

# Synthesis and validation of a virtual anthropometric user population of German civilians based on an up-to-date representative dataset

S. Wischniewski<sup>1</sup>, A. Grötsch<sup>2</sup>, D. Bonin<sup>1</sup>, M. Parkinson<sup>2</sup>

## baua: Focus

Representative anthropometric data are necessary for the virtual planning, design and construction of safe and ergonomic products and work systems and therefore for various product and manufacturing engineering processes. In this context, the human variability as well as the knowledge of correlations and multivariate relationships of different anthropometric parameters have to be integrated into early planning phases.

## Content

1	Introduction.....	1
2	Materials and Methods.....	3
2.1	Unweighting data.....	3
2.2	Creating the sample pool.....	3
2.3	Preparing the DEGS1 data.....	4
2.4	Unweighting / Anonymizing Procedure.....	4
2.5	Validation of the synthesized data.....	4
3	Results.....	3
4	Discussion/Outlook.....	8
5	Conclusions.....	8
	References.....	9

## 1 Introduction

Human-centered design for safe, healthy and competitive workplaces, person specific designs for socio-technical work systems as well as adaptive work place assistance systems require representative anthropometric data sets for the selected user population. For the prospective planning of such scenarios, Digital Human Modeling Systems (DHM-systems) are used to account for individual characteristics of the future employee in early stages of the workplace planning and production process. For a comprehensive and wholesome design of human-centered workplaces, corresponding anthropometric data is needed to be integrated into DHM-systems and the virtual planning and design process (Frohmut and Parkinson 2008; Wischniewski 2013). Publicly available web 2.0 technology opens up promising new opportunities for analyzing anthropometric data.

<sup>1</sup> Federal Institute for Occupational Safety and Health (BAuA)

<sup>2</sup> Department of Engineering Design, Mechanical Engineering and Industrial Engineering

One example is the multivariate accommodation calculator from the OPEN Design Lab at Penn State University ([www.dfnv.org](http://www.dfnv.org)).

However, source data from real populations are often not publicly available. This can be due to the proprietary nature of the data, privacy concerns, regulatory requirements, etc. In many cases, the best source of data on overall body size and shape are government-sponsored studies into the health of citizens (e.g., NHANES, DEGS1). However, the utility of these data for design is usually not an original objective in the work and their use in this manner is not covered by the data use agreements with participants. To overcome this issue, approaches are needed to create synthesized data sets that are statistically equivalent to those with restricted use. For the creation of virtual anthropometric data sets various mathematical concepts exist depending on factors and variables such as the structure and distribution, correlation information, and level of statistical detail of the given source (Parkinson et al. 2009; Nadadur, et al. 2016). Furthermore, Wischniewski et al. (2015) presented a copula based synthesis approach for the German working population (aged 18-67 years) from 1998 based on a representative anthropometric dataset.

This paper introduces a more sophisticated concept using up to date data of the German working population from 2012. Based on copulas and evolutionary algorithms, a virtual user population is created, which is highly comparable to the representative original dataset. A secondary objective is the „unweighting“ of weighted data. Many surveys are designed to be representative of a specific population. To that end, each individual in the sample is given a survey weight equivalent to their representative power within the population of interest. Loosely, the weight identifies the number of people in the target population that an individual represents within the sample. When the weights are not normalized, they sum to the number of the people in the target population.

When analysis is conducted using data from a weighted sample, the weights must be included or the results (e.g., calculated means or percentiles) are incorrect. Since many are unfamiliar with the appropriate use of statistical weights, it is often advantageous to have a representative population for whom analysis without the weights can be conducted. The objective, then, is to identify a population for whom unweighted statistical analysis produces the same results as a weighted analysis of the source population.

These two objectives, unweighting and anonymizing the data, can be worked toward simultaneously.

## 2 Materials and Methods

In this section the used methodological concepts for the aforementioned approach are presented consisting of the synthesis of the virtual population and the validation of the synthesized dataset. The measures are based on the first wave of the DEGS1 (German health and examination survey for adults) data collection (DEGS1, 2008-2011). Included are male and female participants within the range of the working age in Germany (18-67 years). The DEGS1 data is part of the national health monitoring conducted by the Robert Koch Institute (RKI) in Germany (Scheidt-Nave et al. 2012; Robert Koch Institute 2015). For the presented approach, only complete datasets for stature and weight were used and the BMI was calculated accordingly. Overall, 2.680 male and 2.963 female datasets were included in the reference source.

### 2.1 Unweighting data

One approach to unweighting data is to select individuals from a large pool of candidates, then calculate their statistical properties, compare to the source, and repeat until they are satisfactorily similar. For large studies such as NHANES, the number of sampled individuals within the source may be large enough that it can serve as its own source. Since the goals

of the NHANES effort are varied, there are many more individuals in the source than are statistically necessary to model the anthropometric variance well. For this particular situation, however, the source data (from German DEGS1) could not be used since the use agreements explicitly forbid the publication of any part of the data.

An alternate strategy is to draw individuals from an existing – but different – population. For example, individuals for the unweighted population could be selected from ANSUR, NHANES, or CAESAR. The individuals would be selected such that they match the stature, mass, and BMI summary statistics for the source population. Using this strategy ensures that the combinations of parameters – each individual in the new population – is a realistic combination. However, the combinations may differ from German civilians in critical ways. For example, the masses associated with a given stature in the ANSUR military population might be leaner than those in a civilian population.

## 2.2 Creating the sample pool

To mitigate the issues associated with unweighting data, candidate sets of anthropometry (i.e., stature, mass and BMI) can be synthesized for males and females using a copula-based technique developed at Penn State. This approach leverages the known relationships in the source data and can generate an arbitrary number of randomly selected individuals. This resulting pool has similar covariance to the source population. To ensure that additional variability is not introduced into the model, synthesized individuals with measures beyond those in the observed population (e.g., taller than the tallest individual, lighter weight than the lightest individual) are removed.

For the present work, the copula-based approach was used to generate 40,000 men and 40,000 women. The copulas synthesized stature and BMI. Mass was calculated from these two measures ( $BMI = \text{weight (kg)} / \text{stature (m)}^2$ ). Individuals will be drawn from this pool in the unweighting/anonymizing method outlined below.

## 2.3 Preparing the DEGS1 data

Candidate unweighted, anonymized target populations must be evaluated against the source weighted population. To this end, the percentiles from 1 to 99 were calculated for the weighted data. This was done for both male and female samples. The min and max (0th and 100th percentile) were omitted due to the sparsity of data in the tails.

## 2.4 Unweighting / Anonymizing Procedure

Optimization was used to systematically evaluate candidate populations and identify the best match. Specifically, the genetic algorithm (GA) within Matlab was used. Evolutionary algorithms are stochastic and usually slower than their deterministic counterparts. However, they are well-suited to large, discontinuous problems. Since the goal of the present work is to identify the 4,000 individuals that best match the requirements, there are 4,000 design variables (a relatively high number).

For each candidate population selected by the GA, the percentiles from 1-99 were calculated for each of stature, mass, and BMI. These were then compared to the target values for the source population. The objective function consisted of a measure of the absolute, normalized difference between the two populations. The GA evaluated new populations iteratively until the best one (e.g., the one that minimized the aggregate error) was identified. The parameters were set such that just over two million candidate pools were typically considered.

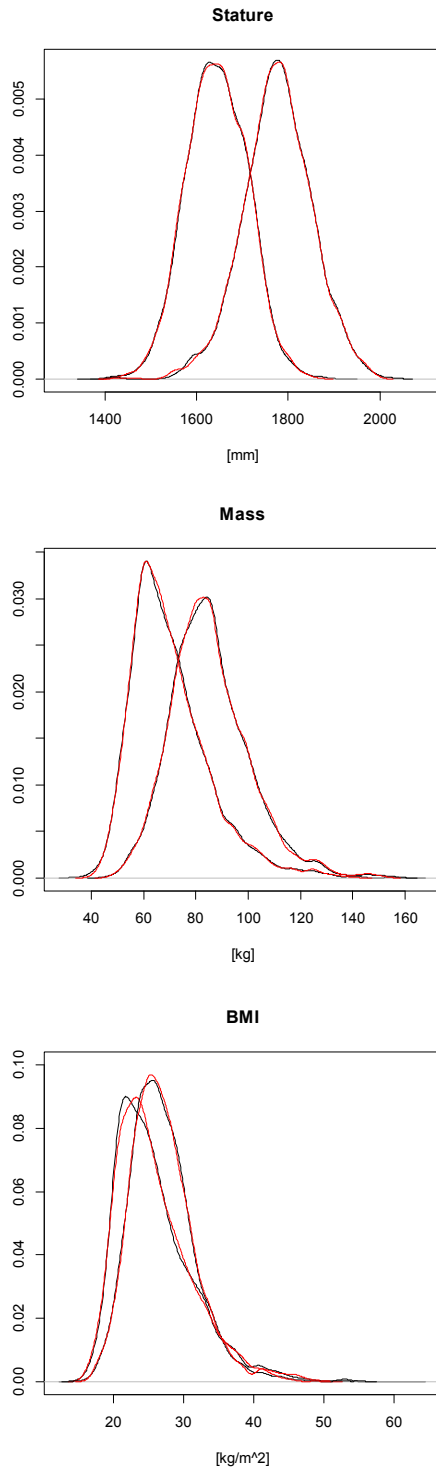
## 2.5 Validation of the synthesized data

A validation procedure was applied to compare the multivariate data sets as described in Wischniewski et al. (2015) using the statistical computing and graphics software package R (R Core Team 2013). It is based on the calculation and comparison of the difference between the total accommodation levels of source and synthesized data under predefined selection

criteria. The validation was performed for 10.000 trials, where the lower limit was set at 1st percentile and the upper limit at 99th percentile.

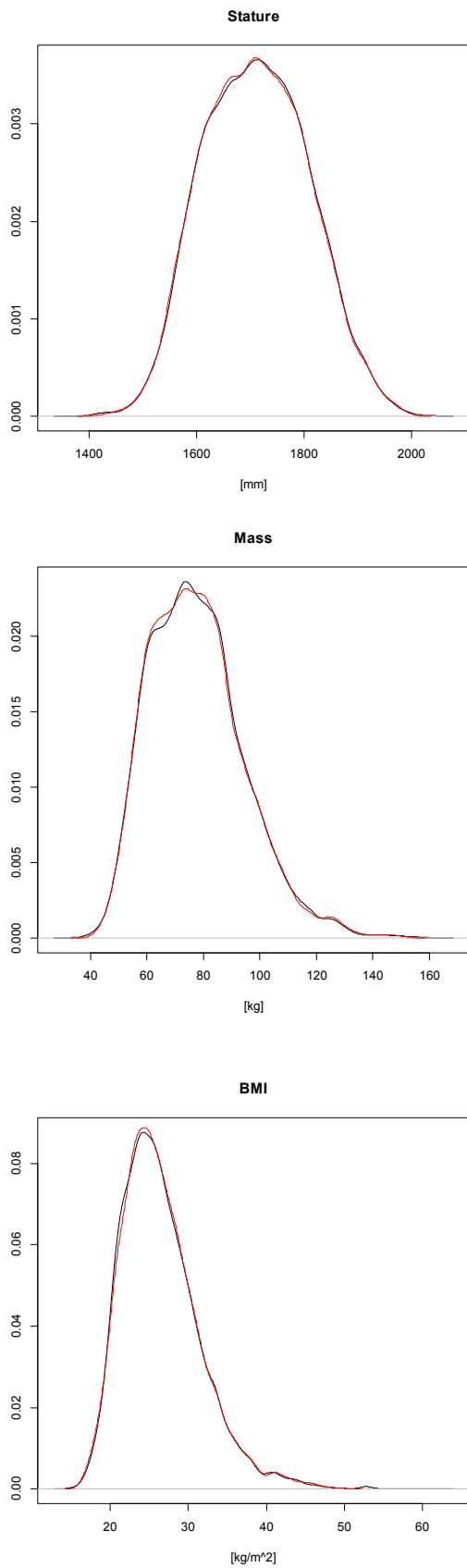
### 3 Results

Data sets of 4.000 synthesized males and synthesized females were created. Figures 1 shows the density plots for weighted source (DEGS1) and the synthesized male and female data-sets.



**Fig. 1** Density plots of weighted source (black) and synthesized dataset (red) for males (right curves) and females (left curves).

Figure 2 shows the plots for the joint male and female population.



**Fig. 2** Density plots for the combined male and female datasets of weighted source (black) and synthesized (red).

Table 1 shows the spearman correlation coefficients of the multivariate source datasets (m=male, f=female, b=both).

**Table 1** Correlation matrix of the source data (DEGS1)

	STATURE	MASS	BMI
<b>Stature</b>	1.000	.370** (m)	-.100** (m)
		.231** (f)	-.170** (f)
		.528** (b)	.020 (b)
<b>Mass</b>	.370** (m)	1.000	.863** (m)
	.231** (f)		.903** (f)
	.528** (b)		.839** (b)
<b>BMI</b>	-.100** (m)	.863** (m)	1.000
	-.170** (f)	.903** (f)	
	.020 (b)	.839** (b)	

\*\*significance level 0.01

Table 2 shows the according spearman correlation coefficients of the multivariate synthesized datasets.

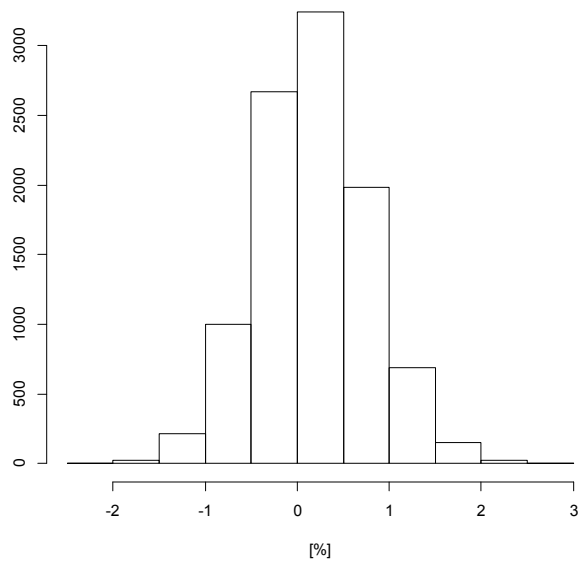
**Table 2** Correlation matrix of the synthesized data

	STATURE	MASS	BMI
<b>Stature</b>	1.000	.386** (m)	-.095** (m)
		.214** (f)	-.188** (f)
		.529** (b)	.011 (b)
<b>Mass</b>	.386** (m)	1.000	.857** (m)
	.214** (f)		.903** (f)
	.529** (b)		.834** (b)
<b>BMI</b>	-.095** (m)	.857** (m)	1.000
	-.188** (f)	.903** (f)	
	.011 (b)	.834** (b)	

\*\*significance level 0.01

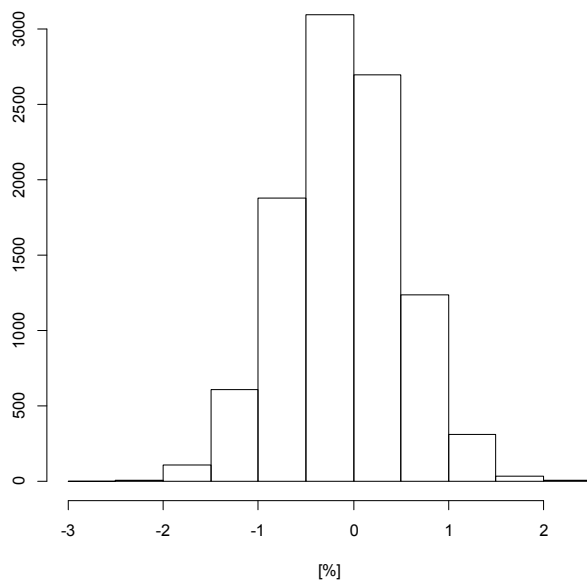
As mentioned before, another validation based on upper and lower parameter limits and resulting accommodation levels for weighted source and synthesis on top of the GA was performed for 10.000 trials; minimum lower limit was set at 1st percentile, the maximum upper limit at 99th percentile. The procedure is described in detail at Wischniewski et al. (2015). In this validation the weighted source was used.

Figure 3 shows the histogram of the calculated accommodation level differences for the 10.000 trials using the male populations. The mean difference was 0.18 % with a standard deviation of 0.6 %, minimum of -2.04 % and maximum of 2.66 %.



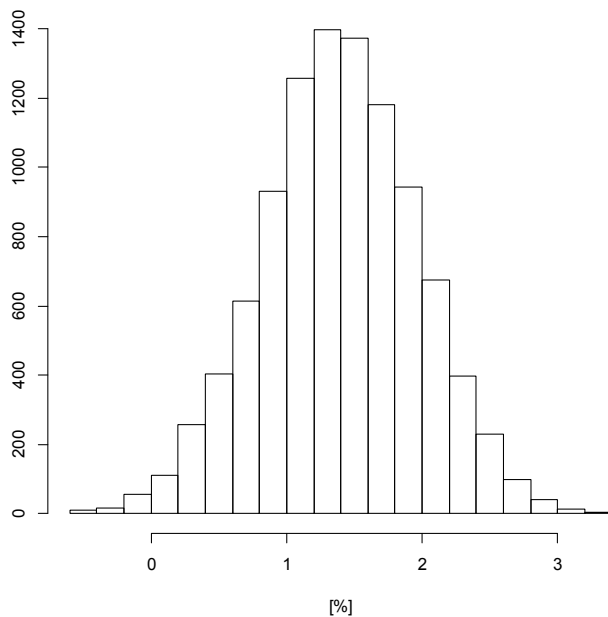
**Fig. 3** Difference in accommodation levels in % for the male populations

Figure 4 shows the histogram of the calculated accommodation level differences for the 10.000 trials using the female populations. The mean difference was 0.1 % with a standard deviation of 0.61 %, minimum of -2.54 % and maximum of 2.38 %.



**Fig. 4** Difference in accommodation levels in % for the female populations

Figure 5 shows the histogram of the calculated accommodation level differences for the 10.000 trials using the combined populations. The mean difference was 1.4 % with a standard deviation of 0.57 %, minimum of -0.6 % and maximum of 3.31 %.



**Fig. 5** Difference in accommodation levels in % for the combined populations

## 4 Discussion / Outlook

The presented method shows a high reliability of the validation and good concordance of the synthesized data compared to the source. The validation included the parameter set stature, mass and BMI due to the available data structure of the source. Accordingly, the quality of the synthesized data is assumed to be accurate to be made publicly available, e.g. at web 2.0 applications such as the aforementioned tool of the Open Design Lab at Penn State University. For the user, these online available data serve as a basis for the implementation of up-to-date and representative anthropometric data sets in his DHM-systems. It helps to guarantee aspects such as the realistic scaling of manikins and digital human models respectively for prospective workplace design and planning processes. The presented work serves a first step and an extended amount of parameters will be synthesized based on anthropometric data being currently collected by means of 3D-bodyscans in collaboration with the University of Greifswald in Germany. This reference dataset will contain measures listed within the international standard ISO 7250-2, the technical requirements of the 3D-bodyscanner correspond with the needs of the international standard ISO 20685.

## 5 Conclusions

This paper presents a method for creating virtual anthropometric datasets, based on a representative dataset for German civilians. Stature, mass and BMI were chosen for reference calculations. Based on the multivariate statistical correlations of the source data, the results show good comparability of the virtual population and the source dataset. The proposed approach combines the use of the concept of copulas and evolutionary algorithms for the



synthesis of an anthropometric dataset that is highly comparable to the representative dataset of the German working age population. The generated anthropometric data and its public accessibility can serve the DHM user as a practical and application orientated source for the implementation in work places and product design as well in process and work planning. Current research continues and addresses the validation of this method including an extended set of relevant anthropometric parameters.

## Acknowledgement

The authors would like to thank the Department of Epidemiology and Health Monitoring of the Robert Koch Institute for providing the source data for the presented research.

## References

- [1] Fromuth RC, Parkinson MB, 2008. Predicting 5th and 95th percentile anthropometric segment lengths from population stature. In: Proceedings of DETC08 ASME International Design Engineering Technology Conferences, New York City, USA.
- [2] International Organization for Standardization ISO 7250. Basic human body measurements for technological design.
- [3] International Organization for Standardization ISO 20685. 3-D scanning methodologies for internationally compatible anthropometric databases.
- [4] Nadadur, G, Raschke, U, and Parkinson, MB, 2016. A quantile-based anthropometry synthesis technique for global user populations. *International Journal of Industrial Ergonomics* 53:167-178.
- [5] Parkinson MB, Reed MP, 2009. Creating virtual user populations by analysis of anthropometric data. *International Journal of Industrial Ergonomics* 40, 106-111.
- [6] R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- [7] Robert Koch Institute, Department of Epidemiology and Health Monitoring, 2015. German Health Interview and Examination Survey for Adults (DEGS1). Public Use File 1. Version. doi: 10.7797/16-200812-1-1-1, <http://dx.doi.org/10.7797/16-200812-1-1-1>
- [8] Scheidt-Nave C, Kamtsiuris P, Gößwald A, Hölling H, Lange M, Busch MA, Dahm S, Döller R, Ellert U, Fuchs J, Hapke U, Heidemann C, Knopf H, Laussmann D, Mensink GBM, Neuhauser H, Richter A, Sass AC, Rosario AS, Stolzenberg H, Thamm M, Kurth BM, 2012. German health interview and examination survey for adults (DEGS) – design, objectives and implementation of the first data collection wave. *BMC Public Health* 12:730.
- [9] Wischniewski S, 2013. Delphi Survey: Digital Ergonomics 2025. In: Proceedings of the 2nd International Symposium on Digital Human Modeling (DHM). Ann Arbor, Michigan, USA.
- [10] Wischniewski S, Bonin D, Grötsch A, 2015. Virtual anthropometry – synthesis and visualisation of virtual anthropometric populations for product and manufacturing engineering. In: Proceedings of the 19th Triennial Congress of the IEA, Melbourne, Australia.

This paper was reviewed and accepted for presentation at the 4th International Digital Human Modeling Symposium (DHM2016) held on June 15-17, 2016 in Montréal, Québec, Canada by École de Technologie Supérieure (ÉTS) in association with the International Ergonomic Association (IEA) Technical Committee on Human Simulation and Virtual Environments (TC HSVE).