A General Nonlinear Least Squares Data Reconciliation and Estimation Method for Material Flow Analysis

Grant M. Kopec, Julian M. Allwood, Jonathan M. Cullen, and Daniel Ralph

Keywords:

data reconciliation
industrial ecology
material flow analysis (MFA)
nonlinear constrained optimization
Sankey diagram
steel

Supporting information is available on the *JIE* Web site

Summary

The extraction, transformation, use, and disposal of materials can be represented by directed, weighted networks, known in the material flow analysis (MFA) community as Sankey or flow diagrams. However, the construction of such networks is dependent on data that are often scarce, conflicting, or do not directly map onto a Sankey diagram. By formalizing the forms of data entry, a nonlinear constrained optimization program for data estimation and reconciliation can be formulated for reconciling data sets for MFA problems where data are scarce, in conflict, do not directly map onto a Sankey diagram, and are of variable quality. This method is demonstrated by reanalyzing an existing MFA of global steel flows, and the resulting analytical solution measurably improves upon their manual solution.

Introduction

As noted by Brunner and Rechberger (2004), material flow analysis (MFA) data are normally aggregated from a range of different sources, including both direct and proxy measurements, using diverse methods of data collection and processing with a range of qualities. This is particularly true when MFAs cross many industries and data collection entities. Cullen and Allwood (2013, 3060) document the sources of uncertainty in their MFA of global aluminium flows "... misinterpretation of survey questions and terminology used to describe process and materials; unintentional or deliberate misreporting of data in the surveys; incomplete coverage of ... facilities, requiring data to be scaled; calculation errors in the aggregation of data; and miscommunication of data in published reports."

The reconciliation of material flows for which there are contradictory and incomplete data is often completed through the use of ad-hoc methods of data choice, estimation, and manipulation to ensure that such networks are self-consistent. A formal, quantitative method to reconcile material flows, making best possible use of all known data, constraints, and expert opinion, is needed.

A number of techniques have been documented for data reconciliation of nonmatching data points, and estimation of the value of variables where no data exists, in MFA. However, such methods are generally restricted to the use of basic classifications of flow data and linear constraints in the reconciliation. We demonstrate that "unconventional" data forms may exist as linear and nonlinear constraints to the MFA analysis. This study extends current MFA data reconciliation and estimation methods by formulating:

1. A mathematical classification system for incorporating "unconventional" forms of data that do not directly map onto Sankey diagrams.
2. A method for data reconciliation and estimation based on a least squares minimization of total error in a Sankey diagram using nonlinear optimization, with quantitative weighting of data source quality and allowing for multiple data points for any variable.

Address correspondence to: Julian M. Allwood, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom. *Email*: jma42@cam.ac.uk

DOI: 10.1111/jiec.12344
Editor managing review: Reinout Heijungs

Volume 00, Number 0

A method to extend MFA methodology by classifying diverse input data forms, and estimating and reconciling data, is presented here, using the MFA of global steel production and use by Cullen and colleagues (2012b) as an example of its application.

Literature Review

The basis of an MFA analysis is the construction of a weighted, direct network. Such networks consist of a series of processes connected by material flows. In the network literature, the terminology changes to "nodes connected by edges." Data on processes are normally given as a quantity over a unit of time (e.g., kilograms per day), and material flows are defined as either a quantity or as a fraction of their source process, the latter known as a transfer coefficient (TC). As defined in Schmidt (2008a, 2008b), this article will use the term Sankey diagram to refer to weighted and directed networks for MFA analysis and will use the terminology "nodes connected by edges" to describe parts of the Sankey diagram.

MFAs may have three distinct data problems: (1) There may not be a complete set of data for the problem of interest; (2) there may be nodes or TCs where multiple, conflicting data points exist; and (3) the assembled data set for nodes or TCs would violate the principle of conservation of mass because of errors in the data. These data problems result in considerable uncertainty in many MFA analyses, with Brunner and Rechberger (2004) and Fischer-Kowalski and colleagues (2011) both asserting that uncertainties of ±10% are common in MFA analyses.

Perhaps the most common strategy for data reconciliation to ensure that mass is conserved is to use a qualitative assessment of the relative quality of the available data sources, and then either manipulating the data quantity or choosing the data or data source that appears to fit best or that is most consistent. Cullen and colleagues (2012b), Brunner and Rechberger (2004), and Bringezu and colleagues (2003) all acknowledge manipulation based on qualitative assessment of data sources. In particular, Bringezu and colleagues (2003) note that, with data collected through Eurostat, direct material input and direct material consumption data do not match and must be manually manipulated to form the basis of coherent analyses. The documented errors in data used for MFA analysis are indicative of the need for systematic treatment of these errors. Based on their review of economy-wide material flow accounting, Fischer-Kowalski and colleagues (2011, 868) conclude that "a major problem concerning the integration of physical trade data into MFA accounts . . . is the lack of standardized procedures to handle the manifold data gaps and flaws in the primary data."

Laner and colleagues (2014) present a systematic review of quantitative methods for data reconciliation in MFAs. However, the choice of reconciliation method will normally be determined by the forms of data available for a particular analysis. This literature review is organized to associate the available data reconciliation methods with the forms of data that they accept. In doing so, it can be demonstrated that existing methods do not necessarily incorporate the full range of data types that may be commonly available for MFA analyses.

Based on this review of data reconciliation methods for MFA-style weighted and directed flows, a procedure may have some combination of the following capabilities for reconciling data forms:

(1) Reconciliation with node data;
(2) Reconciliation with TC (flow fraction) data;
(3) Reconciliation with flow quantity (edge width) data;
(4) Reconciliation with external or "unconventional" data forms that may not directly map onto the Sankey diagram format;
(5) Reconciliation that may incorporate multiple data points for each variable;
(6) Reconciliation that incorporates information on upper and lower bounds for the variables;
(7) Reconciliation that incorporates information on the likely probability distributions of variables;
(8) Reconciliation that incorporates measures of data quality.

Formal, quantitative data reconciliation methods appropriate for MFA analysis can be classified into four categories: (1) linear methods; (2) constrained optimization; (3) Bayesian techniques; and (4) the RAS family of input-output (I-O) matrix data reconciliation techniques.

A linear method for data reconciliation is documented in Brunner and Rechberger (2004), where a linear least squares regression is used to reconcile node data to conform to conservation of mass constraints, while assuming that all TCs are constants, and thus free of error. However, errors in TC data may be of similar magnitude to those errors in node data and should not in general be ignored. Matyus and colleagues (2003) document an MFA data reconciliation procedure for node data and generally assume that TCs are constants, but describe a two-optimization procedure to include a TC data point with a residual. The first optimization solves the problem without the TC data or TC equation, and the second optimization assumes that the TC value is constant, calculated as a mean of the data and solved values. However, this procedure results in a TC value that is only an average of the data point and the value from the first optimization, so the value may not be a best-fit value. Additionally, the procedure would be unlikely to produce a valid solution for an MFA with a large number of TC data constraints because of the elimination of a large number of TC equations. These methods do not provide means of quantitatively comparing the quality of data sources that may be used in an analysis. Linear least squares methods described here make the implicit assumption that data for an MFA analysis will only take the form of nodes or TCs.

A number of constrained optimization approaches have also been formulated, primarily derived from the work of Van der Ploeg (1985, 1988). These are generally extensions of linear least squares methods. Fellner and colleagues (2011) structure a nonlinear constrained optimization data reconciliation technique with bilinear equations to reconcile mass balance data in fuels, though the formulation does not include inequality constraints, incorporation of external data, or information about data quality. This technique is also used in the STAN material flow analysis software, where it is further described by Cencic and Rechberger (2008). The STAN software package data reconciliation algorithm reconciles TC and flow quantity data and bounds, and allows for additional linear relationships (ALRs) to be defined between flows, but does not incorporate data quality measures (besides bounds), unconventional data, errors in ALRs, or multiple data points for each variable. This technique also assumes that all variables follow a normal probability distribution, though often a variable may follow another probability distribution, if it is known at all.

Narasimhan and Jordache (2000) formulate a general least squares approach to data reconciliation in problems with nonlinear constraints. This formulation relates measurements of flow rate, temperature, and pressure, so the specific equations may be less appropriate for MFA data reconciliation, though the basic methodology is applicable to MFA. However, this formulation does not allow for multiple data points per variable, nor does it explicitly incorporate measures of data quality or specific equations for unconventional data entry.

Cencic and Frühwirth (2014) describe a Bayesian technique for data reconciliation in MFA that accepts probability distributions other than the normal distribution for variables. Such probability distributions need to be known or estimated before reconciliation. However, the technique is only valid for linear equality constraints, and any data quality information must be reflected in the known previous joint probability distribution.

These techniques both take the form of weighted and directed networks, so techniques used to balance economic and social accounting matrix I-O tables (IOTs) can be applied to MFA analysis. The RAS (named for "Richard A. Stone") family of matrix balancing data reconciliation techniques, commonly used to balance IOTs, ensures that row and column sums (the nodes/processes in an MFA) are consistent with the sum of the I-O matrix elements (the "flows" in MFA). Lenzen and colleagues (2009) provide descriptions of RAS and its variants, where the original RAS algorithm is a biproportionate scaling of an I-O matrix of initial guesses of either transfer coefficients or real flows, using known column and row sums, which would represent the process values in MFA. Further developments of the RAS method include MRAS, where some transfer coefficients are known with certainty, and KRAS, the most developed of the RAS family of algorithms. The RAS family of algorithms, particularly KRAS, is well suited to MFA data reconciliation problems where data are available as a set of process values, an estimated set of either TCs or flow data, any external linear constraints on process values, as documented by Lenzen and colleagues (2014), and some data quality measure. However, the inclusion of nonlinear data constraints is precluded, and a user must choose between using TC or flow data for the reconciliation.

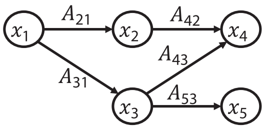

Figure 1 Example of a material flow system.

This review indicates that, although there are many MFA data reconciliation procedures, those that can incorporate a wide selection of different forms of data, multiple data points per variable, as well as measures of data quality, are relatively underdeveloped. In particular, the use of unconventional data forms is often not acknowledged in the literature on MFA data reconciliation, and where it is, such unconventional data, such as the ALRs in the STAN software, are assumed to be constants rather than variables with their own attendant error. In this article, we will provide a formulation of a general data reconciliation procedure that can incorporate measures of data quality as well as a wider array of data forms than current methods.

Material Flow Networks

The Sankey diagram provides a basis for structuring MFA data. It is a weighted, directed graph where nodes represent material or energy in a particular state and are connected by edges to represent the flow or transformation of material from the node representing a process i to another node representing process j. The width of an edge is scaled to the size of the flow. The basic structure of a Sankey diagram is defined in equation (1), where the value of node x_j is determined by the set of TCs A_{ji} and the set of nodes x_i, where TCs define the fraction of material from a node that is transferred by an edge. In order to conserve mass, equation (2) holds. Equations (1) and (2) are the complete set of equations necessary to describe the structure of a Sankey diagram, the structural constraint equations. The Sankey diagram is assumed to be a highly structured network, where all nonzero nodes and TCs are known to be nonzero and are defined by equation (1).

$$x_j = \sum_i A_{ji} x_i, \ \forall j \tag{1}$$

$$\sum_j A_{ji} = 1, \ \forall i \tag{2}$$

We present a simple methodological example of such an MFA in figure 1. In this system, there are five processes, numbered x_1 through x_5, and five flows between those processes. This example will be used throughout this article to describe the formulation of unconventional forms of data and the setup and solution of the data reconciliation procedure.

Conventional and Unconventional Data Input

Once the Sankey diagram structure for a particular problem is defined, data are entered both for nodes and TCs. A data point for x_i or A_{ji} is denoted by $\hat{x}_{i,q}$ or $\hat{A}_{ji,r}$, respectively, where q and r are particular data points. Given that data points $\hat{x}_{i,q}$ and $\hat{A}_{ji,r}$ will have some error, each has an associated normalized residual $r_{i,q}$ or $r_{ji,r}$, respectively. Equations (3) and (4) describe the relationship between the reconciled value of x_i or A_{ji} and their data points $\hat{x}_{i,q}$ and $\hat{A}_{ji,r}$.

$$x_i = \hat{x}_{i,q}(1 + r_{i,q}) \tag{3}$$

$$A_{ji} = \hat{A}_{ji,r}(1 + r_{ji,r}) \tag{4}$$

Equations (3) and (4) are the data constraint equations, which describe the conventional data that populate the Sankey diagram, whereas the structural constraint, equations (1) and (2), describe the form of a Sankey diagram. As reflected in the form of equations (3) and (4), all variables are assumed to be independent.

Information that is useful to populate a Sankey diagram may exist in forms other than nodes or TCs and is represented by the variable γ_k. Perhaps the most common of these 'unconventional' data forms is data about a flow magnitude (or edge width) w_{ji}, where $\gamma_k = w_{ji} = A_{ji}x_i$. Alternatively, information may exist that indicates that the sum of a set of nodes x_i, $\alpha \le i \le \beta$ is equal to a quantity γ_k, such that $\sum_{i=\alpha}^{\beta} x_i = \gamma_k$. Data of forms that do not conform to the conventional MFA node or TC data types are designated unconventional data forms. Though the literature review covered a number of data reconciliation methods used in MFA that accept data of various types, many unconventional forms of data were classified in the data set used by Cullen and colleagues in their global steel analysis. Table 1 classifies seven linear or nonlinear unconventional data variables γ_k and their relationships between the conventional node or TC variables of a Sankey diagram.

Equation (5) specifies the relationship between γ_k and data point $\hat{\gamma}_{k,s}$, where s is the data point, analogous to equations (3) and (4). Equation (6) specifies any lower bounds l or upper bounds u on node, TC, or unconventional data constraint variables.

$$\gamma_k = \hat{\gamma}_{k,s}(1 + r_{k,s}) \tag{5}$$

$$\begin{aligned} &x_i \ge 0,\ \forall i \\ &0 \le A_{ji} \le 1,\ \forall\, i, j \\ &l_i \le x_i \le u_i,\ \textit{for some } i \\ &l_{ji} \le A_{ji} \le u_{ji},\ \textit{for some } i, j \\ &l_k \le \gamma_k \le u_k,\ \textit{for some } m \end{aligned} \tag{6}$$

Thus, the constraint equations for an MFA consist of a set of structural equations (1) and (2), a set of conventional and unconventional data constraint equations (3) through (5), any upper and lower bounds on variables in equation (6), and a set of unconventional data constraint equations from table 1. Likely forms of unconventional data constraints are documented in table 1, though the construction of other unconventional data constraint forms beyond those in table 1 is possible, depending on the data available for a particular analysis.

Table 2 in the next section documents the conventional and unconventional data forms used for the methodological example. The variable alt_{21} defines the width of the edge (amount of the flow) going from x_1 to x_2, whereas the variable alt_{42} defines the contribution of flow A_{42} to the value of node x_4. These unconventional data constraints make the problem infeasible for data reconciliation methods documented in the literature review.

A Nonlinear Program for Data Reconciliation and Estimation

Once a material flow network is designed, and data are collected and classified into equations (3), (4), (5), and (6), as well as the unconventional constraint forms in table 1, a nonlinear constrained optimization program can be constructed to reconcile data points and fill data gaps, deviating as little as possible from the available data. Therefore, the objective function is constructed as a least squares-type function to minimize the sum squared residuals r from equations (3), (4), and (5). The equations in table 1 form the set of constraints for the nonlinear program.

Owing to the diversity of data sources that will be used in an analysis, it is useful to assign a quantitative quality measure $\Phi_{i,q}$, $\Phi_{ji,r}$, or $\Phi_{k,s}$ to each data point, where $1 \le \Phi \le 100$. $\Phi_{i,q}$ represents the data quality of the data point $\hat{x}_{i,q}$ for node i, $\Phi_{ji,r}$ is the quality for data point $\hat{A}_{ji,r}$ for TC A_{ji}, and $\Phi_{k,s}$ is the quality for data point $\hat{\gamma}_{k,s}$ of unconventional data constraint k. In this case, data points that are qualitatively judged to be of higher quality are assigned a larger-quality value. Thus, the objective function is given in equation (7), which normalizes the objective function about each variable by dividing by the number of data points for each variable so that no variable is disproportionately weighted because of the total number of data points that exist for a node (Q_i being the total number of data points for node x_i), TC (R_{ji} being the total number of data points for TC A_{ji}), or unconventional data variable (S_k being the total number of data points for unconventional data constraint γ_k). I, J, and K are the set of nodes, TCs, and unconventional data constraints with any associated data points, respectively.

$$\begin{aligned} f(r) = &\sum_I \frac{\sum_{q=1}^{Q_i}\left(\Phi_{i,q}\cdot(r_{i,q})^2\right)}{Q_i} \\ &+\sum_J \frac{\sum_{r=1}^{R_{ji}}\left(\Phi_{ji,r}\cdot(r_{ji,r})^2\right)}{R_{ji}} \\ &+\sum_K \frac{\sum_{s=1}^{S_k}\left(\Phi_{k,s}\cdot(r_{k,s})^2\right)}{S_k} \end{aligned} \tag{7}$$

Quality values contribute proportionally to the objective function value. Thus, taking two variables x_i and x_j, each with single data points only, then the gradient of the objective function will be larger in the x_i dimension by a factor of $\frac{\Phi_{i,1}r_{i,1}}{\Phi_{j,1}r_{j,1}}$. Thus, where $\mathrm{r}_{i,1} = \mathrm{r}_{j,1}$, the solver will preferentially reduce $\mathrm{r}_{i,1}$ by a factor of $\frac{\Phi_{i,1}}{\Phi_{j,1}}$, and this understanding can be used when assigning relative confidence values to data points.

Table 1 Description of unconventional data constraint forms considered in this study, with examples from the global steel MFA of Cullen and colleagues (2012a)

Description	*Equation form*	*Example from Cullen and colleagues (2012a)*
Linear constraints		
a. Sums of variable subsets	$\sum_{i=\alpha}^{\beta} x_i = \gamma_k$ (for a node sum) $\sum_{i,j=\alpha,\beta}^{A,B} A_{ji} = \gamma_k$ (for an allocation sum) $\sum_{l=\alpha}^{\beta} \gamma_l = \gamma_k$ (for an unconventional data sum, e.g., edge widths or inverse allocations)	Sum data constraints describe the value of a sum of nodes, TCs, edge widths, or other constraints. For example, it is estimated that total production of all tubes, bars, rods, and sections in figure S1 in the supporting information on the Web is 553 Mt.
Nonlinear constraints		
b. Edge widths	$\gamma_k = A_{ji} x_i = w_{ji}$	Edge width constraints define the width of an edge going from a node x_i to another node x_j. For example, the amount of material flowing from blast furnaces to electric arc furnaces in figure S1 is thought to be 44.6 Mt.
c. Inverse TCs	$\gamma_k x_j = A_{ji} x_i = w_{ji}$	The inverse TC describes the portion of a node x_j that is derived from a contributing node x_i. For example, 95% of inputs into section mills are thought to come from blooms in figure S1.
d. Additional linear relations (ALRs)	$x_j = \gamma_k x_i,\ j \neq i$ $A_{ji} = \gamma_k A_{lm},\ j \neq l \, and \, i \neq m$ $\gamma_m = \gamma_k \gamma_l,\ j \neq i$	Nodes, TCs, or unconventional constraints may have a relationship that does not involve the transfer of material, so will not be reflected in the structural constraints in equation (1). This is equivalent to the additional linear relations defined by Matyus and colleagues (2003). For example, it is estimated that losses from direct reduction are the same as those from blast furnaces in figure S1, so in this case, $\gamma_k = 1$.
e. Percentage of sums of variable subsets	$x_l = \gamma_k \sum_{i=\alpha}^{\beta} x_i$ (for a node sum) $A_{lm} = \gamma_m \sum_{i,j=\alpha,\beta}^{A,B} A_{ji}$ (for an allocation sum) $\gamma_m = \gamma_k \sum_{l=\alpha}^{\beta} \gamma_l$ (for an unconventional constraint sum)	A percentage of a sum constraint defines the value of a node, TC, edge width, or other constraint as a percentage of the sum of a set. For example, in figure S1, it is estimated that, of the sum of losses and internal recycling flows for ingots, 25% is losses and 75% is internally recycled. These constraints are also used where a TC is defined as a fraction of the nonloss quantity, e.g., 30% of the nonloss output (yield) from hot strip mills is sent to galvanizing plants. Where the desired value is part of the set (e.g., $l \in i$ for nodes), the sum is inclusive, and where it is not (e.g., $l \notin i$ for nodes), the sum is exclusive. Inclusive vs. exclusive percentage of sum constraints need to be differentiated because their gradients take different forms in the optimization program.
f. Pro-rata inverse allocations	$\gamma_k = \gamma_l,\ k \neq l$ $\Rightarrow \frac{A_{ji}}{x_j} = \frac{A_{lm}}{x_l},\ i \neq m$	Pro-rata constraints describe data where two inverse TCs are equal. For example, liquid steel from electric arc furnaces flows to blooms and billets on a pro-rata basis, with the percent of blooms and billets that come from electric arc furnace-produced steel equal.
g. Sequential multiplications	$x_i = \gamma_k \gamma_l,\ k \neq l$ $A_{ji} = \gamma_k \gamma_l,\ k \neq l$	Sequential multiplication constraints describe many types of data that are multiplied together. For example, the aggregate efficiency of a blast furnace in figure S1 is derived from a number of small steps, each with a separate efficiency. These efficiencies were multiplied together to create an aggregate efficiency for blast furnaces.

Note: MFA = material flow analysis; TCs = transfer coefficients; Mt = megatonnes.

With the objective function and the constraints from table 1, the constrained nonlinear program is defined as:

$$\text{minimise}: \sum_{I} \frac{\sum_{q=1}^{Q_i} \left(\Phi_{i,q} \cdot (r_{i,q})^2 \right)}{Q_i} + \sum_{J} \frac{\sum_{r=1}^{R_{ji}} \left(\Phi_{ji,r} \cdot (r_{ji.r})^2 \right)}{R_{ji}} + \sum_{K} \frac{\sum_{s=1}^{S_k} \left(\Phi_{k,s} \cdot (r_{k,s})^2 \right)}{S_k}$$

subject to: **structural constraints**

$$x_j = \sum_i A_{ji} x_i, \ \forall j$$

$$\sum_j A_{ji} = 1, \ \forall i$$

and **data constraints**

$$x_i = \hat{x}_{i,q}(1 + r_{i,q}), \qquad \textit{for some } i$$

$$A_{ji} = \hat{A}_{ji,r}(1 + r_{ji,r}), \qquad \textit{for some } i, \ j$$

$$\gamma_k = \hat{\gamma}_{k,s}(1 + r_{k,s}), \qquad \forall m$$

and unconventional data constraints from *Table* 1.

Table 2 List of the variables, data points with data quality, and reconciled solution for the MFA in figure 1

Variable	*Type*	*Equation form*	*Data points (quality)*	*Final value*	*Residuals*
x_1	Node	n/a	none	25.2	
x_2	Node	n/a	None	6.8	
x_3	Node	n/a	15 (60) | 18 (90)	18.4	0.23 | –0.03
x_4	Node	n/a	None	15.3	
x_5	Node	n/a	10 (60)	10.0	0.00
A_{21}	TC	n/a	None	0.27	
A_{31}	TC	n/a	None	0.73	
A_{42}	TC	n/a	None	1.00	
A_{43}	TC	n/a	None	0.46	
A_{53}	TC	n/a	0.6 (90)	0.54	–0.1
alt_{21}	Edge width	$alt_{21} = A_{21}x_1$	20 (30)	6.8	–0.66
alt_{42}	Inverse TC	$alt_{42}x_4 = A_{42}x_2$	0.4 (90)	0.45	0.12

Note: MFA = material flow analysis; TC = transfer coefficient; n/a = not applicable.

and **boundary constraints**

$$\begin{aligned} &0 \le A_{ji} \le 1,\ \forall i, j \\ &x_i \ge 0,\ \forall i \\ &l_i \le x_i \le u_i, \qquad \textit{for some } i \\ &l_{ji} \le A_{ji} \le u_{ji},\ \textit{for some } i, j \\ &l_k \le \gamma_k \le u_k, \qquad \textit{for some } m \end{aligned}$$

Because any data type may have error associated with it, some of the constraints in this program contain multiplicative nonlinearities, generally bilinear, including the mass flow structural constraints and some unconventional constraints. Thus, the set of constraints is not generally a convex set. The pooling problem, as defined by Greenberg (1995) and others, formulates mass flow as a set of nonlinear (bilinear) constraints similar to the formulation in this article. Thus, this problem is a bilinear, global, nonconvex program. However, pooling problem formulations generally assume perfect data and use the objective function to optimize a cost metric for the TC of source material to a set of intermediate pools and final products. The formulation of pooling problems under conditions of stochastic or uncertain data is limited to the use of data scenarios, as Li and colleagues (2011) describe in their work on design of natural gas networks.

When the form of the objective function is included, the program is a quadratically constrained quadratic program, as described by Mehanna and colleagues (2015). Owing to nonconvexity, global optimization methods must be used to solve the constrained nonlinear program. Because such a program may have many local minima, an initial set of guesses g_o is supplied so that the numerical constrained nonlinear optimization solver starts in a well of attraction, increasing the probability that the solution will be a local minimum that is superior to a manual solution. For this implementation, Matlab's fmincon algorithm, documented in the Matlab documentation (Matlab 2013) and by Byrd and colleagues (1999), was used. The fmincon solver uses an interior point algorithm to find a minimum of the objective function. As noted by Misener and Floudas (2010), similar solvers are used in the "pooling problem" formulation.

The validity of a solution is assessed by examining whether the maximum constraint violation is within limits. Once the validity of a result has been determined, the quality of the result is based on how well the solution fits the available data and can be assessed according to: (1) the value of the objective function, which can be compared between solutions, and (2) the value of individual residuals as well as the average and standard deviation of the absolute value of the residual set.

In summary, the proposed methodology consists of the following steps:

(1) Design a material flow network.
(2) Collect data to populate the node, TC, and any unconventional data constraints, recording relative quality values for the data.
(3) Place the structural and data constraint equations, as well as the objective function, into forms that can be used by a numerical nonlinear optimization solver and solve.
(4) Review the results of the optimization. If the solver cannot find a feasible solution that satisfies all constraints, or if a feasible solution is found that does not conform to expectations, there are several options: (a) data quality values may be updated; (b) initial values g_o may be added so that the solver starts in a well of attraction that is more likely to provide a satisfactory solution; or (c) structural or data constraints may be missing and should be added to the program. The MFA can then be iterated starting at step 1 or 2, as appropriate.

The methodological example contains 15 constraint equations (nine linear and six nonlinear), as seen in table 2, and a feasible solution was found in approximately 1 second. As would be expected, the solution deviates further from the data points with the lowest-quality ratings.

Results of the Global Steel Reanalysis

The effectiveness of this reconciliation method was tested against the "manual" MFA of global steel flows from Cullen and colleagues (2012a,b). The manual analysis represents the researchers' best estimate, through qualitative selection and

Table 3 Global steel reanalysis problem setup

a)	
Type of variables	*Quantity*
Node variables	211
TC variables	389
Unconventional variables	58
Total variables[a]	**658**
b)	
Type of constraints	*Quantity*
Structural constraints	**390**
Node data constraints	33
TC data constraints	121
Unconventional data constraints	58
Total constraints[b]	**602**

The number of a) variables and b) constraints for the global steel flow Sankey nonlinear constrained optimization program.

[a] Excludes residual variables.

[b] Excludes boundary constraints.

TC = transfer coefficient.

Table 4 Results

Value	*Cullen and colleagues*	*This study*
Objective function, equation (7)	145.62	99.05
Objective function without confidence values	1.95	1.35
Maximum residual value	0.887	0.692

Note: Comparison of the results of this study with results from Cullen and colleagues.

manipulation of data, of the flow of global steel. In the main article and the supplementary information, the researchesr document their manipulation of data points in order to satisfy the structural constraints of mass flow and conservation of mass. For example, data from the World Steel Association (2010) indicates that 44.5 million tonnes (Mt) of steel were used for welded tube in 2009, but the value that Cullen and colleagues (2012a) assign is 62.4 Mt in order to ensure that conservation of mass constraints are met. The Sankey diagram produced by Cullen and colleagues (2012a) is shown in figure S1 in the supporting information available on the Journal's website.

Structuring this analysis according to the system of nodes, TCs, and unconventional data constraints gives the problem in table 3, while also providing a consistent structure for the Sankey diagram in Cullen and colleagues (2012b) and the network diagram in the supplementary information of Cullen and colleagues (2012a). The analysis by Cullen and colleagues involves a large number of nodes, TCs, and data sources. It was clearly a large undertaking and is likely at the limit of what is possible using a manual method for MFA construction and reconciliation.

The Sankey diagram structure used in this analysis is functionally the same as that described in the supplementary information of Cullen and colleagues (2012a), though some nodes and edges have been added to ensure mathematical consistency. The system of equations is underdetermined, with more variables to solve for than constraints, demonstrating the utility of the proposed method, given that such a system potentially has many solutions unless an objective function is used. Of 600 node and TC variables to solve for, data were directly provided for only 154 nodes or TCs.

The reconciliation was implemented using only the data sources cited by Cullen and colleagues in their article and supplementary information (2012a, 2012b) so the solutions are comparable. Quality values Φ for each data point were determined through consultation with the original authors, based on the trustworthiness of each data source. Data points from sources considered to be more reliable and representative of the global steel industry, such as the World Steel Association statistical yearbook (2010), were given relatively higher-quality values ($\Phi_{WSA} = 90$) than data points from regional sources that might not be representative of the global industry, or sources considered to be less reliable. For example, data points from the Japan Iron and Steel Federation might not be representative of the global steel industry, so were given a lower-quality value ($\Phi_{JISF} = 60$).

The initial values g_0 used in this reanalysis were, in decreasing order of preference, single data points, averages of multiple data points, and the solution of Cullen and colleagues (2012a). Thus, the set of initial values did not constitute a feasible solution, but did represent the best information available. It was found that, for a nonlinear program of this size, reasonable initial values need to be provided in order to arrive at a feasible solution.

The final value of the objective function was calculated for both this method and for Cullen and colleagues' manual solution, with the result shown in table 4. The value of the objective function is the determinant of the quality of the solution, and the value of the objective function from this study is superior to that from the manual solution of Cullen and colleagues (2012b). As a robustness check, the value of the objective function with all quality values set to 1 is also calculated for the solution of the weighted problem, and the value from this study also is superior to Cullen and colleagues.

Figure 2 shows the individual and cumulative values of the residuals for five histogram bins, and it shows that the final value of the objective function for Cullen and colleagues' solution was dominated by residuals of value greater than 0.25, whereas the final value of the objective function for the numerical solution was dominated by the residuals greater than 0.5. Thus, the algorithm was able to find a better solution, as defined by the objective function, by reducing large (>0.25) value residuals.

The histogram in figure S2 in the supporting information on the Web shows distribution of the values of the residuals for the numerical answer and Cullen and colleagues. For Cullen and colleagues, the majority of residuals are zero, given that most data points were used "as is," whereas a relatively small number of data points had large residuals to reflect conflicting data sources or where manual adjustments were made to ensure that conservation of mass constraints were met. In contrast, the

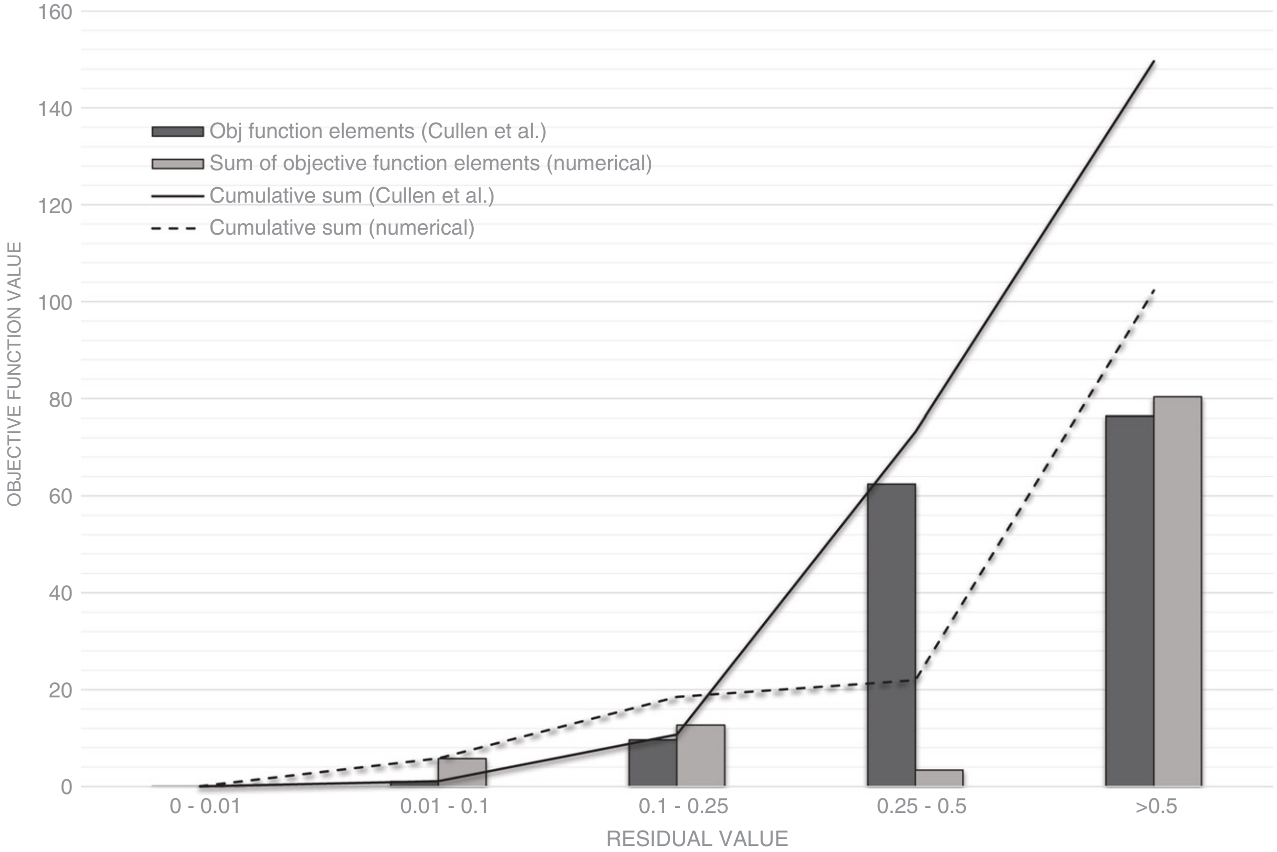

Figure 2 Sum of objective function values. A comparison of the contribution to the objective function value from residuals in histogram bins.

numerical solution has many residuals with small values that do not contribute significantly to the final objective function value, but are necessary in order to meet all constraints. The least squares-type objective function ensures that small residual values are penalized less than large ones, so it follows that the optimized solution would redistribute error so that a larger number of variables have small residuals, whereas the manual solution would have a relatively higher number of variables with large residuals.

The final results for the global steel analysis show that this solution surpasses Cullen and colleagues' solution with respect to the objective function metric. This yields additional flexibility and capability for interactive analysis of data inconsistency. Table S1 in the supporting information on the Web shows that the sets of variables in this solution and Cullen and colleagues' solution that have a deviation of greater than 10% from the data wholly overlap. This total intersection indicates that the method can produce superior results, as defined by the objective function and constraints, compared to manual methods while incorporating concerns about certain data sources that an MFA author can faithfully reflect in the constraints, quality values, and objective function of this method.

Whereas the approximate solution time for the methodological example of 18 variables and 15 constraints was approximately 1 second, the time for the objective function to converge for the global steel reanalysis with 878 variables (including residuals) was approximately 10 minutes, with both analyses being run using Matlab version 2014b, run on a standard laptop computer with an Intel® Core^{TM} i5-3360M CPU with 8.00 gigabytes of RAM. The objective function value and the maximum constraint violation value for the global steel reanalysis converged after approximately 600 iterations of the solver, as seen in figure 3. Figure 3 shows that the solution starts off infeasible, as measured by the maximum constraint violation, because the initial values g_0 are, where possible, set to data points that violate conservation of mass and other constraints. The solver then finds a feasible solution where the maximum constraint violation is acceptable and the objective function value is superior to Cullen and colleagues' objective function value.

Discussion

The method proposed here is able to incorporate a wide variety of both linear and nonlinear data constraints into the data reconciliation procedure. Table 5 compares the capacity of this method to incorporate these data types with a number of other commonly used data reconciliation methods in MFA. Any data reconciliation procedure will have advantages and disadvantages. For example, the RAS family of algorithms is

Table 5 Summary of data types that can be handled by data reconciliation methods commonly used in MFA

	Data type											
Technique	*Node*	*TCs*	*Edge width*	*ALRs*	*Inverse TCs*	*Sums of variable subsets*	*Percentage of sums*	*Pro-rata data*	*Sequential multiplications*	*Bounds*	*Probability distributions*	*Data quality*
RAS		X[a]	X[a]									
KRAS	X	X	X	X		X				X		X
Brunner and Rechberger (2004)	X			X								
Matyus and colleagues (2003)	X	X[b]		X								
Fellner and colleagues (2011) (STAN software)		X	X	X[c]					X	X[d]	X[e]	
Cencic and Frühwirth (2014)	X			X		X				X	X	
This method	X	X	X	X	X	X	X	X	X	X		X

[a] Must reconcile based on either TCs or edge widths.
[b] Error for a limited number of TCs only can be reconciled.
[c] ALRs may be defined, but the multiplier is a constant rather than a variable.
[d] Defined as $\pm 2\sigma$, assuming a normal probability distribution.
[e] Only normal probability distributions are allowed.
MFA = material flow analysis; TCs = transfer coefficients; ALRs = additional linear relations.

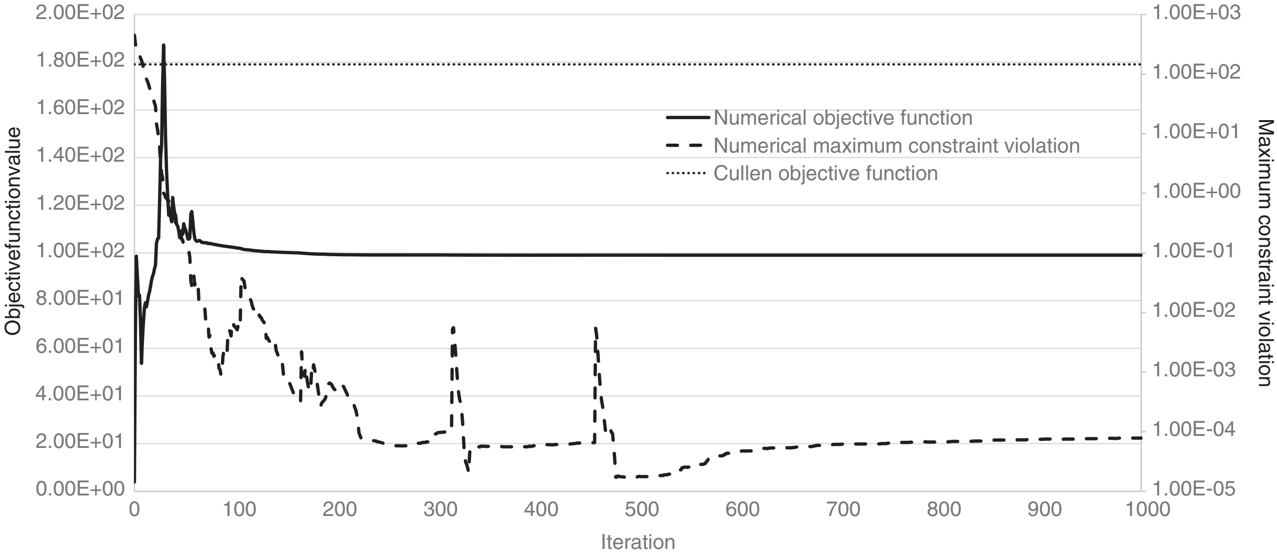

Figure 3 Graph of the objective function value and maximum constraint violation versus the number of iterations of the solver for the global steel reanalysis. The objective function value for Cullen and colleagues' analysis is also shown on the graph.

able to reconcile data from large numbers of variables, and the class of problems to which it is usually applied will normally have conventional node or TC data points only. Similarly, the techniques of Brunner and colleagues and STAN are appropriate for MFAs where most data are available in conventional forms and has normal probability distributions.

In general, the choice of data reconciliation method will depend upon problem structure. The method proposed in this article is likely to be most appropriate in problems of medium size for which (1) a significant proportion of the available data exist in unconventional forms; (2) problems where the number of unknowns is significantly greater than the number of constraint equations; and (3) where the structure of the problem is amenable to an iterative approach of assuming reasonable starting values for variables which lack data. Problems for which the RAS family of algorithms are normally used, where there may be significantly more than 1,000 variables, may be too large for standard nonlinear solvers to solve. Smaller problems that contain data in standard node and TC formats can be reconciled using algorithms that are simpler and less computationally expensive to implement. The method presented here can only implicitly incorporate probability distributions through use of upper and lower bounds at an appropriate multiple of the standard deviation.

Thus, this reconciliation method is generally able to incorporate more data than other quantitative reconciliation methods and is likely superior to, and quicker than, the manual adjustment method used by Cullen and colleagues (2012a). Despite the metrics listed in table 4, it is impossible to prove that this method provides an overall better solution than a manual solution, for two reasons. First, the numerical solver used finds a local, rather than global, minimum for the objective function. The second reason is model error, for example, constraints are improperly formulated or data exist that are not expressed in constraint equation form. This method is dependent on the assumption of reasonable starting values for larger problems in particular, so that there is potential to use other data reconciliation procedures, such as those discussed in the literature review, for providing those starting values, although this has not been investigated in this study. The value of the inclusion of unconventional data forms is apparent because running the reconciliation using this procedure, but without any unconventional data points, results in a solution where 53% of node variables have a value that deviates more than 5% from either Cullen and colleagues' solution or the solution that includes the unconventional data constraints.

The numerical solution might also be improved upon if local numerical linearization methods are used. The usefulness of such methods would be highly dependent on finding good starting values for the optimization. A large literature on relaxation or linearization of the pooling problem exists (e.g., Narasimhan and Jordache 2000; Gounaris et al. 2009; Faria and Bagajewicz 2012). A future article will explore relaxations and linearizations of this particular program.

This analysis has led to four lessons to ensure that future implementations produce solutions that are superior to manual solutions. First, numerical nonlinear constrained optimization solvers cannot guarantee that a solution is a global, rather than local, minimum of the objective function. Thus, particularly for larger problems, it is useful to supply a reasonable guess for the initial value of all variables, even those for which no data are available, increasing the chance that the initial values, and thus the solution, will be in the attractor well that corresponds to the global minimum, or at least to a sensible local solution, of the objective function.

Second, it is useful to run the optimization with only structural and conventional data constraints before running an optimization that includes the alternative data constraint forms. The first optimization solution is useful because it is easy to identify obvious errors in the structural or node or TC data constraints.

Third, it may be beneficial to place common sense bounds on some variables in order to constrain the solution to reasonable values. In the global steel flows analysis, the only known bound on the amount of iron ore extracted per year is that it must not be less than zero ($x_i \geq 0$). However, one might reasonably infer that true value is 1,000 Mt ± 25%. Replacing the original non-negative lower bound with this inference ($750 \geq x_i \geq 1250$) will further constrain the solution and make it more likely that the final value of the objective function is a global, rather than local, minimum. If the intended confidence interval of the analysis is 95%, then the bounds for certain variables could be set at $\pm 2\sigma$.

Fourth, the method described in this article is iterative. The objective function, node, TC, and residual values should be examined to ensure that they are sensible. A relative error or average relative error value that is large may indicate an insufficient number of, or wrongly applied, structural or data constraints. If the solution is not sensible, new structural constraints, data constraints, bounds, and quality values may be added to the optimization program. In particular, new forms of unconventional data constraints may be considered if information exists that is not compatible with the forms listed in table 1.

Conclusions and Future Work

The MFA data reconciliation method described here is (1) able to accept a comprehensive array of nonlinear conventional and unconventional data forms than other reconciliation methods and (2) formulates a nonlinear data reconciliation program without the requirement to assume that some pieces of data are constants while incorporating quantitative data quality measures and multiple data points for the same variable.

The methodology described in this study can be expanded upon in two directions. First, the list of alternative constraints in table 1 can be expanded to account for additional data types. In particular, the availability of remotely sensed data as well as "big data" availability means that new data sets, particularly translations between data at different spatial and temporal scales, could be incorporated into MFA analysis through the addition of new data constraint equations. Second, the bilinear problem here can be linearized to decrease the difficulty in finding a global minimum of the program.

Acknowledgments

The authors thank Dr. Elizabeth Curmi and Dr. Andrew Skelton of the University of Cambridge for their assistance in the preparation of this work. The authors gratefully acknowledge the financial support of BP as part of its Energy Sustainability Challenge for this research.

References

Bringezu, S., H. Schütz, and S. Moll. 2003. Rationale for and interpretation of economy-wide materials flow analysis and derived indicators. *Journal of Industrial Ecology* 7(2): 43–64.

Brunner, P. H. and H. Rechberger. 2004. *Practical handbook of material flow analysis*. Boca Raton, F L, USA: Lewis.

Byrd, R. H., M. E. Hribar, and J. Nocedal. 1999. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization* 9(4): 877–900.

Cencic, O. and R. Frühwirth. 2014. A general framework for data reconciliation—Part I: Linear constraints. *Computers & Chemical Engineering*. http://dx.doi.org/10.1016/j.compchemeng.2014.12.004. Accessed 2 January 2015.

Cencic, O. and H. Rechberger. 2008. Material flow analysis with software STAN. *Environmental Informatics and Industrial Ecology* 2008: 440–447. ISBN 978-3-8322-7313-2.

Cullen, J. M. and J. M. Allwood. 2013. Mapping the global flow of aluminium: From liquid aluminium to fabricated goods: Supplemental information. *Environmental Science & Technology* 47(7): 3057–3064.

Cullen, J. M., J. M. Allwood, and M. D. Bambach. 2012a. Mapping the global flow of steel: From steelmaking to end-use goods. *Environmental Science & Technology* 46(24): 13048–13055.

Cullen, J. M., J. M. Allwood, and M. D. Bambach. 2012b. Mapping the global flow of steel: From steelmaking to end-use goods: Supplemental information. *Environmental Science & Technology* 46(24): 13048–13055.

Faria, C. and M. J. Bagajewicz. 2012. A new approach for global optimization of a class of MINLP problems with applications to water management and pooling problems. *Process Systems Engineering* 58(8): 2320–2335.

Fellner, J., P. Aschenbrenner, O. Cencic, and H. Rechberger. 2011. Determination of the biogenic and fossil organic matter content of refuse-derived fuels based on elementary analyses. *Fuel* 90(11): 3164–3171. http://linkinghub.elsevier.com/retrieve/pii/S0016236111003619. Accessed 22 December 2014.

Fischer-Kowalski, M., F. Krausmann, S. Giljum, S. Lutter, A. Mayer, S. Bringezu, Y. Moriguchi, H. Schütz, H. Schandl, and H. Weisz. 2011. Methodology and indicators of economy-wide material flow accounting. *Journal of Industrial Ecology* 15(6): 855–876. http://doi.wiley.com/10.1111/j.1530-9290.2011.00366.x. Accessed 15 March 2012.

Gounaris, C. E., R. Misener, and C. A. Floudas. 2009. Computational comparison of piecewise—Linear relaxations for pooling problems. *Industrial & Engineering Chemistry Research* 48(12): 5742–5766.

Greenberg, H. J. 1995. Analyzing the pooling problem. *ORSA Journal on Computing* 7(2): 205–217.

Laner, D., H. Rechberger, and T. Astrup. 2014. Systematic evaluation of uncertainty in material flow analysis. *Journal of Industrial Ecology* 18(6): 859–870. http://doi.wiley.com/10.1111/jiec.12143. Accessed 18 December 2014.

Lenzen, M., B. Gallego, and R. Wood. 2009. Matrix balancing under conflicting information. *Economic Systems Research* 21(1): 23–44. www.tandfonline.com/doi/abs/10.1080/09535310802688661. Accessed 24 March 2013.

Lenzen, M., D. D. Moran, A. Geschke, and K. Kanemoto. 2014. A non-sign-preserving RAS variant. *Economic Systems Research* 26(2): 197–208.

Li, X., E. Armagan, T. Asgeir, and P. I. Barton. 2011. Stochastic pooling problem for natural gas production network design and operation under uncertainty. *AIChE Journal* 57(8): 2120–2135.

Matlab. 2013. fmincon. *Matlab R2013b Documentation*. www.mathworks.co.uk/help/optim/ug/fmincon.html. Accessed 24 September 2013.

Matyus, T., A. Gleiss, K. Gruber, and G. Bauer. 2003. Data reconciliation, structure analysis and simulation of waste flows: Case study vienna. *Waste Management & Research* 21(2): 93–109.

Mehanna, O., K. Huang, B. Gopalakrishnan, A. Konar, and N. D. Sidiropoulos. 2015. Feasible point pursuit and successive approximation of non-convex QCQPs. *Signal Processing Letters, IEEE* 22(7): 804–808.

Misener, R. and C. A. Floudas. 2010. Global optimization of large-scale generalized pooling problems: Quadratically constrained MINLP models. *Industrial & Engineering Chemistry Research* 49(11): 5424–5438. http://pubs.acs.org/doi/abs/10.1021/ie100025e. Accessed 3 December 2014.

Narasimhan, S. and C. Jordache. 2000. *Data reconciliation & gross error detection: An intelligent use of process data*. Houston, TX, USA: Gulf.

Ploeg, F. Van Der. 1985. FIML estimation of dynamic econometric systems from inconsistent data. *International Journal of Systems Science* 16(1): 1–29.

Ploeg, F. Van der. 1988. Balancing large systems of national accounts. *Computer Science in Economics and Management* 1(1): 31–39.

Schmidt, M. 2008a. The Sankey diagram in energy and material flow management—Part I: History. *Journal of Industrial Ecology* 12(1): 82–94.

Schmidt, M. 2008b. The Sankey diagram in energy and material flow management—Part II: Methodology and current applications. *Journal of Industrial Ecology* 12(2): 173–185.

World Steel Association. 2010. *Steel statistical yearbook 2009*. Brussels: World Steel Association. www.worldsteel.com/.

About the Authors

Grant M. Kopec is a Ph.D. candidate, **Julian M. Allwood** is a professor, and **Jonathan Cullen** is a lecturer, all at the University of Cambridge Department of Engineering in Cambridge, United Kingdom. **Daniel Ralph** is a professor at the University of Cambridge Judge Business School.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Supporting Information S1: This supporting information provides two figures and one table: Figure S1 is a Sankey diagram of the global steel MFA from Cullen and colleagues (2012a). Figure S2 is a histogram of the individual residual values that comprise the objective function values of both this study and that from Cullen and colleagues (2012a). Finally, table S1 presents a detailed breakdown of the high-value residuals in the analysis.