ON MATERIALS INFORMATICS AND KNOWLEDGE DISCOVERY: MECHANICAL CHARACTERIZATION OF VAPOR-GROWN CARBON NANOFIBER/ VINYL ESTER NANOCOMPOSITES

Osama Abuomar[1,2], Sasan Nouranian[2], Roger King[1,2], Tom Lacy[3]

[1] Department of Electrical and Computer Engineering, Mississippi State University, 406 Hardy Road, Mississippi State, MS 39762, USA;
[2] Center for Advanced Vehicular Systems (CAVS), 200 Research Blvd., Starkville, MS 39759, USA;
[1] Department of Aerospace Engineering, Mississippi State University, Mississippi State, MS 39762, USA;

Keywords: Materials informatics, Data mining, Viscoelastic / Compression / Tension/ Flexural vapor-grown carbon nanofiber, Vinyl ester, Unsupervised learning.

ABSTRACT

In this study, data mining and knowledge discovery techniques were employed to acquire new information about the viscoelastic, flexural, compression, and tension properties for vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposites. These properties were used to design a unified VGCNF/VE framework solely from data derived from a designed experimental study. Formulation and processing factors (curing environment, use or absence of dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time) and testing temperature were utilized as inputs and the true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, 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 (particularly 30^{o}C) and tan delta had the most significant effects on the output responses followed by VGCNF high shear mixing time and sonication time. SOMs also showed how to produce optimal responses using a certain combination(s) of inputs. A clustering technique, i.e., fuzzy C-means algorithm (FCM), 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 (where 30^{o}C and 120^{o}C are the most dominant), tan delta, high shear mixing time, and sonication time features as well as to place the viscoelastic VGCNF/VE specimens that have the same storage and loss moduli and tested at the same temperature in separate clusters. FCM results also showed that all nanocomposites structures in the new framework are essential but the viscoelastic VGCNF/VE data is the most important. Most importantly, this work highlights the significance and utility of data mining and knowledge discovery techniques in the context of materials informatics by discovering certain patterns and trends that have not been known before.

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 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.

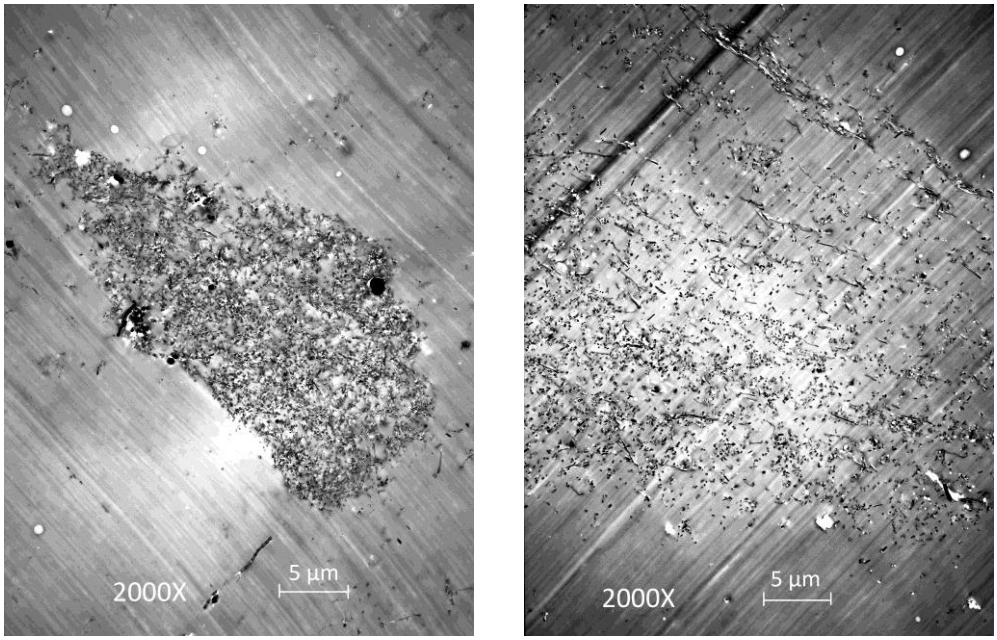
Brilakis *et al.*[13] presented an automated and content-based construction site image retrieval method based on the recognition of material clusters in each image. Under this method, the pixels of each image were grouped into meaningful clusters and were subsequently matched with a variety of pre-classified 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]. Polymer nanocomposites have recently been widely investigated[17,18] and AbuOmar *et al.*[19] applied data mining and knowledge discovery techniques, as a proof of concept and as a first attempt to utilize the concept of materials informatics, to a thermosetting viscoelastic vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposite system[20-23] where 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)[24,25] and clustering techniques[26,27]. 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[28]. 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[19]. These results are consistent with previous response surface characterizations of the viscoelastic nanocomposite system.

This study seeks to expand the viscoelastic nanocomposite dataset into a unified framework which includes more VGCNF/VE structures and then apply data mining and knowledge discovery techniques to the resulting dataset. The new expanded framework consists of the viscoelastic VGCNF/VE data, VGCNF/VE compression data, VGCNF/VE tension data, and flexural properties of VGCNF/VE[29]. This is the first time that such framework is designed, studied and analyzed and the major contribution of this paper is to apply data mining and knowledge discovery techniques in order to discover new knowledge, properties, and trends that have not been known *a priori* using this framework, thereby aiding the nanocomposite design, fabrication, and characterization without the need to conduct expensive and time-consuming experiments.

VGCNFs are commercially viable nanoreinforcements with superb mechanical properties[30]. VEs are thermosetting resins suitable for automotive structural composites due to their superior properties in comparison with unsaturated polyesters[21-23, 31, 32]. 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.



a) b)

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).

In this study, several unsupervised knowledge discovery techniques were used to explore an expended VGCNF/VE framework[20, 29]. The dataset in the framework consisted of 565 data points each corresponding to the combinations of eight input design factors and nine output responses, i.e., a total of seventeen "dimensions." The dimensions in data mining are the combination of both inputs and outputs of the developed model. The dimensions of the new VGCNF/VE framework are curing environment, use or absence of dispersing agent, mixing method, VGCNF fiber loading (or sometimes referred to as VGCNF weight fraction), VGCNF type, high shear mixing time, sonication time, temperature, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta (ratio of loss to storage modulus), where the last nine dimensions correspond to measured macroscale material responses. Kohonen maps[24,25], 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[26,27] to characterize the physical/mechanical properties of the new VGCNF/VE nanocomposites framework.

MATERIALS AND METHODS

All data used in this work were generated using various statistical experimental designs, such as a general mixed level full factorial and central composite design, and are described in detail elsewhere[20-23, 29]. Different datasets were merged into a larger one incorporating 240 viscoelastic data points, 60 flexural data points, 172 compression data points, and 93 tension data points for

variously formulated and processed VGCNF/VE nanocomposites. Therefore, the new larger dataset has a total of 565 data points. Each data point corresponds to combinations of eight input design factors and nine output responses. The input factors of the new VGCNF/VE dataset are curing environment (air vs. nitrogen), use or absence of a dispersing agent, mixing method (ultrasonication, high-shear mixing, and combination of both), VGCNF weight fraction, VGCNF type (pristine vs. oxidized), high-shear mixing time, sonication time, and temperature. The output factors (i.e., measured properties) are true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta. Different data interpolation techniques were used to replace some of the missing and unknown data fields in the new dataset[34]. These techniques include linear interpolation which is a method of curve-fitting using linear polynomials, and spline interpolation where the interpolant is a spline (piecewise polynomial). However, spline interpolation is more precise than regular polynomial interpolations because of its low interpolation error regardless of the polynomial degree used for the spline. In addition, spline interpolation avoids the problem of Runge's phenomenon, which occurs when using high degree polynomials for the interpolation process[34].

THEORY/CALCULATION

This study incorporates eight input design factors, i.e., curing environment (nitrogen, oxygen), use or absence of dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time, and testing temperature and nine output responses, i.e., true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta. Hence, the dataset represents a seventeen-dimensional (17-D) space for analysis. Since curing environment, use or absence of dispersing agent, mixing method, and VGCNF type are considered *qualitative* factors, they are represented by a numeric code for analysis purposes. For two-level factors (curing environment, use or absence of dispersing agent, and VGCNF type), 0 and 1 are the coded values for the first and second levels, respectively. For the three-level factor (mixing method), -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[24,25], PCA[28], and the FCM clustering algorithm[26,27] were used with the 565 treatment combination dataset to discover data patterns and trends for the expanded nanocomposites framework and to identify the different system features related to the specific material properties. SOMs were created with respect to VGCNF fiber loading, high shear mixing time, sonication time, temperature, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, 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 new framework because it has the highest impact on the resulting responses. VGCNF high shear mixing time and sonication time were also important features. 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 high shear mixing time or sonication time values. In addition, after analyzing the clustering results, it has been found that the viscoelastic VGCNF/VE data is very important in the newly designed VGCNF/VE framework.

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

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[24]. 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[25].

Learning in an ANN can occur in either a supervised or an unsupervised fashion[24]. 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[24]. 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.

Self-Organizing Maps (SOMs)

Kohonen[25] 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)[25]. 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[35] 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[24]. 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.

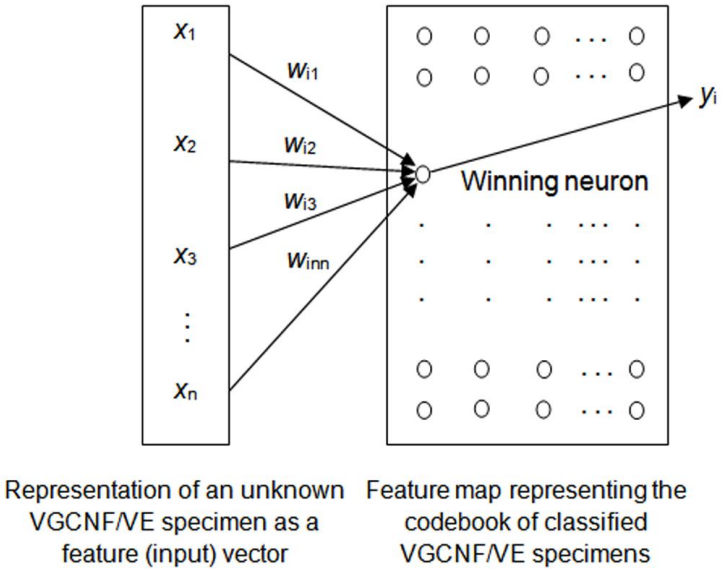


Fig. 2. Representation of the VGCNF/VE data analysis using ANN and a SOM. In the processing unit, the input vector *x* is multiplied by the weight vector *w* to create a mapping to the output vector *y*.

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[24].

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 winning neighborhood is defined as the one located around the best matching neuron, m[24].

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 [24]. As the training progresses, the radius of N_m decreases with time (t) such that $N_{m(t_1)} > N_{m(t_2)} > N_{m(t_3)} > ... > N_{m(t_n)}$, where $t_1 < t_2 < t_3 < ... < t_n$. 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.

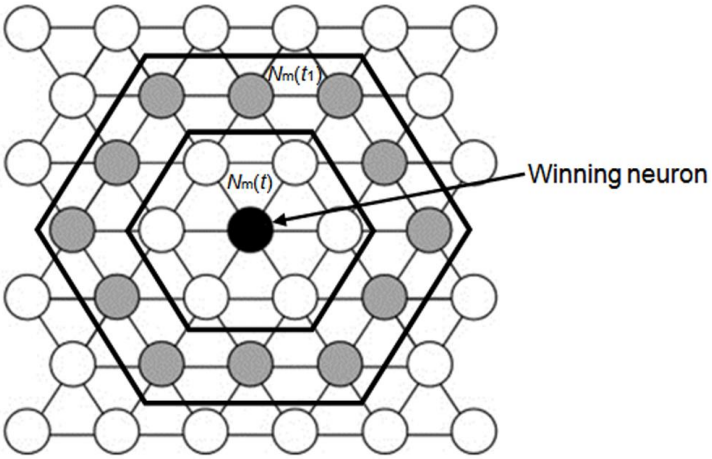


Fig. 3. Hexagonal grid used for SOMs showing 4 nearest neighbors

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.

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[26, 27] is presented. Clustering is often associated with the "membership" matrix U[27], 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 $2 \le C \le (N-1)$.

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

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

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

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

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[26,27] 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

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

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 $(1 \le Q < \infty)$ 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,

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

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

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

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*:

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

Therefore,

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

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

The GK distance measure in Equation 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[27]. 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

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

End for
End for

//Now, $t = 1$
$t = t + 1$;
//Prototypes Update
for $j = 1$ to C
solve:

$$\sum_{i=1}^{N} u_{ji}^{Q}(t-1) \frac{\partial d(x_i, v_j)}{\partial v_j} = 0 \quad ; \text{with respect to } v_j \text{ and set } v_j \text{ equal to the computed}$$

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.

RESULTS AND DISCUSSION

In Fig. 4, a 10×10 SOM resulting from the 565 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, at the middle, and at the lower left corner of the map, whereas specimens tested at 90°C and 120°C tend to cluster at the lower right corner. Most importantly, since the SOM contains many specimens which were tested at 30°C, this gives an indication that 30°C is the temperature that has the highest impact on the characteristics and properties of the studied nanocomposites specimens in the designed framework. The testing temperature of 120°C is also important (since the SOM has a small cluster of 120°C in the lower right corner) but not as critical as 30°C.

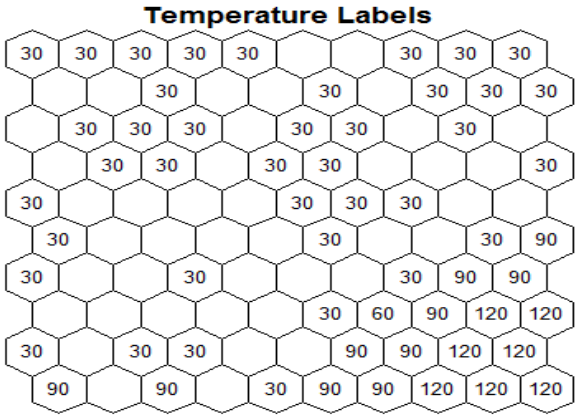


Fig. 4. A 10×10 SOM with respect to temperature for the 565 nanocomposite specimens used in the study (with all seventeen dimensions). The specimens tested at the same temperature tend to cluster together and 30°C is the testing temperature that drives the characteristics of all specimens in the designed framework.

In Figs. 5, 6, and 7 three 10×10 SOMs for the VGCNF high shear mixing time, VGCNF sonication time, and the tan delta response are shown, respectively. In Fig. 5, specimens with the same or close values of high shear mixing time tend to cluster together especially at the top and at the middle of the SOM. This means that high shear mixing time is important in the newly designed VGCNF/VE framework. However, this tendency is not consistent and is less than the clustering tendency shown in Fig. 4 for temperature. Similarly, in Fig. 6, specimens with the same or close values of sonication time tend to cluster together. However, compared with Fig. 5, the clustering tendency of the specimens based on the sonication time is more pronounced than that of the high shear mixing time but less than the clustering tendency for temperature in Fig. 4. In Fig. 7, VGCNF/VE specimens with the same tan delta response values tend to cluster together and the clustering tendency is more consistent than that of high shear mixing time and sonication time. This leads to the conclusion that both tan delta and temperature are dominant features for the treatment combinations and have the highest impact on the responses for all the specimens in the framework followed by the sonication time and then by the high shear mixing time.

Another observation that can be seen from Fig. 7 is that most specimens in the SOM have a tan delta of 0.00 which were clustered at the top and at the middle of the map. This means that the specimens with no delta values, i.e. compression, tension specimens and flexural specimens are essential components in the new framework. In addition, by comparing Fig. 4 with Fig. 7, most specimens of testing temperature 30°C have a corresponding 0.00 tan delta response value. This leads to the conclusion that compression and tension specimens and flexural specimens treated at 30°C are very important in the new nanocomposites framework.

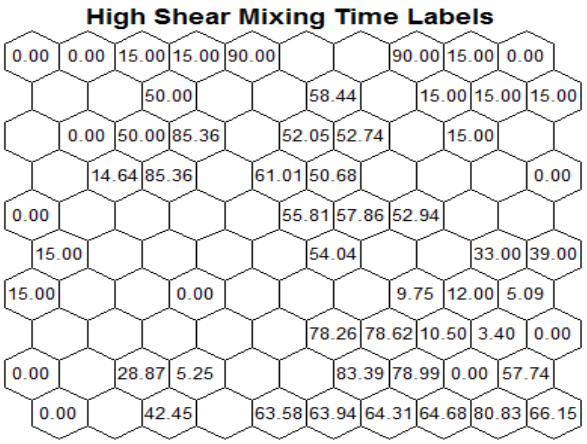


Fig. 5. A 10×10 SOM with respect to VGCNF high shear mixing time. The clustering tendency is less than that of the temperature in Fig. 4 and can be seen clearly at the top and at the middle of the SOM.

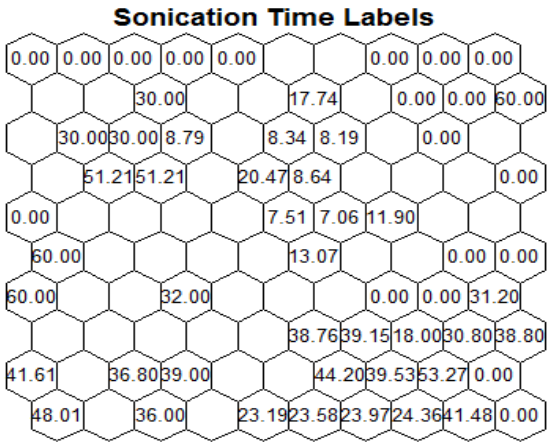


Fig. 6. A 10×10 SOM with respect to VGCNF sonication time. The clustering tendency is less than that of the temperature in Fig. 4 but more than that of high shear mixing time in Fig. 5. This can be clearly seen at the top, and the middle, and at the lower right corner of the map

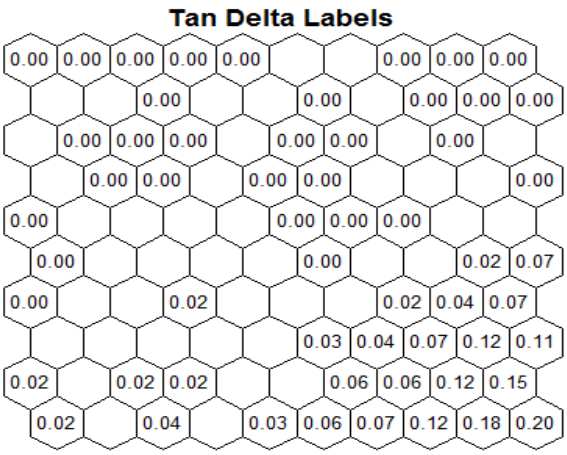


Fig. 7. A 10×10 SOM with respect to tan delta values. The clustering tendency is similar to that of the temperature in Fig. 4. However, specimens treated at 30°C have a corresponding 0.00 tan delta value. This means that Compression and tension specimens and flexural specimens tested at 30°C are essential components in the new framework.

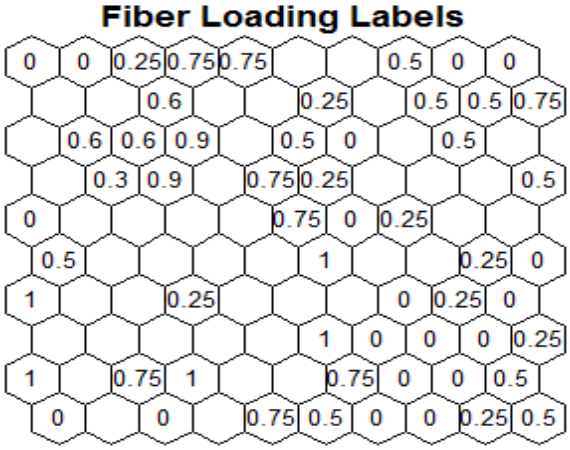


Fig. 8. A 10×10 SOM with respect to VGCNF fiber loading (VGCNF weight fractions) values. The clustering tendency is inconsistent throughout the SOM and so VGCNF fiber loading is *not* dominant for the treatment combinations of the newly designed framework

In Fig. 8, a 10×10 SOM for the VGCNF fiber loading (sometimes referred to as VGCNF weight fraction) is shown. Unlike the sound impact of VGCNF weight fraction on the viscoelastic properties of VCCNF/VE based on the study conducted by AbuOmar *et. al*[19], this feature does not have that much impact on the new nanocomposites framework as the clustering tendency is inconsistent throughout the SOM.

In Figs. 9, 10, 11, 12, 13, 14, 15, and 16, eight 10×10 SOMs for the responses of true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, and loss modulus are shown, respectively. From Figs. 9, 10, 11, and 12 specimens that have no true ultimate strength responses, no true yield strength responses, no engineering elastic modulus responses, and no engineering ultimate strength responses are important in the new framework. This can be seen in four large clusters of 0.00 in Figs 9, 10, 11, and 12. This means that VGCNF/VE flexural and viscoelastic specimens are essential in the new framework. Similarly, from Figs. 13 and 14, one can see that specimens with no flexural modulus and flexural strength responses are also important in the new framework. Particularly, compression and tension specimens and viscoelastic specimens are essential in the new design. In addition, from Figs. 15 and 16, specimens that don't have any responses for storage modulus and loss modulus are important in the whole framework; namely the compression and tension specimens and the flexural specimens. This leads to the conclusion that all the components (i.e., VGCNF/VE compression, tension, flexural, and viscoelastic specimens) of the newly designed nanocomposites framework are essential. Therefore, the validity of the selection of these nanocomposites structures in the framework is confirmed by this observation.

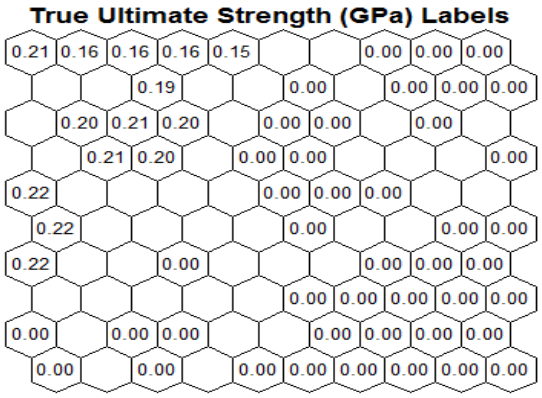


Fig. 9. A 10×10 SOM based on the true ultimate strength response. The values are rounded for simplicity.

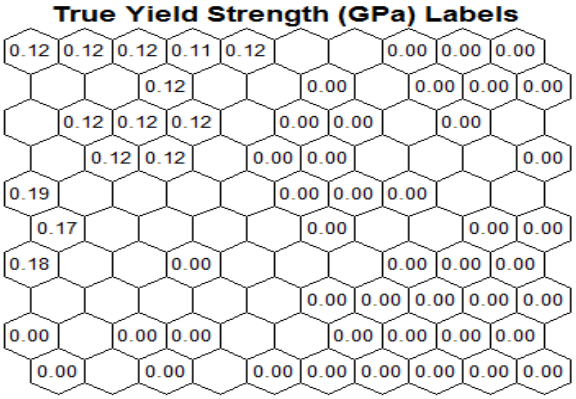


Fig. 10. A 10×10 SOM based on the true yield strength response. The values are rounded for simplicity.

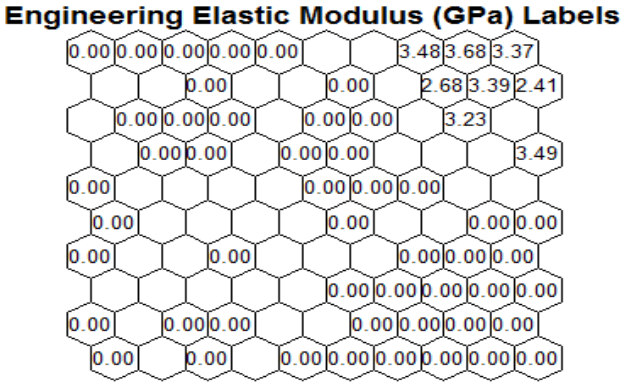


Fig. 11. A 10×10 SOM based on the engineering elastic modulus response.

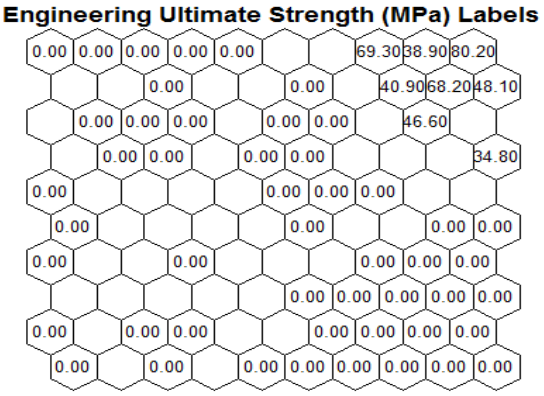


Fig. 12. A 10×10 SOM based on the engineering ultimate strength response.

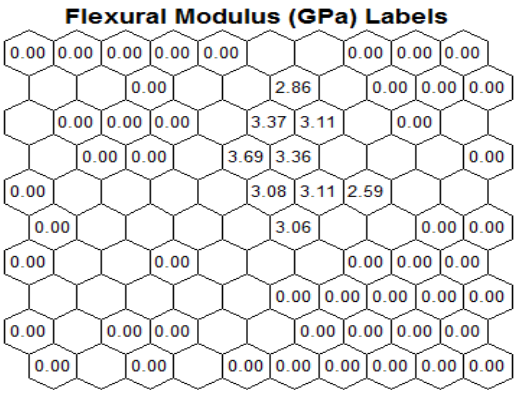


Fig. 13. A 10×10 SOM based on the flexural modulus response.

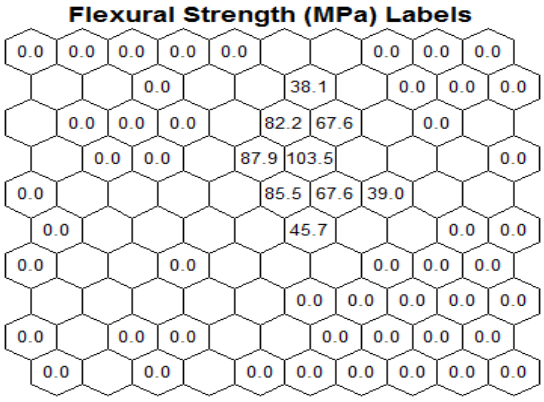


Fig. 14. A 10×10 SOM based on the flexural strength response.

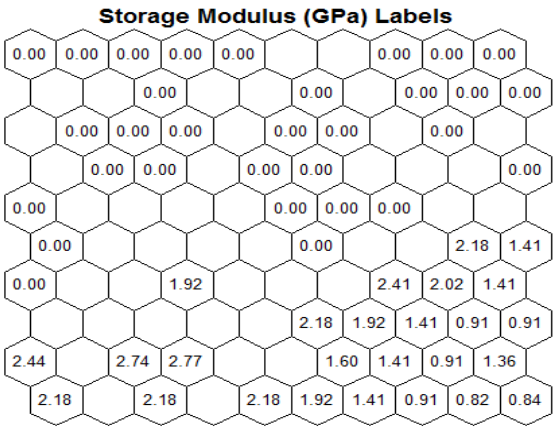


Fig. 15. A 10×10 SOM based on the storage modulus response.

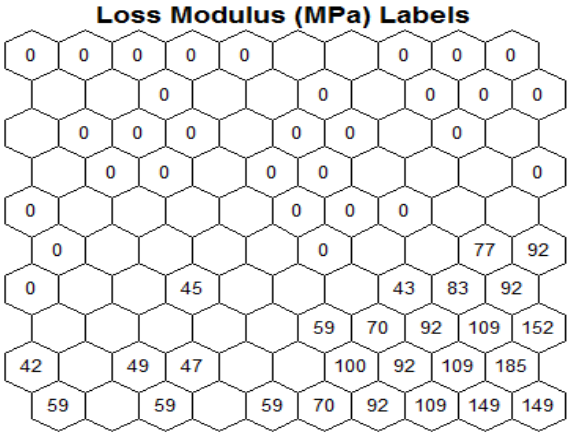


Fig. 16. A 10×10 SOM based on the loss modulus response. The values are rounded to the nearest integer for simplicity.

In addition to the sensitivity analysis inferred from SOMs, the different conditions needed to produce a particular optimal (highest) response can also be determined. In Fig. 17, 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 565 with specific values of curing environment, use of a dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time, temperature, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta. The indices in Fig. 17 can be used to extract information linking the different dimensional combinations that produce the optimal response values.

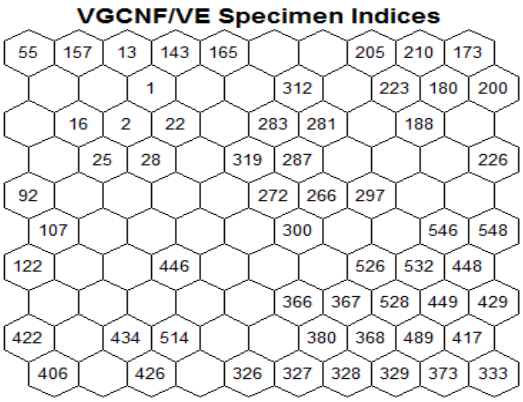


Fig. 17. A 10×10 SOM illustrating the indices (numeric orders) of the 565 nanocomposite specimens of the new framework[20, 23, 29].

From Fig. 9, a group of three specimens have the highest true ultimate strength of about 0.22 GPa, located at the fifth, sixth, and seventh rows of the SOM. In Fig. 17, these values correspond to specimen indices 92, 107, and 122. Clearly, different input properties can be determined to produce the same 0.22 GPa response value. These properties are shown in Table 2. Nanocomposite designers can use such information in the selection of input factor levels.

Table 2: Different dimensional (factorial) combinations required to produce an optimal true ultimate strength of about 0.22 GPa for the three specimens.

Optimal ultimate strength response value for the three specimens = 0.22 GPa	
Curing environment	Oxygen, Oxygen, Oxygen
Use of a dispersing agent	Yes, Yes, Yes
Mixing method	HS, US/HS, US/HS
VGCNF fiber loading (phr)	0.00, 0.50, 1.00
VGCNF type	Pristine, Pristine, Pristine
High shear mixing time (min)	0.00, 15.00, 15.00
Sonication time (min)	0.00, 60.00, 60.00
Temperature (°C)	30.00, 30.00, 30.00

From Fig. 10, one specimen has the highest true yield strength of about 0.19 GPa, located at the fifth row of the SOM. In Fig. 17, this value corresponds to specimen index of 92 and different input properties can be determined to produce this response value. These properties are shown in Table 3.

Table 3: Different dimensional (factorial) combinations required to produce an optimal true yield strength of about 0.19 GPa.

Optimal true yield strength response value = 0.19 GPa	
Curing environment	Oxygen
Use of a dispersing agent	Yes
Mixing method	HS
VGCNF fiber loading (phr)	0.00
VGCNF type	Pristine
High shear mixing time (min)	0.00
Sonication time (min)	0.00
Temperature (°C)	30.00

From the sensitivity analysis that was conducted based on SOMs above, it was confirmed that temperature, sonication time, and high shear mixing time are the most dominant factors in the new nanocomposites framework. Therefore, one can focus on the quantities of these factors when determining the optimal conditions required to achieve the optimal response value(s). From Table 2, high shear mixing time can be 0.00 or 15.00 minutes and sonication time can be 0.00 or 60 minutes and temperature must be at 30^{o}C. However, since the specimen of index 92 achieves both the optimal values of true ultimate strength and true yield strength, then very low values of high shear mixing and sonication times must be used at the testing temperature of 30^{o}C (Tables 2 and 3).

From Fig. 11, one specimen has the highest engineering elastic modulus of about 3.68 GPa, located at the first row of the SOM. In Fig. 17, this value corresponds to specimen index of 210 and different input properties can be determined to produce this response value. These properties are shown in Table 4.

Table 4: Different dimensional (factorial) combinations required to produce an optimal engineering elastic modulus of about 3.68 GPa.

Optimal engineering elastic modulus response value = 3.68 GPa	
Curing environment	Nitrogen
Use of a dispersing agent	Yes
Mixing method	HS
VGCNF fiber loading (phr)	0.00
VGCNF type	Oxidized
High shear mixing time (min)	0.00
Sonication time (min)	0.00
Temperature (^{o}C)	30.00

From Fig. 12, one specimen has the highest engineering ultimate strength of about 80.20 MPa, located at the first row of the SOM. In Fig. 17, this value corresponds to specimen index of 173 and different input properties can be determined to produce this response value. These properties are shown in Table 5.

Table 5: Different dimensional (factorial) combinations required to produce an optimal engineering ultimate strength of about 80.20 MPa.

Optimal engineering ultimate strength response value = 80.20 MPa	
Curing environment	Nitrogen
Use of a dispersing agent	No
Mixing method	US
VGCNF fiber loading (phr)	0.00
VGCNF type	Oxidized
High shear mixing time (min)	0.00
Sonication time (min)	0.00
Temperature (^{o}C)	30.00

From Tables 4 and 5, in order to produce specimens with optimal values of engineering elastic modulus and engineering ultimate strength, the high shear mixing time and the sonication time must be very low and the testing temperature of 30^{o}C must be used.

From Fig. 13, one specimen has the highest flexural modulus of about 3.69 GPa, located at the fourth row of the SOM. In Fig. 17, this value corresponds to specimen index of 319 and different input properties can be determined to produce this response value. These properties are shown in Table 6.

Table 6: Different dimensional (factorial) combinations required to produce an optimal flexural modulus of about 3.69 GPa.

Optimal flexural modulus response value = 3.69 GPa	
Curing environment	Oxygen
Use of a dispersing agent	Yes
Mixing method	HS
VGCNF fiber loading (phr)	0.75
VGCNF type	Oxidized
High shear mixing time (min)	61.01
Sonication time (min)	20.47
Temperature (ºC)	30.00

From Fig. 14, one specimen has the highest flexural strength of about 103.5 MPa, located at the fourth row of the SOM. In Fig. 17, this value corresponds to specimen index of 287 and different input properties can be determined to produce this response value. These properties are shown in Table 7.

Table 7: Different dimensional (factorial) combinations required to produce an optimal flexural strength of about 103.5 MPa.

Optimal flexural strength response value = 103.5 MPa	
Curing environment	Oxygen
Use of a dispersing agent	Yes
Mixing method	HS
VGCNF fiber loading (phr)	0.25
VGCNF type	Pristine
High shear mixing time (min)	50.68
Sonication time (min)	8.64
Temperature (ºC)	30.00

From Tables 6 and 7, in order to achieve the optimal values of flexural modulus and flexural strength, generally low sonication time and relatively high values of high shear mixing time must be used. Testing temperature must be at 30ºC.

From Fig. 15, two specimens have the highest storage modulus of about 2.76 GPa, located at the ninth row of the SOM. In Fig. 17, this value corresponds to specimen indices of 434 and 514 and clearly different input properties can be determined to produce this response value. These properties are shown in Table 8.

Table 8: Different dimensional (factorial) combinations required to produce an optimal storage modulus of about 2.76 GPa for the two specimens.

Optimal storage modulus response value for the two specimens = 2.76 GPa	
Curing environment	Oxygen, Oxygen
Use of a dispersing agent	Yes, Yes
Mixing method	HS, US/HS
VGCNF fiber loading (phr)	0.50, 0.50
VGCNF type	Pristine, Pristine
High shear mixing time (min)	28.87, 5.25
Sonication time (min)	36.80, 39.00
Temperature (ºC)	30.00, 30.00

From Fig. 16, two specimens have the highest loss modulus of about 149 MPa, located at the tenth row of the SOM. In Fig. 17, this value corresponds to specimen indices of 333 and 373 and clearly different input properties can be determined to produce this response value. These properties are shown in Table 9.

Table 9: Different dimensional (factorial) combinations required to produce an optimal loss modulus of about 149 MPa for the two specimens.

Optimal loss modulus response value for the two specimens = 149 MPa	
Curing environment	Oxygen, Oxygen
Use of a dispersing agent	No, No
Mixing method	US, US
VGCNF fiber loading (phr)	0.25, 0.25
VGCNF type	Pristine, Oxidized
High shear mixing time (min)	66.15, 80.83
Sonication time (min)	25.91, 41.48
Temperature (ºC)	120.00, 120.00

From Tables 8 and 9, in order to produce specimens with optimal storage modulus values, the high shear mixing time must be low with moderately high sonication time and the testing temperature must be at 30ºC. On the hand, the optimal values of loss modulus are obtained by using relatively high values of high shear mixing time and lower sonication time. The testing temperature in this case must be higher at 120ºC.

A PCA was run on the VGCNF/VE data in the newly designed nanocomposite framework. Fig. 18 shows a graphical representation for the PCA of the data. PCA reduced the number of data dimensions from seventeen 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 and discover certain physical/mechanical behavior associated with the data without running additional experiments.

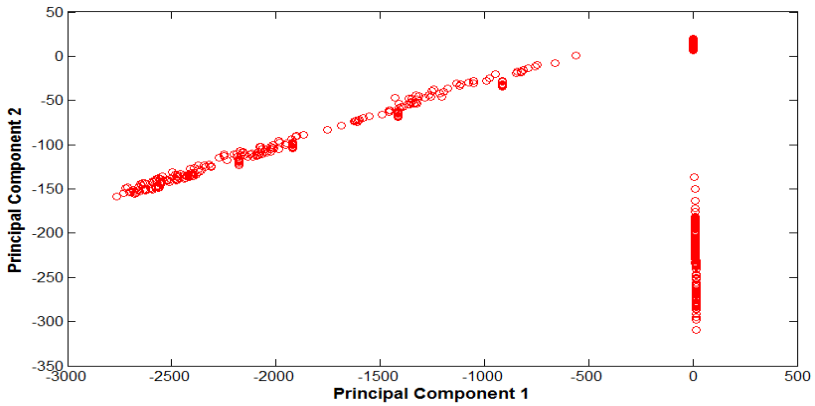
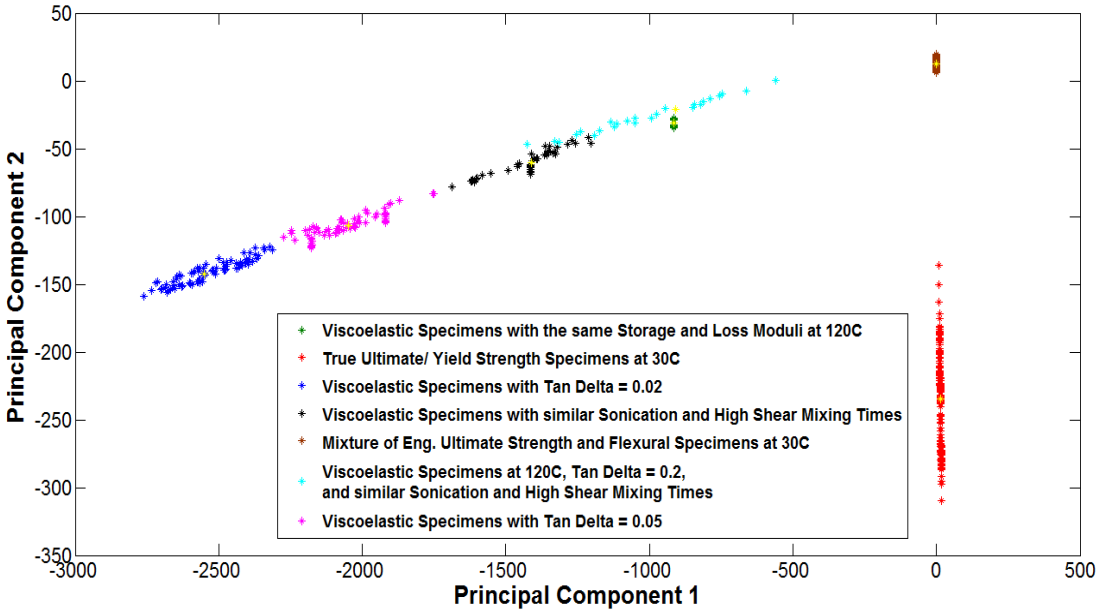
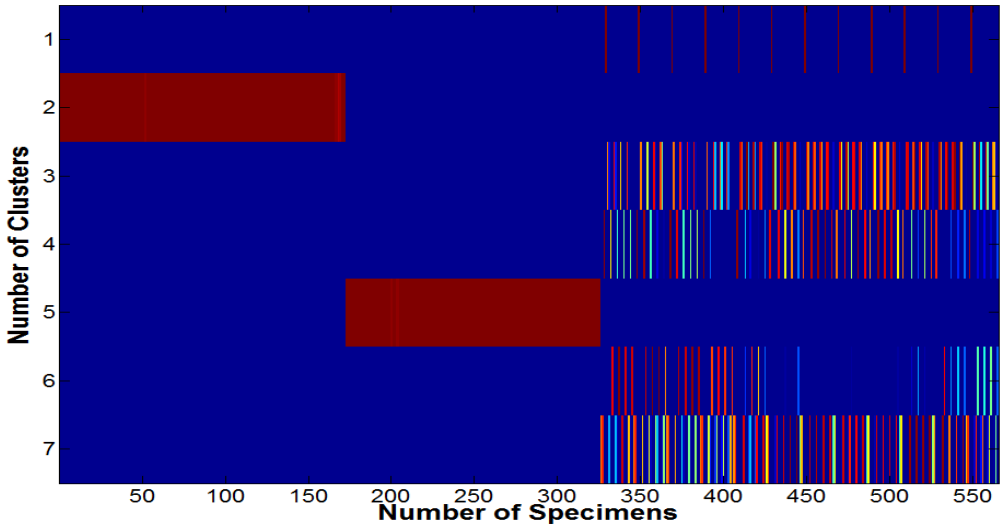


Fig. 18. A 2-D graphical representation of the VGCNF/VE nanocomposite specimen data in the newly designed framework (illustrated by circle points) using the PCA technique. This technique maps the data from a 17-D space down to a 2-D space so that different clustering algorithms can be applied. The values 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.

The FCM was applied to the new VGCNF/VE nanocomposite data using the GK distance measures. In Fig. 19, the FCM results are illustrated, where seven clusters are chosen to represent the data using the GK distance measure. In Fig. 19a, viscoelastic VGCNF/VE specimens were divided into seven different clusters. Particularly, one cluster with viscoelastic specimens that have the same tan delta response = 0.02, one cluster with viscoelastic specimens that have the same tan delta response = 0.05, one cluster with viscoelastic specimens with similar sonication and high shear mixing times, one cluster with viscoelastic specimens that have the same storage and loss moduli responses at the testing temperature of 120^{o}C, and one cluster with viscoelastic specimens tested at 120^{o}C with the same tan delta response value = 0.2 and have similar sonication and high shear mixing times. This leads to the conclusion that VGCNF/VE viscoelastic data is an essential component in the new framework. In addition, tan delta response is a dominant feature in this material system followed by the sonication time and high shear mixing time as some specimens were placed in some clusters based on their sonication and high shear mixing times. Moreover, the testing temperature of 120^{o}C is important in the framework as some viscoelastic specimens were placed in two clusters based on that temperature. In addition, specimens with engineering ultimate strength response and flexural data tested at 30^{o}C were clustered together in one cluster and specimens with true ultimate and yield strength responses tested at 30^{o}C were also placed in a separate cluster. This means that the testing temperature of 30^{o}C is a dominant feature in the framework as well as both flexural and compression and tension specimens are also important in this material system. In Fig. 19b, a "scale data and display image (imagesc) object" plot is presented to indicate the number of clusters, 7 in this case, (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. 19a. These findings prove that temperature is a dominant feature for the whole dataset.



a)



b)

Fig. 19. a) Clustering results after applying the FCM algorithm and the GK distance measure, when $C = 7$. b) In the "scale data and display image object (imagesc)" plot, seven bands representing seven clusters can be identified.

Using the GK distance measure, FCM works better for the 565 VGCNF/VE specimens when the selected number of clusters equals seven. For this case, specimens tested at different temperatures (particularly at 30°C and 120°C) and have the same tan delta responses tend to be located in separate clusters that distinguish each of these temperatures and tan delta values. In addition, when the number of clusters equals to seven, more features that have pronounced effect in the new nanocomposite framework can be identified. For example, sonication time and high shear mixing time has come out to be important in the framework after applying FCM when seven clusters were selected. Also, viscoelastic specimens tested at 120°C and have the same storage modulus and loss modulus responses have similar physical and mechanical behavior as they were placed in one separate cluster. These results confirmed some of the SOM findings above in that tan delta, temperature, sonication time, and high shear mixing time are the most dominant features in this material system and suggest that the FCM algorithm was able to identify VGCNF/VE specimens in the framework that have similar properties and placed them into different clusters based on tan delta, temperature, sonication time, high shear mixing time, and specimens type/ structure.

The SOM analysis allows a preliminary visual identification of the different existing groups[27]. 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[25]. 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 specimens and their associated viscoelastic, flexural, compression, and tension 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 material system in the newly designed nanocomposite framework.

SUMMARY AND CONCLUSIONS

Knowledge discovery and data mining techniques were applied to a unified framework of different vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposite structures as a case study for materials informatics. This dataset had been generated by a full factorial experimental design with 565 different design points representing three different nanocomposite structures, VGCNF/VE viscoelastic data, flexural data, compression, and tension data. Each treatment combination in the design consisted of seventeen feature dimensions corresponding to the design factors, i.e., curing environment, use or absence of dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time and testing temperature were utilized as inputs and the true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta were selected as outputs. Self-organizing maps (SOMs) were created with respect to temperature, tan delta, high shear mixing time, sonication time, VGCNF fiber loading, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus and loss modulus. After analyzing the SOMs, temperature and tan delta were identified as the most dominant features for the newly designed VGCNF/VE nanocomposites framework having the highest impact on the material responses in the framework. Sonication time and high shear mixing time were also important. 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 similar tan delta values. The cluster with the highest number of specimens in the "temperature labels" SOM is the 30^{o}C cluster. This means that 30^{o}C is the most important temperature as it drives the behavior of all specimens in the newly designed framework. Analyzing the SOMs with respect to true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus and loss modulus demonstrated that VGCNF/VE specimens with different features could be designed to match an optimal value of VGCNF/VE compression response and/or VGCNF/VE tension response and/or VGCNF/VE flexural response and/or VGCNF/VE viscoelastic response.

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, tan delta values as well as sonication time and high shear mixing time. However, the testing temperature of 30^{o}C and 120^{o}C were the most important temperatures as specimens were clustered based on these two particular temperatures. In addition, the FCM was able to recognize the viscoelastic specimens tested at 120^{o}C and have the same storage and loss modulus values and placed them in one cluster. This reflects the fact that the mechanical and physical properties of these specimens are similar. In addition, when seven clusters were selected and the GK distance measure was applied, there was one cluster that had only VGCNF/VE compression and tension specimens, one cluster that had a mixture of engineering ultimate strength and flexural specimens, and five clusters that had VGCNF/VE viscoelastic specimens with different properties. This means that all nanocomposites structures in the framework are important and the VGCNF/VE viscoelastic specimens are the most important structure. Moreover, the FCM algorithm worked better when the number of clusters equals seven, because high shear mixing time came out to be an important feature in the new framework. In addition, when the number of

clusters equal seven, the viscoelastic specimens that have the same storage and loss moduli tend to be placed in one cluster.

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 new VGCNF/VE system, whether they are input design factors or output responses.

- Developing a tool for identifying VGCNF/VE specimen designs leading to the optimal (highest) responses of true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus and loss modulus. This will facilitate tailoring of nanocomposite viscoelastic, compression, tension, and flexural properties and, in turn, minimize fabrication costs and increase the production efficiency by the domain experts.

- Developing a methodology to better identify cogent patterns and trends in VGCNF/VE data in the new framework. Each cluster can be identified based on one or more of the input design factors of the new 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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] I. WITTEN, E. FRANK, and M.A. HALL, *Data Mining: Practical Machine Learning Tools and Techniques*, 2011, Elsevier Science.

[2] Z.Q. Lu: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, *Royal Statistical Society: Series A (Statistical Society)*, 173 (2010), 693-694.

[3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth: From Data Mining to Knowledge Discovery in Databases, *AI magazine*, 17 (1996), 37.

[4] D.T. LAROSE, *An Introduction to Data Mining*, 2005, Traduction et adaptation de Thierry Vallaud.

[5] K. Rajan: Materials Informatics, *Materials Today*, 8 (2005), 38-45.

[6] K.F. Ferris, L.M. Peurrung and J.M. Marder: Materials Informatics: Fast Track To New Materials, *Advanced Materials and Processes*, 165 (1) (2007), 50-51.

[7] C. Suh, K. Rajan, B.M. Vogel, B. Narasimhan and 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 and 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 and M.F. Horstemeyer, On Dislocation-Based Artificial Neural Network Modeling Of Flow Stress, *Materials and Design*, 31 (2010), 3683-3689.

[11] T. Sabin, C. Bailer-Jones and P. Withers: Accelerated Learning Using Gaussian Process Models to Predict Static Recrystallization in an Al-Mg Alloy, *Modelling and Simulation in Materials Science and Engineering*, 8 (2000), 687.

[12] A. Javadi and 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 and Y. Shinagawa: Construction Site Image Retrieval Based On Material Cluster Recognition, *Advanced Engineering Informatics*, 20 (2006), 443-452.

[14] A. Ullah and 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*, 2006, McGraw-Hill, New York, NY.

[16] J. Garces, D.J. Moll, J. Bicerano, R. Fibiger and D.G. McLeod: Polymeric Nanocomposites for Automotive Applications, *Advanced Materials*, 12 (2000), 1835-1839.

[17] F. Hussain, M. Hojjati, M. Okamoto and 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 and T.W. Chou: Nanocomposites in Context, *Composites Science and Technology*, 65 (2005), 491-516.

[19] O. Abuomar, S. Nouranian, R. King, J.L. Bouvard, H. Toghiani, T. E. Lacy and C. U. Pittman Jr: Data Mining and Knowledge Discovery in Materials Science and Engineering: A Polymer Nanocomposites Case Study, *Advanced Engineering Informatics*, 27 (2013), 615-624

[20] S. Nouranian: Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites: Designed Experimental Study of Mechanical Properties and Molecular Dynamics Simulations, (2011), Mississippi State University, *PhD Dissertation*, Mississippi State, MS USA.

[21] S. Nouranian, H. Toghiani, T.E. Lacy, C.U. Pittman and 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.

[22] S. Nouranian, T.E. Lacy, H. Toghiani, C.U. Pittman Jr and 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.

[23] G.W. Torres, S. Nouranian, T.E. Lacy, H. Toghiani, C.U. Pittman Jr and 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.

[24] R. King, A. Rosenberger and L. Kanda: Artificial Neural Networks and Three-Dimensional Digital Morphology: A Pilot Study, *Folia Primatol*, 76 (2005), 303-324.

[25] T. KOHONEN, *Self-Organization and Associative Memory*, 1988, Springer-Verlag.

[26] S. MIYAMOTO, H. ICHIHASHI and K. HONDA, *Algorithms for Fuzzy Clustering: Methods in C-Means Clustering with Applications*, 2008, Springer.

[27] J.C. Bezdek and R. Ehrlich: FCM: The Fuzzy C-Means Clustering Algorithm, *Computers and Geosciences*, 10 (1984), 191-203.

[28] I.T. JOLLIFFE, *Principal Component Analysis*, 2002, Springer.

[29] J. Lee, S. Nouranian, G.W. Torres, T. E. Lacy, H. Toghiani, C. U. Pittman and J. L. DuBien: Characterization, Prediction, and Optimization of Flexural Properties of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites by Response Surface Modeling, *Journal of Applied Polymer Science*, 130 (2013), 2087-2099. doi: 10.1002/app.39380

[30] G.Tibbetts, M. Lake, K. Strong and B. Rice: A Review of the Fabrication and Properties of Vapor-Grown Carbon Nanofiber/Polymer Composites, *Composites Science and Technology*, 67 (2007), 1709-1718.

[31] A. Plaseied, A. Fatemi, and M. Coleman: Influence of Carbon Nanofiber Content and Surface Treatment on Mechanical Properties of Vinyl Ester, *Polymers and Polymer Composites*, 16 (2008), 405-413.

[32] A. Plaseied and A. Fatemi: Tensile Creep and Deformation Modeling of Vinyl Ester Polymer and Its Nanocomposite, *Journal of Reinforced Plastics and Composites*, 28 (2009), 1775.

[33] D.C. MONTGOMERY, *Design and Analysis of Experiments*, 2009, 7th ed., John Wiley & Sons, Hoboken, NJ.

[34] *MATLAB Mathematics and Interpolation*, Release 2012a, The MathWorks, Inc., Natick, Massachusetts, United States.

[35] J.W. Sammon Jr: A Nonlinear Mapping for Data Structure Analysis, *IEEE Transactions on Computers*, 100 (1969), 401-409.