Environ Health Prev Med (2012) 17:98–108
DOI 10.1007/s12199-011-0223-0

REGULAR ARTICLE

MEM spectral analysis for predicting influenza epidemics in Japan

Ayako Sumi · Ken-ichi Kamo

Received: 30 September 2010 / Accepted: 15 May 2011 / Published online: 7 June 2011

Abstract
Objectives The prediction of influenza epidemics has long been the focus of attention in epidemiology and mathematical biology. In this study, we tested whether time series analysis was useful for predicting the incidence of influenza in Japan.
Methods The method of time series analysis we used consists of spectral analysis based on the maximum entropy method (MEM) in the frequency domain and the nonlinear least squares method in the time domain. Using this time series analysis, we analyzed the incidence data of influenza in Japan from January 1948 to December 1998; these data are unique in that they covered the periods of pandemics in Japan in 1957, 1968, and 1977.
Results On the basis of the MEM spectral analysis, we identified the periodic modes explaining the underlying variations of the incidence data. The optimum least squares fitting (LSF) curve calculated with the periodic modes reproduced the underlying variation of the incidence data. An extension of the LSF curve could be used to predict the incidence of influenza quantitatively.
Conclusions Our study suggested that MEM spectral analysis would allow us to model temporal variations of influenza epidemics with multiple periodic modes much more effectively than by using the method of conventional time series analysis, which has been used previously to investigate the behavior of temporal variations in influenza data.

Keywords Influenza · Prediction analysis · Time series analysis · Surveillance · Epidemiology

A. Sumi (✉)
Department of Hygiene, Sapporo Medical University School of Medicine, S-1, W-17, Chuo-ku, Sapporo 060-8556, Japan
e-mail: sumi@sapmed.ac.jp

K. Kamo
Center for Medical Education, Department of Liberal Arts and Sciences, Sapporo Medical University School of Medicine, Sapporo 060-8556, Japan

Introduction

For preventing and predicting influenza epidemics, it is necessary to investigate temporal variations of the disease morbidity data in detail [1–5]. To elucidate temporal variational structures in the morbidity data of influenza, many studies have been carried out by using conventional time series analysis [6–12], such as a Gaussian random process for the modeling of influenza epidemics [12] and an autoregressive model (AR) including a seasonal autoregressive-integrated moving average model [10].

On the other hand, recently, researchers have tried to interpret the behavior of temporal variations in the morbidity of influenza in terms of nonlinear dynamics which causes multiple periodic structures with characteristic fluctuations [13–16]. However, the Gaussian random process and the AR model using random noise are not robust for interpreting the multiple periodic structures caused by nonlinear dynamics [17]. Thus, in order to investigate temporal variations in the morbidity of influenza for predicting the disease incidence, it is necessary to establish a new method of time series analysis.

We have already proposed a newly devised method of time series analysis, which enables us to identify multiple periodicities in the temporal variations of time series [18–20]. The series of analysis in the present study combines spectral analysis based on the maximum entropy method (MEM) in the frequency domain with the nonlinear

least squares method (LSM) in the time domain. MEM spectral analysis is useful to investigate the periodicities of time series of short data length, such as the morbidity data of infectious diseases. The validity of the result obtained from MEM spectral analysis is confirmed by calculating the optimum least squares fitting (LSF) curve to the time series with the LSM. With this method of time series analysis, we previously proposed a new analysis method for the prediction of epidemics with a clear criterion of adequate prediction [24]. The present method is based on the most traditional method of prediction analysis, which uses an extrapolation curve corresponding to underlying variations of time series in the future. A key point of the method is an estimation of the underlying variation of time series. The present method has been used successfully for the time series generated from a susceptible/exposed/infectives/recovers model [21], which is a well-known nonlinear dynamical system for analyzing epidemics of infectious diseases including influenza. Satisfactory results were also obtained for the morbidity data of measles [18, 21–25], which, because of the comparative simplicity of infection and immunity of measles, is useful as a model of another infectious diseases [26].

Regarding influenza, Kakehashi et al. [8] separated the morbidity data in Japan into a seasonal component, a quadratic trend, and an AR process. On the other hand, our preceding work on influenza epidemics in Japan [27] identified a periodic structure of morbidity data that changes temporally because of the effect of influenza pandemics and vaccine programs in Japan. Based on this result, in the present study, we further investigated periodic structures of influenza morbidity data in Japan in detail, and attempted to predict future values of the morbidity of influenza quantitatively.

Materials and methods

Incidence data

The time series data analyzed in the present study represent the monthly reported numbers of influenza cases per 100,000 population. The data were obtained from *Statistics of Communicable Disease* in Japan [28]. The monthly incidence data were gathered over 612 months from January 1948 to December 1998, covering the periods of pandemics in 1957, 1968, and 1977. A detailed description of the incidence data is given in our previous work [27].

We also analyzed the data for the weekly incidence of influenza in Japan from January 1987 to October 2010 obtained from the *Infectious Diseases Weekly Report Japan* (IDWR) [29]. The weekly incidence data represent the weekly reported numbers of influenza cases per sentinel clinic and hospital.

Time series analysis

Theoretical background

We take any time series data $\{x(t)\}$ (t time) to represent discrete data at $t = k\Delta t$ ($k = 1, 2, 3, \ldots N$) where Δt is the time interval and N the length of the time series. The data are divided into two parts: analysis range $\{x_A(t)\}$ and prediction range $\{x_P(t)\}$. We investigate the periodic structure of the data in $\{x_A(t)\}$, and use it to indicate the data in $\{x_P(t)\}$ which follows the analysis range behind.

The data in the analysis range, $\{x_A(t)\}$, are assumed to be composed of systematic and fluctuating parts [30]:

$$\{x_A(t)\} = \text{systematic part} + \text{fluctuating part}. \quad (1)$$

The systematic part in Eq. 1 is regarded as an underlying variation of $\{x_A(t)\}$. The fluctuating part in Eq. 1, resulting from a dynamic mechanism such as chaos dynamics existing behind the data and/or random noise caused by measurement error, is obtainable by subtracting the underlying variation from $\{x_A(t)\}$. We can use the extrapolation curve of the underlying variation for prediction [31]. A key point for prediction analysis is the estimation of the underlying variation.

The underlying variation is assumed to be described as the function $x_{UV}(t)$ given by the linear combination of sine and cosine functions,

$$x_{UV}(t) = a_0 + \sum_{n=1}^{S} \{a_n \sin(2\pi f_n t) + b_n \cos(2\pi f_n t)\}, \quad (2)$$

which is calculated using LSM for $\{x_A(t)\}$ with unknown parameters S, f_n, a_0, a_n, and b_n ($n = 1, 2, \ldots, S$) where S is the total number of components, f_n (=$1/T_n$, T_n: its period) the frequency of the n-th periodic component, a_n and b_n the amplitudes of the n-th component, and a_0 a constant which indicates the average value of the time series.

The LSM using Eq. 2 must be nonlinear. Linearization of this nonlinearity is required to obtain unique optimum values of these parameters. In the present study, linearization is achieved by using the value of f_n estimated by MEM spectral analysis. $x_{UV}(t)$ thus obtained as the optimum LSF curve is extended to the data in the prediction range. As a result, future values are indicated quantitatively by the extrapolation curve of $x_{UV}(t)$. The procedure for the estimation of $x_{UV}(t)$ is constructed in five steps (I, II, III, IV, and V).

Step I: Setting up the incidence data for analysis

Logarithmic transformation and/or removing long-term trend of the data are performed, if the frequency histogram

for the data is apart from the normal distribution required for conventional spectral analysis.

Step II: Determination of f_n (spectral analysis)

To estimate f_n in Eq. 2, we conduct MEM spectral analysis for $\{x_A(t)\}$ (Eq. 1), and obtain the power spectral density (PSD). From the PSD, we can obtain the power representing the amount of amplitude of the data at each frequency [30]. An outline of MEM spectral analysis is given in the Appendix.

Step III: Determination of the fundamental modes and the value of S

Based on the result of periods estimated by MEM spectral analysis, we must assign fundamental modes f_n that construct the periodic structure of $x_{UV}(t)$ of $\{x_A(t)\}$ (Eq. 2). To assign fundamental modes f_n, we define the "contribution ratio" to define a criterion for the evaluation of adequacy of $x_{UV}(t)$ to $\{x_A(t)\}$. Detailed explanations of the contribution ratio are included in the Appendix. Based on the result of the contribution ratio, we can safely determine the fundamental modes constructing $x_{UV}(t)$ of $\{x_A(t)\}$ and the optimum value of S in Eq. 2.

Step IV: Determination of a_0, a_n, and b_n (LSM)

The optimum values of parameters a_0, a_n, and b_n ($n = 1, 2, \ldots, S$) in Eq. 2 are exactly determined from the optimum LSF calculation using Eq. 2 with the values of S and f_n.

Step V: Prediction of the incidence

The extrapolation of the optimum LSF curve can be used for prediction of the incidence because the optimum LSF curve is regarded as the predictable part [31]. $x_{UV}(t)$ determined in Step IV is extended to the prediction range.

Setting up the incidence data for the analysis

The monthly incidence data of influenza ($N = 612$) are plotted in Fig. 1a. The histogram of the incidence data (Fig. 1a′) is apart from the normal distribution required for conventional spectral analysis. Thus, first, we carried out logarithmic transformation of the incidence data, where 122 zero values were replaced by small positive random values. Next, the long-term oscillatory trend of the log-data was removed by the LSM. As a result, the residual data were obtained (Fig. 1b). The frequency histogram for the residual data (Fig. 1b′) approximates to the normal

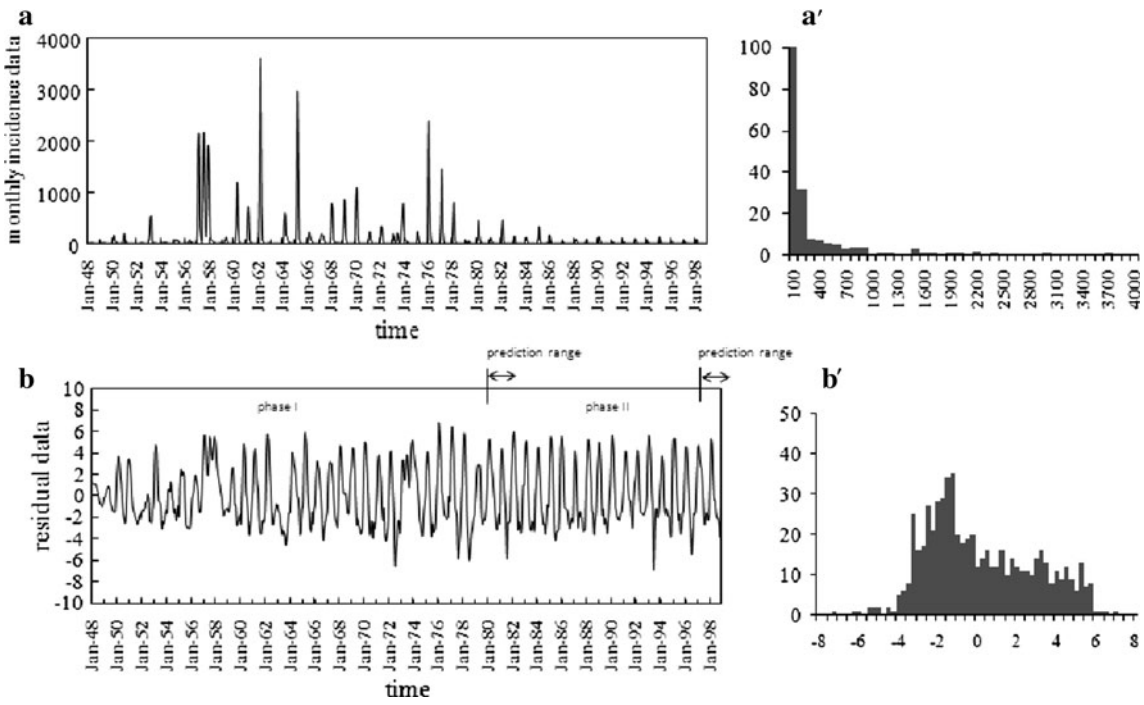

Fig. 1 Monthly incidence data of influenza in Japan from January 1948 to December 1998: **a** the original data; **a′** histogram of the original data; **b** the residual data, obtained by subtracting the least squares fitting (LSF) curve from the logarithmically transformed original data; **b′** histogram of the residual data. *Small vertical lines* in **b** indicate the boundaries of phase I (January 1948–December 1979) and phase II (January 1980–December 1996)

distribution required for conventional spectral analysis. The details of the procedure for setting up the incidence data (Fig. 1a) for the analysis are discussed in our previous work [27].

Setting up the prediction and analysis ranges of the incidence data of influenza

In our preceding work [27], it was confirmed that the periodic structures of the residual data (Fig. 1b) during 1948–1979 and the data during 1980–1998 were different from each other. It was considered that this difference was due to the occurrence of two influenza pandemics ("Asian flu" in the year 1957 and "Hong Kong flu" in the year 1968/1969) and the start of vaccine programs in 1962 and 1976. Thus, in the present study, we divided the residual data (Fig. 1b) into two ranges (phases I and II) and set the boundary of the residual data at the end of 1979 as phase I (January 1948–December 1979) and phase II (January 1980–December 1996). By calculating $x_{UV}(t)$ of the residual data in phase I and phase II, we attempted to predict the residual data during January 1980–December 1981 and that during January 1997–December 1998, respectively.

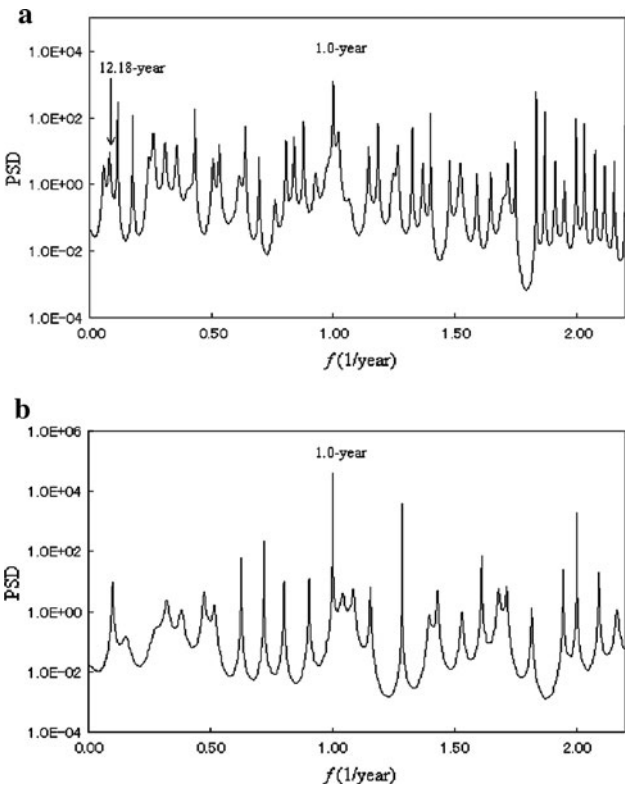

Fig. 2 Power spectral density (PSD) obtained by maximum entropy method (MEM) spectral analysis for two ranges of the residual data ($f < 2.2$): **a** phase I (January 1948–December 1979) and **b** phase II (January 1980–December 1996)

Results

Periodic structure of the incidence data in the analysis range

The PSDs, $P(f)$'s (f frequency), for the residual data in the analysis ranges of phases I and II were calculated. The semi-log plots of the PSDs ($f \leq 2.2$) are shown in Fig. 2a and b for phases I and II, respectively (unit of f: 1/year). For both phases, many well-defined spectral lines are clearly observed as dominant peaks in Fig. 2. Ten spectral peak-frequency modes were selected, in descending order of the power of the spectral peak, and these are summarized, with the corresponding periods and intensities (powers) of the spectral peaks, in Table 1. We calculated the powers of the PSD from integrating the PSD over the peak area.

As seen in Fig. 2, the common prominent peaks are observed at $f = 1.0$ (1 year) corresponding to an annual cycle of epidemics in both PSDs. In the PSD for phase I (Fig. 2a), the spectral line at $f = 0.08$ (12.18-year) may be related to the interval between the influenza pandemics in 1957 and 1968.

Table 1 Characteristics of the ten dominant spectral peaks shown in Fig. 2

Phase I (1948–1979)			Phase II (1980–1996)		
f	Period (year)	Power	f	Period (year)	Power
0.12	8.70[a]	0.25	0.10	10.35	0.05
0.26	3.81[a]	0.28	0.32	3.13	0.06
0.31	3.23	0.17	0.38	2.65	0.03
0.36	2.80	0.13	0.47	2.11	0.06
0.43	2.31[a]	0.33	0.63	1.60	0.06
0.64	1.56	0.14	0.72	1.39[a]	0.11
0.84	1.19	0.11	0.80	1.25	0.02
0.88	1.14	0.16	0.90	1.11	0.03
1.00	1.00[a]	3.20	1.00	1.00[a]	6.04
1.19	0.84	0.15	2.00	0.50	0.46

[a] The assigned fundamental modes

Determination of fundamental modes

The trends of the contribution ratio against the value of S are shown in Fig. 3a and b for phases I and II, respectively. The value of the contribution ratio of each S value is listed in Table 2, shown as 'a' for phase I and 'b' for phase II. For phase I (Fig. 3a; Table 2 [a]), the contribution ratios from $S = 1$ to $S = 5$ in the prediction range kept large values of around 0.6, and the values were almost the same as those in the analysis range. The contribution ratio at $S = 4$ in the prediction range had the largest value (Table 2 [a]). Thus, we assigned four periodic modes, constructing $x_{UV}(t)$ at $S = 4$ (8.70, 3.81, 2.31, and 1.00 years, as listed in Table 1). The values of the contribution ratio at $S = 4$ in the analysis and prediction ranges were 0.714 and 0.658, respectively (Table 2 [a]).

For phase II (Fig. 3b; Table 2 [b]), the contribution ratios in the analysis and prediction ranges kept large values of around 0.8–0.9 for all S values. The contribution ratio at $S = 2$ in the prediction range had the largest value. Thus, two periodic modes could be assigned as fundamental modes for the LSF curve at $S = 2$ (1.39 and 1.00 years, as listed in Table 1). The values of the contribution ratio at $S = 2$ in the analysis and prediction ranges were 0.852 and 0.874, respectively.

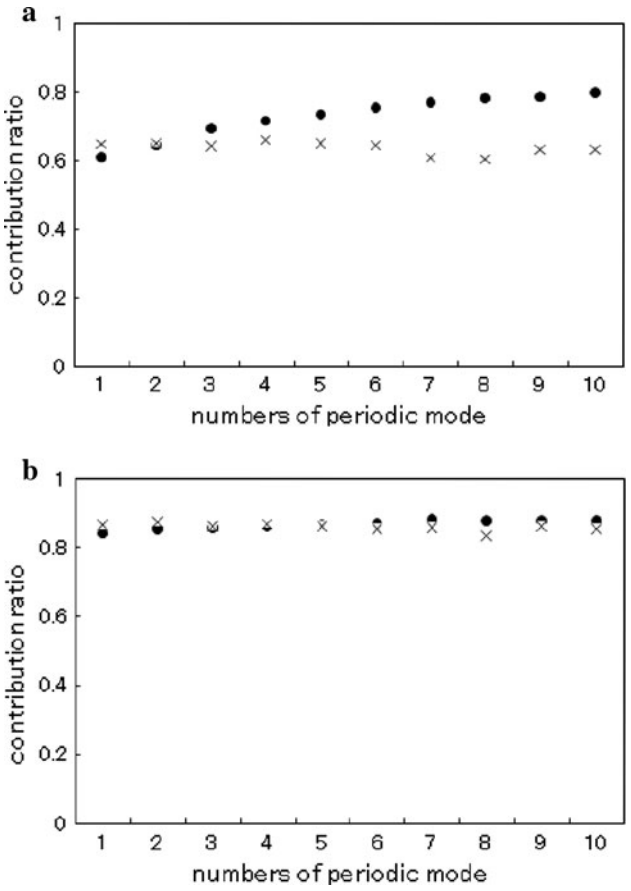

Fig. 3 Contribution ratios in the analysis and prediction ranges (*filled circles* and *multiplication symbols*, respectively): **a** phase I (January 1948–December 1979) and **b** phase II (January 1980–December 1996)

Calculation of the LSF curve

The optimum LSF curves calculated with the fundamental modes, $x_{UV}(t)$, are shown in Fig. 4a and b for phases I and II, respectively. For phase I (Fig. 4a), $x_{UV}(t)$ basically reproduces the underlying variation of $\{x_A(t)\}$, but large deviations between $x_{UV}(t)$ and $\{x_A(t)\}$ were observed in the range of 1960–1963, for example. For phase II (Fig. 4b), $x_{UV}(t)$ fairly well reproduces the underlying variation of $\{x_A(t)\}$. The values of period, amplitude and time of acrophase for the fundamental modes are listed in Table 3.

Prediction of the incidence of influenza

$x_{UV}(t)$ for phase I was extended from the analysis range (January 1948–December 1979) to the prediction range (January 1980–December 1981). As seen in Fig. 5a,

Table 2 The values of the contribution ratios shown in Fig. 3a,b

Number of periodic mode	(a) Contribution ratio		(b) Contribution ratio	
	Analysis range January 1948–December 1979	Prediction range January 1980–December 1981	Analysis range January 1980–December 1996	Prediction range January 1997–December 1998
1	0.605	0.648	0.839	0.865
2	0.643	0.648	0.852	0.874
3	0.692	0.641	0.856	0.862
4	0.714	0.658	0.861	0.867
5	0.732	0.648	0.865	0.862
6	0.753	0.643	0.869	0.852
7	0.768	0.605	0.881	0.858
8	0.780	0.602	0.875	0.832
9	0.785	0.630	0.877	0.862
10	0.798	0.631	0.877	0.852

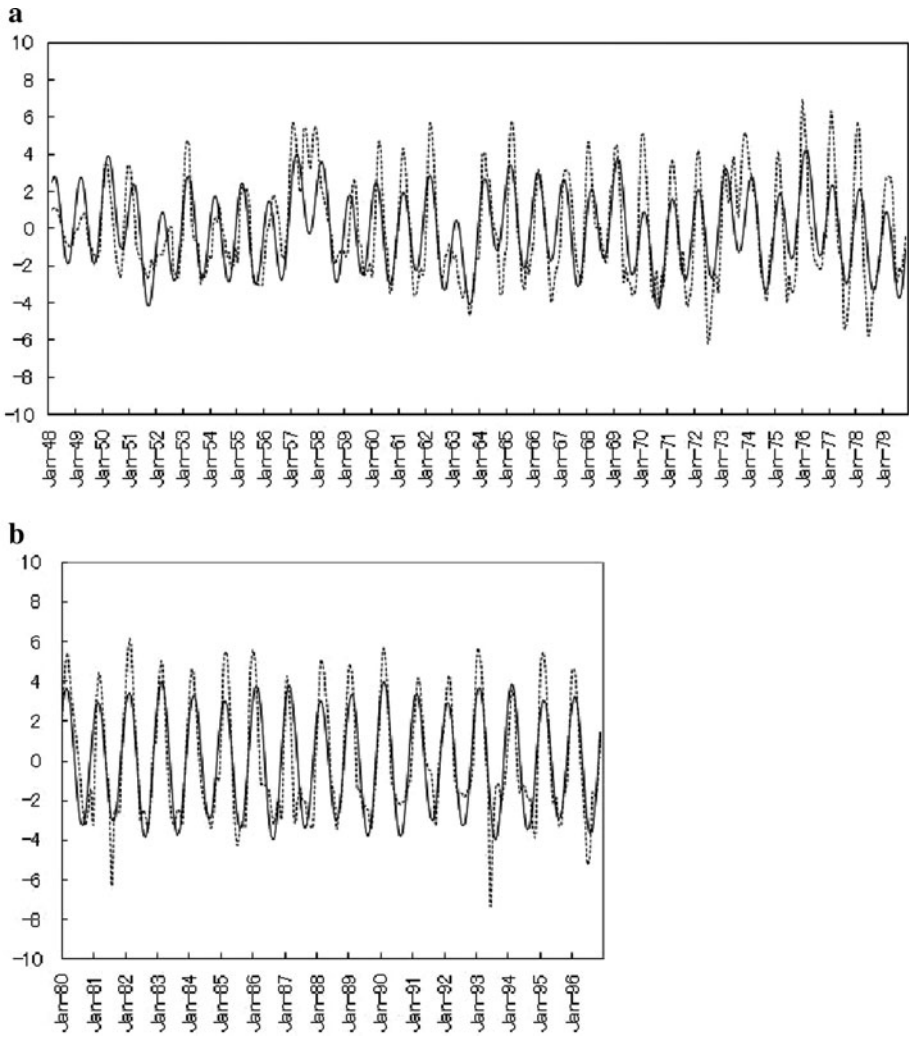

Fig. 4 Comparison of the optimum LSF curve (*solid lines*) with the residual data (*dotted lines*) in the analysis range: **a** phase I (January 1948–December 1979) and **b** phase II (January 1980–December 1996)

$x_{UV}(t)$ in the prediction range reproduces the underlying variation of the residual data well to allow the inclusion of 95% confidence intervals.

In phase II (Fig. 5b), $x_{UV}(t)$ extended from the analysis range (January 1980–December 1996) to the prediction range (January 1997–December 1998) fits within the 95% confidence interval, reproducing the underlying variation of the residual data well. For both phases I and II, almost all data points of the residual data fit within the 95% confidence interval (CI) tested by t distribution, $x(t) = Y(t) \pm t_{0.05}s$, where $Y(t)$ is given by the estimated regression line by the plotting of $x_{UV}(t)$ against $\{x_P(t)\}$, and s indicates standard error.

Effect of the long-term trend and data-length of analysis range of the incidence data on prediction analysis

To predict the residual data during January 1997–December 1998, we further investigated the value of the contribution ratio for the longer length of analysis range of the residual data: January 1948–December 1996. For this case, Fig. 6a indicates the contribution ratio versus S. The contribution ratios of the residual data in the analysis range had large values of around 0.7–0.8 for all S values, but the values were smaller than those in the shorter length of analysis range of the residual data, January 1980–December 1996, shown in Fig. 3b. In addition, we investigated the contribution ratio for the log-data that included the

Table 3 Parameters of fundamental modes

	Period (year)	Amplitude	Time of acrophase
Phase I (1948–1979)	8.70	0.65	09 March 1949
	3.81	0.82	21 January 1950
	2.31	0.75	23 May 1948
	1.00	2.47	21 March 1948
Phase II (1980–1996)	1.39	0.52	26 May 1980
	1.00	3.49	23 February 1980

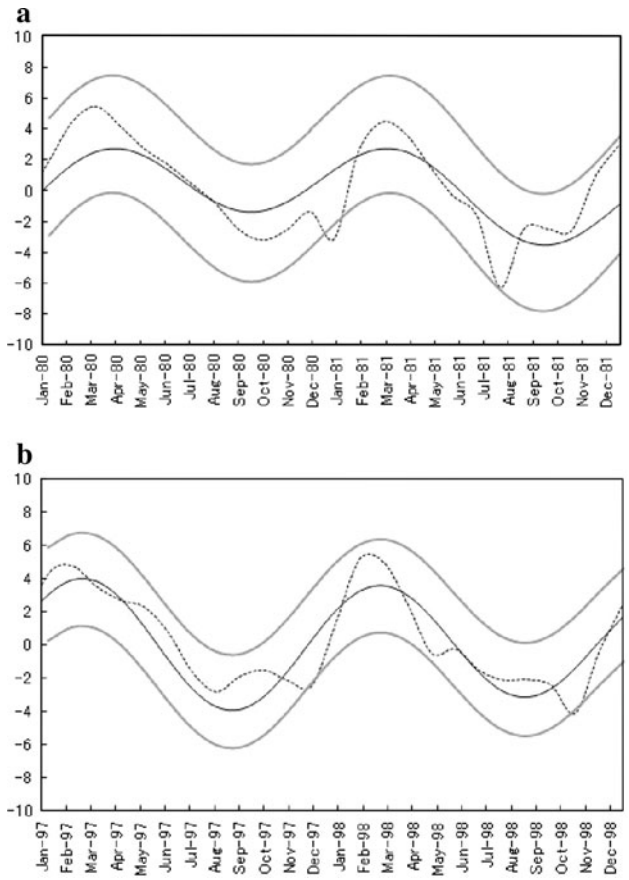

Fig. 5 Comparison of the optimum LSF curve (*solid lines*) with the residual data (*dotted lines*) in the prediction range: **a** phase I (January 1980–December 1981) and **b** phase II (January 1997–December 1998). *Gray lines*: 95% confidence interval

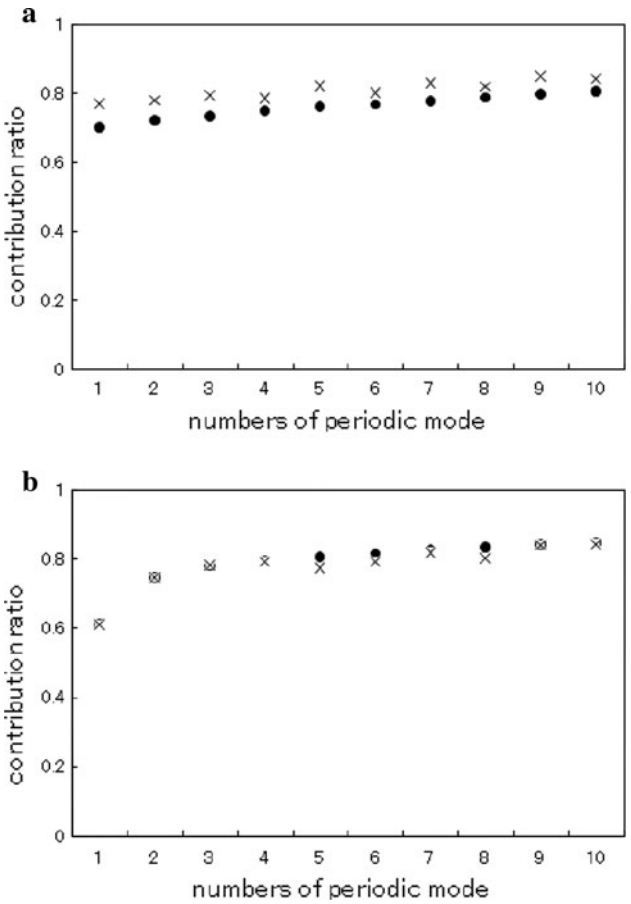

Fig. 6 Two cases of the trend of the contribution ratio for prediction of the residual data during January 1997–December 1998 (*filled circles* and *multiplication symbols*, respectively): **a** the case that the residual data during January 1948–December 1996 is used for the analysis range. **b** The case that the log-data, including long-term trend, during January 1948–December 1996 is used for the analysis range

long-term trend [27]. In this case, as seen in Fig. 6b, the values of the contribution ratio of the log-data in the analysis range (1948–1996) gradually increase as the value of S increases and keep large values of around 0.6–0.8, but the values are smaller than the those for the residual data in the analysis range (Fig. 3b). Thus, it can be said that removing the long-term trend of the log-data conducted in Step I and dividing the residual data into two phases are appropriate approaches in conducting prediction analysis for the incidence data of influenza.

Long-term predictability of influenza epidemics

For investigating the predictability of influenza incidence, we analyzed the weekly incidence data of influenza from January 1987 to October 2010 (Fig. 7a). Using the same procedure as that used for the monthly data, we set up the weekly data (Fig. 7a) for analysis. We obtained the residual data (Fig. 7b: dotted line) and divided the data into an analysis range (January 1987–December 1996) and a prediction range (January 1997–October 2010). The curve of the contribution ratio against S for the residual data in the analysis and prediction ranges is shown in Fig. 7c (filled circles and open circles, respectively). The contribution ratio of each S value is listed in Table 4. The contribution ratio in the prediction range increases with increasing S, but is smaller than the contribution ratio in the analysis range for all S values. Thus, we could not find the optimum value of S which would have been suitable to use for calculating the LSF curve. Thus, for the residual data (Fig. 7b), we investigated the contribution ratio versus S for a shorter-term prediction range (January 1997–December 1998) within the prediction range (January 1997–October 2010). The result obtained is shown in Fig. 7c (multiplication symbol) and Table 4. The contribution ratio gradually increases in the region of small S from 1 to 4 and the contribution ratio approximates to a constant in the region of S from 5 to 8. At $S = 4$, the contribution ratio in the prediction range has the largest

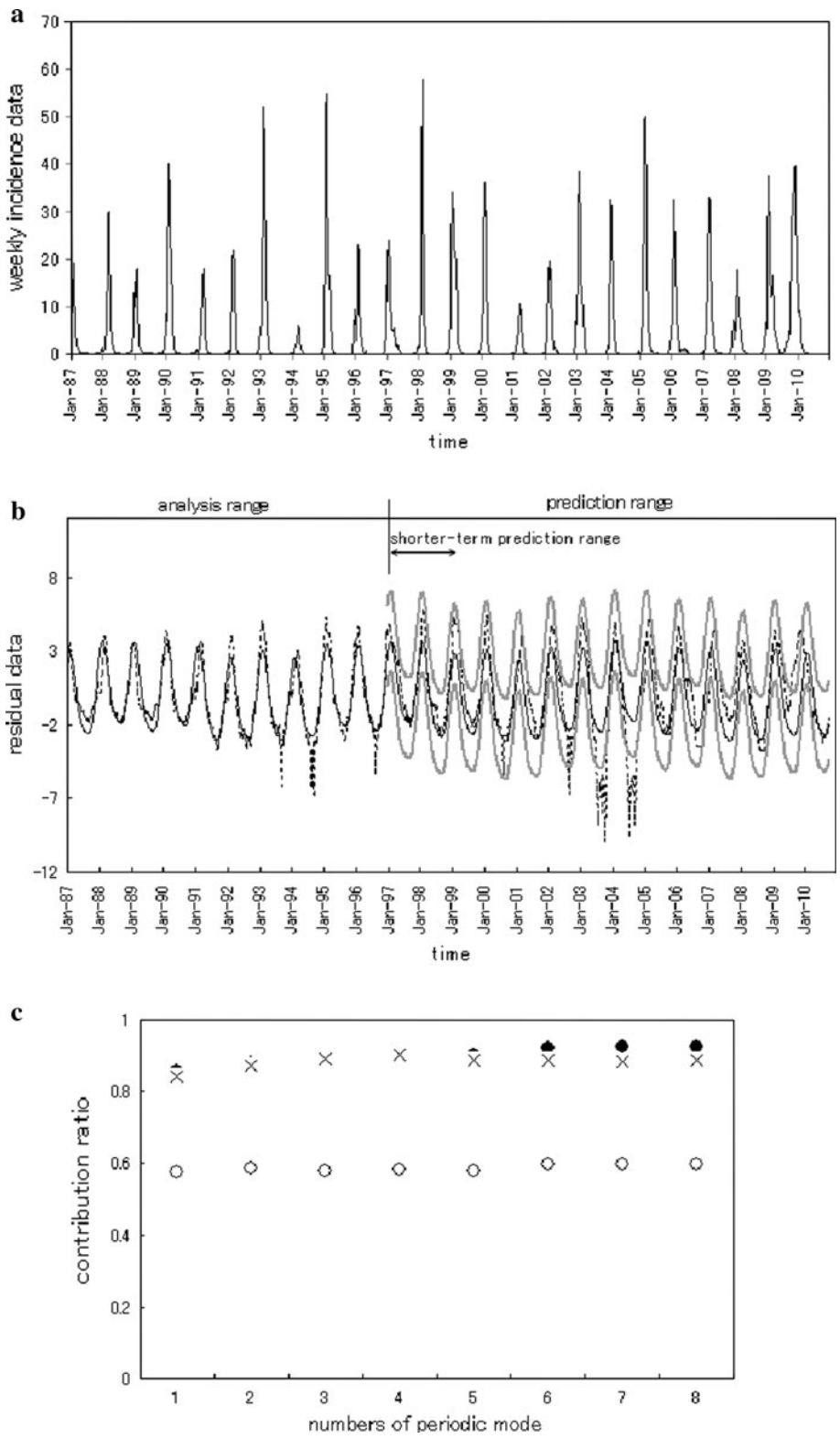

Fig. 7 Weekly incidence data of influenza in Japan from January 1987 to October 2010: **a** the original data, **b** comparison of the LSF curve (*solid line*) and the residual data of the original data (*dotted line*). *Small vertical line* indicates the boundary between the analysis and prediction ranges. *Gray lines*: 95% confidence interval. **c** Contribution ratios in the analysis range (January 1987–December 1996), prediction range (January 1997–October 2010), and shorter-term prediction range (January 1997–December 1998) (*filled circles*, *open circles*, and *multiplication symbols*, respectively)

value (0.90, as listed in Table 4). Thus, the four periodic modes used for the LSF curve could be assigned as fundamental modes at $S = 4$ (7.69, 2.31, 1.00, and 0.50 years). The LSF curve calculated with the four fundamental modes is shown in Fig. 7b (solid line). In the figure, it can be confirmed that the extension of the LSF curve to the prediction range (January 1997–October 2010) indicates short-term predictability; that is, the LSF curve reproduces the

Table 4 The values of the contribution ratios shown in Fig. 7

Number of periodic mode	Contribution ratio		
	Analysis range January 1987–December 1996	Prediction range January 1997–October 2010	Shorter-term prediction range January 1997–December 1998
1	0.856	0.574	0.840
2	0.880	0.586	0.872
3	0.893	0.580	0.893
4	0.900	0.582	0.900
5	0.903	0.580	0.888
6	0.921	0.598	0.885
7	0.924	0.596	0.882
8	0.925	0.596	0.888

underlying variation of the residual data (dotted line) in the shorter-term prediction range (January 1998–December 1999). The reason for this short-term predictability can be considered as follows: the underlying variation of the residual data in the prediction range (January 1997–October 2010) is different temporally from that in the analysis range (January 1987–December 1996).

Discussion

In the present study, we investigated periodic structures of the incidence of influenza in Japan by using MEM spectral analysis. We successfully assigned the fundamental modes constructing the underlying variations of the incidence data in the analysis range $x_{UV}(t)$ (Eq. 2) (Table 1). As a result, we obtained $x_{UV}(t)$ with good fitness to the original incidence data $x(t)$ (Fig. 4a, b), although some prominent peaks in phase I could not be reproduced well. This disagreement between $x_{UV}(t)$ and $x(t)$ (Fig. 4a) is thought to mean that the underlying variation in phase I is temporally different because of the effects of the introduction of vaccine programs (in 1962 and 1976) and the effects of influenza pandemics (in 1957 and 1968).

The dominant spectral line at the 8.7-year period in the PSD for phase I (Fig. 2a; Table 1) is approximately consistent with the period in Scotland, that is, an 8.0-year period [32]. It is widely accepted that many infectious diseases show a certain periodicity in prevalence. For example, the biennial cycle of measles epidemics has long attracted the attention of epidemiologists and mathematical biologists [33, 34]. So far, researchers have suggested that such periodicities in measles epidemics might be caused by extrinsic factors, as reflected in periodic transmission rates, e.g., seasonality, or they may be caused by time delays, age structure, or non-seasonality in incidence rates [35]. The periodic modes of 8.7 years assigned for influenza in the present study may be explained by the dynamics of pandemics (in 1957 and 1968) and epidemics of influenza [36]. That is, after a major antigenic shift resulting in a pandemic, increasing numbers of members of a population come to posses the appropriate antibodies, and subsequent epidemics, decreasing in intensity and occurring at increasing intervals, are due to minor antigenic changes. Eventually another major antigenic shift occurs, and the cycle is repeated. It is notable that, for phases I and II in our study, dominant spectral lines were observed at frequencies corresponding to periodic modes around 2–4 years, as listed in Table 1 (3.81 and 2.31 years for phase I and 3.13 and 2.11 years for phase II, for example). Based on the study of Plotkin et al. [37], it can be considered that periodic modes around 2–4 years for phases I and II might correspond to the average duration of cross-reactive immunity against the influenza A virus.

In the present study, the precise determination of fundamental modes constructing the underlying variations of the incidence data of influenza enabled us to conduct prediction analysis (Figs. 5, 7). The prediction curve of the incidence data could be quantitatively indicated by the extension of $x_{UV}(t)$ to the prediction range. This reproducibility for the incidence data was considered to have come about for the following reason: the fundamental modes constructing $x_{UV}(t)$ (Table 1) were substantially well assigned by MEM spectral analysis and reconstruct the periodic structure of the underlying variation of the data in the prediction range. It is anticipated that the present method of time series analysis consisting of MEM spectral analysis and LSM will contribute to further development in the field of prediction analysis of epidemics of influenza.

Acknowledgments This study was supported by a Grant-in-Aid for Scientific Research (No. 20590609) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Appendix

On maximum entropy method (MEM) spectral analysis

The power spectral density (PSD) obtained by MEM spectral analysis for time series data under analysis, with equal sampling interval Δt (=1 month, in the present study), can be calculated from

$$P_m(f) = \frac{P_m \Delta t}{\left| 1 + \sum_{k=-m}^{m} \gamma_{m,k} \exp[-i2\pi f k \Delta t] \right|^2}, \quad \text{(A1)}$$

where P_m is the output power of a prediction-error filter of order m and $\gamma_{m,k}$ the corresponding filter coefficient, $m = 0, 1, 2, \ldots, M$; M is the optimum filter order. P_m and $\gamma_{m,k}$ are determined by solving the following Yule-Walker equations with the use of Burg's procedure:

$$\begin{bmatrix} C_0 & C_1 & \cdots & C_m \\ C_1 & C_0 & \cdots & C_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_m & C_{m-1} & \cdots & C_0 \end{bmatrix} \begin{bmatrix} 1 \\ \gamma_{m,1} \\ \vdots \\ \gamma_{m,m} \end{bmatrix} = \begin{bmatrix} P_m \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \text{(A2)}$$

where C_k ($k = 0, 1, \ldots, m$) is autocorrelation function of time series data $\{x(k\Delta t)\}$ described by

$$C_k = \frac{1}{N} \sum_{i=1}^{N-|k|} \{x(i+k) - \mu\}\{x(i) - \mu\}, \quad \text{(A3)}$$

where μ is the mean value of time series data, and N the length of time series data. By setting $m = M$, we obtain $P(f)$ ($=P_M(f)$) from Eq. A1.

On the determination of the value of M

In the present study, the value of M for the residual data was determined on the basis of the investigation of the lag dependence of MEM-estimated periods. The large value of M for calculating MEM–PSD is necessary for extracting the fundamental mode from the time series data, and recently this has been supported theoretically [38].

On the determination of the value of the "contribution ratio"

The determination of S in Eq. 2 is made via the following procedure. Based on the result of periods estimated by MEM spectral analysis, we must assign fundamental modes f_n that construct the periodic structure of $x_{UV}(t)$ of $\{x_A(t)\}$ (Eq. 2). Then, we investigate the contribution of ten dominant periods estimated by MEM spectral analysis to the LSF curve in the analysis and prediction ranges: (a) the LSF curve in the analysis range is calculated with the variation S, by the ten modes being added to the LSF curve one by one in the order of magnitude of the power of the spectral peak frequency, (b) the LSF curve calculated with each S is extended to the prediction range, and (c) the evaluation of the LSF curve is performed. Procedure (c) is divided into four steps [(c)-1, (c)-2, (c)-3, and (c)-4]. In procedure (c)-1, the power of each periodic mode is evaluated by the square of amplitude, A_i^2, of the ith mode constituting the LSF curve [28]. In procedures (c)-2 and (c)-3, we estimate R_j corresponding to the power of time series, which is obtained by subtracting the LSF curve from the original time series. As a result, the total powers of the original time series in the analysis and prediction ranges (Q_A and Q_P, respectively) are obtained by

$$Q_j = \sum_{i=1}^{S} A_i^2 + R_j, \quad j = \begin{cases} \text{A}: \text{ analysis range} \\ \text{P}: \text{ prediction range} \end{cases} \quad \text{(A4)}$$

When both sides of Eq. A4 are divided by Q_j, we obtain the following normalized relation:

$$\frac{\sum_{i=1}^{S} A_i^2}{Q_j} + \frac{R_j}{Q_j} = 1, \quad j = \begin{cases} \text{A}: \text{ analysis range} \\ \text{P}: \text{ prediction range} \end{cases} \quad \text{(A5)}$$

where $\sum_{i=1}^{S} A_i^2 \big/ Q_j$ and R_j/Q_j correspond to the contribution of $\sum_{i=1}^{S} A_i^2$ and R_j to Q_j, respectively. Then, in procedure (c)-4, we define the first term of the left-hand side of Eq. A5, the "contribution ratio", which means the contribution $\sum_{i=1}^{S} A_i^2$ normalized by Q_j. If $\sum_{i=1}^{S} A_i^2 \big/ Q_j$ in the first term becomes large, then the second term R_j/Q_j becomes small.

References

1. Cliff A, Haggett P, Smallman-Raynor M. An exploratory method for estimating the changing speed of epidemic waves from historical data. Int J Epidemiol. 2008;37:106–12.
2. Brachman PB. Surveillance. In: Evans AL, Brachman PS, editors. Bacterial infections of humans: epidemiology and control. 2nd ed. New York: Plenum Medical Book Co.; 1991. p. 49–61.
3. Cliff AD, Haggett P. Statistical modelling of measles and influenza outbreaks. Stat Methods Med Res. 1993;2:43–73.
4. MacMahon B, Pugh TF. Epidemiology: principles and methods. Boston: Brown Company; 1970.
5. Mugglin AS, Cressie N, Gemmell I. Hierarchical statistical modeling of influenza epidemic dynamics in space and time. Stat Med. 2002;21:2703–21.
6. Choi K, Thacker SB. An evaluation of influenza mortality surveillance, 1962–1979. I. Time series forecasts of expected pneumonia and influenza deaths. Am J Epidemiol. 1981; 113:215–26.
7. Crighton EJ, Mineddin R, Mamdani M, Upshur REG. Influenza and pneumonia hospitalizations in Ontario: a time series analysis. Epidemiol Infect. 2004;132:1167–74.

8. Kakehashi M, Tsuru S, Seo A, Amran A, Yoshinaga F. Statistical analysis and prediction on incidence of infectious diseases based on trend and seasonality. Jpn J Hyg. 1993;48:578–85.
9. Pease CM. An evolutionary epidemiological mechanism, with applications to type A influenza. Theor Popul Biol. 1987; 31:422–52.
10. Quénel P, Dab W. Influenza A and B epidemic criteria based on time-series analysis of health services surveillance data. Eur J Epidemiol. 1998;5:285–93.
11. Urashima M, Shindo N, Okabe N. A seasonal model to simulate influenza oscillation in Tokyo. Jpn J Infect Dis. 2003;56:43–7.
12. Šaltytė Benth J, Hofoss D. Modelling and prediction of weekly incidence of influenza A specimens in England and Wales. Epidemiol Infect. 2008;7:1–9.
13. Liu W, Levin SA. Influenza and some related mathematical models. In: Levin SA, Hallan TG, Gross LJ, editors. Applied mathematical ecology. New York: Springer; 1989. p. 235–52.
14. Mills CE, Robins JM, Lipsitch M. Transmissibility of 1918 pandemic influenza. Nature. 2004;432:904–6.
15. Stillianakis NI, Perelson AS, Hayden FG. Emergence of drug resistance during an influenza epidemic: insights from a mathematical model. J Infect Dis. 1998;177:863–73.
16. Xu Y, Allen LJS, Perelson AS. Stochastic model of an influenza epidemic with drug resistance. J Theor Biol. 2007;248:179–93.
17. May RM. The chaotic rhythms of life. In: Hall N, editor. The New Scientist guide to chaos. London: IPC Magazines New Scientist; 1991. p. 82–94.
18. Sumi A, Kamo K, Ohtomo N, Kobayashi N. Study of the effect of vaccination on periodic structures of measles epidemics in Japan. Microbiol Immunol. 2007;51:805–14.
19. Sumi A, Hemilä H, Mise K, Kobayashi N. Predicting the incidence of human campylobacteriosis in Finland with time series analysis. APMIS. 2009;117:614–22.
20. Ohtomo K, Kobayashi N, Sumi A, Ohtomo N. Relationship of cholera incidence to El Ninõ and solar activity elucidated by time-series analysis. Epidemiol Infect. 2009;19:1–9.
21. Sumi A, Ohtomo N, Tanaka Y. Study on chaotic characteristics of incidence data of measles. Jpn J Appl Phys. 1997;36:7460–72.
22. Sumi A, Ohtomo N, Tanaka Y, Koyama A, Saito K. Comprehensive spectral analysis of time series analysis of time series data of recurrent epidemics. Jpn J Appl Phys. 1997;36:1303–18.
23. Sumi A, Olsen LF, Ohtomo N, Tanaka Y, Sawamura S. Spectral study of measles epidemics: the dependence of spectral gradient on the population size of the community. Jpn J Appl Phys. 2003;42:721–33.
24. Sumi A, Ohtomo N, Tanaka Y, Sawamura S, Olsen LF, Kobayashi N. Prediction analysis for measles epidemics. Jpn J Appl Phys. 2003;42:7611–20.
25. Sumi A. Time series analysis of surveillance data of infectious diseases in Japan. Hokkaido J Med Sci. 1998;73:343–63.
26. Cliff A, Haggett P, Smallman-Raynor M. Measles: an historical geography of a major human viral disease from global expansion to local retreat, 1840–1990. Oxford: Blackwell; 1993.
27. Sumi A, Kamo K, Ohtomo N, Mise K, Kobayashi N. Time series analysis of incidence data of influenza in Japan. J Epidemiol. 2011;21:21–9.
28. Statistics and Information Department, Minister's Secretariat, Ministry of Health and Welfare in Japan. Statistics of communicable diseases. Tokyo: Health and Welfare Statistics Association; 1968–1999 (in Japanese).
29. National Institute of Infectious Diseases. Infectious diseases weekly report. http://idsc.nih.go.jp/idwr/index.html (1987–2009). Accessed 9 May 2011 (in Japanese).
30. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th ed. Oxford: Blackwell, Science; 2002.
31. Gershenfeld NA, Weigend AS. The future of time series: learning and understanding. In: Weigend AS, Gershenfeld NA, editors. Time series prediction: forecasting the future and understanding the past. New York: Addison-Wesley; 1994. p. 1–70.
32. Clegg EJ. Infectious disease mortality in two Outer Hebridean islands: 1. measles, pertussis and influenza. Ann Hum Biol. 2003;30:455–71.
33. Anderson RM, Grenfell BT, May RM. Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. J Hyg Camb. 1984;93:587–608.
34. Noah ND. Cyclic patterns and predictability in infection. J Hyg Camb. 1989;102:175–90.
35. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. London: Oxford University Press; 1991.
36. Kilbourne ED. Epidemiology of influenza. In: Kilbourne ED, editor. The influenza viruses and influenza. New York: Academic Press; 1973. p. 483–538.
37. Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci USA. 2002;99:6263–8.
38. Tokiwano K, Ohtomo N, Tanaka Y. Saidai Entropy-ho ni-yoru Jikeiretsu Kaiseki: Memcalc no Riron to Jissai (Time series analysis by maximum entropy method: theory and practice of MemCalc). Sapporo: Hokkaido University Press; 2002. (in Japanese).