Mathematical Population Studies
2000, Vol. 8(1), pp. 1–29
Reprints available directly from the publisher
Photocopying permitted by license only

RECONSTRUCTING SUSCEPTIBLE AND RECRUITMENT DYNAMICS FROM MEASLES EPIDEMIC DATA

GEORGIY V. BOBASHEV*[,a], STEPHEN P. ELLNER[b], DOUGLAS W. NYCHKA[b] and BRYAN T. GRENFELL[c]

[a]*Statistics Research Division, Research Triangle Institute, Research Triangle Park, North Carolina, 27709*
[b]*Biomathematics Graduate Program, Department of Statistics Box 8203 North Carolina State University, Raleigh, North Carolina, 27695-8203*
[c]*Department of Zoology, Cambridge University, Downing St., Cambridge, CB2 3EJ, UK*

Dynamical epidemic studies are often based on the reported number of cases. For various purposes it would be helpful to have information about the numbers of susceptibles, but these data are rarely available. We show that under general theoretical assumptions it is possible to reconstruct, up to linear scaling parameters, the dynamics of the susceptible class, as well as the rate of recruitment to the susceptible class, based only on case report data. We demonstrate that susceptible data reconstructed by our method improve the performance of forecasting models. Our estimate of susceptible class dynamics also can be used to estimate the age distribution of recruitment into the susceptible class, if the birth rate is known from independent data. Simulation experiments show that the reconstruction is robust to errors in the reporting scheme. This work was motivated by measles in large developed-world cities prior to immunization

Acknowledgements: We thank Ben Bolker for numerous valuable discussions and assisting us in collating the data sets. This research was partially supported by NSF Grant DMS 92-17866 to Nychka, Ellner, and A.R. Gallant, and was a portion of the senior author's Ph.D. thesis in the Biomathematics Graduate Program at North Carolina State University.

* Corresponding author. Tel.: (919) 541-6167, Fax: (919) 541-5966.
E-mail: bobashev@rti.org

programs; our theoretical assumptions are empirically justified for measles but should also be applicable to some other diseases with permanent immunity.

KEYWORDS: Mathematical epidemiology; measles; susceptibility; forecasting; modeling

1. INTRODUCTION

In the present study we estimate the dynamics of population susceptible to measles and some other childhood infectious diseases using case report data. We illustrate how knowledge about susceptible population can improve epidemic forecasting models and provide additionl insight to the problems of disease transmission. Mechanistic models of childhood infectious diseases usually classify individuals according to disease stage, as either Susceptible, Exposed (latent), Infected or Recovered (SEIR). After birth an individual sequentially passes through each stage with certain transition probabilities. Such models are often termed SEIR models (Anderson and May, 1991; Ellner *et al.*, 1995; London and Yorke, 1973; Tidd *et al.*, 1993). These models can also include other demographic information such as age structure, e.g. the Realistic Age Structure (RAS) model (Anderson and May, 1991; Bolker and Grenfell, 1993). The study of mechanistic epidemic models is usually done qualitatively, because neither the exact values of the parameters, nor the exact functional form of all of the rate processes, is definitely known (Ellner *et al.*, 1995; Fine and Clarkson, 1982a, b; London and Yorke, 1973; Tidd *et al.*, 1993). Parameter estimation, and comparison of fitted models with data, are severely hindered by the fact that typically only one state variable of the real system is observable, namely the numbers of infectives. In some cases, serological data give crude estimates of the number of susceptibles (Grenfell and Anderson, 1985); however, the vast majority of data are case reports reflecting the number of infectives.

Among numerous infectious diseases, measles keeps an exceptional place in attracting investigators, both epidemiologists and mathematicians. In addition to the relative biological simplicity of the disease, the data are some of the best among studies of biological population dynamics (Grenfell *et al.*, 1995). Some typical examples of the available data are shown in Figure 1. Not surprisingly, these data sets are widely studied and have often been used for testing new hypotheses and models. In this paper, we use pre-vaccination measles data to develop and evaluate a new method for reconstructing the dynamics of the susceptible population, and of the recruitment rate into the

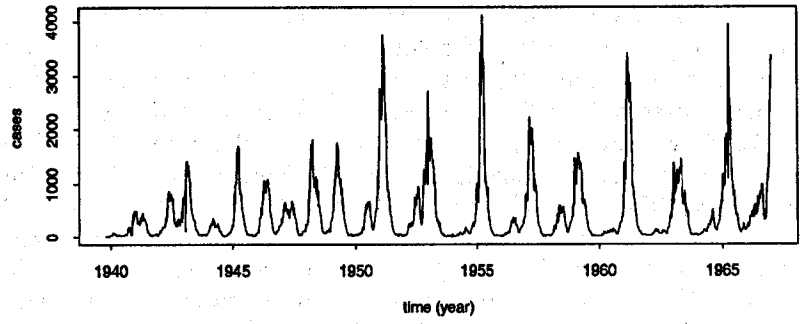
susceptible class. The time-delay between birth and recruitment to the susceptible class reflects the fact that a pre-school child, although immunologically susceptible to the disease, may have very limited contact with the infected school-age population and thus may not become truly susceptible for some time. For the remainder of this paper, susceptibility will always mean full susceptibility in this sense.

We show that the additional information about the susceptible dynamics is useful for model-fitting and forecasting, and also for gaining additional insight into the mechanistic causes of the observed dynamics. In particular, it can help to clarify the roles of fluctuations in birth rates (Grenfell *et al.*, 1995), and of family structure (Black, 1959). The effects of family structure on recruitment have been observed in serological studies (Black, 1959). In (Anderson and May, 1985, 1991; Grenfell *et al.*, 1995) the authors presented simple models of these effects, however, these models require detailed information on the percentages of families with different numbers of children, and assume that only children of certain ages are recruited. Having an estimate of recruitment makes it possible to directly relate changes in recruitment rates with changes in epidemic dynamics, and to examine how family structure affects the distribution of time between birth and the onset of full susceptibility.

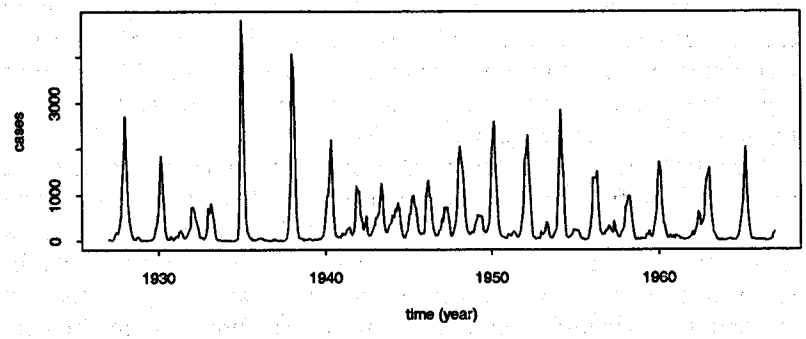
Most statistical forecasting models in epidemiology use state-space reconstruction from lagged case report data (Ellner *et al.*, 1995; Ellner and Turchin, 1995; Schaffer *et al.*, 1990). It is unclear, however, how many lags are needed in the model, and how the fitted model can be interpreted biologically. In order to identify a meaningful state space, it is useful to incorporate some mechanistic structure into the phenomenological statistical models. The number of cases in the near future certainly depends on the number of susceptibles as well as the number of cases at the present time (Anderson and May, 1991; Fine and Clarkson, 1982a). Therefore, models based on both time series together – cases and susceptibles – are likely to yield improvements in forecasting.

One of the first attempts to reconstruct susceptible dynamic was by Hedrich (1933), who used a simple model to estimate the dynamics of susceptibles under age 15 in Baltimore from 1899 to 1931. He used a mass-balance model where the increment in the total number of susceptibles was given by the sum of gains and losses due to births, immigration, infection, death, emigration, and "retirement" from the disease at the age of 15. This approach requires, however, a lot of additional demographic information. The model also assumes that all children become susceptible at the age of 6, when they first go to

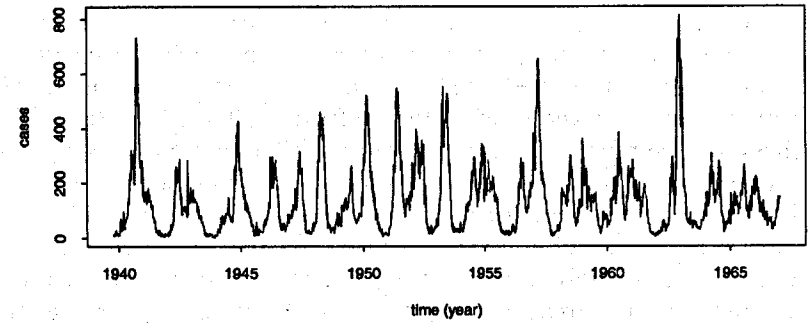
(a)


cases
0
1000
2000
3000
4000
1940
1945
1950
1955
1960
1965
time (year)

(b)


cases
0
1000
3000
1930
1940
1950
1960
time (year)

(c)


cases
0
200
400
600
800
1940
1945
1950
1955
1960
1965
time (year)

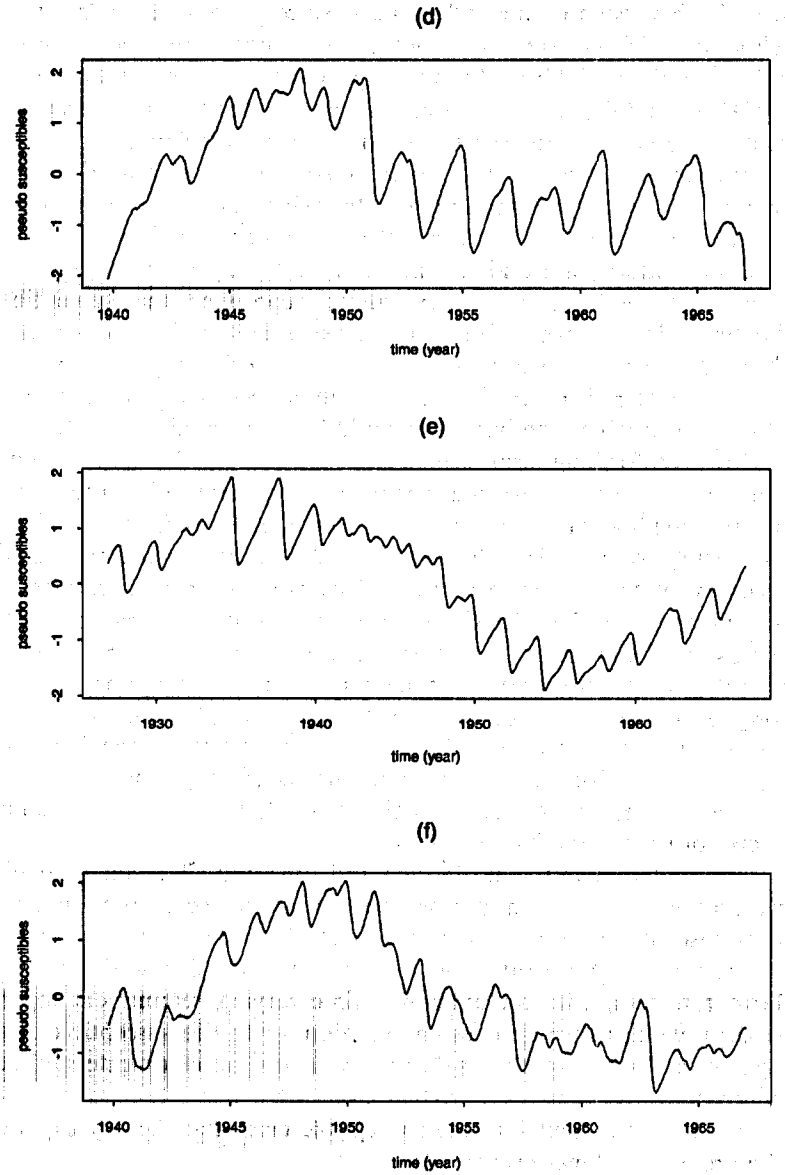


FIGURE 1 Measles case report data for (a) London, (b) Copenhagen, and (c) Liverpool, and (d, e, f) the corresponding pseudo-susceptibles $\hat{S}_t$ calculated from equation (7). London and Liverpool data from Registrar General for England and Wales, Weekly Reports; Copenhagen data from L.F. Olsen (*personal communication*).

school. This assumption is rather crude, since pre-school children with school-age siblings are also exposed to infections prevalent among older school-age children. A similar approach was taken by Cliff *et al.* (1981) who used detailed demographic accounting to reconstruct susceptible populations in their studies of measles in Iceland.

Fine and Clarkson (1982a) used a similar model that does not require as much information as Hedrich's, but it also assumed immediate transfer of newborns into the susceptible class. Knowing case-notification data and the initial value of susceptibles, Fine and Clarkson's model can be iterated to reproduce the susceptible dynamics. In their work, Fine and Clarkson (1982a, b) reconstructed the susceptible dynamics in order to estimate the force of infection (the infection rate per susceptible), using only several-year-long segments of data, over which the birth rate could be considered constant.

However, for long term reconstruction this method has limitations: it assumes a perfect reporting scheme and a constant birth rate. Testing this method on simulated data from the deterministic SEIR or stochastic RAS models with constant birth rate, results in almost exact recovery of the susceptible values. However, when it is applied to actual case data, the Fine-Clarkson method produces large long-time trends in the number of susceptibles, whereas in the SEIR/RAS models, neither constant nor time-varying birth rates produce such long-term trends. The discrepancy between the SEIR/RAS and the Fine-Clarkson-Hedrich models makes it unclear whether the dynamics of the susceptible class *really* contains large long-time trends, or whether these reconstruction methods are failing to describe some aspect of the susceptible dynamics.

As an alternative to the Hedrich and Fine and Clarkson's models, we propose here a method based on the rate of recruitment into the fully susceptible class (as defined above). We remove the assumption of the constant birth rate, and replace Fine and Clarkson's constant birth rate term with an unknown, time-varying recruitment to the susceptible class, which is estimated along with the susceptible class dynamics. We show that this change leads to a more accurate method, which not only estimates the dynamics of the susceptible class, but also estimates recruitment into the susceptible class and thereby explains the presence of long-term trends.

If we only have case report data, then our method estimates scaled susceptibles and recruitment (up to additive and multiplicative constants). With additional information about the actual numbers of susceptibles and infectives at two or more different times, we can

estimate the actual susceptible values. This information can sometimes be obtained from serological studies as in (Fine and Clarkson, 1982b; Grenfell and Anderson, 1985; Roden and Heath, 1977). With knowledge of the real birth rate, it is then possible to estimate the delay-kernel transforming birth rate into recruitment rate. Although the present study is focused on measles, the results can be extended to other childhood diseases with permanent immunity.

The paper is structured as follows. Section 2 contains the description of the mathematical model used for reconstruction, and then in section 3 we derive the reconstruction method. Applications of the reconstruction method are presented in section 4. In 4.1 we compare forecasting model with and without susceptibles as a state variable, and show that knowledge of estimated susceptibles improves the forecasting accuracy; in 4.2 we estimate the distribution of ages at which individuals recruit into the susceptible class, and discuss how family structure can influence recruitment. Finally, we summarize the results and consider applications to other diseases in section 5.

2. THE MODEL

We consider a standard SEIR model with compartments S, E, I, and R representing the numbers of susceptible, exposed (latent), infected and recovered individuals respectively. After birth, an individual is assumed to pass through S, E, I, and R classes sequentially, and eventually dies in the recovered class.

We require several additional assumptions, that are based on known measles epidemiology and previous analysis of measles data.

a) *Reported cases are the numbers of new infectives reduced by a constant fraction of under-reporting.*

Analysis of case notification, serological and birth data shows that although there might be significant short-term fluctuations in reporting rate (such as postal delays during Christmas vacations) the under-reporting coefficient is fairly stable for each particular city (Fine and Clarkson, 1982b; London and Yorke, 1973; Mollison and Ud Din, 1993).

b) *The latent and infectious stages have a fixed, definite length.*

This assumption is different from the usual SEIR assumption of exponentially-distributed stage durations, and also more realistic.

Keeling and Grenfell (1997) show that models with constant stage durations are better able to account for high-frequency oscillations in measles case-report data than models with exponential stage durations. For measles the latent and infectious stages last 6–9 days, and 6–7 days, respectively (Black, 1989). We assume that the 2–3 day variation in the lengths of the latent and infectious stages are negligibly small compared to the reporting intervals considered in the derivation of our method.

c) *The seasonal pattern of variation in the contact rate intensity is the same each year, and is nearly constant around the time of epidemic peaks.*

As was shown in (Fine and Clarkson, 1982(a, b); Grenfell and Anderson, 1985; London and Yorke, 1973; Mollison and Ud Din, 1993), the rate of contacts between susceptible and infective individuals has a strong seasonal pattern. In the same papers, it was estimated that the year-to-year variation in the seasonal pattern is minor (around $\pm 10-15\%$) compared to the amount of within-year seasonal variation. Epidemic peaks in case report data occur within a few weeks or months after Christmas. Because this is in the middle of school terms, the contact rate is constant or almost constant for at least several weeks around the epidemic peak.

d) *Mortality in the susceptible, exposed and infected classes is insignificant.*

This assumption is justified by the fact that measles is usually contracted before the age of 15, and on average the maximum effect of births on measles occurs at the age of 6 (Anderson and May, 1991; Black, 1989; Grenfell and Anderson, 1985).

e) *Net migration is negligible.*

We assume that the changes in the sizes of these classes caused by migration is negligible compared to the changes caused by infection and recovery within each individual city (Fine and Clarkson, 1982b).

f) *Variation in the number of individuals entering the infective compartment (i.e. the number of new cases) around an epidemic peak, is negligible in comparison with the size of the peak.*

This assumption essentially re-states the fact that a smooth function has a horizontal tangent at its maximum. This assumption is also supported by weekly and monthly case report data (see Figure 1(a-c)).

Our reconstruction method, like prior methods, is based on mass balance for the susceptible compartment, which can be written as

$$S_{t+\tau} = S_t - E_{t,\tau} + Q_{t,\tau}. \tag{1}$$

Here S_t is the number of susceptible individuals at time t, $E_{t,\tau}$ is the number of individuals moving into the Exposed compartment between times t and $t+\tau$, and $Q_{t,\tau}$ is the number of new recruits to the susceptible class between times t and $t+\tau$.

If the reporting period τ for cases is much longer than the duration of the latent period (e.g. monthly totals), then most new infections are counted during the reporting period when they occur. This implies that $E_{t,\tau} \approx I_{t,\tau}$. Similarly, if the reporting period is close to the latent period (weekly totals), then $E_{t,\tau} \approx I_{t+I,\tau} \approx I_{t,\tau}$. The only circumstance in which these arguments would fail is when the disease incidence is changing very rapidly. For the data considered here, and either weekly or monthly reporting periods, the changes are not rapid enough to create significant differences between $E_{t,\tau}$ and $I_{t,\tau}$ (Figure 2).

Then using the approximation $E_{t,\tau} \approx I_{t,\tau}$, and the assumption that a constant fraction of cases are reported, we can then re-write (1) as

$$S_{t+\tau} = S_t - \alpha C_{t,\tau} + Q_{t,\tau} \tag{2}$$

where $C_{t,\tau}$ is the number of new cases reported between t and $t+\tau$, and $1/\alpha$ is the fraction of cases that are reported.

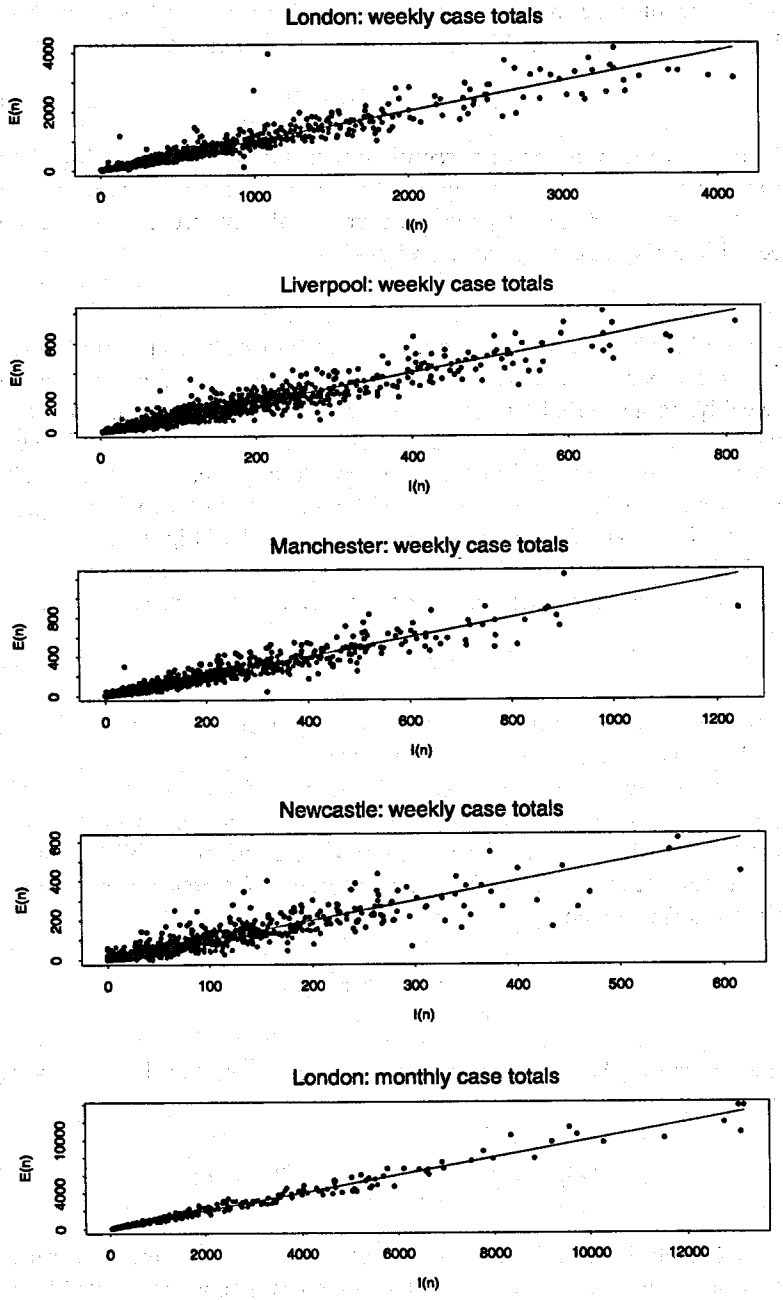
The recruitment to the susceptible class is composed of past births, and we assume that it can be expressed as a convolution of the birth rate $B(t)$, with a delay kernel $G(a)$ giving the probability that an individual becomes fully susceptible at age a. The recruitment rate at time t is then given by

$$\int_0^\infty G(t-u)\, B(u) du, \tag{3}$$

and therefore the total recruitment between times t and $t+\tau$ is given by

$$Q_{t,\tau} = \int_t^{t+\tau} \int_0^\infty G(v-u)\, B(u) du dv. \tag{4}$$

In (Black, 1989) a fixed six-year delay between birth and recruitment to susceptibility was assumed. This corresponds to a generalized kernel


London: weekly case totals
E(n)
0
2000
4000
0
1000
2000
3000
4000
I(n)
Liverpool: weekly case totals
E(n)
0
200
600
0
200
400
600
800
I(n)
Manchester: weekly case totals
E(n)
0
400
800
0
200
400
600
800
1000
1200
I(n)
Newcastle: weekly case totals
E(n)
0
200
400
600
0
100
200
300
400
500
600
I(n)
London: monthly case totals
E(n)
0
4000
10000
0
2000
4000
6000
8000
10000
12000
I(n)

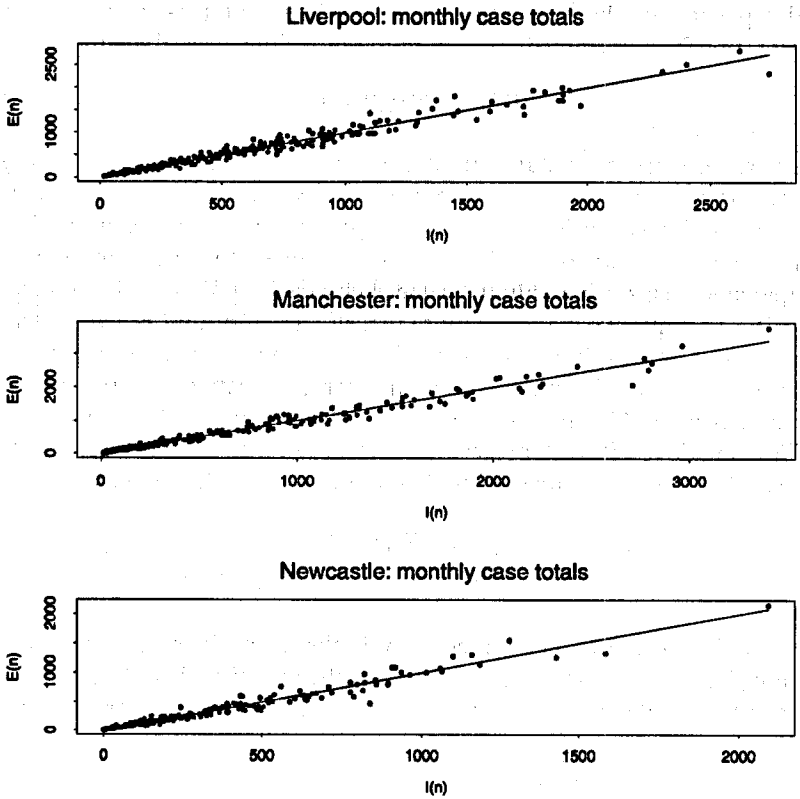


FIGURE 2 Evidence for the approximation $E_{t,\tau} \approx I_{t,\tau}$ when the reporting period is weekly or monthly. For cities with case data reported weekly, we can use the known duration of the latent period (approx. 1 week) to infer the number of new infections each week. Based on the reported cases and estimated number of infections, we compare E and I for weekly and monthly case totals. These results are typical of those obtained for large cities with weekly case data available.

function G that is infinite value at age six and zero everywhere else, yet the integral of G is over all ages is exactly 1 (a kernel of this type is called a Dirac δ-function).

The final ingredient in our method is the assertion that the number of susceptibles is roughly the same at all epidemic peaks. To motivate the assertion we present the argument first for a simplified SIR model; this is followed by a more detailed argument using the assumptions of

this paper. For the conventional SIR model with mass-action contact rate, this follows immediately from the infectious class dynamics

$$dI/dt = \beta(t)SI - I/d_I,$$

where d_I is the mean duration of the infective stage. At a peak of $I(t)$ we have $dI/dt = 0$ and therefore $S(I) = (d_I\beta(t))^{-1}$. Because $\beta(t)$ changes slowly during the times of year when peaks occur and is roughly the same each year (as noted above), the value of $S(t)$ at epidemic peaks is roughly the same each year.

In this paper we make the more realistic assumption that latent and infectious stages have constant duration, but a similar argument applies. At an epidemic peak the number of infectives is large, so (as mentioned in Hamer, 1906; Fine and Clarkson, 1982a; Ellner and Turchin, 1995) it is appropriate to use the mass-action contact rate $\beta(t)SI$. Around the times of epidemic peaks we therefore have

$$\begin{aligned} R_E &= \beta(t)S(t)I(t) \\ R_I &= R_E(t - d_L), \end{aligned}$$

where let R_E and R_L denote respectively the rates of flow of individuals into the exposed (latent) class, and into the infective class. Because the latent and infective stages have fixed durations (by assumption), we then have

$$\begin{aligned} I(t) &= \int_{t-d_I}^{t} R_I(u)du = \int_{t-d_I}^{t} R_E(u - d_L)du \\ &= \int_{t-d_I-d_L}^{t-d_L} \beta(u)S(u)I(u)du = d_I\beta(t^*)S(t^*)I(t^*), \end{aligned} \tag{5}$$

where t^* is some time between $t - d_I - d_L$ and $t - d_L$ (i.e. t^* is within the time interval when the individuals who are infective at the peak where first infected). As noted above, near peak times of $I(t)$, $\beta(t)$ does not vary much from some constant β^*, and moreover $I(t)$ is not rapidly changing because $dI/dt = 0$ at the peak. We can therefore conclude from (5) that

$$S(t^*) \approx (d_I\beta^*)^{-1}, \tag{6}$$

with inexactness coming from the fact that $\beta(t)$ is not exactly constant, and $I(t)$ and $I(t^*)$ are not exactly equal.

Equations (2), (4) and (6) are the basis for our reconstruction method. We would like to re-emphasize that these equations are not general properties of SEIR or similar model. Their derivations make use of the model but also depend on the epidemiology of measles and on empirical properties of the measles case report data that we are analyzing (pre-vaccination, large cities in developed countries). Thus, we would not necessary expect them to apply to all other diseases that can be modeled in the SEIR framework. Our delay-kernel representation of recruitment (equations 3 and 4) is an attempt to capture an essential feature of age-structure for our purposes, without having to construct an explicitly age-structured model with many more parameters. Underpinning this approach is an assumption that for school-age or younger children, age mainly affects their chances of having established a permanent contact, either directly or via older siblings, with the school population that is the primary source of infection. Once an individual has become fully susceptible in this sense, it is no longer necessary to keep track of their age.

3. RECONSTRUCTION METHOD

We introduce a new variable $\hat{S}_t$ by an equation structurally similar to equation (1):

$$\begin{aligned} \hat{S}_{t+\tau} &= \hat{S}_t - C_{t,\tau} + \bar{C}_{t,\tau}, \\ \hat{S}_0 &= 0, \end{aligned} \tag{7}$$

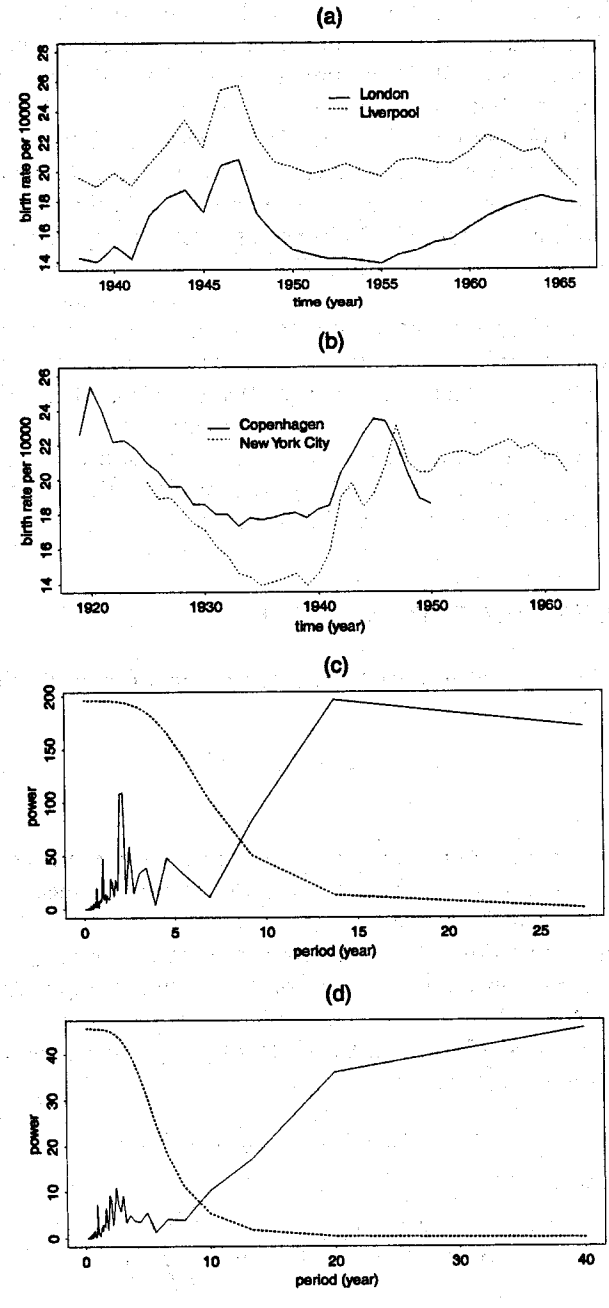
where $\bar{C}_{t,\tau}$ is the average number of reported cases over the entire data set. Given the case report data $C_{t,\tau}$, $\hat{S}_t$ values are obtained simply by iterating equation (7) forward from $t = 0$.

Averaging equation (1) over a period of time much longer than the timestep τ, and considering the approximate equality $E_{t,\tau} \approx I_{t,\tau}$ we get that $\bar{I}_{t,\tau} = \bar{Q}_{t,\tau}$, where the bar denotes the average value. Combining equations (1) and (7) with (2) and denoting $\tilde{Q}_{t,\tau} = Q_{t,\tau} - \bar{Q}_{t,\tau}$ after some rearrangement yields:

$$S_{t+\tau} - S_t = (\hat{S}_{t+\tau} - \hat{S}_t)\alpha + \tilde{Q}_{t,\tau}, \tag{8}$$

and therefore

$$\hat{S}_t = \left(S_t + \sum_{n=0}^{n=t/\tau} \tilde{Q}_{n,\tau} + S_0 \right) \Big/ \alpha. \tag{9}$$


(a)
London
Liverpool
birth rate per 10000
14
16
18
20
22
24
26
28
1940
1945
1950
1955
1960
1965
time (year)
(b)
Copenhagen
New York City
birth rate per 10000
14
16
18
20
22
24
26
1920
1930
1940
1950
1960
time (year)
(c)
power
0
50
100
150
200
0
5
10
15
20
25
period (year)
(d)
power
0
10
20
30
40
0
10
20
30
40
period (year)

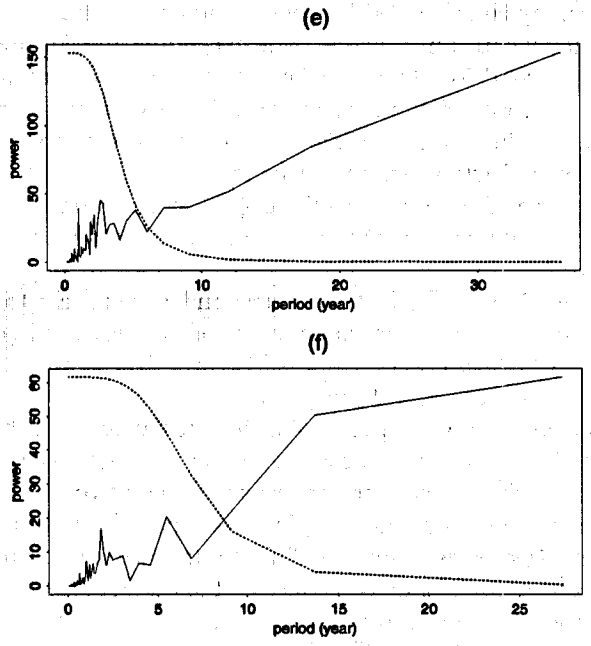


FIGURE 3 (a)-(b) Birth rate data for London, Copenhagen, Liverpool, and New York City. (c)-(f) Spectral densities of $\hat{S}_t$ (solid lines) and the shape of the spline high-pass filters (broken lines) for London, Copenhagen, Liverpool, and New York City with cutoff frequency corresponding to a period of 7 years. London and Liverpool data from Registrar General for England and Wales, Monthly Reports; data for New York City extracted from U.S. Census data by Ben Bolker (*personal communication*); Copenhagen data from L.F. Olsen (*personal communication*).

Figure (1(d-f)) shows $\hat{S}_t$ obtained from data using equation (7) with $S_0 \equiv 0$. In each case it is evident that $\hat{S}_t$ contains a low frequency trend (period > 15 years) and high frequency components (period < 10 years). However, since the value of S at some time t^* near an epidemic peak is the same for every epidemic as discussed above, and large epidemics occur every second or third year, trends with period longer than 2–3 years cannot be present in the actual dynamics of susceptibles. Hence it follows from equation (9) that the low-frequency component in $\hat{S}_t$ is due to $\sum_{n=0}^{n=t/\tau} \hat{Q}_{n,\tau}$, the recruitment into the susceptible class.

For all of the cities studied, the significant changes in birth rate and immigration occur over periods of 8–15 years, and there are no large shorter-period fluctuations (e.g. Figure 3(a-b)). This observation was

already used by Hedrich (1933), where he approximated the birth rate by an annually constant function. The exceptions are European cities during the Second World War when migration from urban to rural areas and back was fast and significant. The distributed transition of newborns into the susceptible class would be expected to smooth out high-frequency fluctuations in the birth rate, and integration of the recruitment (i.e. the summing of the $\tilde{Q}_{n,\tau}$ in equation (9)) provides additional smoothing. Thus the high-frequency component in $\tilde{S}_t$ must come from S_t.

Therefore, S_t and $\sum_{n=0}^{n=t/\tau} \tilde{Q}_{n,\tau}$ correspond respectively to the high frequency and low frequency components of $\hat{S}_t$. They can be estimated separately by high-pass and low-pass filtering of $\hat{S}_t$, up to the scale and location parameters α and S_0 in (9).

As expected, the power spectra of $\hat{S}_t$ for the studied cities have a gap between high and low frequencies, at a period of about 8 years (Figure 3(c-f)). For each city we therefore separate these two components in the frequency domain by constructing a low-pass filter with cutoff (or half-power) frequency corresponding to the minimal value of the power spectrum in the gap. Since neither the real reporting rate α, nor any real susceptible or recruitment numbers are fully known for these populations, both $\hat{S}_t$ and $\sum_{n=0}^{n=t/\tau} \tilde{Q}_{n,\tau}$ are reconstructed only up to additive and multiplicative constants. $\tilde{Q}_{t,\tau}$ is then obtained by differencing the low-frequency component. In Figure 4 we illustrate the main steps of the reconstruction method. Details of the filtering procedure are given in the Appendix.

The accuracy of the reconstruction method was evaluated using both the real data and simulated data. We used an SEIR system of ordinary differential equations as in Engbert and Drepper (1993) to generate simulated susceptible and case data. We also used finite-population simulations of the model with periodic variations in the birth rate. The finite-population model treats the right-hand sides of the differential equation as specifying transition rates between the compartments, and was implemented by using a two-hour step and Poisson-distributed transitions with mean given by (transition rate) × (time step). We modeled the measurement (reporting) errors as in Ellner *et al.* (1995); the data (real or simulated) were perturbed by lognormal errors, with coefficient of variation based on the assumption that cases are reported or not in pairs. This generates a higher error variance than if cases are reported independently, and gives a better fit to the high-frequency tail of the power spectrum of the data Ellner *et al.* (1995). To be conservative we used the error variance for a 50%

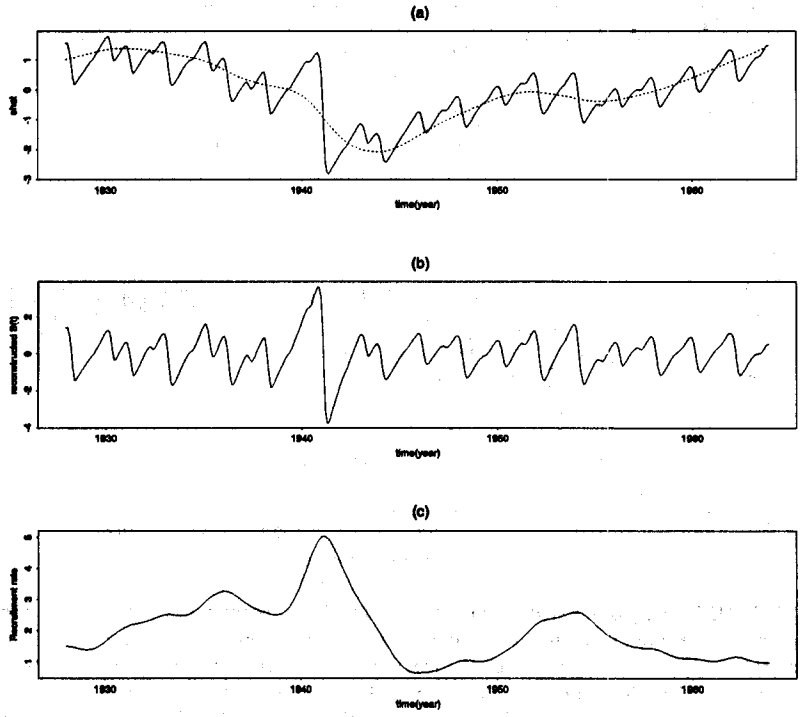


FIGURE 4 The steps of the reconstruction method. (a) The solid line shows the "pseudo-susceptibles" $\hat{S}_t$ from the New York City data, and the broken line is the low frequency component obtained by the smoothing spline method. (b) The reconstructed scaled susceptibles are given by the residuals from the spline fit, $\hat{S}_t$ – (low frequency component). (c) The reconstructed recruitment rate, obtained by differencing the low frequency component.

reporting rate (which maximizes the error variance), and we also adjusted the noise variance to values beyond the maximum possible under our reporting model.

The results are, first of all, that susceptible reconstruction from SEIR differential equations case data is almost perfectly accurate ($r^2 = 0.98$ between true and reconstructed monthly susceptibles; Figure 5). Second, the reconstruction is not affected much by measurement errors: for both the real and simulated data, there is a high correlation between values reconstructed from case data with and without measurement errors, even if the error variance is substantial. The finite-population simulations incorporated two features absent

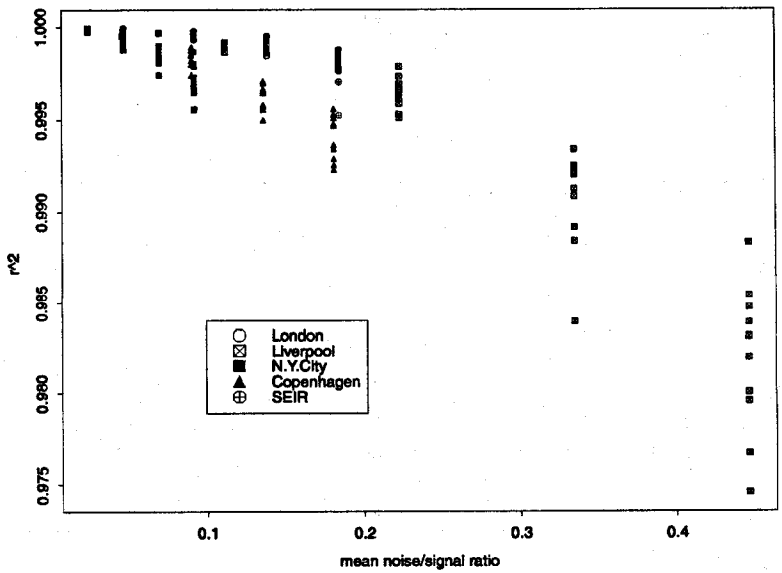


FIGURE 5 Effect of noisy data on the accuracy of the reconstruction. For SEIR, the plotted values are the squared correlations are between reconstructed susceptibles from noisy data (exact data plus simulated measurement errors), and the exact susceptible time series from the simulation. For the real data, the correlations are between reconstructed susceptibles from data with and without the addition of simulated measurement errors (on top of those already present in the data). The measurement error level is measured by the ratio of noise variance to data variance. The maximum possible Poisson variance, corresponding to 50% reporting rate, was chosen as the initial level. We increased the variance by factors of 2, 3, and 4, and for each data set and variance level we generated ten noisy data sets.

from the differential equation situations: demographic stochasticity, and periodic variation in birth rates. The reconstruction accuracy was therefore slightly lower, but was still quite good in most simulations (Figure 6).

4. APPLICATIONS

4.1. Forecasting

Given the reconstructed susceptible dynamics, it is natural to base epidemic forecasts on models which make use of this information. In addition, if such models make more accurate forecasts than models

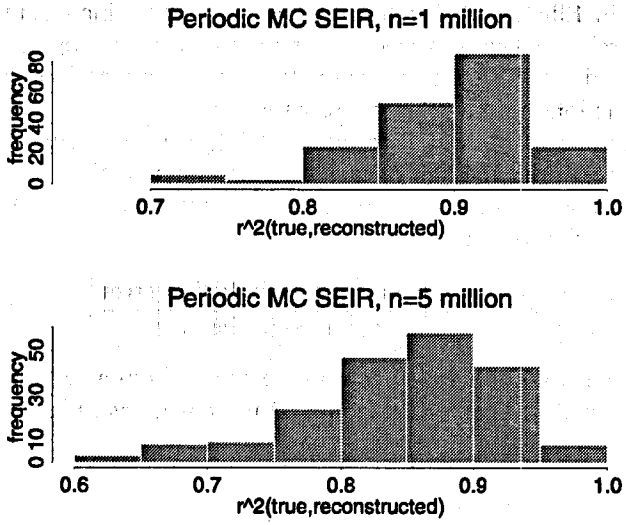


FIGURE 6 Correlation coefficient (r^2) between true and reconstructed susceptibles from finite-population Monte Carlo simulation of SEIR model. Susceptible dynamics were reconstructed from 200 sets of 420 months (35 years) of case report totals with simulated reporting errors as described in the text (50% reporting rate, cases reported or not in clusters of size 2). Top panel shows results for a city of size 1 million, birth rate varying periodically by $\pm 15\%$ with period 15 years; bottom panel shows results for city size of 5 million, birth rate varying periodically by $\pm 25\%$ with period of 15 years.

that do not use the reconstructed susceptibles, that is evidence that the reconstructed "data" are accurate enough to be useful.

Our model incorporates the qualitative structure of transmission in SEIR/RAS models, in which the number of new infectives depends only on the current numbers of infectives and susceptibles, and the seasonally varying contact rate between them. However, we do not assume *a priori* the conventional mass-action equation. Thus, we consider a model of the form

$$C_{t+\tau} = F(C_t, S_t, \sin(t), \cos(t)) + \varepsilon_t, \tag{10}$$

where the periodic components represent seasonal forcing (with time measured in years), and ε_t is a random exogenous noise. The function F is fitted by a flexible or fully non-parametric nonlinear regression model. We compared the accuracy of this model with the model

$$C_{t+\tau} = F(C_t, C_{t-l}, \ldots, C_{t-ml}, \sin(t), \cos(t)) + \varepsilon_t, \tag{11}$$

as used in Ellner and Turchin (1995). This model incorporates seasonal forcing but uses lagged case report data as surrogates for the unobserved state variables, motivated by the method of attractor reconstruction in time-delay coordinates.

The comparison, as in Tidd *et al.* (1993), was conducted by fitting each model to the first half of each data series, and making predictions for the second half. The accuracy of the predictions is based on a *pseudo* r^2 criterion, given by

$$r^2 = 1 - \frac{\text{mean square prediction error}}{\text{variance of the data}}. \tag{12}$$

Because we are not actually computing the r^2 from a regression, it is possible for the r^2 computed from (12) to be negative. Figure 7 shows

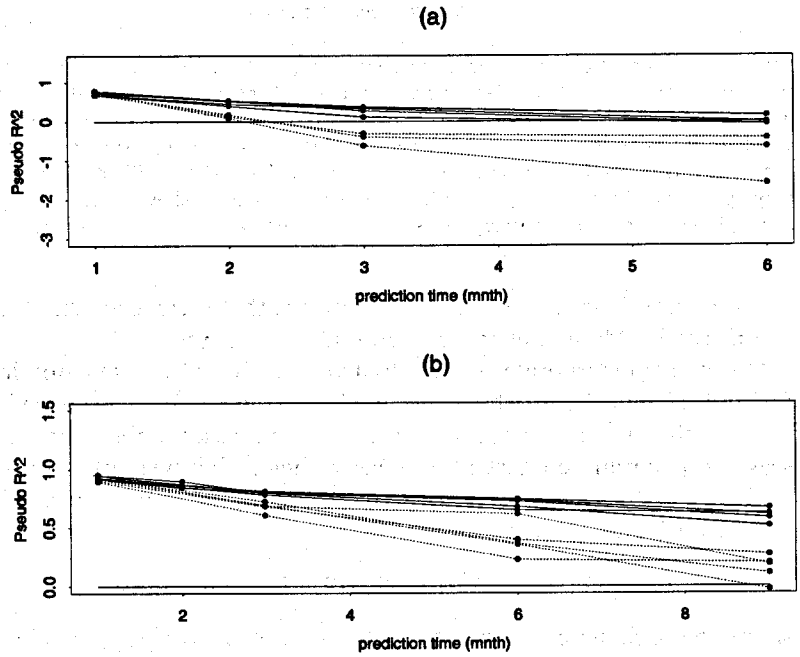


FIGURE 7 Prediction of the Liverpool (a) and London (b) data several months ahead. The solid lines are for predictions using model (10) that incorporates the reconstructed susceptibles, the dashed lines using model (11). Neural network models of different complexity (number of hidden units varied from one to four) were fitted to the first half of the data, and were then used to predict the second half.

that model (10), which incorporates the reconstructed susceptibles, produces more accurate forecasts for all of our data sets.

Predictions can also be made by assuming a particular mechanistic model (i.e. specifying an *a priori* form for F in equation (10)), and forecasting by solving the model either exactly or approximately. This alternative forecasting approach is considered in Grenfell *et al.* (1995).

4.2. Family Structure and Recruitment into the Fully Susceptible Class

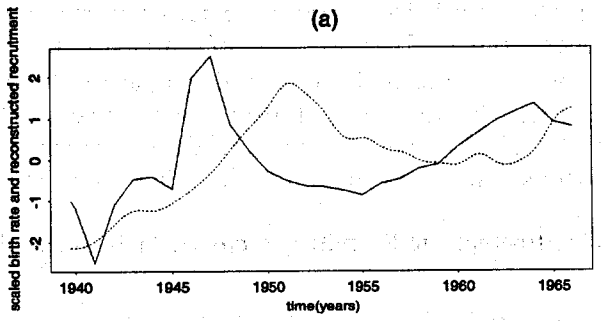
Comparing the reconstructed recruitment rate to the raw birth rate (Figure 8(a)), it appears that the recruitment rate can be interpreted as the result of time-delay and smoothing applied to the birth rate. This observation agrees with the mechanistic assumption that a child has a much higher chance of getting the disease after attaining school age. Thus, pre-school children who have a smaller chance of getting the disease form a temporarily "protected class" that acts as an accumulator for future recruitment to the susceptible class. However, not all of the pre-school children are "protected" in this sense; in particular, children with school-age elder siblings would be exposed to secondary infection by close contact with their siblings (Anderson and May, 1991; Grenfell *et al.*, 1995). Thus we would expect to see a distribution of residence times in the protected class, that would be affected by family size and the age differences between siblings.

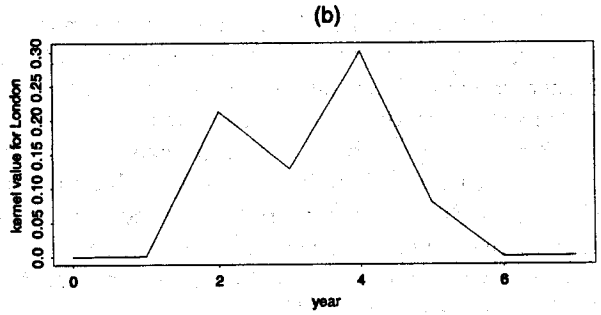
In Grenfell *et al.* (1995) the recruitment rate into the susceptible class Q_t in year t was modeled by using data on previous births and the proportions of families with different numbers of children:

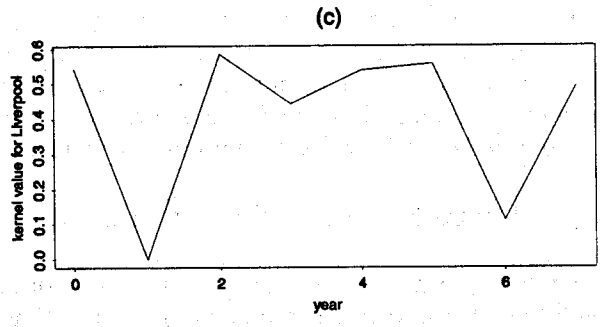
$$Q_t = g_{0,t-5}B_{t-5} + g_{1,t-3}B_{t-3} + g_{2,t-1}B_{t-1}, \tag{12}$$

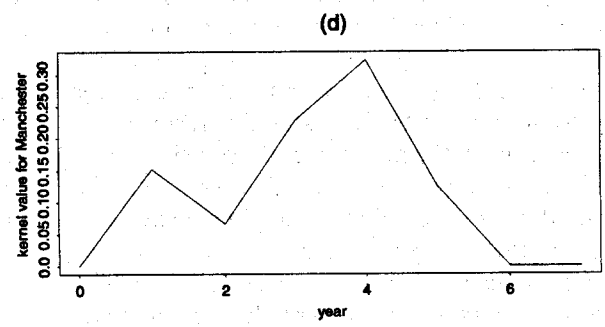
where B_t is the birth rate in year t, and $g_{j,t}$ is the proportion of the families with j children at year t. The values of $g_{j,t}$ were estimated from the annual reports of the Registrar General for England and Wales. Here, we have estimated Q_t from the epidemic and birth rate data. We can then estimate the average contribution of each age to the recruitment, by using birth rate data and the reconstructed recruitment. These estimates have the advantage of not using data on family sizes, and therefore provide a way to test the hypothesis that differences in family size affect the distribution of time-delays between birth and recruitment to the susceptible class.

Because significant variations in recruitment rate occur on a time scale of several years, we can assume that recruitment rate is constant


(a)
scaled birth rate and reconstructed recruitment
-2
-1
0
1
2
1940
1945
1950
1955
1960
1965
time(years)


(b)
kernel value for London
0.0
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
year


(c)
kernel value for Liverpool
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
year


(d)
kernel value for Manchester
0.0
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
year

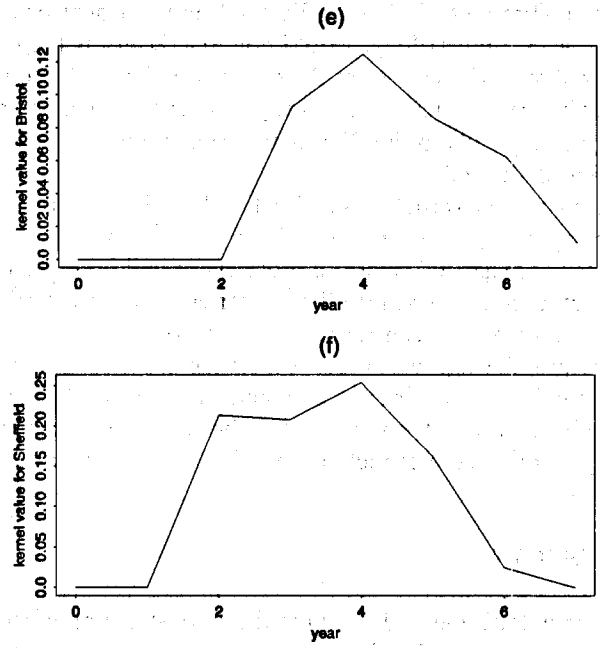


FIGURE 8 (a) Real birth rate (solid line) and reconstructed recruitment (broken line) for London. Recruitment and birth rates are scaled to zero mean and unit variance. (b)-(f) Reconstructed recruitment delay kernels for measles in London, Liverpool, Manchester, Sheffield and Bristol.

during the shorter time interval τ corresponding to the timestep of our data. Thus, equation (4) can be rewritten as:

$$Q_{t,\tau} = \tau \int_0^\infty G(t-u)B(u)du. \tag{13}$$

We estimated the shape of a kernel $G(s)$, by fitting a discrete analog of equation (13),

$$\bar{Q}_t = \sum_{j=1}^{7} G_j \bar{B}_{t-j}, \tag{14}$$

where the overbars indicate annual average values of the recruitment and birth rates. Values of G_j were estimated by least-squares subject to the constraint $G_j \geq 0$, using the POWELL routine for function

minimization (Press *et al.*, 1992). We used only postwar case data in order that the estimates of $Q_{t,\tau}$ would not be affected by the migration during World War II. Because of the delay kernel we need to use birth data preceding recruitment by several years. Even though part of our birth data corresponds to the War years we assume that this data was not significantly biased.

The reconstructed kernel shapes for five British cities are presented in Figure 8(b-f). The shapes of these kernels suggest that after attaining school age (6–7 years) almost every child has become susceptible, but pre-school susceptibility varies in different cities. For example, in Liverpool a considerable percent of children become susceptible before the age of two, while in London this percentage is negligibly small. This difference presumably reflects a larger average family size in Liverpool than in London, because having more siblings in a family increases the chance of an infant getting infected.

5. DISCUSSION

We have presented a method allowing the reconstruction of susceptible and recruitment dynamics from case report data, and applied the method to pre-vaccination measles data. We have defined the conditions under which this method works, and showed that the method is robust to reporting errors.

Two main applications are suggested. For epidemic prediction, we showed that knowledge of the susceptible dynamics makes the models more accurate and more interpretable, in that the fitted model corresponds to the rate of disease-transmitting contacts between susceptibles and infectives. Second, the reconstruction makes it possible to estimate the dynamics of recruitment into the susceptible class. These estimates indicate that recruitment rate is not simply the birth rate with a discrete time delay, but that the age at recruitment is variable. This suggests that SEIR-type models with variable birth rate may need to be modified by introducing an additional "protected class". The resulting PSEIR model considers that most newborn children are initially protected from infection, and until reaching school age have an age-dependent probability of recruitment into the susceptible class.

We have not discussed estimation of the scaling parameters α and S_0 in the reconstructed susceptibles. Estimations of these parameters can be found in (Fine and Clarkson, 1982a; Grenfell and Anderson, 1985; Mollison and Ud Din, 1993). Knowledge of these two parameters and,

thus, the exact susceptible values yields the possibility of a closer study of the functional form of the force of infection, especially at low infective population values. For a qualitative study or forecasting, however, knowledge of the scaled dynamics may be nearly as useful as knowledge of the exact values.

The methods presented in this paper can be applied directly to some other childhood diseases such as chickenpox and mumps in developed countries, as long as they fit assumptions (a-f) (Weller, 1989; Katz, 1997), with only simple modifications to account for differences in stage durations. For example, Figure 9 compares the recruitment rate to the susceptible class estimated from measles case reports, and from chickenpox case reports in Copenhagen. On mechanistic grounds we can safely assume that the two curves are essentially identical. The fact that the estimates coincide thus supports the accuracy of the method for both diseases. However more fundamental modifications of the method would be needed for diseases with only temporary immunity, or with non-negligible disease-induced mortality. The necessary changes are easily accomplished in principle, but the accuracy of the method in the face of these complexities is an open question.

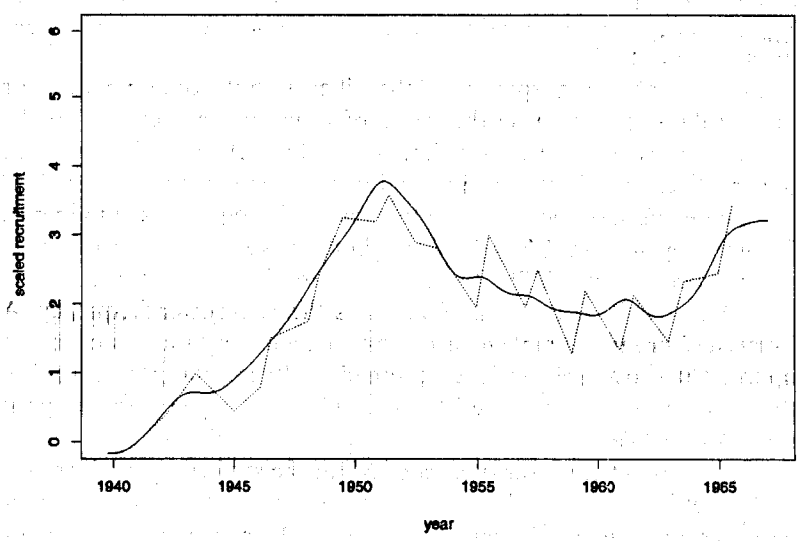


FIGURE 9 Comparison of scaled reconstructed recruitment rates for measles (solid line) and chickenpox (broken line) in Copenhagen.

APPENDIX: FILTERING HIGH AND LOW FREQUENCY COMPONENTS FROM $\hat{S}_t$

For long data sets it is convenient to use smoothing splines as the low-pass filter. The shape of the filter for smoothing splines is a function of the smoothing parameter λ. For long enough time series (e.g. roughly 500 or more weekly data points) the filter is approximately fourth order, i.e. its weight function $\Theta(v, \lambda)$ has the form

$$\Theta(v, \lambda) = \frac{1}{\alpha_1(\lambda) + v^4 \alpha_2(\lambda)}, \tag{A1}$$

where v is frequency, and α_1, α_2 are constants, depending on the smoothing parameter λ (Nychka, 1991). The power spectrum of $\hat{S}_t$ and shape of the spline filter are illustrated in Figure 3.

By adjusting λ, it is possible to choose the half-power (cutoff) frequency of a smoothing spline filter so that it corresponds to the middle of the gap in the power spectrum of $\hat{S}_t$. After λ is chosen, we fit the smoothing spline to the $\hat{S}_t$ series and obtain

$$\hat{S}_t = \text{smooth spline fit} + \text{residuals}.$$

The smooth spline is the low-frequency component (proportional to $\sum_{n=0}^{n=t/\tau} \tilde{Q}_{t,\tau}$), and the residuals are the high-frequency component (proportional to S_t).

Use of smoothing splines as the filter is not recommended for smaller data sets, because the value of v will be smaller and so the filtering is less effective (Nychka, 1991). In such cases it would be preferable to filter in the frequency domain, using an equation such as (A1) to decompose the FFT into low and high frequency components, followed by inverse FFT. FFT method results here were obtained using *fftpack* from NETLIB and a low-pass filter similar to equation (A1), $\Theta(\nu) = 1/(1 + (\nu/\nu_0)^4)$, where ν_0 is the half-power frequency. A Fortran77 program implementing this method for data sets of length up to 5000 is available on request (email to ellner@stat.ncsu.edu). For the data considered here, the FFT and spline filters give essentially the same results (Table 1).

Even though the power spectra of $\hat{S}_t$ for the studied cities all consist of distinct high and low-frequency components, it might not always be clear where to place the cutoff frequency. In addition, one can use different filtering techniques (splines vs. FFT, and abrupt vs. gradual frequency cutoff functions in FFT methods). A comparison of spline

TABLE 1

Correlation between susceptible data reconstructed by spline and FFT methods, and by FFT methods with different cutoff frequencies chosen in a spectral gap. "Minimum correlation in spectral gap" is the minimum pair-wise correlation coefficient between susceptible time series reconstructed by the FFT method with different cutoff frequencies within the spectral gap, over all frequencies corresponding to periods of an integer number of years

City	Correlation between spline and FTT method	Spectral gap (period in years)	Minimum correlation in spectral gap
London	0.97	5–10	0.96
Liverpool	0.96	5–8	0.95
Copenhagen	0.97	5–10	0.95
New York City	0.96	5–10	0.94

and FFT methods applied to our data sets is presented in Table 1. As long as the cutoff point is chosen to be within the spectral gap between high and low frequency components, the choice of filtering techniques and cutoff frequency does not appear to significantly influence the results.

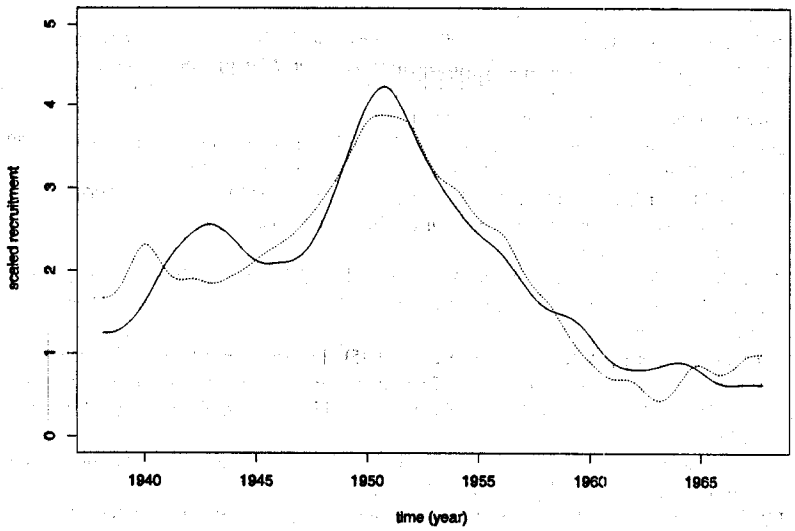


FIGURE 10 Scaled recruitment rate for London reconstructed by the FFT method (solid line) and reconstructed roughly from estimates at the annual peaks of susceptibles (broken line).

Recruitment also can be reconstructed roughly in a simpler way without employing filtering. Since S_t differ from $\hat{S}_t$ by a low frequency trend, local extrema of S_t and of $\hat{S}_t$ occur at nearly the same times. Thus when $\hat{S}_t$ reaches an extremum, equation (1) leads to $I_{t,\tau} = Q_{t,\tau}$. Under the assumption that $I_{t,\tau} = \alpha C_{t,\tau}$, the case report data therefore give scaled recruitment at the times when $\hat{S}_t$ has an extremum. Figure 10 demonstrates the match between recruitment rates reconstructed in this manner, and recruitment estimated by filtering.

REFERENCES

Anderson, R.M. and May, R.M. (1985). Age related changes in the rate of disease transmission: implications for the design of vaccination programmes. *Journal of Hygiene (Cambridge)*, **94**: 365–436.

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

Bartlett, M.S. (1957). Measles periodicity and community size. *Journal of Royal Statistical Society A*, **120**: 48–70.

Black, F.L. (1959). Measles antibodies in the population of New Haven, Connecticut. *Journal of Immunology*, **82**: 74–83.

Black, F.L. (1989). Measles, in *Viral Infections of Humans: Epidemiology and Control*, Evans, A.S. (ed.) New York: Plenum.

Bolker, B.M. and Grenfell, B.T. (1993). Chaos and biological complexity in measles dynamics. *Proceedings of the Royal Society. London. Biology*, **251**: 75–81.

Capasso, V. (1993). Mathematical structure of epidemic systems. *Lecture Notes in Biomathematics 97*. Berlin: Springer-Verlag.

Center for Disease Control, Measles and School Immunization Requirements (1978). United States. Morbidity and Mortality Weekly Reports 27–391.

Cliff, A.D., Ord, J.K., Haggett, P. and Versey, G.R. (1981). *Spatial Diffusion and Historical Geography of Epidemics in an Island Community*. Cambridge: Cambridge Geographical Studies.

Dietz, K. and Schenzle, D. (1985). Mathematical models for infectious disease statistics, in *A Celebration of Statistics*, Atkinson, A.C., Fienberg, S.E. (eds.) New York: Springer-Verlag,

Ellner, S., Gallant, A.R. and Theiler, J. (1995). Detecting nonlinearity and chaos in epidemic data, in *Epidemic Models: Their Structure and Relation to Data*, Mollison D. (ed.) Proceedings of NATO ARW on Epidemic Models, Cambridge: Cambridge University Press.

Ellner, S. and Turchin, P. (1995). Chaos in a noisy world: new methods and evidence from time-series analysis. *American Naturalist*, **145**: 343–375.

Ellner, S., Bailey, B., Bobashev, G.V., Gallant, A.R., Grenfell, B. and Nychka, D.W. (1998). "Noise and Nonlinearity in Epidemics: Combining Statistical and Mechanistic Modeling to Characterize and Forecast Population Dynamics." *American Naturalist*, **151**(5): 425–440.

Engbert, R. and Drepper, F.R. (1993). Qualitative analysis of unpredictability: a case study from childhood epidemics, in *Predictability and Nonlinear Modeling in Natural Sciences and Economics*. Wageningen.

Fine, P.E.M. and Clarkson, J.A. (1982a). Measles in England and Wales-I: An analysis of factors underlying seasonal patterns. *International Journal of Epidemiology*, **11**: 5–14.

Fine, P.E.M. and Clarkson, J.A. (1982b). Measles in England and Wales-II: The impact of the measles vaccination programme on the distribution of immunity in the population. *International Journal of Epidemiology*, **11**: 15–29.

Grenfell, B.T. and Anderson, R.M. (1985). The estimation of age-related rates of infection from case notification and serological data. *Journal of Hygiene (Cambridge)*, **95**: 419–36.

Grenfell, B.T., Kleczkowski, A., Ellner, S. and Bolker, B.M. (1995). Non-linear forecasting and chaos in ecology and epidemiology: Measles as case study, in *Nonlinear Time Series and Chaos, Vol. 2*. Proceedings of Royal Society Discussion Meeting, Tong, H. (ed.) Singapore: World Scientific.

Hamer, W.H. (1906). Epidemic disease in England – the evidence of variability and of persistency of type. *Lancet*, **1**: 733–79.

Hedrich, A.W., (1933). Monthly estimates of the child population 'susceptible' to measles, 1900–1931, Baltimore. *American Journal of Hygiene*, **17**: 613–36.

Katz, S.L. (1997). Mumps, in *Viral Infections of Humans: Epidemiology and Control*, Evans, A.S. and Kaslow, R.A. (eds.) New York: Plenum.

Keeling, M. and Grenfell, B.T. (1997). Disease extinction and community size: modeling the persistence of measles. *Science*, **275**: 65–67.

London, W.P. and Yorke, J.A. (1973). Recurrent outbreaks of measles, chickenpox and mumps. I. Seasonal variation in contact rates. *American Journal of Epidemiology*, **98**: 453–68.

Mollison, D. and Ud Din, S. (1993). Models for seasonal variability of measles. *Mathematical Biosciences*, **117**: 155–177.

Nychka, D.W. (1991). Choosing a range for the amount of smoothing in nonparametric regression. *Journal of American Statistical Association*, **86**(415): 653–664.

Press, W.H., Teukolsky, S.A., Vettrling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in Fortran*. Cambridge: Cambridge University Press.

Olsen, L.F., Truty, G.L. and Schaffer, W.M. (1988). Oscillations and chaos in epidemics: a nonlinear dynamic study of six childhood diseases in Copenhagen, Denmark. *Theoretical Population Biology*, **33**: 344–370.

Roden, A.T. and Heath, W.C. (1977). Effects of vaccination against measles on the incidence of the disease and on the immunity of the child population in England and Wales. *Health Trends*, **9**: 69–72.

Schaffer, W.M., Olsen, L.M., Truty, G.L. and Fulmer, S.L. (1990). The case for chaos in childhood epidemics, in *The Ubiquity of Chaos*, Krasner, S. (ed.) Washington, D.C.: AAAS.

Spector, P. (1994). *An Introduction to S and S-Plus*. Belmont: Duxbury Press.

Sugihara, G. and May, R.M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, **344**: 734–741.

Tidd, C.W., Olsen, L.F. and Schaffer, W.M. (1993). The case or chaos in childhood epidemics. II. Predicting historical epidemics from mathematical models, *Proceedings of the Royal Society, London. B*, **254**: 257–273.

Weller, T.H. (1989). Varicella-Herpes Zoster Virus, in *Viral Infections of Humans: Epidemiology and Control*, Evans, A.S. (ed.) New York: Plenum.