ON-LINE SOFT SENSOR FOR PREDICTING RUBBER PROPERTIES IN A MIXTURE PROCESS BASED ON REGRESSION MODELS WITH FEATURE SELECTION

Sodupe Ortega, E.; Urraca Valle, R.; Antoñanzas Torres, J.; Alía Martínez, M. J.; Sanz García, A.; Martínez de Pisón Ascacíbar, F. J.

Universidad de La Rioja

This communication deals with the complex behavior of rubber mixture processes and the more accurate estimation of some properties of resulting rubber bands. The main issue is to develop an on-line soft sensor for estimating significant parameters related to rubber properties. The sensor would be able to avoid the continual discard of defective material, reducing its high costs associated. This can be achieved by detecting the unexpected process variations or even bad operating set points.

The system is based on a “wrapper” scheme. First, a feature selection routine (backwards selection) is use to find the optimum feature subset from mixture process attributes, which will be utilized as inputs of linear regression model.

Those attributes that better explain the dependent variables are determined in an iterative process and the most accurate solution will be finally selected. Our proposed sensor has several advantages, i.e. the use of a linear model provides wider and deeper knowledge of the industrial process and the backwards selection techniques allow us to obtain better parsimony models. Eventually, we demonstrate that the soft sensor is also able to establish the clear relations between the independent variables and rheometric parameters of rubber.

Keywords: *Soft sensor; Rubber properties; Rubber mixture process; Regression models; Feature selection*

SENSOR ON-LINE PREDICTIVO DE PROPIEDADES EN EL PROCESO DE MEZCLADO DE GOMA MEDIANTE MODELOS DE REGRESIÓN CON SELECCIÓN DE VARIABLES

Esta comunicación aborda el complejo comportamiento de los procesos de mezclado de gomas y la estimación más precisa de propiedades de las bandas de goma producidas. El objetivo es desarrollar un sensor virtual on-line que estime los parámetros significativos relacionados con las propiedades finales de la goma. El sensor sería capaz de evitar el continuo desecho de material defectuoso, reduciendo los altos costes asociados. Esto se consigue detectando variaciones no esperadas en el proceso o puntos de operación erróneos. El sistema está basado en un “wrapper”. Una selección de variables (backwards selection) es utilizada para encontrar el subconjunto de atributos óptimo de los parámetros del proceso de mezclado que serán entradas de los modelos de regresión lineal. Aquellas variables que mejor explican las variables dependientes son determinadas mediante un proceso iterativo que finaliza con la solución que genere la mayor precisión en el resultado. La ventaja de usar modelos lineales es un conocimiento más amplio y profundo del proceso industrial. También las técnicas de selección de variables permiten obtener modelos más parsimoniosos. El sensor también es capaz de establecer relaciones claras entre las variables independientes y los parámetros reométricos de la goma.

Palabras clave: *Sensor virtual; Propiedades de la goma; Proceso de mezclado de gomas; Modelos de regresión; Selección de características*

Correspondencia: enrique.sodupeo@unirioja.es

1. Introduction

Rubber extrusion of complex profiles is one of the most critical manufacturing processes in automotive components industry. The production process is mainly characterized by a high variability on its working conditions and the need of a continuous readjusting of the most relevant control parameters involved. In addition, automotive industry nowadays requires higher quality standards and companies are constantly implementing more stringent controls.

To assure quality of final products, the rheological curve of rubber has a key role in the whole production process. Despite being such an important factor, it measurement and the capacity to be directly modified are very complex issues.

Parameters that define rheological curve can only be obtained by means of laboratory tests. These tests have to carry out after rubber compound has been extruded, therefore obtained results cannot be used to detect failures during the extrusion process. These large delays in acquisition of useful data forced to analyze it afterwards, creating a problem to plant engineers.

There are two principal phases in a rubber extrusion line, mixing and extrusion. Rubber mixing phase has a key role to obtain a good quality final product, but research on this phase is not so wide than the second one. Variations occurred in this section of the process will be decisive to avoid wasting raw material, resulting in a significant decrease of production costs. For that reason, as much useful information is provided online, as more reliable and ease control of this specific part. Development of prediction models based on historical data from the mixing process is a real possibility to contribute to achieve these advantages.

Prediction models have previously been used in the rubber extrusion industry. However, rubber mixing process is characterized by high complexity and influenced by a large number of variables (Zhang, Song et al. 2012). This is the main reason because data-driven models are preferred to solve the complexity problem (Martínez-de-Pisón, Yang, Liu et al. 2009; Gao, Ji et al. 2010).

In this study, four types of linear regression models (multiple linear regression, rpart, M5Rules and cubist) are proposed instead of using non linear black-box models to achieve a better interpretability of results without losing accuracy. To this end, a better understanding of the most important variables of the process should be obtained. Implementation of a soft-sensor in the mixing process will help operator with the decision making process and will avoid the lack of information while the laboratory test results are available. This proposal have been widely used in industry (Kadlec, Gabrys et al. 2009) and several authors have demonstrated that can provide useful knowledge for controlling setting points of rubber mixers .

Two factors have direct impact on parameters from the rheological curve of rubber. On the one hand, input conditions and properties of raw materials used for the compounds; and on the other hand, setting points of the mixer that needs to be changed when a new batch of raw material is loaded.

It is well known that obtaining constant properties in raw materials is a challenging task. This directly affects the correct behavior of the mixer. Each time a new batch of raw material is loaded to the machine, the operator needs to change all the setting points of the mixer in order to readjust its proper behavior. This is done in order to get the same rheological results that the quality department of the company fixes for each rubber profile.

This is a process that entirely depends on the good eye and experience of plant operators, but this does not guaranty that the adopted decisions will lead to the expected results. In order to solve this problem and help the operator with the decision making process, we

propose the use of a data-driven regression models in a wrapper approach with a selection of the input variables. The aim of variable selection is to improve predictors performance, provide faster and more cost-effective predictors and a better understanding of the underlying process that generated the data (Guyon and Elisseeff, 2003).

These models have the opportunity to show plant operators which are the process variables that have more influence in order to achieve better final products.

2. Database

The database used to train and test the performance of the four prediction models selected was obtained from a real rubber extrusion process. It includes a total of 20 variables. A number of 15 were used as possible inputs of models and the remaining variables were the rheological curve parameters to predict. The input variables were: “dureme”, “pocome”, “predu1”, “predu2”, “prerci1”, “prerci2”, “terefi”, “terein”, “vca1”, “vca2”, “vca3”, “vexten”, “vne1”, “vne2” and “vne3”. The output variables of the rheological curve were: “ml”, “ts1”, “tc50”, “tc90” and “mh”. All predictors used belong to rubber mixing phase.

Six different rubber formulae were included in the creation of the data base with a total of 1240 samples, corresponding to six different rubber compounds. Further database explanations about composition and variables used can be found in (Marcos, Espinoza et al. 2007).

3. Methodology proposed

In industrial processes, many parameters can be measured. However, this does not mean that all these variables are necessary to train the best prediction models. This is due to the fact that some variables may not explain the issue under study. Using these irrelevant variables as inputs, model accuracy can be reduced. For that reason, a reduction of the number of inputs should be carried out before the training phase. Here is where the parsimony concept comes into play: finding models where a variable reduction technique is applied and its loss of accuracy is feasible respect the original model, will lead to better understanding of the problem. In addition, a reduction of the resources involved such as industrial process sensors and measuring times can be achieved.

Another important factor that directly affects the correct behavior of models and selection of the proper subset size is variable importance. Usually determine importance of input variables is a hard decision process and depends on the experience of plant engineers. The problem is that not always these advices are available and decisions about which inputs are included in the model must be taken.

Due to the high variability of rubber extrusion process, it is very important to acquire as much information as possible. A feature selection routine was used as a “wrapper” in order to know the optimum subset size of variables used in the linear models. Indeed, variable importance was also analyzed in order to know which variables have a greater influence in the predicted outputs. A scheme about wrapper steps can be seen in Figure 1. Four linear models were implemented in order to predict output variables of rheological curve: multiple linear model, rpart, M5Rules and cubist.

Wrapper used consists on a recursive featuring elimination (RFE) with a backwards selection routine. The RFE algorithm gives the best number of variables to use and which of these variables have more significant importance. A resampling method (bootstrap) was used inside the wrapper in order to have a greater amount of data to analyze. Wrapper structure can be described as follows:

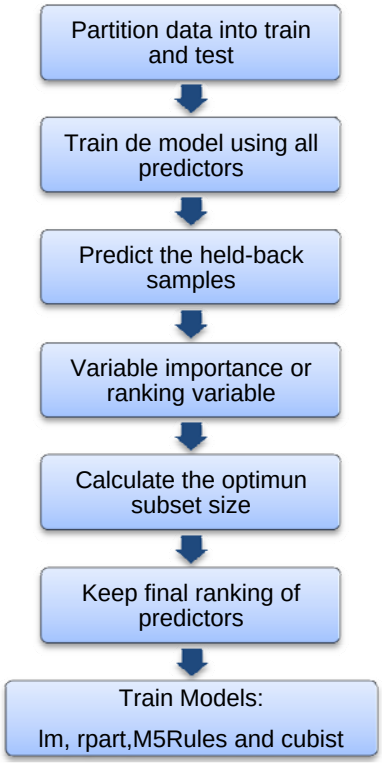
1. Partition data into train and test

2. Model is tuned with the training set of data using all inputs
3. Calculate variable importance or variable ranking
4. Backwards selection routine is applied in order to know the optimum subset size of the model. Algorithm starts training the model with all the p available inputs. For each iteration, the input with a lower score is excluded and model is trained again with (p-1) inputs. There will be as many iterations as inputs wanted be analyzed. An optional recalculation of variable rankings can be done after each predictor is excluded.
5. Best subset size is determined by some measure performance. In this study RMSE was used to adjust models performance. RMSE is computed over all models with different number of inputs. A tolerance is chosen by the client in order to determine the acceptable subset size to be picked. This tolerance determines the acceptable difference percentage between the best subset size and the reduced one. This tolerance is calculated according the Equation (1). Optimum RMSE (RMSEopt) corresponds to the minimum error obtained for an specific subset size.

$$\mathrm{RMSEtol} = 100 \cdot \frac{\mathrm{RMSE} - \mathrm{RMSEopt}}{\mathrm{RMSEopt}} \qquad (1)$$

6. Once the best subset size is determined, next step is to select the list of inputs to keep in the final model. This selection is carried out according to the previous variable ranking.

Figure 1: Basic procedure to set up wrapper scheme



Two previous steps are carried out before execution of wrapper code, normalization of the original data base and partition of data intro train and test. The first step is needed when there are different scales in the database. In the second step training dataset is used for

tuning models, but testing dataset is not used in order to check the generalization ability of models predicting new data.

Once the wrapper scheme calculates the best subset size and which inputs should be included, this information is used in order to train the four models studied. In order to be able to compare the performance of all models, same inputs should be used and the same wrapper was applied to all of them. Once the four predicted models are obtained, a denormalization of data is applied.

The performance of the models was measured using following:

$$\mathrm{RMSE} = \left\{ \frac{\sum_{k=1}^{n} [y(k) - \hat{y}(k)]^2}{n} \right\}^{1/2} \quad (2)$$

Where $y(k)$ is the target output, $\hat{y}$ is the prediction of the model and n is the total number of instances;

$$\mathrm{MAE} = \frac{\sum_{k=1}^{n} |y(k) - \hat{y}(k)|}{n} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{k=1}^{n} (y - \bar{y})^2}{\sum_{k=1}^{n} (\hat{y} - \bar{y})^2} \quad (4)$$

Where $\bar{y}$ is the mean of the observed data.

These performance measurement values were calculated in both training and testing. A repeated k-fold cross validation (CV) technique was used in order to generate a larger number of estimates, so a more reliable performance of models is obtained.

Mathematical and statistical analyses were carried out with the open source software R-Project 2.15 (http://www.r-project.org), running on a dual quadcore Opteron server with Linux SUSE 11.2. The preprocessing and post processing of data was also carried out with the same statistical software and all models were implemented using the following R-project packages: “Cubist”, “rpart” and ”RWeka”. Finally, the wrapper approach was integrated in the design process using “caret” package.

4. Overview of the models

Models implemented to predict output variables of rheological curve are described below:

- Rpart is an iterative process of splitting data intro separate sub-groups, using a two stage procedure. The algorithm recursively chooses the split that partitions the data into two parts such as to minimize the sum of the squared deviations from the mean in the separate parts. This partition process is done until a minimum size is achieved or no improvement can be made. After the complete tree is built a cross validation technique is applied in order to prune or simplify it. A further explanation of rpart can be found in (Therneau and Atkinson, 1997).
- M5rules is a method to generate rules from model trees. It is a basic separate-and-conquer strategy for learning rules. However, instead of building a single rule, a full model tree at each stage is built and taking its best branch as a rule. All instances covered by the rule are removed from the dataset. The process is applied recursively to the remaining instances and terminates when all instances are covered by one or

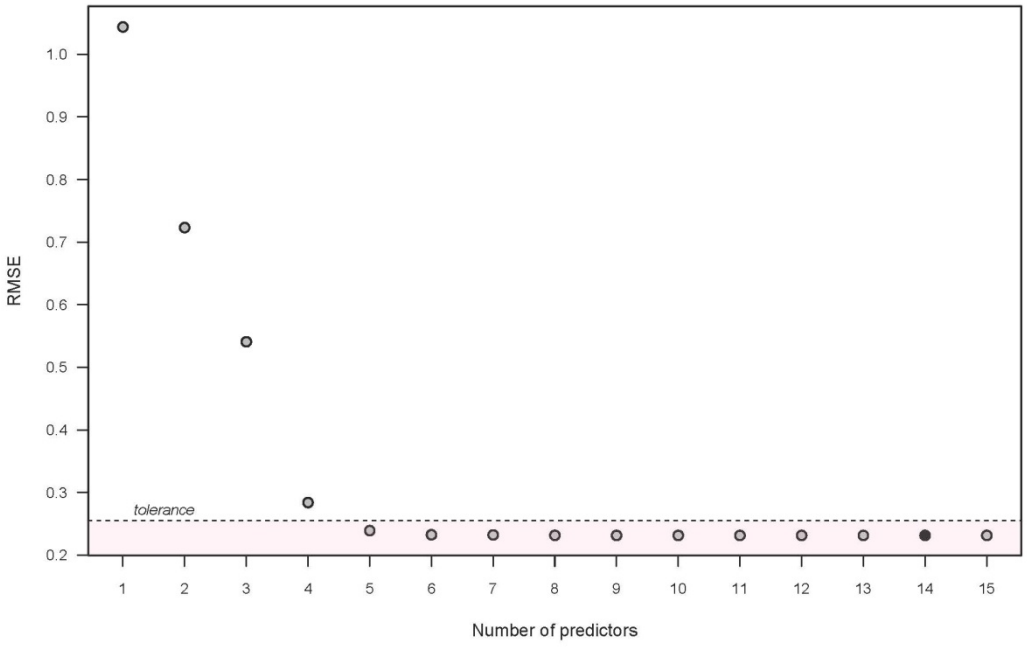
more rules. This avoids potential for over-pruning. In contrast to PART (Partial Decision Trees), which employs the same strategy for categorical prediction, M5'Rules builds full trees instead of partially explored trees (Küüksille, Selba et al. 2011).

- Cubist is a rule based model that is an extension of Quinlan's M5 model tree. A tree is grown where the terminal leaves contain linear regression models. These models are based on the inputs used in previous splits. Also, there are intermediate linear models at each step of the tree. A prediction is made using the linear regression model at the terminal node of the tree, but is \smoothed" by taking into account the prediction from the linear model in the previous node of the tree (which also occurs recursively up the tree). The tree is reduced to a set of rules, which initially are paths from the top of the tree to the bottom. Rules are eliminated via pruning and/or combined for simplification. A further explanation can be found in (Quinlan, 1992).

6. Results and discussion

Figure 2 shows a visual explanation of subset size selection and tolerance technique carried out. Tolerance used for all the experiments was 10% and calculated according Equation (1). Each point of the figure represents a model with a specific number of inputs in abscises axis and RMSE performance in ordinates axis. The point with the black background represents the model with a specific number of inputs that has a lower RMSE or optimum RMSE. This graphic does not remain constant, each new iteration tolerance value and best subset size are computed.

Figure 2: Selection of best subset size



Those models located under the tolerance line have a similar accuracy to the optimum model, with a lower use of inputs and less complexity. For instance, choosing only five inputs at least 90% of accuracy can be achieved respect optimum model (see Figure 2). A further explanation of this technique can be found in (Hastie, Tibshirani and Friedman, 2003). Those models within assumed tolerance with a lowest number of inputs were chosen to form the best subset size.

Table 1: Subset size used for each model

	Number of predictors																	
	ml				ts1				tc50				tc90				mh	
	4	5	6	7	5	6	7	8	5	6	7	8	5	6	7	8	5	6
lm	0	843	156	1	93	670	236	1	3	692	289	16	0	478	522	0	551	449
rpart	20	793	186	1	213	663	122	2	51	650	282	17	7	766	227	0	656	344
M5Rules	7	851	141	1	193	702	104	1	59	608	316	17	2	704	293	1	698	302
cubist	28	822	150	0	185	731	83	1	43	659	280	18	2	713	285	0	691	309

In Table 1 it shown the subset size used to train models for each output variable of the rheological curve. A considerable reduction of the original number of inputs is achieved. Most used subset sizes have five and six inputs. Therefore, a simplification of models can be achieved without a significant loss of accuracy.

Tables 2, 3, 4 and 5 show the performance of four models (lm, rpart, M5Rules and cubist) for each output variable are represented. Both tables provide mean values of the 100 x 10-fold cross validation technique carried out. The feature selection was not applied in Tables 2 and 3. It is clear that only test results for both tables are represented. Results obtained within the framework of this study reveal cubist is the model with the highest accuracy, followed by M5Rules and lm. As expected, models trained with the complete set of inputs have better accuracy than models which wrapper approach is applied. This tendency is reflected in the four models studied and all rheological parameters predicted. Despite there is a reduction predicting accuracy using wrapper scheme, this reduction mostly is not observed until the second decimal. In addition, this loss of accuracy is offset by a higher parsimony of models and a higher interpretability of results.

Table 2: Mean errors of variables ml, ts1 and tc50 without feature selection

	ml			ts1			tc50		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
lm	0,2312	0,1796	0,9690	1,9807	1,5593	0,9484	2,2663	1,7575	0,9513
rpart	0,2577	0,2019	0,9625	2,2052	1,6301	0,9505	2,3787	1,8113	0,9530
M5Rules	0,2287	0,1674	0,9797	2,0176	1,4571	0,9489	2,3120	1,6808	0,9460
cubist	0,2228	0,1646	0,9558	1,9597	1,4287	0,9744	2,2776	1,6359	0,9790

Table 3: Mean errors of variables tc90 and mh without feature selection

	tc90			mh		
	RMSE	MAE	R^2	RMSE	MAE	R^2
lm	1,4801	1,1563	0,9586	1,5716	1,2368	0,9653
rpart	1,5655	1,1535	0,9588	1,7302	1,3702	0,9537
M5Rules	1,4266	1,0677	0,9547	1,5596	1,1811	0,9691
cubist	1,4469	1,0646	0,9656	1,5665	1,1380	0,9809

Table 4: Mean errors of variables ml, ts1 and tc50 with feature selection

	ml			ts1			tc50		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
lm	0,2334	0,1804	0,9686	2,0933	1,6425	0,9423	2,3039	1,7751	0,9522
rpart	0,2747	0,1930	0,9602	2,0581	1,5910	0,9487	2,3534	1,7905	0,9521
M5Rules	0,2282	0,1722	0,9495	2,0775	1,5886	0,9294	2,3625	1,7715	0,9351
cubist	0,2260	0,1709	0,9622	2,0792	1,5783	0,9495	2,4338	1,7537	0,9637

Table 5: Mean errors of variables tc90 and mh with feature selection

	tc90			mh		
	RMSE	MAE	R^2	RMSE	MAE	R^2
lm	1,6380	1,2654	0,9524	1,5961	1,2499	0,9609
rpart	1,5511	1,1492	0,9565	1,7427	1,3760	0,9518
M5Rules	1,4445	1,1139	0,9487	1,5335	1,1998	0,9625
cubist	1,4821	1,1163	0,9570	1,5428	1,1965	0,9705

Wrapper approach can provide useful information as shown in Table 6, where influence of inputs for each output variable is represented. It can be observed which variable have a more direct effect explaining rheological parameters and which variables are irrelevant in order to understand its behavior. For example, variables “predu1” and “vne3” are always included as inputs for all models.

These results show that more strict control of the setting points in the mixture machine will lead to a better control of the output parameters. This is not the case with those variables that were excluded or not included in any models.

Table 6: Percentage of inputs used for each output variable

	Predictors														
Output variable	dureme	pocome	predu1	predu2	prerci1	prerci2	terefi	terein	vca1	vca2	vca3	vexten	vne1	vne2	vne3
ml	0	0	100	100	0	2	0.2	0.1	0	0	49.4	100	65.1	0	99
ts1	18.4	0.6	100	100	25.1	71.5	23.1	18.3	0	100	0	0	57.5	0	100
tc50	16.5	0	100	100	0	4.8	99.7	10	0	100	0	1.2	99.7	0	99.9
tc90	45.6	100	100	99.9	0	7.5	99.2	0	0	100	0	0	0	0	100
mh	0	100	100	3.5	14.7	10.5	0.8	0	0	0	84.5	100	32.8	0	98.1

8. Conclusions

This paper deals with the development of regression models to explain the complex behavior of rubber mixing process where physical models have been widely used. First, a wrapper scheme has been introduced in order to obtain parsimonious models and also better understanding of the underlying process of mixing phase. This dimensionality reduction did not generate a significant reduction on accuracy. Indeed, it is only observed a variation in the

number of the second decimal in performance measures. The developed wrapper provided additional useful information for rubber mixer operators and plant engineers. A higher interpretability of results and a better understanding of the most relevant setting points were achieved. All in all, this shows that using regression models in wrapper schemes is an interesting technique for modeling soft sensors. We consider them a promising technique that can be used in many industrial applications.

9. Acknowledgements

The authors are grateful for financial support provided by the University of La Rioja via grant FPI-2012 and for support provided by the Autonomous Government of La Rioja under its 3er Plan Riojano de I+D+I via project FOMENTA 2010/13.

10. References

Gao, Y. C., J. Ji, et al. (2010). Adaptive least contribution elimination kernel learning approach for rubber mixing soft-sensing modeling.

Guyon and Elisseeff (2003). “An introduction to variable and feature selection” J. Mach. Learn. Res., Vol. 3, pp. 1157-1182

Hastie, Tibshirani and Friedman (2003). “The Elements of Statistical Learning: Data Mining, Inference, and Prediction In The Elements of Statistical Learning”.

Kadlec, P., B. Gabrys, et al. (2009). "Data-driven Soft Sensors in the process industry." Computers & Chemical Engineering 33(4): 795-814.

Küüksille, E. U., R. Selba, et al. (2011). "Prediction of thermodynamic properties of refrigerants using data mining." Energy Conversion and Management 52(2): 836-848.

Marcos, A. G., A. V. P. Espinoza, et al. (2007). "A neural network-based approach for optimising rubber extrusion lines." International Journal of Computer Integrated Manufacturing 20(8): 828-837.

Martínez-de-Pisón, F. J., C. Barreto, et al. (2008). "Modelling of an elastomer profile extrusion process using support vector machines (SVM)." Journal of Materials Processing Technology 197(1–3): 161-169.

Quinlan (1992). Learning with continuous classes In Proceedings of the Australian Joint Conference on Artificial Intelligence, pp. 343-348

Therneau and Atkinson (1997). “An introduction to recursive partitioning using the RPART routines”

Yang, D., Y. Liu, et al. (2009). Online prediction of Mooney viscosity in industrial rubber mixing process via adaptive kernel learning method.

Zhang, Z., K. Song, et al. (2012). "A novel nonlinear adaptive Mooney-viscosity model based on DRPLS-GP algorithm for rubber mixing process." Chemometrics and Intelligent Laboratory Systems 112(0): 17-23.