simple model defined in the previous section it was assumed that individuals mix uniformly with each other, but reality is of course nearly always more complicated, which hence should be taken into account in modelling and statistical analysis. For human diseases there are mainly two types of mixing heterogeneities that has been accounted for: households and networks. The first and most important is the relevance of household structure for many diseases: for diseases like influenza the risk of transmitting to a specific household member is much higher than the risk of transmitting to a (randomly selected) individual in the community. This can be modelled by assuming a transmission rate λ_H to each individual of the same household, and another "global" transmission rate λ_G/n (of different order) to each individual outside the household. The effect of such additional transmission within household is that infected individuals will tend to cluster in certain households leaving other households unaffected (e.g. Ball et al., 1997), and the higher

 ISSN: 2102-6238

λ_H is, the more will infected individuals be clustered. This can be used when inferring model parameters including reproduction numbers as illustrated by Ball et al. (1997), but also more recently in e.g. Fraser (2007).

For temporal data the two different transmission rates may be disentangled more directly by comparing the current fraction of infectives in a household whenever infection occurs (cf. Fraser, 2007). For a model having constant infectious rates throughout the infectious period, the log-likelihood contribution relevant for estimating λ_G and λ_H equals

$$\sum_{i,j} \log[S_i(t_{ij}-)(\lambda_H I_i(t_{ij}-) + \frac{\lambda_G}{n} I(t_{ij}-))] - \int_0^{t_{obs}} \lambda_H \left(\sum_i S_i(u) I_i(u) \right) + \frac{\lambda_G}{n} S(u) I(u) du,$$

where $\{t_{ij}\}, t_{ij} \in (0, t_{obs})$ are the observed infection times in household i and t_{obs} is the end of the observation period, $I_i(t)$ and $I_i(t-)$ denote the number of infectives in household i at t or just before t respectively, and similar for $S_i(t)$ and $S_i(t-)$, and where (as before) $S(t) = \sum_i S_i(t)$ and $I(t) = \sum_i I_i(t)$ are the corresponding totals. This likelihood can be used (assuming the rare situation where infection times are actually observed) to infer the transmission parameters λ_H and λ_G, i.e. it enables distinction between if most transmission is within or between households. If only final size data is available it is still possible to determine if most transmission takes place within or between households by fitting parameters to the final size likelihood using recursive equations (cf. Ball et al., 1997). This method also enables estimation of a reproduction number R_*, which now both is more complicated to interpret and is a more complicated function of model parameters. A similar structure to households, having higher transmission within the groups than between, is that of schools and, for domestic animals, herds. These units are larger thus allowing some large population approximations such that each herd may have its own R_0. A complicated inference problem lies in estimating the contact rates between herds using transportation data (e.g. Lindström et al., 2009).

A different type of mixing heterogeneity which has received a lot of attention in the modelling community over the last 10-15 years is where the community is treated as a social network and where transmission takes place only (or mainly) between neighbours of the network (e.g. Newman, 2003). Both the structure of the network as well as the transmission dynamics taking place "on" the network are important for inferring the potential of an outbreak (R_0) and effects of various preventive measures. A big difference from the household setting just discussed is that usually the underlying network is rarely observed. At best, certain local properties of the network, such as the mean degree, the degree distribution, the clustering coefficient and/or the degree-degree correlation, may be known or estimated. From such local data more global structures determining the potential of disease outbreaks are usually not identifiable (cf. Britton and Trapman, 2013).

3.3. Spatial models

Infectious disease epidemics in populations are inherently spatial because infectious agents are spread by contact from an infectious host to a susceptible host that is "nearby". Heterogeneity in space may play an important role in the persistence and dynamics of epidemics. For example, localised extinctions may be more common in smaller subpopulations whilst coupling between subpopulations may lead to reintroduction of infection into disease-free areas. Understanding the

 ISSN: 2102-6238

spatial heterogeneity has important implications in planning and implementing disease control measures such as vaccination.

One way to account for spatial heterogeneity is to extend the general epidemic model by partitioning the population into spatial subunits of the hosts: nearby hosts are grouped together and interact more strongly than the ones that are further apart. These are the so-called meta-population models (or patch models) and they have been used also to investigate aspects of global disease spread in measles, SARS... A simple two-patch spatial model where hosts move between the two patches at some rate m independent of a disease status would be as follows:

$$\begin{aligned}
\frac{dS_1(t)}{dt} &= -\lambda S_1(t)I_1(t)/n_1 + m(S_2(t) - S_1(t)) \\
\frac{dI_1(t)}{dt} &= \lambda S_1(t)I_1(t)/n_1 - \gamma I_1(t) + m(I_2(t) - I_1(t)) \\
\frac{dS_2(t)}{dt} &= -\lambda S_2(t)I_2(t)/n_2 + m(S_1(t) - S_2(t)) \\
\frac{dI_2(t)}{dt} &= \lambda S_2(t)I_2(t)/n_2 - \gamma I_2(t) + m(I_1(t) - I_2(t))
\end{aligned}$$

where S_i, I_i and $n_i, i = 1,2$ are the number of susceptibles, infected and the community size in patch i respectively. The degree of mixing between groups can be specified, relaxing the assumption of uniform mixing of all individuals.

Time series data sets of infectious disease counts are now increasingly available with spatially explicit information. Some work has been done on time series susceptible-infected-recovered (TSIR) model (Finkenstädt et al., 2002) and its extensions as epidemic metapopulation model assuming gravity transmission between different communities (Xia et al., 2004; Jandarov et al., 2014). According to a generalized gravity model, the amount of movement between the patches (communities) i and j is proportional to $n_i^{\tau_1} n_j^{\tau_2}/d_{ij}^{\rho}$ with $\rho, \tau_1, \tau_2 > 0$ and d_{ij} is the distance between the patches. The transient force of infection by infecteds in location i on susceptibles in location j is $m_{i\to j,t} \propto \frac{n_{j,t}^{\tau_1} I_{i,t}^{\tau_2}}{d_{ij}^{\rho}}$.

4. Statistical analysis of emerging outbreaks

One of the most urgent problems in infectious disease epidemiology over the last decade has been to quickly learn about new diseases (or new outbreaks of old diseases). Examples include SARS (Lipsitch et al., 2003; Riley et al., 2003), foot and mouth disease (Ferguson et al., 2001), H1N1-influenza, (Yang et al., 2009; Fraser et al., 2009) and, most recently, the Ebola outbreak in West Africa (WHO Ebola response team, 2014). A difference from the situation discussed above is that here, in order to identify efficient control measures, estimations are urgent *during* the outbreak. It is not possible to wait until the end of the outbreak and use final size data to infer R_0 and related parameters. Instead inference has to be performed during the early growing stage of the outbreak. Beside having less data this also introduces the risk of producing biased estimates from the fact that individuals that are infected during early stages of an outbreak are usually not representative for the community at large. As an example, the early predictions of the HIV

 ISSN: 2102-6238

outbreak in the 1980's predicted tens of millions of infected within a couple of years, predictions which turned out to be way too high. One partial explanation to this and similar situations is that in a heterogeneous community highly susceptible individuals will get infected early in the epidemic and if predictions are based on the whole community being equally susceptible as the initial group of infected the predictions will overestimate the final size.

As described in ealier sections, the basic reproduction number R_0 carries information about the potential of the epidemic and hence also how much preventive measures are needed to stop an outbreak. During an emerging outbreak, the data (such as weekly reports of new cases) carry information about the exponential growth rate r of the epidemic (also known as the Malthusian parameter), so estimates of r are easily obtained. However, there is no direct relation between r and R_0; for example, a disease with twice as high transmission *and* recovery rate has the same R_0 but larger growth rate r. It is the so-called *generation time* that determines r, the generation time is defined as the time between infection of an individual to the (random) time of infection of one of the individuals he/she infects. The Malthusian parameter r is defined as the solution to the Lotka-Volterra equation

$$\int_0^\infty e^{-rt}\mu(t)dt,$$

where $\mu(t)$ determines the expected generation time and is defined as the average rate at which an infected individual infects new individuals t time units after he/she was infected. The shape of $\mu(t)$ is very influential on the value of r, and the duration and variation of the latent as well as infectious periods have a large impact on r, and thus on what can be inferred also about R_0 in an emerging epidemic outbreak. See Wallinga and Lipsitch (2007) for more about the connection between r, the generation time and R_0.

In most emerging outbreaks the distribution $\mu(t)$ of the generation time is not known and inference methods are needed. However, very rarely infections times, end of latency periods and end of infectious periods are observed. Instead, some related events, such as onset of symptoms and end of symptoms are at best observed. The time between such successive observable events, e.g. the time between onset of symptoms of an infected and the time of onset of symptoms of one of the individuals infected by him/her, is denoted the *serial time*. As has been thoroughly investigated by Svensson (2007), generation time and serial time need not have the same distributions, the latter typically has more variation. As a consequence, even though inference about the serial times is possible from observable data it cannot be used directly to infer the generation time.

A final complicating matter when inferring r and R_0 using data from an emerging outbreak is that the generation time, the "forward" process defined above, is often estimated using data from the corresponding "backward" process. By backward process we mean that infected individuals are contact-traced backwards in time aiming at finding the infection time of their infector (e.g. WHO Ebola response team, 2014). By looking backward in time, short generation time will be over-represented because infections from longer generation times might not have yet occurred (cf. Scalia Tomba et al., 2010). If this bias is not accounted for, predictions based on the backward intervals will be biased in that the predicted number of cases will be over-estimated.

As it has just been explained, there are several potential pitfalls when estimating R_0 and effects of preventive measures from an ongoing emerging outbreak, the reason being that the observed/estimable growth rate r is not directly related to R_0 but only indirectly through the generation time, and the latter is sensitive to usually unknown latent and infectious period

 ISSN: 2102-6238

distributions. But suppose this complicating problem is somehow under control. Is then estimation of R_0 straightforward? The immediate answer is that heterogeneities in the community also play a role when inferring R_0 in an emerging outbreak. However, Trapman et al. (2015) show that for the most commonly studied heterogeneities: multitype epidemics, network epidemics and household epidemics, their effect is very minor. More precisely, estimating R_0 assuming a homogeneous community when in fact it is a multitype epidemic gives exactly the correct estimate of R_0, estimating R_0 assuming a homogeneous community when in fact it comes from a (configuration) network epidemic makes the estimate of R_0 slightly biased from above (the conservative, "better" direction), and finally estimation of R_0 assuming homogeneity when the outbreak agrees with a household epidemic will make the estimate of R_0 close to the correct value and most often conservative. As a consequence, when the relevant heterogeneities make up a combination of the above heterogeneities the simpler estimate assuming homogeneity will slightly overestimate R_0, see Trapman et al. (2015) for more on this topic.

5. Estimation methods (for partially observed epidemics)

As mentioned in Section 2, the main difficulty in parameters estimation for epidemic models is that the infection process is only partially observed and observed quantities may be aggregated (e.g. weekly, monthly etc...). Therefore, the likelihood may become very difficult to evaluate, especially when considering temporal data, since evaluating the likelihood typically involves integration over all unobserved quantities, which is rarely analytically possible. Data imputation methods embedded into statistical inference techniques, such as the expectation-maximisation (EM) algorithm and Markov chain Monte Carlo (MCMC) have been used to estimate the unknown parameters in epidemic models.

The EM algorithm has been considered for epidemic inference problems by e.g. Becker (1997). If we denote with Y the observed data, with Z the augmented data (latent or missing) and with θ the parameter (vector) to estimate, the EM algorithm seeks to find the maximum likelihood estimate of the marginal likelihood by iteratively applying the following two steps: the E-step (expectation step) and the M-step (maximisation step). Once an initial parameter θ_0 is chosen, the E-step and M-step are performed repeatedly until convergence occurs, that is until the difference between successive iterates is negligible. The E-step consists of computing the expected value of the complete data log-likelihood conditional on the observed data and the parameter estimate $\theta^{(t)}$ at iteration t, i.e. $Q(\theta|\theta^{(t)}) = \mathrm{E}_{Z|Y,\theta^{(t)}}\left[\log L(\theta;Y,Z)\right]$ and the M-step requires maximising the expectation calculated in the E-step with respect to θ to obtain the next iterate. The latent data should be chosen such that the log-likelihood of the complete data is relatively straightforward. However, the evaluation of the expectation step can be rather complicated.

Data-augmented MCMC can be used to explore the joint distribution of parameters and latent variables in a similar fashion. Especially in the Bayesian context, the approach is straightforward and it consists in specifying an "observation level" model $P(Y|Z,\theta)$, a "transmission level" model $P(Z|\theta)$ and a prior $p(\theta)$, as outlined by Auranen et al. (2000), resulting in $P(Y,Z,\theta) = P(Y|Z,\theta)P(Z|\theta)p(\theta)$. One drawback with this approach is that it requires high memory for large-scale systems and in addition, designing efficient proposal distributions for the missing data may be challenging. Therefore, applications of data augmentation in MCMC have been mainly concerned with the situation in which data arise from a single large outbreak of a disease (Gibson

 ISSN: 2102-6238

and Renshaw, 1998; O'Neill and Roberts, 1999) or data on small outbreaks across a large number of households (O'Neill et al., 2000).

For large epidemics in large populations, another option is to find analytically tractable approximations of the epidemic model. In epidemic time series data a natural choice is to approximate continuous-time models by discrete-time models (Lekone and Finkenstädt, 2006). An important constraint in those models is that one observation period must effectively capture one generation of cases. This may be achieved only if the generation time of the disease is equal to the length of observation periods, or is a multiple of it. In the latter case, the data must be further aggregated, which may lead to an additional loss of information.

Cauchemez and Ferguson (2008) propose a statistical framework to estimate epidemic time-series data tackling the problem of temporal aggregation (and missing data), by augmenting the data with the latent state at the beginning of each observation period and introducing a diffusion process that approximates the SIR dynamic and has an exact solution. See also Guy et al. (2015) in this journal issue.

Ionides et al. (2006) formulates the inference problem for epidemic models in terms of nonlinear dynamical systems (or state-space models) which consist of an unobserved Markov process Z_t i.e. the state process and the observation process Y_t. The model is completely specified by the conditional transition density $f(Z_t|Z_{t-1},\theta)$, the conditional distribution of the observation process $f(Y_t|Y_{t-1},Z_t,\theta) = f(Y_t|Z_t,\theta)$ and the initial density $f(Z_0|\theta)$. The basic idea is to consider the parameter θ as a time varying process θ_t, i.e. a random walk in $R^{dim(\theta)}$ so that $E(\theta_t|\theta_{t-1}) = \theta_{t-1}$, $Var(\theta_t|\theta_{t-1}) = \sigma^2\Sigma$, $E(\theta_0) = \theta$ and $Var(\theta_0) = \sigma^2 c^2\Sigma$ with σ and c scalar quantities. Then, the objective is to obtain estimate of θ by taking the limit as $\sigma \to 0$. The authors use iterated filtering to produce maximum likelihood estimates with a Sequential Monte Carlo (SMC) method.

A general technique that alleviates the problems generated by likelihood evaluation and that is growing in popularity in epidemiology is the so-called Approximate Bayesian Computation (ABC). ABC utilizes the Bayesian paradigm in the following manner: if M represents the model of interest, then the observed data Y are simply one realization from M, conditional on θ. For a given set of candidate parameters θ', drawn from the prior distribution, we can simulate a data set Y' from M. If $\rho(s(Y'),s(Y)) \leq \varepsilon$, where ρ is a similarity metric, $s(\cdot)$ is a set of lower dimensional (approximately) sufficient summary statistics and ε is chosen small, then θ' is a draw from the posterior. ABC (or likelihood-free computation) can be used with rejection sampling (McKinley et al., 2009), MCMC (Marjoram et al., 2003) or SMC routines (Toni et al., 2009). A general criticism of this method concerns the level of approximation generated by: the choice of metric ρ, the tolerance ε and the number of simulations to obtain estimates.

For stochastic models where simulation is time consuming, it may not be possible to use likelihood-free inference. Learning about parameters in a complex deterministic or stochastic epidemic model using real data can be thought of as a "computer model emulation/calibration" problem (Farah et al., 2014). Emulators are statistical approximations of a complex computer model, which allow for simpler and faster computations. An emulator may consist of a regression model allowing for model discrepancy and measurement error and can be easily fitted to the reported epidemic data. Recent work in emulation and calibration for complex computer models for fitting epidemic models include Jandarov et al. (2014) where a Gaussian process approximation is chosen to mimic the disease dynamics model using key biologically relevant summary statistics obtained from simulations of the model at different parameter values.

 ISSN: 2102-6238

6. Statistical models for infectious diseases surveillance

Infectious disease data are often collected for disease surveillance purposes and information is typically available as incidence counts aggregated over regular intervals (e.g. weekly). As a consequence, individual information is often lost. Also, the number of susceptibles in a population is rarely available. The typical goal in a surveillance setting is to monitor disease incidence, detecting outbreaks prospectively. Due to the lack of detailed information mentioned above, this is rarely achieved by fitting epidemic stochastic models to data, i.e. by explicitly modelling the transmission process.

Commonly the problem is formulated as statistical analysis for detecting an anomaly (step increase) in univariate count data time series $\{y_t, t = 1,2,\ldots\}$. The first approach dates back to Farrington et al. (1996) who compared the observed count in the current week with an expected number, which is calculated based on observations from the past, i.e. similar weeks from the previous years from a set of so-called reference values. An upper threshold is then derived so that an outbreak alarm is triggered once the current observation exceeds this threshold. At time s, $\mathbf{y}_s = \{y_t; t \leq s\}$ the statistic $r(\cdot)$ is calculated on the basis of $\mathbf{y}_s$ compared to a threshold value g. This results in the alarm time $T_a = min\{s \geq 1 : r(\mathbf{y}_s) > g\}$. Several variations/extensions of the Farrington's method exist, (Salmon et al., 2014), based on a two-step procedure: first, a Generalized Linear (Additive) Model (Poisson or Negative Binomial) is fitted to the reference values, and then the expected number of counts μ_s is predicted and used (with its variance) to obtain an upper bound g_s: the alarm is raised if $y_s > g_s$. Other model generalizations allow the detection of sustained shifts through cumulative sum methods (Höhle and Paul, 2008). Applications are in both human and veterinary epidemiology, see e.g. Kosmider et al. (2006).

Sometimes infectious disease data are available at a finer geographical scale (cases are geo-referenced). In these situations the problem of spatio-temporal disease surveillance can be formulated in terms of point-process models (Diggle et al., 2005). The focus is predicting spatially and temporally localised excursions over a pre-specified threshold value for the spatially and temporally varying intensity of a point process $\lambda^*(x,t)$ in which each point represents an individual case. In Diggle et al. (2005), the point process model is a non-stationary log-Gaussian Cox process in which the spatio-temporal intensity has a multiplicative decomposition into two components: one describing purely spatial $\lambda_0^*(x)$ and the other purely temporal variation $\mu_0(t)$ in the normal disease incidence pattern, and an unobserved stochastic component representing spatially and temporally localised departures from the normal pattern $\Phi(x,t)$. Hence, the spatio-temporal incidence is $\lambda^*(x,t) = \lambda_0^*(x)\mu_0(t)\Phi(x,t)$ for t in the prespecified observation period $[0,T], T > 0$, and observation region $S \in R$. Within this modelling framework, anomaly is defined as a spatially and temporally localised neighbourhood within which $\Phi(x,t)$ exceeds an agreed threshold, g, via the predictive probabilities $p(x,s;g) = P(\Phi(x,s) > g|\text{data until time } s)$.

Statistical models as the above mentioned, can also be used for the study of spatio-temporal correlations and patterns explaining the statistical variability in incidence counts. In fact, as a consequence of the disease transmission mechanism, the observations are inherently time and space dependent and appropriate statistical models have to account for such feature in the data. Geographic information can be available at different scales. For example, as in Diggle et al. (2005), an entire region is continuously monitored. A (marked) point pattern model representation has a branching process interpretation and therefore allows the calculation of the expected number

 ISSN: 2102-6238

of secondary infections generated by an infective within its range of interaction (proxy for R_0), see Meyer et al. (2014). A second possibility is that infections are obtained at a discrete set of units at fixed locations followed over time, as farms during livestock epidemics (Keeling and Rohani, 2008). In this case, an SIR modelling approach can be pursued. A third case, probably the most common one, is to have individual data aggregated over some administrative regions and convenient period of time (e.g. week, month etc...).

A general statistical framework for modelling such data can be found in Paul et al. (2008) that extends the model previously proposed by Held et al. (2005). The model is based on a Poisson branching process with immigration and can be seen as an approximation to a chain-binomial model without information on the number of susceptibles to the disease. Previous counts enter additively on the conditional mean counts that is decomposed in two parts: the *endemic* part and the *epidemic* part. The former explains a baseline rate of cases that is persistent with a stable temporal pattern, while the latter accounts for occasional outbreaks. In particular, the number of cases observed at unit i at time t, $i = 1,\dots,m$, $t = 1,\dots,T$ is denoted by y_{it}. The counts can be assumed to follow a Negative Binomial distribution $y_{it}|y_{it-1} \sim NegBin(\mu_{it},\phi)$ with conditional mean $\mu_{it} = \lambda' y_{it-1} + \exp(\eta_{it})$ and conditional variance $\mu_{it}(1+\phi\mu_{it})$ where $\phi > 0$ is an overdispersion parameter and λ' is an unknown autoregressive parameter. The *epidemic* component is represented by $\lambda' y_{it-1}$ and the *endemic* part is $\exp(\eta_{it})$. The inclusion of previous cases allows for temporal dependence beyond seasonal patterns within a unit. To explain the spread of a disease across units, the *epidemic* component can be formulated as $\lambda' y_{it-1} + \gamma_i \sum_{j\neq i} w_{ji} y_{j,t-l}$ where $y_{j,t-l}$ denotes the number of cases observed in unit j at time $t-l$ with lag $l \in 1,2,\dots$ and w_{ji} are suitably chosen weights. To model seasonality, the *endemic* component can be specified as $\nu_{it} = \alpha_i + \sum_{s=1}^{S} \beta_s sin(\omega_s t) + \delta_s cos(\omega_s t)$ where ω_s are Fourier frequencies. The parameter α_i allows for different incidence levels in each of the m units.

Statistical models for surveillance are evaluated and selected in terms of predictive performance in one step ahead-prediction. Strictly proper scoring rules are generally used for this purpose (Gneiting and Raftery, 2007; Czado et al., 2009), the most popular for count data being the logarithmic score.

Most of the statistical models mentioned above are implemented in the R package `surveillance` (Höhle, 2007). Bayesian extensions are fitted via Integrated Nested Laplace approximation (Rue et al., 2009).

7. Concluding remarks

In this paper we have presented results for the general stochastic epidemic model and shown how to infer the most important epidemiological parameters, R_0 and v_c under different data scenarios (final size data or temporal data). The general stochastic epidemic model assumes a finite population that mixes homogeneously and a constant infection rate λ during the infectious period. In Sections 3 and 4 we elaborate some model extensions, e.g. individual heterogeneity, heterogeneous mixing and spatial models discussing how estimation changes.

However, there are other features that affect the disease spread (and therefore other model extensions to account for them) that have not been treated in this work. For example, the probability of getting infected with a disease is usually not constant in time: some diseases are seasonal e.g. common cold viruses. Also an "external" change e.g. the implementation of a control measure,

 ISSN: 2102-6238

may affect either contact rates or infectiousness (or both). One way to account for that is to let the infection rate λ change in time, e.g. as a periodic function (Cauchemez and Ferguson, 2008).

Epidemic models can also be used to derive estimators for the efficacy of control measures such as vaccine, using data generated by field trials and observational studies. Understanding the relation between disease dynamics and interventions is essential particularly for vaccination programs. In fact, vaccines can have protective effects in reducing susceptibility, infectiousness or both and efficacy estimation has to be performed accordingly (Halloran et al., 2010).

Over the last few years, an alternative approach for modelling infectious disease outbreaks has focused on phylodynamics, the integration of phylogenetic methods to analyze the genetic variation of the pathogen and epidemic models (Grenfell et al., 2004). This approach offers new insights into the dynamics of disease outbreak with the aim of inferring transmission routes and times of infection, see e.g. Volz et al. (2009) or Soubeyrand (2015) in this special issue.

Acknowledgments

Both authors are grateful to the Swedish Research Council (grant 340-2013-5003) for financial support.

References

Anderson, R. M. and May, R. M. (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 New York.

Auranen, K., Arjas, E., Leino, T., and Takala, A. K. (2000). Transmission of pneumococcal carriage in families: a latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association*, 95(452):1044–1053.

Ball, F. and Clancy, D. (1993). The final size and severity of a generalised stochastic multitype epidemic model. *Advances in Applied Probability*, 25(4):721–736.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability*, 7(1):46–89.

Becker, N. G. (1989). *Analysis of infectious disease data*. CRC Press.

Becker, N. G. (1997). Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. *Statistical Methods in Medical Research*, 6(1):24–37.

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):287–307.

Britton, T. (2004). Epidemic models, inference. *In Encyclopedia of Biostatistics*, pages 1667–1671.

Britton, T. and Trapman, P. (2013). Inferring global network properties from egocentric data with applications to epidemics. *Mathematical Medicine and Biology*, 10.1093/imammb/dqt022.

Cauchemez, S. and Ferguson, N. M. (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(25):885–897.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

Diekmann, O., Heesterbeek, H., and Britton, T. (2013). *Mathematical tools for understanding infectious disease dynamics*. Princeton University Press.

Diggle, P., Rowlingson, B., and Su, T.-l. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16(5):423–434.

Farah, M., Birrell, P., Conti, S., and De Angelis, D. (2014). Bayesian emulation and calibration of a dynamic epidemic model for H1N1 influenza. *Journal of the American Statistical Association*, To appear.

Farrington, C., Andrews, N., Beale, A., and Catchpole, M. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159:547–563.

 ISSN: 2102-6238

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001). The foot-and-mouth epidemic in great britain: pattern of spread and impact of interventions. *Science*, 292(5519):1155–1160.

Finkenstädt, B. F., Bjørnstad, O. N., and Grenfell, B. T. (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics*, 3(4):493–510.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS One*, 2(8):e758.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., et al. (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*, 324(5934):1557–1561.

Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15(1):19–40.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332.

Guy, R., Larédo, C., and Vergu, E. (2015). Approximation and inference of epidemic dynamics by diffusion processes. *Journal de la SFDS*.

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

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199.

Höhle, M. (2007). `Surveillance`: An R package for the monitoring of infectious diseases. *Computational Statistics*, 22(4):571–582.

Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics & Data Analysis*, 52(9):4357–4368.

Ionides, E., Bretó, C., and King, A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443.

Jandarov, R., Haran, M., Bjørnstad, 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 (Applied Statistics)*, 63(3):423–444.

Keeling, M. J. and Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Klinkenberg, D., De Bree, J., Laevens, H., and De Jong, M. (2002). Within-and between-pen transmission of classical swine fever virus: a new method to estimate the basic reproduction ratio from transmission experiments. *Epidemiology and infection*, 128(02):293–299.

Kosmider, R., Kelly, L., Evans, S., and Gettinby, G. (2006). A stastistical system for detecting salmonella outbreaks in british livestock. *Epidemiology and infection*, 134(05):952–960.

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177.

Lindström, T., Sisson, S. A., Nöremark, M., Jonsson, A., and Wennergren, U. (2009). Estimation of distance related probability of animal movements between holdings and implications for disease spread modeling. *Preventive Veterinary Medicine*, 91(2):85–94.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H., et al. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300(5627):1966–1970.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

McKinley, T., Cook Alex, R., Robert, D., et al. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1):1–40.

Meyer, S., Held, L., and Höhle, M. (2014). Spatio-temporal analysis of epidemic phenomena using the R package `surveillance`. *ArXiv preprint*, `arXiv:1411.0416v1`.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M., and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(4):517–542.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of*

 ISSN: 2102-6238

the Royal Statistical Society: Series A (Statistics in Society), 162(1):121–129.

Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27(29):6250–6267.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L.-M., Lam, T.-H., Thach, T. Q., et al. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627):1961–1966.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Salmon, M., Schümacher, D., and Höhle, M. (2014). Monitoring count time series in R: Aberration detection in public health surveillance. *ArXiv preprint*, `arXiv:1411.1292v1`.

Scalia Tomba, G., Svensson, Å., Asikainen, T., and Giesecke, J. (2010). Some model based considerations on observing generation times for communicable diseases. *Mathematical Biosciences*, 223(1):24–31.

Soubeyrand, S. (2015). Construction of semi-markov genetic-space-time SEIR models and inference. *Journal de la SFDS*.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Trapman, P., Ball, F., Dhersin, J. S., Tran, V. C., Wallinga, J., and Britton, T. (2015). Robust estimation of control effort in emerging infections. *Submitted*.

Volz, E. M., Pond, S. L. K., Ward, M. J., Brown, A. J. L., and Frost, S. D. (2009). Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430.

Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604.

WHO Ebola response team (2014). Ebola virus disease in West Africa - the first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, 371:1481–1495.

Xia, Y., Bjørnstad, O. N., and Grenfell, B. T. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, 164(2):267–281.

Yang, Y., Sugimoto, J. D., Halloran, M. E., Basta, N. E., Chao, D. L., Matrajt, L., Potter, G., Kenah, E., and Longini, I. M. (2009). The transmissibility and control of pandemic influenza A (H1N1) virus. *Science*, 326(5953):729–733.

 ISSN: 2102-6238