Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Power Sources 122 (2003) 77–84
JOURNAL OF POWER SOURCES
www.elsevier.com/locate/jpowsour

Multivariate statistical methods in battery research

Paul Hagan*, Dymphna Fellowes

Department of Biological Sciences, University of Lincoln, Brayford Pool, Lincoln LN6 7TS, UK

Received 18 July 2002; received in revised form 20 January 2003; accepted 10 February 2003

Abstract

The use of multivariate statistical methods in battery research is developed with examples drawn from the literature and unpublished work by the authors.

The techniques discussed may be described in general as, data reduction, cluster analysis and regression methods for prediction. Individually or collectively these represent the three main areas of interest to battery researchers.

Data reduction permits the visualization of the relationship between samples which are characterized by multiple measured variables. Cluster analysis extends this process to examine any natural groupings existing in the samples, based on the variables measured, and multivariate prediction is a calibration technique permitting the modelling of complex non-linear systems.

Keywords: Statistics; Cluster analysis; Principal component analysis; Battery; Regression; Prediction

1. Introduction

As we are all aware it is a relatively straightforward procedure to represent the relationship between two variables, such as capacity and time for a cell, on a scatterplot in two dimensions.

As an alternative, this relationship can be represented by vectors in two dimensions, where the cosine of the angle between the vectors expresses the correlation between the variables. For a perfect correlation between the two variables the vectors have an angle of 0° between them with a cosine of one. For two uncorrelated variables the angle between the vectors is 90° and the cosine is zero, these vectors are said to be orthogonal. With a third variable, say temperature, we could present the scatterplot as a 3D projection in two dimensions. However, for four or more variables, there is no direct way of showing the relationship.

We could, of course, examine the scatterplots for all pairs of data as a simple way of looking at the data. For a number of variables and a number of cells this would rapidly become very confusing. For one cell with seven measured variables we would need to examine 21 scatterplots.

An examination of such plots can be misleading as only examining the variables in pairs could easily obscure any structure present in the original multi-dimensional space.

The solution lies in a variety of techniques known as mapping or ordination techniques. These techniques seek to represent the original multi-dimensional space in a reduced number of dimensions while, as far as possible, retaining the original structure. These techniques are used to produce a two-dimensional map of the original multi-dimensional space. It is possible to represent the original multi-dimensional space in three dimensions, but until quite recently this has been uncommon.

Consider the data of Fig. 1. FeS_2 (pyrite) has been proposed as a cathode material in lithium secondary battery systems [1]. Pyrite is an abundant mineral and its commercial extraction usually presents little serious mining problems. The quality of pyrite can vary dramatically among sources and, just as likely, between different lots from the same source. This is, in part, due to the variable genesis of the mineral, even between geologically related sites. Differences in trace element composition may vary spectacularly between pyrite deposits from geologically related sites separated by their geographical age. The major advantage of natural pyrite as a cathode material is its low cost. The reproducibility of the cathode behaviour is improved with pyrite of a uniform physical size and chemical composition.

The data in Fig. 1 is plotted in three-dimensional space. For n variables n-dimensional space would be required. The samples are scattered in this space. By a manner analogous to linear regression in n-dimensions a vector is found which describes as much of the difference (variance) between the samples, as possible. This is called the first

* Corresponding author.
E-mail address: phagan@lincoln.ac.uk (P. Hagan).

0378-7753/03/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0378-7753(03)00344-6

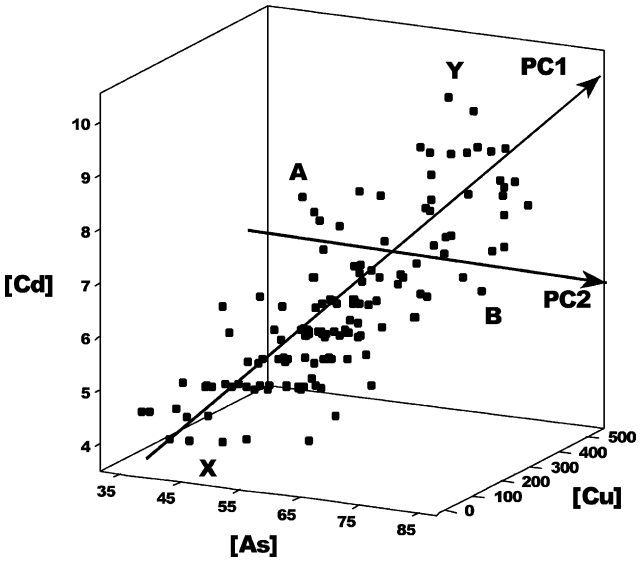

Fig. 1. Representation of trace element impurities in a naturally occurring FeS_2 cathode material. Data points (■) are plotted for the copper, arsenic and cadmium concentrations of pyrite samples from different locations.

principal component or PC1. Samples X and Y differ most along PC1. Each sample can then be projected onto PC1 to identify a co-ordinate along this axis. PC1 is not usually sufficient to describe all the variations between the samples. A second principal component, PC2, is required to describe the differences between samples A and B. There are two constraints on PC2:

(a) PC2 should be perpendicular to PC1.
(b) PC2 should explain as much as possible of the remaining differences.

Further principal components may be constructed within the constraints that they should all be orthogonal and explain as much as possible of the remaining variance. Theoretically there are as many principal components as variables. The aim of principal component analysis (PCA) is to explain as much of the data structure as possible with as few principal components as possible. The construction of principal components removes the covariance between variables, and reduces the number of variables needed to model any response. If the analysis can explain the differences between samples in two or three principal components we have succeeded in data reduction and this may be considered to be a major attraction of PCA.

The main disadvantage of PC's is that they have no direct physical meaning. This will, of course, make visualisation and interpretation of score plots difficult, i.e. we can see if samples differ from each other but not why.

Loadings plots are the link between the measured variables and principal component space. In constructing a score plot the position of the samples relative to each other is presented in principal component space rather then in the original variable space. Loadings plots are essentially projections of the unit vectors of the original variable space onto principal component space. Variables which appear close together in the loadings plot will be highly correlated whereas variables which appear at opposite ends of the origin in a loadings plot will be negatively correlated. There is no correlation between variables which are well separated in a loadings plot. Like score plots, loadings plots are normally presented as two-dimensional plots which represent the analyst's window into PC space. It is not immediately obvious that points which appear close together in a two-dimensional plot may be well separated along a third dimension.

2. Principal component analysis

To help clarify these points we will employ an unusual, but highly relevant example, drawn from unpublished work by the authors, on the identification of possible new cathode materials drawn from naturally occurring minerals.

Our data set will comprise 137 naturally occurring minerals with the following input variables:

1. relative molecular mass, RMM (kg);
2. entropy of formation, entropy ($J\,K^{-1}\,mol^{-1}$);
3. molar volume (m^3);
4. enthalpy of formation, ΔH ($J\,mol^{-1}$);
5. free energy of formation, ΔG ($J\,mol^{-1}$);
6. 2θ values for the three main XRD peaks;
7. total oxidation numbers of the cationic elements (all summed as positive values).

An examination of the score plot for PC1 and PC2 in Fig. 2 shows that certain minerals and compounds already well known as cathode materials in a variety of battery systems may be found grouped on the right hand side of the plot. This may have been influenced by including the free energy

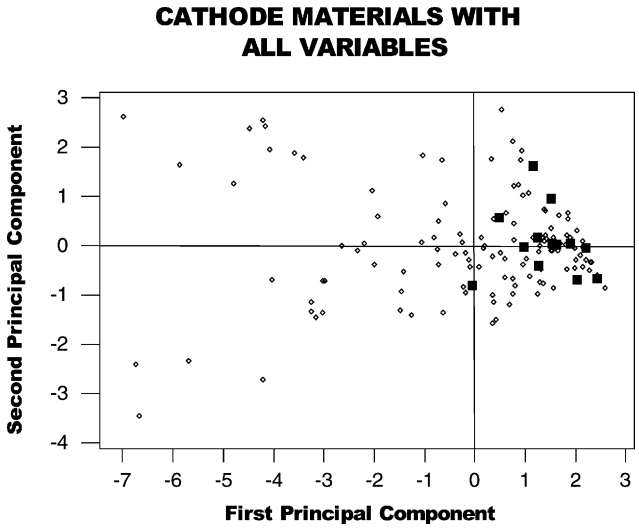

Fig. 2. Scores plot of PC1 vs. PC2 for proposed cathode materials. Recognised cathode materials are identified as solid squares (■).

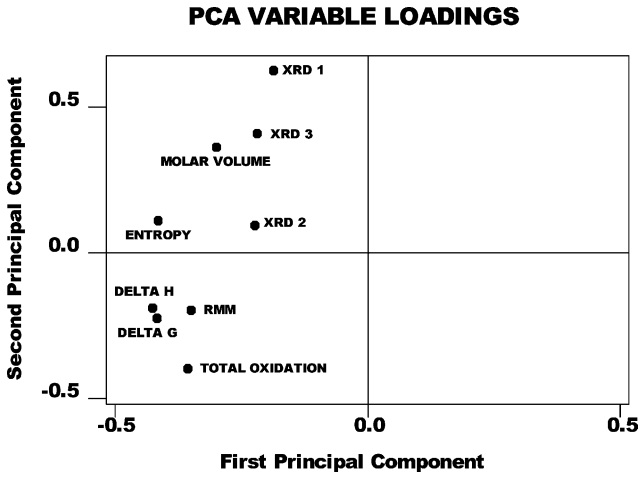

Fig. 3. Loadings plot for PC1 and PC2.

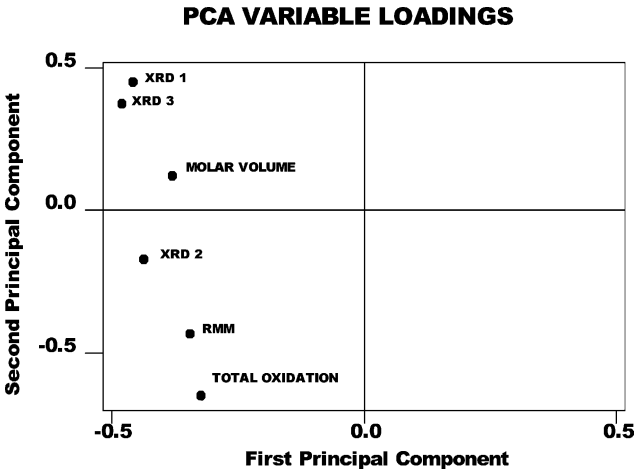

Fig. 5. Loadings plot for PC1 and PC2.

of formation in the analysis. The loadings plot in Fig. 3 shows, as to be expected, that the free energy of formation and the enthalpy of formation are highly correlated (hardly surprising in view of the thermodynamic function which relates them). In fact all the variables have negative loadings along PC1 and they all serve to separate the minerals along PC1. The spread of samples along PC2 is explained by the positive and negative loadings of the variables above and below the origin.

If the analysis is repeated without the three thermodynamic variables a similar score plot is generated in Fig. 4. The loadings plot shown in Fig. 5 shows that the variable 'XRD 3' has the most negative loading along PC1 followed closely by the remaining variables. The loadings of the variables along PC2 are greatest for 'XRD peak 1' (positive loading) and for 'total oxidation' (negative loading).

We have established that the scores plot for PC1 and PC2 in Fig. 2 may be useful in identifying possible cathode materials based on a set of variables which are readily available from published data sets. Because of the expected effect including the thermodynamic variables might have on the groupings in the scores plot it is reasonable to re-analyse the data and perform another PCA with the thermodynamic variables eliminated from the data set.

It is apparent from an examination of Figs. 2 and 4 that excluding the thermodynamic variables from the analysis has had little effect on the grouping of the cathode materials in the scores plot. The next stage in a search for viable new cathode materials would be an identification of the other minerals situated close to the identified cathode materials in the score plot and an evaluation of their performance in prototype cells.

A recent application of PCA concerns the characterization of expanders for lead–acid batteries [2] using 128 measured variables including chemical composition and physical properties.

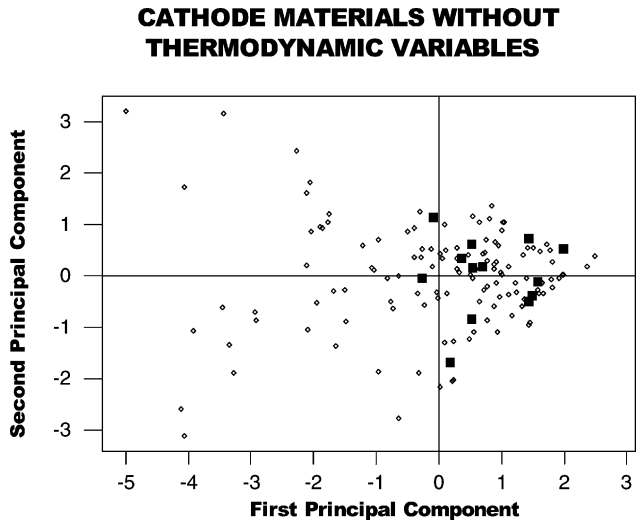

Fig. 4. Scores plot of PC1 vs. PC2 for proposed cathode materials. Recognised cathode materials are identified as solid squares (■).

3. Cluster analysis

The basic principle upon which all clusters analyses are based is very simple. All of them attempt to group samples or objects into groups of similar objects called clusters. Objects are placed into different clusters such that members of any cluster are more similar to each other in some way than they are to members of any other cluster.

The major problem associated with cluster analysis is that the techniques always produce clusters even in circumstances where there are no natural groupings in the data. The analysis actually imposes a cluster structure on the data. The success of the method will depend entirely on knowing whether the clusters produced are real ones or simply artefacts of the method.

An example, drawn from unpublished work by the authors, on an investigation into trace element concentrations in pyrite and their possible effect on battery performance

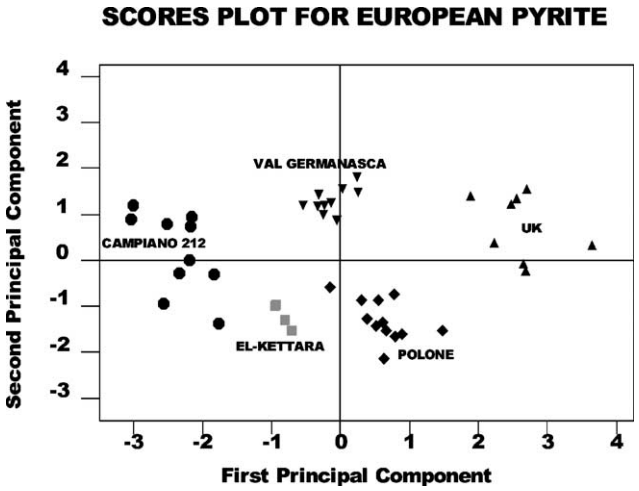

Fig. 6. Scores plot for European pyrite.

will be valuable in a consideration of cluster analysis. It has been reported previously [1] that the quality of pyrite samples is wholly dependent on their geological age and location. A large number of pyrite samples from locations all over Western Europe were analysed for trace impurities by the authors.

The scores plot in Fig. 6 shows that pyrite from different locations, and of different age may be grouped based on a PCA analysis with concentrations of trace elements as variables. There are a number of well defined clusters which, when the points in the score plot are identified by different symbols, are quite apparent. It is questionable if some of the clusters would have been so readily identified if all samples had the same symbol in the score plot.

It is usual to be rather more quantitative when analysing samples for clustering. The most straightforward method employs hierarchical agglomerative methods in which members are merged using a single-linkage rule. The dendrogram in Fig. 7 was produced using a single linkage rule (sometimes referred to as the nearest neighbour rule). In the dendrogram, at the lowest level, all the items being clustered are independent and at the highest level all are joined into one group. If n items are to be clustered, all agglomerative methods require $n - 1$ steps to complete the clustering.

All hierarchical agglomerative methods produce non-overlapping clusters which are nested; that is, each cluster is included or subsumed in large clusters at higher levels of similarity. This is clearly illustrated in Fig. 7. where a single cluster (all samples in the data set) is produced at the 41.28 level and between the 80.43 and 60.85 levels there are a number of clusters identified which may be compared with the easily identified clusters in Fig. 6.

Only a single pyrite sample (marked (◆) on the dendrogram) from the group ‘Campiano 212’ has been incorrectly identified as a member of the ‘El Kettara’ group. In fact, by simple visual inspection, the sample is less representative of the Campiano 212’ group than it is of the ‘El Kettara’ group.

One of the main problems in the validity of cluster analysis is, whether the clusters ‘identified’ by the method represent natural groups or whether the clusters are merely *produced* (rather than identified) as a consequence of the clustering rules used. A cluster solution should be. viewed with caution and great care should be exercised before claiming the discovery of any natural clusters.

4. Prediction

Partial least squares (PLS), alternatively known as projection of latent structures, is a powerful multivariate statistical linear regression technique which extracts the relationship between an array of output variables and an array of input

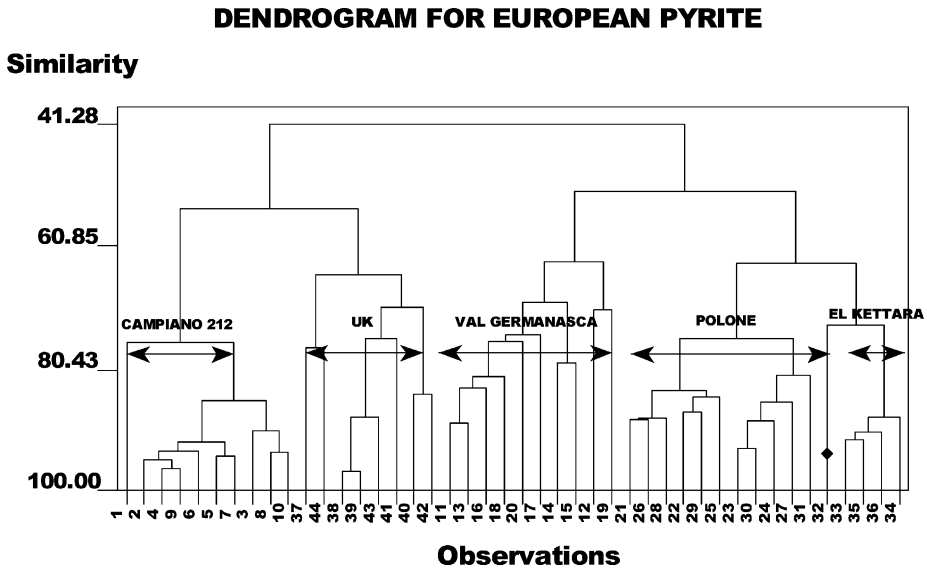

Fig. 7. Dendrogram for European pyrite.

variables, where a high degree of correlation exists between the output and input variables. PLS differs from principal component regression (PCR) in that the reduction in the dimensionality of the raw data is based on both the input (henceforth referred to as the *X* matrix) as well as the output matrix (henceforth referred to as the *Y* matrix) (for PLS) and not just for the input data (for PCR). As a result, the main principal components in PLS cannot be arrived at in one singular value decomposition (SVD) step as can be done in PCR. Instead the principal component that lines up most of the deflated *X* and *Y* data matrices is then extracted. The cycle repeats itself until enough PLS principal components have been extracted.

A valuable example of using PLS in power sources work is the prediction of 'state of charge' (SOC) from measurements of the electrochemical impedance spectra (matrix *X*) and state of charge (matrix *Y*) [3] for nickel–metal hydride batteries.

Results from this method, using a cross validation procedure to test the predictive capability of the method suggest a root mean square error of prediction (RMSEP) of 7%. Unfortunately the predictive power of the model decreases at SOC values of less than 10%.

Fuzzy logic methodology has been used to determine the 'state of health' of primary lithium/sulfur dioxide cells [4] from impedance data. Other techniques of value in predicting failure in secondary batteries are the use of artificial neural networks [5], which are inspired by the biological behaviour of neurones.

The power of both these latter prediction methods lies in their ability to model complex non-linear systems without the need for explicit mathematical models [6].

Multivariate methods have been applied to the lifetime and performance prediction of lead–acid [7–9] and Ni–Cd [10] batteries. New metal hydrides have been designed for NiMH batteries using PLS-pattern recognition methods [11].

All of these methods are extremely powerful and provide a link between multivariate data, the impedance spectrum or cycling data, and SOC or cycle life respectively.

An example of SOC prediction using impedance spectra, from unpublished work by the authors, with a reduced data set for clarity will serve to illustrate the method for a small number of samples and variables.

The prediction problem concerns the estimation of the relationship between a measured *X* and a property *Y* and the use of this relationship to estimate an unknown *Y* from an observed *X*. At no point in the calibration or prediction procedure does the model require the input of an explicit mathematical relationship between *X* and *Y*. The multivariate character appears when the impedance is measured at many different frequencies jointly and the overall impedance spectrum corresponds to the measured *X*. The power of a PLS prediction model lies in its ability to predict, with acceptable confidence, the value *Y* when it has a non-linear relationship with *X*. If the aim were to predict the SOC from batteries or cells stored or discharged at alternate temperatures then PLS models would need to be constructed from a calibration set of batteries stored or discharged under a similar temperature regime. The effects of ageing and abuse have not been investigated in this communication but work is proceeding, in the authors laboratory, to determine the effects of variable degrees of abuse and variable age on the predictive power of PLS models. In practice, many more samples would be required for construction of a statistically robust model but limiting this example to six samples and one test cell allows for a more easily viewed data-set. The measurements were performed on six standard 2400 mAh NiMH cells using a Solartron 1287 electrochemical interface coupled to a Solartron 1255 Frequency Response Analyser for electrochemical impedance spectroscopy. Following initial cycling at 0.25 C (based on the nominal capacity) the impedance measurements were recorded at different SOC's for a discharge load of 0.25 C.

The frequency range used for the data analysis was between 0.5 and 250 Hz whereas the impedance measurements were measured over a much wider range. Impedance measurements were made with an ac signal of amplitude 100 mA in galvanostatic mode.

Lower frequency measurements were avoided as they are slow and the SOC may drop appreciably during the scan.

Additionally for rapid 'in service' determinations faster higher frequency measurements are desirable. SOC's at the conclusion of each impedance measurement were calculated from the total time of discharge. Changes in the SOC of $<0.75\%$ were avoided by omitting low frequency measurements at the discharge load of 0.25 C. Fig. 8 demonstrates that there is an obvious difference in the impedance responses, for the real part of the impedance, at different SOC's for a discharge load of 0.25 C.

In the following illustrative example of the PLS regression and prediction method it is important to note that no information regarding cell voltage or discharge load have been used in formulating the multivariate model.

In most cases knowledge of the cell potential and discharge current for a galvanostatic discharge would be sufficient to determine the SOC. In practice this would be difficult for a flat discharge profile and for a cell with an unknown history. Our goal will be to predict the SOC from an impedance spectrum and to further validate the predictive power of the model using a test spectrum, which was not used in the construction of the original model. The method used to construct the regression equation is known as full cross validation. The frequency spectrum for a single battery is automatically removed from the data set and the SOC for that battery is predicted using a model constructed from the remaining batteries. The sample is removed to ensure that the measured SOC does not influence the prediction. An iterative procedure then repeats this procedure for all the objects in the data set. The result of this modelling is presented in Fig. 9. Essentially this modelling by a 'take one out' procedure produces a best fit for all the samples in the calibration set. The root mean square error, between the measured and

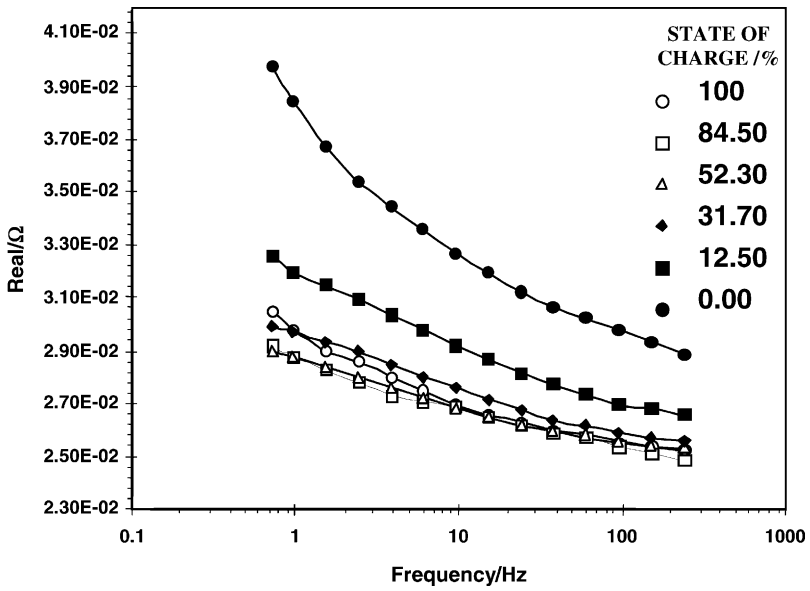

Fig. 8. The real part of the impedance for different states of charge measured at 0.25 C.

predicted values indicates that this model predicts the SOC of the cell, without SOC and discharge load input parameters, with an average error of 1.5%. The broken line in Fig. 9 indicates the situation if the predicted value equalled the measured value while the solid line is a linear least squares fit for the 'predicted versus measured' SOC values. In this particular example SOC of less than 10% were poorly predicted by the PLS model and, consequently were omitted from the modelling and prediction. It may be that the poor prediction of SOC towards the end of discharge is an indicator of extreme non-linear behaviour. An alternative model may be possible for acceptable predictions of SOC less than 10%.

This RMSEP value of 1.5% gives an indication of the predictive power of the model with a self-contained data set which is implicit in any method involving cross validation. No model can be considered of value in the sense of prediction until it has been validated with a test set of batteries which were not used in producing the calibration model. In order to test the prediction power of the model our test set will comprise a single cell of known SOC which will allow

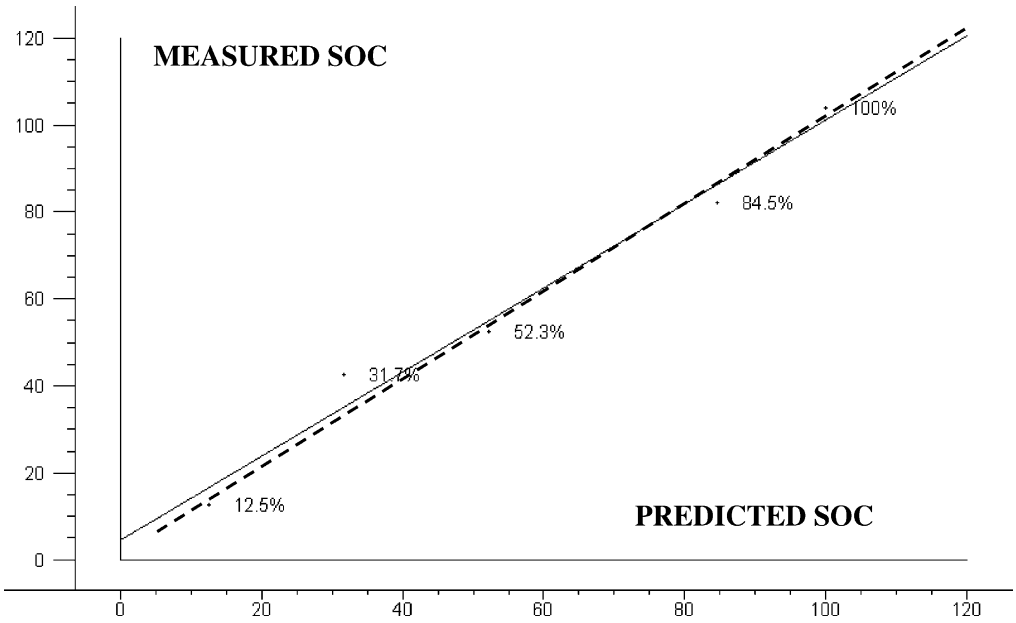

Fig. 9. Full cross-validated prediction for SOC with the complete data set.

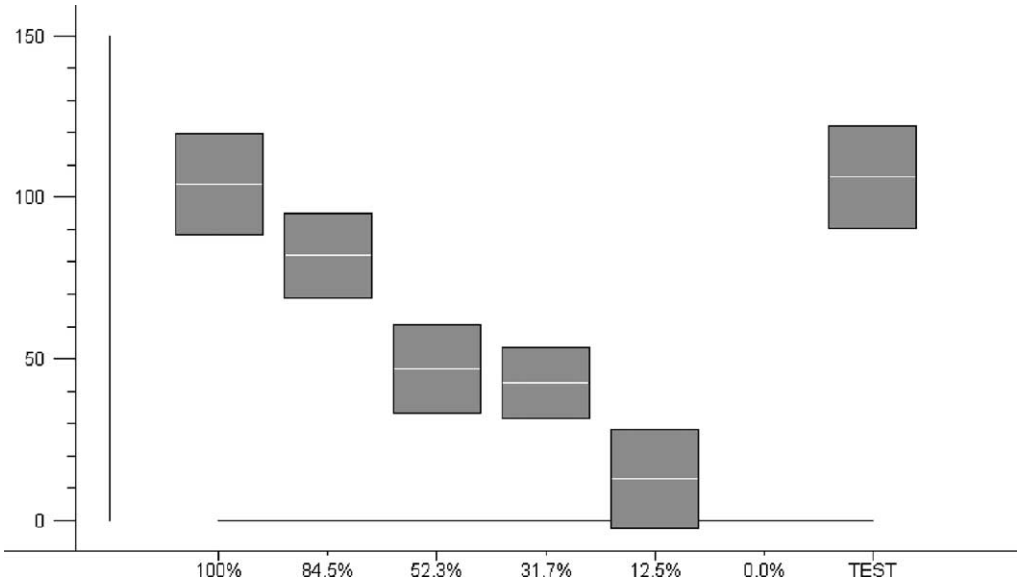

Fig. 10. Prediction of the SOC of the data set cells with a test cell of known SOC discharged at the 0.25 C rate.

Table 1
Predicted state of charge and standard deviations from the NiMH PLS model

Sample (%)	Predicted SOC (%)	Deviation SOC (%)
100.00	104.03	15.56
84.50	82.08	13.12
52.30	46.95	13.67
31.70	42.71	11.01
12.50	12.77	15.35
Test	106.39	15.83

us to compare the predicted SOC with the known value. The test cell (Table 1) had a known SOC of 93.5% which clearly lies within the limits of the prediction (SOC (predicted) = $106 \pm 15.8\%$). This result shown in Fig. 10 is quite satisfactory in view of the very small data set used to construct the model.

5. Conclusions

There is valuable information hidden in your current or historical data and multivariate analysis may be the key, which unlocks it. The use of multivariate techniques is expanding rapidly in a very large number of areas in the research and development, manufacturing and marketing sectors and it would seem appropriate that there is scope for a more widespread use of the armoury of data reduction and prediction techniques which are widely available through many commercially available software packages.

The three techniques most widely used in multivariate analysis

(i) principal component analysis,
(ii) cluster analysis,
(iii) partial least squares prediction,

have all been shown to have application in the research and development of non-mechanical electrical power sources. Principal component analysis may be a valuable tool in examining the viability of existing naturally occurring materials as potential electrode materials. Principal component analysis has also been used to characterize expander materials employed in the plates of lead–acid batteries. Cluster analysis has been demonstrated as a means by which natural pyrite may be classified into groups, characterized by geological age and location, based on analyses of trace metal impurity concentrations. Partial least squares prediction has been successfully used by a number of investigators to predict the state of charge, of a variety of battery systems, from their impedance spectra.

There is no fixed path for those who venture into multivariate work but rich rewards await those who look in the right place for results. Decisions have to be made continuously, and they have to be made with good judgement and usually on electrochemical or battery grounds.

The first decision relates to the variables to be studied, and here may lie the first problem. If a variable is omitted which ought to be included an important vector may be trivialised to the point were it appears arbitrary or unintelligible when, in its full form, it has a ready interpretation. Conversely if a variable, which adds little to the investigation, is included, but is nevertheless highly correlated with the other variables, a clear result may be obscured.

A final cautionary remark that should be part of the awareness of anyone employing multivariate analysis in their work is an appropriate conclusion to this examination of multivariate methods in battery research.

The hypotheses evolved by multivariate methods are like any other and should only be accepted if they can be co-ordinated with other knowledge and can be confirmed by experimental evidence.

References

[1] E. Strauss, G. Ardel, V. Livshits, L. Burstein, D. Golodnitsky, E. Peled, J. Power Sources 88 (2000) 206–218.
[2] B.O. Myrvold, D. Pavlov, J. Power Sources 85 (2000) 92–101.
[3] K. Bundy, M. Karlson, G. Lindbergh, A. Lundquist, J. Power Sources 72 (1998) 118–125.
[4] A.J. Salkind, C. Fennie, P. Singh, T. Atwater, D.E. Reisner, J. Power Sources 80 (1999) 293–300.
[5] M. Urquidi-Macdonald, N.A. Bomberger, J. Power Sources 74 (1998) 87–98.
[6] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[7] S.P. Perone, R. Petesch, P. Chen, W.C. Spindler, S.L. Deshpande, J. Power Sources 37 (1992) 379.
[8] S.P. Perone, J. Power Sources 13 (1984) 23.
[9] A.L. de Azevedo, F.B. Diniz, B.B. Neto, J. Power Sources 52 (1994) 87.
[10] W.A. Byers, S.P. Perone, J. Electrochem. Soc. 126 (1979) 720.
[11] H. Liu, J. Guo, N. Chen, T. Huang, Anal. Lett. 29 (1966) 341.