Canadian Studies in Population 38, No. 3–4 (Fall/Winter 2011):37–58.

Support Vector Machines as tools for mortality graduation

Anastasia Kostaki
Department of Statistics
Athens University of Economics and Business
kostaki@aueb.gr

Javier M. Moguerza
Department of Statistic and Operational Research
Rey Juan Carlos University, Madrid

Alberto Olivares
Department of Statistic and Operational Research
Rey Juan Carlos University, Madrid

Stelios Psarakis
Department of Statistics
Athens University of Economics and Business

Abstract

A topic of interest in demographic and biostatistical analysis as well as in actuarial practice, is the graduation of the age-specific mortality pattern. A classical graduation technique is to fit parametric models. Recently, particular emphasis has been given to graduation using non-parametric techniques. Support Vector Machines (SVM) is an innovative methodology that could be utilized for mortality graduation purposes. This paper evaluates SVM techniques as tools for graduating mortality rates. We apply SVM to empirical death rates from a variety of populations and time periods. For comparison, we also apply standard graduation techniques to the same data.

Keywords: *mortality pattern, graduation techniques, support vector machines, kernel regression estimators.*

Résumé

L'ajustement des modèles de mortalité par âge est un sujet d'intérêt à la fois en analyse démographique et biostatistique et en pratique actuarielle. Une technique d'ajustement classique consiste à adapter les modèles paramétriques. Dernièrement, on accorde une attention spéciale à l'ajustement au moyen de techniques autres que les techniques paramétriques. Les machines vectorielles de support (SVM Support Vector Machines) représentent une méthode novatrice pouvant servir à l'ajustement des taux de mortalité. Cet article évalue ces techniques en tant qu'outils d'ajustement de taux de mortalité. C'est ainsi que nous utilisons ces techniques pour les taux de mortalité empiriques de plusieurs populations et périodes. À des fins de comparaisons, nous utilisons les techniques d'ajustement normales pour les mêmes données.

Mots-clés : *modèle de mortalité, techniques d'ajustement, machines vectorielles de soutien, estimateurs de régression kernel.*

Introduction

Representing the age-specific mortality pattern of a population is of particular interest in demographic analysis, biostatistics, and actuarial practice. For nearly two centuries demographers, biostatisticians, actuaries, and social workers have shown great interest in the means of representing the age-specific mortality patterns of populations. Demographers want to describe and project the mortality pattern of a population for the purpose of mortality analysis, as well as to provide population projections. Biostatisticians need a basis for making mortality forecasts. Actuaries need a mortality basis suitable for calculations in life insurance and in designing social security systems. Social planning also requires estimations and projections of age-specific mortality.

In order to estimate the unknown age-specific probabilities of dying that underline the empirical measures, we can use graduation techniques applied to empirical death rates, under the assumption that the true probabilities follow a smooth pattern through age. For the purpose of graduation, several parametric and non-parametric techniques have been proposed. Parametric functions of age, commonly known in demography as *mortality laws,* have been in use for more than a century. The earliest attempt to provide such a formula was by de Moivre in 1725, while the most widely known law of mortality was proposed by Gompertz in 1825. Keyfitz (1982) provides a review of these historical laws. In modern times, many authors have contributed to the theory of parametric models of mortality, and to the problem of estimating their parameters (e.g., Heligman and Pollard 1980; Keyfitz 1982; Forfar et al. 1988; Kostaki 1992; Hannerz 1999; Karlis and Kostaki 2000).

Recently the utilization of non-parametric smoothing techniques for graduation purposes has gained attention. Among these techniques, special attention is given to kernels (Cobas and Haberman 1983). An evaluation of kernels as tools for graduating the mortality pattern is provided by Kostaki and Peristera (2005).

Support Vector Machines (SVM) is a modern non-parametric graduation methodology 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 et al. 2007), they can probably serve as an equally useful tool for mortality graduation purposes. Regarding demographic data, SVM have shown an interesting performance when applied to the graduation of age-specific fertility patterns (Kostaki et al. 2009). These techniques are easy to adjust, which implies they can be easily applied by demographers who may lack a thorough background on Statistical Learning Theory or pattern recognition.

This work provides an evaluation of the SVM methodology in the context of mortality graduation. Section 2 provides a summary description of proven graduation techniques, i.e., kernels and parametric models. Section 3 is devoted to a presentation of the SVM methodology. Then, in Section 4 an evaluation is provided of the utilization of SVM methodology for the graduation of age-specific death rates. Namely, we apply SVM to empirical death rates for several populations and time periods. Additionally, for comparison purposes, kernels are also applied, and the Heligman-Pollard model (Heligman and Pollard 1980), is fitted to the same datasets. Finally, in Section 5 some concluding remarks are provided.

Graduation techniques

Laws of mortality

Parametric modeling is widely used in demography for graduation purposes since it provides results with the highest degree of smoothness. Detailed presentations of the features of parametric models are given by Keyfitz (1982), Kostaki (1992), Congdon (1993), and Karlis and Kostaki (2000). A huge variety of mortality laws has been presented in the literature since 1725. Among them, the most successful attempt to describe the mortality pattern for total life span through a parametric model might be the one proposed by Heligman and Pollard (1980). This model is described by the formula

$$\frac{q_x}{p_x} = A^{(x+B)^C} + De^{-(E(\ln x - \ln F)^2} + GH^x \ ,$$

where q_x is the probability of dying within a year, $p_x = (1 - q_x)$, and A to H are parameters to be estimated. It includes eight parameters, all of them having demographic interpretation. The first additive term of the right-hand side of the formula describes mortality of the childhood ages. It includes three parameters: *A,* which reflects the level of childhood mortality; *C*, related to the rate of mortality decrease in childhood ages; and *B*, which is indicative of the mortality level at age zero. The middle term reflects accident mortality and it also includes three parameters: *D*, related to the severity of the accident hump; *E*, related to its spread; and *F*, indicating the location of the hump. Finally, the third term includes two parameters: *G*, reflecting the level of later adult mortality; and *H*, related to the rate of mortality increase at the later adult ages.

Heligman and Pollard (1980) estimated these parameters using a least-squares approach, in order to minimize the sum of squares

$$S^2 = \sum_z \left[\frac{\hat{q}_x}{q_x} - 1 \right]^2 ,$$

where $\hat{q}_x$ is the fitted value at age x and q_x is the observed mortality rate.

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:

$$y_i = m(x_i) + \varepsilon_{i,} \quad i = 1, \ldots, p \ ,$$

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:

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

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 to representing the weight function W_h is to describe its shape by a density function called the *kernel function*, with a scale parameter h, i.e., the bandwidth, which 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, positive function, which peaks at zero and decreases monotonically as the bandwidth parameter 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 Kostaki and Peristera (2005). Among the alternative estimators, Kostaki and Peristera (2005) have shown that the one by Gasser-Muller (Gasser and Muller 1979; 1984) has proved the most adequate in the context of mortality graduation.

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

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

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

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 Hardle (1990; 1991) and Kostaki and Peristera (2005). One approach to selecting the bandwidth parameter is to construct a direct plug-in estimator of the optimal smoothing parameter h_{opt}. Gasser et al. (1991) give expressions for the h_{opt} appropriate to the Gasser-Muller estimator, and describe how the unknown quantities can be effectively estimated. An important issue for the selection of bandwidth is the choice between global and local. Local bandwidth selection allows obtaining a bandwidth that adapts for local efficiencies 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 increases in areas of low density. Brockmann et al. (1993) and Hermann (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, Hermann (1997) developed an iterative plug-in algorithm that is a generalization of the global iterative plug-in algorithm of Gasser et al. (1991). A description of this algorithm can be found in Hermann (1997), where the advantage of this approach over the cross-validation method and the global plug-in rule is highlighted.

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 et al. 2007). Support Vector Machines are part of regularization methods that also include *Splines* (Moguerza and Muñoz 2006). In fact, there is a close relation between both methodologies, SVM and Splines (Pearce and Wand 2006). Next we provide a description of the regression version of SVM and its main features.

Support Vector Machines for regression

Presenting the geometrical interpretation of SVM for regression, we note that from a practical point of view, regression SVM can be formulated as a convex quadratic optimization problem (therefore, without local minima) of the form

$$\begin{aligned}
&\min_{w,b,\xi,\xi'} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{p}\left(\xi_i + \xi_i'\right)\\
&\text{s. t. } \left(w^T\phi\left(x_i\right)+b\right)-y_i \le \varepsilon - \xi_i, \ \ i=1,\dots,p,\\
&\qquad y_i-\left(w^T\phi\left(x_i\right)+b\right) \le \varepsilon - \xi_i', \ \ i=1,\dots,p,\\
&\qquad \xi_{\mathrm{i}}, \xi_i' \ge 0, \qquad\qquad\qquad\quad i=1,\dots,p,
\end{aligned}$$

where (x_i, y_i), $i=1,\dots,p$ are a set of data with $x_i \in R^n$ and $y_i \in R$, ξ_i, and ξ'_i are slack variables which permit the violation of a boundary determined by ε. $\Phi\colon R^n \rightarrow R^m$ is a mapping defining the kernel function $K{:}X \times X \rightarrow R$ (for instance, the space X may be defined as R^n), such that $K(x,y) = \Phi(x)^T\Phi(y)$. In this way, geometrically Φ maps the data from the so-called "input space" (that is, R^n) into the "feature space" (that is, R^m). 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." The three most widely used kernels are: the linear kernel $K(x,y) = x^Ty$, which corresponds to the identity mapping; the polynomial kernel $K(x,y) = (c + x^Ty)^d$, where c and d are constants, which maps the data into a finitely dimensional space; and the Gaussian kernel

$$K\left(x,y\right) = e^{\frac{-\|x-y\|^2}{\sigma}},$$

where σ is a positive constant, which maps the data into an infinitely dimensional space. The role of the kernel is crucial within the SVM methodology. Depending on the kernel used, the approximation capacity of the methodology will be different. In this way, the linear kernel (the simplest one) will be useful for the approximation of linear functions, while the Gaussian kernel will be suitable for the approximation of nonlinear functions. Given its approximation capacity, the Gaussian kernel is the most extensively used in the literature (for a complete set of examples, see Moguerza and Muñoz 2006).

It can be shown (see Moguerza and Muñoz 2006) that

$$f^*(x)=\sum_{i=1}^{p}\alpha_i K(x,x_i)+b=\left(w^*\right)^T\Phi(x)+b^*,$$

where w^* and b^* are the values of w and b at the solution of the quadratic optimization problem. 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:

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

$$\text{s. t.}\ \sum_{i=1}^{p}\left(\lambda_i-\lambda_i^{'}\right)=0,$$

$$0\le\lambda_i\le C,\ i=1,\ldots,p,$$

$$0\le\lambda_i^{'}\le C,\ i=1,\ldots,p.$$

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

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

where $\alpha_i=\lambda_i^*-\lambda_i'^*$, being λ_i^* and $\lambda_i'^*$ 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.

Piecewise Support Vector Machine (PSVM)

The standard SVM described above can be specialized in order to treat functions whose derivatives take large values within some intervals of the range of support values, and small values within other intervals of the range of support values. With this aim we define the Piecewise Support Vector Machine (PSVM) method. The key point of his method is to train a SVM for each predefined interval, and then calculate the breakpoints between intervals as a function of the piecewise smoothers. In the case of mortality data, two intervals of the same length have been considered in order to divide age x. The first interval corresponds to the subset of the curve domain with stationary points, that is, points where the first derivative equals zero. The second interval corresponds to the subset of the curve domain where the function has an increasing behaviour, that is, where the first derivative of the curve is approximately constant. Suppose $x\in[l,u]$, where l and u denote the lower and upper ages; we then compute $f^*=f_l^*+f_b^*+f_2^*$, where f_l^* equals the SVM solution for $x\in[l,x_b)$ and equals 0 otherwise; f_2^* equals the SVM solution for $x \in (x_b, u]$, and equals 0 otherwise; and $f_b^* = \Gamma(f_l^*, f_2^*)$ for the breakpoint x_b and equals 0 otherwise, where Γ is computed as a function of f_l^* and f_2^*.

Evaluation and comparisons

Our calculations are based on the empirical age-specific mortality rates of the male and female populations of Sweden, for the periods 1981–5, 1984–8, and 1991–5, as well as France and Japan for the years 1990, 1991, and 1995. The Swedish datasets are taken from Statistics Sweden, while the French and Japanese ones are parts of the Berkeley Mortality Database, available in the web via the address http://www.demog.berkeley.edu/wilmoth/mortality.

For kernel applications, the subroutine "glkerns" of the library "glkern" from the R-package is used for the calculation of Gasser-Müller estimators with bandwidth parameter. This is available at http://www.unizh.ch/biostat/software. In order to select the bandwidth for a Gaussian kernel regression estimator, trials were made using a direct plug-in technique (Ruppert et al. 1995)—in particular, the one implemented in the KernSmooth library—and the R-package. However, this methodology has been discarded given the overfitting observed above. Therefore, the bandwidth parameter has been computed by cross-validation, leading to a value of 2.3849 for all the estimated curves. In this way, we have a unique model for all the datasets.

The parameters in Heligman-Pollard model are estimated using an iterative routine of the Nag library that is based upon a modification of the Gauss-Newton algorithm, described by Gill and Murray (1978).

For the SVM applications, the subroutine "svm" of the library "e1071" of the R-package is used to derive the SVM and the PSVM model parameters. This is available at 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 curve, 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. As a first step, the ranges of parameters ε, σ, and C are determined. Then, in the second step, the best combination of the three parameters is computed using cross-validation techniques. In particular, the values $\varepsilon = 0.02$, $\sigma = 125$, and $C = 2{,}200$ were obtained for the SVM implementation. For the PSVM implementation, the values $\varepsilon = 0.11$, $\sigma = 111.1$, and $C = 3{,}900$ were obtained for the first interval, and values $\varepsilon = 0.008$, $\sigma = 175.4$, and $C = 50$ were obtained for the second interval, while the solution for the breakpoint x_b were calculated as an average function of f_l^* and f_2^*. It can be observed that the parameters for the SVM implementation are approximately an average of the parameters obtained for each interval of the PSVM implementation. The parameters change so drastically between the two intervals because the structure of the curve is significantly different within each interval. In this way, with the PSVM we are able to capture in a better way the local structure of the curves.

In this application, the values for the corresponding dimensions in the SVM model are $n = 1$, $m = 1$ (given that this is the dimension induced by the Gaussian kernel; see Moguerza and Muñoz 2006), and $p = 83$, that is, the number of data within each set. We should note here again that the same set of parameter values is used for all the datasets. In this way, we are able to make fair comparisons of these results with those produced by kernels.

A mortality graduation can be considered successful if the graduated rates progress smoothly from age to age, and at the same time accurately reflect the underlying mortality pattern while avoiding systematic deviations and random variations. In this sense, we are going to evaluate the effectiveness of different adopted approaches for the graduation of our mortality datasets.

Although graphical representation of the observed and the graduated rates is a useful way to derive conclusions, we also use statistical criteria in order to evaluate the performance of the alternative estimators. For that, we use a chi-square criterion to check the closeness of the graduated rates to the observed ones. Then in order to evaluate smoothness of the results we calculate the sum of the absolute values of the third differences for each graduation.

The chi-square criterion, used for evaluating adherence of the results to the observed rates, is defined as

$$\chi_n^2 = \sum_x \frac{E_x}{(1-q_x)q_x}\left(\hat{q}_x - q_x\right)^2 ,$$

where E_x is the exposed-to-risk population at age x, q_x is the observed death rate at age x, $\hat{q}_x$ is the graduated one, and $E_x/[q_x(1-q_x)]$ are the reciprocals of the variances of the observed q_x.

Finally in order to check for smoothness of the resulting probabilities, we examine the third-order differences of the graduated values. We therefore calculate the sum of the absolute values of the third differences in each graduated set of values, i.e., the quantity

$$\sum\left|\Delta^3\hat{q}_x\right| ,$$

multiplied by 100,000 in order to have an easier interpretation of the results.

The values of the two criteria for all the datasets used, and all graduation techniques used, are presented in Tables A1–A3 (Appendix A). Table A4 presents average results for the overall data. Examining these values, one can easily observe that the SVM graduation proves adequate in terms of goodness of fit, as well as in terms of smoothness. Considering the values of χ^2 quantity, for the Swedish and the Japanish datasets, these are in almost all cases lower for the SVM than for the HP8 and kernels. However, for the French datasets the results for the two SVM techniques, and especially those for the PCVM one are clearly superior to those obtained for the other two techniques. Considering the overall values of the χ^2 criterion presented in Table A4, we conclude that both SVM techniques prove superior to the other two methodologies.

Considering smoothness, the values of the sum of third-order differences, in almost all cases, and overall were lower for the two alternative SVM techniques than for the other two methods.

Comparing the values of both SVM and PSVM criteria, we conclude that PSVM proves superior to SVM in terms of goodness of fit. However in terms of smoothness, SVM in many cases provides somewhat better results than PSVM.

Figures B1–B6 (Appendix B) illustrate the results for some chosen cases. As clearly observable in these illustrations, SVM and PSVM show a successful performance, especially in the most difficult parts of the age interval, i.e., the early adult ages. Figures B7–B18 illustrate the results of each technique separately for some chosen cases. It is clear in these figures that the results of the SVM techniques are closer to the empirical data than those of the Heligman-Pollard formula, the latter exhibiting some systematic deviations in the early adult ages. It is also clear that SVM techniques provide better results than kernels regarding both goodness of fit and smoothness.

Remarks

In this paper we proposed the application of Support Vector Machines techniques as tools for graduating age-specific mortality patterns. For evaluation purposes we applied SVM methodology to empirical datasets of a variety of populations and time periods. In addition, for comparison we also applied kernels and fit the Heligman-Pollard formula to the same datasets. The results of our calculations indicate that SVM techniques prove to be adequate, and in most cases superior, to the other two graduation techniques, providing results that are closer to the empirical values when compared to the Heligman-Pollard model and kernels, and smoother than those provided by kernels. An advantage of non-parametric graduation techniques compared to parametric modeling is that these are more flexible and can adequately be applied to all datasets. Meanwhile, in datasets with distorted patterns the use of standard models is inadequate; more complicated formulae are required in such cases. Furthermore, regulation of the degree of smoothness by the user can also be considered 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 easily be used as a multivariate model, providing a promising area for further research on demographic problems.

References

Brockmann, M., T. Gasser, and E. Herrmann. 1993. Locally Adaptive Bandwidth choice for kernel regression estimators. *Journal of the American Statistical Association* 88(424):1302–9.

Copas, J.B., and S. Haberman. 1983. Non-parametric graduation using kernel methods. *Journal of the Institute of Actuaries* 110:135–56.

Forfar, D.O., J.J. McCutcheon, and A.D. Wilkie. 1988. On graduation by mathematical formula. *Journal of the Institute of Actuaries* 115:1–135.

Gasser, T., and H.G. Muller. 1979. Kernel estimation of regression functions, in *Smoothing Techniques for Curve Estimation.* Lecture Notes in Mathematics 757. New York: Springer-Verlag, pp. 23–68.

———. 1984. Estimating regression functions and their derivatives by the Kernel Method. *Scandinavian Journal of Statistics* 11:171–85.

Gasser, T., A. Kneip, and W. Kohler. 1991. A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association* 86(415):643–52.

Gill, P., and W. Marray. 1978. Algorithms for the solution of a non-linear least squares problem. *Journal of Numerical Analysis SIAM* 15:977–92.

Hannerz, H. 1999. *Methodology and Applications of a New Law of Mortality*. Lund (Sweden): Department of Statistics and University of Lund.

Hardle, W. 1990. *Applied Non-parametric Regression*. Cambridge (UK): Cambridge University Press.

———. 1991. *Smoothing Techniques with Implementation in S*. New York: Springer-Verlag.

Hartmann, M. 1987. Past and recent attempts to model mortality at all ages, *Journal of Official Statistics* 3:19–36

Heligman, M., and J.H. Pollard. 1980. The age pattern of mortality. *Journal of the Institute of Actuaries* 107:49–80.

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

Karlis, D., and A. Kostaki. 2000. Bootstrap techniques for mortality models. *Biometrical Journal* 44(7):850–66.

Keyfitz, N. 1982. Choice of the function for mortality analysis: Effective forecasting depends on a minimum parameter representation. *Theoretical Population: Biology* 21:329–52.

Kostaki, A. 1992. Nine-parameter version of the Heligman Pollard Formula. *Mathematical Population Studies* 3(4):277–88.

Kostaki, A., and P. Peristera. 2005. Graduating mortality data using Kernel techniques: Evaluation and comparisons. *Journal of Population Research* 22(2):185–97.

Kostaki, A., J.M. Moguerza, A. Olivares, S. Psarakis. 2009. Graduating the age-specific fertility pattern using Support Vector Machines. *Demographic Research* 20:599–622.

Kronmal, R.A., and M.E. Tarter. 1968. The estimation of probability densities and cumulatives by Fourier series methods. *Journal of American Statistical Association* 63:952.

Lokern library for R software. 1997. www.unizh.ch/biostat/software. Data accessed 2002.

Mack, Y.P. 1981. Local properties of k-NN regression estimates. *SIAM Journal of Algebraic and Discrete Methods* 2:311–23.

Muñoz, A., and J.M. Moguerza. 2005. Building smooth neighbourhood kernels via Functional Data Analysis. *Lecture Notes in Computer Science* 3697:631–6.

Moguerza, J.M., and A. Muñoz. 2006. Support Vector Machines with applications. *Statistical Science* 21(3):322–36.

Moguerza, J.M., A. Muñoz, and S. Psarakis. 2007. Monitoring nonlinear profiles using Support Vector Machines. *Lecture Notes in Computer Science* 4789:574–83.

Pearce, N.D., and M.P. Wand. 2006. Penalized splines and reproducing kernel methods. *The American Statistician* 60(3):233–40.

Ruppert, D., S.J. Sheather, and M.P. Wand. 1995. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90(432):1257–70.

Schölkopf, B., A.J. Smola, R.C. Williamson, and P.L. Bartlett. 2000. New support vector algorithms. *Neural Computation* 12:1207–45.

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

Appendix A. Tables A1–A4.

Table A1. Values of the two criteria for Swedish data.

SWEDEN		Kernel	HP8	SVM	PSVM
Females					
1981–1985	χ^2	2842	950	725	255
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	381	624	258	513
1984–1988	χ^2	1817	861	293	190
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	352	518	431	443
1991–1995	χ^2	2507	1468	882	234
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	321	435	169	390
Males					
1981–1985	χ^2	3813	180	717	427
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	534	73	619	629
1984–1988	χ^2	3125	191	485	314
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	543	695	534	602
1991–1995	χ^2	3340	268	490	341
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	625	578	550	506

Table A2. Values of the two criteria for Japanese data.

JAPAN		Kernel	HP8	SVM	PSVM
Females					
1990	χ^2	1767	4370	453	179
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	348	346	284	318
1991	χ^2	1859	3849	568	234
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	360	347	285	312
1995	χ^2	1601	3516	320	205
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	315	316	261	279
Males					
1990	χ^2	2219	1140	495	370
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	455	410	398	400
1991	χ^2	2047	951	300	277
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	472	403	440	405
1995	χ^2	2023	542	394	430
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	467	406	447	382

Table A3. Values of the two criteria for French data.

FRANCE		Kernel	HP8	SVM	PSVM
Females					
1990	χ^2	3508	2887	594	381
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	570	581	359	478
1991	χ^2	2897	1995	639	366
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	474	557	306	459
1995	χ^2	1839	879	366	330
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	487	405	498	351
Males					
1990	χ^2	4685	983	786	658
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	771	771	756	674
1991	χ^2	4625	687	999	470
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	759	771	638	686
1995	χ^2	2697	987	1117	485
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	788	511	462	504

Table A4. Average values of the two criteria for the overall data.

		Kernel	HP8	SVM	PSVM
Sweden					
Females	χ^2	2388,67	1093,00	633,33	226,33
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	351,33	525,67	286,00	448,67
Males	χ^2	3426,00	213,00	564,00	360,67
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	567,33	448,67	567,67	579,00
Total	χ^2	2907,33	653,00	598,67	293,50
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	459,33	487,17	426,83	513,83
Japan					
Females	χ^2	1742,33	3911,67	447,00	206,00
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	341,00	336,33	276,67	303,00
Males	χ^2	2096,33	877,67	396,33	359,00
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	464,67	406,33	428,33	395,67
Total	χ^2	1919,33	2394,67	421,67	282,50
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	402,83	371,33	352,50	349,33
France					
Females	χ^2	2748,00	1920,33	533,00	359,00
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	510,33	514,33	387,67	429,33
Males	χ^2	4002,33	885,67	967,33	537,67
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	772,67	684,33	618,67	621,33
Total	χ^2	3375,17	1403,00	750,17	448,33
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	641,50	599,33	503,17	525,33
OVERALL TOTAL	χ^2	2733,94	1483,56	590,17	341,44
	$\sum\left\|\Delta^3\hat{q}_x\right\|$	501,22	485,94	427,50	462,83

Appendix B. Figures B1–B18.

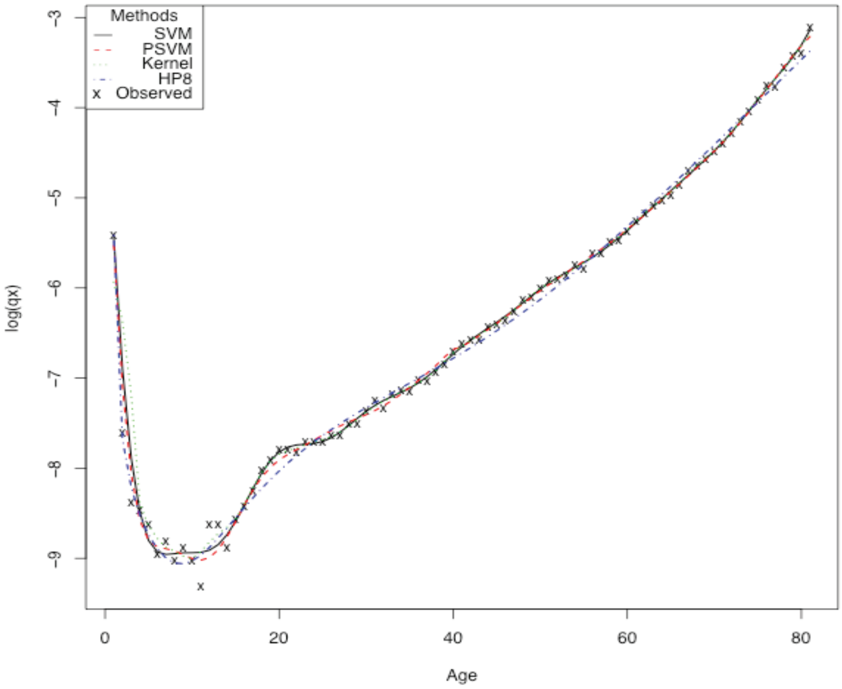

Figure B1. Empirical and graduated q_x-values, French females, 1995.

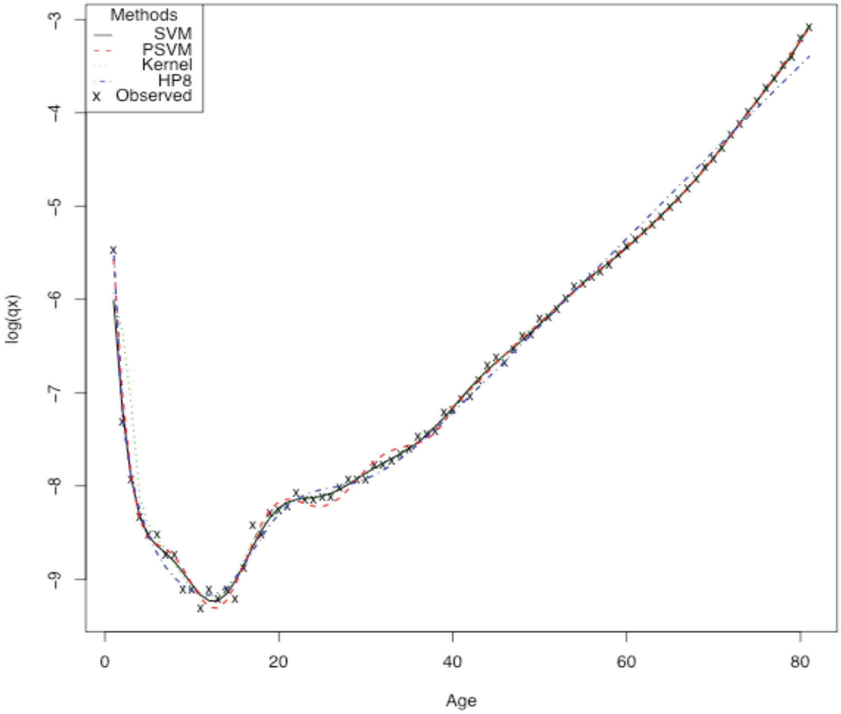

Figure B2. Empirical and graduated q_x-values, Japanese females, 1991.

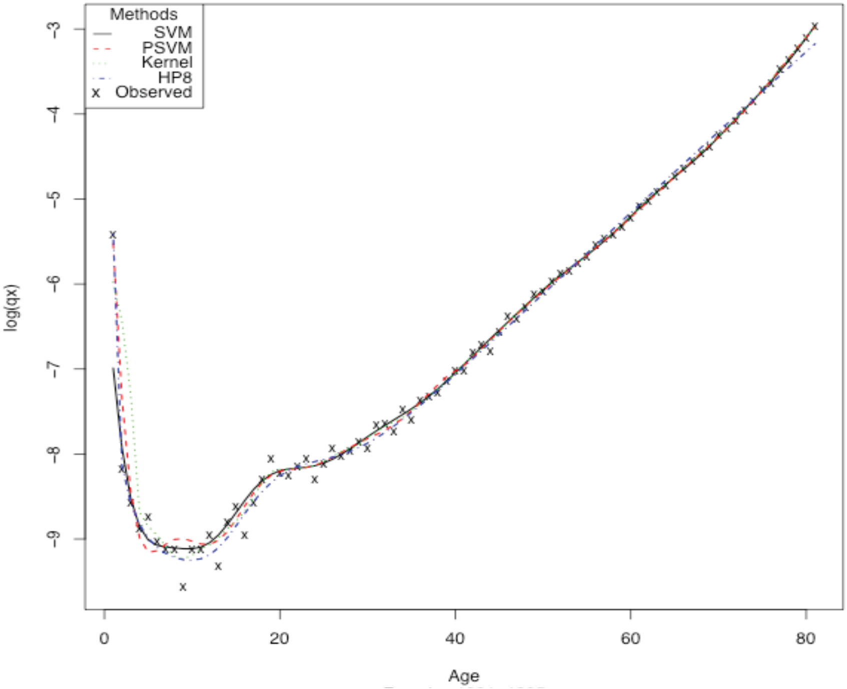

Figure B3. Empirical and graduated q_x-values, Swedish females, 1991–5.

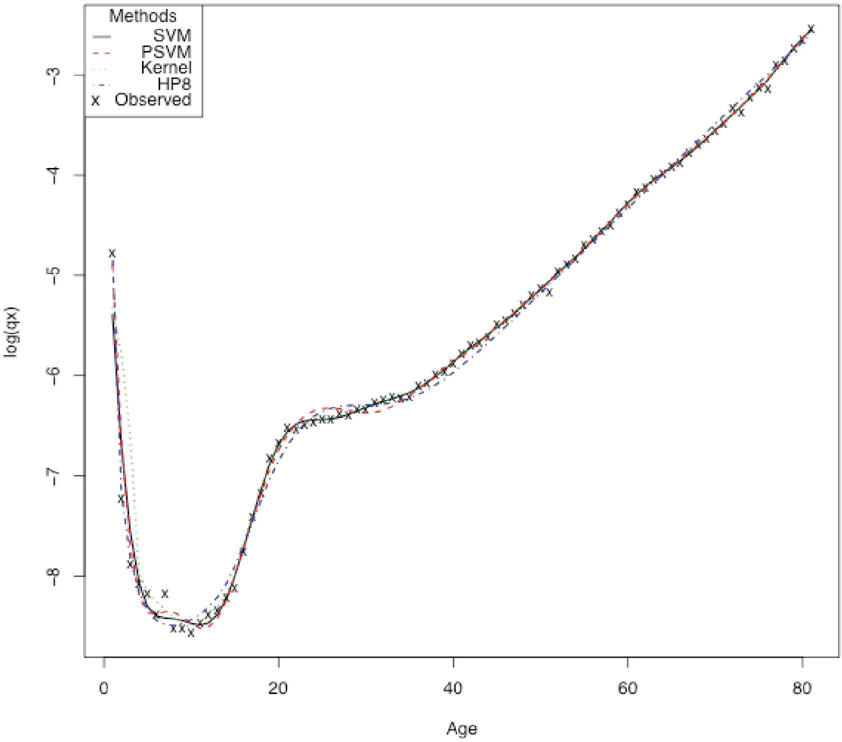

Figure B4. Empirical and graduated q_x-values, French males, 1991.

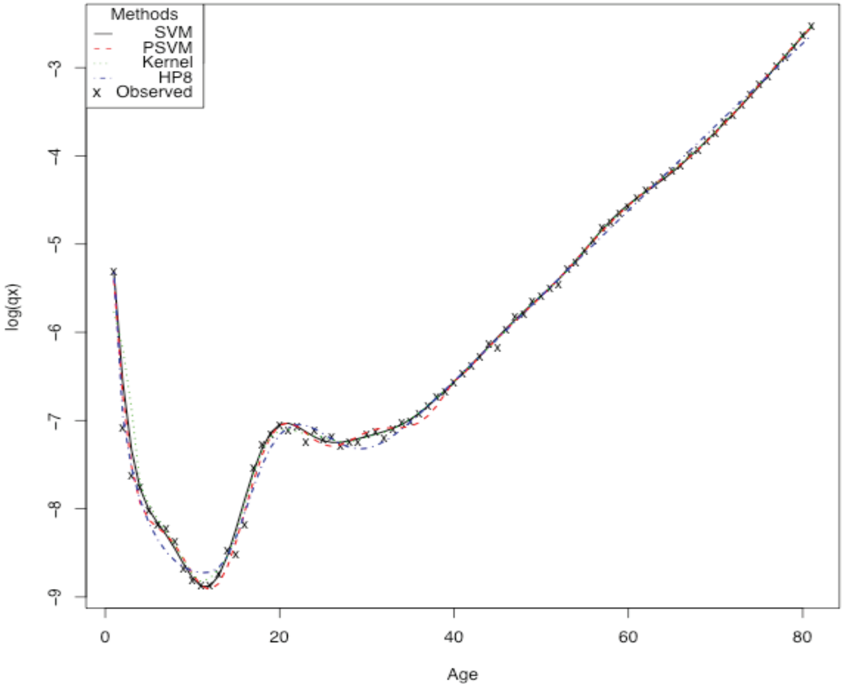

Figure B5. Empirical and graduated q_x-values, Japanese males, 1990.

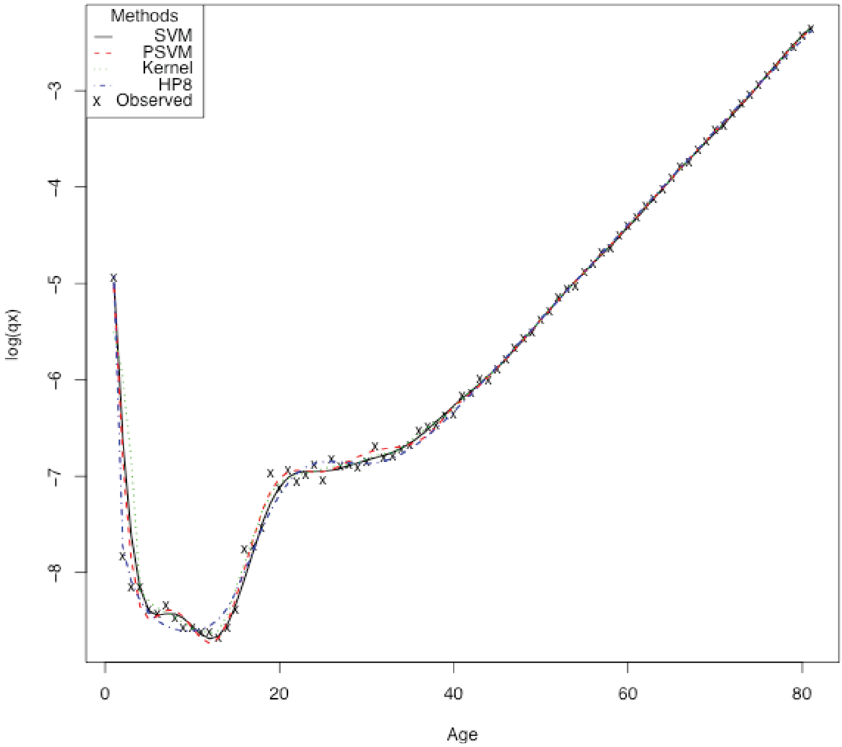

Figure B6. Empirical and graduated q_x-values, Swedish males, 1981–5.

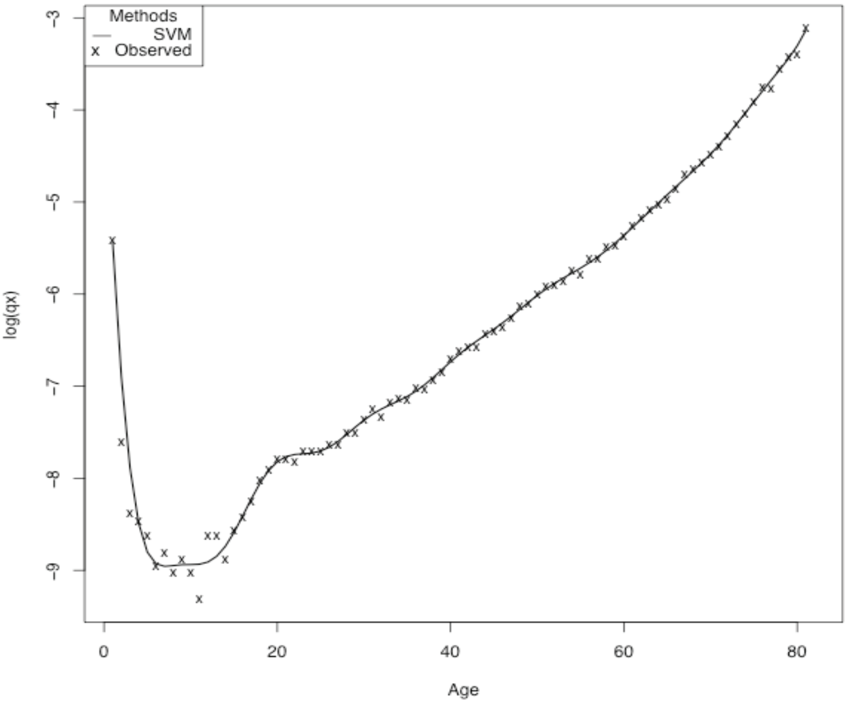

Figure B7. Empirical and graduated q_x-values, French females 1995.

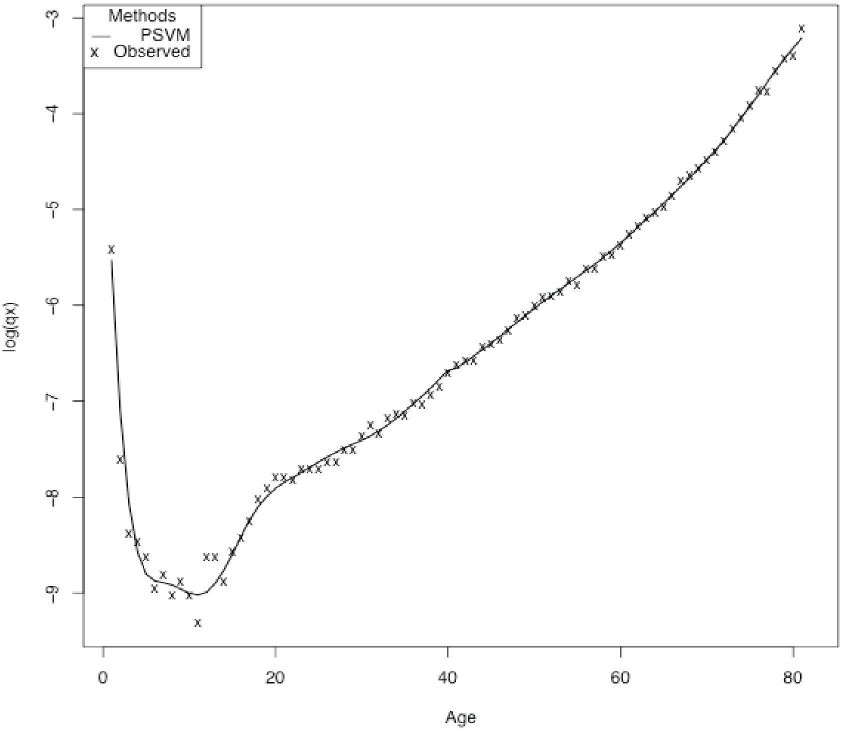

Figure B8. Empirical and graduated q_x-values, French females 1995.

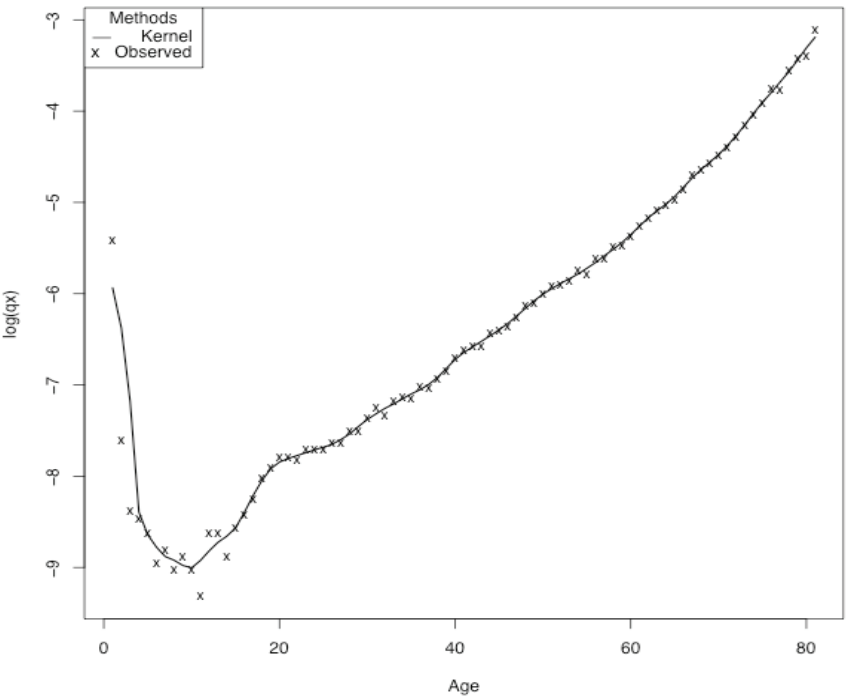

Figure B9. Empirical and graduated q_x-values, French females 1995.

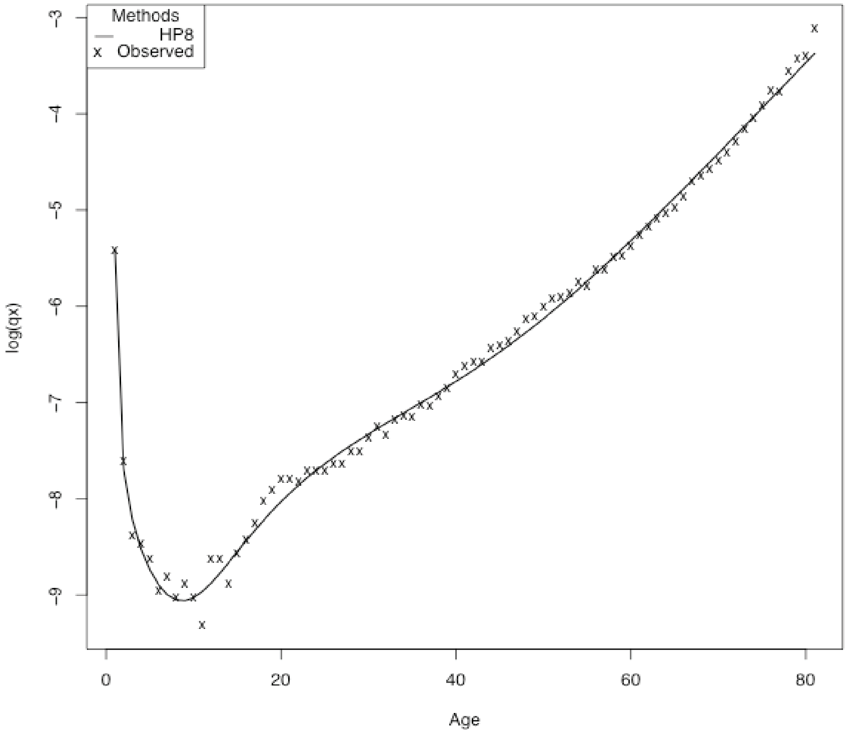

Figure B10. Empirical and graduated q_x-values, French females 1995

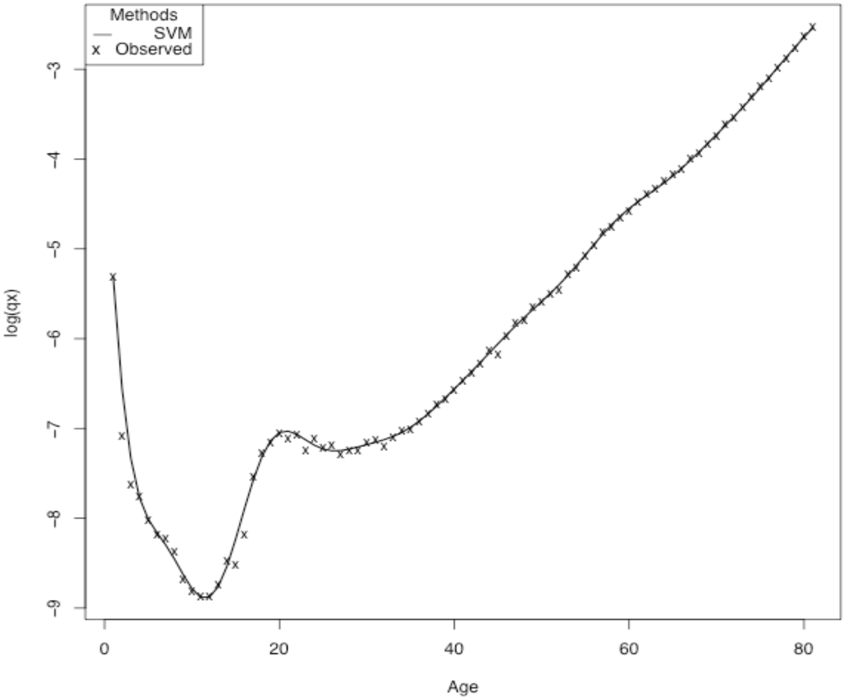

Figure B11. Empirical and graduated q_x-values, Japanese males 1990.

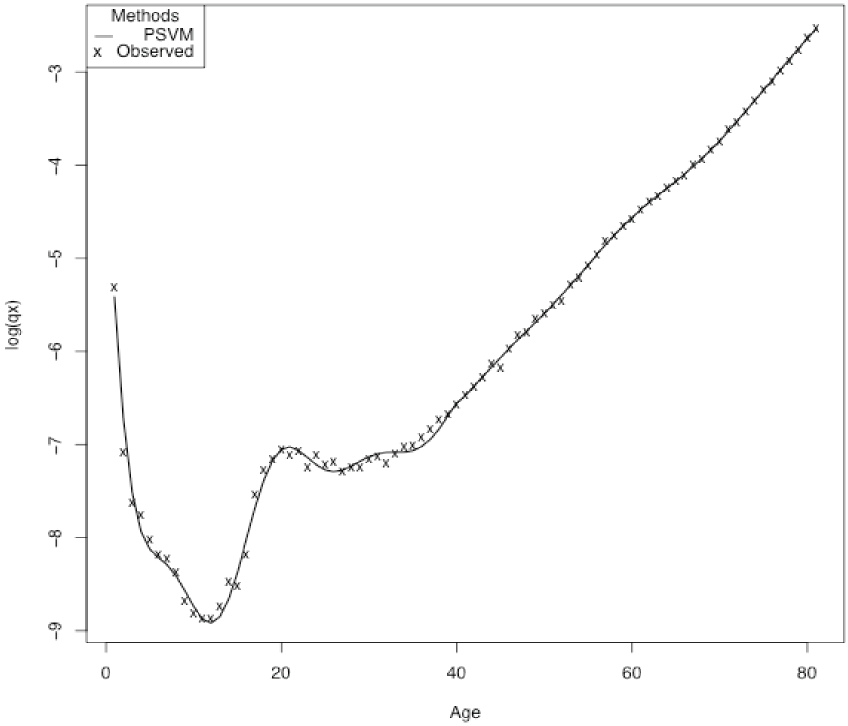

Figure B12. Empirical and graduated q_x-values, Japanese males 1990.

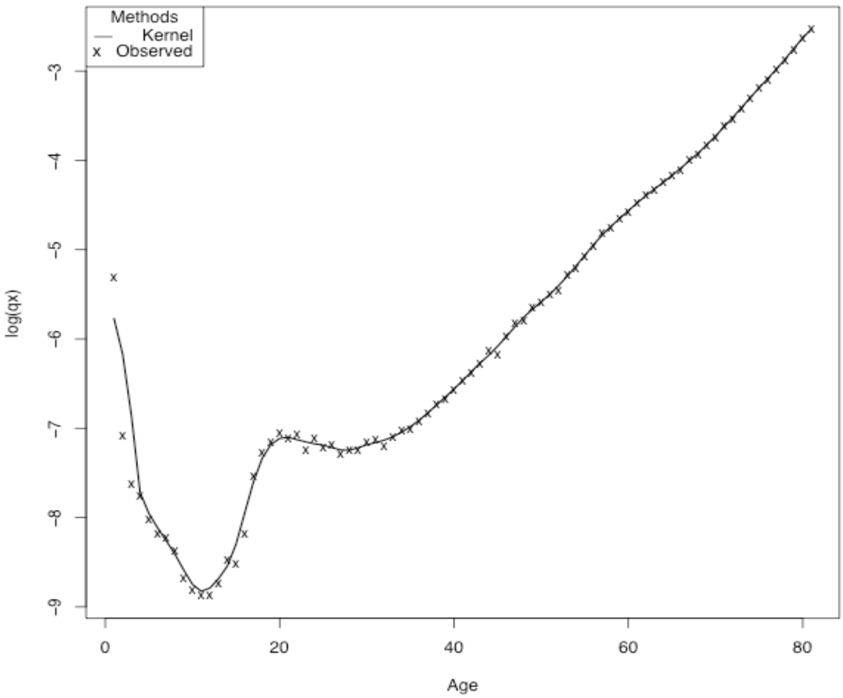

Figure B13. Empirical and graduated q_x-values, Japanese males 1990.

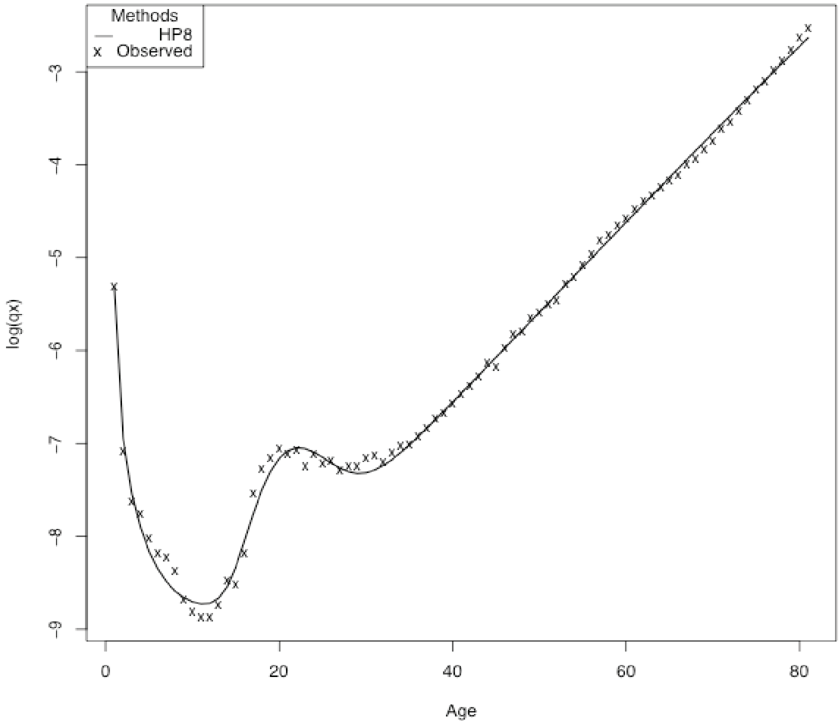

Figure B14. Empirical and graduated q_x-values, Japanese males 1990.

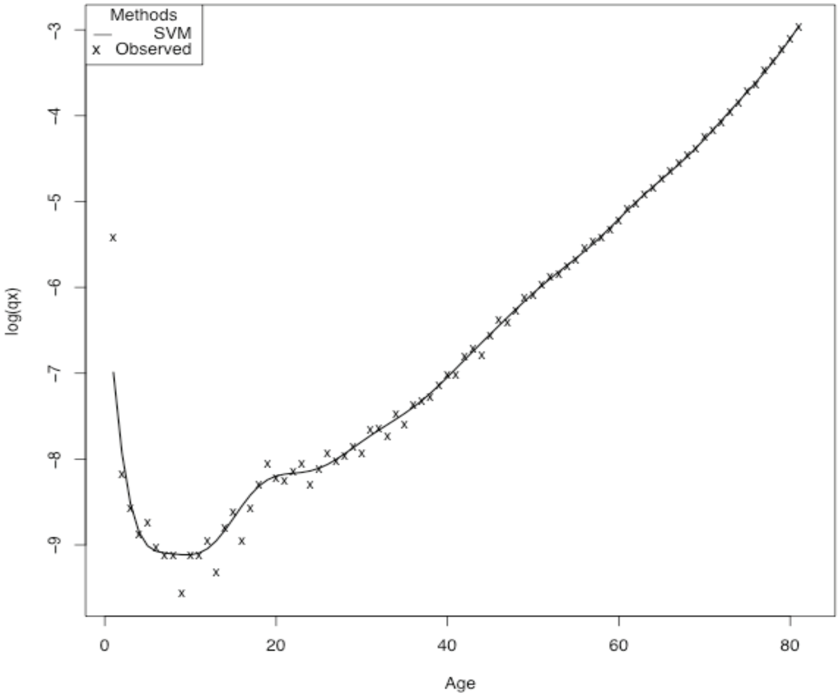

Figure B15. Empirical and graduated q_x-values, Swedish females 1991–1995.

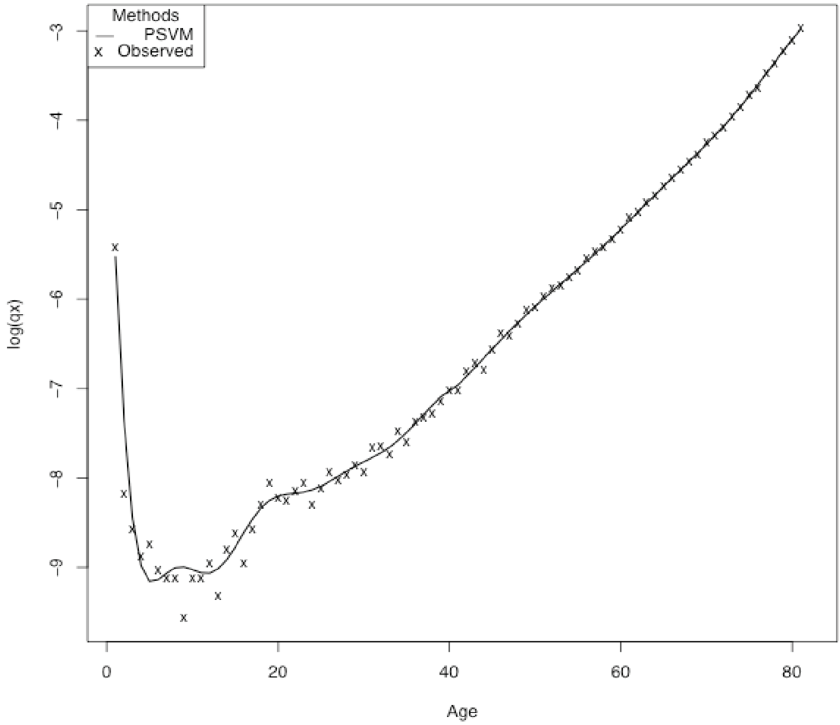

Figure B16. Empirical and graduated q_x-values, Swedish females 1991–1995.

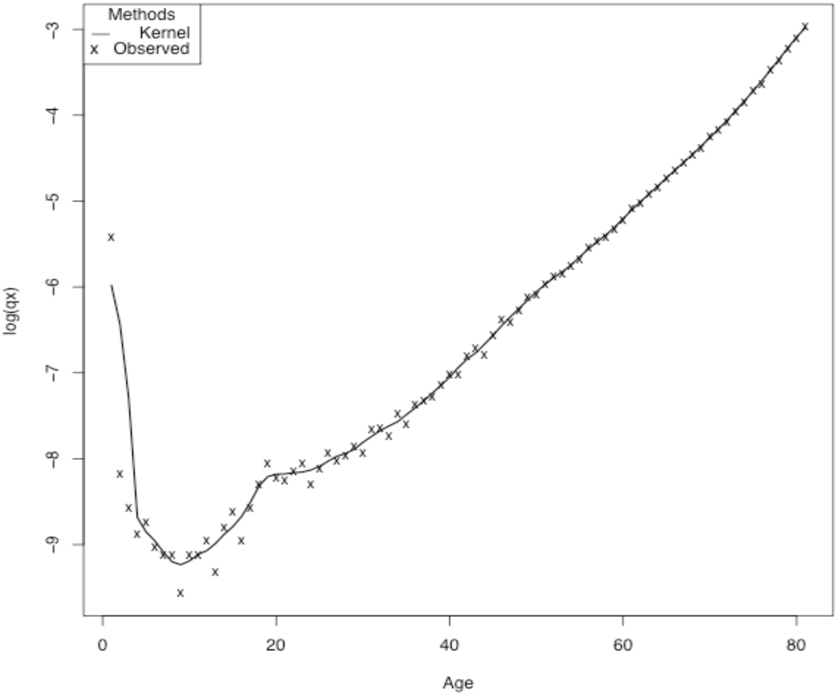

Figure B17. Empirical and graduated q_x-values, Swedish females 1991–1995.

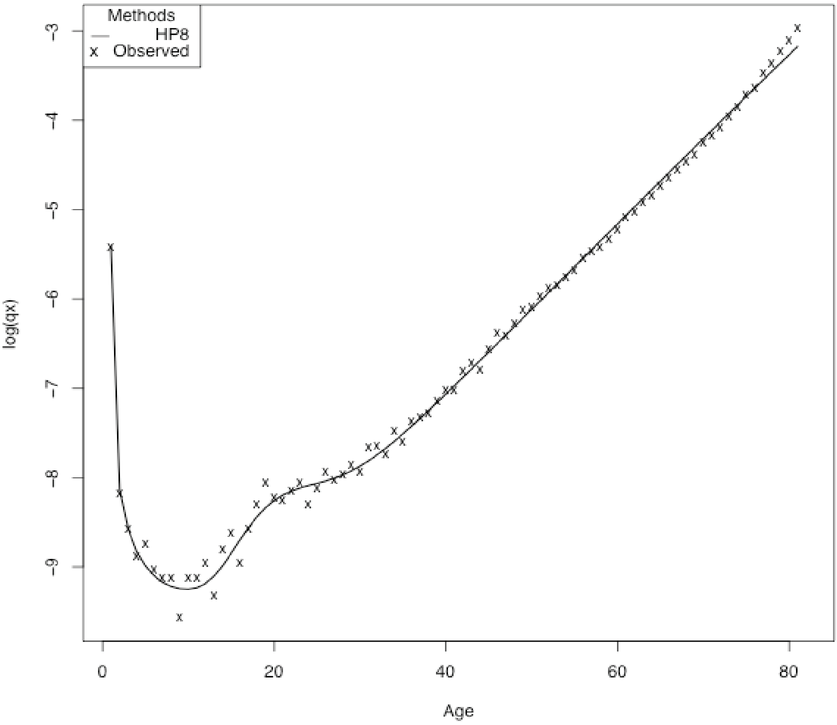

Figure B18. Empirical and graduated q_x-values, Swedish females 1991–1995.