Statistical methods to estimate the mean annual Emission Factor of a fuel gas stream with variable composition.

DRAFT Stijn Bierman, GSNL-PTX D/S

May 25, 2020

Abstract

Trustworthy and precise estimates of greenhouse gas emissions of refineries are essential to measure improvements in terms of reduced net emissions, and for regulatory compliance. For example, refineries in The Netherlands which want to take part in CO_2 emissions trading are required by the Dutch Emissions Authority (NEA) to report estimates of the annual average Emission Factor (EF; the amount of produced CO_2 per amount of combusted fuel gas) of each refinery fuel gas.

The composition of refinery fuel gases, and therefore also the EF, varies with time. The EF of a sample, taken from the flow of a refinery fuel gas, can be measured in the laboratory. The cost per sample is relatively high, and a large number of samples is typically required to estimate the annual average EF with sufficient precision. For the same number of laboratory samples, more precise estimates can be obtained if an auxiliary variable is available which correlates strongly enough with laboratory measurements of EF and can be measured at low cost and high frequency. Ideally, an auxiliary variable can be measured continuously using an online measurement device which is permanently in place.

An overview is given of statistical recipes (estimators) which yield estimates of the annual average EF and associated precision, based on laboratory measurements alone or on a combination of laboratory measurements and one or more continuously monitored auxiliary variables. Estimators are compared in terms of their uncertainty, measured by the width of the 95% confidence interval of the mean, and their coverage probability, i.e. the proportion of cases in which the confidence interval contains the true average EF. This is done using a simulation study.

The uncertainty of the estimate of the annual average EF depends on sample size as well as on the process variability in EF The use of an auxiliary variable greatly increases the precision of estimates. Bootstrap estimators have good performance, and can readily be extended to multiple regression in case multiple auxiliary variables are required to get predicted EF which correlates strongly with EF.

In a study of a fuel gas in Pernis, EF was found to correlate strongly with Stoichiometric Air Requirement (SAR) and Lower Heating Value (LHV). Models which combine either SAR or LHV with molar mass can predict EF to a high precision. Molar mass is currently already measured online. If additional online measurements can be done which correlate with either SAR or LHV (or which measure these variables directly), such as a calorimeter or oxygen content, then this could greatly improve the estimates of annual average EF for fuel gas in Pernis.

Contents

List of Figures

List of Tables

1 Introduction

Trustworthy and precise estimates of greenhouse gas emissions of refineries and petrochemical plants are essential to measure reductions in net emissions per amount of produced product and to demonstrate regulatory compliance. For example, refineries in The Netherlands which want to take part in CO_2 emissions trading, are required by the Dutch Emissions Authority (NEA) to report estimates of the annual average Emission Factor (EF; the amount of produced CO_2 per amount of combusted fuel gas; see e.g. Shires et al. [2009]) of each fuel gas. For each fuel gas stream, the annual average EF can be multiplied with the annual total amount of combusted fuel gas to yield an estimate of the total amount of CO_2 produced. The NEA requires these estimates of annual average EF to have an associated minimum precision.

The EF of a single sample, taken from the flow of a fuel gas, can be measured in the laboratory. The cost per sample is relatively high, and a large number of samples may be needed to estimate the annual average EF with sufficient precision. A well known and potentially highly efficient way to more precise estimates, for the same number of laboratory samples, is to use one or more auxiliary variables which correlate with laboratory measurements of EF and which can be measured at lower cost and higher frequency (see e.g. chapters 6 and 7 in Cochran [1977] and chapters 7 and 8 in Thompson [2012]). Ideally, an auxiliary variable correlates strongly with EF and can be measured continuously using an online measurement device which is permanently in place. The auxiliary variables are not of direct interest themselves, but can help in obtaining estimates with higher precision.

A large body of scientific literature exists on how to estimate means or totals based on samples and auxiliary variables. Excellent overviews are given by Cochran [1977] and Thompson [2012]. Few people are familiar with this literature, and there are no universally applicable rules which produce valid and efficient estimates in all situations. It is not straightforward to select an appropriate statistical procedure for a given situation, and to transparently demonstrate that the assumptions are met for the method to yield valid estimates. It seems worthwhile therefore to build up a body of knowledge on appropriate statistical methodology which may be used to obtain such estimates.

This report provides and overview of statistical recipes (estimators) which yield estimates of the annual average EF and associated precision, based on laboratory measurements alone or on a combination of a continuously monitored auxiliary variable and laboratory measurements. The estimators are applied to synthetic data sets, generated under scenarios which capture a range of different situations which may be encountered in practice. The scenarios are defined in terms of:

- The number of laboratory measurements
- The correlation (strength of relationship) between the auxiliary variable and the laboratory measurements
- The variability in EF, auxiliary variable and flow rate measurements (due to a combination of process variability and measurement errors)
- The correlation (strength of relationship) between flow rate and EF, and between flow rate and auxiliary variable

Estimators are compared in terms of their precision, measured by the width of the 95% confidence interval of the mean, and their coverage probability, i.e. the proportion of cases in which the confidence interval contains the true average EF. If an estimator produces confidence intervals with a coverage probability (substantially) below 95%, then the precision estimates based on this estimator are unjustifiably high and as such potentially misleading. For each scenario, the most efficient estimator yields an estimate with the highest precision whilst maintaining an acceptable coverage probability.

2 Estimators

Let y_i be the EF as measured in the laboratory on sample i; $i = 1, 2, \ldots, n$, where n denotes the number of samples taken from the fuel gas flow during the year (the sample size). Triples of laboratory (EF; y_i), auxiliary (x_i) and flow rate (b_i) measurements at the time that sample i was taken are denoted by: (y_i, x_i, b_i). The assumption is that measurements of the flow rate and auxiliary variable are always available at the time a sample was taken. Let X_j and B_j be a measurement of an auxiliary variable and flow rate respectively, where $j = 1, 2, \ldots, k$ is the total number of measurements of this auxiliary variable taken on the fuel gas flow during the year. Typically, $k \gg n$, because the auxiliary variable and flow rate can be measured at high frequency at relatively low cost.

Below, a number of statistical procedures (estimators) are given for estimating the annual average EF, $\bar{y}$, and the lower and upper bound of the 95% confidence interval, $\bar{Y}_L$ and $\bar{Y}_U$ respectively. The width U of the confidence interval is given by $U = \bar{Y}_L - \bar{Y}_U$. The relative precision,%U, of the estimate $\bar{y}$ is given by half the width of the confidence interval divided by the estimate of the annual average: $\%U = 100\frac{U}{2\bar{y}}$.

2.1 Without auxiliary variable: Simple Random Sampling (SRS)

If no auxiliary variable is available, and assuming that a representative sample of the fuel gas flow has been taken, the annual average EF can be estimated as:

$$\bar{y} = \Sigma_{i=1}^{n}(y_i w_i), \tag{1}$$

where w_i is given by:

$$w_i = \frac{b_i}{\Sigma_{i=1}^{n} b_i}. \tag{2}$$

Samples can be taken at regular time intervals, as long as it can be demonstrated that there are no periodic trends in the EF that coincide with the sampling frequency. A simple time series plot of EF measurements is informative: if there are no apparent time-trends then more or less regular sampling during the year is acceptable. If there are large gaps during the year, e.g. of one or more months without any samples being taken and if there are (or may be) longer-term trends (lasting weeks and months) in EF, then the set of collected samples may no longer be representative for the annual average.

An estimate of the standard error of the weighted mean given by equation 1 is given by (see Gatz and Smith [1995] and Cochran [1977]):

$$s_{\bar{y}} = \sqrt{\frac{n}{n-1}\Sigma_{i=1}^{n}[w_i^2(y_i - \bar{y})^2]} \tag{3}$$

The upper and lower bound of the 95% confidence interval are given by:

$$\bar{y}_L = \bar{y} - t_{0.975;n-1} s_{\bar{y}} \tag{4a}$$

$$\bar{y}_U = \bar{y} + t_{0.975;n-1} s_{\bar{y}}, \tag{4b}$$

where $t_{0.975;n-1}$ is the 97.5 percentile of the t distribution with $n - 1$ degrees of freedom.

2.2 Without auxiliary variable: Bootstrap

A non-parametric Bootstrap (see Efron and Tibshirani [1993]) method can be used to estimate the standard error or confidence interval of $\bar{y}$. Synthetic data sets are created by drawing samples at random with replacement from the observed set of samples. Each synthetic data set is of equal size to the observed data set, and consists of measurements that were present in the original (actual) data set, although not every measurement in the original data set may have been drawn for inclusion in the synthetic data set and other measurements may have been chosen more than once. The weighted mean is computed for each synthetic data set using equation 1). This way, a statistical distribution of means is formed from which the 2.5 an 97.5 percentiles are taken as the lower and upper limits respectively of the 95% confidence interval. Pseudocode with an outline of the Bootstrap procedure is given in Algorithm 1.

Advantages of the Bootstrap:

- The methodology is generic: any estimator for the mean may be used instead of the weighted mean.
- It is relatively straightforward to implement in a common data science language such as *R*, *Python* or *Matlab*.
- If the sample size is large enough (a rule of thumb is more than $n = 50$ measurements for a reasonably symmetric distribution of values), and the measurements are independent and representative draws from the population, then this method will tend to yield reasonable estimates with good coverage probability.

```
Algorithm 1: Bootstrap estimate of the confidence interval of a weighted sample mean
Data: The set of $n$ pairs of measurements $(y_i, b_i)$
Initialization: Generate a sequence of integers $r = 1, 2, \ldots, n$
repeat $M$ times
    Generate a random set $r^*$ of indicators by drawing $n$ values at random with replacement from the set $r$
    Create a synthetic data set of $n$ pairs of values $(y_r, b_r)$, where the indicators $r$ are taken from the set $r^*$ which was created in the previous step
    Based on the synthetic data, compute the weighted sample mean $\bar{y}^* = \frac{\Sigma(y_r b_r)}{\Sigma b_r}$, and store this value in a vector
end
Rank the vector of size $M$ with sample means $\bar{y}^*$ from smallest to largest and take the 2.5 and 97.5 percentiles as estimates of the lower and upper bounds of the confidence interval of the mean respectively.
```

The main disadvantages of the Bootstrap are:

- It is a computationally intensive method.
- This method does not work well with small sample sizes (e.g. less than $n = 50$ samples).
- The assumptions under which the results are valid are less explicit compared to a fully parametric method.

2.3 With a single auxiliary variable: regression estimator as in Cochran, 1977

The linear regression estimator is designed to increase precision by use of an auxiliary variable x which is correlated with the actual variable of interest y (Cochran [1977], page 189). For this estimator, it is assumed that the annual average of the auxiliary variable $\bar{X}$ is known:

$$\bar{X} = \frac{\Sigma_{j=1}^{k} B_j X_j}{\Sigma_{j=1}^{k} B_j} \tag{5}$$

This is reasonable only when both the auxiliary variable and flow rate are monitored continuously and at high frequency during the entire year, and a value is stored for example every hour or every 15 minutes.

The linear regression estimate of the annual average EF is given by:

$$\bar{y}_{lr} = \bar{y} + a_1(\bar{X} - \bar{x}), \tag{6}$$

where the subscript lr denotes *linear regression* and a_1 is the slope of a linear regression line as an estimate of the increase in y for a unit increase in the value of x (see equation 7.1 in Cochran [1977]). In equation 6, $\bar{x}$ is the weighted sample mean of the auxiliary variable:

$$\bar{x} = \Sigma_{i=1}^{n}(x_i w_i), \tag{7}$$

with the w_i as given in equation 2.

The parameter a_1 is obtained using the usual least squares estimator:

$$a_1 = \frac{\Sigma_{i=1}^{n}[(y_i - \hat{y})(x_i - \hat{x})]}{\Sigma_{i=1}^{n}[x_i - \hat{x}]^2}, \tag{8}$$

where $\hat{y} = n^{-1}\Sigma y_i$ and $\hat{x} = n^{-1}\Sigma x_i$ are simple unweighted arithmetic means.

The rationale behind the estimator as given in equation 6 is that if due to random sampling variability $\bar{x}$ is below the annual average $\bar{X}$, then the expectation is that estimate $\bar{y}$ will be below average by an amount $a_1(\bar{X} - \bar{x})$ (Cochran [1977], page 189). We note that the estimator given in equation 6 is not identical to equation 7.1 in Cochran [1977] because of the use of a weighted arithmetic mean instead of a simple arithmetic mean.

An estimate of the standard error of $\bar{y}_{lr}$ is given by equation 7.29 in Cochran [1977]:

$$s_{\bar{y}_{lr}} = \sqrt{\frac{1}{n-2}\Sigma_{i=1}^{n}[(y_i - \bar{y}_{lr})(x_i - \bar{x})]^2} \tag{9}$$

The upper and lower bound of the 95% confidence interval are given by:

$$\bar{y}_L = \bar{y}_{lr} - t_{0.975;n-2} s_{\bar{y}_{lr}} \tag{10a}$$

$$\bar{y}_U = \bar{y}_{lr} + t_{0.975;n-2} s_{\bar{y}_{lr}}, \tag{10b}$$

where $t_{0.975;n-2}$ is the 97.5 percentile of the t distribution with $n-2$ degrees of freedom.

We note that the estimator of the standard error given by equation 9 does not take the variability in the flow rate measurements into account.

2.4 With a single auxiliary variable: regression estimator as in van Zanten, 2016

A regression estimator for the average annual EF is described in van Zanten [2016]:

$$\bar{y}_{vz} = a_0 + a_1 \bar{X}, \tag{11}$$

where the subscript vz refers to the author of the technical report van Zanten [2016], $\bar{X}$ is given by equation 5.

The parameters a and b are the intercept and slope respetively of a linear regression model, $y_i = a_0 + a_1 x_j + z_j$, where z_j are stochastic variables that give the deviations between measured and predicted values. The z_j are assumed to be independent and Normally (Gaussian) distributed. The slope parameter a_1 is given by equation 8. The intercept a_0 is given by:

$$a_0 = \hat{y} - a_1 \hat{x}, \tag{12}$$

where $\hat{y} = n^{-1}\Sigma y_i$ and $\hat{x} = n^{-1}\Sigma x_i$ are simple unweighted arithmetic means.

The estimator for the standard error of $\bar{y}_{vz}$ is given in equation 7 in van Zanten [2016]:

$$s_{\bar{y}_{vz}} = s_{re}\sqrt{n^{-1}\frac{\Sigma_{i=1}^{n}[x_i - \bar{X}]^2}{\Sigma_{i=1}^{n}[x_i - \hat{x}]^2} + m^{-1}\left(1 + \frac{k^{-1}\Sigma_{j=1}^{k}[B_j - \hat{B}]^2}{(\hat{B})^2}\right)}, \tag{13}$$

where $\hat{y} = n^{-1}\Sigma y_i$, $\hat{x} = n^{-1}\Sigma x_i$ and $\hat{B} = k^{-1}\Sigma B_j$ are simple unweighted arithmetic means. The standard deviation of the residuals of the linear regression model s_{re} is given by:

$$s_{re} = \sqrt{\frac{\Sigma_{i=1}^{n}[z_i - \hat{z}]^2}{n-2}}, \tag{14}$$

where $\hat{z} = n^{-1}\Sigma z_j$ is the simple unweighted arithmetic mean of the model residuals.

The upper and lower bound of the 95% confidence interval are given by:

$$\bar{y}_L = \bar{y}_{vz} - t_{0.975;n-2} s_{\bar{y}_{vz}} \tag{15a}$$

$$\bar{y}_U = \bar{y}_{vz} + t_{0.975;n-2} s_{\bar{y}_{vz}}, \tag{15b}$$

where $t_{0.975;n-2}$ is the 97.5 percentile of the t distribution with $n-2$ degrees of freedom.

It is not assumed that $\bar{X}$ is known without error. The variability in flow rate measurements is taken into account in the estimate of the standard error of the mean.

2.5 With a single or multiple auxiliary variables: regression estimator with Bootstrap

The Bootstrap method, introduced in section 2.2, may also be applied to estimate the confidence interval of the mean based on a linear regression equation. Two variations of the algorithm are discussed:

- A Bootstrap algorithm which assumes that the flow rate is not correlated with the EF and the auxiliary variable. In this algorithm, the EF is modelled as a linear function of the auxiliary variable only.
- A Boostrap algorithm in which the EF is modelled as a linear function of the auxiliary variable and flow rate. This is applicable only in situations where the flow rate correlates with the EF and auxiliary variable.

The second variation of the algorithm, with the linear regression of EF on both the flow rate and auxiliary variables, is included because the flow rate plays an important role in computing the weighted means and because the estimators in sections 2.3 and 2.4 do not take this possibility into account. A question of interest is therefore what the performance, in terms of coverage probabilities, of the estimators is when the flow rate correlates with the EF. Pseudocode for the Bootstrap algorithm is given in Algrotihm 2.

The Bootstrap procedures outlined in this section can easily be extended to include more variables and slope parameters.

Thompson [2012] (chapter 8) outlines a number of estimators based on multiple regression, and mentions that more research is needed on the topic (page 117 in Thompson [2012]). For this reason, it may best to use the Bootstrap to estimate uncertainties for multiple regression models.

Algorithm 2: Bootstrap estimate of the confidence interval of an estimate of the mean based on a linear regression equation of EF on the auxiliary variable (version 1) or both the auxiliary variable and flow rate (version 2).

```
Data: $n$ triples of measurements $(y_i, x_i, b_i)$; $k$ pairs of measurements $(X_j, B_j)$; $k$ weights $W_j = \frac{B_j}{\Sigma_{j=1}^{k} B_j}$
Initialization: Generate a sequence of integers $r = 1, 2, \ldots, n$
repeat $M$ times
    Generate a random set $r^*$ of indicators by drawing $n$ values at random with replacement from the set $r$
    Create a synthetic data set of $n$ pairs of values $(y_r, x_r)$, where the indicators $r$ are taken from the set $r^*$ which was create in the previous step
    if regression model based on auxiliary variable then
        Based on the synthetic data compute the intercept $a_0^*$ and slopes $a_1^*$ of the linear regression equation $y_r = a_0^* + a_1^* x_r + z_r$ (using the standard least squares estimator for these parameters)
        Compute the standard deviation of the residuals of the regression model, $s_{re}^*$
        Create a synthetic data set of $j = 1, 2, \ldots, k$ estimates of EF, $q_j$: $q_j = a_0^* + a_1^* X_j + z_j^*$, where the $z_j^*$ are drawn at random from a Normal distribution with mean zero and standard deviation $s_{re}^*$
    end
    if regression model based on auxiliary variable and flow rate then
        Based on the synthetic data compute the intercept $a_0^*$ and slopes $a_1^*$ and $a_2$ of the linear regression equation $y_r = a_0^* + a_1^* x_r + a_2^* b_r + z_r$ (using the standard least squares estimator for these parameters)
        Compute the standard deviation of the residuals of the regression model, $s_{re}^*$
        Create a synthetic data set of $j = 1, 2, \ldots, k$ estimates of EF, $q_j$: $q_j = a_0^* + a_1^* X_j + a_2^* B_j + z_j^*$, where the $z_j^*$ are drawn at random from a Normal distribution with mean zero and standard deviation $s_{re}^*$
    end
    Based on the synthetic set of measurements $q_j$ compute the weighted mean $\bar{y}^* = \Sigma_{j=1}^{k}[W_j q_j]$ and store this value in a vector.
end
Rank the vector of size $M$ with sample means $\bar{y}^*$ from smallest to largest and take the 2.5 and 97.5 percentiles as estimates of the lower and upper bounds of the confidence interval of the mean respectively.
```

3 Model Validation

3.1 Representativeness of the sample

Samples can be taken at regular time intervals, as long as it can be demonstrated that there are no periodic trends in the EF that coincide with the sampling frequency. A simple time series plot of EF measurements is informative: if there are no apparent time-trends then more or less regular sampling during the year is acceptable. If there are large gaps during the year, e.g. of one or more months without any samples being taken and if there are (or may be) longer-term trends (lasting weeks and months) in EF, then the set of collected samples may no longer be representative for the annual average.

3.2 Validity of the regression model

More precise estimates can be obtained with auxiliary variables. However, the validity of these estimates depends to a large extent on the validity of the regression model (see for example chapter 4 in van Zanten [2016]). Deviations of measurements from the regression line are assumed ot be independent and Normally distributed variables, with constant variance across the range of predicted values. A number of diagnostic plots can be used to assess whether these assumptions are reasonable, or to see if there are indications that they are violated.

- Time series plots of model residuals should show no apparent patterns. If there are pattterns, then it is worth investigating whether or nor the residuals correlate with other variables (if available).
- A Quantile-Quantile plot to check if the model residuals are (at least approximately) normally distributed
- A scatterplot of predicted versus observed values, to check for potential outliers, and to see of the variability (the "scatter" around the 1:1 line) is roughly constant across the predicted range of values.

A good graphical analysis of the residuals is key for a good appraisal of the validity of the model.

4 Assessment of the performance of estimators using a simulation study

4.1 Simulation method

A simulation study, in which synthetic data with characteristics that are as close as possible to those that may occur in real life, may be used to better understand of performance of the estimators presented in chapter 2 in terms of:

- The dependence of the precision (the width of the 95% confidence interval) of the estimates, as a function of sample size, variability in the key process parameters (EF and flow rate) and the strength of the relationship between auxiliary variable and EF
- The coverage probability of the 95% confidence intervals, measured by the proportion of times that the confidence intervals contain the true value of the annual average EF used to generate the synthetic data sets.

An estimator yields valid results only if its associated coverage probability is (close to) 95%. A good, or efficient, estimator yields, for a given sample size and scenario, a small width of the confidence interval whilst maintaining good coverage probability. If a high precision (small width of confidence interval) is associated with a coverage probability that is too low, then the precision estimate is misleading.

Scenarios are defined by a combination of the following:

- The number of laboratory measurements
- The correlation (strength of relationship) between the auxiliary variable and the laboratory measurements
- The variability in EF, auxiliary variable and flow rate measurements (due to a combination of process variability and measurement errors)
- The correlation (strength of relationship) between flow rate and EF, and between flow rate and auxiliary variable

It is assumed that the flow rate and auxiliary variable is monitored continuously, so that the uncertainty regarding for example the average, or weighted average, of the flow rate and auxiliary variable is negligible. At the point of writing this report, only a single auxiliary variable is used in the simulation study.

Many synthetic data sets are created, each with $4 \times 24 \times 365 = 35040$ triples of measurements (Y_j, B_j, X_j), representing a hypothetical situation in which all variables of interest were monitored every 15 minutes during the entire year. The true value of the annual average EF is taken to be:

$$\mu_{true} = \frac{\Sigma[Y_j B_j]}{\Sigma B_j}.$$

The triples of measurements are drawn from a Multivariate Normal (MVN) distribution:

$$\begin{bmatrix} Y \\ X \\ B \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu_Y \\ \mu_X \\ \mu_B \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & \rho_{YX}\sigma_Y\sigma_X & \rho_{YB}\sigma_Y\sigma_B \\ \rho_{YX}\sigma_Y\sigma_X & \sigma_X^2 & \rho_{XB}\sigma_X\sigma_B \\ \rho_{YB}\sigma_Y\sigma_B & \rho_{XB}\sigma_X\sigma_B & \sigma_B^2 \end{bmatrix} \right), \tag{16}$$

where (μ_y,σ_Y^2), (μ_x,σ_X^2) and (μ_B,σ_B^2) are parameters for the mean and variance of the EF, auxiliary variable and flow rate respectively, to be specified to define a scenario. The parameters ρ_{YX}, ρ_{YB} and ρ_{XB} denote the correlation between the pairs of variables. In the simulation study, ρ_{YX} and ρ_{YB} can be set independently, but ρ_{XB} is set to be $\rho_{XB} = \rho_{YX} * \rho_{YB}$. The mean and variance of the auxiliary variable are of no consequence to the estimates as long as this variable is monitored continuously.

The sampling of the fuel gas flow is simulated by drawing n triples (y_j, b_j, x_j) at random without replacement from the set of k triples (Y_j, B_j, X_j).

An outline of the simulation procedure is given in pseudocode in Algorithm 3.

The reliability of the coverage probability C depends on the number of simulations M. To assess the uncertainty in the coverage probability due to small number of simulations, a 95% confidence interval is computed based on the 2.5 and 97.5 percentiles of the Beta distribution: Beta$(I, M - I)$.

4.2 web-App to explore the expected precision of estimates for different sccenarios

A simple web-App has been created which allows the user to specify the scenario in terms of the key parameters of interest (see a screen shot of the web-App in figure 1):

- The sample size n.

Algorithm 3: Simulation study to estimate the relative precision and coverage probability of the estimators.

```
Initialize: Choose values for the sample size $n$, the pairs of values for the mean ad variance of the
  process parameters ($\mu_y$,$\sigma_Y^2$), ($\mu_x$,$\sigma_X^2$) and ($\mu_B$,$\sigma_B^2$), and the correlation parameters $\rho_{YX}$ and $\rho_{YB}$. Set
  $I \leftarrow 0$
repeat $M$ times
    Generate a synthetic data set of 35040 triples $(Y_j, B_j, X_j)$ using equation 16.
    Compute the true value $\mu_{true}$.
    Generate a synthetic sample data set of $n$ triples $(y_j, b_j, x_j)$ by sampling at random without
      replacement.
    Apply the estimators described in chapter 2 and store the estimated relative precision in a vector.
    if $\mu_{true}$ is smaller than the upper bound and larger than the lower bound of the 95% confidence
      interval then
        $I \leftarrow I + 1$
    end
end
Rank the vector of size $M$ with precisions from smallest to largest and take the 2.5 and 97.5 percentiles
  as estimates of the range of values that the relative precision could take under the defined scenario.
  The coverage probability $C$ (expressed as a percentage) is given by $C = 100I/M$.
```

- The mean and standard deviation (defining process variability) of the EF, μ_Y and σ_Y.
- The mean and standard deviation (defining process variability) of the flow rate, μ_B and σ_B.
- The correlation between the auxiliary variable and the EF, ρ_{XY}.
- The correlation between the flow rate and the EF, ρ_{YB}.

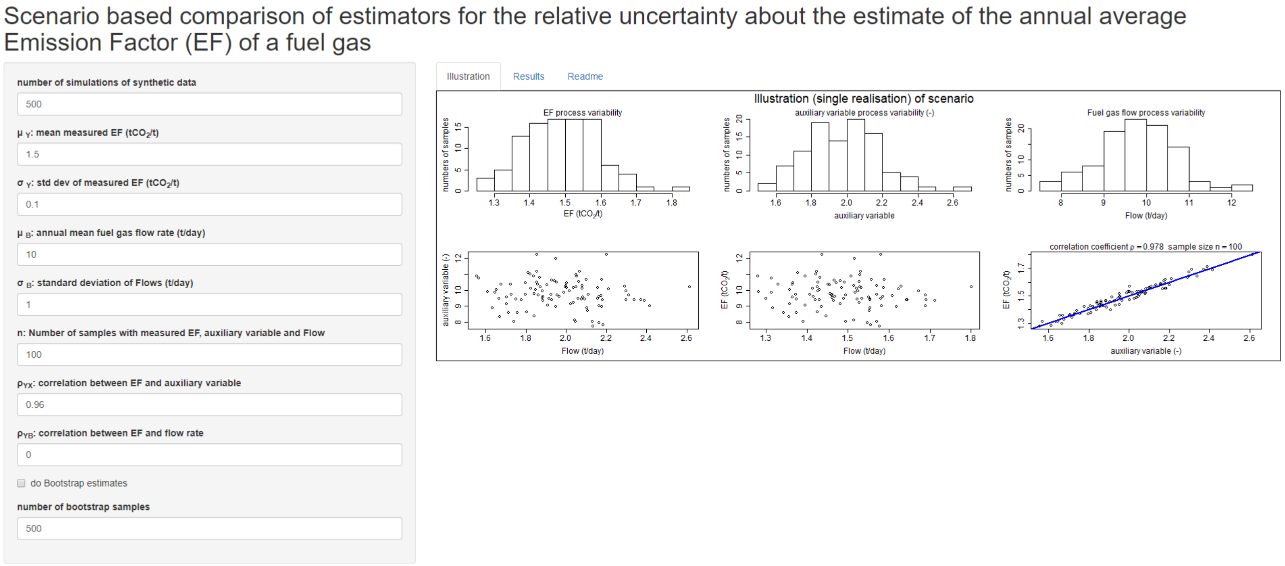
the use must choose values for these parameters so that synthetic data sets are created that closely resemble the real-life situations.

Additionally, the following must be specified:

- The number of simulations (synthetic data sets). The larger the number of simulations, the more precise the estimates of the performance of the estimators will be. For the relative precision, 500 simulations is typically sufficient. For a good estimate of coverage probability, typiclly at least 1000 simulations are required.
- A tick box to indicate whether or not Bootstrap estimators should be included. This is computer intensive, and it may take several minutes for the results to be generated.
- The number of simulations in the Bootstrap. For a reasonable assessment of the performance of the Bootstrap, at least 1000 simulations are required. The larger the number of Bootstrap simulations and synthetic data sets, the longer it will take for the results to be generated.

An illustration of the results of the web-App, for a given scenario with $n = 100$ and $\rho_{YX} = 0.94$ is given in figure 2. For this particular scenario, the estimators without auxiliary variable yield relative uncertainty estimates that are between 2.18% and 3%, well above the target set by the NEA of 0.5%. The corresponding coverage probability is on target: close to 95%. There is good agreement between the Bootstrap (section 2.2) and Simple Random Sample (SRS; section 2.1) estimators. The estimators with auxiliary variables (sections 2.5,2.3 and 2.4 all agree well in terms of both relative uncertainty and coverage, and yield uncertainty estimates in the range 0.62% - 0.85%, somewhat above the target set by the NEA.

Another illustration of the results of the web-App, for a given scenario with $n = 150$ and $\rho_{YX} = 0.97$ is given in figure 3. For this particular scenario, the estimators with auxiliary variables (sections 2.5,2.3 and 2.4 all agree well in terms of both relative uncertainty and coverage, and yield uncertainty estimates in the range 0.46% - 0.59%; depending on chance the esitmate may come in just below or just above the target set by the NEA.



Scenario based comparison of estimators for the relative uncertainty about the estimate of the annual average Emission Factor (EF) of a fuel gas

number of simulations of synthetic data: 500

μ_Y: mean measured EF (tCO_2/t): 1.5

σ_Y: std dev of measured EF (tCO_2/t): 0.1

μ_B: annual mean fuel gas flow rate (t/day): 10

σ_B: standard deviation of Flows (t/day): 1

n: Number of samples with measured EF, auxiliary variable and Flow: 100

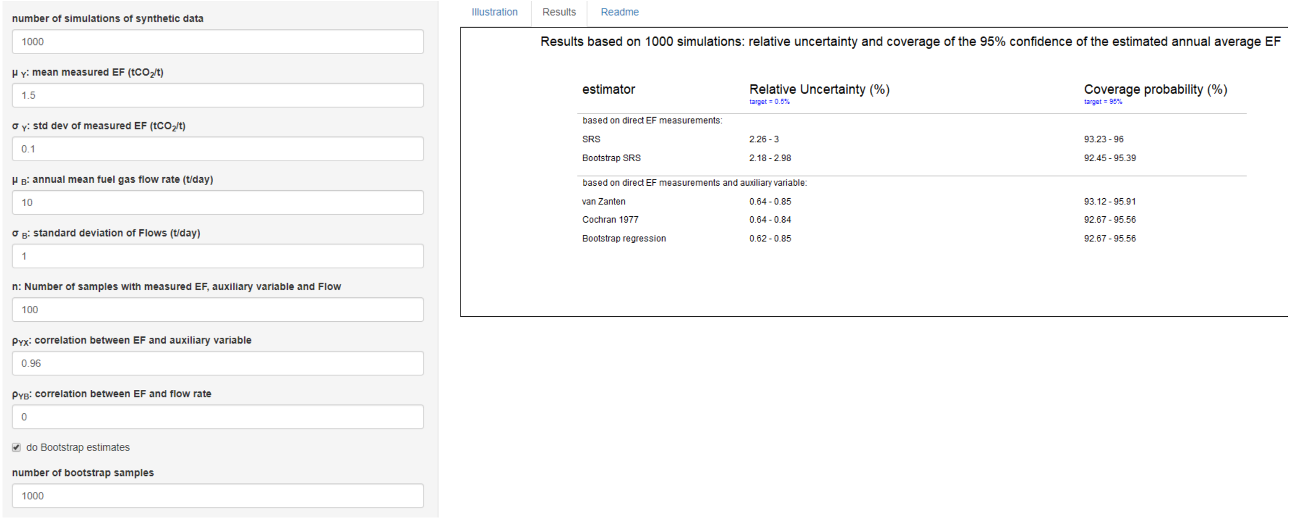
ρ_{YX}: correlation between EF and auxiliary variable: 0.96

ρ_{YB}: correlation between EF and flow rate: 0

□ do Bootstrap estimates

number of bootstrap samples: 500

Figure 1: Screen shot of the web-App. On the left are the inputs that define the scenario as well as the number of required simulations. On the right are a number of graphs that illustrate, for a single synthetic data set, what the synthetic data look like.



number of simulations of synthetic data: 1000

μ_Y: mean measured EF (tCO_2/t): 1.5

σ_Y: std dev of measured EF (tCO_2/t): 0.1

μ_B: annual mean fuel gas flow rate (t/day): 10

σ_B: standard deviation of Flows (t/day): 1

n: Number of samples with measured EF, auxiliary variable and Flow: 100

ρ_{YX}: correlation between EF and auxiliary variable: 0.96

ρ_{YB}: correlation between EF and flow rate: 0

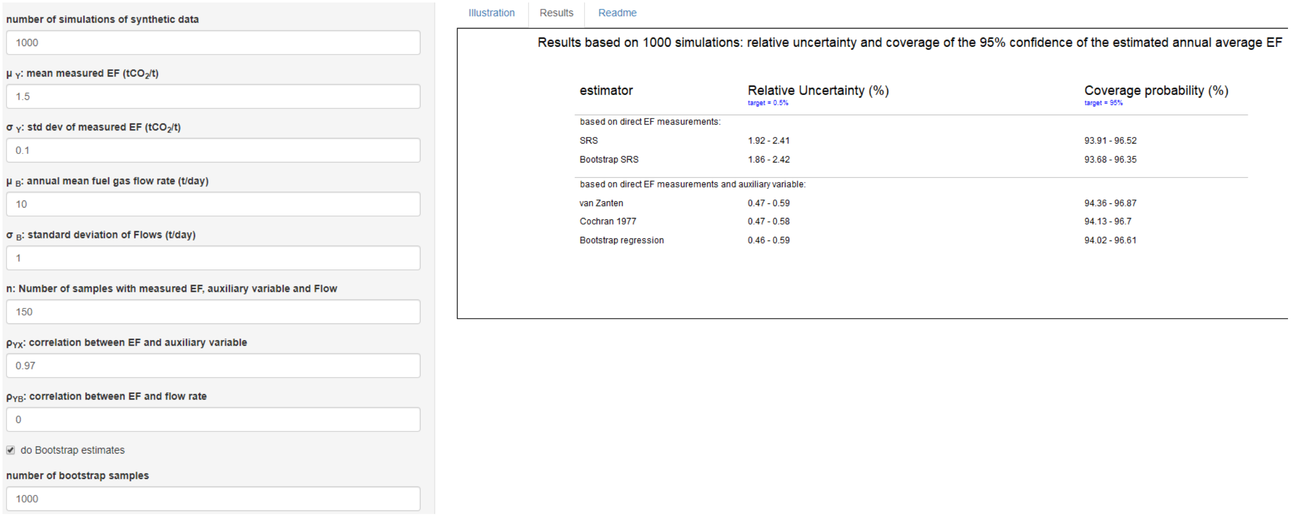
☑ do Bootstrap estimates

number of bootstrap samples: 1000

Results based on 1000 simulations: relative uncertainty and coverage of the 95% confidence of the estimated annual average EF

estimator	Relative Uncertainty (%) target = 0.5%	Coverage probability (%) target = 95%
based on direct EF measurements:		
SRS	2.26 - 3	93.23 - 96
Bootstrap SRS	2.18 - 2.98	92.45 - 95.39
based on direct EF measurements and auxiliary variable:		
van Zanten	0.64 - 0.85	93.12 - 95.91
Cochran 1977	0.64 - 0.84	92.67 - 95.56
Bootstrap regression	0.62 - 0.85	92.67 - 95.56

Figure 2: Screen shot of the web-App, with results. On the left are the inputs that define the scenario as well as the number of required simulations. On the right are the results in terms of the relative precisions and coverage probabilities. Note that these results took approximately 20 minutes to be generated.



number of simulations of synthetic data

1000

μ_Y: mean measured EF (tCO_2/t)

1.5

σ_Y: std dev of measured EF (tCO_2/t)

0.1

μ_B: annual mean fuel gas flow rate (t/day)

10

σ_B: standard deviation of Flows (t/day)

1

n: Number of samples with measured EF, auxiliary variable and Flow

150

ρ_{YX}: correlation between EF and auxiliary variable

0.97

ρ_{YB}: correlation between EF and flow rate

0

☑ do Bootstrap estimates

number of bootstrap samples

1000

Results based on 1000 simulations: relative uncertainty and coverage of the 95% confidence of the estimated annual average EF

estimator	Relative Uncertainty (%) target = 0.5%	Coverage probability (%) target = 95%
based on direct EF measurements:		
SRS	1.92 - 2.41	93.91 - 96.52
Bootstrap SRS	1.86 - 2.42	93.68 - 96.35
based on direct EF measurements and auxiliary variable:		
van Zanten	0.47 - 0.59	94.36 - 96.87
Cochran 1977	0.47 - 0.58	94.13 - 96.7
Bootstrap regression	0.46 - 0.59	94.02 - 96.61

Figure 3: Screen shot of the web-App, with results. On the left are the inputs that define the scenario as well as the number of required simulations. On the right are the results in terms of the relative precisions and coverage probabilities. Note that these results took approximately 20 minutes to be generated.

5 Combining measurements from multiple fuel streams

If multiple fuel streams are sampled, then measurements may be combined to obtain an estimate of the average annual EF for all fuel streams combined. A lower relative uncertainty can be obtained if the measurements can be obtained. The statistical methodology, including the optimal allocation of sample sizes across fuel streams to minimize the relative uncertainty, is outlined in chapter 5.5 in Cochran [1977].

6 Data analysis Pernis Refinery

In this section, a descriptive analysis of EF measurements from a fuel gas in the Pernis refinery is presented. From 2014 to 2020, the fuel gas stream has been sampled at more or less regular intervals albeit with certain longer periods without any measurements. For each sample, the EF is measured based on a full analysis of the composition of the fuel gas in the laboratory. There is considerable variability in the EF over time (figure 4), which is believed to be a reflection of the variability in process conditions, as the measurement uncertainty is small. The annual variability in EF, due to temporal variability in the composition of the fuel gas, is given in figure 5.

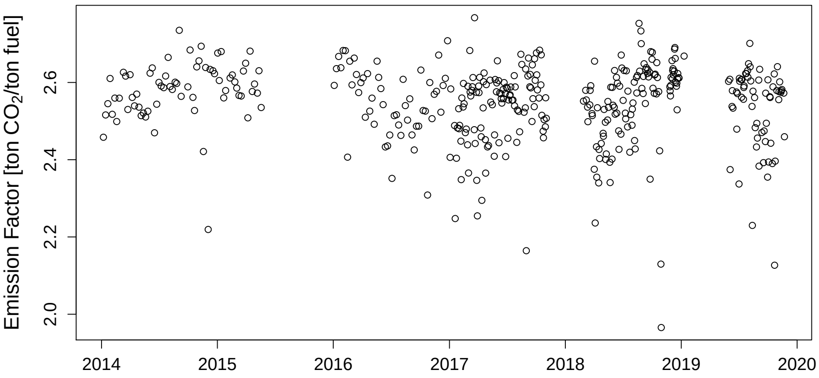


Figure 4: Time series plot of EF, measured on samples taken from the flow of a fuel gas at the Pernis refinery.

The EF correlates very strongly with the Stoichiometric Air Requirement (SAR; figure 6) and less strongly with Lower Heating Value (LHV; figure 7). Estimates of SAR and LHV are based on the composition of the sample of fuel gas as determined in the laboratory. As such, these variables are therefore not directly useful as auxiliary variables. However, these relationships may provide clues to potential candidates for high-frequency online measurements that may correlate well with EF. SAR may be directly estimated based on a mass spectrometry device (see e.g. Merriman and Lewis [2016]), and it is worth investigating if oxygen analyzers or calorimeters correlate well with either SAR or LHV.

The residuals of a linear model of EF on SAR correlate strongly with molar weight as measured by an analyser which is already online in the fuel gas stream at the Pernis refinery (figure 8). A multiple linear regression model of EF on SAR and molar weight yields predictions that correlate very strongly with measured EF (figure 9). This is encouraging because, if a device can be found which can measure online a variable which correlates with SAR or LHV, then this may result in a model with good predictive performance in combination with online molar weight measurements.

A time series plot of the residuals of the multiple linear regression model of EF on SAR and molar weight indicates that there are still some unexplained trends in the EF measurements (figure 10). For example, there is a consistent deviation between predicted and measured EF in 2019. This, however, does not have to lead to a deterioration of the uncertainty of the estimate of the annual average, as long as the annual average EF is based on a combination of direct EF measurements in the laboratory and the auxiliary variable(s).

An exploratory data analysis (not shown in this report) indicated that the residuals of the multiple linear regression model of EF on SAR and molar weight correlate with ethylene and propylene concentrations in the fuel gas. Based on these two additional variables, the predictive model for EF could be further improved. However, this is not further discussed or illustrated here.

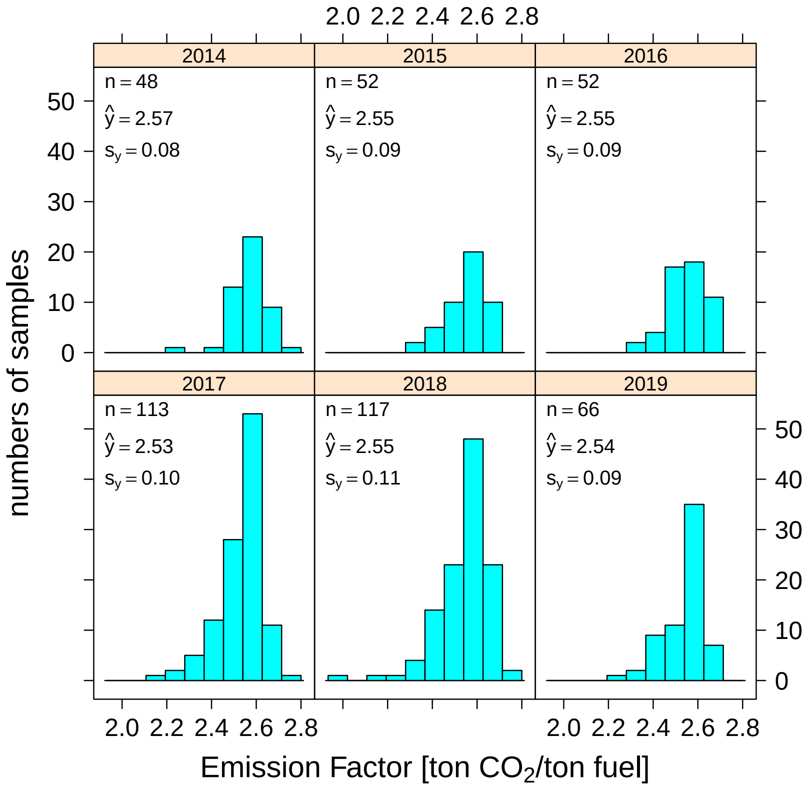


Figure 5: Frequency histograms of EF, measured on samples taken from the flow of a fuel gas at the Pernis refinery. Indicated are the annual number of samples n, the annual standard deviation in measured EF (σ_y) and the annual average EF $(\hat{y})$.

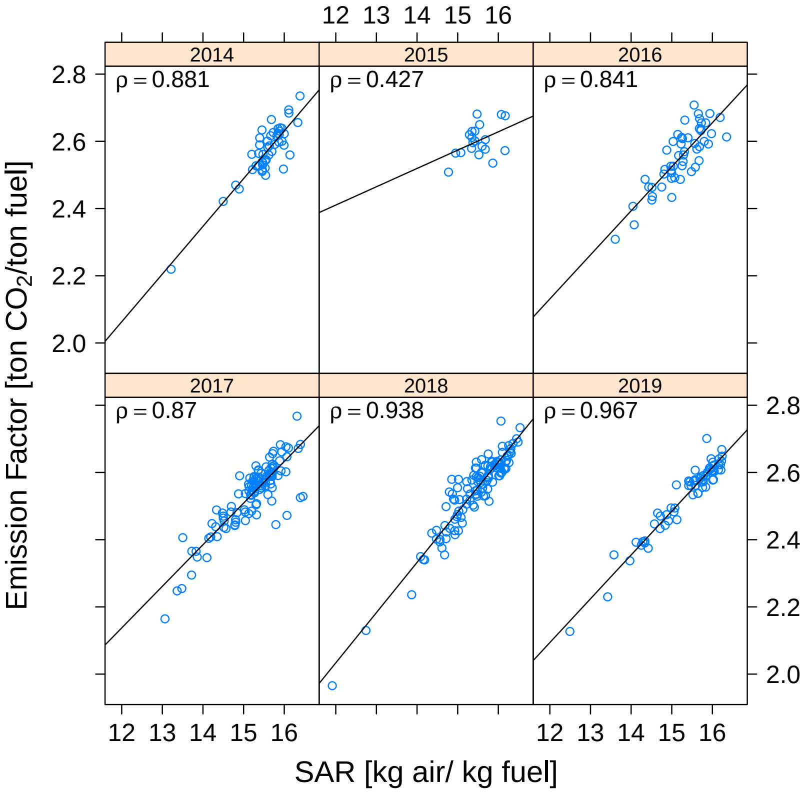


Figure 6: Scatterplot of EF against Stoichiometric Air Requirement (SAR), with Pearson correlation coefficients (ρ).

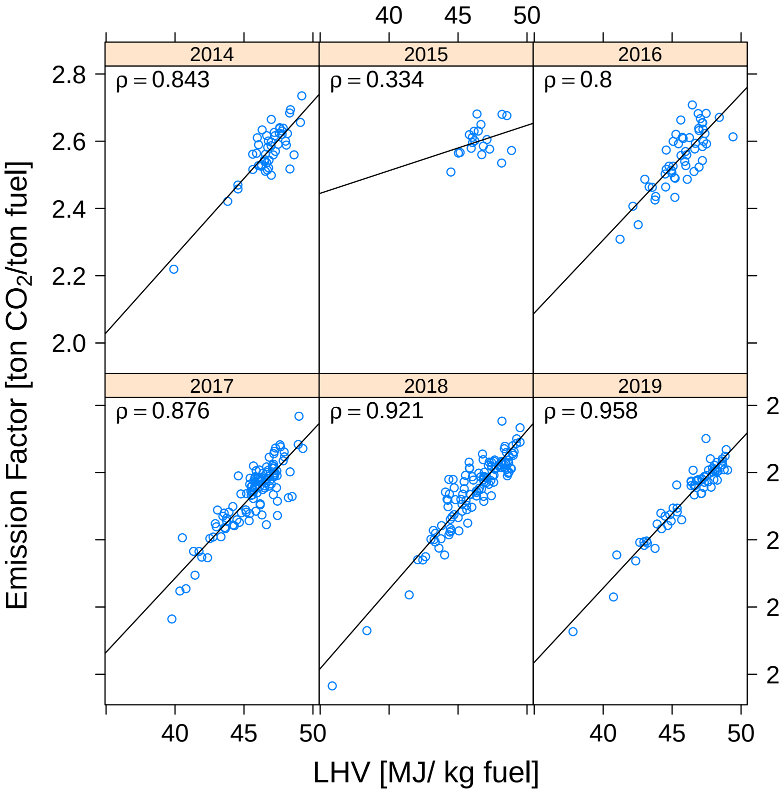


Figure 7: Scatterplot of EF against Lower Heating Value (LHV), with Pearson correlation coefficients (ρ).

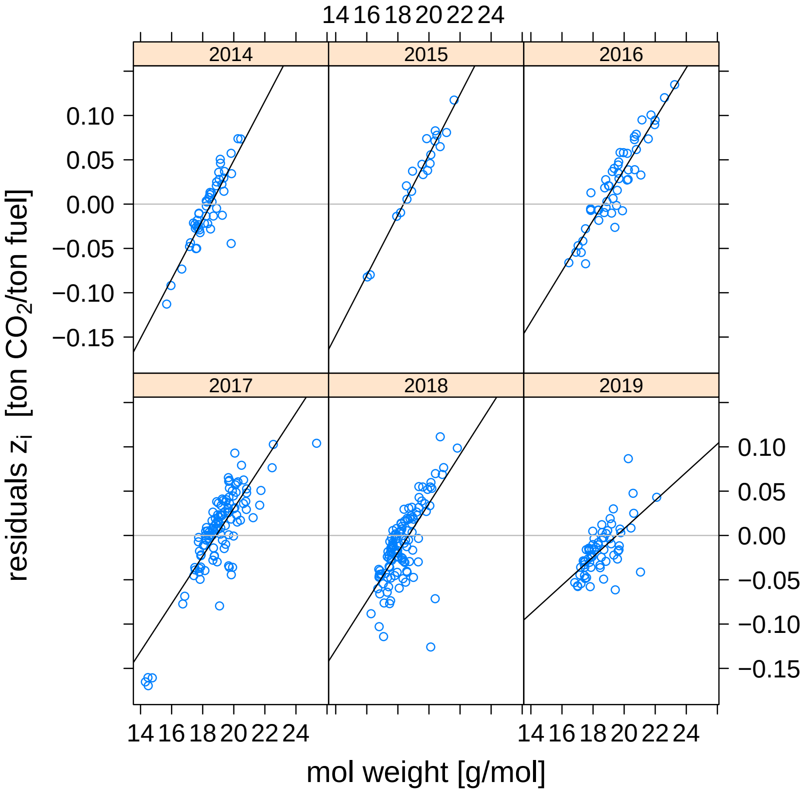


Figure 8: Scatterplot of the residuals of a linear regression model of EF on SAR, versus molar weight.

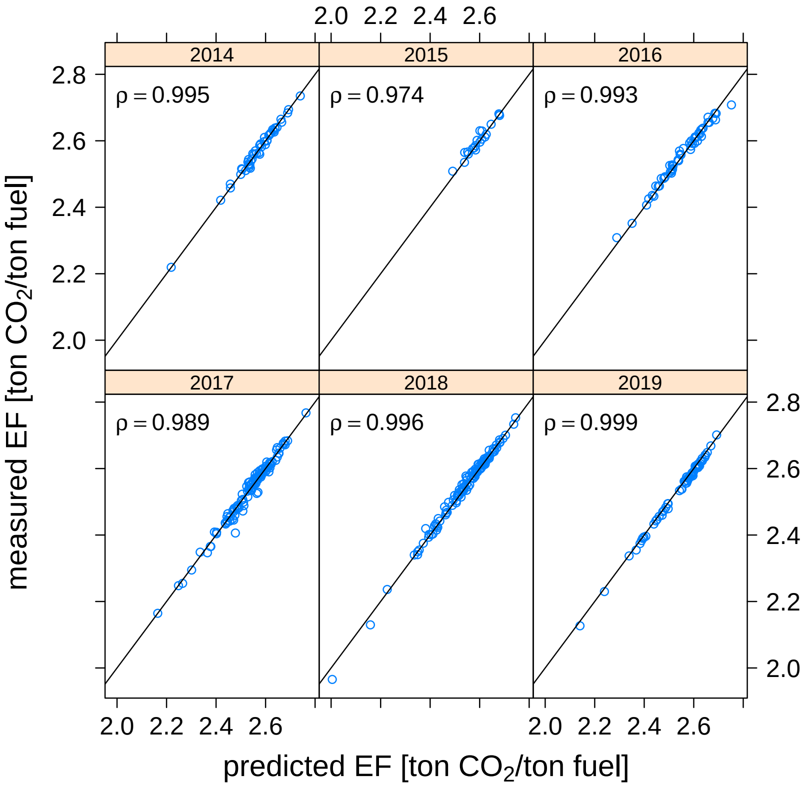


Figure 9: Scatterplot of measured versus predicted EF. The EF was predicted using a linear regression model with SAR and molar weight as covariates. Indicated on the graphs are Pearson correlation coefficients (ρ).

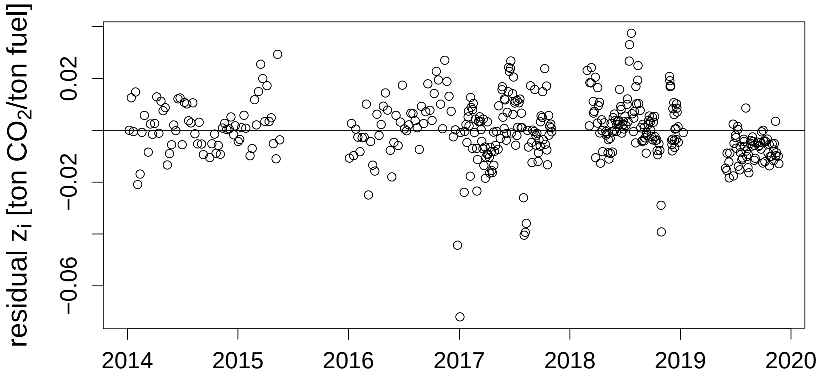


Figure 10: Time series plot of the residuals of a multiple linear regression model of EF on SAR and molar weight.

References

William G. Cochran. *Sampling Techniques, 3rd Edition.* 1977.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap.* Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

Donald F. Gatz and Luther Smith. The standard error of a weighted mean concentration—i. bootstrapping vs other methods. *Atmospheric Environment*, 29(11):1185 – 1193, 1995. ISSN 1352-2310. doi: https://doi.org/10.1016/1352-2310(94)00210-C. URL `http://www.sciencedirect.com/science/article/pii/135223109400210C`.

D. Merriman and G. Lewis. *Fast On-Line Monitoring of Fuel Gases with the Thermo Scientific Prima PRO Process Mass Spectrometer.* Winsford Cheshire UK, 2016.

Theresa M. Shires, Christopher J. Loughran, Stephanie Jones, and Emily Hopkins. *Compendium of greenhouse gas emissions methodologies for the oil and natural gas industry.* American Petroleum Institute, Austin, Texas, 2009.

Steven K. Thompson. *Sampling, 3rd Edition.* 2012.

J.H. van Zanten. *Het bepalen van CO2-emissies met behulp van correlatiemodellen.* Amsterdam StatLab, Amsterdam, Nederland, 2016.