Predictive Analytics in the Petrochemical Industry: Research Octane Number (RON) forecasting and analysis in an Industrial Catalytic Reforming Unit

Tiago Dias[a b], Rodolfo Oliveira[b], Pedro Saraiva[ac], Marco S.Reis[a]

[a] Univ Coimbra, CIEPQPF, Department of Chemical Engineering, Rua Sílvio Lima, Pólo II - Pinhal de Marrocos, 3030-790 Coimbra, Portugal

[b] Petrogal, SA, Rua Belchior Robles, 4451-852 Leça da Palmeira, Portugal

[c] Dean of NOVA IMS, Campus de Campolide, 1070-312 Lisboa, Portugal

This is the accepted author manuscript of the following article published by Elsevier:

Dias, T., Oliveira, R., Saraiva, P., & Reis, M. S. (2020). Predictive analytics in the petrochemical industry: Research Octane Number (RON) forecasting and analysis in an industrial catalytic reforming unit. Computers and Chemical Engineering, 139, [106912]. https://doi.org/10.1016/j.compchemeng.2020.106912

Predictive Analytics in the Petrochemical Industry: Research Octane Number (RON) forecasting and analysis in an Industrial Catalytic Reforming Unit

Tiago Dias[1,2], Rodolfo Oliveira[2], Pedro Saraiva[1,3], Marco S. Reis[1]

[1] *CIEPQPF, Department of Chemical Engineering, University of Coimbra, Pólo II – Rua Sílvio Lima, 3030-790 Coimbra, Portugal*

[2] *Petrogal, SA, Rua Belchior Robles, 4451-852 Leça da Palmeira, Portugal*

[3] *Dean of NOVA IMS, Campus de Campolide, 1070-312 Lisboa, Portugal*

Abstract

The Research Octane Number (RON) is a key parameter for specifying gasoline quality. It assesses the ability to resist engine knocking as the fuel burns in the combustion chamber. In this work we address the critical but complex problem of predicting RON using real process data in the context of a catalytic reforming process from a petrochemical refinery. We considered data collected from the process over an extended period of time (21 months). RON measurements are obtained offline, by laboratory analysis, with a significant delay and at much lower rates when compared to process measurements. The proposed workflow covers all the way from data collection, cleaning and pre-processing to data-driven modelling, analysis and validation for a real industrial refinery located in Portugal. The accuracy achieved with the best soft sensors open up perspectives for industrial applications and the results obtained also provide relevant information about the main RON variability sources.

Keywords: Predictive Data Analytics; Soft sensors; Research Octane Number; Catalytic Reforming; Big Data.

1. Introduction

In the Industry 4.0 and Big Data era there is an increasing interest in the exploitation of the huge amounts of industrial data that are being routinely collected and stored. In the particular case of the petrochemical industry, significant gains can be anticipated given the high leveraging of mass production for even small/moderate improvements arising from data-driven process diagnosis of problems and improvement opportunities. Industrial processes and refineries are equipped with a large diversity of sensors that record process variables, such as flow rates, temperatures, pressures, pH or conductivities, primarily for the purposes of real-time monitoring and control (Fortuna et al., 2007; Lin et al., 2007; Seborg and Thomas F. Edgar, 2011; Souza et al., 2016). Product quality variables tend to be registered less frequently, due to the fact that they require complex protocols, time and cost to extract and process samples in the laboratories, an activity which is often associated with time-consuming procedures, expensive equipment and highly trained staff. This leads to rather different acquisition rates, creating sparse data structures in the plant databases, also known as multirate data structures. These data sources can now be integrated and further explored through advanced process analytics methodologies, for developing new predictive, monitoring and diagnosis solutions.

To mitigate the aforementioned effects of the long delays and low sampling rates for the product quality variables (such as poor process control, reduced process efficiency, higher variability of product quality and off-spec levels), there has been an increase of interest to apply process analytics methods for developing inferential models that are able to provide real-time estimates of their values. They are known as soft sensors (Chéruy, 1997; Geladi and Esbensen, 1991) and require access to several information sources, including data and process knowledge. In their development, a coherent analytical workflow should be set in place according to which an appropriate model structure for the soft sensor is selected taking into account the phenomena to be addressed and the structure of data that is available to estimate and implement the models, as the goals to be achieved by using them.

These models can be derived from first principles and/or from process/product quality data, thus following mostly mechanistic or data-driven approaches, respectively. The model-based methodology chosen is dependent on the availability of knowledge about the process phenomena and all related parameters, which can be quite scarce in many industrial environments. Data-driven methods take advantage of the information extracted from data in order to develop predictive models. They critically depend on the existence of potentially informative data collectors, a requirement that has been improving over time and even more recently, with the emergence of new smart and remote sensing technologies connected to the industry 4.0 paradigm. Examples of data-driven techniques that can be applied, include, among others, the following ones: principal component regression (PCR) (Jackson, 1991, 1980; Jolliffe, 2002; Krzanowski, 1988; Martens and Naes, 1989; Wold et al., 1987) partial least-squares (PLS) (De Jong, 1993; Facco et al., 2009; Geladi, 1988; Geladi and Kowalski, 1986; Kaneko et al., 2009; Lindgren et al., 1993; Naes et al., 2004; Wold et al., 2001), artificial neural networks (ANN) (Anderson, 1997; McAvoy et al., 1989; Rumelhart et al., 1986; Venkatasubramanian et al., 1990; Willis et al., 1991), neuro-fuzzy systems (NFS) (Jang et al., 2005; Jianxu and Huihe, 2003; Wang and Mendel, 1992; Zeng and Singh, 1995), regression trees (Cao et al., 2010; Strobl et al., 2009), support vector machines and several machine learning algorithms (Fu et al., 2008). Quantitative structure–activity relationships (QSPR) constitute another type of data-driven approaches, where molecular structural parameters are correlated with macroscopic properties (e.g., molar refractivity, critical pressure, temperature and volume) (Dehmer et al., 2012; Hansch et al., 1996).

Many good examples have been reported over the years on how such data-driven Process Systems Engineering approaches can be useful in different industrial applications. However, when one has to deal with real plant data collected from the Chemical Process Industries, a number of additional important challenges need to be faced right from the early stages of data analysis. Among them, one finds usually included the following: existence of missing data, multiple sampling rates (multirate), presence of outliers and noise, as well as strong correlations and multicollinearity. Multirate data often arise when considering simultaneously process data and measurements for product quality variables (Lu et al., 2004; Wu and Luo, 2010). Outliers are values that significantly deviate from the usual ranges and can be originated by communication errors, sensor malfunction or process upsets (Chiang et al., 2003). Missing data occurs when there is no value stored for a variable in a specific sampling instant due to a sensor malfunction (process variable) or problems in the laboratory (product quality variable). All of these issues will be addressed in this article, as described in Section 2.3, dedicated to the Data Analysis Workflow, and their relevance should never be underestimated when we are dealing with real plant data.

In the specific context of refining processes, soft sensors are of critical importance. They have been applied in a range of applications, including: estimation of kerosene drying point through the use of bootstrap aggregated PLS (Zhou et al., 2012); prediction of the quality of vacuum gas oil from catalytic hydrocracking using a probabilistic framework (Lababidi

et al., 2011); prediction of NO_X emissions from a combustion unit in industrial boilers using a dynamic artificial neural network (Shakil et al., 2009); estimation of RON for a catalytic reformer unit through recursive PLS (Qin, 1998); prediction of the gasoline absorbing rate in a FCC unit by least squares support vector machines (Feng et al., 2003); estimation of the light diesel freezing point through a neuro-fuzzy system based on rough set theory and genetic algorithms (Luo and Shao, 2006); prediction of the quality of crude oil distillation using an approach based on the evolving Takagi-Sugeno method (Macias et al., 2006); fault diagnosis in a FCC reactor using a neural network (Yang et al., 2000); and the estimation of the distillate and bottom compositions for a distillation column through PLS and dynamic PLS (Kano et al., 2001).

Gasoline is one of the most consumed crude oil derivatives. The Research Octane Number (RON) is a product quality parameter for gasoline, assessing its ability to resist engine knocking. Knock occurs when the mixture of fuel and air explodes in the cylinder, instead of burning in a controlled way. If the octane number does not lie within a specific range, the engine does not work properly, thus causing a significant power loss and increased pollution emissions. This property is assessed by running a sample in a motor under standard and well controlled conditions, which take considerable preparation and execution times.

This paper addresses the critical challenge of predicting RON using just readily available process data collected from a real plant. With such a tool available, plant operators can anticipate the necessary corrective and troubleshooting actions to be taken, instead of waiting hours/days for the laboratorial analysis outcomes, with the inherent potential detrimental consequences. The ultimate benefits of the existence of the soft sensor are namely the following ones: reduction of energy consumption in the reforming unit (operators do not need to take conservative actions of increasing temperature to maintain RON levels at the target because of the lack of real-time information about their values), increase of the catalyst lifetime (that degrades faster at higher temperatures) and reduction of CO_2 emissions (as temperatures are better managed, the fuel burned in the furnaces is just the necessary to achieve the desired RON).

Hardware sensors, like online analysers, have been used to measure RON. However, they are expensive, require proper calibration and, perhaps most importantly, there are non-trivial maintenance issues in their operation. As an alternative, expedite analytical methods can be run in the laboratories, such as Near-Infrared (NIR), Raman and Nuclear Magnetic Resonance (NMR) spectroscopy, which also require expensive equipment, specific data analysis procedures and still lead to slower acquisition rates (even though better than the reference motor method). Nonetheless, among these techniques, NIR analysis does have some advantages, since it is a well-known analytical method for the study of petroleum products, as well as being fast, non-destructive, also requiring little sample preparation, and being able to capture multi-component information. It also requires the use of chemometric methods, such as PLS, to predict the target quality parameters (Amat-Tosello et al., 2009; Balabin et al., 2007; Bao and Dai, 2009; He et al., 2014; Kardamakis and Pasadakis, 2010; Lee et al., 2013; Mendes et al., 2012; Voigt et al., 2019).

In order to overcome the aforementioned limitations of hardware sensors, several “software sensors” or soft sensors were developed to predict RON in real-time, speeding up operational decisions, with the corresponding benefits referred above. Most published works concerning the prediction of RON rely on correlating it with other properties that are measured in the laboratory, like feed composition, distillation curves and density. Therefore, they do not solve the fundamental problem of reducing the analysis time delay. Thus, this work focus on developing a model to predict RON based only on readily

available process variables, which are collected online and at high sampling rates, as this was also found to be a key priority for further process improvement to be achieved at the refinery we have been collaborating with.

In the specific context of RON prediction, artificial intelligence frameworks such as neural networks (Moghadassi et al., 2016; Sadighi et al., 2015), fuzzy logic (Vezvaei et al., 2011) and hybrid methods (Ahmad et al., 2018) are the most often used methodologies. They rely on a mixture of process variables and measurements collected from the laboratory. Looking at the high number of regression methods currently available, just a few of them were tested on this topic of interest. Therefore, it is an additional goal of this work to apply and comparatively assess representatives from different classes of prediction frameworks regarding their capability to predict RON based only on process variables, as well as to provide information on the most relevant sources of RON variability. Future possible application scenarios for this soft sensor will also include monitoring and control of catalyst deactivation.

The rest of this article is structured as follows: Section 2 describes the data set collected from the refinery, as well as the overall data analysis workflow that was chosen; Section 3 reports in detail the different regression methods that were tested and compared; Section 4 presents the main results obtained for the prediction of RON, which are further discussed in Section 5. Finally, the main conclusions are summarized in Section 6.

2. Data Collection and Analysis Workflow

In this section, we present a brief description of the industrial unit from which data were collected. The subsequent data analysis workflow is also clarified, including the steps of data collection, cleaning, pre-processing, model development, comparison and final validation.

2.1 Catalytic Reforming Unit

Gasoline is one of the main products from oil industry. It is produced through the mixture of process streams in the distillation range of naphtha. The streams usually involved in gasoline production are straight run naphtha, cracked naphtha, coke naphtha (i.e. after hydrotreatment) and reformate naphtha. Reformate is produced in a processing unit called catalytic reforming (Fig. 1). The main goal of the catalytic reforming unit is to produce a stream with high aromatic hydrocarbons content that can be fed into the gasoline pool or to produce other petrochemical intermediates, such as benzene, toluene and xylenes, according to market demands. Due to the high content of aromatics compounds, reformate can raise significantly the octane number in the gasoline.

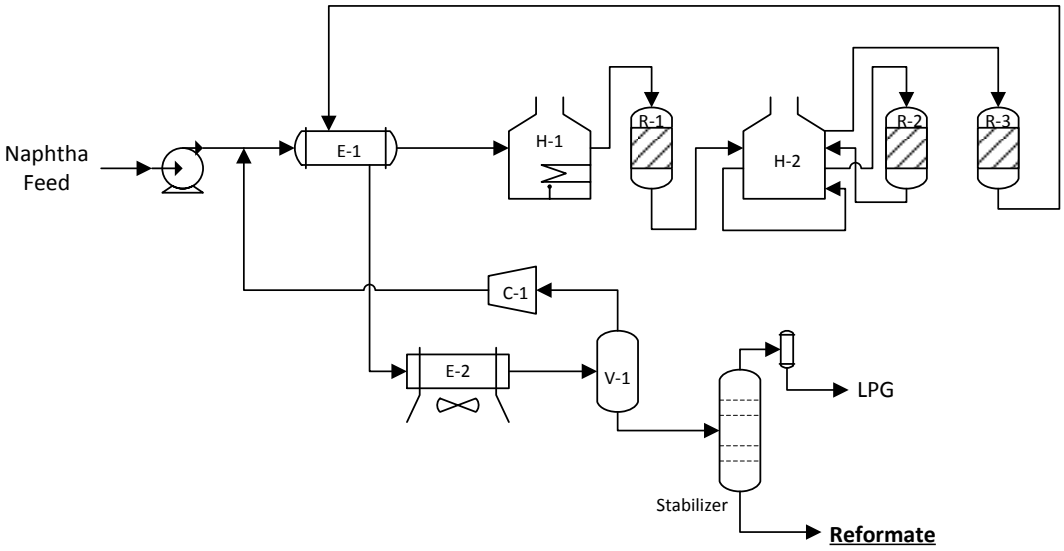


Fig. 1. Catalytic reforming process.

2.2 Data Collection

Data were collected and stored into a Microsoft® Excel worksheet using a Visual Basic for Applications (VBA) code. The analysis was carried out in the MATLAB® environment (The Mathworks, Inc.).

The data set used in this work concerns an extended operation period of 1 year and 9 months. It contains information about the catalytic reforming process coming from process variables, as well as product quality data concerning the response variable, RON. The information collected from the process corresponds essentially to feed and hydrogen flow rates, average temperatures and pressures in the reactors, temperatures and pressures in the distillation columns. The process variables have a sampling rate of one minute, while RON (the response variable) has a sampling rate on the order of days. Therefore, our data set contains a multirate structure, with missing data and outliers also being present.

RON measurements were obtained using the most accurate methodology available in the plant, the Combustion Fuel Research (CFR) engine, also known as the knock engine. The CFR engine measures octane by combusting the fuel and physically finding out the knock that occurs, as prescribed by ASTM Method D2699.

Overall, 41 process variables were employed in this study for understanding and predicting the RON of gasoline.

2.3 Data Analysis Workflow

The industrial data set described in the previous section was subject to a sequence of processing stages, starting with Data Cleaning and Pre-Processing. Data Cleaning concerns the detection and removal of outliers, plant shutdowns and records associated with sensor malfunctions. In the Pre-Processing stage, we defined the time granularity level of the analysis (or resolution). Then, missing data imputation techniques were applied to fill in the empty slots of data at the selected resolution. After these two stages, we obtained a final data set considered to be suitable for model building purposes.

The general data analytics workflow adopted in this work, from data collection to model development and validation, is presented in Fig. 2 and will be briefly described in the following subsections.

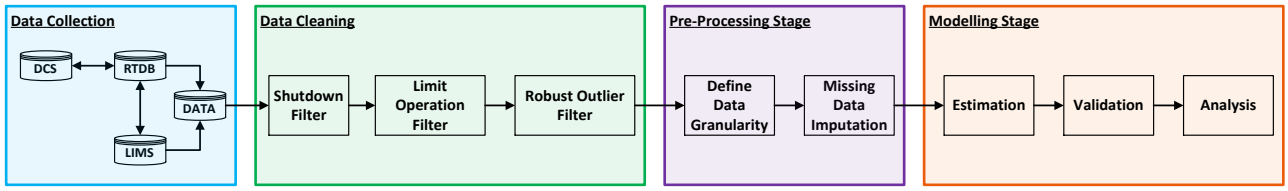


Fig. 2. Data analysis workflow.

2.3.1. Data Cleaning

Outliers correspond to values that diverge, in a more or less obvious way, from the expected typical range (Chiang et al., 2003). Outliers may be caused, for example, by excessive sensor noise, sensor degradation, errors in communication and/or process disturbances. Outliers can be classified as obvious or non-obvious (Kadlec et al., 2009; Qin, 1997). Obvious outliers are values that are inconsistent with the technical parameters of the sensor. In order to detect this type of outliers, prior information about the sensor operating ranges is needed. Non-obvious outliers are more difficult to identify because they respect the technological and physical limitations, but nonetheless are outside the typical values of the variables (not reflecting their normal behaviour). The detection of outliers is also a crucial part on data cleaning, since they can have a negative impact on the performance of the models developed or even bias the conclusions reached. This activity is usually implemented by running sets of rules over the raw data. However, the blind implementation of these rules is not advisable, and one should always validate the outcome to check if there are outliers that were not detected or observations that were incorrectly classified as outliers (Chiang et al., 2003). Observations that present more subtle deviations from the overall patterns, such as distinct correlation patterns, should be detected and analysed during model development, using multivariate methodologies such as the Mahalanobis distances, Principal Component Analysis or the diagnostic tools associated to the modelling frameworks (leverages, influence metrics such as the Cook's distance, etc.).

Typical rules to identify outliers are based on statistical descriptors of historic data. One of the simplest outlier detection rules is the 3σ algorithm (Pearson, 2002). This method assumes a Gaussian distribution for the data and classifies as an outlier any value that lies outside the range of $\hat{\mu}_x \pm 3\hat{\sigma}_x$, where $\hat{\mu}_x$ and $\hat{\sigma}_x$ are the estimated mean and standard deviation of variable x, respectively. However, this method is sometimes inefficient precisely due to the presence of outliers that tend to inflate the standard deviation, causing some of them to fall inside the acceptable range. The Hampel identifier approach is more robust, because it replaces the mean with the median and the standard deviation with median absolute deviation from the median (MAD) (Fortuna et al., 2007; Scheffer, 2002; Souza et al., 2016), as described in Equation (1):

$$s_{\text{MAD}} = 1.4826 \times \text{median}\left\{\left|x_i - \text{median}\left(x\right)\right|\right\} \tag{1}$$

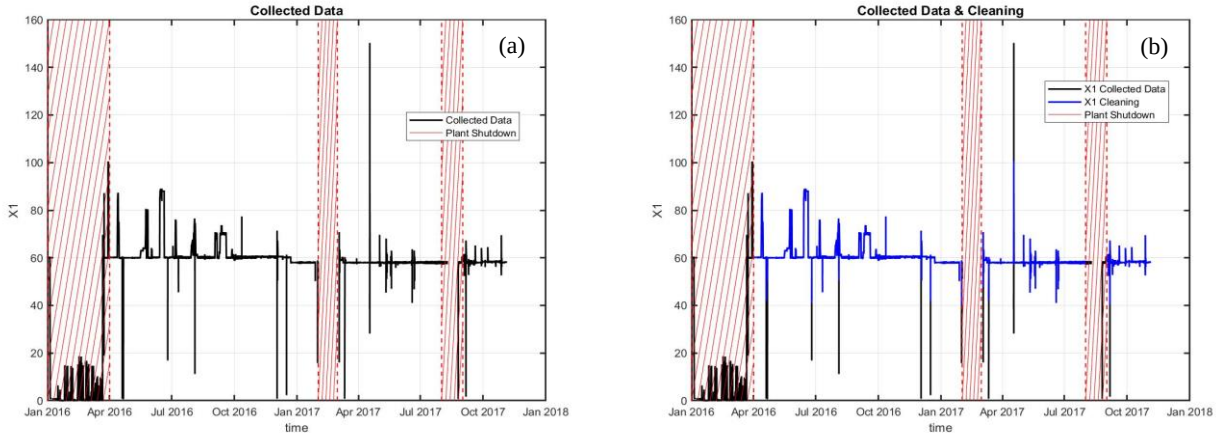
where the coefficient 1.4826 is chosen such that the expected MAD corresponds to the standard deviation $\hat{\sigma}_x$ for normally distributed data.

In order to make the approach adaptive to local characteristics of data, it is possible to apply the Hampel identifier over a moving window. This approach has two tuning parameters: the cut-off threshold and the width of the time window. We will call it the adaptive Hampel identifier method.

For illustration purposes, Fig. 3 presents some of the results obtained in the cleaning stage of the data analysis workflow. Fig. 3.a illustrates raw data for process variable X_1 (process variables are anonymized for protecting critical industrial information). The red dashed lines identify time periods corresponding to plant shutdowns, and therefore should be removed from the analysis. Other abnormal values can be identified by applying simple variable-dependent thresholds, called operation limits. Each process variable, therefore, has its own operation limits and Fig. 3.b presents data with these two elimination stages applied, where the blue line corresponds to the cleaned data after application of the operation limits. From Fig. 3.b it is possible to observe that there are still data points left to be removed. In order to accomplish this last cleaning stage, three different types of filters were tested: (i) 3σ rule; (ii) Hampel identifier; (iii) adaptive Hampel identifier.

From Fig. 3.c it is also possible to verify that the upper and lower thresholds obtained from the 3σ rule do not identify most of the data points considered to be outliers. The Hampel identifier leads to similar results as the 3σ filter, failing to detect and remove most of the data points classified as outliers.

The moving window technique diminishes the impact of an outlier in the calculation of the new thresholds, because it does not consider the data set as a whole, but only the local variability. Regarding the application of the adaptive Hampel identifier, for each variable, the window size was varied using the following alternative lengths: 50, 500, 1000, 2000, 3000 and 5000. The selection of the best window size in each case was performed by visual inspection, in order to validate that the points identified were in fact outliers and that no relevant information was eliminated.



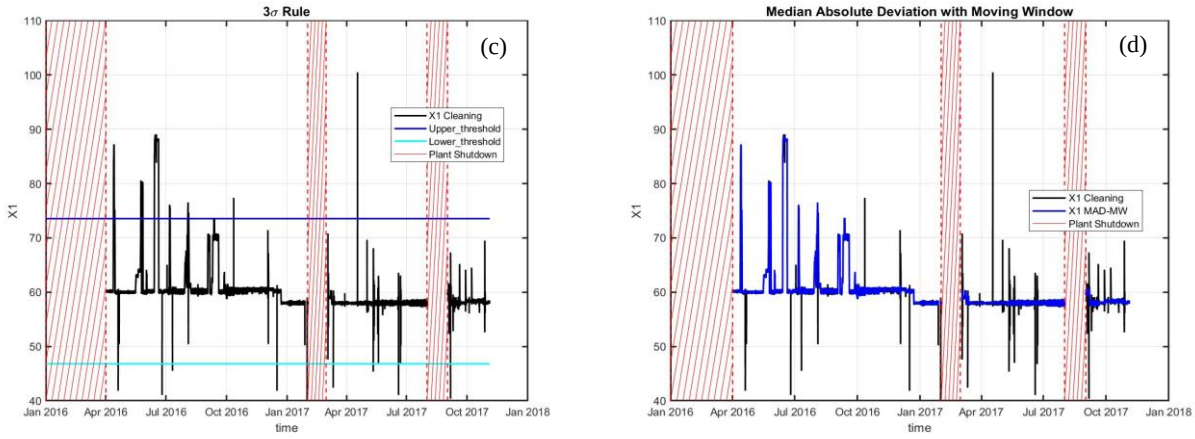


Fig. 3. Comparison of various cleaning steps over the same variable, X_1: (a) No cleaning filter; (b) Shutdown filter; (c) 3σ filter; (d) 3σ Hampel identifier with moving window technique. Black lines represent collected data, while horizontal lines represent the 3σ thresholds.

Fig. 3.d presents the results obtained after applying the adaptive Hampel filter for a size window of 2000, where it is possible to observe that most of the outliers were identified and removed with this filter. The blue line represents the data at the end of the cleaning stage. Note that, for this variable, by increasing the window size we would not be able to identify the remaining outliers.

Similar approaches were applied to all the available process variables, with identical results, and it is important to stress how critical this step of data cleaning is when we need to address complex sets of real industrial data.

2.3.2. Pre-processing

The data pre-processing stage is divided into two steps: (i) data granularity selection; (ii) missing data imputation. Both of them were found to be quite important for the analysis outcome and therefore should be carefully conducted.

2.3.2.1. Granularity Selection for Data Analysis

After the data cleaning stage, it is necessary to establish the granularity level for both process and quality variables. This decision depends on the analysis goal and structure of the available data. It is a very important stage, often overlooked in process analytics activities. Granularity is given by the length of the non-overlapping time windows over which measurements are aggregated by averaging. This selection is goal dependent. The main goal of this work is not only to develop a predictive model with good accuracy but also to provide information about the key variability drivers and could be applied to the study of catalyst deactivation. As this phenomenon spans periods of months and the RON measurements are collected only a few times per week, in this work we have set the granularity level to the day (24h). This means that process variables and laboratory measurements from one day were aggregated and saved for further analysis. This choice also mitigates the sparse multirate structure of the raw data table and reduces the need for missing data imputation. More specifically, as the target response variable, RON, is only available once per day (approximately), an analysis conducted at a finer granularity would lead to data tables lacking the required response values, a problem that is greatly mitigated by

setting the resolution at 1 day. This type of data aggregation will also help reduce our data set volume, while keeping the information accurate enough for the intended goal. Table 1 presents the pseudo-algorithm used for establishing the granularity level in all process and laboratory variables. The granularity is defined by the parameter *r*, which represents the time support of the aggregation operations.

Table 1 Pseudo-algorithm for establishing data granularity/resolution.

Algorithm Averaging at a resolution *r*

Input a data set (**X)**, selected resolution (*r*)

1. Obtain the original (*n*) number of samples of **X** and the new number of samples after the aggregation operations (*nnew*);
2. For each variable (*j*), compute the coarser resolution values by applying the median in each aggregation window.

Output a data set at resolution *r*: **X_new.**

```
[n,m] = size(X);
nnew = floor(n/r);
for j = 1:m
    for k = 1:nnew
        idx_aux = (k-1)*r+1:k*r;
        X_new = nanmedian{X(idx_aux,j)};
    End
End
```

Fig. 4 illustrates the result of selecting the granularity of 1 day, when applied to variable X_1, thus forming 523 new observations, corresponding to the medians of the aggregation periods. After this operation, variable X_1 still has 7 missing values, and such an issue needs to be handled in the next step of the workflow: missing data imputation.

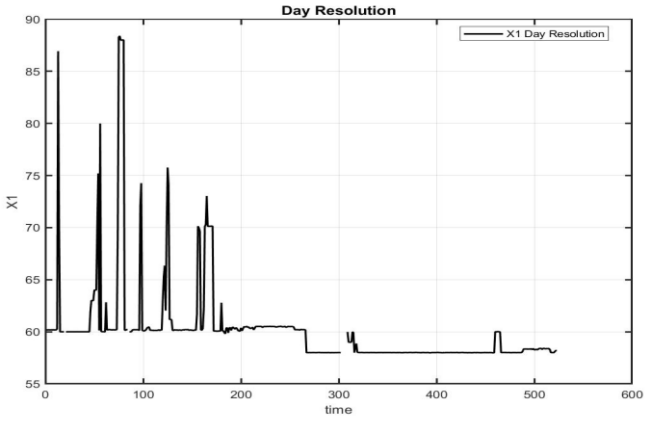


Fig. 4. Variable X_1 represented with a granularity level of 1 day.

2.3.2.2. Missing Data Imputation

Selecting a coarser data granularity mitigates the existence of missing data. Despite that, some empty records may remain, thus requiring the application of missing data imputation techniques. Fig. 5 provides the percentage of missing data for each variable, when adopting a granularity level of 1 day, for our industrial case study. For example, variable X_1, at the granularity level of 1 day, has 7 missing records out of 532 (1.34 % of of missing data).

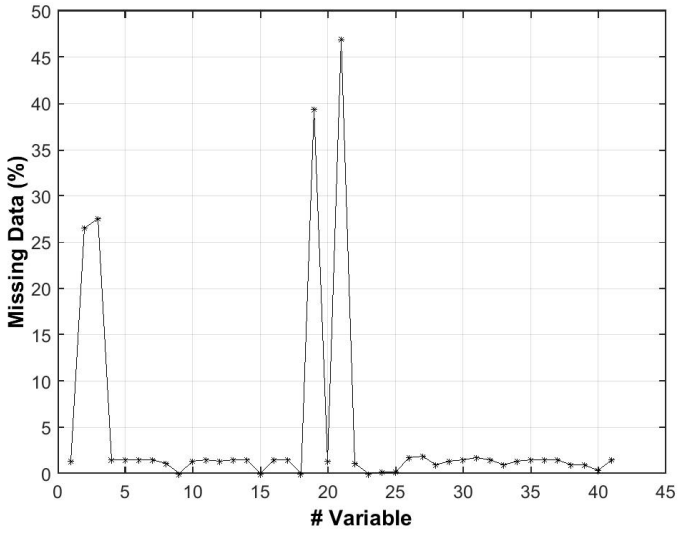


Fig. 5. Percentage of missing data for all variables, at the granularity of 1 day.

Data imputation exploits some type of potential redundancy in the data: either variables' mutual correlation, autocorrelation or both. In case the dominant correlation pattern occurs between variables, a variety of methods were proposed to exploit it for missing data estimation under MAR (Missing at Random) and MCAR (Missing Completely at Random) scenarios. Often they implement EM (Expectation-Maximization) and regression approaches, both for small and large data sets (Arteaga and Ferrer, 2002; Little and Rubin, 2002; Nelson et al., 1996; B Walczak and Massart, 2001; B. Walczak and Massart, 2001). The other source of redundancy often found in industrial data sets occurs across the observations' mode, due to the dynamic patterns of the variables over time, i.e., their autocorrelation. The existence of large process units with consequently large inertial elements and large time constants, from which data are collected at fast acquisition rates, create conditions for such strong autocorrelation patterns to naturally emerge. The autocorrelation function represents the degree of similarity between a given time series and a lagged (i.e., time shifted) version of itself, for different lags. More specifically, given the sequence of measurements, $Y_1, Y_2, ..., Y_N$, obtained at regular instants, $t_1, t_2, ..., t_N$, the autocorrelation function of Y for lag k is defined as:

$$r_k = \frac{\sum_{i=1}^{N-k}\left(Y_i - \overline{Y}\right)\left(Y_{i+k} - \overline{Y}\right)}{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2} \tag{2}$$

Fig. 6.a, presents the autocorrelation for variable X_1, and its correlation with other variables, where it is clear the strong autocorrelation for small lags.

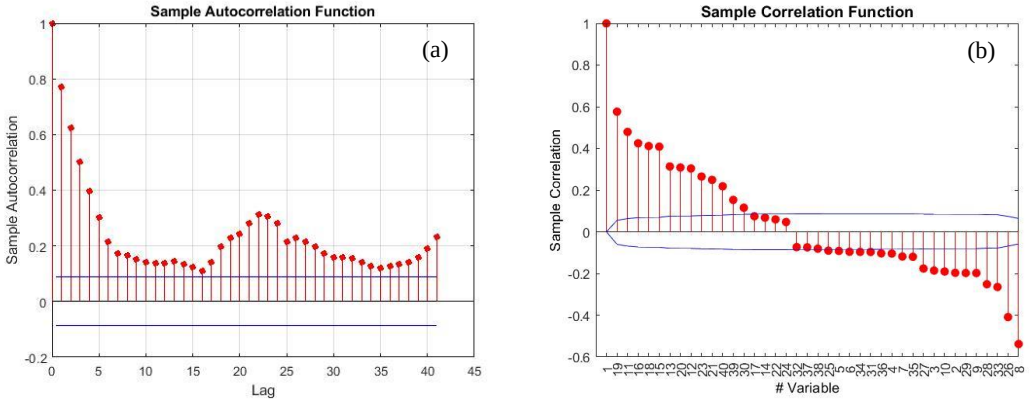


Fig. 6 (a)Autocorrelation function for variable X_1 at the granularity level of 1 day; **(b)** Pearson correlation between X_1 and other process variables.

When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also closely associated. The correlation of X_1 and the rest of the variables present in Fig. 6.b is not so strong as the autocorrelation (note that the value of 1 concerns the correlation of X_1 with itself). Therefore, as the dominating correlation structure in our dataset corresponds to variables' autocorrelation, we have adopted robust interpolative methods to estimate missing records. More specifically, the imputation was performed via a moving window median approach, where missing data at the centre is replaced by the median of the data points falling within the moving window. This process is repeated for each variable, after the length of the window is specifically adjusted. Like for the outlier detection approach, after the imputation there is a need to validate visually the imputation results for all the variables (Chiang et al., 2003).

In summary, in this work we have mainly exploited the variable autocorrelation to estimate missing data, using interpolation schemes. This is the more reliable way, as the variable-wise correlation structure was not as strong as their autocorrelation. EM methods could also be used for exploiting correlation and autocorrelation, but the simpler interpolative methods showed satisfactory robustness and accuracy, and were therefore adopted.

After missing data imputation it is also possible to conduct a PCA analysis and use its capabilities to identify unusual observations in the PCA subspace or around it, namely through the well-known Hotelling's T^2 statistic of the scores and Q (or *SPE* – Square Prediction Error) statistic of the residuals, respectively. We have implemented such an approach and detected some outlying observations. However, since only a small fraction of the total number of observations are possible outliers, and there was no information in the refinery records referring they were not valid measurements, we decided to follow a more conservative approach and did not remove them, even though we are aware that this would lead to a lower prediction performance of the methods. Observations such as these happen regularly in process plants, and we think it is sensible to include them in the assessment of the methods (which also puts into test their robustness to deviating observations).

2.4 Modelling Stage

Selecting the best type of modelling approach is an important task in the design of an inferential model. However, there are no general guidelines for defining a priori what the most adequate methodology should be. Therefore, our proposal here is to study the prediction ability of different methodologies arising from different corners of the analytical spectrum. A brief description of the eleven methods considered in this work will be presented in Section 3. It is possible to verify from this sample of methodologies that one of our concerns was to have representatives from the main classes of methods for handling large scale complex data sets, such as those arising from many industrial scenarios, like the one addressed here.

In our data analysis workflow, the predictive methods are compared with resort to a robust procedure which is based on a Monte Carlo double cross-validation (Geisser and Eddy, 1979; Krzanowski, 1982; Stone, 1974; Wold, 1978, 1976), as shown in Fig. 7. The specific procedure adopted consists of implementing several Monte Carlo runs, where data are randomly partitioned into two groups: (i) a training set that is used to select the hyper-parameter(s) for the model and to estimate the model; (ii) a testing set used to predict the response variable, by applying the model derived with the training set. This process is repeated several times, and the prediction accuracy of each model is evaluated using the $RMSE^{test}$ and R^2_{test} scores (the hyper-parameter(s) considered for each method, range of values and the strategy used to select them in Step 2 are presented in Table A.1 of the Appendix).

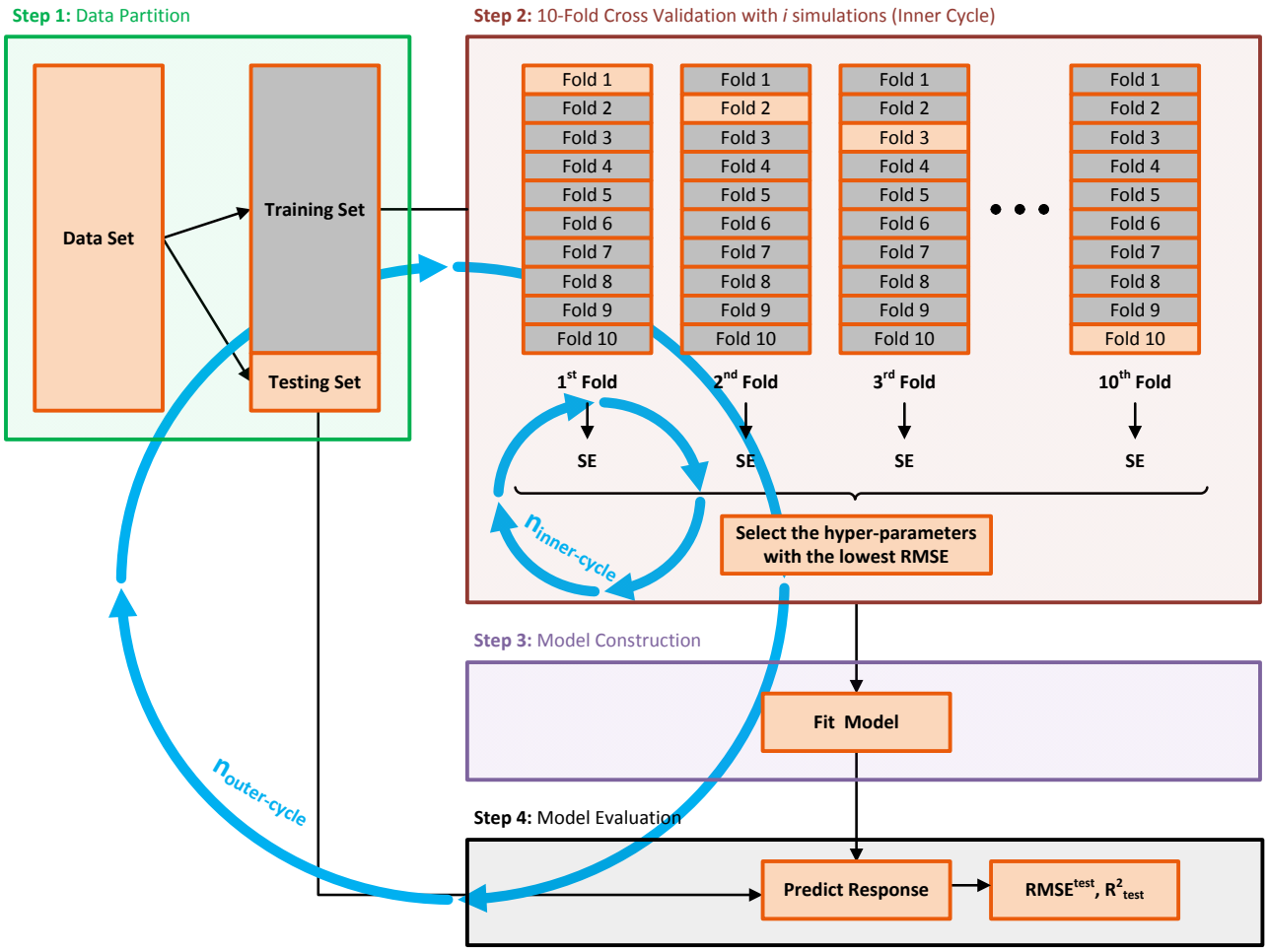


Fig. 7. Schematic illustration of the Monte Carlo double cross-validation methodology.

More specifically, the data partition (Step 1) will create a training set that corresponds to 80% of the number of data records from the original data set and a testing set with the remaining 20% data records. This split of data can occur in three ways: (i) order split (the first 80% samples go to the training set and the remaining ones to the testing set); (ii) random split; (iii) stratified sampling split. The stratified sampling split consists on splitting the response variable into a pre-selected number of intervals based on its percentiles (e.g., 0-25th, 25th-50th, 50th-75th and 75th-100th percentiles). From each group, 80% of the data will be randomly selected to form the training set, and the rest goes to the testing set, this being the approach that we adopted here.

The training set is used to optimize the selection of the hyper-parameter of each model. Since in some industrial processes it may be difficult to obtain sufficient historical data to develop a model, it is advantageous to use the K-Fold cross-validation technique (Geisser, 1993; Kohavi, 1995) in this phase. In K-Fold cross-validation (represented by Step 2 of Fig. 7), from the existing K folds, $(K-1)$ are retained to train a model, and the remaining fold is used to validate it. This process is repeated K times, ensuring that each fold is used once and only once, in the validation process. The Root Mean Square Error (RMSE) in the left-out fold, obtained for each realization of the hyper-parameter, is saved and the realization leading to the lowest RMSE is adopted for model development purposes.

After the selection of the hyper-parameter(s), the predictive model is estimated (Step 3) using the entire training set. Then, it is applied to the test set, for predicting the response variable over these new observations (Step 4), and their performance is assessed by the corresponding $\mathrm{RMSE}^{\mathrm{test}}$ and coefficient of determination ($\mathrm{R}^2_{\mathrm{test}}$) scores, with $\mathrm{RMSE}^{\mathrm{test}}$ and $\mathrm{R}^2_{\mathrm{test}}$ calculated as:

$$\mathrm{RMSE}^{\mathrm{test}} = \sqrt{\frac{\sum_{i=1}^{n_{\mathrm{test}}} \left(\hat{y}_i - y_i^{\mathrm{test}} \right)^2}{n_{\mathrm{test}}}} \tag{3}$$

$$\mathrm{R}^2_{\mathrm{test}} = 1 - \frac{\sum_{i=1}^{n_{test}} \left(y_i^{\mathrm{test}} - \hat{y}_i \right)^2}{\sum_{i=1}^{n_{test}} \left(y_i^{\mathrm{test}} - \bar{y} \right)^2} \tag{4}$$

where y_i^{test} is the i^{th} observed response of the testing set, $\bar{y}$ is the mean of y, $\hat{y}_i$ is the i^{th} estimated response and n_{test} is the number of observations belonging to the testing set.

This entire procedure (from Step 1 to Step 4) is repeated $n_{\text{outer-cycle}}$ times, in order to avoid any bias due to the random assignment of data records between the training and test sets. For each run of this outer cycle, the $\mathrm{RMSE}^{\mathrm{test}}$ and $\mathrm{R}^2_{\mathrm{test}}$ scores are saved and will be used to make a final comparison between the different methods according to a formal statistical testing procedure, namely a pairwise student's t-test, in order to perform a robust and precise comparative evaluation of performance between the different methods used in the analysis. More specifically, the statistical test indicates if there is a statistically significant difference in performance ($\mathrm{RMSE}^{\mathrm{test}}$) between each pair of methods (significance level set to $\alpha = 0.05$). If so, the best method (the one with lower $\mathrm{RMSE}^{\mathrm{test}}$) receives 2 points, and the other 0. In case the prediction performance of the two methods over the outer test data trials is not statistically distinct, a tie has occurred, and both methods receive one point. Applying this assessment system to all pairwise combinations of methods, and summing the scores for each method leads to their Key Performance Index (KPI) values. A high value of KPI indicates that the regression method is consistently better than the others (in a statistically significant way). The maximum value of KPI is thus $2 \times (m-1)$ when using *m* methods for comparison purposes.

Since some of the regression methods under comparison are sensitive to scale, the variables were normalized, or autoscaled. This consists of centering all variables to zero mean and scaling them to unit variance (Jolliffe, 2002; Naes et al., 2004). This procedure was not applied to the entire data set, but just to the training data sets and the scaling parameters were stored, so that they could be applied to scale the testing data set.

3. Predictive Modelling Methodologies

There are many regression methodologies currently available to perform predictive modelling. They can be just variants of the same basic approach or present strong differences in their assumptions (e.g., regarding collinearity, sparsity, non-linearity, etc.), estimation procedures and final outcomes. We adopt here the systematic grouping of scalable regression methods into four categories proposed elsewhere (Rendall and Reis, 2018), where methods were classified as: variable selection methods (Section 3.1); penalized methods (Section 3.2); latent variable methods (Section 3.3) and tree-based ensemble methods (Section 3.4). Methods from each of the four groups will be considered in our study, guaranteeing in this way that the analytical landscape is covered in a balanced fashion, and that different modelling approaches are explored during model development, and later on that the corresponding performances can be compared.

The classical Multiple Linear Regression (MLR) approach was also included in this study, since it is well-known, widely used and has been implemented with several variable selection methods (Draper and Smith, 1998; Montgomery et al., 2012). The MLR model, $Y = b_0 + \sum_{j=1}^{p} b_j X_j + \varepsilon$, assumes that only the response variable carries a sizeable error, which is additive and homoscedastic (constant variance). The regression coefficients are found by least squares fitting, as described in Equation (5):

$$\hat{\mathbf{b}}_{\mathbf{MLR}} = \underset{b=[b_0 \ldots b_P]^{\mathrm{T}}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \right\} \tag{5}$$

where $\hat{\mathbf{b}}_{\mathbf{MLR}}$ is a vector containing the regression coefficients, n is the number of observations, y_i is the i^{th} observed response value and $\hat{y}_i$ the respective estimated response. MLR faces problems when predictors present moderate-high levels of collinearity, because the estimation of the regression coefficients becomes unstable (high variance). In that case, other approaches are available that lead to more stable predictions.

In the next subsections, the four groups of methods mentioned above will be briefly introduced (Sections 3.1 to 3.4).

3.1. Variable Selection Methods

Variable selection methods assume that a subset of all the predictors does provide significant information concerning the response variable (predictors' sparsity). Examples of methods that fall under this group are the following: forward stepwise regression (FSR) (Andersen and Bro, 2010; Montgomery and Runger, 2003; Murtaugh, 1998); MLR coupled with genetic algorithms for variable selection (GA) (Goldberg, 1989; Leardi, 2007, 2003; Leardi et al., 1992; Sofge, 2002); and best subsets (BS) (Anzanello and Fogliatto, 2014; Zhang, 2016). In this work, only the first methodology was tested. FSR begins with no predictors selected and adds them step by step if they bring a statistically significant improvement to the explanatory capability of the model. Predictors that were first introduced can also be removed later on, if they are no longer relevant (i.e. statistically significant) when others have also been added to the model. The final estimates for the regression coefficients $\left(\hat{\mathbf{b}}_{\mathbf{FSR}}\right)$ are obtained by applying the MLR method to the selected predictors.

3.2. Penalized Regression Methods

Methods belonging to the penalized regression group have in common the introduction of a regularization strategy to decrease the model's variance and stabilize the estimation of the regression coefficients. This class of methods includes, for instance: ridge regression (RR) (Draper and Smith, 1998; Hoerl and Kennard, 1970; Yan, 2008); least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996); and elastic net (EN) (Hastie et al., 2009; Hesterberg et al., 2008; Zou and Hastie, 2005) RR is an extension of linear regression, obtained by introducing a L2-norm penalization for large regression coefficients into the original MLR formulation, as described in Equation (6):

$$\hat{\mathbf{b}}_{\mathbf{RR}} = \underset{b=[b_0 \ldots b_P]^{\mathbf{T}}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \sum_{j=1}^{p} b_j^2 \right\} \tag{6}$$

where λ is a hyper-parameter responsible for the regularization.

This type of penalization enforces the regression coefficients to be small, but hardly zero. Note that if λ is set to zero RR becomes MLR.

LASSO regression is similar to RR but using a L1-norm penalization. This difference in the penalization strategy does have an important consequence: with LASSO, coefficients from irrelevant variables are effectively shrunk to zero. For that reason, LASSO is known to perform simultaneously stabilization of parameter estimation and variable selection, leading to stable models with fewer predictors than the original data set. The regression coefficients for LASSO are calculated as:

$$\hat{\mathbf{b}}_{\mathbf{LASSO}} = \underset{b=[b_0 \ldots b_P]^{\mathbf{T}}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \sum_{j=1}^{p} \left| b_j \right| \right\} \tag{7}$$

where λ is a hyper-parameter responsible for the regularization factor.

EN regression is a generalization of RR and LASSO, having them as particular cases. Coefficients are estimated by solving Equation (8), where one can find the penalty terms of RR and LASSO. Now, two hyper-parameters need to be tuned: α and λ, where λ is the hyper-parameter that controls the bias-variance trade-off and α is a hyper-parameter that weights the L1-norm (LASSO, $|b_j|$) and L2-norm (RR, b_j^2) penalties:

$$\hat{\mathbf{b}}_{\mathbf{EN}} = \underset{b=[b_0 \ldots b_P]^{\mathbf{T}}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} \left| b_j \right| + \frac{1-\alpha}{2} \sum_{j=1}^{p} b_j^2 \right) \right\} \tag{8}$$

From Equation (8) it is possible to obtain RR and LASSO techniques by manipulating the hyper-parameter α. RR is obtained by setting $\alpha = 0$, which means that only the squared penalty is applied. On the other hand, LASSO is obtained by setting $\alpha = 1$.

3.3. Latent Variable Methods

Latent variable methods are based on the assumption that a few underlying unobserved variables (called latent variables), are responsible for the observed variability on both the $\mathbf{X}$ and $\mathbf{Y}$ variables. The latent variables are estimated through linear combinations of the measured original variables. This class of methods is appropriate to handle data sets with high levels of collinearity, since in this case the reduced set of latent variables is able to efficiently extract the main patterns of variation and explanation. From this group, three methodologies were analysed here: principal component regression (PCR) (Jackson, 1991, 1980; Jolliffe, 2002; Krzanowski, 1988; Martens and Naes, 1989; Wold et al., 1987), principal component regression with a forward stepwise selection strategy (PCR-FS) and partial least squares (PLS) (De Jong, 1993; Geladi, 1988; Geladi and Kowalski, 1986; Haaland and Thomas, 1988; Helland, 2001, 1988; Höskuldsson, 1996; Jackson, 1991; Lindgren et al., 1993; Wold et al., 1984, 2001).

These methods have been used in a wide variety of chemical engineering applications, including process control (Dayal and MacGregor, 1997); property prediction (Qin, 1998) and statistical process monitoring (Kourti and MacGregor, 1996; Kresta et al., 1991; MacGregor and Kourti, 1995; Negiz and Çinar, 1998; Nomikos and MacGregor, 1995); and fouling detection (Kourti and MacGregor, 1995).

PCR (Jackson, 1991; Jolliffe, 2002; Wold et al., 1987) consists of applying principal component analysis (PCA) to the $\mathbf{X}$ variables and then using the resulting components (also called scores) as regressors to predict the target response. PCA finds the directions (also mentioned as principal components) that maximize the variance of $\mathbf{X}$, which are obtained by building linear combinations of the original predictor variables. The principal components are not correlated among themselves and a small number of PC (p_{PCR}) is often good enough to describe the overall variability of $\mathbf{X}$. The PCA decomposition is presented in Equation (9):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{9}$$

where $\mathbf{T}_{n\times p}$ is the matrix of PCA scores (the i^{th} column corresponds to the scores for the i^{th} PC), $\mathbf{P}_{m\times p}$ is the matrix of PCA loadings (the i^{th} column contains the loadings of the variables for the i^{th} PC), $\mathbf{E}_{n\times m}$ is the residual matrix and p stands for the number of PC selected. The final model is built with the first p principal components, which define the hyper-parameter space for this method, and its value can be selected by methodologies such as cross-validation.

The PCR-FS methodology is comparable to PCR, but in this case principal components are added to the model, based on the forward stepwise methodology (see Section 3.1), and only the principal components $\left(p_{\text{PCR-FS}}\right)$ which are statistically significant to explain $\mathbf{Y}$ do enter the model. It is important to mention that this is the only difference between PCR and PCR-FS, since the $p_{\text{PCR-FS}}$ components are not necessarily the same as the p_{PCR} components for PCR.

The last method included from this category is PLS (Geladi and Kowalski, 1986; Park and Han, 2000; Wold et al., 2001, 1984; Zamprogna et al., 2004). PLS shares some similarities with PCR due to the fact that it also searches for directions maximizing a certain criterion. However, instead of maximizing the explanation capability for $\mathbf{X}$ variability, PLS finds directions (also known as latent variables) that maximize the covariance between the $\mathbf{X}$ and $\mathbf{Y}$ latent variables. In other words, the subspace selected in PCR is the one explaining the most of the variability in $\mathbf{X}$, but this is not necessarily the

subspace most explicative of the variability in $\mathbf{Y}$. As in the case of PCR, the number of latent variables (p_{PLS}) is the hyper-parameter for this model and it can be selected for instance using cross-validation.

3.4. Tree-Based Ensemble Methods

The ensemble methods based on regression trees do form the last class of regression approaches included in this study. Regression Trees essentially split the $\mathbf{X}$ space into a set of hyper-rectangular regions. This partition is performed with the objective of grouping observations with similar response values (Breiman et al., 1984; Dietterich, 2000; Hastie et al., 2009; Strobl et al., 2009). Firstly, the algorithm finds the splitting variable and the splitting point, aiming to reduce the squared error between the predicted and observed values for the response. In each split, two decisions need to be made: (i) the predictor variable used for the split (splitting variable); (ii) the splitting value for that variable. Each region can be further sub-divided, always with the target of minimizing the squared errors, until a certain stopping criteria is met. These models are very flexible and capable of capturing non-linear relationships, discontinuities, etc.

A major concern during the implementation of tree-based methods is the variance of the trees created, since a small change in the training data can originate a different tree, with different variable and value splits. To overcome this problem, ensemble methods are employed: instead of using a single regression tree, many of them are generated and used simultaneously.

Bagging of regression trees (BRT) combines bootstrapping and averaging techniques to create an ensemble model, and the predicted value is the mean value computed over all the trees of the ensemble. The Random forests (RF) approach is comparable to BRT, with the difference that each tree of the ensemble only considers a random subset of the available predictors. For both BRT and RF, the number of trees (t_{BRT}) and (t_{RF}) is selected by cross-validation. The last method considered in this class is boosting regression trees (BT). Boosting is a method where the model is sequentially constructed by fitting a simple regression model to the current residuals that were not fitted by previous models (Cao et al., 2010; Elith et al., 2008). The iterative boosting process only considers the current residual of $\mathbf{Y}$ and fits the residual of $\mathbf{Y}$ with the original $\mathbf{X}$. The number of trees (t_{BT}) is also selected by cross-validation.

No method from the classes referred above is expected to perform always better than the others and claim overall predictive superiority (Murphy, 2012). Therefore the final decision about which method to use should be based on a rigorous consideration of the options available. When the choice is not obvious, the decision process benefits from a robust comparative analysis. Thus, a state-of-the-art comparison methodology based on Monte Carlo double cross-validation, as described in Section 2.4, was implemented, in order to establish rankings of the best methods to adopt for addressing a particular problem, like the one we are handling here.

4. Results

In this section, we report a summary of the main results obtained in this work, with a special focus on the comparison of performances for the different predictive methodologies used to estimate industrial RON values for a major petrochemical facility.

4.1 Prediction Accuracy

As described previously, $RMSE^{test}$ and R^2_{test} are the two parameters employed to evaluate the prediction capabilities of the different methods tested, which are presented in Table 2. The coefficient of determination R^2 between predicted and observed response values, obtained over a large number of cross-validation trails, assesses the methods' prediction accuracy in terms of a normalized parameter. The $RMSE$ is another common measure of accuracy that estimates the standard error of prediction obtained for the different models used.

The values of R^2 range from 0 (the dependent variable cannot be predicted from the regressor variables) to 1 (the response variability can be fully captured by the model). The closer this parameter is to 1, the better the model is on predicting the response variable. The $RMSE$ is a non-negative quantity, and the lower its value, the better the prediction performance of the corresponding model.

Table 2. Average $RMSE^{test}$ and R^2 in test conditions over all cross-validation trials, $\overline{RMSE^{test}}$ and $\overline{R^2_{test}}$, for each regression method tested.

Parameter	MLR	FSR	RR	LASSO	EN	PCR	PCR-FS	PLS	Bagging	Random Forests	Boosting
$\overline{RMSE^{test}}$	0.76	0.77	0.73	0.72	0.72	0.93	1.09	0.73	0.41	0.40	0.50
$\overline{R^2_{test}}$	0.71	0.70	0.73	0.74	0.74	0.57	0.41	0.74	0.91	0.92	0.87

From Table 2, it is possible to verify that all methods present acceptable performances in terms of prediction accuracy, taking into account that they were developed from real plant data, with all the common variability sources and uncertainties that are usually present under these real world operation scenarios. These results also point to a certain advantage of using tree-based ensemble methods over the remaining linear modelling approaches tested. Fig. 8 presents the Observed versus Predicted scatterplots for such a class of non-linear methods, where it can also be observed that random forests regression is indeed the best method, potentially capturing a non-linear behaviour in the dependence of RON from process variables.

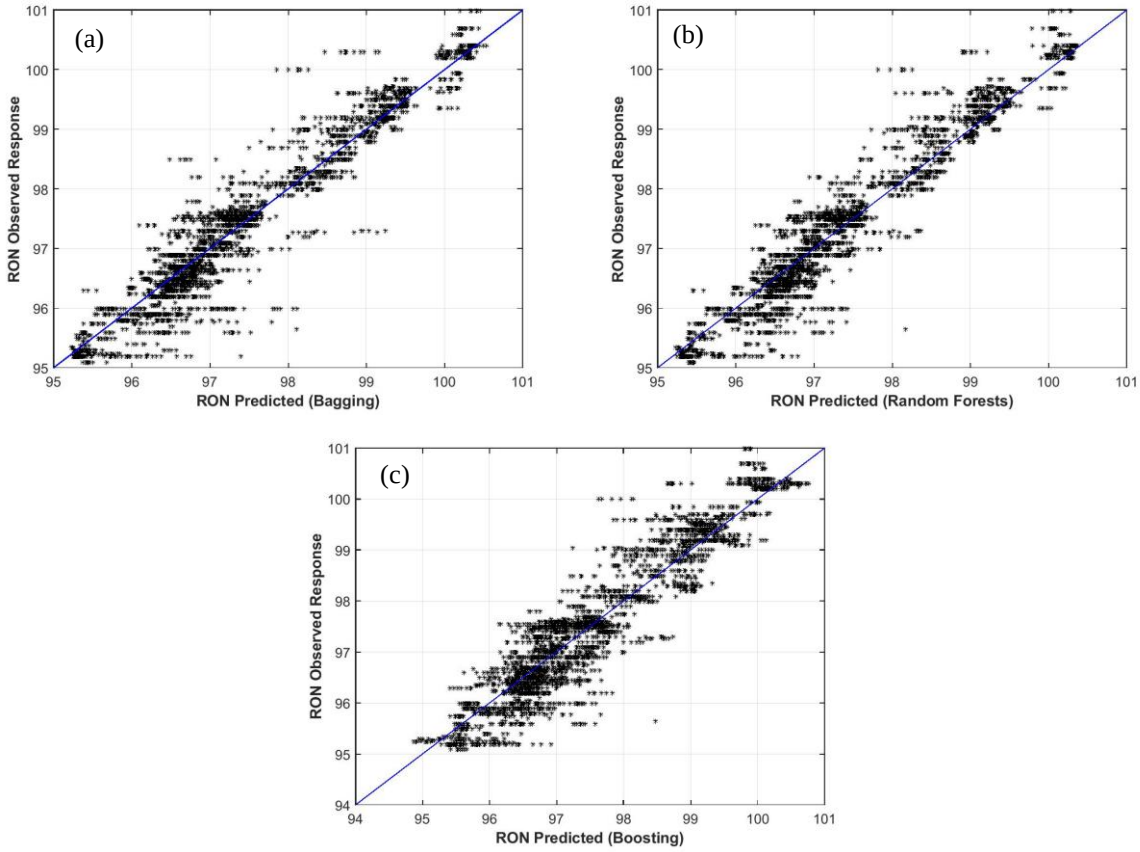


Fig. 8. Observed vs Predicted scatterplots for the RON under testing conditions in all outer cycles of the double cross-validation comparison procedure for: (a) bagging of regression trees; (b) random forests; (c) boosting of regression trees.

As referred in Section 2.4, since all the methods were applied using exactly the same train and test data sets for each Monte Carlo run, it is possible to perform a pairwise statistical comparison for all pairs of methods, and summarize the overall results in a Key Performance Indicator (KPI) parameter obtained for each method, where the higher KPI the better the method is going to be, when compared with others.

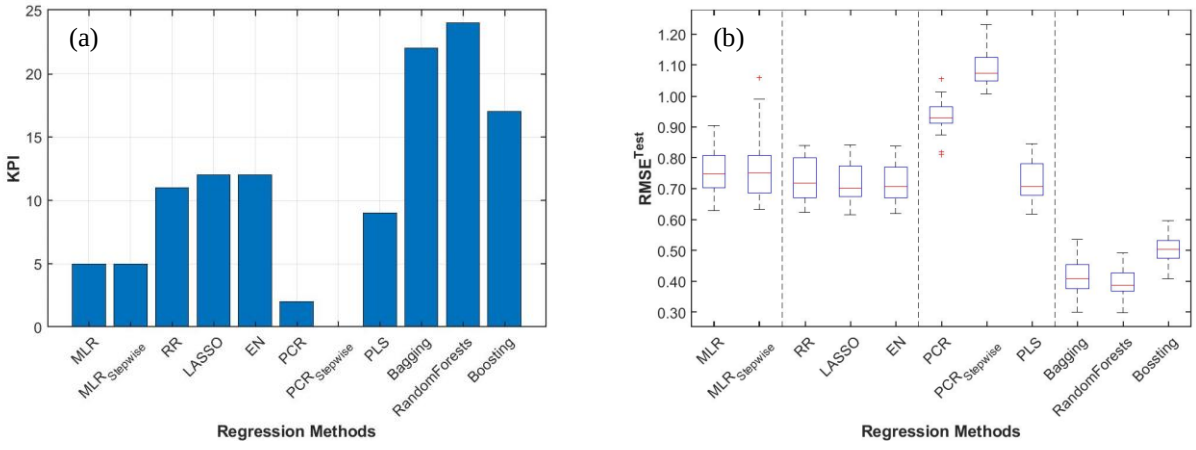


Fig. 9. (a) KPI obtained for each regression method; (b) Box plots of $RMSE^{test}$ considering all the cross-validation trials.

Fig. 9 confirms, once again, the superiority of non-linear methods and regression tree-based approaches as the class with the most promising results (highest KPI scores), with random forests obtaining the best scores for this industrial set of data.

The overall good performance presented by the ensemble methods may indicate that the relation between RON and the predictors is indeed non-linear and that we may have different zones of the model space where different relationships do prevail.

Regarding now linear approaches, the class of penalized methods is the one leading to highest KPI scores. Fig. 10 presents some results for LASSO, including the Observed versus Predicted scatterplot (a) and a ranked bar plot that indicates the relative importance of the variables in the model (b) (described in more detail in the next section).

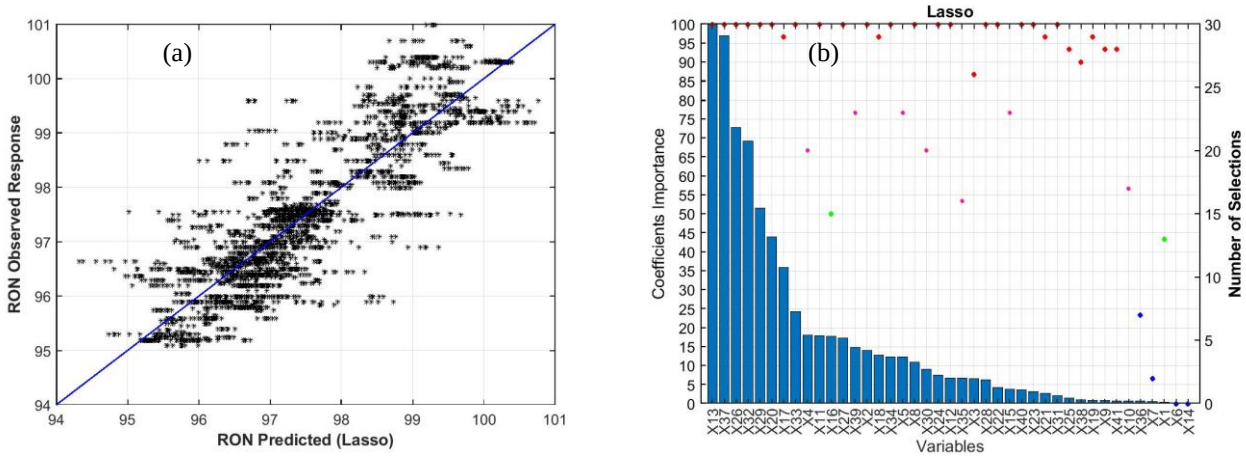


Fig. 10. Results from the LASSO model: (a) Observed vs Predicted values of RON and (b) bar plot of coefficients' importance.

The practical interest of deriving regression methods is not limited to provide accurate estimates of the response variable. Another relevant insight that can be obtained from them is the diagnostic of critical/important variables. From Fig. 10.b, one can notice that, for the specific context of LASSO, the two most important variables are X_{13} and X_{37}. The coloured dots are related to the number of times each variable entered the LASSO model during the cross-validation runs. The blue dots are variables that were selected less than 25% of the times, green from 25% to 50%, magenta between 50% and 75%, and red when variables were selected more than 75% of the times. These results are also consistent with the following further discussion and analysis of the importance of the several independent variables under consideration over RON behaviour.

4.2 Analysis of Variables' Importance

To further assess the importance of each variable, a diagnosis of the relative importance of each independent variable for each method was also conducted. Furthermore, a methodology was devised for obtaining the global importance of each variable, under the scope of all tested methods. This methodology is based on the analysis of the regression vectors obtained with the several linear regression methodologies contemplated in our study. The rational is that variables' importance mainly concerns the relevance of main effects in a regression model and these are well estimated in such linear frameworks. However, alternative methods do exist (for instance Random Forests have a built-in feature for assessing variables). The procedure followed is presented in this section.

The importance of variable i according to the method j is related to the magnitude of the corresponding regression coefficient, $\left|\beta_{i,j}\right|$ (note that data is autoscaled, and therefore the magnitude of the coefficients reflects the importance of the respective variables, independently of their original units and scales of measurement). Since there are k outer cycles (Monte Carlo simulations) the cumulative score, $B_{i,j}$, for each variable and method is calculated by Equation (10):

$$B_{i,j} = \sum_{k=1}^{n_{\text{outer-cycle}}} \left|\hat{\beta}_{i,j}^{(k)}\right| \tag{10}$$

The relative normalized importance of a variable for a specific method is computed by Equation (11). This expression assigns values between 0 and 1 to each variable. A relative importance of 0 means that the variable is the least important for a respective method, while a relative importance of 1 corresponds to the most important variable for that particular method:

$$I_R\left(X_{i,j}\right) = \frac{B_{i,j} - \mathbf{MinB}}{\mathbf{MaxB} - \mathbf{MinB}} \tag{11}$$

where $\mathrm{I_R}\left(\mathrm{X_{i,j}}\right)$ stands for the relative importance of variable i for method j; $\mathbf{MinB} = \min_i\left\{B_{i,j}\right\}$ and $\mathbf{MaxB} = \max_i\left\{B_{i,j}\right\}$, representing the minimum and maximum in the set of all $B_{i,j}$ for method j, respectively.

The relative importance of a variable under the scope of a given method, as defined by Equation (11), can be used to derive an overall assessment of variable importance under the global scope of all the methods under analysis. The procedure adopted consists of computing a weighted average of the relative importance for method j, weighted by some function of the quality of model j. In fact, since some methods perform better in predicting the RON than others, this fact was taken into account and methods with better performance are given more credibility in their assessment of the variables' importance. The weight for method j is given by $\mathrm{norm}R^2_{\text{test},j}$, which is the coefficient of determination under test conditions, $\mathrm{R^2_{test}}$ (see Table 3), after normalization to fall under the range between 0 and 1. Normalizing the $\mathrm{R^2_{test}}$ will ensure that the best method will contribute to the global importance with a weight of 1, while the worst one will do so with a weight of 0. The normalized coefficients of determination are computed via Equation (12) (see results in Table 3 and Fig. 11):

$$\mathrm{norm}R^2_{\text{test},j} = \frac{\mathbf{maxR^2_{test}} - R^2_{\text{test},j}}{\mathbf{maxR^2_{test}} - \mathbf{minR^2_{test}}} \tag{12}$$

where $\mathrm{R^2_{test,j}}$ is the coefficient of determination for testing conditions of method j, the term $\mathbf{maxR^2_{test}} = \max_j\left\{R^2_{\text{test},j}\right\}$ and $\mathbf{minR^2_{test}} = \min_j\left\{R^2_{\text{test},j}\right\}$ represent the maximum and minimum values of $\mathrm{R^2_{test,j}}$.

The global importance of each independent variable is then assessed by applying Equation (13):

$$I_G\left(X_i\right)=\frac{\sum_{j=1}^{n_{\text{methods}}} I_R\left(X_{i,j}\right)\times \text{norm}R^2_{\text{test},j}}{n_{\text{methods}}} \tag{13}$$

where $\mathrm{I_G}\left(\mathrm{X_i}\right)$ is the global importance for variable i, $\text{normR}^2_{\text{test},j}$ is the $\mathrm{R}^2_{\text{test},j}$ normalized for method j and n_{methods} stands for the number of methods considered in the study.

Table 3. $\mathrm{R}^2_{\text{test}}$ normalized values for all the regression methods tested.

Parameter	MLR	FSR	RR	LASSO	EN	PCR	PLS
$\text{normR}^2_{\text{test}}$	0.84	0.75	0.95	0.99	1.00	0.00	0.97

The global importance of each variable thus obtained is shown in Fig. 11, with an indication of the industrial section where the process variables came from (more specific information cannot be disclosed given its industrial strategic relevance for the company).

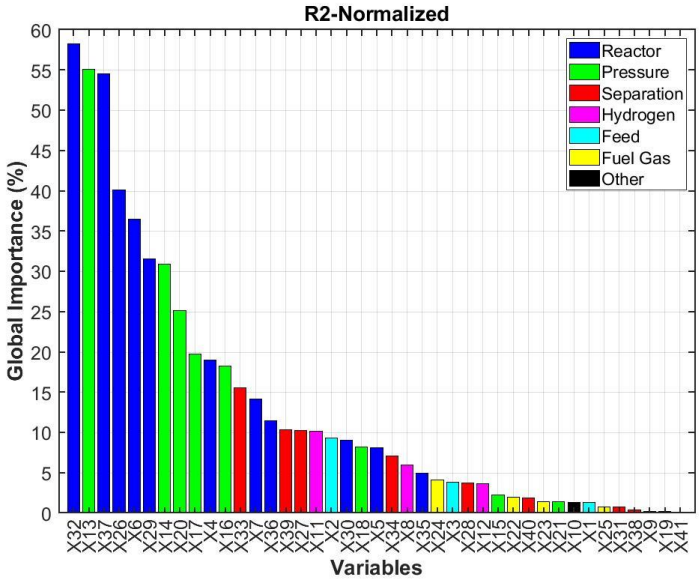


Fig. 11. Global importance for all the predictors.

As can be observed from Fig. 11, variables $\mathrm{X_{32}}$, $\mathrm{X_{13}}$ and $\mathrm{X_{37}}$ show up as the three most important ones, when taking into consideration all the methods under analysis. It is also possible to observe that the most important variables are the ones related to the reaction zone of the plant. This happens because the RON is highly dependent on the composition of aromatic and branched paraffinic compounds, which are produced precisely in the reactor section of the plant. Although, as stated before, no further details can be disclosed about the process, the knowledge obtained from our data analysis was validated by the plant process engineers and provided helpful insights for understanding RON's behaviour and its underlying causes of variation.

5. Discussion

Analysing the results obtained with the different categories of predictive methodologies studied, we can conclude that for this particular industrial problem and data set the class of ensemble methods based on regression trees provides the best performances. These ensemble methods are able to model non-linear relationships and the results obtained suggest that a non-linear relationship, together with possible different relationships across the modelling space, can indeed be present between the predictors and the RON spaces. Predictions made under testing conditions confirm the model accuracy and robustness, achieving high $\overline{R^2_{test}}$ levels of 0.91, 0.92 and 0.87 for bagging, random forests and boosting, respectively. From our experience, these values deserve particular consideration, given the existence of numerous unmeasured sources of variation in a large-scale industrial process such as a refinery, which introduce non predictive components in the data, as well as possibly some missing elements and noise. For the company itself, and its process engineers, such a prediction capability, together with the corresponding process insights extracted from the set of industrial data under exploration, were found to be quite promising, powerful and useful. Therefore, the application of alternative non-linear predictive methodologies will be explored with more detail in our future work.

Furthermore, our results were also able to diagnose the section of the plant and the specific variables having more impact over RON variability and estimation capability. This aspect is also critical for the daily operation of the plant and for sustaining process improvement initiatives aimed at achieving increased performance and better control, as well as to reduce the influence of disturbances over final product quality. It was thus possible to conclude that the most important variables for predicting RON are associated with the reaction zone of the process, something that is in line with the available engineering knowledge about the process, because aromatic and branched paraffinic compounds are produced precisely in the reaction area of the plant, and they are believed to have an important impact over RON values.

6. Conclusions

In this work, a detailed data analytics workflow was presented and applied for addressing the challenging and complex problem of predicting RON in the Catalytic Reforming Unit of an oil refinery, using only process data and identifying the most relevant sources of RON variability. This workflow is now implemented as a generic data analysis platform and includes a resolution/granularity definition stage that is goal dependent. In our work, the resolution was set to be that of the day (24 hour aggregation period), which is adequate for identifying the main variability drivers and achieve models with good prediction accuracy namely for the purposes of RON prediction, monitoring and controlling catalyst deactivation – a phenomenon that spans months of operation in the plant.

A rich variety of predictive methods representative of different classes of regression methodologies was studied (eleven methods overall) and compared for the task of RON prediction. The comparison methodology was based on a double cross-validation approach, in order to assure for accurate and robust assessments of their relative predictive merits.

The best results were obtained for methods coming from the class of ensembles of regression trees. This may indicate the presence of non-linear relationships between predictors and the response variable, as well as different types of relationships to prevail across different parts of the independent variables space. Therefore, new methods from this class should be brought to the analytics workflow in the future and tested, such as Kernel methods, support vector regression with non-linear kernels and neural networks combined with latent variable models (e.g., non-linear principal component regression).

In terms of key predictors, the analysis of the outcomes of estimated models revealed that the most important variables are concentrated in the reaction area of the plant. This diagnostic provided critical insights on how the process can be monitored and controlled in order to better deal with catalyst deactivation.

Many studies of data-driven models are conducted under simulated or well controlled conditions. Therefore, they are not faced with some of the additional challenges that one has to address when dealing with real industrial data collected from an existing plant. In this work, we have shown that even under such more realistic settings, the adoption of appropriate statistical tools for data collection, cleaning, pre-processing and modelling can indeed lead to quite powerful results and conclusions, supporting, in this case, the development of models that are able to estimate quite well the RON values, and therefore to support process improvement efforts, as well as extract useful process knowledge and insights. Such conclusions are also of practical interest and were highly valued and regarded as such by process engineers from the plant where our studies have been conducted.

Acknowledgements

Tiago Dias acknowledges FCT and Galp Energia for the support given to the Doctoral Program through project PB/BDE/128562/2017.

References

Ahmad, I., Ali, G., Bilal, M., Hussain, A., 2018. Virtual sensing of catalytic naphtha reforming process under uncertain feed conditions, in: 2018 International Conference on Computing, Mathematics and Engineering Technologies (ICoMET). IEEE, pp. 1–6. https://doi.org/10.1109/ICOMET.2018.8346447

Amat-Tosello, S., Dupuy, N., Kister, J., 2009. Contribution of external parameter orthogonalisation for calibration transfer in short waves - Near infrared spectroscopy application to gasoline quality. Anal. Chim. Acta 642, 6–11. https://doi.org/10.1016/j.aca.2009.01.003

Andersen, C.M., Bro, R., 2010. Variable selection in regression - a tutorial. J. Chemom. 24, 728–737. https://doi.org/10.1002/cem.1360

Anderson, J.A., 1997. An Introduction to Neural Networks, 3rd ed. MIT Press, Cambridge.

Anzanello, M.J., Fogliatto, F.S., 2014. A review of recent variable selection methods in industrial and chemometrics applications. Eur. J. Ind. Eng. 8, 619–645. https://doi.org/0.1504/EJIE.2014.065731

Arteaga, F., Ferrer, A., 2002. Dealing with missing data in MSPC: several methods, different interpretations, some examples. J. Chemom. 16, 408–418. https://doi.org/10.1002/cem.750

Balabin, R.M., Safieva, R.Z., Lomakina, E.I., 2007. Comparison of linear and nonlinear calibration models based on near

infrared (NIR) spectroscopy data for gasoline properties prediction. Chemom. Intell. Lab. Syst. 88, 183–188. https://doi.org/10.1016/j.chemolab.2007.04.006

Bao, X., Dai, L., 2009. Partial least squares with outlier detection in spectral analysis: A tool to predict gasoline properties. Fuel 88, 1216–1222. https://doi.org/10.1016/j.fuel.2008.11.025

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. CRC Press.

Cao, D.S., Xu, Q.S., Liang, Y.Z., Zhang, L.X., Li, H.D., 2010. The boosting: A new idea of building models. Chemom. Intell. Lab. Syst. 100, 1–11. https://doi.org/10.1016/j.chemolab.2009.09.002

Chéruy, A., 1997. Software sensors in bioprocess engineering. J. Biotechnol. 52, 193–199. https://doi.org/10.1016/S0168-1656(96)01644-6

Chiang, L.H., Pell, R.J., Seasholtz, M.B., 2003. Exploring process data with the use of robust outlier detection algorithms. J. Process Control 13, 437–449. https://doi.org/10.1016/S0959-1524(02)00068-9

Dayal, B.S., MacGregor, J.F., 1997. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. J. Process Control 7, 169–179. https://doi.org/10.1016/S0959-1524(97)80001-7

De Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. Chemom. Intell. Lab. Syst. 18, 251–263.

Dehmer, M., Varmuza, K., Bonchev, D., 2012. Statistical Modelling of Molecular Descriptors in QSAR/QSPR. Wiley-Blackwell.

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning, in: Multiple Classifier Systems. Springer, Berlin, pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, 3rd ed, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, New York. https://doi.org/10.1002/9781118625590

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Facco, P., Doplicher, F., Bezzo, F., Barolo, M., 2009. Moving average PLS soft sensor for online product quality estimation in an industrial batch polymerization process. J. Process Control 19, 520–529. https://doi.org/10.1016/j.jprocont.2008.05.002

Feng, R., Shen, W., Shao, H., 2003. A Soft Sensor Modeling Approach Using Support Vector Machines, in: Proceedings of the 2003 American Control Conference. IEEE, Denver, Colorado, USA, pp. 3702–3707. https://doi.org/10.1109/ACC.2003.1240410

Fortuna, L., Graziani, S., Rizzo, A., Xibilia, M.G., 2007. Soft Sensors for Monitoring and Control of Industrial Processes, 1st ed. Springer-Verlag London, London. https://doi.org/10.1007/978-1-84628-480-9

Fu, Y., Su, H., Zhang, Y., Chu, J., 2008. Adaptive Soft-sensor Modeling Algorithm Based on FCMISVM and Its Application in PX Adsorption Separation Process. Chinese J. Chem. Eng. 16, 746–751. https://doi.org/10.1016/S1004-9541(08)60150-0

Geisser, S., 1993. Predictive inference: An Introduction, 1st ed. New York: Chapman & Hall.

Geisser, S., Eddy, W.F., 1979. A Predictive Approach to Model Selection. J. Am. Stat. Assoc. 74, 153–160.

Geladi, P., 1988. Notes on the history and nature of partial least squares (PLS) modelling. J. Chemom. 2, 231–246. https://doi.org/10.1002/cem.1180020403

Geladi, P., Esbensen, K., 1991. Regression on multivariate images: Principal component regression for modeling,

prediction and visual diagnostic tools. J. Chemom. 5, 97–111. https://doi.org/10.1002/cem.1180050206

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning, 1st ed. Addison-Wesley Publishing Company, Inc.

Haaland, D.M., Thomas, E. V., 1988. Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. Anal. Chem. 60, 1193–1202. https://doi.org/10.1021/ac00162a020

Hansch, C., Hoekman, D., Gao, H., 1996. Comparative QSAR: Toward a deeper understanding of chemicobiological interactions. Chem. Rev. 96, 1045–1075. https://doi.org/10.1021/cr9400976

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed. Springer, New York.

He, K., Cheng, H., Du, W., Qian, F., 2014. Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. Chemom. Intell. Lab. Syst. 134, 79–88. https://doi.org/10.1016/j.chemolab.2014.03.007

Helland, I.S., 2001. Some theoretical aspects of partial least squares regression. Chemom. Intell. Lab. Syst. 58, 97–107. https://doi.org/10.1016/S0169-7439(01)00154-X

Helland, I.S., 1988. On the Structure of Partial Least Squares Regression. Comun. Stat. - Simul. Comput. 17, 581–607.

Hesterberg, T., Choi, N.H., Meier, L., Fraley, C., 2008. Least angle and L1 penalized regression: A review. Stat. Surv. 2, 61–93. https://doi.org/10.1214/08-SS035

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 55–67. https://doi.org/10.2307/1267351

Höskuldsson, A., 1996. Prediction Methods in Science and Technology. Thor Publishing.

Jackson, J.E., 1991. A User ' s Guide To Principal Components, 1st ed. John Wiley & Sons.

Jackson, J.E., 1980. Principal Components and Factor Analysis: Part I—Principal Components. J. Qual. Technol. 12, 201–213. https://doi.org/10.1080/00224065.1980.11980967

Jang, J.O., Chung, H.T., Jeon, G.J., 2005. Saturation and Deadzone Compensation of Systems using Neural Network and Fuzzy Logic, in: American Control Conference. IEEE, pp. 1715–1720. https://doi.org/10.1109/acc.2005.1470215

Jianxu, L., Huihe, S., 2003. Soft sensing modeling using neurofuzzy system based on Rough Set Theory, in: American. pp. 543–548. https://doi.org/10.1109/acc.2002.1024863

Jolliffe, I.T., 2002. Principal Component Analysis, 2nd ed, Springer Series in Statistics. Springer-Verlag, New York. https://doi.org/10.1007/b98835

Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven Soft Sensors in the process industry. Comput. Chem. Eng. 33, 795–814. https://doi.org/10.1016/j.compchemeng.2008.12.012

Kaneko, H., Arakawa, M., Funatsu, K., 2009. Development of a New Soft Sensor Method Using Independent Component Analysis and Partial Least Squares. AIChE J. 55, 87–98. https://doi.org/10.1002/aic.11648

Kano, M., Showchaiya, N., Hasebe, S., Hashimoto, I., 2001. Inferential control of distillation compositions: Selection of model and control configuration. IFAC Proc. Vol. 34, 347–352. https://doi.org/10.1016/S1474-6670(17)33848-X

Kardamakis, A.A., Pasadakis, N., 2010. Autoregressive modeling of near-IR spectra and MLR to predict RON values of gasolines. Fuel 89, 158–161. https://doi.org/10.1016/j.fuel.2009.08.029

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, in: 14th International Joint Conference on Artificial Intelligence. pp. 1137–1143.

Kourti, T., MacGregor, J.F., 1996. Multivariate SPC methods for process and product monitoring. J. Qual. Technol. 28, 409–428. https://doi.org/10.1080/00224065.1996.11979699

Kourti, T., MacGregor, J.F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemom. Intell. La 82, 3–21. https://doi.org/https://doi.org/10.1016/0169-7439(95)80036-9

Kresta, J. V., Macgregor, J.F., Marlin, T.E., 1991. Multivariate statistical monitoring of process operating performance. Can. J. Chem. Eng. 69, 35–47. https://doi.org/10.1002/cjce.5450690105

Krzanowski, W.J., 1988. Principles of multivariate analysis : a user's perspective. Oxford University Press, New York.

Krzanowski, W.J., 1982. Between-Group Comparison of Principal Components - Some Sampling Results. J. Stat. Comput. Simul. 15, 141–154. https://doi.org/10.1080/00949658208810577

Lababidi, H.M.S., Chedadeh, D., Riazi, M.R., Al-Qattan, A., Al-Adwani, H.A., 2011. Prediction of product quality for catalytic hydrocracking of vacuum gas oil. Fuel 90, 719–727. https://doi.org/10.1016/j.fuel.2010.09.046

Leardi, R., 2007. Genetic algorithms in chemistry. J. Chromatogr. A 1158, 226–233. https://doi.org/10.1016/j.chroma.2007.04.025

Leardi, R., 2003. Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Data handling in science and technology. Elsevier B.V.

Leardi, R., Boggia, R., Terrile, M., 1992. Genetic algorithms as a strategy for feature selection. J. Chemom. 6, 267–281. https://doi.org/10.1002/cem.1180060506

Lee, S., Choi, H., Cha, K., Chung, H., 2013. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. Microchem. J. 110, 739–748. https://doi.org/10.1016/j.microc.2013.08.007

Lin, B., Recke, B., Knudsen, J.K.H., Jørgensen, S.B., 2007. A systematic approach for soft sensor development. Comput. Chem. Eng. 31, 419–425. https://doi.org/10.1016/j.compchemeng.2006.05.030

Lindgren, F., Geladi, P., Wold, S., 1993. The Kernel algorithm for PLS. J. Chemom. 7, 45–59.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, 2nd ed. Wiley Series in Probability and Statistics, Hoboken, New Jersey.

Lu, N., Yang, Y., Gao, F., Wang, F., 2004. Multirate dynamic inferential modeling for multivariable processes. Chem. Eng. Sci. 59, 855–864. https://doi.org/10.1016/j.ces.2003.12.003

Luo, J.X., Shao, H.H., 2006. Developing soft sensors using hybrid soft computing methodology: a neurofuzzy system based on rough set theory and genetic algorithms. Soft Comput. 10, 54–60. https://doi.org/10.1007/s00500-005-0465-0

MacGregor, J.F., Kourti, T., 1995. Statistical process control of multivariate processes. Control Eng. Pract. 3, 403–414. https://doi.org/10.1016/0967-0661(95)00014-L

Macias, J., Angelov, P., Zhou, X., 2006. A Method for Predicting Quality of the Crude Oil Distillation, in: 2006 International Symposium on Evolving Fuzzy Systems. IEEE, Ambleside, pp. 214–220. https://doi.org/10.1109/ISEFS.2006.251167

Martens, H., Naes, T., 1989. Multivariate Calibration. Wiley, Chichester UK.

McAvoy, T.J., Wang, N.S., Naidu, S., Bhat, N., Hunter, J., Simmons, M., 1989. Interpreting Biosensor Data via Backpropagation, in: International 1989 Joint Conference on Neural Networks. Washington, DC, pp. 227–233.

Mendes, G., Aleme, H.G., Barbeira, P.J.S., 2012. Determination of octane numbers in gasoline by distillation curves and partial least squares regression. Fuel 97, 131–136. https://doi.org/10.1016/j.fuel.2012.01.058

Moghadassi, A., Beheshti, A., Parvizian, F., 2016. Prediction of research octane number in catalytic naphtha reforming unit of Shazand Oil Refinery. Int. J. Ind. Syst. Eng. 23, 435. https://doi.org/10.1504/IJISE.2016.077696

Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to Linear Regression Analysis, 5th ed. New Jersey.

Montgomery, D.C., Runger, G.C., 2003. Applied Statistics and Probability for Engineers, 3rd ed. John Wiley & Sons, New York.

Murphy, K.P., 2012. Machine Learning - A Probabilistic Perspective, Chance Encounters: Probability in Education. MIT Press, Cambridge, MA. https://doi.org/10.1007/978-94-011-3532-0_2

Murtaugh, P.A., 1998. Methods of variable selection in regression modeling. Commun. Stat. - Simul. Comput. 27, 711–734. https://doi.org/10.1080/03610919808813505

Naes, T., Isakson, T., Fearn, T., Davies, T., 2004. A user-friendly guide to Multivariate Calibration and Classification. NIR Publications, Chichester UK.

Negiz, A., Çinar, A., 1998. Monitoring of multivariable dynamic processes and sensor auditing. J. Process Control 8, 375–380. https://doi.org/10.1016/s1474-6670(17)43139-9

Nelson, P.R.C., Taylor, P.A., Macgregor, J.F., 1996. Missing data methods in PCA and PLS: Score calculations with incomplete observations. Chemometfics Intell. Lab. Syst. 35, 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

Nomikos, P., MacGregor, J.F., 1995. Multivariate SPC charts for monitoring batch processes. Technometrics 37, 41–59. https://doi.org/10.1080/00401706.1995.10485888

Park, S., Han, C., 2000. A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns. Comput. Chem. Eng. 24, 871–877. https://doi.org/10.1016/S0098-1354(00)00343-4

Pearson, R.K., 2002. Outliers in process modeling and identification. IEEE Trans. Control Syst. Technol. 10, 55–63. https://doi.org/10.1109/87.974338

Qin, J.S., 1998. Recursive PLS algorithms for adaptive data modeling. Comput. Chem. Eng. 22, 503–514. https://doi.org/10.1016/S0098-1354(97)00262-7

Qin, S.J., 1997. Neural Networks for Intelligent Sensors and Control - Practical Issues and Some Solutions, in: Omidvar, O., Elliott, D.L. (Eds.), Neural Systems for Control. Elsevier, pp. 213–234. https://doi.org/10.1016/B978-012526430-3/50009-X

Rato, T.J., Reis, M.S., 2017. Multiresolution Soft Sensors: A New Class of Model Structures for Handling Multiresolution Data. Ind. Eng. Chem. Res. 56, 3640–3654. https://doi.org/10.1021/acs.iecr.6b04349

Reis, M.S., Rendall, R., Chin, S.T., Chiang, L., 2015. Challenges in the Specification and Integration of Measurement Uncertainty in the Development of Data-Driven Models for the Chemical Processing Industry. Ind. Eng. Chem. Res. 54, 9159–9177. https://doi.org/10.1021/ie504577d

Rendall, R., Reis, M.S., 2018. Which regression method to use? Making informed decisions in "data-rich/knowledge poor" scenarios – The Predictive Analytics Comparison framework (PAC). Chemom. Intell. Lab. Syst. 181, 52–63. https://doi.org/10.1016/j.chemolab.2018.08.004

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning Internal Representations by Error Propagation, in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations. MIT Press, Cambridge, MA, pp. 319–362.

Sadighi, S., Mohaddecy, R.S., Norouzian, A., 2015. Optimizing an Industrial Scale Naphtha Catalytic Reforming Plant Using a Hybrid Artificial Neural Network and Genetic Algorithm Technique. Bull. Chem. React. Eng. Catal. 10, 210–220. https://doi.org/10.9767/bcrec.10.2.7171.210-220

Scheffer, J., 2002. Dealing With Missing Data. Res. Lett. Inf. Math. Sci. 3, 153–160.

Seborg, D.E., Edgar, T.F., Mellichamp, D.A., 2011. Process Dynamics and Control, 3rd ed. John Wiley & Sons.

Shakil, M., Elshafei, M., Habib, M.A., Maleki, F., 2009. Soft Sensor for NOx and O2 Using Dynamical Neural Network. Comput. Electr. Eng. 35, 578–586. https://doi.org/10.1016/j.compeleceng.2008.08.007

Sofge, D.A., 2002. Using Genetic Algorithm Based Variable Selection to Improve Neural Network Models for Real-World Systems, in: Proceedings of the 2002 International Conference on Machine Learning and Applications. pp. 16–19.

Souza, F.A.A., Araújo, R., Mendes, J., 2016. Review of soft sensor methods for regression applications. Chemom. Intell. Lab. Syst. 152, 69–79. https://doi.org/10.1016/j.chemolab.2015.12.011

Stone, M., 1974. Cross-validatory Choice and Assessment of Statistical Predictions. J. R. Stat. Soc. Ser. B 36, 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Strobl, C., Malley, J., Gerhard T, 2009. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B (Statistical Methodol. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Venkatasubramanian, V., Vaidyanathan, R., Yamamoto, Y., 1990. Process fault detection and diagnosis using neural networks - I. steady-state processes. Comput. Chem. Eng. 14, 699–712. https://doi.org/10.1016/0098-1354(90)87081-Y

Vezvaei, H., Ordibeheshti, S., Ardjmand, M., 2011. Soft-Sensor for Estimation of Gasoline Octane Number in Platforming Processes with Adaptive Neuro-Fuzzy Inference Systems (ANFIS). Int. J. Chem. Mol. Nucl. Mater. Metall. Eng. 5, 913–917. https://doi.org/10.5281/zenodo.1058211

Voigt, M., Legner, R., Haefner, S., Friesen, A., Wirtz, A., Jaeger, M., 2019. Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field1H NMR@80 MHz, handheld RAMAN and benchtop NIR. Fuel 236, 829–835. https://doi.org/10.1016/j.fuel.2018.09.006

Walczak, B., Massart, D.L., 2001. Dealing with missing data: Part I. Chemom. Intell. Lab. Syst. 58, 29–42. https://doi.org/10.1016/S0169-7439(01)00131-9

Walczak, B, Massart, D.L., 2001. Dealing with missing data: Part I. Chemom. Intell. Lab. Syst. 58, 15–27. https://doi.org/10.1016/S0169-7439(01)00131-9

Wang, L.X., Mendel, J.M., 1992. Fuzzy Basis Functions, Universal Approximation, and Orthogonal Least-Squares Learning. IEEE Trans. Neural Networks 3, 807–814. https://doi.org/10.1109/72.159070

Willis, M.J., Di Massimo, C., Montague, G.A., Tham, M.T., Morris, A.J., 1991. Artificial neural networks in process engineering. IEE Proc. D Control Theory Appl. 138, 256–266. https://doi.org/10.1049/ip-d.1991.0036

Wold, S., 1978. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics 20, 397–405. https://doi.org/10.1080/00401706.1978.10489693

Wold, S., 1976. Pattern recognition by means of disjoint principal components models. Pattern Recognit. 8, 127–139. https://doi.org/10.1016/0031-3203(76)90014-5

Wold, S., Esbensen, K., Geladi, P., 1987. Principal Component Analysis. Chemom. Intell. Lab. Syst. 2, 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wold, S., Ruhe, A., Wold, H., Dunn, W.J., 1984. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. J. Sci. Stat. Comput. 5, 735–743. https://doi.org/10.1137/0905052

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wu, Y., Luo, X., 2010. A novel calibration approach of soft sensor based on multirate data fusion technology. J. Process Control 20, 1252–1260. https://doi.org/10.1016/j.jprocont.2010.09.003

Yan, X., 2008. Modified nonlinear generalized ridge regression and its application to develop naphtha cut point soft sensor. Comput. Chem. Eng. 32, 608–621. https://doi.org/10.1016/j.compchemeng.2007.04.011

Yang, S.H., Chen, B.H., Wang, X.Z., 2000. Neural network based fault diagnosis using unmeasurable inputs. Eng. Appl. Artif. Intell. 13, 345–356. https://doi.org/10.1016/S0952-1976(00)00005-1

Zamprogna, E., Barolo, M., Seborg, D.E., 2004. Estimating product composition profiles in batch distillation via partial least squares regression. Control Eng. Pract. 12, 917–929. https://doi.org/10.1016/j.conengprac.2003.11.005

Zeng, X.J., Singh, M.G., 1995. Approximation Theory of Fuzzy Systems—MIMO Case. IEEE Trans. Fuzzy Syst. 3, 219–235. https://doi.org/10.1109/91.388175

Zhang, Z., 2016. Variable selection with stepwise and best subset approaches. Ann. Transl. Med. 4, 1–6. https://doi.org/10.21037/atm.2016.03.35

Zhou, C., Liu, Q., Huang, D., Zhang, J., 2012. Inferential estimation of kerosene dry point in refineries with varying crudes. J. Process Control 22, 1122–1126. https://doi.org/10.1016/j.jprocont.2012.03.011

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodol. 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

Table A.1 Hyper-parameters for each method used during the model training phase. The strategy for selecting the methods' hyper-parameters is based on a 10-fold cross-validation.

Method	Hyper-parameter(s)	Possible Values(s)	Selection Strategy
MLR	-	-	-
FSR	p_{enter} p_{out}	0.05 0.1	-
RR	α γ	0 [0.001 0.01 0.1 1 10]	10-fold cross-validation
LASSO	α γ	1 [0.001 0.01 0.1 1 10]	10-fold cross-validation
EN	α γ	[0 0.1667 0.333 0.500 0.6667 0.8333 1.000] [0.001 0.01 0.1 1 10]	10-fold cross-validation
PCR	α_{PCR}	min(20, *n*, *p*)	10-fold cross-validation
PCR_FS	p_{enter} p_{out}	0.05 0.1	-
PLS	α_{PLS}	min(20, *n*, *p*)	10-fold cross-validation
BRT	T_{BRT}	[50 100 500 1000 5000]	10-fold cross-validation
RF	T_{RF}	[50 100 500 1000 5000]	10-fold cross-validation
BT	T_{BT}	[50 100 500 1000 5000]	10-fold cross-validation

Abbreviations: MLR - Multiple Linear Regression; FSR - Forward Stepwise Regression; RR - Ridge Regression; LASSO - Least Absolute Shrinkage and Selection Operator; EN - Elastic Net; PCR - Principal Component Regression; PCR_FS - Principal Component Regression with scores added in a Forward Stepwise fashion; PLS - Partial Least Squares; BRT - Bagging of Regression Trees; RF - Random Forests; BT - Boosting of Regression Trees.