THE GENERATIVE TOPOGRAPHIC MAPPING AS A PRINCIPAL MODEL FOR DATA VISUALIZATION AND MARKET SEGMENTATION: AN ELECTRONIC COMMERCE CASE STUDY

Alfredo Vellido¹, Paulo J.G. Lisboa¹ and Karon Meehan²

¹School of Computing and Mathematical Sciences, Liverpool John Moores University, Byrom St.
Liverpool L3 3AF, UK
²Business School, Liverpool John Moores University, 98 Mount Pleasant, Liverpool L3 5UZ, UK

e-mail: A.Vellido@livjm.ac.uk, P.J.Lisboa@livjm.ac.uk, k.meehan@livjm.ac.uk

Accepted in the final form: 15 November, 2000

Abstract. The process of extracting knowledge from data involves the discovery of patterns of interest which may be implicit, for instance, in specific clusters of data points. In the context of Internet retailing, finding clusters of typical consumer types is among the most important uses of data mining techniques. Cluster-based market segmentation models, grounded on surveys of customer opinion, can give the online retailer a competitive edge, forming the basis for effective targeting and enabling the redirection of made-to-measure content towards the customer. The Generative Topographic Mapping (GTM) is proposed as a statistically principled technique for cluster-based market segmentation. In this non-linear latent variable model, a posterior probability of cluster membership can be defined for each individual, providing a robust framework for the visualization of high dimensional data and the segmentation to different levels of granularity. The advantages of the GTM over the well-known Self-Organizing Map (SOM), to which it is an alternative, are described and this new model is applied in a business-to-consumer e-commerce case study. In addition, an entropy-based measure is defined to quantify the information content of the GTM unsupervised maps about an externally imposed class label.

Keywords: Data clustering and visualization, market segmentation, Generative Topographic Mapping, electronic commerce, entropy, cluster-based information measures.

1 Introduction

The process of extracting knowledge from data involves the discovery of those patterns of interest which they might contain. One type of such patterns is the clusters in which the data points are grouped [7]. A typical cluster analysis application such as market segmentation frequently combines quantitative and qualitative methods, and it has recently been acknowledged that its design and deployment can be more clearly grounded in a sound statistical framework [36]. Also, a key issue in the design of statistical frameworks for data analysis is the reproducibility and robustness of the results obtained. Non-linear models, in particular, do not replicate well when applied repeatedly even to the same data, unless they are implemented within a statistically constrained framework.

This paper introduces the Generative Topographic Mapping (GTM), a non-linear latent variable model developed by Bishop, Svensén and Williams [3,4], as a stable and statistically principled model for data clustering and visualization which addresses critical demands posed by market segmentation.
This model is an alternative to the widely used Kohonen’s Self-Organizing Map (SOM) [18,19] and it is capable of accommodating segmentation strategies of different granularity, from micro-segmentation to the traditional aggregate segmentation. The GTM preserves, in the low-dimensional latent visualization space, the topographic ordering of the points in the multi-dimensional data space. Therefore, it can be used for outliers detection [35] and, in the presence of information over time, to trace and predict the evolution of individuals through the segmentation map. Moreover, the probability density generated by the GTM in data space, if inverted using Bayes’ theorem, produces a posterior probability of cluster/segment membership for each individual, as a result of which this model can be used for fuzzy clustering. Although the proposed methodology for knowledge discovery with the GTM is, in this paper, focused on a particular case study of current interest, as a methodology is widely applicable to the analysis of large databases.

The case study concerns business-to-consumer electronic commerce, utilizing data from the 9th GVU’s WWW Users Survey, produced by the Graphics, Visualization & Usability Center [17]. Part of these data is concerned with the Internet users’ opinions of online shopping. The segmentation study adopts a previously defined data representation using linear factors that have been shown [34] to be predictive of the propensity to buy on-line. The results obtained complement those of a previous study that utilized the SOM model [33].

Turning now to the application area, the interactive nature of the Internet requires a readjustment of the assumed relationship between marketers and customers. They become bound to establish both an economic and a social contract, as the transactions in this medium cannot be discrete but necessarily relationship exchanges [15]. Consumers are empowered by the information richness of the medium, which enables unprecedented availability of access to most aspects of the shopping process. Comparison of products, services and their prices are within everybody’s keystroke reach, whereas opening times and geographical locations are hardly a limitation anymore. In this new commercial framework, those issues entail deep transformations of the marketing mix [5,10] and on-line vendors have to make an extra effort to differentiate their offer and attract the on-line browsers towards their electronic outlets. Cluster-based market segmentation, grounded on surveys of customer opinion, can give the online retailer a competitive edge: the identification of such segments can be the basis for effective targeting, enabling the redirection of made-to-measure content towards the customer. In the context of Internet retailing, the identification of clusters of consumer types has been stated as “the most important use of data mining, as this type of information is useful in a myriad of other planning and development tasks” [31].

The paper is organized in the following manner: The GTM model is described in the second section, where its main advantages over the SOM are summarized. Section 3 outlines the rationale for proposing the GTM as a principled model for data visualization and market segmentation. The case study on e-commerce, including the implementation of the experiment and the presentation and discussion of the results, is analyzed in Section 4. The paper is then closed by a section of general conclusions.

2 The generative topographic mapping model

In this section, the main principles and properties of the Generative Topographic Mapping model are briefly described, followed by a summary of its advantages over the model to which it intends to be a principled alternative: the Self-Organizing Map [18,19].

2.1 Principles of the generative topographic mapping

The Generative Topographic Mapping (GTM) [3,4] is a non-linear latent variable model that generates a probability density in the multi-dimensional data space, using a set of latent variables of smaller dimension. This non-linear mapping is described by the generalized linear regression model

\[ y = W\Phi(u) \]  

(1)
where \( \mathbf{u} \) is an \( L \)-dimensional vector of latent variables, \( \mathbf{W} \) is the matrix that generates the explicit mapping from latent space to an \( L \)-dimensional manifold embedded in data space, and \( \Phi \) is a set of \( R \) basis functions which, in this study, are chosen to be Gaussians. For the non-linear mapping to remain analytically and computationally tractable, and also to elaborate a principled alternative to the SOM, the prior distribution of \( \mathbf{u} \) in latent space is defined as a discrete grid, similar in spirit to the grid of units of the SOM

\[
p(\mathbf{u}) = \frac{1}{M} \sum_{i=1}^{M} \delta(\mathbf{u} - \mathbf{u}_i)
\]

where \( M \) is the number of nodes in the grid. Since the data do not necessarily lie in an \( L \)-dimensional space, it is necessary to make use of a noise model for the distribution of the data points \( \mathbf{x} \). The integration of this data distribution over the latent space distribution, gives

\[
p(\mathbf{x} | \mathbf{W}, \beta) = \int p(\mathbf{x} | \mathbf{u}, \mathbf{W}, \beta) p(\mathbf{u}) \, d\mathbf{u} = \frac{1}{M} \sum_{i=1}^{M} \left( \frac{\beta}{2\pi} \right)^{D/2} \exp \left\{ -\frac{\beta}{2} \| \mathbf{m}_i - \mathbf{x} \|^2 \right\}
\]

where \( D \) is the dimensionality of the data space, and \( \mathbf{m}_i = \mathbf{W} \Phi(\mathbf{u}_i) \) for the discrete node representation (2), according to expression (1). Using the SOM terminology, \( \mathbf{m}_i \) can be considered as reference vectors, each of them the centre of an isotropic Gaussian distribution in data space [4]. A log-likelihood can now be defined as

\[
L(\mathbf{W}, \beta) = \sum_{n=1}^{N} \ln p(\mathbf{x}^n | \mathbf{W}, \beta)
\]

for the whole input data set \{\( \mathbf{x}^n \}\).

The distribution (3) corresponds to a constrained Gaussian mixture model [12], hence its parameters, \( \mathbf{W} \) and \( \beta \), can be determined using the Expectation-Maximization (EM) algorithm [8], details of which can be found in [3]. As part of the Expectation step, the mapping from latent space to data space, defined by (1), can be inverted using Bayes’ theorem so that the posterior probability of a GTM latent space node \( i \), given a data-space point \( \mathbf{x}^n \), is defined as

\[
R_i^n \equiv p(\mathbf{u}_i | \mathbf{x}^n) = \frac{\exp \left\{ -\frac{\beta}{2} \| \mathbf{m}_i - \mathbf{x}^n \|^2 \right\}}{\sum_i \exp \left\{ -\frac{\beta}{2} \| \mathbf{m}_i - \mathbf{x}^n \|^2 \right\}}
\]

This is known as the responsibility taken by each node \( i \) for each point \( n \) in the data space. It will prove itself extremely useful, given a 2-dimensional latent space, for data visualization and also for cluster analysis in the context of market segmentation. The complexity of the mapping generated by the GTM model is mainly controlled by the number and form of the basis functions \( \Phi \). Further control of this effective complexity can be achieved with the addition of a regularization term to the error function (4), in such a way that the training of the GTM would consist of the maximization of a penalized log-likelihood

\[
L(\mathbf{W}, \beta) = \sum_{n=1}^{N} \{ \ln p(\mathbf{x}^n | \mathbf{W}, \beta) \} + \frac{\alpha}{2} \| \mathbf{w} \|^2
\]

where \( \mathbf{w} \) is a vector shaped by concatenation of the different column vectors of the weight matrix \( \mathbf{W} \) and \( \alpha \) is a regularization coefficient. This regularization term is effectively preventing the GTM to fit the noise in the data and is used under the assumption that there exist an underlying data generator which is a combination of the density functions for each of the segments.

The optimum values for all these complexity-controlling parameters should ideally be evaluated in a continuous space of solutions. Given that the GTM is formulated within a probabilistic framework, this can be accomplished using the Bayesian formalism and, more specifically, the evidence approximation [23]. The application of this methodology produces update formulae for the regularisation coefficient.
\( \alpha \) and for the inverse variance of the noise model \( \beta \) [4]. Once the parameters \( \alpha \) and \( \beta \) have been adaptively optimized, the best GTM model (in the sense that it reaches the best compromise between fitting the data and representing the underlying distribution from which the data were generated) can be obtained by experimenting with different combinations of the number of Gaussian basis functions and its width, \( \sigma \). The GTM training process is summarised in Figure 1.

\[ \text{INITIALIZE PARAMETERS and SOM TOPOLOGY} \]  
\[ \text{EXPECTATION} \quad \text{Curves values of the parameters} \quad \text{and used to calculate} \quad \text{convergence (5)} \]  
\[ \text{MAXIMIZATION} \quad \text{Recalculating of} \quad \text{by} \quad \text{maximizing the complete-data log-likelihood} \]  
\[ \text{BAYESIAN REFERENCE} \quad \text{Recalculating of} \quad \text{and posterior for steps of the EM algorithm} \]  
\[ \text{ASSESS the GTM using the EM algorithm, within the Bayesian approach} \]

Figure 1: Summary of the GTM training process.

2.2 Advantages of the GTM over the SOM model

The main advantage of the GTM over the SOM model is that the former generates a density distribution in the input data space so that the model can be described and developed within a principled probabilistic framework. An example of development of the GTM is the use of a Bayesian approach to automatic regularization and smoothing of the resulting mapping. As part of this process, the GTM learning parameters calculation is grounded in a sound theoretical basis. The GTM also provides the well-defined objective function (4), whereas the SOM training does not involve the minimisation of any error function; its maximisation using either standard techniques for non-linear optimisation or the EM-algorithm has been proved to converge, unlike in the case of the SOM. Finally, the magnification factor for the GTM (described in Section 3.2) can be calculated as a continuous function of the latent variables, avoiding the discrete approximation to which the SOM is limited.

3 The GTM as a principled model for data visualization and market segmentation

In this section we intend to put the Generative Topographic Mapping (GTM) forward as a principled model for data visualization and market segmentation. We address the fundamental questions of statistical validation of the clustering model and its suitability to the requirements of the segmentation techniques.

3.1 Justification of the GTM as a principled model for data visualization and market segmentation.

The GTM, as a non-linear latent variable model based on a constrained mixture of Gaussians, whose parameters are estimated with the EM algorithm, is defined within a sound statistical framework. Referring generally to mixture of distribution models, Wedel and Kamakura [36,p.33] argue that “the statistical approach clearly is a major step forward in segmentation research.” That is in contrast with traditional non-overlapping clustering algorithms used for segmentation, grouped in two main types: hierarchical (e.g. Ward’s method) and nonhierarchical (e.g. K-means). These algorithms, as
well as a neural network-based model such as the SOM, are limited by its heuristic nature and their results can not be justified by standard statistical theory.

One of the consequences of the definition of the GTM within a statistical framework is that the posterior probability (5) of each component of the mixture of Gaussians (or node in the latent space), given the data, can be calculated. Each of the nodes in latent space can be interpreted as a cluster, as with the SOM model [28]. In the context of market segmentation, these clusters correspond to segments of the market. Therefore, the GTM model defines a posterior probability of cluster membership for each cluster and each data point. Assuming that each data point belongs only to one cluster (non-overlapping clustering) but the information it conveys is insufficient to uniquely assess its cluster membership [26, 36], the GTM can be said to perform a type of fuzzy-clustering. This posterior probability also provides the model with the data visualization capabilities that other models that project high-dimensional data into a visualization space possess [35].

The mapping defined by (1) ensures the insightful property of topographic ordering: any two points that are close in latent space will be mapped to points that are necessarily close in data space. Consequently, neighboring clusters in latent space correspond to groups of data that are also close in data space. Should the data sets contain information for the same entities at different moments in time, the continuity of this mapping could be used for visualizing their evolution through the map of clusters and even for hypothesizing about their most likely future segment trajectories.

This preservation of the topographic order naturally brings about another feature of the GTM, which is most relevant for its implementation as a market segmentation tool: besides the definition of each node in latent space as a cluster / segment, these nodes can be aggregated in a principled way to form macro-clusters. Therefore, the GTM provides a way to address the always contentious issue of the level of detail or granularity with which the segmentation should be carried out. This matter confronts two opposite views: advocates of personalized and one-to-one marketing argue that it is possible, in the Internet medium, to take to its extreme the segmentation rationale, reaching and targeting segments of one. This should be placed in the context of a post-modern model of markets in increasing fragmentation [9] and ultimately justified by the expected benefits of “reaching individual consumers in order to satisfy their unique needs and wants in the best way” [16]. Against this view, several of its shortcomings have been highlighted, as for instance the lack of economic viability that entails the creation of massive amounts of customized content, and the costs of maintaining, over time, the services for personalized interaction. Also, the potentially counterproductive breach of the privacy sphere involved in the pursuit of detailed information about individuals. The GTM can bridge the gap between these controversial points of view, as it can accommodate cluster-based segmentation strategies of different levels of segment detail. The way the GTM enables the aggregated segmentation to happen is described next.

3.2 Unifying micro- and macro-segmentation strategies: magnification factors and cumulative responsibility for the GTM:

The GTM maps points from a regular grid in latent space into the multi-dimensional data space. In doing so, regions in latent space undergo distortions. The regularity of the latent grid and hence of the visualization map will not necessarily reflect the separation of natural groupings as it is in the data space. The same problem affects the SOM algorithm and, in order to tackle it for this model, Ultsch [32] and Kraaijveld, Mao and Jain [20] suggested visualization strategies based on maps of pseudo-colour levels representing distances between code-book vectors (similar to the GTM reference vectors). Definite macro-clusters are expected to be indicated by separate groupings of close code-book vectors, which can be visually assessed. In the context of the GTM, Bishop, Svensén and Williams [2] have addressed this problem making use of the concept of the magnification factors. It is at this point that the GTM definition of a probability density, in the form of a lower-dimensional manifold embedded in data space, reveals its importance: the local magnification factor can now be calculated, resorting to differential geometry, as a continuous function of the latent space variables.
For the sake of brevity, only the main results from [2] will be presented here. Given a two-dimensional latent space, the GTM will map an infinitesimal rectangle with area \( dA = \prod_i dx_i \) from this space into another infinitesimal rectangle in the manifold embedded in data space and defined by (1), with area. The relation between those areas and, therefore, the magnification factor is proved to be

\[
\frac{dA'}{dA} = |g_{ij}|^{\frac{1}{2}}
\]  

(7)

where \( g_{ij} \) is the metric tensor defined as

\[
g_{ij} = \delta_{kl} \frac{\partial y^k}{\partial x^i} \frac{\partial y^l}{\partial x^j}
\]  

(8)

In matrix form, and making use of (1), the expression (7) can thus be written as

\[
\frac{dA'}{dA} = \left| \Psi^T W^T W \Psi \right|^{\frac{1}{2}}
\]  

(9)

where \( \Psi \) is a matrix of partial derivatives of the basis functions \( \Phi \) with respect to the latent variables. Now the magnification factor \( \frac{dA'}{dA} \) can be plotted in the latent space visualization map, using a pseudo-colour representation. The map areas indicate varying mapping distortions, with extreme shades in a grey-scale representation representing very large or very small distortions. These can be cautiously associated to inter-cluster or within-cluster regions. As a result, the macro-clusters that are a prerequisite for aggregate segmentation can be defined.

In this paper, we propose a new methodology to define macro-clusters from the trained GTM. This methodology allows us to use all the information conveyed by the posterior probability of cluster membership defined in expression (5). Let us define the cumulative posterior probability of cluster membership or cumulative responsibility for a node \( i \) in latent space as

\[
R_{CUM,i} = \sum_n R_{n,i}^{(n)} = \sum_n p(u_i | x^n) = \sum_n \frac{\exp \left\{ -\frac{\beta}{2} \| \mathbf{m}_i - x^n \|^2 \right\}}{\sum_{n'} \exp \left\{ -\frac{\beta}{2} \| \mathbf{m}_i - x^{n'} \|^2 \right\}}
\]  

(10)

Regions of the map with high values of \( R_{CUM,i} \) will roughly correspond with regions of high mode occurrence, unless the distribution is strongly multi-modal. A further advantage of this criterion is that it also indicates the approximate extension of the macro-clusters. The reason is that sparse clusters of points in the multidimensional space are compacted in small areas of the map. Consequently, the posterior probability of a point in such type of cluster is expected to be narrow and concentrated in only a few map units. On the other hand, the mapping of points from compact clusters in multidimensional space will not entail large distortions. These points will occupy larger regions of the map and their posterior probabilities will be broad. Overall, a landscape representation of (10) will be characterized by broad plateaus, corresponding to compact clusters, and narrow high peaks, corresponding to sparse groupings in the data.

With this, we have a reasonable idea of where the macro-clusters lie in the map. Now we intend to confirm that idea using a clustering algorithm on the reference vectors \( \mathbf{m}_i = W \Phi (u_i) \) with no a priori specification of the cluster centroids. Following Murtagh [27], a contiguity-constrained agglomerative algorithm is proposed. As described in previous sections, neighboring conditions will be preserved in the latent space representation, hence the imposition of a contiguity condition. The algorithm will proceed as follows:

1. Each node \( i \) of the map is initialized as a macro-cluster with its center at \( \mathbf{m}_i \), and uniquely labeled.
2. At every step of the algorithm, the two closest neighboring macro-clusters are merged. This merger entails substituting the previous centers with their mean.
3. Repeat step 2 until the macro-cluster partition approximates the representation in Figure 3.
The distances in the algorithm are taken to be Euclidean. The contiguity or neighboring condition in step 2 consists of only considering those macro-clusters that contain neighbor nodes in the map as candidates to be merged. For nodes neither in the edges nor in the corners of the map, the eight surrounding nodes are considered as their neighbors.

Summing up, it has been shown that the GTM addresses the problem of “continuous distribution of heterogeneity versus market segments” [36, p.331] that confronts two different views. The first being where the heterogeneity amongst consumers is not so pronounced that the possibility of grouping them into segments is rejected. The second would consider the partition of the consumer representation continuum into segments as an artifact [1], thus assuming that markets are perfectly heterogeneous. The GTM grid of nodes / clusters in latent space can be modified at will: either augmented so that the distribution of cases in the visualization map becomes sparser (although the differences between nodes / clusters will be smaller), or reduced so that all cases agglomerate in a few clusters (more heterogeneous). On the other hand, the magnification factors and the cumulative responsibility, as defined in this section, help to define macro-clusters of data that can be associated to aggregate market segments.

4 Segmentation of the business-to-consumer electronic commerce market using the GTM

The real-world data used for the study are described in this section, including a summary of results from previous studies. This is followed by the evaluation of the application of the GTM to these data.

4.1 Description of the data and its pre-processing

This study makes use of publicly available data from the web-based 9th GVU’s WWW User Survey [17]. From the first two questions of its “Internet Shopping (Part 1) Questionnaire” (“general opinion of using the WWW for shopping as compared to other means of shopping” and “opinion of Web-based vendors compared to vendors in other forms of shopping”), 44 items were selected.

These items, in a previous study [34], were subjected to factor analysis, for dimensionality reduction and latent variable exploration, and a nine-factor model was obtained following standard goodness-of-fit test procedures. The study intended to assess which factors were more predictive of the propensity to buy online. The data sample contained records from 2180 Internet users and information (in the form of a binary dependent variable) of whether the respondent had or had not ever purchased online. Subsequently, a variable selection method, Automatic Relevance Determination (ARD) [24], associated to a MLP trained within the Bayesian approach, was applied to the resulting factor scores. The five most relevant factors were labeled according to previously published qualitative studies, and they are summarized in Table 1. The application of ARD to these factors produced a ranking, according to which factor 2 (from now on referred to in this study as Risk Perception) in Table 1 bears most of the predictive power of the model, followed by factors 1 and 3 (referred to as Compatibility and Affordability) and finally, by the less relevant factors 4 and 5 (referred to as Ease of use and effort/Responsiveness).

<table>
<thead>
<tr>
<th>FACTOR</th>
<th>DESCRIPTION</th>
<th>ATTRIBUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shopping experience: Compatibility</td>
<td>Control and convenience</td>
</tr>
<tr>
<td>2</td>
<td>Consumer risk perception / Environmental control</td>
<td>Trust and security</td>
</tr>
<tr>
<td>3</td>
<td>Affordability</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>Shopping experience: Effort</td>
<td>Ease of use</td>
</tr>
<tr>
<td>5</td>
<td>Shopping experience / Customer service</td>
<td>Effort / Responsiveness and empathy</td>
</tr>
</tbody>
</table>

Table 1: Descriptive summary of the 5 factors selected by the ARD model as the most predictive of online purchasing behaviour
This previous study also showed that, from a hold-out sample, a MLP trained within the Bayesian approach is able to correctly discriminate between the classes of purchasers and non-purchasers with a maximum of over 81% accuracy.

The practice of factor analyzing the observable data, followed by the clustering of the obtained factor scores is known, in the marketing literature, as the tandem approach towards segmentation. Factor analysis helps to overcome the limitations associated with survey-based segmentation studies, i.e. noisy data, poorly measured variables based on subjective ratings, and unbalanced item selection across the domain of the surveyed constructs [11].

Green & Krieger [11] and Schaffer and Green [29], acknowledge that a good validation of the segmentation results might stem from the relation of alternative clusterings to some profit-based measure. This provides a justification for our use of the tandem approach: in this study, the segmentation is carried out according to factors which have been shown to discriminate, quite accurately, the classes of purchasers and non-purchasers, each of which offers a profit/non-profit opportunity to the marketer. Therefore, although the GTM is a descriptive clustering tool, the whole model is closer to the criterion-based or predictive types of segmentation models such as CHAID [25] or clusterwise regression [37]. The GTM is expected to generate clusters that discriminate between consumers that belong to each of the classes defined by the dependent variable. This property has been observed in the SOM model [30,33]. The resulting cluster solution has an explicitly profit-based interpretation, as each cluster can be described and targeted in a marketing campaign with the combined knowledge of the latent factors that shape it and the propensity to buy which they predict.

4.2 Presentation and discussion of the results

This section gives a brief account of the implementation of the GTM, followed by the presentation and discussion of the segmentation results.

4.2.1 Implementation of the GTM model

The model by which we illustrate the use of the GTM consists of a grid of 5x5 basis functions with a common width $\sigma = 1$. The original sample of 2180 individuals was not balanced in terms of class-membership, so that a class-balanced data set of 778 individuals was randomly selected from which two-thirds of the data were used to train the GTM. The values for the regularization coefficient and the inverse variance of the noise model after training were, in turn, $\alpha = 1.12$ and $\beta = 1.57$. It has to be borne in mind that $\alpha$ and $\beta$ are the result of the GTM training, not fixed or initial values. The purpose of the Bayesian approach is, precisely, the automatic calculation of the optimum values of these parameters. The value of $\sigma$ was set following the guidelines in Bishop et al. (1998b), i.e. selecting the model with a value of $\sigma$ such that it yields the highest log-evidence. A fixed grid in latent space of 15x15 nodes was selected as a compromise on the level of detail or granularity of the cluster solution.

4.2.2 Class membership-related segmentation results

The expression (5) provides a posterior probability of cluster/segment membership for each individual. In order to visualize that probability for a complete set of data, the information has to be somehow summarized. This can be done [3] by calculating the mean of the distribution for each point $x^a$ in data space

$$\langle u | x^a, W^*, \beta^* \rangle = \sum_{i=1}^{M} R_i^n u_i$$

(11)

where $W^*$ and $\beta^*$ are the values of $W$ and $\beta$ for the trained GTM. The mode of the distribution, given by

$$t_{\max} = \arg \max_{\{i\}} R_i^n$$

(12)
for \( i=1, \ldots, M \), can also be used. In this case, the information is visualized in the way that is usual with the SOM model.

Figure 2 represents the whole data set by the mode (12), as mapped onto the nodes of the trained GTM. Each individual has been assigned a color, black or white, depending on its true class membership as described in Section 4.1. Grey nodes indicate that individuals from different classes have been mapped onto them. Figure 2(left) shows that the GTM, without any prior information on class membership has managed to separate, quite clearly, the classes of purchasers and non-purchasers. Such a result partially justifies our semi-predictive approach (as described in Section 4.1) of using factors shown to predict the propensity to buy on-line as the bases for segmentation. This approach would be fully justified if it were shown that the original, observable variables, can not separate both classes with more accuracy than the factor description of the data. A criterion has to be defined to quantify that capability.

The expression (13) will be evaluated for the three different GTM models resulting from the use of: a) the 44 original observable variables mentioned in Section 4.1; b) the complete 9-factor solution prior to factor selection [34]; c) the 5-factor selection used in this study.

\[
S_1(C_1, C_2) = - \sum_{\text{clusters}} P(\text{clusters}) \sum_{\text{classes}} P(\text{class} | \text{clusters}) \ln P(\text{class} | \text{clusters}) = (13)
\]

where \( C_1 \) and \( C_2 \) are, in turn, the classes of purchasers and non-purchasers, and \( p_i^1 \) and \( p_i^2 \) are the ratios of purchasers \( N_i^1 \) and non-purchasers \( N_i^2 \) mapped into the latent-space node \( i \). \( N_i \) is the total number of individuals mapped into node \( i \), so that \( N = \sum_{i=1}^{M} N_i \) is the number of individuals in the complete data set. \( M \) is the total number of latent-space nodes. The minimum entropy value is 0, which corresponds to the ideal case in which every node/cluster has only members of one of the classes mapped into it. The maximum entropy is \( \ln(2) \approx 0.69 \).

The expression (13) will be evaluated for the three different GTM models resulting from the use of: a) the 44 original observable variables mentioned in Section 4.1; b) the complete 9-factor solution prior to factor selection [34]; c) the 5-factor selection used in this study.
Even more important than the entropy measure for the trained GTM itself, is the quantification of the generalization capabilities of this model, i.e. to what extent is it able to discriminate the classes of purchasers and non-purchasers from a sample of consumers with which the model has not been trained (a hold-out sample, with one-third of the data). Although the GTM can be a very powerful exploratory and descriptive tool, it would be of little help unless it was able to generalize, giving the marketer the possibility to assess the propensity to buy on-line of any future potential customers. Therefore, the entropy expression (13) for the hold-out sample will be calculated for the three GTM models described above. The results for both the training and the generalization models are summarized in Table 2.

The use of the original 44 variables yields slightly better training entropy results than any of the factor models. This can be put down to the fact that those 44 variables convey more information than any of the factor models. Nevertheless, the 5-factor selection model is shown to produce a lower generalization entropy, which can be justified on the basis that all the information that is redundant in terms of class-discrimination has been removed from this model. Therefore, the use of the tandem approach as described in Section 4.1 can now be argued to be an appropriate segmentation methodology for market profit optimization.

<table>
<thead>
<tr>
<th>ENTROPY MEASURES</th>
<th>MODELS</th>
<th>5-factor</th>
<th>9-factor</th>
<th>44-variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td></td>
<td>0.3302</td>
<td>0.3504</td>
<td>0.3187</td>
</tr>
<tr>
<td>Test</td>
<td>0.2317</td>
<td>0.2882</td>
<td>0.2512</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Values of the entropy measure (13) for the trained GTM (first row) and for the hold-out sample (second row). The latter are lower because the measure is sensitive to the total number of samples in the map and the size of the hold-out sample was, in this case, half the size of the training sample. The best results have been highlighted.

Figure 3: Responsibility for the trained GTM. Light shades of grey correspond to peaks, i.e. areas of data concentration. Dark shades correspond to valleys, i.e. regions of low data occurrence. This probability distribution can be used as a benchmark for the contiguity constrained clustering algorithm.

4.2.3 Agglomerative segmentation results

Figure 2(right) conveys information about the size of each of the nodes / micro-clusters (number of individuals whose mode corresponds to each of them) showing agglomerations of data in specific
areas. This could provide a starting point for the definition of the macro-clusters of an aggregate segmentation strategy [22]. We apply now the strategy described in Section 3.2. First, the cumulative responsibility is displayed in Figure 3. Notice the resemblance of these contours with the magnification factors displayed in Figure 4. In fact, the cumulative responsibility implicitly contains the information provided by the magnification factors map. These maps provide us with visual clues about the placement and dimensions of the aggregate clusters or macro-segments.

Expression (13) has been introduced to provide a criterion to quantify the class-separation capabilities of cluster representations. Nevertheless, this is not enough to compare solutions with different number of clusters, as those produced by the new cluster-aggregation strategy based on the cumulative responsibility, proposed in Section 3.2. Let us define the entropy due to the cluster partition as

\[ S2 = - \sum_{\text{clusters}} P(\text{cluster}) \ln P(\text{cluster}) = -\sum_{i=1}^{M} \frac{N_i}{N} \ln \frac{N_i}{N} \]  

and the entropy associated to the prior class distributions as

\[ Sl = - \sum_{\text{classes}} P(\text{class}) \ln P(\text{class}) \]  

The behavior of this entropy metric is illustrated with several extreme cases in Table 3. It is clear that a one-to-one correspondence between clusters and classes reduces \( S \) to its absolute minimum of zero, given that the numerators are information divergences, hence positive semi-definite. Therefore, a minimum of \( S \) will correspond to a parsimonious cluster partition.
1. All the data contained in a single cluster  
   \[ S_1 = S_l \]  
   \[ S_2 = 0 \]  
   \[ S = \infty \]  

2. Each cluster is a separate class  
   \[ S_1 = 0 \]  
   \[ S_2 = S_l \]  
   \[ S_1 + S_2 - S_l = 0 \]  
   \[ S = 0 \]  

3. Each data observation is in a separate cluster  
   \[ S_1 = 0 \]  
   \[ S_2 = \ln N \]  
   \[ S = \frac{\ln N - S_l}{S_l} \]  

4. Clusters and classes are completely independent  
   \[ S_1 = S_l \]  
   \[ S_2 = - \sum_{i=1}^{M} \frac{N_i}{N} \ln \frac{N_i}{N} \]  
   \[ S = \frac{S_l}{S_2} + \frac{S_2}{S_l} \]  

Table 3: Extreme cases of the entropy-based measure (17)

Figure 5 shows the values of the entropy for the trained GTM across a range of cluster solutions. The 3-segment solution displayed in Figure 6 generates a minimum (optimum) of \( S \). It strongly resembles the map in Figure 2(left), with two segments in the sides roughly corresponding to the class of non-purchasers, and a segment occupying the centre of the map mainly corresponding to the class of purchasers.
Figure 6: A 3-segment solution as an example of the clustering procedure described in the text. The squares of the map with a dot in the center correspond to GTM nodes with no data point mapped into them.

Figure 7: Reference maps for the trained GTM, associated with each of the factors in the 5-factor selection, in a pseudo-colour representation. Light shades of grey correspond to high values of the elements of the reference vectors, whereas dark shades correspond to low values. The interpretation of those high and low values is reported in Table 4.
<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Compatibility: Control / Convenience</th>
<th>POSITIVE VALUES</th>
<th>NEGATIVE VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>People who perceive shopping on the WWW as compatible and convenient, feeling they are in control of the shopping process.</td>
<td>People who do not perceive shopping on the WWW as either compatible or convenient, or do not feel in control of the shopping process.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor 2</th>
<th>Risk perception: Trust / Security</th>
<th>POSITIVE VALUES</th>
<th>NEGATIVE VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>People who do not perceive that shopping online entails a differential risk. They find online vendors trustworthy.</td>
<td>People who perceive security and economic risks as major deterrents to shop online. They find online vendors untrustworthy.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor 3</th>
<th>Affordability</th>
<th>POSITIVE VALUES</th>
<th>NEGATIVE VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>People who find the whole prerequisites for online shopping as affordable.</td>
<td>People who find the whole prerequisites for online shopping as unaffordable.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor 4</th>
<th>Shopping Experience: Effort / Ease of use</th>
<th>POSITIVE VALUES</th>
<th>NEGATIVE VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>People who find shopping online easy and unproblematic.</td>
<td>People who find shopping online as a complicated undertaking, difficult to learn, that requires a lot of mental effort.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor 5</th>
<th>Shop.Exp.;Customer Service Effort / Responsiveness</th>
<th>POSITIVE VALUES</th>
<th>NEGATIVE VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>People who consider that online vendors provide a responsive customer service, reducing the effort involved in the online shopping process.</td>
<td>People who do not consider that online vendors provide a responsive customer service.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Interpretation of the low and high (negative / positive) values of the factor scores (inputs to the model) and reference vector elements of the trained GTM.

Those segments have to be interpreted in terms of the segmentation bases. For that purpose and, as defined in Section 2.1, the reference vectors \( \mathbf{m}_i = \mathbf{W} \Phi(\mathbf{u}_i) \) in data space, for each node \( i \) in the latent visualization space, will be used. Each of the variables shaping these vectors can be visualized using a reference map, which is a pseudo-color representation of its numerical values. The reference maps for our trained GTM are shown in Figure 6. The meaning of the corresponding numerical values (white for high, black for low in a grey-shaded palette) is described in Table 4.

The segment on the left-hand side of Figure 6 is dominated by low values of risk perception and effort/responsiveness and high values of affordability (See also Table 1 for the description of the factors). It might be characterized as a segment of consumers who, seeing shopping on-line as affordable, are deterred by perceptions of security and economic risks. For the segment on the right-hand side of the map the risk perception seems to be attenuated, whereas the affordability becomes very low; ease of use and effort/responsiveness are also locally very low. The main characteristic of this segment seems to be that the consumers in it do not find shopping on-line economically or otherwise compatible with their current situation. Finally, the central segment, very compact according to the cumulative responsibility and the magnification factor in Figures 3 and 4, has medium to high values for all the factors, indicating that, in general, it corresponds to consumers convinced of the advantages of shopping on-line. That is consistent with the fact that most of those individuals belong to the class of purchasers. A finer segmentation (not optimal from the class-separation point of view but, perhaps, more actionable from a marketing standpoint), using the same aggregation strategy, is exemplified by the 7-segment solution in Figure 8. In this case, it can be seen that the segment solution resembles closely the cumulative responsibility map in Figure 3. The interpretation of the segments according to the reference maps in Figure 7, their labeling and sizes are summarized in Table 5.

Although, for the sake of brevity, not all the results are reported here, different segment solutions from 5 to 10 segments were investigated. A 10-segment solution, for instance, would split the lower part of segment 4 to produce an extra segment that, according to the reference maps, could be labeled as Security and Cost Confident. Furthermore, segment 5 would be split in 3 small groups, two of them
in an intermediate position with respect to segments 3 and 6. This possible split of segment 5 will be further analyzed in the next paragraphs.

![Figure 8: A 7-segment solution obtained with the agglomerative clustering procedure described in the text. It is interpreted in Table 5 with the help of the reference maps from Figure 7. The squares of the map with a dot in the center correspond to those GTM nodes into which no data point has been mapped.]

<table>
<thead>
<tr>
<th>SEGMENT DESCRIPTION</th>
<th>SEGMENT LABEL and SIZE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Very low on Compatibility and rather low on Perception of risk, although rather high on Affordability. (20.2% purchasers – 79.8% non-purchasers)</td>
<td>Unconvinced (14.0%)</td>
</tr>
<tr>
<td>2 High Compatibility compounded with low values of Perception of risk. (29.4% purchasers – 70.6% non-purchasers)</td>
<td>Security conscious (13.1%)</td>
</tr>
<tr>
<td>3 Similar to the Unconvinced segment but scoring higher in the factor of Compatibility. (51.5% purchasers – 48.5% non-purchasers)</td>
<td>Undecided (8.7%)</td>
</tr>
<tr>
<td>4 All factors present medium-to-high values. Values of Perception of risk and Affordability are specially high. (84.3% purchasers – 15.7% non-purchasers)</td>
<td>Convinced (35.0%)</td>
</tr>
<tr>
<td>5 Low Compatibility and very low Ease of use. Medium to high Affordability. (17.1% purchasers – 82.9% non-purchasers)</td>
<td>Complexity avoiders (14.3%)</td>
</tr>
<tr>
<td>6 Most factors present medium values except Affordability, which is very low. (45.2% purchasers – 54.8% non-purchasers)</td>
<td>Cost conscious (10.8%)</td>
</tr>
<tr>
<td>7 This small group scores very low in Effort / responsiveness, but medium-to-high in the remaining factors. (50% purchasers – 50% non-purchasers)</td>
<td>Customer service wary (4.1%)</td>
</tr>
</tbody>
</table>

Table 5: Description of the 7-segment solution. Segments are numbered, according to Figure 8, and described in the second column according to Table 4 (This column also includes the percentages of purchasers and non-purchasers present in each segment). The relative size and proposed label of each segment are included in the third column.
4.2.4 Segment profiling

There are several criteria that define the feasibility of a market segmentation strategy. Amongst others, the substantiability and actionability of the segment solution, which are met by the ones provided in this study. Another two criteria, identifiability and accessibility might sometimes depend on the availability of secondary information, such as demographic and socio-economic data. This type of information has been shown [36] to be neither the most effective in developing segments nor a good predictor of the propensity to buy on-line. Nevertheless, it might sometimes be the only kind of data available to the marketer.

We profile now the 7-segment solution provided in the previous section, making use of the following variables: age, household income, years of Internet experience, average of hours a week of Internet usage and gender. Similar segmentation variables have also been used in other studies [10]. Many conclusions can be drawn from these results. Table 6, though, is quite self-explanatory, so that only some general ideas will be sketched here. Firstly, all the background bases show significant differences over the clusters, as measured by a $\chi^2$ test. Segment 4 (Convinced), mainly composed of purchasers, can be characterized as mostly male in their late twenties to early forties, with high-band income and Internet-savvy. Most of the segments with a majority of non-purchasers tend to be dominated by females, an interesting feature given the level of significance of the gender basis reported in the table. This is especially relevant in the case of the core non-purchaser segment 1, the Unconvinced, because it reveals women as the type of customers that find buying online incompatible with their life and shopping styles (See reference maps, Figure 7).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Whole Sample</th>
<th>Seg 1</th>
<th>Seg 2</th>
<th>Seg 3</th>
<th>Seg 4</th>
<th>Seg 5a</th>
<th>Seg 5b</th>
<th>Seg 5c</th>
<th>Seg 6</th>
<th>Seg 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (A) *</td>
<td>A ≤ 25</td>
<td>26.1</td>
<td>27.5</td>
<td>25.5</td>
<td>35.3</td>
<td>19.8</td>
<td>16.0</td>
<td>24.6</td>
<td>52.4</td>
<td>32.1</td>
</tr>
<tr>
<td></td>
<td>25 &lt; A ≤ 35</td>
<td>25.2</td>
<td>22.0</td>
<td>26.5</td>
<td>30.9</td>
<td>29.3</td>
<td>12.0</td>
<td>16.9</td>
<td>4.8</td>
<td>26.2</td>
</tr>
<tr>
<td></td>
<td>35 &lt; A ≤ 45</td>
<td>23.4</td>
<td>17.4</td>
<td>25.5</td>
<td>19.1</td>
<td>24.2</td>
<td>48.0</td>
<td>23.1</td>
<td>28.6</td>
<td>19.1</td>
</tr>
<tr>
<td></td>
<td>A &gt; 45</td>
<td>25.3</td>
<td>33.1</td>
<td>22.5</td>
<td>14.7</td>
<td>26.7</td>
<td>24.0</td>
<td>35.4</td>
<td>14.2</td>
<td>22.6</td>
</tr>
<tr>
<td>Hours (H) **</td>
<td>H ≤ 10</td>
<td>34.1</td>
<td>31.2</td>
<td>44.1</td>
<td>38.2</td>
<td>22.7</td>
<td>48.0</td>
<td>49.2</td>
<td>52.4</td>
<td>35.7</td>
</tr>
<tr>
<td></td>
<td>10 &lt; H ≤ 20</td>
<td>31.9</td>
<td>35.8</td>
<td>30.4</td>
<td>35.3</td>
<td>28.9</td>
<td>40.0</td>
<td>29.2</td>
<td>23.8</td>
<td>38.1</td>
</tr>
<tr>
<td></td>
<td>H &gt; 20</td>
<td>34.0</td>
<td>33.0</td>
<td>25.5</td>
<td>26.5</td>
<td>48.4</td>
<td>12.0</td>
<td>21.6</td>
<td>23.8</td>
<td>26.2</td>
</tr>
<tr>
<td>Years (Y) **</td>
<td>Y ≤ 1</td>
<td>18.2</td>
<td>23.8</td>
<td>22.5</td>
<td>13.2</td>
<td>6.6</td>
<td>20.0</td>
<td>36.9</td>
<td>42.9</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>1 &lt; Y ≤ 3</td>
<td>38.2</td>
<td>35.8</td>
<td>37.2</td>
<td>42.6</td>
<td>35.5</td>
<td>32.0</td>
<td>43.1</td>
<td>38.1</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>3 &lt; Y ≤ 5</td>
<td>22.4</td>
<td>21.1</td>
<td>24.5</td>
<td>26.5</td>
<td>26.0</td>
<td>28.0</td>
<td>13.8</td>
<td>4.8</td>
<td>16.7</td>
</tr>
<tr>
<td></td>
<td>Y &gt; 5</td>
<td>21.2</td>
<td>19.3</td>
<td>15.8</td>
<td>17.7</td>
<td>31.9</td>
<td>20.0</td>
<td>6.2</td>
<td>14.2</td>
<td>13.1</td>
</tr>
<tr>
<td>Income (I) *</td>
<td>I &lt; 30K</td>
<td>26.1</td>
<td>30.3</td>
<td>24.5</td>
<td>27.9</td>
<td>20.1</td>
<td>4.0</td>
<td>32.3</td>
<td>52.4</td>
<td>33.3</td>
</tr>
<tr>
<td></td>
<td>30K ≤ I &lt; 50K</td>
<td>28.1</td>
<td>23.8</td>
<td>36.3</td>
<td>25.0</td>
<td>23.1</td>
<td>28.0</td>
<td>29.2</td>
<td>23.8</td>
<td>40.5</td>
</tr>
<tr>
<td></td>
<td>50K ≤ I &lt; 70K</td>
<td>18.8</td>
<td>22.0</td>
<td>16.7</td>
<td>16.2</td>
<td>23.8</td>
<td>28.0</td>
<td>12.3</td>
<td>9.5</td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td>I &gt; 70K</td>
<td>27.0</td>
<td>23.9</td>
<td>22.5</td>
<td>30.9</td>
<td>33.0</td>
<td>40.0</td>
<td>26.2</td>
<td>14.3</td>
<td>15.5</td>
</tr>
<tr>
<td>Gender (G) **</td>
<td>Male</td>
<td>51.6</td>
<td>42.2</td>
<td>47.1</td>
<td>47.1</td>
<td>64.5</td>
<td>60.0</td>
<td>40.0</td>
<td>47.6</td>
<td>42.9</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>48.4</td>
<td>57.8</td>
<td>52.9</td>
<td>52.9</td>
<td>35.5</td>
<td>40.0</td>
<td>60.0</td>
<td>52.4</td>
<td>57.1</td>
</tr>
<tr>
<td>Segment Size (abs.)</td>
<td>778</td>
<td>109</td>
<td>102</td>
<td>68</td>
<td>272</td>
<td>25</td>
<td>65</td>
<td>21</td>
<td>84</td>
<td>32</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>14.0</td>
<td>13.1</td>
<td>8.7</td>
<td>35.0</td>
<td>3.2</td>
<td>8.4</td>
<td>2.7</td>
<td>10.8</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Table 6: The bases age, average of hours a week of Internet usage, years of Internet experience, household income and gender are represented, in the left-side of the table, as Age, Hours, Years, Income (in US $) and Gender. *$\chi^2$ test, significant at p<0.05; **significant at p<0.001. All the figures in the table, corresponding to those 5 bases, are percentages of the segment size (or the sample size in the case of the first column). All the segment sizes and the percentage of the sample size they represent are shown in the last row of the table.
The division of segment 5, the Complexity Avoiders, into 3 sub-segments seems strongly justified by their very different profiles. Sub-segment 5a is the top-left part of segment 5 and is close to segment 3, the Undecided, which makes it a somehow “softer” target for potential marketing campaigns. Its profile is most interesting: quite older than the rest of segment 5, more Internet-savvy and with very high average income level. It is also more male-dominated. Sub-segment 5c, in the top-right corner of the map is the hard core of the segment, with a profile that is extremely young, barely Internet experienced and with the lowest average income in the map. Sub-segment 5b is half-way between the other two sub-segments and presents a strongly female profile.

Segment 6, the Cost Conscious, add rather lower incomes (which is consistent with the low values of the affordability factor that the individuals in this segment present) to lack of Internet experience. Segment 3 (Undecided) are similar to the Cost Conscious, with a young profile and average-to-low Internet usage and experience. Their main difference lies in the higher income distribution of the Undecided, which makes this segment highly attractive to marketers. Segment 7 (Customer Service Wary) is almost a variant on the Unconvinced, only younger and with lower Internet usage.

The levels of Internet usage and experience seem to be related to the membership of segments dominated by purchasers, which is reinforced by the results of the test for the corresponding background variables reported in Table 6. This can be understood in the light of the flow construct, developed for the Web computer-mediated environment by Hoffman and Novak [13] to describe the state of mind induced by the Web navigation. The consequences of this flow include “increased learning, increased exploratory and participatory behaviors, and more positive subjective experiences” [14]. The authors characterize consumers’ Web navigation behavior as either goal-directed or exploratory. The first is a “directed search mode ( . . . ) in which the consumer is extrinsically motivated to find a particular site or piece of information in a site”, whereas the second “corresponds to a nondirected, exploratory search mode” [14]. It is argued that early Web users’ flow experiences are generally of the exploratory kind, whereas only experience will lead users to achieve flow experiences in a goal-directed task such as purchasing online.

5 Conclusions

A fully quantitative knowledge discovery methodology has been proposed, which centres on the Generative Topographic Mapping as a tool for data clustering and visualization. It addresses the need to identify distinct clusters within a continuously varying data set [36,p.338], by using statistically principled methods both to organize the data into clusters and to set the cluster boundaries. The probabilistic formulation of this non-linear latent variable model provides estimates for the posterior probability of cluster/segment-membership for each individual, enabling it to be used as a robust tool for micro-segmentation. It also has the useful property of preserving topographic ordering, which, together with the definitions of the magnification factor and the cumulative responsibility, makes the GTM an aggregate segmentation model. Moreover, an entropy-based measure has been defined to quantify the information content of the GTM unsupervised maps about an externally imposed class label. This measure is the sum of the appropriately normalized additional information from adding the external label to the cluster structure, and from imposing the clusters onto the prior class labels. It achieves a global minimum when the two structures match identically.

The application of this methodology to data concerning consumers’ opinions on online shopping demonstrates several useful features of this approach. The GTM shows considerable discrimination of the two main groups of on-line purchasers and non-purchasers without the need of a priori information about class membership of the individuals in the sample. This reinforces similar results obtained with the SOM model [33] and with supervised models [34]. According to the entropy measure, it has been shown that the parsimonious description of the original data, represented by a 5-factor selection model, results in a trained GTM that optimizes the separation of purchasers and non-purchasers in a hold-out sample. This comes to justify the profit-based tandem approach to market segmentation utilized in this study. The entropy measure has been generalized to allow the comparison between
data partitions with different number of segments.

Figures 3 to 8 and Tables 3 to 6 illustrate the use of the GTM for aggregate segmentation. Some patterns of managerial interest can be found here: from the magnification factor and cumulative responsibility maps, together with Figure 2, the group of purchasers seems to be compact and homogeneous, corresponding to the Convinced (segment 1) in Figure 8. Instead, the non-purchasers seem to be much more heterogeneous in nature. Segments shaped mainly by those non-purchasers might be targeted with diverse marketing strategies that address the specific concerns affecting the individuals within them. For instance, and following the definitions in Table 5, Complexity Avoiders may become engaged in the on-line purchasing process through the simplification of its procedures and the utilization, at all steps of the transaction, of user-friendly interfaces. The Undecided might become engaged if the marketer made every effort to develop exchange relationships based upon trust, which, as remarked by Hoffman, Novak and Peralta [15], are more likely to succeed. Strategies of this kind could also be designed to target the rest of those segments with strong non-purchasers presence.

Further research is required in the application of the GTM model with time-dependent market data, given that the most accessible information to Internet retailers is precisely time-dependent Web log data [6]. Therefore, there is a need to accurately model the dynamics of consumers across the map of segments. The calculation of their most likely trajectories would enable the marketer to develop dynamic segmentation forecasting techniques and assess the stability of existing segments.

References


[35] J. Vesanto, SOM-Based Data Visualization Methods, Intelligent Data Analysis, 3(2):111-126, 1999
