Data Reconciliation under Fuzzy Constraints in Material Flow Analysis

Didier Dubois[1], Hélène Fargier[1] Dominique Guyonnet[2]

[1]IRIT, CNRS & Université de Toulouse, France
[2]BRGM-ENAG, Orléans, France

Abstract

Data reconciliation consists in modifying noisy or unreliable data in order to make them consistent with a mathematical model (herein a material flow network). The conventional approach relies on least squares minimization. Here we show that the setting of fuzzy sets provides a generalized approach that is more flexible and less dependent on oftentimes debatable probabilistic justifications. Moreover the proposed setting also encompasses constraint-based formulations using intervals.

Keywords: Material flow analysis, data reconciliation, least squares, fuzzy constraints

1. Introduction

Material flow analysis consists in calculating the quantities of a certain product transiting a network of local entities referred to as processes, considering input and output flows and including the presence of material stocks. The unknowns to be determined are the values of the flows and stocks. The basic principle that provides constraints on the flows is that what goes into a process must come out, up to the variations of stock. Flows and stocks must be balanced, through a set of linear equations. The mass-balance equation relative to a process with n flows in, k flows out and a stock level s is written:

$$\sum_{i=1}^{n} IN_i = \sum_{j=1}^{k} OUT_j + \Delta s \qquad (1)$$

where Δs is the amount of stock variation (positive if $\sum_{j=1}^{k} OUT_j < \sum_{i=1}^{n} IN_i$ and negative otherwise).

Such flow balancing equations in a process network define a linear system of the form $Ay^t = B$, y being the vector of N flows and stock variations. In order to evaluate balanced flows and stocks, data is collected regarding the material flow transiting the network and missing flow or stock variation values are calculated. But this task may face several difficulties:

- There may not be sufficient information to determine all the missing flows or stock variations.
- There may be on the contrary too much information available and the system of balance equations may not have any solution.
- The available information is often not sufficiently reliable and precise.

In this paper we address the second and third cases. If the data are in conflict with the mass-balance equations, it may be because they are erroneous and should be corrected: this is the problem of data reconciliation, a well-known problem in statistics ever since its origins. As early as the end of the XVIIIth century, this question was addressed using the method of least-squares. It still is the case today, but the justification is statistical and usually based on the Central Limit Theorem and the principle of maximum likelihood. In this paper, we examine the limitations of this approach and propose a preliminary discussion on alternative approaches that take into account more explicitly data uncertainty, using intervals or fuzzy intervals. Following the latter approaches, the problem can then be solved using crisp or fuzzy linear programming. We outline a general framework based on fuzzy intervals, understood as flexible constraints, that encompasses the least-squares method.

2. Data reconciliation

Data reconciliation consists in modifying measured or estimated quantities in order to balance the mass flows in a given network. The vector of flows y is subdivided into two sub-vectors x and u, i.e., k informed quantities x_i and $N - k$ totally unknown quantities u_j, to be determined. We denote by $\hat{x}$ the vector of available measurements $\hat{x}_i$. In general, the system $A(xu)^t = B$ has no solution such that $x = \hat{x}$. This absence of solution is assumed to be due to measurement errors or information defaults. The problem to be solved is to modify x, while remaining as close as possible to $\hat{x}$, such that the mass balance equations $Ay^t = B$, with $y = (xu)$, are satisfied.

2.1. The least-squares approach

The traditional approach to data reconciliation [16] considers that data come from measurements, and measurement errors follow a Gaussian distribution with zero average and a diagonal covariance matrix. The precision of each measurement $\hat{x}_i$, understood as a mean value, is characterized by its standard deviation σ_i. Data reconciliation becomes a prob-

lem of optimization under linear constraints. In the simplest case (assuming no u):

$$\text{Find } x \text{ minimizing} \sum_{i=1}^{k} w_i(x_i - \hat{x}_i)^2$$
$$\text{such that } Ax^t = b$$

It is the method of weighted least-squares used in many data reconciliation packages such as STAN [2]. The solution is known to be of the form [16]:

$$x^* = \hat{x} - W^{-1}A^t(AW^{-1}A^t)^{-1}A(\hat{x} - b),$$

where W is a diagonal matrix containing terms $1/w_i$. Weights are often of the form $w_i = (\sigma_i)^{-2}$.

Such packages sometimes also reconcile variances as explained in [16]. It assumes that the vector of estimated values $\hat{x}$ has a multivariate normal distribution characterized, by a covariance matrix C generalizing W, whose diagonal contains the variances σ_i^2. The balance flows being linear, the reconciled values x^* depend to the estimated values via a linear transformation, say $x^* = B\hat{x}$ hence also have a normal distribution. The covariance matrix of x^* is then of the form $C^* = BCB^t$.

2.2. Limitations of the approach

The method of least-squares is often justified based on the principle of maximum likelihood, applied to normal distributions, in turn justified by the Central Limit Theorem (CLT). If p_i is the probability density function associated with error $\epsilon_i = x_i - \hat{x}_i$, the maximum likelihood is calculated on the function $L(x) = \prod_{i=1}^{k} p_i(x_i - \hat{x}_i)$. If the p_i's are normal with average 0 and standard deviation σ_i, then $p_i(x_i - \hat{x}_i)$ is proportional to $e^{-\frac{(x_i-\hat{x}_i)^2}{\sigma_i^2}}$. As a consequence, the maximum of $L(x)$ coincides with the solution to the least squares method. The Gaussian assumption seems to be made because of the popularity of Gauss law (that computes an approximation of the standard deviation of a function $f(x_1, \ldots x_k)$ in the vicinity of a measurement point based on the linear part of its Taylor expansion). The universal character of this approach, albeit reasonable in certain situations, is nevertheless dubious:

- It is not consistent with the history of statistics [18]. The least-squares method, developed by Legendre (1805) and Gauss (end of XVII-Ith century), was discovered prior to the CLT, as is the Gauss function. Invented precisely to solve a problem of data reconciliation in astronomy, the least squares method sounded natural since it was in accordance with the Euclidean distance. Moreover, it led to solutions that could be calculated analytically and it could justify the use of the average in the estimation of quantities based on several independent measures. The normal law was discovered by Gauss as the only error function that was compatible with the average estimator. However, CLT is a mathematical result obtained independently by Laplace, who later on made the connection between his mathematical result and the least squares method.
- The CLT presupposes a statistical process with a finite average E and standard deviation σ. In this case, the average of n random variables v_i has standard deviation $\sigma/\sqrt{n}$ and the distribution of the average $\frac{\sum_{i=1}^{n} v_i - nE}{\sqrt{n}}$ is asymptotically Gaussian as n increases. The fundamental hypothesis behind the normal distribution is the existence of a finite σ. In practice, this implies that for N observations a_i of v, the empirical variance $msd = \frac{2\sum_{i<j}(a_i - a_j)^2}{N(N-1)}$ remains bounded as N increases, which is neither always true nor easily verifiable.
- The Gaussian hypothesis is only valid in the case of an unbounded random variable. If v_i is positive or bounded, assuming that the quantity $E_n = \frac{\sum_{i=1}^{n} v_i}{n}$ assymptotically follows a normal distribution with standard deviation $\sigma/\sqrt{n}$ is an approximation that may be useful in practice but does not constitute a general principle.

Based on the remarks above, it is natural to look for alternative methods for reconciling data that do not come from a statistical measurement process. A first alternative consists in representing error-tainted data by means of intervals and checking the compatibility between these intervals.

3. Interval reconciliation

In practice, information on mass flows is seldom precise: the data-gathering process often relies on subjective expert knowledge or on scarce measurements published in various documents that moreover might be obsolete. Each flow value provided by a source can be more safely represented by an interval $\hat{X}_i$, which in a first stance, can be considered as encompassing the actual flow value: of course, the less precise the available information, the wider the interval. Missing values u_i can also be taken into account: we then select as its attached interval the domain of possible values of the corresponding parameter (for example, the unknown grade of an ore extracted from a mine and sent to the treatment plant can, by default, be represented by the interval [0 ,100]%). In the least-squares approach to data reconciliation, we use weights to reflect the assumed variance of a Gaussian phenomenon; if such information on variances σ_i^2 are available, we can set $\hat{X}_i = [\hat{x}_i - 3\sigma_i, \hat{x}_i + 3\sigma_i]$, as the distribution of x_i is often assumed to be Gaussian. Thus each of the N variables y_i of the vector $y = xu$ is delimited a priori by an interval $\hat{Y}_i$.

The representation of flow data by intervals leads

us to consider the reconciliation as a problem of constraint satisfaction; the mass balance equations must be satisfied for flux and stock values that lie within the specified intervals - or, to be more precise, we can restrain these intervals to the sole values that are compatible with the balancing model given the possible values of other variables. Formally, the reconciliation problem can be expressed as follows:

For each $i = 1, \dots N$, find the smallest and largest values for y_i, such that :

$$Ay^t = B$$
$$y_i \in \hat{Y}_i, i = 1, \dots, N$$

The calculation of consistent minimum and maximum values of y_i is sufficient: since all the equations are linear, we can show that if there exists two flow vectors y and y', each being a solution to the above system of equations, then any vector v lying between y and y' componentwise is a solution of the system of equations $Ay^t = B$. The problem can of course be solved using linear programming.

Due to the linearity of the constraints, it may also be solved by methods based on interval propagation. For each variable y_i, equation j of the system $Ay^t = B$ can be expressed as

$$y_i = \frac{\sum_{k \neq i} b_j - a_{jk} y_k}{a_{ji}}, i = 1, \dots, N.$$

We can then project this constraint on y_i and find the possible values of y_i consistent with it. Due to the m linear constraints, the values of y_i can be restricted to lie in the interval:

$$Y_i = \hat{Y}_i \cap (\cap_{j=1\dots,m} \frac{\sum_{k \neq i} b_j - a_{jk} \hat{Y}_k}{a_{ji}}),$$

where $\frac{\sum_{k \neq i} b_j - a_{jk} \hat{Y}_k}{a_{ji}}$ is calculated according to the laws of interval arithmetic [12]; if the new interval of possible values of y_i is more precise ($Y_i \subset \hat{Y}_i$), it is in turn propagated to the other variables. This procedure, known as "arc consistency", is iterated until intervals are stabilized; it converges within a finite number of steps to a unique set of intervals [13] (for additional details on interval propagation techniques, see [1, 11]). Note that contrary to the statistical approach, here the model constraints and the imprecise data are handled on a par.

4. Modeling data using fuzzy intervals

In the framework of measurement problems, Mauris [14] has suggested that in the case of competing error functions (empirical probability distributions $p_i, i = 1 \dots k$), one may refrain from choosing one of them and consider a family of probabilities $\mathcal{P}$ instead, to represent our knowledge about x, where $p_i \in \mathcal{P}, \forall i$. In general such a representation can be extremely complex. For instance, $\mathcal{P}$ should be convex, typically the convex hull of $\{p_i, i = 1 \dots k\}$. However a very simple and convenient representation is via possibility distributions having the shape of a fuzzy interval [4].

A possibility distribution is a mapping $\pi : \mathbb{R} \to [0, 1]$ such that $\pi(r^*) = 1$ for some $r^* \in \mathbb{R}$. It represents the current information on a quantity x. The idea is that $\pi(r) = 0$ if and only if $x = r$ is impossible, while $\pi(r) = 1$ if $x = r$ is a totally normal, expected, unsurprizing value. One rationale for this framework is that the set $I_\alpha = \{r, \pi(r) \geq \alpha\}$ (α-cut) contains x with level of confidence $1 - \alpha$, that can be interpreted as a lower probability bound. In particular, it is sure that $x \in \{r, \pi(r) > 0\} = S(\pi)$, the support of the possibility distribution.

A fuzzy interval is a possibility distribution whose α-cuts I_α are closed intervals. They form a nested family of intervals containing the core $C(\pi) = \{r, \pi(r) \geq \alpha\} = 1$ and contained in the support. The simplest representation of a fuzzy interval is a trapezoid defined by the core and the support. Note that this format is very convenient to gather information from experts in the form of confidence intervals.

Given a possibility distribution π, the degree of possibility of an event A is $\Pi(A) = \sup_{r \in A} \pi(r)$. The degree of certainty of event A is $N(A) = 1 - \Pi(A^c)$, where A^c is the complement of A. A possibility distribution can be viewed as encoding a convex probability family $\mathcal{P}(\pi) = \{P, P(A) \geq N(A), \forall A \text{ measurable}\}$ (see [4] for references). Functions Π and N can be shown to compute exact probability bounds in the sense that

$$\Pi(A) = \sup_{P \in \mathcal{P}(\pi)} P(A) \quad \text{and} \quad N(A) = \inf_{P \in \mathcal{P}(\pi)} P(A).$$

When several error functions are possible, one may choose to represent them by a possibility distribution that encompasses them. This is the idea developed by Mauris [14]. Probabilistic inequalities is one example. For instance, knowing the mean value and the standard deviation of a random quantity, Chebyshev inequality gives a possibility distribution that encompasses all probability distributions having such characteristics [7]. Gauss inequality also provides such possibility distributions encompassing probability distributions with fixed mode and standard deviation (see [15]). It yields a triangular (bounded) fuzzy interval if probability distributions have bounded support. Hence a possibility distribution may account for incomplete statistical data.

Conversely, if an expert provides a probability distribution that represents subjective belief, it is possible to reconstruct a possibility distribution that fits the Laplace principle of indifference. When the available knowledge is an interval $[a, b]$, and the expert is forced to propose a probability distribution, the most likely proposal is a uniform distribution over $[a, b]$ due to symmetry. If the available knowledge is a possibility distribution π, this symmetry argument leads to replace π by a probability

distribution constructed by (i) picking at random a threshold $\alpha \in [0,1]$ and (ii) a number at random in the α-cut I_α of π. One may argue that we should bet on the basis of this probability function in the absence of any other information. Conversely, a subjective probability provided by an expert can be represented by the (unique) possibility distribution that would yield this probability distribution using this two-stepped random Monte-Carlo process [10].

In summary, fuzzy intervals, and specifically triangular or trapezoidal possibility distributions, may account for various kinds of uncertain information.

5. Fuzzy interval reconciliation

The interval approach does not yield the same type of answer as the least-squares method because it provides only intervals rather than precise values. Such intervals can be compared to reconciled variances provided by current software for MFA and data reconciliation like STAN [2]. A natural way to obtain both reconciled values and intervals is to tolerate a certain level of flexibility on the flow estimates using the notion of fuzzy interval: the more-or-less possible values of each flow or stock y_i will be limited by a fuzzy interval $\tilde{Y}_i$. For some of these quantities, these constraints will be satisfied to a certain degree, rather than simply either satisfied or violated. The problem of searching for a possible solution then becomes an optimization problem - we seek an optimal position within all the (fuzzy) intervals of possible values. If no solution provides entire satisfaction for all intervals, some will be relaxed if necessary [5].

5.1. A general framework

In this approach, the linear equations describing the material flow for each process are considered as integrity constraints that must necessarily be satisfied, but the information relative to possible values of each flow or stock quantity y_i is represented, not as a strict interval $\hat{Y}_i$ but as a fuzzy interval $\tilde{Y}_i$, interpreted as a possibility distribution π_i. This interval may still coincide with the domain of the quantity, in the case of total ignorance.

An assignment y for all y_i is possible, provided it satisfies all the constraints. In other words, the degree of plausibility of an assignment $y = xu$ can be obtained by a conjunctive aggregation of the local satisfaction degrees. Namely it is $\star_{i=1}^{N}\pi(y_i)$ if y satisfies the integrity constraints, and 0 otherwise - the operation $\star$ being associative, commutative and increasing on $[0,1]$ (a t-norm). We may then calculate the most plausible reconciled vectors and the associated degree of possibility by solving the following problem:

Find the values $y = xu$ that maximize:

$$\pi_\star(y) = \star_{i=1}^{N}\pi_i(y_i) \quad \text{such that } Ay^t = B$$

Rather than providing the user with one amongst several optimal solutions, it is often more informative to have reconciled flows in the form of fuzzy intervals obtained by projection on the domain of each y_i:

$$\max_{y \text{ s.t. } y_i = v \text{ and } Ay^t = B} \star_{j=1}^{N}\pi_j(y_j)$$

The operator models the fact that the y_i must be placed as close as possible to the cores of the fuzzy intervals - to their center if we use triangular (or trapezoidal) representations. In the case of "classical" intervals, i.e. when degrees of membership to the $\tilde{Y}_i$'s are 0 or 1, the $\star$ operator performs a simple conjunction and we come back to the formulation of Section 3. The cases $\star = \min$ and $\star =$ product constitute two basic modeling choices. We may also use the Łukasiewicz t-norm : $\max(0, a+b-1)$, which eliminates as impossible some vectors y that have albeit positive but too low scores. The first operator, the minimum, nevertheless presents advantages from a computation viewpoint: it allows the use of tools from linear programming or interval propagation.

5.2. The max-min approach

The optimization problem to be solved takes the form:

Find y^* that maximizes

$$\pi_{\min}(y) = \min_{i=1}^{N}\pi_i(y_i) \quad \text{where } Ay^t = B$$

with $Ay^t = B$. Let $\alpha^* = \min_{i=1}^{N}\pi_i(y_i^*)$ where y^* is an optimal solution. This implies that we cannot aim at a plausibility value $\alpha > \alpha^*$, since there will be no simultaneous choice of the y_i in the α-cuts of $\tilde{Y}_i$ that will form a consistent vector in the sense of the network defined by $Ay^t = B$, whereas there exists at least one consistent assignment of flows at the level α^*. Note that there may exist several solutions that allow a level of satisfaction α^*.

Once α^* is known, we can assign to each flow an interval of optimal values $(\tilde{Y}_i)_{\alpha^*} = \{y_i : \pi_i(y_i) \geq \alpha^*\}$ by solving for each y_i the following interval reconciliation problem: Find the minimum (resp. maximum) values of y_i such that $Ay^t = B$ and:

$$\pi_j(y_j) \geq \alpha^*, j = 1, \ldots, N.$$

The fuzzy reconciliation problem fails if $\alpha^* = 0$. The optimal supports of the optimal fuzzy intervals containing the y_i's can be obtained if we use the supports of the $\tilde{Y}_i$ in the procedure of the previous section. This program contains on the one hand the mass flow model $Ay^t = B$ which, as seen previously, is linear, and then we force the y_i to belong to the supports $[\underline{s}_i, \overline{s}_i], i = 1, \ldots N$ of the fuzzy intervals $\tilde{Y}_i$'s.

5.3. Resolution methods

From a technical standpoint, the fuzzy interval reconciliation problem can be solved using three alternative approaches:

Using a fuzzy interval propagation algorithm As in the crisp case, fuzzy intervals of possible values $\tilde{Y}_i$ can be improved by projecting the fuzzy domains of other variables over the domain of y_i via the balancing equations:

$$\tilde{Y}_i' = \tilde{Y}_i \cap (\cap_{j=1\ldots,m} \frac{\sum_{k\neq i} b_j - a_{jk}\tilde{Y}_k}{a_{ji}}),$$

where $\frac{\sum_{k\neq i} b_j - a_{jk}\tilde{Y}_k}{a_{ji}}$ is a fuzzy interval $\tilde{A}_j$ that can be easily obtained by means of fuzzy interval arithmetics [8] since equations are linear. Note that $\tilde{Y}_i \cap (\cap_{j=1\ldots,m}\tilde{A}_j)$ has possibility distribution $\pi_i' = \min(\pi_i, \min_{j=1\ldots,m}\pi_{\tilde{A}_j})$.

The propagation algorithm iterates these updates by propagating the new fuzzy intervals on all the neighboring y_i's, until their domains no longer evolve. This procedure presupposes efficient fuzzy interval representation schemes must be used. Typically we should use piecewise linear fuzzy intervals [17] including subnormalized ones. Eventually, optimal (maximally precise) fuzzy intervals $\tilde{Y_i}^*$ are obtained as resulting fuzzy domains of the reconciled flows. These fuzzy domains may be subnormalized: at least one of them has height $h_i = \sup_{y_i}\pi_i^*(y_i) = \alpha^*$ that may be less than 1 and may contain a single value. However the heights h_j of other optimal fuzzy intervals may be greater that α^*. Their h_j-cuts contain the optimal values y_i^*. However, this method will only provide the fuzzy intervals with possibility distributions $\min(\pi_i, \alpha^*)$ since fuzzy arithmetic methods applied to fuzzy intervals of various heights only preserve the least height [8].

Using α-cuts In order to take advantage of the calculation power of modern linear programming packages, a simple solution is to proceed by dichotomy on the α-cuts of the fuzzy intervals : once each $\tilde{Y}_i$ is cut at a given level α, we obtain a system of equations as in Section 3, replacing $\hat{Y}_i$ by the interval $(\tilde{Y}_i)_\alpha$; this system can therefore be solved by calling an efficient linear programming solver. If the solver finds a solution, the level α is increased; if not, i.e., if it detects an inconsistency in the system of equations, the value α is decreased, etc. until the maximum value α^* is obtained with sufficient precision, along with the corresponding intervals $(\tilde{Y}_i)_{\alpha^*}$.

Using fuzzy linear programming When the fuzzy intervals are triangular or trapezoidal (or even homothetic, as in the case of L-R fuzzy numbers), it is possible to write a (classical) linear program in order to obtain the value of α^*, then obtain the optimal ranges $(\tilde{Y}_i)_{\alpha^*}$'s for the reconciled flows. It is necessary to model the fact that the global degree of plausibility of the optimal reconciled values is the least among the local degrees of possibility, i.e., we should maximize a value less than all the $\pi_i(y_i)$, hence we should write N constraints $\alpha \leq \pi_i(y_i), i = 1, \ldots, N$. When the original fuzzy intervals are triangular with core $\hat{y}_i$, each constraint is written in the form of two linear inequalities, one for each side of the fuzzy intervals. All these equations being linear, we can then use a linear solver to maximize the value α such that:

$$\begin{aligned} Ay^t &= B \\ \underline{s}_i \leq y_i &\leq \overline{s}_i, i = 1, \ldots, N \\ \alpha(\hat{y}_i - \underline{s}_i) &\leq y_i - \underline{s}_i, i = 1, \ldots, N \\ \alpha(\overline{s}_i - \hat{y}_i) &\leq \overline{s}_i - y_i, i = 1, \ldots, N \end{aligned}$$

The same type of modeling yields the inf and sup limits of the α^*-cuts for the reconciled intervals $\tilde{Y}_i^*$ (maximizing and minimizing y_i, letting $\alpha = \alpha^*$ in the constraints above). By virtue of the linearity of the system of equations and of the membership functions, we can reconstruct the reconciled $\tilde{Y}_i^*$ up to possibility level α^* by linear interpolation between the cores and the optimal supports obtained by deleting the third and fourth constraints in the above program (although strictly speaking the reconciled fuzzy intervals might only be piecewise linear).

It is possible (and recommended) to iterate the above procedure and refine the optimal intervals $(\tilde{Y}_i)_{\alpha^*}$ by instanciating the quantities y_i such that $(\tilde{Y}_i)_{\alpha^*}$ reduces to a singleton $\{y_i^*\}$ as described in [6], leaving other variables in their optimal α^*-cuts. Namely, let $V_1 = \{i : (\tilde{Y}_i)_{\alpha^*} = y_i^*\}$. We can solve the problem of maximizing the value α such that

$$\begin{aligned} Ay^t &= B \\ \underline{s}_i \leq y_i &\leq \overline{s}_i, i \notin V_1 \\ y_i &= y_i^*, i \in V_1 \\ \alpha(\hat{y}_i - \underline{s}_i) &\leq y_i - \underline{s}_i, i \notin V_1 \\ \alpha(\overline{s}_i - \hat{y}_i) &\leq \overline{s}_i - y_i, i \notin V_1 \\ \alpha &\geq \alpha^* \end{aligned}$$

Then we get an optimal value $\alpha_1^* > \alpha^*$, and we can look for the optimal ranges $(\tilde{Y}_i)_{\alpha_1^*}, i \in V_1$ some of which again reduce to singletons. So, at this second step we have instanciated a set $V_2 \supset V_1$ of variables. We can iterate this procedure until all variables are instanciated, at various levels of optimal possibilities. Eventually, it delivers precise reconciled values along with possibility distributions around them. These values are Pareto-optimal in the sense of the vector-maximisation of the vectors $(\pi_1(y_1), \ldots \pi_N(y_N))$.

Among the three approaches, the latter based on fuzzy linear programming looks like the most convenient one.

	y_1	y_2	y_3	y_4
Least squares method				
Original data	24 ± 2	16 ± 3	15 ± 4	22 ± 5
Reconciliated values	$23,8$	$15,5$	$15,9$	$23,4$
Reconciliated fuzzy intervals				
Triangular fuzzy intervals	$(22, 24, 26)$	$(13, 16, 19)$	$(11, 15, 19)$	$(17, 22, 27)$
$\alpha^* : \frac{11}{14}$				
Reconciliated cores	$23 + 4/7$	$15 + 5/14$	$15 + 6/7$	$23 + 1/14$
Reconciliated supports	$[22, 26]$	$[13, 19]$	$[11, 19]$	$[17, 27]$

Table 1: Reconciliated flows for Example 1

6. Some examples

We present simple examples in order to compare the statistical and fuzzy approaches

6.1. One-process case

We consider the example illustrated in Figure 1, which is composed of four flows (y_1, y_2, y_3, y_4) and one process (P1). Flows y_1 and y_2 enter the process, while y_3 and y_4 exit the process. There are no stocks. In this example we have symmetric triangular fuzzy intervals $\tilde{Y}_1 = 24 \pm 2, \tilde{Y}_2 = 16 \pm 3, \tilde{Y}_3 = 15 \pm 4, \tilde{Y}_4 = 22 \pm 5$. In the least-squares method, we interpret the half-length of the interval as a standard deviation.

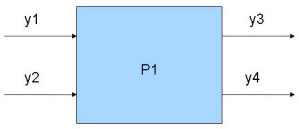

Figure 1: Example 1

With the fuzzy interval approach, the calculation of α^* using linear programming is obtained by solving the following linear problem: Maximize α such that:

$$
\begin{aligned}
y_1 + y_2 &= y_3 + y_4 \\
22 \leq y_1 &\leq 26 \\
\alpha \cdot (26 - 24) &\leq 26 - y_1 \\
\alpha \cdot (24 - 22) &\leq y_1 - 22 \\
13 \leq y_2 &\leq 19 \\
\alpha \cdot (19 - 16) &\leq 19 - y_2 \\
\alpha \cdot (16 - 13) &\leq y_2 - 13 \\
11 \leq y_3 &\leq 19 \\
\alpha \cdot (19 - 15) &\leq 19 - y_3 \\
\alpha \cdot (15 - 11) &\leq y_3 - 11 \\
17 \leq y_4 &\leq 27 \\
\alpha \cdot (27 - 22) &\leq 27 - y_4 \\
\alpha \cdot (22 - 17) &\leq y_4 - 17
\end{aligned}
$$

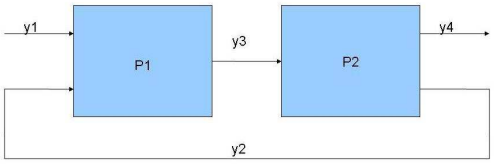

Figure 2: Example 2

The results obtained using the two methods (least-squares and fuzzy interval reconciliation) are provided in Table 1. We note that the alpha-cuts of the fuzzy intervals at level α^* after propagation are singletons and that the maximum deviation between the initial and reconciled values is smaller in the case of the fuzzy method than with the least-squares method.

6.2. Two-process example

We consider the example in Figure 2, composed of four flows (y_1, y_2, y_3, y_4) and two processes (P1 and P2). Flows y_1 and y_2 both enter process P1; y_3 exits P1 to enter P2, two flows exit P2: y_4 and y_2, while the latter is recycled into P1. In this example $\tilde{Y}_1 = 20 \pm 3, \tilde{Y}_2 = 10 \pm 2, \tilde{Y}_3 \in 20 \pm 4, \tilde{Y}_4 = 16 \pm 3$.

For the approach using fuzzy intervals, the calculation of α^* by linear programming is obtained by solving a system of equations similar to that of the previous case. We obtain $\alpha^* = 1/3$. We can also obtain this value by calculating the height of $\tilde{Y}_1 \cap \tilde{Y}_4$. Indicated in Table 2 are the cuts at level $1/3$ and the supports of the reconciled fuzzy intervals. We note that reconciled values obtained by least-squares are at the center of the supports of the reconciled intervals obtained using the fuzzy interval method.

However, it is possible to refine the remaining intervals. If we retain the information $y_1^* = y_4^* = 18$ and run the fuzzy interval propagation procedure again, we verify that the intersection $\tilde{Y}_2 \cap (\tilde{Y}_3 - 18)$ has a height of unity, obtained for $y_2 = 10$. We

can also fix $y_3 = 28$ considering $\tilde{Y}_3 \cap (\tilde{Y}_2 + 18)$. We can therefore verify that $\pi_1(18) = \pi_4(18) = 1/3, \pi_2(10) = \pi_3(28) = 1$ and therefore that the least-squares solution coincides in this particular example with the Pareto-optimal solution of the fuzzy data reconciliation problem. The former example shows that this is not always the case.

6.3. Comparing reconciled values: a generic example

Consider a single process with n inputs x_i and a single output $x_0 = \sum_{i=1}^{n} x_i$. Suppose all measured inputs are $\hat{x}_i = a > 0$ while $\hat{x}_0 = ka > 0$. One may argue that, assuming the x_i's have the same variance 1, x_0 has variance equal to n. It is easy to obtain least squares estimates, minimizing $\sum_{i=1}^{n}(x_i - a)^2 + \frac{(x_0-a)^2}{n}$ under the balancing constraint. It is easy to find that $x_0^{LS} = \frac{a(k+n)}{2}$ and $x_i^{LS} = \frac{a}{2} + \frac{ak}{2n}$. Note that $\lim_{n\to\infty} x_i^{LS} = a/2$ and in fact $\frac{a}{2} < x_i^{LS} \leq \frac{a(k+1)}{2}$. All reconciled flows linearly increase to infinity if k increases.

In the fuzzy interval approach we can assume triangular membership functions : $\tilde{X}_i$ has mode a and support $[a-\alpha, a+\beta]$, where the magnitudes of α, β depend on the available knowledge. Suppose that the relative error of the data is everywhere the same so that $\hat{X}_0$ has mode ka and support $[k(a-\alpha), k(a+\beta)]$. The reconciled value for x_0 is obtained as the value for which the intersection $\hat{X}_0 \cap n\tilde{X}_i$ has maximal positive possibility degree. There are two cases

$$x_0^* = \begin{cases} \frac{nka(\alpha+\beta)}{n\alpha+k\beta} \text{ if } k \leq n \text{ and } k(a+\beta) > n(a-\alpha) \\ \frac{nka(\alpha+\beta)}{k\alpha+n\beta} \text{ if } k \geq n \text{ and } k(a-\alpha) < n(a+\beta). \end{cases}$$

It can be checked that the least squares solution is encompassed by the fuzzy interval approach. If $k \leq n$, $x_0^* = x_0^{LS}$ if and only if α, β are chosen such that $n\alpha = k\beta > \frac{a(n-k)}{2}$ (the inequality makes the fuzzy reconciliation problem feasible). Likewise, if $n \geq k$, the condition is $k\alpha = n\beta > \frac{a(n-k)}{2}$.

Finally we check when x_0^* is closer to the estimated value ka than x_0^{LS}. For instance, if $k \leq n$, $x_0^* > ka$ and $x_0^{LS} > ka$, and $x_0^{LS} > x_0^* > ka$ provided that $k\beta < n\alpha$.

7. A unified framework for least squares and fuzzy interval reconciliation

If we select the product for operation $\star$ in the general formulation of Section 5.1, the reconciliation problem boils down to maximizing the expression $\pi_\odot(y) = \prod_{i=1}^{N} \pi_i(y_i)$ under constraints $Ay^t = B$. If in addition we choose to use Gaussian shapes $\pi_i(y) = e^{-\frac{(y_i-\hat{y}_i)^2}{\sigma_i^2}}$ for the fuzzy intervals, it becomes clear that this formulation brings us precisely back to the maximum likelihood expression $L(x)$. Therefore the fuzzy interval approach defined in Section 5.1 captures the least-squares method as a special case, minimizing the distance to estimated values in the sense of the l_2 norm.

With $\star = \min$ and triangular fuzzy intervals $\tilde{Y}_i$ centered around measured values $\hat{y}_i$, solving the max-min fuzzy constraint problem, reduces to minimizing the maximal weighted absolute deviation:

$$e_\infty(y) = \max_{i=1,\dots,N} \frac{|y_i - \hat{y}_i|}{\sigma_i},$$

using an l_∞ norm instead of the Euclidean l_2 norm. Here σ_i is interpreted as the spread of the fuzzy interval $\tilde{Y}_i$.

Similarly, choosing $a \star b = \max(0, a+b-1)$ under the same hypotheses comes down to minimizing a weighted sum of absolute errors, i.e., use the l_1 norm:

$$e_1(y) = \sum_{i=1,\dots,N} \frac{|y_i - \hat{y}_i|}{\sigma_i}.$$

More generally recent works on penalty-based aggregation [3] may help us find an even more general setting for devising reconciliation methods in terms of general penalty schemes when deviating from the measured data flows.

8. Conclusion

In the context of the mass flow reconciliation problem, we often deal with scarce data of various origins, pertaining to different quantities, that we can hardly assume to be generated by a standard random process. It seems more natural to concentrate efforts on the choice of a distance (l_1, l_2, l_∞, ...) for minimizing the error rather than to invoke the CLT to justify the least-squares method. A fuzzy-set approach to data reconciliation has been proposed. Its advantages are:

- Its general character: in a formal sense, it generalizes the least-squares method without betraying the principle of maximum likelihood. Indeed, it is well-known that a likelihood function is a special case of a possibility distribution [9] and we see clearly that the likelihood $L(x)$ is a special case of $\pi_\star(x)$, with $\star =$ product.
- Its clear conceptual framework, both for representing uncertainty pervading the data (statistical or subjective) and for leaving the choice of the distance that enters the error function to the user. The reconciled ranges around the reconciled values are also more easy to interpret than the reconciled variances, as they result from standard interval propagation.
- The opportunity of solving the problem in the max-min case, using standard methods and software.

However, our framework does not encompass the probabilistic method of variance reconciliation described in Section 2.1, since the latter views the normal distribution on flow measurements as a data

	y_1	y_2	y_3	y_4
Least squares method				
Original data	20 ± 3	10 ± 2	28 ± 4	16 ± 3
Reconciliated values	18	10	28	18
Reconciliated fuzzy intervals				
Triangular fuzzy intervals	$(17, 20, 23)$	$(8, 10, 12)$	$(24, 28, 32)$	$(13, 16, 19)$
$\alpha* : \frac{1}{3}$				
Reconciliated cores	$[18, 18]$	$[8 + \frac{2}{3}, 12 - \frac{2}{3}]$	$[26 + \frac{2}{3}, 29 + \frac{1}{3}]$	$[18, 18]$
Reconciliated supports	$[17, 19]$	$[8, 12]$	$[25, 31]$	$[17, 19]$
Reconciliated cores: 2d round	$[18, 18]$	$[10, 10]$	$[28, 28]$	$[18, 18]$

Table 2: Reconciliated flows For Example 2

generation process and not as an additional constraint. In contrast one could consider reconciling variances as a data fusion problem.

Mind that minimizing a sum of absolute valued deviations (and to a lesser extent quadratic), runs the risk of making certain values of x_i deviate significantly from the data $\hat{x}_i$, whereas the max-min approach is designed to keep all of them as close as possible to the initial data. The latter approach seems to be more reasonable if the data come as single estimate from an expert or other sources for each quantity. This approach is currently being studied for analyzing the material flow of rare earth elements in the anthroposphere of the EU-27.

Acknowledgements This work is supported by the French National Research Agency (ANR), as part of Project ANR-11-ECOT-002 ASTER "Systemic Analysis of Rare Earths - flows and stocks".

References

[1] F. Benhamou, L. Granvilliers, F. Goualard. Interval Constraints: Results and Perspectives. *New Trends in Constraints*, LNAI 1865, pages 1-16, Springer, 2000. *22nd European Symposium on Computer Aided Process Engineering* (I.D. Lockart Bogle, M. Fairweather, Eds.), Elsevier 2012, 122-126

[2] P.H. Brunner and H. Rechberger (2004). Practical Handbook of Material Flow Analysis. Lewis Publishers.

[3] T. Calvo, G. Beliakov Aggregation functions based on penalties, *Fuzzy Sets and Systems*, 161(10), 2010, 1420-1436.

[4] D. Dubois Possibility theory and statistical reasoning *Comput. Stat. & Data Anal.*, 51, 47-69, 2006

[5] D. Dubois, H. Fargier, H. Prade. Possibility theory in constraint satisfaction problems: Handling priority, preference and uncertainty. *Applied Intelligence*, 6, 287-309, 1996.

[6] D. Dubois, P. Fortemps. Computing improved optimal solutions to max-min flexible constraint satisfaction problems. *Eur. J. of Operation Research*, 118, 95-126, 1999.

[7] Dubois D. Foulloy L. Mauris G. Prade H. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing.* 2004. 10, 273-297

[8] D. Dubois, E. Kerre, R. Mesiar, H. Prade. Fuzzy interval analysis. In: *Fundamentals of Fuzzy Sets*, Dubois, D. Prade, H., Eds: Kluwer, Boston, Mass, The Handbooks of Fuzzy Sets Series, 483-581, 2000.

[9] D. Dubois, S. Moral, H. Prade. A semantics for possibility theory based on likelihoods. *J. of Math. Anal. Appl.*, 205, 359-380, 1997.

[10] D. Dubois, H. Prade, Philippe Smets: A definition of subjective possibility. *Int. J. Approx. Reasoning* 48(2): 352-364 (2008)

[11] L. Granvilliers, F. Benhamou. Algorithm 852: RealPaver: an interval solver using constraint satisfaction techniques. *ACM Trans. on Mathematical Software* 32(1), 138-156, 2006.

[12] L. Jaulin, M. Kieffer, O. Didrit, E. Walter. *Applied Interval Analysis*, Springer, London, 2001.

[13] O. Lhomme. Consistency Techniques for Numeric CSPs. *Proc. Int. Joint Conf on Artificial Intelligence*(IJCAI 1993), pages 232-238.

[14] G. Mauris: Expression of Measurement Uncertainty in a Very Limited Knowledge Context: A Possibility Theory-Based Approach. *IEEE T. Instrum. and Meas.* 56(3): 731-735 (2007)

[15] G. Mauris: Possibility distributions: A unified representation of usual direct-probability-based parameter estimation methods, *Int. J. of Approx. Reasoning*, 52(9), 1232-1242 (2011).

[16] S. Narasimhan, C. Jordache, Data reconciliation and gross error detection: an intelligent use of process data, Gulf Publishing Company, Houston, 2000.

[17] H. Steyaert, F. Van Parys, R. Baekeland and E. Kerre. Implementation of piecewise linear fuzzy quantities, *Int. J. Intelligent Systems*, 10, 1049-1059, 1995.

[18] S. M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900.* Belknap Press/Harvard University Press, 1990.