Jul 1st, 12:00 AM

Enhanced Presentation and Analysis of Uncertain LCA Results with Principal Component Analysis

P. J. Notten

J. G. Petrie

Enhanced Presentation and Analysis of Uncertain LCA Results with Principal Component Analysis

P.J. Notten[a] **and J.G. Petrie**[b]

[a] *Dept. of Chemical Engineering, University of Cape Town, South Africa; now at: 2.-0 LCA consultants, Amagertorv 3,2, 1160 Copenhagen K, Denmark, pin@lca-net.com*

[b] *Dept. of Chemical Engineering, University of Sydney, Australia*

Abstract: A significant challenge of an uncertainty assessment is the presentation of the results, since a quantitative uncertainty analysis dramatically increases the already considerable amount of data that needs to be communicated in an LCA study. This paper investigates three graphical options to interpret output samples from quantitative uncertainty analyses. The output samples are from case studies within the coal-fired power generation sector, and include an assessment of empirical uncertainty from a stochastic uncertainty assessment and an assessment of uncertainty in decision variables from a parametric sensitivity analysis. Two commonly used representations of probabilistic samples are demonstrated, namely "box and whisker" plots and plots of the cumulative probability density function, as well as the multivariate geometric technique, principal component analysis (PCA). Cumulative probability plots are useful representations of uncertainty where a quantitative estimate of the relative uncertainty between options is required, but they become extremely tedious (many pair-wise combinations) and difficult to interpret when a large number of options are compared over many criteria. In such cases, PCA can be used to provide a valuable overview of the results, where it is able to clearly present any trade-offs that have to be made between selection criteria, and the "spread" of the options under consideration over the decision space. Box and whisker plots are good at representing the relative importance of empirical parameter uncertainty and the uncertainty arising from the choice of decision variables, and show the degree of shifting between the options as well as the full range over which the options potentially act. The three representations of uncertainty are found to complement each other, as each enhances different aspects of the results. The most appropriate graphical presentation method is found to depend on the particular decision context and the particular stage of the analysis.

Keywords: Stochastic results; Uncertain LCA results; Principal Component Analysis; Presentation

1. INTRODUCTION

A quantitative analysis of the uncertainty is becoming an increasingly accepted component of a life cycle assessment (LCA) study. The use of stochastic models and the presentation of results in ranges or as confidence intervals have been shown by a number of authors to enhance the decision support capabilities of an LCA study (e.g. Meier [1997]; Maurice et al. [2000]). Nonetheless, considerable challenges remain with regard to incorporating quantitative uncertainty analyses into life cycle assessment [Notten, 2002], notably with characterising the uncertainty of the input parameters, modelling correlated inputs, and analysing the considerable volume of data resulting from the analysis. This paper focuses only on the latter aspect of the uncertainty analysis.

Presenting and analysing the large data sets resulting from an LCA study is already a demanding process. These demands increase considerably when the results are extended to include a consideration of uncertainty, with each data point replaced by an output sample, as well as an increase in the number of scenarios requiring consideration. This paper therefore looks at ways of analysing and communicating uncertain results, including a novel technique using principal component analysis (PCA). The paper demonstrates the use of the methods with respect to a case study looking at technology options in coal-fired power generation, and draws conclusions as to the relative strengths of the different methods.

2. GRAPHICAL REPRESENTATIONS OF UNCERTAINTY

2. 1 Commonly used statistical methods

Uncertain data are often defined using common statistical measures, such as the variance, confidence intervals etc. Whilst such statistics are efficient at summarising the uncertainty of the data sample, graphical methods are typically the most effective in communicating insights into uncertain data sets. Three basic ways of presenting probabilistic results are:

- The probability density function (PDF),
- The integral of the PDF, the cumulative density function (CDF), and
- Displaying selected fractiles, as in box and whisker plots.

Examples of these commonly used graphical representations of uncertainty can be found in the following case study section. Each of the representations emphasise different aspects of the probability distribution of the output sample. PDFs give the relative probabilities of the different values and show the shape of the distribution. Box and whisker plots emphasise confidence intervals and means, and are thus a simple way to represent uncertain results when the range and not the distribution shape is primarily of interest.

The CDF also provides little information on the shape of the distribution, but is the best option if information on the fractiles of the distribution is required (i.e. the probability that the actual value of the variable is less than a particular value). The CDF is often preferred to the PDF when presenting stochastic output samples, because it looks a lot less noisy with the equivalent sample size. Comparative studies often share a number of sub-processes for which identical data have been used, with the resultant correlation between the output samples removed by basing the analysis on the normalised difference between the output samples [Coulon et al., 1997; Meier, 1997]. This has the added benefit of producing an easy to interpret CDF plot, in that the y-intercept of a CDF plot of the difference between two options shows the degree of confidence that can be held that the one option always performs better than the other.

A significant draw-back of CDF plots is that they are limited in the number of dimensions they can display. Conclusions have to be drawn across a large number of single plots if many options are to be assessed across many different environmental indicators. It is therefore difficult to get an overview of the results. A potential solution is to use the multivariate data analysis technique principal component analysis (PCA), which is able to reduce the dimensionality of the data set by producing a planar view of the data. The use of PCA for the presentation and analysis of LCA results has been demonstrated by Le Teno [1999], and its use explored in a number of complex case study situations in Notten [2002]. The following section discusses the main features of PCA with respect to its use in LCA.

2. 2 Principal component analysis

The goal of PCA is to represent the variation present in many variables in a small number of factors (or principal components), which are found via a mathematical manipulation of the data matrix. A new space in which to view the data is constructed by redefining the axes using these factors, instead of the original variables. The new axes allow the analyst to view the true multivariate nature of the data in a relatively small number of dimensions, allowing the identification of structures in the data that were previously obscured. The principal components are the eigenvalues of a correlation or covariance matrix of the input data, whilst the co-ordinates of the transformed variables on the principal component plane are given by the eigenvectors. The theory of principal component analysis can be found in most multivariate data analysis textbooks, e.g. Murtagh and Heck [1987].

The results of a PCA are best analysed graphically (see Figure 1). Stochastic output samples plot as clouds of points, which can be interpreted as “zones of confidence” [Le Téno, 1999] (see Figure 1). In this figure it is the structures or patterns in the data that are of interest, where the distances between the clouds of points determine similarities (or differences) between the options, and the overlap between the clouds visually identifies the significance of the rankings between the options. The axes do not have any physical meaning, and are merely measures of proximity that are interpreted as similarity. It is thus only the relative distance between the clouds of points, and their size and degree of overlap that are of note.

To interpret the principal component plot it is necessary to look at the factor-variable correlations given by the eigenvectors. These provide a measure of each variable’s contribution to the principal components, and indicate which variables are best at discriminating between the options under investigation. In Figure 1, the eigenvectors are represented by the lines emanating from the origin, and their length and orientation indicate which variables have the

greatest influence in “pulling” the data apart to create the spatial arrangement of the clouds of points. The eigenvectors also provide insights into the data structure. Highly correlated variables plot close together, thereby pointing to redundant selection criteria, whilst the relative lengths of the lines provide a measure of the relative ability of the variables to discriminate between the options (e.g. impact categories showing no significant differences between the options plot with short lines). Thus an analysis of the eigenvectors identifies the minimum set of impact categories useful for distinguishing between the options.

PCA is based on correlations between the data points, and thus finds where the greatest variations are occurring between the options, not where the largest absolute changes in indicator scores occur. It therefore does not provide information on the relative importance of the potential impacts. This also means that there is no need to correct for the problem of common data elements by basing the analysis on a difference between samples. However, it may be necessary to normalise the data so that the analysis is not skewed by variables operating on very different scales.

The following section demonstrates these features of PCA in a case study, and shows how it complements an analysis using other common representations of uncertain output samples.

3. CASE STUDY

The following case study is an excerpt from a larger study looking at technology options to refurbish coal-fired power plants in South Africa [Notten, 2002]. In particular, possibilities to utilise discard coal, a waste product from coal beneficiation, are investigated. An average year’s performance is chosen as the functional basis for comparison, because of the need to capture those environmental interventions (notably those from solid waste dumps), that only become evident after the plant has been operating for a number of years. The study includes all major processes in the coal-electricity supply chain (mining, coal preparation, coal combustion, flue gas cleaning and solid waste disposal), as well as the production and transport of major ancillary materials (liquid fuels, treatment chemicals etc.).

A rigorous assessment of uncertainty is undertaken in the study, including a parametric sensitivity analysis that systematically investigates model parameter uncertainty (i.e. the choice of operating conditions for the technology options), as well as a probabilistic assessment of empirical uncertainty. An iterative method based on successively refining the distributions of the most influential input parameters is followed, where the probability distributions take into account the certainty and completeness of the data source used to define the parameter, as well as an assessment of the adequacy of the data source in the particular context of the case study. Full details of the uncertainty assessment framework developed in the study can be found in Notten [2002].

The primary aim of this paper is to demonstrate methods for the presentation and analysis of uncertain results. Thus, in order to clearly present the extensive results of this case study in a suitable format (i.e. small, black and white graphics), it has been necessary to reduce the number of options considered, as well as to reduce the scatter of the uncertainty samples (although conclusions are still based on the full results). The results presented here are therefore generated from a random re-sampling of the interquartile range of the output distributions. Also, the results are normalised with respect to a "base case" option (the technology currently in operation), and presented on a relative scale on which the "base case" is assigned a score of 100 for each impact category considered (i.e. a score of less than 100 represents an improvement relative to the currently employed technology, whilst a score greater than 100 represents a higher contribution to that particular impact potential).

Figure 1 presents the results of a principal component analysis on the uncertainty samples of the option sets summarised in Table 1. The principal component "loadings" (the eigenvectors) and the principal component "scores" (the transformed data points) are calculated using standard statistical algorithms (available in most statistical software packages, see Notten [2002] for a summary of the underlying calculations).

Table 1. Primary decision variables causing the observed differences between the options (four plant configurations with fluidised bed combustion (FBC), and one with pulverised fuel (PF) combustion).

Option	PF	FBC A	FBC B	FBC C	FBC D
Boiler	PF	FBC	FBC	FBC	FBC
Cooling	wet	wet	wet	dry	wet
Fuel	coal	discard	discard	discard	blend
Sorbent	-	lime	dolomite	lime	dolomite
SO_2 removal	-	90%	40%	90%	40%

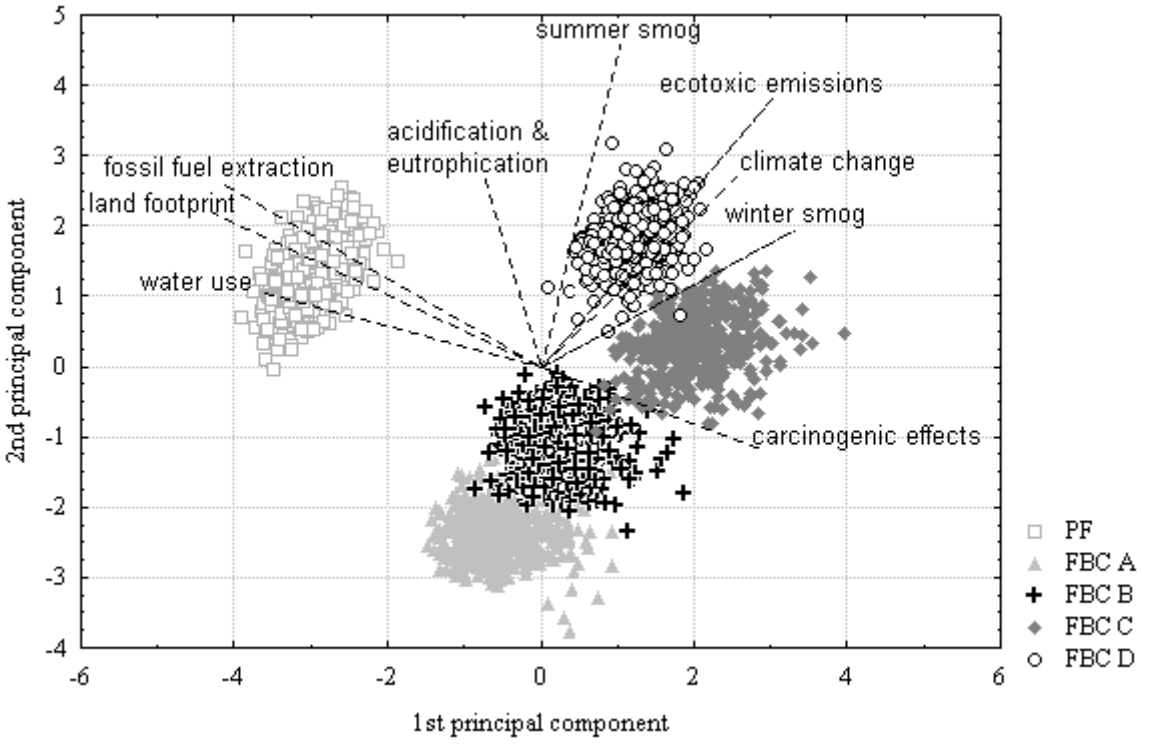

Figure 1. Normalised output samples transformed onto the 1st and 2nd principal component plane (points), with the eigenvectors or principal component "loadings" superimposed (labelled dashed lines).

In Figure 1 the output samples are shown transformed onto the principal component plane (i.e. placed in the new multivariate space defined by the principal components instead of the original variables), together with the relative contributions of each variable to the 1st and 2nd principal components (the lines emanating from the origin). If the lines are thought of as arrows, sample values plotting strongly in the direction of the arrow indicate a poor performance against that criterion, and the reverse for options plotting away from the direction of the arrow (for all indicators considered, a lower contribution means better performance). The relative length of the line indicates the strength of the observed difference between the options. For example, Figure 1 shows there is a certain and substantially higher contribution to fossil fuel extraction by the pulverised fuel (PF) option than the options with fluidised bed combustion (FBC). Whilst among the FBC options, there is a less certain and smaller observed difference between option D and the other FBC options (indicating its slightly higher contribution to fossil fuel extraction).

PCA captures the major sources of variability between the options, and shows which impact indicators are responsible for the observed variations. Although PCA efficiently summarises all sources of variability, Figure 1 only shows the first two principal components. Thus when interpreting Figure 1 it is important to bear in mind the sources of variance not captured by the first two principal components. From Table 2 it can be seen that the first two principal components, accounting for 63% of the variability, capture the large differences exhibited between the PF option and the three FBC options, and the greatest sources of variability between the FBC options (contributions to summer smog and ecotoxicity).

Table 2. Percentage contributions of the first four principal components to the overall variance, and the percentage contributions from the variables to these four principal components.

	PC 1	PC 2	PC 3	PC 4
% contribution to overall variance	35	28	19	7.5
Carcinogenic effects	10	2.1	2.9	85
Summer smog	1.2	31	1.2	0.7
Winter smog	13	5.6	20	3.0
Climate change	7.5	11	15	0.1
Ecotoxic emissions	11	21	1.1	0.1
Acidification and eutrophication	0.6	11	35	0.1
Impacted land footprint	21	7.0	4.5	5.6
Fossil fuel extraction	20	10	1.5	4.9
Water use	15	1.6	19	0.6

The difference in SO_2 removal efficiency between the options causes a different ranking between the FBC options to that found by the 1st principal component. This is responsible for the relatively high contribution to the overall variance by the 3rd principal component, whilst the contribution by the 4th principal component is a result of the very high uncertainty in predicting carcinogenic emissions (i.e. variability within rather than between the output samples).

A fairly extensive overlap between the output distributions is evident in Figure 2. Thus a different representation is required to determine the degree of confidence that can be held in the observed differences. For example, before selecting option A, a decision maker may wish to know how certain he/she can be that option A does indeed have a lower contribution to winter smog than option C. A CDF plot of the difference between their output samples provides this (see Figure 3), where the y-intercept gives the degree of confidence that can be held in the observed difference (shown by the arrow in Figure 3). In this case, there is a high degree of confidence in the relative performance of the options, with 94% of the sample predicting option C to have a higher contribution to winter smog than option A (i.e. their difference is less than zero for only 6% of the sample values). The precise representation of uncertainty in Figure 3 is extremely useful, but since it requires a pairwise comparison of a single indicator at a time, it was first necessary to narrow the option set under consideration.

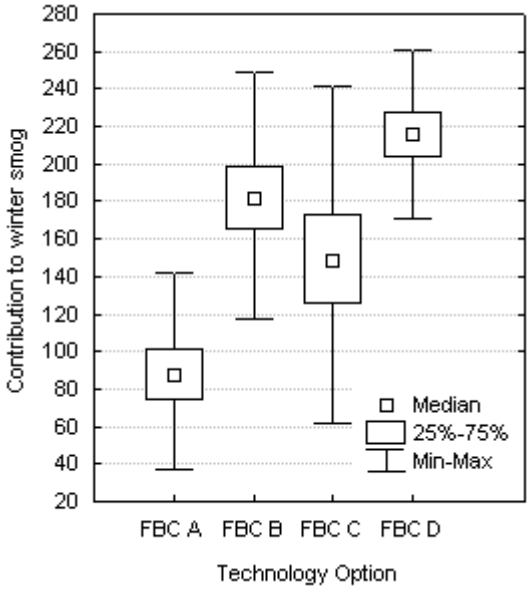

Figure 2. Contributions of the four fluidised bed combustion options to winter smog, relative to a "base case" option with a contribution of 100.

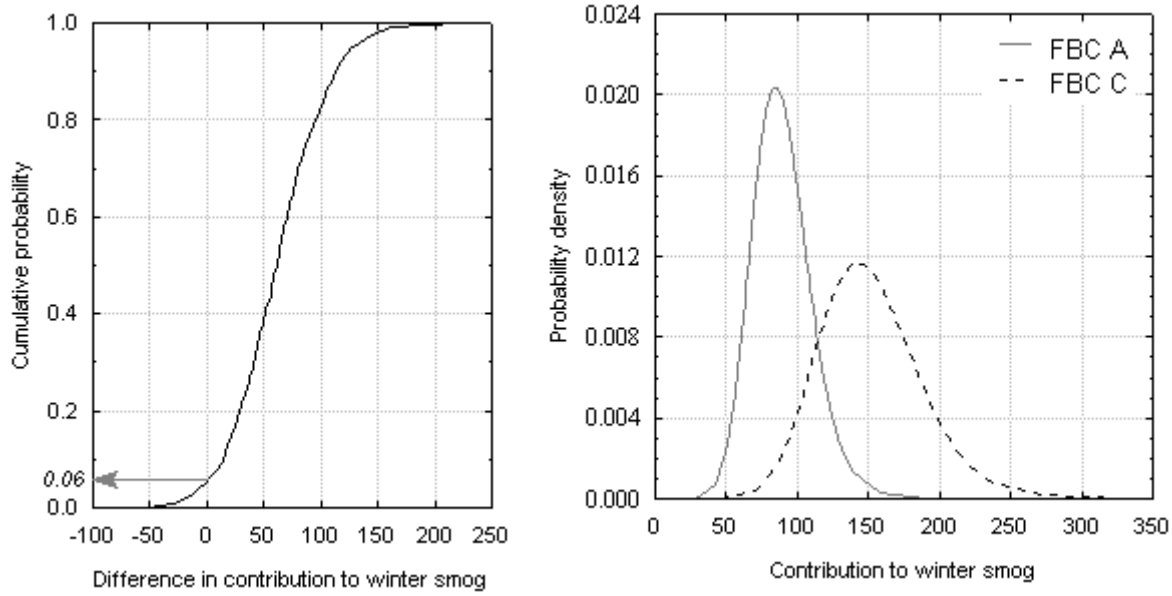

Figure 3. Cumulative probability function (CDF) of the difference between options A and C, and the probability density functions of these two options. The y-intercept of the CDF (shown by the arrow) gives the degree of confidence that option C performs better than option A.

4. CONCLUSIONS

Principal component analysis provides a valuable overview of LCA results, capable of highlighting where the most significant differences between options are occurring, and the impact categories responsible for the differences. This is extremely useful when a large number of scenarios need to be considered, giving a quick indication of the significance of the scenarios with respect to each other. PCA is thus most useful in a strategic type assessment, where a large number of decision variables are likely to require assessment as part of the uncertainty analysis. Even in studies considering a smaller option set, PCA still provides valuable assistance by making explicit the tradeoffs between the impact potentials. Furthermore, PCA provides useful information on the underlying structure of the result sample, particularly with respect to the strength and independence of the criteria chosen to evaluate the systems. However, PCA is best used in combination with other graphical representations of probabilistic samples. The three methods explored in the case study (PCA, box and whisker plots and cumulative probability plots) are found to complement each other, as each is able to enhance different aspects of the result sample.

The principal component representation allows a visualization of the trade-offs that have to be made between impacts, and the spread of the options over the operating space. Box and whisker plots are good at representing the relative importance of decision variable and empirical parameter uncertainty, whilst CDF plots are useful when a quantitative estimate of the relative uncertainty between options is required (i.e. the degree of confidence in the observed difference between the options). Whilst giving the most precise information, CDFs become extremely tedious and difficult to interpret when a large number of options are involved (many pair-wise combinations). In this case, PCA is invaluable, as it enables the full data set to be displayed on a single plot (provided a sufficiently high percentage of the overall variance is displayed by the first two principal components). Box and whisker plots provide a level of information between these two methods. The three representations of uncertainty are therefore most useful when used successively to narrow the option set.

REFERENCES

Coulon R., V. Camobreco, H. Teulon, and J. Besnainou, Data Quality and Uncertainty in LCI, International Journal of Life Cycle Assessment, 2(3), 178-182, 1997.

Le Téno J-F., Visual Data Analysis and Decision Support Methods for Non-Deterministic LCA. *International Journal of Life Cycle Assessment*, 1(4), 41-47, 1999.

Maurice, B., R. Frischknecht, V. Coelho-Schwirtz, and K. Hungerbühler, Uncertainty Analysis in Life Cycle Inventory. Application to the Production of Electricity with French Coal Power Plants, *Journal of Cleaner Production*, 8, 95-108, 2000.

Meier M.A., Eco-Efficiency Evaluation of Waste Gas Purification Systems in the Chemical Industry, LCA Documents, Vol. 2, Eco-Informa Press, Bayreuth, 1997.

Murtagh, F., and A. Heck, Multivariate Data Analysis, Reidel Publishing Company, Dordrecht, 1987.

Notten, P., Life Cycle Inventory Uncertainty in Resource Based Industries - A Focus on Coal-Based Power Generation, Department of Chemical Engineering, University of Cape Town, 2002.