# Who are our clients: consumer segmentation through explorative data mining

Viktor, Herna L.; Peña, Isis; Paquet, Eric



**Publisher's version / Version de l'éditeur:**
https://doi.org/10.1504/IJDMMM.2012.048109
*International Journal of Data Mining, Modelling and Management, 4, 3, pp. 286-308, 2012*

# Who are our clients: Consumer Segmentation through Explorative Data Mining

Isis Peña[1], Herna L Viktor[1] and Eric Paquet[1,2]
[1] School of Information Technology and Engineering, University of Ottawa
[2] National Research Council of Canada

---

The apparel industry aims to produce comfortable and aesthetically pleasing garments that fit populations well. However, repeated studies of apparel customers' levels of satisfaction indicate that their needs are often not being met. In order to produce better fitting clothes, it is crucial to understand the body profiles of typical consumers. Exploring the demographic profiles of correctly identified customer segments holds obvious benefit for production, marketing and ensuring return on investments. To this end, we explore a database, containing anthropometric measurements, demographic profiles and 3-D body scans, of samples of the North American, Italian and Dutch populations. Through the use of objective interestingness measures-based feature selection and feature extraction, we accurately discover the relevant subsets of body measurements that require special care when designing clothes. Furthermore, we apply association analysis to the demographic data, in order to target customers and thus potentially defining new market opportunities.

Categories and Subject Descriptors: H 2.8. Data mining, I 1.2.6. Learning, J 7. Consumer products
General Terms: Algorithms, Design, Experimentation
Additional Key Words and Phrases: Utility-based data mining, anthropometry, interestingness measures, consumer segmentation, marketing

---

## 1. INTRODUCTION

One of the main challenges for the apparel industry is to produce garments that fit the customers properly, are aesthetically pleasing and comfortable, and subsequently sell. In order to produce garments that fit us well, better characterizations of our populations are needed. In the context of tailoring, the aim is to cover the maximum number of individuals with the minimum number of groupings. It follows that these different groupings, or sizes, should correspond to real bodies, within and across various populations. However, repeated studies of the degree of satisfaction with apparel fit show that consumers' needs are not being met (Shofield and LaBat, 2005; Ashdown, Loker and Rucker, 2007). For example, a North American study found that about 50% of women and 62% of men cannot find satisfactorily fitting clothes (DesMarteau, 2000).

Recent work has, to some extent, addressed these concerns. Anthropometric surveys such as the CAESAR™ Project (Robinette, et al., 2002), SizeUSA (2009) and SizeChina (2009) have been carried out on civilian populations. Also, some recent studies attempt to find the most important aspects to be taken into account when designing garments. Viktor, Paquet and Guo (2006) find body size groupings in a sample of the North American male population. Veitch, Veitch and Henneberg (2007) aim to produce a well-fitting bodice for Australian women. Hsu, Lin and Wang (2007) identify three body types

and thirty-eight sizes for the female adult Taiwanese population by applying principle component analysis (PCA) on eleven anthropometric measures.

Although the abovementioned work attempts to address the problem of identifying the main aspects that should be considered for the design of garments, they only focus either on a specific body part, or on a gender. Moreover, they do not account for the economic factors of the data mining process. Importantly, they do not attempt to find the subset of body measurements with the highest utility within this domain. That is, they fall to focus on obtaining those measurements that would be of the most interest when designing apparel for different sizes within each population and gender. Also, they do not explore the link between anthropometric data and demographic profiling, which holds obvious benefits for streamlining production line, focusing marketing and targeted advertising. This paper addresses this need using an interestingness measures-based methodology. In order to achieve this, we aim to understand the typical consumers' body profile by identifying the natural body size groups and their distinctive characteristics, using clustering techniques. Also, we attempt to find the most important body measurements that define each size, and study how these measurements interrelate. To this end, we employ interestingness measures to identify the minimal sets of body measurements that are relevant for the different sizes, or different customer segments, within each population and gender. Next, we perform association analysis to further understand the demographic profile of the various segments. In this way, we obtain reduced body measurements (both anthropometric and 3-D) of high utility, to be used to optimize apparel design and act as a guide for marketing and advertisement campaigns.

The remaining of this paper is organized as follows. Section 2 introduces our case study, namely the CAESAR$^{TM}$ anthropometric database, which contains subjects from North America, Italy and the Netherlands. This is followed by Section 3 which discusses utility-based data mining and, in particular, objective interestingness measures, which are used to guide our data mining process. Section 4 explains our methodology and results when characterizing the populations. In this section, we describe the cluster analysis of the anthropometric measurements. In Section 5, we present the approach followed when reducing the number of body measurements, through the use of interestingness measures-based feature selection together with feature extraction. Section 6 contains results when exploring the demographic profiles of the target populations and Section 7 concludes the paper.

## 2. THE CAESAR™ DATABASE

CAESAR™ is an anthropometric database containing up-to-date information about European and North American civilian populations (Robinette, et al., 2002). Anthropometric data refer to a collection of physical dimensions of a human body. This database includes traditional anthropometric measurements of a large number of individuals from North America, Italy and the Netherlands. The numbers of anthropometric measurements are forty-four (44) for the males, and forty-five (45) for the females, since recording the under bust circumference is not appropriate for the male subjects. These measurements include height, weight, acromial height, waist circumference, thigh circumference and foot length, amongst others, which were recorded by domain experts. Additionally, the shape of each person was scanned in three dimensions using a full body scanner. That is, a laser scanning device measured and recorded detailed geometry of the subjects' body surface. The 3-D body scans were described using a global shape-based descriptor, which is an abstract and compact representation of the three-dimensional shape of the corresponding body. In essence, each scan is represented by a set of three histograms, which constitutes a 3-D shape index or descriptor for the human body (Paquet, Robinette and Rioux, 2000). This index characterizes the radial and angular distribution of the surface elements associated with a given body, and is designed to be orientation invariant and robust against pose variation. In our experiments, we use these 3-D scans to visually validate our anthropometric data mining results. To this end, we employ the Cleopatra system, which is able to navigate through, and retrieve similar, 3-D scans based on their 3-D shape (Paquet, Robinette and Rioux, 2000).

A unique characteristic of the CAESAR™ database is that it also contains demographic details, such as age, income and gender, about all survey participants. This information, when combined with the anthropometric measures, provides us with the opportunity for customer segmentation based on not only their physical characteristics, but also their income levels and lifestyle choices. Through the use of this information, we are thus able to determine inventory levels and possible sales venues, identify groups of individuals who would respond well to direct marketing, and guide advertising campaigns. A crucial issue is, therefore, to find those patterns that are of interest and of high utility, as discussed next.

## 3. FINDING INTERESTING PATTERNS

Utility-based data mining aims to consider all utility aspects in the data mining process, and maximize the utility of the entire process (Weiss, Saar-Tsechansky and Zadrozny, 2005; Weiss and Tian, 2008; Zadrozny, Weiss and Saar-Tsechansky, 2006). In particular, a subset of utility-based approaches focuses on reducing the time and mining space cost by using *interestingness measures*. Interestingness measures are "measures" that narrow the search space and find the truly interesting, and actionable, patterns to the user (Geng and Hamilton, 2006).

In our case study, we aim to discover and typify our consumer segments, in order to (a) guide apparel design and production, (b) better manage inventories and distribution and (c) correctly target potential consumer markets. To achieve this objective, we employ interestingness measures during our data mining process, as follows. We consider the results of the data mining process to be interesting if it is able to correctly and accurately characterize the bodies of different sizes, within each population and gender. The body measurements which are crucial when designing a garment for a size, and therefore need special consideration, are thus considered to be interesting. In addition, we employ a combination of interestingness measures, to discover actionable associations, when analyzing the demographic profiles of the customer segments as identified during anthropometric data mining.

Table I. Objective Interestingness measures for Association and Classification rules

| Name | Formula | Description |
|---|---|---|
| Support | $P(AC)$ | Support is a measure of significance of the rule. Represents the fraction of records that the given rule satisfies. |
| Confidence | $\frac{P(AC)}{P(A)}$ | Confidence measures the reliability of the inference made by the rule. It is defined as the probability of seeing the rule's consequent under the condition that the record also contains the antecedent. |
| Lift | $\frac{P(AC)}{P(A)P(C)}$ | Lift is a correlation measure that indicates how many times more often $A$ and $C$ occur together than expected if they were statistically independent. |
| Conviction | $\frac{P(A)P(\neg C)}{P(A\neg C)}$ | Conviction is a measure of independence of $A$ and $C$ that is sensitive to rule direction since also uses the information of the absence of the consequent. |
| Leverage | $P(AC)-P(A)P(C)$ | Leverage measures the difference of $A$ and $C$ appearing together in the dataset and what would be expected if $A$ and $C$ were statistically independent. |

In the literature, a number of criteria have been proposed to determine whether a pattern is interesting. These criteria are: *Generality*, *Reliability*, *Conciseness*, *Peculiarity*,

*Diversity*, *Novelty*, *Surprisingness*, *Utility* and *Actionability* (Geng and Hamilton, 2006; McGarry, 2005). Since generality, reliability, conciseness, peculiarity and diversity depend only on the data patterns, these criteria are considered objective. The criteria novelty, surprisingness, utility and actionability depend on the user, and therefore are considered subjective.

Table II. Objective Interestingness measures for Feature Selection

| Name | Formula | Description |
|---|---|---|
| Information Gain | Suppose the attribute $A$ has $v$ distinct values $\{a_1,a_2,...,a_v\}$ and splits $D$ into $v$ subsets $\{D_1,D_2,...,D_v\}$. Then *InfoGain* is calculated as: $$entropy(D)-\sum_{j=1}^{v}\frac{\lvert D_j\rvert}{\lvert D\rvert}*entropy(D_j)$$ Where $D_j$ contains those records in $D$ that have outcome $a_j$ and *entropy* is defined as: $$\sum_{j=1}^{m} p_j \log_2(p_j)$$ Where $p_j$ is the probability that a record in $D$ belongs to class $C_j$ | Information Gain measures how much we gain by selecting the attribute $A$. That is, the expected reduction of the information requirement causing by knowing the value of $A$. |
| Gain Ratio | The Gain Ratio for attribute $A$ over dataset $D$ is: $$\frac{InfoGain(A)}{SplitInfo(A)}$$ Where *SplitInfo* is defined as: $$-\sum_{j=1}^{v}\frac{\lvert D_j\rvert}{\lvert D\rvert}*\log_2\left(\frac{D_j}{D}\right)$$ | Gain ratio represents the potential information generated by selecting the attribute $A$. |
| Chi Squared | Let $(A_i,C_j)$ denote the event that attribute $A$ takes the value $a_i$, and attribute $C$ takes de value $c_j$. That is, where $(A=a_i, C=c_j)$. The $\chi^2$ is calculated as: $$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$ Where $o_{ij}$ is the observed frequency (i.e., actual number) and $e_{ij}$ is the expected frequency of $(A_i,C_j)$. The expected frequency is calculated as: $$\frac{count(A=a_i)*count(C=c_j)}{\lvert D\rvert}$$ Where $count(A = a_i)$ is the number of records having value $a_i$ for $A$, and $count(C = c_j)$ is the number of records having value $c_j$ for $C$. | Chi-squared measures the correlation between two attributes, $A$ and $C$. $A$ represents a conditional attribute and $C$ represents either another conditional attribute or the class attribute. |

Objective measures are based on the statistical strengths or properties of the discovered patterns. Little knowledge about the user or application is initially required, in order to assess the degree of interestingness of a pattern. In this research, our aim is to find the natural groupings in the data, based on the subjects' anthropometric measurements; i.e. groupings that are concise, reliable and general. Therefore, we focus

our attention on objective interestingness measures. Table I shows the most widely used objective interestingness measures in our context. In order to make the measures comparable, all measures are defined using probabilities. *A* represents the antecedent, *C* represents the consequent and *N* denotes the total number of records. $P(A) = {}^{n(A)}\!/_{N}$ denotes the probability of *A*.

Interestingness measures such as Information gain, Gain ratio and Chi-squared (as shown in Table II) are not only used as heuristics to select attribute-value pairs for inclusion in classification rules, but are also used for feature selection. (The notation used in Table II is as follows. Let *D* be the dataset; suppose the class attribute has *m* distinct values defining *m* different classes, $C_i$ (for $i = 1, \ldots, m$). Let $C_i$, *D* be the set of records of class $C_i$ in *D*. Let |D| and $|C_i, D|$ denote the number of records in *D* and $C_i$, *D*, respectively). The idea in feature selection is to remove the attributes with little or no predictive information. This reduces the set of attributes to be used as well as the time and mining space, thus improving both efficiency and accuracy (Geng and Hamilton, 2006; Han and Kamber, 2006), as will be shown in Section 5 of this paper.

## 4. CHARACTERIZATION OF THE POPULATION

As stated earlier, one of the most important challenges for the clothing industry is to produce garments with quality fit. In order to produce better fitting garments, accurate and up-to-date measurements need to be further analyzed in order to be able to better characterize the population (Ashdown, Loker and Rucker, 2007) and target consumers. It is therefore crucial to obtain an accurate profile of the body profiles of the various consumer segments.

To address the aforementioned issue, we aim to find the natural body size groupings using the anthropometric measurements as contained in the CAESAR™ anthropometric database. From these groups we identify size *archetypes* and their most important characteristics. All our experiments are implemented in WEKA, a collection of machine learning algorithms for data mining tasks (Witten and Frank, 2005). We validate our results using a 3-D information retrieval system, called Cleopatra (Paquet, Robinette and Rioux, 2000). Cleopatra is a visual data mining tool that allows to navigate through the index space and to visualize the associated human bodies. In our experiments, it is used to explore, the possible grouping of the various individuals. In this study we consider the American female population as well as the Italian and Dutch, both male and female populations. (Interested readers are referred to (Viktor, Paquet and Guo, 2006) for a

discussion of the results obtained when analyzing the American male population.) The data was first separated based on the gender of the subjects. The resulting sets consist of 256 American females, 413 Italian males, 388 Italian females, 567 Dutch males and 700 Dutch females.

### 4.1 Clustering of Anthropometric Measurements

In order to identify the natural body size groupings, as a first step to facilitate customer segmentation, we apply cluster analysis techniques to the anthropometric data. Cluster analysis is an unsupervised learning data mining technique used to partition a set of physical or abstract objects into subsets or clusters based on data similarity (Han and Kamber, 2006; Witten and Frank, 2005).

In the context of tailoring and clothing manufacturing, the ideal scenario is to cover the greatest number of people with the fewest number of sizes (Hsu, Lin and Wang, 2007). Therefore, we aim to find the minimum number of clusters that fully characterize the population. Since three clusters is the minimum number that makes sense from a tailoring point of view, i.e. *small*, *medium* and *large*, we start partitioning the data into three clusters. Then, by inspecting the cluster distribution we decide whether is worthwhile to split the clusters, as described in (Witten and Frank, 2005). This process is repeated until the clusters appear well-defined.

The measurements are normalized, but not with the standard deviation. Recall that we aim to cluster the bodies as a whole. If the measurements are normalized with their respective standard deviations, they become scale invariant. This is not a suitable feature in our case, due to the following reason. Consider a distance related to a relatively small portion of the body, e.g. the finger length. This measurement should not be considered on an equal footing with a distance related to a relatively large proportion of the body, like e.g. the height, since we are attempting to cluster the bodies as a whole. For this reason, the measurements are normalized relative to a characteristic scale which, in practice, is application dependant.

Our objective is to cluster on the entire space of measurements. Naturally, by using features selection (as described in the next section), one may reduce the number of dimensions and alleviate the curse of dimensionality. Nevertheless, the negative effect associated with the high dimensionality of our space is reduced by the fact that our measurements are not noisy. Remember that these measurements were precise measurements made by a group of experts and that each one of them was validated by at least another independent expert. Since our measurements are not noisy and are relevant

for the body description, in the sense that they are the standard measurements used in anthropometry to characterize the human body, they may potentially increase the contrast (Houle, et al., 2010). The fact also that the measurements are normalized, at a particular scale, implies that only the dimensions which are relevant for that particular scale or resolution have a high impact on the clustering process.

A number of clustering algorithms were considered. These included partitioning, hierarchical, density-based, model-based and grid-based approaches. By inspection of the cluster distribution and through the analysis of the results using Cleopatra, we found that the American female population is best characterized with five clusters. This is also the case for the Italian and Dutch male populations, while the Italian and Dutch female populations are better characterized with six clusters as shown in (Peña, Viktor and Paquet, 2009a; Peña, Viktor and Paquet, 2009b). We also observed that the best partition for the American female population is achieved by the k-means algorithm, while the best partition for the Italian and Dutch both males and females is achieved using a *density-based* algorithm with k-means components. In a density-based algorithm, clusters are grown, starting from a seed value, as long as the density of data points exceeds some threshold value (Han and Kamber, 2006). It follows that the clusters are regarded as regions of high density which are separated by regions of low object density (noise).

As stated above, we visually validate the clusters produced by using the Cleopatra system (Paquet, Robinette and Rioux, 2000), as follows. The Cleopatra system enables us to retrieve the 3-D body scans associated with each subject in the CAESAR™ database. Each cluster Centroid or archetype is used as a "seed", and we proceed to find the *n* most similar bodies, in terms of 3-D shape, from the database. Here, *n* is user defined and corresponds to the number of subjects belonging to the same cluster. The similarity is measured using the Euclidian distance. Our choice of the Euclidian distance is motivated by the fact that this distance is the most commonly used in anthropometry, when comparing anthropometric measurements. Using another metric would lead to results that are not comparable with previous results, as obtained in the field, and would not be suitable for practical and industrial applications such as design and ergonomic (Robinette, et al., 2002; Hsu, Lin and Wang, 2007).

We determine whether the *n* nearest bodies fall within the same anthropometric cluster. This extra step allows us to double-check that the anthropometric results also "make sense" for a 3-D shape point of view. Tables III to VII indicate some of the characteristics of the Centroids of the male and female populations. Shown in the table are the mean (in cm), the standard deviation in parenthesis, and the number of subjects on

each cluster. Fig. 1 to Fig. 5 show the 3-D body scans of the human subjects that correspond to these measurements, highlighting the difference in body types of the clusters. Each subject is visualised using four views, namely their fronts, backs, as viewed sideways and seen from the top. These figures are in high resolutions and readers are invited to zoom in to see the details. Thus, by inspecting the values in Tables III to VII, the cluster distribution, and through the analysis of the results using Cleopatra, we observe that the anthropometric clusters discriminate between the different body sizes.

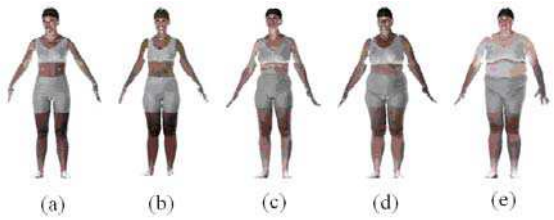



Fig. 1. American female Archetypes: (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large

Table III. Body Measurements of American Female archetypes

| | Small | Medium | Large | X-Large | XX-Large |
|---|---|---|---|---|---|
| Bust | 87.7 (5.8) | 88.8 (4.6) | 94.4 (4.5) | 105.4 (6.9) | 127.1 (9.6) |
| Waist | 69.1 (6.1) | 70.8 (5.0) | 77.4 (6.3) | 88.4 (9.1) | 110.5 (5.6) |
| Hip | 96.0 (5.0) | 99.4 (5.1) | 105.5 (5.3) | 113.9 (7.7) | 133.2 (9.8) |
| Shoulder to Wrist | 53.4 (2.0) | 57.1 (1.7) | 60.5 (2.3) | 56.9 (1.7) | 63.6 (2.7) |
| Stature | 155.9 (4.9) | 163.7 (3.8) | 171.9 (5.0) | 162.0 (4.0) | 180.6 (5.2) |
| Shoulder Breadth | 41.2 (1.7) | 42.4 (1.8) | 44.5 (1.8) | 46.8 (2.2) | 55.4 (4.1) |
| Weight (lbs) | 119.0 (12.5) | 130.4 (10.8) | 154.7 (11.9) | 178.4 (22.3) | 278.9 (23.7) |
| Num. of Subjects | 53 (21%) | 99 (39%) | 60 (23%) | 39 (15%) | 5 (2%) |

By inspecting the body measurements of the American female population, we observe that the range of measurements of the top body is wide. This indicates that the torso is the most variable feature (in all the sizes) for the American females. This is observed in the American Centroids in Fig. 1, where the most noticeable differences between sizes are in the torso area. It may also be observed that the *Small* size subjects have short torso. This makes sense, since *Small* subjects are usually shorter than the subjects of other sizes. Consequently, when designing clothes for the American population, this has to be taken into account to produce garments that fit the population properly.

For the Italian female population, we observe that the main difference between the *X-Small* and *Small* sizes is the height. Since the *Small* size subjects are taller, it follows that they have longer arms, as may be observed in Table IV. This again represents a major

consideration when designing garments such as jackets, as opposed to bust circumference or shoulder breadth, which are very similar for both sizes.

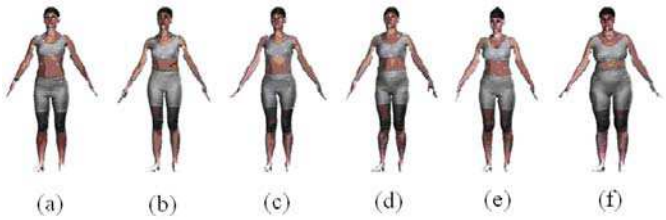



Fig. 2. Italian female Archetypes. (a) X-Small, (b) Small, (c) Medium, (d) Large, (e) X-Large and (f) XX-Large

Table IV. Body Measurements of Italian Female archetypes

| | X-Small | Small | Medium | Large | X-Large | XX-Large |
|---|---|---|---|---|---|---|
| Bust | 83.8 (4.6) | 82.9 (3.4) | 86.9 (4.2) | 91.9 (5.7) | 92.8 (4.8) | 108.0 (10.6) |
| Waist | 68.5 (5.5) | 70.8 (4.8) | 74.2 (4.7) | 76.8 (5.8) | 78.8 (5.4) | 92.3 (9.9) |
| Hip | 91.4 (3.9) | 92.4 (3.6) | 96.4 (3.7) | 99.1 (4.0) | 102.5 (4.1) | 114.6 (5.0) |
| Shoulder to Wrist | 53.8 (1.4) | 57.7 (1.2) | 59.3 (1.8) | 56.3 (1.5) | 61.3 (2.1) | 58.4 (2.1) |
| Stature | 152.7 (3.3) | 159.2 (2.6) | 164.8 (2.8) | 157.1 (3.2) | 169.3 (4.3) | 160.9 (4.7) |
| Shoulder Breadth | 38.2 (1.6) | 39.0 (1.6) | 40.3 (1.5) | 41.0 (1.8) | 42.2 (1.6) | 45.0 (2.1) |
| Weight (lbs) | 106.7 (8.2) | 110.2 (6.1) | 124.2 (8.5) | 129.5 (9.8) | 143.9 (9.7) | 177.5 (18.5) |
| Num. of Subjects | 51 (13%) | 65 (17%) | 110 (28%) | 82 (21%) | 56 (14%) | 24 (6%) |

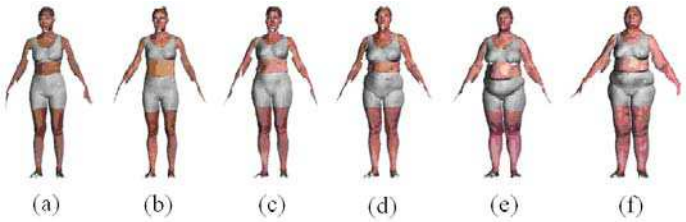



Fig. 3. Dutch female Archetypes. (a) X-Small, (b) Small, (c) Medium, (d) Large, (e) X-Large and (f) XX-Large

For the Dutch female population (Table V) is noticeable that the tallest individuals are among the *Medium* and *XX-Large* sizes. Also, the longest arm length is among the *Medium* and *XX-Large* populations, surpassing importantly the *Large* and *X-Large*-sized subjects. This may be interpreted as follows. The *Large* and *X-Large* subjects do not have the same constitution as the individuals that wear size *Medium*. They are shorter and more robust than the *Medium*-sized persons. This provides valuable information about garment design, for different sizes. For the size *Medium*, the clothes have to be designed mainly for tall subjects, while in the design of garments for the *Large* and *X-Large* sizes, the girths are the primary aspect to take into account.

Table V. Body Measurements of Dutch Female archetypes

| | X-Small | Small | Medium | Large | X-Large | XX-Large |
|---|---|---|---|---|---|---|
| Bust | 90.7 (5.9) | 92.1 (5.0) | 98.4 (6.2) | 106.1 (6.7) | 117.5 (8.3) | 120.6 (9.2) |
| Waist | 74.1 (6.4) | 76.4 (5.9) | 83.5 (7.1) | 90.7 (8.0) | 103.3 (9.4) | 107.9 (11.0) |
| Hip | 97.2 (5.1) | 101.2 (5.4) | 106.4 (5.1) | 110.1 (5.8) | 117.5 (7.7) | 121.6 (10.4) |
| Shoulder to Wrist | 55.1 (2.4) | 58.5 (1.8) | 62.0 (2.2) | 56.8 (1.9) | 58.6 (2.2) | 63.0 (2.0) |
| Stature | 160.0 (4.7) | 169.7 (3.7) | 176.9 (4.8) | 162.4 (5.1) | 167.1 (4.3) | 176.9 (5.6) |
| Shoulder Breadth | 40.3 (1.7) | 41.7 (1.7) | 44.0 (2.1) | 43.6 2.4) | 47.0 (2.9) | 47.6 (2.9) |
| Weight (lbs) | 126.5 (13.1) | 141.1 (12.4) | 165.3 (14.5) | 171.1 (13.3) | 210.0 (21.8) | 234.3 (28.9) |
| Num. of Subjects | 130 (19%) | 198 (28%) | 125 (18%) | 125 (18%) | 83 (12%) | 39 (6%) |

For the Italian males (Table VI), we observe that the chest, waist, and neck base circumference of the *Large* and *X-Large* subjects are similar. The main feature that makes these sizes different is that the *X-Large* subjects have longer arms. Notice the impact this fact has on clothing design. Consider for instance designing jackets: for persons who are wearing *Large* and *X-Large* garments, the main consideration becomes the arm length. Other measurements such as the chest circumference remain practically the same in both sizes. We also notice that even though the measurements of the *X-Large* subjects are similar to the *Large*-sized individuals, the former are taller than the *Large* and *XX-Large* subjects. Thus, when designing, for example, pants for the *X-Large* subjects, these need to have a longer leg length.

Table VI. Body Measurements of Italian Male archetypes

| | Small | Medium | Large | X-Large | XX-Large |
|---|---|---|---|---|---|
| Chest | 90.3 (5.1) | 91.2 (4.6) | 98.8 (5.6) | 98.9 (5.9) | 108.4 (7.1) |
| Waist | 78.2 (5.2) | 79.7 (3.8) | 87.5 (6.2) | 86.9 (5.6) | 97.8 (9.6) |
| Hip | 93.0 (4.1) | 93.9 (4.0) | 100.0 (3.9) | 100.5 (3.8) | 108.7 (7.8) |
| Neck Base | 45.8 (1.7) | 46.6 (1.4) | 48.2 (1.6) | 48.6 (1.4) | 50.2 (1.7) |
| Shoulder to Wrist | 60.7 (1.9) | 64.2 (1.8) | 61.7 (1.8) | 67.2 (2.1) | 64.9 (1.9) |
| Stature | 166.8 (4.2) | 175.3 (3.8) | 170.3 (4.1) | 182.6 (5.0) | 176.8 (4.3) |
| Shoulder Breadth | 43.7 (2.1) | 45.0 (1.9) | 46.2 (1.9) | 47.5 (2.3) | 49.5 (2.0) |
| Weight (lbs) | 136.6 (12.5) | 147.3 (11.6) | 166.7 (11.8) | 174.6 (14.2) | 203.8 (24.8) |
| Num. of Subjects | 88 (21%) | 107 (26%) | 107 (26%) | 69 (17%) | 42 (10%) |

Fig. 4. Italian male Archetypes. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large

For the Dutch male Centroids shown in Table VII, it may be observed that the *XX-Large* subjects are generally shorter than the *X-Large* persons. This is, again, an important feature to consider when designing, for example, pants; the legs should have short lengths for the *XX-Large* size. We also observe that the *Large* and *X-Large* subjects have similar waist circumference, but the former have slightly wider chests. Subsequently, when designing jackets or shirts for the *Large* size, these should be wider and shorter than the ones designed for the *X-Large* subjects.

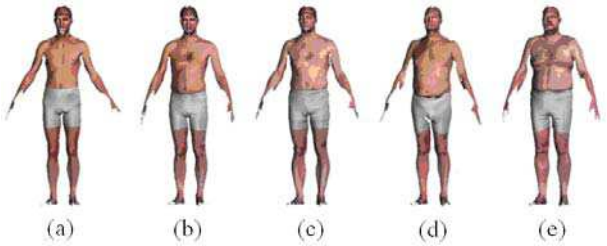


Fig. 5. Dutch male Archetypes. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large

Table VII. Body Measurements of Dutch Male archetypes

| | Small | Medium | Large | X-Large | XX-Large |
|---|---|---|---|---|---|
| Chest | 93.1 (6.0) | 96.5 (4.8) | 107.8 (6.7) | 104.9 (6.2) | 121.2 (7.0) |
| Waist | 82.5 (6.9) | 86.5 (5.7) | 96.3 (7.4) | 97.4 (5.7) | 112.5 (9.1) |
| Hip | 95.3 (4.6) | 98.5 (3.6) | 103.1 (4.5) | 108.1 (4.0) | 116.3 (6.2) |
| Neck Base | 45.4 (2.4) | 47.4 (2.5) | 50.3 (2.7) | 51.0 (2.3) | 55.2 (3.4) |
| Shoulder to Wrist | 60.3 (3.3) | 65.4 (2.6) | 62.4 ( 3.0) | 67.2 (3.2) | 65.6 (3.6) |
| Stature | 173.0 (6.6) | 184.8 (4.6) | 176.1 (5.6) | 193.2 (6.2) | 188.1 (8.3) |
| Shoulder Breadth | 44.0 (1.6) | 46.6 (1.8) | 47.8 (2.0) | 48.8 (2.0) | 52.0 (3.4) |
| Weight (lbs) | 149.9 (14.8) | 171.5 (12.2) | 198.9 (17.5) | 214.4 (16.9) | 264.1 (29.2) |
| Num. of Subjects | 126 (22%) | 173 (31%) | 139 (25%) | 82 (14%) | 47 (8%) |

### 4.2 Analysis of the Three Populations

Using the anthropometric clustering results, we analyze the three populations of our study, and discuss some similarities and differences for both the male and female populations. In our analysis, we consider the Centroids or archetypes, since these are representatives of the other subjects that belong to the same cluster. In the analysis of the

male population we also consider the results for the American male population as presented in (Viktor, Paquet and Guo, 2006). Fig. A1, as contained in Appendix A, shows some body measurements and indicates how the three male populations are positioned in the measurement range.

Recall that our aim is to determine the natural groupings within each population, and also to find possible differences between the three populations, which may be of interest to streamline inventory, target consumers and when designing the clothes. When considering the data in Tables VI and VII, and Fig. A1, the next observations are worth mentioning. Even though, the three populations are described by five sizes, these are not comparable, per se, to one another. For example, if we consider the *Small* sizes for the Americans, Italians and the Dutch, we observe the Italians are considerably thinner, while both the Americans and Dutch tend to be more robust. In general we observe the Italians are thinner than the Americans and the Dutch. It may be seen also that the tallest population is the Dutch, while the shortest population is the Italians. Moreover, in the three populations, the *XX-Large* subjects are shorter than the *X-Large* subjects. This is an important feature to consider when designing, for example, pants; the legs should have short lengths for the *XX-Large* size. Furthermore, if we examine the measurements and height of the American and Dutch archetypes of the corresponding sizes, we observe that the range of measurements are similar, but in general the Dutch are taller, making the Americans the most robust population.

Fig. A2 in Appendix A shows how the three female populations are distributed when considering some typical body measurements. Since the number of sizes for the Americans is five, while for the Italians and Dutch they are six, again we cannot compare them directly. However, from Tables III to V, and Fig. A2, it may be observed that the Italians are the thinnest and shortest population. We also observe that the Dutch *X-Small* subjects are more robust and taller than the American *Small*; the Dutch size *Small* individuals have wider chests and hips, and are taller than the American *Medium*. If we continue in this way, it follows that the Dutch *XX-Large* individuals are taller and more robust than the American *XX-Large*, but surprisingly, the American *XX-Large* subjects are taller and more robust than the Dutch *XX-Large*.

Moreover, we notice some relationships among the different sizes of the three populations. For instance, we notice the body measurements that correspond to the Dutch *Large*, American *X-Large* and Italian *XX-Large* sizes are very similar. Furthermore, the size *Small* of the Dutch is comparable to the *X-Large* size of the Italians. The Dutch

seems to be larger than the Americans for the smaller sizes, but the American *XX-Large* resulted to be the tallest and most robust of all populations, as shown in Fig. 1 and Fig. 3.

## 5. FINDING THE MOST IMPORTANT MEASUREMENTS

In the previous section, we identified archetypes of our consumers, using the total number of body measurements. We showed that the sizings of the three populations used in this study vary. We now utilize interestingness measures to reduce the number of body measurements. Reducing the number of measurements not only increases the efficiency of the learning process, enhances comprehensibility of the learned results and improves the learning performance (predictive accuracy) (Han and Kamber, 2006). Rather, the reduced sets of body measurements also help identify the body measurements that require special attention when designing, manufacturing and subsequently marketing garments. This holds obvious benefits. Firstly, by using these measures as "constraints" during manufacturing, we ensure that we do not produce clothes that simply do not sell within a population, due to poor fit. Furthermore, it allows us to differentiate between sizes, focusing on details such as e.g. variations of typical sleeve lengths between different segments of populations. In terms of marketing, an "our shirt sleeves will fit you comfortably" advertising campaign comes to mind. In this way, one may ensure that the claims of the advertisement campaign are, indeed, matched by the characteristics of the product.

To this end, we perform *feature selection*, as introduced in Section 3, to remove features or attributes in the data that are statistically uncorrelated with the class labels. Recall that interestingness measures are used in feature selection to remove the attributes with little or no predictive information. In our case study, this means that we use interestingness measures to identify the subset of the body measurements which is of most importance when describing the archetype of each size, or customer segment. For feature selection we thus apply Information Gain, Gain Ratio, Chi Squared, the Consistency subset evaluator and the CFS subset evaluator. These are measures that have been widely used in the context of feature selection and have been found to produce good results (Cunningham, 2007).

In order to verify our results, we consider the full set of anthropometric measurements (forty-four for the males, and forty-five for the females since under bust circumference was only recorded for the males) and the subsets produced by the feature selection and feature extraction. In order to perform feature selection and feature extraction, we first constructed a number of classifiers, where the clusters we discovered during the

characterization phase acted as class labels. For our experimentation, we consider three different classifiers, namely *RIPPER*, *C4.5* and *PART*. Our choice of classifiers is motivated by the fact that we are interested in finding set of rules, in order to identify those attributes that are of importance within the various populations.

The results of applying feature selection on the anthropometric data for the American, Italian and Dutch male and female populations are shown in Tables VIII to XIII. Shown are the predictive accuracy and, in parenthesis, the number of attributes in the subset.

Table VIII. Results of the Attribute Reduction for the American Male Population

| | Original | Info Gain | Gain Ratio | $\chi^2$ | Consis Subset | CFS Subset |
|---|---|---|---|---|---|---|
| PART | 78.7% | 82.6% (8) | 81.2% (5) | 82.6% (8) | **83.1% (8)** | 80.0% (26) |
| Ripper | 79.2% | 82.9% (6) | 81.2% (8) | 82.9% (15) | **79.7% (8)** | 78.7% (26) |
| C4.5 | 80.1% | 83.8% (6) | 81.9% (7) | 83.1% (7) | **84.5% (8)** | 79.5% (26) |

Table IX. Results of the Attribute Reduction for the American Female Population

| | Original | Info Gain | Gain Ratio | $\chi^2$ | **Consis Subset** | CFS Subset |
|---|---|---|---|---|---|---|
| PART | 75.8% | 82.0% (24) | 77.3% (11) | 80.9% (27) | **82.4% (8)** | 79.7% (25) |
| Ripper | 76.6% | 80.5% (19) | 79.3% (19) | 81.3% (12) | **77.7% (13)** | 78.1% (25) |
| C4.5 | 76.2% | 80.1% (6) | 78.5% (7) | 80.1% (8) | **83.6% (8)** | 77.7% (21) |

Table X. Results of the Attribute Reduction for the Dutch Male Population

| | Original | Info Gain | **Gain Ratio** | $\chi^2$ | Consis Subset | CFS Subset |
|---|---|---|---|---|---|---|
| PART | 80.3% | 82.7% (13) | **82.2% (14)** | 81.1% (14) | 80.3% (12) | 81.3% (25) |
| Ripper | 80.6% | 81.3% (9) | **82.4% (14)** | 81.3% (13) | 82.0% (8) | 80.3% (25) |
| C4.5 | 78.5% | 81.8% (9) | **82.2% (6)** | 80.6% (13) | 80.4% (12) | 80.3% (31) |

Table XI. Results of the Attribute Reduction for the Dutch Female Population

| | Original | Info Gain | Gain Ratio | $\chi^2$ | Consis Subset | CFS Subset |
|---|---|---|---|---|---|---|
| PART | 81.1% | 83.9% (12) | 82.3% (13) | **84.7% (7)** | 83.1% (7) | 81.9% (35) |
| Ripper | 77.9% | 83.4% (13) | 82.9% (17) | **82.7% (7)** | 81.3% (7) | 81.9% (25) |
| C4.5 | 77.4% | 82.9% (13) | 83.3% (11) | **82.7% (7)** | 81.7% (7) | 79.3% (25) |

For each of the populations, we select those subsets that provided a good trade-off between predictive accuracy and number of measures. For example, for the American male population (Table VIII) we observe that the subsets produced by Information gain,

Gain Ratio, Chi Squared and Consistency subset evaluator considerably improve the predicted accuracy with less body measurements. This is especially evident for the subset produced by the Consistency subset evaluator, where the accuracy is higher than 83%, when using PART and C4.5. This subset then contains the eight most important measures to define the body sizes for the male population. For each population, the choice is indicated in **bold**.

Table XII. Results of the Attribute Reduction for the Italian Male Population

| | Original | Info Gain | Gain Ratio | $\chi^2$ | Consis Subset | CFS Subset |
|---|---|---|---|---|---|---|
| PART | 73.6% | 82.1% (18) | **82.1% (15)** | 81.1% (12) | 79.7% (20) | 78.7% (25) |
| Ripper | 74.3% | 78.5% (15) | **80.2% (10)** | 77.5% (14) | 78.9% (20) | 78.0% (25) |
| C4.5 | 73.9% | 77.0% (9) | **78.2% (8)** | 76.3% (8) | 75.8% (20) | 76.3% (25) |

Table XIII. Results of the Attribute Reduction for the Italian Female Population

| | Original | **Info Gain** | Gain Ratio | $\chi^2$ | Consis Subset | CFS Subset |
|---|---|---|---|---|---|---|
| PART | 78.1% | **83.8% (7)** | 83.3% (10) | 81.7% (7) | 81.2% (12) | 80.2% (24) |
| Ripper | 75.8% | **81.7% (7)** | 81.4% (7) | 82.5% (8) | 80.2% (8) | 78.1% (31) |
| C4.5 | 76.8% | **81.7% (5)** | 81.7% (8) | 83.0% (6) | 79.1% (12) | 79.9% (24) |

The reduced set of measurements indicate that, for the American males, the most important body measurements are the acromial height and knee height together with the length of the arm. Special attention should be paid to the knee and acromial heights when designing long or short pants, in order to thus take the position of the knee into consideration. Furthermore, the length of the arm is important when designing shirts that fit this population well.

Table XIV. Reduced Anthropometric Body Measurements for the American Population

| Males | Females |
|---|---|
| Acromial Height Sitting | Arm Length (Shoulder-Wrist) |
| Arm Length (Shoulder-Wrist) | Arm Length (Shoulder-Elbow) |
| Arm Length (Spine-Wrist) | Bust Circumference under Bust |
| Hand Length | Buttock Knee Length |
| K nee Height Sitting | Stature |
| Stature | Subscapular Skinfold |
| Thumb Tip Reach | Thumb Tip Reach |
| Weight | Weight |

For the American females, the circumference under the bust and the buttock knee length become crucial when defining the body size. Hence, when designing clothes for the American females, the circumference under the bust should receive more attention

than other measurements that are mainly used in garment design, such as the bust circumference. Moreover, the subscapular skinfold, a measurement of subcutaneous fat accumulation, is considered in the reduced set of body measurements for the American females. This confirms our previous results that the Americans are the most robust population.

For the Dutch males, the reduced set of measurements indicates that the most significant measurements are the waist circumference, the chest girth at scye and the vertical trunk circumference. When tailoring shirts, sweaters or jackets, for the male population, these measurements should be considered carefully to produce garments that fit this population properly. For the Dutch females, the most important measurements are the bust circumference and, as in the case of the males, the chest girth at scye and the vertical trunk circumference. Therefore, when tailoring clothes for the Dutch females, the bust circumference requires special attention in order to design garments that fit the population better.

Table XV. Reduced Anthropometric Body Measurements for the Dutch Population

| Males | Females |
|---|---|
| Chest Girth at Scye | Arm Length (Spine-Wrist) |
| Hip Breadth Sitting | Bust Circumference |
| Stature | Chest Girth at Scye |
| Vertical Trunk Circumference | Stature |
| Waist Circumference | Thumb Tip Reach |
| Weight | Vertical Trunk Circumference |
| | Weight |

Table XVI. Reduced Anthropometric Body Measurements for the Italian Population

| Males | | Females |
|---|---|---|
| Arm Length (Shoulder-Wrist) | Hip Circ Max Height | Arm Length (Shoulder-Wrist) |
| Arm Length (Spine-Wrist) | Knee Height Sitting | Arm Length (Spine-Wrist) |
| Buttock Knee Length | Stature | Knee Height Sitting |
| Chest Circumference | Thumb Tip Reach | Stature |
| Chest Girth at Scye | Waist Circumference | Thumb Tip Reach |
| Crotch Height | Waist Height | Vertical Trunk Circumference |
| Hip Breadth Sitting | Weight | Weight |
| Hip Circumference | | |

For the Italian females, the reduced set of measurements considers the vertical trunk circumference and the knee height, which are relevant when, for example, tailoring blouses, skirts or pants. The vertical trunk circumference is important when deciding what the length of a jacket or a blouse should be, in order to produce garments that are not too short or long for this population. For the Italian males, the most important

measurements are the chest, waist and hip circumferences along with the crotch, waist and hip heights. The measurements, then, address both the height and girths. This indicates that not only the height, but also the chest, waist and hip circumferences should receive special attention when designing clothes for the Italian males. This, again, confirms our results that the main characteristics of the Italian population are related to height and girths. That is, the Italians are the shortest and thinnest population.

## 6. DETERMINING THE TARGET POPULATION

It follows that the reduced set of anthropometric measurements of the three sets of population archetypes, as discussed in the previous section, provides us with useful knowledge regarding the typical body shapes of consumers. The next step, then, is to determine our target markets. It follows that understanding the demographic nature of the individuals within each size, or then each customer segment, may point towards new market opportunities. To this end, when identifying a target market, demographic aspects such as age, gender, income and lifestyle choices need to be considered. This demographic profile may be used to determine, for example, the potential customers to be targeted by an expensive clothing line. The main lifestyle characteristics of such a group may then be used to identify venues where advertisements should be placed, so as to obtain maximum results. Specifically, we are interested in finding not only the relationship between demographic attributes, but also to link them to the different anthropometric sizes or segments. That is, we are interested in linking a person's lifestyle preferences and other demographic indicators to his/her physical characteristics, in order to facilitate advertising and direct marketing campaigns.

We performed association analysis on the demographic data associated with each customer size, as identified earlier. In our analysis of demographic data we used Apriori (Agrawal and Srikant, 1994) and Predictive Apriori association rule learners (Scheffer, 2001). The default measure for evaluating the interestingness of a rule in Apriori is confidence. Confidence is the proportion of the records covered by the antecedent that are also covered by the consequent. A problem with confidence is the fact that is sensitive to the frequency of the consequent. Consequents with higher support produce higher confidence values even if there is no association between the items. We then also use the interestingness measures lift, conviction and leverage. As shown by Sheikh, Tanveer and Hamdani (2004), a single measure should not be used to determine the interestingness of a rule. Instead, a combination of different measures has to be used in order to get the rules that are really interesting. We select then the association rules

whose lift and conviction values are greater than 1 and leverage value greater than 0. We also filter out rules with less than 85% confidence. For the Predictive Apriori algorithm, we select only the rules with 75% or greater accuracy.

### 6.1 Demographic data analysis

Recall that the CAESAR$^{TM}$ database includes demographic data such as fitness, education, occupation, marital status, income, etc. and data about the subject's car was also recorded. The demographic data was collected through a demographic questionnaire that was filled out by the participant. Even though the information was clarified with the participant in case of apparent inconsistencies, the demographic data provided depend on the truthfulness and objectivity of the participants.

The sampling of the population is another issue that should be considered when analyzing the demographic data. Although the goal of stratified sampling is to have an equal number of subjects on each stratum, for the minority groups this was not achieved. This minority groups are not representative and should be interpreted with caution. This is especially evident in the American and Italian sample. The sampling in the Dutch population shows the most balanced number of subjects per strata and it is not biased towards any age range. For the Italian population (both males and females) the number of rules was very small. This is caused by the bias of the data towards young people. Actually, 65% of the Italian population is 30 years or younger. Most of them are students that are single with no children, and low income. Therefore, the rules generated from these datasets mainly indicate this pattern. Moreover, defining a demographic profile of the Italian population is additionally difficult by the fact that there is around 67% of the income values and about 50% of the automotive values that are missing. Therefore, we decided not to include the results of the Italian population in our findings.

### 6.2 Demographic profiles

Using the information provided by the association rules, we now analyze the interrelationships between age, fitness, education, occupation and income, amongst others. This demographic profile allows a better understanding of the customer and some of their preferences.

Firstly, we considered the relationship between the level of fitness and the marital status and number of children, for **American males**. In general, single males with no children have higher level of fitness than married males. Also, we observe that a 30 aged or older male is usually married, while a male younger than 30 years is most likely to be

single. In general married males have higher income than single males. This is especially evident for the Medium sized males, where a married Engineer earns significantly more than his single counterpart. Moreover, for the Large sized group, a 31-35 aged married male holding a master's degree with high level of fitness have higher income than an unfit person with the same background. For the XX Large sized males, we found that there is a large segment of 46-50 aged married males that have high income. Recall that, in general, XX Large people are difficult to characterize and therefore to produce garments that fit them well. Again, a marketing effort may want to study further the characteristics of the individuals within this group in order to produce exclusive garments that target this segment of the population. We also noticed that there is a strong relationship between the males whose occupation is management and the Infiniti luxury model car. The same relationship is observed between Engineers and Ford automobiles. Moreover, males holding a master's degree tend to prefer Nissan cars. This information may be used to further define the clothing preferences of the customer. For example, a person driving a sport car would be more likely to buy a more casual suit than a person driving a more classic car.

For the **American female** population we observe that they have some patterns in common with the male population. We found that the income of a married female is higher than her single counterpart. Unlike the male population where a male younger than 30 years is single, 25 aged or older females are usually married. We also noticed that the fitness level of a single female is higher than a married female. Moreover, a single female holding a bachelor degree usually has medium or high level of fitness. A clothing company then may want to design e.g. sportswear for this group; or a gym may target them for advertising a special fitness program.

As observed in the male population, and as may be expected, there is a pattern indicating that females whose occupation is management have high income. Once more, a clothing company may design expensive professional clothes to target this segment of the population. Unlike the male population, a married female holding a master's degree has an annual income over $100,000. A further study of this segment of the population may reveal lifestyle choices that may make them potential customers for certain products. Another interesting pattern holding for the female population is that women with annual income equal or higher than $60,000, is a potential customer for buying a new car.

In general, married **Dutch males** have higher income than single males. An exception to this pattern may be observed for the Small sized males, where a 20-30 single male have a similar income than a 36-40 aged married male. The key factors for this exception

may be the difference in age and level of education. We also observe, in contrast to the American males, that for the Medium sized males, a married 26-30 aged male holding a bachelor degree earns significantly more than a married male with the same background but 31-35 aged. For the XX Large sized males the following patterns were observed. A married male with middle-level education and medium level of fitness have higher income than a male with the same background but with lower level of fitness. Moreover, some rules indicate that an XX Large sized and 56-60 aged male has a high probability of being unemployed. Unlike the American male population, for the Dutch males there is no clear relationship between the level of fitness and the marital status or the number of children. That is, fitness seems to be independent from the marital status or the number of children. Another interesting pattern for the Dutch male population is that some rules indicate that the income is related to the place of residence. In general, people with high annual income live in Utrecht, a useful fact for targeted marketing.

For the **Dutch female** population, we observe there is some relationship between the size and the age. In general younger females tend to be smaller in size, while older females are larger in size. Similar to the Dutch male population, there is no clear relationship between the level of fitness and the number of children or the marital status. The pattern between the marital status and the income is also present for the females. That is, single females earns on average less than the married counterpart. Also, it can be seen that a married female who is 30 years or older and is XSmall or Medium in size has a high income. A clothing company may wish to investigate further this segment of the population and design exclusive garments for them. Similar to the Dutch male population, there is a strong pattern between the annual income and the place of residence. Again, Utrecht is the place where people have higher income. This is the case for married Medium sized females whose income is in the range of $77,000-103,000.

## 7. DISCUSSION

Our results indicate that the three populations used in our study have different body profiles, which should be taken into account when designing clothes. Furthermore, the crucial reduced body measurements where not the same. Our evaluation of the demographic profiles of the various consumer segments, in order to facilitate better targeted marketing and advertising campaigns, also pointed to differences in the demographic natures and lifestyle choices of the different populations.

It follows that even the most esthetically pleasing clothes will not sell, if they do not fit the population. If a clothing company does not understand their market, they may

overproduce clothing, or produce clothes for a non-existing market. By streamlining their markets by combining anthropometric and demographic profiling, clothing returns are minimized, consumer satisfaction is enhanced and a better correspondence between advertisement and consumer satisfaction is achieved. Also, marketing campaign focusing on typical consumers' profiles holds obvious benefit.

For example, consider a scenario where we aim to design clothes for e.g. XX Large sized married Dutch males aged 45-50. From our study we know XX-Large individuals are difficult to characterize and therefore to produce garments that fit them properly. From the demographic analysis we found XX-Large married Dutch males aged 45-50 reported high income and may be the ideal market for expensive tailor-made suits. We may then subsequently decide to conduct an initial direct marketing and advertising campaigns in Utrecht, to target possible new consumers.

## 8. CONCLUSIONS

This paper described a study aiming to find, and typify, apparel consumers. In order to produce well-fitting clothes, the different sizes must correspond to real body shapes, in the sense that one or more archetypes should represent the individuals belonging to the same size accurately. Consequently, it is important to define clusters that may be characterized by one archetype, i.e. a truly representative of all other individuals that belong to the same cluster. Based on the verified assumption that the cluster has a convex or quasi convex symmetry, the archetype then corresponds to the closest individual to the Centroid of the cluster. We might also choose one of the individuals belonging to the sub-region with the highest density in terms of number of individuals. If the resulted clusters are not convex, more than one archetype might be necessary to fully characterize the cluster. In the context of tailoring, however, the optimal scenario is to cover the largest number of people with the fewest number of sizes. In this context then, it is preferred to have only one archetype, since each new size or sub-size involves more tailoring and increases the complexity in the manufacturing.

The method we utilize in this work satisfies the aforementioned requirements, since we were able to group the individuals into clusters with a well defined Centroid. Our verification, when using the Cleopatra system, indicates that the cluster membership corresponds to the reality, in the sense that the bodies correspond to our expectations of the cluster membership. Also, our results indicate that the number of body measurements may be significantly reduced by applying interestingness measure-based feature selection and feature extraction. Moreover, these new sets of reduced body measurements improve

the predictive accuracy. These sets therefore contain the most important body measurements for defining the body sizes, and may be used in garment design to identify those body measurements that require special attention, when tailoring clothes for a specific population and gender. We also analyzed the demographic data to better understand the individuals within each size. This analysis allowed us to identify potential target markets for different member of the clothing industry, ranging from haute couture to mass market manufacturers.

Future work will involve investigating the use of alternative data clustering approaches, such as quantum evolutionary algorithms (Ramdane, Meshoul, Batouche and Kholladi, 2010), dynamically growing self-organizational trees such as DGSOT (Luo, et al., 2004), and subspace k-means clustering algorithms such as SMART (Jing, et al., 2009). In this paper, we employed wrapper methods for feature selection. We are also interested in considering the use of alternative feature selection approaches based on correlation and independence analyses (Witten and Frank, 2005).

## REFERENCES


Agrawal, R. and Shirkant, R., 1994. Fast algorithms for mining association rules in large databases. In VLDB '94, pp. 487–499.

Ashdown, S. Loker, S. and Rucker, M., 2007. Improved Apparel Sizing: Fit and Anthropometric 3-D Scan Data. *Annual Report NTC Project: S04-CR01-07*, National Textile Center

Cunningham, P., 2007. Dimension Reduction. *Technical Report UCD-CSI-2007-7*, University College Dublin, pp. 1–24

Desmarteau, K., 2000. CAD: Let the Fit Revolution Begin. *Bobbin*, vol. 42, pp. 42-56

Geng, L, and Hamilton, H.J., 2006. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.* vol. 38, no. 3, pp. 1-32

Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*, San Francisco, USA: Morgan Kaufmann

Houle, M. E. et al., 2010. "Can Shared-Neighbour Distances Defeat the Curse of Dimensionality?", SSDBM 2010: 21th International Conference on Scientific and Statistical Database Management, LNCS 6189, pp. 482-500.

Hsu, C.H., Lin, H.F. and Wang, M.J. 2007. Developing Female Size Charts for Facilitating Garment Production by Using Data Mining. *Journal of Chinese Institute of Industrial Engineers*, vol. 24, no. 3, pp. 245–251

Jing, L. et al., 2009. SMART: a subspace clustering algorithm that automatically identifies the appropriate number of clusters, Int. J. Data Mining, Modelling and Management, Vol. 1, No. 2, pp. 149-177.

Kim, Y., Street, W.N. and Menczer, F. 2003. Feature Selection in Data Mining. *Data mining: opportunities and challenges*, pp. 80–105

Luo, F. et al., 2004. A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles. Bioinformatics, Vol. 20, No. 16, pp. 2605-2617.

Mcgarry, K. 2005. A Survey of Interestingness Measures for Knowledge Discovery. *Knowl. Eng. Rev.*, vol. 20, no. 1, pp. 39–61

Paquet, E., Robinette, K.M. and Rioux, M. 2000. Management of Three-Dimensional and Anthropometric Databases: Alexandria and Cleopatra. *Journal of Electronic Imaging*, vol. 9, pp. 421–431

Peña, I., Viktor, H.L. And Paquet, E. 2009A. Finding Clothing that fit through Cluster Analysis and Objective Interestingness Measures. *DaWaK 2009*, LNCS 5691, pp. 216–228

Peña, I., Viktor, H.L. And Paquet, E. 2009B. Explorative Data Mining for the Sizing of Population Groups, KDIR 2009, pp. 152-159

Ramdane, C., Meshoul, S., Batouche, M. and Kholladi, M-K. ,2010. A quantum evolutionary algorithm for data clustering, Int. J. Data Mining, Modelling and Management, Vol. 2, No. 4, pp. 369-387.

Robinette, K.M. et al., 2002. Civilian American and European Surface Anthropometry Resource (CAESAR). Final Report, Volume I: Summary. AFRL-HE-WP-TR-2002-0169, United States Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division

Scheffer, T. 2001. Finding association rules that trade support optimally against confidence. *In PKDD' 01*, pp.424–435

Schofield, N.A. and Labat, K.L. 2005. Exploring the Relationships of Grading, Sizing and Anthropometric Data. *Clothing and Textiles Research Journal*, vol. 23, no. 1, pp. 13–27

Sheikh, L.M., Tanveer, B. and Hamdani, M.A. 2004 Interesting measures for mining association rules. INMIC 2004, pp. 641–644.

SizeChina. 2009. China National Sizing Survey. Resource available at http://www.sizechina.com/

SizeUSA. 2009. The US National Size Survey. Resource available at http://www.sizeusa.com/

Veitch, D., Veitch, L. and Henneberg, M. 2007. Sizing for the Clothing Industry Using Principal Component Analysis - An Australian Example. *Journal of ASTM International (JAI)*, vol. 4, no. 3, 12 pp

Viktor, H.L., Paquet, E. and Guo, H. 2006. Measuring to Fit: Virtual Tailoring Through Cluster Analysis and Classification. *PKDD 2006: Knowledge Discovery in Databases*, pp. 395–406

Weiss, G., Saar-Tsechansky, M. and Zadrozny, B. 2005. Report on UBDM-05: Workshop on Utility-Based Data Mining. *SIGKDD Explor. Newsl.*, vol. 7, no. 2, pp. 145–147

Weiss, G. and Tian, Y. 2008. Maximizing classifier utility when there are data acquisition and modeling costs. *Data Min Knowl Disc*, vol. 17, no. 2, pp. 253-282

Witten, I.H. and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, San Francisco, USA: Morgan Kaufmann

Zadrozny, B., Weiss, G. and Saar-Tsechansky, M. 2006. UBDM 2006: Utility-Based Data Mining 2006 Workshop Report. *SIGKDD Explor. Newsl.*, vol. 8, no. 2, pp. 98–101.

## APPENDIX A

Analysis of the measurements of the archetypes for the three population groups.

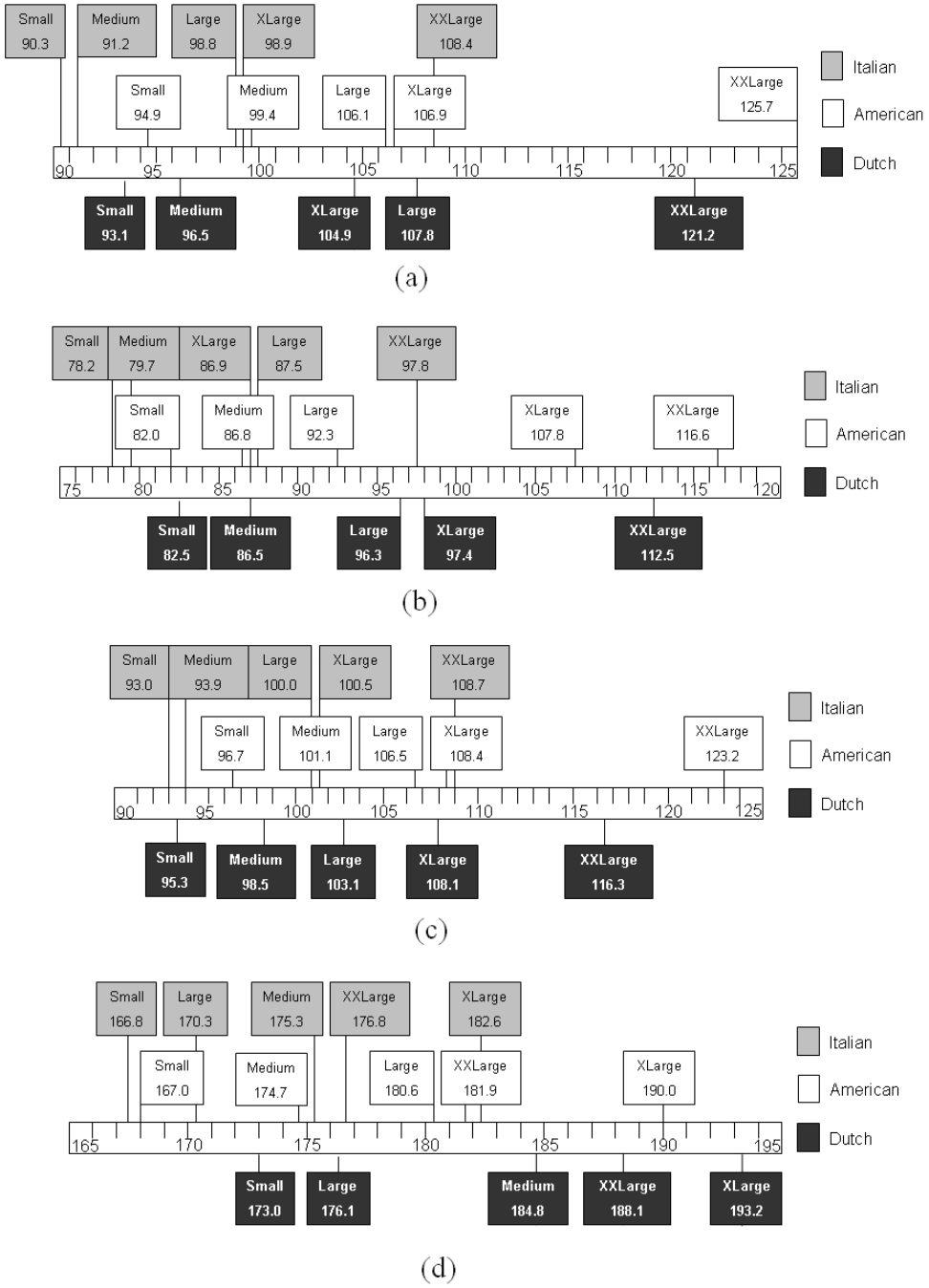
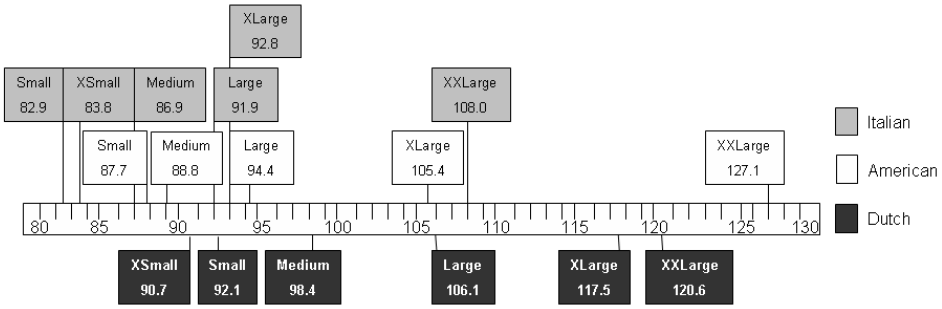
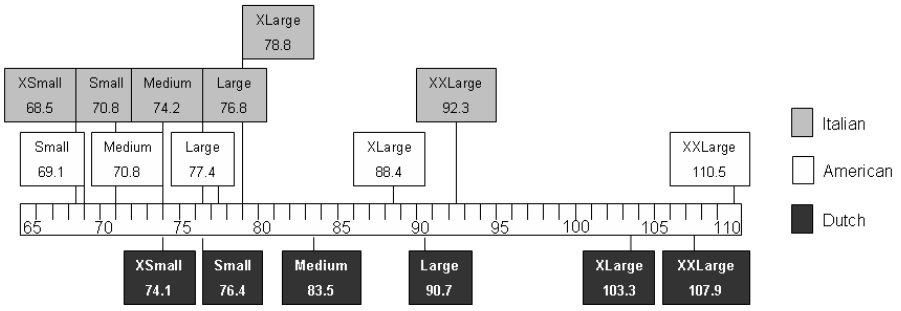


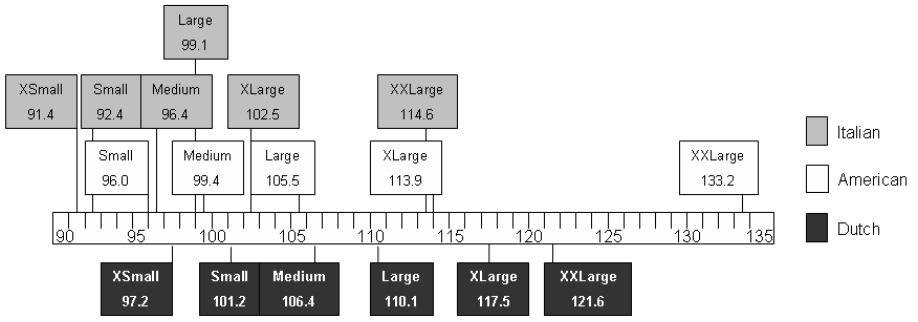

Fig. A1. Analysis of the male population. The figure shows the mean values corresponding to each Centroid or Archetype according to: (a) Chest Circumference, (b) Waist Circumference, (c) Hip Circumference and (d) Stature



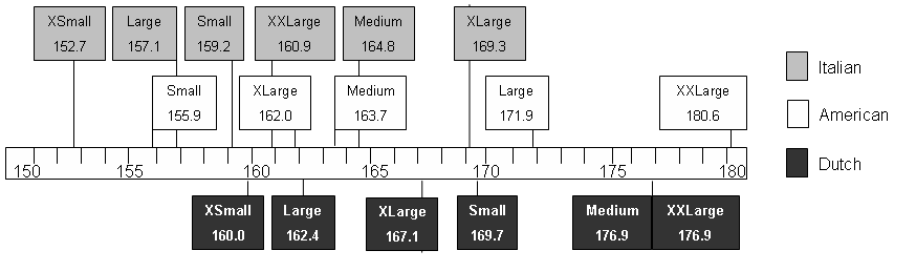

(a)




(b)




(c)




(d)

Fig. A2. Analysis of the female population. The figure shows the mean values corresponding to each Centroid or Archetype according to: (a) Bust Circumference, (b) Waist Circumference, (c) Hip Circumference and (d) Stature