View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CiteSeerX

Comparison of Time Series Decomposition Methods

[1]V. Svendova, [1,2]J. Holcik

[1]Institute of Biostatistics and Analyses, Brno, Czech Republic,
[2]Institute of Measurement Science, SAS, Bratislava, Slovakia
Email: svendula@mail.muni.cz

Abstract. *In this paper we presented some of the classical methods for the decomposition of a time series. We used moving average/median methods for removing trend and combined them with averaging and recursive methods for removing a seasonal component. We applied these methods to medical data of colorectal cancer incidence in the Czech Republic and results were compared using Fisher's g test statistics. Moving median in combination with either of the season removing methods proved to be the most effective method for the cancer incidence data.*

Keywords: time series decomposition, moving average, moving median, recursive method, Fisher's g test of periodicity

1. Introduction

There are number of methods for eliminating trend (*T*), seasonal (*S*) and random (ε) components of a time series. We applied some of the most used methods and compared suitability of their application on the data we work with, which describes dynamics of monthly incidence of colorectal cancer. We used additive decomposition model, signed as

$$y_t = T_t + S_t + \varepsilon_t \tag{1}$$

2. Subject and Methods

Trend estimation

Trend estimation is based on suppression of seasonal and random fluctuations. Method of centered moving averages, also called filtering, is one of the classical approaches for obtaining trend. Let us denote filter coefficients as a vector:

$$f = (f_{-n}, f_{-n+1}, \dots, f_0, \dots, f_{n-1}, f_n) \tag{2}$$

The trend estimation T_t of a time series y_t by centered moving averages, is computed as

$$T_t = \sum_{i=-n}^{n} f_i y_{t+i}\,, \quad \text{where} \sum_{i=-n}^{n} f_i = 1 \tag{3}$$

Order and values of the filter coefficients depend on character of a series, smoothness requirements and expected period of a seasonal component. The order of a filter is defined by the width of a time series segment which is smoothed by a polynomial or other function. We assume that our monthly data show deterministic trend and are periodic with a period of one year. Therefore the frequency is 1/12 = 0.0833 [month^{-1}] and a filter needs to be of the length $12k + 1, k \in \mathbb{Z}$ (odd length for practical reasons, see [1]). The longer the filter is, the smoother the trend is going to be.

We compared performance of *simple moving average (SMA)* [1], length of 13:

$$f_{SMA} = \frac{1}{12}\left(\tfrac{1}{2}, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \tfrac{1}{2}\right), \tag{4}$$

triangle moving average (TMA), length of 25, which represents serial connection of two SMA filters:

$$f_{TMA} = \frac{1}{144}\left(\tfrac{1}{4}, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 11.5, 11, 10, 9, \ldots, 3, 2, 1, \tfrac{1}{4}\right), \qquad (5)$$

polynomial moving average (PMA) [1], length of 13:

$$f_{PMA} = \frac{1}{143}(-11, 0, 9, 16, 21, 24, 25, 24, 21, 16, 9, 0, -11) \qquad (6)$$

and *moving median (MM)* [1], length of 13:

$$T_t = median(y_{t-6,}y_{t-5}, \ldots, y_t, y_{t+1}, \ldots, y_{t+6}) \qquad (7)$$

Seasonal estimation

Seasonal components describe periodical changes in a time series. Computer calculations demand to use as easiest methods as possible, so that the results can be misleading. The most common method is *averaging (AV)* [1] which is based on the arithmetic means of individual months ($m = 1, \ldots, 12$), calculated over k-periods (years):

$$S_m^{AV} = \frac{1}{k}\sum_{i=0}^{k} y_{m+iT} \qquad (7)$$

where *T* is a period of the seasonal component (in our case 12 months).

Simplicity of this method is in an assumption of stationarity of the seasonal component.

An alternative method to the above mentioned averaging can be a *recursive method (R)* defined by a difference equation

$$S_m^R = 0.1\, y_m + 0.9\, S_{m-12}^R, \qquad (8)$$

that represents a recursive comb filter with pass-bands at frequencies corresponding to fundamental frequency of a seasonal component and its integer multiples. This method does not assume a stationary character of the seasonal component and responses to instantaneous variations in its frequency content.

Evaluation of the method performance

We used Fisher's exact g test of periodicity [3] for comparison of results of separation of the time series additive components, where

$\boldsymbol{H_0}$: series is a Gaussian white noise

$\boldsymbol{H_1}$: series contains a deterministic periodic component

with level of significance $\boldsymbol{\alpha = 0.05}$. This test is based on the periodogram spectral estimator and rejects the null hypothesis that the periodogram contains a value significantly larger than the average value.

Fisher's g test statistic (in [3] called *g statistic*) for a series of length *n* is computed as

$$fisher(n) = \frac{\max\limits_{1 \le i \le n}\{\mathrm{pgram_i}\}}{\sum_i pgram_i},$$

where $pgram_i$ is an ith value of periodogram of the time series.

Fisher value of a given series was compared with a *fisher* value of Gaussian noise, labelled as *IdealFisher*.

We compared the *fisher* values after the removal of a trend component, as well as after the removal of both the trend and the seasonal component.

3. Results

Data for the experiment were taken from the Czech National Cancer Registry [4]. We constructed the time series of a normalized incidence of a colorectal cancer (monthly values over almost 30 years).

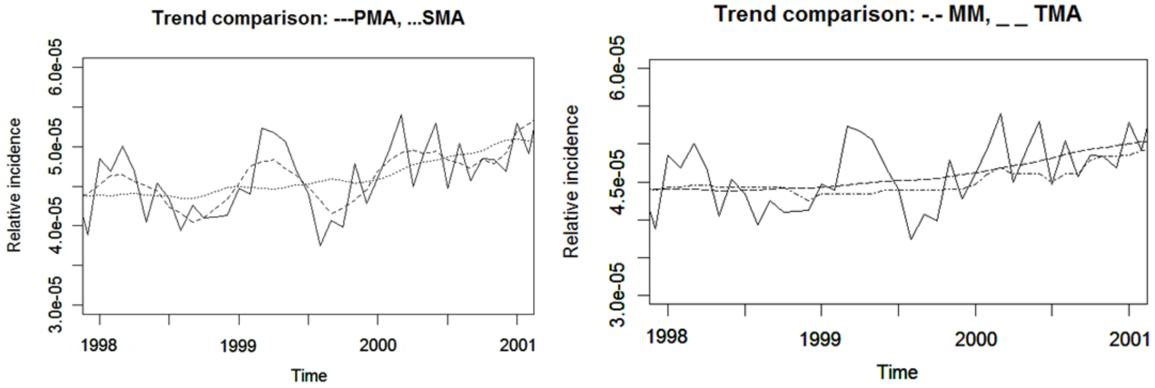

Fig. 1. Comparison of trends: solid line for the original data, dashed for polynomial (PMA), dotted for simple (SMA), longdashed for triangle (TMA) moving average and dotdashed for moving median (MM).

We see (Fig.1) that the triangle method produces significantly smoother trend than the others. Periodograms of the data with removed trend are shown in Fig.2.

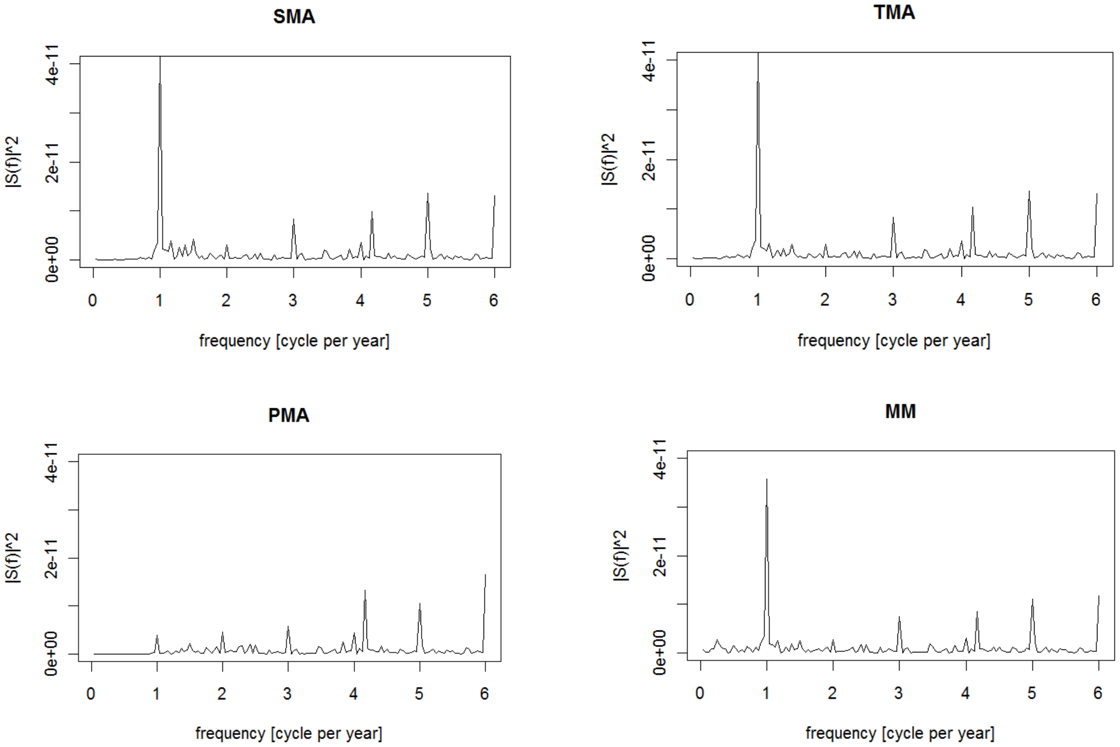

Fig. 2. Periodograms of time series with removed trend.

Table 1. Fisher's statistics for different processing of the time series. 'Season+random' stands for the original data after removing trend by the method in corresponding column (SMA,PMA,TMA,MM); 'random AV'and 'random R' stands for the original data after removing both trend and the seasonal component by AV, and R method, resp. (The closer to IdealFisher the fisher statistics is, the closer is a series to the white noise.) **IdealFisher = 0.038** (for a series of 288 values).

	SMA	**PMA**	**TMA**	**MM**
season+random	0.264	0.121	0.269	0.237
random AV	0.108	0.176	0.117	0.089
random R	0.104	0.164	0.107	0.086

4. Discussion

Due to the transients in computations of a filter responses, the original data were shortened from 372 to 288 values. Fisher's test rejected all null hypothesis on Gaussian noise, which indicates that there is still a significant periodic value in the processed data series after removing trend and seasonal component by the methods described here. This led us to application of the Fisher's statistics from the Fisher's exact g test and comparison *IdealFisher* value with a *fisher* value for the specific time series, as a rate of proximity to the white noise. From the Table 1, it is obvious that the best results (after removing the seasonal component) were obtained by means of moving median method (MM). MM method suppresses influence of outlying values, which seems to be important for our data. PMA filter, while used for removing trend, suppressed also a significant amount of periodic component at its fundamental frequency. This can be seen in Fig. 1, where trend determined by PMA closely matches the original series, as well as from the low value of $|\boldsymbol{S}(\boldsymbol{f})|^{\mathbf{2}}$ at the frequency $\mathbf{f} = \mathbf{1}\ [\boldsymbol{cycle\ per\ year}]$ in the periodogram in Fig. 2. We can see (Table 1, method PMA) the value of *fisher* statistics for the 'random R' series higher than that for 'season+random' series. This indicates the fact that application of the R method to the series describing colorectal cancer incidence with no significant seasonal component with a year periodicity, introduces false periodicity instead of removing it. Unfortunately, the seasonal component of the processed data does not satisfy a condition of stationarity. Further, signal to noise ratio for the given type of data is rather low and the number of repetitions of the seasonal component is relatively small. That is why the results obtained by the averaging method appeared even worse than in case of the recursive filter.

Acknowledgements

The research was granted by the project of the Czech Science Foundation No.102/09/H083: "Information Technology in Biomedical Engineering".

References

[1] Cipra T, Finanční ekonometrie. EKOPRESS,s.r.o., 2008, 257-325.

[2] Cryer J D, Chan K-S, Time Serie Analysis with applications in R, Springer, 2008.

[3] Ahdesmaki M, Lahdesmaki H, Yli-Harja O, *Robust Fisher's Test for Periodicity Detection in Noisy Biological Time Series,* 2007

[4] Klinická onkologie, The Journal of the Czech and Slovak Oncological Societies, Supplement 2007