Raport Badawczy

Research Report

RB/28/2016

Learning in national inventory reporting: a bivariate approach

J. Jarnicka, Z. Nahorski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6
01-447 Warszawa
tel.: (+48) (22) 3810100
fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Zbigniew Nahorski

Warszawa 2016

Learning in National Inventory Reporting: A Bivariate Approach

Jolanta Jarnicka, Zbigniew Nahorski

Abstract We analyze the uncertainty in National Inventory Reports (NIR) on greenhouse gases (GHG) emission, provided annually by cosignatories to the UNFCCC and its Kyoto Protocol. Each report contains data on GHG emission from a given year and revisions of past data, recalculated due to improved knowledge and methodology. We consider them realizations of a discrete-time non-stationary stochastic process being a sum of two other processes. Given a data matrix of realizations, we aim to estimate and then analyze the mean values and variances of the component processes as functions of time. Both existence and uniqueness of a solution to this problem are investigated, and algorithms for estimating the mean values and variances (standard deviations) are proposed. The results are presented for a few selected EU-15 countries.

1 Introduction

National Inventory Reports (NIRs) are prepared annually by the Annex I countries since 2001. Some countries started inventorying even a few years earlier, around the mid-90s. The calculations were performed back until 1990, i.e. back to the (unified) basis year of the Kyoto Protocol. It is also a common practice to compile revisions of all or some figures from the earlier years along with the newly prepared inventories. The latter one is recommended, but it is not obligatory.

The aforementioned inventories and revisions can be analyzed to answer several questions. Nahorski & Jęda (2007) used the inventories to assess the inventory uncertainties. The authors showed that, the estimates obtained had quite a good agreement with uncertainty estimates reported at that time by some countries. Hamal (2010), and Marland et al. (2009), raised a question, if inventories and revisions can be used to estimate the time evolution of uncertainties of the NIR emission esti-

Jolanta Jarnicka and Zbigniew Nahorski
Systems Research Institute, Polish Academy of Sciences,
e-mail: `Jolanta.Jarnicka@ibspan.waw.pl`, e-mail: `Zbigniew.Nahorski@ibspan.waw.pl`

mates. This problem has been approached from two directions. Hamal (2010) and Žebrowski et al. (2016) analyzed data for consecutive years, gathering data from different revisions in a year. Nahorski & Jarnicka (2010), Jarnicka & Nahorski (2015, 2016a) analyzed revision sequences from different years, fitting a parametric model.

In this study we present an approach, where the data are analyzed both by years and by revisions. This idea has been already signalized in a conference paper by Jarnicka & Nahorski (2016b), where only the mean values of the errors are analyzed. In the present paper full analysis of both the mean values and standard deviations is provided and their estimates for seven selected EU-15 countries: Austria, Belgium, Denmark, Germany, Finland, Ireland, and the UK are given.

2 Problem motivation and statement

Emissions vary in time and, as time series data, need to be de-trended to fully analyze their properties, e.g. to assess their uncertainty. Following Nahorski & Jęda (2007), this can be done by calculating deviations from the smoothing splines fitted to the most recently revised data series, separately for each country. From the statistical point of view, it means that the results obtained are conditional on the last available inventory report. This way, new sequences of deviations for each revision are obtained. Each of these sequences can be considered a realization of a stochastic nonstationary discrete-time process, denoted by S_y^t, where y identifies a realization and t denotes the discrete time.

Let us assume that, realizations S_y^t are not equally long, i.e. they have the same start time t_0 (i.e. 1990), but various end times y, where $y < Y$, and Y denotes the end time of the longest realization considered. Hence, for realization y it holds $t_0 \leq t \leq y$, and $y \leq Y$. The structure of realizations is graphically presented in Table 1.

$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\ldots$	S_y^{t-1}	S_y^t	S_y^{t+1}	$\ldots$	S_y^y	none	$\ldots$	none
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\ldots$	S_Y^{t-1}	S_Y^t	S_Y^{t+1}	$\ldots$	S_Y^y	S_Y^{y+1}	$\ldots$	S_Y^Y

Table 1 Indexing the realization data.

We call shortly the 'realization y' a realization which ends at time y. The index t determines the place of an element in realization S_y^t, and is called in the sequel the 'position t'. In particular, end time y of the realization y is on the position y. We allow for the lack of some intermediate realizations of the process, as well as the lack of some data points in available realizations, assume however some restrictions.

For simplicity we consider the following notation. By $\mathcal{T}_y$ we denote a set of indices t of non-missing entries in the realization y. By $\mathcal{Y}_t$ we denote a set of indices y of non-missing entries in the position t from all existing realizations. The set $\mathcal{Y}$ consists of all indices y of existing realization, and the set $\mathcal{T}$ – of all indices t, for which there exists a non-missing entry in at least one realization, i.e. $\mathcal{T} = \bigcup_{y \in \mathcal{Y}} \mathcal{T}_y$.

Assume that, the process considered is the sum of two other processes, such that

$$S_y^t = V^t + H_y \tag{1}$$

where V^t and H_y are (possibly dependent) random variables with finite first and second moments. The assumed data structure is illustrated in Table 2.

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
…	$V^{t-1}+H_y$	V^t+H_y	$V^{t+1}+H_y$	…	V^y+H_y	none	…	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
…	$V^{t-1}+H_Y$	V^t+H_Y	$V^{t+1}+H_Y$	…	V^y+H_Y	$V^{y+1}+H_Y$	…	V^Y+H_Y

Table 2 The assumed data structure.

This data structure is motivated by analysis of GHG inventories. An inventory consists of sums of emissions from all atom sources. According to the UNFCCC guidelines, the emission of the atom source E_i is calculated as the product of the source activity A_i, and the source emission factor EF_i

$$E_i = EF_i \cdot A_i.$$

Assuming small deviations ΔEF_i and ΔA_i of both variables, the relative error of the emission can be expressed as the sum of relative errors of the emission factor and the activity

$$\frac{\Delta E_i}{E_i} = \frac{\Delta EF_i}{EF_i} + \frac{\Delta A_i}{A_i}.$$

Thus we consider here two types of uncertainty, the one related to preparing the rough data for reporting, called in the sequel *the data gathering errors*, and the one related to compilation of the final inventory from the rough data, called *the data processing errors*.

To simplify the notation, we put $E(H_y) \overset{\text{def}}{=} \overline{H}_y$, and $E(V^t) \overset{\text{def}}{=} \overline{V}^t$. The estimation problems can be stated as follows.

Problem P1

Given realizations $\{S_y^t; t \in \mathcal{T}_y, y \in \mathcal{Y}\}$ of discrete-time process (1), estimate the mean values $m_t = \overline{V}^t$, $t \in \mathcal{T}$ and $m_y = \overline{H}_y$, $y \in \mathcal{Y}$.

Problem P2

Given realizations as in P1, estimate the variances $d_t = \hat{\sigma}^t$, $t \in \mathcal{T}$ and $d_y = \hat{\sigma}_y$, $y \in \mathcal{Y}$, and covariances $\hat{\rho}_{V^t, H_y}$, $t \in \mathcal{T}_y, y \in \mathcal{Y}$.

3 Existence and uniqueness of a solution to P1

It is quite obvious that, a solution to the Problem P1 may not exist, due to too many missing entries. Therefore we assume that, an additional condition, which we call the *Existence Condition* (EC), is satisfied.

Existence Condition EC1

Let S_{y}^{t}, be realizations of (1). Then the following conditions are met

(i) for all $y \in \mathcal{Y} - Y$ there exists S_{Y}^{y},

(ii) for all $t \notin \{y : y \in \mathcal{Y}\}$ and for all $y \in \mathcal{Y} - Y$ there exists S_{y}^{t}.

EC1 requires that, in addition to the endpoints of realizations, the indicated entries are non-missing. Condition (i) guaranties the existence of all entries of the longest realization, and (ii) assumes the existence of at least one entry for all intermediate positions t in any realization except the longest one.

Remark 1. EC1 is sufficient but not necessary. If EC1 is not satisfied, existence of a solution to P1 can be shown for some practical cases. In particular, if either (i) or (ii) is not satisfied, a partial solution exists.

Observe that, decomposition (1) is not unique. Indeed, for any $C \in \mathbb{R}$ we get

$$S_{y}^{t} = (V^{t} + C) + (H_{y} - C) = (V')^{t} + (H')_{y}. \qquad (2)$$

Since $\overline{V^{t}}$ and $\overline{H}_{y}$ depend on arbitrary constant C, the dependence (2) results in infinitely many solutions to P1. An additional condition is therefore required. We call it the *Uniqueness Condition* (UC).

Uniqueness condition UC1

Let H_{Y} be a random variable, such that $\overline{H}_{Y} < \infty$. Then

$$\overline{H}_{Y} = 0. \qquad (3)$$

The assumption in (3) was motivated by the application considered in Section 8.

Remark 2. Specifying (3), UC1 determines the value of a constant C in (2), provided that decomposition (1) exists.

Proposition 1. *Under* EC1 *and* UC1, *determination of mean values* $m_{t} = \overline{V}^{t}$ *and* $m_{y} = \overline{H}_{y}$ *from* S_{y}^{t} *is unique.*

The proof is straightforward. From the condition UC1 it holds $\overline{H}_{Y} = 0$, so for all $t \in \mathcal{T}_{Y}$ values V^{t} are given. Following the assumptions of the condition EC1 all other value can be found as well.

4 Estimation method

Consider all realizations S_y^t, satisfying EC1 and UC1. We subtract $\overline{V}^t$ from all non-missing S_y^t, $t \in \mathscr{T}$, $y \in \mathscr{Y}$. Then $\overline{H}_y$ can then be estimated as arithmetic means of the differences $S_y^t - \overline{V}^t$, $t \in \mathscr{T}_y$. Similarly, $\overline{V}^t$ can be estimated as arithmetic means of $S_y^t - \overline{H}_y$ over appropriate $y \in \mathscr{Y}_t$. Having subtracted both means $\overline{V}^t$ and $\overline{H}_y$ from each non-missing entry S_y^t, all arithmetic means calculated for each realization and each position in the residual structure should be equal zero. This motivates the following iterative algorithm, consisting of three steps.

Algorithm 1 Estimating mean values $\overline{H}_y$ and $\overline{V}^t$.

```
$\Delta S_y^{t,(0)} \leftarrow S_y^t$
$k \leftarrow 0$
repeat                                                    ▷ Step 1. Compute alternatively mean values
    $k \leftarrow k+1$                                              ▷ over $t$ and $y$, and subtract them
    for all $y \in \mathscr{Y}$ do                                   ▷ from $\Delta S_y^t$ in subsequent iterations
        calculate the means $\overline{H}_y^{(k)}$ from $\Delta S_y^{t,(k)}$, $t \in \mathscr{T}_y$   ▷ until the difference is close to zero.
        $\Delta S_y^{t,(k)} \leftarrow \Delta S_y^{t,(k-1)} - \overline{H}_y^{(k)}$
    end for
    $k \leftarrow k+1$
    for all $t \in \mathscr{T}$ do
        calculate the means $\overline{V}^{t,(k)}$
        $\Delta S_y^{t,(k)} \leftarrow \Delta S_y^{t,(k-1)} - \overline{V}^{t,(k)}$
    end for
until $|\overline{H}_y^{(k)}|, y \in \mathscr{Y}$ and $|\overline{V}^{t,(k)}|, t \in \mathscr{T}$ are small enough
for all $y \in \mathscr{Y}$ do                                       ▷ Step 2. Compute initial estimates
    $\bar{\overline{H}}_y \leftarrow \sum_{\text{odd}\,k} \overline{H}_y^{(k)}$                  ▷ by summing up the partial mean values.
end for
for all $t \in \mathscr{T}$ do
    $\bar{\overline{V}}^t \leftarrow \sum_{\text{even}\,k} \overline{V}^{t,(k)}$
end for
for all $y \in \mathscr{Y}$ do                                       ▷ Step 3. Compute final estimates
    $\hat{\overline{H}}_y \leftarrow \bar{\overline{H}}_y - \bar{\overline{H}}_Y$                                    ▷ that satisfy UC3.
end for
for all $t \in \mathscr{T}$ do
    $\bar{\bar{\overline{V}}}^t \leftarrow \bar{\overline{V}}^t + \bar{\overline{H}}_Y$
end for
```

Proposition 2. *Under assumptions of Proposition 1, Algorithm 1 converges to the solution.*

Proof. To prove the convergence of iterations in Step 1, we first observe that, in each iteration the sum of squares of all entries diminishes. Let $\Delta S_y^{t,(k)}$ be the (t,y)th entry in an iteration starting with an odd index k. Denote the arithmetic mean of all entries of the realization y in the step k by $m_y^{(k)}$, $y \in \mathscr{Y}$, and the one of all entries on the position t by $m_t^{(k)}$, $t \in \mathscr{T}$.

For the sum of squares of all non-missing entries of the realization y in the step $k+1$ we have

$$\sum_{n\in\mathcal{T}_y}(\Delta S_y^{t,(k+1)})^2=\sum_{t\in\mathcal{T}_y}(S_y^{t,(k)}-m_y^{(k)})^2=\sum_{t\in\mathcal{T}_y}(\Delta S_y^{t,(k)})^2-(m_y^{(k)})^2. \tag{4}$$

Similarly, for the sum of squares of all non-missing entries on the position t in the step $k+2$ we get

$$\sum_{y\in\mathcal{Y}_t}(\Delta S_y^{t,(k+2)})^2=\sum_{y\in\mathcal{Y}_t}(\Delta S_y^{t,(k+1)}-m_t^{(k+1)})^2=\sum_{y\in\mathcal{Y}_t}(\Delta S_y^{t,(k+1)})^2-(m_t^{(k+1)})^2. \tag{5}$$

Let us now consider the sum of squares of all non-missing entries after the step $k+2$. From (5) we obtain

$$\sum_{(t,y)\in\mathcal{T}\times\mathcal{Y}}(\Delta S_y^{t,(k+2)})^2=\sum_{t\in\mathcal{T}}\sum_{y\in\mathcal{Y}_t}(\Delta S_y^{t,(k+2)})^2=\sum_{t\in\mathcal{T}}\{\sum_{y\in\mathcal{Y}_t}(\Delta S_y^{t,(k+1)})^2-(m_t^{(k)})^2\}.$$

Proceeding this way and taking into account (4), we get

$$\sum_{y\in\mathcal{Y}}\sum_{t\in\mathcal{T}_y}(\Delta S_y^{t,(k+1)})^2-\sum_{t\in\mathcal{T}}(m_t^{(k)})^2=\sum_{y\in\mathcal{Y}}\{\sum_{t\in\mathcal{T}_y}(\Delta S_y^{t,(k)})^2-(m_y^{(k+1)})^2\}-\sum_{t\in\mathcal{T}}(m_t^{(k)})^2$$

and finally

$$\sum_{(t,y)\in\mathcal{T}\times\mathcal{Y}}(\Delta S_y^{t,(k+2)})^2=\sum_{(t,y)\in\mathcal{T}\times\mathcal{Y}}(\Delta S_y^{t,(k)})^2-\sum_{y\in\mathcal{Y}}(m_y^{(k+1)})^2-\sum_{t\in\mathcal{T}}(m_t^{(k)})^2.$$

Thus

$$\sum_{(t,y)\in\mathcal{T}\times\mathcal{Y}}(\Delta S_y^{t,(k+2)})^2\leq\sum_{(t,y)\in\mathcal{T}\times\mathcal{Y}}(\Delta S_y^{t,(k)})^2.$$

As the sums are nonnegative, the sequence of the sums for increasing k converges. Let us denote the limit value by

$$\sum_{(t,y)\in\mathcal{T}\times\mathcal{Y}}(\Delta S_y^{t,(\infty)})^2.$$

It is easy to notice that, the arithmetic means of all realizations and all positions in $\Delta S_y^{t,(\infty)}$, $t\in\mathcal{T},y\in\mathcal{Y}$ are equal to zero. Otherwise, we could continue iterations to eventually get smaller value of the sum of squares. Hence, the mean values $m_y^{(\infty)}$, $y\in\mathcal{Y}$, and $m_t^{(\infty)}$, $t\in\mathcal{T}$, can be found as the arithmetic means of the differences

$$\Delta S_y^{t,(0)}-\Delta S_y^{t,(\infty)}\quad t\in\mathcal{T},y\in\mathcal{Y}$$

They can also be expressed in the equivalent form, as the series

$$m_y^{(\infty)}=\sum_{k=1}^{\infty}m_y^{(2k-1)}\qquad y\in\mathcal{Y} \tag{6}$$

$$m_t^{(\infty)}=\sum_{k=1}^{\infty}m_t^{(2k)}\qquad t\in\mathcal{T}. \tag{7}$$

Hence, given (6) and (7), the mean values

$$\overline{H}_y=m_y^{(\infty)},\ \text{and}\ \overline{V}^t=m_t^{(\infty)}.$$

According to Proposition 1 the solution is unique, which completes the proof.

5 Estimation of variances and correlations in P2

The variance of the sum (1) is given by the following formula

$$(\sigma_y^t)^2 = (\sigma^t)^2 + (\sigma_y)^2 + 2cov(V^t, H_y), \tag{8}$$

where $(\sigma_y^t)^2$ is the variance of S_y^t, $(\sigma^t)^2$ – the variance of the variable V^t, $(\sigma_y)^2$ – the variance of the variable H_y, and $cov(V^t, H_y)$ is the covariance of the variables V^t and H_y. The variables V^t and H_y are strongly correlated, so it is unreasonable to assume that, their covariance is equal to zero. As the number of values to be estimated is greater than the number of values of $(\sigma_y^t)^2$, the use of conventional methods of estimation of the right hand values in (8) is problematic.

Observe that, any change of $(\sigma_y^t)^2$ or $(\sigma_y)^2$ in (8) can be compensated by the appropriate change of the covariance to keep the left hand value unchanged. To cope with this problem, we consider the least-squares minimum-covariance solutions, in the least squares sense. That is, we propose to look for solutions minimizing the following criterion

$$J_1 = \sum_t \sum_y (cov(V^t, H_y))^2 = \frac{1}{4} \sum_{t \in \mathscr{T}} \sum_{y \in \mathscr{Y}_t} \left[(\sigma_y^t)^2 - (\sigma^t)^2 - (\sigma_y)^2\right]^2.$$

Unfortunately, a solution to this problem is non-unique, since a constant C, (where $C \in \mathbb{R}$) can be added to one of the variances and subtracted from the other one. Taking into account the relationship

$$cov(V^t, H_y) = \rho_y^t \sigma^t \sigma_y,$$

where ρ_y^t denotes the correlation coefficient of variables V^t and H_y, we get

$$(\sigma_y^t)^2 = (\sigma^t)^2 + (\sigma_y)^2 + 2\rho_y^t \sigma^t \sigma_y. \tag{9}$$

Equation (9) will be used to find the estimates of $(\sigma^t)^2$ and $(\sigma_y)^2$ as *the least-squares minimum-correlation solution.*

Problem P2.1.

Find values $(\hat{\sigma}^t)^2$, $t \in \mathscr{T}$, and $(\hat{\sigma}_y)^2$, $y \in \mathscr{Y}$, that minimize over $(\sigma^t)^2$, $t \in \mathscr{T}$, and $(\sigma_y)^2$, $y \in \mathscr{Y}$, the criterion

$$J = \frac{1}{4} \sum_{t \in \mathscr{T}} \sum_{y \in \mathscr{Y}_t} (\rho_y^t)^2 = \frac{1}{4} \sum_{t \in \mathscr{T}} \sum_{y \in \mathscr{Y}_t} \left(\frac{(\sigma_y^t)^2 - (\sigma^t)^2 - (\sigma_y)^2}{\sigma^t \sigma_y}\right)^2 =$$

$$= \frac{1}{4}\sum_{t\in\mathscr{T}}\sum_{y\in\mathscr{Y}_t}\frac{\left((\sigma_y^t)^2-(\sigma^t)^2-(\sigma_y)^2\right)^2}{(\sigma^t)^2(\sigma_y)^2} = \frac{1}{4}\sum_{y\in\mathscr{Y}}\sum_{t\in\mathscr{T}_y}\frac{\left((\sigma_y^t)^2-(\sigma^t)^2-(\sigma_y)^2\right)^2}{(\sigma^t)^2(\sigma_y)^2} \tag{10}$$

subject to $\sigma^t > 0,\ t \in \mathscr{T}$, and $\sigma_y > 0,\ y \in \mathscr{Y}$.

Remark 3. In the minimization domain the criterion function (10) is continuous and bounded below by 0. The limits on the coordinates (for $\sigma^k \to 0$ or/and $\sigma_y \to 0$) are $+\infty$. When additionally sufficiently large upper boundaries, $\sigma^k \leq M$, $\sigma_y \leq M$, are added, a minimum exists. However, it may not be unique. Although the criterion function J seems to be symmetric with respect to σ^t and σ_y, the domains for t and y are different, so they cannot be simply exchanged.

6 Computational procedure

Computing the first partial derivatives of the criterion function J in (10), we have

$$\frac{\partial J}{\partial(\sigma^t)^2} = -\frac{1}{4}\left\{\frac{1}{(\sigma^t)^4}\sum_{y\in\mathscr{Y}_t}\frac{\left((\sigma_y^t)^2-(\sigma_y)^2\right)^2}{(\sigma_y)^2} - \sum_{y\in\mathscr{Y}_t}\frac{1}{(\sigma_y)^2}\right\} \qquad t\in\mathscr{T}; \tag{11}$$

$$\frac{\partial J}{\partial(\sigma_y)^2} = -\frac{1}{4}\left\{\frac{1}{(\sigma_y)^4}\sum_{t\in\mathscr{T}_y}\frac{\left((\sigma_y^t)^2-(\sigma^t)^2\right)^2}{(\sigma^t)^2} - \sum_{t\in\mathscr{T}_y}\frac{1}{(\sigma^t)^2}\right\} \qquad y\in\mathscr{Y}, \tag{12}$$

from where the necessary condition of extreme can be written as follows

$$(\sigma^t)^2 = \sqrt{\frac{\sum_{y\in\mathscr{Y}_t}\frac{\left((\sigma_y^t)^2-(\sigma_y)^2\right)^2}{(\sigma_y)^2}}{\sum_{y\in\mathscr{Y}_t}\frac{1}{(\sigma_y)^2}}} \qquad t\in\mathscr{T} \tag{13}$$

$$(\sigma_y)^2 = \sqrt{\frac{\sum_{t\in\mathscr{T}_y}\frac{\left((\sigma_y^t)^2-(\sigma^t)^2\right)^2}{(\sigma^t)^2}}{\sum_{t\in\mathscr{T}_y}\frac{1}{(\sigma^t)^2}}} \qquad y\in\mathscr{Y}. \tag{14}$$

Obviously, the expressions under the roots are nonnegative, so the real roots (13) - (14) exist.

The necessary conditions form systems of nonlinear equations, and their solution is a rather complicated task. However, for each known set of the values $(\sigma^t)^2, t \in \mathscr{T}$ from (14) we can calculate corresponding optimal values of $(\sigma_y)^2, y \in \mathscr{Y}$ and similarly of $(\sigma^t)^2, t \in \mathscr{T}$.

The above property is needed in the Gauss-Seidel optimization scheme, where, starting with an initial solution, the improved solutions are found by iterative alternative optimization with respect to both groups of variables. In our case, a small modification is possible, i.e. that optimization with respect to separate variables is done analytically, which simplifies and accelerates computations.

The Gauss-Seidel procedure converges to the optimal solution, if it is unique, see e.g. [8]. However, it is not know if a solution to systems obtained from the necessary condition is unique. Moreover, if there are more stationary points, there may exist also points other then the minimum values. In such a case the minimization procedure proposed above may fail.

Given the estimates $\hat{\sigma}_y^t$ and the (sub)optimal solutions $\hat{\sigma}^t, t \in \mathscr{T}; \hat{\sigma}_Y, y \in \mathscr{Y}$, the estimates of ρ_y^t can be calculated from (9) as

$$\hat{\rho}_y^t = \frac{1}{2}\left(\frac{(\hat{\sigma}_y^t)^2}{\hat{\sigma}^t \hat{\sigma}_y} - \frac{\hat{\sigma}^t}{\hat{\sigma}_y} - \frac{\hat{\sigma}_y}{\hat{\sigma}^t}\right). \tag{15}$$

7 Analysis of convergence

To analyze the convergence, we calculate the second derivatives. We have

$$J_{yy} = \frac{\partial^2 J}{\partial((\sigma_y)^2)^2} = \frac{1}{2}\frac{1}{(\sigma^t)^2}\sum_{y \in \mathscr{Y}_t} \frac{\left((\sigma_y^t)^2 - (\sigma_y)^2\right)^2}{(\sigma_y)^2} \qquad y \in \mathscr{Y} \tag{16}$$

and at the optimum, where (12) is satisfied it holds

$$J_{yy} = \frac{(\hat{\sigma}_y)^2}{2}\sum_{t \in \mathscr{T}_y} \frac{1}{(\hat{\sigma}^t)^2} \qquad y \in \mathscr{Y}. \tag{17}$$

Similarly

$$J_{tt} = \frac{\partial^2 J}{\partial((\sigma^t)^2)^2} = \frac{1}{2}\frac{1}{(\sigma_y)^2}\sum_{t \in \mathscr{T}_y} \frac{\left((\sigma_y^t)^2 - (\sigma^t)^2\right)^2}{(\sigma^t)^2} \qquad t \in \mathscr{T} \tag{18}$$

and at the optimum

$$J_{tt} = \frac{(\hat{\sigma}^t)^2}{2}\sum_{y \in \mathscr{Y}_t} \frac{1}{(\hat{\sigma}_y)^2} \qquad t \in \mathscr{T}. \tag{19}$$

The mixed second order derivatives are equal to zero

$$J_{yz} = \frac{\partial^2 J}{\partial(\sigma_y)^2(\sigma_z)^2} = 0 \qquad z \neq y \quad y,z \in \mathscr{Y} \tag{20}$$

$$J_{ts} = \frac{\partial^2 J}{\partial(\sigma^t)^2(\sigma^s)^2} = 0 \qquad s \neq t \quad t,s \in \mathscr{T} \tag{21}$$

$$J_{yt} = J_{ty} = \frac{\partial^2 J}{\partial(\sigma^t)^2(\sigma_y)^2} = \frac{1}{4}\frac{(\sigma_y^t)^4 - (\sigma^t)^4 - (\sigma_y)^4}{(\sigma^t)^4(\sigma_y)^4} \qquad t \in \mathscr{T}_y, y \in \mathscr{Y} \tag{22}$$

First, observe that, optimizing only with respect to $(\sigma_y)^2$, the Hessian matrix of the second partial derivatives is diagonal with the positive values J_{yy} on the diagonal. The matrix J_{yy} is positive defined and then the point satisfying the necessary conditions for $y \in \mathscr{Y}$ is the unique optimum. Similarly with optimization with respect to $t \in \mathscr{T}$. Hence, in each step of the above modification of the Gauss-Seidel method the unique optimum with respect to either the variables indexed by $t \in \mathscr{T}$ or $y \in \mathscr{Y}$ is obtained.

The Hessian matrix of the second derivatives $\mathbf{H}$ for all variables has the block structure

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix} \tag{23}$$

where the matrices $\mathbf{A}$ and $\mathbf{D}$ are diagonal with the values I_{tt} and I_{yy} on the diagonals, respectively, while the matrix $\mathbf{B}$ has entries $J_{ty} = J_{yt}$ for all $t \in \mathscr{T}_y, y \in \mathscr{Y}$. The positive definiteness of the matrix $\mathbf{H}$ can be analyzed using several methods, for example the Sylvester criterion or Cholesky factorization. However, this cannot be done using analytic methods.

8 Estimation results

The NIR data on CO_2 emissions without LULUCF (Land-use, Land-use-change, and Forestry) for seven EU-15 countries: Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK, were analyzed. These data span from 1990 to 2014, with revisions made every year from 2001 to 2014. The smoothing splines were used to de-trend the last emission data series. Figure 1 presents the trajectories of the most recent emission data reported (i.e. the longest realizations, for the year $Y = 2014$) with fitted smoothing splines.

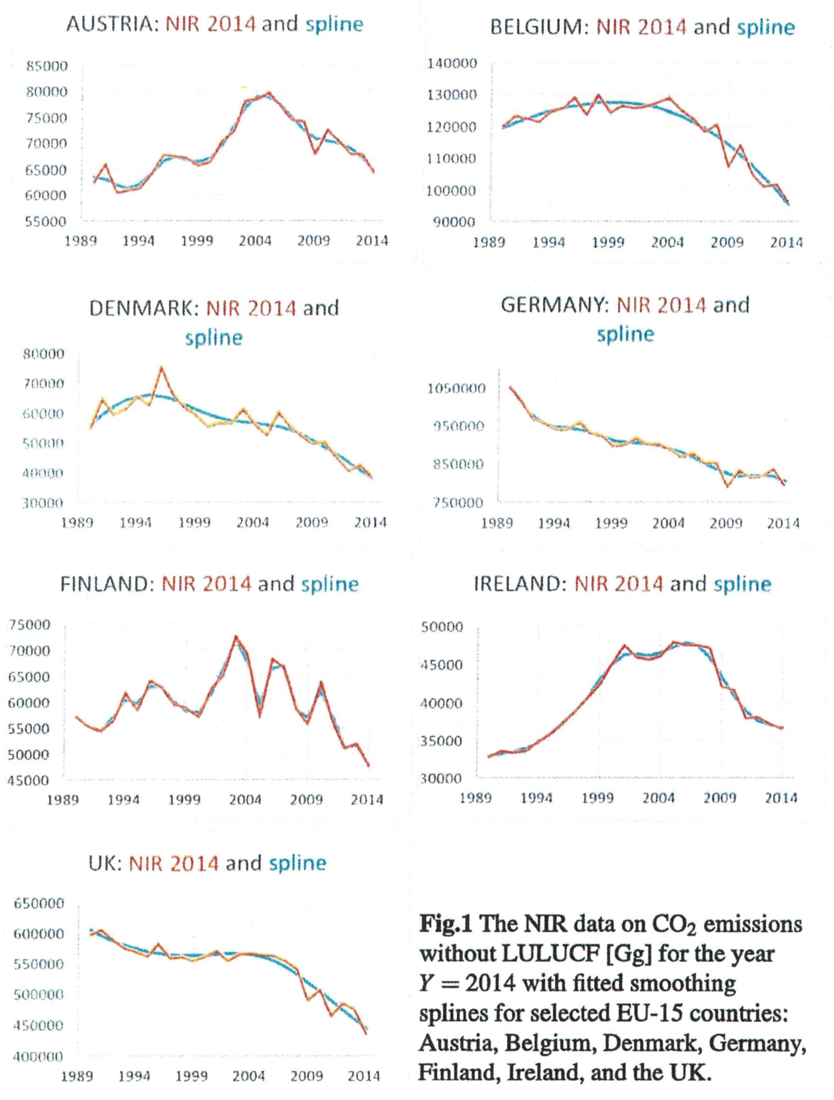


Fig.1 The NIR data on CO_2 emissions without LULUCF [Gg] for the year $Y = 2014$ with fitted smoothing splines for selected EU-15 countries: Austria, Belgium, Denmark, Germany, Finland, Ireland, and the UK.

For each of the countries considered, the data matrix was then obtained by subtracting the smoothing splines from all the earlier revisions, giving S_y^t, for $y < Y$. The converted matrices satisfy EC1, so Algorithm 1 from Section 4 was applied. This way, estimates of mean values $\overline{V}^t$ for positions t from 1990 to 2014, and of mean values $\overline{H}_y$ for revisions y from 2001 to 2014 were obtained. It could be observed that, the values $m_y^{(k)}$ and $m_t^{(k)}$ in consecutive iterations converged quickly to zero.

After five to seven iterations the estimates of $m_y^{(k)}$ were less than 0.1 and of $m_t^{(k)}$ practically equaled zero. Hence, only a few iterations provided sufficient accuracy. The trajectories of estimated mean values $\overline{V}^{t}$, i.e. of the data gathering errors, are depicted in Figure 2, while the estimates of $\overline{H}_y$, i.e. of the data processing errors, are presented in Figure 3.

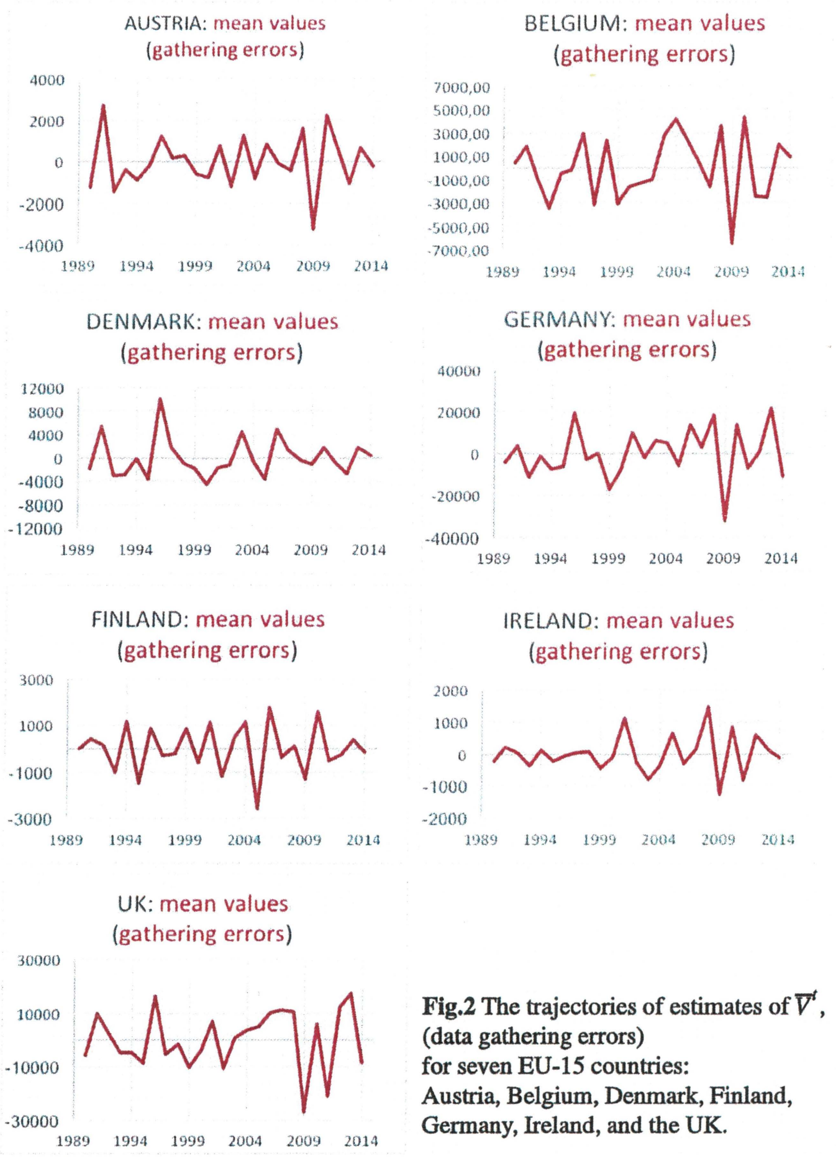


Fig.2 The trajectories of estimates of $\overline{V}^{t}$, (data gathering errors) for seven EU-15 countries: Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK.

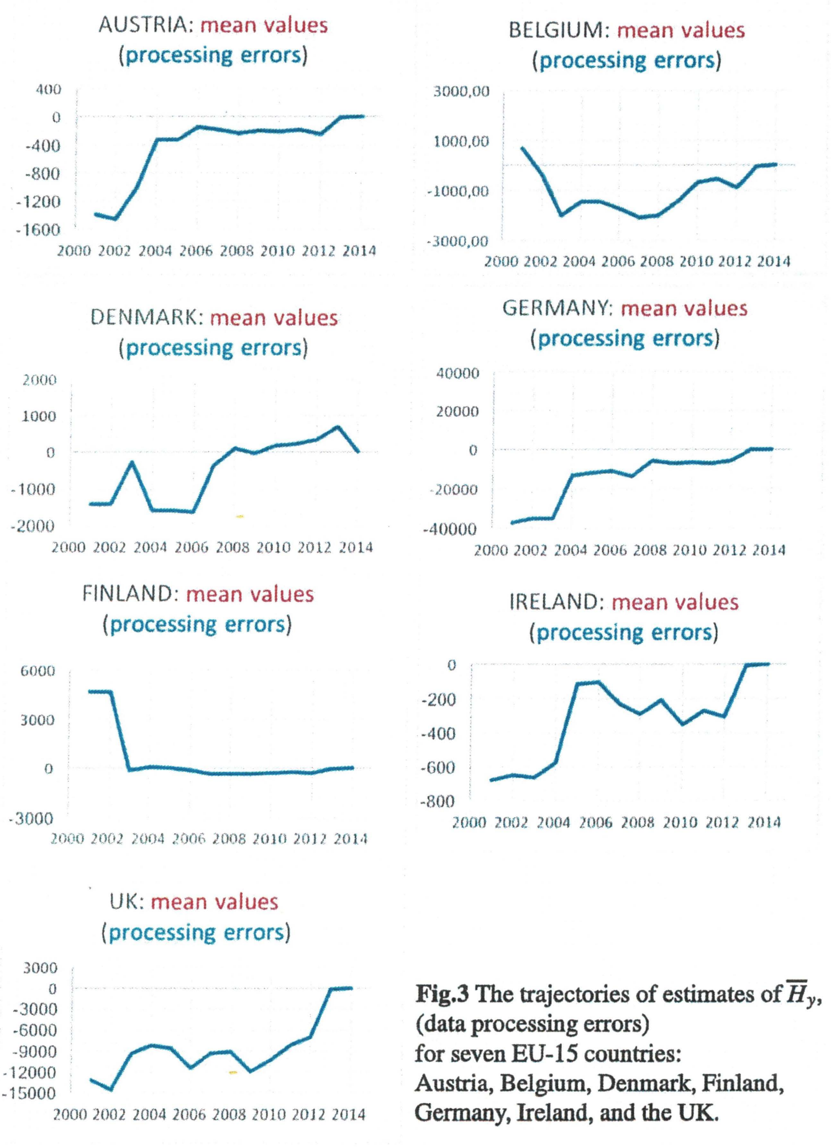


Fig.3 **The trajectories of estimates of $\overline{H}_y$, (data processing errors) for seven EU-15 countries: Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK.**

The second part of the analysis was devoted to estimate standard deviations σ_y (gathering errors) and σ_t, using the method described in Sections 6 and 7. At first, the squared residuals after de-meaning the detrended data S_y^t were treated as $(\sigma_y^t)^2$. The starting point of the algorithm was there one, were the estimates of $(\sigma_y)^2$ equal to half of the maximum value of $(\sigma_y^t)^2$, $t \in \mathcal{T}_y$.

Applying the method, we could observe that, in the first two-three iterations the minimization of the criterion J was very fast. After that, the algorithm switched to maximization. Convergence of the criterion values became much slower, while estimates of the variances were still changing (mainly decreasing). The shape of the curves did not changed much, and in some cases the algorithm switched again to the minimization, but the criterion value was never smaller than the one obtained nafter early iterations. This may suggest that, the minimized function has a rather complicated shape, posibly with more local minimum values, although the curve in the neighbourhood of the minimum values may be rather flat.

The estimates of the standard deviation for the best early minimum values are depicted in Figure 4 for the gathering variables and in Figure 5 for the processing ones.

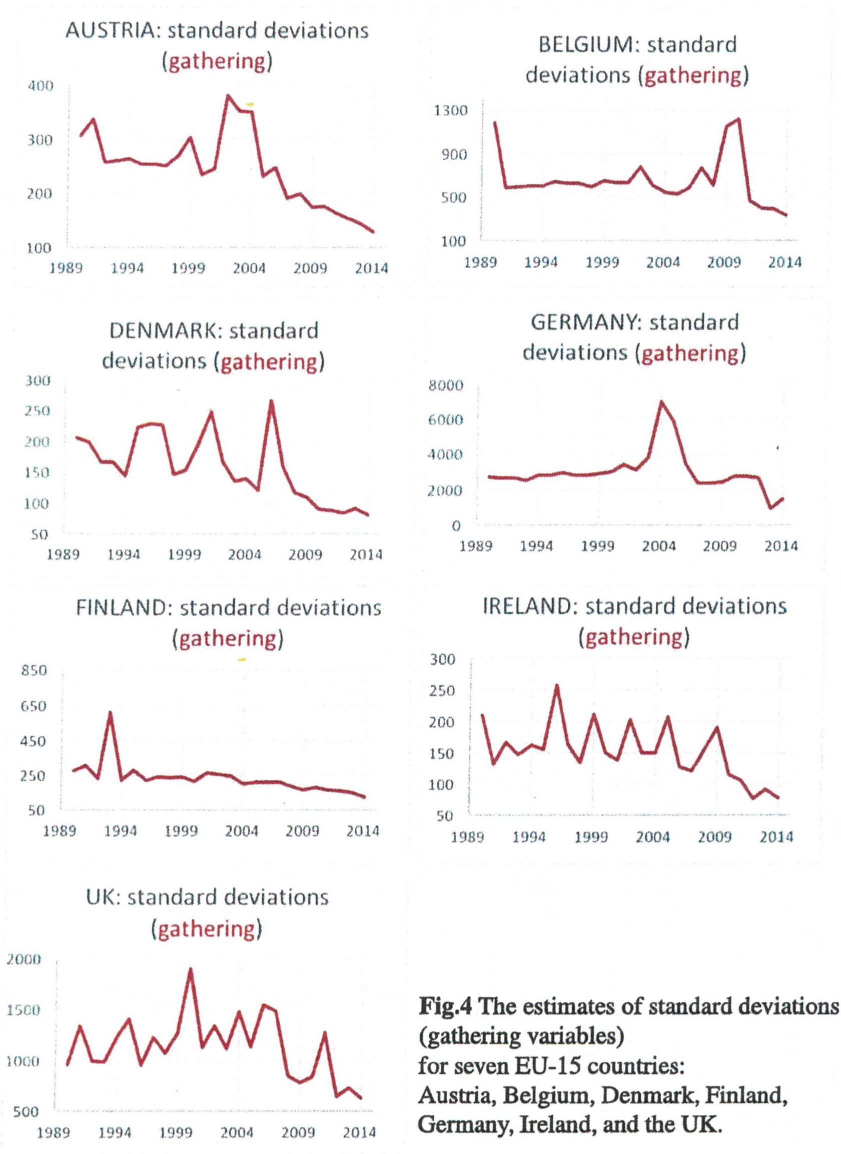


Fig.4 The estimates of standard deviations (gathering variables) for seven EU-15 countries: Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK.

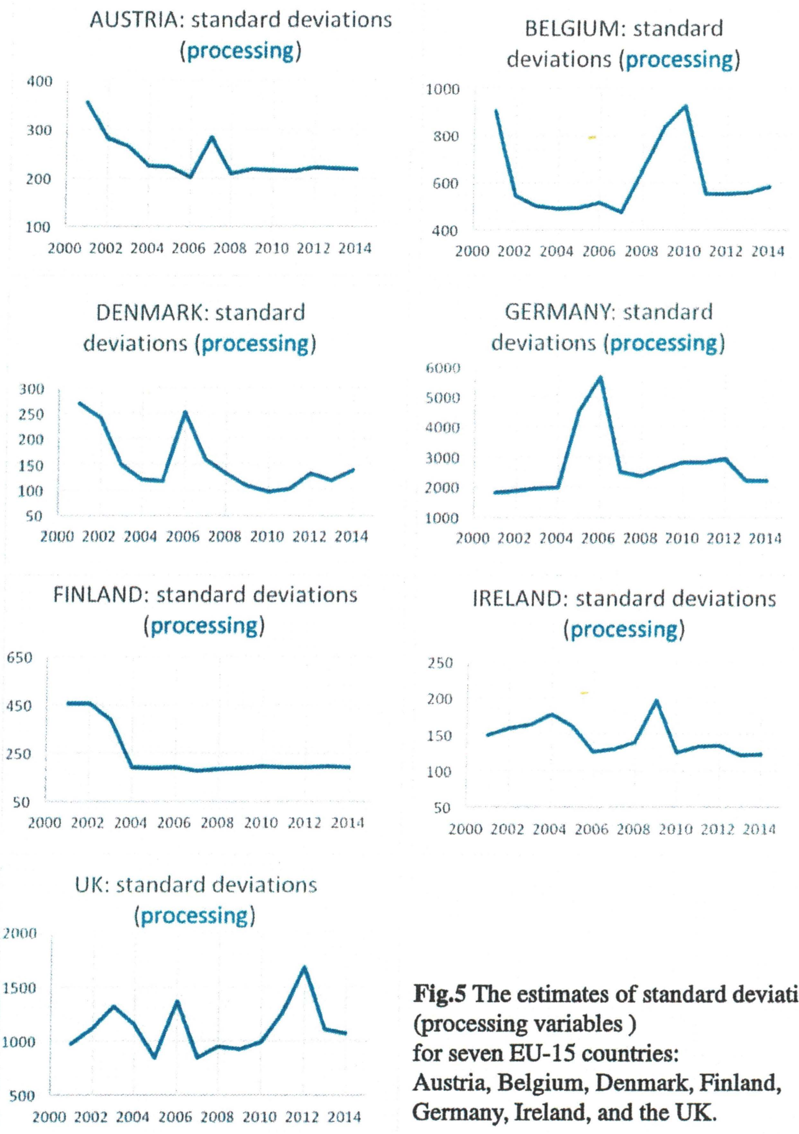


Fig.5 The estimates of standard deviations (processing variables)
for seven EU-15 countries:
Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK.

Additionally, in Figures 6 and 7 the mean values estimated above, along with subtracted and added estimates of standard deviations are presented. The results depicted there can be considered a sort of confidence intervals.

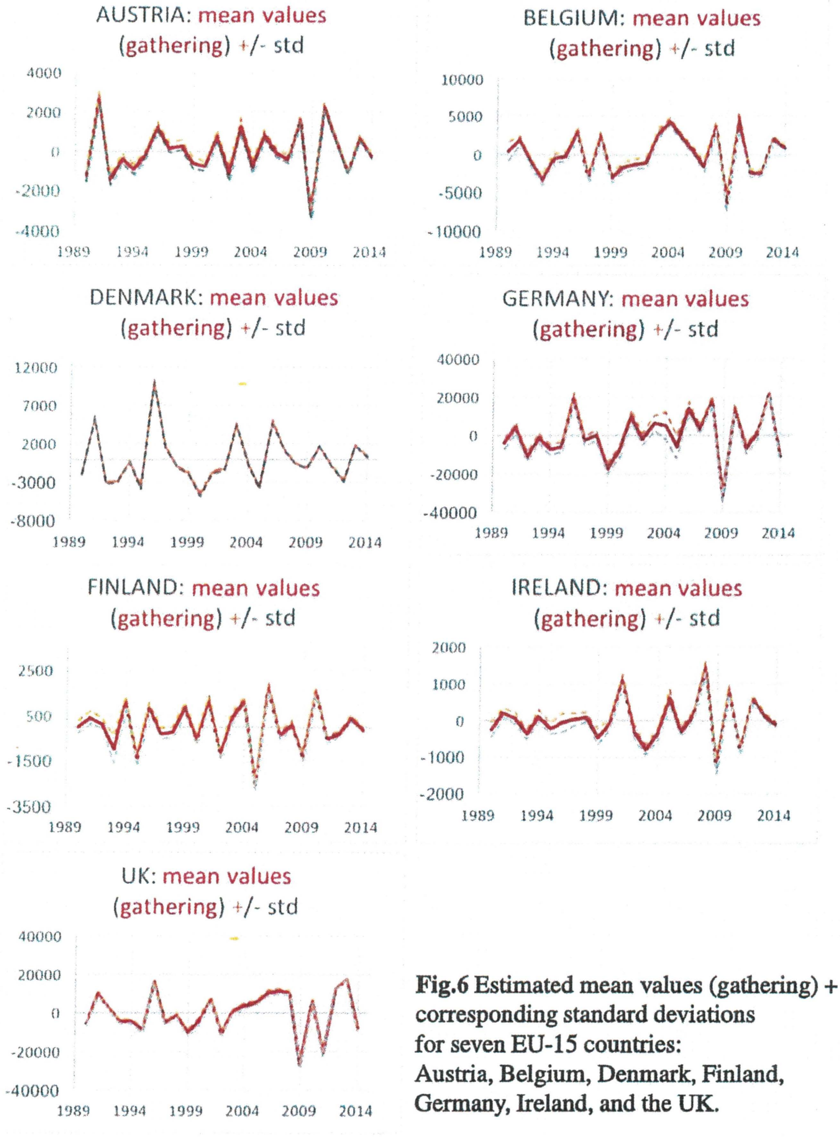


Fig.6 Estimated mean values (gathering) + - corresponding standard deviations for seven EU-15 countries: Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK.

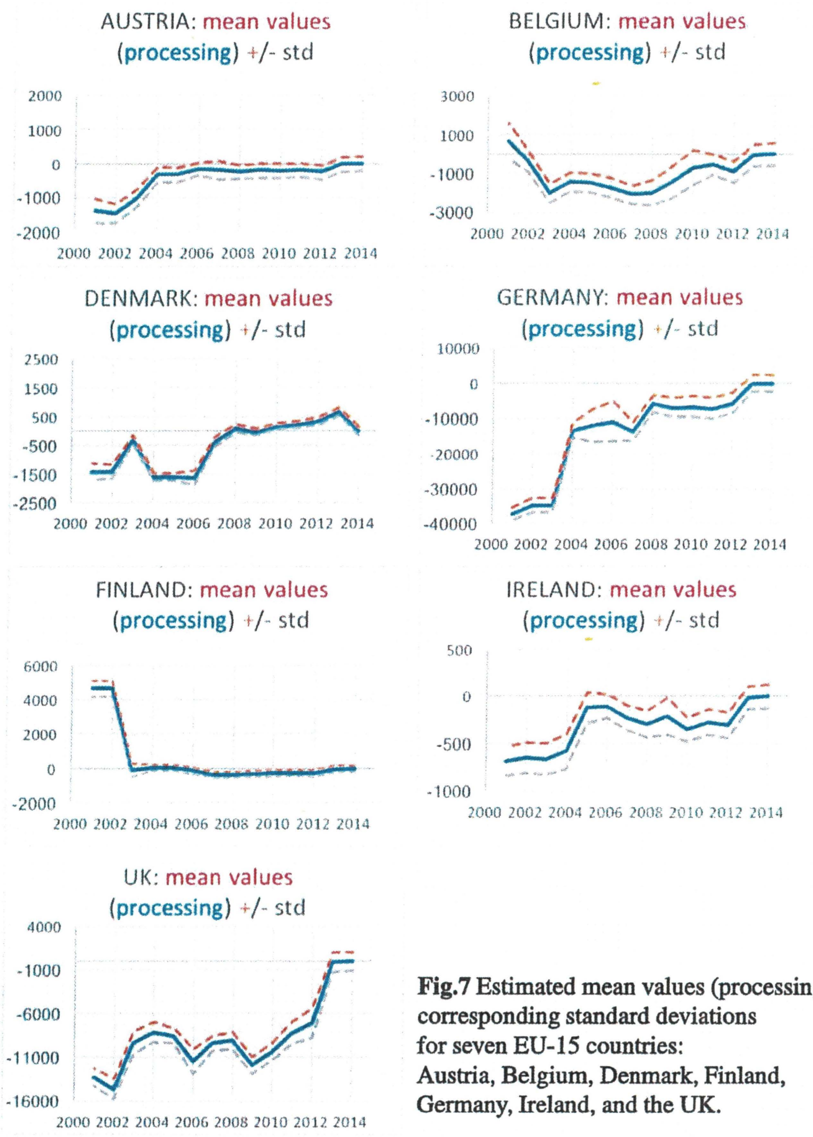


Fig.7 Estimated mean values (processing) + - corresponding standard deviations for seven EU-15 countries: Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK.

9 Discussion of results

While the estimates of $\overline{H}_y$ (for data processing) presented in Figure 3 form reasonably smooth curves, the estimates of $\overline{V}^t$ for data gathering (Figure 2) seem to be purely random. Due to the uniqueness condition UC1, it holds

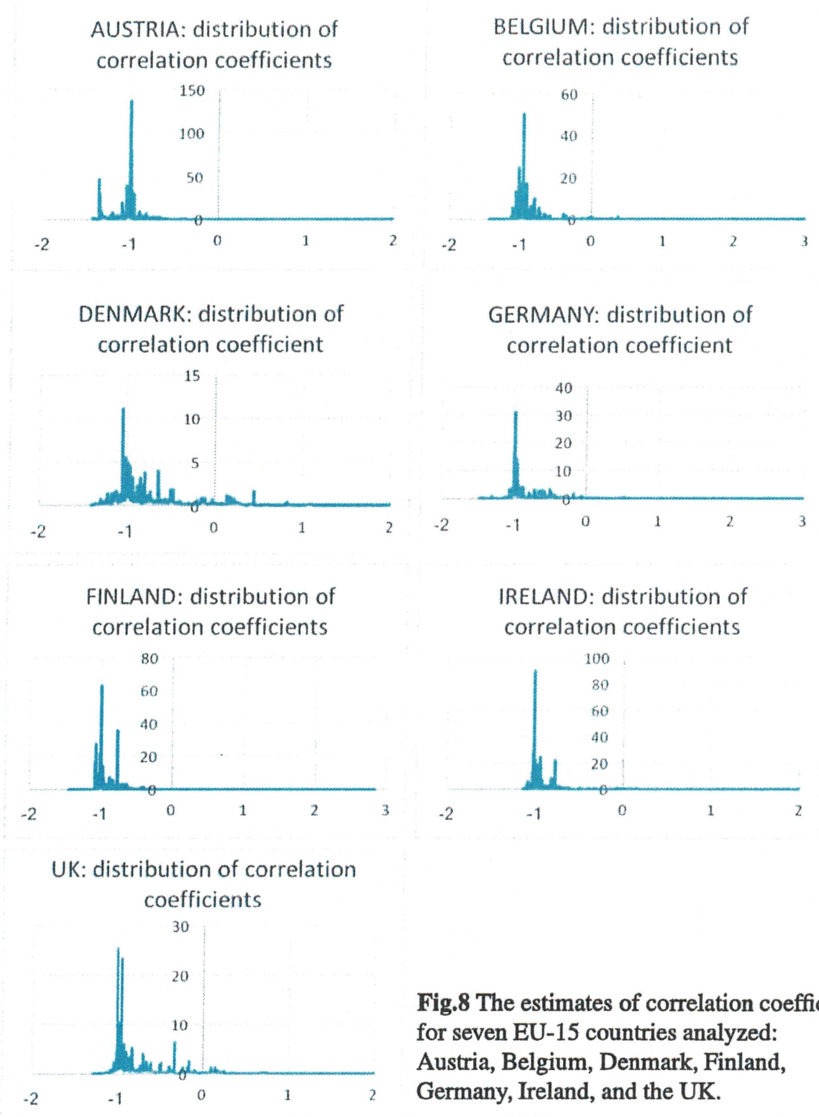
$$\sum_{t \in \mathcal{T}_Y} V^t = 0.$$

It is easy to observe that is satisfied for $\hat{\overline{V}}^t$ in all cases. Hence, analyzing the mean values, no learning feature was observed in gathering the rough data. On the other hand, some learning can be attributed to the process of processing the rough data, which is more or less visible in all cases. It is worth mentioning that, for most countries the mean values of processes reach the final value from below. A noticeable exception in that rule is Finland, and to some extend Denmark.

Slightly opposite, the standard deviations of data processing often decrease in the early 2000th, but then start to be rather constant (Figure 5). The standard deviations of data gathering (Figure 4) may be quite volatile, but tend to decrease in recent years.

In other words, data processing became more precise in time, but its accuracy did not improve since beginning of 2000th. Precision of data gathering did not change, but its accuracy improved in 2000th. In both cases, some learning is manifested, although of different nature: in precision for data processing, and in accuracy in data gathering.

The results presented for the standard deviations are, however, still tentative. The minimized functions were not yet explored well enough and it is not clear if the minimum values obtained are global or only local. Moreover, the results obtained from estimation of the correlation coefficients need some interpretation. Distributions of the estimates, depicted in Figure 8, show that the coefficients gather close to -1. It may suggest that, there exist some relationship between variables in the minimized function. Answering these questions however requires further studies.



AUSTRIA: distribution of correlation coefficients

BELGIUM: distribution of correlation coefficients

DENMARK: distribution of correlation coefficient

GERMANY: distribution of correlation coefficient

FINLAND: distribution of correlation coefficients

IRELAND: distribution of correlation coefficients

UK: distribution of correlation coefficients

Fig.8 The estimates of correlation coefficients for seven EU-15 countries analyzed: Austria, Belgium, Denmark, Finland, Germany, Ireland, and the UK.

10 Conclusions

We presented methods of estimating mean values and variances (standard deviations) of the data gathering and data processing errors in the National Inventory Reports on GHG emissions, containing data from a given year and revisions of past data. We aimed at estimating time evolution of the uncertainty, taking into account all data revised in consecutive years. For this, we calculated deviations of revisions from the smoothing spline fitted to the latest data reported, treating the deviations obtained as realizations of a non-stationary process. We focused on estimation of the mean values and variances of these deviations.

It is assumed that uncertainties related to data gathering and processing, enter additively into the combined data, which is well motivated by the way the uncertainties appear in the inventory reports. Moreover, it is assumed that, only the latter uncertainty is affected during preparation of new revisions, while the former ones are only attributed to the year when the emissions were inventoried for the first time.

The problem was first analyzed from mathematical point of view. It was pointed that, in general a solution to the separation-of-means problem may not exist, and if it does, it is not unique. Conditions for existence and uniqueness of mean values were introduced, and the algorithm for solving the problem was proposed.

To separate the variances, the least-squares minimum-correlation estimates were introduced. This idea leads to minimization of a nonlinear criterion function consisting of the sum of squares of the correlation coefficients. It was possible for it to find separately the unique optimal solutions for each argument from solving the necessary optimization conditions. This encouraged us to use a variant of the Gauss-Seidel optimization method. However, analysis of the Hessian matrix of the second order derivatives proved too complicated to draw any decisive conclusions on global solutions characterization.

The results of application to the NIR data for Austria, Belgium, Denmark, Germany, Finland, Ireland, and the UK, were presented. The mean-separation method worked well and converged to a solution close enough to the optimal one in only a few iterations. The variance-separation method was able to find a local minimum in two or three iterations. But the shape of the minimized function proved to be complicated, possibly with many local minimum values. This part of the study requires further research.

References

1. K. Hamal (2010) Reporting GHG emissions: Change in uncertainty and its relevance for detection of emission changes. IIASA, Laxenburg, Interim Report IR-10-003.
2. J. Jarnicka and Z. Nahorski (2015) A method for estimating time evolution of precision and accuracy of greenhouse gases inventories from revised reports, *in Proc. 4th Intl Workshop on Uncertainty in Atmospheric Emissions,* Kraków, Poland, 97–102.

3. J. Jarnicka and Z. Nahorski (2016a) Estimation of temporal uncertainty structure of GHG inventories for selected EU countries, *Federated Conference on Computer Science and Information Systems*, Gdańsk, Poland, 11 - 14 September.
4. J. Jarnicka and Z. Nahorski (2016a) Estiation of means in a Bivariate Discrete-Time Process. Prepr. 14th Conference of the Polish Operational and Systems Research Conference, Warsaw, 12-14 Oct 2016.
5. G. Marland, K. Hamal, and M. Jonas (2009), How uncertain are estimates of CO_2 emissions? *Journal of Industrial Ecology*, **13**, 1, 4–7.
6. Z. Nahorski and J. Jarnicka (2010) Modeling uncertainty structure of greenhouse gases inventories, SRI PAS, Warsaw, Report RB/11/2010, unpublished.
7. Z. Nahorski and W. Jęda (2007), Processing national CO_2 inventory emission data and their total uncertainty estimates, *Water, Air, and Soil Pollution: Focus*, 7, 4-5, 513–527.
8. I.Z. Zangwill (1969) *Nonlinear Programming: A Unified Approach.* Prentice-Hall, Englewood Cliffs, NJ.
9. P. Żebrowski, M. Jonas, and E. Rovenskaya (2015) Assessing the improvement of greenhouse gases inventories: can we capture diagnostic learning? *in Proc. 4th Intl Workshop on Uncertainty in Atmospheric Emissions,* Kraków, Poland, 90–96.

Appendix

Here we present an alternative analysis of uniqueness, which uses an analytic geometry approach. The derivations do not lead to a final conclusion, so it is still rather another motivation of a suspicion that the solution of the problem P2.1 may not be unique.

Before going to analysis of the full case, we start with a simplified one. To simplify notation, we denote $x_t = (\sigma^t)^2$, $z_y = (\sigma_y)^2$, $S_{ty} = (\sigma_y^t)^2$. We know that a minimum of the criterion function (10) exists. Let us assume that the value of a (t,y) element of the sum for this minimal point is μ_{ty}. We look for values x_t, z_y satisfying the equation

$$\frac{1}{4}\frac{(S_{ty}-x_t-z_y)^2}{x_t z_y} = \mu_{ty} \qquad \mu_{ty} \geq 0 \tag{24}$$

After few simple manipulations we get

$$(x_t - S_{ty})^2 + (z_y - S_{ty})^2 + 2(1-2\mu_{ty})x_t z_y = S_{ty}^2$$

Then, the points that give the same value μ_{ty} of the criterion satisfy an algebraic equation of the second degree. Hence, they lie on a conic curve. Translation of coordinates $v_{ty} = x_t - S_{ty}$ and $w_{ty} = z_y - S_{ty}$ yields

$$v_{ty}^2 + w_{ty}^2 + 2(1-2\mu_{ty})v_{ty}w_{ty} + 2(1-2\mu_{ty})S_{ty}(v_{ty}+w_{ty}) = [1-2(1-2\mu_{ty})]S_{ty}^2 \tag{25}$$

Let us notice that for $\mu_{ty} < \frac{1}{4}$ the left hand side is negative while the right hand side is positive. Hence, the above equation has no real solutions in this case. This also means that it is impossible to get the zero value of the criterion function (10).

Calculating the discriminant we have

$$\triangle = 1 - (1 - 2\mu_{ty})^2 = 4\mu_{ty}(1 - \mu_{ty}) \tag{26}$$

Hence:
if $\triangle > 0$ (or $\frac{1}{4} \leq \mu_{ty} < 1$), then the curve is an ellipse;
if $\triangle = 0$ (or $\mu_{ty} = 1$), then the curve is a parabola;
if $\triangle < 0$ (or $\mu_{ty} > 1$), then the curve is a hyperbola.
Hence, for $\mu_{ty} \geq \frac{1}{4}$ there is continuum of points satisfying (25).

Now, considering the general case, we have

$$\sum_{y \in \mathscr{Y}} \sum_{t \in \mathscr{T}_y} \left[\frac{1}{4} \frac{(S_{ty} - x_t - z_y)^2}{x_t z_y} - \mu_{ty} \right] = 0 \tag{27}$$

which can be transformed to

$$\sum_{y \in \mathscr{Y}} \sum_{t \in \mathscr{T}_y} \frac{(x_t - S_{ty})^2 + (z_y - S_{ty})^2 + 2(1 - 2\mu_{ty})x_t z_y - S_{ty}^2}{x_t z_y} = 0$$

and then to

$$\sum_{y \in \mathscr{Y}} \sum_{t \in \mathscr{T}_y} \left\{ \left[(x_t - S_{ty})^2 + (z_y - S_{ty})^2 + 2(1 - 2\mu_{ty})x_t z_y - S_{ty}^2 \right] \prod_{t \in \mathscr{T}_y} x_t \prod_{y \in \mathscr{Y}_t} z_y \right\} = 0 \tag{28}$$

Hence, the set of points minimizing the criterion function forms now a complicated hypersurface. In particular, the intersection of all quadric surfaces (25) belongs to it, but possibly not only. From the assumption, at least one real point satisfying the equation exists. The question is how many they are. Analysis of this case seems to be a complicated task, but it seems very likely, that the minimal points form a set of the cardinality of continuum.