Chinese Business Review, ISSN 1537-1506
April 2011, Vol. 10, No. 4, 311-316

# Applying the Non-homogeneous Stochastic Gompertz Process for Modeling Populations

María Dolores Huete-Morales, Francisco Abad Montes
University of Granada, Granada, Spain

It is a well known fact that studies on growth primarily take into account human populations, although currently, many scientific fields (biology, economics, etc.) also use growth models to reflect behaviours in diverse phenomena. These deterministic models are difficult to apply in real populations since, as we know, the volume of a human population depends intrinsically on a large number of other socio-economic variables, including changes in fertility patterns, improvements in living conditions, individual health factors which produce an increase or decrease in the number of years lived, the state of economic well-being, or changes in migratory fluxes. In this study, we have examined the stochastic Gompertz non-homogenous diffusion process, analysing its transition probability density function and conducting inferences on the parameters of the process through discrete sampling. All of the results are applied to the population of Andalusia with data disaggregated by sex during the period of 1981 to 2002, taking purely demographic variables as exogenous factors: life expectancy at birth, foreign immigration to Andalusia and total fertility rate.

*Keywords:* Gompertz diffusion process, exogenous factors, demography, population

## Introduction

It becomes necessary to use frameworks in order to make provision for population adjustment, including diffusion processes, which are widely used in growth models (Suddhendu, 1988). The inclusion of exogenous factors in such models has a clear advantage, since they allow us to consider variables which influence population growth, which in turn allows for a clear improvement in the modelling of phenomena. In any case, this process proves to be an innovative way to establish or adjust population growth since normally, deterministic growth models are used (which are dependent to a large degree on population growth rates). As far as the Gompertz process is concerned, this model was introduced by Ricciardi (1977) who considered applications in the field of biology, and by Crow and Shimizu (1998). Later, Gutiérrez, Gutiérrez-Sánchez, Nafidi, Rom_an and Torres (2005) conducted inferences on the said process and examined discrete trajectories. In Ferrante (2000), continuous trajectories of the process were considered and applied to tumour growth. The non-homogenous case in the Gompertz process through the use of exogenous variables has been defined by Nafidi (1997) in a general context and has been applied more recently by Gutiérrez in dealing with the problem of inference by considering

María Dolores Huete-Morales, lecturer, Department of Statistics & O. R., Faculty of Sciences, University of Granada.
Francisco Abad Montes, lecturer, Department of Statistics & O. R., Faculty of Sciences, University of Granada.

discrete trajectories in the said process.

We examine the non-homogenous univariate Gompertz process, which includes a series of exogenous variables within the trends. Initially, the likelihood function is obtained, which brings with it about the problem of the need for knowing the implicit expression of the functions related to the exogenous factors in order to be able to conduct inference about the parameters. For this reason, a specific case is considered, thus facilitating estimates on these parameters. Finally, an exhaustive study is conducted into the application of previous theoretical results. The Andalusian population is taken as an endogenous variable; the number of immigrants proceeding from foreign countries, life expectancy at birth and the synthetic indicator of fertility, taken in all cases for men and women (1981-2002), are used as exogenous variables.

## The Non-homogenous Gompertz Diffusion Process

### Characterisation

Let $\{X(t), t_0 \leq t \leq T\}$ be a one-dimensional diffusion process, *R*-valued and with transition distribution function:

$$P(y,t|x,s) = P(X(t) = y|X(s) = x)$$

If we consider as infinitesimal moments (drift and diffusion coefficients) of the process respectively:

$$a(t,x) = g(t)x - h(t)x\log(x),\ b(t,x) = \sigma^2 x^2$$

with *h* and *g* as two continuous and parametric functions, which, in other words, may depend on a certain number of parameters and $\sigma > 0$, we have the unidimensional Gompertz diffusion process with exogenous factors, with the following diffusion equations:

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial y}\big((g(t)y - h(t)y\log(y))p\big) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2 y^2 p\right)$$

$$\frac{\partial p}{\partial s} = -\big((g(s)x - h(x)x\log(x))\big)\frac{\partial p}{\partial x} - \frac{1}{2}\left(\sigma^2 x^2\right)\frac{\partial^2 p}{\partial x^2}$$

where *p* is the density of transition function. The transition distribution results:

$$P(y,t|x,s) = \left(2\pi\sigma^2 e^{-2\int_s^t h(z)dz}\int_s^t e^{2\int_s^\theta h(z)dz}d\theta\right)^{-\frac{1}{2}} y^{-1}$$

$$\exp\left(-\frac{1}{2}\frac{\left[\log(y) - \log(x)e^{-\int_s^t h(z)dz} - e^{-\int_s^t h(z)dz}\int_s^t k(\theta)e^{-\int_\theta^t h(z)dz}d\theta\right]^2}{\sigma^2 e^{-2\int_s^t h(z)dz}\int_s^t e^{2\int_s^\theta h(z)dz}d\theta}\right)$$

with the above, we can deduce that the *r*-order conditional moments of the endogenous variable:

$$E\left(X^r(t)|X(s) = x\right) = \exp\left(r\log(x)e^{-\int_s^t h(z)dz} + re^{-\int_s^t h(z)dz}\int_s^t k(\theta)e^{-\int_\theta^t h(z)dz}d\theta + \right.$$

$$\left. + \frac{r^2\sigma^2}{2}e^{-2\int_s^t h(z)dz}\int_s^t e^{2\int_s^\theta h(z)dz}d\theta\right)$$

and by taking *r* = 1, we immediately obtain the first-order conditional moment (conditioned trend function, CTF):

$$E\left(X(t)\middle|X(s)=x\right)=\exp\left(\log(x)e^{-\int_s^t h(z)dz}+e^{-\int_s^t h(z)dz}\int_s^t k(\theta)e^{-\int_\theta^t h(z)dz}d\theta+\right.$$
$$\left.+\frac{\sigma^2}{2}e^{-2\int_s^t h(z)dz}\int_s^t e^{2\int_s^\theta h(z)dz}d\theta\right)$$

**Inference in the Model**

In order to find estimators of the process parameters, we use the maximum likelihood method and we consider discrete sampling, in other words, a realization of the same in the instants $(t_0,t_1,\dots,t_n)$, $X(t_0)=x_0;X(t_1)=x_1;\dots,X(t_n)=x_n$ with the initial condition $P[X(t_0)=x_0]=1$. If we indicate:

$$m_\alpha=\log(x_{\alpha-1})e^{-\int_{t_{\alpha-1}}^{t_\alpha}h(z)dz}-e^{-\int_{t_{\alpha-1}}^{t_\alpha}h(z)dz}\int_{t_{\alpha-1}}^{t_\alpha}k(\theta)e^{-\int_\theta^{t_\alpha}h(z)dz}d\theta$$

the associated log-likelihood function will be:

$$\log(L)(x_0,x_1,...,x_n)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{1}{2}\sum_{\alpha=1}^{n}\log\left(e^{-2\int_{t_{\alpha-1}}^{t_\alpha}h(z)dz}\int_{t_{\alpha-1}}^{t_\alpha}e^{2\int_{t_{\alpha-1}}^{\theta}h(z)dz}d\theta\right)-\sum_{\alpha=1}^{n}\log(x_\alpha)$$
$$-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}\frac{[\log(x_\alpha)-m_\alpha]^2}{\left(e^{-2\int_{t_{\alpha-1}}^{t_\alpha}h(z)dz}\int_{t_{\alpha-1}}^{t_\alpha}e^{2\int_{t_{\alpha-1}}^{\theta}h(z)dz}d\theta\right)}$$

Logically, in order to be able to minimise this function with regards to the unknown parameters, we need to know the form of the functions *h* and *g*, which is not always possible; so, we consider that the functions *h* and *g* to be $h(t)=\beta$ and $g(t)=\alpha_0+\sum_{i=1}^{q}\alpha_i g_i(t)$, where the exogenous variables $g_i(t)$ are continuous functions in $[t_0,T]$. In this way, the stochastic differential equation which characterizes the process is:

$$dx(t)=\{g(t)x(t)-\beta x(t)\log x(t)\}dt+\sigma x(t)dw(t) \qquad (1)$$

On differentiating the log-likelihood in relation to *a*, $\sigma^2$ and *β*, the following equations appear:

$$U_\beta v_\beta=U_\beta U'_\beta a$$
$$n\sigma^2=(v_\beta-U'_\beta a)'(v_\beta-U'_\beta a)$$
$$\left(v_\beta{}^{-1}e^{-\beta}l'_x-a'\frac{\partial U_\beta}{\partial\beta}\right)(v_\beta-U'_\beta a)=0$$

where $l'_x=(\log(x_1),\dots,\log(x_n))'$ and $\frac{\partial U_\beta}{\partial\beta}$ the matrix formed by those derived from the elements of $U_\beta$ in relation to *β*. In this way, we obtain the maximum likelihood estimators of *a* and $\sigma^2$:

$$\hat{a}=\left(U_{\hat\beta}U'_{\hat\beta}\right)^{-1}\left(U_{\hat\beta}v_{\hat\beta}\right) \qquad (2)$$

$$n\hat\sigma^2=v'_{\hat\beta}H_{U,\hat\beta}v_\beta \qquad (3)$$

with $H_{U,\hat\beta}=I_n-U'_{\hat\beta}\left(U_{\hat\beta}U'_{\hat\beta}\right)^{-1}U_{\hat\beta}$ idempotent symmetric matrix. The estimator of *β* is obtained by substituting equation (2) and equation (3) in the third likelihood equation, leaving the following expression:

$$\left(v_\beta{}^{-1}e^{-\beta}l'_x-v_{\hat\beta}U'_\beta\left(U_{\hat\beta}U'_{\hat\beta}\right)^{-1}\frac{\partial U_\beta}{\partial\beta}\right)H_{U,\beta}v_\beta=0 \qquad (4)$$

Due to the fact that in equation (4), the functions $g_j(t)$ appear, it is not possible to have an explicit estimator expression of *β*, since such functions may only be known as a result of discrete observations of the exogenous variables $y_{ij};i=1,...,n;j=1,...,q$. For this reason, exogenous factors are normally constructed from observed values of the variables through polygonal functions:

$$g_j(t) = y_{i-1,j} + \left(y_{i,j} - y_{i-1,j}\right)\left(t - t_{i-1}\right) \tag{5}$$

thus, if we indicate

$$z_{ij}(\beta) = y_{i-1,j} + \left(y_{i,j} - y_{i-1,j}\right)\frac{\beta - 1 + e^{-\beta}}{\beta\left(1 - e^{-\beta}\right)}$$

we can state:

$$\int_{t_{i-1}}^{t_i} g_j(\xi)e^{-\beta(t_i-\xi)}d\xi = \gamma_\beta z_{ij}(\beta)$$

## Application to the Andalusian Population

The collected variables (or exogenous factors) to adjust the non-homogenous model to the Andalusian population disaggregated by sex for the 1981-2002 period consist of foreign immigrants, life expectancy at birth (e0) or mean number of remaining years of life of new-born children, and a synthetic indicator of fertility, total fertility rate (TFR), or number of children per mother at a fertile age. The population by sex has been obtained from the Andalusian Institute of Statistics (Instituto de Estadística de Andalucía, IEA). The information on the number of foreign immigrants is provided by the National Institute of Statistics (Instituto Nacional de Estadística, INE) through the Residential Variations Statistics (it should be noted here that until the year 1983, disaggregation by sex is only estimated by the INE since, up until that time, the sex of immigrants was not recorded). The values for life expectancy at birth have been elaborated here through the construction of biannual mortality tables based on information provided by the IEA. In the following applications, the TFR indicator for both female and male populations has been used as an exogenous variable. It is true that the male TFR may be calculated, but this measurement is not frequently employed, and there is no significant variation between the indicator for males and for females. The female Andalusian population has been used as an endogenous variable and the number of foreign female immigrants, female life expectancy at birth and total fertility rate have been employed as exogenous variables. As has been previously mentioned, the period of observation is from 1981 to 2002, although the final two years have been reserved in order to carry out predictions and to test the goodness of the model. The exogenous factors have been constructed from the said observations, taking into account polygonal functions of the type (5). The $\beta$ parameter is estimated by means of numeric procedures on equation (4) using the "Mathematica" software package; this value is used with equation (2) and equation (3) with the objective of obtaining the remaining estimated parameters. These estimates are shown in Table 1.

Table 1

*Estimated Parameters*

| Parameter | Estimated value (women) | Estimated value (men) |
|---|---|---|
| $\beta$ | 0.0378191 | 0.0797127 |
| $\alpha_0 - \sigma^2/2$ | 0.0507695 | 0.1031180 |
| $\alpha_1$ | -0.0005712 | -0.0029721 |
| $\alpha_2$ | 1.2287000 | 1.0394100 |
| $\alpha_3$ | 0.0066227 | 0.0271248 |
| $\sigma^2$ | $2.91204 \times 10^{-6}$ | $3.32719 \times 10^{-6}$ |

With these values, the trend function and the conditioned trend of the process are estimated, which are graphically represented in Figure 1 and Figure 2.

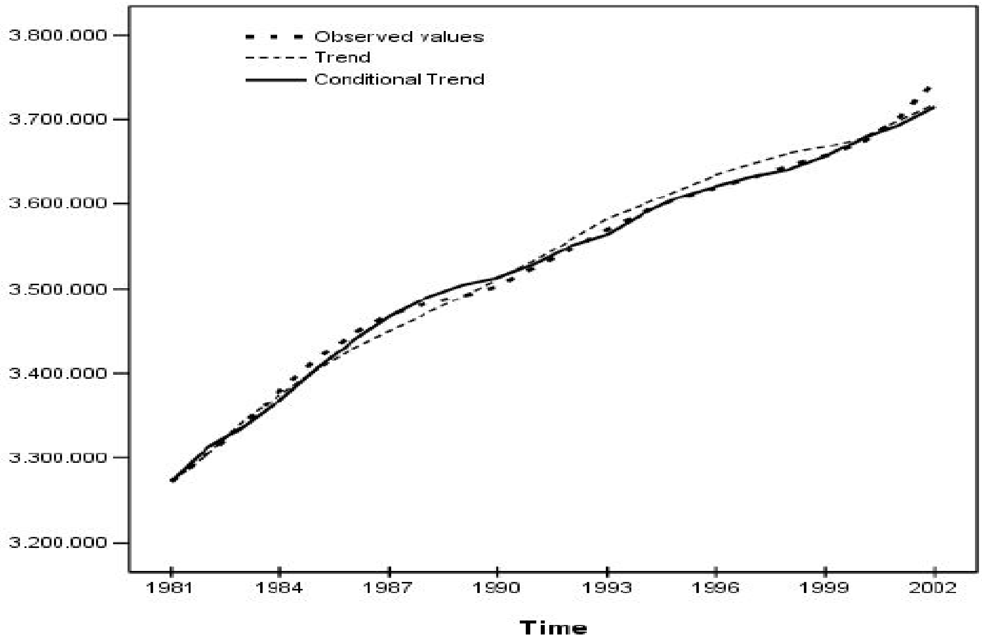

*Figure 1.* Estimated trend functions (women).

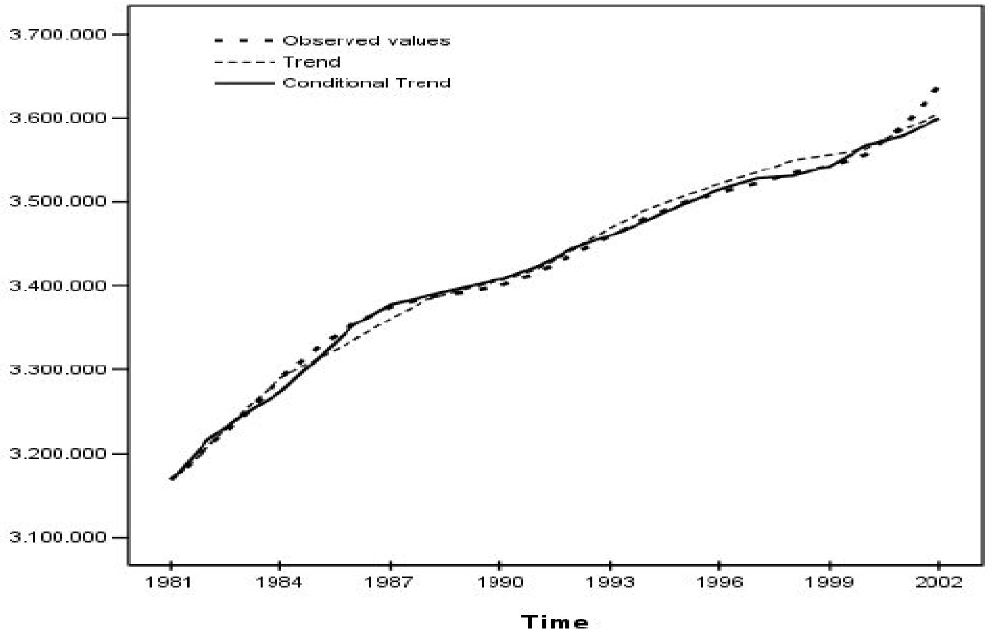

*Figure 2.* Estimated trend functions (men).

In relation to male Andalusian population, we again assume the number of foreign male immigrants, male life expectancy at birth and the total fertility rate as exogenous variables. The estimated parameters in the model are shown in Table 1 and the estimated trend functions in Figure 2.

## Conclusions

The conditioned trend function represents the population observations better than the non-conditioned trend function (NCTF). However, the use of the NCTF, while it does not register small fluctuations, is more advantageous in terms of making future population predictions because in obtaining the NCTF in one instant, it is not necessary to know the observed value of the previous instant. In order to do this, we only need to establish a series of hypotheses for the values of the exogenous variables. In fact, in any projection of population it is necessary to have previously carried out suppositions on the behaviour of the basic demographic indicators (life expectancy, TFR, migrations). In this way, a wide range of possibilities is opened, since the behaviour of the population at each age (or age intervals) and predictions made by age can be studied. From the absolute value of the conditioned trend function and the observed population, the errors committed in each year of observation have been calculated (%) as a coefficient of the difference between the observed population value and the estimated population value. The differences between the observed and estimated values in a small number of years reach 0.5% of the observed population (only in the year 2002, which is reserved for predictions, does it go beyond this percentage level). This is an indicator that this non-homogenous Gompertz model can acceptably represent population behaviour.

## References

Crow, E. L., & Shimizu, K. (1988). *Lognormal distribution theory and application.* New York: Dekker.

Ferrante, L., Bompadre, S., Possati, L., & Leone, L. (2000). Parameter estimation in a gompertzian stochastics model for tumor growth. *Biometrics, 56,* 1076-1081.

Gutiérrez, R., Gutiérrez-Sánchez, R., Nafidi, A., Rom_an, P., & Torres, F. (2005). Inference in gompertz-type nonhomogeneous stochastic systems by means of discrete sampling. *Cybernetics and Systems, 36,* 203-216.

Huete, M. D.(2006). El modelo estocástico de Gompertz. Modelización de datos sociodemográficos (Ph.D. thesis, University of Granada).

Nafidi, A. (1997): Difusiones Lognormales con factores exógenos. Extensiones a partir proceso de difusión de Gompertz (Ph.D. thesis, University of Granada).

Ricciardi, L.M. (1977). Diffusion processes and related topics in biology. *Lect.Notes Biomath, 14.* Springer Verlag.

Suddendun, B. (1988). *Stochastic processes in demography and applications.* New Delhi: Wiley Eastern Limited.