Demographic Research a free, expedited, online journal
of peer-reviewed research and commentary
in the population sciences published by the
Max Planck Institute for Demographic Research
Konrad-Zuse Str. 1, D-18057 Rostock · GERMANY
www.demographic-research.org

DEMOGRAPHIC RESEARCH

VOLUME 20, ARTICLE 25, PAGES 599-622
PUBLISHED 03 JUNE 2009
http://www.demographic-research.org/Volumes/Vol20/25/
DOI: 10.4054/DemRes.2009.20.25

Research Article

Graduating the age-specific fertility pattern using Support Vector Machines

Anastasia Kostaki

Javier M. Moguerza

Alberto Olivares

Stelios Psarakis

Table of Contents

Demographic Research: Volume 20, Article 25
Research Article

Graduating the age-specific fertility pattern using Support Vector Machines

Anastasia Kostaki [1]

Javier M. Moguerza [2]

Alberto Olivares [3]

Stelios Psarakis [4]

Abstract

A topic of interest in demographic literature is the graduation of the age-specific fertility pattern. A classical graduation technique extensively used by demographers is to fit parametric models that accurately reproduce it. Standard non parametric statistical methodology, as kernels and splines, might alternately be used for this graduation purpose. Support Vector Machines (SVM) is an innovative non parametric methodology that could also be used for fertility graduation purposes. This paper evaluates SVM techniques as tools for graduating fertility rates. To that end, we apply these techniques to empirical age-specific fertility rates from a variety of populations and time periods. Additionally, for comparison reasons we also fit parametric models and kernels to these empirical data sets.

[1] Department of Statistics, Athens University of Economics and Business. 76, Patission St. 10434, Athens, Greece. E-mail: kostaki@aueb.gr
[2] Department of Statistic and Operational Research, Rey Juan Carlos University, Spain
[3] Department of Statistic and Operational Research, Rey Juan Carlos University, Spain
[4] Department of Statistics, Athens University of Economics and Business

1. Introduction

Statistical graduation techniques are useful tools in demographic research, producing estimations of the true patterns of age-specific demographic rates. Therefore such techniques can serve in order to provide a clear description of the real shape of age-specific patterns and also consequently serve as a clear basis for population projections.

In fertility analysis, in order to estimate the unknown age-specific fertility rates which underlie the empirical measures, some graduation technique can be applied to the latter, under the assumption that the true rates follow a smooth pattern through age. A standard technique used for graduating the empirical rates is to provide a model that presents the age-specific birth rates as a parametric function of age. Modeling fertility curves has attracted the interest of demographers for many years. A variety of parametric models presenting the fertility rates as a function of age have been proposed in order to describe the age-specific fertility pattern. Some of them provide nice fits to the one year age-specific fertility rate distribution (Hoem et al. 1981; Peristera and Kostaki 2007). Recently the utilization of non parametric techniques in smoothing problems has gained attention. Alternatively, standard non parametric statistical methodology, such as kernels and splines, might also be used for this graduation purpose.

Support Vector Machines (SVM) is a modern non parametric graduation technique that appeared in the mid nineties in the framework of Vapnik's Statistical Learning Theory (Vapnik 1995; Moguerza and Muñoz 2006). Since SVM techniques have shown very successful results in smoothing noisy data, such as neighbourhood curves (Muñoz and Moguerza 2005) or nonlinear profiles (Moguerza, Muñoz, and Psarakis 2007), they can probably serve as useful tools for fertility graduation purposes too. For this application, SVMs are used as general purpose smoothers that enforce a degree smoothness that is chosen by the modeler.

This work provides an evaluation of the SVM methodology in the context of fertility graduation. In the next section, a review of existing parametric models for fitting fertility data is given. Section 3 provides a brief presentation of kernel techniques, while Section 4 is devoted to a presentation of SVM methodology. Then, in Section 5, the results of our calculations fitting parametric models and applying kernels and SVM to a variety of empirical data sets are presented, while finally, in Section 6, the main findings of our calculations are briefly discussed.

2. Parametric models of fertility

A variety of parametric models for fitting the age-specific fertility curve have been proposed in the literature. Among these models, several have been proved to provide accurate

fits to one year age-specific fertility distributions. At the outset a presentation of these models is provided.

The Hadwiger function (Hadwiger 1940; Gilje 1969) is expressed by,

$$f(x) = \frac{ab}{c}\left(\frac{c}{x}\right)^{\frac{3}{2}} \exp\left\{-b^2\left(\frac{c}{x} + \frac{x}{c} - 2\right)\right\},$$

where x is the age of the mother at birth and a, b, c are the three parameters to be estimated. Chandola, Coleman and Horns (1999) argued that the parameters of the model may have a demographic interpretation as follows. Parameter a is associated with total fertility, parameter b determines the height of the curve, parameter c is related to the mean age of motherhood, while the term $\frac{ab}{c}$ is related to the maximum age-specific fertility rate (or modal age-specific fertility rate).

The Gamma function (Hoem et al. 1981) is given by,

$$f(x) = R\frac{1}{\Gamma(b)c^b}(x-d)^{b-1}\exp\left\{-\frac{x-d}{c}\right\}, \qquad \text{for } x > d$$

where d represents the lower age at childbearing, while the parameter R determines the level of fertility. The parameters b and c have no direct demographic interpretation, but Hoem et al. (1981) have made substitution using these by the mode m, the mean μ and the variance σ^2 of the density, so that $c = \mu - m$ and $b = \frac{\mu - d}{c} = \frac{\sigma^2}{c^2}$.

The Beta function also proposed by Hoem et al. (1981) is given by the formula,

$$f(x) = R\frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)}(\beta-\alpha)^{-(A+B-1)}(x-\alpha)^{A-1}(\beta-x)^{B-1}, \qquad \text{for } \alpha < x < \beta.$$

Its parameters are related to the mean ν and the variance τ^2 through the relations

$$B = \left\{\frac{(\nu-\alpha)(\beta-\nu)}{\tau^2} - 1\right\}\frac{\beta-\nu}{\beta-\alpha} \qquad \text{and} \qquad A = B\frac{\nu-\alpha}{\beta-\nu}.$$

As Hoem et al. (1981) mention, the parameters α and β are frequently interpreted as the lower and upper age limits of fertility, respectively. The parameter R determines the overall level of fertility.

Schmertmann (2003) proposed an alternative model for representing age-specific fertility schedules. This is obtained by defining three index ages that describe the shape of the age-specific fertility using a piecewise quadratic spline function. This model describes the shape of the age-specific fertility rates in terms of the ages at which some certain characteristic points are reached; α is the youngest age at which fertility rises above zero, R is the age at which fertility reaches its peak level, and H is the youngest age above R at which fertility falls to half of its peak level.

The model proposed is given by

$$f(x) = \begin{cases} R\sum_{k=0}^{4}\theta_k(x-t_k)_+^2, & \alpha \le x \le \beta \\ 0 \qquad\qquad\qquad\qquad\quad , & \text{otherwise.} \end{cases}$$

Knots $t_0 < t_1 \ldots < t_4$ fall in the interval between ages α and β, where $t_0 = a$, (the lowest age of childbearing) and $(x-t_k)_+ \equiv \max[0, x-t_k]$.

As Schmertmann (2003) mentions the quadratic spline model can be useful for describing the shape of many fertility schedules but it requires thirteen parameters to be estimated, while their meaning is somewhat opaque. Therefore, he constructed a spline model in which the three index ages $[\alpha, P, H]$ determine the shape function $f(x)$, while the parameter R determines the level of fertility. The reduction of the number of parameters is achieved by determining knot positions from the index ages and by imposing mathematical restrictions so that the spline function mimics common features of the age-specific fertility rates.

Recently, as Peristera and Kostaki (2007) describe, the fertility pattern in some developed countries exhibits a deviation from its classical shape. Recent data sets of the United Kingdom, Ireland and the USA show distortions in terms of a bulge in fertility rates of younger women. Furthermore, in countries with distorted fertility, the pattern of first births also exhibits an intense hump in younger ages, stronger than that of the total fertility pattern. This heterogeneity initially discovered in the recent fertility distributions of the English speaking European countries and the USA, but recently even in data from other European populations as Spain and Norway, might be related to marital status, religion, educational level and differences in social and economic conditions, as well as to ethnic differences in the timing and the number of births. As expected, the existing models are unable to describe the new shape of the fertility pattern, and therefore the use of more appropriate representations is required.

For describing the new shape of the fertility pattern, Chandola, Coleman and Horns (1999) developed a two-component mixture model of Hadwiger functions which is given by the following expression,

$$\begin{aligned} f(x) = am\left(\frac{b_1}{c_1}\right)\left(\frac{c_1}{x}\right)^{\frac{3}{2}} \exp\left\{-b_1^2\left(\frac{c_1}{x}+\frac{x}{c_1}-2\right)\right\} \\ + (1-m)\left(\frac{b_2}{c_2}\right)\left(\frac{c_2}{x}\right)^{\frac{3}{2}} \exp\left\{-b_2^2\left(\frac{c_2}{x}+\frac{x}{c_2}-2\right)\right\}, \end{aligned}$$

where x is the age of the mother at birth. This model requires the estimation of six parameters: m is the mixture parameter that determines the relative sizes of the two component distributions and α, b_1, c_1, b_2, c_2 the other parameters of the model. According to the

authors, these parameters may also be demographically interpreted. Parameter α is correlated with the overal fertility level, c_1 and c_2 are related to the level and the trend of the mean ages of births outside and inside marriage.

Recently, Peristera and Kostaki (2007) proposed a flexible model for describing the fertility pattern which in each different version captures both the classical and the distorted fertility pattern. The simpler version of this model (hereafter P-K model) is

$$f(x) = c_1 \exp\left[-\left(\frac{x-\mu}{\sigma(x)} \right)^2 \right]$$

where $f(x)$ is the age-specific fertility rate at mother age x, c_1, μ, σ are parameters to be estimated, while $\sigma(x) = \sigma_{11}$ if $x \leq \mu$, and $\sigma(x) = \sigma_{12}$ if $x > \mu$. The parameter c_1 describes the base level of the fertility curve and is associated with the total fertility rate, μ reflects the location of the distribution, i.e. the modal age, while σ_{11}, σ_{12} reflect the spread of the distribution before and after its peak, respectively.

An alternative version of this model (hereafter P-K mixture model), which captures the new shape of the fertility pattern mentioned above, is

$$f(x) = c_1 \exp\left[-\left(\frac{x-\mu_1}{\sigma_1(x)} \right)^2 \right] + c_2 \exp\left[-\left(\frac{x-\mu_2}{\sigma_2} \right)^2 \right]$$

where $f(x)$ is the age-specific fertility rate at mother age x, while $\sigma_1(x) = \sigma_{11}$ if $x \leq \mu_1$, and $\sigma_1(x) = \sigma_{12}$ if $x > \mu_1$, and c_1, c_2, μ_1, μ_2, σ_{11}, σ_{12}, σ_2 are parameters to be estimated.

The parameters c_1 and c_2 express the levels of total fertility of the first and the second hump respectively, μ_1 and μ_2 are related to the mean ages of the two subpopulations; the one with earlier fertility and the other with fertility at later ages. The parameters σ_{11}, σ_{12} reflect the spread of the distribution of the most intense hump before and after each peak, and σ_2 reflects the variance of the less intense one.

3. Kernel techniques

Consider a set of observations of two variables X and Y, i.e. data of the form (x_i, y_i), $i = 1, \ldots, p$ which are related via an unknown regression function m as follows:

$$y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \ldots, p,$$

where the ε_i are independent random variables with zero mean and constant variance.

The problem now consists in estimating the unknown function m. In order to estimate m at a point x, the values of the response variable are locally averaging. The width

of the neighbourhood over which averaging is performed, called bandwidth, controls the smoothness of the resulting estimator. Hence, an estimator of the function m of the following type is used:

$$\hat{m}_h(x) = n^{-1} \sum W_h(x; X_1, X_2, \ldots, X_n) \cdot Y_i,$$

where W_h is a weight function depending on the bandwidth parameter h and the set of variables $X_1, \ldots, X_n$.

A conceptually simple approach for the representation of the weight function W_h is to describe its shape by a density function, called the kernel function, with a scale parameter h, the bandwidth, that adjusts the size and the form of the weights near x. Therefore, kernel regression estimators are local weighted averages of the response variable whose weights are determined by the kernel function K, while the size of the weights depends on the bandwidth parameter h.

Generally, the kernel function K has the fundamental properties of a probability density. In the regression context the kernel function is generally a smooth, symmetric, positive function which peaks at zero and decreases monotonically as the bandwidth parameters increases in size.

Several formulae have been proposed for the kernel estimator $\hat{m}$ of the regression mean function m, depending on the type of the kernel regression estimator used. An extensive presentation of these formulae is provided in Peristera and Kostaki(2005). Among the alternative estimators, Peristera and Kostaki (2005) have shown that the Gasser-Müller estimator (Gasser and Müller 1979, 1984) has proved the most adequate alternative in the context of mortality graduation.

At a point x, the Gasser-Müller estimator is given by the following formula,

$$\hat{m}_{GM}(x) = \sum_{i=1}^{n} Y_{[i]} \int_{(x_{(i)}+x_{(i-1)})/2}^{(x_{(i+1)}+x_{(i)})/2} K_h(x - x_i)dx,$$

where $x_0 = -\infty$, $x_n = +\infty$, $x_{(i)}$ denotes the ith largest value of the observed covariate values and $Y_{[i]}$ is the corresponding response value.

The appropriate selection of the bandwidth parameter is of great importance since it controls the degree of smoothness, and consequently influences the resulting estimator. A presentation of bandwidth selection techniques can be found in Härdle (1990, 1991), and Peristera and Kostaki (2005). An approach for the selection of the bandwidth parameter is to construct a direct plug-in estimator of the optimal smoothing parameter h_{opt}. Gasser, Kneip and Kohler (1991) give expressions for the h_{opt} appropriate to the Gasser-Müller estimator and describe how the unknown quantities can be effectively estimated. An important issue for the selection of bandwidth is the choice between a global or a local

one. Local bandwidth selection permits one to obtain a bandwidth that adapts for local efficiency in different parts of the design points, which means that a smaller bandwidth is used in areas of high density while the value of the bandwidth increase in areas of low density. Brockmann, Gasser and Herrmann (1993) and Herrmann (1997) have mentioned the advantage of using kernel regression estimators with a local bandwidth instead of a global one. The main idea of the plug-in method is to estimate the optimal bandwidths by estimating the asymptotically optimal mean integrated squared error bandwidths. For the selection of a local bandwidth, Herrmann (1997) developed an iterative plug-in algorithm that is a generalization of the global iterative plug-in algorithm of Gasser, Kneip and Kohler (1991). A description of this algorithm can be found in Herrmann (1997), in which the advantage of this approach over both the cross-validation method and the global plug-in rule, is highlighted.

4. Support Vector Machines

Support Vector Machines (SVMs) appeared in the middle nineties in the framework of Vapnik's Statistical Learning Theory (Vapnik 1995; Moguerza and Muñoz 2006), providing very successful results for the smoothing of noisy data such as neighbourhood curves (Muñoz and Moguerza 2005) or nonlinear profiles (Moguerza, Muñoz and Psarakis 2007). Support Vector Machines are regularization methods. These methods also include Splines (Moguerza and Muñoz 2006). In fact, there is a close relation between SVM and splines (Pearce and Wand 2006). Next we provide a description of the regression version of SVM and its main features.

4.1 Regularization Theory

Regularization methods (Tikhonov and Arsenin 1977), allow the construction of smooth functions by solving an optimization problem of the form:

$$\min_{f \in H_K} \frac{1}{p} \sum_{i=1}^{p} L(f(x_i) - y_i) + M\|f\|_K^2,$$

where $(x_i, y_i), i = 1, \ldots, p$ are a set of data with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, L is a loss function, $M > 0$ is a constant, H_K is a Reproducing Kernel Hilbert Space[5] (RKHS, see Aronszajn 1950; Moguerza and Muñoz 2006) generated from a kernel $K : X \times X \longrightarrow \mathbb{R}$ (for instance, the space X may be defined as $\mathbb{R}^n$), and $\|f\|_K$ is the norm of f in the RKHS. Notice that, for a fixed value of z, $K(x, z)$ defines a function of x. Roughly speaking, a

[5] Wikipedia has readable descriptions of both Hilbert Spaces and Reproducing Kernel Hilbert Spaces.

RKHS is a space made up of linear combinations of functions $K(x, z_i)$, and their limits. For the case of regression SVM, the loss function L is defined as:

$$L(x) = \begin{cases} |x| - \varepsilon, & \text{if } |x| \geq \varepsilon, \\ 0 \qquad\quad , & otherwise, \end{cases}$$

where $\varepsilon > 0$ is a constant. The idea is to find a smooth function $f^* \in H_K$ that solves the optimization problem above. This function, which, as already stated, belongs to the RKHS H_K, will have the form $f^*(x) = \sum_{i=1}^{p} \alpha_i K(x, x_i) + b^*$, where α_i and b^* are constants, $K(x, y) = \Phi(x)^T \Phi(y)$ is the kernel function that generates H_K and $\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is a mapping defining K. In this way, geometrically, Φ maps the data from the so-called "input space" (that is, $\mathbb{R}^n$) into the "feature space" (that is, $\mathbb{R}^m$). The constant $M > 0$ penalizes non-smoothness of the possible solutions to the problem.

4.2 Geometrical Interpretation of Support Vector Machines

Although the previous formulation is the one that provides the best theoretical properties, from a practical point of view regression SVM can be presented from its geometrical interpretation. It can be shown (Moguerza and Muñoz 2006) that the regularization problem can be formulated as a convex quadractic optimization problem (therefore, without local minima) of the form:

$$\begin{aligned} & \min_{w,b,\xi,\xi'} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{p} \left(\xi_i + \xi_i'\right) \\ \text{such that} \quad & \left(w^T \phi(x_i) + b\right) - y_i \leq \varepsilon - \xi_i, \quad i = 1, \ldots, p, \\ & y_i - \left(w^T \phi(x_i) + b\right) \leq \varepsilon - \xi_i', \quad i = 1, \ldots, p, \\ & \xi_i, \xi_i' \geq 0, \qquad\qquad\qquad\qquad i = 1, \ldots, p, \end{aligned}$$

where ξ_i and ξ_i' are slack variables which permit the violation of a boundary determined by ε. It can be shown (see Moguerza and Muñoz 2006, for the details) that $f^*(x) = \sum_{i=1}^{p} \alpha_i K(x, x_i) + b = (w^*)^T \Phi(x) + b^*$, where w^* and b^* are the values of w and b at the solution of the quadratic optimization problem. One of the key issues of SVM is how to use $\phi(x)$ to map the data into a higher-dimensional space. To achieve this task, a kernel approach is used in order to operate in the feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. Three are the most widely used kernels: the linear kernel $K(x, y) = x^T y$, which corresponds to the identity mapping; the polynomial kernel $K(x, y) = \left(c + x^T y\right)^d$, where c and d are

constants, which maps the data into a finitely dimensional space; and the Gaussian kernel $K(x,y) = e^{\frac{-\|x-y\|^2}{\sigma}}$, where σ is a positive constant which maps the data into an infinitely dimensional space. The Gaussian kernel, given its approximation capacity, is the most extensively used (see Moguerza and Muñoz 2006, for a complete set of examples). In practice, the optimization problem to solve is not the primal formulation shown above. For practical purposes, the problem to solve is the "dual problem" (Schölkopf et al. 2000), that is:

$$\max_{\lambda,\lambda'} -\frac{1}{2}\sum_{i,j=1}^{p}(\lambda_i-\lambda'_i)(\lambda_j-\lambda'_j)K(x_i,x_j) - \varepsilon\sum_{i=1}^{p}(\lambda_i+\lambda'_i) + \sum_{i=1}^{p} y_i(\lambda_i-\lambda'_i)$$

$$\text{such that} \sum_{i=1}^{p}(\lambda_i-\lambda'_i) = 0,$$
$$0 \le \lambda_i \le C, \quad i = 1,\ldots,p,$$
$$0 \le \lambda'_i \le C, \quad i = 1,\ldots,p.$$

It can be shown that both problems, primal and dual, are equivalent, and that:

$$f^*(x) = \sum_{i=1}^{p}(\lambda_i^*-\lambda_i'^*)K(x,x_i) + b^* = \sum_{i=1}^{p}\alpha_i K(x,x_i) + b^*,$$

where $\alpha_i = \lambda_i^* - \lambda_i'^*$, λ_i^* and $\lambda_i'^*$ being the values of λ_i and λ'_i at the solution of the dual problem. Therefore, in practice, the estimated parameters are the α coefficients, whose number is p, that is, the number of data. In this way, the relationship between kernels and SVM is clear: only the closed form of the kernel K is needed, and not the explicit mapping Φ. Notice that this distinctive peculiarity allows, for instance, the use of the Gaussian Kernel in order to evaluate $f^*(x)$. Moreover, in practice, only a small percentage of the α coefficients will differ from zero, which makes simpler the evaluation of this function (this is one of the advantages of SVM, see Moguerza and Muñoz 2006), and reduces the number of estimated parameters.

5. Results

In order to evaluate the performance of SVM techniques in graduating age-specific fertility patterns, we apply this technique as well as kernels, while we also fit the parametric models mentioned above to period single year age-specific fertility rates for the populations of Sweden 1996 and 2000, Norway 1992 and 2000, Denmark 1992 and 2000, Belgium 1993 and 1995, Greece 1995 and 2000, Italy 1995 and 2000, UK 1992 and 2000,

and Ireland 1995 and 2000, as well as for the white and black populations of the USA 2003 and the cohorts of 1942 and 1963 for Spain The empirical data sets were obtained from *Eurostat New Cronos* database. Additionally, single year age-specific fertility rates for the US were derived from the 2003 *Natality Data Set*, obtained by request from the US National Centre of Health Statistics. Cohort data for Spain for the generations born from 1942 to 1963, obtained from the *Eurostat New Cronos* database. It should be noted that even for cohorts not yet completed, *Eurostat* provides estimates of the fertility rates for older women by using the rates observed for previous generations, without waiting for the cohort to reach the end of the reproductive period.

The fits of the parametric models presented at the outset were initially calculated by Peristera and Kostaki (2007). In populations with no apparent early-age hump, the Hadwiger, Gamma, Beta, P-K, and quadratic spline models (Schmertmann 2003) are fitted, while in cases of distorted fertility distributions, the Hadwiger (Chandola, Coleman, and Horns 1999, 2002) and the P-K mixture models (Peristera and Kostaki 2007) are fitted. The simple models used previously in data sets without distortions had a rather disappointing performance in the distored data sets.

In order to avoid heterogeneity we also use data differentiated by order of birth, and cohort and period data sets. Finally in the case of the USA, the fits of the alternative models are provided for the white and black population separately.

In order to fit the alternative parametric models, it is generally accepted that the most efficient procedure is to use weighted least squares, with weights equal to the reciprocals of the variances of the empirical rates. However, as pointed out by Hoem (1976) and Hoem et al. (1981), a weighted estimating procedure would give too much attention to the low fertility ages in the tails and especially to the higher ones in the upper tail, while giving too little attention to the high fertility ages in the middle, and thus is not desirable. Therefore, for the estimation of the parameters of the various models, a non-linear unweighted least-squares procedure is adopted. The models are fitted by means of a Gauss-Newton optimization scheme. The Matlab built-in routine for non-linear parameter estimation *lsqnonlin* is used in order to find the unconstrained minimum of the unweighted residual sum of squares.

The quadratic Spline estimates are obtained using the program provided by Schmertmann (2003) at the web page http://mailer.fsu.edu/ schmert/qsfit/qsfit.htm.

For kernel applications, the subroutine "lokerns" of the library "lokern" for the R-package is used for the calculation of Gasser-Müller estimators with local bandwidth parameter. This is available in http://www.unizh.ch/biostat/software. In order to select bandwidth for a local linear Gaussian kernel regression estimator, trials are made using a direct plug-in technique (Ruppert, Sheather, and Wand 1995). There, we use the KernSmooth library and the R package. However, this methodology has been discarded given the overfitting observed. Therefore, the bandwidth parameter has been computed

by cross-validation leading to a value of 1.9066 for all the estimated curves. In this way, we have a unique model for all the data sets.

For the SVM techniques, the subroutine *svm* of the library *e1071* for the R-package is used. This is available in http://cran.r-project.org/. A two-step simulation procedure is used to select the parameters ε, σ and C of the ε - regression procedure: ε is used to fix the width of a band around the fitted curved, σ plays the role of a variance, and C is an upper bound for the λ coefficients in the dual optimization problem and, at the same time, penalizes the values of the slacks corresponding to those points lying outside of the band determined by ε in the primal optimization problem. In a first step, the range of parameters ε, σ and C are determined. Then in a second step, the best combination of the three parameters is computed using R flow sentences. In particular, the values $\varepsilon = 0.0001$, $\sigma = 40$ and $C = 1.8$, have been chosen for the SVM implementation. Additionally, in this application, the values for the corresponding dimensions in the SVM model are $n = 1$, $m = \infty$ (given that this is the dimension induced by the Gaussian kernel, see Moguerza and Muñoz 2006) and $p = 34$, that is, the number of data within each set. We should notice here again that we use the same set of parameter values for all the data sets. In this way, we are able to make fair comparisons of these results with those produced by kernels.

The values of the sums of squares of the differences between the empirical and the resulting values for all the data sets used, and all graduation techniques used, are provided in Tables 1 and 2. The results of fitting the parametric models were first presented in Peristera and Kostaki (2007).

Figures 1-6 provide illustrations for some chosen cases. In all the cases we are using ages ranging from 15 to 48, so each schedule has 34 rates.

Table 1: Values of the minimization criterion multiplied by 100.000, at the exit of the estimation procedure for P-K model, Beta model, Gamma model, Hadwiger model, quadratic Spline model, kernels and SVM

SSE·10^6	**P-K Model**	**Beta Model**	**Gamma Model**	**Hadwiger Model**	**Quadratic Spline Model**	**Kernel**	**SVM**
Period Data							
Sweden							
1996	115	108	132	326	174	67	72
2000	117	181	321	689	174	30	11
Norway							
1992	242	175	265	656	263	65	61
2000	233	225	640	329	287	40	10
Denmark							
1992	103	107	130	383	169	54	20
2000	225	363	575	1073	287	51	6
Belgium							
1993	401	396	380	540	462	68	15
1995	346	374	376	558	525	78	30
Greece							
1995	190	137	184	289	101	26	14
2000	34	114	491	617	55	14	13
Italy							
1995	20	58	139	352	49	18	11
2000	47	71	524	908	82	14	3
Cohort Data							
Spain							
1943	732	1005	1159	1547	5450	452	562
1962	295	259	1113	184	3720	69	67

Table 2: Values of the minimization criterion, multiplied by 100.000, at the exit of the estimation procedure for P-K mixture model, Hadwiger mixture, kernels and SVM, for the US data

SSE$\cdot 10^6$	**P-K Mixture Model**	**Hadwiger Mixture Model**	**Kernel**	**SVM**
Period Data Total Births				
UK				
1992	154	35	37	14
2000	99	22	40	14
Ireland				
1995	437	97	62	90
2000	78	177	65	43
Spain				
1999	29	17	30	12
2000	23	15	31	6
Cohort Data Total Births				
Spain				
1963	77	85	59	62
Period Data First Births				
UK				
2004	5	8	47	4
Ireland				
2000	73	53	61	62
Period Data Second Births				
UK				
2004	4	5	45	3
Ireland				
2000	31	31	25	28
USA 2003				
Total	150	28	63	58
White	28	156	63	51
Black	39	190	103	86

Figure 1: Observed and estimated period age-specific fertility rates for Denmark, 2000

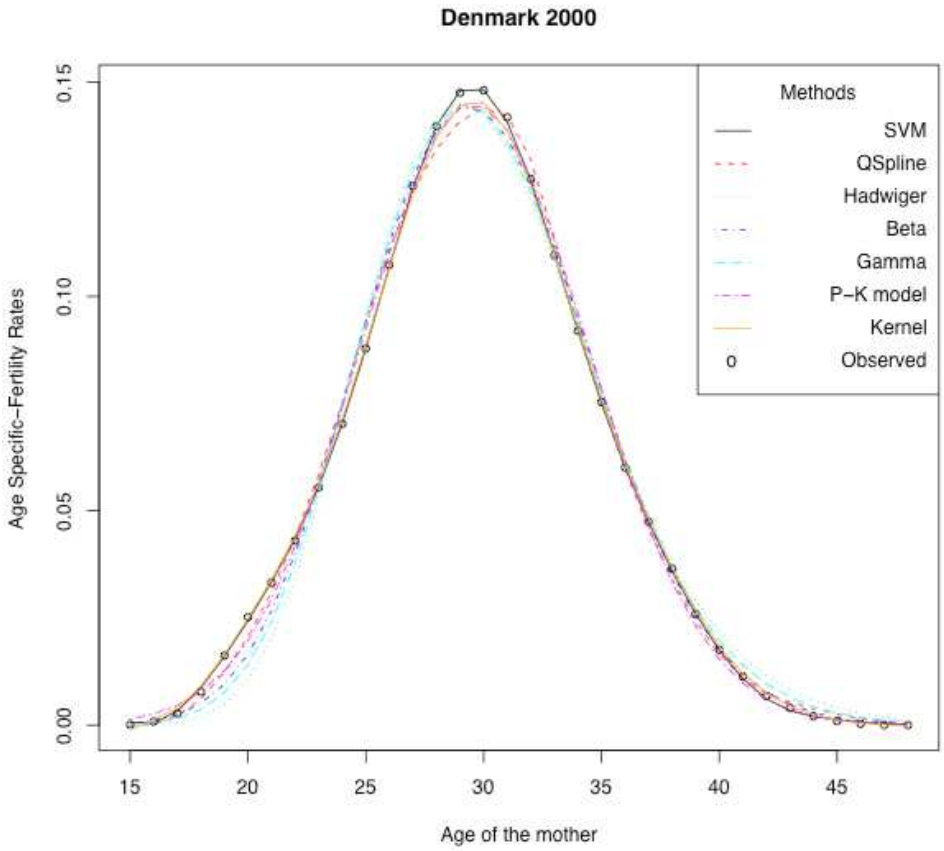


Figure 2: **Observed and estimated cohort age-specific fertility rates for Spain, 1943**

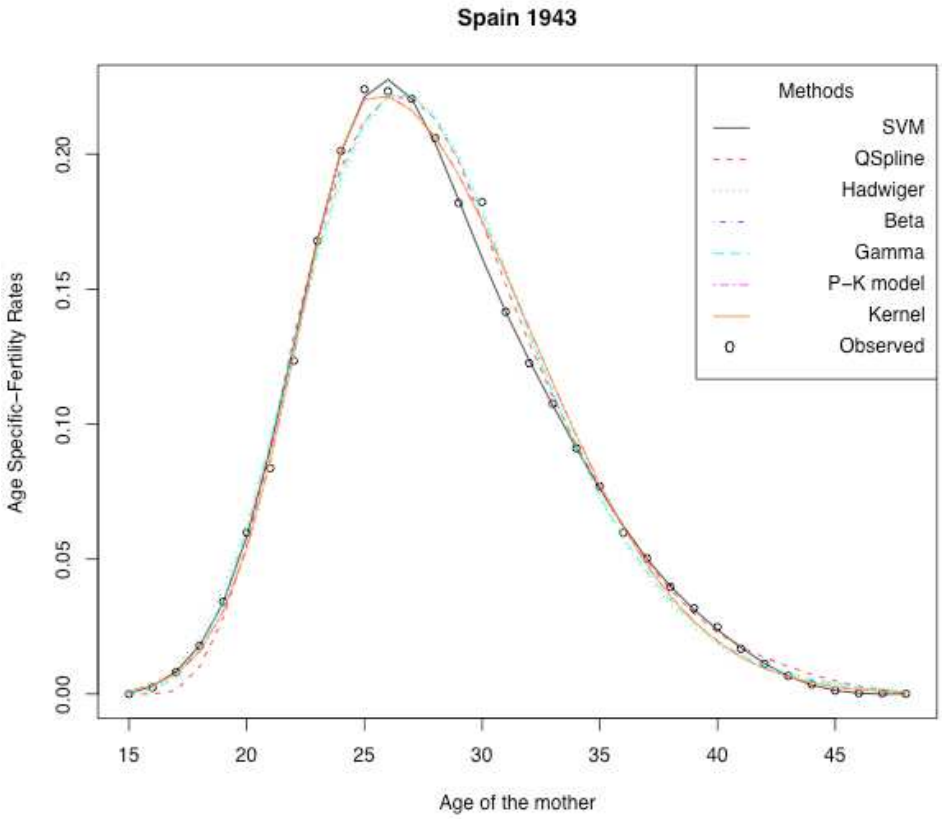


Figure 3: Observed and estimated cohort age-specific fertility rates for Spain, 1963

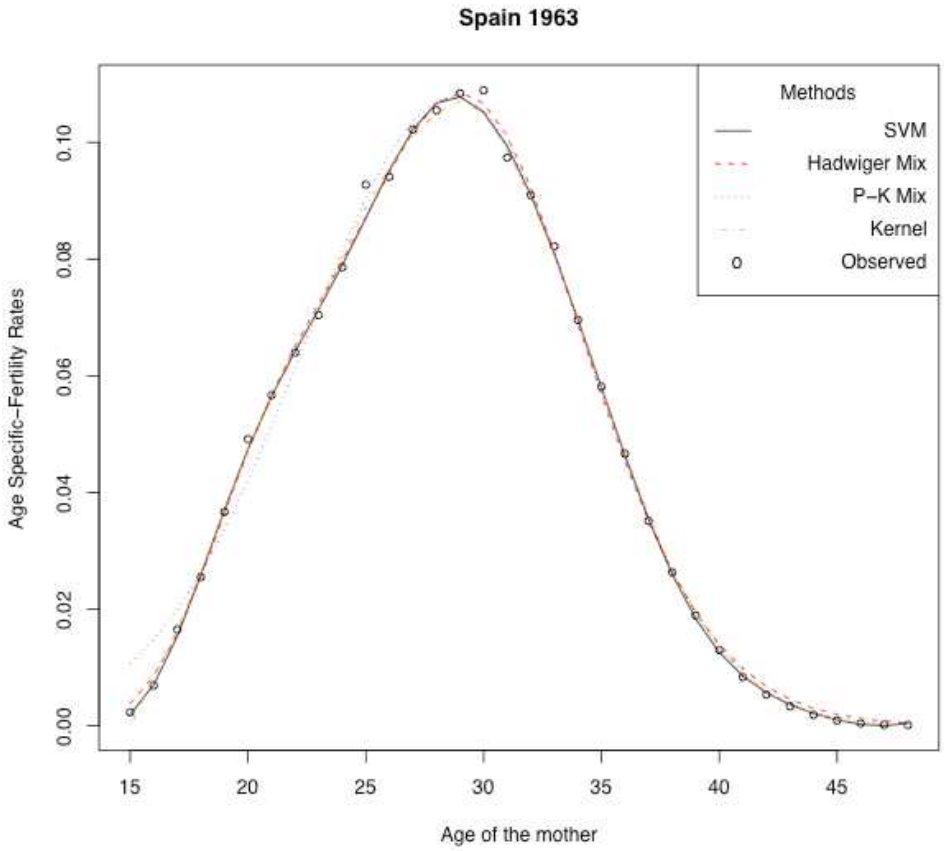


Figure 4: **Observed and estimated age-specific fertility rates of Ireland, 2000. First births.**

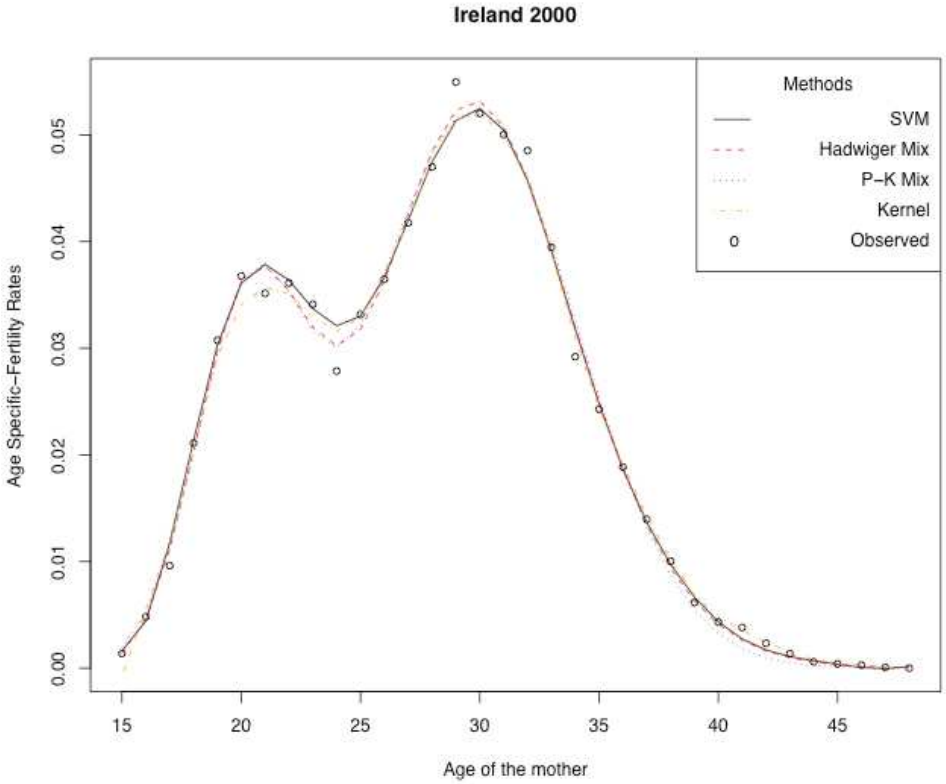


Figure 5: **Observed and estimated age-specific fertility rates of US, 2003. White population.**

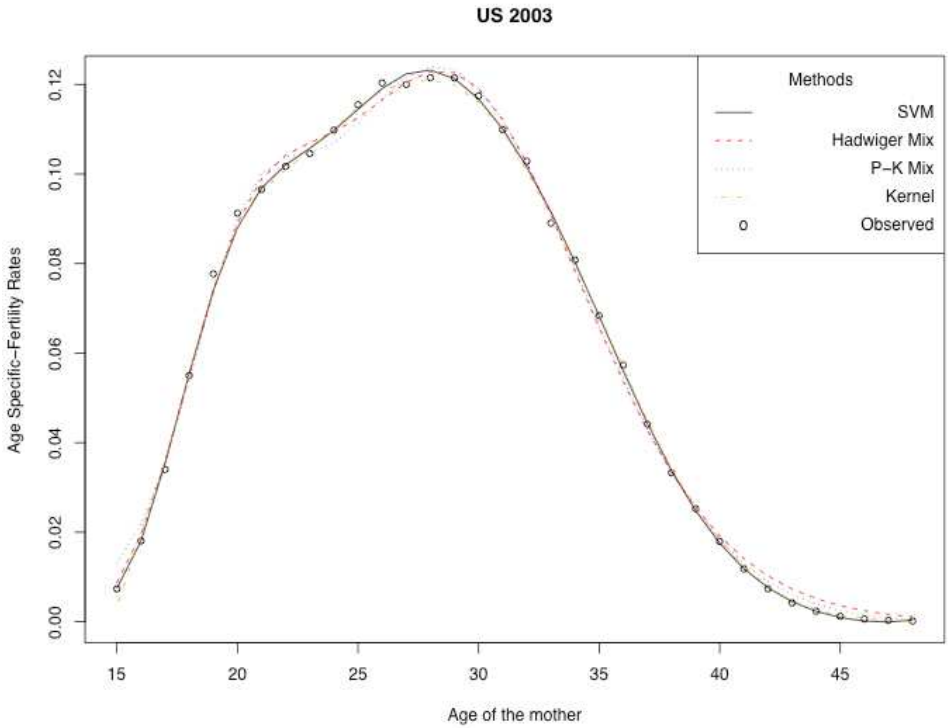
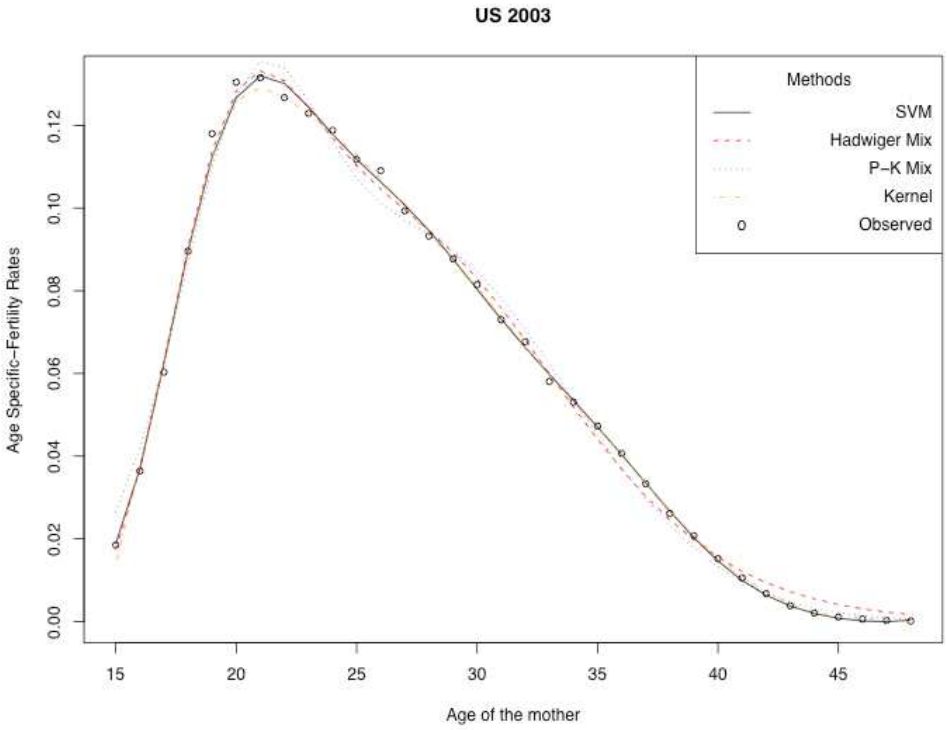


Figure 6: Observed and estimated age-specific fertility rates of US, 2003. Black population.



6. Findings

In this paper we propose the application of Support Vector Machines for graduating age-specific fertility rates. In order to evaluate the performance of SVM we apply this technique to a variety of empirical cohort and period data sets of alternative populations. In addition, for comparison reasons, we also fit parametric models and apply kernels to these data sets.

According to the values of the minimizing criterion, the results for the two non parametric techniques are apparently closer to the empirical values than those provided by the parametric models. This can partly depend on the fact that parametric models provide highest smoothness. A higher degree of smoothness might result from larger distances between the empirical and the graduated values. Turning now to the comparison between the two non parametric techniques, the results provided by the SVM are in most cases associated with lower values of the minimizing criterion.

As is obvious in tables and figures, SVM show a successful performance in graduating the empirical rates in both simple and distorted data sets, producing results that, in the vast majority of cases, are closer to the empirical rates than the other methods. Regarding the figures, one can observe that especially for the ages in the peak and the tails of the fertility curve, the results of SVM were closer to the empirical values than those of most of the other methods.

An advantage of non parametric graduation techniques in comparison with the parametric modeling is that these can be adequately applied to all data sets, while in data sets with distorted patterns, the use of standard models is inadequate and more complicated formulae are required. Furthermore, the regulation of the degree of smoothness by the user can also be considered as an advantage, allowing the user to choose the optimal degree of smoothness depending on the purpose of graduation at hand and also avoiding oversimplification of age patterns.

Regarding future extensions of this work, SVM can be easily used as a multivariate model, providing a promising area for further research in demographic problems.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* 68: 337–404. doi: 10.2307/1990404.

Brockmann, M., Gasser, T., and Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *Journal of the American Statistical Association* 88(424): 1302–1309. doi: 10.2307/2291270.

Chandola, T., Coleman, D. A., and Horns, R. W. (1999). Recent european fertility patterns: fitting curves to 'distorted' distributions. *Population Studies* 53(3): 317–329. doi: 10.1080/00324720308089.

Chandola, T., Coleman, D. A., and Horns, R. W. (2002). Distinctive features of age-specific fertility profiles in the English-speaking world: Common patterns in Australia, Canada, New Zealand and the United States, 1970-98. *Population Studies* 56: 181–200. doi: 10.1080/00324720215929.

Eurostat New Cronos Database (2006). *Europa database: Population and Social Conditions (electronic resource)*. Manchester UK. http://www.esds.ac.uk/international/support/user_guides/eurostat/Cronoswe.asp.

Gasser, T., Kneip, A., and Kohler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association* 86(415): 643–652. doi: 10.2307/2290393.

Gasser, T. and Müller, H. (1979). Kernel estimation of regression functions. In: *Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics 757*, pp. 23–68. New-York: Springer-Verlag. doi: 10.1007/BFb0098489.

Gasser, T. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics* 11: 171–185.

Gilje, E. (1969). Fitting curves to age-specific fertility rates: some examples. *Statistical Review of the Swedish National Central Bureau of Statistics III* 7: 118–134.

Hadwiger, H. (1940). Eine analytische reprodutions-funktion fur biologische gesamtheiten. *Skandinavisk Aktuarietidskrift* 23: 101–113.

Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics* 6(1): 35–54. doi: 10.2307/1390723.

Hoem, J. M. (1976). The statistical theory of demographic rates: A review of current developments (with discussion). *Scandinavian Journal of Statistics* 3: 169–185.

Hoem, J. M., Madsen, D., Nielsen, J. L., Ohlsen, E., Hansen, H. O., and Rennermalm,

B. (1981). Experiments in modelling recent Danish fertility curves. *Demography* 18: 231–244. doi: 10.2307/2061095.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S*. New York: Springer-Verlag.

Moguerza, J. and Muñoz, A. (2006). Support vector machines with applications. *Statistical Science* 21(3): 322–336. doi: 10.1214/088342306000000493.

Moguerza, J., Muñoz, A., and Psarakis, S. (2007). Monitoring nonlinear profiles using support vector machines. *Lecture Notes in Computer Science* 4789: 574–583. doi: 10.1007/978-3-540-76725-1_60.

Muñoz, A. and Moguerza, J. (2005). Building smooth neighbourhood kernels via functional data analysis. *Lecture Notes in Computer Science* 3697: 631–636. doi: 10.1007/11550907.

Pearce, N. and Wand, M. (2006). Penalized splines and reproducing kernel methods. *The American Statistician* 60(3): 233–240. doi: 10.1198/000313006X124541.

Peristera, P. and Kostaki, A. (2005). An evaluation of the performance of kernel estimators for graduating mortality data. *Journal of Population Research* 22(2): 185–197. doi: 10.1007/BF03031828.

Peristera, P. and Kostaki, A. (2007). Modeling fertility in modern populations. *Demographic Research* 16(6): 141–194.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90: 1257–1270. doi: 10.2307/2291516.

Schölkopf, B., Smola, A., Williamson, R., and Bartlett, P. (2000). New support vector algorithms. *Neural Computation* 12: 1207–1245. doi: 10.1162/089976600300015565.

Schmertmann, C. P. (2003). A system of model fertility schedules with graphically intuitive parameters. *Demographic Research* 9(5): 82–110. doi: 10.4054/DemRes.2003.9.5.

Tikhonov, A. and Arsenin, V. (1977). *Solutions of ill-posed problems*. John Wiley and Sons.

US National Centre of Health Statistics (2003). Natality data set (electronic resource). http://www.cdc.gov/nchs/. Maryland.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.