Trends in Computational and Applied Mathematics, **23**, N. 2 (2022), 213-222
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029
www.scielo.br/tcam
doi: 10.5540/tcam.2022.023.02.00213

Universal Approximators: New Approach for the Curve Fit of the COVID-19 Infected Population

A. M. A. BERTONE[1*], J. B. MARTINS[2] and R. S. M. JAFELICE[3]

Received on December 10, 2020 / Accepted on September 6, 2021

ABSTRACT. Fuzzy systems that include Takagi-Sugeno inference method with linear outputs, are widely known to have the ability to uniformly approximate any polynomial with high precision and, as a consequence, any continuous function, by applying the approximation theorem of Weierstrass. There is one more advantage for these methods which is to obtain an explicit expression of the defuzzified output as a function of the system's inputs. The purpose of this study is to describe the dynamics of a data set collected through the behavior of the tangential envelope and the local concavity of a curve to be adjusted. The functions that define the envelope and its concavity are identified by means of a hybrid system that combines a fuzzy clustering with the qualities of the Takagi-Sugeno inference method. The analyzed data set represents the world population of confirmed infected people by the infectious disease caused by the coronavirus of severe acute respiratory syndrome, named COVID-19. The proposed fuzzy method, in two versions, first and second order, are compared with curves built through the least square method with the maximum of the absolute value of the difference between the fit values and the data, normalized at each instant. In these comparisons, both fuzzy approaches proposed in this study are the ones that best match the data collected, being the fuzzy approximation of second order the best of all.

Keywords: mathematical modeling, fuzzy clustering, Takagi-Sugeno inference method, regression analysis.

1 INTRODUCTION

The universal approach is the basis for theoretical research and practical applications of fuzzy systems, particularly for those using the Takagi-Sugeno (TS) inference method. There is currently a line of research in this direction that includes the most diversity investigations on the accuracy of TS fuzzy systems [1, 2, 6, 9].

*Corresponding author: Ana Maria Amarillo Bertone – E-mail: amabertone@ufu.br
[1]Faculdade de Matemática (FAMAT), Universidade Federal de Uberlândia, Av. João Naves de Ávila, 2121, 38408-100, Uberlândia, MG, Brazil – E-mail: amabertone@ufu.br https://orcid.org/ 0000-0003-4370-9506
[2]Departamento de Computação, Instituto Federal do Triângulo Mineiro, Rua João Batista Ribeiro, 4000, 38064-790, Uberaba, MG, Brazil – E-mail: jefferson@iftm.edu.br http://orcid.org/0000-0001-6804-9802
[3]Faculdade de Matemática (FAMAT), Universidade Federal de Uberlândia, Av. João Naves de Ávila, 2121, 38408-100, Uberlândia, MG, Brazil – E-mail: rmotta@ufu.br https://orcid.org/0000-0001-8489-3974

The proposed method in this work is based on a fuzzy identification technique whose process has two components: fuzzy clustering and a TS fuzzy inference [8]. In fact, a data set collected is organized in inputs and output for the first stage of the method, consisting in a clustering of the data set by a fuzzy similarity. The Gustafson and Kessel algorithm is chosen for this purpose [4]. This stage of the process determines the fuzzy sets that are the antecedents of the TS fuzzy inference. This second stage provides a defuzzification output which is the explicit function that fits the data set collected. This data set comes with two inputs, time t, that represents one day, and $x(t)$, representing the number of infected people by a disease. The data set is reorganized adding one output calculated as the first order finite variation of $x(t)$. Then, the fuzzy method is applied to obtain the envelope fit curve. In the case of the concavity fit curve, the inputs are the same, adding an output which is the calculation of the second order finite variation. In short, the advanced and centralized finite difference methods [3] are used to organize the data set for the fuzzy identification.

The pandemic known as COVID-19: CO for coronavirus, VI for virus, D for disease, and 19 for the year it was discovered, has had an outbreak unleashed in China on December 31, 2019. The data have being recorded daily on the Worldometers' portal [10]. The data set collected for this study provides confirmed numbers referring to the population of infected people by the coronavirus, is used to obtain a curve that interprets the dynamics of the disease's infection spread. Samples for this investigation have been collected from January 22 until November 15 of 2020.

In order to measure the precision of the fuzzy identification methodology, the result obtained by the proposed method is compared with the fit curve of polynomial, exponential, Gaussian, and power type, obtained through the least squares method [3]. Exponential and Gaussian curves are the most accepted to interpret the type of dynamics of the collected data from January to November of 2020 [7]. To measure the accuracy of the methods for their comparison, is used the maximum of the absolute value of the difference of the fit curve values and the data, both normalized at each instant.

This work is organized as follows: in Section 2 the theoretical foundations of the methodology are explained; in Section 3 the methodology is developed; in Section 4 the results are presented to conclude, in Section 5, with final considerations.

2 THEORETICAL BACKGROUND

In general, clustering is an unsupervised classification of data, resulting in groups of elements called clusters. The purpose of this technique is to organize the patterns represented by vectors or points in the multidimensional space, according to a mathematical measure of similarity. The fuzzy clustering allows the elements of the data to belong to all clusters simultaneously with different membership degrees. There are many techniques that refer to fuzzy clustering; this study uses the algorithm developed by Gustafson and Kessel (GK) [4]. At the beginning of the GK algorithm the following elements are considered:

- Euclidean distance as a measure of similarity;

- an initial set of clusters centers, chosen from the equally distributed data set;
- a specific number of clusters, m, and a tolerance, tol, as a stopping criterion.

The GK algorithm stands out for changing the geometry of the cluster in each iteration, interpreting the similarities of the clusters more precisely.

Let $z_i = (t_i, x(t_i), s(t_i)), i = 1, \ldots, N$ a given data set where $(t_i, x(t_i))$ are the inputs and $s(t_i)$ the output, that has been clustering. As a consequence of this process, four fundamental elements are obtained for the construction of the TS inference method, the next step of the methodology. These elements are:

- the centers of clusters, v_j, $j = 1, 2, \ldots, m$, that will be the centers of the Gaussian-type membership functions, corresponding to the antecedents of the TS inference method;
- μ_{ij}, entries of the membership matrix that contains the membership degree of the element $z_i = (t_i, x(t_i), s(t_i)), i = 1, \ldots, N$ to each cluster;
- the projections of the highest levels of memberships of each cluster over the t axis, that is, the sets given by

$$A_j = \left\{ t_i, \mu_{ij}(t_i, x(t_i), s(t_i))) \geq \alpha \right\}, \ j = 1, \ldots, m,$$

  where α is determined in the interval $[0.5, \ 1[$ by an optimization process in order to approximate the points of the projected cluster to a Gaussian function [5];
- the standard deviation of the Gaussian membership function which is given by $\sigma_j = \beta(\max A_j - \min A_j)$, where β is determined by the same optimization process as aforementioned.

These four elements are essential for building the antecedents of the fuzzy TS inference method. The consequences of this inference are obtained by means of a local multivariate regression of the data outputs. The fuzzy rules are then established by:

Rule^j: If t is A_j then $D^j(t) = \theta_{j0} + \theta_{j1} t + \theta_{j2} x(t)$.

The final step is the defuzzification of the TS inference method, which is a weighted average of the product of the membership degrees of the inference's output values. This final step provides an explicit expression $\delta(t)$ that relates data inputs with the data outputs, giving by:

$$\delta(t) = \sum_{j=1}^{m} \left(\frac{\exp\left(-\frac{(t-v_j)^2}{2\sigma_j^2} \right) \left(\theta_{j0} + \theta_{j1} t + \theta_{j2} x(t) \right)}{\sum_{j=1}^{m} \exp\left(-\frac{(t-v_j)^2}{2\sigma_j^2} \right)} \right). \tag{2.1}$$

3 CURVE FITTING THROUGH FUZZY IDENTIFICATION METHODOLOGY

The Worldometer data set [10], $\{t_i, x(t_i)\}$ $i = 1,2,\ldots,N$, $N = 300$ represent the total number of days from January 22 to November 15, and $x(t_i)$ the number of confirmed infected individuals in the world at time t_i. Such data are organized in two matrix with three columns each, being $D_f = [t_i, x(t_i), \Delta_f(x(t_i))]$ and $D_c = [t_i, x(t_i), \Delta_c(x(t_i))]$, $i = 1,2,\ldots,N$, where:

- t_i is the first column of each matrix. The data is provided every day, therefore t_i represents one day.
- $x(t_i)$ is the second column of both matrix representing the infected number of people in day t_i.
- $\Delta_f(x(t_i)) = x(t_{i+1}) - x(t_i)$ is the third column of the matrix D_f, representing the day variation for a unitary step. The last line is filled with the same value as the penultimate line. The reason for repeating this value is that the difference between data set's consecutive entries can be considered minimal for one day of the disease's spread. In summary, it has been applied the advanced finite difference method [3].
- $\Delta_c(x(t_i)) = x(t_{i+1}) - 2x(t_i) + x(t_{i-1})$, is the third column of matrix D_c representing the variation of the variation. For the same reasons as aforementioned, the first line is filled with the result of $i = 2$ and the last line is filled with the same value as the penultimate line. The method is the centralized finite difference of order two [3].

Using the fuzzy identification method [8], the following parameters are obtained for each data:

- D_f: $m = 6$ for the number of clusters, $tol = 0.005$, $\alpha = 0.93$ and $\beta = 0.37$.
- D_c: $m = 6$ for the number of clusters, $tol = 0.6$, $\alpha = 0.5$ and $\beta = 0.85$.

As a result, are determined two functions, denoted by $\delta_f(t) = \delta_f(t, x(t))$ and $\delta_c(t) = \delta_c(t, x(t))$, through the formula of Equation (2.1). The function δ_f represents the values of the tangential envelope of the fit curve to be determined. The parameters of function δ_f and δ_c are detailed in Table 1 and Table 2, respectively.

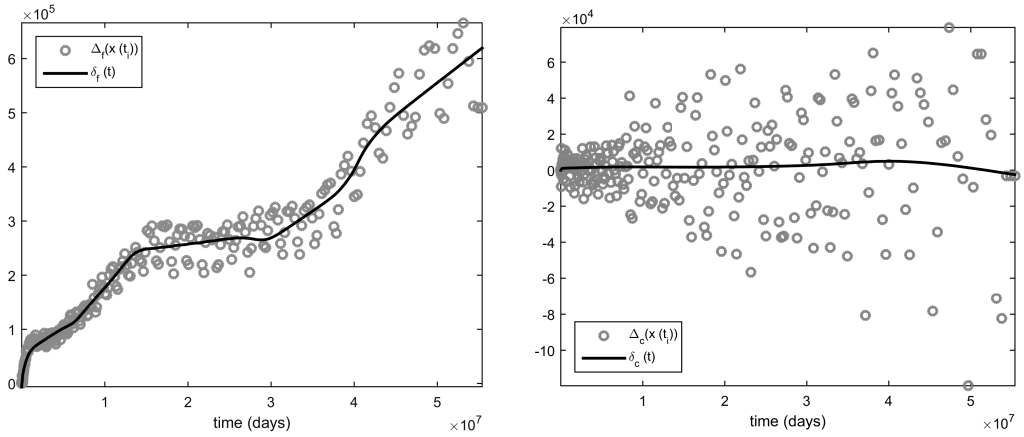
The function δ_f that fits data $D_f = (t_i, x(t_i), \Delta_f(x(t_i)))$ is shown in the Figure 1 (a) along with data and, in Figure 1 (b), is shown the data $D_c = (t_i, x(t_i), \Delta_c(x(t_i)))$ and the graph of δ_c.

Table 1: Elements of the δ_f function of the Equation (2.1).

Cluster	v_j^f	σ_j^f	θ_{j2}^f	θ_{j1}^f	θ_{j0}^f
1	31.47	21.84	332.48	0.03	-6348.11
2	97.15	19.99	857.70	0	14.20
3	153.24	10.74	118.73	0.01	1.21
4	203.95	15.55	1754.03	0	14.21
5	253.46	9.63	-317.38	0.01	-2.20
6	287.14	8.52	-189.62	0.01	-0.97

Table 2: Elements of the δ_c function of the Equation (2.1).

Cluster	v_j^c	σ_j^c	θ_{j2}^c	θ_{j1}^c	θ_{j0}^c
1	38.58	65.46	19.11	0	-5.56
2	103.84	50.16	17.67	-9.36	-42.14
3	152.60	35.71	23.51	0	0.23
4	202.38	44.21	7.82	1.13	0.06
5	252.29	36.56	4.83	5.30	0.03
6	291.82	20.41	247.55	0	1.36



(a) (b)

Figure 1: The function $\delta_f(t)$ and $\delta_c(t)$.

In the following, are detailed the steps of the algorithm for the calculation of the formula that approximates the values of $x(t_i)$, formula that includes the functions $\delta_f(t)$ and $\delta_c(t)$. In fact, assume that the curve, $g(t)$ that is going to fit the data is smooth enough, more precisely $\mathscr{C}^\infty(]\sqcup_\infty - \varepsilon, \sqcup_\backslash + \varepsilon[)$, $\varepsilon > 0$ small and arbitrary, and define $g(t_1) = x(t_1)$, $x(t_1)$ the first data value corresponding to the number of infected individuals at $t = 1$. The following steps are taken to reach the explicit formula for the first order approximation $g(t)$:

Step 1: Consider first Taylor's polynomial of $g(t)$ at point t_1, denoted by $p(t)$ and with the expression $p(t) = g(t_1) + g'(t_1)(t - t_1)$. Thus, since $\delta_f(t)$ is the curve representing $g(t)$ derivative, an approximation for the value $x(t_2)$ can be calculated as follows:

$$p(t_2) = g(t_1) + g'(t_1)(t_2 - t_1) \approx x(t_1) + \delta_f(t_1)(t_2 - t_1) = x(t_1) + \delta_f(t_1), \tag{3.1}$$

recalling that $t_{i+1} - t_i = 1$ for all $i = 1, \ldots, N$ corresponding to the collected data set. Therefore:

$$x(t_2) \approx x(t_1) + \delta_f(t_1). \tag{3.2}$$

Step 2: An approximation for $x(t_3)$ is calculated the same manner, using the first degree of Taylor's polynomial of $g(t)$ at t_2 to obtain:

$$x(t_3) \approx x(t_2) + \delta_f(t_2). \tag{3.3}$$

Thus, replacing (3.2) in (3.3), the approximate value for $x(t_3)$ follows the formula:

$$x(t_3) \approx x(t_1) + \delta_f(t_1) + \delta_f(t_2).$$

Step k: An inductive process is carried out on the following observation moments to conclude:

$$x(t_k) \approx x(t_1) + \sum_{i=1}^{k-1} \delta_f(t_i), \quad k = 2, \ldots, N. \tag{3.4}$$

It is noteworthy that the formula of Equation (3.4) can be extended for the calculation of intermediate values or for prediction purposes. Indeed, for instance, given $t \in]t_{k-1}, t_k[$, that value for $g(t)$ is calculated as

$$g(t) = x(t_1) + \sum_{i=1}^{k-1} \delta_f(t_i) + \delta_f(t)(t - t_{k-1}). \tag{3.5}$$

Following the same reasoning applied to the second degree polynomial of Taylor, another approximation for $x(t_k)$ is obtained as being:

$$x(t_k) \approx x(t_1) + \sum_{i=1}^{k-1} (\delta_f(t_i) + \frac{1}{2}\delta_c(t_i)), \quad k = 2, \ldots, N. \tag{3.6}$$

As aforementioned the formula (3.6) can be extended to other intermediate points $t \in]t_{k-1}, t_k[$ by a similar manner as (3.5), namely

$$g^2(t) = x(t_1) + \sum_{i=1}^{k-1} \delta_f(t_i)(t - t_{k-1}) + \frac{\delta_c(t)}{2}(t - t_{k-1})^2, \tag{3.7}$$

where $g^2(t)$ is the fuzzy approximation of second order for data collected.

The flowchart of the whole procedure is described in Figure 2.

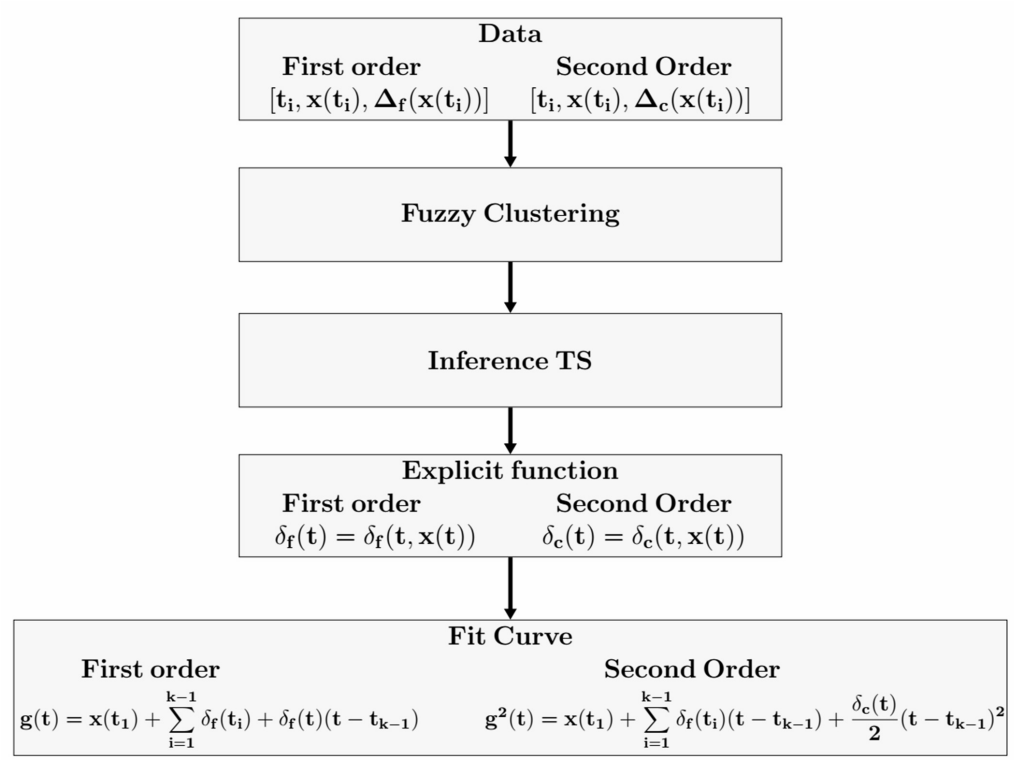


Figure 2: Flowchart of the fuzzy fit curves methodology.

Another use of formulas (3.4) - (3.6) and (3.5) - (3.7) is to predict future values or estimate retarded values of the data collected. This investigation is in progress. Furthermore, by knowing the smoothness of the curve candidate to fit the data observed, the same formulas can be generalized for the r-degree of smoothness.

4 RESULTS

The methodology described in Section 3 is applied to the data set of the confirmed numbers of infected population from COVID-19, as well as the curves obtained through the following types of regression methods:

1. a polynomial fit of degree 4 given by the function

$$a_1(t) = -0.0004776t^4 + 0.94t^3 + 424.9t^2 - 2.372 \cdot 10^4 t + 2.575 \cdot 10^5;$$

2. an exponential fit given by the function

$$a_2(t) = 1.525 \cdot 10^6 \exp(0.01218t);$$

3. a Gaussian fit given by the function

$$a_3(t) = 6.655 \cdot 10^7 \exp(-((t - 384.3)/164.6)^2);$$

4. a power fit given by the function

$$a_4(t) = 28.67t^{2.531}.$$

These curves a_l, $l = 1,2,3,4$ are obtained using the least squares method [3] for the data collected.

To compare the result of the proposed fuzzy curve fit method with the curves a_l, $l = 1,2,3,4$, an accuracy measure is chosen defined by

$$Error = \max_{t_i} \left\{ \left| \frac{x(t_i)}{M_I} - \frac{a_l(t_i)}{M_I} \right| \right\}, \tag{4.1}$$

where

$$M_I = \max_{t_i} \{x(t_i)\} = 55,388,299,$$

that is, the values $x(t_i)$ and $a_l(t_i)$ are normalized. The comparison result is shown in Table 3. It should be noted that the lowest value of the $Error$ of Equation (4.1) corresponds to the fuzzy curve fit methodology of second order that has been exposed in this work. This best goodness of fit is followed by the curve constructed through the same methodology, being of the first order.

In Figure 3 is shown the graphs of the curves a_l, $l = 1,2,3,4$ along with the graph of the functions of first order and second order obtained by the new approach developed in this research.

It is important to notice that the complete procedure of clustering and fuzzy inference, in both cases (first and second order approximation) in a computer Intel I7 processor, 60GB of RAM, 256 GB SSD memory drive, takes between 0.009085 and 0.015856 seconds. That means, no computational effort for a standard computer.

Table 3: Comparison of the goodness of fit resulting from the measure (4.1) under the different methodologies.

Metodology	$Error$
Polynomial Fit: $a_1(t)$	0.045034
Exponential Fit: $a_2(t)$	0.063565
Gaussian Fit: $a_3(t)$	0.075695
Power Fit: $a_4(t)$	0.03705
Fuzzy Identification First Order Fit: $g(t)$	0.0066691
Fuzzy Identification Second Order Fit: $g^2(t)$	0.0033225

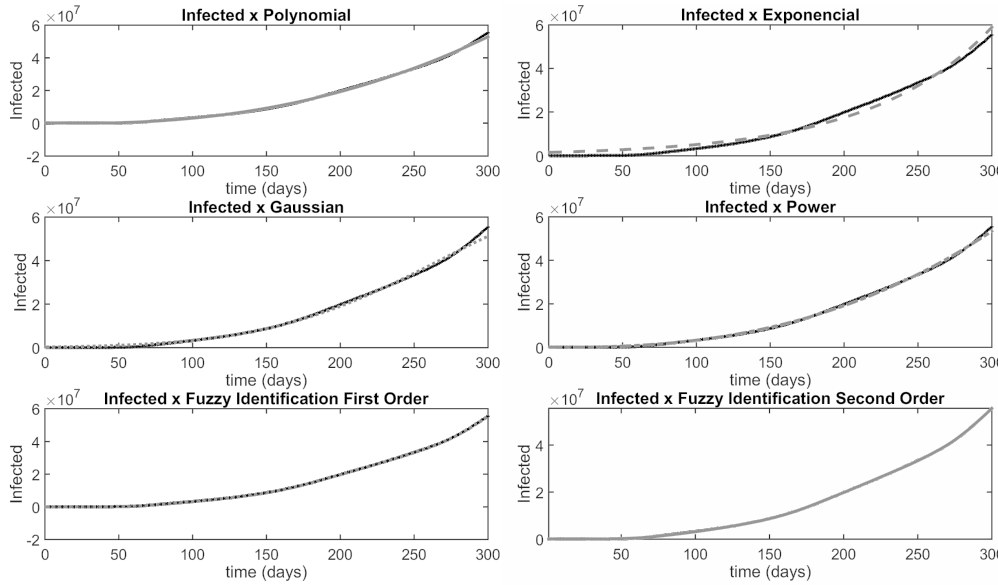


Figure 3: Comparison of the methodologies with the data collected.

5 CONCLUSION

A new fuzzy approach for fit curve method is presented with promising results compared to the classic least squares method, in four types of curves. The comparison is made through data corresponding to confirmed numbers of infected individuals from COVID-19 in the world, using a chosen accuracy measure. Two types of fuzzy approximations are provided that represent the tangential envelope and the local concavity of the curve that would fit the data collected. The smallest errors obtained from the comparative study are given for the two curves built through the fuzzy identification system. The result is due to the intrinsic characteristics of the fuzzy methodology in which the inference method applied is the Takagi-Sugeno. These systems are proven to be universal approximators, in the sense that the methodology is able to approach continuous functions with high precision, in addition to provide an explicit continuous function

as a defuzzified output. With respect to future work, is in process to use this methodology as a tool for prediction modeling.

REFERENCES

[1] K. Bart. Fuzzy Systems as Universal Approximators. *Trans. on Computers*, **43** (1994), 1153–1162.

[2] J.J. Buckey. Sugeno type controllers are universal controllers. *Fuzzy Sets Systems*, **53**(3) (1993), 299–303.

[3] R.L. Burden & J.D. Faires. "Numerical Analysis". Brooks/Cole Cengage Learning, United States, 9 ed. (2011).

[4] D.E. Gustafson & W. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In "Proceedings of the IEEE Control and Decision Conference" (1979), p. 761–766.

[5] R.M. Jafelice & A.M.A. Bertone. "Biological Models via Interval Type-2 Fuzzy Sets". Springer Briefs in Mathematics (2020).

[6] J. Kim, K. Koo & J. Lee. Monotonic fuzzy systems as universal approximators for monotonic functions. *Intell. Autom. Soft Comput.*, **18**(1) (2012), 13–31.

[7] E.Z. Martinez, D.C. Aragon & A.A. Nunes. Short-term forecasting of daily COVID-19 cases in Brazil by using the Holt's model. *Revista da Sociedade Brasileira de Medicina Tropical*, **53** (2020). URL `https://preprints.scielo.org/index.php/scielo/preprint/view/667`.

[8] J.B. Martins, A.M.A. Bertone & K. Yamanaka. Novel Fuzzy System Identification: Comparative Study and Application for Data Forecasting. *IEEE Latin America Transactions*, **17** (2019), 1793–1799.

[9] L. Wang & J. Mendel. Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. *IEEE Trans. Neural Netw.*, **3**(5) (1992), 807–814.

[10] Worldometer (2020). URL `https://www.worldometers.info/coronavirus`.