



Using regression models for predicting the product quality in a tubing extrusion process

Vicente García¹ · J. Salvador Sánchez² · Luis Alberto Rodríguez-Picón³ · Luis Carlos Méndez-González³ · Humberto de Jesús Ochoa-Domínguez⁴

Received: 23 April 2017 / Accepted: 26 March 2018 / Published online: 31 March 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Quality in a manufacturing process implies that the performance characteristics of the product and the process itself are designed to meet specific objectives. Thus, accurate quality prediction plays a principal role in delivering high-quality products to further enhance competitiveness. In tubing extrusion, measuring of the inner and outer diameters is typically performed either manually or with ultrasonic or laser scanners. This paper shows how regression models can result useful to estimate both those physical quality indices in a tube extrusion process. A real-life data set obtained from a Mexican extrusion manufacturing company is used for the empirical analysis. Experimental results demonstrate that k nearest-neighbor and support vector regression methods (with a linear kernel and with a radial basis function) are especially suitable for predicting the inner and outer diameters of an extruded tube based on the evaluation of 15 extrusion and pulling process parameters.

Keywords Regression models · Extrusion process · Product quality prediction · Support vector regression · K nearest-neighbor

Introduction

Tubings and pipes are manufactured using different methods, but extrusion is probably the most efficient one. This complex thermoforming process involves heating a raw material (usually plastic, metal, polymer, concrete or ceramic) and forming a final ring-shaped product. A primary advantage of extrusion over other manufacturing processes is its capability to create a very complex cross-sectional profile object (Oberg et al. 2012). However, extrusion processing compresses many interdependent input parameters (both process and system

variables) and output parameters. Process variables refer to the operating conditions that can be controlled and manipulated directly, whereas the system variables are determined by the process parameters and have an impact on the output parameters (Chevanan et al. 2007).

Unlike the cyclic techniques such as injection molding or blow molding, extrusion is a steady-state or continuous process. This means that, for example, a change in the parameters of the extruder will disrupt the steady-state process condition with a non-negligible effect on the quality of the extruded product. Therefore, all input parameters must be identified, controlled and monitored to guarantee success in the extrusion process since manufacturing quality prediction, control, and monitoring are critical (Khan et al. 2014). Common deficiencies of extruded products are related to visual or geometrical characteristics (e.g., diameter variations, color changes and rough surface) and physical or mechanical properties (e.g., elasticity and rigidity).

Several intelligent and soft computing models (Witten et al. 2011) have been applied to a large variety of manufacturing processes, such as production, fault detection, process planning and monitoring, machine maintenance, and quality prediction and control (Charaniya et al. 2010; Choudhary et al. 2008; Harding et al. 2006; Köksal et al. 2011; Kusiak 2006; Pratihari 2015; Yin et al. 2015). In particular, the use of

✉ Vicente García
vicente.jimenez@uacj.mx

¹ División Multidisciplinaria en Ciudad Universitaria, Departamento de Ingeniería Eléctrica y Computación, Universidad Autónoma de Ciudad Juárez, C. José de Jesús Macías Delgado, 18100 Ciudad Juárez, Chihuahua, Mexico

² Institute of New Imaging Technologies, Universitat Jaume I, Av. Vicent Sos Baynat s/n, 12071 Castelló de la Plana, Spain

³ Departamento de Ingeniería Industrial y Manufactura, Universidad Autónoma de Ciudad Juárez, Av. del Charro 450 Norte, 32310 Ciudad Juárez, Chihuahua, Mexico

⁴ Departamento de Ingeniería Eléctrica y Computación, Universidad Autónoma de Ciudad Juárez, Av. del Charro 450 Norte, 32310 Ciudad Juárez, Chihuahua, Mexico

these techniques for machinery fault detection and product quality prediction has received increasing attention over the last years.

Krömer et al. (2010) showed the ability of genetic programming to evolve fuzzy classifiers on a real-world problem for detecting faulty products in an industrial production process. Multi-layer perceptron neural networks were employed to predict errors in mold surface roughness (Erzurumlu and Oktem 2007) and the product quality in a wave soldering process (Liukkonen et al. 2009). Support vector machines (Jiang et al. 2013) and radial basis function neural networks (Zhang et al. 2014) were used to predict the quality of propylene polymerization in industrial processes. Chien et al. (2007) applied the K-means clustering algorithm and decision trees for the detection of defects in semiconductor manufacturing. The rough set approach was applied to find out solder defects in printed circuit boards (Kusiak and Kurasek 2001). Quality prediction in plastic injection molding processes was tackled using back-propagation neural networks (Sadeghi 2000), support vector machines (Ribeiro 2005) and genetic algorithms (Meiabadi et al. 2013). A combined method based on artificial neural network and particle swarm optimization was proposed to improve the mechanical performance of polymer products (Xu et al. 2015). Adly et al. (2015) presented a simplified subspace regression algorithm for accurate identification of defect patterns in semiconductor wafer maps. Two evolutionary fuzzy ARTMAP neural networks were designed by Tan et al. (2015) to deal with the class imbalance problem in semiconductor manufacturing operations. Ghorai et al. (2013) developed a visual inspection system to localize defects on hot-rolled steel surfaces employing some kernel classifiers, such as the support vector machine and the vector-valued regularized kernel function approximation. Wu et al. (2017) introduced a method based on random forests for tool wear prediction and compared its performance with that of support vector regression and feed-forward back-propagation neural networks. Wang et al. (2018) presented a comprehensive survey of deep learning algorithms for smart manufacturing.

With regards to the particular case of product quality prediction in extrusion processes, we can pay attention to a set of works that have employed some soft computing techniques. For instance, Wu and Hsu (2002) combined the finite-element approach, a polynomial network and a genetic algorithm to develop a method for the design of the optimal shape of an extrusion die. Li et al. (2004) adopted the cooperation between a three-layer back-propagation neural network and a genetic algorithm to set up the system and optimize the technical parameters in the semi-solid extrusion of composite tubes and bars. Yu et al. (2004) proposed a strategy based on a fuzzy neural-Taguchi network and a genetic algorithm to determine the optimal die gap programming of extrusion blow molding processes. Oke et al. (2006) optimized the flow

rate of the plastic extrusion process in a plastic recycling plant with the application of a neuro-fuzzy model. González Marcos et al. (2007) introduced improvements in the rubber extrusion process by predicting the characteristics of rubber with a multi-layer perceptron neural network. Sharma et al. (2009) suggested a model of forward mapping for hot extrusion process using the ANFIS neuro-fuzzy approach. Hsiang et al. (2012) investigated the optimal process parameters that maximize the multiple performance characteristics index for hot extrusion of magnesium alloy bicycle carriers through a fuzzy-based Taguchi method. Ramana and Reddy (2013) proposed to make use of clustering, naïve Bayes, and decision trees to predict and improve the final product quality in a plastic extrusion process. Zhao et al. (2013) employed a Pareto-based genetic algorithm for optimization of port-hole extrusion die. Support vector regression models and multi-layer perceptron neural networks were compared for the prediction of specific properties of rubber extruded mixtures (Urraca Valle et al. 2013). Carrano et al. (2015) employed an evolutionary computing algorithm to optimize the operational and screw geometrical parameters of a single screw polymer extrusion system. One-class classification methods were used by Kohlert and König (2015) for yield optimization of an extrusion process in a polymer film industry. Chondronasios et al. (2016) introduced a feature extraction technique based on gradient-only co-occurrence matrices to detect blisters and scratches on the surface of extruded aluminum profiles using a two-layer feed-forward artificial neural network.

The main purpose of this paper, therefore, is to analyze the performance of some regression models in the prediction of product quality (regarding the inner and outer diameters) in a tubing extrusion process. From an application perspective, the novelty of this study is on the specific solution proposed for product quality control in a plastic tube manufacturing plant. To the best of our knowledge, there are no previous reports that analyze the use of parameters taken from the extrusion and pulling processes to predict the inner and outer diameters of an extruded tube using the regression methods considered here.

Henceforth the paper is organized as follows. “Description of the tubing extrusion process” section describes the tubing extrusion process of a Mexican manufacturing company, which provided us with the database used for the subsequent empirical analysis. “Regression models” section introduces the bases of the regression models that will be explored in this study. Next, “Experimental set-up” section presents the experimental set-up and the performance evaluation criteria used in the experiments, while the results are given and discussed in “Results and discussion” section. Finally, “Conclusions and future work” section summarizes the main conclusions and outlines some possible avenues for future research.

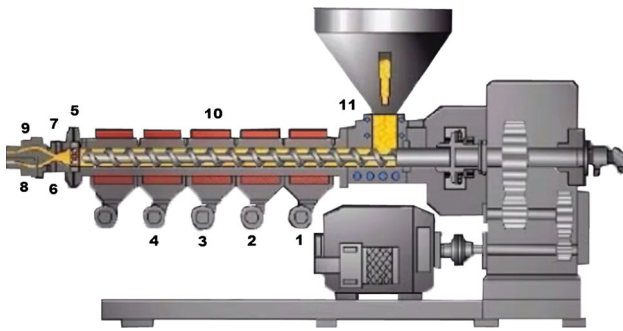


Fig. 1 Extrusion process. Process parameters that determine the quality of the tube: [1–4] zone temperatures, [5] melting temperature, [6–9] die temperature, [10] revolutions per minute of the screw, and [11] base hopper temperature

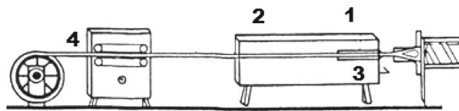


Fig. 2 Pulling process. Process parameters that determines the quality of the tube: [1–2] tank temperature, [3] vacuum pressure, and [4] tension of the pulling mechanism

Description of the tubing extrusion process

This section provides a general description of the tube extrusion process used by a manufacturing company located in Ciudad Juárez (Chihuahua, Mexico). Thus the extrusion process consists of two stages. In the initial phase, the plastic is fed into the heating chamber of the extruder to melt it (see Fig. 1). Once molten, the plastic is pushed by a screw device through the shaped die, which forms the plastic into a tube-shaped form.

In the second phase depicted in Fig. 2, the extruded tube is pulled by a mechanism through a water tank or a blowing system to cool it down and get the final form.

A defect can be defined as a deviation of the product characteristics from the specifications set up by the manufacturing process (Khan et al. 2014), or the difference between the desired product and the resulting product (Dhafr et al. 2006). It can be caused by a single source or the cumulative effect of several factors, which may arise at any stage of the extrusion process. Some defects can be found in extruded parts such as the rough surface, the extruder surging, the thickness variation, the uneven wall thickness, the diameter variation, and the centering problem. In this work, the extruded tube quality was defined regarding the inner diameter (ID) and the outer diameter (OD), as shown in Fig. 3. Although other characteristics could affect the quality of the product (e.g., length of the tube, wall thickness, or color uniformity), the only functional requirements for this application correspond to the inner and outer diameters because these are the critical characteristics that were stated by the customer.

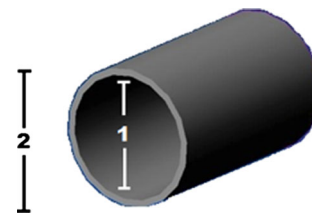


Fig. 3 Extruded tube. Quality indices: inner and outer diameters

To guarantee the quality of the manufactured tube (i.e., the inner and outer diameters have to be within design specifications set by the customer), every process parameter must be identified, controlled, and monitored throughout the extrusion process. For example, in the extruder zone, there are several input parameters that might yield significant deviations in the characteristics of the product: the base hopper temperature (BHT), the zone temperature (ZT), the die temperature (DT), the melting temperature (MT), and the revolutions per minute of the screw (SRPM). In the case of the pulling stage, the set of parameters are the tank temperature (TT), the vacuum pressure (VP), and the tension of the pulling mechanism (TPM).

In total, there are 15 process parameters that may produce deviations in the functional requirements of the extruded tube: four zone temperatures, four die temperatures, the melting temperature, the revolutions per minute of the screw, the base hopper temperature, two tank temperatures, the vacuum pressure, and the tension of the pulling mechanism. As a result, each sample will be described by these 15 input parameters and the two output variables mentioned in the previous paragraph (ID and OD).

Table 1 reports the main characteristics of the database used in the empirical analysis: the attribute number, the attribute description and some statistics, such as the minimum and maximum values of the attribute, the mean and the standard deviation.

It is important to point out that the different input parameters were measured and recorded using specific sensors during the extrusion and pulling processes, and an operator collected the data at a fixed time. Analogously, the input and output diameters of tubes were measured manually with a vernier caliper by the operator. At each shift, these tasks were carried out three times, thus obtaining a data set as the one shown in the example of Table 2.

Regression models

In this section, we briefly introduce the regression methods that will be further applied to product quality prediction for the tubing extrusion process just described.

Table 1 Characteristics of the tube-extrusion data set used in the experiments

No.	Description	Minimum	Maximum	Mean	Std. Dev.
<i>Extrusion process</i>					
1	Zone 1 temperature (°F)	435	448	440.68	2.39
2	Zone 2 temperature (°F)	432	445	440.53	2.45
3	Zone 3 temperature (°F)	429	463	433.69	4.09
4	Zone 4 temperature (°F)	425	440	433.48	3.05
5	Melting temperature (°F)	436	450	441.78	2.96
6	Die 1 temperature (°F)	442	454	447.99	2.88
7	Die 2 temperature (°F)	441	459	448.39	3.28
8	Die 3 temperature (°F)	445	456	448.97	2.31
9	Die 4 temperature (°F)	444	458	449.10	2.46
10	Revolutions per minute of the screw (rpm)	52	75	62.00	5.94
11	Base Hopper temperature (°F)	46	80	58.50	7.84
<i>Pulling process</i>					
1	Tank 1 temperature (°F)	49	71	64.40	5.70
2	Tank 2 temperature (°F)	0.2	10.1	1.01	6.23
3	Vacuum pressure (inHg)	59	65.6	61.42	1.42
4	Tension of the pulling mechanism (ft/min)	40	56	50.34	2.13
<i>Output variables</i>					
1	Inner diameter (in)	0.227	0.233	0.230	0.001
2	Outer diameter (in)	0.228	0.331	0.328	0.014

Table 2 An example of collected data for the tubing extrusion process

Time	ZT ₁	ZT ₂	ZT ₃	ZT ₄	MT	DT ₁	DT ₂	DT ₃	DT ₄	SRPM	BHT	TT ₁	TT ₂	VP	TPM	ID	OD
1	438	437	431	434	442	448	451	452	455	64	71	62	1.58	63.4	47	0.228	0.327
2	441	439	452	437	450	452	457	453	455	68	78	56	3.05	61.7	51	0.230	0.311
:																	
<i>t</i>	440	443	461	429	444	445	452	448	448	70	74	60	1.83	64.0	52	0.229	0.330

Let $T = \{(\mathbf{x}_1, a_1), \dots, (\mathbf{x}_n, a_n)\} \in (\mathbf{x} \times a)^n$ be a data set of n independent and identically distributed (i.i.d.) random pairs (\mathbf{x}_i, a_i) , where $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iD}]$ represents an instance in a D -dimensional feature space and a_i denotes the continuous target value associated to it. The aim of regression is to learn a function $f: \mathbf{y} \rightarrow a$ to predict the value a for a new sample $\mathbf{y} = [y_1, y_2, \dots, y_D]$.

Nearest neighbor regression

One of the most popular and successful supervised learning methods corresponds to the nearest neighbor (NN) rule due to its algorithmic simplicity and high prediction performance. This non-parametric technique works under the assumption that new samples share similar properties with the set of stored instances and therefore, it predicts the output of a new sample based on its closest neighbor.

The concept of the NN rule can be generalized for regression because the nearest neighbor method assigns a new sample \mathbf{y} the same target value as the closest instance in T , according to a particular dissimilarity measure (generally, the Euclidean distance). An extension of this procedure is the k -NN decision rule, in which the algorithm retrieves the k closest instances in T . When $k = 1$, the target value assigned to the input sample is the target value indicated by its nearest neighbor. For $k > 1$, the k -NN regression model (k -NNR) estimates the target value $f(\mathbf{y})$ of a new input sample \mathbf{y} by averaging the target values of its k nearest neighbors (Biau et al. 2012; Guyader and Hengartner 2013; Kramer 2011; Lee et al. 2014):

$$f(\mathbf{y}) = \frac{1}{k} \sum_{i=1}^k a_i \quad (1)$$

where a_i denotes the target value of the i -th nearest neighbor.

Distance-weighted k-NN regression

When the basic k-NN algorithm estimates the target value for the new sample, it ignores some relevant information that each of the k nearest neighbors might provide regarding their distance (Batista and Silva 2009). To overcome this shortcoming, Dudani (1976) proposed a weighting function, which weights more heavily closer neighbors than distant neighbors, depending on their corresponding distances to the new sample.

In general, a weighting function has to work based upon the premise that the weights should decrease with increasing sample-to-neighbor distance (Dudani 1976). Let $\mathbf{x}_i (i = 1, \dots, k)$ be the closest instances to an input sample \mathbf{y} , and let $d_i = d(\mathbf{x}_i, \mathbf{y})$ be the distance between \mathbf{x}_i and \mathbf{y} . A common weighting technique computes w_i for the i -th nearest neighbor as the inverse of its distance (Dudani 1976):

$$w_i = \frac{1}{d_i} \quad d_i \neq 0 \tag{2}$$

Another possible weighting function (Batista and Silva 2009) can be defined as

$$w_i = 1 - d_i \tag{3}$$

Once the weights w_i have been computed, the distance-weighted k-NN approach for regression (k-NNRw) estimates the target value as follows (Hall et al. 2009):

$$f(\mathbf{y}) = \frac{\sum_{i=1}^k w_i a_i}{\sum_{i=1}^k w_i} \tag{4}$$

Linear regression

Multiple linear regression (LR) attempts to model the relationship between two or more independent variables (in this case, the input attributes reported in Table 1) and an output or response variable by fitting a linear equation to the observed data (Draper and Smith 1998). Every value of the independent variable is associated with a value of the response variable. The general form of the multiple linear regression equation can be written as follows:

$$f(\mathbf{y}) = \alpha + \sum_{j=1}^D \beta_j y_j + \epsilon \tag{5}$$

where α is a constant (the point where the regression line intercepts the Y -axis), β_j are the regression coefficients on the independent variables y_j , and ϵ is the residual or fitted error.

The regression coefficients β_i are estimated by curve fitting based on the least square method with the aim of minimizing the fitted error (the difference between the observed and estimated values). Equation 5 indicates how the average response of the output variable changes with the independent variables. Thus the LR model can be used to predict the target value a from new observed values of \mathbf{y} .

Support vector regression

The foundations of support vector machines are well-known for both classification and regression problems. Smola and Scholkopf (2004) published an excellent tutorial on support vector machines for regression (SVR). The objective of the SVR model is to define a linear regression function to map the input data to a high-dimensional feature space, in which input data can be separated easier than in the original input space (Chou et al. 2017; Ma et al. 2003),

$$f(\mathbf{x}) = \mathbf{W}^T \Phi(\mathbf{x}) + b \tag{6}$$

where \mathbf{W} is a weight vector, $\Phi(\mathbf{x})$ maps the input sample \mathbf{x} to the high-dimensional feature space, and b is a bias term.

The \mathbf{W} and b can be obtained by solving an optimization problem (Ma et al. 2003):

$$\begin{aligned} & \underset{\mathbf{W}, b}{\text{minimize}} && \frac{1}{2} \mathbf{W}^T \mathbf{W} + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ & \text{subject to} && a_i - (\mathbf{W}^T \Phi(\mathbf{x}) + b) \leq (\xi_i + \xi_i^*) \\ & && (\mathbf{W}^T \Phi(\mathbf{x}) + b) - a_i \leq (\xi_i + \xi_i^*) \\ & && \xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, n \end{aligned} \tag{7}$$

where C is a regularization parameter, ξ_i and ξ_i^* are non-negative slack variables to penalize for errors that are greater than ϵ in magnitude

By introducing the Lagrange multipliers α, α^* , and a kernel function K , the model form in the dual space can be written as:

$$f(\mathbf{x}) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \tag{8}$$

The use of a kernel function allows to deal with feature spaces of arbitrary dimensionality without having to compute the mapping function $\Phi(\mathbf{x})$ explicitly (Yang and Shieh 2010). The kernels most commonly used are linear, polynomial, sigmoid, and radial basis functions.

Experimental set-up

As already stated, this study aims to evaluate the performance of some regression models for product quality prediction in

the tubing extrusion process of a manufacturing plant. Thus we conducted a pool of experiments on a data set with 260 samples that were collected using the procedure described in “Description of the tubing extrusion process” section. It has to be remarked that all input attribute values (process parameters) were normalized to the range [0, 1].

We focused our study on the simple k-NNR (no weighting), two weighted versions of k-NNR using Eqs. 2 and 3, in the sequel called k-NNRw1 and k-NNRw2 respectively, the LR model, the SVR technique with three different kernels and the multi-layer perceptron (MLP) neural network. The kernels used in the SVR model were a linear function (SVR-1), a polynomial function of degree 2 (SVR-2) and a radial basis function (SVR-RBF). For the regression algorithms based on the k-NN rule, twenty-five odd values of k (1, 3, . . . , 29) were tested. All regression models were taken from the WEKA toolkit (Hall et al. 2009).

Following the standard strategy used to evaluate the performance of regression models when databases are small- or medium-sized, the 10-fold cross-validation method was adopted (Buza et al. 2015; Hall et al. 2009; Hu et al. 2014). The original data set was randomly divided into ten parts of size $n/10$ (where n denotes the total number of samples in the data set); for each fold, nine blocks were gathered as the training set for learning the model, and the remaining fold was used as an independent test set. Additionally, with the aim of increasing the statistical significance of the experimental scores, ten repetitions were run for each trial and the results from predicting the output of test samples were averaged across the 100 runs.

Evaluation criteria

In the framework of regression, the purpose of most performance evaluation metrics is to estimate how much the predictions (p_1, p_2, \dots, p_n) deviate from the actual target values (a_1, a_2, \dots, a_n). These metrics are minimized when the predicted value for each test sample agrees with their true value (Caruana and Niculescu-Mizil 2004). Two of the most popular performance measures that have frequently

been employed to assess the model performance in regression problems are the root mean square error (RMSE),

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - a_i)^2} \quad (9)$$

and the mean absolute error (MAE),

$$MAE = \frac{1}{n} \sum_{i=1}^n |p_i - a_i| \quad (10)$$

Both these metrics show how far away the predicted values p_i are from the target values a_i by averaging the magnitude of individual errors without taking care of their sign.

Results and discussion

Since the quality product was predicted here using the inner and outer diameters of the extruded tubes, the experiments and the subsequent analysis of results were performed according to these two physical quality indices. Hence, for each database (i.e., the outer and inner diameter databases), we compared the average of the two performance measures (RMSE and MAE) achieved by each regression method.

Outer diameter database

Table 3 reports the average results in terms of RMSE and MAE across the 100 runs for each regression technique. In the case of the k-NN methods, the values correspond to those of the best k . Based on the root mean square error, one can observe that the three k-NN algorithms, the linear support vector (SVR-1) and the SVR-RBF obtained the lowest error rates (very close to 0) when predicting the outer diameter. In addition, the same behavior can be viewed in terms of MAE.

If we consider that the output values of the outer diameter are in a range from 0.228 to 0.331, then all these regression models appear to be suitable for predicting the quality

Table 3 Average results in terms of RMSE and MAE (with standard deviations) on the outer diameter database

	RMSE	MAE
k-NNR	0.0096912371 ± 0.0075398296	0.0030848789 ± 0.0021710948
k-NNRw1	0.0096563665 ± 0.0072890820	0.0030848789 ± 0.0021710948
k-NNRw2	0.0096912371 ± 0.0075398296	0.0030848789 ± 0.0021710948
LR	0.0138590570 ± 0.0194352190	0.0062838398 ± 0.0045738531
SVR-1	0.0099222666 ± 0.0101735078	0.0029027012 ± 0.0024793246
SVR-2	0.2603216450 ± 0.9783462937	0.0523638108 ± 0.1920644156
SVR-RBF	0.0095693860 ± 0.0103128854	0.0027710505 ± 0.0024827469
MLP	0.0130784745 ± 0.0076458498	0.0066302173 ± 0.0053236816

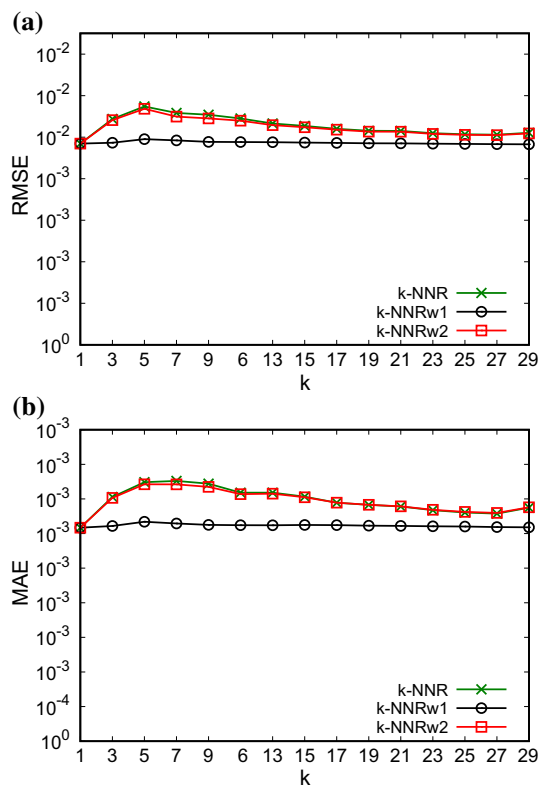


Fig. 4 Outer diameter database: performance regression measures with *k*-NN regression models when varying *k*. (a) Root Mean Square Error. (b) Mean Absolute Error

of extruded tubes. However, the small differences in both RMSE and MAE results do not allow us to draw significant conclusions about which method is the best numerical prediction technique. In fact, even the LR and MLP models could be applied to tackle this problem because their errors were also close enough to 0.

Figure 4 shows the performance measures for the three versions of *k*-NN regression when varying the value of *k* from 1 to 29. The graphical results suggest that when *k* increases, the *k*-NNRw1 shows a steady behavior along the *X*-axis. In the case of *k*-NNR and *k*-NNRw2 models, their error rates

increase along with the value of the parameter *k*. In summary, it appears that *k*-NNRw1 performed the best with all values of *k* > 1, demonstrating the benefits of applying this technique to predict the quality of extruded tubes concerning their outer diameter.

Inner diameter database

As in the previous section, we analyzed the behavior of the regression models to predict the inner diameter of an extruded tube. Table 4 shows the results of RMSE and MAE averaged across the 100 runs for each technique. Results are conceptually similar to those of the outer diameter database: (i) the methods based on *k*-NN, the SVR-1 and the SVR-RBF yielded very low error values (≈ 0); (ii) here MLP also appears to be among the best performing algorithms; and (iii) except the SVR-2 method, differences in the results of the regression models seem not to be significant.

Figure 5 depicts the performance results for all versions of the *k*-NN regression models as a function of *k*. One can observe that *k*-NNRw1 achieved very similar performance results regardless of the *k* value. In the case of plain *k*-NNR and *k*-NNRw2, when *k* increases, the error rates decrease slightly. These results suggest that the three *k*-NN regression models are suitable for predicting the inner diameter of extruded tubes, although the *k*-NNR and *k*-NNRw2 approaches seem to be the best techniques.

Conclusions and future work

The present paper has focused on predicting two quality indices in a tubing extrusion process. A thoroughly experimental study has been carried out on a real-life data set provided by an extrusion tube manufacturing plant located in Ciudad Juárez (Chihuahua, Mexico). More specifically, three *k*-NN regression methods (the straightforward algorithm and two distance-weighted approaches), the linear regression model, three SVR configurations (SVR-1,

Table 4 Average results in terms of RMSE and MAE (with standard deviations) on the inner diameter database

	RMSE	MAE
<i>k</i> -NNR	0.0010767157 ± 0.0001233918	0.0008922620 ± 0.0001050170
<i>k</i> -NNRw1	0.0012095069 ± 0.0001328387	0.0010122028 ± 0.0001120801
<i>k</i> -NNRw2	0.0010783177 ± 0.0001229690	0.0008940031 ± 0.0001048994
LR	0.0011882847 ± 0.0009202389	0.0009338954 ± 0.0002309976
SVR-1	0.0013502332 ± 0.0011949720	0.0009922225 ± 0.0002923969
SVR-2	0.1471915327 ± 0.5658384933	0.0296721074 ± 0.1110184843
SVR-RBF	0.0010781128 ± 0.0001270223	0.0008608249 ± 0.0001151820
MLP	0.0013352608 ± 0.0002263802	0.0010851685 ± 0.0001695045

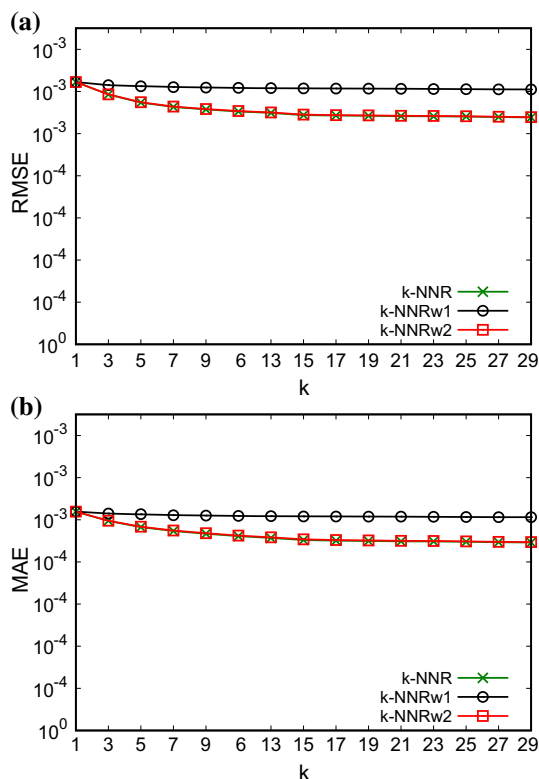


Fig. 5 Inner diameter database: Performance regression measures with k -NN regression models when varying k . (a) Root Mean Square Error. (b) Mean Absolute Error

SVR-2, and SVR-RBF), and a multi-layer perceptron have been used to predict the inner and outer diameters of an extruded tube based on the evaluation of 15 process parameters.

Experimental results suggest that distance-weighted k -NN regression models along with the linear and the RBF-based support vector regression methods were the most effective techniques for the prediction of extruded tube quality, achieving RMSE and MAE rates close to 0. From our analysis when varying the k values, we found out that when k increases, the performance regression rates are (almost) stable.

Future research will be mainly addressed to incorporate a feature selection phase to remove any attribute that might be considered noisy or irrelevant. Another avenue for further investigation concentrates on developing some regression algorithms based on the surrounding neighborhood concept. Finally, we are also interested in analyzing the behavior of ensembles of regression models.

Acknowledgements The authors would like to acknowledge the financial support from the Spanish Ministry of Economy, Industry and Competitiveness [TIN2013-46522-P], and the Generalitat Valenciana [PROMETEOII/2014/062].

References

- Adly, F., Alhussein, O., Yoo, P. D., Al-Hammadi, Y., Taha, K., Muhaidat, S., et al. (2015). Simplified subspace regression network for identification of defect patterns in semiconductor wafer maps. *IEEE Transactions on Industrial Informatics*, 11(6), 1267–1276.
- Batista, G. E. A. P. A., & Silva, D. F. (2009). How k -nearest neighbor parameters affect its performance. In *Argentine symposium on artificial intelligence*, Mar de Plata, Argentina (pp. 1–12).
- Biau, G., Devroye, L., Dujmović, V., & Krzyzak, A. (2012). An affine invariant k -nearest neighbor regression estimate. *Journal of Multivariate Analysis*, 112, 24–34.
- Buza, K., Nanopoulos, A., & Nagy, G. (2015). Nearest neighbor regression in the presence of bad hubs. *Knowledge-Based Systems*, 86, 250–260.
- Carrano, E. G., Coelho, D. G., Gaspar-Cunha, A., Wanner, E. F., & Takahashi, R. H. (2015). Feedback-control operators for improved Pareto-set description: Application to a polymer extrusion process. *Engineering Applications of Artificial Intelligence*, 38, 147–167.
- Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining*, New York, NY (pp. 69–78).
- Charaniya, S., Le, H., Rangwala, H., Mills, K., Johnson, K., Karypis, G., et al. (2010). Mining manufacturing data for discovery of high productivity process characteristics. *Journal of Biotechnology*, 147(3–4), 186–197.
- Chevanan, N., Muthukumarappan, K., & Rosentrater, K. A. (2007). Neural network and regression modeling of extrusion processing parameters and properties of extrudates containing DDGS. *Transactions of the American Society of Agricultural and Biological Engineers*, 50(5), 1765–1778.
- Chien, C. F., Wang, W. C., & Cheng, J. C. (2007). Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications*, 33(1), 192–198.
- Chondronasios, A., Popov, I., & Jordanov, I. (2016). Feature selection for surface defect classification of extruded aluminum profiles. *The International Journal of Advanced Manufacturing Technology*, 83(1), 33–41.
- Chou, J. S., Ngo, N. T., & Chong, W. K. (2017). The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate. *Engineering Applications of Artificial Intelligence*, 65, 471–483.
- Choudhary, A. K., Harding, J. A., & Tiwari, M. K. (2008). Data mining in manufacturing: A review based on the kind of knowledge. *Journal of Intelligent Manufacturing*, 20(5), 501–521.
- Dhafr, N., Ahmad, M., Burgess, B., & Canagassabady, S. (2006). Improvement of quality performance in manufacturing organizations by minimization of production defects. *Robotics and Computer-Integrated Manufacturing*, 22(5–6), 536–542.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. Hoboken, NJ: Wiley.
- Dudani, S. A. (1976). The distance-weighted k -nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(4), 325–327.
- Erzurumlu, T., & Oktem, H. (2007). Comparison of response surface model with neural network in determining the surface quality of moulded parts. *Materials & Design*, 28(2), 459–465.
- Ghorai, S., Mukherjee, A., Gangadaran, M., & Dutta, P. K. (2013). Automatic defect detection on hot-rolled flat steel products. *IEEE Transactions on Instrumentation and Measurement*, 62(3), 612–621.
- González Marcos, A., Pernía Espinoza, A. V., Alba Elías, F., & García Forcada, A. (2007). A neural network-based approach for opti-

- missing rubber extrusion lines. *International Journal of Computer Integrated Manufacturing*, 20(8), 828–837.
- Guyader, A., & Hengartner, N. (2013). On the mutual nearest neighbors estimate in regression. *The Journal of Machine Learning Research*, 14, 2361–2376.
- Hall, M., Frank, E., & Holmes, G. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 10–18.
- Harding, J. A., Shahbaz, M., Srinivas, S., & Kusiak, A. (2006). Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering*, 128(4), 969–976.
- Hsiang, S. H., Lin, Y. W., & Lai, J. W. (2012). Application of fuzzy-based Taguchi method to the optimization of extrusion of magnesium alloy bicycle carriers. *Journal of Intelligent Manufacturing*, 23(3), 629–638.
- Hu, C., Jain, G., Zhang, P., Schmidt, C., Gomadam, P., & Gorka, T. (2014). Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery. *Applied Energy*, 129, 49–55.
- Jiang, H., Yan, Z., & Liu, X. (2013). Melt index prediction using optimized least squares support vector machines based on hybrid particle swarm optimization algorithm. *Neurocomputing*, 119, 469–477.
- Khan, J. G., Dalu, R. S., & Gadekar, S. S. (2014). Defects in extrusion process and their impact on product quality. *International Journal of Mechanical Engineering and Robotics Research*, 3(3), 10–18.
- Kohlert, M., & König, A. (2015). Large, high-dimensional, heterogeneous multi-sensor data analysis approach for process yield optimization in polymer film industry. *Neural Computing and Applications*, 26(3), 581–588.
- Köksal, G., Batmaz, I., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, 38(10), 13,448–13,467.
- Kramer, O. (2011). Unsupervised K-nearest neighbor regression. ArXiv e-prints [arXiv:1107.3600](https://arxiv.org/abs/1107.3600).
- Krömer, P., Snášel, V., Platoš, J., & Abraham, A. (2010). Evolving fuzzy classifier for data mining—An information retrieval approach. In *Proceedings of the 3rd international conference on computational intelligence in security for information systems*, León, Spain (pp. 25–32).
- Kusiak, A. (2006). Data mining: Manufacturing and service applications. *International Journal of Production Research*, 44(18–19), 4175–4191.
- Kusiak, A., & Kurasek, C. (2001). Data mining of printed-circuit board defects. *IEEE Transactions on Robotics and Automation*, 17(2), 191–196.
- Lee, S. K., Kang, P., & Cho, S. (2014). Probabilistic local reconstruction for k-nn regression and its application to virtual metrology in semiconductor manufacturing. *Neurocomputing*, 131, 427–439.
- Li, H. J., Qi, L. H., Han, H. M., & Guo, L. J. (2004). Neural network modeling and optimization of semi-solid extrusion for aluminum matrix composites. *Journal of Materials Processing Technology*, 151(1–3), 126–132.
- Liukkonen, M., Hiltunen, T., Havia, E., Leinonen, H., & Hiltunen, Y. (2009). Modeling of soldering quality by using artificial neural networks. *IEEE Transactions on Electronics Packaging Manufacturing*, 32(2), 89–96.
- Ma, J., Theiler, J., & Perkins, S. (2003). Accurate on-line support vector regression. *Neural Computation*, 15(11), 2683–2703.
- Meiabadi, M. S., Vafaesefat, A., & Sharifi, F. (2013). Optimization of plastic injection molding process by combination of artificial neural network and genetic algorithm. *Journal of Optimization in Industrial Engineering*, 6(13), 49–54.
- Oberg, E., Jones, F., Horton, H., Ryffel, H., & McCauley, C. (2012). *Machinery's handbook*. New York, NY: Industrial Press.
- Oke, S. A., Johnson, A. O., Charles-Owaba, O. E., Oyawale, F. A., & Popoola, I. O. (2006). A neuro-fuzzy linguistic approach in optimizing the flow rate of a plastic extruder process. *International Journal of Science & Technology*, 1(2), 115–123.
- Pratihari, D. K. (2015). Expert systems in manufacturing processes using soft computing. *The International Journal of Advanced Manufacturing Technology*, 81(5), 887–896.
- Ramana, E. V., & Reddy, P. R. (2013). Data mining based knowledge discovery for quality prediction and control of extrusion blow molding process. *The International Journal of Advanced Manufacturing Technology*, 6(2), 703–713.
- Ribeiro, B. (2005). Support vector machines for quality monitoring in a plastic injection molding process. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 35(3), 401–410.
- Sadeghi, B. H. M. (2000). A BP-neural network predictor model for plastic injection molding process. *Journal of Materials Processing Technology*, 103(3), 411–416.
- Sharma, R. S., Upadhyay, V., & Raj, K. H. (2009). Neuro-fuzzy modeling of hot extrusion process. *Indian Journal of Engineering and Materials Sciences*, 16, 86–92.
- Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
- Tan, S. C., Watada, J., Ibrahim, Z., & Khalid, M. (2015). Evolutionary fuzzy ARTMAP neural networks for classification of semiconductor defects. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5), 933–950.
- Urraca Valle, R., Sodupe Ortega, E., Antoñanzas Torres, J., Alonso García, E., Sanz García, A., & Martínez de Pisón Ascacibar, F. J. (2013). Comparative methodology of non-linear models for predicting rheological properties of rubber mixtures in industrial lines. In *Proceedings of the 17th international congress on project management and engineering*, Logroño, Spain (pp. 1346–1357).
- Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*. <https://doi.org/10.1016/j.jmsy.2018.01.003>.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann Publishers.
- Wu, C. Y., & Hsu, Y. C. (2002). Optimal shape design of an extrusion die using polynomial networks and genetic algorithms. *The International Journal of Advanced Manufacturing Technology*, 19(2), 79–87.
- Wu, D., Jennings, C., Terpenney, J., Gao, R. X., & Kumara, S. (2017). A comparative study on machine learning algorithms for smart manufacturing: Tool wear prediction using random forests. *Journal of Manufacturing Science and Engineering*, 139(7), 071,018–071,026.
- Xu, Y., Zhang, Q., Zhang, W., & Zhang, P. (2015). Optimization of injection molding process parameters to improve the mechanical performance of polymer product against impact. *The International Journal of Advanced Manufacturing Technology*, 76(9), 2199–2208.
- Yang, C. C., & Shieh, M. D. (2010). A support vector regression based prediction model of affective responses for product form design. *Computers & Industrial Engineering*, 59(4), 682–689.
- Yin, S., Li, X., Gao, H., & Kaynak, O. (2015). Data-based techniques focused on modern industry: An overview. *IEEE Transactions on Industrial Electronics*, 62(1), 657–667.

- Yu, J. C., Chen, X. X., Hung, T. R., & Thibault, F. (2004). Optimization of extrusion blow molding processes using soft computing and Taguchi's method. *Journal of Intelligent Manufacturing*, *15*(5), 625–634.
- Zhang, Z., Wang, T., & Liu, X. (2014). Melt index prediction by aggregated RBF neural networks trained with chaotic theory. *Neurocomputing*, *131*, 368–376.
- Zhao, G., Chen, H., Zhang, C., & Guan, Y. (2013). Multiobjective optimization design of porthole extrusion die using pareto-based genetic algorithm. *The International Journal of Advanced Manufacturing Technology*, *69*(5), 1547–1556.