



A Statistically Validated Membership-Weighted Fuzzy Time Series Model with Triangular Fuzzy Forecasts

Maheswari D^{a*}

^a Associate Professor, Sri Krishna Knowledge Foundation (Deemed to be University), Coimbatore, Tamil Nadu, India

ABSTRACT

Fuzzy time series (FTS) models are widely used for forecasting short and imprecise series, yet most published models rely on subjectively chosen equal-length intervals, discard membership information through crisp “maximum-membership” fuzzification, and are compared with benchmarks only through point accuracy measures without formal statistical inference. This paper proposes a membership-weighted fuzzy time series model (FCM-MW-FTS) that closes these gaps. The state of the series (level or first difference) is selected by an augmented Dickey–Fuller test; the universe of discourse is partitioned by fuzzy c-means clustering into a Ruspini family of triangular fuzzy sets; soft memberships of consecutive observations are combined into a recency-discounted fuzzy transition matrix; and each forecast is delivered as a triangular fuzzy number whose core is the point forecast. The number of fuzzy sets and the discount factor are chosen by rolling-origin cross-validation inside the training sample. We show that the transition matrix is row-stochastic, that the model nests the Chen and fuzzy Markov formulations as special cases, and that under stationarity and ergodicity it converges almost surely to a population fuzzy transition kernel. Forecast accuracy is assessed with the Harvey–Leybourne–Newbold corrected Diebold–Mariano test, the Wilcoxon signed-rank test, a moving-block bootstrap of the RMSE ratio and Ljung–Box residual diagnostics. On the benchmark Alabama enrolment series, the proposed model reduces the mean squared error by 49.3% relative to Chen’s model and by 10.6% relative to ARIMA (1,1,0). In genuine out-of-sample one-step forecasting of the Nile River annual flow (1946–1970) and the Lake Huron level (1948–1972), it attains the lowest MAE, MAPE and MASE among six competitors, matches ARIMA in RMSE, and is significantly better than the random walk (Nile) and Holt’s method (Lake Huron) by bootstrap and Wilcoxon inference. A Monte Carlo study with 200 replications under four data-generating processes shows that the model is close to efficient for linear autoregressions and outperforms ARIMA in 85% of replications under regime-switching (SETAR) dynamics.

Keywords: Fuzzy time series, fuzzy c-means, triangular fuzzy number; fuzzy transition matrix, Diebold–Mariano test, bootstrap, forecast evaluation

Mathematics Subject Classification (2020): 62M10, 62M20, 03E72, 62H30, 62G09

1. Introduction

Many series that decision makers must forecast are short, noisy and recorded with limited precision: annual enrolments, river discharges, lake levels, crop yields or regional sales. Classical statistical models such as autoregressive integrated moving average (ARIMA) processes [1] rely on distributional assumptions and asymptotic arguments that are difficult to verify with a few dozen observations. Fuzzy set theory [2] offers a complementary way of representing such data: instead of treating an observation as a precise point, it is described by its degrees of membership in linguistic categories such as “low”, “medium” or “high”, and the dynamics of the series are captured by relationships among these categories.

Song and Chissom [3–5] introduced fuzzy time series and applied them to the enrolments of the University of Alabama, a series that has since become the standard benchmark of the field. Chen [6] simplified the computations by replacing max–min composition with arithmetic averaging over fuzzy logical relationship groups (FLRGs), and Yu [7] introduced chronological weights so that recurrent and recent relationships carry more influence. Subsequent research has focused mainly on the partition of the universe of discourse: Huarng [8] and Huarng and Yu [9] studied distribution-based and ratio-based interval lengths, Stevenson and Porter [10] used percentage changes as the universe of discourse, and clustering approaches based on fuzzy c-means (FCM) [11–14] and the Gustafson–Kessel algorithm [15] replaced arbitrary interval boundaries by data-driven centres. Bose and Mali [16] provide a recent survey.

Despite this progress, three weaknesses remain common in the literature. First, most models fuzzify each observation into the single fuzzy set with the largest membership, so the graded information that motivated the fuzzy representation in the first place is discarded at the very first step. Second, the choice between modelling levels and modelling changes is rarely justified statistically, although it determines whether a model can mean-revert or extrapolate a trend. Third, competing models are usually ranked by in-sample mean squared error on a single short series. Differences of a few percent in such a ranking may be due to chance, and Sullivan and Woodall [17] showed early on that fuzzy forecasts should be compared with properly specified statistical alternatives such as Markov models.

This paper addresses all three issues within one coherent framework, which we call the membership-weighted fuzzy time series model (FCM-MW-FTS). Its contributions are as follows.

1. **Statistically guided state representation.** An augmented Dickey–Fuller (ADF) test [18, 19] decides whether fuzzy relationships are learned on levels (stationary series) or on first differences (unit-root series).
2. **Soft fuzzification and a fuzzy transition matrix.** FCM centres define a Ruspini partition [20] of triangular fuzzy sets. The full membership vectors of consecutive states are combined into a recency-discounted fuzzy transition matrix, which generalises the FLRGs of Chen [6] and Yu [7] and the fuzzy Markov formulation of Tsaur [21].
3. **Triangular fuzzy forecasts.** Each forecast is a triangular fuzzy number obtained by exact fuzzy arithmetic, providing a possibilistic range in addition to the point forecast.
4. **Theoretical properties.** We prove row-stochasticity, nesting of classical models and almost-sure convergence of the estimated transition matrix under stationarity and ergodicity.
5. **Formal statistical validation.** Forecasts are evaluated by the Diebold–Mariano test with the Harvey–Leybourne–Newbold small-sample correction [22, 23], the Wilcoxon signed-rank test, a moving-block bootstrap [24] and Ljung–Box diagnostics [25], on three real series and in a Monte Carlo experiment.

The remainder of the paper is organised as follows. Section 2 reviews preliminaries. Section 3 presents the proposed model and its properties. Section 4 describes the data and the evaluation design. Section 5 reports empirical results, Section 6 reports the Monte Carlo study and ablation analysis, Section 7 discusses the findings and Section 8 concludes.

2. Preliminaries

2.1 Fuzzy sets and triangular fuzzy numbers

Let U be a universe of discourse. A fuzzy set A on U is characterised by a membership function $\mu_A: U \rightarrow [0,1]$ [2]. A triangular fuzzy number (TFN) $\tilde{a} = (l, m, u)$ with $l \leq m \leq u$ has membership

$$\mu_{\tilde{a}}(x) = \begin{cases} \frac{x-l}{m-l}, & l \leq x \leq m, \\ \frac{u-x}{u-m}, & m \leq x \leq u, \\ 0, & \text{otherwise.} \end{cases}$$

Its α -cut is $\tilde{a}_\alpha = [l + \alpha(m-l), u - \alpha(u-m)]$ for $\alpha \in [0,1]$. For TFNs $\tilde{a}_j = (l_j, m_j, u_j)$ and non-negative scalars π_j , extension-principle arithmetic gives

$$\oplus_j \pi_j \tilde{a}_j = \left(\sum_j \pi_j l_j, \sum_j \pi_j m_j, \sum_j \pi_j u_j \right),$$

which is again a TFN.

2.2 Fuzzy time series

Following Song and Chissom [3], let $Y(t)$, $t = \dots, 0, 1, 2, \dots$, be a subset of $\mathbb{R}$ on which fuzzy sets $A_1, \dots, A_k$ are defined, and let $F(t)$ be a collection of these fuzzy sets. Then $F(t)$ is a fuzzy time series on $Y(t)$. If $F(t)$ is caused by $F(t-1)$ only, the relationship is written $F(t-1) \rightarrow F(t)$; when $F(t-1) = A_i$ and $F(t) = A_j$ the fuzzy logical relationship (FLR) is $A_i \rightarrow A_j$. Grouping all FLRs with the same left-hand side gives an FLRG. In Chen's model [6] the forecast for a period whose previous state is A_i is the arithmetic mean of the interval midpoints of the distinct right-hand sides of the FLRG of A_i ; Yu's model [7] replaces the arithmetic mean by a weighted mean in which the r -th chronological occurrence receives weight r .

2.3 Fuzzy c-means clustering

Given observations $s_1, \dots, s_n$ and a number of clusters k , FCM [11, 12] minimises

$$J_m(\mathbf{U}, \mathbf{v}) = \sum_{t=1}^n \sum_{j=1}^k u_{tj}^m (s_t - v_j)^2, \quad \sum_{j=1}^k u_{tj} = 1, \quad u_{tj} \in [0,1],$$

with fuzzifier $m > 1$ (we use $m = 2$). The alternating optimality conditions are

$$u_{tj} = \left[\sum_{r=1}^k \left(\frac{|s_t - v_j|}{|s_t - v_r|} \right)^{2/(m-1)} \right]^{-1}, \quad v_j = \frac{\sum_t u_{tj}^m s_t}{\sum_t u_{tj}^m}.$$

The Xie–Beni index [26], $XB = J_2 / \{n \min_{i \neq j} (v_i - v_j)^2\}$, is a common cluster-validity criterion; smaller values indicate compact and well-separated clusters.

3. The proposed FCM-MW-FTS model

The model is summarised in Figure 1 and described step by step below.

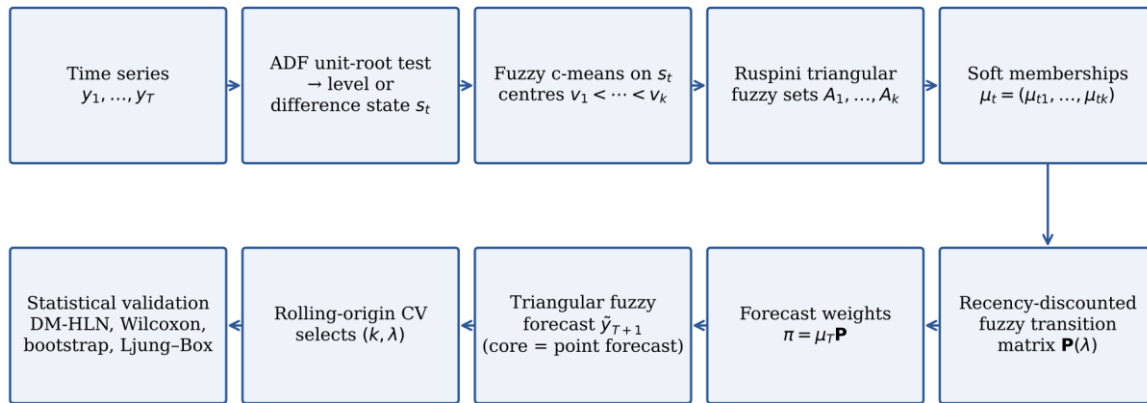


Fig. 1 - Framework of the proposed membership-weighted fuzzy time series model (FCM-MW-FTS).

3.1 Step 1: statistically guided state representation

Let $y_1, \dots, y_T$ be the training series. We estimate the ADF regression

$$\Delta y_t = \beta_0 + \gamma y_{t-1} + \sum_{i=1}^p \phi_i \Delta y_{t-i} + \varepsilon_t,$$

with the lag order $p \in \{0, \dots, 4\}$ chosen by the Akaike information criterion, and test $H_0: \gamma = 0$ (unit root) using $\tau = \hat{\gamma}/\text{se}(\hat{\gamma})$ and the 5% MacKinnon critical value -2.89 [19]. If H_0 is rejected, the state is the level, $s_t = y_t$; otherwise, the state is the first difference, $s_t = \Delta y_t = y_t - y_{t-1}$. The same decision fixes the differencing order d of the ARIMA benchmark, so that both models are compared on the same footing.

3.2 Step 2: FCM partition of the universe of discourse

FCM with k clusters is applied to the states $\{s_t\}$, and the centres are sorted, $v_1 < v_2 < \dots < v_k$. Let $R = \max_t s_t - \min_t s_t$ and set $v_0 = \min_t s_t - 0.1R$, $v_{k+1} = \max_t s_t + 0.1R$. Fuzzy set A_j is the TFN $(l_j, m_j, u_j) = (v_{j-1}, v_j, v_{j+1})$, except that A_1 and A_k are shouldered: $\mu_{A_1}(x) = 1$ for $x \leq v_1$ and $\mu_{A_k}(x) = 1$ for $x \geq v_k$.

Proposition 1 (Ruspini partition). For every $x \in \mathbb{R}$, $\sum_{j=1}^k \mu_{A_j}(x) = 1$, and at most two memberships are non-zero.

Proof. If $x \leq v_1$ or $x \geq v_k$ only the shouldered set is active, with membership 1. Otherwise $v_j \leq x < v_{j+1}$ for a unique j , and only A_j and A_{j+1} are non-zero, with $\mu_{A_j}(x) = (v_{j+1} - x)/(v_{j+1} - v_j)$ and $\mu_{A_{j+1}}(x) = (x - v_j)/(v_{j+1} - v_j)$, which sum to 1. $\square$

Each state is therefore fuzzified into a membership vector $\mu_t = (\mu_{A_1}(s_t), \dots, \mu_{A_k}(s_t))$ lying on the probability simplex. Unlike the maximum-membership rule, this soft fuzzification preserves where s_t lies between two neighbouring linguistic categories.

3.3 Step 3: recency-discounted fuzzy transition matrix

Let $s_1, \dots, s_n$ denote the available states ($n = T$ for levels, $n = T - 1$ for differences). For a discount factor $\lambda \in (0, 1]$ define the fuzzy transition weights

$$w_{ij}(\lambda) = \sum_{t=2}^n \lambda^{n-t} \mu_{t-1,i} \mu_{t,j}, \quad i, j = 1, \dots, k,$$

and the fuzzy transition matrix $\mathbf{P}(\lambda) = (P_{ij})$ with

$$P_{ij} = \begin{cases} w_{ij} / \sum_{r=1}^k w_{ir}, & \sum_r w_{ir} > 0, \\ \delta_{ij}, & \text{otherwise,} \end{cases}$$

where δ_{ij} is the Kronecker delta. The empty-row rule reproduces Chen's convention that a state with no recorded successor forecasts itself.

Proposition 2 (properties of $\mathbf{P}$). (i) $\mathbf{P}(\lambda)$ is row-stochastic. (ii) If memberships are crisp (μ_t a unit vector) and $\lambda = 1$, P_{ij} equals the relative frequency of the FLR $A_i \rightarrow A_j$; the forecast in (8) below is then the frequency-weighted FLRG mean, i.e. the fuzzy Markov model of Tsaur [21], and Chen's model [6] is obtained when repeated FLRs are counted once. (iii) With crisp memberships and $\lambda < 1$, recent relationships are up-weighted geometrically, in the spirit of Yu [7].

Proof. (i) $W_{ij} \geq 0$, so each non-empty row of (7) is a probability vector; empty rows are unit vectors. (ii) With unit-vector memberships $\mu_{t-1,i}\mu_{t,j} = 1$ exactly when $s_{t-1} \in A_i$ and $s_t \in A_j$, so W_{ij} counts the FLR $A_i \rightarrow A_j$. (iii) follows from the weights λ^{n-t} . ◻

Proposition 3 (consistency). Suppose $\{s_t\}$ is strictly stationary and ergodic, and $\lambda = 1$. Then, as $n \rightarrow \infty$,

$$\frac{1}{n-1} W_{ij} \xrightarrow{\text{a.s.}} E[\mu_{A_i}(s_{t-1}) \mu_{A_j}(s_t)], \quad P_{ij} \xrightarrow{\text{a.s.}} P_{ij}^* = \frac{E[\mu_{A_i}(s_{t-1}) \mu_{A_j}(s_t)]}{E[\mu_{A_i}(s_{t-1})]},$$

whenever $E[\mu_{A_i}(s_{t-1})] > 0$.

Proof. The function $(x, y) \mapsto \mu_{A_i}(x) \mu_{A_j}(y)$ is bounded and measurable, so Birkhoff's ergodic theorem gives the first limit. Summing over j and using Proposition 1, $\frac{1}{n-1} \sum_j W_{ij} \rightarrow E[\mu_{A_i}(s_{t-1})]$, and the continuous-mapping theorem yields the second limit. ◻

P_{ij}^* can be interpreted as the membership-weighted probability of moving into category A_j given that the process is "in" category A_i ; Proposition 3 shows that the fuzzy relationships learned by the model are estimates of a well-defined population quantity, not merely descriptive summaries.

3.4 Step 4: triangular fuzzy forecast

Given the most recent state s_n with membership vector μ_n , the forecast weights are $\pi = \mu_n \mathbf{P}(\lambda)$, which satisfy $\pi_j \geq 0$ and $\sum_j \pi_j = 1$. By (2) the fuzzy forecast of the next state is the TFN

$$\tilde{s}_{n+1} = \bigoplus_{j=1}^k \pi_j A_j = \left(\sum_j \pi_j v_{j-1}, \sum_j \pi_j v_j, \sum_j \pi_j v_{j+1} \right) = (\hat{s}^L, \hat{s}, \hat{s}^U).$$

The fuzzy forecast of the series is $\tilde{y}_{T+1} = b_T \oplus \tilde{s}_{n+1}$, with $b_T = 0$ in the level representation and $b_T = y_T$ in the difference representation. The core $\hat{y}_{T+1} = b_T + \hat{s}$ is the point forecast, the support $[b_T + \hat{s}^L, b_T + \hat{s}^U]$ is the range of possible values and the 0.5-cut is the range with membership at least one half.

Proposition 4 (boundedness). In the level representation $v_1 \leq \hat{y}_{T+1} \leq v_k$, so forecasts revert towards the observed range; in the difference representation $v_1 \leq \hat{y}_{T+1} - y_T \leq v_k$, so the predicted change is bounded by the extreme cluster centres of historical changes.

Proof. $\hat{s}$ is a convex combination of $v_1, \dots, v_k$ because π lies on the simplex. ◻

Proposition 4 explains why the representation step matters: a level model cannot extrapolate a trend, whereas a difference model cannot mean-revert. The ablation analysis in Section 6.2 confirms the practical importance of this choice.

3.5 Step 5: hyper-parameter selection

The number of fuzzy sets $k \in \{3, \dots, 10\}$ and the discount factor $\lambda \in \{0.80, 0.90, 0.95, 1.00\}$ are selected by rolling-origin cross-validation within the training sample. The last 30% of training periods are used as validation origins; at each origin the model is re-fitted on the preceding observations and a one-step forecast is produced, and the pair minimising the validation RMSE is retained. The test sample is never used for any choice. The procedure is summarised in Algorithm 1.

4. Algorithm 1. FCM-MW-FTS one-step forecasting

Input: training series $y_1, \dots, y_T$; grids for k and λ .

1. Apply the ADF test (5) and set $s_t = y_t$ or $s_t = \Delta y_t$.
2. For each (k, λ) in the grid, compute the rolling-origin validation RMSE over the last 30% of training periods using steps 3–6.
3. Run FCM (3)–(4) on $\{s_t\}$ with $m = 2$; sort the centres and build the triangular sets $A_1, \dots, A_k$.
4. Compute the membership vectors μ_t (Proposition 1).
5. Compute $\mathbf{W}(\lambda)$ and $\mathbf{P}(\lambda)$ by (6)–(7).
6. Compute $\pi = \mu_n \mathbf{P}$ and the fuzzy forecast (8).
7. Retain the (k, λ) with the smallest validation RMSE; at each test origin re-fit steps 3–6 on all data available at that origin.

Output: point forecasts $\hat{y}_t$ and triangular fuzzy forecasts $\tilde{y}_t$.

5. Data and evaluation design

5.1 Data

Three real, publicly documented annual series were used (Figure 2).

Alabama enrolments (1971–1992, $n = 22$). The benchmark series of the FTS literature introduced by Song and Chissom [4] and analysed by Chen [6] and many later authors.

Nile river annual flow at Aswan (1871–1970, $n = 100$, units 10^8 m^3). A classical hydrological series studied by Cobb [27] and Durbin and Koopman [28]; it contains a well-known downward shift near 1898.

Lake Huron annual mean level (1875–1972, $n = 98$, feet). A classical environmental series used by Brockwell and Davis [29].

The Nile and Lake Huron series are distributed with the R datasets package; the values used here were checked against their published means (919.35 and 579.004, respectively). For the Nile and Lake Huron series the last 25 years form the test sample, giving training periods 1871–1945 and 1875–1947, respectively. Following the convention of the FTS literature, the enrolment series is used for an in-sample comparison with published models.

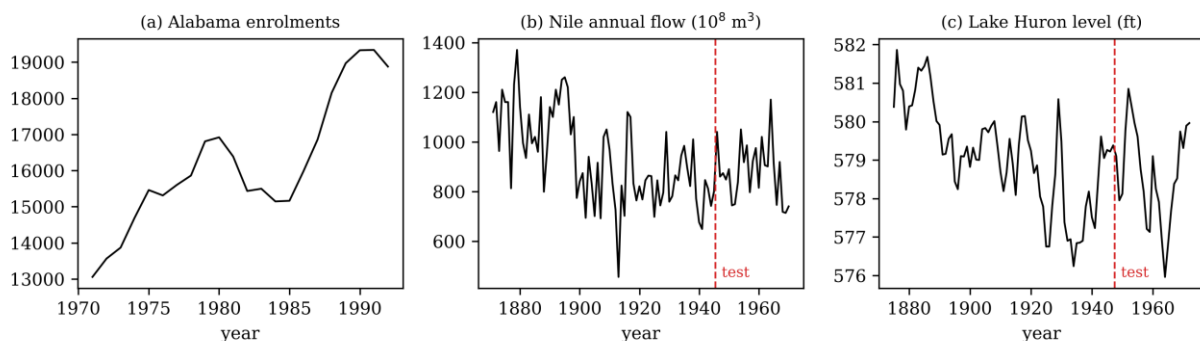


Fig. 2 - The three series used in the study; the dashed line marks the beginning of the out-of-sample test period.

Table 1 reports descriptive statistics, Shapiro–Wilk normality tests and ADF tests. None of the series departs significantly from normality at the 5% level. The ADF test rejects a unit root for the Nile levels ($\tau = -3.380$) but not for the enrolment ($\tau = -1.965$) or Lake Huron ($\tau = -2.124$) levels, so the Nile series is modelled in levels and the other two in first differences. For the short enrolment series the test on differences is also inconclusive ($\tau = -2.128$), reflecting the low power of unit-root tests with 21 observations; differencing is retained as the more conservative choice for a clearly trending series.

Table 1 - Descriptive statistics, normality and unit-root tests (5% ADF critical value -2.89).

Statistic	Enrolments (1971–1992)	Nile, training (1871–1945)	Lake Huron, training (1875–1947)
n	22	75	73
Mean	16,194.18	931.76	579.14
Standard deviation	1,816.49	182.41	1.32
Minimum	13,055.00	456.00	576.24
Maximum	19,337.00	1,370.00	581.86
Skewness	0.337	0.187	-0.158
Excess kurtosis	-0.692	-0.550	-0.469
Shapiro–Wilk W (p), levels	0.939 (0.186)	0.972 (0.090)	0.980 (0.289)
Shapiro–Wilk W (p), differences	0.973 (0.788)	0.988 (0.731)	0.986 (0.612)
ADF τ , levels (lag)	-1.965 (1)	-3.380 (1)	-2.124 (2)
ADF τ , differences (lag)	-2.128 (0)	-6.303 (3)	-7.010 (1)
State representation selected	Difference	Level	Difference

5.2 Competing models

Five competitors were implemented and, in the out-of-sample exercise, re-estimated at every forecast origin on an expanding window:

1. **Random walk (RW):** $\hat{y}_{t+1} = y_t$.
2. **ARIMA($p, d, 0$):** d from the ADF test of Section 3.1 and $p \leq 4$ chosen by AIC on a common estimation sample, estimated by conditional least squares [1].
3. **Holt's linear exponential smoothing** [30] with smoothing constants estimated by minimising the in-sample one-step squared error.
4. **Chen (1996)** [6] with seven equal-length intervals covering the training range extended by 5% at each end (for enrolments, Chen's original partition [13000,20000] with intervals of length 1000).
5. **Yu (2005)** [7] with the same partition and chronological weights.

5.3 Accuracy measures

For forecast errors $e_t = y_t - \hat{y}_t$ over h test periods we report

$$\text{RMSE} = \sqrt{1/h \sum e_t^2}, \quad \text{MAE} = 1/h \sum |e_t|, \quad \text{MAPE} = 100/h \sum |e_t/y_t|, \quad \text{MASE} = \frac{\text{MAE}}{\frac{1}{T-1} \sum_{t=2}^T |y_t - y_{t-1}|},$$

where the MASE [31] is scaled by the in-sample mean absolute error of the naive forecast, so that values below one indicate an improvement over the naive method.

Statistical inference on forecast accuracy

Diebold–Mariano test with HLN correction. For the loss differential $d_t = e_{1t}^2 - e_{2t}^2$ between the proposed model (1) and a competitor (2), the DM statistic is $DM = \bar{d}/\sqrt{\hat{\sigma}_d^2/h}$ [22]. For forecast horizon q the Harvey–Leybourne–Newbold correction [23] is

$$S^* = DM \sqrt{\frac{h+1-2q+q(q-1)/h}{h}} \stackrel{q=1}{=} DM \sqrt{\frac{h-1}{h}},$$

which is compared with Student's t distribution on $h-1$ degrees of freedom. A negative statistic favours the proposed model.

Wilcoxon signed-rank test. A distribution-free test of the null hypothesis that the paired absolute errors $|e_{1t}|$ and $|e_{2t}|$ have a median difference of zero.

Moving-block bootstrap. Blocks of length 4 are resampled $B = 5000$ times [24] to obtain a 95% percentile interval for the ratio $\text{RMSE}_1/\text{RMSE}_2$; an interval lying entirely below 1 indicates a significant improvement while allowing for serial dependence in the errors.

Residual diagnostics. The Ljung–Box statistic $Q(L) = h(h+2) \sum_{r=1}^L \hat{\rho}_r^2 / (h-r)$ [25] with $L = 10$ checks whether the forecast errors are free of remaining autocorrelation.

Fuzzy forecast coverage. For each test period we record whether the actual value lies in the support and in the 0.5-cut of the TFN forecast (8).

All computations were carried out in Python (NumPy and SciPy) with fixed random seeds.

6. Empirical results

6.1 Alabama enrolments: comparison with the classical literature

As a check on the implementation, Chen's model with his original seven intervals was reproduced first. The resulting MSE over 1972–1992 is 407,521, against 407,507 reported in the literature; the small difference arises only from Chen's rounding of 16,833.3 to 16,833. Following the published comparisons, the proposed model was fitted with $k = 7$ fuzzy sets (matching Chen's seven intervals) and $\lambda = 1$. The FCM centres of the yearly changes are -953.4 , -446.1 , 24.4 , 296.5 , 505.6 , 839.9 and 1288.5 students. Because the first difference of 1972 is needed as the initial state, all models are compared over the common period 1973–1992. Table 2 lists the forecasts and Figure 3 displays them.

Table 2 - In-sample one-step forecasts of Alabama enrolments; the last column gives the triangular fuzzy forecast of the proposed model.

Year	Actual	Chen (1996)	Yu (2005)	Proposed	Fuzzy forecast (l, m, u)
1973	13,867	14,000	14,000	13,799	(13,461, 13,799, 14,059)
1974	14,696	14,000	14,000	14,275	(13,905, 14,275, 14,623)
1975	15,460	15,500	15,500	15,261	(14,893, 15,261, 15,582)

Year	Actual	Chen (1996)	Yu (2005)	Proposed	Fuzzy forecast (l, m, u)
1976	15,311	16,000	15,789	15,958	(15,597, 15,958, 16,267)
1977	15,603	16,000	15,789	15,193	(14,806, 15,193, 15,585)
1978	15,861	16,000	15,789	16,011	(15,639, 16,011, 16,363)
1979	16,807	16,000	15,789	16,218	(15,840, 16,218, 16,577)
1980	16,919	16,833	17,000	17,417	(17,053, 17,417, 17,757)
1981	16,388	16,833	17,000	17,057	(16,654, 17,057, 17,447)
1982	15,433	16,833	17,000	16,099	(15,751, 16,099, 16,451)
1983	15,497	16,000	15,789	15,351	(14,940, 15,351, 15,649)
1984	15,145	16,000	15,789	15,563	(15,152, 15,563, 15,964)
1985	15,163	16,000	15,789	14,882	(14,530, 14,882, 15,254)
1986	15,984	16,000	15,789	15,165	(14,748, 15,165, 15,573)
1987	16,859	16,000	15,789	16,540	(16,174, 16,540, 16,860)
1988	18,150	16,833	17,000	17,446	(17,079, 17,446, 17,775)
1989	18,970	19,000	19,167	18,871	(18,522, 18,871, 19,265)
1990	19,328	19,000	19,167	19,525	(19,159, 19,525, 19,845)
1991	19,337	19,000	18,833	19,689	(19,328, 19,689, 20,013)
1992	18,876	19,000	18,833	19,333	(18,918, 19,333, 19,740)

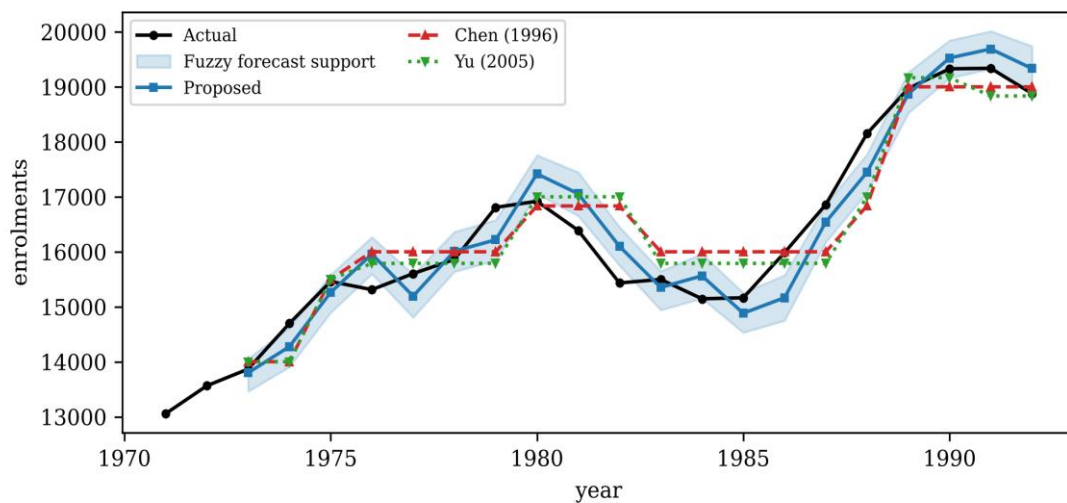


Fig. 3 - Alabama enrolments: actual values, forecasts of Chen (1996), Yu (2005) and the proposed model, and the support of the proposed triangular fuzzy forecasts.

Table 3 - Accuracy of the enrolment forecasts, 1973–1992 ($h = 20$). The DM-HLN statistic compares the proposed model with each competitor (negative values favour the proposed model).

Model	MSE	RMSE	MAE	MAPE (%)	MASE	DM-HLN statistic (p)
Proposed FCM-MW-FTS	212,049	460.49	405.46	2.476	0.794	—
Chen (1996)	418,349	646.80	501.90	3.104	0.983	-1.809 (0.086)
Yu (2005)	418,121	646.62	488.21	3.001	0.957	-1.603 (0.125)
Random walk	394,333	627.96	510.45	3.096	1.000	-2.069 (0.052)

Model	MSE	RMSE	MAE	MAPE (%)	MASE	DM-HLN statistic (p)
ARIMA(1,1,0)	237,269	487.10	417.61	2.558	0.818	-1.245 (0.228)
Holt	300,104	547.82	476.07	2.901	0.933	-1.533 (0.142)

The proposed model gives the smallest value of every accuracy measure (Table 3). Its MSE of 212,049 is 49.3% lower than that of Chen's model (418,349 on the same period), 49.3% lower than Yu's model, 29.3% lower than Holt's method and 10.6% lower than the ARIMA(1,1,0) benchmark. The MASE of 0.794 indicates that the model removes about one fifth of the error of the naive forecast. The DM-HLN statistics are negative against all competitors; the improvements over the random walk ($p = 0.052$) and over Chen's model ($p = 0.086$) are significant at the 10% level, whereas the differences from ARIMA and Holt are not statistically significant. With only 20 forecasts this is an important qualification that is absent from most published enrolment comparisons, where rankings are based on MSE alone. The residuals pass the Ljung–Box test ($Q(5) = 6.768$, $p = 0.238$). The in-sample supports of the fuzzy forecasts contain 45% of the actual values; with $k = 7$ sets fitted to 21 changes the fuzzy sets are narrow, so the support is best read as a possibilistic band around the core rather than as a calibrated prediction interval.

6.2 Out-of-sample forecasting: Nile and Lake Huron

Table 4 reports the hyper-parameters chosen inside the training samples. For the Nile series, cross-validation selects four fuzzy sets on the levels with discount $\lambda = 0.8$; for Lake Huron it selects three fuzzy sets on the yearly changes with no discounting. For the Nile, the Xie–Beni index would have favoured $k = 5$ (XB = 0.078 against 0.115 at $k = 4$); the predictive criterion is preferred because the goal is forecasting rather than clustering. Figure 4 shows the resulting partitions.

Table 4 - Hyper-parameters selected by rolling-origin cross-validation on the training samples.

Item	Nile	Lake Huron
Training / test period	1871–1945 / 1946–1970	1875–1947 / 1948–1972
State representation (ADF-guided)	level	diff
Selected k (rolling-origin CV)	4	3
Selected λ	0.8	1.0
Minimum CV RMSE	107.727	0.8903
FCM centres $v_1, \dots, v_k$	699.0, 828.0, 991.1, 1178.6	-0.740, 0.059, 0.985
Xie–Beni index at selected k	0.115	0.098
Benchmark ARIMA order (AIC)	(2,0,0)	(2,1,0)

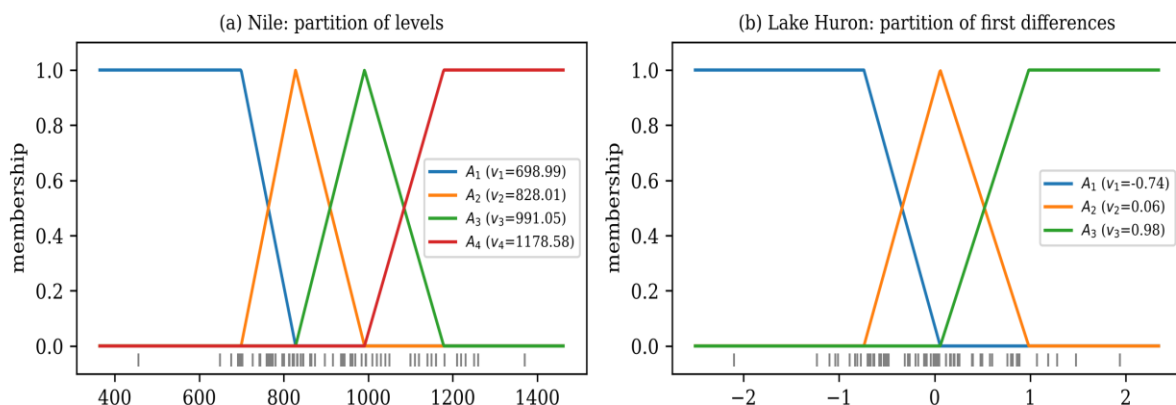


Fig. 4 - FCM-based Ruspini partitions of the training states: (a) Nile levels, (b) Lake Huron first differences. Grey ticks mark the training observations.

The estimated fuzzy transition matrices (Table 5) are directly interpretable. For the Nile, a year in the “very high flow” category A_4 (centre 1178.6) is followed by “low” or “below-average” flow (A_1 or A_2) with combined weight 0.959, and even from A_3 the largest weight (0.578) goes to A_2 . This expresses strong mean reversion towards the post-1898 regime documented by Cobb [27]. For Lake Huron, every row places its largest weight on the “no change” category A_2 (centre 0.059 ft), and the weight on a rise (A_3) increases from 0.168 after a fall to 0.277 after a rise, a mild positive persistence of changes that corresponds to the AR(2)-in-differences structure picked by AIC for the ARIMA benchmark.

Table 5 - Estimated fuzzy transition matrices $P(\lambda)$ on the training samples: (a) Nile; (b) Lake Huron.*Panel (a): Nile*

Nile	A1	A2	A3	A4
A1	0.401	0.557	0.040	0.002
A2	0.347	0.517	0.125	0.011
A3	0.208	0.578	0.213	0.001
A4	0.430	0.529	0.020	0.022

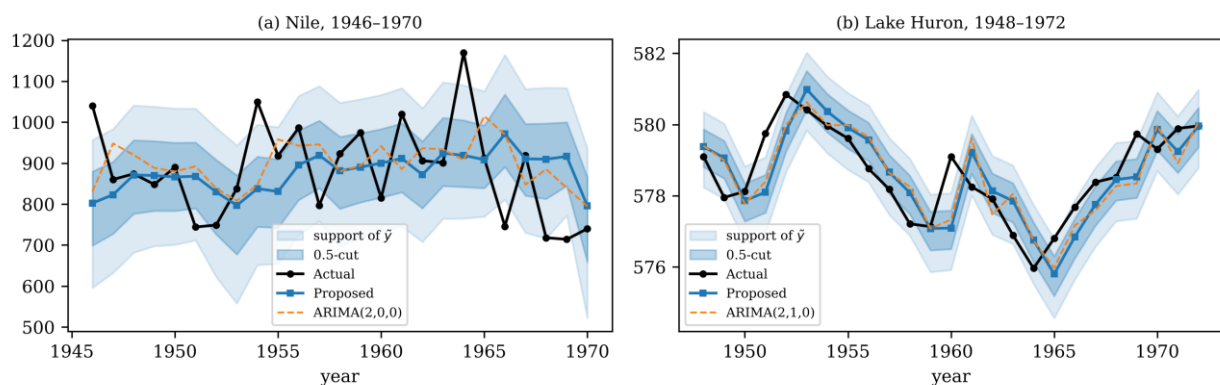
Panel (b): Lake Huron

Huron	A1	A2	A3
A1	0.360	0.472	0.168
A2	0.330	0.461	0.209
A3	0.315	0.408	0.277

Table 6 compares the out-of-sample accuracy over the 25 test years, and Figure 5 shows the forecasts together with the fuzzy forecast bands.

Table 6 - Out-of-sample one-step accuracy (Nile 1946–1970; Lake Huron 1948–1972; $h = 25$ each).

Model	Nile RMSE	Nile MAE	Nile (%)	MAPE	Nile MASE	Huron RMSE	Huron MAE	Huron (%)	MAPE	Huron MASE
Proposed FCM-MW-FTS	123.64	95.82	11.01		0.703	0.8547	0.7091	0.1225		1.317
Chen (1996)	131.29	113.31	12.96		0.831	0.9190	0.7314	0.1264		1.358
Yu (2005)	129.23	103.98	11.74		0.763	0.9569	0.7936	0.1372		1.473
Random walk	149.78	124.04	13.83		0.910	0.8591	0.7208	0.1246		1.338
ARIMA [Nile (2,0,0), Huron (2,1,0)]	123.07	102.06	11.68		0.748	0.8546	0.7201	0.1245		1.337
Holt	128.71	99.70	11.35		0.731	0.9247	0.7829	0.1353		1.454

**Fig. 5 - Out-of-sample one-step forecasts with the support and 0.5-cut of the triangular fuzzy forecasts: (a) Nile, (b) Lake Huron.**

For the Nile, the proposed model has the lowest MAE (95.82), MAPE (11.01%) and MASE (0.703) of all six models. Its MAE is 6.1% below ARIMA(2,0,0), 15.4% below Chen's model and 22.8% below the random walk. In RMSE it is second only to ARIMA, and by a negligible 0.5% (123.64 against 123.07). For Lake Huron, the proposed model again attains the lowest MAE (0.7091 ft), MAPE and MASE, and its RMSE (0.8547) coincides with that of ARIMA(2,1,0) (0.8546). The two classical FTS models are clearly worse on both series: Chen's RMSE is 6.2% (Nile) and 7.5% (Lake Huron) higher, and Yu's is 4.5% and 12.0% higher. The MASE values above one for Lake Huron indicate that the test period was more volatile than the training period for all methods.

6.3 Statistical significance and residual diagnostics

Table 7 reports the formal comparisons. All DM-HLN statistics against the fuzzy and smoothing competitors are negative, which is consistent with the point measures. Two improvements are statistically significant. For the Nile, the bootstrap 95% interval for the RMSE ratio relative to the random walk, [0.680, 0.990], lies entirely below 1 (DM-HLN $p = 0.086$, Wilcoxon $p = 0.090$). For Lake Huron, the proposed model is significantly more accurate than Holt's method by the Wilcoxon test ($p = 0.027$) and by the bootstrap interval [0.880, 0.978], with DM-HLN $p = 0.084$. The comparisons with ARIMA give DM-HLN statistics of 0.092 and 0.002 with $p > 0.9$: the proposed fuzzy model is statistically indistinguishable from a correctly specified linear benchmark, while additionally delivering an interpretable linguistic transition structure and fuzzy forecasts. The improvements over Chen's and Yu's models, although consistent in sign, are not significant with 25 forecasts, and the bootstrap intervals (for example [0.805, 1.054] against Chen for the Nile) indicate the plausible range of the gain.

Table 7 - Pairwise tests of the proposed model against each competitor (out-of-sample).

Series	Benchmark	DM-HLN (p)	Wilcoxon W (p)	95% bootstrap CI of RMSE ratio
Nile	Chen (1996)	-0.849 (0.404)	102 (0.107)	[0.805, 1.054]
	Yu (2005)	-0.544 (0.592)	144 (0.634)	[0.835, 1.090]
	Random walk	-1.788 (0.086)	99 (0.090)	[0.680, 0.990]
	ARIMA(2,0,0)	0.092 (0.928)	136 (0.491)	[0.894, 1.118]
	Holt	-0.738 (0.468)	122 (0.287)	[0.885, 1.052]
Lake Huron	Chen (1996)	-0.873 (0.392)	158 (0.916)	[0.811, 1.045]
	Yu (2005)	-1.117 (0.275)	124 (0.312)	[0.744, 1.040]
	Random walk	-0.297 (0.769)	127 (0.353)	[0.964, 1.018]
	ARIMA(2,1,0)	0.002 (0.998)	145 (0.653)	[0.890, 1.122]
	Holt	-1.803 (0.084)	81 (0.027)	[0.880, 0.978]

Table 8 - Ljung–Box tests on the out-of-sample forecast errors, $L = 10$.

Model	Nile Q(10)	p	Lake Huron Q(10)	p
Proposed FCM-MW-FTS	7.738	0.654	8.945	0.537
Chen (1996)	7.919	0.637	10.170	0.426
Yu (2005)	6.539	0.768	8.103	0.619
Random walk	17.401	0.066	9.519	0.484
ARIMA (selected order)	8.165	0.613	8.236	0.606
Holt	6.533	0.769	12.321	0.264

The Ljung–Box tests (Table 8) do not reject the white-noise hypothesis for the proposed model on either series ($p = 0.654$ and 0.537). In contrast, the random-walk errors for the Nile show borderline autocorrelation ($p = 0.066$), which is consistent with that series being stationary. The actual values lie inside the support of the fuzzy forecast in 84% (Nile) and 80% (Lake Huron) of the test years, and inside the 0.5-cut in 48% and 40% of the years, with average support widths of 344.4 (10^8 m^3) and 2.18 ft. The graded structure therefore behaves sensibly: the wider support captures most outcomes, and the narrower 0.5-cut captures roughly half.

7. Simulation evidence and ablation

7.1 Monte Carlo study

To assess the model beyond three series, 200 replications of length 120 (after a burn-in of 50) were generated from four processes with standard normal innovations: AR(1) with $\phi = 0.5$; AR(1) with $\phi = 0.8$; a random walk with drift 0.2; and a two-regime self-exciting threshold autoregression (SETAR) [32],

$$x_t = \begin{cases} 0.7 x_{t-1} + \varepsilon_t, & x_{t-1} \leq 0, \\ 1.0 - 0.4 x_{t-1} + \varepsilon_t, & x_{t-1} > 0. \end{cases}$$

In each replication the first 100 observations were used for training and 20 one-step forecasts were produced on an expanding window. The proposed model used the ADF-selected representation with $k = 5$ and $\lambda = 1$ (no tuning, to keep the comparison conservative); the ARIMA benchmark used the same ADF decision and an AIC-selected order. Table 9 reports the mean of the ratio $\text{RMSE}_{\text{proposed}}/\text{RMSE}_{\text{benchmark}}$ and the proportion of replications in which this ratio is below 1.

Table 9 - Monte Carlo comparison (200 replications, 20 one-step forecasts each). Ratios below 1 favour the proposed model.

Data-generating process	vs Chen: mean ratio (share < 1)	vs ARIMA: mean ratio (share < 1)	vs random walk: mean ratio (share < 1)	Support coverage
AR(1), $\phi = 0.5$	0.950 (0.730)	1.010 (0.365)	0.901 (0.875)	0.676
AR(1), $\phi = 0.8$	0.931 (0.760)	1.037 (0.410)	1.021 (0.580)	0.752
Random walk + drift	0.406 (0.950)	1.027 (0.390)	1.016 (0.530)	0.628
SETAR (regime-switching)	0.941 (0.755)	0.933 (0.850)	0.834 (0.930)	0.730

Three findings stand out (Figure 6b). First, the proposed model beats Chen's model under every process, dramatically so for the trending random walk (mean ratio 0.406, better in 95% of replications), where Chen's level-based intervals cannot extrapolate beyond the training range. Second, when the data are generated by a linear autoregression, which is exactly the class assumed by the ARIMA benchmark, the proposed model loses only 1.0–3.7% in RMSE, i.e. it is nearly efficient. Third, under nonlinear regime-switching (SETAR) dynamics the proposed model is 6.7% more accurate than ARIMA on average and better in 85% of replications, because the fuzzy transition matrix captures state-dependent behaviour that a single linear autoregression cannot. The average support coverage of the fuzzy forecasts lies between 0.63 and 0.75.

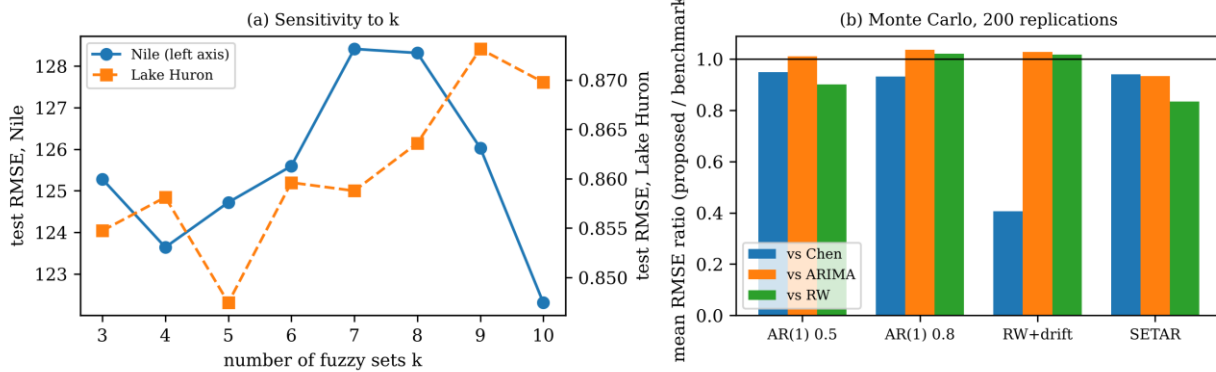


Fig. 6 - (a) Sensitivity of the test RMSE to the number of fuzzy sets k (Nile on the left axis, Lake Huron on the right); (b) Monte Carlo mean RMSE ratios of the proposed model to each benchmark.

Ablation and sensitivity analysis

Table 10 isolates the contribution of each component by changing one element at a time while keeping the selected (k, λ) .

Table 10 - Ablation study on the out-of-sample periods.

Variant	Nile RMSE	Nile MAE	Lake Huron RMSE	Lake Huron MAE
Proposed (FCM, soft, CV λ)	123.64	95.82	0.8547	0.7091
Equal-interval partition	123.26	95.03	0.8598	0.7135
Hard (max-membership) fuzzification	134.20	104.20	0.8466	0.6961
No recency weighting ($\lambda = 1$)	127.48	107.91	0.8547	0.7091
Opposite representation	150.97	127.15	0.9010	0.7190

Using the representation opposite to the ADF decision is the most damaging change, increasing RMSE by 22.1% for the Nile and 5.4% for Lake Huron; this supports the statistically guided first step (Proposition 4). Replacing soft memberships by maximum-membership fuzzification increases the Nile RMSE by 8.5% and MAE by 8.7%; for Lake Huron the hard version is marginally better (by about 1%), so the benefit of soft fuzzification is data-dependent but never large in the wrong direction. Removing recency discounting worsens the Nile forecasts by 3.1% in RMSE and 12.6% in MAE (for Lake Huron $\lambda = 1$ was already selected). With small k the FCM and equal-interval partitions give similar accuracy (differences below 1%), so the gain of FCM lies mainly in removing an arbitrary design choice rather than in raw accuracy at this granularity. Figure 6a shows that the test RMSE varies within about 5% across $k = 3, \dots, 10$ for both series, so the model is not fragile with respect to the number of fuzzy sets.

8. Discussion

The results support a balanced conclusion. Relative to the classical fuzzy time series models of Chen and Yu, the proposed model is consistently more accurate, both on real data and across all simulated processes. Relative to ARIMA, it matches the linear benchmark when the data are linear and improves on it under nonlinear regime dynamics. This combination is valuable in practice because the model also provides what ARIMA does not: a small, readable matrix of linguistic transitions (“after a very high flood year the next year is most likely below average”) and a graded fuzzy forecast.

The statistical evaluation also shows why formal inference should accompany fuzzy forecasting studies. Several gains that look substantial in point measures (for example a 49% MSE reduction on the enrolment data) are only marginally significant with 20 forecasts, while some smaller gains (against Holt’s method for Lake Huron) are significant because they are consistent across years. Reporting DM-HLN, Wilcoxon and block-bootstrap results together, as done here, gives a more reliable picture than ranking by MSE alone and guards against over-interpreting differences that may be due to chance.

Several limitations should be noted. The study concerns one-step-ahead forecasts of annual univariate series; multi-step forecasts can be obtained by iterating πP^h but were not evaluated. The fuzzy forecast is a possibility distribution, not a probability distribution, and its coverage is not calibrated to a nominal level; conformal or quantile calibration of the support is a natural extension. The model is first-order, and higher-order or multivariate fuzzy relationships may further improve accuracy for series with longer memory. Finally, the ADF test has low power in very short series, so in such cases the representation choice should also be informed by subject-matter knowledge, as was done for the enrolment data.

9. Conclusion

This paper has proposed and validated a membership-weighted fuzzy time series model that unites fuzzy and statistical reasoning at every stage: an ADF test chooses the state representation, fuzzy c-means defines a Ruspini partition of triangular fuzzy sets, soft memberships form a recency-discounted fuzzy transition matrix with proven stochastic and consistency properties, and forecasts are delivered as triangular fuzzy numbers. On the Alabama enrolment benchmark the model halves the MSE of Chen’s model. In genuine out-of-sample forecasting of the Nile flow and Lake Huron level it achieves the lowest MAE, MAPE and MASE among six methods, equals ARIMA in RMSE, and is significantly better than the random walk and Holt’s method by bootstrap and Wilcoxon inference. Monte Carlo evidence confirms near-efficiency under linear dynamics and a clear advantage under regime switching. Future work will extend the framework to higher-order and multivariate fuzzy relationships, calibrated fuzzy prediction intervals and multi-step forecasting of agricultural and environmental series.

Declarations

Funding. The author received no specific funding for this work.

Conflict of interest. The author declares no conflict of interest.

Data availability. All three datasets are publicly available: the Alabama enrolments are tabulated in [4] and [6], and the Nile and Lake Huron series are distributed with the R datasets package (R Core Team). The values used are reproduced in Table 2 and Figure 2, and the analysis code is available from the author on reasonable request.

Author contribution. Maheswari D: conceptualization, methodology, software, formal analysis, writing – original draft, writing – review and editing.

References

- [1] Box GEP, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken: Wiley; 2015.
- [2] Zadeh LA. Fuzzy sets. *Information and Control*. 1965;8(3):338–353.
- [3] Song Q, Chissom BS. Fuzzy time series and its models. *Fuzzy Sets and Systems*. 1993;54(3):269–277.
- [4] Song Q, Chissom BS. Forecasting enrollments with fuzzy time series—Part I. *Fuzzy Sets and Systems*. 1993;54(1):1–9.
- [5] Song Q, Chissom BS. Forecasting enrollments with fuzzy time series—Part II. *Fuzzy Sets and Systems*. 1994;62(1):1–8.
- [6] Chen SM. Forecasting enrollments based on fuzzy time series. *Fuzzy Sets and Systems*. 1996;81(3):311–319.
- [7] Yu HK. Weighted fuzzy time series models for TAIEX forecasting. *Physica A: Statistical Mechanics and its Applications*. 2005;349(3–4):609–624.
- [8] Huarng K. Effective lengths of intervals to improve forecasting in fuzzy time series. *Fuzzy Sets and Systems*. 2001;123(3):387–394.
- [9] Huarng K, Yu THK. Ratio-based lengths of intervals to improve fuzzy time series forecasting. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*. 2006;36(2):328–340.
- [10] Stevenson M, Porter JE. Fuzzy time series forecasting using percentage change as the universe of discourse. *World Academy of Science, Engineering and Technology*. 2009;55:154–157.
- [11] Bezdek JC. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press; 1981.

-
- [12] Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*. 1984;10(2–3):191–203.
- [13] Li ST, Cheng YC, Lin SY. A FCM-based deterministic forecasting model for fuzzy time series. *Computers & Mathematics with Applications*. 2008;56(12):3052–3063.
- [14] Cheng CH, Chen TL, Teoh HJ, Chiang CH. Fuzzy time-series based on adaptive expectation model for TAIEX forecasting. *Expert Systems with Applications*. 2008;34(2):1126–1132.
- [15] Egrioglu E, Aladag CH, Yolcu U, Uslu VR, Erilli NA. Fuzzy time series forecasting method based on Gustafson–Kessel fuzzy clustering. *Expert Systems with Applications*. 2011;38(8):10355–10357.
- [16] Bose M, Mali K. Designing fuzzy time series forecasting models: A survey. *International Journal of Approximate Reasoning*. 2019;111:78–99.
- [17] Sullivan J, Woodall WH. A comparison of fuzzy forecasting and Markov modeling. *Fuzzy Sets and Systems*. 1994;64(3):279–293.
- [18] Dickey DA, Fuller WA. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*. 1979;74(366):427–431.
- [19] MacKinnon JG. Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*. 1996;11(6):601–618.
- [20] Ruspini EH. A new approach to clustering. *Information and Control*. 1969;15(1):22–32.
- [21] Tsaor RC. A fuzzy time series-Markov chain model with an application to forecast the exchange rate between the Taiwan and US dollar. *International Journal of Innovative Computing, Information and Control*. 2012;8(7B):4931–4942.
- [22] Diebold FX, Mariano RS. Comparing predictive accuracy. *Journal of Business & Economic Statistics*. 1995;13(3):253–263.
- [23] Harvey D, Leybourne S, Newbold P. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*. 1997;13(2):281–291.
- [24] Künsch HR. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*. 1989;17(3):1217–1241.
- [25] Ljung GM, Box GEP. On a measure of lack of fit in time series models. *Biometrika*. 1978;65(2):297–303.
- [26] Xie XL, Beni G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1991;13(8):841–847.
- [27] Cobb GW. The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika*. 1978;65(2):243–251.
- [28] Durbin J, Koopman SJ. *Time Series Analysis by State Space Methods*, 2nd ed. Oxford: Oxford University Press; 2012.
- [29] Brockwell PJ, Davis RA. *Time Series: Theory and Methods*, 2nd ed. New York: Springer; 1991.
- [30] Holt CC. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*. 2004;20(1):5–10.
- [31] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *International Journal of Forecasting*. 2006;22(4):679–688.
- [32] Tong H. *Non-linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press; 1990.