

# PROTOCOL FOR:

Linking early warning signals to the temporal epidemiology of measles in Nigerien cities

*Andrew Tredennick (atredenn@gmail.com)*

*10 September, 2018*

## Authors

- Andrew Tredennick
- Eamon O’Dea
- Pejman Rohani
- John Drake

## Background

Theory shows that epidemic transitions can be anticipated by trends in the statistical properties of disease time series (AERO papers). The existence of statistical trends in the data that precede critical transitions, so-called ‘early warning signals’ (EWS), imply that we may be able to anticipate disease emergence and outbreaks. The end goal is a model-independent detection system, where statistical properties of disease surveillance data can trigger warnings of impending outbreaks without the need to fit mechanistic models of disease transmission (Han and Drake 2017).

However, there is currently a gap between the theoretical work, which has relied on knowing the underlying disease dynamics, and the eventual goal of applying EWS in real-world situations where the underlying disease dynamics may be unknown. Theoretical development of EWS has focused on anticipating when the population becomes supercritical, when  $\mathcal{R}_0 > 1$ , after which an outbreak is inevitable, perhaps with some bifurcation delay (Dibble et al. 2016). Knowing the value of  $\mathcal{R}_0$  through time makes it possible to test the accuracy of EWS that are estimated from state variables alone. Empirical application of EWS does not require knowing the value of  $\mathcal{R}_0$  through time, meaning that “tests” require making assumptions about when critical transitions occur. Whether EWS track and/or anticipate underlying dynamics of real disease time series remains unknown, and is a critical knowledge gap that must be filled before EWS can confidently be deployed.

To fill this gap we will fit a mechanistic model to incidence data of measles in Niger to estimate the temporal epidemiology of the disease, yielding the very same parameters that are known in data-free modeling studies. In particular, we are interested in the correlation between EWS and the time-varying reproductive ratio, known as the effective reproductive ratio ( $\mathcal{R}_E$ ). If EWS and  $\mathcal{R}_E$  are significantly and positively correlated, then we have empirical evidence that EWS are applicable in real-world settings. If EWS and  $\mathcal{R}_E$  are negatively correlated or not significantly positively correlated, then we have evidence that EWS may not be applicable in certain settings.

## Research questions

1. Do model-independent early warning signals *track* and/or *anticipate* the effective reproductive ratio?
2. Do some early warning signals outperform others in terms of their ability to track or anticipate the effective reproductive ratio?
3. Does the degree to which early warning signals track or anticipate the effective reproductive ratio depend on the magnitude or dynamics of the effective reproductive ratio itself? For example, the temporal correlation between an EWS and  $\mathcal{R}_E$  might be weak when  $\mathcal{R}_E < 1$ , where the dynamics are subcritical and the observed time series is mainly driven by environmental or demographic stochasticity.

## Study design

### I. Fit a mechanistic SIR model to measles incidence data from 4 cities in Niger

The goal is fit a model to estimate:

- The latent  $S$ ,  $I$ , and  $R$  latent states
- Time-varying rate of transmission,  $\beta_t$
- Time-varying effective reproductive ratio,  $\mathcal{R}_{E(t)}$

#### The mechanistic SIR model

The model is a continuous time SIR model with limited demography, specified as a set of stochastic differential equations (SDEs):

$$dS = \mu_t N_t dt - \lambda_t S_t dt + dU_t \quad (1)$$

$$dI = \psi_t dt + \lambda_t S_t dt - \gamma I_t dt + dV_t \quad (2)$$

$$dR = \gamma I_t dt + dW_t, \quad (3)$$

where  $\mu_t$  is birth rate at time  $t$ ,  $\gamma$  is the time-invariant recovery rate,  $\psi_t$  is rate of imported infections, and  $\lambda_t$  is the time-varying force of infection. We do not include a death process in the model because we expect death rates from the susceptible and infectious classes to be minimal relative to births and we are not interested in robust estimates of the recovered class. Excluding deaths means we can avoid making further assumptions about demographic rates – we are already making assumptions about birth rates (e.g., the rate is the same across cities, but with city-specific population size). We model stochasticity in births and imported infections by drawing time-specific values from Poisson distributions. Transitions in the model are shown in Table 1. We model  $\lambda_t$  as a function of a seasonal transmission rate subject to additional environmental variability through time:

$$\lambda_t = \frac{\beta_t I_t}{N_t}, \quad (4)$$

$$\beta_t = \beta \left( 1 + \sum_{i=1}^6 q_i \xi_{it} \right) \Gamma_t, \quad (5)$$

where  $\beta_t$  is the rate of transmission at time  $t$ ,  $\beta$  is the mean transmission rate,  $\psi$  accounts for measles infections from external sources that are not part of the local dynamics, and the term  $\sum_{i=1}^6 q_i \xi_{it}$  is a B-spline to model seasonality in transmission. The B-spline bases ( $\xi_{it}$ ) are periodic with a 1 year period. The transmission rate ( $\beta_t$ ) is also subject to stochastic process noise at each time step,  $\Gamma_t$ , which we model as a gamma-distributed white (temporally uncorrelated) noise with mean 1 and variance  $\sigma^2$  (Bretó and Ionides 2011). In this model, the effective reproductive ratio at time  $t$  is:  $\mathcal{R}_{E(t)} = \beta_t / \gamma$ .

We assume observed case reports ( $\mathbf{y}$ ) are drawn from a Poisson distribution subject to a constant reporting fraction ( $\rho$ ),

$$y_t \sim \text{Poisson}(\rho I_t). \quad (6)$$

Table 1: Transitions in the SIR model. We show the deterministic transmission rate for clarity, but our model uses the stochastic transmission rate.

Transition	$(\Delta S, \Delta I, \Delta R)$	Propensity
birth	$(1, 0, 0)$	$N_t \mu$
imported infections	$(0, 1, 0)$	$\psi_t$
transmission (deterministic)	$(-1, 1, 0)$	$SI\beta_t/N_t$
transmission (stochastic)	$(-k, k, 0)$	$\frac{S}{k} \sum_{j=0}^k \binom{k}{j} (-1)^{k-j+1} \tau_f^{-1} \ln(1 + (\beta_t I/N_t)) \tau_f(S - j)$
recovery	$(0, -1, 1)$	$I\gamma$

### Model fitting

We will fit the model to data via iterated filtering to maximize the likelihood, as implemented in the R package **pomp**. We seek to estimate 14 parameters for each city (Table 2).

Table 2: List of model parameters to be estimated.

Parameter	Description
$\beta$	Mean rate of transmission
$\gamma$	Constant rate of recovery
$\rho$	Constant reporting fraction
$q_i$ for $i \in \{1, 2, 3, 4, 5, 6\}$	B-spline coefficients for seasonality
$\psi$	External infections
$\sigma^2$	Variance of gamma-distributed environmental noise
$S_0$	Initial conditions for susceptible class
$I_0$	Initial conditions for infected class
$R_0$	Initial conditions for recovered class

Model fitting will proceed in three steps, independently for each city:

~~Generate initial parameter sets from a large Latin hypercube search space of 5000 parameter sets. Run 10 replicate particle filters at each of these parameter sets, with 2000 particles each. Retain the 500 parameter sets with the highest likelihood.~~

1. Conduct a semi-global parameter search using MIF, starting from 1000 parameter sets from a Latin hypercube. After convergence, assessed every 50 MIF iterations, we will retain the 500 estimates with the highest likelihood for the next MIF step. At this stage, we will use the following **pomp** parameters, `pomp::mif2(Np = 2000, Nmif = 50, cooling.fraction.50 = 1, cooling.type = "geometric")`, and initial parameter conditions allowed to take random walks with a standard deviation of 0.02. For states (i.e.,  $S_0, I_0, R_0$ ), we will set the initial conditions random walk standard deviation to 0.1 and the random walk will stop after 1 year (52 weeks) of the simulated dynamics.
2. Conduct a local parameter search using the retained final 500 parameter sets from above (the 500 with the highest likelihood). This parameter set will serve as starting conditions for 500 instances of iterated filtering using the following **pomp** parameters, `pomp::mif2(Np = 2000, Nmif = 50, cooling.fraction.50 = 0.95, cooling.type = "geometric")`, and initial parameter conditions allowed to take random walks with a standard deviation of 0.02. For states (i.e.,  $S_0, I_0, R_0$ ), we will set the initial conditions random walk standard deviation to 0.1 and the random walk will stop after 1 year (52 weeks) of the simulated dynamics. We will consider the parameter set with the highest likelihood at this stage to be the MLEs ( $\hat{\theta}$ ).

3. To complete our inference from the model, we will perform particle MCMC using  $\hat{\theta}$  as the initial conditions. The MCMC analysis allows us to estimate the uncertainty around all parameters and to generate estimates of the latent states and the  $\mathcal{R}_E$  for the observed data. That is, we will end up with a time series of  $\mathcal{R}_E$ , including uncertainty, that corresponds to the observation time series.

- Can we just get this from `pfilter(save.states)`?

~~Conduct a local parameter search using the the 50 parameter sets from the global search with the highest likelihoods. MHF parameters will be the same as above, except the cooling factor will be `cooling.fraction.50` = 0.9. We will consider the parameter set with the highest likelihood at this stage to be the MLEs ( $\hat{\theta}$ ).~~

## II. Calculate a suite of EWS from the observed data

We will consider 10 candidate EWS (Table 3). The EWS will be calculated using the function `spaero::get_stats()` in R for each of the case report time series shown in Figure 1 (below). We will use a “backward looking” bandwidth of 52 weeks. “Backward looking” means that the rolling window over which the EWS are calculated covers the 35 weeks prior to the focal time point  $t$ , rather than the window being centered on  $t$ .

Table 3: List of candidate early warning signals and their estimating equations. Note that  $b$  denotes the bandwidth. See Brett et al. (2018) for details.

EWS	Estimator	Theoretical Correlation with $\mathcal{R}_E(t)$
Mean	$\mu_t = \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{X_s}{2b-1}$	Positive
Variance	$\sigma_t^2 = \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s - \mu_s)^2}{2b-1}$	Positive
Coefficient of variation	$CV_t = \frac{\sigma_t}{\mu_t}$	Null
Index of dispersion	$ID_t = \frac{\sigma_t^2}{\mu_t}$	Positive
Skewness	$S_t = \frac{1}{\sigma_t^3} \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s - \mu_s)^3}{2b-1}$	Positive
Kurtosis	$K_t = \frac{1}{\sigma_t^4} \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s - \mu_s)^4}{2b-1}$	Positive
Autocovariance	$ACov_t = \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s - \mu_s)(X_{s-\delta} - \mu_{s-\delta})}{2b-1}$	Positive
Autocorrelation	$AC_t = \frac{ACov_t}{\sigma_t \sigma_{t-\delta}}$	Positive
Decay time	$\bar{\tau}_t = -\delta / \ln[AC_t(\delta)]$	Positive
First differenced variance	$\Delta\sigma_t^2 = \sigma_t^2 - \sigma_{t-\delta}^2$	Positive

## III. Calculate the correlation between EWS lags and ( $\mathcal{R}_E$ )

Answering our Research questions requires calculating the correlation between each EWS and  $\mathcal{R}_E$  through time. We are interested in the correlation between EWS and  $\mathcal{R}_E$  because these correlations tell us whether a particular EWS can anticipate a change in  $\mathcal{R}_E$ . We will calculate the temporal correlation of  $[EWS(t), \mathcal{R}_E(t)]$  for all EWS with Spearman’s  $\rho$  using the R function `cor(..., method = "spearman")`. Pearson’s correlation coefficient is not suitable for testing nonlinear associations and Kendall’s  $\tau$  does not take into account the distance between discordant pairs, so we will use Spearman’s rank correlation.<sup>1</sup> As shown in Table 2, we expect all EWS to positively covary with  $\mathcal{R}_E$  over time, a prediction we can test by calculating the correlations and their significance using the function `cor.test(..., method = "spearman")`.

There often exists a lag between when the dynamical system goes critical and when an outbreak or epidemic

<sup>1</sup>Note that Spearman’s  $\rho$  does not handle ties very well, so we may have to fall back on Kendall’s  $\tau$  if there are many ties and we do not want to rely on approximate p-values.

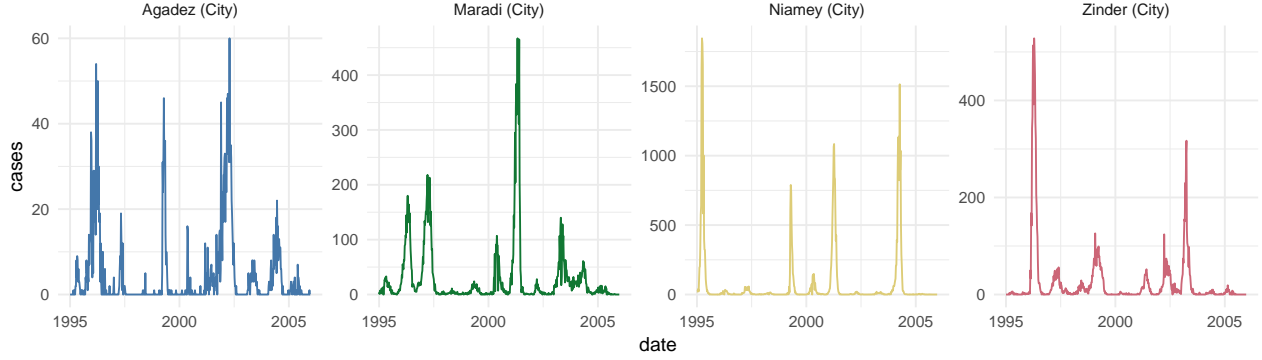


Figure 1: Time series of weekly measles case reports from four cities in Niger.

occurs, so-called bifurcation delay (Dibble et al. 2016). Therefore, EWS that correlate well with underlying dynamics (i.e.,  $\mathcal{R}_E(t)$ ) should be useful for anticipating outbreaks.

## Data sources

We will use weekly measles case report data from four Nigerien cities, collected over an 11 year period (1995-2005) (Figure 1). These data are ideal for stress testing EWS because each city has different population sizes, has different dynamics in terms of size of outbreaks and length of inter-epidemic periods, and each time series has different amounts of “noise” (though, the difference in variability is probably just reflective of population size differences). The data come from [somewhere/someone], and used here with permission from [somewhere/someone].

## Analysis and expected results

The results will start with a figure of example time series of  $\mathcal{R}_E$  and an EWS (e.g., variance), as well as a scatterplot of their correlation (all in one multipanel plot, e.g., Figure 1).

The main figure will show the correlation between each EWS and  $\mathcal{R}_E$  for each city (Figure 2). Figure 2 answers question one (see Research questions): do EWS track and/or anticipate  $\mathcal{R}_E$ ? If, in general, correlations are significant and positive, as expected by theory, then the answer is “yes.” If, in general, correlations are weak (not significant), null, or negative, then the answer is “no.” We can also indicate which correlations are significant, either in this figure or in a supplementary table.

The second question, “Do some EWS outperform others?”, will be answered by simply ranking each EWS by correlation strength (also visible in Figure 2).

The third question, “Does the degree to which early warning signals track or anticipate the effective reproductive ratio depend on the magnitude or dynamics of the effective reproductive ratio itself?”, will be answered by conducting a *post hoc* statistical analysis. First, the  $\mathcal{R}_E(t)$  time series will be broken up into clusters according to whether  $\mathcal{R}_E(t) > 1$  or not. Then, we will run the correlation analysis separately for each cluster (e.g., a single time series might have 10 sections where  $R < 1$ , and 10 sections where  $R > 1$ , yielding a sample size of 20 time series). Lastly, we will conduct an ANOVA to determine if the threshold value of  $\mathcal{R}_E$  determines the correlation between  $\text{EWS}_t$  and  $\mathcal{R}_E(t)$ . One expectation is that the correlation will be low/weak when  $\mathcal{R}_E \lesssim 1$  but high/strong when  $\mathcal{R}_E \gtrsim 1$ . An example figure is shown below (Figure 3), in which whether  $\mathcal{R}_E(t)$  is above/below the threshold of 1 does impact its correlation with each EWS.

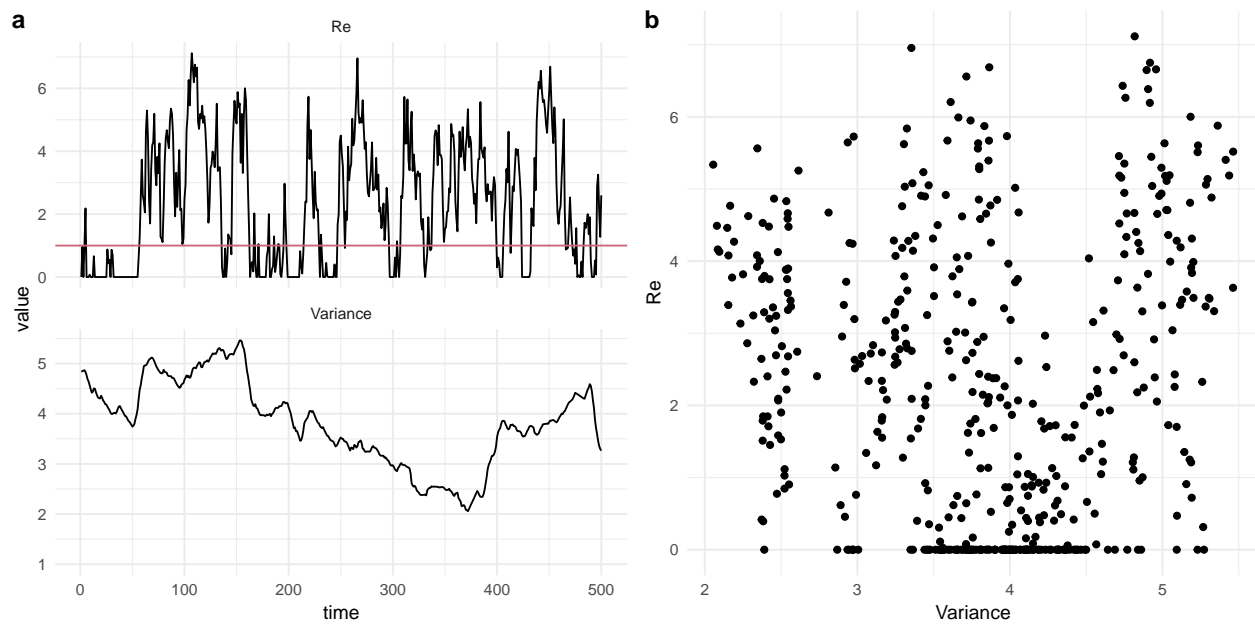


Figure 2: Time series of  $\mathcal{R}_E$  and the variance of case reports (a). Scatterplot showing the correlation between  $\mathcal{R}_E$  and the variance of case reports (b).

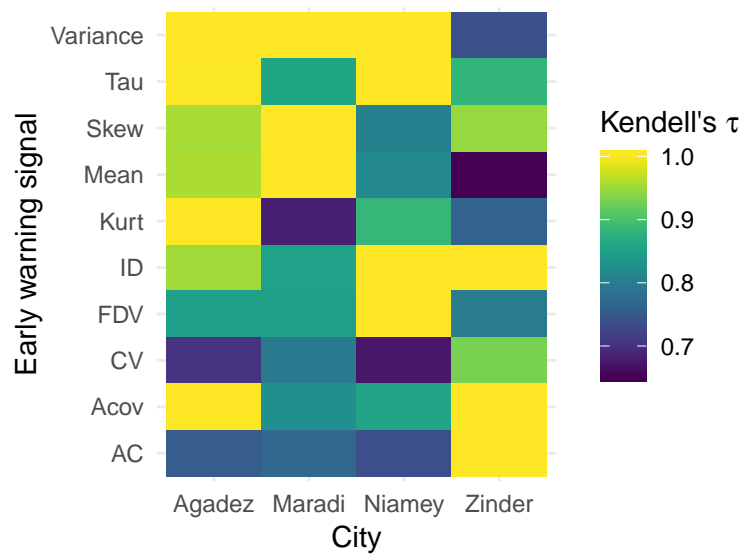


Figure 3: Heatmap of temporal correlations of EWS and  $R_E$  through time for each city.

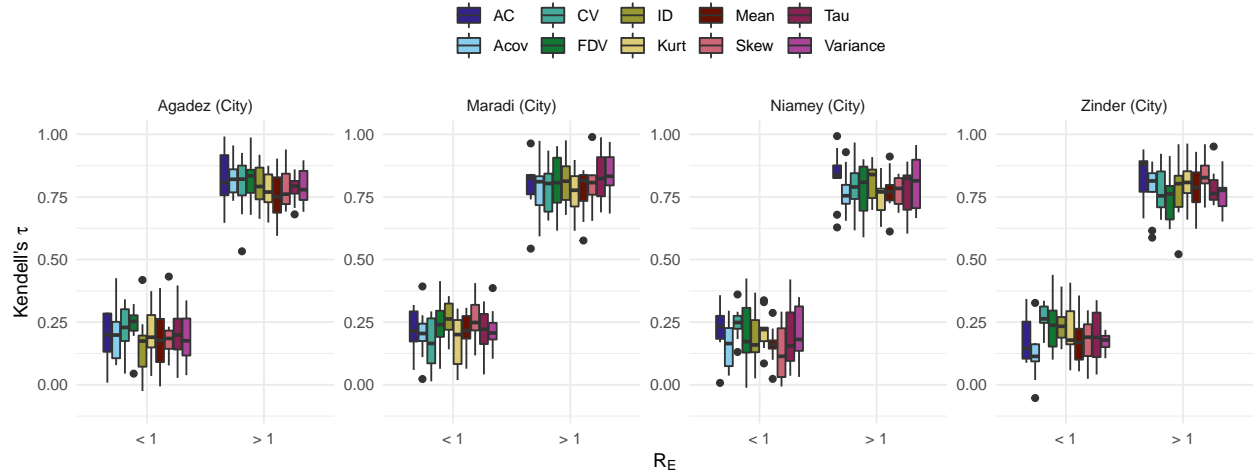


Figure 4: Boxplots of temporal correlations between EWS and  $R_E$  through time, split apart by the value of  $R_E$ .

## Checklist

- Create cleaned dataset of just the four focal cities from Niger
- Write R code to define the `pomp` model object (SDE)
- Simulate time series from the `pomp` object with parameters that give us irregular dynamics like the observed data
- Do test analysis of steps II-III using the simulated data
- Perform iterated filter parameter search for MLEs
  - Get set up on UGA cluster
  - Get set up on MIDAS cluster
- Calculate EWS on the observed data
  - Work with Eamon on bandwidth, etc.
- Calculate correlations with real data and empirically-derived  $R_E$  and lags
- Write the paper!

## Changelog

- **09-10-2018**
  - We are going to forego the proposed initial search via particle filtering because doing so over a large enough parameter space to be meaningful is too computationally expensive. Instead, after conferring with Pej, the new plan is to start with MIF from 1,000 starting parameter sets from a Latin Hypercube design. Then we'll gradually take subsets of those.
  - Imported infections are now stochastic: `psi[t] ~ rpois(psi)`.
  - Likelihood is now poisson. Initial fitting kept showing very little overdispersion, so we're going to use Poisson to speed up computation.