Tracking Measles Infection through Non-linear State Space Models

Shi Chen

Department of Entomology, The Pennsylvania State University, University Park, PA 16802, USA.

John Fricks

Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA.

Matthew J. Ferrari

The Center for Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA.

Summary. Estimating the burden of infectious disease is complicated by the general tendency for underreporting of cases. When the reporting rate is unknown, conventional methods have relied on accounting methods that do not make explicit use of surveillance data or the temporal dynamics of transmission and infection. State space models are a framework for various methods that allow dynamic models to be fit with partially or imperfectly observed surveillance data. State space models are an appealing approach to burden estimation as they combine expert knowledge in the form of an underlying dynamic model, but make explicit use of surveillance data to estimate parameter values, to predict unobserved elements of the model, and to provide standard errors for estimates.

1. Introduction

The estimation of the national and global burden of infectious diseases is essential for setting future targets and funding priorities, as well as evaluating the impact of public health measures (Stein et al. (2003),Wolfson et al. (2007)). The problem of burden estimation, however, is complicated by a paucity of detailed surveillance data and a general tendency for under-reporting of cases (Dabbagh et al. (2007)). Many of the methods that have been developed for estimating disease burden have been based on an informal combination of expert opinion, in the form of either a static or dynamic epidemic model and surveillance data, when available (Stein et al. (2003),Wolfson et al. (2007), Crowcroft et al. (2003)). Here we discuss a formal combination of epidemic dynamics with surveillance data using a state space model to provide prediction and bounds for the annual burden of measles based on annual reporting of disease at the national scale.

The global reduction of the burden of morbidity and mortality due to measles is a triumph of modern public health (Wolfson et al. (2007)). Immunization programs focused on routine delivery of measles vaccine and supplemental pulsed campaigns have led to a greater than 85% reduction in global measles mortality since 1980. However, measles remains a leading cause of vaccine-preventable death in children under 5 years in much of the world. Each year, significant resources are allocated to measles vaccination through the World Health Organization (WHO) Expanded Programme on Immunization and the supplemental vaccination campaigns at the country level. Setting goals for vaccination

programs, allocating resources, and evaluation of program success are all complicated by the inherent challenge of assessing the burden of measles disease and mortality at both the national and global level.

Measles cases are routinely reported at the national level for the 193 WHO member states each year. It is generally assumed, however, that these reports represent a severe under-reporting of true measles incidence. As such, programmatic decisions tend to be made based on estimated or corrected estimates of measles incidence (Stein et al. (2003),Wolfson et al. (2007)). In countries where the available measles surveillance is unreliable, estimates of measles burden have been calculated using so-called natural history methods, which combine the available demographic (population size and birth rate) and vaccination data (timing and coverage) to arrive at estimates of the unobserved true incidence (Stein et al. (2003),Wolfson et al. (2007), Crowcroft et al. (2003)). Natural history models for disease burden suffer from two major shortcomings. First, they often require *ad hoc* methods to arrive at confidence intervals for burden. Second, there is no clear connection between the estimated value and the numbers reported by national surveillance programs. Here we present a method to combine a standard natural history type model for estimating the incidence of measles with the annual reported incidence data at the national level in the context of a state space model.

State space, or hidden Markov, models are a framework for methods that fit dynamic models with partially or imperfectly observed data (such as under-reported surveillance data). State space models are characterized by two inter-related parts: a state equation and an observation equation. In the terminology of these methods, the state of the system, X, varies in time according to a mathematical expression describing the evolution from the previous time step, $f(X)$, that is governed by a set of unknown parameters. The observed data, Y, are (possibly transformed) measures of the states through time, which might contain some observation error. This transformed measure of the state, $g(X)$ describes the expected relationship between the unobserved states, X, and the observed data Y. If the state model is a Markov process, such that the value of the states at time t depend only on the value of the states at time $t-1$, then we can more easily carry out statistical inference on the parameters and predict the unknown states based on the conditional probability densities $h_S(X_t|X_{t-1})$ and $h_O(Y_t|X_t)$ given the system parameters.

The primary example of a state space model is the Kalman filter or linear state space model. The original application of such models was to identify the system states given the observations in the case of a known system. However, in a variety of contexts the goal of using such a model may be to infer the parameters of the full system such as in the estimation of ARMA process (Brockwell and Davis (1991)). For the goal of estimating disease burden, we are interested in first estimating parameters of the underlying state and observation models to enable us to predict the states themselves (i.e. the unobserved incidence of disease). Other authors have such as Breto et al. (2009) and Ionides et al. (2006) have given a more modern approach to fitting these models by using methods such as particle filters and computational Bayesian techniques; however, a large gap remains between these methods and what is done in practice. One goal of the present work is to take a simple approach that continues to be in the more general framework of these more modern approaches, i.e. filtering, yet may be more accessible to many practitioners. In addition, we will demonstrate that our simpler approach is comparable to these more computationally intensive methods.

Here we present a specific application of state space modeling to predict unobserved measles burden from under-reported national case reporting. In the sections below, we first

describe the basic state model for the progression of measles cases through time and then discuss an algorithm for predicting the unobserved incidence using the extended Kalman filter (EKF) and potential extensions to these methods. We then explore the performance of this algorithm on simulated data and provide examples of the application to real surveillance data.

2. A Model for Measles Burden

The relatively simple natural history of measles is well described by a family of non-linear epidemic models known as susceptible-infected-recovered (SIR) models (Anderson and May (1991), Bjornstad et al. (2002)). The population is divided into 3 compartments comprising individuals who are susceptible (S) after birth and eventually become infected (I), and then following a period of approximately 2 weeks, recover (R) and are immune to subsequent infection. The transmission of infection between infectious and susceptible individuals is an often complex, non-linear function of the contact process between individuals (McCallum et al. (2001). However, the epidemic dynamics of measles for a variety of settings (Ferrari et al. (2008), Bjornstad et al. (2002), Anderson and May (1991), Metcalf et al. (2009)) are well represented by a simple set of coupled ordinary differential equations:

$$\begin{aligned} \frac{dS}{dt} &= \alpha N - \beta_t S\frac{I}{N} - \mu S \\ \frac{dI}{dt} &= \beta_t S\frac{I}{N} - \gamma I - \mu I \\ \frac{dR}{dt} &= \gamma I - \mu R \end{aligned} \tag{1}$$

where β is the transmission rate (which may vary seasonally),γ is the recovery rate (here 1/14), α is the birth rate, μ is the mortality rate, and N is the total population size. Vaccination is targeted at young children and is dynamically equivalent to a reduction in the effective birth rate.

A key feature of SIR models is the phenomenon of herd immunity (Anderson and May (1991)). That is, the per capita transmission rate is positively related to the proportion of the population that is susceptible. Thus, when the proportion of the population that is susceptible falls low enough, the per capita transmission rate falls below 1 and the population is said to be at herd immunity: i.e. infection cannot spread in the population. As a consequence, infection tends to occur in epidemics, which extinguish themselves as the population becomes increasingly immune. Subsequent outbreaks then occur when the susceptible population is sufficiently replenished by births.

At the annual time scale, we can write the number of susceptibles at year t as the susceptibles in the previous year, minus those that became infected, plus the birth cohort.

$$S_t = S_{t-1} - I_{t-1} + X_t \tag{2}$$

Here the birth cohort X_t is the total number of births, B_t, minus those that were vaccinated. The efficacy of the measles vaccine is approximately 0.85 with a single dose and 0.99 with two doses (Uzicanin and Zimmerman (2011)). Thus, if $V1_t$ and $V2_t$ are the vaccine coverage with one and two doses respectively at time t, then the non-immune birth cohort at time t

is

$$X_t = B_t(1 - 0.85V1_t(1 - V2_t) - 0.99V1_tV2_t) \tag{3}$$

The number of individuals that become infected in year t is the number that were susceptible, S_t, times an annualized infection rate β. Thus we can rewrite 2 as

$$S_t = S_{t-1} - \beta_{t-1}S_{t-1} + X_t \tag{4}$$

Of course, from the SIR model above, the annualized infection rate is likely to be a complicated function of the contact pattern and the level of susceptibility in the population. We propose to model β_t as an increasing function of the proportion of the population that is susceptible S_t/N_t, which phenomenologically reflects the behavior of herd immunity in the standard SIR model. Thus we write equation (4) as

$$S_t = S_{t-1} - \left(1 - \exp\left(-\theta_1 \frac{S_{t-1}}{N_{t-1}}\right)\right) S_{t-1} + X_t \tag{5}$$

The annualized infection rate goes to 0 as the proportion of susceptibles declines and increases to 1 in a fully susceptible population at a rate governed by the parameter θ_1. Note that measles is highly infectious and the classic SIR models would predict that $> 95\%$ of individuals would become infected in a naive (i.e. fully susceptible) population (Anderson and May (1991)). Thus the assumption that the annualized infection rate goes to 1 is not unjustified. Routinely, large-scale vaccination campaigns are conducted to supplement routine vaccination programs. These campaigns are designed both to provide an opportunity for a second dose in areas where only one routine dose is available, and to provide an opportunity for a first dose in areas where routine coverage is low. As such these campaigns reduce the total susceptible pool, rather than just the birth cohort. If the coverage for such a campaign in year t is Y_t, then the entire susceptible pool in the subsequent year is reduced,

$$S_t = [S_{t-1} - \left(1 - \exp\left(-\theta_1 \frac{S_{t-1}}{N_{t-1}}\right)\right) S_{t-1} + X_t](1 - Y_{t-1}) \tag{6}$$

and the number of incident cases in each year is

$$I_(t-1) = \left(1 - \exp\left(-\theta_1 \frac{S_{t-1}}{N_{t-1}}\right)\right) S_{t-1} \tag{7}$$

To initiate the model, we need the number of susceptible individuals prior to the first observation. As this is generally not known, we set $S_0 = \theta_2 N_0$, where θ_2 is the initial proportion of the population that is susceptible to measles, and N_0 is the population size at that time. National reports of measles cases are generally negatively biased as not all infected cases present with severe enough symptoms to seek medical care. We assume that there is time invariant, under-reporting of cases such that the expected number of cases at time t is $C_t = \theta_3 I_t$.

3. Inference Using State Space Models

We now have a plausible model for the evolution of the number of susceptibles and the number of infections given several input variables such as the number of births and childhood vaccination coverage, which together give X_t above, the population size, and the year and

coverage of pulsed vaccination campaigns and parameter values. How then can we infer the parameter values and also predict the true number of infections given the reported number of infections and the input variables? In some ways we would like to regress the number of infections on the input variables; however, the number of reported infections is a function of susceptibles instead of being directly dependent on the system parameters. We can treat the number of susceptibles as a hidden variable, important to the system but essentially not observable.

We write the model then as follows. The observation equation is

$$C_t = \theta_3 \left(1 - \exp\left(-\theta_1 \frac{S_{t-1}}{N_{t-1}}\right)\right) S_{t-1} + \epsilon_t \tag{8}$$

and the system equation is given as

$$S_t = [S_{t-1} - \left(1 - \exp\left(-\theta_1 \frac{S_{t-1}}{N_{t-1}}\right)\right) S_{t-1} + X_t](1 - Y_{t-1}) + \eta_t, \tag{9}$$

where ϵ_t and η_t are independent in time and with respect to one another, Gaussian random variables with mean 0 and variance θ_4^2 and θ_5^2 respectively.

3.1. State Space Models

We start with a brief discussion of state space models which will form the basis of our approach to parameter estimation. Linear state space models have a history dating back to the 1960's and are commonly used to estimate parameters in discrete index, Gaussian time series (Harvey and Phillips, 1979; Brockwell and Davis, 1991). A closely related concept to the linear state space model is the Kalman filter. A filter in this context is the conditional density of the state of a system (S_t in our case) at time t given the observed part of the system up to time t $(C_1, ..., C_t)$. If we assume that the system is Gaussian, as we do in the linear setting, then the conditional mean (which we denote by $\hat{S}_t$) and variance (which we denote by P_t) is sufficient to describe that density. In addition to the filter, a smooth may also be calculated where all of the observed values both past and future states are used to predict S_t.

Not only does the calculation of the Kalman filter allow us a "best guess" for the state at time t given the observations through the conditional expectation, evaluation of the filter provides the value of the likelihood function of the system for a given set parameter values. This later fact will allow us to use the Kalman filter to numerically optimize this likelihood function to give maximum likelihood estimators for the parameters. As the proposed model for measles burden is non-linear (equation 9), we will use a standard linear approximation approach to the non-linear system in order to fit our proposed model; known as the extended Kalman filter (Elliott et al. (1995)). Since there are a number of excellent resources for linear state space models, we omit the details and direct the reader to Shumway and Stoffer (2000), Brockwell and Davis (1991), or Elliott et al. (1995) for more details.

The extended Kalman filter (EKF) makes two important simplifications

(a) That the use of a linear approximation to the non-linear state model will not severely bias the estimates, and
(b) That the error process is Gaussian.

Clearly, the second assumption does not represent the observation process well; observations are restricted to whole numbers, greater than zero and more likely to be either binomial or Poisson. However, in practice it is unclear the relative impact of these two assumptions on estimation of the state variables.

Below, we will compare the EKF to two alternate particle filter approaches to quantify the impact of the simplifications. By using particle filters to approximate the likelihoods, we can both simulate directly from the non-linear state model and assume a non-Gaussian observation model. One comparison will assume the same underlying model as the EKF, but approximate the filter and thus the likelihood function using Monte Carlo as opposed to a linearization. The other comparison will contrast the EKF with a particle filter technique using a binomial error distribution. When using the particle filters, we will use a Bayesian framework to estimate the parameters. The reason for this is computational; while the particle filters allow us to evaluate the approximate likelihood function in a different way, optimization over quantities that include Monte Carlo sampling error can be difficult.

Another advantage of state space models in general is the simplicity of handling missing data. The state space framework allows for prediction for future values of the underlying system, S_t in this context. Since there is no additional observation error, then there are obvious ways to update the filter. For a more detailed account. For example, see Brockwell and Davis (1991). For the input variables, we also encounter missing values. However, since we are viewing these as exogenous deterministic quantities, we use linear interpolation to handle these missing values.

3.2. Estimation and Prediction

Using the extended Kalman filter, we can then calculate an approximate likelihood function of our time series for a given set of parameter values. Our system has five parameters, and non-linear optimization even over so few parameters can be difficult (see the Appendix for detailed discussion of the algorithm we used). It is worth noting here that there are modern methods that allow for a possibly more precise calculation of the likelihood using simulation based methods (See Breto et al. (2009) and Ionides et al. (2006)), and we will use these as a comparison. However, one goal of this project was to design a method that was sufficiently tractable to be used by policy makers. Thus, we propose a method that requires only evaluation of the likelihood, rather than approximation of the likelihood through simulation, which may be computationally quite intensive; we provide a comparison to simulated particle filters below.

In our example, estimating the variance parameters of the random perturbations proved to be delicate. Through simulation studies we discovered that naive optimization using a quasi-Newton method (Byrd et al. (1995)) would cause either one or the other variance parameter to go to zero. By plotting the likelihood surface for those two parameters we saw that a particular geometry gave rise to such a situation (Figure 1).

When using the linear Kalman filter, there is a simple expression for the likelihood estimator of the variance of the observation error if we could observe the entire system

$$\hat{\sigma}_w^2 = \frac{1}{n}\sum_{t=1}^{n}(C_t - g\hat{S}_t)^2 + g^2 P_t,$$

where $\hat{S}_t$ is the one step ahead predictor for S_t and P_t is the one step ahead prediction error. (This can obviously be modified for the EKF.) If this were the only unknown param-

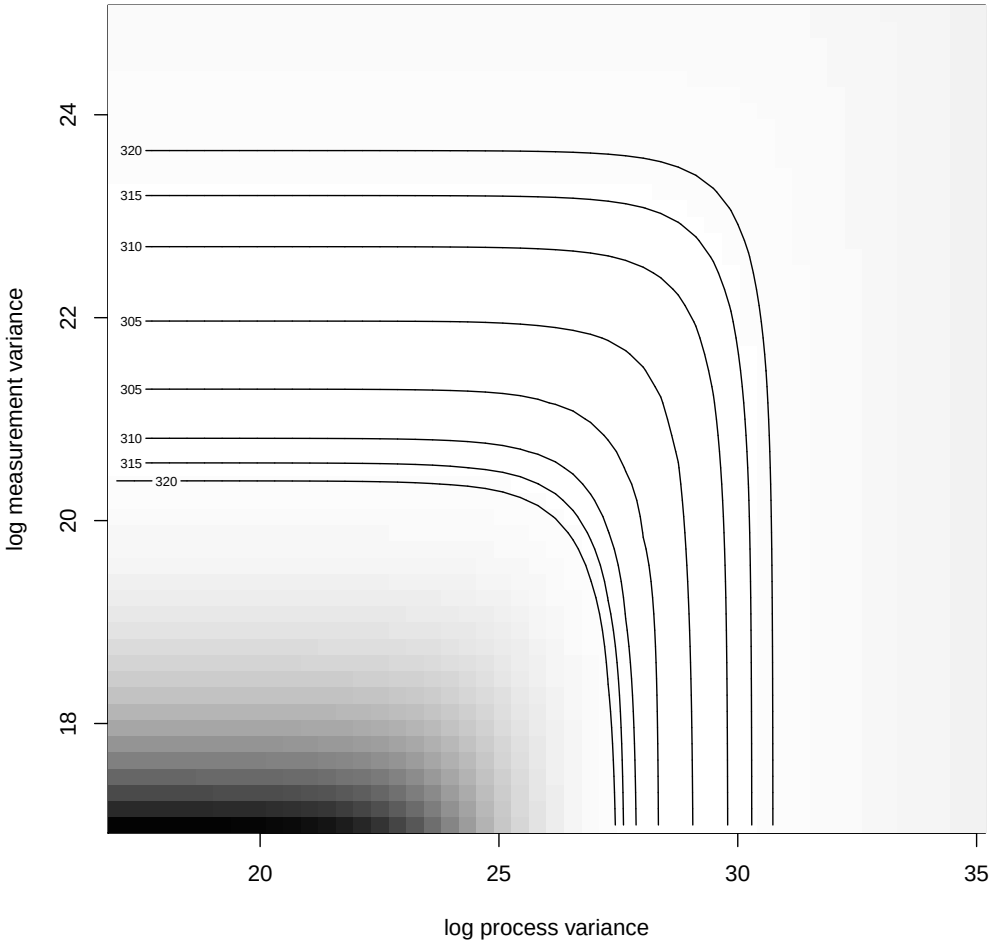

Fig. 1. Example profile negative log likelihood surface for the process (x-axis) and measurement (y-axis) components for an example dataset (Nigeria, see below). Shading (dark is large) and contours give negative log likelihood .

eter, we could perform a simplified EM algorithm to estimate this by alternating between evaluating the smoother and obtaining this estimate for σ_w^2 and using this estimate in the next evaluation of the smoother.

By doing this, we optimize the likelihood over σ_ϵ^2 using non-linear numerical optimization, but for each evaluation of the likelihood we estimate the variance of the observation equation using the above method. Moreover, we have found that the estimation of the other parameters are somewhat robust to the values of the variance parameters. So, we first estimate the non-variance estimators. Then, optimize over the variance parameters and iterate until convergence. In practice, only a few iterations are needed for convergence. One of the benefits of likelihood (or more accurately in this case approximate likelihood) estimation is the ability to obtain approximate standard errors by evaluating the hessian of the log likelihood function at the optimal point.

As previously discussed, estimating the parameters is a necessary step for our primary goal of predicting the true number of incident cases in the population. This was given in equation 7 and is a function of the number of susceptibles for which our method gives a prediction, $\hat{S}_t$, for each time. Moreover, our proposed method also has a prediction variance implying that we have an approximate 95% prediction interval of

$$\left(\hat{S}_t + z_{0.025}\sqrt{P_t}, \hat{S}_t - z_{0.025}\sqrt{P_t}\right)$$

where $z_{0.025}$ is the 2.5th percentile of a standard normal distribution.

Since the incident cases is an increasing, invertible function (over the relevant range) of $\hat{S}_{t-1}$, we can transform this interval to the corresponding interval for I_t. For simplicity we define this increasing function as

$$I_t = h(S_{t-1}) = \left(1 - \exp\left(-\theta_1 \frac{S_{t-1}}{N_{t-1}}\right)\right) S_{t-1}$$

Then, we can write the prediction interval for I_t as

$$\left(h\left(\hat{S}_t + z_{0.025}\sqrt{P_t}\right), h\left(\hat{S}_t - z_{0.025}\sqrt{P_t}\right)\right)$$

4. Examples

4.1. Application to Measles Surveillance Data

To illustrate the application of the EKF to annual surveillance data we applied the method to annual time series of measles incidence from 1980 to 2007 for 4 countries: Nigeria, Bolivia, Cambodia, and Pakistan (Figure 2). We chose these 4 countries to reflect a range of dynamics from highly endemic measles in Nigeria, to Bolivia, where measles has been largely extirpated through vaccination; thus these examples cover a full range of dynamic behavior as well as demography and vaccination history. All member states of the WHO are requested to report their annual cases counts and vaccination coverage for diseases and vaccines included in their national immunization program (WHO (2009a)), and data are available at (WHO (2009b)).

In practice, we first apply the optimization algorithm (above) to estimate the parameters θ_1 through θ_5 for each country (Table 1). Given those parameter estimates, we apply the EKF at the estimated parameter values to produce predictions of the unobserved time series of susceptibles, $\hat{S}_t$, and the associated variances P_t. We construct approximate 95%

Table 1. Parameter estimates for 4 example countries.

Parameter	θ_1	θ_2	θ_3	$log(\theta_4)$	$log(\theta_5)$
Nigeria	5.02	0.20	0.03	20.0	21.6
Bolivia	1.34	0.50	0.01	25.0	14.3
Cambodia	1.0	0.50	0.05	22.6	18.65
Pakistan	79.0	0.01	0.005	19.9	19.1

prediction intervals for the susceptible time series as $\hat{S}_t \pm z_{0.025}\sqrt{P_t}$. We can then transform the predictions and bounds for the unobserved susceptibles into predictions and intervals for the unobserved measles cases using the transformation $I_t = \left(1 - \exp\left(-\hat{\theta_1}\frac{S_t}{N_t}\right)\right) S_t$; where $\hat{\theta}_1$ is the maximum likelihood estimator of θ_1.

One of the reasons for using maximum likelihood is to reduce bias and to use the standard errors derived from the maximized likelihood; however, we are far from the asymptotic regime. To gauge the reliability of standard errors, we simulated 10000 iterations of the fitted, annual model (e.g. equations 8 and 9), using the observed levels of population size and vaccination as covariates from the 4 countries described above: Nigeria, Bolivia, Cambodia, and Pakistan. For each of the simulated iterations, we fit the model (above) using the EKF and compare the resulting fitted parameter values to the true values. For all 4 example countries, the estimates of θ_1 tend to be positively biased, and in 3 of the 4 example countries (Cambodia, Nigeria, and Pakistan) the estimates of θ_2 tended to be negatively biased (Figure 3). However, the observation rate, θ_3 was in general well estimated and the variation in the estimates was comparatively low (Figure 3).

4.2. Performance on Simulated Data

Clearly, measles transmission dynamics occur on a much faster scale than the 1-year time step of the process model described above. The unobserved epidemic process is assumed to follow the TSIR epidemic model, which has been previously shown to replicate measles dynamics well for a variety of settings (Bjornstad et al. (2002), Ferrari et al. (2008), Metcalf et al. (2009). Previous methods have been proposed to estimate under reporting (Finkenstadt and Grenfell (2000)) or fit a full state-space model (Morton and Finkenstadt (2005)) with bi-weekly surveillance data. However, in practice, such highly resolved data are only available for a handful of settings, and large-scale burden estimation must be based on annual surveillance. Thus, a more rigorous test of the method is its ability to estimate incidence from dynamics that are simulated at the bi-weekly scale, but aggregated at the annual scale to reflect the true observation process. Let I_t, be the true number of newly infected individuals at time step t, and S_t the true number of susceptible individuals. I_{t+1} depends stochastically on the number of infectious and susceptible hosts at time t such that $I_{t+1} \sim$ binomial(S_t,$1 - exp(-\beta_t I_t^{\alpha})$), where β_t is the time-specific transmission rate and α is a parameter to account for non-linearities in transmission. The transmission rate, β_t, was assumed to be a sinusoidal function of time ($\beta_t = 20(1 + .2cos(2\pi t/26))$) to account for the commonly observed seasonality in measles transmission. We chose a value of $\alpha = 0.97$ which is consistent with the range seen for a variety of settings (Bjornstad et al. (2002), Ferrari et al. (2008)). We use a time step of two weeks, which is the average infectious generation time (the time from infection to recovery and immunity) for measles. New infections are drawn from the pool of susceptible individuals, which is, in turn, replenished by births. Thus, the number of susceptibles at time t is given by $S_{t+1} = S_t - I_t + 1 + (1 - V_t)B_t$, where

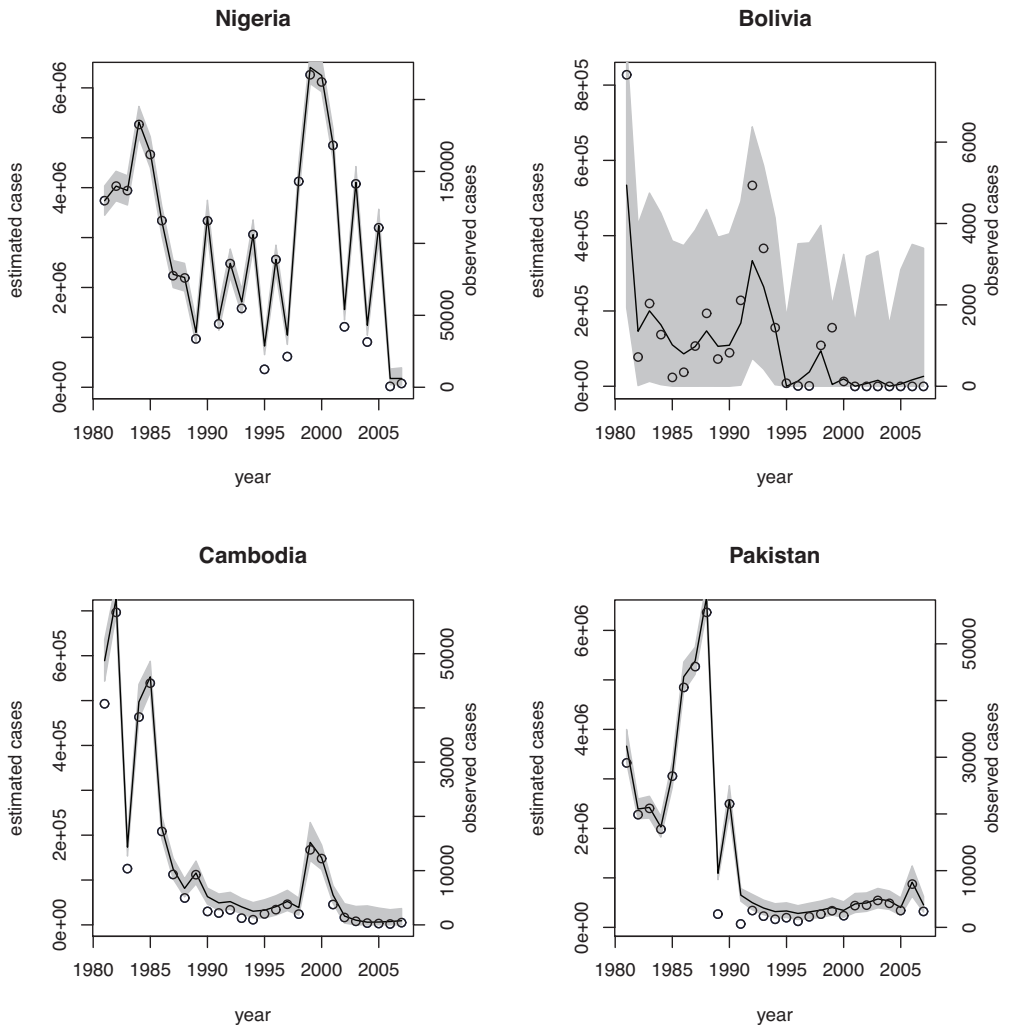

Fig. 2. Filtered state estimates of the unobserved measles cases for 4 example countries. A) Nigeria, B) Bolivia, C) Cambodia, D) Pakistan. Black curves give the reconstructed estimate, the shaded grey region indicates an approximate 95% prediction region. The circles in each panel indicate the reported measles cases divided by the estimated reporting fraction.

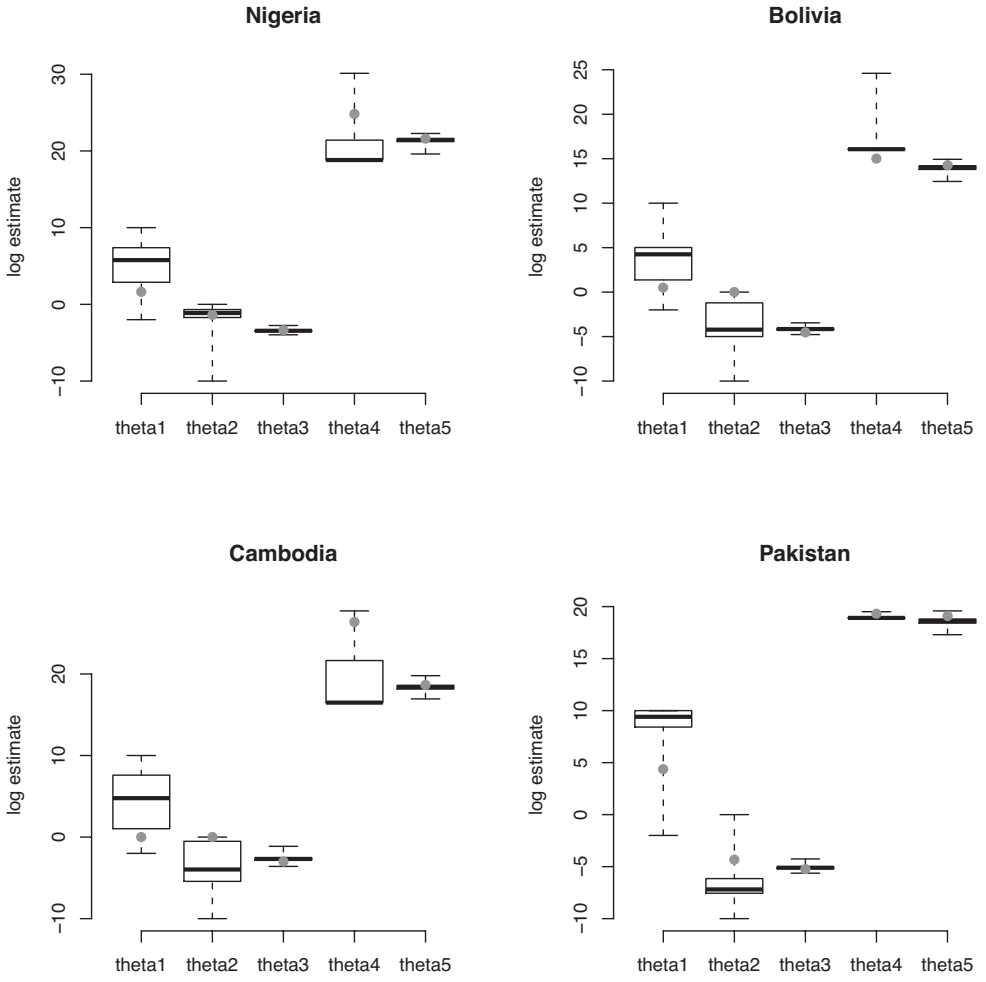

Fig. 3. The distribution of log parameter estimates for 10000 simulated time series based on the fitted values for the 4 example countries. Boxes indicate the central 50% of estimates, whiskers extend to the range of estimates. A)Nigeria, B) Bolivia, C) Cambodia, D) Pakistan. The grey circles indicate the true value of each parameter in the simulation.

B_t is the number of births and V_t is the proportion of the birth cohort that is vaccinated. We conducted these simulations under a large and small country sceniario. For the former we assumed that the population size, and birth rate were the same as for Nigeria: total population of 50 million, and 3406000 births per year, equally distributed throughout the year. For the latter we assumed a population with 10% of Nigeria's population size and annual births. Simulations were run for 56 years, with only the incidence in the second half of the time series uses for estimation. Annual vaccination coverage was assumed to be 50% for the first 28 years of simulation. Vaccination increased linearly from 50% to 80% in years 29-43 and was held constant at 85% for the final 14 years to mimic the trend of increasing vaccine coverage seen in many countries around the world (Figure 4 A).

The true number of infected cases in each year, I_{year}, of the simulation was the sum of the 26 2-week periods in the year. Reported cases in any given year were then assumed to be under-reported at a rate p_t, where p_t is a beta distributed random variable with mean θ_3 and variance θ_5. We simulated 5000 iterations of the bi-weekly model at each of 5 levels of observation rate, θ_3, ranging from 0.05 to 0.25 and corresponding variances, $\theta_5 = 0.01\theta_3$. For each we fit the EKF model to evaluate the estimates of the true number of cases and reporting rate.

The annual state space model appears to provide reasonably unbiased predictions of the unobserved time series of cases generated from the bi-weekly model, even under a range of observation rates (Figures 4 and 5). The estimated reporting rate trends positively with the true reporting rate, θ_3 (Table 2). The variance in residuals between the true and reconstructed time series tends to increase with lower reporting rate. The large residuals in the early part of the time series reflect the uncertainty in the initial state values. Thus, for application to burden estimation, it is advisable to use time series with observations earlier than the period of interest so that the uncertainty due to initial conditions is resolved by the time period of interest.

We further compared the prediction of true cases and reporting rate from the EKF to those from a) a particle filter with Gaussian system and observation model corresponding directly to the EKF and b) a particle filter with a Gaussian system model and a binomial observation model where the observation rate in any given year is a beta distributed random variable with constant mean (i.e. the same form as the simulation model). The comparison to the former allows us to evaluate the impact of the linear approximation to the process model made in the EKF because, in the particle filter, the state model is simulated directly from the underlying non-linear model (equation 9). The comparison to the latter allows us to evaluate the impact of the assumption of a Gaussian observation model. As the implementation of the particle filters is computationally intensive, we compare the fits using 500 simulated time series. The implementation of the particle filters is described in Appendix.

While the transmission parameter for the 2-week model and annualized model are not directly comparable (due to the differences in model structure) we can directly compare the estimated reporting rate from the EKF and the particle filters to the true value from the simulations (Table 2). For both the small and the large population simulations the estimated reporting rate appears accurate for simulations with reporting rate ranging from 0.05 to 0.25. Over 500 simulated time series the variance in the estimated reporting rate was largest for the EKF and smallest for the binomial particle filter (Table 2). We calculated the sum of the absolute differences between the true time series of incidence to the reconstructed time series of incidence from the EKF, the Gaussian particle filter, and the binomial particle filter. For the large population simulations, under all three fitting methods, the sum of

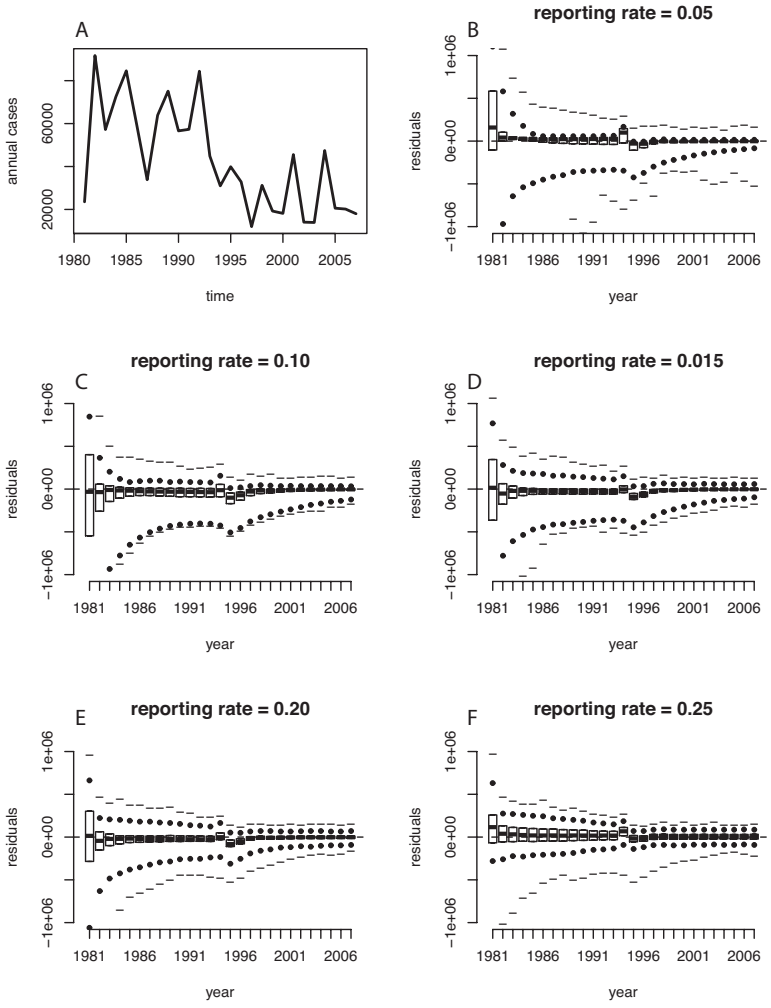

Fig. 4. A. A typical realization of the annual reported measles cases from one run of the large-population simulation, with 10% reporting. B-F. The distribution of the residuals between the true number of measles cases and the reconstructed values of from the EKF shown as a function of the time point in the simulation. Each boxplot shows the distribution from 5000 simulation runs. The panels B-F give the results of simulation runs at 5 levels of the reporting rate, θ_3, ranging from 0.05 to 0.25. Boxes show the central 50% of residuals, solid circles give the central 95% of residuals, and dashes give the range.

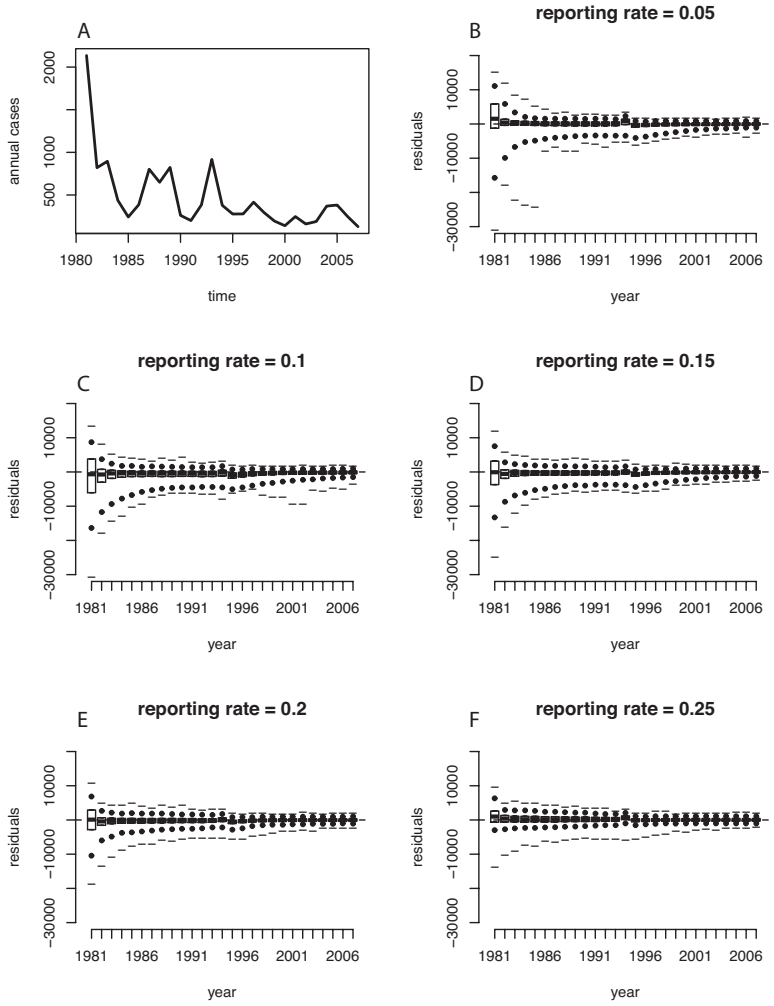

Fig. 5. A. A typical realization of the annual reported measles cases from one run of the small-population simulation, with 10% reporting. B-F. The distribution of the residuals between the true number of measles cases and the reconstructed values of from the EKF shown as a function of the time point in the simulation. Each boxplot shows the distribution from 5000 simulation runs. The panels B-F give the results of simulation runs at 5 levels of the reporting rate, θ_3, ranging from 0.05 to 0.25. Boxes show the central 50% of residuals, solid circles give the central 95% of residuals, and dashes give the range.

Table 2. Estimated reporting rate for the large and small populations using the EKF, the Gaussian particle filter (PF), and the binomial PF for time series generated under the bi-weekly model. Values in brackets are 95% confidence intervals for the EKF and the 0.025 and 0.975 quantiles of the posterior distribution for the PF examples.

		True Reporting Rate				
		0.05	*0.10*	*0.15*	*0.20*	*0.25*
large population	EKF	0.051 (0.04-0.06)	0.094 (0.06-0.12)	0.145 (0.10-0.16)	0.195 (0.15-0.23)	0.254 (0.23-0.28)
	Gaussian PF	0.049 (0.04-0.06)	0.098 (0.08-0.11)	0.147 (0.13-0.17)	0.194 (0.17-0.22)	0.245 (0.22-0.27)
	Binomial PF	0.051 (0.04-0.06)	0.10 (0.09-0.11)	0.15 (0.13-0.16)	0.198 (0.18-0.22)	0.247 (0.23-0.27)
small population	EKF	0.051 (0.04-0.06)	0.093 (0.06-0.12)	0.145 (0.10-0.17)	0.196 (0.156-0.228)	0.254 (0.22-0.28)
	Gaussian PF	0.049 (0.04-0.06)	0.098 (0.08-0.11)	0.147 (0.13-0.17)	0.196 (0.17-0.22)	0.245 (0.22-0.27)
	Binomial PF	0.048 (0.03-0.07)	0.093 (0.07-0.14)	0.140 (0.11-0.20)	0.185 (0.14-0.27)	0.229 (0.22-0.28)

absolute difference between the true and predicted time series decreased for higher reporting rates (Figure 6A). Across all levels of reporting, the sum of absolute difference was lower for the two particle filter methods, and the distribution of the sum of absolute difference was indistinguishable between the two particle filter methods (Figure 6A). This suggests, that some precision is lost when using the EKF as a result of the linear approximation, but very little precision is lost by approximating the observation error process as Gaussian. The results were similar when the three methods were used to fit the annualized model to data simulated from a population of 10% the size. One relevant difference was that, in the small population simulations, the binomial particle filter tended to under-estimate the reporting rate (Table 2). This led to a consequent bias in the estimated true incidence and a slightly higher sum of absolute error (Figure 6B). In the application of the binomial particle filter, simulated values of the true incidence that were lower than the observed incidence have non-finite likelihood weights. As such, the binomial observation model results in severe particle depletion which may result in this slight bias. It is possible that this could be overcome with much larger computational effort (i.e. increasing the number of particles).

5. Conclusions

In many applications, surveillance data are analyzed with the goal of conducting inference on the underlying dynamics; i.e. the parameters of the state model (Ferrari et al. (2008), Breto et al. (2009), Cauchemez et al. (2008)). In that setting, the unobserved states are treated as nuisance parameters. The problem of burden estimation is peculiar in that inference on the unobserved states is the end goal, and we integrate over our uncertainty in the underlying model parameters. The current, natural history based methods for estimating disease burden also make use of an underlying model, but make the implicit assumption that the parameters of that model are fixed and known (Stein et al. (2003), Wolfson et al. (2007)). Thus, statements of uncertainty in burden estimates rely on *ad hoc* incorporation of variation to account for parameter uncertainty (i.e. see Wolfson et al. (2007)). The key benefit of state space models is that they provide a transparent and repeatable framework for calculating uncertainty in the prediction of the unobserved states and in the estimation of the underlying model parameters. And, because the state space model is rooted in observed surveillance data, the resulting estimates are not dependent on subjective statements of parameter uncertainty.

The model we have presented here performs well at reconstructing the broad-scale trends in annual incidence, though fails to account for some of the fine scale dynamics seen in the

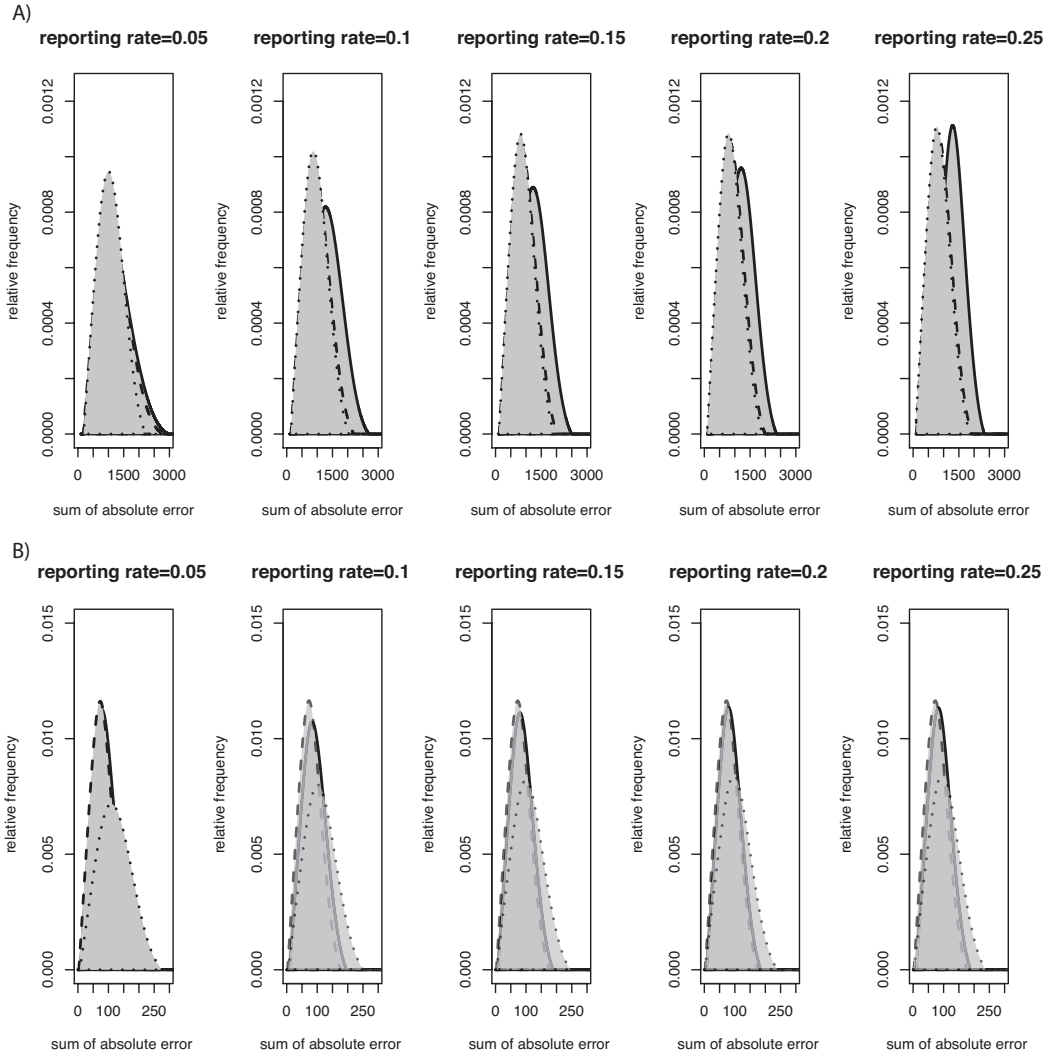

Fig. 6. The distribution of the sum of absolute errors between the true annual cases and the predicted annual cases from the EKF (solid line), the Gaussian particle filter (dashed line), and the binomial particle filter (dotted line) for simulations from A) the large population setting and B) the small population setting. Panels from left to right correspond to simulations with observation rates 0.05, 0.10, 0.15, 0.20, 0.25 respectively.

bi-weekly model. The estimator of the reporting rate, θ_3 was generally unbiased and had low standard error. While the EKF is computationally more tractable to apply, it does come a cost of slightly reduced precision in the estimates. The Gaussian particle filter, which does not make a linear approximation to the process model, and the binomial particle filter, which does not make the linear approximation and uses the true observation model, resulted in slight improvements in the precision of the estimate of the reporting rate and the uncertainty in the estimates of the unobserved time series of incidence cases. Thus, for evaluating large scale trends in measles incidence at the national or global scale, the EKF provides accurate estimates of the reporting rate and predicted unobserved annualized incidence despite the linear approximation and the assumption of Gaussian error in the observation model.

These annualized models are likely too coarse to capture the underlying non-linear dynamics that give rise to the complex behavior that is often seen in fine scale models of measles dynamics (Earn et al. (2000)). Thus, it is difficult to make a direct comparison between the annualized transmission function and the fine-scale transmission rate in the two-week model. In particular, given the seasonal nature of measles transmission (Ferrari et al. (2008), Bjornstad et al. (2002)) it may be challenging to disentangle the impact of seasonal variation in transmission relative the mean transmission rate without data at a finer resolution. While it is possible to fit a fine-scale model using analogous state-space models (Ferrari et al. (2008), Breto et al. (2009), Cauchemez et al. (2008)) to annualized data, it is uncertain what the use of fine-scale models would add to the estimation of burden. Fine-scale models are clearly important for the goal of understanding the non-linear dynamics of transmission and for making forward predictions. While it is beyond the scope of this work, the extent to which the signature of fine-scale, non-linear dynamics can be detected in annualized surveillance remains and interesting question. The real practical benefit of the state space model presented here is the explicit framework for combining a dynamic model for incidence with surveillance data to obtain objective predictions of burden and their associated uncertainty.

State space approaches to burden estimation present three important programmatic benefits for policy applications; they are rooted in surveillance data, objective, and flexible. The availability and quality of disease surveillance data is a significant challenge in public health. Because the inference for state space models is grounded in the surveillance data, rather than expert opinion alone, the data quality should be reflected in the prediction bounds on the unobserved states. The nature of the state reconstruction, in the Kalman filter and other methods, is such that addition of new data improves inference of the parameters and thus prediction at all time steps. Thus, state space models provide an incentive to continue and improve surveillance.

In general, the data are collected at a local (provincial or national) scale, reported from health facilities to subnational and national authorities, and are principally analyzed at a regional (national or international) scale. The explicit use of the local scale data in estimating burden and the resulting policy has the benefit of facilitating communication with the local levels. Prior burden estimation methods have imposed subjective classifications on high reporting and low reporting countries and conducted different analyses for each (Stein et al. (2003), Wolfson et al. (2007)). While justifiable, these methods are subject to criticism by the local (national) constituents as to the classifications used. The state space model approach we have proposed here is objective and equitable in the sense that the same algorithm is applied to all data, and the reconstruction of the unobserved states (and the associated estimates of reporting) are based only on the input data.

As with any analysis, the inference is only as good as the underlying model and quantity

of data will allow. We have presented fairly simple state and observation equations for the sake of illustration. In particular, it seems unlikely that the reporting rate, θ_3, would be constant through time. Valid arguments could be made that the reporting rate should be higher during outbreaks due to heightened awareness, or should scale with the vaccine coverage rate (i.e. increased investment in public health), or transitions to new surveillance protocols. While addressing this question is beyond the scope of this work, it is worth noting that analysis of such candidate models for, say, the functional relationship between reporting rate and vaccine coverage is a straightforward extension of these methods. Clearly there are many candidate models for both the state and observation equations that could be considered. Indeed, it may be that the appropriate model has not yet been developed. A benefit of state space models is that they are based in the formal framework of likelihood, which allows for an objective comparison of candidate models and a basis for the adoption of new models as they are proposed.

The reconstruction of imperfectly observed time series using state space models presents a useful tool for policy. These methods provide a solution to the problem of developing an objective correction for under-reported surveillance data. Further, the explicit use of surveillance data to derive both predictions with error bounds places a measurable value on collecting high quality surveillance data and facilitates the communication of policy decisions. Measles is peculiar in that the relatively simple epidemiology limits the number of unobserved states to only susceptible and infected individuals. Developing state space models for other pathogens that might have long latency periods or unobserved carriers (e.g. pertussis and meningitis) would be more difficult, though is at least theoretically amenable to these methods.

6. Conclusions

The authors would like to acknowledge support from World Health Organization and the Inference for Mechanistic Models Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (grant no. DEB-0553768), the University of California, Santa Barbara and the State of California.

References

Anderson, R. M. and R. M. May (1991). *Infectious diseases of humans: dynamics and control*. Oxford: Oxford University Press.

Bjornstad, O. N., B. F. Finkenstadt, and B. T. Grenfell (2002). Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series sir model. *Ecological Monographs 72*(2), 169–184.

Breto, C., D. H. He, E. L. Ionides, and A. A. King (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics 3*(1), 319–348.

Brockwell, P. and R. Davis (1991). *Time Series: Theory and Methods*. Springer.

Byrd, R., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing 16*, 1190–1208.

Cauchemez, S., A. J. Valleron, P. Y. Boelle, A. Flahault, and N. M. Ferguson (2008). Estimating the impact of school closure on influenza transmission from sentinel data. *Nature 452*(7188), 750–U6.

Crowcroft, N. S., C. Stein, P. Duclos, and M. Birmingham (2003). How best to estimate the global burden of pertussis? *Lancet Infectious Diseases 3*(7), 413–418.

Dabbagh, A., R. Eggers, S. Cochi, V. Dietz, P. Strebel, and T. Cherian (2007). A new global framework for immunization monitoring and surveillance. *Bulletin of the World Heatlh Organization 85*(12), 904.

Doucet, A., N. De Freitas, and N. Gordon (2001). *Sequential Monte Carlo methods in practice*. Springer Verlag.

Earn, D. J. D., P. Rohani, B. M. Bolker, and B. T. Grenfell (2000). A simple model for complex dynamical transitions in epidemics. *Science 287*(5453), 667–670.

Elliott, R., L. Aggoun, and J. Moore (1995). *Hidden Markov models: estimation and control*. Springer.

Ferrari, M. J., R. F. Grais, N. Bharti, A. J. K. Conlan, O. N. Bjornstad, L. J. Wolfson, P. J. Guerin, A. Djibo, and B. T. Grenfell (2008). The dynamics of measles in sub-saharan africa. *Nature 451*(7179), 679–684.

Finkenstadt, B. and B. Grenfell (2000). Time series modelling of childhood diseases: a dynamical systems approach. *Applied Statistics 49*(2), 187–205.

Harvey, A. and G. Phillips (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika 66*(1), 49–58.

Ionides, E. L., C. Breto, and A. A. King (2006). Inference for non-lineaer dynamical systems. *Proceedings of the National Academy of Sciences 103*, 18438–18443.

McCallum, H., N. Barlow, and J. Hone (2001). How should pathogen transmission be modelled? *Trends in Ecology & Evolution 16*(6), 295–300.

Metcalf, C. J. E., O. N. Bjornstad, B. T. Grenfell, and V. Andreasen (2009). Seasonality and comparative dynamics of six childhood infections in pre-vaccination copenhagen. *Proceedings of the Royal Society B-Biological Sciences 276*(1676), 4111–4118.

Morton, A. and B. Finkenstadt (2005). Discrete time modelling of disease incidence time series by using markov chain monte carlo methods. *Applied Statistics 54*(2), 575–594.

Shumway, R. and D. Stoffer (2000). *Time series analysis and its applications*. Springer Verlag.

Stein, C. E., M. Birmingham, M. Kurian, P. Duclos, and P. Strebel (2003). The global burden of measles in the year 2000 - a model that uses country-specific indicators. *Journal of Infectious Diseases 187*, S8–S14. Suppl. 1.

Uzicanin, A. and L. Zimmerman (2011). Field effectiveness of live attenuated measles-containing vaccines: a review of published literature. *Journal of Infectious Diseases. (in press)*.

WHO (2009a). Who vaccine-preventable diseases: monitoring system - 2009 global summary. Technical report, WHO.

WHO (2009b, December). World health organization: Immunization surveillance, assessment and monitoring. Website.

Wolfson, L. J., P. M. Strebel, M. Gacic-Dobo, E. J. Hoekstra, J. W. McFarland, and B. S. Hersh (2007). Has the 2005 measles mortality reduction goal been achieved? a natural history modelling study. *Lancet 369*(9557), 191–200.

7. Appendix

In this appendix, we describe the particle filter framework that was used as a comparison. The basic steps of prediction followed by update that are seen in the extended Kalman filter are also used here. See Doucet et al. (2001) for a basic introduction to particle filters. However, the starting point for the recursion of the filter, the conditional distribution of the system at time $t-1$ given the data from 1 to $t-1$, is represented by a cloud of particles instead of a mean and variance. In this way, the representation of the distribution is more robust. We will use M particles to represent the filter at each time step. So, we again need to describe the recursion to obtain a new cloud of particles representing the distribution of S_t given $C_1, ..., C_t$ from the cloud of points representing the distribution of S_{t-1} given $C_1, ..., C_{t-1}$.

So, we start with M particles, $S^1_{t-1|t-1}, ..., S^M_{t-1|t-1}$, which represent the distribution, S_{t-1} given $C_1, ..., C_{t-1}$. Then we perform the following steps.

(a) Perform one step ahead prediction by simulating from the transition probability density for the system for each of the particles, $S^m_{t-1|t-1}$. We then have a new set of particles $S^1_{t|t-1}, ..., S^M_{t|t-1}$.
(b) Create weights using the current observation, C_t, for each of these predictions using the observation distribution of C_t given S_t, $w_m = p(C_t|S^m_{t|t-1})$.
(c) Standardize the weights, $w_m / \sum_{i=1}^M w_i$, and use these probabilities to sample m observations with replacement from $S^1_{t|t-1}, ..., S^M_{t|t-1}$. We denote the resulting particles by $S^1_{t|t}, ..., S^M_{t|t}$.

One important thing to note is that the second step allows us to approximate the contribution to the likelihood function at the tth time point. Specifically,

$$L(\Theta) = \prod_{t=1}^{T} p(C_t|C_{t-1}, ..., C_1) = \prod_{t=1}^{T} \int p(C_t|s)p(s|C_{t-1}, ..., C_1)ds \approx \prod_{t=1}^{T} \sum_{m=1}^{M} p(C_t|S^m_{t|t-1})$$

Note that maximizing this likelihood, especially over several parameters, could be quite challenging given that the evaluation of the likelihood contains sampling error from the Monte Carlo integration. Therefore, we use a Bayesian framework and simulate from the posterior distribution. To do this, we simulate a sample from the prior distribution, weight this sample with the likelihood value corresponding to that draw, and resample using these weights to arrive at a sampling from the posterior. For each simulated time series, we evaluated the posterior using 5000 draws from the prior distribution and evaluated the

Table 3. Comparison of the mean and quantiles (0.025th and 0.975th) of the prior and posterior distributions for the observation rate for the large population example.

	True Reporting Rate				
	0.05	*0.10*	*0.15*	*0.20*	*0.25*
Prior	0.1	0.1	0.1	0.1	0.1
	(0.007-0.35)	(0.007-0.35)	(0.007-0.35)	(0.007-0.35)	(0.007-0.35)
Post. Gaussian PF	0.049	0.098	0.147	0.196	0.245
	(0.04-0.06)	(0.08-0.11)	(0.13-0.17)	(0.18-0.21)	(0.23-0.26)
Post. Binomial PF	0.049	0.095	0.140	0.188	0.232
	(0.04-0.06)	(0.09-0.12)	(0.13-0.15)	(0.17-0.20)	(0.22-0.25)

likelihood for each using 1000 particles. We present the mean of the posterior distribution as the parameter estimate. We then evaluate the particle filter at the posterior means for all parameters, using 1000 particles, to estimate the unobserved annual incidence.

For both the large and small population simulations we chose priors that were uniform on the log transformed parameters for θ_1, θ_4, and θ_5 and uniform on the logit transformed parameters for θ_2 and θ_3. The prior bounds for the Gaussian particle filter were for θ_1 to θ_5 respectively were (0, 5), (-5, 3), (-5, -.5), (0, 20), (1, 20). The prior bounds for the Binomial particle filter for θ_1 to θ_5 respectively were (0, 5), (-5, 3), (-5, -.5), (8, 20), (0, 4.5). For both models and both population sizes, the posterior distribution was significantly different from prior distribution for the reporting rate (see Table 3 for a comparison of the posterior and prior for the large population simulations).