The reconciliation of multiple conflicting estimates: Entropy-based and axiomatic approaches

MDPI

Article

The reconciliation of multiple conflicting estimates: entropy-based and axiomatic approaches

João F. D. Rodrigues [1]* and Michael L. Lahr [2]

[1] Institute of Environmental Sciences CML, Leiden University, Einsteinweg 2, 2333 CC Leiden, the Netherlands

[2] Edward J. Bloustein School of Planning & Public Policy, Rutgers, The State University of New Jersey, 33 Livingston Avenue, New Brunswick, NJ 08901-1982, USA

* Correspondence: E-mail: j.rodrigues@cml.leidenuniv.nl; Tel.: +31 645 062 746

Academic Editor: name
Version October 18, 2018 submitted to Entropy

Abstract: When working with economic accounts it may occur that multiple estimates of a single datum exist, with different degrees of uncertainty or data quality. This paper addresses the problem of defining a method that can reconcile conflicting estimates, given best guess and uncertainty values. We proceeded from first principles, using two different routes. First, under an entropy-based approach, the data reconciliation problem is addressed as a particular case of a wider data balancing problem, and an alternative setting is found in which the multiple estimates are replaced by a single one. Afterwards, under an axiomatic approach, a set of properties is defined, which characterizes the ideal data reconciliation method. Under both approaches, the conclusion is that the formula for the reconciliation of best guesses is a weighted arithmetic average, with the inverse of uncertainties as weights, and that the formula for the reconciliation of uncertainties is a harmonic average.

Keywords: Uncertainty modelling; economic accounts; conflicting estimates; entropy-based approach; axiomatix approach.

1. Introduction

With improvements in information technology, the world has become more unified and interconnected. Information is now typically shared quickly and easily from all over the globe, such that barriers formed by linguistic and geographic boundaries essentially have been torn down. This has enabled people from disparate cultures and backgrounds to share ideas and information. One outcome of this regime change has been a boosting of the perceived benefits of statistical information. While some benefits of such statistical information have been known since at least Quetelet's (1835) tome on so-called "social physics" was published, today's massive socio-economic statistical repositories in Europe, North America, and East Asia are enabling a data revolution of sorts. Indeed, the fields of data mining and data analytics are fast becoming important fields of academic study. Mirroring the rise of data availability and the nature of some of the data itself, the term "big data" has been coined [1] to refer to the extremely voluminous and complex data sets that require specialized processing application software to deal with them.

The most prominent stewards of socio-economic data are government statistical agencies, which focus on producing and disseminating data products secured via surveys (for example the American Community Survey), censuses (such as Japan's 2015 Population Census), and administrative procedures (like information needed to get an academic promotion in Spain). As a result, data storage is now ubiquitously electronic, replicated offsite to guard against storage failure, and measured in petabytes. Electronic storage enables low-cost dissemination of data. It also facilitates the integration

of records across disparate databases – for example, into a system of national accounts, which is what countries use to generate their estimates of gross domestic product, as suggested by the United Nations. Both lead to concerns about confidentiality of data and how it can be protected [2].

Our point in broaching the above is that data producers, disseminators, and users alike can run into the problem of having access to multiple estimates for a single quantity of interest. In the particular experience of the authors, which motivated the present study, these multiple estimates are a consequence of non-disclosure by a statistical office in order to ensure confidentiality. Hence we act as what Duncan *et al.* [2] called "data snooper". For concreteness, in such instances we are interested in obtaining number of employees at county level from the U.S. Bureau of Statistics's Quarterly Census of Employment and Wages (QCEW) data and U.S. Bureau of the Census's County Business Patterns. In both datasets these figures are suppressed for selected sectors in some counties and even states. But information is provided for larger spatial and sectoral units, so it is possible to use this higher-level information to obtain multiple estimates of the quantities of interest. It is common for official statistical data to have a hierarchical structure so this problem is quite general. Garfinkel *et al.* [3] note that the increasing ability of data snoopers is making ever more data stewards reluctant to provide certain data products because they are finding it increasingly difficult to ensure confidentiality to the agents from whom they obtain the data. This is despite some use of noise as a disclosure limitation [4].

In this paper, we focus on the problem of combining such multiple estimates into a single value. In the case of economic accounts Miller and Blair [5, pp. 384-6] have called it "the reconciliation issue". The reconciliation issue considered here should not be confused with the more general problem of data balancing, in which a set of multiple data points need to satisfy a set of constraints: that problem is addressed in other studies, such as Kruithof [6], Stone *et al.* [7], Byron [8], Van der Ploeg [9], Lahr and Mesnard [10], Chen [11]. General solutions to confidentiality disclosure or data censoring issues are provided by [12] and [13]. Herein we set out to assist current and future data snoopers and miners, by identifying what a data reconciliation method should be from first principles when a fairly general formulation of the reconciliation constraints is possible.

In particular, we consider that the multiple estimates for a particular datum can be characterized by a best guess and uncertainty. If we interpret each estimate as a random variable with an underlying probability distribution, the best guess is the expected value and the uncertainty is standard deviation. In the case of multiple data sources, the conflict enabling the multiple estimates is self-evident. When numbers are published with some data censored and for which estimates can be obtained using partial information [14], the conflict can arise from a higher (or lower) hierarchical spatial or sectoral level (e.g., average employee number if the number of establishments is available). To the best of our knowledge no first-principle approach to this problem has yet been published, although more heuristic approaches can be found in Bourque *et al.* [15], Miernyk *et al.* [16], Jensen and McGaurr [17], Gerking [1819], Weale [20], Boomsma and Oosterhaven [21], Rassier *et al.* [22]. We tackle the same problem from two different angles.

Using concepts and techniques from Bayesian inference [23] and in particular the minimum cross-entropy method [24], we first address the problem of data reconciliation as a particular case of more general data balancing [25]. That is, we consider there are two or more initial estimates for a particular datum, but this datum is itself embedded in a set of constraints connecting it to other data that are potentially unbalanced. We look for simplifications of the general setting under which this original problem can be transformed into another balancing problem where the multiple estimates are replaced by a single one. We prove that, if the initial uncertainty estimates are close to one another, the data reconciliation method of best guesses is a weighted arithmetic average and the data reconciliation method of uncertainties is a harmonic average.

Afterwards we address the same problem from an axiomatic perspective, laying out the desirable properties of a data reconciliation method. Such an approach has roots in different fields, from table deflation [26] and supply-use transformations [27] to environmental responsibility [28]. It turns out that the canonical data reconciliation method, i.e., the one that satisfies all required properties, is none

other than a suitable generalization of the entropy-based method derived earlier. That generalization centers on the introduction of the number of previously combined priors and a ranking of estimates by their relative quality.

2. Entropy-based approach

2.1. Basic concepts

Bayesian inference was first developed by Laplace [29] and later expanded by others, such as Jeffreys [30], Jaynes [31] and Jaynes [23]. According to the Bayesian paradigm, a probability is a degree of belief about the likelihood of an event, and should reflect all relevant available information about that event. According to Weise and Woger [32], if an empirical quantity is subject to measurement errors, it must be described by a random variable, whose expectation is the best guess and whose standard-deviation is the uncertainty estimate.

More formally, a *prior* datum θ_i is characterized by a probability distribution $\pi(q_i)$, which expresses the degree of belief that the datum takes realization q_i. The *best guess* is $\mu_i = \mathrm{E}[\theta_i]$ and the *uncertainty* is $\sigma_i = \sqrt{\mathrm{Var}[\theta_i]}$. When multiple data are considered, e.g., θ_i and θ_j, it is necessary to introduce the *correlation* between them, $\rho_{ij} = \mathrm{Cov}[\theta_i, \theta_j]/\sigma_i\sigma_j$. Rodrigues [33] further provides a series of rules to determine the properties of a strictly positive prior datum, using the maximum-entropy principle [34].

The type of data we are interested in are connected to one another through *accounting identities* of the form:

$$\theta_0 = \sum_{i=1}^{n} \theta_i,$$

where θ_0 is an *aggregate* datum and the θ_i's are *disaggregate* data. If the set of data is arranged in a vector $\boldsymbol{\theta}$ of length n_T, the set of n_K accounting identities can be defined through a concordance matrix $\mathbf{G}$, where, for a given accounting identity i, $G_{ij} = 1$ if θ_j is a disaggregate datum, $G_{ij} = -1$ if θ_j is an aggregate datum and $G_{ij} = 0$ otherwise.

If the prior configuration is unbalanced, then $\mathbf{G}\boldsymbol{\theta} \neq \mathbf{0}$, where $\mathbf{0}$ is a vector of zeros. Rodrigues [25] derives an analytical solution and a series of approximations that, given a concordance matrix and prior configuration, provide a *posterior* configuration, $\mathbf{t}$, such that $\mathbf{G}\mathbf{t} = \mathbf{0}$. The notational convention used here is that Greek letters refer to priors while Latin cognates will refer to posteriors, i.e., m_i, s_i and r_{ij} are, respectively, the best guess and uncertainty of t_i and correlation between t_i and t_j.

2.2. Problem formulation

We are now in position to formulate the data reconciliation problem. Given *initial priors* θ' and θ'' and a system with $n_T + 1$ numerical data $\{\theta_1, \ldots, \theta_{n_T-1}, \theta', \theta''\}$, and $n_K + 1$ accounting identities, where accounting identity $n_K + 1$ takes the form $\theta' = \theta''$, our goal is to determine the *final prior* θ, in a new system with n_T numerical data, $\{\theta_1, \ldots, \theta_{n_T-1}, \theta\}$ and the n_K first accounting identities of the original system, in which the posteriors $\{t_1, \ldots, t_{n_T-1}\}$ are identical in both data balancing problems, and $t = t' = t''$. Conceptually, we are approaching data reconciliation as a form of preliminary data balancing, as illustrated in Figure 1. The conflicting estimates are initial priors of the same datum, and the reconciled value is a final prior. Note the following notational convention: while other variables (and their properties) are denoted with subscripts, initial priors/posteriors (and their properties) are denoted with one ($'$) or two ($''$) primes, and the final prior/posterior is denoted with neither subscripts nor primes.

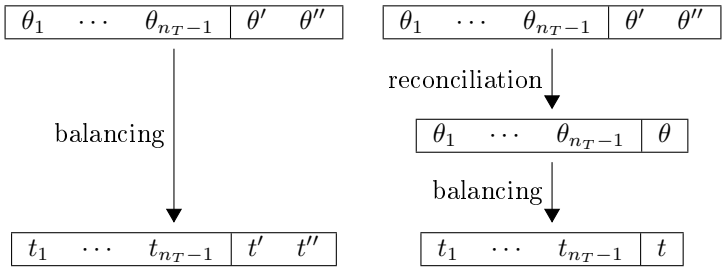

Figure 1. On the left-hand side balancing in a single step, with multiple initial estimates (priors) of the same datum, θ' and θ'', balanced to the same quantity (posterior), $t' = t''$. On the right-hand side balancing in two steps: first the reconciliation procedures combines the multiple initial estimates (initial priors), θ' and θ'', into a final prior, θ; afterwards the full system is balanced, leading to posterior t. We impose that the result from both procedures is the same, $t = t' = t''$.

Three situations emerge: either the datum to be reconciled is only a disaggregate datum; it is only an aggregate datum; or it is both a disaggregate and an aggregate datum, in different accounting identities. We will deal with the three cases separately.

We now present simple systems to illustrate the three possible cases. As a benchmark consider a tabular system (i.e., with data organized in rows and columns) with no multiple estimates consisting of a 2×3 table $\mathbf{A}$ with row sums $\mathbf{b}$ and columns sums $\mathbf{c}$. Furthermore, consider that the sum of both $\mathbf{b}$ and $\mathbf{c}$ is known as d. If $\mathbf{i}$ is a vector of ones of appropriate length, all vectors are in column format by default, and prime ($'$) adjoined to a matrix or vector denotes transpose, then the previous set of constraints means that:

$$\begin{aligned} \mathbf{A}\mathbf{i} &= \mathbf{b}; \\ \mathbf{A}'\mathbf{i} &= \mathbf{c}; \\ \mathbf{b}'\mathbf{i} &= d; \\ \mathbf{c}'\mathbf{i} &= d. \end{aligned}$$

The vectorized form of this system and the concordance table is presented in Table 1. In the baseline system there is a total of twelve variables (columns of the concordance matrix $\mathbf{G}$) and seven constraints (rows thereof). The first six variables are disaggregate values (corresponding to the initial $\mathbf{A}$ matrix), the following five are mixed (row and column sums $\mathbf{b}$ and $\mathbf{c}$), and the last one is an aggregate datum (d). The first two constraints (rows of $\mathbf{G}$) are the row sums of $\mathbf{A}$, the following three are its columns sums, and the last two are the sums of $\mathbf{b}$ and $\mathbf{c}$. To understand how $\mathbf{G}$ is constructed let us consider the first constraint, which is the row sum of $\mathbf{A}$. Formally, this is:

$$A_{11} + A_{12} + A_{13} - b_1 = 0,$$

hence in the first row of $\mathbf{G}$ the entries corresponding to the columns of A_{11}, A_{12} and A_{13} have 1s, the entry corresponding to the column of b_1 has -1 and all entries are zero.

We are now in position to formalize the three situations of multiple estimates of a single datum as variants of Table 1 in which an additional row and column has been added to $\mathbf{G}$.

The case of disaggregate datum occurs if the datum for which multiple estimates exist is an interior point, which for concreteness we consider to be element A_{23}: the set of constraints is shown in Table 2. As an illustration of the case of there being two estimates of an aggregate datum consider it to

Table 1. Prior vector and concordance matrix, with no multiple estimates.

$\boldsymbol{\theta}'$	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A_{23}	b_1	b_2	c_1	c_2	c_3	d
G	1	1	1	0	0	0	-1	0	0	0	0	0
	0	0	0	1	1	1	0	-1	0	0	0	0
	1	0	0	1	0	0	0	0	-1	0	0	0
	0	1	0	0	1	0	0	0	0	-1	0	0
	0	0	1	0	0	1	0	0	0	0	-1	0
	0	0	0	0	0	0	1	1	0	0	0	-1
	0	0	0	0	0	0	0	0	1	1	1	-1

be d: the set of constraints is shown in Table 3. Finally, consider as example of an element that is both aggregate and disaggregate that of b_1: the set of constraints is shown in Table 4.

Table 2. Prior vector and concordance matrix, with multiple estimates of A_{23}.

$\boldsymbol{\theta}'$	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A'_{23}	b_1	b_2	c_1	c_2	c_3	d	A''_{23}
G	1	1	1	0	0	0	-1	0	0	0	0	0	0
	0	0	0	1	1	1	0	-1	0	0	0	0	0
	1	0	0	1	0	0	0	0	-1	0	0	0	0
	0	1	0	0	1	0	0	0	0	-1	0	0	0
	0	0	1	0	0	0	0	0	0	0	-1	0	1
	0	0	0	0	0	0	1	1	0	0	0	-1	0
	0	0	0	0	0	0	0	0	1	1	1	-1	0
	0	0	0	0	0	1	0	0	0	0	0	0	-1

Table 3. Prior vector and concordance matrix, with multiple estimates of d.

$\boldsymbol{\theta}'$	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A_{23}	b_1	b_2	c_1	c_2	c_3	d'	d''
G	1	1	1	0	0	0	-1	0	0	0	0	0	0
	0	0	0	1	1	1	0	-1	0	0	0	0	0
	1	0	0	1	0	0	0	0	-1	0	0	0	0
	0	1	0	0	1	0	0	0	0	-1	0	0	0
	0	0	1	0	0	1	0	0	0	0	-1	0	0
	0	0	0	0	0	0	1	1	0	0	0	-1	0
	0	0	0	0	0	0	0	0	1	1	1	0	-1
	0	0	0	0	0	0	0	0	0	0	0	1	-1

Table 4. Prior vector and concordance matrix, with multiple estimates of b_1.

$\boldsymbol{\theta}'$	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A_{23}	b'_1	b_2	c_1	c_2	c_3	d	b''_1
G	1	1	1	0	0	0	-1	0	0	0	0	0	0
	0	0	0	1	1	1	0	-1	0	0	0	0	0
	1	0	0	1	0	0	0	0	-1	0	0	0	0
	0	1	0	0	1	0	0	0	0	-1	0	0	0
	0	0	1	0	0	1	0	0	0	0	-1	0	0
	0	0	0	0	0	0	0	1	0	0	0	-1	1
	0	0	0	0	0	0	0	0	1	1	1	-1	0
	0	0	0	0	0	0	1	0	0	0	0	0	-1

It is perhaps instructive to describe how the reconciliation problems differ from the features of the baseline system. The three variants of the baseline are constructed by adding a single variable, the conflicting estimate, which by convenience is always appended to the original system. It is also necessary to add an extra constraint, connecting the two conflicting estimates. Finally, the baseline system is also changed so that in one of the original occurrences of the datum to be reconciled is the first conflicting estimate and the second occurrence is the other conflicting estimate.

Note that in this simple example there are only two constraints affecting each datum, but that naturally is not generally the case. The number of constraints per datum is arbitrary and can be either one or larger than two. An example of what this system might represent is employment count by region and sector, with an extra dimension being type of ownership (private or local, state, or federal government), as reported in the QCEW database.

2.3. From balancing to reconciliation

Rodrigues [25] shows that if the posterior configuration is balanced, then its first- and second-moment constraints are:

$$\mathbf{0} = \mathbf{G}\mathbf{m}; \tag{1}$$

$$\mathbf{0} = \operatorname{diag}\left(\mathbf{G}\mathbf{S}|\mathbf{G}|'\right), \tag{2}$$

where $\mathbf{m}$ and $\mathbf{S}$ are the posterior best-guess vector and covariance matrix, and the latter is defined as $\mathbf{S} = \hat{\mathbf{s}}\mathbf{R}\hat{\mathbf{s}}$, where $\mathbf{s}$ is the vector of posterior uncertainties and $\mathbf{R}$ is the vector of posterior correlations, and $\hat{}$ denotes diagonal matrix. Likewise $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the prior best guess vector and covariance matrix, and the latter is defined as $\boldsymbol{\Sigma} = \hat{\sigma}\mathbf{P}\hat{\sigma}$.

The analytical solution of the data-balancing problem is:

$$\tilde{\mathbf{S}}^{-1} = \tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{G}'\hat{\boldsymbol{\beta}}|\mathbf{G}|; \tag{3}$$

$$\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{m}} = \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} + \mathbf{G}'\boldsymbol{\alpha}. \tag{4}$$

Notice that Equations 3-4 contain symbols adjoined with $\tilde{}$ (which we refer to as Gaussian parameters) while Equations 1-2 do not. The connection between the Gaussian parameters and the corresponding observable quantities is described in Rodrigues [25]: when relative uncertainty, σ_j/μ_j or s_j/m_j, is low, then the Gaussian parameter and the observable are identical. When relative uncertainty is high, the best guess Gaussian parameter tends to $-\infty$ and the uncertainty Gaussian parameter tends to ∞, in such a way that if relative uncertainty is unitary, $-\tilde{\mu}_j/\tilde{\sigma}_j^2 = 1/\mu_j = 1/\sigma_j$ and $-\tilde{m}_j/\tilde{s}_j^2 = 1/m_j = 1/s_j$. There is no closed-form expression between observables and Gaussian parameters in the multivariate case.

If both the prior uncertainty of aggregate data and initial prior correlations are high, we obtain a simplified weighted least-squares (WLS) method in which the weights are prior uncertainties:

$$\mathbf{m} = \boldsymbol{\mu} + \hat{\sigma}\mathbf{G}'\boldsymbol{\alpha}, \tag{5}$$

and posterior correlations are set by considering that relative uncertainty is constant, $\mathbf{s} = \mathbf{m} \odot \boldsymbol{\sigma} \oslash \boldsymbol{\mu}$, where $\odot$ and $\oslash$ are Hadamard (or entrywise) product and division, and the update takes place in small steps.

This WLS method is a generalization of the standard biproportional balancing method (RAS) for arbitrary structure and uncertainty data [25]. However, it is in a way too simple for the data reconciliation problem, because it keeps relative uncertainty constant. In the data reconciliation problem this assumption is untenable, whenever the relative uncertainty of the initial priors differs.

Thus, we now look for a simplification of the general solution (Equations 3-4) that is still feasible and that allows both for best guess and uncertainty reconciliation. Let us consider that correlations change little from prior to posterior, so that only uncertainties are adjusted. Equations 3-4 become:

$$\hat{\mathbf{s}}^{-1}\mathbf{R}^{-1}\mathbf{s}^{-1} = \hat{\sigma}^{-1}\mathbf{P}^{-1}\sigma^{-1} + \mathbf{G}'\boldsymbol{\beta};$$
$$\hat{\mathbf{s}}^{-1}\mathbf{R}^{-1}\hat{\mathbf{s}}^{-1}\mathbf{m} = \hat{\sigma}^{-1}\mathbf{P}^{-1}\hat{\sigma}^{-1}\boldsymbol{\mu} + \mathbf{G}'\boldsymbol{\alpha},$$

where we dropped the $\tilde{}$, meaning that all variables are observables. If correlations are not adjusted, then $\mathbf{R} = \mathbf{P}$, and if variances change little $\mathbf{s} \simeq \sigma$ The previous expressions become:

$$\mathbf{s}^{-1} = \sigma^{-1} + \mathbf{P}\hat{\sigma}\mathbf{G}'\boldsymbol{\beta};$$
$$\hat{\mathbf{s}}^{-1}\mathbf{m} = \hat{\sigma}^{-1}\boldsymbol{\mu} + \mathbf{P}\hat{\sigma}\mathbf{G}'\boldsymbol{\alpha}.$$

For convenience, consider now that a datum corresponding to entry (i, j) in the tabular matrix is t_{ij}, while the sums of row or column i is t_i, and the Lagrange parameters of a row sum or column sum are adjoined with superscript R or C. For a particular entry, the previous matrix equation reads:

$$\frac{1}{s_{ij}} = \frac{1}{\sigma_{ij}} + \sigma_i^R\beta_i^R + \sigma_j^C\beta_j^C + \sum_{k\neq i,j}\left(\sigma_{ik}\beta_k^C + \sigma_{kj}\beta_k^R\right);$$
$$\frac{1}{s_i} = \frac{1}{\sigma_i} + \sigma_i\left(\beta_i^R + \beta_i^C\right),$$

where $\sigma_i^R = \sum_j \sigma_{ij}$ and $\sigma_i^C = \sum_j \sigma_{ji}$. If the adjustment from prior to posterior is small, then $\sigma_i^R \simeq \sigma_i^C \simeq \sigma_i$. If $\beta_i^* = \sigma_i\beta_i^R$, $\sigma_i \gg \sigma_{ij}$ and $\sigma_i \gg \sigma_{ji}$, then the previous expression matrix expressions simplify to:

$$\mathbf{s}^{-1} = \sigma^{-1} + \mathbf{G}'\boldsymbol{\beta}^*; \tag{6}$$
$$\hat{\mathbf{s}}^{-1}\mathbf{m} = \hat{\sigma}^{-1}\boldsymbol{\mu} + \mathbf{G}'\boldsymbol{\alpha}^*, \tag{7}$$

where the derivation of Equation 7 follows along identical lines to that Equation 6. We now use these expressions to obtain a tentative solution of the data reconciliation problem, even though they were derived under rather strict assumptions.

2.4. A tentative solution

We now examine the implications of applying Equations 6-7 to different to the different data reconciliation configurations described in Section 2.2: multiple estimates of (a) an aggregate datum; (b) a disaggregate datum; and (c) a datum that is both aggregate and disaggregate. We shall see that the same expression applies to all these problems.

For clarity, the analysis is carried out using scalar expressions, and, for brevity, only to the case of two constraints per datum. The strategy of the proof is the same for all configurations: to derive constraints connecting prior and posterior in the original problem and in a modified problem in which there is only a single datum where originally there were the conflicting estimates.

2.4.1. Aggregate datum

Consider that there are two initial priors of a datum, θ_0' and θ_0'' and that the datum is involved in two accounting identities, the first summing over elements 1 to n' and the second summing over $n'+1$ to n'':

$$
\begin{aligned}
t'_0 &= \sum_{i=1}^{n'} t_i; \\
t''_0 &= \sum_{i=n'+1}^{n'+n''} t_i; \\
t'_0 &= t''_0,
\end{aligned}
$$

where each t_i, for $i > 0$, can be affected by other accounting identities. The Lagrange parameters associated with these three expressions in Equation 6 are denoted, respectively, by β'_0, β''_0 and β_0. We wish to determine a final prior θ_0, such that:

$$
\begin{aligned}
t_0 &= \sum_{i=1}^{n'} t_i; \\
t_0 &= \sum_{i=n'+1}^{n'+n''} t_i.
\end{aligned}
$$

Equation 6 reads, for the original problem:

$$
\begin{aligned}
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta'_0 + \ldots \quad \text{if} \quad 1 \leq i \leq n'; \\
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta''_0 + \ldots \quad \text{if} \quad n'+1 \leq i \leq n''; \\
\frac{1}{s'_0} &= \frac{1}{\sigma'_0} - \beta'_0 + \beta_0; \\
\frac{1}{s''_0} &= \frac{1}{\sigma''_0} - \beta''_0 - \beta_0,
\end{aligned}
$$

where ... refers to other Lagrange parameters. And in the modified problem:

$$
\begin{aligned}
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta'_0 + \ldots \quad \text{if} \quad 1 \leq i \leq n'; \\
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta''_0 + \ldots \quad \text{if} \quad n'+1 \leq i \leq n''; \\
\frac{1}{s_0} &= \frac{1}{\sigma_0} - \beta'_0 - \beta''_0.
\end{aligned}
$$

Notice that for every datum $i > 0$ the original and modified problem are identical. Because the posteriors of the aggregate datum are all identical, $s'_0 = s''_0 = s_0$, we can write:

$$
\begin{aligned}
2\frac{1}{s_0} &= 2\left(\frac{1}{\sigma_0} - \beta'_0 - \beta''_0\right) \\
&= \frac{1}{\sigma'_0} + \frac{1}{\sigma''_0} - \beta'_0 - \beta''_0 + \beta_0 - \beta_0.
\end{aligned}
$$

A similar expression can be obtained from Equation 7 for the final prior best guess, leading to the solution:

$$\frac{1}{\sigma_0} = \frac{1}{2}\left(\frac{1}{\sigma'_0} + \frac{1}{\sigma''_0}\right);$$
$$\frac{\mu_0}{\sigma_0} = \frac{1}{2}\left(\frac{\mu'_0}{\sigma'_0} + \frac{\mu''_0}{\sigma''_0}\right).$$

Thus, both the final prior of the absolute uncertainty, σ, and the relative uncertainty, σ/μ, are obtained as the harmonic average of the initial prior absolute and relative uncertainties.

2.4.2. Disaggregate datum

Consider now that there are two initial priors of an interior point, θ'_1 and θ''_1, which is affected by two accounting identities, such that the posteriors satisfy:

$$t'_0 = t'_1 + \sum_{i=2}^{n'} t_i;$$
$$t''_0 = t''_1 + \sum_{i=n'+1}^{n'+n''} t_i;$$
$$t'_1 = t''_1.$$

The Lagrange parameters associated with these three expressions are, as before, β'_0, β''_0 and β_0. We wish to determine a final prior θ_1, such that:

$$t'_0 = t_1 + \sum_{i=2}^{n'} t_i;$$
$$t''_0 = t_1 + \sum_{i=n'+1}^{n'+n''} t_i.$$

Equation 6 reads, for the original problem:

$$\frac{1}{s_i} = \frac{1}{\sigma_i} + \beta'_0 + \ldots \quad \text{if} \quad 2 \leq i \leq n';$$
$$\frac{1}{s_i} = \frac{1}{\sigma_i} + \beta''_0 + \ldots \quad \text{if} \quad n'+1 \leq i \leq n'';$$
$$\frac{1}{s'_0} = \frac{1}{\sigma'_0} - \beta'_0 + \ldots;$$
$$\frac{1}{s''_0} = \frac{1}{\sigma''_0} - \beta''_0 + \ldots;$$
$$\frac{1}{s'_1} = \frac{1}{\sigma'_1} + \beta'_0 + \beta_1;$$
$$\frac{1}{s''_1} = \frac{1}{\sigma''_1} + \beta''_0 - \beta_1,$$

and in the modified problem:

$$\begin{aligned}
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta'_0 + \ldots \quad \text{if} \quad 2 \leq i \leq n';\\
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta''_0 + \ldots \quad \text{if} \quad n'+1 \leq i \leq n'';\\
\frac{1}{s'_0} &= \frac{1}{\sigma'_0} - \beta'_0 + \ldots;\\
\frac{1}{s''_0} &= \frac{1}{\sigma''_0} - \beta''_0 + \ldots;\\
\frac{1}{s_1} &= \frac{1}{\sigma_1} + \beta'_0 + \beta''_0.
\end{aligned}$$

As before, the data for which there are no conflicting estimates (t'_0, t''_0 and t_i with $i > 1$) are subject to the same set of constraints in the original and in the modified problem. Because the posteriors of the disaggregate datum are all identical, $s'_1 = s''_1 = s_1$, we can write:

$$\begin{aligned}
2\frac{1}{s_1} &= 2\left(\frac{1}{\sigma_1} + \beta'_0 + \beta''_0\right)\\
&= \frac{1}{\sigma'_1} + \frac{1}{\sigma''_1} + \beta'_0 + \beta''_0 + \beta_1 - \beta_1.
\end{aligned}$$

At this stage it becomes clear that we will encounter exactly the same solution as in the case of an aggregate datum:

$$\begin{aligned}
\frac{1}{\sigma_1} &= \frac{1}{2}\left(\frac{1}{\sigma'_1} + \frac{1}{\sigma''_1}\right);\\
\frac{\mu_1}{\sigma_1} &= \frac{1}{2}\left(\frac{\mu'_1}{\sigma'_1} + \frac{\mu''_1}{\sigma''_1}\right).
\end{aligned}$$

2.4.3. Mixed datum

Consider now that there are two initial priors, θ'_1 and θ''_1, of a datum that is both aggregate and disaggregate, in different accounting identities, and whose posteriors satisfy:

$$\begin{aligned}
t_0 &= t'_1 + \sum_{i=2}^{n'} t_i;\\
t''_1 &= \sum_{i=n'+1}^{n''} t_i;\\
t'_1 &= t''_1.
\end{aligned}$$

As before the Lagrange parameters are denoted as β'_0, β''_0 and β_1. We wish to determine a final prior θ_1, such that:

$$\begin{aligned}
t_0 &= t_1 + \sum_{i=2}^{n'} t_i;\\
t_1 &= \sum_{i=n'+1}^{n''} t_i.
\end{aligned}$$

Equation 6 reads, for the original problem:

$$\begin{aligned}
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta'_0 + \dots \quad \text{if} \quad 2 \leq i \leq n';\\
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta''_0 + \dots \quad \text{if} \quad n'+1 \leq i \leq n'';\\
\frac{1}{s_0} &= \frac{1}{\sigma_0} - \beta'_0 + \dots;\\
\frac{1}{s'_1} &= \frac{1}{\sigma'_1} + \beta'_0 + \beta_1;\\
\frac{1}{s''_1} &= \frac{1}{\sigma''_1} - \beta''_0 - \beta_1,
\end{aligned}$$

and in the modified problem:

$$\begin{aligned}
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta'_0 + \dots \quad \text{if} \quad 2 \leq i \leq n';\\
\frac{1}{s_i} &= \frac{1}{\sigma_i} + \beta''_0 + \dots \quad \text{if} \quad n'+1 \leq i \leq n'';\\
\frac{1}{s_0} &= \frac{1}{\sigma_0} - \beta'_0 + \dots;\\
\frac{1}{s_1} &= \frac{1}{\sigma_1} + \beta'_0 - \beta''_0.
\end{aligned}$$

As has become routine, for datum 0 and for every datum $i > 1$ the original and modified problem are identical. Because $s'_1 = s''_1 = s_1$, we can write:

$$\begin{aligned}
2\frac{1}{s_1} &= 2\left(\frac{1}{\sigma_1} + \beta'_0 - \beta''_0\right)\\
&= \frac{1}{\sigma'_1} + \frac{1}{\sigma''_1} + \beta'_0 - \beta''_0 + \beta_1 - \beta_1.
\end{aligned}$$

Thus, it is clear that the solution is again identical.

3. Axiomatic approach

3.1. Axiomatic formulation

In Section 2 we obtained a data reconciliation algorithm from first principles, as an operation of data balancing under a particular structure. However, we can also reason about the data reconciliation algorithm in terms of its properties, i.e., we will not determine what it is, but what it ought to be.

If θ' and θ'' are two initial priors, the data reconciliation algorithm is a function $f(\cdot)$ that generates a final prior $\theta = f(\theta', \theta'')$, where each prior θ is characterized by a best guess, μ, an absolute uncertainty, σ, and a relative uncertainty, $u = \sigma/\mu$, which can take values in the range $0 \leq u \leq 1$. Let $x_{\min} = \min\{x', x''\}$ and $x_{\max} = \max\{x', x''\}$, where x can be μ, σ or u.

We now propose a series of properties that define the data reconciliation method.

Property 1 (Lower and upper bounds)**.** *The parameters of the final prior lie within the range set by the parameters of the initial priors,* $\mu_{\min} \leq \mu \leq \mu_{\max}$, $\sigma_{\min} \leq \sigma \leq \sigma_{\max}$ *and* $u_{\min} \leq u \leq u_{\max}$.

Property 2 (Commutativity). *The order in which the initial priors are combined does not matter,* $f(\theta', \theta'') = f(\theta'', \theta')$.

Property 3 (Associativity). *Several initial priors can be combined and the resulting final prior is invariant to the order of reconciliation,* $f(\theta', f(\theta'', \theta''')) = f(f(\theta', \theta''), \theta''')$.

Property 4 (Identity). *If the initial prior best guesses are identical,* $\mu' = \mu''$ *then the final prior best guess is identical,* $\mu = \mu' = \mu''$. *If the initial prior uncertainties are identical,* $\sigma' = \sigma''$ *then the final prior uncertainty is identical,* $\sigma = \sigma' = \sigma''$.

Property 5 (Monotonicity). *The relative adjustment from initial to final prior increases with the relative magnitude of initial uncertainty:*

$$\frac{\mu - \mu'}{\mu'' - \mu} = g\left(\frac{\sigma'}{\sigma''}\right); \tag{8}$$

$$\frac{\sigma - \sigma'}{\sigma'' - \sigma} = h\left(\frac{\sigma'}{\sigma''}\right). \tag{9}$$

where $dg(x)/dx > 0$ *and* $dh(x)/dx > 0$.

Property 6 (Absorption). *If initial prior* θ' *is known with minimal uncertainty,* $u' = 0$, *and* θ'' *is not,* $u'' > 0$, *then the final prior is identical to the first initial prior,* $f(\theta', \theta'') = \theta'$. *If initial prior* θ' *is known with maximal uncertainty,* $u' = 1$, *and* θ'' *is not,* $u'' < 1$, *then the final prior is identical to the second initial prior,* $f(\theta', \theta'') = \theta''$.

We believe that these six properties are uncontroversial and self-explanatory. However, it turns out that the problem as formulated here has no solution, i.e., no formula can satisfy all of the above properties. We later overcome this hurdle by generalizing the problem formulation, to include two additional concepts: a hierarchy of data quality and the number of combined priors.

3.2. The canonical data reconciliation method

The properties outlined in Section 3.1 constrain the range of data reconciliation algorithms but do not define a unique solution. However, Equations 8-9 suggests how it may be possible to obtain a solution. Let us consider that $g(x)$ and $h(x)$ take the simple yet flexible form of $g(x) = ax^b$ and $h(x) = cx^d$.

The condition of identity (Property 4), in the case of $\mu' = \mu''$ and $\sigma' = \sigma''$ leads to the indeterminacy:

$$\frac{\mu - \mu'}{\mu'' - \mu} = \frac{0}{0}.$$

But if the limit is approached as $\mu' = \mu - \delta$ and $\mu'' = \mu + \delta$, when $\delta \to 0$, then:

$$\frac{\mu - \mu'}{\mu'' - \mu} = \frac{\delta}{\delta} = 1.$$

Thus, under the condition of identity, Equations 8-9 imply that:

$$1 = a1^b;$$
$$1 = c1^d,$$

so $a = c = 1$. Let us further consider the simplest possible case $b = d = 1$, so that $g(\cdot)$ and $h(\cdot)$ are the identity lines. Applying $g(x) = x$ and $h(x) = x$ to Equations 8-9 leads to:

$$\frac{\mu - \mu'}{\sigma'} = \frac{\mu'' - \mu}{\sigma''};$$
$$\frac{\sigma - \sigma'}{\sigma'} = \frac{\sigma'' - \sigma}{\sigma''}.$$

Rearranging terms:

$$\mu\left(\frac{1}{\sigma'} + \frac{1}{\sigma''}\right) = \frac{\mu'}{\sigma'} + \frac{\mu''}{\sigma''};$$
$$\sigma\left(\frac{1}{\sigma'} + \frac{1}{\sigma''}\right) = \frac{\sigma'}{\sigma'} + \frac{\sigma''}{\sigma''}.$$

Recalling that $u = \sigma/\mu$ we obtain the canonical data reconciliation method as:

$$\mu = \left(\frac{1/\sigma'}{1/\sigma' + 1/\sigma''}\right)\mu' + \left(\frac{1/\sigma''}{1/\sigma' + 1/\sigma''}\right)\mu''; \tag{10}$$
$$\frac{1}{\sigma} = \frac{1}{2}\left(\frac{1}{\sigma'} + \frac{1}{\sigma''}\right). \tag{11}$$

Equation 10 can be be expressed in two other ways:

$$\mu = \frac{\sigma}{2}\left(\frac{\mu'}{\sigma'} + \frac{\mu''}{\sigma''}\right); \tag{12}$$
$$\frac{1}{u} = \frac{1}{2}\left(\frac{1}{u'} + \frac{1}{u''}\right). \tag{13}$$

Thus, if the ratio of relative adjustment of best guesses and uncertainties is identical to the ratio of absolute uncertainties of the initial priors, the best-guess data reconciliation method is a weighted average, where the weights are proportional to the inverse of absolute uncertainty, and the absolute and relative uncertainty data reconciliation methods are harmonic averages.

Does this data reconciliation method satisfy the properties of Section 3.1? It is trivial to check that Properties 1, 2, 4 and 5 are satisfied. But this is not the case for Properties 3 and 6. In the following subsections we present suitable extensions of the canonical data reconciliation method to address these problems.

3.3. The number of combined priors

The canonical data reconciliation method is not associative. The properties of $f(\theta', f(\theta'', \theta'''))$ are:

$$\frac{1}{u} = \frac{1}{2}\left(\frac{1}{u'} + \frac{1}{2}\left(\frac{1}{u''} + \frac{1}{u'''}\right)\right) = \frac{1}{2u'} + \frac{1}{4u''} + \frac{1}{4u'''}$$
$$\frac{1}{\sigma} = \frac{1}{2}\left(\frac{1}{\sigma'} + \frac{1}{2}\left(\frac{1}{\sigma''} + \frac{1}{\sigma'''}\right)\right) = \frac{1}{2\sigma'} + \frac{1}{4\sigma''} + \frac{1}{4\sigma'''}.$$

While the properties of $f((\theta', \theta''), \theta''')$ are:

$$\frac{1}{u} = \frac{1}{2}\left(\frac{1}{2}\left(\frac{1}{u'} + \frac{1}{u''}\right) + \frac{1}{u'''}\right) = \frac{1}{4u'} + \frac{1}{4u''} + \frac{1}{2u'''}$$
$$\frac{1}{\sigma} = \frac{1}{2}\left(\frac{1}{2}\left(\frac{1}{\sigma'} + \frac{1}{\sigma''}\right) + \frac{1}{\sigma'''}\right) = \frac{1}{4\sigma'} + \frac{1}{4\sigma''} + \frac{1}{2\sigma'''}.$$

Thus, $f(\theta', f(\theta'', \theta''')) \neq f((\theta', \theta''), \theta''')$. But upon some reflection, this result is in fact reasonable. The final prior is the combination of two initial priors with equal weights. If some of these initial priors is itself a combination of other initial priors, this information has to be considered explicitly.

Let us introduce a new quantity, n, as the *number of combined priors*, so that now a prior θ is defined by a best guess, μ, an absolute uncertainty, σ, and n. Consider the following data reconciliation rule:

$$\mu = \left(\frac{n'/\sigma'}{n'/\sigma' + n''/\sigma''}\right)\mu' + \left(\frac{n''/\sigma''}{n'/\sigma' + n''/\sigma''}\right)\mu''; \tag{14}$$
$$\frac{1}{\sigma} = \frac{1}{n}\left(\frac{n'}{\sigma'} + \frac{n''}{\sigma''}\right); \tag{15}$$
$$n = n' + n''. \tag{16}$$

As before, Equation 14 can be be expressed in two other ways:

$$\mu = \frac{\sigma}{n}\left(\frac{n'}{\sigma'}\mu' + \frac{n''}{\sigma''}\mu''\right); \tag{17}$$
$$\frac{1}{u} = \frac{1}{n}\left(\frac{n'}{u'} + \frac{n''}{u''}\right). \tag{18}$$

This data reconcilation rule satisfies the first five properties of Section 3.1.

3.4. Ranking of data quality

The canonical data reconciliation method satisfies the absorption property of minimal uncertainty. If $\sigma' = 0$ and $\sigma'' > 0$, then:

$$\frac{1}{\sigma'} + \frac{1}{\sigma''} \simeq \frac{1}{\sigma'},$$

and Equations 10-11 become:

$$\mu \simeq \left(\frac{1/\sigma'}{1/\sigma'}\right)\mu' + \left(\frac{1/\sigma''}{1/\sigma'}\right)\mu'' = (1)\mu' + (0)\mu'';$$
$$\sigma \simeq 2\sigma' = 0,$$

so $\mu = \mu'$ and $\sigma = \sigma'$. However, it does not satisfy the absorption property of maximal uncertainty. If $\sigma' = \mu'$ and $\sigma'' < \mu''$, then $u' = 0$ and Equations 10, 11 and 13 become:

$$\mu = \left(\mu'' + \sigma''\right) \frac{\mu'}{\mu' + \sigma''};$$
$$\sigma = 2\sigma'' \frac{\mu'}{\mu' + \sigma''};$$
$$u = 2\frac{u''}{1 + u''},$$

and thus $\mu \neq \mu''$ and $\sigma \neq \sigma''$.

In order to ensure that the absorption of maximal uncertainty is satisfied, we use the concept of *data quality*, introduced in Rodrigues [25]. The idea is that, besides an uncertainty estimate, which formalizes quantitatively a degree of confidence in the accuracy of the best guess of a datum, it is also possible to formalize qualitatively a degree of confidence in the accuracy of a datum relative to others.

For the purpose of data balancing, Rodrigues [25] suggests that a datum that is considered to be of higher quality should be kept fixed while lower quality data are adjusted. The natural corollary, in the problem of data reconciliation, is to consider that when one wishes to combine two initial priors of differing levels of data quality, the prior of lower quality should be disregarded.

If a datum has unitary relative uncertainty, then it is maximally uninformative, and it is reasonable to disregard it. After all, a maximally uninformative prior should only be used if no better alternative is available. We therefore suggest that, if $\sigma' = \mu'$ and $\sigma'' < \mu''$, then $\theta = \theta''$ directly, without using Equations 10-11.

3.5. Summary

We now present the expressions for the combination of n initial priors, θ_i, with $i = 1, \ldots, n$ into a single final prior θ. Addressing this problem requires the specification, for each prior, θ_i, of its best guess, μ_i, its absolute uncertainty, σ_i, and the number of previously combined priors, n_i.

If all relative uncertainties, $u_i = \sigma_i/\mu_i$, are in the range $0 < u_i < 1$, then the final prior properties are defined as:

$$\mu = \sum_{i=1}^{n} \left(\frac{n_i/\sigma_i}{n/\sigma}\right) \mu_i; \tag{19}$$
$$\frac{1}{\sigma} = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{n_i}{\sigma_i}\right); \tag{20}$$
$$n = \sum_{i=1}^{n} n_i. \tag{21}$$

Equation 19 can be expressed as:

$$\frac{1}{u} = \frac{1}{n}\sum_{i=1}^{n} \frac{n_i}{u_i}. \tag{22}$$

If some initial priors have zero relative uncertainty, $u_i = 0$, then all other initial priors should be disregarded. If some initial priors have unitary relative uncertainty, $u_i = 1$, then it is they which should be disregarded.

In Figure 2 we illustrate the behaviour of Equation 19, when $n = 2$ and $n_1 = n_2 = \mu_2 = 1$. The plot shows different curves of the combined posterior best guess μ as a function of the prior best guess

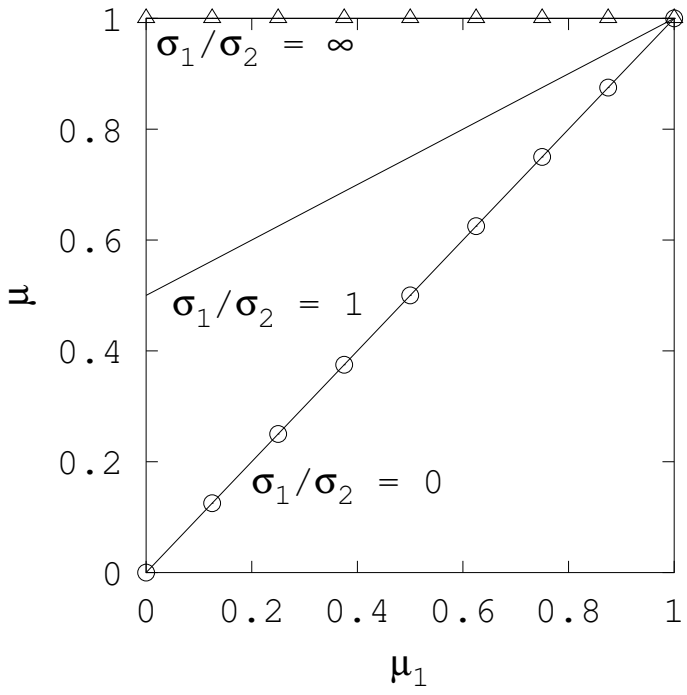

Figure 2. Best guess of the combined prior as a function of initial prior best guess, μ_1, when $\mu_2 = 1$, for diferent values of σ_1/σ_2: Solid line (–) when $\sigma_1/\sigma_2 = 1$, solid line with circle markers (–○–) when $\sigma_1/\sigma_2 = 0$ and solid line with triangle markers (–△–) when $\sigma_1/\sigma_2 = \infty$.

μ_1, where each curve corresponds to a different ratio of uncertainties, σ_1/σ_2. When both uncertainties are identical, the posteriod best guess is the arithmetic average of the two prior best guesses. When the uncertainties differ, the best guess prior with the largest uncertainty contributes the least to combined best guess: in the limit case in which $\sigma_1 \gg \sigma_2$ the prior μ_1 is ignored and the posterior is similar to μ_2; when $\sigma_1 \ll \sigma_2$ the reverse occurs and $\mu \simeq \mu_1$.

In turn, Figure 3 we illustrate the behaviour of Equation 20. We still consider that $n = 2$ and $n_1 = n_2$ but now no explicit assumption about best guesses is necessary. Instead, the uncertainty of the second variable is fixed, $\sigma_2 = 1$, and the curve shows the value of the combined posterior as a function of prior uncertainty σ_1. The figure describes an arc slightly above the diagonal line. When both prior uncertainties are identical ($\sigma_1 = 1$), then the posterior equals the priors, as expected. As σ_1 becomes smaller than σ_2 the combined prior becomes closer to σ_1 than to σ_2, but always larger, $\sigma > \sigma_1$, except in the limit case $\sigma_1 \to 0$, in which case $\sigma \to \sigma_1$.

4. Conclusions and discussion

Herein we investigated using two distinct pathways the problem of reconciling multiple conflicting estimates in the course of database development. We assume that the developer (data snooper) is tooled with a best guess and uncertainty for each of those conflicting estimates.

First, we apply a maximum-entropy Bayesian inference method, under the limiting condition that the adjustment from prior to posterior uncertainties is small. Second, we obtain a canonical data reconciliation method through an axiomatic approach that is as simple as possible but satisfied important qualitative properties. Each approach verifies the other.

The resulting formula for the best guess, Equation 19, is a weighted average showing that, as the count of conflicting priors underlying a particular prior rises, the value of that prior increases in importance in terms of obtaining a solution. We get a similar result with the inverse of the uncertainty,

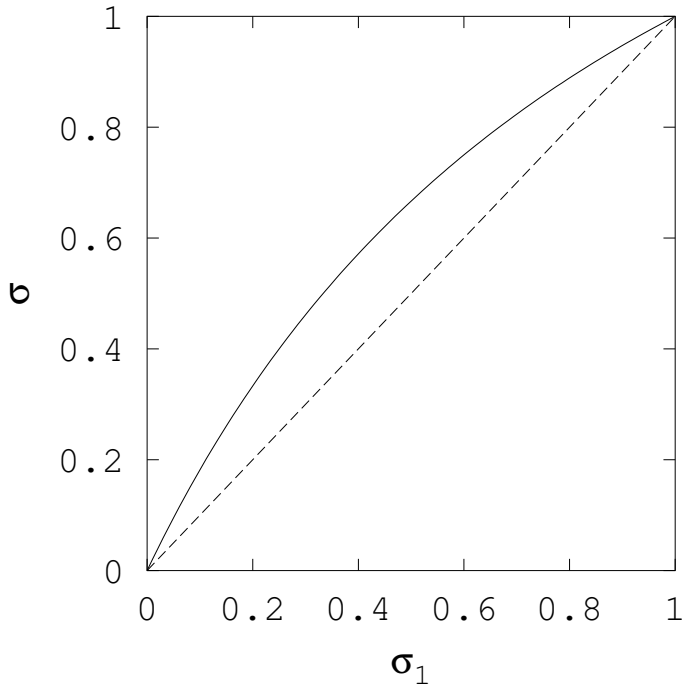

Figure 3. Solid line (–) and relative uncertainty, uncertainty of the final prior, σ, as a function of initial prior uncertainty, σ_1, when $\sigma_2 = 1$. Dashed line (- -) is the identity line.

that is, the narrower the uncertainty of an estimate the more it contributes to the final solution. The resulting formula for the uncertainty, Equation 20, is a harmonic average where the same factors are present: as the count of conflicting priors underlying a particular prior rises, the value of that prior increases in importance; and the narrower the uncertainty of a prior, the more it contributes to the final solution.

Of course, limitations to our approach must be mentioned. And the key one is certainly that, in some practical applications, the data snooper will lack information on either or both best guess and uncertainty. It may be that instead, one only has upper and lower bounds for the datum of interest to inform its best guess and uncertainty. This is certainly the case in some instances when data are censored, e.g., the anti-suppression problem of Gerking *et al.* [13] and Isserman and Westervelt [12]. Future work using variable ranges with externally informed priors would be a natural extension of what is presented here. Indeed, some initial forays into this line of investigation are already underway, see, e.g., Makarkina and Lahr [35].

It should be mentioned that although the focus of attention here was on conflicting estimates arising from economic accounts there are other circumstances in which a formally identical problem arises, for example in expert elicitation [36].

Author Contributions: J.R. performed the conceptualization and formal analysis, M.L. provided the motivation and both J.R. and M.L wrote the manuscript.

Funding: Funding to be specified when accepted.

Acknowledgments: Any errors the paper may contain are the sole responsibility of the authors.

Conflicts of Interest: The authors declare no conflict of interest.

1. Lohr, S. The Origins of 'Big Data': An Etymological Detective Story. *New York Times* **2013**. Available online in January 2018 at https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.
2. Duncan, G.T.; Elliot, M.; Salazar-Gonzalez, J.J. *Statistical Confidentiality: Principles and Practice*; Springer: New York, 2011.
3. Garfinkel, R.; Gopal, R.; Goes, P. Privacy Protection of Binary Confidential Data against Deterministic, Stochastic, and Insider Attack. *Management Science* **2002**, *48*, 749–764.
4. Evans, T.; Zayatz, L.; Slanta, J. Using Noise for Disclosure Limitation of Establishment Tabular Data. *Journal of Official Statistics* **1998**, *14*, 537–551.
5. Miller, R.E.; Blair, P.D. *Input-Output Analysis: Foundations and Extensions*; Cambridge University Press: Cambridge, UK, 2009. Second Edition.
6. Kruithof, R. Telefoonverkeersrekening. *De Ingenieur* **1937**, *52*, E15–E25.
7. Stone, R.; Meade, J.E.; Champernowne, D.G. The Precision of National Income Estimates. *The Review of Economic Studies* **1942**, *9*, 111–125.
8. Byron, R. The Estimation of Large Social Account Matrices. *Journal of the Royal Statistical Society, Series A* **1978**, *141*, 359–367.
9. Van der Ploeg, F. Reliability and the Adjustment of Sequences of Large Economic Accounting Matrices. *Journal of the Royal Statistical Society, Series A* **1982**, *145*, 169–194.
10. Lahr, M.L.; Mesnard, L.d. Biproportional Techniques in Input-Output Analysis: Table Updating and Structural Analysis. *Economic Systems Research* **2004**, *16 (2)*, 115–134.
11. Chen, B. A Balanced System of U.S. Industry Accounts and Distribution of the Aggregate Statistical Discrepancy by Industry. *Journal of Business and Economic Statistics* **2012**, *30*, 202–211.
12. Isserman, A.; Westervelt, J. 1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data. *International Regional Science Review* **2006**, *29*, 311–335.
13. Gerking, S.; Isserman, A.; Hamilton, W.; Pickton, T.; Smirnov, O.; Sorenson, D. Anti-suppressants and the Creation and Use of Non-Survey Regional Input-Output Models. Regional Science Perspectives in Economic Analysis: A Festschrift in Memory of Benjamin H. Stevens; Lahr, M.; Miller, R., Eds.; Elsevier: New York, 2001; pp. 379–406.
14. Rodrigues, J.; Marques, A.; Wood, R.; Tukker, A. A network approach for assembling and linking input-output models. *Economic Systems Research* **2016**, *28*, 518–538.
15. Bourque, P.J.; Chambers, E.J.; Chiu, J.S.Y.; Denman, F.L.; Dowdle, B.; Gordon, G.; Thomas, M.; Tiebout, C.; Weeks, E.E. *The Washington Economy: An Input-Output Study*; University of Washington: Seattle, 1967.
16. Miernyk, W.H.; Shellhammer, K.L.; Brown, D.M.; Coccari, R.L.; Gallagher, C.J.; Wineman, W.H. *Simulating Regional Economic Development: An Interindustry Analysis of the West Virginia Economy*; D.C. Heath and Co.: Lexington, 1970.
17. Jensen, R.C.; McGaurr, D. Reconciliation of purchases and sales estimates in an Input-Output table. *Urban Studies* **1976**, *13*, 59–65.
18. Gerking, S. Reconciling 'rows only' and 'columns only' coefficients in an Input-Output model. *International Regional Science Review* **1976**, *1*, 623–626.
19. Gerking, S. Reconciling reconciliation procedures in regional Input-Output Analysis. *International Regional Science Review* **1979**, *4*, 23–36.
20. Weale, M. The reconciliation of values, volumes and prices in national accounts. *Journal of the Royal Statistical Society. Series A* **1988**, *151 (1)*, 211–221.
21. Boomsma, P.; Oosterhaven, J. A double-entry method for the construction of bi-regional Input-Output tables. *Journal of Regional Science* **1992**, *32*, 269–284.
22. Rassier, D.; Howells, T.; Morgan, E.; Empey, N.; Roesch, C. Implementing a reconciliation and balancing model in the U.S. industry accounts. Working Paper WP2007-4, Bureau of Economic Analysis, 2007.
23. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
24. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Annals of Mathematical Statistics* **1951**, *22 (1)*, 79–86.

25. Rodrigues, J.F.D. A Bayesian Approach to the Balancing of Statistical Economic Data. *Entropy* **2014**, *16 (3)*, 1243–1271.
26. Persons, W.M. Fisher's Formula for Index Numbers. *The Review of Economics and Statistics* **1921**, *3*, 103–113.
27. Kop Jansen, P.; ten Raa, T. The Choice of Model in the Construction of Input-Output Coefficients Matrices. *International Economic Review* **1990**, *31*, 213–227.
28. Rodrigues, J.; Domingos, T.; Giljum, S.; Schneider, F. Designing an indicator of environmental responsibility. *Ecological Economics* **2006**, *59*, 256–266.
29. Laplace, P.S. *Essai Philosophique sur les Probabilités*; Courcier Imprimeur: Paris, France, 1814.
30. Jeffreys, H. *Theory of Probability*; Clarendon Press: Oxford, UK, 1939.
31. Jaynes, E.T. *Papers on Probability, Statistics and Statistical Physics*; Kluwer Academic Publishers: Dordrecht, Netherlands, 1983.
32. Weise, K.; Woger, W. A Bayesian theory of measurement uncertainty. *Measurement Science and Technology* **1992**, *4 (1)*, 1–11.
33. Rodrigues, J.F.D. Maximum-Entropy Prior Uncertainty and Correlation of Statistical Economic Data. *Journal of Business & Economic Statistics* **2016**, *34*, 357–367.
34. Jaynes, E.T. Information Theory and Statistical Mechanics I. *Physical Review* **1957**, *106*, 620–630.
35. Makarkina, G.V.; Lahr, M.L. Estimating Nationwide Impacts using an Input-Output Model with Fuzzy Parameters. 25th International Input-Output conference. Atlantic City, NJ, 2017. https://rucore.libraries.rutgers.edu/rutgers-lib/54916/.
36. A., E.; Mikhailov, L.; Keane, J. Inconsistency reduction in decision making via multi-objective optimisation. *European Journal of Operational Research* **2018**, *267*, 212–226.