

# Worm plot: a simple diagnostic device for modelling growth reference curves

Stef van Buuren<sup>1,\*</sup>,† and Miranda Fredriks<sup>2,3</sup>

<sup>1</sup>*Department of Statistics, TNO Prevention and Health, Leiden, P.O. Box 2215, 2301 CE Leiden, The Netherlands*

<sup>2</sup>*Department of Paediatrics, Leiden University Medical Center (LUMC), P.O. Box 9600, 2300 RC Leiden, The Netherlands*

<sup>3</sup>*Child Health Division, TNO Prevention and Health, Leiden, The Netherlands*

## SUMMARY

The worm plot visualizes differences between two distributions, conditional on the values of a covariate. Though the worm plot is a general diagnostic tool for the analysis of residuals, this paper focuses on an application in constructing growth reference curves, where the covariate of interest is age. The *LMS* model of Cole and Green is used to construct reference curves in the Fourth Dutch Growth Study 1997. If the model fits, the measurements in the reference sample follow a standard normal distribution on all ages after a suitably chosen Box–Cox transformation. The coefficients of this transformation are modelled as smooth age-dependent parameter curves for the median, variation and skewness, respectively. The major modelling task is to choose the appropriate amount of smoothness of each parameter curve. The worm plot assesses the age-conditional normality of the transformed data under a variety of *LMS* models. The fit of each parameter curve is closely related to particular features in the worm plot, namely its offset, slope and curvature. Application of the worm plot to the Dutch growth data resulted in satisfactory reference curves for a variety of anthropometric measures. It was found that the *LMS* method generally models the age-conditional mean and skewness better than the age-related deviation and kurtosis. Copyright © 2001 John Wiley & Sons, Ltd.

## 1. INTRODUCTION

The last decade has witnessed an upsurge in methods for constructing age-related reference curves; see the comparison by Wright and Royston [1] for an overview. The major task in centile construction is to smooth the reference distribution in two directions simultaneously, between age and within age. Though this problem can be solved in a variety of ways, all approaches need to specify the amount of smoothness that provides a reasonable trade-off between parsimony of the curves and the fidelity to the data. Different choices lead to different reference values and to dissimilar appearances of the curves. It is therefore sensible to use

\*Correspondence to: Stef van Buuren, Department of Statistics, TNO Prevention and Health, Leiden, P.O. Box 2215, 2301 CE Leiden, The Netherlands

† E-mail: S.vanBuuren@pg.tno.nl

diagnostics that guide the choice of smoothness parameters in fitting a particular set of data. The present paper introduces a simple, general and flexible graphical method, termed worm plot, to support such modelling decisions.

The method can be used in situations where smoothing within age relies on a theoretical distribution. For example, the *LMS* method [2, 3] assumes that the reference distribution at a given age is normal after a Box–Cox transformation. Other possibilities include normalizing transformations like the shifted log-function [4], and other distributions like the Johnson distribution [5]. The worm plot can be applied to other transformations and distributions, but this paper focuses on the use of diagnostics in conjunction with the *LMS* method.

The text introduces the Fourth Dutch Growth Study, and then briefly reviews the *LMS* model and some diagnostic tools. Next, we introduce the worm plot, highlight its role in modelling reference curves, and apply it to the Fourth Dutch Growth Study. Finally, the text discusses some choices in the worm plot, and its relation to other methods for choosing smoothness parameters.

## 2. FOURTH DUTCH GROWTH STUDY 1997

The Fourth Dutch Growth Study [6, 7] is a cross-sectional study that measures growth and development of the Dutch population between ages 0 and 21 years. The study is a follow-up to earlier studies performed in 1955 [8], 1965 [9] and 1980 [10], and its primary goal is to update the 1980 references. Children with diagnosed growth disorders, those on medication known to interfere with growth, and those without a West European parent were excluded from the population definition. Like the previous studies, the sample was stratified by province, municipal size, sex and age. The planned sample size was equal to 16 188, and based on the objective to detect at least a 1.8 cm final height difference between the 1997 and 1980 studies with a power of 99 per cent. Age groups were chosen as follows: six age groups in year 1; four in year 2; two in year 3; three over the period 3–8; and two per year between ages 9 and 20. The age group interval of girls older than 17 years was a year instead of a half year. The realized sample size was  $n = 14\,500$ . Table I contains the composition of the sample.

The study measured, among other variables, height, weight and head circumference. Until the age of two, length of infants was measured to the nearest 0.1 cm in the supine position. A ‘microtoise’ was used to measure height of children older than 2 years. Children younger than 4 years were measured by 24 ‘Well Baby Clinics’ during the regular periodical health examination. Children between the ages of 4 and 9 were measured by 25 ‘Municipal Health Services’ during regular health assessments. Older children received a personal invitation based on a stratified sample from the Municipal Register Office. Non-response (children who refused to show up at the health clinic or refused a measurement) varied between 20 per cent in ages of 11 to more than 60 per cent in those over age 17. Of a random sample of non-responders ( $n = 230$ ), 170 returned a questionnaire. No significant differences from the study sample were found.

In order to obtain a sufficiently large sample, additional measurements were done at high schools, universities, a youth festival, and during medical examinations for joining the army. No statistically significant differences in height were found between the original and additional sample. The distributions of the combined sample for age, sex, municipal size, family size and

Table I. Sample size of the Fourth Dutch Growth Study 1997 (Dutch children) by age and sex.

Age	Boys	Girls
0-1	1219	1219
1-2	807	797
2-3	464	454
3-4	295	314
4-5	84	84
5-6	134	137
6-7	66	63
7-8	142	140
8-9	110	108
9-10	334	320
10-11	350	366
11-12	367	364
12-13	381	395
13-14	432	470
14-15	414	392
15-16	407	400
16-17	355	240
17-18	350	183
18-19	333	217
19-20	271	172
20-21	153	171
21-22	14	12
Total	7482	7018

child education were similar to national figures obtained from Statistics Netherlands [11]. The only exception was geographical region for girls over age 18. A weighted analysis, beyond the scope of this paper, was performed to correct the height references for this effect.

### 3. *LMS* MODEL

The *LMS* method [2, 3] describes a variable  $y$  as a semi-parametric regression function of a time-dependent variable  $t$ , so that the distribution of  $y$  changes gradually when plotted against  $t$ . The distribution is summarized by three time-varying natural spline curves: the Box-Cox power that converts  $y$  to normality ( $L$ ); the median ( $M$ ); the coefficient of variation ( $S$ ). Let  $L(t)$ ,  $M(t)$  and  $S(t)$  stand for the value of the  $L$ ,  $M$  and  $S$  curves at age  $t$  for a given *LMS* model. The standard deviation score  $z$  of a particular measurement  $y$  at age  $t$  can be computed as

$$z = ((y/M(t))^{L(t)} - 1)/L(t)S(t), \quad \text{if } L(t) \neq 0$$

or

$$z = \log(y/M(t))/S(t), \quad \text{if } L(t) = 0$$

The number of effective degrees of freedom (EDF) [12] is a convenient parameter that expresses the amount of adjustment necessary for smoothing a set of data. In the *LMS* model, the smoothness of the *L*, *M* and *S* curve is characterized by three scalar EDF-parameters:  $E_L$ ,  $E_M$  and  $E_S$ . An EDF of zero constrains the entire curve to a given value, and an EDF of 1 corresponds to a constant value whose location is to be estimated from the data. An EDF of 2 yields a straight line, while larger EDFs allow for increasingly more flexibility in the fitted curves. Note that this definition of EDFs may differ from that of other authors, as EDFs are strictly not defined for values less than 2. Also, EDFs are sometimes presented reduced by one, so that the EDF for a straight line is 1 rather than 2. Cole and Green [3] argued that the distributions of  $E_L$ ,  $E_M$  and  $E_S$  in the *LMS* model are largely independent of each other, implying that one EDF can be optimized while fixing the other two.

In the following, we will use an abbreviated notation for *LMS* models as 'LMMSX', where 'L' stands for  $E_L$ , 'MM' for  $E_M$ , 'S' for  $E_S$  and where 'X' is a transformation option ('space' if none, 'R' if rescaled, 'P' if power transformation). Thus, model 4096R has  $E_L=4$ ,  $E_M=9$ ,  $E_S=6$ , with the rescale option set. Transformation options are a specific feature of the *LMS* program that will transform the time axis. The rescale transformation ('R') fits the *LMS* model on a rescaled time axis that stretches periods of rapid growth (for example, infancy and puberty), and that compresses periods with lower growth velocities (mid-childhood or adulthood). In effect, the distribution is allowed to change more rapidly at locations where the *M* curve is steep. After fitting in transformed time, the results are scaled back to the original time scale. The power transform ('P') option allows the user to specify the two parameters (offset and power) to rescale the time axis. Cole *et al.* [13] describe these options in more detail. The judicious use of the options may substantially improve the fit.

### 3.1. Diagnostics

For a given choice of  $E_L$ ,  $E_M$  and  $E_S$ , the *LMS* program maximizes the penalized likelihood. Several types of diagnostic skills and tools are helpful for inspecting the quality of the solution:

- (i) *Visual inspection of the shape of the reference curves.* Experienced researchers may recognize the appropriateness of a given set of reference curves based on subtle features in the shape, like a 'pubertal belly' in cross-sectional data. In general, substantial exposure to reference curves is needed to develop the necessary skills.
- (ii) *Centiles plotted onto the individual data points.* This type of plot is useful for inspecting outliers and for detecting gaps in the data and gross errors in the model, but its resolution is too limited to be helpful in choosing among different models. The measurements can be visualized in both the original and in the SD scale, but the latter is often clearer.
- (iii) *Empirical and fitted centiles plotted on top of each other.* This is an old and quite accurate technique in which the observations are divided into age groups. Empirical centiles are computed for each group, and these are plotted together with the fitted curves. If everything is right, the fitted curves should be close to the point estimates (that is, within sampling error). Various choices are possible for the vertical scale (raw, standardized for mean and/or standard deviation). A disadvantage of the raw data plot is that if the standard deviation changes with age, the same distance means different things

at different ages. Van Wieringen [14] pioneered a standardized graph under the heading of 'graphical graduation'. Care is needed in computing extreme percentiles, as some interpolation is needed. The algorithms implemented in SPSS and SAS can give odd results (mostly estimates that are too wide, irrespective of the interpolation algorithm), and we prefer the S-plus function *quantile()* for this purpose. As the display does not contain individual observations, it may be insensitive to subtle deviations between the fitted and empirical distributions, especially if the number of centiles is small.

- (iv) *Observed and expected counts.* Healy *et al.* [15] suggested comparing the observed and expected frequencies of observations within defined centile and age groups. This can only be done if one assumes a distribution of the measurements for each age group. One must choose cutpoints for centile and age groups, thus leading to somewhat arbitrary comparisons. The Kolmogorov–Smirnov test and the Q–Q plot, both described below, evade the choice of centile cutpoints, and thus compare the entire observed and expected age-related distribution.
- (v) *Statistical tests.* If the distribution of the measurements is known, a statistical test can be used to test the fit of the solution. In the *LMS* model,  $z$  should be distributed as  $N(0, 1)$  at all ages. Normality at different age groups can be checked by means of, for example, the Shapiro–Wilk  $W$  test. This test is sensitive in picking up any skewness, but is less powerful in detecting kurtosis [16]. For other distributions, one could apply a Kolmogorov–Smirnov test. A disadvantage of tests in general is that they do not tell *how* the empirical and theoretical distributions differ. Techniques based on statistical significance may overfit the curves in large samples. Purists might say that the application of inferential tests for modelling does not comply with the orthodox Neyman–Pearson criteria (since the same data are repeatedly used), and the interpretation of non-significant tests as evidence for the model is not without problems [17].
- (vi) *Quantile–quantile plot (Q–Q plot) of the z-scores.* Q–Q plots [18] can be applied if the measurements are supposed to follow a known distribution. The display plots the quantiles of the theoretical distribution (on the horizontal axis) against those of the empirical distribution (on the vertical axis). The Q–Q plot for normal data, also known as the normal probability plot, is best known, but it can be adapted to other distributions. The plot yields insight into structural characteristics (for example, skewness, kurtosis) of empirical deviations from the assumed distribution. In its detrended form, the Q–Q plot is very sensitive to subtle deviations [19]. Detrended means that each empirical quantile is subtracted from its corresponding unit normal quantile. As will be demonstrated below, the use of the Q–Q plot as a global diagnostic is limited though.
- (vii) *Worm plot.* The worm plot consists of a collection of detrended Q–Q plots, each of which applies to one of successive age groups. The vertical axis of the worm plot portrays, for each observation, the difference between its location in the theoretical and empirical distributions. The data points in each plot form a worm-like string. The shape of the worm indicated how the data differ from the assumed underlying distribution, and when taken together, suggests useful modifications to the model. A flat worm indicates that the data follow the assumed distribution in that age group.

Note that the application of the latter four approaches require distributional assumptions, whereas the first three do not. In practice, one will typically apply a combination of diagnostics as in, for example, the recent paper by Royston and Wright [20].

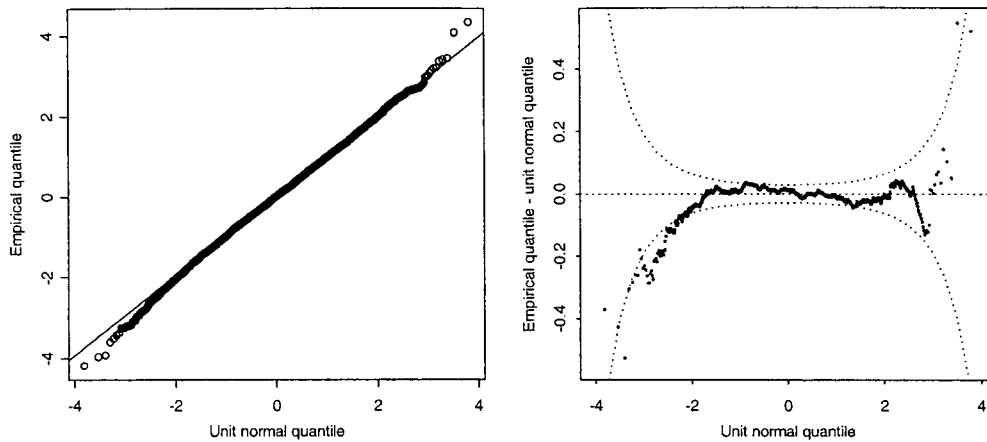


Figure 1. Conventional (left) and detrended Q–Q plot (right) of the z-scores of height of Dutch boys (all ages combined, *LMS* model 0051R). The detrended Q–Q plot improves upon the resolution of the conventional display and includes 95 per cent confidence intervals. Both figures convey the misleading message that model 0051R fits reasonably well.

### 3.2. Illustration

Figure 1 contains the conventional Q–Q plot and the detrended Q–Q plot for a normal distribution of z scores of over 7000 boys in the fourth Dutch Growth Study 1997. All ages are combined here. The detrended plot on the right contains the 95 per cent confidence interval of the unit normal quantiles. For a given quantile  $z$  with associated probability  $p$  and a sample size  $n$ , the 95 per cent confidence interval is computed as  $\pm 1.96 \times f(z)^{-1} \sqrt{(p(1-p)/n)}$ , where  $f(z)$  is the normal density function [21]. Owing to scarce data, the interval becomes broader towards the extremes, so in the tails larger differences between theoretical and empirical quantiles are tolerated. Except for the area below  $-2$  SD, the empirical quantile points are all located near the main diagonal. The marginal z scores, that is the z scores of all age groups combined, thus closely follows a normal distribution. However, this apparent fit does not imply that the model even remotely fits the data.

Figure 2 is a worm plot of the same data. The data are split into 16 age groups of equal size, and the relevant computations are done in each group separately. The exact age boundaries are given in each panel of the plot. Figures 1 and 2 provide dramatically dissimilar views on the same data. In fact, Figure 2 shows that the model fits badly at almost all ages. The only reasonable fit occurs in age group 9.1–10.4. For other ages, the worms move around in all directions, indicating the existence of gross errors in the fit of the statistical model. The modelling problem is now to ‘tame the worms’, so that each of them becomes as flat as possible and aligns up neatly along the horizontal axis. The 95 per cent confidence interval gives an impression of the sampling variation, and delineates the region where the worm should be located most of the time if the empirical and theoretical distributions agree. The shape of the worms communicates the type of misfit between model and data. Table II summarizes several aspects of the distribution. Each shape describes a different aspect of the model fit.

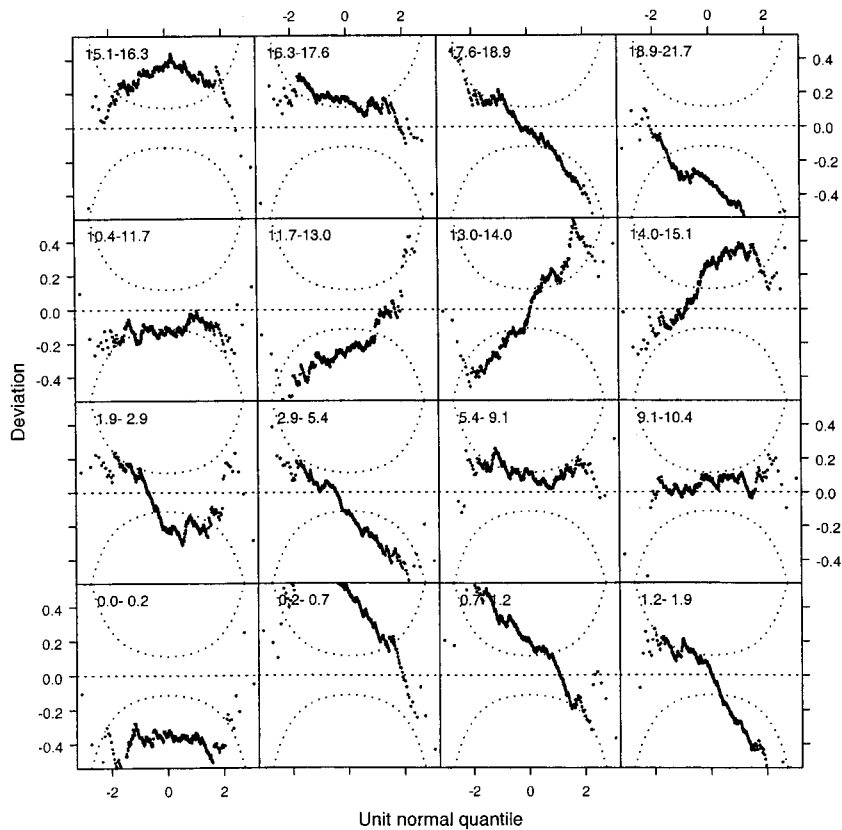


Figure 2. Worm plot of the z-scores for height of Dutch boys (*LMS* model 0051R, same as Figure 1). The plot consists of detrended Q-Q plots in 16 age groups of equal size, ordered from the lower-left panel to the upper-right panel. Model 0051R fits badly in almost all ages.

Table II. Interpretation of various patterns in the worm plot.

Shape	Moment	If the	Then the
Intercept	Mean	worm passes above the origin, worm passes below the origin,	fitted mean is too small. fitted mean is too large.
Slope	Variance	worm has a positive slope, worm has a negative slope,	fitted variance is too small. fitted variance is too large.
Parabola	Skewness	worm has a U-shape, worm has an inverted U-shape,	fitted distribution is too skew to the left. fitted distribution is too skew to the right.
S-curve	Kurtosis	worm has an S-shape on the left bent down, worm has an S-shape on the left bent up,	tails of the fitted distribution are too light. tails of the fitted distribution are too heavy.

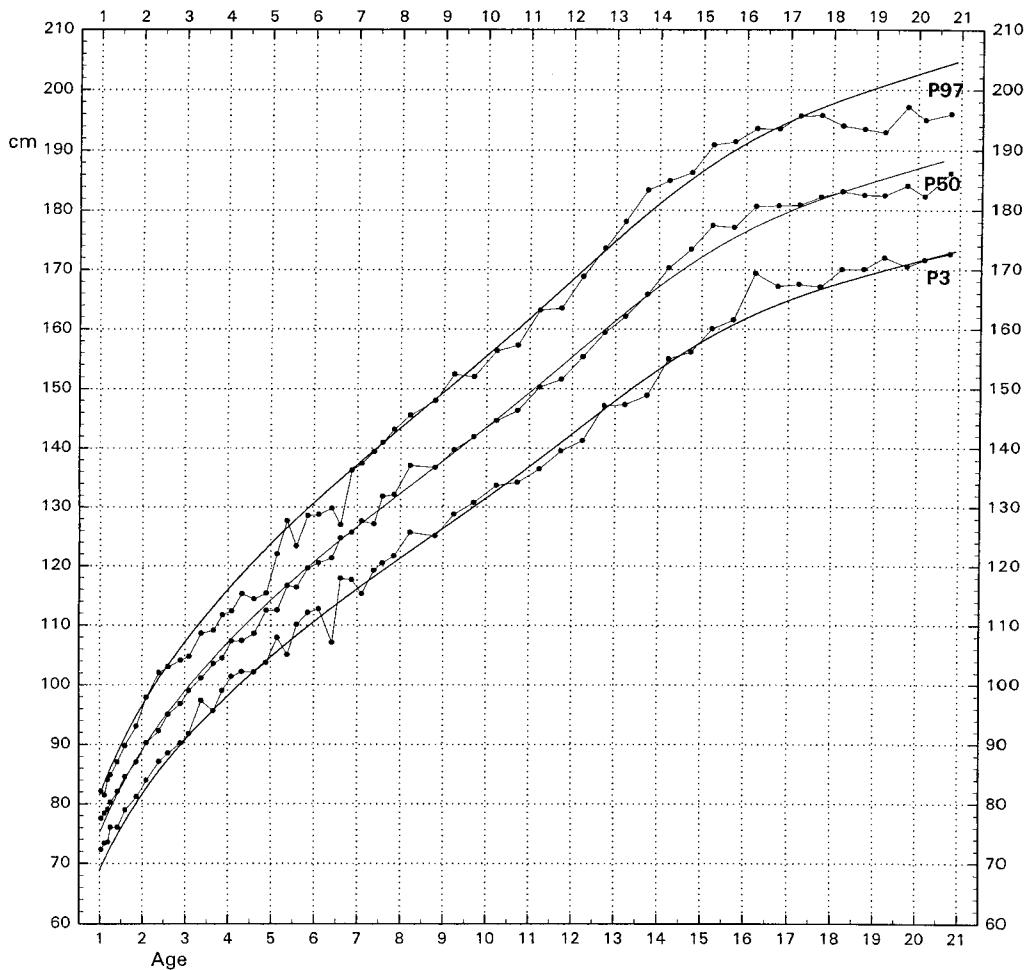


Figure 3. Raw and fitted percentiles for height of Dutch boys (all ages combined, *LMS* model 0051R).

#### 4. MODELLING STRATEGY

The number of EDFs in Figures 1 and 2 is equal to  $E_L=0$ ,  $E_M=5$  and  $E_S=1$  for the  $L$ ,  $M$  and  $S$  curves, respectively, compactly written as 0051R. This model corresponds to a normal distribution of constant variation and a moderate spline for the  $M$  curve. Figure 2 suggests that the model is too inflexible for the data. This is confirmed in Figure 3, which draws the raw and fitted percentiles on the same diagram. The P50 does not follow the raw median and misses the bend at about 15–16 years. Until the age of 6 and above the age of 17, the P3 and P97 reference curves are too wide, while they are too narrow during puberty.

Cole provides guidelines on obtaining optimal values for  $E_L$ ,  $E_M$  and  $E_S$  (see the documentation of the *LMS* FORTRAN program in <ftp://ftp.statlib.edu/lms>). Starting with values



$4 \leq E_M \leq 6$  and  $E_L = E_S = 1$ , his strategy is to optimize  $E_M$  first by increasing  $E_M$  progressively by 1 until the change in the penalized likelihood becomes small. Let this change be denoted as the *D*-statistic, defined as  $D(v, w) = 2(l(w) - l(v))$ , where  $v$  is the more restrictive model nested inside model  $w$ , and where  $l(v)$  and  $l(w)$  are the corresponding penalized log-likelihood values. It is common to assume that  $D$  has an asymptotic  $\chi^2$ -distribution with  $d$  degrees of freedom, where  $d$  is the number of additional free parameters in the less restricted model [12]. A typical cut-off point of  $D$  is 2, but the precise choice depends also on sample size, where larger samples need larger cut-off points. The final decision on the  $E_M$ -parameter will depend on the appearance of the  $M$  curve. The process is repeated for the  $S$  and  $L$  curves, fixing the previous optimal values of  $E_M$  and  $E_S$ . Cole suggested skipping the model with  $E_L = 2$  and  $E_S = 2$  in order to evade 'silly values at the extremes'. In addition to the  $M$  step, one could experiment with alternative transformations of the time axis, which may help to reduce the complexity of the  $M$  curve.

This conditional optimization approach is simple and relatively easy to perform. The sequence of steps (first  $M$ , then  $S$ , and then  $L$ ) is sensible because the  $M$  curve describes the most important variation, while the influence of  $L$  is relatively small. Subjective elements in the procedure include the choice of the cut-off point and the visual assessment of the  $M$  curve. With sample sizes around 7000, we frequently found that a change of say 5 or 10 units did not appear to have any influence on the shape of the curves. It sometimes happened that increasing  $E_M$  introduced spurious wiggles.

It is often difficult to see what actually happens to the fitted curves when an EDF changes. Also, it is hard to assess how well the curves actually fit the data. The worm plot can be used in a visual analogue to the conditional optimization strategy, and remedies these two deficits. This is done by the following steps:

1. increase  $E_M$  such that each worm passes through the origin of the plot;
2. then, increase  $E_S$  such that each worm has more or less a zero slope;
3. then, increase  $E_L$  such that quadratic shapes (U-shapes) disappear.

By aligning worm plots of two models side by side, it is easy to see at what points the *LMS* optimization changes the curves. This requires the same cognitive skills as needed for the children's game 'find the 10 differences between two pictures'. The worm plot gives a visual impression of the fit between data and model at different ages.

## 5. MODELLING THE FOURTH DUTCH GROWTH STUDY 1997

Figure 4 is the worm plot for model 0101R for the Dutch male height data, thus where  $E_M$  has increased from 5 to 10. The vertical distance between worm and origin is now small everywhere, which indicates that the  $M$  curve does a reasonable job in modelling mean height. Increasing  $E_M$  to 11 did not appreciably improve upon the display, so  $E_M = 10$  was considered an appropriate choice.

Figure 5 is a similar display for models with  $E_S = 6$  instead of  $E_S = 1$ . The worms are flattened in comparison to Figure 4, signalling that the new  $S$  curve is a better description of the differences in height variation across ages. Differences between models 0106R and 0107R were deemed insignificant, so  $E_S = 6$  was taken as the final choice.

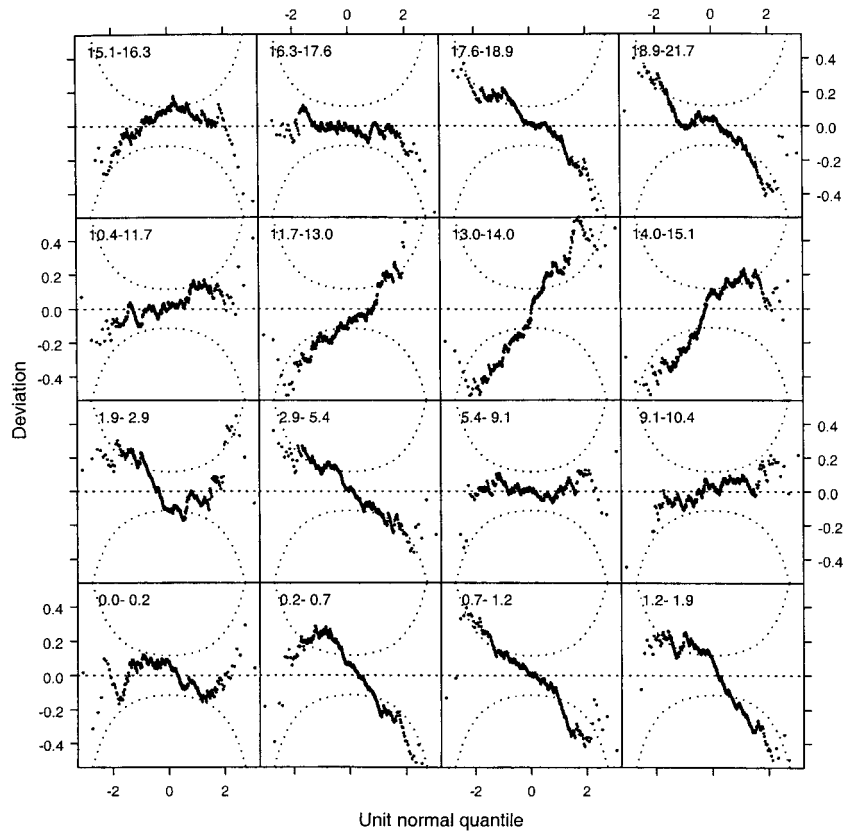


Figure 4. Worm plot for height of Dutch boys (*LMS* model 0101R). Worms are close to the origin, indicating a reasonable fit of the *M* curve.

The second [9] and third [10] Dutch growth studies found a skewed conditional height distribution during the first years of puberty. Does this also hold in the present data? Several values for  $E_L$  (0,1,3,4) were tried, thus correcting for age-related skewness. The overall impression is that the effect of increasing  $E_L$  is quite small. Figure 6 displays the worm plot corresponding to model 4106R. This model fits slightly better than the normal model with  $E_L = 0$ . If there were a difference of only one degree of freedom between both models, we would prefer the more complex model over the simpler normal model. However, here the models differ by four degrees of freedom. The worm plots for models with  $E_L = 1$  and  $E_L = 3$  lie in between those in Figures 5 and 6, and the transitions are small. Figure 7 is a diagram of the raw and fitted percentiles for models 0106R and 4106R. It appears that the effect of increasing  $E_L$  is very small. Note that there is a rising linear trend in height after age 19. This reflects the strength of the spline approach compared to a parametric curve (for example, the Jolicoeur model) which assumes that height reaches an asymptote in adulthood. Model 0106R was used to update the Dutch growth references, and the official height reference values [6] are based on this model.

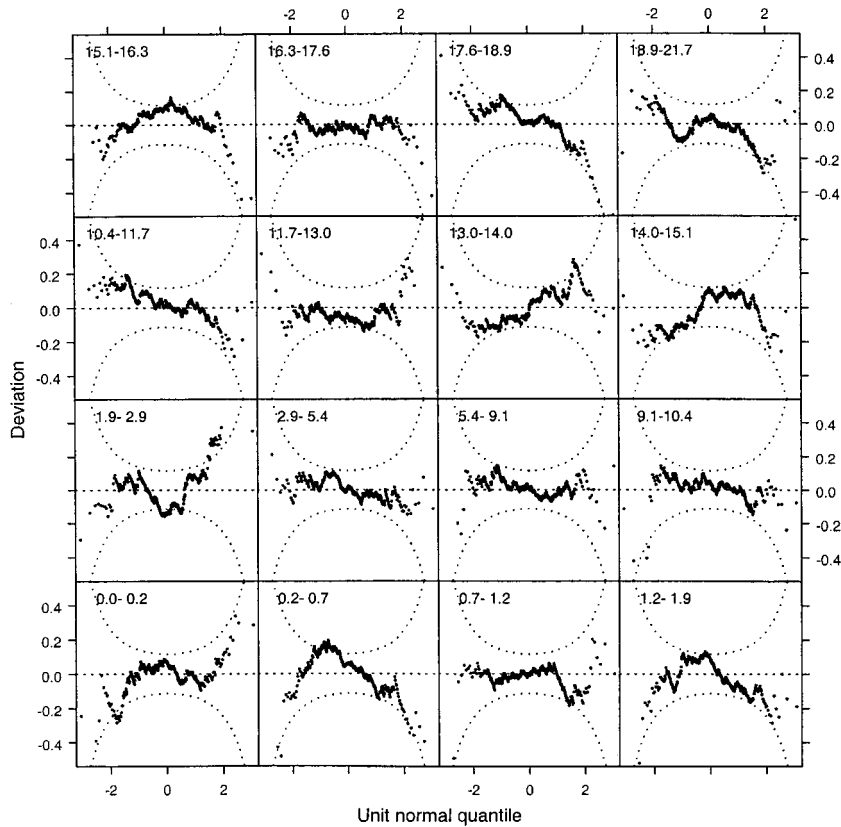


Figure 5. Worm plot for height of Dutch boys (*LMS* model 0106R). Worms are relatively free of linear trend, indicating a reasonable fit of the *S* curve.

## 6. PROPERTIES

### 6.1. Shape estimates

It is possible to quantify the basic features of the worm shape. Polynomial regression of the empirical on the theoretical quantiles gives numerical estimates of various aspects of the discrepancy between the observed and theoretical distributions. Suppose that the results are scaled according to an equation of the form  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$ , where  $Y$  denotes the vector of detrended ordered observations and where  $X$  denotes the vector of quantiles. The idea is that the shape coefficient  $\beta_0$  measures the amount of misfit of the *M* curve,  $\beta_1$  measures the amount of misfit of the *S* curve, and so on. The correspondence between the  $\beta$ 's and the moments of the empirical distribution relates to the inverse Cornish–Fisher expansion [22, 23]. Some statistical literature suggests that  $\beta_0$  is equal to the difference of the theoretical and empirical means, and that  $\beta_1$  measures differences in variation [24, 25]. The inverse Cornish–Fisher expansion suggests that  $\beta_2$  and  $\beta_3$  measure differences in skewness and kurtosis, respectively, but the precise relation between these coefficients and more common

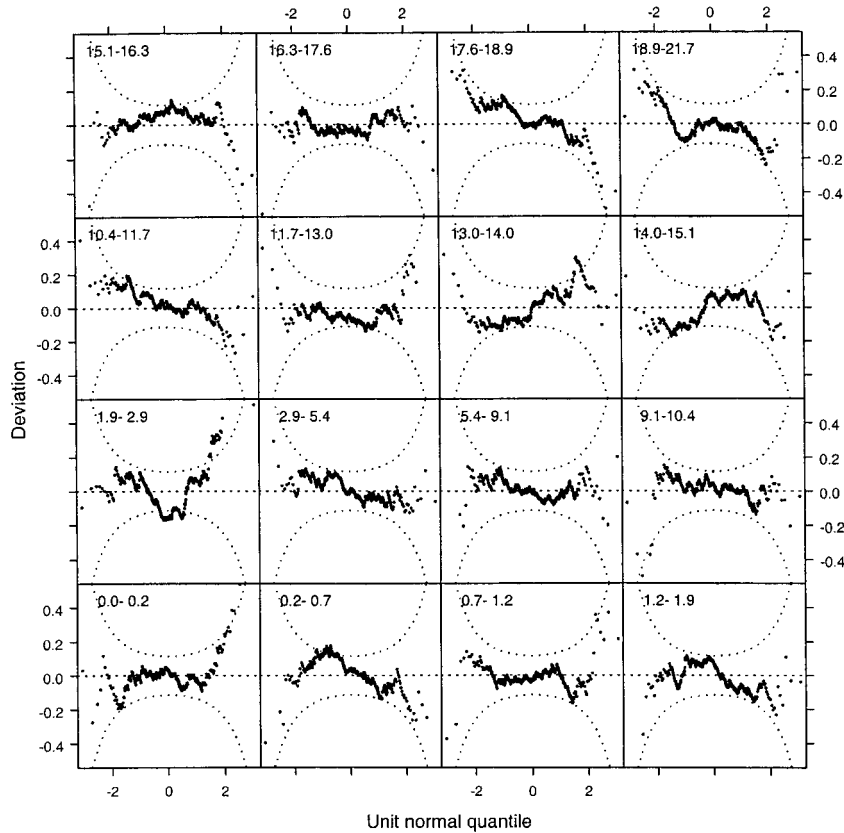


Figure 6. Worm plot for height of Dutch boys (*LMS* model 4106R). Adding skewness parameters hardly improves the fit.

measures of skewness and kurtosis is not yet clear. The HRY method [15] implicitly relies on the properties, as it describes the form of the age-conditional distribution as polynomial functions of the unit normal quantiles, as above. Healy *et al.* note that increasing the degree of these polynomials from 1 to 2 allows for skewness, and increasing it to 3 allows for skewness and kurtosis.

Shape coefficients can be used for quantitative assessments of model fit. Coefficients of the same type (for example, all  $\beta_0$ ) can be compared across models to see the effect of model alterations. Shape coefficients  $\beta_0$  and  $\beta_1$  are approximately on the same scale and can be compared with each other. Coefficients  $\beta_2$  (quadratic) and  $\beta_3$  (cubic) are on smaller scales. To give some idea of their interpretation, we categorize solutions where the absolute values of  $\beta_0$  or  $\beta_1$  are in excess of 0.10 as misfits. For  $\beta_2$  we use a threshold of 0.05, and for  $\beta_3$  we take 0.03.

### 6.2. Choices in the worm plot

Some details of the worm plot, like the number of age groups and the choice of scales, have not yet been discussed. When restricted to a square layout, convenient choices include the

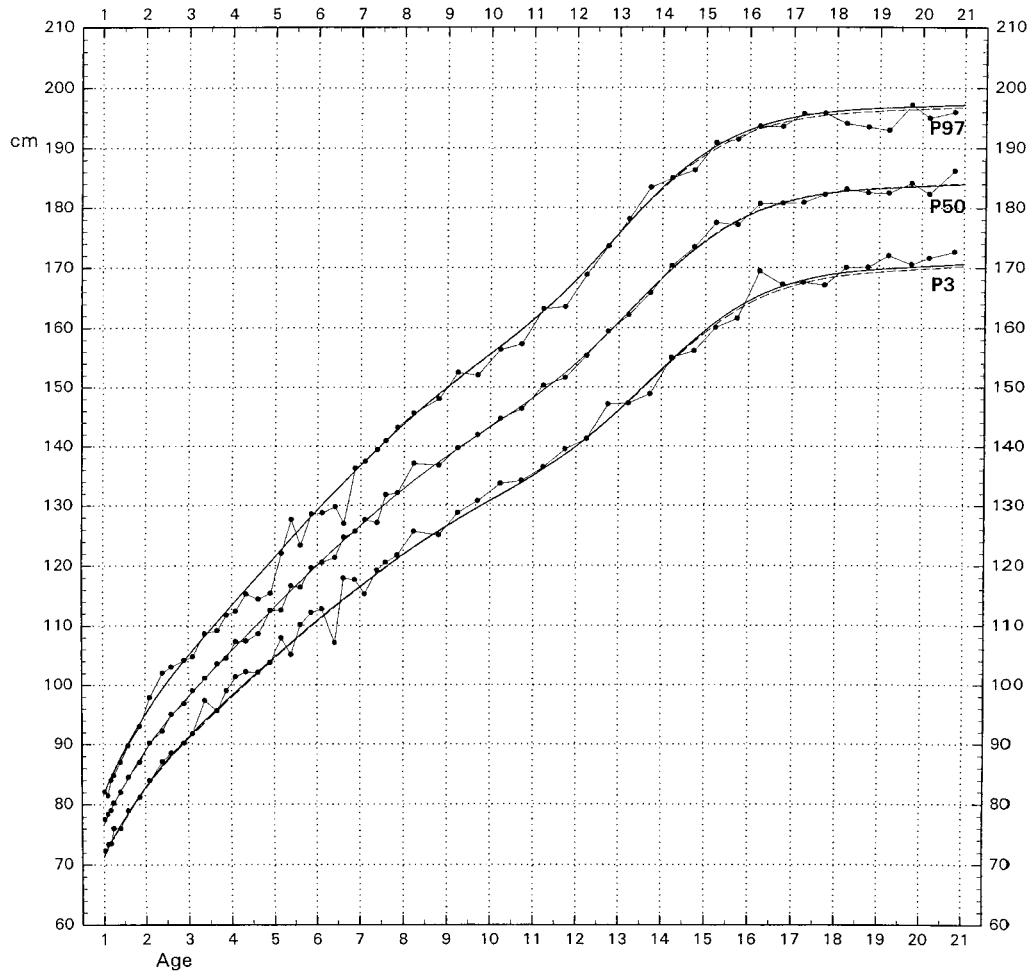


Figure 7. Raw and fitted percentiles for height of Dutch boys (*LMS* models 0106R and 4106R). Model 0106R is the official 1997 height reference update.

$3 \times 3$ , the  $4 \times 4$  and the  $5 \times 5$  plotting grid, thus defining 9, 16 and 25 age groups, respectively. In general, increasing the number of age groups provides a more detailed, but less stable plot. As a rough guideline, at least 200–300 points per group are needed for a reasonably stable picture. In our experience, using nine age groups might obscure important deviations from normality, like those concerning the whole age range as in Figure 1. On the other hand, the display becomes somewhat crowded if 25 or more groups are formed, especially if side-by-side comparisons are being made. In addition, the number of points can become quite low. The numbers of 16 groups seems to be a good compromise, but it is also useful to experiment with other resolutions.

The scale of the  $y$ -axis was chosen as the range  $-0.5$  to  $+0.5$  SD for all panels. This range is a compromise between an adequate display of the volatility of the worms and the

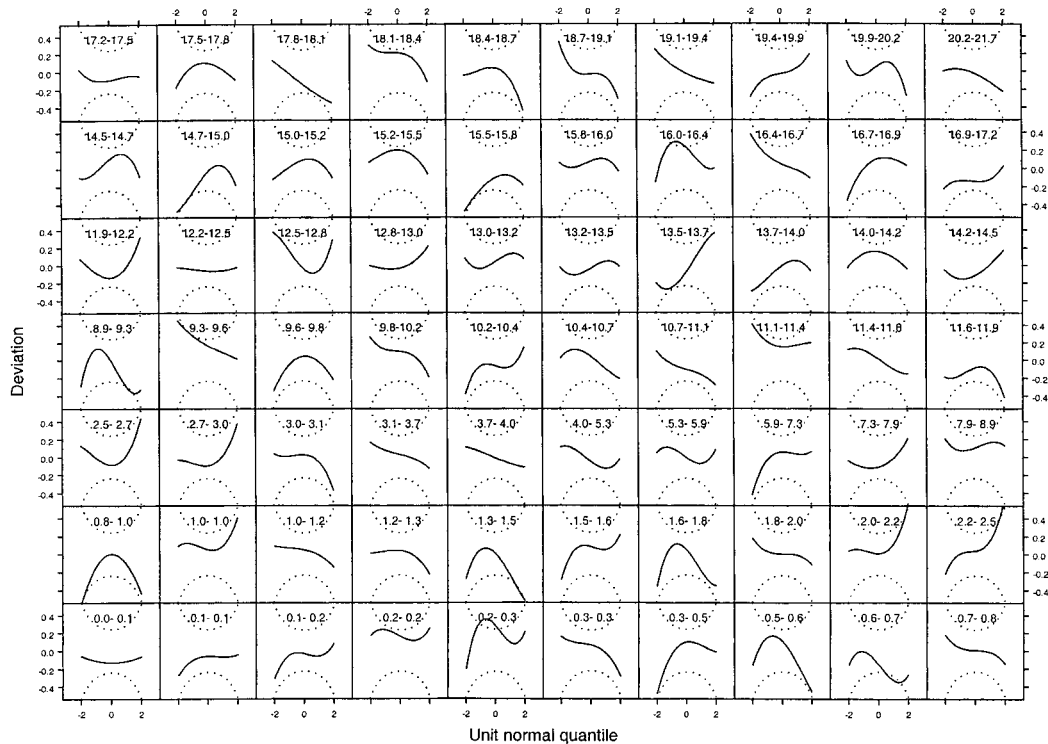


Figure 8. Cubine plot with 70 panels for height of Dutch boys (*LMS* model 0106R).

objective of minimizing the number of points outside the active plotting area. The setting may occasionally produce entire empty panels if the model fits very badly.

It is sensible to choose the cut-off points on age such that each group approximately has the same expected increment in the measurement. It is, however, erroneous to directly cluster heights into groups, since that would inadvertently destroy the age-conditional normal distribution. We therefore divided the observations into groups of equal size. Since groups at higher velocities are sampled more often, this procedure approximates the objective of 'same expected increments'. In this way, the mean height increment per age group varies between 7 cm (ages 0.0–0.2 years) to 22 cm (ages 5.4–9.1 years) to 2 cm (above 16 years). We experimented with non-overlapping and overlapping age groups, where the observation appears in two adjacent panels, and found that either possibility led to similar model choices. An advantage of overlap is that the display is relatively insensitive to the exact location of the boundary points. A disadvantage is that it cannot be handled so easily with summary statistics like those that were recently proposed [26]. All plots in this paper were made with non-overlapping groups, so every observation appears just once.

Figure 8 is an example of a *cubine plot*, a stylized version of the worm plot. 'Cubine' stands for 'cubic line', that is, the line predicted by the four-parameter polynomial model of Section 6.1. The interpretation of the cubine is identical to that of the worm. The 95 per cent confidence interval of the cubine is also plotted. One could check whether the cubine is

Table III. Optimal *LMS* models for various types of reference diagrams in the Fourth Dutch Growth Study, and the total of misfits (out of 16) for each polynomial shape.

Measure	Ages	Boys				Girls					
		Model	Misfits				Model	Misfits			
			$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$		$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$
Height for age	0–21	0106R	0	1	0	0	0105R	0	1	0	0
Weight for age	0–21	4085R	0	5	0	2	4086R	0	4	0	2
Weight for height	0–16	3074R	0	5	2	3	3074R	0	1	0	0
Head circumference for age	0–21	0095R	0	5	4	7	0074R	1	6	2	7
Body mass index for age	0–21	5135P	0	3	0	2	5116P	0	3	0	0

Note: Using a power transformation for body mass index with offset 0 and powers 0.33 (boys) and 0.25 (girls).

located within the interval. If it is, this suggests that differences between the empirical and theoretical distributions for that age group is due to random variation.

The cubine plot is useful to assess finer details in the case where the number of panels is large, that is, for smaller age groups. For example, we were concerned that the rather large height increment in the worm plot in ages 5–9 years would obscure important deviations. The corresponding cubines for the ages in Figure 8 are quite regular, indicating that the model fits well here and that the fit is independent of the age grouping. If certain shapes repeat in successive panels, for example, three consecutive U-shaped cubines, such repetitions could be used to detect detailed misfits. Cubines could be used instead of worms, but the view on the raw data will be lost. Also, they will not display shapes more complex than the cubic.

### 6.3. Relation with the *D*-statistic

The application of the worm plot may lead to model choices that differ from those obtained by the difference between penalized likelihood values. For height, the worm plot usually suggests larger cut-off points in terms of likelihood differences, thus resulting in models with fewer parameters. For example, the likelihood difference between models 0101R and 0111R is equal to  $D(0101R, 0111R) = 27.9$ , while  $D(0106R, 0107R) = 12.0$ . In both cases the more complex model is a statistically significant improvement over the simpler one, yet the worm plots do not indicate noticeable differences between the solutions. Employing a cut-off point of  $D \leq 10$  would result in model 1137R, and a cut-off of  $D \leq 5$  would produce model 1169R. The latter model clearly overfitted the data and produced wiggly curves.

One might be inclined to think that the worm plot is less sensitive than the *D*-statistic, thus leading to overly simplistic models. Though both the worm plot and the *D*-statistic assess similar aspects of the model fit, this conjecture is inaccurate. For example, in modelling the *S* curves for head circumference, the worm plot indicates appreciable changes in the conditional circumference distribution, while the *D*-statistic is quite small (for example,  $D(0093R, 0094R) = 2.8$ ,  $D(0094R, 0095R) = 9.0$ ). Thus, the worm plot is not simply a coarse version of the *D*-statistic, but provides a different and more informative view on the data.

Table III gives an overview of the final models fitted for the Fourth Dutch Growth Study, as well as the total of misfits for each basic shape, as defined in Section 6.1. Note that the fit of the mean curve is quite good in almost all cases. An exception is model 0074R for head

circumference of the girls. To a lesser extent, this also holds for the quadratic shape. Height curves appear to be relatively easy to fit with the *LMS* model. The worm plot of weight for age contains clear S-shapes for ages 14–17 years, indicating that the fitted tails might be too thin in this age range. The story of weight for height is somewhat mixed as the curves for the girls fit substantially better than those for the boys. Head circumference appears to have thicker tails than the normal distribution during the first year of life. The fit of the reference values for body mass index (BMI) is quite good.

## 7. CONCLUSION

The worm plot is a diagnostic tool to describe salient features of the age-conditional  $z$ -score distribution. It aids in finding proper smoothing values for  $E_M$ ,  $E_S$ , and  $E_L$  of the *LMS* method. There is a close correspondence between these smoothing parameters and particular shapes of the worms. These basic shapes can be estimated numerically by polynomial regression. The worm plot assesses whether a particular *LMS* model leaves any important unexplained structure in the residuals. The *LMS* model generally adequately describes the median and the skewness of the data, but has more difficulty in modelling deviation. The *LMS* model assumes that there is not any kurtosis.

The worm plot can be used in conjunction with other methods than *LMS*. In fact, the normal worm plot can assess the fit of any model based on conditional normality, including a large variety of linear and non-linear regression models. The tool seems especially useful in cases where inspection of marginal normality is misleading, as in Figure 1. Using the worm plot in conjunction with the *LMS* model is particularly instructive since different parameters influence different aspects of the worm. Other growth models that will probably work quite well include the HRY model [13], the fractional polynomial model [20] and the quantile model [27]. Chambers *et al.* [21] give general formulae for estimating the confidence intervals, so the worm plot can also be applied to distributions other than the normal. Repetition of shapes in the cubine plot might be investigated in a formal way by computing and testing the autocorrelation of the shape coefficients across age groups.

The summary of the shape estimates in Table III may act as a rough guideline for users that fit *LMS* models to other data. One should keep in mind that the results are based on the analysis of one data set. In future studies, it could be useful to quantify and study the variation in optimal EDFs derived from other populations.

Cole [2] remarked that producing centile charts has always been something of a black art. His *LMS* method combined ideas of the method of Roede and Van Wieringen [10] and Van 't Hof *et al.* [28], and paved the way for modern methods that give reproducible results. However, the inner workings of modern centile fitting methods are not so obvious: black art was replaced by a black box. We think that our worms can contribute in opening up this black box, and hope that they provide fertile soil for further development.

## APPENDIX: S-PLUS 4.5 FUNCTION FOR DRAWING THE WORM PLOT

```
# S-plus 4.5 functions for plotting the worm plot.
# Author: S. van Buuren, TNO Prevention and Health (1999).
```



```

read.lms <- function(filename)
{
# function to read the z-scores from the .lms output file generated
# by the LMS software of T.J. Cole (version June 1998)
  lms.par <- scan(filename, n = 8)
  print(lms.par)
  lms.skip <- lms.par[8] + 2
  read.table(filename, skip = lms.skip, col.names = c("age", "val", "z"))
}
wp <- function(data, layout = c(4,4), overlap = 0, worms = T, cubines = F,
coefsave = F, labels = T, hor = T, vert = F, ci = T, sub = paste (deparse(
  substitute(data )), deparse(substitute(overlap))))
# function for plotting the worm plot on the active graphics device
{
  panel <- function(x, y)
  {
    qq <- as.data.frame(qqnorm(y, plot = F))
    qq$y <- qq$y - qq$x
    plot (qq$x, qq$y, type = "n", ylim = c(-0.5, 0.5), xlim = c(-3, 3),
lab = c(3, 5, 7), tck = -0.01)
    if (hor) abline(0, 0, lty = 2, col = 1)
    if (vert) abline(0, 100000, lty = 2, col = 1)
    if(worms)
      points(qq$x, qq$y, col = 1, pch = 1, mkh = 0, cex = 0.25)
    if(cubines | coefsave)
      fit <- lm(y ~ x + x^2 + x^3, data = qq)
    if(cubines) {
      s <- spline(qq$x, fitted(fit))
      flags <- s$x > -2 & s$x < 2
      lines(list(x = s$x[flags], y = s$y[flags]))
    }
    if(coefsave) {
      est <- coef(summary(fit))[, 3]
      assign(".est", c(.est, est), frame = 0)
    }
    if (ci) ciplot(sum(!is.na(qq$y)))
  }
}
agetext <- function (classes, layout = c(4, 4), cex=0.6, dx = 0.06, dy=0.02)
# function for adding age group text to the worm plot panels
{
  txt <- apply(format(round(summary(classes)$intervals,1)),
1,paste,collapse = "-")
  x <- rep((0:(layout[1]-1))/layout[1]+dx,layout[2])
  y <- rep((1:(layout[2]))/layout[2]-dy,each = layout[1])
  text(x, y, txt, cex=cex)
}

```

```

assign("panel", panel, frame = 1)
assign("worms", worms, frame = 1)
assign("cubines", cubines, frame = 1)
assign("coefsave", coefsave, frame = 1)
assign("hor", hor, frame = 1)
assign("vert", vert, frame = 1)
assign("ci", ci, frame = 1)
assign(".est", NULL, frame = 0)
if(length(layout) == 1) layout <- rep(layout, 2)
n <- prod(layout)
classes <- equal.count(data$age, n, overlap = overlap)
if(n == 1) form <- ~ data$z
else form <- ~ data$z | classes
print.trellis(qqmath(form, layout = layout, aspect = 1, strip = F,
sub = list(sub, cex = 0.5), xlab = list("Unit normal
quantile", cex = 0.75), ylab = list("Deviation",
cex = 0.75), panel = panel))
if (labels) agetext(classes, layout, cex = 0.6, dx = 0.06, dy = 0.02)
return(list(classes = classes, .est = get(".est", frame = 0)))
}
ciplot <- function(n, level = 0.95, lz = -2.75, hz = 2.75, dz = 0.25) {
# adds confidence interval to Q-Q plot panel
z <- seq(lz, hz, dz)
p <- pnorm(z)
se <- (1/dnorm(z)) * (sqrt(p*(1-p)/n))
low <- qnorm((1 - level)/2)*se
high <- -low
lines(z, low, lty=2)
lines(z, high, lty=2)
}

```

#### ACKNOWLEDGEMENTS

We thank Professor J. C. van Wieringen for his thoughtful and detailed comments on an earlier version of this paper. Dr C. G. M. Oudshoorn and Professor W. R. van Zwet were helpful in the interpretation of the mathematical properties of the shape coefficients. We are grateful to Professor T. J. Cole for letting us use his LMS software. We thank the editor and two anonymous reviewers for their expert contributions that helped us to improve the presentation. The Fourth Dutch Growth Study was executed in co-operation with the Well Baby Clinics and the Municipal Health Services, and was sponsored by the Ministry of Health, Welfare and Sports, the Praeventiefonds, Nutricia and Pharmacia and Upjohn.

#### REFERENCES

1. Wright EM, Royston P. A comparison of statistical methods for age-related reference intervals. *Journal of the Royal Statistical Society, Series A* 1997; **160**:47–69.

2. Cole TJ. Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society, Series A* 1998; **151**:385–418.
3. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalised likelihood. *Statistics in Medicine* 1992; **11**:1305–1319.
4. Royston JP. Estimation, reference ranges and goodness of fit for the three-parameter lognormal distribution. *Statistics in Medicine* 1992; **11**:897–912.
5. Thompson ML, Theron GB. Maximum likelihood estimation of reference centiles. *Statistics in Medicine* 1990; **12**:539–548.
6. Fredriks AM, van Buuren S, Burgmeijer RJF, Meulmeester JF, Beuker RJ, Brugman E, Roede MJ, Verloove-Vanhorick SP, Wit JM. Continuing positive secular change in The Netherlands 1955–1997. *Pediatric Research* 2000; **47**:316–323.
7. Fredriks AM, van Buuren S, Wit JM, Verloove-Vanhorick SP. Body index measurements in 1996–7 compared with 1980. *Archives of Childhood Diseases* 2000; **82**:107–112.
8. de Wijn JF, de Haas JH. Groeidiagrammen van 1–25 jarigen in Nederland (Growth diagrams for 1–25 year olds in the Netherlands). Verhandelingen Nederlands Instituut voor Praeventieve Geneeskunde, Leiden, 1960.
9. Van Wieringen JC, Wafelbakker F, Verbrugge HP, de Haas JH. *Growth Diagrams 1965 Netherlands. Second National Survey on 0–24-year-olds*. Wolters Noordhoff: Groningen, 1971.
10. Roede MJ, van Wieringen JC. Growth diagrams 1980: Netherlands third nationwide survey. *Tijdschrift voor Sociale Gezondheidszorg* 1985; **63**:supplement.
11. Statistics Netherlands. *Demographic Statistics* 1996. Centraal Bureau voor de Statistiek: Voorburg/Heerlen, 1997.
12. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman and Hall: London, 1990.
13. Cole TJ, Freeman JV, Preece MA. British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statistics in Medicine* 1998; **17**:407–429.
14. Van Wieringen JC. Secular changes of growth: 1964–1966 height and weight surveys in the Netherlands, in historical perspective. Thesis, Netherlands Institute for Preventive Medicine TNO, Leiden, 1972.
15. Healy MJR, Rasbash J, Yang M. Distribution-free estimation of age-related centiles. *Annals of Human Biology* 1988; **15**:17–22.
16. Royston JP. An extension of Shapiro and Wilk's  $W$  test for normality to large samples. *Applied Statistics* 1982; **31**:115–124.
17. Wang C. *Sense and Nonsense of Statistical Inference: Controversy, Misuse and Subtlety*. Marcel Dekker: New York, 1993.
18. Hoaglin DC. Using quantiles to study shape. In *Exploring Data Tables, Trends, and Shapes*, Hoaglin DC, Mosteller F, Tukey JW (eds). Wiley: New York, 1985; 417–459.
19. Friendly M. *SAS System for Statistical Graphics*, 1st edn. SAS Institute: Cary, NC, 1991.
20. Royston JP, Wright EM. A method for estimating age-specific reference intervals ('normal ranges') based on fractional polynomials and exponential transformation. *Journal of the Royal Statistical Society, Series A* 1998; **161**:79–101.
21. Chambers JM, Cleveland WS, Kleiner B, Tukey PA. *Graphical Methods for Data Analysis*. Wadsworth Publishing Company: Belmont, CA, 1983.
22. Cornish EA, Fisher RA. Moments and cumulants in the specification of distributions. *Revue de l'Institut Statistique Internationale* 1937; **5**:1–14 (reprinted in the collected works of R.A. Fisher, vol. 4).
23. Kendall M, Stuart A, Ord JK. *Kendall's Advanced Theory of Statistics, Volume 1*, 5th edn. Charles Griffin: London, 1987.
24. Benard A, Bos-Levenbach EC. Het uitzetten van waarnemingen op waarschijnlijkheidspapier (Plotting observations on probability paper). *Statistica Neerlandica* 1953; **7**:163–173.
25. Van Zwet WR. Convex transformations of random variables. Dissertation, University of Amsterdam, 1964.
26. Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Statistics in Medicine* 2000; **19**:1831–1847.
27. Heagerty PJ, Pepe MS. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics* 1999; **48**:533–551.
28. Van't Hof MA, Wit JM, Roede MJ. A method to construct age references for skewed skinfold data using Box-Cox transformations to normality. *Human Biology* 1985; **57**:131–139.