JAISCR, 2011, Vol.1, No.4, pp. 283–300

ON CHOOSING THE FUZZINESS PARAMETER FOR IDENTIFYING TS MODELS WITH MULTIDIMENSIONAL MEMBERSHIP FUNCTIONS

Andreas Kroll

Measurement and Control Department, Mechanical Engineering,
University of Kassel Mnchebergstrasse 7, D-34125 Kassel,
andreas.kroll@mrt.uni-kassel.de

Abstract

Fuzzy clustering is a well-established method for identifying the structure/fuzzy partitioning of Takagi-Sugeno (TS) fuzzy models. The clustering algorithms require choosing the fuzziness parameter *m*. Prior work in the area of pattern recognition shows, that a suitable choice of *m* is application- dependent. Yet, the default of *m*=2 is commonly chosen. This paper examines the suitable choice of *m* for identifying TS models. The focus is on models that use the classifiers resulting from fuzzy clustering as multi-dimensional membership functions or their projection and approximation. At first, the differentiability and grouping properties of the fuzzy classifiers are analyzed to make a general recommendation of choosing $m\in(1;3)$. Besides, the effect of the cluster number *c* on the classification fuzziness is examined. Finally, requirements that are specific to TS modeling are introduced, which narrow down the suitable range for *m*. Building on algorithm analysis and four case studies (function approximation, a vehicle engine and an axial compressor application for nonlinear regression), it is demonstrated that choosing $m\in(1;1.3)$ for local and $m\in(1;1.5)$ for global estimation will typically provide for good results.

1 Introduction

This paper examines the choice of the uncertainty parameter in clustering-based Takagi-Sugeno (TS) fuzzy modeling. Theoretical constraints and experience-based narrower bounds will be presented.

Takagi-Sugeno (TS) fuzzy modeling [1] can be used to provide quantitative nonlinear models for simulation, prediction, control, diagnostics and other applications. TS models can be derived from a physical model by the sector nonlinearity method [2],[3] or by Taylor series expansion in several points and suitable partitioning [4],[5]. However, physical models are not always available and their derivation can be costly. Alternatively, TS models can be identified from input/output data in case suitable observations of the system behavior are available. Simple grid-based partitioning often results in a large amount of rules due to the curse of dimensionality. An alternative is to use a greedy growth strategy such as the axis-parallel dividing LOLIMOT algorithm [6]. The advantage is its simplicity and the prediction- quality-guided division process. The drawback is that it will not produce parsimonious models in case the nonlinearities are not axis-aligned. A non-axis-aligned dividing version of the algorithm has been proposed, recently [7].

Alternatively, fuzzy clustering in the input, output or product space can be used for identifying the fuzzy partitions. The resulting (truly) multi-dimensional partitioning (in case of input or product space clustering) permits to develop parsimonious model representations with excellent approximation capability [8],[9]. The major disadvantage is that this results in difficulties to interpret models. For this reason, the membership functions (MF) are often projected to the coordinate axis and approximated by scalar functions such as trapezoids or Gaussians. This provides for better interpretability at the cost of inducing approximation errors.

Fuzzy-*c*-means (FCM), Gustafson-Kessel (GK) and Gath-Geva (GG) algorithms are often used for clustering-based fuzzy identification. These algorithms require an a-priori specification of the fuzziness parameter (weighting exponent) *m* and the number of clusters *c*. For determining *c,* cluster validity indices [10-13] or cluster-merging algorithms [12],[14] can be used, and are widespread in use. The fuzziness parameter adjusts the degree of fuzziness of the clusters. In case of TS modeling, fuzzy clustering is in general carried out using m=2 without further consideration: Sin and de Figuero [15] and Laukonen et al. [16] for FCM, Babuska [11:p.58; 87] and Nelles [6: p.146] for FCM- and GK-, Setnes [17] for GK- and Abonyi et al. [18,19] for GG-based approaches.

It is known, that the resulting cluster centers depend little on *m* if the clusters have similar geometry and density [20]. However, the shape of the MF depends significantly on *m*, such that *m* can majorly affect the quality of the resulting fuzzy model - no matter whether the clustering MF are used as is or projected and approximated. If MF shape and location are succeedingly optimized, the commonly applied derivative-based optimization routines still rely on good initial values due to the risk of local convergence. This motivates to look for available methods on choosing *m*. In the area of fuzzy clustering and pattern recognition this problem has been addressed for grouping and general classification problems. A literature review shows that for this field of application no common theoretical basis exists. A wide range of values between 1+ and 101 is used. m=2 is often specified as appropriate (default) choice [10],[21-24]. However, it was recognized that this may not be suitable for all applications [21], [25]. The dependency of the fuzziness of the partitioning, not only on *m*, but also on the number of clusters [26],[27], makes the choice more difficult.

This contribution addresses choosing *m* in clustering-based TS system identification. It is shown that choosing $m\in(1;3)$ provides for continuously differentiable MF. This property transfers to the TS system, as the local models are continuously differentiable. Continuous differentiability simplifies the application of derivative-based optimization methods. These can be used for identifying nonlinear output error (NOE)-type fuzzy models or to tune the partitioning to reduce the prediction error of the model [28],[29]. Moreover it will be shown that $m\in(1;3)$ provides for membership functions with a flat top at the center, which is desired for partitioning application. For $m>3$ the function has a sharp peak at the center. Besides these general considerations, specific requirements for TS model identification can be considered to narrow down the suitable range for *m*: "Reactivation" effects [7],[22] (the membership declines with rising distance from a center, but then starts to incline again) are reduced by a small choice of *m*. The same is true for unintuitive interpolation effects in "V-type situations" [30] of the local models. As *m* adjusts the fuzziness, a small choice of *m* means little interaction between the local models. That permits to interpret the local models as local linearization of the original system. On the contrary, a larger value of *m* permits better approximation if global estimators are used. Such analysis will be used to derive a reduced range for *m*. Summarizing, the optimal choice of *m* depends on the application, and using a default of m=2 may not exploit the full potential of the modeling approach.

Table 1. Acronyms

Acronym	Meaning
FCM	Fuzzy-*c*-means (algorithm)
GE	Global (parameter) estimation
GG	Gath-Geva (algorithms)
GK	Gustafson-Kessel (algorithm)
LE	Local (model parameter) estimation
MF	Membership function
MSE	Mean squared error
NOE	Nonlinear output-error (model)
PWA	Piecewise affine (model/system)
TS	Takagi-Sugeno (model/system)
WLS	Weighted least squares (algorithm)

This paper is organized as following: The next section summarizes FCM and GK clustering algorithms. Section 3 introduces the basics of TS modeling. Section 4 surveys prior work on choosing

m. Section 5 addresses limiting behavior, grouping strategy, classifier differentiability and TS model specific requirements dependent on m. Constraints and guidelines for an appropriate choice of m are presented, which are the major results of this contribution. Results from numerical case studies on nonlinear function approximation and regression are presented in section 6. The final section summarizes the work. The appendix contains the proof for the differentiability theorem. Table I records the acronyms used in this paper.

2 Fuzzy-c-Means and Gustafson-Kessel Algorithm

Given a set X of N observations $x_k \in \Re^M, k = 1,...,N$, each covering M features, the FCM algorithm [10],[31] locates c cluster centers $v_i \in \Re^M, i = 1,...,c$ and determines all $c \times N$ memberships $\mu_i(x_k) := \mu_{i,k} \in [0;1]$ for the chosen fuzziness parameter $1 < m < \infty$ such that the cost function (weighted within groups sum of squared errors)

$$J_{\text{FCM}} = \sum_{k=1}^{N} \sum_{i=1}^{c} (\mu_{i,k})^m \|x_k - v_i\|_{A_i}^2 \tag{1}$$

is minimized subject to

$$\sum_{i=1}^{c} \mu_{i,k} = 1 \forall k \tag{2}$$

$$\sum_{k=1}^{N} \mu_{i,k} > 0 \forall i. \tag{3}$$

The purpose of the equality constraint (2) is to avoid the trivial solution $\mu_{i,k} = 0 \forall k, i$. The inequality constraint (3) prevents empty clusters. The optimization problem is solved iteratively by alternately solving two reduced problems: For fixed $\mu_{i,k}$, the optimal cluster centers

$$v_i = \frac{\sum_{k=1}^{N} (\mu_{i,k})^m \cdot x_k}{\sum_{k=1}^{N} (\mu_{i,k})^m} \forall i \in \{1,...,c\} \tag{4}$$

result from solving the corresponding unconstrained optimization problem. The optimal memberships

$$\mu_{i,k} = \left[\sum_{j=1}^{c} \left(\frac{||x_k - v_i||_{A_i}^2}{||x_k - v_j||_{A_j}^2} \right)^{\frac{1}{m-1}} \right]^{-1} \quad \forall i \in \{1,..,c\}, k \in \{1,...,N\} \tag{5}$$

are obtained by fixing the v_i and solving the corresponding constrained optimization problem due to (2). If x_k coincides with v_i full membership results, while coinciding with v_j means no membership to the i-th cluster. The FCM uses the same distance norm $|| \cdot ||_{A_r}$for all clusters. Typical choices are Euclidean, Mahalanobis or Minkowski/L$_p$ distance. The GK algorithm minimizes the same cost function as the FCM, but individually adapts the distance norm for each cluster: It uses an inner product norm, where the form matrices A_r are computed from the local fuzzy scatter (or fuzzy covariance) matrices. The prototypes v_i and the memberships $\mu_{i,k}$ are updated with the same formulae as the FCM (4), (5). The $c \times N$ memberships $\mu_{i,k}$ can be noted as a matrix $U = [\mu_{i,k}]$ (c-partitioning) and the cluster centers as $V = [v_r]$. For details on FCM and GK algorithm the reader is referred to e.g. [10],[11],[12].

3 Takagi-Sugeno Fuzzy Models and Identification

Takagi-Sugeno (TS) fuzzy models [1] offer a fuzzy-rule-based description for systems. A TS rule of an input/output model is given as: If $\mathbf{x}$ is $\mathbf{A}_i$ Then y_i=f$_i$($\mathbf{x}$). The model output y is inferred by taking the weighted average of the rule outputs y_i:

$$y(x) = \frac{\sum_{i=1}^{c} \alpha_i(x) \cdot y_i(x)}{\sum_{i=1}^{c} \alpha_i(x)} =: \sum_{i=1}^{c} \phi_i(x) \cdot y_i(x) \tag{6}$$

where c is the number of rules and ϕ_i the i-th fuzzy basis function (FBF). The degree of fulfillment/rule activation is simply the membership degree $\alpha_i = \mu_i$ [11]. TS models in state space representation follow correspondingly. Input/output models in discrete-time representation are used for system identification purposes in general. Each TS rule's antecedent defines the fuzzy validity region of the corresponding local model y_i. Strictly speaking, each local model is weighted by its FBF.

Data-driven modeling includes the major tasks of partition and conclusion identification. Commonly, multi-input-multi-output (MIMO) problems are treated as separate multi-input-single-output (MISO) problems: Nonlinear cause-effect relationships may be different for the outputs such that decomposition into MISO models enables parsimonious models. The local models are usually chosen as affine functions

$$\hat{y}_i(x) = [a_{1,i}; ...; a_{n,i}; a_{0,i}][x^T; 1]^T = \Theta_i^T \tilde{x}. \quad (7)$$

due to the better approximation capability than for linear ones. If the partitions have been determined, the structure of the conclusion function is selected. Then the parameters Θ_i of the local models are identified. This can be done separately for each local model by solving the problem

$$\hat{\Theta}_i : \arg\min_{\Theta i} \sum_{k=1}^{N} w_i(k) \cdot (y(k) - \hat{y}_i(k, \Theta_i))^2 \quad (8)$$

This is referred to as 'local estimation' (LE). The ordinary weighted least squares (WLS) method can be used and provides for the solution

$$\hat{\Theta}_i = [\Phi^T W_i \Phi]^{-1} \Phi^T W_i Y \quad (9)$$

with $\Phi = [\tilde{x}_k^T], W_i = diag[w_{i,k}], Y = [y_k]$ and $k = 1, ..., N$. Alternatively, the parameters of all local models can be computed simultaneously by solving the problem

$$\hat{\Theta} : \arg\min_{\Theta} \sum_{k=1}^{N} (y(k) - \hat{y}(k, \Theta))^2 ; \hat{\Theta} := \left[\hat{\Theta}_1; ...; \hat{\Theta}_c\right]. \quad (10)$$

This is referred to as 'global estimation' (GE). The WLS can be used to solve problem (10). Weighting strategies include using the memberships assigned by fuzzy clustering $w_{i,k} := \mu_{i,k}$. Alternatively, the center positions resulting from clustering can be kept, m be changed in (5) and the resulting memberships be used as weights. Typically, the clustering would be conducted with the same or a larger m than used for the TS model. Alternatively, binary memberships can be assigned by using α-level sets of the memberships resulting from clustering. Excluding data from the interpolation regime may improve the local estimation. More sophisticated methods for (dynamic) system identification are described e.g. in [32].

The choice of m for clustering affects the grouping philosophy: If a cluster should only comprise very similar data, m should be chosen larger: In the neighborhood of its center, the membership (5) to the cluster declines the faster the larger m is chosen. For $m \to \infty$, all memberships except for the centers approach $1/c$. Therefore, larger values of m result in small regions with high membership around each cluster center. A membership of $\approx 1/c$ to all clusters is assigned to data outside these regions. This means that the data is not significantly assigned to a cluster in the majority of the feature space. If clustering is used to partition X into c regions as e.g. required for TS modeling, regions of high membership should predominate. This is achieved by choosing m close to 1.

4 Prior Methods for Choosing the Fuzziness Parameter

This section surveys methods to choose m in fuzzy clustering that were identified in a literature review. Table II provides for a summary.

Bezdek et al. [33] suggest $m \in [1; 30]$ and in particular $m \in [1.5; 3]$ from experience with FCM.

The FCM cost functional decreases strictly monotone with m (evaluated at optimal pairs **U,v**), i.e. its derivative

$$\frac{\partial J(m, U, v)}{\partial m} = \sum_{k=1}^{N} \sum_{i=1}^{c} (\mu_{i,k})^m \log(\mu_{i,k}) \|x_k - v_i\|^2 \quad (11)$$

is negative for all $m > 1$, however the rate of change is not uniform [10, p. 73]. McBratney and Moore [34] suggest choosing the argument m that maximizes $(-\partial J/\partial m)\sqrt{c}$. Multiplication with $\sqrt{c}$ takes care of their expectation that m should be smaller for larger c. In the 2-dim. case studies with both well and not well separated clusters, a value of $m \in [1.9; 2.8]$ was determined to be useful.

Choe and Jordan [35] propose choosing m for the FCM by using fuzzy decision theory. Maximizing the number of data points within a cluster is defined as a fuzzy goal and minimizing the sum of squared errors within a cluster is defined as a fuzzy constraint. m is chosen as the argument that corresponds to the maximum of the membership function resulting from intersecting all fuzzy part-objectives. In a 2-dim. case study with two poorly separated classes m= 1.1...40 was tested and m=12 determined as optimum.

Pal and Bezdek [27] state that very high or low values of m may influence cluster validity indices that use memberships from the FCM. A usual range of $m \in (1; 5)$ is mentioned with m=2 being the most common choice. In the 4-dim. case studies with 3

Table 2. Overview On Methods to Choose Fuzziness

Ref.	Bounding of m	Assessed information	Cluster Algorithm	m tested in case studies	m determined case studies	Application context
[33]	IV[1.5;3]	EX	FCM	1.25. . . 2	-	General grouping
[34]	SV	CR	FCM	1.1. . . 3.5	1.9. . . 2.8	Classification
[35]	SV	CR	FCM	1.1. . . 40	12	Decision function estimation
[27]	IV[1.5;2.5]	EX	FCM	1.005. . . 7	-	General grouping
[36]	SV	RD	FCM	1. . . 7	2.92; 2.5	Mamdani model rule generation
[37]	SV	CR	FCM	≥ 1.5	5	Partitioning for TS model
[38]	IV[1.5;2.5]	CR	FCM	1.5. . . 3.3	-	General grouping
[39]	IV[1.6;3], SV	RD	SMFC	≥ 1.2	-	Classification (image)
[25]	UB, SV	RD	FCM	1. . . 1.2; 1.7; 2.5	1.112; 1.17; 1.25	Grouping/classification (gene expression)
[45]	SV	CR	FCM, MCV	1. . . 8	4.47	Mamdani model rule generation
[21]	UB	RD	FCM	-	1.6. . . 3.2	Grouping/classification
[26]	SV	CR	FCM	1.2. . . 2.5	1.67; 1.9	Mamdani model rule generation
[41]	SV	CR	FCM	1.05. . . 3	1.05	Grouping/classification (gene expression)
[42]	SV	CR	FCM	1.1. . . 3	1.1. . . 2	Classification (image)
[40]	UB	CR	FCM	1.02. . . 2	1.2. . . 1.55	Grouping/classification (gene expression)
[46]	SV	CR	FCM	-	1.41; 1.71; 101	Modeling/approximation
Here	IV (1; 3)	CD	FCM, GK, PCM, PGK	1. . . 10	1.05. . . 1.3/1.5	TS modeling/approximation

UB = Upper bound, IV = Interval, SV = Single value; EX = Experience, RD = Raw data, CR = Clustering result, CD = Classifier description; FCM = Fuzzy-c-Means, SMFC = Supervised Mahalanobis fuzzy classifier, GK = Gustafson-Kessel algorithm, PCM = Possibilistic FCM, PGK = Possibilistic GK, MCV = Minimum Volume Clustering and 4 not well separated classes $m \in [1.5;2.5]$ is determined as suitable choice. Emami et al. [36] used FCM and determine the trace of the scatter matrix

$$K = \text{trace}\left(\sum_{k=1}^{N}\left[(x_k - \bar{x}) \cdot (x_k - \bar{x})^T\right]\right); \bar{x} = \sum_{k=1}^{N} x_k. \quad (12)$$

They suggest choosing a value of m that corresponds to approximately $K/2$. Mamdani-type fuzzy modeling is addressed: In 2- and 3-dim. case studies with 8 and 6 clusters, they determined m=2.92 and m=2.5, respectively.

Chen and Wang [37] used FCM for determining partitions of TS models with Gaussian membership functions. They suggest increasing m in 0.1 increments starting from 1.5 until the standard deviations of the clusters (calculated from the fuzzy covariance matrices of the clusters) are large enough such that the resulting membership functions will have sufficient overlap. In 1- and 2-dim. case studies on function approximation with 2 and 4 clusters m=4 and m=5 are determined, respectively.

Gao et al. [38] propose choosing m such that both J_{FCM} and the partition entropy H_m are minimized:

$$m* = \arg_{m}(\max(\min(\mu_G(m), \mu_C(m))))\, with \quad (13)$$

$$\mu_G(m) = \exp(-\alpha \cdot J_{FCM}/\max_{m} J_{FCM}), e.g. \alpha = 1.5 \quad (14)$$

$$\mu_C(m) = (1 + \beta \cdot H_m/\max_{m} H_m)^{-1}, e.g. \beta = 10. \quad (15)$$

Requiring a minimum clustering tendency in the data and excluding cases with well separated clusters, they suggest $m \in [1.5;2.5]$. In two case studies $m* \in [1.5;3.3]$ and $m* \in [1.7;2.1]$, respectively, are determined. They conclude that the better the separability of classes is, the higher the value of m that should be chosen.

Deer and Eklund [39] require that the memberships of a pixel in an image should reflect the true proportions of the contributing classes in the pixels to determine a value for m. They conduct studies under a linear mixing assumption and derive values $m \in [1.3;3]$. Their studies refer to a FCM that uses a local Mahalanobis distance metric for each class.

Dembl and Kastner [25] analyze complex gene expression/microarray data and make the assumption that the cluster centers will be close to some genes. They studied the set

$$Y_m = \{(d^2(x_i;x_k))^{1/(m-1)}; k \neq i; k,i = 1;2;...;N\} \quad (16)$$

of distances between the data sets and observed that values of m that provide FCM memberships close to $1/c$ lead to a coefficient of variation of Y_m close to 0.03 dim($\mathbf{x}$). Starting from m=2, they numerically search for an upper bound m_{ub} for m such that

$$m_{\text{ub}} : cv\{Y_m\} = \sigma(Y_m)/\bar{Y}_m \approx 0.03 \cdot \dim(x) \quad (17)$$

holds. Secondly, they suggest choosing $1 < m \leq 2$ to obtain high memberships for data that are strongly related to clusters:

$$\begin{array}{l} m_{\text{ub}} \geq 10: \quad m := 2 \quad \text{and} \\ m_{\text{ub}} < 10: \quad m := 1 + m_{\text{ub}}/10 \end{array} \quad (18)$$

They demonstrated their method in three case studies with $\dim(x) \in \{13;16;60\}, c \in \{10;16;20\}$ and determined $m \in \{1.25;1.17;1.112\}$, respectively.

Mller [40] argues that too large values of m cause empty/missing clusters. He suggests to upper bound m by the value for which all lower values do not lead to missing clusters. In case studies on microarray/gene expression data, FCM was applied and upper bounds between 1.2 and 1.55 are determined. He refers to problems with high-dim. data, small sample size and low values of m (<1.5…2).

Yu et al. [21] derive two rules for upper bounding m by analyzing the optimality properties of fixed points of the FCM:

$$\begin{array}{l} \text{Rule } 1: m \leq \frac{\min(M,N-1)}{\min(M,N-1)-2}, \\ \text{if } \min(M,N-1) \geq 3 \end{array} \quad (19)$$

with the number of observations N and the dimensionality of the data M and

$$\begin{array}{l} \text{Rule } 2: m \leq (1 - 2 \cdot \lambda_{\max}(F_{U*}))^{-1}, \\ \text{if } \lambda_{\max}(F_{U*}) < 0.5 \end{array} \quad (20)$$

with

$$\begin{array}{l} F_{U*} = \frac{H^T H}{N}; \\ H = \left[\frac{x_1-\bar{x}}{||x_1-\bar{x}||}; \frac{x_2-\bar{x}}{||x_2-\bar{x}||}; \cdots; \frac{x_N-\bar{x}}{||x_N-\bar{x}||}\right] \end{array} . \quad (21)$$

For $\lambda_{\max}(F_{U*}) \geq 0.5$ any $m > 1$ is admissible. They tested their rules with data sets from the UCI Repository of Machine Learning data base with 3 to 26 classes. Upper bounds in the range of $m \in [1.6; 3.2]$ were obtained.

Zeinali and Notash [26] argue that *m* should be chosen to reflect the level of uncertainty in the data. The FCM is used to derive trapezoidal MF and complete a Mamdani-type fuzzy model. The initial location of the cluster centers is kept and *m* is chosen to minimize the mean squared prediction error of the model. In two modeling case studies with 8 clusters *m*=1.67 and *m*=1.9 are determined.

Yang et al. [41] used Simulated Annealing to determine *c* and *m* for gene expression/microarray data. The optimization target is the sum of the FCM cost functional and one of 4 validity indices such as the Xie-Beni index. A case study searches in the range of $m \in [1.05; 3]$ and determines a lower bound of *m*=1.05.

Okeke and Karnieli [42] suggest choosing *m* such that the original data set can be well predicted from the result of a FCM classification. They increment *m* starting from *m*=1.1 until a chosen upper bound for *m* such as 3, 5, 7 (up to 30) is reached. In case studies values $m \in [1.1; 2]$ were determined.

Kung and Su [43] examine the limiting behavior of Fuzzy-*c*-Regression Modeling (FCRM) for $m \to 1_+$ and $m \to \infty$. In case studies they use fixed values $m \in \{1.1; 2; 7\}$.

Sugeno and Yasukawa [44] used *m*=2 with the FCM in case studies on identification of Mamdani-type fuzzy models and refer to $m \in [1.5; 3]$ as the usual choice. Zarandi and Esmaeilian [45] use a Genetic Algorithm to determine *m* for the use in Mamdani-type fuzzy models. In a case study, *m*=4.47 was computed as the optimum within the bounds of $m \in [1; 8]$.

In the context of Takagi-Sugeno modeling, fuzzy clustering is in general carried out using *m*=2 without investigating the impact of this choice, e.g. [6],[11],[15],[16],[17],[18],[19]. Alata et al. [46] used the FCM with *m*=2 and succeedingly a Genetic Algorithm to optimize *m* wrt. the prediction error. In three modeling case studies the optimized values are $m \in \{1.4; 1.7; 101\}$. In [8] the impact of choosing $m \in (1; 10]$ on the prediction error is examined. For FCM- or GK-based identification of dynamical models, the general trend of the approximation error decreasing with *m* was shown.

Different approaches are taken in [22],[47]: The FCM-cost functional for *m*=2 is extended by an additional term in [47] to enforce broader areas of distinct memberships. In [22], the objective function of the clustering algorithm is changed to amend problems with reactivation. As the fuzziness parameter is globally effective, in [48] a method for local adaptation of *m* is proposed in order to answer locally deviating requirements that can occur when modeling heterogeneous systems. In [20], type-II-fuzzy sets are used to manage uncertainty in *m*; for this purpose an upper and lower limit for *m* can be specified.

5 Theoretical Constraints and Recommendations on Choosing m

The previous paragraph recorded different methods to choose *m* from various application areas. In this section, choosing for clustering-based TS fuzzy model identification is discussed.

5.1 Introductory Example

Eq. (5) shows that the fuzziness does not only depend on *m*, but also on *c*: Each cluster contributes a summand. Hence, if moving away from a prototype, μ declines earlier and faster if *c*. is increased. Evidently, the impact of *c* on the fuzziness increases with *m* and becomes significant for larger values. A simple example is used to illustrate this: A succeeding number of prototypes is added around a prototype $v_1 = 0$ while the effect on the membership assignment is monitored. Arrangements with $c = 2; 3; 4; 9; 27$ prototypes are studied for $m = 1.5; 2; 3; 5$. Figure 1 records the results if

- (a) one additional center is placed at $x_1 = 1$ (1-dim. Case, "+"),
- (b) two additional centers are placed at $x_1 = \pm 1$ (1-dim. Case, "*"),
- (c) 3 additional centers are placed at $x = (1,0), (1,1), (0,1)$ (2-dim. Case, "•"),
- (d) 8 additional centers are placed at x = (1,0), (1,1), (0,1), (-1,1), (-1,0), (-1,-1), (0,-1), (1,-1) (2-dim. Case, "⋄"), and

- (e) the centers are as in d) plus the same 3x3 grid of centers being copied and shifted by 1 and -1 in direction of x_3 (3-dim. Case, "○").

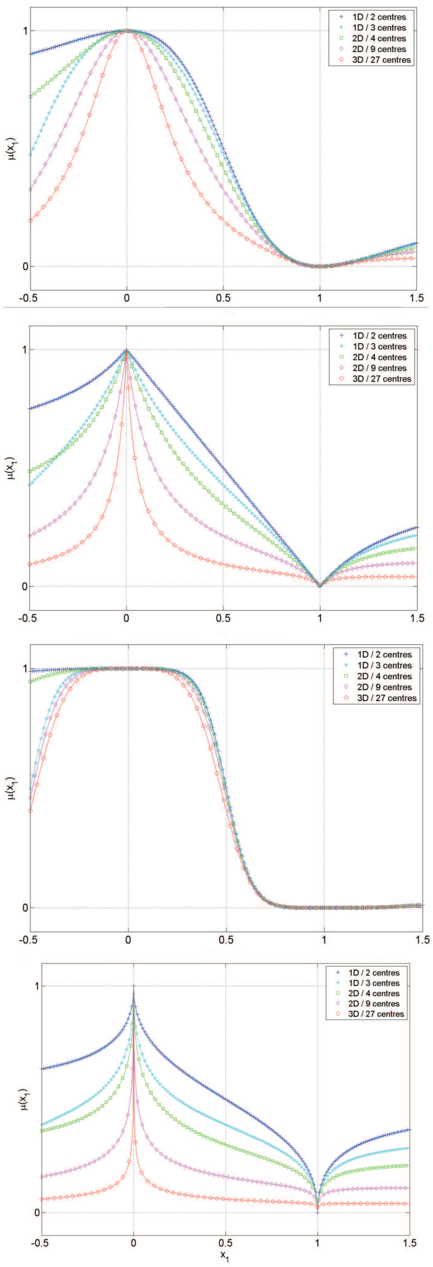

Figure 1. Shape of membership function with center in 0 for different numbers of neighbouring centers and different choices of fuzziness parameter: $m = 1.5; 2; 3; 5$(top to bottom).

The Euclidean distance norm is used. For m=1.5 (or smaller), distinct memberships prevail and areas with indistinct membership have minor extension. Reactivation and the dependency of the classifier fuzziness on c are negligible. For m=2, increasing c (for a given prototype distribution) causes the membership functions to narrow down: Along the connection line between both centers in Figure 1, the membership to $v_1$has halved at mid distance for c=2 but already at quarter distance for c=27. This effect is in general not desirable. For m>2, the described effects occur more pronounced.

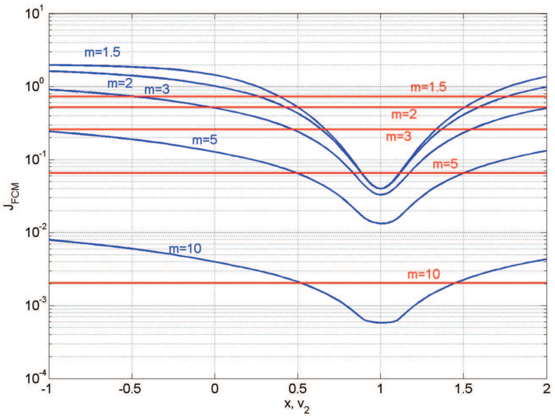

Figure 2. J_{FCM} for different choices of m: for fixed position of $v_1 = 0$ and varying position of v_2 (blue) or for placing both v_1 and v_2 at $\bar{X} = 0.5$ (red)

5.2 Limiting Behavior

The FCM approaches the (hard) c-means algorithm for $m \to 1$. For $m \to \infty$, all memberships except at the centers approach $1/c$: $\lim_{m\to\infty} \mu_i(x) = 1/c \forall x \neq v_r; r = i, j$. The statement that all centers v_i "approach" the (total) mean $\bar{X}$ of the observations X [27] is however not valid in general, as counterexamples show: Consider the case in which a data set X that consists of a point at 0 and one at 1 has to be clustered into two groups. Placing a cluster center at each point, provides exactly for J_{FCM}=0 while placing both centers into $\bar{X} = 0.5$ yields J_{FCM}>0 for finite m. Then consider $X = \{-0.1; 0.1; 0.9; 1.1\}$for c=2 clusters. Placing cluster centers at 0 and 1, results in a lower J_{FCM} for finite m than placing them at $\bar{X} = 0.5$, see Figure 2. This means that the cluster centers do not necessarily approach $\bar{X}$with rising value of m.

5.3 Distance Weighting and Grouping Strategy

Probabilistic MF (5) assess a relative distance $\tilde{d_{i,j}} := d_i/d_j$, which is weighted by an exponential term $(\tilde{d_{i,j}})^{2/(m-1)} =: (\tilde{d_{i,j}})^{\beta}$. For $1 < \beta < \infty$ (i.e. $1 < m < 3$) the exponent decreases the effective distance for $0 < \tilde{d}_{i,j} < 1$ and increases it for $\tilde{d}_{i,j} > 1$ as compared with the unweighted distance, i.e. an exponent of 1 (β=1, m=3) (Figure 3). The half-way inter-center distance $\tilde{d}_{i,j} = 1$ is not affected by the weighting. Hence, points closer to the reference v_i than to $v_j$are assigned a higher membership to v_i than in the case of unweighted distance, and points further away a lower one. This provides for an enlarged region of significant membership assignments for smaller choices of m. Hence, choosing a small value of m supports fuzzy partitioning as required for fuzzy modeling. On the contrary, for 0< β<1 (i.e. 3<m< ∞) the exponent increases the effective distance for $0 < \tilde{d}_{i,j} < 1$ and decreases it for $\tilde{d}_{i,j} > 1$. Hence, points closer to the reference v_i than to v_j are assigned a lower membership to the reference than in the unweighted case. This provides for faster declining membership values with rising distance from the prototype. Such a parameterization supports the objective to just group data of high similarity. The limit behavior is that for β → *0* (i.e. $m \to \infty$) a cluster's extension reduces just to its prototype.

5.4 Differentiability-based Criterion

Understanding the differentiability property of MFs and therefore of the entire TS model dependent on m sets the base for applying derivative-based parameter optimization methods. These can be used to efficiently identify NOE-models or a-posteriori tune model parameters. Besides, differentiability has a geometrical interpretation: Losing the property of continuous differentiability means losing the property of a smooth function graph. This motivated the derivation of the following

Theorem 1: The function

$$\mu_i(x) = \left[\sum_{j=1}^{c} \left(\frac{||x-v_i||^2_{A_i}}{||x-v_j||^2_{A_j}} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{22}$$

with $x, v_i, v_j \in \Re^M, c \in N_+, m \in \Re$ and an inner product norm

$$||x-v_r||^2_{A_r} = (x-v_r)^T \cdot A_r \cdot (x-v_r) \tag{23}$$

with a positive definite matrix $A_r = [a^r_{h,l}] \in \Re^{M \times M}$ is continuously differentiable wrt. x for m∈(1;3).

The proof is recorded in the appendix.

Remark 1: As both FCM and GK algorithm use MF (22), the result holds for both.

Remark 2: Similarly it can be shown that MF (22) that use a L$_p$-norm are continuously differentiable wrt. x for m∈(1;3).

Remark 3: Analogously, it can be shown that μ_i is continuously differentiable wrt. all v_r and to m for m∈(1;3). This also holds if a L$_p$-norm is used.

A similar result can be derived for MF that are used by possibilistic fuzzy cluster algorithms:

Theorem 2: The function

$$\mu_i(x) = \frac{1}{1 + \left(\eta_i^{-1} ||x-v_i||^2_{A_i} \right)^{\frac{1}{m-1}}} \tag{24}$$

with $x_k, v_i \in \Re^M, \eta, m \in \Re_+$with an inner product norm

$$||x-v_i||^2_{A_i} = (x-v_i)^T A_i \cdot (x-v_i) \tag{25}$$

with a positive definite matrix $A_i = [a^i_{h,l}] \in \Re^{M \times M}$ is continuously differentiable wrt. x for m∈(1;3).

The proof is recorded in the appendix.

Remark 1: As the possibilistic c-means (PCM) and the possibilistic Gustafson-Kessel (PGK) algorithm use both membership functions of type (24) [51], the result is applicable to both algorithms.

Remark 2: Similarly it can be shown that μ_i using a L$_p$-norm are continuously differentiable wrt. x for m∈(1;3).

Remark 3: Analogously it can be shown that μ_i is continuously differentiable wrt. all v_r,m and η_i for m∈(1;3).

The example in Figure 1 illustrates the geometrical interpretation of these results: The graph of $\mu_1(x_1)$ is smooth for $m \in (1;3)$ and has a flat central region around the MF center. As shown in the proof, for m=3 the derivate is discontinuous but finite in the centers. In the function graph, this provides for a peak with finite opening angle at the centers. For m>3, the derivative has a pole in the centers, providing for a sharp peak in the function graph at the centers.

5.5 TS Modeling Specific Requirements

The proposed clustering-based TS fuzzy model identification strategy requires choosing m in three places: clustering, weighting during local model parameter estimation and model evaluation. All can be chosen individually. The location of the prototypes determined by FCM and GK clustering little depend on the value of m (if the clusters have similar geometry and density) [20]. The shape of the MFs and therefore the weighting of individual training data significantly depend on m, see Figure 1.

If global estimation (GE) is used, a consistent situation during identification and evaluation is important, as GE also considers the interpolation regions and fits the model tighter to the data. It is recommended to use the same value of m during estimation and evaluation. A value next to 1 is advisable, if the system is approximately piecewise affine and the model partitioning well reproduces the true partitioning. In case of smooth nonlinearities, the approximation quality benefits from giving the estimator more effective degrees of freedom. A choice of $1 < m < 1.5$ typically provides for good results. GE can create models that cannot be interpreted as local linearization anymore. If the latter is required, choosing m close to 1 limits the possible extend (for $m \to 1$ GE converts to LE). The issues of reactivation and unintuitive interpolation in TS models have been addressed in the sections 1 and 5.1. Their severity is reduced by a small choice of m. However, GE considers these effects, which limits their impact on the approximation error.

A local estimator (LE) conceptually ignores the interpolation between the local models during identification. Therefore, a small value of $1 < m < 1.3$ reduces negative effects of data in the interpolation region on the estimation. As in contrast to GE, the situations during identification and evaluation differ anyway, for model evaluation m can possibly be chosen slightly larger than for estimation to "soften" the interpolation region. A small value of m is key for the local interpretability, meaning firstly membership function with distinct central areas and secondly local models that can be interpreted as local linearization of the true system. Reactivation and unintuitive interpolation effect are not "managed" by LE, which requires an m close to 1 to restrict their severity.

Good results are obtained with product space clustering, as it uses more information on the system characteristics than input space clustering. However, MFs defined on the product space have to be projected to the input space for model evaluation, which causes changes in assigned memberships of the data during estimation and evaluation. Therefore, particularly for GE the best strategy will be case-dependent. It is remarked that in case of dynamic system modeling, the system output is delayed and fed back as input such that the input space includes information on the output.

The undesired reactivation effect can also be addressed by changing the clustering algorithm: A noise cluster can be added, to which data far away from the other clusters are assigned [49]. Also, the objective function of the clustering algorithm can be changed [42], [47]. As reactivation does not occur for MFs with finite support, MF (5) can be projected and approximated accordingly (e.g. by trapezoids). The drawback is the resulting approximation error. Finally it is remarked that partitions and local models can also directly be derived from the result of GK-product-space- clustering [11].

5.6 Summary

The majority of the methods developed to adjust m addresses grouping or general classification problems. This paper addresses fuzzy-partitioning applications for TS models imposing altered requirements. It is assumed that membership functions (5) are used straight in the TS model or projected and approximated.

With respect to *properties of the fuzzy partition*, it was proven (independent from the application) that a desirable flat central area around the partition centers results for $1 < m < 3$ and gets lost for larger m. This represents a hard upper bound. Secondly, on example it was shown that for about $m > 1.5$ (which is a soft upper bound) several "side effects" due to the design of function (5) become significant: Reactivation increases. Adding centers around given ones changes the memberships in the "enclosed area". This statement is supported by several parametric case studies, of which some are presented in the next section. These results do not depend on the method used to obtain the partitioning. As $m = 1.5$ already causes noticeably extended interpolation regimes, a strong demand on interpretability requires a choice closer to 1.

With respect to the *performance of the identification strategy*, having distinct memberships avoids impairing local estimations by peripheral data that does not "belong" to the local model. Experience shows that $m < 1.3$ in case of local and $m < 1.5$in case of global estimation will typically provide for good results. Continuously differentiable membership functions for $1 < m < 3$simplify the application of derivate-based optimization. This is useful to estimate NOE models or to optimize the partitioning regarding approximation quality.

6 Numerical Examples

6.1 Smooth Nonlinear Function Approximation

The approximation of several basic generic nonlinear functions (step, saturation, sigmoid, trigonometric, semicircle. . .) by TS models was studied regarding the optimal choice of m. As example, the results for a semi-circle as smooth nonlinearity are presented. For this purpose, training data sets were generated by evaluating $y = \sqrt{1-x^2}$ for $N =$ 81equidistant arguments in the interval $x \in [-1;1]$. Studies were made with noiseless and noisy data. As they provided for similar results only the results for ideal data are presented in the following.

The FCM was applied with Euclidean distance norm. For LE, memberships defined in the product space were used as weights. In case of GE, memberships were defined in the input space. Different choices of m for clustering and identification on the one hand as well as for model evaluation on the other hand were examined. Figure 3 shows the mean squared prediction error $E_{\mathrm{MSE}} = N^{-1} \cdot \sum_{k=1}^{N} (y_k - \hat{y}_k)^2$ in a qualitative manner by visualizing it as a matrix of grey scale values. The value of m was varied in increments of 0.1 within the interval [1.1; 3]. The choice of m during clustering is not varied independently as it hardly affects the resulting position of the prototypes; m is changed in the membership functions used for weighting in the parametric studies. The columns record top-down the results for simultaneously increasing the value of m in clustering and estimation. The rows note the results for, from left to right, increasing value of m in model evaluation. Hence, the top left cell stands for the crispiest configuration.

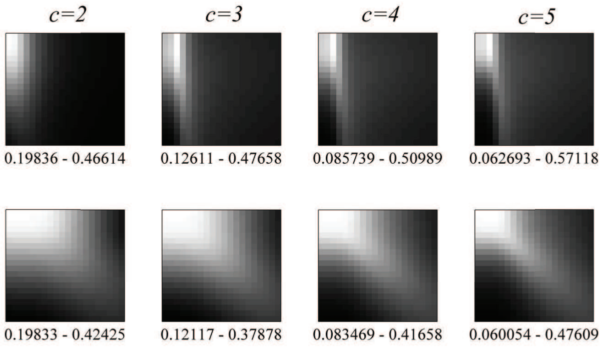

Figure 3. TS model performance (EMSE) for approximating a semi-circle function: top (bottom) row shows results for LE (GE). Columns refer to m used for clustering and modelling, rows refer to m used for model evaluation. m was varied in increments of 0.1 between 1.1 and 3.0 with the top left cell corresponding to the lowest values. Light (dark) colour indicates a low (high) EMSE. The range of encountered EMSE numbers is recorded below each matrix.

Using LE with m=1.3 provides optimal results for model building (except for c=2, which favors m=1.1). Regarding model evaluation, $m \in [1.1;1.4]$ provides optimal results. A small value of m restricts the range of data effectively used by an LE and therefore avoids that peripheral data (from the local model's perspective) impairs the results. GE considers the interpolation effects and a larger fuzziness increases the effective degrees of freedom for the estimator. This explains that for model building m should be chosen higher than for LE, e.g. $m \in [1.1;1.4]$. A design for consistency during training and evaluation is preferred, which is indicated by the good results on the main diagonal in the figure. This means it is recommended to use the same m for estimation and evaluation. A comparison of E_{MSE} value for LE (top) and GE (bottom) in Figure 3 shows that the best achieved performance is similar for LE and GE. However, LE yields good models only for choosing m in a narrow range. GE provides for good modeling for a wider range of m.

6.2 Discontinuous Piecewise Affine Function Approximation

This subsection studies identifying a piece-wise affine (PWA) model for a PWA system, as this permits to compare estimated and true parameter sets. Moreover, it is an example of a discontinuous original system. Consider a system that is composed of c=3 truly affine local models

$$y_i(x_1;x_2) = [x_1;x_2;1]\cdot[a_{1,i};a_{2,i};a_{0,i}]^T \\ =: [x^T;1]\Theta_i$$

given as

$$\begin{array}{l} y_1 = -4x_1 + 4x_2 - 2 =: [x_1;x_2;1]\cdot\Theta_1 \\ y_2 = 4x_1 - 2x_2 - 4 =: [x_1;x_2;1]\cdot\Theta_2 \\ y_3 = 2x_1 + x_2 + 1 =: [x_1;x_2;1]\cdot\Theta_3 \end{array} \quad . \tag{26}$$

Let the bounds of local models be defined by assigning a center to each model

$$\begin{array}{l} v_1 = [0.5;0.5]^T; v_2 = [0.5;1.5]^T; \\ v_3 = [1.5;1]^T \end{array} \tag{27}$$

and by carrying out a Dirichlet decomposition of the data space for these centers using the Euclidean norm. This defines the true system, which is to be approximated by a TS model.

An input data set X of N=90 observations is constructed as follows: Around each $\mathbf{v}_i$ (27) 30 normal distributed points $x_k$are generated using a uniform variance of σ^2=0.25. The points are assigned to the local model, in which partition it is located. Eqs. (26) are used to compute the y_k for all x_k. Figure 5 (semitransparent) shows the graph of the "true" system. To obtain training data with simulated noise, a mean-free normal distributed random number with variance of 0.25 (N(0;0.25)) is added to each y_k.

The FCM is applied with c=3 and Euclidean. m is incremented by 0.01 in $[1.1;1.5]$ and 0.1 in $[1.5;10]$. The FCM is terminated if $\Delta\mu_{i,k}|_{\max} < 10^{-8}$or if 100 iterations are completed. Five random initializations are tested for each choice of m, but the initial value dependence was negligible. Clustering in the input space provided for prototypes close to the true centers v_i, so did product space clustering. The parameter vectors are estimated using LE with the MF resulting from clustering as weights. Figure 4 shows J_{FCM}, E_{MSE}, the maximum absolute error $E_{\max} = \max_k |y_k - \hat{y}_k|$and the Euclidean norm of the deviation of LE-estimated and true parameters (22) $||\Delta\Theta|| := \sum_{q,i}\sqrt{(\hat{\Theta}_{q,i} - \Theta_{q,i})^2}$ dependent on m. For high model quality, m should be small. On the contrary, J_{FCM} declines with rising m. This results from $\partial J/\partial m < 0 \forall m > 1$, see (11), and is application-independent. The original function is well approximated for m=1.1: Figure 5 compares the graphs of true and identified models. Figure 6 shows the estimated TS model that considers the interpolation between the local models due to the soft MF in contrast to the hard switching original system. It illustrates that TS models with m close to 1 can be used to approximate discontinuous systems.

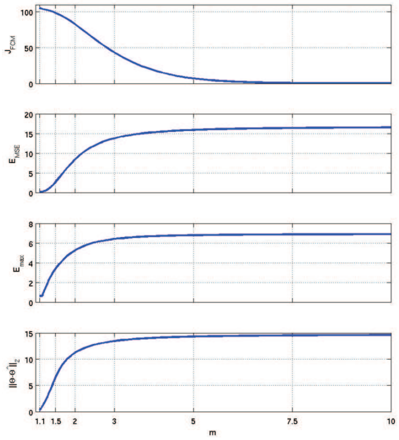

Figure 4. Performance criteria dependent on m determined on training data

6.3 Vehicle Engine Characteristic Curve

The identification of characteristic curves of a 3l-FSI-Audi-gasoline engine from data recorded at a test stand (Figure 7) provides for a smooth nonlinear regression problem. Air mass flow y, manifold pressure (x_1), speed (x_2) and throttle opening (x_3) were measured. Plotting y against x_1 and $x_2$or x_2 and x_3, respectively, exhibits a locally approximately plan distribution of the measurements. Plotting y against $x_1$and x_3, however, reveals a "helix type" characteristic, which is difficult to capture with parsimonious, local affine TS models. The available N=689 measurement data were normalized to the unit interval and then randomly divided into equally sized training and a test data set. Different identification methods and model candidates were examined. Clustering with the GK algorithm provided for one magnitude smaller value of J_{FCM} than using FCM with Euclidean, but the prediction quality is about the same. On the contrary, the parameter count of a GK-based model nearly doubles for the same c. Using the FCM with Mahalanobis norm yielded significantly worse results than for the Euclidean wrt. both clustering and prediction quality related criteria. Assessing the model alternatives with information criteria such as the Bayesian information criterion, BIC [6], promotes using FCM and Euclidean. The results for this choice will be presented in more detail.

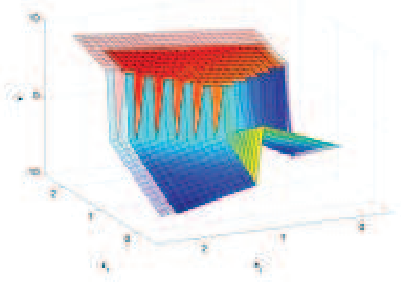

Figure 5. True (noiseless) PWA system (semitransparent) and graphs of local models identified by LE for m=1.1 using noisy data (full colour)

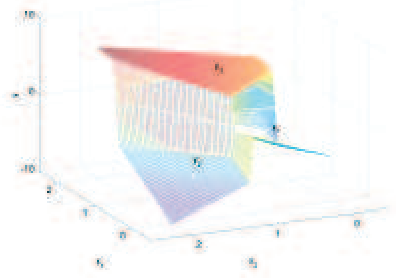

Figure 6. Transfer characteristic of identified TS model for PWA model

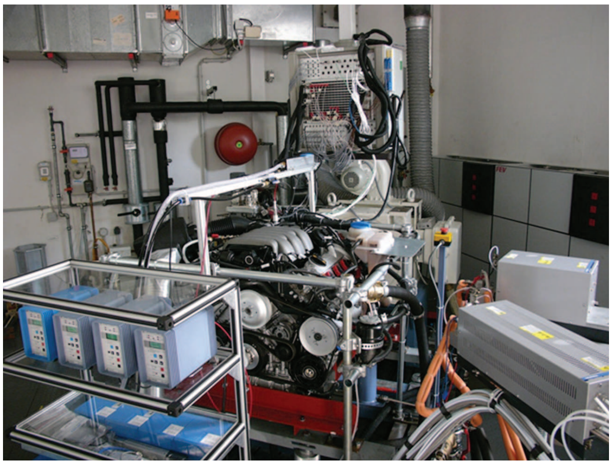

Figure 7. Engine test stand used to record the data.

A 2-step-clustering procedure is used where the result from the first clustering with larger fuzziness (m=3) initializes a 2^{nd} clustering with smaller fuzziness. The identification procedure was repeated 30 times with random initialization for each choice of design parameters to avoid inappropriate local convergence of the FCM. The result with minimum J_{FCM} was chosen. This procedure was experienced to be moderately less prone to inappropriate local convergence. The m-values in Figure 8 refer to the choice for the second clustering step. This value is used also during parameter estimation and model evaluation. The scores for E_{MSE} and E_{max} in Figure 8 indicate 1.1?m?1.5 as a suitable range. Given the nature of the problem, there is no “true” number of clusters. The choice depends on the required approximation quality. Figure 9 compares measurement data and prediction of a TS model with c=9 and m=1.1 on test data. This model was the best of the tested FCM models, see also Figure 8. Evidently the TS model can well approximate the helix structure.

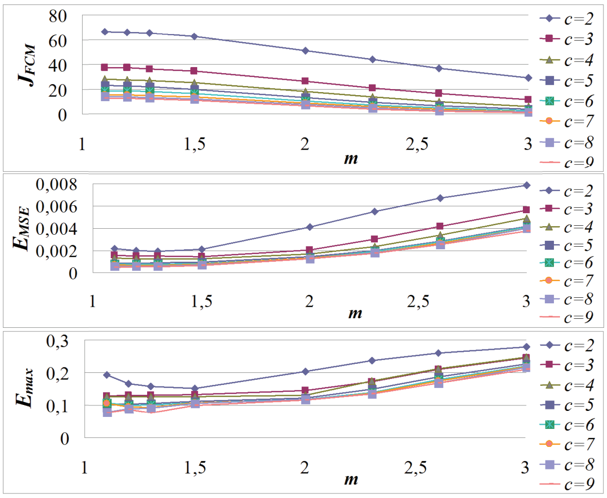

Figure 8. Engine model performance dependent on fuzziness m and number of clusters c evaluated for training data.

6.4 Compressor Characteristic Curve

The identification of a characteristic curve of an axial compressor (NASA CR-72694) provides for a heterogeneous nonlinear regression problem: As the original data in Figure 10 show, both smooth nonlinearities showing locally varying curvature and crisp behavioral changes appear. Moreover, the crisp change does not follow a straight line. The objective is to determine a parametric model for the mass flow (y) dependent on isentropic efficiency (x_1) and pressure ratio (x_2). The mean was removed from the data and the result divided by the maximum deviation from the mean. 500 of the

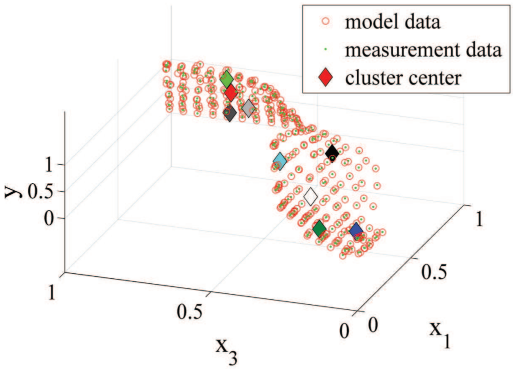

Figure 9. Air mass flow y dependent on manifold pressure (x$_1$) and throttle opening (x$_3$): measurements and prediction of engine model with *c = 9, m = 1.1* (normalized data).

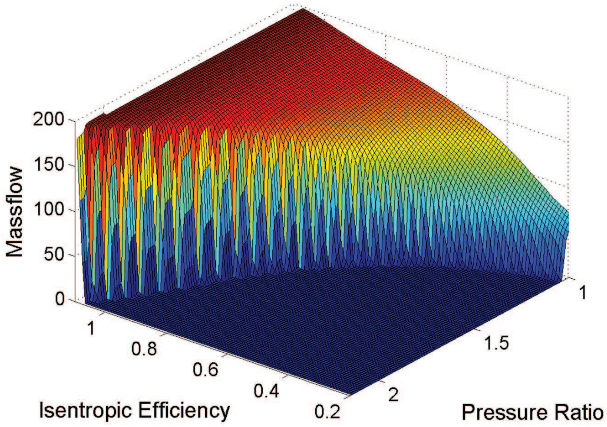

Figure 10. Reference compressor curve (original data).

available 56248 data sets were randomly chosen for identification, another 500 randomly for model validation. Partitioning was done in the product space using the FCM with Euclidean. Clustering and parameter identification were repeated 20 times for all parameterizations of c and m in order to avoid inappropriate local convergence of the FCM. Parameters were determined using LE with product space membership functions. For GE the MFs were projected to the input space. Varying m in increments of 0.01 and c indicates m=1.2 and c=6 to be a suitable choice (Figure 11). Investigating the shape of the resulting (soft) partition boundaries revealed that the course of the crisp change in the original behavior is not well reproduced yet. Therefore the identification algorithm was augmented by an a-posteriori numerical optimization step that adjusted all $\mathbf{v}_i$and m in order to minimize J_{MSE}. Figure 12 shows the resulting model. The optimization reduced the identification error by a factor of 4 and the validation error by about a third. The reason for this improvement is that the optimization moved the $\mathbf{v}_i$ such that the resulting partitioning well reproduced the bended contour of the "step change" of the graph. The optimized fuzziness was m=1.06. A comparison with the piecewise affine model in [50] favors the presented approach due to a smaller overall prediction error and less local models. Further details are provided in [48].

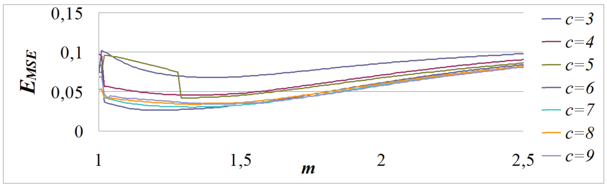

Figure 11. E$_{M}SE$ on test data for different choices of m and c in the compressor study (normalized data).

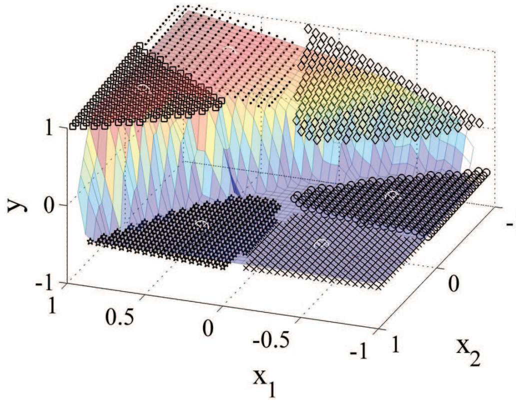

Figure 12. TS model with 6 local models (dotted planes), corresponding partition centres (circles) and overall output (mesh) (normalized data).

7 Conclusions

Fuzzy clustering has successfully been deployed in many areas including TS model identification. A key design parameter is the weighting exponent/fuzziness parameter m. Different recommendations were given on the choice with m=2 often being the default for grouping and partitioning applications. This contribution proposed guidelines for choosing m for clustering-based TS model identification, where the membership functions resulting from clustering are used as is for the TS model or projected and approximated before.

Bounding $m \in (1;3)$ is proposed from analyzing differentiability and grouping concept of the classifier function. It was shown that additional, model-specific requirements on choosing m have to be adhered to derive models of good quality. This provided for tighter bounds: Optimal application-specific value of m can typically be found within the interval $(1;\ 1.3]$ for local and $(1;\ 1.5]$ for global estimation – where the upper bound is meant to be fuzzy. The guidelines were demonstrated in four case studies, which also have shown the significant effect of m on the quality of the identified TS model. As the optimal choice of m depends on the application, the common choice of m=2 could be optimal. However, it has not been a good choice in any of the case studies carried out by the author. Therefore, it is advised to carefully consider the choice of m for a given problem.

Appendix

Proof of theorem 1: Given arguments $x = [x_1;...;x_f;...;x_M]^T \in \Re^M$, prototypes $v_r = [v_{r,1};...;v_{r,M}] \in \Re^M, r \in \{1;2;...;c\}$, and an inner product norm

$$d_r^2 := ||x-v_r||_{A_r}^2 = (x-v_r)^T A_r \cdot (x-v_r) \qquad (28)$$

with form matrix $A_r = [a_{h,l}^r] \in \Re^{M\times M}$, the partial derivative $\partial\mu_i/\partial x_f$is obtained as:

$$\frac{\partial\mu_i}{\partial x_f} = \frac{\frac{-1}{m-1}\sum_{\substack{j=1\\ j\neq i}}^{c}\left(\frac{d_i^2}{d_j^2}\right)^{\frac{2-m}{m-1}}\frac{\left(\frac{\partial d_i^2}{\partial x_f}\right)d_j^2-d_i^2\left(\frac{\partial d_j^2}{\partial x_f}\right)}{d_j^4}}{\left(1+\sum_{\substack{j=1\\ j\neq i}}^{c}\left(\frac{d_i^2}{d_j^2}\right)^{\frac{1}{m-1}}\right)^2} \qquad (29)$$

with

$$\frac{\partial d_r^2}{\partial x_f} = \sum_{l=1}^{M}(a_{f,l}^r + a_{l,f}^r)(x_l - v_{r,l}). \qquad (30)$$

Potential points of discontinuity are the cluster centers v_r and the further analysis can be restricted to those. Hence, it remains to show that the partial derivatives of μ_i to all $x_f$are continuous in all v_r. It will be analyzed for what m left- and right-hand limit of $\partial\mu_i/\partial x_f$are identical if x_f approaches $v_r$from above or below. As μ_i takes a strict maximum of 1 in v_i and a strict minimum of 0 in any $v_j, j \neq i$, this is only the case if $\partial\mu_i/\partial x_f \to 0$ for left- or right-hand approaching a prototype. Theoretically, $m < 1$ is possible, but would not provide for meaningful membership functions.

Approaching v_i: Within a sufficient small neighborhood v_i is approached in direction of x_f. Then $d_j > 0 \forall j \neq i$ and $d_i^2 = (x_f - v_{i,f})^2 a_{f,f}^i$. Inserting the latter into eq. (29) provides for

$$\frac{\partial\mu_i}{\partial x_f} = \frac{\frac{-1}{m-1}}{\left(1+\sum_{\substack{j=1\\ j\neq i}}^{c}\left(\frac{(x_f-v_{i,f})^2 a_{f,f}^i}{d_j^2}\right)^{\frac{1}{m-1}}\right)^2}\cdot$$
$$\cdot\sum_{\substack{j=1\\ j\neq i}}^{c}\left(\frac{(x_f-v_{i,f})^2 a_{f,f}^i}{d_j^2}\right)^{\frac{2-m}{m-1}}\cdot\frac{1}{d_j^4}\cdot \qquad (31)$$
$$\cdot\left(2(x_f-v_{i,f})a_{f,f}^i d_j^2 - \ (x_f-v_{i,f})^2)\right)\cdot$$
$$\cdot a_{f,f}^i \textstyle\sum_{l=1}^{M}(a_{f,l}^j + a_{l,f}^j)(x_l - v_{j,l})$$

After collecting the $(x_f - v_{i,f})$-terms it can be concluded that $\partial\mu_i/\partial x_f \to 0$ if $(x_f - v_{i,f})^{\frac{3-m}{m-1}} \to 0$ and $(x_f - v_{i,f})^{\frac{2}{m-1}} \to 0$ for $x_f \to v_{i,f}$. This is the case if $m \in (1;3)$. Eq. (31) also shows that different non-zero finite left- and right-hand limits result for $m = 3$. A pole results in v_i for $m \leq 1$ or $m > 3$.

Approaching $v_j; j \neq i$: Within a sufficient small neighborhood, $v_h \in \{v_1, v_2, ..., v_c\}\backslash v_i$[1] is approached in direction of x_f. Rearranging eq. (29) provides for:

$$\frac{\partial\mu_i}{\partial x_f} = \frac{\frac{-1}{m-1}\left(d_h^2\right)^{\frac{2}{m-1}}\left(d_i^2\right)^{\frac{2-m}{m-1}}}{\left(\left(d_h^2\right)^{\frac{1}{m-1}}+\left(d_i^2\right)^{\frac{1}{m-1}}+\sum_{\substack{j=1\\ j\neq i,h}}^{c}\left(\frac{d_i^2}{d_j^2}d_h^2\right)^{\frac{1}{m-1}}\right)^2}\cdot$$
$$\cdot\sum_{\substack{j=1\\ j\neq i}}^{c}\left(\frac{1}{d_j^2}\right)^{\frac{m}{m-1}}\left(\left(\textstyle\sum_{l=1}^{M}(a_{f,l}^i + a_{l,f}^i)(x_l - v_{i,l})\right)d_j^2 - \right.$$
$$\left. -d_i^2\left(\textstyle\sum_{l=1}^{M}(a_{f,l}^j + a_{l,f}^j)(x_l - v_{j,l})\right)\right) \qquad (32)$$

Then $d_i > 0, d_j > 0 \forall j \neq i$and $d_h^2 = (x_f - v_{h,f})^2 a_{f,f}^h$. Inserting the latter into eq. (32) and collecting the critical terms $(x_f - v_{h,f})$ provides for $(x_f - v_{h,f})^{\frac{2}{m-1}}$ and $(x_f - v_{h,f})^{\frac{3-m}{m-1}}$. These terms have to approach 0 for $x_f \to v_{h,f}$, which is the case for $m \in (1;3)$.

[1] *Subindex h isintroducedforthecenterofinterest,asthereare c* -1*centers v j in*(29).

Proof of theorem 2: Given arguments $x = [x_1;...;x_f;...;x_M]^T \in \Re^M$, prototypes $v_i = [v_{i,1};...;v_{i,M}] \in \Re^M, i \in \{1;2;...;c\}$, and an inner product norm

$$d_i := ||x - v_i||_{A_i}^2 = (x - v_i)^T A_i \cdot (x - v_i) \tag{33}$$

with form matrices $A_r = [a_{h,l}^r] \in \Re^{M \times M}$, the partial derivative $\partial\mu_i/\partial x_f$is obtained as:

$$\frac{\partial \mu_i}{\partial x_f} = \frac{\frac{-1}{m-1}\left(\frac{1}{\eta_i}\right)^{\frac{1}{m-1}}\left(d_i^2\right)^{\frac{2-m}{m-1}}}{\left(1+\left(\frac{d_i^2}{\eta_i}\right)^{\frac{1}{m-1}}\right)^2} \cdot \frac{\partial d_i^2}{\partial x_f} \tag{34}$$

Inserting $d_i^2 = (x_f - v_{i,f})^2 a_{f,f}^i$ in eq. (34) shows that $\partial\mu_i/\partial x_f \to 0$ if $(x_f - v_{i,f})^{\frac{3-m}{m-1}} \to 0$for $x_f \to v_{i,f}$. This is the case if $m \in (1;3)$. Eq. (32) also shows that different non-zero but finite left- and right-hand limits result for $m = 3$. A pole results in v_i for $m \leq 1$ or $m > 3$.

Acknowledgment

The author is grateful to A. Schrodt, A. Drrbaum and S. Soldan for their support on the application studies and to A. Geiger for his help with typesetting. Thanks are due to Dr. M. Ayeb, Faculty of Electrical Engineering and Computer Science, University of Kassel for providing the engine measurement data.

References

[1] Takagi T. and Sugeno M., "Fuzzy identification of systems and its application to modelling and control", *IEEE Trans. Systems, Man and Cybernetics,* Vol. 15, No. 1, pages 116–132, 1985.

[2] Tanaka K. and Wang H.O., *Fuzzy Control Systems design and analysis: a linear matrix inequality approach*. New York: Wiley, 2001.

[3] Kawamoto S., Tada K., Ishigame A. and Taniguchi T., "An approach to stability analysis of second order fuzzy systems," Proc. *IEEE Conf. on Fuzzy Systems*, San Diego, pages 1427–1434, 1992.

[4] Wang H.O., Tanaka K. and Griffin M.F., "An approach to fuzzy control of nonlinear systems: stability and design issues," *IEEE Trans. on Fuzzy Systems*, Vol. 4, No. 1, pages 14–23, 1996.

[5] Johansen T.A., Hunt K.J., Gawthrop P.J. and Fritz H., "Off-equilibrium linearization and design of gain-scheduled control with application to vehicle speed control," *Control Eng. Practice*, Vol. 6, pages 167–180, 1998.

[6] Nelles O., *Nonlinear system identification*. Berlin, Germany: Springer, 2001.

[7] Hartmann B. and Nelles O., "Automatic adjustment of the transition between local models in a hierarchical structure identification algorithm," in *Proc. European Control Conference (ECC'2009)*, Budapest, Hungary, pages 1599–1604, August 2009.

[8] Kroll A., "Identification of functional fuzzy models using multidimensional reference fuzzy sets," Fuzzy Sets and Systems, Vol. 80, No. 2, pages 149–158, 1996.

[9] Kroll A., "On modeling discontinuous and heterogeneous nonlinear systems using Takagi-Sugeno systems," (in German) in *Proc. GMA Workshop 'Computational Intelligence'*, Dortmund, Germany, pages 64–79, Dec. 2010.

[10] Bezdek J. C., *Pattern recognition with fuzzy objective function algorithms*, New York: Plenum, 1981.

[11] Babuska R., *Fuzzy modeling for control*, Boston: Kluwer, 1998.

[12] Hppner F., Kruse F., Klawonn F. and Runkler T., *Fuzzy cluster analysis*, Chichester, England: Wiley, 1999.

[13] Abonyi J., *Fuzzy model identification for control*. Berlin: Birkhuser, 2003.

[14] Krishnapuram R. and Freg C.-P., "Fitting an unknown number of lines and planes to image data through compatible cluster merging", *Pattern Recognition*, Vol. 25, No. 4, pages 385–400, 1992.

[15] Sin S.-K., and de Figuero R. J. P., "Fuzzy system design through fuzzy clustering and optimal predefuzzification," in *Proc. 2nd IEEE Conf. on Fuzzy Systems*, San Francisco, pages 190–195, 1993.

[16] Laukonen E. G., Passino K. M., Krishnaswami V., Luth G.-C. and Rizzoni G., "Fault detection and isolation for an experimental internal combustion engine via fuzzy identification," *IEEE Trans. Control Systems Technology*, Vol. 3, No. 3, pages 347–355, Sep. 1995.

[17] Setnes M., "Supervised fuzzy clustering for rule extraction", *IEEE Trans. Fuzzy Systems*, Vol. 8, No. 4, pages 416–424, August 2000.

[18] Abonyi J., Babuska R. and Szeifert F., "Modified Gath-Geva fuzzy clustering for identification of Takagi-Sugeno fuzzy models," *IEEE Trans. Systems, Man, and Cybernetics, Part B*: *Cybernetics*, Vol. 32, No. 5, pages 612–621, Oct. 2002.

[19] Feil B., Abonyi J. and Szeifert F., "Model order selection of nonlinear input-output models - a clustering based approach," *Journal of Process Control*, Vol. 14, pages 593–602, 2004.

[20] Hwang C. and Rhee F. C.-H., "Uncertain fuzzy clustering: interval type-2 fuzzy approach to c-means", *IEEE Trans. Fuzzy Systems*, Vol. 15, No. 1, pages 107–120, Feb. 2007.

[21] Yu J., Cheng Q. and Huang H., "Analysis of the weighting exponent in the FCM", *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 34, No. 1, pages 634–638, Feb. 2004.

[22] Klawonn F. and Hppner F., "What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier," Proc. *5^{th} Int. Symposium on Intelligent Data Analysis IDA 2003*, Berlin, Germany, pages 254–264, Aug. 2003.

[23] Hathaway R.J. and Bezdek J.C., "Fuzzy c-means clustering of incomplete data," *IEEE Trans. Systems, Man, and Cybernetics — Part B: Cybernetics*, Vol. 31, No. 5, pages 735–744, Oct. 2001.

[24] Bezdek J.C., Keller J., Krishnapuram R. and Pal N.R., *Fuzzy models and algorithms for pattern recognition and image processing*, Boston: Kluwer, 1999.

[25] Dembl D. and Kastner P., "Fuzzy c-means method for clustering microarray data," *Bioinformatics*, Vol. 19, pages 973–980, Aug. 2003.

[26] Zeinali M. and Notash L., "A systematic method of adaptive fuzzy logic modeling, using an improved fuzzy c-means clustering algorithm for rule generation," in *Proc. 2005 IEEE Conf. Control Applications*, Toronto, Canada, pages 84–89, Aug. 2005.

[27] Pal N. R., Bezdek J. C., "On cluster validity for the fuzzy c-means model", *IEEE Trans. Fuzzy Systems*, Vol. 3, pages 370–379, Aug. 1995.

[28] Bernd T., Kroll A. and Schwarz H., "Approximating nonlinear dynamic processes with optimized functional fuzzy models," (in German) in *Proc. GMA Workshop 'Fuzzy Control'*, Dortmund, Germany, pages 179–190, Nov. 1997.

[29] Zimmerschied R., "Identifying nonlinear processes with dynamic local-affine models – means to reduce bias and variance," (in German), Ph.D. dissertation, Univ. of Darmstadt, July 2008.

[30] Babuka R., Fantuzzi C., Kaymak U. and Verbruggen H.B., "Improved inference for Takagi-Sugeno models," in *Proc. 5^{th} IEEE International Conference on Fuzzy Systems*, 1996, pages 1642–1647.

[31] Dunn J. C., "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, Vol. 3, pages 32–57, 1973.

[32] Zimmerschied R. and Isermann R., "Regularization techniques for identification using local-affine models," (in German) at – *Automatisierungstechnik*, Vol. 56, No. 7, pages 339–349, 2008.

[33] Bezdek J. C., Ehrlich R. and Full W., "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, Vol. 10, No. 2-3, pages 191–203, 1984.

[34] McBratney A. B. and Moore A. W., "Application of fuzzy sets to climatic classification," *Agricultural and Forest Meteorology*, Vol. 35, No. 1–4, pages 165–185, 1985.

[35] Choe H. and Jordan J. B., "On the optimal choice of parameters in a fuzzy c-means algorithm," in *Proc. IEEE Int. Conf. on Fuzzy Systems*, San Diego, USA, pages 349-354, March 1992.

[36] Emami M. R., Trksen I. B. and Goldenberg A. A., "Development of a systematic methodology of fuzzy logic modeling," *IEEE Trans. Fuzzy Systems*, Vol. 6, No. 3, pages 346–361, Aug. 1998.

[37] Chen M. S. and Wang S. W., "Fuzzy clustering analysis for optimizing fuzzy membership functions," *Fuzzy Sets and Systems*, Vol. 103, No. 2, pages 239-254, April 1999.

[38] Gao X., Lie J. and Xie W., "Parameter optimization in FCM clustering algorithms," in *Proc. 5^{th} Int. Conf. on Signal Processing Proceedings (WCCC-ICSP'2000)*, Vol. 3, pages 1457–1461, 2000.

[39] Deer P. J.and Eklund P., "A study of parameter values for a Mahalanobis distance fuzzy classifier," *Fuzzy Sets and Systems*, Vol. 137, No. 2, pages 191-213, 2003.

[40] Mller U., "Missing clusters indicate poor estimates or guesses of a proper fuzzy exponent," in *Applications of Fuzzy Sets Theory*. Berlin, Germany: Springer, pages 161–169, 2007.

[41] Yang W., Rueda L. and Ngom A., "A simulated annealing approach to find the optimal parameters for fuzzy clustering microarray data," in *Proc. 25^{th} Int. Conf. Chilean Computer Science Society*, Valdivia, Chile, pages 45–55, 2005.

[42] Okeke F., Karnieli A., "Linear mixture model approach for selecting fuzzy exponent value in fuzzy c-means algorithm," *Ecological Informatics*, Vol. 1, pages 117–124, Jan. 2006.

[43] Kung C. C., Su J. Y., ”A study of cluster validity criteria for the fuzzy c-regression models clustering algorithm,” in *Proc. 2007 IEEE Int. Conf. Systems, Man and Cybernetics*, Montreal, Canada, pages 853–858, Oct. 2007.

[44] Sugeno M., Yasukawa T., “A fuzzy logic approach to qualitative modeling,” *IEEE Trans. Fuzzy Systems*, Vol. 1, No. 1, pages 7–31, Feb. 1993.

[45] Zarandi M. H. F., Esmaeilian M., Zarandi M. M. F., ”A systematic fuzzy modeling for scheduling of textile manufacturing system,” *Int. Journal of Management Science and Engineering Management*, Vol. 2, No. 4, pages 297–308, 2007.

[46] Alata M., Molhim M., Ramini A.: ”Optimizing of fuzzy c-means clustering algorithm using GA,” *World Academy of Science, Engineering and Technology*, Vol. 39, pages 224–229, Mar. 2008.

[47] Hoppner F., Klawonn F. , ”A new approach to fuzzy partitioning,” in *Proc. Joint 9th IFSA World Congr. and 20th NAFIPS Int. Conf.*, Vancouver, Canada, pages 1419–1424, 2001.

[48] Kroll A., Soldan S., “On data-driven Takagi-Sugeno modeling of heterogeneous systems with multidimensional membership functions”, in *Proc. 18th World Congress of the Int. Federation of Automatic Control (IFAC)*, Milano, Italy, pages 4803–4808, August 2011.

[49] Dav R. N. , “Characterization and detection of noise in clustering,” *Pattern Recognition Letters*, Vol. 12, pages 657–664, 1991.

[50] Mnz E. , “Identification and diagnosis of hybrid dynamical systems”, (in German) Ph.D. dissertation, Univ. of Karlsruhe, Universittsverlag Karlsruhe, 2006.

[51] Krishnapuram R., Keller J., ”A possibilistic approach to clustering”, *IEEE Trans. Fuzzy Systems*, Vol. 1, No. 2, pages 98–110, May 1993.