**A** **Systematic Approach for Creating ad-hoc Customized Fetal Growth Curves**

***Abstract***

*Scanning for potential abnormalities in the development process of fetal growth is one of the main purposes of antenatal care and, if reliable, it may be a first alarm signal to predict cases of intrauterine growth restriction or macrosomia.*

*Centile charts, which show fetal biometric measurements plotted against gestational ages (typically measured in weeks), are widely used as a screening test in order to recognize those foetuses with borderline biometric measurements that may be associated with structural or chromosomal abnormalities or simply related to constitutional smallness.*

*Due to the substantial plethora and heterogeneity of the methodologies adopted to derive centile charts, an appropriate standardization process may be crucial to a better quality care.*

*In this perspective, starting from the examination of the usually adopted statistical methods, we propose a systematic approach in charge to facilitate the construction of ad-hoc customized fetal growth curves, which are becoming strictly necessary in order to provide accurate diagnoses and to reduce the occurrence of false positive and/or false negative.*

***Keywords***

Fetal biometry, fetal intrauterine growth, fetal ultrasound, ultrasonographic growth curves, fetal growth curves, neonatal anthropometry, statistical methods, least square technique, outlier detection, percentiles, Gaussian distribution, reference intervals, customization, longitudinal study, cross-sectional study.

***Authors***

Lucia Vaira, Anthony S. Chapman, Mario A. Bochicchio, …..

**A Systematic Approach for Creating ad-hoc Customized Fetal Growth Curves**

**Introduction**

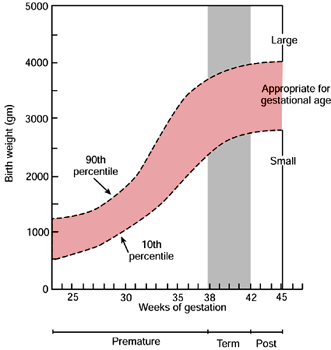
The optimal practice in prenatal care requires an accurate assessment of the growth and the wellbeing of the foetus by analysing fetal biometric parameters trends along with gestational age.

Perni et al. [1] in 2004 demonstrated that all measurements of fetal biometry are highly reproducible both by the same and by different operators. This information is of great importance since major decisions are taken daily based on biometric results.

Biometric measurements (e.g. Femur Length, Abdominal Circumference, Head Circumference, …) are often plotted and compared with reference growth curves and are considered to be [2]:

* normal or *Appropriate for Gestational Age* (AGA), with values between the 10th and the 90th percentile;
* too low or *Small for Gestational Age* (SGA), with values below the 5th percentile;
* too high or *Large for Gestational Age* (LGA), with values above the 95th percentile.

In Figure 1 the typical fetal growth curve is depicted, with the possible foetus status: pre-term, full-term and post-term according to the gestational age (GA in the following).



**Figure** 1**: Fetal growth curve concept**

Typically, the status of the pregnancy is determined in a retrospective fashion, and there is scope for controversy in deciding which pregnancy should be defined as abnormal and excluded from the reference sample [3].

Several approaches to improve the ability of fetal biometry to detect potential high-risk foetuses have been proposed and developed in literature:

* parameters can be compared with a given customized reference curve [4], [5];
* parameters can be integrated into a formula (e.g. for estimation of fetal weight) [6];
* parameters can be assessed each other (proportionality index) [7], [8], [9];
* parameters can be measured and assessed longitudinally (speed or rate of growth) [10], [11].

However, the use of cross-sectional growth curves remains the most widely used method of screening for fetal growth abnormalities.

A variety of techniques have been proposed in literature for constructing centile charts. The collection of fetal biometric data necessary to develop such charts represents an important part of the whole diagnostic process since the choice of the chart that will be adopted for diagnosis has a considerable effect on the interpretation of biometric data.

In particular, the choice of a suitable methodology, which yields the reference charts, is particularly crucial since it could avoid and, in general, decrease the overall number of misclassified foetuses (false positive and/or false negative), improving the accuracy of diagnosis as well.

It is commonly accepted that the quality of fetal measurement depends on three main factors [12]:

* machines and measurement systems (callipers, ellipses...) maintenance and accuracy;
* appropriateness of measurement charts for local population and analysis methods;
* quality of image plane acquisition and measurement.

This implies that “data quality check”, which has to be performed on the gathered data, is highly recommended in order to ensure that such methods can work when applied to the general population.

Different statistical approaches for obtaining reference intervals and centile charts has been published [13], [14], [15], [16] but the widespread used method is due to Altman in 1993 [17] which is mainly based on regression analysis to fit curves that better represent the growth model.

For modelling the growth phenomenon, different candidate functions have been proposed in literature. A sophisticated example is the Gompertz curve, which was used in [18]. A polynomial of adequate degree was used in [19], and the lambda-mu-sigma (LMS) method was discussed by Cole in [20]. Other authors [21] proposed a multidimensional view of the standard for the foetuses’ development process by using a logistic function with three parameters and a nonlinear regression method.

**The general problem**

As any other observation phenomenon, the representation of fetal growth data suffers from “dispersion” around an ideal behaviour, which can be described by analytical or numerical models of the phenomenon itself.

In order to correlate the experimental data with the numerical ones, it is necessary to use appropriate methods to “mediate” the effect of such dispersion. One of the most adopted techniques for obtaining the interpolating curve to the experimental data is that of least mean squares that is mainly based on the minimization of the sum of squares of the deviations related to each measure by the approximating curve, using a polynomial of any order to obtain the so called “best fit”.

The fetal growth assessment and the consequential computation of reference centiles, can be seen as a typical “inverse problem”, which consists of inferring the values of the parameters that characterize the system starting from the gathered data (representing the experimental measurements).

Least squares are popular for solving inverse problems because they lead to the easiest computations. Their only drawback is their lack of robustness, i.e. their strong sensitivity to a small number of large errors (outliers) in a data set, and in the case of collection of fetal biometric data, the outliers are often present. This can be due to different reasons (human mistakes, wrong equipment calibration, …), which can bias the estimation results.

The usage of appropriate and adequate fetal growth charts represents the preliminary process to any other quality control policy.

A multiplicity of statistical methods for constructing reference intervals and centile charts has been presented in literature but, in our knowledge, does not exist a systematic approach, which includes the outliers handling as well. We feel that outlier detection should be considered as a preliminary step for a coherent data analysis needed to obtain “clean data” to which apply the classical statistical manipulations.

Although outliers are often considered as an error or noise, they should not be regarded as a pejorative term since they may often carry important and valuable information about the process under investigation or the data gathering and recording process. This means that they should be investigated carefully and hence, before considering the possible elimination of these points from the data, when we have a measurement that is radically different than the others in a series of measurements, we have to decide why it is different. That is, we have to decide if the anomalous measurement is the result of some mistake or whether it is representative of the population being sampled or measured and should be included with all the other measurements.

As Taylor [22] points out, omitting any data is controversial. This means that an appropriate methodology, able to split the original data set into two portions (good and bad values) and also to study the potential goodness of the “strange” data, has to be adopted in order to achieve a consistent analysis.

In literature, other authors [23], [24] exclude only “*extreme outliers*”, values whose combinations were implausible (more simply detected at earlier gestational ages). This unlikelihood is mainly based on the rule of thumb related to the interquartile range (i.e. the positive difference the first and third quartile, a range which contains the middle 50% of all the data values): an observation x is categorized as an extreme outlier if it lies three times the interquartile range.

It has been recognized that fetal growth curves are affected by several factors, which have to be taken into consideration when a foetus is assessed. They include: ethnic group; genetic factors; maternal characteristics such as nutrition, education, pregnancy at a very early age, unhealthy lifestyle; complications of pregnancy; physical and socio-economic position; disease states; physical work during pregnancy; tobacco and alcohol consumption.

In our recent paper we have defined all the essential factors to take into account in the construction of the fetal biometric charts suitable for diagnostic purposes. In particular, in [25] we have proposed a collaborative approach that by means of an online system is able to build personalized fetal growth curves embracing all the physiological and pathological variables affecting fetal growth.

In this paper, we describe the statistical approach in detail and we demonstrate its main properties by analysing a true dataset of biometric data coming from clinical practise in order to create ad-hoc fetal growth curves by handling also the possibility of outliers’ existence.

**The proposal: material and methods**

*Study design*

Dealing with such a kind of statistical analysis, one of the most heated debates is related to the study design, in particular to the option about measuring each involved foetus once (cross-sectional study) or more times (longitudinal study).

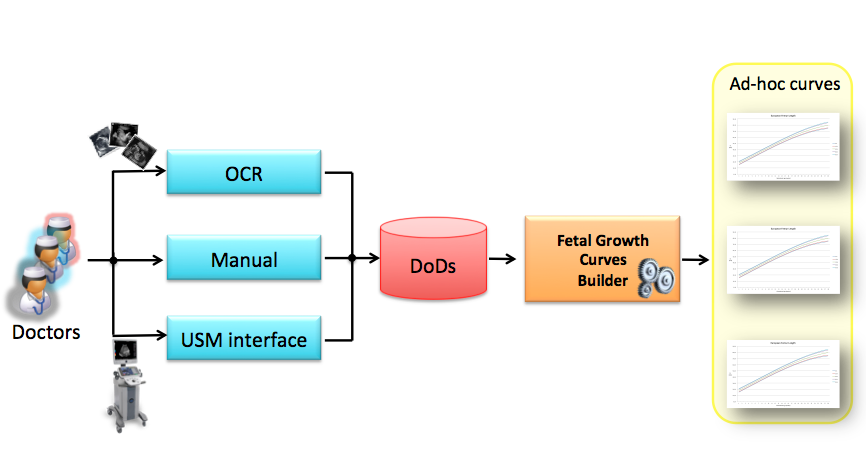
Serial measurements in order to assess fetal growth were first introduced by Willocks in 1962 [26] and the first longitudinal chart for biparietal diameter was presented by Campbell and Newman in 1971 [27]. However, many obstetricians use cross-sectional charts for their serial fetal examinations, even though cross-sectional studies are not designed to evaluate growth. Royston points out that when using size-charts instead of growth charts to determine growth, the obstetrician may be misled [3]. In contrast to cross-sectional data, longitudinal data have a hierarchical structure based on two levels of variation: within foetuses between gestational ages (level 1), and variation between foetuses (level 2) [28][29][30].

In this paper we will ignore the serial nature of the data, calculating the percentiles as if the data were derived from separate individuals.

*Data collection*

As illustrated in Figure 2, the data gathering process can be implemented in three main manners:

* Data can be manually inserted by means of direct input. This is the widely adopted method by clinicians who typically handle such information on paper and hence they transcribe hand-written worksheets.
* Data can be collected directly at the source, i.e. at the output of the medical equipment used for the assessment of fetal biometric parameters, such as the traditional ultrasound machine. In this case, the standard adopted for the distribution and viewing of medical images is the DICOM (Digital Imaging and Communications in Medicine) standard, which allows obtaining discrete values directly from its headers and payload.
* Data can be inserted by uploading ultrasound pictures coming from the clinical practice (doctors’ printed archives). In this case the images are acquired, scanned and converted in the TIFF format. An Optical Character Recognition (OCR) subsystem will be then in charge to analyse and extract textual data directly from the acquired pictures which are typically complemented by measurements (biometric parameters indication and the corresponding values), calculations (gestational age measured in weeks which can be obtained started from the Last menstrual Period for example), and descriptive data (ultrasound machine settings, exam date, patient’s data, and so on).



**Figure** 2**: Involved data sources to gather biometric data from clinical practise**

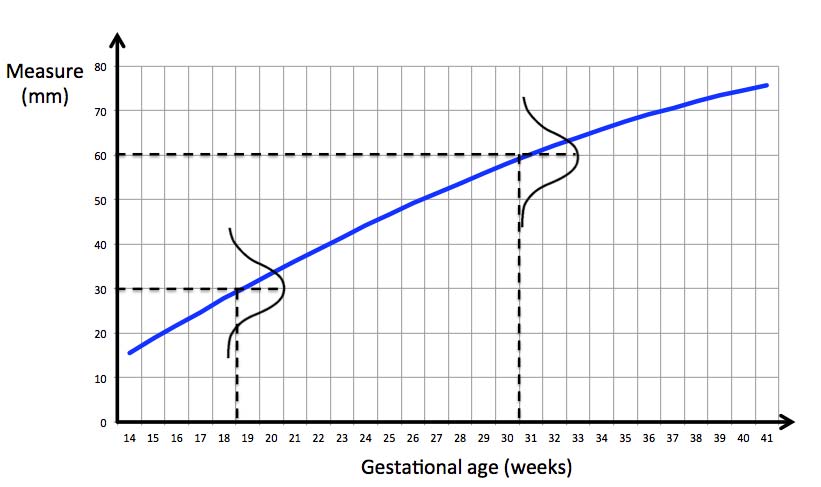
Independently from the nature of the data source, the biometric measurements coming from clinical practise will contribute to feed and enrich the Database of Databases (DoDs), as depicted in Figure 2, which will be adopted by the Fetal Growth Curves Builder (FGCB in the following) that represents the heart of the approach that we want to propose, which will be in charge to create ad-hoc fetal growth curves.

Assuming that, at each GA the measurement of interest has a Gaussian distribution with a mean and a standard deviation (SD in the following) and that, in general, both vary smoothly with gestational age, a centile curve can be calculated using the well-known formula:

*Centile = mean + K \* SD (1)*

where K is the desired normal equivalent deviate (NED) that takes a value corresponding to the proportion of the standard normal distribution (with mean of 0 and SD of 1) lying to the left of it. For instance, the 50th centile (with a proportion of 0.5 of the standard normal distribution to the left of it) has an NED of 0, while the determination of a 90% reference range (i.e. the 5th and 95th centile curves) would require K = ±1.645.

In Figure 3 is showed how, by fixing certain values of GA, the corresponding value representing the biometric parameter measure denotes the mean on which the Gaussian is centred (situated on the regression curve).

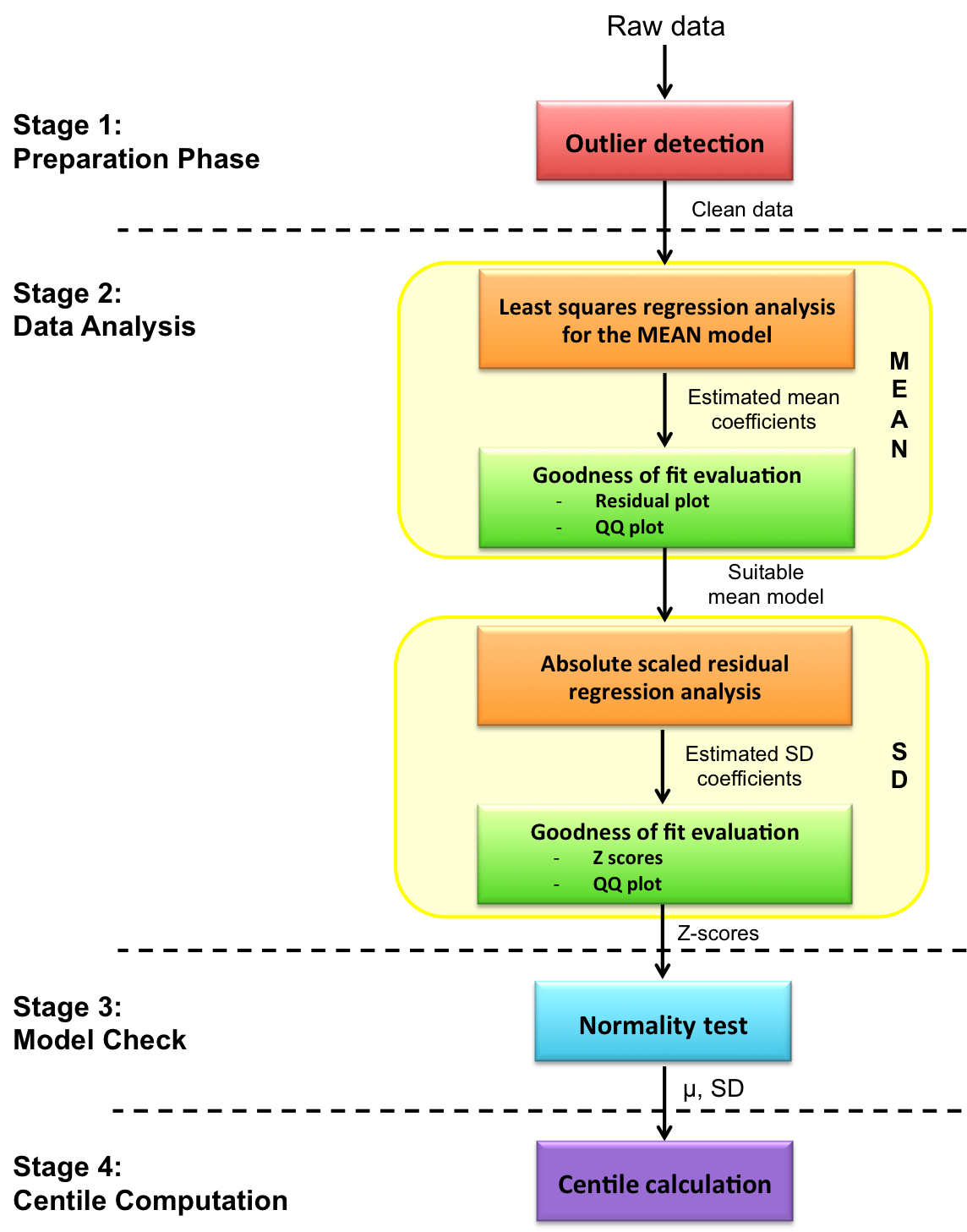


**Figure** 3**: Any fetal growth curve can be seen as a set of Gaussians**

*Procedure*

The proposed approach is based on 4 main steps, as depicted in Figure 4:

* stage 1- preparation phase: involves the outlier detection starting from the raw data;
* stage 2 - data analysis: comprises the modelling and checking phases both for the mean and for the SD by adopting least squares regression analysis;
* stage 3 - model check: includes the normality test needed to check the model;
* stage 4 - centile computation: includes the centile calculation starting from the estimated mean and SD obtained in the previous steps.



**Figure** 4**: Framework block diagram**

*Stage 1: Preparation phase*

In this context we assume that the only qualitative and quantitative variables needed for centile computation purposes are those related to the different biometric parameters and the corresponding gestational age. Furthermore we consider the data processing phase (including merging, subsetting and transforming operations) as if it were already concluded.

This first step of the approach (the preparation phase) allows the raw data to be “cleaned” by means of the outlier detection method.

The literature on outliers is quite extensive and it covers not only statistical sectors but also other area of science. Nevertheless, an exact definition of outlier does not exist, this because it often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins [31] defines an outlier as “*an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*”. Barnet and Lewis [32] affirm that “*an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs*”, similarly, Johnson [33] defines an outlier as “*an observation in a data set which appears to be inconsistent with the remainder of that set of data*”.

Being the outliers the most extreme observations, they may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

Most methods used to detect outliers assume a known distribution of the data (i.e. often independent and identically distributed) and often assume that the distribution parameters and the type of expected outliers are also known [32]. Clustering algorithms are also adopted in many cases but they are mostly optimized to find clusters rather than outliers. In addition, accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters. Outlier detection is subjective, as there does not seem to be any universally accepted mathematical criteria for defining which observations can be considered regular and which are outliers.

However, at the current time, it is possible to identify at least three fundamental approaches to the problem of outlier detection [34]:

* approach 1: it determines the outliers without any prior knowledge of the trends of the observations (this is a learning approach analogous to unsupervised clustering). The approach is predominantly retrospective and is comparable to a batch-processing system and employs two sub-techniques: diagnostic approach (it highlights the potential outlying points and once detected, it may remove them) and accommodation approach (it incorporates the outliers into the distribution model generated and employs a robust classification method);
* approach 2: it models both normality and abnormality (this approach is similar to the supervised classification) and requires that each observation in the data is pre-tagged as “normal” or “abnormal” data;
* approach 3: it models only normality or, in a very few cases, models abnormality (it is analogous to a semi-supervised recognition or detection).

Common outlier identification techniques include the Chauvenet’s criterion and the Dixon test.

The Chauvenet’s criterion [35] has found applications in astronomy, nuclear technology, geology, epidemiology, molecular biology, radiology and many other fields of physical science. Taylor’s statement [22] of Chauvenet’s criterion is:

*“If you make N measurements of a single quantity x and if one of the measurements (, say) is suspiciously different from all the others, Chauvenet’s criterion gives a simple test for deciding whether to reject this suspect value. First, compute the mean and standard deviation of all N measurements and then find the number of standard deviations by which differs from*

*Next, find the probability (assuming the measurements are normally distributed about with width of ) of getting a result as deviant as , and hence, the number of measurements expected to deviate this much,*

*If n < 0.5, then according to Chauvenet’s criterion, you can reject the value .”*

Another kind of technique is the Dixon's Q test [36],which is a used for identification and rejection of outliers, but it should be used sparingly and never more than once in a data set and this is not our case, since in our collection could be more then one outlier. It is hence recommended for use in small populations and for situations where data are normally distributed but the mean or variance change slowly over time.

Another method that can be used to screen data for outliers is the Z-score, using the mean and the standard deviation. The basic idea of this rule is that if the sample follows a normal distribution, N (μ, σ2), then Z-scores follow a standard normal distribution, N (0, 1), and Z-scores points that exceed 3 in absolute value are generally considered as outliers. Since no Z-score exceeds 3 in a sample size less than or equal to 10, the Z-score method is not very good for outlier labelling, particularly in small data sets [37]. Another limitation of this rule is that the standard deviation can be inflated by a few or even a single observation having an extreme value. Thus it can cause a masking problem, i.e., the less extreme outliers go undetected because of the most extreme outlier(s), and vice versa.

What is assured is that when there are outliers, the estimation result can be biased. Least squares are not robust to such outliers, since they ponder too much on the model, in fact a single outlier could have an arbitrarily large effect on the final estimate.

The *breakdown point* of an estimator represents the minimum fraction of outliers that are sufficient to produce an arbitrary large bias, that is the percentage of data that can be changed without the model, and hence the parameters are changed. It was introduced by Hampel in 1971 [38] and it’s well known that the least square estimators breakdown point is 0%.

In order to obtain a robust and reliable evaluation, this first step of the whole approach has to be viewed a two-stage process:

1. Classify data as outliers and inliers
2. Perform the least squares technique to only inliers

Although the approach may seem simple enough, it is necessary to consider that an outlier detection method that is based on an initial not robust measure, can suffer for the effect of masking, i.e. the capacity that a group of outliers may mask the other and then escape to the control [39].

To exclude outliers and perform the least squares technique, the mechanism of the *consensus set* (the largest set of values compatible with a certain pattern) could be considered. In particular, the RANSAC (RANdom SAmple Consensus) algorithm [40] could be adopted in order to produce a model which is calculated considering only the inliers and that generates a correct result with a given probability, which increases with the increase of the allowed iterations. In particular, the RANSAC algorithm is an iterative method that can be summarized in the following steps:

1. Randomly choose minimal subset of data points necessary to fit the model (a sample)
2. Points within some distance threshold t of model are a *consensus set*. Size of consensus set is model’s support
3. Repeat for N samples; model with the biggest support is most robust fir
   1. Points within distance t of best model are inliers
   2. Fit final model to all inliers

RANSAC is not a deterministic algorithm, thus the results are produced correctly only with a certain probability, which increases with the growing of iterations.

The algorithm iteratively selects a random subset of input points and estimates the parameters of mathematical model. Estimated model is evaluated by the test against the input data and the subsequent computation of inliers.

An advantage of the method is about its ability to achieve robust estimation of the model parameters, i.e., it can estimate the parameters with a high degree of accuracy even when significant amount of outliers are present in the data set.

A disadvantage of RANSAC is that there is no upper bound on the time it takes to compute these parameters. When an upper time bound is used (a maximum number of iterations) the solution obtained may not be the optimal one, it may not even be one that fits the data in a good way. A reasonable model can be produced by RANSAC only with a certain probability, a probability that becomes larger the more iterations that are used.

RANSAC can tolerate more than 50% outliers, but it requires the user to specify an error tolerance which is not known a priori in many practical environments. The value of this tolerance has a significant influence on the performance of RANSAC.

Another disadvantage of RANSAC is that it requires the setting of problem-specific thresholds.

Outliers may be due to experimental mistakes, but sometimes they rather may be the result of biological variation, or differences in some other variable that is not included in the predicted model. Hence, the presence of an outlier poses a critical dilemma: it could be the most interesting finding in the study and it would be a big mistake to automatically exclude it without further experimentation.

Every point of the analysed samples lies on a continuous spectrum from “normal data” to noise, and finally to anomalies, as illustrated in Figure 5. The separation of the different regions of this spectrum is often not precisely defined and is chosen on an ad-hoc basis according to application-specific or domain-specific criteria. Some authors use the terms weak outliers and strong outliers in order to distinguish between noise and anomalies [41], [42]. The interpretability of an outlier detection model is extremely important since is often desirable to determine why a particular data point is an outlier in terms of its relative behaviour with respect to the remaining data.

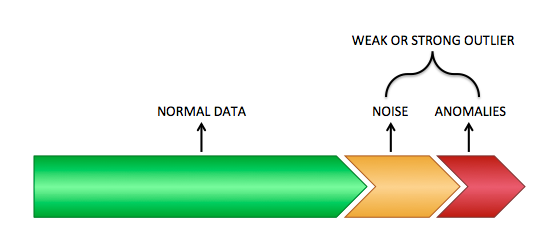


Figure 5: The spectrum from normal data to outliers (from left to right)

Most outlier detection algorithms output an outlier score and a threshold on this score is used in order to declare data points as outliers. If the threshold is picked too restrictively in order to minimize the number of declared outliers, then the algorithm will miss true outlier points (false negatives). On the other hand, if the algorithm declares too many data points as outliers, then it will lead to too many false positives. This trade-off can be measured in terms of precision and recall and by analysing the Precision-Recall (PR) curve or the more easily interpretable Receiving Operating Characteristics (ROC) curve [43].

Motulsky and Brown [44] in 2006 proposed a Robust regression and Outlier removal (ROUT) method, following three main steps:

1. adoption of robust nonlinear regression method to fit a curve that is not influenced by outliers;
2. residuals analysis of the robust fit to identify any outliers;
3. outliers removal and ordinary least-squares regression performance on the remaining data.

There are schools of thought that are adamantly opposed to deletion of outliers. Rather than remove outliers, an alternative approach is to fit all the data (including any outliers) using a robust method that accommodates outliers so they have minimal impact [46], [48]. Robust fitting can find reasonable best-fit values of the model’s parameters but cannot be used to compare the fits of alternative models. Moreover no robust nonlinear regression method provides reliable confidence intervals for the parameters or confidence bands for the curve. So robust fitting, alone, is a not yet an approach that can be recommended for routine use.

*Stage 2: Data Analysis*

The second stage allows obtaining the coefficients describing both the model of the mean and that of the SD, which are necessary for the centile computation.

These coefficients can be computed starting from a set of experimental data in order to find the best model (in terms of simplicity and interpretability), which can better represent the data trend, by adopting automatic learning methods. The choice of the appropriate model is a critical issue that most of the time requires experience and knowledge of the considered specific field.

Traditionally, this can be achieved by means of interpolation, which could yet produce very complex functions especially with the increase of the number of values (this is because any kind of information is embedded in data, even a possible noise; a good model, instead, should include only the components of the phenomenon to be analysed).

Another approach is that of statistical regression, which however, seeks to optimize the parameters for a pre-specified model structure, that is it attempts to estimate the parameters of a function provided in advance.

The genetic programming is closer to the aim, avoiding imposing a priori assumptions and inferring the model from the data, but being originally devoted to solve more general problems, it might not effectively exploit the intrinsic characteristics of the problem.

In 1990 Koza [47] started to use genetic algorithms in order to directly discover the mathematical model exploiting just the data. The basic idea is very simple: a series of functions is generated and a test of goodness is run on it. According to some specific rules, another set of functions is then generated. These rules attempt to reward those functions which are shown the best, so that the next group of functions is on average better with respect to the last, increasing hence the probability that the desired function exists in that set. This approach (also known as symbolic regression) is however very expensive in terms of required power machine.

Why do we adopt statistical regression?

The number of possible and different statistical techniques for the construction of ad-hoc customized fetal growth curves has mushroomed in literature in the last three decades but, generally, these techniques can be divided into two main categories: parametric methods, which are based on modelling the data distribution and non-parametric or, equivalently, empirical methods.

In the most common parametric method, the fundamental assumption is that at each GA the measurement of interest has a normal distribution (as stated beforehand). Regression analysis technique is applied for estimating both the mean and the SD of the different biometric parameters across GA, as detailed by Altman and Chitty [48] and Royston and Wright [3]. In particular, the mean has to be estimated by the fitted values from an appropriate polynomial regression curve of the measurement of interest on GA, which is adequate to represent the relationships between them. Least squares regression analysis can be used to estimate both the mean and the SD curves as polynomial functions of GA, where the biometric parameter is the outcome (or dependent) variable and GA is the predictor (explanatory or independent) variable.

One of the most delicate aspects related to the multiple linear regression is about the selection of explanatory variables in order to obtain a satisfactory final model. In fact, the estimation of an initial model that includes all k possible regressors almost certainly will produce a result in which some regressors have a significant *p-value* while others will not. This raises the question of the correct/appropriate selection of the subset of regressors.

In this context does not exist an ideal algorithm that allows to achieve the final optimal model.

In order to establish which of the available models better represents the real data, the *matryoshka* method is often adopted: a sequence of classes of models in which each class includes within it the previous classes. A concrete example of such a method is the approach recommended by Royston and Wright [3], which suggest the initial use of a cubic polynomial (, where, for simplicity, GA is represented by t). If the cubic coefficient, d, is not significantly different from zero (approximately if d is less than twice its SD), a quadratic polynomial () should be fitted with the same assessment made of the quadratic coefficient, c.

The process should be repeated until no further removal of terms is possible.

While quadratic or cubic curves will often give a good fit to the data, Altman and Chitty [48] suggest the linear-cubic model () as a good alternative for fetal size data. It is advocated that the choice of curve be based not only on statistical significance, but also that the quality of fit to the data and aesthetic appearance, especially at the extremes of GA, should be taken into account.

Once a suitable mean model has been decided upon, attention can turn to the variability in the data.

The SD can be estimated starting from the residuals (the differences between the observed values and those predicted by the regression equation).

Altman [17] showed that, provided the residuals for any given gestational age followed a normal distribution, the absolute value of the residual, without the + or – sign, can be used to estimate the SD. It follows that the residuals from the mean model should also be normally distributed.

This in turn means that the absolute residuals (residuals with the sign removed) have a half normal distribution. As the mean of a half standard normal distribution is , the mean of the absolute residuals multiplied by , is an estimate of the SD of the residuals. Hence if the SD is not reasonably constant over GA, predicted values from a regression of the absolute residuals on age multiplied by , will give age-specific estimates of the SD of the residuals, and hence of y.

An alternative formulation for Altman’s approach favoured by Royston and Wright [3] is to produce “scaled absolute residuals” (SARs) by multiplying the absolute residuals by , The SARs are then regressed on GA, the predicted values from which again estimate the SD of the residuals.

Under either formulation, if the absolute residuals, be they scaled or unscaled, show no trend with GA, the SD is estimated as the SD of the unscaled original residuals (observed value minus predicted value). If there is a trend, polynomial regression is needed to estimate an appropriate curve in the same way as for the mean. Altman suggests that it is unlikely that a curve more complex than quadratic is required for a satisfactory fit to the SD [17]. Superimposing ±1.645 × SD on the residual plot is useful to see how well the SD has been modelled, as approximately 90% of the observed residuals should fall within these limits.

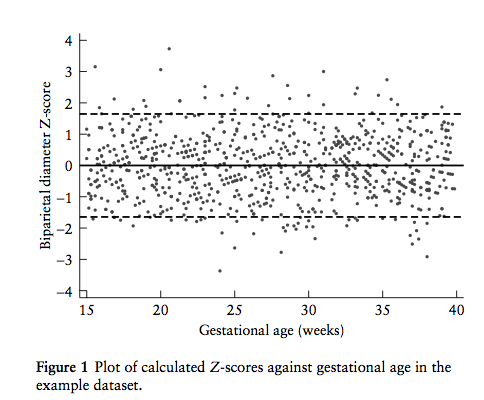
*Stage 3: Model’s Check*

A useful tool in assessing model fit is the adoption of Z-scores (also known as SD scores), which are defined as:

(2)

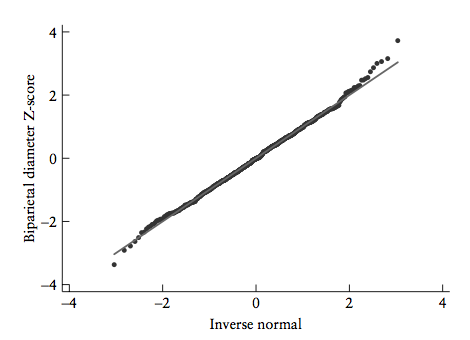
where and are, respectively, the mean and SD given by the model for the GA at which the observation is made. Hence Z-scores represent the observed values expressed on a standard normal scale (with a mean of 0 and SD of 1), with the mean and SD adjusted for GA.

Altman and Chitty [48] recommend three methods of evaluation for the goodness of fit. Firstly, a plot of the Z-scores against GA should be checked for the existence of any patterns. The Z-scores should be randomly scattered about zero at all GAs, with any deviation from this indicating that the mean curve may require modification. This is shown in Figure 6 considering an example dataset, with the Biparietal Diameter Z-scores appearing to adhere to this stipulation.



**Figure** 6**: Plot of calculated Z-scores against gestational age in the example dataset**

Secondly, a normal plot (which is essentially a scatterplot of the actual data values plotted against the ‘ideal’ values from a normal distribution) can be used to check that the Z-scores have a close to normal distribution. This is signified by a roughly straight line but can be confirmed more formally using the Shapiro-Wilk W test or Shapiro-Francia W’ test. For explanatory purposes, Figure 7 shows that in the example dataset the Biparietal Diameter Z-scores do have a close to normal distribution and this is corroborated by both the Shapiro-Wilk W and Shapiro-Francia W’ tests having P of 0.998.



**Figure** 7**: Normal plot of calculated Z-scores in the example dataset**

Finally, the appropriate proportion of observations should fall between and outside fitted centiles, for example approximately 90% of Z-scores should lie between Z = −1.645 and Z = +1.645. Deviation from this may imply that a higher-order polynomial curve for the SD is needed. A brief examination suggests that approximately 90% of the data lie between the lines, with calculations confirming that 4.9% of the data lie below Z = −1.645 and 4.2% above Z = +1.645 (compared to an expected 5% for each). It is unlikely that the values will both be exactly 5%, so figures such as these indicate an adequate level of fit.

*Stage 4: Centile Computation*

Once a satisfactory model has been determined, the centile curves for the desired reference interval may be calculated by substituting the expressions for the mean and SD into equation (1). The Z-score for any new individual may be calculated using equation (2) and its centile obtained using the inverse normal distribution. Finally, the calculated centiles should be superimposed on the scatter diagram of observed values against GA to ensure a suitable fit.

**Results: a case study**

In order to show the feasibility of the proposed systematic approach, we have performed a test on the field considering about 500 ultrasound pictures including the Femur Length (FL in the following) parameter of foetuses related to Italian women undergoing ultrasound examination between the 11th and 41th weeks of gestation.

The FL measures are available in table form, as shown in Figure 8, where each single observation (measured in mm) is related to a specific gestational age (measured in weeks).

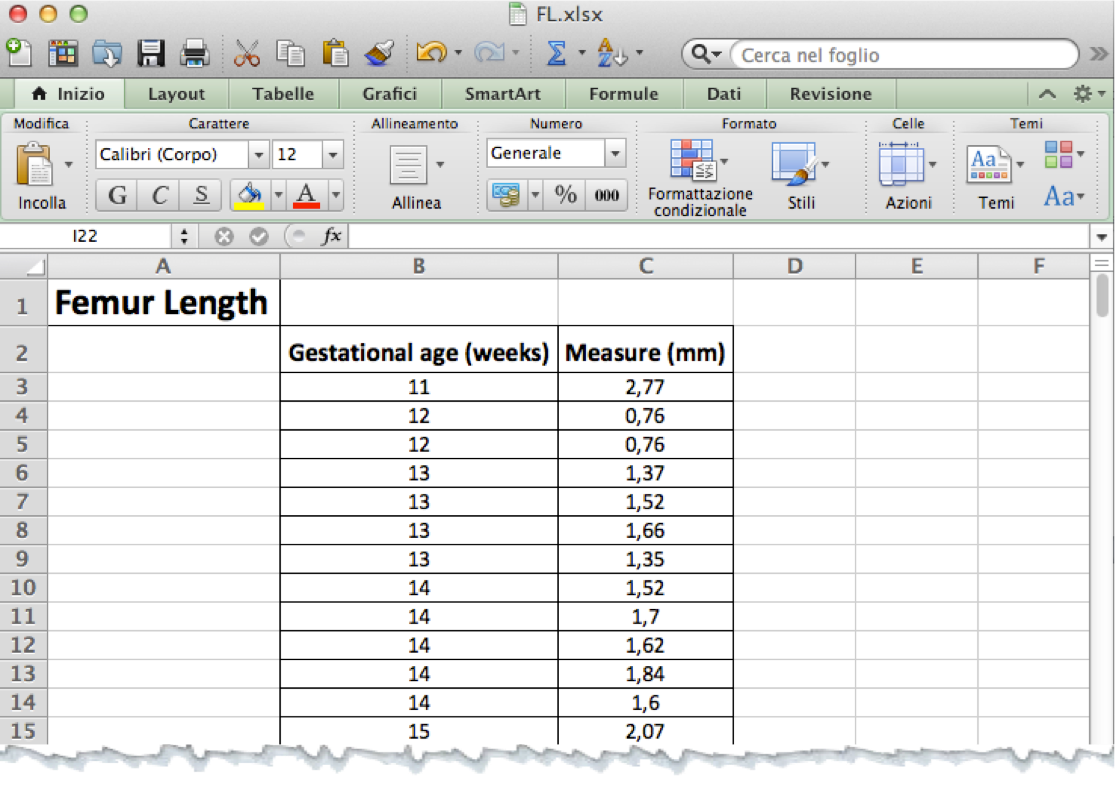


Figure 8: Femur Length raw data

*Case study – stage 1 – preparation phase*

There is no single universally applicable or generic outlier detection approach. The algorithm should be suitable for the specific data set in terms of correct distribution model, correct attribute types, speed, and so on. The decision will be also centred on the type of the chosen approach (clustering approach, classification approach or a novelty approach) and this depends on the data type, whether the data is pre-labelled, the ground truth of the labelling and finally on which will be the decision about the handling of the detected outliers.

We will exclusively focus on “unsupervised learning” techniques, reducing to finding sparse regions in large multidimensional data sets.

We choose RANSAC? Why?

We decided to adopt the RANSAC method to identify any outlier since our aim is to obtain an estimator characterized by a breakdown point greater than 0 and with a reasonable efficiency.

Applying the RANSAC algorithm (whose pseudo code is indicated in Appendix D) to our experimental data, we have obtained two outliers in all, depicted in Figure 9 as red star points.

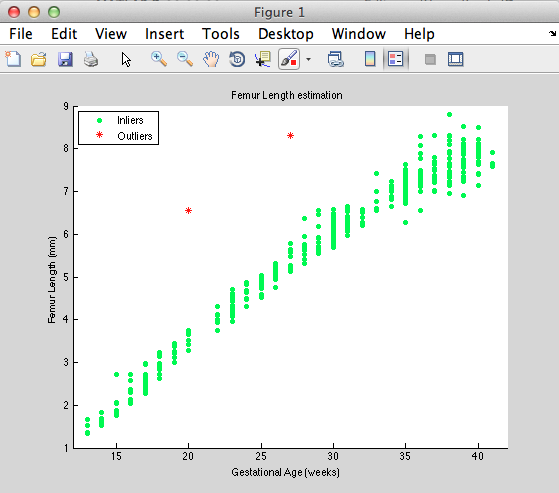


Figure 9: Sample trend (outliers and inliers)

Once the outliers have been detected, the sample has to be cleaned before the regression analysis.

This because when analysing data, outlying observations cause problems since they may strongly influence the result.

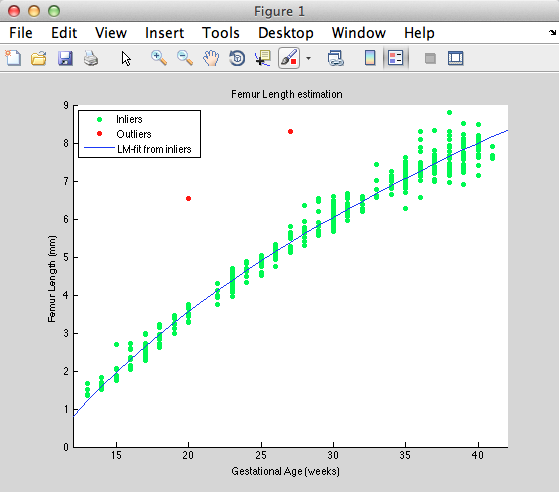
Outliers may be due to experimental mistakes or incorrect measurements, but sometimes they rather may be the result of biological variation, or differences in some other variable that is not included in the predicted model (in other words it could represent a rare event). Hence, the presence of such outliers could be the most interesting finding in the study, and it would be a big mistake to automatically exclude them without further experimentation.

What threshold?

a threshold is needed for deciding when a point is far enough from the curve to be declared an outlier.

This means that outliers are removed and ordinary least-squares regression is performed on the remaining data.

The solid line in fig. XXX represents the fit

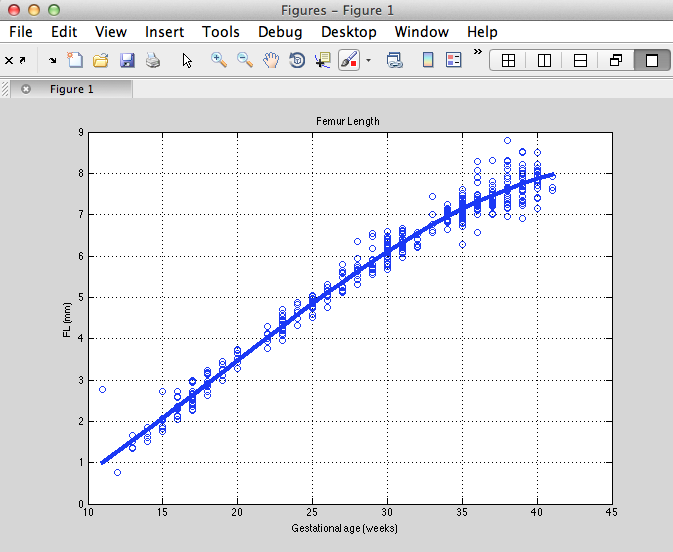


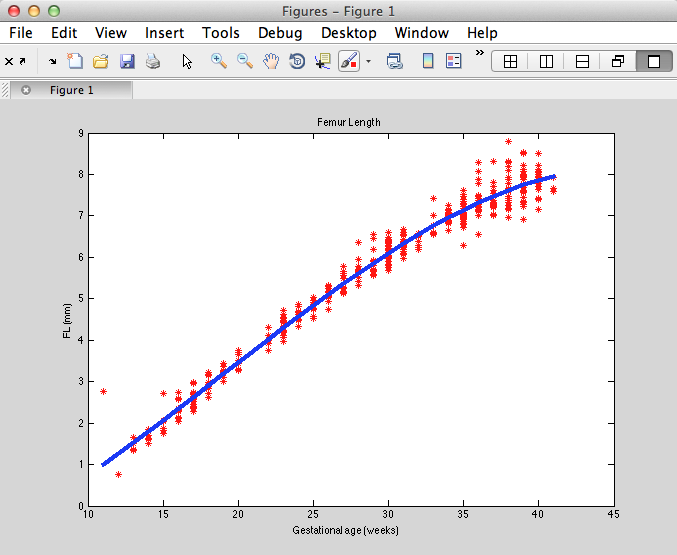
*Case study – stage 2 – data analysis*

Looking to the trend showed in the graph, it follows that the data can be represented by a third degree polynomial:

The coefficients can be computed by means of the Least Mean Square method obtaining a system constituted by 444 equations in 4 variables:

the corresponding polynomial will be:





*Case study – stage 3 – model check*

*Case study – stage 4 – centile computation*

**Conclusion**

The fetal growth assessment is a relevant matter in the perinatal care and, in some cases, it’s poorly handled in clinical practise, underestimating hence its diagnostic power that can have a significant effect already in the early stages of pregnancy.

Several studies and methodologies have been presented in literature, but none of them deals with the potential existence of outliers, which can greatly skew the result.

In our analysis we have included this crucial aspect and, moreover, we have outlined a systematic approach mainly based on standard statistical techniques, which adequately represent the flow of operations that has to be considered by anyone who want to approach to this topic. We are conscious that different technical choices exist: for instance, the outlier detection, can be implemented in several ways; the regression analysis can drive to different solutions, and so on, but a general methodological guide is always needed for a significant statistical analysis.

The adoption of cross-sectional or longitudinal data affects on the statistical method that has to be accordingly applied. Methodological aspects of cross-sectional studies were considered by Altman and Chitty, whereas in this paper we have ignored the serial nature of the data and we have calculated the different growth percentiles as if the data came from separate individuals considering, hence, the sample size as given by the number of different observations.

We feel that the tight integration between suitable data gathering processes and appropriate statistical techniques could be a step forward in the right direction to overcome the concerns related to the heterogeneity of methodologies and in order to provide accurate diagnoses.

**References**

1. Perni S.C., Chervenak F.A., Kalish R.B., Magherini-Rothe S., Predanic M., Streltzoff J., Skupski D.W. Intraobserver and interobserver reproducibility of fetal biometry. Ultrasound Obstet Gynecol 2004; 24: 654–658
2. Sananes N., Guigue V., Kohler M., Bouffet N., Cancellier M., Hornecker F., Hunsinger M.C., Kohler A., Mager C., Neumann M., Schmerber E., Tanghe M., Nisand I., Favre R. Use of Z-scores to select a fetal biometric reference curve. Ultrasound Obstet Gynecol 2009; 43: 404-409
3. Royston P. and Wright E.M. How to construct ‘normal ranges’ for fetal variables. Ultrasound Obstet Gynecol 1998; 11:30-38
4. Gardosi J., Chang A., Kalyan B., Sahota D., Symonds E.M. Customised antenatal growth charts. Lancet 1992; 339: 283–287
5. Mongelli M., Wilcox M., Chang A. An adjustable fetal weight standard. Ultrasound Obstet Gynecol 1995; 6: 168 – 174
6. Hadlock F.P., Harrist R.B., Sharman R.S., Deter R.L., Park S.K. Estimation of fetal weight with the use of head, body, and femur measurements – a prospective study. Am J Obstet Gynecol 1985; 151: 333–337
7. Campbell W.A., Vintzileos A.M., Rodis J.F., Turner G.W., Egan J.F., Nardi D.A. Use of the transverse cerebellar diameter/abdominal circumference ratio in pregnancies at risk for intrauterine growth retardation. J Clin Ultrasound 1994; 22: 497–502
8. Santolaya-Forgas J., Meyer W.J., Gauthier D.W., Kahn D. Intrapartum fetal subcutaneous tissue/femur length ratio: an ultrasonographic clue to fetal macrosomia. Am J Obstet Gynecol 1994; 171: 1072–1075
9. Campbell W.A.,Vintzileos A.M., Rodis J.F., Ciarleglio L., Craffey A. Efficacy of the biparietal diameter/femur length ratio to detect Down syndrome in patients with an abnormal biochemical screen. Fetal Diagn Ther 1994; 9: 175–182
10. Owen P., Donnet M.L., Ogston S.A., Christie A.D., Howie P.W., Patel N.B. Standards for ultrasound fetal growth velocity. Br J Obstet Gynaecol 1996; 103: 60–69
11. Owen P., Khan K.S. Fetal growth velocity in the prediction of intrauterine growth retardation in a low risk population. Br J Obstet Gynaecol 1998; 105: 536–540
12. Dudley N.J., Chapman E. The importance of quality management in fetal measurement. Ultrasound Obstet Gynecol 2002; 19(2): 190-196
13. Hynek M. Approaches for constructing age-related reference intervals and centile charts for fetal size. European journal for biomedical informatics, 2010, roč. 6, č. 1, s. 43-52. ISSN: 1801-5603
14. Silverwood R.J., Cole T.J. Statistical methods for constructing gestational age-related reference intervals and centile charts for fetal size. Ultrasound Obstet Gynecol. 2007 Jan; 29(1):6-13.
15. Bonellie S.R., Raab G.M. A comparison of different approaches for fitting centile curves to birthweight data. Stat Med. 1996 Dec 30; 15(24):2657-67.
16. Kiserud T., Johnsen S.L. Biometric assessment. Best Pract Res Clin Obstet Gynaecol. 2009 Dec; 23(6):819-31. doi: 10.1016/j.bpobgyn.2009.06.007. Epub 2009 Jul 26.
17. Altman D.G. Construction of age-related reference centiles using absolute residuals. Stat Med. 1993 12, 917-924.
18. Luecke R.H., Wosilait W.D., Young J.F. Mathematical representation of organ growth in the human embryo/fetus. International Journal of Bio-Medical Computing 1995; 39:337-347
19. Guihard-Costa A.M., Menez F., Delezoide A.L. Organ weights in human fetuses after formalin fixation: standards by gestational age and body weight. Pediatric and Developmental Pathology 2002; 5:559-578
20. Cole T.J., Freeman J.V., Preece M.A. 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
21. Naito K., Udagawa J., Otani H. Multidimensional standard curve for the development process of human fetuses. Stat Med. 2010 Sep 20; 29(21): 2235 – 45
22. Taylor J. R., 1997, An Introduction to Error Analysis: The Statistical Study of Uncertainties in Physical Measurements, University Science Books
23. Roberts C. L. and Lancaster P. A. L. Australian national birthweight percentiles by gestational age. Med J Aust 1999; 170 (3): 114-118
24. Arbuckle T.E., Wilkins R., Sherman G.J. Birth weight percentiles by gestational age in Canada. Obstet Gynecol. 1993 Jan;81(1):39-48
25. Bochicchio M.A., Vaira L., Longo A., Malvasi A., Tinelli A. Quantitative Fetal Growth Curves Comparison: A Collaborative Approach. Information Technology in Bio- and Medical Informatics. Lecture Notes in Computer Science Vol. 8649, 2014, 60-74
26. Willocks J. The use of ultrasonic cephalometry. Proc R Soc Med 1962; 55:640
27. Campbell S., Newman G.B. Growth of the fetal biparietal diameter during normal pregnancy. J Obstet Gynaecol Br Commonw 1971; 78(6): 513–9
28. Altman D.G. and Chitty L.S. Design and analysis of studies to derive charts of fetal size. Ultrasound Obstet Gynecol 1993; 3(6): 378–84
29. Royston P. and Altman D. G. Design and analysis of longitudinal studies of fetal size. Ultrasound Obstet Gynecol 1995; 6(5): 307–12
30. Goldstein H. Efficient statistical modelling of longitudinal data. Ann Hum Biol 1986; 13(2): 129–41
31. Hawkins, D. 1980. Identification of Outliers. Chapman and Hall
32. Barnett V., Lewis T., Outliers in Statistical Data. John Wiley, 1994
33. Johnson R., Applied Multivariate Statistical Analysis. Prentice Hall, 1992
34. Hodge V. J. and Austin J. A. Survey of Outlier Detection Methodologies, Artificial Intelligence Review 22, 85-126. Kluwer Academic Publisher. 2004
35. Chauvenet W.. A Manual of Spherical and Practical Astronomy V. II. 1863. Reprint of 1891. 5th ed. Dover, N.Y.: 1960. pp. 474–566
36. Dixon W.J., 1953, Processing data for outliers: Biometrics, 9(1), 74–89
37. Iglewicz, B. and Hoaglin, D. How to detect and handle outliers. ASQC Quality Press, 1993
38. Hampel, F.R. (1971) A General Qualitative Definition of Robustness. Annals of Mathematical Statistics, 42, 1887-1896
39. Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. John Wiley and Sons, New York
40. Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 6 (June 1981), 381-395
41. Aggarwal C. C. and Yu P.S. Outlier detection in High dimensional Data, ACM SIGMOD Conference, 2001
42. Knorr E. and Ng R. Finding Intensional Knowledge of Distance-Based Outlier, VLDB Conference, 1999
43. Aggarwal C. Outlier Analysis, Springer, January 2013
44. Motulsky H.M. and Brown R.E. Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate, BMC Bioinformatics 2006, 7:123
45. Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. Robust Statistics: the Approach Based on Influence Functions. New York: John Wiley and Sons; 1986
46. Hoaglin D.C., Mosteller F., Tukey J.W. Understanding Robust and Exploratory Data Analysis. New York: John Wiley and sons; 1983
47. Koza, J.R. (1990). Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems, Stanford University Computer Science Department technical report STAN-CS-90-1314
48. Altman D.G., Chitty L.S. Charts of fetal size: 1. Methodology. Br J Obstet Gynaecol 1994; 101: 29 – 34