Modelling Fertility:

A Semi-Parametric Approach

Walter Oberhofer and Thomas Reichsthaler

June 2004

Discussion Paper of the University of Regensburg Vol. 396

Modelling Fertility: A Semi-Parametric Approach

ABSTRACT

This article presents a categorical model of fertility based on the statistical theory of the Generalised Linear Model (GLM). Focussing on the individual probability of giving birth to a child, we derive distributions which can be embedded in a GLM framework. A major advance of that methodology is the knowledge of the distribution of the random variable, which leads to a Maximum Likelihood estimation procedure.

The approach takes into account the smooth shapes of parameter development over the age of the mother as well as over time. The estimation of this semi-parametric approach is done using the Local-Likelihood-method. The presented method provides stable results of the fertility, especially for smaller populations. This is illustrated by using a data set which consists of less than 100,000 inhabitants.

Adress:

Chair for Econometrics,

Department of Economics

University of Regensburg

93040 Regernsburg

Tel: +49-941-943-2736

Fax: +49-941-943-4917

Modelling Fertility: A Semi-Parametric Approach

INTRODUCTION

Finding areas in today's economic life which are not influenced crucially by topics from demography is hard, perhaps non-existent. This is why decision makers of economy, politics and administration have a considerable demand for demographic analysis. This might be on aspects of pension systems, health care policy, unemployment problems or other fields of economic and social interest.

Demographers are usually asked for quantitative analysis about certain compositions of a historical, recent or future population. Fertility, as one of the three major demographic components (Cannan 1895), has a remarkable impact on these types of questions. When it comes to the subject of long term structural changes, the analysis of childbearing behaviour plays a critical role in this context. Considering the impact function of demographic analysis for policy aspects, structural changes of fertility over time are often the focus in most cases.

Due to the considerably strong influence of different social, economical or ethnical aspects on this demographic subject, the directions of explaining fertility patterns have become highly multi-dimensional.

Since most of the extensive scientific literature deals with theoretical explanations of different fertility structures, there is, considering the extraordinary practical relevance of the problem, a remarkable gap between theory and the possibilities of practical applications. While researchers usually do not take into consideration the existing data restrictions of statistical sources, the interchange between these two parties is sometimes not as intensive as it should be.

The goal of our contribution is therefore to set down a conceptual statistical model of the childbearing behaviour of a population, using generally available data about the mother's age and over different census dates. Our approach restricts itself deliberately to data which are available in nearly all administration systems, at least in local administrative districts in developed countries. Less developed countries usually have administrative systems which can provide appropriate data for our approach although the aggregational level is usually higher. Aspects of changing behavioural patterns play a crucial role in many demographic analyses. We have therefore developed a model which derives information about fertility characteristics over time and particularly for smaller populations where random influences usually overlap inherent structural changes.

This sort of demographic analysis is not new. There are huge numbers of theoretical as well as data driven contributions that claim to deal with exactly these problems.

Usually, childbearing behaviour is modelled for a fixed census year over the age of the mother using a parametric function. Therefore, the age specific relative frequencies of a sample are estimated for every census date parametrically. In a second step the results of these estimations have to be analysed over time. Our contribution expands that approach in two directions.

First, we develop a model which does not restrict itself on a heuristic formulation of relative fertility frequencies but provides an easy and consistent closed stochastic approach to describe probabilities of childbearing dependent on time and age of the mother for a given population. This approach is embedded into the theory of Generalised Linear Models (GLM). It is shown that a large number of contributions in recent demographic literature can be nested into that methodology. Further, a new functional description of the fertility curve is suggested.

Secondly, using a varying-coefficient approach with simultaneous estimation over mother's age and time we offset consistency problems which usually arise from the separation of these two dimensions. We choose the technique of Local Likelihood estimation.

As a result of our efforts we develop an approach which:

- models age specific fertility based on a theoretically derived distribution function
- fulfils the requirements for these probabilities regarding development over time and over mother's age
- data requirements can be met by official statistics nearly everywhere, and
- provides results that can used to describe structural changes in fertility patterns over time, especially for smaller populations which usually exhibit a large influence of random noise effects.

This leads to the following structure of the article. In the next section we embed the fertility decision into a categorical stochastic model. This leads directly to the theory of GLM. The modelling of fertility rates is done parametrically over mother's age and non-parametrically over time.

The third section deals with the estimation method based on the Local Likelihood principle.

We then apply the method on two data sets where the German population is used as an example of a big population (60 million, as we restrict ourselves to the area of west Germany before reunification). The city of Bolzano in Italy is used as an example of a small population (less than 100,000 inhabitants). As modelling fertility is used in many cases as input for population projections, we address some problems in following this approach. Although fashionable for decades, we are suspicious of whether time series models are an appropriate method to deal with these problems.

Finally, we conclude our findings and address some further research directions. In the appendix we deal with the mathematical derivation of the M.L. function.

A CATEGORICAL MODEL OF FERTILITY

We consider a woman of childbearing age at the beginning of a certain year. Let p be the probability of a woman giving birth to a child during that year. From a theoretical point of

view, we exclude the possibility of the bearing of twins and also the fact of becoming pregnant twice a year. Both cases can be taken into account on the practical side via a rise of the probability.

The situation can be described statistically introducing an indicator variable for every woman with the value 1 if the woman has a child during this year, and 0 otherwise; that means the indicator is a Bernoulli variable with probability *p* for the value 1 and (1-*p*) for the value 0. *p* depends upon time and a number of individual characteristics, which are denoted by the vector *z: p(z).*

If there are *N* women of childbearing age, they can be numerated from 1 to *N*. In this way we can assign the vector $\mathbf{z}_j$ and the indicator S_j for each woman *j*,$1 \le j \le N$. Now we take a closer look at the vector $\mathbf{S} = (S_1, S_2, \ldots, S_N)$ of all indicator variables. Assuming independence between the fertility probabilities, the distribution of *S* can be written as

$$\mathrm{P}(\mathbf{S} = (n_1, n_2, \cdots n_N)) = \prod_{j=1}^{N} p(z_j)^{n_j} (1 - p(z_j))^{1-n_j}, \qquad (1)$$

where $n_j \in \{0,1\}$.

Note that every woman can be identified through her position in *S* which is essential in the classical categorical regression approach.

In the next step one has to decide which characteristics should be used to determine a woman's fertility behaviour. Contributions addressing this question are as numerous as manifold. Most of the scientific literature, as Leibenstein (1957), Becker (1960) or Easterlin, Pollak and Wachter (1980) to name just very few from a huge list, argue on microeconomic levels. Macroeconomic approaches also try to derive socioeconomic theories about this subject. Overviews regarding this question can be found in Robinson (1997), Schultz (1997) or Birdsall (1988).

Despite intensive research in this area, only few of the findings are usable for an applied general analysis because in most cases, the need for empirical data on a micro level is very

high. Due to the fact that fertility is such an essential decision in every woman's life, the potential determinants come from a variety of very different aspects of life e.g., political, economical, social, ethnical or religious origin. Modelling these effects in an empirical framework requires sound information about the population or at least a sample on an individual level. This is only possible by means of timely and costly procedures of gathering data, only partly done by administrative statistical institutions. However, one of our goals was to provide a method based on readily available data from official statistics. We therefore do not take into account a broad information range for modelling but restrict ourselves to the age of the mother as the single determinant of fertility for a given census. Further we incorporate structural development over time.

Of course, further information is available from the mentioned data sources. An example is marital status which undoubtedly plays a central role in the childbearing behaviour of women, all other determinants being equal. Another additional information pattern is the ethnicity of the mother which is also a major fertility differential between women. Information about whether the new born child is the first born baby or not is stored in some cases, which could also contribute to the explanation of the fertility phenomena.

The incorporation of these aspects would lead to a remarkable improvement in the estimation quality (see e.g. Chandola, Coleman and Hiorns 1999). The method presented here would not be affected heavily, but the formulation of the underlying distribution functions would grow in complexity, namely in its numerical optimisation. Further, it would lead to an even more complicated notation in the following mathematical analysis. Therefore, these dimensions were not taken into account. It should be mentioned that this information basket must be available at an individual level in order to be implemented into the model. This should be the case in only very few official statistical administrations.

Without considering this additional information, every woman of the same age i $(1 \le i \le a)$ has the same vector z and so the same probability of giving birth to a child, say p_i. Note that a

is an age where childbearing is no longer possible in biological terms. The age years 1 to 12 are also biologically irrelevant. However, for easiness of notation, we start the index with age year 1. It is useful to summarise the women of equal age *i* and to introduce the scalar random variable S_i^* denoting the number of births by these women during the specified year. In doing so, we assume the number of women at age *i*, say N_i as given.

This leads to a binomial distribution for S_i^*

$$\mathrm{P}\left(S_i^* = h_i\right) = \binom{N_i}{h_i} p_i^{h_i} \left(1 - p_i\right)^{N_i - h_i}, \ \left(1 \le i \le a\right), \tag{2}$$

where h_i gives the number of births of the women at age *i*.

In the case of independence of fertility over age we get the joint probability of $\mathbf{S}^* = \left(S_1^*, S_2^*, \ldots, S_a^*\right)$ from (2)

$$P\left(\mathbf{S}^* = \left(h_1, h_2, ..., h_a\right)\right) = \prod_{i=1}^{a} \binom{N_i}{h_i} p_i^{h_i} \left(1 - p_i\right)^{N_i - h_i} . \tag{3}$$

The probability of (3) equals (1), although in the latter notation there is no longer a difference between women of the same age. For the statistical description of our fertility model the random vector S^* is easier to handle than *S*.

Usually, we have a sample containing data over several years, denoted by the index $t\left(1 \le t \le T\right)$. By additionally assuming independence over time, the joint distribution of $\left(S_{t1}^*, S_{t2}^*, \ldots, S_{ta}^*\right)$ can be written as

$$P\left(\left(S_{t1}^*, S_{t2}^*, ..., S_{ta}^*\right) = \left(h_{t1}, h_{t2}, ..., h_{ta}\right), 1 \le t \le T\right) = \prod_{t=1}^{T} \prod_{i=1}^{a} \binom{N_{ti}}{h_{ti}} p_{ti}^{h_{ti}} \left(1 - p_{ti}\right)^{N_{ti} - h_{ti}} . \tag{4}$$

By assumption, this expression contains only factors with $N_{ti} \neq 0$. It describes the probability distribution of the observed births over age as well as over time completely.

Without additional constraints on p_{ti}, M.L. estimation of (4) obviously leads to the relative frequencies $\hat{p}_{ti} = f_{ti} = h_{ti} / N_{ti}$.

FITTING OVER AGE AND TIME

Following this plausible approach we yet have not considered two aspects: First, the probability of childbearing p_{ti} has a smooth development over the mother's age *i*. For a homogenous population this function should be unimodal. Note that the function cannot be considered as distribution, although it is often named as fertility distribution over age. From a biological point of view, the theoretical probability of childbearing below a certain lower boundary age, say 14 years, as well as over an upper boundary age, say 50 years does not need to equal 0 but should be on a very low level and reaching 0 beyond biological borders of age. Secondly, the probability of childbearing p_{ti} should have a smooth development over time. Unlike the first case, there is no more *a priori* information available. Obviously, the last 50 years of demographic analysis have seen a certain amount of structural changes in fertility behaviour. But a brief look at some of the fertility parameters over time of populations show here have been ups and downs in these demographic factors which makes it even harder to explain the directions of structural changes. Even more difficult, and in our opinion almost impossible, is modelling these movements parametrically in the hope of getting information about future development of the parameters. The figures 1a) to 1c) present the Total Fertility Rate (TFR) $=\sum_{i=1}^{a} f_{ti}$, the age mode, a young age fertility index (fertility for age younger than 18 years, $FI_{18}^{y}=\sum_{i=1}^{17} f_{ti}$) and an old age fertility index (fertility for age higher than 39 years, $FI_{39}^{o}=\sum_{i=40}^{a} f_{ti}$) for the observation years 1953 to 1999 for Western Germany. Note that the mode has been estimated calculating a polynomial function of degree two over age with the

highest relative frequency and both neighbouring values. The mode has been identified by setting the first order condition of the polynomial to 0.

It is obvious that these parameters have a constant development over certain periods while in other periods they change the trend or the direction in an unpredictable way. These changes can be explained in the historic context but at the time of their occurrence they have been rather unpredictable. Note that these information patterns have a smooth behaviour over time due to the large size of the underlying population. To illustrate the requirement of smoothness over both dimensions age and time, figure 2 gives an idea how the functional form of fertility over age and time could look. The figure is based on data from west Germany, estimated with the technique described in the following section.

The fertility approach in the context of Generalised Linear Models (GLM)

A major problem in empirical demographic work is that both aspects mentioned are hard to find in empirical raw data for smaller populations. In this case the influence of random errors can bias demographic structures. Therefore, we need to apply two different ways to deal with that problem.

While the development of childbearing over mother's age can be described with different mathematical functional forms due to the demographic knowledge about the shape of the curve, this is not the case for the development over time. A very general way of formulating the problem would be in the GLM-context (e.g. Mc Cullagh 1980)

$$p_{ti} = g\left(pr\left(i, b_{t1}, .., b_{tM}\right)\right),$$

where *g* denotes the *response function* and $pr\left(i, \boldsymbol{b}\right)$ the *predictor*, which depends upon age *i* and a $\left(1 \times M\right)$ parameter vector *b*, depending on time *t*. Usually the predictor is assumed to be linear, in our case we do not necessarily need linearity.

Usually, one of the following response functions is used in the literature concerning GLM. The exponential function $g(x) = \exp(x)$ guaranteeing positive probabilities. A possible shortcoming could be seen in the fact that theoretically values above 1 are possible. The logit function $g(x) = \frac{exp(x)}{1 + exp(x)}$ covers the advantage that only values can occur between 0 and 1.

For our needs, both approaches can be used because the relative frequencies do not exceed 0.2. Bearing in mind that the exponential function allows easier formal derivations, we proceed with this functional form. Note that in other demographic areas like household formation analysis, it might be necessary to focus on the logit function (see Haupt, Oberhofer and Reichsthaler 2003). However, the results of the estimation do not react very sensitively to the choice of the response function.

Fitting over Age

Considerably more important than the response function is the choice of an appropriate predictor. The typical development of age specific fertility, as shown in figure 1a, is similar to a slightly modified normal distribution. It is worth mentioning that the function is not necessarily symmetric, while the skewness could appear in both directions. One way of modelling these aspects parsimoniously would be the use of the exponential functional form for the response and a parametric spline function for the predictor:

$$p_{ti} = \exp\left(b_{t1} - b_{t2}(i - b_{t3})^2 - b_{t4}(i - b_{t3})^2 d_{ti}\right), \qquad (5)$$

where the dummy d_{ti} is defined as

$$d_{ti} = \begin{cases} 0 \text{ for } i \le b_{t3} \\ 1 \text{ for } i > b_{t3} \end{cases}.$$

By reparameterising (5) one can interpret the resulting parameters in terms of demographic patterns:

c_{t1} = young age fertility Index = $FI^{y}_{i_l} = \sum_{i=1}^{i_l} \exp\left(b_{t1} + b_{t2}\left(i - b_{t3}\right)^2\right)$, $i_l < b_{t3}$

c_{t2} = old age fertility Index = $FI^{o}_{i_u} = \sum_{i=i_u}^{a} \exp\left(b_{t1} + b_{t2}\left(i - b_{t3}\right)^2 + b_{t4}\left(i - b_{t3}\right)^2 d_{ti}\right)$, $i_u > b_{t3}$

c_{t3} = total fertility rate TFR = $\sum_{i} \exp\left(b_{t1} + b_{t2}\left(i - b_{t3}\right)^2 + b_{t4}\left(i - b_{t3}\right)^2 d_{ti}\right)$

$c_{t4} = b_{t3}$ = mode

As lower and upper boundary ages i_l and i_u different age values can be defined. In industrialised countries it is plausible to set the lower age boundary at about 17 years and the higher age boundary at about 40 years. This leads to indicators for two demographic aspects which are of crucial interest for describing recent changes of fertility, namely the childbearing behaviour of very young and older women. While fertility underlies certain biological age restraints, there are strong social and ethnical tendencies which influence fertility at ages near these biological borders. Therefore these patterns claim special interest. Additionally, the mode of the fertility curve also serves as an indicator for fertility adjustment towards sociodemographic changes. Note that the four parameters c_{t1}, c_{t2}, c_{t3} and c_{t4} can unambiguously be transferred to b_{t1}, b_{t2}, b_{t3} and b_{t4}.

Undoubtedly our choice of the predictor function is one of a huge number of alternatives. Modelling fertility curves over the mother's age has attracted the attention of many demographers. This has led to a huge number of different specifications. Beta and Gamma functions have been used (see e.g. Hoem et al. 1981, Thompson et al. 1989) as well as the Hadwiger function (Gilje 1969). More recent approaches use mixed models. For example, mixed Hadwiger functions which lead to a more flexible form but incorporates a much higher number of parameters (Chandola et al. 1999).

Interestingly, a number of methods described in the literature can be seen in the GLM context. Bloom (1982) applied the approach of a double-exponential function originally developed by

Coale and McNeil (1972) to the context of fertility analysis. Although the observed variable was modelled in a rather different way – they used the cohort approach and focussed on the first birth of women. Knudsen et al. (1993) applied this functional form on fertility with an exponential function both as predictor and response. However, in most cases the considerations were made in a purely descriptive context.

Given the mother's age and observation year, we have full information about the distribution of the number of births, and can apply Maximum Likelihood estimation. In most cases GLS or OLS estimations are used without considering the true distribution of the random variables. It can be shown that under certain circumstances the results of the M.L. procedure can be seen as a Weighted Least Squares estimation, where the weights arise due to statistical reasons (see Oberhofer and Reichsthaler, 2000, 27).

Fitting over Time

While the distribution of fertility over the mother's age is rather easy to handle demographically as well as statistically, the knowledge about and the handling of the evolution of fertility over time is more ambiguous. From the demographical side of the problem we have some ideas about the structural behaviour of certain parameters over time. One example is the suspicion that the functional mode over the mother's age would move slightly towards a higher age. On the other hand, a brief look at figure 1 demonstrates that neither monotony nor direction of such a structural change remains constant over longer periods. Therefore it seems adequate to describe the movement of the parameters over time in a non parametric approach. Further, one needs to keep in mind the requirement of a smooth development of the parameters over time. Both aspects can be fulfilled using a smoothing procedure. To keep as much information as possible in the estimation procedure, this smoothing approach will not be applied in a second step of the estimation process, but will be carried out simultaneously using the local likelihood procedure (Tibshirani and Hastie 1987).

This leads to a Likelihood function for a time window W_t

$$\prod_{s\in W_t}\prod_{i=1}^{a}\binom{N_{si}}{h_{si}}p_{ti}^{h_{si}}\left(1-p_{ti}\right)^{N_{si}-h_{si}} \tag{6}$$

t = time index

i = age index (from lower age to upper age boundary)

N = number of women

h = observed relative frequencies of births

W_t= time window, $W_t=\{s|t-w\le s\le t+w\}$ with $w+1\le t\le T-w$

t = midway of the window

w = parameter of windows width

$p_{ti}=g\left(pr\left(i,b_{t\circ}\right)\right)$, where $b_{t\circ}=\left(b_{t1},b_{t2},b_{t3},b_{t4}\right)$.

Minimisation of (6) leads to the M.L. estimators $\hat{b}_{t\circ}$.

The estimated probabilities can be written as $\hat{p}_{ti}=g\left(pr\left(i,\hat{b}_{t\circ}\right)\right)$.

The first order conditions of the M.L function to be used for the optimisation as well as the iteration steps can be found in the appendix.

APPLICATION

The Data Set

In this analysis our aim is the estimation of age-specific fertility probabilities. We will illustrate our findings on the basis of a data set for the city of Bolzano in South Tyrol (Italy) from 1990 to 2000. This data set can clearly be described as a small region with less than 100,000 inhabitants. The data are recorded for women of single ages from 14 to50 years.

The data are obtained from the essentially complete reporting of births. Nevertheless, from a statistical point the number of births has to be viewed as a sample. In our model we assumed the probability of bearing a child for every woman to be *p*, this means the total number of

births of all women is stochastic (see e.g. Brillinger 1986:697). This is also true if the number of women in the group is sufficiently high or all women of the region are incorporated in the observation. Thompson et al. (1989) wrote: “Our data contain full age-specific detail and are obtained from the essentially complete reporting of births in the Vital Statistics Registration System, so there is no sampling error”. This statement is misleading, as argued above. We prefer to follow the view of Keyfitz (1966) who suggests that “a census may be regarded as a sample drawn in time from all the times in which substantially the same conditions prevailed”.

Note that we argue in a context of period specific age profiles instead of cohort specific ones. This is due to the fact that one needs a sufficient long observation record of at least 50 years to derive 10 to 15 cohorts with a closed fertility cycle. This data requirement can be fulfilled by some administrative statistical institutions, but an even greater number do not have that long range of data. Accordingly, we use period specific fertility rates. Further, our age specific data are defined as age of the mother at childbearing. This is a slightly different conceptual approach than the widely used method of using the age-group of the mother.

To illustrate the problematic data situation of a small population, we compare it with fertility data from west Germany which is a large population with about 60 million people. Both data sets come from the official Statistical Institutions. As can be seen in figure 3, the data from Bolzano presents a more irregular schedule which is caused by the small age specific groups and underlines the sample character of the data.

Comparing both fertility curves over time, one can derive a bundle of demographic parameters which become more unstable the smaller the relevant sample pool is (see figures 4a-d). This becomes obvious when the shape of the parameter curves from the Bolzano data are compared with the German ones. Besides the visual analysis we calculate the relative

mean deviation of the variables with its lagged values to derive a measure of irregularity of the curves, $RMD = \frac{\sqrt{\frac{1}{T-1}\sum_{t=2}^{T}(x_t - x_{t-1})^2}}{\bar{x}}$.

While the tfr curve (figure 4a) for Bolzano provides a RMD of 0.059, the German one is 0.024.

This is similar to the picture with the age mode (figure 4b), where the Bolzano curve with an RMD of 0.083 is about eight times higher than the German RMD value of 0.010. It is the same with the Old Age (FI^o_{40}, figure 4c) and Young Age Fertility Indexes (FI^y_{18}, figure 4d), where Bolzano has the values 0.253 for the RMD of FI^o_{40} and 0.463 for FI^y_{18} whereas the German values are significantly lower with 0.046 for the FI^o_{40} and 0.0544 for the FI^y_{18}.

The structural development of parameters obviously gets more and more biased by random influences the smaller the population is. This could be seen clearly when comparing both "fertility mountains" over age and time for Germany (big population) and Bolzano (small population). While the first one contains a smooth shape over both dimensions (see figure 5a), the latter has sharp ups and downs which are a clear contradiction against demographic aspects (see figure 5b).

Results

Due to the smoothing procedure, the results do not cover the whole observation period, but are cut off on both borders of the time series. The raw data set for Bolzano originally contains data from 1990 to 2001 while the range of the estimation is from 1992 to 1999. In particular, missing estimation results for the most recent years could be unsatisfying in certain circumstances. This could be remedied by various estimation techniques which are not presented here. As a result of the estimation procedure, the development of the parameters over time is much smoother than the original values, as noted in figures 4a-d.

As mentioned before, the data of the smaller population of Bolzano is more influenced by random effects than a larger sample like Germany. This also means that the estimation and smoothing procedure of our approach has a stronger effect on the Bolzano data than on the German data where the estimation was also applied to compare the smoothing effects.

To demonstrate these effects we provide some measures of fit which have to be interpreted in a rather uncommon direction. While the Mean Absolute Percentage Error (MAPE) and the Mean Algebraic Percentage Error (MALPE) (see e.g. Smith 1987) are commonly used to describe the goodness of fit of an estimation to raw data, the results of both tests can also be interpreted in terms of intensity of the smoothing process of our approach.

Both measures presented in figure 6 show clear evidence that the smoothing of the Bolzano data is much stronger (the mean MAPE value over the nine observation years is 14.3% whereas the German mean MAPE is 5.3%). There is no bias in the smoothing procedure since the MALPE values move around the zero line in both cases.

Additionally, we introduce a measure of the degree of smoothness of a three dimensional function. We therefore calculate the Mean Absolute Percentage Smoothness Index (MAPSI) between p_{sj} and $\tilde{p}_{sj}$, $1 \le \mathrm{i} \le \mathrm{a}$, $1 \le \mathrm{t} \le \mathrm{T}$. We approximate in the window $\{t-1 \le s \le t+1,\ i-1 \le j \le i+1\}$ p_{sj} by a quadratic function $\alpha_0 + \alpha_1 s + \alpha_2 j + \alpha_3 sj$ using least squares. Denoting the approximation by $\tilde{p}_{sj}$ the sum of squared residuals $\sum_{j=i-1}^{i+1} \sum_{s=t-1}^{t+1} \left[p_{sj} - \tilde{p}_{sj}\right]^2$ can be regarded as a measure of relative smoothness in the point (*t*, *i*).

This seems to be an adequate measure of smoothness over both dimensions. To analyse the smoothing effect of the procedure, we need to calculate the Smoothing Quotient $SQ = 1 - \frac{MAPSI_{estimation}}{MAPSI_{rawdata}}$.

In the case of Germany, for the years 1990 to 1996 SQ_{GER} has the value of 0.43 while in Bolzano for the period 1992 to 1998 the parameter is 0.89 (see table 1). This indicates that the smoothing procedure has a much stronger impact on the Bolzano data than on the German data. The resulting fertility mountain does not contain any irregularities as seen in figure 7.

A comparison of the estimated (and smoothed) mountains (figure 7a and 7b) with the mountains of the raw data (figure 5a and 5b) gives an idea of the impact of the smoothing intensity to be measured with the SQ parameter.

The method derives theoretically consistent fertility parameters over time and mother's age, also for smaller populations where random influences become stronger than the underlying structural change patterns. The presented estimation method allows analysis of the change of fertility parameters over time without losing too much information from the raw data. This would be the case if the estimation procedure was applied without considering distributional information or simultaneous estimation over both dimensions.

All four demographic parameters - the total fertility rate, the mode of the fertility function, the development of old age as well as young age childbearing – need to be interpreted cautiously to derive sound statements about underlying changes of biological, social, political, ethical or economic origin. This can only be done by reducing original data from samples to its structures and filtering random noise out. In our opinion the presented method can claim those requirements.

TOWARDS PROJECTIONS

The formulation of a fertility model usually leads to a statement about future development of childbearing behaviour. Many authors use the analysis of historic aspects as a starting point for demographic projections.

Since the mid eighties, the scientific literature concerning this subject has been dominated by models arising from the time series approach of Box and Jenkins (1970).

In our opinion, this is the reason why most demographic contributions about modelling fertility prefer time series instead of non-parametric smoothing methods to model structural changes over time. The results of our method are demographic parameters about fertility which are consistent over the mother's age as well as over time. In particular, the last property can identify them as interesting input for further time series projection approaches.

Despite widespread use, the success of this complex methodology is doubtful (Stoto 1983, Smith 1987); moreover, resulting confidence intervals are disappointingly large which diminishes the meaningfulness of the models. As Lee (1993:187) points out, "...many of these methods…are ingenious and there has been some isolated success. On the whole, however, the state of our knowledge and understanding in this important area is discouraging, despite the substantial resources that have been devoted to it"

Approaches primarily using time series analysis supply accurate estimations of the demographic rates in between the sample survey, but no valid extrapolations could be made for mid-term periods of about 10 to 20 years. This arises from the fact that time series models can produce a good fit for any function, at least if they are heavily parameterised. However, they have no ability to find demographic causalities, which are fundamental for an appropriate statement about future development. An interesting example is the fertility in east Germany which has diminished by about a half over about five years since the German reunification. Therefore, it seems necessary that population projections should be based on social theories, linking social, cultural and economic factors to demographic behaviour (Keilman 1990). This means that the "objective" statistical data-driven approach is repelled by a more "subjective" but causality driven one.

Moreover, the opinion in literature concerning the use of socioeconomic knowledge in projections is thoroughly ambiguous. While Keyfitz (1981) found positive influences on the

accuracy of demographic predictions, other research came to contrasting results (Alho 1990, Alho and Spencer 1990), which are caused by the “assumption drag” (Ascher 1985); assumptions that are too conservative(Ahlburg and Land 1992). Other findings emphasise that additional factors should not be too restrictive for the development of the time series model (Alho 1997).

CONCLUDING REMARKS

Describing age-specific fertility has attracted the attention of huge numbers of demographers over the last decades. Despite its inherent interdisciplinarity, this topic seems to contain a remarkable gap between demographic theory and statistical methodology.

In this broad field of diverging aspects our contribution tries to offer an easy model of fertility based on statistical as well as demographic theory. Accordingly, we generated a micro analytically founded model of fertility, obtaining the number of births in a population, given the age of the mother, as a Bernoulli random variable. This is a clear deviation from the usually heuristic analysis of fertility patterns.

Due to the requirement of developing an approach using only data available from official statistics, only three inherently given information have been used – gender, age and observation time. As the distribution of the childbearing probability is known, the model can be nested into the widely known family of GLM. Subsequently, the age specific fertility has been modelled parametrically. A parsimoniously parameterised function has been used. Each parameter can be transformed unambiguously into a demographic attribute which makes the approach much more comprehensive.

Instead of choosing the response and the predictor function due to purely mechanistic and statistical aspects which is more or less usual for GLM approaches, we defined both functions according to demographic facts.

To derive appropriate smoothness over both dimensions age and time, we specified a Local Likelihood function, which is a nonparametric procedure over time. The simultaneous estimation over both dimensions contains much more information than the ordinary sequential procedure.

The result of the approach was shown with a data set for the Italian city of Bolzano with less than 100,000 people serves as an example of a smaller population. As a result of our estimation, we got a smooth fertility mountain where random effects were filtered but

structural changes of fertility behaviour remain in the data set. The presented approach has the major advantage that the distribution of the number of births, given the mother's age, is known. Therefore we do not need to estimate the parameters with an *ad hoc* OLS or GLS method, but can use the much more sophisticated M.L. technique. This leads to an additional information profit compared to ordinary approaches.

The crucial improvement of the presented approach lies in the fact that we do not argue with heuristic facts, but define a plausible, statistically and demographically well founded and consistent model. It should be noted that with this approach not all information available from administrative statistical institution data sources was exploited. In most cases there is additional information available like marital status or ethnic origin, which undoubtedly has an impact on a woman's childbearing behaviour.

Our model can be extended in these dimensions. While the estimation procedure would not be affected heavily by such an extension, the distribution of the random variable would become far more complex. Additionally, one needs to answer the question of modelling the inferences between different determinants such as age with marital status, ethnic origin with time and so on. A lot of work remains in answering these questions. In our opinion, the presented approach is an interesting alternative to the usual methods and can serve as an initial point to deal with new, demanding demographic as well as statistical questions.

APPENDIX: THE MAXIMUM LIKELIHOOD ESTIMATION

The first order conditions of the log likelihood for a fixed t

$$\nabla L_t(b) = \sum_{s\in W_t}\sum_{i=1}^{a} \frac{[h_{si} - N_{si}\, p_{ti}]}{1-p_{ti}} \frac{\partial pr(i,b)}{\partial b_j} = 0, \quad 1 \le j \le 4 \qquad \text{(a1)}$$

lead to the M.L. estimators $\hat{b} = (\hat{b}_1, ..., \hat{b}_4)$ and for every t $(w+1 \le t \le T-w)$ results a different estimator.

Note that pr(i, b) is a quadratic function in b_j which helps a lot in calculating the first order conditions.

The four equations (a1) can be solved iteratively using the Fisher Score approach

$b^{(l+1)} = b^{(l)} - H_t^{-1}(b^{(l)}) \nabla L_t(b^{(l)}),\ l = 0,1,2,...$ where $H_t^{-1}(b) = \left(\frac{\partial^2 L_t(b)}{\partial b_j\, \partial b_k} \right)$ and

$$\frac{\partial^2 L_t(b)}{\partial b_j\, \partial b_k} = \sum_{s\in W_t}\sum_{i=1}^{a} \frac{(h_{si} - N_{si} p_{ti})}{(1-p_{ti})^2} \frac{\partial pr(i,b)}{\partial b_j} \frac{\partial pr(i,b)}{\partial b_k} + \sum_{s\in W_t}\sum_{i=1}^{a} \frac{(h_{si} - N_{si} p_{ti})}{1-p_{ti}} \frac{\partial^2 pr(i,b)}{\partial b_j\, \partial b_k} .$$

The Iterations finishes if the difference between two following estimations $b^{(l+1)} - b^{(l)}$ is sufficiently small.

As mentioned before, this approach delivers no estimators for $t = 1,...,q$ and $t = T-q,...,T$.

One could use methods to derive values for these data borders, but we do not follow this path here.

REFERENCES

Ahlburg, D. A. and Land, K. C. 1992. „Population forecasting: guest Editors` introduction," *International Journal of Forecasting* 8:289-299.

Alho, J. M. 1990. "Stochastic methods in population forecasting", *International Journal of Forecasting* 6: 521-530.

Alho. J. M. 1997. "Scenarios, Uncertainty and Conditional Forecasts of the World Population", *Journal of the Royal Statistical Society* 160, Part 1: 71-85.

Alho, J. M. and B.D. Spencer. 1990. "Error Models for Official Mortality Forecasts", *Journal of the American Statistical Association* 85:609-616.

Ascher, W. 1978. *Forecasting: An Appraisal for Policy Makers and Planners*, Baltimore: Johns Hopkins University Press.

Becker, G. S. 1960. "An economic Analysis of Fertility." in: *Demographic and Economic Change in Developed Countries*. Princeton: Princeton University Press.

Birdsall, N. 1988. "Economic Approaches To Population Growth." In *Handbook of Development Economics*, edited by H. Chenery and T.N. Srinivasn. Amsterdam: North Holland.

Bloom, D. E. 1982. „What's Happening to the Age at First Birth in the United States? A Study of Recent Cohorts," *Demography* 19(3):351-370.

Box, G. E. P. and G.M. Jenkins.1970. *Time Series Analysis – Forecasting and Control*, San Francisco: Holden Day.

Brillinger, D.R. 1986. "The Natural Variablity of Vital Rates and Associated Statistics." *Biometrics* 42:693-734.

Cannan, E. 1895. "The probability of a cessation of the growth of population in England and Wales during the next century", *Economic Journal* 5:505-615.

Chandola, T., D.A. Coleman and R.W. Hiorns. 1999. "Recent European Fertility Patterns: Fitting to 'Distorted' Distributions." *Population Studies* 53(3):317-329.

Coale, A. J. and D.R. McNeil. 1972. „The distribution by Age of the Frequency of First Marriage in a Female Cohort,“ *Journal of the American Statistical Association* 67:743-749.

Easterlin, R.; R. Pollak, and M. Wachter. 1980. “Toward a More General Economic Model of Fertility Determination.” Pp. 81-150 in *Population and Economic Change in Developing Countries,* edited by R. Easterlin. Chicago: Chicago Press.

Gilje, E. 1969. “Fitting Curves to Age-Specific Fertility Rates.” *Statistical Review of the National Census Bureau of Statistics of Sweden*, Third Series, Vol. 7:118-134.

Haupt, H., W. Oberhofer and T. Reichsthaler. 2003. „A Varying-coefficient Approach to Estimation and Extrapolation of Household Size“, *Mathematical Population Studies* 10:249-273.

Hoem, J., D. Madsen, J.L. Nielsen, E.-M. Olsen, H.O. Hansen and B. Rennermalm. 1981. „Experiments in Modelling Recent Danish Fertility Curves.” *Demography* 18(2):231-244.

Keyfitz, N. 1966. “Sampling variance of standardized mortality rates”, *Human Biology* 38:309-317.

Keyfitz, N. 1981. “The Limits of Population Forecasting.“ *Population and Development Review* 7:579-593.

Knudsen Ch., R. McNown and A. Rogers. 1993. “Forecasting Fertility: An Application of Series Methods to Parameterized Model Schedules”. *Social Science Research* 22:1-23.

Fahrmeir, L. and C. Kredler. 1984. “Verallgemeinerte lineare Modelle”, in: *Multivariate statistische Verfahren*, edited by L. Fahrmeir and A. Hamerle. Berlin: deGruyter.

Lee, R. D. 1993. „Modelling and Forecasting the Time Series of US Fertility: Age Distribution, Range, and Ultimate Level.” *International Journal of Forecasting* 9:187-202.

Leibenstein, H.M. 1957. *Economic Backwardness and Economic Growth*. New York: Wiley.

McCullagh, P. 1980. “Regression models for ordinal data,” *Journal of the Royal Statistical Society* 42, Part B:109-127.

Oberhofer, W. 1998. “The expected Populations Development for South Tyrol” (German and Italian version only). Bozen: Landesinstitut für Statistik, vol. 58.

Oberhofer, W. and T. Reichsthaler. 2000. „Dealing with Fertility: About the Pitfalls of Models and Forecasts.” Discussion Paper No.345. University of Regensburg.

Robinson, W.C. 1997. „The Economic Theory of Fertility over Three Decades.“ *Population Studies* 51:63-74.

Schultz, T.P. 1997. „Demand for Children in Low Income Countries.“ Pp. 349-430, in *Handbook of Population and Family Economics*, edited by M.R. Rosenzweig and O. Stark. Amsterdam: Elsevier Science B.V.

Smith, S.K. 1987. „Tests of Forecast Accuracy and Bias for Country Population Projections,“ *Journal of the American Statistical Association* 82:991-1012.

Stoto, M.A. 1983. „The Accuracy of Population Projections,“ *Journal of the American Statistical Association* 78:13-20.

Thompson, P.A., W.R. Bell, J.F. Long and R.B. Miller. 1989. „Multivariate Time Series Projections of Parameterised Age-specific Fertility Rates,“ *Journal of the American Statistical Association* 84:689-699.

Tibshirani, R. and T. Hastie. 1987. “Local Likelihood Estimation.” *Journal of the American Statistical Association* 82:559-567.

TABLE 1. MEAN ABSOLUTE PERCENTAGE SMOOTHING INDEX (MAPSI) FOR WESTERN GERMANY AND BOLZANO: 1990-1998.

	Western Germany			Bolzano		
	MAPSI[a,b] in %			MAPSI[a,b] in %		
Year	Raw Data	Fitted Data	SQ[b]	Raw Data	Fitted Data	SQ[b]
1990	3.26	1.37	0.58			
1991	1.75	1.23	0.30			
1992	1.56	1.27	0.19	17.34	1.87	0.89
1993	2.13	0.94	0.56	14.59	1.72	0.88
1994	1.94	1.05	0.46	14.02	1.28	0.91
1995	1.95	1.32	0.32	18.99	1.61	0.92
1996	1.98	1.05	0.47	14.57	1.38	0.91
1997				14.21	2.07	0.86
1998				17.62	2.03	0.88
	Mean Value over observation period of six Years					
	2.09	1.18	0.43	15.87	1.71	0.89

[a] Calculated for the ages 15 to 43.

[b] See text for detailed information about the indexes.

FIGURE 1a. TOTAL FERTILITY RATE FOR WEST GERMANY; 1954 TO 1999

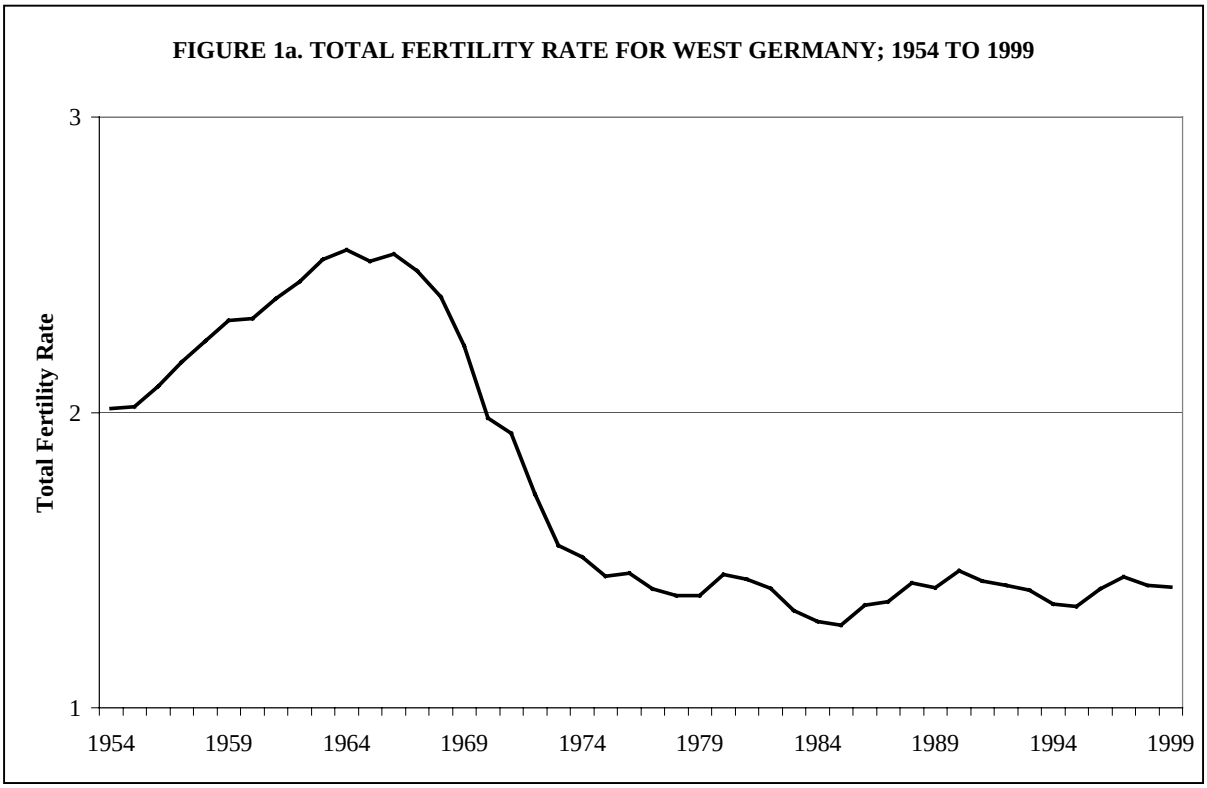

FIGURE 1b. MODE OF MOTHERS' AGE FOR WEST GERMANY; 1954 to 1999

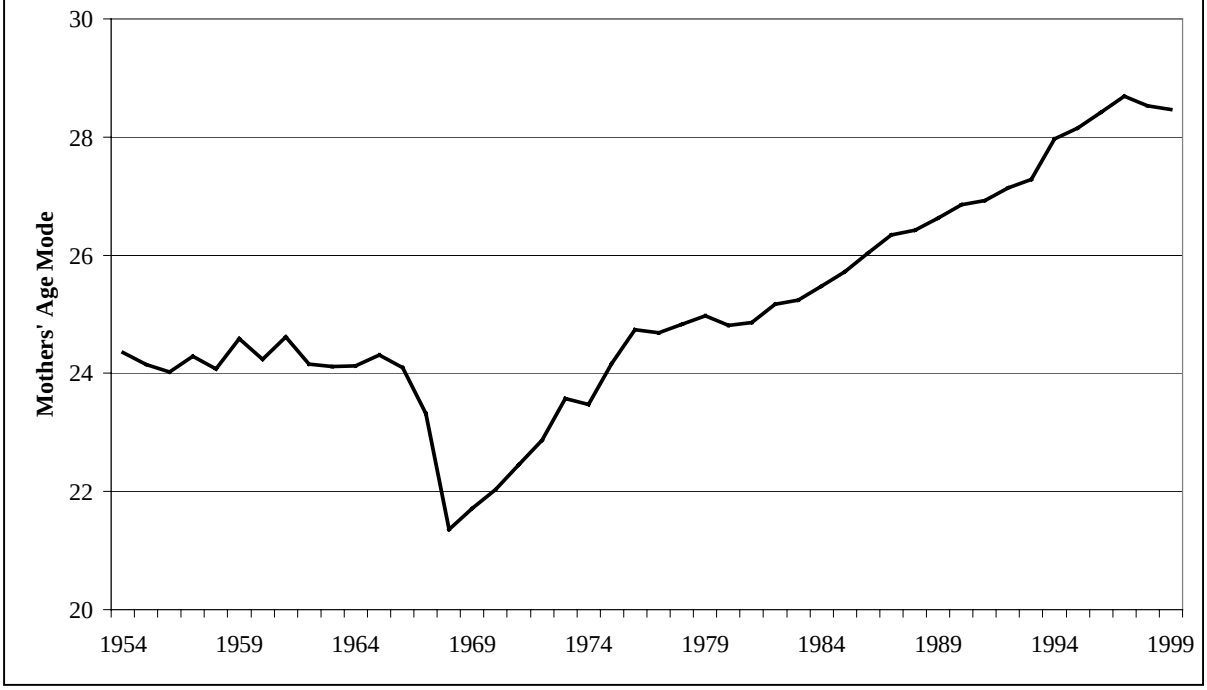

FIGURE 1c. YOUNG AGE AND OLD AGE FERTILITY INDEX FOR WEST GERMANY; 1954 TO 1999

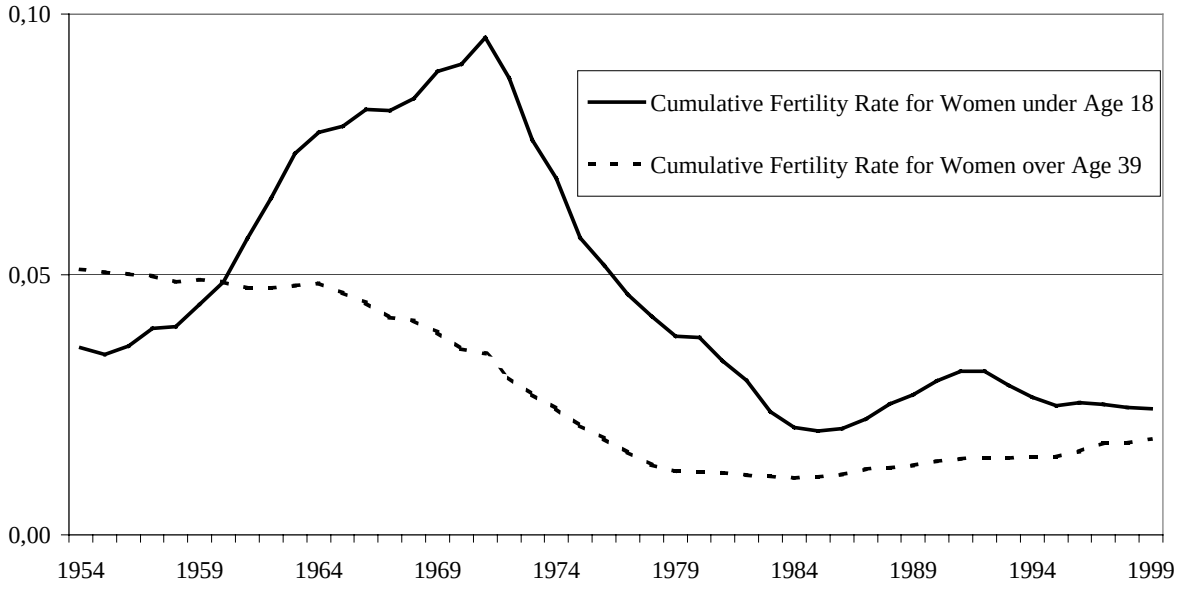

FIGURE 2. SMOOTH FERTILITY MOUNTAIN OVER WOMEN'S AGE AND TIME; EXAMPLE BASED ON DATA FOR WEST GERMANY

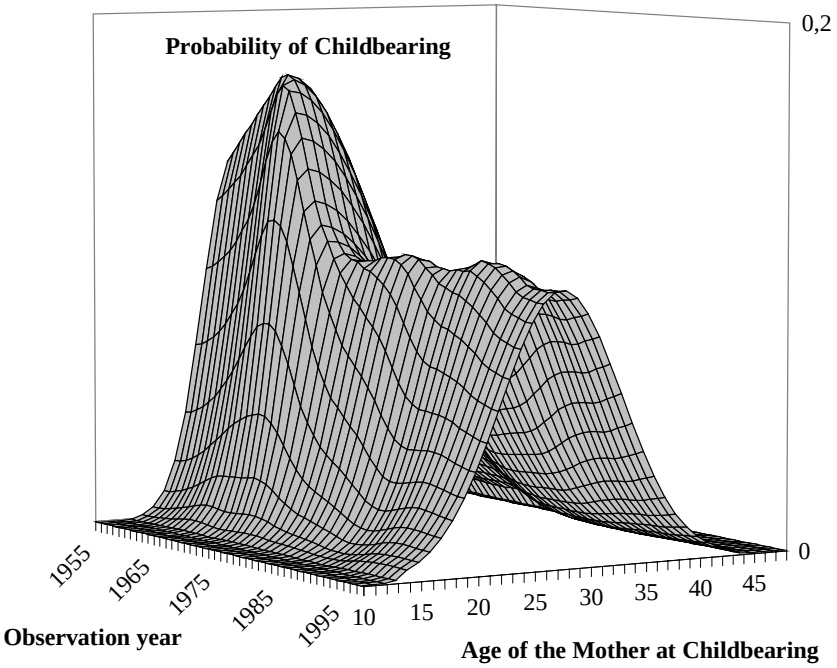

FIGURE 3. FERTILITY FUNCTION OVER MOTHERS' AGE FOR WEST GERMANY AND BOLZANO; 1999

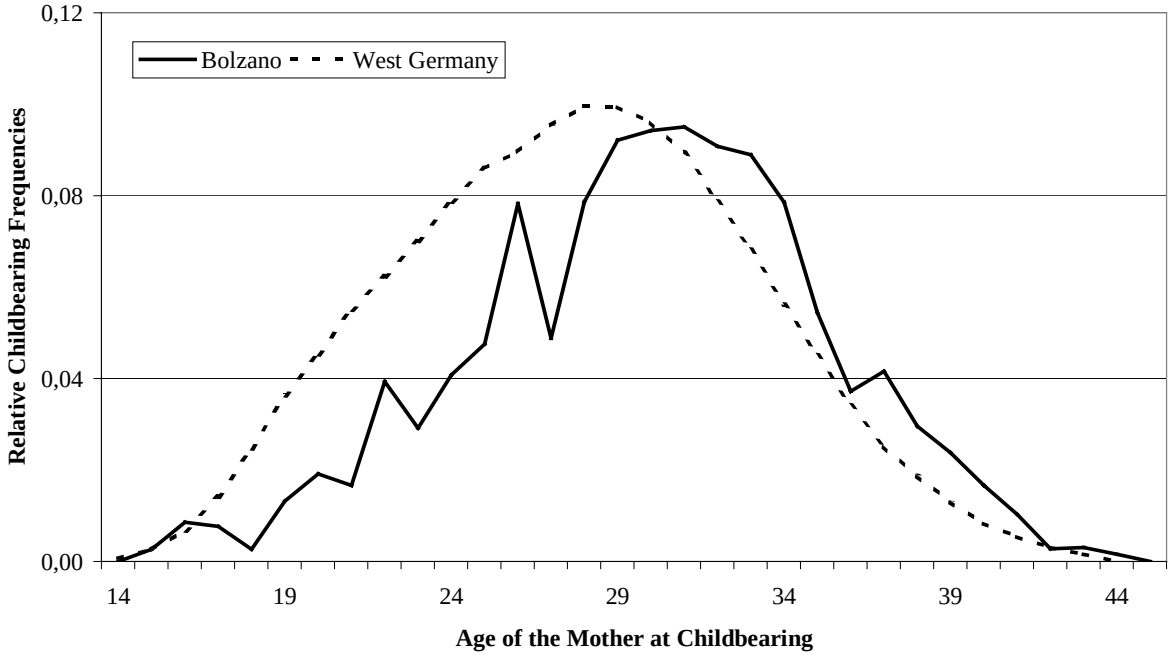

FIGURE 4a. TOTAL FERTILITY RATE FOR WEST GERMANY AND BOLZANO; 1990 TO 2001

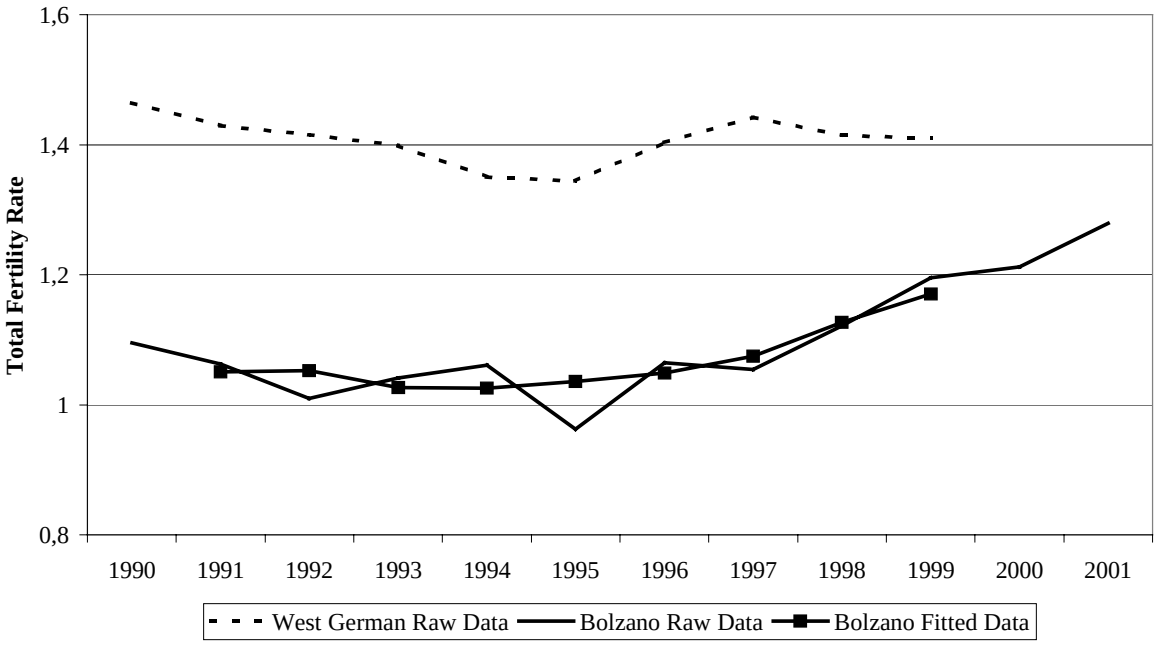

FIGURE 4b. MODE OVER MOTHERS' AGE FOR WEST GERMANY AND BOLZANO; 1990 TO 2001

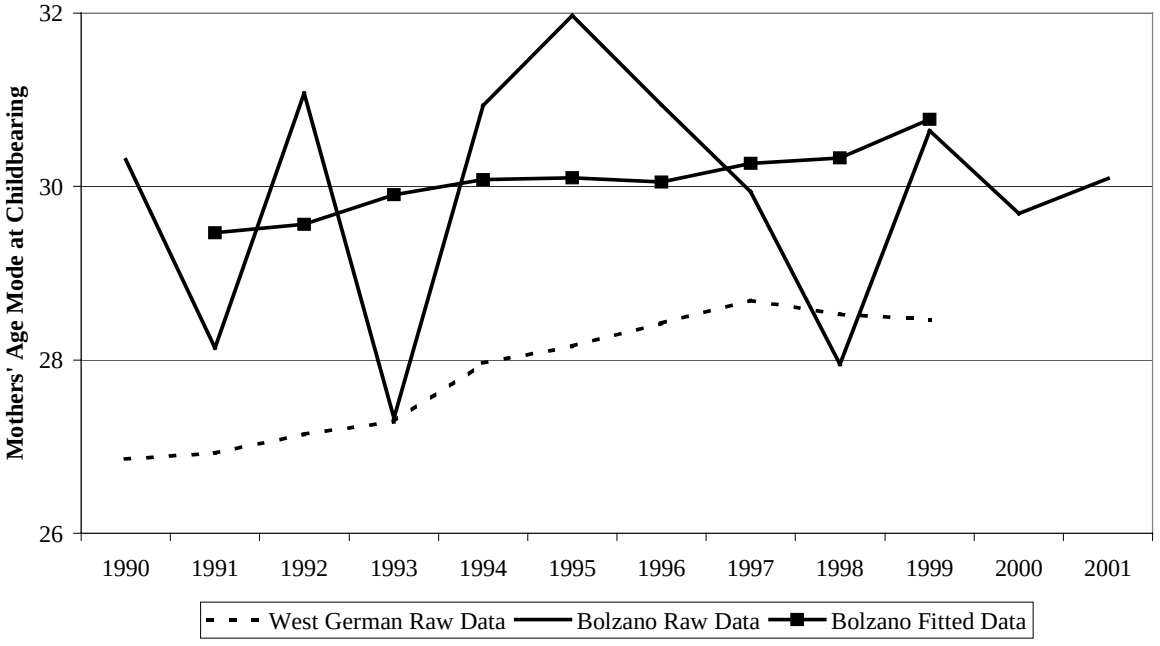

FIGURE 4c. OLD AGE FERTILITY INDEX FOR WEST GERMANY AND BOLZANO; 1990 TO 2001

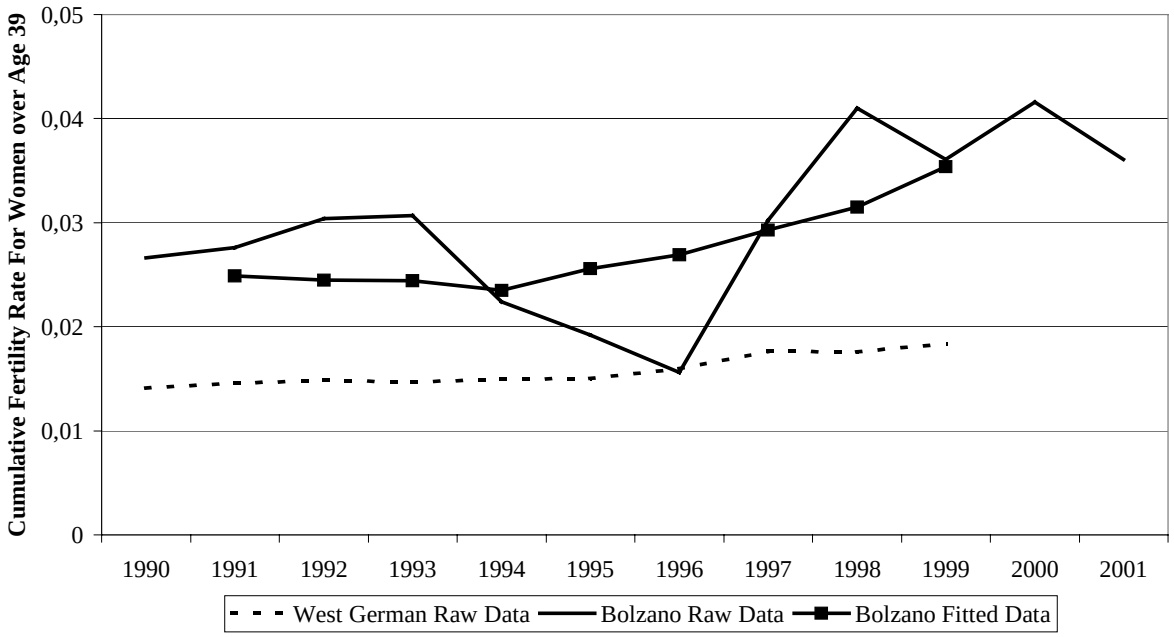

FIGURE 4d. YOUNG AGE FERTILITY INDEX FOR WEST GERMANY AND BOLZANO; 1990 TO 2001

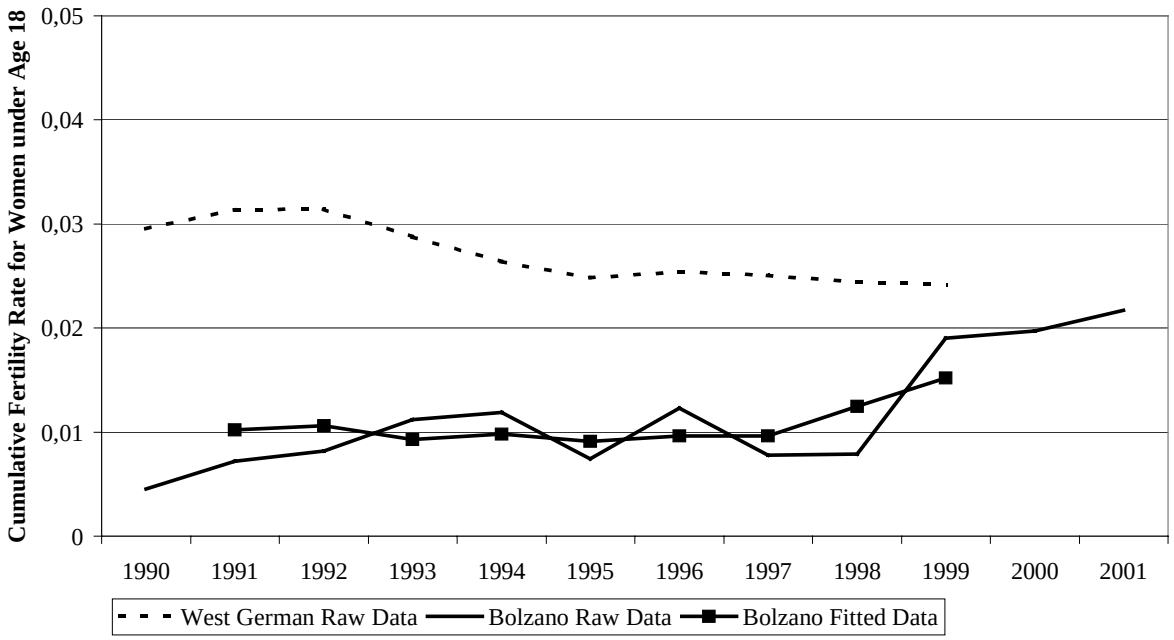

FIGURE 5a. FERTILITY MOUNTAIN OVER WOMEN'S AGE AND TIME FOR WEST GERMANY; 1990 TO 1999.

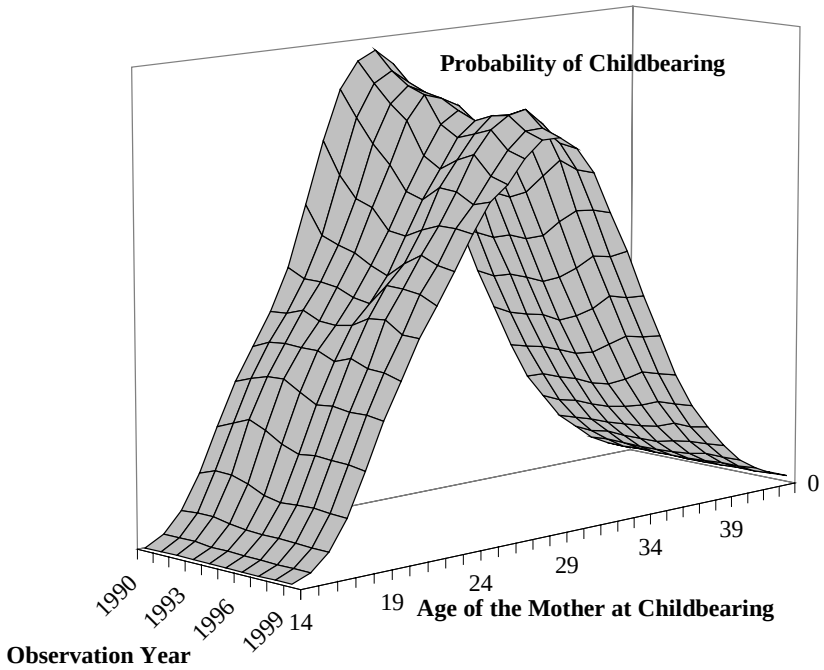

FIGURE 5b. FERTILITY MOUNTAIN OVER WOMEN'S AGE AND TIME FOR BOLZANO; 1990 TO 2001.

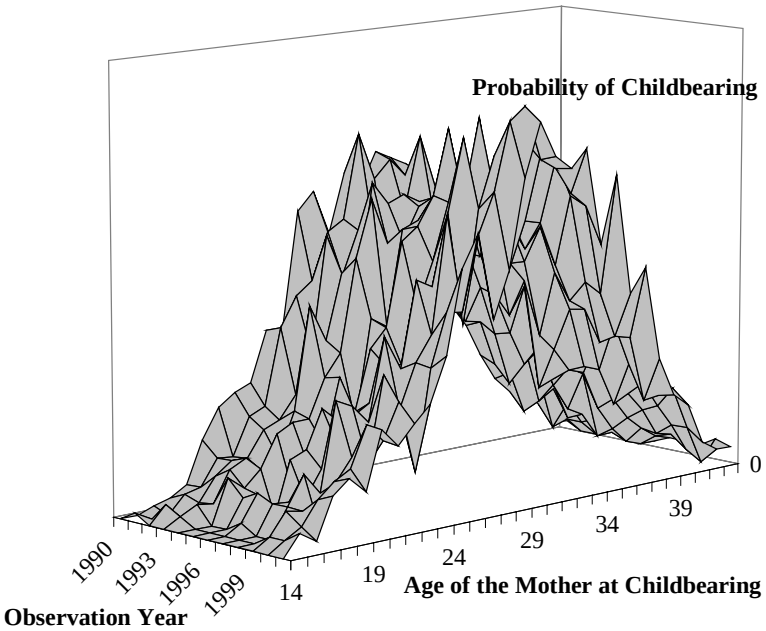

FIGURE 6. GOODNESS OF FIT MEASURES FOR WEST GERMANY AND BOLZANO; 1988 TO 1999.

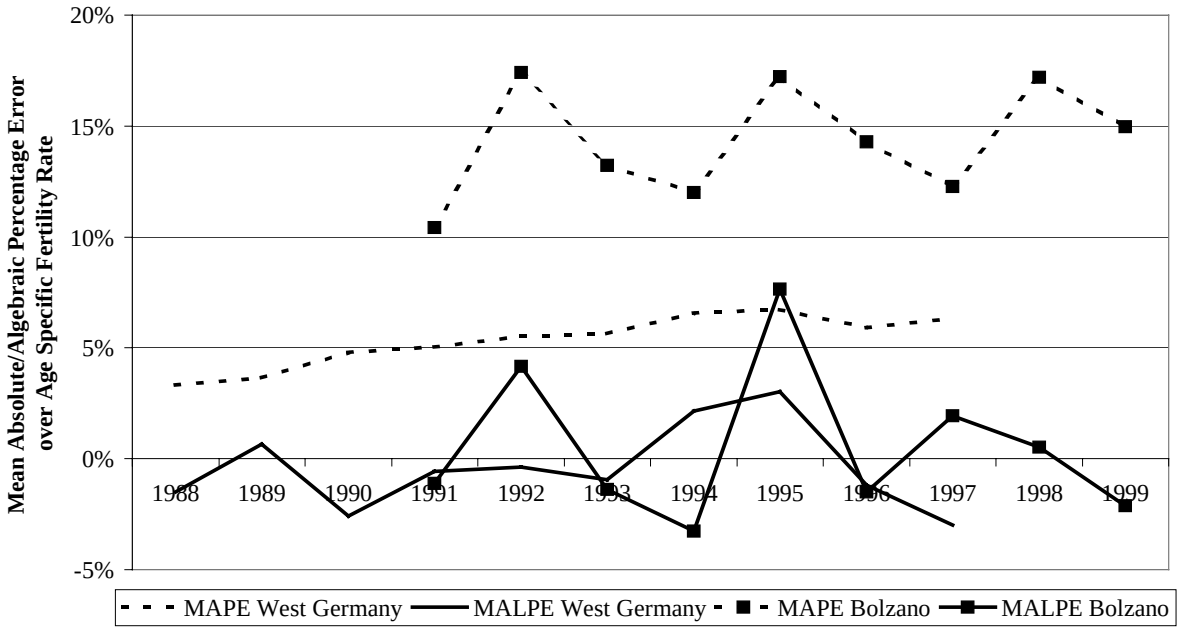

FIGURE 7a. SMOOTHED FERTILITY MOUNTAIN OVER WOMEN'S AGE AND TIME FOR WEST GERMANY; 1990 TO 1997.

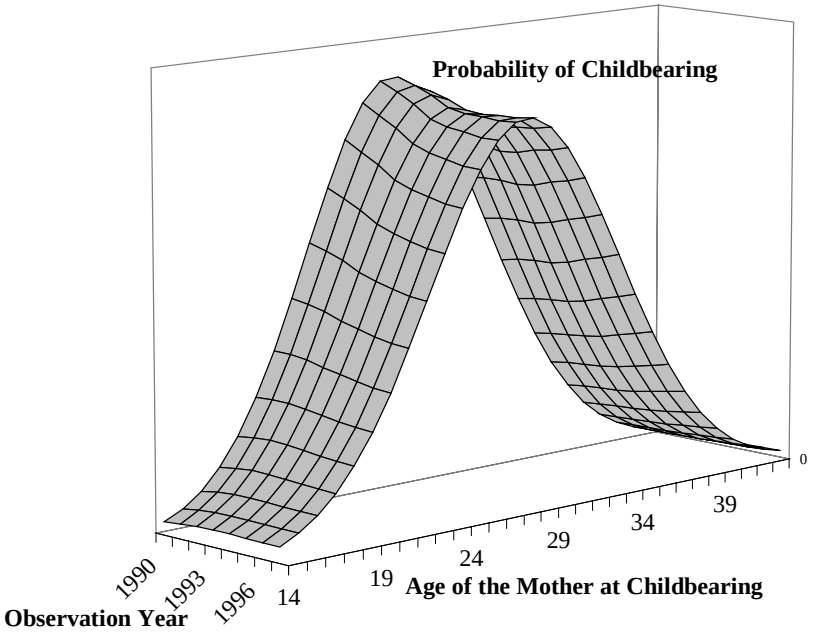

FIGURE 7b. SMOOTHED FERTILITY MOUNTAIN OVER WOMEN'S AGE AND TIME FOR BOLZANO; 1991 TO 1999.