Noname manuscript No.
(will be inserted by the editor)

Improved Quantification of Important Beer Quality Parameters based on Non-linear Calibration Methods applied to FT-MIR Spectra

Carlos Cernuda · Edwin Lughofer* · Helmut Klein · Clemens Forster · Marcin Pawliczek · Markus Brandstetter

Received: 29.04.2016

Abstract During the production process of beer, it is of utmost importance to guarantee a high consistency of the beer quality. For instance, the bitterness is an essential quality parameter which has to be controlled within the specifications already at the beginning of the production process in the unfermented beer (wort) as well as in final products such as beer and beer mix beverages. Nowadays, analytical techniques for quality control in beer production are mainly based on manual supervision, i.e. samples are taken from the process and analyzed in the laboratory. This typically requires significant lab technicians efforts for only a small fraction of samples to be analyzed, which leads to significant costs for beer breweries and companies. Fourier transform mid-infrared (FT-MIR) spectroscopy was used in combination with non-linear multivariate calibration techniques to overcome (i) the time consuming off-line analyses in beer production and (ii) already known limitations of standard linear chemometric methods **, like partial least squares (PLS),** for important quality parameters [1][2] such as bitterness, citric acid, total acids, free amino nitrogen, final attenuation or foam stability. The calibration models are established with enhanced non-linear techniques based (i) on a new *piece-wise linear version of PLS* by employing fuzzy rules for local partitioning the latent variable space and (ii) on extensions of *support vector regression variants* (ε-PLSSVR and ν-PLSSVR), for overcoming high computation times in high-dimensional problems and time-intensive and inappropriate settings of the kernel parameters. Furthermore, we introduce a *new model selection scheme* based on bagged ensembles in order to improve robustness and thus predictive quality of the final models. The approaches are tested on real-world calibration data sets for wort and beer mix beverages, and successfully compared to linear methods, as showing a clear out-performance in most cases and being able to meet the model quality requirements defined by the experts at the beer company.

Keywords quality control of beer · FT-MIR spectroscopy · non-linear PLS · flexible fuzzy systems · support vector regression variation · bagged model selection

C. Cernuda: BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
Department of Knowledge-Based Mathematical Systems, Johannes Kepler University Linz, Austria
E. Lughofer*: Department of Knowledge-Based Mathematical Systems, Johannes Kepler University Linz, Austria
H. Klein and C. Forster: BrauUnion GmbH, Poschacherstrasse 35, 4020 Linz, Austria
M. Pawliczek and M. Brandstetter: RECENDT GmbH, Linz, Austria
*Corresponding Author Email: edwin.lughofer@jku.at

1 Introduction

1.1 Motivation and State-of-the-Art

A high quality of beer and its spin-offs (e.g., beer mix beverages), one of the most heavily consumed beverages in the world and the typical 'national drink' in Middle European countries, is of great importance in order to satisfy the consumers and the whole alcoholic drink market. For instance, the bitterness is an essential quality parameter which has to be controlled within the specifications already at the beginning of the production process in the unfermented beer (wort) as well as in final products such as beer and beer mix beverages [3]: it is the key parameter for achieving a certain taste the beer should have in order to fall within the common classification boundaries [4]. A high quality can be only guaranteed by permanent supervision of the liquid during its production.

By the application of an analytical method on the basis of FT-MIR spectroscopy in combination with suitable chemometric methods it is possible to significantly reduce time consuming laboratory analysis. Instead of measuring the relevant quality parameters — such as bitterness, free amino nitrogen, final attenuation, citric acid, total acid and foam stability — with six different analytical methods sequentially, it is possible to have all quality parameters simultaneously analyzed in less than 15 minutes in case of 10 samples drawn from the liquid after production. In comparison, a manual analysis of the most relevant parameters, namely bitterness, final attenuation and free amino nitrogen requires operators efforts of about four hours and an overall duration for final attenuation of about 24 hours in sum. This usually causes significant costs for beer breweries and companies.

Current analytical methods for quality control of beer rely on time and resource consuming chemical analysis in the laboratory where for each quality parameter an individual method and equipment is needed. Recently, spectroscopic methods are being developed in order to determine relevant quality parameters simultaneously in much shorter time and strongly reduced effort for sample preparation [5] [6]. In this context, chemometric methods are employed to gain mathematical models for quantification of the analytes [7] and process parameters [8]. Currently, most of these approaches and resulting models are based on linear calibration methods (not being able to resolve any non-linearities contained in the production process adequately with sufficient accuracy), mainly on the basis of partial least squares regression [9] and especially without the usage of robust model selection techniques. A non-linear approach can be found in [2] where neural networks have been used for predicting the content of acetic acid, however it does not address important beer parameters for the end consumer such as bitterness, final attenuation or foam; moreover, no robust model selection strategies are embedded for appropriately addressing calibration problems based on a very low number of samples.

1.2 Our Approach

Our approach aims on compensating current shortcomings in beer quality analysis and goes significantly beyond state-of-the-art in terms of the following aspects:

- It enables the fully automatic quantification of several important beer parameters in wort as well as in the final products (beer, beer mix beverages), such as bitterness, final attenuation, free amino nitrogen, citric acid, total acid and foam stability.

- It employs a FT-MIR spectrometer equipped with an automatic sampler for the purpose to draw samples from probes and to overcome the time consuming off-line analyses in beer production.
- It applies enhanced non-linear calibration modeling methods to overcome already known limitations of standard linear chemometric methods (such as PLS) for the essential parameters mentioned above: one is based on a variation of support vector regression (SVR), the other one on a batch version of *Gen-Smart-EFS* **(short for *Generalized Smart Evolving Fuzzy Systems*)** for extracting generalized fuzzy rules in a fast single-pass manner; it is coupled with PLS in order to achieve a kind of a piece-wise linear (thus overall non-linear) version of PLS with fuzzy transitions.
- It embeds a new, robust model selection scheme based on bagged model ensembles which are constructed from multiple bags; the selection is carried out on a bunch of possible model candidates, which are obtained due to various learning parameter combinations (parameter grid) used in SVR and fuzzy rule base extraction. **The bagged model selection scheme has been mainly motivated due to the availability of a very low number of samples for calibration, as bagging explores a sparse sample space in a nice way, thus increasing robustness of calibration [10]**.

We evaluate our approach on three data sets, two drawn from wort and one from beer mix beverages production. Thereby, we report on both, the cross-validation (CV) error as well as on the error on a separate validation set. Results show that there is a clear improvement in CV errors over classical linear state-of-the-art methods when applying enhanced non-linear techniques for most of the targets achieving finally errors within the limits of the company's requirements, whereas this can be also confirmed for the separate validation data in case of beer mix beverages. For beer mix beverages data, the application of bagging for model selection brings much improvement for providing robust models on separate validation data with lower over-fitting proneness in case of bitterness and foam stability (the two most essential parameters); in fact, without bagging no useful results within the acceptable error ranges could be achieved for these two parameters.

2 The Setup

2.1 Data Acquisition

Spectroscopic data of the wort and beer mix beverages samples were acquired off-line using a Nicolet iZ10 FT-IR-spectrometer with CETAC ASX-520 autosampler. Besides the spectrometer core, this instrument contains a programmable logic controller (PLC) and an engine for the evaluation of chemometric models. The optics of the spectrometer contains a monolithic Michelson interferometer, which helps with temperature stability [11]. The resolution and the measurement rate of the instrument are configurable.

For mid infrared spectroscopic measurement, we used a transmission flow cell with an optical path length of 15μm or 20μm and CaF2 windows. Data collection was set to 16 scans per sample with a resolution of 4cm^{-1} in the spectral range 400 to 4000cm^{-1}. Absorbance spectra of the investigated samples were calculated according to Beer's law [12] using pure water for recording the background single beam spectra. A schematical sketch of the measurement setup is shown in Figure 1.

Wort samples were filtered with Kieselgur and beer samples were degassed by ultrasonic treatment. In order to obtain appropriate and reproducible spectra, the filling of the flow cell was optimized by rinsing the flow cell between the replica of the sample and back-washing with deionized water after each sample. Reference data for the target parameters to be super-

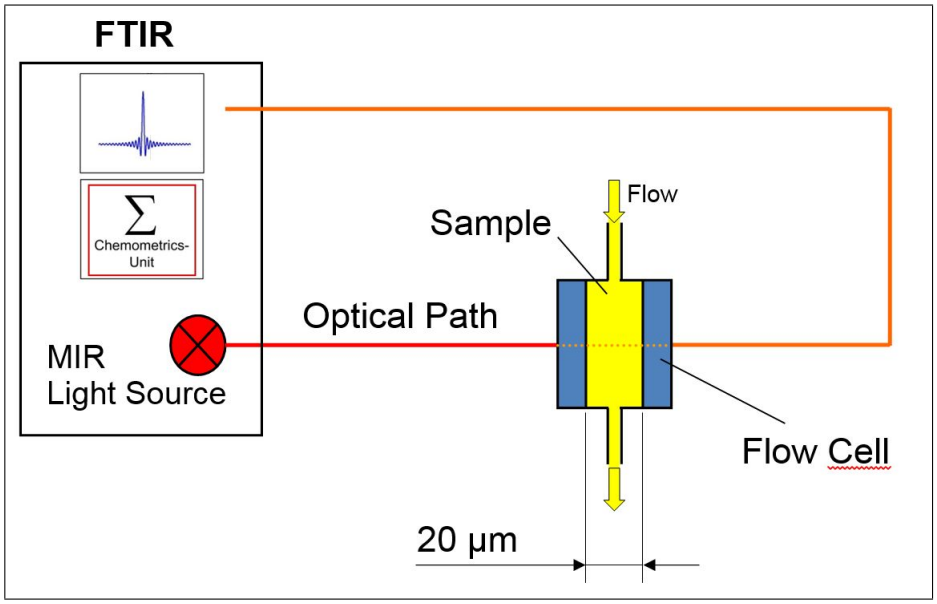


Fig. 1: Schematic view of the data acquisition framework as installed at beer production systems

vised (bitterness, final attenuation, free amino nitrogen, types of acidities and foam stability) have been obtained by manual analysis of wort and beer mix beverages probes taken at random from the production process. Using the reference data and the corresponding FT-MIR absorption spectra the (non-linear) chemometric models could be established, see Section 5.

3 Non-Linear Calibration Methods

3.1 Non-Linear PLS with the Usage of Flexible Fuzzy Inference Systems

Classical PLS Partial least squares regression [13] is one of the most widely used calibration method in today's chemometric modeling tasks and applications [14] [15] [16]. The core concept of PLS is the transformation of the original input feature space — in case of Chemometrics, it is typically the space spanned by the wavelengths [17] or at least partial connected pieces in form of wavebands [18] contained in the spectra — into a reduced input space for best explaining the variance contained in the target (which is typically a continuous numerical output when dealing with regression problems). Partial least squares is used to find the fundamental relations between two matrices (input $\mathbf{X}$ and output $\mathbf{Y}$), i.e. a latent variable approach to model the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the input space that explains the maximum multidimensional variance direction in the output space. It emphasizes a rotation of the input space in order to best explain (the variance of the) target by means of linear relations/mappings. The resulting transformed space is characterized by eigenvectors which forms the so-called *latent variable space*. Their corresponding eigenvalues can be sorted in descending order in order to achieve a ranking of latent variables based on which a selected subset due

to a variance-explained cut-off is typically used for model calibration (as, e.g., typically used within the well-known and widely-used PLS-Toolbox [1]).

Non-Linear PLS (Variants) through Kernel Transformation **Even though PLS respects the target concept during space transformation, it is still a linear method, i.e. it emphasizes rotations to best represent covariance structures in the data in a linear sense. In order to establish a non-linear variant of PLS for emphasizing the best variance explanation in a non-linear sense, the *kernel-based PLS* (K-PLS) is typically employed [19]. Firstly, it applies the kernel trick (in the same way as done in support vector regression, see below) in order to perform a non-linear transformation of the original data set into an S-dimensional feature space. Secondly, it performs the conventional partial least squares algorithm on the transformed kernel Gram matrices K1 (for the input space) and K2 (for the output space), with entries $\mathbf{K1}_{ij} = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{K2}_{ij} = \mathbf{K}(\mathbf{y}_i, \mathbf{y}_j)$ and K the multi-dimensional kernel and $\mathbf{x}_i, \mathbf{x}_j$ the input vectors of the i-th and j-th sample (so, the kernel function is applied to all sample pairs).** The disadvantage of K-PLS becomes immediate when the number of samples N available for regression is large, because then the kernel Gram matrix from input space explodes in size ($N \times N$).

Our Non-Linear PLS Version In this paper, we abandon the disadvantage of non-linear K-PLS and apply a non-linear version recently introduced in [20] and successfully applied for establishing calibration models from **Fourier transform near-infrared (FT-NIR)** spectra in melamine resin production (for cloud point prediction and supervision). Its basic idea lies in the partitioning of the latent variable space (after transformation with classical PLS) into several (C) local pieces represented by fuzzy rules, i.e. dividing it into different partial principal component directions along the target. Each fuzzy rule embeds a local linear hyper-plane for local trend estimation, thus it results in piece-wise local linear PLS predictors, which are combined through a weighted linear combination, where the weights are rule activation levels in form of multi-dimensional Gaussian kernels. This assures smoothness of the whole regression surface as the piece-wise linear predictors are 'kernel-smoothened' across their transitions [21].

Our Fuzzy Rules Learning Engine Our engine for extracting the appropriate number and positioning of the fuzzy rules from data acts in a single-pass manner directly in the PLS space, i.e. each single sample taken from the calibration set is first transformed to the latent variable space due to the loadings and then sent into the fuzzy rule learning process. Single-pass capability assures very fast learning speeds of the whole fuzzy systems, as the rule base grow and the parameters are recursively updated based on single samples (loaded one-by-one into the memory), leading to a method whose computational complexity and virtual memory requirement is linear with the number of samples in the calibration set. This makes it very attractive for calibrating models over larger parameter grids within time-intensive cross-validation procedures, see Section 3.3. In particular, our learning engine is based on the *Gen-Smart-EFS* approach [22], whose core functionality (without the concepts regarding rule merging and dynamic feature weighting for dimension reduction) is used to find the appropriate number of rules in single-pass evolution steps and also to estimate the kernel functions forming the antecedents of the rules; in this way, each rule antecedent is associated with a triplet $(\mathbf{c}, \Sigma^{-1}, \mathbf{r})$ with $\mathbf{c}$ its center, Σ^{-1} the inverse covariance matrix defining its multivariate ellipsoidal shape and $\mathbf{r}$ its tolerance radius (statistical range of influence also termed as prediction

[1] http://www.eigenvector.com/software/pls_toolbox.htm

interval [23]), which is automatically extracted from data and steers rule evolution versus rule update, see below. In this sense, one fuzzy rule reads as

$$\text{IF } \mathbf{x} \text{ IS (about) } \mu_i \text{ THEN } l_i(\mathbf{x}) = w_{i0} + w_{i1}x_1 + w_{i2}x_2 + ... + w_{ip}x_p \tag{1}$$

with $\mu_i = exp(-\frac{1}{2}(\mathbf{x}-\mathbf{c_i})^T\Sigma_i^{-1}(\mathbf{x}-\mathbf{c_i}))$ denoting the multivariate Gaussian distribution.

The single-pass rule evolution and antecedent learning steps are as follows (with $C = 0$ initially):

1. Load a new sample $\mathbf{x}$; if it is the first one, Goto Step 5 (there, ignoring the if-part);
2. Elicit the winning rule, i.e. the rule closest to the current sample, which is then denoted as $\mathbf{c_{win}}$; for the distance calculation, standard Mahalanobis distance is used [24] (as on the right hand side in (2) below).
3. Check whether the following criterion is met (the *rule evolution criterion*):

$$min_{i=1,...,C}\sqrt{(\mathbf{x}-\mathbf{c_i})^T\Sigma^{-1}(\mathbf{x}-\mathbf{c_i})} > r_i \qquad r_i = vigi * p^{1/\sqrt{2}} * \frac{1.0}{(1-1/(k_i+1))^m} \tag{2}$$

   with p the dimensionality of the input feature space and *vigi* an a priori defined parameter, steering the tradeoff between stability (update of an old cluster) and plasticity (evolution of a new cluster); k_i the support of the ith rule and m a tuning parameter per default set to 4. This is the only sensitive parameter and is varied during the model evaluation phase, see Section 4.3 — for further explanation of this criterion, please refer to [22].
4. **If** (2) **is not met**, the centre of the winning rule is updated by

$$\mathbf{c_{win}}(N+1) = \mathbf{c_{win}}(N) + \eta_{win}(\mathbf{x} - \mathbf{c_{win}}(N)) \tag{3}$$

   and its inverse covariance matrix by (the index *win* neglected due to transparency reasons):

$$\Sigma^{-1}(k+1) = \frac{\Sigma^{-1}(k)}{1-\alpha} - \frac{\alpha}{1-\alpha}\frac{(\Sigma^{-1}(k)(\mathbf{x}-\mathbf{c}))(\Sigma^{-1}(k)(\mathbf{x}-\mathbf{c}))^T}{1+\alpha((\mathbf{x}-\mathbf{c})^T\Sigma^{-1}(k)(\mathbf{x}-\mathbf{c}))} \tag{4}$$

   with N the number of samples seen so far and $\alpha = \frac{1}{k_{win}+1}$ with k_{win} the number of samples seen so far for which $\mathbf{c_{win}}$ has been the winning rule (cluster). The former stems from the idea in vector quantification [25] by minimizing the expected squared quantization error; the learning gain η_{win} is thereby set in a way that it fulfills the Robbins-Monroe conditions. The latter is a recursive exact update without requiring the original covariance matrix, which is analytically derived with the usage of the Neumann series, see [26] for full details.
5. **If** (2) **is met**, a new rule is evolved as covering a new region in the feature space (i.e. having sufficient *novelty content*) by setting its center $\mathbf{c}_{C+1}$ to the coordinates of $\mathbf{x}$ and initialize its inverse covariance matrix Σ_{win}^{-1} by setting it to a diagonal matrix with entries 1 divided by a small fraction, i.e. 1/100, of the variable ranges (= initial rule spreads); increase the number of rules $C = C+1$.
6. If there have not yet been all samples in the calibration set processed, Goto Step 1, otherwise Stop.

Once these are formed, the consequent parameters $l_1, ..., l_C$ for all C rules are estimated through fuzzily weighted least squares [27] in order to assure local learning which has several advantages over global learning, see [21], Chapter 2 for a detailed analysis. **A block diagram summarizing the procedure can be seen in Figure 2**

A special case comes up when the inverse covariance matrix Σ^{-1} is used as a diagonal matrix (ignoring the co-variances between the inputs). Then, axis-parallel fuzzy rules are triggered and the steps in the itemization above end up in the classical ***flexible fuzzy inference systems (FLEXFIS)*** approach [28].

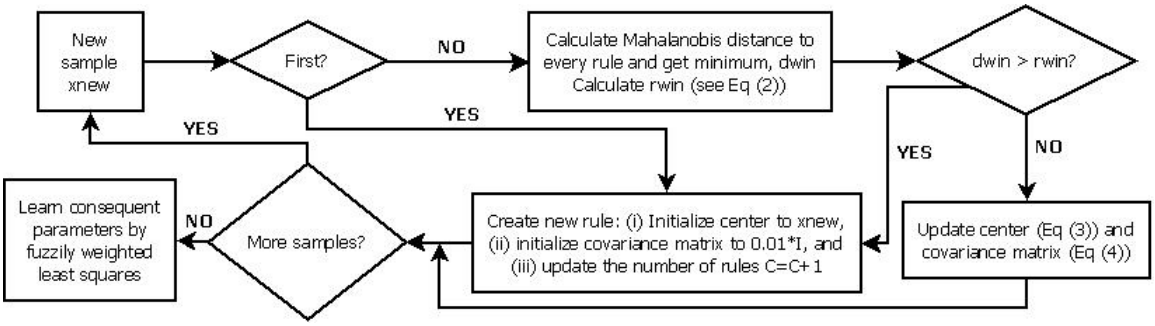


Fig. 2: Block diagram summarizing the fuzzy rules learning engine

3.2 Support Vector Regression Variation

Support vector machines **(SVM)** [29] [30] is a well known non-linear classification method, based on the calculation of hyper-planes in the input feature space to separate classes with maximal margin. The samples closest to the decision boundary, i.e. defining the positioning of the hyper-planes are called the *support vectors*. It employs the kernel trick [31] for performing a non-linear transformation of the original data into a linearized space, where then the conventional linearized concept of separating hyper-planes with margin maximization can be again applied. There is a regression version, support vector regression (SVR) [32], with two variants called ε-SVR, and ν-SVR. The general principle behind SVR is the following: a mapping ϕ maps the data $\mathbf{X}$ to a m-dimensional feature space, where a linear model is generated

$$f(\mathbf{X},\omega) = \sum_{j=1}^{m} \omega_j \phi_j(\mathbf{X}) + b \tag{5}$$

where b is the bias term (null for centered data), ϕ_j are the non-linear transformations, and ω_j are the model coefficients. The quality of the estimation is then measured by an ε-insensitive loss function, meaning that any loss below ε is neglected.

Finally, the coefficients calculation depend on the two variants of SVR. For ε-SVR, the coefficients are the solution of the quadratic problem

$$\begin{aligned} &min\ \tfrac{1}{2}\|\omega\|^2 + \tfrac{C}{n}\textstyle\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &s.t.\ \begin{cases} y_i - \omega^T x_i - b \le \varepsilon + \xi_i \\ \omega^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i,\ \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{6}$$

where n is the number of inputs, x_i the inputs, y_i the targets, C is the cost parameter, and ξ_i and ξ_i^* are slack variables.

For ν-SVR, the coefficients are the solution of the quadratic problem

$$\begin{aligned} &min\ \tfrac{1}{2}\|\omega\|^2 + C\nu\varepsilon + \tfrac{C}{n}\textstyle\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &s.t.\ \begin{cases} y_i - \omega^T x_i - b \le \varepsilon + \xi_i \\ \omega^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i,\ \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{7}$$

The cost parameter C controls how relevant is to fall out of the ε-insensitivity zone, and the parameter ν is used to bound the noise. Indeed, ν is an upper bound on the fraction of errors, and a lower bound on the number of support vectors. Figure

Online Resource 1 shows an example of the ε-insensitive areas (in between the dotted lines), and the loss for the only two points outside. Therefore, the total loss is the sum of the loss for those two points in this case.

There are plenty of kernel functions that could be used as non-linear transformations. The most commonly used one is the Gaussian kernel, given by

$$K(u,v) = e^{-\gamma|u-v|^2} \tag{8}$$

where γ is the spread of the Gaussian function. The most interesting parameters to be tuned are C and γ in both SVR approaches.

A drawback for SVR is the computational cost for high-dimensional data sets. Therefore we propose a variant for both ε-SVR, and ν-SVR. It consists on including a previous step, in which the goal is to reduce the dimensionality in advance by compressing the data by means of PLS. We denote these variants by ε-PLSSVR, and ν-PLSSVR. Therefore a new parameter arises, that is the number of latent variables to be used.

3.3 Intervened Non-Linear Modeling and Evaluation Scheme (for all Methods)

Assuming to have N calibration samples available (drawn by the spectroscopic equipment as described in Section 2.1), our modelling procedure together with the full evaluation performs the following steps:

1. Calculate latent variables $lat_1, ..., lat_{all}$ with all the number of wavelengths contained in the spectra, and ordered according to their importance. Notice that our approach includes always a previous data compressing step by means of PLS. Therefore, the latent variables from PLS are always needed.
2. Define parameter grid: Parameter selection for PLS, FLEXFIS-PLS, and PLSSVR is based on a grid search including a cross validation procedure. There are two parameter selection approaches: the classical CV selection based on the minimum CVRMSE (=cross-validated root mean squared error), and a robust model selection based on bagging CV, see Section 3.4. The parameter grids are different for each of the algorithms:
   - For PLS, use $dim = \{1, ..., a\}$ for the number of latent variables to be included into the calibration model, achieving a vector of grid points $\mathbf{g}_i = dim_i$.
   - For ridge regression, use different regularization parameters λ, with grid points $\mathbf{g}_i = \lambda_i$.
   - For generalized linear models with elastic net **(GLMNet)**, use the coefficient α that controls the convex combination between Lasso and ridge regression, and the regularization parameter λ. **This results in a matrix of grid points $\mathbf{G}_{ij} = (\alpha_i, \lambda_j)$.** See Section 4.2 for further details on GLMNet.
   - For fuzzy systems, use the number of latent variables dim_i and define the vigilance parameter $vigi$ inside the interval $(0,1)$ that steers the rule evolution criterion in (2) and thus controls the level of non-linearity applied [28]. This results in a matrix of grid points $\mathbf{G}_{ij} = (dim_i, vigi_j)$.
   - For the SVR approaches, the number of latent variables is fixed, taken from the applied model selection performed for PLS. The parameters to be tuned are the cost C and the width of the Gaussian kernel function γ. The matrix of grid points is $\mathbf{G}_{ij} = (C_i, \gamma_j)$, default grid suggested by the authors of the guidance for using Lib-SVM [33], the most widely-used library for SVM.
3. For all grid points, perform 10-fold cross-validation [34], in both the classical and the bagged versions, and store the cross-validation error: $\mathbf{CVerr}_i$ respectively $\mathbf{CVerr}_{ij}$. See 3.4 for further details on the bagged version.

4. Perform model selection. Complexity is measured in different ways in each of the considered algorithms, existing in some cases a relationship between the complexity and the parameters in the grids. For instance, it increases with the number of latent variables in PLS. For fuzzy systems learning coupled with PLS, it increases in a direct way with dimensionality and in an inverse way with vigilance because the lower the vigilance the higher the number of rules (as (2) is more often fulfilled). For SVR, the complexity can be measured in terms of the number of support vectors. Thus, the higher the number, the higher the complexity. There is a direct relationship with the cost C, because a high cost means a high penalization for non-separable points, thus higher number of support vectors would be stored in order to diminish the number of non-separable points. There is also a relation between the complexity and the width γ, as a higher value induces a lower kernel width, i.e. steeper surfaces and thus a higher non-linearity. Then, our model selection procedure selects the parameters corresponding to the grid point for which the corresponding model has lower CVRMSE, after being penalized according to their complexity ($\mathbf{CVerr}(pen)$):

$$\mathbf{CVerr}_{ij}(pen) = \mathbf{CVerr}_{ij} \cdot e^{\alpha\ param1_i + \beta\ (1 - param2_j)} \tag{9}$$

with **param1** related to dimensionality in case of fuzzy modelling and to cost in case of SVR, and **param2** to the vigilance in case of fuzzy modeling and to γ in case of SVR. α and β are normalization factors which are set to 0.05 in our case, 0.5 respectively.

5. Perform a final model training on the whole training set with the obtained optimum parameters $(param1_i^*, param2_j^*)$, and test it on a separate validation set (if available).

A block diagram summarizing the procedure can be seen in Figure 3

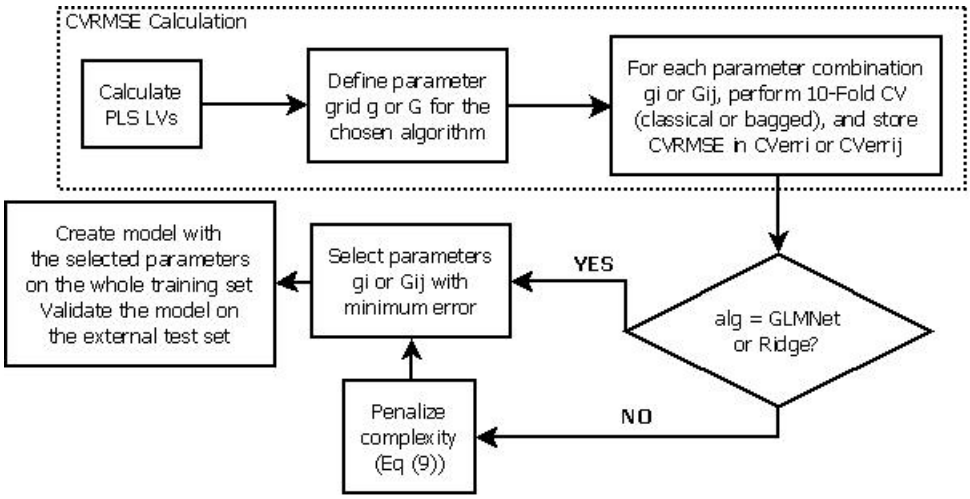


Fig. 3: Block diagram summarizing the standard model selection approach

3.4 Bagged Selection for Increasing Robustness of Non-Linear Modelling

Bagging [35] stands for bootstrap aggregating. The basic idea behind is the creation of M bags of N' training samples each by means of sampling with replacement (bootstrap sampling [36]). The bagged algorithm is performed for all M bags, and the M outputs are aggregated according to certain aggregation function, depending on the algorithm. Theoretically the diversity brought by the bootstrap sampling lead to M models that are not necessarily good, but lead to a good final aggregated model. Notice that the usual size N' of the bags coincide with the number of available samples N. In that concrete case, the expected percentage of unique samples in each bag is 63.2% [37].

We use bagging for the specific purpose of model selection, thus including the following steps:

1. Create M bags with N samples in each bag.
2. For the k-th bag, perform the classical cross validation for all parameters combinations in the grid, depending on the regression approach under consideration. Store the errors $\mathbf{CVerr}_{ij}^{k}$ for the parameters $(param1_i, param2_j)$.
3. Aggregate all k cross validation errors using the average as aggregation function.
4. Penalize the errors according to the complexity, using equation (9).
5. Select the optimum parameters $(param1_i^*, param2_j^*)$, for which the penalized error is minimum.

A block diagram summarizing the procedure can be seen in Figure 4

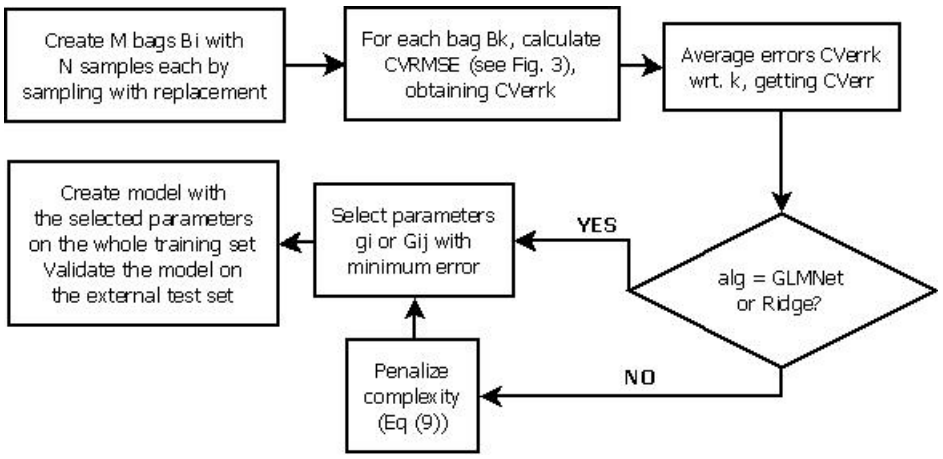


Fig. 4: Block diagram summarizing the bagged model selection approach

Please note, to take the average as aggregation function in Step 3 is the standard way as used in many other bagged modeling approach, such as, for instance, random forests [38]. It is well-known that it produces more robust predictions, especially in case of a low number of samples due to its characteristics to explore the sample space through the bootstrapped bags well, as e.g. analyzed in [10]. A low number of calibration samples is expected in our application, as the real targets have to be manually elicited by the experts, whereas such a manual analysis requires an effort of several hours for only a couple of samples. In this sense, the usage of bagging for our application is well motivated.

4 Case Study Configuration for Evaluation of Calibration Methods

4.1 Data Sets Characteristics and Pre-Treatment

Three data sets have been made available from beer production with manual target measurements. Two of them correspond to unfermented beer (wort), independently recorded in 2014 and 2015 with different parts in the measurement equipment (15μm and 20μm cells), and the third one to beer mix beverages beer production, the latter including a separate validation set recorded several weeks later. Due to the differences in the composition and in the final product, the parameters that are relevant for the product quality, and will be therefore monitored, are not the same. The concrete parameters and their acronyms (coming from German language) are

- For wort: (i) Bitterness (EBU), (ii) final attenuation (FA), and (iii) free amino nitrogen (FAN).
- For beer mix beverages: (i) Bitterness, (ii) foam stability (S), (iii) citric acid (CA), and (iv) total acid (TA).

The most relevant one is bitterness, which is known to show quite non-linear behavior, thus is a good motivation for non-linear approaches. Moreover, within a pre-study conducted by BrauUnion, it turned out that linear approaches failed to reach an acceptable accuracy for these targets.

Figures 5 and 6 show, respectively, the absorbance spectra for the one wort data set and the beer mix beverages data set. It can be seen in Figure 6 that the external validation set shows severe extrapolation, indicating a hard validation benchmark case, which can be indeed weakened by appropriate pre-processing methods, see below, but not completely avoided. Just by visual inspection it is clear that not all wavelengths are relevant and constructive for the modeling process (e.g., sudden peaks should be removed). Thus, subsets of the original 1790 wavenumbers have been selected by an expert. Those subsets contain between 200 and 500 wave-numbers each, depending on the data set and target. The selections have been tested against several stochastic and non-stochastic variable selection methods, **e.g. using uninformative variable elimination [39][40], forward selection [41] and genetic algorithms [18];** and have been found to be optimal for those subsets sizes.

In each data set the number of samples available for each target varies. Spectral data is continuously being recorded, but some targets require longer time to be measured than others. Due to the high effort for manual analysis of probes drawn in order to obtain the target values, the number of samples have been restricted to 31 for beer mix beverages, 47 for 2014 wort and 37 for 2015 wort data sets. After cleaning, this number reduces further to the values for the several targets as shown in Table 1. According to this very low number of samples, bagging which explores well the sample space, can be expected to provide more robust models (model selection) than the classical CV.

Besides, several well known preprocessing methods [42] have been employed in order to find a preprocessing strategy that behaves well for all targets. For operational reasons, as the spectral data are the same for all the targets, a single strategy for all targets of each data set is required. The chosen one for all data sets has been the 2-steps strategy consisting on first applying standard normal variate [43], and then mean centering.

4.2 State-of-Art Methods used for Comparison

It is known by the experts, that most of the parameters we are interested on show, to some extent, some non-linear behavior. Nevertheless, basic linear methods have been applied at the company's production site. In order to check (improved) per-

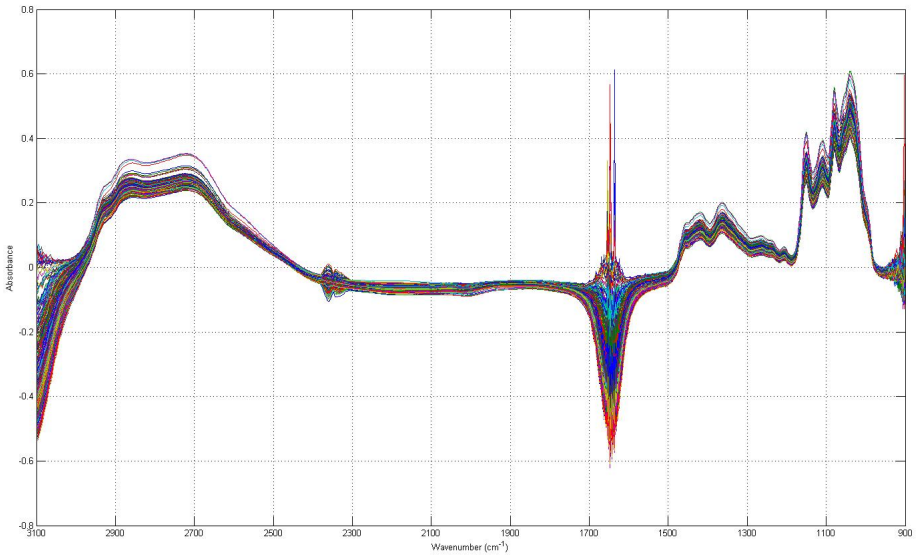


Fig. 5: Absorbance spectra for wort (2014).

Dataset	Target	Unit	Relevant spectral regions cm^{-1}	# of samples calib/valid
Wort 2014	Bitterness	EBU	1346-1564	47/-
Wort 2014	Final attenuation	%	976-1363	47/-
Wort 2014	Free amino nitrogen	mg/l	1012-1475, 2517-2980	47/-
Wort 2015	Bitterness	EBU	1134-1499	37/-
Wort 2015	Final attenuation	%	1070-1421	37/-
Wort 2015	Free amino nitrogen	mg/l	1012-1437	37/-
Beer mix beverages	Bitterness	EBU	1138-1443	31/11
Beer mix beverages	Foam stability	s	1207-1437, 2748-2960	31/11
Beer mix beverages	Citric acid	g/l	1148-1495	31/11
Beer mix beverages	Total acids	g/l	1051-1128, 1168-1495, 2748-2931	31/11

Table 1: Data sets characteristics used for calibration and validation

formance achievable with non-linear methods, we will compare our non-linear methods with the following state-of-the-art linear methods:

Partial Least Squares (PLS): **It is a linear method, used to find the fundamental relations between two matrices (input X and output Y), i.e. a latent variable approach to model the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the input (X) space that explains the maximum multidimensional variance direction in the output (Y) space. It emphasizes a rotation of the input space in order to best explain (the variance of the) target by means of linear relations/mappings. We have used it for comparison**

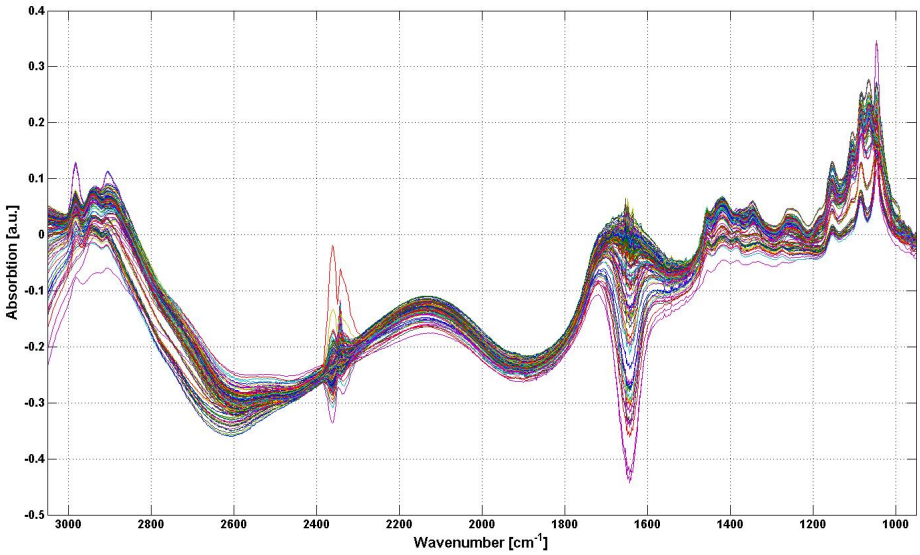


Fig. 6: Absorbance spectra for beer mix beverages, including both calibration and validation data.

purposes, because it is the most widely used state-of-the-art method in chemometrics and especially in automatic beer parameter analytics. For a compact summary of its principal concepts, please refer to the beginning of Section 3.1.

Generalized linear models with elastic net (GLMNet): The Lasso method [44] and ridge regression [43] are approaches included in the family of shrinkage methods that can be seen as regression algorithms including an ℓ_1 and an ℓ_2 penalties respectively. The *elastic net* [45] includes a penalty based on a combination of both ℓ_1 and ℓ_2 penalties, looking for some elasticity in the regularization.

Generalized linear models [43] is a generalization of ordinary linear regression that provides flexibility in the sense that the distribution of the errors is not necessarily supposed to be normal, as happens in ordinary linear regression.

The combination of the elastic net with generalized linear models is a regression algorithm based on generalized least squares that uses cyclical coordinate descent [46] in a path-wise fashion [47] in order to select the optimum elasticity in the regularization via the elastic net.

Ridge regression: Despite it is a particular case of GLMNet, when the lasso part of the elastic net is ignored, ridge regression deserves its own separate spot. In MLR we determine the best regression vector $\hat{\mathbf{b}}$, according to a minimum least squares criterion, when trying to solve the regression problem $\mathbf{y} = \mathbf{X} \cdot \mathbf{b}$ with $\mathbf{X}$ the regression matrix. Then, it is well known that the regression vector is

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$

When handling variables that are highly correlated, problems of singularities arise when it comes to calculating the inverse of $\mathbf{X}^T\mathbf{X}$. A way to deal with this problem is *regularization*. It consists on adding a regularization term in the

least squares minimization problem. Ridge regression uses $\alpha\mathbf{IX}$ as regularization term, where $\lambda > 0$ is a parameter to be tuned. Then, the regression vector becomes

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y}$$

The regularization parameter will be tuned with a grid, see below.

4.3 Evaluation Scheme and Parametrization

The evaluation scheme is performed differently for the unfermented beer and beer mix beverages beer, due to the characteristics of the data. For unfermented beer, we have performed the classical cross validation model selection, as stated in Section 3.3 for both data sets separately, so that we can compare the performance of the regression methods. As a hard benchmark, we used the final models trained on the 2014 data for validation on the 2015 data (different measurement equipments), just to check how far our models are able to reliably extrapolate into the future. For beer mix beverages, the availability of validation data offers the possibility of comparing also the classical and bagged cross validation model selections, in order to see how close those model selection approaches are to the best possible parameter combination for the external validation set (which is not accessible during CV selection).

The proposed parameter grids are the following:

- PLS: The number of latent variables **(coded as** $P1$ **in the tables in Section 5)** varies from 1 to 15.
- Ridge: The regularization parameter λ **(coded as** $P2$**)** goes from 0.01 to 0.95 in steps of 0.05. This grid has been successfully used in previous studies [48], in which the data were obtained under similar circumstances in similar real world problems.
- GLMNet: The regularization parameter λ **(coded as** $P1$**)** has been set from 0.01 to 0.09 in steps of 0.01, in order to leave the default value suggested by the proposers of the method, 0.05, in the middle of the grid. The parameter α **(coded as** $P2$**)**, responsible for playing with the elasticity in the elastic net takes the values from 0.1 to 1.0 in steps of 0.1. Notice that $\alpha = 1$ is equivalent to use pure lasso, and $\alpha = 0$ would be pure ridge (excluded here because it has its own spot).
- FLEXFISPLS: The dimensionality **(coded as** $P1$**)** varies in the same way as the number of latent variables in PLS, and the vigilance **(coded as** $P2$**)** takes the values between 0.1 and 0.9, with steps of length 0.1.
- ε-PLSSVR, ν-PLSSVR: As mentioned in Section 3.2, the number of latent variables used is not tuned, but fixed as the selection made for PLS. Besides, the cost and spread parameters **(coded as** $P1$ **and** $P2$ **respectively)** take the values suggested by the Lib-SVM library developers. Thus, C takes the values in $\{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$, and γ in $\{2^{-15}, 2^{-13}, \ldots, 2^{3}\}$.

5 Results

The results section is structured according to the two validation schemes we have conducted for performance evaluation:

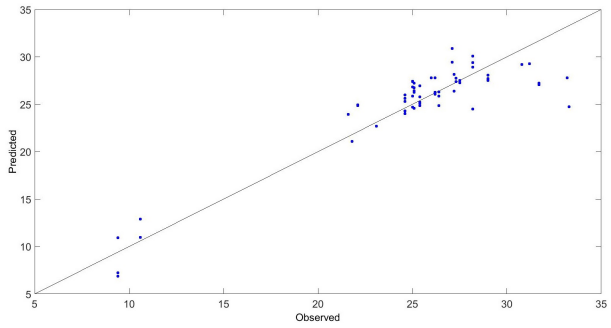
1. A classical and enhanced (employing bagging) cross-validation procedure on each of the training data sets for wort 2014, wort 2015 and beer mix beverages data.
2. Validation on a separate available test data set in case of beer mix beverages, as well as validation of the final models trained on wort 2014 data on the wort 2015 data (hard benchmark).

In the following two subsection we will visually show the results and perform a detailed interpretation of them.

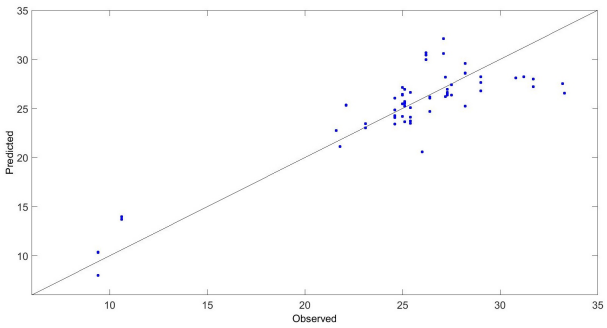
5.1 Cross-Validation Performance

When it comes to unfermented beer, the most relevant characteristics to be monitored are bitterness, final attenuation, and free amino nitrogen.

Figure 7 shows the results for EBU in the data set for wort beer from year 2014; we can see: 7a, and 7b the correlation plots corresponding to, respectively, the best non-linear and linear methods; 7c a summary table containing the selected parameters **, the CVRMSE and the average R^2 of the predictions in all folds (CVR2)** for each calibration method; and 7d the observed vs predicted plot for the method achieving the lowest CVRMSE (highlighted in bold font in 7c). Analogously, the results for FA and FAN targets are shown in Figures Online Resource 2 and Online Resource 3 respectively.



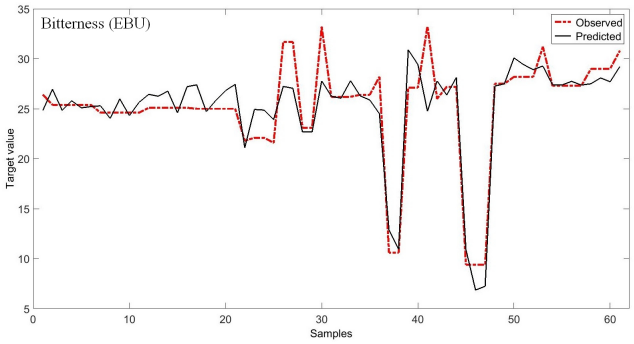
(a) Best non-linear.



(b) Best linear

	P1	P2	CVRMSE	CVR2
PLS	5	—	3.9036	0.6439
GLMNet	1.0	0.04	**3.8533**	0.6516
Ridge	—	0.96	4.0770	0.6133
FLEXFIS	5	0.4	3.6296	0.6942
ε-SVR	2^3	2^7	2.4779	0.8628
ν-SVR	2^3	2^7	**2.4721**	0.8622

(c) Summary table

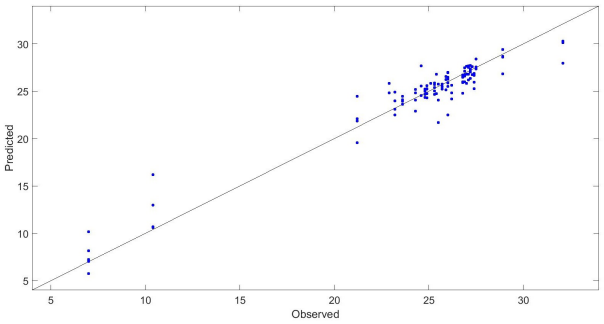


(d) Observed vs predicted (best)

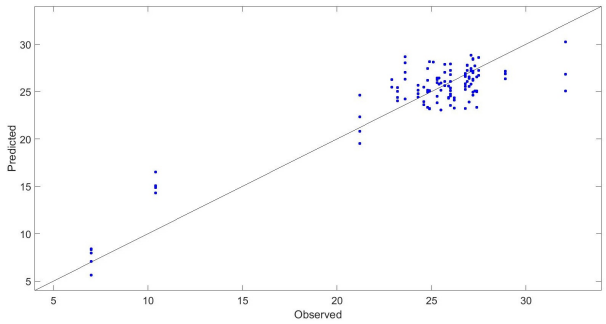
Fig. 7: **CV summary results for bitterness in the data set for wort beer from year 2014. We can see: 7a, and 7b the correlation plots corresponding to, respectively, the best non-linear and linear methods; 7c a summary table containing the selected parameters, the CVRMSE and the average R^2 of the predictions in all folds (CVR2) for each calibration method; and 7d the observed vs predicted plot for the method achieving the lowest CVRMSE (highlighted in bold font in 7c).**

Notice that for final attenuation (see Figure Online Resource 2) the performance of the non-linear methods is indeed quite similar to the performance of PLS, which is linear, thus theoretically less prone to over-fitting. Besides, it is good to

see that the prediction ability for samples close to the targets' extreme values is high, despite the lack of balance in the data. For FAN (Figure Online Resource 3), both SVR approaches show around 15% lower CVRMSE than the rest. Just by ocular inspection, comparing Online Resource 3 (a) and (b), we can see that the SVR approach performs well in both upper and lower boundaries, and GLMNet does not. Thus this explains the difference in the CVRMSE. In case of bitterness, the most important parameter for wort supervision (as being responsible for the final taste for customers), the improvement achieved by SVR compared to the best linear method GLMnet is about 36%, finally achieving the company's goal to stay within the error range limit of 3 (an error of 2.47 is achieved), which is not the case for linear methods (and error of 3.85 is achieved).



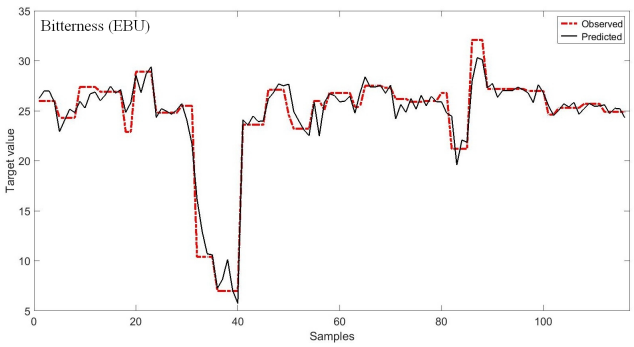
(a) Best non-linear.



(b) Best linear

	P1	P2	CVRMSE	CVR2
PLS	8	—	2.1702	0.8150
GLMNet	1.0	0.01	**2.0737**	0.8319
Ridge	—	0.56	2.1307	0.8232
FLEXFIS	6	0.4	1.6494	0.8934
ε-SVR	2^3	2^7	1.3170	0.9341
ν-SVR	2^3	2^7	**1.2891**	0.9368

(c) Summary table

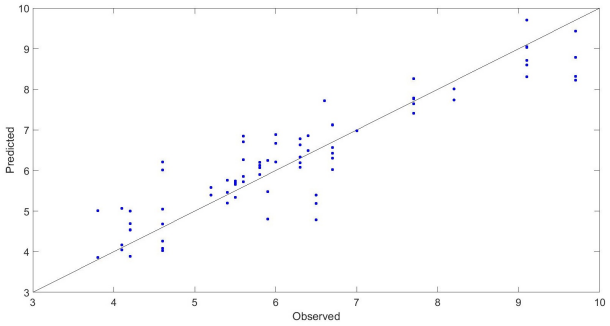


(d) Observed vs predicted (best)

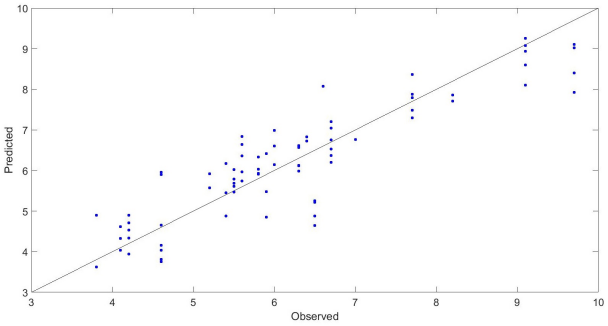
Fig. 8: **CV summary results for bitterness in the data set for wort beer from year 2015. We can see: 8a, and 8b the correlation plots corresponding to, respectively, the best non-linear and linear methods; 8c a summary table containing the selected parameters, the CVRMSE and the average R^2 of the predictions in all folds (CVR2) for each calibration method; and 8d the observed vs predicted plot for the method achieving the lowest CVRMSE (highlighted in bold font in 8c).**

When it comes to the data set for wort beer from 2015 (flow cell with 20 μm), the structure of the results is similar for both, the linear and non-linear methods. Again, there is a clear outperformance of linear methods by non-linear ones in case of bitterness (see Figure 8) and FAN (see Online Resource 5), but this time also for final attenuation (Online Resourse 4). The overall conclusions are a good extrapolation behavior, little risk of over-fitting for both SVR approaches, and a much

lower one for FLEXFIS, which is noticeable when we see that the dimensionality is lower than in PLS and the vigilance is pretty high, at least 0.3, which is usually an indicator of a low non-linearity degree. If vigilance is below 0.3, that is an indicator for very high non-linearity in our model, thus high risk of over-fitting.



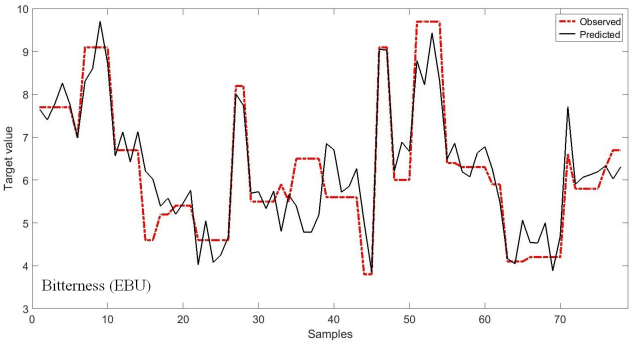
(a) Best non-linear.



(b) Best linear

	P1	P2	CVRMSE	CVR2
PLS	7	—	0.6910	0.8080
GLMNet	0.1	0.01	0.6951	0.8059
Ridge	—	0.31	**0.6667**	0.8209
FLEXFIS	6	0.6	0.6785	0.8210
ε-SVR	2^3	2^5	**0.3676**	0.9462
ν-SVR	2^3	2^5	0.3788	0.9422

(c) Summary table



(d) Observed vs predicted (best)

Fig. 9: **CV summary results for bitterness in the data set for beer mix beverages. We can see: 9a, and 9b the correlation plots corresponding to, respectively, the best non-linear and linear methods; 9c a summary table containing the selected parameters, the CVRMSE and the average R^2 of the predictions in all folds (CVR2) for each calibration method; and 9d the observed vs predicted plot for the method achieving the lowest CVRMSE (highlighted in bold font in 9c).**

For beer mix beverages beer, it is noticeable that in both, citric acid (see Online Resources 7) and total acids (Online Resources 8), the performance of both linear and non-linear approaches is similar. In both targets FLEXFIS is the worst algorithm, but the situations are different. The parameters for the total acids look coherent, but in the case of citric acid, it seems that the CV model selection aimed to a parameter combination with two huge clusters (dimensionality equals 2, much lower than the amount of LVs in PLS, because the vigilance is the lowest possible). In case of bitterness (Figure 9), non-linear methods can again outperform linear ones significantly (as is the case for wort data) — whether this is a matter of over-fitting or not (because of the high parameter values in SVR), will be clarified in the subsequent section when illuminating the results on the separate validation data set. When it comes to foam stability (Online Resources 6), the difference between the number

of LVs (latent variables) from PLS and the dimensionality for FLEXFIS is quite big. Nevertheless, it has to be understood in terms of non-linearity degree. PLS needs more dimensions in order to catch part of the non-linearity, but FLEXFIS can do it with lower ones (vigilance indicates a mid-high degree of non-linearity).

5.2 Performance on Separate Validation Data

Regarding the data corresponding to beer mix beverages beer, the bagging approach has also been applied for model selection as an alternative to standard cross validation. Those results are not shown in previous section, because the final error used in the bagging approach has a different purpose. That error measure is an average made from CVRMSEs coming from the different bags, thus comparing that error measure with the usual CVRMSE makes no sense. Nevertheless, the aim is not to use that error measure as an estimation of the future performance on unseen data, but to take advantage of the robustness provided by the use of diverse bags for a better, more robust model selection, in order to obtain a lower root mean square error of prediction (on separate test data) (termed RMSEP) and to reduce the over-fitting effect.

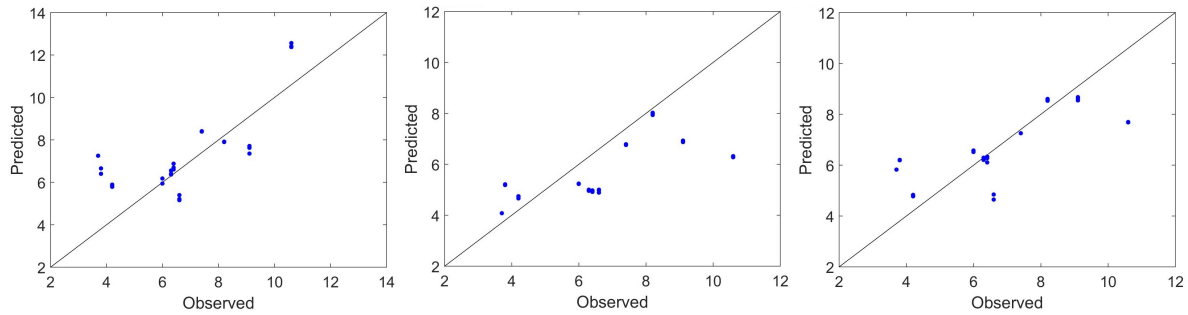
In order to check the performance of the two model selection approaches (classical and bagged), we have calculated the RMSEP for all possible combinations of the model learning parameters (see Section 4.3), so the parameter combination corresponding to the minimum RMSEP is considered as the best possible model selection overall,

$$\left(\widehat{P}_1, \widehat{P}_2\right) = argmin_{(P_1,P_2)\in \mathbf{G}_1\times\mathbf{G}_2}\left(RMSEP_{P_1,P_2}\right)$$

In this way we can see how close our model selections are from the best overall possible model selection.

Figures 10, Online Resources 9, Online Resources 10 and Online Resources 11 show the results respectively for bitterness, foam stability, citric acid, and total acid. The structure of each Figure is: (a) correlation plot for the best method according to the classical CV selection, i.e. the parameters corresponding to the minimal CV error are selected (thus, a direct comparison to the figures for the CV results is possible), (b) correlation plot for the best method according to the new bagged selection, (c) correlation plot for the best method according to the (theoretically) best possible selection (corresponding to minimal entry in the right half of the tables), and (d) summary table containing the selected parameters for each algorithm (in both classical CV and bagged selections, the latter indicated in the method name by the appended term 'Bag'), the root mean square error of prediction, and the corresponding R^2 for our selections (columns 2-5) and the best possible selections (columns 6-9).

When it comes to bitterness, Figure 10, we can see that in both the best model selection and the best possible selection there is a very good prediction ability in the important range $[5,8]$. Besides, the errors are systematically below 2, that is a requirement from the company. With one single exception, ε-SVR, the use of the bagged model selection improve the classical one, leading the selection to models with lower complexity (lower number of LVs in PLS, lower number of rules in FLEXFIS, and lower number of support vectors in SVR), while being closer to the best possible models in terms of error performance. In fact, in most cases the **RMSEP of the model selected by grid search (RMSEPGS)** can be significantly reduced, especially in case of fuzzy modeling (using FLEXFIS) down to 1.55, clearly outperforming all state-of-the-art methods. In the case of SVR approaches, there is margin for improving the selection process. The most promising action would be to estimate/compute the adequate number of LVs to be used, instead of using the ones obtained for PLS. One more thing should be noticed for ν-SVR, that is the improvement that bagging brings — not clearly visible in the RMSEP, but in



(a) Best classical model selection. (b) Best bagged model selection. (c) Best overall

	P1GS	P2GS	RMSEPGS	R2PGS	P1B	P2B	RMSEPB	R2PB
PLS	7	—	1.7145	0.4532	3	—	1.6114	0.5487
PLS-Bag	2	—	1.6597	0.4978	3	—	1.6114	0.5487
GLMNet	0.1	0.01	**1.7004**	0.4541	1	0.02	1.6317	0.5193
GLMNet-Bag	1	0.09	1.6467	0.5126	1	0.02	1.6317	0.5193
Ridge	—	0.31	1.7138	0.4513	—	0.96	1.6882	0.4650
Ridge-Bag	—	0.01	1.6986	0.4430	—	0.96	1.6882	0.4650
FLEXFIS	6	0.6	1.8536	0.4608	3	0.1	**1.4066**	0.5748
FLEXFIS-Bag	2	0.9	**1.5597**	0.4978	3	0.8	**1.4066**	0.5748
ε-SVR	2^3	2^5	1.9497	0.2166	2^1	2^3	1.6225	0.4107
ε-SVR-Bag	2^{-15}	2^{-5}	2.0571	0.2299	2^1	2^1	1.6225	0.4107
ν-SVR	2^3	2^5	2.0702	0.1943	2^1	2^3	1.5947	0.4321
ν-SVR-Bag	2^{-15}	2^{-5}	2.0456	0.4453	2^1	2^1	1.5947	0.4321

(d) Summary table

Fig. 10: **Summary results of external validation for bitterness for the data corresponding to beer mix beverages. The four parts correspond to: (a) correlation plot for the best method according to the classical CV selection, i.e. the parameters corresponding to the minimal CVRMSE), (b) correlation plot for the best method according to the new bagged selection, (c) correlation plot for the best method according to the (theoretically) best possible selection (corresponding to minimal entry in the right half of the summary table), and (d) summary table containing the selected parameters for each algorithm (in both classical CV and bagged selections, the latter indicated in the method name by the appended term 'Bag'), the root mean square error of prediction, and the corresponding R^2 for our selection (columns 2-5) and the best possible selection (columns 6-9).**

the R^2. The reason for it is the presence of some isolated high error peaks in the boundaries of the range that penalize the error, but not the correlation.

For foam stability (see Online Resources 9), similar observations as in case of bitterness can be made, whereas the improvement achieved by bagging is even more intense in case of non-linear methods (e.g., reduction of more than 50% error in case of ε-SVR down to an error of around 26). In this sense, this variant in combination with bagging is the most feasible option. Compared to the CV results, the errors are indeed significantly worse but with the help of bagging still lying

within the company's upper limit of 30 (which is not achievable with classical CV selection). The best possible selection (right half of the table in Figure Online Resources 9) does not really further improve the error on separate validation data. Hence, the bagged selection already achieves the optimum performance during CV, which is the ideal situation as the separate test data set is generally not accessible during the training phase.

For citric acid and total acids (Online Resources 10 and 11 respectively), bagging helps only in the non-linear methods. The reason for that is clear, the higher the risk of overfitting, the bigger the advantage of bagged approaches, but the non-bagged variants already perform pretty well (close to the CV results) and clearly in-line the upper error limit of 0.3. It is known by the experts that the most non-linear target is bitterness. This fact is confirmed by the results, in which linear methods seem to be the best ones, being GLMNet the preferred of those. Besides, fuzzy modeling with FLEXFIS behaves better than all linear models, despite the model selection cannot see it. The reason is the flexibility of FLEXFIS to adapt to any degree of non-linearity, even light non-linearity like in the case of both citric acid, and total acid. In the case of citric acid, the classical selection for FLEXFIS is working badly, leading to the lowest possible vigilance. The consequence of that is a high number of rules. Bagging selects the same dimensionality, but with a higher vigilance (= a lower number of rules), that is closer to the best possible selection and expected to be more robust (as less prone to over-fitting) for prediction on future data.

Finally, the validation of the 2014 wort models with the 2015 wort data, which we checked by incidence (thus have not been a requirement by the company) did not bring any reasonable results for bitterness and FAN, as the errors raised to significantly above 4 in case of bitterness and to above 14 in case of FAN (both significantly above the requested upper limits), also when taking into account the best possible parameter/model selection. However, for FA they stayed in the same range as achieved through cross-validation, which is a remarkable result due to the fact that they have been recorded with two different measurement equipments.

6 Conclusion and Outlook

This paper proposes two non-linear modeling techniques for calibrating models to predict important parameters during beer production. The supervision of them is necessary in order to guarantee a high level of beer quality, to assure that a beer tastes in the same way as used to within small boundaries of variation and thus that it satisfies the customers' expectations. Current state-of-the-art chemometric methods based on spectroscopic measurements does not meet the minimal prediction error requirement provided by the company for all the important parameters (especially not for bitterness and final attenuation), which, however, can be resolved with two non-linear modeling techniques, 1.) the first one relying on a non-linear version of PLS with the usage of Takagi-Sugeno fuzzy systems for obtaining piecewise linear predictors, and 2.) the second one (with even higher performance) relying on a variation of support vector regression. In particular, an error reduction of about 35% up to 45% in case of bitterness and of about 50% in case of final attenuation could be achieved. Furthermore, in case of beer mix beverages, the new, robust model selection scheme based on bagged ensembles lead to significant error reduction on separate validation data for foam stability and bitterness: especially, in case of foam the error can be reduced from 64 down to below the upper allowed limit of 30, which is remarkable. In case of the acids for mix beverages, no significant improvement could be made, as the linear models already performed very well on them.

Future work includes the usage of enhanced genetic algorithms for wavelength selection in the context of differential evolution and co-evolution (as having been successfully applied before on FT-NIR spectra data from another application [18] by the main authors of this paper) as well as the application of more advanced ensemble methods such as, e.g., boosting or random forests for a better stability of prediction errors on separate validation data. Additionally, more important parameters for different types of alcoholic and non-alcoholic beverages will be analyzed by our non-linear modeling techniques.

Acknowledgements

Financial support was provided by (i) the Austrian research funding association (FFG) under the scope of the COMET programme within the research project Industrial Methods for Process Analytical Chemistry - From Measurement Technologies to Information Systems (imPACts) (contract #843546), (ii) the Basque Government through the ELKARTEK and BERC 2014-2017 programs, and (iii) the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa accreditation SEV-2013-032. This publication reflects only the authors' views.

Compliance with Ethical Standards

The authors declare that they have no conflict of interest. This paper does not contain any studies with human participants or animals performed by any of the authors.

References

1. R. A. Speers, P. Rogers, and B. Smith. Non-linear modelling of industrial brewing fermentations. *J I Brewing*, 109(3):229–235, 2003.
2. Y. Zhang, S. Jia, and W. Zhang. Predicting acetic acid content in the final beer using neural networks and support vector machine. *J I Brewing*, 118(4):361–367, 2012.
3. M. McMurrough, V. Lynch, F. Murray, and M. Kearney. A comparison of alternative high-performance liquid chromatographic systems for measuring bitterness in beer. *J Am Soc Brew Chem*, 45:6–13, 1987.
4. D. de Keukeleire. Fundamentals of beer and hop chemistry. *Quim Nova*, 23(1):108–112, 2000.
5. E. Polshin, B. Aernouts, W. Saeys, F. Delvaux, F.R. Delvaux, D. Saison, M. Hertog, B.M. Nicolai, and J. Lammertyn. Beer quality screening by FT-IR spectrometry: Impact of measurement strategies, data pre-processings and variable selection methods. *J Food Eng*, 106(3):188–198, 2011.
6. D.W. Lachenmeier. Rapid quality control of spirit drinks and beer using multivariate data analysis of fourier transform infrared spectra. *Food Chem*, 101(2):825–832, 2007.
7. J. Christensen, A.M. Ladefoged, and L. Nrgaard. Rapid determination of bitterness in beer using fluorescence spectroscopy and chemometrics. *J I Brewing*, 111(1):3–10, 2012.
8. S. Grassi, J.M. Amigo, C.B. Lyndgaard, R. Foschino, and E. Casiraghi. Beer fermentation: monitoring of process parameters by FT-NIR and multivariate data analysis. *Food Chem*, 155:279–286, 2014.

9. M. Haenlein and A.M. Kaplan. A beginner's guide to partial least squares (PLS) analysis. *Und Stat*, 3(4):283–297, 2004.
10. P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta. *Metalearning*. Springer, Berlin Heidelberg, 2009.
11. Z. Bleier, C. Brouillette, and R. Carangelo. A monolithic interferometer for FT-IR spectroscopy. *Spectroscopy*, 14(10):46–49, 1999.
12. P.R. Griffiths and J.A. De Haseth. *Fourier Transform Infrared Spectrometry, 2nd Edition*. John Wiley & Sons, New Jersey, 2007.
13. I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, Berlin Heidelberg New York, 2002.
14. R.G. Brereton. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons, Hoboken, New Jersey, 2003.
15. K. Varmuza and P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Boca Raton, 2009.
16. M. Otto. *Chemometrics, 2nd Edition*. John Wiley & Sons, Hoboken, New Jersey, 2007.
17. H. Mark and J. Workman. *Chemometrics in Spectroscopy*. Academic Press, Elsevie, The Netherlands, Amsterdam, 2007.
18. C. Cernuda, E. Lughofer, P. Hintenaus, and W. Märzinger. Enhanced waveband selection in NIR spectra using enhanced genetic operators. *J Chemometr*, 28(3):123–136, 2014.
19. R. Rosipal. Kernel partial least squares for nonlinear regression and discrimination. *Neural Netw World*, 13(3):291–300, 2003.
20. C. Cernuda, E. Lughofer, P. Hintenaus, W. Märzinger, T. Reischer, M. Pawlicek, and J. Kasberger. Hybrid adaptive calibration methods and ensemble strategy for prediction of cloud point in melamine resin production. *Chemometr Intell Lab*, 126:60–75, 2013.
21. E. Lughofer. *Evolving Fuzzy Systems — Methodologies, Advanced Concepts and Applications*. Springer, Berlin Heidelberg, 2011.
22. E. Lughofer, C. Cernuda, S. Kindermann, and M. Pratama. Generalized smart evolving fuzzy systems. *Evol Sys*, 6(4):269–292, 2015.
23. K. Krishnamoorthy and T. Mathew. *Statistical Tolerance Regions: Theory, Applications, and Computation*. John Wiley & Sons, Hoboken, New Jersey, 2009.
24. P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India, vol. 2 (1)*, pages 49–55, 1936.
25. R.M. Gray. Vector quantization. *IEEE ASSP Mag*, 1(2):4–29, 1984.
26. E. Lughofer and M. Sayed-Mouchaweh. Autonomous data stream clustering implementing incremental split-and-merge techniques — towards a plug-and-play approach. *Inform Sciences*, 204:54–79, 2015.
27. P.P. Angelov and D. Filev. An approach to online identification of Takagi-Sugeno fuzzy models. *IEEE T Syst Man Cy B*, 34(1):484–498, 2004.
28. E. Lughofer. FLEXFIS: A robust incremental learning approach for evolving TS fuzzy models. *IEEE T Fuzzy Syst*, 16(6):1393–1410, 2008.
29. V. Vapnik. *Statistical Learning Theory*. Wiley and Sons, New York, 1998.

30. B. Schölkopf and A.J. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, London, England, 2002.
31. T. Hofmann, B. Scholkopf, and A.J. Smola. Kernel methods in machine learning. *Ann Stat*, 36(3):1171–1220, 2009.
32. A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Stat Comp*, 14:199–222, 2004.
33. C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification, 2010.
34. M. Stone. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc*, 36(1):111–147, 1974.
35. Leo Breiman. Bagging predictors. *Mach Learn*, 24(2):123–140, 1996.
36. L. P. Bras, M. Lopes, A.P. Ferreira, and J.C. Menezes. A bootstrap-based strategy for spectral interval selection in pls regression. *J Chemometr*, 22(11–12):695–700, 2008.
37. B. Efron and R. Tibshirani. Improvements on cross-validation: The .632 + bootstrap method. *J Am Stat Assoc*, 92(438):548–560, 1997.
38. L. Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.
39. V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, and S. Sterna. Elimination of uninformative variables for multivariate calibration. *Anal Chem*, 68(21):3851–3858, 1996.
40. W. Cai, Y. Li, and Shao X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometr Intell Lab*, 90:188–194, 2008.
41. C.R. Andersen and R. Bro. Variable selection in regression - a tutorial. *J Chemometr*, 24(11-12):728–737, 2010.
42. A. Rinnan, F. van den Berg, and S. B. Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *Trend Anal Chem*, 28(10):1201–1222, 2009.
43. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction (Second Edition)*. Springer, New York Berlin Heidelberg, 2009.
44. R. Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc*, 58:267 – 288, 1996.
45. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J R Stat Soc*, 67(2):301–320, 2005.
46. T. Hastie, R. Tibshirani, and J. Friedman. Regularized paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1), 2010.
47. T. Hastie, R. Tibshirani, and J. Friedman. Pathwise coordinate optimization. *Ann Appl Stat*, 1(2):302–332, 2007.
48. C. Cernuda, E. Lughofer, W. Maerzinger, and J. Kasberger. NIR-based quantification of process parameters in polyether-acrylat (PEA) production using flexible non-linear fuzzy systems. *Chemometr Intell Lab*, 109(1):22–33, 2011.