Time series analysis and forecasting intellectual methods using in epidemiology supervision system

Olga Radivonenko, Tetiana Korchak

Informatics Department, National Aerospace University, Chkalova Str., 17, Kharkiv, 61070, UKRAINE,
E-mail: ORadivonenko@gmail.com, kotavi@i.ua

Abstract – In this paper the problem of time series analysis is considered. The importance of a collaborative approach to the prevention and control of enteric infection are highlighted. Time series forecasting methods are described. Basic methods of statistical, regression and fractal analysis are used. Also the problem of epidemic thresholds modeling was considered. The implementation of the fuzzy clustering procedure in the method of epidemic thresholds calculation was proposed. This feature allows exclude epidemic information from the calculation data, which will improve trueness and certainty of results substantially.

Key words – R/S analysis, Hurst exponent, fuzzy clustering, epidemiology supervision.

I. Introduction

Main characteristic of the epidemic clinical behavior are almost annually nascent epidemics, which not only hurt to health for people but also have substantial influence in life of society, reflected on national economy. It requires implementation of all possible measures of prophylaxis for maximal diminishment of the annually inflicted hurt.

Efficiency of fight against infections in the epidemic period in a great deal depends on quality of the prophylactic measures and timeliness of realization of the program of antiepidemic measures conducted in a preepidemic period. Early determination of beginning of epidemic is needed for this purpose, that is the basic task of epidemiology supervision. The common chart of the epidemiology supervision system is depicted at the fig. 1.

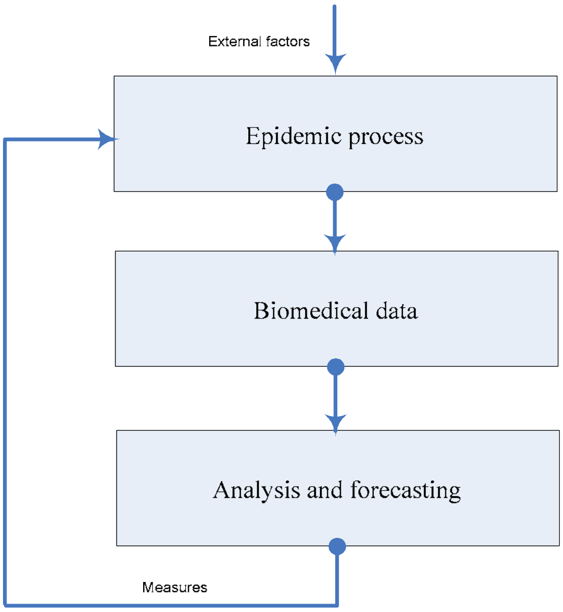

Fig.1. Epidemiology supervision system common chart

Epidemiological service carries out the daily monitoring of level of morbidity and dynamics of development of epidemic process by means of indexes of epidemic thresholds. Epidemic thresholds are estimated on the basis of long-term indexes of morbidity for all age groups of population. Within the framework of the epidemiology supervision a daily and weekly calculation and systematic analysis of morbidity in different age groups of population is conducted. If current morbidity exceeds the set threshold (tolerance limit for middle unepidemic morbidity), it is the sign of beginning epidemic.

To solve this problem the theory of time series analysis and forecasting has been considered. An important application of time series forecasting is to prevent undesirable events to occur by applying corrective or contention measures.

II.Time series analysis and forecasting methods

Analysis and forecasting methods of dynamic time series are connected with a research of the parameters isolated from each other. Each parameter consists of two elements: the deterministic forecast component and the random forecast component. If the basic tendency of evolution is certain then development of the first forecast does not represent great difficulties and its further extrapolation is possible. The forecast of the random components are more complex, as its occurrence can be estimated only with some probability.

The principal goal of time series analysis is to develop quantitative methods which allow us to characterize time series, in particular, to say quantitatively how two time series differ or how they are related.

Time series data can be considered as deterministic, random and chaotic (fig. 2).

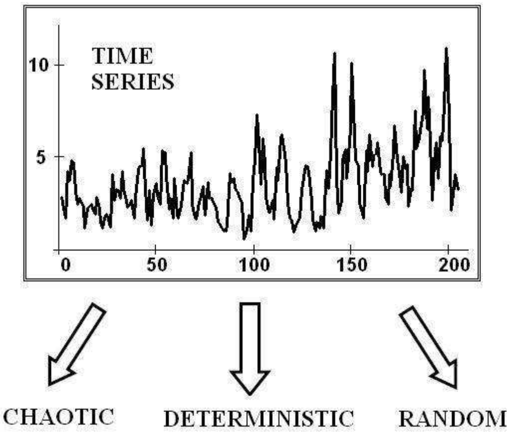

Fig.2. Time series behavior

Deterministic – can be predicted by an explicit mathematical relationship; random – the exact value of the future cannot be predicted based on the known observations. In truth, many time series lie somewhere in between strictly deterministic and random.

To define whether time series is normally distributed or not the normality test is presented. An informal approach to testing normality is to compare a histogram of the residuals to a normal probability curve. Based on the Fig. 2 hypothesis of normal distribution is rejected.

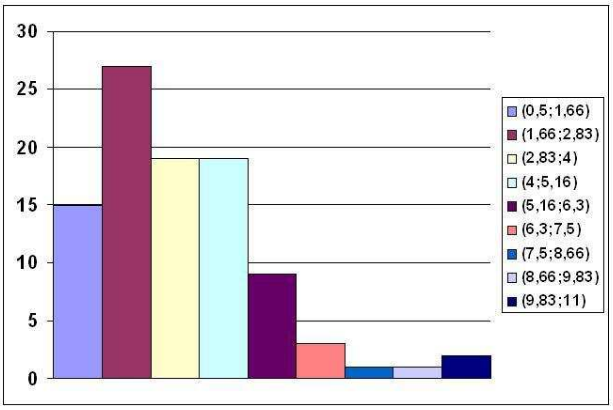

Fig.3. Histogram of the normality test

Time series data was analyzed for the autocorrelation function (Fig. 4)

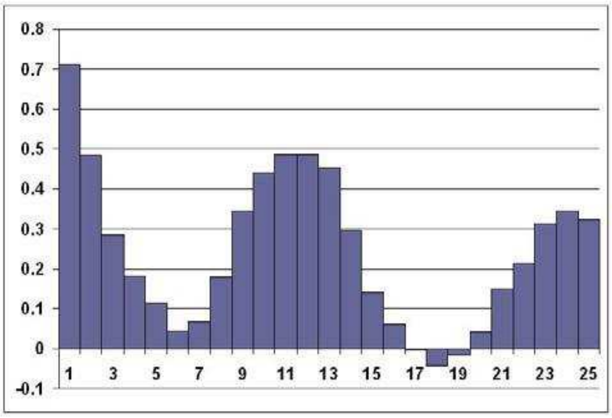

Fig.4. Histogram of the autocorrelation function analysis

R/S analysis was presented. It is non-parametric analysis, meaning there is no assumption or requirement of the shape of the underlying distribution. It was developed by H. E. Hurst who observed an unexpected behavior of natural time series. They have become known as the Hurst phenomenon.

In [1] the algorithm of calculating of the Hurst's exponent was shown. It starts with calculating of the standard deviation:

$$S(\tau)=\sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}\left(x_t-\bar{x}_\tau\right)^2}\,, \qquad (1)$$

where $\bar{x}_\tau=\frac{1}{\tau}\sum\limits_{t=1}^{\tau}x_t$.

The self-adjusted range $R(\tau)$ is defined

$$R(\tau)=\max_{t=1}^{\tau}X(t,\tau)-\min_{t=1}^{\tau}X(t,\tau)\,. \qquad (2)$$

Finally, next formula (3) defines the value that is equal to H, which is called Hurst's exponent.

$$R(\tau)/S(\tau)=\left(\tau\!\big/\!_2\right)^H. \qquad (3)$$

$H=0.5$ implies an independent process.

$0.5<H\le 1$ imply a persistent time series characterized by long memory effects.

$0\le H<0.5$ imply an anti-persistent time series, which covers less distance than a random process. Such behavior is observed in mean-reverting processes, although that assumes that the process has a stable mean.

Fig. 5 shows the result of calculating Hurst's exponent for the former time series.

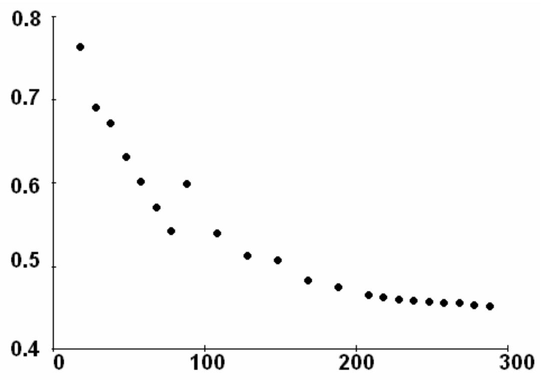

Fig. 5. Values for Hurst's exponent

Based on the time series analysis that was conducted in this work the next assumption can be provided. The former time series can be reconstructed by the next rule.

$$X=\begin{pmatrix}X^1\\ X^2\\ \dots\\ X^{12}\end{pmatrix}=\begin{pmatrix}x_i^1 & x_{i+L}^1 & x_{i+2L}^1\dots x_{i+nL}^1\\ x_i^2 & x_{i+L}^2 & x_{i+2L}^2\dots x_{i+nL}^2\\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots & & \\ x_i^{12} & x_{i+L}^{12} & x_{i+2L}^{12}\dots x_{i+nL}^{12}\end{pmatrix},$$

where X_i^n – state of the system in discrete time i, n – the number of years, L – lag or shift reconstructed.

The next formula (4) is used for forecast implementation.

$$X \quad\rightarrow\quad \begin{pmatrix}x_{i+(n+1)L}^1\\ x_{i+(n+1)L}^2\\ \dots\dots\\ x_{i+(n+1)L}^{12}\end{pmatrix} \qquad (4)$$

According to R/S analysis and test for normality the statement can be assumed that reconstructed time series can provide us with better forecasting result.

III. Forecasting approaches

Single Exponential smoothing is a very popular forecasting method for some reason:

- it is easy to use;
- requires very little computation effort;
- needs only a few data to produce future prediction.

It is recommended for short or immediate term prediction, for stationary data or when there is a slow growth or decline over time. The method bases on the following formula [3]:

$$y_t^* = \alpha y_t + (1-\alpha)y_t^*,$$

where y_t^* is a forecast, y_t is a time series, α is between zero and one.

The Holt's Three Parameters Exponential Smoothing Model is similar to Brown's model as it estimates the trend and uses it in forecasting. The equations are as follows [3]:

$$S_t = \alpha X_t + (1-\alpha)(S_{t-1} + T_{t-1} + R_{t-1/2})$$

$$T_t = \beta dS_t + (1-\beta)T_{t-1}$$

$$R_t = \gamma d2S_t + (1-\gamma)R_{t-1}$$

$$dS_t = S_t - S_{t-1},\ d2S_t = dS_t - dS_{t-1}$$

Finally, the forecast can be found as:

$$F_{t+m} = S_t + T_t m + 1/2 R_t m^2$$

where m is the number of periods ahead to be forecast.

The smoothing constants α, β, γ must be specified to minimize the forecast errors over a past time horizon. During the smoothing constants selection procedure a compromise between two wishes should be taken into consideration. On the one hand it is recommended to follow changes in the pattern of the data. On the other hand a method that can distinguish between random fluctuations and changes in the basic pattern of the data is highly required.

Takagi-Sugeno models are adaptive fuzzy logic application for a forecasting problem epidemic growth of disease parameters has following advantages [4]:

– adaptive fuzzy models are easily construed after training by the person;

– some fuzzy models (Mamdani type) are less exacting to an experimental data volume, than neural networks or networks TSK;

– conflicting data can be processed by the fuzzy logic models;

– fuzzy models exactness can be improved by an addition of expert rules.

IV. Forecast results

As a result of numerical experiments prediction methods was estimated (Table 1), where RME and MAPE – mean square error (5) and mean absolute error in percents (6), respectively.

$$RME = \sqrt{\frac{1}{T^*}\sum_{t=1}^{T^*}\left(y_t^* - y_t\right)^2} \quad (5)$$

$$MAPE = \frac{1}{T^*}\sum_{t=1}^{T^*}\left(\frac{y_t^* - y_t}{y_t}\right)\cdot 100 \quad (6)$$

where y_t – actual value of t; y_t^* – predicted value for t; T^* – forecasting horizon.

TABLE 1

	RME	MAPE
Braun method	1.5	27%
Adaptive ES	2.068	32%
Holt-Vinters	2.33	69%
Least square method	2.66	40
TS-models	3.94	61

As a result Fig. 6 was presented with graphs of results of used forecasting methods.

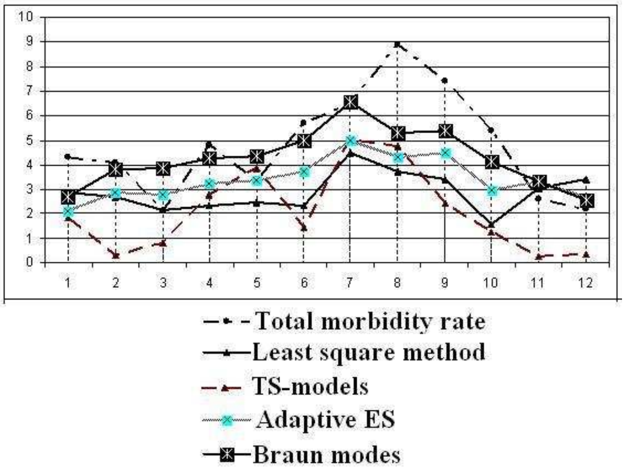

Fig. 6. Forecasting result

V. Epidemic thresholds calculation method

Usually, the method, based on the determination empirical series statistics is used for the estimation of epidemic thresholds [5,6]. Epidemic thresholds are the upper tolerant limits of unepidemic morbidity.

At the sufficient number of supervisions $N_i \geq 5$ upper tolerant limit was calculated:

$$X'_в = \overline{X}_i + Q_{N_i-2}\sqrt{\frac{N_i - 1}{N_i - 2 + Q^2{}_{N_i-2}}}\cdot S_i,$$

where Q_{N_i-2} – value of the Student's test for the 95% confidence probability and $N_i - 2$ degree of freedom, $\overline{X}_i = \frac{1}{N_i}\sum_{n=1}^{N_i} X'_n$ – mean value of morbidity for calculation period, $S_i = \sqrt{\frac{1}{N_i - 1}\sum_{n=1}^{N_i}(X'_n - \overline{X}_1)^2}$ – standard deviation.

The approximate method with the variation coefficients calculation was used for weeks, where the number of the supervisions $N_i < 5$. Results of calculation are epidemic threshold graphs for each of age groups.

According to the mentioned methodic, it is necessary to exclude data which were represented during the period of epidemics in the city, and data, which were admitted as ambiguous from basic data. Usually, epidemiologist performs it. The using of fuzzy clustering and next aggregation of results with epidemiologist's subjective resolution allow select the class of epidemic information more accurately. As the clustering procedure, hybrid fuzzy clustering algorithm was used, described in [7]. Proposed approach is based on a fuzzy relation similarity model and supplemented with clusters merging algorithm. Application of fuzzy clustering allows remove the form and positional relationship uncertainty of clusters, quantity of clusters uncertainty and partitioning fuzziness.

Verification of results was conducted by means of real statistical information and epidemic thresholds comparison for weeks, during which the epidemic of flu was registered in the city (Fig. 7).

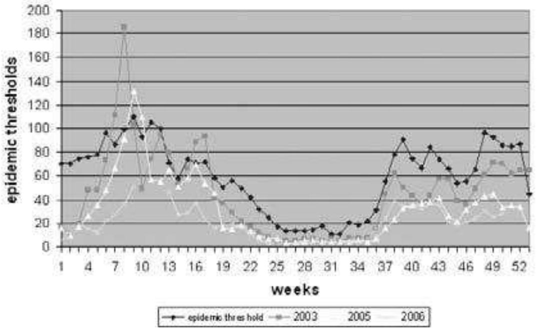

Fig. 7. Epidemic threshold and morbidity

The computational experiments were carried out. According to results, average epidemic threshold was reduced 19% in comparison with classical method (Fig. 8).

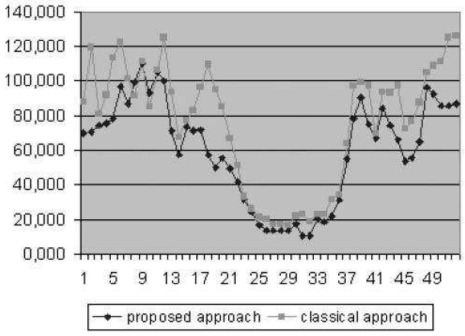

Fig. 8. Comparison results of the classical and proposed approach for an age group 7-14 years (average epidemic threshold reduction 19%)

On the base of the considered methods developed and deployed the «Epidemiology supervision information system on flu "Epid_Observe"» [8].

VI. Conclusion

In the paper comparison of fuzzy logic methods and traditional methods of time series forecasting was conducted. Computing experiments showed that not all the offered prediction approaches provide high accuracy of forecasting. Hurst's R/S-analysis provided the research with a new approach of reconstructing of the time series trajectory.

Also the epidemic thresholds calculation method using hybrid fuzzy clustering procedure was represented. The method allows exclude epidemic information from the calculation data, and demonstrates efficiency of proposed approach in comparison with classical method. The developed epidemiology supervision information system for monitoring and analysis of epidemic morbidity allows simplify the task of doctor-epidemiologist, and reduce time for the calculation of epidemic thresholds for all of age-dependent groups.

References

[1] R. L. Bras, I. Rodríguez-Iturbe, "Random Functions and Hydrology", Technology & Engineering, 1993, – 559 p.

[2] E. E. Peters, "Fractal market analysis: Applying Chaos Theory to Investment and Economics", Wiley Finance, 1994 – 350 p.

[3] C. C. Holt, Forecasting Trends and Seasonals by Exponentially Weighted Moving Averages, Carnegie Institute of Technology, Pittsburgh, Pennsylvania, 1957.

[4] Tanaka K., Sugeno M. (1992) Stability analysis and design of fuzzy control systems. Fuzzy Sets and Systems, vol. 45, 135-156.

[5] Methodical recommendations on the operative analysis and prognostication of epidemiology situation on a influenza and other acute respiratory viral infections //Moscow-Saint-Petersburg, 2006. – 72 p.

[6] Sokolov A.Yu. Time series analysis mathods in the tasks of infectious diseases eruptions prognostication / A.Yu. Sokolov, O.S. Radivonenko, T.V.Korchak // Radioelectronic and computer systems. – Kh.: NASU Khai, 2007. – № 2 (21) . – P. 36-41.

[7] Sokolov A. Fuzzy clustering algorithm for data segmentation under uncertainty conditions / A. Sokolov, O. Radivonenko // Proceedings of East West Fuzzy Colloquium 2008. – Zittau: IPM, – 2008. – P. 234-238.

[8] A. c. 23449, Ukraine,.Computer program «Epidemiology supervision information system on flu "Epid_Observe"» / O.Yu. Sokolov, T.O. Chumachenko, L.O. Kleschar, O.S. Radivonenko, V.V. Yakubovs'kiy, Ye.V. Ganchin. – 17.01.2008.