British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

defra
Department for Environment Food and Rural Affairs

Methodology for the determination of normal background contaminant concentrations in English soils

Land Use Planning and Development Programme

Commissioned Report CR/12/003 N

BRITISH GEOLOGICAL SURVEY

LAND USE PLANNING AND DEVELOPMENT PROGRAMME
COMMISSIONED REPORT CR/12/003N

Methodology for the determination of normal background contaminant concentrations in English soils

MR Cave, CC Johnson, EL Ander and B Palumbo-Roe

Contributor

R M Lark

B G Rawlins

Keywords

Contaminated land, soil, background, geochemistry.

Bibliographical reference

Cave, MR, Johnson, CC, Ander, EL and Palumbo-Roe, B. 2012. Methodology for the determination of normal background contaminant concentrations in English soils. *British Geological Survey Commissioned Report*, CR/12/003. 42pp.

Keyworth, Nottingham British Geological Survey 2012

BRITISH GEOLOGICAL SURVEY

The full range of our publications is available from BGS shops at Nottingham, Edinburgh, London and Cardiff (Welsh publications only) see contact details below or shop online at www.geologyshop.com

The London Information Office also maintains a reference collection of BGS publications, including maps, for consultation.

We publish an annual catalogue of our maps and other publications; this catalogue is available online or from any of the BGS shops.

The British Geological Survey carries out the geological survey of Great Britain and Northern Ireland (the latter as an agency service for the government of Northern Ireland), and of the surrounding continental shelf, as well as basic research projects. It also undertakes programmes of technical aid in geology in developing countries.

The British Geological Survey is a component body of the Natural Environment Research Council.

British Geological Survey offices

BGS Central Enquiries Desk

Tel 0115 936 3143 Fax 0115 936 3276

email enquiries@bgs.ac.uk

Kingsley Dunham Centre, Keyworth, Nottingham NG12 5GG

Tel 0115 936 3241 Fax 0115 936 3488

email sales@bgs.ac.uk

Murchison House, West Mains Road, Edinburgh EH9 3LA

Tel 0131 667 1000 Fax 0131 668 2683

email scotsales@bgs.ac.uk

Natural History Museum, Cromwell Road, London SW7 5BD

Tel 020 7589 4090 Fax 020 7584 8270

Tel 020 7942 5344/45 email bgslondon@bgs.ac.uk

Columbus House, Greenmeadow Springs, Tongwynlais, Cardiff CF15 7NE

Tel 029 2052 1962 Fax 029 2052 1963

Maclean Building, Crowmarsh Gifford, Wallingford OX10 8BB

Tel 01491 838800 Fax 01491 692345

Geological Survey of Northern Ireland, Colby House, Stranmillis Court, Belfast BT9 5BF

Tel 028 9038 8462 Fax 028 9038 8461

www.bgs.ac.uk/gsni/

Parent Body

Natural Environment Research Council, Polaris House, North Star Avenue, Swindon SN2 1EU

Tel 01793 411500 Fax 01793 411501

www.nerc.ac.uk

Website www.bgs.ac.uk

Shop online at www.geologyshop.com

Foreword

This report presents the results from the third work package of a Defra-funded Science and Research project to establish normal background contaminant concentrations in the soils of England. The project (Project reference SP1008: *Establishing data on normal/background levels of soil contamination in England*) commenced 5th October 2011 and is scheduled to end 31st March 2012. Work package 1 (WP1) was concerned with a review of existing data and Work package 2 (WP2) an exploration of the data. A methodology to determine these concentrations, Work package 3 (WP3), is reported here. Technical guidance in the use of normal background concentrations will be written for contaminants for which NBCs can be determined by the end March 2012 (Work Package 4, WP4).

Acknowledgements

The approach to this work is based on proposals prepared before this project commenced and the authors are grateful to the valuable contribution made by Murray Lark and Barry Rawlins to some of the ideas and concepts used in the methodology presented here.

Contents

FIGURES

TABLES

Summary

The land surface of England has been divided into domains for purposes of defining background soil concentrations of arsenic (As), lead (Pb) and benzo[*a*]pyrene (BaP). Within any domain for a particular substance there may be one or more factors (anthropogenic or geogenic) which guide us to expect elevated background concentrations of a substance. In the case of As, the Ironstone and Mineralisation Domains have elevated concentrations from geogenic sources. Other sites, with no particular factors causing elevated concentrations, constitute the Principal Domain. In the case of Pb there are elevated concentrations from anthropogenic sources in the Urban Domain, geogenic sources in the Mineralisation Domain, and a Principal Domain where elevated concentrations are not expected. Because there are much less data for BaP, both England and Britain are divided into just two domains: Urban and Principal (*i.e*. non-urban).

For each contaminant, the normal background concentration (NBC), for each of the three contaminants within their domains, have been determined in a systematic and robust statistical manner which is summarised in a methodology flow diagram. First, the statistical distributions of contaminant concentrations are characterised for each domain. Histograms or density plots and summary statistics (the skewness coefficient and the octile skewness coefficient) are used to judge whether the data can assumed to come from a Gaussian variable and whether outliers are present in the data set. If the distribution is not Gaussian then the data are transformed so that the distribution becomes Gaussian by either taking the natural logarithm or in some instances a Box-Cox transform of the data. The NBC is set by taking the upper 95% confidence limit of the 95th percentile of the distribution.

1 Introduction

1.1 BACKGROUND

The work described here is part of the process to simplify the contaminated land regime for England and Wales where there is a legacy of land contamination from industrial activity and urbanisation. Statutory guidance is issued by the Secretary of State for Environment, Food and Rural Affairs (Defra) in accordance with section 78Y of the Environmental Protection Act 1990. Section 57 of the Environment Act 1995 created Part 2A of the Environment Protection Act 1990 establishing a legal framework for dealing with contaminated land (DETR, 2000; Defra, 2006). The Statutory Guidance is intended to explain how the contaminated land regime should be implemented. However, the Guidance, which is supposed to explain when land does (and does not) need to be remediated, has created significant uncertainties. Therefore, revision of the Statutory Guidance intends to make it more usable for those working with contaminated land and remediation (Defra, 2011a). This has been previously described in the first Project report (Ander *et al*., 2011) and a methodology to determine normal background contaminant concentrations in soils is part of the revision process.

1.2 PROJECT OBJECTIVES

The objectives covered by this report are detailed in the Project proposals (BGS, 2011) which form part of the contract of work and specifically relate to Work Package 3 (WP3). This follows on from the work of Work Packages 1 (WP1 – Review of existing data) and 2 (WP2 - Data Exploration) both of which have been described in detail in Ander *et al*., 2011. The objective behind WP3 is to present a statistically robust methodology for the determination of normal background concentrations (NBCs) of contaminants in English soils. The NBC will be a representation of the "normal levels" described in the Statutory Guidance. The initial ideas for this methodology were presented in the Project proposal document (BGS, 2011) and were also discussed at the Project Workshop held at BGS Keyworth on 22nd November 2011. The methodology is tested on three contaminants – arsenic (As), lead (Pb) and Benzo[*a*]pyrene (BaP). These represent contaminants with varied natural and diffuse anthropogenic origin and have differing amounts of available data on which to base NBCs estimates. Asbestos was investigated in WP1 and WP2 but, in view of the scarcity of available information on this contaminant, it is considered inappropriate to apply any statistical methodology to determine NBCs for such a contaminant.

1.3 APPROACHES TO DETERMINING BACKGROUND CONCENTRATIONS

The definition of the term "normal background" as applied to contaminant concentrations in soil was explored as part of WP1 (Ander *et al*., 2011) and in its simpliest sense refers to the levels of a contaminant one might expect in a sample of soil from a specified location. For this Project the normal background concentration is the level of a contaminant in a soil arising from a combination of natural processes that characterise a soil, including diffuse source anthropogenic inputs. The term background is defined in many different ways and varies from discipline to discipline. Matschullat *et al*. (2000) investigated the concept of a geochemical background and how it can be calculated and comment that the numerous citations demonstrate the need for such a term. They also revealed the lack of a clear definition or agreement in its use. The revised Statutory Guidance, Section 3, (Defra, 2012) usefully discusses how "normal" presence of contaminants arises and also states what they are not.

Methodologies to determine background concentrations are as numerous as there are definitions for background. There are important strategic, economic and legislative drivers for understanding and quantifying soil element/contaminant concentrations. From an economic and strategic point of view, the exploration and development of economic metalliferous mineral deposits by geochemical exploration has meant a lot of resources and investigations have gone into methodologies to determine backgrounds. For more than sixty years, statistical methods have been used to distinguish between anomalous and background concentrations of the chemical elements in soils in order to locate buried mineralisation (*e.g.* Lovering *et al*., 1950; Hawkes and Bloom, 1955; Tidball *et al*., 1974). Some of these techniques are described in detail by Matschullat *et al*. (2000) with application examples and include:

- The Lepeltier method
- Relative cumulative frequency curves
- Normality of sample ranges
- Regression techniques
- Mode analysis
- 4-σ outlier test
- Interactive 2 σ technique
- Calculated distribution function

Grunsky (2010), in a more recent review of interpreting geochemical survey data, discusses graphical methods for differentiating geochemical background from anomalies. The mean plus 2σ is a commonly used approach and for a normal data distribution this would represent approximately 97% of a data population.

Legislation concerned with healthy and sustainable environments is also now a significant driver for information on background contaminant concentrations. For this purpose, documents such as the British Standard guidance on the determination of background values (ISO 2011) have been published. ISO 19258:2011 covers the prerequisites of sampling, analysis and data handling and outlines some essentials of statistical evaluation of data. A good example of the statistical analysis of a large soil data set is that of Oliver *et al*. (2002) who present a statistical and geostatistical analysis of the National Soil Inventory (NSI *aqua regia*) of England and Wales.

Matschullat *et al*. (2000), ISO 19258:2011 (ISO 2011) and Grunsky (2010) serve to illustrate the fact that the determination of background requires good quality data on element/contaminant concentrations in soil and a statistical methodology to deliver estimates for background concentrations. Work package 1 dealt with the availability and robustness of data. The proposed statistical methodology is the subject of this report.

The majority of research to date has focused on methods for providing typical background concentrations of potentially harmful elements (PHEs) in soil (*e.g*. Appleton, 1995; Appleton *et al*., 2008). Geochemists express the geochemical baseline (a spatially fluctuating chemical environment at a given point in time) in terms of the natural baseline (stable over long periods of time) with an overprint of the anthropogenic baseline (one or many contributing sources) that changes over a relatively short period of time (Johnson and Ander, 2008). An understanding of what constitutes the natural baseline enables the contribution of the anthropogenic component to be estimated. These approaches to determining “backgrounds” are largely based on soil sampling and analyses over different parent material groups which have been shown to exert a dominant control on topsoil chemistry in England (Rawlins *et al*.,

2003). Alternative approaches have been investigated based on associations with particle size fractions across England and Wales (Zhao *et al*., 2007) or globally based on statistical relationships with total soil iron or manganese (Hamon *et al*., 2004). An alternative approach, proposed by Appleton *et al*. (2008) is to estimate typical background concentrations from a statistical measure (*e.g*. the geometric mean) based on existing soil analyses within soil parent material polygons. The background concentrations for a particular PHE are mapped using delineations of the parent material polygons.

In the short time available to the Project, it has not been possible to investigate all the different approaches used with regard to national legislation in other countries around the world. However, it is interesting to note some of the methodology being used to address background estimations in some selected countries around Europe. Paterson *et al.* (2003), in a report commissioned by the Scottish Environment Protection Agency (SEPA), have looked at background levels of contaminants is Scottish soils. Simple statistical analysis of small data populations (289 samples) produced basic information that could be used as an indication of background but only rural soil samples were used in the analysis.

In Italy, APAT-ISS (2006) gives government guidance for the determination of background values of metals and metalloids in Italian soils. The national guidance adopts the ISO 19258:2011 (ISO 2011) definition of natural background concentration, described as the concentration of a substance in the soil that is derived from geological or pedological processes including also diffuse contributions. The stepwise approach for deriving background values involves the collection of data, the statistical analysis of the data and the determination of the background value. The selection of the sampling sites follows the typological approach (based on parent material, soil type and land use), choosing sites within homogeneous areas. The statistical analysis is carried out on data sets, each representative of homogenous typologies. The descriptive statistics for data distribution include the minimum, maximum, median, percentile, standard deviation, skewness, kurtosis and graphical representations such as box plots, histograms and percentage cumulative frequency plots. The guideline describes in detail a series of statistical tests to identify outliers and to define the distribution type of the data (normal, lognormal, gamma, non-parametric distribution) among which are the W test, D'Agostino Test, Normal Q-Q Plot, Lilliefors Test, Gamma Q-Q Plot, Kolmogorov-Smirnov Test and Anderson Darling Test. The background value is defined as the 95th percentile of the population. In the assessment processes of contaminated land for a particular site the site-specific data are compared with the background data population (the two populations having common parent material, soil type). For comparison of the background and site-specific data the guideline also indicates different statistical test on the basis of the distribution type of the populations (*e.g* Slippage test, Quantile test, Wilcoxon rank sum test, Gehan test, t-student test and t-Satterthwaite test).

In Finland, a Government Decree on the Assessment of Soil Contamination and Remediation Needs (214/2007) (Finnish Government Decree, 2007) became legislation on 1 June 2007 (Jarva *et al*., 2010; Tarvainen and Jarva, 2011). The decree defines a geochemical baseline as being the natural geochemical background concentration and superimposed diffuse anthropogenic input of elements in the topsoil. Backgrounds are assessed on a local investigation of the geochemical baseline rather than on national values and the upper limit of geochemical baseline variation for element X (BL_X) is calculated as follows:

$$BL_X = P_{75} + 1.5(P_{75} - P_{25})$$

Where P_{75} is the 75th percentile of element X concentrations and P_{25} is the 25th percentile of element X concentrations.

An important point made by many of the accounts looking at methodologies for background determinations (*e.g*. Matschullat *et al*., 2000; Reimann and Garrett, 2005; Tarvainen and Jarva, 2011) is that estimations are very dependent on location and scale. It is for this reason that the domain approach explored in WP2 (Ander *et al*., 2011) forms an important part of the NBC classification described in the following section.

2 Methodology

Work Package 2 defined the domains for the selected contaminants (As, Pb and BaP) from either natural or diffuse pollutions sources (Ander *et al*., 2011) and the next stage is to define the population of values of a contaminant in each domain and an upper limit for the population which could define the boundary of where the natural or diffuse pollution signature in the soil ends. Soils with contaminant concentrations above this value are then considered to be derived from point source pollution. This boundary for each domain is the NBC.

2.1 STATISTICAL ANALYSIS METHODOLOGY

The available data are assumed to conform to a linear mixed model of the following form.

Z = fixed effects + continuous random variation + point contamination

where Z is a random variable that represents the contaminant data derived from WP2. The fixed effects are sources of variation in the observed concentrations that are attributable to geogenic sources or diffuse anthropogenic activities. The identification of domains in WP2 should capture this, with the domain mean representing the fixed effect. The continuous random variation (typically Gaussian (normal) or log-Gaussian) represents the typical variation arising from geogenic or diffuse anthropogenic sources within the defined domains. The point contamination is assumed to introduce outlying values into the data. The equation above can be re-written informally as:

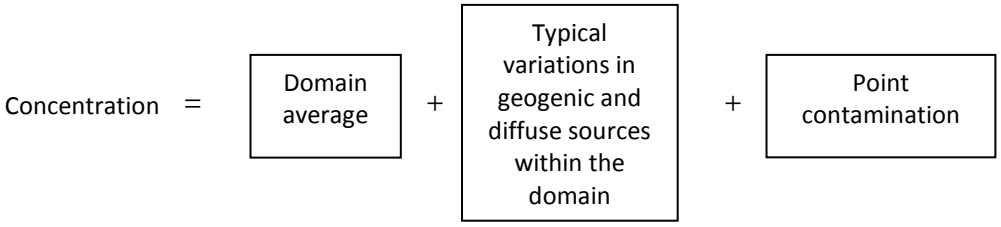

Figure 1: Conceptual model of the contaminant concentration in soil

For any contaminant the first two terms (domain average + typical variations) give rise to the normal range of values or normal variation of the contaminant. The objective of the procedure is to characterise this normal variation in terms of a statistical distribution. Some percentile of this distribution will then be defined as a limit on the typical variation of the contaminant for the domain, used to decide whether a particular value represents point contamination. The following methods allow the typical distribution from available data to be characterised, given that these may contain point contamination data.

2.1.1 Exploratory analysis

An initial exploratory analysis is necessary in order to identify an appropriate form for the distribution of the continuous random variation, to decide whether or not point contamination appears in the data and then to derive appropriate statistics to define the typical range of variation. This is achieved using histograms, density plots and summary statistics.

The fixed effects will be represented by the domains of the classification selected to represent geogenic and anthropogenic sources of normal concentrations of the specific contaminant in soil defined in WP2. Each domain will be examined in turn.

2.1.2 Skewness

The shape of the data distribution within each domain can be used to discriminate between the normal random variation and the point source variation. A key descriptive parameter of the distribution that can be used for this purpose is the skewness of the distribution. A distribution is said to be skewed if one of its tails is longer than the other. The first distribution shown in Figure 2 has a positive skew. This means that it has a long tail in the positive direction. The second distribution has a negative skew since it has a long tail in the negative direction. Finally, the third distribution is symmetric and has no skew.

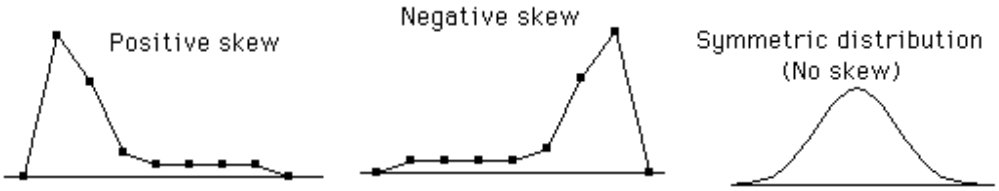

Figure 2: Examples of skewed distributions

The skewness coefficient (SC) of a distribution can be calculated as:

$$SC = \sum(x^i-\mu)^3/N\sigma^3$$

where μ is the mean, σ is the standard deviation, x_i is the i^{th} value of data set and N is the number of data points. The normal distribution has an SC of 0 since it is a symmetric distribution.

As a general rule, the mean is larger than the median in positively skewed distributions and less than the median in negatively skewed distributions. Distributions with positive skew are, in general, more common in geoscience than distributions with negative skews (Reimann and Filzmoser, 2000). In geochemical studies it is a common rule of thumb that the analysis of data can proceed on the assumption that they come from a normal random variable on the basis of inspection of the histogram and a value of the SC is in the interval [-1, 1] then a Gaussian (normal) distribution will give a good fit to the data (Rawlins *et al*., 2005). The conventional SC coefficient is, however, highly influenced by outliers in the data so that the inclusion of a few outlier data points to one tail of a substantially Gaussian data set gives rise to large SCs indicating non-Gaussian behaviour. A measure of skewness which is less susceptible to outliers is the octile skewness (OS) coefficient (Brys *et al.,* 2003) which is defined as:

$$OS = ((Q_{0.875} - Q_{0.5})-(Q_{0.5} - Q_{0.125}))/(Q_{0.875} - Q_{0.125})$$

where Q_n is the n^{th} quantile of the data set.

The OS does not include the extreme tails of the distribution in its calculation and is therefore not influenced by outliers. A rule of thumb for interpretation of the OS (comparable the SC), is that we can treat data as symmetrically distributed if SC falls in the interval [–0.2,0.2]. (Rawlins *et al*., 2005).

The different properties of the SC and the OS can be exploited to highlight the possible occurrence of outliers in a data set. If the OS and SC of a given geochemical data set are calculated and the SC is in the range [-1, 1] and the OS is in the range [-0.2, 0.2] then the data set can be treated as symmetrically distributed. If, however, the SC >1 but the OS is <0.2 then it is likely that the data are Gaussian but there are outliers in the right hand tail of the distribution. If the SC >1 and the OS >0.2 then the distribution is likely to have a non-Gaussian distribution.

2.1.3 Data transformation

Skewness in the shape of a distribution may result from: (i) an underlying non-Gaussian distribution (systematic skewness), or (ii) presence of outlying values, or (iii) both of these. It is important to distinguish the situation as far as possible when planning further analysis of data to characterise a domain. This is because we can most efficiently characterise a variable from sample data when it is from a Gaussian random variable, and if it is not then a transformation of the data is desirable. Data may be transformed to logarithms, or in some cases by the Box-Cox transform of which the log-transform is a special case. However, we should only transform data if they are systematically skewed, and not if skewness is entirely due to outliers. If, for example, our data have an OS which is in the interval [-0.2,0.2] but a SC larger than 1 then this suggests that the underlying distribution is symmetrical, and a transformation is not appropriate. If, however, the OS were larger than 0.2, then a transformation should be considered.

Where it is possible (perhaps after transformation) to assume a particular form for the distribution of the typical variation of a substance, then it is generally preferable to use such an estimated distribution to define limits of typical variation (since empirical percentiles, particularly of small samples, can be erratic). Where this is not possible then empirical percentiles of the distribution may be used, although larger percentiles may be influenced by outliers. Given a robust estimate of parameters of the frequency distribution (Gaussian or transformed) of typical (geogenic and anthropogenic) concentrations for a given domain, it is possible to partition our observations into those most likely to represent typical variation and those most likely to represent point contamination (Lark, 2002; Rawlins *et al*., 2005).

2.1.4 Setting limits for normal concentrations

Having defined the shape of the data distributions within each domain for contaminants from either natural or diffuse pollutions sources, the next stage is to define a value, based on the distributions, which can be used in to assess whether the contaminant concentration in a soil found within a given domain is from normal background or from point source contamination.

Before moving on to the evaluation of the data distribution of a given domain, it is necessary to assess whether the data is representative of the true population of soil contaminant concentrations within the domain. The ISO 19258:2011 standard “Soil quality: Guidance on the determination of background values” (ISO 2011) recommends a minimum number of 30 samples. In addition to this, the samples should be spatially distributed to ensure they are representative of the underlying true data population.

A value as defined by the median (50th percentile) of a soil contaminant concentration data set gives a measure of the central tendency of the distribution but, on its own, would not be helpful as, by definition, 50% of the data will be above this value. We need to ask the question "What is the highest concentration of contaminant in this domain that is likely to come from normal background?" In this case it would be sensible to use a percentile which encompasses most of the data. In probabilistic risk assessment and other approaches to defining background (APAT-ISS, 2006) the 95th percentile of the data distribution is used and has been used in the past in the CLEA (Contaminated Land Exposure Assessment) risk assessment tool for soil contamination (Defra-EA, 2002).

2.1.5 Statistical analysis

The statistical analysis of the contaminant concentrations domains has been carried out using the R programming language (R Development Core Team, 2011). Figures 3a and b show the flow chart for the statistical procedure used to derive NBC. Part I of the process (Figure 3a) is the essentially the data gathering and exploration phase (WP1&2) in which the contaminant results are attributed to domains. Question 1 asks if the contaminant is suitable for a NBC. Asbestos and manufactured organic contaminants with no natural origin, for example, fail this question. The data exploration (Ander *et al*., 2011) identifies the areas (domains) where there are clearly identifiable controls on high concentrations of a specified contaminant. The contaminant data set is then subdivided into domain data sets. A minimum of 30 results are considered necessary to determine a NBC (see Section 2.1.4). The initial statistical analysis of a domain data set is to plot the data distributions (density distribution and histogram plots) and calculate the skewness coefficient (SC) and octile skewness coefficient (OS) (see Section 2.1.2). It must be emphasised that this is not an automated procedure for generating NBCs and in addition to calculating statistical measures it is important to inspect distribution plots of untransformed and transformed data throughout the procedure. The R code used in this statistical analysis is available as a Project resource from the BGS Project website[1].

Figure 3b shows parts II – IV of the procedure which is a series of skewness testing steps (box 3 in Figure 3b) in which the SC and OS, along with the data density plot, are used to assess a whether a Gaussian fit is appropriate, the presence of outliers and whether a data transformation is necessary. The central philosophy of the method requires viewing of the data distribution as a histogram or data density plot and using judgement of the shape of the plot along with the SC and OS to decide whether the data are consistent with the assumption of an underlying Gaussian random variable. The steps applied to the data are:

i) If the data are symmetrically distributed (SC is <1 and the OS is <0.2, *i.e.* TEST 2 in Figure 3b) then the data are consistent with the assumption of a Gaussian distribution and the parametric percentiles are fitted based on the mean and standard deviation of the data.

ii) If the data show a mostly symmetrical distribution with potential outliers in the distribution tail (SC >1 but OS <0.2, *i.e.* TEST 3 in Figure 3b) then the data are consistent with the assumption of a Gaussian distribution and the parametric percentiles are fitted using median and the median absolute deviation (MAD) in place of the mean and standard deviation as these measures are robust to outliers (Reimann and Filzmoser, 2000).

iii) If the data distribution is skewed (SC is >1 and the OS is >0.2, *i.e.* TEST 1 in Figure 3b) then the data is not suitable for fitting to a Gaussian model and the data need to be

[1] http://www.bgs.ac.uk/gbase/NBCDefraProject.html

transformed to using either a logarithmic or Box-Cox transform (Reimann and Filzmoser, 2000). After transform the distribution is re-examined using parts III and IV. After calculation of the percentiles the data are back transformed to their original units;

iv) Finally, if the data cannot be made to be consistent with a Gaussian distribution (even after transform) the empirical percentiles for the data set are calculated.

In practice the empirical, parametric and robust percentiles have been reported for each domain to check for consistency between methods. The methodology assumes that data for a given domain comes predominantly from a single population of data and that the data are either normally distributed or have a positive SC. For the contaminants and domains considered in this report these assumptions hold true (see Section 3).

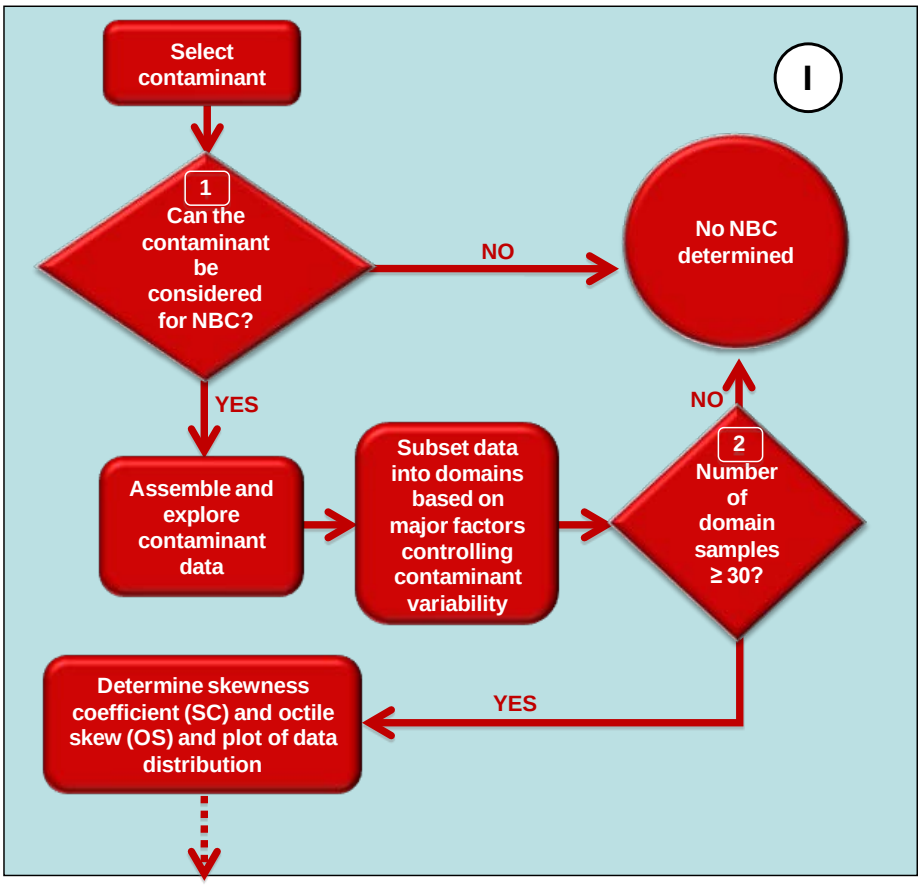

Figure 3a: Flow chart for the calculation of the NBC for a given contaminant – part I: Data gathering and exploration (see text for explanation)

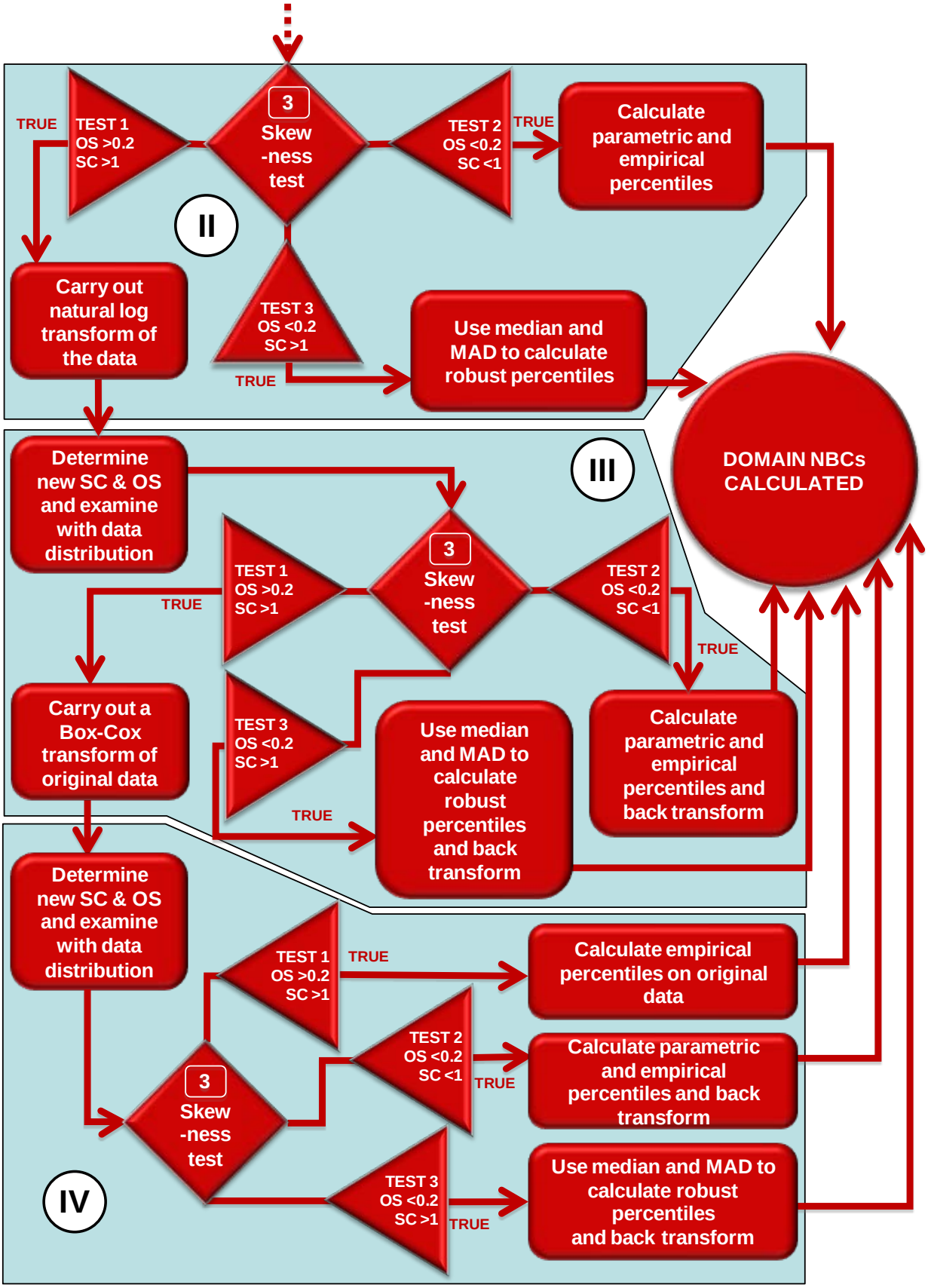

Figure 3b: Flow chart for the calculation of the NBC for a given contaminant – parts II - IV: Skewness testing and transformations (see text for explanation). MAD is the mean absolute deviation.

2.1.6 Uncertainty in normal background concentrations (NBCs)

For each contaminant domain percentile values from the 50th to the 95th in steps of 5 have been calculated based on empirical portioning of the raw data, the percentiles based on the percentiles of a Gaussian fit to the data set (taking into account the need for data transformation), and the percentiles of a Gaussian fit calculated using the mean and the MAD of the data (taking into account the need for data transformation). The percentile values are subject to uncertainty based on the number of data points and the shape of the distribution. An assessment of the uncertainty on the percentiles was calculated by empirical, parametric and robust parametric methods and has been included in the statistical calculations using a bootstrap resampling routine implemented using the "boot" package within the R programming language, based on the seminal work of Davison and Hinkley (1997). The bootstrap routine used 1000 resamples of the original or transformed data providing a 95% percentile confidence interval on the calculated percentile (Efron, 1987).

In Section 2.1.4 a justification was made for using the 95th percentile of the distribution to define the upper boundary of normal contamination for a given contaminant and domain. Calculation of the uncertainty using bootstrapping gives an upper limit to the calculated 95th percentile. Examination of the uncertainty on the calculations shown in Section 3 confirms that the parametric uncertainties from the Gaussian fit are much less erratic than the empirical values, then the parametric limits have been chosen as being a better representation of the uncertainty on the percentiles.

As it has been argued that the NBC should represent the highest concentration of contaminant in this domain that is likely to come from normal background, then the NBC should be the upper end of the confidence interval. **The NBC is defined as the upper 95% confidence limit of the 95th percentile (taking into account data transformations).**

3 Results of statistical analysis

For each contaminant and domain, density plots of the raw and transformed data are presented along with the SC and OS values. Plots of the empirical and parametric and robust percentiles are also provided along with the relative uncertainties for both types of percentile. The percentiles are also given in a table giving the empirical, parametric Gaussian and robust Gaussian percentiles along with their 95% confidence limits.

3.1 ARSENIC

3.1.1 Ironstone Domain (As)

The raw data are positively skewed, after transformation to logarithms the distribution of the data seems consistent with the assumption of an underlying Gaussian random variable (Figure 4), no outliers were found. The different types of percentile show reasonable agreement (Figure 5) although the empirical uncertainties are always higher than the parametric values (Figure 5 and Table 1). The 95th percentile with confidence limits is shown in Figure 6.

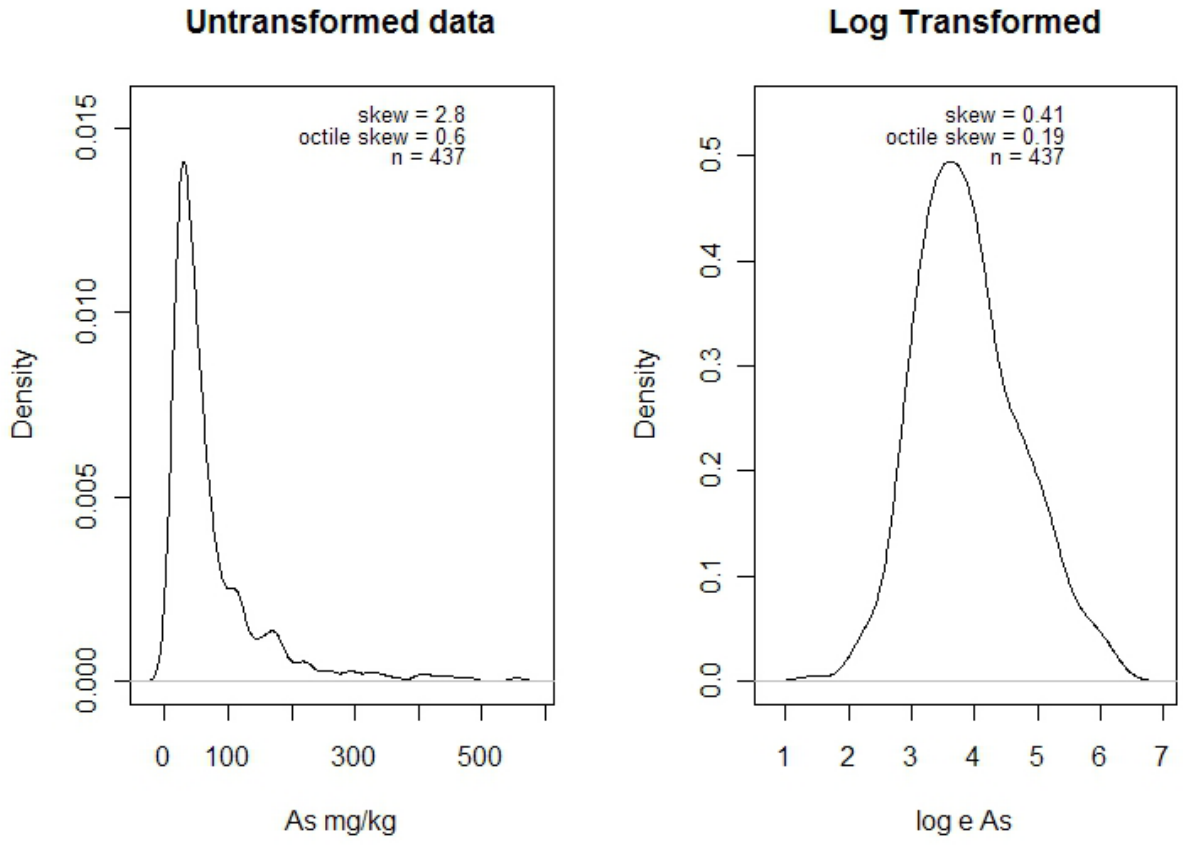

Figure 4: Density distributions for the raw data and the log transformed data for As in the Ironstone Domain (n = number of samples)

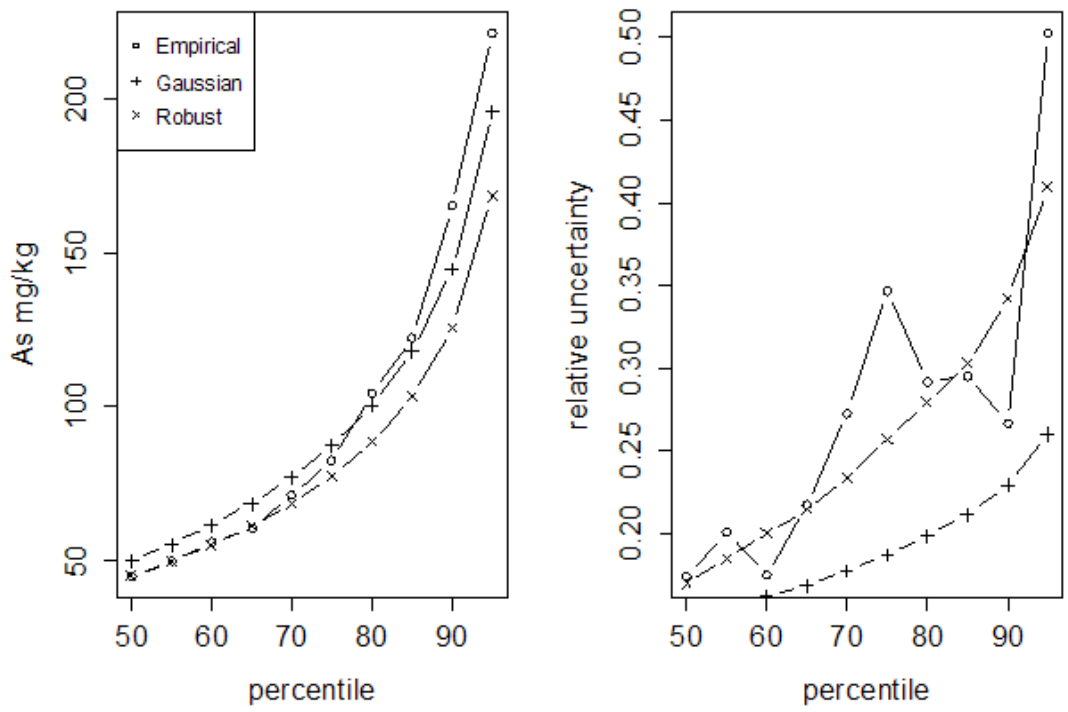

Figure 5: Comparison of empirical, Gaussian and Robust percentiles for As in the Ironstone Domain

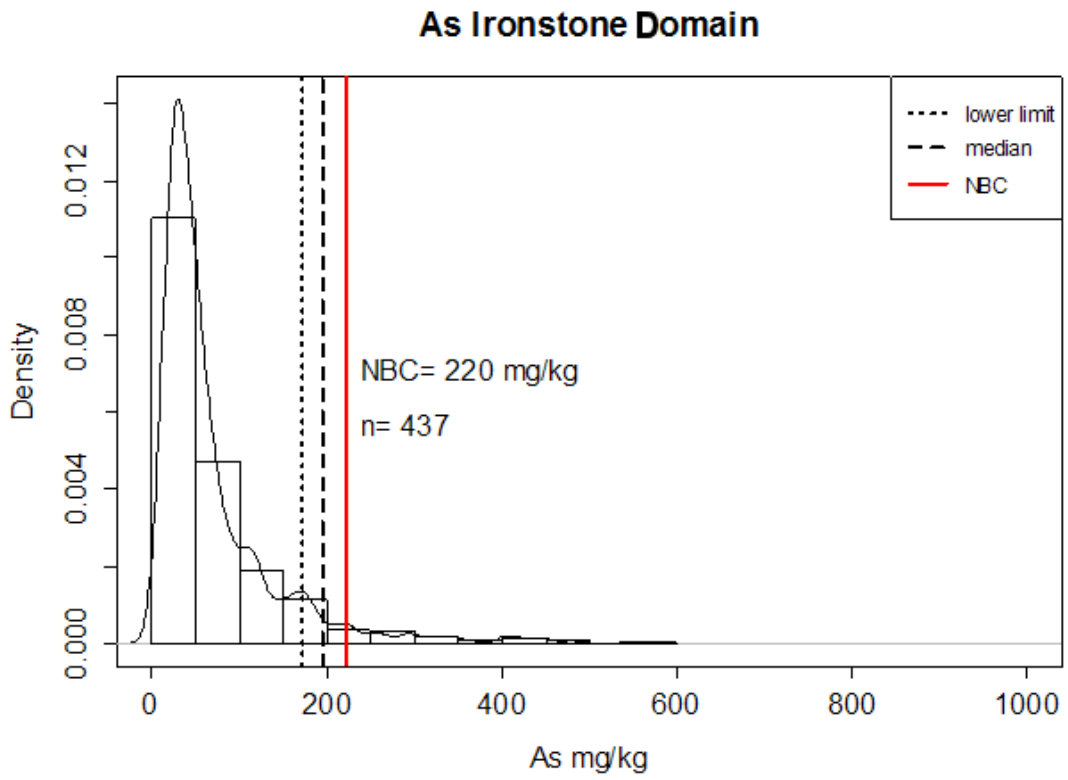

Figure 6: Summary density plot and histogram of the distribution of As in the Ironstone Domain showing the NBC and confidence interval (n = number of samples)

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	45.0	41.7	49.4	**50.0**	**46.2**	**53.7**	45.0	41.3	49.4
55	49.7	45.3	55.8	**55.5**	**51.1**	**59.7**	49.8	45.5	55.1
60	56.0	50.2	60.0	**61.7**	**56.7**	**66.6**	55.2	49.8	61.6
65	60.9	56.7	69.6	**68.8**	**63.2**	**74.5**	61.4	54.8	68.7
70	71.4	61.7	81.0	**77.3**	**70.6**	**84.1**	68.6	60.7	77.5
75	82.7	72.6	97.7	**87.5**	**79.6**	**95.9**	77.4	67.7	88.7
80	104.8	86.4	116.6	**100.6**	**91.1**	**111.0**	88.5	76.4	102.0
85	122.7	108.9	147.1	**118.3**	**106.4**	**131.7**	103.5	88.1	120.7
90	165.6	140.7	179.3	**145.1**	**129.1**	**163.5**	126.1	105.2	149.3
95	221.7	179.7	291.4	**196.2**	**171.6**	**223.9**	168.8	136.6	205.4

Table 1: Empirical (Emp), Parametric Gaussian (P) and Robust Gaussian (R) Percentile values for As in the Ironstone Domain (concentrations in mg/kg). L and H values represent confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

3.1.2 Mineralisation Domain (As)

For As in the Mineralisation Domain, initial $\log_e$ transform does not bring either the SC or OS within the specified limits for a Gaussian fit (Figure 7). However, there is evidence for an outlier in the data in the right tail of the distribution.

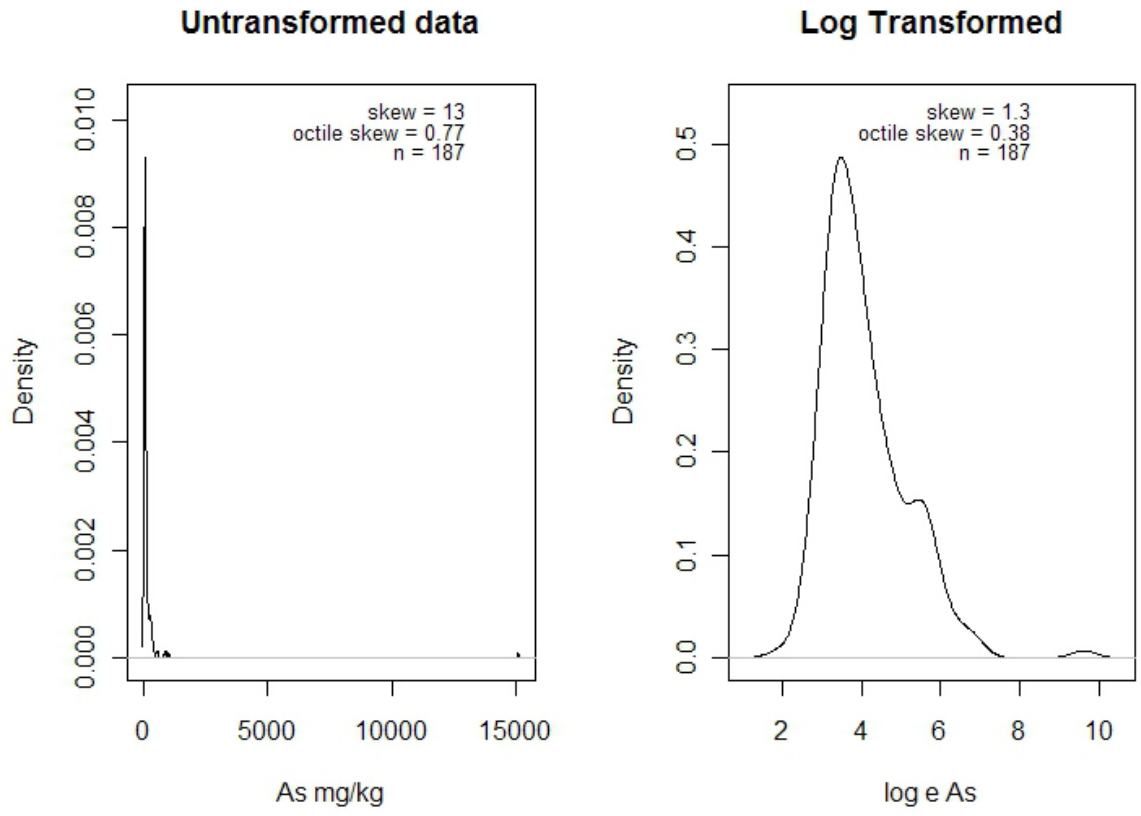

Figure 7: Density distributions for the raw data and the $\log_e$ transformed data for As in the Mineralisation Domain (n = number of samples)

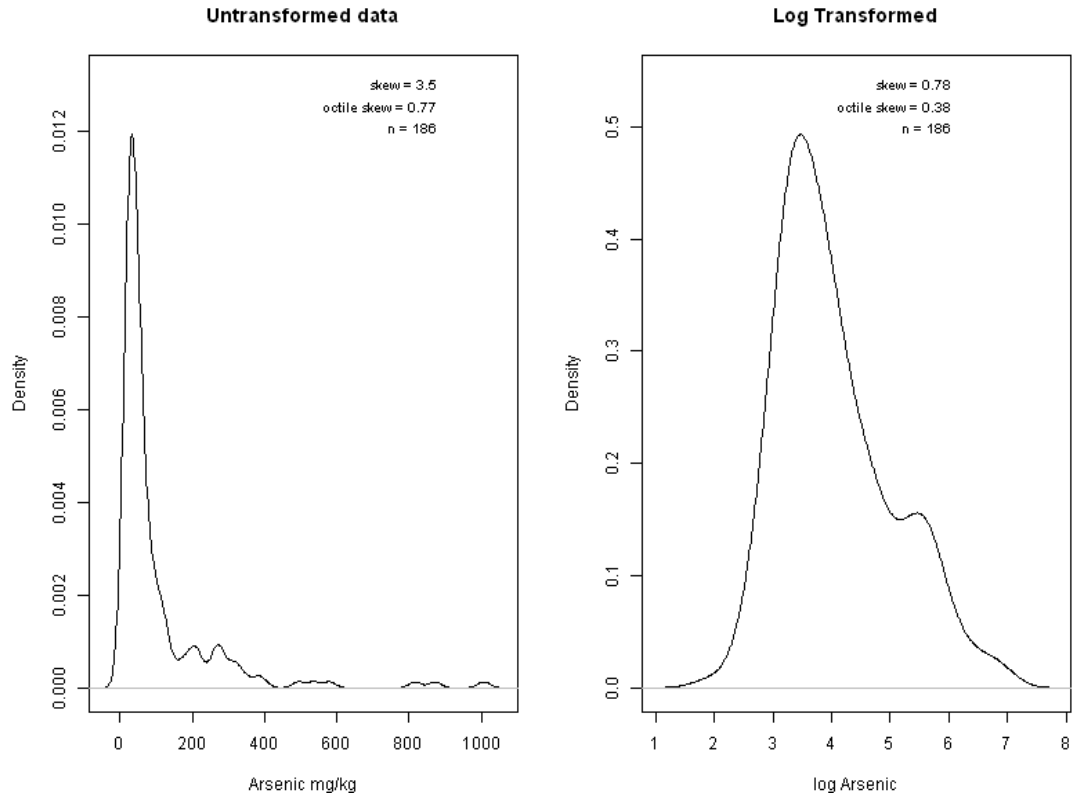

Figure 8: Density distributions for the raw data and the $\log_e$ transformed data for As in the Mineralisation Domain with outlier removal (n = number of samples)

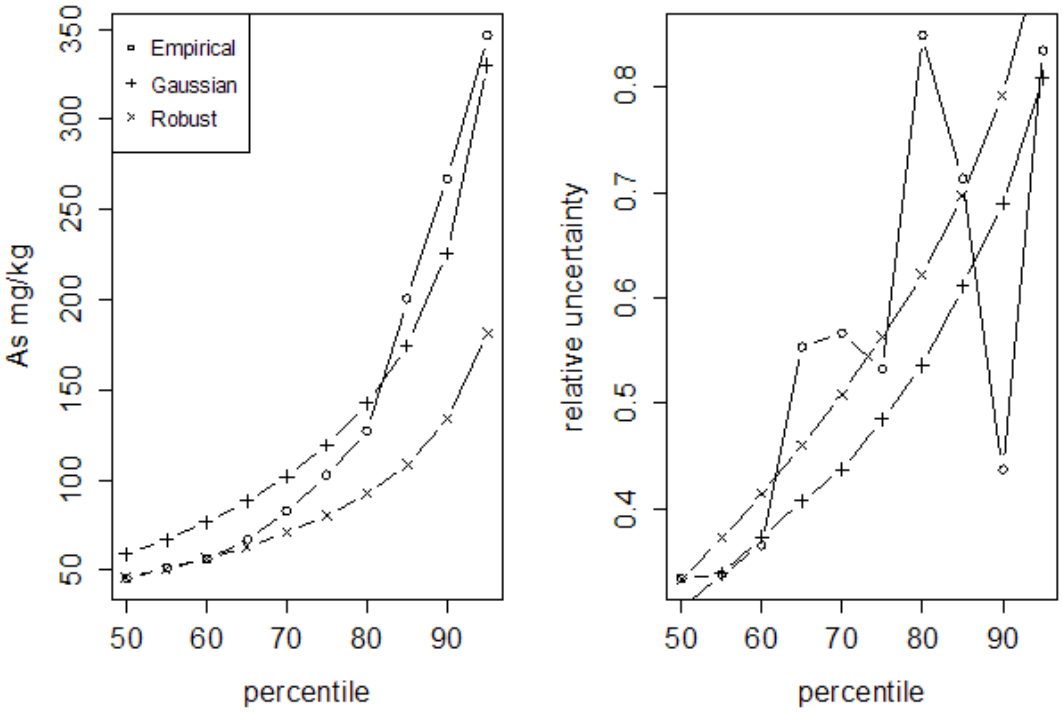

Figure 9: Comparison of empirical, Gaussian and Robust percentiles for As in the Mineralisation Domain

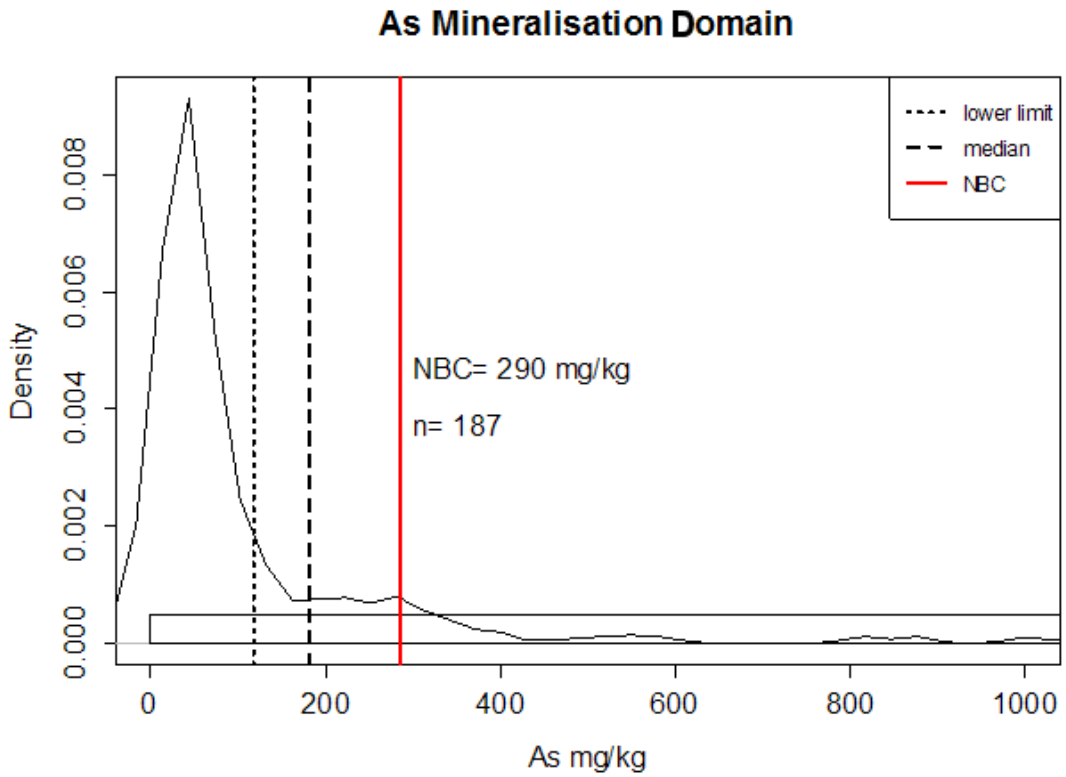

Figure 10: Summary density plot and histogram of the distribution of As in the Mineralisation Domain showing the NBC (n = number of samples)

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	45.6	38.9	54.2	58.9	50.6	68.8	**45.6**	**38.9**	**54.2**
55	51.6	44.1	61.0	67.2	56.5	79.5	**50.7**	**42.6**	**61.4**
60	56.1	50.3	72.9	76.8	63.8	92.2	**56.4**	**46.5**	**69.8**
65	66.5	54.4	89.3	88.2	71.9	107.7	**63.0**	**50.7**	**79.7**
70	82.9	61.2	107.4	102.0	81.8	126.7	**70.8**	**55.6**	**91.4**
75	102.5	78.5	133.0	119.4	93.8	150.1	**80.3**	**61.6**	**106.2**
80	127.1	97.1	204.7	142.3	109.3	183.5	**92.4**	**68.9**	**125.3**
85	200.3	123.9	266.0	174.5	130.7	232.1	**108.8**	**78.4**	**152.6**
90	267.4	202.5	319.7	225.6	162.6	312.2	**133.6**	**92.4**	**197.0**
95	346.6	277.1	536.2	330.1	224.6	484.6	**181.1**	**118.3**	**286.3**

Table 2: Empirical (Emp), Parametric Gaussian (P), and Robust Gaussian (R) Percentile values for As in the Mineralisation Domain (concentrations in mg/kg). L and H values represent confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

When the extreme value is removed the SC of the data is reduced to 0.78, although the OS is still higher than 0.2 (Figure 8). There is some evidence for two populations of data in this domain (Figure 7 and Figure 8) with the robust percentiles giving lower values than either the empirical or parametric values (Figure 9). Continuing on the basis of one data population and a $\log_e$ transformation using the robust percentiles calculated from the median and the MAD the NBC of 290 mg/kg looks reasonable relative to the overall data set (Figure 10 and Table 2).

3.1.3 Principal Domain (As)

The raw data is positively skewed but a $\log_e$ transformation makes the distribution consistent with the assumption of an underlying Gaussian random variable (Figure 11), with no outliers indicated.

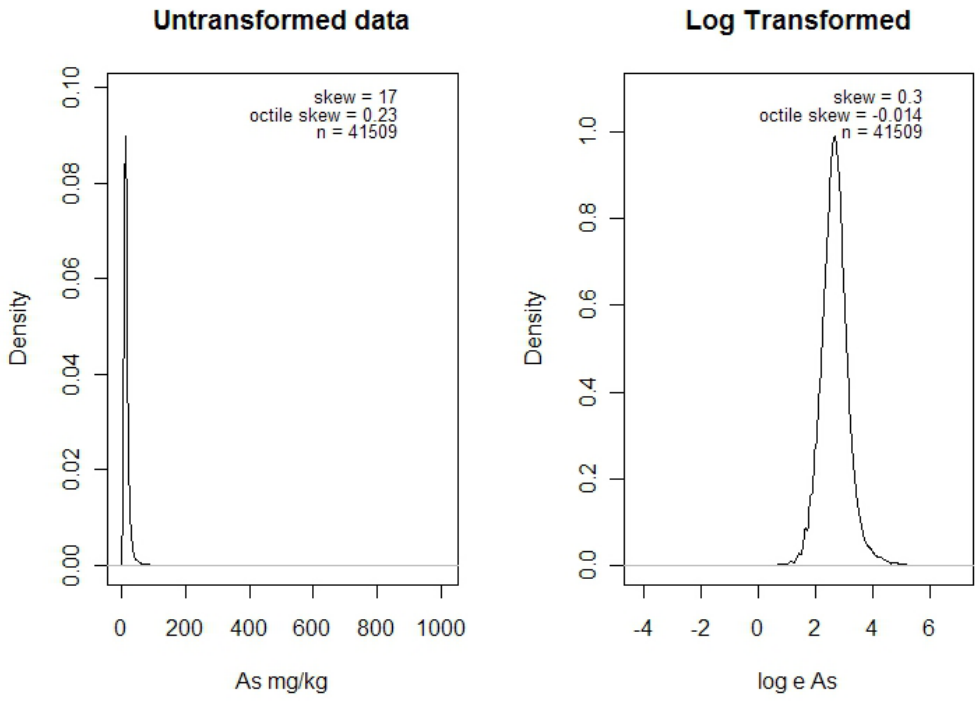

Figure 11: Density distributions for the raw data and the $\log_e$ transformed data for As in the Principal Domain (n = number of samples)

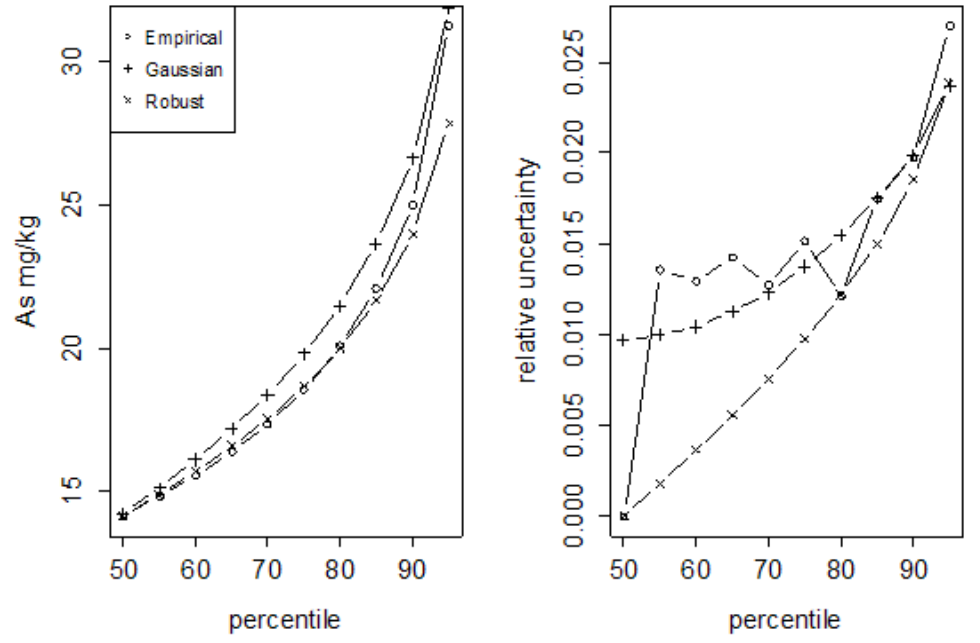

Figure 12: Comparison of empirical, Gaussian and Robust percentiles for As in the Principal Domain

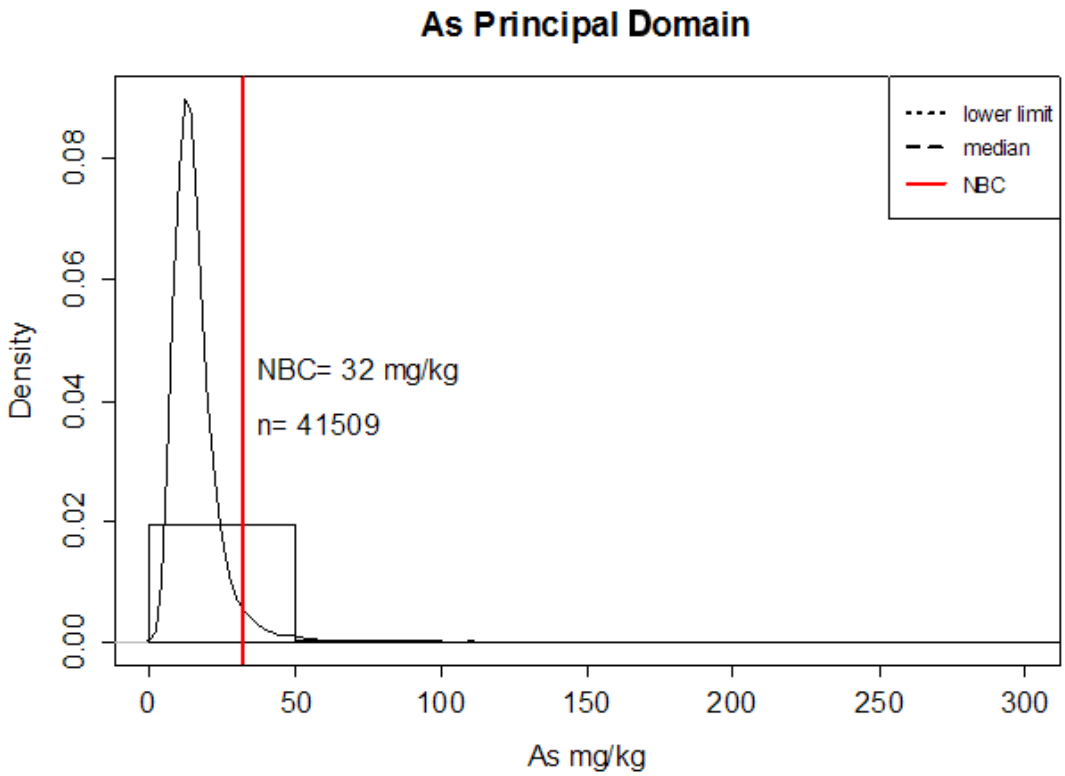

Figure 13: Summary density plot and histogram of the distribution for As in the Principal Domain showing the NBC (n = number of samples)

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	14.1	14.1	14.1	**14.2**	**14.1**	**14.3**	14.1	14.1	14.1
55	14.8	14.7	14.9	**15.1**	**15.0**	**15.2**	14.9	14.9	14.9
60	15.5	15.4	15.6	**16.1**	**16.0**	**16.2**	15.7	15.6	15.7
65	16.4	16.2	16.5	**17.2**	**17.1**	**17.3**	16.6	16.5	16.6
70	17.3	17.2	17.4	**18.4**	**18.3**	**18.5**	17.5	17.5	17.6
75	18.6	18.4	18.7	**19.8**	**19.7**	**19.9**	18.7	18.6	18.8
80	20.1	19.9	20.2	**21.5**	**21.3**	**21.6**	20.0	19.9	20.1
85	22.1	21.8	22.2	**23.6**	**23.4**	**23.8**	21.7	21.5	21.8
90	25.0	24.7	25.2	**26.6**	**26.4**	**26.9**	24.0	23.8	24.2
95	31.2	31.0	31.9	**31.8**	**31.5**	**32.2**	27.8	27.5	28.2

Table 3: Empirical (Emp), Parametric Gaussian (P), and Robust Gaussian (R) Percentile values for As in the Principal Domain (concentrations in mg/kg). L and H values represent confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

The three percentile estimation methods show close agreement and low relative uncertainty (Figure 12). The NBC for As in this domain has a very clearly defined as 32 mg/kg with very tight confidence interval of 0.7 mg/kg As (Figure 13 and Table 3).

3.2 LEAD

3.2.1 Mineralisation Domain (Pb)

The raw Pb data is positively skewed but a $\log_e$ transformation makes the distribution consistent with the assumption of an underlying Gaussian random variable (Figure 14), no outliers were indicated. The three percentile estimation methods show close agreement (Figure 15) with relative uncertainty on the 95 percentile ranging from 0.4 to 0.5. The NBC for Pb in this domain is 2400 mg/kg (Figure 16 and Table 4).

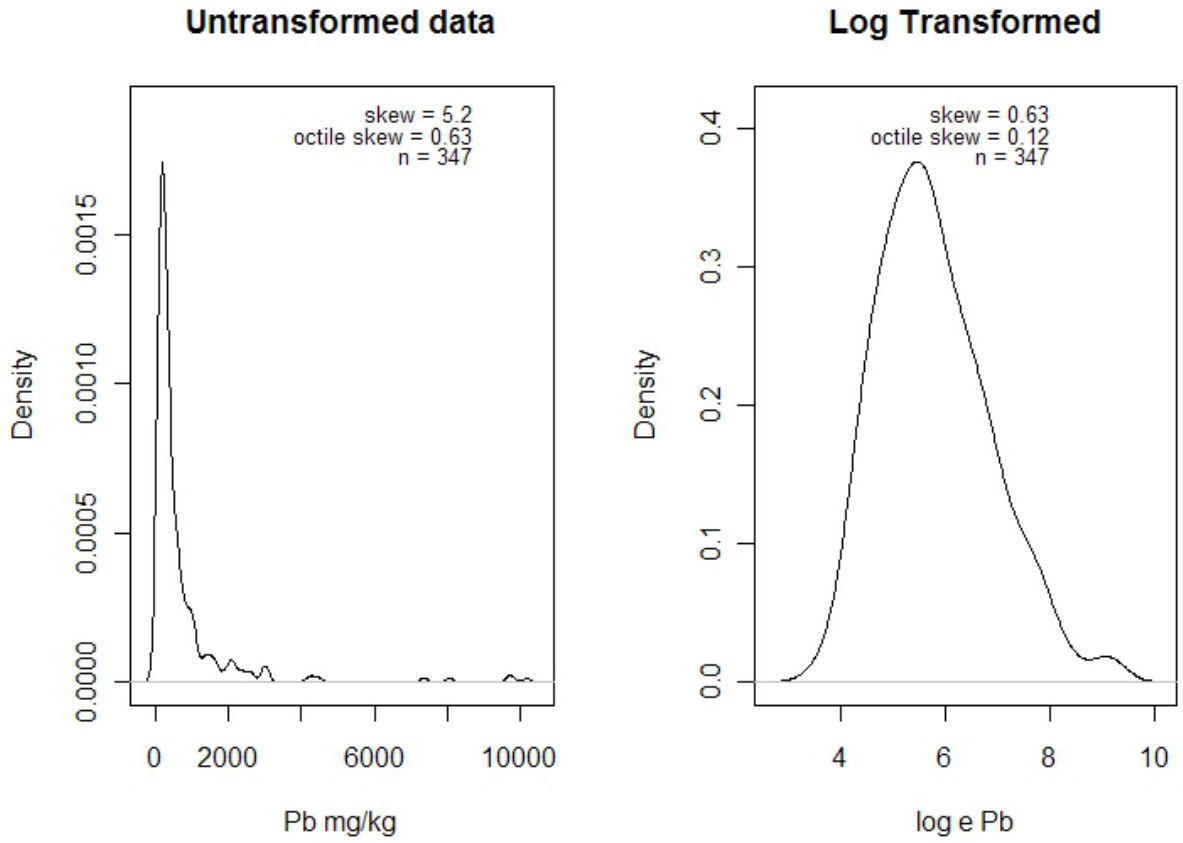

Figure 14: Density distributions for the raw data and the $\log_e$ transformed data for Pb in the Mineralisation Domain (n = number of samples)

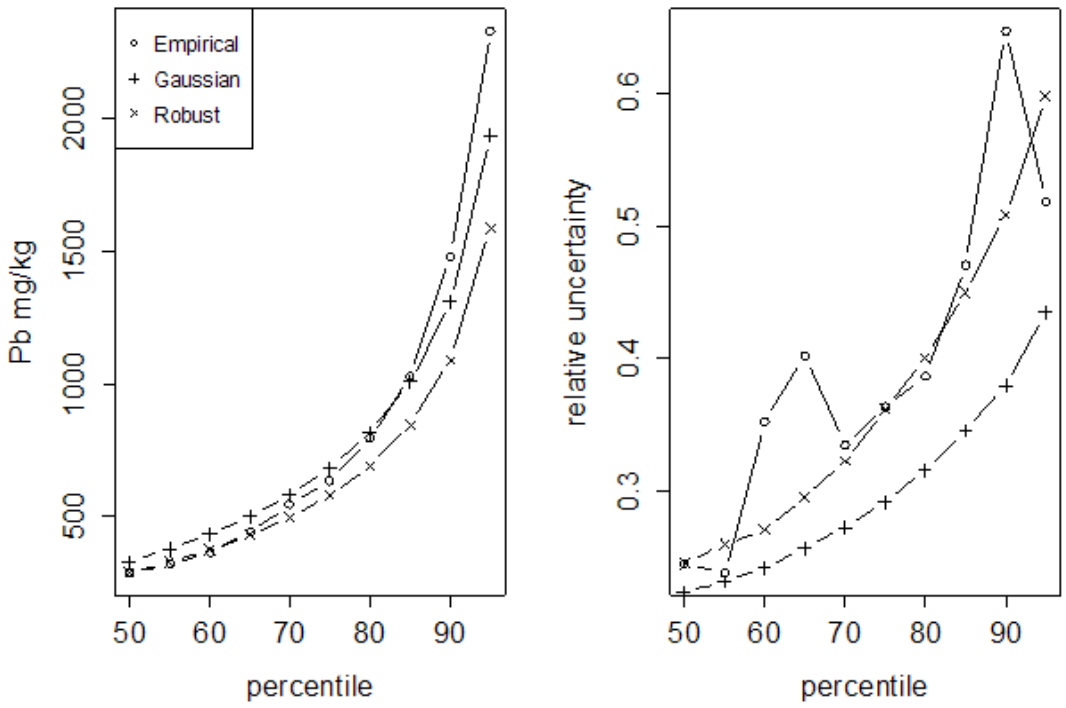

Figure 15: Comparison of empirical, Gaussian and Robust percentiles for Pb in the Mineralisation Domain

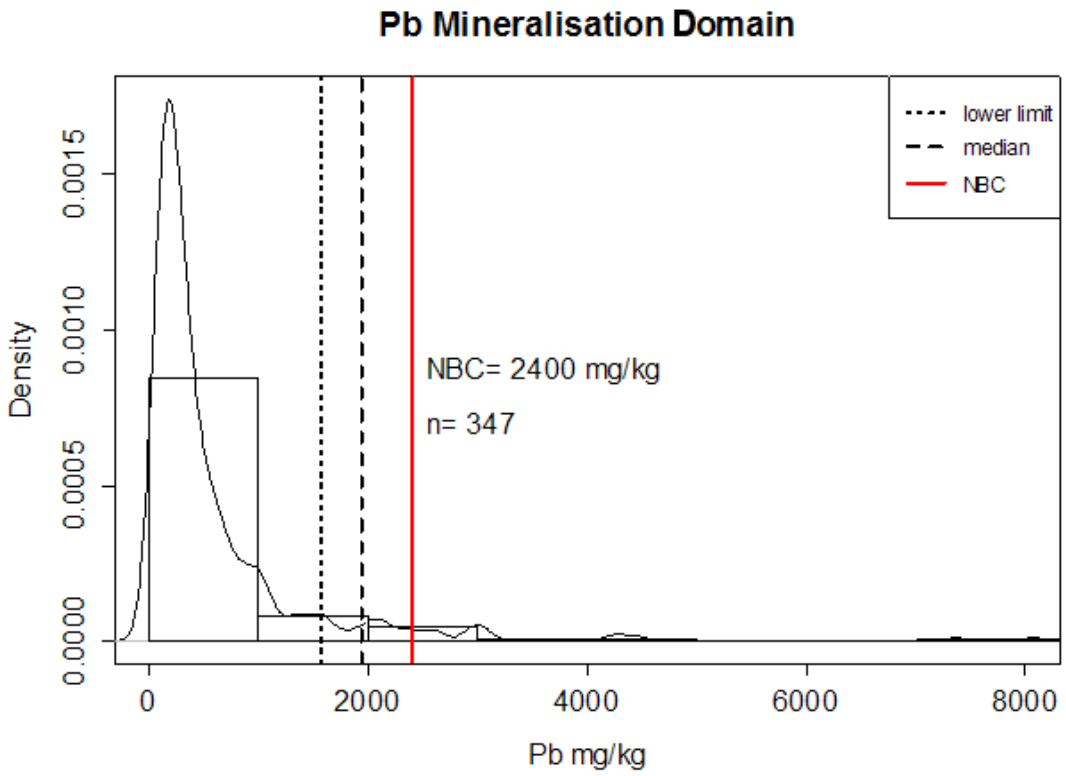

Figure 16: Summary density plot and histogram of the distribution of Pb in the Mineralisation Domain showing the NBC (n = number of samples)

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	290	252	323	**332**	**296**	**370**	290	252	322
55	323	290	366	**380**	**337**	**425**	330	286	370
60	367	322	444	**435**	**385**	**489**	377	325	425
65	444	366	543	**501**	**442**	**567**	432	370	491
70	547	445	631	**582**	**509**	**662**	499	427	575
75	635	553	789	**684**	**594**	**785**	582	493	682
80	798	654	966	**818**	**703**	**949**	692	574	825
85	1030	858	1305	**1008**	**855**	**1185**	847	692	1034
90	1482	1060	1995	**1312**	**1093**	**1567**	1091	875	1371
95	2332	1801	2991	**1937**	**1582**	**2377**	1589	1237	2099

Table 4: Empirical (Emp), Parametric Gaussian (P) and Robust Gaussian (R) Percentile values for Pb in the Mineralisation Domain (concentrations in mg/kg). L and H values represent confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

3.2.2 Urban Domain (Pb)

The raw Pb data is positively skewed but a $\log_e$ transformation makes the distribution consistent with the assumption of an underlying Gaussian random variable (Figure 17), with no outliers indicated.

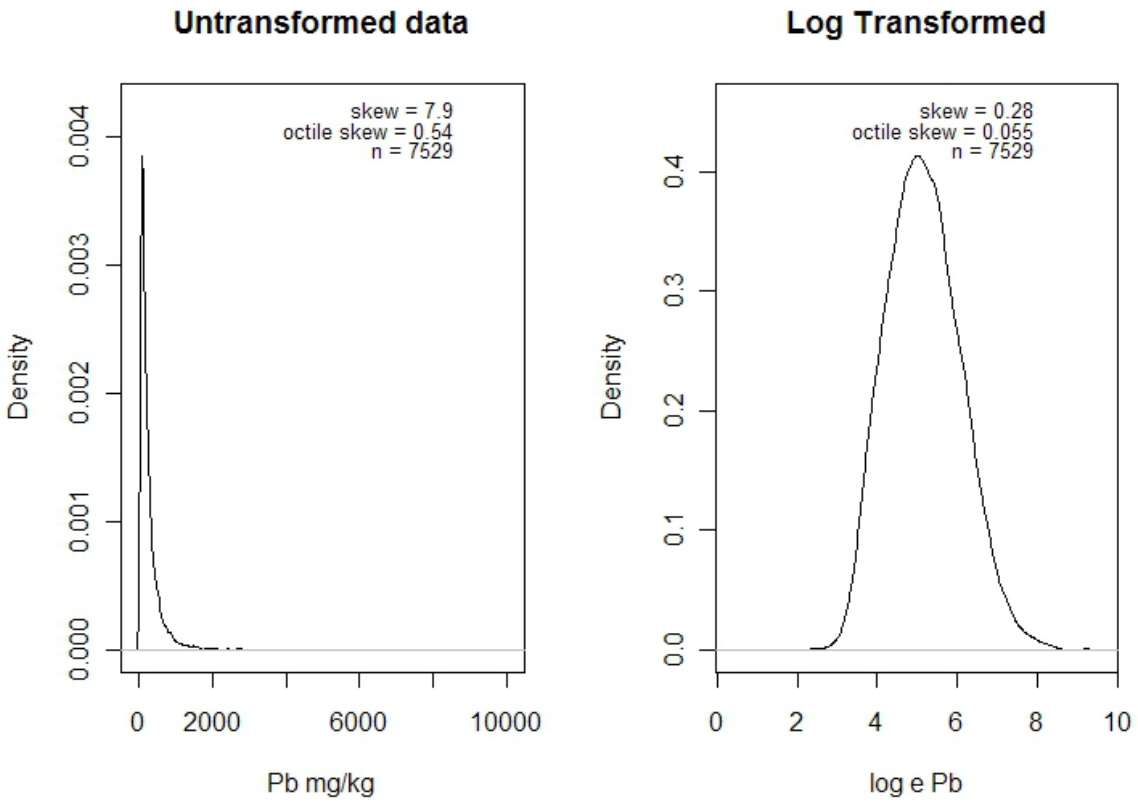

Figure 17: Density distributions for the raw data and the $\log_e$ transformed data for Pb in the Urban Domain (n = number of samples)

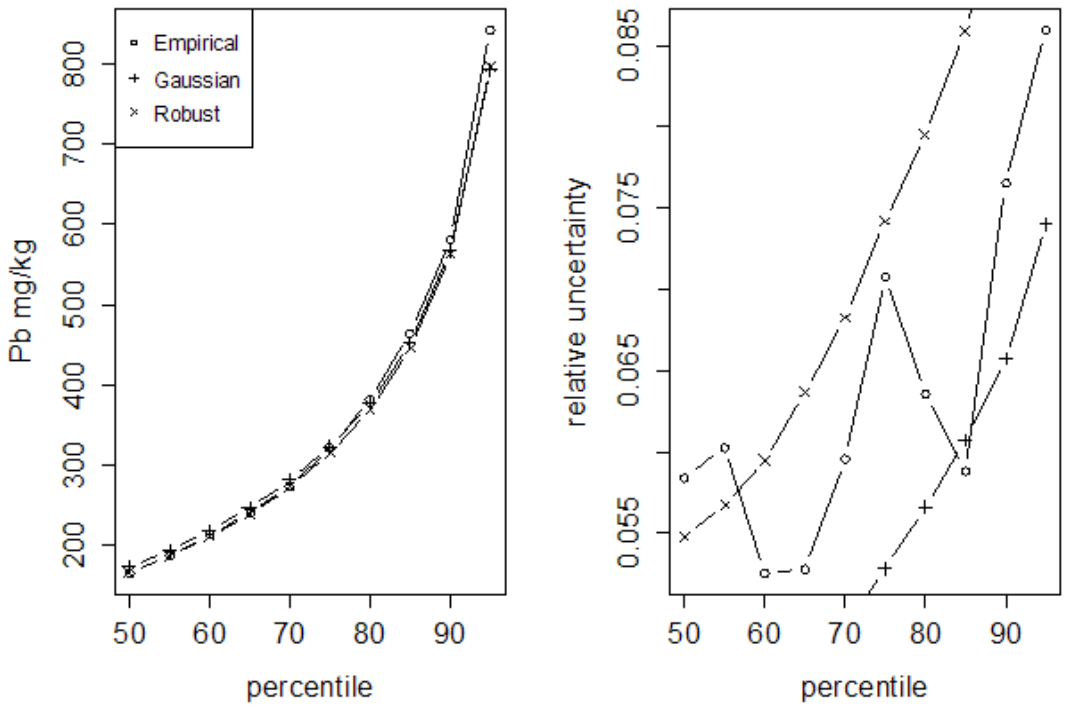

Figure 18: Comparison of empirical, Gaussian and Robust percentiles for Pb in the Urban Domain

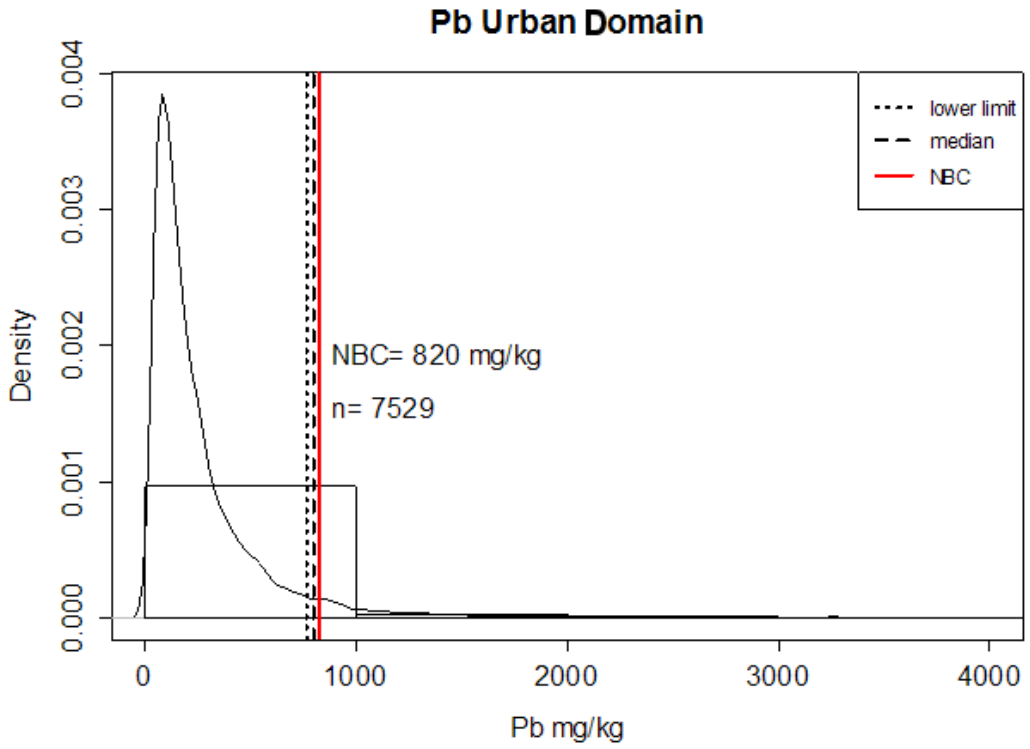

Figure 19: Summary density plot and histogram of the distribution of Pb in the Urban Domain showing the NBC (n = number of samples)

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	166	162	171	**174**	**170**	**177**	166	161	171
55	187	182	192	**195**	**191**	**199**	187	181	193
60	213	207	217	**219**	**214**	**224**	211	205	218
65	242	236	248	**248**	**242**	**253**	239	232	247
70	274	268	283	**282**	**275**	**288**	273	264	282
75	322	310	330	**324**	**316**	**331**	316	304	326
80	382	370	395	**378**	**368**	**387**	370	356	383
85	464	451	478	**452**	**439**	**464**	446	426	463
90	582	562	608	**567**	**549**	**584**	564	536	588
95	843	806	875	**794**	**765**	**821**	797	751	838

Table 5: Empirical (Emp), parametric Gaussian (P) and Robust Gaussian (R) Percentile values for Pb in the Urban Domain (concentrations in mg/kg). L and H values represent confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

The three percentile estimation methods show close agreement (Figure 18). The NBC for Pb in this domain is clearly defined as 820 mg/kg (Figure 19 and Table 5).

3.2.3 Principal Domain (Pb)

The raw Pb data is positively skewed and even after a $\log_e$ transformation the SC is >1 and the OS is >0.2 (Figure 20).

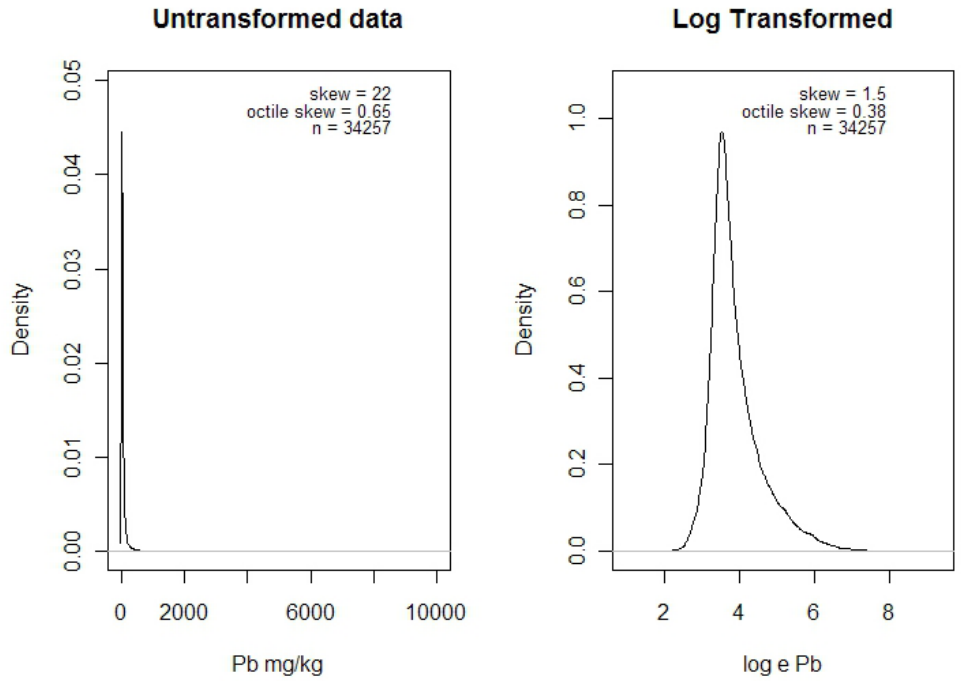

Figure 20: Density distributions for the raw data and the $\log_e$ transformed data for Pb in the Principal Domain (n = number of samples)

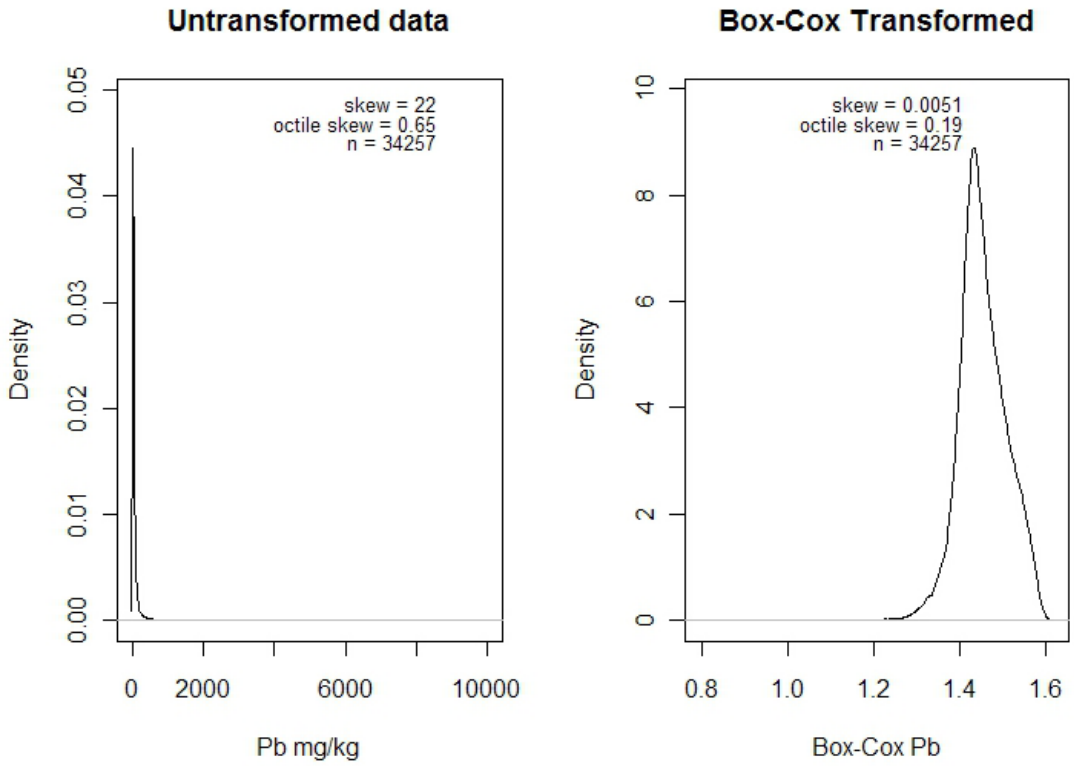

Figure 21: Density distributions for the raw data and the Box-Cox transformed data for Pb in the Principal Domain (n = number of samples)

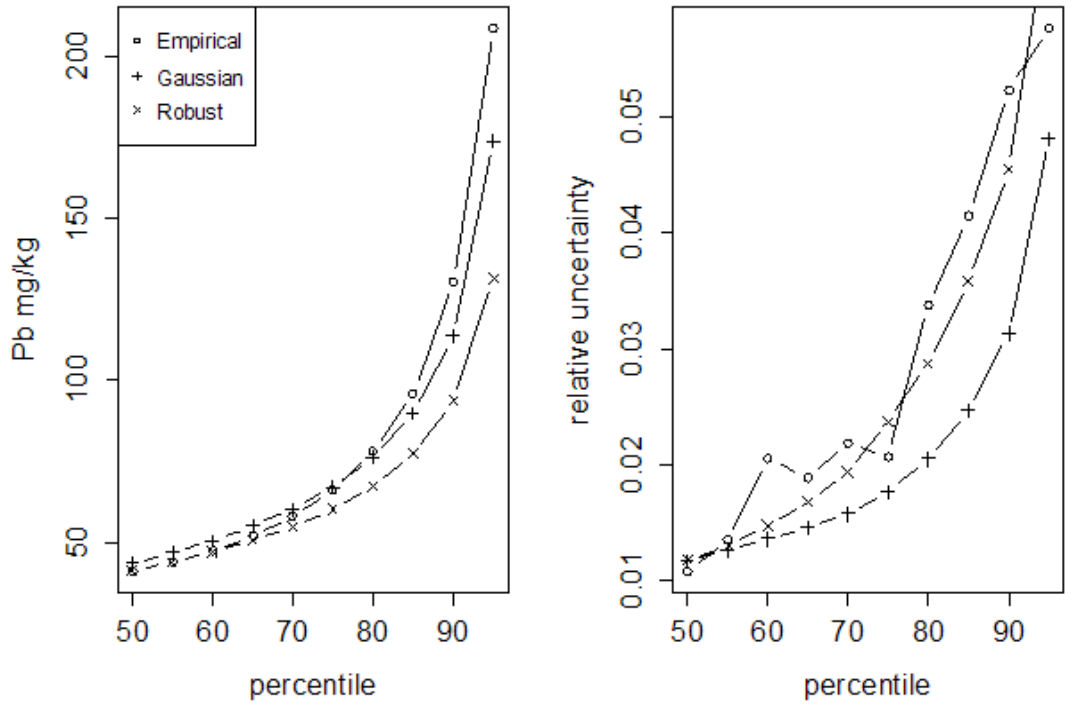

Figure 22: Comparison of empirical, Gaussian and Robust percentiles for Pb in the Principal Domain

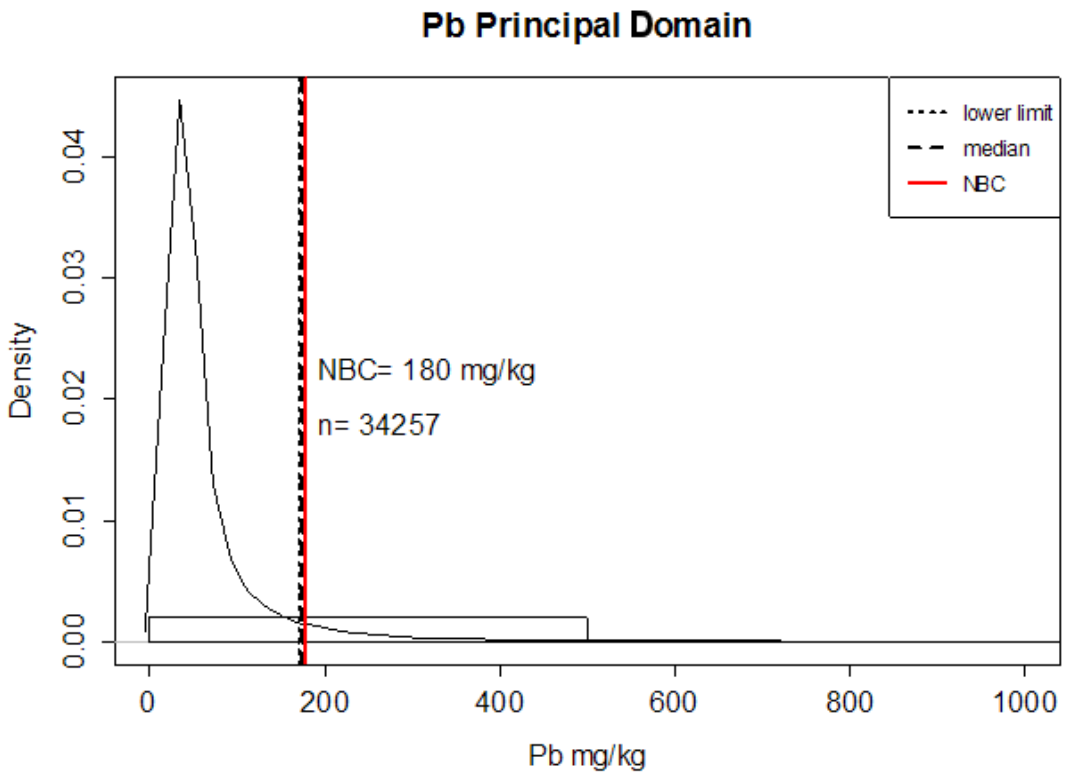

Figure 23: Summary density plot and histogram of the distribution for Pb in the Principal Domain showing an example NBC (n = number of samples)

In this case a Box-Cox transform has been applied to the data (Figure 21) which brings both the SC and OS below 1 and 0.2 respectively so the Box-Cox transformed data is consistent with the assumption of an underlying Gaussian random variable. No outliers have been identified. The three percentile estimation methods show close agreement (Figure 22). The NBC for Pb in this domain is clearly defined as 180 mg/kg (Figure 23 and Table 6).

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	41.1	40.7	41.2	**43.5**	**43.2**	**43.7**	41.1	40.7	41.2
55	43.9	43.4	44.0	**46.7**	**46.5**	**47.0**	43.8	43.3	43.9
60	47.3	46.9	47.9	**50.5**	**50.2**	**50.8**	46.9	46.3	47.1
65	51.9	51.6	52.6	**54.9**	**54.5**	**55.3**	50.5	49.8	50.7
70	57.9	57.3	58.6	**60.3**	**59.8**	**60.8**	54.8	54.0	55.1
75	66.0	65.4	66.8	**67.1**	**66.5**	**67.7**	60.2	59.1	60.5
80	77.8	76.7	79.0	**76.3**	**75.6**	**77.2**	67.2	65.8	67.7
85	96.0	94.3	98.0	**90.0**	**88.9**	**91.2**	77.2	75.2	78.0
90	130.3	127.6	133.6	**113.6**	**111.9**	**115.6**	93.7	90.8	95.1
95	208.6	202.7	214.8	**173.6**	**169.6**	**178.2**	131.5	126.0	134.5

Table 6: Empirical (Emp), Parametric Gaussian (P) and Robust Gaussian (R) Percentile values for Pb in the Principal Domain (concentrations in mg/kg). L and H values represent confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

3.3 BENZO[*a*]PYRENE

For BaP an Urban (n=11) and Principal (n=165) Domains have been identified in WP2. The number of data points in these two domains for England is very much less than those available for As and Pb. For the Urban Domain there are only 11 data points which, although reasonably well distributed spatially, is insufficient to produce an NBC (see Section 2.1.4). In order to increase the data support, particularly for the Urban Domain, NBCs for the Urban (n=32) and Principal (n=371) have been calculated for Britain which includes data points in Wales and Scotland. For sampling points in Wales and Scotland, however, the GLUD land use data is not available so the classification of land use into Urban and Principal has been made using the site descriptions provided by the soil surveys for this data reported in WP2.

3.3.1 Urban Domain (BaP)

The raw data is positively skewed but a $\log_e$ transformation makes the distribution consistent with the assumption of an underlying Gaussian random variable (Figure 24). There is some evidence that there may be more than one population in this data set although this may be a function of the low number of samples being considered. The three percentile estimation methods show reasonable agreement but as the number of data points is low (n=32) the uncertainties on the percentiles are high (Figure 25). The NBC for BaP in this domain is 3.6 mg/kg with a wide confidence interval spanning 2.4 mg/kg BaP (Figure 26 and Table 7).

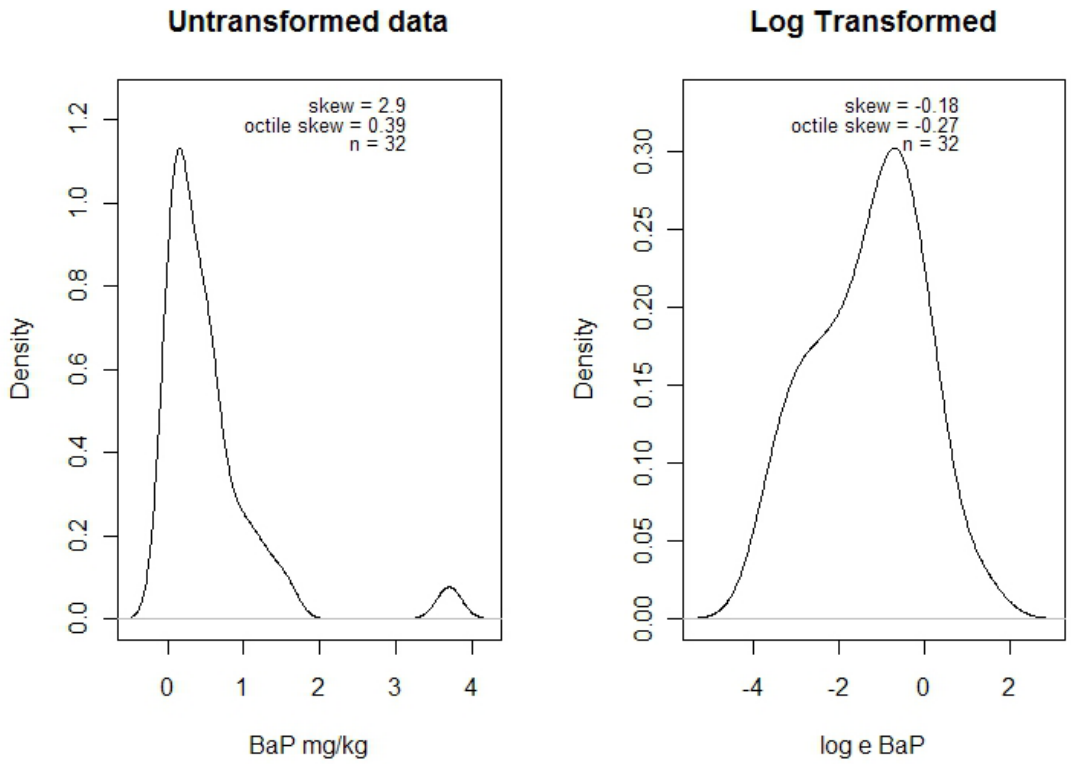

Figure 24: Density distributions for the raw data and the $\log_e$ transformed data for BaP in the Urban Domain (n = number of samples)

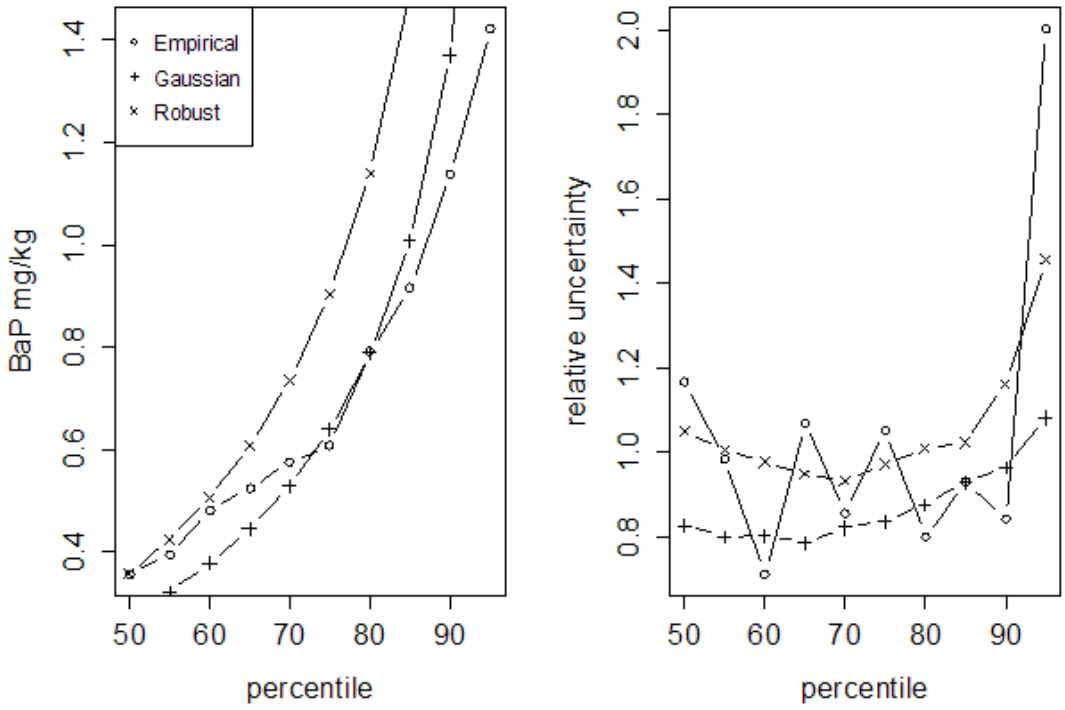

Figure 25: Comparison of empirical, Gaussian and Robust percentiles for BaP in the Urban Domain

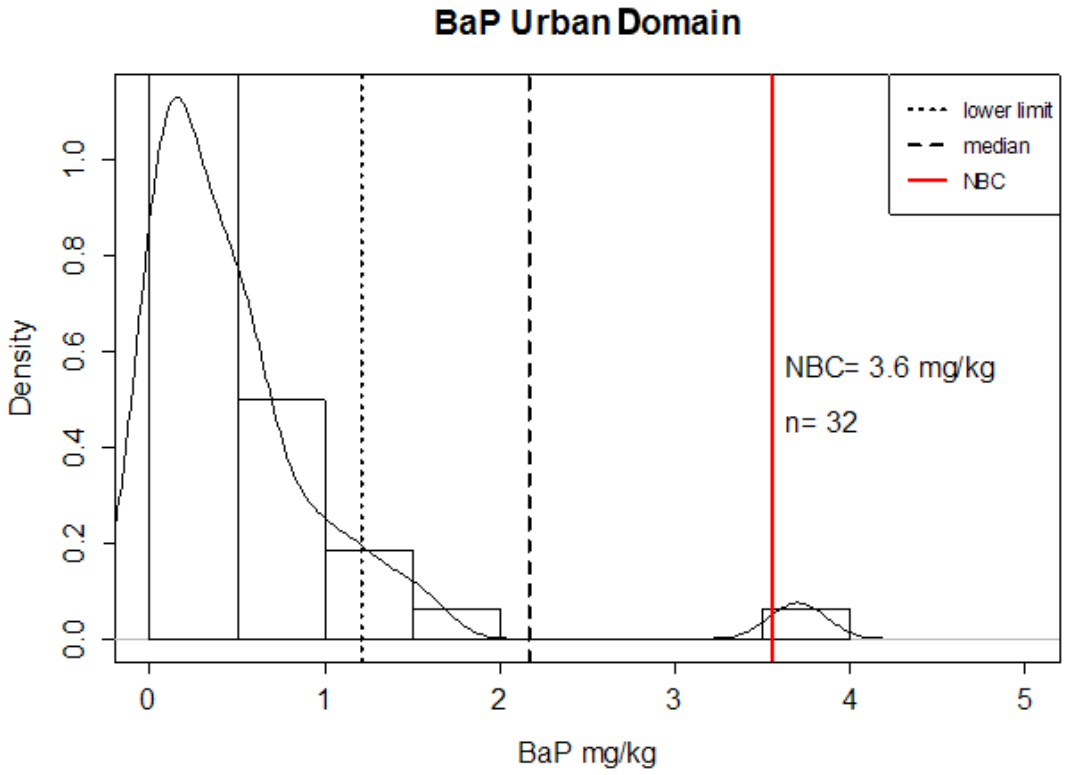

Figure 26: Summary density plot and histogram of the distribution of BaP in the Urban Domain showing the NBC (n = number of samples)

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	0.36	0.17	0.53	**0.27**	**0.18**	**0.43**	0.36	0.13	0.55
55	0.39	0.21	0.59	**0.32**	**0.21**	**0.50**	0.42	0.16	0.63
60	0.48	0.26	0.60	**0.38**	**0.25**	**0.58**	0.51	0.21	0.73
65	0.52	0.35	0.84	**0.44**	**0.29**	**0.68**	0.61	0.26	0.87
70	0.57	0.38	0.86	**0.53**	**0.34**	**0.81**	0.73	0.31	1.05
75	0.61	0.44	1.06	**0.64**	**0.41**	**0.99**	0.90	0.38	1.32
80	0.79	0.52	1.28	**0.79**	**0.50**	**1.23**	1.14	0.47	1.72
85	0.92	0.56	1.40	**1.01**	**0.62**	**1.58**	1.49	0.61	2.38
90	1.14	0.61	1.56	**1.37**	**0.81**	**2.19**	2.09	0.82	3.65
95	1.42	0.85	3.70	**2.17**	**1.21**	**3.59**	3.46	1.16	6.87

Table 7: Empirical (Emp), Parametric Gaussian (P) and Robust Gaussian (R) Percentile values for BaP in the Urban Domain (concentrations in mg/kg). L and H values represent confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

3.3.2 Principal Domain (BaP)

The raw data is positively skewed but a $\log_e$ transformation makes the distribution consistent with the assumption of an underlying Gaussian random variable (Figure 27), no outliers were indicated but there is some evidence for more than one data population.

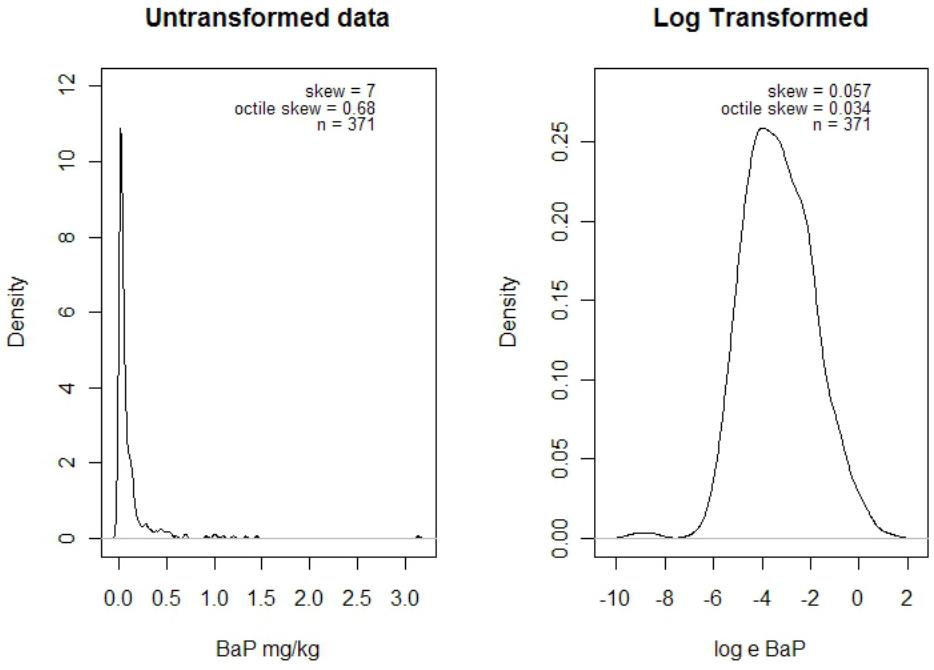

Figure 27: Density distributions for the raw data and the $\log_e$ transformed data for BaP in the Principal Domain (n = number of samples)

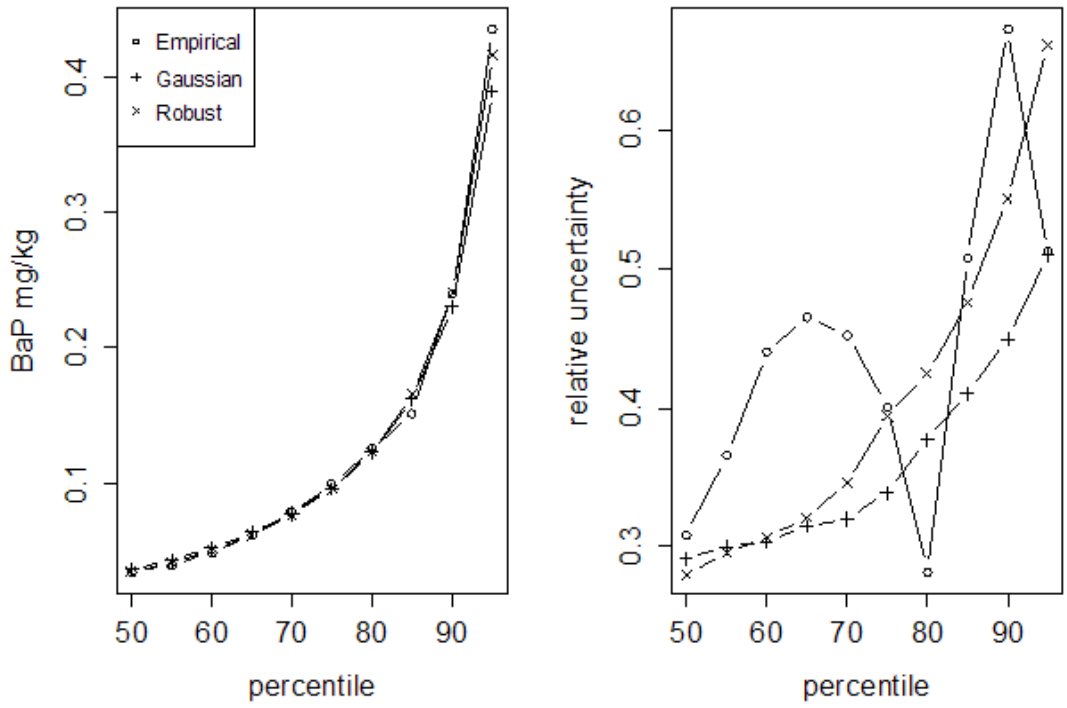

Figure 28: Comparison of empirical, Gaussian and Robust percentiles for BaP in the Principal Domain

The three percentile estimates show good agreement (Figure 28) with the empirical uncertainties showing rather erratic behaviour compared to the parametric values (Figure 28). The calculated NBC is 0.5 mg/kg (Figure 29 and Table 8).

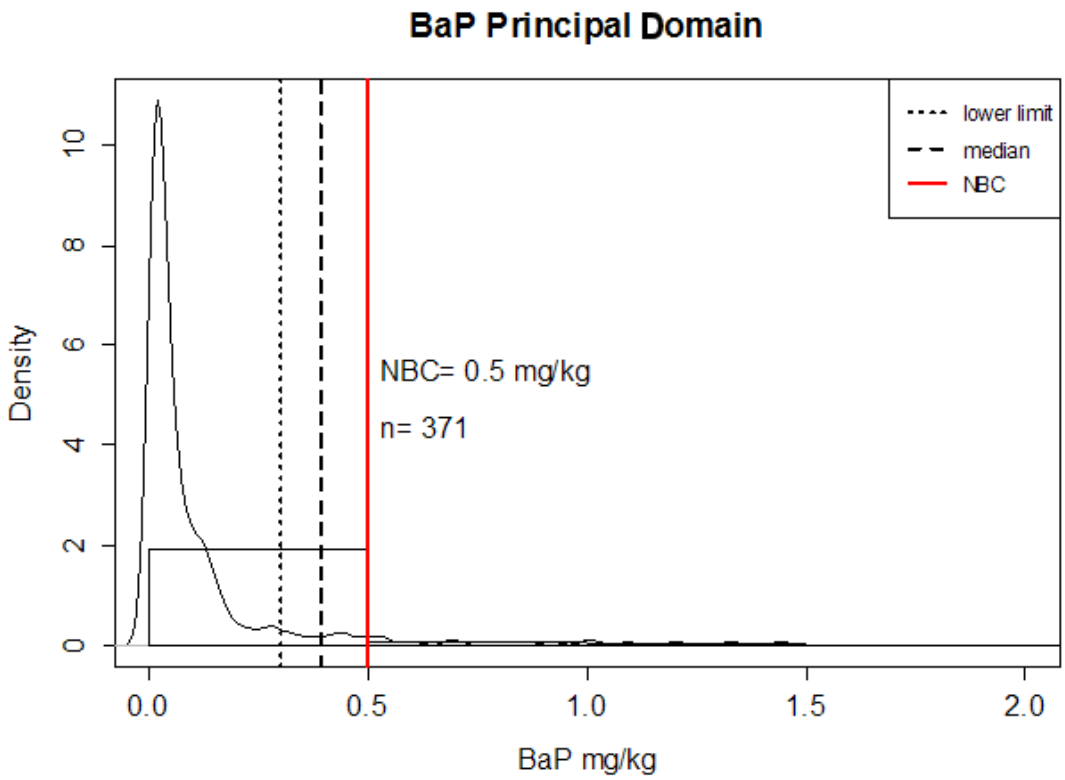

Figure 29: Summary density plot and histogram of the distribution of BaP in the Principal Domain showing the NBC (n = number of samples)

Percentile	Empirical	Emp L	Emp H	Parametric (P)	P L	P H	Robust (R)	R L	R H
50	0.035	0.029	0.040	**0.037**	**0.031**	**0.042**	0.035	0.029	0.040
55	0.040	0.035	0.050	**0.044**	**0.038**	**0.051**	0.042	0.035	0.049
60	0.049	0.040	0.063	**0.053**	**0.045**	**0.061**	0.051	0.042	0.059
65	0.062	0.050	0.078	**0.064**	**0.054**	**0.075**	0.062	0.050	0.072
70	0.079	0.064	0.099	**0.078**	**0.067**	**0.092**	0.076	0.061	0.090
75	0.099	0.081	0.120	**0.096**	**0.082**	**0.114**	0.096	0.075	0.114
80	0.126	0.109	0.145	**0.123**	**0.103**	**0.148**	0.124	0.095	0.150
85	0.152	0.128	0.205	**0.162**	**0.134**	**0.198**	0.166	0.124	0.206
90	0.240	0.170	0.332	**0.231**	**0.188**	**0.286**	0.241	0.173	0.306
95	0.436	0.313	0.539	**0.390**	**0.306**	**0.504**	0.417	0.283	0.562

Table 8: Empirical (Emp), Parametric Gaussian (P), Robust Gaussian (R) Percentile values for BaP in the Britain Principal Domain (concentrations in mg/kg). L and H values represent 95% confidence intervals around the median. Shaded/bold values indicate data used to calculate NBC

4 Concluding remarks on calculated normal background concentrations

Table 9 shows the NBC for a given contaminant in a given domain (with its associated area expressed in km^2 and the relative proportion of England it covers) based on the procedures discussed in Sections 2.1.4 and 2.1.5, and the results of the statistical analysis given in Section 3. This value is the highest contaminant concentration in soil likely to be derived from normal background concentrations; values exceeding this are assumed to be from point source pollution. Given the sampling and analytical uncertainties on the measurements (for example, see Johnson (2011) for discussion on precision and accuracy of British Geological Survey G-BASE soil sample analyses), it is only appropriate to give the NBCs to two significant figures (Table 9).

The NBCs are calculated for the whole of England, there may be instances where they are not appropriate on a local scale, for example because the national values have been calculated on a relatively small data set (*e.g*. BaP) or the national scale data does not take into account a localised normal background population. In these instances it may be more appropriate to set localised NBC based on a localised sampling scheme and application of the procedures outlined in this report.

The earlier Project report (Ander *et al*. 2011) described how domains were identified for each contaminant. At a more local scale it was described how contaminant variability would occur within a defined domain and especially the Principal Domain. This variability can be seen from maps provided in other Work Packages of this Project (*e.g.* the national interpolated images in the supplementary information of the Technical Guidance Sheets). When using NBCs at a local more site specific scale, a first stage in any investigation should be to ask how appropriate is the domain NBC to the local setting. This can be further informed by looking at information other than that for soils such as the high density BGS stream sediment mapping (Johnson *et al*., 2005). This may define the surface chemical environment at a sufficiently large scale in areas where soil data may be much sparser. Using the methodology described in this report, NCBs can be defined for local areas providing there is enough systematically collected data available for the statistical analysis.

Contaminant	Domain	Area (km^2)	Area (%)	95th Percentile			n
				Lower limit	Median	Upper limit	
As	Ironstone	1352	1%	170	200	**220**	437
	Mineralisation	2250	2%	120	180	**290**	187
	Principal	129350	97%	32	32	**32**	41509
Pb	Urban	5432	4%	770	790	**820**	7529
	Mineralisation	2871	2%	1600	1900	**2400**	347
	Principal	124648	94%	170	170	**180**	34257
BaP	Urban	-	-	1.2	2.2	**3.6**	32
	Principal	-	-	0.31	0.39	**0.5**	371

Area represents the domain area for England, the % area being the area of England covered by that domain. n is the number of data points used.

Table 9: NBC values (in red/bold) and other information for As, Pb and BaP (all results in mg/kg)

References

Ander, E.L., Cave, M.R., Johnson, C.C. and Palumbo-Roe, B. 2011. Normal background concentrations of contaminants in the soils of England. Available data and data exploration. *British Geological Survey Commissioned Report*, CR/11/145. 124pp.

APAT-ISS. 2006. Protocollo Operativo per la determinazione dei valori di fondo di metalli/metalloidi nei suoli dei siti d'interesse nazionale. Revisone 0. Agenzia per la Protezione dell'Ambiente e per i Servizi Tecnici and Istituto Superiore di Sanita.

Appleton, J. D. 1995. Potentially harmful elements from natural sources and mining areas: characteristics, extent and relevance to planning and development in Great Britain. British Geological Survey. Technical Report WP/95/3. Commission Report for the Department of Environment (DoE), UK.

Appleton, J.D., Rawlins, B.G. and Thornton, I., 2008. National-scale estimation of potentially harmful element ambient background concentrations in topsoil using parent material classified soil:stream–sediment relationships. *Applied Geochemistry*, 2008(9), 2596-2611.

BGS, 2011. Proposal Form (EVID2) for "Establishing data on normal/background levels of soil contamination in England". Defra Research Project Proposal call CTE1118, British Geological Survey, Keyworth, Nottingham, UK. 5th September 2011. *Commercial in Confidence*.

Brys, G., Hubert, M. and Struyf, A. 2003. A comparison of some new measures of skewness. In: *Developments in Robust Statistics*. Dutter, R., Filzmoser, P., Gather, U., and Rousseeuw, P.J. (editors). (Heidelberg: Physica-Verlag.)

Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap Methods and their Applications*. Cambridge University Press.

Defra, 2006. Environmental Protection Act 1990: Part IIA Contaminated Land. Defra Circular 01/2006. Department for Environment, Food and Rural Affairs (Defra), UK. September 2006. 200pp.

Defra, 2011a. Simplification of contaminated land statutory guidance. Impact Assessment. IA No: Defra1133, Department for Environment, Food and rural Affairs (Defra)/Welsh Assembly Government, UK. 6th October 2011.

Defra, 2012. Environmental Protection Act 1990: Part 2A Contaminated Land Statutory Guidance. April 2012. Department for Environment, Food and Rural Affairs (Defra). HM Government. Available on-line at: http://www.defra.gov.uk/environment/quality/land/ last accessed 1st May 2012.

Defra-EA, 2002. Potential contaminants for the assessment of land, Department for Environment, Food and Rural affairs and Environment Agency (UK).

DETR, 2000. Environmental Protection Act 1990: Part IIA Contaminated Land. DETR Circular 02/2000, Department of the Environment, Transport & the Regions (DETR), UK. 20th March 2000. 162pp. London: HMSO.

Efron, B. 1987. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397), 171-185.

Finnish Government Decree, 2007. Government Decree on the Assessment of Soil Contamination and Remediation needs (214/2007). In Finnish – unofficial translation available.

Grunsky, E.C. 2010. The interpretation of geochemical survey data. *Geochemistry: Exploration Environment Analysis*, 10, 27-74.

Hamon, R.E., McLaughlin, M.J., Gilkes, R.J., Rate, A.W., Zarcinas, B., Robertson, A., Cozens, G., Radford, N. and Bettenay, L., 2004. Geochemical indices allow estimation of heavy metal background concentrations in soils. *Global Biogeochemical Cycles*, 18, 1-6.

Hawkes, H.E. and Bloom, H. 1955. Heavy metals in stream sediment used as exploration guides. *Mining Engineer*, 8, 1121-1126.

ISO 19258:2011. 2011. Soil quality: Guidance on the determination of background values. ISO 19258:2005. International Organisation for Standardisation. Corrigendum 31st August 2011.

Jarva, J., Tarvainan, T., Reinikainen, J. and Eklund, M. 2010. TAPIR – Finnish national geochemical baseline database. *Science of The Total Environment*, 480, 4385-4395.

Johnson, C.C. 2011. Understanding the Quality of Chemical Data from the Urban Environment – Part 1: Quality Control Procedures. Chapter 5, 61-76. In: *Mapping the Chemical Environment of Urban Areas*. Johnson C.C., Demetriades, A., Locutura, J. and Ottesen, R.T. (editors). John Wiley & Sons, Ltd., Chichester, UK.

Johnson, C.C**.** and Ander, E.L., 2008. Urban Geochemical Mapping Studies: How and why we do them. *Environmental Geochemistry and Health*, 30, 511-530.

Johnson, C.C., Breward, N., Ander, E.L. and Ault, L. 2005. G-BASE: Baseline geochemical mapping of Great Britain and Northern Ireland. Geochemistry: Exploration, Environment, Analysis, 5(4), 347-357.

Lark, R.M. 2002. Modelling complex soil properties as contaminated regionalized variables. *Geoderma*. **106**, 171–188.

Lovering, T.S., Huff, L.C. and Almond, H. 1950. Dispersion of copper from the San Manuel copper deposit, Pinal County, Arizona. *Economic Geology*, 45, 493-514.

Matschullat, J., Ottenstein, R. and Reimann, C. 2000. Geochemical background - can we calculate it? *Environmental Geology*, 39, 990 - 1000.

Oliver, M.A., Loveland, P.J., Frogbrook, Z.L., Webster, R. and McGrath, S.P., 2002. Statistical and Geostatistical Analysis of the National Soil Inventory of England and Wales. Project SP0124, Defra (formerly MAFF) Soil Programme Technical Report.

Paterson, E., Towers, W., Bacon, J.R. and Jones, M. 2003. Background levels of contaminants in Scottish soils. The Macaulay Institute. *Final Contract Report to SEPA, 60pp.*

R Development Core Team, 2011. *R: A language and environment for statistical computing*., R Foundation for Statistical Computing, Vienna, Austria.

Rawlins, B.G., Webster, R. and Lister, T.R. 2003. The influence of parent material on top soil geochemistry in eastern England. *Earth Surface Processes and Landforms*, 28, 1389-1409.

Rawlins, B.G., Lark, R.M., O'Donnell, K.E., Tye, A.M. and Lister, T.R. 2005. The assessment of point and diffuse metal pollution of soils from an urban geochemical survey of Sheffield, England. *Soil Use and Management*, 21(4), 353-362.

Reimann, C. and Filzmoser, P. 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology*, 39(9), 1001-1014.

Reimann, C. and Garrett, R.G. 2005. Geochemical background – concept and reality. *Science of the Total Environment*, 350, 12-27.

Reimann, C., Filzmoser, P., Garrett, R.G. and Dutter, R. 2008. *Statistical Data Analysis Explained*. Applied Environmental Statistics with R. Wiley.

Tarvainan, T. and Jarva, J. 2011. Using geochemical baselines in the assessment of soil contamination in Finland. Chapter 15, 223-231. In: *Mapping the Chemical Environment of Urban Areas*. Johnson C.C., Demetriades, A., Locutura, J. and Ottesen, R.T. (editors). John Wiley & Sons, Ltd., Chichester, UK.

Tidball, R.R., Erdman, J.A. and Ebens, R.J. 1974. Geochemical baselines for sagebrush and soil, Powder River Basin, Montama Wyoming. *US Geological Survey Open-file Report*, Vol. 74-250, 6-13.

Zhao, F.J., McGrath, S.P. and Merrington, G., 2007. Estimates of ambient background concentrations of trace metals in soils for risk assessment. *Environmental Pollution*, 148, 221-229.