Soft Comput (2007) 11:479–488
DOI 10.1007/s00500-006-0111-5

FOCUS

Visualization of fuzzy clusters by fuzzy Sammon mapping projection: application to the analysis of phase space trajectories

Balazs Feil · Balazs Balasko · Janos Abonyi

Published online: 25 July 2006

Abstract Since in practical data mining problems high-dimensional data are clustered, the resulting clusters are high-dimensional geometrical objects, which are difficult to analyze and interpret. Cluster validity measures try to solve this problem by providing a single numerical value. As a low dimensional graphical representation of the clusters could be much more informative than such a single value, this paper proposes a new tool for the visualization of fuzzy clustering results. By using the basic properties of fuzzy clustering algorithms, this new tool maps the cluster centers and the data such that the distances between the clusters and the datapoints are preserved. During the iterative mapping process, the algorithm uses the membership values of the data and minimizes an objective function similar to the original clustering algorithm. Comparing to the original Sammon mapping not only reliable cluster shapes are obtained but the numerical complexity of the algorithm is also drastically reduced. The developed tool has been applied for visualization of reconstructed phase space trajectories of chaotic systems. The case study demonstrates that proposed FUZZSAMM algorithm is a useful tool in user-guided clustering.

1 Introduction

In our society the amount of data doubles almost every year. Hence, there is an urgent need for a new generation of tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of data. Among the wide range of data-mining tools, the clustering-based computational intelligence methods are becoming increasingly popular, as they are able to learn the mapping of functions and systems, as well as explore structures and classes in data.

Clustering algorithms always fit the clusters to the data, even if the cluster structure is not adequate for the problem. To analyze the adequateness of the cluster prototypes and the number of the clusters, cluster validity measures are used. However since validity measures reduce the overall evaluation to a single number, they cannot avoid a certain loss of information. To avoid this problem, this paper proposes a new tool for the visualization of fuzzy clustering results, since the low dimensional graphical representation of the clusters could be much more informative than such a single value of the cluster validity.

The impact of visualization of fuzzy clustering results has been already realized in Klawonn et al. (2003), when the membership values of the data obtained by the clustering algorithm were simply projected into the input variables, and the resulted plots served for the same purpose as validity measures. Among the wide range of clustering tools, the self-organizing map (SOM) is often visualized by principal component analysis (PCA) and Sammon mapping to give more insight to the structure of high dimensional data. Usually, with the use of these tools the cluster centers (the codebook of the SOM) are mapped into a two-dimensional space (Vesanto 2000). Fuzzy *c*-means cluster analysis has also been combined with similar mappings and successfully applied to map the distribution of pollutants and to trace their sources to access potential environmental hazard on a soil database from Austria (Hanesch et al. 2001).

B. Feil (✉) · B. Balasko · J. Abonyi
Department of Process Engineering, University of Veszprem,
P.O. BOX 158, 8201 Veszprem, Hungary
e-mail: feilb@fmt.vein.hu
URL: http://www.fmt.vein.hu/softcomp

While PCA attempts to preserve the variance of the data during the mapping, Sammon mapping tries to preserve the interpattern distances (Mao and Jain 1995; Pal and Eluri 1998). Hence, this paper focuses on the application of Sammon mapping for the visualization of the results of clustering, as the mapping of the distances is much closer to the task of clustering than preserving the variances. There are two main problems encountered in the application of Sammon mapping to the visualization of fuzzy clustering results:

- The aim of cluster analysis is the classification of objects according to similarities among them, and organizing data into groups. In metric spaces, similarity is often defined by means of distance from a data vector to some prototypical object of the cluster. The prototypes are usually not known beforehand, and are sought by the clustering algorithm simultaneously with the partitioning of the data. The prototypes may be vectors (centers) of the same dimension as the data objects, but they can also be defined as "higher-level" geometrical objects, such as linear or non-linear subspaces or functions. Hence, classical projection methods based on the variance of the data (PCA) or based on the preservation of the Euclidian interpoint distance of the data (Sammon mapping) are not applicable when the clustering algorithm does not use the Euclidian distance norm.
- As Sammon mapping attempts to preserve the structure of high n-dimensional data by finding N points in a much lower q-dimensional data space, such the interpoint distances measured in the q-dimensional space approximate the corresponding interpoint distances in the n dimensional space, the algorithm involves a large number of computations as in every iteration step it requires the computation of $N(N-1)/2$ distances. Hence, the application of Sammon mapping becomes impractical for large N (de Ridder and Duin 1997).

To avoid these problems this paper proposes a new algorithm. By using the basic properties of fuzzy clustering algorithms the proposed tool maps the cluster centers and the data such that the distances between the clusters and the data-points will be preserved. During the iterative mapping process, the algorithm uses the membership values of the data and minimizes an objective function that is similar to the objective function of the original clustering algorithm.

The usefulness of the proposed algorithm is presented by a special case study: trajectories in the reconstructed state space of chaotic systems are visualized. The reason why it is *possible* to demonstrate the approach through this example is that there are methods to estimate the number of state variables of nonlinear (chaotic) systems based on (fuzzy) clustering algorithms (see e.g. (Jiang and Adeli 2003)). The authors developed a general clustering based nonlinear MIMO model identification method, which can be applied for this purpose. However, the main purpose of this paper is to present a new tool to visualize clustering results, therefore the deep description of the model identification method and the state space reconstruction approach is not presented, just a brief introduction is given with whose help the interested reader can step forward. On the other hand, the reason why this example *was* chosen is that this method applies a relatively complex cluster prototype, and it gives a great opportunity to demonstrate the effectiveness of the modified Sammon mapping. Through the results it can be seen that the joint of the two methods can be an effective tool in the analysis of chaotic systems as well.

In the following, in Sect. 2, the new visualization tool will be described. Section 3 contains a case study to illustrate the usefulness of the proposed methods with an introduction to state space reconstruction and crystallizer modeling. Section 4 concludes the paper.

2 FUZZSAM: visualization of fuzzy clusters

2.1 Fuzzy clustering

The objective of clustering is to partition the set of N multivariable data points, $\mathbf{x}_k = [x_{k1}, \ldots, x_{kn}]^{\mathrm{T}}$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$ into c clusters. In fuzzy clustering the objects are not forced to fully belong to one of the clusters; they are rather assigned membership degrees between 0 and 1 indicating their partial memberships. The data set, $\mathbf{X}$, is thus partitioned into c fuzzy subsets. The $N \times c$ matrix $\mathbf{U} = [\mu_{ki}]$ represents the fuzzy partitions, where μ_{ki} denotes the degree of the membership of the kth observation belongs to the ith cluster.

The objective of fuzzy clustering is to minimize the sum of the weighted squared distances, $d^2(\mathbf{x}_k, \eta_i)$, between the data points and the cluster prototypes. The corresponding objective function is

$$J(\mathbf{X}, \mathbf{U}, \boldsymbol{\eta}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ki})^m d^2(\mathbf{x}_k, \eta_i), \tag{1}$$

where $\mathbf{U}$ represents the fuzzy partition matrix of the data set $\mathbf{X}$, $\boldsymbol{\eta}$ represents the parameters of the cluster prototypes (η_i, e.g. cluster centers, covariances, etc.), and

m is a weighting exponent that determines the fuzziness of the resulting clusters, often chosen as $m = 2$. Different cluster shapes can be obtained with different distance measures. Using points as prototypes results in spherical clusters (fuzzy c-means clustering), while using fuzzy covariance matrices results in ellipsoids (e.g. Gustafson–Kessel and Gath–Geva clustering), or with different kinds of prototypes [e.g., fuzzy linear varieties (FCV)], where the clusters are linear subspaces of the feature space.

Cluster validity refers to the problem whether a given fuzzy partition fits to the data. The clustering algorithm always tries to find the best fit for a fixed number of clusters and the parameterized cluster shapes. However this does not mean that even the best fit is meaningful at all. Either the number of clusters might be wrong or the cluster shapes might not correspond to the groups in the data, if the data can be grouped in a meaningful way at all. Cluster validity measures are used to validate the clustering result in general and also to determine the number of clusters (Bezdek 1981). They are useful to estimate the required number of the clusters and decide about merging of two clusters, etc. However, a low dimensional graphical representation of the clusters could be much more informative than such a single value. This is especially true when the user wants to verify whether a suitable cluster prototype and distance measure were chosen (e.g., FCM vs. Gustafson–Kessel clustering, etc.). In the remaining part of this section, we propose a new tool for the visualization of fuzzy clustering results based on Sammon mapping.

2.2 Sammon mapping

The Sammon mapping is a multi-dimensional scaling method. It is a well-known procedure for mapping data from a high n-dimensional space onto a lower q-dimensional space by finding N points in the q-dimensional data space, such that the interpoint distances $d^*_{ij} = d^*(\mathbf{y}_i, \mathbf{y}_j)$ in the q-dimensional space approximate the corresponding interpoint distances $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ in the n-dimensional space. This is achieved by minimizing an error criterion, called the Sammon's stress, E:

$$E = \frac{1}{\lambda} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\left(d_{ij} - d^*_{ij}\right)^2}{d_{ij}}, \tag{2}$$

where $\lambda = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}$.

The minimization of E is an optimization problem in Nq variables y_{il}, $i = 1, 2, \ldots, N$, $l = 1, 2, \ldots, q$, as $\mathbf{y_i} = [y_{i1}, \ldots, y_{iq}]^{\mathrm{T}}$. Sammon applied the method of steepest decent to minimizing this function. Introduce the estimate of y_{il} at the tth iteration

$$y_{il}(t+1) = y_{il}(t) - \alpha \left[\frac{\frac{\partial E(t)}{\partial y_{il}(t)}}{\frac{\partial^2 E(t)}{\partial^2 y_{il}(t)}} \right], \tag{3}$$

where α is a nonnegative scalar constant (recommended $\alpha \simeq 0.3 - 0.4$), i.e., the step size for gradient search in the direction of

$$\begin{aligned} \frac{\partial E(t)}{\partial y_{il}(t)} &= -\frac{2}{\lambda} \sum_{k=1, k\neq i}^{N} \left[\frac{d_{ki} - d^*_{ki}}{d_{ki} d^*_{ki}} \right] (y_{il} - y_{kl}), \\ \frac{\partial^2 E(t)}{\partial^2 y_{il}(t)} &= -\frac{2}{\lambda} \sum_{k=1, k\neq i}^{N} \frac{1}{d_{ki} d^*_{ki}} \left[(d_{ki} - d^*_{ki}) \right. \\ &\quad \left. - \left(\frac{(y_{il} - y_{kl})^2}{d^*_{ki}} \right) \left(1 + \frac{d_{ki} - d^*_{ki}}{d_{ki}} \right) \right]. \end{aligned} \tag{4}$$

It is not necessary to maintain λ for a successful solution of the optimization problem, since the minimization of $\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(d_{ij} - d^*_{ij}\right)^2 / d_{ij}$ gives the same result.

When the gradient-descent method is applied to search for the minimum of Sammon's stress, a local minimum in the error surface can be reached. Therefore a significant number of runs with different random initializations may be necessary. Nevertheless, the initialization of y can be based on information which is obtained from the data, such as the first and second norms of the feature vectors or the principal axes of the covariance matrix of the data (Mao and Jain 1995).

A disadvantage of the original Sammon mapping is that when a new data point has to be mapped, the whole mapping procedure has to be repeated (Pal and Eluri 1998). This means computational load, because in each iteration $N \times (N-1)/2$ distances as well as the error derivatives, must be calculated. Hence, the application of Sammon mapping becomes impractical for large N. To avoid this problem, in the following subsection we introduce some modifications in order to tailor Sammon mapping for the visualization of fuzzy clustering results.

2.3 Modified Sammon mapping

By using the basic properties of fuzzy clustering algorithms where only the distance between the data points and the cluster centers are considered to be important, the modified algorithm takes into account only $N \times c$ distances, where c represents the number of clusters, weighted by the membership values similarly to Eq. (1):

$$E_{\mathrm{fuzz}} = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ki})^m \left(d(\mathbf{x}_k, \eta_i) - d^*_{ki} \right)^2 \tag{5}$$

where $d(\mathbf{x}_k, \eta_i)$ represents the distance between the $\mathbf{x}_k$ datapoint and the η_i cluster center measured in the original n-dimensional space, while $d^*_{ki} = d^*(\mathbf{y}_k, \mathbf{z}_i)$ represents the Euclidian distance between the projected cluster center $\mathbf{z}_i$ and the projected data $\mathbf{y}_k$. This means that in the projected space, every cluster is represented by a single point, regardless of the form of the original cluster prototype, η_i. The application of the simple Euclidian distance measure increases the interpretability of the resulted plots (typically in two dimensions, although three-dimensional plots can be used as well). If the type of cluster prototypes is properly selected, the projected data will fall close to the projected cluster center represented by a point resulting in an approximately spherically shaped cluster.

The resulting algorithm is similar to the original Sammon mapping, but in this case in every iteration after the adaptation of the projected data points, the projected cluster centers are recalculated based on the weighted mean formula of the fuzzy clustering algorithms:

- **[Input]**: Desired dimension of the projection, usually $q = 2$, the original data set, $\mathbf{X}$; and the results of fuzzy clustering: cluster prototypes, η_i, membership values, $\mathbf{U} = [\mu_{ki}]$, and the distances $D = \left[d_{ki} = d(\mathbf{x}_k, \eta_i)\right]_{N\times c}$.
- **[Initialize]** the projected data points by $\mathbf{y}_k$ PCA based projection of $\mathbf{x}_k$, and compute the projected cluster centers by

$$\mathbf{z}_i = \frac{\sum_{k=1}^{N} (\mu_{ki})^m \mathbf{y}_k}{\sum_{k=1}^{N} (\mu_{ki})^m} \tag{6}$$

  and compute the distances with the use of these projected points $D^* = \left[d^*_{ki} = d(\mathbf{y}_k, \mathbf{z}_i)\right]_{N\times c}$.
- **[While]** ($E_{fuzz} > \varepsilon$) and (t$\leq$ maxstep)
  {for ($i = 1 : i \leq c : i++$
  {for ($j = 1 : j \leq N : j++$
  $\left\{ \text{Compute } \frac{\partial E(t)}{\partial y_{il}(t)}, \frac{\partial^2 E(t)}{\partial^2 y_{il}(t)}, \Delta y_{il} = \Delta y_{il} + \left[\frac{\frac{\partial E(t)}{\partial y_{il}(t)}}{\frac{\partial^2 E(t)}{\partial^2 y_{il}(t)}} \right] \right\}$
  }
  $y_{il} = y_{il} + \Delta y_{il}, \forall i = 1, \ldots, N,\ l = 1, \ldots, q$
  Compute $\mathbf{z}_i = \sum_{k=1}^{N} (\mu_{ki})^m \mathbf{y}_k / \sum_{k=1}^{N} (\mu_{ki})^m$
  $D^* = \left[d^*_{ki} = d(\mathbf{y}_k, \mathbf{z}_i)\right]_{N\times c}$
  }
  Compute E_{fuzz} by Eq. (5).

where the derivatives are

$$\frac{\partial E(t)}{\partial y_{il}(t)} = -2 \sum_{j=1}^{c} \sum_{k=1}^{N} \left[\frac{d_{kj} - d^*_{kj}}{d^*_{kj}} (\mu_{kj})^m \right] (y_{il} - z_{jl}),$$

$$\frac{\partial^2 E(t)}{\partial^2 y_{il}(t)} = -2 \sum_{j=1}^{c} \sum_{k=1}^{N} \frac{(\mu_{kj})^m}{d^*_{kj}} \left[(d_{kj} - d^*_{kj}) - \frac{(y_{il} - z_{jl})^2}{d^*_{kj}} \left(1 + \frac{d_{kj} - d^*_{kj}}{d^*_{kj}} \right) \right]. \tag{7}$$

The resulted two dimensional plot of the projected data and the cluster centers is easily interpretable since it is based on a normal Euclidian distance measures between the cluster centers and the data points. Based on these mapped distances, the membership values of the projected data can be also plotted based on the classical formula of the calculation of the membership values:

$$\mu^*_{ki} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d^*(\mathbf{x}_k, \eta_i)}{d^*(\mathbf{x}_k, \eta_j)} \right)^{\frac{2}{(m-1)}}}. \tag{8}$$

Of course, the resulting 2D plot will only approximate the original high dimensional clustering problem. The quality of the approximation can easily be evaluated based on the mean square error of the original and the recalculated membership values.

$$P = \|\mathbf{U} - \mathbf{U}^*\|, \tag{9}$$

where $\mathbf{U}^* = [\mu^*_{ki}]$ represents the matrix of the recalculated memberships.

Of course there are other tools to get information about the quality of the mapping of the clusters. For example, the comparison of the cluster validity measures calculated based on the original and mapped membership values can also be used for this purpose.

2.4 Advantages of fuzzy Sammon mapping

The method presented above has been tested by several simple synthetic and other well-known multidimensional benchmark examples (Iris, Wisconsin etc.) and with several clustering algorithm with different prototypes (fuzzy c-means, Gustafson–Kessel and Gath–Geva algorithms). It has to be mentioned again that the proposed tool is able to handle various cluster prototypes and exactly this is its main advantage. The main purpose of the proposed method is to visualize the (fuzzy) clustering results and not the multidimensional data themselves. It follows from this that

1. The distance measure of the cluster prototype is "transformed" into Euclidian distance in the projected two dimensional space. Therefore a difficult cluster prototype can be represented by a single point in two dimensions *regardless of the form of the original cluster prototype*, η_i. The application of

the simple Euclidian distance measure increases the interpretability of the resulted plots.

2. In case of a properly selected cluster prototype (e.g. there are ellipsoid shape clusters with the same size, the number of clusters are known a priori and the applied clustering method is Gustafson-Kessel algorithm), then the projected data will fall close to the projected cluster center represented by a point resulting in an approximately spherically distributed cluster. If the resulted clusters in two dimensions are distorted, then it can be determined that the applied cluster prototype cannot be used to represent a 'real' cluster in the multidimensional space (or the number of clusters is wrong). Hence, the proposed method can be used to validate the clustering results by a human inspector.

For the sake of comparison, the data and the cluster centers are also projected by PCA and the standard Sammon projection. Beside the visual inspection of the results the mean square error of the recalculated membership values, P, see Eq. (9), the difference between the original and the recalculated partition coefficient (one of the well known cluster validity measures (Bezdek 1981)), and the Sammon stress coefficient Eq. (2) were analysed. It is proven that fuzzy Sammon mapping gives (far) better results than the other two visualization methods (for more details see Kovacs and Abonyi 2002).

3 Case study

In this section the advantages of the proposed method are presented by a complex problem to show how effective the method can be: the phase space trajectories of a chaotic crystallizers are analyzed by clustering algorithm and the result are visualized. In the following — for the sake of completeness –, the state space reconstruction approach (Sect. 3.1) and the model of a continuous crystallizer (Sect. 3.2) is discussed, the results and the discussion can be found in Sect. 3.3. It has to be mentioned that the main purpose of this paper is to present a new visualization tool, the methods discussed in Sect. 3.1 and 3.2 were already published in the literature. Section.3.3 can be understood without deep elaboration of the details in Sects. 3.1 and 3.2.

3.1 State space reconstruction

Clustering is applied in the reconstructed space defined by the lagged measured variables x_k:

$$\mathbf{x}_k = [x_k, x_{k-\tau}, \ldots, x_{k-\tau(d_e-1)}]^{\mathrm{T}}. \tag{10}$$

The number of components d_e is usually referred to as *embedding dimension*. Although, the data samples are embedded in a d_e dimensional space, they do not necessarily fill that space. The system defines a nonlinear hypersurface in which the state variables reside. The dimension of this hyper-surface is referred to as *intrinsic, topological or local dimension*, d_l. First, the lag time τ is chosen by using the average mutual information or autocorrelation function. The key step of the approach is the clustering of the data. A model-based clustering algorithm was developed for the identification of operating regimes and parameters of Gaussian mixture models using the expectation maximization (EM) method (Abonyi et al. 2004). The embedding dimension is inferred from the one-step ahead prediction performance of the local models. The method enables us to estimate the intrinsic dimension of the reconstructed space simultaneously by analyzing the eigenvalues of the fuzzy cluster covariance matrices.

To identify a model that can be used for prediction of a time series global, local and semi-local methods can be used Lillekjendlie et al. (1994). The modeling framework that is based on combining a number of local models, where each local model has a predefined operating region in which it is valid, is called *operating regime based model* (Babuška and Verbruggen 1991). This approach is advantageous in the modeling of complex nonlinear systems, since while it may not be possible to find a model that is universally applicable to describe the unknown MIMO system, with the application of the divide and conquer paradigm it is possible to decompose the complex problem into a set of smaller identification problems where standard linear models give satisfactory performance.

This operating-regime based model is formulated as follows:

$$\begin{aligned}\mathbf{x}_{k+1} &= \sum_{i=1}^{c} \beta_i(\mathbf{x}_k) \underbrace{(\mathbf{A}_i\mathbf{x}_k + \mathbf{b}_i)}_{f_i} \\ &= \sum_{i=1}^{c} \beta_i(\mathbf{x}_k) \left[\mathbf{x}_k^{\mathrm{T}}\, 1\right] \boldsymbol{\theta}_i^{\mathrm{T}} \\ &= \sum_{i=1}^{c} \beta_i(\mathbf{x}_k) f_i(\mathbf{x}_k, \boldsymbol{\theta}_i),\end{aligned} \tag{11}$$

where the function $\beta_i(\mathbf{x}_k)$ describes the operating region, and $\boldsymbol{\theta}_i = [\mathbf{A}_i\, \mathbf{b}_i]$ is the parameter matrix of the ith local model (rule). The output of the ith local model is denoted by $\mathbf{x}_{k+1}^i = f_i(\mathbf{x}_k, \boldsymbol{\theta}_i)$.

The main advantage of the developed framework is its transparency. The operating regions of the local models can be represented by fuzzy sets Babuška

and Verbruggen (1991). This representation is appealing, since many systems change their behavior smoothly as a function of the operating point, and the soft transition between the regimes introduced by the fuzzy set representation captures this feature in a natural way.

The bottleneck of the data-driven identification of operating regime based models is the identification of the functions that can represent the operating regimes of the models (membership functions in the terminology of fuzzy models), which requires nonlinear optimization. In this paper Gaussian functions are used to represent the operating regimes of the linear models:

$$\beta_i(\mathbf{x}_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}_k - \mathbf{v}_i)^{\mathrm{T}}\mathbf{F}_i^{-1}(\mathbf{x}_k - \mathbf{v}_i)\right)}{\sum_{j=1}^{c}\exp\left(-\frac{1}{2}(\mathbf{x}_k - \mathbf{v}_j)^{\mathrm{T}}\mathbf{F}_j^{-1}(\mathbf{x}_k - \mathbf{v}_j)\right)}, \quad (12)$$

where $\mathbf{v}_i$ is the center and $\mathbf{F}_i$ is the covariance matrix of the multivariate Gaussian function.

A new clustering-based technique for the identification of these parameters was developed by the authors in Abonyi et al. (2004). The objective is to partition the identification data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$ into c clusters to reveal the underlying structure of the data. The patterns belong to clusters with degrees that are in inverse proportion to the distance from the respective cluster prototypes. The basic idea of the proposed algorithm is to define the cluster prototype such that it locally approximates the MIMO function $\mathbf{x}_{k+1} = \mathbf{f}_r(\mathbf{x}_k)$. In this way, the algorithm simultaneously partitions the data (i.e., identifies the operating regimes of the local models) and determines the cluster prototypes (i.e., identifies the local model parameters). The fuzzy partition is represented by the $\mathbf{U} = [\mu_{ki}]_{N\times c}$ matrix, where the element μ_{ki} represents the membership degree of $\mathbf{x}_k$ in cluster i.

The clustering is based on minimizing the following cost function:

$$J(\mathbf{X}, \mathbf{U}, \eta) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ki})^m\, d^2(\mathbf{x}_k, \eta_i), \quad (13)$$

where the squared distance $d^2(\mathbf{x}_k, \eta_i)$ is given by the probability that the data belong to a given cluster:

$$\begin{aligned}\frac{1}{d^2(\mathbf{x}_k, \eta_i)} &= p(\eta_i)p(\mathbf{x}_k|\eta_i)p(\mathbf{x}_{k+1}|\mathbf{x}_k, \eta_i)\\ &= p(\eta_i)\underbrace{\frac{1}{(2\pi)^{\frac{d_e}{2}}\sqrt{|\mathbf{F}_i|}}\exp\left(-\frac{1}{2}(\mathbf{x}_k - \mathbf{v}_i)^{\mathrm{T}}\mathbf{F}_i^{-1}(\mathbf{x}_k - \mathbf{v}_i)\right)}_{p(\mathbf{x}_k|\eta_i)}\\ &\quad\times\underbrace{\frac{1}{(2\pi)^{\frac{d_e}{2}}\sqrt{|\mathbf{P}_i|}}\exp\left(-\frac{1}{2}(\mathbf{x}_{k+1} - \mathbf{x}_{k+1}^i)^{\mathrm{T}}\mathbf{P}_i^{-1}(\mathbf{x}_{k+1} - \mathbf{x}_{k+1}^i)\right)}_{p(\mathbf{x}_{k+1}|\mathbf{x}_k, \eta_i)}.\end{aligned} \quad (14)$$

The term $p(\eta_i)$ represents the a priori probability of the ith cluster defined by the η_i set of parameters. The Gaussian distribution $p(\mathbf{x}_k|\eta_i)$ defines the domain of influence of a cluster (see Eq. (12)). The third term is based on the performance of the local linear models where $\mathbf{P}_i$ is the weighted covariance matrix of the modeling error of the ith local model. The minimization of the functional (Cao 1997) represents a non-linear optimization problem which can be solved by the method developed by the authors (Abonyi et al. 2004). The cluster prototype is $\eta_i = [p(\eta_i), \mathbf{v}_i, \mathbf{F}_i, \mathbf{P}_i, \boldsymbol{\theta}_i]$ in this case.

3.2 Model of the crystallizers

Most of the developments in the field of nonlinear dynamics assume that one has a complete description of the dynamic system under consideration. The practical application of these results, in principle, requires the simultaneous measurement of all state variables. In the case of crystallizers, however, in-line measurement of crystal size distribution, that is one of the most important properties of crystallization processes, is a difficult task so that application of the finite dimensional moment equation model, computed from the crystal size distribution, often appears to be troublesome. Instead, observing a time series of one or more observables of the system the dynamics of the unknown deterministic finite dimensional system can be reconstructed from this scalar time series as it was shown by Takens (1981); Sauer et al. (1991). A number of algorithms have been proposed to handle this problem numerically (Ashkenazy 1999; Ataei et al. 2004; Cao 1997; Jiang and Adeli 2003) but the method presented above, based on fuzzy clustering of chaotic time series Abonyi et al. (2004), allows solving three tasks simultaneously: selection of the embedding dimension, estimation of the intrinsic dimension, and identification of a model that can be used for prediction of chaotic time series.

Consider two continuous isothermal MSMPR (mixed suspension, mixed product removal) crystallizers connected in cascade series where the first crystallizer is forced by sinusoidally varied solute input, while the second crystallizer is forced with the output solute signal of the first one. Let the crystallizers be identical in the sense that all kinetic and process parameters are of the same value. Further, let us assume that the working volumes are constant during the course of the operation, all new crystals are formed at a nominal size $L_n \approx 0$, crystal breakage and agglomeration are negligible, no growth rate fluctuations occur, the overall linear growth rate of crystals is size-independent and has the form of the power law expression of supersaturation, and the nucleation rate is described by Volmers model.

The population balance model is an adequate mathematical description of crystallization processes. This model consists of a mixed set of ordinary and partial integral-differential equations even in the simplest case of MSMPR crystallizers, and the state space of a crystallizer is given by the Descartes product $\mathbb{R}^k \times \mathbb{N}$, where k is a positive natural number, of some vector space $\mathbb{R}^k$ of concentrations and temperatures, and the function space $\mathbb{N}$ of population density functions. Consideration of dynamical problems of crystallizers in this product space, however, seems to be quite complex, so that we usually concentrate on a reduced case, approximating the distributed parameter system by a finite-dimensional state space model based on the moments of crystal size. Taking into account only the first four leading moments, the resulted space even in the simplest case of isothermal MSMPR crystallizers becomes six-D. It is suitable for studying the dynamic phenomena of crystallizers (Abonyi et al. 2002; Lakatos 1994), but it often appears to be too complex to apply for their model based control (Abonyi et al. 2002). Consequently, it is reasonable to generate an appropriate reduced-dimensional model e.g. for control purposes. Then, the moment equation model of this system of crystallizers is a closed set of six ordinary differential equations by each crystallizer, and their connection can be described by algebraic equations. There are ranges of parameters in which the system behavior is chaotic and it can also be observed by real crystallization processes.

In the next section it is presented how the fuzzy clustering-based algorithm can be applied for reduction of the moment equation model of continuous isothermal crystallizers using the chaotic time series generated under some operation conditions, and how the visualization can be used in this case.

3.3 Results and discussion

In this analysis, the zero order moment of the second crystallizers is used to predict the embedding and local dimension of the system with simultaneous identification of a model that can be used for prediction. Using only this single time-series, the embedding and local dimension want to be estimated by the method described in Sect. 3.1. By the simulation, the sampling time was equal to 0.01 times of the dimensionless time unit ξ. By the state-space reconstruction, the sampling time was ten times greater, so as to reduce the time and memory demand. The other parameters of the method: number of points: 1,500; number of clusters: $c = 10$; termination tolerance: $\epsilon = 10^{-4}$ and the weighting exponent: $m = 2$. According to our experience, the proposed approach is quite robust with respect to the choice of the clustering algorithm parameters.

The lag time was chosen as the first minimum of the average mutual information, and in this case it is equal to 0.6ξ. The embedding dimension was run from 1 to 10, and by each value the model identification was evaluated to get the one-step-ahead prediction error. By $d_e = 4$ the mean square error is low enough and significant reduction in the model error cannot be determined by greater dimensions, either. After the correct embedding dimension is chosen, it is possible to estimate the local dimension by the analysis of the rate of the cumulative and the total sum of eigenvalues weighted by the a priori possibility of the clusters. It can be determined that a 2-dimensional subspace of the reconstructed 4-dimensional state space contains the trajectories with high, 95% possibility.The disadvantage of this method is that it is only able to estimate integer local dimensions. To avoid this, fractal dimension of the trajectories can also be computed. In this case the so-called correlation dimension (Grassberger and Procaccia 1983) is equal to 2.0009 which confirms the result of the clustering based method.

On Fig. 1, the measured data are depicted with solid line embedded in the 4-dimensional reconstructed state space. To validate the model in some sense, a free-run simulation was evaluated whose results can also be seen on Fig. 1 plotted with dotted line. On the left side of Fig. 1 the first, on the right the second three dimensions are depicted. (To visualize all of four dimensions a scatterplot matrix can be used, which is a 4×4 matrix in this case.) It can be seen that although the system is chaotic, the trajectories are similar and take the same subspace, so the approximation is very good. The predicted one dimensional data are depicted on Fig. 2. It can be seen that there are ranges in the space that cannot be modeled as well as other ranges, it causes that the predicted trajectories draw away from the measured ones.

It is possible to depict the 'relationship' and 'order' of clusters in the 4-dimensional space in some sense. In the analyzed case there is a dynamic system so it can be computed how the data points are 'wandering' in the state space from one cluster to another. This enables us to order clusters. The procedure is the following: the points with high probability $\{k\colon \mu_{ki} > 0.99$ by the current cluster $i\}$ should be found and determined by which cluster the membership value $\mu_{k+m,j}, j \neq i$ will be first high again 'in the future', $m > 0$. In this way it can be computed how much percentage of the points in the current cluster wanders to which clusters. In Fig. 3 a this 'path' can be seen. In this figure only the paths with ratio greater than 40% are depicted because of transparency (the width of arrows is proportional to the ratio). It has

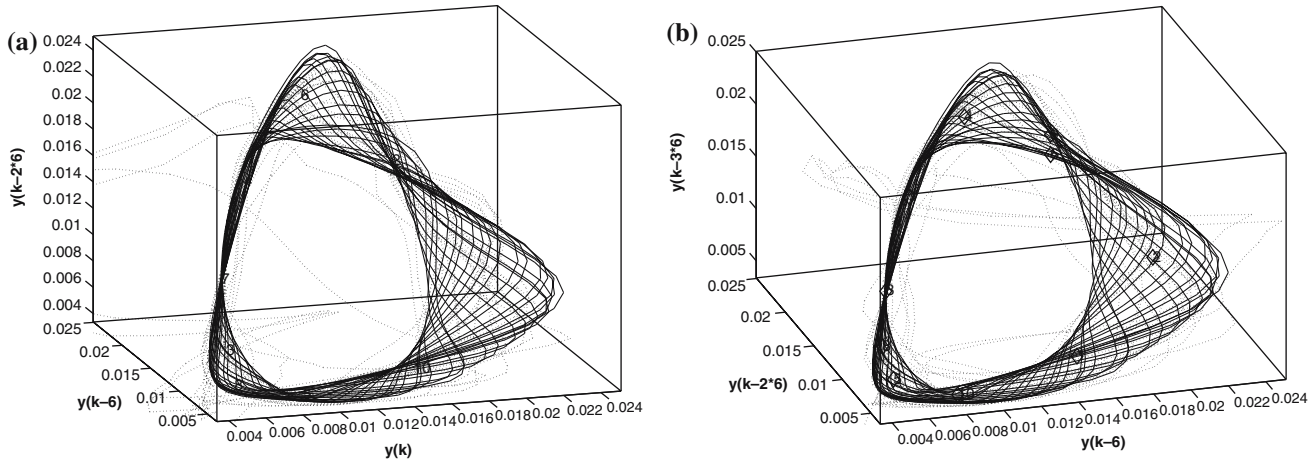

Fig. 1 Free run simulation in the state space

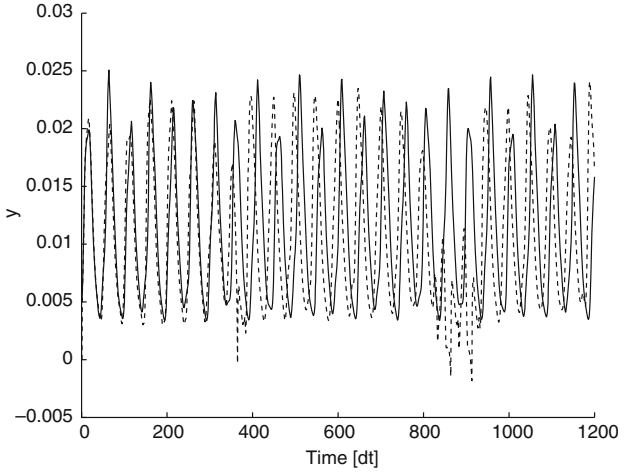

Fig. 2 Free run simulation

Table 1 The a priori probability of clusters

Cluster	Probability (%)
1	12.87
2	10.40
3	11.14
4	12.54
5	10.32
6	6.95
7	11.91
8	8.30
9	7.08
10	8.45

to be mentioned that the size of the clusters is also very important from the viewpoint of the path because relatively small clusters have the same effect as large ones. However, in this case this effect is negligible because the size of clusters (the *a priori* probability $p(\eta_i)$) are nearly the same as it can be seen in Table 1. This kind of figure can be useful to analyze the relationship of clusters in the original multidimensional space but it cannot be useful to visualize the clustering results because it does not contain any information about that. For this purpose a visualization technique has to be used. At first, consider the result of PCA that is depicted in Fig. 3 b. It can be determined that the order of the cluster is completely the same as in Fig. 3 a and the 'torsion' of the trajectories (see in Fig. 1) can also be observed.

PCA is a linear method. It means that it finds the 2-dimensional linear subspace of this 4-dimensional space that fits data the best. If the data do not lie close to a linear 2-dimensional subspace, the mapping gives bad results. In this case, it is advisable to apply nonlinear techniques. Results given by Sammon mapping can be seen in Fig. 4: on the left initialized by PCA and on the right with random initialization. It can be determined that these results are nearly the same and they are very similar to PCA results, only the data points and clusters have turned round in some directions or have been mirrored. It is caused by the procedure of Sammon mapping itself because it tries to preserve the distance between each data pair, so there is no 'fix point' in the projected space. Because of the similarity of results from linear PCA and nonlinear Sammon mapping, it can be determined that not only the local dimension of the trajectories is equal to 2, but also this subspace is nearly a linear one. It can also be represented by index-numbers (Table 2). The numerical results summarized in Table 2 show that the proposed FUZZSAM tool outperforms the linear method and the classical Sammon projection tools. The P error of the membership values between are much smaller, and the original and recalculated cluster validity measures (the partition coefficient in this section) F and F^* are similar when the projection is based on the proposed FUZZSAM mapping. It also shows that PCA

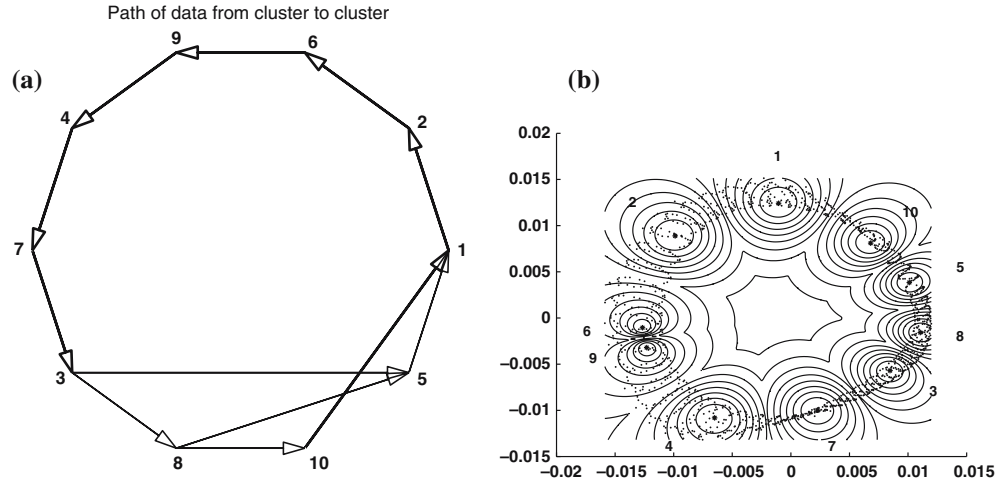

Fig. 3 Projection by PCA

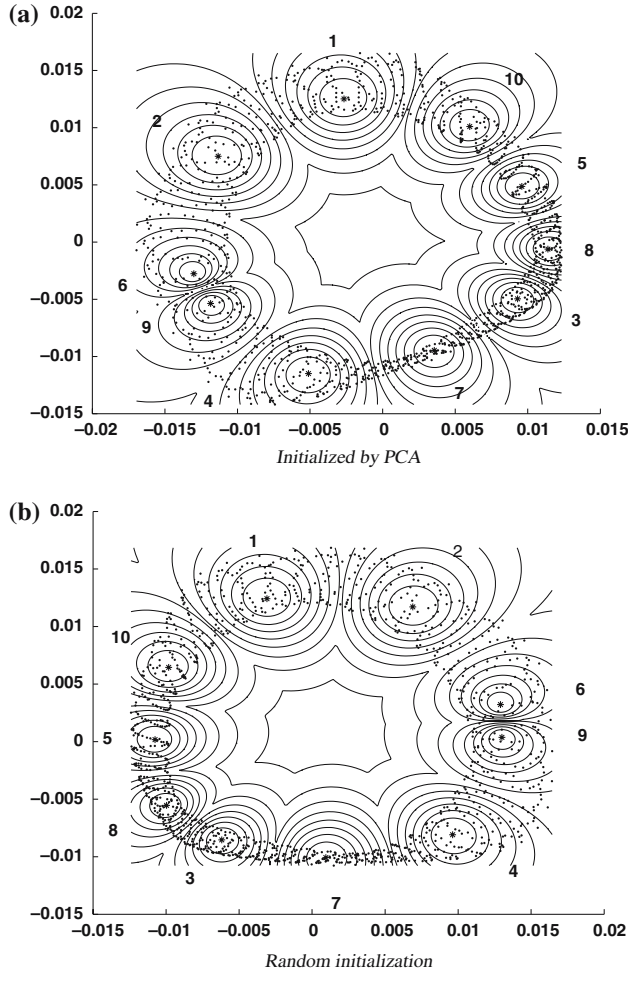

Fig. 4 Projection based on Sammon mapping

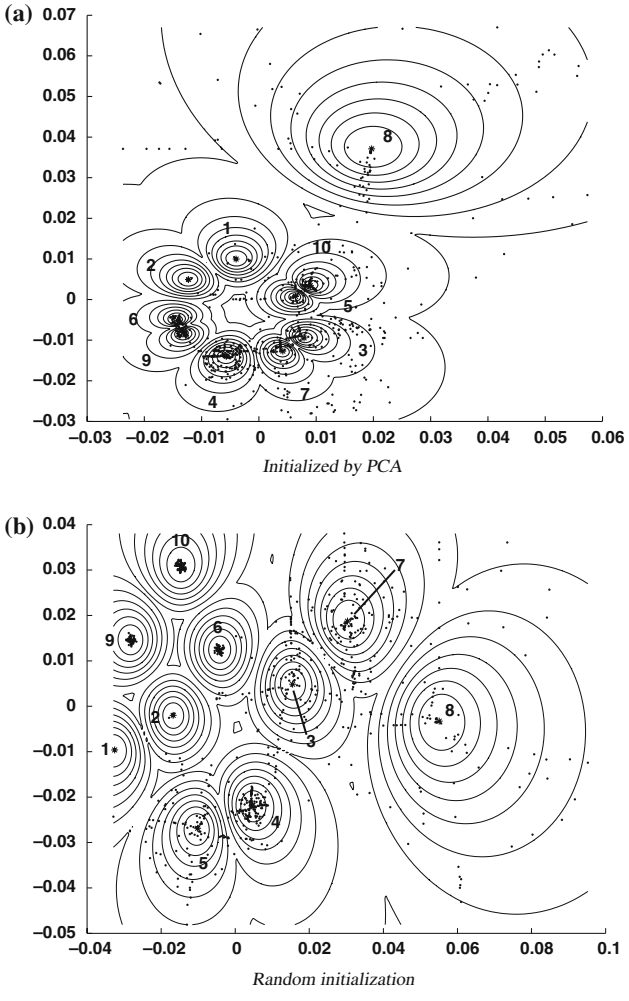

Fig. 5 Projection based on fuzzy Sammon mapping

and Sammon mapping give nearly the same result in the viewpoint of visualization of clustering results.

The proposed fuzzy Sammom mapping gives completely different results as it can be seen in Fig. 5. As it was the case in Fig. 4, the results on the left is initialized by PCA and the ones on the right given by random initialization. Figure 5 a shows similar structure of clusters

Table 2 Comparison of the performance of the mappings (line of *FUZZSAM (rand)* contains the average of results by 20 random initialization

Method	*P*	*F*	*F**	*E*
PCA	0.0859	0.9911	0.5920	0.0035
SAMMON	0.0813	0.9911	0.5717	0.0018
FUZZSAM (PCA)	0.0625	0.9911	0.7004	2.6737
FUZZSAM (rand)	0.0408	0.9911	0.7727	10.1837

to PCA results (Fig. 3). The order of clusters is similar as well (see e.g. 1-2-6-9-4-7), but in the latter figure it can be seen which clusters are less compact, so less accurate: mainly cluster 8, but there are data points far from cluster 3, 5 and 7 as well. It has to be mentioned that also Fig. 3 (a) shows cluster 8 as a 'critical point'. Fuzzy Sammon mapping with random initialization gives similar results but no such structure of clusters as in the previous examples. It can be seen that there are data points scattered in the range of cluster 3, 7 and 8, since the other clusters are very compact with similar a priori probability (Table 1). The proposed FUZZSAM tool outperforms the linear method and the classical Sammon projection tools. It is interesting that FUZZSAMM with random initialization gives better results than with PCA initialization, because in the latter case PCA 'forces' FUZZSAMM algorithm to preserve the original structure of clusters in some sense.

4 Conclusions

This paper presented a tool that gives the user feedback about the result of fuzzy clustering. The FUZZSAM method generates two dimensional informative plots about the quality of the cluster prototypes and the required number of clusters. This tool uses the basic properties of fuzzy clustering algorithms to map the cluster centers and the data such that the distances between the clusters and the data-points are preserved. The numerical results show superior performance over PCA and the classical Sammon projection tools. As the case study proves, the resulted tool is useful not only for interactive data mining, but also for analysis of dynamical systems.

Acknowledgements The support of the Cooperative Research Center (2005-III-1), the Hungarian Science Foundation (T037600 and T049534) is gratefully acknowledged.

References

Abonyi J, Lakatos BG, Ulbert ZS (2002) Modelling and control of isothermal crystallizers by self organising maps. Chem Eng Trans 1(3):1149

Abonyi J, Feil B, Babuška R (2004) State-space reconstruction and prediction of chaotic time series based on fuzzy clustering. IEEE SMC 2004, October 10–13

Ashkenazy Y (1999) The use of generalized information dimension in measuring fractal dimension of time series. Physica A 271:427–447

Ataei M, Lohmann B, Khaki-Sedigh A, Lucas C (2004) Model based method for estimating an attractor dimension from uni/multivariate chaotic time series with application to Bremen climatic dynamics. Chaos Solitons Fractals 19:1131–1139

Babuška R, Verbruggen HB (1991) Fuzzy set methods for local modeling and identification. Multiple model approaches to nonlinear modeling and control. Taylor & Francis, London, pp 75–100

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York

Cao C, (1997) Practical method for determining the minimum embedding dimension of a scalar time series. Physica D 110:43–50

Casdagli M (1989) Nonlinear prediction of chaotic time series. Physica D 35:335–356

Farmer JD, Sidorowich JJ (1987) Predicting chaotic time series. Phys Rev Lett 59(8):845–848

Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208

Hanesch M, Scholger R, Dekkers MJ (2001) The application of fuzzy c-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites. Phys Chem Earth 26:885–891

Hunter Jr, Norman F (1992) Nonlinear prediction of speach signals. In: Casdagli M, Eubanks S (eds) Nonlinear modelling and forecasting. Addison-Wesley, pp 467–492

Jiang X, Adeli H (2003) Fuzzy clustering approach for accurate embedding dimension identification in chaotic time series. Integrated Comput Aided Eng 10:287–302

Klawonn F, Chekhtman V, Janz E (2003) Visual inspection of fuzzy clustering results. Adv Soft Comput Eng Des Manuf 65–76

Kovacs A, Abonyi J (2002) Vizualization of fuzzy clustering results by modified Sammon mapping. In: Proceedings of the 3rd international symposium of Hungarian researchers on computational intelligence, pp 177–188

Lakatos BG (1994) Stability and dynamics of continuous crystallizers. Comput Chem Eng 18:427–431

Lakatos BG (1995) Sustained oscillation in isothermal CMSMPR crystallizers: effect of size-dependent crystall growth rate. Sapundzhiev TsJ (1995) ACH Models Chem 132:379–394

Lerner B, Guterman H, Aladjem M, Dinstein I, Romem Y (1998) On pattern classification with Sammon's nonlinear mapping an experimental study. Pattern Recognit 31:371–381

Lillekjendlie B, Kugiumtzis D, Christophersen N (1994) Chaotic time series — Part II: System identification and prediction. Modeling Identification Control 15(4):225–243

Mao J, Jain AK (1995) Artifical neural networks for feature extraction and multivariate data projection. IEEE Trans Neural Netw 629–637

Pal NR, Eluri VK (1998) Two efficient connectionist schemes for structure preserving dimensionality reduction. IEEE Trans Neural Netw 9:1143–1153

de Ridder D, Duin RPW (1997) Sammon's mapping using neural networks: a comparison. Pattern Recognit Lett 18:1307–1316

Sauer T, Yorke JA, Casdagli M (1991) Embedology. J Stat Phys 65:579–615

Takens F (1981) Detecting strange attractor in turbulence. In: Rand DA, Young LS (eds) Dynamical systems and turbulence. Springer, Berlin Heidelberg New York, pp 366–381

Vesanto J (2000) Neural network tool for data mining: SOM Toolbox. In: Proceedings of symposium on tool environments and development methods for intelligent systems (TOOLMET2000), pp 184–196