Interval-Valued Data Learning for Robustness of Energy Recovery Systems

Rachid Ouaret, Pascal Floquet, Jean-Pierre Belaud, Stéphane Négny

Metin TÃ¼rkay Metin Türkay, Erdal Aydin (Eds.)
Proceedings of the 31th European Symposium on Computer Aided Process Engineering
June 6th to 9th, 2021, Istanbul, Turkey.

Interval-Valued Data Learning for Robustness of Energy Recovery Systems

Rachid Ouaret[a,*], Pascal Floquet [a], Jean-Pierre Belaud[a] and Stéphane Negny[a]

[a]*Chemical Engineering Laboratory, Université de Toulouse, CNRS, Toulouse, France, LGC UMR 5503, 4 allée Emile Monso, 31030 Toulouse*
rachid.ouaret@toulouse-inp.fr

Abstract

Modeling of energy recovery systems, Heat Exchanger Networks (HEN) as an example, are complex processes due to many parameters involved and the lack of knowledge on the impact of operational variables on the response. These variables are often affected by different types of uncertainty as to measurement errors, computation errors, or imprecision related to the underlying method. The uncertainty in the data may be treated by considering, rather than a single value for each variable, the interval of values in which it may fall, histograms, or other multivalued data: symbolic data. This work aims to use, when possible, the symbolic data analysis to adapt the classical mathematical HEN models. It deals with the study of continuous interval data through suitable Principal Component Analyses and Regression for two purposes: clustering exchanges (i) classification of exchangers to detect those impacted by uncertainty factors and (ii) evaluation of the relationship between the different process parameters (inlet temperature, heat transfer coefficient, etc.) on interval data. The new method has been tested on a real data set and the numerical results are reported. The symbolic approach provides a simple way to study a great number of scenarios.

Keywords: Machine Learning for Symbolic Data, Interval-valued data, Flexibility and Robustness, Heat Exchanger Networks (HEN)

1. Introduction

Energy saving is an important issue for both industries and society. In the industrial chemical process, Heat Exchanger Networks (HEN) are widely used techniques for reducing external heating and cooling utilities. Data generated by those complex systems has increased drastically over the past few years. Suppose we have 20 exchangers on which 4 variables are measured (2 input and 2 output hot and cold input streams). If we assume that we have hourly values of these variables for 1 year, then each exchanger is described by 35040 data. If, on the other hand, certain characteristics of the exchangers (let's say 3) vary over time, then a tabular representation conventionally used in data analysis would contain 2 102 400 values ($35040 \times 20 \times 3$). This is not very large compared to what could be provided by the industry in real-time.

Even such data is ubiquitous for a larger scale of HEN systems, data-oriented based approaches are often analyzed with simplified models by aggregating data. During this operation, some information on the variability aspects is lost. The robustness assessment of HEN is therefore affected since its flexibility is conditioned by the variability of the uncertain parameters. When the size n of entries (exchangers) and p number of variables (features) are very large, classical analysis can be problematic.

To address this problem, we use a more complex object description to capture the variability of measured parameters on each exchanger. When dealing with quantitative variables, complete information can be achieved by describing a set of statistical units in terms of interval data, histogram, ... rather than a single-valued variable. Mathematically, interval-valued data with measurements on p random variables, are p-dimensional hyperrectangles in $\mathbb{R}^p$. Such data need to be visualized, synthesized, and compared on factor spaces.

This paper is in a complementary perspective to Floquet et al's work (Floquet et al., 2016). They have initiated the robustness analysis of a simple exchange networks using interval arithmetic. They pointed out the "*butterfly effect*" of the alteration of characteristics of some heat exchangers on the operation of the HEN. Indeed, a maintenance operation of an exchanger, alter the variability of the other parameters of the network. Reducing fouling would imply variations in pressure and shift flows between parallel branches, all changing over time in a way that is difficult to predict (Macchietto et al., 2018). The purpose of this paper is to check whether there are groups of exchangers characterized by the same properties in terms of their responses to external fluctuations. Following the same idea (interval-valued data), the Symbolic Data Analysis (SDA) (Billard and Diday, 2006; Bock and Diday, 2012) is used to study how uncertainty in the output of a model can be apportioned to different sources of uncertainty.

2. Machine Learning for symbolic data

Figure 1 outlines the design and methodological scheme of the proposed method. This methodology draws on two main components: (B) Symbolic data through interval arithmetic, and (C) machine learning for symbolic data. Part A corresponds to the HEN model to be studied or simulated. The interval-valued data is then constructed in part (B).The core function of the proposed method included in this paper is part (C). The originality of this research lies in the combination of a traditional robustness analysis of HEN with SDA.

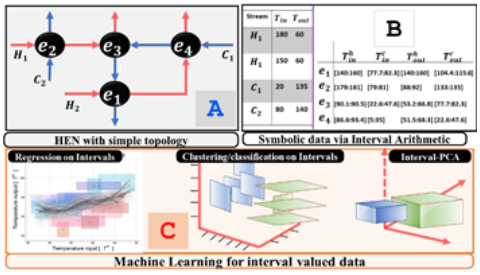

Figure 1: Overview of the SDA integration and data-driven processing for HEN robustness analysis.

2.1. Symbolic Data and statistics of interval-valued variables

In classical statistics, $n \times p$ data matrix $X = X_{ij}$ is defined between n individuals and p variables, where each cell (i, j) contains a **unique** value x_{ij}. A symbolic objects are more complex than a simple valued variable description, symbolic data can contain internal variation of the features representing *imprecise knoweledge* and can be structured. The symbolic analysis generalizes the classical data analysis, e.g. $x = c$, $c \in \mathbb{R}$ is equivalent to the symbolic interval $\xi = [c, c]$. A full conceptualization of symbolic objects can be found in Bock and Diday (2012). Let Ω be a set of individuals, $\mathscr{D}$ containing the descriptions of individuals and the descriptions of classes of individuals, a, a mapping defined from Ω into $\mathscr{D}$ which associates to each $\omega \in \Omega$ a description $d \in \mathscr{D}$ by using intervals, histograms, etc. More formally, symbolic object is a triplet $s = (a, \mathscr{R}, d)$ where $\mathscr{R}$ is a relation between descriptions, d is a description, and a is a mapping defined from Ω in $\mathscr{L}$ depending on $\mathscr{R}$

and d. For instance $\mathscr{L} = \{true, false\}$ or $\mathscr{L} = [0,1]$ and $\mathscr{R}$ may be one of the relations in $\{=, \equiv, \leq, \subseteq, \Longrightarrow\}$.

To illustrate this point with a simple HEN example, if the outlet temperature of two exchangers ω_1, ω_2 is given by $T_{out}(\omega_1) = 90^{\circ}$C, $T_{out}(\omega_2) = 120^{\circ}$C the description of the class $\mathscr{D} = (\omega_1, \omega_2)$ obtained by a generalization process can be $[90, 120]$. The symbolic object s is defined by a triple $s = (a, \mathscr{R}, d)$ where $d = [90, 120]$, $\mathscr{R} = {}' \in {}'$ and a is the mapping: $\Omega \rightarrow \{true, false\}$ such that a is the true value of $T_{out}(\omega) \mathscr{R} d$,written $a(\omega) = T_{out}(\omega) \in [90, 120]$. An individual ω is in the extent of s if and only if $a(\omega) = true$. Simple statistical descriptions (mean, variance, ...) for interval-valued variables have been defined in (Bertrand and Goupil, 2000). Let consider $Y_j \equiv Z$ be the j^{th} interval-valued random variable, and $Z(\omega_u) = [a_u, b_u]$ is a realization of Z for the observation ω_u over the observed interval $[a_u, b_u]$. The empirical distribution function, $F_Z(\xi)$, is the distribution function of a mixture of m distributions $\{Z(\omega_u), u = 1, 2, \ldots, m\}$. The central and dispersion parameters of a variable all derived from a strong assumption: the inherent fluctuation within random intervals and rectangles is uniformly distributed: $f(\xi) = \frac{1}{m} \sum_{u: \xi \in Z(u)} \left(\frac{1}{(b_u - a_u)} \right)$. The symbolic sample mean for interval-valued data is given by

$$\overline{Z} = \frac{1}{2m} \sum_{u} (b_u + a_u), \tag{1}$$

and the sample variance is given by

$$S^2 = \frac{1}{3m} \sum_{u} \left(b_u^2 + b a_u + a_u^2 \right) - \frac{1}{4m^2} \left[\sum_{u} (b_u + a_u) \right]^2. \tag{2}$$

2.2. PCA for interval-valued data

A principal component analysis is designed to reduce p-dimensional observations into s-dimensional components (where $s \ll p$) in an interpretable way, such that most of the information in the data is preserved. Let $i = 1, \ldots, n$ denote n objects (exchangers) described by p features (or variable) $Y_1, \ldots, Y_p$ (temperatures, ...). The symbolic data matrix used for interval PCA is given by $\underline{X} = \begin{pmatrix} \xi_{11} & \cdots & \xi_{11} \\ \vdots & \ddots & \vdots \\ \xi_{n1} & \cdots & \xi_{np} \end{pmatrix}$, where $\xi_{ij} = \left[\underline{x_{ij}}, \overline{x_{ij}} \right]$ is the interval of possible values of variable j for the exchager i, and the symbolic data vector can be denoted by $\boldsymbol{x}_i = (\xi_{i1}, \ldots, \xi_{ip}) = \left(\left[\underline{x_{i1}}, \overline{x_{i1}} \right], \ldots, \left[\underline{x_{ip}}, \overline{x_{ip}} \right] \right)$. The data point is represented in $\mathbb{R}^p$ space by hyperrectangles R_i with 2^p vertices. There are mainly two methods to solve the algebraic mapping to lower dimension: Vertices and Centers methods. To find the factorial axes for Centers method, a classical PCA is applied to the centers $c_i \in \mathbb{R}^p$ of the n hyperrectangles R_i. The coordinates of i^{th} center c_i is denoted by x_{ij}^c, where x_{ij}^c is computed for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. From that case, the centers of hyperrectangles in its rows is an $n \times p$ data matrix, denoted by $\underset{\sim}{X}$ and the j^{th} column of $\underset{\sim}{X}$ is denoted by the feature Y_j^c. The interval principal components values are obtained by computing first the υ^{th}principal component of the center c_i is given by

$$\psi_{i\upsilon}^c = \sum_{j=1}^{p} \left(x_{ij}^c - \overline{x_{ij}^c} \right) \cdot \eta_{j\upsilon} \tag{3}$$

where $\eta_{j\upsilon} = (\eta_{1\upsilon}, \ldots, \eta_{p\upsilon})$ is the υ^{th}eigenvector of S (sample covariance matrix associated with the dataset). It is possible to find an $\left[\underline{\psi_{i\upsilon}}, \overline{\psi_{i\upsilon}}\right]$ of the possible values of υ^{th}principal component $\psi^c_{i\upsilon}$ of c_i (Cazes et al., 1997). For the object i :

$$\overline{\psi_{i\upsilon}} = \sum_{j=1}^{p} \max_{\underline{x_{ij}} \le x^r_{ij} \le \overline{x_{ij}}} \left(x^r_{ij} - \overline{x^c_j}\right) \cdot \eta_{j\upsilon} \tag{4}$$

$$\underline{\psi_{i\upsilon}} = \sum_{j=1}^{p} \min_{\underline{x_{ij}} \le x^r_{ij} \le \overline{x_{ij}}} \left(x^r_{ij} - \overline{x^c_j}\right) \cdot \eta_{j\upsilon} \tag{5}$$

3. Application to HEN data

3.1. Data simulation design and preliminary analysis

The starting point for this application is the data table from step B in Figure 1, which is the result of HEN model analysis by interval arithmetic. In order to evaluate the behavior of the exchanges on a more complex network, simulated data on the initial table were undertaken and the simulation procedure is as follows:

1. For each variable ($T^h_{in}\left[\min^h_{in}, \max^h_{in}\right]$, $T^c_{in}\left[\min^c_{in}, \max^c_{in}\right]$, $T^h_{out}\left[\min^h_{out}, \max^h_{out}\right]$, $T^c_{out}\left[\min^c_{out}, \max^c_{out}\right]$) and for all exchagers, we look for the minimum and the maximum.
2. Simulate 40 points (exchangers) using random variable from the uniform distribution. For example, for the exchanger No. 30 (E30), the minimum input temperature for the cold stream is obtained by the following probabilistic simulation scheme: $T^c_{in}(E30) \sim \mathscr{U}\left(\min\left(\min^c_{in}\right), \max\left(\min^c_{in}\right)\right)$. Retrieve the interval matrix of the 4 variables (T^h_{in},T^c_{in},T^h_{out},T^c_{out}).

Statistical description for the interval-valued data of the initial matrix and the simulated one, using equations 1 and 2, is presented in the following Table 1:

Table 1: Descriptive statistics (mean and standard deviation *s.d*) for the interval-valued data of HEN

		$\left[T^h_{in}\right]$	$[T^c_{in}]$	$\left[T^h_{out}\right]$	$[T^c_{out}]$
Initial data	Mean	[124.20:131.42]	[46.32:61.30]	[70.32:84.30]	[84.68:94.95]
	sd	[44.47:45.42]	[38.25:23.67]	[20.81:12.03]	[47.34:38.20]
Simulation	Mean	[112.14:134.02]	[45.34:42.93]	[68.95:81.42]	[76.98:97.71]
	sd	[20.77:28.33]	[22.91:23.85]	[11.38:9.02]	[33.34:25.32]

3.2. Principal Component Analysis for interval-valued data

The interpretation of the position of the interval-valued data in the principal plane is the same as in the classical principal component analysis situation. In Figure 2, we show the

results with respect to the first three axes, achieved by the Symbolic PCA using 4 and 5 (centers method). Notice that the 61% of the total inertia is explained by the first two axes in the case of simulation (40 exchangers), and 98.5% of in the case of initial data (4 exchangers). In Figure 2, closeness among clusters exchanger mainly influenced by the same descriptors.

Only one group consisting of exchangers 1 and 3 can be identified. The constituent elements of this cluster are influenced by the same main factors. Exchangers 2 and 4 are detached from the cluster from a 3D perspective. This observation is in line with the conclusions of the initial study (Floquet et al., 2016).

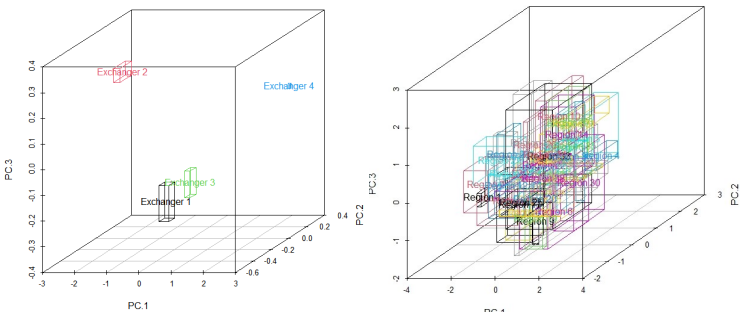

Figure 2: Principal 3D-space with data of interval type of HEN. Factorial for 4 Exchangers (left) and for 40 Exchangers (right).

For simulated data, the same analysis can be achieved using additional information, and at this stage, it is also difficult to give an interpretation of the similarity in size and shape among exchangers. In order to shed new light on the variability of outlet temperatures as a function of input streams, we propose a linear regression on interval values.

3.3. Regression and prediction for interval-valued data

The model proposed here is based on the classical formation of HEN models. However, a notable difference lies in the integration of symbolic objects in the model. To understand how the input (cold and hot) temperature stream affects the output temperature, we propose a regression model for interval-valued temperature. For a hot stream:

$$T^h_{out}\left[\min^h_{out},\max^h_{out}\right] = \beta_0 + \beta_1 \cdot T^h_{in}\left[\min^h_{in},\max^h_{in}\right] + \beta_2 \cdot T^c_{in}\left[\min^c_{out},\max^c_{out}\right] + \varepsilon \quad (6)$$

The following results are obtained using the simulated data in the Table 2.

Table 2: Results of interval-valued data regression

	Dependent variable	
	$T^h_{out}\left[\min^h_{out},\max^h_{out}\right]$	$T^c_{out}\left[\min^c_{out},\max^c_{out}\right]$
$\widehat{\beta_0}$	63.22	37.46
$\widehat{\beta_1}$	0.0046	0.17
$\widehat{\beta_2}$	0.244	0.6

To evaluate this approach, two datasets were used: *(i) training dataset* which is a set of the first 35 exchangers used to fit the parameters of the model 6 and *(ii) test dataset*

which is a set of the rest 5 exchangers used to provide an unbiased evaluation of the estimated model. The Table 3 shows the results of the predictions using the fitted model and corresponding data test. The size of the coefficient for each independent variable ($T_{in}^{h}\left[\min_{in}^{h}, \max_{in}^{h}\right]$ and $T_{out}^{c}\left[\min_{out}^{c}, \max_{out}^{c}\right]$) gives the size of the effect that variable is having on the dependent variable T_{out}^{h} and T_{out}^{c}. The estimated coefficients tells how much the output stream is expected to increase when that inputs streams (hot and cold) increases by one. For example the coefficients for the model 6, T_{in}^{h} differed by $1°C$ (and T_{in}^{c} did not differ) T_{out}^{c} will differ by $0.17°C$ units, on average. The estimated intercept $\widehat{\beta}_0$, is the expected mean value of T_{out}^{h} and T_{out}^{c} when all inputs are 0. In our experiments, the T_{in}^{h} and T_{out}^{c} never comes close to 0, then intercept has no meaningful interpretation.

Table 3: The prediction results

Exchangers	**Predictions**		**Test**	
	$\left[\widehat{T_{out}^{h}}\right]$	$\left[\widehat{T_{out}^{c}}\right]$	$\left[T_{out}^{h}\right]$	$\left[T_{out}^{c}\right]$
E36	[64.21:75.01]	[61.41:80.81]	[73.61:67.64]	[39.89:119.97]
E37	[73.25:82.21]	[76.12:104.68]	[69.72:89.61]	[115.31:89.80]
E38	[68.06:76.77]	[63.75:87.63]	[57.47:91.65]	[46.96:104.48]
E39	[64.73:77.73]	[69.71:80.77]	[88.25:85.72]	[113.54:66.64]
E40	[60.93:72.54]	[52.76:75.85]	[52.70:88.04]	[83.85:100.02]

4. Conclusion

Symbolic Data Analysis (SDA) extends statistics and multivariate data analysis to deal with data structured in a distributional form with complex internal variations. In this paper, comprehensive modeling via SDA has been presented that moves substantially beyond the traditional modeling in HEN robustness analysis. It proposed some new approaches which intended to redefine the robustness of HEN based on interval data. First, the detection of Exchangers cluster which would be affected by common factors has been modeled by Symbolic Principal Component Analysis. Second, the relation between inherent variation, expressed by intervals, of input and output temperature has been modeled using the linear regression method for interval-valued variables. Future research should include histogram valued-data.

References

P. Bertrand, F. Goupil, 2000. Descriptive statistics for symbolic data. In: Analysis of symbolic data. Springer, pp. 106–124.

L. Billard, E. Diday, 2006. Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley. Chichester.

H.-H. Bock, E. Diday, 2012. Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer Science & Business Media.

P. Cazes, A. Chouakria, E. Diday, Y. Schektman, 1997. Extension de l'analyse en composantes principales à des données de type intervalle. Revue de Statistique appliquée 45 (3), 5–24.

P. Floquet, G. Hétreux, R. Hétreux, L. Payet, 2016. Analysis of operational heat exchanger network robustness via interval arithmetic. In: Computer Aided Chemical Engineering. Vol. 38. Elsevier, pp. 1401–1406.

S. Macchietto, F. Coletti, E. D. Bejarano, 2018. Energy recovery in heat exchanger networks in a dynamic, big-data world: Design, monitoring, diagnosis and operation. In: Computer Aided Chemical Engineering. Vol. 44. Elsevier, pp. 1147–1152.