

What actually limits an IT2FCM-Markov chain fuzzy time series model? Type reduction, directional adjustment, and the forecast horizon

Ali Bardadi¹, Budi Warsito^{1,2*}, Bayu Surarso^{1,3}, Wibowo Harry Sugiharto⁴

¹*Doctoral Program of Information Systems, Postgraduate School, Universitas Diponegoro, Semarang, Indonesia*

²*Department of Statistics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia*

³*Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia*

⁴*Department of Informatics Engineering, Faculty of Engineering, Universitas Muria Kudus, Kudus, Indonesia*

Abstract Hybrid fuzzy time series models that combine Interval Type-2 Fuzzy C-Means (IT2FCM) partitioning with Markov chain transition weighting lag behind the series at abrupt spikes. Two explanations are available in the literature and have never been separated: the approximate type reduction used inside the clustering loop contracts the footprint of uncertainty, and the probability-weighted forecast is a convex combination of the cluster centres and so cannot leave their span. This study separates them by factorial ablation on a published architecture [10], under an expanding-window protocol, on two ambient carbon monoxide series from one station in Semarang, Indonesia: 48 readings spanning [132, 946] and 288 half-hourly readings spanning [2063, 26934], the second placing 13.7 observations in each cluster against 3.0 in the first. Neither mechanism survives scrutiny. Replacing exact Karnik-Mendel type reduction by the Nie-Tan surrogate changes RMSE by 0.07% on the short series and 1.0% on the long one, neither significant, while the exact iteration costs three to five times as much per fit. The directional term lowers RMSE by 17% on the short series and raises it on the long one, where the dynamics are more strongly mean-reverting. A third finding subsumes both. At one step ahead the full pipeline is indistinguishable from persistence on both series ($p = 0.4137$ and 0.4551), so the ablated configurations are collectively no better than copying the last observation forward. Extending the chain to P^h shows where the model does earn its place: from two to twelve hours ahead it improves on persistence by 6.7-10.2% in RMSE ($p \leq 0.0015$), and at that horizon the differences between all six configurations fall below significance. A corollary finding is that an equal-width partition without any interval type-2 machinery attains the lowest overall RMSE on the larger series at approximately 1/400 of the computational cost of the full pipeline, suggesting that the sophistication of the clustering stage does not determine forecasting accuracy in this class of model. For half-hourly monitoring, the choice of type-reduction operator is not what limits these models; the choice of forecast horizon is.

Keywords Fuzzy time series, Interval type-2 fuzzy C-means, Karnik-Mendel type reduction, Markov chain, Forecast horizon, Naive baseline, Air quality

AMS 2010 subject classifications 62M10, 03E72

DOI: 10.19139/soic-2310-5070-4411

1. Introduction

Carbon monoxide is the most abundant traffic-derived pollutant in a dense urban corridor, and it is dangerous because it is undetectable without instruments. It binds to haemoglobin in preference to oxygen, and the resulting carboxyhaemoglobin impairs oxygen delivery to the heart and brain; the World Health Organization sets its guideline levels so that carboxyhaemoglobin does not exceed 2%, a threshold derived from reduced exercise tolerance and earlier onset of angina in people with cardiovascular disease [1]. The 2021 revision of the global

*Correspondence to: Budi Warsito (Email: budiwarsito@live.undip.ac.id). Department of Statistics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang 50275, Indonesia.

air quality guidelines added a 24-hour guideline of 4 mg m^{-3} alongside the existing short-averaging limits, reflecting evidence that harm occurs at concentrations previously considered acceptable [1]. Because the symptoms of mild exposure-headache, fatigue, impaired concentration-resemble ordinary complaints, the population at risk is generally unaware of it, and management depends entirely on measurement and, where possible, on anticipation.

Anticipation is difficult because of how the pollutant behaves in time. Street-level CO is driven by incomplete combustion in vehicle engines, so its concentration tracks traffic density and inherits its structure: a bimodal diurnal cycle with morning and evening commuting peaks, modulated by the growth and collapse of the planetary boundary layer, which dilutes the morning peak as the day warms and traps the evening one as the layer collapses [2]. Superimposed on that cycle are excursions of a different character. Dense sensor deployments show that the interquartile range of hourly CO near an arterial road roughly triples during rush hour relative to midday [3], so a monitoring series is not a smooth periodic signal with noise but a regime-switching one: long stretches of low, slowly varying concentration punctuated by sharp surges whose timing depends on traffic incidents and meteorology rather than on the clock alone. These are the moments that matter for a health advisory, and they are also the moments that a forecasting model is most likely to miss.

The measurement side of this problem has largely been solved. Wireless sensor networks now report street-level CO continuously at intervals of seconds to minutes, at a spatial resolution that fixed regulatory stations cannot approach, and are explicitly proposed as the basis for early warning of pollution episodes [3]. What such a network does not supply is lead time: it reports what the concentration is, not what it will be. Earlier implementations of CO monitoring systems in this region, including web-based real-time displays driven by sensor acquisition, demonstrate that continuous measurement infrastructure is already in place; what remains missing is the forecasting layer that converts current readings into actionable warnings [39]. That gap is where forecasting enters, and it also sets a condition on what a useful forecast must look like, to which we return below.

Among the methods available for a short, irregular, regime-switching series, the Fuzzy Time Series (FTS) framework of Song and Chissom [4, 5], simplified into an arithmetic rule-based form by Chen [6], is attractive because it assumes neither a distribution nor stationarity nor a large sample conditions that ARIMA and its seasonal variants require and that monitoring data rarely satisfy [7, 8]. FTS represents observations as fuzzy sets over a partitioned universe of discourse and forecasts by rules learned from the transitions between them. Its accuracy has been pursued along two lines: making the partition data-driven rather than heuristic, which led through fuzzy c -means to interval type-2 clustering, and resolving the ambiguity of a state with several observed successors, which led to Markov weighting of the transition rule. Section 2 traces both. Our own earlier work joined them, applying interval type-2 clustering to the ambient CO series used here [9] and then combining it with a first-order chain [10].

That combined architecture reproduces the stable regime of a CO series well and responds to a surge only after the peak has passed. The lag is not a detail: it is a failure precisely at the excursions described above, the ones an advisory would be issued for. Two explanations for it are available, and they have never been separated.

The first blames type reduction. Every interval type-2 clustering iteration must collapse the interval memberships into a crisp cluster centre, and the exact solution the iterative Karnik-Mendel algorithm [11] is widely judged too costly to run inside a clustering loop, so closed-form surrogates are substituted in practice. All of them contract the footprint of uncertainty towards its centre. The argument that follows is that the boundary clusters, the ones framing extreme values, are pulled inwards, leaving the outermost intervals too narrow to accommodate a spike when it arrives.

The second explanation is structural and has attracted far less attention. The Markov-weighted forecast $\sum_j P_{s_t j} v_j^*$ is a convex combination of the cluster centres. It is therefore confined to the closed interval between the smallest and the largest centre, whatever the transition probabilities happen to be. A model so constructed cannot produce a forecast beyond the span of its own centres, and at a spike the relevant target frequently lies outside that span. This bound holds regardless of how the centres were obtained, and so is entirely independent of the type-reduction question.

Both explanations predict the same symptom, which is why attributing the lag to type reduction has gone untested. This paper tests it, taking the published architecture of [10] as the baseline and varying two factors independently: the type-reduction operator inside the clustering loop, and the presence of a directional adjustment term that releases the forecast from the convex hull of the centres. A third question follows from the monitoring

context itself and shapes the evaluation throughout. If a station already reports every thirty minutes, a one-step forecast is superseded by observation before any decision can act on it; the horizon at which a forecast earns its place is therefore an empirical question, and one this literature does not usually ask. Answering it also resolves a prior question: whether configurations that differ by a few per cent in one-step RMSE differ at all on a task where a model is actually useful. The contributions are:

- *A factorial ablation that separates the two candidate mechanisms.* Exact KM type reduction and the directional adjustment are switched on and off independently, so that the contribution of each is measured rather than assumed.
- *A direct test of the type-reduction hypothesis.* The exact KM iteration is compared against the Nie-Tan surrogate with every other component held fixed, which is the comparison the contraction argument implies but which the literature does not report.
- *A leak-free evaluation protocol.* All models are assessed under an expanding window in which the clustering, the interval boundaries and the transition matrix are re-estimated from the training window at every step, so that no quantity derived from a test point enters the fit that predicts it.
- *Responsiveness measured where it is claimed.* Absolute error is reported separately at spike points and away from them, since a lag at surges is the behaviour under investigation.
- *A naive baseline and a horizon sweep.* Every result is placed against persistence, the baseline this literature routinely omits, and the chain is extended to P^h so that accuracy is reported as a function of lead time. This is what identifies the horizon band in which the model is worth running.
- *Replication on a second series, and a cost account.* Every configuration is re-run on an independent 288-point series from the same station whose partition carries four times as many observations per cluster, which is the regime in which the contraction argument should bite hardest. Wall-clock cost per partitioning fit is reported alongside accuracy, since an operator that buys nothing must still be paid for.
- *A counter-intuitive partitioning result with cost implications.* Equal-width partitioning, without any interval type-2 machinery, attains the lowest spike error on Series A and the lowest overall RMSE on Series B, significantly so, at roughly 1/400 of the computational cost of the full pipeline. This finding challenges the premise that clustering sophistication governs forecasting accuracy in this class of model and is reported as a finding in its own right, not merely as an ablation control.

The result is negative on the hypothesis it set out to test and, on the second series, negative on the correction we introduced to replace it. The horizon question turns out to matter more than either. Section 2 positions the study against existing work. Section 3 specifies the model and the factorial design. Sections 4 and 5 present the protocol, results and discussion. Section 6 concludes.

2. Related works

Fuzzy time series methods have accumulated a substantial applied literature over three decades, mapped bibliometrically by Palomero et al. [12], with air quality among the recurring application domains-Folaponmile et al. [13] forecast urban pollution in Kaduna with a Chen-type model, and Pribadi et al. [14] apply the Cheng variant to a volatile financial series with the same motivation of avoiding distributional assumptions. Two largely independent lines of work feed the architecture examined here. The clustering line begins with type-1 FCM applied to interval construction, where the partition becomes objective but still rests on a fuzzifier fixed by trial; Chen et al. [15] pair it with justifiable granularity and particle swarm search, Gong and Feng [16] with an improved membership update, and Satriani et al. [17] with a heuristic hybrid. Hwang and Rhee [18] generalised the algorithm to interval type-2, replacing the point fuzzifier with a pair and giving each observation an interval of memberships whose width is the footprint of uncertainty [19]; Dang et al. [20] extended it to collaborative clustering, and Yin et al. [21] carried the construction into an FTS pipeline and supplied the fuzzifier interval adopted in Section 3. The transition line is separate. Odah [22] applies a chain directly to a rate series; Alyousifi et al. [23] weight FLRG consequents by empirical transition probability and [24] extend the scheme to PM_{10} ; Alhendi et al. [25] embed

a chain in a neural pipeline. Xian and Cheng [26] join the two lines in a Pythagorean fuzzy setting, which is the closest published analogue to the model dissected below.

Type reduction itself has been treated almost exclusively as a computational problem. Karnik and Mendel [11] gave the exact iterative solution together with the switch-point characterisation on which Eq. (3) rests. Wu and Mendel [27] cut its iteration count without moving its fixed point. Nie and Tan [28] took a different route, weighting each observation by the arithmetic mean of its upper and lower memberships in a single pass; that quantity is the initialisation of the KM iteration rather than its fixed point, so the centroid interval it returns is systematically narrower than the true one. Rhee [29] and Rubio and Castillo [30] review the practical consequences of these choices, and He et al. [31] embed the exact iteration inside an adaptive interval-valued clustering scheme for land cover classification, where its cost is tolerated because the objective is accuracy of assignment.

What none of this work reports is whether the contraction matters downstream. The operators are compared on centroid accuracy and on runtime, not on the forecasting error of a pipeline that consumes their output, and the inference that a narrower centroid interval degrades that error plausible, and repeated often has to our knowledge never been measured against a controlled alternative. Supplying that measurement is the first factor of the design in Section 3.2. Taken together, the clustering and transition lines share a structural gap: components are validated in isolation but the evidence for their joint contribution to forecasting accuracy is thin, non-comparable across studies due to the absence of a common evaluation protocol, and confined to $h = 1$ accuracy against no naive baseline.

The second explanation has no comparable literature, which is itself informative. Weighting schemes proliferate probabilistic [23], granularity-based [15, 16], possibilistic [30] but every one of them redistributes influence among the cluster centres and therefore leaves the forecast inside their convex hull. Xian and Cheng [26] observe the resulting ceiling in a Pythagorean setting without acting on it. Corrections that do escape the hull exist in the FTS literature as trend or adjustment terms, but they are introduced as accuracy improvements rather than tested against the mechanism they are supposed to fix. Isolating that mechanism is the second factor of the design.

The availability of modern alternatives ARIMA and its seasonal variants, gradient-boosted trees, and transformer-based forecasters does not invalidate the FTS approach for this setting; it contextualises it. The series studied here are short (48 and 288 observations), sampled at irregular or short fixed intervals, and structurally non-stationary in a way that distributional assumptions do not accommodate. FTS models are selected not because they are expected to outperform deep learning on a large benchmark but because they operate without those assumptions and produce transition rules whose structure can be examined by a domain practitioner. Recent probabilistic and transformer-based forecasting advances, while powerful in data-rich regimes, have not been evaluated under the constraints of a street-level air quality sensor network and were not designed for them; recent work in this journal has demonstrated both the complexity of short-term CO monitoring [40] and the difficulty of extracting gains from increased architectural complexity in multivariate air quality settings [41]. The open question this paper addresses whether the components of the FTS pipeline contribute independently to accuracy, and at which horizon the pipeline adds value over doing nothing is a prior question to any comparison with alternative methods.

Table 1 situates the present study. The prior architecture [10] appears as a separate row, since this work is a controlled dissection of it rather than an alternative to it.

2.1. Relationship to prior work

This paper is a direct continuation of two of our own studies and shares their data. In [9] IT2FCM was applied to the same ambient CO series to obtain a data-driven partition, without a probabilistic transition stage. In [10] the partition and a first-order Markov chain were combined into the IT2FCM-MC-FTS architecture; in that implementation the fuzzifier interval was collapsed to its centroid ($\bar{m} = 1.95$) before clustering, so no upper and lower memberships were formed, no footprint of uncertainty was constructed, and no type reduction was performed.

Three things distinguish the present study. It restores the interval fuzzifier and performs type reduction, exactly and approximately, as an experimental factor rather than an implementation detail. It adds the directional adjustment term of Eq. (9), which is absent from [10]. And it evaluates every configuration including a re-run of [10] under identical conditions with an expanding-window protocol, so the numbers reported here are not comparable

Table 1. Comparative summary of existing methods and the present study

Method	Components	Limitations
Type-1 Fuzzy (T1FM) [32–34]	Markov Chain	Type-1 fuzzy sets over a discrete-time chain
Interval Type-2 Chain (IT2FM) [32–34]	Fuzzy Markov	Interval type-2 fuzzy sets; type reduction to recover a crisp state
Adaptive IT2FCM [31]	(A-IT2FCM)	Interval type-2 clustering with Karnik-Mendel reduction
General Type-2 FCM [35]	(GT2 FCM)	α -planes and zSlices over a linguistic fuzzifier
Hybrid FTS-MC with C-Means [36]	Type-1 C-Means	Partition feeding a Markov transition rule
Credal C-Means (CCM) clustering [37]	IT2FCM-MC-FTS [10]	Credal partition with meta-clusters for ambiguous objects
		Fuzzifier interval collapsed to its centroid, Markov chain, volatility heuristic
This study		Factorial ablation of type reduction and directional adjustment under a leak-free protocol
		Single-valued memberships absorb neither measurement noise nor misspecification of the fuzzifier
		Type reduction dominates the cost, and the choice of reduction operator is left open
		KM is applied to a classification objective, not to a forecasting partition; cost scales with sample size
		Every plane requires its own reduction, multiplying an already heavy computation
		Partition depends on an arbitrary initialisation, and the intervals need not track data density
		Assignment stays unstable for points lying between clusters
		No FOU is formed and no type reduction is performed; forecast confined to the convex hull of the centres
		Univariate; a single 48-point series; the two factors are tested only in combination with a first-order chain

with those in [10] and supersede them for this series. We reserve the label IT2FCM-MC-FTS for the architecture published in [10] and denote the full model examined here as KM-IT2FCM-MC-FTS.

3. Model and factorial design

3.1. Pipeline

The pipeline runs in four stages: an interval type-2 clustering step that fixes the partition, a fuzzification step that converts observations into a state sequence, a counting step that estimates the transition matrix, and an inference step that issues the one-step forecast. Figure 1 shows how they connect. Each stage is specified below at the level of detail required to reimplement it, with every constant taken from the code that produced the results of Section 5.

3.1.1. Stage 1: partitioning IT2FCM generalises FCM by replacing the single fuzzifier m with an interval $[m_1, m_2]$, so that every observation receives an interval of membership; the width of that interval is the footprint of uncertainty. Following Yin et al. [21] we set $m_1 = 1.8$ and $m_2 = 2.2$. For an observation y_k and a cluster centre v_i , let $d_{ik} = \|y_k - v_i\|$. The membership evaluated at a fuzzifier m is

$$u_{ik}(m) = \left[\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}} \right]^{-1}. \quad (1)$$

Evaluating Eq. (1) at both fuzzifiers gives the bounds of the interval membership:

$$\bar{u}_{ik} = \max \{u_{ik}(m_1), u_{ik}(m_2)\}, \quad \underline{u}_{ik} = \min \{u_{ik}(m_1), u_{ik}(m_2)\}. \quad (2)$$

Equation (2) is the definition; the condition-based formulation of Hwang and Rhee [18] is an equivalent closed-form rule for deciding which fuzzifier yields the larger value.

Because the memberships are intervals, each cluster centre is an interval and must be reduced to a crisp value. With the observations sorted ascending, the exact endpoints satisfy

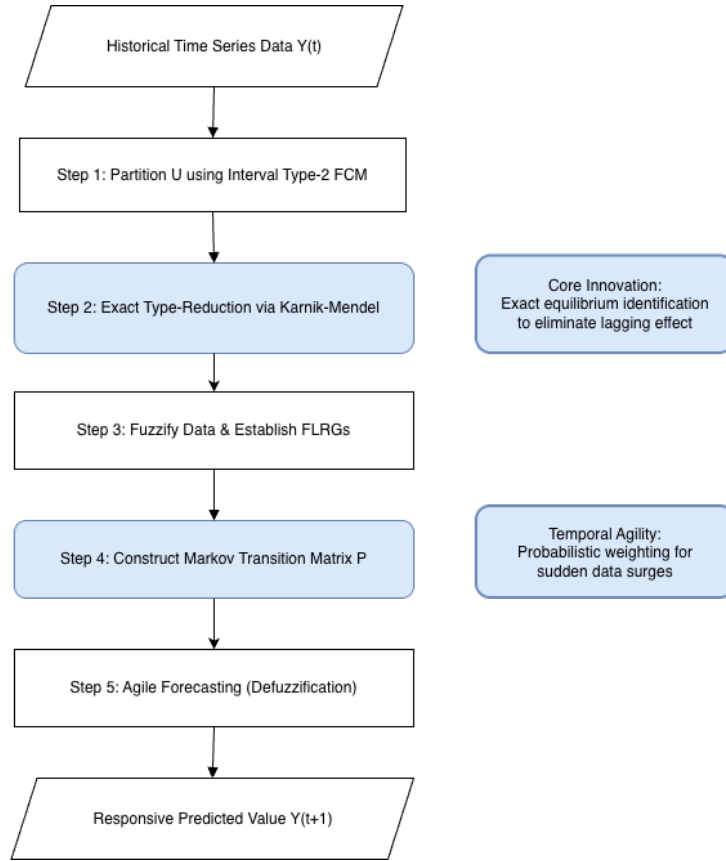


Figure 1. Structure of the pipeline. The two stages varied in the ablation are the type reduction inside Stage 1 and the directional term in Stage 4.

$$v_i^L = \frac{\sum_{k=1}^L y_k \bar{u}_{ik} + \sum_{k=L+1}^n y_k \underline{u}_{ik}}{\sum_{k=1}^L \bar{u}_{ik} + \sum_{k=L+1}^n \underline{u}_{ik}}, \quad v_i^R = \frac{\sum_{k=1}^R y_k \underline{u}_{ik} + \sum_{k=R+1}^n y_k \bar{u}_{ik}}{\sum_{k=1}^R \underline{u}_{ik} + \sum_{k=R+1}^n \bar{u}_{ik}}, \quad (3)$$

where L and R are switch points defined by $y_L \leq v_i^L \leq y_{L+1}$ and $y_R \leq v_i^R \leq y_{R+1}$. The KM procedure locates them by iteration rather than assuming them. The crisp centre is the midpoint of the exact interval,

$$v_i^* = \frac{v_i^L + v_i^R}{2}. \quad (4)$$

Factor A type reduction. Three settings are compared. *Exact KM* iterates Eq. (3) to its fixed point with an inner tolerance of 10^{-6} . *Nie-Tan* [28] takes $v_i^L = v_i^R = \sum_k y_k \theta_k / \sum_k \theta_k$ with $\theta_k = (\bar{u}_{ik} + \underline{u}_{ik})/2$, that is, the initialisation of the KM iteration and not its fixed point. *Type-1* collapses the fuzzifier interval to its centroid $\bar{m} = 1.95$ before clustering and runs ordinary FCM, reproducing the partitioning stage of [10].

Equations (1)-(4) are iterated until the largest change in any centre falls below $\varepsilon = 0.005$, subject to a ceiling of 1000 iterations. Cluster initialisation uses a fixed seed (1852 for the interval variants, 7521 for the type-1 replication of [10]).

The number of clusters is derived from the volatility of the training window rather than fixed a priori:

$$vol = \frac{1}{n-1} \sum_{t=2}^n |y_t - y_{t-1}|, \quad c_{\text{raw}} = \text{round}\left(\frac{\max Y - \min Y}{vol}\right), \quad c = \max\left\{2, \min(c_{\text{raw}}, \lfloor n/2 \rfloor)\right\}. \quad (5)$$

The lower bound guarantees at least two states; the upper bound guarantees, on average, at least two observations per cluster, the minimum for the transition frequencies of Stage 3 to be estimable.

3.1.2. Stage 2: fuzzification Clusters are sorted by their equilibrium points and boundaries are placed midway between adjacent ones, so the partition is dense where the data are dense:

$$B_i = \frac{v_i^* + v_{i+1}^*}{2}, \quad i = 1, \dots, c-1, \quad B_0 = \min Y, \quad B_c = \max Y. \quad (6)$$

Each Y_t is mapped to the fuzzy set A_i whose interval $(B_{i-1}, B_i]$ contains it, producing the state sequence $S = \{s_1, \dots, s_n\}$. Consecutive pairs form Fuzzy Logical Relationships $A_{t-1} \rightarrow A_t$, grouped by antecedent into FLRGs.

3.1.3. Stage 3: transition weighting Let N_{ij} count the transitions $A_i \rightarrow A_j$ in the training window. The first-order transition matrix is the row-wise normalisation

$$P_{ij} = \frac{N_{ij}}{\sum_{k=1}^c N_{ik}}. \quad (7)$$

If state i is never followed by any state in the training window which occurs for extreme states early in a rolling forecast the row is undefined and we set $P_{ii} = 1$, so the model defaults to persistence rather than failing.

3.1.4. Stage 4: forecast and directional adjustment Let s_t be the state of the most recent observation. The forecast is

$$\hat{Y}_{t+1} = \sum_{j=1}^c P_{s_t j} v_j^* + D_t. \quad (8)$$

The first term is a convex combination of the cluster centres and therefore lies in $[\min_j v_j^*, \max_j v_j^*]$ for every possible probability vector. On the present series the final training window gives $[141.55, 930.41]$ against an observed maximum of 946: the first term alone cannot reach the extremes of the data, and more importantly cannot overshoot the current centre in the direction of a developing surge. The second term releases it:

$$D_t = \frac{l_{s_t}}{2} \text{sgn}(Y_t - Y_{t-1}), \quad l_{s_t} = B_{s_t} - B_{s_t-1}. \quad (9)$$

The magnitude is half the width of the interval currently occupied and the sign follows the most recent observed change, so the correction is proportional to the local granularity of the partition and directed by the local trend.

Factor B directional adjustment. D_t is either included as in Eq. (8) or set to zero.

3.2. Factorial design

Table 2 lists the configurations. Rows 1-4 vary the two factors of interest with everything else fixed; rows 5-6 are the conventional single-component ablations, retained for comparability with the FTS literature.

Table 2. Configurations compared. Every row uses the same volatility heuristic, the same boundary rule and the same expanding-window protocol

Configuration	Type reduction	Transition rule	D_t
IT2FCM-MC-FTS [10]	type-1 ($\bar{m} = 1.95$)	Markov	no
+ D_t only	type-1 ($\bar{m} = 1.95$)	Markov	yes
+ exact KM only	exact KM	Markov	no
Nie-Tan variant	Nie-Tan	Markov	yes
FTS-IT2FCM	exact KM	simple averaging	yes
FTS-MC	equal-width partition	Markov	yes
KM-IT2FCM-MC-FTS	exact KM	Markov	yes

4. Experimental setup

4.1. Data

Two series of ambient carbon monoxide concentration from the Mijen air quality monitoring station in Semarang City, Indonesia, are used. Series A is the 48-point record on which [9, 10] were based:

827, 907, 946, 937, 945, 885, 766, 625, 539, 487, 411, 385, 385, 570, 770, 774, 690, 612, 525, 409, 292, 235, 209, 177, 146, 141, 133, 132, 137, 143, 163, 198, 218, 181, 192, 170, 140, 138, 133, 144, 186, 256, 452, 437, 432, 459, 520, 564.

Series B is an independent record of 288 half-hourly readings covering six consecutive days (1-6 October 2020), concatenated chronologically. Table 3 contrasts them. The two differ in every respect that the contraction argument of Section 2 implicates. Series B spans a range thirty times wider, is sampled sixteen times more finely, and because the volatility heuristic returns 21 states for 288 observations against 16 for 48 places 13.7 observations in each cluster where Series A places 3.0. If approximate type reduction degrades a forecast by displacing the equilibrium points of sparsely populated boundary clusters, Series B is where the exact iteration should earn its cost.

Table 3. The two validation series

	Series A	Series B
Observations	48	288
Sampling interval	irregular	30 minutes
Universe of discourse	[132, 946]	[2063, 26934]
Range	814	24871
Mean absolute first difference	50.87	1184.08
States returned by Eq. (5)	16	21
Observations per cluster	3.0	13.7
Transitions exceeding the spike threshold	25.5%	15.7%
Forecasts evaluated (default window)	38	228

4.2. Protocol

Every forecast is one step ahead and out of sample. The window expands rather than slides: at step t the clustering, the interval boundaries and the transition matrix are all re-estimated from $Y_1, \dots, Y_{t-1}$ alone, and the resulting $\hat{Y}_t$ is scored against the withheld Y_t . No quantity derived from the future of the series enters the fit at any point. The initial training window is 10 observations on Series A and 60 on Series B, giving 38 and 228 evaluated forecasts; Section 5.7 reports the sensitivity of every conclusion to that choice on both series.

This protocol matters here. Fitting the partition and the transition matrix once on the complete series and then scoring in sample inflates apparent accuracy substantially and is not a forecast; results obtained that way are not comparable with those below. The multi-step forecasts at horizon h , reported in Section 5.8, are obtained by raising the transition matrix estimated at step t to the power P^h ; no future observations enter this computation, and the expanding-window constraint is satisfied at every horizon.

4.3. Metrics

Two error measures are reported. RMSE is scale-bound and penalises the large one-step residuals that arise at spikes, which is where the responsiveness question lies. SMAPE is scale-relative and exposes behaviour in the low-concentration regime, which RMSE dominated by the high end of a series spanning 132 to 946 would conceal. For the pairwise comparisons the Shapiro-Wilk test is applied first to each absolute error series; normality fails for all of them, so the comparisons rest on the Wilcoxon signed-rank test at $\alpha = 0.05$, under the null of no difference between the paired absolute errors. Because six pairwise comparisons are made against the proposed model on each series, Holm-Bonferroni correction is applied to control the familywise error rate. The correction does not overturn any raw conclusion: on Series A all six adjusted p -values exceed 0.41; on Series B the adjusted p -values for FTS-IT2FCM (0.0025) and FTS-MC (0.0404) remain significant, while the remaining four comparisons including the published baseline and the Nie-Tan variant do not fall below α after correction (adjusted $p \geq 0.63$). Bootstrap 95% confidence intervals for RMSE (5000 replicates) are reported alongside each point estimate in Tables 4 and ?? . Two model-free baselines are carried throughout Section 5.8: persistence, $\hat{Y}_{t+h} = Y_t$, and a seasonal naive rule that repeats the value observed at the same time of day on the previous day.

Responsiveness is measured directly. A spike point is a step at which $|Y_t - Y_{t-1}|$ exceeds 1.5 times the mean absolute first difference of the series in question: 76.31 on Series A, where eight of the 38 evaluated steps qualify, and 1776.11 on Series B, where 37 of 228 qualify. Mean absolute error is reported separately over those steps and over the remainder. The threshold factor of 1.5 is a design choice. Running the same analysis at factors of 1.0 (thresholds 50.9 and 1184.1) and 2.0 (thresholds 101.7 and 2368.2) gives the same qualitative ordering: spike error exceeds non-spike error by a factor of five to six across all configurations at every threshold, and the ranking of configurations at spikes is stable. The conclusions on spike responsiveness do not depend on the specific choice of 1.5.

5. Results and discussion

Three findings organise what follows. First, type reduction does not determine forecasting accuracy: exact Karnik-Mendel and the Nie-Tan surrogate are statistically indistinguishable on both series while differing by a factor of three to five in runtime, and the null holds across all fixed cluster counts tested ($c \in \{5, 10, 21\}$). Second, the directional term improves accuracy on the short, spiky series but degrades it on the long, mean-reverting one, which is the signature of a correction committed to the sign of the last observed change in a series that frequently reverses. Third, and most consequentially, the entire pipeline is indistinguishable from persistence at one step ahead, but improves significantly on persistence from two to twelve hours ahead. At that useful horizon, none of the six ablated configurations differs from another. An equal-width partition with no type-2 machinery achieves the lowest overall RMSE on Series B at roughly 1/400 of the computational cost.

5.1. Ablation on Series A

Table 4 reports every configuration on Series A under the common protocol.

Table 4. Series A: forecasting error and comparison against the full model (38 one-step forecasts). Wilcoxon statistics compare the absolute error series of each row against KM-IT2FCM-MC-FTS

Configuration	RMSE	SMAPE (%)	Wilcoxon W	p
IT2FCM-MC-FTS [10]	94.53	18.22	202.0	0.0136
+ D_t only	78.44	14.29	331.0	0.5757
+ exact KM only	89.52	19.85	158.0	0.0035
Nie-Tan variant	74.51	15.07	294.0	0.2738
FTS-IT2FCM (simple averaging)	76.09	15.07	10.0	0.9165
FTS-MC (equal-width partition)	80.55	15.89	341.0	0.6775
KM-IT2FCM-MC-FTS	74.46	15.23		

Three readings follow directly.

The full model improves on the published baseline, and the improvement is significant. RMSE falls from 94.53 to 74.46 and SMAPE from 18.22% to 15.23%, with $p = 0.0136$. This is the one comparison in the table that supports a claim of progress over [10] on this series. Section 5.5 shows that it does not survive replication.

The improvement is carried by the directional term, not by type reduction. Adding D_t alone to the published baseline recovers 78.44 of the 74.46 attained by the full model, and the difference between that configuration and the full model is not significant ($p = 0.5757$). Adding exact KM alone recovers only 89.52, and that configuration is significantly worse than the full model ($p = 0.0035$). Decomposing the total reduction of 20.07 in RMSE, the directional term accounts for 16.09 on its own and exact type reduction for 5.01, with the two overlapping rather than adding.

Exact type reduction is not distinguishable from its closed-form surrogate. With every other component held fixed, replacing the KM iteration by the Nie-Tan operator changes RMSE from 74.46 to 74.51 and leaves SMAPE marginally lower, at $p = 0.2738$. The contraction of the footprint of uncertainty that Nie-Tan is known to produce is real, but on this series it does not propagate into a measurable loss of forecasting accuracy. That is the central negative result of this study.

Two conventional ablations complete the picture. Replacing the IT2FCM partition with equal-width intervals raises RMSE to 80.55, a difference that is not significant ($p = 0.6775$). Replacing Markov weighting with simple averaging gives 76.09, also not significant but this comparison rests on only six non-tied pairs out of 38, because the two rules coincide wherever an FLRG has a single consequent, which is the majority of rows on a series this short. Little weight should be placed on it.

Figure 2 shows the trajectories. The published baseline is visibly flatter than the series throughout the descending limb; the full model tracks the turning points more closely but still trails the sharpest of them.

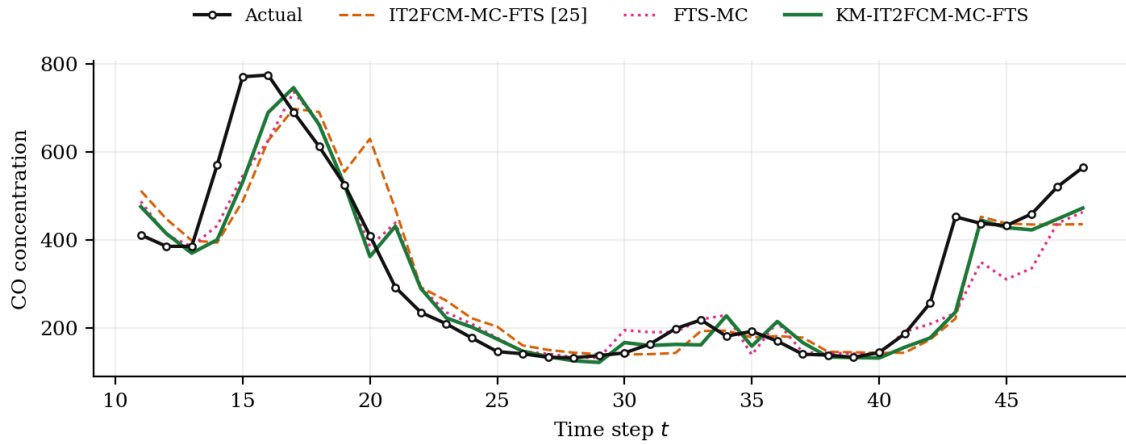


Figure 2. Actual series and one-step forecasts under the expanding-window protocol.

5.2. Error distribution

Aggregate error conceals the shape of the error distribution, which Figure 3 makes visible. Panel (a) places the full model and the Nie-Tan variant almost on top of one another and both below the published baseline on median absolute error and interquartile spread. Panel (b) shows how the separation accumulates. The baseline falls behind from the first forecast onwards rather than at any single turning point: the cumulative gap against the full model is already 36 at $t = 11$ and 181 at $t = 16$, and it widens most sharply between $t = 19$ and $t = 23$, over the steep descent from 525 to 209. By the end of the window it stands at 504, roughly a quarter of the baseline's total accumulated error.

Table 5. Series A: normality of the absolute error series (Shapiro-Wilk)

Configuration	W	p
IT2FCM-MC-FTS [10]	0.7809	4.36×10^{-6}
+ D_t only	0.7413	8.13×10^{-7}
+ exact KM only	0.8158	2.23×10^{-5}
Nie-Tan variant	0.7544	1.39×10^{-6}
FTS-IT2FCM	0.7456	9.70×10^{-7}
FTS-MC	0.8291	4.34×10^{-5}
KM-IT2FCM-MC-FTS	0.7561	1.50×10^{-6}

Normality is rejected for every configuration (Table 5), which is expected: absolute error on a spiking series is bounded below by zero and has a long right tail. The Wilcoxon test used throughout Table 4 is the appropriate choice.

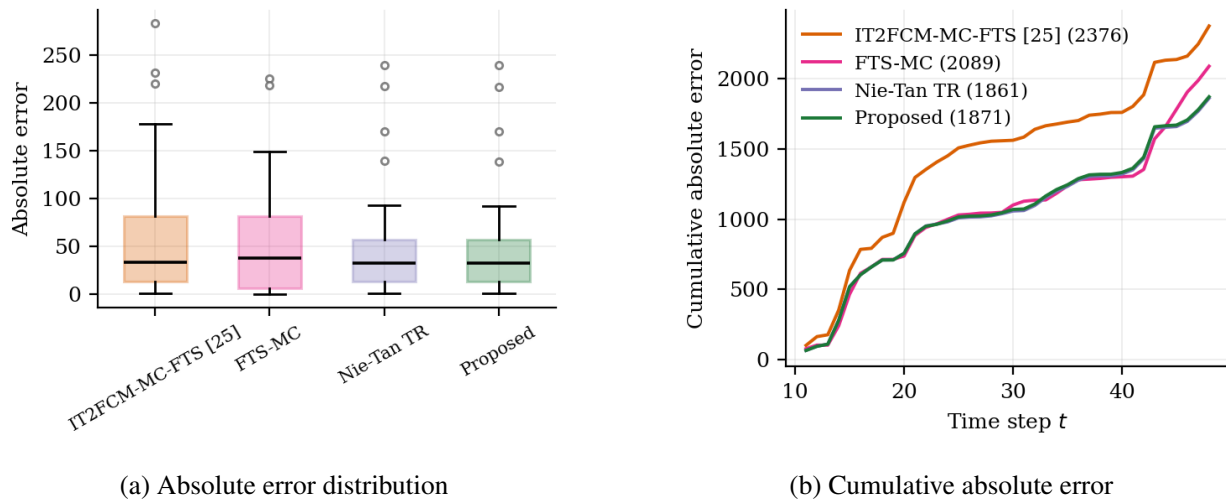


Figure 3. Distribution and accumulation of absolute error.

5.3. Responsiveness at spikes

Table 6 and Figure 4 address the behaviour the study set out to explain.

Table 6. Series A: mean absolute error at spike points ($|\Delta Y| > 76.31$, eight steps) and elsewhere (thirty steps)

Configuration	At spikes	Elsewhere
IT2FCM-MC-FTS [10]	150.31	39.11
+ D_t only	130.86	28.53
+ exact KM only	123.64	46.38
Nie-Tan variant	114.76	31.44
FTS-IT2FCM	118.74	31.13
FTS-MC	106.38	41.27
KM-IT2FCM-MC-FTS	114.58	31.83

The full model reduces spike error from 150.31 to 114.58, a 23.8% reduction, and non-spike error from 39.11 to 31.83. Both factors contribute at spikes, and here unlike in the aggregate exact type reduction does something: it lowers spike error from 150.31 to 123.64 on its own. But it does so while raising error everywhere else, from 39.11 to 46.38, which is why it does not improve RMSE overall. Widening the outermost intervals helps when the series leaves them and hurts when it does not.

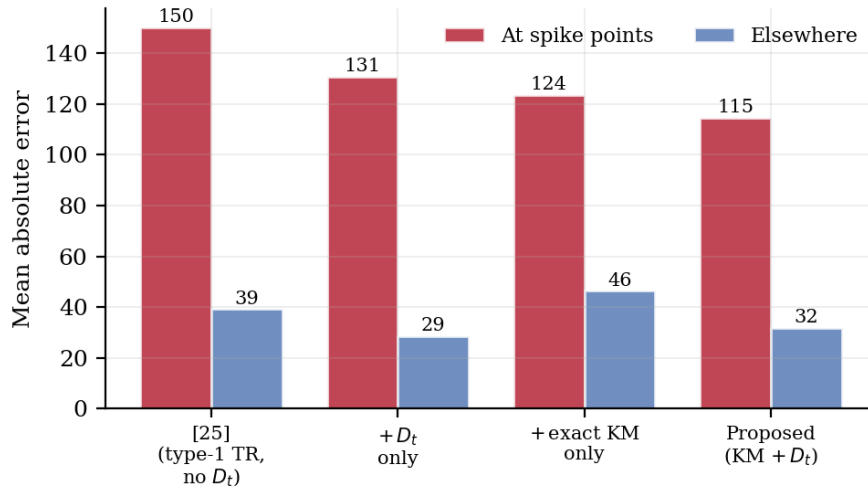


Figure 4. Decomposition of mean absolute error at spike points and away from them, across the four configurations of the primary factorial design. The two conventional ablations of Table 6 are omitted here for legibility.

One row of Table 6, not shown in Figure 4, deserves emphasis because it cuts against the framing of this literature. FTS-MC, with equal-width intervals and no interval type-2 machinery at all, attains the lowest spike error in the table (106.38). Its aggregate RMSE is worse (80.55) because it is correspondingly poor in the stable regime (41.27). Equal-width partitioning places wide intervals at the sparse high end of the series, which is exactly what a spike needs; adaptive partitioning narrows them there because few observations live there. The responsiveness advantage sought from a footprint of uncertainty is available more cheaply, and with a different trade-off, from a partition that ignores density altogether.

5.4. Residual diagnostics

Figure 5 examines what the full model leaves behind. Panel (a) shows no systematic drift the residuals do not trend with time (Spearman $\rho = -0.14$, $p = 0.41$) but the spread is plainly not constant. Residual magnitude scales with the level of the series ($\rho = 0.64$, $p < 0.001$): mean absolute residual is 21.77 over the steps below the median concentration and 76.73 above it, a factor of 3.5. The two visible clusters of large residuals, around $t = 14$ -21 and at $t = 43$, coincide with the spike points of Table 6.

This heteroscedasticity is expected rather than anomalous on a series spanning 132 to 946, and it is the reason SMAPE is reported alongside RMSE throughout: an error measure that is not scale-relative attributes almost all of the aggregate to the high-concentration episodes. It does, however, qualify the residual diagnostics. Panel (b) is centred close to zero and roughly symmetric, so the forecast is not biased in either direction on average, but a symmetric distribution pooled across two regimes with very different error scales says less than the same plot would on a homoscedastic series.

5.5. Replication on Series B

Series A supports one positive claim and one negative one. Series B, four times more densely populated per cluster and thirty times wider in range, tests both. Table 7 reports the same seven configurations under the same protocol.

The negative result replicates, in the regime that should have refuted it. Exact Karnik-Mendel and Nie-Tan differ by 24.87 in RMSE, one per cent of the error, at $p = 0.1576$. This is the comparison Section 5.9 of an earlier draft would have set aside as untested on well-populated clusters: with 13.7 observations per cluster the exact iteration still buys nothing measurable. Across the four initial windows tried (30, 60, 96, 144) the comparison reaches the five per cent level once, at the shortest evaluation set, and the sign of the difference is not stable.

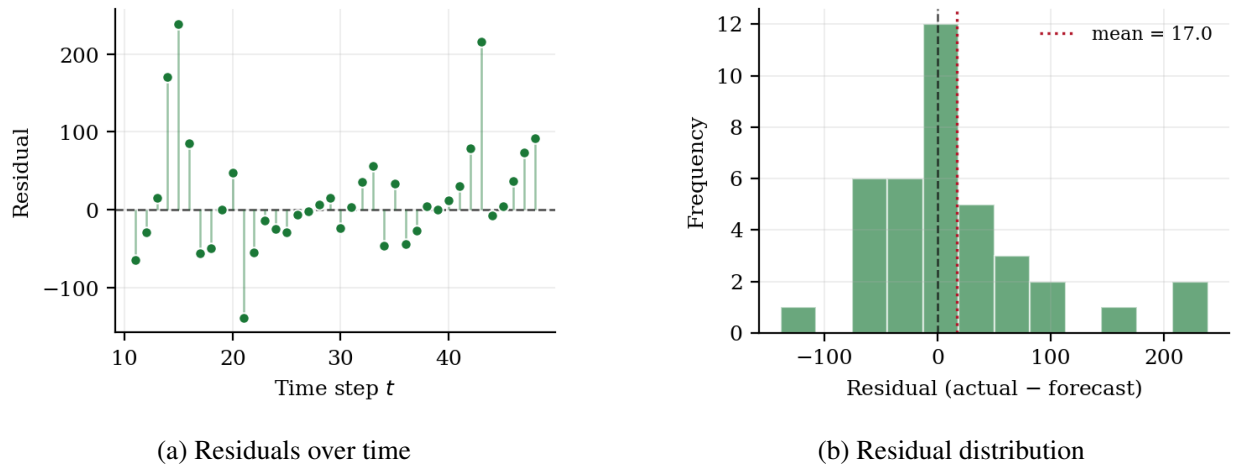


Figure 5. Residual diagnostics for KM-IT2FCM-MC-FTS.

Table 7. Series B: forecasting error and comparison against the full model (228 one-step forecasts, initial window 60)

Configuration	RMSE	SMAPE (%)	Wilcoxon W	p
IT2FCM-MC-FTS [10]	2356.63	20.33	12524.0	0.5957
+ D_t only	2374.50	20.78	11956.0	0.2712
+ exact KM only	2393.69	20.02	11633.0	0.9264
Nie-Tan variant	2402.32	20.02	11644.0	0.1576
FTS-IT2FCM (simple averaging)	2390.60	22.02	7311.0	0.0004
FTS-MC (equal-width partition)	2247.18	21.56	10412.0	0.0081
KM-IT2FCM-MC-FTS	2377.45	20.00		

The positive result does not replicate. The full model does not improve on the published baseline here. Its RMSE is marginally higher (2377.45 against 2356.63) and the difference is nowhere near significant ($p = 0.5957$); the same holds at every window tried, with p between 0.33 and 0.60. The cause is identifiable rather than mysterious: the directional term, which carried the entire gain on Series A, is actively harmful on Series B. Adding D_t alone to the baseline raises RMSE from 2356.63 to 2374.50, and does so at every window. Section 5.9 of this paper anticipated exactly this failure mode a fixed half-interval step commits to the direction of the last observed change, which overshoots whenever the series reverts and Series B, sampled every thirty minutes, reverts far more often than the irregularly sampled Series A. What survives is the SMAPE ordering: the full model is best on Series B at 20.00% against 20.33% for the baseline, so the benefit persists in the low-concentration regime where the metric is scale-relative, and is swamped in RMSE by the high-concentration episodes.

Figure 6 shows the trajectories over the first 120 evaluated steps. All three configurations track the diurnal shape and all three undershoot the peaks; the visual difference between them is slight, which is what Table 7 reports numerically.

The equal-width finding strengthens. FTS-MC attains the lowest RMSE on Series B by a clear margin (2247.18), significantly better than the full model ($p = 0.0081$), and the result holds at every window. On Series A the same configuration was best only at spike points; here it is best overall. Table 8 shows the decomposition.

The pattern of Table 6 recurs: the full model is best away from spikes and mid-field at them, and no configuration comes close to eliminating the lag spike error remains roughly six times the error elsewhere for every model in the table.

5.6. Computational cost

An operator that buys no measurable accuracy must still be paid for. Table 9 reports the wall-clock cost of a single partitioning fit, the quantity that the rolling protocol repeats at every step.

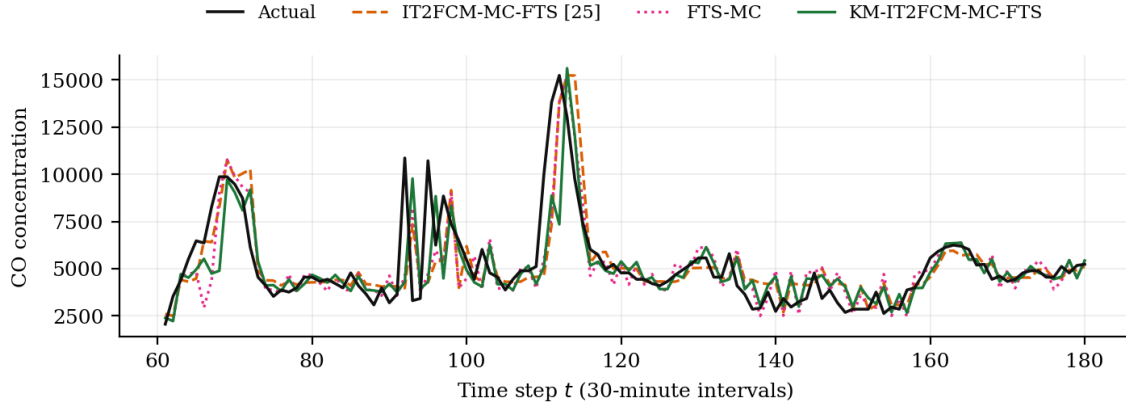


Figure 6. Series B: actual values and one-step forecasts over the first 120 evaluated steps (2.5 days), under the expanding-window protocol.

Table 8. Series B: mean absolute error at spike points ($|\Delta Y| > 1776.11$, 37 steps) and elsewhere (191 steps)

Configuration	At spikes	Elsewhere
IT2FCM-MC-FTS [10]	4459.40	760.64
+ D_t only	4463.61	769.37
+ exact KM only	4236.62	766.94
Nie-Tan variant	4268.75	738.95
FTS-IT2FCM	4045.85	895.79
FTS-MC	4217.69	796.71
KM-IT2FCM-MC-FTS	4261.67	727.11

Table 9. Cost of one partitioning fit, in seconds (best of three runs, single core). c is the state count returned by Eq. (5)

n	c	Exact KM	Nie-Tan	Type-1	KM / Nie-Tan
48	9	0.0289	0.0053	0.0021	5.5
96	8	0.0360	0.0087	0.0039	4.1
144	13	0.0772	0.0196	0.0185	3.9
192	15	0.1288	0.0397	0.0494	3.2
288	21	0.4451	0.1268	0.0493	3.5

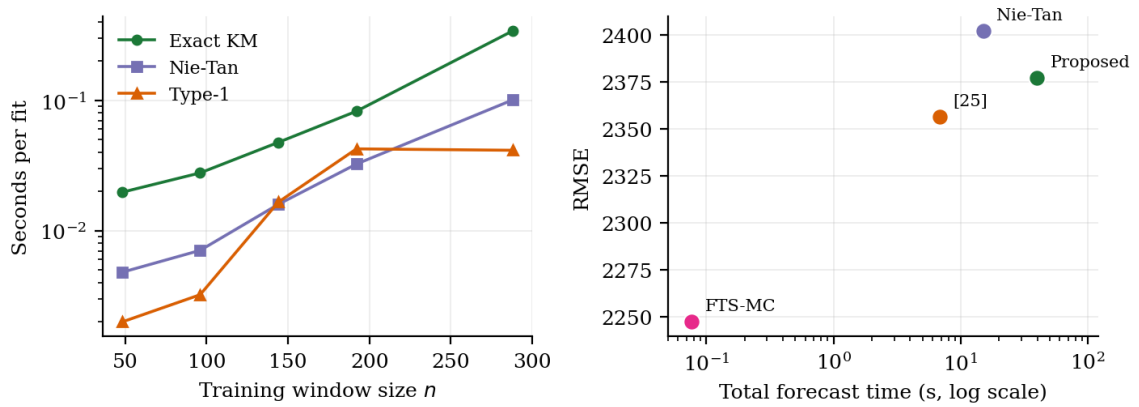


Figure 7. Left: cost of one partitioning fit against training window size (log scale). Right: accuracy against total forecast time on Series B; the cheapest configuration attains the lowest RMSE.

Exact type reduction costs three to five times a Nie-Tan fit throughout, and the gap does not close as n grows: from 48 to 288 observations the KM fit slows by a factor of 15.4 against 5.6 for the type-1 fit, because the switch-point search runs over a sorted array whose length grows with n , once per cluster, inside every outer iteration. Over the full rolling forecast on Series B the totals are 39.8 seconds for the proposed model, 15.0 for the Nie-Tan variant, 6.9 for the published baseline and 0.1 for equal-width partitioning a factor of roughly 400 between the most and least expensive configurations in the study, for differences in RMSE of a few per cent that mostly favour the cheap end. All timings were obtained on a single core of an Apple M2 processor (arm64, macOS 15.5) using Python 3.11.4, numpy 1.26.4 and scipy 1.14.1; results on other hardware will differ in absolute terms but the relative ratios between operators are determined by algorithmic structure and should be stable.

Table 10. Cost-accuracy trade-off on Series B (228 evaluated forecasts). Total wall-clock time is the sum over all expanding-window fits; RMSE is the one-step-ahead value. Efficiency ratio: lower RMSE per unit time is better (lower value preferred).

Configuration	RMSE	Total time (s)	RMSE $\times$ time
IT2FCM-MC-FTS [10]	2356.6	6.9	16 261
+ D_t only	2374.5	6.9	16 384
+ exact KM only	2393.7	39.8	95 269
Nie-Tan variant	2402.3	15.0	36 035
FTS-IT2FCM (simple averaging)	2390.6	5.2	12 431
FTS-MC (equal-width)	2247.2	0.1	225
KM-IT2FCM-MC-FTS	2377.5	39.8	94 625

The cost-accuracy trade-off in Table 10 makes the case starkly. The RMSE $\times$ time product for the proposed model (94,625) is more than 400 times that of FTS-MC (225), yet FTS-MC attains the lowest RMSE in the study. The Nie-Tan variant offers a middle ground: it costs 15.0 seconds against 39.8 for the proposed model while matching it within 1% in RMSE. For a deployment where computational budget matters, the Nie-Tan variant dominates the proposed model in this efficiency metric.

5.7. Sensitivity to the initial training window

The initial window is arbitrary, so every configuration was re-run for windows of 6, 10, 16, 24 and 30 observations on Series A and 30, 60, 96 and 144 on Series B. On Series A the ordering of the four primary configurations is stable: the full model and the Nie-Tan variant are always within 0.7 RMSE of one another, the D_t -only configuration is always closer to the full model than the KM-only configuration, and the published baseline is always worst. The significance of the improvement over [10] weakens as the window lengthens and the number of forecasts falls ($p = 0.002$ at a window of 6, 0.014 at 10, 0.111 at 16, 0.584 at 24), which is a statement about power rather than about effect.

On Series B the three conclusions of Section 5.5 hold at every window: the comparison against [10] is never significant (p between 0.33 and 0.60), D_t raises RMSE in every case, and FTS-MC has the lowest RMSE in every case. The exact-versus-surrogate comparison reaches the five per cent level once, at the longest window and hence the smallest evaluation set ($p = 0.0423$ at 144 forecasts against 0.1576, 0.1763 and 0.1640 elsewhere), which is the pattern of a null result sampled repeatedly rather than of an effect emerging.

5.8. Forecast horizon and the naive baseline

Everything to this point forecasts one step ahead, which is the convention in the FTS literature and the setting in which [9, 10] were evaluated. That convention deserves scrutiny. An air quality station reporting every thirty minutes will simply observe Y_{t+1} half an hour after Y_t , so a one-step forecast buys almost no decision time, and the question of whether it buys any accuracy over doing nothing is rarely asked. We ask it here.

The comparison is against persistence, $\hat{Y}_{t+1} = Y_t$, which requires no model at all. Table 11 reports it, together with a seasonal naive baseline $\hat{Y}_{t+h} = Y_{t+h-48}$ that repeats the value observed at the same time of day on the previous day.

Multi-step forecasts are obtained by raising the transition matrix to the required power, $\hat{Y}_{t+h} = \sum_j (P^h)_{s,t,j} v_j^*$, which is the natural extension of Eq. (8) and requires no additional machinery. The directional term is retained.

Table 11. Series B: accuracy against naive baselines across forecast horizons. Wilcoxon p -values compare the proposed model against persistence. Boldface marks the lower RMSE in each row

h	Lead time	n	Proposed		Persistence		Seasonal RMSE	p
			RMSE	SMAPE	RMSE	SMAPE		
1	30 min	228	2377.4	20.00	2255.6	19.17	3587.6	0.4137
2	1 h	227	2922.8	24.84	2950.3	25.59	3533.1	0.1717
4	2 h	225	3705.1	29.81	3969.2	33.64	3851.4	0.0015
6	3 h	223	4274.0	33.42	4742.6	40.86	4166.1	0.0000
12	6 h	217	4461.9	35.24	4971.0	42.79	4172.0	0.0003
24	12 h	205	4439.8	32.98	4900.5	46.11	4451.2	0.0000
48	24 h	181	3913.6	31.60	3845.1	34.40	4153.2	0.3887

At one step the model does not beat doing nothing. Persistence attains an RMSE of 2255.6 against 2377.4 for the proposed model, and a lower SMAPE as well. The difference is not significant ($p = 0.4137$), which is the weaker of the two possible readings: the entire pipeline interval type-2 clustering, exact type reduction, Markov weighting, directional correction produces one-step forecasts that cannot be distinguished from copying the last observation forward. The same holds on Series A, where persistence gives 70.97 against 74.46 ($p = 0.4551$). Every ablation in Sections 5-5.5 therefore compares configurations that are, collectively, no better than a baseline none of them was tested against.

The model earns its place from two to twelve hours ahead. From $h = 4$ onwards the ordering reverses and the margins are significant: the proposed model improves on persistence by 6.7% in RMSE at two hours, 9.9% at three hours, 10.2% at six hours and 9.4% at twelve hours, with $p \leq 0.0015$ throughout, and by considerably more in SMAPE. At $h = 48$ the advantage disappears, because a full day ahead the previous day's value at the same clock time is a strong predictor and the seasonal baseline becomes competitive with both. Figure 8 shows the crossover.

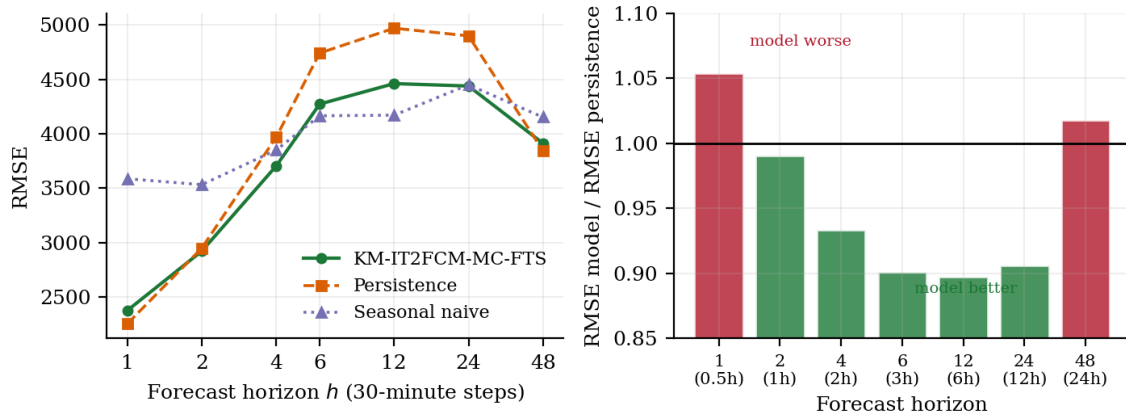


Figure 8. Series B: accuracy against forecast horizon. Left, RMSE of the proposed model and the two naive baselines. Right, the ratio of model to persistence RMSE; values below one indicate the model is more accurate. The useful band runs from two to twelve hours.

At the useful horizon the ablation distinctions vanish. Repeating the full factorial design at $h = 6$ gives 4470.7 for the published baseline, 4464.9 with D_t , 4331.9 with exact KM alone, 4267.9 for the Nie-Tan variant, 4505.9 for equal-width partitioning and 4274.0 for the full model. No configuration differs significantly from the full model (p between 0.2275 and 0.9347), and the spread across all six is 5.6% of the error. What matters at this horizon is having a transition matrix that can be iterated at all; which type-reduction operator produced the cluster centres, and whether a directional term was added, no longer registers.

5.9. Discussion

The hypothesis this study set out to test was that approximate type reduction is responsible for the lag observed in IT2FCM-Markov hybrids. It is not, on either series. Nie-Tan and exact KM produce forecasts differing by 0.07% of RMSE on Series A and 1.0% on Series B, neither significant, while the exact iteration costs three to five times as much per fit. The contraction of the footprint of uncertainty is genuine, but the equilibrium points are the only quantity that survives into Stage 2, and displacing them by a few units moves interval boundaries that are tens or hundreds of units wide. The mechanism is real and the effect is negligible. An information-theoretic reading of this supports the same conclusion: the centroid interval $[v_i^L, v_i^R]$ produced by KM and the narrower interval from the Nie-Tan initialisation both collapse to a single crisp midpoint v_i^* before passing to Stage 2. Whatever contraction the surrogate introduces, it affects only the width of the interval that is immediately discarded; the boundary positions that determine state assignment and hence the transition matrix depend on the midpoints alone. A small displacement of a midpoint within a partition whose intervals span tens to hundreds of units cannot materially change state membership or transition counts.

The strength of this conclusion rests on Series B. On Series A alone the null could be attributed to sparse boundary clusters; Series B places 13.7 observations per cluster over a range thirty times wider, and returns the same result. The conditions under which exact type reduction would pay for itself in a forecasting pipeline, if they exist, are not these.

The alternative explanation fares better as a diagnosis than as a prescription. That a probability-weighted average of the cluster centres is trapped in their convex hull is not in doubt; it is arithmetic, and it applies to any FTS model whose defuzzification is a convex combination, including the Pythagorean and intuitionistic variants that share the structure [26, 38]. What the two series show is that escaping the hull with a fixed half-interval step is not a reliable remedy. On Series A the term lowers RMSE by 17% and carries the entire improvement over the published baseline; on Series B it raises RMSE at every window tried. The difference between the series is the direction of the mechanism's error: a correction that commits to the sign of the last observed change pays off when moves persist and costs when they revert, and a half-hourly urban CO series reverts often. The convexity bound is worth attacking; D_t as defined in Eq. (9) is not the way to attack it.

The failure mode of D_t on Series B points toward the design of a successor. The term as defined commits unconditionally to the sign of the most recent observed change; it pays off when moves persist and costs when they revert. A regime-aware correction would first estimate whether the series is currently trending or mean-reverting, then scale or suppress D_t accordingly. The lag-1 autocorrelation of the training window, or the Hurst exponent estimated over a rolling window, provides a low-cost signal for this: positive autocorrelation supports the directional adjustment; negative autocorrelation suggests it should be zeroed. On a half-hourly urban CO series, the sign of lag-1 autocorrelation in the training window is an adequate proxy for local persistence. We did not implement this modification because the present paper tests existing components rather than proposing new ones; it is, however, the natural extension. The key design requirement is that the detection threshold itself be calibrated on training data and not on future observations.

One result runs consistently against the framing of this literature and deserves to be stated plainly. Equal-width partitioning, with no interval type-2 machinery at all, attains the lowest spike error on Series A and the lowest overall RMSE on Series B, significantly so ($p = 0.0081$, Holm-Bonferroni adjusted $p = 0.040$), at roughly 1/400 of the computational cost. The reason is visible in the construction: adaptive partitioning places narrow intervals where observations are dense and wide ones where they are sparse, which is exactly backwards for accommodating an excursion into the sparse tail. The interval widths confirm this numerically. On Series B, the 21-state IT2FCM partition produces two tail intervals covering the spike-value region that average 4,959 concentration units wide, against the uniform equal-width benchmark of 1,184 units. A value arriving during a spike must therefore be assigned to a state whose fuzzy set is more than four times wider than the equal-width alternative, reducing the resolution with which the ensuing transition is specified. On Series A the effect is smaller in absolute terms (IT2FCM tail-2 average 61 units vs. equal-width 51) but structurally identical. The sparsity problem also manifests in cluster occupancy: on Series B, seven of the 21 IT2FCM states contain fewer than five observations in the final training window, meaning the corresponding rows of the transition matrix are estimated from fewer than five counts and are correspondingly unstable. Equal-width states, by construction, distribute observations more uniformly and

reduce the incidence of near-empty boundary rows. The sophistication of the partitioning stage is not what limits these models, and on the evidence here it is not obviously what improves them either.

The horizon result reframes all of this. A literature that evaluates only at $h = 1$ has been comparing models on a task where none of them beats persistence. Once persistence is in the table the ranking question changes shape: the proposed model is not better than doing nothing at thirty minutes; it is meaningfully better than doing nothing at three to twelve hours; and at that horizon none of the six configurations differs significantly from another. The mechanism that earns the improvement is the transition matrix and the ability to iterate it, not the machinery upstream of it.

This has a direct bearing on deployment. Continuous monitoring networks report at intervals of one to thirty minutes, so a one-step forecast on a half-hourly series is superseded by the next observation before it can inform anything; the lead time it offers is shorter than the response time of any intervention it might trigger. The horizons at which this class of model does add information two to twelve hours are also the horizons at which air quality decisions are actually taken: issuing a health advisory for the afternoon, or restricting traffic in a corridor before an evening peak. We would therefore argue that reporting $h = 1$ accuracy alone, as [9, 10] and much of the surrounding literature do, understates and misdirects the case for these models at once. For a practitioner operating a half-hourly air quality network, the practical recommendation is direct: at one step ahead, persistence is not only free but statistically indistinguishable from the full pipeline; the model earns its running cost only from two to twelve hours ahead, which is also the window in which an advisory or a traffic restriction can still be acted upon. This finding aligns with the observation in [41] that increased architectural complexity does not necessarily improve short-term air quality forecasting accuracy, and underscores the importance of reporting multi-horizon performance rather than a single $h = 1$ error.

The practical implication remains narrow even so. Spike error is roughly six times non-spike error for every configuration on Series B, and no model here misses a surge by less than 4000 concentration units against an observed range of 24871. At three hours ahead the proposed model still carries a SMAPE of 33%. Forecasts of this class can characterise typical diurnal conditions and support policy evaluation over time; none is accurate enough at surges to trigger a public alert on its own.

6. Conclusion

Four findings emerge from the factorial ablation of two candidate mechanisms against a leak-free expanding-window protocol on two ambient carbon monoxide series.

Neither candidate mechanism holds up. Exact Karnik-Mendel type reduction is statistically indistinguishable from the Nie-Tan surrogate on both series including the data-rich regime the contraction argument implicates while costing three to five times as much per fit. The directional term lowers RMSE by 17% on the 48-point series and raises it on the 288-point series, because a fixed step in the direction of the last observed change overshoots when the series reverts. The improvement over [10] on the short series ($p = 0.0136$) does not replicate on the long one ($p = 0.5957$); it is a series-specific result.

The finding that subsumes both is about horizon. At one step ahead the complete pipeline is indistinguishable from persistence on both series, so the configurations compared in the ablation are collectively no better than copying the last observation forward a baseline this literature does not normally report. Iterating the transition matrix to P^h locates where the model does add information: from two to twelve hours ahead it improves on persistence by 6.7 to 10.2% in RMSE, significantly in every case, and the advantage lapses again at twenty-four hours where the previous day's value at the same clock time becomes a strong predictor. At the useful horizon the six configurations of the factorial design span 5.6% of the error and none differs significantly from another.

Three limitations bound these conclusions. Both series come from a single monitoring station and a single pollutant. Carbon monoxide in an urban corridor has characteristic dynamics a bimodal diurnal cycle, impulsive traffic-driven spikes, and a mean-reverting half-hourly structure that are not shared by $\text{PM}_{2.5}$, NO_2 , or pollutants measured in non-urban or high-altitude environments. The null finding on type reduction and the failure of D_t on Series B may reflect properties specific to these dynamics rather than to IT2FCM architectures generally;

replication on a different pollutant, a different city, or a different sampling regime is required before drawing conclusions about the model class. The fuzzifier interval is inherited from [21] rather than tuned, and the state count follows a volatility heuristic designed for neither sample size; the sensitivity analysis in Section 5.7 shows the ordering of conclusions is stable to the initial training window, but sensitivity to the state count itself or to a fixed rather than heuristic state count has not been reported here and represents a further assumption. The multi-step results rest on iterating a first-order chain, which assumes not only that the transition structure is Markovian but that the matrix estimated from the training window at step t remains valid at horizons up to $h = 48$. A directly estimated h -step matrix, fitted from pairs (s_t, s_{t+h}) in the training window, imposes no stationarity assumption across horizons but requires a sample large enough to populate all c^2 cells of each matrix; on Series A this is not feasible. A higher-order chain would capture multi-step dependencies at the cost of a state space that grows exponentially in order. Both alternatives are natural successors to the first-order design used here and represent the most tractable direction for future work.

The most useful next steps follow from the negative results rather than any positive one. Type-reduction operators are not where the accuracy of these pipelines is decided, and effort spent refining them is unlikely to be repaid; the defuzzification step, where the convexity bound lives, is the better target, though the failure of D_t on Series B shows that a correction must adapt to the local persistence of the trend rather than assume it. More consequentially, this class of model should be evaluated and reported at the horizons where it beats a naive baseline, with that baseline shown. On the evidence here that band is two to twelve hours for half-hourly urban CO which is also, and not coincidentally, the window in which an air quality advisory or a traffic restriction can still be acted upon.

Acknowledgement

This work was supported by the Doctoral Dissertation Research Grant (Hibah Penelitian Disertasi Doktor, PDD) of the Ministry of Higher Education, Science and Technology of the Republic of Indonesia, under principal contract no. 224/C3/DT.05.00/PL-BARU/2026 and derivative contract no. 254-12/UN7.D2/PP/IV/2026. The authors also thank the Environmental Agency of Semarang City (Dinas Lingkungan Hidup Kota Semarang) for providing access to the ambient concentration records analysed here.

Data and code availability

Series A is embedded in the experiment script. Series B, the complete implementation that produced every number and figure in this paper (including the re-run of [10], the multi-horizon sweep, and the naive-baseline comparison), are publicly available at Zenodo: <https://doi.org/10.5281/zenodo.22686800>. The repository contains `km_it2fcm_mc_fts.py` (main experiment), `make_figures.py` (figure generation), `konsentrasi.csv` (Series B data), and a `README.md` with step-by-step reproduction instructions. All results are reproducible with `python km_it2fcm_mc_fts.py` using only `numpy` and `scipy`.

REFERENCES

1. World Health Organization, *WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*, World Health Organization, Geneva, 2021.
2. F. Noronha, D. Ribeiro, P. Salinas, J. Soares, A. P. de Oliveira, R. M. de Miranda, L. Antunes, and T. Souza, *The evolution of temporal and spatial patterns of carbon monoxide concentrations in the metropolitan area of Sao Paulo, Brazil*, *Advances in Meteorology*, vol. 2016, 8570581, 2016.
3. T. H. Wen, J. A. Jiang, C. H. Sun, J. Y. Juang, and T. S. Lin, *Monitoring street-level spatial-temporal variations of carbon monoxide in urban settings using a wireless sensor network (WSN) framework*, *International Journal of Environmental Research and Public Health*, vol. 10, no. 12, pp. 6380-6396, 2013.
4. Q. Song, and B. S. Chissom, *Forecasting enrollments with fuzzy time series Part I*, *Fuzzy Sets and Systems*, vol. 54, no. 1, pp. 1-9, 1993.
5. Q. Song, and B. S. Chissom, *Forecasting enrollments with fuzzy time series Part II*, *Fuzzy Sets and Systems*, vol. 62, no. 1, pp. 1-8, 1994.

6. S. M. Chen, *Forecasting enrollments based on fuzzy time series*, Fuzzy Sets and Systems, vol. 81, no. 3, pp. 311-319, 1996.
7. U. M. Sirisha, M. C. Belavagi, and G. Attigeri, *Profit prediction using ARIMA, SARIMA and LSTM models in time series forecasting: A comparison*, IEEE Access, vol. 10, pp. 124715-124727, 2022.
8. A. K. Dubey, A. Kumar, V. García-Díaz, A. K. Sharma, and K. Kanhaiya, *Study and analysis of SARIMA and LSTM in forecasting time series data*, Sustainable Energy Technologies and Assessments, vol. 47, 101474, 2021.
9. A. Bardadi, B. Warsito, B. Surarso, and W. H. Sugiharto, *Interval type-2 fuzzy C-means (IT2FCM) for robust forecasting of carbon monoxide concentrations in urban air quality monitoring*, Mathematical Modelling of Engineering Problems, vol. 12, no. 6, pp. 2021-2030, 2025.
10. A. Bardadi, B. Warsito, B. Surarso, and W. H. Sugiharto, *Managing partitioning uncertainty and transition ambiguity in fuzzy time series: A robust IT2FCM-Markov chain approach*, Mathematical Modelling of Engineering Problems, vol. 12, no. 12, pp. 4406-4412, 2025.
11. N. N. Karnik, and J. M. Mendel, *Centroid of a type-2 fuzzy set*, Information Sciences, vol. 132, no. 1-4, pp. 195-220, 2001.
12. L. Palomero, V. García, and J. S. Sánchez, *Fuzzy-based time series forecasting and modelling: A bibliometric analysis*, Applied Sciences, vol. 12, no. 14, 6894, 2022.
13. A. Folaponmle, S. F. Kolawole, and S. N. John, *Air pollution forecasting using fuzzy time series models for Kaduna Metropolis, Nigeria*, FUOYE Journal of Engineering and Technology, vol. 8, no. 2, pp. 142-146, 2023.
14. T. Pribadi, D. Siregar, A. Amalia, A. S. Lubis, T. A. Parluhutan, F. Kumalasari, and J. L. Marpaung, *Application of the Cheng fuzzy time series model for stock price forecasting: A case study in the energy sector*, Mathematical Modelling of Engineering Problems, vol. 12, no. 8, pp. 2741-2753, 2025.
15. H. L. Chen, X. D. Gao, and Q. Wu, *An enhanced fuzzy time series forecasting model integrating fuzzy C-means clustering, the principle of justifiable granularity, and particle swarm optimization*, Symmetry, vol. 17, no. 5, 753, 2025.
16. Z. T. Gong, and J. D. Feng, *A novel multifactor type-2 fuzzy time series model based on improved fuzzy C-means algorithm and justifiable granularity for stock index forecasting*, Soft Computing, vol. 29, pp. 3719-3732, 2025.
17. R. B. A. Satriani, F. Farikhin, and B. Surarso, *Hybrid heuristic model and fuzzy C-means for stock forecasting using type 2 fuzzy time series*, Interdisciplinary Social Studies, vol. 4, no. 1, pp. 27-36, 2024.
18. C. Hwang, and F. C. H. Rhee, *Uncertain fuzzy clustering: Interval type-2 fuzzy approach to C-means*, IEEE Transactions on Fuzzy Systems, vol. 15, no. 1, pp. 107-120, 2007.
19. P. Hernández, S. Cubillo, and C. Torres-Blanc, *A complementary study on general interval type-2 fuzzy sets*, IEEE Transactions on Fuzzy Systems, vol. 30, no. 11, pp. 5034-5043, 2022.
20. T. H. Dang, L. T. Ngo, and W. Pedrycz, *Interval type-2 fuzzy C-means approach to collaborative clustering*, in 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istanbul, Turkey, pp. 1-7, 2015.
21. Y. Yin, Y. H. Sheng, and J. R. Qin, *Interval type-2 fuzzy C-means forecasting model for fuzzy time series*, Applied Soft Computing, vol. 129, 109574, 2022.
22. M. H. Odah, *Unemployment rate forecasting via Markov chains*, International Journal of Sustainable Development and Planning, vol. 16, no. 6, pp. 1191-1194, 2021.
23. Y. Alyousifi, M. Othman, I. Faye, R. Sokkalingam, and P. C. L. Silva, *Markov weighted fuzzy time-series model based on an optimum partition method for forecasting air pollution*, International Journal of Fuzzy Systems, vol. 22, pp. 1468-1486, 2020.
24. Y. Alyousifi, M. Othman, A. Husin, and U. Rathnayake, *A new hybrid fuzzy time series model with an application to predict PM10 concentration*, Ecotoxicology and Environmental Safety, vol. 227, 112875, 2021.
25. A. Alhendi, A. S. Al-Sumaiti, M. Marzband, R. Kumar, and A. A. Z. Diab, *Short-term load and price forecasting using artificial neural network with enhanced Markov chain for ISO New England*, Energy Reports, vol. 9, pp. 4799-4815, 2023.
26. S. D. Xian, and Y. Cheng, *Pythagorean fuzzy time series model based on Pythagorean fuzzy c-means and improved Markov weighted in the prediction of the new COVID-19 cases*, Soft Computing, vol. 25, pp. 13881-13896, 2021.
27. D. Wu, and J. M. Mendel, *Enhanced Karnik-Mendel algorithms*, IEEE Transactions on Fuzzy Systems, vol. 17, no. 4, pp. 923-934, 2009.
28. M. Nie, and W. W. Tan, *Towards an efficient type-reduction method for interval type-2 fuzzy logic systems*, in 2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence), Hong Kong, China, pp. 1425-1432, 2008.
29. F. Chung-Hoon Rhee, *Uncertain fuzzy clustering: Insights and recommendations*, IEEE Computational Intelligence Magazine, vol. 2, no. 1, pp. 44-56, 2007.
30. E. Rubio, and O. Castillo, *Interval type-2 fuzzy clustering algorithm using the combination of the fuzzy and possibilistic C-mean algorithms*, in 2014 IEEE Conference on Norbert Wiener in the 21st Century (21CW), Boston, MA, USA, 2014.
31. H. He, X. Yu, and D. Hu, *Land cover classification based on adaptive interval-valued type-2 fuzzy clustering analysis*, in Knowledge Science, Engineering and Management (KSEM 2015), pp. 712-720, 2015.
32. J. C. Figueroa-García, D. Kalenatic, and C. A. Lopez, *Interval type-2 fuzzy Markov chains: Type reduction*, in 7th International Conference on Advanced Intelligent Computing Theories and Applications, Zhengzhou, China, pp. 211-218, 2011.
33. J. C. F. García, *Interval type-2 fuzzy Markov chains: An approach*, in 2010 Annual Meeting of the North American Fuzzy Information Processing Society, Toronto, ON, Canada, 2010.
34. J. C. Figueroa-García, *Interval type-2 fuzzy Markov chains*, in Advances in Type-2 Fuzzy Sets and Systems, pp. 49-64, 2013.
35. O. Linda, and M. Manic, *General type-2 fuzzy C-means algorithm for uncertain fuzzy clustering*, IEEE Transactions on Fuzzy Systems, vol. 20, no. 5, pp. 883-897, 2012.
36. Y. Alyousifi, and M. Othman, *Forecasting PM10 concentration based on a hybrid fuzzy time series model*, in the 1st International Conference on Artificial Intelligence for Smart Community (AISC 2020), Universiti Teknologi Petronas, Malaysia, pp. 177-184, 2022.
37. Z. G. Liu, Q. Pan, J. Dezert, and G. Mercier, *Credal c-means clustering method based on belief functions*, Knowledge-Based Systems, vol. 74, pp. 119-132, 2015.
38. O. C. Yolcu, and U. Yolcu, *A novel intuitionistic fuzzy time series prediction model with cascaded structure for financial time series*, Expert Systems with Applications, vol. 215, 119336, 2023.

39. W. H. Sugiharto, A. A. Riadi, and M. I. Ghazali, *Web based information system of carbon monoxide pollution*, *E3S Web of Conferences*, vol. 73, Art. No. 05026, 2018.
40. T. Herlambang, Z. Othman, and M. R. Arief, *Design and implementation of an IoT-based indoor air quality monitor with ensemble and unscented Kalman filter forecasting*, *Statistics, Optimization & Information Computing*, vol. 16, no. 4, pp. 3734-3746, 2026.
41. L. Anggraini, E. Noersasongko, A. Marjuni, and Purwanto, *Multivariate hourly air quality forecasting with MES-LSTM as the core framework*, *Statistics, Optimization & Information Computing*, vol. 16, no. 1, 2026.