ELSEVIER

Fuzzy Sets and Systems 123 (2001) 387–394

FUZZY sets and systems

www.elsevier.com/locate/fss

Effective lengths of intervals to improve forecasting in fuzzy time series

Kunhuang Huarng

Department of Finance, Chaoyang University of Technology, 168 GiFeng E. Rd., WuFeng, Taichung County, Taiwan, ROC

Received 14 December 1998; received in revised form 12 November 1999; accepted 9 March 2000

Abstract

Length of intervals affects forecasting results in fuzzy time series. Unfortunately, the issue of how to determine effective lengths of intervals has not been touched in previous studies. This study proposes distribution- and average-based length to approach this issue. Distribution-based length is the largest length smaller than at least half the first differences of data. Average-based length is set to one half the average of the first differences of data. Empirical analyses show that distribution- and average-based lengths are simple to calculate and can greatly improve forecasting results; in particular, they are superior to the randomly chosen lengths used in previous studies. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Enrollments; Forecasting; Fuzzy sets; Fuzzy time series; Stock market indices

1. Introduction

The concept of fuzzy time series was first proposed by Song and Chissom [2]. Since then, several other studies have appeared [1,3,4]. However, there are still many critical issues open. The determination of effective lengths of intervals is one of these.

Length of intervals greatly affects forecasting results in fuzzy time series. Hence, an effective length of intervals can significantly improve the forecasting results. This study proposes distribution- and average-based length to approach this issue. Chen's model gave the best results among previous studies [1,3,4] and hence is selected as the target for comparison. The yearly data on enrollments at the University of Alabama as well as the daily data from Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) are used to demonstrate the impact of effective lengths of intervals on forecasting results. Empirical analyses show that both distribution- and average-based lengths are simple to calculate and can greatly improve on the forecasting results obtained from previous models.

Section 2 briefly introduces fuzzy time series. Section 3 explains the relevant definitions of lengths of intervals and proposes approaches to determine effective lengths. Sections 4 and 5 compare the forecasting results of university enrollments and TAIEX by various lengths of intervals. Section 6 offers some conclusions.

2. Fuzzy time series

Let U be the universe of discourse, where $U = \{u_1, u_2, \ldots, u_n\}$. A fuzzy set A_i of U is defined by

$$A_i = f_{Ai}(u_1)/u_1 + f_{Ai}(u_2)/u_2 + \cdots + f_{Ai}(u_n)/u_n,$$

E-mail address: huarng@mail.cyut.edu.tw (K. Huarng).

0165-0114/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0165-0114(00)00057-9

where f_{Ai} is the membership function of fuzzy set A_i, $f_{Ai}: U \to [0,1]$. u_k is the element of fuzzy set A_i, and $f_{Ai}(u_k)$ is the degree of belongingness of u_k to A_i. $f_{Ai}(u_k) \in [0,1]$ where $1 \leqslant k \leqslant n$.

The concept of fuzzy time series was first proposed by Song and Chissom [2]:

Definition 1. $Y(t)$ $(t = \ldots, 0, 1, 2, \ldots)$, is a subset of R. Let $Y(t)$ be the universe of discourse defined by fuzzy set $f_i(t)$. If $F(t)$ consists of $f_i(t)$ $(i = 1, 2, \ldots)$, $F(t)$ is defined as a fuzzy time series on $Y(t)$ $(t = \ldots, 0, 1, 2, \ldots)$.

Following Definition 1, relevant definitions are proposed.

Definition 2. If there exists a fuzzy relationship $R(t-1,t)$, such that $F(t) = F(t-1) \times R(t-1,t)$ where × represents an operator, then $F(t)$ is said to be caused by $F(t-1)$. (Note that the operator can be either max–min [3], min–max [4], or arithmetic operator [1].) When

$$F(t-1) = A_i \quad \text{and} \quad F(t) = A_j,$$

the relationship between $F(t-1)$ and $F(t)$ (called a fuzzy logical relationship in [3]) is denoted by

$$A_i \to A_j.$$

Definition 3. Fuzzy logical relationships with the same fuzzy set on the left-hand side can be further grouped into a fuzzy logical relationship group [1]. Suppose there are fuzzy logical relationships such that

$$A_i \to A_{j1},$$
$$A_i \to A_{j2},$$
$$\ldots$$

They can be grouped into a fuzzy logical relationship group

$$A_i \to A_{j1}, A_{j2}, \ldots$$

Following Chen's model, the same fuzzy sets can only show up once on the right-hand side of the fuzzy logical relationship group.

3. Lengths of intervals

For enrollment forecasting, Song and Chissom choose 1000 as the length of intervals, without specifying any reason [3]. Since then, 1000 has been used as the length of intervals in further studies [4,1]. How the lengths of intervals affect forecasting results was left unanswered in these studies. In fact, different lengths of intervals may lead to different forecasting results.

A time-series forecasting example is given with two lengths of intervals to show that different lengths of intervals may lead to different forecasting results and forecasting errors.

Suppose we have the following time-series data:

$$6, 10, 12, 6, 4$$

The range U is set as $[3, 13]$. If the length of intervals is chosen as 5, there are two intervals: $u_1 = [3, 8]$ and $u_2 = [8, 13]$. According to Chen's model, the forecasting mean squared error (MSE) is 10. On the other hand, if the length of interval is set to 2, there are five intervals: $u_1 = [3, 5]$, $u_2 = [5, 7]$, $u_3 = [7, 9]$, $u_4 = [9, 11]$, and $u_5 = [11, 13]$. The MSE is 4.5. Obviously, different lengths of intervals result in different forecasting errors.

Hence, the determination of the lengths of intervals, especially effective ones, is never trivial in the forecasting of fuzzy time series. An efficient way to choose effective lengths of intervals is therefore critical to improve forecasting in fuzzy time series. A key point in choosing effective lengths of intervals is that they should not be too large or small. When an effective length of intervals is too large, there will be no fluctuations in the fuzzy time series. On the other hand, when the length is too small, the meaning of fuzzy time series will be diminished. In oder to reflect fluctuations properly and to keep fuzzy time series meaningful, the heuristic is set in such a way that at least half the fluctuations in the time series are reflected by the effective lengths of intervals.

The fluctuations in fuzzy time series can be represented by the absolute value of the first differences of any two consecutive data (the first differences hereafter). Hence, the heuristic can reflect at least half the first differences. Based on this idea, two approaches are proposed: distribution- and average-based length. Distribution-based length is calculated

Table 1
Base mapping table

Range	Base
0.1–1.0	0.1
1.1–10	1
11–100	10
101–1000	100

according to the distribution of the first differences of data. Average-based length is set to the largest length that is smaller than half the first differences.

Algorithm for distribution-based length
1. Calculate all the absolute differences between A_{i+1} and A_i $(i = 1,\dots,n-1)$, as the first differences and the average of the first differences.
2. According to the average, determine the base for length of intervals by following Table 1.
3. Plot the cumulative distribution of the first differences. The base determined in step 2 is used as interval.
4. According to the base determined in step 2, choose as the length of intervals the largest length that is smaller than at least half the first differences.

To show how to determine effective length of intervals using distribution-based length, another example is given. Suppose we have the following time series data: 30, 50, 80, 120, 100, and 70. The algorithm for distribution-based length is implemented step by step below:

1. The first differences are

   $20, 30, 40, 20, 30$

   The average of the first differences is 28.
2. From Table 1, the base for the length of intervals is 10.
3. The number of the first differences larger than 30 is 1. The number of the first differences larger than 20 is 3.
4. Because 20 is the largest length we can have, which is still smaller than at least half the first differences, 20 is chosen as the length of intervals.

The second approach is based on the average of the first differences of data, so it is called average-based length. Since the average of the first differences may not necessarily fulfill the heuristic (at least half the first differences should be reflected), the average-based length is set to one half of the average of the first differences.

Algorithm for average-based length
1. The same as step 1 in the algorithm for distribution-based length.
2. Take one half the average (in step 1) as the length.
3. According to the length (in step 2), determine the base for length of intervals by following Table 1.
4. Round the length according to the determined base as the length of intervals.

To demonstrate how effective length of intervals can be determined by average-based length, the same time series example as above is given. The time series data are 30, 50, 80, 120, 100, and 70. The algorithm for average-based length is implemented step by step below:

1. The first differences are

   $20, 30, 40, 20, 30$

   The average of the first differences is 28.
2. Take half of the average as the length, which is 14.
3. According to the length (in step 2), the base for length of intervals is determined as 10 by following Table 1.
4. Round the length 14 by the base 10, which is 10. So 10 is chosen as the length of interval.

4. Enrollment forecasting

Enrollment forecasting for the University of Alabama was used in the previous studies of fuzzy time series [1,3,4]. This case is also used here to show the disadvantages of randomly chosen lengths. In terms of MSE, Chen's model gave better results than the other models. So Chen's model is used as the target model to compare the effects of various lengths of intervals.

4.1. Chen's model with randomly chosen length of intervals

Forecasting by Chen's model with a range 13 000–20 000 and length of intervals 1000 is illustrated below.

Step 1: *Defining the universe of discourse and intervals*. As in [1], $U = [13\,000, 20\,000]$; the length of the intervals is 1000. Hence, there are intervals u_1, u_2, u_3, u_4, u_5, u_6, u_7, where $u_1 = [13\,000, 14\,000]$, $u_2 = [14\,000, 15\,000]$, $u_3 = [15\,000, 16\,000]$, $u_4 = [16\,000, 17\,000]$, $u_5 = [17\,000, 18\,000]$, $u_6 = [18\,000, 19\,000]$, $u_7 = [19\,000, 20\,000]$.

Step 2: *Defining fuzzy sets* A_i. In this case, the linguistic variable is "enrollment"; A_i $(i = 1, 2, \ldots)$ as possible linguistic values of "enrollment". Each fuzzy set A_i is assigned to a linguistic term: $A_1 =$ (not many), $A_2 =$ (not too many), $A_3 =$ (many), $A_4 =$ (many many), $A_5 =$ (very many), $A_6 =$ (too many), $A_7 =$ (too many many). Each A_i is defined by the intervals u_1, u_2, u_3, $\ldots, u_7$:

$$A_1 = 1/u_1 + 0.5/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7,$$

$$A_2 = 0.5/u_1 + 1/u_2 + 0.5/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7,$$

$$A_3 = 0/u_1 + 0.5/u_2 + 1/u_3 + 0.5/u_4 + 0/u_5 + 0/u_6 + 0/u_7,$$

$$A_4 = 0/u_1 + 0/u_2 + 0.5/u_3 + 1/u_4 + 0.5/u_5 + 0/u_6 + 0/u_7,$$

$$A_5 = 0/u_1 + 0/u_2 + 0/u_3 + 0.5/u_4 + 1/u_5 + 0.5/u_6 + 0/u_7,$$

$$A_6 = 0/u_1 + 0/u_2 + 0/u_3 + 0/u_4 + 0.5/u_5 + 1/u_6 + 0.5/u_7,$$

$$A_7 = 0/u_1 + 0/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0.5/u_6 + 1/u_7.$$

Table 2 lists the enrollments at the University of Alabama from 1971 to 1992 and their corresponding fuzzy enrollments A_i.

Step 3: *Establishing fuzzy logical relationships and fuzzy logical relationship groups*. From A_i in Table 2, the fuzzy logical relationships can be obtained, as in Table 3. The fuzzy logical relationships can be rearranged to fuzzy logical relationship groups, as in Table 4.

Table 2
Enrollment data sets

Year	Enrollment	Fuzzy enrollment A_i
1971	13 055	A_1
1972	13 563	A_1
1973	13 867	A_1
1974	14 696	A_2
1975	15 460	A_3
1976	15 311	A_3
1977	15 603	A_3
1978	15 861	A_3
1979	16 807	A_4
1980	16 919	A_4
1981	16 388	A_4
1982	15 433	A_3
1983	15 497	A_3
1984	15 145	A_3
1985	15 163	A_3
1986	15 984	A_3
1987	16 859	A_4
1988	18 150	A_6
1989	18 970	A_6
1990	19 328	A_7
1991	19 337	A_7
1992	18 876	A_6

Table 3
Enrollment fuzzy logical relationships

$A_1 \to A_1$	$A_1 \to A_2$
$A_2 \to A_3$	$A_3 \to A_3$
$A_3 \to A_4$	$A_4 \to A_4$
$A_4 \to A_3$	$A_4 \to A_6$
$A_6 \to A_6$	$A_6 \to A_7$
$A_7 \to A_7$	$A_7 \to A_6$

Table 4
Enrollment fuzzy logical relationship groups

$A_1 \to A_1, A_2$
$A_2 \to A_3$
$A_3 \to A_3, A_4$
$A_4 \to A_4, A_3, A_6$
$A_6 \to A_6, A_7$
$A_7 \to A_7, A_6$

Step 4: *Forecasting*. Forecasting is conducted by the following rules:

Rule 1: If the current fuzzy set is A_i, and the fuzzy logical relationship group of A_i is empty, i.e., $A_i \to$,

Table 5
Enrollment forecasting using various models with length 1000

	Song and Chissom's time invariant model [3]	Song and Chissom's time variant model [4]	Chen's model [1]
MSE	423 027	775 686	407 521

the forecast is m_i, the midpoint of u_i.

$Forecasting\_Value_i = m_i$.

Rule 2: If the current fuzzy set is A_i, and the fuzzy logical relationship group of A_i is one-to-one, i.e., $A_i \rightarrow A_{p1}$, the forecast is m_{p1}, the midpoint of u_{p1}.

$Forecasting\_Value_i = m_{p1}$.

Rule 3: If the current fuzzy set is A_i, and the fuzzy logical relationship group of A_i is one-to-many, i.e., $A_i \rightarrow A_{p1}, A_{p2}, \ldots, A_{pk}$, the forecast is equal to the average of $m_{p1}, m_{p2}, \ldots, m_{pk}$, the midpoints of $u_{p1}, u_{p2}, \ldots, u_{pk}$, respectively.

$$Forecasting\_Value_i = \frac{\sum_{x=1}^{k} m_{px}}{k}.$$

This study uses MSE to compare the forecasting performance.

$$MSE = \frac{\sum_{i=1}^{n}(Forecasting\_Value_i - Actual\_Value_i)^2}{n},$$

where there are altogether n data.

By Chen's model, the MSE is 407 521. Enrollment forecasting results by the other models with the same range and length of intervals are summarized in Table 5. The MSE by Song and Chissom's time invariant model is 423 027. The MSE by Song and Chissom's time variant model is 775 686. As far as MSE is concerned, Chen's model forecasts best.

4.2. Chen's model with various lengths of intervals

Since Chen's model forecasts best among various models, various lengths of intervals are applied to Chen's model. First, distribution-based length is calculated:

1. The first differences are calculated and their average is 510.33.
2. From Table 1, the base for the length of intervals is 100.
3. Plot the cumulative distribution of the first differences as in Fig. 1. There are 21 first differences in total. The number of the first differences larger than 900 is 3; 800 is 7; 700 is 8, etc.
4. From the distribution, there are 10 first differences less than 400. 400 is the largest length smaller than at least half the first differences. Hence, 400 is chosen as the length of intervals.

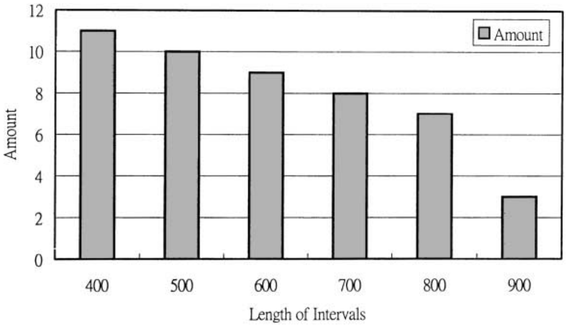

Fig. 1. Cumulative distribution of first differences for university enrollments.

When the length 400 is applied to Chen's model, the MSE is 124 707.

Second, average-based length is calculated:

1. The first differences are calculated and their average is 510.33.
2. Take half the average as the length, which is 255.17.
3. According to the length chosen in step 2, the base for length of intervals is set as 100 by following Table 1.
4. Round the length 255.17 by base 100, which is 300. So 300 is chosen as the length.

When average-based length is applied to Chen's model, the MSE is 78 792. Concerning MSE, the

Table 6
Enrollment forecasting using Chen's model with different lengths of intervals

Length of intervals	200[a]	300[b]	400[c]	500	600	700	800	900	1000[d]
MSE	104 640	78 792	124 707	173 453	254 592	222 557	365 045	246 892	407 521

[a] The second best results.
[b] Average-based length; the best results.
[c] Distribution-based length; the third best results.
[d] The length of intervals in the previous studies.

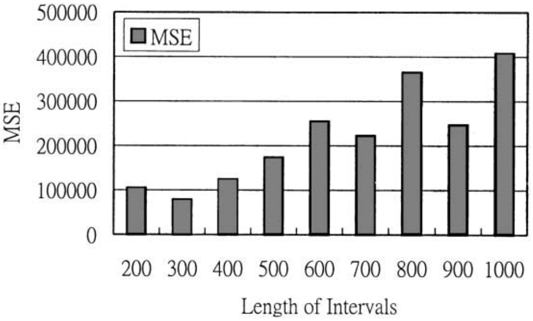

Fig. 2. Enrollment forecasting using Chen's model with different lengths of intervals.

average-based length gives better result than both distribution-based length and randomly chosen length 1000.

Third, other lengths of intervals ranging from 200 to 900 (100 based) are chosen as the lengths of intervals for Chen's model. Among the various lengths, it is found that average- and distribution-based lengths give the best and third best forecasts, respectively. The results are shown in Table 6 and compared in Fig. 2.

It is worth noting that "the smaller the lengths, the better the forecasting results" does not necessarily hold true. In this example, the smallest length 200 does not forecast better than 300.

5. TAIEX forecasting

In order to provide a large amount of data for forecasting, data sets from Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) were chosen. Two data sets were used: (1) from January 4, 1996 to December 31, 1996 and (2) from January 4, 1997 to December 31, 1997. Again, the forecasting results of using Chen's model with various interval lengths are compared with those derived from distribution- and average-based lengths.

5.1. 1996 forecasting

For 1996, the range for the fuzzy sets is from 4600 to 7000. First, distribution-based length is calculated:

1. The first differences are calculated and their average is 48.
2. From Table 1, the base for the length of intervals is 10.
3. Plot the cumulative distribution of the first differences. There are 288 first differences in total.
4. From the distribution, there are 134 first differences smaller than 30. 30 is the largest length smaller than half the first differences. Hence, 30 is chosen as the length of intervals.

When the length 30 is applied to Chen's model, the MSE is 3312.

Second, average-based length is calculated:

1. The first differences are calculated and the average of the first differences is 48.
2. Take half the average as the length, which is 24.
3. Given the length chosen in step 2, the base for length of intervals is determined as 10 by following Table 1.
4. Round the length 24 by base 10, which is 20. 20 is chosen as the length.

When average-based length is applied to Chen's model, the MSE is 3004.

Third, other lengths of intervals ranging from 20 to 100 (10 based) are chosen as the lengths of intervals for Chen's model. It is found that average- and

Table 7
1996 TAIEX forecasting using Chen’s model with different lengths of intervals

Length of intervals	20[a]	30[b]	40	50	60	70	80	90	100
MSE	3004	3312	3634	3817	4243	4623	4133	4969	5246

[a] Average-based length; the best results.

[b] Distribution-based length; the second best results.

distribution-based lengths provide the best and second best forecasting results among those lengths. The results are shown in Table 7 and Fig. 3.

5.2. 1997 forecasting

For 1997, the range for the fuzzy sets is set from 6800 to 10 200. First, distribution-based length is calculated:

1. The first differences are calculated and their average is 94.4.
2. From Table 1, the base for the length of intervals is 10.
3. Plot the cumulative distribution of the first differences.
4. From the distribution, there are 97 first differences (287 in total) smaller than 60. 60 is the largest length smaller than half the first differences. Hence, 60 is chosen as the length of intervals.

When the length 60 is applied to Chen’s model, the MSE is 13 189.

Second, average-based length is calculated:

1. The first differences are calculated and the average of the first differences is 94.4.
2. Take half the average as the length, which is 47.2.
3. Given the length chosen in step 2, the base for length of intervals is determined as 10 by following Table 1.
4. Round the length 47.2 by base 10, which is 50. Hence 50 is chosen as the length.

When average-based length is applied to Chen’s model, the MSE is 12 356.

Third, other lengths of intervals ranging from 40 to 120 (10 based) are chosen as the lengths of intervals for Chen’s model. Average- and distribution-based lengths provide the second and the third best forecasting results among those lengths as in Table 8 and Fig. 4.

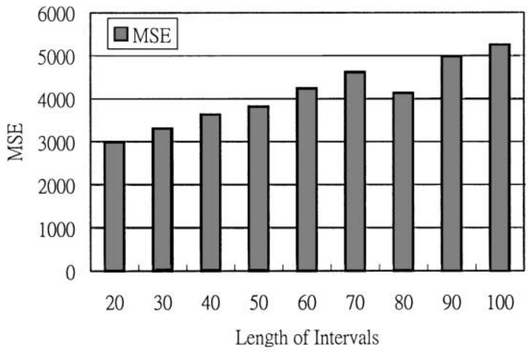

Fig. 3. 1996 TAIEX forecasting using Chen’s model with different lengths of intervals.

6. Conclusion

The determination of effective lengths of intervals is critical for forecasting in fuzzy time series. The objective of this study is to provide effective lengths of intervals to improve forecasting. Average- and distribution-based length are proposed. Various types of time series data (yearly and daily) from two domains (enrollment and stock indices) are used for empirical analysis. Results show that, first, Chen’s model with distribution- and average-based lengths forecast better than the same model with randomly chosen lengths for enrollment forecasting. Second, using Chen’s model, distribution- and average-based lengths forecast better than many randomly chosen lengths for TAIEX forecasting. In addition, both distribution- and average-based lengths are simple to calculate; hence, both can be used as effective lengths to improve forecasting in fuzzy time series.

Table 8
1997 TAIEX forecasting using Chen's model with different lengths of intervals

Length of intervals	40[a]	50[b]	60[c]	70	80	90	100	110	120
MSE	11 509	12 356	13 189	13 416	13 832	16 724	15 313	16 957	16 587

[a] The best results.

[b] Average-based length; the second best results.

[c] Distribution-based length; the third best results.

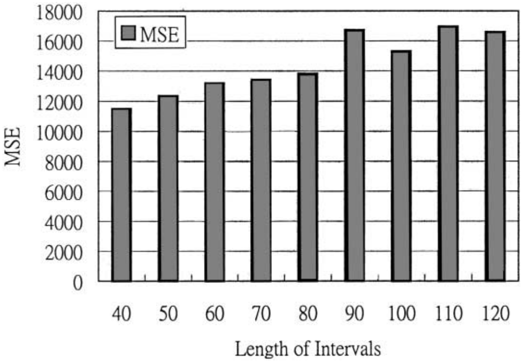

Fig. 4. 1997 TAIEX forecasting using Chen's model with different lengths of intervals.

References

[1] S.-M. Chen, Forecasting enrollments based on fuzzy time series, Fuzzy Sets and Systems 81 (1996) 311–319.

[2] Q. Song, B.S Chissom, Fuzzy time series and its models, Fuzzy Sets and Systems 54 (1993) 269–277.

[3] Q. Song, B.S. Chissom, Forecasting enrollments with fuzzy time series – part 1, Fuzzy Sets and Systems 54 (1993) 1–9.

[4] Q. Song, B.S. Chissom, Forecasting enrollments with fuzzy time series – part 2, Fuzzy Sets and Systems 62 (1994) 1–8.