View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Open Archive Toulouse Archive Ouverte

Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 13201

To link to this article : DOI:10.1080/03081079.2014.920840
URL : http://dx.doi.org/10.1080/03081079.2014.920840

To cite this version : Dubois, Didier and Fargier, Hélène and Ababou, Meissa and Guyonnet, Dominique *A fuzzy constraint-based approach to data reconciliation in material flow analysis.* (2014) International Journal of General Systems, vol. 43 (n° 8). pp. 787-809. ISSN 0308-1079

Any correspondance concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

A Fuzzy Constraint-based Approach to Data Reconciliation in Material Flow Analysis

Didier Dubois[a*], Hélène Fargier[a], Meïssa Ababou[b] Dominique Guyonnet[b]

[a] IRIT, CNRS & Université de Toulouse, France
[b] BRGM-ENAG, Orléans, France

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 use a fuzzy-set-based approach, replacing Gaussian likelihood functions by fuzzy intervals, and a leximin criterion. We show that the setting of fuzzy sets provides a generalized approach to the choice of estimated values, that is more flexible and less dependent on oftentimes debatable probabilistic justifications. It potentially encompasses interval-based formulations and the least squares method, by choosing appropriate membership functions and aggregation operations. This paper also lays bare the fact that data reconciliation under the fuzzy set approach is viewed as an information fusion problem, as opposed to the statistical tradition which solves an estimation problem.

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

1. Introduction

Material flow analysis (MFA) consists in calculating the quantities of a certain product transiting within a defined system made up of a network of local entities referred to as processes, considering input and output flows and including the presence of material stocks. This method was developed in the sixties to study the metabolism of urban systems, like (Wolman 1965) for water networks. A material flow system is defined by a number of related processes. Material conservation is the basis of material flow analysis: constraints are typically related to conservation laws such as steady-state material, energy and component balance. In material flow analysis, the unknowns to be determined are the values of the flows and stocks at each process. These flows and stocks must be balanced, through a set of linear equations. 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. This is translated into mass-balance equations relative to a process with n flows in, k flows out and a stock level s is written:

*Corresponding author. Email: dubois@irit.fr

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

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 are collected regarding the material flow transiting the network and missing flow or stock variation values are calculated. But this task may face two opposite kinds of 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 is incompatible with the available data. This is because the available information is often not sufficiently reliable.

In this paper we address the second case. 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 science as early as the end of the 18th century when this question was addressed using the method of least squares. The idea was to find solutions to a system of linear equations as close as possible to measured values according to the Euclidean distance. The same method is still used today, but the justification is statistical and usually based on the Central Limit Theorem, the use of Gaussian functions and the principle of maximum likelihood.

Data reconciliation has been defined as "a technique to optimally adjust measured process data so that they are consistent with known constraints" by Kelly (2004). According to Crowe (1996), such adjustment consists in a "constrained minimization problem that is usually one of constrained least squares"; see also (Narasimhan and Jordache 2000). Ayres and Kneese (1969) extended the application of MFA to national economies while Baccini and Brunner (1991) used it to study the metabolism of the anthroposphere, i.e., that portion of the environment that is made or modified by humans for use in human activities and human habitats. Material flow analysis has become an important tool in the field of industrial ecology, e.g. (Frosch and Gallopoulos 1989). More recently, researchers have applied MFA to study the global flows and stocks of metals, e.g., Graedel et al. (2004), Bonnin et al. (2013). Some have generalized material flow analysis to several periods of time via a dynamic approach (Bai et al. 2006). Data reconciliation software are now available such as STAN (Brunner and Rechberger 2004) or BILCO (Durance et al. 2004).

In this paper, we examine the limitations of the classical approach and propose an alternative one that takes into account data uncertainty more explicitly, using intervals or fuzzy intervals [1]. Under this view, imprecise data are considered as (flexible) constraints, to the same extent as balance equations, contrary to the statistical methodology that considers the former as random variables. Under this approach, the problem can then be solved using crisp or fuzzy linear programming. The idea

[1] A short preliminary version of this paper (Dubois et al. 2013) was presented at the EUSFLAT conference in Milano, September 2013.

of using fuzzy intervals to address data reconciliation problems can be traced back to a paper by Kikuchi (2000) in connection with traffic modeling, and has not been much used in MFA since then, one exception being life cycle inventory analysis (Tan et al. 2007). The aim of this paper is:

- To better position the least square approach in a historical perspective questioning its usual justifications in the MFA context.
- To motivate the need to reconciliation methods different from the least square approach in this context.
- To present and improve the fuzzy interval approach to data reconciliation.
- To outline a general framework based on fuzzy intervals for the derivation of precise estimates, that encompasses the least squares estimate as a special case.
- To show that the problems addressed by the statistical and fuzzy approaches are fundamentally different despite the fact that the computed estimates in each approach are special cases of a more general setting.

The paper is organized as follows: in Section 2, we recall the formulation of the material flow problem and its least squares solution. We question the appropriateness of the statistical justification of this technique in the data reconciliation problem. Section 3 assumes imprecise data are modelled by means of intervals, and presents the interval reconciliation problem where model constraints and imprecise data are handled as crisp constraints. Section 4 extends the interval-based reconciliation method to fuzzy intervals. Section 5 provides simple examples where the specificity of the fuzzy data reconciliation problem can be highlighted. Section 6 describes the application of the proposed methodology to an example inspired by a real case of copper flow analysis. Finally, in Section 7, we compare the statistical and the fuzzy approaches both on the issue of estimating plausible flow values and on that of computing the resulting uncertainty on the flow values. The appendix recalls definitions pertaining to the modelling of uncertain data by fuzzy sets.

2. Data reconciliation via least squares: a discussion

As recalled above, data reconciliation in the MFA context consists in modifying measured or estimated quantities in order to balance the mass flows in a given network. We denote by y the vector of flows and stocks and subdivide it into two subvectors 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 this paper we focus on the case when 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 defects. The problem to be solved is to modify x, while remaining as close as possible to $\hat{x}$, so that the mass balance equations $A(xu)^t = B$ are satisfied.

2.1. *The least squares approach*

The traditional approach to data reconciliation (Narasimhan and Jordache 2000) 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 is then formulated as a problem of quadratic optimization under linear constraints. In the simplest case, assuming no

variables u (that is, some piece of information is available for all flows and stocks):

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

$$\text{such that } Ax^t = b$$

The solution is known to be of the form (Narasimhan and Jordache 2000):

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

where W is a diagonal matrix containing terms $1/w_i$. Weights are often of the form $w_i = (\sigma_i)^{-2}$. It is the method of weighted least squares used to reconcile data in several material flow analysis tools such as STAN (Brunner and Rechberger 2004) or BILCO (Durance et al. 2004).

Such packages sometimes also reconcile variances as explained in (Narasimhan and Jordache 2000). 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 on the estimated values via a linear transformation, say $x^* = B\hat{x}$ - hence, the x^* also have a normal distribution and the covariance matrix of x^* is 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. The shape of the latter is 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 mean 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. The universal character of this approach, albeit reasonable in certain situations, is nevertheless dubious:

- It is not consistent with the history of statistics (Stigler 1990). The least squares method, developed by Legendre (1805) and Gauss (end of 18th century), was discovered prior to the CLT, and the normal law was found independently. 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 intuitively appealing average in the estimation of quantities based on several independent measures. The normal law was discovered by Gauss as the only error function compatible with the average estimator. However, the CLT is a mathematical result obtained independently by Laplace, who later on made the connection between his mathematical result and the least squares method, based on Gauss finding.
- The CLT presupposes a statistical process with a finite mean value 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 variable $\frac{\sum_{i=1}^{n} v_i - nE}{\sqrt{n}}$ 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. This assumption is neither always true nor easily verifiable; but it is obviously true if the measured quantity is bounded from below and from above due to physical constraints (but then its distribution is not Gaussian, strictly speaking - even if it is often approximated by a Gaussian function).

- 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 the repetitive use of a single measurement process. Indeed, in the reconciliation problem, we rather face the case of having single assessments of many quantities, rather than many measurements of a single quantity. Given the fact that these assessments are not assumed to come from physical sensors (they may come from documents or experts), and that actual values are not independent since related via balance equations, applying some kind of ergodicity that would justify the classical estimation method makes little sense here. In fact, other probabilistic techniques can be envisaged when the Gaussian assumption does not apply. For instance, Gottschalk et al. (2010) use a Bayesian approach to represent uncertain data, in the form of various distributions (e.g. the uniform one in case of total uncertainty within limits); they apply Monte-Carlo methods to solve the reconciliation problem. Alhaj-Dibo et al. (2008) propose a mixture of two Gaussians to account for noise and gross errors in a separate way, which enable the latter to be coped with in the reconciliation process.

An alternative, more straightforward approach consists in representing error-tainted data by means of intervals and checking the compatibility between these intervals and the material flow model. This is quite different from the standard statistical approach, where the least squares solution is taken for granted and variance reconciliation is the result of a kind of probabilistic sensitivity analysis around it.

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. Or the flow information deals with various products grouped together. Each flow value provided by a source can thus be more safely represented, as a gross approximation, by an interval $\hat{X}_i$ that 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 weighted least squares approach to data reconciliation, weights reflect the assumed variance of a Gaussian phenomenon; if such information on variances σ_i^2 is available, we can set $\hat{X}_i = [\hat{x}_i - 3\sigma_i, \hat{x}_i + 3\sigma_i]$ as a realistic interval containing x_i. This choice captures 99.74% of the normal distribution. Actually,

the distribution of x_i is often assumed to be Gaussian for practical reasons, even when the actual parameter is known to be positive or bounded due to physical constraints. Thus, knowledge about each of the N variables y_i of the vector $y = xu$ can be approximately modelled by an interval $\hat{Y}_i$. In this setting, there is clearly a uniform treatment of balance equations and data pertaining to measured or non measured flow values.

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 feasibility ranges of other variables in the form of intervals. Formally, the reconciliation problem can be expressed as follows:

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

$$\begin{aligned} Ay^t &= B \\ y_i &\in \hat{Y}_i, i = 1, \ldots, N \end{aligned}$$

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 exist 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 (Benhamou et al. 2000). For each variable y_i, equation j of the system $Ay^t = B$ can be expressed as

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

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:

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

where $\frac{\sum_{k \neq i} b_j - a_{jk} \hat{Y}_k}{a_{ji}}$ is calculated according to the laws of interval arithmetic (Jaulin et al. 2001); if the new interval of possible values of y_i has become 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; when there are no disjunctive constraints, it converges within a finite number of steps to a unique set of intervals (Lhomme 1993) (for additional details, see (Benhamou et al. 2000; Granvilliers and Benhamou 2006)). This approach has actually been applied to reconciliation problems in the area of measurement in the early 2000's (Ragot and Maquin 2004; Ragot et al. 2005). Again, contrary to the statistical approach, the model constraints and the imprecise data are handled on a par, the latter being viewed as unary constraints.

4. Fuzzy interval reconciliation

The interval approach of Section 3 does not yield the same type of answer as the least squares method because it provides intervals rather than precise values. Such intervals may look similar to reconciled variances provided by current software for MFA and data reconciliation like STAN (Brunner and Rechberger 2004) (but as we shall see later, this comparison is misleading).

A natural way to obtain both reconciled values and intervals is to enrich the representation of the information pertaining to 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. In practice, it means that for each informed quantity, not only an interval should be provided, but also a plausible value (or a shorter interval thereof). Such information can be modelled by means of a triangular or trapezoidal fuzzy interval. See the appendix, for an introductory discussion of fuzzy intervals as representing incomplete information, and the basic definitions used in this section.

The problem of searching for a possible solution then becomes an optimization problem - we seek an optimal value within all the (fuzzy) intervals of possible ones. If no solution provides entire satisfaction for all intervals, some of them will be relaxed if necessary (Dubois et al. 1996).

4.1. *The max-min setting*

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 now represented in the form of a fuzzy interval $\tilde{Y}_i$, understood as a possibility distribution π_i that expresses a flexible unary constraint. This fuzzy interval may coincide with the domain of the quantity, in the case of total ignorance.

An assignment y for all y_i is feasible, 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. The simplest approach is to use the minimum as the standard fuzzy conjunction, following the pioneering paper of Bellman and Zadeh (1970). It has the merit of being insensitive to possible dependencies between the involved fuzzily assessed quantities.

The corresponding optimization problem for determining a most plausible estimate was already formulated some years ago by Kikuchi (2000): Find y^* that maximizes

$$\pi_{\min}(y) = \min_{i=1}^{N} \pi_i(y_i) \quad \text{where } Ay^t = B \tag{2}$$

with $Ay^t = B$. Let $\alpha^* = \min_{i=1}^{N} \pi_i(y_i^*)$ be the maximal plausibility value and y^* an optimal solution. The value α^* can be interpreted as the degree of consistency of the flexible constraint problem. This implies that we cannot hope for 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 y^* at level α^*. This approach was in fact already used in an algorithm for tuning the cutting parameters of machine-tools, based on expert preference, under crisp constraints pertaining to

the production rate (Dubois 1987).

Note that, contrary to what examples in the paper by Kikuchi (2000) may suggest, there may exist several solutions y^* that achieve a global level of satisfaction α^*. Once α^* is known, we can indeed 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:

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

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 $\tilde{Y}_i^*$ obtained by projection of $\pi_{\min}$ on the domain of each y_i:

$$\forall v \in S(\tilde{Y}_i), \mu_{\tilde{Y}_i^*}(v) = \max_{y \text{ s.t. } y_i = v \text{ and } Ay^t = B} \min_{j=1}^{N} \pi_j(y_j)$$

The supports of the fuzzy intervals $\tilde{Y}_i^*$ 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 mathematical program contains on the one hand the mass flow model $Ay^t = B$ which, as seen previously, is linear; moreover we force the y_i's to belong to the supports $[\underline{s}_i, \overline{s}_i], i = 1, \ldots N$ of the fuzzy intervals $\tilde{Y}_i$'s. Finding the fuzzy reconciled flows $\tilde{Y}_i^*$ provides both plausible ranges and uncertainty around them. These fuzzy domains are subnormalized if $\alpha^* < 1$: they all have heights $h_i = \sup_{y_i} \pi_i^*(y_i) = \alpha^*$ and at least one of them contains a single value y_i^*, while the α^*-cuts of others are intervals of optimal values.

The fuzzy reconciliation method fails to deliver a solution if $\alpha^* = 0$. In that case, we may consider that the data are inconsistent with the material flow model, and we must either re-assess the validity of data items (deleting some of them considered unreliable, or relaxing the tolerance distributions) or revise the material flow model. The possibility of this occurrence contrasts with the least squares method which always provides a solution. But this solution may yield values far away from the original estimates in case the data is strongly conflicting with the balance equations. It will correspond to a very low likelihood value in the Gaussian interpretation. While the possible failure of the fuzzy constraint-based reconciliation method may be regarded as a drawback, one may on the contrary see it as a virtue as it is capable of warning the user when an inconsistency occurs without having to prescribe an arbitrary likelihood global threshold on the [0, 1] scale.

4.2. *Resolution methods*

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

4.2.1. 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 (Dubois et al. 2000) since equations are linear. Expressions such as $\tilde{Y}_i \cap (\cap_{j=1...,m}\tilde{A}_j)$ have possibility distribution $\pi'_i = \min(\pi_i, \min_{j=1...,m}\pi_{\tilde{A}_j})$.

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 (Steyaert et al. 1995) including subnormalized ones. Eventually, the optimal (maximally precise) fuzzy intervals $\tilde{Y}_i^*$ (fuzzy domains of the reconciled flows defined in the previous subsection) are obtained with heights not greater than α^*. Indeed, fuzzy arithmetic methods applied to fuzzy intervals of various heights only preserve the least height (Dubois et al. 2000).

4.2.2. 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 $(\tilde{Y}_i^*)_{\alpha^*}$.

4.2.3. 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 $\alpha \leq \pi_i(y_i), i = 1, \ldots, N$ (a trick as old as (Zimmermann 1978)). When the original fuzzy intervals are triangular with core $\hat{y}_i$ and support $[\underline{s}_i, \overline{s}_i]$, each constraint is written in the form of two linear inequalities, one for each side of the fuzzy intervals, as already proposed in (Kikuchi 2000; Tan et al. 2007). 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 the reconciled fuzzy intervals might only be piecewise linear).

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

4.3. Iterating the optimisation process

It is possible (and recommended) to iterate the method and update again some of the fuzzy ranges $\tilde{Y_i}^*$. Namely, one may refine the optimal intervals $(\tilde{Y_i}^*)_{\alpha^*}$ not reduced to a single value yet, and obtain more precise plausible estimates. The idea, described in (Dubois and Fortemps 1999), is that, while some intervals $(\tilde{Y_i})_{\alpha^*}$ reduce to singletons $\{y_i^*\}$ that can be considered as fully determined flows, other intervals $(\tilde{Y_i})_{\alpha^*}$ obtained after the previous optimization step can be further reduced to precise values as well.

Namely, let $V_1 = \{i : (\tilde{Y_i}^*)_{\alpha^*} = y_i^*\}$ be the indices of parameters whose values are fixed by considering α^*-cuts. This set is not empty for otherwise, since the fuzzy sets $\tilde{Y_i}$ are of triangular shape, one could still raise the level α^* without creating an inconsistency, which by assumption is not the case as α^* is maximal. So, we define a second optimization problem, where we assign their optimal values y_i^* to flows $y_i \in V_1$, and leave other values free in their original fuzzy ranges. We thus solve the following partially instantiated program: maximize the value β such that

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

Then we get a new optimal value $\beta^* > \alpha^*$ that pertains to flows not in V_1. Indeed, there are several possible values $y_i \in (\tilde{Y_i}^*)_{\alpha^*}$, when $i \notin V_1$, and the new optimisation problem tends to select the ones that have higher membership grades inside $(\tilde{Y_i}^*)_{\alpha^*} \cap \tilde{Y_i}$. We thus get narrower optimal ranges $Y_i^2 \subseteq (\tilde{Y_i}^*)_{\alpha^*} \cap (\tilde{Y_i})_{\beta^*}, i \notin V_1$ some of which (forming a subset V_2 of flows) again reduce to singletons. So, at this second step we have instantiated a set $V_2 \cup V_1$ of variables. We can iterate this procedure until all variables y_i are instantiated, at various levels of optimal possibility $\alpha_i^*, i = 1, \ldots k,$, with $\alpha_k^* > \alpha_{k-1}^* > \cdots > \alpha_2^* = \beta^* > \alpha_1^* = \alpha^*$. Eventually, it delivers for each variable y_i a 4-tuple $(\underline{s}_i, \overline{s}_i, y_i^*, \alpha_j^*)$, assuming a precise value y_i^* was found at step j. It can be approximated by a triangular fuzzy interval $\tilde{Y_i}^{**}$ such that $[\underline{s}_i, \overline{s}_i]$ is its support as well as the support of $\tilde{Y_i}^*$ (found in the first pass), y_i^* is its core, and α_i^* its height, that is, precise reconciled values along with their maximal range of possible values around them[2].

The plausible estimates obtained at the end of this recursive procedure are Pareto-optimal in the sense of the vector-maximisation of the vectors $(\pi_1(y_1), \ldots \pi_N(y_N))$, and leximin-optimal for the maximisation. Namely, there does not exist another tuple of values y such that $\pi_i(y_i) \geq \pi_i(y_i^*), \forall i = 1, \ldots, N,$

[2]In fact we also possess all cuts $(\tilde{Y_i}^*)_{\alpha_\ell}, \ell < j$.

and $(\pi_1(y_1), \ldots \pi_N(y_N)) \neq (\pi_1(y_1^*), \ldots \pi_N(y_N^*))$, on the one hand, and moreover, $(\pi_1(y_1^*), \ldots \pi_N(y_N^*))$ is maximal for the leximin order defined by $(a_1, \ldots, a_N) \geq_{lmin} (b_1, \ldots, b_N)$ if and only if $(a_{\sigma(1)}, \ldots, a_{\sigma(N)})$ is lexicographically greater than $(b_{\tau(1)}, \ldots, b_{\tau(N)})$ [3] where $(a_{\sigma(1)}, \ldots, a_{\sigma(N)})$, and $(b_{\tau(1)}, \ldots, b_{\tau(N)})$ are the two vectors reshuffled in the increasing order: $a_{\sigma(1)} \leq \cdots \leq a_{\sigma(N)}$, and $b_{\tau(1)} \leq \cdots \leq b_{\tau(N)}$ (see (Dubois and Fortemps 1999) for details on the leximin order in the setting of max-min optimisation).

5. Some examples

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

5.1. *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 $\tilde{Y}_1 = 24 \pm 2, \tilde{Y}_2 = 16 \pm 3, \tilde{Y}_3 = 15 \pm 4, \tilde{Y}_4 = 22 \pm 5$.

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

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

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 (no need for a second pass) and that the maximum distance between the initial and reconciled values is smaller in the

[3] That is, there exists an index k, such that $a_{\sigma(i)} = b_{\tau(i)}, \forall i < k$, and $a_{\sigma(k)} > b_{\tau(k)}$.

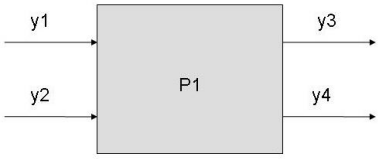

Figure 1. Example 1: one process with 4 flows

	y_1	y_2	y_3	y_4
Least squares method				
Original data	24 ± 0.67	16 ± 1	15 ± 1.33	22 ± 1.67
Reconciliated values and S.D's	23.8 ± 0.54	15.5 ± 0.91	$15,9 \pm 1.12$	23.4 ± 1.22
Fuzzy set method				
Original fuzzy intervals	$(22, 24, 26)$	$(13, 16, 19)$	$(11, 15, 19)$	$(17, 22, 27)$
$\alpha^* : \frac{11}{14}$				
Reconciliated values	$23 + 4/7$	$15 + 5/14$	$15 + 6/7$	$23 + 1/14$
Reconciliated supports	$[22, 26]$	$[13, 19]$	$[11, 19]$	$[17, 27]$

Table 1. Reconciliated flows for Example 1

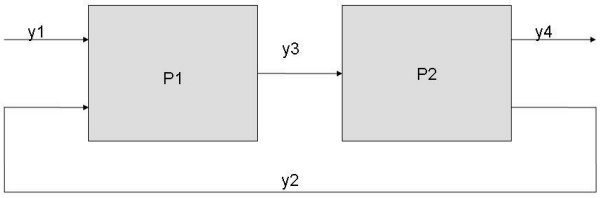

Figure 2. Example 2: two processes with 4 flows

case of the fuzzy method than with the least squares method, which is expected since the aim of the max-min approach is precisely to minimize the largest deviation from initial values. Moreover, in this example the reconciled supports are left unchanged by the reconciliation procedure.

5.2. *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 $\tilde{Y}_1 = 20 \pm 3, \tilde{Y}_2 = 10 \pm 2, \tilde{Y}_3 \in 28 \pm 4, \tilde{Y}_4 = 16 \pm 3$.

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

	y_1	y_2	y_3	y_4
Least squares method				
Original data	20 ± 1	10 ± 0.67	28 ± 1.33	16 ± 1.67
Reconciliated values	18 ± 0.64	10 ± 0.61	28 ± 0.79	18 ± 0.64
Fuzzy set method				
Original fuzzy intervals	$(17, 20, 23)$	$(8, 10, 12)$	$(24, 28, 32)$	$(13, 16, 19)$
$\alpha* : \frac{1}{3}$				
Reconciliated supports	$[17, 19]$	$[8, 12]$	$[25, 31]$	$[17, 19]$
Reconciliated cores (1st round)	$[18, 18]$	$[8.66, 11.44]$	$[26.66, 29.33]$	$[18, 18]$
Reconciliated cores: 2d round	$[18, 18]$	$[10, 10]$	$[28, 28]$	$[18, 18]$

Table 2. Reconciliated flows For Example 2

solution coincides in this particular example with the Pareto-optimal solution of the fuzzy data reconciliation problem, due to the symmetry of the network and of the fuzzy intervals. The next example shows that this is rather seldom the case.

5.3. *Comparing reconciled values: a simple 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, x_0 has a variance n times larger. This is what is assumed in the following.

It is easy to obtain least squares estimates, minimizing $\sum_{i=1}^{n}(x_i - a)^2 + \frac{(x_0 - ka)^2}{n}$ under the balancing constraint. It is easy to find that

$$x_0^{LS} = \frac{a(k+n)}{2} \text{ and } x_i^{LS} = \frac{a}{2} + \frac{ak}{2n}.$$

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 general 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 latter inequality makes the fuzzy reconciliation problem feasible).
- Likewise, if $n \geq k$, the condition is $k\alpha = n\beta > \frac{a(n-k)}{2}$.

These findings are at odds with the least squares method. Indeed note that in order to get the same estimates in the two approaches, we cannot assume the triangular fuzzy intervals are symmetric, while the translation of normal laws into symmetric fuzzy intervals ($\alpha = \beta$ with spreads $\alpha = 3\sigma$) would enforce symmetry.

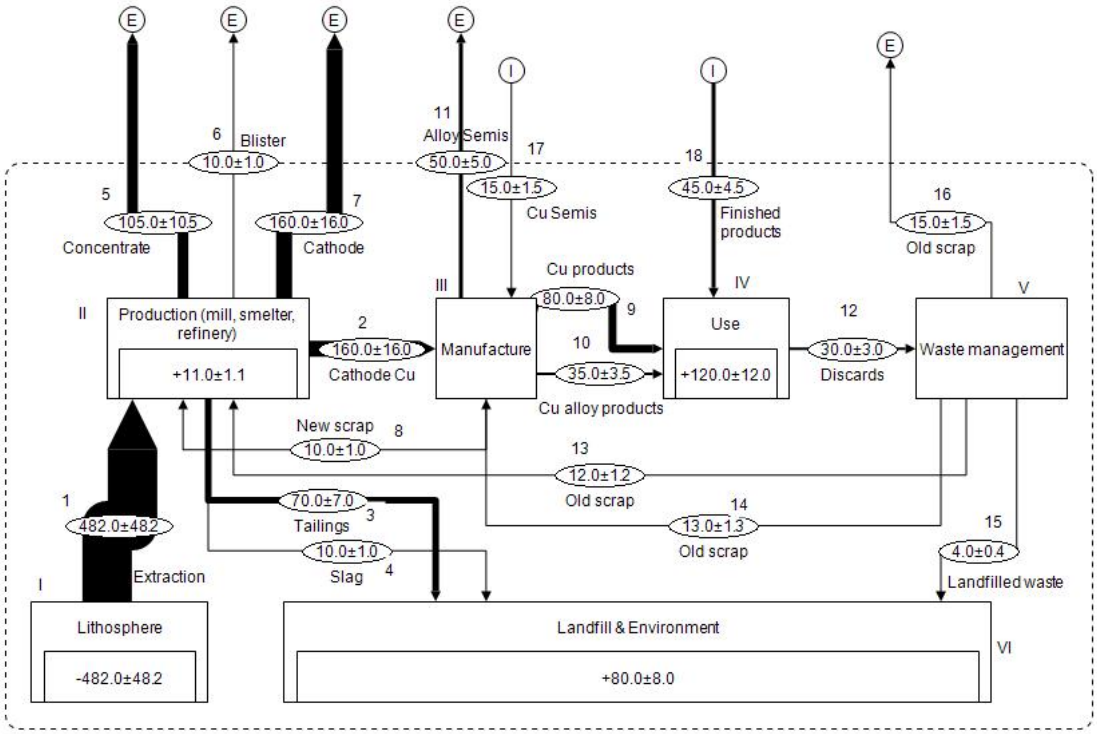

Figure 3. Initial flows and stocks of copper in the Australian economy (van Beers et al. 2007). Numbers in thousand metric tons.

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

6. A case study

The least squares and fuzzy reconciliation approaches were compared using data adapted from van Beers et al. (2007) relative to flows and stocks of copper in Australia in the mid-90s (see Figure 3 adapted from Figure 5A in that paper). The processes considered by these authors in their analysis of the major flows of copper over the entire copper life-cycle in Australia are: extraction from the lithosphere (the subsurface), treatment (production) of the copper ore, manufacturing of semi- and finished products (e.g. copper wire, tubing, etc.), use of these products in the economy, waste management and finally landfill and the environment, overall 24 quantities to be reconciliated.

For the purpose of the application, flows and stocks in Figure 5A of van Beers et al. (2007) were shifted arbitrarily from their initial values so that the MFA is no longer balanced. This initial MFA, which served for the reconciliation, is depicted in the Sankey diagram of Figure 3, which was constructed using the STAN software (Brunner and Rechberger 2004). Sankey diagrams (Baccini and Brunner 1991) are a specific type of flow diagram, in which the widths of the arrows are proportional to the flow quantities. Such diagrams are typically used to visualize energy or material transfers between processes. They are particularly useful to help identify potentials for recycling.

For example, Figure 3 suggests that the yearly change in stock in the "Landfill

Flow	Rec. Mean	Rec. σ	Rec. Support	Optimal Cut First Pass	Final Leximin
1. Extraction	491.1	18.1	[337.4, 626.6]	[390.8, 573.2]	478.6
2. Cu cath to Mnf.	150.6	7.7	[112, 208]	[129.7, 185.4]	154.1
3. Tailings	68.1	5.2	[49, 91]	[56.8, 83.2]	68.8
4. Slag	10.0	1.0	[7, 13]	[8.1, 11.9]	9.8
5. Export Cu con.	104.1	10.1	[73.5, 137]	[85.1, 124.9]	105.7
6. Export of blister	10.0	1.0	[7, 13]	[8.1, 11.9]	10.1
7. Exp.Cu cath.	158.0	14.7	[112, 207]	[129.7, 190.3]	161.1
8. New scrap to Prod.	10.0	1.0	[7, 13]	[8.1, 11.9]	10.4
9. Cu products	80.9	6.4	[56, 104]	[64.8, 95.1]	80.8
10. Cu alloy products	35.2	3.4	[24.5, 45.5]	[28.4, 41.6]	35.4
11. Export alloy	50.7	4.8	[35, 65]	[40.5, 59.5]	52.2
12. Discards	38.7	1.8	[31, 39]	35.7	35.7
13. Old scrap to Prod.	10.6	1.1	[8.4, 15.6]	9.7	9.7
14. Old scrap to Mnf.	11.3	1.2	[9, 17]	10.5	10.5
15. Landfilled waste	3.8	0.4	[2.8, 5.2]	3.2	3.2
16. Export of old scrap	12.9	1.4	[10.5, 18.7]	12.2	12.2
17. Import Cu semis	14.9	1.5	[10.5, 19.5]	[12.2, 17.8]	14.4
18. Import finished prod.	44.7	4.3	[31.5, 58.5]	[36.5, 53.5]	43.5
I. Ch. In Litho. Stock	491.1	18.1	[337.4, 626.6]	[390.8, 573.2]	478.6
II. Ch. In Prod. Stock	11.0	1.1	[7.7, 14.3]	[8.9, 13.1]	10.9
IV. Change in Use stock	122.1	7.2	[84, 156]	[97.3, 142.7]	124.1
VI. Change in Landfill Stock	81.9	5.3	[58.8, 104]	[68.1, 95.1]	81.9
Total Imports (I)	59.7	4.5	[42, 78]	[48.6, 71.4]	57.9
Total Exports (E)	344	19	[238, 441]	[275.7, 398.6]	341.5

Table 3. Results of reconciliation using the least squares and the possibilistic methods. Flow and process numbers refer to Fig. 3

and Environment" process is significant when compared to the copper extracted from the subsurface. Such data and diagrams can help motivate efforts with respect to so-called "landfill mining" operations, e.g., (Jain et al. 2013)). Figure 3 suggests that in the mid-90s, approximately 500 kilotons of copper were extracted each year as copper ore from Australian mines. This ore was processed in mills, smelters and refineries to produce intermediate copper products (concentrate, blister, copper cathode). Such processing generated discards in the form of tailings and slags that ended up in the "Landfill & Environment" process. The intermediate copper products were sent to manufacturing processes, located within Australia, to generate finished products that entered the economy to be used. Some finished products were exported outside Australia, the limits of which are symbolized by the dashed line in Figure 3. End-of-life products (discards) entered the waste management system, which generated old scrap that was either exported or else recycled within the domestic production and manufacturing processes. Waste containing copper also ended up in landfills.

Figure 3 also provides the means and standard deviations of the flows and stocks used for the least squares reconciliation. Since no information pertaining to standard deviations is provided by van Beers et al. (2007), standard deviations (before reconciliation) were assumed to be equal to 10% of the mean. For the possibilistic case, the original possibility distributions were assumed to be triangular. The means provided by Figure 3 were assumed to represent the central preferred values of the distribution; the supports were taken as plus or minus three times the standard deviations (see Section 3).

Comparative results of the least squares and possibilistic reconciliation methods are presented on Table 3. The results of the possibilistic reconciliation are those of the max-min method with leximin iteration. The first pass delivers the supports of

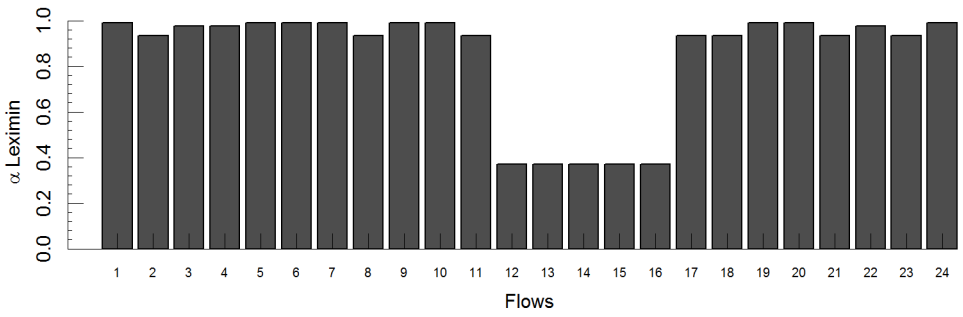

Figure 4. Convergence of the leximin iterative method

the fuzzy intervals, and the global consistency level (α^*) for the fuzzy constraint reconciliation, which is close to 0.4, as shown on Figure 4.3. It reflects a moderate conflict between original items of information. On this figure, we can see that the flows 12 to 16 are set to this consistency levels. They are critical and have thus precise first pass reconciliated values that can be seen on the 2d column from the right on Table 3. These flows are inputs and outputs of the waste management process on Fig. 3, which indicates the location of the most conflicting information. The other variables are still assigned intervals corresponding to the optimal consistency value. The three runs needed to reach precise estimates are patent from Figure 4, each run corresponding to a higher possibility value. The right-hand column provides precise estimates resulting from several leximin iterations. Table 3 also illustrates the point that with the possibilistic method, ranges around the preferred values are not necessarily symmetrical, unlike the least squares method. Work is currently under way to identify the most appropriate graphical representation of such results in a Sankey diagram.

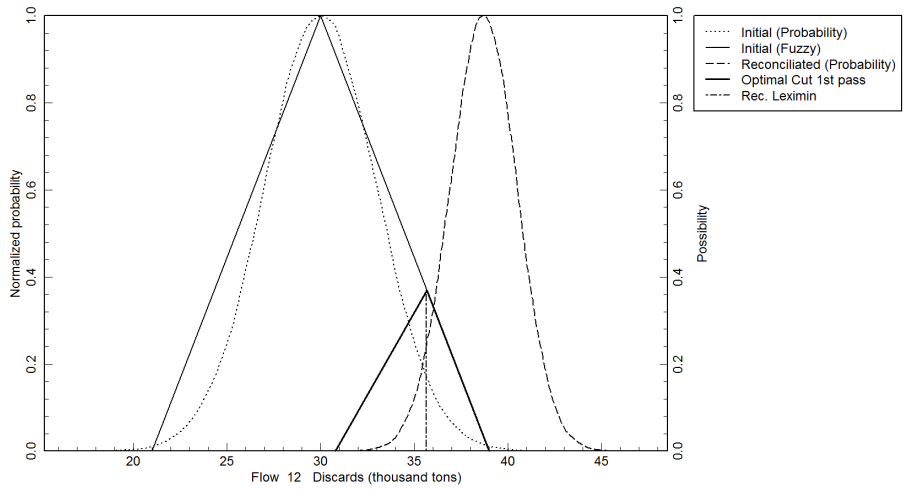

Figure 5. Comparison of Discards flow from Use, obtained using the least squares and possibilistic methods

While reconciled flows are generally close to initial flows, there are some significant differences; as for example in the case of "Discards from Use to Waste management", which vary by nearly 30% compared to the initial value. The values for this flow before and following reconciliation are depicted in Figure 5. For the purpose of the comparison, the probability density functions were normalized to unity. As can be

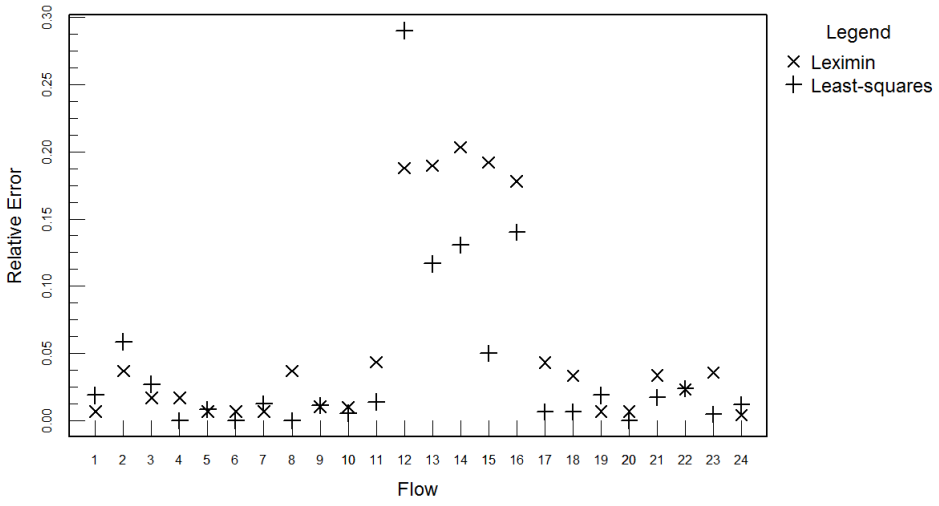

Figure 6. Relative distances of the LS and Leximin solutions to original values

seen in this figure, in the case of least squares reconciliation, the distribution is shifted laterally and moves outside the support of the initial interval, whereas in the case of the fuzzy reconciliation method, the reconciled flow and its possibility distribution (always) remain within the boundaries of the initial distribution. This suggests the fact that the distribution of least squares estimated values may overlap a domain of values considered impossible or very unlikely, in any case inconsistent with the original imprecise data.

Besides, Figure 6 pictures the relative differences between reconciliated values and original values. It lays bare the fact that the method keeps many least square estimates close to original values, at the risk of letting some be far away: especially the critical flow 12 "Discards from Use to Waste management" is the worst result in relative value, while the fuzzy set approach does better; the latter yields slightly worse results for critical flows 13, 14, 15,16. This is because the least square method clearly suggests some initial values of parameter are outliers, while the fuzzy approach tries to build a trade-off between all initial estimates, considered as valuable so long as $\alpha^* > 0$. This view may be considered more natural if initial estimates come from experts and not from measurement devices subject to gross errors.

7. Discussion: least squares or fuzzy set approach

Beyond comparing the results obtained by the two data reconciliation methods on practical examples, it is interesting to see to what extent these methods differ in their principles and the problems they address. The least squares estimation has two possible readings: a distance-based one and a statistical one. The distance-based one turns out to be close to the fuzzy set-based approach in the derivation of the most plausible reconciled values, and as shown below both can be put in the same general formal setting. However, the variance-reconciliation step requires a statistical understanding of the least squares procedure, and it does not consider measurement data as (unary) constraints in the same sense as balance equations. As explained in this section, beyond the possibility of obtaining different results, the conceptual frameworks underlying the statistical approach to the least squares method and the fuzzy constraint approach are radically different.

7.1. *A unified framework for reconciliation*

The max-min formulation (2) of Section 4.1 of the fuzzy constraint approach can be extended, replacing the minimum by a more general fuzzy conjunction. Namely, we may instead consider a likelihood function of the form $L(y) = \star_{i=1}^{N} \pi(y_i)$, where the operation $\star$ is associative, commutative and increasing on $[0, 1]$ - a t-norm (Klement et al. 2000). In fact, it is well-known that a likelihood function is a special case of a possibility distribution (Dubois et al. 1997). 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:

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

Restricting to continuous Archimedean t-norms, maximising $\pi_\star(y)$ comes down to minimizing a sum of the form $\sum_{i=1}^{N} g(\pi_i(y_i))$ where g is a t-norm generator (a continuous decreasing mapping from $[0, 1]$ to $[0, +\infty)$ with $g(1) = 0$ (Klement et al. 2000)). It comes close to generalised forms of least squares discussed for instance in (Alhaj-Dibo et al. 2008). Under suitable choice of functions π_i and g, the composition $g(\pi(y))$ is of the form $(\frac{y_i - \hat{y}_i}{\sigma_i})^p$ for some value $p > 1$: the problem comes down to minimizing an l_p norm.

If we select the product for operation $\star$ in the general formulation, 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 standard maximum likelihood expression of the least squares method. Therefore the general fuzzy interval framework captures the least squares estimation method as a special case, minimizing the Euclidean distance to estimated values.

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:

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

using a Chebyshev l_∞ norm (i.e., $\lim_{p\to\infty} e_p(y) = (\sum_{i=1,\ldots,N} \frac{|y_i - \hat{y}_i|^p}{\sigma_i^p})^{1/p}$) instead of the Euclidean l_2 norm. Here σ_i is interpreted as half the support length of the fuzzy interval $\tilde{Y}_i$. The precise estimates obtained by repeating the max-min optimisation step (Subsection 4.3) yields the strict Chebyshev norm solution already known in numerical analysis (Descloux 1963; Rice 1962). In his paper, Rice (1962) even describes a recursive procedure similar to the one we outline.

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:

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

More generally recent works on penalty-based aggregation (Calvo and Beliakov 2010) 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. It is well-known that applied to the estimation of a single quantity

for which several measurements are available, the three methods respectively yield the average (l_2 norm) the median (l_1 norm) and the mid-point between extreme measurement values (l_∞ norm). So the approach using l_1 norm is insensitive to outliers while the one based on the l_∞ norm is very sensitive to extreme values. This state of facts could be viewed as a major drawback for the fuzzy maxmin approach in some context as repeated measurements with gross errors. However, our context is not the one of repeated measurements of a single quantity, but the case of single imprecise expert information items about several quantities related by linear constraints. In this context it is perfectly reasonable to provide estimates of each quantity that respect as much as possible the opinion of each expert (which is what the fuzzy approach does). However the fuzzy approach does detect the presence of outlier experts, when the collected information is inconsistent with the model equations. But then, outlier elimination is the result of a reasoned process, not of an automatic data processing method.

7.2. *The reconciliation problem : estimation vs. information fusion*

Despite the above formal unification of methods computing reconciled values, the statistical approach seems to solve a problem that is radically different from the problem solved by the fuzzy approach, when it comes to modelling the resulting uncertainty on estimated values:

- The statistical approach envisages data reconciliation as an estimation problem, i.e., that of finding the ideal unbiased estimate (namely, the least squares solution) and computing its distribution as induced by the distributions of the data.
- In the fuzzy approach, the aim is to find the widest fuzzy ranges for reconciled values by projecting the result of merging model constraints and data constraints. The reconciled values are then maximal likelihood estimates *resulting* from these reconciled local fuzzy ranges.

So the statistical approach first computes estimates, while the fuzzy approach primarily computes fuzzy ranges resulting from an information fusion process. The choice of a paradigm (estimation or fusion) actually does not depend on the (probabilistic or not) formal setting. Indeed, one could use a fusion approach in the probabilistic setting and an uncertainty propagation approach in the possibilistic setting.

The probabilistic counterpart of the fuzzy approach, that is a probabilistic information fusion method, may run as follows. Let $\mathcal{D}$ denote the domain encompassed by the flow and stock balance equations, and $P(x)$ be the joint distribution of the measured data:

(1) Condition the joint distribution $P(x)$ of data on the balanced flow domain $\mathcal{D}$;
(2) Compute the projections of $P(x|\mathcal{D})$ on the parameter ranges;
(3) Extract means for all parameters, and the covariance matrix.

Clearly, if $P(x)$ is a multidimensional Gaussian function, it may fail to be the case for the resulting distributions (e.g. if the domain is $\mathcal{D}$ bounded, or for instance the symmetry of distributions may be lost). On the contrary the usual statistical approach to the reconciliation process preserves the Gaussian nature of the inputs when the model equations are linear (Narasimhan and Jordache 2000). In any case, computing the distribution of the maximum likelihood estimate is different from projecting the conditional probability over the domain of reconciled values on

each parameter space. We have seen in the case study that the distribution of best least squares estimates may fail to fit inside the ranges of the input data, when they are approximated by Gaussian functions.

Conversely, one can envisage possibilistic reconciliation in the spirit of an estimation procedure followed by a sensitivity analysis step:

- Choose a preferred norm (via a t-norm and a shape of π_i) and form the corresponding error criterion
- Compute a vector of optimal reconciled values y^* as a function, say f, of measured values $\hat{x}$.
- Compute the possibilistic uncertainty on reconciled values by sensitivity analysis using the imprecision of measured values: $\tilde{Y}^* = f(\tilde{X})$ where $\tilde{X}$ is the fuzzy set vector around the initial estimate $\hat{x}$.

Just as in the statistical method, the resulting fuzzy intervals are not necessarily upper bounded by the fuzzy sets originally assigned to input values. The study and implementation of these alternative approaches to reconciliation is left for further research.

8. Conclusion

In the context of the material 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 treat the problem as one of information fusion than as a pure statistical estimation based on random measurements. As a consequence, it sounds more reasonable to practically justify the choice of a distance (l_1, l_2, l_∞, ...) for minimizing the error rather than to invoke the Central Limit Theorem to justify the least squares method. A fuzzy-set approach to data reconciliation has been proposed. Its advantages are:

- Its flexible setting for representing various kinds of imprecise information items.
- Its clear conceptual framework as an information fusion problem. The reconciled ranges around the reconciled values are also more easy to interpret than the reconciled variances, as they result from the conjunctive merging of all available information items.
- Its general framework: in a formal sense, it recovers the least squares method by a proper choice of a shape for membership functions and of a conjunction operation, without betraying the principle of maximum likelihood.
- The possibility of solving the problem in the max-min case using standard linear programming methods and software.

However, this fuzzy constraint-based data reconciliation framework is conceptually at odds with the usual probabilistic reconciliation methods where the flow measurements are viewed as random variables affecting the optimal estimates, and not as additional constraints to be merged with the flow model. Further developments are needed in order to

- Study more examples where the max-min and the least squares approaches provide disagreeing results, so as to refine the comparison outlined here.
- Compare the information fusion and the estimation approaches inside the probabilistic and possibilistic paradigms respectively, so as to better understand when one approach is more cogent than the other one.

A software for fuzzy constraint-based approach has been built with a view to apply it to the analysis of material flow of rare earth elements in the anthroposphere of the EU-27.

Appendix: Modeling data using fuzzy intervals

Intervals have a limited expressive power. One is led to a dilemma between safeness and precision. Namely, short intervals are unreliable, and large intervals are uninformative. However a very simple and convenient, yet much more expressive, generalisation of intervals consists of fuzzy intervals (Dubois 2006) representing possibility distributions on the real line. A possibility distribution is a mapping $\pi : \mathbb{R} \to [0, 1]$ such that $\pi(r^*) = 1$ for some $r^* \in \mathbb{R}$: it is a normal fuzzy set (Zadeh 1978). 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 (Dubois et al. 2004). 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 $C(\pi) = \{r, \pi(r) = 1\}$ and contained in the support. The simplest representation of a fuzzy interval is a trapezoid defined by its core and its support. Note that this format is very convenient to gather information from experts in the form of nested confidence intervals, or more basically in the form of one safe interval and a plausible value.

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 (Dubois 2006) for references. Functions Π and N can be shown to compute exact probability bounds in the sense that:

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

In fact, it can be shown (Dubois et al. 2004) that $\mathcal{P}(\pi)$ is characterised by the α-cuts of π:

$$\mathcal{P}(\pi) = \{P : P(\{r : \pi(r) \geq \alpha\}) \geq 1 - \alpha, \forall \alpha > 0\},$$

thus suggesting that a possibility distribution is a kind of two-sided cumulative probability distribution. Probabilistic inequalities yield examples of such possibility distributions. 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 (Dubois et al. 2004). Gauss inequality also provides such possibility distributions encompassing probability distributions with fixed mode and standard deviation as pointed out in (Mauris 2011). It yields a triangular (bounded) fuzzy interval if probability distributions have bounded support. Hence a possibility distribution may account for incomplete statistical data (Dubois et al. 2004).

In the framework of measurement problems, Mauris (2007) has suggested that

in the case of competing error functions (empirical probability distributions $p_i, i = 1 \dots k$, such as Gaussian, uniform, double exponential, etc.), 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, in the setting of imprecise probability theory, $\mathcal{P}$ should be convex, typically the convex hull of $\{p_i, i = 1 \dots k\}$ (Walley 1991). Alternatively, 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 (2007). This representation is much simpler, even if more imprecise. This remark gives some foundation to the idea of using fuzzy intervals for representing measurement-based imprecise statistical information.

Conversely, if an expert provides a probability distribution that represents subjective belief, it is possible to reconstruct a possibility distribution by reversing the Laplace principle of indifference (Dubois et al. 2008). 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 (Yager 1982). 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 (Dubois et al. 2008). Note that the symmetric triangular possibility distribution over a bounded interval encompasses the uniform distribution on this interval (it is the most precise choice that retains symmetry) (Mauris 2007).

In summary, fuzzy intervals, and specifically triangular or trapezoidal possibility distributions, may account for uncertain information coming from various origins.

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". Thanks to Romain Leroux for his help on the copper example.

References

Alhaj-Dibo M., Maquin D. and Ragot J. 2008. Data reconciliation: a robust approach using a contaminated distribution. *Control Engineering Practice* 16 (2): 159-170.

Ayres R.U. and Kneese A.V. 1969. Production, consumption and externalities. American Economic Review 59(3): 282-297.

Baccini P. and Brunner H.P. 1991. *Metabolism of the Anthroposphere* Springer, New York.

Bellman R. E. and Zadeh L. A. 1970. Decision making in a fuzzy environment, *Management Science*, 17: B141-B164.

Benhamou F., Granvilliers L. and Goualard F. 2000. Interval Constraints: Results and Perspectives. *New Trends in Constraints*, LNAI 1865, 1-16, Springer, Berlin.

Bai S., Thibault J. , and McLean D.D. 2006. Dynamic data reconciliation: Alternative to Kalman filter *Journal of Process Control*, 16: 485-498.

Bonnin M., C. Azzaro-Pantel, *et al.* 2013. Development of a Dynamic material flow analysis model for French copper cycle, *Chemical Engineering Research and Design*, 91(8):1390-1402.

Brunner P.H. and Rechberger H. 2004. *Practical Handbook of Material Flow Analysis*. Lewis Publishers.

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

Crowe C. 1996. Data reconciliation - progress and challenges. *Journal of Process Control*, 6: 89-98.

Durance M.-V., Brochot S. and Mugabi M. 2004. Material balance approach for parameter determination in bioleaching process. In: *Computer Applications in Biotechnology* (M.-N. Pons and J.F. Van Impe, Eds). Proc. 9th IFAC International Symposium, Nancy, France.

Descloux J. 1963. Approximations in l_p and Chebyshev approximations, *J. Soc. Indust. Appl. Math.* 11: 1017-1026.

Dubois D. 1987. An application of fuzzy arithmetics to the optimization of industrial machining processes", Mathematical Modelling, 9, 461-475.

Dubois D. 2006.Possibility theory and statistical reasoning *Computational Statistics & Data Analysis*, 51, 47-69.

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

Dubois D., Fargier H., and Guyonnet D. 2013. Data Reconciliation under Fuzzy Constraints in Material Flow Analysis. In: J. Montero et al. (Eds.) *Proc. European Society For Fuzzy Logic And Technology Conference* (EUSFLAT 2013), Milan, 25-32, Atlantis Press.

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

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

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

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

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

Frosch R.A. and Gallopoulos N.E. 1989. Strategies for Manufacturing. *Scientific American* 261(3), 144-152.

Gottschalk F., Scholz R. W. and Nowack B. 2010 Probabilistic material flow modeling for assessing the environmental exposure to compounds: Methodology and an application to engineered nano-TiO2 particles, *Environmental Modelling & Software* 25(3): 320-332.

Graedel T.E., van Beers D. *et al.* 2004. The multilevel cycle of anthropogenic copper. *Environmental Science & Technology* 38: 1253-1261.

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

Jain P. ,Townsend TG. and Johnson P. 2013. Case study of landfill reclamation at a Florida landfill site. *Waste Management*, 33(1): 109-116.

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

Kelly J.D. 2004. Techniques for solving industrial nonlinear data reconciliation problems. *Computers and Chemical Engineering*, 28: 2837-2843,

Kikuchi S. 2000. A method to defuzzify the fuzzy number: transportation problem application, *Fuzzy Sets and Systems*, 116(1): 3-9.

Klement E.P., Mesiar R. and Pap E. 2000. Triangular Norms, Kluwer, Dordrecht.

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

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

Mauris G. 2011. Possibility distributions: A unified representation of usual direct-

probability-based parameter estimation methods, *Int. J. of Approx. Reasoning*, 52(9): 1232-1242.

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

Ragot J. and Maquin D. 2004. Reformulation of data reconciliation problem with unknown-but-bounded errors. *Industrial and Engineering Chemistry Research*, 43(6): 1530-1536,

Ragot J., Maquin D and Alhaj-Dibo M. 2005. Linear mass balance equilibration: a new approach for an old problem. *ISA Transactions*, 44(1): 23-35,

Rice J. 1962. Tschebyschev approximation in a compact metric space, *Bull. Amer. Math. Soc.* 68: 405-410.

Steyaert H. , Van Parys F., *et al.* 1995. Implementation of piecewise linear fuzzy quantities, *Int. J. Intelligent Systems*, 10: 1049-1059.

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

Tan R. R. , Briones L. M. A. and Culaba A.B. 2007 Fuzzy data reconciliation in reacting and non-reacting process data for life cycle inventory analysis, *Journal of Cleaner Production*, 15(10): 944-949.

van Beers D., van Berkel R. and Graedel T.E. 2005. The application of material flow analysis for the evaluation of the recovery potential of secondary metals in Australia. *4th Australian LCA Conference*, Sydney, Australia.

Walley P. 1991, *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London.

Wolman A. 1965. The metabolism of cities. *Scientific American*, 213(3): 179-190.

Yager R.R. 1982. Level sets for membership evaluation of fuzzy subsets. In : *Fuzzy Sets and Possibility Theory : Recent Developments* (R.R. Yager, ed.), 90-97, Pergamon Press, Oxford.

Zadeh L.A. 1978. Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets and Systems*, 1:1–28.

Zimmermann H. -J. 1978. Fuzzy programming and linear programming with several objective functions *Fuzzy Sets and Systems* 1: 45-55.