Quaderni di Statistica
Vol. ?, 200?

An analysis of the long-term cycle in the spread of measles in Italy

Eugene M. Cleur
Facoltà di Economia, Università di Pisa
E-mail: cleur@ec.unipi.it

Summary: The number of reported cases of measles, in the pre-vaccination period in large communities, are generally characterized by the presence of a seasonal component and a long-term cyclical component which ranges, in general, between two and four years and which is better known as the inter-epidemic period. Researchers tend to resort to causal mathematical models in order to explain the inter-epidemic period, paying less attention, with some rare exceptions, to stochastic statistical models. This paper illustrates how simple time series autoregressive moving average models could provide a better fit to the long-term dynamics of the spread of measles and then proceeds to an analysis of the long term component in each series in an attempt to discover whether clusters of Regions having the same time path can be defined.

Keywords: Infectious diseases, Compartmental mathematical models, Autoregressive moving average models, Parametric estimate of spectrum.

1. Introduction

Many time series reporting the cases of childhood infectious disease on a weekly or monthly basis exhibit a trend, a 12 month cycle and a longer cycle, which may be taken as a proxy for the so called *inter-epidemic period*, of varying length; bi-weekly data for English and Welsh urban centres exhibit a surprisingly regular two year cycle (see Bjornstad et al. (2002), Grenfell et al. (2002)), whereas Italian data collected for each of the 20 Regions appears to have an *inter-epidemic period* which, on average, is slightly longer than three years (see Manfredi et al. (2002)). Many explanations are possible for these differences, beginning from the different geographical areas considered, but before that can be seriously done, it is fundamental that a better understanding of the Italian data be gained. In Manfredi et al. (2002) an analysis of the cyclical behaviour of the monthly regional time series was carried out. In this paper the long-term cycle is scrutinised in more detail by means of time series analyses carried out mostly in the frequency domain.

Before being analysed, the series were log transformed, then detrended using a deterministic function of time and finally deseasonalised using monthly dummies; several tests, including those by Dicky and Fuller (1979) and by Philips and Perron (1988) rejected the unit root hypothesis. A value of one was added to each data point before the log transformation to avoid the problem of log(zero) for those months, few in number, when no cases were reported. A numerical illustration which justifies the detrending and deseasonalisation methods used here may be found in Cleur et al.(2003).

We begin by identifying the peak frequencies of the long-term cycle through a non-harmonic Fourier analysis and then compare these results with those provided by a deterministic compartmental mathematical model of the SEIR type (Manfredi et al. (2002)). The paper then proceeds with the estimation of univariate autoregressive (AR) models and their transformation to the frequency domain in order to obtain the parametric spectra. The ability of the parametric spectra to reproduce the cycles identified through the non-harmonic Fourier analysis is taken as a criterion for evaluating these time series models. Finally, the long-term components are estimated via the demodulation-remodulation technique and an attempt is made to classify their evolution in time into distinct groups.

This study considers only the pre-vaccination data, i.e. January 1949 to December 1976, for the simple reason that the true dynamics of the disease are not affected and altered by an external regulator such as vaccination. The Italian regions are treated as separate epidemiological units.

2. Identification of the peak frequency of the inter-epidemic period

Research into the transmission dynamics of infectious diseases such as measles has tended to favour a deterministic mathematical models approach focusing on the causal mechanisms underlying the infection process.

Such an approach relies on deterministic compartmental models (Anderson and May, 1991) often having the SEIR structure by which individuals who experience the infection are assumed to move from the susceptible (S) state to the exposed state (E), in which individuals are infected but not yet infectious, then to the infected (I) state, in which individuals are capable of transmitting the infection, and finally recover from the disease, by entering the removed (R) state, in which they are permanently immune. Statistical modelling, on the other hand, has rarely been taken into account with very few significant exceptions such as Anderson et al. (1984), Bjornstad et al. (2002), Finkenstadt et al. (1998a, 1998b, 2000) and Grenfell et al. (2001, 2002) in which demographic variables such as age structures, birth rates and population size play a fundamental role in explaining the dynamics of the number of infected cases.

A simple form of the SEIR model, for details see Manfredi et al. (2002), provided among other results estimates of the inter-epidemic period, reported in the first column of *Table 1*, for each of the 20 Italian Regions. These results, in contrast with those obtained for England and Wales, reveal dynamics which are not common to all Regions and this makes it extremely difficult to understand the mechanisms which generate the spread of the disease.

Table 1. *Periodicity (in months) of long term cycle identified by non - parametric spectral analysis, by direct estimation of the peak frequency, and by the parametric estimate of the spectrum.*

	Periodicity of long-term cycle				Lag AR model
	SEIR model 1971-76	Nonparamet spectrum	Direct method	Parametric spectrum	
Val d'Aosta	37.2	42	38.55	38.57	27
Piemonte	37.2	30	28.69	29.92	30
Lombardia	34.8	30	28.43	28.43	9
Veneto	36.0	36-42	38.55	38.55	8
Trentino	34.8	36	34.15	34.33	23
Friuli	36.0	36	33.24	33.78	33
Emilia-Rom.	37.2	42	40.02	40.02	32
Liguria	39.6	60	64.11	58.72	36
Toscana	39.6	36	34.15	34.15	17
Umbria	38.2	42	40.80	41.07	16
Marche	38.2	42	39.27	39.27	23
Lazio	37.2	42	38.79	40.80	7
Abruzzo	37.2	48	40.80	40.27	25
Molise	36.0	60	70.60	77.57	14
Campania	34.8	36	36.32	35.70	34
Puglia	33.6	48	45.53	45.53	35
Basilicata	33.6	48	44.88	30.65	17
Calabria	34.8	36-42	41.34	40.54	34
Sicilia	33.6	36	37.62	37.62	30
Sardegna	34.8	48	45.53	45.53	27

As against this procedure, based on causal relationships, results from a univariate time series approach based on spectral analysis were also reported in Manfredi et al. (2002). Briefly, it was observed that the non-parametric estimate of the spectral density for each series was concentrated in three narrow frequency bands: (in order of magnitude) around frequency $\pi/6$ corresponding to the annual cycle, around zero frequency which was identified as being due to a gradually decreasing trend and, finally, around a frequency with periodicity varying between 3 and 5 years. As often happens, the spectral analysis was able to establish only tentatively the peak frequency at which the long-term cycle was present. A more precise knowledge of the peak frequency is obtained in this paper via the following non-harmonic analysis (see Damsleth and Spjotvoll (1982)) instead of the classical periodogram analysis which assumes that the periodicities of the cyclical components are dependent on the length of the series, *T*, an assumption not often confirmed by empirical data.

Suppose that λ is the true peak frequency of a cycle in the series analysed, X_t ; since λ need not be a function of the length of the series analysed, the cycle often corresponds to a non harmonic function. Then one might estimate λ by minimizing the quantity

$$Q(\alpha, \beta, \lambda) = \sum_t [X_t - \mu - \alpha\cos(\lambda t) - \beta\sin(\lambda t)]^2 \qquad (1)$$

Where μ is the mean of the series, α and β are unknown Fourier coefficients and λ is the desired peak frequency.

The spectral density may be used for identifying a starting value for λ. This procedure encounters problems if many cycles are present at nearby frequencies. For this reason, the procedure followed in this paper was to apply ordinary least squares in minimizing (1) with respect to α and β, over a grid of frequencies centred around the starting value suggested by the spectral density estimate. This grid search method finally provides estimates of λ as well as of the corresponding α and β. As can be seen from *Table 1* the starting value, in column two, and the final result, in column three, are not too far apart. The peak frequencies correspond to periods varying from approximately 28 months in Piemonte to 70 months in Molise; Molise is a rather special case due to it being a very small region with a very sparsely distributed population. On the other hand, in most cases there is little agreement between the periodicity identified this way and that predicted by the SEIR model which consequently obliges us to conclude that the mathematical modelling carried out, although useful for causal analysis, needs to be refined if we are to have a clearer understanding of the phenomena.

Having ascertained that a deterministic compartmental mathematical model of the SEIR type very often failed in correctly predicting the long-term cycle (it should be stated that this was not one of the objectives of that analysis although, *a posteriori*, it evidences one of the shortcomings of the estimated mathematical model), we proceed with an investigation into the use of AR models as a more reliable alternative. The question to be answered is whether AR models can provide a better prediction of the long-term cycle.

3. AR models for the inter-epidemic period.

For investigating the capability of an AR model in predicting the presence of a long-term cycle in the spread of measles, resort is again made to a frequency domain analysis and, in particular, to the parametric spectrum (see Hamilton (1994), Priestley (1981)) corresponding to the estimated AR model. If the estimated parametric spectrum indicates the possible presence of a cyclical component which has a periodicity near that identified above from an application of the Damsleth and Spjotvoll (1982) procedure, we will conclude that the parametric model for which the spectrum was calculated provides a reliable prediction of the inter-epidemic period.

The models have been estimated from the pre-whitened (detrended and deseasonalised) data and the criterion for determining the order is the following:

$$\underset{p}{Min}(dm - arsp(p))^2$$

where *dm* is the periodicity determined by the Damsleth and Spjotvoll method and *arsp(p)* is the periodicity corresponding to the peak frequency in the autoregressive spectrum calculated from an autoregressive model of lag *p*. The maximum lag was arbitrarily set to 36. The relevant results are summarised in the last two columns of *Table 1*.

The parameter estimates and their standard errors in each AR model are not reported here in order avoid problems connected with the presentation of a notable amount of information in a single Table and limited space, but they may be obtained shortly from the authors web site www.cleur.ec.unipi.it.

Since the parametric spectrum may be estimated at any frequency in the interval [0, π] this makes it possible to identify the peak frequency up to any desired degree of precision.

As can be seen from *Table 1*, the long-term cycles reproduced by the spectra of the identified AR models, with the exceptions of Liguria and Molise, are in close agreement with those identified in each series using a direct estimate of the peak frequency. This suggests that, if good fitting is an objective of the analysis, then high order AR models, although lacking the capacity to explain causal relationships, could provide a good alternative to the more popular mathematical models used in this field of research. We could also reasonably go on to conclude that given the better prediction of the long-term cycle, the AR models should provide better forecasts for the individual series.

4. Isolating the long-term cycle

Attention will now be given to a more detailed analysis of the long-term component in each series. To do so we must first be able to extract this component through filtering.

It is widely accepted that cyclical behaviour in phenomena like the spread of measles is never regular, but varies in periodicity as well as in amplitude; for instance, the annual cycle identified and estimated by time series methods is a sort of average of oscillations with periodicity varying around 12 months, i. e. in a frequency band around frequency $\pi/6$. Hence when attempting to estimate the annual cycle, we are never interested in the component at frequency $\pi/6$, but rather in components varying in a narrow band of frequencies around frequency $\pi/6$. The same is true of the long-term cycle. This objective may be realised by applying demodulation-remodulation techniques (see Granger and Hatanaka (1964) for a clear and simple description). Briefly the method consists in the following:

Suppose we are interested in isolating the component around frequency λ_0 in the series X_t. First demodulate by forming the series

$$\mathrm{Y}_t^c = X_t \cos(\lambda_0 t) \qquad \text{and} \qquad \mathrm{Y}_t^s = X_t \sin(\lambda_0 t)$$

Next, filter the series Y_t^c and Y_t^s using a low pass filter, $F[.]$, with a narrow bandwidth to form the series

$$\widetilde{Y}_t^c = F[\mathrm{Y}_\mathrm{t}^\mathrm{c}] \qquad \text{and} \qquad \widetilde{Y}_t^s = F[\mathrm{Y}_\mathrm{t}^\mathrm{s}]$$

The quantity

$$f_t(\lambda_0) = 2[(\widetilde{Y}_t^c)^2 + (\widetilde{Y}_t^s)^2]$$

is the instantaneous power of the spectrum of $\mathrm{X_t}$, at time point t, and is useful for discovering changes in the power of the spectrum over time.

An estimate of the desired component may be obtained by remodulating, i. e. by forming

$$\mathrm{X_t}(\lambda_0) = 2\widetilde{Y}_t^c \cos(\lambda_0 \mathrm{t}) + 2\widetilde{Y}_t^s \sin(\lambda_0 \mathrm{t})$$

Often moving averages are used in defining the low pass filter. This has the big disadvantage of resulting in losses of averages at the beginning and end of the series. This problem is overcome here by applying kernel smoothing as a low pass filter. In particular, if $\mathrm{X_t}$ is to be smoothed, the following kernel smoothing is carried out (see Wand et al.(1995), Shumway(2000)):

$$y_t = \sum_{t=1}^{T} W_t(i) X_t$$

where $$W_t(i) = K\left(\frac{t-i}{b}\right) \Big/ \sum_{j=1}^{T} K\left(\frac{t-j}{b}\right)$$

is the Naradaya-Watson density estimator with *i, t* = 1, 2, … , *T*; $K(\cdot)$ is the standard normal kernel function with *b*= 40 which was set after experimenting a number of values.

INSERIRE TABELLA n.2

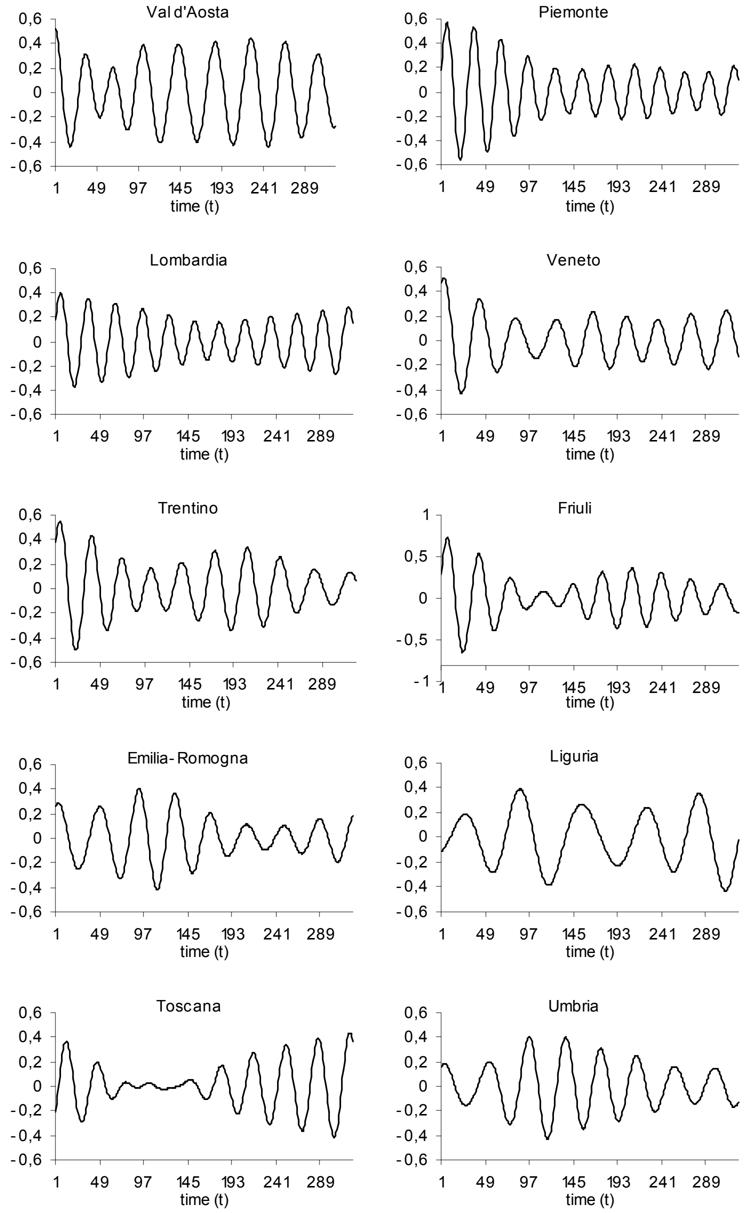

Figure 1. Long-term components estimated by demodulation-remodulation. Northern and Central Italian Regions.

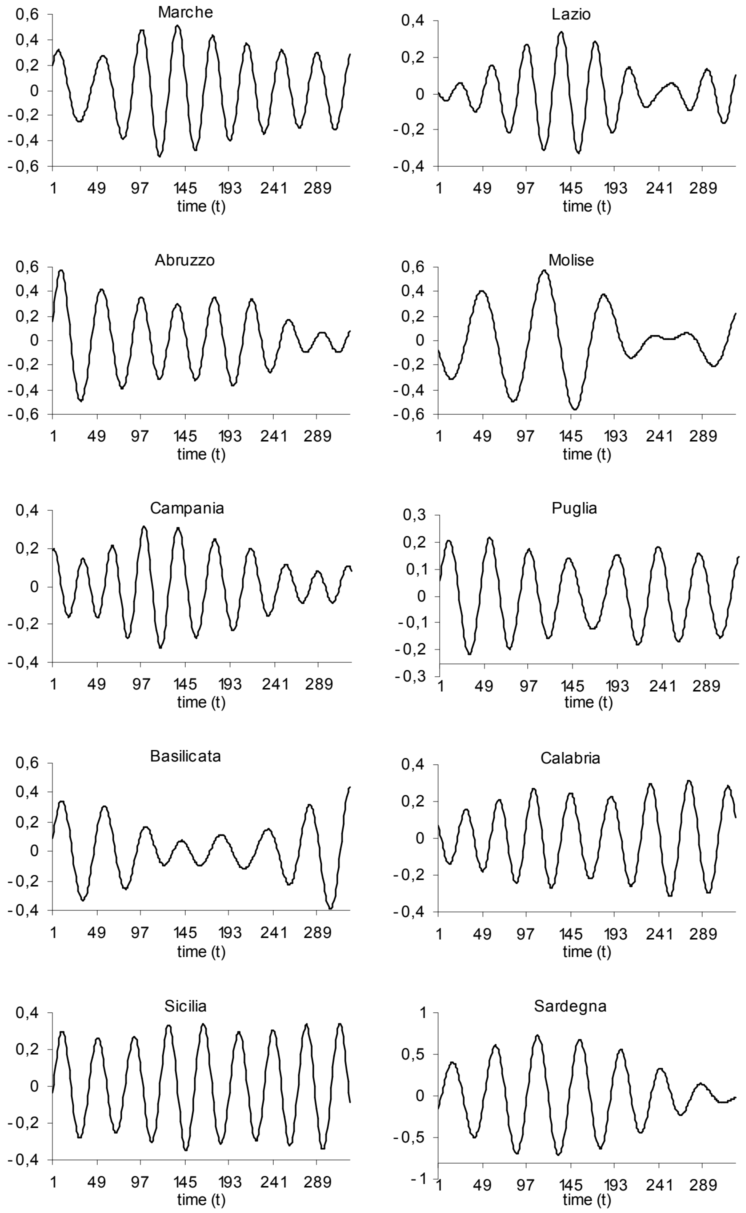

Figure 1 (contd.). Long-term components estimated by demodulation-remodulation. Central and Southern Italian Regions.

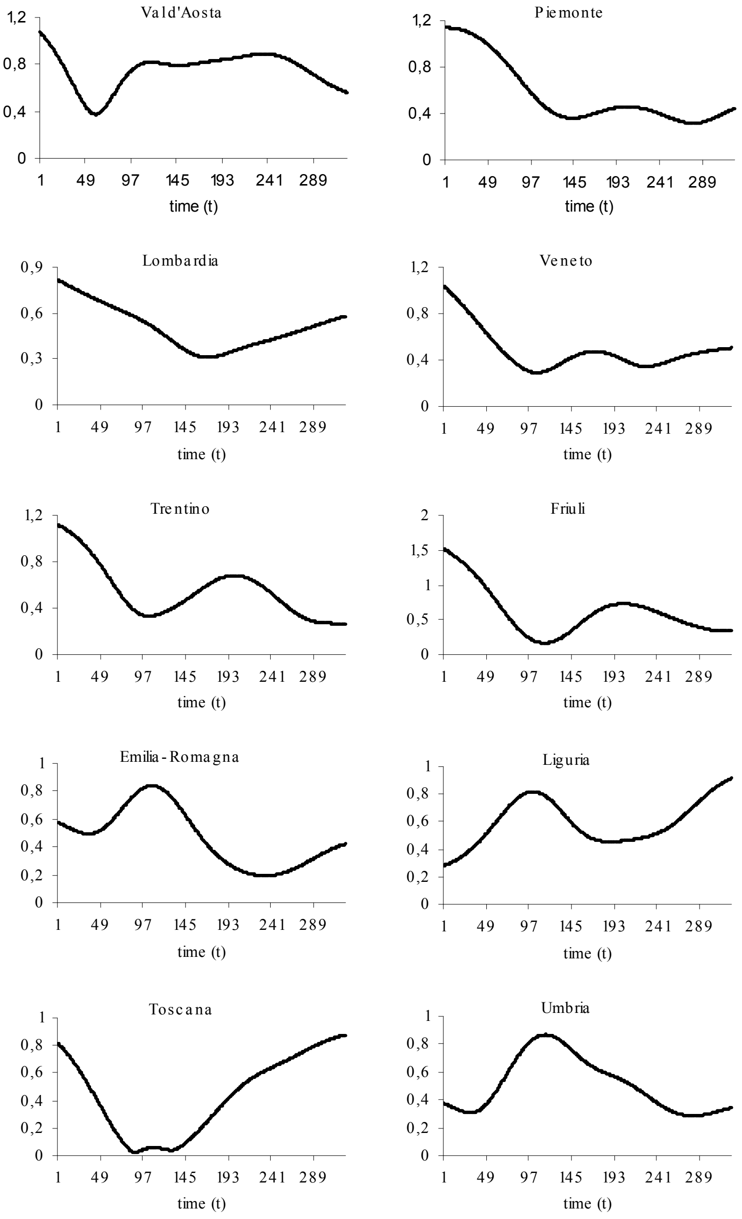

Figure 2. Instantaneous Spectrum of long-term component. Northern and Central Regions

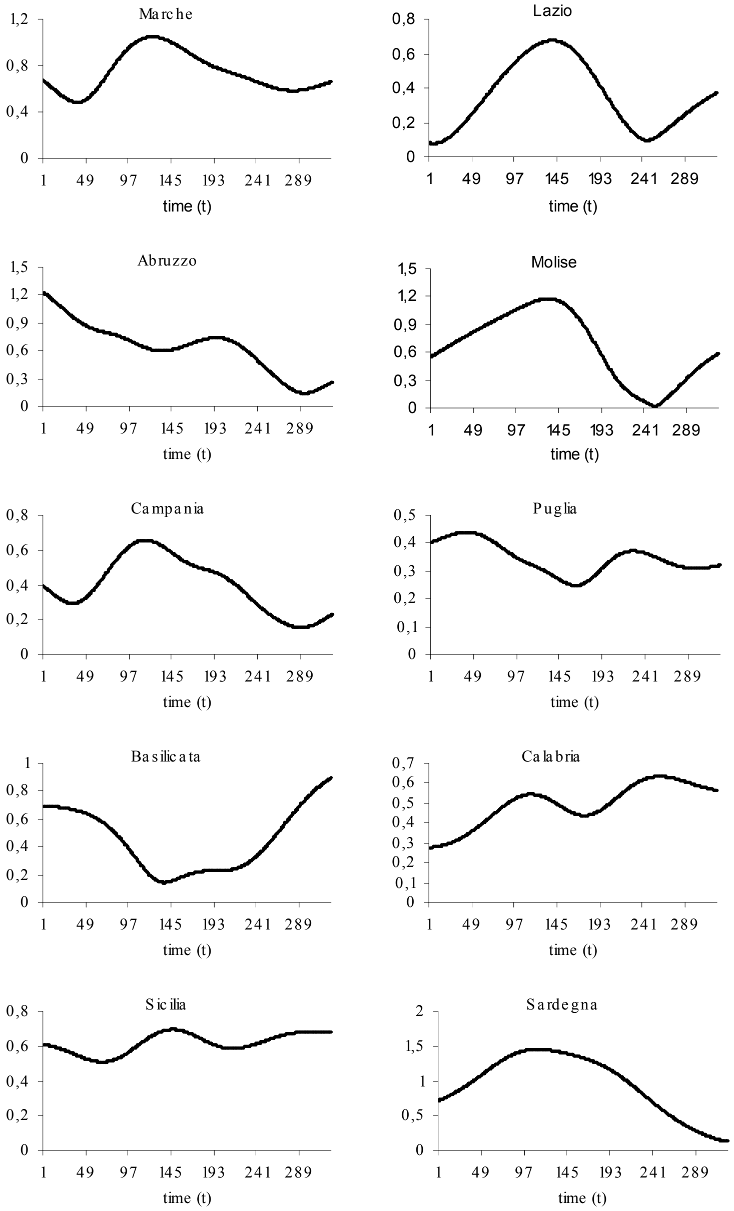

Figure 2 (contd). Instantaneous Spectrum of long-term component. Central and Southern Regions

The estimated long-term components in each regional series and the corresponding instantaneous spectra are shown in *Figures 1* and *2* respectively. Comments on single graphs would be of little use and will therefore be avoided. The amplitudes of the long-term cycles are clearly non stationary and do not follow the same pattern for all the Regions. However, from a visual examination of the graphs in *Figures 1* and *2*, where t=1 corresponds to January 1949, and from the correlation matrix of the long-term components reported in *Table 2*, the following similarities may be observed:

- Umbria, Marche and Lazio, form the core of a cluster of Regions also comprising Abruzzo, Emilia-Romagna and Campania;
- a smaller cluster of Regions is formed by Friuli and Trentino together with Veneto;
- there are also two isolated high correlations between Piemonte and Lombardia (0.91) and between Puglia and Basilicata (0.83).

A further observation may be made on examining *Figure 1*. It appears that when the number of reported cases reaches an unusual (for the particular series) low (see the trajectories for Friuli, Emilia-Romagna, Toscana, Molise), it takes some time before the phenomenon returns to it's previous level. Such a behaviour often suggests dynamics that are amplitude dependent.

Confirmation of the similarities observed above is given by a principal components analysis carried out on the correlation matrix of the estimated long-term components. There is no universally established and reliable method for determining the number of principal components to take into consideration, so the widely used empirical rule of considering only those which together account for approximately 80% of the total variance will be applied here; the first five principal components account for 76% of the total variance. The correlations between each of these components and the 20 Regions are reported in *Table 3* and are useful for purposes of identification. In fact, paying attention to correlations which are greater than 0.70, this too is a common practice, leads to the following conclusions: the first principal component (27% of total variation) may be associated with the Regions Umbria, Marche, Lazio, Abruzzo, Campania and Emilia-Romagna, and the second principal component (15% of total variation) with Regions Veneto and Friuli (the correlation with Trentino is just below 0.70). In other words, the first two principal components in order of importance identify the same cluster of Regions obtained above from an examination of the correlation matrix between the estimated long-term components and the visual examination of the instantaneous spectra and time paths of the long-term component in each Region. If we ignore Emilia-Romagna, we may conclude that the analysis carried out so far has identified a relationship in the long-term spread of measles in nearby areas.

Table 3. Correlations between first 5 principal components and the Regions (p.c. = principal component)

Region	p.c. 1	p.c. 2	p.c. 3	p.c. 4	p.c. 5
Val d'Aosta	-0,0104	-0,0965	0,8238	-0,1645	-0,0406
Piemonte	0,3042	0,4606	0,585	0,0569	0,0853
Lombardia	0,2971	0,4058	0,5527	0,0578	0,0896
Veneto	0,107	0,9006	-0,1564	0,0692	-0,0321
Trentino	0,5087	0,695	0,1961	0,0605	0,0298
Friuli	0,3117	0,7394	0,1568	0,0549	0,0061
Emil-Rom	0,784	0,1958	-0,4251	-0,0446	0,0557
Liguria	0,0309	-0,0766	-0,0345	-0,1326	0,8922
Toscana	0,4576	-0,1793	-0,1912	0,2314	0,0393
Umbria	0,8749	-0,2055	-0,0001	-0,2992	-0,0618
Marche	0,9414	-0,2622	-0,0235	-0,0013	-0,0089
Lazio	0,787	-0,2739	-0,0772	-0,2305	-0,0132
Abruzzo	0,8075	-0,2998	0,0001	0,0825	-0,1094
Molise	0,0391	0,04	0,058	0,2169	-0,8451
Campania	0,7272	-0,22	0,4832	-0,2169	-0,0429
Puglia	0,3748	-0,2036	0,0109	0,7851	0,1673
Basilicata	0,3683	-0,2342	0,0694	0,8543	0,0696
Calabria	-0,2752	-0,1061	0,67	0,425	0,1136
Sicilia	0,2385	0,4952	-0,6025	0,3684	0,0499
Sardegna	-0,3663	-0,1777	0,0436	0,5759	-0,0016

The remaining three principal components taken into consideration indicate three additional and distinct patterns among the long-term components, i. e., the distinct trajectory for Valle d'Aosta which is correlated with the third principal component (13% of total variation), the common path for Basilicata and Puglia (remember the correlation coefficient between these two components in *Table 2* is 0.83) which identifies the fourth principal component (12% of total variation), and, finally, the contrasting trajectories present in the long-term components for Liguria and Molise, both with a relatively long periodicity of approximately five years, and associated with the fifth principal component (8% of total variation).

5. Conclusion.

Epidemiologist as well as sanitary authorities are interested in the mechanisms which might generate the spread of infectious diseases. It is natural that causal models, which often take into account demographic variables, should constitute the main theoretical tool of analysis. However, evidence at times suggests shortcomings in such an approach and this gives rise to the need for alternative analyses. There is not much evidence in the published literature of the methods of time series analysis applied in this paper although their usefulness is well illustrated here.

We have shown that AR models, identified through a comparison of the parametric and nonparametric spectra, are capable of providing better predictions of the long-term cycle in the regional spread of measles during a pre-vaccination period in Italy than one of the commonly used causal mathematical model. This leads to the possibility of obtaining better forecasts, although a causal explanation is missing.

Unlike England and Wales where the long-term cycle has a constant periodicity of two years, the Italian regions are characterised by long-term cycles which vary from just over two years to more than five years which means that inter-regional relationships are limited or not simple to model. However, the estimates of the trajectories of the long-term cycles have enabled us to identify, principally, two cluster of regions with similar dynamics. The interesting point with this result is that the regions involved are all neighbouring Regions in central Italy and in the north east.

Finally, the information contained in *Figures 1* and *2*, such as the changes in the amplitudes of the long-term oscillations as well as changes in the power of the spectrum in time, should be useful to the model builder in selecting those exogenous variables which would result useful in explaining such dynamics.

Acknowledgements: A special thanks must go to S. Salmaso and M. Ciofi of the Istituto della Sanità in Rome for providing the data, and to P. Manfredi for innumerable and invaluable discussions on modelling the spread of infant diseases.

References

Anderson R.M., Grenfell B.T., May R.M. (1984), Oscillatory fluctuations in the incidence of infectious diseases and the impact of vaccination: time series analysis, *Journal of Hygiene Cambridge*, 93, 587-698

Anderson R.M., May R.M. (1991), *Infectious diseases of humans: dynamics and control*, Oxford University Press, Oxford.

Bjornstad. O. N., B. F. Grenfell, Finkenstadt B.F. (2002), Dynamics of Measles Epidemics: Estimation of Scaling of Transmission Rates using a Time Series SIR Model. *Ecological Monographs*, 72, 169-184.

Cleur E. M., Manfredi P., Williams J. R. (2003), The pre and post vaccination regional dynamics of measles in Italy: Insights from time series analysis, *Working Paper 244, Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa.*

Damsleth E., Spjotvoll E. (1982), Estimation of Trigonometric Components in Time Series, *Journal of the American Statistical Association*, 77, 381-87.

Dicky, D. A., Fuller W. A. (1979), Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association.*, 74, 427-31.

Finkenstadt B. F., Grenfell. B. T. (1998a), Empirical determinants of measles metapopulation dynamics in England and Wales, *Proc. R. Soc. London*, 265, 211-220.

Finkenstadt B. F., Keeling M. and Grenfell. B. T. (1998b), Patterns of density dependence in measles dynamics, *Proc. R. Soc. London,* 265, 753-762.

Finkenstadt B. F., Grenfell. B. T. (2000), Time series modelling of childhood diseases: a dynamical systems approach., *Journal of the Royal Statistical Society, Series C*, 49, 187-205.

Granger C. W. J., Hatanaka J. (1964), *Spectral Analysis of Economic Time Series*, Princeton Univ. Press, New Jersey.

Grenfell B. F., Bjornstad O. N, Kappey, J. (2001), Travelling waves and spatial hierarchies in measles epidemics, *Nature*, 414, 716-723.

Grenfell B. F., Bjornstad O. N., Finkenstadt B. F. (2002), Dynamics of Measles Epidemics: Scaling Noise, Determinism, and Predictibility with the TSIR Model, *Ecological Monographs*, 72, 185-202.

Hamilton J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.

Manfredi P., Williams J. W., Cleur E. M., Salmaso S., and Ciofi M (2002), The pre-vaccination regional landscape of measles in Italy: contact patterns and related amount of needed eradication efforts (and the "EURO" conjecture), *Working Paper 230, Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa* .

Phillips P. C. B., Perron P. (1988), Testing for a unit root in time series regression, *Biometrika*, 75, 335-346.

Priestley M. B. (1981), Spectral Analysis and Time Series: Volume 1, *Academic Press Inc.*, London.

Shumway R. H., Stoffer. D. S. (2000), Time Series Analysis and its Applications, *Springer-Verlag*, New York.

Wand M. P., Jones M. C. (1995), Kernel Smoothing, *Chapman and Hall/CRC*, London.