Environmental Research 142 (2015) 319–327

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Time series regression model for infectious disease and weather

Chisato Imai [a,*], Ben Armstrong [b], Zaid Chalabi [b], Punam Mangtani [c], Masahiro Hashizume [a]

[a] Department of Pediatric Infectious Diseases, Institute of Tropical Medicine, Nagasaki University, 1-12-4 Sakamoto, Nagasaki 852-8523, Japan
[b] Department of Social and Environmental Health, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH UK
[c] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

ARTICLE INFO

Article history:
Received 21 January 2015
Received in revised form
26 June 2015
Accepted 28 June 2015
Available online 16 July 2015

Keywords:
Time series
Method
Infectious disease
Weather
Climate

ABSTRACT

Time series regression has been developed and long used to evaluate the short-term associations of air pollution and weather with mortality or morbidity of non-infectious diseases. The application of the regression approaches from this tradition to infectious diseases, however, is less well explored and raises some new issues.

We discuss and present potential solutions for five issues often arising in such analyses: changes in immune population, strong autocorrelations, a wide range of plausible lag structures and association patterns, seasonality adjustments, and large overdispersion.

The potential approaches are illustrated with datasets of cholera cases and rainfall from Bangladesh and influenza and temperature in Tokyo. Though this article focuses on the application of the traditional time series regression to infectious diseases and weather factors, we also briefly introduce alternative approaches, including mathematical modeling, wavelet analysis, and autoregressive integrated moving average (ARIMA) models.

Modifications proposed to standard time series regression practice include using sums of past cases as proxies for the immune population, and using the logarithm of lagged disease counts to control autocorrelation due to true contagion, both of which are motivated from "susceptible-infectious-recovered" (SIR) models. The complexity of lag structures and association patterns can often be informed by biological mechanisms and explored by using distributed lag non-linear models. For overdispersed models, alternative distribution models such as quasi-Poisson and negative binomial should be considered. Time series regression can be used to investigate dependence of infectious diseases on weather, but may need modifying to allow for features specific to this context.

1. Introduction

Time series regression (TSR) is widely used among environmental epidemiologists to examine associations between environmental predictors and adverse health outcomes. The method has been developed to evaluate the associations of air pollution and weather with all-cause mortality or morbidity in places where this is overwhelmingly due to non-infectious diseases (e.g. cardiovascular diseases). More recently, TSR approaches from this tradition have been applied to communicable diseases (Hashizume et al., 2008; Jusot and Alto, 2011; Lin et al., 2013; Luque Fernandez et al., 2009; Mangtani et al., 2006). However, the use of TSR in this context is less well explored and raises some new issues (Imai and Hashizume, 2015).

This article aims to discuss and present solutions to the most important issues arising for studies using TSR models to investigate associations of weather with infectious diseases. Though few of the issues we discuss are unique to infectious diseases, they are posed in ways that require some adaptation of the approaches developed for non-infectious diseases, and our main aim is to describe such adaptations. We make reference to alternatives to TSR that have also been considered from mathematical modeling, signal processing, or econometric traditions in particular when aspects of them can be incorporated into a TSR approach, but those methods are not described in detail.

Where we propose solutions, we illustrate them using datasets of influenza in Tokyo and cholera in Bangladesh (see Supplemental material pages 2 and 7 for details of the data). These two infectious diseases demonstrate short term immunity (or diseases with frequent changes in antigenic strains or subtypes) and long term

* Corresponding author.
E-mail addresses: chisato.imai@gmail.com (C. Imai), ben.armstrong@lshtm.ac.uk (B. Armstrong), Zaid.Chalabi@lshtm.ac.uk (Z. Chalabi), punam.mangtani@lshtm.ac.uk (P. Mangtani), hashizum@nagasaki-u.ac.jp (M. Hashizume).

http://dx.doi.org/10.1016/j.envres.2015.06.040

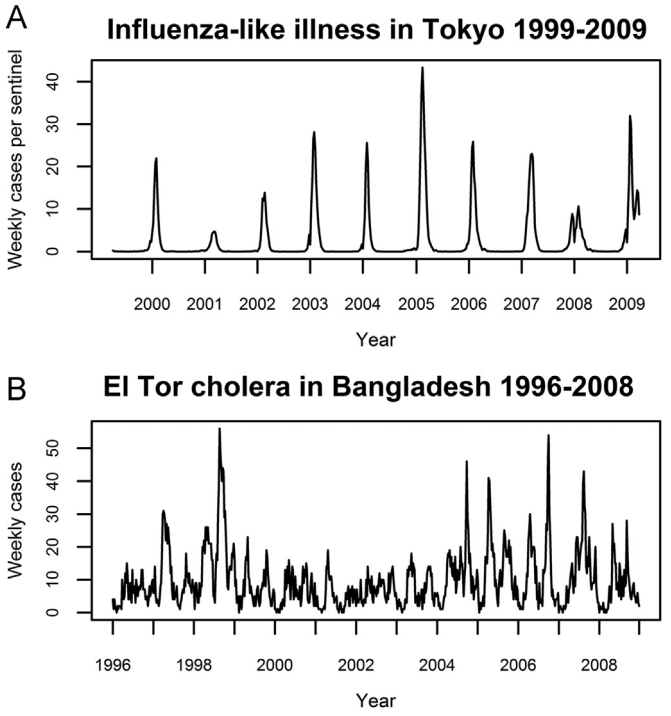

Fig. 1. (A) Weekly total influenza-like illness cases per sentinel medical facility in Tokyo, 1999–2009. (B) Weekly total El Tor cholera cases from laboratory confirmed infections from the hospital at ICDDR, B in Dhaka, Bangladesh, 1996–2008.

immunity respectively. These datasets are graphed in Fig. 1. Brief summary results are presented in the main text. More detailed results, R code, and the influenza data (the cholera data are not public) are available in Supplemental material.

This article begins with brief summaries of the TSR model typically used for non-infectious diseases in environmental epidemiology and a time series susceptible-infectious-recovery (SIR) model from the mathematical modeling tradition. These are then followed by five sections, each of which addresses an issue arising in the application of TSR to infectious diseases, and a discussion.

1.1. Overview of the time series regression model

The traditional TSR analysis seeks to identify how measured time-varying factors x_t (e.g. temperature) explain variation in an outcome series Y_t, usually daily counts of disease occurrence. The Poisson model is the most common TSR model, and can be presented as

$$\begin{aligned} Y_t &\sim Poisson(\mu_t) \\ \log(\mu_t) &= \beta_0 + \beta x_t + \sum_p \beta_p z_{p,t} + f(t) \end{aligned} \tag{1}$$

where $f(t)$ is a smooth function of time t designed to model and so avoid confounding by season and long term trend, x_t denotes an observed time varying variable of interest such as temperature, and $\{\beta_0, \beta, \beta_p\}$ are regression coefficients. Other measured risk factors are denoted as $z_{p,t}$. This model, in particular the choice of a suitable time function $f(t)$, has been reviewed in a recent tutorial paper (Bhaskaran et al., 2013). As in general in the TSR tradition, Bhaskaran focuses on acute effects to non-infectious disease outcomes.

1.2. Overview of the time series susceptible-infectious-recovery model

A feature of infectious diseases is that survivors of the disease are often immune to re-infection for some time. This causes potentially rapid changes in the population susceptible to infection. In particular, one possible explanation for the waning course of epidemics after the peak is that the susceptible population is exhausted, or at least, given herd immunity, susceptible contacts of infected cases become too sparse for infection propagation.

The SIR model is based on this and other known mechanisms for the dynamics of immunity and transmission among population. When combined with time series data this approach is called the TS-SIR (or sometimes TSIR) model. One variant of this model, simplified from Koelle (Koelle and Pascual 2004; Koelle et al., 2005), can be written in discrete time as

$$Y_t = \theta_{t-1} Y_{t-1}^{\alpha} \left(\frac{S_{t-1}}{N_{t-1}}\right)^{\gamma} \varepsilon_t \tag{2}$$

where, N_t is the total population size, S_t is the number of susceptible individuals, θ_t is pathogen transmissibility at time t and ε_t is multiplicative noise. α and γ are parameters associated with the type of mixing between individuals. S_t is not observed, but estimated from subtracting the sum of fractions of past incident cases where fractions immune κ_i are assumed to smoothly decline with the intervening time step $t - i$:

$$S_t = N_t - \sum_{i=0}^{m} Y_{t-i}\kappa_i \tag{3}$$

where m is the total duration of immunity (in time steps). At first sight this seems quite different from the traditional TSR framework, but taking logarithms and making a Taylor series approximation (details in Koelle and Pascual (2004)) reveal a strong similarity

$$\log(Y_t) \cong \log(\theta_{t-1}) + \alpha \log(Y_{t-1}) - \frac{\gamma}{N_t} \sum_{i=0}^{m} Y_{t-i}\kappa_i + \log(\varepsilon_t) \tag{4}$$

Koelle used this approach to estimate parameters θ_{t-1} for each time point (smoothed and separated from the seasonal component) and then considered associations with weather (and other explanatory factors) in a second stage, but there seems no reason why direct incorporation of explanatory variables in a single stage, as in the TSR would not be possible. However, given the large number of constrained parameters (the θ_{t-1} and κ_i), the complete Koelle model does not quite fall within the traditional TSR framework, therefore we decided not to pursue it here, though this approach may be an interesting subject for future research.

2. Topic 1: Immune population

The traditional TSR generally assumes that the population at risk of the outcome under study is more or less constant; however, as noted in the SIR overview above, immunity to infectious diseases causes variation in the susceptible population. Unless allowed for, such variations in the underlying population at risk may bias estimates of the associations with the weather.

If the size of the susceptible population were known at each time point, it could be allowed for in the model, but this information does not always exist. We review below some approaches to this problem. Choice of the approach is likely to depend on the specific infectious disease, given the large variations among infections in the duration of immunity.

2.1. Rely on smooth function of time to model changes in immunity

Much of the effect of changes in immune fractions of populations is often to induce seasonal and other long term variations of diseases (Grassly and Fraser, 2006; Pascual and Dobson, 2005).

Thus, within the TSR framework, the allowance for smooth changes over time (see topic 4) to some extent allows for changes in immunity too, but not necessarily completely, for example if the changes in immunity are faster than the smooth change function allows.

2.2. Counts of past cases as explanatory variable

Lopman et al. (2009), studying the dependence of norovirus on weather and other factors, allowed for variation in the number of susceptible individuals from year to year by including as an explanatory variable named a "population immunity factor" that was essentially the number of cases in the previous year, which assumes an immunity duration acquired by virus infection of about 1 year. Lopman's ad hoc approach can thus be seen as a special case of Koelle's TS-SIR model (Koelle and Pascual, 2004), in which the number of immune persons is $\frac{\gamma}{N_t}\sum_{i=0}^{m} Y_{t-i}\kappa_i$, where the κ_i is assumed to be 1 (last year) otherwise 0, and the population size N_t is constant. This suggests a range of "simplified Koelle's TS-SIR" approaches in which sums of past counts of disease are included as explanatory variables in the TSR.

In our influenza dataset, once the strong autocorrelation had been allowed for (see Topic 2), using an explanatory variable of the sum of cases in the same season ("immune factor") significantly predicted fewer cases (55% case reduction at the maximum immune factor observed) and changed the estimated temperature effect (details in Supplemental Material, Table S3). For El Tor cholera, for which immunity lasts longer, this immunity factor was substantially correlated with the variables controlling for a long term time trend and had little impact on the results. (Supplemental material, Table S1).

2.3. Onset analysis using binary outcomes

Rather than using counts of disease in each time period (Y_t), this approach focuses on timing of onset of epidemics (the time when Y_t exceeds some threshold point), and investigates the association of the timing of onset or peaks of epidemics with weather conditions up to that time. By assuming that the susceptible population is relatively constant up to time of onset, the need to estimate the population is avoided. For example, a study found that low absolute humidity predicted onset of wintertime influenza epidemics four weeks later (Shaman et al., 2010). This onset approach is particularly attractive for diseases for which the immunity levels of the population are reliably low at the start of the disease season, due to short immunity or new emerging subtype of the disease (e.g. norovirus and influenza).

Difficulties include the definition of the onset threshold count, and the need for many epidemic episodes to achieve adequate statistical power, since each epidemic episode is effectively one data point. Most published analyses investigating the association of onset times with weather have been descriptive, for example using correlations (Sultan et al., 2005), but survival analysis approaches are also possible, either assuming a parametric form for time to onset using a logistic model (in influenza season) or non-parametrically using a Cox model (Table 1).

2.4. Analysis of incidence data after excluding the onset or peak of epidemics

Although this type of analysis has not yet been published, an approach closely related to the timing of onsets or peaks would be to use a conventional TSR of disease counts, but to exclude all times beyond the points in epidemics when the proportion of population susceptible is thought likely to become importantly low, such as the peak, or (more conservatively) onset of the "take-off" slope of the epidemic.

In our ten-year Tokyo influenza dataset, the associations of time to onset and peak were indeed imprecisely estimated compared to approaches using counts due to a lack of statistical power as there were only 10 onset and peak points. Results from TSR analyses of counts excluding weeks after onset or peaks were

Table 1
Model comparison for Tokyo influenza.

Models	Outcome (Y)	Predictors		Distribution[a]	Temperature effect	Dispersion parameter[c]
		AC term	Immune term		Estimate % (95% CI)[b]	
TSR with quasi-Poisson						
Standard TSR	count	–	–	QP	−5.8 (−10.9, −0.5)	349.4
+ Autocorrelations (AC)[d]	count	$residual_{t-1}$	–	QP	−3.7 (−6.9, −0.5)	118.4
+ AC	count	Y_{t-1}	–	QP	−4.8 (−8.6, −0.8)	188.7
+ AC	count	$\log(Y_{t-1}+1)$	–	QP	−5.5 (−7.5, −3.4)	48.4
+ AC	count	$\log(Y_{t-1}+0.5)$	–	QP	−5.5 (−7.5, −3.4)	49.0
+ AC+immune term	count	$\log(Y_{t-1}+1)$	Σ(cases so far in season)	QP	−6.7 (−8.7, −4.6)	47.1
Up to onsets	count	$\log(Y_{t-1}+1)$	–	QP	−5.7 (−12.0, 1.0)	9.0
Up to peaks	count	$\log(Y_{t-1}+1)$	–	QP	−4.5 (−8.9 to 0.1)	67.7
TSR with different distribution models						
Negative binomial	count	$\log(Y_{t-1}+1)$	Σ(cases so far in season)	NB	−6.6 (−10.0, −3.0)	na
Linear regression	log(count+1)	$\log(Y_{t-1}+1)$	Σ(cases so far in season)	Gaussian	−5.2 (−8.6, −1.6)	na
Non-TSR models[e]						
Onset: logistic regression	onset (1,0)	$\log(Y_{t-1}+1)$	–	Bernoulli	−19.5 (−49.9, 29.4)	na
Onset: cox regression	onset (1,0)	$\log(Y_{t-1}+1)$	–	CB	−16.0 (−46.7, 32.4)	na
Peak: logistic regression	peak (1,0)	$\log(Y_{t-1}+1)$	–	Bernoulli	−50.7 (−75.3, −1.7)	na
Peak: cox regression	peak (1,0)	$\log(Y_{t-1}+1)$	–	CB	−51.6 (−78.3, 7.7)	na

[a] Parameter φ estimated by (Pearson)/(residual d.f.), lower values represent better fit.
[b] QP: quasi-Poisson, NB: negative binomial. CB: conditional Bernoulli.
[c] Temperature effect estimate (%)=Excess Relative Risk/100=[exp(β)−1]/100=% increase in cases per 1 °C increase of temperature.
[d] We chose to explore autocorrelation first because exploratory analyses suggested it as the most important factor.
[e] For these models the temperature coefficient is the change in log odds of onset or peak per degree.

broadly similar to those including all the weeks following onsets or peaks, though with greater standard errors (Table 1). We did not attempt these analyses for cholera, for which the definition of onset or peaks would be problematic due to less obvious epidemic episodes.

3. Topic 2: Strong autocorrelations caused by disease transmission

Most TSR analyses of non-infectious diseases find that autocorrelation of disease counts is small after seasonality and long term trend is adjusted for. Residual autocorrelation, if any, is usually attributed to failures to include risk factors that vary slowly in time and allowed for by fitting lagged residuals from a standard model as "autocorrelation terms" (Brumback et al., 2000). For residuals from TSR of infectious diseases, autocorrelation is often much stronger than for non-infectious diseases, and consideration of mechanism suggests that "true contagion" (Cameron and Trivedi, 2013) rather than omitted important risk factors is likely to be the primary cause. That is, the number of cases occurring at time t will be directly dependent on the size of the infectious population, which is the number of cases that occurred in the recent past. For non-infectious diseases, failure to take account of autocorrelation is rarely a major problem and usually biases standard errors more than regression coefficients (e.g. weather-disease associations) (Campbell, 2005; Schwartz et al., 1996), but this depends on the data rather than being a mathematical general result, so may not be the case for infectious diseases.

3.1. Modification motivated by the time series SIR model

SIR models such as that described in Eq. (2) suggest that for infectious diseases, rather than including lagged residuals as a term in the model, it matches the likely mechanism better to include the logarithm of lagged outcome counts, $\log(Y_{t-1})$. This approach of including $\log(Y_{t-1})$ falls within the class of transitional regression models considered by Brumback et al. (Brumback et al., 2000) and is suggested explicitly by Cameron and Trivedi (Cameron and Trivedi, 2013).

The choice of lag (e.g. 1, 2, or both) should be informed by the time from onset of one case to another infected from the first (serial interval). With weekly data which is typical for infectious diseases, for instance, 1 week lag would be appropriate for diseases with short incubation periods such as cholera which is usually 2–3 days (Heymann, 2008).

The inclusion of past cases would cause downward bias (i.e. overadjustment (Schisterman et al., 2009)) in the total effect of a longer lagged weather exposure if part of the effect was through a causal path involving a more immediate impact of that exposure causing higher counts and hence infectious persons at time $t-1$. However if like Koelle we are interested in how the weather exposure impacts on disease transmissibility θ, the effect mediated by this pathway should be excluded and for this Koelle's model Eq. (4), suggests the transmission term $\alpha \log(Y_{t-1})$ is required. Nevertheless, clarifying this more definitively, perhaps by simulation, would be desirable.

In both the influenza and cholera examples, residual autocorrelation of the standard TSRs was high (first order $r=0.88$, 0.63 respectively). Including $\log(Y_{t-1})$ as covariate predicted outcomes better by reducing residual dispersion than either residuals or unlogged Y_{t-1}, left less auto-correlation in residuals (see Table 1 and Supplemental material, Figs. S2 and S5), and altered the weather effect estimates.

4. Topic 3: A wide range of plausible lag structure and association patterns

Associations between weather variability and health outcome are rarely linear. In addition, the effects of weather are typically delayed. These features are present in non-infectious diseases, but infectious diseases can display patterns of associations and lag structures that are not generally considered in non-infectious diseases. One reason is that, for infectious diseases, the casual pathway can be more complicated than for non-infectious diseases. A striking example is the association between Lyme disease incidence with rainfall in the spring two years before (Subak, 2003). This complex association is plausible because of the biology of the tick vector, rodent host population, and human behavior in response to the weather variability. In another example of mechanisms leading to complex associations and lag structures, precipitation can drive much malaria activity in Africa, as it builds the water pools necessary for vector breeding and larval habitats. However, heavy precipitation has a flushing effect and thus can remove habitats (Reiter, 2001). Drought may also eliminate mosquitoes in area with little standing water, but at the same time, create safe havens for mosquitoes in areas with abundant water by eliminating predators (Lafferty, 2009). Choice of broad lag structures and association patterns in infectious diseases should be specific to the diseases and study locations.

4.1. Biological plausibility and likely mechanisms

Choice of broad lag structures and association patterns in infectious diseases should be specific to the diseases and study locations. Biological plausibility and preceding studies can often help inform the reasoning for the primary focus for the duration of delayed effects and the shapes of the association patterns (linear, U, J shaped, etc.). If the biology suggests long lags, then the range of possible models for the control of long term trends and seasonality is also determined by this information (see Topic 4).

4.2. Distributed nonlinear lag estimation analysis

Distributed lag non-linear models (DLNMs), implemented in the flexible R package dlnm (Gasparrini, 2011; Gasparrini et al., 2010) allows researchers to explore non-linearity of various form (spline curves, "hockey stick", step functions) combined with flexible lag structures (associations with varying delays). Studies have used this method to simultaneously explore and evaluate the delayed effects and association patterns of the weather variability of interest (Eisenberg et al., 2013; Kim et al., 2012; Zhao et al., 2014). Fig. 2 shows a DLNM fitted to the rain-cholera data which suggests a leveling off of the effect of rain above 400 mm/week. The delayed association was spread over five weeks but peaked at two weeks after high rain.

5. Topic 4: Controlling for seasonality and long term trends

The same techniques for controlling for seasonality and long term trends for non-infectious diseases (Bhaskaran et al., 2013) apply to infectious diseases, although the nature of the patterns may be different. That is, some combination of splines of time, sine-cosine (Fourier) terms, and month or season indicator variables. Here we confine the discussion to some particular considerations for some infectious disease studies.

5.1. Omit season and trend to avoid unnecessary adjustment

In a previous review study, several studies with no adjustments

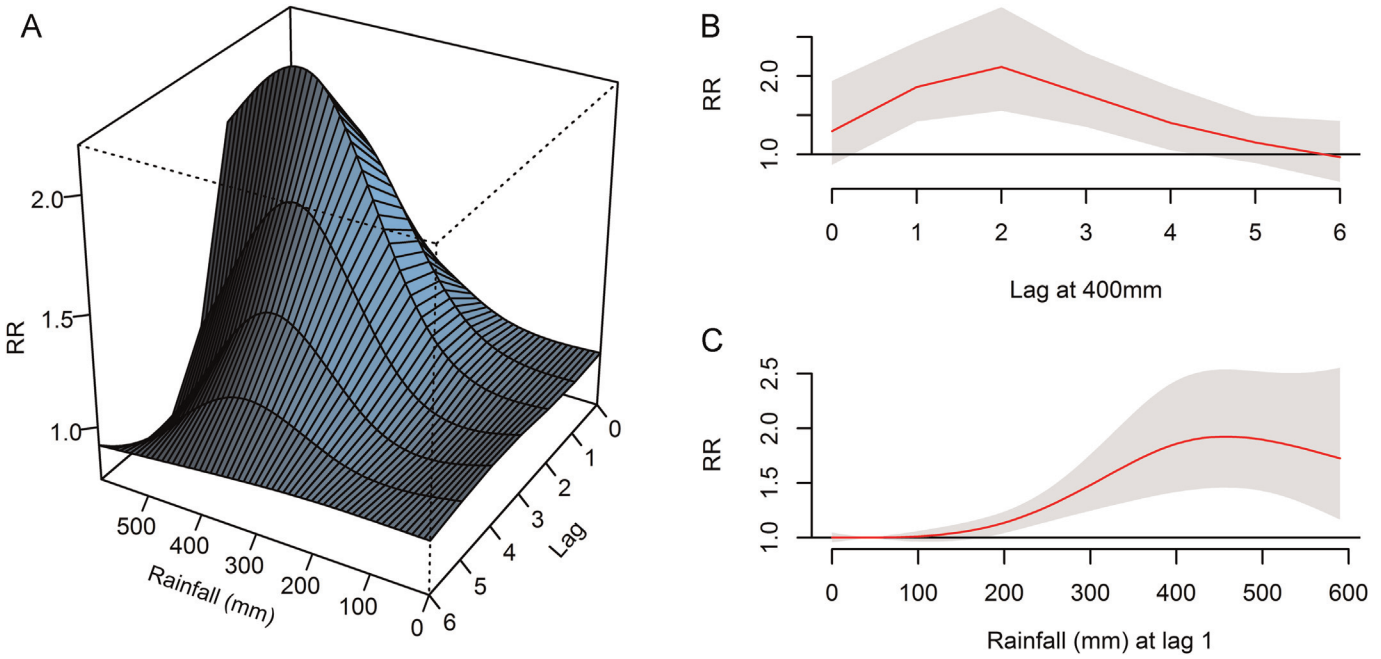

Fig. 2. (A) 3 dimensional plot of relative risk (RR) of cholera by rainfall per week and lag. (B) The two-dimensional plot "slice" at 400 mm rainfall. (C) The two-dimensional plot slice at lag 1 week.

for seasonality and trends were reported (Imai and Hashizume, 2015). A possible reason for this is that the researchers believed that these patterns could be entirely explained by environment, weather, biological systems and human pathogenesis, which are already in the models. If indeed they are so explained, there would be no need for seasonal and trend model terms, and including such terms would diminish power and reduce precision in estimates of parameters of interest (i.e. unnecessary adjustment (Schisterman et al., 2009)).

Whether inclusion of such terms (i.e. $f(t)$ in Eq. (1), typically Fourier functions or spline basis variables) can cause over-adjustment bias is controversial. On the one hand, it might be argued that calendar time (t) is constructed to follow the rotation of the earth round the sun and thus time, through season, is very closely linked causally to weather and climate, so could be on the causal pathway between weather and disease. In that case adjusting for season would give only the part of the association unmediated by seasonality. On the other hand, though season is clearly causally linked to weather, the direction of the causal chain could be argued to be *season* → *weather* → *disease* rather than *weather* → *season* → *disease*, in which case season would not be between weather and disease, but upstream of weather. This argument can also be put algebraically. The explanatory variables in question $f(t)$ are functions of time t alone, so cannot be affected by weather, suggesting that they cannot be on the causal pathway as mediators. In this case their inclusion, even if unnecessary to control confounding, would not cause bias in the coefficients of weather variables though may lose precision in them. Further work to clarify this would be useful.

In light of the controversy about bias, there is a case for researchers to include results unadjusted for season, but it seems to us unwise to present no results adjusted for season except possibly if residual seasonal patterns beyond those explicable by the included weather terms are tested for and found not to be present.

5.2. Allow potential complexity in seasonal patterns

There may be drivers of seasonal patterns for infectious diseases that are different from non-infectious diseases. School terms and national holidays may be important in determining population mixing and hence infection. Thus, standard models for non-infectious diseases may not be adequate for infectious diseases. Following Bhaskaran et al. (2013) and Schwartz et al. (1996) we favor choice of season/trend model being informed by *a priori* considerations as well as whether the terms are significant with the data. Further, because model choice in this (and other respects) is uncertain, we favor undertaking sensitivity analyses in which the impacts of alternative model choices on the results of interest are tabulated.

5.3. Separate functions for seasonal and long term patterns for long lag effects

The most common practice for non-infectious disease is to include a spline function of time, typically with 4–12 degrees of freedom (d.f.) per year, to allow for both seasonal and other long term trends. This works well for estimating acute weather effects (e.g. lags by two or three weeks on daily based data), but depletes precision for identifying longer lag associations because the time spline competes with the longer lag weather terms in explaining outcome variation. If longer lag effects need to be considered, using annual cyclic sine-cosine pairs and to model season and a low-order time spline (e.g. 1–2 d.f. per year) or year indicator to control inter-annual variation are alternatives that reduce the competition between the season/trend terms and the weather terms.

In the example datasets (both weekly), we examined associations on the time-scale of weeks or months but not longer anticipated substantial seasonality, and were not confident that between-year variation would be smooth. We thus used annual period sine-cosine pairs and 6 harmonics for the cholera example and year indicators as the base model. We found that these season and trend terms were indeed significant, indicating that some form of control was indicated, though a full analysis would need to explore sensitivity to these specific model assumptions.

6. Topic 5: Overdispersion

Infectious disease counts often have a variance above that

expected in a Poisson distribution, much more than for non-infectious diseases. The selection of a method for allowing for overdispersion is, therefore more important for infectious diseases. This selection of alternative models affects estimates of standard errors more than estimates of associations of interest (coefficients) themselves (Cox, 1983; Lee et al., 2012). Here we briefly mention only the most popular options, because more extensive discussions have been published elsewhere (Zeileis et al., 2008).

6.1. Quasi-Poisson and negative binomial

Most commonly used methods to overcome overdispersion are the quasi-Poisson and negative binomial models. Both have an overdispersion parameter (denoted by φ and ψ in the following equations), but assume a different relationship of variance (σ^2) to mean (μ): quasi-Poisson assumes $\sigma^2 = \varphi\mu$ whereas the negative binomial assumes $\sigma^2 = \mu + \mu^2/\psi$ (O'Hara and Kotze, 2010). In most cases, those two models yield similar estimates of regression coefficients, but not always, in particular if overdispersion is high. For choosing between the two models, some model fit statistics such the Akaike information criteria (AIC), Bayesian information criteria (BIC), or quasi-AIC (QAIC) are often considered, but these methods are questionable for comparing models with quasi-likelihood and regular likelihood (Ver Hoef and Boveng, 2007). Although there is no generally established approach for this aspect of model selection, Ver Hoef proposed plotting the sums of squared residuals, $\sum_t (Y_t - \mu_t)^2$, against fitted counts μ_t. A straight line suggests quasi-Poisson whereas a quadratic curve suggests negative binomial (Ver Hoef and Boveng, 2007). Both our influenza and cholera datasets showed substantial overdispersion in a quasi-Poisson model (47.1 and 1.9 respectively). Ver Hoef's method showed the quasi-Poisson model fitted the influenza data better, but both models fitted the cholera data about equally well (Fig. 3).

6.2. Zero-inflated Poisson and negative binomial

Suitable for data with excessive zero count outcomes, this is a two-component mixture model combining probability of having a zero count (a point mass at zero) with a count distribution following Poisson or negative binomial (Lambert, 1992). This model can separate how explanatory variables affect the chance of having zero cases from how they affect the mean of cases on non-zero periods. The interpretation of results, in return, could be complicated (Loeys et al., 2012).

6.3. Gaussian linear model for $\log(Y_t)$

Where counts are consistently reasonably large (say mainly greater than 10) (Peduzzi et al., 1996), the Gaussian linear model has many attractions, as it allows many features available in linear regression. Its essential assumption, that the residual variance does not vary, in particular with the fitted value, can be easily checked. Zeroes present a problem which is usually resolved by working with $\log(Y_t + 0.5)$ or $\log(Y_t + 1)$. Some authors have pointed out limitations in this approach (Cameron and Trivedi, 2013; O'Hara and Kotze, 2010), but their practical implication for this context is not clear.

Residual analyses of the linear models for $\log(Y_t)$ indicated poor fit in both our datasets (Supplemental material, Figs. S3 and S8), but the weather effect estimates were broadly similar to the Poisson model (Table 1 and Supplemental material, Table S2).

7. Summary of models fitted to the example datasets

In the influenza analysis, the estimated temperature effect was appreciably different in magnitude (though not in direction) between TSR and non-TSR models (Table 1). Applying a range of models in sensitivity analyses would therefore seem a sensible precaution.

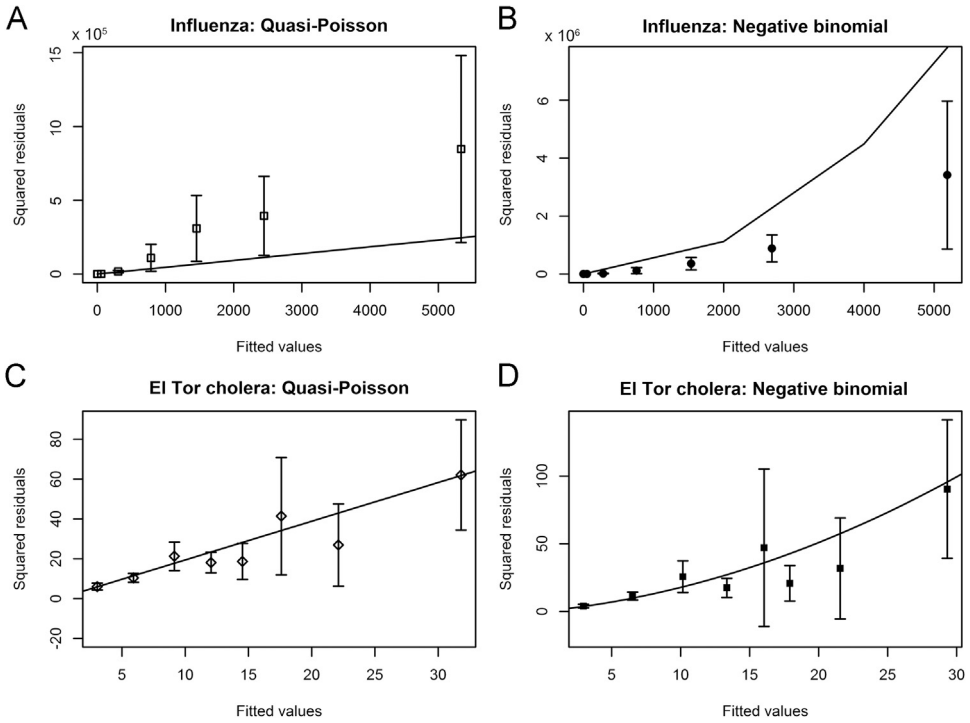

Fig. 3. Comparison of fit of quasi-Poisson and negative binomial models using Ver Hoef's method. The error bars are the 95% CI around the group means of squared residuals $(Y_i - \mu_i)^2$, plotted against mean expected values μ_i where subscript i denotes the ith observation. The observed μ_i were classified into groups divided at quantiles 0.25, 0.5, 0.7, 0.8, 0.85, 0.9. The lines present the variances expected for the distribution (A) and (B) Influenza, and (C) and (D) El Tor cholera.

Table 2
Summary table.

Issues	Potential approaches	Further discussion and alternative methods
Immune population	• Rely on smooth function of time to model changes in immunity. • Include counts of past cases as an explanatory variable. • Consider determinants of time of epidemic onset each year as in a survival analysis. • Analyze the incidence after excluding data after the onset or peak of epidemics.	• Consider integrating TSR and TS-SIR model traditions.
Strong autocorrelations by disease transmission	• Include $\log(Y_{t-1})$ as a covariate.	• Consider analysis models such as autoregressive integrated moving average (ARIMA) and wavelet analysis.
A wide range of plausible lag structure and association patterns	• Be informed by biological plausibility and likely mechanisms (*a priori* approach). • Use distributed lag non-linear models (DLNMs).	
Controlling for seasonality and long term trends	• Omit season and trend to avoid unnecessary adjustment • Consider potential complexity in seasonal patterns. • Consider using separate functions for seasonal and long term patterns for long lag effects.	• Consider how much seasonality and other time control is required and appropriate.
Overdispersion	• Consider quasi-Poisson, negative binomial, zero-inflated, and transformed Gaussian models. Use Ver Hoef's method to compare negative binomial and quasi-Poisson models graphically. For series with few small counts, use linear models with log(count) as the dependent variable.	• Consider how to compare goodness of fit of models with different distributional assumptions (non-nested).

Within the TSR models of both influenza and cholera datasets, the inclusion of the proposed autocorrelation and immune terms improved fit of the model considerably compared to the standard TSR model (Table 1 and Supplemental material, Table S2).

8. Alternative methods of investigating associations of weather with infectious diseases

Although the primary focus of this paper is on using TSR approaches, we also briefly describe here two alternative approaches: Auto-Regressive Integrated Moving Average (ARIMA) models and wavelet models.

8.1. ARIMA models

ARIMA models are popular in econometrics and environmental time series analysis and have been used in infectious disease modeling (Unkel et al., 2012). A simple special case of an ARIMA model is the autocorrelation model as described above including $\log(Y_{t-1})$ as an explanatory variable, but it also allows more complex forms of dependence on past outcome counts. A key assumption of ARIMAs model is that the observed time series $\{Y_t\}$ is stationary (the expected count and variation about it does not change over time). Non-stationarity could be incorporated into ARIMA models along the lines discussed under "seasonal and trend" above, but distinct methods have been developed in the ARIMA tradition, often removing trends from the series before fitting the model, and including the "seasonal ARIMA (SARIMA)" group of models. Another assumption, that autocorrelation decays exponentially, can be relaxed by using fractional ARIMA models (Hussain et al., 2005).

ARIMA models can be used to investigate the associations of weather with infectious diseases. In ARIMA models this is usually done by incorporating explanatory "exogenous" variables into the model, which can be written schematically in the form of expression (1) as

$$log(Y_t) = \beta_0 + \beta x_t + [\text{ARIMA terms}] \quad (5)$$

8.2. Wavelet models

A more radical and novel approach to non-stationarity and the outbreak nature of infectious diseases is the use of wavelet transforms (Cazelles et al., 2007). The advantage of such wavelet analysis is that is can capture outbreaks well, by decomposing the time series in two-dimensions (time and frequency)

Wavelet cross-spectrum and wavelet coherence can be used to determine the degree to which the infectious disease incidence and weather are related. The coherence function is unity when there is a perfect linear relationship between the two time series at a particular frequency and time and zero when they are completely independent. For example, using wavelet analysis, Cazelles et al. (2007) found that the incidence of cholera in Ghana was "weakly coherent" with the El Niño Southern Oscillation time series.

A limitation of ARIMA and wavelet models for sparse count data is that methods, other than by log transformation of the outcome series, have not yet been developed to allow their application. Statistical analysis using wavelet models, even of transformed counts, is quite complicated (Cazelles et al., 2014).

9. Discussion

We have discussed ways in which conventional TSR techniques for investigating associations of weather with adverse health events may need to be extended when considering infectious disease counts. The potential approaches we have discussed are

summarized in Table 2; some have been previously proposed and some are new.

To our knowledge, this is the first published discussion of these issues. We have focused on a broad overview of the available methods, but future studies may develop individual methods in more details. In general the solutions we suggest, in particular as regards allowing for immunity and variation in susceptible population, require further development and evaluation of their validity to be completely confident of them. Meanwhile, we suggest sensitivity is assessed to any assumptions which are in doubt by fitting alternative models also.

We briefly reviewed approaches other than the conventional TSR models, which poses the question of whether it may be more appropriate to use one of these methods than TSR. While this may sometimes be the case, the familiarity of many epidemiologists with the traditional TSR, its focus on the environment-health association, and the relatively direct interpretation of results in public health terms (relative risks, attributable burdens) give TSR advantages in this context.

We found a considerable overlap in published time series SIR (TS-SIR) models with TSR models. Indeed two of our proposed extensions of TSR (i.e. using sums of past cases to allow for immunity and logged immediate past cases to allow for autocorrelation due to true contagion) were motivated by similar terms in TS-SIR models. The limitations of the full TS-SIR approaches published so far, apart from their complexity, are that regressors (e.g. weather factors) are not usually incorporated directly. An interesting direction for future research would be to incorporate weather dependence in TS-SIR models while formulating them in a way allowing estimation by TSR software, thus forging a convergence between TS-SIR and TSR approaches.

In conclusion, TSR models may be used to investigate the dependence of infectious disease on weather, but are likely to require modifying to allow for features specific to this context.

Funding

None declared.

Conflicts of interest

None declared.
No animals and human subjects involved in this study.

Acknowledgments

We thank Yoonhee Kim for her support with demonstration analysis, ASG Faruque for sharing cholera data with us, and Katia Koelle for explaining her research to us that provided important insights to our study. Chisato Imai is also supported by JSPS fellowship. This study was supported by the School of Tropical Medicine and Global Health, Nagasaki University.

Appendix A. Suplementary Information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.envres.2015.06.040.

References

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., Armstrong, B., 2013. Time series regression studies in environmental epidemiology. Int. J. Epidemiol. 42, 1187–1195. http://dx.doi.org/10.1093/ije/dyt092.

Brumback, B.A., Ryan, L.M., Schwartz, J.D., Neas, L.M., Stark, P.C., Burge, H.A., 2000. Transitional regression models, with application to environmental time series. J. Am. Stat. Assoc. 95, 16–27.

Cameron, A.C., Trivedi, P.K., 2013. Regression Analysis of Count Data. Cambridge University Press, New York.

Campbell, M.J., 2005. Time Series Regression. In: Encyclopedia of Biostatistics. John Wiley & Sons, Ltd., USA.

Cazelles, B., Chavez, M., Magny, G.C., Guegan, J.F., Hales, S., 2007. Time-dependent spectral analysis of epidemiological time-series with wavelets. J. R. Soc. Interface 4, 625–636. http://dx.doi.org/10.1098/rsif.2007.0212.

Cazelles, B., Cazelles, K., Chavez, M., 2014. Wavelet analysis in ecology and epidemiology: impact of statistical tests. J. R. Soc. Interface 11, 20130585. http://dx.doi.org/10.1098/rsif.2013.0585.

Cox, D.R., 1983. Some Remarks on Overdispersion. Biometrika 70, 269–274.

Eisenberg, M.C., Kujbida, G., Tuite, A.R., Fisman, D.N., Tien, J.H., 2013. Examining rainfall and cholera dynamics in Haiti using statistical and dynamic modeling approaches. Epidemics 5, 197–207. http://dx.doi.org/10.1016/j.epidem.2013.09.004.

Gasparrini, A., Armstrong, B., Kenward, M.G., 2010. Distributed lag non-linear models. Stat. Med. 29, 2224–2234. http://dx.doi.org/10.1002/sim.3940.

Gasparrini, A., 2011. Distributed lag linear and non-linear models in R: the package dlnm. J. Stat. Softw. 43, 1–20.

Grassly, N.C., Fraser, C., 2006. Seasonal infectious disease epidemiology. Proc. Biol. Sci. 273, 2541–2550.

Hashizume, M., Armstrong, B., Hajat, S., Wagatsuma, Y., Faruque, A.S., Hayashi, T., et al., 2008. The effect of rainfall on the incidence of cholera in Bangladesh. Epidemiology 19, 103–110.

Heymann, D.L., 2008. Control of Communicable Diseases Manual, 19th ed. American Public Health Association, Washington DC.

Hussain, S., Harrison, R., Ayres, J., Walter, S., Hawker, J., Wilson, R., et al., 2005. Estimation and forecasting hospital admissions due to Influenza: Planning for winter pressure. The case of the West Midlands, UK. J. Appl. Stat. 32, 191–205.

Imai, C., Hashizume, M., 2015. A systematic review on methodology: time series regression analysis for environmental factors and infectious diseases. Trop. Med. Health 43, 1–9. http://dx.doi.org/10.2149/tmh.2014-21.

Jusot, J.F., Alto, O., 2011. Short term effect of rainfall on suspected malaria episodes at Magaria, Niger: a time series study. Trans. R. Soc. Trop. Med. Hyg. 105, 637–643. http://dx.doi.org/10.1016/j.trstmh.2011.07.011.

Kim, Y.M., Park, J.W., Cheong, H.K., 2012. Estimated effect of climatic variables on the transmission of Plasmodium vivax malaria in the Republic of Korea. Environ. Health Perspect. 120, 1314–1319.

Koelle, K., Pascual, M., 2004. Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. Am. Nat. 163, 901–913.

Koelle, K., Rodo, X., Pascual, M., Yunus, M., Mostafa, G., 2005. Refractory periods and climate forcing in cholera dynamics. Nature 436, 696–700.

Lafferty, K.D., 2009. The ecology of climate change and infectious diseases. Ecology 90, 888–900.

Lambert, D., 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Lee, J.H., Han, G., Fulp, W.J., Giuliano, A.R., 2012. Analysis of overdispersed count data: application to the Human Papillomavirus Infection in Men (HIM) study. Epidemiol. Infect 140, 1087–1094. http://dx.doi.org/10.1017/S095026881100166X.

Lin, H., Zou, H., Wang, Q., Liu, C., Lang, L., Hou, X., et al., 2013. Short-term effect of El Nino-Southern Oscillation on pediatric hand, foot and mouth disease in Shenzhen, China. PLoS One 8, e65585. http://dx.doi.org/10.1371/journal.pone.0065585.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., 2012. The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. Br. J. Math. Stat. Psychol. 65 (1), 163–180.

Lopman, B., Armstrong, B., Atchison, C., Gray, J.J., 2009. Host, weather and virological factors drive norovirus epidemiology: time-series analysis of laboratory surveillance data in England and Wales. PLoS One 4, e6671. http://dx.doi.org/10.1371/journal.pone.0006671.

Luque Fernandez, M.A., Bauernfeind, A., Jimenez, J.D., Gil, C.L., El Omeiri, N., Guibert, D.H., 2009. Influence of temperature and rainfall on the evolution of cholera epidemics in Lusaka, Zambia, 2003–2006: analysis of a time series. Trans. R. Soc. Trop. Med. Hyg. 103, 137–143.

Mangtani, P., Hajat, S., Kovats, S., Wilkinson, P., Armstrong, B., 2006. The Association of respiratory syncytial virus infection and influenza with emergency admissions for respiratory disease in london: an analysis of routine surveillance data. Clin. Infect Dis. 42, 640–646.

O'Hara, R.B., Kotze, D.J., 2010. Do not log-transform count data. Methods Ecol. Evol. 1, 118–122. http://dx.doi.org/10.1111/j.2041-210X.2010.00021.x.

Pascual, M., Dobson, A., 2005. Seasonal patterns of infectious diseases. PLoS Med. 2, e5.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R., 1996. A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49, 1373–1379.

Reiter, P., 2001. Climate change and mosquito-borne disease. Environ. Health Perspect. 109 (Suppl 1), 141–161.

Schwartz, J., Spix, C., Touloumi, G., Bacharova, L., Barumamdzadeh, T., le Tertre, A., et al., 1996. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. J. Epidemiol. Commun. Health 50 (Suppl 1), S3–S11.

Schisterman, E.F., Cole, S.R., Platt, R.W., 2009. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 20, 488–495.
Shaman, J., Pitzer, V.E., Viboud, C., Grenfell, B.T., Lipsitch, M., 2010. Absolute humidity and the seasonal onset of influenza in the continental United States. PLoS Biol. 8, e1000316. http://dx.doi.org/10.1371/journal.pbio.1000316.
Subak, S., 2003. Effects of climate on variability in Lyme disease incidence in the northeastern United States. Am. J. Epidemiol. 157, 531–538.
Sultan, B., Labadi, K., Guegan, J.F., Janicot, S., 2005. Climate drives the meningitis epidemics onset in west Africa. PLoS Med. 2, e6.
Unkel, S., Farrington, C.P., Garthwaite, P.H., Robertson, C., Andrews, N., 2012. Statistical methods for the prospective detection of infectious disease outbreaks: a review. J. R. Stat. Soc. Ser. A Stat. Soc. 175, 49–82. http://dx.doi.org/10.1111/j.1467-985X.2011.00714.x.
Ver Hoef, J.M., Boveng, P.L., 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 88, 2766–2772.
Zeileis, A., Kleiber, C., Jackman, S., 2008. Regression models for count data in R. J. Stat. Softw. 27.
Zhao, X., Chen, F., Feng, Z., Li, X., Zhou, X.-H., 2014. The temporal lagged association between meteorological factors and malaria in 30 counties in south-west China: a multilevel distributed lag non-linear analysis. Malar. J. 13, 57. http://dx.doi.org/10.1186/1475-2875-13-57.