# A novel SOM-based method for profile generation: theory and an application in direct marketing

Alex Seret and Sébastien Versailles

19/05/2011

**Abstract**

**The field of direct marketing is constantly searching for new data mining techniques in order to analyze the increasing available amount of data. Self-organizing maps (SOM) have been widely applied and discussed in the literature, since they give the possibility to reduce the complexity of a high dimensional attribute space while providing a powerful visual exploration facility. Combined with clustering techniques and the extraction of the so-called salient dimensions, it is possible for a direct marketer to gain a high level insight about a dataset of prospects. In this paper, a SOM-based profile generator is presented, consisting of a generic method leading to value-adding and business-oriented profiles for targeting individuals with predefined characteristics. Moreover, the proposed method is applied in detail to a concrete case study from the concert industry. The performance of the method is then illustrated and discussed and possible future research tracks are outlined.**

## 1 Introduction

The explosive growth of the amount of available data and the reliance on data mining techniques have led to the creation of a myriad of new business models and opportunities. The field of direct marketing is not an exception and explores ways of getting competitive advantages by supporting research on the development of innovative and value-adding techniques. Self-organizing maps (SOM) are one of these techniques and have been applied for as many purposes as domains. Giving a powerful encapsulated facility for the analysis of complex databases by reducing the curse of dimensionality, this technique provides the direct marketer with the required tools to take accurate, quick and value-adding decisions. The inexhaustible source of applications has been widely discussed in the literature and has been combined with existing techniques in order to verify the statement that *The whole is greater than the sum of its parts.* Combined with clustering techniques, it is then possible to widen the scope of the analysis and obtain a better insight about the studied data. The extraction of so-called salient dimensions permits the direct marketer to identify and segment its prospects.

In this paper, the authors propose a generic method aiming at generating profiles based on the SOM technology and the extraction of salient dimensions, enabling the direct marketer to formalize his feelings and insights on a dataset while generating value-adding and business-oriented profiles which target individuals with predefined characteristics. The developed generic method is applied to a real life case study, conducted in cooperation with Ticketmatic, a Belgian provider of ticketing solutions. Data from the concert industry is analyzed, and the performance of the proposed method is discussed and challenged in order to evaluate its potential while identifying further interesting research topics.

This paper is structured as follows. Section 2 provides the necessary background about customer segmentation and direct marketing, introducing the concepts of segmentation bases, customer profitability, the RFM framework and two techniques of segmentation, namely self-organizing maps and salient dimensions extraction. In Section 3 the SOM-based profile generator is presented and completed with ad hoc definitions of performance measures. Section 4 presents an application of the proposed method and an analysis of the performance of the generated profiles. A more extensive discussion of the impact of different parameters on the performance, the managerial aspects and different topics for further research is to be found in Section 5.

## 2 Customer segmentation & Direct marketing

In the field of direct marketing, many techniques have been used to identify the most profitable customers, or which of the customers are most likely to respond to a specific campaign. However, such analyses only enable the direct marketer to predict the behavior of the already known customers. A more interesting goal for customer segmentation is the identification of customer profiles, so that one can predict the behavior of unknown customers. With such customer profiles, interesting applications in direct marketing emerge, such

| Paper | Segmentation Bases | | | | | Main technique(s) |
|---|---|---|---|---|---|---|
| | Demographic | Geographic | Psychographic | Behavioral | | |
| | | | | RFM | Other | |
| Chan (2008) | x | | | x | | Genetic algorithm |
| Suh et al. (1999) | x | x | | x | x | Neural network, Logistic regression |
| Kim et al. (2006) | x | | x | | x | Decision tree |
| Füller and Matzler (2008) | x | | x | | x | Cluster analysis |
| Hwang et al. (2004) | x | | | | x | Logistic regression, Neural network, Decision tree |
| Jonker et al. (2004) | | | | x | | Genetic algorithm, Chi-squared automatic interaction detector |
| Lee and Park (2005) | x | | | | x | Self-organizing maps, C4.5 |
| Baesens et al. (2002) | | | | x | x | Bayesian neural network |
| Baesens et al. (2004) | | | | | x | Bayesian neural network |
| Hsieh (2004) | x | x | | x | | Self-organizing maps, Apriori |
| Sohrabi and Khanlari (2007) | | | | x | x | K-means |
| Liu and Shih (2005) | | | | x | | K-means |
| McCarty and Hastak (2007) | x | | x | x | | Chi-squared automatic interaction detector, Logistic regression |

Table 1: Segmentation bases and techniques in the customer segmentation literature.

as targeting specific geographic zones or social groups (e.g. readers of a certain journal, listeners of a certain radio channel, etc.). Whether or not the main goal of the segmentation is to build customer profiles, two major characteristics have to be defined: the segmentation bases and the technique used to identify segments. Table 1 presents an overview of the customer segmentation literature, focused on these two characteristics.

## 2.1 Segmentation bases

As presented in Table 1, many variables have been used to serve as bases for customer segmentation. Four categories of segmentation variables are identified by Kotler et al. (2006):

- *Demographic variables* are the most used variables, as they are the easiest to collect and generally provide satisfying results. Among others, this category includes information about the age, gender, income and religion of the customer.
- A second commonly used category gathers *geographic variables* about the client, including the distance between the client and the buying place, his country, or his address.
- The third category contains *psychographic variables* such as the lifestyle, the personality and the attitudes of the customer. These are much harder to measure, and are much less used in the literature.
- The last, but not the least important category identified by Kotler et al. (2006) is the one including *behavioral variables.* With the emergence of information systems, enterprises have much information about a client's purchases at their disposal, and it has become much easier to infer client behavior. Among these behavioral variables, Kotler et al. (2006) and Michiels (2008) distinguish preferences of the customer, benefice sought, loyalty and profitability, the latter being discussed more deeply in Section 2.2.

## 2.2 Customer profitability & The RFM framework

Kim et al. (2006) identify three ways to segment a customer list using a specific measure of profitability, the Lifetime Value (LTV), which can be generalized to any measure of profitability.

1. The first and most simple use of such a measure is to order the customer list by descending order of profitability and target the marketing spending on the first customers of the list.
2. A second way is to perform segmentation with the different dimensions of the concerned measure as segmentation bases. (Hwang et al. (2004); Werner and Kumar (2000))
3. The third method, identified by Kim et al. (2006), uses a profitability measure as one of the dimensions of the segmentation, along with other variables such as e.g. demographic variables. As it allows for capturing more complex correlations, this approach has been adopted in the remainder of this paper.

All of these practices require a method to measure profitability. There exist numerous methods in the literature and, in particular, the concept of LTV has been defined, implemented, and discussed in many occasions and ways (Hwang et al. (2004) and Sohrabi and Khanlari (2007) are two excellent examples).

The method that will be used to measure customer profitability in this study is the Recency-Frequency-Monetary (RFM) framework. It is a well known technique widely used and studied in the literature (See Table 1). It is based on the following three variables:

- Recency: When was the last purchase of the customer?

- Frequency: How often has the customer bought?
- Monetary: How much money has the customer spent?

Although these concepts are not precisely defined, the RFM framework is well appreciated because it is easy to use and interpret. Different operationalizations of the three variables can be found in the literature as discussed by e.g. Jonker et al. (2004), and Baesens et al. (2002). Other notable points of divergence in the use of the RFM framework are: (1) the relative weights attributed to the three variables and (2) the decision whether or not to aggregate the three variables.

## 2.3 Techniques for customer segmentation

An overview of techniques for customer segmentation which are used in the literature is presented in Table 1. This section will focus on the two major techniques used in this study, namely self-organizing maps (SOM) and salient dimension extraction.

### 2.3.1 Self-organizing maps

Kohonen maps, also called self-organizing maps (SOM), have been introduced in 1981 by Kohonen. Fields like data exploratory analysis, industrial and medical diagnostics, speech analysis and corruption analysis (Huysmans et al. (2006)) are contemporary examples of SOM analysis applications and successes. This section is based on Kohonen (1995) and aims at giving a theoretical background to the reader. An application of the technique can be found in Section 4.1. The main objective of the SOM algorithm is the representation of a high dimensional input dataset on lower dimensional maps. This gives the possibility to explore the data and to use techniques like visual correlation analysis or clustering analysis in an intuitive manner. To do so, a feedforward Neural Network (NN) is trained on the input data. The output layer is a map with a lower dimensionality and a given number of neurons. During each iteration of the algorithm, an input data vector $n_i$ is compared with the neurons $m_r$ of the output map using Euclidian distances. The neuron $m_c$ with the smallest distance with regard to the input vector is identified as the Best Matching Unit (BMU):

$$\|n_i - m_c\| = \min_r\{\,\|n_i - m_r\|\}. \tag{1}$$

The weights of the BMU are then modified in the direction of the input vector, leading to a self-organizing structure of the neurons. A learning rate $\alpha(t)$ and a neighborhood function $h_{cr}(t)$ are defined as parameters of the learning function:

$$m_r(t+1) = m_r(t) + \alpha(t)h_{cr}(t)[n(t) - m_r(t)]. \tag{2}$$

The learning-rate will influence the magnitude of the BMU's adaptation after matching with an input vector $n_i$, whereas the neighborhood function defines the range of influence of the adaptation. In order to guarantee the stability of the final output map, decreasing learning rates and neighborhood functions are often used at the end of the training. An exhaustive discussion of the influence of the parameters such as the number of neurons, the shape of the map, or the initial weights of the neurons is to be found in Kohonen (1995).

### 2.3.2 Salient dimensions extraction

Extracting salient dimensions (SD) for automatic SOM labeling is a methodology developed by Azcarraga et al. (2005) and aims at identifying salient dimensions for clusters of SOM nodes. These salient dimensions are then used to label a SOM in an unsupervised way. The methodology is based on five main stages and starts with the training of a SOM using preprocessed data normalized within an input range of 0 to 1, followed by the clustering of the resulting nodes using any clustering technique. Pruning the nodes within the different clusters will lead to more homogeneous clusters and is the aim of the second step. This pruning phase is based on the mean and the standard deviation of the Euclidian distance between the centroid and the neurons of the different clusters. A parameter $z_1$ is used to identify the neurons to be pruned (the outliers or unlabeled neurons) and the neurons to be kept. The higher the value of $z_1$, the smaller the number of neurons pruned. The third step consists of identifying two sets for each cluster. On one hand the in-patterns set is defined and gathers all the individual training patterns belonging to the cluster. On the other hand, the out-patterns set consists of all the individual training patterns belonging to the other clusters or being attached to an unlabeled neuron identified in the second step. Using the sets defined in the previous step, the salient dimensions can then be identified for the clusters using a measure of deviation in the statistical sense of the term. A difference factor is calculated for each dimension of all clusters and is used to identify the salient dimensions. A second parameter, $z_2$, is used to build a confidence interval around the mean of the difference factors of a cluster. A salient dimension will then be a dimension $d$, belonging to the set $D$ gathering all the dimensions, for which the difference factor differs too much with regard to other dimensions within a cluster:

$$|df(k,d) - \mu_{df}| \geq z_2\sigma_{df}(k), \tag{3}$$

with $df(k,d)$ being the difference factor for the dimension $d$ of the cluster $k$, and $\mu_{df}(k)$ and $\sigma_{df}(k)$ respectively the mean and the standard deviation of the difference factors of the cluster $k$. The smaller the value of $z_2$, the larger the number of salient dimensions identified. The final step uses the different salient dimensions to label clusters with input from domain-specific experts. The result gives the possibility to label a new pattern using the label of the cluster to which it is attached. The formulas leading to the different statistics are to be found in detail in Azcarraga et al. (2005) and are discussed and adapted in Section 3.1.

# 3 SOM-based profile generator

This section is composed of Section 3.1 which proposes a generic method aiming at generating profiles based on the SOM approach and the salient dimensions extraction and Section 3.2 which defines different measures of performance applicable to the generated profiles.

## 3.1 The method

The general idea of the SOM-based profile generator consists of 5 main steps: (1) The generation of indices; (2) SOM training; (3) Clustering and SD extraction; (4) Generation of profiles; and (5) Ranking of profiles. Figure 1 schematizes these steps which are discussed in detail in what follows.

The first step consists of the preparation of the different indices that will be used during the training of the SOM. Categorical variables are preferred because of the way of using the extraction of the salient dimensions. Using continuous variables, it can only be defined whether a variable has high or low values with regard to other clusters. It is then better to categorize the continuous variables, giving the possibility to identify one or more of these categories as salient for a given cluster. Thus, a set $N$ of input vectors $n_i$ with $|D|$ dimensions is obtained. The value assigned to a dimension of $n_i$ is either 1, if the input vector is characterized by the given dimension, or 0 if not.

During the second step, a SOM is trained using normalized values in the range of 0 to 1 as described in Azcarraga et al. (2005). The reader interested in the details of the parametrization of a SOM analysis is referred to Kohonen (1995).

In the third step, the output map of the previous step is clustered using any clustering technique, e.g. the $k$-means clustering widely discussed in the literature (e.g. Tan et al. (2006)). The method for extracting salient dimensions is a special case of the method developed in Azcarraga et al. (2005) with the first parameter, $z_1$, tending to infinity and the second parameter, $z_2$, being equal to zero. It corresponds to a case where no pruning of the clusters is performed and where all dimensions are either positive or negative salient dimensions. The in-patterns set $\phi_{in}(k)$ of the cluster $k$ is defined as the set of all individual training patterns belonging to the cluster $k$:

$$n_i \in \phi_{in}(k) \Leftrightarrow \forall j \in K, \min_j(dist(c_j, n_i)) = dist(c_k, n_i), \tag{4}$$

with $dist(c_j, n_i)$ the Euclidian distance between the centroid of cluster $j$ and the individual training pattern $n_i$, and $K$ the set of all clusters identified using $k$-means clustering. The out-patterns set $\phi_{out}(k)$ of cluster $k$ is computed by subtracting the in-patterns set of cluster $k$ from the set $N$ of all individual training patterns:

$$\phi_{out}(k) = N \backslash \phi_{in}(k). \tag{5}$$

In order to identify the salient dimensions, the following steps have to be processed for each cluster (steps 1, 2 and 3 being adapted from Azcarraga et al. (2005)):

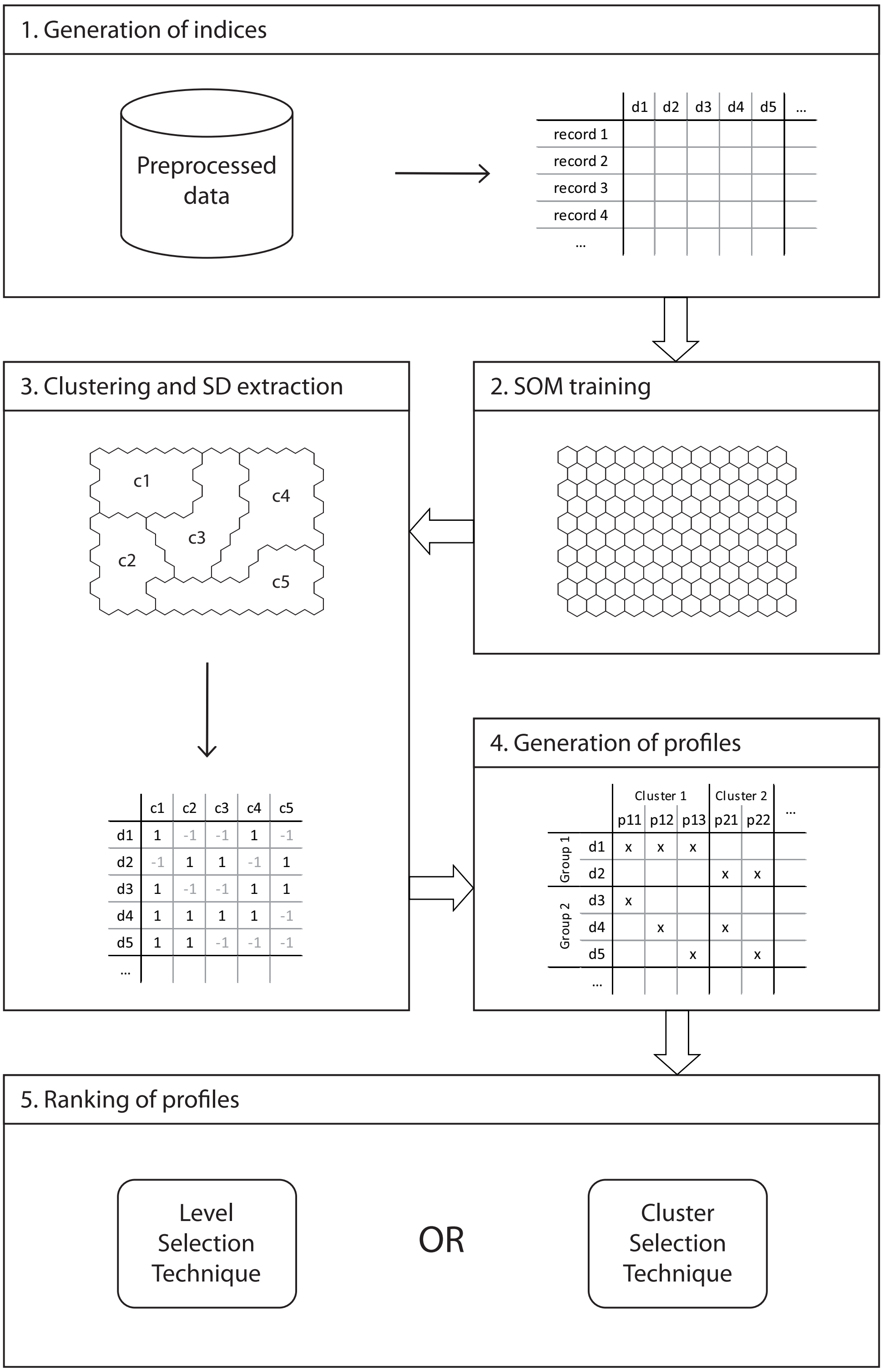
1. For each dimension $d$, compute $\mu_{in}(k,d)$ and $\mu_{out}(k,d)$ as respectively the mean input value for the set of in-patterns $\phi_{in}(k)$ and out-patterns $\phi_{out}(k)$, where $n_{id}$ is the $d$th component of the input vector $n_i$:

$$\mu_{in}(k,d) = \frac{\sum_{n_i \in \phi_{in}(k)} n_{id}}{|\phi_{in}(k)|}, \tag{6}$$

$$\mu_{out}(k,d) = \frac{\sum_{n_i \in \phi_{out}(k)} n_{id}}{|\phi_{out}(k)|}. \tag{7}$$

2. Compute the difference factor $df(k,d)$ of each dimension $d$ as:

$$df(k,d) = \frac{\mu_{in}(k,d) - \mu_{out}(k,d)}{\mu_{out}(k,d)}. \tag{8}$$



1. Generation of indices
Preprocessed data
2. SOM training
3. Clustering and SD extraction
c1
c2
c3
c4
c5
4. Generation of profiles
5. Ranking of profiles
Level Selection Technique
OR
Cluster Selection Technique

| | d1 | d2 | d3 | d4 | d5 | ... |
|---|---|---|---|---|---|---|
| record 1 | | | | | | |
| record 2 | | | | | | |
| record 3 | | | | | | |
| record 4 | | | | | | |
| ... | | | | | | |

| | c1 | c2 | c3 | c4 | c5 |
|---|---|---|---|---|---|
| d1 | 1 | -1 | -1 | 1 | -1 |
| d2 | -1 | 1 | 1 | -1 | 1 |
| d3 | 1 | -1 | -1 | 1 | 1 |
| d4 | 1 | 1 | 1 | 1 | -1 |
| d5 | 1 | 1 | -1 | -1 | -1 |
| ... | | | | | |

| | | Cluster 1 | | | Cluster 2 | | ... |
|---|---|---|---|---|---|---|---|
| | | p11 | p12 | p13 | p21 | p22 | |
| Group 1 | d1 | x | x | x | | | |
| | d2 | | | | x | x | |
| Group 2 | d3 | x | | | | | |
| | d4 | | x | | x | | |
| | d5 | | | x | | x | |
| | ... | | | | | | |

Figure 1: Figure schematizing the five steps of the SOM-based profile generator.

3. Compute the difference factors mean $\mu_{df}(k)$ over all dimensions $d$ as:

$$\mu_{df}(k) = \frac{\left(\sum_{d=1}^{D} df(k,d)\right)}{|D|}. \tag{9}$$

4. The salient dimension sign $sds(k,d)$ of the cluster $k$ for the dimension $d$ can then be computed as:

$$sds(k,d) = 1 \text{ if } df(k,d) \geq \mu_{df}(k), \tag{10}$$

or

$$sds(k,d) = -1 \text{ if } df(k,d) < \mu_{df}(k). \tag{11}$$

Based on the salient dimension signs, the profiles, each consisting of a set of dimensions, for a given set of targeted dimensions $T$ are generated by using Algorithm 1, newly proposed in this paper. This is the fourth step of the method.

**Algorithm 1** Generation of profiles

```
Define a set $PD$ as a subset of $D$, containing the dimensions to be involved in the profile generation.
Define a set $G$ of non overlapping groups[1]of dimensions from $PD$ so that there is no dimension of $PD$ not belonging to one group of $G$ and no dimension of $PD$ belonging to two different groups of $G$.
Define a set $T$ of targeted dimensions such that each group of $G$ is at most represented by one dimension in $T$.
Define a set of targeted groups $TG$ composed of all groups having one dimension in $T$.
Define a set of untargeted groups $UG$ composed of all groups of $G$ not belonging to $TG$.
Define a set of selected clusters $SC$ composed of the clusters having a positive salient dimension sign for all the targeted dimensions of the set $T$.
Create a list $LSC$ composed of the clusters of the set $SC$.
Assign a score computed as the sum of the difference factors of the dimensions in $T$ for each cluster $k \in LSC$ and rank the clusters in $LSC$ in a decreasing order of their respective scores.
for all cluster $k$ in $SC$ do
    Identify a set $PC^k$ of all the possible combinations of the dimensions belonging to the groups of $UG$ and having positive salient dimension signs, with maximum one dimension in each group of $UG$ and minimum one dimension in a group of $UG$ if there is at least a dimension in that group with a positive salient dimension.
    Create a list $LPC^k$ composed of the combinations of $PC^k$.
    Assign a score to each combination in $LPC^k$ computed as the sum of the difference factors of the dimensions involved in each combination and rank the combinations in $LPC^k$ in a decreasing order of their respective scores.
end for
```

The fifth and last step of the method consists of the ranking of the profiles generated in the previous step using one of the two priority rules implemented by Algorithms 2 and 3, Cluster selection technique (CST) and Level selection technique (LST).

**Algorithm 2** Cluster selection technique (CST)

```
 1: An empty list of selected profiles $LSP$ is created.
 2: if the list $LSC$ is empty then
 3:     The algorithm stops and the list of selected profiles $LSP$ is returned.
 4: else
 5:     Select the first cluster sc of $LSC$.
 6:     while $LPC^k$ is not empty do
 7:         Add the first combination of $LPC^k$ to $LSP$.
 8:         Remove the first combination of $LPC^k$ from $LPC^k$.
 9:     end while
10:     Remove sc from $LSC$ and go back to line 2.
11: end if
```

[1] A group is defined as a set of dimensions having a real-world meaning to the user. An example of such a group could be the different dimensions resulting from the categorization of a variable (e.g. the variable would be the age variable, whereas the categories, such as [18..25], [26..35], [36..50], [51..65], and [66..], would be the dimensions of the group).

**Algorithm 3** Level selection technique (LST)

```
1:  An empty list of selected profiles $LSP$ is created.
2:  if the list $LSC$ is empty then
3:      The algorithm stops and the list of selected profiles $LSP$ is returned.
4:  else
5:      Select the first cluster $k$ of $LSC$.
6:      Add the first combination of $LPC^k$ to $LSP$.
7:      Remove the first combination of $LPC^k$ from $LPC^k$.
8:      if $LPC^k$ is not empty then
9:          Rank $k$ at the last position of $LSC$.
10:     else
11:         Remove $k$ from $LSC$.
12:     end if
13:     Go to line 2.
14: end if
```

Step four and five are contributions of this paper and allow for the identification of profiles which have certain characteristics. The input for these steps is the output of a SOM analysis and SD extraction. The final result is a list $LSP$ of combinations of dimensions whose groups belong to $UG$. These combinations are the profiles containing the required targeted dimensions and are ranked in $LSP$ according to some idea of importance expressed in the chosen ranking technique.

## 3.2 Performance measures

The performance measure used to evaluate the performance of the generated profiles is expressed as the ratio between the degree of matching of the profiles generated in the previous section with a testing dataset $TN$ and the degree of matching with $TN$ of randomly generated profiles used as benchmark. Different performance measures are needed in order to express this gain.

A matching function $\theta\left(n_i, p\right)$ is used to express the similarity between a given input vector $n_i$, element of $TN$, and a profile $p$, element of the list of selected profiles $LSP$:

$$\theta\left(n_i, p\right) = 1 \Leftrightarrow \text{ all the dimensions of p are equal to 1 in } n_i, \tag{12}$$

or

$$\theta\left(n_i, p\right) = 0 \Leftrightarrow \text{ at least one of the dimensions of p is equal to 0 in } n_i. \tag{13}$$

A second matching function $\lambda\left(n_i, sLSP\right)$ is defined and expresses whether or not at least one profile $p$ belonging to a subset $sLSP$ of $LSP$ matches with an input vector $n_i$:

$$\lambda\left(n_i, sLSP\right) = 1 \Leftrightarrow \exists p \in sLSP : \theta\left(n_i, p\right) = 1, \tag{14}$$

or

$$\lambda\left(n_i, sLSP\right) = 0 \Leftrightarrow \neg\exists p \in sLSP : \theta\left(n_i, p\right) = 1. \tag{15}$$

A matching ratio $\alpha\left(TN, sLSP\right)$ is then defined and returns the proportion of input vectors in $TN$ matching with at least one profile $p$ in $sLSP$:

$$\alpha\left(TN, sLSP\right) = \frac{\sum_{n_i \in TN} \lambda\left(n_i, sLSP\right)}{|TN|}. \tag{16}$$

The random performance $\chi\left(p\right)$ is defined as the probability of matching when using a randomly generated profile having 1 dimension in each group of $UG$ for which $p$ has a dimension:

$$\chi\left(p\right) = \prod_{d \in p} \frac{1}{|g(d)|}, \tag{17}$$

with $d$ a dimension of $p$, and $g(d)$ the group to which $d$ belongs. Using the random performance function $\chi\left(p\right)$, a second random performance function $\beta\left(sLSP\right)$ is defined as:

$$\beta(sLSP) = \sum_{p \in sLSP} \chi\left(p\right). \tag{18}$$

Finally, the gain $\pi(sLSP)$ obtained when using a given subset $sLSP$ on a testing dataset $TN$ can be computed as:

$$\pi(TN, sLSP) = \frac{\alpha\left(TN, sLSP\right)}{\beta(sLSP)}. \tag{19}$$

| Variable name | Description | Type |
|---|---|---|
| ticketID | The ID of the ticket. | integer |
| creationDate | The date of the purchase. | date |
| customerID | The ID of the customer related to the ticket. | integer |
| basketID | The ID of the basket related to the ticket | integer |
| totalAmountBasket | The price of the basket the ticket is related to. | float |
| concertName | The name of the concert the ticket gives access to. | string |
| geoLong | The longitude from which the customer purchased the ticket. | float |
| geoLat | The lattitude from which the customer purchased the ticket. | float |
| birthdate | The birthdate of the customer. | date |
| gender | The gender of the customer. | char |

Table 2: Summary of the information about sold tickets.

# 4 Market segmentation in the concert industry: an application of the SOM-based profile generator

This section presents a case study carried out in collaboration with Ticketmatic, one of the leading ticketing software companies in Europe. To apply the SOM-based profile generator, online ticket sales data of one concert organizer collected during the period 2007-2010 was used. The data consisted of 63.000 records gathering information about sold tickets as summarized in Table 2.

The application programming interface (API) provided by the *last.fm* website[2] was used to gather tags about artists involved in the different concerts. These tags are provided by the *last.fm*'s users and are gathered in a database accessible via the API.

The preprocessing consisted of the selection of all records having values for all the attributes of Table 2 combined with a rudimentary outlier detection procedure using the mean and standard deviation of the geographic coordinates to identify geographically isolated customers. A further preprocessing step was to prune all records related to concerts of which the artists were not tagged in the *last.fm* API.

## 4.1 Application of the profiling method

The objective was to profile the customers of the available dataset in order to predict the profiles of the customers potentially interested in a specific future concert, which will be called *The Concert* in the remainder of this paper. To do so, the preprocessed dataset was divided in two subdatasets. The first one was composed of all the records related to tickets sold for *The Concert* and was used as test dataset while the second one was composed of all records related to tickets sold for other concerts and was used as training dataset. Starting with the preprocessed data about the sold tickets described in Table 2, three categories of indices were developed for all customers of both subdatasets. The three RFM variables were constructed for each customer, giving the possibility to rank them based on a score ranging from 1 to 5. An index *total RFM* was then computed by summing the three values of the RFM variables, leading to a score between 3 and 15 for each customer. Four categories of customers were defined based on their respective *total RFM* indices using the following intervals: [3..5], [6..8], [9..11] and [12..15]. The birthdate of the different customers was used to generate an index capturing the age. Five categories were defined using the following intervals in years: [18..25], [26..35], [36..50], [51..65] and [66..]. The gender of the customers was used to build two extra categories. The information captured under the variables *geolong* and *geolat*, using the IP addresses of the booking computers, gave the possibility to obtain an index representing the geographic distance separating the customer from the concert infrastructure. Categories were defined using following intervals in km: [0..5], [6..10], [11.15], [16..25], [26..50] and [51..]. In order to define an index capturing the interest of a given customer for a tag, an artist or a concert, the *last.fm* API data was used in combination with data of the previous concerts involving the given customer. Considering that a concert consists of a series of artists characterized by a series of tags, it is then possible to rank the customers according to their score for a given concert as described in Algorithm 4. Note that this interest-based

[2] http://www.lastfm.fr/api

| Dimension | Index category | Original variable | Range | Index name | Value |
|---|---|---|---|---|---|
| D1 | Demographic | Sex | M | Sex Man | $\{0,1\}$ |
| D2 | Demographic | Sex | F | Sex Woman | $\{0,1\}$ |
| D3 | Demographic | Age | 18..25 | Age 18-25 | $\{0,1\}$ |
| D4 | Demographic | Age | 25..35 | Age 25-35 | $\{0,1\}$ |
| D5 | Demographic | Age | 35..50 | Age 35-50 | $\{0,1\}$ |
| D6 | Demographic | Age | 50..56 | Age 50-56 | $\{0,1\}$ |
| D7 | Demographic | Age | 65.. | Age 65-more | $\{0,1\}$ |
| D8 | Demographic | Distance | 0..5 | Distance 0-5 | $\{0,1\}$ |
| D9 | Demographic | Distance | 5..10 | Distance 5-10 | $\{0,1\}$ |
| D10 | Demographic | Distance | 10..15 | Distance 10-15 | $\{0,1\}$ |
| D11 | Demographic | Distance | 15..25 | Distance 15-25 | $\{0,1\}$ |
| D12 | Demographic | Distance | 25..50 | Distance 25-50 | $\{0,1\}$ |
| D13 | Demographic | Distance | 50.. | Distance 50-more | $\{0,1\}$ |
| D14 | RFM | Total rfm | 3..5 | Total rfm 1 | $\{0,1\}$ |
| D15 | RFM | Total rfm | 6..8 | Total rfm 2 | $\{0,1\}$ |
| D16 | RFM | Total rfm | 9..11 | Total rfm 3 | $\{0,1\}$ |
| D17 | RFM | Total rfm | 12..15 | Total rfm 4 | $\{0,1\}$ |
| D18 | Interest-based | The Concert | 0.. | The Concert | [0..1] |

Table 3: Summary of the indices generated in the first step of the SOM-based profile generator.

variable is based solely on tags of previously attended concerts, and not on information of *The Concert* itself. In the application, the score for a given concert, *The Concert*, was used as last index and, combined with the other indices of the previous categories, led to a total of 18 dimensions as summarized in Table 3.

**Algorithm 4** Score of a given customer for a given concert

```
Define the sets $C$, $R$, $A$ and $T$ as respectively the sets of all the customers, concerts, artists and tags.
Select a concert $r$ from $R$.
Select a customer $c$ from $C$ the score of which must be calculated for the concert $r$.
for all $t \in T$ do
    Define a set $A^t$ composed of all the artists in $A$ having $t$ as one of their tags.
    Define a set $R^t$ composed of all the concerts in $R$ having minimum one artist in $A^t$.
end for
Define a set $R^c$ composed of all the concerts in $R$ customer $c$ attended.
Define a set $A^r$ composed of all the artists in $A$ related to $r$.
for all $a \in A^r$ do
    Define a set $T^a$ composed of all the tags $t$ from $T$ related to $a$.
end for
Compute the score of the customer $c$ for the concert $r$ as: $score(c,r) = \sum_{a \in A^r} \sum_{t \in T^a} \frac{|R^t \cap R^c|}{|R^c|}$
return $score(c,r)$
```

A 10x12 SOM was trained, using as input vectors the normalized values in the range of 0 to 1 of the 18 dimensions for each customer of the training dataset. A $k$-means clustering was performed on the neurons of the generated SOM starting with $k$ equal to 10 and selecting the best $k$ using the Davies-Bouldin index[3] value as decision criterion. The Davies-Bouldin index led to the decision to choose a $k$ of 9 as shown in Figure 2. Next, a salient dimension analysis was performed using the adapted method presented in Section 3.1. Table 4 shows the results of the difference factor computation for the different clusters and dimensions, to be compared with the mean difference factor of each cluster as explained in Section 3.1. Using Formulas 10 and 11 of Section 3.1, the $sds(k,v)$ values were computed for each dimension and for each cluster as shown in Table 4 where a bold number represents a $sds(k,v)$ equal to one and a number in normal script a $sds(k,v)$ equal to zero. Then, Algorithm 1 was applied using the results of the previous steps. A detailed sequence of its application is to be found in what follows:

1. $PD = \{D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}, D_{14}, D_{15}, D_{16}, D_{17}, D_{18}\}$
2. $G = \{\{D_1, D_2\}, \{D_3, D_4, D_5, D_6, D_7\}, \{D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}\}, \{D_{14}, D_{15}, D_{16}, D_{17}\}, \{D_{18}\}\}$

[3] The interested reader is referred to Davies and Bouldin (1979) for more information about the Davies-Bouldin index.

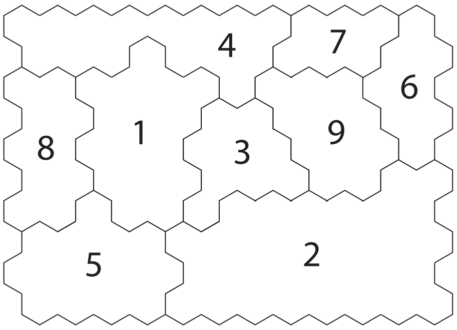


Figure 2: Visualization of the clustering of the 10x12 SOM leading to nine clusters.

| | Clusters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dim | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| D1 | **0,73** | -0,93 | **0,49** | **0,79** | -0,94 | **0,67** | **0,70** | **0,52** | 0,57 |
| D2 | -0,93 | **3,74** | -0,71 | -1,00 | **2,03** | -0,95 | -1,00 | -0,73 | -0,78 |
| D3 | -0,72 | -0,40 | -0,87 | **0,38** | **1,77** | -0,54 | **0,57** | -0,30 | **1,07** |
| D4 | **1,46** | -0,28 | **1,26** | -0,55 | **0,14** | -1,00 | -0,99 | **1,08** | -0,72 |
| D5 | -0,88 | **1,22** | -0,66 | **0,37** | -1,00 | **1,97** | **1,24** | -1,00 | -0,31 |
| D6 | -0,62 | -0,57 | -0,88 | **0,71** | -0,11 | **0,27** | -0,15 | -0,22 | **3,29** |
| D7 | -0,74 | -0,83 | -0,61 | **0,49** | -0,21 | -0,21 | **0,45** | -0,82 | **7,14** |
| D8 | -0,05 | 0,13 | **0,28** | -0,64 | 0,01 | **0,16** | -0,52 | **0,48** | 0,13 |
| D9 | -0,31 | **0,53** | -0,26 | 0,01 | -0,32 | **0,48** | -0,12 | -0,44 | 0,08 |
| D10 | -0,21 | 0,07 | -0,19 | **0,17** | -0,06 | **0,18** | 0,01 | **-0,09** | 0,09 |
| D11 | -0,03 | 0,01 | **-0,03** | 0,04 | -0,07 | -0,23 | **0,23** | **0,20** | -0,01 |
| D12 | **0,09** | -0,14 | **-0,01** | **0,65** | **0,09** | -0,20 | **0,37** | -0,61 | -0,10 |
| D13 | **0,33** | -0,34 | -0,23 | **0,21** | **0,16** | -0,12 | **0,27** | **0,37** | -0,19 |
| D14 | **2,14** | 0,07 | -0,53 | -0,59 | -0,57 | -0,38 | **1,96** | -0,86 | -0,30 |
| D15 | -0,61 | -0,44 | -0,74 | **1,51** | **1,58** | -1,00 | -0,74 | **1,78** | -0,78 |
| D16 | -0,74 | **0,60** | **1,89** | -0,91 | -1,00 | **1,89** | -0,66 | -1,00 | **1,63** |
| D17 | **0,81** | **0,85** | -0,39 | -0,85 | -0,69 | **0,89** | 0,05 | -0,94 | 0,10 |
| D18 | -0,13 | 0,10 | **-0,05** | -0,02 | **0,09** | 0,00 | -0,03 | **-0,02** | -0,05 |
| $\mu_{df}$ | -0,02 | 0,19 | -0,12 | 0,04 | 0,05 | 0,10 | 0,09 | -0,14 | 0,60 |

Table 4: Table of the difference factors for each dimension and for each cluster, with bold numbers indicating positive salient dimensions signs.

3. $T = \{D_{18}\}$

4. $TG = \{D_{18}\}$

5. $UG = \{\{D_1, D_2\}, \{D_3, D_4, D_5, D_6, D_7\}, \{D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}\}, \{D_{14}, D_{15}, D_{16}, D_{17}\}\}$

6. $SC = \{3, 5, 8\}$

7. $LSC = (3, 5, 8)$

8. $score(3) = -0.051$
   $score(5) = 0.093$
   $score(8) = -0.016$

   $LSC = (5, 8, 3)$

9. (a) $PC^3 = \{\{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{11}, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\}\}$
   (b) $LPC^3 = (\{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{11}, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\})$
   (c) $score(\{D_1, D_4, D_8, D_{16}\}) = 3,920$
   $score(\{D_1, D_4, D_{11}, D_{16}\}) = 3,607$
   $score(\{D_1, D_4, D_{12}, D_{16}\}) = 3,626$

   $LPC^3 = (\{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\}, \{D_1, D_4, D_{11}, D_{16}\})$

   (a) $PC^5 = \{\{D_2, D_3, D_{12}, D_{15}\}, \{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\}, \{D_2, D_4, D_{13}, D_{15}\}\}$
   (b) $LPC^5 = (\{D_2, D_3, D_{12}, D_{15}\}, \{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\}, \{D_2, D_4, D_{13}, D_{15}\})$
   (c) $score(\{D_2, D_3, D_{12}, D_{15}\}) = 5,470$
   $score(\{D_2, D_3, D_{13}, D_{15}\}) = 5,539$
   $score(\{D_2, D_4, D_{12}, D_{15}\}) = 3,833$
   $score(\{D_2, D_4, D_{13}, D_{15}\}) = 3,903$

   $LPC^5 = (\{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_3, D_{12}, D_{15}\}, \{D_2, D_4, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\})$

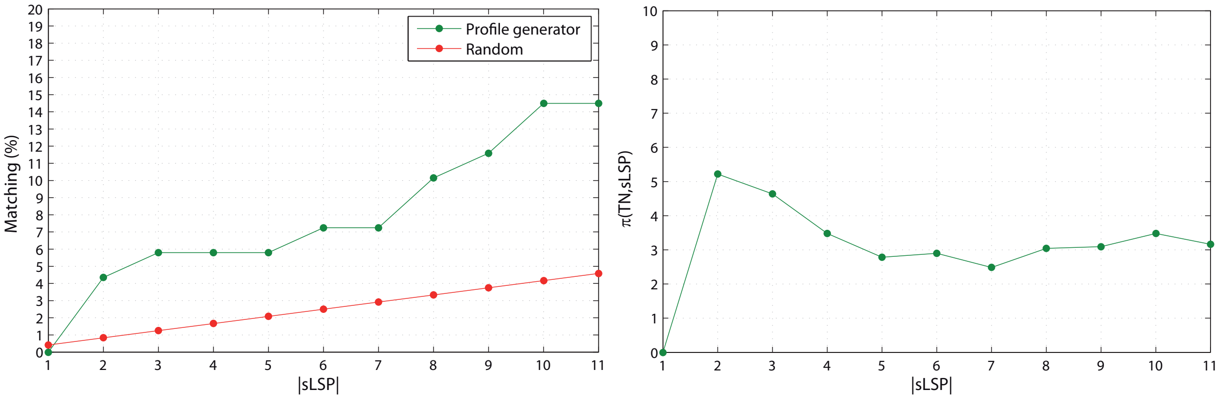
   (a) $PC^8 = \{\{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\}\}$
   (b) $LPC^8 = (\{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\})$
   (c) $score(\{D_1, D_4, D_8, D_{15}\}) = 3,856$
   $score(\{D_1, D_4, D_{10}, D_{15}\}) = 3,287$
   $score(\{D_1, D_4, D_{11}, D_{15}\}) = 3,580$
   $score(\{D_1, D_4, D_{13}, D_{15}\}) = 3,752$

   $LPC^8 = (\{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\})$

After this step, the three lists $LPC^3$, $LPC^5$ and $LPC^8$ contained the generated profiles targeting the dimensions $d$ belonging to $T$. So far, the profiles of the customers interested in *The Concert* had been generated and ranked within each cluster and the clusters had been ordered. The final step consisted of selecting the profiles in order to generate a priority list $LSP$ as described in Section 3.1. The interested reader can apply both ranking techniques on the output of the previous step and compare the results with the following resulting priority lists of selected profiles $LSP$'s, generated using respectively CST and LST.

1. Cluster Selection Technique: $LSP = (\{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_3, D_{12}, D_{15}\}, \{D_2, D_4, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\}, \{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\}, \{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\}, \{D_1, D_4, D_{11}, D_{16}\})$

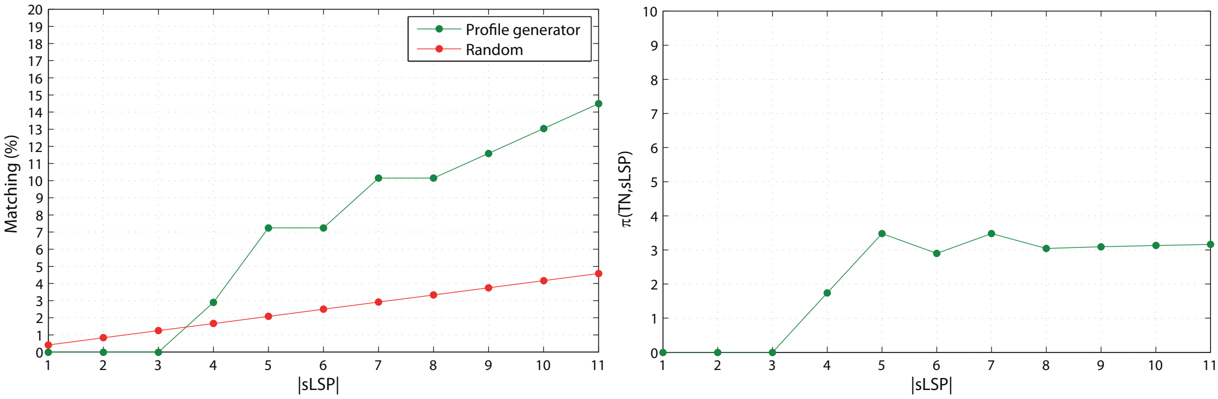
2. Level Selection Technique: $LSP = (\{D_2, D_3, D_{13}, D_{15}\}, \{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_8, D_{16}\}, \{D_2, D_3, D_{12}, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\}, \{D_1, D_4, D_{12}, D_{16}\}, \{D_2, D_4, D_{13}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{11}, D_{16}\}, \{D_2, D_4, D_{12}, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\})$



(a) Performance for The Concert using the LST

(b) Gain for The Concert using the LST

Figure 3: Performance and gain for The Concert using the LST.



(a) Performance for The Concert using the CST

(b) Gain for The Concert using the CST

Figure 4: Performance and gain for The Concert using CST.

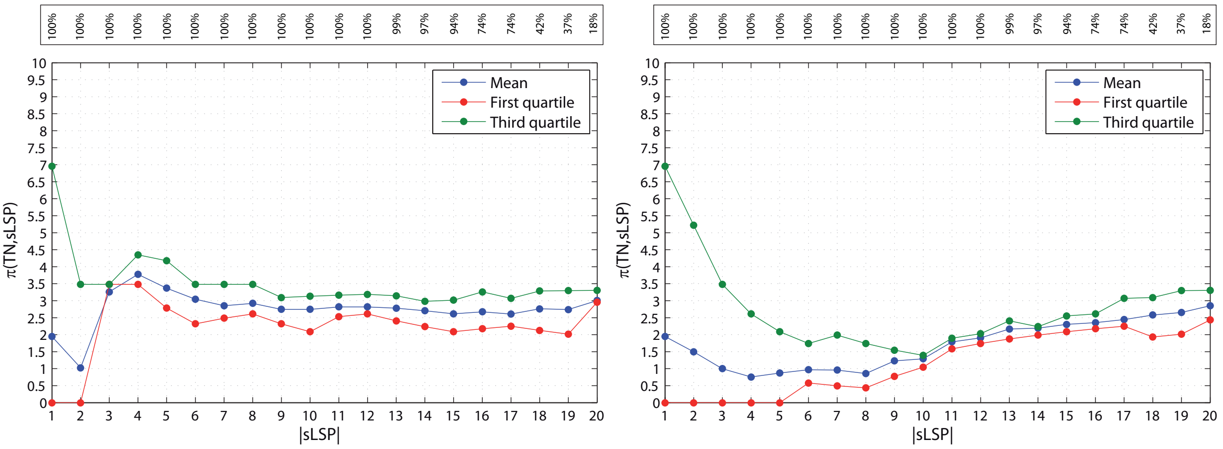
## 4.2 Performance of the generated profiles

The aim of this section is to evaluate the profiles generated by the method developed in this paper. As introduced in Section 4, the profiles were generated using the dataset not incorporating tickets sold for the concert we targeted in order to test the algorithm on an independent test set. The second dataset will now be used to evaluate the quality of the generated profiles. The generation of indices is performed on the subdataset gathering the records of tickets sold for *The Concert* in the way described in Section 4. The 18th dimension concerning the interest for the concert is not considered in this section for obvious reasons leading to a test set $TN$ of input vectors with 17 dimensions corresponding to all the customers related to the tickets sold for *The Concert*.

The performance of the profile generator can be compared with the performance of a random tool using the defined measures of performance. Figure 3(a) shows the matching function $\alpha\left(TN, sLSP\right)$ and the random performance function $\beta\left(sLSP\right)$ for different subsets $sLSP$. The number of profiles in $\beta\left(sLSP\right)$ is given on the X axis, whereas the Y axis represents the difference between $\alpha\left(TN, sLSP\right)$ and $\beta\left(sLSP\right)$, giving the added value of the profile generator with regard to a random prediction. The ratio of both $\alpha\left(TN, sLSP\right)$ and $\beta\left(sLSP\right)$ on the other hand gives the gain $\pi(TN, sLSP)$, expressing how many times the profile generator is better, or worse, than the random profile generator. Figure 3(b) is a graph of the gain $\pi(TN, sLSP)$ for the different numbers of selected profiles in $sLSP$. A value greater than 1 implies an improvement of the prediction power when the profile generator is used. Figures 3(a) and 3(b) show that, using the LST, two profiles are needed to outperform a random generation of the profiles.

Figures 4(a) and 4(b) show the same functions as in Figure 3(a) and 3(b) when the profile generator is used combined with CST. It is obvious that the order of selection of the profiles influences the performance as can be seen when comparing Figures 3(b) and 4(b). When CST is used, four profiles are needed to outperform the random instead of the two needed when using LST.

The previous graphs showed the performance of the profiles generated using the SOM-based profile generator while targeting *The Concert*. However, a certain variability of the results is to be captured because of the clustering technique used in our application, the $k$-means clustering. The number of clusters $k$ and the form



(a) Gain for The Concert using the LST

(b) Gain for The Concert using the CST

Figure 5: Gain for The Concert using LST and CST.

of the clusters have indeed a huge impact on the generated profiles and the $k$-means technique provides no guarantee concerning the stability of the clusters generated. The point is thus to know whether the obtained results are average results or singularities. In order to give an answer to this question, the method was applied 100 times using the same data and targeting the same concert. The results are shown on Figures 5(a) and 5(b), for LST and CST respectively. The mean gain obtained for a subset of profiles, the size of which is given on the X axis, and the first and third quartile values are plotted. So far, an important consideration is that some clustering iterations lead to different numbers of profiles. In order to capture the fact that some statistics are not representative for the 100 iterations, this representativity is expressed as a percentage of the executions of the method having generated at least a number $|sLSP|$ of profiles, and is indicated on the top of each subfigure for each number of profiles $|sLSP|$. For this representativity reason, the statistics up to a $|sLSP|$ of 20 profiles have been plotted. A discussion on the impact of factors such as the selection technique used or the amount of available data is to be found in the next section.

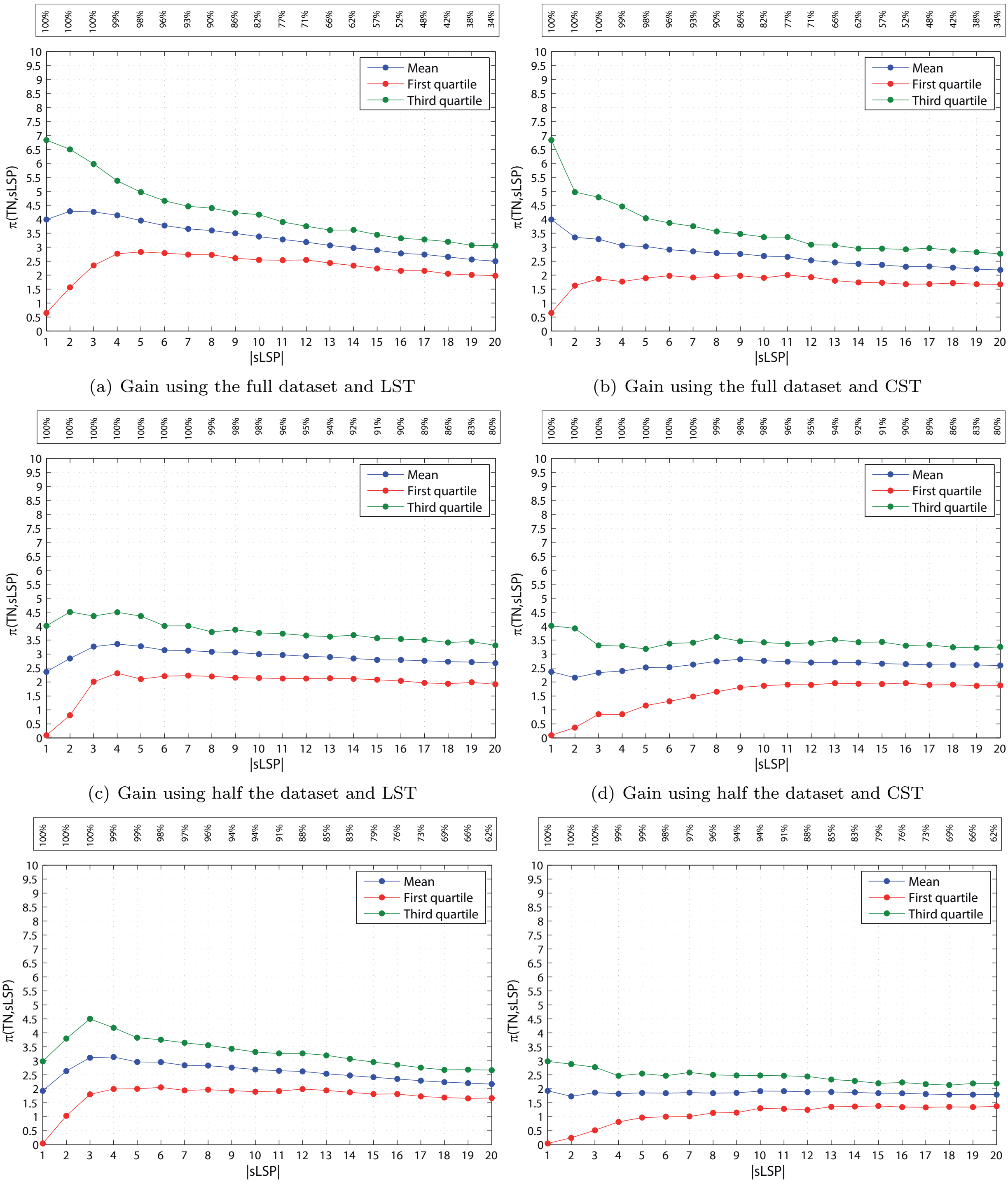
# 5 Discussion

This section will focus on three main points. First of all, the impact of the amount of available data and the selection technique used will be discussed in Section 5.1 and will lead to more insight on the proposed method. In Section 5.2, the managerial aspects related to the use of the method will be discussed, together with the feedback given by Ticketmatic, the company which provided the data. Finally, Section 5.3 will introduce potential interesting topics for future research with regard to the proposed method.

## 5.1 Impact of the parameters

The objective of this section is to answer the two following questions: (1) Is LST better or worse than CST, and (2) does more data lead to better results? To do so, an experiment has been set up using the data described in Section 4. The tested factors are the amount of data, consisting of the full dataset, half the dataset, or a fourth of the dataset, and the selection technique used being either LST or CST. The combination of these factors leads to six different experiments involving all the concerts having minimum 100 tickets in the preprocessed dataset in order to have an acceptable amount of data in the test set, leading to 107 concerts meeting the requirements. For each of these experiments, the SOM-based profile generator was applied 100 times for each of the 107 concerts taking as targeted dimension the interest for the given concert as applied in Section 4.1. The 100 iterations of the method are needed in order to capture the variability introduced by the $k$-means technique as mentioned in Section 4.2. Figure 5.1 shows the results of the six experiments where, for each subfigure representing one of the experiments, the relevant statistics are averaged over the 107 concerts.

Figure 6 summarizes the 64.200 executions of the method needed to perform the experiment, and enables to answer the two questions introduced in this section. Comparing Figures 6(a), 6(c), and 6(e) with 6(b), 6(d), and 6(f) respectively, a clear outperformance of LST can be seen. It must be noted that the values for the three statistics for the first subset $sLSP$ and the last one (not plotted here) are the same whether LST or CST is used because the start and the finish points of both techniques are the same. The concave curvature of the mean gain, when LST is used, leads then to better results than the more convex curvature resulting from the usage of CST. This conclusion is verified for the three levels of the tested factor *Amount of data*, leading to a preference for LST, thus answering the first question of the preceding paragraph. Concerning the impact on



(a) Gain using the full dataset and LST

(b) Gain using the full dataset and CST

(c) Gain using half the dataset and LST

(d) Gain using half the dataset and CST

(e) Gain using a fourth of the dataset and LST

(f) Gain using a fourth of the dataset and CST

Figure 6: Output of the six experiments testing the factors *Amount of available data* and *Ranking technique*.

the performance of the available amount of data, comparing Figures 6(a) and 6(b) with Figures 6(c) and 6(d) respectively, leads to the conclusion that using the full dataset results in better gains for both LST and CST with regard to the usage of half the dataset. Moreover, comparing Figures 6(c) and 6(d) with Figures 6(e) and 6(f) respectively, leads to the conclusion that using half the dataset results in better gains for both LST and CST with regard to the usage of a fourth of the dataset. Both these conclusions can then be generalized by the following: using more data as input for the SOM-based profile generator leads to better profiles with regard to the gain they provide, hereby answering the second question of the previous paragraph.

## 5.2 Managerial aspects

Some managerial considerations that came up in the discussions and in the feedback sessions during the realization of this study will be discussed in this section.

It must be emphasized that every step of the presented application of the profile generator is value-adding for the company. Indeed, as it will be presented in what follows, every step brings insight to the marketer and generates knowledge. This knowledge enables the marketer to perform more efficient actions, e.g. more targeted campaigns, at lower costs and with higher expected benefits. By computing indices in the first step of the method, it is possible to sort customers in function of a certain characteristic, such as e.g. their profitability. The SOM analysis gives the marketer more insight in the available dataset, allowing him to visualize correlations between numerous indices. The clustering in the third step serves to obtain analytic confirmation of what was visualized in the SOM analysis, while the salient dimensions extraction facilitates the understanding of the obtained clusters. Finally, the fourth step delivers profiles which define individuals with specific characteristics, and the last step ranks these profiles to facilitate their use. These profiles allow the marketer to identify specific groups with well defined characteristics as target in certain marketing campaign (e.g. if men between 18 and 25 years old are identified, the marketer could focus on a certain magazine or TV channel with a high reader or viewer density for this group).

The influence of the quantity of data on the performance of the profile generator has been discussed in Section 5.1. Having more data enables to make more powerful predictions, which leads to even more efficient marketing actions. Therefore, this data has a potential market value. Indeed, it would probably be profitable for event organizers to purchase, sell or exchange data with one another, or to set up an aggregated database. The construction of a business model based on ticket sales data would fall outside the scope of this study, but it is certainly an interesting potential business opportunity.

Self-organizing maps are used in many domains, and could be used in countless others. In any of these domains, the profile generation method presented in this paper enables the user of the SOM technology to gain even more valuable knowledge. Indeed, the SOM approach is subject to certain limitations, such as the lack of readability when the dimensionality increases or the subjectivity of the interpretation of such a topographic map. This method goes beyond these limits, by formalizing the analysis of the maps and generating concrete and value-adding profiles.

## 5.3 Further research

In this section, the different steps of the method proposed in Section 3.1 will be challenged, leading to different tracks for further research.

The first step of the method, the generation of indices, has been introduced in Section 3.1 and extended in Section 4 for the purpose of the application. A wider research topic could be the analysis of the impact of the curse of dimensionality on the performance of the proposed method. As mentioned in Section 3.1, categorical variables are preferred because of the information provided by the salient dimension analysis. However, the granularity of the categories should be studied in order to propose a method leading to a definition of an optimal setup for the usage of the proposed SOM-based profile generator. Moreover, in this paper a rudime

The second step of the method, the SOM training, is introduced in this paper as a black box. A more extensive analysis of the impact on the generated profiles of SOM parameters such as the number of neurons, the shape of the SOM, the chosen learning rate and the used neighboring function, should lead to a better insight and a definition of best practices with regard to the way of training the SOM as a step of the proposed method.

The third step, clustering and salient dimension analysis, is introduced in Section 3.1 and the impact of the clustering technique used is discussed in Section 4.2 and Section 5.1 where the variability introduced by the clustering technique is analysed. It should now be clear for the reader that the impact of the clustering technique is not to be neglected and further research could focus on the granularity of the clusters generated and the way of generating them. The salient dimension analysis presented in Azcarraga et al. (2005) has been adapted in Section 3.1 in order to capture more information while suppressing the arbitrary fixing of the values of $z_1$ and $z_2$. Further research could identify another way to determine the sensibility of the salient dimensions analysis based on a more formal and statistical approach.

The fourth step, the generation of profiles, is presented in Section 3.1 with Algorithm 1. The class of evolutionary algorithms could be studied in combination with the generated profiles in order to increase the generated population.

Finally, the fifth step, the ranking of the profiles, should offer a multititude of tracks for further research. A lot of priority rules are indeed conceivable, including LST and CST, and could increase the performance of the proposed method, depending on the application domain and the chosen parameters in the previous steps of the method.

# 6 Conclusion

The SOM-based profile generator proposed in this paper is to be used in order to go further than a classical SOM analysis. As illustrated in Section 4, the interest of the method resides in its capacity to generate value-adding profiles targeting given dimensions using the SOM technology and the extraction of salient dimensions. However, as introduced in Section 3, a SOM-based analysis, which itself provides valuable output given its visualization power, is not to be replaced but reinforced by the generated profiles.

The performance of the proposed method has been illustrated by an application in the concert industry, showing a real added value while identifying factors, such as the available amount of data or the ranking technique used, being potential factors for improvement. The results of Section 5.1 concerning the importance of the available amount of data should be an incentive for companies, which are aiming at building customer profiles, to aggregate their data. By translating the gain in prediction power of the generated profiles in terms of money, it is possible to assess the value of the available data and create new business models. Moreover, it should be clear for the reader that the two ranking techniques proposed in this paper are a starting point for future improvement of the method. It can already be concluded that LST outperforms CST, leading to the expectancy that other techniques could increase the power of the proposed method.

The main contribution of this paper to the literature is the development of a generic method for profile generation which is applicable in all cases where the SOM technology is used. The method enables to formalize intuitive feelings and insights resulting from the combination of a SOM analysis and the extraction of salient dimensions.

# References

A.P. Azcarraga, M.H. Hsieh, S.L. Pan, and R. Setiono. Extracting salient dimensions for automatic som labeling. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 35(4):595–600, 2005.

Bart Baesens, Stijn Viaene, Dirk Van den Poel, Jan Vanthienen, and Guido Dedene. Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research 138*, pages 191–211, 2002.

Bart Baesens, Geert Verstraeten, Dirk Van den Poel, Michael Egmont-Petersen, Patrick Van kenhove, and Jan Vanthienen. Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research 156 issue 2*, pages 508–523, 2004.

Chu Chai Henry Chan. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications 34*, pages 2754–2762, 2008.

David L. Davies and Donald W. Bouldin. A cluster seperation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 224–227, 1979.

Johann Füller and Kurt Matzler. Customer delight and market segmentation: An application of the three-factor theory of customer satisfaction on life style groups. *Tourism Management 29*, pages 116–126, 2008.

Nan-Chen Hsieh. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications 27*, pages 623–633, 2004.

Johan Huysmans, David Martens, Bart Baesens, Jan Vanthienen, and Tony Van Gestel. Country corruption analysis with self organizing maps and support vector machines. *Intelligence and Security Informatics*, pages 103–114, 2006.

Hyunseok Hwang, Taesoo Jung, and Euiho Suh. An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications 26*, pages 181–188, 2004.

Jedid-Jah Jonker, Nanda Piersma, and Dirk Van den Poel. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert Systems with Applications 27*, pages 159–168, 2004.

Su-Yeon Kim, Tae-Soo Jung, Eui-Ho Suh, and Hyun-Seok Hwang. Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications 31*, pages 101–107, 2006.

Teuvo Kohonen. *Self-Organizing Maps.* Springer, 1995.

Philip Kotler, Kevin Lane Keller, Bernard Dubois, and Delphine Manceau. *Marketing Management.* Prentice Hall, 12th edition, 2006.

Jang Hee Lee and San Chan Park. Intelligent profitable customers segmentation system based on business intelligence tools. *Expert Systems with Applications 29*, pages 145–152, 2005.

Duen-Ren Liu and Ya-Yueh Shih. Integrating ahp and data mining for product recommendation based on customer lifetime value. *Information & Management 42*, pages 387–400, 2005.

John A. McCarty and Manoj Hastak. Segmentation approaches in data mining: A comparison of rfm, chaid, and logistic regression. *Journal of Business Research 60*, pages 656–662, 2007.

Ian Michiels. Customer analytics - segmentation beyond demographics. *Aberdeen Group*, 2008.

Babak Sohrabi and Amir Khanlari. Customer lifetime value (clv) measurement based on rfm model. *Iranian Accounting & Auditing Review, Vol. 14 No. 47*, pages 7–20, 2007.

E.H. Suh, K.C. Noh, and C.K. Suh. Customer list segmentation using the combined response model. *Expert Systems with Applications 17*, pages 89–97, 1999.

P.N. Tan, M. Steinbach, V. Kumar, et al. *Introduction to Data Mining.* Pearson Addison Wesley Boston, 2006.

J. Reijnartz Werner and V. Kumar. On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *The Journal of Marketing, Vol. 64, No. 4*, pages 17–35, 2000.