Talanta 85 (2011) 2599–2604

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Recovering incomplete data using Statistical Multiple Imputations (SMI): A case study in environmental chemistry

Theresa G. Mercer[a,*], Lynne E. Frostick[b], Anthony D. Walmsley[c]

[a] Centre for Adaptive Science and Sustainability, Department of Geography, University of Hull, HU6 7RX, United Kingdom
[b] Department of Geography, University of Hull, HU6 7RX, United Kingdom
[c] Department of Chemistry, University of Hull, HU6 7RX, United Kingdom

ARTICLE INFO

Article history:
Received 31 January 2011
Received in revised form 8 August 2011
Accepted 10 August 2011
Available online 17 August 2011

Keywords:
Environmental data
Missing values
Statistical Multiple Imputation

ABSTRACT

This paper presents a statistical technique that can be applied to environmental chemistry data where missing values and limit of detection levels prevent the application of statistics. A working example is taken from an environmental leaching study that was set up to determine if there were significant differences in levels of leached arsenic (As), chromium (Cr) and copper (Cu) between lysimeters containing preservative treated wood waste and those containing untreated wood. Fourteen lysimeters were setup and left in natural conditions for 21 weeks. The resultant leachate was analysed by ICP-OES to determine the As, Cr and Cu concentrations. However, due to the variation inherent in each lysimeter combined with the limits of detection offered by ICP-OES, the collected quantitative data was somewhat incomplete. Initial data analysis was hampered by the number of 'missing values' in the data. To recover the dataset, the statistical tool of Statistical Multiple Imputation (SMI) was applied, and the data was re-analysed successfully. It was demonstrated that using SMI did not affect the variance in the data, but facilitated analysis of the complete dataset.

1. Introduction

Missing values are a common occurrence in environmental studies due to the natural unpredictability of experimental conditions and shortcomings in data acquisition [1] and experimental design. Where missing values occur in a dataset, there are usually accompanying problems with computation and application of statistical analysis. This may result in the inability to analyse and interpret the data. A commonly used solution to this problem is to omit data taken from a sample with missing values from the dataset [1,2], although this leads to bias of the data and loss of information [2]. Alternatively, the missing values must be filled (imputed) with representative values so that the entire dataset can be used.

Statistical Multiple Imputation (SMI) is a technique that can be used to overcome missing values. It has its origins in governmental census and questionnaires [3] where it was developed to cope with two common situations: missing data, where values simply were not entered and with-held data, where the person filling out the census was actively hiding some data (yearly salary, earning etc.) Both these situations would cause the statistical treatment of the data to be somewhat skewed. Of course, when the number of missing or incorrect values is low, then the leverage effect on the data is small; but as the number rises, then the effect increases. At the core of the SMI is the replacing of the missing values with a credible value, and to perform this replacement multiple times – thus the resulting summary statistics are based upon multiple imputations rather than just one single set (single imputation). This addresses the uncertainty linked to the missing data [1]. Of course, the problem is now shifted to the choice of the credible values. The need to be able to determine credible values means that for a correct application of SMI there is a need for a significant quantity of useable data to estimate the distribution for the missing values. Typically, environmental data is well suited to this, as it normally has large datasets with small and random sets of missing values or data.

Few examples of the use of SMI for environmental data exist in the literature, although it has a wide applicability in this area. Hui et al. [4] used the technique to fill in missing gaps in eddy covariance flux data taken from the atmosphere. By using this method, the authors were able to calculate annual estimations and to improve the analysis of uncertainties. Sartori et al. [2] used the technique in a cancer mortality analysis with estimated exposure dose to account for missing body index mass values in some participants. They found that the technique corrected for bias and led to increased efficiency in the data. Davison and Hemphill [5] used SMI to use observed levels of related air pollutants to impute missing ozone observations where there was incomplete days of data. This allowed for analysis of the data.

---

[*] Corresponding author.
E-mail address: T.Mercer@hull.ac.uk (T.G. Mercer).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.08.022

In this paper we take the example of an environmental leaching study looking at heavy metal release from preservative treated wood into soil. This particular study illustrates the type of missing values that may be experienced in environmental studies and the successful application of SMI to allow for application of statistics. The preservative wood used in the investigation was treated with chromated copper arsenate (CCA). CCA has been the most commonly used in the UK up until the introduction of the Marketing and Use EU Directive 2003/02/EC restricting its use. The preservative prolongs the service life of wood by 20–50 years by making it resistant to biological attack [6]. As such, in-service CCA-treated wood is expected to be a major component of the UK wood waste stream in the future. Concerns over the impact of the chemical constituents of this treatment on the environment and human health have prompted the introduction of legislation to ensure that such waste is disposed of safely in Hazardous Waste Landfills. Despite this, studies have shown that it can still enter into the landscape mulch market due to inadequate detection equipment and increasing societal pressures to recycle [7,8].

The aim of this study was to assess the leaching of arsenic, chromium and copper from CCA-treated wood by comparing it to leaching from untreated wood. Statistical Multiple Imputations (SMI) were performed on gaps in the data to allow for PCA and standard multivariate statistical analysis of the dataset.

2. Experimental

2.1. Leaching trays (lysimeters) setup

Lysimeters are a standard method used for water-balance and waste studies although they are also applicable to measuring behaviour of trace elements, heavy metals and microorganisms [9]. A lysimeter study allows exposure to natural conditions but with some degree of control over the factors studied. They also allow for collection and measurement of leachate at the base once it has passed through the soil column. In these experiments fourteen plastic lysimeters were constructed. All lysimeters were filled with 20 kg of commercially bought topsoil. Seven of these were overlain with 4.5 kg of chipped CCA-treated wood waste mulch. The other seven were overlain with chipped untreated wood of the same species. Seven replicates for each treatment was considered to be sufficient for redundancy in statistical analysis. The lysimeters were all placed in a secure outdoor location (University of Hull Botanic & Experimental Garden, Cottingham) and exposed to natural weather conditions (Fig. 1).

The specific preservative used in the investigation was CCA-type C consisting of 34% As_2O_5, 47.5% CrO_3 and 18.5% CuO [10]. The resultant leachate (measured in ppb) represents a relative scale, as it contains the ratio of the variables. This type of data is referred to as compositional data and requires further treatment as standard multivariate analysis requires the scale to be absolute. This type of data is also very common in many environmental fields. For example, in petrology, geochemical compositions of rocks are expressed as percentages of weight of major oxides [11]. In sedimentology, ternary diagrams are commonly used to represent the proportions of sand, silt and clay [10]. In social sciences using surveys and questionnaires, answers to multiple choice questions will result in compositional data. Korhoňová et al. [12] discuss the further treatment of compositional data, in this case coffee aroma measured with six volatile compound markers. In their study they compared PCA and cluster analysis on the raw data against centred log-ratio transformations of the data and found that the transformation provided better results and interpretation of the dataset. Other recent examples where compositional data analysis has provided improved results includes research into the male mating preference of cichlids [13] and multivariate analysis of sedimentary chemofacies [14].

Fig. 1. Experimental layout.

2.2. Sample retrieval and analysis

The experimental period began on the 9th of May 2007 and ran for 21 weeks. Leachate samples were collected after periods of rainfall. These were immediately returned to the laboratory and stored at 4 °C for total metal analysis. Other environmental indicators such as redox potential and pH were measured from the leachate but total metal levels were selected as a target [15].

In some cases, collection bottles were not available due to the sheer quantity required. Furthermore, there were some inevitable losses of leachate due to broken or lost sample tubes. In addition, not all boxes had leachate available to be collected. As such a full range of samples was not available at every sampling interval. This variability in sample availability (and therefore data availability) is often a problem in environmental studies. Therefore, this limitation dictated the experimental set-up ensuring that there were adequate replicates for each treatment.

All leachate samples underwent pressurised microwave assisted acid digestion to extract total metals from the samples. Leachate samples were measured out w/w and diluted with 70% concentrated HNO_3 to 10% concentration [16]. Blanks and spiked samples were included in the process to ascertain spike recovery from the digestion process and cleanliness of the microwave vessels. Methods were set-up in the Microwave Accelerated Reaction System (MARS^{TM}) CEM Corporation, USA. Following dissolution and cooling down of the digests, the solution was transferred to a 50 ml volumetric flask and diluted to the required volume using UHQ water [17].

All the digestates were analysed using an inductively coupled plasma optical emission spectrometer (ICP-OES) (Perkin Elmer Optima 5300DV, Connecticut, USA). A limit of measurability (LOM) experiment was carried out on an inductively coupled plasma mass spectrometer (ICP-MS) when it was discovered that the metal concentrations in some of the preliminary leachate samples were at very low levels (ppb) and below the limit of detection (LOD) of the ICP-OES. For As, Cr and Cu, the minimum measurable levels were 8.8, 2.6 and 3.8 ppb, respectively. For the LOM study, calibration standards were created between 1 and 10 ppb for As, Cr and Cu (calibration curve r^2 for As = 0.999; Cr = 0.999, Cu = 0.999). Spiked acid and equivalent matrix matched samples were created within the range of 1–8 ppb to run as samples. The results of the experiment can be seen in Table 1. These illustrate the limiting effects of sample matrix and digesting of samples on measurability of data compared with clean acid.

A certain degree of method development was required in the preliminary stages of analysis to ensure that the ICP-OES (Perkin Elmer Optima 5300DV, Connecticut, USA) was running

Table 1
Limit of measurability.

Concentrations	As		Cr		Cu	
	Matrix	Acid	Matrix	Acid	Matrix	Acid
1 ppb	−3.36	3.18	2.47	1.23	0.66	1.32
2 ppb	−2.10	1.67	3.53	2.30	9.55	2.56
3 ppb		3.58		3.38		3.5
4 ppb	−3.21	6.76	5.41	4.41	19.01	4.72
5 ppb		8.71		5.49		5.91
6 ppb		7.68		6.82		6.79
7 ppb		15.41		7.86		8.23
8 ppb	4.59	10.86	9.38	8.52	16.41	9.01

appropriately for the samples and concentrations measured. This helped reduce spectral, physical, chemical and memory interferences as outlined by the EPA [18]. The units of measurement were selected depending on sample type and expected concentration levels (e.g. ppm for soil and woodchips, ppb for leachate). Integration time was lengthened from 2 to 10 s to allow for stabilisation of the spectral signal by averaging background noise over a longer time. The purge gas flow was changed from normal to high to account for the wavelength of arsenic (<200 nm). ICP plasma optimisation was carried out regularly with 1 ppm manganese. This serves to align the view and provide a maximum signal to background ratio [18]. The view was changed from radial to the more sensitive axial view for samples with lower concentrations of Cu, Cr and As. Rinse time was also increased from 60 to 80 s to ensure all elements were washed out of the system between samples. All samples were run interspersed with blanks and standards to monitor for instrumental drift. Replicate readings were taken for each sample and averaged. The selected wavelength readings taken from the samples were based on the recommendations of Perkin Elmer for Optima^{TM} Series ICP-OES Spectrometers with As measured at 188.979 nm, Cr at 267.716 nm and Cu at 327.393 nm. Quantitative determination of element concentrations was calculated using standards to create calibration curves of intensity versus concentration [19].

2.3. Data processing

2.3.1. Calculation of metal concentrations

The calculations of the metal concentrations from the ICP-OES data were conducted in Microsoft Excel 2007. Firstly, the LOD was calculated as three standard deviations (SD) of the blanks. For each metal, a good agreement between the wavelengths was required. Readings were taken from wavelengths that exceeded the LOD and the concentration in the original samples calculated with Eq. (1) where m is the metal concentration in solution and DF is the dilution factor.

$$\text{Metal concentration in original solution} = m \times DF \tag{1}$$

2.3.2. Calculation and theory of missing value replacements

Due to the missing values in the data, standard statistical analysis of the dataset was not possible. The extent of the missing values is shown in Fig. 2 (also refer to supplementary data Tables S-1 and S-2). SPSS (v. 16) was used to calculate both mixed and repeat analysis of variances (ANOVA). The missing values across the dataset caused only six acceptable cases out of a possible seven per batch to be taken for the lysimeters containing untreated wood

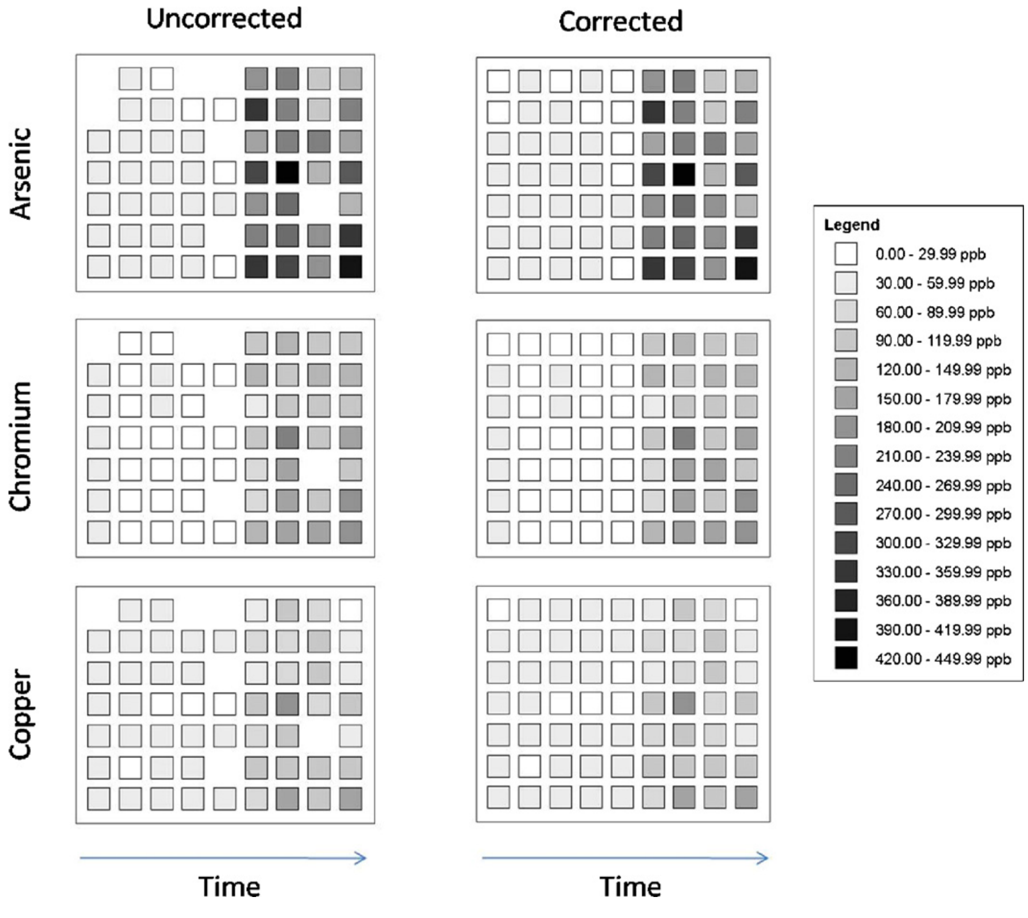

Fig. 2. Uncorrected and corrected leachate data.

and only three acceptable cases out of a possible seven per batch for the lysimeters containing CCA-treated wood. As such, the ANOVA could not be carried out due to reduced replication and therefore variation.

There are a number of missing-data mechanisms that need to be considered in relation to underlying values of the variables in the dataset. These are described by Little and Rubin [20]. The first termed missing at random (MAR), occurs when each missing value depends only on the components that are observed and not on the components that are missing [20,21]. A second more restrictive example is called missing completely at random (MCAR), where the missing data are a random sample of the whole dataset and the 'missingness' is independent of the data values. When the probability of a missing value depends on the missing part of the data, the mechanism is known as not missing at random (NMAR) [20].

A compounding problem arises with compositional data, where all the components should be positive real numbers as they are constituent parts of a whole [11]. Therefore, it is necessary to replace both types of zero with an appropriate value [11]. In this work there are two types of missing values – both similar to the problems for which SMI was developed. The first concerning zero values (or rounded zeros [10,11]) apply to analyte concentrations that are below the limit of detection. These are classed as NMAR [21]. With rounded zero values, traditionally a zero would be recorded, but statistically zero is the least likely value for the analysis. The credible value in this situation will be distributed between zero and the LOM. To enable the LOM to be determined, a series of experiments were performed with real samples, as the LOM is not the same as the theoretical LOD.

There are a number of methods that can be used to correct for rounded zeros. These are usually based on substitution or imputation strategies. With any method, the replacements should not significantly distort the structure of the data or the covariance structure of the components [22]. For compositional data, there are four replacement strategies. The first; Additive Replacement Strategy involves replacing the zeros by a small value and then normalizing the imputed compositions [23]. However, this method can distort the covariance structure of the data set [23]. A second strategy; Multiplicative Replacement Strategy imputes exactly the same value in all the zeros of the compositions. This replacement introduces artificial correlation between components which have zero values in the same composition. When the number of rounded zeros is quite large, parametric methods such as the Expectation Maximisation (EM) algorithm [24] and multiple imputation method are recommended [21]. The EM method uses the relationship between the observations and variables for estimating the missing values [25]. However, EM regression-based imputation methods require at least one column of data to be complete, and preferably to have data across many columns. In this work there are only three columns in the data, for the 3 metals, and none of the data is complete throughout the set. This lack of complexity in the data lends itself to a simpler imputation approach, such as the one used in this work.

The LOM results indicate that for metal concentrations below 8 ppb, the data did not reflect the spiked concentrations. As a result, readings below the LOD were recorded as zero. It was then possible to correct the zeros in the dataset using the following method. In order to carry out the multiple imputations, the data were divided into columns of the single metals measured and the values of zero were removed. They were then used to check for normal distribution. Because the data was skewed, random numbers were used for multiple imputation calculations. Random numbers were generated between zero and the minimum value measured for each metal on the ICP-OES. The random numbers were then multiplied by the minimum values to produce a Hot Deck of suitable numbers. For each metal, the data is then copied into 10 separate columns and where there is a zero value, it is replaced with a value from the Hot Deck. The mean of each column is calculated, and then the column means are then themselves averaged to produce the corrected mean. The same method can be applied to calculate the corrected standard deviation. At the end of each row (sample), the averages of the row are calculated and these figures are the imputed data averages.

The second type of missing values occurred where there was no sample generated (also known as essential zeros). These are classed as MAR [22] and are suited to the application of observed data likelihood such as in the multiple imputation method. The case where the sample is unavailable is rather more unusual – but did occur with a fairly high degree of frequency in this research. As the lysimeters were exposed to natural meteorological conditions, the amount of rainfall that they experienced was random. During periods with little rainfall there was often no leachate to collect, and subsequently no analysis could take place. To obtain a credible value in this instance, a value for the analyte is chosen that lies within the distribution of the concentrations for the other replicate lysimeters. This approach should ensure that the dataset does not become distorted. In these cases where there was insufficient leachate for ICP-OES analysis, a hot deck of suitable values was produced from the distribution of values found for the replicate lysimeters in each batch to generate randomly distributed numbers in the desired range. These values were then used to replace the missing ones and the summary statistics were computed.

The imputed data was validated by checking the variances of the imputed data sets and by using several performance indices described in Junninen et al. [26]. These determine the success of the imputation methods made against known values. The first index performed was the index of agreement (d_2), which is more useful than the commonly used R^2 to overcome the potential discrepancies between the predicted and observed values. The second index, root mean squared error (RMSE), was used to calculate the average error of the model. The final index mean absolute error (MAE) is a more sensitive measure of residual error than RMSE [26]. The performance indicators were tested by selecting four continuous sample dates for one of the metals where all data was available. An observed datapoint was then removed from each date and imputed back using the method described above.

Following the replacement of the missing values, log-transformations were applied to the corrected data. These should allow easier mathematical handling than using ratios based on the components. Furthermore, they allow the use of multivatriate techniques on the data whilst preserving the relative magnitude of the components and the compositional vectors in real space [21]. Those that are typically applied to compositional data are the additive log-ratio (alr), isometric log-ratio (ilr) and centred log-ratio (clr) [21]. All three log transformations were performed on the corrected data using the CoDaPack program [27] developed at the University of Girona. However, the transformed data resulted in negative numbers and missing data. This further prevented the data from being analysed with PCA and multivariate statistics. As such, statistical analyses were applied directly to the corrected data.

2.3.3. Multivariate analysis

The research question posed with the experiment was to determine if there is a difference in the leaching of metals from the treated and untreated samples, and the traditional method for this analysis is the analysis of variance (or ANOVA). Once the main effects have been determined by ANOVA, it is common to then explore the data with tools such as principle component analysis (PCA) to elucidate features such as groupings and interactions among datasets. Both these methods are based upon the variance in the data, and as such they are both badly affected by incomplete datasets. Both these techniques were carried out using SPSS v. 16.

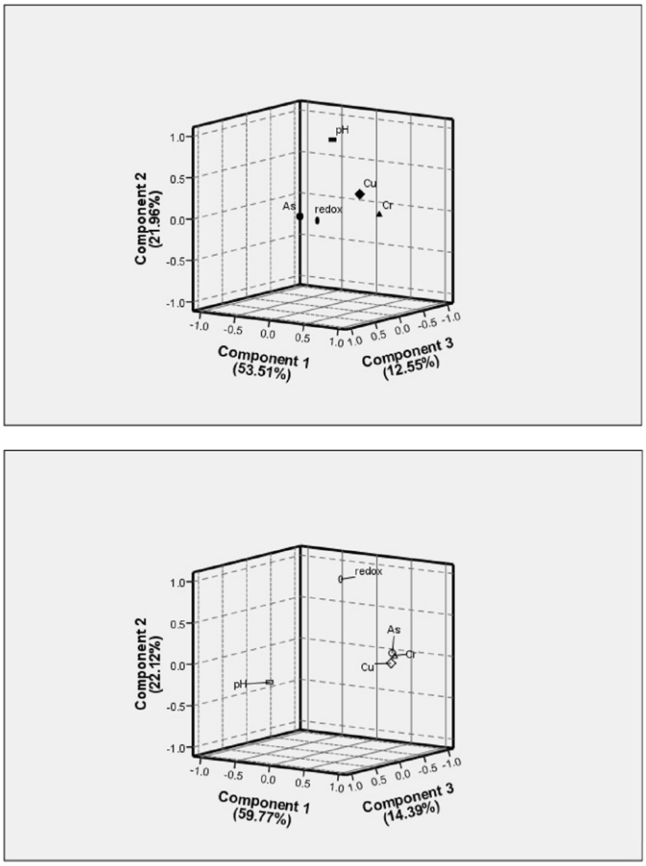

Fig. 3. Principle component plots for data from untreated wood lysimeters (top) and CCA-treated wood lysimeters (bottom). Percentage of variance explained is listed after each component.

3. Results and discussion

The variances of the imputed datasets were checked to see if they differ significantly using an *F*-test. They were found not to be significantly different at the 5% level. The performance indicators also illustrated the success of the imputations ($d_2 = 0.506$; RMSE = 6.215; MAE = 5.673). The d_2 results fall well within the range of the Junninen et al. [26] study, whilst the RMSE and MAE are lower than the values and as such denote improved performance. Furthermore, the outcomes of the ANOVA were not significantly different between the generated datasets ($p < 0.05$). Therefore, using the imputed dataset did not make a difference to the statistical analyses applied to it. However, without undertaking this simple procedure, the data could not be analysed or interpreted fully (i.e. only 3 cases out of a possible 14 could be used in the ANOVA).

PCA was carried out on the corrected data from the untreated and treated samples (Fig. 3). In the lysimeters containing untreated wood, the three metals behaved differently from each other with arsenic falling in principal component 1 and chromium and copper in principal component 2. pH and redox were found to be inversely or negatively correlated. In the lysimeters containing CCA-treated wood waste, the three elements are grouped very closely, suggesting that they vary in the same way throughout the experiment. Redox and pH vary independently of metal concentrations and all the results cluster together (Fig. 4). The elements in the leachate from the lysimeters containing CCA-treated wood waste vary consistently with each other. However, this contrasts with the results for lysimeters with untreated wood, where arsenic varies differently from copper and chromium.

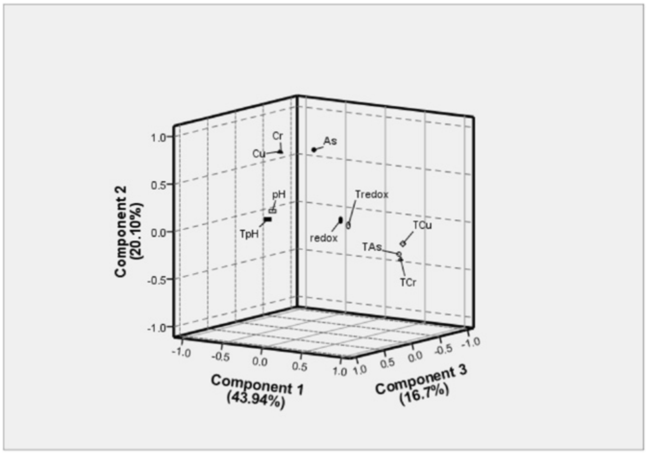

Fig. 4. Combined principal component analysis plot (T prefix denotes leachate results taken from the CCA-treated wood and soil lysimeters. No prefix denotes leachate results taken from the untreated wood and soil lysimeters). Percentage of variance explained is listed after each component.

Table 2
ANOVA data.

Source of variation	*d.f.*	SS	MS	Variance ratio
Time	8	282,767.531	35,345.941	6.355
Metals	2	137,773.759	68,886.8795	27.757
Time × metals	16	116,416.441	7276.0275	11.975
Error (time)	104	578,419.621	5561.727	
Error (metals)	26	64,526.088	2481.773	
Error (time × metals)	208	126,383.225	607.612	

A repeat design ANOVA was carried out on the data to better define these differences (Table 2). It was found that temporal trends were significant ($p < 0.05$) as well as there being significant differences between the two treatments ($p < 0.05$). Furthermore, there was a significant interaction between the metals and time ($p < 0.05$). The 'time × metals' interaction had double the variance of time. The relationship suggests that the sampling period had a different effect on each metal. It is evident that treated wood has an effect on both the variance of the metal analytes and the concentration of analytes detected.

The use of Statistical Multiple Imputation has been shown to be a very useful tool deal with missing values within a known distribution and can be used in environmental research to deal with natural datasets. However, this technique is not a substitute for good experimental design [28] and was only effective because an adequate number of replicate samples were available to determine the actual distribution range for the analytes under investigation. Combined with the limit of measurability study which accurately determined the realistic detection limits of the method used, it has been possible to correct a dataset with a relatively simple procedure, without distorting the data as a whole (viz. the summary statistical analysis). The results of the statistical calculations show that the lysimeters containing CCA-treated wood waste behave very differently from those containing untreated wood. Temporal effects were significant and the high levels of leaching of metal(loid)s observed in the CCA-treated wood and soil lysimeters led to exceedances of UK environmental quality standards and WHO drinking water guidelines by an order of magnitude.

Acknowledgements

This research was funded by an EPSRC doctoral studentship. We would like to thank Gervais Sawyer for his help with species identification, Bob Knight for his help with the laboratory analysis and Richard Middleton for his help with the GIS work. We are also very grateful to the reviewers for their comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.talanta.2011.08.022.

References

[1] A. Plaia, A.L. BondÌ, Atmos. Environ. 40 (2006) 7316–7330.

[2] N. Sartori, A. Salvan, K. Thomaseth, Comput. Stat. Data Anal. 49 (2005) 937–953.

[3] E.B. Rubin, Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York, 1987.

[4] D. Hui, S. Wan, B. Su, G. Katul, R. Monson, Y. Luo, Agric. Forest Meteorol. 121 (2004) 93–111.

[5] A.C. Davison, M.W. Hemphill, Atmos. Environ. 21 (1987) 629–639.

[6] I.V. Christensen, A.J. Pedersen, L.M. Ottosen, A.B. Ribeiro, Sci. Total Environ. 364 (2006) 45–54.

[7] M. Blassino, H. Solo-Gabriele, T. Townsend, Waste Manage. Res. 20 (3) (2002) 290–301.

[8] T.G. Townsend, H. Solo-Gabriele, T. Tolaymat, K. Stook, Sci. Total Environ. 309 (2003) 173–185.

[9] T. Takamatsu, M.K. Koshikawa, M. Watanabe, H. Hou, T. Murata, Eur. J. Soil Sci. 58 (2007) 329–334.

[10] H. Hasan, L. Hu, H.M. Solo-Gabriele, L. Fyber, Y. Cai, T. Townsend, Environ. Pollut. 158 (2010) 1479–1486.

[11] J. Atchison, The Statistical Analysis of Compositional Data, The Blackburn Press, New Jersey, 2003.

[12] M. Korhoňová, K. Hron, D. Klimčíková, L. Müller, P. Bednář, P. Barták, Talanta 80 (2) (2009) 710–715.

[13] M.E.R. Pierotti, J.A. Martin-Fernández, O. Seehausen, Evolution 3 (9) (2009) 2372–2388.

[14] J.C. Montero-Serrano, J. Palarea-Albaladejo, J.A. MartÍn-Fernández, Sediment. Geol. 228 (2010) 218–228.

[15] T.G. Mercer, An Experimental Study of the Environmental Impacts of CCA-Treated Wood Waste Land Application, Ph.D. Dissertation, University of Hull, 2010.

[16] Environmental Protection Agency. Method 3015: Microwave Assisted Acid Digestion of Aqueous Sample and Extracts (online), 1994, available: http://www.epa.gov/sw-846/pdfs/3015.pdf (accessed 13.03.07).

[17] Q.J. Song, Elemental Speciation in the Composting Process by ICPMS. Ph.D. Dissertation, University of Hull, 2003.

[18] Environmental Protection Agency. Method 6010B: Inductively Coupled Plasma-Atomic Emission Spectrometry (online), 1996, available: http://www.uhu.es/tamara.garcia/quimI/apuntes/EPA-ICP.pdf (accessed 20.03.07).

[19] C.B. Boss, K.J. Fredeen, Concepts, Instrumentation and Techniques in Inductively Coupled Plasma Optical Emission Spectrometry, 3rd ed., Perkin Elmer, USA, 2004.

[20] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, 2nd ed., John Wiley & Sons, New York, 2002.

[21] J. Palarea-Albaladejo, J.A. MartÍn-Fernández, J. Gómez-GarcÍa, Math. Geol. 39 (2007) 625–645.

[22] J.A. MartÍn-Fernández, C. Barceló-Vidal, V. Pawlowsky-Glahn, Math. Geol. 35 (3) (2003) 253–278.

[23] R. Hijazi, Proceedings of the 4th International Workshop on Compositional Data Analysis, Girona, 2011.

[24] J. Palarea-Albaladejo, J.A. Martín-Fernández, Comput. Geosci. 34 (2008) 902–917.

[25] K. Hron, M. Templ, P. Filzmoser, Comput. Stat. Data Anal. 54 (2010) 3095–3107.

[26] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, Atmos. Environ. 38 (2004) 2895–2907.

[27] M. Comas-Cufi, S. Thió-Henestrosa, in: J.J. Egozcue, R. Tolosana-Delgado, M.I. Ortego (Eds.), CoDaWork'11: 4th International Workshop on Compositional Data Analysis, Sant Feliu de GuÍxols, 2011.

[28] D.J. Finney, Exp. Agric. 24 (1988) 421–432.