THIS MANUSCRIPT IS UNDER REVIEW IN THE JOURNAL OF INDUSTRIAL ECOLOGY ID 24-JIE-8438.R2

When Correlation Matters: A Practical Guide to Dealing with Uncertainty in the Case of Data Disaggregation

Simon Schulte[1,2], Arthur Jakobs[3], Rick Lupton[4]

[1] Life Cycle Sustainability, Department of Sustainability and Planning, Aalborg University, 9000 Aalborg, Denmark.
[2] Industrial Ecology Freiburg, University of Freiburg, Freiburg, Germany.
[3] Center for Energy and Environmental Sciences & Center for Nuclear Engineering and Sciences, Laboratory for Energy Systems Analysis, Technology Assessment Group, Paul Scherrer Institut, Villigen PSI, Switzerland.
[4] Centre for Sustainable Energy Systems, Institute of Sustainability and Climate Change, University of Bath, Bath, UK.
Correspondence: Simon Schulte, simonsc@plan.aau.dk, Aalborg University, 9000 Aalborg, Denmark.

Article Type: Methods Article

Conflict of Interest Statement: The authors declare no conflict of interest.

Abstract

Correctly modelling the relationships between correlated, uncertain input data is crucial for producing accurate uncertainty estimates of model results. This requires both an uncertainty analysis that accounts for correlations and the appropriate communication of the results, so that other analysts can correctly interpret the reported uncertainties. However, neither is common practice in industrial ecology modelling. A typical case for correlated results is the disaggregation of a total value into uncertain shares, for which we present a practical yet robust approach to model the uncertainty. Our approach is based on two standard and two generalised Dirichlet distributions, and it uses the maximum entropy principle to choose minimally biased distribution parameters in the absence of specific known values. We discuss how correlation should be communicated to preserve accurate uncertainty information and provide examples to quantify the difference it makes to the results

when the correlation is simplified or completely neglected. The proposed procedure will improve the accuracy of uncertainty quantification in Material Flow Analysis (e.g. where allocation coefficients split flows to sectors), Input Output Analysis (e.g. where aggregated environmental impact data has to be disaggregated to detailed economic sectors), and some instances in Life Cycle Assessment (e.g. where market shares are uncertain). Last but not least, to lower the technical barrier to applying these approaches, we provide easy-to-use Python and R packages which automate the approach.
KEYWORDS: industrial ecology, Dirichlet, maximum entropy, MFA, IO, LCA

1 Introduction

Having solid knowledge of model uncertainty is paramount for both robust decision-making (Morgan et al., 1990; Reale et al., 2017) and for efficiently "prioritising data collection efforts" (Groen and Heijungs, 2017). However, Industrial Ecology (IE) research frequently lacks quantitative uncertainty estimates (Laner et al., 2014; Groen and Heijungs, 2017; Zhang et al., 2019). Accurate uncertainty quantification is a particular challenge in the common case that IE model results are formed from the aggregated sums of many individual data points: for example, the contribution of different sectors' footprint multipliers in Input-Output Analysis (IOA); the total material stock in a Material Stock and Flow Analysis (MSFA) being composed of stocks in different sectors and products; or the overall environmental impacts from contributions of multiple processes in a Life Cycle Assessment (LCA).

What these situations have in common is that the uncertainty in the aggregated total depends strongly on how the uncertainty in the individual elements is correlated. Negative correlation tends to cancel out individual variations when aggregated, while positive correlations tend to exaggerate them (Groen and Heijungs, 2017; Heijungs et al., 2019; Solazzo et al., 2021). Ignoring correlations has direct implications on decision-making, for example when decision makers have to decide between two correlated uncertain variables (such as comparing the impact of renewable e-fuels vs. hydrogen to fuel ships), as well as implications on the results of models which use these correlated variables as inputs (Figure 1).

Figure 2 illustrates this, showing two disaggregated random variables A and B in two extreme cases, one with a strong positive correlation (top row) and one with a strong negative correlation (bottom row). Ignoring correlations could lead to misleading conclusions about the likelihood of one option being preferable to another: Whereas in the top row the marginal distributions of A and B look almost identical, in the bottom case the marginal distributions of A and B (and the sample means and percentiles) are more distinct, suggesting that B is expected to be larger. In fact, due to the strong positive correlations (top row) the probability that B is larger A is actually 100% (that is, even though both A and B vary, $B > A$ for all N draws), while the negative correlations (bottom row) lead to a probability that B is larger than A of only 64%. When correlated variables are used as inputs in another model, the uncertainty of the results of this model might be biased. In the simple case of summing two correlated random variables, the uncertainty of that aggregate is overestimated in case of negative correlations and underestimated in case of positive correlations (see Figure 2, right column).

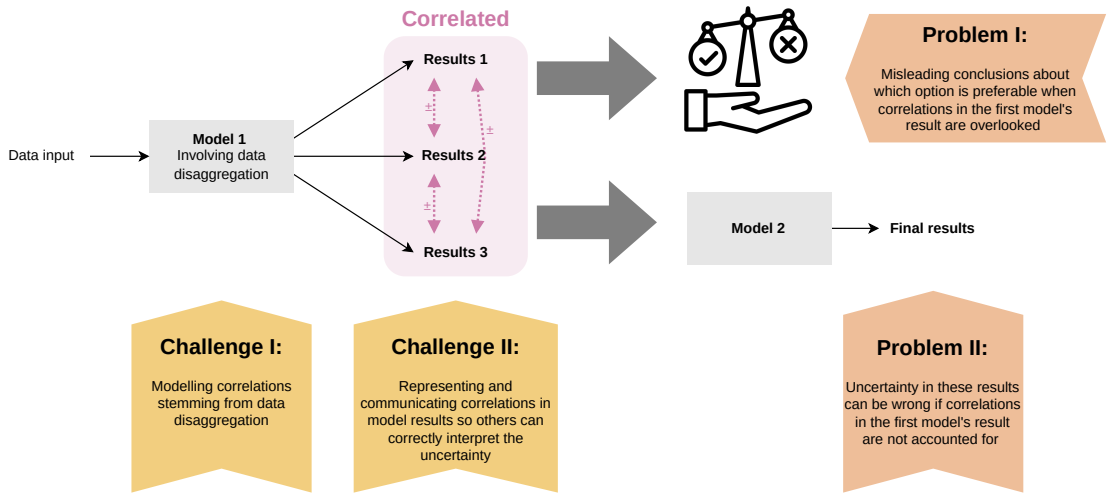

Figure 1: A schematic representation of the two challenges this paper addresses to overcome two common problems that might arise when ignoring correlations.

In general, the direction and size of the error will depend on the details of the problem, but as an illustrative real-world example, an analysis of disaggregation uncertainty in an IOA of the carbon footprint of Germany (in the Supplementary Information, Section D) shows that neglecting correlations can cause the level of uncertainty (measured as standard deviation) in individual sector's carbon multiplier to deviate by -34 to +130%, and the uncertainty in the national footprint to be overestimated by 46%.

As illustrated in Figure 1 there are two challenges to overcome to avoid making these errors in the uncertainty quantification of IE models:

1. Modelers must keep track of correlations introduced into model outputs due to the model structure and assumptions; and

2. Modelers must communicate correlations in model outputs, and make use of information about any correlations in model input data.

For the first problem, while in general there could be many reasons for correlations to be present in model results, we focus on a very common case for IE models: when researchers have to deal with the challenge of disaggregating single data points because the data is unavailable at the required resolution. This usually involves splitting one data point into several disaggregates using additional proxy or auxiliary data (both terms are used interchangeably in this article) or assumptions, and is a challenge common in all three of the model families dominant in IE research (Figure 3). While a few IE studies (reviewed in more detail in Section 2) use Dirichlet or other probability distributions and Monte Carlo sampling to track the correlations and uncertainty introduced during disaggregation (Lupton and Allwood, 2018; Paoli et al., 2018; Helbig et al., 2022; Charpentier Poncelet et al., 2022), it is not widespread in IE practice. Moreover, the basic methods applied in the literature do not handle various technical cases that are important in real-world problems, such as when some subset of the disaggregated shares is more uncertain than others, or when information about some shares and their uncertainty is missing altogether.

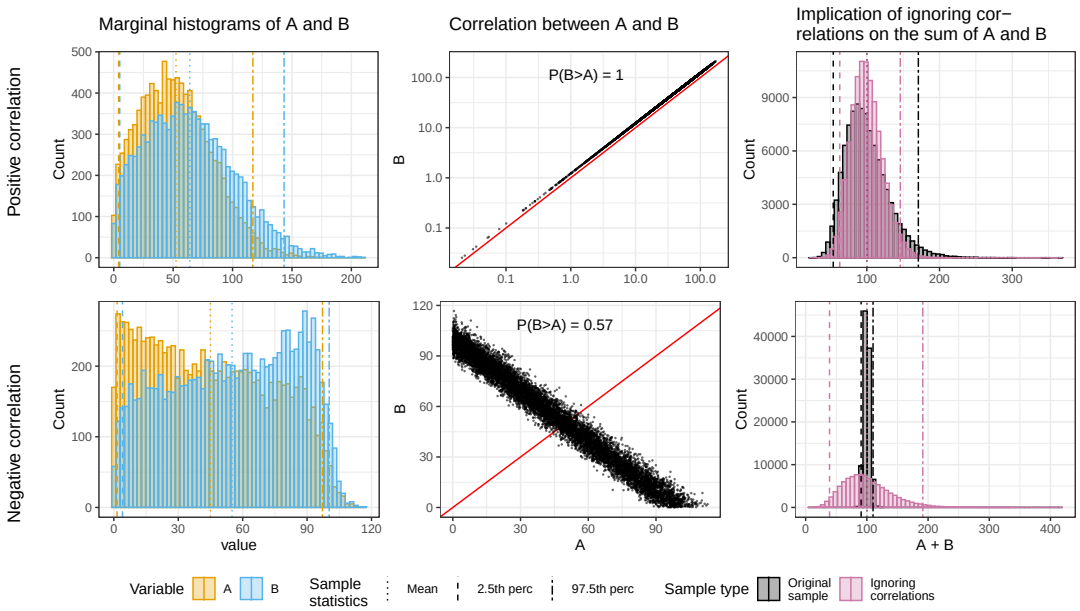

Figure 2: A illustrative example showing two random variables A and B that are strongly positively (top row) and negatively (bottom row) correlated. Both A and B are the result of disaggregating a common aggregate, with different assumptions about the uncertainty in the shares and the initial aggregated total. Left column shows the marginal histograms of A and B and middle column the correlation between them. Even though in the top case the means and 2.5th/97th percentiles between A and B are very similar (45 vs. 55 for the means) due to the strong positive correlation, for all $N = 10000$ draws B is larger than A (i.e. $P(B > A) = 1$). In contrast, even though in the bottom case the means of A and B are the same as above, the negative correlations lead to a much lower probability that B is larger than A. The right column illustrate *Problem 1* from Figure 1(a): The error introduced in the results of Model 2 when the correlations of the results of Model 1 are ignored. Ignoring negative correlations leads to overestimating uncertainties (bottom row) while ignoring positive correlations tend to underestimate uncertainties (top row). To check the code to reproduce the figures and the data behind it, please see the "Data and code availability" statement.

We therefore present in this paper a review of current practices and limitations for dealing with uncertainty due to disaggregation (Section 2), and introduce the theory and a practical guide that IE modelers can use to track correlation in model outputs, in the context of MFA, LCA or IOA (Section 3).

For the second problem – communicating information about correlations in model outputs – we are not aware of any specific guidance on this for IE modelers. We therefore discuss the ways that correlation in uncertain model results may be communicated, with different trade-offs between complexity, storage space and accuracy, and the effect on subsequent analysis if correlations are not accurately preserved (Section 4).

Finally, we discuss the relevance of our presented approach for IE researchers and show its limitations and avenues for further research and refinements (Section 5).

2 Background and literature review

2.1 Data disaggregation under uncertainty

Consider the examples shown in Figure 3. All have in common that we have an aggregate flow y_0, which is known, such as the total amount of steel manufactured in a given time and geography. What we do not know, but are interested in, are the K disaggregate flows $y_1, ..., y_K$ (also called components), such as the different end-use sectors where the manufactured steel ends up. Even though we do not know the values of $y_1, ..., y_K$, our model structures in IE demand that the individual components y_i's need to sum to the known aggregate flow y_0 to respect the mass-, energy-, stoichiometric-, or economic balance of the model:

$$y_0 = \sum_{i=1}^{K} y_i \tag{1}$$

Equation 1 is also called an *accounting identity*.

To get estimates of the disaggregate flows, one usually uses proxy data to calculate shares (ratios/fractions) of the respective disaggregate units $x_1, ..., x_K$. In order to allocate the entire aggregate flow without leaving any residual (thus to respect the system balance), those shares need to sum to one:

$$\sum_{i=1}^{K} x_i = 1 \tag{2}$$

Disaggregate flows are calculated as

$$y_i = x_i y_0, \forall i \in \{1, ..., \mathrm{K}\}. \tag{3}$$

In the MFA example above, proxy data might be monetary steel purchases by the car manufacturing and construction sectors. The LCA modeler might take the energy content or economic value of the three functional flows as proxies for allocation, or production volumes to split aggregated measurements of energy use across processes (we discuss later the difference between uncertainty in normative choices such as allocation, and

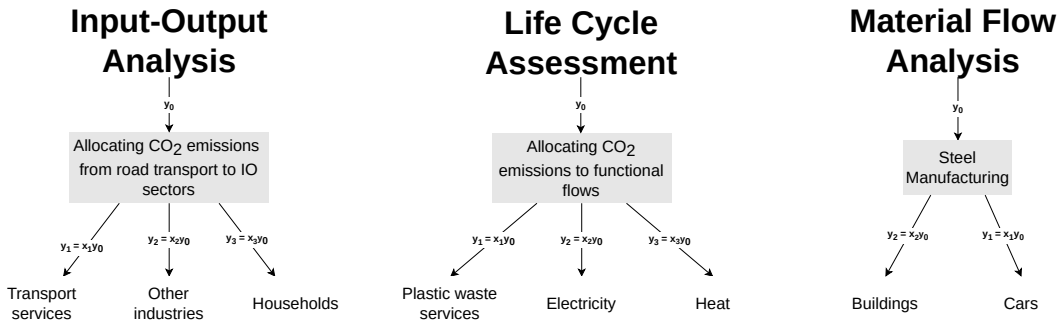

Figure 3: (Simplified) examples of data disaggregation from the three model families dominant in Industrial Ecology research. For example, an MFA modeler interested in the flow of steel through the economy must know the fractions of steel that end in cars and buildings, respectively, for a given time and location (Streeck et al., 2023). An LCA modeler assessing the environmental impacts of a plastic incineration plant for combined heat and electricity production needs to allocate CO_2-emissions to different functional flows, i.e., to the two products leaving the plant (electricity and heat) and the waste product entering the plant (plastic waste) (Guinée et al., 2021), or when measured total electricity consumed in a factory must be shared between the modeled sub-processes. Meanwhile, compilers of environmentally-extended IO databases must allocate CO_2-emissions from road transport to the IO sectors (Schulte et al., 2024).

disaggregation of empirical data). Meanwhile, the IO modeler might use the economic size of the different sectors.

When modelers aim to understand the robustness or uncertainty of their models beyond just average values (or point estimates), they have to use uncertainty propagation, assessing how uncertainty propagates from model inputs to model outputs.[1] Two options for propagating uncertainty exist: analytical or simulation-/sampling-based methods. Analytical uncertainty propagation usually involves using calculus, "applying a local derivative of the mathematical function that specifies how inputs are transformed into outputs" (Heijungs and Lenzen, 2014). This approach requires a good understanding of the processes involved in the model and becomes inaccurate for large uncertainties and/or non-linear models.

Thus, in IE, researchers often apply simulation or sampling-based methods. The most commonly used method is Monte-Carlo (MC) sampling. MC sampling methods propagate uncertainty from model inputs to outputs by repeatedly and randomly sampling from the probability distribution of the model inputs and calculating the model results for each iteration. This results in a sample of model results from which several summary statistics such as mean, standard deviation, quantiles, or the covariance matrix can be computed. This method is more flexible than the analytical method but at the expense of additional computational efforts. Because the complexity of models in IE requires this flexibility, we base this paper on the sampling methods, although we note the problem of estimating the uncertainty of disaggregates can also be approached analytically (Jung et al., 2014) or based on numerical approximations (Lenzen and Murray, 2010; Min and

[1]Note that in this article, we only deal with "parameter uncertainty." For other types of uncertainty common in IE, please refer to Huijbregts (1998); Laner et al. (2014)

Rao, 2018; Rodrigues, 2016). We provide a literature review and a comparison between these different approaches in the Supporting Information.

Uncertainty propagation is straightforward when applied to data *aggregation* especially when sampling-based methods are used: the components are simply added together for each sample to produce the samples of the aggregate. In the case of data *disaggregation*, however, accurately dealing with uncertainty becomes significantly more complex. In particular, the specification of the probability distributions to draw from becomes more challenging. This complexity arises primarily from the accounting identity (Equation 1) introducing inherent correlations between the disaggregates. Ignoring those correlations, in turn, can lead to under- or overestimating uncertainty (Groen and Heijungs, 2017; Heijungs et al., 2019; Solazzo et al., 2021). The accounting identity constraint naturally introduces a negative correlation between the x_i's and either positive or negative correlations between the disaggregates y_i's depending on the size of the uncertainties of the aggregates and the shares (Rodrigues, 2016).

2.2 Usage of the Dirichlet distribution to sample shares

Although data disaggregation is common in IE research, only few studies assess the uncertainty of this disaggregation. Of those that do, most sample the shares from the Dirichlet distribution (e.g. Meyer et al., 2017; Paoli et al., 2018; Lupton and Allwood, 2018; Helbig et al., 2022; Charpentier Poncelet et al., 2022; Kim et al., 2025). A few studies used different distributions to sample shares, e.g. Bornhöft et al. (2016) who used a normalized uniform distribution, whose shortcomings have already been discussed in Lupton and Allwood (2018). The Dirichlet distribution, in turn, is noted for its practical suitability in sampling shares due to its property of samples summing to one (Igos et al., 2019). Moreover, samples generated from the Dirichlet distribution are naturally negatively correlated. It is important to note that those correlations are solely determined by the properties of the data. Hence, the correlations we discuss in the following solely refer to the statistical correlations needed to ensure results are consistent with the accounting identities that govern them. Whenever modelers have prior estimates of correlations (e.g. from expert knowledge), "these alternative priors should be used for as long as they are properly justified and mutually consistent" (Rodrigues, 2016).

While few in number, the IE studies using the Dirichlet distribution highlight the versatility of its use cases. Helbig et al. (2022) apply the Dirichlet distribution to sample allocation parameters of their dynamic MFA model to trace the fate of seven metals at a global level. In a similar context, Charpentier Poncelet et al. (2022) use the Dirichlet distribution to estimate the uncertainty from their model on estimating losses and lifetimes of metals in the economy. In their uncertainty analysis on useful energy balances for the UK, Paoli et al. (2018) use the Dirichlet distribution to randomly allocate energy flows to energy end-uses. In their Bayesian MFA study Lupton and Allwood (2018) define prior distributions for transfer coefficients (i.e. "the fractions of a process's output that flows to different destinations") using Dirichlet distributions. Meyer et al. (2017) study different techniques to include noise damage in LCA. The authors use the Dirichlet distribution to sample all model parameters representing a percentage in their uncertainty analysis. Recently, Kim et al. (2025) use the Dirichlet distribution to sample relative shares of activities in Life Cycle Inventory data that must sum to a whole – such as different power

sources in national electricity mixes or inputs in implicit market datasets. Santos et al. (2022) use the Dirichlet distribution in their stochastic IO analysis to sample the columns of the technology matrix $\boldsymbol{A}$.

However, the way the Dirichlet distribution has been used to sample variables with a "sum-to-one-constraint" so far in IE research has two major limitations, as outlined in the following.

2.2.1 Choosing parameters for the Dirichlet distribution

The first limitation relates to how the parameters of the Dirichlet distribution are chosen. The Dirichlet distribution is parameterized by a vector of positive-valued parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_\mathrm{K})$ such that $\sum_{i=1}^{K} \alpha_i = 1$, and a positive-valued concentration parameter $\gamma > 0$.[2]

Most existing studies choose the concentration parameter γ of the Dirichlet distribution in ways that either ignore available uncertainty information or distort it. Yet, since γ controls how spread out the disaggregated shares can be (see Figure 4), the choice of the γ parameter might have a large effect on the overall uncertainty of the model results.

Two problematic approaches dominate in the literature. First, some studies make arbitrary choices: Santos et al. (2022) simply set γ to 1, while Helbig et al. (2022) tests three arbitrary values without justification. Second, other studies compromise the uncertainty information for some components to match targets for others. Paoli et al. (2018) parameterize the Dirichlet distribution in such a way to "match the uncertainty in the largest part [i.e. share] to the specified value" acknowledging this "exaggerate[s] the uncertainty of small parts". Similarly, Lupton and Allwood (2018) determine the concentration parameter γ to give a desired variance for one of the disaggregates, preventing independent control over other components' uncertainties. Going one step further, Kim et al. (2025) optimize γ (which they call λ) for exchanges "with higher production volume shares" by first estimating how much uncertainty each component would need on its own and then setting γ to the simple average of those "uncertainty scores" for the exchanges that are larger than the market average. Meyer et al. (2017) (implicitly) choose the γ parameter so that the Dirichlet distributed shares exhibit exogenously assumed standard deviations.

Only Charpentier Poncelet et al. (2022) attempt a systematic approach using pedigree matrices, but apply the same variance measure to all shares regardless of their individual uncertainty characteristics.

Against this background, a principled method is needed to choosing the parameters of a Dirichlet distribution.

2.2.2 Lack of flexibility of the Dirichlet distribution

The second limitation of the sampling-based approach using the standard Dirichlet distribution is its lack of flexibility. The standard Dirichlet distribution does not allow for cases in which the uncertainty differs between the components, nor for cases with partially missing information on the components. However, outside IE research approaches

[2]Note that in the literature the parameterization of the Dirichlet distribution differs. Alternative parameterizations found are listed in the Supporting Information

have been suggested and applied to increase the flexibility of the standard Dirichlet distribution. To allow for handling different uncertainties between components, several generalizations of the Dirichlet distribution have been suggested (Plessis et al., 2010; Lingwall et al., 2008). To address partially missing information, several forms of hybrid Dirichlet sampling procedures have been suggested (Plessis et al., 2010; Luedeker, 2022; Ng et al., 2011).

2.3 Summary of gaps in current approaches

To summarize, we identify the following key gaps in the current approaches and implementation of uncertainty analysis involving data disaggregation:

- First, even though data disaggregation is a common problem faced by IE scholars, yet despite its apparent advantages within the sampling based approach, the Dirichlet distribution has rarely been applied in IE research to estimate model uncertainty. This highlights the need for an clear and easy to follow recipe on how to use the Dirichlet distribution within IE research.
- Second, the few IE studies using a Dirichlet distribution to sample variables with a "sum-to-one-constraint" rely on somewhat arbitrary assumptions concerning selecting the concentration parameter γ. A more principled method for choosing parameter values in situations with incomplete information is therefore needed.
- Third, the sampling approach based on the standard Dirichlet distribution lacks flexibility with regard to mixed or partially missing information.

3 Practical approach to modeling uncertainty in disaggregation

Against those gaps in current approaches, in the following we introduce the theory and a practical guide that IE modelers can use to track correlations stemming from data disaggregation in their models. To ease the application by IE modelers, we implemented the proposed approach in an R and Python package, respectively (see links below).

3.1 Notation

We use uppercase letters to denote a random variable. Hence, the aggregate is denoted Y_0, the disaggregates (components) $Y_1, ..., Y_{\mathrm{K}}$, and the shares/fractions $X_1, ..., X_{\mathrm{K}}$. We use x_i and y_i to denote a particular realization of a random variable. Aggregate variables or realizations thereof are denoted with a subscript 0, such as X_0. Disaggregates and shares/fractions with a subscript $i \in 1, ..., \mathrm{K}$, e.g., X_i or X_3. Further, we denote the best-guess (in the form of the expected value) of variable i as m_i and the uncertainty (in the form of a standard deviation) as s_i.

3.2 General procedure

Uncertainty propagation using MC requires first to assign probability distributions to the model inputs. In a process that has the form illustrated in Figure 3 (i.e., which involves data disaggregation), there is one explicit model input Y_0, and K implicit model inputs in the form of the vector of best-guesses of the shares $\boldsymbol{m} = (m_1, ..., m_{\mathrm{K}})$ stemming from proxy data.

By repeatedly (N-times) and randomly sampling from those probability distributions, we can generate sets of N random variants of both model variables $\{y_0^1, y_0^2, ..., y_0^N\}$ and $\{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^N\}$. Using Equation 3 we calculate N random samples of the K disaggregates $\{y_1^1, ..., y_1^N\}, \{y_2^1, ..., y_2^N\}, ..., \{y_{\mathrm{K}}^1, ..., y_{\mathrm{K}}^N\}$.

3.3 Assigning probability distributions to the model inputs

The first step in uncertainty analysis is assigning probability distributions to the model inputs. This would be straightforward if the modeler had good statistical information on the 'real' distribution of the model input. In that situation this information should of course be used directly, but in many cases only more limited information is available (e.g., the mean, standard deviation, lower/upper bounds), which must then be used to choose an assumed probability distribution. To make this choice as objective as possible, Jaynes (1957) introduced the Maximum Entropy (MaxEnt) principle. According to Jaynes (1957), among all probability distributions that align with a given set of constraints and information, the one with the maximum entropy should be selected. The MaxEnt principle implies that the chosen distribution is maximally uninformative about what is unknown and maximally informative about what is known. Consequently, the MaxEnt distribution provides the least biased estimation consistent with the provided constraints and information. In this paper, we aim to adhere to the MaxEnt principle where possible.

We cover different cases that differ in the kind of information a modeler has concerning the aggregate variable (four cases, Table 1) and the shares (four cases, Table 1), which represent the most common cases from our own work and the IE literature.

3.3.1 Aggregate

Choosing the MaxEnt distribution for an isolated datum such as the aggregate Y_0 is straightforward and well-studied. Others have done the theoretical derivation of the MaxEnt distributions for all of the cases covered in this work, and we refer to the existing literature here. Table 1 lists the four cases covered in this paper, differing in terms of information available along with their MaxEnt distribution. We consider cases where modelers have information on upper and/or lower bounds (a_0 and b_0, respectively), a best-guess (or expectation) m_0 and/or an uncertainty estimate in the form of the standard deviation s_0. Since variables in IE modeling often reflect physical or monetary flows, they are usually constrained to be positive. In that case the lognormal distribution is a common option in IE research (Qin and Suh, 2017; Laner et al., 2014). The lognormal distribution is the MaxEnt distribution when the mean and the variance of the natural logarithm of a random variable X are known.

We note that in IE research a variety of other distributions can and are used for isolated random variables (Heijungs, 2024). The choice of distribution should always be guided by knowledge we have about our data at hand. In particular, if a researcher has knowledge of the distribution of the aggregate then they should use this instead of the MaxEnt solution listed here.

3.3.2 Shares

For the shares, we consider four different cases which differ in the amount of information available to the modeler (see Table 1):

1. No specific information is available apart from the number of processes (K) an aggregate quantity is linked to.
2. Some proxy or auxiliary data exist, allowing the modeler to calculate the best guess $\boldsymbol{m}$ of the shares.
3. Information on both the shares' best-guesses $\boldsymbol{m}$ and uncertainties on these 'best-guesses' $\boldsymbol{s}$.
4. Incomplete information on the shares' best-guesses and/or uncertainties (for example, we might only have information on the uncertainty of one component).

Finding probability distributions for these four cases is more challenging than for the aggregate quantity. As seen in the literature review, the multivariate Dirichlet distribution is a good choice for sampling shares because samples from the Dirichlet distribution always sum to one. Moreover, Vlad et al. (2001) showed that the Dirichlet distribution is the distribution with the maximum entropy for the constraints of positivity and sum-to-one. For the third and fourth cases, however, the standard Dirichlet distribution lacks the flexibility required to introduce different uncertainties for each disaggregate item. That is why we use a generalized variant of the Dirichlet distribution (Plessis et al., 2010) for case three and a hybrid Dirichlet distribution for case four. However, for these last two cases, we diverge from the MaxEnt principle as we are neither aware of any work on finding the MaxEnt distribution given the information and constraints, nor were we able to find the solution (if there is any) by ourselves within the scope of this paper.

The standard Dirichlet distribution

Formally expressed, the Dirichlet distribution describes $K \geq 2$ random variables $X_1, ..., X_\mathrm{K}$ such that each $x_i \in (0, 1)$ and $\sum_{i=1}^{K} x_i = 1$. In its most commonly used version, the Dirichlet distribution is parameterized as follows:

$$x_1, ..., x_\mathrm{K} \sim Dir^*(\mu_1, ..., \mu_\mathrm{K}), \tag{4}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_\mathrm{K})$ is a vector of K positive reals.

The expected value is given by

$$E[X_i] = \frac{\mu_i}{\sum_{k=1}^{K} \mu_k}. \tag{5}$$

Table 1: The cases this paper addresses, differing in terms of the amount of information available on both aggregate data and shares, and their respective probability distributions. The aggregate cases and their MaxEnt probability distributions are based on Rodrigues (2016), while the share cases and their probability distributions are based on Plessis et al. (2010). For aggregate cases: μ = Mean of X, σ = Standard deviation of X, μ^*/σ^* = Mean/Standard deviation of lnX, a = minimum value of X, b = maximum value of X. For share cases: Dir = standard Dirichlet distribution, $Dirg$ = generalized Dirichlet distribution, $Dirh$ = hybrid Dirichlet distribution, $\boldsymbol{\alpha}$ = Mean of sector shares/branching ratios (vector), $\boldsymbol{\beta}$ = Standard deviations of sector shares/branching ratios (vector), γ = concentration parameter, γ^* = fitted concentration parameter to MaxEnt. $\boldsymbol{\{1/K\}}_K$ denotes a vector of $(1/K, 1/K, ...)$ of size K.

case id	Information available	Probability distribution
aggregate 1	a_0, b_0	$Unif(a_0, b_0)$
aggregate 2	m_0, s_0	$Norm(\mu = m_0, \sigma = s_0)$
aggregate 3	$m_0, a_0 = 0$	$Exp(\lambda = 1/m_0)$
aggregate 4	m_0^*, s_0^*	$Lognorm(\mu^* = m_0^*, \sigma^* = s_0^*)$
shares 1	K	$Dir(\alpha = \boldsymbol{\{1/K\}}_K; \gamma = K)$
shares 2	$m_1, \ldots, m_{\mathrm{K}}$	$Dir(\alpha = \boldsymbol{m}; \gamma = \gamma^*)$
shares 3	$m_1, \ldots, m_{\mathrm{K}}$; $s_1, \ldots, s_{\mathrm{K}}$	$Dirg(\boldsymbol{m}; \boldsymbol{s})$
shares 4	$m_1, \ldots, \mathrm{NA}, \ldots, m_{\mathrm{K}}; s_1, \ldots, \mathrm{NA}, \ldots, s_{\mathrm{K}}$	$Dirh(\boldsymbol{m}; \boldsymbol{s})$

Thereby, $\sum \mu_i$ provides the "concentration" of the variables. Since we want to model this concentration explicitly in the paper, we use a slightly modified version of the Dirichlet distribution which makes the concentration parameter γ explicit:

$$x_1, ..., x_{\mathrm{K}} \sim Dir(\alpha_1, ..., \alpha_{\mathrm{K}}; \gamma), \tag{6}$$

which is parameterized by a vector of positive-valued parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{\mathrm{K}})$ such that $\sum_{i=1}^{K} \alpha_i = 1$, and an additional positive-valued concentration parameter $\gamma > 0$.

Both parameterization forms yield the same distribution if

$$\mu_i = \gamma \alpha_i, \forall i \in \{1, ..., \mathrm{K}\} \tag{7}$$

The Dirichlet distribution Dir has the useful property that the expected values for each variable X_i equal the parameter value α_i:

$$E[X_i] = \alpha_i, \forall i \in \{1, ..., \mathrm{K}\}. \tag{8}$$

The concentration parameter γ, on the other hand, controls the variance(s) of $\boldsymbol{X} = (X_1, ..., X_{\mathrm{K}})$. This is illustrated in Figure 4 showing histograms of 10000 Dirichlet distributed random numbers with the same average sector shares $\boldsymbol{\alpha} = (0.1, 0.3, 0.6)$ but with different values of γ. From that, we see that the variance decreases with increasing γ. In other words, the distributions become more concentrated with higher values of γ.

Since for cases 1 and 2 we assume not to know the uncertainties of the shares, we apply the MaxEnt principle to determine the value of γ that maximizes the entropy of the

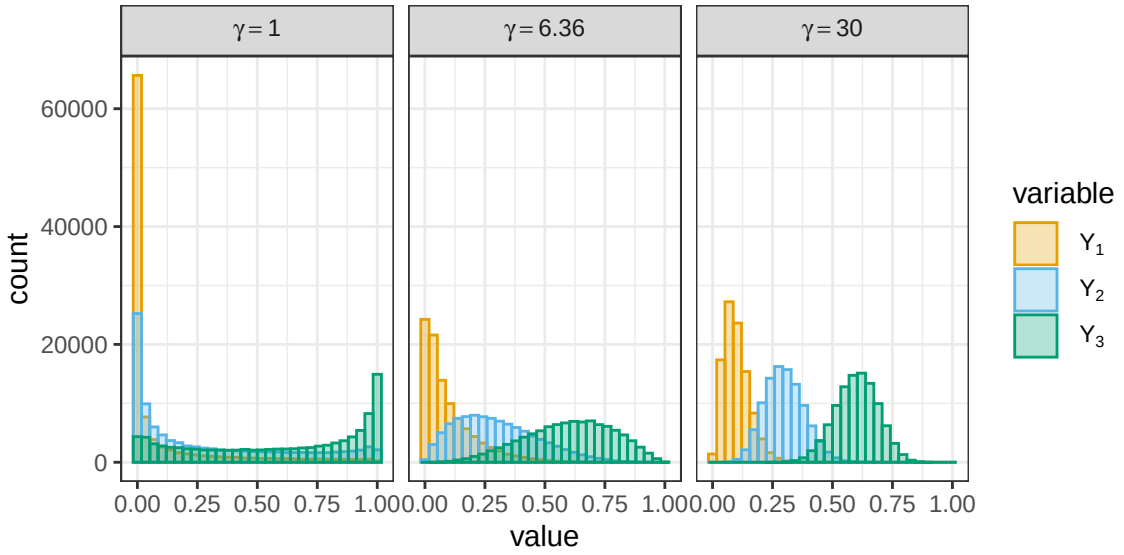

Figure 4: Histograms of shares sampled from Dirichlet distributions with three different values of γ (N = 10000, $\alpha = (0.1, 0.3, 0.6)$). $\gamma = 6.36$ marks the MaxEnt solution for that specific case. To check the code to reproduce the figure and the data behind it, please see the "Data and code availability" statement.

Dirichlet distribution given the $\boldsymbol{m}$ provided (more details in Section B.4 of the Supporting Information). For case 1, which assumes no knowledge of shares, the MaxEnt Dirichlet distribution is

$$Dir(\alpha = \{\frac{\mathbf{1}}{\boldsymbol{K}}\}_K; \gamma = K), \tag{9}$$

where $\{\frac{\mathbf{1}}{\boldsymbol{K}}\}_K$ denotes a vector of only $\frac{1}{K}$'s of size K. This distribution is called a flat Dirichlet or the maximal uninformative Dirichlet.

In the second case, the modeler has some proxy or auxiliary data to calculate a best-guess of the shares. In this case, optimizing the γ to find the Dirichlet distribution with the maximum entropy depends on the values of the $\boldsymbol{\alpha}$. Hence, there is not one unique solution like in case 1. We propose an optimization procedure to find $\hat{\gamma}$ (the value for γ that maximizes the entropy of the Dirichlet distribution) which is explained further in Section B.4 of the Supporting Information.

The Generalized Dirichlet distribution

With the γ parameter of the standard Dirichlet distribution, we can only adjust the concentration of *all* sampled shares $\boldsymbol{X}$ *simultaneously*. Case 3, however, in which we have not only information on the shares but also on the uncertainty of those shares, which can differ between the different components, demands more flexibility. For this case, several generalizations of the Dirichlet distribution have been proposed. Here, we apply the one formulated by Lingwall et al. (2008) which is parameterized as:

$$x_1, ..., x_{\mathrm{K}} \sim Dirg(\alpha_1, ..., \alpha_{\mathrm{K}}; \beta_1, ..., \beta_{\mathrm{K}}), \tag{10}$$

where, following the notation from the the standard Dirichlet distribution above, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_\mathrm{K})$ is a vector with the best estimates of the shares and $\boldsymbol{\beta} = (\beta_1, ..., \beta_\mathrm{K})$ is an additional vector of positive-valued parameters, describing the input uncertainties on the α's. While the PDF is given in Lingwall et al. (2008, equation 2 and 3), we use the sampling algorithm provided in Appendix 2 of Plessis et al. (2010) to generate random numbers from this distribution. To do so one first defines two parameters: a shape parameter α_i^* and a scale parameter β_i^*:

$$\alpha_i^* = \left(\frac{\alpha_i}{\beta_i}\right)^2 \text{ and } \beta_i^* = \frac{\alpha_i}{\beta_i^2} \,, \tag{11}$$

which are then used to generate K independent samples $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_\mathrm{K}$ of length N by sampling from gamma distributions:

$$\mathrm{gamma}(\alpha_\mathrm{i}^*, \beta_\mathrm{i}^*) = \frac{\beta_i^{*\alpha_i^*}}{\Gamma(\alpha_i^*)} x_i^{\alpha_i^*-1} e^{-\beta_i^* x_i} \,, \tag{12}$$

which are then normalised via

$$x_{ji} = \frac{\tilde{x}_{ji}}{\sum_i^n \tilde{x}_{ji}} \,, \tag{13}$$

to adhere to the 'sum-to-one' constraint $\sum_i x_i = 1$.

As already noted by Lingwall et al. (2008), due to the sum-to-one constraint, the uncertainty of the sampled shares $x_1, ..., x_\mathrm{K}$ will, in general, be close but not exactly equal to the desired uncertainty $\beta_1, ..., \beta_\mathrm{K}$.

3.3.3 The hybrid Dirichlet distribution

In some cases, however, a modeler only has partial information on a composition's best guesses or uncertainties. For example, we might only have information on the best-guesses of some components, while for others not at all. Here, we propose and implement a hybrid Dirichlet sampling approach, which shares the same general form as the generalized Dirichlet:

$$x_1, ..., x_\mathrm{K} \sim Dirh(\alpha_1, ..., \alpha_\mathrm{K}; \beta_1, ..., \beta_\mathrm{K}), \tag{14}$$

but allows for missing elements both in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

In this hybrid Dirichlet approach, the disaggregates are divided into different parts based on the available information, each of which is then sampled semi-independently before all parts are combined again. The hybrid Dirichlet algorithm is described by the following stepwise procedure:

1. **Handling missing means:** If elements in $\boldsymbol{\alpha}$ are missing, the remaining mass is evenly distributed among the missing components so that $\sum \alpha_i = 1$.

2. **Truncated Beta draw for the components with SDs:** All components i for which an uncertainty estimate β_i is available are independently sampled from a Beta distribution with shape parameters α_i and β_i, resulting in an array of size (N, N_β), where N_β is the number of components we have a β value for. Next we reject and resample any row whose sum exceeds 1.

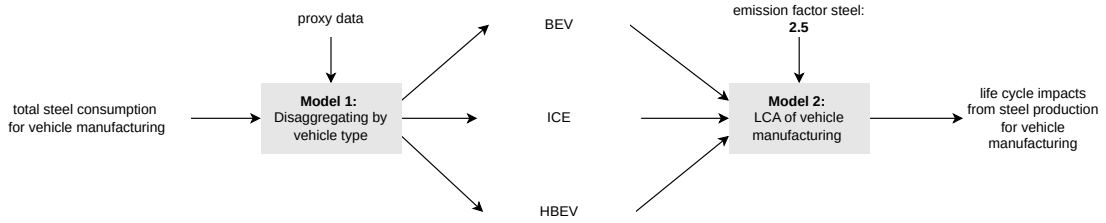

Figure 5: The overall procedure for an illustrative example with two simple models: *Model 1* which disaggregates total steel consumed for vehicle manufacturing by vehicle type (ICE, BEV, and HBEV) based on production volume proxies. Those disaggregated figures are then used by *Model 2* to estimate the life cycle impacts from steel production for vehicle manufacturing.

3. **Bias check and automatic downgrade (iterative):** The rejection criteria in Step 2 can lead to a truncation of some Beta distributions which have either a high α_i (high share of the total) or a high β_i (high uncertainty). The truncation of the distribution tail leads to bias in the expectation value. Therefore we compute the relative bias as $\text{bias}_i(X_i, \alpha_i) = \frac{|\bar{x}_i - \alpha_i|}{\alpha_i}$ for all components drawn in Step 2. Whenever bias_i exceeds the user-set tolerance (which is 0.1 in our case), ignore the corresponding β_i (set to `NA`) and repeat Steps 2–3. The loop stops as soon as every Beta-sampled component satisfies the bias criterion or no β_i remain.

4. **Maximum-entropy Dirichlet draw for the remainder:** All remaining components, whose β_i is `NA`, from a standard Dirichlet distribution with concentration γ chosen so that the entropy of the distribution is maximized.

5. **Row-wise rescaling:** Finally, the Dirichlet part (leaving the Beta draws untouched) is rescaled so that each sample satisfies $\sum_{i=1}^{K} x_i = 1$.

With this approach we make sure that the possible bias caused by the rejection of Beta draws in step 2 stays within an acceptable – user defined – range.

So, this methods gives a practical solution for sampling from Dirichlet distributions with partial information with reasonable parameters, falling back gracefully to a feasible distribution when the exact distribution cannot be sampled.

3.3.4 A simple example

Having introduced the theory of how to sample disaggregates, we next apply the procedure to a very simple example that follows the schema introduced in Figure 1 in the introduction. We assume that an IE researcher has data on total steel consumption for vehicle manufacturing but needs to disaggregate this figure by vehicle type (ICE, BEV, and HBEV) based on production volume proxies (Figure 5).

In Figure 6 we show three different scenarios, which differ in terms of what information is available on the aggregate (total steel consumption) and the shares (by vehicle type). In all three scenarios the modeler has a best-guess and standard deviation for the aggregate. For the shares, scenario 1 depicts the case that only best-guesses are available, scenario 2 shows the case where for one component both a best guess and a standard

deviation is available while for the other two no information at all is available. Lastly, scenario 3 depicts the case where best-guesses and standard deviations are available for all components.

Following the procedure described above, the aggregates and shares are sampled from the respective distributions. In scenario 1 the samples of the shares samples are generated from a standard Dirichlet distribution with the γ fitted for MaxEnt, in scenario 2 from a hybrid Dirichlet and in scenario 3 from a generalized Dirichlet. The aggregate samples come in all three cases from a lognormal distribution.

In the the R and Python packages which accompany this paper the choice of distribution for the aggregate and the shares, the determination of the parameters (in particular γ^*), and the sampling itself is all automated so that the user only has to specify the information available.

As can be seen in the right column of Figure 6, the three different configurations lead to very distinct correlation patterns. In scenario 1 a low uncertainty of the aggregate with a maximal wide uncertainty of the shares leads to negative correlations between all components. In scenario 2 a higher uncertainty of the aggregate with a mix of low uncertainty of one component (ICE) and a high uncertainty of the other two, leads to a slightly negative correlation between BEV and HBEV and positive correlations between the ICE and BEV and ICE and HBEV. In scenario 3 a high uncertainty of the aggregate together with a very small uncertainty on the shares leads to strong positive correlations between the shares of all three vehicle types.

4 On the challenge of sharing correlated data

So far, we have discussed generating samples for disaggregates based on the information provided. Integrating this sampling approach into uncertainty propagation allows modelers to obtain samples of their model results. The challenge then becomes how to communicate and share those results that include uncertainty effectively. Unlike deterministic results, which can be shared as a simple number (if results are 1-dimensional) or as (multidimensional) numeric arrays, Monte Carlo (MC) samples are more complex. While sharing full MC samples provides complete information (Lesage et al., 2018), it can lead to data storage issues, especially for large models like Multi-Regional Input-Output (MRIO) databases.

To address this, modelers often share summary statistics of the sample, typically mean and standard deviation (Lenzen et al., 2013). Researchers can use these statistics to propagate uncertainty in their subsequent analyses by independently sampling the different elements of the initial model results. Independent sampling is usually carried out using univariate distributions where each variable is treated separately, ignoring how they might be related to each other.

However, if data disaggregation is involved at any step of the model, samples of the elements of the model results are naturally correlated (Figure 6). Hence, when sharing only the mean and standard deviation, one loses information on correlations between model elements. As already pointed out in the introduction, this can lead to misleading conclusions about the likelihood of one option being preferable to another, and to inconsistencies when reconstructing disaggregates, potentially over- or underestimating

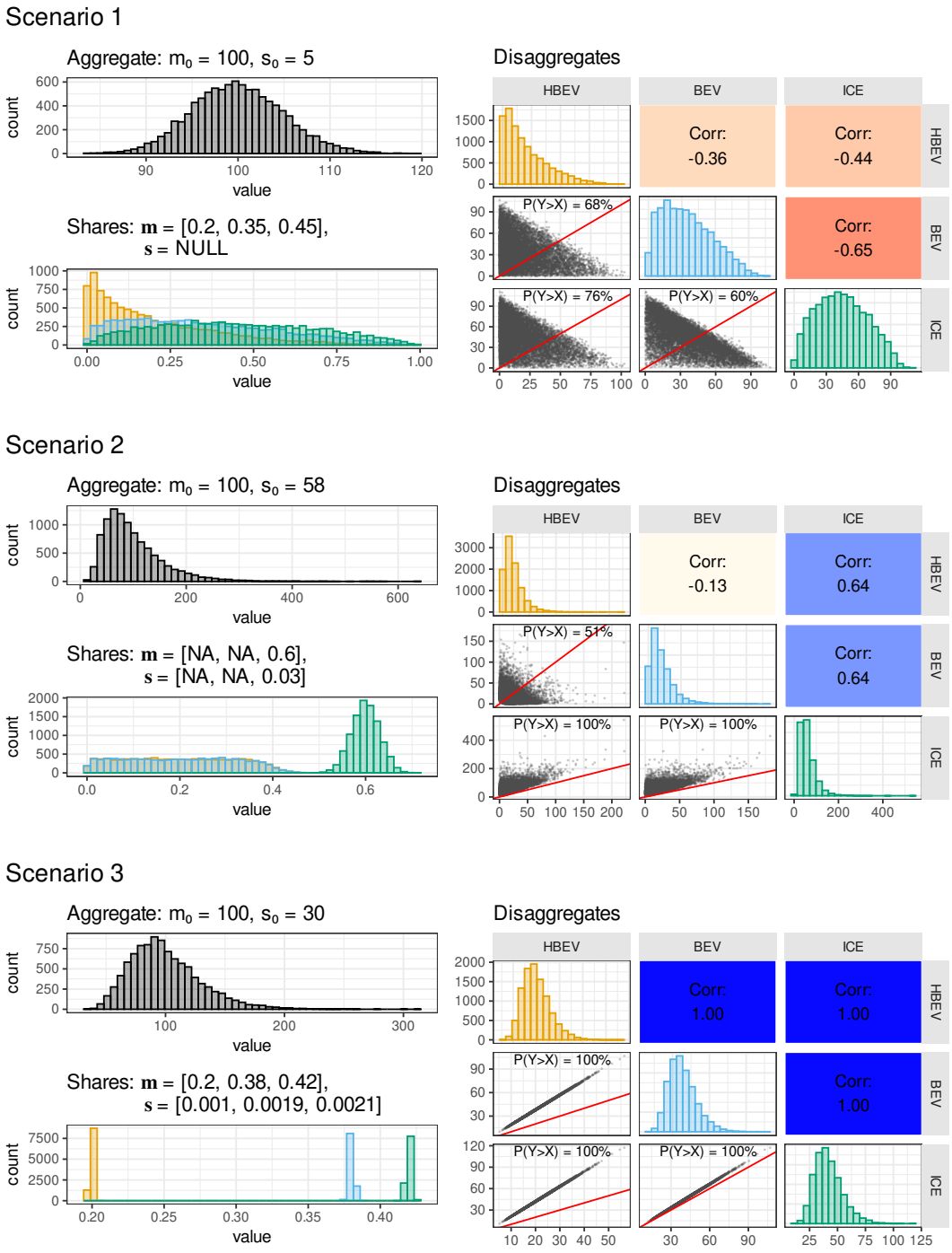

Figure 6: Three different scenarios for the example from Figure 5, which differ in terms of what information is available on the aggregate (total steel consumption) and the shares (by vehicle type). Each scenario leads to different correlation patterns. To check the code to reproduce the figure and the data behind it, please see the "Data and code availability" statement.

aggregate uncertainty.

Let us illustrate this with the example from before on steel consumed by vehicle type. Imagine a second modeler using the results from Model 1 (the steel consumption by vehicle type) to calculate the life cycle emissions from steel production for vehicle manufacturing, as a weighted average across the different vehicle types (let's call this Model 2). Depending on the uncertainty information shared by Modeler 1 and the correlation between the results from Model 1 (end use flows of steel to vehicle type), the uncertainty of the average life cycle emissions (Model 2) may be over- or underestimated. Figure 7, shows the results of Model 2 for the different correlation scenarios from the above example (Figure 6). Each column in Figure 7 corresponds to a correlation scenario (negative, mixed, positive) and shows the uncertainty distributions for three different ways of communicating the uncertainty in the results from Model 1.

First, the black histograms/boxplots show the uncertainty of the average life cycle emissions if the full uncertainty distributions from the MC sampling in Model 1 are used (assuming they were shared by Modeler 1). In the following, we will consider this as the 'truth'.

Second, the yellow histograms/boxplots in Figure 7 show the case in which only the mean and standard deviation of the end use flows are shared and used in a univariate Gamma distribution to sample from in Model 2. We see that if correlations are negative as in scenario 1, neglecting them leads to an overestimation, and vice versa for positive correlations (scenario 3). Ignoring the mixed correlations from scenario 2 leads to no particular bias.

To improve the results for the cases when model results are correlated, sharing the covariance matrix along with the Mean values provides more complete information compared to sharing the univariate standard deviations. The Mean values and the covariance matrix can be fed into a multivariate distribution, which enables all elements to be sampled simultaneously, taking correlations into account. This case is illustrated as the third option in blue in Figure 7, which show results where the values of the end use flows from Model 1 were sampled from a multivariate gamma distribution using the means and covariance matrix. The Gamma distribution was chosen because for our data, the values are always positive, and this distribution performed better than alternatives in our tests (see the Supporting Information).

We can see that this "intermediate" approach better captures the relationships between variables than independent sampling, with less data needed than sharing the full set of sampled values, though it slightly overestimates aggregate uncertainty when the disaggregates are negatively correlated (scenario 1). This happens because gamma distributions, unlike normal distributions, have mathematical constraints that make it difficult to model strong negative relationships between positive-valued variables (Minhajuddin et al., 2004).

To summarize, if sampled model results contain correlations, the best option is always to share the full uncertainty samples from the Monte Carlo Simulation to preserve all information. If that is not feasible due to storage or other constraints, the second-best option is to share the covariance matrix and mean or median values. Additionally, modelers could provide an analysis on which multivariate distribution fits their data best (like we did in the Supporting Information where we compared three different distributions), and estimate the error introduced when re-sampling. The least best option, which can

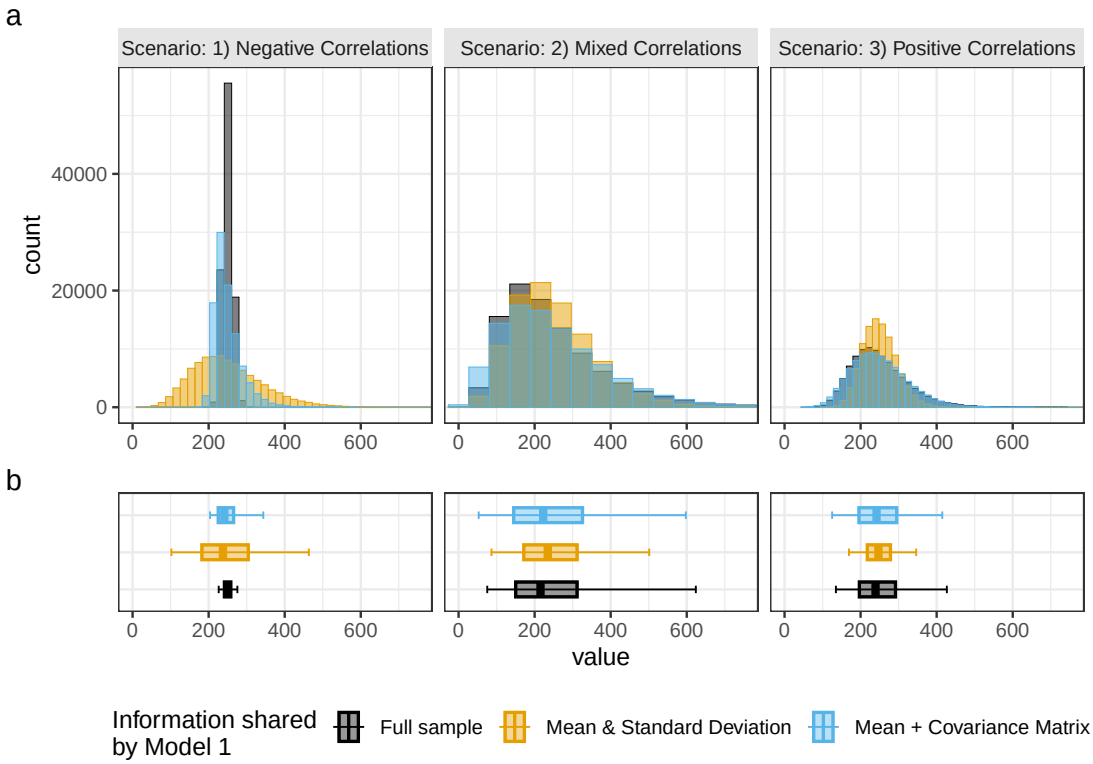

Figure 7: A comparison of the uncertainty estimation in Model 2, for different levels of available information on the uncertainty of the input data (output of Model 1 in the example in Figure 6). The three columns show the different correlation scenarios and the different color histograms **(a)** represent the uncertainty distributions for the results of Model 2 for each level of available uncertainty information. The black case assumes that the full samples from the Monte Carlo simulation in Model 1 have been shared, for the yellow case only the means and SDs were used (sampled from a univariate Gamma distribution), and the blue case assumes the availability of the means and covariance matrix (sampled from a multivariate Gamma distribution. The Boxplots **(b)** show the median (middle bar), 25th and 75th percentiles (box), and 2.5th and 97.5th (whiskers) of the distributions in **(a)**. Note: The x-axis of the Scenario 2 plot has been cut at 300 due to the distributions' long tails extending until almost 1000. To check the code to reproduce the figure and the data behind it, please see the "Data and code availability" statement.

lead to a serious misestimation of model uncertainties, is to only share the mean and SD, without any information on correlations.

5 Discussion and conclusion

This paper presents an approach to conducting an uncertainty analysis for models that include data disaggregation. The approach builds on different variants of the Dirichlet distribution to sample shares with the inherent sum-to-one property. The approach is flexible concerning information available on the input data. It can handle different levels of available information, both on the aggregate and the shares to sample from, while inherently accounting for statistical correlations.

Regarding sharing and reusing model results involving data disaggregation, we show that ignoring correlations can theoretically lead to both under- and overestimation of uncertainty. The actual importance of neglecting correlations varies between models; while in the main paper we have used simple examples to illustrate the issues and solutions, a more realistic case study on compiling German CO_2 satellite accounts shows that it also makes substantial differences in practice to the level of uncertainty estimated, most often leading to overestimation. Sharing and using the covariance matrix in addition to the sample mean leads to considerably more accurate results. However, inaccuracies persist, at least when the data is constrained (e.g., to be non-negative), since sampling from a multivariate distribution, which is only defined for a constrained space, fails in perfectly matching negative correlations. Against this background, we recommend sharing and using the entire MC sample to retain all information on data dependencies. If not possible due to constrained storage capacity, we recommend sharing at least the mean and covariance matrices so that succeeding users of the model results can resample the data from a multivariate distribution.

Further research could focus on identifying those elements that contribute most to the overall model uncertainty, e.g., by applying a global sensitivity analysis (Kim et al., 2022) or alternative approaches (Qin and Suh, 2021) so that data gathering can be prioritized more efficiently.

Building on the core sampling methods we present here, there are three areas where further development would be useful. First, the sampling methods in this paper reflect "statistical" correlations, in the sense that they are only determined by the properties of the data. There can also be "real-world" (physical) correlations, i.e., those that exist due to dependencies present in the real world, such as the relation between heat and electricity produced by a combined heat power plant. Further developments could include prior information on such "physical" correlations between individual shares or an aggregate and one/several shares. This would involve a more flexible alternative to the Dirichlet distribution, such as the logistic-normal distributions (Aitchison and Shen, 1980). However, generating random numbers from a multivariate logistic-normal distribution so that the sample means equals the best-guesses on the shares is hardly possible since there is no analytical for the mean or the SD (though there might be ways to solve this numerically). Another option to include prior correlations would be sampling based on the Monte Carlo Markov Chain (MCMC) approach (Andrieu et al., 2003), which, however, is very computationally intensive. Moreover, we consider that IE modelers very rarely possess

information on those dependencies. Second, although our analysis is mainly based on the principle of MaxEnt, we deviate from it in two cases: Neither for the case where we assume information on both best-guesses and uncertainties of the shares nor for the case where we assume partly missing information, we cannot exclude that there exist other distributions than the one we proposed (generalized/hybrid Dirichlet) that have a higher entropy given the information and constraints. Future work could be carried out on refining those sampling approaches.

Third, our algorithm for sampling from Dirichlet distributions with partial information is approximate, and while it works well for parameter values likely to be encountered in practice, there could be alternative ways to describe partial information about shares which represent better the modeler's knowledge in highly uncertain situations. Nonetheless, we find the sampling methods illustrated in the paper are already good enough for most use cases, and a substantial improvement to current practice.

The sampling approach can similarly be used for any task that involves data disaggregation, or the one-to-many allocation in general. As outlined in the introduction, in IE research, this includes sampling transfer coefficients in MFA, and disaggregation of broader categories of environmental impact to more detailed economic sectors in IOA. In LCA, the sum-to-one type constraints can occur in different stages of the analysis. In the inventory analysis stage, there may be only aggregated measured data available which must be shared between the specific processes modeled (for example, total electricity consumption of a factory is measured and must be split between several processes within the factory). Similar to the MFA examples, this can be achieved via some proxy data (perhaps the total mass output of each process, or their input power ratings) which implies correlations. Another case is the allocation of impacts to different functional flows in LCA in the case of multifunctionality. While this allocation is a normative choice, and so not exactly the same as the disaggregation of empirical measurements, it is still useful to be represent uncertainty in allocation factors either because the modeler wishes to include alternative choices as an element of the total modeling uncertainty, or because the allocation factors themselves are uncertain (e.g. prices are imperfectly known when applying economic allocation). Jung et al. (2014) has discussed the modeling of uncertainty in allocation factors using an analytical approach, and the sampling-based methods discussed in this paper expand on that to allow more flexibility in the nature of the uncertainty. Despite uncertainty analysis being relatively common in LCA (compared to IO and MFA), the uncertainty of those allocations is currently not included in most uncertainty analyses. Kim et al. (2025) who model the uncertainty of market mixes in LCA marks a very recent exception, yet their sampling could be made more coherent and flexible with the procedure presented in this paper e.g. by allowing different uncertainties for different shares using the generalized Dirichlet distribution.

Through the implementation in the form of the Python package *maxent_disaggregation* and R-package *MaxentDisaggregation* accompanying the paper, the approach can easily be incorporated into most MC workflows. With that, we hope to contribute to lowering the technical barrier to conduct uncertainty analysis of IE models, transitioning uncertainty assessment from an optional add-on to a standard practice in IE studies.

Authors' contributions

SiS and AJ conceived of the presented idea. SiS, AJ and RL conceptualized and designed the research. SiS and AJ performed the computations and analyzed the results. All authors discussed the results. SiS wrote the paper with inputs from all the authors.

Data and code availability

The data and code that support the findings of this study are all openly available:

- The Python version of the package is available via PyPi and Anaconda, with the source code on github `https://github.com/jakobsarthur/maxent_disaggregation`, as well as on Zenodo with DOI: 10.5281/zenodo.15606672 and package documentation at `https://maxent-disaggregation.readthedocs.io/en/latest/`.
- The R-package created and used for this paper is available on Github `https://github.com/simschul/MaxentDisaggregation` and Zenodo with `https://zenodo.org/records/15611532`
- The R-code to reproduce all figures and the results from our case study is available on Github: `https://github.com/simschul/uncertainty_disaggregation`
- The data needed to reproduce the case study is available on Zenodo: DOI:10.5281/zenodo.13806019
- The data behind Figure 2, 4, 6 and 7 is available on Zenodo: DOI:10.5281/zenodo.15746585

Acknowledgment

We thank three anonymous reviewers for their valuable feedback on the paper and our sampling approach, in particular for pointing out the problems related to truncated distributions in an earlier draft. Moreover, we thank Stefan Pauliuk for the constructive discussion of the research idea. Simon Schulte receives funding from the KR foundation. AJ receives funding from the ETH Board in the framework of the Joint Initiative SCENE.

References

Aitchison, J. and Shen, S. M. (1980). Logistic-Normal Distributions: Some Properties and Uses. Biometrika, 67(2):261–272.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. Machine Learning, 50(1):5–43.

Bornhöft, N. A., Sun, T. Y., Hilty, L. M., and Nowack, B. (2016). A dynamic probabilistic material flow modeling method. Environmental Modelling & Software, 76:69–80.

Charpentier Poncelet, A., Helbig, C., Loubet, P., Beylot, A., Muller, S., Villeneuve, J., Laratte, B., Thorenz, A., Tuma, A., and Sonnemann, G. (2022). Losses and lifetimes of metals in the economy. Nature Sustainability, 5(8):717–726.

Groen, E. A. and Heijungs, R. (2017). Ignoring correlation in uncertainty and sensitivity analysis in life cycle assessment: What is the risk? Environmental Impact Assessment Review, 62:98–109.

Guinée, J., Heijungs, R., and Frischknecht, R. (2021). Multifunctionality in Life Cycle Inventory Analysis: Approaches and Solutions. In Ciroth, A. and Arvidsson, R., editors, Life Cycle Inventory Analysis : Methods and Data, LCA Compendium – The Complete World of Life Cycle Assessment, pages 73–95. Springer International Publishing, Cham.

Heijungs, R. (2024). Probability, Statistics and Life Cycle Assessment: Guidance for Dealing with Uncertainty and Sensitivity. Springer International Publishing, Cham.

Heijungs, R., Guinée, J. B., Mendoza Beltrán, A., Henriksson, P. J. G., and Groen, E. (2019). Everything is relative and nothing is certain. Toward a theory and practice of comparative probabilistic LCA. The International Journal of Life Cycle Assessment, 24(9):1573–1579.

Heijungs, R. and Lenzen, M. (2014). Error propagation methods for LCA—a comparison. The International Journal of Life Cycle Assessment, 19(7):1445–1461.

Helbig, C., Kondo, Y., and Nakamura, S. (2022). Simultaneously tracing the fate of seven metals at a global level with MaTrace-multi. Journal of Industrial Ecology, 26(3):923–936.

Huijbregts, M. A. J. (1998). Application of uncertainty and variability in LCA. The International Journal of Life Cycle Assessment, 3(5):273.

Igos, E., Benetto, E., Meyer, R., Baustert, P., and Othoniel, B. (2019). How to treat uncertainties in life cycle assessment studies? The International Journal of Life Cycle Assessment, 24(4):794–807.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. Physical Review, 106(4):620–630.

Jung, J., von der Assen, N., and Bardow, A. (2014). Sensitivity coefficient-based uncertainty analysis for multi-functionality in LCA. The International Journal of Life Cycle Assessment, 19(3):661–676.

Kim, A., Mutel, C., and Hellweg, S. (2025). Global sensitivity analysis of correlated uncertainties in life cycle assessment. Journal of Industrial Ecology, n/a(n/a).

Kim, A., Mutel, C. L., Froemelt, A., and Hellweg, S. (2022). Global Sensitivity Analysis of Background Life Cycle Inventories. Environmental Science & Technology, 56(9):5874–5885.

Laner, D., Rechberger, H., and Astrup, T. (2014). Systematic Evaluation of Uncertainty in Material Flow Analysis. Journal of Industrial Ecology, 18(6):859–870.

Lenzen, M., Moran, D., Kanemoto, K., and Geschke, A. (2013). Building eora: A global multi-region input–output database at high country and sector resolution. Economic Systems Research, 25(1):20–49.

Lenzen, M. and Murray, J. (2010). Conceptualising environmental responsibility. Ecological Economics, 70(2):261–270.

Lesage, P., Mutel, C., Schenker, U., and Margni, M. (2018). Uncertainty analysis in LCA using precalculated aggregated datasets. The International Journal of Life Cycle Assessment, 23(11):2248–2265.

Lingwall, J. W., Christensen, W. F., and Reese, C. S. (2008). Dirichlet based Bayesian multivariate receptor modeling. Environmetrics, 19(6):618–629.

Luedeker, B. (2022). Compositional Datasets and the Nested Dirichlet Distribution. Statistical Science Theses and Dissertations. 30. https://scholar.smu.edu/hum_sci_statisticalscience_etds/30.

Lupton, R. C. and Allwood, J. M. (2018). Incremental Material Flow Analysis with Bayesian Inference. Journal of Industrial Ecology, 22(6):1352–1364.

Meyer, R., Benetto, E., Igos, E., and Lavandier, C. (2017). Analysis of the different techniques to include noise damage in life cycle assessment. A case study for car tires. The International Journal of Life Cycle Assessment, 22(5):744–757.

Min, J. and Rao, N. D. (2018). Estimating Uncertainty in Household Energy Footprints. Journal of Industrial Ecology, 22(6):1307–1317.

Minhajuddin, A. T. M., Harris, I. R., and Schucany, W. R. (2004). Simulating multivariate distributions with specific correlations. Journal of Statistical Computation and Simulation, 74(8):599–607.

Morgan, M. G., Henrion, M., and Small, M. (1990). Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press.

Ng, K. W., Tian, G.-L., and Tang, M.-L. (2011). Dirichlet and Related Distributions: Theory, Methods and Applications. John Wiley & Sons.

Paoli, L., Lupton, R. C., and Cullen, J. M. (2018). Useful energy balance for the UK: An uncertainty analysis. Applied Energy, 228:176–188.

Plessis, S., Carrasco, N., and Pernot, P. (2010). Knowledge-based probabilistic representations of branching ratios in chemical networks: The case of dissociative recombinations. The Journal of Chemical Physics, 133(13):134110.

Qin, Y. and Suh, S. (2017). What distribution function do life cycle inventories follow? The International Journal of Life Cycle Assessment, 22(7):1138–1145.

Qin, Y. and Suh, S. (2021). Method to decompose uncertainties in LCA results into contributing factors. The International Journal of Life Cycle Assessment, 26(5):977–988.

Reale, F., Cinelli, M., and Sala, S. (2017). Towards a research agenda for the use of LCA in the impact assessment of policies. The International Journal of Life Cycle Assessment, 22(9):1477–1481.

Rodrigues, J. D. F. (2016). Maximum-Entropy Prior Uncertainty and Correlation of Statistical Economic Data. Journal of Business & Economic Statistics, 34(3):357–367.

Santos, J. R., Tapia, J. F. D., Lamberte, A., Solis, C. A., Tan, R. R., Aviso, K. B., and Yu, K. D. S. (2022). Uncertainty Analysis of Business Interruption Losses in the Philippines Due to the COVID-19 Pandemic. Economies, 10(8):202.

Schulte, S., Jakobs, A., and Pauliuk, S. (2024). Estimating the uncertainty of the greenhouse gas emission accounts in global multi-regional input–output analysis. Earth System Science Data, 16(6):2669–2700.

Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., and Janssens-Maenhout, G. (2021). Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases. Atmospheric Chemistry and Physics, 21(7):5655–5683.

Streeck, J., Pauliuk, S., Wieland, H., and Wiedenhofer, D. (2023). A review of methods to trace material flows into final products in dynamic material flow analysis: From industry shipments in physical units to monetary input–output tables, Part 1. Journal of Industrial Ecology, 27(2):436–456.

Vlad, M. O., Tsuchiya, M., Oefner, P., and Ross, J. (2001). Bayesian analysis of systems with random chemical composition: Renormalization-group approach to Dirichlet distributions and the statistical theory of dilution. Physical Review E, 65(1):011112.

Zhang, H., He, K., Wang, X., and Hertwich, E. G. (2019). Tracing the Uncertain Chinese Mercury Footprint within the Global Supply Chain Using a Stochastic, Nested Input–Output Model. Environmental Science & Technology, 53(12):6814–6823.