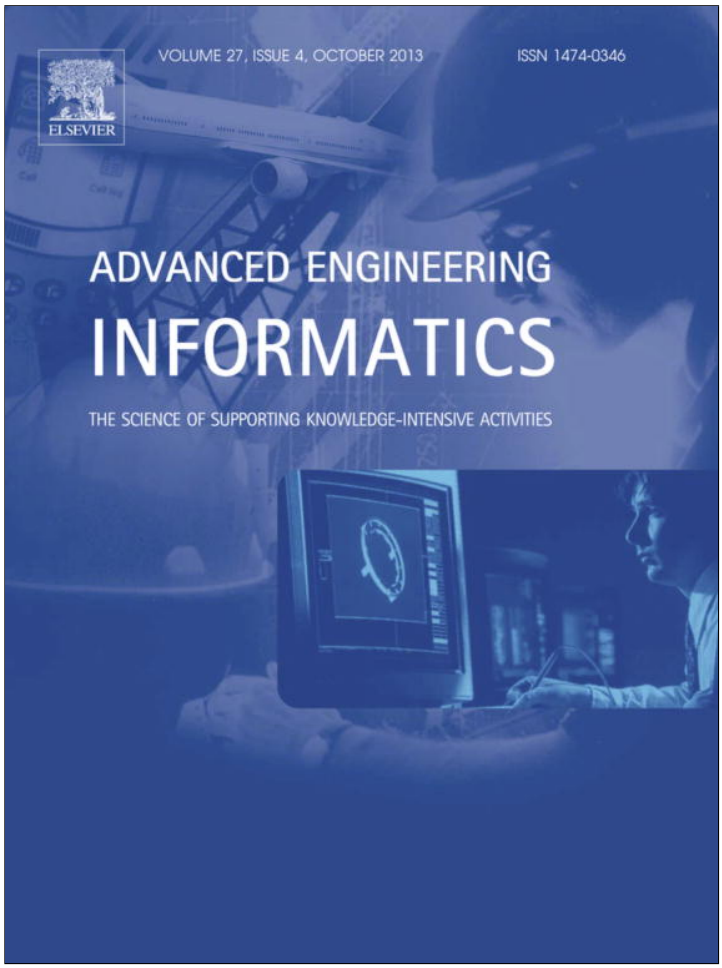

Advanced Engineering Informatics 27 (2013) 615–624

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Data mining and knowledge discovery in materials science and engineering: A polymer nanocomposites case study ☆

O. AbuOmar [a,b], S. Nouranian [b], R. King [a,b,*], J.L. Bouvard [c], H. Toghiani [d], T.E. Lacy [e], C.U. Pittman Jr. [f]

[a] Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[b] Center for Advanced Vehicular Systems (CAVS), Mississippi State, MS 39762, USA
[c] Center for Material Forming (CEMEF), Mines ParisTech, 06904 Sophia Antipolis Cedex, France
[d] The Dave C. Swalm School of Chemical Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[e] Department of Aerospace Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[f] Department of Chemistry, Mississippi State University, Mississippi State, MS 39762, USA

ARTICLE INFO

Article history:
Received 10 December 2012
Received in revised form 23 May 2013
Accepted 13 August 2013
Available online 5 September 2013

Keywords:
Materials informatics
Data mining
Vapor-grown carbon nanofiber
Vinyl ester
Unsupervised learning

ABSTRACT

In this study, data mining and knowledge discovery techniques were employed to validate their efficacy in acquiring information about the viscoelastic properties of vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposites solely from data derived from a designed experimental study. Formulation and processing factors (VGCNF type, use of a dispersing agent, mixing method, and VGCNF weight fraction) and testing temperature were utilized as inputs and the storage modulus, loss modulus, and tan delta were selected as outputs. The data mining and knowledge discovery algorithms and techniques included self-organizing maps (SOMs) and clustering techniques. SOMs demonstrated that temperature had the most significant effect on the output responses followed by VGCNF weight fraction. SOMs also showed how to prepare different VGCNF/VE nanocomposites with the same storage and loss modulus responses. A clustering technique, i.e., fuzzy C-means algorithm, was also applied to discover certain patterns in nanocomposite behavior after using principal component analysis as a dimensionality reduction technique. Particularly, these techniques were able to separate the nanocomposite specimens into different clusters based on temperature and tan delta features as well as to place the neat VE specimens (i.e., specimens containing no VGCNFs) in separate clusters. Most importantly, the results from data mining are consistent with previous response surface characterizations of this nanocomposite system. This work highlights the significance and utility of data mining and knowledge discovery techniques in the context of materials informatics.

1. Introduction

Data mining is a field at the intersection of computer science and modern mathematical analysis [1–4]. It is used for discovering patterns in large datasets using predictive modeling techniques, where hidden data trends can be found [2]. The overall goal of the data mining process is to extract information from a large complex dataset and transform it into an understandable structure, thus enabling knowledge discovery. This transformation of massive amounts of structured and unstructured data into information and then into new knowledge using a myriad of data mining techniques is one of the great challenges facing the engineering community. The use of data mining techniques in the context of materials science and engineering is considered an important extension of materials informatics [5–8]. This interdisciplinary study integrates computer science, information science, and other domain areas to provide new understanding and to facilitate knowledge discovery. Materials informatics is a tool for material scientists to interpret vast amounts of experimental data through the use of machine learning approaches integrated with new visualization schemes, more human-like interactions with the data, and guidance by domain experts. It can also accelerate the research process and guide the development of new materials with select engineering properties. Material informatics is being fueled by the unprecedented growth in information technology and is driving the interest in the application of knowledge representation/discovery, data mining, machine learning, information retrieval, and semantic technology in the engineering disciplines.

There are several recent published applications utilizing material informatics and data mining. Hu et al. [9] used material

☆ Handled by C.-H. Chen.

* Corresponding author at: Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762, USA. Tel.: +1 662 325 2189.
E-mail address: rking@cavs.msstate.edu (R. King).

1474-0346/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aei.2013.08.002

informatics to resolve the problem of materials science image data sharing. They presented an ontology-based approach that can be used to develop annotation for non-structured materials science data with the aid of semantic web technologies. Yassar et al. [10] developed a novel computational model based on dislocation structures to predict the flow stress properties of 6022 aluminum alloy using data mining techniques. An artificial neural network (ANN) model was used to back-calculate the *in situ* non-linear material parameters and flow stress for different dislocation microstructures [10]. Sabin et al. [11] evaluated an alternative statistical Gaussian process model, which infers a probability distribution over all of the training data and then interpolates to make predictions of microstructure evolution arising from static recrystallization in a non-uniform strain field. Strain, temperature, and annealing time were the inputs of the model and the mean logarithm of grain size was its output. Javadi and Rezania [12] provided a unified framework for modeling of complex materials, using evolutionary polynomial regression-based constitutive model (EPRCM), integrated in finite element (FE) analysis, so an intelligent finite element method (EPR–FEM) was developed based on the integration of the EPR-based constitutive relationships into the FE framework. In the developed methodology, the EPRCM was used as an alternative to the conventional constitutive models for the material. The results of the analyses were compared to those obtained from conventional FE analyses. The results indicated that EPRCMs are able to capture the material constitutive behavior with a high accuracy and can be successfully implemented in a FE model.

Brilakis et al. [13] presented an automated and content-based construction site image retrieval method based on the recognition of material clusters in each image. Under this method, the pixels of each image were grouped into meaningful clusters and were subsequently matched with a variety of pre-classified material samples. Hence, the existence of construction materials in each image was detected and later used for image retrieval purposes. This method has allowed engineers to meaningfully search for construction images based on their content. Sharif Ullah and Harib [14] presented an intelligent method to deal with materials selection problems, wherein the design configurations, working conditions, as well as the design-relevant information are not precisely known. The inputs for this method were: (1) a linguistic description of the material selection problems (expressing the required levels of material properties/attributes and their importance), and (2) the material property charts relevant to the linguistic description of the problem. The method was applied to select optimal materials for robotic links and it was found that composite materials were better than metallic materials for robotic links.

A class of advanced materials, nano-enhanced polymer composites [15], have recently emerged among the more traditional structural metals. Polymer nanocomposites have been used in a variety of light-weight high-performance automotive composite structural parts where improved specific properties and energy absorption characteristics are required [16]. Though polymer nanocomposites have recently been widely investigated [17,18], they have never been studied in the context of material informatics. Therefore, the purpose of this study is to apply data mining and knowledge discovery techniques, as a proof of concept, to a thermosetting vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposite system. Nouranian et al. [19–21] and Torres et al. [22] developed a relatively large dataset for this material system suitable for data mining. This study seeks to use this dataset to demonstrate the usefulness of knowledge discovery and data mining techniques for nanocomposite material property characterization.

VGCNFs are commercially viable nanoreinforcements with superb mechanical properties [23]. VEs are thermosetting resins suitable for automotive structural composites due to their superior properties in comparison with unsaturated polyesters [20–22, 24,25]. Incorporating VGCNFs into VEs may provide improved mechanical properties relative to the neat matrix. These mechanical properties, however, are dependent on the degree of VGCNF nanodispersion in the matrix achieved during the mixing stage of the process. Examples of good and poor nanofiber dispersion in the matrix are given in Fig. 1, where two transmission electron micrographs of VGCNF/VE specimens are compared. Large nested groups of nanofibers (agglomerates) are a sign of poor VGCNF dispersion in the matrix, often resulting in inferior mechanical properties.

Data mining and knowledge discovery techniques can help discover and map patterns in the physical, mechanical, and system properties of VGCNF/VE nanocomposites, thereby aiding the nanocomposite design, fabrication, and characterization without the need to conduct expensive and time-consuming experiments.

In this study, several unsupervised knowledge discovery techniques were used to explore a large VGCNF/VE dataset [19]. The dataset consisted of 240 data points each corresponding to the combinations of five input design factors and three output responses, i.e., a total of eight "dimensions." The dimensions in data mining are the combination of both inputs and outputs of the developed model. The dimensions of the VGCNF/VE dataset are VGCNF type, use or absence of dispersing agent, mixing method, VGCNF weight fraction, temperature, storage modulus, loss modulus, and tan delta (ratio of loss to storage modulus), where the last three dimensions correspond to measured macroscale material properties. Kohonen maps [26,27], or self-organizing maps (SOMs), were applied to the dataset in order to conduct a sensitivity analysis of all of these factors and responses. In addition, principal component analysis (PCA) [28] was used to provide a two-dimensional (2-D) representation of nanocomposite data. This facilitated application of the fuzzy C-means (FCM) clustering algorithm [29,30] to characterize the physical/mechanical properties of VGCNF/VE nanocomposites.

2. Materials and methods

A brief summary of the statistical experimental design and testing procedures to generate the VGFCNF/VE dataset is given here. A more detailed discussion can be found in [19–21].

2.1. Statistical experimental design

The effect of five input design factors on the viscoelastic properties (storage and loss modulus) of VGCNF/VE nanocomposites were investigated using a general mixed-level full factorial experimental design [31]. These carefully selected factors, based on the state-of-the-art formulation and processing procedures, included: (1) VGCNF type (designated as A), (2) use of a dispersing agent (B), (3) mixing method (C), (4) VGCNF weight fraction in parts per hundred parts of resin (phr) (D), and (5) the temperature (E) used in dynamic mechanical analysis (DMA) testing. Experimental design factors and their associated levels are given in Table 1.

A total of $2 \times 2 \times 3 \times 5 \times 4 = 240$ "treatment combinations" (different combinations of the factor levels in Table 1) were randomized to eliminate bias in preparing the specimens. Each treatment combination resulted in three specimens prepared from the same material batch [20,21]. Each specimen was tested using a dynamic mechanical analyzer (single cantilever/flexure mode) to measure average storage modulus, loss modulus, and tan delta for each treatment combination. Storage and loss moduli are dynamic mechanical properties and indicative of the polymer nanocomposite's stiffness and energy dissipation capability, respectively.

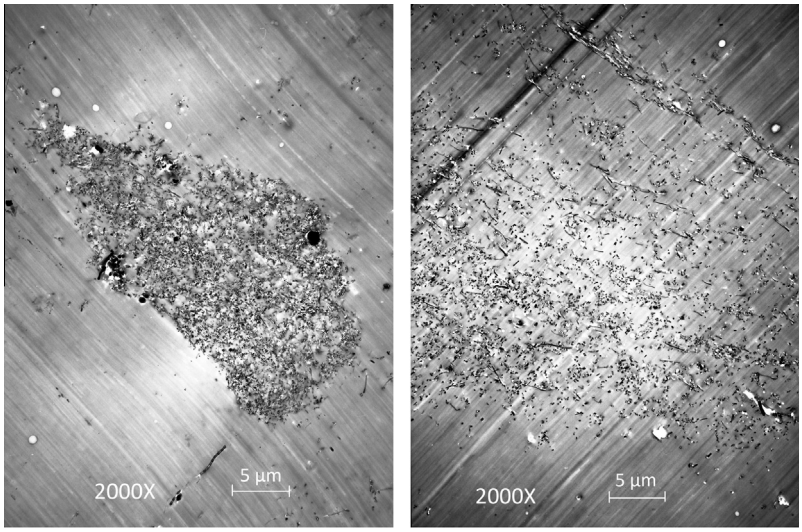

Fig. 1. Transmission electron micrographs of two VGCNF/VE specimens, where a nested VGCNF structure (agglomerate) is shown in (a), indicating a poor VGCNF dispersion in the matrix, and a better-dispersed system is shown in (b).

Table 1
The experimental design factors and their levels [19,21].

Factor designation	Factors	Level				
		1	2	3	4	5
A	VGCNF type	Pristine	Oxidized	–	–	–
B	Use of dispersing agent	Yes	No	–	–	–
C	Mixing method	US[a]	HS[b]	HS/US	–	–
D	VGCNF weight fraction (phr[c])	0.00	0.25	0.50	0.75	1.00
E	Temperature (°C)	30 °C	60 °C	90 °C	120 °C	–

[a] Ultrasonication.
[b] High-shear mixing.
[c] Parts per hundred parts of resin.

2.2. Materials and processing

A low styrene content (33 wt%) VE resin (Ashland Co., Derakane 441–400) and two VGCNF commercial grades, i.e., pristine PR-24-XT-LHT and surface-oxidized PR-24-XTLHT-OX (Applied Sciences Inc.) were utilized for nanocomposite specimen preparation [20,21]. In addition, methyl ethyl ketone peroxide (MEKP) (US Composites Inc.) and 6% cobalt naphthenate (CoNaph) (North American Composites Co.) were selected as initiator and crosslinking promoter, respectively. Air release additives BYK-A 515 and BYK-A 555 (BYK Chemie GmbH) were used to remove air bubbles introduced during mixing. A commercial dispersing agent BYK-9076 (BYK-Chemie GmbH) was employed to improve VGCNF dispersion in the resin.

Test specimens were prepared from a batch of resin comprising 100 parts resin, 0.20 phr 6% CoNaph, 0.20 phr BYK-A 515, 0.20 phr BYK-A 555, 0.00–1.00 phr VGCNFs (based on the design given in Table 1), and a 1:1 ratio of BYK-9076 to VGCNFs. The VGCNF/resin blend was mixed by either an ultrasonicator (ultrasonic processor GEX750-5C, Geneq Inc.), high-shear mixer (model L4RT-A, Silverson Machines Ltd.), or a combination of both, as dictated by the design given in Table 1. Then the nanofiber/resin blend was degassed under vacuum for 5–15 min at pressures of 8–10 kPa. The blend was thermally cured for 5 h at 60 °C followed by 2 h post-curing at 120 °C.

2.3. Dynamic mechanical analysis (DMA)

Test specimens were cut from cured specimens for DMA and polished using sandpaper. The storage and loss moduli were measured over a temperature range of 27–160 °C using a dynamic mechanical analyzer (TA Instruments, Model Q800) in the single cantilever mode at an amplitude of 15 μm, a fixed frequency of 10 Hz, and a heating rate of 5 °C/min.

3. Theory/Calculation

The average storage and loss moduli from three repeat tests for each of the 240 treatment combinations are given in [19]. This study incorporates five input design factors, i.e., VGCNF type (A), use of a dispersing agent or not (B), mixing method (C), VGCNF weight fraction (D), and DMA testing temperature (E) and three output responses, i.e., storage modulus, loss modulus, and tan delta. Hence, the dataset represents an eight-dimensional (8-D) space for analysis. Since factors A, B, and C are considered *qualitative* factors, they are represented by a numeric code for analysis purposes. For two-level factors A and B, 0 and 1 are the coded values for the first and second levels, respectively. For the three-level factor C, −1, 0, and 1 are the coded values for the first, second, and third levels, respectively (Table 1).

The logic behind data mining can be summarized as follows: (1) identify dominant patterns and trends in the data by utilizing the SOMs to conduct a sensitivity analysis; (2) apply a dimensionality reduction technique, such as PCA, to the data in order to enable the FCM clustering analysis of the data; (3) perform the FCM analysis of the data; and (4) transfer the findings of data mining techniques to the domain experts to validate the discovered data patterns and trends.

On the basis of the above discussion, SOMs [26,27], PCA [28], and the FCM clustering algorithm [29,30] were used with the 240 treatment combination dataset to discover nanocomposite data patterns and trends and to identify the different system features related to the specific material properties. SOMs were created with respect to temperature, VGCNF weight fraction, storage modulus, loss modulus, and tan delta. After analyzing the SOMs, temperature was identified as the most important input feature for the VGCNF/VE nanocomposites because it has the highest impact on the resulting storage and loss moduli responses. VGCNF weight fraction was also an important feature. In addition, it was inferred from the SOMs that some specimens tested at the same temperature tended to have several sub-clusters (groups). Each sub-cluster had the same tan delta or VGCNF weight fraction values.

Before applying these techniques, a brief explanation of ANN and unsupervised learning is presented.

3.1. Artificial neural networks (ANNs) and unsupervised learning

ANNs are a host of simple processors (neurons) that are interconnected in an organized fashion (architecture) and associated with a learning algorithm that emulates a biological process [26]. There are numeric values (weights) associated with the interconnections of the simple processors that are adjusted over time to emulate learning. These weights encode knowledge about the problem domain. The architectures (neurons and their interconnections) provide a computational structure for simulating a biological neural network. Therefore, many of the architectures, including the one used in this study, are based on findings from the field of neuroscience [27].

Learning in an ANN can occur in either a supervised or an unsupervised fashion [26]. A supervised approach uses a learning algorithm that creates an input/output mapping based on a labeled training set; thus, creating a mapping between an n-dimensional input space and m-dimensional output space. In this case, the network will learn a functional approximation from the input/output pairings and will have the ability to recognize or classify a new input vector into a correct output vector (generalization). An unsupervised learning architecture, in contrast, presents the network with only a set of unlabeled input vectors from which it must learn. In other words, the unsupervised ANN is expected to create characterizations about the input vectors and to produce outputs corresponding to a learned characterization (i.e., knowledge discovery).

ANNs that use unsupervised learning will determine natural clusters or feature similarity within the input dataset and to present results in a meaningful manner [26]. Since no labeled training sets are used in this approach, the outputs from the unsupervised learning network must be examined by a domain expert to determine if the classification provides any new insight into the dataset. If the result is not reasonable, then an adjustment is made to one of the training parameters used to guide the network's learning, and the network is presented the patterns again.

3.2. Self-organizing maps (SOMs)

Kohonen [27] has proposed that humans process complex information by forming reduced representations of the relevant facts. An important aspect of this reduction in dimensionality is the ability to preserve the structural inter-relationships between input and output factors. He proposed that the brain accomplished this by a spatial ordering of neurons within the brain. This procedure did not involve movement of neurons, but was achieved through a change in the physiological nature of the neuron.

Kohonen maps are utilized to map patterns of arbitrary dimensionality into 2-D or three-dimensional (3-D) arrays of neurons (maps) [27]. A SOM may be thought of as a self-organizing cluster. The basic components of a 2-D SOM for assessing VGCNF/VE feature data are shown in Fig. 2. The inputs are the dimensions of the dataset being analyzed. Note that each element of the input vector x is connected to each of the processing units on the map through the weight vector w_{ij}. After training, the SOM will define a mapping between the nanocomposite input data space and the 2-D map of neurons. The nanocomposite feature output y_i of a processing unit is then a function of the similarity between the input vector and the weight vector. The nonlinear mapping of the SOM utilizes a technique developed by Sammon [32] that preserves the higher dimensional closeness on the map. In other words, if two vectors are close to each other in the higher dimensional space, then they are close to each other on the map.

In Fig. 2, a trained feature map and its response to a winning output neuron, when excited by an original training pattern or an unknown similar input vector pattern, is shown [26]. This figure is a general illustration to show the logic of the SOM and the ANN techniques. Knowledge about the significance of the area around the winning neuron will then help the domain expert in knowledge discovery.

The SOM training algorithm is typically implemented on a planar array of neurons as shown in Fig. 3 with spatially defined neighborhoods (e.g., hexagonal or rectangular arrays, with six or four nearest neighborhoods, respectively). Also, the map must contain some method of compressing the data into a manageable form. One important attribute of a SOM is that it performs data compression without losing information regarding the relative distance between data vectors. A SOM typically uses the Euclidean distance to determine the relative nearness or similarity of data [23].

The idea of a spatial neighborhood, N_m, is used in measuring the similarity between the input vector and values of the reference vector represented by the vector of weights between the input layer and all of the neurons on the map. Before training begins, the weights are randomized and a learning rate and neighborhood size are selected. Then, when a training vector is presented to the network the neuron on the map with the most similar weight values is found. The weights of the winning neuron and the neighborhood neurons are then adjusted (learning) to bring them closer to the training vector. Over the course of the iterative training process, the neighborhood size and learning rate are independently decreased until the map no longer makes significant adjustments. The result is that the neurons within the currently winning neighborhood undergo adaptation at the current learning step while the weights in the other neighborhoods remain unaffected. The

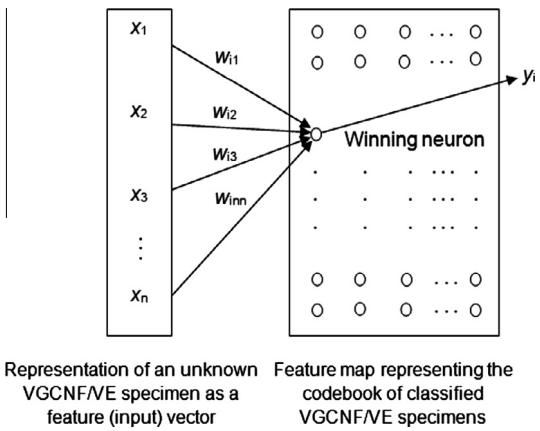

Fig. 2. Representation of the VGCNF/VE data analysis using ANN and a SOM. In the processing unit, the input vector x is multiplied by the weight vector w to create a mapping to the output vector y.

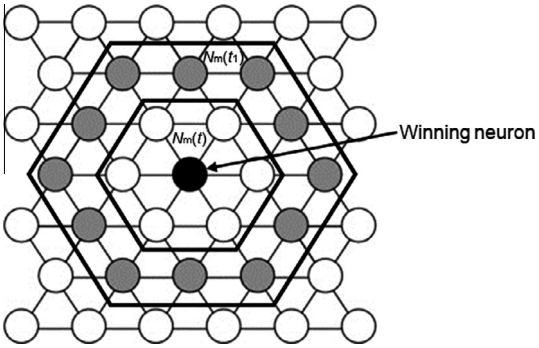

Fig. 3. Hexagonal grid used for SOMs showing 4 nearest neighbors.

winning neighborhood is defined as the one located around the best matching neuron, m [23].

The operation of the SOM algorithm progresses as follows. First, for every neuron i on the map, there is associated a parametric reference vector w_i. The initial values of $w_i(0)$ are randomly assigned. Next, an input vector x (R^n) is applied simultaneously to all of the neurons. The smallest of the Euclidean distances is used to define the best-matching neuron; however, other distance metrics may be explored to determine their efficacy in clustering the codebook vectors [23]. As the training progresses, the radius of N_m decreases with time (t) such that $N_{m(t_1)} > N_{m(t_2)} > N_{m(t_3)} > \ldots > N_{m(t_n)}$, where $t_1 < t_2 < t_3 < \ldots < t_n$. In other words, the neighborhood of influence can be very large when learning begins, but towards the end of the learning process, the neighborhood may involve only the winning neuron. The SOM algorithm also uses a learning rate that decreases with time.

In summary, the self-organization of the map proceeds as follows: (1) the map is presented with a sufficient number of training patterns; (2) weights are only adjusted on the neurons in the winning neighborhood; and (3) the adjustment is in proportion to the activation received by each neuron in the neighborhood. This weight adjustment enhances the same responses to a sufficiently similar subsequent input. As a result, a map is obtained with weights coding the stationary probability density function of the pattern vectors used for the training. The map also displays the data from a different viewpoint; instead of viewing the data as an n-dimensional vector, it can be viewed as a 2-D plot. This is where expert human analysis is enhanced. Instead of looking at the n-dimensional input vector of a sample and trying to determine what its meaning is, one needs only to look at the location of the sample on the map [24].

3.3. Principal component analysis (PCA)

Principal component analysis (PCA) is a method of identifying patterns in data and expressing this data to highlight similarities and differences [28]. These patterns can be hard to find in data of higher dimensions, where visual representations are not available. Therefore, PCA can be used as a powerful tool for analyzing data, identifying patterns, and data compression.

After performing PCA, the number of dimensions will be reduced without much loss of the embedded information. PCA includes four main data processing steps. First, the mean, i.e., the average across each dimension, is calculated. Second, the mean is subtracted from each of the data dimensions. Third, the covariance matrix [28] is calculated along with its eigenvalues and eigenvectors. Finally, these eigenvectors and eigenvalues can be used to choose the principal components and form a *feature* vector in order to derive the new low-dimensional dataset.

3.4. Fuzzy C-means (FCM) clustering algorithm

Once data dimensions have been reduced to a 2-D or 3-D graphical representation via PCA, several clustering algorithms can be applied to discover patterns in the data. In the following section, a summary of the FCM clustering algorithm, developed by Bezdek and Ehrlich [30] is presented. Clustering is often associated with the "membership" matrix U [30], which specifies the degree by which a certain data vector x belongs to a particular cluster c. The size of U is $C \times N$, where C is the number of clusters and N is the number of data vectors in the dataset. C is set initially to be $2 \leqslant C \leqslant (N-1)$.

$$U = \begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1N} \\ u_{21} & u_{22} & \ldots & u_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ u_{C1} & u_{C2} & \ldots & u_{CN} \end{bmatrix}, \tag{1}$$

$$\text{where } u_{ij} = \begin{cases} 1 & \text{if } x_j \in A_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

u_{ij} is called a crisp 0–1 matrix and x_j and A_i represent the data vector j and the class i, respectively. The number of elements in a cluster is given by the sum across a row of U, and

$$\sum_{i=1}^{C} u_{ij} = 1 \quad \text{for all} \quad j = 1, 2, \ldots, N, \tag{3}$$

Clustering can be described using an optimization scheme, which involves formulating a cost function and then using iterative and alternate estimations of the function. For example, the cluster centers and membership matrix U can be initially computed and then iteratively recalculated and updated.

FCM was created by Bezdek and Ehrlich [30] and is considered an objective function-based clustering technique. Each cluster using FCM has a prototype v_i that distinguishes cluster i, where the initial values of v_i can be set randomly or by picking the furthest points in the dataset or by picking exemplars from the dataset. Thus, the overall prototype vector V has a size of $(1 \times C)$ and can be denoted as

$$V = \{v_1, v_2, \ldots, v_c\} \tag{4}$$

The FCM cost function can be written as

$$J(U, V) = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^{Q} d(x_k, v_i), \tag{5}$$

where Q is a weighting exponent ($1 \leqslant Q < \infty$) and $d(x_k, v_i)$ is the distance measure between the data vector x_k and the cluster i (represented by prototype i). Therefore,

$$u_{rs} = \frac{1}{\sum_{i=1}^{C} \left(\frac{d(x_s, v_r)}{d(x_s, v_i)} \right)^{\frac{1}{Q-1}}}. \tag{6}$$

For the Euclidean distance measure,

$$d_{ik}^2(x_k, v_i) = d_{ik}^2 = (x_k - v_i)^T (x_k - v_i) = x_k^T x_k - 2 x_k^T v_i + v_i^T v_i. \tag{7}$$

Therefore,

$$v_j = \frac{\sum_{k=1}^{N} (u_{jk})^Q x_k}{\sum_{k=1}^{N} (u_{jk})^Q}. \tag{8}$$

Now, for the Gustafon–Kessel (GK) distance measure,

$$d_{ik} = \left(|\Sigma_i|^{\frac{1}{D}} \left((x_k - v_i)^T \Sigma_i^{-1} (x_k - v_i) \right) \right)^{\frac{1}{2}}, \tag{9}$$

where d_{ik} is scaled by the hyper-volume approximation denoted by $|\Sigma_i|^{\frac{1}{D}}$. Σ_i is the covariance matrix for class i:

$$\frac{\partial d_{ik}^2}{\partial v_i} = -2|\Sigma_i|^{\frac{1}{D}}\Sigma_i^{-1}(x_k - v_i), \tag{10}$$

Therefore,

$$v_i = \frac{\sum_{k=1}^{N}(u_{ik})^Q x_k}{\sum_{k=1}^{N}(u_{ik})^Q}, \tag{11}$$

$$\Sigma_i = \frac{\sum_{k=1}^{N}(u_{ik})^Q (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{N}(u_{ik})^Q}, \tag{12}$$

The GK distance measure in Eq. (9) uses a cluster-specific covariance matrix, so as to adapt various sizes and forms of the clusters. Thus, clustering algorithms that utilize GK distance measures try to extract much more information from the data than the algorithms based on the Euclidean distance measure [30]. Hence, the GK distance measure was used in this study. Based on this development, the *pseudo code* of the FCM algorithm is given as follows:

Compute $C \times N$ distance matrix;
Choose $v_j(0)$ as initial estimates of v_j, $j = 1,\ldots,C$;
//Initial value of the iteration counter, t
$t = 0$;
//Update the membership matrix U
Repeat:
  for $i = 1$ to N
    for $j = 1$ to C

$$u_{ji} = \frac{1}{\sum_{k=1}^{C}\left(\frac{d(x_i, v_j)}{d(x_i, v_k)}\right)^{\frac{1}{Q-1}}};$$

  End for
End for
//Now, $t = 1$
$t = t + 1$;
  //Prototypes Update
  for $j = 1$ to C
  solve:
  $\sum_{i=1}^{N} u_{ji}^{Q}(t-1)\frac{\partial d(x_i, v_j)}{\partial v_j} = 0$; with respect to v_j and set v_j equal to the computed solution
End for
Test for convergence:Select termination criteria using, for example, particular number of iterations or the difference from t to $t - 1$ of the sum of prototype differences or other appropriate criteria.

---

4. Results and discussion

In Fig. 4, a 10×10 SOM resulting from the 240 data points is shown. Nanocomposite specimens tested at the same DMA temperature tend to cluster together. For example, specimens tested at 30 °C tend to cluster at the top of the map, whereas specimens tested at 120 °C tend to cluster at the bottom. A mixture of specimens tested at 60 °C and 90 °C are located in the middle of the map.

In Figs. 5 and 6, two 10×10 SOMs for the VGCNF weight fraction and the tan delta response are shown, respectively. In Fig. 5, specimens with the same weight fraction tend to cluster together, but this tendency is not consistent and is less than the clustering

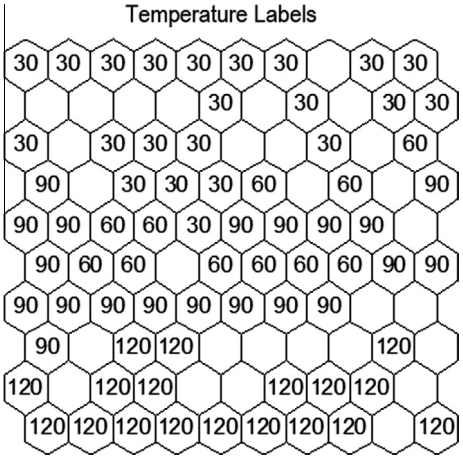

Fig. 4. A 10×10 SOM with respect to temperature for the 240 nanocomposite specimens used in the study (with all eight dimensions). The specimens tested at the same temperature tend to cluster together.

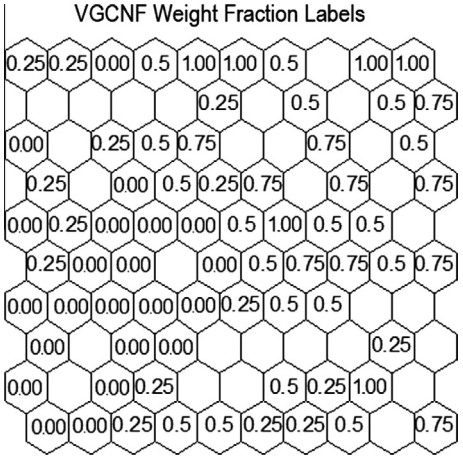

Fig. 5. A 10×10 SOM with respect to VGCNF weight fractions. The clustering tendency is less than that of the temperature in Fig. 4. However, within a certain temperature cluster, the existence of sub-clusters with the same VGCNF weight fraction is confirmed.

tendency shown in Fig. 4 for temperature. However, if the cluster at one temperature (say 30 °C on the top of Fig. 4) is considered and compared to the corresponding cluster in Figs. 5 and 6, sub-clusters with similar VGCNF weight fractions can be identified. For example, the first seven 30 °C labels from the left in Fig. 4 have corresponding weight fractions of 0.25, 0.25, 0, 0.50, 1.00, 1.00, and 0.5 phr and corresponding tan delta values of 0.03, 0.02, 0.03, 0.02, 0.02, 0.02, and 0.02. This means that within the nanocomposite specimens tested at 30 °C, there are some specimens with similar VGCNF weight fractions that tend to cluster together. For example, specimens with a weight fraction of 0.25 phr as well as 1.00 phr are mapped together (Fig. 5). Similarly, specimens that have a tan delta value of 0.02 are mapped together (Fig. 6). This explains why some of the specimens tested at the same temperature are separated by blank hexagons from each other. Each group of specimens in Figs. 5 and 6 that were tested at the same temperature tend to have similar VGCNF weight fractions or tan delta values. However, in Fig. 6, the clustering for tan delta is more pronounced than that of the VGCNF weight fraction and less than that of the temperature. This leads to the conclusion that temperature is the dominant feature for the treatment combinations and has the highest impact on the responses.

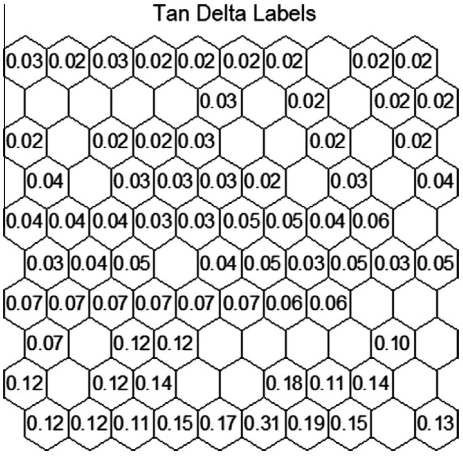

Fig. 6. A 10 × 10 SOM with respect to tan delta values. The clustering tendency is less than that of the temperature in Fig. 4. However, within a certain temperature cluster, the existence of sub-clusters with the same tan delta value is confirmed.

In addition to the sensitivity analysis inferred from SOMs, the different conditions needed to produce a particular response can also be determined. In Fig. 7, a 10 × 10 SOM is shown indicating the indices, which represent the numeric orders of the specimens mapped. Each index corresponds to one treatment combination out of 240 with specific values of VGCNF type, use of a dispersing agent, mixing method, VGCNF weight fraction, testing temperature, storage modulus, loss modulus, and tan delta. The indices in Fig. 7 can be used to extract information linking the different dimensional combinations that produce certain response values. For example, in Figs. 8 and 9, SOMs for the storage and loss moduli are illustrated, respectively.

In Fig. 8, the storage modulus response values are shown. A group of three specimens have a storage modulus of about 2.6 GPa, located at the third and fourth rows of the SOM. In Fig. 7, these values correspond to specimen indices 73, 65, and 25. Clearly, different dimensional properties can be determined to produce the same 2.6 GPa response value. These properties are shown in Table 2, where the third row (highlighted) has a lower tan delta value and higher storage modulus than the other two specimens. Nanocomposite designers can use such information in the selection of input factor levels.

Similarly, the loss moduli for a group of three specimens are all about 104 MPa in the sixth and seventh rows of the SOM in Fig. 9.

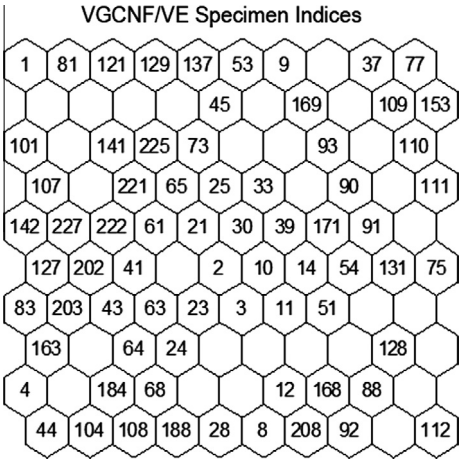

Fig. 7. A 10 × 10 SOM illustrating the indices (numeric orders) of the 240 nanocomposite specimens [19].

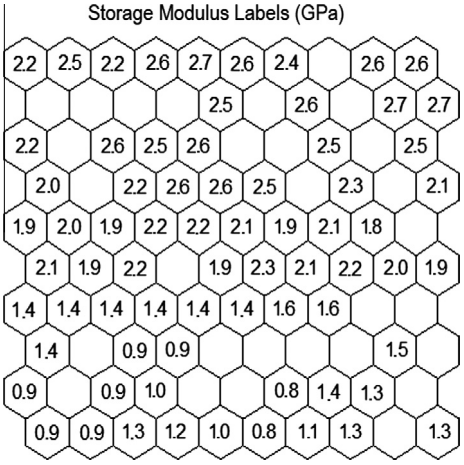

Fig. 8. A 10 × 10 SOM based on the storage modulus response.

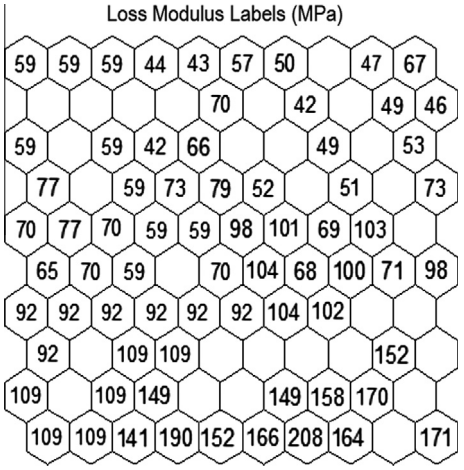

Fig. 9. A 10 × 10 SOM based on the loss modulus response; the values are rounded to the nearest integer for simplicity.

These correspond to indices 10, 11, and 51 (Fig. 7). Again, different dimensional properties can be prescribed to produce the 104 MPa responses. These properties are shown in Table 3, where the first row (highlighted) has a lower tan delta value and higher storage modulus response than the other two specimens.

A PCA was run on the VGCNF/VE nanocomposite data. Fig. 10 shows a graphical representation for the PCA of the data. PCA reduced the number of data dimensions from eight to two and each specimen was a given a specific 2-D representation (principal component 1 and 2 axes) so that specimens that have similar properties were mapped together in the 2-D space. Thus, there are no specific units associated with the abscissa and ordinate. This step is fundamental so that clustering algorithms (Section 3.4) can be applied to identify certain patterns in these nanocomposite data. Such patterns can be used to explain certain physical/mechanical behavior associated with the data without running additional experiments.

The FCM was applied to the VGCNF/VE nanocomposite data using the GK distance measures. In Fig. 11, the FCM results are illustrated, where four clusters are chosen to represent the data using the GK distance measure. The data points are divided into four different clusters, each shown with a different color. In Fig. 11a, the nanocomposite specimens tested at 90 °C and 120 °C each form a separate cluster with average tan delta values of 0.049 and 0.148, respectively. The rest of the nanocomposite

Table 2
Different dimensional (factorial) combinations required to produce a storage modulus of about 2.6 GPa.

VGCNF Type (A)	Use of a Dispersing Agent (B)	Mixing Method (C)	VGCNF Weight Fraction (D) (phr)	Temperature (E) (°C)	Storage Modulus (GPa)	Loss Modulus (MPa)	Tan Delta
Pristine	Yes	US[1]	0.25	30	2.577	79	0.031
Oxidized	Yes	US	0.25	30	2.566	73	0.028
Oxidized	Yes	US	0.75	30	2.641	66	0.025

[a] Ultrasonication.

Table 3
Different dimensional (factorial) combinations required to produce a loss modulus of about 104 MPa.

VGCNF Type (A)	Use of a Dispersing Agent (B)	Mixing Method (C)	VGCNF Weight Fraction (D) (phr)	Temperature (E) (°C)	Storage Modulus (GPa)	Loss Modulus (MPa)	Tan Delta
Pristine	No	US[1]	0.5	60	2.276	104	0.046
Pristine	No	US	0.5	90	1.621	104	0.064
Oxidized	No	US	0.5	90	1.614	102	0.063

[a] Ultrasonication.

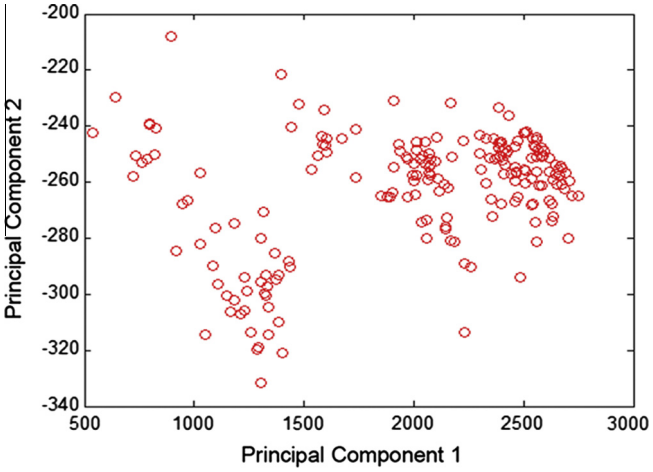

Fig. 10. A 2-D graphical representation of the VGCNF/VE nanocomposite specimen data (illustrated by circle points) using the PCA technique. This technique maps the data from an 8-D space down to a 2-D space so that different clustering algorithms can be applied. The values associated with the principal dimensions 1 and 2 are random, but each specimen was given a 2-D coordinate so that specimens with similar properties would be mapped together in the 2-D space.

specimens tested at 30 °C and 60 °C, along with neat VE specimens tested at 30 °C, form a single cluster with an average tan delta value of 0.025. The remaining neat VE specimens tested at 60 °C, 90 °C, and 120 °C form the fourth cluster. In Fig. 11b, a "scale data and display image (imagesc) object" plot is presented to indicate the number of clusters (each distinct set of bands in a row) and the bands associated with each cluster. The bands reflect the densities of data points within each cluster and correspond to the distances between the data points in Fig. 11a. These findings prove that temperature is a dominant feature for the whole dataset.

In Fig. 12, the FCM results are illustrated where five clusters are chosen to represent the data using the GK distance measure. The nanocomposite and neat VE specimens tested at 30 °C form a cluster with an average tan delta value of 0.025. Included in this cluster is a fraction of the nanocomposite specimens tested at 60 °C. The remainder of the nanocomposite and neat VE specimens tested at 60 °C, along with a fraction of the nanocomposite specimens tested at 90 °C, are contained in a separate cluster with an average tan delta value of 0.039. The rest of nanocomposite specimens tested at 90 °C and a fraction of nanocomposite specimens tested at

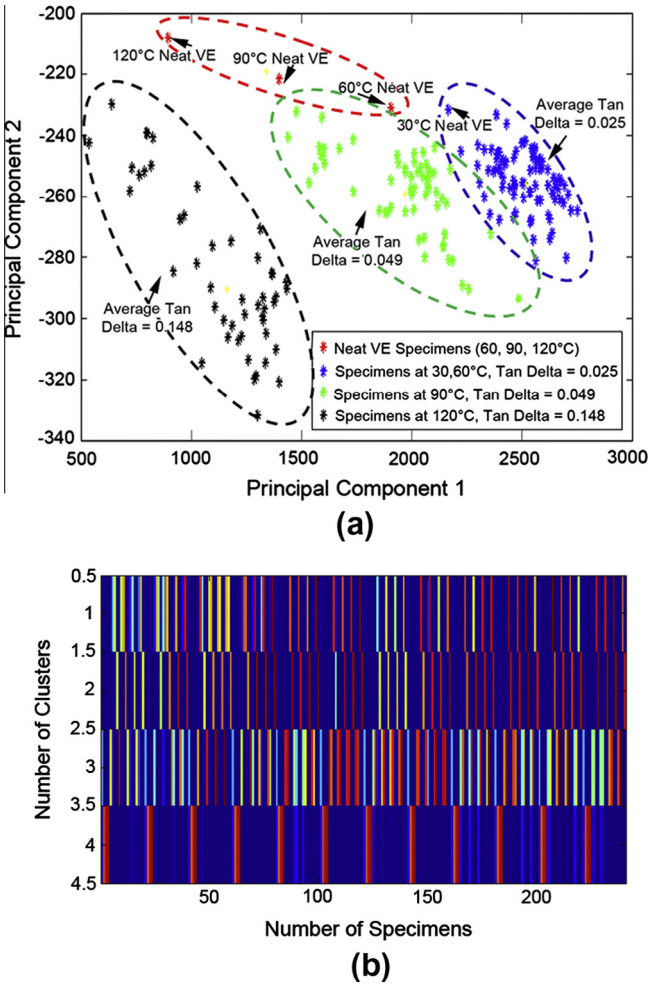

Fig. 11. (a) Clustering results after applying the FCM algorithm and the GK distance measure, when $C = 4$. Temperature labels are also included. (b) In the "scale data and display image (imagesc) object" plot, four bands representing four clusters can be identified.

120 °C form a third unique cluster with an average tan delta value of 0.099. The rest of the nanocomposite specimens tested at 120 °C form a fourth cluster with an average tan delta value of 0.148. Lastly, Fig. 12a includes a fifth separate cluster that contains the

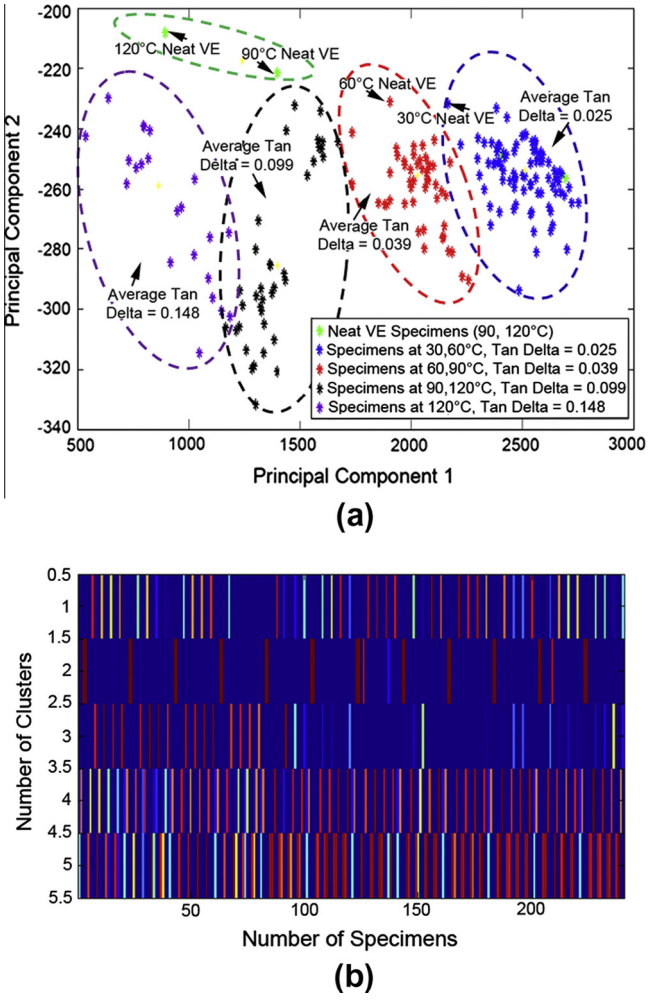

Fig. 12. (a) Clustering results after applying the FCM algorithm and the GK distance measure, when $C = 5$. Temperature labels are also included. (b) In the "scale data and display image object (imagesc)" plot, five bands representing five clusters can be identified.

neat VE specimens tested at 90 °C and 120 °C. In Fig. 12b, an imagesc plot is presented, where five clusters can be identified. Again, these results demonstrate that temperature is a dominant feature.

Using the GK distance measure, FCM works better for the 240 VGCNF/VE specimens when the selected number of clusters equals four. For this case, specimens tested at different temperatures tend to be located in separate clusters that distinguish each of these temperatures. In addition, neat VE data specimens tested at 60–120 °C tended to cluster together. These results suggest that the FCM algorithm was able to identify VGCNF/VE specimens that have similar properties and placed them into different clusters.

The SOM analysis allows a preliminary visual identification of the different existing groups [24]. In contrast, the FCM clustering approach identifies existing clusters and provides a mechanism to assign VGCNF/VE specimens to the appropriate cluster. Furthermore, FCM allows objects to belong to several clusters simultaneously, with different degrees of membership. This feature is not available in SOMs [27]. Hence, SOMs can be more helpful in identifying the dominant feature(s)/dimension(s) in the dataset. Other clustering algorithms (e.g., FCM) can be used to better identify cogent patterns and trends in VGCNF/VE data. In addition, different VGCNF/VE and/or neat VE specimens and their associated viscoelastic properties can be identified and categorized within their respective clusters. Each cluster can be identified based on one or more of the input design factors of the VGCNF/VE system.

5. Materials informatics and validation of the results

The findings from this data mining research confirm the trends established previously using a response surface methodology [20,21] and are consistent with the viscoelasticity theory for polymers [33]. In general, nanocomposite storage moduli drop steadily as temperature increases up to the glass transition temperature (T_g), where a sharp drop of several orders of magnitude occurs. Analogously, the loss modulus increases with increasing temperature, reaching a maximum at T_g. The clustering of VGCNF/VE nanocomposite data for specimens tested at different temperatures (i.e., 30, 60, 90, 120 °C) suggests distinguishably different viscoelastic material behaviors at these temperatures (Figs. 4, 11 and 12). Of course, the effect of temperature on nanocomposite storage and loss moduli is much greater than that of the other factors and, hence, clustering based on temperature is readily apparent.

The effect of VGCNF weight fraction on nanocomposite storage and loss moduli was significant in previous studies [20,21]. Since VGCNFs reinforce and stiffen the matrix, the nanocomposite storage modulus increased with increasing VGCNF weight fraction until a peak was reached near an optimal VGCNF weight fraction of 0.50 phr. However, due to the presence of large VGCNF agglomerates caused by incomplete nanodispersion at higher weight fractions, a steady increase in the storage modulus was not realized once the nanofiber weight fraction exceeded 0.50 phr [20]. The loss modulus typically decreases with increasing VGCNF weight fraction [20,21]. However, due to VGCNF agglomeration and poor dispersion in the polymer matrix, stress concentrations and frictional sliding in the entangled VGCNF networks cause a more complex viscoelastic nanocomposite behavior [21]. Using data mining, the effect of nanofiber weight fraction on nanocomposite viscoelastic properties was clearly identified (Fig. 5) consistent with previous response surface model results [20,21]. The clustering of neat VE specimens indicates that a sharp difference exists between the viscoelastic responses of these specimens versus that for the VGCNF/VE nanocomposites. Using a response surface methodology [21], a ~20% increase in the storage modulus was observed by introducing VGCNFs into neat VE.

6. Summary and conclusions

Knowledge discovery techniques were applied to a vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposite dataset as a case study for materials informatics. This dataset had been generated by a full factorial experimental design with 240 different design points. Each treatment combination in the design consisted of eight feature dimensions corresponding to the design factors, i.e., VGCNF type, use of a dispersing agent, mixing method, VGCNF weight fraction, and testing temperature as the inputs and storage modulus, loss modulus, and tan delta as the output responses. Self-organizing maps (SOMs) were created with respect to temperature, tan delta, VGCNF weight fraction, storage modulus, and loss modulus. After analyzing the SOMs, temperature was identified as the dominant feature for the VGCNF/VE nanocomposites having the highest impact on the viscoelastic material responses. VGCNF weight fraction was also a dominant feature. In addition, it was inferred from the SOMs that some specimens tested at the same temperature tended to have several sub-clusters. Each sub-cluster had the same tan delta or VGCNF weight fraction values. Analyzing the SOMs with respect to storage and loss moduli demonstrated that VGCNF/VE specimens with different features could be designed to match a desired storage and/or loss modulus.

Finally, another data analysis was performed using the principle component analysis (PCA) technique. Then, the fuzzy C-means

(FCM) algorithm with the Gustafon-Kessel (GK) distance measure was applied to the resulting new dataset. The FCM clustered the specimens based on temperature as well as tan delta values. In addition, the FCM was able to recognize neat VE specimens tested at 30, 60, 90, and 120 °C and placed most of them in one cluster. In other words, when four clusters were selected and the GK distance measure was applied, neat VE specimens tested at 60–120 °C were placed in one cluster. In contrast, when five clusters were selected and the GK distance measure was applied, neat VE specimens tested at 90 and 120 °C were placed in one cluster. This reflects the fact that the viscoelastic properties of each neat VE specimen in both groups are similar. However, the FCM algorithm worked better when the number of clusters equals four, because more neat VE specimens tend to cluster together at the selected temperatures.

In summary, the main contributions of this study are:

- Developing a sensitivity analysis structure using SOMs in order to discover the most and least dominant features of the VGCNF/VE system, whether they are input design factors or output responses.
- Developing a tool for identifying VGCNF/VE specimen designs leading to the same storage and loss moduli. This will facilitate tailoring of nanocomposite viscoelastic properties and, in turn, minimize fabrication costs by the domain experts.
- Developing a methodology to better identify cogent patterns and trends in VGCNF/VE data. Each cluster can be identified based on one or more of the input design factors of the VGCNF/VE system.

The knowledge discovery techniques applied here demonstrate the dominant features in the nanocomposite data without the need to conduct additional expensive and time-consuming experiments. This highlights the feasibility of data mining and knowledge discovery techniques in materials science and engineering and the rising field of Materials Informatics.

Acknowledgment

This work was supported in part by the US Department of Energy under contract DE-FC26-06NT42755.

References

[1] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier Science, 2011.
[2] Z.Q. John Lu, The elements of statistical learning: data mining, inference, and prediction, Journal of the Royal Statistical Society: Series A (Statistics in Society 173 (2010) 693–694.
[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine 17 (1996) 37.
[4] D.T. Larose, An introduction to data mining, Traduction et adaptation de Thierry Vallaud (2005).
[5] K. Rajan, Materials informatics, Materials Today 8 (2005) 38–45.
[6] K.F. Ferris, L.M. Peurrung, J.M. Marder, Materials informatics: fast track to new materials, Advanced Materials & Processes 165 (1) (2007) 50–51.
[7] C. Suh, K. Rajan, B.M. Vogel, B. Narasimhan, S.K. Mallapragada, Informatics methods for combinatorial materials science, Combinatorial Materials Science (2007) 109–119.
[8] Q. Song, A preliminary investigation on materials informatics, Chinese Science Bulletin 49 (2004) 210–214.
[9] C. Hu, C. Ouyang, J. Wu, X. Zhang, C. Zhao, NON-structured materials science data sharing based on semantic annotation, Data Science Journal (2009). 904220065.
[10] R.S. Yassar, O. AbuOmar, E. Hansen, M.F. Horstemeyer, On dislocation-based artificial neural network modeling of flow stress, Materials & Design 31 (2010) 3683–3689.
[11] T. Sabin, C. Bailer-Jones, P. Withers, Accelerated learning using Gaussian process models to predict static recrystallization in an Al-Mg alloy, Modelling and Simulation Materials Science and Engineering 8 (2000) 687.
[12] A. Javadi, M. Rezania, Intelligent finite element method: an evolutionary approach to constitutive modeling, Advanced Engineering Informatics 23 (2009) 442–451.
[13] I.K. Brilakis, L. Soibelman, Y. Shinagawa, Construction site image retrieval based on material cluster recognition, Advanced Engineering Informatics 20 (2006) 443–452.
[14] A. Ullah, K.H. Harib, An intelligent method for selecting optimal materials and its application, Advanced Engineering Informatics 22 (2008) 473–483.
[15] J.H. Koo, Polymer nanocomposites: processing, characterization, and applications, McGraw-Hill, New York, NY, 2006.
[16] J. Garces, D.J. Moll, J. Bicerano, R. Fibiger, D.G. McLeod, Polymeric nanocomposites for automotive applications, Advanced Materials 12 (2000) 1835–1839.
[17] F. Hussain, M. Hojjati, M. Okamoto, R.E. Gorga, Review article: polymer-matrix nanocomposites, processing, manufacturing, and application: an overview, Journal of Composite Materials 40 (2006) 1511–1575.
[18] E.T. Thostenson, C. Li, T.W. Chou, Nanocomposites in context, Composites Science and Technology 65 (2005) 491–516.
[19] S. Nouranian, Vapor-grown carbon nanofiber/vinyl ester nanocomposites: Designed experimental study of mechanical properties and molecular dynamics simulations, Mississippi State University, PhD Dissertation, Mississippi State, MS USA, 2011.
[20] S. Nouranian, H. Toghiani, T.E. Lacy, C.U. Pittman, J. Dubien, Dynamic mechanical analysis and optimization of vapor-grown carbon nanofiber/vinyl ester nanocomposites using design of experiments, Journal of Composite Materials 45 (2011) 1647–1657.
[21] S. Nouranian, T.E. Lacy, H. Toghiani, C.U. Pittman Jr., J.L. Dubien, Response surface predictions of the viscoelastic properties of vapor-grown carbon nanofiber/vinyl ester nanocomposites, Journal of Applied Polymer Science (2013). DOI: 10.1002/app.39041.
[22] G.W. Torres, S. Nouranian, T.E. Lacy, H. Toghiani, C.U. Pittman Jr., J. Dubien, Statistical Characterization of the Impact Strengths of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites Using a Central Composite Design, Journal of Applied Polymer Science 128 (2013) 1070–1080.
[23] G.G. Tibbetts, M.L. Lake, K.L. Strong, B.P. Rice, A review of the fabrication and properties of vapor-grown carbon nanofiber/polymer composites, Composites Science and Technology 67 (2007) 1709–1718.
[24] A. Plaseied, A. Fatemi, M.R. Coleman, Influence of carbon nanofiber content and surface treatment on mechanical properties of vinyl ester, Polymers & Polymer Composites 16 (2008) 405–413.
[25] A. Plaseied, A. Fatemi, Tensile creep and deformation modeling of vinyl ester polymer and its nanocomposite, Journal of Reinforced Plastics and Composites 28 (2009) 1775.
[26] R.L. King, A. Rosenberger, L. Kanda, Artificial neural networks and three-dimensional digital morphology: a pilot study, Folia Primatologica 76 (2005) 303–324.
[27] T. Kohonen, Self-Organization and Associative Memory, Springer- Verlag, 1988.
[28] I.T. Jolliffe, Principal Component Analysis, Springer, 2002.
[29] S. Miyamoto, H. Ichihashi, K. Honda, Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications, Springer, 2008.
[30] J.C. Bezdek, R. Ehrlich, FCM: the fuzzy c-means clustering algorithm, Computers & Geosciences 10 (1984) 191–203.
[31] D.C. Montgomery, Design and Analysis of Experiments, seventh ed., John Wiley & Sons, Hoboken, NJ, 2009.
[32] J.W. Sammon Jr., A nonlinear mapping for data structure analysis, IEEE Transactions on Computers 100 (1969) 401–409.
[33] J.D Ferry, Viscoelastic Properties of Polymers, Wiley, 1980.