Epidemiol. Infect. (2014), **142**, 520–529. © Cambridge University Press 2013
doi:10.1017/S0950268813001283

Spectral analysis based on fast Fourier transformation (FFT) of surveillance data: the case of scarlet fever in China

T. ZHANG[1], M. YANG[2], X. XIAO[1], Z. FENG[3], C. LI[1], Z. ZHOU[4], Q. REN[1] AND X. LI[1]*

[1] *West China School of Public Health, Sichuan University, Chengdu, Sichuan, China*
[2] *School of Community Health Sciences, University of Nottingham, Nottingham, UK*
[3] *Disease Control and Emergency Response Office, Chinese Centre for Disease Control and Prevention, Beijing, China*
[4] *School of Mathematics, Sichuan University, Chengdu, Sichuan, China*

Received 21 May 2012; Final revision 16 April 2013; Accepted 2 May 2013; first published online 10 June 2013

SUMMARY

Many infectious diseases exhibit repetitive or regular behaviour over time. Time-domain approaches, such as the seasonal autoregressive integrated moving average model, are often utilized to examine the cyclical behaviour of such diseases. The limitations for time-domain approaches include over-differencing and over-fitting; furthermore, the use of these approaches is inappropriate when the assumption of linearity may not hold. In this study, we implemented a simple and efficient procedure based on the fast Fourier transformation (FFT) approach to evaluate the epidemic dynamic of scarlet fever incidence (2004–2010) in China. This method demonstrated good internal and external validities and overcame some shortcomings of time-domain approaches. The procedure also elucidated the cycling behaviour in terms of environmental factors. We concluded that, under appropriate circumstances of data structure, spectral analysis based on the FFT approach may be applicable for the study of oscillating diseases.

Key words: Fast Fourier transformation, national notifiable disease surveillance data of China, scarlet fever, time-series analysis.

INTRODUCTION

The notion that many infectious diseases such as influenza, measles and chickenpox [1] exhibit repetitive or regular behaviour over time is of vital importance. In order to guide planning for future disease outbreaks, there has been considerable interest in applying mathematical and statistical models to elucidate the underlying mechanism of cyclical behaviour. Time-domain methods [2–4], a regression of the present based on the past, such as the seasonal autoregressive integrated moving average (SARIMA) model [5] and the generalized autoregressive conditional heteroskedasticity (GARCH) model [6], appear elegant in practice since they provide a view of nature in terms of intuitive linear forms. However, these models are at risk of over-differencing [7] (which indicates the abuse of differencing methods leading to heavy loss of information) and over-fitting [5] (which means fitting too many redundant parameters). More importantly, these modelling approaches assume

* Author for correspondence: Professor Xiaosong Li, Department of Medical Statistics, West China School of Public Health, Sichuan University, No. 17, Section 3, South Renmin Road, Chengdu, Sichuan, 610041, P.R. China.
(Email: lixiaosong1101@126.com)

linearity of variables [8], whereas the inherent structure of epidemic mechanisms is often nonlinear [9]. Thus, an alternative method for time-series analysis is needed.

Spectral analysis is based on the decomposition of an empirical series into regular components [10]. From the view of regression, spectral analysis models may be considered as a regression of the present on periodic sines and cosines. A main advantage of this model lies in its capability to explain the dynamics of some infectious diseases. For example, José & Bishop [4] compared a SARIMA model and the power spectral density method to characterize the overall dynamics of rotavirus infections as a whole and that of serotypes G1, G2, G3, G4 and G9 individually. According to that study, although the SARIMA model detected no obvious discernible pattern of dynamics except for the annual cycle, the spectral analysis did, in fact, capture seasonal, biannual and quinquennial periods.

As spectral analysis is becoming increasingly indispensable in biomedicine and epidemiology [11–13], some barriers to this method have been noted that may affect its application. First, some spectral approaches are too empirical to be appropriate. For example, the cyclical regression models [14] require frequency parameters to be preset. This is inappropriate, especially when the seasonal patterns remain elusive. Second, some analyses focus on the separate estimation method [15], which, in fact, consists of two separate steps: (1) to obtain the frequency parameter by means of the maximum entropy method (MEM); and (2) to utilize least squares fitting (LSF) to estimate the incidence curve. Such methods are simple, but their efficiency is yet to be demonstrated [16]. The joint estimation procedure seems more efficient. However, to our knowledge, joint estimation remains a daunting task in the frequency domain [17, 18]. Third, as claimed by Luo *et al.* [19]: with its origins in electrical engineering science, spectral analysis requires adequate mathematical and physical expertise which, unfortunately, presents an obstacle for many epidemiological researchers and practitioners.

Therefore, to overcome these shortcomings, we implemented a simple and efficient spectral analysis, based on the fast Fourier transformation (FFT) approach, in order to evaluate the epidemic dynamic of scarlet fever incidence in China. This method was shown to generate more valid, meaningful and even simpler explanations than time- and frequency-domain analyses, especially for periodic variations caused by biological, physical, or environmental phenomena [20].

We modelled data on the surveillance incidence of scarlet fever in China from 2004 to 2010. Scarlet fever is caused by erythrogenic toxin released by *Streptococcus pyogenes*. Nowadays, although scarlet fever is rare and generally mild, serious sequelae which threaten the heart and kidneys can still occur, especially in school children [21, 22]. The surveillance database of scarlet fever, together with the established knowledge of the cyclical behaviour of the infection, makes this disease an ideal example for studying epidemic dynamics using a FFT approach. Furthermore, we also include examples to show that this method can be easily applied to achieving various goals such as prediction and influencing factor exploration.

MATERIALS AND METHODS

Scarlet fever and environmental data

As prescribed by the Law of the P.R. China on the Prevention and Treatment of Infectious Diseases, physicians who find pathogen carriers or patients suspected of scarlet fever infection must report their findings to the local health and anti-epidemic agency within a specified time limit, next the health administration department under the State Council will promptly release information on and publicly announce the true epidemic situation. The original incidence data was obtained from the National Infectious Diseases Reporting System, Centers for Disease Prevention and Control, China. The incidence observations (represented by cases per100 000) to be analysed were aggregated by month from January 2004 to December 2010, for the whole nation and each region of mainland China (22 provinces, five autonomous regions, four municipalities, excluding Hong Kong, Macao SAR and Taiwan province). Moreover, we also excluded Jiangxi (code = 36) and Hainan (code = 46) provinces because there were too many zeros in the corresponding incidence time-series. Thus, there were 30 incidence time-series (each had 84 observations) analysed in our study. In addition, the environmental data (i.e. monthly sunshine hours, average relative humidity, average temperature and precipitation for major cities) were collected from the National Bureau of Statistics of China [23].

All statistical analyses and graphs in this paper were performed in R (R Foundation, Austria) which is a

free software environment for statistical computing and graphics.

Methods of analysis

We denote $X(t)$ as the number of scarlet fever incidences observed at time t. Based on the classical decomposition in time-series analysis [5], incidence series $\{X(t)\}$ are assumed to be represented as realization of the process:

$$\{X(t)\} = \text{trend component} + \text{periodic component} + \text{random noise component}, \quad (1)$$

In equation (1), the trend component, describing the long-term changes in data, is the polynomial function of time t. The periodic component is referred to as a function with known period, and the random noise component represents random errors, which is commonly treated as Gaussian white noise [15].

In the spectral analysis, the trend component can be easily obtained via maximum-likelihood estimation or least squares estimation; hence, the key point for analysis lies in the estimation of the periodic component. This component is described by the function $X_{PC}(t)$, which is assumed to be a mixture of cosine or sine functions with multiple frequencies and amplitudes.

$$X_{PC}(t) = A_0 + \sum_{i=1}^{N} A_i \cos(2\pi f_i t + \phi_i), \quad (2)$$

where $f_i(=1/T_i$; T_i is the period) is the frequency, A_0 is a constant indicating the average value of the periodic component, N is the total number of components, A_i is the amplitude and ϕ_i the phase that determines the starting point of the cosine function. All A_0, A_i, f_i and $\phi_i (i=1, \ldots, N)$ are parameters to be estimated in the model. In addition, through the derivative calculations in equation (2), we can obtain the maximum value and maximum point, $t_{\max}$ (at which the maximum value is reached), which respectively indicates the underlying peak and peak time in terms of epidemiology.

Although many spectral analysis methods take the form of equation (1) [15, 24], there are diverse estimations for parameters as stated in the previous section. Instead of being either too empirical or complicated, we aimed to establish the model simply with the help of FFT [25]. As a consequence, our method involves the following two steps.

Step I. Data pre-processing

Time-series were rearranged from the original dataset by month for each region. Outliers were detected by hypothesis tests beforehand [26–28]. Two types of outliers were taken into account: addictive outliers and level shifts. Usually an addictive outlier is caused by a recording error and a level shift may be the result of an outbreak or control. Thus, addictive outliers should be studied carefully to check whether there is any justification for smoothing or discarding them. If any level shifts exist, it is advisable to analyse the series by first breaking it into homogeneous segments at the corresponding time point. After outlier detection, the trend component is fitted by a polynomial function of time t and then removed. This procedure, subtracting the fitted function from the time-series, is also known as detrending. It is recommended that the order of the estimated polynomial function for the trend component is determined by the shape of incidence curve and hypothesis test. Logarithmic transformation is performed for the detrended data if the frequency histogram is separate from the normal distribution required for spectral analysis.

Step II. FFT approach

This step is the core of our method. We take $\{X_t^*\}$ to represent the pre-processed data after the first step. The concept of spectral analysis expresses the underlying dynamics in terms of periodic variations as Fourier frequencies being driven by sines and cosines. In this sense, FFT is employed as an efficient approach which transforms the data from the time domain (which can be considered as the function of time) into the frequency domain (i.e. the function of frequency):

$$d(j/N) = N^{-\frac{1}{2}} \sum_{t=1}^{N} X_t^* \exp(-2\pi i t j/N), \quad (3)$$

where j/N is designated the Fourier or fundamental frequency. Since i in equation (3) is the imaginary unit denoted as $i^2 = -1$, the result of FFT is a complex number. Given the Fourier frequency, the FFT value can be calculated by equation (3). It is also guaranteed mathematically that the phase can be calculated through the arc-tangent function of the FFT value while the amplitude equals its module [29]. Thus, the amplitude-frequency curve and phase-frequency curve can be plotted. From the former curve the prominent frequencies, which correspond to the highest amplitudes, are identified, and from the latter curve

the corresponding phases can then be estimated. Thus, the parameters in equation (2) can be estimated by FFT. The algorithms above are available in the STATS package of R software.

Furthermore, in order to take into account the variability of the parameters and the autocorrelation within time-series, the block bootstrap technique and permutation test were adopted [30]. The corresponding period of Fourier frequency (i.e. 1/Fourier frequency) was chosen to be the block length so that the autocorrelation structure within seasonal blocks is preserved. First, we simulated 10000 replications under the null hypothesis of absence of seasonality by block bootstrap sampling, and then obtained the P value by comparing the initially observed statistics (peak or peak time) with the distribution of the 10000 simulated replications. In addition, the confidence intervals of parameters were obtained with the bootstrap percentile method (level of significance = 0·05).

The two steps above constitute the main body of our method. If continuous periodic oscillations are identifiable, then our method should be able to predict and explain the cycling behaviour in terms of extrinsic or intrinsic factors [31]. However, in effect, these explorations may belong to the spectral analysis itself by definition, although they can be viewed as the derivatives of our approach.

RESULTS

Our analyses focused both on incidence data in each region and the general situation over the whole country. For clarity of interpretation, the raw incidence data for the whole nation, presented in Figure 1*a*, is used as an illustration.

Main results of the method

Step I

The national incidence series consisted of 84 observations and no outlier was detected ($P > 0{\cdot}05$). As shown in Figure 1*a* no sign of trend component was observed, and neither the linear ($P = 0{\cdot}54$) nor quadratic ($P = 0{\cdot}70$) regression curve of incidence on time was statistically significant. Figure 1*b* shows that the frequency histogram of the national incidence data reasonably resembles the normal distribution with $P = 0{\cdot}07$ from the Kolmogorov–Smirnov test, suggesting that the data are suitable for a FFT approach. Otherwise, logarithmic transformation is recommended before analysis.

Step II

We used FFT to further characterize the periodic component of the pre-processed data. Similarly to a triangular prism which decomposes the light into different frequencies in the colour spectrum, the FFT approach offers a straightforward means to isolate the periodic components oscillating at various frequencies. As mentioned above, we used the amplitude-frequency curve (Fig. 1*c*) to identify the prominent frequency and subsequently determined the phase through the phase-frequency curve (Fig. 1*d*).

In addition, we used a suitable bandpass filter [20], which is a mode of constraining those frequencies within a certain range (e.g. $0{\cdot}01 \leqslant f \leqslant 0{\cdot}50$) and rejecting frequencies outside that range, to reconstruct the periodic component in equation (1). Such an approach is reasonable because periods that are either too long or too short are considered inappropriate for the FFT approach. In Figure 1*c*, the seasonality of scarlet fever in China was identified by an annual pattern (frequency = 0·0833, period ≈12 months) and a semi-annual pattern (frequency = 0·1667, period ≈6 months), respectively. As a result, the periodic component was able to be expressed in the following equation, where all the parameters were statistically significant (permutation tests, $P < 0{\cdot}05$):

$$\begin{aligned}\hat{X}_{PC}(t) = {} & 0\cdot1759 + 0\cdot0942^{*}\cos(2\pi^{*}0\cdot1667^{*}t \\ & +1\cdot6871) + 0\cdot0314^{*}\cos(2\pi^{*}0\cdot0833^{*}t \\ & -1\cdot6808),\end{aligned} \tag{4}$$

Finally, as there was no significant trend component in the data, the estimation for model (1) of the national situation could be expressed in a model with a single component as:

$$\hat{X}(t) = \hat{X}_{PC}(t), \tag{5}$$

Based on equation (5), with parameter estimates as in equation (4), it can be concluded, by maximization and block bootstrapping, that the first peak occurred between March and April (peak time 1 = 4·02, 95% CI 3·88–4·81 months) with a peak value 0·19, while the second peak occurred between October and December (peak time 2 = 11·31, 95% CI 10·80–12·39 months).

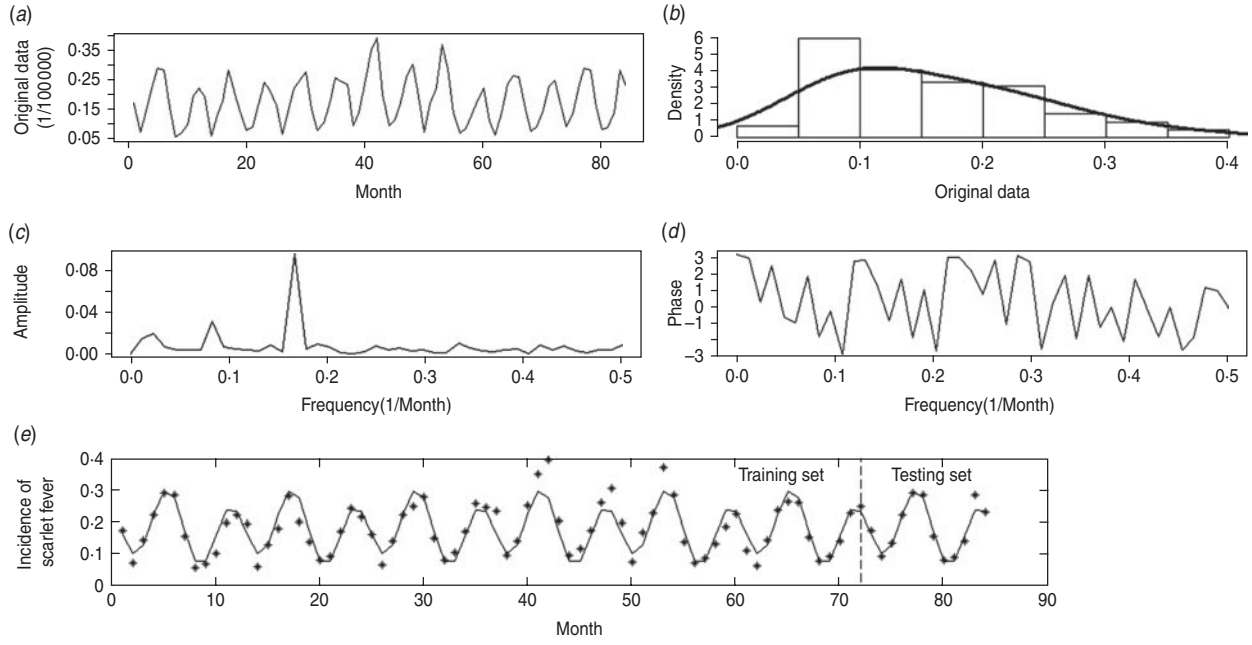

Fig. 1. Monthly incidence data of scarlet fever in China from 2004 to 2010. (*a*) The original time-series; (*b*) histogram of original area; (*c*) estimated amplitude-frequency curve; (*d*) estimated phase-frequency curve; (*e*) original data and fast-Fourier-transform forecast-fitted curve, with a vertical line splitting the training and testing periods.

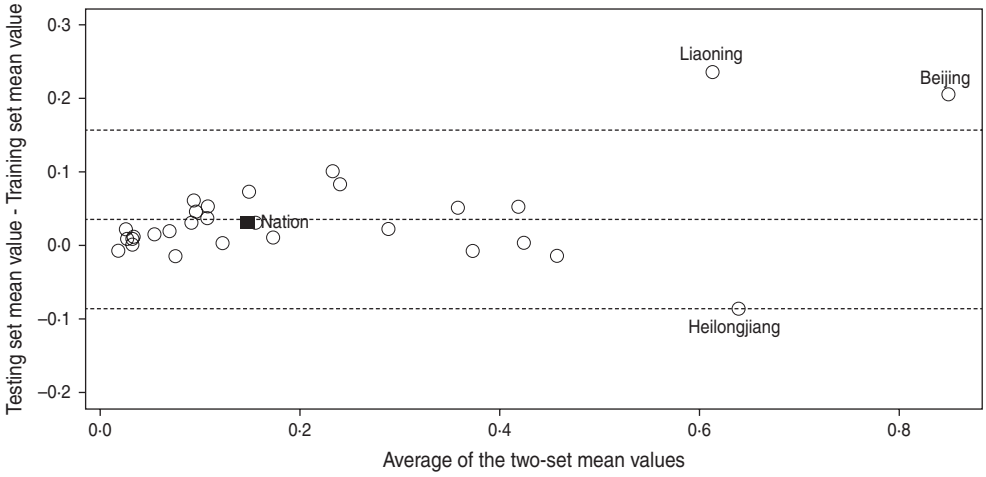

Fig. 2. Bland–Altman plot for the training and testing set mean values. The open symbols (○) represent the region-specific incidence time-series and the solid symbol (■) represents the national incidence time-series. The top and bottom dashed horizontal lines represent the 95% limits of agreement for each comparison, and the central dashed line represents the average of the difference between the two-set mean values.

Evaluation and application of the method

Prediction and validation

One of the most important applications of time-series analysis is prediction. The extrapolation of the epidemic curve fitted by equation (4) can be used for prediction of the incidence because it is regarded as the predictable part. For each incidence time-series, we split the data into the training set (January 2004–December 2009) and testing set (January 2010–December 2010), and used the first set for model fitting and the second set for making predictions.

To confirm the agreement between these two sets, we first calculated the mean values of training and testing sets separately for each incidence time-series, and then applied the Bland–Altman plot (see Fig. 2).

Table 1. *Comparison of errors for FFT spectral analysis and the SARIMA model**

Time-series label	Administrative code	In-sample error		External sample error		No. of outliers	
		FFT	SARIMA	FFT	SARIMA	AO	LS
Mainland of P.R. China		0·0245	0·0278	0·0120	0·0136	0	0
North							
Beijing	11	0·2333	0·2878	0·3294	0·5881	0	1
Tianjin	12	0·1394	0·1706	0·1745	0·2947	3	0
Hebei	13	0·0614	0·0564	0·0594	0·0992	5	0
Shanxi	14	0·0640	0·0746	0·0823	0·1729	1	0
Inner Mongolia	15	0·0931	0·1345	0·0763	0·0815	1	0
Northeast							
Liaoning	21	0·2289	0·2416	0·2889	0·4813	2	0
Jilin	22	0·2216	0·1397	0·1740	0·3741	3	0
Heilongjiang	23	0·2698	0·2202	0·3342	0·4336	3	0
East							
Shanghai	31	0·1216	0·1164	0·1403	0·1685	3	0
Jiangsu	32	0·0194	0·0223	0·0105	0·0203	0	0
Zhejiang	33	0·0292	0·0462	0·0461	0·0686	2	0
Anhui	34	0·0094	0·0140	0·0190	0·0187	2	0
Fujian	35	0·0127	0·0156	0·0181	0·0182	5	0
Jiangxi	36	—	—	—	—	—	—
Shandong	37	0·0603	0·0423	0·0679	0·0515	6	0
Central South							
Henan	41	0·0184	0·0213	0·0334	0·0833	2	0
Hubei	42	0·0107	0·0121	0·0127	0·0192	3	0
Hunan	43	0·0060	0·0065	0·0092	0·0177	4	0
Guangdong	44	0·0103	0·0117	0·0117	0·0250	2	0
Guangxi	45	0·0134	0·0223	0·0129	0·0689	3	0
Hainan	46	—	—	—	—	—	—
Southwest							
Chongqin	50	0·0267	0·0375	0·0353	0·0350	5	0
Sichuan	51	0·0325	0·0348	0·0365	0·0380	3	0
Guizhou	52	0·0162	0·0246	0·0240	0·0323	1	0
Yunnan	53	0·0360	0·0566	0·0571	0·0390	5	0
Tibet	54	0·1525	0·2253	0·1607	0·1347	3	0
Northwest							
Shanxi	61	0·0425	0·0549	0·0343	0·1088	3	0
Gansu	62	0·0386	0·0444	0·0473	0·0481	1	0
Qinghai	63	0·1071	0·2715	0·1569	0·1257	3	1
Ningxia	64	0·1896	0·2013	0·2189	0·3383	1	0
Xinjiang	65	0·1149	0·1727	0·1906	0·3358	5	0
Average	—	0·0801	0·0936	0·0958	0·1445	—	—

FFT, Fast Fourier transformation; SARIMA, seasonal autoregressive integrated moving average; AO, addictive outlier; LS, level shift.

* Mean absolute deviation is calculated as an error measure.

In Figure 2, each incidence time-series is represented by assigning the average of the two mean values as the abscissa (x axis) value, and the difference between the two values as the ordinate (y axis) value. It can be seen from Figure 2 that despite just a few exceptions (i.e. Beijing, Liaoning, Heilongjiang), the training and testing sets are nearly identical within each incidence time-series. This was also verified by the paired two-sample t test ($t = 0{\cdot}6477$, $P = 0{\cdot}5198$).

Most of these series exhibit seasonal nonlinearity variation structure, so it seems plausible to introduce spectral analysis rather than a time-domain approach.

Table 2. *Seasonal patterns of scarlet fever incidence time-series in P.R. China by region, 2004–2011**

Location/region	Administrative code	Latitude (degrees)	Peak 1 (1/100 000)	Peak time 1 (month)	Peak 2 (1/100 000)	Peak time 2 (month)
North						
Beijing	11	40·3	0·7030	5·2	0·8140	11·8
Tianjin	12	39·4	0·5316	6·1	0·4099	10·8
Hebei	13	39·3	0·2132	4·6	0·1727	12·2
Shanxi	14	37·7	0·3137	6·0	0·2610	11·1
Inner Mongolia	15	45·4	0·4401	4·5	0·5041	12·3
Northeast						
Liaoning	21	41·1	0·6308	5·1	0·8242	12·1
Jilin	22	43·6	0·3837	4·8	0·3934	12·0
Heilongjiang	23	48·5	0·6093	5·2	0·7483	11·6
East						
Shanghai	31	31·1	0·3398	5·0	0·2724	11·9
Jiangsu	32	32·9	0·1228	5·7	0·1015	10·9
Zhejiang	33	29·2	0·1277	6·0	0·1108	11·2
Anhui	34	32·0	0·0412	6·3	0·0335	11·3
Fujian	35	26·0	0·0426	4·9	0·0399	14·0
Jiangxi	36	27·3	—	—	—	—
Shandong	37	36·3	0·1429	5·1	0·1162	11·4
Central South						
Henan	41	33·9	0·0666	5·0	0·0658	12·2
Hubei	42	31·2	0·0379	6·1	0·0327	11·0
Hunan	43	27·4	0·0171	5·1	0·0199	11·6
Guangdong	44	22·9	0·0341	4·6	0·0344	12·0
Guangxi	45	23·9	0·0368	4·1	0·0402	12·1
Hainan	46	19·2	—	—	—	—
Southwest						
Chongqing	50	30·2	0·0936	4·7	0·0732	12·3
Sichuan	51	30·2	0·1404	5·6	0·1163	11·2
Guizhou	52	26·9	0·0816	4·9	0·0684	11·4
Yunnan	53	25·2	0·1493	3·9	0·1226	11·9
Tibet	54	31·7	0·1586	5·2	0·1558	11·5
Northwest						
Shanxi	61	35·6	0·1951	4·9	0·1610	11·8
Gansu	62	37·7	0·2028	5·0	0·1714	11·1
Qinghai	63	35·4	0·3719	4·5	0·3027	12·0
Ningxia	64	37·3	0·5507	5·2	0·4428	11·7
Xinjiang	65	41·8	0·4092	5·0	0·4141	12·1

* The focus is on the mainland of the P.R. China, which does not contain Hong Kong, Macao SAR and Taiwan. Also excluded in analysis are Jiangxi (code = 36) and Hainan (code = 46) provinces because there are too many zeros in the corresponding time-series. The first digit of administrative code refers to the location: 1, North; 2, Northeast; 3, East; 4, Central South; 5, Southwest; 6, Northwest.

To confirm this, we performed two analyses. First, we compared the internal and external validities of our method with those of the SARIMA model of time-domain approach. To take into account the relatively low incidence, we chose the mean absolute deviation (MAD) [32] as an error measure. Table 1 lists the results of the whole nation as well as each region. As can be seen, even in the presence of outliers, the application of FFT takes on a lower MAD than the SARIMA model both for the in-sample and external sample errors in most cases.

Second, we compared the average in-sample error and average out-of-sample error by each method. The values were 0·08 and 0·09, resprectively, for FFT (error difference $\approx$0·01), while the values were 0·09 and 0·14, respectively, for the SARIMA model

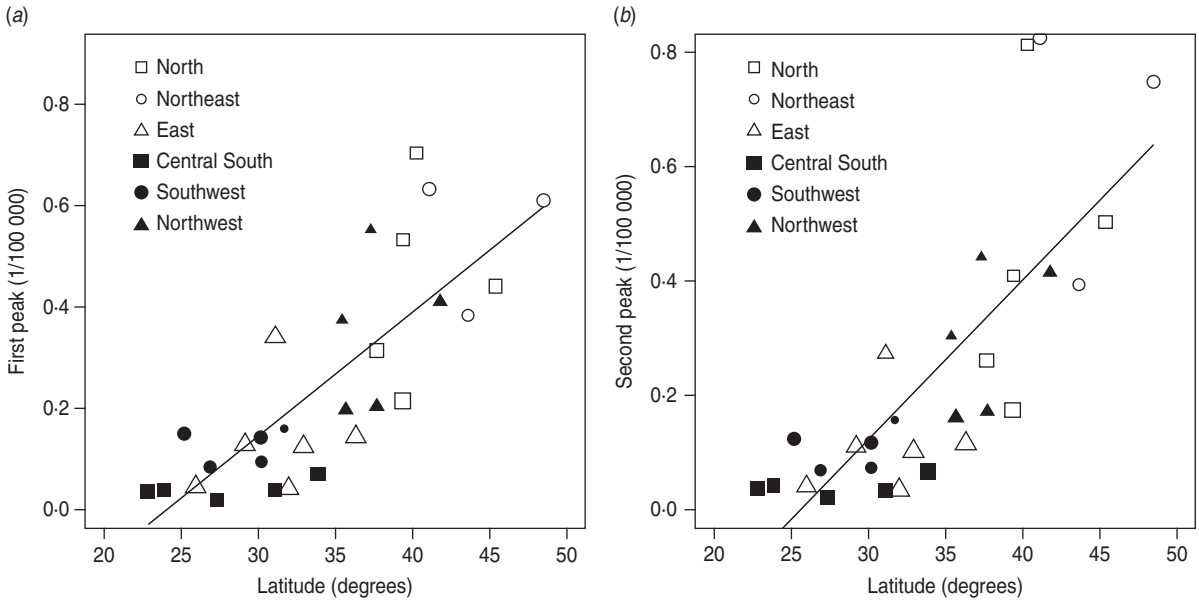

Fig. 3. Trends of peak variations in scarlet fever incidence by latitude across China. (*a*) The first peak against the latitude of regions; (*b*) the second peak against the latitude of regions. Each combination of symbols (triangle, square, circle) and colour (black for south, white for north) of the graph refers to the location of the corresponding region, while the size of the graph is proportional to the population of each region.

(error difference $\approx 0{\cdot}05$). The larger error difference for the SARIMA model may be caused by over-fitting [5]. Thus, the results for FFT appear to show a better performance than the SARIMA model.

Influencing factor exploration

Table 2 presents peaks and peak times extracted from each region (province, autonomous region, municipality). We found that both the first and second peaks were significantly large for the region-specific time-series. To examine whether latitude/longitude by region had an impact on the peaks and peak times, we further employed univariate regression models with the LSF method in which the peaks and peak times derived from each series were separate dependent variables, and the latitude/longitude with respect to each province were capital independent variables. It was found that the relationship between first/second peak and latitude was significant (first peak: $R^2 = 0{\cdot}6077$, $P < 0{\cdot}001$; second peak: $R^2 = 0{\cdot}5936$, $P < 0{\cdot}001$), while others were not. Figure 3(*a, b*) shows the plots of first/second peak values against the latitudes of province capitals. It is intriguing to find that the peak values (both first and second peaks) varied significantly along a latitudinal gradient. Namely, the peak values were highest in the northern zones while becoming attenuated southwards.

The geographical pattern in China as discussed above, has also been stated to be correlated with a variety of environmental factors (e.g. sunshine hours, precipitation, relative humidity, temperature) in previous studies [33, 34]. To confirm this, Pearson's product-moment correlation analysis was used. We found that the peak values were significantly positively correlated with sunshine hours ($r = 0{\cdot}23$, $P < 0{\cdot}01$), and negatively correlated with precipitation ($r = -0{\cdot}21$, $P < 0{\cdot}01$), relative humidity ($r = -0{\cdot}33$, $P < 0{\cdot}01$), and temperature ($r = -0{\cdot}23$, $P < 0{\cdot}01$). These results provided a clue for further investigation.

DISCUSSION

In this paper, we suggest that the regularity in time-series can be expressed in terms of periodic variations of the underlying phenomenon that produce the series, expressed as Fourier frequency driven by sines and cosines. We present a rather simple and efficient spectral analysis approach that has a tendency to outperform SARIMA models in both model-fitting and prediction.

In the previous section, we took the incidence data of the whole nation as an example. Such series may become an ideal paradigm since there is neither trend component nor outlier. However, other situations can be more complicated. For example, we

detected a level shift change at $t = 55$ (July 2008) in the incidence series of Beijing. This is in accord with other studies, suggesting that the incidence of scarlet fever had fallen in Beijing prior to the 2008 Olympic Games [35]. Thus, we split the data at $t = 55$ and removed the trend term from each segment separately. As the detrended data were skewed, logarithmic transformation was then performed. After these adjustments, we applied spectral analysis to derive the periodic component; the results are shown in Table 1.

When comparing FFT and the SARIMA model, only five (16·7%) of the total of 30 series had a larger FFT in-sample error than the SARIMA model. As inferred by the larger number of outliers, the corresponding regions (i.e. Hebei, Liaoning, Heilongjiang, Shanghai, Shandong) are typically hotspots for scarlet fever outbreaks in China [36]. There are four regions (i.e. Shandong, Tibet, Qinghai, Yunnan) where the external sample error for FFT is obviously greater than it is for the SARIMA model. By checking the original data, we found that scarlet fever incidence in these regions changed from 20% to 50% in 2010 compared to the incidence over the last 6 years while the national average level remained unchanged. These results imply that, on the one hand, changes in data structure can have unavoidable influences on spectral analysis. However, by contrast, empirical evidence with FFT showing many more minimal errors and lower error difference seems to affirm the validity and robustness of spectral analysis compared to that of the time-domain approach.

China is the world's third largest country, extending about 50 degrees of latitude, encompassing diverse regional climates, terrain, population densities and social customs, etc. The application of the results of our method suggest that environmental forces (precipitation, relative humidity, temperature) play an important role in scarlet fever epidemics in China, which coincides with epidemics reported in previous studies in the Czech Republic and Russia, as well as some cities in China [33, 34, 37, 38]. The application sheds light on the underlying practical value of our method. Since spectral analysis can provide different perspectives compared to conventional time-series models, it is reasonable to expect more applications of the method in future investigations on scarlet fever and other infectious diseases to be performed.

Time-series analysis is a data-driven technique. To our knowledge, epidemics of streptococcal infection including scarlet fever, as well as other diseases such as chickenpox, are fundamentally determined by the mechanism of the noisy limit cycle [9], which leads to the temporal changes shown as seasonal variations. Therefore, we believe that, under appropriate conditions (e.g. normality and absence of outliers) of data structure, our procedure will contribute to further studies of many other periodically oscillating diseases. More studies on transmission dynamics are still required.

ACKNOWLEDGEMENTS

The authors thank the anonymous referees for their constructive comments on the manuscript. We are extremely grateful to Dr Liu Yuanyuan (Sichuan University, China) for inspiring discussions and valuable advice. Our thanks are also due to Katrina Seymour (University of Technology, Sydney) and Dr Bao Huanchen (University of Virginia) for revision of the paper. This study was supported by the National Natural Science Foundation (grant no. 30571618) and National Special Foundation for Health Research (grant no. 200802133) of China.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Becker NG.** *Analysis of Infectious Disease Data*. New York: Chapman & Hall, 1989, pp. 164.
2. **Liu Q, *et al.*** Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. *BMC Infectious Diseases* 2011; **11**: 218.
3. **Zeger SL, Irizarry R, Peng RD.** On time series analysis of public health and biomedical data. *Annual Review of Public Health* 2006; **27**: 57–79.
4. **José MV, Bishop RF.** Scaling properties and symmetrical patterns in the epidemiology of rotavirus infection. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 2003; **358**: 1625–1641.
5. **Box GEP, Jenkins GM.** *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day, 1976, pp. 33–35.
6. **Bollerslev T.** Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 1986; **31**: 307–327.
7. **Cai XH.** Time series analysis of air pollution CO in California south coast area, with seasonal ARIMA model and VAR model (dissertation). Los Angeles, CA, USA: University of California, 2008, 84 pp.
8. **Ture M, Kurt I.** Comparison of four different time series methods to forecast hepatitis A virus infection. *Expert Systems with Applications* 2006; **31**: 41–46.

9. **Olsen LF, Schaffer WM.** Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics. *Science* 1990; **249**: 499–504.
10. **Shumway RH.** *Time Series Analysis and its Applications*. New York: Springer Press, 2006, pp. 174.
11. **Cai W, Xu Z, Baumketner A.** A new FFT-based algorithm to compute Born radii in the generalized Born theory of biomolecule solvation. *Journal of Computational Physics* 2008; **227**: 10162–10177.
12. **Jakoby B, Vellekoop MJ.** FFT-based analysis of periodic structures in microacoustic devices. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 2000; **47**: 651–656.
13. **Dillenseger JL, Esneault S.** Fast FFT-based bioheat transfer equation computation. *Computers in Biology and Medicine* 2010; **40**: 119–123.
14. **Lui KJ, Kendal AP.** Impact of influenza epidemics on mortality in the United States from October 1972 to May 1985. *American Journal of Public Health* 1987; **77**: 712–716.
15. **Sumi A, Kamo KI.** MEM spectral analysis for predicting influenza epidemics in Japan. *Environmental Health and Preventive Medicine* 2011; **17**: 98–108.
16. **Tsay RS.** *Analysis of Financial Time Series*, 2nd edn. Hoboken, New Jersy: Wiley & Sons, Inc., 2005, pp.146.
17. **Ma X.** Joint estimation of time delay and frequency delay in impulsive noise using fractional lower order statistics. *IEEE Transactions on Signal Processing* 1996; **44**: 2669–2687.
18. **Lin DD, *et al.*** Joint estimation of channel response, frequency offset, and phase noise in OFDM. *IEEE Transactions on Signal Processing* 2006; **54**: 3542–3254.
19. **Luo T, *et al.*** Time series analysis based by spectral power distribution-maximum entropy method (MEM) [in Chinese]. *Chinese Journal of Health Statistics* 2010; **5**: 477–484.
20. **Peyton ZP.** *Probability, Random Variables, and Random Signal Principles*, 4th edn. New York: McGraw Hill Companies, Inc., 2001, pp. 25.
21. **Lamden KH.** An outbreak of scarlet fever in a primary school. *Archives of Disease in Childhood* 2011; **96**: 394–397.
22. **Barnett BO, Frieden IJ.** Streptococcal skin diseases in children. *Seminars in Dermatology* 1992; **11**: 3–10.
23. **National Bureau of Statistics.** Database (http://www.stats.gov.cn/). Accessed 9 April 2012.
24. **Alonso WJ, *et al.*** Seasonality of influenza in Brazil: a traveling wave from the Amazon to the subtropics. *American Journal of Epidemiology* 2007; **165**: 1434–1442.
25. **Cochran WT, *et al.*** What is the fast Fourier transform? *Proceedings of the IEEE* 1967; **55**: 1664–1674.
26. **Tsay RS.** Outliers, level shifts, and variance changes in time series. *Journal of Forecasting* 1988; **7**: 1–20.
27. **Chen C, Liu LM.** Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association* 1993; **8**: 284–297.
28. **Chang I, Tiao GC, Chen C.** Estimation of time series parameters in the presence of outliers. *Technometrics* 1988; **30**: 193–204.
29. **Bracewell RN.** *The Fourier Transformation and its Application*, 3rd edn. New York: McGraw Hill Companies, Inc., 1999.
30. **Davision AC, Hinkley DV.** *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press, 1997, pp. 23.
31. **Breban R, *et al.*** Is there any evidence that syphilis epidemics cycle? *Lancet Infectious Diseases* 2008; **8**: 577–581.
32. **Goh C, Law R.** Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management* 2002; **23**: 499–510.
33. **Wang J, *et al.*** Epidemiological investigation of scarlet fever in Hefei City, China, from 2004 to 2008. *Tropical Doctor* 2010; **40**: 225–226.
34. **Li XY, *et al.*** Correlative study on association between meteorological factors and incidence of scarlet fever in Beijing. *Practical Preventive Medicine* 2007; **14**: 1435–1437.
35. **Qian HK, *et al.*** Spatial-temporal scan statistic on scarlet fever cases in Beijing, 2005–2010. *Disease Surveillance* 2011; **26**: 435–238.
36. **Liu Z, Wang BX, Wang SC.** The analysis of dynmaics of scarlet fever in China from 2003–2008 [in Chinese]. *Journal of public health and preventive medcine* 2009; **20**: 21–22.
37. **Briko NI, *et al.*** Epidemiological pattern of scarlet fever in recent years. *Zhurnal mikrobiologii, epidemiologii, i immunobiologii* 2003; **5**: 67–72.
38. **Hubalek Z.** North Atlantic weather oscillation and human infectious diseases in the Czech Republic, 1951–2003. *European Journal of Epidemiology* 2005; **20**: 263–270.