Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR WORKING PAPER WP 2010-013
FEBRUARY 2010 (REVISED DECEMBER 2011)

Probabilistic Forecasting using Stochastic Diffusion Models, with Applications to Cohort Processes of Marriage and Fertility

Mikko Myrskylä (myrskyla@demogr.mpg.de)
Joshua R. Goldstein (goldstein@demogr.mpg.de)

Working papers of the Max Planck Institute for Demographic Research receive only limited review. Views or opinions expressed in working papers are attributable to the authors and do not necessarily reflect those of the Institute.

Probabilistic Forecasting using Stochastic Diffusion Models, with Applications to Cohort Processes of Marriage and Fertility

Mikko Myrskylä [1]

Joshua R. Goldstein [2]

[1] Corresponding author. Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, 18057 Rostock, Germany. Phone: +49 (0)381 2081-118, Fax: +49 (0)381 2081-418, Email: myrskyla@demogr.mpg.de

[2] Max Planck Institute for Demographic Research. Email: goldstein@demogr.mpg.de

Acknowledgements. We wish to credit Ron Lee, who first suggested to JRG that the Hernes diffusion models be fit using a time-series, stochastic diffusion approach.

ABSTRACT

We study prediction and error propagation in the Hernes, Gompertz, and logistic stochastic diffusion models and use them to forecast demographic cohort processes. We develop a unified framework in which the models are linearized with respect to cohort age and predictions are derived from an underlying linear process. For prediction variance we develop a Monte Carlo estimator which can be used for a wide class of underlying linear processes. For the case of random walk with drift we develop an analytic prediction variance estimator. The variance estimators allow the forecaster to make precise the level of within-model prediction uncertainty. In addition, the analytic variance estimator provides insights into the sources of prediction uncertainty. Applications to marriage and fertility rates illustrate the usefulness of the new methods, and extend them to simultaneous forecasting of multiple cohorts and to processes restricted by factors such as declining fecundity.

INTRODUCTION

Forecasting uncompleted cohort experience is a key task in demography. Diffusion models, which describe how a population adopts a new innovation, technology, or behavior, are potentially useful in this respect. We analyze the Hernes, Gompertz, and logistic innovation diffusion models and develop a unifying framework for time-series based probabilistic forecasting of cohort processes with these models. We introduce the concept of stochastic diffusion, which both expands the theoretical coverage of the models to include period effects and allows evaluation of the probabilistic forecast uncertainty. Applications to marriage and fertility rates illustrate the usefulness of these new methods, and extend the methods to simultaneous forecasting of multiple cohorts and to processes restricted by factors such as declining fecundity.

BACKGROUND

The Hernes, Gompertz, and logistic diffusion models are commonly used in the social sciences. The Hernes model has been developed and applied for forecasting cohort marriage patterns (Hernes 1972, Goldstein and Kenney 2001, Li and Wu 2008). For cohort fertility forecasts, the Gompertz model previously used to fit period fertility (Hoem Madsen et al. 1981, Pollard and Valkovics 1992) can also be used to predict cohort rates (Goldstein 2010). The logistic diffusion model has not been used often for modeling cohort schedules (cf. Ike 2002), but is the standard model of population growth subject to constraints (Pearl and Reed 1920; Preston, Heuveline and Guillot 2001). Furthermore, in the economic literature the logistic model had been used extensively to forecast innovation diffusion (Gruber and Verboven 2001; Harvey 1984; Mar-Molinero 1980; Meade and Islam 2006).

Although developed in a deterministic setting, the logistic, Gompertz, and Hernes models can all be extended to a stochastic setting. We do this by allowing random shocks to influence the processes. Introducing randomness is appealing because it acknowledges that the model under consideration is not the only influence on behavior. Introducing random shocks not only incorporates the possibility of other influences but, in our formulation, quantifies the importance of these outside factors to uncertainty in forecasts. A further advantage of introducing randomness is that it allows inclusion of influences that may extend across periods or affect cohorts in similar or correlated ways. In short, we see the stochastic models we introduce here as a step forward in making diffusion models broader and more realistic.

Our approach builds on the stochastic forecasting framework pioneered by Alho, Lee, Tuljapurkar and others (Alho 1990; Lee 1993; Lee and Tuljapurkar 1994). We take from these approaches the idea of modeling temporal change as a single or set of univariate stochastic time series, with the difference being that our approach is applied to cohort processes. An innovation of our approach is that we introduce stochastic elements in the context of behavioral cohort models.[1] Cohort forecasting is of particular interest to those studying the behavioral basis of demographic rates, as it relates to the life course behavior of individuals. For example, cohort fertility or marriage behavior is that experienced by real individuals as opposed to the synthetic nature of period indices.

[1] In the economic literature, additive shocks at the level of the directly observed non-linear process are occasionally incorporated into diffusion models (Meade and Islam 1995, 2006). Such an approach may be unrealistic in demographic applications in which the process stabilizes and uncertainty decreases with age and level of the process.

The models we explore were all based in their original formulation on differential equations, wherein the levels and previous rates of change influence the subsequent evolution of the process. We introduce a stochastic element to these differential equations. In physics, finance, and many other fields, the new field of stochastic differential equations has allowed the introduction of random perturbations into previously deterministic models (for example, the Ito equations with wide range of applications (Øksendael 2003), or the Black-Scholes (1973) option pricing equations). Here we take a first step at introducing a similar conceptualization to demographic models.[2]

Our approach is based on linearization of the models. When forecasting is the goal, linearization has certain advantages over alternative methods such as fitting the diffusion curve to observed cumulative proportions, or change in the proportions (Billari and Toulemon 2006; Goldstein and Kenney 2001; Hernes 1972; Martin 2004). In particular, linearization makes the estimation easy and allows the incorporation of stochastic shocks in a straightforward additive, rather than multiplicative fashion. [3]

[2] The conceptual similarity between stochastic differential equations (SDE) and our results is masked by the fact that we work in an exclusively discrete set-up. Consequently, the mathematics look different. The conceptualization, however, is not. For example, it would be straightforward to combine the stochastic linear processes (introduced in the next section) with corresponding behavioral differential equations to get what are called Langevin equations in the SDE language.

[3] The tendency to treat complex processes as linear is occasionally criticized, as in the "General Linear Reality" paradigm the timing and order of events are often irrelevant for the outcome, and there is no feedback from the outcome to the effect (Abbott 1988). In the world of diffusion processes we are able to relax these assumptions since the timing and sequence is critical, and the diffusion process allows for dynamic feedback from the process to the effect.

Linearizing the diffusion model, in itself, is not new. Winsor showed in 1932 how the logistic and Gompertz models can be linearized with respect to time. Harvey (1984) took the next step by showing how the predictions of a logistic model can be constructed from an autoregressive integrated moving average (ARIMA) time series model fit to the underlying linear process. In much of the research, however, the linear process has been modeled as a deterministic time trend (Frances 1994; Li and Wu 2008). This is contrary to the idea of diffusion since in the deterministic time trend model the effect of perturbations, or period shocks, vanishes over time[4]. For example, Li and Wu (2008) use the Hernes model to predict first marriages, and base the predictions on a deterministic underlying process which is modeled using a linear regression line. If a more dynamic difference stationary structure is allowed, as in Harvey (1984), no attempt to derive prediction variance has been made. [5]

We build on prior research on modeling cohort processes with diffusion models by i) treating the underlying linear process as a dynamic non-stationary process; ii) showing how Monte Carlo simulation allows prediction interval estimation for a wide range of underlying linear processes; and iii) deriving an analytical prediction interval estimator for the case of random walk with drift as the underlying linear process. The approach allows the user to estimate and understand the sources of the probabilistic uncertainty in the predictions, a topic which is becoming increasingly important in

[4] In the trend stationary specification, the effect of perturbations vanishes over time; in the difference stationary specification, perturbations have a long-lasting effect (Raffalovich 1994). In diffusion processes the past influences the future. Thus the difference stationary specification, which has “long memory”, seems to fit better for diffusion processes.

[5] For example, in a logistic analysis of the growth of a stock variable – the number of tractors in Spain – Harvey (1984: 644) writes that “Unfortunately, finding a suitable prediction interval for the stock is not as straightforward. Various approximations can be derived, but a study of their properties has not been attempted here.”

demography (Alho et al. 2006; Keilman and Pham 2004; Lee 1998; Lutz and Goldstein 2004; Lutz, Sanderson and Scherbov 2001). Empirical applications extend the methods to simultaneous forecasting of correlated cohorts and to processes restricted by factors such as declining fecundity, and illustrate that the new methods are useful in quantifying the prediction uncertainty.

Our work is distinct from the large literature that deals with diffusion in various scientific contexts. In spatial and network analysis, the word diffusion is often used to describe the influence or feedback between neighboring or linked observations (Anselin 1988; Valente 1995). In spatial analysis, the problem often reduces to the specification and estimation of linear regression models which describe how regions are linked in time and space (Doreian 1980; Land, Deane and Blau 1991; Tolnay, Deane and Beck 1996). In network analysis, the central focus is on describing the structure of the linkages between individuals, and analyzing how the linkages influence the flow of ideas or behaviors (Christakis and Fowler 2008; Cowan and Jonard 2004; Marsden and Friedkin 1993). In both spatial and network analysis, the issues of how non-linear behavioral diffusion processes can be linearized and estimated, and how probabilistic forecasts and forecast intervals can be derived from the underlying linear process, are as far as we know largely absent.

In demography and sociology, diffusion models such as Hernes are often used to analyze and forecast the adoption of new ideas in the same spirit we do. The statistical issues that are confronted, however, are mainly about the estimation of the model parameters. The model itself is seen as deterministic, and the only source of prediction uncertainty comes from the uncertainty in the data and in the model parameters (Goldstein and Kenney 2001, Li and Wu 2008). When within-model stochasticity is allowed, this often pertains to individual level uncertainty. For example, in microsimulation and in agent-based modeling of demographic processes the diffusion arises from micro level interactions (Billari and Prskawetz 2003; Hammel, Mason and Wachter 1990; Wachter 1987). At the individual level there is uncertainty regarding the outcomes, but the macro-level uncertainty arises mainly from the simulation and is often seen more as a nuisance that needs to be averaged out rather than a feature characterizing the process.

Also in the survival formulation of the diffusion models, the stochasticity is only at the individual level (Diekmann 1989). These formulations may be very useful for estimating the model parameters, in particular because standard statistical packages often allow flexible estimation of survival models. However, the uncertainty in predictions (if such are made) is then limited to the uncertainty in the model paramaters. The Coale-McNeil model for first marriages (Coale and McNeil 1972) also incorporates stochasticity only at the individual level.

Another advantage of the consistent stochastic framework that we propose is that we allow shocks indexed by time to influence the cohort processes. This extends the reach of cohort models to allow period influences and to be consistent with stochastic period models (Alho 1990; Lee 1993; Lee and Tuljapurkar 1994). The framework we put forward may be useful in analyzing Gompertz mortality models based on a declining stock of vitality with age, or for testing the hypothesis of an invariant rate of aging (Vaupel 2010). Our framework potentially allows researchers to also test the behavioral assumptions of diffusion models by seeing if outside shocks propagate over time.

The paper is organized as follows. The next three sections show how estimation, prediction and prediction interval estimation can be done in the Hernes, Gompertz, and logistic cohort diffusion models using the dynamic time series approach. The first section on Hernes is the most detailed as the Gompertz and logistic cases are highly analogous to the Hernes case. The derivations of the analytical variance estimators are given in the Appendix. Following the introduction of the methods, we illustrate the techniques by applying them to marriage and fertility. To anticipate the results, Table 1 summarizes the key results by showing the model equations, linearizations, prediction equations and prediction variance estimators.

THE HERNES DIFFUSION MODEL

The Model and Its Linearization

Let P_t be the proportion in a cohort that by age t has adopted the innovation under study. Assume that $P_0, P_1, ..., P_t$ are observed and that $P_{t+1}, P_{t+2}, ..., P_{t+k}$ are being predicted using the Hernes model. The Hernes diffusion model (Hernes 1972) for a proportion P_t is

$$\frac{dP_t}{dt} = ab^t P_t (1 - P_t). \qquad (1)$$

A behavioral interpretation of the model is that a person's chance of 1st marriage depends on a peer-pressure effect proportional to the fraction already married P_t, the choice of remaining eligible partners (1-P_t), and an age effect b^t allowing for the lessening attractiveness of marriage with age due either to heterogeneity or to the actual process of getting older. The model can be solved for P_t as

$$P_t = \frac{1}{1 + \frac{1 - P_0}{P_0} \exp\left(\frac{a - ab^t}{\ln b} \right)}, \qquad (2)$$

where P_0 is the initial value at the 1st age of marriage. The model can be linearized with respect to cohort age *t* with

$$\ln\left(\frac{dP_t}{dt} \frac{1}{P_t (1 - P_t)} \right) = \ln a + t \ln b\,. \qquad (3)$$

Following Li and Wu (2008), we accommodate the model for discrete data by approximating the derivative as $\frac{dP_t}{dt} \approx \frac{P_{t+1} - P_{t-1}}{2}$. This gives us

$$\ln a + t \ln b \approx \ln\left(\frac{P_{t+1} - P_{t-1}}{2} \frac{1}{P_t(1 - P_t)} \right) \equiv g_t . \qquad (4)$$

We call the process g_t the underlying linear process.[6] This is a special case of what Li and Wu (2008) call the latent function of the Hernes model. If g_t is assumed to be a deterministic process, then P_t is also deterministic, and the source of prediction uncertainty is the uncertainty in the model parameters. This is the approach of Li and Wu (2008), where the latent function is modeled as a deterministic linear regression model. Here we take a different approach and assume that the underlying process g_t is a time series process, for example, an autoregressive integrated moving average process (ARIMA). As a special case, we consider the random walk with drift model $g_t = g_0 + \delta t + \sum_{i=1}^{t} \varepsilon_i$, where ε_t is independent, normal and zero mean shock with variance σ_ε^2. This is a potentially useful representation of g_t as the model is parsimonious but allows the past influence the future, consistent with the nature of the diffusion concept, and allows for arbitrary random shocks, perhaps due to a new period effect. The parameters $(\delta, \sigma_\varepsilon^2)$ are estimated by

[6] Deviations from linearity in *g* signal deviations from model assumptions. In principle, one can use standard methods to test the linearity (Hinich 1982, Harvey and Leybourne 2007). In demography the number of observations is typically small, resulting in low test power. Therefore visual inspection of the process may be preferred over formal testing, as in Wu (1990). These remarks apply also to the Gompertz and logistic models.

$$\hat{\delta} = \frac{g_{t-1} - g_1}{t-2} \quad \text{and} \quad \hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{t-1}\left(g_i - g_{i-1} - \hat{\delta}\right)^2}{t-3}. \qquad (5)$$

Here, the first observation is at time 0, and there are t+1 observations of the process P. The differencing in equations (4) and (5) makes the denominators in (5) t-2 and t-3, respectively.

Prediction

One-step and k-step ahead predictions $\hat{P}_{t+1}$ and $\hat{P}_{t+k}$ are based on predictions for the underlying linear process g. Under random walk with drift, these are $\hat{g}_{t+1} = g_t + \hat{\delta}$ and $\hat{g}_{t+k} = g_t + \hat{\delta}k$. The predictions $\hat{P}_{t+1}$ and $\hat{P}_{t+k}$ can be derived in several ways; we use the prediction equation

$$\hat{P}_{t+k} = \hat{P}_{t+k-1} + \hat{P}_{t+k-1}\left(1 - \hat{P}_{t+k-1}\right)\exp\left(\hat{g}_{t+k}\right). \qquad (6).$$

The equation (6) is obtained using the approximation

$$\frac{P_{t+1} - P_{t-1}}{2}\frac{1}{P_t\left(1-P_t\right)} = \exp\left(g_t\right) \approx \left(P_t - P_{t-1}\right)\frac{1}{P_{t-1}\left(1-P_{t-1}\right)} \qquad (7)$$

and solving P_t in terms of P_{t-1} and g_t. The core of this approximation is the assumption that a one-step change is close to the average change over two periods, that is, $0.5 \cdot \left(P_{t+1} - P_{t-1}\right) \approx P_t - P_{t-1}$.

The prediction equation (6) is preferred because of its simplicity and linearity in $\exp\left(g_t\right)$. Alternatives include $\hat{P}_{t+k} = 1/[1+\exp[-\exp(\hat{g}_{t+k})]\cdot(1-\hat{P}_{t+k-1})/\hat{P}_{t+k-1}]$, which follows directly from (2), and a prediction that is based on solving $\hat{P}_{t+k}$ from $\exp\left(g_t\right)\hat{P}_{t+k}^2 + \left[1-\exp\left(g_t\right)\right]\hat{P}_{t+k} = \hat{P}_{t+k-1}$. This quadratic equation arises from the approximation $\exp(g_t) \approx (P_t - P_{t-1})/[P_t(1-P_t)]$. Simulation experiments indicated that these alternatives are not more accurate than (6). In fact, all three prediction equations resulted in a small downward bias. This is because a discrete growth factor $\exp(\hat{g}_{t+i+1})$ is applied to $\hat{P}_{t+i}$, whereas optimally one would apply a continuous growth factor from $\hat{P}_{t+i}$ to $\hat{P}_{t+i+1}$. The bias could be reduced by a mid-point correction, analogous to the Euler method for solving differential equations numerically (Griffiths and Smith 1991). In this correction, the growth factor is based on the average of the current and one-step-ahead prediction of the underlying linear process. The bias, however, is small if the step length is small, and was negligible in simulations in which the step length corresponded to one year. Therefore we do not use the mid-point correction. These remarks apply also to the Gompertz and logistic models: A small bias resulting from the discretization is present and could be reduced using the mid point correction, but in practice with one year age groups such a correction is not needed.

Prediction Variance

Here we describe the analytical variance estimator in the case where the underlying linear process is random walk with drift, and discuss how Monte Carlo simulation can be used to estimate the variance for a more general class of processes. The derivations of the analytical variance estimator are given in the Appendix.

Analytical variance estimator. Under random walk with drift as the underlying linear process, the prediction variance for a k-step ahead prediction $\hat{P}_{t+k}$ is

$$V\left(\hat{P}_{t+k}\right) = \sigma_\varepsilon^2 \exp(2g_t) \sum_{i=1}^{k} \sum_{j=1}^{k} \min(i,j) \exp[\delta(i+j)] \gamma_{t+i-1} \gamma_{t+j-1} , \tag{8}$$

where σ_{ε}^{2} is the variance of the error term ε; g_t is the value of the underlying linear process at last observation; and $\gamma_{t+i-1} = \hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)$.

Estimator (8) reveals important facts about the sources of the prediction uncertainty. First, the multiplying factor σ_{ε}^{2} shows that the prediction variance grows linearly with the variance of the error term ε in the underlying linear process. Second, the factor $\exp(2g_t)$ implies that if the predictions are made at a late age, the prediction variance is small. This is because the drift in g_t is negative, so late age (large t) results in small g_t and small $\exp(2g_t)$. Conversely, if the predictions are made at an early age, the variance is large. Third, the term $\exp(\delta)$ implies that if the drift in g is large, that is the diffusion takes place rapidly, the variance is small. Conversely, if the drift is close to 0 and diffusion happens at a slow pace and, the variance is large. Finally, remembering that $\gamma_{t+i-1} = \hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)$, we see that if the proportion P is close to the upper bound 1, then the coefficients γ are small and additional contribution to the variance also small. Similar remarks apply also to the Gompertz and logistic diffusion models (discussed in the next two sections).

Monte Carlo variance estimator. A straightforward Monte Carlo variance estimator can be based on simulated paths of g. For the special case of a random walk with drift and normally distributed innovations, $g_{t+1}, g_{t+2}, \ldots, g_{t+k}$ are simulated by

$$g_{t+j} = g_t + \hat{\delta} j + \sum_{i=1}^{j} \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \hat{\sigma}_{\varepsilon}^{2}\right). \tag{9}$$

The simulated g-paths are transformed to predictions $\hat{P}$ using (6). The variance and (non)parametric prediction intervals can be calculated from the simulated distribution of P. In the Monte Carlo setting, one is not obliged to model the innovations as being normal. Alternative distributions could equally well be used if the data suggests non-normality. One could also resample from the observed innovations, instead of assuming a known distribution.

Table 1 summarizes the key results of this section: The Hernes diffusion model.

THE GOMPERTZ DIFFUSION MODEL

The Model and Its Linearization

Unlike the Hernes model, the Gompertz model is valid for repeated events such as non-parity specific fertility. Thus the variable that is being modeled need not be restricted to the unit interval. To keep the notation consistent, however, we continue to denote the cumulative rate for the process that is being modeled by P_t, where t is the age. Values $P_0, P_1, ..., P_t$ are observed and $P_{t+1}, P_{t+2}, ..., P_{t+k}$ are predicted. The Gompertz growth model for P_t is

$$\frac{dP_t}{dt} = a \exp(-bt) P_t \tag{10}$$

and the solution for the cumulative rate is $P_t = k \exp[-\frac{a}{b}\exp(-bt)]$.

For a behavioral interpretation of the Gompertz model see Goldstein (2010). Log of the log-derivative linearizes the model to $\ln a - bt$. To accommodate the model for discrete data, we use the discretization $\frac{d \ln P_t}{dt} \approx \frac{1}{P_t}\frac{P_{t+1} - P_{t-1}}{2}$, proposed by Li and Wu (2008) in the context of the Hernes model. With this linearization we have

$$\ln a - bt \approx \ln\left(\frac{1}{P_t}\frac{P_{t+1} - P_{t-1}}{2}\right) \equiv g_t. \tag{11}$$

We model the underlying linear process g_t as a time series process. In the case of a random walk with drift, the model parameters are estimated using (5).

Prediction and Prediction Variance

One-step and k-step ahead predictions $\hat{P}_{t+1}$ and $\hat{P}_{t+k}$ are based on predictions for the underlying linear process. Under random walk with drift, these are $\hat{g}_{t+1} = g_t + \hat{\delta}$ and $\hat{g}_{t+k} = g_t + \hat{\delta}k$. To derive the predictions $\hat{P}_{t+1}$ and $\hat{P}_{t+k}$ we use the approximation $0.5 \cdot (P_{t+1} - P_{t-1}) \approx P_t - P_{t-1}$, which was used also in the Hernes case. We proceed in deriving the predictions as follows. First note that for the Gompertz model $\exp(g_t)$ describes proportional change. This can be approximated by

$$\exp(g_t) = \frac{1}{P_t}\frac{P_{t+1} - P_{t-1}}{2} \approx \frac{1}{P_t}(P_t - P_{t-1}) = 1 - \frac{P_{t-1}}{P_t}. \tag{12}$$

The right hand side expression for g_t in (12) allows expressing P_t in terms of the previous observation P_{t-1} and current value of g_t: $P_t \approx P_{t-1} /[1 - \exp(g_t)]$. This leads to recursive prediction:

$$\hat{P}_{t+1} = \frac{P_t}{1 - \exp(\hat{g}_{t+1})} \quad \text{and} \quad \hat{P}_{t+k} = \frac{\hat{P}_{t+k-1}}{1 - \exp(\hat{g}_{t+k})} . \tag{13}$$

Prediction variance can be obtained analytically or by Monte Carlo simulation. The Monte Carlo method is identical to the Hernes case. The analytical variance estimator for a k-step ahead prediction $\hat{P}_{t+k}$ is derived in the Appendix; the result is

$$V(\hat{P}_{t+k}) = \sigma_\varepsilon^2 \exp(2g_t) \sum_{i=1}^{k} \sum_{j=1}^{k} \min(i, j) \cdot \exp[\delta(i + j)]. \tag{14}$$

As in the Hernes case, the estimator (14) reveals the sources contributing to prediction uncertainty. First, the prediction variance grows linearly with the variance of ε. If the predictions are made at a late age, so that g_t is small, the prediction variance is small. If the diffusion is rapid so that the absolute value of δ is large, the variance is small. The only major difference with respect to the Hernes variance equation is that (14) does not include terms of the type $\gamma_{t+i-1} = \hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)$, which implied that as P gets closer to one, increase in variance is small. These terms are absent here since the Gompertz process is not limited to the unit interval.

Table 1 summarizes the key results of this section: The Gompertz diffusion model.

THE LOGISTIC DIFFUSION MODEL

The Model and Its Linearization

As in the Hernes case, P_t is the proportion in a cohort that has by age t adopted the innovation under study, $P_0, P_1, ..., P_t$ are the observed proportions and $P_{t+1}, P_{t+2}, ..., P_{t+k}$ are predicted. The logistic diffusion model for the proportion P_t is

$$\frac{dP_t}{dt} = bP_t\left(1 - \frac{P_t}{a}\right) \tag{15}$$

and the solution for the cumulative proportion is $P_t = a/[1 + \exp(a - bt)]$. For a behavioral interpretation of the logistic diffusion model see Mansfield (1963). The model is linearized by $\ln\left(\frac{dP_t}{dt}\frac{1}{P_t^2}\right) = \ln(b/a) + a - bt$. To accommodate the model for discrete data, we use the discretization $\frac{dP_t}{dt} \approx \frac{P_{t+1} - P_{t-1}}{2}$. This gives us

$$\ln(b/a) + a - bt \approx \ln\left(\frac{P_{t+1} - P_{t-1}}{2}\frac{1}{P_t^2}\right) \equiv g_t\,. \tag{16}$$

We model the underlying linear process g_t as a time series process. In the case of a random walk with drift, the model parameters are estimated using (5).

Prediction and Prediction Variance

Predictions $\hat{P}_{t+k}$ are based on predictions for the underlying linear process. In order to express P_t in terms of P_{t-1} and g_t, we use the approximation

$$\frac{P_{t+1}-P_{t-1}}{2}\frac{1}{P_t^2} \approx \left(P_t - P_{t-1}\right)\frac{1}{P_{t-1}^2}. \tag{17}$$

Since $\exp\left(g_t\right) = \frac{P_{t+1}-P_{t-1}}{2}\frac{1}{P_t^2}$, we can approximate P_t in terms of P_{t-1} and g_t: $P_t = P_{t-1} + P_{t-1}^2 \exp\left(g_t\right)$. The predictions can then be constructed recursively as

$$\hat{P}_{t+1} = P_t + P_t^2 \exp\left(\hat{g}_{t+1}\right) \quad \text{and} \quad \hat{P}_{t+k} = \hat{P}_{t+k-1} + \hat{P}_{t+k-1}^2 \exp\left(\hat{g}_{t+k}\right). \tag{18}$$

Harvey (1984) uses a similar logistic model but does not provide uncertainty estimates as “finding a suitable prediction interval for the stock is not as straightforward” (Harvey 1984: 645). In the Appendix we derive the variance estimator for the case of random walk with drift; the result is

$$V\left(\hat{P}_{t+k}\right) = \sigma_\varepsilon^2 \exp(2g_t)\sum_{i=1}^{k}\sum_{j=1}^{k}\min(i,j)\exp[\delta(i+j)]\gamma_{t+i-1}\gamma_{t+j-1} \tag{19}$$

where $\gamma_{t+i-1} = \hat{P}_{t+i-1}^2$. The interpretation of (19) is analogous to the Hernes case. If the underlying linear process is not a random walk with drift, Monte Carlo simulation can be used to estimate the variance; the process is the same as in the Hernes case.

Table 1 summarizes the key results of this section: The logistic diffusion model.

APPLICATIONS AND EXTENSIONS

We use simulated data and the Hernes model to illustrate the linearization, prediction and variance estimation under controlled conditions. We also consider two real data applications that extend the models presented above. In the first application we use the Hernes model to forecast the proportion ever married among the French 1965-1975 female cohorts and estimate the probability of a cohort cross-over. In this application, we extend the single-cohort Hernes model to allow cohorts to be correlated in time. In the second application we forecast Dutch fertility for the 1960-1977 the Gompertz model and introduce a method for correcting the predictions for age-specific fecundity decline.

Illustration with the Hernes model

We generate Hernesian data using the random walk with drift model for the linear process. Specifically, we generate random walk observations $g_t = g_{t-1} + \delta + \varepsilon_t$ with parameter values $\delta = -0.15, \sigma_\varepsilon = 0.1$ and $g_0 = 0$. We set the initial value $P_0 = 0.001$ and generate observations $P_1, P_2, ..., P_{35}$ using the Hernesian updating equation $P_t = 1/(1 + \exp\left[-\exp(g_t)\right](1 - P_{t-1})/P_{t-1})$. We then "observe" this data up to age 20 and predict $P_{21}, P_{22}, ..., P_{35}$.

FIGURE 1 ABOUT HERE

Figure 1 Panels A-C illustrate the process. Figure 1 Panel A shows the simulated shocks ε_t for one sample path, simulated 95% prediction interval for *t*>20 from 1,000 sample paths that take off at *t*=20, and estimated 95% prediction interval based on first 20 observations. The standard deviation estimate $\hat{\sigma}_\varepsilon$ is 0.092, a value close to the true value 0.1.

Figure 1 Panel B shows one full realization of the simulated linear process $g_t = g_{t-1} + \delta + \varepsilon_t$ and estimated 95% prediction interval based on first 20 observations. The estimate for the drift is $\hat{\delta} = -0.154$, close to the true drift $\delta = -0.15$. The prediction intervals are estimated in a standard way using $k\hat{\sigma}_\varepsilon^2$ as the variance estimate for k-step ahead prediction $\hat{g}_{20} + k\hat{\delta}$. Due to the discretization, g_{20} is not observed.

Figure 1 Panel C illustrates how the linear process g_t is retransformed to P_t. The Figure shows one full realization of the simulated P_t and predictions and the estimated 95% prediction interval for P_t, *t*>20. The predictions are constructed using the equation (6) and prediction interval using the equation (8). Both are based on data that is observed only up to *t*=20. The difference between predicted P_t and the one sample path at *t*=35 is small. More importantly, the analytic variance estimator accurately captures the within-model uncertainty: for the 1,000 simulated P_t paths that take off at *t*=20, the coverage rate for the variance estimator at *t*=35 was 92.6% at the 95% nominal level.

French First Marriages and a Cohort Cross-Over

The Hernes model was developed and is often used to predict the proportion married within a cohort (Goldstein and Kenney 2001; Hernes 1972; Li and Wu 2008). Hernes (1972) gave no uncertainty bounds for the predictions. Goldstein and Kenney (2001) base their marriage rate predictions on survey data and estimate only the uncertainty arising from sampling variation.[7] Li and Wu (2008) use

[7] Billari and Toulemon (2006) use the Hernes model to forecast cohort childlessness and base the forecast uncertainty on uncertainty in the model parameters, in the same spirit as Goldstein and Kenney (2001).

a deterministic time trend model in conjunction with the Hernes model, and base their prediction intervals on a mix of within-model and model parameter uncertainty.

We use the Hernes model to predict the proportion ever married among the French female 1965, 1970 and 1975 cohorts, and analyze the likelihood that the younger cohorts' would catch up with the older cohorts' in proportion ever married. For the 1965 cohort data is observed up to age 40; for the 1970 cohort up to age 35; and for the 1975 cohort up to age 30. We use the full observed data to construct the predictions.

To analyze the likelihood of a cohort cross-over, we first construct the predictions and prediction intervals for each of the cohorts using the random walk with drift specification developed earlier in this paper. We then extend the basic single-cohort Hernes model to a correlated-cohorts model which allows a more realistic analysis of the cohort cross-over. Figure 2 shows the results of the single-cohort approach in which no correlation between the cohorts is allowed.

FIGURE 2 ABOUT HERE

Figure 2 shows that the prediction bounds for the 1965 cohort do not overlap with those for the 1970 and 1975 cohorts. Thus it seems unlikely that the 1970 or 1975 cohorts would catch up with the 1965 cohort. The prediction interval for the 1975 cohort, however, overlaps with that of the 1970 cohort, suggesting that the 1975 cohort may catch up with 1970 cohort. Such inference, however, assumes that the cohort processes are independent. In reality, period fluctuations may influence cohorts simultaneously, creating correlations across cohorts. The relevant correlation is the correlation in the innovations in the underlying random walk with drift processes. If the correlation in cohort processes is assumed to be zero, Figure 2 gives a reasonable picture of the likelihood of a cohort cross-over.

We, however, estimate the correlation for years 1995-2005 (ages 25-35 for cohort 1970 and ages 20-30 for cohort 1975) to be 0.31.

We use the Monte Carlo method in conjunction with the estimated cohort correlation to analyze the likelihood of a cohort cross-over between the 1970 and 1975 cohorts under the assumption that the correlation stays the same at ages not yet observed. We first estimate the drift and variance parameters for the underlying random walk with drift processes for the 1970 and 1975 cohorts. We then simulate two sets of cohort processes with 1,000 simulated marriage paths in each: in the first set, the correlation in the future innovations in the underlying random walk with drift processes for the 1970 and 1975 cohorts is zero. These processes are transformed to predictions which start at age 35 for the 1975 cohort and at age 30 for the 1970 cohort. The results mimic those shown in Figure 2. In the second set of simulations, we allow the cohorts to be correlated by generating the future innovations in the underlying linear processes from a bivariate normal distribution with correlation 0.31, and transform these processes to proportions ever married.

In the simulation without correlation, 0.8% of the 1970 and 1975 cohorts' marriage paths had crossed by age 40; 2.5% by age 42; and 5.4% by age 45. In the simulation with correlated cohorts, 0.2% of the marriage paths had crossed by age 40; 1.8% by age 42; and 4.2% by age 45. Thus when taking the correlation in the cohort processes into account the likelihood of a cohort cross-over drops from above 0.05 to below 0.05.

Dutch Completed Fertility and the Gompertz Model

Kohler (2001) and Bernardi (2003) show that social interaction influences fertility. Consistent with the social interaction theories, Goldstein (2010) shows that the Gompertz model works well in predicting first births and fertility if applied to cohort data. At older ages, and especially for the later cohorts, however, there may be departures from the model. Without prediction intervals, however, it is difficult to assess what is a departure from the model and what is within-model fluctuation.

One of the factors which may result in a departure from model is declining fecundity. At ages above 30 declining fecundity may influence fertility, so that the Gompertz model – which as such does not factor in fecundity – is at risk of overpredicting fertility (for example, Goldstein 2010). We use the Gompertz model to forecast completed cohort fertility, and introduce a method for taking declining fecundity into account. We use cohort fertility data for Dutch female cohorts born from 1960 to 1970 (Human Fertility Database 2010) to first explore the fecundity decline and then estimate an infecundity correction. We test the correction to out-of-sample data (cohorts 1950 and 1955) and then forecast cohort fertility for the 1965-1977 birth cohorts.

Figure 3a shows the estimated underlying process g_t for the Dutch female birth cohorts born in 1960-1970. If the Gompertz model held for these cohorts, the process g_t should be approximately linear. As the Figure 3a shows, the decline in g_t accelerates with age. The acceleration is present for all cohorts and starts at an approximately same age, close to 30, suggesting that the force behind the acceleration is physiological rather than cohort- or period-specific. We exploit this observation to develop an infecundity correction which influences the rate of fecundity decline in the Gompertz diffusion model. The correction is based on the empirical observation that for cohort fertility and ages 30 and above, the underlying process g_t departs from linearity in a predictable manner.

FIGURES 3a, 3b ABOUT HERE

We model the fecundity decline by using a two-stage model in which g_t is a random walk with drift up to age 30. For ages above 30 we assume that for each additional year of age, the pace of the decline accelerates at a constant rate. That is, at ages 30 and above $g_t = g_{t-1} + \delta \cdot IFC^{(t-30)} + \varepsilon$, where IFC is the infecundity correction. We estimate the IFC from the 1960-1970 cohorts as follows. We first estimate the drift parameter for each cohort using data up to age 30. Denote this cohort specific drift parameter by δ_c. For each cohort we construct predictions $\hat{g}_{c,t} = \hat{g}_{c,t-1} + \delta_c \cdot IFC^{(t-30)}$ for ages 31-45. We estimate the parameter IFC by minimizing the sum of weighted squared errors $\sum_{c,t} w_t \left(g_{c,t} - \hat{g}_{c,t}\right)^2$, where the weights are defined as $w_t = (45 - t - 1)/(15 \cdot 8)$. With this specification the weights decline linearly so that $w_{31} = 1/8$ and $w_{31} = 1/(15 \cdot 8)$. Such a weighting gives more weight to young ages whose contribution to fertility matters more than that of older ages. Using a grid search with IFC ranging from 0 to 1.500 with step length 0.001 we estimate the IFC to be 1.118. Figure 3b shows the predicted g_t for the 1960-1970 cohorts with this IFC (observations: ages 15-30; predictions: ages 31-45).

Since the IFC was estimated from the data, it is not surprising that the model fits well for the 1960-1970 cohorts. We test the external validity of the IFC = 1.118 for two birth cohorts preceding the data from which the parameter was estimated, the 1950 and 1955 birth cohorts. We use observations up to age 30; estimate the underlying process g_t and its parameters; and predict g_t with $\hat{g}_{c,t} = \hat{g}_{c,t-1} + \delta_c \cdot IFC^{(t-30)}$, with IFC = 1.118. The predictions for cohort fertility are derived from g_t using (13) and 95% prediction intervals are calculated using (14). For comparison, we estimate the predictions also without the infecundity correction (that is, set IFC = 1).

FIGURES 4a, 4b ABOUT HERE

Figures 4a and 4b show the observations, forecasts with and without the infecundity correction, and 95% prediction interval for the forecasts with the infecundity correction. For the 1950 cohort, the observed cohort fertility at age 45 is 1.90. Without the infecundity correction, the prediction for cohort fertility at age 45 is 2.02 and with the correction, the prediction matches with the true value 1.90. More importantly, the 95% prediction interval [1.82-1.94] covers the true value. For the 1955 cohort, completed fertility at age 45 is 1.87. Without the infecundity correction, the prediction at age 45 is 2.11. With the infecundity correction, the prediction is 1.82 and the 95% prediction interval [1.73-1.90] captures the true value 1.87.[8] Thus without the infecundity correction, cohort fertility is overestimated. With the infecundity correction the predictions seem reasonably accurate.

FIGURE 5 ABOUT HERE

The developed methods allow probabilistic forecasting of fertility by age over cohorts. Figure 5 shows cohort fertility forecasts for ages 30, 35 and 45 for Dutch female cohorts 1950-1977 (1950-1975 with 5 year birth intervals and the 1977 cohort). The data is observed up to year 2008 so cohort fertility is fully observed at age 30 for all cohorts. At age 35, predictions are needed for the 1975 and 1977 cohorts. The graph shows that the prediction uncertainty grows rapidly as we move to more recent cohorts. For the 1977 cohort, predicted cohort fertility at age 35 is 1.55 (95% PI 1.51-1.59) and at age 45, 1.85 children per woman (95% PI 1.72-1.98). In other words, we know that the 1977 cohort had cumulative fertility 1.06 by year 2008. The prediction intervals suggest that by 2012, at

[8] The prediction interval for 1955 cohort is 42% wider than for cohort 1950 even though the predictions start at the same age 30, and the ultimate completed fertility rates are similar. The reason for this is that the 1955 cohort had children later than the 1950 cohort, which means that from the model's perspective, predictions for the 1955 cohort start earlier.

age 35, we expect the 1977 cohort to have on average 0.45-0.54 children more, and by age 45, 0.67-0.92 children more than in 2008.

The point estimates also suggest that cohort fertility is on the increase as the prediction 1.85 for the 1977 cohort is higher than for any cohort born after 1960. The prediction interval, however, shows that the uncertainty in the predictions is high and grows rapidly as we move to later cohorts. For the 1970 cohort (for which fertility is observed up to age 38), the length of the prediction interval is 0.04 units (1.72-1.77). For the 1975 cohort, the length of the prediction interval increases to 0.22 (1.69-1.91) and for the 1977 cohort the interval length is 0.26 units (1.72-1.98). Thus the data is consistent with the hypothesis that cohort fertility is increasing, but the alternative hypothesis – stable or decreasing fertility – can not be rejected.

DISCUSSION

This paper studied prediction and error propagation in the Hernes, Gompertz and logistic innovation diffusion models and applied the methods to demographic processes. We developed a unifying framework in which the predictions and prediction intervals can be derived from an underlying linear process. We showed how Monte Carlo simulation can be used to estimate prediction uncertainty for a wide range of underlying processes, and derived and analytic closed form variance estimator for the case of a random walk with drift. The analytic variance estimator revealed the role of different sources in contributing to total uncertainty, most importantly that the earlier the predictions are made and the slower the diffusion, the larger the uncertainty in the predictions. In the empirical analyses we further extended the methods in two dimensions. First, we showed how one can move from the single-cohort specification to a multi-cohort setting in which the cohort processes are correlated in time, allowing a realistic analysis of the probability of a cohort cross-over. Second, we developed a method that allows correcting cohort fertility forecasts for declining fecundity.

Simulation studies and empirical applications to first marriages and cumulative fertility showed that the developed methods are useful in quantifying the uncertainty in the predictions: They give a precise sense of the within-model error, and allow the forecasters a new ability to characterize the uncertainty. We showed that if the model assumptions hold less than perfectly, as in the case cumulative fertility where advanced age fertility seems to be constrained by extra-model factors such as declining fecundity, the models can be modified to take such factors into account. When accommodating the model for the full complexity of reality is not possible, the constructed prediction intervals give a lower bound for the total uncertainty.

This paper considers only the within-model error in the prediction uncertainty. To assess the magnitude of total error, future studies will need to expand the range of fitted populations, incorporating fertility and marriage data from the United States and other European countries, as well as historical data, to compare the relative importance of the within-model error to the total error. It is especially interesting to see where the methods do not work – for example in the case of fertility postponement, the tendency of the uncorrected Gompertz model to overpredict fertility at oldest ages is likely to be an indication of sterility, a phenomenon the model is not built to capture. Departures from the model may provide means of indirectly estimating the magnitude of lost fertility due to sterility. More generally, if the within-model error accounts for a large fraction of the total error, our methods can be used as gages of the uncertainty in forecasts. If, however, the within-model error is small, then we would recommend characterizing our methods as providing a lower-bound on uncertainty, to which a substantial amount of model specification uncertainty would need to be added.

The new methods give raise to several further applications and research questions. First, the methods may prove useful in predicting period fertility rates which could be done by combining adjoining cohorts. Second, the models allow testing the assumptions of the social diffusion framework by looking if a shock in fertility at certain age influences the cohort's subsequent fertility at older ages. Third, our method for correcting completed fertility forecasts for infecundity may provide a way for estimating the total infecundity, and fertility lost due to infecundity, in modern populations. Fourth, it will be fruitful to look at the correlation in model parameters across cohorts and over time and space. The correlations across these dimensions may be used in increasing the accuracy of the estimated model parameters, and in providing a richer description of past marriage and fertility changes.

APPENDIX

Here we derive variance estimators for the Hernes, Gompertz, and logistic models for the case where the underlying linear process is a random walk with drift. For more general processes, one can use the Monte Carlo method, as discussed in the main text.

Hernes Prediction Variance. In the Hernes model, one step ahead prediction is $\hat{P}_{t+1} = P_t + P_t(1-P_t)\exp(\hat{g}_{t+1})$. Here P_t is a constant, so the prediction variance is

$$V\left(\hat{P}_{t+1}\right) = \left[P_t(1-P_t)\right]^2 V\left[\exp\left(\hat{g}_{t+1}\right)\right] \tag{A1}$$

The delta method approximation for $V\left[\exp\left(\hat{g}_{t+1}\right)\right]$ is

$$V\left[\exp\left(\hat{g}_{t+1}\right)\right] = V\left(\hat{g}_{t+1}\right)\left[\frac{d\exp\left[E\left(\hat{g}_{t+1}\right)\right]}{dx}\right]^2 \tag{A2}$$

We assume that the contribution of the uncertainty in the drift estimate is small.[9] Then

$$V\left(\hat{g}_{t+1}\right) = E\left(g_t + \hat{\delta} - g_t - \delta - \varepsilon_{t+1}\right)^2 \approx E\left(\varepsilon_{t+1}\right)^2 = \sigma_\varepsilon^2 \tag{A3}$$

and

$$\frac{d\exp\left[E\left(\hat{g}_{t+1}\right)\right]}{dt} = \exp\left[E\left(\hat{g}_{t+1}\right)\right] = \exp\left(g_t + \delta\right). \tag{A4}$$

[9] Chatfield (1993, 2001) discusses the contribution of parameter uncertainty on prediction intervals and concludes that "given all other uncertainties, it is usually adequate to compute PIs by substituting parameter estimates into the true-mode PMSE [prediction mean squared error]" (2001: p481)

Combining (12)-(15) we get the one-step ahead prediction variance:

$$V\left(\hat{P}_{t+1}\right)=\left[P_t\left(1-P_t\right)\right]^2\sigma_\varepsilon^2\exp\left(2g_t+2\delta\right). \qquad \text{(A5)}$$

Variance (A5) is estimated by replacing σ_ε^2 and δ by their estimators, given in (5).

For a k-step ahead prediction, the recursive nature of the Hernes prediction equation (6) means that one has to take into account the cumulation of uncertainty. We approach the problem by approximating the Hernes predictions with

$$\hat{P}_{t+k}\approx P_t+\sum_{i=1}^{k}\hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)\exp\left(\hat{g}_{t+i}\right). \qquad \text{(A6)}$$

Variance of (17) can be approximated by the double sum of the covariances:

$$V\left(\hat{P}_{t+k}\right)\approx V\left[\sum_{i=1}^{k}\gamma_{t+i-1}\exp\left(\hat{g}_{t+i}\right)\right]=\sum_{i=1}^{k}\sum_{j=1}^{k}\gamma_{t+i-1}\gamma_{t+j-1}\operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right),\exp\left(\hat{g}_{t+j}\right)\right], \qquad \text{(A7)}$$

where $\gamma_{t+i-1}=\hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)$. The diagonal elements of the covariance matrix can be estimated using the delta method as

$$V\left[\exp\left(\hat{g}_{t+i}\right)\right]=i\sigma_\varepsilon^2\exp\left(2g_t+2i\delta\right). \qquad \text{(A8)}$$

The off-diagonal elements $\operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right),\exp\left(\hat{g}_{t+j}\right)\right]$, $i\neq j$ result from double-counting of the errors: shocks ε_t up to $t=i$ influence both g_{t+i} and g_{t+j}, provided that $j\geq i$. Simulation experiments indicated that these off-diagonal elements have a non-negligible variance contribution. We approximate the off-diagonal using first order Taylor series approximation as

$$\operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right),\exp\left(\hat{g}_{t+j}\right)\right] \approx \min\left(i,j\right)\cdot\sigma_{\varepsilon}^{2}\cdot\exp\left(g_t+i\delta\right)\exp\left(g_t+j\delta\right). \tag{A9}$$

The interpretation for (A9) is the following. There are $\min\left(i,j\right)$ common shocks ε_t in both g_{t+i} and g_{t+j}, each contributing σ_{ε}^{2} to the covariance. The terms of the form $\exp\left(g_t+i\delta\right)$ scale the covariance proportionally to the size of $\exp\left(\hat{g}_{t+i}\right)$. For $i=j$, the equation (A9) for the off-diagonal elements reduces to the equation (A8) for the diagonal elements.

The k-step ahead prediction variance estimator is obtained by combining (A7)-(A9):

$$V\left(\hat{P}_{t+k}\right)=\sigma_{\varepsilon}^{2}\exp(2g_t)\sum_{i=1}^{k}\sum_{j=1}^{k}\min(i,j)\exp[\delta(i+j)]\gamma_{t+i-1}\gamma_{t+j-1}\,, \tag{A10}$$

where $\gamma_{t+i-1}=\hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)$. This is the estimator given in the equation (8)

Gompertz Prediction Variance. We first linearize the predictions and then approximate the variance with the delta method. For small $\exp\left(\hat{g}_{t+k}\right)$, the predictions (13) are approximated by

$$\hat{P}_{t+1}\approx P_t+\exp\left(\hat{g}_{t+1}\right)\quad\text{and}\quad\hat{P}_{t+k}\approx P_t+\sum_{i=1}^{k}\exp\left(\hat{g}_{t+i}\right). \tag{A11}$$

For a one-step ahead prediction the variance is $V\left(\hat{P}_{t+1}\right)=V\left[\exp\left(\hat{g}_{t+1}\right)\right]$. We already derived the variance estimator for $\exp\left(\hat{g}_{t+1}\right)$ in the Hernes case (equations A2-A4). Here the estimator is the same. Thus the one-step ahead prediction variance for the Gompertz model is

$$V\left(\hat{P}_{t+1}\right)=\sigma_{\varepsilon}^{2}\exp\left(2g_t+2\delta\right). \tag{A12}$$

The parameters of (29) are estimated by (6). For the k-step ahead prediction, we base the estimator on a linearization $\hat{P}_{t+k} = P_t + \sum_{i=1}^{k} \exp\left(\hat{g}_{t+i}\right)$ whose variance is

$$V\left[\sum_{i=1}^{k} \exp\left(\hat{g}_{t+i}\right)\right] = \sum_{i=1}^{k}\sum_{j=i}^{k} \operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right), \exp\left(\hat{g}_{t+j}\right)\right]. \tag{A13}$$

The elements of the covariance matrix (30) are identical to the Hernes case (equations A8-A9). By combining (A8)-(A9) and (A13) we get the k-step ahead prediction variance estimator:

$$V\left(\hat{P}_{t+k}\right) = \sigma_{\varepsilon}^{2} \exp\left(2g_t\right) \sum_{i=1}^{k}\sum_{j=1}^{k} \min\left(i, j\right) \cdot \exp\left[\delta\left(i+j\right)\right]. \tag{A14}$$

First order Taylor series approximation would deliver the same estimator. Note that (A14) is identical to the Hernes estimator (A10), with the exception that here $\gamma_{t+i-1} = 1$.

Logistic Prediction Variance. The logistic prediction equation (18) looks very much like the Hernes prediction equation (6). The key difference is that the logistic equation does not have the terms $(1-\hat{P}_{t+k-1})$ in the updating equation. Following the steps we took to get the Hernes variance estimator (A10), we get the k-step ahead variance estimator for the logistic model:

$$V\left(\hat{P}_{t+k}\right) = \sigma_{\varepsilon}^{2} \exp\left(2g_t\right) \sum_{i=1}^{k}\sum_{j=1}^{k} \min\left(i, j\right) \exp\left[\delta\left(i+j\right)\right] \gamma_{t+i-1}\gamma_{t+j-1}\ , \tag{A15}$$

where $\gamma_{t+i-1} = \hat{P}_{t+i-1}^{2}$. Note that in the Hernes case, $\gamma_{t+i-1} = \hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)$.

REFERENCES

Abbot, A. 1988. "Transcending General Linear Reality." *Sociological Theory* 6:169-186.

Adsera, A. 2004. "Changing fertility rates in developing countries: the impact of labor market institutions." *J Popul Econ* 17:17–43.

Alho, J., M. Alders, H. Cruijsen, N. Keilman, T. Nikander, and D.Q. Pham. 2006. "New forecast: Population decline postponed in Europe." *Statistical Journal of the United Nations Economic Commission for Europe* 23(1):1-10.

Alho, J.M. 1990. "Stochastic methods in population forecasting." *International Journal of Forecasting* 6(4):521-530.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Baum, C.F. 2006. *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.

Bernardi, L. 2003. "Channels of Social Influence on Reproduction." *Population Research and Policy Review* 22(5-6):527-555.

Billari, F.C.and A. Prskawetz. 2003. "Agent-Based Computational Demography." Heidelberg, Germany: Physica Verlag.

Billari, F.C.and L. Toulemon. 2006. "Cohort Childlessness Forecasts and Analysis Using the Hernes Model." in *European Population Conference 2006*. Liverpool, UK.

Black, F.and M. Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81(3):637-654.

Chatfield, C. 1993. "Calculating Interval Forecasts." *Journal of Business & Economic Statistics* 11(2):121-135.

—. 2001. "Prediction Intervals. A review of principles for calculating prediction intervals when forecasting." in *Principles of Forecasting: A Handbook for Researchers and Practitioners*, edited by J.S. Armstrong. Norwell, MA: Kluwer.

Christakis, N.A.and J.H. Fowler. 2008. "The Collective Dynamics of Smoking in a Large Social Network." *New England Journal of Medicine* 358(21):2249-2258.

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.

Cowan, R.and N. Jonard. 2004. "Network structure and the diffusion of knowledge." *Journal of Economic Dynamics and Control* 28(8):1557-1575.

Diekmann, A. 1989. "Diffusion and Survival Models for the Process of Entry Into Marriage " *Journal of Mathematical Sociology* 14:31–44.

Doreian, P. 1980. "Linear Models with Spatially Distributed Data: Spatial Disturbances or Spatial Effects?" *Sociological Methods & Research* 9:29–60.

Frances, P.H. 1994. "A Method to Select Between Gompertz and Logistic Trend Curves." *Technological Forecasting and Social Change* 46:45-49.

Goldstein, J.R. 2010. "A behavioral Gompertz model for cohort fertility schedules in low and moderate fertility populations." in *MPIDR Working Papers*. Rostock, Germany: Max Planck Institute for Demographic Research.

Goldstein, J.R.and C.T. Kenney. 2001. "Marriage Delayed or Marriage Forgone? New Cohort Forecasts of First Marriage for U.S. Women." *American Sociological Review* 66(4):506-519.

Griffiths, D.V.and I.M. Smith. 1991. *Numerical methods for engineers: a programming approach.* . Boca Raton: CRC Press.

Gruber, H.and F. Verboven. 2001. "The diffusion of mobile telecommunications services in the European Union." *European Economic Review* 45(3):577-588.

Hammel, E.A., C. Mason, and K.W. Wachter. 1990. "SOCSIM II, a sociodemographic microsimulation program, rev. 1.0, operating manual." in *Graduate Group in Demography Working Paper No. 29. Berkeley, California, University of California, Institute of International Studies, Program in Population Research.*

Harvery, D.I.and S.J. Leybourne. 2007. "Testing for time series linearity." *Econometrics Journal* 10:149–165.

Harvey, A.C. 1984. "Time series forecasting based on the logistic curve." *Journal of the Operational Research Society* 35:641-646.

Hernes, G. 1972. "The Process of Entry into First Marriage." *American Sociological Review* 37(2):173-182.

Hinich, M.J. 1982. "Testing for Gaussianity and linearity of a stationary time series. ." *Journal of Time Series Analysis* 3(3):169-176.

Ike, S. 2002. "A non-stationary stochastic process model of completed marital fertility in Japan." *The Journal of Mathematical Sociology* 26(1-2):35-55.

Keilman, N.and D.Q. Pham. 2004. "Time series based errors and empirical errors in fertility forecasts in the Nordic countries." *International Statistical Review* 72(1):5-18.

Kohler, H.-P. 2001. *Fertility and Social Interaction: An Economic Perspective*: Oxford University Press.

Land, K.C., G. Deane, and J.R. Blau. 1991. "Religious Pluralism and Church Membership: A Spatial Diffusion Model." *American Sociological Review* 56:237-249.

Lee, R.D. 1993. "Modeling and Forecasting the Time Series of U.S. Fertility: Age Patterns, Range, and Ultimate Level." *International Journal of Forecasting* 9(2):187-202.

—. 1998. "Probabilistic Approaches to Population Forecasting." *Population and Development Review* 24:156-190.

Lee, R.D.and S. Tuljapurkar. 1994. "Stochastic Population Forecasts for the United States: Beyond High, Medium, and Low." *Journal of the American Statistical Association* 89(428):1175-1189.

Li, N.and Z. Wu. 2008. "Modeling and Forecasting First Marriage: A Latent Function Approach." in *Population Association of America Annual Conference*. New Orleans, LA, USA.

Lutz, W.and J.R. Goldstein. 2004. "Introduction: How to Deal with Uncertainty in Population Forecasting?" *International Statistical Review* 72(1):1-4.

Lutz, W., W. Sanderson, and S. Scherbov. 2001. "The end of world population growth." *Nature* 412(6846):543-545.

Mansfield, E. 1963. "The speed of response of firms to new techniques." *Quarterly Journal of Economics* 77:290–311.

Mar-Molinero, C. 1980. "Tractors in Spain: A logistic analysis." *Journal of the Operational Research Society* 31:141-152.

Marsden, P.V.and N.E. Friedkin. 1993. "Network Studies of Social Influence." *Sociological Methods & Research* 22(1):127-151.

Martin, S.P. 2004. "Reassessing delayed and foregone marriage in the United States." in *Russel Sage Working Paper*.

Meade, N.and T. Islam. 1995. "Prediction Intervals for Growth Curve Forecasts." *Journal of Forecasting* 14:413-430.

—. 2006. "Modelling and forecasting the diffusion of innovation - A 25-year review." *International Journal of Forecasting* 22(3):519-545.

Øksendael, B. 2003. *Stochastic Differential Equations. 6th edn*. Berlin: Springer-Verlag.

Pearl, R.and L.J. Reed. 1920. "On the Rate of Growth of the Population of the United States and Its Mathematical Representation " *Proc. Nat. Acad. Sci.* 6:275-288.

Preston, S.H., P. Heuveline, and M. Guillot. 2001. *Demography: Measuring and Modeling Population Processes*: Oxford: Blackwell Publishers.

Raffalovich, L.E. 1994. "Detrending Time Series: A Cautionary Note." *Sociological Methods and Research* 22:492-519.

Tolnay, S.E., G. Deane, and E.M. Beck. 1996. "Vicarious violence: spatial effects on southern lynchings, 1890-1919." *American Journal of Sociology* 102(3):788-815.

Valente, T.W. 1995. *Network Models of the Diffusion of Innovations* Cresskill: Hampton Press.

Vaupel, J.W. 2010. "Biodemography of human ageing." *Nature* 464(7288):536-542.

Wachter, K.W. 1987. "Microsimulation of household cycles." Pp. 215-227 in *Family Demography: Methods and Their Application*, edited by J. Bongaarts, T. Burch, and K.W. Wachter. Oxford: Clarendon Press.

Winsor, C.P. 1932. "The Gompertz Curve as a Growth Curve." *Proceedings of the National Academy of Sciences of the United States of America* 18(1):1-8.

Wu, L.L. 1990. *Simple graphical goodness-of-fit tests for hazard rate models*. Madison, WI: University of Wisconsin Press.

Figure 1. Simulated Hernes data with random walk with drift as the underlying process. Panel A: Shocks ε (epsilon); Panel B: random walk with drift; Panel C: observations and predictions P.

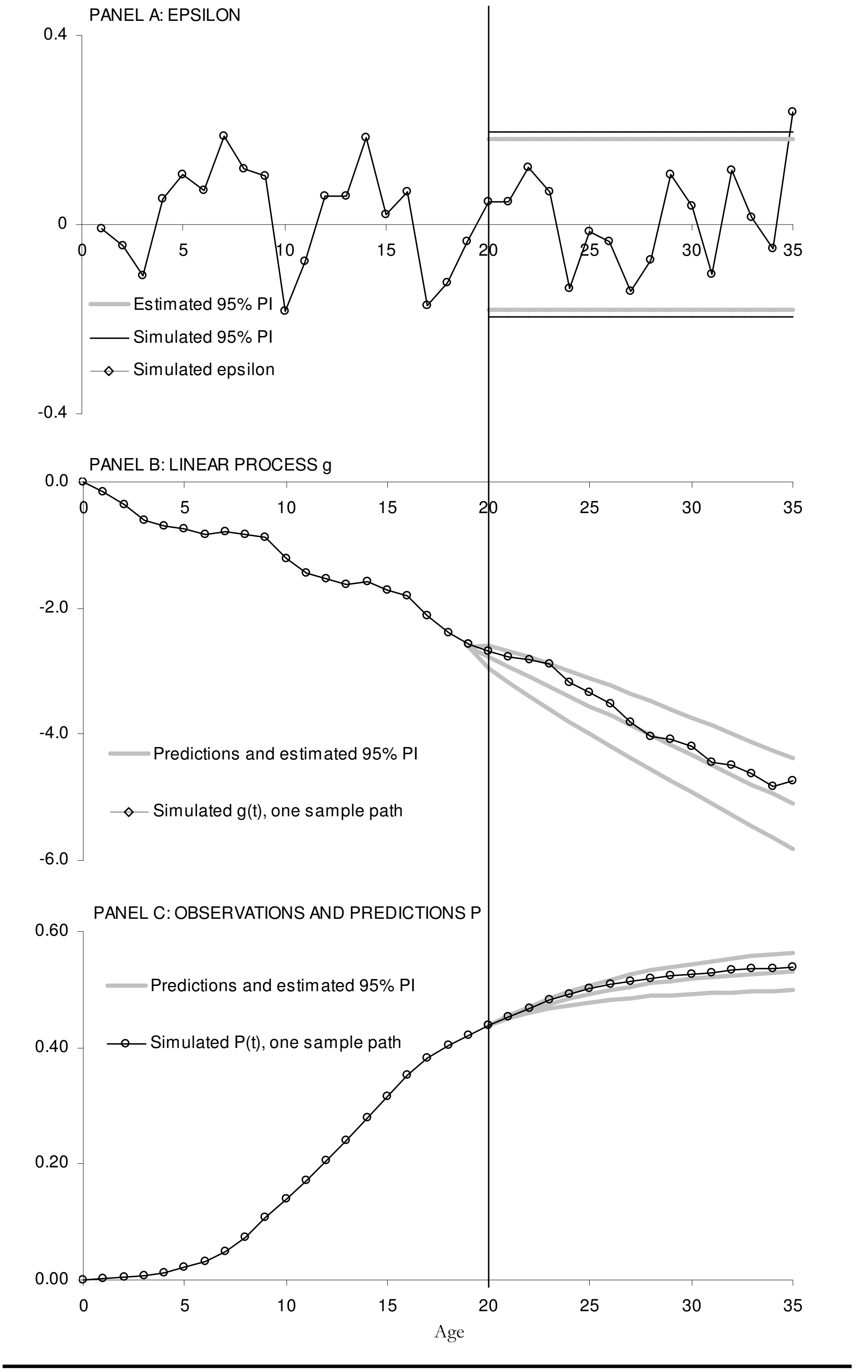
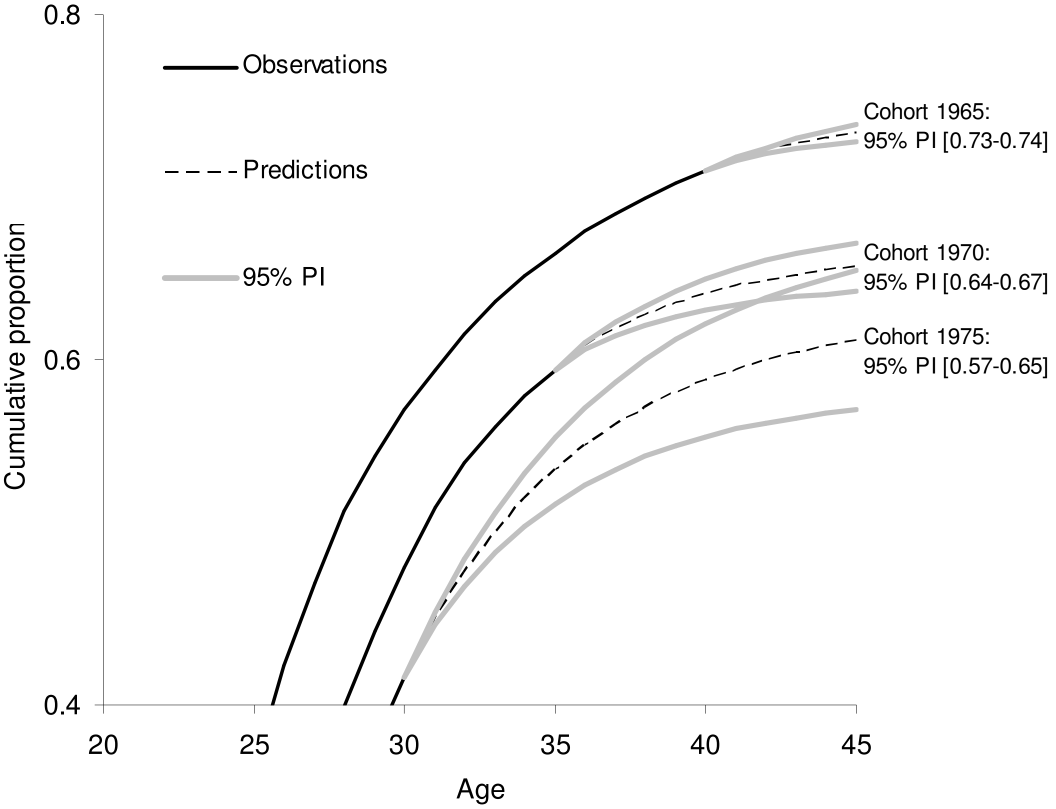
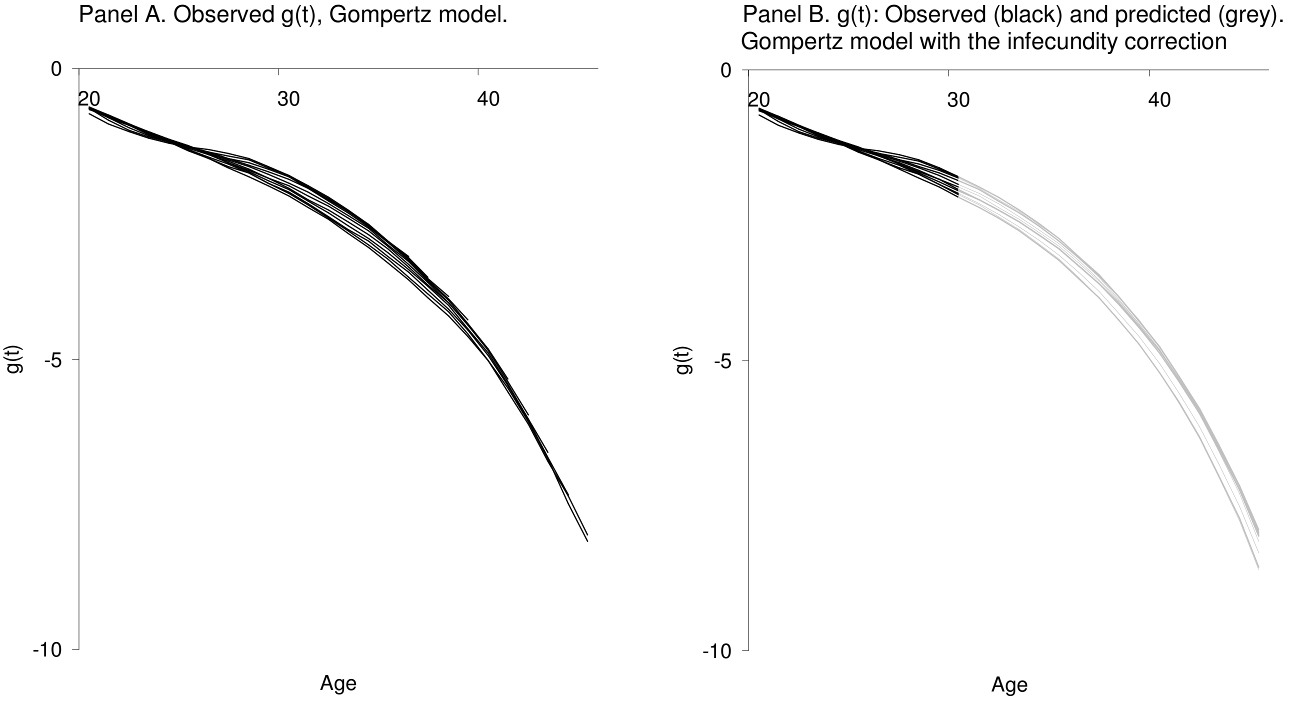


Figure 2. Forecasts and 95% prediction intervals (PI) for the cumulative proportion ever married by age; French female cohorts 1965, 1970 and 1975. Predictions and 95 % PIs are based on the Hernes model with random walk with drift as the underlying linear process.



Notes: The overlapping prediction intervals can only be used to make inferences about cross-over under independent trajectories. See text for discussion of correlated trajectories.

Figure 3. Underlying process g(t) in Gompertz model for cohort fertility and the infecundity correction (IFC). Data: Dutch female cohorts 1960-1970. Panel A: Observed g(t). Panel B: Predicted g(t) based on the estimated IFC and data up to age 30. Source: Cohort TFR Human Fertility Database (2010), g(t) and IFC own calculations.

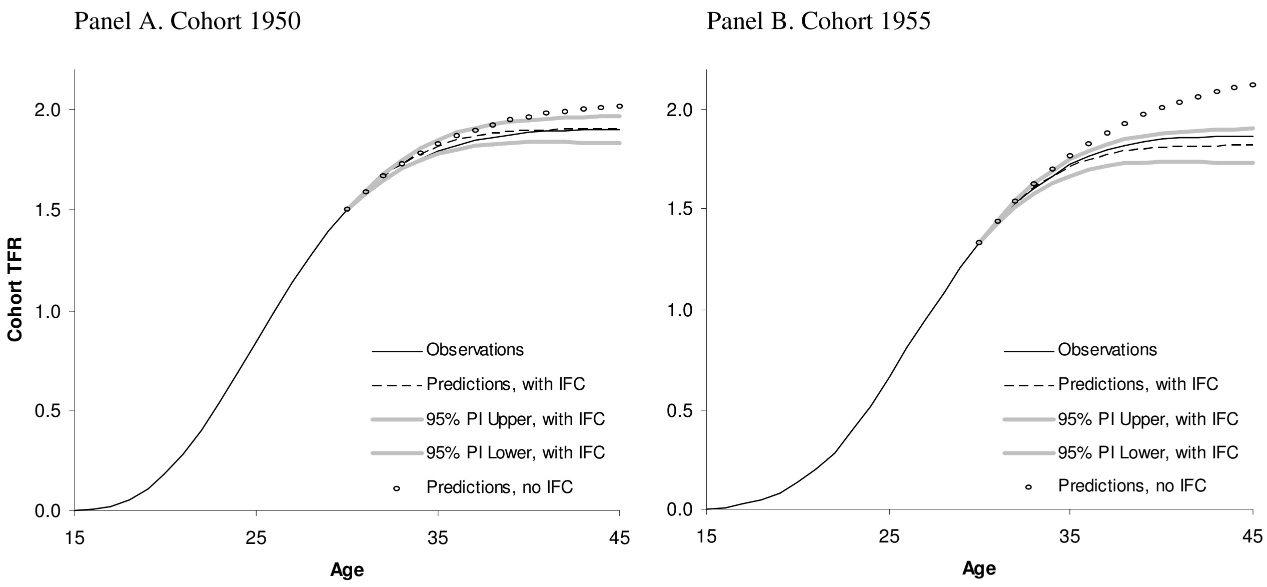


Panel A. Observed g(t), Gompertz model.

Panel B. g(t): Observed (black) and predicted (grey). Gompertz model with the infecundity correction

Notes: We estimate the IFC by first estimating cohort specific drifts for ages up to 30. We then fit a model $g_{c,t} = g_{c,t-1} + \delta_c \cdot IFC^{(t-30)}$ to ages above 30 and estimate IFC by minimizing the weighted sum of squared errors. The weights $w_t = (45 - t + 1)/(15 \cdot 8)$ decline linearly so that $w_{31} = 1/8$ and $w_{45} = 1/(15 \cdot 8)$. This specification gives more weight for younger ages whose fertility contribution is larger than that of older ages. Note that while the drifts are cohort specific, IFC is shared by the cohorts. This estimated IFC is 1.118; that is, for each additional year above age 30, the pace of the decline in g(t) accelerates by 11.8%.

Figure 4. Predictions and 95% prediction intervals (PI) for cohort fertility with and without the infecundity correction (IFC); Dutch female cohorts 1950 and 1955. Predictions use data up to age 30 and the Gompertz model with random walk with drift as the underlying process. Data: Human Mortality Database (2010).



Panel A. Cohort 1950

Panel B. Cohort 1955

Figure 5. Predictions and 95% prediction intervals (PIs) for completed cohort fertility at ages 30, 35 and 45 for Dutch female cohorts 1950-1977. Predictions and 95% PIs are based on the Gompertz model with infecundity correction and random walk with drift as the underlying process. For each cohort, data is used until the last point of observation (year 2008).

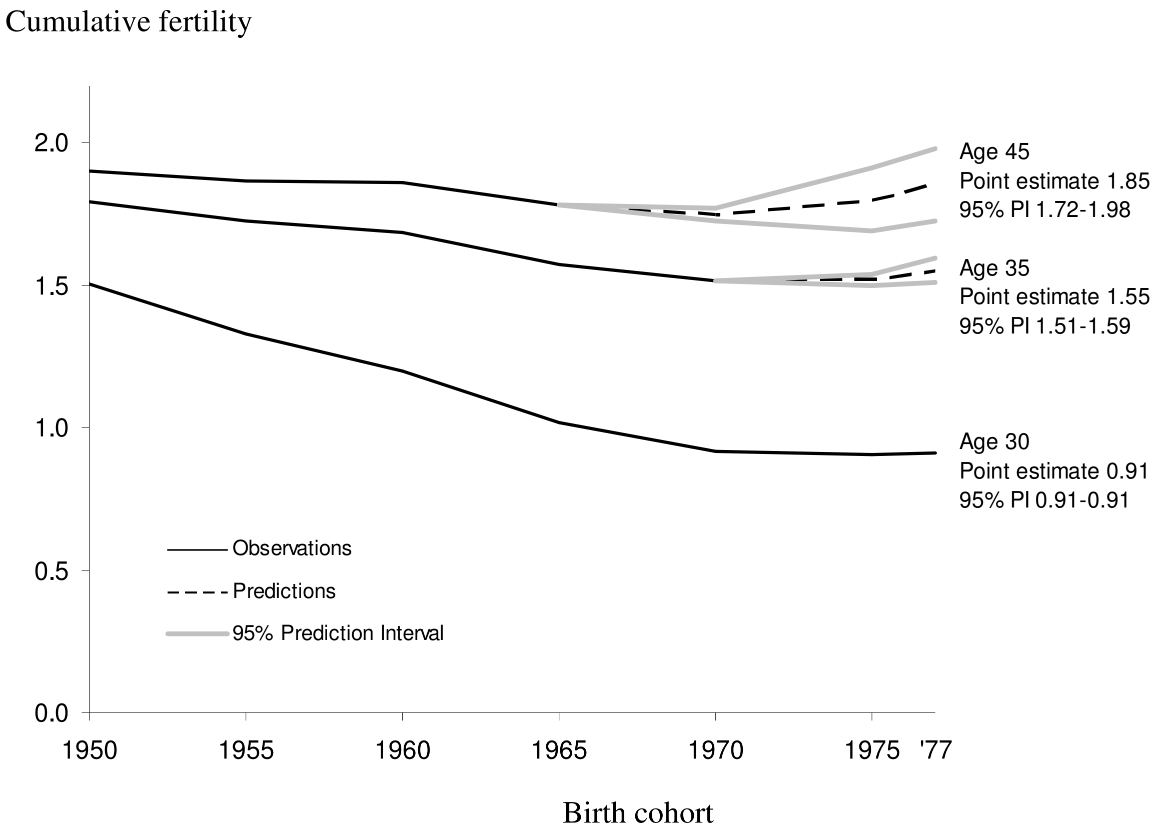


Table 1. Summary of the Hernes, Gompertz and logistic models with a random walk with drift as the underlying process.

	Hernes	**Gompertz**	**Logistic**
1. Model equation	$\frac{dP_t}{dt} = ab^t P_t(1-P_t)$	$\frac{dP_t}{dt} = a\exp(-bt)P_t$	$\frac{dP_t}{dt} = bP_t\left(1-\frac{P_t}{a}\right)$
2. Linearization	$\ln\left(\frac{dP_t}{dt}\frac{1}{P_t(1-P_t)}\right) = \ln a + t\ln b \equiv g_t$	$\ln\left(\frac{d\ln P_t}{dt}\right) = \ln a - bt \equiv g_t$	$\ln\left(\frac{dP_t}{dt}\frac{1}{P_t^2}\right) = \ln(b/a) + a - bt \equiv g_t$
3. Model for g; estimators $\hat{\delta}$, $\hat{\sigma}_\varepsilon^2$; predictions $\hat{g}_{t+k}$	Model: $g_t = g_0 + \delta t + \sum_{i=1}^{t}\varepsilon_i$	Estimators: $\hat{\delta} = \frac{g_{t-1} - g_1}{t-2}$; $\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{t-1}\left(g_i - g_{i-1} - \hat{\delta}\right)^2}{t-3}$	Predictions: $\hat{g}_{t+k} = g_t + \hat{\delta}k$
4. Predictions $\hat{P}_{t+k}$	$\hat{P}_{t+k} = \hat{P}_{t+k-1}$ $+\hat{P}_{t+k-1}\left(1-\hat{P}_{t+k-1}\right)\exp\left(\hat{g}_{t+k}\right)$	$\hat{P}_{t+k} = \frac{\hat{P}_{t+k-1}}{1-\exp(\hat{g}_{t+k})}$	$\hat{P}_{t+k} = \hat{P}_{t+k-1} + \hat{P}_{t+k-1}^2\exp\left(\hat{g}_{t+k}\right)$
5. Variance $V\left(\hat{P}_{t+k}\right)$	General form: $V\left(\hat{P}_{t+k}\right) = \sigma_\varepsilon^2\exp(2g_t)\sum_{i=1}^{k}\sum_{j=1}^{k}\min(i,j)\exp[\delta(i+j)]\gamma_{t+i-1}\gamma_{t+j-1}$, where γ_{t+i-1} is:		
	$\gamma_{t+i-1} = \hat{P}_{t+i-1}\left(1-\hat{P}_{t+i-1}\right)$	$\gamma_{t+i-1} = 1$	$\gamma_{t+i-1} = \hat{P}_{t+i-1}^2$