Measles in small populations : predictability in highly stochastic systems

Q. Caudron[1,*], A. S. Mahmud[2], C. J. E. Metcalf[1], M. Gottfreðsson[3], B. T. Grenfell[1]
1 Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, 08544, USA
2 Office of Population Research, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ, 08544, USA
3 Landspítali, Hringbraut 101, Reykjavík, Iceland
∗ E-mail: qcaudron@princeton.edu

Abstract

A standard assumption in the modelling of epidemic dynamics is that of a certain level of homogeneity. The well-known and oft-used SIR model, arguably the most important compartmental model in theoretical epidemiology, assumes that the disease being modelled is strongly immunising, directly transmitted, and has a well-defined period of infection, in addition to these homogeneity assumptions. Childhood infections, such as measles, are prime examples of diseases that fit the SIR-like mechanism, and these infections have been well studied for many systems. Populations are typically assumed to be large and well-mixed, and the data assumed to contain a high signal-to-noise ratio. Differential equations can then be used to model the disease's dynamics, generally with positive results, both in terms of explaining the underlying mechanisms for infection, and for the prediction of future time series data on disease incidence. Here, we consider a setting where populations are small and heterogeneous, and where the dynamics of infection are driven by extinction-recolonisation events. Using a TSIR model, we fit prevaccination measles incidence and demographic data in Bornholm, the Faroe Islands, and four districts of Iceland, between 1901 and 1965. The datasets for each of these countries suffers from different levels of data heterogeneity and sparsity. We explore the potential for prediction of this model : given historical incidence data and up-to-date demographic information, and knowing that a new epidemic has just begun, can we predict how large it will be ? We show that, despite a lack of significant seasonality in the incidence of measles cases, and potentially severe heterogeneity at the population level, we are able to estimate the size of upcoming epidemics, conditioned on the first time step, to within reasonable confidence. One more sentence would be nice.

Introduction

Rewrite, expand. General train of thought – SIR is our go-to model, but it makes some assumptions. Many of those are taken care of in measles, but then our analysis is typically restricted to rich datasets in large, well-mixed populations. CITE anything done in small, heterogeneous populations. Insist on stochasticity in our data – extinction-recolonisation. Good ideas in Metcalf, Bjornstad, Grenfell and Andreasen, Proc B, 2009

Measles is a highly contagious and strongly immunising infection of the respiratory system. Due to its extreme transmissibility, its epidemiology is conditional on the birth of susceptible individuals. As such, the temporal dynamics of measles are typically strongly oscillatory, driven seasonally by the increased contact rate amongst young children during school periods, assuming the population is large enough to sustain the disease [1]. These dynamics have been well studied (Grenfell papers, others), and many modelling efforts have successfully explained the biennial cycle exhibited in prevaccination records of measles incidence in Europe and elsewhere (papers ?).

In small populations, where the number of individuals is much smaller than the critical community size required to support an endemic infection, however, the dynamics of measles cases are vastly different.

Susceptible individuals accumulate when measles is absent; then, driven by stochastic recolonisation, an epidemic may sweep through a large fraction of the susceptible population very quickly, only to go extinct abruptly as susceptibles become depleted. This results in very sharp, spiky epidemics, the timing of which may be impossible to predict, but the size and duration of which may be a function of historical data.

In this paper, we address the question of predictability of measles epidemics in small populations, based on records of past incidence and on demographic data. We present data on the demographics and disease incidence in prevaccination-era Bornholm, the Faroe Islands, and four districts in Iceland.

reconstruct the dynamics of susceptible individuals and infer the rate of reporting of cases using the TSIR model [2] in prevaccination

sizes are, by training a model on records of past epidemics and demographic data. Using the TSIR model [2], we explore the dynamics of measles in prevacinnation Bornh

Any seasonality ?

Methods

Data

Measles incidence data were obtained for Iceland, the Faroe Islands, and Bornholm (CONFIRM?), from 1900 to 1965, from [3]. For Iceland, this dataset collects monthly figures for measles cases reported in 47 medical districts, originally sourced from *Heilbrigðisskýrslur* (Public Health in Iceland). Major revisions to the boundaries of medical districts took place twice during the study period : in 1907 and 1932. For the Faroe Islands and Bornholm, the data were originally published WHERE?.

Demographic data for Iceland were obtained from two sources. Annual data on population and number of live births for the entire country were taken from [**?**]. Decadal population data from 1901 to 1965, for 262 municipalities, were obtained from online publication in [4]. Municipality boundaries changed from three to five times during the study period. In addition, many municipalities had missing data. Medical districts and municipalities were matched based on names, and confirmed using latitude and longitudes. Using this procedure, we were able to confidently match four medical districts with municipalities. Several matched districts were discarded either due to missing population data, or lack of confidence in the matching of the geographical boundaries. It is also worth noting that matched medical district – municipality pairs may not encompass the exact same area, but one may be a (potentially partial) subset of the other. With the data available, we were able to match four district – municipality pairs : Akureyri, Reykjavík, Hafnarfjörður, and Vestmannaeyjar.

Data on the demographics of the Faroe Islands were taken from the Statistical Yearbooks of Denmark published by [5], and from Statistics Faroe Islands [6]. Annual data on population and births from 1901-1965 were obtained in aggregated form for all of the islands in the Faroe archipelago.

Demographic data for Bornholm was collected from several publications from [5]. Annual population data for Bornholm were obtained from [5], which contains detailed statistical information collected by Statistics Denmark. Pre-1930 annual birth data were obtained from the *Ægteskaber, Fødte og Døde* (Marriages, Births and Death) publications of Statistics Denmark. Post-1930 annual birth data were obtained from *Befolkningsudvikling og sundhedsforhold 1901-60* (Population, Development and Health 1901 - 1960), also from Danmarks Statistik.

All locations experienced a slight decline in birth rates just prior to 1940, followed by a sharp increase in births after 1940. In Bornholm, births declined steadily from around 1945 onwards. In Iceland and the Faroe Islands, aggregate births increased steadily from 1940s onwards, with a slight decline in Iceland around 1960.

Figure 1 shows the reported incidence for Bornholm, the Faroe Islands, and four districts of Iceland.

The TSIR model

For systems with small, heterogeneous populations, epidemic dynamics are driven by stochasticity in the timing of disease recolonisations and in population mixing. As such, dynamical models such as the well established SIR model are unable to adequately represent the underlying biological processes which dictate the evolution of disease incidence over time. The time-series SIR model [2] is a discrete-time, stochastic model of disease progression written in terms of a set of difference equations. Assuming that the infection is fully immunising and that the infectious period is well-defined, then the evolution of the number of infected cases, I, can be written,

$$\mathbb{E}\left[I_{t+1}\right] = r_t \, S_t \, I_t^{\alpha}, \tag{1}$$

where S_t is the number of susceptible individuals at time t, seasonal contact rates are represented by the periodic parameter $r_t = r_{t+P}$ for P time steps per year, $0 < \alpha < 1$ is an inhomogeneity parameter, and where $\mathbb{E}\left[\,\cdot\,\right]$ denotes the expectation operator. The time step is set as the generation time of the infection. Then, the number of susceptible individuals is defined by,

$$S_{t+1} = S_t + B_{t-d} - I_t + u_t, \quad \mathbb{E}\left[u_t\right] = 0. \tag{2}$$

Here, u_t is zero-mean additive noise. B_{t-d} is the number of births born d time step prior to t, the delay d assumed due to maternal immunity, and set at four months [7].

The observed number of cases, C_t, is assumed to be underreported by a reciprocal reporting rate $\rho_t \geq 1$, such that the true number of infected cases at time t is given by $I_t = \rho_t \, C_t$. If the number of susceptible individuals S_t fluctuates around a mean $\bar{S}$ such that $S_t = \bar{S} + Z_t$, then, from equation (2), the dynamics of the susceptible individuals are given by

$$Z_{t+1} = B_{t-d} + Z_t - \rho_t \, C_t + u_t. \tag{3}$$

A major assumption made by the TSIR model is that all individuals will eventually become infected. As such, the incidence of infected cases should track births. Successive iteration of equation (3) yields,

$$Z_{t+1} = \sum_{i=1}^{t} B_{i-d} - \sum_{i=1}^{t} \rho_i \, C_i + \sum_{i=1}^{t} u_i + Z_0. \tag{4}$$

If $\rho_t = \rho$ is a constant, and u_t is small, then equation (4) reflects a linear relationship between the cumulative births and the cumulative incidence. However, as Z_{t+1} depends on Z_t, it can be shown that the reporting rate need not be a constant, and that it could be estimated using locally linear regression methods. Then, Z_t can be found as the residuals of this regression.

Fitting

The time step in the difference equations (1) and (2) is fixed at the generation time of the infection. For measles, the period of time from infection to recovery is approximately two weeks [7]. Due to the very spiky nature of the reported incidence data (whose derivatives are non-smooth due to low sampling rates), interpolation must be done such that peaks in the data are not missed or reduced. As such, a linear interpolant with an integer multiple of the number of points per year was used. This yielded 24 time points per year, thus maintaining the maximum values of the peaks in the data, and fixing the generation time at just over fifteen days.

Population and live births, assumed to be smooth, were interpolated cubically to 24 time points per year. Despite large intervals between some of the reported demographics data, Finkenstädt and Grenfell [2] report that the regression for reconstructing susceptibles is robust to pronounced changes in birth rates.

The reporting rate ρ_t was estimated using Gaussian process regression, given the births and reported cases. Unlike splines or locally-weighted regression methods, Gaussian process regressions do not optimise smoothness of the fitted values, but instead yield the best unbiased estimates, assuming the error in the underlying process is normally distributed. Once found, ρ_t is the derivative of the Gaussian process prediction for the cumulative number of births, with respect to the cumulative number of cases, and Z_t are the residuals of the regression.

The mean number of susceptibles $\bar{S}$ was estimated marginally by profiling the likelihood of the logarithmic form of equation (1) :

$$\ln\left(\mathbb{E}\left[I_{t+1}\right]\right) = \ln(r_t) + \ln\left(\bar{S} + Z_t\right) + \alpha \ln\left(I_t\right), \tag{5}$$

after which the seasonal contact rates r_t were estimated conditionally on $\bar{S}$. The inhomogeneity parameter was fixed at $\alpha = 0.97$, as in [8], implying a small, nonlinear inhomogeneity, yet not significantly impacting transmission dynamics between large and small epidemics. Not sure this is relevant here – there are no epidemic troughs because we just go to zero. Perhaps best to try inferring this once more, now that sensitivity thresholds are better established.

Predictions

Using the TSIR model as defined by the system of equations (1, 2), predictions for epidemic dynamics were made by sampling the incidence I_{t+1} from a binomial distribution :

$$I_{t+1} \sim \text{Bin}\left(S_t,\, 1 - \mathrm{e}^{-\lambda}\right), \tag{6}$$

where the number of Bernoulli trials is given by the number of susceptible individuals, S_t. The probability of a successful infection is defined by the force of infection, $\lambda = r_t\, I_t^{\alpha}$, cumulated over one biweek period. This assumes that each individual spends an exponentially-distributed period of time in the infected class.

Due to the abundance of zeros in the incidence time-series, initial conditions cannot simply be taken as the point (I_0, S_0). Instead, each epidemic must be simulated independently, with initial conditions given by the data at the time that the epidemic begins. For each epidemic, we fix the number of infected cases and of susceptible individuals as per the data, and allow the simulation to continue until the next epidemic begins. Thus, we always simulate the same number of epidemics as given by the incidence data, where each epidemic is simulated conditioned on the data available at the beginning of that epidemic. Is this clear, or does it need rewording ?

In order to clearly establish when this time is, a sensitivity threshold must be set. Let $\tau \in \mathbb{Z}^+$ define the number of reported infected cases necessary for any particular biweek period to be considered part of an epidemic. A choice of $\tau = 1$ ensures that all available non-zero data is used. However, many potential epidemics go extinct before propagating through the population, especially in highly heterogeneous populations. As such, using $\tau = 1$ would cause a large number of strongly overestimated epidemics, which in turn would deplete the susceptible pool, and underestimate future epidemics. As such, we treat τ as a sensitivity parameter, and fit it by selecting the sensitivity threshold which yields the highest correlation between the mean predicted epidemic traces and the incidence data, as defined by Pearson's R^2. Then, the first point in a sequence of time steps whose incidence is greater than or equal to the threshold is considered the beginning of that epidemic.

Results

Dynamics

After fitting parameters as described above, predicted epidemic time-series were generated for each of the six localities, using the sensitivity thresholds reported in Table 1. The mean of fifty thousand simulations, and the inferred seasonalities, are shown in Figure 2, with their respective 95% confidence intervals. The reported zero-corrected correlation coefficient is simply Pearson's R^2 computed between the mean prediction and the observed incidences, with points where both time-series are zero left out, to reduce inflation of the correlation due to the large number of zeroes in the time-series.

A significant number of predicted epidemics have a right shoulder, where the model predicts that epidemics take longer to go extinct than those observed. Depending on locality, many of these shoulders are small (Akureyri, Hafnarfjörður, and Vestmannaeyjar), whilst for other localities, predicted epidemics may fail to go extinct entirely, demonstrating cyclical behaviour until the beginning of the next epidemic (Bornholm, the Faroe Islands, and Reykjavík). This may indicate that the inhomogeneity parameter, fixed at $\alpha = 0.97$ for these simulations, is an overestimate. Thoughts on this ? Inferred values are between 0.8 and 0.92, but resulting predictions are fairly poor.

The inferred seasonalities have wide distributions, demonstrated by their large confidence intervals. This is potentially due to the highly stochastic nature of measles recolonisations into their respective localities, which is the primary driver for when epidemics occur. Need to expand, or is it clear that this wouldn't be the case in large populations, where measles can be sustained ?

What else should be said here about Figure 2 ?

Predictability in Epidemic Sizes

Rather than considering a point-wise comparison between the predicted and observed epidemic time-series, a potentially more robust measure of predictability is the total number of infected cases that a particular epidemic will generate. We define the size of an epidemic as the sum of reported cases C_t for observed data, or I_t/ρ_t for predicted data, from the first time point in an epidemic to the time point before the next epidemic begins. Figure 3 shows the mean predicted epidemic size for each observed epidemic for the six localities. Several of these localities show a strong linear relationship, with near-zero intercepts and gradients around one. Fill in numbers after simulations.

Any more to be said in this section ?

Discussion

Predictions on epidemic sizes can be made with a significant level of certainty, despite sparse demographic data for all localities, mismatching incidence and demography information in Iceland, and strong spatial barriers to population mixing in the Faroe Islands.

TAU

Interpolated our monthly data for 24 periods - discuss effects of this. Trajectory matching vs simple interpolation.

Reconstructed susceptibles in six small populations (discuss Sbar / N).

Inferred reporting rate - may be indicator of bad dataset if above one ?

Low to no seasonal effects found - do we want a histogram of epidemic periods throughout the year ? No relatable seasonal effects within Iceland. Do we prefer 12-period of 24-period seasonality ? Could r be representative of some simply random effect that is not really a seasonal factor ? Can we do as well in predicting using flat r ?

SOME predictability found - given the simple model, we do OK at reconstructing epidemics. Reasonable correlation between the epidemic sizes. Duration much less predictable - possible due to how we define an epidemic and cut them off as they run into the next one.

In all cases, our epidemics are longer in time - we don't go extinct as much as we should. Can we suggest an improvement in the model to ameliorate this ?

Intercept on size plot - we should be seeing zero intercept, gradient one. We're seeing greater gradient, so we're overestimating size; positive intercept may indicate improvement needed in the model

Note Bornholm (single island, clean data effect) vs Faroe (many small islands, aggregate incidence data) vs Icelandic areas (disconnected demographics vs incidence borders). In good locations, we can certainly give a good prediction of epi size. Tend to do better for large epidemics. Gradients all ABOUT one, some slightly over (right shoulder effect ?). Large variability in error bars - some epidemics are easier to predict than others (why ?)

ROMAIN :

Why did this method work on other datasets and not this one ? Demography

What other methods could work on this data ?

What can be done to improve the quality of the data ?

Materials and Methods

Acknowledgments

References

1. Black FL (1966) Measles endemicity in insular populations: critical community size and its evolutionary implication. Journal of Theoretical Biology 11: 207–11.

2. Finkenstädt BF, Grenfell BT (2000) Time series modelling of childhood diseases: a dynamical systems approach. Journal of the Royal Statistical Society: Series C (Applied Statistics) 49: 187–205.

3. Cliff AD, Haggett P, Ord JK, Versey GR (1981) Spatial Diffusion: An Historical Geography of Epidemics in an Island Community. Cambridge University Press.

4. Iceland. Statistics Iceland. URL `www.statice.is`.

5. Denmark. Danmarks Statistik. URL `www.statistikbanken.dk`.

6. Faroe islands. Hagstova Føroy a. URL `www.hagstova.fo`.

7. Anderson RM, May RM (199wdwd1) Infectious diseases of humans: dynamics and control. Oxford University Press.

8. Metcalf CJE, Munayco CV, Chowwell G, Grenfell BT, Bjørnstad ON (2010) Rubella metapopulation dynamics and importance of spatial coupling to the risk of congenital rubella syndrome in Peru. Journal of the Royal Society Interface 8: 369–376.

Figure Legends

Tables

Figures

Locality	τ
Akureyri	19
Bornholm	15
Faroe Islands	15
Hafnarfjörður	8
Reykjavík	18
Vestmannaeyjar	7

Table 1. Sensitivity thresholds τ for each locality, fit by maximising the correlation between the mean simulated epidemic time-series and the reported incidence data.

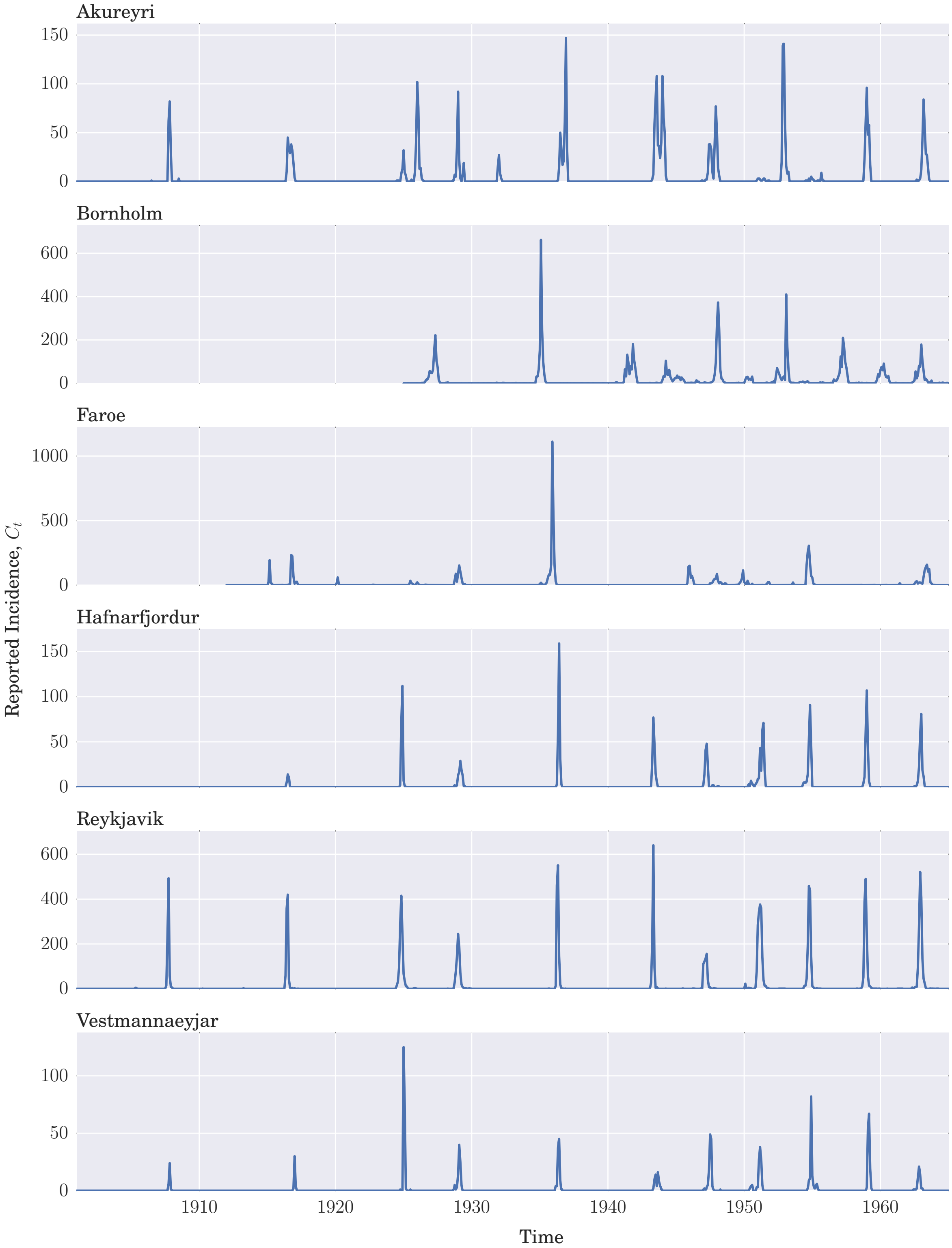

Figure 1. Reported incidence for Bornholm, the Faroe Islands, and four localities in Iceland. High temporal synchronicity can be seen in the Icelandic localities.

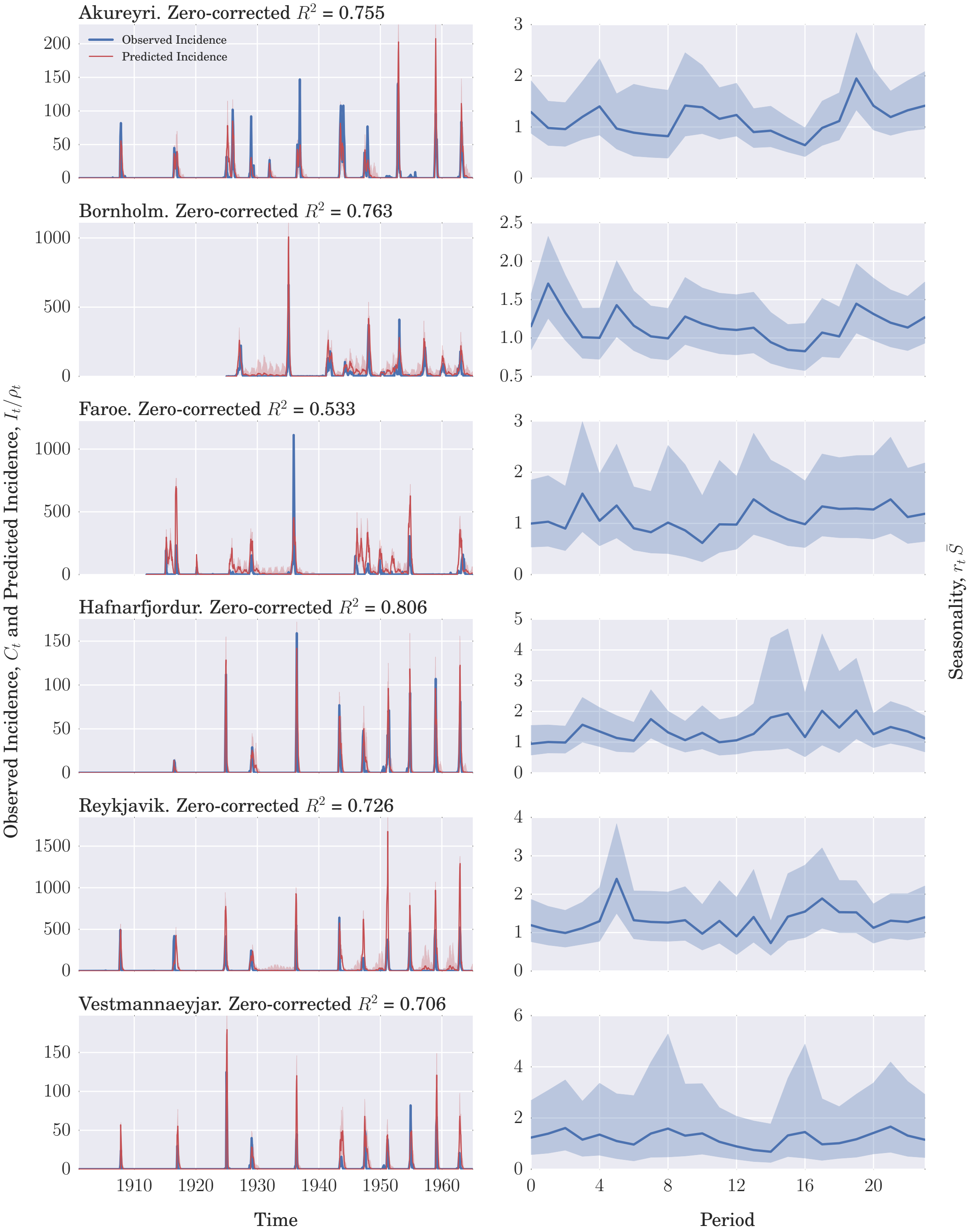

Figure 2. Predicted incidence and inferred seasonal trends. For the predicted time-series, the mean value of incidence simulations is plotted as a dark red line, with 95% confidence intervals given in light red. Seasonality is plotted as a function of the biweek, with 95% confidence intervals in light blue.

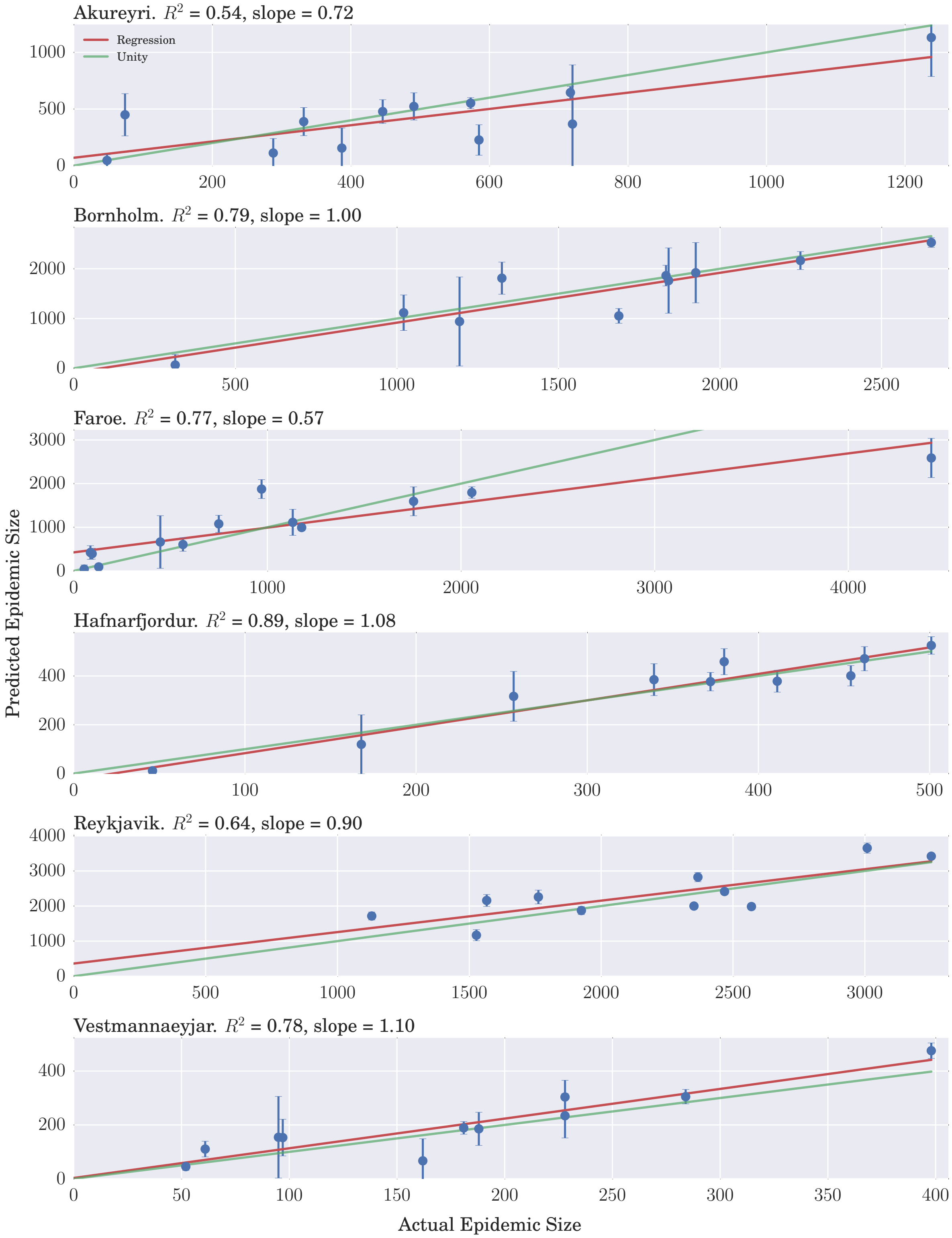

Figure 3. Predictability of epidemic sizes. The predicted size of each observed epidemic is given as a function of its observed size. Do we want to force the fits through the origin, or report the best regression line with floating intercept as we are now ?