Paper 1772-2018

Business Customer Value Segmentation for strategic targeting in the utilities industry using SAS

Spyridon Potamitis, Centrica; Paul Malley, Centrica

ABSTRACT

Numerous papers have discussed the importance of businesses understanding the value of their customers, taking advantage of various segmentation techniques to target customers more efficiently, enhance business processes and improve the customer journey. The process of creating a segmentation is often considered as a combination of science and art to deliver meaningful and useful results for the business. Additionally, traditional approaches of dynamically calculating customer value are not always achievable when the data required for the calculations are sparse, involve too much complexity or not accessible. The aim of this paper is to present methodology that can be used to create a Customer Value Segmentation that combines business knowledge with demographic, behavioural and pricing data, utilising SAS® Enterprise Guide and SAS® Enterprise Miner, that will deliver a simple and interpretable analytical solution that can enable useful insight into customer value that can drive appropriate strategies based on their findings.

INTRODUCTION

At a simple starting point, customer value could be defined by a business as how much a customer is worth to them, taking into account a range of incomes and costs. This definition increasingly becomes more complex dependent on what time period and level a business user is interested in defining value, for example, if the value is measured at point-of sale, in-year profit, the duration of a customer's contract, or the entire "lifetime" of the customer's relationship with the business. A commonly used approach to define value is calculating Life Time Value (LTV) for a customer, where typically over a three or five year timescale, expected incomes, costs and discount rate are taken into account.

As well as a monetary perspective, other metrics a business may wish to use to define customer value can include less straightforward quantities, such as customer satisfaction, interactions with the business via digital channels, and so on. For example, a happier customer is more likely to stay with a business, and those customers using digital channels to interact with a business may incur less cost. How these all interconnect into an easily understood customer value segmentation is a non-trivial challenge.

The utilities sector is no different, especially in business-to-business supply where the sizes of customers and the complexity of their energy needs vary enormously compared to the largely flat residential value profile, defining customer value as a combination of energy consumption, how many services a customer has, how much they cost to serve, how happy they are with quality of service, to name but a few. The level of complexity in data adds to the challenge, as some metrics may be specific to the customer, the site energy or services are applied, some customers interact with intermediaries, data on the business' hierarchy also adds complication eg the parent company level, account level, meter level, etc.

Another challenge is, even within the same organization, different teams may define value in their own ways. For example, a Commercial department may utilize LTV methodology in order to validate pricing decisions at the point of sale. The estimate of lifetime value at this point in the customer's journey with the business may not accurately reflect value as time progresses. At later touchpoints, such as point of renewal or marketing campaigns tailored to the customer's specific needs, the customer LTV may not be as useful a tool.

We present here a solution to resolve the aforementioned challenges of defining customer value using SAS Enterprise Guide and SAS Enterprise Miner. The specific example used throughout the paper will focus on B2B (business to business) UK energy and services customers. To complement Pricing's view of lifetime value and to minimize the possibility of multiple versions of the truth, we use Pricing's LTV as

an input variable to our segmentation along with a variety of other key customer metrics. The following example will demonstrate it is possible to create a dynamic solution for effectively identifying the relative value of our business customers that can be used across multiple teams within a business and across the entire customer lifecycle. The familiar K-means clustering algorithm forms the starting point of our Customer Value segmentation, and we discuss the importance of thorough validation of the clusters recommending several criteria in order to select the optimal solution.

DATA AND METHODOLOGY

As with any analytical project, we would typically commence with a project scoping meeting. This is where the analytics team and business stakeholders discuss the purpose, approach and timescales of the project. We recommend holding workshops with business stakeholders to truly understand what customer value means to them. At one such workshop, we created a shortlist of potential variables that could serve as inputs to our segmentation. This also helps business areas buy-in to the segmentation, making them feel comfortable with what the model may consist of and the model outputs. The variables selected were a mixture of numeric and categoric, and the Clustering node used in SAS Enterprise Miner can accommodate both data types.

The next step involves data extraction. Using SAS Enterprise Guide, a representative sample of data including 50,000 customers was selected as of January 2017 along with all the potential input variables that were mentioned above. Once the data has been extracted, we then move to the exploratory data analysis phase, which results in transforming the data if necessary, handling missing data, removing outliers, normalizing data and so on. As a health warning, from our experience, the next step towards creating a useful segmentation involves the most importand part of “art”. It is entirely possible that each variable is handled differently, and how one chooses to transform your potential input variables may impact on the different analytical solutions Through personal preference, we would then recommend importing this data into Enterprise Miner. One of the next steps would be to use the Variable Clustering node to minimize the risk of feeding in correlated variables into the following clustering solution. Variable clustering also helps reduce the dimensionality of data within the model.

The final variables that were pre-selected ready for the Clustering algorithm can be seen in Table 1.

Variable Name	Type	Description
Pricing LTV	Continuous	Pricing LTV is calculated at point of acquisition for each meter and doesn’t change over time
Risk Score	Continuous	Refers to the risk score of a customer
Customer Tenure	Continuous	The time in years the customer has been with us
Customer level of contact	Ordinal	How many times a customer has recently contacted us
Online Account flag	Binary	If a customer has an online account or not
Number of employees	Ordinal	The number of company employees
Intermediary flag	Binary	If a customer has joined via an intermediary or not
Services flag	Binary	If a customer has an active contract with services or not
Industry Type	Nominal	Industry Type of customer’s company

Table 1: Table showing the reduced number of potential variables with their data types and brief descriptions.

Pre-selection of variables is also another part of the “art” before the “science”. We recommend applying logical filters to the potential input variables to avoid forming clusters based on variables that will be difficult aid interpretation, or do not contribute in a meaningful or useful way in the chosen analytical solution.

K-MEANS ALGORITHM

The K-means algorithm is a well known and widely used clustering algorithm. K-means performs well with large datasets but the inputs should be standardised if the analyst requires all the variables to have equal weights, thus allowing the ratios of change of one variable to be comparable to the change in another variable [1]. The Cluster node within SAS Enterprise Miner has the functionality built in for variables’ standardisation and effectively tackles this issue. The K-means algorithm involves the following steps:

1) Randomly select “k” cluster centers (also know as centroids).

2) Calculate the distance between each data point and cluster centres. Then assign each data point to the cluster center with the minimum distance.

3) Recalculate the new cluster centres, and then also recalculate the distance between each data point and the newly obtained cluster centers.

4) Repeat step 2 & 3 until no data point can be reassigned.

The algorithm has been previously discussed in the SAS Global Forums of 2008 and 2013 where the authors suggest that K-means will find the optimal clustering solution based on the number of -clusters that the analyst specifies [2][3]. In most cases, where the analyst does not know beforehand the exact number of clusters they want to create, the analyst has two options:

1. Experiment with different number of k clusters in order to find the optimal solution that is interpretable and meaningful for the business.

2. Use the “Auto” setting in the Clustering node within SAS Enterprise Miner that will find the optimal number of k clusters based on the natural patterns in the data, based on the Cubic Clustering Criterion.

If the “Auto” setting is selected, SAS Enterprise Miner will first perform hierarchical clustering (using the Average Linkage method, Centroid method or Ward method to calculate distances between clusters) to select the optimal number of k before running the K-means process.

We have to point out that even if the “Auto” setting achieves identifying the optimal numbers for k and comes up with well separated and clearly defined clusters, if this approach does not answer the business problem in question, the solution will have very little purpose in business [4].

RESULTS

For the reasons mentioned above we experimented with different number of clusters (Figure 1), while we also tried the “Auto” settings so that we can have a view of the CCC plot (Figure 2).

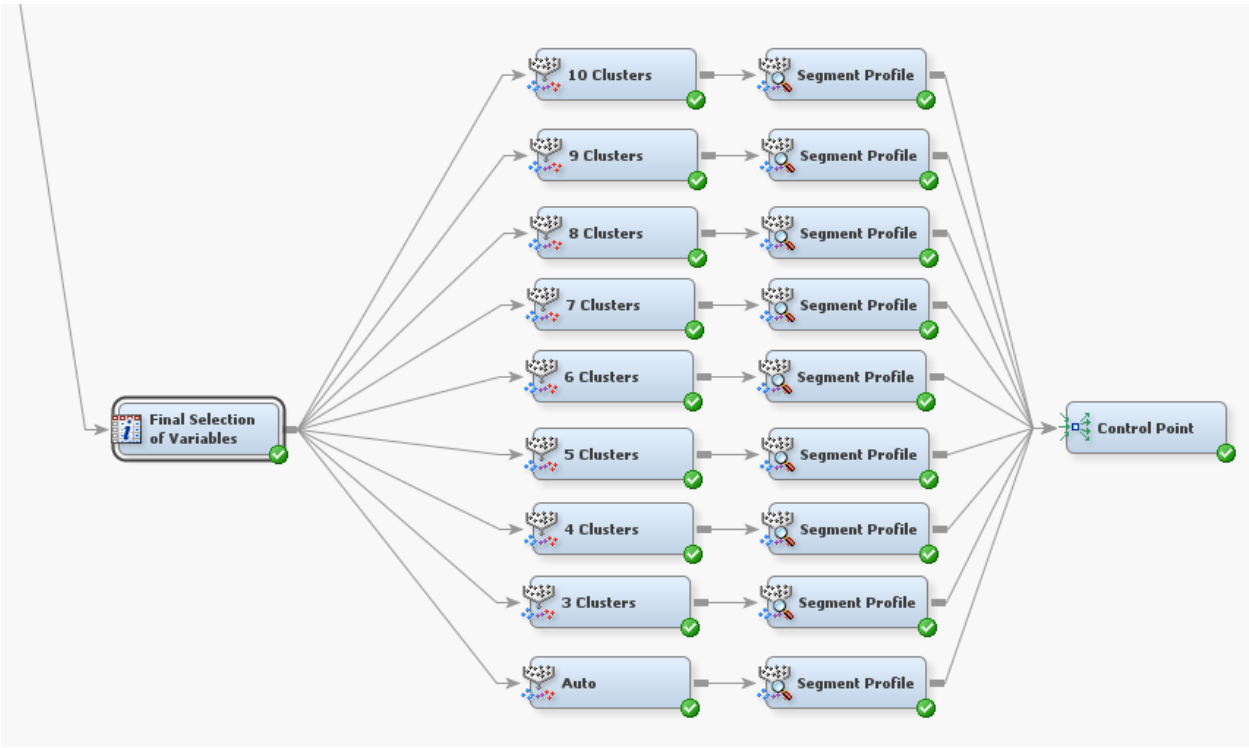

Figure 1: Partial view of Enterprise Miner diagram, starting with the final selection of potential input variables, followed by several Clustering nodes with different value for k clusters, put a Clustering node Auto using the “Auto” setting.

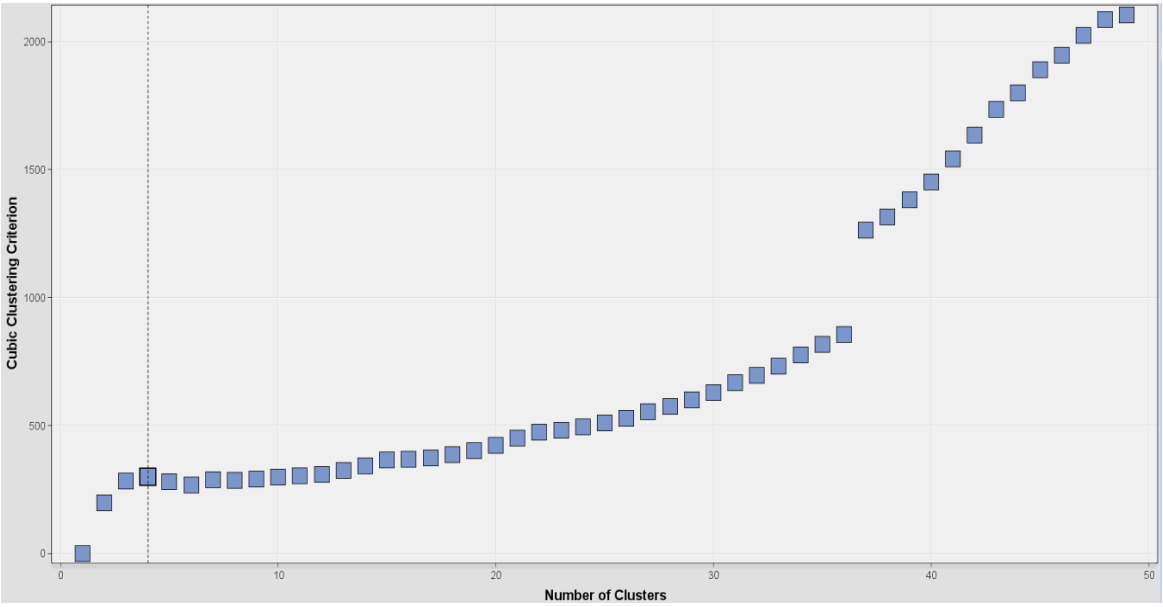

Figure 2: CCC plot in SAS Enterprise Miner derived when using the Auto settings in the Clustering node.

The number of clusters selected using the CCC was four, as it reaches a local maximum (Figure 2) and does not meet any of the following criteria[5]:

- The number of clusters is greater than or equal to the Minimum specified in Selection Criterion properties.

- The number of clusters has CCC values that are greater than the CCC Cut-off specified in the Selection Criterion properties.
- The number of clusters is less than or equal to the final maximum value.
- A peak in the number of clusters exists.

The percentage of the sample that falls into each cluster is visualised with a pie chart as seen in Figure 3. The corresponding percentages are given in Table 2.

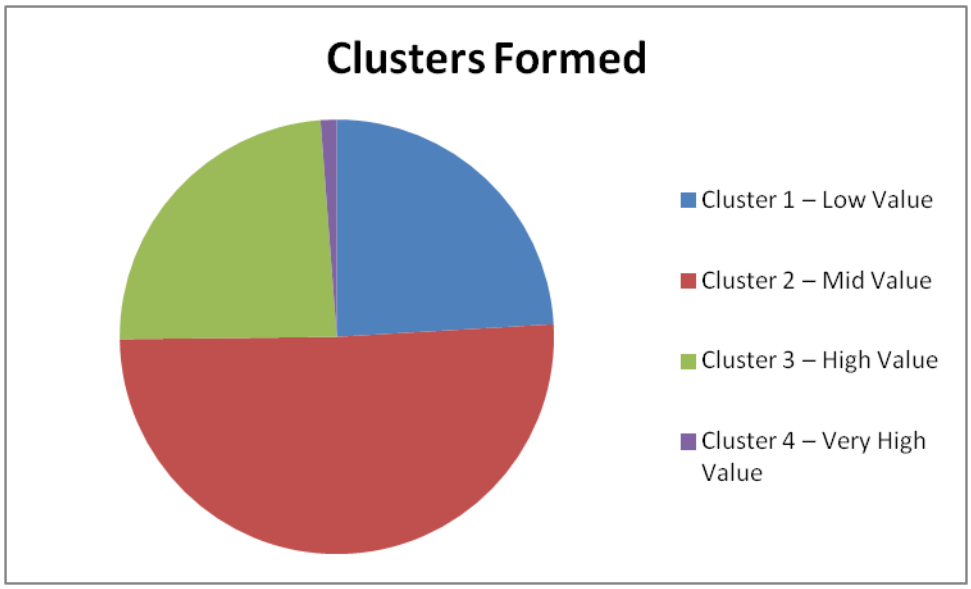

Figure 3: Pie chart showing distribution of volume from the customer segmentation.

Segment	Percentage
Cluster 1 – Low Value	24.1%
Cluster 2 – Mid Value	50.8%
Cluster 3 – High Value	23.9%
Cluster 4 – Very High Value	1.2%

Table 2: Clusters of the segmentation with corresponding percentages of the sample.

Although the Cluster node provides statistics that can be useful in giving the analyst a rough flavour of how the variables' values in the clusters are distributed, in our experience we recommend focusing on interpreting the clusters with the aid of Segment Profile node.

In Figure 4 we can observe the results from the Segment Profile node which shows the distribution of variables in each cluster (blue bars) in comparison with the distribution of the same variables in the overall population (red outline bars). Additionally, the order of the variables in Figure 4 shows the relative importance of each variable in forming the clusters.

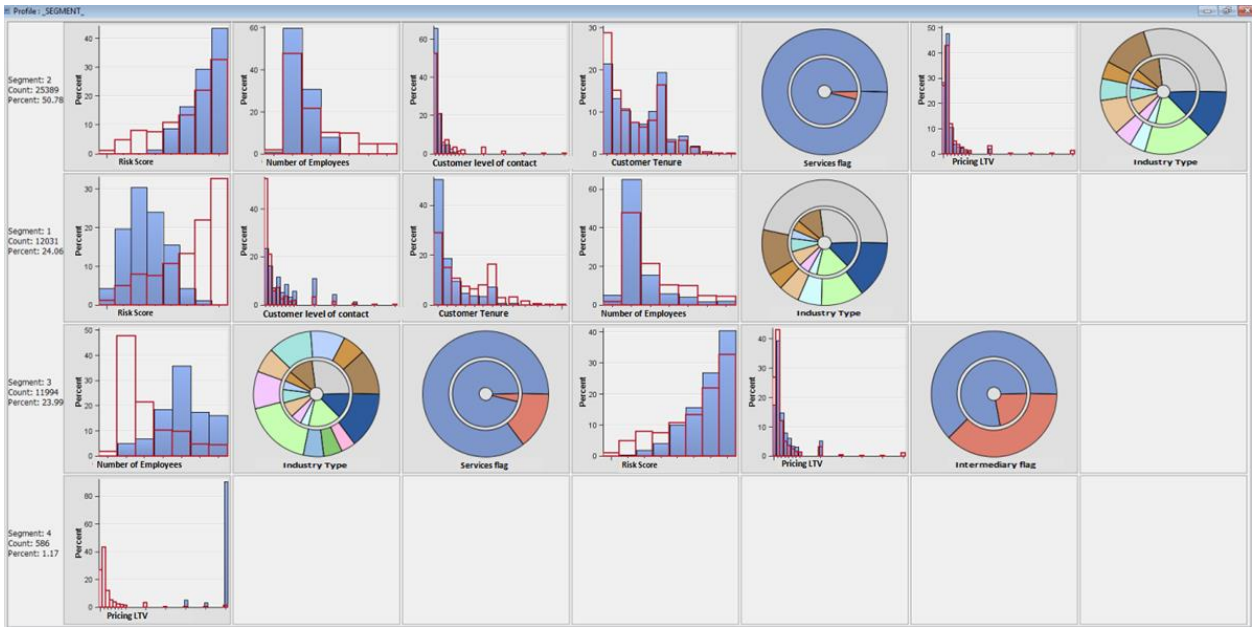

Figure 4: Summary output of the segmentation from the Segment Profile node in SAS Enterprise Miner.

The Segment Profile node is a great tool when trying to give meaningful interpretation to the clusters. By observing each cluster more closely we can get a real sense of who the customers are within each segment, allowing us to label them from Low to Very High value. More specifically:

Cluster 1 – Low Value

- Low Risk Scores indicates that there are more indebted customers or customers that are more probable to be seriously indebted in the future.
- High customer level of contact suggests that they are high cost to serve.
- Customers' tenure is smaller than the overall distribution.
- Most of these customers have low number of employees, suggesting they are small users by commercial energy standards.
- Their industry types are unknown to us in approximately 45% of the cases, significantly higher than the overall sample.

Cluster 2 – Mid Value

- High Risk Scores indicates the majority of customers are paying on time and more regularly.
- Low customer level of contact suggests that they are relatively low cost to serve.
- Customers' tenure is larger than the overall distribution.
- Customers typically do not have additional business services.

Cluster 3 – High Value

- Significantly higher number of employees than the overall distribution.

- Good representation in specific Industry Types known to be very reputable according to our business experts, and also with no issues of missing data.
- High representation of customers with services compared to the overall distribution.
- High involvement with intermediaries which typically represents a higher level of customer engagement with their energy.

Cluster 4 – Very High Value

- Additional analysis showed that the majority of this segment are multi-site customers with significantly high consumption.

From the above analysis we conclude that a sensible segmentation based on the relative value of our customers has been achieved. We stress again the importance of careful consideration of the input variables and how you choose to transform them, as both points do contribute to how meaningful the results are for the business. One example includes the LTV margin variable – the original variable was not normalized but instead a maximum cap was used as we wanted to obtain a cluster formed by extreme and erroneous but possible outliers. It consequently occurs us to develop future strategies based on the customer value segmentation.

How the customer value segmentation can be used is entirely dependent on the creativity of the various business areas of your organization. For example, some possible strategies that could be implemented for the groups of customers that were identified could include:

- Providing cost effective channels of communication for “Low Value” and “Mid Value” segments, given these customers typically prefer a simpler relationship with their energy provider;
- Targetting “High Value” and “Very High Value” segments with offers (e.g. loyalty schemes) to increase their overall retention, again tying in with the customers’ needs and how they wish to manage their energy and services;
- Work with intermediaries that have customers in our more valuable segments to renew early and prioritise them based on the value of customers they bring to the company;
- Enhancing customer experience for the high value segments offering a “red carpet” service.

MONITORING

A crucial component of implementing a customer value segmentation is knowing how the segmentation should be monitored on a regular basis. It is entirely possible that over time the segmentation created may become out of date as the UK market changes and more importantly your customer portfolio changes. Therefore, monitoring on a regular basis, say monthly, is recommended to ensure that the segmentation is still fit for purpose. There are multiple ways to monitor the performance and stability of the segmentation. The metrics that we explored include:

1. Calculating the proportions of population in each segment. As a rule of thumb, we recommend that if we notice a percentage increase of more than 5% in a segment we should investigate further and change accordingly.

2. Calculating the mean, range and standard deviation of the distances (that are produced by the SAS Scoring code, which can be obtained from Cluster node’s results) of the observations from their cluster seeds. If there is a significant difference when we compare the stats above with the results we obtained when we first developed the segmentation, then there is a need to update the solution.

3. Conducting characteristic analysis of the variables that were used in the clustering solution. While suggested methods 1 and 2 would identify that something is not going as expected with the clustering solution, it would be difficult to identify which variable is causing these changes. In order to obtain this metric the variables' values should first be binned and the percentage of population that fall into each bin should be compared with the development dataset. A typical approach could be to look at Characteristic K-S stability metric. If a significant change is identified in a variable, this would mean that the characteristics of the population are changing and we should revisit/update the analytical solution.

4. Conducting further analysis on how the customers migrate from one segment to another over time as some of their characteristics change. This analysis could raise certain practical alerts for the clustering solution (.for example it may appear odd if customers appear to fluctuate as Very High value one month and as Low value the next month) while it could also be particularly interesting in business. In relation to the value segmentation that we developed, it would clearly be of interest to see which variables would influence customers classified as Low value to migrate to the Mid value segment. One possible example could be customer level of contact – if there's an observed association between the number of times a customer contacts us and their classification to Low or Mid value segments, it may be of interest to do a root cause of analysis to understand how those customers could be more valuable due to implementing different contact strategies.

CONCLUSION

In this paper a segmentation based on the relative value of customers was discussed which combines a mixture of data elements using the K-means algorithm that is implemented in SAS Enterprise Miner. The solution is based on the effective collaboration between the analytics team and the relevant business areas to first understand that customer value means to them, and then to create a segmentation that is meaningful, easily interpretable, and actionable with a range of potential applications in the business. The interpretation can come alive when ensuring the candidate input variables are selected thoughtfully, and then transformed carefully before running any algorithms. Our customer value segmentation is another timely reminder that the power of analytics is a healthy mixture of the art and the science.

REFERENCES

[1] Collica, R. S. (2007). Customer segmentation and clustering using SAS Enterprise Miner(Third ed.). Cary, NC: SAS Institute.

[2] Cross G.; Thompson W. 2008 "Understanding Your Customer: Segmentation Techniques for Gaining Customer Insight and Predicting Risk in the Telecom Industry". Proceedings of the SAS Global Forum 2008 Conference. Cary, NC: SAS Institute Inc.

http://www2.sas.com/proceedings/forum2008/154-2008.pdf

[3] Poulsen, R (2013) "Multivariate Statistical Analysis in SAS: Segmentation and Classification of Behavioural Data". Proceedings of the SAS Global Forum 2013 Conference. Cary, NC: SAS Institute Inc.

http://support.sas.com/resources/papers/proceedings13/447-2013.pdf

[4] Linoff, G., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. Indianapolis, IN: Wiley Pub.

[5] Tip: Guidelines for Choosing a Clustering Method in the Cluster Node 2015. SAS Institute, Inc. Cary, NC.

https://communities.sas.com/t5/SAS-Communities-Library/Tip-Guidelines-for-Choosing-a-Clustering-Method-in-the-Cluster/ta-p/223483

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Spyridon Potamitis
sp.potamitis@gmail.com

Paul Malley
paulmalley@hotmail.com