AN EXPERT SYSTEM APPROACH TO LCI DATABASE MANAGEMENT

Notten P and Weidema B
2.-0 LCA consultants
Borgergade 6, 1., 1300 Copenhagen K., Denmark
pin@lca-net.com

ABSTRACT: Tools to check for errors and inconsistencies in LCI databases are currently lacking, as are systems to aid in the selection of LCI data. This is because the large and aggregated nature of LCI databases make them difficult to evaluate. However, if viewed as a multivariate problem, where the product and processes are interpreted as the independent variables and the environmental exchanges as the dependent variables, it is possible to take advantage of the many statistical and mathematical techniques developed to gain insight into multivariate systems. In this paper, an expert system combining data retrieval and multivariate data analysis techniques is proposed to enhance database management. Through the use of exploratory pattern recognition techniques, such as correlation analysis and principal component analysis (PCA), the expert system will be able to alert a user to unexpected differences between inventories of the same or similar processes, to identify redundant entries, and to warn of possible errors in the data or mistakes in the data entry process. The expert system can be applied both to a single database and to a network of databases, thus providing a means for integration of diverse data sources.

Keywords: data, database management, expert system

1 INTRODUCTION

LCI databases are difficult to evaluate. Whilst significant steps have been made towards developing a consistent format in which to report and document LCI data (e.g. ISO 14041 technical standard [1]), tools to check for errors and inconsistencies in LCI data are generally not available. The fact that there is a considerable need for such tools is attested by the startlingly high degree of variability found across inventories of seemingly identical products [2; 3; 4]. For a particular product, a factor 10-100 (or higher) variation can be expected for process-specific emissions [4], whilst considerable inconsistencies are likely as to which life cycle stages and emission parameters are included [2]. The way in which the data are typically presented in LCI databases (large tables with very many entries) means that these large variations and inconsistencies are unlikely to readily identified by a user.

The lack of stringent requirements on the amount of meta-data that should be reported with LCI data means that many LCI databases are still published without adequate supporting documentation. In such cases it is very difficult, if not impossible, to make judgements on the quality of the data and its adequacy for use in a specific study. Even when LCI databases are well documented (which is all too often not the case) it is difficult to get a "feel" for data presented in vast tables with possibly hundreds of data entries. This disconnect between user and data is compounded by the fact that LCI data is most often reported summed and normalised, or otherwise aggregated to some high degree. This obscures the underlying information used to construct the inventory, leaving a user no option than to rely completely on the often scanty documentation included with the inventory to interpret its quality.

The volume of LCI data in circulation is steadily increasing, with a number of private and national database initiatives being undertaken around the world [5]. These are of varying degrees of quality and completeness, and often built upon existing data sources. The need for guidance in selecting the best data for a particular application is therefore becoming increasingly necessary. Furthermore, with the movement towards uniformity in format, and the existence of large data sets, it is increasingly feasible that an expert system could be developed to fulfill these needs. In particular, a system is envisaged that is able to contrast and compare inventories, thereby highlighting inconsistencies and possible errors in the data.

2 PROPOSED APPROACH

Vast tables of numbers may appear very difficult to interpret from an LCA practitioner viewpoint, but if viewed as a multivariate problem, where the products and processes are interpreted as the independent variables and the environmental exchanges as the dependent variables, it is possible to take advantage of the many statistical and mathematical techniques developed to gain insight into multivariate systems. These techniques have been developed precisely for such systems, where the high dimensionality of the data matrix (e.g. many variables measured over many processes) means that the data can no longer simply be interpreted "by eye" (e.g. in 2-D or 3-D graphical plots of the data). Instead, exploratory statistical methods aim to use the information content of the data to understand something of interest about the systems from which the data has been collected. In essence, the methods attempt to recognise patterns in the data which can provide useful information about the sample from which the data is collected [6].

The potential of using statistical methods to analyse LCI databases has been demonstrated by Huele and van den Berg (1998), who successfully apply simple correlation analyses to find identical and redundant entries in a large LCI database [7]. The use of the multivariate data analysis technique, Principal Component Analysis (PCA), has also been demonstrated in LCA studies, where it is shown to be a valuable aid in interpreting LCA results, capable of providing a unique visualisation of the structure of the results [8; 9].

These studies hint at the powerful insights that may be possible with more sophisticated multivariate data analysis techniques. In particular, the factor-based or

cluster analysis techniques, with their ability to recognise patterns within the data, have considerable potential to provide insights into LCI data that are at present not easily achievable, e.g. exposing actual from apparent differences between inventories by identifying inventories dominated by processes for which identical data sets have been used. The pattern recognition techniques are also able to alert users to possible errors and inconsistencies in the data sets by highlighting large or unexplained differences between inventories of the same or similar products (see following example). Given a sufficiently large number of varied and independent life cycle inventories, it is proposed that by using "supervised learning" tools, such as SIMCA (Soft Independent Modelling of Class Analogies), the insights gained through the analysis of those inventories for which sufficient meta-information is available, can be used to provide information on those inventories for which detailed documentation is lacking.

2.2 PCA example

Figure 1 presents an example of the sort of insights into large data sets able to be gained through the use of multivariate statistical methods. The example is only intended to provide a feel for the sort of results possible, and it is not within the scope of this paper to fully describe the method illustrated (principal component analysis). The theory of PCA can be found in most multivariate statistical analysis and chemometrics textbooks, e.g. [10; 6].

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 the factors found, rather than with 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 him/her to identify structures in the data that were previously obscured.

It is these structures or patterns in the data that are of interest in Figure 1. The distances between the points representing the various LCIs are what concern us, as these are used to define similarities and differences between the inventories. The axes do not have any physical meaning, they are merely measures of proximity that are interpreted as similarity. It is thus the relative distance between the points that is of note.

Figure 1 shows the results of an analysis of the LCIs of four different chemicals (ammonia, sodium hydroxide, ethylene and benzene) from four different data sources: ETH (Okoinventare fur Energiesysteme), APME (Ecoprofiles of the European Plastics and Polymer Industries, Reports 4, 6 and 14), Tenside (Petrochemical Intermediates, TS1, and Sulphur and Caustic Soda, TS3) and IDEA (An International Database for Ecoprofile Analysis) (data referenced as in the source, the LCA software model PEMS [11]).

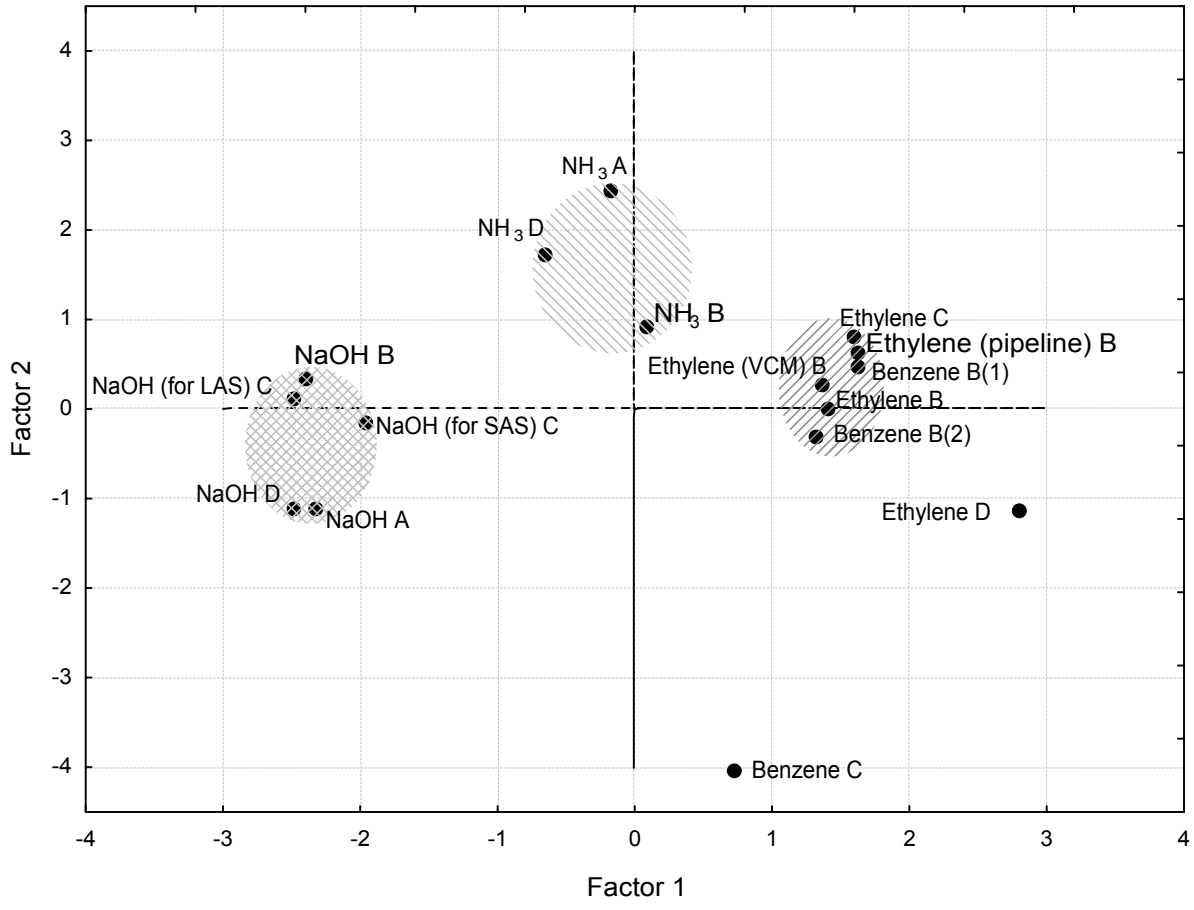

Figure 1. Principal component "scores" (data points projected on to the 1st and 2nd principal component axes) from an analysis of the inventories of four different chemicals using four different data sources. A: ETH, B: APME, C: Tenside and D: IDEA.

The four data sources have very different levels of completeness and little uniformity in format, so for this simple example, the analysis is based on only ten inventory items (see Figure 2). Even with this very reduced data set, three distinct clusters can be discerned in Figure 1, corresponding to the inventories for ammonia, sodium hydroxide and the organic chemicals (benzene and ethylene) (see hatched areas in Figure 1).

Immediately evident in Figure 1 is that Benzene from data source C shows characteristics significantly different to the inventories of the other organic compounds (it falls well outside the cluster of like inventories). To determine what is causing this significant deviation from the other inventories it is necessary to look at the principal component "loadings". These are a measure of each variable's contribution to the principal components, and indicate which variables are best at discriminating between the cases under investigation. In Figure 2, the length of the lines and their orientation indicate which variables have had the greatest influence in "pulling" the data apart to create the patterns in Figure 1. Interpreting Figure 1 and Figure 2 simultaneously (imagine the two overlain), indicates BOD to have been influential in pulling Benzene C away from the other inventories. A check with the data confirms this, and finds that the BOD measurements for this inventory are more than two orders of magnitude greater than the others. This is considerably larger than the variation shown by any of the other inventory items (which show at most a factor 5 variation), indicating that the BOD value may be an error. The PCA analysis finds where the greatest variations are occuring, not where the largest absolute changes in emissions occur. This is because it is based on correlations between the data points. The analysis is thus not influenced by the relative magnitude of the various inventory items, and can also include inventory items in different units (e.g. MJ and kg).

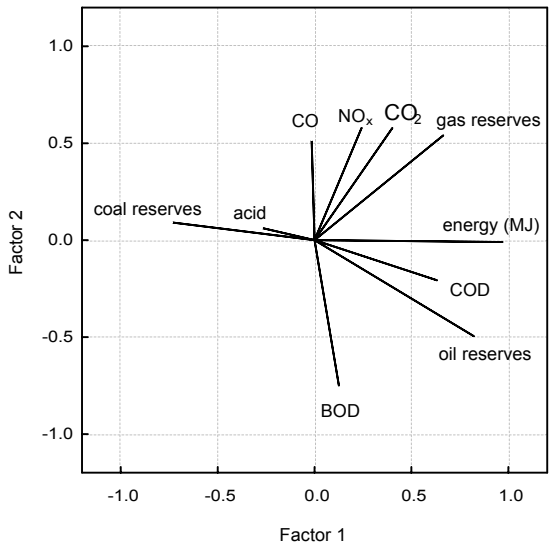

Figure 2. Principal component "loadings" (the contribution of each variable to the 1st and 2nd principal component axes) in the analysis of the inventories of four different chemicals using four different data sources (see Figure 1).

Clearly there is a need for an expert system to translate the results of the analysis to information useful to the user (i.e. a user is given a warning regards the anomalous BOD values without needing to understand PCA or to be able to interpret Figures 1 and 2). An expert system is necessary so that a user does not need to understand the intricacies of the statistical methods used to benefit from the insights they give into the data.

3 PRACTICE AND LIMITATIONS

The expert system will enhance traditional database search features by providing information regards the quality and completeness of the data. For example, a search for an inventory of a specific product/technology from a specific region would return the inventory for that region, but inform the user of gaps or significant differences between the inventory and others it contains for that or similar products. As the expert system is only making apparent what is hidden in the data, it is up to the user to determine whether to accept or reject the advice. For example, if a specific product inventory is seen to have significant differences between it and others in the database, this does not necessarily mean these differences are errors. A very large difference in a single inventory item might indicate a data entry error, however a pattern of differences would require closer examination. If indications are that all other inventories for this product have been built upon the same data (e.g. their dominant emissions are all from electricity production where identical data has been used), the inventory showing a different pattern of emissions may in fact be of higher quality. The expert system will present the differences with the possible underlying causes, but ultimately the course of action (e.g. to use another inventory, adjust the suspect inventory item, etc.) is up to the user.

It is important to note that the information the expert system provides can only be as good as the database upon which it rests. For example, where a user is interested in an inventory from a particular region, and the system shows large differences between this inventory and those from other regions, instead of merely alerting a user to these differences, the expert system would ideally be able to evaluate them in light of what it knows of regional variability. However this would only be possible if the database provided a number of inventories from different regions to allow for such an analysis. The success of the expert system therefore hinges on it being provided sufficiently large and well documented data sets in a common format. Whilst the expert system is able to inform a user about an inventory for which detailed documentation is lacking, this is only possible after the system has been "trained" and validated using well documented data.

The expert system is intended not to be limited to any one particular database, and ideally will be able to be used across a network of databases. However, a consistent format across the databases is a key requirement, since the expert system is only able to evaluate data in the same format as that it in which it has been "trained". Similarly, the criteria able to be used by the expert system to search a database are limited by the format of the database, and how complete and detailed the documentation is for the database. In other words, for the expert system to return and evaluate inventories closest to meeting a set of particular criteria, at a minimum, the inventories must have completed fields for those criteria (e.g. technology, region etc.). The more well specified the database, the better the expert system will be able to function.

4 CONCLUSIONS

Given a sufficiently large and well documented LCI database, it is proposed that exploratory mathematical and statistical techniques can be used to "unlock" some of the information in LCI databases obscured by layers of aggregation and normalisation. By using advanced statistical methods to explain unexpected variations, gaps and clustering of inventories, powerful insights into the data can be passed on to a user to assist with data selection. The expert system is not intended to choose the data for the user, but to help the user make a more informed decision about the best data to use in his/her particular study.

The expert system will combine conventional database search features with information on the quality and completeness of the data. In a specific search, the expert system will thus not only return the data that is closest to meeting a user's specific search criteria, but also evaluate the returned data in light of the other data in the database. The expert system thus allows a user to get an overview of the data not presently possible in LCI databases, and to get a feel for variations in emission values, levels of completeness, etc.

A prerequisite for the success of the expert system is a consistent data format and well specified data entries. The degree to which the expert system can interpret the variations it observes in the data rests on the availability of sufficiently large and well documented data sets with which to "train" and validate the system.

REFERENCES

1. ISO (2002): Environmental management - Life cycle assessment - Data documentation format. ISO/TS 14048:2002. Geneva, ISO
2. De Smet, B, Stalmans, M (1996): LCI Data and Data Quality: Thoughts and Considerations. Int J LCA **1**(2): 96-104.
3. Hanssen, O, Asbjørnsen, O (1996): Statistical Properties of Emission Data in Life Cycle Assessments. J Cleaner Prod **4**(3-4): 149-157.
4. Finnveden, G, Lindfors, L-G (1998): Data Quality of Life Cycle Inventory Data - Rules of Thumb. Int J LCA **3**(2): 65-66.
5. Norris, G, Notten, P (2002): Current Availability of LCI Databases in the World, UNEP/SETAC life cycle initiative, inventory programme.
6. Kramer, R (1998): Chemometric Techniques for Quantitative Analysis. New York, Marcel Dekker, Inc.
7. Huele, R, van den Berg, N (1998): Spy Plots: A Method for Visualising the Structure of LCA Data Bases. Int J LCA **3**(2): 114-118.
8. Le Téno, J-F (1999): Visual Data Analysis and Decision Support Methods for Non-Deterministic LCA. Int J LCA **1**(4): 41-47.
9. Notten, P, Petrie, J (2003): The Presentation and Analysis of Uncertain Results in LCA. paper under preparation.
10. Murtagh, F, Heck, A (1987): Multivariate Data Analysis. Dordrecht, Reidel Publishing Company.
11. PIRA (1996): PEMS (PIRA Environmental Management Software). Surrey, PIRA International.