Multi-source Heterogeneous Data Fusion for Toxin Level Quantification

Eugeniu Strelet*, Zhenyu Wang, You Peng**, Ivan Castillo**, Ricardo Rendall**, Bea Braun**, Mark Joswiak**, Leo Chiang**, Marco S. Reis***

** Univ Coimbra, CIEPQPF, Department of Chemical Engineering, Rua Sílvio Lima, Pólo II – Pinhal de Marrocos, 3030-790 Coimbra, Portugal (e-mail: marco@eq.uc.pt).*
***Dow Inc, Freeport, USA (e-mail:HChiang@Dow.com)*

Abstract: The operational management of wastewater treatment plants (WWTP) is a complex activity due to the biological phenomena' intricate nature. This complexity hinders the adoption of first principles approaches, which lack the necessary accuracy to be adopted in practice. Data-driven methodologies also face significant challenges in processing the different information sources available. In this work, we present a data-driven and model-agnostic data-fusion framework to estimate the concentration level of a toxin in the effluent, using heterogeneous data (sensor data, images, laboratory measurements) collected at different locations in the process. Single- and multi-source modeling approaches are applied and compared. Among the methodologies tested, Bayesian fusion stands out as presenting a good balance in terms of accuracy, stability, and flexibility.

Keywords: Heterogeneous data; Irregular sampling; Data fusion; Variable selection; Data-driven modeling.

1. INTRODUCTION

The operational management of wastewater treatment plants (WWTP) presents many challenges due to the complex nature of the biological processes, namely their non-stationarity, intrinsic variability, and possible non-linear behavior. Furthermore, data collected from such systems also present several challenges of their own. For example, the target quality parameter(s) are often only available at low frequencies. Also, inferential sensors, Process Analytical Technology (PAT) and imaging instrumentation do not sample frequently enough, are often not synchronized, and present limitations on reliability and accuracy. On the other hand, first principles models are not available to describe the system with enough accuracy or have too many unknown or unreliably estimated parameters. In this context, data collected from different sources, at different frequencies, and with distinct structures, even though difficult to handle, provide the only viable source of information to address WWTP management and optimization.

In this work, we report a novel solution to estimate a regulatory controlled toxin in processed wastewater, from multiple data sources with different structures and collected at different acquisition rates. One of the goals of the WWTP is to treat the wastewater so that the toxin level does not exceed the Environmental Protection Agency (EPA) compliance for disposal. The main processing stages of the WWTP are presented in Figure 1. They consist of a settling unit, a flotation unit, and finally, a filtering unit, after which the effluent is transferred to the environment. The key step for toxin removal happens at the flotation unit where additives are mixed with the water streams to form flocculates containing the toxin, which are then separated from the clean liquid. Wastewater samples taken from the bulk below the flocculates are sent for laboratory analysis for toxin level quantification approximately 1-3 times per week (sampling points shown in Figure 1). To effectively control the toxin concentration and secure it does not exceed the compliance limit, it is highly desirable to have information on a more frequent basis, and to know how it relates with the additives used in the process. However, more frequent testing is not possible due to the high cost and intensive labor involved. Therefore, a data-driven framework is proposed in this work to estimate the toxin level in the effluent, using data from multiple sources, including the flotation process and features extracted from images collected in two positions (after the flotation and settling units).

At the beginning of the analysis, very little was known about the structure of the available data sets and their relationship with the toxin. Furthermore, there was a lack of understanding about the biological processes taking place mainly at the settling unit, due to the complexity of the bioreactor media. In addition, the different data sources available present structural heterogeneity, different time resolutions (granularity ranges from several hours to one day) and their collection is not synchronized. All these challenges to data analysis were addressed in the framework proposed in this article.

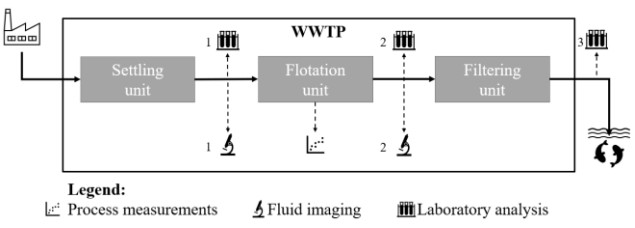

Figure 1 - Schematic representation of the Wastewater Treatment Plant.

In order to handle the aforementioned challenges and to achieve a robust and accurate estimation of the toxin level, an information fusion methodology was conceived. Data/information fusion methods offer flexible solutions for

handling the integration of different data sources (Alyannezhadi *et al.*, 2017; Azimirad & Haddadnia, 2015; Castanedo, 2013; Diez-Olivan *et al.*, 2019; Sansana *et al.*, 2020; Wang & Chiang, 2019). The success of the fusion platform depends on the careful design of its components, including the development of inferential models for the different data sources (Mitchell, 2012; Sidek & Quadri, 2012). Important to the development of these models is the quality of data and the information generated by integrating data with analytics for achieving a given goal, i.e., the InfoQ (Kenett & Shmueli, 2014). The InfoQ concept was adapted to the development of solutions for the Chemical Processing Industry (CPI) (Reis & Kenett, 2018), and its principles will guide the development of the fusion platform proposed here.

This paper is organized as follows. In Section 2 we present the data sets analyzed and the workflow followed. In Section 3, the modelling results obtained are presented in detail. A discussion of the results is made in Section 4, where the main conclusions are also summarized, and ideas for future work are shared. Due to confidentiality issues, all numerical data used were normalized and variables renamed.

2. DATA SETS AND METHODS

In this section we briefly introduce the data sets used and the proposed methodology.

2.1 Data sets description

Three data sets from different locations in the WWTP were collected and used to predict the toxin level at the effluent of the filtering unit: (i) process measurements at the flotation unit, (ii) particle images taken at the settling and flotation units, and (iii) laboratory measurement of toxin level collected after the settling and flotation units. Process data (i) are available as daily averages. Toxin and image data are collected 2-4 times per week (irregular sampling). The process data set consists of 115 variables (*e.g.,* temperatures, turbidities, flowrates, etc.) and 659 observations. Image data (ii) are obtained using equipment that capture images of individual particles (microbiological media) in the waste water samples. Images are collected in two positions, namely after the flotation process and after the settling unit; the corresponding data sets contain 156 and 106 observations, respectively. Additionally, the toxin level (iii) is also measured during the process (153 and 106 samples collected after the settling and flotation units, respectively; the target response is the toxin level after the flotation unit). Data collection covers 2 years of operation.

Images are used for image-based solid particle quantification: several images are collected for each sample and processed for feature extraction (features consist of summary statistics of geometrical properties of objects identified in the images). A total of 1302 image features are extracted per sample.

2.2 Exploratory data analysis

The structure of data acquired influences the performance of the inferential models derived from them. Therefore, the correlation structure of all data sets was first visualized using heat maps of the absolute Pearson's correlation coefficient and Principal Component Analysis (PCA) (Jackson, 1991; Jolliffe, 2002). Furthermore, the capability of each feature to predict the toxin level was also analyzed, in order to detect potential strong predictors as well as assess the existence of sparsity in the regressors and infer the need to apply filters for predictors screening. This exploratory step is also useful for detecting outliers and eliminate segments of bad data.

2.3 Model development

In the absence of prior information about the best modelling approach, the source-dependent models were developed following a model-agnostic approach. No modelling methodology was assumed *a priori* to be superior, and their merits were assessed and compared using real data, under a robust and systematic training/testing scheme. Here, we have considered six modelling methods arising from different corners of the analytics landscape that have the potential to cope with the characteristics of our data sets (*i.e.* high dimensional feature space): Partial Least Squares (PLS) (Wold *et al.*, 2001), LASSO (Tibshirani, 1996), Elastic Net (EN) (Zou & Hastie, 2005), Random Forests (Breiman, 2001), Boosting, and Multi-Layer Perceptron (MLP) (Koskela *et al.*, 1996).

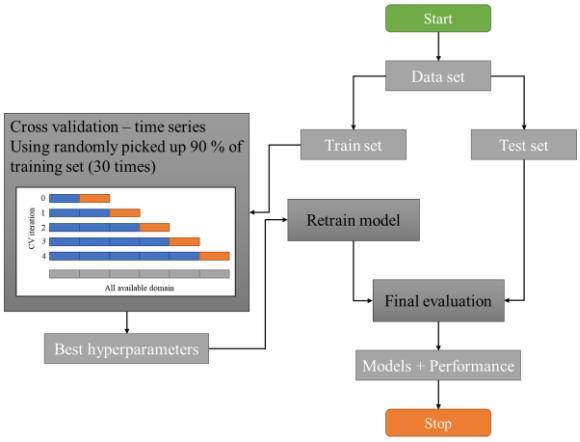

Figure 2 - Training and testing procedure carried out for each modelling methodology.

To assess the performance of the single-source models, each data set was split into two subsets (a training set with the initial 80% of the observations and a test set with the remaining 20%, located at the end; this splitting is more appropriate for assessing the performance of the methods in the future). The training set was used to estimate the model, where the Repeated Prequential (RP) method (Cerqueira *et al.*, 2019) was applied to tune the hyper-parameters (number of components, regularization constant, etc.), before estimating the remaining model parameters.

The RP method is a cross-validation approach designed for time series data. In this work, we implemented this method several times, by removing 10% of the training data (aligned by the response). At each time, the new estimated model was applied to the same test set (the 20% left out test set), obtaining the corresponding performance estimation. The procedure is summarized in Figure 2.

Given this nested cross-validation procedure, in this work we distinguish two performance indicators expressed in terms of

Root Mean Squared Error (RMSE, Equation 1): the Cross-Validation RMSE (RMSECV) using the RP methodology (inner cycle); and the Prediction RMSE (RMSEP) obtained in the test set (outer cycle).

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

2.4 Variable selection

The objective of variable selection is to narrow down the candidate set of predictors, selecting those with the highest potential to generate high quality quality information, *i.e.* for maximizing InfoQ (Reis & Kenett, 2018). Variable selection (VS) plays a critical role during model development and can significantly affect the predictive performance by mitigating the effects of over-fitting.

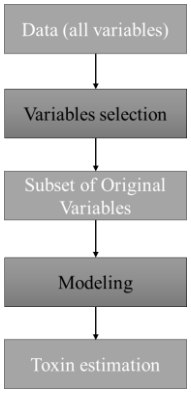

Figure 3 – Model development procedure.

In the present work, features were selected based on variable importance metrics directly linked to each regression methodology under analysis:

- Variables Importance in Projection (VIP): PLS
- Regression Coefficients: LASSO and EN
- Features Importance (FI): RF and Boosting
- Permutation Importance (PI): MLP

According to the appropriate variable importance measure for each type of model, variables were selected prior to the modeling stage (Figure 3; an exhaustive search of the best feature selection methodology is beyond the scope of the present work).

2.5 Data Fusion

Two methodologies were tested for fusing multi-source data: a multiblock methodology represented by the concatenated method (Figure 4) (Campos *et al.*, 2017) and Bayesian fusion (Figure 5).

The concatenated method combines all data sets into an extended feature matrix and then trains a model with the entire concatenated set. In the present work, a preliminary stage of variable selection was made in each data set, so that only variables with potential predictive value are integrated in the concatenated method to estimate the toxin level.

On the other hand, Bayesian fusion combines information from different sources while explicitly considering their uncertainty. The predictions from each source are then combined to generate the final toxin level estimate. The uncertainty of each source determines the weights used for combining the predictions obtained from their respective models. The weights are updated dynamically, using the residuals from the last five observations (RMSEP).

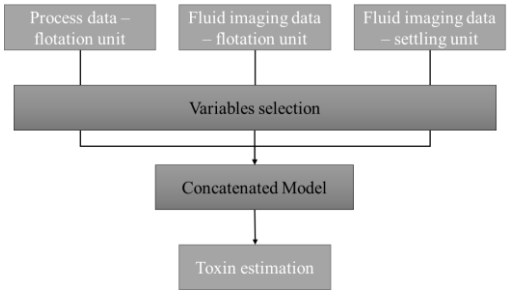

Figure 4 - Scheme for the concatenated method.

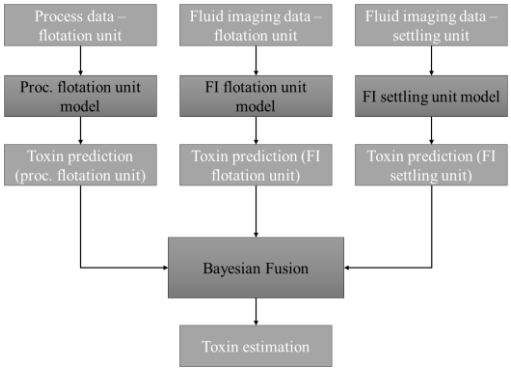

Figure 5 – Scheme of the Bayesian fusion method.

3. RESULTS

We now present the results obtained from implementing each stage of the proposed workflow described in the previous section.

3.1 Exploratory data analysis

In this stage, special focus was given to the analysis of the correlation structure of the data sources, as well as the relationship between individual features and toxin level. These aspects are related to the predictors' multicollinearity and sparsity issues, and therefore it is desirable to extract insights about them before constructing the models.

The Pearson's pair-wise correlation coefficient was computed for all data sets. The absolute value of the coefficients for the process variables from the flotation unit are presented in the heatmap shown in Figure 6, where it is possible to verify the existence of some blocks of highly correlated variables (e.g., flows and pump speeds in the flotation unit data). Collections of highly correlated variables were also found analyzing cross-correlation heatmaps for fluid imaging data (e.g., average geometrical features of the objects in the images, such as those related to area and diameter).

Predictor's sparsity, *i.e.*, the existence of a few strong predictors in the middle of many weak or noisy ones, was also

explored. In Figure 7, we present the *p*-value from the statistical test for the significance of the correlation coefficient between each process variable and the toxin level. Considering a significance level of 0.05, a *p*-value below this threshold indicates a significant correlation with the toxin. On the other hand, variables with *p*-values above 0.05 are weakly associated with the target response. A similar analysis was carried out for the image-based features from both the flotation and settling units.

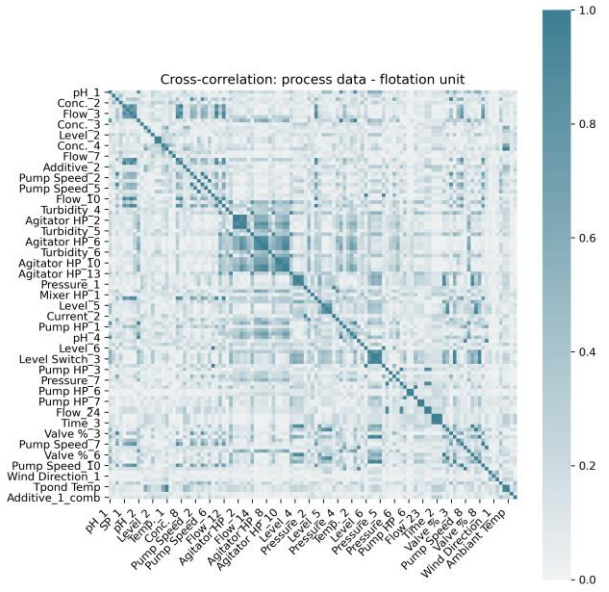

Figure 6 - Heatmap of absolute value of Pearson's correlation coefficient between predictors, process data from flotation unit.

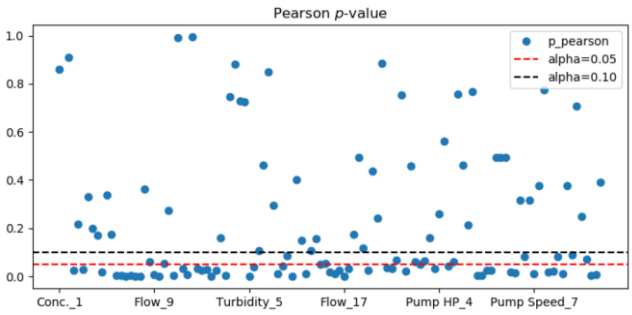

Figure 7 – P-values of the Pearson's correlation of process data from the flotation unit and the toxin level.

Then, PCA was applied to each data set separately. Results for two data sets are presented in Figure 8. The first two PCs capture between 25% to 35% of the overall variability, in each data set. Even though these values may seem low, they are common in industrial data, especially from biological systems where a large amount of variability is unique to each predictor and uncorrelated with other sources of variability. Even so, visual analysis of the first two principal components, PC 1 and PC 2, still conveys valuable information about the main patterns of variation in the data sets and the possible existence of clustered observations, trends, outliers, etc. Figure 8a presents the scatter plot between the scores of the first two PCs for the process data set, revealing the existence of two clusters, which are likely related to different operational conditions during the analysis period. Indeed, these two clusters correspond to a change of one piece of equipment in the process. However, the quality of the models built for each period did not show significant differences when compared to the model built on the entire period.

Figure 8b presents the scatter plot of the scores from the first two PCs of the image data from the settling unit (representing 33% of the original variation). As no obvious clusters are observed, the two operational periods are not considered separately in the subsequent analysis.

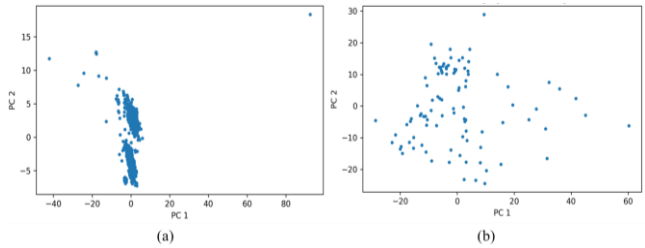

Figure 8 - Scatter plot of the scores for PC 1 and PC 2: a – process data from flotation unit, b – fluid imaging data from settling unit.

3.2 Model Development

The WWTP biological processes may present relevant dynamic modes. Thus, the use of past data may improve the model's predictive accuracy. In order to explore this possibility, lagged variables were included in the analysis as additional features (for process data from the flotation unit). Models were trained using EN and RF methods and compared with those obtained without the inclusion of lagged variables. Here, EN and RF were selected due to the embedded variables selection capabilities of the methodologies so that they could provide a good indication of the potential value of introducing lagged-variables in the analysis. (This does not imply however that EN and RF will be the selected models in the end.)

Table 1 - Median RMSECV for the considered lags (process data set from the flotation unit). Best results are highlighted.

Meth.	k	k-1	k-2	k-3	k-10	k-25	k-30
EN	1.954	1.036	1.056	1.047	1.093	0.995	0.957
RF	1.843	1.025	1.023	1.043	1.068	1.019	0.963

In the study conducted, lags of up to 30 days were considered based on the known dynamics of the biological system. Table 1 summarizes the results of model performance from cross-validation (RMSECV) using process data with lags for the EN and RF model, respectively. Here, *k-n* indicates that variables with lags from 0 up to *n* days back in the past, were also included for model development. Results revealed that an improvement was achieved after introducing just a one-day lag in the predictor; however, the best performance was obtained when lags up until 30 days were used. Thus, they included in the variable set for estimation of the toxin with the process data from the flotation unit (note that the final variable/lag selection will be done later on; this is just to define the past horizon to be included for variable selection and model training). In this way, the number of variables was increased from 115 to 3565 (115 × 31, from day k until day k-30).

Adding lagged variables to the image data sets from flotation and settling units raises implementation issues, because they are collected at irregular sampling rates. Therefore, information from the past were incorporated using a time windowed aggregation approach, where features were averaged over moving windows of 7, 14, 21, and 28 days. The cross-validated RMSE of models with different averaging windows are shown in Table 2.

Table 2 - Median RMSECV for aggregated data (imaging data from the flotation and settling units). Best results are highlighted.

Source	Meth.	D	7D av.	14D av.	21D av.	28D av.
Flotation unit	EN	1.092	**0.973**	1.044	1.038	1.046
	RF	1.017	0.991	1.033	**0.972**	0.974
Settling unit	EN	1.014	**0.981**	0.998	1.038	1.043
	RF	0.988	**0.945**	1.036	1.088	1.005

For image data from the flotation unit, EN model has the best performance with a 7 days average included in the predictors (see Table 2). In the case of RF, the optimal choice was 21 days. Thus, for this data set, the 7 and 21 days averaged data will be considered for toxin estimation. This implies doubling the number of variables (from 1302 to 2604), because features replicates in two aggregation windows are now considered.

Table 3 - Median RMSECV for the methods developed from a single source of data (models with "VS" prefix are built on selected variables subsets). Best results are highlighted.

	Process data flotation	Fluid imaging flotation	Fluid imaging settling
PLS	*	*	1.088
VS PLS	0.808	0.889	0.779
LASSO	1.024	1.033	1.061
VS LASSO	0.762	0.81	0.761
EN	1.021	0.937	0.981
VS EN	0.767	0.812	0.673
RF	0.956	0.896	0.948
VS RF	**0.681**	**0.756**	**0.635**
Boosting	0.957	0.893	0.952
VS Boosting	0.774	0.816	0.673
MLP	*	*	1.083
VS MLP	0.875	0.873	0.682

* Unreliable predictions were obtained in some cross-validation runs.

As for the image data from the settling unit (Table 2), EN and RF results are relatively consistent. The best results are obtained using a 7 days average and therefore these variables are included as additional predictors for building single-source models. Here, the original variables were transformed into the aggregated ones, so the number of variables remains the same (1302).

With the lagged structure defined, we now present the results obtained from all six estimation methods considered in this work for predicting toxin level using each data source: PLS, LASSO, EN, RF, Boosting, and MLP.

The results for model development without/with a variable selection step can be found in Table 3, where Random Forests with Variable Selection (VS RF) presents the best performance for all data sources.

The next step is to integrate/fuse all data sources together. Recall there are two approaches, concatenated and Bayesian fusion. VS RF was also used to build the concatenated model.

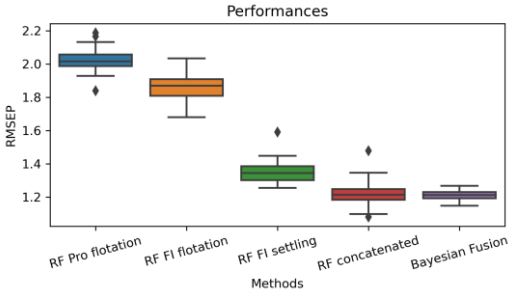

Figure 9 - Test results for the single-source (RF Pro Flotation - RF using the Process data from Flotation unit; RF FI flotation - RF derived from imaging data from the Flotation unit; RF FI settling – RF using data from settling unit) and multi-source (RF Concatenated and Bayesian Fusion).

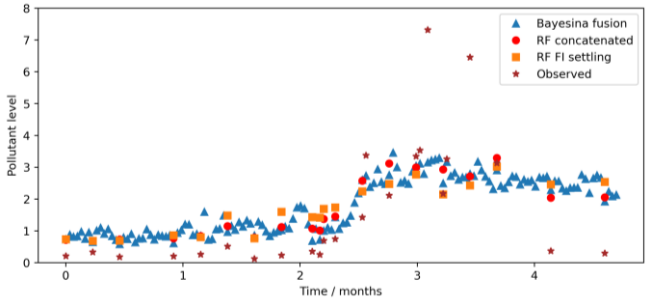

Figure 10 - Time series plot of observed and predicted toxin levels for various methods using fluid image data from settling unit (RF FI settling), concatenated RF, and Bayesian fusion.

Figure 9 presents the performance of the test data using three single-source and two multi-source methods. Multi-source methods perform better than single-source methods. Comparing the results from the two fusion strategies, the concatenated RF fusion method gives quite similar performance to the Bayesian fusion methodology. However, Bayesian fusion performance is apparently more stable, presenting less variability.

From Figure 9, it is also possible to verify that Bayesian fusion and the concatenated RF methods offer the best solution among those studied in this work. The Bayesian approach is also able to provide estimates as long as one source is available, without requiring any extra processing step of missing data imputation.

Figure 10 shows the observed and predicted values for the test set of the toxin as time series. There are 5 observations in test domain that neither the concatenated method nor the best single source models were able to predict, due to lack of imaging data from the settling unit.

4. DISCUSSION AND CONCLUSIONS

Accurate and timely estimation of the target effluent propery are important factors for the operation management of a WWTP. In this work, several data-driven methodologies were used to estimate the toxin level, based on data collected from different sources, with distinct structures (image/sensor data, regular/irregular sampling). Model performance was enhanced

through variable selection and machine learning techniques together with fusion schemes. As shown in the results section, the multi-source solutions generally show better performance than single-source models, due to their ability to integrate information from different sources and explore their complementary synergies.

While the concatenated and Bayesian fusion methods were found to perform similarly, the latter one does present some advantages. Its performance is more consistent and stable, and it considers the accuracy of each source (updated over time), in contrast to the concatenated approach. Even if the variables and blocks are scaled, this will not make any effect on the RF model. Another positive aspect of the Bayesian approach is that it does not require data from all sources to be simultaneously available. The last issue is important in practice, as process data is acquired on a daily basis whereas image data is only acquired 2-3 times per week.

The multi-rate structure, mentioned above, affects single source models (namely those based on image data from both flotation and settling units), as well as the concatenated fusion strategy. This issue is clearly visible in Figure 10, where the frequency of predicted values is higher for the Bayesian Fusion strategy. On the other hand, the single-source model based on process data from the flotation unit can provide predictions at the required frequency (daily), but its performance is comparatively worse (Figure 9).

In future work, alternative fusion methods will be explored. Also, approaches that simultaneously perform feature engineering and selection, such as Network-Induced Supervised Learning (Reis, 2013b; Reis, 2013a), will be explored. Finally, the problem of adjusting the additive dosing in the flotation unit will be considered.

REFERENCES

Alyannezhadi, M. M., Pouyan, A. A. and Abolghasemi, V. (2017). An efficient algorithm for multisensory data fusion under uncertainty condition. *Journal of Electrical Systems and Information Technology*, 4, 269–278.

Azimirad, E. and Haddadnia, J. (2015). The Comprehensive Review On JDL Model In Data Fusion Networks: Techniques and Methods. *International Journal of Computer Science and Information Security*, 13, 53–60.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.

Campos, M. P., Sousa, R., Pereira, A. C. and Reis, M. S. (2017). Advanced predictive methods for wine age prediction: Part II – A comparison study of multiblock regression approaches. *Talanta*, 171, 132–142.

Castanedo, F. (2013). A Review of Data Fusion Techniques. *The Scientific World Journal*, 2013, 1–19.

Cerqueira, V., Torgo, L. and Mozetic, I. (2019). Evaluating time series forecasting models: An empirical study on performance estimation methods. *arXiv:1905.11744 [cs, stat]* (on-line). http://arxiv.org/abs/1905.11744. Accessed 6 May 2020.

Diez-Olivan, A., Del Ser, J., Galar, D. and Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Information Fusion*, 50, 92–111.

Jackson, J. E. (1991). *A user's guide to principal components*. New York, Wiley.

Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. *Encyclopedia of Statistics in Behavioral Science*, 30, 487.

Kenett, R. S. and Shmueli, G. (2014). On information quality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177, 3–38.

Koskela, T., Lehtokangas, M., Saarinen, J. and Kaski, K. (1996). Time Series Prediction with Multilayer Perceptron, FIR and Elman Neural Networks. In: *In Proceedings of the World Congress on Neural Networks*. Press. 491–496.

Mitchell, H. B. (2012). *Data Fusion: Concepts and Ideas*. Springer, Berlin.

Reis, M. S. (2013a). Applications of a new empirical modelling framework for balancing model interpretation and prediction accuracy through the incorporation of clusters of functionally related variables. *Chemometrics and Intelligent Laboratory Systems*, 127, 7–16.

Reis, M. S. (2013b). Network-induced supervised learning: Network-induced classification (NI-C) and network-induced regression (NI-R). *AIChE Journal*, 59, 1570–1587.

Reis, M. S. and Kenett, R. (2018). Assessing the value of information of data-centric activities in the chemical processing industry 4.0. *AIChE Journal*, 64, 3868–3881.

Sansana, J., Rendall, R., Wang, Z., Chiang, L. H. and Reis, M. S. (2020). Sensor Fusion with Irregular Sampling and Varying Measurement Delays. *Industrial & Engineering Chemistry Research*, 1–13.

Sidek, O. and Quadri, S. A. (2012). A review of data fusion models and systems. *International Journal of Image and Data Fusion*, 3, 3–21.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Wang, Z. and Chiang, L. (2019). Monitoring Chemical Processes Using Judicious Fusion of Multi-Rate Sensor Data. *Sensors*, 19, 2240.

Wold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109–130.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.