



**HAL**  
open science

## Uncertainties of Domestic Road Freight Statistics: Insights for Regional Material Flow Studies

Jean-Yves Courtonne, Pierre-Yves Longaretti, Denis Dupré

### ► To cite this version:

Jean-Yves Courtonne, Pierre-Yves Longaretti, Denis Dupré. Uncertainties of Domestic Road Freight Statistics: Insights for Regional Material Flow Studies. *Journal of Industrial Ecology*, 2018, 22 (5), pp.1189-1201. <10.1111/jiec.12651>. <hal-01610261>

**HAL Id: hal-01610261**

**<https://hal.science/hal-01610261v1>**

Submitted on 4 Oct 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# Uncertainties of domestic road freight statistics: insights for regional material flows studies

Jean-Yves Courtonne, Pierre-Yves Longaretti, Denis Dupré

## Abstract

Freight statistics are at the core of many studies in the field of industrial ecology because they depict the physical inter-dependencies of territories and allow to link worldwide productions and consumptions. Recent studies have been increasingly focusing on subnational scales, often relying on domestic freight data. In this perspective, this article analyses the uncertainties of the French domestic road freight survey, road being by far the most common mode of transport in the country. Based on a statistical analysis of the survey, we propose a model to estimate the uncertainty of any given domestic road transport flow. We also assess uncertainty reduction when averaging the flows over several years, and obtain for instance a 30% reduction for a 3-year average. We then study the impact of the uncertainties on regional material flow studies such as the Economy-Wide Material Flow Analysis of the Bourgogne region. Overall the case studies advocate for a systematic assessment of freight uncertainties, as neither the disaggregation level nor the quantities traded are good enough predictors. This justifies the need for an easy-to-implement estimation model. Finally, basic comparison with the German and Swedish surveys tend to indicate that the main conclusions presented in this article are likely to be valid in other European countries.

## 1 Introduction

Material flow analysis (MFA) is a systematic assessment tool used to assess the flows and stocks of a system during a period of time. Either focusing on specific substances or on the whole economy, it has been widely applied to countries and more recently to subnational scales (Binder et al., 2004; Kovanda et al., 2009; Niza et al., 2009; Browne et al., 2011; Moore et al., 2013; Theobald et al., 2016). Regarding these studies, Rechberger et al. (2014) raise two key questions: “*What level of precision do we achieve, and what level of precision do we require?*” These questions can be related to the remarks of Binder et al. (2009) regarding the lack of implementation of MFA in policy making: indeed, if the precision of results is not properly analyzed, the reliability of the studies may be questioned. On the contrary, analyzing uncertainties helps preventing premature conclusions and points to the main lacks of information. For instance, it would not make sense to try to explain a 10% difference on the per-capita consumption of two regions if it can be explained by an overlapping of confidence intervals. Laner et al. (2014) review the existing MFA literature and propose a classification of uncertainty management methods into three types of approaches: qualitative and semi-quantitative approaches (e.g., Graedel et al. 2004), approaches based on data quality classification (e.g., Weidema and Wesnæs 1996; Hedbrant and Sörme 2001; Danius 2002) and statistical approaches (e.g., Cencic and Rechberger 2008; Dubois et al. 2014). They insist on the distinction between “random” uncertainty (that cannot be reduced) and “epistemic” uncertainty (that is due to a lack of knowledge). Patricio et al. (2015) provide uncertainty quantifications of Economy-Wide Material Flow Analysis (EW-MFA) for

different spatial levels in Sweden: national, regional and urban levels<sup>1</sup>. Uncertainties of input data are estimated based on available statistical information, expert judgment and empirical estimates to account for imputation errors, for instance when a proxy is used. Domestic transport is identified as a major source of uncertainty when tackling regional and urban levels. Generally speaking, they underline that the lack of direct information implies larger uncertainties as the spatial resolution increases. They also insist on the fact that evaluating uncertainties is more common in Substance Flow Analysis (SFA) studies than in EW-MFA where they are generally treated qualitatively. In a previous article, the authors qualitatively assessed the relative reliability of different sources based on the data estimation technique (Courtonne et al., 2015): cross-checking of several surveys, some of which exhaustive (e.g., employment, agriculture production), declaration and punctual control (e.g., customs, companies' communications), modeled or downscaled data, extrapolation of a statistical survey on subpopulations (e.g., road freight). This article is a first attempt of moving from qualitative to quantitative uncertainty assessment, putting the focus on domestic road freight statistics.

In France, the statistical service of the ministry of ecology (SOeS) compiles information from various sources to provide centralized data updated every year on both international and domestic freight (SitraM database). As about 85% of merchandises transported between French regions travel by road<sup>2</sup>, a specific survey, called the TRM<sup>3</sup> survey, is dedicated to domestic road freight. It consists of a poll on French trucks (more details are provided in the methods section). Uncertainties being inherent to polls, the SOeS assesses every year the precision of the survey for the total aggregated result, that is, whatever the good transported and whatever the regions of loading and unloading. For instance, the 2010 quality assessment indicates a precision of about 1.5% on a total of about 2 Gt, which can be interpreted as: *there is a 95% chance for the real value to belong to an interval between 1970 Mt and 2030 Mt*<sup>4</sup>. Until now, however, uncertainties have never been estimated for disaggregated flows whereas these information are becoming more and more used. For instance, among the 22 French regions<sup>5</sup>, 4 have already undertaken an EW-MFA and several other are launching studies. Typically, domestic freight data can be obtained by NST 2007 positions (382 categories) up to the NUTS 3 levels (96 French départements) for loading and unloading. Extrapolations for each flow are provided along with the number of observations they are based on (see the glossary next section for a proper definition). The SOeS warns that results based on less than 10 observations shouldn't be used because they are probably not significant. This article is aiming at a more precise assessment of the uncertainties. Especially, we try to:

- Quantify uncertainties of any domestic road freight flow based on anonymised detailed results of the surveys (accessed upon special research request),
- Check whether or not this quantification is in line with the threshold of 10 observations proposed by the SOeS,
- Propose a simpler model of uncertainty quantification only based on publicly available information (e.g., number of observations, see glossary).

We believe this information can directly benefit ongoing and future MFA studies in France, and also foster research on similar topics in other countries, especially in Europe. Indeed, following the Council

---

<sup>1</sup>Note that the uncertainties reported in their article correspond to one standard deviation (68% confidence interval), whereas here we chose to express them as two standard deviations (95% confidence interval).

<sup>2</sup>Domestic inter-regional freight is about twice as large as international imports. Although it would be interesting to estimate uncertainties of customs data, this can not be done with the methodology developed in this article.

<sup>3</sup>*Enquête permanente sur le Transport Routier de Marchandises*: permanent survey on road freight.

<sup>4</sup>See the *uncertainty* section in the glossary for a more thorough explanation.

<sup>5</sup>Note that since January 1st 2016, metropolitan France is divided into 14 regions.

Regulation No 1172/98 on statistics on the carriage of goods by road, EU countries as well as candidate countries must carry surveys on road freight similar to the TRM survey. If uncertainty calculation seems to be a common practice in some countries (as Sweden and Germany as we show in the case studies), it may not be the case everywhere and it is especially unlikely for uncertainties to be available for any possible aggregation level of the flows. Moreover, to our knowledge, the question of uncertainty reduction through a multi-year average MFA has not been fully explored yet, while in some cases providing more accurate results for a 3-year average (for instance) may be more relevant than providing less precise annual results.

The article is structured into three sections. Data sources and methods are depicted in a first section. A second section is dedicated to theoretical results, and provides formulae to compute uncertainties for a single year and for a multi-year average. Because the precision required of course depends on the problem, case studies ranging from EW-MFA to MFA focusing on specific products are presented in a third section. This last section also includes a basic comparison of the French, Swedish and German surveys. Finally, a summary of the main findings and perspectives for future research are proposed in the conclusion.

## 2 Materials and methods

### 2.1 Glossary

For the sake of clarity, we provide definitions of the following terms that are used throughout the article:

- *Population*. All vehicles-weeks in France (typically 52 times the number of vehicles). The population total size is  $N$  (a quantity that varies from year to year), and elements of this population are denoted  $i$  ( $1 \leq i \leq N$ ). Vehicles-weeks are elements obtained by considering that each vehicle provides an independent elementary unit each week of the year. An element is therefore one vehicle surveyed for a single week.
- *Subpopulation*. Vehicles-weeks that meet specific criteria for at least one of their travels: specific loading area, specific unloading area and specific type of merchandise transported. We are interested in well-defined subpopulations, identified by a given origin ( $o$ ), destination ( $d$ ) and category of goods ( $c$ ) transported. These subpopulations are symbolically represented by an index  $\delta_i^c$ , taking two values:  $\delta_i^c = 1$  if  $i$  belongs to the desired  $c = (o, d, c)$  subpopulation, and  $\delta_i^c = 0$  otherwise.
- *Sample*. Vehicles-weeks surveyed. A sample is noted  $s$  and is a collection of elements of the population, usually of predetermined size  $S$ . The number of times a given element  $i$  is represented in a sample is noted  $S_i$ . For a sampling method without replacement (as is the case for the TRM survey),  $S_i = 0$  or  $1$ .
- *Subsample*. Subpopulation restricted to a sample; subsample sizes are denoted  $n$ . Note that the selection of a subsample is a post-selection procedure once the sample is drawn, and not a subsequent resampling of the sample. This definition does not conform with the conventional use of the word.
- *Number of observations*. Number of elementary operations of loading/unloading meeting specific criteria (specific loading/unloading/good); a single vehicle-week can report several observations of the same good between the same loading and unloading places.
- *Characteristic quantity*. Any quantity of interest on the total population, for example the total quantity transported per year (tonnes) corresponding to specific criteria (loading-unloading-good). Quantities

of interest are denoted  $y$  in a generic way. The most important quantity one tries to estimate through survey sampling is the total of a quantity of interest on the whole population

$$T = \sum_{i=1}^N y'_i, \quad (1)$$

where  $y'_i = N_w y_i$  and  $N_w = 52$  is the number of weeks of the annual survey. The number of weeks scaling is due to the fact that elements are vehicles-weeks and not vehicles-year (see the *Supplementary material* first section for more details). In the probabilistic description adopted here,  $y_i$  are parameters, not random variables.

For a subpopulation, the associated total is

$$T^c = \sum_{i \in c} y'_i = \sum_{i=1}^N \delta_i^c y'_i. \quad (2)$$

- *Inclusion probability.* The inclusion probability  $\pi_i$  relates to element  $i$  and defines its probability of being present in the sample. By construction the probability of absence is  $\bar{\pi}_i = 1 - \pi_i$ . Unequal probability sampling methods are characterized by the fact the  $\pi_i$  is not constant (not independent of  $i$ ); it is the form of sampling adopted in the TRM survey.
- *Coefficient of variation.* Also known as relative standard deviation, it is a measure of dispersion equal to the ratio of the standard deviation sigma ( $\sigma$ ) to the mean mu ( $\mu$ ).
- *Uncertainty.* In this article, only uncertainties due to the sampling methodology are taken into account<sup>6</sup>. They are provided in the form of 95% confidence intervals, that is nearly twice (1.96) the coefficient of variation, given an hypothesis of Gaussian distribution<sup>7</sup>. We therefore use the notation  $\mu \pm 2\sigma/\mu$  (expressed in %). Uncertainties above 100% mean the extrapolation ( $\mu$ ) is not significant and shouldn't be used.
- *Extrapolation.* procedure allowing us to estimate the quantities of interest in the whole population from a sample.

## 2.2 Description of the methodology of the TRM survey

The TRM survey focuses on motor vehicles registered in France, less than 15 years old, belonging to one of the two following categories: lorries of loading capacity (LC) higher than 3.5 tonnes and road tractors of maximum permissible laden weight (MPLW) above 5 tonnes.

The design experiment is constructed with an unequal probability sampling method, taking advantage of known contextual information about the vehicles. The unequal drawing probabilities are defined based on three factors: the age of the vehicle, the loading capacity of the vehicle and the activity of the owner. The goal is to draw more “young” vehicles and vehicles which have a high loading capacity because they have

<sup>6</sup>This excludes for instance mistakes in the answers from truck drivers, possible errors during the treatment of the forms or bias due to the handling of nonresponse. These sources of error are quite difficult to quantify. Nonresponse is handled by a simple correction factor.

<sup>7</sup>The computation of the standard deviation itself is independent from any distribution hypothesis; the Gaussian hypothesis is made in order to translate the standard deviation in terms of confidence interval. This hypothesis is used for the same purpose by the French statistical office handling the survey.

a higher contribution in the total tonnes and tonnes-kilometers transported. Half of the sample is renewed every year in order to make the year-to-year evolution more reliable.

Once the answers are collected, sampling weights are corrected for non-response and improved by a calibration procedure (CALMAR method), aiming at reconciling data estimated from the survey with otherwise known totals, for instance the number of trucks registered in each region, the number of trucks of each type etc. More information about this procedure can be found in CGDD-SOeS (2011).

The owners (or users) of the vehicles drawn for the survey are asked about the use of the vehicle during one week (the answer is mandatory). The unit of the survey is therefore the vehicle-week.

In 2010, the survey had the following characteristics (CGDD-SOeS, 2011):

- population size: 527,403
- sample size: 77,921
- number of usable forms: 49,933
- sampling rate: (sample size / population size)  $\times$  1/52 weeks = 0.28%

### 2.3 Formulae for totals and associated uncertainty estimation

We only collect here the relevant expressions for total estimators and their variance. Some details about these expressions and their justification is provided in the first section of this paper's *Supplementary material*.

The extrapolation of the annual characteristic totals is performed with the help of the Horvitz-Thomson (HT) estimator:

$$\widehat{T} = \sum_{i \in s} \frac{y'_i}{\pi_i}, \quad (3)$$

The HT estimator variance can itself be estimated by

$$\widehat{V}_s(\widehat{T}) = \sum_{i \in s} (1 - \pi_i) \left( \frac{y'_i}{\pi_i} - \frac{\sum_{k \in s} (1 - \pi_k) y'_k / \pi_k}{\sum_{k \in s} (1 - \pi_k)} \right)^2. \quad (4)$$

For a subpopulation  $c$ , the HT estimator can be expressed as

$$\widehat{T}^c = \sum_{i \in s} \frac{y'_i \delta_i^c}{\pi_i}, \quad (5)$$

with the associated variance estimator

$$\widehat{V}_s(\widehat{T}^c) = \sum_{i \in s} (1 - \pi_i) \left( \frac{y'_i \delta_i^c}{\pi_i} - \frac{\sum_{k \in s} (1 - \pi_k) y'_k \delta_k^c / \pi_k}{\sum_{k \in s} (1 - \pi_k)} \right)^2 \quad (6)$$

The 95% confidence limit is readily obtained from the variance of the estimator under a Gaussian distribution assumption<sup>8</sup>

$$\widehat{e}_{95\%} = 2 \frac{\sqrt{\widehat{V}_s(\widehat{T})}}{\widehat{T}}. \quad (7)$$

---

<sup>8</sup>The service producing the survey also uses a 95% confidence interval to compute the uncertainty on the total (all flows aggregated).

## 2.4 Data processing

The study is based on data from the French road freight surveys from 2002 to 2010; the data have been made anonymous for confidentiality purposes. For each truck of the samples, the following three types of information are given:

- targeted information: location of loading, location of unloading, category of good transported, reported tonnes, reported tonnes-kilometers,
- sampling information: statistical weight of each truck,
- contextual information (used for the sampling): age of the vehicle, type of vehicle, activity of the transporter, region of registration.

From this disaggregated information, we recreated all existing combinations of aggregation levels, distinguishing between:

- 3 levels of loading: 1 country (nuts 1), 22 régions (nuts 2), 96 départements (nuts 3),
- 3 levels of unloading (same),
- 3 levels of product category: all goods, 10 groups of goods (NSTR between 2002 and 2008) / 20 groups of goods (NST 2007 between 2009 and 2010), 176 detailed goods (NSTR between 2002 and 2008) / 382 detailed goods (NST 2007 between 2009 and 2010).

An illustration of such combinations is provided in the accompanying *Supplementary Material* file. We computed confidence intervals for each extrapolation (about 900,000 extrapolations in total for the period 2002-2010) with a C++ program implementing Eq. (7).

## 3 Theoretical results

### 3.1 Estimating uncertainties based on subsample sizes

We built a file based on the surveys from 2002 to 2010 containing the following fields for each extrapolation:

- $n$ : size of the subsample used for the extrapolation (see *subsample* in the glossary),
- $e$ : 95% error (see *uncertainty* in the glossary), computed from Eq. (7).

Extrapolations were first sorted out in 15 classes depending on the size of the subsample they are based on. 5%, 50% and 95% quantiles of uncertainties were then computed for each class<sup>9</sup>.

Power law fits to the data are proposed in figure 1. Applying these models, table 1 provides the values for lower, median and upper uncertainty estimations for different values of subsample sizes. The  $R^2$  of the median model, that is the proportion of variability it explains, is equal to 0.87, which is satisfactory. We conducted similar studies for two other predictor variables, that are available in the supplementary material. First, we took the number of observations (see glossary) as predictor variable, since it is currently the default

---

<sup>9</sup>The combinatorial effect implies that the vast majority of the studied extrapolations have small subsample sizes. Therefore data has to be organized into subsample size classes to prevent a ponderation bias during the analysis. More details are available in the *Supplementary Material* file.

variable provided by the French statistical office. The derived model has an  $R^2$  of 0.77, which is still acceptable but indicates that the size of the subsample is a better predictor and should be preferred when available. Finally, we tested the hypothesis “the higher the quantity transported the smaller the uncertainty”, using the number of tonnes transported as the explicative variable. This last model is the least efficient, with an  $R^2$  of only 0.63<sup>10</sup>.

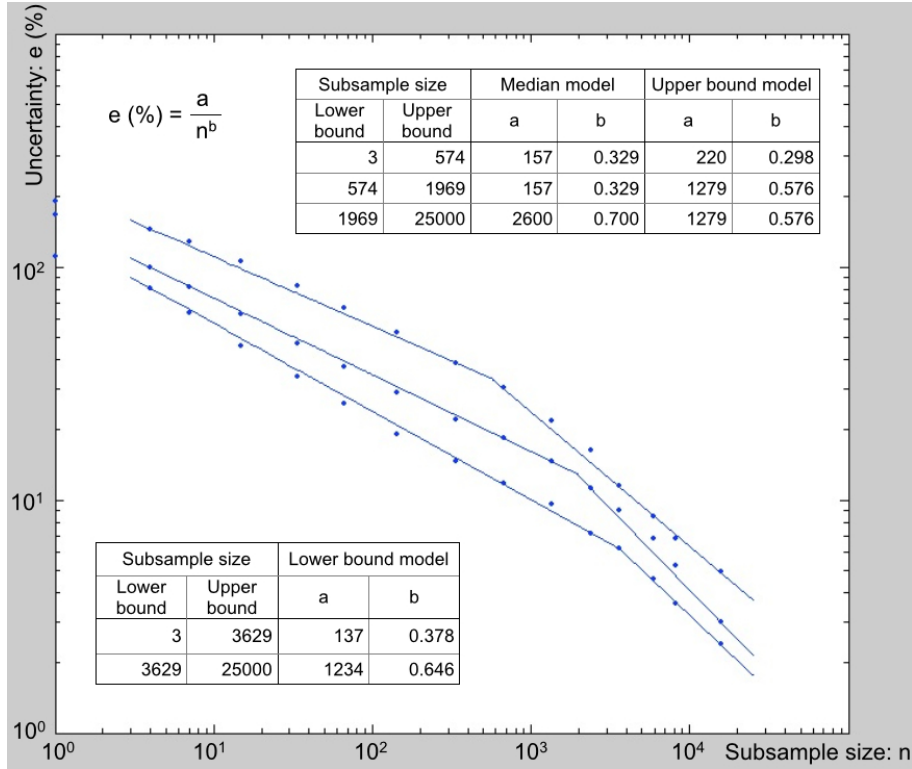


Figure 1: Piecewise models for estimating lower bound, median and upper bound uncertainties based on the size of the subsamples (NB: this is a log-log plot). Extrapolations were classified in size classes as explained in the text: each class is represented by three dots (aligned vertically). Lower dots correspond to 5% percentiles, middle dots to medians and upper dots to 95% percentiles. Explicative models were calibrated in order to properly fit the dots. The piecewise models  $e (\%) = a/n^b$  are drawn in blue lines and the parameters are provided in the two tables. The  $R^2$  of the median model is 0.87. It is interesting to notice that for the three models, the slope increases when the size of the subsample reaches a threshold: upper bounds are the first to decrease, followed by median (for nearly 2000 individuals) and lower bounds.

<sup>10</sup>All the  $R^2$  indexes are computed with the formula  $R^2 = 1 - SS_{residuals}/SS_{total}$ .

Subsample size	Uncertainty (%)		
	Lower bound	Median	Upper bound
5	75	92	136
10	57	73	110
15	49	64	98
20	44	58	90
50	31	43	68
100	24	34	56
150	21	30	49
200	19	27	45
500	13	20	34
1000	10	16	24
2000	8	13	16
5000	5	7	9
10000	3	4	6
20000	2	3	4

Table 1: Lower, median and upper uncertainty estimations for given subsample sizes (rows). We see that below a size of 15 vehicles-week, there is risk for the statistic not to be significant (above 100% of uncertainties), although it can be significant in some cases. Above 150 vehicles-week, there is a 95% chance that the uncertainty is lower than 50%, the median estimation being 30%.

### 3.2 Estimating the reduction of uncertainty when averaging over several years

Let  $T_i$  be the estimation of tonnes transported for a given subsample (given loading and unloading locations, and transported product) during the year  $\alpha$ . The average over several years  $n_y$  is:

$$\bar{T} = \frac{1}{n_y} \sum_{\alpha} T_{\alpha}. \quad (8)$$

Assuming  $T_{\alpha}$  are independent variables, we would have:

$$V(\bar{T}) = \sum_{\alpha} V\left(\frac{T_{\alpha}}{n_y}\right) = \frac{1}{n_y} \frac{\sum_{\alpha} V(T_{\alpha})}{n_y} = \frac{1}{n_y} \bar{V}(T), \quad (9)$$

where  $\bar{V}(T)$  is the mean of the  $V(T_{\alpha})$  over  $n_y$  years. Then:

$$\sigma(\bar{T}) = \left(\frac{\bar{V}(T)}{n_y}\right)^{1/2}, \quad (10)$$

and the associated error (twice the coefficient of variation):

$$e = 2 \frac{\sigma(\bar{T})}{\bar{T}} = \frac{2}{\bar{T}} \left(\frac{\bar{V}(T)}{n_y}\right)^{1/2}. \quad (11)$$

In reality one cannot assume that the annual totals  $T_{\alpha}$  are independent variables because every year, half of the sample is reconducted. To recover statistical independence, we chose to build alternative independent samples by removing redundant vehicles-weeks. If we call  $p$  the size of a full sample for one year and consider that this size is stable over the  $n_y$  years considered, we see that the full sample over  $n_y$  years will

have a size of  $n_y p$  while any independent sample over  $n_y$  years will have a size of  $p + (n_y - 1)p/2$  (see the table 3 of the *Supplementary Material* file and its accompanying comment).

This argument suggests to substitute  $[p + (n_y - 1)p/2]/p$  to  $n_y$  in the last expression of Eq. (11), leading to the following predictor for the error  $e$ :

$$e = \frac{2}{\sqrt{1 + (n_y - 1)/2}} \frac{\sqrt{\sum_{\alpha} V(T_{\alpha})/n_y}}{\sum_{\alpha} T_{\alpha}/n_y}. \quad (12)$$

This relation is plotted on figure 7 of the *Supplementary Material* for an average over 3 years, where the uncertainty  $e$  is represented as a function of  $x = 2\sqrt{\bar{V}(T)/\bar{T}}$  with  $\bar{V}(T) = \sum_{\alpha} V(T_{\alpha})/n_y$  and  $\bar{T} = \sum_{\alpha} T_{\alpha}/n_y$ . The linear behavior holds with good accuracy, with the correct slope predicted by Eq. (12). To obtain this plot, we sorted out observations depending on the value of the predictor and in each class and computed the median of the uncertainties obtained by applying Eq. (6) to the independent sample.

The results displayed in table 2 allow us to validate this model over all the possible number of years one could use for this averaging procedure. To construct this table, we have constructed the samples as explained above for each averaging option, and compared a fitted slope with the predicted one; the difference between the two is clearly negligible for all practical purposes. Theoretically, the uncertainty of the average could still be lowered because the part of the sample removed is not 100% correlated with the part kept. However, it is not possible to know the level of correlation between these sample parts, so that we advise to break the sample over several years as explained in this section and use Eq. (12) as a basic predictor.

Average over $n_y$ years	Observed slope (fitted model)	Suggested model $1/\sqrt{1 + (n_y - 1)/2}$	$R^2$ of the fitted model	$R^2$ of the suggested model
n=2 (2006-2007)	0.82	0.82	0.97	0.97
n=3 (2005-2007)	0.70	0.71	0.98	0.97
n=4 (2004-2007)	0.60	0.63	0.98	0.96
n=5 (2003-2007)	0.55	0.58	0.98	0.96
n=6 (2002-2007)	0.52	0.53	0.98	0.97

Table 2: Fitted slopes versus modeled slopes in a  $e = slope \times 2\sqrt{\bar{V}(T)/\bar{T}}$  model. The  $R^2$  index of the fitted model shows that  $[\bar{V}(T)]^{1/2}/\bar{T}$  is a very good predictor of the error. The suggested model produces slopes very close to that of the fitted model and its  $R^2$  index stays very high. Note that the value  $1 - slope$  can be interpreted as the reduction of uncertainty obtained thanks to the averaging operation (e.g. for a 3 year-average, we can expect a reduction of uncertainty of about 30%).

## 4 Case studies

### 4.1 Implications for regional EW-MFA: the case of the Bourgogne region

The methodology for Economy-Wide Material Flow Analysis was standardised by Eurostat (2001). Its objective is to quantify the physical inputs into an economic system, material accumulation in this system and outputs to other economies or back to nature, as illustrated by table 3. This table shows the material balance of the Bourgogne region as computed by Alterre-Bourgogne (2013).

We only consider the uncertainties of road freight statistics and we use Gauss's law of propagation to capture their impact on EW-MFA results. Of course this requires the use of a correspondence table between

freight and EW-MFA classifications: we use the table provided by the French official guide for regional EW-MFA (CGDD, 2014).

Regional (Eurostat nuts 2 level) results are presented in table 3. We can see that at this level road freight uncertainties are very low for total imports and exports (about 5%) and low to moderate when we start disaggregating by product categories (between 5% and 21%). Also note that with 95% intervals of confidence of respectively 2.5% and 6% DMI (Domestic Material Input) and DMC (Domestic Material Consumption) indices seem robust to road freight uncertainties, although DMC is more subject to variations. This conclusion is confirmed by the results for the four départements of Bourgogne (nuts 3 levels) displayed in table 4: while DMI uncertainties vary between 5% and 8%, DMC uncertainties vary between 12% and 21%. At this geographical level, DMCs should therefore be compared with great care and intervals of confidence should be given. This higher sensitivity of the DMC index to input data uncertainties was already underlined in Patricio et al. (2015).

Flow type	Flow	Value (Mt)	Uncertainty
Inputs (Mt)	Balancing inputs	15.8	
	Domestic extraction (unused)	20.1	
	Domestic extraction (used)	27.8	
	Imports	27.1	+/- 5%
	products from agriculture and fisheries	5.4	+/- 11%
	products from sylviculture	1.4	+/- 16%
	metallic minerals and derived products	2.4	+/- 5%
	non metallic minerals and derived products	4.9	+/- 21%
	fossil fuels and derived products	3.8	+/- 7%
	other products	9.2	+/- 7%
	Indirect flows associated to imports	131.4	
System (Mt)	Net addition to stock	11.0	
	Recycling	4.0 to 5.5	
Outputs (Mt)	Balancing outputs	15.1	
	Domestic extraction (unused)	20.1	
	To air	11.2	
	To nature (others)	6.6	
	Exports	26.8	+/- 4%
	products from agriculture and fisheries	8.0	+/- 7%
	products from sylviculture	1.6	+/- 16%
	metallic minerals and derived products	2.1	+/- 17%
	non metallic minerals and derived products	5.4	+/- 11%
	fossil fuels and derived products	0.8	+/- 13%
other products	8.8	+/- 7%	
	Indirect flows associated to exports	120.6	
Indicators	DMI	33.5 t/cap	+/- 2.5%
	DMC	17 t/cap	+/- 6%

Table 3: Economy-Wide Material Flow Analysis of the Bourgogne region (Alterre-Bourgogne, 2013). Uncertainties of trade flows (95% interval of confidence) are computed by the authors. Uncertainties of DMI and DMC indicators only incorporate the effect of domestic road freight uncertainties (other data sources are not studied here). Taking other sources of uncertainty into account would of course lead to higher uncertainties for these indicators.

Flow type	Flow	Côte d'Or	Nièvre	Saône et Loire	Yonne
Inputs (Mt)	Balancing inputs	4.5	2.5	6	2.8
	Domestic extraction (unused)	7.1	3.1	4.4	5.5
	Domestic extraction (used)	7.8	5.9	7.8	6.3
	Imports	9.4 +/- 6%	3.8 +/- 17%	12.8 +/- 9%	6.5 +/- 14%
System (Mt)	Net addition to stock	2.5	1.8	4.0	2.7
Outputs (Mt)	Balancing outputs	3.2	3.1	6.3	2.5
	Domestic extraction (unused)	7.1	3.1	4.4	5.5
	To nature	5.3	2.6	6.3	3.5
	Exports	10.7 +/- 8%	4.6 +/- 18%	10.0 +/- 7%	7.0 +/- 10%
Indicators	DMI	33 t/cap +/- 5%	44 t/cap +/- 8%	37 t/cap +/- 7%	37 t/cap +/- 8%
	DMC	12 t/cap +/- 15%	23 t/cap +/- 21%	19 t/cap +/- 12%	17 t/cap +/- 20%

Table 4: Economy-Wide Material Flow Analysis of the départements (nuts 3 levels) of Bourgogne (CGDD, 2014). Uncertainties (95% interval of confidence) of trade flows and of the indicators are computed by the authors. Uncertainties of DMI and DMC indicators only incorporate the effect of domestic road freight uncertainties (other data sources are not studied here).

## 4.2 The *disaggregation effect*: a focus on agro-products and wheat

In this section, we show the results for two categories of products: all products from agriculture, forestry and fishery on the one hand and wheat on the other hand (in the NST 2007 classification these correspond respectively to the codes 01 and 0111). In addition, we distinguish between three levels of spatial disaggregation: total domestic road freight, total road imports and total road exports of each French region and inter-regional road freight. The results presented in figure 2 clearly show that geographical and product disaggregation rapidly lead to loss of precision given the decrease in the size of the subsample. The most disaggregated results (inter-regional trade of wheat) show that there is a very large variability of uncertainties depending on the flows (here, depending on loading and unloading regions), ranging from 31% to 183%.

This clearly shows that the disaggregation level is not enough to characterize uncertainty, but that there is a large regional variation. As a consequence, uncertainties should be evaluated in each MFA study to check whether or not road freight is reliable enough depending on the question tackled. Data reconciliation techniques that couple transport data with production, transformation and consumption data seem well adapted to take advantage of available trade data while still taking its (lack of) precision into account, as shown by the authors in a previous article (Courtonne et al., 2015).

## 4.3 Uncertainty propagation: studying wheat supply areas of French regions

Another common way of using transport databases is to estimate the origin or the destination of products (see, e.g., Kastner et al. 2011; Billen et al. 2009; Courtonne et al. 2016 to name only a few). In this section, we analyze the case of supply areas.

Following Kastner et al. (2011), we define a matrix  $R$  whose terms  $R_{ij}$  represents the percentage of region  $i$  supply initially originating from region  $j$ . It is computed as follows:

$$R = (I - Z\hat{x}^{-1})^{-1}\hat{p}, \quad (13)$$

$$x = p + Z \cdot \mathbb{I}. \quad (14)$$

In these relations,  $I$  is the identity matrix;  $p$  the vector of regional productions;  $Z$  the transport matrix (with

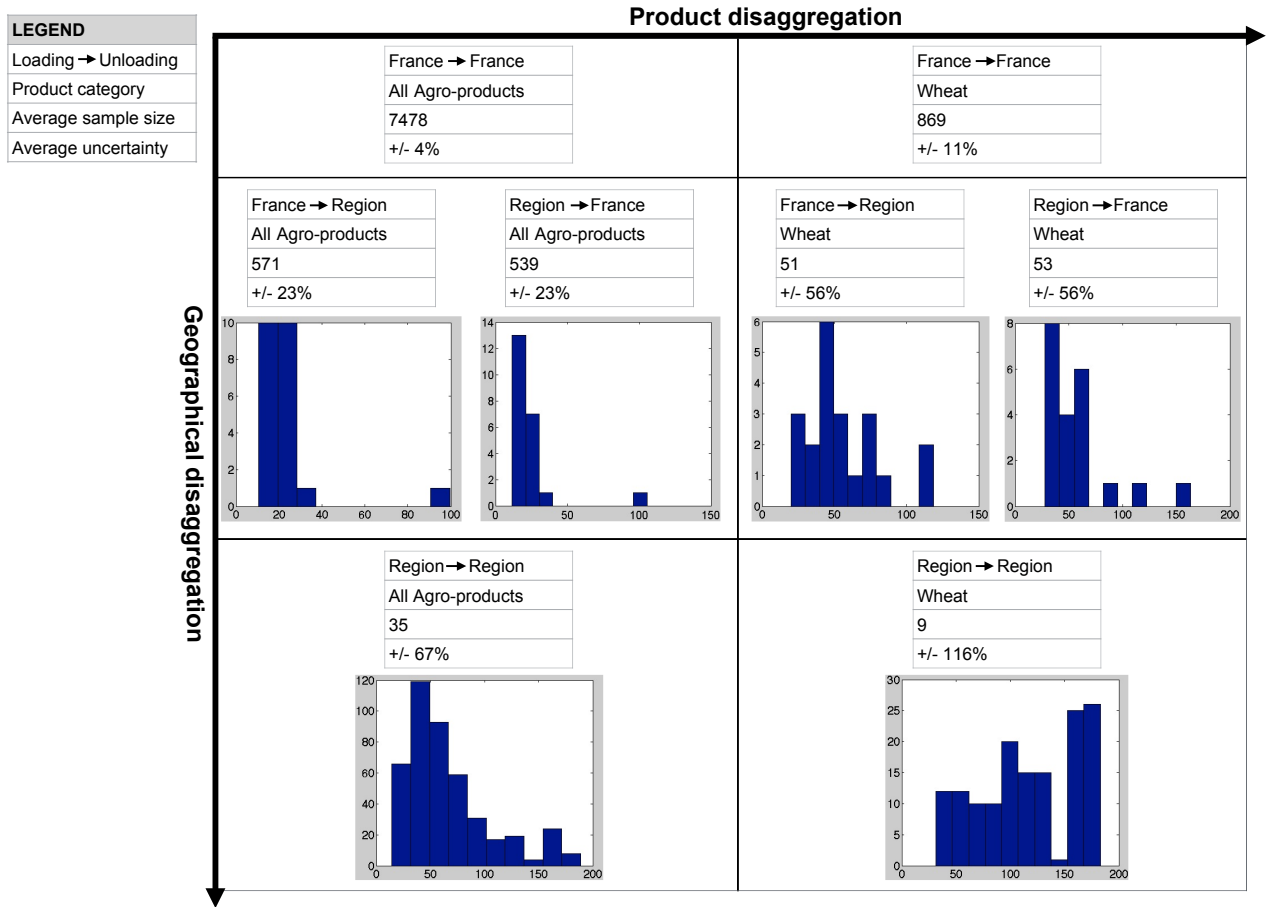


Figure 2: Analysis of uncertainties of road freight at various spatial and product disaggregation levels. Distributions are shown, with the  $x$  axis representing uncertainties computed for the year 2010 and the  $y$  axis representing the number of occurrences. Distributions are not drawn for total domestic freight (loading and unloading in France, whatever the region) since there is only one occurrence per product category. Uncertainties above 100% correspond to non-significant extrapolations. Flows between French départements (nuts 3 levels) are not shown but display an even larger level of uncertainty.

$Z_{ij}$  the exports of country  $j$  to country  $i$ );  $\mathbb{1}$  a vector of ones so that  $Z\mathbb{1}$  is the vector of total imports for each region; and  $\hat{x}^{-1}$  a diagonal matrix whose  $i$ th diagonal elements is  $x_i^{-1}$ . Note that term  $R_{ij}$  takes into account an infinite regression of imports and exports between all regions (e.g., region  $j$  exports to region  $k$  which in turn exports partly to region  $i$  while importing its own supply from other regions, etc), under an assumption of perfect blend<sup>11</sup>.

For the study of uncertainty propagation we use Monte-Carlo simulations and only consider uncertainties of domestic road freight statistics<sup>12</sup>. For each simulation, road freight data (symbolized by the  $r$  superscript on matrix  $Z$ ) are generated and satisfy the four following constraints:

- all terms must remain positive,
- random draws of trade between two regions are generated from a Gaussian distribution centered on the estimated value and within the 99% confidence interval (3 sigmas).
- total imports and total exports of a region computed with the random draws of inter-regional trade must be within the 99% confidence interval of the estimated total imports and total exports
- total domestic trade computed with the random draws of inter-regional trade must be within the 99% confidence interval of the estimated total domestic trade.

The distributions mean value are extracted from the TRM survey extrapolated values and the confidence interval from our error analysis.

These constraints are necessary due to the fact that aggregated data is more reliable than disaggregated data. They translate into the following inequalities:

- For each couple of regions  $(i, j)$ ,  $Z_{ij}^r \geq 0$ , with  $Z^r$  the random draw of road freight,
- $|Z_{ij}^r - Z_{0ij}^r| \leq 3\sigma_{ij}^{(0)}$  with  $Z_0^r$  the initial road freight data, of estimated standard deviation  $\sigma^{(0)}$ ,
- $|(Z^r\mathbb{1})_i - (Z_0^r\mathbb{1})_i| \leq 3\sigma_i^{(1)}$ ,
- $|(\mathbb{1}'Z^r)_j - (\mathbb{1}'Z_0^r)_j| \leq 3\sigma_j^{(2)}$ ,
- $|\mathbb{1}'Z^r\mathbb{1} - \mathbb{1}'Z_0^r\mathbb{1}| \leq 3\sigma^{(3)}$ .

Here,  $\sigma^{(0)}$  is a matrix,  $\sigma^{(1)}$  and  $\sigma^{(2)}$  are vectors and  $\sigma^{(3)}$  is a number.

Figure 3 shows the supply areas for wheat of two French region (Haute-Normandie and Alsace), chosen for their high level of inter-regional imports. These two examples show that the supply uncertainty is often but not always small: for instance, the supply of Alsace coming from Franche-Comté is estimated inside of a wide 2%-22% interval. Turning to the analysis of all regions, we find that the average uncertainty for regions contributing to more than 5% to another region's supply is equal to 28% (the median is 18%) but the dispersion is quite high. Again, this suggests that uncertainty evaluation for input data as well as for output results should be generalized as emphasized by Laner et al. (2014).

<sup>11</sup>This hypothesis states that the proportion of production going to local consumption (which is unknown) is the same as the proportion of total supply going to local consumption (which is known). The same applies to the proportion of production directed to exports. This assumes constant stocks on the long run.

<sup>12</sup>This means that uncertainties of production, international trade and domestic trade by other modes of transport are not considered. It is however very likely that domestic road freight data are the less precise.

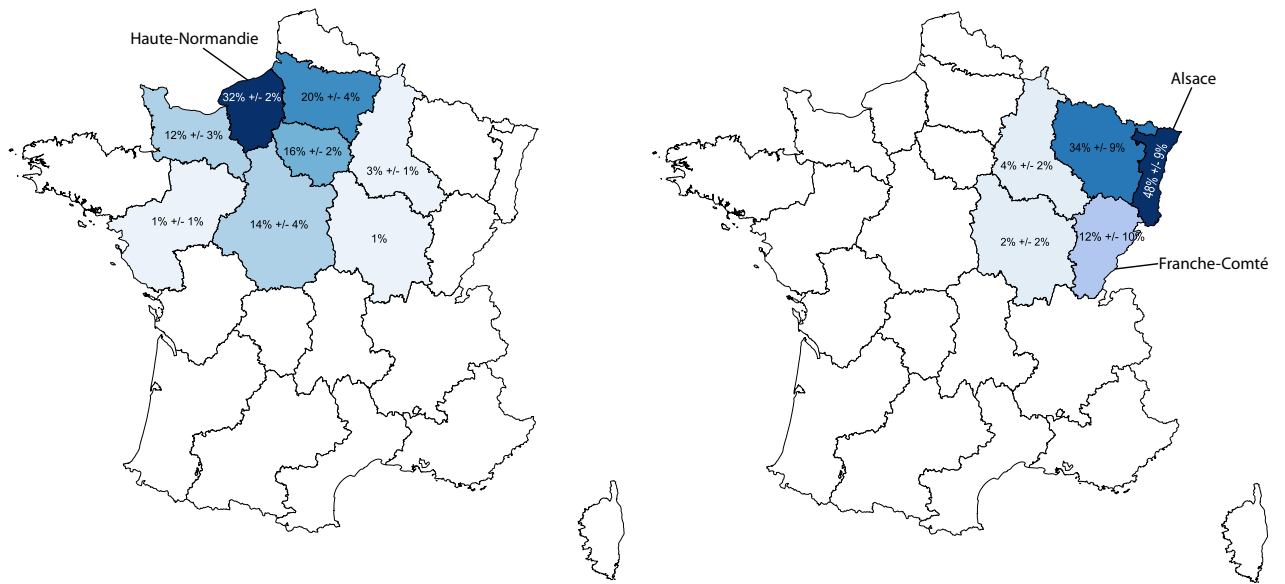


Figure 3: Wheat supply areas of region Haute-Normandie (left) and Alsace (right). Note that for each region  $x$ , the first percentage gives the fraction of supply of region  $y$  (Haute-Normandie or Alsace) produced in region  $x$ . 95% Confidence intervals are shown in the second percentage and are expressed in absolute value (which is a percentage since the quantity of reference is a percentage). Most of the uncertainties do not twist the conclusions that could be made based on central values alone, except for the case of supply of Alsace originating from Franche-Comté which can in fact vary between 2% and 22%.

## 4.4 Comparison between the French, Swedish and German surveys

It is interesting to compare freight error analysis between countries, when possible. For Sweden, following Patricio et al. (2015), we retrieved data from the Swedish statistics office and built table 5. Note that the uncertainties presented correspond to total domestic freight (a regional resolution will have higher uncertainties as we saw previously). For Germany, the German statistics office in charge of the survey only provides extrapolations meeting minimum quality criteria: the subsample size must be above 35 and the uncertainty below 40%. It also highlights extrapolations that are based on less than 50 trucks or whose uncertainties are above 20%. The uncertainty of the total in Sweden is about four times larger than in France, the latter being itself about twice as large as in Germany. We conclude from these examples that uncertainties in European countries are likely to be comparable in order of magnitude to the ones reported here. If one wishes to use disaggregated domestic freight data for MFA purposes in European countries, we therefore advise to conduct uncertainty studies similar to the one exposed in this work.

## 5 Discussion and outlook

This article addresses the complex and technical issue of uncertainties of domestic road freight data in the context of material flow studies. Below, we start by summarizing the major quantitative results before pointing out the importance of such studies from a broader viewpoint.

### 5.1 Major quantitative results

Since uncertainties of domestic road freight data are not currently computed by the French statistics office, we have conducted our own uncertainty analysis of the sampling procedure performed in France, and proposed a model to estimate the resulting errors, based on available predictor variables, the subsample size or the number of observations. We also proposed a computation rule to assess uncertainty reduction when averaging over  $n_y$  years. Results suggest that conducting MFA studies on period of several years instead of a single year would lead to more reliable results. For instance, we can expect a 3-year average to reduce the uncertainty of road freight by 30%.

Turning to specific material flows, we illustrated the *data disaggregation effect* on the case of agro-products, and then *the propagation effect* of determining regional supply areas for wheat. Results show it is not possible to assert up to which disaggregation level the data can be exploited: computing uncertainties should be done on every flow studied (which justifies the need for reliable and easy-to-implement computation rules). We show that the subsample size (or the number of observations) is a much better predictor than the total tonnes estimated, that is, one should not directly assume that large quantities are necessarily more reliable than smaller ones.

We focused on the French case for definiteness but the methodology developed in this paper should be of more general validity. Although specific results may differ, many trends are likely to be valid in other European countries since these countries make use of the same kind of surveys for assessing road freight. The fact robustness increases significantly when using time period averages is typically a result that will be true in other countries.

### 5.2 Broader outlook

Estimating uncertainties of MFA data is important for at least three reasons. First, as for any scientific endeavor, MFA results can not be properly interpreted without providing information on their reliability.

Product category	Uncertainty of the 2010 French survey	Uncertainty of the 2010 Swedish survey	Uncertainty of the 2013 German survey
01 - Products of agriculture, hunting, and forestry; fish and other fishing products	4%	10%	< 20%
02 - Coal and lignite; crude petroleum and natural gas	18%	114%	20% < e ≤ 40% or 35 < # trucks ≤ 50
03 - Metal ores and other mining and quarrying products; peat; uranium and thorium ores	4%	16%	< 20%
04 - Food products, beverages and tobacco	4%	13%	< 20%
05 - Textiles and textile products; leather and leather products	22%	56%	< 20%
06 - Wood and products of wood and cork (except furniture); articles of straw and plaiting materials; pulp, paper and paper products; printed matter and recorded media	7%	19%	< 20%
07 - Coke and refined petroleum products	7%	30%	< 20%
08 - Chemicals, chemical products, and man-made fibers; rubber and plastic products; nuclear fuel	8%	25%	< 20%
09 - Other non-metallic mineral products	4%	24%	< 20%
10 - Basic metals; fabricated metal products, except machinery and equipment	10%	21%	< 20%
11 - Machinery and equipment n.e.c.; office machinery and computers; electrical machinery and apparatus n.e.c.; radio, television and communication equipment and apparatus; medical, precision and optical instruments; watches and clocks	8%	29%	< 20%
12 - Transport equipment	10%	31%	< 20%
13 - Furniture; other manufactured goods n.e.c.	12%	62%	< 20%
14 - Secondary raw materials; municipal wastes and other wastes	8%	26%	< 20%
15 - Mail, parcels	9%	21%	< 20%
16 - Equipment and material utilized in the transport of goods	14%	15%	< 20%
17 - Goods moved in the course of household and office removals; baggage and articles accompanying travelers; motor vehicles being moved for repair; other non-market goods n.e.c.	15%	43%	< 20%
18 - Grouped goods: a mixture of types of goods which are transported together	4%	8%	< 20%
19 - Unidentifiable goods: goods which for any reason cannot be identified and therefore cannot be assigned to groups 01-16	17%	-	< 20%
20 - Other goods n.e.c.	55%	39%	e > 40% or # trucks ≤ 35
TOTAL (all product categories)	1.5%	6%	0.8%

Table 5: Comparison of the 95% uncertainties in the 2010 French and Swedish road freight surveys and the 2013 German survey for the main categories of products (NST 2007 classification). Uncertainties of the French survey are computed by the authors as described in methods section, except for the total which is provided in CGDD-SOeS (2011). Uncertainties of the Swedish survey are computed based on table 7A of Swedish-Transport-Administration (2011). Uncertainties of the German survey were taken from the table p. 60 of Kraftfahr-tBundesamt (2014), except for the total which was directly provided to the authors by the statistics office.

Conversely, uncertainties are useful to analyze the level of detail researchers can currently aim at, that is, given existing statistical data. For instance, *can we model material flow down to the département level in a satisfactory way, or should we only focus on the regional level? Will results be reliable enough if we study flows of durum wheat? Or should we rather study all wheat flows? Or aggregated cereal flows?* Of course it is often the case that too much aggregation leads to less interpretable results; knowledge about uncertainties can help set the cursor right. Finally, without any robustness assessment, results are unlikely to be used to guide policy making even if the original research was action oriented.

With the large amounts of data and computational power available nowadays, it is tantalizing to elaborate evermore sophisticated tools without taking the time to question to what extent they provide useful new contributions to our knowledge and understanding. Analyzing uncertainties helps putting such results in perspective, avoiding premature conclusions, pointing to current lack of information, in line with this famous quote from Read (1920): *“it is better to be vaguely right than exactly wrong”*.

## Acknowledgments

This research was funded by grants from Inria and Artelia Eau & Environnement. The authors would like to thank the freight statistics team from the French Ministry of Ecology, and especially Mr Mahmoud Jlassi, for their kind collaboration, which was essential to the study.

## References

- Alterre-Bourgogne (2013). La bourgogne comptabilise ses flux de matières. *Repères* 64, 1–12.
- Billen, G., S. Barles, J. Garnier, J. Rouillard, and P. Benoit (2009). The food-print of paris: long-term reconstruction of the nitrogen flows imported into the city from its rural hinterland. *Regional Environmental Change* 9, 13–24.
- Binder, C. R., C. Hofer, A. Wiek, and R. W. Scholz (2004). Transition towards improved regional wood flows by integrating material flux analysis and agent analysis: the case of Appenzell Ausserrhoden, Switzerland. *Ecological Economics* 49, 1–17.
- Binder, C. R., E. Van Der Voet, and K. S. Rosselot (2009). Implementing the results of material flow analysis. *Journal of Industrial Ecology* 13, 643–649.
- Browne, D., B. O’Regan, and R. Moles (2011). Material flow accounting in an irish city-region 1992?2002. *Journal of Cleaner Production* 19(9?10), 967 – 976.
- Cencic, O. and H. Rechberger (2008). Material flow analysis with software STAN. *Journal of Environmental Engineering Management* 18, 3–7.
- CGDD (2014). Comptabilité des flux de matières dans les régions et les départements: Guide méthodologique.
- CGDD-SOeS (2011). Fiche qualité de l’enquête transport routier de marchandises 2010.
- Courtonne, J.-Y., J. Alapetite, P.-Y. Longaretti, D. Dupré, and E. Prados (2015). Downscaling material flow analysis: The case of the cereal supply chain in france. *Ecological Economics* 118, 67 – 80.

- Courtonne, J.-Y., P.-Y. Longaretti, J. Alapetite, and D. Dupré (2016). Environmental pressures embodied in the french cereal supply chain. *Journal of Industrial Ecology*.
- Danius, L. (2002). Data uncertainties in material flow analysis. *KTH, Royal Institute of Technology, Sweden*.
- Dubois, D., H. Fargier, M. Ababou, and D. Guyonnet (2014). A fuzzy constraint-based approach to data reconciliation in material flow analysis. *International Journal of General Systems* 43(8), 787–809.
- Eurostat (2001). Economy-wide material flow accounts and derives indicators: a methodological guide.
- Graedel, T., D. van Beers, M. Bertram, K. Fuse, R. B. Gordon, A. Gritsinin, A. Kapur, R. J. Klee, R. J. Lifset, L. Memon, H. Rechberger, S. Spatari, and D. Vexler (2004). Multilevel cycle of anthropogenic copper. *Environmental Science & Technology* 38, 1242–1252.
- Hedbrant, J. and L. Sörme (2001). Data vagueness and uncertainties in urban heavy-metal data collection. *Water, Air and Soil Pollution: Focus* 1, 43–53.
- Kastner, T., M. Kastner, and S. Nonhebel (2011). Tracing distant environmental impacts of agricultural products from a consumer perspective. *Ecological Economics* 70, 1032 – 1040.
- Kovanda, J., J. Weinzettel, and T. Hak (2009). Analysis of regional material flows: The case of the czech republic. *Resources, Conservation and Recycling* 53, 243–254.
- Kraftfahr-tBundesamt (2014). Verkehr deutscher Lastkraftfahrzeuge (VD) - Inlandsverkehr Jahr 2013 - VD 3.
- Laner, D., H. Rechberger, and T. Astrup (2014). Systematic evaluation of uncertainty in material flow analysis. *Journal of Industrial Ecology* 18, 859–870.
- Moore, J., M. Kissinger, and W. E. Rees (2013). An urban metabolism and ecological footprint assessment of metro vancouver. *Journal of Environmental Management* 124, 51 – 61.
- Niza, S., L. Rosado, and P. Ferrão (2009). Urban metabolism. *Journal of Industrial Ecology* 13, 384–405.
- Patricio, J., Y. Kalmykova, L. Rosado, and V. Lisovskaja (2015). Uncertainty in material flow analysis indicators at different spatial levels. *Journal of Industrial Ecology* 19, 837–852.
- Read, C. (1920). *Logic: Deductive and Inductive*. Public Domain.
- Rechberger, H., O. Cencic, and R. Frühwirth (2014). Uncertainty in material flow analysis. *Journal of Industrial Ecology* 18, 159–160.
- Swedish-Transport-Administration (2011). Lastbilstrafik 2010, Swedish national and international road goods transport 2010.
- Theobald, T. F., M. Schipper, and J. Kern (2016). Phosphorus flows in berlin-brandenburg, a regional flow analysis. *Resources, Conservation and Recycling* 112, 1 – 14.
- Weidema, B. P. and M. S. Wesnæs (1996). Data quality management for life cycle inventories: an example of using data quality indicators. *Journal of Cleaner Production* 4, 167 – 174.

## **About the Authors**

Jean-Yves Courtonne is PhD student in the STEEP team of INRIA, Grenoble, France. He is working under the joint supervision of Pierre-Yves Longaretti and Denis Dupré. Pierre-Yves Longaretti is CNRS researcher at IPAG, Joseph Fourier University, Grenoble, France and also member of the STEEP team INRIA. Denis Dupré is professeur at CERAG, Pierre Mendès-France University, Grenoble, France and has recently joined the STEEP team.

## **Corresponding Author**

Jean-Yves Courtonne, INRIA, 655 Avenue de l'Europe, 38033, Montbonnot Saint-Martin, France; email: [jean-yves.courtonne@inria.fr](mailto:jean-yves.courtonne@inria.fr)

*Supplementary material*

# Uncertainties of domestic road freight statistics: insights for regional material flows studies

Jean-Yves Courtonne, Pierre-Yves Longaretti, Denis Dupré

**The document is structured as follow:**

- Section 1 presents and explains the relevant material and results from survey sampling theory that are used in the main text.
- Section 2 is an illustration of data treatment that shows how combinations of aggregates are generated based on basic extrapolations,
- Sections 3 is an annex to the *theoretical results* section of the article, dedicated to model uncertainties for a single year based two other explanatory variables (the number of observations and the tonnes transported),
- Section 4 is also an annex to the *theoretical results* section of the article, but dedicated to the study of uncertainty reduction when averaging over several years: it explains how independent multi-year samples are built and illustrates the results on a dedicated plot.

## **1 Unequal probability survey sampling in a nutshell**

The French road freight survey (TRM survey) is statistical in nature, and makes use of a specific unequal probability sampling method without replacement to estimate totals of interest, such as the total number of tonnes, kilometers travelled or tonnes-kilometers transported by road from any given origin to any given destinations and category of goods.

The TRM survey also makes use of constraints from auxiliary quantities of known totals to improve the desired total estimate, but this form of constraint is ignored here, as it provides only a modest correction to the variance in most cases and as its implementation requires information we do not have. However a simple correction for non-response is included.

The TRM survey only provides the error on the aggregated total (for all origins, destinations and categories) while we are interested in the error on the various disaggregated totals. The object of the present discussion is therefore to provide the reader with relevant results and explanations from survey sampling theory in order to perform an estimation of the errors of these various disaggregated totals. It is also intended to provide a minimal guideline to the relevant literature; as a consequence, literature citations are focused on the few books and articles most directly related to the problem at hand.

The following subsection is devoted to introducing important definitions and notations; some of these are already given in the main text, but reproduced here, sometimes in more detail, for the reader's convenience.

The next one will give relevant results for totals and their variances for samplings without replacement. The last one will provide a heuristic justification for the approximation we use in this work. All relations are demonstrated (unless otherwise stated) for the interested reader to get a feel of what these relations are about, and as the technical literature has a mathematical flavor that can be daunting for the non-mathematician.

## 1.1 Definitions and notations

One considers a population of  $N$  elements (here vehicles-weeks) denoted  $i$  ( $1 \leq i \leq N$ ). Samples of fixed size  $S$  are drawn from this distribution. Typically in the problem at hand,  $N \simeq 5 \times 10^5$  and  $S \simeq 8 \times 10^4$ , while the number of usable answers is  $\simeq 5 \times 10^4$  (this last figure makes non-response — which includes all unusable answers — non negligible and a correction for this is generally applied).

Several quantities need to be defined to specify what is meant by unequal probability sampling without replacement.

**Samples.** Samples are designated by  $s = (s_1, s_2, \dots, s_N)$  where  $s_i$  is the number of times a given element  $i$  is present in the sample. The set of all such samples is denoted  $F$ . A sample without replacement has  $s_i = 0$  or  $1$ , by definition (an element cannot be selected more than once). A sample with replacement has  $0 \leq s_i \leq S$  where  $S$  is the size of the sample. Samples of fixed, predetermined size are the only ones of interest here; then, by construction,  $\sum_{i=1}^N s_i = S$ .

**Sample design.** The sample design is the probability distribution  $p(s)$  defined on the set of samples  $F$ . As a consequence, both  $s_i$  and the inclusion indicators  $\delta_i$  defined right below are probabilistic quantities<sup>1</sup>. Note that in this form of probabilistic description, the elements of the populations are not random variables, as the elementary probabilistic event of interest is the selection of a whole sample, not the successive selections of elements of the population.

**Inclusion indicator and inclusion probabilities.** The inclusion indicators (or inclusion variables or Cornfield variables)  $\delta_i$  are defined by  $\delta_i = 0$  if  $s_i = 0$  and  $\delta_i = 1$  otherwise ( $s_i \geq 1$ ), i.e., an inclusion indicator measures the presence of its associated element in the sample and ignores its possible repetitions.

The first-order inclusion probability  $\pi_i$  relates to element  $i$  and defines its probability of being present in the sample. By construction the probability of absence is  $\bar{\pi}_i = 1 - \pi_i$ . By definition of the sampling design and inclusion indicator:

$$\pi_i = \sum_{s \in F} p(s) \delta_i. \quad (1)$$

Similarly, the probability that both  $i$  and  $j$  are present in the sample is the second-order probability  $\pi_{ij} = \sum_s p(s) \delta_i \delta_j$ . By definition,  $\pi_{ii} = \pi_i$ . For fixed-size and non-replacement sampling designs,  $\sum_{i=1}^N \pi_i = n$ , from Eq. (1).

**Unequal probability sample designs.** Unequal probability sampling designs are characterized by the fact  $\pi_i$  is not constant (not independent of  $i$ ).

---

<sup>1</sup>No distinction of notation is made here between a random variable and its associated realizations.

**Characteristic quantity.** This refers to any quantity of interest on the whole population, for example the quantity transported per year (tonnes) corresponding to specific criteria (loading and unloading locations, category of product transported). Quantities of interest are denoted  $y$  in a generic way. The most important quantity one tries to estimate through survey sampling is the sum total of  $y$  on the whole population

$$T = \sum_{i=1}^N y_i. \quad (2)$$

In the probabilistic description adopted here,  $y_i$  is a parameter, not a random variable.

**Subpopulation index and total.** We are mostly interested in evaluating totals on well-defined subpopulations, identified by a given origin ( $o$ ), destination ( $d$ ) and category of goods ( $c$ ) transported. These subpopulations are symbolically represented by an index  $\delta_i^c$ , taking two values:  $\delta_i^c = 1$  if  $i$  belongs to the desired  $c = (o, d, c)$  subpopulation, and  $\delta_i^c = 0$  otherwise.

It must be noted that the index  $\delta_i^c$  is not a random variable. It is a fixed parameter allowing us to sort out the relevant elements in the total population or in a sample. In particular, we do not redraw a subsample of the considered sample within this category, we just post-select the relevant elements within the drawn sample.

The associated total is

$$T^c = \sum_{i \in c} y_i = \sum_{i=1}^N \delta_i^c y_i. \quad (3)$$

**Subsamples.** The members of a subpopulation belonging to a given sample are called a subsample here, although they result from a post-selection process and not a second sampling stage. A subsample size is denoted  $n$ .

**Sampling algorithms.** A sample design is specified by its probability distribution  $p(s)$  over the sample set  $F$ , and is constructed algorithmically. For the simple equal probability or replacement cases, the first and second order inclusion probabilities can be specified analytically.

In sampling designs without replacement, elements are not reinserted in the original population if selected in one of the successive steps of the sampling algorithm; on the contrary, a given element can be selected any number of times in a sampling design with replacement. Sampling designs without replacement are more complex to implement but their interest comes from the following theorem (see Tillé 2006, section 2.18): sampling designs with replacement are suboptimal in the sense that designs with smaller variance (i.e., more precise, an essential property) and identical first-order inclusion probabilities<sup>2</sup> can always be found.

Unequal probabilities are useful when the quantity of interest is not evenly distributed in the population. In particular, if one can find an auxiliary quantity  $X_i$  that is known to be (at least approximately) correlated to  $y_i$ , one can substantially improve the quality of the estimate of the total  $T$  at constant sample size  $S$  by designing the probability to be proportional to  $X$ . For the TRM survey, the first-order inclusion probabilities are proportional to the vehicle's age and payload, with a correction factor for some specific categories of vehicles. The specification of the first order inclusion probabilities is the first step in the algorithmic construction of a sample design. For more details, see section 2.10 of Tillé 2006.

---

<sup>2</sup>Second-order inclusion probabilities are necessarily different.

There is an infinite number of sampling designs with given first-order inclusion probabilities, and sampling design algorithms are selected inasmuch as possible to have desirable properties for the second-order inclusion probabilities  $\pi_{ij}$ . Quite often indeed, the algorithmic evaluation of these quantities turns out to be unpractical for large samples, or to possess undesirable properties making the variance strongly unreliable or even negative. For example, the ordered systematic sampling design (Tillé 2006, section 7.1) is a widely used algorithm due to its simplicity (it is used in the TRM survey), but has the undesirable feature to produce a (large) number of vanishing second-order inclusion probabilities. As a result, the usual expression of the Horvitz-Thomson variance Eq. (5) below gives biased results and cannot be used, and an approximate expression must be found by other means.

Finding practical sampling algorithms with simple-to-evaluate and well-behaved  $\pi_{ij}$  or finding useful approximations of the variance based only on first-order inclusion probabilities for known algorithms is still an active area of research in this field (see, e.g., chapters 3 and 7 of Tillé 2006 for more details).

More complex sampling designs can be adopted besides unequal probability sampling without replacement, e.g. stratified or clustered sampling, but these methods will not be discussed here (see Cochran 1977 or Fuller 2009 for more details).

## 1.2 The Horvitz-Thomson estimator and its variance

### 1.2.1 Generic relations

The quantity of interest is the total of any characteristic quantity, Eq. (2). The best known and most used estimator is the Horvitz-Thomson one, defined by (Tillé 2006, section 2.17)

$$\widehat{T} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i=1}^N \delta_i \frac{Y_i}{\pi_i}. \quad (4)$$

An estimator is unbiased if its expectation value is equal to the desired quantity. Eq. (4) shows that the expectation value of the HT estimator over the sample space is unbiased. Indeed,  $E_s(\widehat{T}) = \sum_s \widehat{T} p(s) = T$ , from Eq. (1), where  $E_s(X)$  is the expectation of  $X$  over the sample probability distribution.

The variance  $V_s(\widehat{T})$  of the Horvitz-Thomson estimator follows from its definition, and a straightforward calculation gives:

$$V_s(\widehat{T}) = \sum_{i=1}^N \sum_{j=1}^N \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j). \quad (5)$$

In practice, one needs an estimator of this variance that can be evaluated on the sample at hand. As pointed out above, the main difficulty lies in the evaluation of the second-order inclusion probabilities  $\pi_{ij}$ . A common strategy to avoid the problem is to approximate the variance estimator with an expression involving only appropriate combinations of the first-order inclusion probabilities. One of the most simple approximate estimator of the variance  $\widehat{V}(\widehat{T})$  is given by (Tillé 2006, Eq. 7.20 with  $S/(S-1) \approx 1$  due to  $S \gg 1$ )

$$\widehat{V}(\widehat{T}) = \sum_{i \in s} c_i \left( \frac{y_i}{\pi_i} - \frac{\sum_{i \in s} c_i y_i / \pi_i}{\sum_{i \in s} c_i} \right)^2, \quad (6)$$

where

$$c_i = 1 - \pi_i. \quad (7)$$

A justification for this approximate variance estimator is given in section 1.3.

### 1.2.2 Relations for subpopulations

The preceding considerations are straightforwardly transposed to any subpopulation. In particular, an unbiased estimator of the total in the subpopulation, Eq. (3), is given by

$$\widehat{T}^c = \sum_{i \in s} \delta_i^c \frac{y_i}{\pi_i} = \sum_{i=1}^N \delta_i^c \delta_i \frac{y_i}{\pi_i}. \quad (8)$$

One can check that  $E_s(\widehat{T}^c) = T^c$ ; this follows from the fact that  $\delta_c$  is not a sampling random variable, but simply a selection or post-selection one. However, the resulting estimator will be precise enough only if the inclusion probabilities are sufficiently correlated with the characteristic quantity in the subpopulation.

Similarly, an estimator of the variance of this quantity is provided by

$$\widehat{V}(\widehat{T}^c) = \sum_{i \in s} c_i \left( \delta_i^c \frac{y_i}{\pi_i} - \frac{\sum_{i \in s} c_i \delta_i^c y_i / \pi_i}{\sum_{i \in s} c_i} \right)^2, \quad (9)$$

where  $c_i$  is given in Eq. (7). This can be seen by a formal substitution of  $y_i$  with  $y_i \delta_c$  in Eq. (6).

Neglecting the  $n/(n-1)$  factor of Eq. (7.20) of Tillé (2006) in Eq. (7) is not justified for very small subsample sizes  $n$ , but in such cases, the variance is large anyway, and the total estimator unreliable.

### 1.2.3 Rescaling

**Weeks vs year** The quantities of interest are, e.g., tonnes transported per year, or kilometers travelled per year. On the other hand, our population is made of vehicles-weeks, as the survey selects any vehicle for a single week in the year. As a consequence, Eqs. (4) and (8) do not give totals per year, but totals per week. On the other hand, the “week” in question is not any particular week of the year, as polls are conducted every single week. It is in effect some year average week total, inasmuch as the sampling procedure is homogeneous throughout the year. As a consequence, Eqs. (4) and (8) should be multiplied by the number of weeks  $N_w = 52$ .

Because the Horvitz-Thomson is linear, this procedure is obviously equivalent to rescaling the quantities of interest  $y$  as

$$y'_i = N_w y_i. \quad (10)$$

The resulting Horvitz-Thomson estimators Eqs. (4) and (8) have the same formal expression after this substitution, so that the variance will also be the same except for the substitution of  $y'$  to  $y$  in Eqs. (5), (6) and (9). Note however that this substitution does not imply that for any given truck, the total load transported in the year is  $N_w$  times what is transported in the poll week, due to seasonal variations; it just means that this is a reasonable extrapolation on average over all vehicles-weeks.

**Nonresponse** A similar correction is performed to account for nonresponse. Denoting  $S_{nr}$  the number of usable forms, nonresponse is taken into account by substituting a scaled first order inclusion probability to  $\pi_i$ :  $\pi'_i = S \pi_i / S_{nr}$ . A better correction is obtained by performing the same correction by category (origin, destination, product transported) except maybe for the categories with very small number of elements. The inclusion probabilities provided by the public organization (SOeS) in charge of the TRM database are already corrected for nonresponse in this way and the inclusion probabilities used in all formulae are in fact the primed ones, not the original ones, but we have not made the distinction in notation, for simplicity.

### 1.3 Elements of justification of the adopted estimator of the variance

It is possible to provide some justification of Eqs. (6), (7) and (9). This will be done in this section with the help of the following reasoning:

1. First, an unequal probability sampling design without replacement is approximated by an unequal probability sampling design with replacement characterized by the same first-order inclusion probabilities. The usefulness of this approximation comes from the existence of analytic expressions for the total estimator and its variance in sampling designs with replacement.
2. However, it is known that the resulting variance is not optimal (it is generically overestimated). This will be explicitly shown by taking the limit of a simple random sampling design (equal probability) where the sampling with and without replacement can be analytically compared.
3. One can estimate the error introduced by the sampling design substitution procedure by computing the average number of occurrences of any element in a sample. This will allow us to estimate that on average, the number of elements that are multiply present in the equivalent design with replacement is about 1/3 the total size of the sample in the case of the TRM survey. This indicates that the overestimation of the variance is most probably not negligible, so that a tighter estimator would be a welcome improvement.
4. A correction is applied through a generalization of the expression of the variance estimator, subject to the constraint that the correct variance must be formally recovered for a simple random sampling design without replacement. The corrected estimator is known to still overestimate the variance for the ordered systematic sampling design, though, so that we still err on the side of safety.

#### 1.3.1 First step: approximating a sample design with replacement by a sample design without replacement

Let us define a sample design with replacement that has the same population (size  $N$ ), same sample size  $S$ , and same first order inclusion probabilities  $\pi_i$  as the sample design without replacement of interest. The simplest such sampling design makes use of the probability  $p_i$  that element  $i$  has to be drawn in the  $S$  successive draws that constitute the sample. The probability  $p_i$  is directly related to the inclusion probability  $\pi_i$  from the probability of non inclusion in  $S$  successive draws  $\bar{\pi}_i = (1 - \pi_i)^S$ :

$$\pi_i = 1 - (1 - p_i)^S = S p_i - \frac{(S p_i)^2}{2} + \mathcal{O}(S^3 p_i^3). \quad (11)$$

This relation can be inverted for future use:

$$S p_i = \pi_i + \frac{\pi_i^2}{2} + \mathcal{O}(\pi_i^3). \quad (12)$$

The totals of interest in the population can be estimated on the sample with the help of the Hansen-Hurwitz estimator (Tillé 2006, section 2.16)

$$\widehat{T} = \frac{1}{S} \sum_{i \in s} \frac{y_i}{p_i}. \quad (13)$$

This estimator is unbiased. To see this, consider first a single term in the sum:  $x_i = y_i/S p_i$ ; every term of this kind constitutes a realization of the random variable  $X = Y/S p(Y)$ . Its expectation value on the population is

$$E(X) = \sum_{k=1}^N p_k \frac{y_k}{S p_k} = \sum_{k=1}^N \frac{y_k}{S} = \frac{T}{S}. \quad (14)$$

The variance of  $X$  is readily computed :

$$V(X) = \sum_{k=1}^N p_k \left( \frac{y_k}{S p_k} - \frac{T}{S} \right)^2 = \frac{1}{S^2} \left[ \left( \sum_{k=1}^N \frac{y_k^2}{p_k} \right) - T^2 \right] \quad (15)$$

Now,  $\widehat{T}$  is the sum of  $S$  identical random variables identical to  $X$ . The expectation value and variance of the estimator are therefore the sum of the expectation value and variance of  $X$ , leading to

$$E(\widehat{T}) = S E(X) = T \quad (16)$$

$$V(\widehat{T}) = S V(X) = \frac{1}{S} \left[ \left( \sum_{k=1}^N \frac{y_k^2}{p_k} \right) - T^2 \right]. \quad (17)$$

Note that the expectation value and variance  $E$  and  $V$  are computed on the population, and not on the sample space, contrarily to  $E_s$  and  $V_s$ .

From Eq. (17), one obtains an unbiased estimator of the variance:

$$\widehat{V}(\widehat{T}) = \frac{1}{S(S-1)} \sum_{i \in s} \left( \frac{y_i}{p_i} - \widehat{T} \right)^2. \quad (18)$$

One can check that this estimator is unbiased from the following intermediate identity, which can be verified by direct calculus:

$$\widehat{V}(\widehat{T}) = \frac{1}{S(S-1)} \sum_{i \in s} \left( \frac{y_i}{p_i} - T \right)^2 - \frac{1}{S-1} (\widehat{T} - T)^2, \quad (19)$$

from which one obtains:

$$E(\widehat{V}(\widehat{T})) = \frac{1}{S-1} S^2 V(X) - \frac{1}{S-1} V(\widehat{T}) = V(\widehat{T}). \quad (20)$$

Note that as  $\pi_i \simeq S p_i$ , the Hansen-Hurwitz estimator is closely related to the Horvitz-Thompson one:

$$\widehat{T} = \sum_{i \in s} \frac{y_i}{\pi_i}, \quad (21)$$

Furthermore, for large enough  $S$ ,

$$\widehat{V}(\widehat{T}) \approx \sum_{i \in s} \left( \frac{y_i}{\pi_i} - \frac{\sum_{j \in s} y_j / \pi_j}{S} \right)^2 = \sum_{i \in s} \left( \frac{y_i}{\pi_i} - \left\langle \frac{y_i}{\pi_i} \right\rangle_s \right)^2, \quad (22)$$

where  $\langle \cdot \rangle_s$  is the mean observed on the sample. Note that this relation is of the generic form Eq. (6) with  $c_i = 1$ .

### 1.3.2 Interlude: simple random samplings with and without replacement

It is useful to see why Eq. (22) overestimates the variance of a sampling design without replacement. This is most easily seen by considering sampling designs with uniform probability (aka simple random sampling designs), for which all relevant relations can be derived analytically.

Let us examine first a uniform probability sampling design without replacement. In this case, and denoting  $C_p^q = p!/q!(p-q)!$  the binomial coefficient,

$$\pi_i = \frac{C_{N-1}^{S-1}}{C_N^S} = \frac{S}{N} \equiv \tau, \quad (23)$$

$$\pi_{ii}^{wor} = \pi_i, \quad (24)$$

$$\pi_{ij}^{wor} = \frac{C_{N-2}^{S-2}}{C_N^S} = \frac{S(S-1)}{N(N-1)}, \quad (25)$$

where the superscript *wor* refers to “without replacement”. Defining  $\bar{y} = \sum_{i=1}^N y_i/N$  and  $\sigma^2(y) = \sum_{i=1}^N (y_i - \bar{y})^2/N$ , Eq. (5) yields

$$V_s(\widehat{T}) = \frac{N^2}{S} \frac{1-\tau}{1-1/N} \sigma^2(y) \approx \frac{N^2}{S} (1-\tau) \sigma^2(y). \quad (26)$$

Similarly, defining  $\langle y \rangle = \sum_{i=1}^S y_i/S$  and  $\tilde{\sigma}^2(y) = \sum_{i=1}^N (y_i - \langle y \rangle)^2/N$ , Eq. (22) gives

$$V_s(\widehat{V}(\widehat{T})) = \frac{N^2}{S} \tilde{\sigma}^2(y). \quad (27)$$

For large enough samples,  $\langle y \rangle \approx \bar{y}$  and  $\tilde{\sigma}^2(y) \approx \sigma^2(y)$  so that the main difference between Eq. (26) and (27) is the  $1-\tau$  factor, known as the finite population correction factor. This denomination has a double justification: first, this coefficient becomes negligible as  $N \rightarrow \infty$  at constant  $S$ , and second, as shown by Eq. (25), selecting one element in the population affects the second-order inclusion probability, which is then systematically larger in sampling designs with replacement. These arguments explain why Eq. (22) also overestimates the actual variance in generic unequal probability sampling designs without replacement (see, e.g., Tillé 2006, section 2.18 and Berger 1998).

### 1.3.3 Second step: estimating the need for a correction of the replacement sampling design variance estimator

It has just been pointed out above that a potentially important issue of sampling designs with replacement is that they often lead to overestimate the variance of the population totals; corrected variance expressions have been derived in the literature to compensate for this drawback. Before looking into this question, though, it is of some interest to estimate first if such a correction is necessary in the present case; indeed if the number of multiple draws of the same element in a sample is negligible on average, there would be little interest in performing the correction.

To this effect, one must evaluate the average number of multiple draws of any given element  $i$  of the population, from which one can in turn quantify the mean number of multiple draws in samples of size  $S$ . From our notations and definitions,  $s_i$  is the number of draws of element  $i$  in such a sample; let us further define  $\langle s_i \rangle_s$ , the sample average of  $s_i$ ,  $\langle s_i \rangle_s^{>1}$  the sample average number of multiple draws of  $i$ , and  $\langle s_i \rangle_s^{\leq 1} = \langle s_i \rangle_s - \langle s_i \rangle_s^{>1}$ .

Let us also define the probability  $p_i(k)$  that  $i$  is drawn  $k$  times in a sample of size  $S$  ( $0 \leq k \leq S$ ). This probability is obtained by direct counting and reads

$$p_i(k) = C_S^k p_i^k (1 - p_i)^{S-k}, \quad (28)$$

where  $C_S^k$  is the usual binomial coefficient. Defining  $\omega_i = p_i/(1 - p_i)$ , one has

$$\begin{aligned} \langle s_i \rangle_s &= \sum_{k=0}^S C_S^k k p_i^k (1 - p_i)^{S-k} = \frac{p_i}{(1 - p_i)^{S+1}} \frac{d}{d\omega_i} \left( \sum_{k=0}^S C_S^k \omega_i^k \right) \\ &= \frac{p_i}{(1 - p_i)^{S+1}} \frac{d(1 + \omega_i)^S}{d\omega_i} = \frac{S p_i}{(1 - p_i)^{2S}}. \end{aligned} \quad (29)$$

One also finds:

$$\langle s_i \rangle_s^{\leq 1} = S p_i (1 - p_i)^{S-1}, \quad (30)$$

so that

$$\langle s_i \rangle_s^{> 1} = 3\pi_i^2 + O(\pi_i^3), \quad (31)$$

where Eq. (12) has been used.

The quantity of interest here is the population average of  $\langle s_i \rangle_s^{> 1}$ ,  $E(\langle s_i \rangle_s^{> 1})$ :

$$E(\langle s_i \rangle_s^{> 1}) = 3 \sum_{i=1}^N p_i \pi_i^2 = \frac{3N}{S} \overline{\pi_i^3}, \quad (32)$$

where  $\overline{\pi_i^3}$  is the simple population average of  $\pi_i^3$  (i.e., not weighted by the probability  $p_i$ ).

Finally, defining by  $\Delta S$  the mean (population and sample) average number of multiple draws, one has  $\Delta S = S E(\langle s_i \rangle_s^{> 1})$  so that  $E(\langle s_i \rangle_s^{> 1})$  is the relative number of multiple draws in a sample of size  $S$ . Note also that the same relation applies for a subpopulation  $N^c$  and a subsample  $n$ , *mutatis mutandis*.

A similar logic can be applied to evaluate the relative standard deviation of  $\Delta S$ :

$$\frac{\sigma(\Delta S)}{\Delta S} = \frac{\left[ \overline{\pi_i^2 (\pi_i^2 - N \overline{\pi_i^3} / S)^2} \right]^{1/2}}{N^{1/2} \overline{\pi_i^3}}. \quad (33)$$

For the TRM survey, the histogram of  $\pi_i$  is shown on Fig. 1. One also has  $N/S \simeq 6.76$  (in 2010),  $\overline{\pi_i^3} = 0.016$ ,  $E(\langle s_i \rangle_s^{> 1}) \approx 1/3$  and  $\sigma(\Delta S)/\Delta S \approx 5 \times 10^{-4}$ ; i.e., 1/3 of the sample is made of multiple draws in a sampling design with replacement, on average, and with very little variation from sample to sample. This is not a dominant fraction, but it is clearly not negligible either. As a consequence, a correction to Eq. (22) is required to tighten the variance estimates.

### 1.3.4 Third step: correcting the variance

A corrected variance estimator can be obtained by generalizing Eq. (22) under the form given in Eq. (6); a constraint is applied to the parameters  $c_i$  to specify their expression. In the case at hand, the constraint adopted is that the variance for a uniform probability sampling design should be formally identical to the variance of the equivalent design without replacement, Eq. (26).

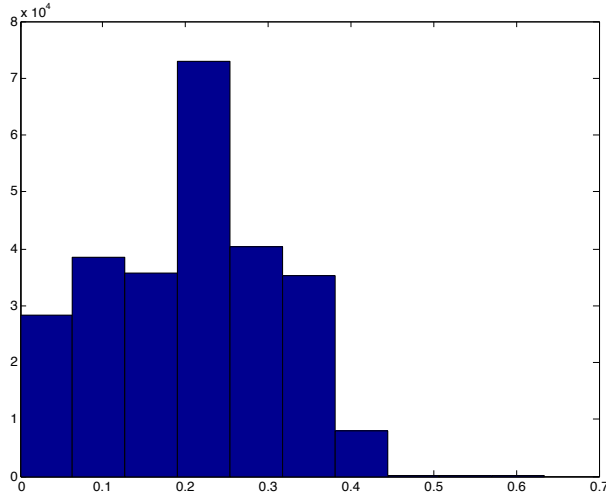


Figure 1: Histogram of the distribution of the first-order inclusion probabilities  $\pi_i$  for the TRM survey.

By construction,  $c_i$  can only depend on  $\pi_i$ :  $c_i = g(\pi_i)$ . For a simple random sampling design discussed in section 1.3.2,  $g(\pi_i) = g(\tau)$  from Eq. (23), and Eq. (6) gives

$$V_s(\widehat{V}(\widehat{T})) = \frac{N^2}{S} g(\tau) \tilde{\sigma}^2(y). \quad (34)$$

Comparison with Eq. (26) shows that the finite population correction factor will be recovered for  $c_i = 1 - \pi_i$ , i.e., Eq. (7).

The form Eq. (22) of the generalized variance is theoretically justified when the sampling design without replacement is “close” to an entropy maximizing sampling design (see Tillé 2006, section 7.5, Tillé 2001, section 6.9 and Berger 1998). This is however not the case for the ordered systematic sampling design adopted in the TRM survey. In fact, it is known that Eqs. (6) and (7) still overestimates the variance (Berger, 2003) for this sampling design, because the design makes numerous second-order inclusion probabilities equal to zero, so that a non negligible number of legitimate samples is not incorporated in the sample population. Eq. (22), on the contrary, implicitly takes into account all possible samples, because the second-order inclusion probabilities can be expressed in terms of the first-order inclusion probabilities in replacement sampling designs<sup>3</sup>, and are never equal to zero.

## 2 Building all combinations of aggregates: an illustrative example

Figure 3 shows it is possible to build 19 aggregates based on the three basic extrapolations from figure 2. Of course, the number of combinations increases when we add new basic flows, levels of products, loading or unloading areas. In the dataset used for the article, about 2.1 million aggregates were built from about 900 thousand basic flows.

<sup>3</sup>This expression is not derived here, as it does not provide useful information to further reduce the variance.

Product	Loading region	Unloading region	Tonnes	Subsample size (# trucks)
Wood	Lorraine	Alsace	30	4
Wood	Bourgogne	Alsace	5	2
Fruits	Alsace	Bourgogne	10	3

Figure 2: Example of 3 basic flows extrapolated from a survey on 9 trucks.

Product	Loading	Unloading	Tonnes	Subsample size
All	France	France	45	9
All	France	Alsace	35	6
All	France	Bourgogne	10	3
All	Lorraine	France	30	4
All	Bourgogne	France	5	2
All	Alsace	France	10	3
Wood	France	France	35	6
Fruits	France	France	10	3
All	Lorraine	Alsace	30	4
All	Bourgogne	Alsace	5	2
All	Alsace	Bourgogne	10	3
Wood	France	Alsace	35	6
Fruits	France	Bourgogne	10	3
Wood	Lorraine	France	30	4
Wood	Bourgogne	France	5	2
Fruits	Alsace	France	10	3
Wood	Lorraine	Alsace	30	4
Wood	Bourgogne	Alsace	5	2
Fruits	Alsace	Bourgogne	10	3

Figure 3: All existing combinations of aggregates based on the 3 basic flows above. In this example, we only distinguish between 2 levels of products, loading and unloading areas.

### 3 Estimating uncertainties for a single year based on explanatory variables

#### 3.1 Estimating uncertainties based on subsample sizes: details on the classes used

Subsample size (n)		Number of class members	Cumulative %	
Lower bound (lb) n > lb	Lower bound (ub) n <= ub			
	0	3	1452697	69%
	3	5	179949	78%
	5	10	182689	87%
	10	25	151391	94%
	25	50	63043	97%
	50	100	33040	98%
	100	250	20745	99.4%
	250	500	6551	99.7%
	500	1000	3508	99.9%
	1000	2000	1749	99.96%
	2000	3000	460	99.98%
	3000	5000	307	99.99%
	5000	7000	53	99.997%
	7000	10000	55	99.999%
	10000	25000	11	100%

Figure 4: Number of class members in each class of defined by the size of the subsample. These 15 classes (rows) are used in the *theoretical results* section of the article to build an explicative model.

#### 3.2 Estimating uncertainties based on the number of observations

By default, the French statistical office does not provide the size of the subsample on which each extrapolation is based. It provides instead the *number of observations* on which they are based (see glossary in the article). We use the same method as the one described in the article to propose an explicative model based on this variable:

- the sample is divided into 29 classes defined by an interval of number of observations,
- for each class, we compute the 5%, 50% (median) and 95% percentiles of uncertainty,
- we propose a model depicting the low, median, high uncertainty estimations depending on the number of observations.

Figure 5 displays the calibrated models with their equations and table 1 shows the lower, median and upper uncertainties for different number of observations.

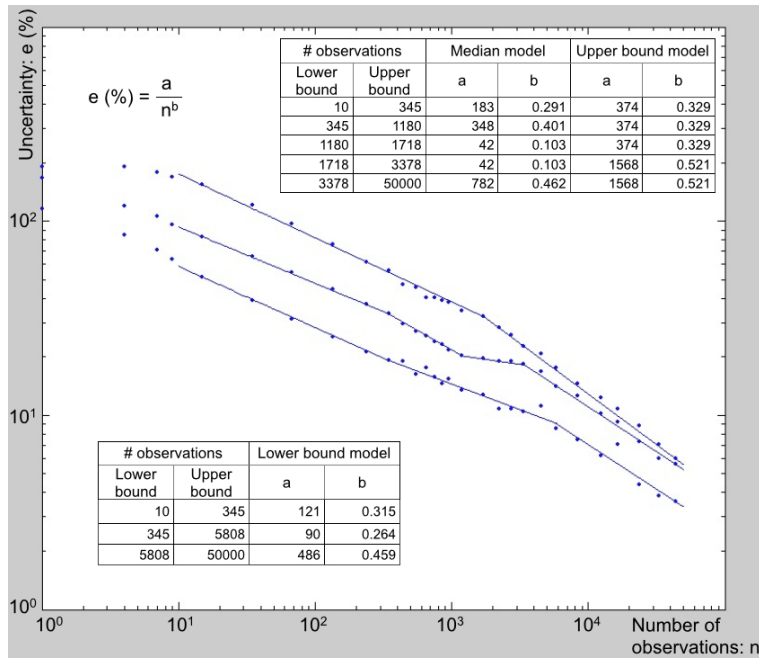


Figure 5: Evolution of lower bound, median and upper bound uncertainties depending on the number of observations (NB: this is a log-log plot). Looking at the median model, it is striking that part of the slope is almost flat between about 1200 and 3400 observations. This does not occur when the explicative variable is the size of the subsample, which is one reason why the latter is a better predictor.

Number of observations	Uncertainty (%)		
	Lower bound	Median	Upper bound
10	59	94	176
20	47	77	140
50	35	59	104
100	28	48	82
200	23	39	66
500	17	29	49
1000	14	22	39
2000	12	19	30
5000	9	15	19
10000	7	11	13
20000	5	8	9
50000	3	5	6

Table 1: Lower, median and upper uncertainty estimations for different number of observations, based on the models depicted in figure 5. We see that below 50 observations, there is risk for the statistic not to be significant (above 100% of uncertainties), although it can be significant in some cases. Above 500 observations, there is a 95% chance that the uncertainty is lower than 50%, the median estimation being 29%.

The number of observation is a predictor of lower quality than the size of the subsample, generally resulting in a wider interval between lower and upper bounds for a given class. The  $R^2$  of this model is 0.77 compared to 0.87 for the subsample size model. This should encourage the statistical office to publish the information on the sample size by default.

### 3.3 Estimating uncertainties based on the quantities (tonnes) estimated

The number of tonnes is in turn tested as explanatory variable for the uncertainty. Results are shown in figure 6. the  $R^2$  index of the median model is 0.63. We can conclude that the hypothesis “*the higher the quantity transported the smaller the uncertainty*” is only partially true. Other explanatory variables, such as the subsample size, should be preferred when it is possible.

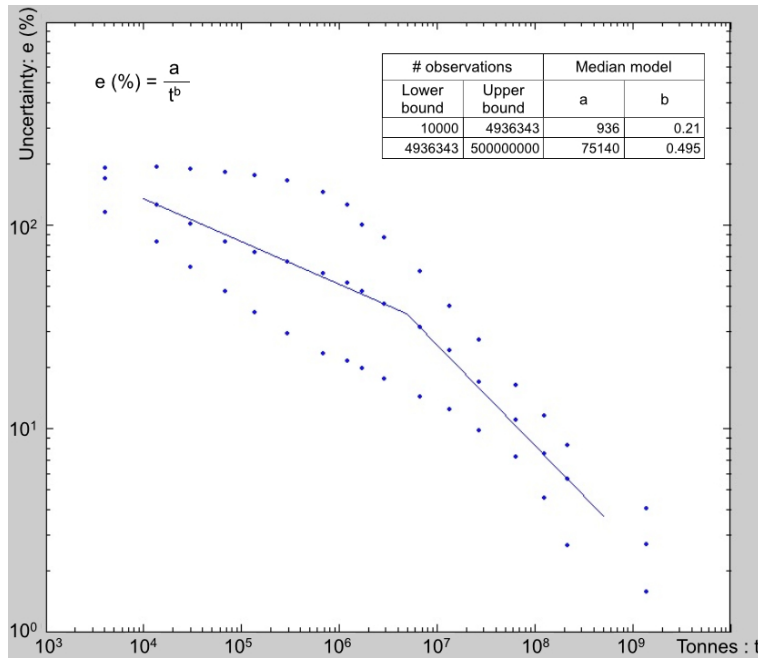


Figure 6: Evolution of lower bound, median and upper bound uncertainties depending on the number of tonnes extrapolated (NB: this is a log-log plot).

## 4 Uncertainty reduction when averaging over several years

### 4.1 Building independent multi-year samples

In the survey, every year, half of the sample is renewed (half the trucks surveyed year  $i$  will also be surveyed year  $i + 1$ ) as shown in table 2. When averaging over  $n_y$  years, there are  $2^{n_y-1}$  independent samples possible. Table 3 shows 1 among the 16 possibilities for an average over 5 years<sup>4</sup>. Among the independent samples generated, we chose the ones of the largest size.

	Sample 2003	Sample 2004	Sample 2005	Sample 2006	Sample 2007
<b>First half</b>	Draw 2002	Draw 2004	Draw 2004	Draw 2006	Draw 2006
<b>Second half</b>	Draw 2003	Draw 2003	Draw 2005	Draw 2005	Draw 2007

Table 2: Composition of the sample for each year of the survey: Vehicles belonging to the draw of year  $i$  are surveyed on years  $i$  and  $i + 1$ .

<sup>4</sup>Note that sample weights have to be re-calibrated in order to leave the total unchanged: for instance if we remove half of the population surveyed in a given year, we have to double the weights of the remaining half population.

	Sample 2003	Sample 2004	Sample 2005	Sample 2006	Sample 2007
First half	Draw 2002	Draw 2004	Draw 2004	Draw 2006	Draw 2006
Second half	Draw 2003	Draw 2003	Draw 2005	Draw 2005	Draw 2007

Table 3: A possibility of independent sample for computing a year-average between years 2003 and 2007. White cells are kept, grey cells are removed from the sample. In total, 16 combinations are possible. Note that, if a one-year sample is composed of  $p$  individuals, the complete sample would be composed of  $n_y p$  individuals while every independent sample is only composed of  $p + (n_y - 1)p/2$  individuals.

## 4.2 Results

Results of averaging over the years 2005-2007 (3-year average) are shown in figure 7. The slope of the model is 0.71 indicating an average reduction of uncertainty of 29%.

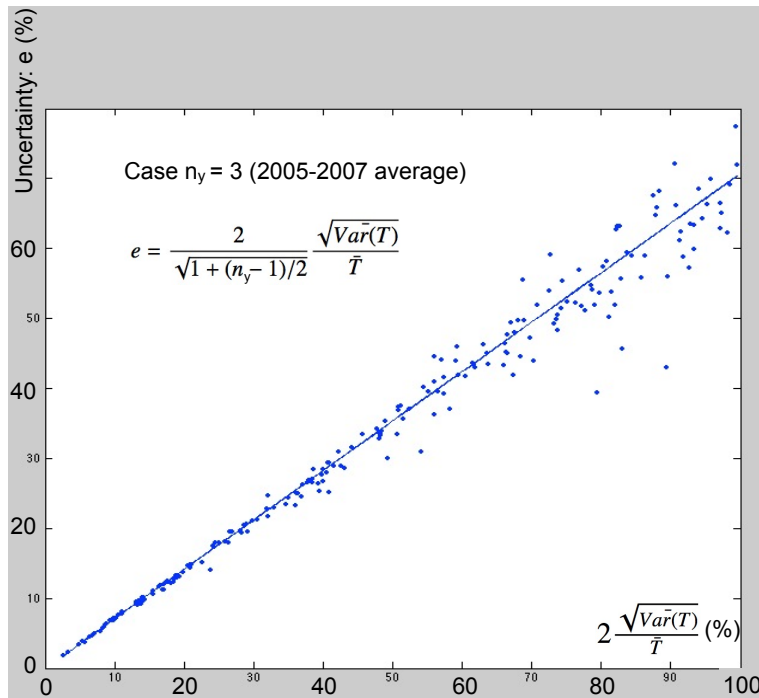


Figure 7: Uncertainties of a 3-year average (2005-2007) depending on  $2\sqrt{\tilde{V}(T)}/\bar{T}$ , with  $\tilde{V}(T) = 1/3 \times (V(T_{2005}) + V(T_{2006}) + V(T_{2007}))$  and  $\bar{T} = 1/3 \times (T_{2005} + T_{2006} + T_{2007})$ . The line corresponding to the model proposed is plotted in blue. The  $R^2$  index of the model is 0.97. About 200 data points are sampled over the 31000 datapoints of the (independent 3-year) sample in order to make the graph readable.

## References

- Berger, Y. G. (1998). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* 74, 149–168.
- Berger, Y. G. (2003). A modified Hájek variance estimator for systematic sampling. *Statistics in Transition* 6, 5–21.
- Cochran, W. G. (1977). *Sampling Techniques, third edition*. Wiley.
- Fuller, W. (2009). *Sampling Statistics*. Wiley.
- Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en population finie*. Dunod.
- Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag.