A probabilistic approach to screen and improve emission inventories

Alain Clappier, Philippe Thunis

Version of Record: https://www.sciencedirect.com/science/article/pii/S1352231020305653
Manuscript_c5ed9aa607781f36d9038f2e9067c2e8

A probabilistic approach to screen and improve emission inventories

A. Clappier[1] and P. Thunis[2].

[1]: Université de Strasbourg, Laboratoire Image Ville Environnement, Strasbourg, France.
[2]: European Commission, Joint Research Centre, Ispra, Italy.

Abstract

Emission inventories are generally identified as the key input to the air quality modelling chain, especially when they are used to support regulatory decisions, such as for air quality planning or for the assessment of concentration levels over a given territory. At the same time, studies point out to emission inventories as the most uncertain factor among the different components of air quality models. In a recent work, Thunis et al. (2016), developed a methodology, supported by a specific screening diagram, to identify discrepancies between emission estimates and target the pollutants and sectors for which improvements should be prioritized. Based only on the total emissions for various pollutants as input, the methodology is able to provide insight on whether these differences arise from issues related to emission factors or activities. In this work we further develop this methodology and show that the use of a probabilistic approach improves its usefulness and relevance. We motivate the use of a probabilistic approach by discussing a series of simple situations to which we apply an "intuitive reasoning". These situations are then used as background to detail the probabilistic methodology and its main assumptions. Tested on a random set of known emission inventories, we show that the methodology performs well in reproducing the expected activities and the associated emission factors. We show that the method becomes more precise when the number of pollutants increases. Given the large differences observed between emission inventories, reducing the discrepancies between them does not only lead to more coherence but it also improves their accuracy as errors can be detected and solved. The approach is mostly designed as a screening to spot the main inconsistencies in the field of atmospheric emissions but the methodology is general and could be applied to other fields, provided that the relationships between variables fulfil similar rules as those described here.

Keywords: emission inventories, activity data, emission factors, probabilistic approach, screening

1. Introduction

Emission inventories are generally identified as the key input to the air quality modelling chain, especially when they are used to support regulatory decisions, such as for air quality planning or for the assessment of concentration levels over a given territory (EEA, 2011; ETC/ACM, 2013). At the same time, studies point out to emission inventories as the most uncertain factor among the different components of air quality models (e.g. meteorology, boundary conditions, model parameters) (Russell and Dennis, 2000; François et al., 2005; Davison et al., 2011; Viaene et al., 2013), therefore limiting their ability to explain the observed variability and trends in atmospheric concentration and composition. The development of accurate emission inventories is hence particularly relevant to air quality applications because this will determine to a large extent the accuracy of the subsequent air quality model results (Tong et al., 2011; Frost et al. 2013; Kuenen et al. 2014; Granier et al. 2011, Guevara et al. 2013).

Emission inventories require information concerning activity factors (e.g. total amount of fuel consumed) and emission factors per activity (e.g. amount of pollutant emitted per activity unit). This calculation is generally made at a very detailed level in terms of sectors and sub-sectors. For a given activity, the emission (e) for a given pollutant (p) is then obtained as the product between the emission factor (f) and the activity data (a).

$$e_p = f_p a \quad (1)$$

The purpose of this decomposition is to distinguish properties that affect all pollutants in the same way (a) from properties that are attached specifically to one pollutant (f_p). In most cases the common property is the activity while the specific properties are the emission factors. For example, the traffic can be defined by an activity (e.g. expressed in km driven) that is similar for all pollutants and emission factors (e.g. expressed in g/km) that differ for each pollutant. It is important to note that the unit or the method used to estimate the activity and emission factors are not relevant here, as these two parameters are not used explicitly in the method.

With the increased use of air quality modelling to support planning, more emission inventories are developed all over the world. As a result, several emission inventories, based on different methodologies (bottom-up or top-down), and/or different proxies are often available for the same region. Comparing these emission inventories is then feasible and brings information about their quality. Of course, if all emission inventories are similar to each other, nothing can be concluded and the emission inventories can either be all very close or very far to the truth. However, the inter-comparison generally highlights huge differences. For example, Trombetti et al. (2018) showed differences that were beyond 100% for some activity sectors among six European emission inventories. In such a case, only one inventory can be close to the truth while at least five inventories are far from it. And in the worst case they are all far from the truth! In any case, the conclusion of the inter-comparison is that something must be done to improve these inventories. The logical process would consist in comparing in details the inventories sector by sector, sub sector by sub sector to identify the origin of the differences. Since differences often exceed 100%, some input data or methods must differ strongly. A regular comparison and updating of the emission inventories would increase and maintain their level of quality and coherence. Given these large differences, it is worth mentioning that a reduction of the discrepancies between two inventories does not only lead to more coherence but it also improves their accuracy as errors ("obvious" mistakes) can be detected and solved. Unfortunately, experience shows that once compared, emission inventories are seldom checked in such details and improved. Two main reasons explain this: 1) given the complexity and the large number of sectors and sub-sectors to cover, several persons are generally involved in the compilation process, 2) because of this complexity, the developers of one inventory have generally no access to the details of other inventories. Under these conditions, identifying the origin of differences between emission inventories is a challenging task that requires time and coordination.

In a recent work, Thunis et al. (2016), denoted as Thunis2016 in the following, developed a methodology to help identifying the reasons for discrepancies between two emission estimates over a given area. One of the main purposes of this screening methodology is to target the pollutants and sectors that show the largest inconsistencies using a very low level of details about the two inventories which have to be compared (only the total emission of the macro sectors is required). The methodology does not provide information about which inventory is closest to the truth but it helps prioritizing the necessary improvements. In particular, the "diamond" diagram proposed in their work aimed at identifying whether differences between

inventories are mostly related to differences 1) in the use of emission factors (properties associated specifically to each pollutant) or 2) in the choice of activity data (property common to all pollutants). The aim of this information is to allow emission inventory developers to focus on the main causes of discrepancies which are likely the causes of errors in their estimates.

Starting from the ratios of emission estimates (e_p^1/e_p^2) from two inventories denoted with the superscripts 1 and 2, the diamond methodology provides information about their activities (a^1/a^2) and emission factors (f_p^1/f_p^2) ratios. Retrieving such information is however only possible if some assumptions are made as the overall problem is characterized by a system of equations that has more unknowns than equations. Thunis2016 assumed that for at least one of the emitted pollutants, additional information on the emission factor was available.

In this work we further develop the diamond methodology and show that the use of a probabilistic approach allows to get rid of the additional assumption made by Thunis2016, and therefore improve the usefulness and relevance of the approach. In section 2, we briefly review the diamond methodology as background to support the description of the improvements discussed in this work. The probabilistic approach is detailed in Section 3 its validation is proposed in Section 4 and the results are discussed in Section 5. We finally illustrate the benefit of the new approach with respect to Thunis2016 with a "diamond" application on two real emission inventories in Section 6.

2. Background and purpose

2.1. The diamond methodology

In Thunis2016, the Authors propose a methodology to compare emission inventories based on emission ratios. This methodology is applied to different pollutants like PPM (primary particulate matter), NO_x, VOC, SO_2, etc… and different activity macro-sectors (i.e. transport, industry, domestic, etc…). In this section, we briefly review the main aspects of this methodology. In a first step, Thunis2016 detail the methodology for a single activity for which emissions can be expressed as the product of an emission factor by that activity. The emission ratio between two inventories is then equal to the product between an emission factor ratio and an activity ratio:

$$\frac{e_1^p}{e_2^p} = \frac{a_1}{a_2}\frac{f_1^p}{f_2^p} \tag{2}$$

in which the subscripts 1 and 2 identify the two inventories for a pollutant *p*.

As detailed in Thunis2016, the diamond methodology aims at quantifying inconsistencies in terms of emission factors and activity ratios (f_1^p/f_2^p and a_1/a_2) only from the limited knowledge we have of the total emission ratio (e_1^p/e_2^p). The emission factor ratio is then used together with the activity ratio as coordinates in a log-log diagram referred as the diamond diagram (Figure 1). In this diagram, each pollutant (for a given activity) is represented by a

specific point. The advantage of the log-coordinates is that emission iso-ratio appears along a diagonal of slope −1 as a result of relation (2) expressed in log scale:

$$\hat{e}^p = \hat{f}^p + \hat{a} \tag{3}$$

in which $\hat{e}^p = log\left(\frac{e_1^p}{e_2^p}\right)$, $\hat{f}^p = log\left(\frac{f_1^p}{f_2^p}\right)$, and $\hat{a} = log\left(\frac{a_1}{a_2}\right)$

The system described by equation (3) contains p equations (one per pollutant) but has $p+1$ unknowns (p emission factor ratios and one activity ratio). In order to locate the points along their diagonals, the Authors propose to assume that one pollutant species (denoted as p^*) would serve as a reference, i.e. that the emission factors for this reference pollutant p^* are equal in the two inventories:

$$\exists p^* \text{ so that } f_1^{p^*}/f_2^{p^*} \approx 1 \text{ and then } \hat{f}^{p*} = 0 \tag{4}$$

Thanks to condition (3) and (4), the emission factor ratio for any other pollutant "p" can then be estimated from the total emission ratio as follows:

$$\hat{f}^p = \hat{e}^p - \hat{a} = \hat{e}^p - \hat{e}^{p*} + \hat{e}^{p*} - \hat{a} = \hat{e}^p - \hat{e}^{p*} + \hat{f}^{p*} = \hat{e}^p - \hat{e}^{p*} \tag{5}$$

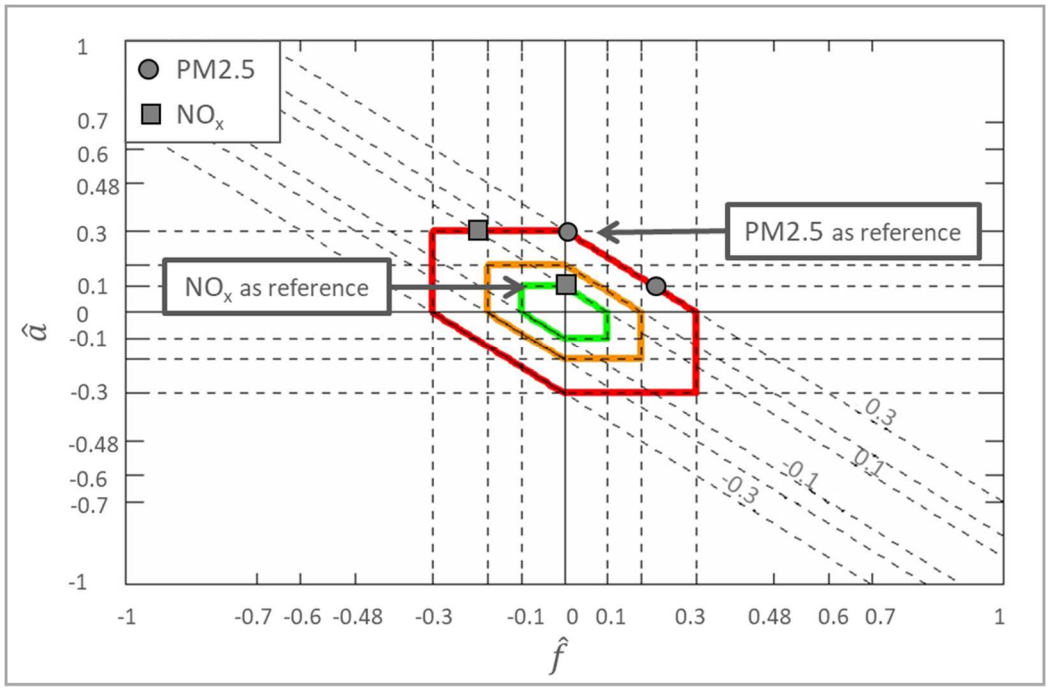

Figure 1: Diamond plot showing an emission comparison for two pollutants, NO_x and PM2.5. The X-and Y-axes indicate the discrepancies between the two inventories in terms of emission factor and activity, respectively. The diagonal ratio iso-lines are indicative of discrepancies in terms of total emissions. The colored iso-lines delimitate the areas where the three factors: emission totals, activity rate, and emission factors are all within a given threshold. See additional explanations in text.

The diamond diagram in Figure 1 provides a simple example with two pollutants (NO_x and PM2.5) for a given activity. Because the log of the total emission ratio $\hat{e}^{NO_x} = 0.1$ (given

input), the NO_x point always lay on the descending 0.1 diagonal while the PM2.5 point is on the 0.3 diagonal as a result of $\hat{e}^{PM2.5} = 0.3$. Because all pollutants for a given activity have the same activity, both pollutants lay on the same horizontal line. In Figure 1, the two horizontal lines represent the same activity but with two different choices for the reference pollutant p^* in equation (4) (top line with PM2.5 and bottom line with NO_x). The reference pollutant always lay on the vertical axis by definition ($\hat{f}^{p*} = 0$, see relation (4)). We see that the interpretation of the diamond diagram in terms of emission factor and activity ratios is strongly affected by the choice of the reference pollutant. Indeed, the log of the activity ratio would be estimated to 0.3 with p^*=PM2.5 but to 0.1 when p^*=NO_x. The reader is referred to Thunis2016 for more details on the methodology. The fact that the interpretation strongly depends on the choice of the reference pollutant is a weakness of the methodology that was identified and discussed in Thunis2016. We introduce in this work an alternative way to build the diamond diagram which does not require assuming a reference pollutant (i.e. assumption (4) disappears).

2.2. Intuitive reasoning

The methodology proposed in this work to overcome the problem of defining a reference pollutant is based on an intuitive reasoning which is often used to estimate the origin of the discrepancy between two emission inventories: "When the comparison of two inventories shows differences that are almost equal for all pollutants, the inconsistency must be caused by a parameter that is common to the calculation of all pollutant emissions. In our case, this must be the activity. When the discrepancy found for one of the pollutants is very different from the discrepancies found for the other pollutants, the inconsistencies should be due to a parameter which is specific to the emission calculation of this pollutant: the emission factor". The following paragraphs aim to verify and codify mathematically this intuitive reasoning and then to use this mathematical analysis to modify the design of the Diamond plot proposed in Thunis2016.
To support our analysis, we start with a few examples based on three pollutants (NO_x, PM2.5 and SO_2) for which we distinguish three cases that cover situations usually met when comparing emissions estimates from two inventories.

Case A: <u>Emission ratios are close to 1 for all three pollutants</u>

$$\hat{e}^{NO_x} \approx \hat{e}^{PM2.5} \approx \hat{e}^{SO_2} \approx 0, \qquad (6)$$

which is similar to state that:

$$\hat{f}^{NO_x} + \hat{a} \approx \hat{f}^{PM2.5} + \hat{a} \approx \hat{f}^{SO_2} + \hat{a} \approx 0 \quad \text{and} \qquad \hat{f}^{NO_x} \approx \hat{f}^{PM2.5} \approx \hat{f}^{SO_2} \approx -\hat{a}.$$

Two possibilities can be distinguished:

<u>Possibility 1</u>: $\hat{a} \approx 0$
This implies that $\hat{f}^{NO_x} \approx \hat{f}^{PM2.5} \approx \hat{f}^{SO_2} \approx 0$, i.e. that both the activity and emission factors are very similar in the two inventories.

<u>Possibility 2</u>: $\hat{a} \neq 0$

In this case, the differences in terms of activities must be compensated by differences in terms of emission factors, and this compensation must be similar for all pollutants ($\approx -\hat{a}$).

If the two inventories were compiled with different input data or methodologies, it would not be surprising to observe substantial differences in terms of activities and emission factors. However, the eventuality that the different input data and methodologies used to construct the two inventories lead to activity ratios compensating for the erroneous emission factor ratios in the exact same way for all pollutants (3 in our example) is very improbable. On the other hand, if the two inventories use similar input data and similar compilation methodologies, it is highly probable that their activities and emission factors will be close in the two inventories. Possibility 1 is therefore more probable than possibility 2.

Case B: Emission ratios are close to each other for all pollutants, but not equal to 1

This case is similar to the previous one, but with a similar bias in the emission ratios for all pollutants, i.e.

$$\hat{e}^{NO_x} \approx \hat{e}^{PM2.5} \approx \hat{e}^{SO_2} \approx \gamma \quad (7)$$

then $\hat{f}^{NO_x} + \hat{a} \approx \hat{f}^{PM2.5} + \hat{a} \approx \hat{f}^{SO_2} + \hat{a} \approx \gamma$

and $\hat{f}^{NO_x} \approx \hat{f}^{PM2.5} \approx \hat{f}^{SO_2} \approx \gamma - \hat{a}$

Similarly, to the previous case, two possibilities may occur:

Possibility 1: $\hat{a} \approx \gamma$ and $\hat{f}^{NO_x} \approx \hat{f}^{PM2.5} \approx \hat{f}^{SO_2} \approx 0$

In this case, emission factors are very similar in both inventories and the differences are entirely explained by the differences in terms of activities

Possibility 2: $\hat{a} \neq \gamma$
Differences between emissions are not (at least not entirely) explained by differences between activities. This implies that the differences in activities must be compensated by differences in emission factors, in the exact same way for all pollutants.

For the reasons exposed in the previous case, Possibility 1 is therefore more probable than possibility 2.

Case C: Emission ratios are very different

$$\hat{e}^{NO_x} \gg \hat{e}^{PM2.5} \gg \hat{e}^{SO_2} \quad (8)$$

then $\hat{f}^{NO_x} + \hat{a} \gg \hat{f}^{PM2.5} + \hat{a} \gg \hat{f}^{SO_2} + \hat{a}$

and $\hat{f}^{NO_x} \gg \hat{f}^{PM2.5} \gg \hat{f}^{SO_2}$

In this case, it is possible to conclude that the emission factor ratios are very different and are ranked in the same order as emission ratios. However, at this stage of the analysis, nothing can be concluded in terms of activities.

In summary, the emission ratios result from the product of emission factors and activity ratios. In principle, an infinity of combinations exists for emission factors and activities that lead to a similar value for the total emissions. Nevertheless, the approach highlighted here with three specific examples shows that some combinations appear more probable than others.

3. Methodology

In this Section, we first discuss important properties that characterize the probability distributions of the activity and emission factors. We then formulate the methodology to identify their most likely values, based on the only knowledge of the total emissions. Note that in the following of this work, for convenience, we will use "LR" to indicate that the variable under discussion is a logarithm of a ratio (e.g. "LR activity" means" logarithm of the activity ratio") and we will denote it with a "hat" symbol.

3.1. Probability distribution

Distributions are centered on zero

The diamond methodology starts from known total emissions of two inventories. Its goal is to provide information about the activity and the emission factors that are not known. We know however that these activities and emission factors can be estimated through different approaches (e.g. top down vs. bottom-up) and are based on different reference and/or datasets (e.g. different versions of COPERT for the emission factors). The emission factor and the activity are therefore considered as two random variables denoted by the uppercase letters F and A. We show hereafter that it is reasonable to assume a normal distribution centered around zero for the LR of each of these variables. Let's start with the activity.
The LR of the activities, can also be considered as a random variable which can be decomposed as a difference between two components:

$$\hat{A} = log(A_1/A_2) = log(A_1) - log(A_2). \quad (9)$$

Each logarithm component in this equation is considered as a random variable centered on the logarithm of the true activity value. We expect that if we pick one inventory among a set of inventories developed independently from each other, the probability of overestimating the activity is the same as underestimating it. This fact is reflected by centering the distribution on the logarithm of the true activity value. If the distribution of A_1 and A_2 are centred on the true activity value, the distribution of A_1/A_2 is then centered on 1 and $\hat{A}$ is therefore centered on 0.

Distributions are normal

Regarding the type of distribution, the Central Limit Theorem tells us that the sum/difference of distributions, regardless of their type, tends towards a normal distribution. Because the LR activity, represents a difference between two activity logarithms, it is reasonable to assume

that its distribution is normal regardless of the type of distribution characterizing the single activity logarithms.
The same reasoning could be done for the LR of the emission factor leading to consider this variable as a random variable centered on zero.

Given these properties, it is therefore reasonable to assume in the following that the probability density functions (PDFs) associated to the probability distributions for $\hat{A}$ and $\hat{F}^p$ can be expressed as normal laws centered around 0:

$$fct_{\hat{A}} \sim \mathcal{N}(0, \sigma_{\hat{A}}) \qquad \text{and} \qquad fct_{\hat{F}^p} \sim \mathcal{N}(0, \sigma_{\hat{F}^p}) \tag{10}$$

with $\sigma_{\hat{A}}$ the standard deviation for $\hat{A}$ and $\sigma_{\hat{F}^p}$ the standard deviation for $\hat{F}_p$. These standard deviations represent the uncertainties that characterize the LR of the activities and emission factors, respectively.

3.2. Probability calculation

The PDFs can be used to calculate the probability that the values of $\hat{A}$ and $\hat{F}^p$ fall within given intervals (Figure 2):

$$P_{\hat{A}}^{\delta} = P\{\hat{A} \in [\hat{a} - \delta/2\,, \hat{a} + \delta/2]\} = \int_{\hat{a}-\delta/2}^{\hat{a}+\delta/2} fct_{\hat{A}}(\hat{A}) d\hat{A} \tag{11}$$

$$P_{\hat{F}^p}^{\delta} = P\{\hat{F}^p \in [\hat{f}^p - \delta/2\,, \hat{f}^p + \delta/2]\} = \int_{\hat{f}^p-\delta/2}^{\hat{f}^p+\delta/2} fct_{\hat{F}^p}(\hat{F}^p) d\hat{F}^p \tag{12}$$

in which $P_{\hat{A}}^{\delta}$ is the probability that $\hat{A}$ takes a value between $\hat{a} - \delta/2$ and $\hat{a} + \delta/2$, and $P_{\hat{F}^p}^{\delta}$ is the probability that $\hat{F}^p$ takes a value between $\hat{f}^p - \delta/2$ and $\hat{f}^p + \delta/2$, δ being the width of the intervals, $\hat{a}$ and $\hat{f}^p$ their centers.

The diamond methodology aims at finding the most likely point's location on the diamond plots, in other words finding the most likely values of $\hat{A}$ and $\hat{F}^p$ among the set (subset of the overall distributions discussed in the previous section) of LR activities and LR emission factors that fulfill the constraint:

$$\hat{e}^p = \hat{F}^p + \hat{A} \tag{13}$$

In which the lower case letter is used for the LR emission totals ($\hat{e}^p$) to indicate that their values are fixed while the LR activity and LR emission factors are upper case letters as they represent random variables. If we take the example of one pollutant (PM2.5), we estimate the probability that the following two events are fulfilled:

$$\xi_1: \qquad \hat{A} \in [\hat{a} - \delta/2\,, \hat{a} + \delta/2] \tag{14}$$

$$\xi_2: \qquad \hat{F}^{PM2.5} \in \left[\hat{f}^{PM_{2.5}} - \delta/2\,, \hat{f}^{PM2.5} + \delta/2\right] \tag{15}$$

The probability that these two events are simultaneously fulfilled is expressed as the probability of the two events intersection: $P^{\delta}_{\hat{A},\hat{F}^{PM2.5}} = P(\xi_1 \cap \xi_2)$. Moreover, it is reasonable to assume that these two events are independent because the activity and the emission factor are computed in different ways. Consequently, the probability of the two events intersection is equal to the product of the single event's probabilities:

$$P^{\delta} = P(\xi_1 \cap \xi_2) = P(\xi_1) \times P(\xi_2) = P^{\delta}_{\hat{A}} \times P^{\delta}_{\hat{F}^{PM2.5}} \tag{16}$$

Thanks to equation (3), $\hat{F}^{PM2.5}$, the interval center for the LR of the emission factors can be related to $\hat{a}$, the interval center for the LR of the activity (Figure 2):

$$\hat{F}^{PM2.5} = \hat{e}^{PM2.5} - \hat{A} \tag{17}$$

The probability of event ξ_2 can therefore be computed using only $\hat{A}$ and δ:

This calculation can be generalized to several pollutants (N). The probability that the LR activity and all LR emission factors fall simultaneously inside intervals of width δ is then expressed as follows:

$$P^{\delta} = P^{\delta}_{\hat{A}} \times \prod_{p=1}^{N} P^{\delta}_{\hat{F}^{p}} \tag{18}$$

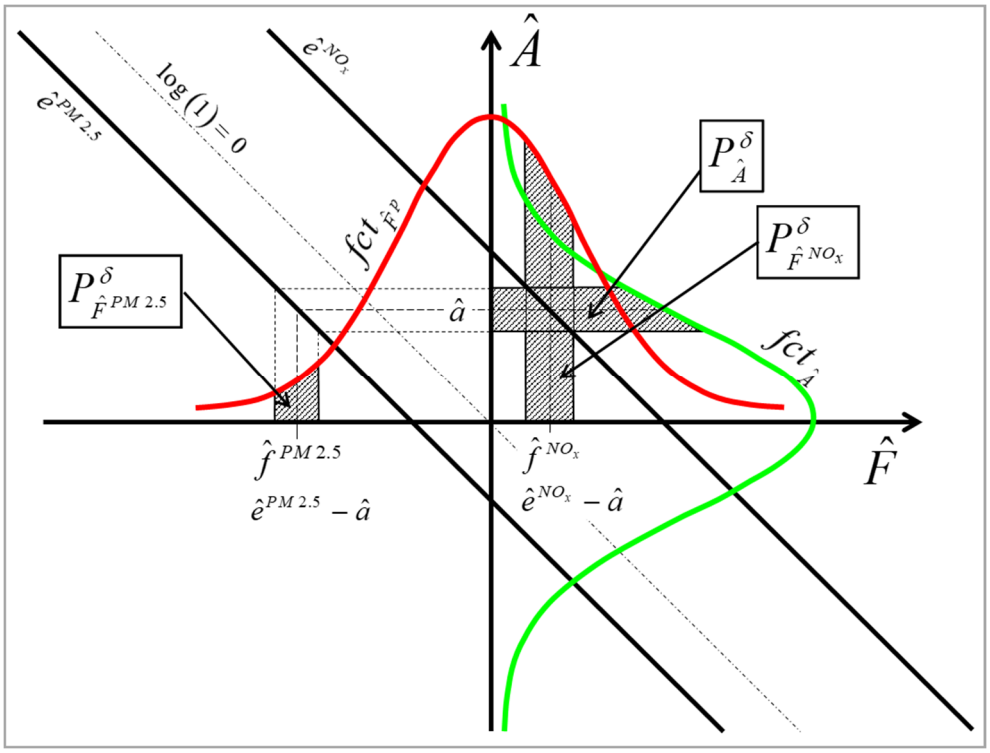

Figure 2: schematic representation of the probability density functions in the case of two pollutants (NO_x and PM2.5). The LR of the total emissions represent the known input data. Although we do not know the exact position of these total emissions in the activity-emission factor ($\hat{A}$-$\hat{F}$) diagram, the NOx and PM2.5 total emissions must lie on their respective diagonals of slope -1 (Equation 4). A probabilistic approach is used to determine the most probable ($\hat{A}$-$\hat{F}$) position for the activity and associated emission factors, along those diagonals. For each interval δ along the LR activity axis ($\hat{A}$), the probability of the activity falling within this interval is calculated (dashed area beneath the green LR activity probability curve). The associated probability for the LR emission factors is calculated similarly (shaded areas beneath the red LR emission factor curve) for both NO_x and PM2.5. For each δ, the overall probability is then estimated as the product of three probabilities (activity, PM2.5 emission factor, NOx emission factor). The most likely value of the activity (and therefore of the associated emission factors) is found when the maximum of this overall probability is reached.

For any value of $\hat{A}$ fulfilling constraint (17), we can consequently calculate its associated probability P^{δ} and deduce that its value follows a Gaussian distribution centered on $\hat{a}_{opt}$ with a variance equal to σ, given by the expressions below (see derivations in Annex 1):

$$\hat{a}_{opt} = \frac{\sum_{p=1}^{N} \hat{e}^p \rho^p}{1+\sum_{p=1}^{N} \rho^p} \quad \text{and} \quad \sigma = \frac{\sigma_{\hat{A}}}{\sqrt{1+\sum_{p=1}^{N} \rho^p}} \tag{19}$$

$$\text{with} \quad \rho^p = \frac{\sigma^2_{\hat{A}}}{\sigma^2_{\hat{F}^p}}$$

The probability distributions of the LR emission factors of this constrained distribution can be directly obtained from the LR activity distribution through equation (17) because the $\hat{e}^p$ are

known. The subscript "opt" stands for "optimal" and represents the activity $\hat{a}$ for which P^{δ} is maximum.

The standard deviations of the constrained distribution of the LR emission factors are all equal to the standard deviation of the LR activity (σ) because $\hat{F}^p$ is equal to a constant ($\hat{e}^p$) minus $\hat{A}$.

The LR emission factors follow therefore a Gaussian distribution centered on:

$$\hat{f}^p_{opt} = \hat{e}^p - \frac{\sum_{p=1}^{N} \hat{e}^p \rho^p}{1 + \sum_{p=1}^{N} \rho^p} \tag{20}$$

In the diamond plot, all pollutants have the same *Y*-coordinate, $\hat{a} = \hat{a}_{opt}$ and specific *X*-coordinates $\hat{f}^p = \hat{f}^p_{opt}$.

4. Validation

In this Section we test the methodology described in the previous Sections with generic numerical examples. We first construct the overall distributions for the LR activities ($\hat{A}$) and LR emission ratios ($\hat{F}^p$) and sum the two to obtain a distribution for the LR of the total emissions ($\hat{E}^p$). All three distributions are normally distributed around zero. While the standard deviations $\sigma_{\hat{A}}$ and $\sigma_{\widehat{F^p}}$ can be freely chosen, the standard deviation of the total emissions is fixed by $\sigma_{\widehat{E^p}} = \sqrt{\sigma^2_{\hat{A}} + \sigma^2_{\widehat{F^p}}}$. Among these overall distributions of total emissions ($\hat{E}^p$), we pick one value per pollutant and construct a set of total emissions ($\hat{e}^p$). Among the subset of ($\hat{A}$) and ($\hat{F}^p$) that fulfills constraint (17), we then apply our methodology (equations 6) to determine $\hat{a}_{opt}$ and σ.

We then calculate the probability associated to each $\hat{a}$ belonging to the constrained distribution of $\hat{A}$ (Figure 3), either by counting its number of occurrences within the distribution (histogram in grey) or via the analytical probability distribution function given by equation 6 (blue line). Both the histogram and analytical probability distribution overlay, and are centered on the value of $\hat{a}_{opt}$ obtained via (19), confirming the derivations presented in the previous sections.

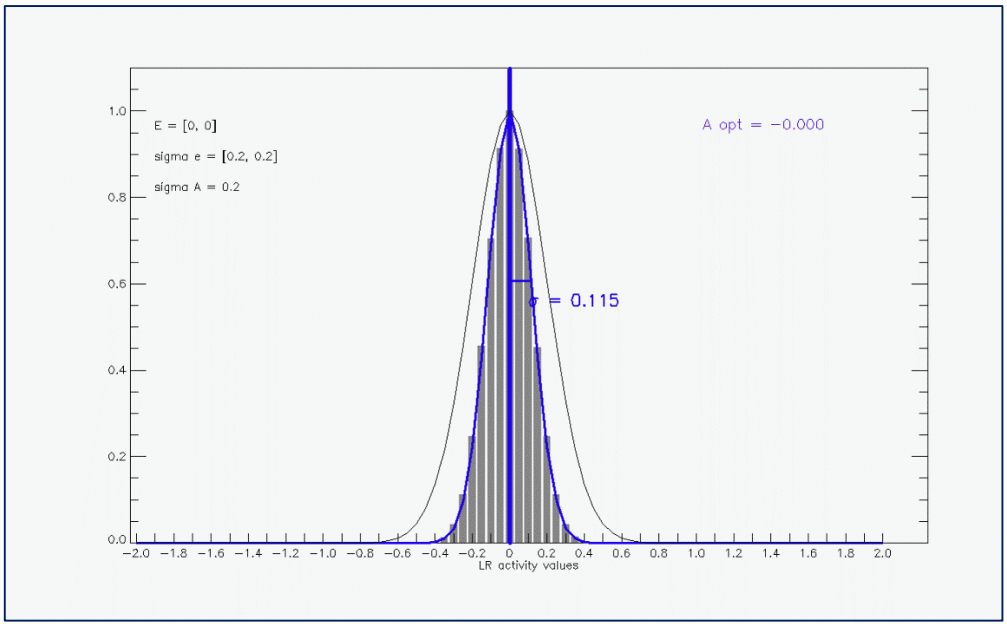

Figure 3: Comparison of the histogram and analytical probability distributions obtained for two pollutants. The known values of the total emissions E are set to 0 and the standard deviation of the original distributions of the LR activity and LR emission factors are set to 0.2. Both the histogram and analytical distributions are centered on the value of a_{opt} given by the theory. The original distribution of the LR activity is over plotted in black to indicate the precision gain between the original and constrained distributions.

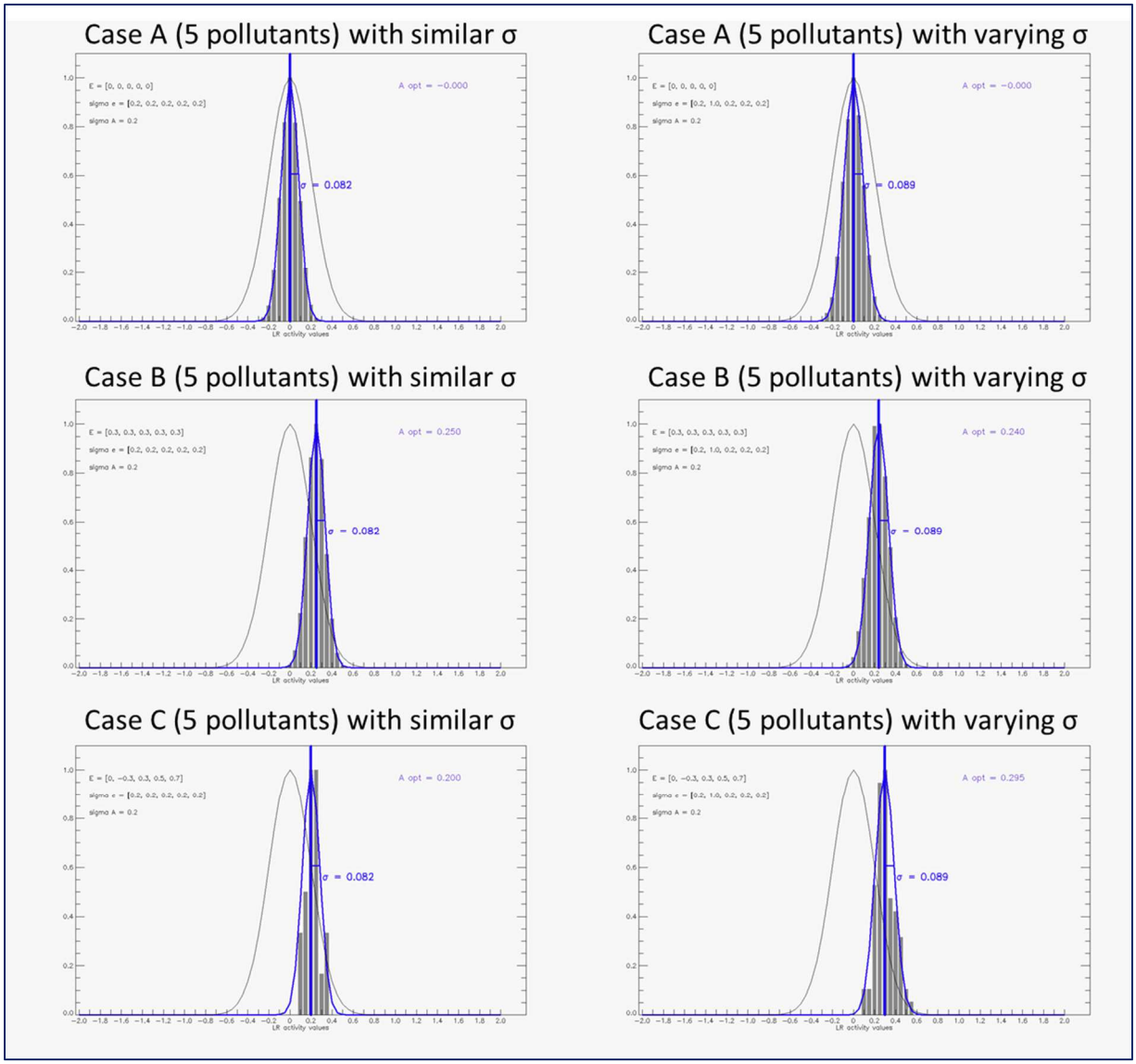

Figure 4: Validation tests cases for the three intuitive situations discussed in Section 2.2. The left and right columns show the results obtained with constant and varying standard deviations respectively. These standard deviations are the ones characterizing the original distributions of the LR activity and emission factors.

From Figure 4, we see that the standard deviation of the distribution of the constrained $\hat{A}$ are similar for cases A, B and C. The standard deviation however decreases with the number of pollutants (as seen from the comparison between Figure 3 and Figure 4 (top-left) and changes with varying original standard deviations (comparison of the left and right columns in Figure 4). We show in Annex 2 that the precision gain (G) between the original standard deviation assumed for $\hat{A}$ ($\sigma_{\hat{A}}$) and the constrained value (σ) is expressed as:

$$G = \frac{\sigma}{\sigma_{\hat{A}}} = \frac{1}{\sqrt{1 + \sum_{p=1}^{N} \rho^p}} \tag{21}$$

Equation (21) confirms that the gain increases with the number of pollutants. It also confirms that our three intuitive cases are equally robust since the gain only depends on the uncertainties associated to the LR emission factors and activity but not on the actual values of the emissions ($\hat{e}^p$). It is interesting to note that Equation (21) also indicates that the gain depends on relative and not on absolute uncertainty levels. Indeed, only the ratios of the activity and emission factors original standard deviations appear in the formulation. Results will remain unchanged even if uncertainties become large as long as these large uncertainties occur for both the activity and the emission factors. Note that if the activity is certain ($\sigma_{\hat{A}} \rightarrow 0$), $\hat{a}_{opt}$ will tend to zero (equation 19) and the robustness gain tends to zero (no gain obtained as the original activity is already known) whereas when once the emission factor of a pollutant species $p*$ is certain ($\sigma_{\hat{F}^{p*}} \rightarrow 0$, for a given p), $\hat{a}_{opt}$ tends to $\hat{e}^{p*}$ (see Annex 1) and the gain tends to infinity. Note that restricting the expression of the gain to the activity is sufficient as activities and emission factors are linked through equation (17).

The results (Figure 3 and Figure 4) show that the methodology is able to estimate the most likely value of the activity with a specific precision gain. In order to check the methodology, we assumed that the original standard deviations were known. In the next Section, we address the importance and relevance of this assumption on the interpretation of the results.

5. Discussion

By construction, points on the diamond plots are located on descending diagonals and their position along this diagonal depends only on their LR activity. Using the mathematical expressions developed in Annexes 1 to 3, we note the following:

- The most likely LR activity value $\hat{a}_{opt}$, does not depend directly on the level of uncertainties attached to the LR activity ($\sigma_{\hat{A}}$) and LR emission factors ($\sigma_{\hat{F}^p}$) but depends only on the ratio of these quantities $\rho^p = \sigma_{\hat{A}}^2 / \sigma_{\hat{F}^p}^2$ (see Equation 6).

- A particular value of $\hat{a}_{opt}$ is obtained when the uncertainties of the LR activity and of the LR emission factors are similar ($\forall p \ \sigma_{\hat{A}} = \sigma_{\hat{e}^p}$). In this case, $\hat{a}_{opt} = \frac{\sum_{p=1}^{N} \hat{e}^p}{N+1}$.

- In Annex 3 it is demonstrated that $\hat{a}_{opt}$ lies within a bounded interval
$$\hat{a}_{opt} \in \left[\min\left(0, \hat{e}^{min}\right), \max(0, \hat{e}^{max})\right]$$
where $\hat{e}^{max}$ and $\hat{e}^{min}$ are the maximum and minimum values of the LR total emissions: $\hat{e}^{max} = \max(\hat{e}^p)$ and $\hat{e}^{min} = \min(\hat{e}^p)$.

We discuss below how the values of $\hat{a}_{opt}$ change in the three cases described in Section 2.2.

Case A : All LR total emissions are equal to 0 ($\forall p \ \hat{e}^p = 0$). Formulations 6 and 7 lead to: $\hat{a}_{opt} = 0$ and $f_{opt}^p = 0$. Moreover $\hat{e}^{min} = \hat{e}^{max} = 0$ implying that the most likely solution for the LR activity and LR emission factors are zero, regardless of their associated uncertainties.

Case B : All LR total emissions are equal to a constant ($\forall p\ \hat{e}^p = e$).
Formulations 6 and 7 lead to: $\hat{a}_{opt} = e\frac{\sum_{p=1}^{N}\rho^p}{1+\sum_{p=1}^{N}\rho^p}$ and $\hat{f}_{opt} = e - e\frac{\sum_{p=1}^{N}\rho^p}{1+\sum_{p=1}^{N}\rho^p} = \frac{e}{1+\sum_{p=1}^{N}\rho^p}$.
$\hat{a}_{opt}$ always lies between 0 and e ($\hat{a}_{opt} \in [\min(0,e), \max(0,e)]$). We can distinguish the following situations:

Extremes: $\hat{a}_{opt} = 0$ and $\hat{a}_{opt} = e$ are two situations for which there is no uncertainty on the LR activity ($\sigma_{\hat{A}} = 0$) and LR emission factors ($\sigma_{\hat{F}^p} = 0$), respectively. When $\hat{a}_{opt} = 0$, a pollutant point in the diamond plot is located on the X-axis, indicating that the LR total emission equals the LR emission factors (Figure 5). When $\hat{a}_{opt} = e$, the pollutant point is located on the Y-axis, indicating that the LR total emission is equal to the LR activity.

Mixed: This includes all situations where some uncertainties are associated to the LR activity and to the LR emission factors. A particular situation occurs when these uncertainties are all similar ($\forall p, \rho^p = 1 \ \Rightarrow \hat{a}_{opt} = e\frac{N}{1+N}$). Because $\hat{f}_{opt}$ is similar for all pollutants, all points are located at the same place on the diamond diagram, in between the two interval bounds (0 and e). When only one pollutant is considered, the single point (white square in Figure 6) is equidistant from the two bounds ($\hat{a}_{opt} = e/2$), indicating that the value of the LR total emission (i.e. the discrepancy between the emission totals in the two inventories) can be explained equally by the LR activity or by the LR emission factor. When the number of pollutants is larger, all pollutant points are collocated at a single place that gets closer to the Y-axis with the increasing number of pollutants ($\hat{a}_{opt} = 2e/3$ for 2 pollutants, $\hat{a}_{opt} = 3e/4$ for 3 pollutants (collocated symbols in Figure 6, etc...), indicating that the value of the LR total emission is more likely to be explained by the LR emission factors than by the activity factor. This is coherent with the conclusions derived from the intuitive reasoning (section 2.2). Without prior information on either the ratio of the LR activity and on the LR emission factors uncertainties, the special case where all uncertainties are equal is a meaningful choice.

To facilitate the interpretation of the diamond plot, we overlay two main diagonals (descending with a slope -1 and ascending with a slope 1 passing through the origin) that delineate four triangle-shaped areas (figure 5). The top and bottom triangles include all locations with a shorter distance to the vertical axis than to the horizontal one, meaning that the LR total emission is most likely explained by the LR activity. On the contrary, the right and left triangles show locations for which the LR total emission is most likely explained by the LR emission factor.

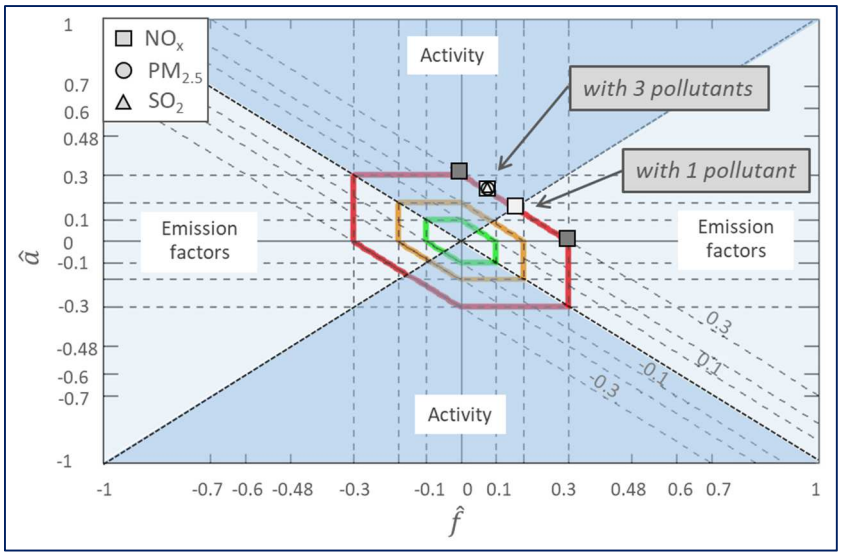

Figure 5 : Diamond plot for case B. Dark grey squares represent the bounds between which the most likely values of $\hat{a}_{opt}$ lie. The light shaded squares represents the special situation when all uncertainties are equal, in the case of one pollutant (white square) and of three pollutants (collocated symbols). The triangles delineate the areas where the uncertainty is likely related to the LR emission factors (left and right) or to the LR activity (bottom and top)

Case C: In comparison to case B, case C is more general. While in case B, all pollutant points on the diamond plot are collocated within bounds defined by 0 and e, the pollutant points are here distributed on a horizontal line (because they all have the same LR activity) at an ordinate $\hat{a}_{opt}$ that lies within an interval bounded by 0, $\hat{e}^{min}$ and/or $\hat{e}^{max}$. The bounding interval can easily become larger than in case B because it is calculated on the basis of more dispersed values of $\hat{e}^{p}$. Nevertheless, without prior information on the ratio of the LR activity and the LR emission factors uncertainties, the specific situation where they are all similar, facilitate the interpretation of the diamond diagram in a similar way to case B. A pollutant point located closer to the X axis than to the Y axis indicates that the LR total emissions is more likely explained by the LR emission factors, for that pollutant (and vice versa). It is interesting to note that when the points are less dispersed (left in Figure 6) case C becomes then close to case B and the diagnostic of the Diamond diagram points out to the activity (NO_x, PM2.5 and SO_2) as the responsible factor for the discrepancies, while when the points are more dispersed (right in Figure 6), the Diamond diagram points to the LR emission factors, at least for some pollutants (NO_x and SO_2). These interpretations are fully coherent with the intuitive reasoning presented in section 2.2.

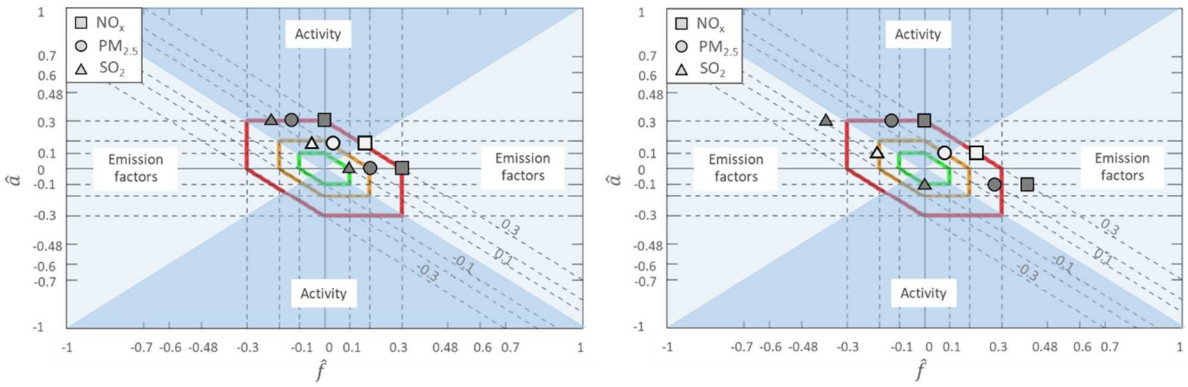

Figure 6: Diamond plot for 3 pollutants (represented by the three symbols) for case C with two situations: (a) all three LR emission totals are relatively close to each other (left) and two of them are close but one more distant (right). The interpretation of the symbols and zones on the diagram is similar to Figure 6.

As mentioned in the introduction, limited discrepancies between two inventories (which correspond to case A) does not imply they are accurate. Indeed, both inventories can agree but be either close or far from the truth. Our analysis however shows that both their activity and emission factors are likely close too (because a compensation that is similar for all pollutant is unlikely). This conclusion is robust because it does not depend on the respective uncertainties associated to the emission factors and activity.
On the contrary, for cases B and C, the high discrepancies that appear help identifying mistakes. Although in these cases, the diagnostic of the Diamond plot can be affected by the relative uncertainties associated to the emission factors and activity, the analysis shows that the points location on the Diamond are bounded. As mentioned above the two extreme bounds represent cases for which we have prior information on one of the LR emission factor or on the LR activity, which is known with certainty. In such a case, the diamond plot loses its usefulness as the issue we want to solve (understand whether uncertainties are mostly associated to one factor or another) is solved beforehand. In addition, this case is not realistic as some level of uncertainties always exist. Moreover, the case in which the emission factors and activity uncertainties would be exactly equal is also unlikely. The question is then: which uncertainty value should one pick when no prior information is available? In the absence of any information, we believe that the best option is to use at start equal uncertainties as default. This has the advantage of locating the points on the Diamond plot at an "average" location from which above or below extrapolations are straightforward. This position constitutes therefore a first guess starting point. Although uncertainties are unlikely exactly equal to each other, they are indeed likely to be close. In the absence of relevant information, assuming that the uncertainties of emission factors and activity are relatively close seems therefore reasonable. Obviously, if information becomes available, it can easily be introduced in the diamond methodology to update its results.

6. Illustrative application

In this section we apply the new methodology and illustrate on a real dataset how to identify the most likely values for the LR activity and LR emission factors for a given activity sector, and graphically represent the results on the diamond plot. We apply the approach on the dataset discussed in Trombetti et al. (2018).

This dataset is composed of 6 inventories and is focusing on four pollutants (NO_x, PM25, VOC and SO_2) in 11 European cities. For this application, we select four specific city-inventory-sector comparisons to illustrate the interpretation of typical situations (Figure 7). We refer to Trombetti et al. (2018) for more details on the dataset.

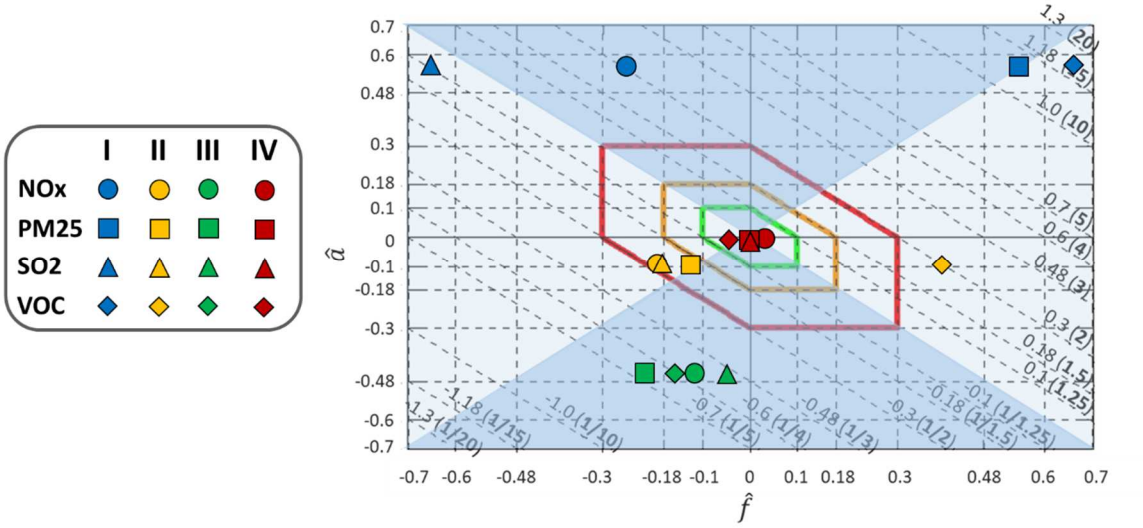

Figure 7: Diamond diagram for four different situation: I (blue) comparison between the CTM4IAM and EDGAR inventories for the residential sector in Paris; II (orange) comparison between the CTM4IAM and MACC3 inventories for the transport sector in Utrecht; III (green) comparison between the EDGAR and MACC2 inventories for the industry sector in Budapest and IV (red) comparison between the EDGAR and MACC2 inventories for the industry sector in Barcelona. Four pollutants are considered: NO_x, VOC, PM2.5 and SO_2). See text for explanations on the positioning of each point in the diagrams.

In case I (corresponding to our intuitive case C), SO_2 shows a low discrepancy in terms of total emissions (value of 1/1.25 on the descending diagonal) whereas NO_x, PM2.5 and VOC show much larger overall discrepancies (factor 3 for NO_x, 10 for PM25 and 15 for VOC). The points for VOC, PM25 and SO2 are located at the transition between the light and dark blue zones on the Diamond plot. It indicates that for these pollutant emission differences are originating from both the activity and emission factors. It does not provide information on priority improvement actions but it provides information that support this process. Although the added value of the diamond diagnostic is limited in this type of situation, it yet shows that: 1) CTM4IAM overestimates EDGAR in terms of residential activity, 2) CTM4IAM overestimates EDGAR for the PM25 and VOC emission factors whereas it underestimates the SO_2 and NO_x ones. Finally, it is interesting to note the good agreement observed originally in terms of SO_2 total emissions hides a large compensation between an overestimation of the activity and an underestimation of the emission factor. This finding is a result of the multi-pollutant approach used for the diamond diagnostic.

In case II, compared to MACC3, the CTM4IAM inventory underestimates the total emissions for NO_x, SO_2 and PM25 (values of 1/2 for NO_x and SO_2, 1/1.3 for PM25 on the descending diagonals) whereas it overestimates the total VOC emissions by a factor 2. In this case, all the points lie in the light blue areas of the Diamond plot indicating that discrepancies are most likely due to emission factors, especially for VOC, that needs to be checked in priority in the two inventories.

In case III (corresponding to our intuitive case B), the EDGAR inventory underestimates the total emissions for all pollutants (values of 1/5 for PM25, 1/4 for VOC and NO_x, and 1/3 for SO_2 on the descending diagonals). In this case, the Diamond plot shows all points in the dark blue areas indicating that discrepancies are most likely due to the activity that needs to be checked in priority in the two inventories.

In case IV (corresponding to our intuitive case A), discrepancies are low for all pollutants (descending diagonal values ranging between 1/1.3 and 1.3). When total emissions show low differences between inventories, the Diamond plot indicates that both the emission factors and activity are likely to be quite similar as well.

Although not used in the context of this application, it is possible to facilitate further the prioritization of actions by associating to each pollutant/sector point in the diagram the value of its total emission. This translates into smaller or larger symbol sizes that help relativizing some issues, e.g. a far-away point in the diagram but with a small size will not be a first priority (see Thunis2016 for details).

7. Conclusion

The Diamond diagram proposed by Thunis2016 represents a simple way of screening the differences between emission inventories. It allows one to identify the pollutants and sectors for which the inconsistencies are the largest. Based only on the total emissions for various pollutants as input, the methodology is able to identify the most problematic pollutants for which further work is needed and provide at the same time some insight on whether these differences arise from issues related to emission factors or activities. Unfortunately, given the larger number of unknowns with respect to the available number of equations in the methodology, Thunis2016 needed to introduce an additional assumption on one of the pollutants to resolve the system (one of the LR emission factor would be known better than the others and could be assumed a value of 0). An alternative approach based on likelihood is proposed here to overcome the Thunis2016 weakness linked to this additional assumption while keeping the main advantages of the approach.

We first motivate the use of a probabilistic approach by discussing a series of simple situations to which we apply an “intuitive reasoning”. These situations are then used as background to detail the probabilistic methodology and its main assumptions, e.g. on the distributions and their uncertainties.

Tested on a random set of known emission inventories, we show that the methodology performs well in reproducing the expected activities (and hence the associated emission factors). The gain of the method, measured as the ratio between the original and solution standard deviation attached to the activity) has a simple formulation that only depends on the number of pollutants.

It is important to stress the fact that the method becomes more precise when the number of pollutants increases. It is therefore important to include all pollutants usually compiled in air quality studies (NO_x, SO_2, VOC, PM2.5, PM10, NH_3) but also other pollutants that are generally not used in the air quality context (e.g. CO_2). Along the same lines, pollutants

emitted by a same source (e.g. shipping) but in other media (e.g. air and water) could be added to the data used for analysis for an increased accuracy.

The diamond approach is mostly designed as a screening to spot the main inconsistencies. It is generally applied when no prior information is known on the uncertainties associated to the emission factors and activities of the different pollutants, in which case the assumption of similar uncertainties remains the most logical one. If this were not the case and that prior information is available, the approach however allows for accounting for the known information.

Although the diamond approach allows differentiating potential inconsistencies in terms of emission factors or activities, it is important to note that it does not require that those activities and emission factors be obtained with similar methods (or be expressed in similar units) in the two inventories. Indeed, differences are made between properties affecting all pollutants in the same way and properties that are attached specifically to one pollutant, regardless of the method or unit used in one or the other inventory. In most cases, the methodologies used in the two inventories will however be relatively similar and the common property will be the activity and the specific properties, the emission factors.

One might advocate for a full access to all internal data underlying emission inventories (emission factors or activities) to facilitate their inter-comparison. Given the number of sectors and pollutants involved, this would however become very demanding. Nevertheless, the diamond approach can be used as a first screening step to rank the sectors that need a priority scrutiny and limit the demand on underlying internal data to the most important sectors.

While the diamond approach is applied in this work to atmospheric emissions, the methodology is general and could be applied to other fields, provided that the relationships between variables fulfil similar rules as those described here.

Acknowledgements
The Authors would like to thanks Bart Degraeuwe and Emanuela Peduzzi for the challenging and fruitful exchanges on the methodology and its associated assumptions.

References

Davison, S., van den Elshout, S., Wester, B., 2011. Integrated Urban Emission Inventories, CiteAIRII, Common Information to European Air, Interreg IVC Programme. Available from: http://www.citeair.eu/fileadmin/Deliverables_and_documents/Guidebook_Integrated_Emission_Inventories_-_final.pdf

EEA, 2011. The application of models under the European Union's Air Quality Directive: A technical reference guide, EEA Technical report 10/2011, Publication Office of the European Union, Luxembourg, ISBN 978-92-9213-223-1. European Environmental Agency.

EEA, 2013. EMEP/EEA air pollutant emission inventory guidebook – 2013. Part B: sectoral guidance chapters. 1.A.4 Small combustion

ETC/ACM, 2013. How to start with PM modelling for air quality assessment and planning relevant to the Air Quality Directive, ETC/ACM Technical Paper 2013/11.

Frost, G.J., Middleton, P., Tarrasón, L., Granier, C., Guenther, A., Cardenas, B., Denier van der Gon, H.A.C., Janssens-Maenhout, G., Kaiser, J.W., Keating, T., Klimont, Z., Lamarque, J.-F., Liousse, C., Nickovic, S., Ohara, T., Schultz, M., Skiba, U., van Aardenne, J., Wang, Y., 2013. New Directions: GEIA's 2020 vision for better air emissions information. Atmospheric Environment, 81, 710-712.

Francois, S., Grondin, E., Fayet, S., Ponche, J.-L., 2005. The establishment of the atmospheric emission inventories of the ESCOMPTE program. Atmospheric Research, 74, 5–35.

Granier, C., Bessagnet, B., Bond, T., D'Angiola, A., Denier van der Gon, H.A.C., Frost, G.J., Heil, A., Kaiser, J.W., Kinne, S., Klimont, Z., Kloster, S., Lamarque, J.-F., Liousse, C., Masui, T., Meleux, F., Mieville, A., Ohara, T., Raut, J.-C., Riahi, K., Schultz, M.G., Smith, S.J., Thompson, A., van Aardenne, J., van der Werf, G.R., van Vuuren, D.P., 2011: Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980-2010 period. Climatic Change, 109, 163-190.

Guevara, M., Martínez, F., Arévalo, G., Gassó, S., Baldasano, J.M., 2013. An improved system for modelling Spanish emissions: HERMESv2.0. Atmospheric Environment, 81, 209-221.

Kuenen, J., Visschedijk, A., Jozwicka, M., Denier van der Gon, H.A.C., 2014. TNO-MACC_II emission inventory: a multi-year (2003-2009) consistent high-resolution European emission inventory for air quality modelling. Atmospheric Chemistry and Physics Discussions, 14, 5837-5869.

Russell, A., Dennis, R., 2000. NARSTO critical review of photochemical models and modeling. Atmospheric Environment, 34, 2283-2324.

Thunis, P., B. Degraeuwe, K. Cuvelier, M. Guevara, L. Tarrason and A. Clappier, 2016. A novel approach to screen and compare emission inventories, Air Qual. Atmos. Health., 9(4), 325-333.

Tong, D. ., Lee, P., Saylor, R. D., 2011. New Direction: The need to develop process-based emission forecasting models. Atmos. Environ. doi: 10.1016/j.atmosenv.2011.10.070.

Trombetti, M., P. Thunis, B. Bessagnet, A. Clappier, F. Couvidat, M. Guevara, J. Kuenen, S. López-Aparicio and B. Degraeuwe, 2018. Spatial inter-comparison of Top-down emission inventories in European urban areas, Atmos. Environ., 173, 142-156.

Viaene, P., Janssen, S., Carnevale, C., Finzi, G., Pisoni, E., Volta, M., Miranda, A., Relvas, H., Gama, C., Martili, A., Douros, J., Real, E., Blond, N., Clappier, A., Ponche, J.-L., Graff, A., Thunis, P., Juda-Rezle, K., 2013. Appraisal project (FP7-ENV CA 303895). D2.3 Air quality assessment and planning, including modelling and measurement. Available from: http://www.appraisal-fp7.eu/site/images/download/APPRAISAL_D23_V1.pdf

Annex 1: Finding the most likely value for the LR activity

Equation (5) shows that the probability P^{δ} is the product of the probability $P_{\hat{A}}^{\delta}$ that the LR activity is within an interval of width δ centered on α and the probabilities $P_{\hat{F}^p}^{\delta}$ that the LR emission factors are within intervals of width δ centered on $\hat{f}^p$. $P_{\hat{A}}^{\delta}$ and $P_{\hat{F}^p}^{\delta}$ represent the PDF areas delimited by the intervals of width δ (Figure 2). For small enough intervals, $P_{\hat{A}}^{\delta}$ and $P_{\hat{F}^p}^{\delta}$ can reasonably be approximated as follows:

$$P_{\hat{A}}^{\delta} \approx \delta \times \frac{1}{\sigma_{\hat{A}}\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma_{\hat{A}}^2}\hat{a}^2\right]$$

$$\text{and} \quad P_{\hat{F}^p}^{\delta} \approx \delta \times \frac{1}{\sigma_{\hat{F}^p}\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma_{\hat{F}^p}^2}\left(\hat{f}^p\right)^2\right] = \delta \times \frac{1}{\sigma_{\hat{F}^p}\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma_{\hat{F}^p}^2}(\hat{e}^p - \hat{a})^2\right]$$

P^{δ} can then be computed as:

$$P^{\delta} \approx \delta \times \frac{1}{\sigma_{\hat{A}}\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma_{\hat{A}}^2}\hat{a}^2\right] \times \prod_{p=1}^{N}\left\{\delta \times \frac{1}{\sigma_{\hat{F}^p}\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma_{\hat{F}^p}^2}(\hat{e}^p - \hat{a})^2\right]\right\}$$

$$\approx \frac{1}{\sigma_{\hat{A}}}\left(\prod_{p=1}^{N}\frac{1}{\sigma_{\hat{F}^p}}\right)\left(\frac{\delta}{\sqrt{2\pi}}\right)^{N+1} \times exp\left\{-\frac{1}{2\sigma_{\hat{A}}^2}\left[\hat{a}^2 + \sum_{p=1}^{N}\rho_p(\hat{e}^p - \hat{a})^2\right]\right\}$$

$$\text{with} \qquad \rho_p = \sigma_{\hat{A}}^2/\sigma_{\hat{e}^p}^2$$

$$P^{\delta} \approx \frac{1}{\sigma_{\hat{A}}}\left(\prod_{p=1}^{N}\frac{1}{\sigma_{\hat{F}^p}}\right)\left(\frac{\delta}{\sqrt{2\pi}}\right)^{N+1} \times exp\left\{-\frac{1}{2\sigma_{\hat{A}}^2}\left[\hat{a}^2\left(1 + \sum_{p=1}^{N}\rho_p\right) - 2\hat{a}\sum_{p=1}^{N}\rho_p\hat{e}^p + \sum_{p=1}^{N}\rho_p(\hat{e}^p)^2\right]\right\} \qquad \text{(A.1)}$$

The maximum value of P^{δ} is reached for $\hat{a} = \hat{a}_{opt}$ when the derivative $dP^{\delta}/d\hat{a}$ is equal to zero, which is expressed as:

$$\frac{d}{d\hat{A}}\left[\hat{a}^2\left(1 + \sum_{p=1}^{N}\rho_p\right) - 2\hat{a}\sum_{p=1}^{N}\rho_p\hat{e}^p + \sum_{p=1}^{N}\rho_p(\hat{e}^p)^2\right] = 0$$

that leads to the following equality:

$$2\hat{a}_{opt}\left(1 + \sum_{p=1}^{N}\rho_p\right) - 2\sum_{p=1}^{N}\rho_p\hat{e}^p = 0$$

from which we obtain the value of $\hat{a}_{opt}$

$$\hat{a}_{opt} = \frac{\sum_{p=1}^{N}\rho_p\hat{e}^p}{1+\sum_{p=1}^{N}\rho_p} \qquad \text{(A.2)}$$

$\hat{a}_{opt}$ depends on $\hat{e}^p$, weighted by the LR activities and LR emissions factors uncertainties.

Annex 2: Variance of the constrained probability distribution

Let's write the P^{δ} distribution given by (A.1) as a function of $\hat{a}_{opt}$. To do this, we define $\hat{a} = x + \hat{a}_{opt}$ and introduce x in (A.1):

$$P^{\delta} \approx \frac{1}{\sigma_{\hat{A}}}\left(\prod_{p=1}^{N}\frac{1}{\sigma_{\hat{F}^p}}\right)\left(\frac{\delta}{\sqrt{2\pi}}\right)^{N+1} \times exp\left\{-\frac{1}{2\sigma_{\hat{A}}^2}\left[\left(x+\hat{a}_{opt}\right)^2\left(1+\sum_{p=1}^{N}\rho_p\right) - 2\left(x+\hat{a}_{opt}\right)\sum_{p=1}^{N}\rho_p\hat{e}^p + \sum_{p=1}^{N}\rho_p(\hat{e}^p)^2\right]\right\}$$

Which with (A.2) leads to:

$$P^{\delta} \approx \frac{1}{\sigma_{\hat{A}}}\left(\prod_{p=1}^{N}\frac{1}{\sigma_{\hat{F}^p}}\right)\left(\frac{\delta}{\sqrt{2\pi}}\right)^{N+1} \times exp\left\{-\frac{1}{2\sigma_{\hat{A}}^2}\left[x^2\left(1+\sum_{p=1}^{N}\rho_p\right) - \frac{\left(\sum_{p=1}^{N}\rho_p\hat{e}^p\right)^2}{1+\sum_{p=1}^{N}\rho_p} + \sum_{p=1}^{N}\rho_p(\hat{e}^p)^2\right]\right\}$$

$$\approx \frac{1}{\sigma_{\hat{A}}}\left(\prod_{p=1}^{N}\frac{1}{\sigma_{\hat{F}^p}}\right)\left(\frac{\delta}{\sqrt{2\pi}}\right)^{N+1} \times exp\left\{-\frac{1}{2\sigma_{\hat{A}}^2}\left[-\frac{\left(\sum_{p=1}^{N}\rho_p\hat{e}^p\right)^2}{1+\sum_{p=1}^{N}\rho_p} + \sum_{p=1}^{N}\rho_p(\hat{e}^p)^2\right]\right\} \times exp\left\{-\frac{1}{2\sigma^2}x^2\right\} \quad \text{(A.3)}$$

$$\text{with} \quad x = \hat{a} - \hat{a}_{opt}$$

$$\text{and} \quad \sigma^2 = \frac{\sigma_{\hat{A}}^2}{1+\sum_{p=1}^{N}\rho_p} \quad \text{(A.4)}$$

Expression (A.3) shows that P^{δ} is a Gaussian distribution centered on $\hat{a}_{opt}$ with a variance equal to σ^2.

Annex 3: Variability associated to the most likely LR activity

The most likely LR activity depends on the variances of the activity and emission factors. In theory, these variances can vary from 0 to infinity and we distinguish the following extreme cases:

- $\sigma_{\hat{F}^{p*}} = 0$ implies that there is no uncertainty associated to a specific LR emission factor for a pollutant $p*$. The knowledge of $\hat{f}^{p*}$allows to know with certainty all the other LR emission factors and the LR activity. This situation is the one assumed in the methodology proposed by Thunis2016.

- $\sigma_{\hat{A}} = 0$ implies that there is no uncertainty associated to the LR activity. The knowledge of the LR activity also sets automatically every emission factor.

- When $\sigma_{\hat{A}}$ is infinite, the uncertainty of the LR activity is extremely large and when the $\sigma_{\hat{F}^p}$ are infinite, the uncertainty on the LR emission factors is extremely large.

Alternatively, to (A.2), $\hat{a}_{opt}$ can also be expressed in terms of the LR emission of a specific pollutant p^* as follows:

$$\hat{a}_{opt} = \frac{\hat{e}^{p*} + \sum_{p \neq p*} \sigma^2_{\hat{F}p*} / \sigma^2_{\hat{F}p} \times \hat{e}^{p}}{1 + \sigma^2_{\hat{F}p*} / \sigma^2_{\hat{A}} + \sum_{p \neq p*} \sigma^2_{\hat{F}p*} / \sigma^2_{\hat{F}p}} \quad \text{(A.5)}$$

(A.5) shows that:

$$\sigma_{\hat{A}} \to 0 \qquad \Rightarrow \ \hat{a}_{opt} \to 0$$

$$\sigma_{\hat{A}} \to \infty \qquad \Rightarrow \ \hat{a}_{opt} \to \frac{\sum_p \sigma^2_{\hat{F}p*} / \sigma^2_{\hat{F}p} \times \hat{e}^{p}}{\sum_p \sigma^2_{\hat{F}p*} / \sigma^2_{\hat{F}p}}$$

$$\exists p^* \ \ \sigma_{\hat{F}^{p*}} \to 0 \qquad \Rightarrow \ \hat{a}_{opt} \to \hat{e}^{p*}$$

and using (A.2) that:

$$\forall p \ \ \sigma_{\hat{F}^{p}} \to \infty \qquad \Rightarrow \ \hat{a}_{opt} \to 0$$

For each pollutant p^*, the LR total emissions are therefore bounded:

$$\hat{e}^{p*} \in \left[\hat{e}^{min}, \hat{e}^{max}\right] \text{ with } \hat{e}^{min} = \min_p(\hat{e}^{p}) \text{ and } \hat{e}^{max} = \max_p(\hat{e}^{p})$$

$$\text{then } \frac{\sum_p \sigma^2_{\hat{F}p*} / \sigma^2_{\hat{F}p} \times \hat{e}^{p}}{\sum_p \sigma^2_{\hat{F}p*} / \sigma^2_{\hat{F}p}} \in \left[\hat{e}^{min}, \hat{e}^{max}\right]$$

and we can finally deduce that $\hat{a}_{opt}$ is bounded as well:

$$\hat{a}_{opt} \in \left[\min\left(0, \hat{e}^{min}\right), \max(0, \hat{e}^{max})\right] \quad \text{(A.6)}$$