Methods for Wavelet-Based Autonomous Discrimination of Multiple Partial Discharge Sources

R. D. Nimmo
Western Power Distribution
Bristol, BS2 0TB

G. Callender and **P. L. Lewin**
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK

ABSTRACT

Recent years have seen increased interest in the application of on-line condition monitoring of medium voltage networks as the need to maintain and operate ageing cable networks increases. Detection and analysis of partial discharge (PD) activity is generally used as an indicator of pre-breakdown processes that may be indicative of insulation degradation over time. A significant challenge for on-line monitoring is discrimination between multiple partial discharge sources that will often naturally exist in the data. To discriminate between PD sources each PD signal is represented as a feature vector and a clustering algorithm is used to identify clusters in the resulting feature vector space, often after dimensional reduction. Each cluster identified in the data corresponds to a distinct PD source. In this work a comparison of clustering algorithms and dimensional reduction techniques is performed to identify clusters for a variety of PD data sets, in all cases the feature vector is created using wavelet decomposition energies. The three clustering algorithms used were Density Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points to Identify Clustering Structure (OPTICS) and Simple Statistics-based Near Neighbour clustering technique (SSNN). The two dimensional reduction techniques considered were Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE). At present the most commonly used combination of dimensional reduction technique and clustering algorithm for PD data are PCA and DBSCAN respectively. From the comparison performed in this work it was found that t-SNE combined with OPTICS or SSNN were the most successful at clustering PD data. For use in practical situations SSNN is preferred over OPTICS as it requires only a single input parameter, which generally be hardcoded, leading to a completely autonomous technique. It is therefore suggested that a combination of t-SNE and SSNN is particularly appropriate for discriminating PD sources.

Index Terms — **Partial Discharge, Condition Monitoring, SSNN, OPTICS, t-SNE.**

1 INTRODUCTION

THE economic incentives of on-line condition based monitoring (CM) of electrical apparatus have encouraged interest and funding among utilities wishing to extend the life of ageing equipment. This is especially true of medium voltage (MV) network operators whose urban cable networks were installed in the middle of the previous century, and are consequently nearing the end of their useful life.

Of the methods used for assessing the condition of electrical apparatus, analysis of the partial discharges (PD) produced by the system has proved most informative and is currently seeing widespread acceptance. A successful PD-based on-line CM tool for high voltage plant must be able to: discriminate between PD sources within the asset, localise each PD source, and assess the likely severity of a potential developing fault. Source discrimination, comprised of feature extraction and classification, is the fundamental step in this process and has received significant research attention [1]. The current method of selecting the most appropriate classification (or clustering) algorithm for a specific data set remain a largely experimental-based approach [2]. Of the plethora of available techniques, PD source classification has been achieved using k-means [1], fuzzy-classifiers [3], Density Based Spatial Clustering of Applications with Noise (DBSCAN) [4], Ordering Points to Identify Clustering

Manuscript received on 13 July 2016, in final form 28 November 2016, accepted 12 January 2017. Corresponding author: G. Callender.

Structure (OPTICS) [5], support vector machines [6], and probabilistic neural networks [7]. By contrast, the feature extraction stage has seen a more restricted selection, with the main methods using either a time-frequency (TF) approach [8], or the use of a discrete wavelet transform (DWT) [9]. More recently, the current methods of feature extraction have been classified into so-called T1, T2, and T3 spaces (using global, mean, and single valued parameters respectively), and an evaluation of their performance carried out [5]. The results of this analysis were that the T2 (such as the DWT) and T3 spaces showed similar performance, although T3 was marginally better.

In this paper an investigation into the application of recently reported source discrimination techniques for PD data is performed. Different approaches to wavelet-based source discrimination and classification have been assessed using three different PD measurement data sets. Firstly, the techniques used to discriminate sources of PD are introduced and a summary of the data sets used in the analysis are provided. The visually intuitive output of OPTICS is then used to evaluate the performance of current wavelet-based discrimination techniques involving the use of Principal Component Analysis (PCA) and t-SNE followed by clustering with DBSCAN. The results of PD source discrimination from the two most promising combinations of dimensional reduction techniques and clustering algorithms are then presented, namely t-SNE with OPTICS and t-SNE with SSNN. One of the key advantage of these methods is that they have a high level of autonomy, and can be employed by a user unfamiliar with the underlying algorithms.

2 FEATURE EXTRACTION

Feature extraction is an important step in source discrimination, because a subsequent clustering technique can only act on variance in the data that actually exist in the feature space generated by the extraction process. The DWT has shown potential in discriminating between PD signals within the noisy environment of on-line applications [9] as well as accurately maintaining distinctions between PD sources from a variety of equipment [5]. It is generally followed by a dimensional reduction technique, typically principal component analysis (PCA), to reduce the dimension of the feature vector to three [1,4,5]. More recently, t-SNE has been proposed as an alternative form of dimension reduction that is especially suited to representing higher dimensional spaces in two or three dimensions. In this section, details regarding the application of the DWT are presented, followed by a brief introduction to PCA and t-SNE.

2.1 DISCRETE WAVELET TRANSFORM

Detected PD signals are inherently transient, aperiodic signals which are best extracted with an asymmetric wavelet basis such as the Daubechies wavelets, Symlet wavelets, and Lemarie wavelets, as this will produce a better match to the shape of the PD pulse [10-11]. More specifically, it has been shown that an order two wavelet is best for analysing an exponential PD pulse whereas the higher order (eight to ten) wavelets are best for damped resonant PDs, typical from an RLC-type detection impedance [11]. In this analysis a higher order Symlet wavelet (sym8) was found to be particularly good at representing the pulses and has been used throughout. For a detailed description of the DWT algorithm the reader is referred to [9].

The choice of the number of decompositions is also vital for reliable feature extraction. It should be noted that there exists an absolute maximum number, $J-1$, of decomposition levels that will allow accurate representation and complete reproduction of the signal [12-13]. The value of J is calculated as

$$J = \mathrm{floor}(\log_2 T), \tag{1}$$

where floor is the floor function and T is the length of the discretised pulse in data points [12].

The wavelet coefficients produced by the DWT possess high dimensionality (a large number of points in each detail level) so it is advisable to use the signal energy in each level rather than the coefficients themselves to describe the PD signal [1, 4, 9]. For the implementation of DWT it is required to specify the number of decompositions and the type of wavelet to use. However, based on the results presented here and elsewhere in the literature [10-11], the number of decomposition levels can be set to their maximum value, $J-1$, calculated using equation 1, and sym8 wavelets are suitable for a variety of PD data. As such both of these parameters can be hard coded and do not need to be considered as user inputs.

2.2 PCA

Dimension reduction using PCA is achieved by means of an orthogonal linear transformation in the direction of the greatest variance exhibited by the data [1]. Formally, PCA tries to find the eigenvectors with the largest eigenvalues of the covariance matrix. As a result of this it tends to prioritise large pairwise distances between data-points instead of small pairwise distances which may still be important; it can therefore reduce sensitivity to subtle differences in data characteristics while exaggerating the more distinctive differences [1].

However, this particular method of dimension reduction is most appropriate for use on data sets where the variables are dependent and inter-correlated [14]. Hence it may be argued that it can hinder the full performance of the DWT, where the feature space is created from independent variables. To investigate this, an alternative dimensional reduction technique is also considered, t-SNE [15]. It should be noted that PCA is completely autonomous and requires no input parameters.

PCA does not require significant computational time and can reduce the dimensions of a data set in the order of MB in less than ten seconds on a standard desktop PC.

2.3 T-SNE

t-SNE is a non-linear technique recently introduced for presenting high-dimensional data in two- or three-dimensions. It uses random walks on neighbourhood graphs to allow the implicit structure of all of the data to influence the way in which a subset of the data is displayed [15].

The method considers the Euclidean distances between data points in the high dimensional space and converts them into conditional probabilities which represent similarities [1]. When moving to the lower dimension space it attempts to maintain these probabilities and so retain the proximity of similar points, with the result that the lower dimensional data retains its original clustered appearance. It should be noted that t-SNE is completely autonomous and requires no input parameters.

Due to the iterative process involved in t-SNE, the computational time involved was significant compared to the discrimination processes. t-SNE reduces the dimension of a data set in the order of MB in minutes.

3 FEATURE DISCRIMINATION

Source discrimination using a wavelet-based approach is dominated in literature by subsequent clustering with DBSCAN, a well-known algorithm that is able to deal with the presence of noise as well as discover clusters of different shapes [16]. More recently, OPTICS has been used to evaluate different spaces for PD signals [5], and showed some promising results.

Recent literature has also shown interest in near-neighbour clustering as alternatives to density based clustering, although not in the field of CM. Of these, one of the most promising in terms of performance is SSNN, a statistics-based algorithm that is able to discover clusters of different shapes and densities, as well as deal effectively with noise [2].

To provide a broad spectrum of enquiry, DBSCAN, OPTICS and SSNN will all be evaluated for PD source classification performance. For all clustering algorithms considered it is necessary to specify the minimum number of points, $MinPts$, required for a group to be considered a cluster. If groups do not meet this criterion they are regarded as outliers and discounted.

3.1 DBSCAN

The use of DBSCAN in PD CM applications is well documented [1,9,17]. DBSCAN targets low dimensional spatial data [18], and relies on two input parameters ε and $MinPts$ [16]. Two points in feature space are considered directly reachable if the distance between them is less than ε. Clusters are formed of points that are all reachable from each other, i.e. any path between any two points in a cluster consists of directly reachable steps between intermediate points. The algorithm does not require significant computational time and can produce clusters from data in the order of MB in less than ten seconds on a standard powerful desktop PC.

The main problem with DBSCAN is its inability to deal with data in which clusters are of different densities. Additionally, clustering performance is very dependent on the input parameter ε whose selection is non-trivial [2]. This limits the reliability of its application to research environments where the characteristics of the data it is being applied to are under investigation.

3.2 OPTICS

OPTICS [19] is broadly similar to DBSCAN but has seen use in CM applications only very recently [5]. It has been chosen as it produces results that are visually interpretable and thus it can provide user insight into a data set that would not be possible with other clustering algorithms.

OPTICS adjusts DBSCAN's shortfalls to provide a more flexible clustering procedure [18]. The only input parameter to the algorithm is $MinPts$. In all these studies, a value of between one and two percent of the number of PD signals analysed was found to be perfectly adequate, which is consistent with previous results [5]. The clustering algorithm works by moving between points in the feature space creating an ordered seed list of reachable points and storing the minimum reachability distance of each point. The reachability distance between two points, p and q, is defined as the maximum of the distance between the points p and q, and the distance between p and the $MinPts^{\text{th}}$ nearest point. An overview of the OPTICS algorithm can be found in [20].

The output from OPTICS, an ordered seed list and minimum reachability distance, is used to plot a reachability graph, which is a useful way of displaying clusters that exist in n-dimensional space. Within this type of plot, clusters are visible as "valleys" and are separated by large "spikes". The higher the peak of these spikes the more "different" the data will be in the clusters; and the lower the valley the "denser" the cluster.

Automatic cluster extraction has been implemented here based on the derivative and magnitude of the reachability graph. A cluster boundary is therefore defined by a large derivative of the reachability distance, coupled with a magnitude that is significantly above the mean within a small window around the boundary. For all data sets considered a suitable window was 1% of the total number of PDs.

The algorithm does not require significant computational time and can produce clusters from data in the order of MB in less than five seconds on a standard desktop PC.

3.3 SSNN

SSNN [2] differs from the previous techniques in that it constructs the clusters based on a statistical understanding of the data. For each point in feature space the distances between it and every other point are calculated. The mean and standard deviation of these distances are used to calculate a distance ε for each point, which can then be used to find clusters in a method similar to DBSCAN. The authors of SSNN also present an algorithm to aid clustering of outlying points, but it was not considered in this work as it was found that this improved speed without loss of necessary performance.

The only input required for SSNN is the minimum number of points needed to form a cluster $MinPts$. Although there is a method to calculate $MinPts$ automatically in [2], it was found that this technique was not reliable for PD data as there can sometimes be a very large variation in the cluster size within a single data set. For the data sets considered $MinPts \approx 10$ was sufficient.

The algorithm does not require significant computational time and can produce clusters from data in the order of MB in less than five seconds on a standard desktop PC.

3.4 DISSIMILARITY MEASURES

To implement these clustering algorithms it is necessary to define the distance between points in feature space. This distance is often referred to as the dissimilarity measure. Density based clustering algorithms for CM applications have exclusively used the pairwise Euclidean distance between points as the dissimilarity measure, as it is the obvious basis on which to form a density-based approach. Of course a Euclidean measure does not have to be used, and studies have shown that non-Euclidean metrics and non-Euclidean non-metrics can both provide informative dissimilarity measures for clustering [21].

The Euclidean distance itself forms part of a set of L_z distance measures which include L_1 (Manhattan), L_2 (Euclidean), L_n (Minkowski with order n) and L_∞ (Chebychev) [22-23]. Metrics may also be based on angular rather than linear methods (as in the Cosine measure) or be developed with specific applications in mind (e.g. the Jaccard measure). Brief investigation by the authors concluded that for PD source discrimination the Euclidean metric is sufficient and there is no advantage to using other metrics. As a result the Euclidean metric is used as the dissimilarity measure throughout this work.

4 PD DATA SETS

The work presented here deals with discrimination of PD sources from PD signal data. Prior to this PD signals must be extracted from the raw acquisition data. For the cases presented here PD signals were extracted by identifying data points that were above a threshold set by the background noise level. Pulses were then extracted by taking data points before and after this “peak” above the background noise level. This threshold was determined automatically by initially setting it to a value equal to the largest data value. The threshold was then gradually reduced, and the number of data points above this threshold was recorded. When the number of data points above the threshold greatly increased with a small reduction in the threshold it was concluded that the threshold had reached the background noise level of the acquisition data. The process of extracting PD signals did not require significant computational time, with PD signals extracted from raw data files in the order of MB in minutes.

The different source discrimination techniques have been evaluated using data from a variety of sources, both field-based and experimental. Field data was measured using commercially available online PD monitoring equipment, details are provided in the literature [6]. The two data sets considered were MV cable circuits in the London area:

1) Tunbridge Wells - Approximately 950 PD pulses were recorded on 9th August 2008 by on-line advanced substation monitors (ASMs) installed on an 11 kV feeder located at Tunbridge Wells. The sampling rate was 50 MSa/s. A detailed visual examination of the data set revealed five distinct clusters, representing five visually different pulse shapes.
2) City Road - Approximately 1200 PD pulses were recorded over 342 AC cycles between the 4th and 13th August 2008 by on-line ASMs installed on an 11 kV feeder located at City Road, London. The sampling rate was 50 MSa/s. A detailed visual examination of the data set revealed five distinct clusters, representing four visually different pulse shapes, along with a small number of very noisy pulses that were considered to be outliers.

Data from a previously reported laboratory experiment was also analysed [6]. In this case the likely source of the data was known:

3) Void in the Crutch - An experiment was undertaken in which a void was intentionally included in the crutch of a three phase 11 kV belted paper insulated lead covered (PILC) cable joint [1]. The overall jointed cable was several meters long, and was allowed to thermally cycle overnight before measurements were taken, mimicking the variable load a cable might experience in the field. The experiment allowed collection of PD from all three phases, but only the phase with the highest PD peak apparent charge magnitude is included here. Due to the setup, it was expected that there would be one source of PD within the cable (i.e. the void in the crutch) and two sources producing PD signals in the cable terminations (one at each end of the cable).

5 EVALUATION OF CURRENT WAVELET-BASED TECHNIQUES

OPTICS produces a reachability graph which can easily display clusters that exist in n dimensions. This graph is particularly useful when analysing the performance of different feature-extraction techniques, as well as the density (and its variation) within clusters. This section uses this tool to evaluate the effectiveness of the dimension reduction techniques outlined in Section 2, as well as demonstrate the main difference in cluster discrimination between DBSCAN and OPTICS.

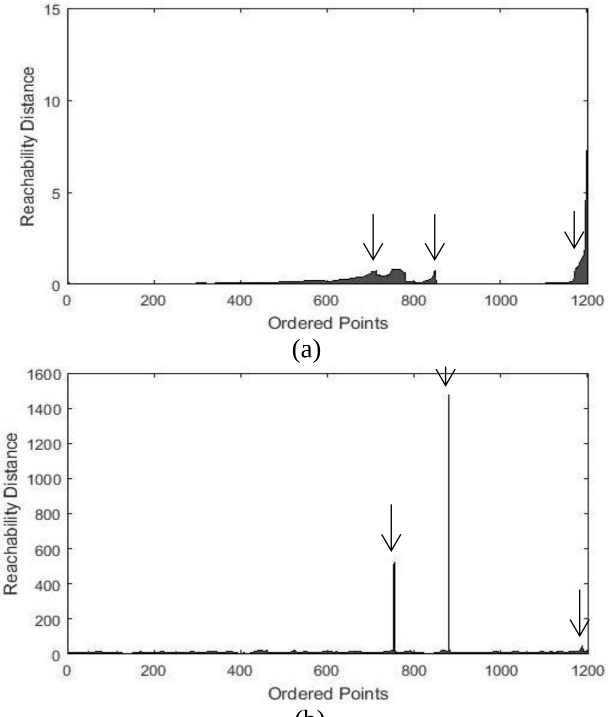
5.1 EVALUATION OF DIMENSION REDUCTION TECHNIQUES

It has been found previously that t-SNE is an improvement over PCA [1], and this has been shown using DBSCAN as a clustering technique. Due to the visually intuitive results that are possible with OPTICS, this study has been performed using this clustering tool in order to gain further insight into the results obtained.

Figures 1a and 1b shows reachability graphs from the City Road data that has been processed using PCA and t-SNE respectively. The difference in the separation of the clusters is visible from the much larger scale of the vertical axis in the t-SNE data. Additionally, it has been found by detailed visual examination of the pulse data that four clusters exist in this data set, which are not distinct in plot (a) but the boundaries for which are very clear in plot (b).

5.2 LIMITATIONS OF DBSCAN

During testing it was found that the performance of DBSCAN is limited by the use of its single-valued parameters ε and $MinPts$. Figure 2 shows the output of OPTICS contrasted with the method used to generate clusters using DBSCAN for the City Road data set. This comparison is possible due to both methods using very similar density based clustering techniques. Four clusters can be identified in Figure 2, but unfortunately DBSCAN overlooks two peaks and concludes that only the two main clusters exist (see horizontal line), because the clusters have significantly different densities. It is noteworthy that it is impossible to extract all four clusters correctly using DBSCAN regardless of input parameter choice.



(a)

(b)

Figure 1. Reachability graphs produced for the City Road data set pre-processed with PCA (a), and t-SNE (b). There are four clusters in this data set (points 0-750, 751-875, 876-1180, and 1181-1200), the boundaries for which have been marked by arrows. The scale on the y axis shows that t-SNE was more successful at separating clusters in the data due to the high magnitude spikes in the reachability distance.

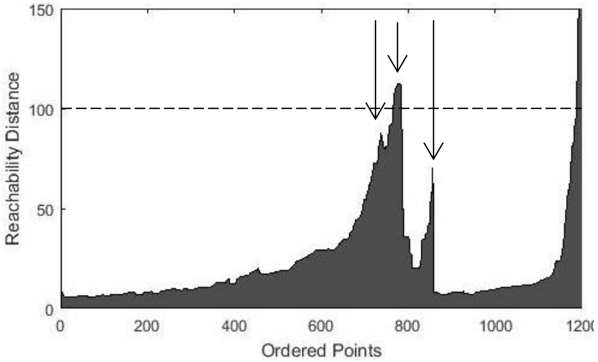


Figure 2. Reachability graph produced by OPTICS showing four cluster for the City Road data set without dimensional reduction. The dashed line shows the effect of using a constant ε, the case in DBSCAN, where only two clusters can be extracted.

6 TWO NEW METHODS FOR PD SOURCE DISCRIMINATION

It has been established that t-SNE is a more accurate method of representing high dimensional wavelet data in three dimensions compared to PCA. Additionally, of the two density-based feature classification methods currently used in literature for PD data clustering, OPTICS has shown greater potential because it overcomes the shortfalls of DBSCAN's single valued clustering parameters. Notwithstanding the general prevalence in CM-related literature of density-based classification techniques, it has been stated in Section 2 that near-neighbour classification techniques have shown an upsurge of interest and good performance. This section will therefore also apply a new statistics-based classification technique, SSNN.

It should be noted that an exhaustive sweep of all combinations of dimensional reduction techniques, clustering algorithms and data sets introduced in this work has been performed by the authors. However, to be concise only the results from the most successful combinations are presented here, which are OPTICS and SSNN, with t-SNE used as the dimensional reduction technique in all cases.

6.1 DISCRIMINATION WITH OPTICS

Investigations were performed on the Tunbridge Wells data set. Figure 3 shows that the clusters are clearly visible in three dimensions. t-SNE has therefore been very successful in maintaining the assumed clustered nature of the data in the higher dimensional wavelet space. Figure 4 shows the reachability graph which shows similarly distinct clusters. Figure 5 shows representative pulses from each cluster together with a box plot of all the clusters in the data set. From this it can be deduced that the clusters found in the reduced dimension wavelet space map to the time domain as well. Figure 6 shows a phase resolved partial discharge (PRPD) pattern of the data.

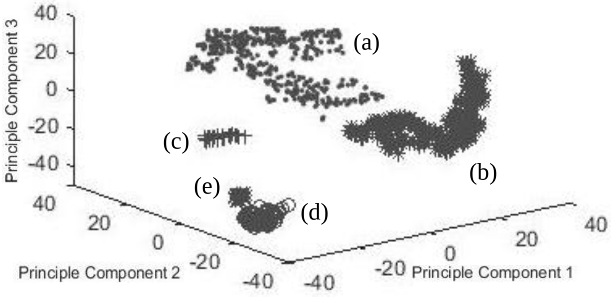


Figure 3: Three-Dimensional scatter showing the clusters produced by OPTICS from the Tunbridge Wells data set labelled (a)-(e).

Following this, the technique was applied to the Void in the Crutch data set. Figure 7 and Figure 8 show the 3D plot of the data and the reachability graph respectively, and the expected three clusters are clearly visible. Figure 9 shows individual pulses and box plots of the clustered data, from which the validity of the three clusters can be further confirmed. Figure 10 shows a PRPD pattern of the data.

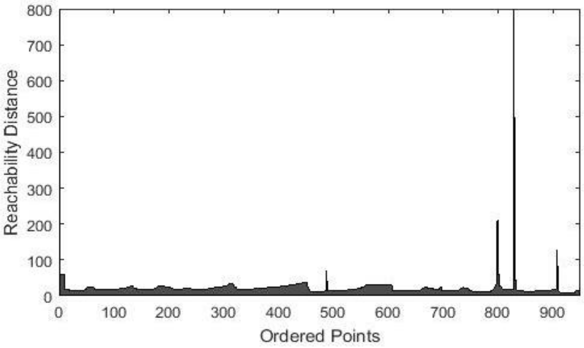
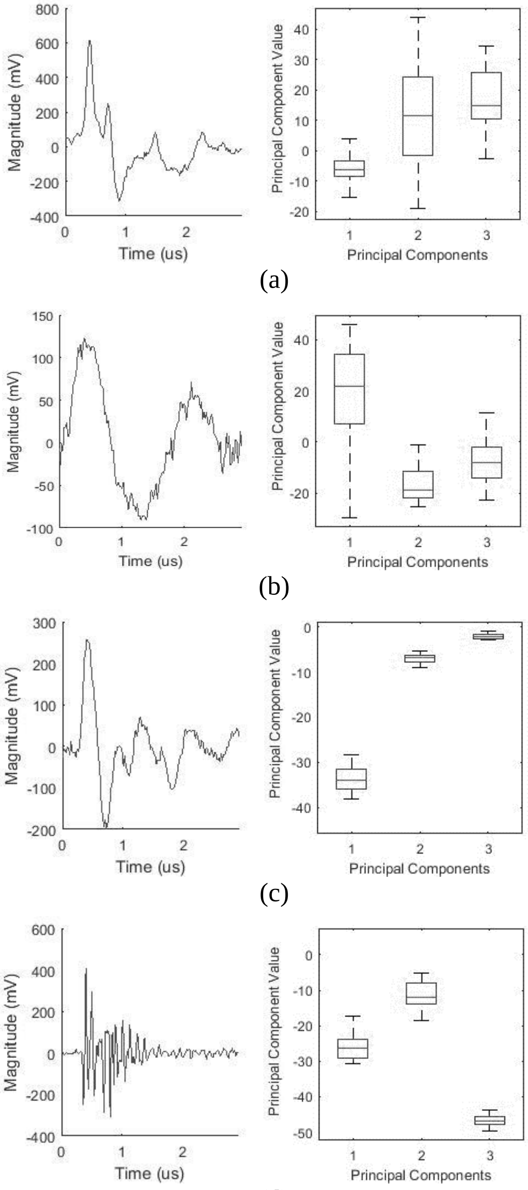


Figure 4: Reachability plot generated by OPTICS from Tunbridge Wells data set.

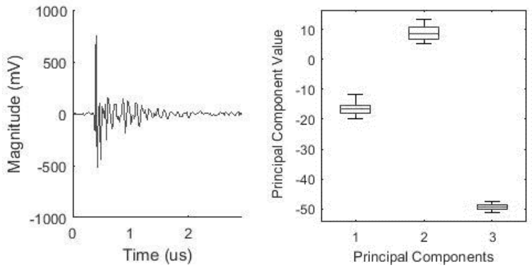


(a)

(b)

(c)

(d)



(e)

Figure 5. Plots of a single pulse from each of the different clusters together with the box plot of all pulses from that cluster from the Tunbridge Wells data set. (a)-(e) corresponds to the same cluster in Figure 3, clusters identified using OPTICS.

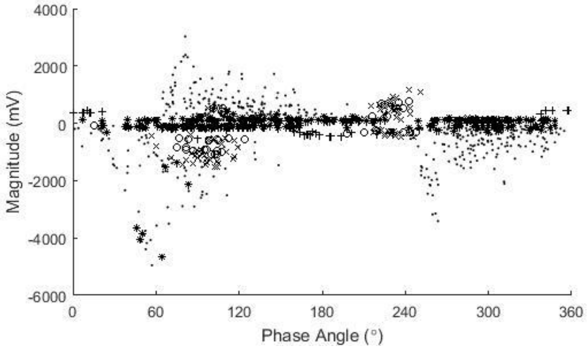


Figure 6. PRPD pattern Tunbridge Wells data set. Marker Types denote distinct clusters identified using OPTICS, markers are the same as those in Figure 3.

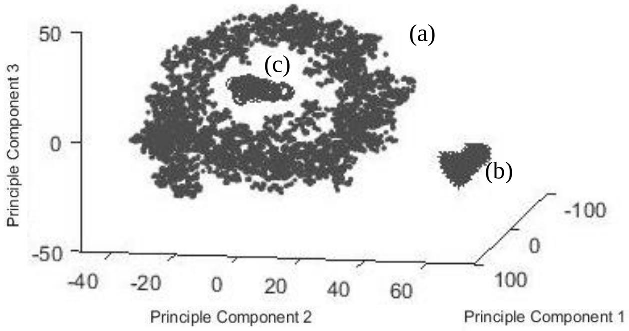


Figure 7. Three-Dimensional scatter showing the clusters produced by OPTICS from the Void in the Crutch data set labelled (a)-(c).

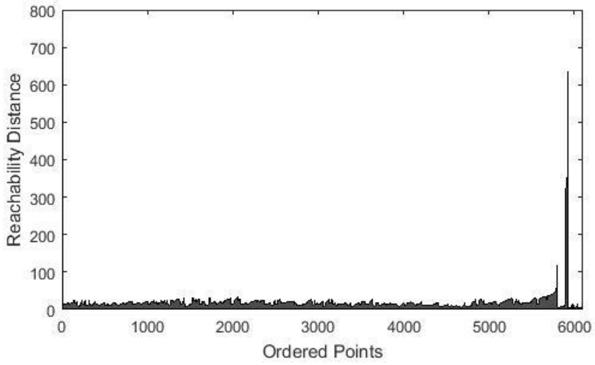
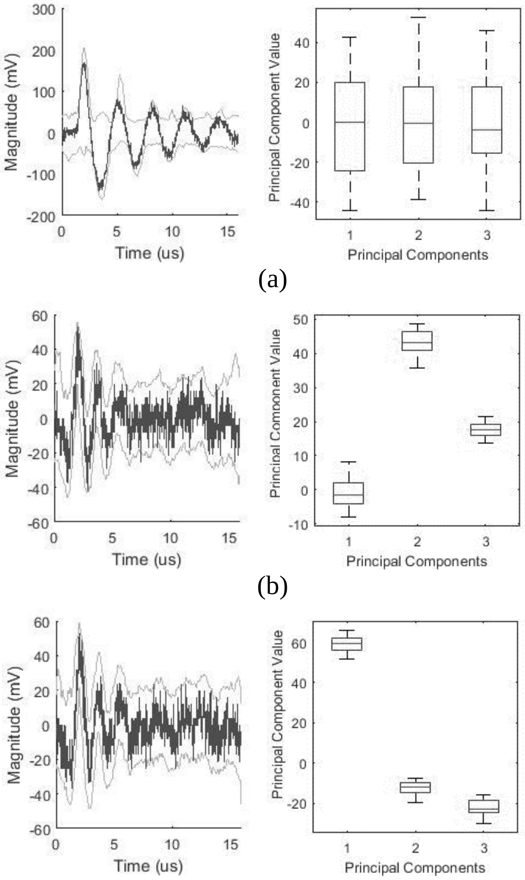


Figure 8. Reachability plot generated by OPTICS from Void in the Crutch data set.

These initial tests have shown that the use of t-SNE and OPTICS have proven a viable alternative to the more prevalent methods involving PCA and DBSCAN. Data of varying density has been correctly clustered easily by the algorithms. Validity of the clusters in the feature space created by t-SNE has been visually confirmed through the use of scatter plots, box plots of the principal components, and in the time domain through the plotting of representative pulses. This is a clear improvement over previous methods. During testing it was noted that OPTICS was sensitive to its input clustering parameter, $MinPts$, although not to the same extent as DBSCAN. OPTICS also requires clustering extraction from a reachability graph; it was found that this could be achieved through a simple algorithm, described in Subsection 3.2. Finally, performance has been good across both field and experimental measurements of PD data, showing the versatility of the approach.



(a)

(b)

(c)

Figure 9. Plots of a single pulse from each of the different clusters together with the box plot of all pulses from that cluster from the Void in the Crutch data set. (a)-(c) corresponds to the same cluster in Figure 7, clusters identified using OPTICS.

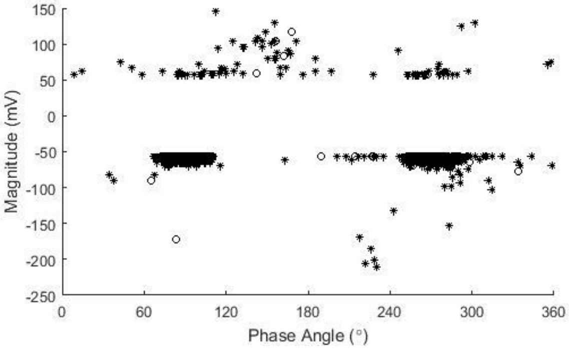


Figure 10. PRPD pattern of the Void in the Crutch data set. Marker Types denote distinct clusters identified using OPTICS, markers are the same as those in Figure 7.

6.2 DISCRIMINATION WITH SSNN

SSNN combined with t-SNE was also tested on MV field data, using the City Road data set. Four clusters were correctly identified, and can be seen in the 3D plot in Figure 11. Representative pulse shapes, along with maximum and minimum lines, and box plots, for each cluster are shown in Figure 12. A PRPD pattern of the data is shown in Figure 13.

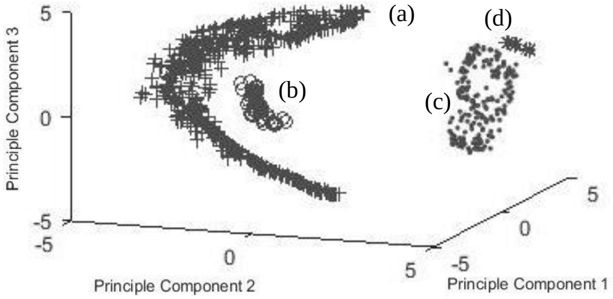
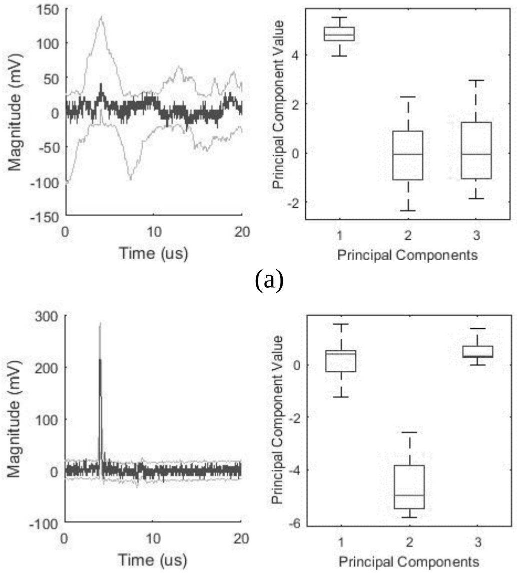
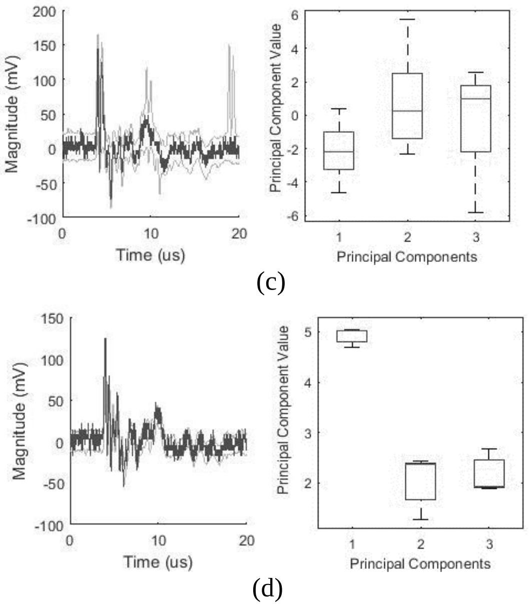


Figure 11. Three-Dimensional scatter showing the clusters produced by SSNN from the City Road data set labelled (a)-(d).



(a)

(b)



(c)

(d)

Figure 12. Plots of a single pulse from each of the different clusters together with the box plot of all pulses from that cluster from the City Road data set. (a)-(d) corresponds to the same cluster in Figure 11, clusters identified using SSNN.

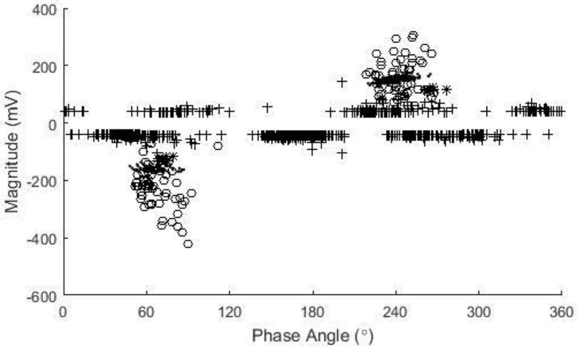


Figure 13. PRPD pattern of the City Road data set. Marker Types denote distinct clusters identified using SSNN, markers are the same as those in Figure 11.

The technique was also applied to the Void in the Crutch data. The results are shown in Figure 14, 15 and 16, which are very similar to the results from the use of OPTICS. The difference in the 3D plot, Figure 14, will result from the reduced number of iterations used in the t-SNE dimension reduction, and the stochastic nature of the algorithm.

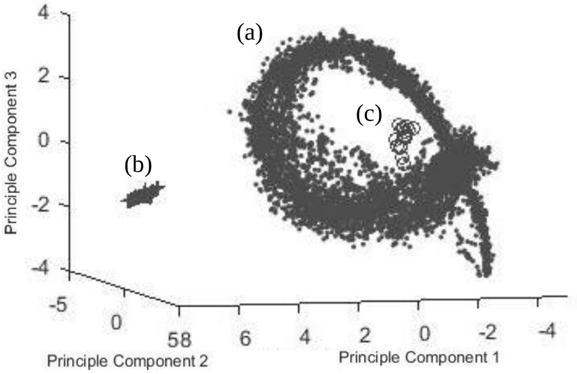
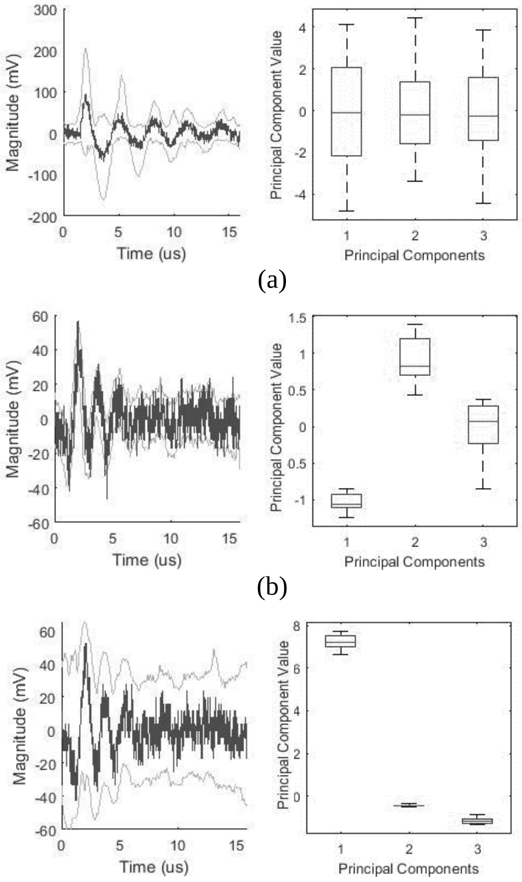


Figure 14. Three-Dimensional scatter showing the clusters produced by SSNN from the Void in the Crutch data set labelled (a)-(c).



(a)

(b)

(c)

Figure 15. Plots of a single pulse from each of the different clusters in the Void in the Crutch data set together with the maxima and minima of all pulses in that cluster, and a box plot of all pulses from that cluster. Clustering was performed with SSNN.

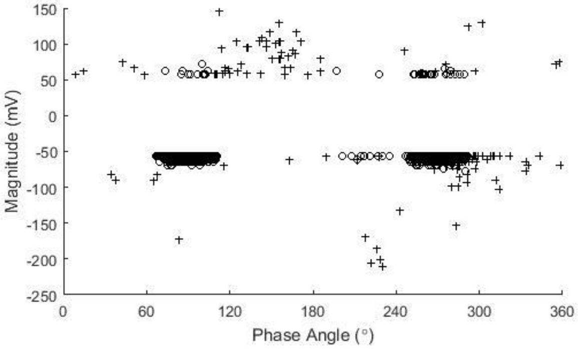


Figure 16. PRPD pattern of the Void in the Crutch data set. Marker Types denote distinct clusters identified using SSNN, markers are the same as those in Figure 14.

6.3 DISCUSSION

The use of OPTICS as a discrimination technique has shown some impressive results, and the method is clearly an improvement over current density-based approaches which suffer from lack of sensitivity, inability to identify clusters of different densities, and reliance on single-valued clustering parameters. Furthermore, with the inclusion of automatic cluster selection from the reachability graph, this method only requires $MinPts$ to be set. Using a value of between one and two percent allowed the technique to become completely autonomous, although fine tuning the value allowed slightly more accurate clustering in some cases, particularly when trying to correctly find very small clusters in a large data set.

The use of SSNN has shown performance that matched that of OPTICS for selectivity, sensitivity, and reliability. Unlike OPTICS, SSNN outputs the clusters directly, rather than through a reachability graph. The reachability graph was extremely useful in assessing the performance of previous signal processing techniques, making it an ideal choice for research based applications. However, SSNN was faster than OPTICS as it did not include these unnecessary steps and required fewer t-SNE iterations. Furthermore, OPTICS requires an additional algorithm to identify peaks in the reachability graph and while this could be achieved simply for the data sets considered here this may not always be the case. Finally, the results of the clustering algorithm are not dependent on the user input $MinPts$ as it only used to exclude small clusters as noise. It is therefore entirely reasonable that when using SSNN $MinPts$ can be hardcoded, thus making the PD source discrimination procedure completely autonomous. The approach using SSNN is therefore advocated for use in systems where speed is important and all that is required is a reliable output without requiring any user input.

Due to the influence of attenuation and dispersion of the PD signal classification of defect type for PD sources is not generally possible for MV cable circuits and this is also the case for the two field-based data sets considered here. However, for the Void in the Crutch data set it is reasonable to conclude the large cluster, (a) in Figures 7 and 14, is from the void in the crutch. The two smaller clusters are likely to be from the cable terminations based on their similar size and shape of PD signals. It is encouraging that three clusters were correctly identified in both OPTICS and SSNN.

This work has shown the shortcomings of the standard method of classifying PD sources; PCA followed by DBSCAN on a feature space. A new dimensional reduction technique, t-SNE, and two new clustering techniques, OPTICS and SSNN, have been shown to address these shortcomings. OPTICS and SSNN were then used successfully to identify PD sources in data from MV cables, which could not be found using the standard method.

An important point is that although the analysis performed here was performed after data acquisition it is possible that the identification of PD sources could be performed simultaneously with measurements. As the raw data is taken, PD signals are identified using the autonomous thresholding detailed in section 4. Each signal could then be decomposed into a feature vector using the DWT and would populate a feature space. Then at regular time intervals dimensional reduction and clustering algorithms could be used to identify new sources of PD within the data. As mentioned previously, a combination of t-SNE and SSNN could perform this routine with complete autonomy.

7 CONCLUSION

Two new methods for discrimination between sources of PD within MV cables have been presented. Both methods rely on the discrete wavelet transform to correctly extract the PD signals, producing a lower dimension feature space. This has been followed with the application of t-SNE as a dimension reduction technique, which allowed the clusters to be visualised in 3D space, and also improved the performance of consequent clustering considerably. It is proposed that SSNN combined with t-SNE is preferable to the existing techniques for the fast and reliable identification of PD sources in complex data sets without requiring any user input.

ACKNOWLEDGMENTS

This work was supported by an EPSRC Doctoral Training Centre grant (EP/G03690X/1).

REFERENCES

[1] J. A. Hunter, *An Investigation into Partial Discharge Activity within Three-phase Belted Cables*, Ph.D. thesis, School Electronics and Computer Science, University of Southampton, UK, 2013.

[2] G. X. Ritter, J.-A. Nieves-Vazquez and G. Urcid, “A simple statistics-based nearest neighbor cluster detection algorithm,” Pattern Recognition, Vol. 43, pp. 918-932, 2015.

[3] G. C. Montanari, A. Cavallini, F. Puletti and A. Contin, “A new methodology for the identification of PD in electrical apparatus: properties and applications,” IEEE Trans. Dielectr. Electr. Insul., Vol. 12, No. 2, pp. 203-214, 2005.

[4] L. Hao, A. Contin, J. A. Hunter, P. L. Lewin, D. J. Swaffield, C. Walton and M. Michel, “A new method for automatic multiple partial discharge classification,” XVII Int’l. Sympos. High Voltage Eng., Hanover, Germany, paper F-021, 2011.

[5] A. Contin, S. Pastore and R. Paganin, “Evaluation of spaces for the separation of signals due to multiple PD sources,” IEEE Electr. Insul. Conf. (EIC), Seattle, 2015, pp. 209-213.

[6] J. A. Hunter, P. L. Lewin, L. Hao, C. Walton and M. Michel, “Autonomous classification of PD sources within three-phase 11kV PILC cables,” IEEE Trans. Dielectr. Electr. Insul., Vol. 20, No. 6, pp. 2117-2124, 2013.

[7] J. A. Hunter, P. L. Lewin, L. Hao, D. Evagorou, A. Kyprianou and G. E. Georghiou, “Comparison of two partial discharge classification methods,” IEEE Int’l. Sympos. Electr. Insul., San Diego, CA, paper 93, 2010.

[8] A. Contin, A. Cavallini, G. C. Montanari, G. Pasini and F. Putelleti, “Digital detection and fuzzy classification of partial discharge signals,” IEEE Trans. Dielectr. Electr. Insul., Vol. 9, No. 3, pp. 335-348, 2002.

[9] L. Hao, J. Hunter, P. L. Lewin and D. J. Swaffield, “Discrimination of multiple PD sources using wavelet decomposition and principal component analysis,” IEEE Trans. Dielectr. Electr. Insul., Vol. 8, No. 5, pp. 1702-1711, 2011.

[10] X. Ma, C. Zhou and I. J. Kemp, “Interpretation of wavelet analysis and its application in partial discharge detection,” IEEE Trans. Dielectr. Electr. Insul., Vol. 9, No. 3, pp. 446-457, 2002.

[11] X. Ma, C. Zhou and I. J. Kemp, “Automated wavelet selection and thresholding for PD detection,” IEEE Trans. Dielectr. Electr. Insul., Vol. 18, No. 2, pp. 37-45, 2002.

[12] H. Zhang, T. B. Ho, Y. Zhang and M.-S. Lin, “Unsupervised feature extraction for time series clustering using orthogonal wavelet transform,” Informatica, Vol. 30, pp. 305-310, 2006.

[13] P. Montero and J. A. Vilar, “An R package for time series clustering,” J. Statistical Software, Vol. 62, no. 1, pp. 1-11, 2014.

[14] A. Herve and J. W. Lynne, “Principall Component Analysis,” Wiley interdisciplinary reviews: computational statistics, vol. 2, no. 4, pp. 433-459, 2010.

[15] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Machine Learning Research, Vol. 9, pp. 2579-2605, 2008.

[16] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” 2nd International Conference on Knowledge Discovery and Data Mining, Portland, pp. 226-231, 1996.

[17] J. Hunter, P. Lewin, L. Hao, C. Walton and M. Michel, “Autonomous classification of PD sources withing three-phase 11kV PILC cables,” IEEE Trans. Dielectr. Electr. Insul., Vol. 20, No. 6, pp. 2117-2142, 2013.

[18] P. Berkhin, “A Survey of Clustering Data Mining Techniques,” in Grouping multidimensional data - recent advances in clustering, Springer Berlin Heidelberg, pp. 25-71, 2006.

[19] M. Daszykowski, B. Walczak and D. L. Massart, “Looking for natural patterns in analytical data (2) tracing local density with OPTICS,” Chem. Inf. Comput. Sci., Vol. 42, pp. 500-507, 2002.

[20] M. Ankrest, M. Breunig, H. Kriegel and J. Sander, “OPTICS: Ordering Points To Identify the Clustering Structure,” Int’l. Conf. Management of Data, 1999, pp. 500-507.

[21] R. P. W. Duin and E. Pękalska, "Non-Euclidean Dissimilarities: Causes and Informativeness," in Structural, Syntactic, and Statistical Pattern Recognition, Berlin, Springer Berlin Heidelberg, 2010, pp. 324-333.

[22] S.-H. Cha, “Comprehensive survey on distance/similarity measures between probability density functions,” Int’l. J. Mathematical Models and Methods in Appl. Sci., Vol. 1, No. 4, pp. 300-307, 2007.

[23] A. R. Webb, “Measures of Dissimilarity,” *in Statistical pattern recognition*, second edition, John Wiley and Sons, pp. 419-429, 2002.

Robbie D. Nimmo was born in South Africa in 1991, but grew up in Malawi and Tanzania. He received the B.Eng. (Hons) degree in electrical engineering and the M.Sc. in energy and sustainability with electrical power engineering from the University of Southampton in 2014 and 2015, respectively. He was awarded the Power Academy Scholarship from 2011-2015, by Western Power Distribution and the IET. During his degrees he has therefore spent time working in the field of distribution network design, protection, maintenance, and management.

George Callender was born in Basildon, UK in 1991. He received M.Sci. (Hons) degree in natural sciences (maths and physics) from the University of Durham, UK in 2013. He is currently a Ph.D. degree student at the University of Southampton. His research interests include partial discharge modelling and partial discharge source discrimination.

Paul L. Lewin was born in Ilford, Essex in 1964. He received the B.Sc. (Hons) and Ph.D. degrees in electrical engineering from the University of Southampton, UK in 1986 and 1994, respectively. He joined the academic staff of the University in 1989 and is Professor of Electrical Power Engineering in the School of Electronics and Computer Science, where he is also head of the Tony Davies High Voltage Laboratory. His research interests are within the generic areas of applied signal processing and control. Within high voltage engineering this includes condition monitoring of HV cables and plant, surface charge measurement, HV insulation/dielectric materials and applied signal processing. In the area of automation he is particularly interested in the practical application of repetitive control and iterative learning control algorithms. Since 1996 he has received funding and grants in excess of £30M, supervised 45 graduate students to successful completion of their doctoral theses and published over 450 refereed conference and journal papers in these research areas. He is a Chartered Engineer, a Fellow of the IET and IEEE and was the general chair of IEEE International Conference on Solid Dielectrics 2007. He is also President of the IEEE Dielectrics and Electrical Insulation Society as well as an Associate Editor of the IEEE Transactions on Dielectrics and Electrical Insulation.