Data-driven Partitioning Approaches for Type-2 Fuzzy Set Based Time Series

Arthur Caio Vargas e Pinto*, Christian Wagner† Frederico Gadelha Guimarães‡, and Eduardo Pestana de Aguiar§
* *Graduate Program in Electrical Engineering, Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil,*
† *School of Computer Science, University of Nottingham, Nottingham, United Kingdom,*
‡ *Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil,*
§ *Department of Mechanical Engineering, Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil*
Email: arthur.vargas@ifmg.edu.br, christian.wagner@nottingham.ac.uk, fredericoguimaraes@ufmg.br, eduardo.aguiar@ufjf.br

***Abstract*—Fuzzy Set Based Time Series (FTS) prediction techniques offer potential advantages in efficient handling of uncertainty in the data. However, such prediction models are commonly challenging to design, requiring careful and application-specific tuning of hyperparameters to provide competitive forecasting performance. Conventional FTS models often rely on predefined partitioning schemes and user-specified parameters, which may introduce subjectivity and limit their adaptability to complex datasets. This work presents novel data-driven interval type-2 fuzzy set based time series models — SODA-T2FTS, ADP-T2FTS, and ADP-T2LIMG — that integrate interval type-2 fuzzy logic with data-driven partitioning techniques. These approaches handle uncertainty and improve predictive performance while reducing reliance on user intervention. The models were evaluated on financial, benchmark, and energy time series datasets, and evaluation was overall performed in terms of the average number of rules (c.f. interpretability), error metrics, execution time, model complexity and noise sensitivity. Results showed that the proposed models demonstrated superior accuracy and competitive interpretability compared to state-of-the-art forecasting techniques, also highlighting that data-driven approaches significantly enhance fuzzy set based time series forecasting by improving partitioning accuracy, reducing subjectivity, and increasing adaptability to different datasets.**

***Index Terms*—Type-2 Fuzzy Systems. Forecasting. Type-2 Fuzzy Set Based Time Series. Data-driven partitioning.**

I. Introduction

Time series forecasting is a technique widely studied in academic research, which consists of analysing data and the sequence of time in order to predict future events, with widespread applications in economics, energy systems, climate monitoring, and beyond. However, uncertainty, challenging behaviors, and patterns within specific time series data might make them challenging to examine and forecast. Dealing with the uncertainty in time series data, in particular, is a recurrent problem in forecasting and decision-making, especially in complex systems where data is often imprecise, incomplete, or affected by noise.

Addressing these uncertainties requires forecasting models to be capable of accounting for vagueness, imprecision, and randomness in time series data. Fuzzy Set Based Time Series (FTS) models have emerged as an effective forecasting approach to handling uncertainty by representing time series data in linguistic terms and incorporating fuzzy logic principles. Unlike conventional statistical models, which rely on precise numerical values, FTS models define time series observations using type-1 fuzzy sets, allowing for a more flexible representation of uncertainty [1]. Fuzzy logic was first proposed in 1965 by Zadeh 2, and then Song and Chissom 3 introduced the FTS method by implementing it on the data of student enrollments at the University of Alabama. Later, Chen et al. 4 introduced a method using simplified arithmetic operations instead of the complex max-min composition operations proposed by Song and Chissom. This model is considered as the milestone in this research area [5].

One major source of uncertainty in FTS methods is the subjective determination of partitioning hyperparameters, which define how the time series data is divided into intervals for the creation of fuzzy sets. For this, some hyperparameters must be defined mainly the partitioning method, the number of partitions and model order [6]. Since FTS models rely on linguistic representation through fuzzy sets, the choices regarding these hyperparameters significantly impact model accuracy and robustness [7].

Although many improvements have been made in FTS forecasting models, there are limitations of type-1 FTS to model and minimize the impact of uncertainties expected in some applications. Therefore, type-2 fuzzy sets were presented by 8 as an extension of the concept of a conventional fuzzy set, providing an additional degree of freedom to model uncertainty and imprecision in a better way [9]. The key advantage of type-2 fuzzy sets (T2FS) over type-1 fuzzy sets (T1FS) is their ability to model uncertainty within the membership functions themselves. In a T1FS, each input value is assigned a crisp membership degree, whereas in a T2FS, the membership function is itself fuzzy, meaning that each input has a range of possible membership values [10].

Despite significant improvements published in the type-2 FTS literature, much still needs to be studied. There are several hyperparameters to be defined in FTS design, specially the partitioning hyperparameters that directly impact in model's accuracy [1], also considering that most of the partitioning/clustering methods used for this task often rely on user-defined parameters for partitioning, introducing subjectivity and limiting adaptability to complex datasets. Therefore, to address the difficulty in FTS partitioning and uncertainty handling, this research aims to investigate FTS forecasting model design by understanding the importance of key design features, to propose new accurate and effective type-2 FTS methods and to validate their performance in different appropriate case studies, as illustrated in Figure 1.

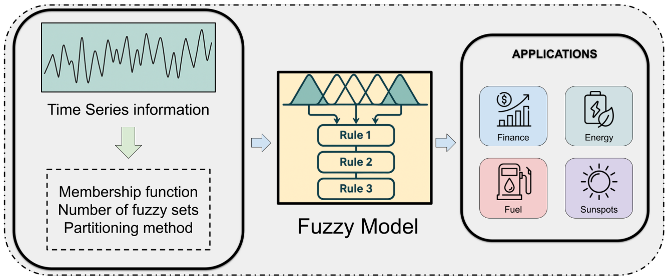

Fig. 1. Overall FTS process in this research.

The main contributions o this research are the Introduction of fully data-driven partitioning algorithms (SODA and ADP) for FTS model design that remove reliance on manual heuristics, thereby addressing a key source of uncertainty in type-2 FTS model design. Also, the development of the SODA-T2FTS and ADP-T2FTS models to address the limitations of user-dependent partitioning methods in conventional FTS, enabling more adaptive, accurate, and uncertainty-aware forecasting for univariate time series. Finally, the extension of ADP-T2FTS into ADP-T2LIMG to capture deeper structural information from the data, improving fuzzy set definition and enhancing forecast accuracy in complex time series.

II. Theoretical Foundations

A. Fuzzy Set Based Time Series Prediction

Unlike classical time series analysis, which depends on crisp numerical information and deterministic models, in Fuzzy Set Based Time Series (FTS) models, time-dependent data is represented using linguistic variables and fuzzy sets. These fuzzy sets may capture the inherent uncertainty in many real-world scenarios, such as economic indicators, weather patterns, and stock prices, where precise numerical values may not fully represent the underlying indicator dynamics. FTS is built upon three fundamental components: the Universe of Discourse - UoD, fuzzy sets, and their respective fuzzy relationships. This UoD defines the range of possible values for the time series, and within this range, fuzzy sets are created. These fuzzy sets are then used to establish fuzzy logical relationships, which model the temporal dependencies between consequent observations.

Type-2 FTS models extend traditional FTS models, which rely on Type-1 fuzzy sets, by incorporating type-2 fuzzy sets. This extension allows for a more nuanced representation of uncertainty, particularly useful when data is noisy, incomplete, or subject to subjective human judgment. A Type-2 fuzzy set differs from a Type-1 fuzzy set in that it has a fuzzy membership function instead of a crisp one. Interval Type-2 Fuzzy Sets (IT2FSs) are a simplified version of Type-2 fuzzy sets, where the secondary membership function is confined to intervals. This makes them computationally efficient while still retaining the ability to model complex uncertainties.

The forecasting procedure that is used by FTS methods presented in this research can be divided into training and forecasting procedures, which are detailed in Appendix I. The training procedure aims to create the fuzzy sets and a knowledge representation of the time series dynamics, both represented by the learned model $\mathcal{M}$.

B. Data-driven Partitioning Approaches

In the context of FTS models, where appropriate partitioning is essential for model performance, data-driven methods for dividing data into distinct intervals without requiring manual intervention are particularly useful as they work directly from the analysis of data distribution, ensuring that the partitions capture meaningful patterns or structures in the dataset while minimizing information loss. In this work, two data-driven partitioning algorithms are explored for UoD partitioning: the Self-Organised Direction Aware algorithm - SODA [11] and the Autonomous Data Partitioning - ADP [12]. These algorithms are designed to use local modes derived from data density peaks to partition datasets into shape-free data clouds. ADP and SODA are grounded in Empirical Data Analysis (EDA), which avoids assumptions inherent in traditional probability-based models. By identifying local density maxima, the algorithm creates data clouds that reflect the natural structure of the data. For both SODA and ADP there is a hyperparameter named *Gridsize* which is set to decide the granularity of the clustering results. The *Gridsize* helps defining the distance threshold for clouds to be considered close enough to be merged, hence directly impacting in model accuracy.

III. The SODA-T2FTS Approach

The SODA-T2FTS model introduces a novel integration between Self-Organized Direction Aware Data Partitioning (SODA) and interval type-2 fuzzy logic. SODA is a density-aware algorithm that considers local and global data directions to define UoD partitions without user intervention. Experiments with financial datasets (TAEIX, NASDAQ and S&P500) (Appendix II). indicate that the proposed model obtained the lowest Root Mean Squared Error (RMSE) in all experiments, including those in which noise was added to the time series, indicating that the proposed forecasting model can predict complex time series with high accuracy using a data-driven approach independent of user interference. Table I shows the results for the NASDAQ time series. In addition to outperforming in accuracy, the proposed approach presented a similar number of fuzzy rules compared to other FTS methods, thus, delivering better performance for comparable interpretability.

IV. The ADP-T2FTS Approach

SODA-T2FTS represents just one pathway to address the broader challenge of automatic and adaptive fuzzy set genera-

TABLE I
CALCULATED RMSE VALUES.

Dataset	Model	Order	Part.	RMSE AVG	RMSE STD	Rank	Time(s)
NASDAQ	ARIMA	1,0,0	-	28.53	11.07	6	0.59
	CFTS	1	10	31.89	11.02	10	2.73
	EWFTS	1	10	29.01	11.02	7	1.59
	FTS	1	15	29.50	9.67	8	4.81
	HOFTS	2	10	30.26	10.94	9	5.68
	Hwang	2	10	36.11	14.79	11	0.81
	IWFTS	1	5	28.31	10.68	5	1.45
	Naive	1	-	41.24	15.15	12	0.03
	PWFTS	3	3	27.38	10.43	2	19.68
	TWFTS	1	8	27.64	10.65	3	2.00
	SODA-T2FTS	**1**	**27**[a]	**24.20**	**8.35**	**1**	**43.40**
	WFTS	1	4	28.22	10.80	4	1.60

[a] For Gridsize = 10.

tion in FTS methods. In this regard, an alternative approach titled ADP-T2FTS is explored, which employs the Autonomous Data Partitioning (ADP) algorithm to define the UoD based on local and global data density. Among other experiments, ADP-T2FTS was used to forecast the TAIEX time series and was compared to state-of-the-art forecasting models designed using type-2 fuzzy logic, neural networks and regression algorithms, studied in 13. The results for all models can be seen in Table II, with the best-performing ones highlighted in bold. ADP-T2FTS outperforms other models for the years 1999, 2000, 2001, and 2004. For 2002, the T2FNS obtains the lowest RMSE values and for 2003, LMNF-D/I is the best performing model.

TABLE II
RMSE VALUES FOR TAIEX FORECASTING

Methods	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
U_R model	164.00	420.00	1070.00	116.00	329.00	146.00
U_NN model	107.00	309.00	259.00	78.00	57.00	60.00
U_NN_FTS model	109.00	255.00	130.00	84.00	56.00	116.00
U_NN_FTS_S model	109.00	152.00	130.00	84.00	56.00	116.00
B_R model	103.00	154.00	120.00	77.00	54.00	85.00
B_NN model	112.00	274.00	131.00	69.00	52.00	61.00
B_NN_FTS model	108.00	259.00	133.00	85.00	58.00	67.00
B_NN_FTS_S model	112.00	131.00	130.00	80.00	58.00	67.00
Cai *et al.*'s method	102.22	131.53	112.59	60.33	51.54	50.33
CNFS-ARIMA	100.01	122.58	115.82	64.34	57.69	55.56
LMNF-D/I	92.19	123.33	116.73	63.66	**50.90**	53.63
T2NFS	97.30	120.90	103.84	**58.10**	51.00	51.73
Jiang's *et al.*'s model	97.61	119.73	113.26	67.39	54.95	56.21
ADP-T2FTS	**91.37**	**118.81**	**100.91**	65.91	57.81	**45.88**

Results showed that ADP is entirely suitable for studying the universe of discourse partitioning, besides being an autonomous partitioning method that does not depend on user interference. In the experiments ADP-T2FTS outperformed type-1 FTS and other advanced forecasting models, also showing that type-2 fuzzy logic may be able to model uncertainty better and minimize its influence in the forecasting process and model design, hence providing higher accuracy.

V. THE ADP-T2LIMG APPROACH

The previous models proposed so far, SODA-T2FTS and especially ADP-T2FTS already present advancements in FTS model design, but the data-driven algorithm method could still be more explored, as only the number of data clouds identified by SODA and ADP was employed to partition the UoD and generate the fuzzy sets. Essential information on shape, endpoints, minimum and maximum values, and other characteristics of the data clouds remained unused, leaving valuable information untapped. To bridge this gap, an updated version of the previous model, termed ADP-T2LIMG, is introduced. This new model leverages the parameters from ADP's output, utilizing this enriched information to more accurately design and define the fuzzy sets in the forecasting model, aligning them more closely with the data distribution. ADP-T2LIMG's algorithm is detailed in Appendix II.

ADP-T2LIMG is compared to the other two proposed models in this research using the TAIEX and SONDA time series. Results shown in Table III highlight ADP-T2LIMG's performance compared to the other two methods, also showing ADP-T2LIMG's longer execution time.

TABLE III
RMSE VALUES FOR TAIEX AND SONDA DATASETS.

Dataset	Model	Gridsize	Order	Part.	RMSE	Time(s)
TAIEX	ADP-T2FTS	4	1	615	62.63	43.19
	SODA-T2FTS	10	1	50	79.53	85.4
	ADP-T2LIMG	5	1	232	**59.86**	580.16
SONDA	ADP-T2FTS	5	1	132	104.23	132.07
	SODA-T2FTS	3	1	6	140.11	56.03
	ADP-T2LIMG	3	1	543	**99.76**	1112.87

Five univariate time series from the PJM hourly energy consumption and the Global Energy Forecasting Competition 2012 (GEFCom 2012) datasets (Appendix III) are then used for evaluation. Table IV showcases the forecasting accuracy of the proposed ADP-T2LIMG model alongside several baseline models detailed in 14.

TABLE IV
MODEL PERFORMANCE COMPARISON (RMSE)

Method	Zone 1	Zone 2	Zone 3	DEOK	AEP
R-HFCM	716.36	5531.02	5837.61	136.78	624.15
PWFTS	939.67	6826.44	7359.91	134.86	662.84
LSTM	4932.35	32610.37	36592.80	2862.96	679.39
CNN	611.85	4685.84	5193.89	94.66	377.45
CNN-LSTM	650.44	5847.70	5935.38	95.02	421.74
ARIMA	5274.91	42207.76	39253.07	603.27	3005.83
ADP-T2LIMG	**487.466**	**3654.04**	**4056.28**	**71.86**	**316.44**

VI. CONCLUSIONS

This research introduced three novel forecasting models — SODA-T2FTS, ADP-T2FTS, and ADP-T2LIMG — each designed to improve fuzzy set based time series forecasting through advanced autonomous data partitioning algorithms and type-2 fuzzy logic. These models, especially ADP-T2LIMG, enhance traditional forecasting approaches by leveraging data-driven partitioning techniques, reducing reliance on predefined parameters, and improving the handling of uncertainty in time series data and model design. As a direct outcome of this research, an open-source Python library, pyT2FTS (https://github.com/arthurcaio92/pyT2FTS), was developed to provide a flexible and accessible framework for fuzzy set based time series forecasting.

REFERENCES

[1] P. Singh, "A brief review of modeling approaches based on fuzzy time series," *International Journal of Machine Learning and Cybernetics*, vol. 8, pp. 397–420, 2017.

[2] L. A. Zadeh, "Fuzzy sets," in *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*. World Scientific, 1996, pp. 394–432.

[3] Q. Song and B. S. Chissom, "Fuzzy time series and its models," *Fuzzy sets and systems*, vol. 54, no. 3, pp. 269–277, 1993.

[4] S.-M. Chen *et al.*, "Forecasting enrollments based on fuzzy time series," *Fuzzy sets and systems*, vol. 81, no. 3, pp. 311–319, 1996.

[5] M. Bose and K. Mali, "Designing fuzzy time series forecasting models: A survey," *International Journal of Approximate Reasoning*, vol. 111, pp. 78–99, 2019.

[6] R. Gao and O. Duru, "Parsimonious fuzzy time series modelling," *Expert Systems with Applications*, vol. 156, p. 113447, 2020.

[7] K. Huarng, "Effective lengths of intervals to improve forecasting in fuzzy time series," *Fuzzy sets and systems*, vol. 123, no. 3, pp. 387–394, 2001.

[8] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning-iii," *Information sciences*, vol. 9, no. 1, pp. 43–80, 1975.

[9] A. C. Vargas Pinto, P. C. Silva, F. G. Guimarães, C. Wagner, and E. P. de Aguiar, "Self-organised direction aware data partitioning for type-2 fuzzy time series prediction," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.

[10] Abhishekh, S. S. Gautam, and S. Singh, "A refined weighted method for forecasting based on type 2 fuzzy time series," *International Journal of Modelling and Simulation*, vol. 38, no. 3, pp. 180–188, 2018.

[11] X. Gu, P. P. Angelov, D. Kangin, and J. C. Principe, "Self-organised direction aware data partitioning algorithm," *Information Sciences*, vol. 423, pp. 80–95, 2018.

[12] X. Gu, P. P. Angelov, and J. C. Príncipe, "A method for autonomous data partitioning," *Information Sciences*, vol. 460, pp. 65–82, 2018.

[13] J.-A. Jiang, C.-H. Syue, C.-H. Wang, J.-C. Wang, and J.-S. Shieh, "An interval type-2 fuzzy logic system for stock index forecasting based on fuzzy time series and a fuzzy logical relationship map," *IEEE Access*, vol. 6, pp. 69 107–69 119, 2018.

[14] O. Orang, F. J. Erazo-Costa, P. C. L. Silva, G. de Alencar Barreto, and F. G. Guimarães, "A large reservoir computing forecasting method based on randomized fuzzy cognitive maps," in *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2024, pp. 1–8.

Appendix I

Fuzzy Set Based Time Series Traiining and Forecasting Procedures are detailed below:

1) Data transformation: Data is differentiated to make the time series stationary or reduce non-stationary effects.

2) Definition of the Universe of Discourse - UoD: First, define the UoD as $UoD = [lower\ bound, upper\ bound]$, IN WHICH *lower bound* and *upper bound* are defined using the minimum ($min(Y(t))$) and maximum ($max(Y(t))$) values observed in the historical time series data, with an additional margin added to accommodate potential fluctuations, usually defined to extend from 10% to 20% the original UoD.

3) UoD partitioning: The UoD is partitioned into k intervals and to each interval a fuzzy set $\tilde{A}_i$ $(i = 0, 1, ..., k)$ is assigned.

4) Fuzzification of time series data: Each interval is associated with a fuzzy set represented by its membership function. The membership function of each fuzzy set assigns a degree of membership to every value in the UoD. The entire training set from the historical time series data is converted into fuzzy values using the defined fuzzy sets in the UoD.

5) Extraction of temporal patterns: Once the training set is fuzzified, temporal relationships between consecutive fuzzy values are analyzed to establish Fuzzy Logical Relationships (FLRs). For instance, if a data point at time $t-1$ is fuzzified to the fuzzy set $\widetilde{A}_4$ and another data point at time t is fuzzified to $\widetilde{A}_7$, the following logical relationship is formed: $\widetilde{A}_4 \rightarrow \widetilde{A}_7$.

After the training procedure is over, the forecasting procedure consists of:

6) Fuzzification: All data points $y(t) \in Y, t-1, ..., n$ from the time series test dataset are fuzzified and their fuzzy values are calculated in respect to the linguistic variable $\widetilde{A}$.

7) Finding compatible rules: After all test samples are fuzzified, their corresponding fuzzy logical relationship group (FLRG) can be retrieved based on their fuzzy state. An FLRG contains all possible future states for a given current state of the test sample, learned from the training historical data during the training phase.

8) Defuzzification: The FLRGs are used to compute the forecast fuzzy set. After that, this IT2FS is reduced to a T1FS in the type-reduction process and then the final crisp forecast value is obtained after the defuzzification process.

Appendix II

Complete algorithm for ADP-T2LIMG:

Algorithm 1: ADP-T2LIMG training and forecasting method

while *Not at the end of time series data* Y **do**
  Extract a segment from Y according to a window size W;
  Split time series data window into training and test sets;
  for *Training data* Y_{trg} **do**
    Compute the UoD;
    Use ADP to partition Y_{trg} into dc distinct data clouds;
    for *Each data cloud* **do**
      Create an UoD_{sub} using the minimum and maximum values of the data cloud;
      Split this UoD_{sub} into $sub_{dc} = dc/2$ partitions;
      Assign an IT2FS $\tilde{A}_{\text{i}}$ $(i = 1, 2, ...dc)$ to each partition ;
    end
    Match all UoD_{sub} and their respective fuzzy sets into the original UoD;
    Filter the fuzzy sets and remove all that are fully overlapped by another set;
    Fuzzify each instance from Y_{trg} by computing membership values across all fuzzy sets;
    Detect temporal patterns (FLRs) between consecutive instances from Y_{trg} ;
    Organize FLRs into FLRGs;
    Build the model $\mathcal{M}$;
  end
  for *The test data* Y_{test} **do**
    Fuzzify each instance $y(t)$ from Y_{test} by computing membership values across all fuzzy sets;
    for *FLRG* $\in \mathcal{M}$ **do**
      if $f(t)$ *matches the LHS* **then**
        rules_matched $\leftarrow$ FLRG;
      end
    end
    Use *rules_matched* in the fuzzy inference procedure and compute $f(t+1)$;
    Obtain $y(t+1)$ after type-reduction and defuzzification;
    Compute and save RMSE metrics;
  end
  Slide the window across Y by D instances;
end

APPENDIX III

Links to datasets:

- National Association of Securities Dealers Automated Quotations - NASDAQ:
  https://github.com/arthurcaio92/pyT2FTS/raw/main/data/NASDAQ.zip

- Standard & Poor's 500 - S&P500:
  https://github.com/arthurcaio92/pyT2FTS/raw/main/data/SP500.zip

- Taiwan Stock Exchange Capitalization Weighted Stock Index - TAIEX:
  https://github.com/arthurcaio92/pyT2FTS/raw/main/data/TAIEX.zip

- Sistema de Organização Nacional de Dados Ambientais - SONDA:
  http://sonda.ccst.inpe.br/

- PJM hourly energy consumption - AEP e DEOK:
  https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption

- Global Energy Forecasting Competition 2012 - GEFCom 2012 Zonas 1, 2 e 3:
  https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting