# Characteristics of the newborn and health status: a congested two-frontier approach

Marie VANDRESSE[1]

**Abstract**

Frontier analysis is extended to evaluate the effect of characteristics of the newborn on its health status. Rather than the expectation of the health status of the newborn conditional on characteristics of the newborn and of its parents, the support of the joint distribution of the two sets of variables is estimated. The model focuses on extreme rather than average values, and highlights the negative influence of extreme characteristics of the newborn or of its parents on its health status. A case study based on Hungarian data is presented. Results show that parental ages, as well as the reproductive history of the mother, are related to the health status of the newborn.

Keywords: health status of the newborn, characteristics of the newborn, extreme values, double-frontier analysis.

---

[1] Docteur en Démographie, Institut de démographie, Université catholique de Louvain, Place Montesquieu, n°1, Bte 17, 1348 Louvain-la-Neuve, Belgium, e-mail : marie.vandresse@uclouvain.be, vm@plan.be

I am grateful to C. Gourbin, G. Masuy-Stroobant, M. Mouchart and G. Wunsch for comments leading to clarifications of the different arguments and to substantial improvements in the presentation. Financial support from the FSR of the Catholic University of Louvain (project: "Fécondité tardive et morbidité/mortalité foeto-infantile") is also acknowledged. I would like to thank Prof. Peter Jozan (HCSO) and Ms Eva Gardos M.Sc., Head of the Population, Health and Welfare Statistics Department (HCSO), for the data set and valuable information on the latter.

## 1. Introduction

Usually, when studying of the determinants of the health status of a child at birth with a structural model, a conceptual or causal diagram is defined and tested with adequate statistical models. The statistical model is based on the expectation of the variable of interest conditional on a set of explanatory variables. The interpretation of the results is based on the mean and the deviation from this mean. Extreme values are not considered. But in certain contexts and in the medical field in particular, the study of the mean values of a population of interest is of lesser interest in the sense that an individual with a normal health status does not require special care involving high financial costs for the society. Let us take one particular example related to the case study presented in this paper. Infants born with a low birth weight (<2500 g), a very low birth weight (<1500 g) or an extremely low birth weight (<1000 g) cost markedly more to the family and to the society than newborns weighing at least 2500 g As an example, Matsuo (2005) mentions that *hospital cost for babies weighing less than 1000 g at birth were 75 percent higher on average than those occurred by babies weighing 1000-1499 g* More, the cost of children with a low birth weight is not restricted to the time spent after birth at the hospital. There are long run costs due to the consequences of low birth weight on the future development of the child. From a public policy point of view, it is consequently of interest to look at the extreme values of the health status and its determinants, rather than the average values, as done among others by classical multivariate regression models. In such a context, an alternative approach focusing on extreme values is even more interesting that the proportion of low birth weight has increased over the last 30 years (Matsuo, 2005).

Furthermore, the traditional approach based on multivariate regression is confronted with problems of exogeneity. Consider a regression estimating the expectation of weight at birth with respect to duration of gestation, potential multiple births, presence of congenital anomalies and an error term including unobservable variables influencing weight, such as alcohol. If alcohol addiction, which is assumed unobservable, is a common cause of both congenital anomalies (explanatory variable) and of the weight at birth (dependent variable), congenital anomalies are not exogenous and the estimated coefficient is biased (Thomas 1997, chap. 8 and 12).

The main objective of the proposed method is not to detect the effect of each characteristic -net of all other effects- of the newborn on her or his health status. This could be done by a multivariate regression controlling for spurious associations (Vandresse, 2008). The objective here is rather to deal with extreme values and endogeneity, for reasons mentioned above. More precisely, we estimate the support of the joint distribution of the variables (smallest closed set whose complement has probability zero) instead of estimating the parameters of a conditional expectation of the variable of

interest. From the estimation of this support, we look at (more details in the next sections):

(a) the concentration of observations, accounting for distances between maximum and minimum attainable health status at birth for each level of a set of characteristics of the newborn (such as parental age or number of previous livebirths).

Considering that not all relevant characteristics of the health status at birth are observable, the frontier approach allows to make use of the available characteristics and evaluate whether each of them is influential or not (in determining the health status at birth). If all characteristics could be observed, the maximum and minimal frontiers would collapse into a unique curve. Indeed, if all determinants of the child at birth are known and observed, and assuming a deterministic process, a given vector of the determinants of the child at birth would correspond to a unique health status. As information decreases (through a lower number of observed characteristics), the distance between the two frontiers increases. The distance can be viewed as a relative measure of the importance of the observed characteristics. The closer the extreme values of the indicator of the health status are, the more influential is the considered characteristic of the newborn on his or her health status at birth. As some of the characteristics of the child at birth are unobserved, for a same level of observed characteristics, the health status might be distributed in the range $[H^*(z), H_*(z)]$ where $H^*(z)$ gives the maximum reachable indicator of health at birth with the given vector of characteristics z and $H_*(z)$ gives the minimum reachable indicator of health with the given vector of characteristics z.

(b) the congestion effect which measures a possible negative effect of a characteristic on the health status at birth. This negative effect appears only above a certain threshold. This method detects also a potential negative effect due to a too low level. The definition of “too high” or “too low” levels depends upon the variable of interest, but in all of the cases, it has the consequence that the variable, for those extreme values, has a detrimental effect on the variable of interest. Consider the effect of gestational age on the health status at birth: a gestational age less than 37 weeks or over 40 weeks has a (potential) negative effect on health status at birth.

The statistical tools to study the support of the distribution are based on the *data envelopment analysis* which is widely used in economics of production in order to evaluate the productivity of the firms. This method is not yet well-know to demographers. To the best of our knowledge, only one research dealing with *data envelopment analysis* has been published in the demographic literature (Berry *et al*., 2002 - found in POPLINE with keywords "data envelopment analysis"). The authors study the demographic

efficiency of a set of countries and justify their approach by the fact that *differences among countries at any level of development may be real, not the results of measurement error or random disturbance*, as assumed by the traditional least squares model. Their article gives consequently additional arguments to show the interests of a frontier approach in demography. Note that Berry and Kim focus their analysis on the upper frontier while the present paper considers a two frontier approach (upper and lower). The two-frontier approach has been developed by Mouchart and Vandresse (2007) in the context of a Belgian survey on the criteria of quality in the modal choice for freight transport. The motivation for developing the two-frontier approach comes from the observation that price and qualities (e.g. door-to-door transport time, flexibility, reliability) of the mode of transport are simultaneously bargained. The price is consequently not a function of the qualities of the mode of transport. The simultaneous bargaining process on price and characterized does not allow, from a statistical point of view, to consider a causal relationship satisfying the exogeneity condition. In the field of demography, the interest for the two-frontier approach should be found among others in the social and financial costs of extreme values such as late fertility, low birth weight or premature birth to be supported by the child, the parents and the society

Section 2 presents the main arguments of the model. The theoretical structure of the model is explained in section 3. Section 4 presents a case study where a health index of the newborn is evaluated with respect to a vector of characteristics of the child and its parents. The weight at birth which is considered as the most important predictor of neonatal mortality (Institute of Medicine, 1985) is selected here as health indicator; mathematically speaking, it is continuous and not upper bounded. The results take also into account the marital status and the activity of the mother, survival of the infant at one year, its gender and if singleton or coming from a multiple delivery.

## 2. Frontiers and Congestion

In order to highlight the role of extreme values, and more particularly distances between extreme values, the model estimates the support of a multivariate distribution composed of the health status at birth and a set of K characteristics of the newborn. The multivariate distribution is considered as a volume $\in R_+^{K+1}$ and is called the Potential Birth Set (PBS). The PBS is delimited by minimal and maximal frontiers. The information of interest is the distance between the two frontiers. The distance between the maximal frontier and the minimal frontier is based on the decomposition, within the space of characteristics, of the volume generated by the set of potential births. As I am more interested in health status variation for a given level of the set of characteristics of the child at birth than in characteristics variation for a given health status, the distance is computed along the axis of the health status.

Compared to an approach where the multivariate distribution would be projected into a two-dimensional space and where distances would be analyzed in this two-dimensional space, the frontier approach does not overwhelms into the two-dimensional space the volume representing a multivariate distribution. The frontier approach estimates first the envelopment surface of the volume composed of the set of observations and then estimates the distance between each observation and the envelopment surface. The distance measures the difference between one particular observation and the corresponding point located on the envelopment surface. This corresponding point is a projection of the observation on the envelopment surface. The projection is made with respect to the axis of interest; in the present case the health status axis.

In the sequel of this section, we assume that the potential births are characterized by the health status at birth and only one observed characteristic of the child at birth. This simplification allows representing the different concepts of the methodology in a two-dimensional space.

## 2.1 Concentration of observations

Let us define $\xi_p = (\eta, \zeta) \in R_+^2$ representing one potential birth where $\eta \in R_+$ is the health status at birth and $\zeta \in R_+$ a characteristic of the child at birth. The set of potential births, the Potential Birth Set (PBS), includes all potential births $\xi_p = (\eta, \zeta)$ (p=1,..,P).

The Potential Births Set and its maximum and minimum frontiers are represented in Figure 1. The maximum health indicator frontier gives the maximum health indicator a child can reach at birth for a given level of the characteristic $\zeta$. Similarly, the minimum health indicator frontier gives the minimum health indicator a child can reach at birth for a given level of the characteristic $\zeta$.

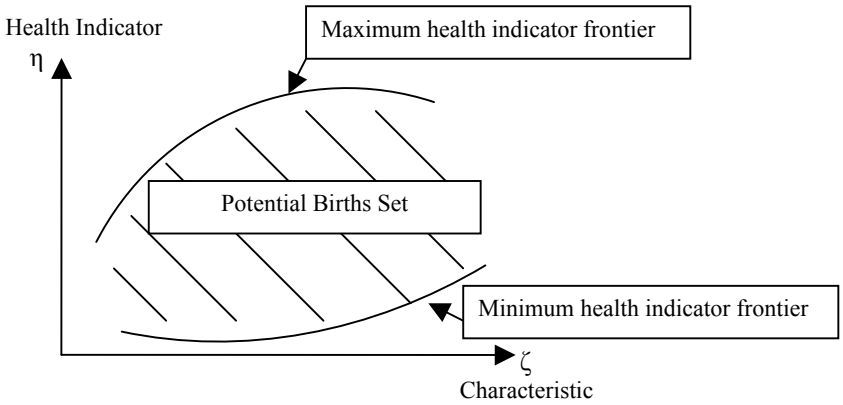


FIGURE 1: Potential Births Set and health frontiers

Let us define $\zeta_1$ and $\zeta_2$ as two different levels of the characteristic $\zeta$ with $\zeta_1 < \zeta_2$. For example, $\zeta$ is paternal age with $\zeta_1$ corresponding to 25 years of age and $\zeta_2$ corresponding to 45 of age. The concentration of observations is represented by the distance $D_j$ (with j representing the level of the characteristic) between the maximum health frontier and the minimum health frontier for a given $\zeta_j$ (Figure 2). A large distance $D_j$ means that the characteristic $\zeta$ at level j has less effect on the health status at birth. At the opposite, a short distance $D_j$ means that the characteristic $\zeta$ at level j influences markedly the level of the health status compared to larger distances. Indeed, a large distance at level j implies a large variance of the health indicator at this level; a short distance at level j implies a small variance at this level. The interpretation of a distance is always relative to the distances for other levels of the same characteristic. Note that there is one exception in the interpretation of the distances between extreme values. If the distances between extreme values are identical for all levels of the characteristic and if the level of the health index is also identical for all levels of the characteristic, this characteristic has no influence on the health indicator, even if the distance between extreme values is low.

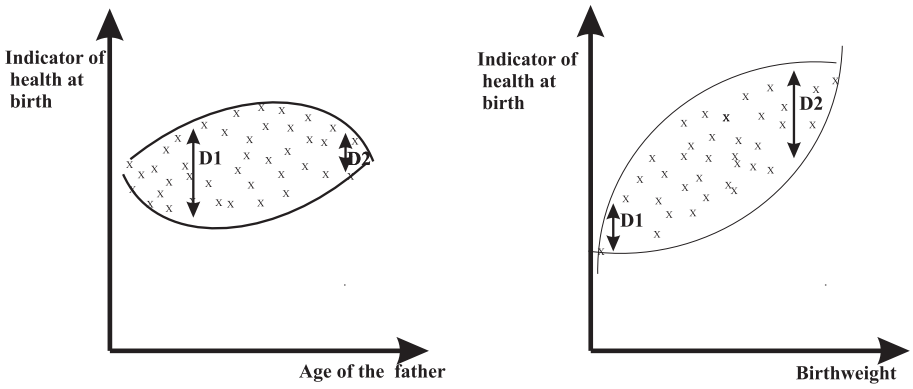


FIGURE 2 : Concentration of the observations

In Figure 2, the first graph represents the joint distribution of the health status at birth and of the paternal age. At lower levels of paternal age, paternal age influences less the health indicator than at an upper level. The distance $D_1$ between the lowest and the highest observations is large. On the contrary, $D_2$ is shorter, meaning that paternal age, at this level, has a more influential effect on the health status at birth compared to young fathers. In the second graph, low birthweight is a crucial characteristic of the health status, all observations are concentrated around a low health status (the variance of the health status is small). On the contrary, a higher weight will be reflected by a higher health status, but with less concentration of the data, which means that, at a high level of weight, other non-observed factors also influence the health status of the baby. From this point of view, the distance between extreme values can also be viewed as the relative importance of the other characteristics of the child at birth on the health status.

## 2.2 The effect of congestion

The risks of child's death or morbidity are repeatedly reported to be higher for the younger and older mothers (Cramer, 1987; Friede *et al.*,1988; Spira *et al.*,1993; Eurocat, 1995). With a frontier approach, such a case is translated into an increasing (beneficial) effect of the characteristics of the child at birth on its health status followed by a decreasing (detrimental) effect. The latter is called the congestion effect characterized by a maximum health frontier which is first increasing, and then decreasing. Let us define the health indicator at birth $I_h$ between 0 and 1 (with 1 representing the best score) and the maternal age MAge (from 15 to 50) as the only observed characteristic of the child at birth (Figure 3). Maternal age should have a positive increasing effect on the health indicator in the age group [15-30], but the effect is negative thereafter, because of congenital anomalies among others. In Figure 3, this congestion effect is represented by the decreasing part of the maximum health indicator frontier.

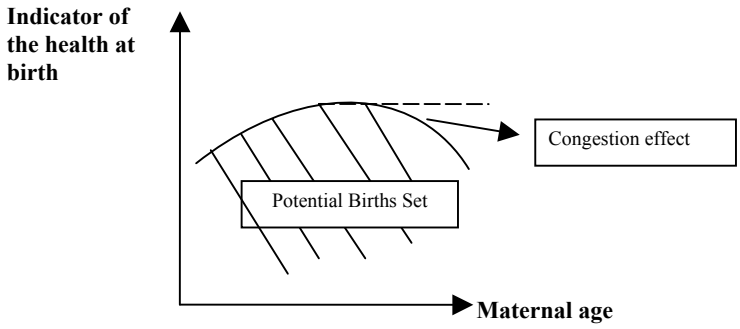


FIGURE 3: The congestion effect

## 2.3 The effect of outlying observations

The estimation of the support of a distribution is sample-sensitive. Indeed, the presence of at least one outlier in the database drastically influences the estimation of the frontiers, and consequently of the distance between the frontiers. In Figure 4(a), the database includes one outlier; in Figure 4(b), the outlier was withdrawn. By withdrawing the outlier all observations are closer to the frontier and the distance between both frontiers is shorter. The question is whether the outlier must be rejected or not from the database. It could be an encoding mistake, a voluntary erroneous response from the interviewed person or a true case which is meaningful but very different from the rest of the sample. In each of the three cases, it makes sense to delete these observations.

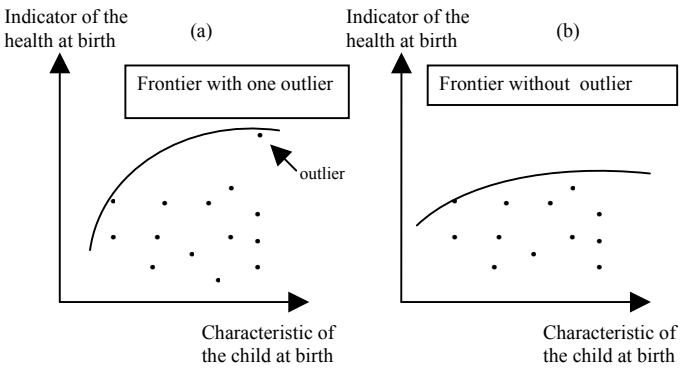


FIGURE 4: Frontier with and without outliers

Different and complementary methods are available for detecting outliers. One is to look at the empirical distribution of each variable searching for the presence of extreme values. A second method is to use statistical criteria as in Simar (2003). The idea is to estimate, for the upper (resp. lower) frontier an expected frontier by expected maximum (resp. minimum) values obtained on the basis of randomly selected sub-samples of the observations. Observations above (resp. below) the upper (resp. lower) expected frontier are considered as outliers. A third method proceeds in two steps. The first step consists in estimating frontiers with the full sample while the second step consists in estimating frontiers with the same sample excluding observations detected in the first step as being on the frontiers. If the distance computed in the first and second steps are very different, the observations detected on the frontiers in the first step are considered as outliers (Mouchart and Vandresse 2007).

## 3. The model

The theoretical framework is an extension of Data Envelopment Analysis (DEA) used in economics for the analysis of production efficiency (Färe et al., 1984; Fried et al., 1993; Simar et al. 2003).

Let us define the Potential Births Set, PBS $\in$ $R_+^{K+1}$, as the convex compact and bounded set of potentially observable births. The potential births set is composed of potential births characterized by the health status at birth, $\eta \in R_+$ and the vector of characteristics of the child at birth, $\zeta \in R_+^K$. The maximal health function, $H^*(\zeta)$, gives the maximum reachable health status with a given vector of characteristics $\zeta$. It is represented by the upper frontier of the potential births set. Similarly, the minimum health function, $H_*(\zeta)$, gives the minimum reachable health status with the vector of characteristics $\zeta$. It is represented by the lower frontier of the potential births set (Figure 1). If we observe a same concentration of the data for all levels of $\zeta$, the estimated health functions should be equidistant for all $\zeta$. On the contrary, different concentrations of the data for each level of the vector of characteristics $\zeta$ should be observed by a distance $D_j = (H^*(\zeta_j) - H_*(\zeta_j))$ between the two health functions depending on $\zeta_j$. The concentration is then represented by the vertical distance between the two health functions.

The two health functions, $H^*(\zeta_j)$ and $H_*(\zeta_j)$, represent the upper and lower frontiers of the convex closure of the set of potential births. The minimum health frontier is composed of the set of births $\xi_* = (\eta_*, \zeta)$ representing the minimal health status $\eta_*$ reached for a vector of characteristics $\zeta$. Similarly, the maximum health frontier is made of the set of births $\xi^* = (\eta^*, \zeta)$ representing the maximal health status $\eta^*$ reached for a vector of characteristics $\zeta$. Formally:

$$H_*(\zeta) = \min\{ \eta \mid (\eta, \zeta) \in \text{PHS}\} \quad (1)$$

$$H^*(\zeta)= \max\{ \eta \mid (\eta, \zeta) \in PHS\} \quad (2)$$

In practice, the estimation of the two frontiers is based on a sample of observed births characterized by the health indicator at birth and the characteristics of the child at birth, modeled as i.i.d. observations from a joint distribution. The potential births $\xi_p= (\eta, \zeta)$ are then replaced, for the estimation, by the set of observed births $x=(y,z)$. For estimating the maximum (resp. minimum) health indicator function, we define, for a given observation $x_i=(y_i,z_i)$, a factor of dilatation (resp. contraction), $\alpha_i$ (resp. $\beta_i$), that brings $y_i$ up to the frontier $H^*(z)=\{x^*=(y^*,z)\}$(resp. $H_*(z)=\{x_*=(y_*,z)\}$):

$$\alpha(y,z)= \max\{a \mid ay \leq \hat{H}^*(z)\} \quad \alpha \in [1,\infty) \text{ and } y \times \alpha(y,z) = y^* \quad (3)$$

$$\beta(y,z)= \min\{b \mid by \geq \hat{H}_*(z)\} \quad \beta \in (0,1] \text{ and } y \times \beta(y,z) = y_* \quad (4)$$

When these observations are identified empirically, the minimum and maximum health indicator functions are interpolated linearly.

For each observation $x_i$, the value of the functions $\alpha(y_i,z_i)$ and $\beta(y_i,z_i)$ can be estimated by linear programming and the distance between the minimum and maximum health functions are estimated as:

$$\begin{aligned} \hat{D}(y_i,z_i) &= \hat{H}^*(y_i,z_i) - \hat{H}_*(y_i,z_i) \\ &= \hat{\alpha}(y_i,z_i) \times y_i - \hat{\beta}(y_i,z_i) \times y_i \\ &= [\hat{\alpha}(y_i,z_i) - \hat{\beta}(y_i,z_i)] \times y_i \end{aligned} \quad (5)$$

The presence of congestion affects the maximal health function and the distance function. The congested maximal health function remains defined by Eq.(2). The difference will appear in the linear programming where the condition of monotonicity is suppressed under the congestion hypothesis.

Remark: For a same health indicator at birth $y_i$ there are as many distances as the number of combinations between the different levels of the set of characteristics of the child at birth.

## 4. Case study

Remark: The data analysis was generated using R 2.1.1 software (Ihaka and Gentleman, 1996).

4.1 The database

The database is composed of 10,215 Hungarian vital registration records of livebirths and infant deaths during one month: June 1984. We rejected 13 observations for which the weight at birth was lower than 500 g. One of these observations is an erroneous response (weight = 2 g). The other twelve observations present a weight at birth between 350 and 450 g. 500 g is the conventional limit fixed by the World Health Organization for national statistics. The final sample includes 10,202 observations (each observation corresponding to one birth) including 193 infants who did not survive their first birthday. A classical health indicator at birth is the APGAR score. However, in order to detect congestion correctly, the health indicator must be continuous and not upper bounded. As for each level of each modality the probability of observing a score of 10 is likely, a congestion effect could not be detected with such a kind of indicator. Consequently, we use the weight at birth as an indicator of the health at birth. Weight at birth is measured in grams. For the characteristics related to the newborn, six variables have been selected on the basis of known risk factors: duration of gestation (Gest), maternal age (MAge) and paternal age (PAge), number of previous livebirths (LB), number of previous (early medium and late) foetal deaths (FD) and number of previous induced abortions (IA). Figure 5 gives the histogram for each characteristic at birth and for the weight at birth.

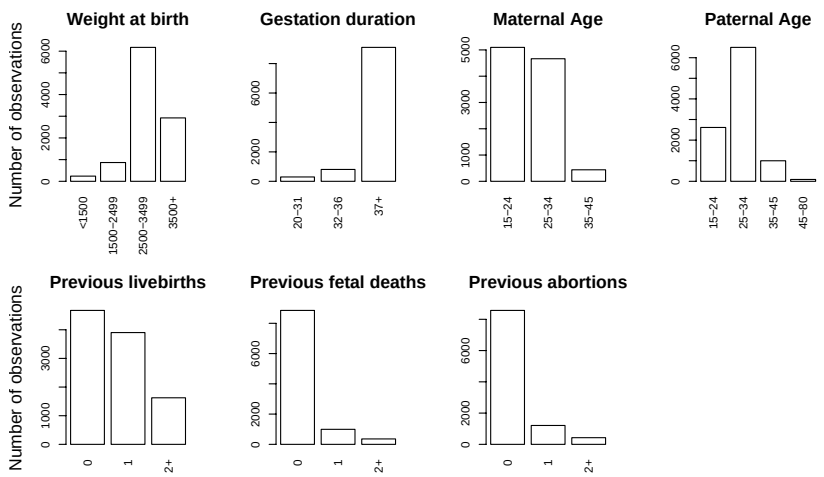


FIGURE 5: Histograms of the variables (data: Hungary 1984-1988)

## 4.2 Concentration and congestion

### 4.2.1 The concentration model

Preliminary remark: in a two-dimensional space, there is one unique distance for each level of the characteristic. In higher dimensional space, there are as many distances as the number of combinations between all levels of all characteristics. In order to interpret the distances for each level of each characteristic, the distributions of the distances for each level of a given characteristic are compared.

The results are presented by means of empirical densities of the different values of interest, with respect to different levels of each characteristic of the child at birth. For parental age, we defined four levels for paternal age (15-24, 25-34, 35-44, 45-80) and three levels for maternal age (15-24, 25-34, 35-44). These classifications are generally those found in the epidemiologic and demographic studies evaluating the effect of parental age on infant morbidity and mortality. For the number of previous foetal deaths, of previous induced abortions and of previous livebirths, we define three levels: 0, 1, 2+. In developed countries, where the total number of pregnancies per women is low, on average, it seems reasonable to make such a classification. Finally, the duration of gestation (in weeks) is divided in three main categories (<32, 32-36, 37+). The distances, measured in the same unit as the weight variables (grams), have been estimated by taking simultaneously the six characteristics into account and are a function in the seven-dimensional space (six characteristics and weight). For each level of each characteristic, the number of estimated distances corresponds to the number of possible combinations between each value of the other five characteristics. This allows us to detect a potential individual effect of the characteristic whatever the level of the other characteristics. This also means that we do not have a unique distance for one level of a characteristics as in the two dimensional case presented in section 2.

*Duration of gestation*. Figure 6 presents the bivariate plot of the distance and the duration of gestation (graph (a)), of the empirical densities of birthweight (graph (b)) and of the empirical densities of the distances with respect to the duration of gestation (graph (c)). Figure 6(b) shows the tendency for low birth weight in case of gestational age less than 37 weeks. The average weight is 1,269 g for a 20 to 31 weeks duration of gestation and 2,489 g for a 32 to 36 weeks duration of gestation. In addition, Figure 6(c) shows that the concentration around the weight is denser for lower gestational age. The average distance is 2,190 for duration of gestation between 20 and 31 weeks, 2,130 for gestation between 32 and 36 weeks and 3,497 for a gestational age over 36 weeks. Both results show the importance of gestational age in the determination of the weight: a higher concentration around a lower weight for duration of gestation lower than 37 weeks. The effect is higher for duration of gestation less than 31 weeks.

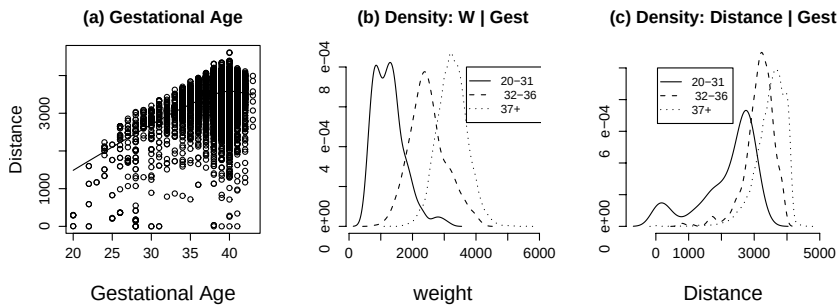


FIGURE 6: Concentration with respect to gestational age

*Parental Age.* Figure 7 presents empirical densities of weight (W) at birth and of distances with respect to paternal and maternal ages (PAge and MAge respectively). It also includes bivariate plots of distances and parental age. For each category of maternal age, the distances between the upper and lower observations are relatively high. The distances increase until 30 years old, then decrease (Figure 7(a)). Looking at the empirical densities of the distances with respect to categories of maternal age (Figure 7(e)), the density for high values is lower for categories 15-24 and 35-44 than for category 25-34. This suggests a higher concentration of the data for the lower and upper categories. Consequently, the effect of being a young (<25) or old (>34) mother not only decreases the weight of the child (Figure 7(c)), but also increases the density around the lower weight. Maternal age seems to be more influential at a lower or upper age. A similar reasoning is done for paternal age (Figures 7(b), (d), (f)). However, the effect of the category 45-80 seems to be quite important on the decrease of the distance (the concentration is even denser). Maternal age and paternal age are highly correlated but it has already been stressed that graphs represent the distribution of the distances conditional on selected characteristics and that distances include simultaneous information on all characteristics. An individual effect is consequently detected by interpreting the different figures, namely comparing the conditional empirical densities of the distance for each category of the selected characteristic.

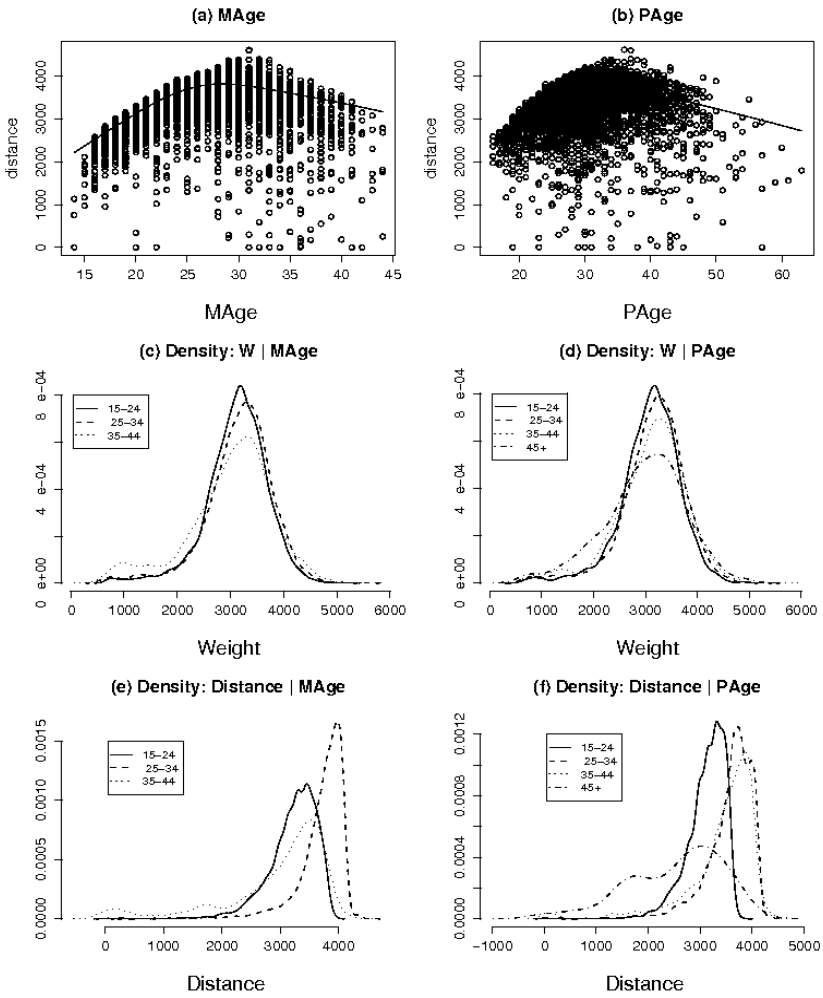


FIGURE 7: Concentration with respect to parental age

*Number of previous livebirths (LB), foetal deaths (FD) and induced abortions (IA).* The number of previous foetal deaths and of previous induced abortions present a similar pattern. The graphs corresponding to these two variables indicate a direct decreasing distance with the number of previous foetal deaths and of previous induced abortions, controlling for the other variables (Figures 8(b) and (c)). The more the number of events is important, the more the characteristic influences the weight level, particularly for events appearing twice or more. Indeed, as shown in Figures 8(h) and (i), the estimated distances are lower for the third category (2+). In Figures 7(e)

and (f), the empirical densities show a decreasing effect of those two variables on the weight. Consequently, the analysis of concentration not only shows that weight is affected by the number of previous foetal deaths or induced abortions, but also that the density around lower weight increases with these events. For the number of previous livebirths, we observe a similar pattern as for parental age. The increasing distance (until 2 livebirths) is followed by a decreasing distance (Figures 8(a) and (g)). Similarly, weight increases from 1 to 2 previous livebirths, then decreases 8(d)). These facts indicate that the extreme values of this characteristic decrease weight at birth and increase the density around this lower weight.

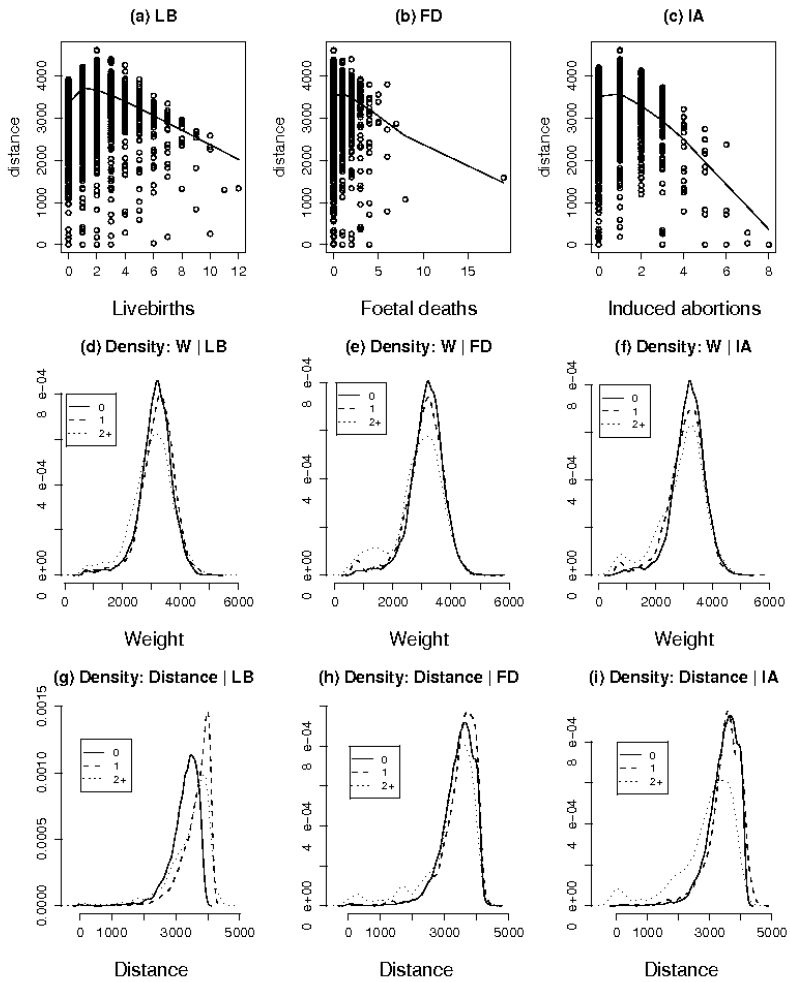


FIGURE 8: Concentration with respect to maternal reproductive history

4.2.2 The congested concentration model

*Duration of gestation*. Figure 9 shows the sensitivity of the gestational age to a congestion effect. Although the effect of gestation on the factor of dilatation (Figure 9(a)) is relatively weak, consequences on the distances are noteworthy (Figure 9(b)). The median of the factor of dilation is 0.69 without congestion and 0.70 with congestion but the median of the distances is 3,619 without congestion and 2,736 with congestion, thus a difference of 883 g.

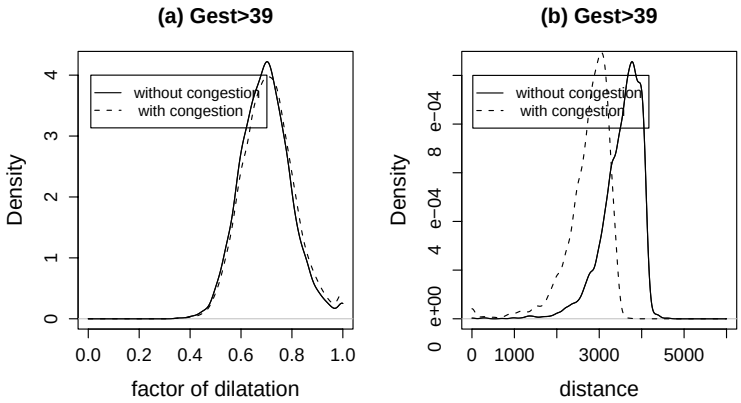


FIGURE 9: Congestion vs non congestion: gestational age

*Parental age*. Figure 10 compares empirical densities of distances and the inverse of the factor of dilation (upper H, Eq.(3)) one without congestion and one with it. Figures 10(a) and (b) represent results for observations characterized by mothers older than 30 years. Figures 10 (c) and (d) consider only fathers older than 38 years. Indeed, a maternal age of 30 years and a paternal age of 38 years have been detected as the starting points of a potential congestion effect.

Presence of maternal congestion should be reflected by smaller factors of dilatation. In order to belong to interval ]0,1], factors of dilatation were replaced by their inverse. Maternal congestion can consequently be observed in Figure 10(a) by a shift of the empirical density of the inverses of the dilatation factors to the right when congestion is taken into account. The median of the inverses of dilatation factor varies from 0.68 without congestion to 0.70 with congestion. The presence of congestion decreases the distances to the upper frontier, and consequently between the two frontiers. This is reflected by a shift of the empirical density of the distances to the left when congestion enters into the model (Figure 10(b)). For the category > 30

years old, the median of the distances without congestion is 3,635 g and 2,676 g when including congestion, i.e. a difference of 959 g.

Concerning paternal age, the effect of congestion seems more important. Figure 10(c) shows that inverses of the factor of dilatation increase for the category [> 38 years old] when congestion is taken in account (shift of the empirical density to the right). The median varies from 0.68 without congestion to 0.75 with congestion. As a consequence, the median of distances, for the category [> 38 years old], decreases from 3,398 g without congestion to 2,242 g with congestion, a difference of 1,156 g. The effect on the distances pushes the empirical density to the left (10(d)).

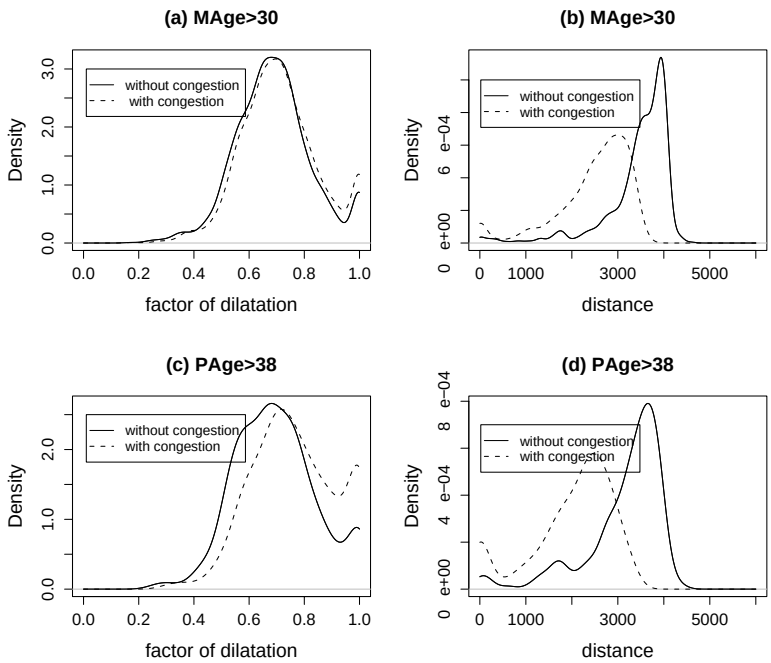


FIGURE 10 : Congestion vs non congestion: parental age

*Number of previous livebirths, foetal deaths and induced abortions.* Figure 11 presents the results of congestion with respect to the number of previous livebirths, the number of previous foetal deaths and the number of previous induced abortions. For each variable, the empirical density of the distances and of the inverses of the factor of dilatations are represented, one with congestion, one without. Concerning previous foetal deaths the median of the distances decreases, for the category "1+", from 3,555 g to 2,743 g (a difference of 812 g). For the number of previous livebirths, the difference between the median of both distances is 828 g (2,767 g without congestion; 1,939 g with congestion). Finally, for the number of previous induced

abortions, the difference between the medians, for the category "1+", is 804 g (3,492 g without congestion; 2,688 g with congestion).

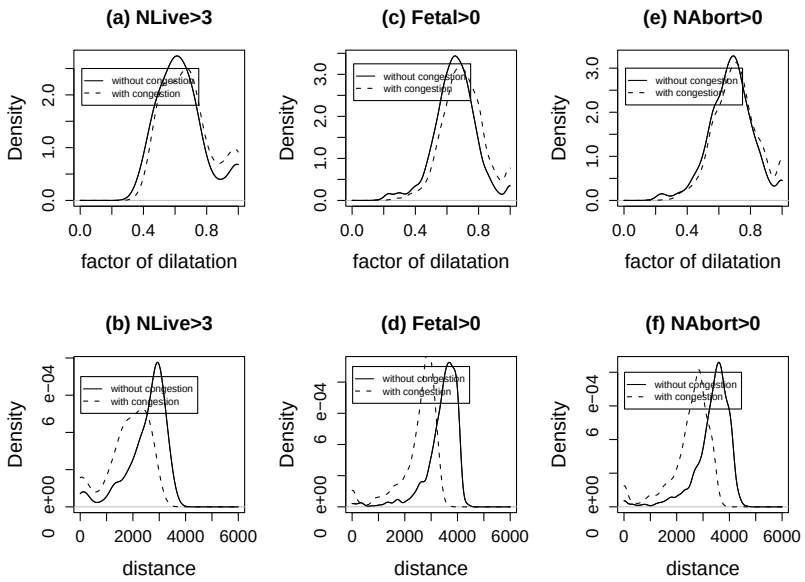


FIGURE 11: Congestion vs non congestion: maternal reproductive history

The presence of congestion might be due to a lack of observations for certain categories of the variables, and for extreme categories in particular. Congestion might be upward biased.

4.3 Effect of outlying data

Considering the low frequency of some events (Table 1), it seems useful to determine the validity of those observations and their effect on the estimation, in the spirit of section 2.3. Observations over six induced abortions or over four previous foetal deaths or over seven previous livebirths were extracted from the data.

TABLE 1: Reproductive history: outlying observations
(data: Hungary June 1984)

| | Total number of previous events | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 19 |
| **Livebirths** | 4674 | 3902 | 1069 | 298 | 125 | 50 | 35 | 23 | 13 | 6 | 4 | 1 | 1 | 0 |
| **Induced abortions** | 8568 | 1207 | 299 | 85 | 19 | 13 | 5 | 3 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Foetal deaths** | 8850 | 993 | 265 | 57 | 22 | 5 | 5 | 2 | 1 | 0 | 0 | 0 | 0 | 1 |

Concerning observations with seven or more previous livebirths they are characterized by an average maternal age of 35 years (for an average of 25 years in the complete sample). Education of both parents is also relatively low in this case for all these observations (no secondary degree). These situations are also characterized by low prenatal care. These facts suggest that observations associated with a lot of previous livebirths are not erroneous (encoding mistake, voluntary untrue answer...) but apparently linked to an unfavorable social context.

Many induced abortions are also associated with older mothers, compared to the complete sample. These observations are also characterized by higher number of previous livebirths. This could reflect the use of induced abortion as a contraceptive method, and consequently not as an erroneous response. Following the same reasoning, we do not find any indication for considering observations with many foetal deaths as erroneous responses (except the case with 19 previous feotal deaths). Many of these events are again linked to older maternal age and we observe a comparatively lower weight at birth for these observations. This suggests that these observations represent mothers with less efficient reproductive capacity.

As already mentioned, outlying data do not include only erroneous data. They also consist of particular situations, less frequently observed but still meaningful. The distances are recomputed without these outlying data. The sub-sample includes 10,153 observations. Results are summarized in Table A.1. For making comparison easier, Table A.1 includes also the statistics for the complete sample. No difference appears between the estimations based on the two different samples, which is not surprising. As we are interested in the variation of the weight with respect to the characteristics of the child, the factors of contraction and dilatation and the distances between frontiers depend on the weight distribution, and not on the distribution of the characteristics of the child. The deletion of the observations does not alter the weight distribution, and the two samples are similar. When events are numerous in the reproductive history of the mother they do not appear as influential.

### 4.4 Biological and socio-economic heterogeneity

As underlined in Mosley *et al.* (1984), biological and socio-economic characteristics do not influence the newborn's health status in the same way. Biological characteristics are defined as proximate determinants in the sense that they have a direct effect on the variable of interest, while socio-economic characteristics have an indirect effect. Several socio-economic variables and some biological variables are qualitative and less immediately included into the model. A possibility consists of evaluating the distance $D_j$ within sub-samples, these sub-samples being defined by socio-economic variables. For example, we separate the sample according to the status of the mother, and then compare the results obtained in each sub-sample with those from the complete sample. By creating sub-samples, we decrease the sample size. In order to simplify the presentation of the article, graphical representations of the results are presented in Appendix B.

#### 4.4.1 Biological heterogeneity

Gender and multiple births are factors of weight differentiation. A strong relation is also observed between weight at birth and infant mortality (Table 2 based on the Hungarian database). Median and average weight at birth are different according to the selected sample. Sub-samples characterized either by single births, by survivals at one year or by male births present an higher average and a higher median than the other sub-samples characterized either by multiple births, by deaths within the first year, or by female births. The difference according to gender seems however less marked. As the sub-samples present different weight distributions, the distances measured in weight between minimal and maximal health functions can accordingly be different. As mentioned previously, discrete variables such as gender or survival at one year cannot be introduced directly into the model. Consequently, discrete variables are integrated into the model using sub-samples. The distances obtained with the complete sample are then compared with six sub-samples: girls, boys; single birth vs. multiple births, infant deaths and infants surviving their first birthday.

TABLE 2: Biological heterogeneity: descriptive statistics
(data: Hungary June 1984)

| Weight at birth (g) | Complete sample | Boys | Girls | Surving 1st birthyear | Infant deaths | Single birth | Multiple birth |
|---|---|---|---|---|---|---|---|
| **Median** | 3,200 | 3,250 | 3,100 | 3,200 | 1,700 | 3,200 | 2,300 |
| **Mean** | 3,135 | 3,203 | 3,062 | 3,159 | 1,883 | 3,157 | 2,178 |
| **Min** | 530 | 550 | 530 | 530 | 530 | 530 | 600 |
| **Max** | 5,600 | 5,600 | 5,200 | 5,600 | 4,370 | 5,600 | 3,600 |
| **St. dev.** | 615 | 628 | 593 | 580 | 976 | 596 | 663 |
| **Num. obs*** | 10,202 | 5,262 | 4,937 | 10,009 | 193 | 9974 | 228 |

*Num. obs. = number of observations in the selected sample
Medians and means between each sub-sample are statistically different

*Infant mortality.* Figure B.1 shows the empirical densities of the distances between the maximum and minimum weights, of the factors of contraction for the minimum health function, and of the factors of dilatation for the maximum health function. For those three elements, the empirical densities are computed separately for the complete sample, for the sub-sample including only infants who survive the first year of life, and for the sub-sample including infants who did not survive the first year of life. These figures indicate that the distribution of distances is more concentrated (the median of the distance is 2,928 for the sample including only infants who did not survive their first year and 4,119 for those who survived) and the weights are closer (resp. more distant) to the minimum (resp. max) health functions. This is especially observed for the median of the factors of contraction (resp. dilatation) for the sample including infants who do not survive (0.51) (resp. 0.50) which is greater (resp. lower) than the one obtained from the sample including infants who did survive (0.24). Infant mortality is then associated with low birth weight. The concentration model shows a higher density around this lower median weight. As the estimation of the distances includes the other five biological characteristics of the child at birth, the association is observed whatever the level of those characteristics. The differences between the complete sample and the sample including surviving infants only are similar. Their empirical densities are identical and overlap in the figures.

*Multiple Birth*. Figure B.2 presents a similar pattern as infant mortality: differences are observed when estimating the distances on sub-samples. The distances between the maximum and minimum health functions are shorter in the case of multiple births (a median of 2,053 for multiple births and of 4,023 for single birth): for multiple births, weights at birth have a lower mean (Table 2) and are more concentrated around this lower mean. Multiple births are consequently a factor of health discrimination, whatever the level of the other child's characteristics at birth.

*Gender.* The distances between maximum and minimum health functions do not show particular differences for boys and girls. The median of the distances is 3,767 for girls and 3,969 for boys. Table 2 does not indicate a marked difference in weight at birth according to gender. These two facts suggest that weight at birth is not discriminated by gender.

4.4.2 Socio-economic heterogeneity

Socio-economic variables, such as parental education, marital status or income are often considered as risk markers of low birth weight. The Hungarian data base includes maternal activity (housewife, active, or other) and the marital status of the mother (single, married, divorced or widowed). Table 3 shows the basic summary statistics of weight at birth with respect to these two socio-economic indicators. Children of single women present, on average, a lower weight at birth than those of married women. The majority of the complete sample is made of married women (97% of the observations). Concerning maternal activity, birthweight is, on average, lower when the mother is a housewife (housewives represent 8 % of the complete sample).

TABLE 3: Biological heterogeneity: descriptive statistics
(data: Hungary June 1984)

| **Weight at birth (g)** | **Complete sample** | **Single** | **Married** | **Housewife** | **Active and other**** |
|---|---|---|---|---|---|
| **Median** | 3,200 | 3,000 | 3,200 | 3,000 | 3,200 |
| **Mean** | 3,135 | 2,903 | 3,140 | 2,952 | 3,152 |
| **Min** | 530 | 700 | 550 | 600 | 530 |
| **Max** | 5,600 | 4,600 | 5,600 | 4,700 | 560 |
| **St. dev.** | 615 | 685 | 611 | 641 | 610 |
| **Num. obs*** | 10,202 | 136 | 9,919 | 903 | 9,299 |

*Num. obs. = number of observations in the selected sample
**: other= student, retired.
Medians and means between each sub-sample are statistically different

*Marital status* The empirical densities of the distances presented in Figure B.4 indicate that children from single women present particular characteristics: weights at birth are, on average, not only lower (Table 3) but also more concentrated around the lower mean than the weights associated with children from married women (the median of the distances is 882 g in the case of single women, 2,897 g in the case of married women and 4,117 g for the complete sample). These observations suggest that being single is a risk marker of low birthweight (or of low health status at birth).

*Mother's activity*. The empirical densities of the distances proposed in Figure B.5 give supplementary indications for interpreting the activity of the mother as a discriminant factor for the weight at birth (or for the health status of the

newborn). Table 3 not only indicates a lower median birthweight associated with being a housewife (3,000 g) compared to active mothers (3,200 g) but the empirical density in Figure B.5 shows that the distances between the minimal and maximal observation are shorter (a median of 1,895 g for housewife and of 2,896 g for active mothers). Being a housewife in this sample has a particularly detrimental effect on the weight of the newborn.

## 5. Conclusion

We computed the distances between extreme observations in order to detect the importance of each level of characteristics on the health indicator. The model of concentration has shown that the distances between the frontiers are relatively important for each characteristic of the newborn, meaning that a unique characteristic cannot determine alone the level of health. For extreme values, the median health status represented by the weight at birth is lower, and observations are more concentrated around this lower median. The extreme values of a characteristic are detrimental for the health status. Concentration can consequently indicate the intensity of the relationship between the child's characteristics and its health status at birth. For parental age and the number of previous livebirths, the distances between the frontiers are first increasing, then decreasing (from 30 years old for the mother, 38 years old for the father and from 2 livebirths onwards). The decreasing effect is more marked for paternal age than for maternal age. The more the number of previous foetal deaths, the higher the concentration of the weight at birth. The median of the weight at birth decreases immediately too. We found that paternal age and the number of previous foetal deaths are the most sensitive characteristics for congestion. Biological heterogeneity shows that multiple births and infant mortality are characterized by a higher density around a lower weight at birth. Sex at birth does not appear as a factor of discrimination. Offspring of single women or of housewives have a lower median weight at birth, with a greater density around this lower median.

This study does not present new or opposite results to those generally reported in the literature on infant morbidity, but this was not the main objective of the research. As already mentioned in the introduction of the chapter, the main objective was to develop a methodology which is in accordance with the problem of interest: the endogeneity of the characteristics of the newborn and the interest for extreme values rather than average values. In this perspective, the results of this study are encouraging for further developments. The absence of contradictory results compared to previous results found in the literature and based on conditional expectations might be viewed as a sort of test of *validity* for this methodology.
The model has some drawbacks. First of all, in the case of the Hungarian data, the sample size is particularly important but for extreme values of the characteristics, the number of observations is markedly lower. Father having their offspring at 45 years or older are far less numerous than those aged 25 to 35. Similarly, it is more likely to observe a woman having none or only

one previous foetal death rather than three. This fact (fewer observations for extreme values) might be part of the explanation for the congestion effect. Congestion might be upward biased. The problem of the number of observations is not proper to the frontier approach. In a classical multivariate regression, a too small number of observations leads to less precision in the estimations and would increase the standard error of the regression parameters. Confidence intervals could be obtained using a bootstrap methodology adapted for nonparametric frontier approaches (Simar *et al*., 1998). However, application of a bootstrap method in this context is not yet adequate and provides wide confidence intervals (Kneip et al., 1998). Secondly, this model has been constructed without a conceptual framework where causal relations are clearly identified; therefore, results should be interpreted in terms of association rather than causality.

As a final comment, and from a methodological point of view, we would like to point out the interest of the frontier approach. Compared to classical statistical methods, and multivariate regression in particular, the frontier approach evaluates all observations with respect to extreme observations while in classical multivariate regression, observations are evaluated with respect to average values. Interest for extreme values in the present context has been highlighted in the introduction. A study of the congestion effect would also not be feasible through classical multivariate regression except if theoretical knowledge suggests introducing non linear effects. This is confirmed, in the present context, by comparing the analysis of the OLS regression of weight at birth conditional to parental age, considering parental age as a continuous variable or as categorical variable. In the first case, congestion is not detected (a linear effect is assumed) while in the second case congestion is detected by comparing the values of the parameters of the regression associated to each category. Multivariate regression deals more with confirmatory analysis while the frontier approach has to be viewed as an exploratory analysis. In the latter case, there is no assumption on the type of relation between variables of interest. Both approaches are complementary to each other.

## References

BERRY, B. J. AND K. HEJA (2002), "Demographic efficiency: Concept and estimation", *Population and Environment*, 23, 3.

CRAMER, J. (1987), "Social factors and infant mortality: identifying high-risk groups and proximate causes". *Demography*, 24 (3), 299–322.

EUROCAT (1995), *Report 6 - Surveillance of congenital anomalies in Europe*, 1980-1992, Institute of Hygiene and Epidemiololgy, Brussels.

FÄRE, R., S. GROSSKOPF, AND C. LOVELL (1984), *The Measurement of Efficiencies of Production*, Kluwer-Nihoff Publishing, Boston.

FRIED, H., C. LOVELL, AND S. SCHMIDT (1993), *The Measurement of Productive Efficiency: Techniques and Applications*, Oxford University Press, New York.

FRIEDE, A., W. BALDWAIN, P. RHODES, J. BUEHLER, AND L. STRAUSS (1988), "Older maternal age and infant mortality in the United States". *Obstetrics and Gynecology*, 72 (2), 152–157.

IHAKA, R., AND GENTLEMAN, R. (1996), R: A Language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, 5(3), 299-314.

INSTITUTE OF MEDICINE (1985), P*reventing low birthweight*, National Academy Press Washington, D.C.

KNEIP, A., B. PARK, AND L. SIMAR (1998), "A note on the convergence of nonparametric DEA estimators for production efficiency scores", *Econometric Theory,* 14, 783–793.

MATSUO, H. (2005), The health consequences of low birth weight: literature review and critique, Document *de travail du SPED*, n°25, Département des Sciences de la Population et du Développement, Université catholique de Louvain, Louvain-la-Neuve.

MOSLEY, H. W. AND L. C. CHEN (1984), "An analytical framework for the study of child survival in developing countries", *Population and Development Review,* Supplements*: Child Survival: Strategies for Research*, 10, 25–45.

MOUCHART, M. AND M. VANDRESSE (2007), "Bargaining powers and market segmentaiton in freight transport", *Journal of Applied Econometrics,* 22(7), 1295-1313.

SIMAR, L. (2003), "Detecting outliers in frontiers models: A simple approach", *Journal of Productivity Analysis*, 20, 391–424.

SIMAR, L. AND P. WILSON (1998), "Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models", *Management Science*, 44, 49–61.

SIMAR, L. AND P. WILSON (2003), *Efficiency Analysis: The Statistical Approach*, Lecture Notes, Institut de statistique, Université catholique de Louvain, Louvain-la-Neuve.

SPIRA, A. ET AL. (1993), "Conception probability and pregnancy outcome in relation to age, cycle regularity, and timing of intercourse", In R. Gray, H. Leridon, and A. Spira (Eds.), *Biomedical and Demographic Determinants of Reproduction*, Clarendon Press, Oxford, UK, 271-274.

THOMAS, R. (1997), *Modern Econometrics*, *An Introduction*. Addison-Wesley, Chichester, England.

VANDRESSE, M. (2008), Estimation of a structural model of the determinants of neonatal mortality in Hungary, 1984-88 and 1994-88, *Population Studies*, 62(1), 85-111.

Appendix:

A Outlying observations: comparative statistics

TABLE A.1: Estimation of distances: comparative statistics with and without outliers, Hungary - June 1984

| | **Complete sample** | | | | | **Sample without outliers** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **mean** | **median** | **sr.dev** | **min** | **max** | **mean** | **median** | **sr.dev** | **min** | **Max** |
| **without congestion** | | | | | | | | | | |
| **B (lower frontier)** | 0.42 | 0.39 | 0.13 | 0.17 | 1 | 0.42 | 0.39 | 0.13 | 0.17 | 1 |
| **1/α (upper frontier)** | 0.66 | 0.66 | 0.11 | 0.17 | 1 | 0.66 | 0.66 | 0.13 | 0.17 | 1 |
| **Distance** | 3432 | 3542 | 534 | 0 | 4614 | 3425 | 3515 | 449 | 0 | 4613 |
| **with congestion** | | | | | | | | | | |
| **B (lower frontier)** | 0.59 | 0.56 | 0.13 | 0.25 | 1 | 0.59 | 0.57 | 0.14 | 0.25 | 1 |
| **1/α (upper frontier)** | 0.68 | 0.68 | 0.11 | 0.18 | 1 | 0.69 | 0.68 | 0.13 | 0.18 | 1 |
| **Distance** | 2749 | 2895 | 607 | 0 | 3730 | 2711 | 2865 | 597 | 0 | 3730 |

**β : factor of contraction ; α : factor of contraction**

B Biological and socio-economic heterogeneity: graphical results

B.1 Biological heterogeneity

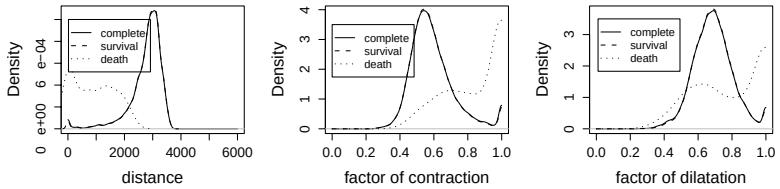


FIGURE B.1: Heterogeneity: survival status at 1 year (Hungary – June 1984)

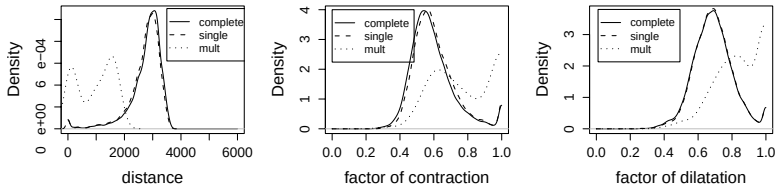


FIGURE B.2: Heterogeneity: multiple births (Hungary – June 1984)

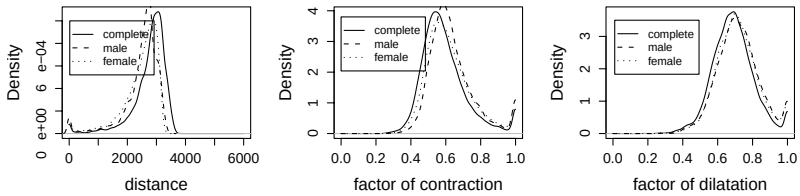


FIGURE B.3: Heterogeneity: gender (Hungary – June 1984)

B.2 Socio-economic heterogeneity

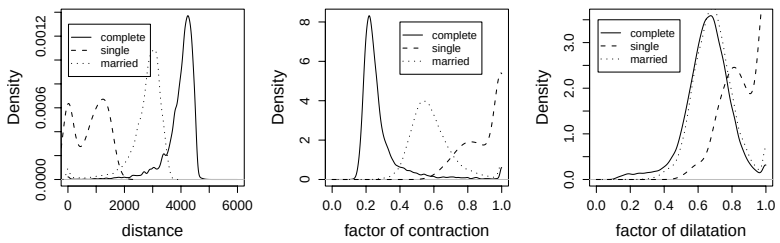


FIGURE B.4: Heterogeneity: marital status (Hungary – June 1984)

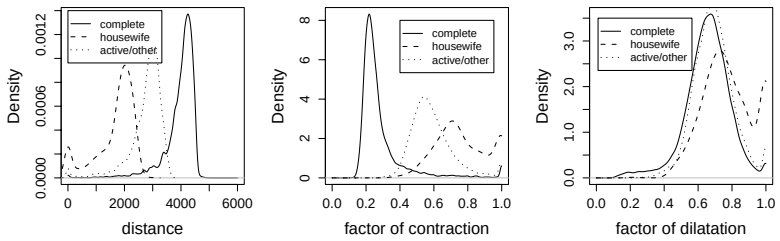


FIGURE B.5: Heterogeneity: mother's activity (Hungary – June 1984)