mathematics

MDPI

Article

Inference with Non-Homogeneous Lognormal Diffusion Processes Conditioned on Nearest Neighbor

Ana García-Burgos [1,*], Paola Paraggio [2], Desirée Romero-Molina [1] and Nuria Rico-Castro [1]

[1] Departamento de Estadística e IO, Universidad de Granada, 18012 Granada, Spain; deromero@ugr.es (D.R.-M.); nrico@ugr.es (N.R.-C.)
[2] Dipartimento di Matematica, Università di Salerno, 84084 Fisciano, Italy; pparaggio@unisa.it
* Correspondence: agburgos@ugr.es

Abstract: In this work, we approach the forecast problem for a general non-homogeneous diffusion process over time with a different perspective from the classical one. We study the main characteristic functions as mean, mode, and α-quantiles conditioned on a future time, not conditioned on the past (as is normally the case), and we observe the specific formula in some interesting particular cases, such as Gompertz, logistic, or Bertalanffy diffusion processes, among others. This study aims to enhance classical inference methods when we need to impute data based on available information, past or future. We develop a simulation and obtain a dataset that is closer to reality, where there is no regularity in the number or timing of observations, to extend the traditional inference method. For such data, we propose using characteristic functions conditioned on the past or the future, depending on the closest point at which we aim to perform the imputation. The proposed inference procedure greatly reduces imputation errors in the simulated dataset.

Keywords: diffusion processes; Gompertz-lognormal; conditioned on future; nearest neighbor; imputation; simulated sample paths; characteristic functions

MSC: 62M20; 62D10; 62P10; 62F10

Citation: García-Burgos, A.; Paraggio, P.; Romero-Molina, D.; Rico-Castro, N. Inference with Non-Homogeneous Lognormal Diffusion Processes Conditioned on Nearest Neighbor. *Mathematics* **2024**, *12*, 3703. https://doi.org/10.3390/math12233703

Received: 8 November 2024
Revised: 22 November 2024
Accepted: 23 November 2024
Published: 26 November 2024

1. Introduction

Lognormal diffusion processes are widely used in various fields of application to model phenomena where the growth rate of a system is proportional to its current state, leading to multiplicative noise. These processes, which are characterized by lognormal distributions, are particularly effective in capturing the dynamics of systems in which values are closely positive, such as stock prices, biological growth, or certain physical processes, making them a key tool in both finance and natural sciences.

Modeling random phenomena through diffusion processes has been widely applied in various fields. One of the most relevant applications of modeling phenomena is its use to study tumor growth. Therefore, various publications explore this research line, as can be seen in [1–9].

Other types of phenomena related to natural growth can also be modeled using non-homogeneous diffusion processes. Some authors have used such models, from the early patterns of population growth [10,11] to growth in rabbits [12], mean weight of swordfish [13], microorganisms in culture [14], pig growth data [15], and the spread of COVID-19 [16].

Recent studies have used lognormal non-homogeneous diffusion processes to model the propagation of fake news [17,18].

As the literature shows, the use of non-homogeneous lognormal diffusion processes has spread to model different growth mechanisms. For all these applications of the lognormal non-homogeneous diffusion process, the authors usually recall known functions, improve different aspects of the inference procedure, including new parametrizations,

functions, and new methodologies, and provide more comprehensive and interpretable approaches.

We usually find studies focused on modeling observed as well as simulated data, which are detected at regular times, usually equidistant from each other and observed at the same instants in all sample paths. Hence, because of the study of processes, it is possible to make inferences about occurrences at moments later than those observed; that is, forecasting the future by knowing the past and present. However, the observation of phenomena that can be modeled as a non-homogeneous lognormal diffusion process does not always respond to this scheme. Moreover, often, these observations are not presented with the same number for different sample paths, nor are they observed at coincident times. This may respond to different observation opportunities that the researcher is not able to control and could complicate the modeling procedure and the comparison between different individuals, given the different times in which they are observed.

Classical methodology assumes that future events can be predicted by considering past information. However, when events are observed in a non-systematic manner (at different times) and with varying frequencies across different sample paths, it is sometimes not possible to make a point estimation using a characteristic function conditioned on past values, as such observations may not exist or be available. This limitation reduces the effectiveness of classical point estimation for values at many time points, particularly those preceding the first observation.

In some cases, a previous observation may be available, allowing for imputation using the classical method. However, if that observation is far in the past, and a posterior observation exists, the classical method disregards the latter. Since the conditioned function only considers past values, it overlooks the potential value of the closer, later observation, thus missing valuable information that could improve the imputation.

We propose a methodology to use non-homogeneous lognormal diffusion processes for modeling and making inferences. Specifically, we propose inference using information not only from the past but also from the future, particularly using the distribution of the process $X(t)$ conditioned on the observed instant of time that is closest to time t. Specifically, if the available information closest to the point of interest is found in a past time, the usual inference is used for an instant t considering the distribution of $X(t)|X(s)$, $s < t$, conditioned from the information at a previous instant s. On the contrary, if the time t whose value $X(t)$ we desire to infer is closer to an available future observed value v ($t < v$), we use the distribution of $X(t)|X(v)$, conditioned on the future, to obtain the inference at t.

For this purpose, we first recall the non-homogeneous lognormal diffusion process and the classical distribution conditional on the past. In second place, we study certain particular non-homogeneous diffusion processes and derive their future-conditioned distributions. Next, we obtain the distribution conditional on the future. Finally, we illustrate the proposal procedure using simulated data generated such that each sample path contains a random amount of data observed at unequal times. The maximum likelihood estimation of the parameters is performed, obtaining point estimations. With them, inference on a set of common points is carried out using three different methods: (a) classically, by employing the process conditioned on the past, (b) contrarily, by employing the process conditioned on the future, and (c) in the proposed way, by alternately employing the process conditioned on the past or the future, contingent upon which instant observed is closest to the instant at which we want to infer the process. We compare the three inference methods for the three conditioned functions of interest: mean, mode, and median. After the comparison, we conclude that the third methodology greatly improves the classical imputation procedure.

2. Non-Homogeneous Lognormal Diffusion Processes

In this section, we recall the definition and some characteristic functions of the non-homogeneous lognormal diffusion process and summarize some particular cases of this process.

2.1. The Process

Following [19] , a diffusion process $X(t)$ with $t \in [t_0, +\infty)$, $t_0 \in \mathbb{R}_0^+$ and state space given by the set of positive real numbers $\mathbb{R}^+$ is a non-homogeneous lognormal diffusion process when its infinitesimal moments are given by the following:

$$A_1(x,t) := h_\theta(t)x \quad \text{and} \quad A_2(x) := \sigma^2 x^2,$$

where $h_\theta(t)$ is a positive, continuous, bounded, and differentiable function in any interval $[0,\tau)$, with $\tau \geq 0$, and $\sigma > 0$. The vector $\theta \in \Theta \subseteq \mathbb{R}^k$, $k \in \mathbb{N}$ contains all the parameters involved in the definition of the process. Such process is the solution of the following stochastic differential equation:

$$\begin{cases} \mathrm{d}X(t) = h_\theta(t)X(t)\mathrm{d}t + \sigma X(t)\mathrm{d}W(t) \\ X(t_0) = X_0, \end{cases} \tag{1}$$

where $W(t)$ represents a standard Wiener process independent on the initial state X_0. We suppose that X_0 is lognormally distributed, i.e.,

$$X_0 \sim \Lambda_1(\mu_0, \sigma_0^2), \tag{2}$$

with $\mu_0, \sigma_0 > 0$, or degenerate, i.e., $\mathsf{P}(X_0 = x_0) = 1$ with $x_0 > 0$. The choice of an initial distribution of this type leads to an explicit and manageable expression for the joint density function.

Thanks to the Itô's formula, it is possible to obtain an explicit solution to Equation (1), given as follows:

$$X(t) = X_0 \exp\big(H_{\xi}(t_0,t) + \sigma(W(t) - W(t_0))\big), \qquad t \geq t_0, \tag{3}$$

where

$$H_{\xi}(s,t) := \int_s^t h_\theta(u)\mathrm{d}u - \frac{\sigma^2}{2}(t-s), \qquad 0 \leq t_0 \leq s \leq t, \tag{4}$$

being

$$\xi = (\theta^T, \sigma^2). \tag{5}$$

As shown in [20], if X_0 has a lognormal distribution $\Lambda_1(\mu_0, \sigma_0^2)$ or if $\mathsf{P}(X_0 = x_0) = 1$, all the finite-dimensional joint distributions of the process are lognormal. Specifically, given $n \in \mathbb{N}$ and $t_0 < t_1 < \cdots < t_n$, the vector $(X(t_1), \ldots, X(t_n))^T$ follows a n-dimensional lognormal distribution, i.e.,

$$(X(t_1), \ldots, X(t_n)) \sim \Lambda_n(\epsilon, \Sigma), \tag{6}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ and $\Sigma = (\sigma_{i,j})_{i,j=1,\ldots,n}$ with

$$\epsilon_i = \mu_0 + H_{\xi}(t_0, t_i), \qquad \sigma_{i,j} = \sigma_0^2 + \sigma^2\big(\min(t_i, t_j) - t_0\big), \qquad i,j \in \{1,\ldots,n\}.$$

Hence, the transition density function is available in closed form and it is given, for $0 \leq t_0 < s < t$, by the following:

$$f(x,t \mid y,s) = \frac{1}{x\sqrt{2\pi\sigma^2(t-s)}} \exp\left(-\frac{\Big(\log(\frac{x}{y}) - H_{\xi}(s,t)\Big)^2}{2\sigma^2(t-s)} \right), \qquad x,y > 0.$$

Thus, it follows that

$$X(t) \mid X(s) = y \sim \Lambda_1\Big(\log y + H_{\xi}(s,t), \sigma^2(t-s)\Big), \qquad t_0 < s < t.$$

We remark that the process $X(t)$ is conditioned on the past, since $s < t$. As the joint distribution is known and is lognormal, other important characteristic functions of the process can be obtained. For example, the n-th moment of $X(t)$, with $n \in \mathbb{N}$, is given by the following:

$$\mathsf{E}[X(t)^n] = \mathsf{E}[X_0^n]\left(\exp\left(H_{\tilde{\xi}}(t_0,t) + \frac{\sigma^2}{2}(t-t_0)n\right)\right)^n, \qquad t \geq t_0.$$

In particular, the expected value of $X(t)$ is given as follows:

$$\mathsf{E}[X(t)] = \mathsf{E}[X_0]\exp\left(H_{\tilde{\xi}}(t_0,t) + \frac{\sigma^2}{2}(t-t_0)\right) = \mathsf{E}[X_0]\exp\left(\int_{t_0}^{t} h_\theta(u)\mathrm{d}u\right),$$

for $t \geq t_0$. From the expressions of the expected value and of the 2nd-order moment of $X(t)$, it is possible to obtain the variance of the process, which is given as follows:

$$\begin{aligned}\mathsf{Var}[X(t)] = \mathsf{E}\left[X_0^2\right]\exp\left(2H_{\tilde{\xi}}(t_0,t) + 2\sigma^2(t-t_0)\right)\\ - \left(\mathsf{E}[X_0]\right)^2\exp\left(2H_{\tilde{\xi}}(t_0,t) + \sigma^2(t-t_0)\right).\end{aligned}$$

Other characteristic functions of interest of the process $X(t)$ are the mode, which is given by

$$\mathsf{Mode}[X(t)] = \mathsf{Mode}[X_0]\exp\left(H_{\tilde{\xi}}(t_0,t) - \sigma^2(t-t_0)\right), \qquad t \geq t_0,$$

and the α-quantile

$$C_\alpha[X(t)] = \exp\left(H_{\tilde{\xi}}(t_0,t) + \mu_0 + z_\alpha\sqrt{\sigma_0^2 + \sigma^2(t-t_0)}\right), \qquad t \geq t_0,$$

where z_α denotes the upper α-quantile of a standard normal distribution. The median can be obtained by setting $\alpha = 0.5$:

$$\mathsf{Med}[X(t)] = \mathsf{Med}[X_0]\exp\left(H_{\tilde{\xi}}(t_0,t)\right), \qquad t \geq t_0.$$

Similarly, the conditional characteristic functions of the process are given by the following:

$$\begin{aligned}&\mathsf{E}[X(t)^n \mid X(s) = y] = \exp\left(n\left(\log y + H_{\tilde{\xi}}(s,t)\right) + \frac{n^2}{2}\sigma^2(t-s)\right),\\ &\mathsf{Mode}[X(t) \mid X(s) = y] = \exp\left(\log y + H_{\tilde{\xi}}(s,t) - \sigma^2(t-s)\right),\\ &C_\alpha[X(t) \mid X(s) = y] = \exp\left(\log y + H_{\tilde{\xi}}(s,t) + z_\alpha\sigma\sqrt{t-s}\right),\\ &\mathsf{Med}[X(t) \mid X(s) = y] = \exp\left(\log y + H_{\tilde{\xi}}(s,t)\right),\end{aligned}$$

with $0 \leq t_0 < s < t$. In the literature, the joint distribution of d sample paths of the process $X(t)$ is available in closed form. For completeness, we recall the expressions given, for example, in [19], by considering a discrete sampling of the process $X(t)$ based on d sample paths. For any sample path, we fix different observation times, denoted by $t_{i,j}$ with $i = 1,\ldots,d$ and $j = 1,\ldots,n_i$. The initial observation time is fixed and equal for any sample path and it is given by $t_{i1} = t_0$, $i = 1,\ldots,d$. Let $\mathbb{X} = \left(\mathbb{X}_1^T,\ldots,\mathbb{X}_d^T\right)^T$ be the matrix containing all the observations with $\mathbb{X}_i = \left(X(t_{i1}),\ldots,X(t_{i,n_i})\right)^T$ for any $i = 1,\ldots,d$. Assuming that the distribution of X_0 is degenerate, i.e., $\mathsf{P}(X_0 = x_0) = 1$, the joint probability density function of $\mathbb{X}$ is given by the following:

$$f_{\mathbb{X}}(x) = \prod_{i=1}^{d} \prod_{j=1}^{n_i-1} \frac{\exp\left(-\frac{\left[\log(x_{i,j+1}/x_{i,j}) - H_{\xi}(t_{i,j},t_{i,j+1})\right]^2}{2\sigma^2 \Delta_i^{j+1,j}} \right)}{x_{i,j+1}\sigma\sqrt{2\pi\Delta_i^{j+1,j}}}, \tag{7}$$

where $x = (x_{1,1}, \ldots, x_{1,n_1-1}, \ldots, x_{d,1}, \ldots, x_{d,n_d-1}) \in \mathbb{R}^n$ with $n = \sum_{i=1}^{d}(n_i - 1)$ and $\Delta_i^{j+1,j} = t_{i,j+1} - t_{i,j}$. From Equation (7), it is possible to obtain the log-likelihood function, which is given by the following:

$$\begin{aligned} L_{\mathbb{X}}(\xi) = &- \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{\left[\log(x_{i,j+1}/x_{i,j}) - H_{\xi}(t_{i,j},t_{i,j+1})\right]^2}{2\sigma^2 \Delta_i^{j+1,j}} - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) \\ &- \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \log\left(x_{i,j+1}\sqrt{\Delta_i^{j+1,j}} \right), \end{aligned}$$

where $H_{\xi}(s,t)$ is defined in Equation (4).

2.2. Particular Cases

The non-homogeneous lognormal diffusion process encompasses a family of processes with different characteristics, whose mean functions have certain shapes and have been widely studied and used in the literature to model different growths that can be found in nature or as an effect of human activity. Although the non-homogeneous lognormal diffusion process is defined from the expression of the stochastic differential equation written in Equation (1), depending on how the function $h_{\theta}(t)$ is considered, different named processes can be reached. Among these specific processes, we highlight some that are especially useful for modeling random phenomena. Therefore, they are valuable tools for decision-making in the fields of public health, ecology, and social sciences.

We consider the Gompertz type, introduced in [12], Gompertz lognormal studied in [21], the Bertalanffy process in the way observed in [13], the logistic-type as authors show us in [14], the Richards-type process seen in [22], the Korf process seen in [23], and the multi-sigmoidal process given in [24]. Table 1 shows which function $h_{\theta}(t)$ should be considered to work with each of them, as well as the mandatory $H_{\xi}(s,t)$ required to obtain the characteristic functions.

Table 1. Particular cases of non-homogeneous lognormal diffusion process.

Name	$h_{\theta}(t)$ Function	$H_{\xi}(s,t)$ with $(s < t)$
Gompertz-type $\xi = (m, \beta, \sigma^2)^T$	$me^{-\beta t}$ with $m, \beta > 0$	$-\frac{m}{\beta}\left(e^{-\beta t} - e^{-\beta s}\right) - \frac{\sigma^2}{2}(t-s)$
Gompertz-lognormal $\xi = (m, \beta, c, \sigma^2)^T$	$me^{-\beta t} + c$ with $m, \beta > 0$	$-\frac{m}{\beta}\left(e^{-\beta t} - e^{-\beta s}\right) - \left(\frac{\sigma^2}{2} - c\right)(t-s)$
Bertalanffy $\xi = (b, c, k, \sigma^2)^T$	$\frac{bck}{e^{kt} - c}$ with $k > 0$, $b \geq 1$, $t_0 \geq \frac{\log c}{k}$	$b\log\left(\frac{e^{kt}-1}{e^{ks}-1}\right) - \frac{\sigma^2}{2}(t-s)$
Logistic-type $\xi = (b, c, \sigma^2)^T$	$\frac{bc}{b + e^{ct}}$ with $b, c > 0$	$-bce^{-c}\log\left(\frac{b + e^c s}{b + e^c t}\right) - \frac{\sigma^2}{2}(t-s)$
Richards-type $\xi = (q, k, \eta, \sigma^2)^T$	$-\frac{qk^t \log k}{\eta + k^t}$ with $q, \eta > 0$, $0 < k < 1$	$q\log\left(\frac{\eta + k^s}{\eta + k^t}\right) - \frac{\sigma^2}{2}(t-s)$
Korf $\xi = (m, \beta, \sigma^2)^T$	$mt^{-(\beta+1)}$ with $m, \beta > 0$	$-\frac{m}{\beta}\left(t^{-\beta} - s^{-\beta}\right) - \frac{\sigma^2}{2}(t-s)$

Table 1. *Cont.*

Name	$h_\theta(t)$ Function	$H_\xi(s,t)$ with $(s<t)$
Multi-sigmoidal $\xi=(\eta,\beta_1,\ldots,\beta_p,\sigma^2)^T$	$\frac{P_\beta(t)e^{-Q_\beta(t)}}{\eta+e^{-Q_\beta(t)}}$, with $Q_\beta(t)=\sum_{i=1}^{p}\beta_i t^i$, $\beta_p>0$, $(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p$, $\eta>0$, $P_\beta(t)=\frac{d}{dt}Q_\beta(t)$	$\log\left[\frac{\eta+e^{-Q_\beta(s)}}{\eta+e^{-Q_\beta(t)}}\right]-\frac{\sigma^2}{2}(t-s)$

3. Non-Homogeneous Lognormal Diffusion Processes Conditioned on the Future

In real-world applications, non-homogeneous lognormal processes are used not only for predictive purposes but also for modeling. In this case, it is sometimes not possible to model at a point in time by conditioning on the past because this information is not available, as the paths of the observed data may be incomplete. Such situations suggest another way to use characteristic functions (mean, mode, and median), conditioning not on past values but on future observed values. This can be useful in various instances, such as to impute missing data for processes that have already been observed, or to infer at instants of time conditional on arbitrary future values under certain scenarios.

In this section, we analyze the main features of the process $X(t)$ conditioned on the future, i.e., $X(s)|X(t)=y$, with $t_0<s<t$ which, as we will see in Section 4, will be immensely useful for imputing missing data and completing the classical estimation method.

In detail, the following proposition provides the expression of the distribution conditioned on the future.

Proposition 1. *The process $X(s)$, given $X(t)=y$ with $t_0<s<t$, follows a one-dimensional lognormal distribution, i.e.,*

$$X(s)\mid X(t)=y\sim\Lambda_1\Big(m_{\xi,\mu_0,\sigma_0^2}(s\mid y,t),s_{\sigma_0^2,\sigma^2}(s\mid t)\Big),$$

where

$$m_{\xi,\mu_0,\sigma_0^2}(s\mid y,t):=\mu_0+H_\xi(t_0,s)+\frac{\sigma_0^2+\sigma^2(s-t_0)}{\sigma_0^2+\sigma^2(t-t_0)}\cdot\big(\log y-\mu_0-H_\xi(t_0,t)\big),\tag{8}$$

and

$$s_{\sigma_0^2,\sigma^2}(s\mid t):=\frac{\big(\sigma_0^2+\sigma^2(s-t_0)\big)\sigma^2(t-s)}{\sigma_0^2+\sigma^2(t-t_0)},\tag{9}$$

where μ_0,σ_0 are the parameters of the initial lognormally distributed state $X(t_0)$, $y>0$, and $H_\xi(t_0,t)$ is defined in Equation (4). (Note that a degenerate distribution is the particular case when taking $\sigma_0^2=0$.)

Proof. It is known that when (X,Y) follows a two-dimensional lognormal distribution and Y a one-dimensional lognormal distribution, then $X\mid Y=y$ also follows a one-dimensional lognormal distribution. More precisely, if

$$(X,Y)\sim\Lambda_2\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix},\begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}\right),\qquad Y\sim\Lambda_1\Big(\mu_Y,\sigma_Y^2\Big),$$

with $\mu_X,\mu_Y,\sigma_{XY}\in\mathbb{R}$, $\sigma_X,\sigma_Y>0$ then

$$X\mid Y=y\sim\Lambda_1\left(\mu_X+\rho\frac{\sigma_X}{\sigma_Y}(\log y-\mu_Y);\sigma_X^2\Big(1-\rho^2\Big)\right),$$

with $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$. Consequently, considering the joint distribution of $(X(s), X(t))$, $t_0 < s < t$, obtained in Section 2, i.e.,

$$(X(s), X(t)) \sim \Lambda_2(\epsilon, \Sigma), \qquad t_0 < s < t,$$

where

$$\epsilon = (\epsilon_1, \epsilon_2) = \left(\mu_0 + H_\xi(t_0, s), \mu_0 + H_\xi(t_0, t)\right)^T,$$
$$\Sigma = \begin{pmatrix} \sigma_0^2 + \sigma^2(s - t_0) & \sigma_0^2 + \sigma^2(s - t_0) \\ \sigma_0^2 + \sigma^2(s - t_0) & \sigma_0^2 + \sigma^2(t - t_0) \end{pmatrix}$$

and the fact that $X(t) \sim \Lambda_1\left(\mu_0 + H_\xi(t_0, t), \sigma_0^2 + \sigma^2(t - t_0)\right)$, we finally obtain the desired result. □

From Proposition 1, it is easy to obtain the explicit expressions of some characteristic functions of interest of the process $X(s)|X(t) = y$ with $t_0 < s < t$, such as the n-th moment, $n \in \mathbb{N}$, the mode, and the α-quantile. More in detail, the n-th moment of $X(s)|X(t) = y$ with $n \in \mathbb{N}$ is given by

$$\mathsf{E}[X(s)^n \mid X(t) = y] = \exp\left(n\, m_{\xi, \mu_0, \sigma_0^2}(s \mid y, t) + \frac{n^2}{2}\, s_{\sigma_0^2, \sigma^2}(s \mid t)\right), \qquad t_0 < s < t,$$

and the expected value of $X(s) \mid X(t) = y$ is given by

$$\mathsf{E}[X(s) \mid X(t) = y] = \exp\left(m_{\xi, \mu_0, \sigma_0^2}(s \mid y, t) + \frac{1}{2}\, s_{\sigma_0^2, \sigma^2}(s \mid t)\right), \qquad t_0 < s < t.$$

Moreover, the mode of $X(s) \mid X(t) = y$ is

$$\mathsf{Mode}[X(s) \mid X(t) = y] = \exp\left(m_{\xi, \mu_0, \sigma_0^2}(s \mid y, t) - s_{\sigma_0^2, \sigma^2}(s \mid t)\right), \qquad t_0 < s < t,$$

the α-quantile is

$$C_\alpha[X(s) \mid X(t) = y] = \exp\left(m_{\xi, \mu_0, \sigma_0^2}(s \mid y, t) - z_\alpha \sqrt{s_{\sigma_0^2, \sigma^2}(s \mid t)}\right), \qquad t_0 < s < t,$$

and the median is given by

$$\mathsf{Med}[X(s) \mid X(t) = y] = \exp\left(m_{\xi, \mu_0, \sigma_0^2}(s \mid y, t)\right), \qquad t_0 < s < t.$$

To determine the expressions of these characteristic functions for the processes summarized in Table 1, it is enough to particularize with the function $H_\xi(t_0, s)$ indicated in the same table.

For the sake of clarity, in Appendix A, Table A1 summarizes the symbols and the notation used, together with the corresponding meaning.

4. Simulation Study

To properly develop our simulation study, we replicate real-life observations. We approach our data to a matrix similar to those provided by real data, as researchers find via cohort studies like Helsinki Birth Court (Finland) [25], The Japan Environment and Children's Study (JECS) [26] (Japan), the Danish National Birth Cohort (Denmark) [27], the data of the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) [28] (USA), the Pelotas (Brazil) birth cohort study [29], the INfancia y Medio Ambiente (Environment and Childhood) (INMA) Project (Spain) [30,31], and many others such as those we can find in [32].

We follow [33], studying the biparietal diameter observed via ultrasound, to generate data with similar behavior to which we found about fetal growth in the birth cohort of the Genetics, Early Life Environmental Exposures and Infant Development in Andalucía (GENEIDA) Project (Spain) (Web site https://www.easp.es/web/geneida/ accessed on 7 November 2024).

Similar observations could be found when using other databases of fetal measurements. Data with the same characteristics could also emerge when studying other phenomena.

The next steps describe the type of non-homogeneous process selected and how the final simulated sample data are obtained.

4.1. The Simulated Process

We simulate data for a Gompertz-lognormal process whose specific expression is available in Section 2.2 with the $h_\theta(t)$ function displayed in Table 1. This specific process includes Gompertz-type diffusion when $c = 0$ and homogeneous lognormal diffusion when $m = 0$, as well as paths changing their curvature, turning from increasing to decreasing or turning from concave to convex shapes (see [21,34]), as can be seen in Figure 1. This fact makes the Gompertz-lognormal diffusion process a very useful tool for modeling a wide range of sample paths, including exponential and sigmoidal growth.

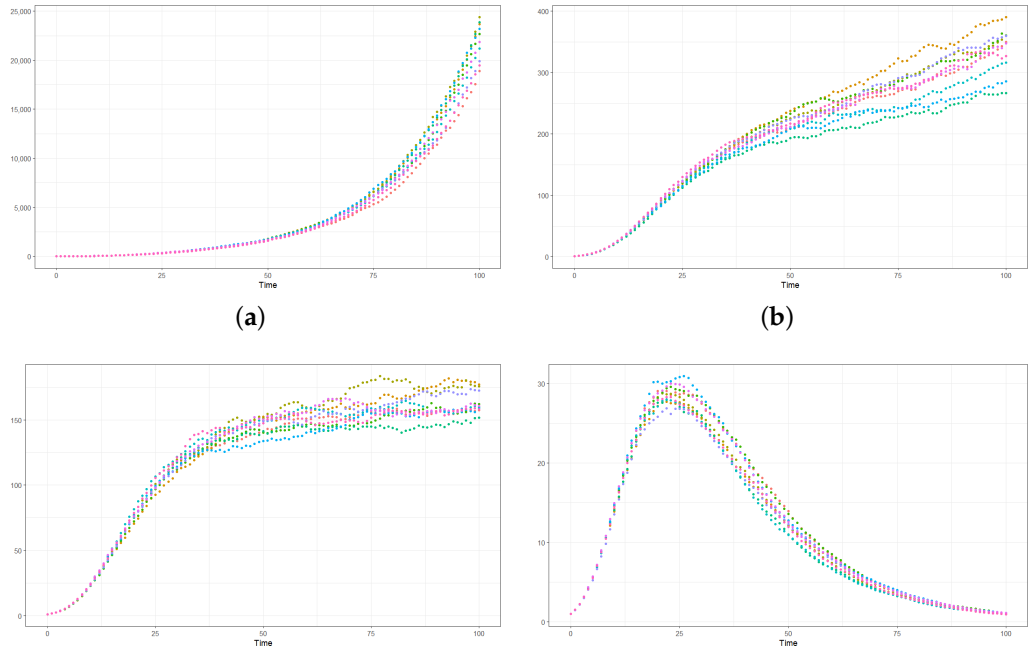


Figure 1. Simulated paths for Gompertz-lognormal diffusion process. (**a**) Exponential shape ($c = 0.05$); (**b**) mixed shape ($c = 0.008$); (**c**) Gompertz shape ($c = 0.001$); (**d**) mixed shape ($c = -0.05$).

Figure 1 displays 100 values of 10 simulated paths of a Gompertz-lognormal process from $X(t_0) = 0.1$ with $t_i - t_{i-1} = 1\ \forall i = 1, \ldots, 100$. Figure 1a shows simulated sample paths with an exponential trend. Figure 1b shows sample paths with a mixture shape, turning from convex to concave shape. Figure 1c shows sample paths with a sigmoidal shape. Figure 1d shows the sample paths with a shape that changes from convex to concave and convex again, turning from increasing to decreasing development. All paths are simulated with $m = 0.5$, $\beta = 0.1$, and $\sigma = 0.01$. The difference between them is the value of parameter c, which changes from 0.05 to -0.05. This wide range of possibilities in the behavior of the process makes it a very useful tool for modeling different kinds of data.

4.2. The Data

The dataset is generated based on the concepts proposed in [35], where a generic simulation technique was described. The observations simulated with this procedure are all obtained in equispaced times of constant and equal jump in each sample path. However, this is not realistic, since in practice, the observations of the phenomena are often not observed at the same time instants, nor are these times equispaced. Specifically, in fetal growth data, it is usual that the observations made on each individual are taken at different time instants, although they are usually around certain values of interest. Also, the number of observations varies in each case, and the observations are not equispaced. To make the simulation study as realistic as possible, we do not work with the simulated observations directly. To create our final sample, we chose data from the original simulated dataset in the following way:

1. The starting point is a matrix with N sample paths, each one containing n data simulated in equispaced times $t_1, t_2, \ldots, t_n$ with $t_i - t_{i-1} = r \; \forall i = 2, \ldots, n$ that we obtain following the software propose in [36].
2. We fix a *mean size* $\lambda \in \mathbb{R}^+$, with λ being the expected number of observations for the final sample paths.
3. We choose p times, called *points of interest*, around which data will be available in the final sample. These points of interest are IP_i with $i = 1, \ldots, p$. For $p = 0$, this step is skipped.
4. For each sample path i, we generate $m_i \in \mathbb{N}$ a random number $m_i \sim Poisson(\lambda)$ with $i = 1, \ldots, n$, enforcing $2 \leq m_i \leq n$, which will be the total available observations for the i-th sample path in the final sample.
5. Each m_i is divided between the number of points of interest, p, to fix how many data will be available around each point of interest for each sample path, obtaining $q_{i,j}$. For $p = 0$, this step is skipped.
6. We randomly choose for the sample path i in the point of interest j a total of $q_{i,j}$ values of the times from the original matrix following a normal distribution with mean IP_j and common chosen variance. We repeat this for all sample paths $i = 1, \ldots, n$ and for all points of interest $j = 1, \ldots, p$. For $p = 0$, we select m_i values via a discrete uniform distribution between t_1 and t_n.
7. The selected times lead us to their corresponding simulated values.
8. The selected times with the selected data are the final simulated sample, where there is no regularity in times nor in the amount of data for different sample paths.

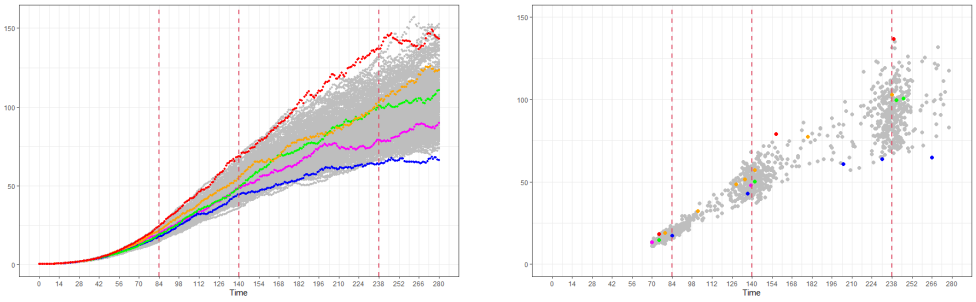
With the objective of simulating 250 sample paths, similar to fetal growth, taking into account the results of [33] regarding biparietal diameter in fetuses, we consider the parameters $m = 0.11$, $\beta = 0.018$, $c = 0.002$, and $\sigma = 0.01$ to simulate the paths. We also consider $X(t_0) = 0.14$ as a possible length of an ovum before it is fertilized [37]. Sample paths begin at $(t_0 = 0, X(t_0) = 0.14)$ and are simulated from t_0 with $t_i - t_{i-1} = 1$, $i = 1, \ldots, 280$, observing the expected duration of a normal pregnancy, as the values of t represent each day of pregnancy.

Our points of interest are $IP_1 = 84$, $IP_2 = 140$, and $IP_3 = 238$, which correspond to the times of interest on the first, second, and third quarters in pregnant women. Then, we generate a number of observations for each case with a Poisson distribution with parameter $\lambda = 4$, starting at $t = 70$ because before this time, it is not easy to take measures via ultrasound. Finally, as variability of the normal distribution, we take a variance of 7 because ultrasound measures are usually scheduled in the same week of the time of interests.

The R code (R software version 4.3.3) for the simulation procedure is available in Appendix B.

In Figure 2, we can see the simulated data. Figure 2a shows the complete paths generated from $t_0 = 0$ to $t_n = 280$ for $N = 250$ individuals. We highlight some random paths in different colors to better observe the behavior of the sample paths and understand the selection of values. The selected final data are depicted in Figure 2b, where we discard

all information that is not easily available in real-world cases. The colored paths in Figure 2a can be seen as colored points in Figure 2b. We also mark the three times of interest ,IP_1, IP_2, and IP_3, with vertical discontinuous red lines.



(**a**) (**b**)

Figure 2. Simulated data. Figure (**a**), on the left, shows a total of 250 simulated sample paths of the Gompertz-lognormal process with $m = 0.11$, $\beta = 0.018$, $c = 0.002$, and $\sigma = 0.01$. Figure (**b**), on the right, shows selected sample of the simulated paths with a mean size of 4 and 3 points of interest around which the samples are taken with a deviation of 7 and from $t = 70$. These data are similar to those for fetal growth measures.

We provide the final simulated dataset represented in Figure 2b, which is available as open data at https://doi.org/10.5281/zenodo.13929734.

4.3. Estimation and Inference

Once we obtain the final dataset, we estimate the parameters of the process via maximum likelihood estimation method using the maxLik package in R [38] with all initial values equal to 0.01. We summarize in Table 2 the values of the parameters and their maximum likelihood estimation.

Table 2. Parameters of the model: values for the simulation and their maximum likelihood estimation.

Parameter	Simulation Value	ML Estimation
m	0.11	0.1095
β	0.018	0.0178
c	0.002	0.0017
σ	0.01	0.0100

With the estimated parameters, we can impute the values for times of interest IP_1, IP_2, and IP_3 using the mean, mode, and median functions, each in the three following cases: (a) conditioning on the previous data, (b) conditioning on the next following data, and (c) conditioning on the nearest available data. In cases (a) and (b), if such necessary observation for the conditioned function does not exist, the imputation is a missing value. We do not take $X(t_0) = 0.14$ as an available observation in any case because it is an arbitrary number rather than a real observation. Later on, we will study how this value affects the imputations.

We consider the Root Mean Square Error ($RMSE$) and the Mean Absolute Error (MAE). Following [39], in this case, MAE is more reliable than $RSME$ due to the leptokurtic distribution of the residuals $\epsilon_{i,j} = x_i(IP_j) - \hat{x}_i(IP_j)$, where $x_i(IP_j)$ is the simulated value for the i-th path at the j-th interest point, and $\hat{x}_i(IP_j)$ the imputed value at the IP_j time.

Table 3 displays the errors in the inference of the three interesting points, considering ($IP_1 = 84$, $IP_2 = 140$, $IP_3 = 238$) as follows:

$$RSME(IP_j) = \sqrt{\frac{1}{N_j}\sum_{i=1}^{N_j}\epsilon_{i,j}^2} \quad \text{and} \quad MAE(IP_j) = \frac{1}{N_j}\sum_{i=1}^{N_j}|\epsilon_{i,j}|$$

where N_j is the number of final imputations and IP_j is the j-th point of interest ($j = 1, \ldots, p$).

Table 4 shows the joint error for the simulated data. Naming $\widehat{N} = \sum_{j=1}^{p} N_j$, we define the joint error measurements as follows:

$$RSME = \sqrt{\frac{1}{\widehat{N}} \sum_{j=1}^{p} \sum_{i=1}^{N_j} \epsilon_{i,j}^2} \quad \text{and} \quad MAE = \frac{1}{\widehat{N}} \sum_{j=1}^{p} \sum_{i=1}^{N_j} |\epsilon_{i,j}|.$$

Tables 3 and 4 also show the number of missing imputations we obtain (NA values) in cases where unavailable data do not allow for the inference procedure. This is the case when one uses the function conditioned to the past or when one uses functions conditioned to a future observation but such observed data do not exist. In order to compare error measurements at the same imputation points, although not all functions have unavailable imputation points, we consider both cases: all possible imputations (NA = 0) and imputations at the points at which all functions can make the forecast (NA > 0).

Table 3. Obtained errors at times of interest IP_1, IP_2, and IP_3: $RSME$ and MAE, number of values that cannot be imputed (NAs) in the inference procedure using mean, mode, and median functions conditioned on the closest past, closest future, and closest observed values.

Error	Time Conditioned on	Mean Function		Mode Function		Median Function	
Imputation at first time of interest $IP_1 = 84$							
		0 NAs	101 NAs	0 NAs	101 NAs	0 NAs	101 NAs
$RSME(IP_1)$	Past	$*$	0.4926	$*$	0.4925	$*$	0.4925
	Future	0.7968	0.8856	0.8021	0.8925	0.7980	0.8872
	Closest	0.5447	0.4657 †	0.5448	0.4641†	0.5446	0.4651 †
$MAE(IP_1)$	Past	$*$	0.2903	$*$	0.2907	$*$	0.2903
	Future	0.5173	0.6021	0.5236	0.6113	0.5190	0.6044
	Closest	0.3123	0.2582 †	0.3130	0.2579 †	0.3125	0.2580 †
Imputation at second time of interest $IP_2 = 140$							
		0 NAs	11 NAs	0 NAs	11 NAs	0 NAs	11 NAs
$RSME(IP_2)$	Past	$*$	2.8285	$*$	2.8368	$*$	2.8288
	Future	$*$	2.6682	$*$	2.7137	$*$	2.6809
	Closest	1.8364	1.8632 †	1.8352	1.8624 †	1.8354	1.8623 †
$MAE(IP_2)$	Past	$*$	1.8817	$*$	1.8739	$*$	1.8763
	Future	$*$	1.6885	$*$	1.7085	$*$	1.6923
	Closest	1.0626	1.0737 †	1.0663	1.0783 †	1.0636	1.0750 †
Imputation at third time of interest $IP_3 = 238$							
		0 NAs	87 NAs	0 NAs	87 NAs	0 NAs	87 NAs
$RSME(IP_3)$	Past	6.1756	6.5106	6.2065	6.5132	6.1741	6.4963
	Future	$*$	3.7365	$*$	3.7184	$*$	3.7300
	Closest	4.2479	3.4014 †	4.2900	3.4011 †	4.2599	3.4011 †
$MAE(IP_3)$	Past	3.6568	4.2782	3.6737	4.2593	3.6584	4.2681
	Future	$*$	1.7851	$*$	1.7777	$*$	1.7826
	Closest	1.8240	1.4672 †	1.8568	1.4727 †	1.8332	1.4688 †

$*$ The error cannot be obtained because the data do not allow all imputations with the conditioned characteristic function. † Smallest error obtained when comparing the three analyzed methods.

Table 4. Obtained errors for all time of interest jointly: $RSME$ and MAE, number of points at which values cannot be imputed (NAs) in the inference procedure using mean, mode, and median functions conditioned on the closest past, closest future, and closest observed values.

Error	Time Conditioned on	NAs	Mean Function	Mode Function	Median Function
$RSME$	past	108	4.2292	4.2489	4.2285
	future	91	2.5091	2.5207	2.5112
	closest †	0	2.6903	2.7123	2.6964
MAE	past	108	2.1935	2.1971	2.1920
	future	91	1.2623	1.2699	1.2636
	closest †	0	1.0663 ††	1.0787 ††	1.0698 ††

† The only method that allows for all desired imputations. †† Smallest MAE obtained when comparing the three analyzed methods.

5. Discussion

5.1. Imputation at First Point of Interest $IP_1 = 84$

When analyzing the information in Table 3 regarding the inference at the first time of interest, we observe that there are two columns for each characteristic function estimating the values at $IP_1 = 84$. These columns summarize the errors for two distinct subsets of individuals: the first, labeled NA = 0, represents cases where all target points can be estimated. The asterisk $*$ indicates that for IP_1, it is not possible to infer all points using the process conditioned solely on past values, as in some instances, past data are unavailable. Consequently, this approach results in 101 missing imputations (NAs) out of the 250 required, amounting to 40.4% of the target data, which renders the procedure unsuitable.

In contrast, when using the process conditioned on future values, all estimations are feasible. Similarly, switching dynamically between the processes conditioned on past and future values—depending on which data are closer to the time point being imputed—also enables complete estimation.

Under NA = 101, we compare the errors for each data input method within the common subset of 149 imputations. Examining these columns and comparing the $RMSE$ and MAE across all functions (mean, mode, and median), we find that the smallest errors occur when the process is conditioned on the nearest available value.

When focusing solely on the two procedures that allow for inference for all individuals, conditioning on the nearest available data consistently outperforms conditioning exclusively on future available data.

For the first point of interest, previously observed data are often unavailable. In such cases, the most effective approach is to impute data by conditioning on the closest available value.

5.2. Imputation at Second Point of Interest $IP_2 = 140$

In this case, we have two columns corresponding to the values of NA: 0 and 11. It is evident that the number of NA values is not significant, with NA = 7 for the process conditioned on the past and NA = 4 for the process conditioned on the future. These represent 2.8% and 1.6% of the data, respectively, indicating that the loss is minimal when using these conditioned methods.

Once again, when we use the process conditioned on the closest available value, we encounter no NA values, as there are always at least two observations for each individual. Furthermore, by comparing the $RMSE$ and MAE for the three characteristic functions at IP_2, we find that this procedure not only avoids generating NA values in the imputation but also results in smaller $RMSE$ and MAE values, indicating that it outperforms both the methods conditioned exclusively on the past or the future.

5.3. Imputation at Third Point of Interest $IP_3 = 238$

When examining the inference for IP_3 in Table 3, because of the allocation of this time, we obtain 87 missing imputations using the process conditioned to the future. This indicates that no further information is available after IP_3 for 87 individuals. Nevertheless, the other two methods produce 0 NA values; however, again, the smallest error measurements are for the process conditioned to the nearest available data.

These observations indicate that the functions conditioned on the closest value should be selected as the best imputation method.

5.4. Comparison Between Characteristic Functions

Moreover, by comparing the three conditional functions—mean, mode, and median—we observe that there is no clear advantage for any of them; at IP_1, we see that the median function is slightly better in $RSME$ conditioning on the closest value for NA = 0 ($RSME$ = 0.5446 versus 0.5447 and 0.5448), but the mean function produces a smaller MAE (0.3123 versus 0.3130 and 0.3125). At this point of interest, the mode function is slightly better when NA = 101 ($RSME$ = 0.4641 versus 0.4657 and 0.4651, MAE = 0.2579 versus 0.2582 and 0.2580), and the mean function outperforms the other functions when using the process conditioned to the future with NA = 0 ($RSME$ = 0.7968 versus 0.8021 and 0.7980, MAE = 0.5173 versus 0.5236 and 0.5190) and with NA = 101 ($RSME$ = 0.7968 versus 0.8021 and 0.7980, MAE = 0.5173 versus 0.5236 and 0.5190).

At the second time of interest IP_2, we notice that the mean function has a smaller $RMSE$ when we use the process conditioned to the past or the future; however, the median function presents a smaller $RSME$ for the function conditioned to the nearest available value, all of them considering NA = 11. When NA = 0, the mode function produces a smaller value ($RSME$ = 1.8352 versus 1.8364 and 1.8354).

This is not consistent with the MAE values, which show the smallest value when using the mode function for the process conditioned on past values (MAE = 1.8739 versus 1.8817 and 1.8763) and for the mean function for the process conditioned on future values, (MAE = 1.6885 versus 1.7085 and 1.6923). If we select the process conditioned on the nearest value, we obtain less MAE if we use the mean function, both considering NA = 0 and NA = 11.

At the third point of interest, we observe a similar situation. The mean function has a smaller $RSME$ for the process conditioned on the future and NA = 87, the mode function produces a smaller $RSME$ for the process conditioned on the future or on the closest available data when $N = 87$, and the median function produces a smaller $RSME$ for the process conditioned on the nearest value and NA = 87 together with the mode function. But looking at the MAE values, we will choose the mean function for imputation with the process conditioned on the nearest value and the mode function if we use the process conditioned on the past or on the future.

Contrary to what happens when we select a method, choosing one function or another does not produce clearly better estimates. When considering the MAE measure and the nearest neighbor method, the mean function seems to be slightly better than the median or mode.

5.5. Joint Error Measurements

Table 4 shows the $RSME$ and MAE for all points of interest using the process exclusively conditioned on the past, exclusively conditioned on the future, and conditioned on the past or the future depending on which observation is closer to the imputation moment. In the first scenario, using the process conditioned to the past, we cannot impute 108 observations (101 on the IP_1 and 7 on IP_2), and we obtain the worst $RSME$ and MAE values for the three functions: mean, mode, and median. For the second approach, using the process conditioned to the future, we have 91 missing imputations and a slightly smaller $RSME$ than the third approach, but a greater MAE. The third method is the only one that allows us to impute all the data, and moreover, it provides the best values for MAE. This measure

is more reliable than $RSME$ because of the leptokurtic distributions of the errors; thus, we clearly find that this third methodology is the best among the three we studied.

For Table 4, we again cannot propose the mean, mode, or median function as better than the others because the functions return very similar values for $RSME$ and MAE and because we do not always find the minimum error value for the same function. However, if we pay attention to characteristic functions conditioned on the nearest observed value, where NA=0, these cases show lower $RSME$ and MAE values for the conditional mean function rather than the mode or median.

5.6. Parameter Choice

The final sample was very similar to the real data of a birth cohort in which the biparietal diameter is observed via ultrasound (see [33]), carefully selecting the parameters that make the dataset appear realistic. We discuss below how this selection might affect the final characteristics of the dataset and the error measures in the imputation procedure.

5.6.1. m, β, and c Parameters of the Infinitesimal Mean of the Process

Changing the values of these three parameters will only affect the shape of the paths, as we can see in Figure 1. It has no effect on the estimation procedure or on the error measurements.

5.6.2. Initial Value $X(t_0)$

Since we have taken an arbitrary value $X(t_0) = 0.14$ and it may be unknown, we carry out the estimation and imputation process with other possible values of $X(t_0) = x_0$. We observe the results when x_0 takes different values between 0.06 and 0.17 following [37], and we observe that the results are independent of the initial value x_0 taken, as long as it is in a plausible range. These results can be seen summarized in Table A2 of Appendix C.

5.6.3. σ Parameter of the Infinitesimal Variance of the Process

The value of σ is related to the noise in the process. High values of σ result in paths with peaks that do not clearly show the trajectory of the process. Low values of σ result in smooth paths whose behaviors can be more clearly susceptible to modeling.

If this value is increased, we will obtain large error measures, due to the fact that the model fits worse than if the value of σ is kept low.

Common values for this parameter are around 0.01 [22]; enough to represent randomness but not so high as to hide the true shape of the paths.

5.6.4. Number and Location of Points of Interest

Three points of interest have been selected to be located in three different segments within the observation period. Thus, the first point of interest, located at the beginning of the observation period, is characterized by a large number of paths that have not been observed before, and it happens when the forward condition is most useful. The second point of interest is located in an intermediate zone, where it is common to find observations before and after the one to be imputed, although this is not always the case. The third point of interest has been placed at the end of the observation period, so that a significant number of subjects do not present observations subsequent to the one to be imputed and therefore cannot be conditioned to the future.

If more or fewer points of interest are taken, they will have similar characteristics to some of the three considered, so that they will not present significant differences with respect to the obtained errors.

5.6.5. Poisson Distribution and λ Parameter

Using the Poisson model in the simulation of the amount of data for each path makes sense given the nature of the values to be obtained. It gives discrete values with no upper bound, with a unimodal probability distribution. Using another distribution or a constant

quantity would only impact the amount of data for the paths in the final dataset. We have selected an extreme case, where the amount of data is low, which fits the actual data that studies have reported in the birth cohorts. No other distribution (or constant value) can result in a dataset with paths that have more missing data, since with these values with many trajectories already have the minimum number of data points, which is 2.

To investigate whether the amount of observations on the paths influences the results, we analyzed different types of paths (with more or fewer values) in the final dataset. The obtained MAE values, summarized using the mean function, are presented in Table A3 in Appendix D.

Firstly, we point out that to compare the three methods, it is necessary to have at least one observation prior to the first point of interest and another after the last. This requirement means that most of the simulated paths cannot be considered in this comparison. For instance, among the 34 trajectories with only $n = 2$ observations, only 3 of them (with 31 having missing values) could be used to determine the comparative error between the methods. This represents less than 10% of the observed trajectories in this case, even though the method of conditioning on the nearest available value is capable of performing all imputations. It is evident that this situation improves as the number of observed data points increases. For $n = 3$, the percentage rises to nearly 20%, reaching over 50% when the number of observed data points is at least five.

For IP_1, it is observed that when n is small, the minimum error is achieved equally by the method conditioned on past values and the method conditioned on the nearest value. This result is logical, as both methods rely on the first observation of the trajectory to predict that point. As the number of observations increases, providing more values that may be closer to IP_1 but observed at a later time, the method conditioned on the nearest value yields the lowest errors.

For IP_3, a similar pattern emerges, but with the methods conditioned on future values and the nearest value. Specifically, their errors are equivalent when the number of observations is small, but the method conditioned on the nearest value achieves lower errors as the number of observations increases.

In the case of IP_2, being an intermediate point in the trajectory, the error is almost always lower for the method conditioned on the nearest value, regardless of the number of paths.

In summary, the prediction method based on the nearest observed value consistently demonstrates errors that are lower than or equal to those of the classical method conditioned on past values and the method conditioned on future values, regardless of the missing values in the trajectory.

5.6.6. Variance of the Normal Distribution

For each point of interest, we choose observations for the final sample using a normal distribution with a mean at the point of interest itself and a standard deviation of 7.

Moving the value of this deviation will only affect how far apart the observations are and may cause the error measures to increase or decrease, but it will not change how much smaller they are relative to each other, nor will it change the fact that with the classical methodology, many imputations are not possible.

5.7. Applicability

The applicability of the proposed method for data imputation is a key aspect of its potential impact, since there is no single universal solution to the missing data problem on real-world datasets [40]. This method is designed to handle datasets with varying characteristics, such as irregular time intervals or missing values, making it suitable for a wide range of real-world applications. By considering both past and future observations, it offers a more comprehensive approach compared to traditional methods. However, its effectiveness depends on the specific nature of the data, and further validation across different datasets is necessary to fully assess its general applicability.

5.8. Conclusions Derived from Simulation Study

In short, we obtain better inference when using the process conditioned to the closest observed value for two reasons: all points can be estimated, contrary to the classical method, and the difference between the real and imputed data is smaller, on average, than those obtained with the classical procedure or only conditioning to future observations. Moreover, the conditional mean function seems to be the characteristic function that provides lower a MAE error with this method, although conditional mode and median functions both give very similar error measurements, close to those given by the conditional mean function.

6. Conclusions

As demonstrated, non-homogeneous lognormal diffusion processes are useful for modeling real situations. This leads us to develop tools that allow us to handle this type of information similarly to what is provided in real situations. Simulation is necessary and useful; however, it typically provides a nonrealistic dataset, where we do not find missing data or irregularities. Therefore, classical inference, which is tested with classical simulated data, employs functions conditioned on the past, considering the future as unknown but capable of being predicted. Classical inference forecasts future values for complete and regular datasets.

However, for many opportunities, we should use the available information of future times. Real-life data are frequently non-systematic and contain many missing values and different sample sizes for individuals. We focus on such situations and propose changing the target from forecasting to imputation. We consider not only the data available for past times but also available data; past and future observations.

Conditioning on future time should be useful in a wide range of situations. For example, the population size may be known at specific time points following the initial time t_0, and it may be desirable to approximate the initial state X_0. In such cases, it can be useful to study the process $X(s)$ conditioned on the future $X(t) = x$ with $t > s \geq t_0$.

We formally obtain the distribution of $X(s)|X(t)$ with $t > s \geq t_0$. With the obtained distribution, we can use the information of future observations for imputation. We also derive the expressions for the characteristic functions of mean, mode, and $alpha$-quantile, which are useful for the imputation procedure.

If we combine both ideas, conditioning on the past and conditioning on the future, we can achieve a method to impute data that improves the classical method in terms of the number of possible imputations and in terms of $RSME$ and MAE. Thereby, conditioning on the nearest available value improves the imputation.

We present the procedure with a simulated dataset that includes the characteristics that we could observe in a birth cohort, where the observations are not at the same time nor in equal quantities. For the simulated data, we determine, as we expected, that the non-homogeneous lognormal process conditioned on the nearest value works better than if the conditioning is only on past values or only on future values.

Author Contributions: Conceptualization, methodology, software, validation, investigation, writing—original draft preparation, writing—review and editing, A.G.-B., P.P., D.R.-M., and N.R.-C.; formal analysis and visualization, A.G.-B. and P.P.; resources, data curation, D.R.-M.; supervision and funding acquisition, D.R.-M. and N.R.-C.; project administration, D.R.-M. and P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out within the framework of the research project PID2021-128261NB-I00 (PROESTEAM), financed by MICIN/AEI/10.13039/501100011033 and by ERDF, EU and Junta de Andalucía grant number FQM-147. This work was also partially supported by "European Union – Next Generation EU" through MUR-PRIN 2022, project 2022XZSAFN "Anomalous Phenomena on Regular and Irregular Domains: Approximating Complexity for the Applied Sciences", and MUR-PRIN 2022 PNRR, project P2022XSF5H "Stochastic Models in Biomathematics and Applications".

Data Availability Statement: The data presented in this study are openly available as open data at https://doi.org/10.5281/zenodo.13929734.

Acknowledgments: The authors thank the Department of Mathematics of the University of Salerno for hosting and supporting A. García-Burgos's stay. The authors also greatly appreciate Y. Román's advice and observations on different aspects of R programming. P. Paraggio is member of the group GNCS of INdAM (Istituto Nazionale di Alta Matematica), and thanks the Departamento de Estadística e IO at the Universidad de Granada for hosting her during her research stay in March/April 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Summary of Key Symbols and Parameters

Table A1. Summary of symbols and notations and their corresponding meanings.

Symbol	Meaning	Reference
μ_0, σ_0^2	Parameters of the initial lognormal distribution.	Equation (2)
$X(t)$	The lognormal diffusion process evaluated at time t.	Equation (3)
σ^2	Parameter related to the infinitesimal variance of $X(t)$.	Equation (3)
$H_{\tilde{\xi}}(t_0, s)$	Integral function related to the drift of the lognormal diffusion process.	Equation (4)
$\tilde{\xi}$	Vector of the parameters of $X(t)$.	Equation (5)
Λ_n	The n-dimensional lognormal distribution.	Equation (6)
$m_{\tilde{\xi},\mu_0,\sigma_0^2}(s \mid y, t)$	First parameter of the conditional distribution of $X(s) \mid X(t) = y$. Depends on the parameters $\tilde{\xi}, \mu_0, \sigma_0^2$ and on the values y, s and t.	Equation (8)
$s_{\sigma_0^2,\sigma^2}(s \mid t)$	Second parameter of the conditional distribution of $X(s) \mid X(t) = y$. Depends on the parameters σ_0^2, σ^2 and on the values s and t.	Equation (9)

Appendix B. R Code for Simulation Procedure

```
library(Sim.DiffProc)

#Code of the GeneralLognormalSimulation function
GeneralLognormalSimulation<-function(N_Sp,n,r,t0,x0,s,method = c(''TransfW
    '',''sde'')){method<-match.arg(method)
process<-PROCESS()
h<-switch(process, An=HFunction(), GT=''m*exp(-beta*t)'', MGL=''m*exp(-
    beta*t)+c'', Be=''b*c*k/(exp(k*t)-c)'', LT=''b*c/(b+exp(c*t))'', T=''b
    *c*q/(b+exp(c*t))'')
env<-switch(process, An=ENV(), GT=ENVGompertzType(),MGL=
    ENVMixGompertzLognormal(), Be=ENVBertalanffy(), LT=ENVLogisticType(),
    RT=ENVRichardsType())
exprh <- as.expression(eval(substitute(substitute(e, env),list(e = parse(
    text = h)[[1]]))))
if (length(x0)>1)
    Initial<-rlnorm(N_Sp,x0[1],x0[2])
else
    Initial<-rep(x0, N_Sp)
if (method==''TransfW''){
    Win<-WienerSimulation(N_Sp,n,r)
    h.t <- function(t) NULL
    body(h.t)<-parse(text=exprh)
    NHLog1<-array(0,c(n+1,N_Sp))
    Time<-seq(t0,length=n+1,by=r)
    AA<-sapply(Time, function(u,l,h) integrate(h,l,u)$value, h=h.t, l=t0)
        -(s^2)*(Time-t0)
    for(i in 1:N_Sp) {
        NHLog1[,i]<-Initial[i]*exp(t(AA)+s*Win[,i])
    }
    NHLog<- cbind(Time,NHLog1)
}
```

```
else {
    exprd<-parse(text=paste(exprh, ``*'', expression(x)))
    exprs<-as.expression(eval(substitute(expression(a * x), list(a = s))))
    NHLog<- snssde1d(N=n,M=N_Sp,x0=Initial,t0=t0,Dt=r,drif=exprd,
        diffusion=exprs)$X
}
invisible(NHLog)}

#Code of the functions used by GeneralLognormalSimulation function to
    simulate sample paths of a standard Wiener process
WienerSimulation<-function(N_Sp,n,r){
    Wiener<-rbind(rep(0,N_Sp), apply(array(rnorm(N_Sp*n,0,sqrt(r)),dim=c(n,
        N_Sp)),2,cumsum))
    invisible(Wiener)
}

#To request the name of the particular diffusion process to be simulated
PROCESS <- function(){
    vector<-c(``GT'',``MGL'',``Be'',``LT'',``RT'',``An'')
    p <- readline(``Process to be simulated (you can choose between GT (
        GompertzType), MGL (Mix Gompertz Lognormal), Be (Bertalanffy), LT (
        LogisticType), RT (RichardsType) or An (Another)) = '')
    while(!is.element(p,vector))
    p <- readline(``The name entered is incorrect. Process to be simulated
         (you can choose between GT (GompertzType), MGL (Mix Gompertz
        Lognormal), Be (Bertalanffy), LT (LogisticType), RT (RichardsType)
        or An (Another)) = '')
    p
}

#To request the values of parameters of a Mix Gompertz Lognormal diffusion
     process
ENVMixGompertzLognormal<-function(){
    Value_m <- readline(``Value of m? '')
    m1<-as.numeric(Value_m)
    Value_beta<- readline(``Value of beta? '')
    beta1<-as.numeric(Value_beta)
    Value_c<- readline(``Value of c? '')
    c1<-as.numeric(Value_c)
    E = list(m=m1, beta=beta1,c=c1)
}

SimulatedSampled<-GeneralLognormalSimulation(N_Sp=250,n=280,r=1,t0=0,x0
    =0.14,s=0.01,method=``TransfW'')
MGL
0.11
0.018
0.002

#Code for selecting random samples of simulated sample paths.
RandomSelectionSamples<-function(data,MeanSize=(nrow(data)-1),
    InterestPoints=c(),variation=1,minimum=1){
    N=ncol(data)-1;#Number of simulated paths
    n=nrow(data)-1; #Number of points in each path
```

```
if(minimum<1|minimum>n){
    stop(``The parameter minimum, which indicates from which time
        instant to sample, must be between 1 and the number of the
        simulated paths data '',n)
}
p=length(InterestPoints);#Number of points of interest. If it is empty
     is because p=0 and there are no points of interest.

#Code to random selection of the sample size between 2 and n with mean
     parameter indicated in MeanSize
m<-rpois(N,MeanSize);
BADmi<-which((m<2)|(m>n))
while(length(BADmi)>0){
    m<-replace(m,BADmi,rpois(1,MeanSize))
    BADmi<-which((m<2)|(m>n))
}
m<-matrix(m,nrow=1,ncol=N)
Maxm<-max(m);

#Code to select observation times of the samples of the simulated
    paths
if((n-minimum)<Maxm){
    stop(``The number of observation times considered is insufficient
        for the required sampling.'')
}
if(p==0){ #if there is not points of interest we select for each
    sample path mi different values for t between the minimun and n-1
    time<-t(apply(m,2,function(m){
        time<-sample(minimum:n,m)
        time<-sort(time)
        while(length(time)<Maxm){
            time<-c(time,NA)
        }
        return(time)
    }))
}
else{ #if there exists points of interest we generate q=mi/p around
    each one until we reach mi values. These are values between the
    minimum and n-1.
    if(min(InterestPoints)<minimum){
        stop(``At least one of the indicated points of interest is
            below the specified minimum value '',minimum)
    }
    if(max(InterestPoints)>n){
        stop(``At least one of the indicated points of interest is
            above the maximum value observed on the simulated path '',n)

    }
    time<-t(apply(m,2,function(m){
        q=1;
        if((m/p)>1){
            q=trunc(m/p);
        }
        time<-c();
```

```
            InterestPoints<-matrix(InterestPoints,nrow=1,ncol=p)
            time<-apply(InterestPoints,2,function(IP){
                obs<-round(rnorm(q,IP,variation))
                BADobs<-which((obs<minimum)|(obs>n))
                while(length(BADobs)>0){
                    obs<-replace(obs,BADobs,round(rnorm(1,IP,variation)))
                    BADobs<-which((obs<minimum)|(obs>n))
                }
                return(obs)
            })
            time<-as.vector(time)
            time<-sort(time)
            time<-unique(time)
            while(length(time)>m){
                A<-sample(1:length(time),1); time<-time[-c(A)]
            }
            while(length(time)<m){
                obs<-sample(minimum:n,1)
                time<-c(time,obs)
                time<-sort(time)
                time<-unique(time)
            }
            while(length(time)<Maxm){
            time<-c(time,NA)
            }
        return(time)
        }))
    }

    #Code to select the samples of the simulated paths
    samplesselected<-apply(time,2,function(t){
        data[t+1,2]
    })

    #Code to export information
    time_samples<-rbind(time,samplesselected)
    time_samples<-matrix(as.vector(time_samples), nrow=nrow(time), ncol=2*
        ncol(time))
    def_time_samples<-matrix(c(1:N),nrow=N,ncol=1)
    def_time_samples<-cbind(def_time_samples,t(m))
    def_time_samples<-cbind(def_time_samples,time_samples)

    return(def_time_samples)
}

y<-RandomSelectionSamples(SimulatedSampled, MeanSize=4,InterestPoints=c
    (84,140,238),variation=7,minimum=70)
write.csv(y,''RandomSelectionSamples.csv'')
```

Appendix C. $RSME$ and MAE Obtained in the Imputation Procedure with the Mean Function Using Different Values for $X(t_0)$

Table A2. Obtained errors at times of interest IP_1, IP_2, and IP_3: $RSME$ and MAE, number of values that cannot be imputed (NAs) in the inference procedure, taking different values of $X(t_0) = x_0$ and using the mean function conditioned on the past, future, and closest observed value.

Error	Time Conditioned on	$x_0 = 0.06$	$x_0 = 0.08$	$x_0 = 0.1$	$x_0 = 0.12$	$x_0 = 0.16$	$x_0 = 0.17$
Imputation at first time of interest $IP_1 = 84$ with 101 NAs							
$RSME(IP_1)$	Past	0.5150	0.5032	0.4966	0.4935	0.4932	0.4940
	Future	0.8892	0.8877	0.8867	0.8860	0.8853	0.8852
	Closest	0.4963	0.4837	0.4753	0.4695	0.4631	0.4623
$MAE(IP_1)$	Past	0.3067	0.2995	0.2948	0.2916	0.2915	0.2922
	Future	0.6067	0.6050	0.6036	0.6027	0.6017	0.6016
	Closest	0.2796	0.2710	0.2652	0.2610	0.2575	0.2572
Imputation at second time of interest $IP_2 = 140$ with 11 NAs							
$RSME(IP_2)$	Past	2.8352	2.8307	2.8287	2.8282	2.8293	2.8299
	Future	2.6555	2.6577	2.6608	2.6642	2.6723	2.6745
	Closest	1.8737	1.8696	1.8667	1.8647	1.8621	1.8617
$MAE(IP_2)$	Past	1.8808	1.8799	1.8799	1.8801	1.8833	1.8840
	Future	1.6779	1.6770	1.6826	1.6852	1.6921	1.6940
	Closest	1.0810	1.0777	1.0758	1.0743	1.0739	1.0740
Imputation at third time of interest $IP_3 = 238$ with 87 NAs							
$RSME(IP_3)$	Past	6.5313	6.5211	6.5151	6.5120	6.5104	6.5108
	Future	3.6906	3.7002	3.7114	3.7237	3.7499	3.7567
	Closest	3.4072	3.4040	3.4023	3.4016	3.4018	3.4022
$MAE(IP_3)$	Past	4.2582	4.2597	4.2649	4.2708	4.2857	4.2897
	Future	1.7646	1.7686	1.7743	1.7798	1.7908	1.7942
	Closest	1.4877	1.4806	1.4751	1.4708	1.4646	1.4641

Appendix D. MAE Obtained in the Imputation Procedure with the Mean Function for Different Values of the Number of Observations in the Paths

Table A3. Obtained errors at times of interest IP_1, IP_2, and IP_3: MAE, number of values that cannot be imputed (NAs) in the inference procedure, splitting the dataset by the number of observations in the path, and using the mean function conditioned on the past, future, and closest observed value.

Size of the Path	Number of Paths	NAs	Time Conditioned on	$MAE(IP_1)$	$MAE(IP_2)$	$MAE(IP_3)$
$n = 2$	34	31	Past	0.3225	4.6208	15.4855
			Future	1.5658	1.4219	1.4961
			Closest	0.3225	4.6208	1.4961
$n = 3$	42	34	Past	0.6206	3.2807	8.1100
			Future	1.2605	2.9085	5.3381
			Closest	0.6206	2.4764	5.3381
$n = 4$	55	34	Past	0.3041	1.8889	5.0255
			Future	0.7458	1.3904	1.0748
			Closest	0.2816	0.6485	1.0556
$n \geq 5$	119	50	Past	0.2327	1.6009	3.4024
			Future	0.4704	1.3703	1.5072
			Closest	0.1895	0.7873	1.2792

References

1. Albano, G.; Giorno, V. A stochastic model in tumor growth. *J. Theor. Biol.* **2006**, *242*, 329–336. [CrossRef] [PubMed]
2. Albano, G.; Giorno, V.; Román-Román, P.; Torres-Ruiz, F. Inferring the effect of therapy on tumors showing stochastic Gompertzian growth. *J. Theor. Biol.* **2011**, *276*, 67–77. [CrossRef] [PubMed]
3. Albano, G.; Giorno, V.; Román-Román, P.; Torres-Ruiz, F. On the therapy effect for a stochastic growth Gompertz-type model. *Math. Biosci.* **2012**, *235*, 148–160. [CrossRef]
4. Román-Román, P.; Torres-Ruiz, F. Inferring the effect of therapies on tumor growth by using diffusion processes. *J. Theor. Biol.* **2012**, *314*, 34–56. [CrossRef]
5. Spina, S.; Giorno, V.; Román-Román, P.; Torres-Ruiz, F. A Stochastic Model of Cancer Growth Subject to an Intermittent Treatment with Combined Effects: Reduction in Tumor Size and Rise in Growth Rate. *Bull. Math. Biol.* **2014**, *76*, 2711–2736. [CrossRef]
6. Albano, G.; Giorno, V.; Román-Román, P.; Román-Román, S.; Torres-Ruiz, F. Estimating and determining the effect of a therapy on tumor dynamics by means of a modified Gompertz diffusion process. *J. Theor. Biol.* **2015**, *364*, 206–219. [CrossRef] [PubMed]
7. Román-Román, P.; Román-Román, S.; Serrano-Pérez, J.J.; Torres-Ruiz, F. Modeling tumor growth in the presence of a therapy with an effect on rate growth and variability by means of a modified Gompertz diffusion process. *J. Theor. Biol.* **2016**, *407*, 1–17. [CrossRef]
8. Giorno, V.; Román-Román, P.; Spina, S.; Torres-Ruiz, F. Estimating a non-homogeneous Gompertz process with jumps as model of tumor dynamics. *Comput. Stat. Data Anal.* **2017**, *107*, 18–31. [CrossRef]
9. Albano, G.; Giorno, V.; Román-Román, P.; Román-Román, S.; Serrano-Pérez, J.J.; Torres-Ruiz, F. Inference on an heteroscedastic Gompertz tumor growth model. *Math. Biosci.* **2020**, *328*, 108428. [CrossRef]
10. Capocelli, R.M.; Ricciardi, L.M. A diffusion model for population growth in random environment. *Theor. Popul. Biol.* **1974**, *5*, 28–41. [CrossRef]
11. Capocelli, R.M.; Ricciardi, L.M. Growth with regulation in random environment. *Kybernetik* **1974**, *15*, 147–157. [CrossRef] [PubMed]
12. Gutierrez-Jaimez, R.; Román, P.; Romero, D.; Serrano, J.J.; Torres, F. A new Gompertz-type diffusion process with application to random growth. *Math. Biosci.* **2007**, *208*, 147–165. [CrossRef]
13. Román-Román, P.; Romero, D.; Torres-Ruiz, F. A diffusion process to model generalized von Bertalanffy growth patterns: Fitting to real data. *J. Theor. Biol.* **2010**, *263*, 59–69. [CrossRef]
14. Román-Román, P.; Torres-Ruiz, F. Modelling logistic growth by a new diffusion process: Application to biological systems. *Biosystems* **2012**, *110*, 9–21. [CrossRef]
15. Ghosh, H.; Prajneshu. Gompertz growth model in random environment with time-dependent diffusion. *J. Stat. Theory Pract.* **2017**, *11*, 746–758. [CrossRef]
16. Barrera, A.; Román-Román, P.; Serrano-Pérez, J.J.; Torres-Ruiz, F. Two Multi-Sigmoidal Diffusion Models for the Study of the Evolution of the COVID-19 Pandemic. *Mathematics* **2021**, *9*, 2409. [CrossRef]
17. Di Crescenzo, A.; Paraggio, P.; Spina, S. Stochastic Growth Models for the Spreading of Fake News. *Mathematics* **2023**, *11*, 3597. [CrossRef]
18. Di Crescenzo, A.; Paraggio, P. Modelling the random spreading of fake news through a two-dimensional time-inhomogeneous birth-death process. *Math. Methods Appl. Sci. arXiv* **2024**, arXiv:2405.06123. [CrossRef]
19. Román-Román, P.; Serrano-Pérez, J.J.; Torres-Ruiz, F. Some notes about inference for the lognormal diffusion process with exogenous factors. *Mathematics* **2018**, *6*, 85. [CrossRef]
20. Gutiérrez, R.; Rico, N.; Román, P.; Torres, F. Approximate and generalized confidence bands for some parametric functions of the lognormal diffusion process with exogenous factors. *Sci. Math. Jpn.* **2006**, *64*, 313–330.
21. Romero, D.; Rico, N.; García-Arenas, M. A new diffusion process to epidemic data. In Proceedings of the Computer Aided Systems Theory—EUROCAST 2013, Las Palmas de Gran Canaria, Spain, 10–15 February 2013; Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 69–76. [CrossRef]
22. Román-Román, P.; Torres-Ruiz, F. A stochastic model related to the Richards-type growth curve. Estimation by means of simulated annealing and variable neighborhood search. *Appl. Math. Comput.* **2015**, *266*, 579–598. [CrossRef]
23. Di Crescenzo, A.; Spina, S. Analysis of a growth model inspired by Gompertz and Korf laws, and an analogous birth-death process. *Math. Biosci.* **2016**, *282*, 121–134. [CrossRef]
24. Di Crescenzo, A.; Paraggio, P.; Román-Román, P.; Torres-Ruiz, F. Statistical analysis and first-passage-time applications of a lognormal diffusion process with multi-sigmoidal logistic mean. *Stat. Pap.* **2023**, *64*, 1391–1438. [CrossRef]
25. Eriksson, J.G.; Sandboge, S.; Salonen, M.K.; Kajantie, E.; Osmond, C. Long-term consequences of maternal overweight in pregnancy on offspring later health: Findings from the Helsinki Birth Cohort Study. *Ann. Med.* **2014**, *46*, 434–438. [CrossRef]
26. Michikawa, T.; Nitta, H.; Nakayama, S.F.; Ono, M.; Yonemoto, J.; Tamura, K.; Suda, E.; Ito, H.; Takeuchi, A.; Kawamoto, T. The Japan Environment and Children's Study (JECS): A Preliminary Report on Selected Characteristics of Approximately 10000 Pregnant Women Recruited During the First Year of the Study. *J. Epidemiol.* **2015**, *25*, 452–458. [CrossRef]
27. Olsen, J.; Melbye, M.; Olsen, S.F.; Sørensen, T.I.; Aaby, P.; Andersen, A.M.N.; Taxbøl, D.; Hansen, K.D.; Juhl, M.; Schow, T.B.; et al. The Danish National Birth Cohort - its background, structure and aim. *Scand. J. Public Health* **2001**, *29*, 300–307. [CrossRef]
28. Eskenazi, B.; Bradman, A.; Gladstone, E.A.; Jaramillo, S.; Birch, K.; Holland, N. CHAMACOS, A Longitudinal Birth Cohort Study: Lessons from the Fields. *J. Child. Health* **2003**, *1*, 3–27. [CrossRef]

29. Victora, C.G.; Barros, F.C.; Lima, R.C.; Behague, D.P.; Gonçalves, H.; Horta, B.L.; Gigante, D.P.; Vaughan, J.P. The Pelotas birth cohort study, Rio Grande do Sul, Brazil, 1982–2001. *Cad. Saúde Pública* **2003**, *19*, 1241–1256. [CrossRef]
30. Guxens, M.; Ballester, F.; Espada, M.; Fernández, M.F.; Grimalt, J.O.; Ibarluzea, J.; Olea, N.; Rebagliato, M.; Tardón, A.; Torrent, M.; et al. Cohort Profile: The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project. *Int. J. Epidemiol.* **2011**, *41*, 930–940. [CrossRef]
31. Iñiguez, C.; Esplugues, A.; Sunyer, J.; Basterrechea, M.; Fernández-Somoano, A.; Costa, O.; Estarlich, M.; Aguilera, I.; Lertxundi, A.; Tardón, A.; et al. Prenatal exposure to NO_2 and ultrasound measures of fetal growth in the Spanish INMA Cohort. *Environ. Health Perspect.* **2016**, *124*, 235–242. [CrossRef]
32. Larsen, P.S.; Kamper-Jørgensen, M.; Adamson, A.; Barros, H.; Bonde, J.P.; Brescianini, S.; Brophy, S.; Casas, M.; Devereux, G.; Eggesbø, M.; et al. Pregnancy and Birth Cohort Resources in Europe: A Large Opportunity for Aetiological Child Health Research. *Paediatr. Perinat. Epidemiol.* **2013**, *27*, 393–414. [CrossRef] [PubMed]
33. García-Burgos, A.; González-Alzaga, B.; Giménez-Asensio, M.J.; Lacasaña, M.; Rico-Castro, N.; Romero-Molina, D. Growth Curves Modelling and Its Application. *Eng. Proc.* **2023**, *39*, 66. [CrossRef]
34. Rico, N.; Romero, D.; García-Arenas, M. Comparing Some Estimate Methods in a Gompertz-Lognormal Diffusion Process. In Proceedings of the Computer Aided Systems Theory—EUROCAST 2013, Las Palmas de Gran Canaria, Spain, 10–15 February 2013; Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 499–506. [CrossRef]
35. Román-Román, P.; Serrano-Pérez, J.J.; Torres-Ruiz, F. First-passage-time location function: Application to determine first-passage-time densities in diffusion processes. *Comput. Stat. Data Anal.* **2008**, *52*, 4132–4146. [CrossRef]
36. Román-Román, P.; Torres-Ruiz, F. The nonhomogeneous lognormal diffusion process as a general process to model particular types of growth patterns. In *Lecture Notes of the Seminario Interdisciplinare di Matematica*; Università degli Studi della Basilicata: Potenza, Italy, 2016; Volume 12, pp. 201–219.
37. Pors, S.E.; Nikiforov, D.; Cadenas, J.; Ghezelayagh, Z.; Wakimoto, Y.; Jara, L.A.Z.; Cheng, J.; Dueholm, M.; Macklon, K.T.; Flachs, E.M.; et al. Oocyte diameter predicts the maturation rate of human immature oocytes collected ex vivo. *J. Assist. Reprod. Genet.* **2022**, *39*, 2209–2214. [CrossRef] [PubMed]
38. Henningsen, A.; Toomet, O. maxLik: A package for maximum likelihood estimation in R. *Comput. Stat.* **2011**, *26*, 443–458. [CrossRef]
39. Karunasingha, D.S.K. Root mean square error or mean absolute error? Use their ratio as well. *Inf. Sci.* **2022**, *585*, 609–629. [CrossRef]
40. Thomas, T.; Rajabi, E. A systematic review of machine learning-based missing value imputation techniques. *Data Technol. Appl.* **2021**, *55*, 558–585. [CrossRef]