Volume 24 (1), pp. 17–36
http://www.orssa.org.za

ORiON
ISSN 0529-191-X

A comparison of various modelling approaches applied to Cholera case data*

F van den Bergh† JP Holloway ‡ M Pienaar § R Koen‡
CD Elphinstone‡ S Woodborne§

Received: 30 July 2007; Revised: 21 January 2008; Accepted: 3 February 2008

Abstract

The application of a methodology that proposes the use of spectral methods to inform the development of statistical forecasting models for cholera case data is explored in this paper. The seasonal behaviour of the target variable (cholera cases) is analysed using singular spectrum analysis followed by spectrum estimation using the maximum entropy method. This seasonal behaviour is compared to that of environmental variables (rainfall and temperature). The spectral analysis is refined by means of a cross-wavelet technique, which is used to compute lead times for co-varying variables, and suggests transformations that enhance co-varying behaviour. Several statistical modelling techniques, including generalised linear models, ARIMA time series modelling, and dynamic regression are investigated for the purpose of developing a cholera cases forecast model fed by environmental variables. The analyses are demonstrated on data collected from Beira, Mozambique. Dynamic regression was found to be the preferred forecasting method for this data set.

Key words: Cholera, modelling, signal processing, dynamic regression, negative binomial regression, wavelet analysis, cross-wavelet analysis.

1 Introduction

The aim in this paper is to record the experience gained from the modelling of cholera outbreak data recorded in the coastal city of Beira, located in the Sofala province of Mozambique. The objective was to model the number of confirmed cholera cases in relation to certain environmental parameters, and to investigate the feasibility of predicting future

*Funded by the CSIR Strategic Research Panel, project number SRP PP TH/2005/053.

†Corresponding author: Remote Sensing Research Unit, Meraka Institute / CSIR, PO Box 395, Pretoria, 0001, South Africa, email: fvdbergh@csir.co.za

‡Logistics and Quantitative Methods, Built Environment, CSIR, PO Box 395, Pretoria, 0001, South Africa.

§Ecosystem Processes and Dynamics, Natural Resources and the Environment, CSIR, PO Box 395, Pretoria, 0001, South Africa.

cholera outbreaks. Two approaches were used, namely *signal processing methods* (singular spectrum analysis and wavelet analysis) and *statistical methods* (dynamic regression and negative binomial regression). Signal processing methods required fewer assumptions on the data than the statistical methods, which proved to be an advantage when formulating descriptive models. In return, the assumptions required by the statistical methods resulted in the ability to assess results objectively via significance testing and other probabilistic methods, which is an advantage in developing prediction models.

The focus in this paper is on the model fitting component, and not on the results of the wider investigation into cholera in Beira. The methods are only discussed at a high level; readers are referred to the technical references provided for more details. Some background on cholera, the data used and an overview of the wider cholera study is provided in §2. Section 3 provides a brief summary of the different techniques used, while some of the results obtained are documented in §4. The paper closes (in §5) with a brief discussion on the usefulness of the different techniques, as applied to the cholera case data.

2 Background and data

Cholera is a bacterial water-borne disease that occurs frequently in many parts of the world, including Southern Africa. A brief discussion of the disease is presented by the World Health Organisation in [30]; a more in-depth discussion is offered by Sack *et al.* [28]. There have been concerns about the recurrence of epidemics of diseases such as cholera, previously thought to be under control [13, p. 72]. Many scientific studies have been undertaken to study cholera and factors that may contribute to its re-occurrence and spread to new areas. Specifically, linkages between environmental conditions and outbreaks of cholera in Bangladesh have been demonstrated by Huq *et al.* [15]. A study by Gil *et al.* [12] indicated a relationship between cholera incidence and elevated sea surface temperatures in Peru, including effects from the 1997–1998 El Niño, while Pascual *et al.* [25] investigated the relationship between El Niño Southern Oscillation (ENSO) and the occurrence of cholera. It has even been suggested by Lobitz *et al.* [18] that remote sensing of sea surface temperature and height can be used as early warning of conditions associated with cholera. Some studies have also been conducted in Africa, with De Magny *et al.* [9] investigating, at a fairly coarse spatial scale, links between environmental variables and cholera outbreaks in Ghana, and Acosta *et al.* [1] studying possible risk factors and the patterns of outbreaks at a localised scale in a rural village in southern Tanzania. Most of these studies reported on the application of statistical techniques or mathematical signal processing techniques to model cholera data, such as Poisson regression [15], nonlinear time series [25], wavelet analysis [9] and singular spectrum analysis [27].

The aim in this paper is to establish mathematical relationships between the number of cholera cases and certain environmental factors that may support the survival and population growth of the cholera bacteria, *Vibrio Cholerae*, in the natural environment and therefore cause cholera outbreaks. The study specifically excluded public health or socio-economic aspects of cholera outbreaks. It made use of recorded cholera case data in Beira, Mozambique, and captured local environmental parameters. The study also investigated whether observed relationships may be used to develop an early warning

system for cholera outbreaks, although it is noted that results are not necessarily applicable to all high prevalence cholera locations.

Beira is a harbour city situated on the estuary of the Pungwe river, and flooding of parts of the city occurs regularly during the rainy season. The warm Mozambique current causes high sea temperatures and warm subtropical weather with associated high temperatures, rainfall and humidity. These conditions are favourable for cholera outbreaks, and there are records of cholera outbreaks during the period prior to independence in 1975, and of major epidemics that occurred in Beira city in 1992/1993 and 1998. Recently, cholera outbreaks occur in Beira virtually every year.

The Beira cholera case data used in this study represents a count of the number of patients treated for cholera per epidemiological week. This is defined as a week running from a Monday to a Sunday, with the first epidemiological week of the year defined as the week containing the first Sunday in January of the year. The case data stretches from the first epidemiological week in 1999 to week 12 (roughly middle March) in 2005.

Environmental parameters were selected on relevance demonstrated in other studies (for example [15] and [12]), and availability. The weather station situated at the Beira airport records daily air temperature, precipitation and humidity. Data were obtained for the period January 1999 to December 2006, and could be matched to the epidemiological week used for the cholera case counts. Remote sensing data were obtained for sea surface temperature and chlorophyll (algae) growth, but the analysis of the remote sensing data is not further discussed in this paper.

3 Analysis and modelling techniques

The modelling process comprises two phases: During the first phase the data are analysed using descriptive techniques, while during the second phase statistical forecasting models are developed, based on the insights gained from the descriptive analyses.

3.1 Descriptive methods

One of the fundamental tools of signal processing is the ability to transform a time-amplitude representation into a frequency-amplitude representation. Usually one defines a function $x(t)$ by specifying the amplitude of the function at time t. However, the same function can be specified as a function of frequency, represented in the sequel as $\hat{x}(f)$, with $-\infty < f < \infty$. Thus, the function $x(t) = \cos t$ can equally well be defined as $\hat{x}(f) = \sqrt{\frac{\pi}{2}}\delta(f-1) + \sqrt{\frac{\pi}{2}}\delta(f+1)$ in the Fourier frequency domain, where $\delta(t)$ is the Dirac delta function denoting an impulse located at time $t = 0$. The frequency representation $\hat{x}(f)$ is a complex number in order to accommodate phase information.

A popular way to gain insight into the dominant frequencies of a signal is to examine its *Power Spectral Density* (PSD) plot. The PSD of a signal represents the *power* that is present at each frequency. The PSD is therefore closely related to the frequency representation of a signal. The one-sided PSD [26, p. 503] is defined as $P_x(f) \equiv 2|\hat{x}(f)|^2$ for real-valued functions $x(t)$.

Various methods may be used to compute estimates of the PSD of a signal; the most widely used method is the *Fourier Transform* (FT), but other methods, such as the *Maximum Entropy Method* (MEM) [26, pp. 577–580] often produce plots that are easier to interpret. The FT of a signal may be computed directly as

$$\hat{x}(f) = \int_{-\infty}^{\infty} x(t) e^{2\pi i f t} \, \mathrm{d}t \tag{1}$$

where $i = \sqrt{-1}$, and f is the frequency at which the transform is computed. In practice, an efficient implementation of the *Discrete Fourier Transform* (DFT), known as the *Fast Fourier Transform* (FFT) [8], is used to compute the FT of a sequence of discrete data points. The switch from a continuous-time signal to discrete-time sampled signal has many ramifications; the reader is referred to Press *et al.* [26, pp. 506–509] for a brief overview of these issues.

The Fourier domain representation of a signal is formed by the superposition of sine and cosine functions of various frequencies. These functions are non-zero over the entire time domain. Therefore the FT provides only *global* frequency information; this topic will be revisited later.

3.1.1 Singular spectrum analysis

Singular Spectrum Analysis (SSA) has been used widely to analyse climatic time series data, including the study of paleoclimatic time series, inter-decadal climate variability analysis, and the analysis of inter-annual and intra-seasonal oscillations. Additional examples and an in-depth discussion of the SSA method is presented by Ghil *et al.* [11].

SSA decomposes a time series into additive components, also referred to as *empirical orthogonal functions*, which form a basis derived directly from the data. The benefits of this approach include that often only a few of these components are required to reconstruct the signal, and that these empirical basis functions may assume non-sinusoidal shapes. These components are classified into three classes: *trend* (slowly varying), *oscillatory* (possibly amplitude-modulated), and *noise* components [14]. It is important to note that in the SSA literature, the term *noise* is used to refer to both stochastic noise, as well as subjective "uninteresting" components of the signal.

The SSA algorithm comprises three phases: decomposition, component selection, and reconstruction. The first phase of the SSA algorithm deals with the decomposition of the time series into its constituent components. This involves the construction of a *trajectory matrix*, which is built by "stacking" lagged copies of the time series into a matrix. The principal components of the covariance matrix are calculated from the trajectory matrix, resulting in an eigenvalue-eigenvector pair for each component. The magnitudes of these eigenvalues correspond to their contribution to the total observed variance in the original time series, — therefore larger eigenvalues are naturally associated with dominant components. If the time series was generated by a system with uncorrelated process noise as well as uncorrelated observation noise, the ordered eigenvalues tend to exhibit a noticeable drop-off in magnitude after a certain point; this usually represents the transition from *trend*/*oscillatory* components to *noise* components. For more complex systems, which

may involve noise generated by autoregressive processes, alternative algorithms such as *Monte-Carlo SSA* (MC-SSA) have been developed to aid with the identification of the noise components [3].

During the third phase, after the important components have been identified, an approximation of the original signal is reconstructed by summing the contributions of each of the selected components. The components previously labeled as noise are thus omitted from this reconstruction, yielding an improved signal-to-noise ratio.

3.1.2 Wavelet analysis

The FT of a signal yields perfect frequency localisation, discarding all time information, so that it is not known when a particular frequency was present. To address this one can apply the FT to short overlapping segments of the signal, instead of the whole signal, resulting in a technique is known as the *windowed Fourier transform*. This has significant drawbacks [29]; a better way of improving the time-localisation of a transform is to use a set of localised basis functions that are (effectively) non-zero only over a finite range. This can be achieved by scaling the harmonic function used to perform the analysis with a compact window function. If this approach is taken to its logical conclusion the result is an optimal multi-resolution analysis method known as the *wavelet transform* [20].

Wavelets allow for the transformation of a time-domain signal to a joint time/frequency representation which preserves information on both the power of a specific frequency in a signal, as well as the time at which this frequency was present. This is achieved by decomposing the original signal at various *scales*; these scales are conceptually similar to the frequencies of the FT, and it is possible to convert between these two representations. For a time series of n discrete points, the wavelet decomposition at scale a yields n transformed values. If the composition is carried out over multiple scales, the result is a two-dimensional set of transformed values, with one axis representing time, and the other axis representing scale (or frequency).

Wavelet analysis employs a wavelet function, such as the *Morlet wavelet*,

$$\Psi_0(t) = \pi^{-1/4} e^{i\omega_0 t} e^{-t^2/2}, \tag{2}$$

where ω_0 is a dimensionless frequency parameter, taken as $\omega_0 = 6$ here to satisfy the admissibility condition [10]. The Morlet wavelet is located at time $t = 0$ and is of scale $a = 1$; this "default" version of the wavelet function is popularly called the "mother wavelet." The wavelet must be translated by a distance b to analyse the signal at time b. In a similar way, the wavelet must be dilated (scaled in time) by a factor a, to analyse properties of the signal at scale a. Incorporating both the translation and scaling into a single step, the *Continuous Wavelet Transform* (CWT) $W_x(b, a)$ at scale a and location b is

$$W_x(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \Psi^* \left(\frac{t - b}{a} \right) \mathrm{d}t, \tag{3}$$

where Ψ is the mother wavelet (2) and * denotes complex conjugation. The subscript 0 is omitted to indicate that Ψ is normalised [29]. The wavelet transform can be applied to discrete time series data comprising n points by using the discrete summation equivalent of

(3) [29]. The convolution at all locations $1 \leq b \leq n$ may be implemented as multiplication in the Fourier domain, so in practice the CWT can be implemented efficiently using the FFT [29].

Peaks in the wavelet power spectrum may be compared to a mean power spectrum (for example, the mean power spectrum of an AR(1)-process) to determine whether they are significantly above the mean spectrum; these hypotheses can be tested statistically at a specified confidence level [21, 29]. Another operation that can be performed on the wavelet transform is to convert the transformed values into a power spectrum, which may be plotted as a contour plot of period *vs.* scale (frequency) to ease the interpretion of the time-frequency content.

The *cross-wavelet spectrum* [29] reinforces the co-varying behaviour found in the power spectrum of independent variables, according to scale, thereby highlighting coherency between two variables. The phase of such coherency is provided by the argument of the cross-wavelet spectrum, from which it is possible to formulate phase synchronisation between the two variables. The cross-wavelet spectrum thus helps to identify which variables contribute the most to the response variable in terms of coherency and phase. The cross-wavelet spectrum between signals $x(t)$ and $y(t)$ is computed as

$$W_{xy}(b, a) = W_x(b, a)W_y^*(b, a). \tag{4}$$

From the cross-wavelet spectrum, the *cross-wavelet power* can be calculated as $|W_{xy}(a, b)|$.

3.2 Statistical methods

Time series analysis techniques were considered to be applicable to the cholera case data, with the environmental parameters (*e.g.,* temperature, rainfall) treated as explanatory variables. In addition, the use of generalised linear models on these data was investigated.

3.2.1 ARIMA models

ARIMA models, developed by Box and Jenkins [5], are a subset of time series analysis techniques that may be used to forecast future values of a time series based on historical values of the time series. ARIMA models can accommodate seasonality, with Makridakis *et al.* [19] giving several seasonal ARIMA examples, and can also handle local seasonality [6] (*i.e.* data that are more related to the same season one or two years previously than the same season several years ago). Furthermore, ARIMA models may be used to analyse time-dependent data (*i.e.* autocorrelation of the series).

A dynamic regression model, a term applied by Pankratz [24] and used by Makridakis [19], uses explanatory variables to forecast the dependent variable, but it still allows one to include the elements of ARIMA to model any patterns that cannot be accounted for by the explanatory variables. They differ from multivariate autoregressive models [19] in that the explanatory variables are leading indicators and are not affected by the dependent variable.

A dynamic regression model for one explanatory variable X can be written in two general

forms, as described in [19], but in the simpler form the forecast variable Y_t takes the form

$$Y_t = a + \frac{\omega(B)}{\delta(B)} X_{t-b} + N_t, \tag{5}$$

where X_t is the explanatory variable, where

$$\omega(B) = \omega_0 - \sum_{i=1}^{s} \omega_i B^i \text{ and } \delta(B) = 1 - \sum_{j=1}^{r} \delta_j B^j$$

in terms of the backward shift operator (*e.g.* $BY_t = Y_{t-1}$) and where N_t is the combined effects of all other factors (*i.e.* noise, modelled as an ARIMA process).

This formula extends naturally to several explanatory variables. In order to calibrate the model for one explanatory variable X, it is necessary to determine the values of r, s and b, as well as the values of p, d and q for the ARIMA(p, d, q) model for N_t. There are various methods for doing this. The method used in this study was suggested by Pankratz [24] and Makridakis [19] and is referred to as the *Linear Transfer Function* (LTF) identification method.

The Box-Jenkins approach to ARIMA modelling consists of three phases, namely: identification, estimation and testing; the dynamic regression modelling approach (as applied to one explanatory variable), given by Makridakis *et al.* [19], is summarised in six steps:

1. Fit a regression model with lagged explanatory variables, using a low-order proxy AR, such as an AR(1), to model the noise component;
2. Test the regression errors for non-stationarity and difference the data if necessary;
3. Establish the values for b, r and s of the explanatory variable X (see Makridakis [19] for guidelines);
4. Identify the relevant ARMA model for the regression errors;
5. Calibrate the final parameter estimates by refitting the new ARMA model for the errors, N_t, and the transfer function model for the explanatory variable X; and
6. Perform diagnostic testing on the final model and omit any unnecessary parameters.

The advantage of using dynamic regression is the ability to regress the cholera data on the environmental variables and then fit an ARMA model to account for residual variability. The strong autocorrelation and seasonality in these data indicate that both the univariate ARIMA models and dynamic regression approach may be useful techniques for forecasting future cholera cases. ARIMA models do not recognise the cholera cases as count data with a minimum of zero; they erroneously allow negative values to be forecast.

Univariate ARIMA models are not appropriate for forecasting far into the future [4, p. 343] but can be very powerful for short forecast horizons. In mitigating cholera outbreaks the objective is typically to forecast only a few weeks ahead and hence univariate forecasts may be useful. The objective of our study, however, was to find environmental factors that may potentially signal the outbreak. Models such as dynamic regression that use explanatory variables to model the forecast variable, while still modelling the autocorrelation in the error terms, are expected to be more appropriate. Such models require forecasts of these environmental drivers for prediction purposes unless there is sufficient lag between the leading environmental indicators and the forecast variable of cholera case counts.

3.2.2 Generalised linear models

GLMs allow the dependent variable to follow a distribution from the family of exponential distributions, which includes the normal, Poisson, binomial, exponential and gamma distributions [23, p. 160, 427]. A GLM is a generalisation of the classical linear models, such as linear regression, since the normal distribution belongs to the class of exponential distributions [22].

Taking into account that count data, like the cholera case data, are always non-negative and often positively skewed (many occurrences of small counts, with only a few large counts), the option of fitting a GLM to the data was investigated. Figure 1 shows the extent of skewness observed in the data.

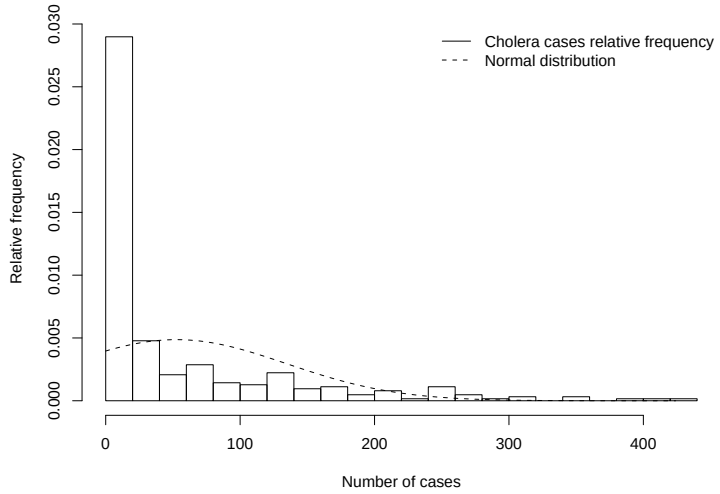


Figure 1: *Histogram of cholera cases.*

The choice of distribution to fit to the dependent variable is important. For counted data not in the form of proportions, the Poisson distribution may be appropriate [22, p. 127]. In a Poisson distribution the variance is expected to be equal to the mean; Byers *et al.* [7] suggest that if the variance is much larger than the mean, a negative binomial distribution may be better suited to the data. The Beira cholera case data exhibit a variance of almost 10 times the mean, thus the negative binomial distribution was deemed more appropriate. Kotz *et al.* [16, Vol 6] provide further comparisons of the negative binomial distribution to other distributions.

Poisson regression has previously been used to model cholera case data in the Bangladesh study of Huq *et al.* [15]. Further details on Poisson regression may be found in McCullagh & Nelder [22], Agresti [2] or Montgomery *et al.* [23]. Negative binomial regression is applied similarly, using a logarithmic link function, with the negative binomial replacing the Poisson distribution. The deviance value may be used to confirm model fit.

4 Results

The process that was followed to explore the data with various signal processing techniques, eventually leading up to the successful modelling of the Beira cholera data using statistical models, is described in this section.

Two main environmental variables used in this study, namely temperature and rainfall, are shown in Figure 2. A plot of the target variable, the number of cholera cases per week, is presented in Figure 4 as the dashed line.

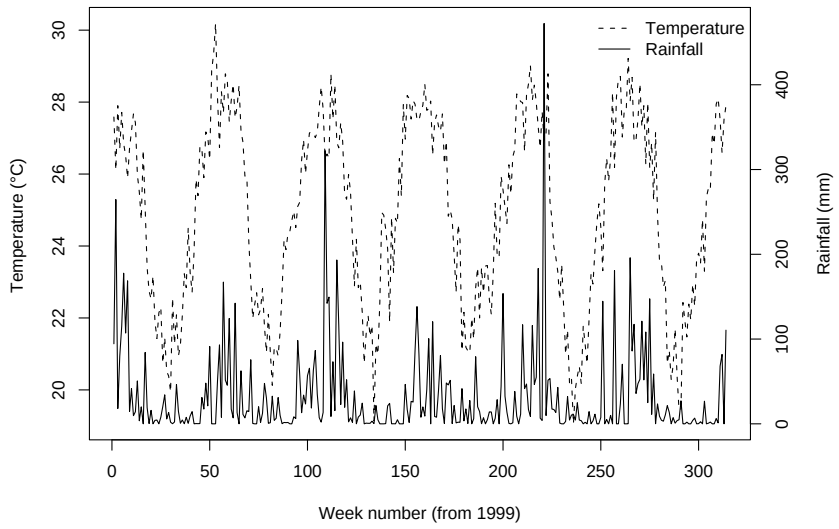
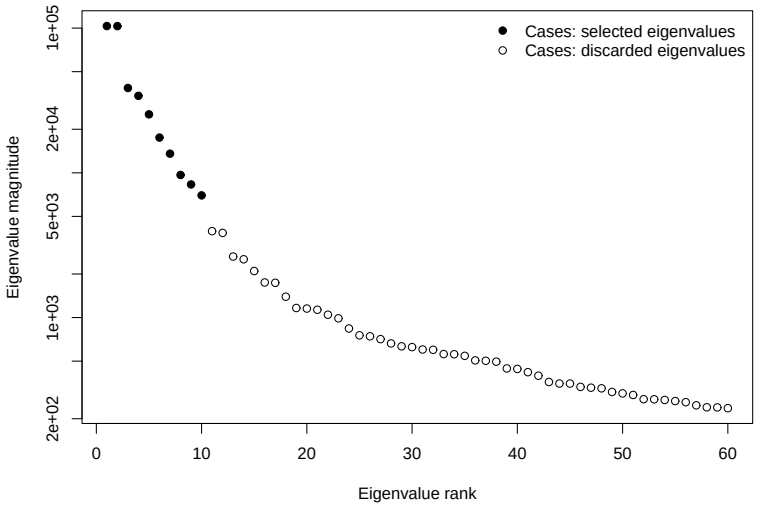


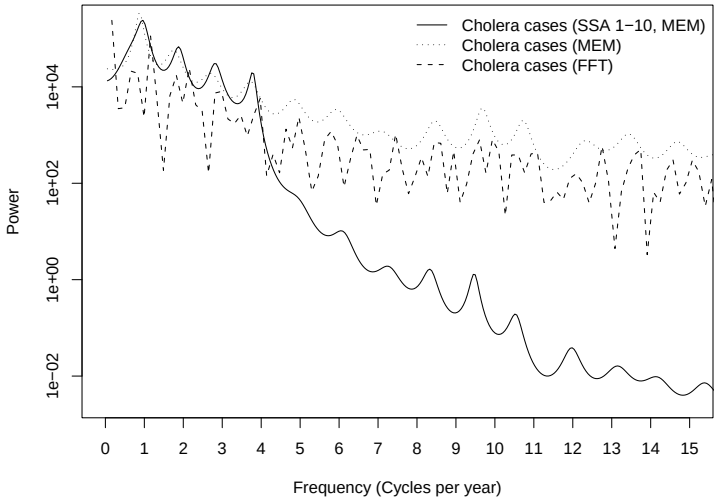
Figure 2: *Temperature and rainfall variables used in the study.*

As a first step towards understanding the data set, a PSD plot of the cholera cases time series was computed to identify the dominant frequencies, using both a standard FFT algorithm and the MEM. The results are presented in Figure 3(b); note how the MEM spectrum estimate (dotted line) is easier to interpret than the FFT spectrum (dashed line). The power spectrum computed using the MEM algorithm can be further improved by enhancing the signal-to-noise ratio; this can be achieved by computing SSA reconstructions of the original series. The SSA algorithm was applied to the cholera case data using a window size of 60, and sixteen principal components were extracted. Figure 3(a) is a plot of the magnitude of the eigenvalues, sorted in descending order.

The solid curve in Figure 3(b) is the power spectrum of the reconstructed cholera case data, using SSA component 1 through 10. This curve drops sharply after frequencies greater than four cycles per year, while the spectrum of the original time series (dashed curve) decays more gradually. The peaks in the lower frequencies, below four cycles per year, appear somewhat sharper in the power spectrum of the reconstructed time series, but there appears to be a small shift in the location of these peaks. Figure 3(b) thus illustrates the benefit of applying SSA to a time series before the application of MEM to



(a) Eigenvalue plot



(b) Power spectrum

Figure 3: *Eigenvalue plot and Power spectrum of cholera case data. The eigenvalues of part (a) were obtained using SSA. The power spectrum of part (b) was obtained by applying MEM to an SSA reconstruction of the cholera case data. The MEM power spectrum of the original cholera case data is provided for reference.*

extract the power spectrum: the significant peaks in the power spectrum are more clearly visible above the background noise.

A reconstruction of the cholera case data using SSA components 1 through to 10 is plotted

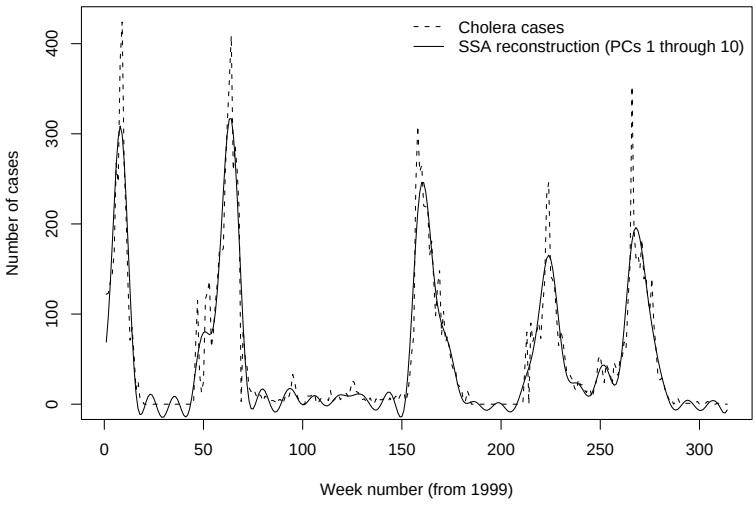


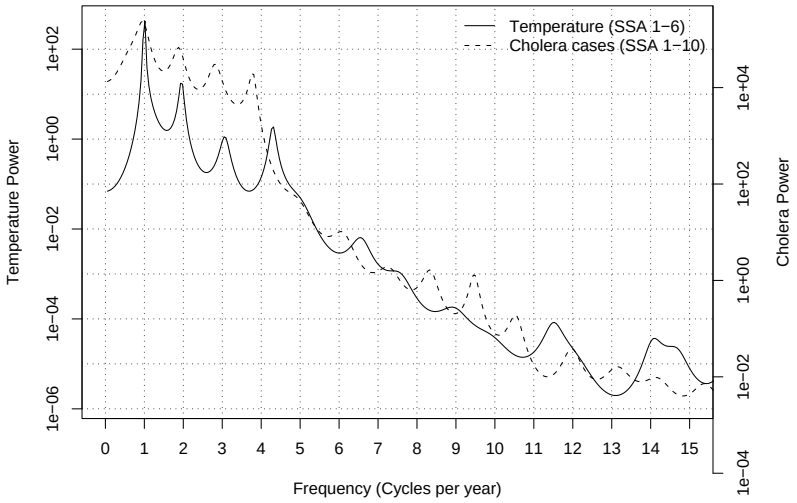
Figure 4: *Cholera cases, reconstructed using SSA.*

alongside the original case data in Figure 4. Some spurious oscillations are visible in the regions where the original case data is near zero. These oscillations are due to aliased frequencies that were caused by the half-wave rectified nature of the cholera case data. These oscillations will be cancelled by higher-frequency harmonics and will gradually disappear if more SSA terms are included in the reconstruction.

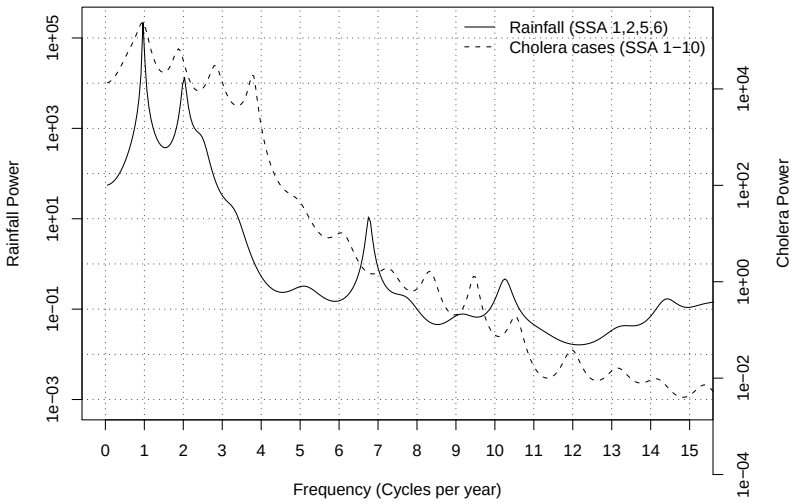
In Figure 5 the MEM power spectrum estimates of the reconstructed cholera case data are compared to the spectra of environmental variables, reconstructed with SSA as indicated. The most salient feature of all these power spectra is that they all have noticeable power at frequencies of approximately one cycle and two cycles per year. What cannot be inferred from the power spectra is the phase differences between the variables, which would indicate lead or lag times.

The meteorological time series data and cholera case data were assessed using cross-wavelet analysis to extract phase information. A Morlet wavelet with a central frequency of 6 was chosen as the analysing wavelet function based on comparable features in the data in the sub-52 period scales. Figure 6 shows the CWT of the cholera case data. Each of the scales a within the CWT is directly inter-comparable because of the normalisation property; furthermore, a scale of a can be compared to a Fourier period of approximately $1.03a$. Caution should be exercised when assessing the sub-52 week periods portrayed in Figure 6. For example, the 26 week period observed in the CWT of Cholera cases is probably an artifact of the half-wave rectified nature of the data, which is similar to the artifacts observed during the SSA analysis.

The cross-wavelet spectra of cholera cases and environmental parameters are presented in Figure 7. The empirical time response between cholera cases and the environmental drivers can be derived from the phase offset calculated in regions of the cross-wavelet spectrum where there is significant power; the phase offset is given by the argument of the



(a) Temperature



(b) Rainfall

Figure 5: *Comparison of power spectra of cholera cases, and two environmental inputs, obtained using the Maximum Entropy Method (MEM). Note that the cholera spectrum power is plotted on its own scale, so absolute power is not comparable in these plots.*

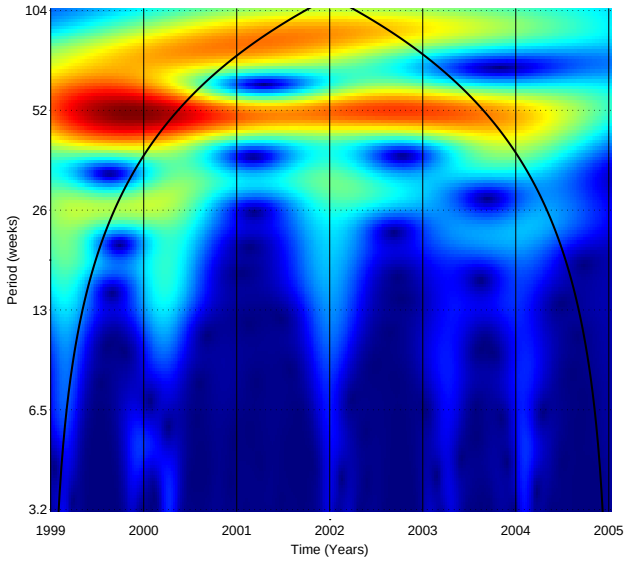


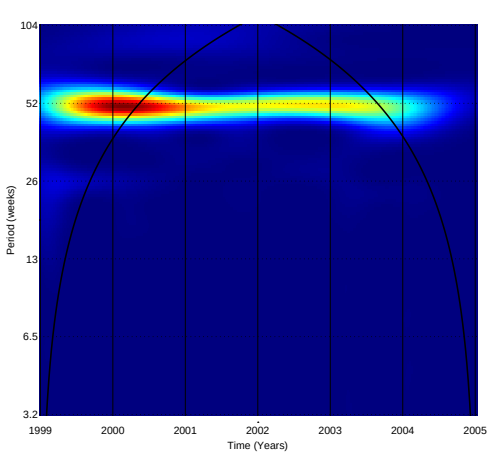
Figure 6: *The power spectrum of the CWT of the cholera case data allows variations to be observed according to scale (period). A nonorthogonal transform is used to produce smooth continuous variations in power. A cone of influence is applied to the power spectrum due to the errors that occur at the beginning and end of the transform. Note in particular the power visible in the 52-week period, visible as the darker horizontal band within the lighter region.*

cross-wavelet transform in those regions.

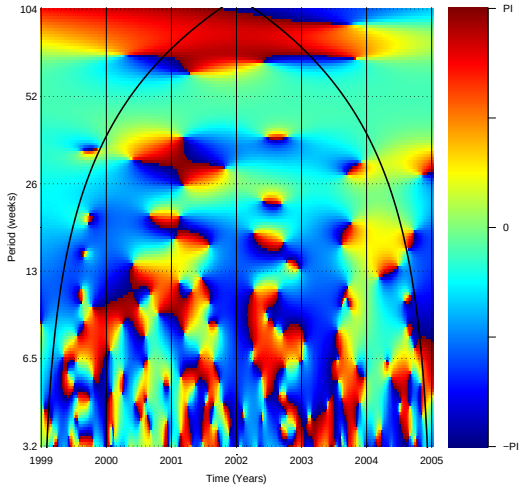
The cross-wavelet spectra between cholera cases and environmental parameters are coherent at a 52-week scale. This signal was isolated from the sub-52 week periods for the phase analysis. The results for humidity (not shown here) suggested that cholera cases lead humidity, which is causally untenable, and so humidity was excluded from further analysis. The phase offset (lead) between air temperature and cholera cases has a frequency modulated response of between 10 and 20 weeks, with an average of 13 weeks, while the phase offset for rainfall varies between approximately 4 and 14 weeks, with an average of 7 weeks. Considering all the scales simultaneously results in a phase offset of between 6 and 9 weeks for air temperature.

This approach suggests that cholera cases may be responding to changes in rainfall and temperature. Using the average lead times determined in the phase analysis of the cross-wavelet spectrum, the cholera cases were assessed against the respective variables. An exponential response was determined for temperature, and a linear response was determined for rainfall. It was assumed that cumulative rainfall would approximate an environmental threshold exceedance that would mimic the lead times returned by the cross-wavelet analysis. A set of cumulative rainfall series was created (with the implicit lead times) and was subjected to the same wavelet and cross-wavelet analysis, with an 8 week accumulation yielding best results.

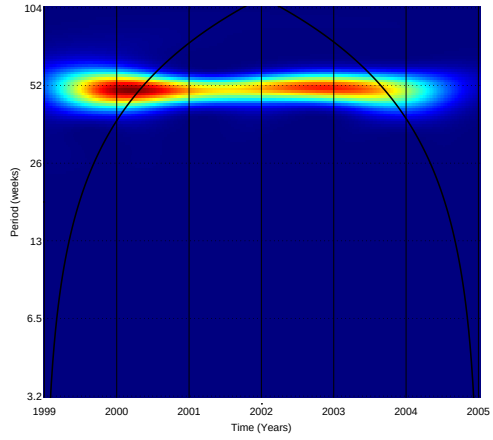
A similar study was conducted on the coherence between temperature and cholera cases, the result of which suggested lead periods of 6 to 8 weeks, depending on the scale con-



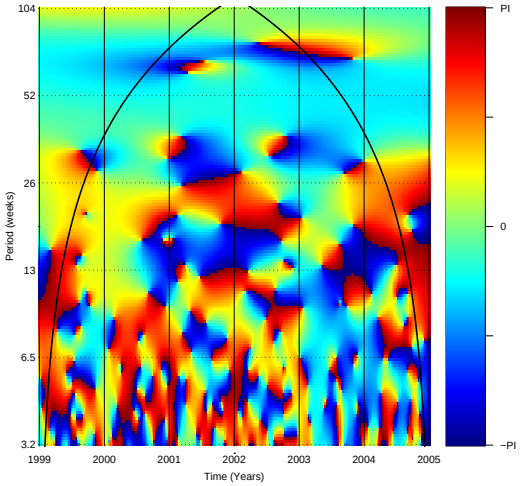
(a) Cholera cases / Rain cross magnitude



(b) Cholera cases / Rain cross phase



(c) Cholera cases / Temperature cross magnitude



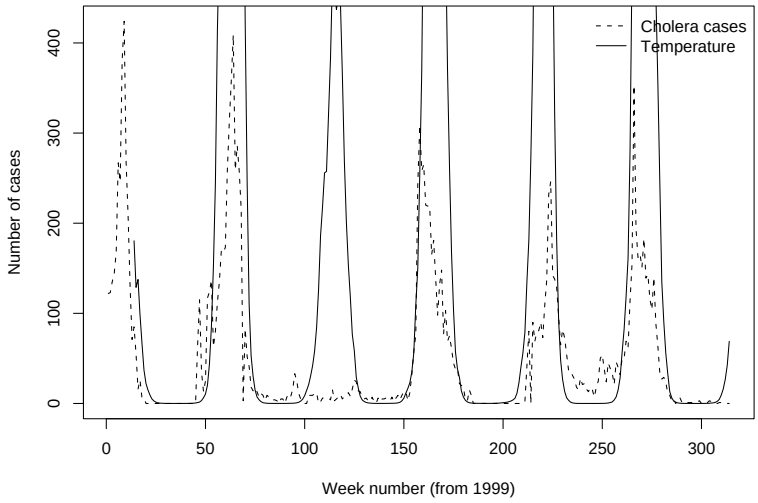
(d) Cholera cases / Temperature cross phase

Figure 7: *Cross-wavelet magnitude* ($|W_{xy}(b,a)|$) *and phase* ($\arg[W_{xy}(b,a)]$).

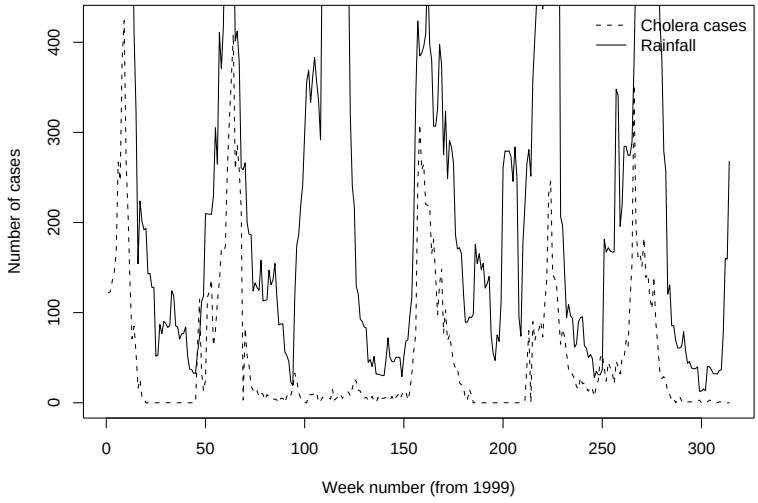
sidered. It was found that the accumulation of temperature smoothed the signal into a sinusoidal shape, which effectively filtered out all information at scales below 52 weeks. Despite this loss of detail, it was found that an accumulation of 14 weeks yielded improved phase difference estimates.

The relationship between cumulative temperature and cholera cases was again found to be exponential, while the response to cumulative rainfall remained linear. The exponential of the 14 week temperature accumulation is presented with the cholera case data in Figure 8(a), along with the 8-week cumulative rainfall in Figure 8(b).

Both SSA and wavelet analysis highlight the seasonality in the Beira cholera data; in



(a) Temperature transform and cholera cases



(b) Cumulative rainfall and cholera cases

Figure 8: *Exponentially transformed cumulative temperature values, and cumulative rainfall plotted against cholera cases. Note how the sudden increases in exponential temperature values closely match the start of cholera outbreaks, and how cumulative rainfall has a similar (but less pronounced) relationship to cholera cases.*

addition, the wavelet analysis indicate relationships between cholera cases and lagged values of environmental variables. This suggests that observed environmental measures may be used to forecast cholera cases a few weeks ahead using statistical forecast models.

The SAS software package was used to carry out all statistical analyses reported in this section. As part of the statistical modelling process, data transformations were considered. No data transformations were required for the negative binomial regression, since the GLM method involved applying a logarithmic link function to the cholera cases implicitly. For ARIMA modelling, data transformations to ensure stationarity (seasonal differencing) were required before application of the model. The data showed strong autocorrelation in the cholera case data, both from time series diagnostic model tests and from applying the Durbin-Watson test for autocorrelation. Results from the wavelet analysis indicated that either exponential or cumulative temperature values and cumulative rainfall values showed stronger relationships with cholera patterns than temperature or rainfall itself, and this finding informed the statistical modelling. The wavelet analysis suggested the use of cumulative rainfall rather than the weekly rainfall, and different cumulation periods were tested. Similarly, the effect of using an exponential transformation of temperature or cumulative temperature, instead of temperature values as measured, was investigated. The outcomes resulting from varying the lags of the variables and derived variables were also considered.

An ARIMA(1,0,0)(0,1,1) model was found to fit the cholera data well, but the inclusion of lagged environmental variables in a dynamic regression model improved the fit. After testing different environmental variables at different lags and carrying out the various diagnostic texts for model adequacy, a number of plausible dynamic regression models remained. To identify the simplest model that fitted the data well, Akaike's Information Criterion (AIC) was calculated. The AIC balances accuracy and complexity by penalising the model based on the number of parameters included in the model [4]. The model with the lowest AIC value is usually chosen as the best fitting model, but in this case the best fitting model included short lags of the environmental variables (lags 0 and 1 of the cumulative rainfall and lag 0 of exponential temperature) which was not conducive to forecasting a few weeks ahead. As a result, a small sacrifice in model fit, as measured by the AIC, was accepted to obtain a dynamic regression model that was more parsimonious and showed better predictive ability. This model contained an autocorrelation component in combination with lags 5, 6 and 7 of exponential air temperature. This model, and all other dynamic regression models fitted to the cholera data, exhibit a lower AIC than the univariate ARIMA model.

In the negative binomial regression study different combinations of variables, in terms of accumulation periods and lags (ignoring lags 0 and 1) were modelled. A model containing rainfall accumulated over 2 weeks and lagged at 6 weeks resulted in the best log-likelihood value. Residual analysis showed that peaks in rainfall translated into cholera case predictions that were too high. Adding in temperature at lag 6 resulted in slightly worse log-likelihood values, but stabilised the predictions, and was therefore considered a better model. This model with both cumulative rainfall and temperature was also slightly better than a model with only temperature at lag 7 and was therefore the preferred negative binomial model. The deviance value of the negative binomial regression model was com-

pared to that of a Poisson regression model, confirming that the negative binomial model was a better fit.

Model	**MSE**	**MAE**
Univariate — ARIMA	1642.9	22.34
Dynamic regression: Exp(temperature) at lags 0,5,6,7,8 + 5 week cumulative rainfall at lags 0,1,4,5 + AR(1)	748.4	14.90
Dynamic regression: Exp(temperature) at lags 5,6 + 2 week cumulative rainfall at lag 4 + AR(1)	847.2	14.53
Dynamic regression: Exp(temperature) at lags 5,6,7 + AR(1)	858.9	14.10
Negative binomial: 2 week cumulative rainfall at lag 6	17152.0	60.28
Negative binomial: Temperature at lag 7	4077.1	37.87
Negative binomial: 2 week cumulative rainfall at lag 6 and temperature at lag 6	3873.3	37.00

Table 1: *Comparative goodness of fit statistics for models (lower values are better).*

In order to compare the fit of the time series models with that of the negative binomial regression, residuals were calculated from which Mean Square Error (MSE) and Mean Absolute Error (MAE) measures were derived. The results are given in Table 1, and show clearly that the dynamic regression gave the best overall fit. Although the Mean Absolute Percentage Error (MAPE) measure is recommended for comparing forecast results between different models, [4, p. 460], the large number of zero cases (actuals) recorded per week made a percentage error difficult to calculate and the MAE was used instead.

The dynamic regression model also provided a visually better fit than the negative binomial regression, as indicated in Figures 9 and 10.

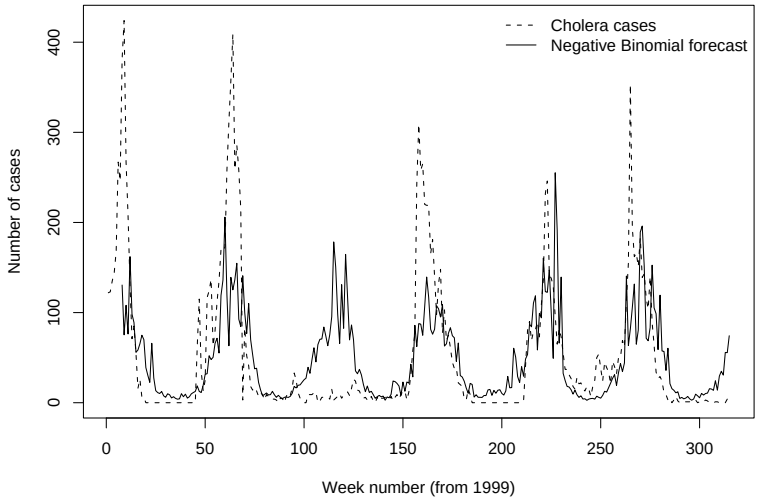


Figure 9: *Model fit of negative binomial model.*

A reasonable model could be developed using negative binomial regression, but this model did not perform as well as dynamic regression. Although a GLM approach has been re-

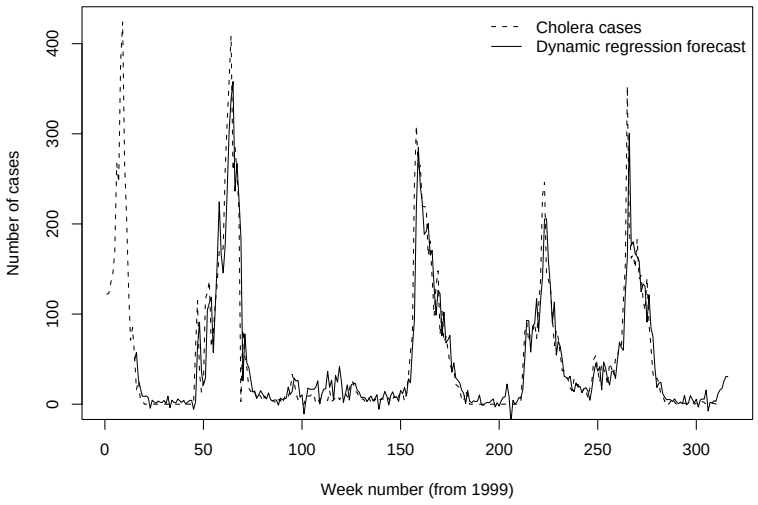


Figure 10: *Model fit using dynamic regression. Note that the results from the model Exp(temperature) at lags 5,6,7 + AR(1) are illustrated.*

ported to be successful for similar data [15], the presence of autocorrelation in the Beira data set is a matter of concern, since GLM requires independent data [22, p. 15]. In terms of underlying assumptions as well as goodness of fit, the dynamic regression model was therefore the preferred method for developing forecast models. The best dynamic regression model required an autoregressive component, which implies that the environmental factors could not adequately explain all the variations in the cholera data.

5 Discussion

Signal processing methods can be used to summarise observed data points accurately, while statistical methods are concerned with determining a model that is consistent with the observed data points, allowing for a random residual effect. This implies that the signal processing methods require fewer assumptions on the data than the statistical methods, with issues such as normality, data transformations and non-negativity of count data not playing an important role. The flexibility of signal processing methods can be an advantage in exploring the data and providing descriptive models. The more stringent assumptions required by the statistical methods results in the ability to assess results objectively via significance testing and other probabilistic methods. Wavelet analysis and the more traditional statistical methods have been used in this manner throughout this study, allowing the signal processing results to influence the choice of explanatory variables in the statistical models. This approach is similar to the one used by Krankowski *et al.* [17] where wavelet analysis was used to determine time delays that were subsequently included in a time series forecasting model.

In terms of stationarity requirements for time series data, the most stringent requirements

are placed by the statistical time series methods. Trends must be estimated and removed, and if there are signs of heteroscedasticity (increasing variance), transformations have to be applied. If such non-stationarities are not removed, other effects may be masked and missed by the model diagnostics. For wavelet analysis there is no requirement of stationarity, and SSA can work with data containing trends provided there is no heteroscedasticity. The statistical GLM methods, however, are designed to deal with heteroscedasticity and trends in data and therefore do not require prior transformations to improve stationarity.

For the final purpose of this study it was also required that some of the techniques be applied to multivariate data. SSA in its original form, as used in this study, is not designed to deal with multivariate data, although multivariate extensions exist. While wavelet analysis typically deals with one variable at a time, cross-wavelet analysis can compare the patterns of two different variables. If comparison is required between more than two variables, successive cross-wavelet analyses comparing each variable with one another is required. While ARIMA models do not allow the comparison of cholera cases with other variables, the strength of both dynamic regression and GLM methods is that it can model the relationship between cholera and other variables (allowing for more than one lag of each variable), while also taking into account the relationships between the variables amongst each other, thereby having the advantage that interactions between all the variables and lags can be modelled well.

The dynamic regression model yielded the best forecasting results for the cholera cases in Beira. The process of using SSA to explore seasonality, and cross-wavelet analysis to explore the coherence and local phase differences of the seasonal components, proved to be very valuable in the design of the final model. During the exploration phase using wavelet analysis, a useful transformation of the data was discovered that contributed to the success of the dynamic regression model forecasts. In our experience the signal processing methods added significant value to the modelling process.

6 Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive feedback.

References

[1] Acosta CJ, Galindo CM, Kimario J, Senkoro K, Urassa H, Casals C, Corachán M, Eseko N, Tanner M, Mshinda H, Lwilla F, Vila J & Alonso PL, 2001, *Cholera outbreak in southern Tanzania: Risk factors and patterns of transmission*, Emerging Infectious Diseases, **7(3)**, [Online], [Cited August 30th, 2006], Available from `http://www.cdc.gov/ncidod/eid/vol7no3\_supp/acosta.htm`.

[2] Agresti A, 1996, *An introduction to categorical data analysis*, John Wiley and Sons, New York (NY).

[3] Allen MR & Smith LA, 1996, *Monte Carlo SSA: Detecting irregular oscillations in the presence of colored noise*, Journal of Climate, **9(12)**, pp. 3373–3404.

[4] Armstrong JS, 2001, *Principles of forecasting*, Kluwer Academic Publishers, Dordrecht.

[5] Box GEP & Jenkins GM, 1970, *Time series analysis: Forecasting and control*, Holden-Day, Inc., San Francisco (CA).

[6] Brocklebank JC & Dickey DA, 2003, *SAS for forecasting time series*, 2nd Edition, John Wiley, Hoboken (NJ).

[7] Byers AL, Allore H, Gill TM & Peduzzi PN, 2003, *Application of negative binomial modeling for discrete outcomes: A case study in aging research*, Journal of Clinical Epidemiology, **56**, pp. 559–564.

[8] Cooley JW & Tukey JW, 1965, *An algorithm for the machine calculation of complex Fourier series*, Mathematics of Computation, **19(90)**, pp. 297–301.

[9] de Magny CC, Cazelles B & Guegan JF, 2006, *Cholera threat to humans in Ghana is influenced by both global and regional climatic variability*, EcoHealth, **3(4)**, pp. 223–231.

[10] Farge M, 1992, *Wavelet transforms and their applications to turbulence*, Annual Review of Fluid Mechanics, **24(1)**, pp. 395–457.

[11] Ghil M, Allen MR, Dettinger MD, Ide K, Kondrashov D, Mann ME, Robertson AW, Saunders A, Tian Y, Varadi F & Yiou P, 2002, *Advanced spectral methods for climatic time series*, Reviews of Geophysics, **40(1)**, p. 1003.

[12] Gil AI, Louis VR, Rivera ING *et al.*, 2004, *Occurrence and distribution of Vibrio cholerae in the coastal environment of Peru*, Environmental Microbiology, **6(7)**, pp. 699–706.

[13] Goldstein BD, 2005, *Emerging and re-emerging infectious diseases: Links to environmental change, Geo Year Book 2004–2005: An overview of our changing environment*, United Nations Publications, New York (NY).

[14] Golyandina NE, Nekrutkin V & Zhigkijavsky AA, 2001, *Analysis of time series structure: SSA and related techniques*, CRC Press, London.

[15] Huq A, Sack RB *et al.*, 2005, *Critical factors influencing the occurrence of Vibrio cholerae in the environment of Bangladesh*, Applied and Environmental Microbiology, **71(8)**, pp. 4645–4654.

[16] Kotz S, Johnson N L & Read C B, 1985, *Encyclopedia of statistical sciences*, John Wiley and Sons, New York (NY).

[17] Krankowski A, Kosek W, Baran LW & Popinski W, 2005, *Wavelet analysis and forecasting of VTEC obtained with GPS observations over European latitudes*, Journal of Atmospheric and Solar-terrestrial Physics, **67**, pp. 1147–1156.

[18] Lobitz B, Beck L, Huq A *et al.*, 2004, *Climate and infectious disease: Use of remote sensing for detecting of Vibrio cholerae by indirect measurement*, Environmental Microbiology, **6(7)**, pp. 699–706.

[19] Makridakis S, Wheelwright SC & Hyndman RJ, 1998, *Forecasting methods and applications*, 3rd Edition, John Wiley, New York (NY).

[20] Mallat SG, 1989, *A theory for multiresolution signal decomposition: The wavelet representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **11(7)**, pp. 674–693.

[21] Maraun D, 2006, *What can we learn from climate data? Methods for fluctuation, time/scale and phase analysis*, PhD dissertation, Universität Potsdam, Potsdam.

[22] McCullagh P & Nelder JA, 1983, *Generalized linear models*, Chapman and Hall, New York (NY).

[23] Montgomery DC, Peck EA & Vining GG, 2006, *Introduction to linear regression analysis*, John Wiley and Sons, New York (NY).

[24] Pankratz A, 1991, *Forecasting with dynamic regression models*, John Wiley, New York (NY).

[25] Pascual M, Rodo X, Ellner S P, Colwell R & Bouma MJ, 2000, *Cholera dynamics and El Niño-southern oscillation*, Science, **289(5485)**, pp. 1766–1769.

[26] Press WH, Teukolsky SA, Vetterling WT & Flannery BP, 2002, *Numerical recipes in C++: The art of scientific computing*, Cambridge University Press, Cambridge.

[27] Rodo X, Pascual M, Fuchs G & Faruque ASG, 2002, *ENSO and Cholera: A nonstationary link related to climate change?*, Proceedings of the National Academy of Sciences, **99(20)**, pp. 12901–12906.

[28] Sack D, Sack RB, Nair GB & Siddique AK, 2004, *Cholera (seminar)*, Lancet, **363**, pp. 223–233.

[29] Torrence C & Compo GP, 1998, *A practical guide to wavelet analysis*, Bulletin of the American Meteorological Society, **79**, pp. 61–78.

[30] WHO, 2000, *Cholera*, World Health Organisation, **N107**, [Online], [Cited January 2007], Available from `http://www.who.int/mediacentre/factsheets/fs107/en/index.html`.