Projection of anthropometric correlation for virtual population modelling

John Rasmussen*
Department of Materials and Production,
Aalborg University,
Aalborg, Denmark
Email: jr@mp.aau.dk
*Corresponding author

Rasmus Plenge Waagepetersen
Department of Mathematical Sciences,
Aalborg University,
Aalborg, Denmark
Email: rw@math.aau.dk

Kasper Pihl Rasmussen
AnyBody Technology A/S,
Niels Jernes Vej 10,
DK-9220 Aalborg, Denmark
Email: kpr@anybodytech.com

Abstract: A new statistical method for generation of virtual populations based on anthropometric parameters is developed. The method addresses the problem that most anthropometric information is reported in terms of summary data such as means and standard deviations only, while the underlying raw data, and therefore the correlations between parameters, are not accessible. This problem is solved by projecting correlation from a data set for which raw data are provided. The method is tested and validated by generation of pseudo females from males in the ANSUR anthropometric dataset. Results show that the statistical congruency of the pseudo population with an actual female population is more than 90% for more than 90% of the possible parameter pairs. The method represents a new opportunity to generate virtual populations for specific geographic regions and ethnicities based on summary data only.

Keywords: anthropometry; human factors; statistics; principal component analysis; PCA; correlation.


Biographical notes: John Rasmussen received his PhD in 1989 from Aalborg University, Denmark, for a thesis on structural design optimisation. He is a Professor at the Department of Materials and Production at the same university. For the past decade, his research interests have been focused on...
musculoskeletal biomechanics and in particular on the development of the Anybody Modeling System and its associated models. However, his interests also include architecture, design optimisation, solid mechanics and tissue engineering.

Rasmus Plenge Waagepetersen received his MSc in 1995 and PhD in 1997 in Mathematical Statistics, both from Aarhus University, Aarhus, Denmark. He was a Scientist at the Danish Institute of Agricultural Sciences, Tjele, from 1997 to 2000, and post doc at Chalmers University of Technology, Gothenburg, Sweden, in the last half of 2000. From 2000 to 2007 he was an assistant and subsequently Associate Professor at the Department of Mathematical Sciences, Aalborg University, Denmark. In 2008–2009 he was a Credit Risk Analyst at Spar Nord Bank, Aalborg. Since 2010 he has been a Full Professor at the Department of Mathematical Sciences, Aalborg University. His main research interests are within spatial and computational statistics. He is a co-author of the Monograph Statistical Inference and Simulation for Spatial Point Processes, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. He is a member of the International Statistical Institute and the Bernouilli Society.

Kasper Pihl Rasmussen received his MSc in Sports Technology from Aalborg University, Aalborg, Denmark, in 2015. His research interests include statistical models of humans for automotive package design. He is currently with AnyBody Technology A/S.

1 Introduction

Anthropometry plays a key role in the fields of ergonomics and industrial product design. Many categories of industrial products derive their value from an ability to interact with humans. Typical examples are vehicles, bicycles, furniture, tools, workplaces, sports equipment and packaging for consumer products. There is also a growing understanding of the importance of compatibility between biomedical products, such as joint replacements and trauma devices, and the bodies into which they are implanted (Palumbo et al., 2012). For most industrial and biomedical products, specific manufacture for a single user is not feasible, and anthropometric compatibility must be obtained by incorporating adjustment into the product design, such as the height of a bicycle saddle, or by production of a series of sizes that will accommodate the majority of the target population, such as shoes. Accurate and updated anthropometric information for the targeted population is crucially important for such products to be successful.

It is possible to investigate the product’s interaction experimentally with a small number of subjects representing the boundary of the targeted population, for instance a 5th percentile female and a 95th percentile male in terms of stature only. Small series of test subjects representing a range is another option. However, neither approach yields a representation of the full anthropometrical variation over the population (Högberg et al., 2015).

For some time, CAD-integrated digital manikin systems, such as RAMSIS®, JACK® and DELMIA®, have been available and have offered the opportunity to incorporate anthropometric variation into computer models. More recently, musculoskeletal modelling systems, such as AnyBody (Damsgaard et al., 2006) and OpenSim (Delp et al., 2007), enable simulation of the detailed biomechanics of the interaction between virtual
humans and products at different scales from full body over specific limbs to single joints or muscles. Although technical problems persist, the goal of digital models is to make virtual experiments simple and inexpensive. If successful, the focus shifts from the execution of the experiments to the generation of a virtual population that validly represents the real population of interest.

This defines the challenge of creating virtual populations representative for particular geographic or ethnic groups. Unfortunately, this task is non-trivial. Anthropometric data collection is expensive and time-consuming, and populations change dimensions over time (de Vries et al., 2010), so data collection must be repeated at regular intervals. It is therefore essential to exploit accessible data as well as possible.

Accessible anthropometric databases vary in collection methods, recorded parameters and presentation of data. In terms of the latter, two categories can be identified.

1. **Raw data** of the measured population, i.e., a table of measured dimensions for each individual.

2. **Summary data**, i.e., means and standard deviations, for the reported dimensions.

The number of accessible anthropometric data sets in the first category, i.e., containing raw data, is quite small. For detailed human anthropometry, the options are in practice limited to the Anthropometric Survey of US Army Personnel (Gordon et al., 1989), better known as ANSUR, and US Civilian and European Surface Anthropometry Resource, CAESAR (Harrison and Robinet, 2002; Robinet and Fleming, 2002). Raw ANSUR data in the form of measured dimensions of individuals have been made accessible on the internet. CAESAR contains measured dimensions as well as three-dimensional body scans from which, in principle, any distinguishable dimension can be derived. It is a commercial product marketed by the society of automotive engineers (SAE). The direct validity of these two databases is naturally limited to the populations and periods in which they were collected.

In contrast to ANSUR and CAESAR, the bulk of accessible databases present information in terms of summary data only, covering different populations and different subsets of measured dimensions. The National Health and Nutrition Examination Survey (NHANES) (Calafat, 2006) is an American data set covering a large number of subjects but rather few reported dimensions. DINED (Molenbroek, 2004) is an anthropometric database representing the Dutch population and DinBELG (Motmans, 2005) similarly for the Belgian population. Adult data, child data and older adult data (Peebles and Norris, 1998) compile data from many different geographic regions, while the Digital Human Research Center AIST (Kouchi et al., 2000) manages a data set for the Japanese population. Finally, human factors textbooks such as Pheasant and Hazelgrave (2005) and Chaffin et al. (2006) contain tables of summary data compiled from a variety of sources.

Common to these data sets is that they report means and standard deviations of each variable, implicitly assuming that the measured variables are normally distributed. The mean values provide a coherent description of the dimensions of the average person in the dataset, and the standard deviations describe the variability of the individual parameters over the population, so it would appear that these summary data describe the population well. However, summary data fail to capture correlations between parameters.

Figure 1 illustrates the strong correlation between two parameters from the ANSUR data set. The central cluster of points represents the measured data, and the two square
outliers represent plausible results of uncorrelated random variations of the two parameters. It is obvious that generation of a pseudo population by random and uncorrelated variation of the two parameters within their respective normal distributions will result in a circularly shaped pseudo cluster of points that is quite different from the real data and therefore does not represent the population validly.

Figure 1 Scatter plot of two dependent parameters from ANSUR, stature and butt height, with their respective univariate distributions indicated on the axes (see online version for colours)

Notes: The circular dots are the actual data points. The square outliers are highly improbable in the population but are probable results of an independent, random variation of the two parameters.

The goal of this paper is to develop an operational procedure that generates valid, virtual populations matching summary, population-specific target data. This is accomplished by harvesting correlation from a detailed reference data set. The underlying assumption is that the correlation matrix from the reference data set can acceptably represent the inaccessible correlation matrix of the target population. Inter-population differences in anthropometry are known to be significant, but it is possible that different populations exhibit similarities in terms of their correlation matrices. For instance, it is likely that all human populations will share a positive correlation between stature and leg length.

The RASH-method (Pheasant and Hazelgrave, 2005) uses scaling to predict unknown anthropometrical parameters by assuming certain linear relationships between pairs of variables within the same population, e.g., a linear relation between sitting height and stature. Later, more sophisticated, hierarchical regression models based on the ANSUR data set have been developed (Jung et al., 2009; You and Ryu, 2005) to enhance the reliability of prediction of unknown parameters from a few known parameters, e.g., stature and body weight. The resulting methods can generate random dimensions of a virtual person with given input parameters, i.e., predict many parameters from a few. A related problem was addressed by Parkinson and Reed (2010) using principal component
analysis (PCA) to enable the projection of rich data sets like ANSUR and CAESAR onto more updated data with fewer parameters, such as NHANES. Similarly, prediction of missing dimensions from known dimensions was done by linear regression (Brolin et al., 2016) and by PCA (Brolin, 2016).

The scope of this paper is different. Rather than predicting unknown parameters from known parameters within the same data set, we consider the problem of generating a valid pseudo population given summary target data and a rich reference data set offering the same parameters as the target. We aim to make the approach general and non-hierarchical, i.e., the approach should not require selection of any particular predictive variables, and it should work for any rich reference and summary target data sets, provided they offer measures of the same dimensions. We validate the method by prediction of known data, i.e., by attempting to generate a virtual female population from a male reference population.

2 Methods

Our approach is divided into two separate steps:

1. Determine an affine transformation that maps the reference data set into a new data set whose means and standard deviations agree with those of the target data set.

2. Use of the mapped data to generate a virtual population with an arbitrary, user-selected number of individuals by means of PCA and assuming a multivariate normal distribution of the variables in the target population.

If the target data set obtained in the first step is deemed sufficient for the analyses of interest, then the second step may be skipped. However, if the target data set is deemed too small in size the second step can be used to up sample the data in order to obtain a better coverage of the sample space. This is of course at the price of the assumption of a multivariate normal distribution. The second step may also be applied for reasons of ethics and data protection, since individuals in the reference data set cannot be inferred from the generated virtual population.

2.1 Mapping a reference data set

We shall assume that we have a reference population, such as ANSUR, defined by a matrix $X$ of $m$ anthropometric dimensions of $n$ measured individuals:

$$X = [x_{ij}], i \in 1 \ldots n, j \in 1 \ldots m$$

For each column of $X$, we compute empirical means and variances:

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \quad s^2_{i,j} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$$

Similarly, we compute the entries of the covariance matrix of $X$:

$$S_{pq} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{iq} - \overline{x}_q)(x_{ip} - \overline{x}_p), \ p, q \in 1 \ldots m$$
from which the entries of the correlation matrix are derived:

\[ C_{i,pq} = \frac{S_{i,pq}}{\sqrt{S_{i,i}^2 \cdot S_{p,q}^2}} \]  

(2)

For the target population, we know the empirical means \( \bar{y}_j \) and variances \( s_{y,j}^2 \) but not individual measurements \( y_{ij} \). However, we introduce the key assumption that to each record \( x_j \) in the reference population there corresponds a record \( y_{ij} \) in the target population given by the affine relation.

\[ y_{ij} = a_{ij}x_j + b_j \]  

(3)

where \( a_{ij} \) and \( b_j \) are unknown coefficients. We determine these coefficients by matching the known target means and variances with those obtained using equation (3):

\[ \frac{1}{n} \sum_{i=1}^{n} (a_{ij}x_j + b_j) = \bar{y}_j \quad \frac{1}{n-1} \sum_{i=1}^{n} (a_{ij}x_j + b_j - \bar{y}_j)^2 = s^2_{y,j} \]

Solving for the coefficients, we obtain

\[ a_{ij} = \frac{\left( \sum x_j \right)^2 - n \sum x_j^2}{\left( \sum x_j^2 \right)^2 - n \sum x_j} \cdot \frac{s^2_{y,j}}{s^2_{x,j}}(1-n)n \]

\[ b_j = \bar{y}_j - a_{ij}\bar{x}_j \]

Having thus determined \( a_{ij} \) and \( b_j \), we can generate an approximation \( Y \) of the target population using the affine relation equation (3), i.e.,

\[ y_{ij} = \frac{s_{y,j}}{s_{x,j}}x_j + \bar{y}_j - \frac{s^2_{y,j}}{s^2_{x,j}}\bar{x}_j = \bar{y}_j + s_{y,j} \frac{x_j - \bar{x}_j}{s_{x,j}} \]  

(4)

Another interpretation of the method is that we transform the \( x_j \) into standardised variables \( z_{y,i} = (x_j - \bar{x}_j)/s_{x,j} \) of mean zero and standard deviation one. Next, the inverse transform is applied to the \( z_{y,i} \) using \( \bar{y}_j \) and \( s_{y,j} \) to obtain \( y_{ij} \) in equation (4).

Using that each \( y_{ij} \) in \( Y \) is given by \( y_{ij} = a_{ij}x_j + b_j \) and that \( s^2_{y,j} = a^2_{ij}s^2_{x,j} \), it is easy to check that the empirical correlation matrix of \( Y \) coincides with that of \( X \) equation (2), i.e., correlation [but not covariance equation (1)] remains constant across an affine transformation. In other words, \( Y \) possesses a realistic correlation structure imported from \( X \).

2.2 Generation of a virtual population from a rich data set

Generation of a virtual population \( Z \), with an arbitrary number of individuals and validly representing the target population, can be accomplished by random sampling. We assume that virtual individuals are multivariate normal distributed with means, standard deviations and correlations identical to those of \( Y \). We then determine an invertible linear transformation that maps the variables of a virtual individual into a set of independent
normal variables. Simulation of these independent transformed variables is straightforward. A simulation of a virtual individual in $Z$ is thus easily obtained by applying the inverse linear transformation to the simulated independent variables. Practically, the linear transformation is constructed by Cholesky decomposition or PCA applied to $Y$. A typical PCA implementation also sorts the aforementioned independent variables in descending order in terms of variance and provides information about the amount of variance in the data set represented by each variable. This carries the additional opportunity of data reduction in the sense that most of the variance can usually be represented by the first few transformed variables. However, data reduction is not the scope of the present work, and we use PCA only in the interest of creating a virtual population by independent sampling.

The widespread use of PCA for multiple purposes has brought about extremely efficient and robust numerical implementations of PCA transformation. The computer language Python has been widely adopted by the scientific community in recent years based on an open source strategy. Different collaborative development projects produce high-quality libraries of algorithms for the Python community. Two of these projects are Pandas (McKinney, 2012) for data structures and data analysis tools and scikit-learn (sklearn) (Pedregosa et al., 2011) for machine learning tasks. PCA is central to the latter, and sklearn consequently provides a robust and efficient set of PCA algorithms. We have used Pandas and sklearn to develop a compact application for generation of a virtual population up-sampled to any number of individuals representative of the ANSUR data set. This will be demonstrated in the results section.

2.3 Validation

For validation purposes, we shall use the ANSUR data set (Gordon et al., 1989) because it is publicly accessible, rich in parameters and contains enough subjects to adequately estimate the correlation structure. ANSUR’s division into male and female populations represents a convenient opportunity for predicting one gender from another while using real data for the predicted gender as the gold standard. Prediction of female anthropometry from male subjects (or vice versa) is relatively challenging, because the anthropometric differences between men and women, even within the same ethnicity, appears to be comparable to or larger than the differences between most ethnic or geographic populations within the same gender.

From the raw ANSUR data, we remove categorical variables, such as ethnicity and platoon numbers, and subjects for whom one or more parameters are missing. This leaves two data sets for males and females with respectively 1,746 and 2,205 subjects. The two data sets share the same 131 anthropometric parameters.

We then compute summary data from the raw female ANSUR data, project the raw male data on them, and compare the result with the original female data. The mapped females are re-sampled by means of PCA, thus generating a female pseudo population, and again compared with empirical data.

To assess the validity of the correlation structure of the virtual population, we compare distributions of individual variables in the empirical female data and the virtual
female population both in terms of tail quantiles and congruency of two-dimensional combinations, which are explained further in the following.

A two-dimensional combination can, for instance, be stature against body mass. Provided all parameters are normally distributed, which is assumed for the reference data and ensured by the re-sampling method for the pseudo-data, combinations of any two parameters, such as illustrated in Figure 1, can be investigated for congruency by considering the relative area of overlap of 95% probability ellipses fitted to each of the clusters of points (Dueser and Shuggart, 1979).

The 131 parameters can be combined into 8,515 different parameter pairs, each of which displays more or less overlap. We shall use results for all combinations as a metric of the validity of the projection method and calculate ellipses and their relative overlaps using the open-source R (R Core Team, n.d.) package SIBER (Jackson et al., 2011).

### 3 Results

Table 1 compares selected tail quantiles between the virtual population and the female data for the variables considered in the pairwise plots below as well as stature.

<table>
<thead>
<tr>
<th>Variable</th>
<th>1%</th>
<th>95%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gluteal F. Ht.</td>
<td>64.8</td>
<td>65.1</td>
<td>67.7</td>
<td>67.4</td>
</tr>
<tr>
<td>Head Breadth</td>
<td>13.3</td>
<td>13.4</td>
<td>13.7</td>
<td>13.6</td>
</tr>
<tr>
<td>Chest Circ. B. B.</td>
<td>67.3</td>
<td>65.3</td>
<td>69.6</td>
<td>68.9</td>
</tr>
<tr>
<td>Chest Circ.</td>
<td>78.5</td>
<td>75.5</td>
<td>81.4</td>
<td>80.7</td>
</tr>
<tr>
<td>Patella M. Ht.</td>
<td>40.0</td>
<td>39.9</td>
<td>41.7</td>
<td>41.5</td>
</tr>
<tr>
<td>Butt height</td>
<td>74.0</td>
<td>73.6</td>
<td>76.8</td>
<td>76.5</td>
</tr>
<tr>
<td>Stature</td>
<td>148.7</td>
<td>148.4</td>
<td>152.9</td>
<td>152.2</td>
</tr>
</tbody>
</table>

Note: For each variable and percentage, first quantile is for data and second for virtual population.

Figures 2 and 3 show the parameter pairs with largest respectively smallest relative overlap of ellipses from the original ANSUR female population and the mapped female population generated from the correlation in the male data set. Similar comparisons between the original females and the resampled pseudo population are depicted in Figures 4 and 5.
Figure 2  Scatter plot of the parameter pair, gluteal furrow height against head breadth, with the best congruency between the ANSUR female subjects, Y (crosses), and the female subjects mapped from males, (dots) (see online version for colours)

Notes: Please refer to the ANSUR documentation for the precise definition of the anthropometric parameters.

Source: Gordon et al. (1989)

Figure 3  Scatter plot of the parameter pair, chest circumference below bust and chest circumference (i.e., bust circumference), with the worst congruency between the ANSUR female subjects, Y (crosses), and the female subjects mapped from males (dots) (see online version for colours)

Notes: Please refer to the ANSUR documentation for the precise definition of the anthropometric parameters.

Source: Gordon et al. (1989)
Figure 4  Scatter plot of the parameter pair, gluteal furrow height against head breadth, with the best congruency between the ANSUR female subjects, Y (crosses), and the resampled pseudo female subjects, z (dots) (see online version for colours)

Notes: Please refer to the ANSUR documentation for the precise definition of the anthropometric parameters.

Source:  Gordon et al. (1989)

Figure 5  Scatter plot of the parameter pair, chest circumference below bust and chest circumference (i.e., bust circumference), with the worst congruency between the ANSUR female subjects, Y (crosses), and the pseudo female subjects, z (dots) (see online version for colours)

Notes: Please refer to the ANSUR documentation for the precise definition of the anthropometric parameters.

Source:  Gordon et al. (1989)
Figure 6  Example of a highly correlated parameter pair, patella mid height and butt height, with a large degree of overlap between the ANSUR female population and the pseudo population (see online version for colours)

Figure 7  For each pair of variables, the plot shows the relative elliptical overlap between the female and the pseudo female population plotted against the correlation coefficient for the pair in the female population (see online version for colours)
The elliptical congruency in Figure 4 is 99.5% while the value for Figure 5 is only 51%. The low congruency of the combination of Figure 5 is an exception; more than 90% of the combinations have congruency better than 90%.

Figures 4 and 5 may indicate that it is more difficult to obtain a high degree of elliptical overlap for parameter combinations with high correlation, but this is not necessarily the case as illustrated by Figure 6, which shows the overlap of a highly correlated parameter pair, namely butt height versus patella mid height; its overlap is 99%.

An overview of the dependency between the parameter pairs' correlation and the ability of the method to predict them correctly is presented in Figure 7 and discussed further in Section 4.

4 Discussion

The agreement between the quantiles in Table 1 is, in general, very good. The variables chest circumference and chest circumference below bust display the larger deviations. If the female variables were perfectly normally distributed, then deviations would only be due to sampling error of the reported quantiles since quantiles of a normal distribution is uniquely determined by the mean and the standard deviation. Indeed, deviations from normality can be observed for chest circumference and chest circumference below bust while, for example, the distribution of stature is very close to normal.

Regarding bivariate distributions, chest circumference and chest circumference below bust further display the statistically worst fit out of 8,515 combinations with only 51% congruency. The data are depicted in Figures 3 and 5. Closer examination reveals that the point cluster of the pseudo population is more narrow, i.e., displays a stronger correlation than the original female data.

Prediction of female anthropometry from male anthropometry appears to be harder for dimensions of large gender differences, and it is important for the assessment of the method to know how often they occur. To this end, Figure 7 reveals that there are only few parameter combinations that have overlaps less than 70%, and more than 90% of the combinations have relative congruency better than 90%. Although poor relative congruency is exclusively found for highly correlated parameter combinations, most of the highly correlated combinations still have congruency better than 90%, such as Figure 6.

The results presented above confirm two findings:

- Despite differences in anthropometry, male and female populations appear to share correlation to a large extent, and it is not infeasible to apply the correlation of one population to another. This is encouraging because the anthropometrical difference between men and women is probably larger than the differences between most ethnic populations.
- Correlation is unaffected by affine mapping of the data, which allows us to generate pseudo populations by projection of existing reference populations on summary data for a target population.
It is tempting to think that the quality of the pseudo population with respect to the real female population that it simulates can be assessed simply by comparing correlation coefficients between the parameter pairs of the two populations, rather than resorting to elliptical overlap. However, Figure 4 illustrates why this is insufficient: Pairs of weakly correlated parameters form almost circular point clusters, which have high overlap despite different correlation coefficients. The pair WAIST_NAT_LNTH and WAIST_HT_SIT_NATURAL is a case-in-point. The correlation of this pair is –0.11 for women and 0.25 for pseudo women, but the point clouds are quite similar with a relative area of overlap of 86%.

5 Conclusions

The findings of this paper are encouraging for the opportunity to create valid pseudo populations for different ethnic and geographic populations based on summary data available from the literature or online databases. The method provides digital human modellers an opportunity to generate pseudo populations for specific ethnic or geographic regions for which only summary data are available. The resulting pseudo populations will benefit from realistic correlations between the anthropometric dimensions stemming from the reference population. Easier access to pseudo populations for digital human models might shift the best practices in ergonomic design from a focus on boundary individuals, typically 5th percentile females and 95th percentile males in terms of stature, to population-wide virtual compatibility tests.

The applicability of the method is not limited to anthropometry in its traditional form. The method makes no assumptions on the input data other than a physical argument for transfer of correlation from one population to another.

Acknowledgements

We thank the anonymous referees for valuable comments and for sharing their insight. This work was supported by Innovation Fund Denmark.

References

Projection of anthropometric correlation for virtual population modelling


