A Multivariate Approach to Utilizing Mid-sequence Process Control Data

Rhett Evans[1,3] and Matthew Boreland[2]

[1]Australian Centre for Advanced Photovoltaics, UNSW, Sydney, NSW, 2052, Australia
[2]School of Photovoltaic and Renewable Energy Engineering, UNSW, Sydney, NSW, 2052, Australia
[3]Solinno Pty Ltd, West Ryde, NSW, 1685, Australia

***Abstract*—While the measurement of cell efficiency is still considered one of the primary assessments of a cell's quality, a modern photovoltaic manufacturing facility will also include a range of metrology to assess the performance at various steps during the manufacturing sequence. These measurements can be used to control individual processes and ensure reliable process interactions, but they are at their most powerful where they can be correlated to the final performance of the cell. Such a relationship is not always easy to establish, particularly when the data collected during the process cannot be parametrized entirely with a single variable. This paper shows how two multivariate approaches can be used to form a relationship between cell lifetime data collected early in a fabrication sequence, and the final cell V_{oc}. While building a model with a high level of predictive accuracy is rarely feasible, it is possible to identify a higher proportion of under-performing product and provide insight into how material type interacts with the manufacturing sequence.**

***Index Terms*—multivariate statistics, manufacturing, photovoltaic cells, process control**

I. Introduction

Many manufacturing sectors use multivariate techniques to improve process outcomes, [1], [2], but to date there is only a limited amount of work in this area published in the field of PV manufacturing [3]–[5]. With the trend towards increased levels of metrology, multivariate techniques can extract a useful and interpretable amount of information from high dimensional data sets. This study examines lifetime data collected early in a solar cell fabrication sequence (mid-sequence metrology) as well as the final performance of those cells (end-of-sequence performance) after fabrication is complete. The lifetime data provides some information on variance up until that point in the fabrication sequence, including information on input material variance, but it is more useful if a relationship can be established to end of sequence performance. One benefit of this is the ability to set a control limit based on mid-sequence metrology and remove product at mid-sequence that is expected to be under-performing by the end of sequence. There is also an opportunity to better understand the influence of downstream processing on final cell performance, and even to detect where problems occur with the measurements themselves. A range of multivariate techniques can be used to establish these relationships and the aim of this work is to introduce two of these through example.

II. Multivariate Statistical Techniques

Most manufacturing industries have in common that the assessment of the quality of the manufactured product cannot be or is not precisely measured. A common use of multivariate statistical techniques is therefore to build relationships between process data and product quality [6]–[8]. For PV cell manufacturing, the situation is a little different. Cell power is the metric of most interest in assessing product quality, and this is simple to measure accurately. Nonetheless, many applications for multivariate techniques still exist, and in this study they are used to express the relationship between mid-sequence metrology and end of sequence performance. The techniques themselves can be divided into two broad groups. Methods such as Principal Component Analysis (PCA) [9], Factor Analysis (FA) [10] and Hotelling statistics [11] are fundamental multivariate techniques used to "summarise" high dimensional data sets and aid in the interpretation of large amounts of data. Another group of techniques map "input" and "output" data sets to each other. A simple example of this is multivariate linear regression [12], but more advanced techniques include Artificial Neural Networks (ANN) [8] and Partial Least Square (PLS) [13]. PCA and PLS often have an advantage in that the structure in the relationships can be better understood from the models [14]. One technique from each of these groups is further examined here for their utility in establishing a relationship between lifetime parameters mid-sequence, and V_{oc} measured at the end of the sequence.

A. Principal Component Analysis

Other resources more fully explain the nuances of the PCA technique [9], but some brief explanations are given here to aid understanding. In the simple case of two dimensional data, PCA is effectively a rotation of the potentially correlated axes of the data set to a new set of orthogonal (i.e. uncorrelated) axes. These new axis are aligned to the direction of maximum variation in the original data set [15]. This is shown visually in figure 1

PCA of high dimensional data sets does the same thing - it rotates the axis to a new set of orthogonal axis. It is much hard to visualize this in high-dimensional spaces. In this more general case, PCA is best explained using linear algebra as summarised below;

- An $m \times n$ data matrix, A, is compiled from m wafers, with n pieces of data for each wafer.
- The data should be mean-centred and normalised in some way so all the data columns are of comparable magnitudes. The details of the normalization used in the present study are explained later.

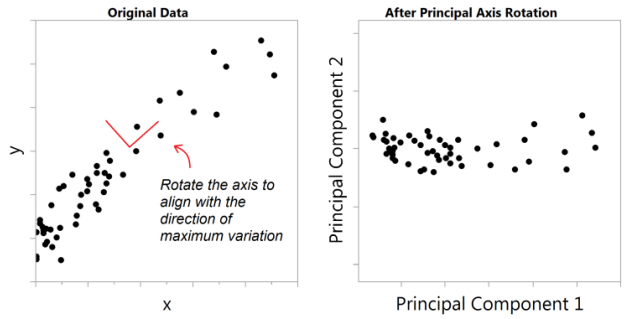

Figure 1: Principal Component Analysis applied to data in two dimensions is effectively a rotation of the axis to align with the directions of maximum variation

- Calculate the n eigenvectors (u_1 through to u_n) of the $n \times n$ covariance matrix $A^T A$ derived from the $m \times n$ data matrix A.
- The first principal component, sometimes called the list of principle component scores, is given by an $m \times 1$ column vector t_1 which is calculated by projecting the data matrix A onto the first eigenvector u_1(as summarised in equation 1). The second principle components is calculated by projecting A onto the second eigenvector and so on.
- The nature of the eigenvectors means that each of the principal components are orthogonal (i.e uncorrelated). It also means that in many situation, a large amount of the variance in the original data matrix is captured in just this first principal component, or the first couple of principal components. This results in the dimensionality of the data set being effectively reduced.

$$\begin{bmatrix} \\ t_1 \\ \\ \end{bmatrix} = \begin{bmatrix} \\ A_{m \times n} \\ \\ \end{bmatrix} \times \begin{bmatrix} \quad u_1 \quad \end{bmatrix} \tag{1}$$

B. Artificial Neural Networks

ANN is a statistical tool that can be trained with large amounts of data to generate relationships between a set of input data and a set of output data. Other resources more fully explain these techniques [8]. In a contemporary context ANN models can be simply generated with a range of common statistical computing packages. Their success is very much determined by the size and quality of the data, but they essentially act like a "black box" models in that no particular understanding about the relationship can be derived from the model. In the case of solar cells where there are many good known physical models to describe their behaviour, this may not be the optimal strategy in all situations. But the models are relatively simple to build and assess, and notwithstanding their disadvantages often represent a "best case scenario" chance for identifying a statistical relationship between any two sets of variables. It is typical practice within any given data set to train the model with some proportion of the data, and assess its predictive capability with the other proportion of the data. If the model is used on an ongoing basis, it should be regularly checked or retrained.

III. Data

The data used in this study comes from a set of 800 cells fabricated in a pilot manufacturing facility. The lifetime data is collected early in the fabrication sequence and measured on a photoconductance decay lifetime tester. The lifetime data is parametrized with seven parameters as described in table I. The J_0 and bulk recombination fitting is done using the Kane-Swanson method [16]. The final V_{oc} data comes from the final measurements of the cells on a flash tester.

Table I: Parametrization of lifetime data. All parameters x_1 through to x_7 are normalized

Parameter	Description
x_1	Implied V_{oc} (iV_{oc}) at one sun
x_2	Sheet Resistance
x_3	Bulk Recombination Fit Parameter
x_4	Surface J_0
x_5	Fit parameter
x_6	Minority Carrier Density @ 1sun / 1E15
x_7	Lifetime at 0.1 suns

The data is normalized to zero mean as is customary in multivariate analysis [11]. Normalization of data is commonly done to protect its commercial sensitivity, but this also acts as a way to place investigative focus onto the statistical analysis technique, rather than the data itself. It is important that this becomes a common and accepted technique within the photovoltaic field to encourage the sharing of improved statistical methodologies. Adjusting the relative variance of the data set is also a crucial tool in multivariate analysis [11]. In the case of the data here, the variance is normalised to a target process variance such that "out-of-range" values are simply shown as being more than three standard deviations from the mean.

IV. Results and Analysis

The statistical package JMP11 was used to build the statistical models described in this section.

A. Single variate relationships

The relationship between normalized V_{oc} (here referred to as nV_{oc}) and parameter x_1 (the normalized iV_{oc}) is shown in figure 2. The data has been normalised to a target standard deviation, the aim being for all the data to fall between ± 3 standard deviations. The "-3" lines are marked for reference on the graph for both quantities. This divides the data in the figure into four quadrants.

Figure 2 shows the difficulty of using a single variate approach - i.e the value of iV_{oc} only - in setting a lower limit for material going through the rest of the process sequence. Based on final V_{oc} testing, quadrant 1 and 2 are pass cells

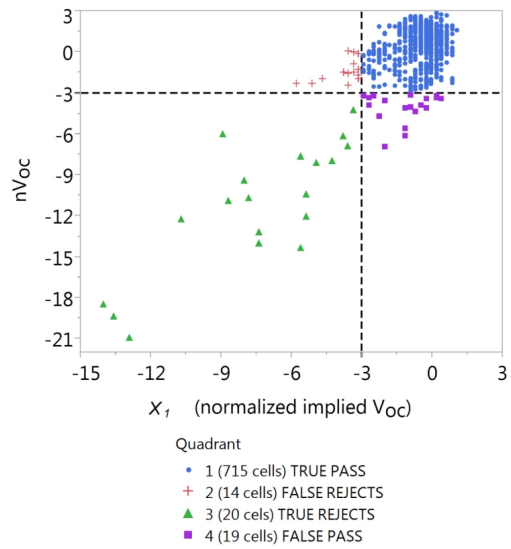

Figure 2: Normalized V_{oc} vs x_1 (Normalized iV_{oc}), divided into four quadrants as defined by the target limits of -3 standard deviations. The classification of TRUE PASS, FALSE REJECT etc is based on how the cells would have been classified based on the lifetime measurement given what is known about the final performance.

and quadrant 3 and 4 are reject cells. If an attempt was made to use only mid-sequence iV_{oc} as a pass/reject criteria, quadrant 2 cells are false rejects and quadrant 4 cells are false passes. All of the lifetime parameters x_1 through to x_7 can be used in a multivariate model to attempt to correctly identify the cells from the four quadrants at the point where the lifetime measurement is done. The success or otherwise of this will also then provide clues about why the quadrant 2 (false reject) samples apparently improved during processing and why the quadrant 4 samples (false passes) apparently got worse. Many methods can be used to develop these multivariate relationships depending on the exact aim and the quality and size of the data sets. In this study, the PCA and ANN approaches are compared.

B. Principal Component Analysis (PCA)

A principal component analysis was done on the data matrix containing the seven lifetime parameters for each of the nearly 800 cells, scaled as per the normalization already described. In some cases it is useful to regress output variables such as final performance against the principle components to uncover correlations. The components themselves are orthogonal and ideal for regression analysis. In some cases it is useful to plot the principal components against each other to show how data with similar characteristics will cluster together in the plot. The latter approach is taken here. The first two principle components are shown plotted in figure 3. The data is once again split by the quadrant location shown in figure 2. The majority of the data from quadrant 3 (true rejects) is now clearly distinct from the rest of the data. The quadrant 2 (false rejects) data is also somewhat distinct from the main cluster of quadrant 1 (true pass) data. The quadrant 4 (false pass) data is not at all distinct from the main data-set.

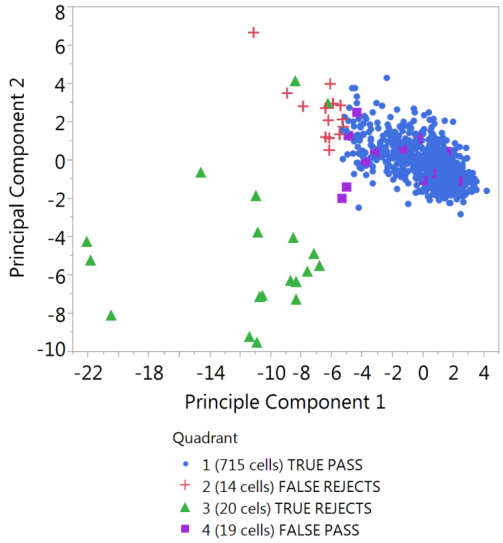

Figure 3: The first two principle components derived from the covariance matrix of the data from the seven lifetime parameters. The cells are all marked as per their quadrant in the original graph of figure 2

These types of relationships can be used in practice as a type of spacial control chart. As cells are processed, their principal components can be calculated and plotted on the graph to ensure that they fall within a target "cluster". It is also possible to look at the the data in three dimensions using the first three principal components or some other combination of principal components. For some data sets, it is sometimes possible to get more distinction between the groups of data in three dimensions. In this case, most of the distinction can already be seen with just the first two principal components. Part of the reason for this can be seen in the scree plot of figure 4. A scree plot shows the value of the eigenvalue associated with each eigenvector. The eigenvalue describes the relative amount of variance explained by each principal component. Figure 4 shows that the first two principal components plotted in figure 3 explain nearly 80% of the overall variance. In this case, plotting figure 3 with a third dimension offers little further discrimination between the groups of data.

The data plotted in figure 3 can also be shown as a bi-plot, as in figure 5. A bi-plot includes the projection of the original parameter axis onto the principal component plane. This can be very useful for monitoring, interpreting and refining the model as the length and direction of the projection gives a graphical representation of which of the parameters have an influence on the principal components. If the projection of the original parameter is parallel or orthogonal to the principal component axis then the parameter is likewise correlated or uncorrelated to the principal component. In figure 5, the projection of these variables onto the first component shows that variance in parameter x_6 has the largest influence on

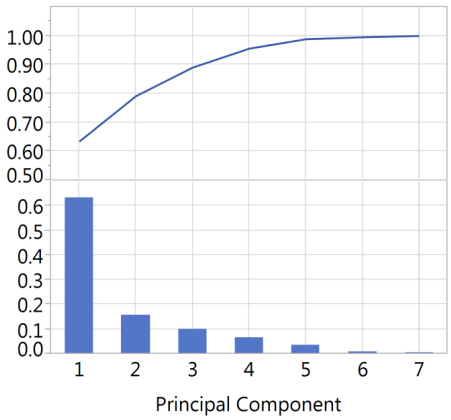

Figure 4: A scree plot showing the relative amount of variance explained by each successive principal component

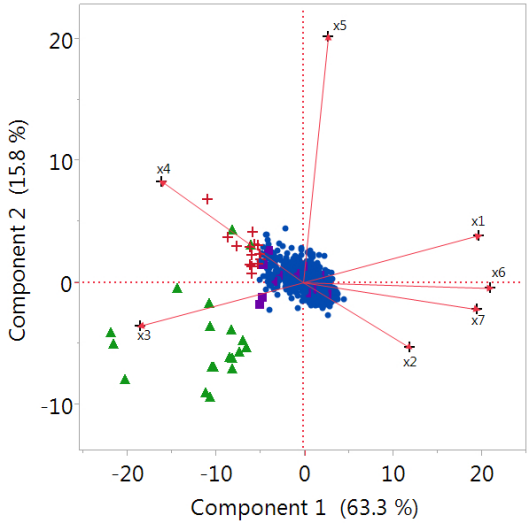

Figure 5: Bi-plot showing the principal component scores as for figure 3 with the projections of the original variable axis onto the principal components.

the first principal component, followed closely by x_1 and x_7. But all the parameters except for x_5, which is almost orthogonal to the first component axis, has a reasonable influence on the first principal component. With regard to the second principal component, parameter x_5 has far and away the greatest influence, as it is nearly parallel to the axis of the second principal component. This important information can be extracted qualitatively from the eigenvectors themselves. Where the parameter axis displayed on the bi-plot have a similar magnitude and direction, as for say x_6 and x_7, it is likely these parameters are highly correlated and maybe one can be removed from the model with little impact. Often though it is worthwhile to leave all of these parameters in the model as it provides for extra redundancy if something goes wrong with one of the parameters. Finally, there are also hints here as to the combination of parameters that may be causing the quadrant 2 (false rejects) cells to improve in processing. These cells are strongly clustered around the x_4(surface J_0) parameter axis projection,with possible some influence of the

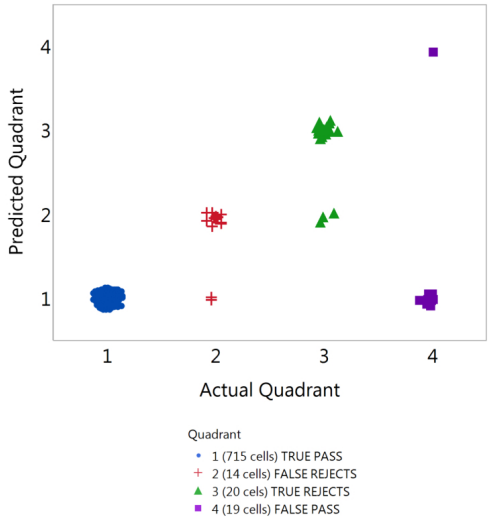

Figure 6: Predicted quadrant graphed against actual quadrant as determined by a neural network model. The meaning of the four quadrants are shown in the figure 2

negative x_2 (surface sheet resistance) parameter axis projection. Although this would need more thorough investigation, it is suggestive of samples with surface issues being amenable to improvement in subsequent processing. Also of note is that regardless of the projection of the data, the quadrant 4 (false pass) samples cannot be made to be entirely distinct from the quadrant 1 (true pass) samples. This suggests that whatever is causing these samples to lose performance during subsequent processing is a potentially random occurrence unrelated to the material properties measured at the lifetime step, or is occurring downstream in the process suggesting a direction for locating additional mid-sequence metrology.

C. Artificial Neural Networks (ANN)

An ANN model was built by training the model using a proportion of data where for each sample the seven lifetime parameters are known as well as the quadrant location of the final V_{oc} data. The predictive capability of this model is then tested on the remaining proportion of the data, by comparing the quadrant predicted by the model to the actual quadrant given by the final V_{oc} test. The predicted quadrant is graphed against the actual quadrant in figure 6. Importantly, once again, the model is able to correctly identify the quadrant 2 (false rejects) samples a significant proportion of the time. A quadrant 2 sample is never misclassified as a quadrant 3 (true rejects) sample. As for the principal component analysis, quadrant 4 (false pass) samples are not distinguishable from quadrant 1 (true pass) samples by the model.

The primary output of a neural network model is a calculated probability for the likelihood that any given set of input parameters x_1 through to x_7 falls into any given quadrant. By default in most models, the chosen outcome is the option that returns the highest probability. There are some applications

where these thresholds can be adjusted to tune the occurrence of false classification. For example, if high process value has been added to the product at the time of the measurement, it may be that it doesn't matter if a quadrant 3 (true reject) samples is passed through the line, but really important that a quadrant 2 (false reject) sample is not prematurely discarded. Threshold probabilities can be adjusted to tune these outcomes.

At this point, the PCA model also offered a lot of extra insight in regards to the relationships between the original variables and the final performance. But this is not so easy to extract from a neural network model short of using a Monte Carlo - style simulation. Consequentially, it is more difficult to determine how the model might be refined or when the model starts behaving differently. This is one of the limitations of the technique as has been previously discussed.

V. Discussion

It is unsurprising that there are limitations in using a single variate data stream like the iV_{oc} to predict the final performance of a cell. However, this works shows that by using the seven parameters collected at the lifetime measurement in a multivariate model, it is possible to predict the final performance with a lower margin of error. Both of the approaches shown here can be useful for generating a set of criteria to more confidently reject under-performing samples at the mid-sequence lifetime measurement. In addition to this, the PCA model is also providing some insight into why some samples may be getting better and some samples may be getting worse during processing. It is also possible to easily monitor the performance of the model itself over time by examining how consistent the projections of the original parameters are onto the principal components, or even where they change in response to process stimulus. The PCA algorithm is itself theoretically very simple, but the outcome is almost entirely dependent on the way the variance in the data set is normalized. There are many opportunities to modify how this normalization is implemented to improve the modeling outcomes. If the main outcome required is to only generate a set of pass / fail criteria at the lifetime measurement, the ANN approach is very useful, although more examination needs to be done with much larger data sets. This will either enable the building of more accurate and robust models, or it will challenge the generality of the approach due to the increased variance in a larger data set. The threshold probabilities can also be varied in the ANN model to tune and tailor the prediction errors. It is likely that good models can be built with many other multivariate techniques. Partial Least Squares particularly offers some appeal for further investigation as it offers some of the advantages of both of the approaches taken here,and also allows for the inclusion of categorical data in the model, which is particularly useful in some circumstances.

VI. Conclusions

Multivariate approaches to process control and process learning will become increasingly important as larger, multidimensional data-sets are generated in PV manufacturing. The investigations here show that when complex measurement data is parametrized and scaled appropriately, it can still be used in its totality to make improved inferences regarding process control and process performance. This has significant implications for improving the understanding and the yield of PV production lines. There is a large body of further work in defining the best, the most robust and the most transferable models of this type.

Acknowledgments

The Australian Center for Advanced Photovoltaics is supported by the Australian Government through the Australian Renewable Energy Agency (ARENA). Responsibility for the views, information or advice expressed herein is not accepted by the Australian Government

References

[1] G. Koeksal, I. Batmaz, and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13 448–13 467, 2011.

[2] S. H. Liao, P. H. Chu, and P. Y. Hsiao, "Data mining techniques and applications - a decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[3] J. L. Coleman and J. Nickerson, "A multivariate exponentially weighted moving average control chart for photovoltaic processess," in *Conference Record of the IEEE Photovoltaic Specialists Conference*, 2005, pp. 1281–1284.

[4] R. Evans, J. Dore, E. V. Voorthuysen, J. Zhu, and M. A. Green, "Data mining photovoltaic cell manufacturing data," in *2014 IEEE 40th Photovoltaic Specialist Conference, PVSC 2014*, 2014, pp. 2699–2704.

[5] R. Evans, J. Dore, E. V. Voorthuysen, and M. A. Green, "Interpreting manufacturing variance using a data mining approach," in *Proceedings of the 29th EUPVSEC*, 2014.

[6] I. G. Chong, S. L. Albin, and C. H. Jun, "A data mining approach to process optimization without an explicit quality function," *IIE Transactions (Institute of Industrial Engineers)*, vol. 39, no. 8, pp. 795–804, 2007.

[7] K. R. Skinner, D. C. Montgomery, G. C. Runger, J. W. Fowler, D. R. McCarville, T. R. Rhoads, and J. D. Stanley, "Multivariate statistical methods for modeling and analysis of wafer probe test data," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, no. 4, pp. 523–530, 2002.

[8] Q. Zhou, Z. Xiong, J. Zhang, and Y. Xu, *Hierarchical neural network based product quality prediction of industrial ethylene pyrolysis process*, ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006, vol. 3973 LNCS.

[9] J. E. Jackson, *A user's guide to principal components*. New York : Wiley, 1991.

[10] H. H. Harman, *Modern factor analysis*. Chicago : University of Chicago Press, 1967.

[11] R. A. Johnson, *Applied multivariate statistical analysis*. Upper Saddle River, N.J. : Prentice Hall ; London : Prentice-Hall International, 1998.

[12] D. C. Montgomery, *Introduction to linear regression analysis*. New York: New York : Wiley, 1982.

[13] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, no. 0, pp. 1–17, 1986.

[14] A. Pratley, E. van Voorthuysen, and R. Chan, "A step by step approach for modelling complex systems with partial least squares," *In Press*, 2014.

[15] N. A. Campbell and W. R. Atchley, "The geometry of canonical variate analysis," *Systematic Biology*, vol. 30, no. 3, p. 268, 1981.

[16] D. E. Kane and R. M. Swanson, "Measurement of the emitter saturation current by a contactless photoconductivity decay method," in *Conference Record of the IEEE Photovoltaic Specialists Conference*, 1985, pp. 578–583.