Environmental Science
Processes & Impacts

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/process-impacts

Assessing the performance of standard methods to predict the standard uncertainty of air quality data having incomplete time coverage †

Richard J. C. Brown*[1], Peter M. Harris[2] and Maurice G. Cox[2]

[1] Analytical Science Division, National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK.

[2] Materials Division, National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK.

* Corresponding author. Tel.: +44 20 8943 6409 Fax: +44 20 8614 0423

E-mail address: richard.brown@npl.co.uk

† Electronic supplementary information available

The validity of standardised equations to calculate the uncertainty arising from missing data during air quality studies is assessed.

Environmental Science: Processes & Impacts Accepted Manuscript

Assessing the performance of standard methods to predict the standard uncertainty of air quality data having incomplete time coverage †

Richard J. C. Brown*[1], Peter M. Harris[2] and Maurice G. Cox[2]

[1] Analytical Science Division, National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK.

[2] Materials Division, National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK.

* Corresponding author. Tel.: +44 20 8943 6409 Fax: +44 20 8614 0423

E-mail address: richard.brown@npl.co.uk

This paper investigates the influence of missing data on annual averages and the uncertainties in these averages produced by UK air quality networks. Whilst the standard methods currently employed produce good results on average, for individual cases the uncertainty in the annual average calculated when data is missing may be appreciably different from that obtained when full knowledge of the distribution of the data is known. These effects become more apparent as the quantity of missing data increases. These outcomes of this study will advance greatly the understanding of the uncertainty of summary air quality statistics produced by air quality networks and will inform the current debate on how best to revise air quality monitoring legislation in Europe and globally to recognise better different levels of confidence in the data produced.

Environmental Science: Processes & Impacts Accepted Manuscript

Assessing the performance of standard methods to predict the standard uncertainty of air quality data having incomplete time coverage †

Richard J. C. Brown*[1], Peter M. Harris[2] and Maurice G. Cox[2]

[1] Analytical Science Division, National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK.

[2] Materials Division, National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK.

* Corresponding author. Tel.: +44 20 8943 6409 Fax: +44 20 8614 0423

E-mail address: richard.brown@npl.co.uk

† Electronic supplementary information available

Abstract

As a result of the complex nature of operating multi-station national air quality networks it is rare that complete data sets are produced from these networks. The reliance of most air quality legislation on the assessment of measured annual average concentrations against target or limit concentrations necessitates the use of methods to calculate an annual average value and the uncertainty in this value in the absence of a complete data set for the year in question. Standard procedures exist for performing these calculations, but it is not clear how effective these are when data having low time resolution are collected and missing data accounts for large periods of the year. This paper investigates the influence of these deficiencies using data from UK air quality networks in the form of monthly average concentrations for polycyclic aromatic hydrocarbons and for metals in the PM_{10} phase of ambient air. Whilst the standard methods currently employed produce good results on average, for individual cases the uncertainty in the annual average calculated when data is missing may be appreciably different from that obtained when full knowledge of the distribution of the data is known. These effects become more apparent as the quantity of missing data increases.

Keywords: air quality, annual average, missing data, weighted average, uncertainty

Environmental Science: Processes & Impacts Accepted Manuscript

Environmental Science: Processes & Impacts Accepted Manuscript

Introduction

Despite improvements in air quality in recent decades, continued concerns about the effects on human health and environmental sustainability of pollutants in ambient air have prompted national and international legislation aimed at limiting the concentrations of harmful substances in the air we breathe. In most cases epidemiological evidence suggests there is no safe level for exposure to harmful air pollutants. Because of this, the majority of legislation is based on limiting allowable annual average concentrations, which establish average exposure over a long period, rather than limiting the time period when concentrations may be above some threshold. An overview of relevant European air quality legislation is available in [1]. Assessment of air quality against legislation is performed in most countries by national or regional air quality networks consisting of fixed monitoring stations.

The nature of these complex endeavours means that the data sets produced do not always cover the whole year. Data is often missing because of, for example, instrument breakdown, invalid measurements, damaged samples, analytical problems, or data excluded at the checking stage. Previously we have considered how best to calculate the annual average concentration in the event of missing data [2]. We showed that considerations of particular interest were pollutants showing seasonality and data missing over extended periods during the year, because in these circumstances the calculated annual average concentration can be sensitive to which data are missing [3,4].

The relevant European Air Quality Directives [5,6] specify the use of ISO 11222 "Air Quality – Determination of the uncertainty of the time average of air quality measurements" [7] for the calculation of the annual average concentration and the uncertainty associated with this annual average. We should note that the use of ISO 11222 assumes that the available data is representative of the temporal variation of the concentration of the pollutant over the defined time period. ISO 11222 uses the simple mean to calculate the time average, $\bar{C}_T$, as:

$$\bar{C}_T = \frac{1}{n}\sum_{i=1}^{n} C_i \qquad (1)$$

where the C_i are the measured values for n measurement periods within the overall time period T, and $n \leq N$, where N is the number of periods for complete coverage of the time period T. Incomplete coverage of the time period T, corresponding to $n < N$, is a further

source of uncertainty associated with the time average $\bar{C}_T$ that is not covered by the uncertainties arising from the sampling and analytical procedures. ISO 11222 states that the standard uncertainty $u_s(\bar{C}_T)$ associated with the time average $\bar{C}_T$ due to incomplete time coverage shall be determined as:

$$u_s^2(\bar{C}_T) = \left(1 - \frac{n}{N}\right)\frac{1}{n}s^2(C_i) \quad (2)$$

where $s^2(C_i)$ denotes the variance of the measured values determined as:

$$s^2(C_i) = \frac{1}{n-1}\sum_{i=1}^{n}(C_i - \bar{C}_T)^2 \quad (3)$$

Equation (2) expresses the uncertainty associated with an estimate of the average of a finite population of N values obtained from incomplete knowledge of that population consisting of n values drawn randomly from the population. (Note that when $n = N$, the uncertainty is zero because in that case the entire population, and its average, are known.) Furthermore, it relies on an estimate of the variance of the population calculated in terms of the n values given by equation (3). To the best of the authors' knowledge the performance of equations (2) and (3) has never been fully examined. It is of particular interest to understand their performance in the most extreme conditions where they might be applied: for instance, where the data collected has low time resolution and missing data accounts for large periods of the year.

This work examines the performance of equations (2) and (3) using data from the UK PAH Monitoring Network ("the PAH Network") from 2008 to 2011 and the UK Urban and Industrial Metals Monitoring Network ("the Metals Network") during 2008, 2009 and 2010. (For a general description of air quality networks in the UK see [8].) Whilst data from the PAH Network shows a seasonal variation at non-industrial stations greater than the random variability arising from meteorological and analytical processes, data from the Metals Network does not show seasonal variation greater than that from random background processes. The seasonality observed in the PAH Network data is caused by increased fuel use in winter for heating resulting in higher concentrations than in summer.

Environmental Science: Processes & Impacts Accepted Manuscript

Environmental Science: Processes & Impacts Accepted Manuscript

Experimental

In the periods 2008 to 2011 and 2008 to 2010, respectively, the 31 monitoring stations of the PAH Network and 24 monitoring stations of the Metals Network produced monthly averaged concentration values for a variety of pollutants. The averaged values are obtained by bulking together filters (exposed for daily periods on the PAH Network and weekly periods on the Metals Network) corresponding to the monthly periods prior to analysis. Details of the sampling and analysis used to produce the data considered in this analysis are given in the relevant network reports [9,10]. The data produced is freely available from Defra's UK-AIR website [11]. We have considered a reduced subset of the data produced by these networks: benzo[a]pyrene (BaP) from the PAH Network and nickel (Ni), cadmium (Cd), arsenic (As) and lead (Pb) from the Metals Network. The allowable concentration of these pollutants in PM_{10} phase of ambient air is regulated by the European legislation previously mentioned [5,6]. The datasets used contained no missing values initially; subsequently, however, data were systematically removed from these datasets to test the performance of equations (2) and (3) as described below.

Let $\{\hat{x}_1, \ldots \hat{x}_N\}$ denote the $N = 12$ measured monthly concentration values in ng/m^3 of benzo[a]pyrene obtained from the PAH Network or of a given metal (e.g. Ni, As, Cd, Pb) obtained from the Metals Network for a given station during a given year.

In practice, only a sample of size n of the N measured monthly values will be available. In this case, if $\{\hat{z}_1, \ldots, \hat{z}_n\}$ denotes the sample of values obtained by making random draws from $\{\hat{x}_1, \ldots, \hat{x}_N\}$ without replacement, an estimate $\hat{y}$ of the average annual concentration, y, of the pollutant in question is calculated as:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_i \qquad (4)$$

The values $\hat{y}$ and $\hat{z}_i$ are estimates (or realized values) of, respectively, the random variables y and z_i that are related by the model:

$$y = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad (5)$$

Applying a result from [12, section 3.44], the standard uncertainty $u_s(\hat{y})$ associated with $\hat{y}$ as a result of incomplete time coverage is given by:

$$u^2(\hat{y}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sigma^2 \quad (6)$$

where σ is the standard deviation of the population $\{\hat{x}_1, \ldots \hat{x}_N\}$. The value of $u(\hat{y})$ reduces with increasing sample size n, and when $n = N$, the largest possible value of n, $u(\hat{y}) = 0$, which corresponds to complete knowledge about the population of N values. In practice, σ is unknown, and is estimated from the available sample of n values, giving the expression adopted by ISO 11222 [7]:

$$u_a^2(\hat{y}) = \left(1 - \frac{n}{N}\right)\frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{z}_i - \hat{y})^2 \quad (7)$$

We now consider whether equations (4), (6) and (7) given above for $\hat{y}$, $u(\hat{y})$ and $u_a(\hat{y})$ provide credible information about the dispersion of the possible values of y. Specifically, given the population of measured monthly concentration values $\{\hat{x}_1, \ldots, \hat{x}_N\}$ with average μ and variance σ^2, for each $n = 2, \ldots, N$, the following calculations are undertaken:

1. Form all possible samples $\{\hat{z}_1, \ldots, \hat{z}_n\}$ of size n by making random draws from the population of N monthly values $\{\hat{x}_1, \ldots, \hat{x}_N\}$ without replacement within a particular sample. The number of possible samples is $M = N!/[n!\,(N-n)!]$.
2. For each sample of n values, evaluate the average $\hat{y}$ using equation (4) and the standard uncertainty $u_a(\hat{y})$ associated with the average given by equation (7).
3. Calculate the mean $m(\hat{y})$ and standard deviation $s(\hat{y})$ of the set of M average values, and the value of $u(\hat{y})$ given by equation (6) assuming knowledge of σ.

Figures 1–4 show the results of these calculations using the monthly measured values of BaP obtained using the PAH network for the Ballymena Ballykeel site during the year 2011. Figure 1 shows the monthly values $\{\hat{x}_1, \ldots, \hat{x}_N\}$ for which a trend in the values is clearly evident. Figure 2 shows the (true) average concentration value μ (solid blue line), and for each sample size n, the average values $\hat{y}$ for all possible samples of size n (black dots), and the mean $m(\hat{y})$ of those average values (red crosses). It is seen that the values $m(\hat{y})$ agree

Environmental Science: Processes & Impacts Accepted Manuscript

closely with the true value μ, and the dispersion of average values $\hat{y}$ is largest when n is small. Indeed, for small n, $\hat{y}$ can under- or over-estimate μ by as much as a factor of three. Figure 3 shows for each sample size n, the standard deviation $s(\hat{y})$ of the set of possible average values (solid blue line), the standard uncertainty $u(\hat{y})$ calculated using equation (6) (blue circles), and the standard uncertainties $u_a(\hat{y})$ for all possible samples calculated using equation (7) (black dots). The square root of the mean of the variances $u_a^2(\hat{y})$ calculated for all possible samples is also shown (red crosses). It is seen that there is good agreement between the values $s(\hat{y})$ and $u(\hat{y})$, indicating that equation (6), which assumes knowledge of σ performs well. In fact, closer inspection of the values $s(\hat{y})$ and $u(\hat{y})$ shows there is a small systematic difference between the values, which arises because the standard deviation $s(\hat{y})$ is calculated using $M - 1$ in the denominator whereas M is assumed in the derivation of $u(\hat{y})$. The difference will be small provided the number M of possible samples of n values from the population of N values is large.

It is also clear that a value $u_a(\hat{y})$ calculated on the basis of a single sample of size n can under- or over-estimate $s(\hat{y})$, and appreciably so when n is small, although "on average" equation (7) provides a value that is in close agreement with $s(\hat{y})$. Figure 4 shows values of $|\hat{y} - \mu|$ plotted against the corresponding values of $u_a(\hat{y})$ calculated using equation (7) (black dots) for all possible samples. Values satisfying $|\hat{y} - \mu| = u_a(\hat{y})$ and $|\hat{y} - \mu| = 2u_a(\hat{y})$ are also shown as, respectively, the upper and lower (blue) lines. Under a Gaussian assumption it would be expected that $|\hat{y} - \mu| \leq 2u_a(\hat{y})$ with 95 % confidence. Consequently, points below the line $|\hat{y} - \mu| = 2u_a(\hat{y})$ indicate samples for which the calculated standard uncertainty (as a measure of the influence of the missing data on the annual average value) may be understated compared to the deviation of the estimated annual average concentration from the true annual average concentration. Similarly, points appreciably above the line correspond to samples for which the standard uncertainty may be overstated. In ISO 11222 [3], a (finite) degrees of freedom is attached to a (individual) calculated value $u_a(\hat{y})$, and is used to describe the "reliability" of the value regarded as an estimate of the dispersion.

Figures ESI 1 – ESI 4 (available in the ESI) show analogous results for the calculations using BaP data for the Newport site obtained during the year 2010. The data for this site is expected to exhibit much less seasonal character due to the influence of nearby industrial sources of

Environmental Science: Processes & Impacts Accepted Manuscript

the pollutant. There is a noticeable (qualitative) difference between the distributions of the average values $\hat{y}$ for the two sites shown in figures 2 and ESI 2 (available in the ESI). For the Newport site, these distributions have a more symmetric character compared to those for Ballymena Ballykeel.

We now consider a similar treatment for data provided by the Metals Network for the measurement of lead at the Belfast site during 2008, and the results are shown in figures ESI 5 – ESI 8 (available in the ESI). In this case, the presence of a single outlying concentration value (in month seven) has considerable influence on the distributions of standard uncertainties $u_a(\hat{y})$ shown in figure ESI 5 (available in the ESI), which have a bimodal character. Indeed, for small n it can happen that there is no individual value $u_a(\hat{y})$ that agrees with $s(\hat{y})$ (or $u(\hat{y})$) that measures the (true) dispersion of the average values $\hat{y}$.

It seems apparent that the expression used by ISO11222 in equation (7) to account for missing values provides on average a good approximation to equation (6), which describes the situation where there is complete knowledge of the population of values (even those that are missing). However, this analysis has shown that when evaluated for an individual sample there can be an appreciable range of uncertainties obtained using equation (7), and these uncertainties may significantly over or understate the true dispersion of average values arising from the influence of missing values.

Finally, figures 5 and 6 provide a form of "meta-analysis" for the two networks. Figure 5 shows values of $|\hat{y} - \mu|$ plotted against the corresponding values of $u_a(\hat{y})$ calculated using equation (7) for all the sites in the PAH network and for the four years (2008 to 2011) for which a complete data set is available. However, results for all possible samples are not shown (as in figures 4, ESI 4 and ESI 8 (available in the ESI)) but only those for a sample size $n = 6$ for which $|\hat{y} - \mu|$ is largest (black crosses), $u_a(\hat{y})$ is largest (blue circles) and $u_a(\hat{y})$ is smallest (red crosses) corresponding to "extreme" cases. Figure 6 shows the same information for the measurement of lead all the sites in the Metals Network (excepting Motherwell) and for the two years 2009 and 2010 for which a complete data set is available. The results indicate that for both data sets cases arise when the deviation of the estimated annual average concentration from the true value is large and the evaluated uncertainty

Environmental Science: Processes & Impacts Accepted Manuscript

associated with the deviation is small. In these cases the uncertainty is not adequate to quantify the effect of an incomplete data set.

Conclusions

The requirement to calculate in the absence of complete data sets annual averages and associated uncertainties for air pollutants whose concentrations are limited by legislation currently relies on published standard methods. In particular ISO11222 defines a strategy for evaluating the additional uncertainty of annual average values owing to missing data.

It is of particular interest in this work to understand the performance of ISO11222 in the most extreme conditions where it might be applied: for instance where data having low time resolution are collected and missing data accounts for large periods of the year. The results have shown that on average the expression used in ISO11222 is accurate, but that for individual cases the evaluated uncertainty component may be significantly larger or smaller than if full knowledge of the distribution of data was known. It seems that this effect is largest for data showing the greatest spread of values throughout the year – either as a result of pronounced seasonality in concentrations, such as for BaP at non-industrial stations, or as a result of significant variability caused by changing meteorological conditions or variations in industrial processes. We have also shown that the presence of a small number of outlying values can have significant effects on the shape of the distribution of average values and on the evaluated uncertainties.

Our meta-analysis has demonstrated the performance of the ISO11222 approach for metals and BaP using data collected by UK air quality networks. These results have shown that when significant quantities of data are missing care must be taken to ensure that the uncertainty associated with the lack of time coverage does indeed provide a proper representation of this effect. On occasions, for instance when considering BaP concentrations that show predictable variation during the year, inspection of the data in the time domain may suggest that uncertainty owing to incomplete data coverage is likely to have been understated, especially if the missing data are expected to be high concentration values at the beginning and end of the year.

It is hoped that this work will assist those performing metals and PAH monitoring to formulate better their uncertainty budgets for annual average values to take better account of

Environmental Science: Processes & Impacts Accepted Manuscript

missing data and also inform policy makers and standardization experts when formulating the next generation of air quality policy.

Environmental Science: Processes & Impacts Accepted Manuscript

Environmental Science: Processes & Impacts Accepted Manuscript

Figure Captions

Figure 1. Measured monthly concentration values of benzo[a]pyrene from the PAH network at Ballymena Ballykeel in 2011.

Figure 2. For Ballymena Ballykeel in 2011, the (true) average benzo[a]pyrene concentration value $\boldsymbol{\mu}$ (solid blue line), and for each sample size $\boldsymbol{n}$, the average values $\hat{\boldsymbol{y}}$ for all possible samples of size $\boldsymbol{n}$ (black dots) and the mean $\boldsymbol{m}(\hat{\boldsymbol{y}})$ of those average values (red crosses).

Figure 3. For Ballymena Ballykeel in 2011, and for each sample size $\boldsymbol{n}$, the standard deviation $\boldsymbol{s}(\hat{\boldsymbol{y}})$ of the set of possible average benzo[a]pyrene values (solid blue line), the standard uncertainty $\boldsymbol{u}(\hat{\boldsymbol{y}})$ calculated using equation (6) (blue circles), and the standard uncertainties $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ for all possible samples calculated using equation (7) (black dots). The square root of the mean of the variances $\boldsymbol{u_a^2}(\hat{\boldsymbol{y}})$ calculated for all possible samples is also shown (red crosses).

Figure 4. For Ballymena Ballykeel in 2011, values of $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}|$ plotted against corresponding values of $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ (black dots). Values satisfying $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| = \boldsymbol{u_a}(\hat{\boldsymbol{y}})$ and $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| = \boldsymbol{2u_a}(\hat{\boldsymbol{y}})$ are also shown as, respectively, the upper and lower (blue) lines: 63 % of points satisfy $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| \leq \boldsymbol{u_a}(\hat{\boldsymbol{y}})$ and 85 % of points satisfy $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| \leq \boldsymbol{2u_a}(\hat{\boldsymbol{y}})$.

Figure 5. Meta-analysis for the PAH Network, showing values of $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}|$ plotted against corresponding values of $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ for all sites in the PAH Network and for the four years 2008 to 2011. Results are shown for a sample size $\boldsymbol{n = 6}$ for which $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}|$ is largest (black crosses), $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ is largest (blue circles) and $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ is smallest (red crosses). Values satisfying $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| = \boldsymbol{u_a}(\hat{\boldsymbol{y}})$ and $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| = \boldsymbol{2u_a}(\hat{\boldsymbol{y}})$ are also shown as, respectively, the upper and lower (blue) lines: 63 % of points satisfy $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| \leq \boldsymbol{u_a}(\hat{\boldsymbol{y}})$ and 87 % of points satisfy $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| \leq \boldsymbol{2u_a}(\hat{\boldsymbol{y}})$.

Figure 6. Meta-analysis for the Metals Network, showing values of $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}|$ plotted against corresponding values of $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ for lead measured at all sites in the Metals Network (except Motherwell) and for the two years 2009 and 2010. Results are shown for a sample size $\boldsymbol{n = 6}$ for which $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}|$ is largest (black crosses), $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ is largest (blue circles) and $\boldsymbol{u_a}(\hat{\boldsymbol{y}})$ is smallest (red crosses). Values satisfying $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| = \boldsymbol{u_a}(\hat{\boldsymbol{y}})$ and $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| = \boldsymbol{2u_a}(\hat{\boldsymbol{y}})$ are also shown as, respectively, the upper and lower (blue) lines: 62 % of points satisfy $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| \leq \boldsymbol{u_a}(\hat{\boldsymbol{y}})$ and 87 % of points satisfy $|\hat{\boldsymbol{y}} - \boldsymbol{\mu}| \leq \boldsymbol{2u_a}(\hat{\boldsymbol{y}})$.

Environmental Science: Processes & Impacts Accepted Manuscript

Acknowledgements

The funding of the National Measurement System's Chemical and Biological Metrology and Innovation Research and Development Programmes by the UK Department for Business Innovation and Skills is gratefully acknowledged.

Environmental Science: Processes & Impacts Accepted Manuscript

References

[1] Air Quality - Existing Legislation: http://ec.europa.eu/environment/air/quality/legislation/existing_leg.htm (Accessed January 2014).

[2] R. J. C. Brown, P. M. Harris, M. G. Cox, Improved strategies for calculating annual averages of ambient air pollutants in cases of incomplete data coverage, *Environ. Sci.: Processes Impacts*, 2013, **15**, 904-911.

[3] R. J. C. Brown, Data loss from time series of pollutants in ambient air exhibiting seasonality: consequences and strategies for data prediction, *Environ. Sci.: Processes Impacts*, 2013, **15**, 545-553.

[4] R. J. C. Brown, A. S. Brown, K. H. Kim, A temperature-based approach to predicting lost data from highly seasonal pollutant data sets, *Environ. Sci.: Processes Impacts*, 2013, **15**, 1256-1263.

[5] European Commission, Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe, *Official Journal of the European Union, L: Legislation (English Edition)*, 2008, **152**, 1–44.

[6] European Commission, Directive 2004/107/EC of the European Parliament and of the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air, *Official Journal of the European Union, L: Legislation (English Edition)*, 2005, **23**, 3–16.

[7] International Standards Organisation, *ISO 11222:2002 Air Quality – Determination of the uncertainty of the time average of air quality measurements*, ISO, Geneva, 2002.

[8] Defra UK-AIR Monitoring Networks: http://uk-air.defra.gov.uk/networks/ (Accessed April 2014).

[9] R. J. C. Brown, D. M. Butterfield, S. L. Goddard, C. L. Mustoe, C. Robins, A. S. Brown, S. Beccaceci, K. J. Whiteside, M. Williams, *NPL Report Number AS 69: Annual Report for 2011 on the UK Heavy Metals Monitoring Network*, Teddington, 2012.

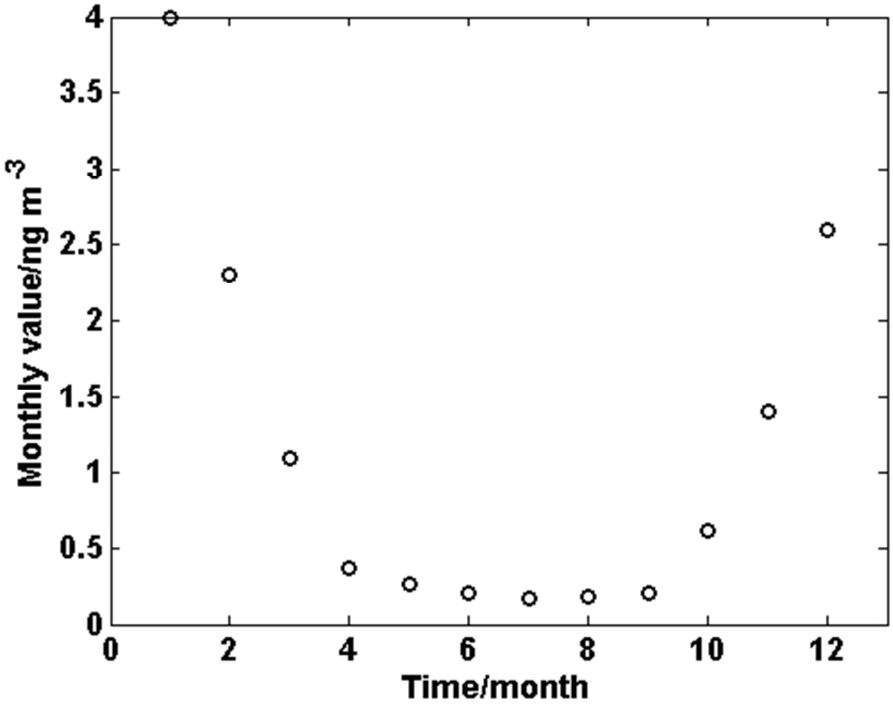
[10] A. S. Brown, D. Sarantaridis, D. M. Butterfield, R. J. C. Brown, K. J. Whiteside, P. Hughey, S. L. Goddard, D. Hussain, M. Williams, *NPL Report Number AS 71: Annual Report for 2011 on the UK PAH Monitoring and Analysis Network*, Teddington, 2012.

[11] UK-AIR: Air information resource. Data archive. http://uk-air.defra.gov.uk/data/ (accessed April 2014).

Environmental Science: Processes & Impacts Accepted Manuscript

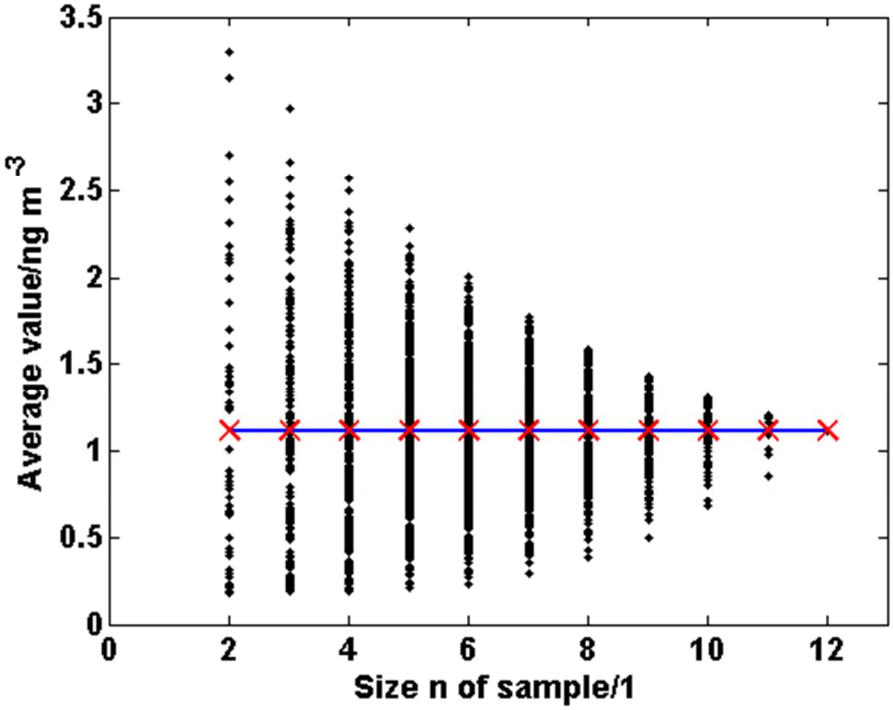
[12] C. A. Bennett and N. L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry*, John Wiley, New York, 1963.

Environmental Science: Processes & Impacts Accepted Manuscript



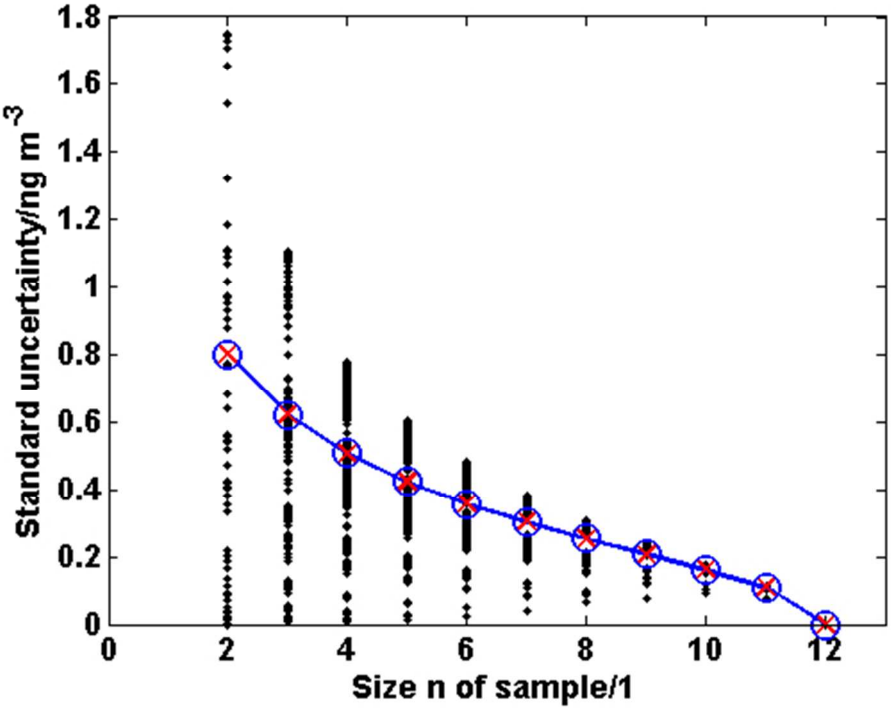
148x110mm (96 x 96 DPI)

Environmental Science: Processes & Impacts Accepted Manuscript



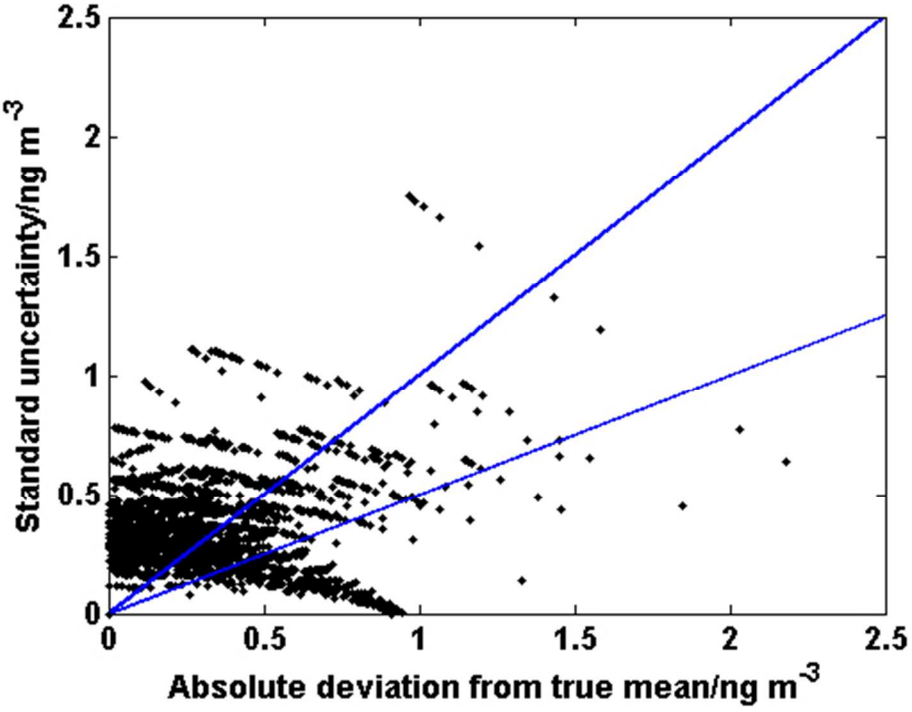
148x110mm (96 x 96 DPI)

Environmental Science: Processes & Impacts Accepted Manuscript



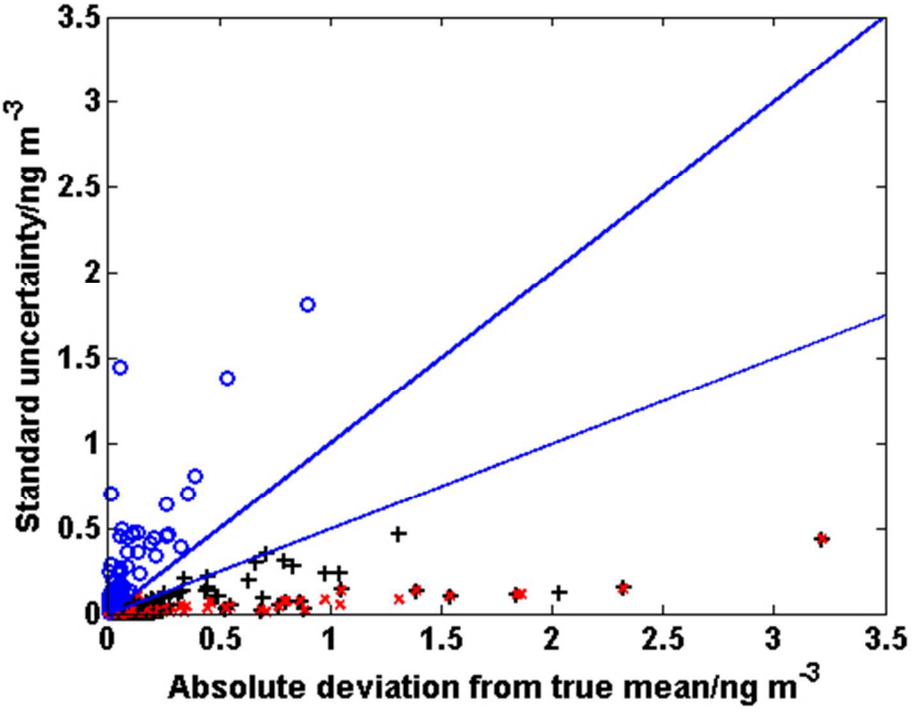
148x110mm (96 x 96 DPI)

Environmental Science: Processes & Impacts Accepted Manuscript



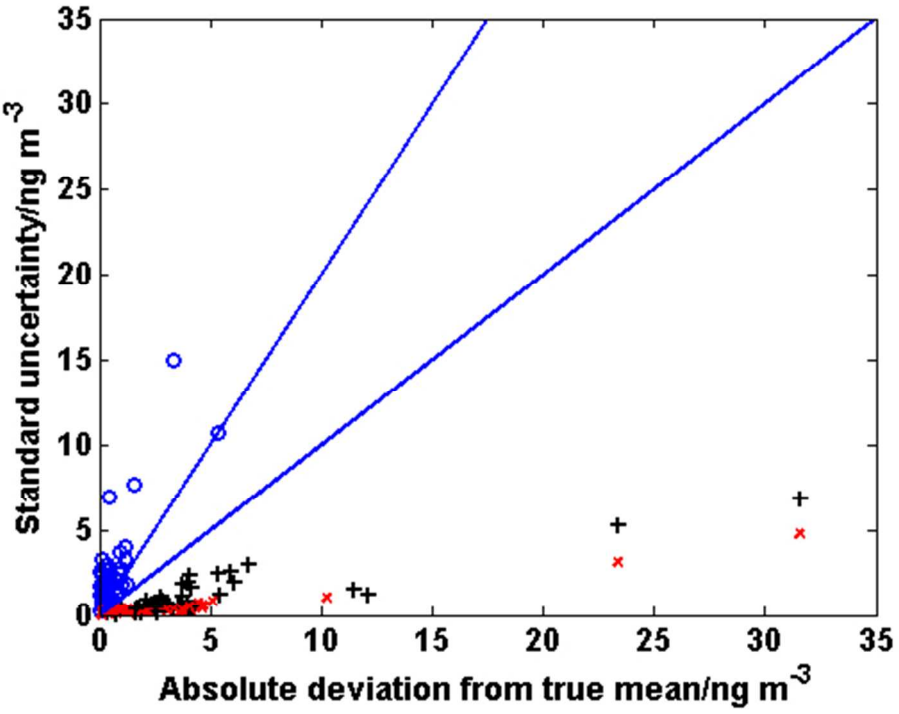
148x110mm (96 x 96 DPI)

Environmental Science: Processes & Impacts Accepted Manuscript



148x110mm (96 x 96 DPI)

Environmental Science: Processes & Impacts Accepted Manuscript



148x111mm (96 x 96 DPI)

Environmental Science: Processes & Impacts Accepted Manuscript